JMIR Medical Education

Impact Factor (2024): 3.2
Volume 12 (2026) ISSN 2369-3762 Editor-in-Chief: Blake J. Lesselroth, MD, MBI, FACP, FAMIA

Contents

Original Papers

Effects of a Generative AI Virtual Patient on Serious Illness Communication Skills: Randomized Controlled Trial (e93034)
Kurtis Haut, Masum Hasan, Thomas Carroll, Ronald Epstein, Taylan Sen, Ehsan Hoque. 22

Students' Perspectives on Digital Psychotherapy—Possible Solutions for Digital Inpatient-Like Care Concepts: Qualitative Interview Study (e82830)
Rebekka Robitzsch, Tania Lalgi, Alexander Diel, Sophie Schulz genannt Menningmann, Lucy Gresser, Patrick Wollenberg, Martin Teufel, Alexander Bäuerle, Anita Robitzsch. 40

AI- vs Human-Based Assessment of Medical Interview Transcripts in a Generative AI–Simulated Patient System: Cross-Sectional Validation Study (e81673)
Hiromizu Takahashi, Kiyoshi Shikino, Takeshi Kondo, Yuji Yamada, Yoshitaka Tomoda, Minoru Kishi, Yuki Aiyama, Sho Nagai, Akiko Enomoto, Yoshinori Tokushima, Takahiro Shinohara, Fumiaki Sano, Takeshi Matsuura, Rikiya Watanabe, Toshio Naito. 56

Evaluating AI-Generated Geriatric Case Studies for Interprofessional Education: Systematic Analysis Across 5 Platforms (e83085)
Nicole Ruggiano, Sudikshya Sahoo, Ava Brashear, Uche Nwatu, Amie Brunson, Hyunjin Noh, Heather Cole, Robert McKinney, C Framil Suarez, Ellen Brown, Suzanne Prevost. 65

Real-World Impact and Educational Effectiveness of an AI-Powered Medical History-Taking System: Retrospective Propensity Score-Matched Cohort Study (e89367)
Yang Liu, Yiying Zhu, Weishan Zhang, Xian Lu, Liping Wu, Minghui Yue, Oudong Xia, Chujun Shi. 79

AI-generated Feedback Following Social Robotic Virtual Patient Interactions and Medical Student Performance: Nonrandomized Quasi-Experimental Study (e90368)
Alexander Borg, Jonathan Schiött, William Ivegren, Cidem Gentline, Viking Huss, Anna Hugelius, Benjamin Jobs, Mini Ruiz, Samuel Edelbring, Carina Georg, Gabriel Skantze, Ioannis Parodis. 97

User Acceptance of an AI-Powered Medical History–Taking Training System Among Undergraduate Medical Students: Mixed Methods Study (e101425)
Yang Liu, Yiying Zhu, Chujun Shi, Xian Lu, Liping Wu, Minghui Yue, Xiaolin Hong, Oudong Xia, Weishan Zhang. 109

Effect of Large Language Model–Powered Virtual Standardized Patients on History-Taking Among Undergraduate Medical Students: Propensity-Matched Cohort Study (e92486)
Yuchen He, Chen Chen, Rong Yin, Wuyang Zhang, Haoli Chang, Wei Yang, Fei Li, Xinhua Li, Zhuying Xia, Xiaoyun Xie, Jing Huang, Qiuming Zeng, Guang Yang, Jing Wu. 128

XSL•FO
RenderX

Perceived Ratings of AI-Generated and Expert-Reviewed vs Clinician-Developed Virtual Cardiology Patient Scenarios Among Medical Students: Randomized Controlled Study (e100839)
Marios Bantidos, Eleni Dafli, Efstratios Karagiannidis, Barbara Fyntanidou, Sotiria Gilou, Athina Nasoufidou, Panagiotis Bamidis. 145

Association of Preexamination AI Virtual-Patient Trainer Scores With Scores on a Local Single-Station History-Taking Objective Structured Clinical Examination: Observational Cohort Study (e109752)
Weishan Zhang, Yang Liu, Yiying Zhu, Xian Lu, Liping Wu, Minghui Yue, Xiaolin Hong, Oudong Xia, Chujun Shi. 160

A Competency Assessment Tool for Liver Contrast-Enhanced Ultrasound Among Sonography Trainees: Development and Multicenter Validation Study (e86730)
Yanping Ma, Bowen Zheng, Jie Ren, Wen He, Lei Shang, Hui Wang, Gaiqin Xue, Fang Nie, Yanli Guo, Yukun Luo, Hongxia Zhang, Xuan Wang, Xiang Jing, Ruifang Zhang, Jianjun Yuan, Rong Wu, Xiaomao Luo, Fenglin Dong, Yan Luo, Qiang Lu, Jintang Liao, Pintong Huang, Qin Chen, Jia'an Zhu, Youfeng Xu, Xiaoyan Xie. 831

The Design and Evaluation of an Online Continuing Medical Education App for Medical Professionals in China: Quantitative Study (e76299)
Xu Zhang, Xianying He, Yuntian Chu, Dongqing Liu, Minzhao Lyu, Weiyi Wang, Haotian Chen, Meihao Ji, Fangfang Cui, Jie Zhao. 849

Global Learner Feedback on Continuing Medical Education–Accredited e-Learning Modules in Pediatric Endocrinology and Diabetes: Cross-Sectional Study (e67332)
Yvonne van der Zwan, Conny van Wijngaard-de Vugt, Abdulsalam Abu-Libdeh, Evangelia Kalaitzoglou, Zacharoula Karabouta, Stenvert Drop, Annemieke Boot, Sze Ng, Jan Idkowiak. 863

Upskilling in Healthy Longevity Medicine and Its Association With Physicians' Implementation Intent and Self-Reported Clinical Confidence: Cross-Sectional Observational Study (e83779)
Evelyne Bischof, Dominika Wilczok, James Kirkland, Bhirau Wilaksono, Christine Huang, Suwanna Suwannaphong, Wanviput Sanphasitvong, Dalila amdži , Carolina Hernandez, Yoko Madea, Hidekazu Yamada, Melissa Fernandes, Ricardo Gaminha Pacheco, Fabiano Serfaty, Fernanda Calvo-Fortes, Amit Goldman, Andrea Maier, Alexey Moskalev, Morten Scheibye-Knudsen, Alex Zhavoronkov. 876

Experiential Course Learning, Wellness, and Higher Education: Qualitative Descriptive Study (e88642)
Adrianna Watson, Neil Peterson, Brandon Thatcher, Michael Thomas, Stacie Hunsaker, Cole Hooley, David Erekson, Adam Simpson, Gregory Snow, Rachel Detrick. 887

Efficacy of an Imaging Anatomy Virtual Simulation Experiment Course in Undergraduate Medical Education: Nonrandomized Controlled Trial (e80393)
ACeng Li, Lina Ma, Rui Peng, Juan Shi, Mingming Zhang, Minwen Zheng, Yayun Wang, Yayu Huang, Jing Ren. 901

Experiences and Perceptions of Clinical and Graduate Medical Students Regarding AI in Syria: Cross-Sectional Study (e84942)
Jouni Aljoudeh, Abdulrahman Al Balkhi, Yahia Ranjous, Abdulrahman Shbani, Danial Takieddin, Mohammad Bashar Izzat. 915

Changes in Self-Reported Empathy After a Basic Medical Communication Course Among Premedical Students: Single-Group Longitudinal Study (e92215)
Seong Jeon, Hyo Yoo. 933

Impact of Community-Oriented Medical Education on Medical Students' Perceptions of Community Health Care: Qualitative Study (e84406)
Kiyoshi Shikino, Kazuyo Yamauchi, Nobuyuki Araki, Naoto Ozaki, Yu Kamata, Shinya Aoki, Yota Katsuyama, Daichi Sogai, Mai Miyamoto, Kensuke Yoshimura, Takeshi Oki, Shoichi Ito. 945

Comparative Efficacy of Simulation-Based and Traditional Training in Ultrasound-Assisted Regional Anesthesia for Medical Students: Randomized Controlled Trial (e77702)
David Sánchez-Poveda, Juan Vicente-Mampel, Belén Curto, Vidal Moreno, Juan García-Esteban, Felipe Hernández-Zaballos, Pablo Alonso-Hernández. 959

XSL•FO
RenderX

Effectiveness of Artificial Intelligence–Assisted Peer Teaching in Orthopedic Clinical Education: Historical Cohort Study (e87959)
Chengcheng Yu, Fangcai Li, Ning Zhang, Hejia Hu, Hanxu Huang, Jingkai Wang, Yiqing Tao, Yinan Wu. 970

Collaborative Simulation-Based Learning to Develop Competencies in Physical Therapy Students: Randomized Controlled Trial (e85442)
Iván Cuyul-Vásquez, Hilda Aravena-Páez, Felipe Araya-Quintanilla, Sergio Guinez-Molinos, Rodrigo Núñez-Cortés, Héctor Gutiérrez Espinoza.
9 8 5

Prescription Support Practice for Pharmacy Students: Pre-Post Educational Intervention Study (e79545)
Fuka Aizawa, Kenta Yagi, Tsukasa Higashionna, Hirofumi Hamano, Shimon Takahashi, Yoshito Zamami, Kazuaki Shinomiya, Takahiro Niimura, Mitsuhiro Goda, Kei Kawada, Keisuke Ishizawa. 998

Youth-Elder Co-Learning Model in Psychiatric Long-Term Care Education: Mixed Methods Evaluation of Communication and Empathy Outcomes (e82812)
Ke-Hsin Chueh, Jeng-Wen Chen, Li-Ang Lee, Hsiang-I Hsu, Yen-Ju Lin, Hsiao-Tzu Wang. 1011

Andragogic Model Curriculum for One-Year ACGME-Accredited Fellowship Programs: Single-Center Educational Improvement Project (e81570)
Romy Yun, Allyson Farran, Echo Rowe, Dina Sheira, Christian Jackson, Thomas Caruso. 1026

Evaluating Large Language Model–Based Automated Scoring in a Voice-Based Virtual Standardized Patient Platform for Medical Students: Cross-Sectional Agreement Study (e95578)
Xiaoxing Gao, Xiaoming Huang, Rongrong Hu, Li Zhang, Huiting Liu, Bingqing Zhang, Chong Wei, Wei Qiu, Mengyu Zhang, Xuefeng Sun. . .
1 0 3 3

Teaching Acute Coronary Syndrome High-Risk ECG Interpretation and Clinical Decision-Making Through FOAMed Videos and Podcast Versus Print-Based Materials Among Emergency Care Providers: Randomized Controlled Mixed Methods Trial (e87587)
Wiebke Turner, Klaus Fessele, Martin Fandler, Jan Ehlers, Julia Nitsche. 1046

Influence of Admission Pathways on Learning Strategies, Assessment Engagement, and Academic Performance Among First-Year Medical Students: Mixed Methods Retrospective Observational and Cross-Sectional Survey Study (e68636)
Issarawan Keadkraichaiwat, Chantacha Sitticharoon, Punyapat Maprapho, Nisa Jangboon, Nadda Wannarat. 1066

Developing and Validating a Coding Scheme for Clinical Reasoning in History Taking Using Generative AI–Based Virtual Patients: Systematic Text Condensation Approach (e84347)
Naping Chen, Luzhen Tang, Yang Liu, Changmin Lin, Zijian Li, Chujun Shi, Mengyu Xia, Dragan Gasevic, Danijela Gasevic, Jinbin Zheng, Yizhou Fan, Xinyu Li. 1093

“Optimizing Learning in Integrated Curriculum”—Comparative Effectiveness of Online and Face-to-Face Formative Assessments: Mixed Methods Study (e84935)
Amira Ismail, Sarah Hussein, Mohamed El-Shafey, Ahmed Al-Neklawy, Ahd Mansour, Shatha Felemban, Shimaa Elaraby. 1107

Virtual Reality and Gamification for Assessing Technical Aptitude, Cognitive Abilities, and Personality Characteristics in Surgical Residency Selection: Validation Study (e82515)
Noa Gazit, Gilad Ben-Gal, Ron Eliashar. 1120

Ambiguity Detection in Medical Exams via Large Language Models: Retrospective Cross-Sectional Pilot Study (e82702)
Romain Lombardi, Alexandre Destere, Jean Dellamonica, Alexandre Gérard, Mathieu Jozwiak. 1144

Exploring the Feasibility of an Examiner-Worn Neck-Mounted Camera for Objective Structured Clinical Examination Assessment: Pilot Feasibility Study (e87483)
Miwa Sekine, Yuji Nishizaki, Amane Endo, Takasuke Ogawa, Yoshihide Takeshita, Motomi Nasu, Hiroo Wada, Chizuko Miyamoto, Michiko Oguro, Chie Otake, Yasuhiko Konishi, Yuichi Tomiki. 1157

XSL•FO
RenderX

Learning Analytics of a National Entrustable Professional Activities Platform: Cross-Sectional Study of System-Level Constraints on Advanced Entrustment in Competency-Based Medical Education (e95066)
Jeng-Wen Chen, Rui-Bin Yu, Kai-Nan Lin, Yukiko Ono, Xuan-Bing Yu, Wei-Chung Hsu, Pa-Chun Wang, Li-Jen Liao, Chun-Hou Liao, Mingchih Chen. . . . 1170

Artificial Intelligence for Assessment and Feedback in Medical Education: Bibliometric Mapping Study and Thematic Evidence Map (e98949)
Zihang Zhao, Zihan Liu, Liang Guo, Teng Pan, Yousheng Zhang, Chenxiang Miao, Yiting Ge, Yipeng Wang, Xin Hu, Xin Wang, Ruipeng Zhang, Zhiyong Hou. . . . 1186

Evaluating the Pediatric Behavior Guidance of Students Based on Actual Clinical Transcripts Scored by Faculty and Large Language Models: Pilot Comparative Study (e83376)
Ishreen Dhillon, Gabriel Lee, Shijia Hu. . . . 1204

Evaluation of the Bridge-in, Objective, Preassessment, Participatory Learning, Postassessment, and Summary–Based Instructional Model in Disaster Nursing: Quasi-Experimental Study (e77728)
Li Pei, Taoyu Jia, Yaxin Liu, Xiao-li Pang, Muzi Li, Haoying Dou. . . . 1211

The Effect of a Traditional Chinese Medicine Course on Western Medicine Students' Attitudes Toward Traditional Chinese Medicine: Self-Controlled Pre-Post Questionnaire Study (e55972)
Haoyu He, Hanjin Cui, Li Guo, Yanhui Liao. . . . 1229

Interactive, Image-Based Modules as a Complement to Prosection-Based Anatomy Laboratories: Multicohort Evaluation (e85028)
Caroline Sumner, Sami Case, Samuel Franklin, Kristen Platt. . . . 1240

Disaster Medicine Training for Medical Students in Lebanon: Quasi-Experimental Comparison of e-Learning and Face-to-Face Modalities (e80409)
Ali Msheik, Ruben Peralta, Zeinab Al Mokdad, Christine Bartulec, Linda Abou-Abbas, Dima Anani, Hussein Al Mawla, Hassan Salame, Mohamad Moussa, Awsan Bahattab. . . . 1250

Dependability of Entrustable Professional Activity Portfolios in Otorhinolaryngology Residency: Nationwide Generalizability and Decision Study (e104628)
Jeng-Wen Chen, Ching-Lin Hsieh, Chun-Hsiang Chang, Ya-Hui Wang, Wei-Chung Hsu, Pa-Chun Wang, Li-Jen Liao, Chun-Hou Liao, Mingchih Chen, Okki Laksmita. . . . 1266

Attitudes Toward and Preconditions for Digital Game–Based Learning Among Public Health and Health Sciences Students: Cross-Sectional Online Survey Study (e83347)
Kamil Wrona, Joanna Albrecht, Anne Krümmel, Leona Aschentrup, Christoph Dockweiler, Timothy Mc Call. . . . 1282

Impact of a Structured Training Program on Medical Student Confidence and Behavior During Their First Radial Arterial Puncture: Comparative Study (e78086)
Camille Rolland-Debord, Lucien Juret, Mathilde Simon, Abdallah El Mouhajer, Cécile Londner, Capucine Morélot-Panzini, Cécile Chenivesse, Thomas Similowski. . . . 1294

Perceptions and Attitudes of Medical Students Toward the Integration of Large Language Models in Medical Education: Cross-Sectional Survey in China (e66381)
Cheng Zhao, Weiqian Yan, Long Wang, Jing Wu, Herve Ndikuriyo, Renhe Yu. . . . 1305

Questionnaires on Perceptions of Artificial Intelligence in Health Care Among Health Care Students: Cross-Cultural Translation Into French and Linguistic Validation (e76572)
Sylvain Kotzki, Calvin Massonnet Turner, Nicolas Vuillerme. . . . 1319

XSL•FO
RenderX

Exploring Ambient Artificial Intelligence to Enhance Learning and Feedback During Operating Room-to-Intensive Care Unit Handoffs: Co-Design and Simulation Study (e85666)
Laleh Jalilian, Federico Barra, Tristan Grogan, Jason Lee, Achuta Kadambi. 1328

Medical Students' Experiences With an Integrated Surgical Suturing Training Module Using Simulation Models and Asynchronous Videos: Interpretive Qualitative Study (e90563)
Álvaro Herrera Alcaíno, Consuelo Vade Martínez, Benjamín Painemal Rivera, Daniela Serey Torres, Sebastián Puentes Bravo, Ignacio Leal Lizama, Josefina Darrigol Parra, Camila Riquelme Bahamondes. 1344

Enhancing Engagement With Endocrine Guidelines and Fostering Medical Student Interest Through Concise Medical Information Cines: Qualitative Co-Design Study (e83711)
Rahul Sagu, Sangamithra Ravi, Aspasia Manta, Punith Kempegowda. 1352

Web-Based Virtual Environment Versus Face-To-Face Delivery for Team-Based Learning of Anesthesia Techniques Among Undergraduate Medical Students: Randomized Controlled Trial (e80097)
Darunee Sripadungkul, Suhattaya Boonmak, Monsicha Somjit, Narin Plailaharn, Wimonrat Sriraj, Polpun Boonmak. 1360

Virtual Reality in Training and Assessment Among Clinical Students and Lecturers at a Nigerian University: A Phenomenological Study (e75021)
Abiola Komolafe, Omotade Ijarotimi, Oluseye Okunola, Olufemi Adetutu, Ayodeji Oluwatope, Olatunde Abiona, Ojo Agunbiade, Adeboye Ayinde, Stephen Aregbesola, Olawale Akinwale, Babatope Kolawole, Lanre Idowu, Alaba Adediwura, Olayinka Otuyemi. 1372

Medical Students' Experiences With Virtual Reality Simulation Training: Qualitative Study (e74301)
Vanshika Sharma, Sohee Park, Alexandra Voinescu, Chris Jacobs. 1384

e-Learning, Distance Education, and Virtual and Augmented Reality in Orthopedic Training: European Cross-Sectional Survey of Trainee Acceptance Guided by the Technology Acceptance Model and Unified Theory of Acceptance and Use of Technology (e79418)
Adam Schlegl, Marko Ostoji , Ines Unterfrauner, Gianluca Ciolli, Panayiotis Megaloikonomos, Vasilios Igoumenou, Michele Mercurio, Thomas Stark, David Kordic, Martijn Dietvorst, Filipe Santos, Viktória Nyakas, Luca Tóth, András Komócsi, Péter Maróti, András Matuz. 1401

Live Podcasting as an Educational Intervention in Dentomaxillofacial Radiology: Controlled Cohort Study (e77980)
Anna-Lena Hillebrecht, Daniel Fritzsche, Thamar Voss, Anne Kruse, Andreas Keßler, Kirstin Vach, Markus Altenburger, Rainer Schmelzeisen, Wiebke Semper-Hogg. 1419

Toward Human-Centered Pedagogy in the Age of Generative AI Among Higher Education Educators: Qualitative Phenomenological Study (e105172)
Zuheir Khlaif, Bilal Hamamra. 1430

Medical Students' Acceptance of Digital Entrustable Professional Activities: Results of a Cohort Study (e87605)
Maximilian Domann, Constanze Richters, Matthias Witti, Matthias Stadler. 1445

Olfactory Cues and Patient Safety Behavior in Virtual Reality Medical Simulation: Controlled Trial (e93148)
Anna Junga, Pascal Kockwelp, Jennifer Dabel, Sönke Scherzer, Benjamin Risse, Markus Holling, Bernhard Marschall, Hendrik Friederichs. .
1 4 5 5

Ambient AI Scribes to Create Educational Feedback Notes for Medical Students: Randomized Trial (e89996)
Jaideep Talwalkar, David Chartash, Lisa Zhang, Michael Makutonin, Conrad Safranek, Anne Sidamon-Eristoff, Lee Schwamm, Donald Wright.
1 4 6 8

Development of Therapeutic Competencies in Health Care Students: Qualitative Focus Group Study Using 360-Degree Video and Virtual Reality Technology (e75776)
Camilla Lauritzen, Charlotte Reedtz, Kjersti Kristensen, Vår Mathisen, Eva Næss, Rigmor Furu, Hege Nermo, Rita Jentoft. 1480

XSL•FO
RenderX

Comparing the Weighted Gain Score and a Rasch-Based Approach for Estimating Learning Outcomes in Medical Education: Quantitative Study (e75516)
Rauf Aliyev, Joy Backhaus, Silke Hammer, Sarah König. . . . 1490

Evaluating the Influence of MCAT Scores on Medical Student Selection and Performance Using a Machine Learning Approach: Retrospective Cohort Study (e94197)
Michele Cherfane, Marc Ghanem, Georges Choueiry, Selim Nasser, Tamina Elias-Rizk, Sola Aoun Bahous. . . . 1501

What Platform Scores Miss: Multidimensional Evaluation of AI Teaching Agents in Medical Education (e96819)
Hui Zhang, Lihui Qu, Jianmin Zheng, Yi Xiong, Hongbo Bai, Ruiying Ji, Guohui Liu, Wanling Chen, Zirui Cheng, Youbang Chen, Chun-tao Yang. . . . 1517

Implementation and Evaluation of a Cancer Immunotherapy Elective for Medical Students: Mixed Methods Descriptive Study (e71628)
Mark Raynor, Rivers Hock, Brandon Godinich, Satish Maharaj, Houriya Ayoubieh, Cynthia Perry, Jessica Chacon. . . . 1529

Effect of an Online Continuing Professional Development Course on Physicians' Intention to Approach a Colleague in Difficulty: Mixed Methods Convergent Study (e80199)
Florence Lizotte, Martin Tremblay, Caroline Biron, Éloi Lachance, Souleymane Gadio, Roberta de Carvalho Corôa, Claude Uwizeye, Sam Daniel, France Légaré. . . . 1546

Impact of Hands-On Pediatric Emergency Training on Wound Suturing, Airway Management, and Orthopedic Casting Skills: Prospective Pre-Post Interventional Study (e86349)
Saeed Alzahrani, Nasser Assiri, Fayez Alharthi, Tariq Al Garni, Asim Alamri, Bader Asiri, Maher Alkuwaiti, Raed Habib. . . . 1567

Developing a Realistic and Cost-Effective Training Model (MaiSurge) for Laparoscopic Hysterectomies to Train and Assess Surgical Skill: Prospective Nonrandomized Controlled Trial (e66369)
Anna Brechter, Roxana Schwab, Christoph Dold, Christine Skala, Maria Schröder, Lina Schiestl, Katharina Gillen, Walburgis Brenner, Annette Hasenburg, Mona Schmidt. . . . 1575

Targeted Educational Intervention Through Game-Based Learning to Promote Rational Antimicrobial Use Among Health Care Learners: Prospective Interventional Study (e72236)
Sumana MN, Supreeta Shettar, Yogeesh Maheshwarappa, GK Megha, Veerabhadra GS, Chinchana Eshwar, Shruthi SC. . . . 1586

Enhancing Therapist Training in the Delivery of Exposure Therapy for Individuals with Anxiety Disorders Using Virtual Reality Simulation: Randomized Feasibility Trial (e80087)
Joshua Kemp, Ariella Rosen, Hyungjin Kim, Margo Adams Larsen, Kristen Benito, Jennifer Freeman, Jason Machan, Peter Tuerk. . . . 1594

Speech- and Text-Based Emotion Recognition in Anesthesiology Residents During Critical Incident Simulation Training: Exploratory Observational Study (e96628)
Sapir Gershov, Itay Bentov, Fadi Mahameed, Aeyal Raz, Shlomi Laufer. . . . 1616

Attitudes Toward Digital Planetary Health Before and After a Lecture Series on Environmental Awareness and Future Visions: Repeated Cross-Sectional Survey Study Regarding Twin Transformation (e89259)
Julia Nitsche, Theresa Busse, Sonja Schmalen, Ina Otte, Jan Ehlers. . . . 1630

The Historical Evolution of Microlearning in Health Professions Education: Bibliometric Analysis (e91616)
Jennie De Gagne, Randall, Elizabeth Blackwood, Katherine Carlson, Hee Kang. . . . 1644

Digital Training for Lay Health Care Workers' Knowledge and Skills in HIV Index Case Testing: Cluster Randomized Trial (e89942)
Tapiwa Tembo, Nora Rosenberg, Katie Mollan, Maria Kim, Sarah Rutstein, Angella Mkandawire, Mike Chitani, Caroline Kumbuyo, Duncan Phiri, Mtisunge Mphande, Elijah Kavuta, Samuel Chilala, Jiayu Wang, Saeed Ahmed, Katherine Simon, Linda-Gail Bekker. . . . 1657

XSL•FO
RenderX

Digital Simulation–Based Ultrasound Training for Physiotherapy Students: Prospective Randomized Controlled Trial (e87897)
Samuel Fernández-Carnero, Belen Diaz-Pulido, Jorge Méndez-Rodriguez, Daniel Pecos-Martin, Santiago Garcia-Miguel, Alexander Achalandabaso-Ochoa, Nicolas Cuenca-Zaldívar, Fermín Naranjo-Cinto Sr. . . . 1670

Combining Problem-Based Learning Methods With the WeChat Platform in Teaching Ophthalmology: Randomized Controlled Trial (e65279)
Fang Fang, Bing Bu, Wenmin Jiang. . . . 1684

Ultrasound-Guided Regional Anesthesia in a Resource-Limited Hospital: Prospective Pilot Study of a Hybrid Training Program (e84181)
Jakob Gamboa, Inge Tamm-Daniels, Roland Flores, Nancy Sarat Diaz, Mario Villasenor, Mitchell Gist, Aidan Hoie, Christopher Kurinec, Colby Simmons. . . . 1696

Development of Virtual Reality Scenarios Addressing Gender-Based Violence in Health Sciences Education: Qualitative Approach (e76098)
Belén Valverde-Alirangues, Marta Benet, Mar Carrió. . . . 1710

Implementation of Emotional Connection Training in Pediatric Primary Care: Mixed Methods Study (e81250)
Elizabeth Erickson, J Amati, Dani Dumitriu, Anna Miller-Fitzwater, Debra Best, Doreet Preiss, Elena Uejo, Adrian Brown, Jennifer Crotty, Leah Gillen, Cherece Grier, Kristina Gustafson, Nicole Shearman, Andréane Lavallée, Jennifer Warmingham, Elena Arduin, Nikki Shearman, Jackie Young, Jessica Sperling. . . . 1723

Exploring the Role of Early Career Medical Professionals From a Digital-Oriented University in Germany in Promoting Digital Health in Professional Settings: Qualitative Interview Study (e86107)
Christopher Koch, Jan Ehlers, Julia Nitsche, Theresa Busse. . . . 1739

Principles for Responsible AI in Health Professions Education, Research, and Care: Health CARE-AI (Contextual, Accountable, Responsible, and Equitable Artificial Intelligence) Framework Delphi Consensus Study (e91626)
Lyn Sonnenberg, David Wiljer, Muhammad Mamdani, Victor Do, Brandon Tang, Babar Haroon, Bizav Jaffer, Jerry Maniate. . . . 1756

Generative AI–Assisted Progressive-Disclosure Case-Based Learning for Clinical Reasoning in Occupational Medicine: Quasi-Experimental Study (e103584)
Peng Su, Min Hu, Chengzhi Chen, Shangcheng Xu. . . . 1768

Development of a Deep Learning–Based Feedback Model to Assist Medical Students Learning Renal Ultrasound Acquisition: Mixed Methods Study (e72110)
Andy Hwang, Rahul Singh, Elizabeth Barrett, Peng Cao, Varut Vardhanabhuti, Pauline Ng, Gordon Wong, Michael Co, Elaine Lee. . . . 1784

Mobile Learning in Medical Education: Quasi-Experimental Realist Evaluation of Usage, Context, and Examination Performance in a Curricular Setting (e85892)
Jan Holzhäuer, Joy Backhaus, Thiên-Tri Lâm, Sarah König. . . . 1795

Mobile Medical Education Apps as Digital Health Information Resources in China and the United States: Longitudinal App Store Metadata Analysis and Topic Modeling of User Review Narratives (e107560)
Junhao Ma, Tian Shen, Peng Xiang, Yingjun Li, Yuanshuai Chen, Zhengyang Lu, Jingkai Zeng, Wei Dai, Hanwen Pan, Liping Yang, Yunfan He, Xingxi Li, Joyce Wu, Qi Yu, Jun Liang, Jianbo Lei. . . . 1811

Learning Participation and Performance Among Social and Academic Learners in a Nursing Massive Open Online Course: Retrospective Cohort Study (e90827)
Jianqing Zeng, Kun Li, Wenxuan Wang, Lingjun Xiao, Yingmin Wang, Xinghui He, Yonghong Qing, Jingjing Xue, Lirong Zhong, Tiebin Yan, Xiuyuan Zheng. . . . 1834

Adoption and Use of Social Media in Health Care Among Medical Residents: Cross-Sectional Study (e83475)
Zineb Zouakia, Aurélia Manns, Thomas Meyre, Anita Burgun, Natacha Kadlub, Rosy Tsopra. . . . 1844

XSL•FO
RenderX

Instagram Video Engagement in Medical Education: Cross-Sectional Study (e79124)
Aapeli Kemppainen, Henrik Nuutinen, Virve Koljonen, Petri Kulmala. . . . 1859

Exploring the Icarus Paradox in Indonesia's Specialist Medical Education System Using the Public Perspective From Online Media: Convergent Mixed Methods Study (e60452)
Faisal Binsar, Mohammad Hamsal. . . . 1872

AI-Enhanced Continuing Professional Development as an Evolving Sociotechnical System: Multimethod Theoretical Framework Development Study (e69156)
Vjekoslav Hlede, Sofia Valanci, G D'Antuono, Heather Dow, Ronan O'Beirne, Richard Wiggins. . . . 1892

The Relationship Between Sense of Community and Collaborative Learning: Quantitative Study in Medical Education (e86893)
Harald Knof, Markus Berndt, Anne Herrmann-Werner, Bernhard Hirt, Thomas Shiozawa. . . . 1908

Design and Evaluation of a Faculty Development Workshop Series on Integrating Generative Artificial Intelligence in Medical Education: Mixed Methods Pilot Study (e89815)
Rajalakshmi Anand, Nicole Bowers, Mange Manyama. . . . 1921

A Competency Framework for Medical AI Education: Mixed Methods Study (e91116)
Chang Cai, Gaoxia Zhu, Shang-Ming Zhou, Olivia Ng, Jamie Duell, Weng Ho, Daisy Chen, Bernett Lee, Fang Li, Siyuan Liu, Vidya Sudarshan, Li Wang, Chanwoo Choi, Xiuiyi Fan. . . . 1937

Systemic Barriers to Curriculum Adaptation for Rapidly Changing Knowledge in Medical Education: Qualitative Study (e96244)
Martin Baumgartner, Matthias Stadler, Matthias Samwald, Martin Fischer. . . . 1956

An AI-Integrated Relational Curriculum Mapping Framework for Alignment and Transparency in Medical Education: Implementation Study (e107727)
Robert Badeau, Jennifer Laury, Danielle Dickey. . . . 1970

Simulation-Based Radiology Training for Medical Technical Assistant Trainees: Quasi-Experimental Study (e97489)
Matthias Witti, Joséphine-Charlotte Coleman, Constanze Richters, Johanna Schönberger, Mary-Anne de Ruijter, Alina Stortz, Melanie Bauer, Raisa Kiriakidou, Matthias Stadler. . . . 1986

Immersive Virtual Reality Training to Improve Novice Physicians' Emergency Response Skills: Randomized Controlled Trial (e71455)
Yeon-Ju Huh, Ju Kim, Narae Yoon, Seoi Jeong, Hyoun-Joong Kong, Sun Myung. . . . 1994

Application of Mixed Reality for Ophthalmic Clinical Skills and Diagnosis: Prospective Study (e71338)
Chun Tan, Wei Yong, Hui'En Lin, Jaslyn Oh, How Yong, Fang Khew, Liang Shen, Yujia Gao, Wei Kow, Yih Tham, Dianbo Liu, Ching-Yu Cheng, Kee Ngiam, Yew Yuen, Ray Manotosh, Eng Khoo, Teck Koh, Woon Tan. . . . 2009

Cost-Effectiveness of a Blended Learning Training Approach for Health Care Workers Conducting HIV Index Case Testing in Malawi: Secondary Economic Analysis of a Cluster-Randomized Controlled Trial (e95450)
Poonam Rawat, Nora Rosenberg, Katie Mollan, Tapiwa Tembo, Dhrutika Vansia, Jiayu Wang, Mike Chitani, Maria Kim, Ireen Chirombo, Caroline Kumbuyo, Victor Mwapasa, Katherine Simon, Saeed Ahmed, Sarah Rutstein. . . . 2021

Effects of an Immersive 180° Video–Based Virtual Reality Instructional Module on Learning Satisfaction, Technology Acceptance, Motivation, and Quiz-Based Learning Achievement in Undergraduate Physical Therapy Students: Randomized Controlled Trial (e94122)
Hyoshil Yoon, Yongwoo Lee. . . . 2036

XSL•FO
RenderX

Augmented Reality–Based Training for a Rapid Blood Transfusion Device Among Emergency Nurses: Randomized Controlled Trial (e98073)
Wooyoung Jang, Wooree Chai, Yeji Kim, Youjung Moon, Suyoung Yoo, Sun Na, Soyeon Shin, Meong Son. 2047

Development and Educational Evaluation of a Radiation Visualization System in a Virtual Reality Training Module for Radiopharmaceutical Administration: Mixed Methods Study of Students' Learning Experiences (e104397)
Akihiro Kakimoto, Daisuke Fujise, Shota Watanabe, Yutaka Otaka, Sakura Minemoto, Ayaka Yasumatsu, Shin Hasegawa. 2059

Refining the Immersive Technology Evaluation Measure for Virtual Reality Clinical Skills Among Native Arabic-Speaking Medical Students: Qualitative Study (e95904)
Afaf Alblooshi, Falah Almarzooqi, Taleb Almansoori, Anmol Punn, Marwa Alameen, Alexander Kieu, Faten AlRadini. 2074

Virtual Reality–Enhanced Training for Trauma-Informed Care Among Residential and Child Mental Health Professionals: Pre-Post Evaluation Study (e86543)
Kostadin Kostadinov, John Goodwin, Gunter Groen, Astrid Jörns-Presentati, Sinéad Heffernan, Áine O'Donovan, Maria O'Malley, Ryan Biskupovic Goulding, James O'Mahony, Stephanie Allen, Margaret Curtin, Satu Haapalainen, Jami Aho, Gergana Petrova, Svetla Ivanova, Valentina Lalova, Joonas Korhonen. 2086

Perspectives of Premedical Students at a Medical University College in Malaysia Regarding Augmented Reality Integration in Anatomy Education: Cross-Sectional Study (e64402)
Fadhilah Idris, Nur Mohd Noor, Faidruz Jam, Nurul Mohd Nor, Siti Mohamad Asri. 2102

Virtual Reality Versus In-Person Simulation of Sepsis for Medical Students: Randomized Comparative Pilot Study (e80316)
Lauren Medwell, Tim Old, Awais Ahmed, Victoria Holloway, Lauren McTaggart, Dafydd Morgan, Callum O'Keeffe, Jemaima Olori, Ellie Payne, Christopher Rainforth, Namritha Ramanujam, Kriti Vaidya, Chris Jacobs. 2121

Remote Augmented Reality Versus Traditional Simulation for Team Leader Assessment in a Cardiac Arrest Scenario: Noninferiority Randomized Controlled Trial (e84367)
Renan Gianotto-Oliveira, Marcos Rojas, Maria Queiroz, Flávia Zanchetta, Anabel Ferrari, Lucas Kojima, Alexandre Paula, Bruno Campos, Dario Cecilio-Fernandes, Thomas Caruso. 2134

Investigating the Impact of a Virtual Reality Experience on Medical Student Empathy: Mixed Methods Study (e76504)
Allen Mundok, Vivian Ho, Lauren Fowler, Ann Kennedy, Shannon Stark-Taylor. 2147

Enhancing Interprofessional Team Performance to Prevent Medication Errors in Emergency Care: Quasi-Experimental Study Using Multimodal Virtual Simulation-Based Interprofessional Education (e66999)
Ora-In Chu, Phanupong Phutrakool, Khrongwong Musikatavorn, Thitiporn Kongchim, Lapol Herabat, Jiraphan Ritsamdang, Krittin Bunditanukul, Kanittha Triamamornwooth, Khuansiri Narajeenron. 2159

Development and Initial Evaluation of Specific Immersive Competence in Virtual Reality–Based Medical Assessments: Exploratory Observational Study (e82136)
Verena Schreiner, Joy Backhaus, Marco Lindner, Melina Heinisch, Sarah König, Sebastian Oberdörfer, Tobias Mühling. 2177

Use of 3D-Printed Models and Augmented Reality in Medical Student Education of Congenital Heart Disease: Randomized Controlled Trial (e85967)
Tyler Langenfeld, Gabriel Bahrami, Yea-Lyn Pak, Susanne Wish-Baratz, Yuxi Zhu, Arpit Agarwal. 2191

Blended Learning Compared With Face-to-Face Learning Among Family Medicine Residents: Randomized Controlled Trial (e86387)
Pierre-Yves Meunier, Sophie Schlatter, Juliette Macabrey, Frédéric Zorzi, Thomas Colleony, Rémy Boussageon, Hubert Maisonneuve, Marion Lamort-Bouché. 2205

XSL•FO
RenderX

Students' Learning Outcomes in a Virtual Microscopic Online and Face-to-Face Anatomy Course: Cross-Over Comparative Study (e82012)
Katharina Sternecker, Christoph Schmitz, Markus Berndt. . . . 2218

Synchronous Remote Teaching for Neurological Examination Training in Fifth-Year Medical Students: Quasi-Experimental Study (e77034)
Chih-Shan Huang, Kung-Pei Tang, Ming-Shun Wu. . . . 2231

Fostering Inductive and Deductive Learning in Oral Microbiology and Immunology With a Dual-Role Duel Card Game: Explanatory Sequential Mixed Methods Study (e80048)
Kawin Sipiyaruk, Ratchapin Srisatjaluk, Ponpentip Chaisiri, Kanokwan Jantawatchai, Jidapha Srihatai, Chutikarn Sonyaem, Thanaphat Traikorakot, Suphasuta Pookboonmee. . . . 2248

Multilevel Determinants of Engagement in Lifelong Learning in Orthodontics From Formal to AI-Supported Approaches: Development of a Conceptual Framework Through a Qualitative Study (e85634)
Theerasak Nakornnoi, Jitpitchaya Wimonchaijit, Rochaya Chintavalakorn, Kawin Sipiyaruk. . . . 2264

Psychomotor Performance in Blended Versus In-Person Emergency Medical Technician Training: Retrospective Cohort Study (e94950)
Cheng-Heng Liu, Ming-Ju Hsieh, Yu-Ting Chen, Chih-Wei Yang. . . . 2277

Medical Students' Reflections on Transitioning to Their First General Practice Placement: Qualitative Descriptive Study (e76599)
Sara Qasrawi, Maryam Tomerak, Shahad Mamalchi, Fatima Atieh, Hasan AlAdraj, Mohamed Abdulla, Salim Fredericks, Ghufran Jassim, Hani Malik, Eric Clarke, Denis Harkin, Shaista Guraya. . . . 2313

AI Perceptions, Professional Identity, and AI-Supported Clinical Decisions Among Medical Students and Clinicians: Cross-Sectional Survey and Quasi-Randomized Vignette Study (e104151)
Julia-Astrid Moldt, Teresa Festl-Wietek, Kay Nieselt, Susanne Zabel, Manfred Claassen, Samuel Wagner, Ulrike Keim, Anne Herrmann-Werner.
2 3 2 5

AI Literacy and Training Needs Among French Midwifery Students: National Cross-Sectional Online Survey and Exploratory Qualitative Follow-Up (e94310)
Sylvain Kotzki, Sonia Knepfler, Lionel Di Marco, Nicolas Vuillerme. . . . 2341

Understanding Clinicians' Informational Needs for AI-Driven Clinical Decision Support Systems: Qualitative Interview Study (e85228)
Simone Mingels, Hannah Piehl, Madeline Therrien, Ekaterina Akhmad, Anniek van Hienen, Johan van Soest, Laura Hochstenbach, Andre Dekker, Olga Damman, Rianne Fijten. . . . 2356

Value, Structure, and Curriculum in US Graduate Health Informatics Programs: Cross-Sectional Study (e87479)
Suhila Sawesi, Diane Dolezel, Pranitha Presingu, Michael Irungu. . . . 2369

Data Science Education for Residents, Researchers, and Students in Psychiatry and Psychology: Program Development and Evaluation Study (e75125)
Hayoung Donnelly, David Mandell, Sy Hwang, Emily Schriver, Ugurcan Vurgun, Graydon Neill, Esha Patel, Megan Reilly, Michael Steinberg, Amber Calloway, Robert Gallop, Maria Oquendo, Gregory Brown, Danielle Mowery. . . . 2387

Effectiveness of Informed AI Use on Clinical Competence of General Practitioners and Internists: Pre-Post Intervention Study (e75534)
Eyad Qunaibi, Ayman Al-Qaaneh, Baraa Ismail, Hussam Muhidat, Farhia Rageh, Najwa Musallam, Alaa Fawzy. . . . 2395

XSL•FO
RenderX

Construct Validity of Frontier Large Language Models on 4 Spanish Health Professions Specialty-Access Examinations (MIR, EIR, PIR, and FIR): Cross-Sectional Benchmarking and Psychometric Study (e106199)
Clemente García-Hidalgo, Manuel Carpio-Salmerón, Pau Benito, Carlos Carazo-Casas, Beatriz Carratalá, Jaime Campos Pavón. 2409

Medical AI Use Intention Among Medical Students and Faculty and Its Associations With AI Literacy, Perceived Benefits, and Risks: Cross-Sectional Survey Study (e105032)
Bomyee Lee, Gwanwook Bang, Su Chae. 2431

Test-Retest Reliability of Large Language Models and Their Concordance With Human Experts in Rubric-Based Scoring of Clinical Summary Statements: Experimental Comparative Study (e94666)
Patrick Stadler, Piotr Bucichowski, Ylva Holzhausen, Inga Hege, Harm Peters, Julius Kaminski. 2464

Multimodal Large Language Models for Dental Chart Image Interpretation: Cross-Sectional Benchmarking Study With Students and Clinicians (e91809)
Ah-Young Cho, Soo-Heang Eo, Mi-Jeong Jeon, Jae-Hoon Kim, Jungjoon Ihm, Deog-Gyu Seo. 2482

Role-Differentiated AI Competencies and Curriculum Implications for Health Professions Education: Qualitative Study (e97608)
Yuyi Park, Yejin Han, Hyeongjo Kim, Solmoe Ahn, Frederick Kron, Jihyun Lee. 2505

AI-Assisted Angoff Standard Setting for Multiple-Choice Examinations in Medical Education: Comparative Study (e98766)
Salah Kassab, Mariam Shadan, Arina Ziganshina, Shifan Khanday, Hossam Hamdy. 2525

Performance of Cloud-Hosted Large Vision-Language Models on the Japanese National Examination for Clinical Laboratory Technicians: Comparative Benchmarking Study (e84266)
Kota Murakami, Ryosuke Matsuzawa, Yuya Tahara-Arai, Haruka Ozaki. 2536

Teaching Model Context Protocol, Retrieval-Augmented Generation, and AI Agents to a Multidisciplinary Hospital Workforce: Single-Group Pre-Post Survey Study (e97822)
Gakyoung Baek, Hyunna Lee, Dong Yang, Minseo Kang, Kun Lee, Minji Choi, Yura Lee, Kye Lee. 2554

Perceptions, Attitudes, and Use of AI by Medical Students: Mixed Methods Study (e91345)
Frédéric Paris, Vincent Garrouste, Laure Abensur Vuillaume. 2573

Enhancing Psychiatry Training Using an Agentic AI Simulated Consultation Tool: Prospective Cohort Study (e88580)
Alice Rueda, Huda Al-Shamali, Zack Cote, Niloy Roy, Reinhard Janssen-Aguilar, Jithin Joseph, Bazen Teferra, Mohammad Kamaleddin, Lisa Burback, Olga Winkler, Bill Kapralos, Andrei Torres, Divya Sharma, Sridhar Krishnan, Guillaume Dumas, Andrew Greenshaw, Alexandre Hudon, Sanjeev Sockalingam, Yanbo Zhang, Adam Dubrowski, Venkat Bhat. 2593

Performance of GPT-4o and Claude in Medical Ethics Scenarios: Comparative Study (e70199)
Karishma Desai, Anna Gorsky, Nicole Zelenski. 2611

Benefits of Chain-of-Thought Prompting for Clinical Record Rubric Evaluation in Undergraduate Medical Education: Experimental Evaluation Study With Medical Faculty (e88652)
Alberto Nogales, Sophia Denizon, Alonso Mateos Rodriguez, Javier Cervera Cordero, Gonzalo Pandelet Barainca, Enrique Aranguren Moliner, Emilio Cervera Barba. 2629

Use, Concerns, and Perspectives on AI in Health Care Among French Health Professionals and Students: Web-Based Cross-Sectional Survey (e89152)
Aurelia Alati, Grégoire Pigné, Carole-Anne Brugère, Jean-Emmanuel Bibault. 2646

Prompt Framing and Evidence Requirement for AI-Generated Educational Responses in Dental Education: Experimental Study (e90736)
Man Hung, Corban Ward, Owen Cohen, Jacob Marx, Zachary Smit, Jacob Newman, Martin Lipsky. 2656

XSL•FO
RenderX

Human-Edited Generative AI-Assisted Multiple-Choice Questions in Postgraduate Family Medicine: Blinded Cross-Sectional Comparative Psychometric Study (e100179)
Kai Sze, Jia Lim, Yng Loke, Teck Ding, Wei Wee, Tongyuan Chen, Jingkai Wu, Vivek Bansal, Qi Fong, Minyang Chow. 2668

Integrating Generative AI Into Clinical Reasoning Education for Medical Students: Mixed Methods Study (e94951)
Cheng-Heng Liu, Yu-Ting Chen, Chiun Hsu, Huey-Ling Chen, Chih-Wei Yang. 2686

AI Integration in Undergraduate Medical Education: Qualitative Study of Faculty Perspectives in the United Arab Emirates (e90192)
Azhar Rahma, Uffaira Hafeez, Munawar Farooq. 2697

Multiagent Large Language Model Framework for Psychotherapy Fidelity Assessment in Motivational Interviewing and Cognitive Behavioral Therapy Training: Cross-Sectional, Simulation-Based Evaluation Study (e92964)
Mohammad Kamaleddin, Mina Mirjalili, Reza Barzegar, Nghia Trung Le, Zachary Cote, Olga Winkler, Lisa Burback, Yanbo Zhang, Osmar Zaiane, Candice Monson, Divya Sharma, Sri Krishnan, Andrew Greenshaw, Richard Zeifman, Peter Selby, Venkat Bhat. 2710

Scaffolded AI-Supported Problem-Based Learning for Medical Interns: Exploratory Retrospectively Registered Randomized Controlled Evaluation With a Voluntary Feasibility Follow-Up (e95039)
Zhengqi Zhao, Baijing Wu, Shuyuan Tian, Haigen Hu, Pengyi Hao. 2767

AI-Generated Versus Human Supervisor Feedback on Medical Students' Clinical Clerkship Logs: Cross-Sectional Convergent Mixed Methods Study (e90064)
Takeshi Kondo, Jeroen Donkers, Hiroshi Nishigori, Sanne Rovers, Sylvia Heeneman. 2784

Impact of a Practical, Hands-On, Continuing Professional Development Course About AI in Health Care Professions Education on the Perceptions and Behaviors of Health Care Educators: Qualitative Case Study (e87381)
Justin Peacock, Jerusalem Merkebu, Anita Samuel. 2808

AI Integration in Spanish Undergraduate Medical Education: National Cross-Sectional Study (e88511)
Ana Enériz Janeiro, Karina Pitombeira Pereira, Julio Mayol, Javier Crespo, Fernando Carballo, Juan B Cabello, Manuel Ramos-Casals, Bibiana Pérez Corbacho, Juan Turnes. 2822

AI-Generated Multiple Mini Interview (MMI) Stations for Medical School Admissions: Psychometric Evaluation (e86208)
Sabryn Hamila, Kyle Birchill, Khoa Cao, Md Hossain, Shane Bullock, Wayne Hodgson, Julia Harrison, Michelle Leech. 2834

Health Professional Students' Use of Generative Artificial Intelligence During Clinical Placements: Cross-Sectional Online Survey Study (e85243)
Sylvain Kotzki, Calvin Massonnet Turner, Kim Gauthier, Mélanie Minoves, Nicolas Vuillerme. 2847

Generative AI's Impact on the Mental Health of Medical Students: Scenario Analysis (e85373)
Nora Arvai, Bertalan Meskó, Gellért Katonai. 2863

An AI-Driven Virtual Patient Platform (CBT Trainer) for Training Cognitive Behavioral Therapy Practitioners Against Competencies: Mixed Methods Pilot Study (e84091)
Tianyu Zhang, Rob Saunders, Stephen Pilling, Ciarán O'Driscoll. 2882

Performance Evaluation of Large Language Models in Multilingual Medical Multiple-Choice Questions: Mixed Methods Study (e81399)
Livia Strasser, Wilma Anschuetz, Fabio Dennstädt, Janna Hastings. 2900

XSL•FO
RenderX

Integrating Large Language Models Into Trauma Education for Medical Students: Randomized Controlled Pilot Trial (e79134)
Joona Gustafsson, Erno Lehtonen-Smeds, Niklas Pakkasjärvi. . . . 2917

ChatGPT versus UpToDate in Preclinical Medical Education: Cross-Sectional Analysis Using Term Frequency–Inverse Document Frequency Cosine Similarity (e82885)
Shankar Thiru, Nicholas Aksu, Matthew Chiang, Daniel Gallagher, Mary Furlong, Elizabeth Prevou, Akhil Khanna. . . . 2928

Comparing AI-Assisted Problem-Solving Ability With Internet Search Engine and e-Books in Medical Students With Variable Prior Subject Knowledge: Cross-Sectional Study (e81264)
Ajiith Xavier, Syed Naeem, Waseem Rizwi, Hiramani Rabha. . . . 2941

AI Literacy Among Chinese Medical Students: Cross-Sectional Examination of Individual and Environmental Factors (e80604)
Chunqing Li, Sian Tsuei, Hongbin Wu. . . . 2949

GPT-4o and OpenAI o1 Performance on the 2024 Spanish Competitive Medical Specialty Access Examination: Cross-Sectional Quantitative Evaluation Study (e75452)
Pau Benito, Mikel Isla-Jover, Pablo González-Castro, Pedro Fernández Esparcia, Manuel Carpio, Iván Blay-Simón, Pablo Gutiérrez-Bedia, Maria Lapastora, Beatriz Carratalá, Carlos Carazo-Casas. . . . 2963

Adaptation of the Japanese Version of the 12-Item Attitudes Towards Artificial Intelligence Scale for Medical Trainees: Multicenter Development and Validation Study (e81986)
Hirohisa Fujikawa, Hirotake Mori, Kayo Kondo, Yuji Nishizaki, Yuichiro Yano, Toshio Naito. . . . 2975

Evaluation of a Problem-Based Learning Program's Effect on Artificial Intelligence Ethics Among Japanese Medical Students: Mixed Methods Study (e84535)
Yuma Ota, Yoshikazu Asada, Saori Kubo, Takeshi Kanno, Machiko Yagi, Yasushi Matsuyama. . . . 2985

AI-Driven Objective Structured Clinical Examination Generation in Digital Health Education: Comparative Analysis of Three GPT-4o Configurations (e82116)
Zineb Zouakia, Emmanuel Logak, Alan Szymczak, Jean-Philippe Jais, Anita Burgun, Rosy Tsopra. . . . 3000

Using AI to Train Future Clinicians in Depression Assessment: Feasibility Study (e87102)
Friederike Holderried, Alessandra Sonanini, Annika Philipps, Christian Stegemann-Philipps, Lea Herschbach, Teresa Festl-Wietek, Stephan Zipfel, Rebecca Erschens, Anne Herrmann-Werner. . . . 3017

Susceptibility of Assessment Types to AI-Generated Content in Digital Health and Health Information Management Education: Quasi-Experimental Pilot Study (e82988)
Tafheem Wani, Michael Liem, Natasha Prasad, Kerin Robinson, Abbey Nexhip, Melanie Tassos, Stephanie Gjorgioski, Urooj Khan, James Boyd, Merilyn Riley. . . . 3031

Faculty Perspectives on AI Integration in Anatomy Education in the United Arab Emirates: Cross-Sectional Survey (e87418)
Prince Zilundu, Pedzisai Mazengenya, Jayaraj Narayanan, LiHua Zhou. . . . 3051

Performance of Large Language Models on the Brazilian National Medical Education Examination: Comparative Benchmark Study (e89839)
Francys Fernandes da Silva, Eduardo Roeder, João Bruneti Severino, Matheus Nespolo Berger, Pedro Basei de Paula, Davi Ferreira, Maria Han Veiga, Thyago de Moraes, Gustavo Lenci Marques. . . . 3068

Impact of an Ambient AI Scribe on Medical Student Objective Structured Clinical Examination Notes: Nonrandomized Clinical Trial (e88264)
Jaideep Talwalkar, Donald Wright, Lee Schwamm, Gary Leydon, Veronika Shabanova. . . . 3087

XSL•FO
RenderX

Training Empathetic Communication Skills in Medical Students With a Role-Prompted GPT-4o Chatbot: Quasi-Experimental Intervention Study (e88092)
Yannik Müller, Hans Lemberg, Paul Gellert, Jan Zoellick. . . . 3100

Voice Cloning Using AI vs Traditional Audio Recording for Prerecorded Courses in Medical Pedagogy: Randomized Controlled Trial (e86569)
Antoine Gavoille, Fabien Subtil, Mikaïl Nourredine. . . . 3112

AI-Generated Microlearning for Plastic Surgery Residency: Single-Arm Pre-Post Feasibility Study (e88695)
Marius Drysch, Sonja Schmidt, Felix Reinkemeier, Flemming Puscz, Alexander Fiedler, Maria Fueth, Marcus Lehnhardt, Alexander Sogorski, Christoph Wallner. . . . 3122

AI-Simulated Patients for Training Shared Decision-Making: Feasibility Study in Medical Education (e100467)
Lea Herschbach, Teresa Festl-Wietek, Alessandra Sonanini, Friederike Holderried, Tobias Albrecht, Bernd Herrmann, Anne Herrmann-Werner.
3 1 3 8

Estimating the Prevalence of Generative AI Use in Medical School Application Essays: Cross-Sectional Study (e96673)
Nicholas Spies, Valerie Ratts, Ian Hagemann. . . . 3152

Large Language Model Performance on Multistep Clinical Cases: Comparative Study Across Question and Case Levels (e95342)
Jiaxue Cha, Yan Zhao, Hui Zong. . . . 3164

Generative AI Factors, Cognitive Engagement, and Patient Safety Learning in Medical Education: Mixed Methods Study (e100828)
Gilbert Golahi, Ferdi Antonio, Devina Annisa. . . . 3174

Use of Large Language Models for Automated Short-Answer Question Grading in Second-Cycle Medical Education: Performance and Consistency Evaluation Study (e89445)
Sabrina Sidali, Pierre-Antoine Soret, Mathilde Wagner, Thibault Voron, Quentin Vanderbecq. . . . 3200

Curriculum-Aligned Conversational AI to Support Pathology Self-Directed Learning in Preclinical Medical Students: Crossover Randomized Pilot Study (e102908)
Kasie Roark, Yumin Choi, Elizabeth Hoy, Noel Tufts, Robert Amevor, Rahul Ghosal, Brian Keisler, Leonardo Bonilha, Ioulia Chatzistamou. . .
3 2 1 2

Medical Interview Performance Using an AI-Based Standardized Patient vs a Human Standardized Patient: Randomized Crossover Trial (e111030)
Yasutaka Yanagita, Daiki Yokokawa, Fumitoshi Fukuzawa, Shiichi Ihara, Ryo Yoshida, Rurika Sato, Takanori Uehara. . . . 3225

Interprofessional Training in Virtual Reality for Health Care: Experimental Study on Procedural Knowledge and Willingness to Collaborate (e85139)
Melanie Bauer, Matthias Witti, Matthias Stadler, Anna Weiß, Martin Fischer, Constanze Richters. . . . 3264

Development and Validation of a Scale for Measuring Leadership and Managerial Competencies of Middle Managers in Health Care and Medical Education in the Gulf Region: Cross-Sectional Study (e77476)
Ahmed Al Ansari, Elizabeth Holloway, Reginald Sequeira. . . . 3278

Profiling Human-AI Regulatory Support in Technology-Enhanced Interprofessional Education Among Health Professions Students: Person-Centered Exploratory Study (e95772)
Fraide Ganotice Jr, John Dizon, Xiaoai Shen. . . . 3289

XSL•FO
RenderX

A Dual Hospital–University Specialist Training Reform to Address Workforce Shortage and Geographic Maldistribution in Indonesia: Integrative Qualitative Policy Process Review (e90270)
Asnawi Abdullah, Yuli Farianti, Lupi Trilaksono, Arianti Anaya, Anna Kurniati, Oos Rosyati, Etik Wiyati, Mimi Sumiarsih, Christa Manik, Nyiayu Sonia, Farizal Muharram, Kevin Simanjuntak, Sydney Tjandra, Akemat Akemat, Else Sihotang, Leni Kuswandari, Siti Yunianti, Fialisa Asriwhardani. . . . 3303

Gender Differences in Medical Students' Self-Assessment: Longitudinal Multi-Cohort Study (e91071)
Lennart Handke, Hendrik Friederichs. . . . 3336

Physical Attractiveness and Chances of Being Invited to Interview With a Medical Residency Program: Retrospective Cohort Study (e81052)
Daria Hunter, Luke Hunter, David Kerner, Aidan Mullan, Derek Vanmeter, Ivan Khapov, Alexander Finch, Sara Hevesi, Ivan Porter II, Colin West, James Homme. . . . 3371

Health and Humanities and Social Sciences Professionals' Perceptions Regarding the Teaching of the Effects of Racism in Medicine: Semistructured Interview Study (e90084)
Mathilde Lambert, Rahmeth Radjack, Adil Mouhab, Malika Mansouri, Marie Moro. . . . 3408

Reviews

Communication Education in Prelicensure and Postlicensure Occupational Therapy: Scoping Review (e95909)
Ryuichi Shimada, Yukiko Sano, Takeshi Sasaki, Astuko Tanimura. . . . 179

Generation of Case-Based Materials Using Large Language Models in Health Professions Education: Scoping Review (e94795)
Zineb Zouakia, Alan Szymczak, Salomé Tordjman, Emmanuel Logak, Thomas Meyre, Matthieu Resche-Rigon, Alain Cariou, Anita Burgun, Rosy Tsopra. . . . 195

Interventions to Enhance COVID-19 Pandemic Health Literacy in Health Professionals: Systematic Review (e70400)
Claudia Hasenpusch, Paula Kuper, Eva-Maria Skiba, Antonia Sprenger, Uwe Matterne, Lena Kannengießer, Esther Grätsch, Jaekyung Shin, Mengtong Li, Christian Apfelbacher. . . . 218

Video Recording of Patient-Clinician Interactions in Health Education: Scoping Review (e70324)
Mariana Morgado, Yago Leira, Tiago Leitão, José Mendes, Luís Proença. . . . 250

AI for Clinical Competency Assessment: Scoping Review of Methods and Applications (e92826)
Wenjia Zhang, Benjamin Daniels, Carol Mita, Hoang Nguyen, David Duong. . . . 272

Application of AI-Generated Content in Medical Education: Systematic Review of the Impact on Critical Thinking Abilities of Medical Students (e79939)
Jinlei Li, Fen Ai, Jueyan Wang, Bingxin Cheng, Yu Li, Zhen Chen. . . . 297

Educational Formats and Content Domains of Interprofessional Education for Licensed Rehabilitation Professionals: Scoping Review (e76189)
Kohei Ikeda, Takao Kaneko, Someka Hijikuro, Natsuki Inoue, Takuto Nakamura, Taisei Takeda, Junya Uchida, Hirofumi Nagayama. . . . 313

Retuning the Premedical Compass in American Programs Worldwide: Scoping Review (e85002)
Michelle Chouairy, Laetitia Bou Samra, Christelle Zeidan, Ola El Zein, Bilal Kaafarani. . . . 329

Use of Digital Technology for Developing Communication Skills in Undergraduate and Postgraduate Medical Education: Scoping Review (e87012)
Princella Seripenah, Heidi Emery, Bakula Patel, Edward Tyrrell, Julie Carson, Jo Leonardi-Bee, Catrin Evans, Emma Wilson, Jaspal Taggar. .
3 4 3

XSL•FO
RenderX

The Effectiveness of Artificial Intelligence in Undergraduate Health Professions Education: Systematic Review and Meta-Analysis of Randomized Controlled Trials (e88933)
Nai Lai, Yin Lim, Min Win, Prabal Bhargava, Paraidathathu Thomas, Qi Ong. . . . 365

Outcomes of Digital Training for Community Health Workers in Low- and Middle-Income Countries: Scoping Review (e82772)
Tapiwa Tembo, Nora Rosenberg, Firaol Ayele, Saeed Ahmed, Linda-Gail Bekker. . . . 389

Augmented Reality in Surgical Training: Systematic Review of Its Impact on Technical Performance in Surgical Trainees (e71572)
Mahmoud El Ashry, Ahmed El Ashry, Hamza Khalique, Yahye Abdalle, Thomas Yeung. . . . 415

Main Recommendations for Developing Education and Awareness Strategies for Rare Diseases: Scoping Review (e79027)
Diego Yamada, Tatiana Komoto, Vinícius Lima, Julio Souza, Michele Seixas, Victor Cassão, Victor Ferraz, Bibiana de Oliveira, Têmis Félix, Domingos Alves. . . . 429

Toolkits to Support Health Care Providers in Health Behavior Change for Patients With Chronic Conditions: Scoping Review (e88653)
Jillian Scandiffio, Robert Simpson, Ashvene Sureshkumar, Jerry Tang, Paloma Fleming, Dina Gaid, Harzaan Gnanakaran, Marina Wasilewski, Nicole Woods, Sarah Donkers, Sarah Munce. . . . 447

Generative AI for Personalized Learning Platforms in Otolaryngology–Head and Neck Surgery Education: Scoping Review (e91785)
Jiang-Tao Zhong, Zai-Zai Cao, Shui-Hong Zhou. . . . 2442

Medical Students' Attitudes, Perceptions, and Self-Reported Familiarity With AI in Health Care: Systematic Review and Meta-Analysis (e89411)
Wei Zhang, Jiaxue Han, Xin Han, Hang Xu, Langkun Wang, Tianhai Lin, Ping Tan, Peng Zhang, Xiaonan Zheng. . . . 2730

Parenthood in US Medical Training Across Specialty Groups: Scoping Review (e87284)
Zuleica Santiago Delgado, Elizabeth Vaughan, Mary Jones, Forrest Jones, Dylan Bridwell, Laura Porterfield. . . . 3347

Gender and Racial Discrimination During Residency Training: Scoping Review (e87524)
Ida John, Liz Dennett, Julie Nguyen, Marghalara Rashid. . . . 3384

Viewpoints

AI-Mediated Assessment of Continuing Medical Education: The Case-based Learning Intelligence Credit System (CLICS) Framework (e99520)
Somkiat Wattanasirichaigoon. . . . 462

The New Gravity: How Generative AI Remapped Innovation and What Academia Must Do Next (e89741)
Hossein Akbarialiabad, Ayman Grada, Sancy Leachman, David Rhew, Alireza Haghighi. . . . 472

Integrating Large Language Models and Metaverse-Related Technologies in Urology Education: Viewpoint (e93235)
Peizhe Li, Huijie Gong, Heyang Liu, Siyu Han, Yiqin Lu, Yongji Yan. . . . 484

AI-Mediated Relational Competence in Medical Education: Construct Rationale and Curriculum Maturity Framework (e105459)
Robert McGrath. . . . 492

XSL•FO
RenderX

Artificial Intelligence as a Core Public Health Competency: Proposal and Perspectives in Education and Research (e92246)
Laura Brunelli, Federico Fonda, Silvio Brusaferro. . . . 504

From Anatomy to Innovation: Embedding Patient-Centered Design in Medical Education (e86230)
Louis Koizia, Pieter Vandekerckhove, Steven Howard, Bettina Maisch, Benjamin Harris. . . . 513

Bridging the Implementation Gap in Medical AI Education: 3-Lever Framework for Concurrent Reform (e88778)
Sherrie Kelly, Melissa Forgie, Mark Walker. . . . 519

From Classroom to Operating Room: Addressing Shortcomings in Gender Affirmation Surgical Education (e71843)
Samuel Robinson, Alireza Hamidian Jahromi. . . . 528

AI-Assisted Formative Assessment in Clinical Education: From Algorithms to Agency (e93710)
Quang Nguyen, Thuy Ha, Tina Mai. . . . 532

AI in UK Medical Education: A Framework for Curriculum Reform (e81953)
Aditya Gaur, Joecelyn Kirani Tan, Medha Sridhar Rao, Muhtasim Fuad, Taha Bhatti, Hareesha Rishab Bharadwaj, Khabab Mohamed Ahmed.
5 4 1

The AI Colleague: Reimagining Interprofessional Collaboration and Education in Medicine (e94034)
Chisung Yuh, Chul-Hyun Cho, Young-Mee Lee. . . . 553

Beyond Pattern Recognition: Call for Functionally Aware AI for Anatomical Illustration (e92700)
Agata Kawalec-Rutkowska, Marian Simka. . . . 563

Promise, Mimicry, and Surveillance: Responsibly Integrating Artificial Intelligence With Socratic Inquiry in Medical Education (e102958)
José Sánchez. . . . 569

Artificial Intelligence in Medical Education: Transformative Potential, Current Applications, and Future Implications (e77127)
Juan Izquierdo-Condoy, Marlon Arias-Intriago, Laura Montero Corrales, Esteban Ortiz-Prado. . . . 580

From Realism to Learner Engagement: Rethinking Fidelity in Simulation-Based Education (e84684)
Julien Pico, Jean-Noel Evain, Christina Aron, Gilles Martin, Ilian Cruz-Panesso, Leonida-Mihai Georgescu, Issam Tanoubi. . . . 594

Voices of Innovation: Reflective Report on Integrating Artificial Intelligence–Simulated Mental Health Patient Scenarios Into Undergraduate Nursing Education in the United Arab Emirates (e78161)
Amina Ahmad, Janisha Kavumpurath, Raheesa Kader, Muna Altamimi, Monia El Hajj, Fatma Ahmed, Muhammad Subu, Taliaa Yafei, Hind Ali, Idil Saleh, Nabeel Al-Yateem. . . . 600

Dopamine, Distraction, and Disruption: Perspectives on How Technology and Generation Z Are Reshaping Medical Education (e80791)
Bruno Andrade, Mariana Araújo Pereira, Katia Avena. . . . 609

Digital Choice Architecture in Medical Education: Applying Behavioral Economics to Online Learning Environments (e86497)
Victoria Ekstrom. . . . 619

Implementing Large Language Models to Support Misconception-Based Collaborative Learning in Health Care Education (e81875)
Brandon Cheah, Shefaly Shorey, Jun Ch'ng, Chee Tan. . . . 629

XSL•FO
RenderX

Preparing Physicians for Tomorrow's Practice: Contemporary Practice Competencies as a Framework for Technology-Driven Change (e90637)
Joseph Safdieh, June Chan, Domenick Falcone, Erika Abramson. . . . 642

Is This Still Medical Education? The Educational Use of Nude Pediatric Clinical Photographs: Bioethical Reassessment (e104023)
Iwona Bujek. . . . 647

The AWARE (AI Use, Why, Attachment, Reality and Risk, and Effect on Functioning) Framework for Assessing Patients' Use of AI in Mental Health Care (e108766)
Alexandre Hudon, Venkat Bhat, John-Jose Nunez, Vincent Paquin, Amylie Malouin-Lachance, François Lespérance. . . . 661

Generative Language Models in Medical Education: From Advent to Entrustment (e105347)
Shiv Patil, Mert Karabacak, Sotirios Bisdas, Konstantinos Margetis. . . . 672

From Competence to Expertise: The Role of Meaningful Experience in Physician Training (e106585)
Nicole Dubosh, Alexander Iyer, Richard Schwartzstein. . . . 681

Defining and Exploring Experience in Health Education (e85039)
Louis-Baptiste Jaunay, Emmanuel Triby, Thierry Pelaccia. . . . 690

Digital Standardized Patients: Conceptual Framework for Generative AI–Empowered Medical Education (e91050)
Chuanlin Jia, Lulu Qi. . . . 2756

Tutorials

Addressing the Engagement Blind Spot: 12-Tip Tutorial to Make Reasoning Audible in Health Professions Education (e104702)
Inthrani Indran, Dujeepa Samarasekera. . . . 700

Co-Designing and Evaluating a 1-Day Quality Improvement Workshop for Medical Students and Resident Physicians: Tutorial on Applying Kern's Curriculum Development Framework (e83657)
Amanda Ling Jie Yee, Mariam Idrissi, Sam Sherratt-Mayhew, Charles Page, Maiar Elhariry, Punith Kempegowda. . . . 713

Empowering Educators to Build AI Chatbots in Health Professions Education: Tutorial on a No-Code Design Workflow (e103399)
Charlene Goh, Eugene Wang, Juanita Kong, Judy Sng. . . . 724

Enhancing Pathology Knowledge for Biomedical Scientists: Pilot Cohort Study of a Comprehensive Self-Paced Tutorial for Bridging Fundamental Concepts With Translational Applications (e84903)
Jennifer Sells, Leticia Clemente, David Farris, Mary Sizemore, Ignacio Wistuba, Rama Soundararajan, Maria Raso. . . . 736

Twelve Tips for Recognizing and Addressing the Adverse Effects of Medical Education Interventions: Tutorial (e89647)
Tadayuki Hashimoto, Mohammad Hasdianda, Kanapa Kornsawad, Shunsuke Kosugi, Makoto Kikukawa. . . . 749

Trust Analysis Canvas for Teaching in the Field of Digital Public Health and Medicine: Tutorial (e79709)
Federica Zavattaro, Clara-Maria Barth, Caroline Brall, Viktor von Wyl, Felix Gille. . . . 757

XSL•FO
RenderX

Developing and Integrating Virtual Reality Courses in Medical Education: Tutorial and Implementation Guideline Informed by Best Practices From the National Project “medical tr.AI.ning” (e80976)
Marvin Mergen, Anna Junga, Henriette Schulze, Philipp Bozdere, Pascal Kockwelp, Leon Pielage, Corbin Sassen, Tina Glückselig, Michael Schmitz, Kathrin Ungru, Norbert Graf, Benjamin Risse, Marcel Meyerheim. . . . 769

Developing a Large Language Model–Based Feedback System for Case Report Writing in Rehabilitation Education: Tutorial (e76924)
Yuuto Tonouchi, Shunsuke Nakai, Kayo Murakami, Yuki Kataoka. . . . 788

How to Conduct Usability Testing of Medical Education Simulations: Methods Tutorial Using a Telemedicine Simulation for Maternal Care (e102697)
Blake Lesselroth, Alexander Douglas, Helen Monkman, Romaric Marcilly, Karalane Bellavia, Mataeo Anderson, Jacob Van Buren, Charles Parsons, Sonakshi Mohanty, Juell Homco, Morgan Richards, Karen Gold. . . . 801

Inclusive Medical Spanish Education to Address Bias in Clinical Communication: Tutorial on Patient-Informed Module Development (e98398)
Alexandra Lopez Vera. . . . 3316

Research Letters

“AACHEN” e-Learning Tool in Augmentative and Alternative Communication for Medical Students in Germany: Cross-Sectional Evaluation Study (e88173)
Jessica Büchs, Christiane Neuschaefer-Rube. . . . 1611

Evaluating Microlearning for Faculty Development in Medical Education: Mixed Methods Pilot Study (e87980)
Darci Lammers, Jeffrey Geske, Jane Linderbaum, Michael Cullen. . . . 3259

A National Exploration of Medical Students’ Views Surrounding Skin Color Terminology in Undergraduate Education (e92420)
Niamh McSwiney, Eliza Hutchison, Stephanie Lax, Paul Leighton, Hannah Wainman. . . . 3330

Letters to the Editor

Rethinking Medical Spanish Competence: Language Proficiency, Clinical Communication, and Inclusion (e110302)
Juli McCarroll. . . . 2291

Author’s Reply: Rethinking Medical Spanish Competence: Language Proficiency, Clinical Communication, and Inclusion (e110709)
Alexandra Lopez Vera. . . . 2293

Authors’ Reply: Methodological Concerns in AI Medical Education Frameworks (e104447)
Juan Izquierdo-Condoy, Marlon Arias-Intriago, Laura Montero Corrales, Esteban Ortiz-Prado. . . . 2295

Authors’ Reply: Big Data, Small Stories: Methodological Considerations for Using Social Media Analytics in Medical Education Research (e97664)
Faisal Binsar, Mohammad Hamsal. . . . 2297

Big Data, Small Stories: Methodological Considerations for Using Social Media Analytics in Medical Education Research (e94825)
Xiaoshuai Li, Rui Wang. . . . 2299

XSL•FO
RenderX

Methodological Concerns in AI Medical Education Frameworks (e101696)
Syeda Fatima, Emaan Fatima, Ragni Kumari. . . . 2301

Documentation Quality and Educational Value in AI-Assisted Feedback (e105029)
Amane Endo, Takeshi Kimura, Yuki Kataoka. . . . 2303

Enhancing Team-Based Learning in Virtual Environments: The Role of Avatar Agency and Immersive Social Presence (e92366)
Hiroki Funao, Motomu Shimaoka, Jun Kako. . . . 2305

Authors' Reply: Enhancing Team-Based Learning in Virtual Environments: The Role of Avatar Agency and Immersive Social Presence (e93060)
Darunee Sripadungkul, Suhattaya Boonmak, Monsicha Somjit, Narin Plailaharn, Wimonrat Sriraj, Polpun Boonmak. . . . 2307

Why Medical Education Without Artificial Intelligence Still Matters: A Neuroscience-Informed Perspective (e94594)
Charles Verdonk. . . . 2309

Authors' Reply: Why Medical Education Without Artificial Intelligence Still Matters: A Neuroscience-Informed Perspective (e95205)
Juan Izquierdo-Condoy, Marlon Arias-Intriago, Laura Montero Corrales, Esteban Ortiz-Prado. . . . 2311

Corrigenda and Addendas

Correction: Performance of GPT-4o and Claude in Medical Ethics Scenarios: Comparative Study (e113933)
Karishma Desai, Anna Gorsky, Nicole Zelenski. . . . 3236

Correction: Upskilling in Healthy Longevity Medicine and Its Association With Physicians' Implementation Intent and Self-Reported Clinical Confidence: Cross-Sectional Observational Study (e114811)
Evelyne Bischof, Dominika Wilczok, James Kirkland, Bhirau Wilaksono, Christine Huang, Suwanna Suwannaphong, Wanviput Sanphasitvong, Dalila amdži , Carolina Hernandez, Yoko Madea, Hidekazu Yamada, Melissa Fernandes, Ricardo Pacheco, Fabiano Serfaty, Fernanda Calvo-Fortes, Amit Goldman, Andrea Maier, Alexey Moskalev, Morten Scheibye-Knudsen, Alex Zhavoronkov. . . . 3238

Correction: Performance of ChatGPT on Nursing Licensure Examinations in the United States and China: Cross-Sectional Study (e105550)
Zelin Wu, Wenyi Gan, Zhaowen Xue, Zhengxin Ni, Xiaofei Zheng, Yiyi Zhang. . . . 3241

Correction: Augmented Reality in Surgical Training: Systematic Review of Its Impact on Technical Performance in Surgical Trainees (e106576)
Mahmoud El Ashry, Ahmed El Ashry, Hamza Khalique, Yahye Abdalle, Thomas Yeung. . . . 3243

Correction: Collaborative Simulation-Based Learning to Develop Competencies in Physical Therapy Students: Randomized Controlled Trial (e112897)
Iván Cuyul-Vásquez, Hilda Aravena-Páez, Felipe Araya-Quintanilla, Sergio Guinez-Molinos, Rodrigo Núñez-Cortés, Héctor Gutiérrez Espinoza.
3 2 4 4

Correction: Attitudes Toward Digital Planetary Health Before and After a Lecture Series on Environmental Awareness and Future Visions: Repeated Cross-Sectional Survey Study Regarding Twin Transformation (e111087)
Julia Nitsche, Theresa Busse, Sonja Schmalen, Ina Otte, Jan Ehlers. . . . 3246

XSL•FO
RenderX

Guest Editorial

Re-Centering Clinical Documentation in the Age of AI Scribes: Four Aims of the Patient Chart Note (e95218)
Chwen-Yuen Chen, Tiffany Leung, Rika Bajra. 3248

Editorial

AI Competency: Current State and Challenges (e86686)
Sian Tsuei. 3253

XSL•FO
RenderX

Effects of a Generative AI Virtual Patient on Serious Illness Communication Skills: Randomized Controlled Trial

Kurtis G Haut[1†*], MSc, PhD; Masum Hasan[1*], MSc, PhD; Thomas Carroll[2], MD, PhD; Ronald M Epstein[3], MD; Taylan Sen[4], PhD; Ehsan Hoque[1], PhD

[1]Department of Computer Science, University of Rochester, 3013 Wegmans Hall, Rochester, NY, United States

[2]Department of Medicine, Divisions of General Medicine and Palliative Care, University of Rochester Medical Center, Rochester, NY, United States

[3]Departments of Family Medicine and Medicine, University of Rochester Medical Center, Rochester, NY, United States

[4]Computer and Information Sciences, Niagara University, Niagara Falls, NY, United States

[†]deceased

[*]these authors contributed equally

Corresponding Author:
Ehsan Hoque, PhD
Department of Computer Science, University of Rochester, 3013 Wegmans Hall, Rochester, NY, United States

Abstract

Background: Serious illness communication (SIC) is a critical skill in medical education, especially in end-of-life care, yet clinicians often lack sufficient opportunities for deliberate practice. Standardized patients (SPs) remain the gold standard for SIC training, but they are limited by cost, availability, and consistency. To support wider access to SIC training, scalable alternatives are needed.

Objective: The aim of this study is to evaluate whether SOPHIE (Standardized Online Patient for Healthcare Interaction Education), an AI-powered virtual patient that provides real-time dialogue and automated feedback, improves SIC skills among clinicians and trainees.

Methods: We conducted a single-blind, parallel-group, randomized controlled trial from June to December 2024 via videoconferencing. Participants were randomized in a 1:1 ratio to the intervention (SOPHIE training) or control (reading module) group. Blinded SP actors and 4 independent raters evaluated outcomes preintervention and postintervention. The intervention group completed 3 interactive SOPHIE modules (Empathize, Be Explicit, and Empower) for 30 minutes, combining dialogue with an interactive virtual patient and automated feedback. Control participants reviewed reading materials covering the same framework for the same duration. The primary outcome was the change in communication skills (Empathize, Be Explicit, and Empower) assessed by SPs and blinded raters using a rubric based on an established clinical communication training framework. Secondary outcomes included self-reported confidence in communication skills.

Results: Fifty-one medical, nursing, and physician assistant students and practicing clinicians (mean age 30.6, SD 13.2 years; women: 40/51, 78.4%) were recruited from academic and clinical settings. Participants were stratified by training level and randomized; all completed the study. Compared with controls, SOPHIE participants demonstrated significantly greater improvement in Empower (β=.11, 95% CI .04-.17; *P*=.002), Be Explicit (β=.08, 95% CI .03-.13; *P*=.002), Empathize (β=.07, 95% CI .01-.13; *P*=.03), and overall communication score (β=.09, 95% CI .04-.14; *P*=.001). Here, β represents the additional pre-to-post improvement in the SOPHIE arm relative to the control arm, estimated from linear mixed effects models. SOPHIE participants also reported higher confidence in their communication skills than controls (mean score of 4.31 vs 3.83 on a 5-point Likert scale; *P*=.009).

Conclusions: Training with an AI-powered virtual patient significantly improved SIC skills and confidence compared with reading modules. Generative AI–based simulations may provide a scalable, accessible, and effective adjunct or alternative to traditional SP training in clinical education.

Trial Registration: ClinicalTrials.gov NCT07409233; http://clinicaltrials.gov/ct2/show/NCT07409233

(JMIR Med Educ 2026;12:e93034) doi:10.2196/93034

KEYWORDS

virtual standardized patients; AI; large language models; generative AI; serious illness communication; palliative care; end-of-life care; randomized controlled trial; medical education; simulation-based training

XSL•FO
RenderX

Introduction

Effective serious illness communication (SIC) in end-of-life care is a universally recognized gap in medical training [1-4]. These are high-stakes conversations in which clinicians must navigate weighty issues, and poor communication has been linked to prognostic misunderstanding [5], perceived lack of support [6], overtreatment [7-9], higher health care costs [10,11], and clinician burnout and malpractice risk [12-14]. National consensus statements on communication in oncology settings prioritize improving SIC [15,16].

Standardized patient (SP) role-play remains a popular training method in which a trained actor assumes the role of a patient, engages in a simulated conversation with a clinician, and provides actionable feedback. Live SP exercises are limited by cost, availability, skill, demographic diversity, and consistency [17-20]. As a result, many clinicians lack access to high-quality, repeated practice, leaving them underprepared. Scalable, easily accessible alternatives are needed to support widespread development of communication skills.

To fill this gap, AI-driven virtual patients offer a promising solution. Prior work has explored conversational agents for providing psychotherapy [21], communication skills training [22], AI-assisted feedback [23], and empathetic response generation [24]. Recent studies have demonstrated the feasibility, usability, and scalability of large language model (LLM)–based virtual patient systems for medical education, including history-taking and simulated clinician-patient dialogue with feedback, primarily through usability studies, systematic reviews, intrinsic evaluations, and trials relying on self-reported or system-based outcomes rather than blinded behavior assessment [25-30]. Despite these promising developments, it is necessary to determine whether AI systems can help cultivate communication skills that clinicians can easily apply in practice. To our knowledge, no published randomized controlled trial (RCT) has evaluated a generative AI–based virtual patient system with personalized, structured feedback and multirater-validated behavior outcomes.

We developed SOPHIE (Standardized Online Patient for Healthcare Interaction Education), a generative AI–powered virtual patient designed to support SIC training. SOPHIE simulates conversations with a patient with advanced cancer and provides real-time, structured feedback on 3 core communication skills—Empathize, Be Explicit, and Empower (3Es)—grounded in the MVP (medical situation, values, and plan) communication framework [31]. Drawing from previous work on clinical communication [32-34] and principles of deliberate practice and behavior learning [35-37], SOPHIE enables repeated, psychologically safe practice through interactive dialogue and automated feedback. Building upon our previous work in schema and in an LLM-based virtual patient system, this work introduces a hybrid dialogue manager that is controllable and versatile. The system was iteratively refined through qualitative clinical evaluation [38] and a controlled pilot study assessing feasibility [39].

In this study, we conducted an RCT to evaluate whether training with SOPHIE leads to greater improvement in SIC skills and self-reported confidence among health professionals compared with written educational materials. We hypothesized that participants assigned to SOPHIE would show greater pre-to-post improvement in Empathize, Be Explicit, Empower, and overall communication scores than participants assigned to the control condition.

Methods

Study Design

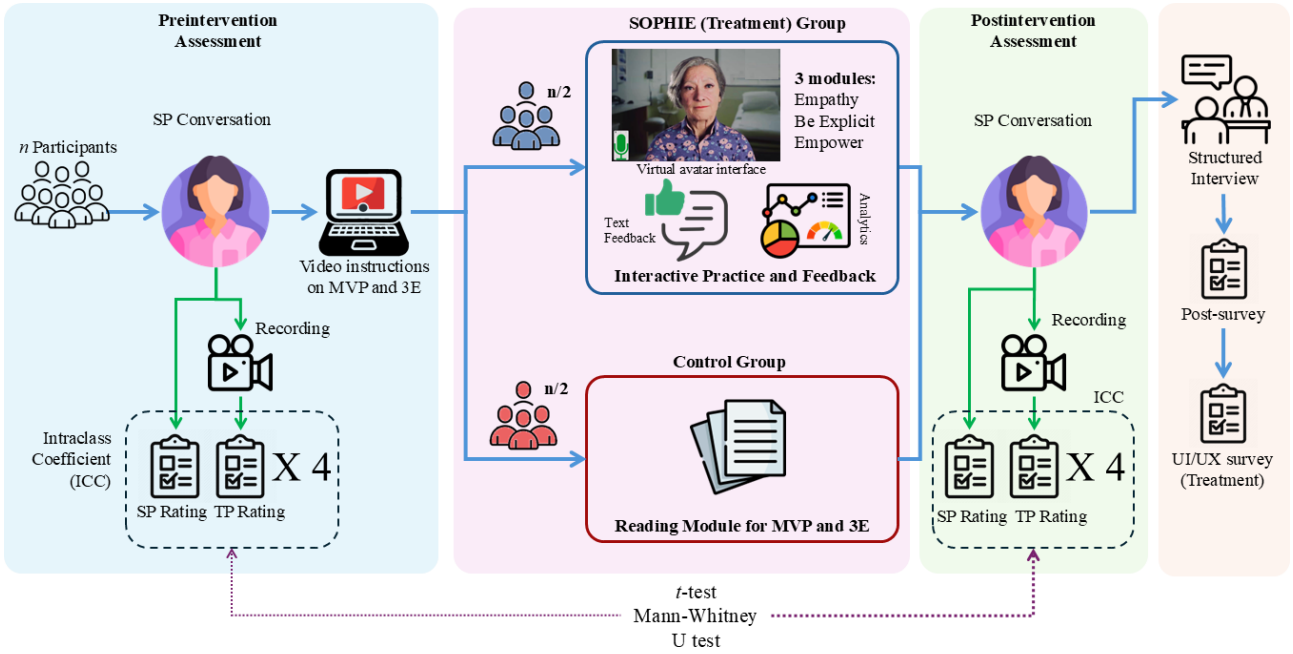
We conducted a single-blind, parallel-group, RCT to evaluate whether an AI-powered virtual patient (SOPHIE) could improve clinician communication skills in serious illness conversations. Participants were randomized in a 1:1 ratio to the intervention (SOPHIE training) or the control (reading modules) group. The study is reported in accordance with the CONSORT 2025 Statement for randomized trials, and the completed CONSORT (Consolidated Standards of Reporting Trials) checklist is provided in Checklist 1.

Sample Size Calculation

We performed a power analysis using pilot data that showed an effect size of 0.82 (Cohen *d*) to determine the target sample size for the study. A 1-tailed, 2-sample *t* test at a 5% type I error rate was selected to detect improvement in the SOPHIE arm relative to the control. We determined that a minimum of 48 participants (24 per arm) was needed to achieve at least 80% power. Sessions were recorded for blinded evaluation. Control participants were offered SOPHIE training after the study. Figure 1 demonstrates the RCT study design.

XSL•FO
RenderX

Figure 1. Randomized controlled study design. Participants were assigned to either the control group (reading module) or the SOPHIE (Standardized Online Patient for Healthcare Interaction Education) group (AI-guided simulation). All participants completed standardized patient (SP) conversations before and after the intervention, which were scored by the SP and 4 third-party (TP) raters. MVP (medical situation, values, and plan) and 3Es (Empathize, Be Explicit, and Empower) [31] are a communication protocol and skills, respectively. UI: user interface; UX: user experience.



Recruitment and Randomization

We enrolled 51 participants, including medical, nursing, and physician assistant students, as well as practicing clinicians, from June to December 2024. Most participants were recruited from the University of Rochester School of Medicine and Dentistry, the University of Rochester School of Nursing, and Strong Memorial Hospital. Recruitment was conducted via email outreach, personal contacts, and campus flyers. Participants provided informed consent before participation. The study took place via the online video-calling platform Zoom (Zoom Communications Inc). Participants were randomized to the SOPHIE intervention (n=26) or the control (n=25). We used stratified block randomization to ensure balance across different participant classes (eg, trainees vs professionals). Prior participation in the SOPHIE pilot study was the only exclusion criterion. Each participant received a US $30 Amazon gift card as compensation. Table 1 presents the demographic breakdown of participants by age, gender, race or ethnicity, and clinical role.

XSL•FO
RenderX

Table . Participant demographics and dataset summary.

Characteristics	Control (N=25)	SOPHIE[a] (N=26)	Total (N=51)
Age (y), mean (SD)	30.0 (11.2)	31.1 (15.1)	—[b]
Gender, n (%)			
Women	18 (72.0)	22 (84.6)	40 (78.4)
Men	7 (28.0)	3 (11.5)	10 (19.6)
Prefer not to say	0 (0.0)	1 (3.8)	1 (2.0)
Race or ethnicity, n (%)			
White	12 (48.0)	17 (65.4)	29 (56.9)
Asian	5 (20.0)	6 (23.1)	11 (21.6)
Black	3 (12.0)	0 (0.0)	3 (5.9)
Latino or Hispanic	3 (12.0)	1 (3.8)	4 (7.8)
Middle Eastern or other	2 (8.0)	2 (7.7)	4 (7.8)
Clinical role, n			
Health care students	14	14	28
Practitioners (physician assistants, nurses, PhDs, and residents)	11	12	23
Dataset summary statistics, n			
Video conversations	—	—	102
Reviews collected	—	—	506
Standardized patient actors	—	—	13
Third-party raters	—	—	4
Reviews per video	—	—	5
Case distribution, n			
Case I: stage IIIa lung cancer	—	—	168
Case II: metastatic lung cancer	—	—	183
Case III: renal cell carcinoma	—	—	155

[a]SOPHIE: Standardized Online Patient for Healthcare Interaction Education.

[b]Not applicable.

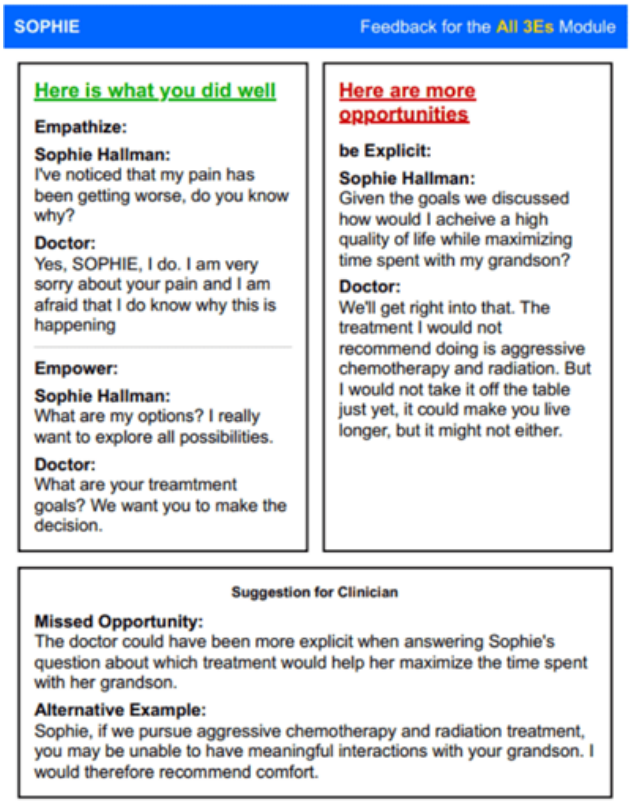
Intervention and Control Conditions

All participants viewed an instructional video on SIC and how to use the MVP framework and 3E skills [31]. The intervention group completed 3 SOPHIE modules, each focusing on 1 of 3 communication skills—Empathize, Be Explicit, and Empower—combining real-time, interactive dialogue with a virtual patient and immediate automated feedback. The Empathize module focused on responding to emotional distress, the Be Explicit module on clearly communicating prognosis and uncertainty, and the Empower module on eliciting values and supporting a patient-centered care plan. After each module, participants received personalized feedback on effective behaviors, missed opportunities, and suggested alternatives for the specific module. Examples of feedback and interface visuals are shown in Figure 2. Figure 3 shows a participant navigating through one of the SOPHIE modules and pre-SP and post-SP interactions.

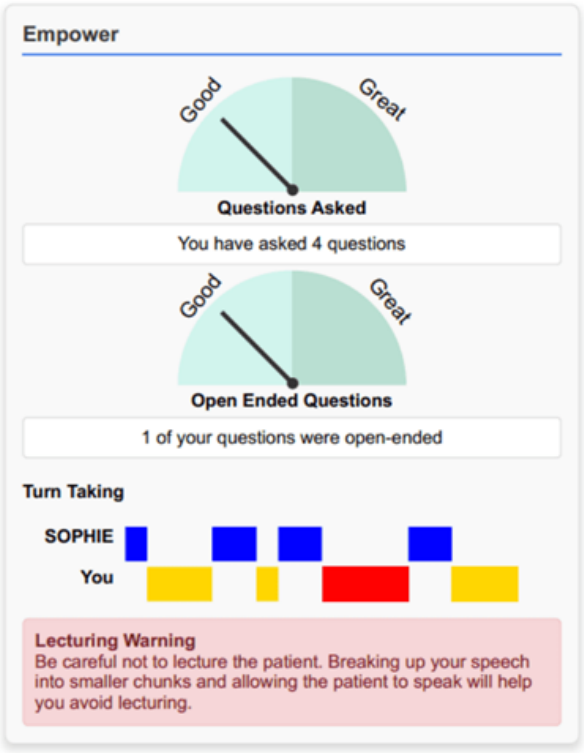
Control participants reviewed reading materials on the same SIC framework for the same equal 30-minute duration. The reading materials used in the control arm covered the same SIC framework and the same 3 target skills addressed in the SOPHIE modules: Empathize, Be Explicit, and Empower. Thus, both groups received the same core educational content, but only the SOPHIE group had the opportunity to practice these skills interactively with a virtual patient and to receive automated, personalized feedback. Control participants reviewed the reading materials for the same 30-minute duration as the SOPHIE training to maintain comparable exposure time across arms.

XSL•FO
RenderX

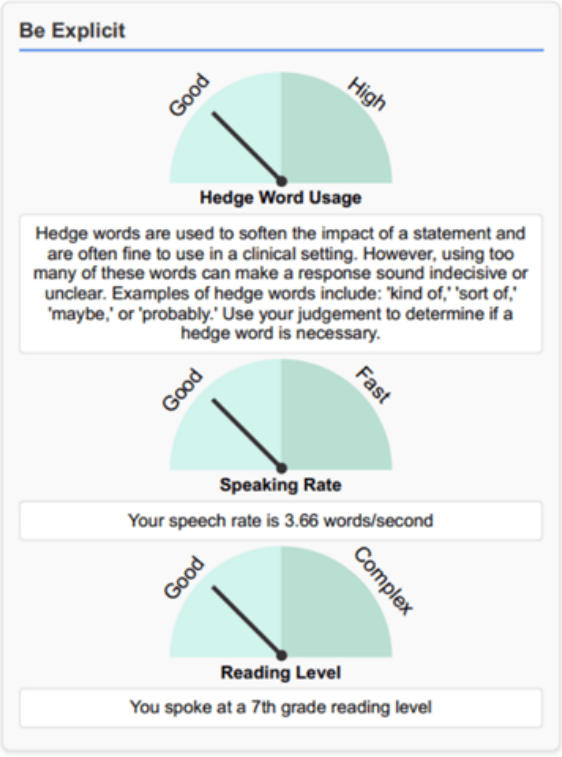
Figure 2. SOPHIE (Standardized Online Patient for Healthcare Interaction Education) feedback interface, simplified for visualization. Panel A shows the main dashboard with skill-based feedback. Panels B to D provide targeted guidance for the Empower, Be Explicit, and Empathize domains.



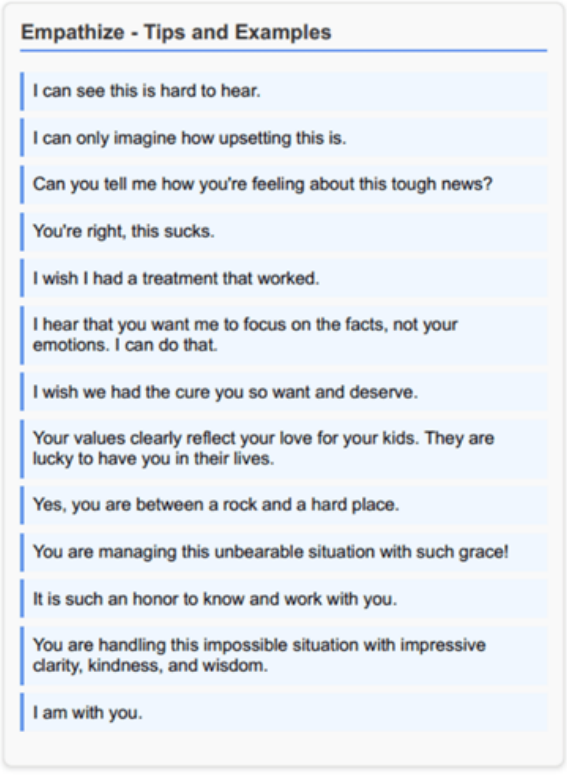
(A) AI-generated descriptive and actionable feedback



(B) Empower feedback



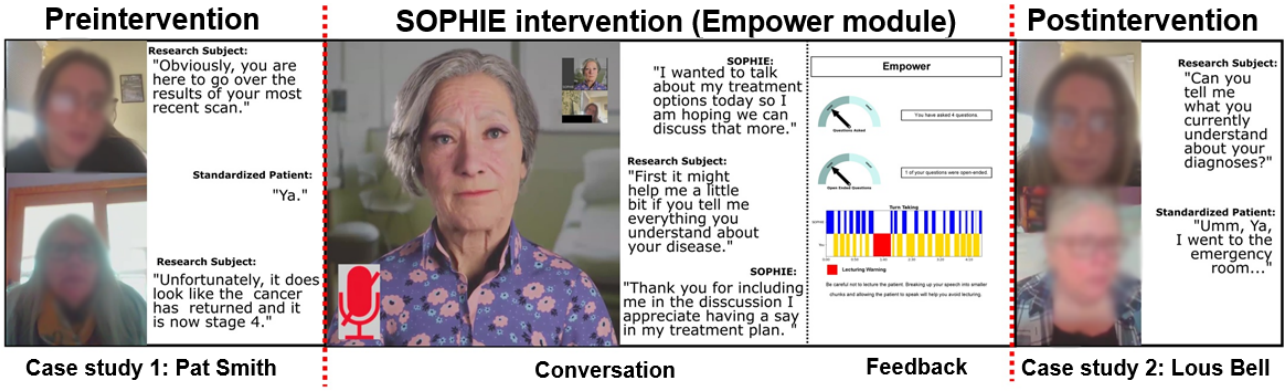
(C) Be Explicit feedback



(D) Empathy suggestions

XSL•FO
RenderX

Figure 3. Illustration of the impact of SOPHIE's (Standardized Online Patient for Healthcare Interaction Education) Empower module on communication. A participant initially delivers difficult news with limited patient engagement. Following SOPHIE's training and targeted feedback, the participant adopts a patient-centered approach, using open-ended questions to foster understanding and shared decision-making.



SP Encounters

All participants completed preintervention and postintervention simulated encounters with a trained SP drawn from 1 of 3 patient scenarios. Each participant was paired with 1 of 13 SPs, who interacted with the participant via video conference and rated the participant's communication skills using the previously described evaluation rubric [31,39]. The 13 SPs who directly interacted with participants were hired through an external service (Professional Communication Simulators). In each interaction, SPs were randomly assigned to 1 of 3 advanced cancer clinical cases (Table 1, case distribution) developed in collaboration with experts in communication training (RME and TC). The SPs were blinded to group assignment (intervention or control) but were aware of the encounter order (pre or post). We recorded a video of each participant-SP encounter and 4 additional actors, who work with the ACT Program (Advanced Communication Training Program, University of Rochester), which is directed by one of the authors (TC), watched the videos and provided a separate rating for both preintervention and postintervention participant-SP encounters. These actors were blinded to both the intervention and the encounter order. Both SPs and third-party (TP) raters completed an online form, described in Table S2 of Multimedia Appendix 1, to evaluate the participant.

Postinteraction Survey and Interview

After the intervention, all participants completed a postinteraction survey and a 10-minute interview (Tables S3-S7 in Multimedia Appendix 1). The shared survey asked both control and SOPHIE participants to rate perceived improvements in Empathize, Be Explicit, Empower, and overall communication skills, as well as stress during the SP interactions. Arm-specific survey questions differed by condition: control participants evaluated the usefulness and engagement of the reading materials, whereas SOPHIE participants evaluated the usefulness, engagement, stress, personalized feedback, and willingness to repeat the SOPHIE modules.

The interview included both structured and open-ended questions. Structured prompts asked all participants what they learned, what messages they found most important, whether the training improved their confidence or competence in the 3E skills, and whether they preferred interactive systems, passive materials, or a combination for communication training. Open-ended prompts allowed participants to describe their emotional experience, expected clinical takeaways, concerns, and suggestions for an ideal computer-based training system. We conducted and recorded the interviews through Zoom immediately after the experiment. These recordings were reviewed, and major themes were extracted from participant responses.

System Design

Overview

The SOPHIE system features a high-fidelity, lifelike virtual patient—a 60-year-old woman with incurable stage 4 lung cancer. Users engage with the system via a face-to-face video call and, after the interaction, receive personalized feedback based on the conversation. Together, the interactive virtual patient encounter and the automated feedback system constitute the SOPHIE intervention. Multimedia Appendix 2 provides a computational cost breakdown for operating our conversation system.

Virtual Avatar Interface

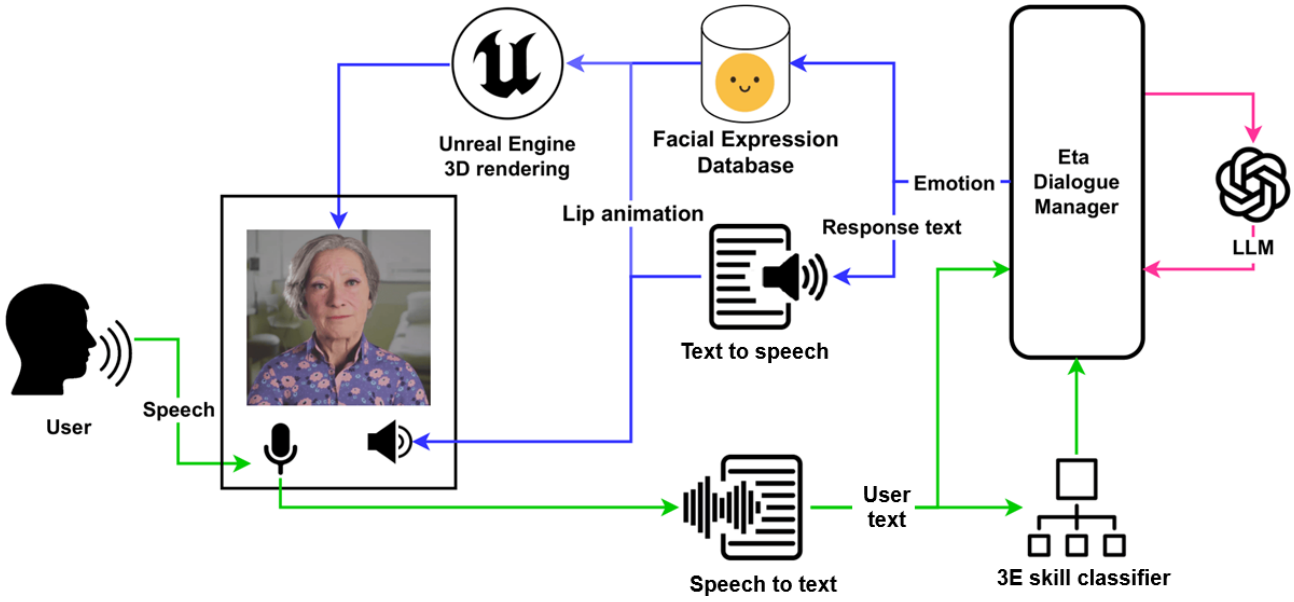
The virtual avatar used in this study was created using the Unreal Engine MetaHuman platform, a 3D computer graphics game engine developed by Epic Games. When the user speaks, their speech is transcribed by a speech-to-text model and processed by the dialogue manager. The dialogue manager generates a response and a context-appropriate emotion tag. The response is converted to audio using a text-to-speech system. For both speech recognition and synthesis, we use a commercial API, the Microsoft Azure Speech SDK. Using the emotion tag, the avatar can express emotions through facial expressions and voice. Alongside a neutral expression, the avatar can express 6 basic emotions through facial expressions: happy, sad, angry, disgusted, surprised, and afraid. Figure 4 displays examples of these emotional facial expressions. The facial expressions are prerecorded motion-capture animations created by the authors; when triggered, the necessary animation file is read from storage and streamed to the avatar. Figure 5 shows the system's data flow diagram.

XSL•FO
RenderX

Figure 4. SOPHIE (Standardized Online Patient for Healthcare Interaction Education) avatar emotional expressions: neutral (left), sad (middle), angry (right). Additional emotions include happy, surprise, fear, and disgust.



Figure 5. SOPHIE (Standardized Online Patient for Healthcare Interaction Education) user speech-to-avatar response generation flow. The user's speech is transcribed using a speech-to-text model, after which a dialogue manager guided by a large language model (LLM) decides the response and emotion. Then the response is played as audio using a text-to-speech model, and facial animation and emotion are displayed with the 3D avatar in Unreal Engine. The pipeline executes in near–real time. 3E: Empathize, Be Explicit, and Empower.



Dialogue Manager: Schema-Guided+LLM Hybrid

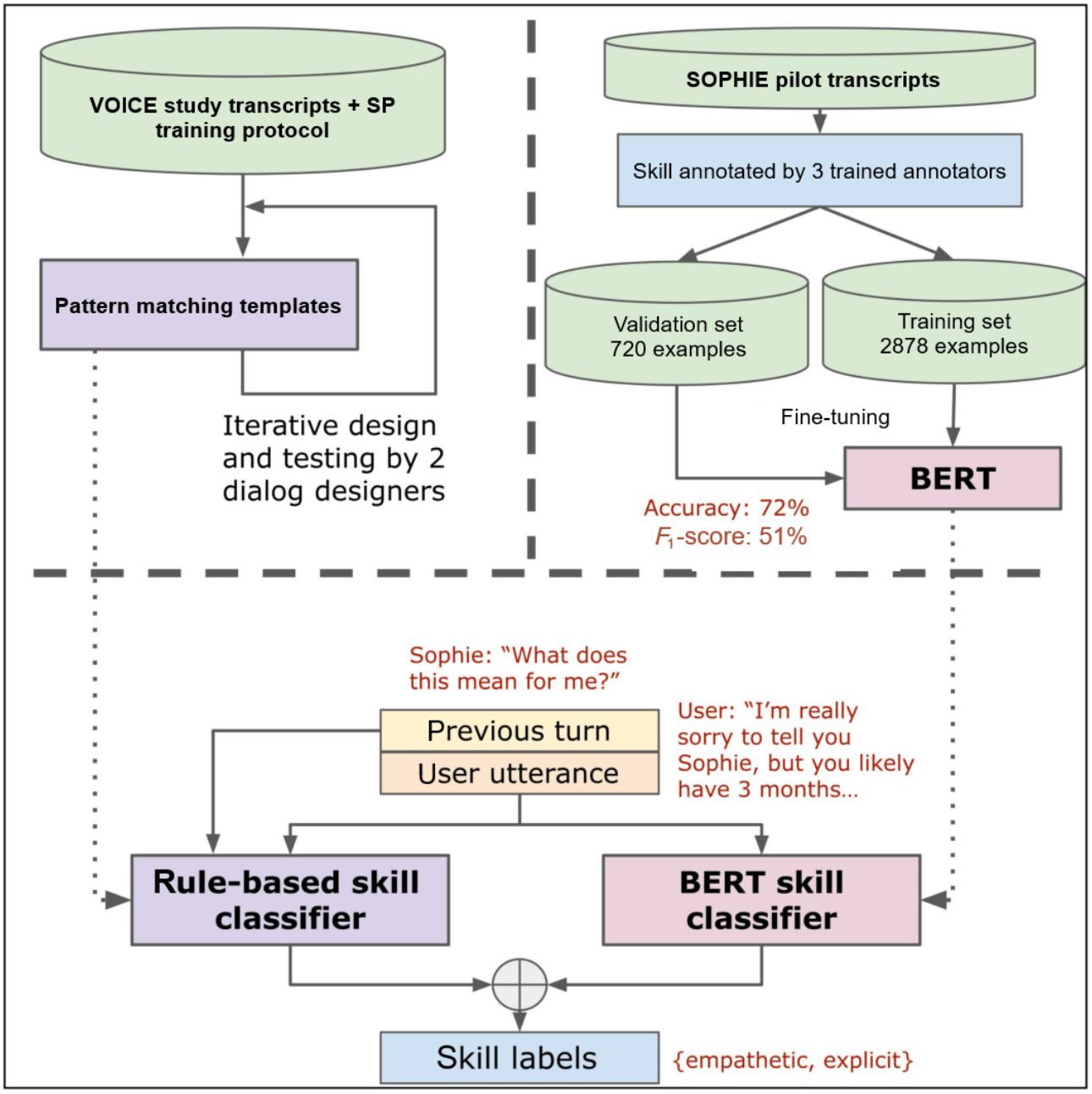
SOPHIE uses a hybrid dialogue system that balances pedagogical control with naturalistic interaction. In our previous work, we have explored using solely a schema-based dialogue system [34,39] and a fully LLM-based dialogue system [40,41], both of which had important limitations. The schema-based dialogue system is more controllable; however, the responses lack diversity and feel scripted. On the other hand, an LLM-based dialogue system, although more natural, is prone to diverging from the topic. In particular, older models such as ChatGPT (GPT-3.5 Turbo; OpenAI) were highly prone to hallucinating and drifting from the main conversation topic.

To create a controllable yet natural conversation system, we developed a hybrid dialog manager that uses a combination of a schema-based system and an LLM. The schema system is designed with the help of actual SIC conversations from the dataset in Epstein et al [33] and grounds the dialogue and provides guardrails. The schema also encodes the expected flow and communication goals per module (Empathize, Be Explicit, and Empower), tracks conversation state, and detects relevant user actions. Figure 6 demonstrates how such a skill is classified based on a user's utterance using a combination of rule-based and machine language–based classifiers. The schema system used in this study is a variation of the dialogue system described in Kane et al [34]. However, the main difference is that when the user's input aligns with one of the predefined schemas, instead of generating a full response, the system generates a structured template for SOPHIE's responses, including an emotion tag representing how the virtual patient is feeling. Next, an LLM (GPT-3.5 Turbo) is instructed to respond naturally to the user, strictly following the template. If the user's input does not match any predefined schema patterns, a response is generated directly by the LLM. This hybrid method gives SOPHIE the controllability and guardrails of schema-based dialogue, while leveraging the versatility of LLMs. The schema

XSL•FO
RenderX

and structured prompting components are intended to reduce dependence on the underlying LLM, although robustness across different models has not yet been evaluated.

Figure 6. A hybrid skill classifier that detects whether a statement from the user falls within the 3E (Empathize, Be Explicit, and Empower) skills. BERT (Bidirectional Encoder Representations from Transformers) is a pretrained deep learning model for text classification [42], and VOICE (Values and Options in Cancer Care) is a dataset consisting of real cancer care conversation transcripts [33]. SOPHIE: Standardized Online Patient for Healthcare Interaction Education; SP: standardized patient.

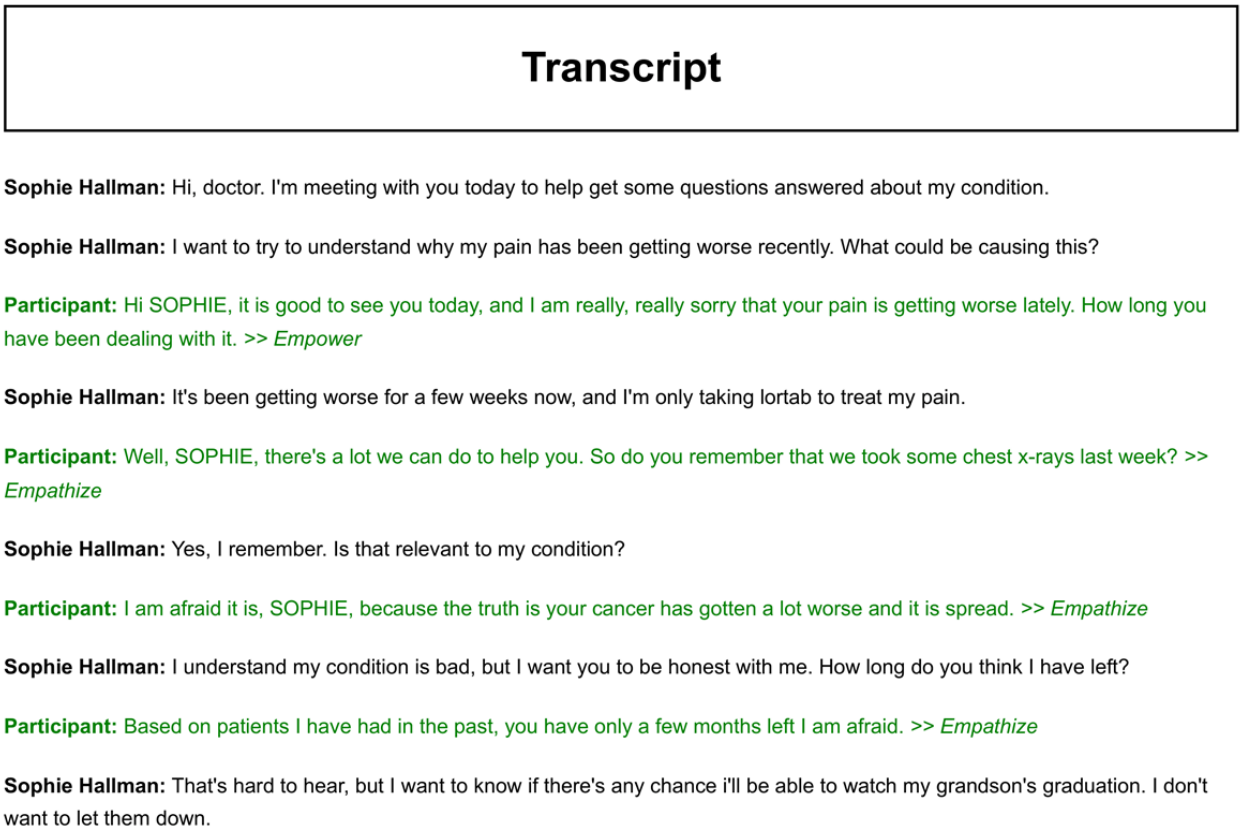


Feedback Delivery

The feedback model is grounded in deliberate practice and behavioral learning theory [35-37], co-designed with practicing clinicians and refined through pilot testing [38,39]. Following each SOPHIE module, users receive structured feedback aligned with the 3 target skills: Empathize, Be Explicit, and Empower. The feedback includes a combination of personalized LLM-generated feedback, examples of what the participant did well and of missed opportunities, suggested alternative phrasing, and skill-specific, structured metrics. The LLM-generated feedback is grounded in the MVP framework and is thus robust to LLM changes. In structured feedback, the Empower module provides feedback on questions asked, open-ended questions, and turn-taking; the Be Explicit module provides feedback on hedge words, speaking rate, and reading level; and the Empathize module provides example phrases for responding to patient emotion. All these metrics were carefully designed in previous studies to elicit each clinical communication skill [38,39]. Figure 2 shows examples of the feedback interface. Figure 7 shows an example of line-by-line feedback on the participant's conversation.

XSL•FO
RenderX

Figure 7. Part of a sample transcript from detailed feedback with the 3E (Empathize, Be Explicit, and Empower) skill classification. SOPHIE: Standardized Online Patient for Healthcare Interaction Education.



Scoring Rubrics to Measure Communication Skills

To our knowledge, there are no widely established benchmark scores for SIC performance. We therefore adopted the MVP framework and converted the 3 target skills—Empathize, Be Explicit, and Empower—into a structured Likert-scale questionnaire to quantify communication behaviors on a numeric scale. The full set of statements used in the rubric is provided in Table S2 of Multimedia Appendix 1. For each of the 3E skills (S), we summed the Likert scores associated with the skill and normalized them to between 0 and 1 using equation 1. For the overall score, questions related to all 3 skills were summed. This process was repeated for the ratings provided by the SP and the 4 TPs, and the resulting scores were averaged:

$$(1)\quad Score_S = \frac{\sum_{P=SP,TP1-4}\left(\sum_{j=\{question_set_S\}} Likert_score_value_j\right)}{max_point_possible_S}$$

We averaged the evaluations across 5 raters to reduce the effects of individual biases and normalized the scores to between 0 and 1 to make them interpretable and comparable across raters.

The rubric was intended to capture within-participant improvement from before to after the intervention, not to define an ideal level of performance or estimate population-level clinical competence. Future studies may use a similar rubric to evaluate whether repeated SOPHIE practice produces larger or more durable gains over time.

Statistical Procedures

We summarized participant demographics using descriptive statistics. Communication skill outcomes were analyzed for each of the 3Es (Empathize, Be Explicit, and Empower) and for the overall score. For descriptive reporting, we calculated pre-to-post change scores by subtracting each participant's first SP encounter score from their second encounter score.

To test whether improvement differed by study arm while accounting for repeated measurements within participants, we fit separate linear mixed effects models for Empower, Be Explicit, Empathize, and overall communication score. Each model included fixed effects for study arm, time (first vs second SP encounter), and the study arm-by-time interaction, with a random intercept for each participant to account for within-participant correlation. Models were estimated using maximum likelihood under an intention-to-treat framework, including all randomized participants, and were specified as:

score ~ study arm × time + (1 | participant).

The study arm-by-time interaction was the primary test of the intervention effect because it estimates the additional pre-to-post improvement in the SOPHIE arm relative to the control arm. Statistical significance was evaluated using α of .05, and 95% CIs were reported for model estimates.

XSL•FO
RenderX

Ethical Considerations

This study was approved by the University of Rochester Institutional Review Board (IRB; study ID: STUDY00006689). All participants provided informed consent after receiving an explanation of the study procedures, risks, and benefits, and their participation was voluntary. The study was conducted remotely using Zoom, and video meetings were recorded only with participants' consent. Study data and recordings were handled according to IRB-approved confidentiality procedures, stored in secure cloud storage, and deidentified before analysis or data sharing. Only aggregate demographic information is reported.

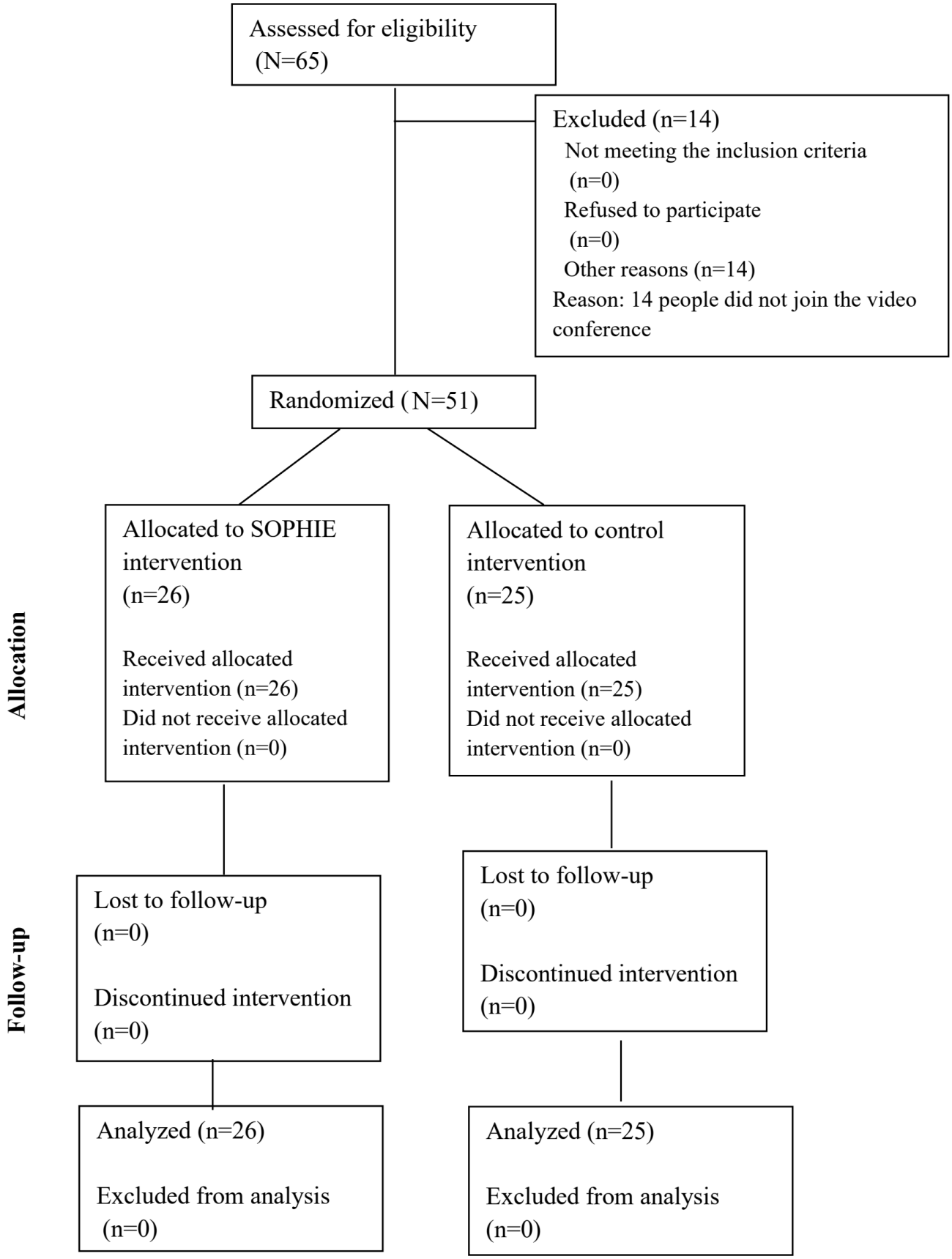
Results

Participant Characteristics

Both participant groups were demographically balanced, with no notable differences in age, gender, or clinical background (Table 1). Logistic regression confirmed no demographic predictors of group assignment. The CONSORT diagram is displayed in Figure 8. Of the 65 eligible participants who signed up, 14 did not show up at the scheduled meeting time. Of the 51 participants who showed up and signed the consent form, all completed the full study. Because the full study took place in a single, continuous session, we achieved a zero-dropout rate.

XSL•FO
RenderX

Figure 8. CONSORT (Consolidated Standards of Reporting Trials) diagram. SOPHIE: Standardized Online Patient for Healthcare Interaction Education.



Interrater Agreement

We examined agreement between the SP ratings and the blinded TP rating using a leave-one-out comparison. The full analysis is reported in Table S1 of Multimedia Appendix 1. For each rater, we compared that rater's score with a reference consensus formed by averaging the remaining raters' scores. SP refers to the SPs who interacted directly with participants, while TP0 to

XSL•FO
RenderX

TP3 refer to the 4 SPs who reviewed the recorded encounters as blinded TP raters. All ratings were normalized at the rater level before analysis to reduce individual scale-use differences.

The SP ratings showed moderate-to-strong agreement with the blinded TP rater consensus (Pearson correlation coefficient r=0.68; Spearman rank correlation ρ=0.67; mean absolute error=0.15; intraclass correlation coefficient [ICC] (A,1)=0.63; 95% CI 0.50 - 0.74). Agreement varied across individual TP raters but was generally within a similar range, with ICC(A,1) values from 0.57 to 0.78. These results suggest that SPs and blinded TP raters captured broadly consistent communication quality while also reflecting meaningful differences in perspective between first-person interaction and TP observation.

Primary Outcome: Communication Skill Improvement

To evaluate communication skill development, we analyzed changes in participants' scores (Empower, Be Explicit, Empathize, and overall) between their first and second SP conversations. Each conversation (n=102) received 5 independent ratings—1 from the participating SP and 4 from blinded TP reviewers. Final scores were averaged across these ratings and change scores (Δ) were calculated by subtracting preintervention from postintervention values and then aggregated as described earlier. Of the expected 510 ratings from 51 participants across 2 SP encounters and 4 independent raters, 4 ratings were missing, resulting in 506 ratings for the final analysis. Interrater reliability was strong, with an ICC of 0.88 (95% CI 0.82 - 0.93) among TP reviewers blinded to intervention or control status before and after the study.

Both groups showed improvement from the first to the second SP encounter, but the SOPHIE group showed larger gains across all communication outcomes (Table 2). For Empower, mean scores increased from 0.49 to 0.55 (Δ=0.06) in the control group and from 0.41 to 0.58 (Δ=0.17) in the SOPHIE group. Similar patterns were observed for Be Explicit (control Δ=0.05; SOPHIE Δ=0.13), Empathize (control Δ=0.07; SOPHIE Δ=0.14), and overall score (control Δ=0.06; SOPHIE Δ=0.15).

To account for the repeated-measures structure of the randomized pre-post design, we analyzed communication outcomes using linear mixed effects models under an intention-to-treat framework. All randomized participants (n=51; 102 observations) were included in the analyses.

The mixed effects models showed significant study arm-by-time interactions for all outcomes. The SOPHIE arm showed greater pre-to-post improvement than the control arm for Empower (β=.11, 95% CI .04-.17; P=.002), Be Explicit (β=.08, 95% CI .03-.13; P=.002), Empathize (β=.07, 95% CI .01-.13; P=.03), and overall communication score (β=.09, 95% CI .04-.14; P=.001). In practical terms, these interaction estimates represent the additional improvement associated with SOPHIE beyond the improvement observed in the control arm.

Table . Communication skill improvement in control and SOPHIE (Standardized Online Patient for Healthcare Interaction Education) arms.

Arm	Control			SOPHIE				P value[a]
	Pre	Post	Δ[b]	Pre	Post	Δ[b]	LLM[c] interaction β[d] (95% CI[e])	
Empower	0.49	0.55	0.06	0.41	0.58	0.17	.11 (.04-.17)	.002
Be Explicit	0.71	0.75	0.05	0.66	0.79	0.13	.08 (.03-.13)	.002
Empathize	0.58	0.65	0.07	0.51	0.65	0.14	.07 (.01-.13)	.03
Overall	0.59	0.65	0.06	0.52	0.67	0.15	.09 (.04-.14)	.001

[a]P indicates whether the difference in improvement between arms was statistically significant.

[b]Δ represents the difference between the second and first encounter mean scores according to the scoring rubric.

[c]LLM: large language model.

[d]The LLM interaction β value represents the additional pre-to-post improvement in the SOPHIE arm relative to the control arm, estimated from linear mixed effects models.

[e]The 95% CI indicates the uncertainty around this estimate.

Secondary Outcomes: Learner Perceptions and System Efficacy

Participants using the SOPHIE system reported markedly greater confidence in their communication skill development compared with controls on a Likert scale ranging from 1 to 5, with average perceived improvement scores of 4.31 (SOPHIE) vs 3.83 (control) on a 5-point scale (P=.009).

The training system itself received strong evaluations, particularly for its actionable feedback on communication skills—the highest-rated feature (mean: 4.7/5, SD: 0.18). Users also praised SOPHIE's usability (mean: 4.2/5, SD: 0.36), though ratings for the virtual human's realism were more moderate (mean: 3.4/5, SD: 0.17).

Qualitative Analysis

Overview

Qualitative feedback from post-session participant interviews was video-recorded and analyzed by the author KGH, revealing 3 dominant themes: the inherent difficulty of serious illness conversations, SOPHIE's value as a training tool, and opportunities for system refinement.

Apprehension in SIC

Participants universally acknowledged the emotional complexity of serious illness dialogues, with many expressing surprise at

XSL•FO
RenderX

their own discomfort. One participant admitted, "I felt nervous—even really, really bad—practicing these conversations," while another emphasized, "Doctors are human too—nobody is born an expert." Participants generally described this discomfort as arising from the emotionally difficult nature of SIC rather than from the SOPHIE system itself. One of the study authors (KGH) remained available throughout each interaction to provide support if needed. No participant discontinued the session or requested additional support because of distress.

Training Value of SOPHIE

The system's accessibility and actionable feedback were widely praised. Participants likened the experience to "being trained in a facility by a specialist, but from my laptop," highlighting its convenience without compromising pedagogical rigor. The specificity of feedback—described as "actionable guidance, not just generic encouragement"—enabled targeted skill development. Notably, the judgment-free environment empowered experimentation: "I tried new strategies because there was no risk of hurting a real patient," one participant explained. Repeated practice translated to tangible confidence gains, with users reporting feeling "more prepared for real conversations."

System Refinement

While effective, SOPHIE's realism faced criticism. Participants noted "creepy" or "uncanny" interactions due to limited emotional range, abrupt speech patterns, and lack of eye contact. Robotic reinforcement phrases (eg, "Thank you for allowing me to express my emotions") further reduced perceived authenticity. To address these gaps, users proposed enhancing multimodal analysis (facial expressions, tone, body language) and integrating game elements (eg, performance scoring and competition).

Beyond health care, participants recognized the 3E framework's potential in business, law, education, and personal relationships.

Discussion

Principal Findings

This RCT tested whether SOPHIE would lead to greater improvement in SIC skills and confidence than did written educational materials. The findings support this hypothesis: participants assigned to SOPHIE showed greater pre-to-post improvement than controls across Empower, Be Explicit, Empathize, and overall communication scores, as well as higher self-reported confidence. These gains occurred after a brief training session of less than 30 minutes, suggesting that AI-based virtual patient training may be feasible within clinical education settings.

Confidence gains were also higher with SOPHIE, indicating that interactive practice not only improved competence but also helped learners recognize their progress. Interviews reinforced this: many participants felt underprepared for the emotional demands of SIC, underscoring the value of deliberate, low-risk practice.

Participants rated SOPHIE's actionable feedback (4.7/5) higher than the avatar's visual realism (3.4/5), suggesting that visual fidelity has room for improvement.

Comparison With Prior Works

Prior work on LLM-based virtual patients has largely focused on system feasibility, usability, and intrinsic quality rather than on behavior learning outcomes. Studies by Elhilali et al [25] and Cook et al [27] demonstrate that LLM-driven patient simulations can generate realistic dialogue and are perceived as usable, but they do not assess whether such systems improve clinician communication skills. A recent systematic review [26] similarly highlights the lack of consistent outcome measures and the limited evidence of skill transfer in the literature. Importantly, very little published work has evaluated AI-based training specifically for SIC, and the available literature is particularly limited in randomized trials with behavior outcomes.

Some randomized studies evaluate AI-based simulations but rely mainly on self-reported outcomes. Tyrrell et al [29] and Roux et al [30] report improvements in perceived skills or confidence, yet do not show clear gains in objectively assessed communication behaviors. Other work using virtual SPs without LLM-driven dialogue has demonstrated improvements in clinical reasoning and diagnostic skills [28], but these systems do not target conversational communication or provide generative, personalized feedback. Thus, although AI-based virtual patient systems are increasingly being studied in medical education, there remains a dearth of evidence on whether such systems can improve SIC skills that transfer to human SP encounters.

In contrast, this study evaluates an LLM-based virtual patient designed for deliberate practice in SIC and measures effectiveness using a randomized controlled design with multirater behavior assessments. This work extends prior studies by providing evidence that generative AI–based patient simulations can lead to measurable improvements in clinician communication behaviors beyond usability or self-reported learning gains.

Future Directions and Scalability

As health systems embrace telemedicine and remote education, scalable and equitable training is essential. SOPHIE leverages generative AI to standardize, measure, and democratize communication training, with the potential to mitigate bias by enabling practice with diverse patient identities. Even modest clinician gains at scale could reduce disparities in outcomes. Applications in conflict resolution, medical error disclosure, and misinformation highlight its versatility. Figure 9 demonstrates diverse appearances of virtual avatars that can be easily integrated into our system.

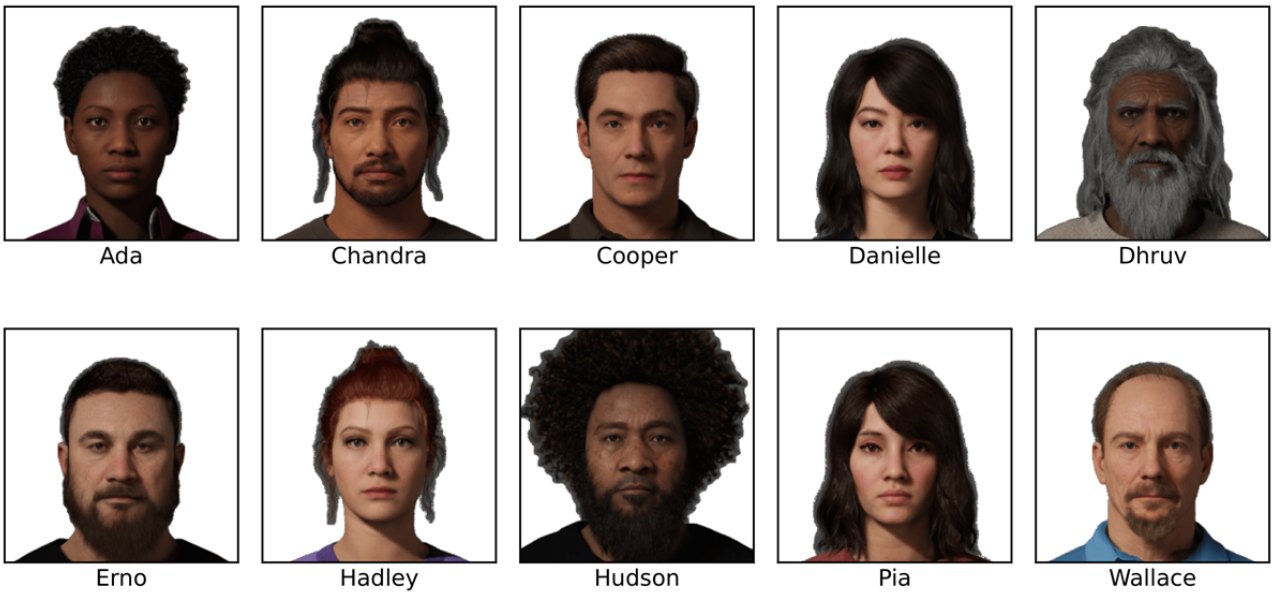
These findings provide emerging empirical support for the hypothesis that AI can serve as an adaptive tutor in complex interpersonal domains [35]. By enabling deliberate practice with structured, immediate feedback [36,37,43], AI systems can support skill development traditionally cultivated through in-person mentorship.

Importantly, our assessment focused not only on participant engagement with the training platform but also on whether skills

XSL•FO
RenderX

translated to performance in encounters with SPs. This approach supports the evaluation of skill transfer beyond the virtual training environment—an important consideration for tools intended to support clinical practice. Looking ahead, systems like SOPHIE may serve as a foundation for the use of AI that strengthens, rather than replaces, the clinician's voice in moments that matter most.

Figure 9. Virtual avatars can be customized to cover a wide range of demographics, including race, gender, age, and cultural background. This could potentially enable clinicians to practice highly sensitive communication in ways that were previously difficult to achieve at scale with real humans.



Limitations

Although our approach has several strengths, the long-term effects of SOPHIE-based training remain unknown. Additionally, the assessment of improvements in communication skills was conducted in a simulated setting; future work is needed to assess its real-world clinical impact. The observed learning effects may also have been influenced by novelty or motivational effects, as many participants were likely interacting with a virtual patient for the first time. In addition, the single-scenario training, the relatively small sample size, and the same-day interaction and evaluation limit the generalizability of the findings. Our control condition measured the effectiveness of SOPHIE against a static reading condition. However, how the SOPHIE intervention would compare with other forms of interactive training, such as SP-based training and feedback, remains still unknown. Future studies should include more diverse virtual patients and clinical cases, larger samples, longer follow-up periods, and comparisons of AI training with human training. Four of the expected 510 ratings were missing; given the small proportion of missing ratings, their effect was likely limited, although some measurement uncertainty remains.

Our dialogue agent captures the user's speech by transcribing it into text and does not take into account the user's facial expressions, prosody, intonation, body language, or other nonverbal cues. While participants recognized emotions such as sadness and concern during interactions, some noted that SOPHIE's expressions felt "uncanny." This highlights a common design challenge: avatars that look too human can be unsettling, while more stylized expressions may actually make emotional cues clearer. Whether intentionally favoring exaggerated, easy-to-read expressions is educationally superior to striving for subtle, lifelike realism should be explored further in future studies.

One important limitation of the SOPHIE system is its dependence on commercial providers for LLMs, speech recognition, and speech synthesis. This reliance can hinder exact reproducibility over time because cloud-based models and APIs may be updated or deprecated. In addition, transmitting conversation data to external services may limit the use of SOPHIE for protected or highly sensitive clinical conversations, even when appropriate safeguards are in place. These concerns could be mitigated by deploying local models, such as the Llama 3 [44] or Qwen 2.5 [45] families for language generation, Whisper [46] for speech recognition, and Kokorofor speech synthesis, allowing all processing to remain within institutional infrastructure. However, such deployments introduce additional engineering complexity and require dedicated hardware resources and ongoing maintenance, potentially increasing the operational burden for academic or health care institutions.

Conclusions

Looking ahead, systems like SOPHIE may serve as a foundation for a new kind of AI-human partnership in health care—one that strengthens, rather than replaces, the clinician's voice in moments that matter most.

XSL•FO
RenderX

Acknowledgments

The authors are grateful to Prof Anson Khang of the University of Rochester for helping to review our statistical analysis plan. The authors used generative AI tools such as Grammarly (Superhuman Platform Inc), ChatGPT (GPT-4o, GPT-5; OpenAI), and Gemini (2.0 Pro, 2.5 Pro; Google AI) as writing helpers and coding assistants throughout the preparation of the manuscript. All generative AI content was used as suggestions and thoroughly reviewed by the authors before being included in the manuscript.

Funding

Part of this work was funded by the Wilmot Team Science Pilot Program at the University of Rochester Medical Center. The funding body played no role in study design, data collection, analysis, inspection, results, or the writing of the manuscript.

Data Availability

The anonymized rating dataset (n=506) used in this study and all analysis code are publicly available at ROC-HCI Lab [47] under the MIT License. These materials are sufficient to reproduce the reported results. Participant and rater identifiers have been removed. Due to institutional review board (IRB) and consent constraints, the original videos and timestamped transcripts are not publicly available but may be shared with qualified researchers upon reasonable request and with appropriate IRB approval.

Authors' Contributions

KGH led the project, including study design, institutional review board approval, participant recruitment, data collection, early analyses, thematic analysis, and manuscript drafting. MH developed the core technical components of SOPHIE's (Standardized Online Patient for Healthcare Interaction Education) verified analyses, created figures and tables, and restructured and refined the manuscript. TC ensured integration of the 3E (Empathize, Be Explicit, and Empower) communication framework, contributed clinical expertise, and supported the design and review. RME co-conceptualized the SOPHIE system, shaped the virtual patient persona and feedback, and contributed to study design and manuscript revisions. TS performed statistical modeling and hypothesis testing and supported data analysis and visualization. EH served as principal investigator, coconceptualized SOPHIE secured funding, supervised the project, and provided strategic oversight and critical revisions.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Additional tables.
[DOCX File, 35 KB - mededu_v12i1e93034_app1.docx]

Multimedia Appendix 2
Additional methods.
[DOCX File, 28 KB - mededu_v12i1e93034_app2.docx]

Checklist 1
CONSORT checklist.
[PDF File, 295 KB - mededu_v12i1e93034_app3.pdf]

References

1. Shilling DM, Manz CR, Strand JJ, Patel MI. Let us have the conversation: serious illness communication in oncology: definitions, barriers, and successful approaches. Am Soc Clin Oncol Educ Book 2024 Jun;44(3):e431352. [doi: 10.1200/EDBK_431352] [Medline: 38788187]
2. Pasricha V, Gorman D, Laothamatas K, Bhardwaj A, Ganta N, Mikkelsen ME. Use of the Serious Illness Conversation Guide to improve communication with surrogates of critically ill patients. A pilot study. ATS Sch 2020 Mar 18;1(2):119-133. [doi: 10.34197/ats-scholar.2019-0006OC] [Medline: 33870276]
3. Paladino J, Sanders JJ, Fromme EK, et al. Improving serious illness communication: a qualitative study of clinical culture. BMC Palliat Care 2023 Jul 22;22(1):104. [doi: 10.1186/s12904-023-01229-x] [Medline: 37481530]
4. Myers J, Steinberg L, Incardona N, Simon J, Sanders J, Seow H. Simplifying serious illness communication: preparing or deciding. Curr Oncol 2024 Sep 28;31(10):5832-5837. [doi: 10.3390/curroncol31100433] [Medline: 39451737]
5. Hagerty RG, Butow PN, Ellis PM, Dimitry S, Tattersall MHN. Communicating prognosis in cancer care: a systematic review of the literature. Ann Oncol 2005 Jul;16(7):1005-1053. [doi: 10.1093/annonc/mdi211] [Medline: 15939716]
6. Korsch BM, Negrete VF. Doctor-patient communication. Sci Am 1972 Aug;227(2):66-74. [doi: 10.1038/scientificamerican0872-66] [Medline: 5044413]

XSL•FO
RenderX

7. Noah BA, Feigenson NR. Avoiding overtreatment at the end of life: physician-patient communication and truly informed consent. Pace Law Rev 2016;36(3):737-787. [doi: 10.58948/2331-3528.1927]
8. Clapp JT, Schwarze ML, Fleisher LA. Surgical overtreatment and shared decision-making-the limits of choice. JAMA Surg 2022 Jan 1;157(1):5-6. [doi: 10.1001/jamasurg.2021.4425] [Medline: 34643671]
9. Shaffer VA, Scherer LD. Too much medicine: behavioral science insights on overutilization, overdiagnosis, and overtreatment in health care. Policy Insights Behav Brain Sci 2018;5(2):155-162. [doi: 10.1177/2372732218786042]
10. Stewart MA. Effective physician-patient communication and health outcomes: a review. CMAJ 1995 May 1;152(9):1423-1433. [Medline: 7728691]
11. Ha JF, Longnecker N. Doctor-patient communication: a review. Ochsner J 2010;10(1):38-43. [Medline: 21603354]
12. Arnsten AFT. Stress signalling pathways that impair prefrontal cortex structure and function. Nat Rev Neurosci 2009 Jun;10(6):410-422. [doi: 10.1038/nrn2648] [Medline: 19455173]
13. Humphrey KE, Sundberg M, Milliren CE, Graham DA, Landrigan CP. Frequency and nature of communication and handoff failures in medical malpractice claims. J Patient Saf 2022 Mar 1;18(2):130-137. [doi: 10.1097/PTS.0000000000000937] [Medline: 35188927]
14. Gruneir A, Mor V, Weitzen S, Truchil R, Teno J, Roy J. Where people die: a multilevel approach to understanding influences on site of death in America. Med Care Res Rev 2007 Aug;64(4):351-378. [doi: 10.1177/1077558707301810] [Medline: 17684107]
15. Gilligan T, Coyle N, Frankel RM, et al. Patient-clinician communication: American Society of Clinical Oncology consensus guideline. J Clin Oncol 2017 Nov 1;35(31):3618-3632. [doi: 10.1200/JCO.2017.75.2311] [Medline: 28892432]
16. Epstein RM, Street RL Jr. Patient-centered communication in cancer care: promoting healing and reducing suffering. : National Cancer Institute, U.S. Department of Health and Human Services, National Institutes of Health; 2007 URL: https://cancercontrol.cancer.gov/sites/default/files/2020-06/pcc_monograph.pdf [accessed 2025-06-09]
17. Bosse HM, Nickel M, Huwendiek S, Schultz JH, Nikendei C. Cost-effectiveness of peer role play and standardized patients in undergraduate communication training. BMC Med Educ 2015 Dec;15(1):183. [doi: 10.1186/s12909-015-0468-1]
18. Menez O, Zagales R, Colleton T, Rodriguez Artze C, Cordero L. Strategies and challenges to diversifying standardized patients at a rural regional campus. Presented at: Indiana University School of Medicine (IUSM) Education Day 2024; Apr 26, 2024 URL: https://scholarworks.indianapolis.iu.edu/server/api/core/bitstreams/43cdac00-861d-46b2-810a-235c5d22e788/content [accessed 2026-02-06]
19. Gillette C, Stanton RB, Rockich-Winston N, Rudolph M, Anderson HG Jr. Cost-effectiveness of using standardized patients to assess student-pharmacist communication skills. Am J Pharm Educ 2017 Dec;81(10):6120. [doi: 10.5688/ajpe6120] [Medline: 29367775]
20. Everett MR, May W, Nowels CT, Main DS. Recruitment, retention, and training of African American and Latino standardized patients: a collaborative study. Med Sci Educ 2005;15:74-80 [FREE Full text]
21. Heinz MV, Mackin DM, Trudeau BM, et al. Randomized trial of a generative AI chatbot for mental health treatment. NEJM AI 2025;2(4). [doi: 10.1056/AIoa2400802]
22. Bowers P, Graydon K, Ryan T, Lau JH, Tomlin D. Artificial intelligence–driven virtual patients for communication skill development in healthcare students: a scoping review. Australas J Educ Technol 2024;40(3):39-57. [doi: 10.14742/ajet.9307]
23. Wang Z, Hassan N, LeBaron V, et al. CommSense: a wearable sensing computational framework for evaluating patient-clinician interactions. Proc ACM Hum-Comput Interact 2024 Nov 7;8:1-31. [doi: 10.1145/3686952]
24. Kearns WR, Bertram J, Divina M, et al. Bridging the skills gap: evaluating an AI-assisted provider platform to support care providers with empathetic delivery of protocolized therapy. AMIA Annu Symp Proc 2024;2023:436-445. [Medline: 38222441]
25. Elhilali A, Ngo ASH, Reichenpfader D, Denecke K. Large language model–based patient simulation to foster communication skills in health care professionals: user-centered development and usability study. JMIR Med Educ 2025 Dec 12;11:e81271. [doi: 10.2196/81271] [Medline: 41385781]
26. Li D, Lebai Lutfi S. Large language model–based virtual patient systems for history-taking in medical education: comprehensive systematic review. JMIR Med Inform 2026 Jan 2;14:e79039. [doi: 10.2196/79039] [Medline: 41481915]
27. Cook DA, Overgaard J, Pankratz VS, Del Fiol G, Aakre CA. Virtual patients using large language models: scalable, contextualized simulation of clinician-patient dialogue with feedback. J Med Internet Res 2025 Apr 4;27:e68486. [doi: 10.2196/68486] [Medline: 39854611]
28. Xu L, Xu Q, Liu C, Chen B, Wang C. Virtual standardized patients for improving clinical thinking ability training in residents: randomized controlled trial. JMIR Med Educ 2025 Dec 8;11:e73196. [doi: 10.2196/73196] [Medline: 41359953]
29. Tyrrell EG, Sandhu SK, Berry K, et al. Web-based AI-driven virtual patient simulator versus actor-based simulation for teaching consultation skills: multicenter randomized crossover study. JMIR Form Res 2025 Nov 20;9:e71667. [doi: 10.2196/71667] [Medline: 41264856]
30. Roux P, Okuya Y, Morel C, et al. Effectiveness of a web-based virtual simulation to train nursing students in suicide risk assessment: randomized controlled investigation. JMIR Serious Games 2025 Aug 1;13:e69347. [doi: 10.2196/69347] [Medline: 40749220]

XSL•FO
RenderX

31. Horowitz RK, Hogan LA, Carroll T. MVP-Medical Situation, Values, and Plan: a memorable and useful model for all serious illness conversations. J Pain Symptom Manage 2020 Nov;60(5):1059-1065. [doi: 10.1016/j.jpainsymman.2020.07.022] [Medline: 32738279]
32. Sen T, Ali MR, Hoque ME, Epstein R, Duberstein P. Modeling doctor-patient communication with affective text analysis. Proc 7th Int Conf Affect Comput Intell Interact (ACII 2017) 2017:170-177. [doi: 10.1109/ACII.2017.8273596]
33. Epstein RM, Duberstein PR, Fenton JJ, et al. Effect of a patient-centered communication intervention on oncologist-patient communication, quality of life, and health care utilization in advanced cancer: the VOICE randomized clinical trial. JAMA Oncol 2017 Jan 1;3(1):92-100. [doi: 10.1001/jamaoncol.2016.4373] [Medline: 27612178]
34. Kane B, Giugno C, Schubert L, Haut K, Wohn C, Hoque E. Managing emotional dialogue for a virtual cancer patient: a schema-guided approach. IEEE Trans Affective Comput 2024;15(3):1041-1052. [doi: 10.1109/TAFFC.2023.3317805]
35. Bereiter C, Scardamalia M. Surpassing Ourselves: An Inquiry into the Nature and Implications of Expertise: Open Court Publishing Company; 1993.
36. Ende J. Feedback in clinical medical education. JAMA 1983 Aug 12;250(6):777-781. [Medline: 6876333]
37. Borrell-Carrio F, Molloy EK, Epstein R. The impact of emotions in feedback. In: Boud D, Molloy E, editors. Feedback in Higher and Professional Education: Understanding It and Doing It Well: Routledge; 2013:50-71.
38. Ali MR, Sen T, Kane B, et al. Novel computational linguistic measures, dialogue system and the development of SOPHIE: Standardized Online Patient for Healthcare Interaction Education. IEEE Trans Affective Comput 2023;14(1):223-235. [doi: 10.1109/TAFFC.2021.3054717]
39. Haut K, Wohn C, Kane B, et al. Validating a virtual human and automated feedback system for training doctor-patient communication skills. Proc 11th Int Conf Affect Comput Intell Interact (ACII 2023) 2023:1-8. [doi: 10.1109/ACII59096.2023.10388213]
40. Hasan M, Ozel C, Potter S, Hoque E. SAPIEN: affective virtual agents powered by large language models *. Proc 11th Int Conf Affect Comput Intell Interact Workshops Demos (ACIIW 2023) 2023:1-3. [doi: 10.1109/ACIIW59127.2023.10388188]
41. Baseman CM, Hasan M, Swinger N, Rauch SAM, Hoque E, Arriaga RI. “Poker with play money”: exploring psychotherapist training with virtual patients. Proc ACM Hum-Comput Interact 2025 Oct 18;9(7):1-29. [doi: 10.1145/3757450]
42. Devlin J, Chang MW, Lee K, Toutanova K. BERT: pre-training of deep bidirectional transformers for language understanding. In: Burstein J, Doran C, Solorio T, editors. Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Association for Computational Linguistics; 2019:4171-4186. [doi: 10.18653/v1/N19-1423]
43. Thomas JD, Arnold RM. Giving feedback. J Palliat Med 2011 Feb;14(2):233-239. [doi: 10.1089/jpm.2010.0093] [Medline: 21314576]
44. Grattafiori A, Dubey A, Jauhri A, Pandey A, et al. The Llama 3 herd of models. arXiv. Preprint posted online on Jul 31, 2024. [doi: 10.48550/arXiv.2407.21783]
45. Yang A, Yang B, Zhang B, et al. Qwen2.5 technical report. arXiv. Preprint posted online on Dec 19, 2024. [doi: 10.48550/arXiv.2412.15115]
46. Radford A, Kim JW, Xu T, Brockman G, McLeavey C, Sutskever I. Robust speech recognition via large-scale weak supervision. In: Krause A, Brunskill E, Cho K, Engelhardt B, Sabato S, Scarlett J, editors. ICML’23: Proceedings of the 40th International Conference on Machine Learning: Association for Computing Machinery (ACM); 2023:28492-28518 URL: https://dl.acm.org/doi/10.5555/3618408.3619590 [accessed 2026-08-06]
47. ROC-HCI/SOPHIE-1.0. GitHub. URL: https://github.com/ROC-HCI/SOPHIE-1.0 [accessed 2026-06-11]

Abbreviations

3Es: Empathize, Be Explicit, and Empower
ACT Program: Advanced Communication Training Program
CONSORT: Consolidated Standards of Reporting Trials
ICC: intraclass correlation coefficient
LLM: large language model
MVP: medical situation, values, and plan
RCT: randomized controlled trial
SIC: serious illness communication
SOPHIE: Standardized Online Patient for Healthcare Interaction Education
SP: standardized patient
TP: third party

XSL•FO
RenderX

Edited by T Mühling; submitted 20.Feb.2026; peer-reviewed by D Bainbridge, M Ikuno; revised version received 12.Jun.2026; accepted 15.Jun.2026; published 20.Oct.2026.

Please cite as:
Haut KG, Hasan M, Carroll T, Epstein RM, Sen T, Hoque E
Effects of a Generative AI Virtual Patient on Serious Illness Communication Skills: Randomized Controlled Trial
JMIR Med Educ 2026;12:e93034
URL: https://mededu.jmir.org/2026/1/e93034
doi:10.2196/93034

© Kurtis G Haut, Masum Hasan, Thomas Carroll, Ronald M Epstein, Taylan Sen, Ehsan Hoque. Originally published in JMIR Medical Education (https://mededu.jmir.org), 20.Oct.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Education, is properly cited. The complete bibliographic information, a link to the original publication on https://mededu.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Students' Perspectives on Digital Psychotherapy—Possible Solutions for Digital Inpatient-Like Care Concepts: Qualitative Interview Study

Rebekka Robitzsch[1,2], MD; Tania Lalgi[1,2], MSc; Alexander Diel[1,2], PhD; Sophie Schulz genannt Menningmann[1,2], BSc; Lucy Ann Gresser[1,2], MD; Patrick Jonas Wollenberg[1,2], MD; Martin Teufel[1,2], MD; Alexander Bäuerle[1,2], PhD; Anita Robitzsch[1,2], MD

[1]Clinic for Psychosomatic Medicine and Psychotherapy, LVR-University Hospital, University of Duisburg-Essen, Virchowstr, 174, Essen, Germany
[2]Center for Translational Neuro- and Behavioral Sciences (C-TNBS), University of Duisburg-Essen, Essen, Germany

Corresponding Author:
Alexander Bäuerle, PhD
Clinic for Psychosomatic Medicine and Psychotherapy, LVR-University Hospital, University of Duisburg-Essen, Virchowstr, 174, Essen, Germany

Abstract

Background: The demand for mental health treatment is increasing, while the availability of treatment remains insufficient to meet the rising demand. Alternative solutions need to be explored to enable access to care for patients who cannot participate in traditional psychotherapeutic settings due to common barriers like place of residence, professional obligations, or physical limitations.

Objective: This study aimed to investigate attitudes toward digital psychotherapy, specifically within a digital inpatient-like therapy setting, among psychology and medical students. These students represent the future generation of therapists and possess the educational background necessary to develop innovative ideas to benefit a digital psychotherapeutic setting.

Methods: We conducted qualitative, semistructured interviews with 20 participants (10 psychology students and 10 medical students). The data were analyzed using an inductive, thematic analysis according to the methodology outlined by Braun and Clarke.

Results: The thematic analysis led to a codebook including 4 overarching categories: (1) evolution of digitalization in medical practice, (2) future directions for digital psychotherapy, (3) technical framework, and (4) artificial intelligence–based psychotherapy.

Conclusions: In the context of mental health, digital psychotherapy is accepted as a viable option when conventional face-to-face therapy is not possible. The primary concerns were potential impairments in the therapeutic relationship and interaction. Artificial intelligence was rejected as a standalone therapy but was considered acceptable as a supplementary tool. Technical problems represent a major obstacle for the consistent and reliable implementation of digital psychotherapy. A successful digital psychotherapeutic concept for inpatient and outpatient settings needs to enable a sufficient interpersonal therapeutic relationship situated within a reliable technical framework.

(JMIR Med Educ 2026;12:e82830) doi:10.2196/82830

KEYWORDS

e-mental health; digital mental health; artificial intelligence; AI psychotherapy; mental health care; psychiatric care; acceptance

Introduction

Mental illnesses are among the most common disorders worldwide. Current data indicate that over 12% of the global population lives with a mental health disorder [1]. Regarding the German population, a report from the Robert-Koch Institute highlights the increasing prevalence of depressive and anxiety symptoms in recent years [2]. Consistent with these findings, the German federal chamber of psychotherapists states that the need for psychotherapeutic treatment has increased in the last 2 decades [3]. However, patients endure waiting time for outpatient therapy from weeks to several months [4-6]. Severe cases or those with certain circumstances necessitate inpatient care [7-9]. From an economic perspective, inpatient care is considerably more expensive. A report from the German Psychotherapists Association estimated the 2019 costs of inpatient mental health care at €25 billion (US $29.4 billion), making it the most expensive specialty in the inpatient treatment sector in Germany [10]. In contrast, the costs for outpatient psychotherapy are significantly lower, at around €2.7 billion (US $3.2 billion) annually [10].

XSL•FO
RenderX

In response to the urgent need for psychotherapeutic treatment while simultaneously aiming to reduce costs for the health care system, the concept of inpatient-equivalent treatment (Stationsäquivalente Behandlung [StäB]) emerges as a realistic solution. This treatment model was implemented in 2018 in Germany [11]. StäB aims to enable mental health treatment by a multidisciplinary team that provides care at the patient's home, which is equivalent to an inpatient treatment in a hospital [12]. Initial results from specialties like psychiatry [13], geriatric psychiatry [14], and child and adolescent psychiatry [15] appear promising. These studies report positive effects on quality of life [14], mental health status, involvement of relatives, efficacy, acceptance of treatment [15], and a reduction in inpatient readmissions [13]. This leads to the conclusion that the StäB concept could be a viable solution to all therapeutic fields. Schwarz et al [16] suggest that utilization of the StäB therapeutic model could be facilitated through digital mental health offerings. This combination could be a promising solution to the sector in general.

Rojas et al [17] suggest that internet-based interventions and digital technologies could help in addressing the shortage and unequal distribution of psychotherapy treatments. Digital options for the treatment of mental health issues were increasingly established during the COVID-19 pandemic because they could be used regardless of time or location [18]. In principle, the regulatory and financial framework in Germany is already set by the Digital Supply Act (Digitale-Versorgungs-Gesetz), which legally enables payment for certain digital health therapeutics by German health insurances [19]. Numerous studies have demonstrated positive effects of digital mental health interventions [20-24], even being supportive to inpatient treatment settings [25,26]. The umbrella term "digital interventions" incorporates a variety of therapeutic offerings delivered through digital media, and these may be accompanied by a therapist to varying degrees but also be available for self-help [27].

The related studies confirm that remotely delivered psychotherapy is accessible, manageable, and cost-effective [28,29], but there is less evidence for video-based psychotherapy than for app-based interventions delivered through a smartphone or the web [30]. Nevertheless, there is some support for video-based therapy [31], with some even calling it comparable [28,32]. Some authors go a step further and ask for the implementation of video-based therapy and digital interventions as an additional support [33], for example, apps, web-based offerings, group therapy, and occasional psychiatric consultations [28].

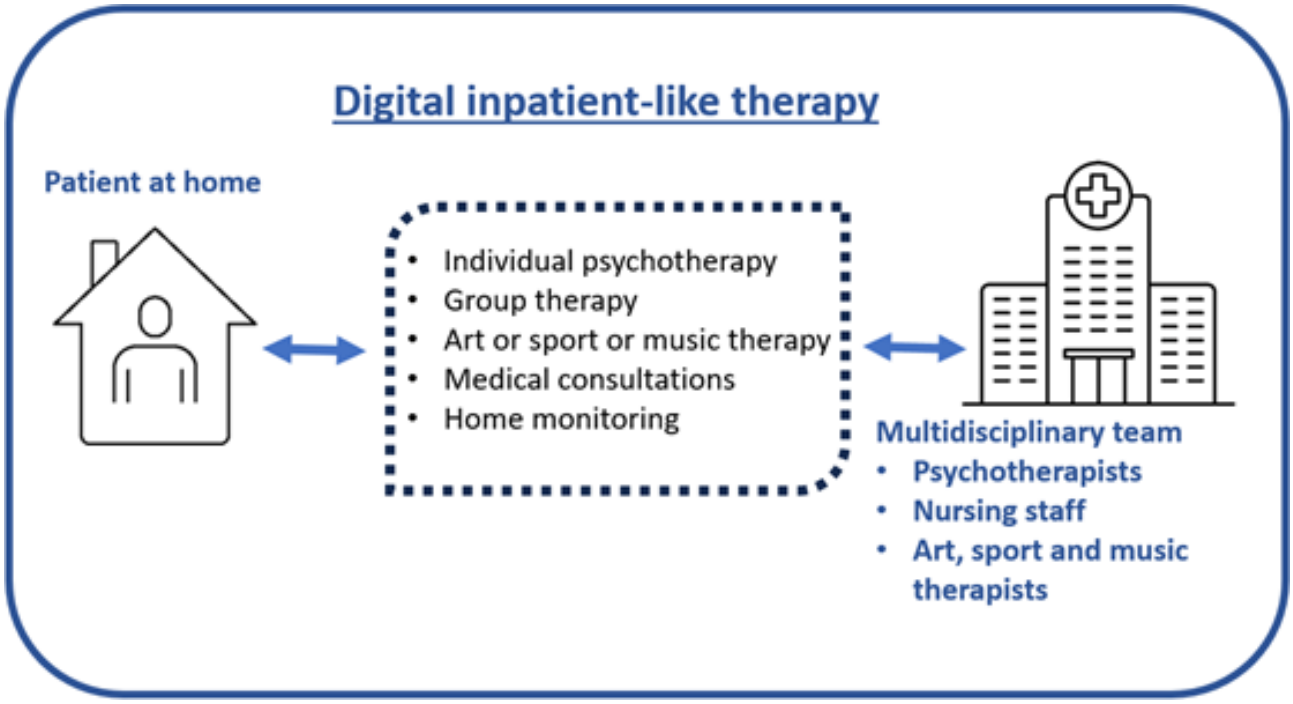
In response to the need for innovative approaches to digital psychotherapy settings, the aim of this study was to investigate the perspectives of medical and psychology students on digital psychotherapy and their opinions on the feasibility of our digital inpatient-like psychotherapy (DIPT) concept. The median age of German students in the last 3 semesters was a few months over 23 years [34], which makes most of them so-called "digital natives," defined as a generation that has used digital media throughout their entire lives [35]. This characteristic could lead to a more intuitive, unprejudiced perspective on the application of digital media and allow them to better see the full potential of digital media compared to older generations. Braun et al [36] recently conducted qualitative interviews with psychology and medical students, investigating their information needs and preferences on digital mental health services. They aimed to consider the perspective of students as future therapists [36].

This particular approach in which the authors focus on students' opinions remains underrepresented in psychotherapy research. Until now, students have often been investigated as the end users of specific interventions such as digital questionnaires [37], online lessons [38], or digital mental health programs [39]. Moreover, there are also studies investigating the effects of digital interventions in general [22,23,40], which include several digital approaches at once. This discrepancy is concerning when considering the unique value of the students' perspective. As young, sometimes self-affected individuals who are well educated in relevant fields, they can provide a holistic perspective on the research question as potential future therapists. Recent findings indicate an open attitude toward the use of e-mental health apps [41].

In this study, semistructured interviews were conducted to capture these unique perspectives regarding possible digital solutions addressing the current unmet needs in psychotherapeutic treatment in Germany. Specifically, the study explored the perspectives of psychology and medical students on a hypothetical DIPT concept. We envisioned a clinic in a virtual environment, where a digital application on an electronic device would provide access to all day-clinic services. All therapeutic modalities should be provided digitally, with representation from all relevant professional disciplines. The preferences for the specific design of such a therapeutic concept were explored in the interviews. Figure 1 provides a schematic overview. Based on these insights, attitudes were examined, along with expected opportunities and barriers. Additionally, we examined attitudes toward the use of artificial intelligence (AI) in psychotherapy, which is emerging as the next advancement in the field [42,43]. Gaining a deeper understanding of these aspects contributes to the development of realistic, patient-centered approaches to digital psychotherapy and also to the investigation of broader digital treatment pathways.

XSL•FO
RenderX

Figure 1. The digital inpatient-like therapy concept.



Methods

Study Design

To address our research objectives, we employed a qualitative interview study design to focus on the lived experiences and perspectives of each respondent [44]. This study method is an effective strategy for developing theories and exploring experiences of social situations and reasons for behavior [45].

Ethical Considerations

Initially, ethical approval was obtained from the Ethics Committee of the Medical Faculty of the University of Duisburg-Essen (23 - 11642-BO). The study was conducted after obtaining informed consent from all participants. Privacy of the participants was safeguarded by pseudonymizing each interview. The audio data were stored on a password-protected device and will be kept for the next 10 years, ensuring adequate confidentiality. Participation was voluntary, and the study participants did not receive any compensation.

Recruitment and Participants

A total of 20 interviews were conducted between March and May 2024, involving 10 students from each of the fields of interest: medicine and psychology. The inclusion criteria were fluency in the German language, adult age (>18 y), and therapeutic-related study subject (human medicine or psychology). Participants were recruited in different ways, including flyers distributed at the university campus, in the clinic setting, on social media, and through direct communication, both verbally and through email, within the study team's environment (professional and personal networks). Every interviewee completed the whole interview.

Semistructured Interview

The majority of qualitative interview formats are semistructured interviews conducted with individuals or groups [46]. A semistructured interview comprises a guide of questions but allows for an individual order of questions, wording, and in-depth inquiry [45]. After reviewing the current relevant literature, we developed an interview guide adapted to our target group. The semistructured questionnaire is provided in Table S3 in Multimedia Appendix 1. Initial pilot testing ensured familiarity with the equipment and the interview process.

The interview was divided into 4 thematic sections. It started with questions about individual sociodemographic data and general and health-related media use. After that, we explored participants' perspectives regarding our DIPT concept. The fourth section addressed experiences and expectations regarding digitalization in medical and psychotherapeutic contexts, including the possibility of AI-guided psychotherapy. Finally, we concluded by offering participants the opportunity to ask questions about the study or provide feedback. The interview guide, which provides a detailed definition of the DIPT concept, is provided in Table S3 in Multimedia Appendix 1.

Data Collection

After a detailed description of the study was provided verbally and in written form, written informed consent was obtained and documented with signature. Interviews were conducted in person at the facilities of the LVR-University Hospital Essen, Clinic for Psychosomatic Medicine and Psychotherapy, Essen, Germany. Audio recordings were obtained with a voice recorder (Sony-ICD-PX470). We selected a sample of 20 interviews, with each subgroup consisting of 10 interviews, because recent studies indicate a saturation of themes with no more aspects emerging at this point [47,48]. Saturation of comprehension requires approximately 20 interviews [48], which was also achieved in this study when subgroups were analyzed separately.

XSL•FO
RenderX

The interview guide was successfully applied. Therefore, all interviews yielded 20 comparable and complete datasets. Thereafter, audio data were transformed into written transcripts using f4x 2024 software [49]. Finally, the data underwent a proofreading process and were carefully collated with the audio recordings to ensure high data quality.

Data Analysis and Quality Control

The average duration of an interview was approximately 29 minutes. The conducted interviews were transformed into pseudonymized transcripts. Data were analyzed semantically using MAXQDA 24 software, a tool that can be used for analyzing qualitative data [50,51]. Three independent researchers analyzed the data: 2 of them coding all transcripts, and the third coder coding 25% of the transcripts. The third researcher also acted as a supervisor and mediated in cases of disagreements, as described by Kuckartz [52].

We conducted a thematic analysis following the instructions established and revised by Braun and Clarke [53,54]. This approach recognizes the active role of the researchers in identifying and interpreting patterns of meaning across the data. The 6-phase framework consists of (1) familiarizing oneself with the data, (2) initial coding, (3) searching for themes, (4) reviewing themes, (5) determining themes, and (6) reporting the results. Themes were generated inductively through an iterative and reflective process, which was data-driven without a preestablished, potentially restrictive codebook [53]. The subjectivity and positionality of the research team (consisting of PhD candidates, research assistants, postdoctoral researchers, and a full-time professor) are considered integral to the analytic process. Thematic analysis is described as a 3-step approach of descriptive coding, interpretive coding, and the final definition of relevant themes [55]. After every step of this analysis, we held a meeting of the 3 coders to ensure the validity of the results [56]. The underlying methodological perspective was a coding reliability approach within reflexive thematic analysis, one of the several subtypes that have evolved from thematic analysis [57]. In this approach, multiple coders work independently and subsequently compare their analysis. The degree of consensus reached enables the reliability of the results [57,58]. Initial coding involved creating short summaries from the raw data (eg, “presence therapy is preferred”). Subsequently, the researchers clustered these codes. For example, summaries related to hardware and software were grouped under the theme “technical infrastructure,” while all mentions of AI were sorted to the theme section “AI-Psychotherapy.” In the following team meeting, these themes were merged and organized into a hierarchical category: theme-subtheme system (axial coding). At this stage, 50% (10/20) of the interviews had been coded. With the developed codebook, the 2 independent researchers then coded the remaining half of the interviews (final coding). In the second meeting, the team discussed and finalized the codebook. Final adjustments were made with the assistance of the supervisor or third coder. To underline sufficient data-reporting quality, we applied the checklist for COREQ (Consolidated Criteria for Reporting Qualitative Research) by Tong et al [59]. The checklist is provided in Checklist 1.

Results

Overview

Participants’ ages ranged from 20 years to 54 years (mean 25.95, SD 6.91 years). Of the 20 participants, 16 (80%) were identified as female, while 4 (20%) were identified as male. Half of the interviewees (10/20, 50%) reported approximately 3 to 6 hours of digital media use per day. Within the context of their study program, 70% (14/20) had some psychotherapeutic experience and 45% (9/20) reported personal experiences in psychotherapy as patients. Most frequently used personal digital devices were smartphones (20/20, 100%) and laptops (14/20, 70%), followed by tablets (12/20, 60%). Digital media was mentioned for various purposes, primarily for participation in social media (17/20, 85%) and entertainment (16/20, 80%). The use of digital health applications ranged from daily (7/20, 35%) to none (6/20, 30%). In this context, digital support was predominantly used for health tracking purposes (12/20, 60%) or in connection with health insurance applications (10/20, 50%). Further information is presented in Table 1.

XSL•FO
RenderX

Table . Sociodemographic data of the participants (N=20).

Characteristics	Values, n (%)
Age (y)	
20 - 25	14 (70)
26 - 30	3 (15)
>30	3 (15)
Study subject	
Psychology	10 (50)
(Human) medicine	10 (50)
Daily media use (h)	
1 - 3	6 (30)
3 - 6	10 (50)
>6	4 (20)
Experiences with psychotherapy	
Voluntary social year	1 (5)
During study program	14 (70)
Personal experience as a patient	9 (45)

Furthermore, we obtained data regarding the individual use of digital devices in general and for health-related purposes. The data are provided in Table S2 in Multimedia Appendix 2. The results indicate a marked usage behavior and high digital literacy.

Codebook

The coding process resulted in the emergence of 4 overarching categories: (1) evolution of digitalization in medical practice, (2) future directions for digital psychotherapy, (3) technical framework, and (4) AI-based psychotherapy. A comprehensive overview of the finalized codebook is provided in Figure 2. For transparency and to support clear thematic interpretation, all participant quotations referenced in the results are included in Table S4 in Multimedia Appendix 3.

XSL•FO
RenderX

Figure 2. Elaborated codebook with categories, themes, and subthemes. The categories are displayed in separate boxes, where solid lines indicate independent sections and dashed lines represent technical aspects that affect the other themes. AI: artificial intelligence.

XSL•FO
RenderX

Category 1: Development of Digitalization in Patient Care

Theme 1.1: Future Perspectives and Potential Usage

This theme included ideas for future digitalization options in somatic medicine. Half of the codes contained a desire to expand existing digital infrastructures, such as providing digital consultation with general practitioners (7/18, 39%) or to make an appointment online (2/18, 11%). In addition, the management and accessibility of digital data emerged as a central concern. Half of the interviewees emphasized the need for the broad implementation of digital medical records, allowing for easier access to patient data across practitioners, involved specialties, or the patients themselves (see quotes 1-3 in Table S4 in Multimedia Appendix 3). Nineteen of 20 (95%) participants generated ideas of digitalization in a medical context, which indicates a general openness to digital health opportunities.

Theme 1.2: Experiences and Challenges

A clear positive example for successful digitalization in the analyzed subgroup was the opportunity for digital somatic and psychotherapy treatments (16/60, 27% codes), the use of apps for health purposes (15/60, 25% codes), and digital prescriptions (7/60, 12% codes). However, participants also reported challenges in the digitalization process in different ways; for example, the poor implementation of concepts (see quote 4 in Table S4 in Multimedia Appendix 3) and an insufficient hospital infrastructure (see quote 5 in Table S4 in Multimedia Appendix 3). As Interviewee 17 summarized, “I wish for more digitalization that functions and does not mean more work” (quote 6 in Table S4 in Multimedia Appendix 3). Another concern was the fear of an overreached digitalization, which “should not somehow get out of hand in this way of course, that in the end you do not see the patient in real life at all” (quote 7 in Table S4 in Multimedia Appendix 3), voiced by 25% (5/20) of the participants. Another aspect was that digital interventions could support certain mental diseases in general (quote 8 in Table S4 in Multimedia Appendix 3) or in specific cases, such as tracking food intake while suffering from an eating disorder (quote 9 in Table S4 in Multimedia Appendix 3).

Category 2: Future Directions for Digital Psychotherapy

Theme 2.1: Requirements for Efficient Implementation

The implementation of a digital psychotherapy setting appeared feasible under certain circumstances. Establishing a therapeutic relationship in person prior to initiating digital therapy was a core requirement of the psychotherapeutic process for 15 of 20 (75%) study participants. All participants expressed a preference for some scheduled in-person meetings, rejecting the idea of a fully digital therapy. An aspect highlighted by 14 participants was the need for other framework conditions to remain unchanged. They emphasized the need of working on the same therapeutic contents (quote 10 in Table S4 in Multimedia Appendix 3), in a calm environment (quote 11 in Table S4 in Multimedia Appendix 3), “That the therapy sessions have the same length” (quote 12 in Table S4 in Multimedia Appendix 3), and “there are still emergency addresses (…). That it is clear, that there is always someone available, if it is necessary” (quote 13 in Table S4 in Multimedia Appendix 3). As interviewee 10 expressed, the setting “should not get to be like a service hotline” (quote 14 in Table S4 in Multimedia Appendix 3). Half of the interviewees highlighted the importance of a prior commitment by patients to engage seriously with the digital format, while 7 out of 20 (35%) participants were concerned about a lack of commitment toward the new concept. Another requirement mentioned was the recruitment of suitable patients for digital therapy. Fifty percent (10/20) assumed that suitability depends on individual patient characteristics. Age was considered a decisive factor by 55% (11/20) of the participants, with older patients probably being less suitable than younger patients. In 16 interviews, it was highlighted that the type of the individual mental disease should be considered in the decision whether a digital therapy is suitable. Thirty percent (6/20) of the participants noted that digital therapy might be particularly beneficial for clients with limited physical mobility.

Theme 2.2: Opportunities and Benefits

A clear advantage seemed to be the provided flexibility of the setting, mentioned by 85% (17/20). Many anticipated that this flexibility would enable easier integration into everyday life (6/20, 30%), particularly for patients with employment obligations (6/20, 30%). Participants viewed the digital concept as adaptable in situations that would otherwise result in therapy cancellations, such as travel-related difficulties (see quote 15 in Table S4 in Multimedia Appendix 3), “In case of illness, or when you have to look after the children” (see quote 16 in Table S4 in Multimedia Appendix 3) or a lack of access to therapists fluent in a client’s native language (see quote 17 in Table S4 in Multimedia Appendix 3). Another opportunity that was mentioned by 9 interviewees is the possibility of receiving emotional support by receiving therapy from home. In total, 14 interviewees expected that digital concepts could help bridge existing treatment gaps by improving access in underserved regions. Furthermore, 65% (13/20) of interviewees highlighted the lower threshold to access psychotherapy as a clear benefit.

Theme 2.3: Critical Aspects

Regarding expected disadvantages, 25% (5/20) of participants alerted that digital therapy might further the possible avoidance of going outside. However, the predominant concern was the expected impairment of interpersonal aspects, mainly concerning the therapeutic alliance between the patient and the therapist, as well as the therapeutic interaction. Seventeen of 20 (85%) respondents believed that digital psychotherapy would impair the therapeutic relationship, making sessions feel less personal (see quote 18 in Table S4 in Multimedia Appendix 3). It seems possible, “that you have concerns or emotions, that only appear, when you are really sitting across from each other” (quote 19 in Table S4 in Multimedia Appendix 3), and it feels better when a therapist is there when you deal with strong upcoming emotions (quote 20 in Table S4 in Multimedia Appendix 3). Participant 15 summarized this by stating “somehow you feel the atmosphere of the person or the chemistry (…). Thus, sometimes I have the feeling, that the screen is more a wall” (quote 21 in Table S4 in Multimedia Appendix 3). Another factor mentioned by 85% (17/20) participants was the reduced insight into nonverbal aspects such as body language, mimic,

XSL•FO
RenderX

and gestures (see quote 22 in Table S4 in Multimedia Appendix 3) and the exclusion of aspects of arriving and leaving a therapy room (quote 23 in Table S4 in Multimedia Appendix 3). In addition, the visible image is limited to the upper body (see quote 24 in Table S4 in Multimedia Appendix 3). Sixteen of 20 (80%) interviewees also emphasized that the relocation to the therapist's office itself provides supporting therapeutic value and would be lost in a digital setting.

Theme 2.4: Transfer of Digital Psychotherapy in an Inpatient-Like Setting

The practical transfer of digital psychotherapy to a day-clinic format was imagined by 40% (8/20) of the participants as a fully digital setting, while 30% (6/20) favored hybrid models with varying degrees of digitalization. Three of 20 (15%) interviewees could imagine both formats. Half of the study participants developed ideas for translating existing therapeutic content in digital form. Moreover, across the interviews, new ideas for the digital psychotherapy concept emerged. Most frequently, participants envisioned a mobile application offering supportive additional therapeutic content (11/20, 55%), particularly to create a successful transfer of therapeutic elements into patients' everyday lives (8/20, 40%), also with regard to the time following discharge from the clinic (see quote 25 in Table S4 in Multimedia Appendix 3). Nevertheless, 40% (8/20) expressed uncertainties or raised open questions regarding the implementation of the digital model.

Category 3: Technical Framework

Theme 3.1: User-Centered Solutions

The subtheme contains aspects that the technical platform for digital therapy should provide. Fourteen of 20 (70%) interviewees imagined therapy through video conferencing. Regarding digital devices, 6 out of 20 (30%) participants underlined the importance of a camera, and 3 out of 20 (15%) participants considered a smartphone to be an inappropriate device for conducting therapy. In 6 interviews, participants described an app as an imaginable platform, and half of them highlighted the importance of a user-friendly design and handling.

Theme 3.2: Essential Technological Competences

In 11 interviews, the need for technological competences to gain access to digital psychotherapy was a topic. Nine emphasized patients' age as an important factor, with older patients probably having more problems with the use of technology, consistent with the concerns about the right client selection. Interviewee 13 described the situation as follows: "I think, you have to be open to the therapy and also have to bring the technical know-how with you. Otherwise, you cannot benefit from it or less, when you are somehow totally overwhelmed with the technology" (quote 26 in Table S4 in Multimedia Appendix 3).

Theme 3.3: Data Security and Privacy Concerns

A secure and confidential transmission of medical data, including psychotherapy sessions, was a recurring concern. Forty percent (8/20) pointed out the importance of a technical platform with provided data security. Two of 20 (10%) participants stated that patients may fear being recorded by the therapist. Data privacy concerns also extended to practitioners; accordingly, "vice versa there has to be a protection for the therapist, that there is no software on the other side, which enables a recording of the sessions" (quote 27 in Table S4 in Multimedia Appendix 3). Privacy concerns also arose from the fact that neither participant of an online meeting can see the other person's entire room (see quotes 28 and 29 in Table S4 in Multimedia Appendix 3). Privacy concerns also appeared in connection with shared patients' surroundings, "When I think about the shared flat and I do not know if that is the safe space for a profound therapy" (quote 30 in Table S4 in Multimedia Appendix 3).

Theme 3.4: Technical Issues

Technical issues emerged in nearly every interview. Seventeen of 20 (85%) interviewees mentioned the importance of a reliable working technological framework. It became clear that various technical disruptions are expected to interfere with the therapy dialogue. The technological side should be "reliably, that the patients can rely on it, that it really takes place every week" (quote 31 in Table S4 in Multimedia Appendix 3). Different concerns showed the vulnerability of a digital therapy session, such as problems with the microphone and internet connection (quote 32 in Table S4 in Multimedia Appendix 3), the quality of the picture of each other (quote 33 in Table S4 in Multimedia Appendix 3), the audio quality (quote 34 in Table S4 in Multimedia Appendix 3), the dependence on a working device (quote 35 in Table S4 in Multimedia Appendix 3), or synchronization problems between audio and video (quote 36 in Table S4 in Multimedia Appendix 3). A more fundamental issue concerned the availability of a working infrastructure. Seven out of 20 (35%) interviewees pointed out that patients might lack the necessary infrastructure (hardware or software); 4 out of 20 (20%) interviewees stated that the clinic might not provide it.

Category 4: AI-Based Psychotherapy

Theme 4.1: Positive Attitudes Toward AI-Based Psychotherapy

It became clear that only a few participants had a positive attitude toward the replacement of a therapist with an AI program. Overall, only 7 codes could be generated over 20 interviews. Expected advantages were a significantly "lower threshold (…), because you can do it 24/7" (quote 37 in Table S4 in Multimedia Appendix 3), with regard to social anxiety, "there is nobody who judges you" (quote 38 in Table S4 in Multimedia Appendix 3), and it seemed to be an option when there is no therapist available locally (quote 39 in Table S4 in Multimedia Appendix 3).

Theme 4.2: Critical Perspectives on AI-Based Psychotherapy

A skeptical, if not outright rejecting, attitude was more prevalent. In 80% (16/20) of the interviews, participants emphasized the essential value of human interaction in therapy. The importance of an interpersonal relationship was clearly highlighted (see quotes 40 and 41 in Table S4 in Multimedia Appendix 3). In conclusion, "I would have a problem with

XSL•FO
RenderX

accepting AI-generated answers (…) for my requests" (quote 42 in Table S4 in Multimedia Appendix 3). Beyond this, general distrust (7/20, 35%) and concerns emerged that the technical side is not developed enough to enable a sufficient therapeutic intervention.

Theme 4.3: Utilization of AI Tools to Support Psychotherapy

In contrast, the use of AI to support making a mental health diagnosis was accepted by 70% (14/20) of the participants. Some even preferred this approach because "you have to be honest, that an AI could maybe reduce something. These two, three sessions diagnostics are never or almost never actually made, especially in the hospital setting, also in the outpatient practice" (quote 43 in Table S4 in Multimedia Appendix 3) and the volume of data could enable an exact diagnosis, the patient could "maybe feel in better hands, because (…) there is a greater data volume to compare" (quote 44 in Table S4 in Multimedia Appendix 3). Furthermore, 75% (15/20) of participants advocated for the use of supportive therapeutic elements like imagination techniques, physical activity, and relaxation instructions or therapeutic diaries.

Discussion

Principal Findings

This study explored medical and psychology students' perspectives on digitalization in medicine and psychotherapy, with a particular focus on our DIPT concept. Participants provided concrete suggestions for supporting digitalization, especially through the expansion of existing structures like teleconsultations, digital appointment scheduling, and digital files. Compared to the existing literature, our findings reflect a more nuanced perspective, containing both enthusiasm and reservation. In recent studies, students' attitudes appear less focused, claiming a general positive [60,61], while others express an overall skeptical attitude [62]. Positive experiences with telemedicine consultations, as reported by our participants, align with previous literature about patients and providers [63,64]. In a similar setting of integrative medicine during the COVID-19 pandemic, Barth et al [65] reported a particularly good working alliance between the therapists and the patients, which raises hope for the psychotherapeutic sector. Nevertheless, the experience of technical problems was reported repeatedly [63,65], consistent with the concerns expressed in our interviews.

The most extensive thematic focus of this study pertained to students' opinions toward digital psychotherapy in general and in the context of a possible inpatient setting. We noticed a clear support for the general idea of digital psychotherapy but at the same time a marked hesitation toward a fully digital psychotherapy setting. This rather conservative result was surprising, given the participants' digital literacy and their frequent use of digital media in general and in health-related contexts. In certain situations—for example, everyday constraints, work-related restrictions, emergency contacts, physical limitations—psychotherapy appears to be accepted; however, fully digital settings seem to be less well-received. Given the assumption that our sample represents a particularly open group with regard to digital treatments, attitudes in the general population may be even more reserved. A closer examination of our study population reveals that our sample consisted of young, predominantly female participants currently enrolled in academic education. This homogeneous study population is consistent with the significantly higher proportion of women in the field of medicine and psychology. The proportion of female students is approximately 70% in medicine [66] and even higher in psychology, at 76.5% [66]. Mackenzie et al [67] reported an association between female gender and a more help-seeking attitude toward mental health treatment, which could consequently lead to a higher response rate of female participants in our study.

Nevertheless, this reported ambivalence toward digital psychotherapy is echoed in current research. In several studies, students show an open attitude toward digital mental health interventions [41,68-70]. Regarding digital psychotherapy, Braun et al [71] reported positive experiences of digital psychotherapy in combination with digital mental health applications. Studies examining students' opinions of digital psychotherapy remain relatively scarce overall. However, a survey from Gbollie et al [72] revealed that only 12.6% of the students preferred digital solutions over a traditional face-to-face therapy. In another survey that accompanied a group therapy setting, students' opinions on therapeutic effects were divided [73]. The specific requirements for digital psychotherapy voiced by our participants are consistent with the literature. Building a therapeutic relationship in person prior to digital sessions seems to be favorable [29,74]. Caution regarding changes beyond therapy location seems warranted. Leukhardt et al [74] found that change of location causes a feeling of uncertainty in the psychotherapeutic process. The idea of specific patient suitability seems valid considering studies on factors that influence engagement in digital mental health interventions: older age [75], severe mental health issues [76,77], or even the lack of knowledge of the existence of this type of care seem to be barriers [78]. Conversely, social affiliation [77] or digital literacy [79] facilitate engagement. The stated chances of digital therapy were in line with prior studies. Digital psychotherapy options are valued for their low threshold, greater flexibility, and reduced travel [80]. Nevertheless, important concerns regarding impaired interpersonal and interactional aspects remain significant. Feasibility studies confirm these concerns: patients experienced digital psychotherapy as less personal [29], with more superficial interaction [74]. Frittgen and Haltaufderheide [81] highlight the changes in sensory and embodied aspects of communication. Given both important critical and beneficial aspects, recommendations for a hybrid therapeutic setting, recommended by several of the participants, emerge as the next logical step. The corresponding data seems promising [82], and some even appear better than traditional therapy [83].

Students also identified the technical framework underlying digital care as a central issue. Studies about digital interventions emphasize the importance of easy usability [69,71], likewise a user-centered design [84]. These features are also mentioned by the participants in the present study. Important problems

XSL•FO
RenderX

highlighted by the current findings are lack of technical competences or issues with the digital hardware and software infrastructure. These arguments cannot be dismissed and have repeatedly been reported as a relevant implementation hurdle for digital interventions or digital psychotherapy [76,77,83,85]. Another aspect in this theme was awareness regarding the handling of sensitive information, which is reflected in previous research works [69,75,79,84].

Considering the latest developments in digital care, AI was also addressed. In our study, it became clear that AI-guided psychotherapy was largely rejected as a standalone psychotherapeutic agent, whereas AI-guided assistance was widely supported. In recent literature, the use of AI in psychotherapeutic settings appears promising in terms of effectiveness [86-88] and may offer low-threshold access [89]. Critical evidence highlights a lack of emotional depth [90], which aligns with our participants' skepticism. Kolding et al [87] state that recent studies about AI need to be interpreted with caution because they often lack quality and have limited clinical relevance. At the European level, the AI Act is a regulation that came into force on August 1, 2024 [91], and is intended to protect recipients of AI health services from physical and psychological harm by preventing the approval of applications with potential adverse effects [91]. The Ethics Committee of the Bundesärztekammer has issued a clear statement on this topic: the responsibility for diagnosis, indication, and therapy lies with the physician [92], and only certified treatments should be used [92]. Similarly, the Kassenärztliche Bundesvereinigung emphasizes that AI should be seen as a support tool, not as a replacement for medical services [93].

Limitations

Despite the insights gained, we still have limitations. The qualitative nature of this interview study can only give results that reflect the outlook and lived experience of the selected study participants. It enables us to understand the investigated themes and develop theories, but the study findings are not generalizable [45]. Nevertheless, we gained valuable knowledge of important factors for mental health treatment and how future mental health professionals want to experience digital psychotherapy, grounded in their life experiences [46]. Although the study group as a whole demonstrated digital literacy and an open-minded attitude toward potential digital psychotherapy treatments, it is noteworthy that only 3 out of 20 (15%) interviewees had received information about digital psychotherapeutic care during their studies, all of whom were psychology students. Given that digital (mental) health treatments are increasingly integrated into routine health care, it is crucial that their application, benefits, and limitations be systematically incorporated into academic curricula at the beginning of professional training. Likewise, Wibberly [94] as well as Kröplin et al [95] argue that the integration of these topics into study curricula is necessary to address the future demands placed on health care providers. In a few cases, study participants were unfamiliar with the concept of a day clinic, and we noticed that this influenced their ability to envision the DIPT concept. Consequently, discussions often shifted toward the aspects of digital outpatient psychotherapy. This context appeared to be more familiar and easier for participants to envision, likely due to their experiences with digital teaching during the COVID-19 pandemic, personal exposure to digital psychotherapy lessons or apps, and contact with digital therapy options addressed during their studies. Furthermore, it is important to keep in mind that researchers, even when following a paradigm of asking open-ended questions without influencing responses, are not free from their own theoretical background and epistemological perspective on the investigated topic [53]. The researchers' different professional backgrounds allowed for a balanced perspective on both medical and therapeutic topics. First coder and interviewer, a medical doctor (26 years old), was trained in an open-ended interview approach. Students were not directly dependent, neither professionally through work as student assistants, nor in the context of academic performance assessments. Second and third coders were student assistants (study subject was psychology), well trained in qualitative research. Overall, we assume that a pressure-free interview atmosphere was maintained. A logical next step would be a feasibility study of a real-life implementation of DIPT. With the insights gained here, a more successful concept can be developed, expected hurdles can be evaded, and identified opportunities can be realized.

Conclusion

The results of this interview study led to the conclusion that students are open to more digitalization in health care. Concerning digital psychotherapy, support is more conditional. Hybrid settings were favored that incorporate establishing a therapeutic alliance in person and/or continuing it through in-person meetings during the therapeutic process. A fully AI-guided psychotherapy was rejected, whereas therapeutic supportive tools were welcomed. Technical problems appear to be a relevant hurdle for committing to a digital setting. The obtained aspects can help in implementing adequate digital therapeutic care in a digital inpatient-like or outpatient setting.

Acknowledgments

The authors thank all the participants of the present study. They thank the Open Access Publication Fund of the University of Duisburg-Essen for supporting the open access publication. The authors also thank Anthea Hammond for language support. DeepL and ChatGPT were only used to improve the language of the text and for translations. No new content was created.

Funding

The authors declared no financial support was received for this work.

XSL•FO
RenderX

Data Availability

The analyzed data are available from the corresponding author on reasonable request.

Authors' Contributions

Conceptualization: RR, LAG, PJW, AB, AR
Data curation: RR
Investigation: RR
Methodology: RR, TL, SSgM, LAG, PJW
Project administration: AB, AR
Supervision: MT, AB, AR
Writing – original draft: RR
Writing – review & editing: TL, AD, SSgM, LAG, PJW, MT, AB, AR

Conflicts of Interest

None declared.

Multimedia Appendix 1
Semistructured interview guide: translated for publication purposes only (originally in German language).
[DOCX File, 17 KB - mededu_v12i1e82830_app1.docx]

Multimedia Appendix 2
Participants' use of digital media.
[DOCX File, 13 KB - mededu_v12i1e82830_app2.docx]

Multimedia Appendix 3
Quotes for themes and subthemes: translated for publication (originally in German language).
[DOCX File, 15 KB - mededu_v12i1e82830_app3.docx]

Checklist 1
COREQ checklist.
[PDF File, 267 KB - mededu_v12i1e82830_app4.pdf]

References

1. Mental disorders. World Health Organization. URL: https://www.who.int/news-room/fact-sheets/detail/mental-disorders [accessed 2025-06-19]
2. NCD-Surveillance-Bericht: Ergebnisse zur Entwicklung Verschiedener Gesundheitsindikatoren in Der Erwachsenen Bevölkerung Bei Hochfrequenter Beobachtung; Stand 2024. : Robert-Koch-Institut; 2024 URL: https://www.rki.de/DE/Themen/Nichtuebertragbare-Krankheiten/Studien-und-Surveillance/Studien/MHS/NCD-Surveillance-Bericht.pdf?__blob=publicationFile&v=2 [accessed 2025-06-19]
3. Krankenkassen Blockieren Sachgerechte Reform Der Bedarfsplanung BPtK: Ländliche Regionen Weiterhin Massiv Benachteiligt. : Bundespsychotherapeutenkammer (BPtK); 2019 URL: https://api.bptk.de/uploads/20190516_pm_bptk_bedarfsplanung_6d4e1d3d6c.pdf [accessed 2025-06-19]
4. Dokumentation: Wartezeiten Auf Eine Psychotherapie – Studien Und Umfragen [Report in German]. : Wissenschaftliche Dienste des Deutschen Bundestages; 2022 URL: https://www.bundestag.de/resource/blob/916578/53724d526490deea69f736b1fda83e76/WD-9-059-22-pdf-data.pdf [accessed 2025-06-19]
5. Reaktion auf Regierungsangaben: Psychotherapeuten beklagen „schwer erträglichen" Versorgungszustand [Article in German]. Ärzte Zeitung. 2023. URL: https://www.aerztezeitung.de/Politik/Psychotherapeuten-beklagen-schwer-ertraeglichen-Versorgungszustand-435767.html [accessed 2025-06-19]
6. Fokus: Ambulante Psychotherapie. GKV-Spitzenverband. 2025. URL: https://www.gkv-spitzenverband.de/gkv_spitzenverband/presse/fokus/psychotherapie.jsp [accessed 2025-06-19]
7. Schuld A. Psychosomatik: Argumente für die stationäre Psychotherapie [Article in German]. MMW - Fortschritte der Medizin 2021 Apr;163(6):50-52. [doi: 10.1007/s15006-021-9668-2]
8. Doering S, Herpertz S, Hofmann T, et al. What kind of patients receive inpatient and day-hospital treatment in departments of psychosomatic medicine and psychotherapy in Germany? Psychother Psychosom 2023;92(1):49-54. [doi: 10.1159/000527881] [Medline: 36516807]
9. Loch AA. Discharged from a mental health admission ward: is it safe to go home? A review on the negative outcomes of psychiatric hospitalization. Psychol Res Behav Manag 2014;7:137-145. [doi: 10.2147/PRBM.S35061] [Medline: 24812527]

XSL•FO
RenderX

10. Report Psychotherapie 2021. 2. Auflage [Report in German]. : DPtV Deutsche Psychotherapeuten Vereinigung e.V; 2021 URL: https://www.dptv.de/fileadmin/Redaktion/Bilder_und_Dokumente/Wissensdatenbank_oeffentlich/Report_Psychotherapie/DPtV_Report_Psychotherapie_2021.pdf [accessed 2025-06-20]
11. Stationsäquivalente psychiatrische Behandlung [Article in German]. Deutsche Krankenhaus Gesellschaft. URL: https://www.dkgev.de/themen/versorgung-struktur/psychiatrie-psychosomatik/stationsaequivalente-psychiatrische-behandlung/ [accessed 2025-06-20]
12. Bramesfeld A. Die Versorgung von Menschen mit psychischen Erkrankungen in Deutschland aus Perspektive des Gesundheits- und Sozialsystems: Aktuelle Entwicklungsbedarfe [Article in German]. Bundesgesundheitsbl 2023 Apr;66(4):363-370. [doi: 10.1007/s00103-023-03671-x]
13. Nikolaidis K, Weinmann S, Döring S, et al. Stationsäquivalente Behandlung (StäB) im Vergleich mit vollstationärer Behandlung: 12-Monats-Follow-up Ergebnisse einer mittels Propensity-Score gematchten retrospektiven Kohortenstudie [Article in German]. Psychiatr Prax 2024 Mar;51(2):92-98. [doi: 10.1055/a-2177-6113]
14. Spannhorst S, Weller S, Thomas C. Stationsäquivalente Behandlung: Eine neue Versorgungsform auch in der Gerontopsychiatrie [Article in German]. Z Gerontol Geriat 2020;53(8):713-720. [doi: 10.1007/s00391-020-01823-1]
15. Boege I, Schepker R, Fegert JM. Vom Hometreatment zur stationsäquivalenten Behandlung (StäB): Ein systematischer Review aufsuchender Behandlung in Deutschland [Article in German]. Z Kinder Jugendpsychiatr Psychother 2020;48(5):393-406. [doi: 10.1024/1422-4917/a000710]
16. Schwarz J, Hemmerling J, Kabisch N, et al. Equal access to outreach mental health care? Exploring how the place of residence influences the use of intensive home treatment in a rural catchment area in Germany. BMC Psychiatry 2022 Dec 26;22(1):826. [doi: 10.1186/s12888-022-04477-y] [Medline: 36572855]
17. Rojas G, Martínez V, Martínez P, Franco P, Jiménez-Molina Á. Improving mental health care in developing countries through digital technologies: a mini narrative review of the Chilean case. Front Public Health 2019;7:391. [doi: 10.3389/fpubh.2019.00391] [Medline: 31921754]
18. Witteveen AB, Young S, Cuijpers P, et al. Remote mental health care interventions during the COVID-19 pandemic: an umbrella review. Behav Res Ther 2022 Dec;159:104226. [doi: 10.1016/j.brat.2022.104226] [Medline: 36410111]
19. Weitzel EC, Quittschalle J, Welzel FD, Löbner M, Hauth I, Riedel-Heller SG. E-mental-health und digitale Gesundheitsanwendungen in Deutschland [Article in German]. Nervenarzt 2021 Nov;92(11):1121-1129. [doi: 10.1007/s00115-021-01196-9] [Medline: 34608535]
20. Sifat MS, Tasnim N, Stoebenau K, Green KM. A qualitative exploration of university student perspectives on mindfulness-based stress reduction exercises via smartphone app in Bangladesh. Int J Qual Stud Health Well-being 2022 Dec;17(1):2113015. [doi: 10.1080/17482631.2022.2113015] [Medline: 35980122]
21. Sasseville M, LeBlanc A, Boucher M, et al. Digital health interventions for the management of mental health in people with chronic diseases: a rapid review. BMJ Open 2021 Apr 5;11(4):e044437. [doi: 10.1136/bmjopen-2020-044437] [Medline: 33820786]
22. Lattie EG, Adkins EC, Winquist N, Stiles-Shields C, Wafford QE, Graham AK. Digital mental health interventions for depression, anxiety, and enhancement of psychological well-being among college students: systematic review. J Med Internet Res 2019 Jul 22;21(7):e12869. [doi: 10.2196/12869] [Medline: 31333198]
23. Ferrari M, Allan S, Arnold C, et al. Digital interventions for psychological well-being in university students: systematic review and meta-analysis. J Med Internet Res 2022 Sep 28;24(9):e39686. [doi: 10.2196/39686] [Medline: 36169988]
24. Petrovic M, Gaggioli A. Digital mental health tools for caregivers of older adults-a scoping review. Front Public Health 2020;8:128. [doi: 10.3389/fpubh.2020.00128] [Medline: 32411643]
25. Zwerenz R, Becker J, Knickenberg RJ, Siepmann M, Hagen K, Beutel ME. Online self-help as an add-on to inpatient psychotherapy: efficacy of a new blended treatment approach. Psychother Psychosom 2017;86(6):341-350. [doi: 10.1159/000481177] [Medline: 29131090]
26. Diel A, Schröter IC, Frewer AL, et al. A systematic review and meta analysis on digital mental health interventions in inpatient settings. NPJ Digit Med 2024 Sep 17;7(1):253. [doi: 10.1038/s41746-024-01252-z] [Medline: 39289463]
27. Gagnon MP, Sasseville M, Leblanc A. Classification of digital mental health interventions: a rapid review and framework proposal. Stud Health Technol Inform 2022 May 25;294:629-633. [doi: 10.3233/SHTI220545] [Medline: 35612165]
28. Fletcher TL, Hogan JB, Keegan F, et al. Recent advances in delivering mental health treatment via video to home. Curr Psychiatry Rep 2018 Jul 21;20(8):56. [doi: 10.1007/s11920-018-0922-y] [Medline: 30032337]
29. Moeller AM, Hansen JP, Andersen PT. Patients' experiences of home-based psychotherapy via videoconference: a qualitative study. Arch Psychiatr Nurs 2022 Aug;39:91-96. [doi: 10.1016/j.apnu.2022.03.004] [Medline: 35688550]
30. Lamb T, Pachana NA, Dissanayaka N. Update of recent literature on remotely delivered psychotherapy interventions for anxiety and depression. Telemed J E Health 2019 Aug;25(8):671-677. [doi: 10.1089/tmj.2018.0079] [Medline: 30300082]
31. Markowitz JC, Milrod B, Heckman TG, et al. Psychotherapy at a distance. Am J Psychiatry 2021 Mar 1;178(3):240-246. [doi: 10.1176/appi.ajp.2020.20050557] [Medline: 32972202]
32. Weightman M. Digital psychotherapy as an effective and timely treatment option for depression and anxiety disorders: implications for rural and remote practice. J Int Med Res 2020 Jun;48(6):300060520928686. [doi: 10.1177/0300060520928686] [Medline: 32527170]

XSL•FO
RenderX

33. Smith KA, Blease C, Faurholt-Jepsen M, et al. Digital mental health: challenges and next steps. BMJ Ment Health 2023 Feb;26(1):e300670. [doi: 10.1136/bmjment-2023-300670] [Medline: 37197797]
34. Hochschulen: Studierende nach Bundesländern [Article in German]. Statistisches Bundesamt (Destatis). 2024. URL: https://www.destatis.de/DE/Themen/Gesellschaft-Umwelt/Bildung-Forschung-Kultur/Hochschulen/Tabellen/studierende-insgesamt-bundeslaender.html [accessed 2024-09-23]
35. Digital native. Cambridge Dictionary. 2024. URL: https://dictionary.cambridge.org/de/worterbuch/englisch/digital-native [accessed 2024-09-23]
36. Braun P, Schwientek AK, Angerer P, et al. Investigating information needs and preferences regarding digital mental health services among medical and psychology students in Germany: a qualitative study. Digit Health 2023;9:20552076231173568. [doi: 10.1177/20552076231173568] [Medline: 37256006]
37. Heesacker M, Perez C, Quinn MS, Benton S. Computer-assisted psychological assessment and psychotherapy for collegians. J Clin Psychol 2020 Jun;76(6):952-972. [doi: 10.1002/jclp.22854] [Medline: 31476251]
38. Keis O, Grab C, Schneider A, Öchsner W. Online or face-to-face instruction? A qualitative study on the electrocardiogram course at the University of Ulm to examine why students choose a particular format. BMC Med Educ 2017 Nov 9;17(1):194. [doi: 10.1186/s12909-017-1053-6] [Medline: 29121902]
39. Harrer M, Adam SH, Baumeister H, et al. Internet interventions for mental health in university students: a systematic review and meta-analysis. Int J Methods Psychiatr Res 2019 Jun;28(2):e1759. [doi: 10.1002/mpr.1759] [Medline: 30585363]
40. Harith S, Backhaus I, Mohbin N, Ngo HT, Khoo S. Effectiveness of digital mental health interventions for university students: an umbrella review. PeerJ 2022;10:e13111. [doi: 10.7717/peerj.13111] [Medline: 35382010]
41. Grüneberg C, Bäuerle A, Karunakaran S, et al. Medical students' acceptance of tailored e-mental health apps to foster their mental health: cross-sectional study. JMIR Med Educ 2025 Jan 24;11:e58183. [doi: 10.2196/58183] [Medline: 39865568]
42. Bhatt S. Digital mental health: role of artificial intelligence in psychotherapy. Ann Neurosci 2025 Apr;32(2):117-127. [doi: 10.1177/09727531231221612] [Medline: 39544658]
43. Dehbozorgi R, Zangeneh S, Khooshab E, et al. The application of artificial intelligence in the field of mental health: a systematic review. BMC Psychiatry 2025 Feb 14;25(1):132. [doi: 10.1186/s12888-025-06483-2] [Medline: 39953464]
44. Kuper A, Reeves S, Levinson W. An introduction to reading and appraising qualitative research. BMJ 2008 Aug 7;337:a288. [doi: 10.1136/bmj.a288] [Medline: 18687727]
45. Bullock A. Conduct one-to-one qualitative interviews for research. Educ Prim Care 2016 Jul;27(4):330-332. [doi: 10.1080/14739879.2016.1176874] [Medline: 27113858]
46. Dicicco-Bloom B, Crabtree BF. The qualitative research interview. Med Educ 2006 Apr;40(4):314-321. [doi: 10.1111/j.1365-2929.2006.02418.x] [Medline: 16573666]
47. Weller SC, Vickers B, Bernard HR, et al. Open-ended interview questions and saturation. PLoS One 2018;13(6):e0198606. [doi: 10.1371/journal.pone.0198606] [Medline: 29924873]
48. Hennink MM, Kaiser BN, Marconi VC. Code saturation versus meaning saturation: how many interviews are enough? Qual Health Res 2017 Mar;27(4):591-608. [doi: 10.1177/1049732316665344] [Medline: 27670770]
49. f4 Login – audiotranskription [Website in German]. f4x automatische Transkription. 2025. URL: https://f4x.audiotranskription.de/anmelden [accessed 2025-03-16]
50. Rädiker S, Kuckartz U. Focused Analysis of Qualitative Interviews with MAXQDA: Step by Step: MAXQDA Press; 2020.
51. MAXQDA [Article in German]. 2025. URL: https://www.maxqda.com/de/ [accessed 2025-06-20]
52. Kuckartz U. Qualitative Inhaltsanalyse Methoden, Praxis, Computerunterstützung [Book in German], 4th edition: Beltz Juventa; 2018.
53. Braun V, Clarke V. Using thematic analysis in psychology. Qual Res Psychol 2006 Jan;3(2):77-101. [doi: 10.1191/1478088706qp063oa]
54. Braun V, Clarke V. Reflecting on reflexive thematic analysis. Qual Res Sport Exerc Health 2019 Aug 8;11(4):589-597. [doi: 10.1080/2159676X.2019.1628806]
55. Bennett D, Barrett A, Helmich E. How to…analyse qualitative data in different ways. Clin Teach 2019 Feb;16(1):7-12. [doi: 10.1111/tct.12973] [Medline: 30474356]
56. Ryan GW, Bernard HR. Techniques to identify themes. Field Methods 2003 Feb;15(1):85-109. [doi: 10.1177/1525822X02239569]
57. Braun V, Clarke V. Can I use TA? Should I use TA? Should I not use TA? Comparing reflexive thematic analysis and other pattern-based qualitative analytic approaches. Couns Psychother Res 2021 Mar;21(1):37-47. [doi: 10.1002/capr.12360]
58. Braun V, Clarke V, Hayfield N, Terry G. Thematic analysis. In: Liamputtong P, editor. Handbook of Research Methods in Health Social Sciences: Springer; 2019:843-860. [doi: 10.1007/978-981-10-5251-4_103]
59. Tong A, Sainsbury P, Craig J. Consolidated criteria for reporting qualitative research (COREQ): a 32-item checklist for interviews and focus groups. Int J Qual Health Care 2007 Dec;19(6):349-357. [doi: 10.1093/intqhc/mzm042] [Medline: 17872937]
60. Veikkolainen P, Tuovinen T, Jarva E, et al. eHealth competence building for future doctors and nurses - attitudes and capabilities. Int J Med Inform 2023 Jan;169:104912. [doi: 10.1016/j.ijmedinf.2022.104912] [Medline: 36356432]

XSL•FO
RenderX

61. Thapa S, Nielsen JB, Aldahmash AM, Qadri FR, Leppin A. Willingness to use digital health tools in patient care among health care professionals and students at a university hospital in Saudi Arabia: quantitative cross-sectional survey. JMIR Med Educ 2021 Feb 19;7(1):e18590. [doi: 10.2196/18590] [Medline: 33605896]
62. Baumgartner M, Sauer C, Blagec K, Dorffner G. Digital health understanding and preparedness of medical students: a cross-sectional study. Med Educ Online 2022 Dec;27(1):2114851. [doi: 10.1080/10872981.2022.2114851] [Medline: 36036219]
63. Barkai G, Gadot M, Amir H, Menashe M, Shvimer-Rothschild L, Zimlichman E. Patient and clinician experience with a rapidly implemented large-scale video consultation program during COVID-19. Int J Qual Health Care 2021 Feb 20;33(1):mzaa165. [doi: 10.1093/intqhc/mzaa165] [Medline: 33313891]
64. Garcia-Huidobro D, Rivera S, Valderrama Chang S, Bravo P, Capurro D. System-wide accelerated implementation of telemedicine in response to COVID-19: mixed methods evaluation. J Med Internet Res 2020 Oct 6;22(10):e22146. [doi: 10.2196/22146] [Medline: 32903195]
65. Barth J, Canella C, Oehler M, Witt CM. Digital consultations during COVID-19: a multiperspective mixed-methods study in an integrative medicine setting in Switzerland. J Altern Complement Med 2021 Jul;27(7):569-578. [doi: 10.1089/acm.2020.0539] [Medline: 33960805]
66. DatenCHECK 3/25: Was studieren Frauen? Was studieren Männer? – Studierende und Studienanfänger*innen nach Geschlecht [Article in German]. CHE Hochschuldaten. URL: https://hochschuldaten.che.de/was-studieren-frauen-was-studieren-maenner/ [accessed 2026-03-07]
67. Mackenzie CS, Gekoski WL, Knox VJ. Age, gender, and the underutilization of mental health services: the influence of help-seeking attitudes. Aging Ment Health 2006 Nov;10(6):574-582. [doi: 10.1080/13607860600641200] [Medline: 17050086]
68. Topooco N, Fowler LA, Fitzsimmons-Craft EE, et al. Digital interventions to address mental health needs in colleges: perspectives of student stakeholders. Internet Interv 2022;28:100528. [doi: 10.1016/j.invent.2022.100528] [Medline: 35378846]
69. Dederichs M, Weber J, Pischke CR, Angerer P, Apolinário-Hagen J. Exploring medical students' views on digital mental health interventions: a qualitative study. Internet Interv 2021;25:100398. [doi: 10.1016/j.invent.2021.100398] [Medline: 34026567]
70. Van Der Poll R, Coetzee B, Bantjes J. Willing and unwilling digital cyborg assemblages: university students talk about mental health apps. Digit Health 2023;9:20552076231210658. [doi: 10.1177/20552076231210658] [Medline: 37915793]
71. Braun P, Atik E, Guthardt L, Apolinário-Hagen J, Schückes M. Barriers to and facilitators of a blended cognitive behavioral therapy program for depression and anxiety based on experiences of university students: qualitative interview study. JMIR Form Res 2023 Apr 12;7:e45970. [doi: 10.2196/45970] [Medline: 37043272]
72. Gbollie EF, Bantjes J, Jarvis L, et al. Intention to use digital mental health solutions: a cross-sectional survey of university students attitudes and perceptions toward online therapy, mental health apps, and chatbots. Digit Health 2023;9:20552076231216559. [doi: 10.1177/20552076231216559] [Medline: 38047161]
73. Hunt X, Jivan DC, Naslund JA, Breet E, Bantjes J. South African university students' experiences of online group cognitive behavioural therapy: implications for delivering digital mental health interventions to young people. Glob Ment Health (Camb) 2023;10:e45. [doi: 10.1017/gmh.2023.39] [Medline: 37854416]
74. Leukhardt A, Heider M, Reboly K, Franzen G, Eichenberg C. Videobasierte Behandlungen in der psychodynamischen Psychotherapie in Zeiten der COVID-19-Pandemie: Interviewstudie mit Psychotherapeut*innen und Patient*innen [Article in German]. Psychotherapeut (Berl) 2021;66(5):398-405. [doi: 10.1007/s00278-021-00532-3] [Medline: 34462623]
75. Mayer G, Gronewold N, Alvarez S, Bruns B, Hilbel T, Schultz JH. Acceptance and expectations of medical experts, students, and patients toward electronic mental health apps: cross-sectional quantitative and qualitative survey study. JMIR Ment Health 2019 Nov 25;6(11):e14018. [doi: 10.2196/14018] [Medline: 31763990]
76. Van Assche E, Bonroy B, Mertens M, et al. E-mental health implementation in inpatient care: exploring its potential and future challenges. Front Digit Health 2022;4(27):1027864. [doi: 10.3389/fdgth.2022.1027864] [Medline: 36588747]
77. Borghouts J, Eikey E, Mark G, et al. Barriers to and facilitators of user engagement with digital mental health interventions: systematic review. J Med Internet Res 2021 Mar 24;23(3):e24387. [doi: 10.2196/24387] [Medline: 33759801]
78. Weitzel EC, Schwenke M, Schomerus G, et al. E-mental health in Germany - what is the current use and what are experiences of different types of health care providers for patients with mental illnesses? Arch Public Health 2023 Jul 17;81(1):133. [doi: 10.1186/s13690-023-01150-y] [Medline: 37461064]
79. Özer Ö, Köksal B, Altinok A. Understanding university students' attitudes and preferences for internet-based mental health interventions. Internet Interv 2024;35:100722. [doi: 10.1016/j.invent.2024.100722] [Medline: 38356613]
80. Meier JV, Noel JA, Kaspar K. Understanding psychology students' perspective on video psychotherapy and their intention to offer it after graduation: a mixed-methods study. Front Psychol 2023;14:1234167. [doi: 10.3389/fpsyg.2023.1234167] [Medline: 37928577]
81. Frittgen EM, Haltaufderheide J. 'Can you hear me?': communication, relationship and ethics in video-based telepsychiatric consultations. J Med Ethics 2022 Jan;48(1):22-30. [doi: 10.1136/medethics-2021-107434] [Medline: 34610979]

XSL•FO
RenderX

82. Ehrt-Schäfer Y, Rusmir M, Vetter J, Seifritz E, Müller M, Kleim B. Feasibility, adherence, and effectiveness of blended psychotherapy for severe mental illnesses: scoping review. JMIR Ment Health 2023 Dec 26;10:e43882. [doi: 10.2196/43882] [Medline: 38147373]
83. Omylinska-Thurston J, Aithal S, Liverpool S, et al. Digital psychotherapies for adults experiencing depressive symptoms: systematic review and meta-analysis. JMIR Ment Health 2024 Sep 30;11:e55500. [doi: 10.2196/55500] [Medline: 39348177]
84. Oti O, Pitt I. Online mental health interventions designed for students in higher education: a user-centered perspective. Internet Interv 2021;26:100468. [doi: 10.1016/j.invent.2021.100468] [Medline: 34703772]
85. Ramos G, Hernandez-Ramos R, Taylor M, Schueller SM. State of the science: using digital mental health interventions to extend the impact of psychological services. Behav Ther 2024 Nov;55(6):1364-1379. [doi: 10.1016/j.beth.2024.04.004] [Medline: 39443071]
86. He Y, Yang L, Zhu X, et al. Mental health chatbot for young adults with depressive symptoms during the COVID-19 pandemic: single-blind, three-arm randomized controlled trial. J Med Internet Res 2022 Nov 21;24(11):e40719. [doi: 10.2196/40719] [Medline: 36355633]
87. Kolding S, Lundin RM, Hansen L, Østergaard SD. Use of generative artificial intelligence (AI) in psychiatry and mental health care: a systematic review. Acta Neuropsychiatr 2024 Nov 11;37:e37. [doi: 10.1017/neu.2024.50] [Medline: 39523628]
88. Zhong W, Luo J, Zhang H. The therapeutic effectiveness of artificial intelligence-based chatbots in alleviation of depressive and anxiety symptoms in short-course treatments: a systematic review and meta-analysis. J Affect Disord 2024 Jul 1;356:459-469. [doi: 10.1016/j.jad.2024.04.057] [Medline: 38631422]
89. Jurblum M, Selzer R. Potential promises and perils of artificial intelligence in psychotherapy - the AI Psychotherapist (APT). Australas Psychiatry 2025 Feb;33(1):103-105. [doi: 10.1177/10398562241286312] [Medline: 39344214]
90. Spytska L. The use of artificial intelligence in psychotherapy: development of intelligent therapeutic systems. BMC Psychol 2025 Feb 28;13(1):175. [doi: 10.1186/s40359-025-02491-9] [Medline: 40022267]
91. Künstliche Intelligenz: AI-Act: Die wichtigsten Fragen zur KI-Verordnung [Article in German]. IHK Region Stuttgart. 2026. URL: https://www.ihk.de/stuttgart/fuer-unternehmen/innovation/innovation-technologie/ce-kennzeichnung/faq-ai-act-ki-verordnung-6086082 [accessed 2026-02-22]
92. Zentrale Ethikkommission. Stellungnahme der Zentralen Kommission zur Wahrung ethischer Grundsätze in der Medizin und ihren Grenzgebieten (Zentrale Ethikkommission) bei der Bundesärztekammer “Entscheidungsunterstützung ärztlicher Tätigkeit durch Künstliche Intelligenz” [Article in German]. Dtsch Arztebl 2021;118(33-34):A1654. [doi: 10.3238/arztebl.zeko_sn_cdss_2021]
93. Themenseite: Künstliche Intelligenz [Article in German]. Kassenärztliche Bundesvereinigung (KBV). 2026. URL: https://www.kbv.de/praxis/digitalisierung/anwendungen/kuenstliche-intelligenz [accessed 2026-02-22]
94. Wibberly KH. Preparing mental health providers for the future: the case for moving beyond the elective telehealth course to integrating telehealth training throughout the curriculum. Front Psychol 2024;14:1301569. [doi: 10.3389/fpsyg.2023.1301569] [Medline: 38235285]
95. Kröplin J, Maier L, Lenz JH, Romeike B. Impact of a “digital health” curriculum on students’ perception about competence and relevance of digital health topics for future professional challenges: prospective pilot study. JMIR Form Res 2025 May 12;9:e58940. [doi: 10.2196/58940] [Medline: 40354637]

Abbreviations

AI: artificial intelligence
COREQ: Consolidated Criteria for Reporting Qualitative Research
DIPT: digital inpatient-like psychotherapy
StäB: Stationsäquivalente Behandlung

Edited by T Gladman; submitted 22.Aug.2025; peer-reviewed by A Hudon, A AL-Asadi, A Quansah; revised version received 17.Mar.2026; accepted 19.Apr.2026; published 01.Jun.2026.

Please cite as:
Robitzsch R, Lalgi T, Diel A, Schulz genannt Menningmann S, Gresser LA, Wollenberg PJ, Teufel M, Bäuerle A, Robitzsch A Students’ Perspectives on Digital Psychotherapy—Possible Solutions for Digital Inpatient-Like Care Concepts: Qualitative Interview Study
JMIR Med Educ 2026;12:e82830
URL: https://mededu.jmir.org/2026/1/e82830
doi:10.2196/82830

XSL•FO
RenderX

© Rebekka Robitzsch, Tania Lalgi, Alexander Diel, Sophie Schulz genannt Menningmann, Lucy Ann Gresser, Patrick Jonas Wollenberg, Martin Teufel, Alexander Bäuerle, Anita Robitzsch. Originally published in JMIR Medical Education (https://mededu.jmir.org), 1.Jun.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Education, is properly cited. The complete bibliographic information, a link to the original publication on https://mededu.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

AI- vs Human-Based Assessment of Medical Interview Transcripts in a Generative AI–Simulated Patient System: Cross-Sectional Validation Study

Hiromizu Takahashi[1], MD, PhD; Kiyoshi Shikino[2], MD, MHPE, PhD; Takeshi Kondo[3,4], MD, MHPE, PhD; Yuji Yamada[5], MD, MPH, PhD; Yoshitaka Tomoda[6], MD, PhD, FACP; Minoru Kishi[7], MD, PhD; Yuki Aiyama[8], MD; Sho Nagai[9], RN, PHN, MSN; Akiko Enomoto[9], RN, PHN, MSN; Yoshinori Tokushima[10], MD, PhD; Takahiro Shinohara[11], MD; Fumiaki Sano[1], MD; Takeshi Matsuura[12], MD; Rikiya Watanabe[13], MD; Toshio Naito[1], MD, PhD

[1]Department of General Medicine, Faculty of Medicine, Juntendo University, 3-1-3 Hongo, Bunkyo Tokyo, 1130033 Japan, Tokyo, Japan

[10]Department of General Medicine, Saga University Hospital, Saga, Japan

[11]Department of General Medicine, Graduate School of Medical and Dental Sciences, Institute of Science Tokyo, Tokyo, Japan

[12]Department of General Medicine, Bibai City Hospital, Hokkaido, Japan

[13]Department of General Internal Medicine, Kita-Harima Medical Center, Hyogo, Japan

[2]Department of Community-Oriented Medical Education, Graduate School of Medicine, Chiba University, Chiba, Japan

[3]Center for Postgraduate Clinical Training and Career Development, Nagoya University Hospital, Nagoya, Japan

[4]The School of Health Professions Education, Maastricht University, Maastricht, The Netherlands

[5]Brookdale Department of Geriatrics and Palliative Medicine, Icahn School of Medicine at Mount Sinai, New York, NY, United States

[6]Department of General Internal Medicine, Itabashi Chuo Medical Center, Tokyo, Japan

[7]Department of Internal Medicine, Nishiwaki Municipal Hospital, Hyogo, Japan

[8]Anesthesiology and Critical Care Medicine, Tenri Hospital, Nara, Japan

[9]Department of Nursing, School of Nursing, University of Human Environments, Aichi, Japan

Corresponding Author:
Hiromizu Takahashi, MD, PhD
Department of General Medicine, Faculty of Medicine, Juntendo University, 3-1-3 Hongo, Bunkyo Tokyo, 1130033 Japan, Tokyo, Japan

Abstract

Background: Generative artificial intelligence (AI) is increasingly used in medical education, including AI-based virtual patients to improve interview skills. However, how much AI-based assessment (ABA) differs from human-based assessment (HBA) remains unclear.

Objective: This study aimed to compare the quality of clinical interview assessments generated via an ABA (GPT-o1 Pro [ABA-o1] and GPT-5 Pro [ABA-5]) with those generated via an HBA conducted by clinical instructors in an AI-based virtual patient setting. We also examined whether AI reduced evaluation time and assessed agreement across participants with different levels of clinical experience.

Methods: A standardized case of leg weakness was implemented in an AI-based virtual patient. Seven participants (2 medical students, 3 residents, and 2 attending physicians) each conducted an interview with the AI patient, and transcripts were scored using the 25-item Master Interview Rating Scale (0‐125). Three evaluation strategies were compared. First, GPT-o1 Pro and GPT-5 Pro scored each transcript 5 times with different random seeds to test case specificity. Processing time was logged automatically. Second, 5 blinded clinical instructors independently rated each transcript once using the same rubric. Third, reliability metrics were applied. For AI, intraclass correlation coefficients (ICCs) quantified repeatability. For humans, the ICC(2,1) was calculated. Agreement was quantified using the Pearson *r*, Lin concordance correlation coefficient, Bland-Altman limits of agreement, Cronbach α, and ICC. Time efficiency was expressed as mean minutes per transcript and relative percentage reduction.

Results: Mean interview scores were similar across methods (ABA-o1: mean 52.1, SD 6.9; ABA-5: mean 53.2, SD 6.8; HBA: mean 53.7, SD 6.8). Agreement between ABA and HBA was strong (r=0.90; concordance correlation coefficient=0.88) with minimal bias (ABA-o1: mean 0.4, SD 2.7; ABA-5: mean 1.5, SD 5.2; limits of agreement: –4.9 to 5.7 for ABA-o1 and –8.6 to 11.7 for ABA-5). The Cronbach α was 0.81 (ABA-o1), 0.86 (ABA-5), and 0.80 (HBA); the ICC(3,1) was 0.77 (ABA-o1) and 0.82 (ABA-5); and the ICC(2,1) was 0.38 (HBA). The coefficient of variation for ABA was approximately half that of HBA (6.6% vs 13.9%). Processing time for 5 runs was 4 minutes, 19 seconds for ABA-o1 and 3 minutes, 20 seconds for ABA-5 vs 10 minutes, 16 seconds for physicians, corresponding to 58% and 67.6% reductions, respectively.

XSL•FO
RenderX

Conclusions: ABA-o1 and ABA-5 produced scores closely matching HBA while demonstrating superior consistency and reliability. In the setting of virtual interview transcripts, these findings suggest that ABA may serve as a valid, rapid, and scalable alternative to HBA, reducing per-assessment time by over half. Applied strategically, AI-based scoring could enable timely feedback, improve efficiency, and reduce faculty workload. Further research is needed to confirm generalizability across broader settings.

(JMIR Med Educ 2026;12:e81673) doi:10.2196/81673

KEYWORDS

medical education; artificial intelligence; AI; virtual patient; clinical interview; ChatGPT; simulation-based learning

Introduction

Background

Effective clinical interviewing is essential for making correct diagnoses and building strong relationships with patients [1]. Traditionally, students learn these skills through supervised practice with real or standardized patients and feedback from faculty [1]. However, this apprenticeship-style approach is time-intensive and limits opportunities for deliberate practice [2].

The assessment component itself also consumes substantial faculty and resident physician (RP) time. In competency-based medical education (CBME), faculty complete numerous workplace-based assessment forms; one Canadian study found a mean of 3 minutes, 6 seconds per entrustable professional activity form, adding approximately 18 minutes of extra documentation time for each staff member every 4-week block [3]. Multiprogram qualitative work further confirms that the cumulative “assessment burden” is now viewed as a major threat to sustainability, prompting programs to redesign processes to reduce administrative load [4].

Recently, generative artificial intelligence (AI) using large language models (LLMs) has enabled the creation of AI-based virtual patients that both converse with learners and automatically evaluate performance [2,5]. Empirical studies have shown promising results for AI assessment in free-text clinical documentation [6], script concordance testing [7], and objective structured clinical examination (OSCE) history-taking stations [8]. Many of these systems use validated rubrics such as the Master Interview Rating Scale (MIRS) to structure feedback [9]. Nevertheless, the reliability and validity of AI-generated ratings remain understudied; therefore, establishing concordance with expert evaluations is a prerequisite for educational or licensure use.

Objectives

This study compared AI-based assessment (ABA) scores of clinical interview performance using GPT-o1 Pro (OpenAI; ABA-o1) and GPT-5 Pro (OpenAI; ABA-5) with human-based assessment (HBA) scores. We hypothesized that ABA scores and HBA scores would exhibit strong concordance and that ABA scoring would serve as a substitute for HBA scoring. We also hypothesized that AI would complete evaluations more rapidly, reducing the assessment burden on clinicians. A secondary aim was to evaluate agreement across participants with differing clinical experience and evaluate whether the use of AI could lead to a measurable reduction in evaluation time, thereby contributing to overall efficiency in assessment processes.

Methods

Study Design and Setting

A cross-sectional validation study was conducted. This study involved 3 medical students (MSs), 3 RPs, and 2 attending physicians (APs) who participated in standardized clinical scenarios.

Virtual Patient Scenario

A man aged 27 years presenting with progressive bilateral leg weakness, particularly proximal, was scripted based on a published case of thyrotoxic periodic paralysis. The scenario, created by a general internal medicine specialist with extensive educational experience drawing directly on prior literature, included relevant clinical history (eg, recent myalgias, tremors, diarrhea, and insomnia), red flag cues (eg, acute onset, muscle weakness, and hypokalemia), and psychosocial factors (eg, recent immigration and use of herbal supplements). The case represented a classic presentation of thyrotoxic periodic paralysis caused by hyperthyroidism. The patient was implemented as an AI-simulated character using ChatGPT’s custom generative pretrained transformers.

Participants

The participants were recruited through convenience sampling complemented by snowball sampling. MSs comprised a third-year and a fifth-year student; RPs comprised 3 postgraduate year 1 residents; and APs comprised 2 board-certified physicians in internal medicine or general internal medicine in Japan, each with ≥5 years of clinical teaching experience.

Each participant conducted a history-taking encounter by speaking with an AI patient. All conversations were recorded and transcribed verbatim. As all interviews took place within the ChatGPT-based simulated patient interface, transcripts were automatically generated from dialogue logs without manual correction.

Scoring Instrument

The MIRS from the University of Tennessee was originally designed to assess 27 items. In this study, 25 of these items were evaluated based on the available conversational recordings. Each item was rated on a scale from 0 to 5 (total possible score: 0‐125) covering domains such as information gathering, organization, empathy, and patient-centered communication.

XSL•FO
RenderX

The excluded items were nonverbal behavior and pace and flow of the interview, which require audiovisual input to evaluate.

Assessment Methods

The main outcome was the comparison of MIRS scores.

For ABA-o1, each transcript was submitted separately to GPT-o1 Pro with a base prompt directing it to rate the encounter using the MIRS and justify each score. This process was repeated 5 times per transcript, and item-level and total scores were averaged across runs. For ABA-5, using the same base prompt, the 7 transcripts were scored in 2 batch submissions rather than individually. Run 1 included MS 1, MS 2, RP 1, and RP 2, and run 2 included RP 3, AP 1, and AP 2. For each batch, the prompt explicitly stated that it contained 4 interview transcripts (run 1) or 3 interview transcripts (run 2). For each participant within a batch, item-level and total MIRS scores were extracted from the model's output. The base prompt and model settings were held constant across runs, and the full prompt is provided in Multimedia Appendix 1 to support reproducibility. For both ABA-o1 and ABA-5, reproducibility was further examined by scoring each transcript 5 times using different random seeds.

For HBA, 5 blinded clinical instructors independently rated each transcript using the same MIRS rubric. All assessors were board certified in general internal medicine or general medicine in Japan, actively involved in medical education, and coauthors of this study (Y Tokushima, TS, RW, TM, and FS). Item-level and total scores were averaged across the 5 raters. To further ensure calibration beyond the preparatory webinar, raters briefly discussed scoring rationales for sample transcripts to reach consensus on the interpretation of rubric items.

The secondary outcome was the comparison of assessment time, which was assessed as follows:

- Physician scoring time for HBA—a stopwatch measured the time from transcript review to completion of scoring.
- AI scoring time for ABA-o1—the elapsed time was automatically recorded for each of the 7 individual submissions from prompt submission to receipt of the complete output.
- AI scoring time for ABA-5—the elapsed time was automatically recorded for each of the 2 batch submissions from prompt submission to receipt of the complete output.

For all 3 methods, mean assessment time and SD were calculated, and absolute and relative time savings of ABA vs HBA were reported.

Statistical Analysis

All analyses used R (version 4.3.1; R Foundation for Statistical Computing). Descriptive statistics (mean and SD and coefficient of variation [CV]) summarized the scores. Agreement was assessed using the Pearson correlation coefficient (*r*) for linear associations; the Lin concordance correlation coefficient (CCC) for both correlation and bias, summarizing overall agreement in a single index; and Bland-Altman analysis for bias and limits of agreement (LoA).

Reliability metrics included the Cronbach α for internal consistency, and intraclass correlation coefficients (ICCs) were calculated to quantify (1) repeatability across the 5 independent GPT-o1 Pro and GPT-5 Pro runs (stability of scores when the same model was applied repeatedly to the same transcript) and (2) interrater reliability across the 5 physician raters (agreement among different human raters). A 2-sided α of <.05 denoted significance.

Ethical Considerations

Ethics approval was obtained from the Juntendo University institutional review board (approval E24-0314-U02). All participants provided written informed consent before taking part. To protect participants' privacy and confidentiality, all interview transcripts and performance scores were deidentified prior to evaluation and analysis by assigning study IDs and removing any potentially identifying information. Only deidentified transcripts were shared with the physician raters, and results are reported in aggregate. Study data were stored on password-protected, access-restricted institutional systems, and only the research team had access. Participants received no financial compensation for participation.

Results

Participant Scores

Table 1 summarizes the interview scores obtained via ABA-o1, ABA-5, and HBA. Across all 7 participants, group-level means were 53.7 (SD 6.8) for HBA, 53.2 (SD 9.2) for ABA-5, and 52.1 (SD 6.9) for ABA-01. Within-participant variability (mean CV percentage) was similar for the 2 automated methods (ABA-o1=6.6%; ABA-5=6.6%) and higher for HBA (13.9%). Individual-level differences were generally small, although notable divergences arose for RP 2 when comparing HBA vs ABA-o1 (46.8 vs 53.4; Δ=6.6) and for AP 2 when comparing ABA-5 vs HBA (67.8 vs 58.8; Δ=9.0) and ABA-5 vs ABA-o1 (67.8 vs 55.6; Δ=12.2).

XSL•FO
RenderX

Table . Mean scores by method and participant (n=7).

Participant	HBA[a]		ABA-o1[b]		ABA-5[c]	
	Score (0-125), mean (SD)	CV[d] (%)	Score (0-125), mean (SD)	CV (%)	Score (0-125), mean (SD)	CV (%)
Medical student 1	48.0 (8.9)	18.5	46.4 (2.4)	5.2	46.0 (1.9)	4.1
Medical student 2	65.0 (9.7)	15.0	63.6 (5.1)	8.1	64.6 (4.2)	6.5
Resident physician 1	47.0 (2.9)	6.2	46.8 (2.9)	6.1	50.0 (2.6)	5.3
Resident physician 2	53.4 (7.2)	13.4	46.8 (3.3)	7.2	51.0 (7.1)	14.0
Resident physician 3	47.2 (3.6)	7.6	47.6 (2.7)	5.7	44.0 (1.0)	2.3
Attending physician 1	56.4 (9.4)	16.7	58.0 (5.4)	9.3	49.2 (2.6)	5.3
Attending physician 2	58.8 (11.7)	19.8	55.6 (2.7)	4.9	67.8 (6.2)	9.1
All	53.7 (6.8)	13.9	52.1 (6.9)	6.6	53.2 (9.2)	6.6

[a]HBA: human-based assessment.

[b]ABA-o1: artificial intelligence–based assessment (ABA) using GPT-o1 Pro.

[c]ABA-5: ABA using GPT-5 Pro.

[d]CV: coefficient of variation.

Agreement and Reliability Across ABA-o1, ABA-5, and HBA

Agreement and reliability were evaluated across the 3 rating methods (ABA-o1, ABA-5, and HBA). Pairwise concordance with HBA was high for both AI variants: ABA-o1 vs HBA showed a Pearson correlation coefficient (*r*) of 0.90 (95% CI 0.78 - 0.96) and CCC of 0.88; ABA-5 vs HBA showed an *r* of 0.87 (95% CI 0.72 - 0.94) and CCC of 0.86. Concordance between the 2 AI pipelines was the highest (ABA-o1 vs ABA-5: *r*=0.98, 95% CI 0.95 - 0.99; CCC=0.98), indicating near interchangeability of the AI variants (Table 2). Internal consistency followed the same pattern: Cronbach α was 0.81, 0.86, and 0.80 for ABA-o1, ABA-5, and HBA, respectively.

Table . Correlation, concordance, and internal consistency between artificial intelligence–based assessment (ABA) and human-based assessment (HBA) scores. Higher values indicate stronger association or consistency.

Comparison	Number of items	Pearson *r* (95% CI)	Lin CCC[a]
ABA-o1[b] vs HBA	25	0.90 (0.78 - 0.96)	0.88
ABA-5[c] vs HBA	25	0.87 (0.72 - 0.94)	0.86
ABA-o1 vs ABA-5	25	0.98 (0.95 - 0.99)	0.98

[a]CCC: concordance correlation coefficient.

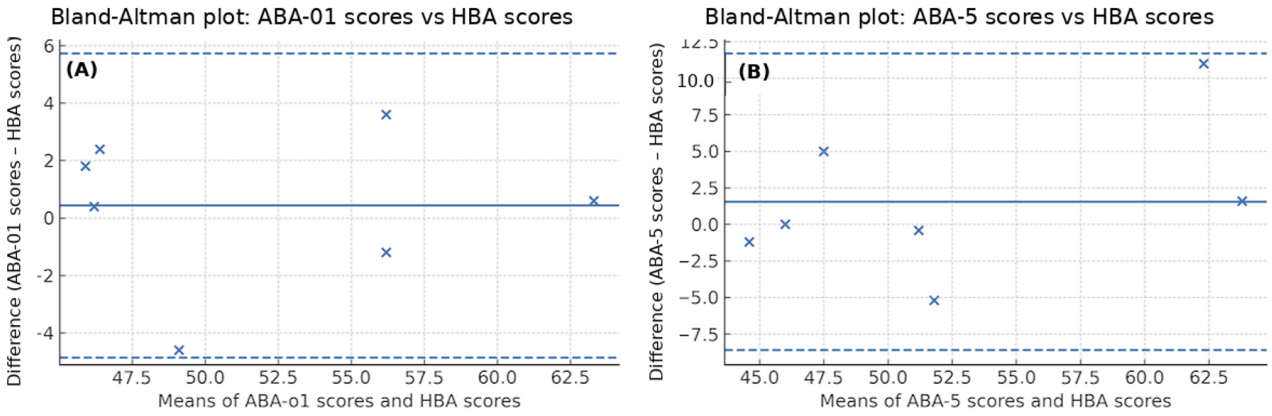
[b]ABA-o1: ABA using GPT-o1 Pro.

[c]ABA-5: ABA using GPT-5 Pro.

All correlations were significant (2-sided *P*<.001). Bland-Altman analyses comparing each ABA with HBA showed small positive mean biases (ABA-o1 vs HBA: +0.43 [SD of differences 2.70]; ABA-5 vs HBA: +1.54[SD of differences 5.17]), with 95% LoA of −4.87 to 5.72 and −8.60 to 11.68, respectively; no proportional bias was observed in either comparison (Figures 1A and 1B).

XSL•FO
RenderX

Figure 1. Bland-Altman plots comparing artificial intelligence–based assessment (ABA) with human-based assessment (HBA): (A) ABA using GPT-o1 Pro (ABA-o1) vs HBA (mean bias 0.43 [SD of differences 2.70]; limits of agreement [LoA]=−4.87 to 5.72) and (B) ABA using GPT-5 Pro (ABA-5) vs HBA (mean bias 1.54 [SD of differences 5.17]; LoA=−8.60 to 11.68). The points indicate participants (×). The solid line shows the mean bias; the dashed lines indicate the LoA.



Repeatability was assessed using the ICC. ABA-o1 showed substantial repeatability across 5 independent runs (ICC(3,1)=0.77; ICC(3,5)=0.94), and ABA-5 likewise showed substantial repeatability (ICC(3,1)=0.82; ICC(3,5)=0.96). In contrast, interrater reliability among the 5 HBA physician raters was only fair on single measures (ICC(2,1)=0.38) and improved when averaging them (ICC(2,5)=0.75). Overall, both AI-based approaches yielded more stable ratings across repeated evaluations than HBA, with ABA-5 slightly more stable than ABA-o1.

Scores by Training Level

Table 3 summarizes mean interview scores and SDs by training level. Across methods, APs had the highest means (HBA: 57.6, SD 1.7; ABA-o1: 56.8, SD 1.7; ABA-5: 58.5, SD 13.2). MSs were next (HBA: 56.5, SD 12.0; ABA-o1: 55.0, SD 12.2; ABA-5: 55.3, SD 13.2), in some cases approximating AP performance. RPs had the lowest means (HBA: 49.2, SD 3.6; ABA-o1: 47.1, SD 0.5; ABA-5: 48.3, SD 3.8). Therefore, the anticipated ordinal pattern (APs>RPs>MSs) was not consistently observed as MS means exceeded RP means across all methods.

Table . Mean interview scores by training level as rated via human-based assessment (HBA) and artificial intelligence–based assessment (ABA).

Group	Participants per group, n	HBA score (0-125), mean (SD)	ABA-o1[a] score (0-125), mean (SD)	ABA-5[b] score (0-125), mean (SD)
Attending physicians	2	57.6 (1.7)	56.8 (1.7)	58.5 (13.2)
Medical students	2	56.5 (12.0)	55.0 (12.2)	55.3 (13.2)
Resident physicians	3	49.2 (3.6)	47.1 (0.5)	48.3 (3.8)

[a]ABA-o1: ABA using GPT-o1 Pro.

[b]ABA-5: ABA using GPT-5 Pro.

Processing Time (35 Cases)

Total processing time was 5 hours, 59 minutes, 35 seconds for the physician benchmark; 1 hour, 56 minutes, 38 seconds for ABA-5; and 2 hours, 31 minutes, 5 seconds for ABA-o1. Average time per case was 3 minutes, 19.9 seconds for ABA-5 (batch-to-batch SD 1 minute, 6 seconds); 4 minutes, 19 seconds for ABA-o1 (SD 3 minutes, 9 seconds); and 10 minutes, 16.4 seconds for the physicians (SD 11 minutes, 9 seconds). Relative to the physicians, total time was reduced by 67.6% with ABA-5 and 58% with ABA-o1 (Table 4).

Table . Analysis time by method (5 independent runs and raters per method). “Batch-to-batch SD” indicates across-run variability. “Time reduction vs physicians” indicates the percentage reduction relative to human-based assessment (HBA).

Method	Total time	Mean time per case (batch-to-batch SD)	Time reduction vs physicians (%)
ABA-5[a]	1 h, 56 min, 38 s	3 min, 20 s (1 min, 6 s)	67.6
ABA-o1[b]	2 h, 31 min, 5 s	4 min, 19 s (3 min, 9 s)	58.0
HBA	5 h, 59 min, 35 s	10 min, 16 s (11 min, 9 s)	—[c]

[a]ABA-5: artificial intelligence–based assessment (ABA) using GPT-5 Pro.

[b]ABA-o1: ABA using GPT-o1 Pro.

[c]Not applicable.

XSL•FO
RenderX

Discussion

Principal Findings

In this validation study comparing 3 rater groups (HBA, ABA-o1, and ABA-5), ABA-o1 and ABA-5 produced interview ratings that were statistically indistinguishable from those produced via HBA yet showed markedly superior psychometric stability relative to HBA (Cronbach α: ABA-o1=0.81, ABA-5=0.86, and HBA=0.80; ICC: ABA-o1=0.77, ABA-5=0.82, and HBA=0.38). Cronbach α values of ≥0.8 indicate good internal consistency [10], and ICC(2,1) values of ≥0.75 denote good interrater reliability [11]. Agreement metrics were likewise robust as evaluative tools: the CCC assesses both correlation and bias in a single index [12], whereas Bland-Altman analysis remains the standard for visualizing bias and LoA [13]. ABA-5 was benchmarked against HBA using the same agreement framework.

Although the observed differences in reliability were significant, they may also have practical implications in educational settings. The consistently higher internal consistency and interrater reliability suggest that ABA scoring (including ABA-o1 and ABA-5) could enhance assessment efficiency and reproducibility. Depending on the context, ABA may serve not only as a scalable adjunct but also as a viable alternative to human raters in transcript-based clinical interview evaluations, although this requires significant larger-scale validation.

Comparison With Prior Work

These findings corroborate previous work in which LLMs matched or exceeded faculty performance when scoring free-text notes [6], designing script concordance tests [7], and evaluating OSCE encounters [8]. A recent study showed that GPT-4o can produce inpatient documentation of comparable quality to that produced by resident physicians while reducing charting time by >50% [14]. In particular, studies of OSCE history-taking and free-text documentation have begun to demonstrate that LLM-based raters can apply communication-focused rubrics in virtual or simulated encounters with performance comparable to that of trained faculty, underscoring the relevance of AI-supported assessment in simulation-based learning contexts [2,6,8]. Consistent with ChatGPT's passing performance on the US Medical Licensing Examination [15], this study suggests that foundation models possess clinically relevant semantic competence even in spoken communication tasks. Moreover, the 58% reduction in analysis time mirrors the 2025 Time for Class survey, where 36% of faculty who used generative AI daily reported a measurable workload decrease [16].

Beyond efficiency, such time savings could play a decisive role in addressing the growing problem of clinician educator burnout and faculty shortages, which are societal challenges that threaten the sustainability of CBME [17,18]. These concerns mirror prior reports that CBME assessment documentation requires several minutes per form and that the cumulative "assessment burden" is perceived as a threat to program sustainability [3,4]. By automating labor-intensive scoring, AI can free physicians to devote more time to high-value coaching and mentorship, thereby enhancing both educator well-being and learner support [17]. Furthermore, the superior scoring consistency observed with LLMs may help curb rater drift and cognitive biases such as leniency, halo, or contrast effects, which multicenter OSCE analyses have identified as long-recognized sources of unreliability and examiner-related variance in workplace-based assessments [19]. Improved fairness and reliability in assessment would advance equity in trainee progression and, ultimately, foster a more competent, patient-centered workforce.

Interpretation and Educational Implications

From an educational perspective, 3 observations are noteworthy when framed across the 3 rater groups (HBA, ABA-o1, and ABA-5).

Consistency vs Nuance

The score distributions for ABA-o1 and ABA-5 suggest that these models apply the rubric more consistently than HBA raters, likely because their underlying embeddings execute the criteria more deterministically once sampling stochasticity is averaged across runs. Consistency is a hallmark of fair assessment; however, the absence of human nuance in ABA-o1 and ABA-5 could miss contextual subtleties (eg, cultural cues and atypical communication styles) that HBA raters may detect. Such subtleties may include culturally patterned ways of showing respect or disagreement, indirect or high-context communication, and unconventional but effective rapport-building strategies that are difficult to fully capture in a text-based rubric. Accordingly, this balance between reproducibility and subtlety is central when integrating ABA into educational programs; in our view, ABA is best used to enhance reproducibility and efficiency, with human raters remaining essential for high-stakes decisions and for cases in which subtle contextual factors are educationally or ethically salient.

Efficiency Gains

Relative to HBA (10 minutes, 16 seconds per case), ABA-5 and ABA-o1 reduced analytic time to 3 minutes, 20 seconds (–67.6%) and 4 minutes, 19 seconds (–58%) per case, respectively, amounting to approximately 240 and 210 faculty minutes saved across 35 encounters, respectively. In throughput terms, this corresponds to an increase in throughput from approximately 6 cases per hour with HBA to 18 cases per hour with ABA-5 and 14 cases per hour with ABA-o1, supporting more timely formative feedback and enabling the reinvestment of AI-derived efficiency gains into coaching rather than grading. When viewed alongside the lower CV and higher reliability indexes for AI-based scoring, these efficiency gains suggest that ABA could support more consistent and sustainable assessment practices within CBME frameworks [3,4,19]. In addition, ABA-5 could process data for 3 to 4 individuals in a single run, reducing the need for repeated prompt inputs and minimizing data handling overhead.

Level-Based Performance

MSs outperformed RPs on the same rubric in this cohort. This pattern may reflect (1) sampling error in a modest cohort, (2) case specificity favoring recently studied content, and/or (3) a rubric that emphasizes foundational communication more than advanced clinical reasoning. In particular, the MIRS prioritizes patient-centered communication behaviors that are heavily

XSL•FO
RenderX

emphasized in undergraduate curricula and may be less sensitive to more advanced diagnostic reasoning skills typically developed during residency. Given the small number of participants and the single standardized case focused on a classic thyrotoxic periodic paralysis presentation, this unexpected pattern should be interpreted as a context-bound, hypothesis-generating finding rather than evidence that MSs generally outperform RPs in broader clinical performance. Replication with larger, more varied case sets and tiered rubrics evaluated across HBA, ABA-o1, and ABA-5 is warranted.

Practical Implications

Practically, programs could deploy an "AI-first, faculty-verified" workflow in which ABA-o1 and ABA-5 provide rapid formative scores and narrative feedback immediately after an encounter and HBA then audits a random subset for quality assurance, similar to double reading in radiology. Such hybridity leverages the speed and reliability of LLMs while retaining human oversight for high-stakes decisions.

Strengths and Limitations

A key strength is the dual evaluation of accuracy (agreement) and efficiency (time), providing a more complete picture of implementation value than through accuracy alone. Nonetheless, several limitations warrant caution:

First, only 7 participants and a single thyrotoxic periodic paralysis scenario were tested, limiting generalizability across learner levels, languages, and clinical contexts. The small and homogeneous sample also restricts the psychometric interpretation of the findings; for example, differences in learner experience, case complexity, and language environment may influence both human and AI scoring behaviors. Therefore, these results should be viewed as preliminary and hypothesis generating rather than confirmatory. Second, convenience sampling and self-selection may have introduced bias toward technology-friendly participants. Third, model and prompt dependence was a limitation; the results correspond to GPT-o1 Pro and GPT-5 Pro with a specific rubric prompt; other LLMs or prompt engineering strategies could alter performance. Fourth, speech-to-text errors were not exhaustively audited and may have influenced ratings. In addition, the evaluation was limited to transcribed textual data; nonverbal cues, vocal tone, and conversational pauses present in the actual interviews could not be assessed. Fifth, there was potentially a systemic bias. High concordance does not preclude shared cognitive blind spots between AI and human raters; fairness audits across sex, accent, and cultural communication styles remain necessary. In practical implementations, this would entail periodic subgroup analyses of score distributions, qualitative review of discrepant cases, and predefined procedures for pausing or adjusting AI-based scoring if systematic disparities are detected.

Future Research

Future studies should (1) evaluate multiple diverse clinical scenarios, including psychosocially complex cases; (2) compare real-time vs postencounter AI feedback; (3) examine learner outcomes such as skill acquisition and satisfaction; (4) conduct cost-effectiveness analyses at scale; and (5) develop and evaluate bias mitigation and explainability techniques—such as routine fairness dashboards, scheduled revalidation against human ratings, and faculty-led oversight processes—to satisfy accreditation requirements.

As this study was limited to transcript-based assessments of simulated encounters, future work is also needed to evaluate how well ABA scores correlate with actual clinical performance and whether AI can reduce assessor burden while maintaining fairness and reliability.

Conclusions

Within the constraints of this pilot, GPT-o1 Pro and GPT-5 Pro matched expert physicians in scoring simulated patient interviews; produced more reliable ratings; and delivered a substantial 58% and 67.6% reduction in analytical time, respectively. These preliminary results indicate that LLMs could serve as a complementary or alternative tool to human raters for transcript-based interview assessments. This approach warrants further investigation as a means to contribute to assessment efficiency in medical education. Careful curricular design and continuous human oversight will be essential to ensure that such tools enhance rather than compromise the validity and equity of learner evaluations.

Acknowledgments

The authors would like to thank the clinical instructors for their time and the participants for engaging in this educational pilot. The authors confirm that ChatGPT (OpenAI) was used during the study and manuscript preparation. Specifically, ChatGPT was used to assist with structuring and organizing the manuscript text (eg, improving clarity and logical flow of sentences and paragraphs) and the evaluation of medical interview transcripts according to predefined assessment criteria. All scientific content, study design, evaluation frameworks, data interpretation, and final editorial decisions were determined and verified by the authors, who take full responsibility for the content of the manuscript.

Funding

This study was supported by Japan Society for the Promotion of Science Grant-in-Aid for Scientific Research JP23K05953; the Japan Medical Education Foundation research grant for financial year 2025; and the "Creating training hubs for advanced medical personnel" program from the Ministry of Education, Culture, Sports, Science, and Technology of Japan.

XSL•FO
RenderX

Conflicts of Interest

None declared.

Multimedia Appendix 1
Prompt used for GPT-o1 Pro and GPT-5 Pro scoring of medical interview transcripts.
[DOCX File, 17 KB - mededu_v12i1e81673_app1.docx]

References

1. Talwalkar JS, Fortin AH, Morrison LJ, et al. An advanced communication skills workshop using standardized patients for senior medical students. MedEdPORTAL 2021 May 27;17:11163. [doi: 10.15766/mep_2374-8265.11163] [Medline: 34124349]
2. Cook DA. Creating virtual patients using large language models: scalable, global, and low cost. Med Teach 2025 Jan;47(1):40-42. [doi: 10.1080/0142159X.2024.2376879] [Medline: 38992981]
3. Cheung K, Rogoza C, Chung AD, Kwan BYM. Analyzing the administrative burden of competency based medical education. Can Assoc Radiol J 2022 May;73(2):299-304. [doi: 10.1177/08465371211038963] [Medline: 34449283]
4. Szulewski A, Braund H, Dagnone DJ, et al. The assessment burden in competency-based medical education: how programs are adapting. Acad Med 2023 Nov 1;98(11):1261-1267. [doi: 10.1097/ACM.0000000000005305] [Medline: 37343164]
5. Takahashi H, Shikino K, Kondo T, et al. Educational utility of clinical vignettes generated in Japanese by ChatGPT-4: mixed methods study. JMIR Med Educ 2024 Aug 13;10:e59133. [doi: 10.2196/59133] [Medline: 39137031]
6. Burke HB, Hoang A, Lopreiato JO, et al. Assessing the ability of a large language model to score free-text medical student clinical notes: quantitative study. JMIR Med Educ 2024 Jul 25;10:e56342. [doi: 10.2196/56342] [Medline: 39118469]
7. Hudon A, Kiepura B, Pelletier M, Phan V. Using ChatGPT in psychiatry to design script concordance tests in undergraduate medical education: mixed methods study. JMIR Med Educ 2024 Apr 4;10:e54067. [doi: 10.2196/54067] [Medline: 38596832]
8. Huang TY, Hsieh PH, Chang YC. Performance comparison of junior residents and ChatGPT in the objective structured clinical examination (OSCE) for medical history taking and documentation of medical records: development and usability study. JMIR Med Educ 2024 Nov 21;10:e59902. [doi: 10.2196/59902] [Medline: 39622713]
9. Master interview rating scale (MIRS).: Eastern Virginia Medical School; 2003. URL: https://www.dmu.edu/wp-content/uploads/Master-Interview-Rating-Scale.pdf [accessed 2025-07-23]
10. Tavakol M, Dennick R. Making sense of Cronbach's alpha. Int J Med Educ 2011 Jun 27;2:53-55. [doi: 10.5116/ijme.4dfb.8dfd] [Medline: 28029643]
11. Koo TK, Li MY. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. J Chiropr Med 2016 Jun;15(2):155-163. [doi: 10.1016/j.jcm.2016.02.012] [Medline: 27330520]
12. Lin LI. A concordance correlation coefficient to evaluate reproducibility. Biometrics 1989 Mar;45(1):255-268. [doi: 10.2307/2532051] [Medline: 2720055]
13. Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. Lancet 1986 Feb 8;1(8476):307-310. [Medline: 2868172]
14. Lu X, Gao X, Wang X, et al. Comparison of medical history documentation efficiency and quality based on GPT-4o: a study on the comparison between residents and artificial intelligence. Front Med (Lausanne) 2025;12:1545730. [doi: 10.3389/fmed.2025.1545730] [Medline: 40438356]
15. Gilson A, Safranek CW, Huang T, et al. How does ChatGPT perform on the United States Medical Licensing Examination (USMLE)? The implications of large language models for medical education and knowledge assessment. JMIR Med Educ 2023 Feb 8;9:e45312. [doi: 10.2196/45312] [Medline: 36753318]
16. Time for class 2025 report: daily AI use linked to reduced faculty workload. D2L Corporation. 2025. URL: https://www.d2l.com/newsroom/tyton_partners_report_examines_ai_in_higher_education/ [accessed 2025-07-23]
17. Banerjee G, Mitchell JD, Brzezinski M, DePorre A, Ballard HA. Burnout in academic physicians. Perm J 2023 Jun 15;27(2):142-149. [doi: 10.7812/TPP/23.032] [Medline: 37309180]
18. AAMC applauds introduction of bill to reduce physician shortage. Association of American Medical Colleges.: AAMC; 2025. URL: https://www.aamc.org/news/press-releases/aamc-applauds-introduction-bill-reduce-physician-shortage-0 [accessed 2025-07-29]
19. Yeates P, McCray G. Investigating the accuracy of adjusting for examiner differences in multicentre objective structured clinical exams (OSCEs): a simulation study of video-based examiner score comparison and adjustment (VESCA). BMC Med Educ 2024 Dec 18;24(1):1466. [doi: 10.1186/s12909-024-06462-3] [Medline: 39695612]

Abbreviations

ABA: artificial intelligence–based assessment
ABA-5: artificial intelligence–based assessment using GPT-5 Pro
ABA-o1: artificial intelligence–based assessment using GPT-o1 Pro

XSL•FO
RenderX

AI: artificial intelligence
AP: attending physician
CBME: competency-based medical education
CCC: concordance correlation coefficient
CV: coefficient of variation
HBA: human-based assessment
ICC: intraclass correlation coefficient
LLM: large language model
LoA: limits of agreement
MIRS: Master Interview Rating Scale
MS: medical student
OSCE: objective structured clinical examination
RP: resident physician

Edited by A Hasan Sapci; submitted 21.Aug.2025; peer-reviewed by A Das, D Cunha, SY Simhadri; revised version received 16.Nov.2025; accepted 14.Dec.2025; published 17.Feb.2026.

Please cite as:
Takahashi H, Shikino K, Kondo T, Yamada Y, Tomoda Y, Kishi M, Aiyama Y, Nagai S, Enomoto A, Tokushima Y, Shinohara T, Sano F, Matsuura T, Watanabe R, Naito T
AI- vs Human-Based Assessment of Medical Interview Transcripts in a Generative AI–Simulated Patient System: Cross-Sectional Validation Study
JMIR Med Educ 2026;12:e81673
URL: https://mededu.jmir.org/2026/1/e81673
doi:10.2196/81673

© Hiromizu Takahashi, Kiyoshi Shikino, Takeshi Kondo, Yuji Yamada, Yoshitaka Tomoda, Minoru Kishi, Yuki Aiyama, Sho Nagai, Akiko Enomoto, Yoshinori Tokushima, Takahiro Shinohara, Fumiaki Sano, Takeshi Matsuura, Rikiya Watanabe, Toshio Naito. Originally published in JMIR Medical Education (https://mededu.jmir.org), 17.Feb.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Education, is properly cited. The complete bibliographic information, a link to the original publication on https://mededu.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Evaluating AI-Generated Geriatric Case Studies for Interprofessional Education: Systematic Analysis Across 5 Platforms

Nicole Ruggiano[1], MSW, PhD; Sudikshya Sahoo[1], MSW; Ava Brashear[2]; Uche Nwatu[1], MSSW; Amie Brunson[1], LMSW, DSW; Hyunjin Noh[1], MSW, PhD; Heather Cole[3], RN, PhD; Robert McKinney[4], MSW, PhD; C Victoria Framil Suarez[5], RN, DNP; Ellen L Brown[5], RN, MS, EdD; Suzanne Prevost[3], RN, PhD

[1]School of Social Work, University of Alabama, Tuscaloosa, AL, United States

[2]Department of Kinesiology, College of Education, University of Alabama, Tuscaloosa, AL, United States

[3]Capstone College of Nursing, University of Alabama, Tuscaloosa, AL, United States

[4]College of Community Health Sciences, University of Alabama, Tuscaloosa, AL, United States

[5]Nicole Wertheim College of Nursing and Health Sciences, Florida International University, Miami, AL, United States

Corresponding Author:
Nicole Ruggiano, MSW, PhD
School of Social Work
University of Alabama
801 University Blvd.
129 Farrah Hall
Tuscaloosa, AL, 35401
United States
Phone: 1 2053484654
Email: nruggiano@ua.edu

Abstract

Background: Simulation-based learning (SBL) has become standard practice in educating health care professionals to apply their knowledge and skills in patient care. While SBL has demonstrated its value in education, many educators find the process of developing new, unique scenarios to be time-intensive, creating limits to the variety of issues students may experience within educational settings. Generative artificial intelligence (AI) platforms, such as ChatGPT (OpenAI), have emerged as a potential tool for developing simulation case studies more efficiently, though little is known about the performance of AI in generating high-quality case studies for interprofessional education.

Objective: This study aimed to generate geriatric case scenarios across 5 AI platforms by a transdisciplinary team and systematically evaluate them for quality, accuracy, and bias.

Methods: Ten geriatric case studies were generated using the same prompt from 5 different generative AI platforms (N=50): ChatGPT, Claude (Anthropic AI), Copilot (Microsoft), Gemini (Google), and Grok (xAI). An evaluation tool was developed to collect evaluative data to assess the content and quality of each case, sociodemographic data of the featured patient, the appropriateness of each case for interprofessional education, and potential bias. Case quality was evaluated using the Simulation Scenario Evaluation Tool (SSET). Each case was evaluated by 3 team members who had experience in SBL education. Assessment scores were averaged, and qualitative responses were extracted to triangulate patterns found in the quantitative data.

Results: While each AI platform was able to generate 10 unique case studies, the quality of studies varied within and across platforms. Generally, evaluators felt that the content in the cases was accurate, though some cases were not realistic. Some patient populations and common conditions among older adults were underrepresented or absent across the cases. All cases were set within traditional health care settings (eg, hospitals and routine medical visits). No cases featured home-based care. Based on the average SSET scores, reviewers assessed ChatGPT to be the highest overall performer (mean 3.27, SD 0.45, 95% CI 2.95-3.59) while Grok received the lowest scores (mean 1.61, SD 1.26, 95% CI 0.71-2.51). Platforms performed best at generating learning objectives (mean 3.35, SD 1.08, 95% CI 3.04-3.65) and lowest on their ability to describe supplies and materials that may be available in hypothetical scenarios (mean 1.27, SD 0.84, 95% CI 1.03-1.51).

XSL•FO
RenderX

Conclusions: This study is the first to systematically evaluate and compare multiple generative AI platforms for case study generation using a validated assessment tool (SSET) and provides evidence-based guidance on selecting and using AI tools effectively. The findings offer practical direction for educators navigating available generative AI tools to enhance training for health care professionals, including specific strategies for prompt engineering that can improve the quality of SBL resources in interprofessional education. These insights enable educators to leverage AI capabilities while maintaining pedagogical rigor.

(JMIR Med Educ 2026;12:e83085) doi:10.2196/83085

KEYWORDS

chatbots; generative artificial intelligence; geriatric nursing; interprofessional education; patient simulations

Introduction

Simulation-Based Learning in Health Care Professional Education

Simulation-based learning (SBL) is an integral component of health care professional education, offering a safe and controlled environment for students to develop clinical skills without risking patient safety [1,2]. SBL is recognized as an effective educational tool that has grown within higher education in many disciplines and can be implemented using a number of approaches, including role-playing, games, computer-assisted virtual instruction, and specially designed skills laboratories [3]. SBL often requires educators to develop highly detailed case studies to guide student learning activities, though these are often challenging to design. Recently, there has been increasing interest in the use of artificial intelligence (AI) platforms to support health care education [4]. While generative AI tools have the potential to generate case studies for use in SBL, little is known about the potential for AI platforms to generate high-quality, effective cases suitable for interprofessional health education. To address this, an interdisciplinary team systematically generated and evaluated geriatric case studies from 5 commercially available generative AI platforms and evaluated their content. The findings have implications for how AI platforms may be appropriately and ethically integrated into SBL.

History of Simulation-Based Learning in Interprofessional Education

The use of simulation in health care dates back decades, with early examples involving standardized patients, who were individuals trained to portray real patients [5]. This innovation allowed medical trainees to practice patient interactions and diagnostic skills in realistic scenarios. Advancements in technology have enhanced the fidelity, effectiveness, and types of simulation tools, leading to widespread adoption across various health care disciplines [6]. For example, the introduction of high-fidelity mannequins and virtual reality platforms has enabled the replication of complex clinical scenarios, thereby enhancing the realism and effectiveness of training sessions [7]. These technological innovations have expanded the scope of SBL, allowing for the practice of rare or high-risk procedures in a risk-free setting, which is crucial for developing competence in various medical specialties. Recent literature underscores the efficacy of SBL in improving clinical competencies among health care professionals. A systematic review by Sawaya et al [8] highlighted that SBL not only enhances immediate knowledge and skill acquisition but also contributes to long-term retention of clinical competencies. Furthermore, the integration of SBL into medical curricula has been associated with improved patient outcomes and a reduction in medical errors, emphasizing its critical role in contemporary health care professional education.

The Benefits and Drawbacks of Simulation-Based Education

One of the primary advantages of SBL is its ability to create a safe learning environment where learners can engage in high-risk scenarios or practice clinical skills without the risks associated with working with actual patients [9,10]. This immersive learning experience is further enhanced using high-fidelity simulators that mimic real patient responses, thereby engaging students emotionally and cognitively and improving the transfer of skills to clinical practice [9,11]. SBL has also been shown to improve learners' self-confidence and satisfaction. For example, previous studies have found that students who participated in SBL in emergency medicine reported higher satisfaction and performed better on assessments compared to those who received traditional instruction [12-14]. This suggests that SBL not only enhances knowledge acquisition but also fosters a positive learning experience. Interprofessional SBL may also emphasize the importance of teamwork in geriatric care, which promotes better collaboration and communication among health care professionals while providing a protected environment where trainees can practice handling complex clinical situations and learn from mistakes in a supportive setting [15-17]. Group-based SBL also promotes active engagement with equipment, simulated patients, and peers, leading to a deeper understanding and increased motivation to learn [18]. This interactive environment stimulates intellectual curiosity, encouraging critical thinking, questioning, and exploration of new knowledge, which aligns with curricula goals to foster curiosity and critical thinking in students.

Despite its numerous benefits, SBL is not without its challenges. One significant drawback is that SBL can be time- and resource-intensive (when using a lab setting) and may also pose challenges when integrated into an already crowded curriculum [9]. Lin et al [19] have asserted that ignoring the implementation and sustainability of SBL can result in higher costs, wasted resources, and the potential failure of educational interventions that are otherwise effective in achieving learning outcomes. Finally, educators may have difficulty in developing scenarios that are realistic and tailored to student learning needs. For example, certain physical findings, such as skin color variations, cannot be adequately represented in certain SBL approaches,

XSL•FO
RenderX

which may hinder the comprehensive training of students [9]. Instructors may wish to tailor scenarios to the varying abilities of students, but the standardized nature of many simulation exercises can hinder this individualized approach. This limitation may result in some learners not receiving the optimal level of challenge or support needed for their development.

Use of AI in Health Professions Education

Over the past decade, there has been increasing interest in identifying potential uses of large language model (LLM) AI tools in education for health care professionals. While the literature has identified opportunities for integrating AI into SBL activities, other potential applications have also been identified [20]. For example, Glauberman et al [20] posit that AI may enhance activities where students explore the impact of social determinants of health on patient care, support tutoring, and provide students with real-time feedback on assignments. However, they also identified potential challenges, such as AI providing hallucinations or biased content, and the risk that students may become over-reliant on AI, which could inhibit critical thinking skills [20].

Given the rising interest, there has been an increase in studies about integrating AI into health education settings. It should be noted that although there are currently multiple commercially available AI LLM platforms, much of the literature on using AI in health education has focused on ChatGPT (OpenAI), due to its rapid adoption among the public, reaching 1 million users in its first 5 days alone [21,22]. In a systematic review of the health education literature, Sallam [23] found that ChatGPT had the potential for creating tailored educational content and communication skill development for students that provides immediate feedback. However, challenges cited in the review included concerns regarding the potential bias and accuracy of information in educational material generated through LLMs.

Rogers et al [24] explored ChatGPT's capacity for generating simulation scenarios for health care education by evaluating 2 patient simulation cases it generated. Rather than prompting the platform to generate a complete case, in their study, they used a series of prompts to create individual components of a larger case (ie, develop a goal statement, create a corresponding scenario, and add specific detail based on existing information). There were several strengths the reviewers identified in the scenarios, including providing clear learning objectives, creating real-world scenarios, highlighting medication dosing guidelines, and describing participants' roles for each case [24]. They also identified several weaknesses in the cases, including inaccurate medication dosing, not providing treatment guidelines, missing details on equipment and supplies that would be needed for the case, objectives that would be too advanced for the learner audience, and inaccurate references [24]. Overall, reviewers reported that the main strengths were in the debriefing content, while inaccuracy posed the greatest problem, citing the need for users to fact-check the output of scenarios generated through ChatGPT.

Rationale for the Study

Geriatric education (or geriatric care) is an increasingly important area of health care professional training, given the growing population of older adults with complex and chronic health needs. As health care systems adapt to meet these challenges, equipping students with the knowledge and skills to provide competent, person-centered care to older adults is essential across disciplines. Generative AI tools have the potential to make some educational tasks more efficient, such as developing scenarios for interprofessional health education simulation exercises. However, the appropriateness of using AI platforms for this task remains unclear. The study by Rodgers et al [24] work evaluated the performance of ChatGPT in creating scenarios; it only evaluated 2 cases and did not compare ChatGPT with other AI platforms. To better understand how generative AI may be used by health educators, this study aimed to evaluate the performance of 5 generative AI platforms in creating high-quality case studies that could be used for interprofessional health education.

Methods

Overview

This study involved a systematic evaluation of geriatric case studies that were generated through AI platforms. The methods and results described below reflect the reporting standards for the Generative Artificial Intelligence Tools in Medical Research (GAMER) statement [25], which can be found through the Enhancing the Quality and Transparency of Health Research Network [26].

The cases were generated in December 2024 from 5 different AI platforms: ChatGPT, Claude (Anthropic AI), Gemini (Google), Grok (xAI), and Copilot (Microsoft). Copilot was chosen because it was the platform recommended by the research team's institution for use in education and research. The other 4 platforms were selected based on their rankings on the LMArena leaderboard, an open-access database of AI platforms developed at the University of California, Berkeley, where users rank various AI platforms on their performance [27]. Before evaluating cases, the team generated 2 cases on each platform's free version. However, it was discovered that 1 platform (Grok) did not have a free version available, and all the platforms varied in what was offered through their free versions (eg, accessing internet data in real time). Therefore, the decision was made to evaluate cases generated through the paid subscription versions of each platform to minimize bias. Table 1 provides a description and comparison of the platforms.

XSL•FO
RenderX

Table 1. Details of the 5 artificial intelligence (AI) platforms selected and their distinguishing qualities.

Platform	Developer	Launch date	Description and distinguishing features
ChatGPT [28]	OpenAI	November 2022 (Public launch)	Built on a generative pretrained transformer (GPT) architecture, the LLM[a] is trained on conversational data to generate human-like responses to queries.
Claude [29,30]	Anthropic AI	March 2023	LLM incorporating "Constitutional AI" principles based on the UN Declaration of Human Rights and AI[b] research ethics; emphasizes safety and neutrality through self-critiquing mechanisms.
Gemini [31,32]	Google	2023 (originally as Bard)	Multimodal LLM that processes words, audio, and pictures; designed for productivity and tasks including image generation.
Grok [33].	xAI (Elon Musk)	November 2023	LLM trained using systems such as Kubernetes, JAX, and Rust, designed to create a more efficient AI platform than those previously developed; has access to information in real time through the platform X (formerly Twitter); programmed to have a witty personality.
Copilot [34]	Microsoft	March 2023	Integrates with Microsoft 365 applications; can access users' personal data to provide tailored responses within a larger productivity suite.

[a]LLM: large language model.

[b]AI: artificial intelligence.

Generation of Case Studies

To assess the performance of AI platforms in generating geriatric case studies, each platform was asked to generate 10 unique case studies, for a total of 50 cases. The decision was made to submit 10 individual requests for cases on each platform rather than a single request for all 10 after discovering that the platforms created more detailed and robust cases when asked to create them one at a time rather than all at once. The prompts used to generate cases were as follows:

- Initial prompt: generate a geriatric case study that can be used in simulation learning for students from various health care professional disciplines, such as medical, nursing, and social work students.
- Subsequent prompts: generate another unique geriatric case study that can be used in simulation learning for students from various health care professional disciplines, such as medical, nursing, and social work students.

Team members uploaded the cases to a secure cloud storage platform. Two team members (NR and Ava Brashear) blinded the cases so that reviewers would not be able to link any individual case to the platform that generated it.

Case Study Evaluation Tool

The team developed an evaluation questionnaire for case study evaluators (ie, social work and nursing faculty) to use in assessing the case studies. The following describes the variables assessed.

Case Quality

Although there are several tools available to evaluate students' performance on SBL activities and the quality of debriefing, little work has been done to create assessment tools to evaluate the quality of simulation case content. The research team used the Simulation Scenario Evaluation Tool (SSET), developed by Hernandez et al [35], to evaluate the quality of case studies. The SSET is a standardized assessment tool that was developed using a modified Delphi approach, a structured, iterative process designed to build consensus among experts. The final version of the SSET includes 20 items that are organized into 5 elements, described in Table 2. The SSET also allows the reviewer to skip elements that are not available in the case study description. For each item, evaluators assess cases using a 5-point Likert scale with 3 anchor points. The anchor points are individualized for each item.

XSL•FO
RenderX

Table 2. Description of the assessment elements included in the Simulation Scenario Evaluation Tool (SSET).

Element	Element description	Number of items
Learning objectives	If learning objectives are included, the extent to which they are: a good fit for students' skill and knowledge levels, specific, measurable, action-oriented, relevant, and reflect different types of knowledge or skills.	7
Clinical context and scenario overview	The amount and quality of information provided about the case facilitate learning outcomes.	2
Critical actions	If the case study describes actions or decisions that the student should be able to demonstrate after reading the case study, the extent to which the actions are observable, support the learning objectives, and are attainable for the students' skill level.	3
Patient states	The details and cues provided to learners can help them carry out the critical actions, including the appropriateness of the case study progression, the ability of learners to take multiple pathways to addressing the case, and how they may help facilitate learning objectives, and require students to take critical actions about the case.	4
Scenario materials and resources	The identification of equipment, supplies, and resources that would normally be available and needed to demonstrate outcomes in clinical skills.	2
Debriefing plan	The extent to which the case guides a postsimulation discussion where learners can receive feedback on their performance.	2

Sociodemographic Data

Sociodemographic information (eg, race, gender, and socioeconomic status) from each of the cases was extracted to evaluate the diversity of patient populations presented across cases.

Appropriate for Interprofessional Education

Reviewers were asked to rate the extent to which cases were clinically accurate using a Likert scale of 1 (Completely Inaccurate) to 4 (Highly Accurate), with the option of "Unsure." They were also asked to rate the extent to which the cases adequately addressed ethical considerations using a Likert scale of 1 (Completely Disagree) to 5 (Completely Agree). Reviewers were asked to indicate whether they would use each case for SBL (Yes, No, Unsure). Open-ended questions allowed evaluators to provide feedback on the quality of cases.

Potential Bias

Open-ended questions were used to allow evaluators to identify content in each case that they perceived as biased or stereotypical of the patient populations.

Process for Evaluating Cases

All case studies were generated in November 2024 and were evaluated between December 2024 and April 2025. Two coauthors (SS and UN) extracted the sociodemographic data from each case (eg, patient gender, race, ethnicity, and living situation). Six coauthors (Amie Brunson, ELB, CVFS, HC, RM, and HN) with expertise in interprofessional education in health care settings applied the SSET to evaluate the quality of cases generated. Each case was randomly assigned to 3 evaluators for review. The evaluation tool was uploaded to Qualtrics (Qualtrics International Inc), and each team member entered their assessments remotely and securely into a single database.

Data Analysis

All data were maintained in a Microsoft Excel spreadsheet for analysis. Descriptive statistics were generated for all of the sociodemographic data, including means, SDs, and CIs. The 3 reviewers' scores on the SSET for each case were averaged. Qualitative responses were extracted by 2 team members with expertise in qualitative health-related research methodologies (NR and HN) and used to triangulate patterns found in the quantitative data.

Post Hoc Examination

After the initial evaluation of the 50 case studies generated by AI for this project, a post hoc analysis was conducted to examine how alternative prompt engineering strategies informed by the study's findings could potentially improve the quality of AI-generated case studies. A detailed prompt was developed that addressed weaknesses and gaps observed in the original analysis and was used to generate a single case study using Claude (Anthropic AI). Details about this analysis and its implications for prompt engineering are provided in the "Discussion" section.

Ethical Considerations

This study did not involve human participants. All data were generated by the AI platforms from information drawn from the internet and other sources on which they had been trained by their developers. No data about real patients were entered into the platforms for this study.

Results

Overall Performance

All of the platforms demonstrated the ability to generate comprehensive case studies, though the quality of the cases varied within and across platforms. All cases were organized into content sections, often including a patient profile and demographics, primary presenting complaints of the patient, lists of current medications, past medical history, social history, and assessment results. Some cases generated learning objectives for students without additional prompting. However, for cases generated by Grok, several were incomplete, sometimes being cut off midsentence. For example, one case study generated by Grok featuring the fictitious patient, Mr Jack O'Connor,

XSL•FO
RenderX

included, “Current Presentation: Mr O’Connor was referred to...” (sic). In another example featuring Mrs Elena Rodriguez, the case included, “Psychosocial Factors:-...” (sic).

Variety of Scenario Details

The cases featured fictitious patients who varied in social and demographic backgrounds. There was some variation in the presenting complaints and medical history, though cases tended to emphasize the most common conditions impacting older adults, such as dementia, diabetes, heart disease, and hypertension. Interestingly, none of the 50 cases addressed COVID-19.

Table 3 provides the breakdown of sociodemographic information provided in the patient cases. There was about equal representation of women (n=26) and men (n=24), as well as a variety of races and ethnicities. It should be noted that none of the cases identified the patient as being a member of the lesbian, gay, bisexual, transgender, queer, intersex, asexual, and more community, having undocumented immigration status, or identifying as Native Hawaiian, Alaska Native, or American Indian. There was also a lack of information regarding social determinants of health, most notably income and insurance information. Qualitative data provided by evaluators revealed that the topics of patient finances, culture, and religion were insufficiently addressed in most cases, which did not provide students with adequate context for addressing patient concerns. For example, patients’ decisions regarding treatment or medication, as well as behaviors such as medication noncompliance, may be influenced by their income, cultural background, or religious beliefs.

XSL•FO
RenderX

Table 3. Sociodemographic characteristics of the fictitious patients featured in the case studies generated by artificial intelligence (AI) platforms.

Characteristic	Value, n (%)
Sex	
Cisgender male	24 (48)
Cisgender female	26 (52)
Race and ethnicity	
Caucasian or White	9 (18)
African American, Black, or Afro-Caribbean	8 (16)
Latino or Latina	5 (12)
Asian or Pacific Islander	9 (18)
Native Hawaiian, Alaska Native, or American Indian	0 (0)
Other	2 (4)
Unknown	16 (32)
English proficiency	
Yes	21 (42)
No (has a family member interpreting)	3 (6)
Unknown	26 (52)
Sexual orientation	
Heterosexual	29 (58)
Unknown	21 (42)
Marital status	
Married	17 (34)
Divorced	5 (10)
Widowed	26 (52)
Committed partner	1 (2)
Unknown	1 (2)
Employment	
Never worked	1 (2)
Retired, not working	45 (90)
Working full-time	1 (2)
Other	3 (6)
Patient household income	
Identified as low-income	3 (6)
Identified as moderate-income	1 (2)
Identified as high-income	0 (0)
Unknown or not provided	46 (92)
Source of payment for health services	
Medicare	11 (22)
Medicare and Medicaid	5 (10)
Medicare and VA[a] benefits	2 (4)
Medicare with supplemental private insurance	2 (4)
Unknown	30 (60)
Living situation	
Lives at home alone	15 (30)

XSL•FO
RenderX

Characteristic	Value, n (%)
Lives with others in their own home	17 (34)
Lives with others in another person's home	3 (6)
Lives in an independent living facility	4 (8)
Lives in an assisted living facility	6 (12)
Lives in a nursing home or a similar setting	3 (6)
Unknown	2 (4)
Religious affiliation	
Christian (Catholic)	1 (2)
Jewish	1 (2)
Hindu	1 (2)
Other	2 (4)
Unknown	45 (90)
Immigration and citizenship	
Immigrant	4 (8)
Unknown	46 (92)

[a]VA: Veterans Affairs.

Case Quality

Accuracy and Best Practices

For most cases, evaluators assessed the information provided as being generally accurate and reflecting best practices. However, they reported discrepancies with some case information, most often due to insufficient detail. For example, in reference to the SBL instructions for students, "*Develop a nursing care plan for hip fracture recovery, emphasizing mobility, prevention of pressure ulcers, and patient education on safety at home post*" in the case of Agness Muller (Grok), one evaluator provided qualitative feedback:

> *This instruction lacks detail and the usual patient progression. For example, a patient living alone with mild cognitive impairment (MCI)—based only on the Montreal Cognitive Assessment, which is not conclusive—and with a hip fracture would not be discharged home.*

However, in a few cases, the evaluators felt that the case information was not realistic. For example, in another case of George Hawkins (Claude), one evaluator commented: "*I do not feel like the patient would be coming in stating that he is forgetting things. This information would be coming from the caregiver.*" Similarly, another evaluator commented on the same case: "*I think having the patient present with complaints of confusion would be a little unrealistic. We would likely see the caregiver reporting these findings to the health care provider.*"

SSET Scores

For most cases, enough details were provided for reviewers to evaluate the 6 elements represented in the SSET assessment. Yet, a small number of cases lacked details on 1 or more elements. Table 4 shows the average scores assigned by reviewers along with their SDs and CIs for each element across AI platforms.

XSL•FO
RenderX

Table 4. Evaluators' average scores, with SDs and 95% CIs, on the Simulation Scenario Evaluation Tool (SSET) and its 6 elements.

Platform	E[a]1, mean, SD, 95% CI	E2, mean, SD, 95% CI	E3, mean, SD, 95% CI	E4, mean, SD, 95% CI	E5, mean, SD, 95% CI	E6, mean, SD, 95% CI	Overall, platform mean, SD, 95% CI
ChatGPT	4.22, 0.29, 4.01-4.42	3.89, 0.82, 3.30-4.47	3.71, 0.63, 3.26-4.15	3.28, 0.95, 2.60-3.96	1.78, 1.05, 1.02-2.53	2.74, 0.84, 2.14-3.34	3.27, 0.45, 2.95-3.59
Claude	3.74, 0.57, 3.33-4.15	3.59, 1.12, 2.79-4.39	3.34, 0.93, 2.68-4.00	3.50, 1.19, 2.65-4.35	1.58, 0.77, 1.03-2.13	1.70, 0.98, 1.00-2.40	2.91, 0.63, 2.46-3.36
Copilot	3.32, 0.42, 3.02-3.63	2.96, 0.83, 2.36-3.55	2.75, 0.93, 2.08-3.42	2.10, 0.67, 1.62-2.58	0.99, 0.28, 0.78-1.19	1.26, 0.71, 0.75-1.77	2.23, 0.39, 1.95-2.51
Gemini	3.30, 0.49, 2.95-3.66	3.15, 0.77, 2.60-3.71	2.87, 0.66, 2.39-3.35	2.39, 0.73, 1.87-2.91	1.32, 0.84, 0.72-1.92	1.20, 0.66, 0.73-1.67	2.37, 0.49, 2.02-2.72
Grok	2.15, 1.69, 0.93-3.37	2.52, 1.97, 1.10-.93	1.90, 1.45, 0.87-2.94	1.53, 1.38, 0.53-2.52	0.70, 0.73, 0.18-1.22	0.84, 1.03, 0.11-1.58	1.61, 1.26, 0.71-2.51
Overall, E[a],	3.35, 1.08, 3.04-3.65	3.22, 1.24, 2.87-3.57	2.91, 1.11, 2.60-3.23	2.56, 1.23, 2.21-2.91	1.27, 0.84, 1.03-1.51	1.55, 1.05, 1.25-1.85	2.48, 0.90, 2.22-2.73

[a]E: element.

Generally, the AI platforms performed best and most consistently at providing details related to element 1 (E1: learning objectives). While objectives varied in focus and scope, the platforms generated learning objectives that were detailed. Except for Grok, the mean scores for the platforms were relatively good (range of 3.30-4.22). ChatGPT particularly excelled at generating learning objectives with consistency, given the narrow CIs (mean 4.22, SD 0.29, 95% CI 4.01-4.42). While Grok received lower scores for E1, there was also variability in their quality (mean 2.15, SD 1.69; CI 0.93-3.37).

In some cases, the AI generated a set of learning objectives tailored to the various disciplines that may be involved in the simulation learning activities. Textbox 1 provides an example generated from Grok.

Textbox 1. Learning objectives generated by Grok for the case of Mrs Isabella Bianchi.

Medical students:

- Manage complex medication regimen for multiple chronic conditions, focusing on interactions and side effects.
- Investigate causes of weight loss in older adults, considering both medical and psychological factors.

Nursing students:

- Implement and teach strategies for medication adherence in patients with mild cognitive impairment.
- Develop a care plan for preventing falls and managing chronic constipation.

Social work students:

- Address the emotional impact of transitioning to assisted living, focusing on preserving dignity and a sense of purpose.
- Explore community resources or programs that could engage Mrs Bianchi's interest in music, potentially improving her mental health.

Interdisciplinary goals:

- Enhance Mrs Bianchi's quality of life by balancing her medical needs with her personal interests and psychological well-being.
- Coordinate a care plan that includes physical activity tailored to her condition, dietary adjustments for heart health and bone density, and social engagement to combat isolation.

The AI platforms also performed moderately well on element 2 (E2: clinical context and scenario overview), as evidenced by mean scores ranging from 2.96 to 2.52. Claude (mean 3.59, SD 1.12) and Grok (mean 2.52, SD 1.97) demonstrated the greatest variability in quality.

Evaluators generally felt that the platforms provided sufficient detail for students to understand the presenting case and address the learning objectives. However, they also noted that most cases described clinical contexts occurring within a health care facility (eg, hospital and primary care office), which may limit applicability for students in professions that often interact with patients in other settings (eg, home visits and pharmacies). There were exceptions, such as the example generated by Gemini in Textbox 2, where a home setting was used as the clinical context.

XSL•FO
RenderX

Textbox 2. Clinical context description for the case Mr Arthur Chen generated by Gemini.

Simulation activities:

- Home visit: students can conduct a simulated home visit to assess Mr Chen's living environment and provide education on disease management and self-care.
- Caregiver support group: students can role-play a caregiver support group to provide Mr Chen with an opportunity to share his experiences and connect with others.
- Interprofessional case conference: students from different disciplines can participate in a case conference to discuss Mr Chen's care plan and coordinate services.
- Medication management simulation: students can practice medication reconciliation and develop strategies to improve adherence.
- Advance care planning role-play: students can role-play a conversation with Mr Chen and his family about advance care planning.

Element 3 (E3: critical actions) and element 4 (E4: patient states) were rated more modestly by evaluators, with average scores of 2.91 (SD 1.11) and 2.56 (SD 1.23), respectively. Evaluators generally felt that the cases offered enough detail for students to make decisions about patient care and carry out critical actions to support patient care needs. In many cases, a progression of the scenario was presented, as exemplified in Textbox 3.

Textbox 3. Scenario progression for the case of Mr Walter Freemen, generated by Claude (UTI: urinary tract infection).

Simulation scenario progression:

- Initial assessment: each discipline conducts its respective assessments in the emergency department or upon admission.
- Team huddle: interdisciplinary team meets to discuss findings and develop an initial care plan.
- Acute management: simulate management of UTI and delirium, including nonpharmacological interventions.
- Medication review: team collaborates on medication reconciliation, considering potential cognitive effects and drug interactions.
- Family meeting: simulate a meeting with Mrs Evelyn Chen (when more lucid) and her husband to discuss diagnosis, treatment plan, and support needs.
- Discharge planning: team develops a comprehensive plan for transition of care, including medication management, follow-up appointments, and caregiver support.
- Follow-up: simulate a postdischarge follow-up appointment to assess resolution of acute issues and manage chronic conditions.

Evaluators rated the cases the lowest for element 5 (E5: scenario materials) and element 6 (E6: debriefing plan). All platforms performed poorly on E5 (mean 1.27, SD 0.84, 95% CI 1.03-1.51) and E6 (mean 1.55, SD 1.05, 95% CI 1.25-1.85). For E5, ChatGPT led with an average score of 1.78 (SD 1.05), followed by Claude (mean 1.58, SD 0.77) and Gemini (mean 1.32, SD 0.84). Copilot (mean 0.99, SD 0.28) and Grok (mean 0.70, SD 0.73) had the lowest scores. Evaluators found it difficult to evaluate the supplies and materials needed for students to complete the simulation cases, as they had to make assumptions about what would typically be available in the specific setting (eg, emergency room). However, they acknowledged that greater detail should be provided by most of the platforms to facilitate learning for students who may be less familiar with the setting. For E6, Grok had the lowest score at 0.84 (SD 1.03), followed by Gemini (mean 1.20, SD 0.66), Copilot (mean 1.26, SD 0.71), and Claude (mean 1.70, SD 0.98). ChatGPT fared slightly better in this area with an average of 2.74 (SD 0.84), but still fell short of high performance. Evaluators noted that many cases lacked a debriefing plan, which would require that the educator create one to coincide with the cases.

Comparison of Platform Performance

When asked whether they would use the presenting case for simulation learning, evaluators most often stated that they would use the cases generated by ChatGPT and were least likely to use cases generated by Copilot. When evaluators were asked to rate the extent to which they agreed that the case addressed ethical issues that students may experience in health care settings, ChatGPT and Gemini generally received the highest scores (mean 3.50, SD 0.61 and mean 3.51, SD 0.83, respectively), corresponding to "somewhat agree." Although evaluators thought many of the cases did not sufficiently address issues of culture, they did not report observing examples of gender or racial stereotypes.

Table 4 provides the summary scores for the SSET across platforms. Overall, ChatGPT performed the best in generating quality case studies, while Grok was rated the lowest, mostly due to the incomplete content generated in many of its cases. However, the platforms varied in their strengths and weaknesses. Among all platforms, ChatGPT performed the best in element 1: learning objectives, with the highest average score of 4.22 (SD 0.29), followed by Claude at 3.74 (SD 0.57). Copilot and Gemini had almost identical scores, with averages of 3.32 (SD 0.42) and 3.30 (SD 0.49), respectively, showing only a marginal difference. Grok, however, had the lowest performance in this element, with an average score of just 2.15 (SD 1.69). For element 2, which focused on understanding the relevance and alignment of the clinical context and scenario across case studies, the average scores of ChatGPT (mean 3.89, SD 0.82),

XSL•FO
RenderX

Claude (mean 3.59, SD 1.12), and Gemini (mean 3.15, SD 0.77) were relatively close. In contrast, Copilot (mean 2.96, SD 0.83) and Grok (mean 2.52, SD 1.97) had lower averages.

Discussion

Principal Findings

This study aimed to evaluate the performance of 5 commercially available AI platforms in their ability to generate geriatric case studies that could be used for simulation learning within interprofessional education settings. Although there was variation in their performance, all 5 platforms evaluated for this study were able to generate comprehensive case studies that generally included accurate information and a variety of patient contexts. ChatGPT was found to be the most reliable of the AI platforms, as demonstrated by consistently high scores and narrower CIs. Grok performed unpredictably and consistently received the lowest scores. It is not surprising that the AI platforms varied in their performance, given that the content generated would be specific to the data used to train each platform and its algorithms. For example, Claude's algorithm emphasizes a Constitutional AI framework, which may result in it integrating different content than other platforms.

Implications

A central benefit of using AI for SBL is that educators have the ability to develop a large number of diverse case studies in a short period of time. However, there were some notable shortcomings of the platforms, including their underrepresentation of some vulnerable patient populations, limited attention to cultural sensitivity and religion in health care, and restricted patient settings. The findings have implications for educators who are interested in using AI tools for SBL within interprofessional education settings.

One of the biggest implications is that educators who generate SBL cases using generative AI must provide oversight of the content produced [36]. Research has found that content generated through AI platforms can, at times, be falsely fabricated (ie, hallucinations) or can draw on content that is outdated or biased [37]. This concern was demonstrated in this study, where evaluators felt that some content provided in cases did not adequately reflect real-world scenarios. Another example is that some vulnerable patient populations and medical conditions (such as COVID-19) were not represented in any of the 50 cases. For example, there was a complete absence of any mention of patients who identify as lesbian, gay, bisexual, transgender, queer, intersex, and asexual in any of the selected platforms. To address these shortcomings of generative AI, educators should seek training in "prompt engineering," which refers to skills in generating prompts for AI tools that are more effective in addressing the goals of the AI user [38]. Meskó [38] indicates that prompt engineering is an important skill for health educators and health care practitioners and requires them to become knowledgeable about how individual AI platforms are designed and function so that they can develop the most effective prompts to meet their needs. In the case of health educators, prompt engineering could support them in developing AI-generated content (SBL case studies or otherwise) that better aligns with the learning objectives they have established for their students.

The Importance of Prompt Engineering

For this study, the same general prompt was used to develop each case study. Prompt writing is recognized as a skill that can support LLM generative AI tools in tailoring responses for the user [39]. In a previous study by Rogers et al [24], the researchers evaluated the performance of ChatGPT in generating 2 case studies. However, rather than using a general prompt to generate a complete case, they used a series of prompts to guide the platform for each section of the case study (eg, presenting problem and social history). Based on findings from this study, prompt engineering skills could allow educators to more efficiently use generative AI tools that address gaps in content (eg, patient populations and conditions addressed) as well as the quality of the cases generated.

For example, in a post hoc examination, we generated an additional case study using Claude that was more specific regarding quality:

> *Generate a geriatric case study that can be used in simulation learning for students from various health professional disciplines, such as medical, nursing, and social work students. The content should reflect the five elements of case studies that are emphasized in the Simulation Scenario Evaluation Tool (SSET), developed by Hernandez et al [35]. For social work students, the activity should address the Educational Policy and Accreditation Standards developed by CSWE.*

The resulting case study organized the content into learning objectives, clinical context, scenario overview, critical actions, patient states and progression, and scenario materials and resources. Thus, the prompt allowed Claude to build the case with established standards in mind. The resulting case study also linked content to specific Council on Social Work Education accreditation standards, such as asking social work students, "*How did you address social justice issues like financial barriers and access to care?(CSWE EPAS Competency 3)*" and "*What advocacy role did you take, and how did you respect Mrs. Martinez's right to self-determination?(CSWE EPAS Competency 2).*"

This approach may help educators optimize the performance of AI tools in generating case studies that meet specific learning goals. Similarly, more tailored prompts could be used to generate cases featuring underrepresented patient populations or specific aspects of case studies that AI platforms may overlook (eg, ethical dilemmas, specific health conditions, or cases in less common health care settings).

Limitations

There were numerous strengths to the methodology used for this study, including the use of paid subscriptions to the AI platforms to generate cases and having 3 evaluators assess each case using a standardized assessment tool. There were also limitations. First, the case studies were all generated in November 2024, and advancements in AI platforms since then may affect current performance. Second, the team used the paid

XSL•FO
RenderX

subscriptions of each AI platform to reduce potential bias from varying access limitations across platforms. However, this may limit generalizability, as most users are more likely to access the free versions of these platforms. It is important to note that both paid and free versions of these platforms have continued to evolve since data collection, making it unclear how results would differ if the study were replicated today with either paid or free versions.

Also, some of the gaps in content that were identified by reviewers across case studies may not have been present if more specific prompts were used (eg, "*Create a case study that features a patient with COVID-19*"). It is also important to highlight that the only perspectives provided about the AI-generated case studies were those of experts in interprofessional health education. The study did not obtain perspectives from students participating in SBL activities. It also did not evaluate how the AI-generated case studies compared to case studies created by professional educators. While this analysis highlighted some of the weaknesses in the AI-generated case studies, they may still be of comparable or higher quality than those generated by humans. Future research should address these gaps.

Conclusion

The functions and accessibility of generative AI have increased dramatically over the past few years. Health care educators and providers are increasingly using these tools to create efficiencies in their work. In the case of interprofessional health education, AI can support SBL by creating a large variety of case studies in a short period of time. However, as with other applications of AI, human oversight is needed to make sure that the output from these platforms appropriately meets the learning needs of students. Although previous studies have examined AI's potential role in SBL, this study is the first to systematically evaluate and compare multiple generative AI platforms for case study generation using a validated assessment tool (SSET). By quantifying quality differences across platforms and identifying systematic deficiencies, this research provides evidence-based guidance for educators on selecting and using AI tools effectively. The findings offer practical direction for educators navigating available generative AI tools to enhance training for health care professionals, including specific evidence-based strategies for prompt engineering that can improve the quality of SBL resources in interprofessional education. These insights enable educators to leverage AI capabilities while maintaining pedagogical rigor, ultimately supporting more efficient development of high-quality simulation training materials.

Acknowledgments

Generative AI was used to generate the case study data evaluated for this project, to assist in editing for grammar, and to assist in editing citations and formatting the content provided in Tables 2 and 4. The data presented in the "Results" section for the cases consisted of the actual text generated by the platforms and were not content created or edited by the research team.

Funding

No external financial support or grants were received for this work.

Data Availability

A copy of all 50 case studies generated by the 5 AI platforms is available as a Multimedia Appendix 1. The case studies are blinded to indicate which platform generated each case.

Conflicts of Interest

None declared.

Multimedia Appendix 1
All case studies generated by artificial intelligence (AI) and subsequently evaluated.
[DOCX File , 446 KB - mededu_v12i1e83085_app1.docx]

References

1. Kim S, Lee H, Connerton TP. How psychological safety affects team performance: mediating role of efficacy and learning behavior. Front Psychol 2020;11(1581):1-15 [FREE Full text] [doi: 10.3389/fpsyg.2020.01581] [Medline: 32793037]
2. Keddington AS, Moore J. Simulation as a method of competency assessment among health care providers: a systematic review. Nurs Educ Perspect 2019;40(2):91-94. [doi: 10.1097/01.nep.0000000000000433] [Medline: 30601272]
3. Sanko JS. A brief history of its use in nursing education. Q Rev Distance Educ 2017;18(2):77-85 [FREE Full text]
4. Zhang W, Cai M, Lee HJ, Evans R, Zhu C, Ming C. AI in medical education: global situation, effects and challenges. Educ Inf Technol 2023;29(4):4611-4633. [doi: 10.1007/s10639-023-12009-8]
5. Flanagan OL, Cummings KM. Standardized patients in medical education: a review of the literature. Cureus 2023;15(7):e42027 [FREE Full text] [doi: 10.7759/cureus.42027] [Medline: 37593270]

XSL•FO
RenderX

6. Englander R, Cameron T, Ballard AJ, Dodge J, Bull J, Aschenbrener CA. Toward a common taxonomy of competency domains for the health professions and competencies for physicians. Acad Med 2013;88(8):1088-1094. [doi: 10.1097/ACM.0b013e31829a3b2b] [Medline: 23807109]
7. Aebersold M. The history of simulation and its impact on the future. AACN Adv Crit Care 2016;27(1):56-61. [doi: 10.4037/aacnacc2016436] [Medline: 26909454]
8. Sawaya RD, Mrad S, Rajha E, Saleh R, Rice J. Simulation-based curriculum development: lessons learnt in global health education. BMC Med Educ 2021;21(1):33 [FREE Full text] [doi: 10.1186/s12909-020-02430-9] [Medline: 33413346]
9. Krishnan D, Keloth AV, Ubedulla S. Pros and cons of simulation in medical education: a review. Int J Med Health Res 2017;3(6):84-87 [FREE Full text]
10. Siew AL, Wong JW, Chan E. Effectiveness of simulated patients in geriatric education: a scoping review. Nurse Educ Today 2021;100:104856. [doi: 10.1016/j.nedt.2021.104856] [Medline: 33740706]
11. Weller JM, Nestel D, Marshall SD, Brooks PM, Conn JJ. Simulation in clinical teaching and learning. Med J Aust 2012;196(9):594. [doi: 10.5694/mja10.11474] [Medline: 22621154]
12. Alanazi A, Nicholson N, Thomas S. The use of simulation training to improve knowledge, skills, and confidence among healthcare students: a systematic review. Internet J Allied Health Sci Pract 2017;15(3) [FREE Full text] [doi: 10.46743/1540-580X/2017.1666]
13. Keiser MM, Turkelson C. Using simulation to evaluate clinical performance and reasoning in adult-geriatric acute care nurse practitioner students. J Nurs Educ 2019;58(10):599-603. [doi: 10.3928/01484834-20190923-08] [Medline: 31573650]
14. Mehdi Z, Ross A, Reedy G, Roots A, Ernst T, Jaye P, et al. Simulation training for geriatric medicine. Clin Teach 2014;11(5):387-392. [doi: 10.1111/tct.12156] [Medline: 25041674]
15. Braude P, Reedy G, Dasgupta D, Dimmock V, Jaye P, Birns J. Evaluation of a simulation training programme for geriatric medicine. Age Ageing 2015;44(4):677-682. [doi: 10.1093/ageing/afv049] [Medline: 25953500]
16. Langdalen H, Abrahamsen EB, Sollid SJM, Sørskår LIK, Abrahamsen HB. A comparative study on the frequency of simulation-based training and assessment of non-technical skills in the Norwegian ground ambulance services and helicopter emergency medical services. BMC Health Serv Res 2018;18(1):509 [FREE Full text] [doi: 10.1186/s12913-018-3325-1] [Medline: 29970079]
17. Lateef F. Simulation-based learning: just like the real thing. J Emerg Trauma Shock 2010;3(4):348-352 [FREE Full text] [doi: 10.4103/0974-2700.70743] [Medline: 21063557]
18. Bland AJ, Tobbell J. Towards an understanding of the attributes of simulation that enable learning in undergraduate nurse education: a grounded theory study. Nurse Educ Today 2016;44:8-13. [doi: 10.1016/j.nedt.2016.05.011] [Medline: 27429323]
19. Lin Y, Cheng A, Hecker K, Grant V, Currie GR. Implementing economic evaluation in simulation-based medical education: challenges and opportunities. Med Educ 2018;52(2):150-160. [doi: 10.1111/medu.13411] [Medline: 28949032]
20. Glauberman G, Ito-Fujita A, Katz S, Callahan J. Artificial intelligence in nursing education: opportunities and challenges. Hawaii J Health Soc Welf 2023;82(12):302-305 [FREE Full text] [Medline: 38093763]
21. Lee H. The rise of ChatGPT: exploring its potential in medical education. Anat Sci Educ 2024;17(5):926-931. [doi: 10.1002/ase.2270] [Medline: 36916887]
22. Wu T, He S, Liu J, Sun S, Liu K, Han Q, et al. A brief overview of ChatGPT: the history, status quo and potential future development. IEEE/CAA J Autom Sinica 2023;10(5):1122-1136. [doi: 10.1109/jas.2023.123618]
23. Sallam M. ChatGPT utility in healthcare education, research, and practice: systematic review on the promising perspectives and valid concerns. Healthcare (Basel) 2023;11(6):887 [FREE Full text] [doi: 10.3390/healthcare11060887] [Medline: 36981544]
24. Rodgers DL, Needler M, Robinson A, Barnes R, Brosche T, Hernandez J, et al. Artificial intelligence and the simulationists. Simul Healthc 2023;18(6):395-399. [doi: 10.1097/SIH.0000000000000747] [Medline: 37747487]
25. Luo X, Tham YC, Giuffrè M, Ranisch R, Daher M, Lam K, GAMER Working Group. Reporting guideline for the use of generative artificial intelligence tools in medical research: the GAMER statement. BMJ Evid Based Med 2025;30(6):390-400 [FREE Full text] [doi: 10.1136/bmjebm-2025-113825] [Medline: 40360239]
26. Reporting guidelines. EQUATOR Network. 2025. URL: https://www.equator-network.org/reporting-guidelines/ [accessed 2026-01-05]
27. Zheng L, Sheng Y, Chiang WL. Leaderboard overview. LMArena. 2024. URL: https://lmarena.ai/leaderboard [accessed 2025-08-13]
28. Ray PP. ChatGPT: a comprehensive review on background, applications, key challenges, bias, ethics, limitations and future scope. Internet Things Cyber-Phys Syst 2023;3:121-154. [doi: 10.1016/j.iotcps.2023.04.003]
29. Introducing Claude. Anthropic. 2023. URL: https://www.anthropic.com/news/introducing-claude [accessed 2025-08-13]
30. What is Claude AI? Claude AI. 2024. URL: https://www.blockchain-council.org/ai/what-is-claude-ai/ [accessed 2025-08-13]
31. An overview of the Gemini app. Gemini. 2024. URL: https://gemini.google/overview-gemini-app.pdf [accessed 2025-08-13]
32. Perera PL, Lankathilake M. Preparing to revolutionize education with the multi-model GenAI tool google gemini? A journey towards effective policy making. J Adv Educ Philos 2023;7(08):246-253. [doi: 10.36348/jaep.2023.v07i08.001]
33. Glover E. Grok: what we know about Elon Musk's AI chatbot. BuiltIn. 2024. URL: https://builtin.com/articles/grok [accessed 2025-08-13]

XSL•FO
RenderX

34. Spataro J. Introducing Microsoft 365 Copilot – your copilot for work. Official Microsoft Blog. 2023. URL: https://blogs.microsoft.com/blog/2023/03/16/introducing-microsoft-365-copilot-your-copilot-for-work/ [accessed 2025-08-13]
35. Hernandez J, Fralliciardi A, Nadir N, Gothard MD, Ahmed RA. Development of a Simulation Scenario Evaluation Tool (SSET): modified delphi study. BMJ Simul Technol Enhanc Learn 2020;6(6):344-350 [FREE Full text] [doi: 10.1136/bmjstel-2019-000521] [Medline: 35515493]
36. Mondal H, Mondal S. Artificial intelligence-generated content needs a human oversight. Indian J Dermatol 2024;69(3):284. [doi: 10.4103/ijd.ijd_163_24] [Medline: 39119306]
37. Fang X, Che S, Mao M, Zhang H, Zhao M, Zhao X. Bias of AI-generated content: an examination of news produced by large language models. Sci Rep 2024;14(1):5224 [FREE Full text] [doi: 10.1038/s41598-024-55686-2] [Medline: 38433238]
38. Meskó B. Prompt engineering as an important emerging skill for medical professionals: tutorial. J Med Internet Res 2023;25:e50638 [FREE Full text] [doi: 10.2196/50638] [Medline: 37792434]
39. Denny P, Leinonen J, Prather J, Luxton-Reilly A, Amarouche T, Becker BA, et al. Prompt problems: a new programming exercise for the generative AI era. 2024 Presented at: SIGCSE 2024: Proceedings of the 55th ACM Technical Symposium on Computer Science Education V. 1; March 07, 2024; Portland, OR p. 296-302 URL: https://doi.org/10.1145/3626252.3630909 [doi: https://doi.org/10.1145/3626252.3630909]

Abbreviations

AI: artificial intelligence
GAMER: Generative Artificial Intelligence Tools in Medical Research
LLM: large language model
SBL: simulation-based learning
SSET: Simulation Scenario Evaluation Tool

Edited by S Brini, T Leung; submitted 27.Aug.2025; peer-reviewed by Y Hu, H Fischle, F Lei; comments to author 27.Oct.2025; revised version received 29.Dec.2025; accepted 29.Dec.2025; published 30.Jan.2026.

Please cite as:
Ruggiano N, Sahoo S, Brashear A, Nwatu U, Brunson A, Noh H, Cole H, McKinney R, Framil Suarez CV, Brown EL, Prevost S Evaluating AI-Generated Geriatric Case Studies for Interprofessional Education: Systematic Analysis Across 5 Platforms JMIR Med Educ 2026;12:e83085
URL: https://mededu.jmir.org/2026/1/e83085
doi:10.2196/83085
PMID:

©Nicole Ruggiano, Sudikshya Sahoo, Ava Brashear, Uche Nwatu, Amie Brunson, Hyunjin Noh, Heather Cole, Robert McKinney, C Victoria Framil Suarez, Ellen L Brown, Suzanne Prevost. Originally published in JMIR Medical Education (https://mededu.jmir.org), 30.Jan.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Education, is properly cited. The complete bibliographic information, a link to the original publication on https://mededu.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Real-World Impact and Educational Effectiveness of an AI-Powered Medical History-Taking System: Retrospective Propensity Score-Matched Cohort Study

Yang Liu[1*], MM; Yiying Zhu[1*], MSc; Weishan Zhang[1*], MSc; Xian Lu[2], BSc; Liping Wu[1], Prof Dr; Minghui Yue[1], MPH; Oudong Xia[3*], Prof Dr Med, MD; Chujun Shi[1*], Prof Dr

[1]Medical Simulation Center, Shantou University Medical College, Shantou, Guangdong, China

[2]Department of Medical Physics and Informatics, Shantou University Medical College, Shantou, Guangdong, China

[3]Shantou University Medical College, Shantou, Guangdong, China

[*]these authors contributed equally

Corresponding Author:
Chujun Shi, Prof Dr
Medical Simulation Center
Shantou University Medical College
No. 22 Xinling Road
Shantou, Guangdong, 515041
China
Phone: 86 754 88900459
Email: cjshi@stu.edu.cn

Abstract

Background: Medical history-taking is a core clinical skill; yet, traditional teaching methods face challenges. We developed an artificial intelligence–powered medical history-taking training and evaluation system (AMTES) and established its technical feasibility as an extracurricular resource. Evidence on whether such tools improve learning outcomes when voluntarily embedded in routine curricula remains limited.

Objective: This study aimed to evaluate the real-world educational effectiveness of AMTES as an opt-in extracurricular tool and examine whether learning gains vary by practice patterns and baseline academic ability.

Methods: We conducted a retrospective cohort study of the 2024-2025 Diagnostics course cohort (N=478) at Shantou University Medical College, China, using total population sampling. Students were categorized as AMTES users (n=205, 42.9%; ≥1 sessions) and nonusers (n=273, 57.1%) based on their voluntary extracurricular adoption of the system during the month preceding a high-stakes final practical skills examination. To address selection bias, we performed 1:1 propensity score matching via logistic regression using age, sex, and 3 previous academic scores as covariates. The average treatment effect on the treated for final examination score (0-70) was estimated with paired *t* tests, and robustness to unobserved confounding was assessed via Rosenbaum sensitivity analysis. Among matched users, practice patterns were identified using K-means clustering on log-derived features, with cluster differences compared using Mann-Whitney *U* tests. Subsequently, we explored aptitude-treatment interaction by testing the interaction between practice intensity and baseline ability using linear and logistic regression models.

Results: Propensity score matching yielded 157 matched pairs (n=314) with excellent covariate balance (|standardized mean difference|<0.1). In the matched cohort, the users outperformed nonusers by 3% (average treatment effect on the treated=2.09, 95% CI 0.75-3.42; *P*=.002). This finding was robust to weak unmeasured confounding (Rosenbaum Γ=1.23). Among users (N=157), cluster analysis of usage logs revealed a low-intensity group (74/157, 47.1%) and a high-intensity group (83/157, 52.9%). The 2 groups reflected differences in both practice quantity and quality. However, the added efforts did not translate into higher scores (mean difference=1.6 points, 95% CI −0.5 to 3.6) or excellence probability (risk difference=7.7 percentage points, 95% CI −5.0 to 20.5). Exploratory aptitude-treatment interaction analyses suggested ability-dependent effects for excellence rate (β_3=1.461; *P*=.04) and marginally for final score (β_3=2.58; *P*=.07), but not for pass rate (*P*=.94).

Conclusions: Building upon previous technical validation, this study contributes real-world effectiveness evidence by evaluating AMTES as a voluntary extracurricular supplement within an authentic, high-baseline curriculum. Unlike previous work focusing on technical feasibility or short-term controlled trials, voluntary extracurricular AMTES use was associated with modest yet

XSL•FO
RenderX

meaningful improvements in summative history-taking performance. Exploratory analyses indicated that the added value of more intensive engagement may be moderated by baseline academic ability. These findings support the scalability of artificial intelligence–enabled supplementary training and inform precision-oriented instructional design.

(JMIR Med Educ 2026;12:e89367) doi:10.2196/89367

KEYWORDS

AI-powered medical history-taking training and evaluation system; medical history-taking; virtual patient; educational effectiveness; retrospective cohort study; propensity score matching; average treatment effect on the treated; self-regulated learning; aptitude-treatment interaction; precision education

Introduction

Medical history-taking is the cornerstone of clinical diagnosis and decision-making. Effective training in this skill is essential for improving diagnostic accuracy [1-3] and is a necessary prerequisite for cultivating competent physicians [4]. Consequently, how to effectively cultivate this practical skill remains a core challenge for medical educators [5]. According to Kolb's experiential learning cycle, skill acquisition requires learners to engage in specific, feedback-driven practice activities to internalize knowledge and effectively develop skills [6]. However, medical education primarily relies on standardized patients (SPs) for this training. The high costs and limited resources associated with SPs [7] restrict the capacity to meet the demands for large-scale, high-frequency, and flexible practice opportunities required for "deliberate practice."

With advancements in digital technologies, particularly revolutionary breakthroughs in large language models (LLMs), LLM-driven virtual patients (VPs) are viewed as a potential innovative solution to resource shortages [8-10]. VPs can function as "interactive practice objects," allowing students to repeatedly drill scenarios and receive contextualized feedback in a safe, on-demand environment [11,12]. However, the direct application of general-purpose LLMs in medical education still faces technical limitations—such as "hallucinations" [13], reliability issues, and inadequate long-context processing capabilities—thereby undermining pedagogical accuracy and trustworthiness. Therefore, how to effectively and equitably integrate customized AI technology into medical education remains a monumental challenge [14].

To address these challenges and ensure pedagogical reliability, our research team developed the artificial intelligence–powered medical history-taking training and evaluation system (AMTES) [15]. This is a custom-designed educational platform powered by the DeepSeek LLM and intended to supplement existing teaching within a blended learning framework. The system features a multiturn dialogue module that simulates realistic, emotionally expressive patient encounters and integrates an automated assessment and feedback system based on clinical expert standards. Notably, our previous prospective multicase validation study established a "three-element" foundation for AMTES as a mature educational tool, that is (1) technical reliability (dialogue accuracy >97%; automated scoring consistency with human experts, intraclass correlation coefficient>0.92), (2) pedagogical appropriateness (training cases meticulously designed and difficulty-validated), and (3) student acceptance (87% found it helpful) [15]. This validation confirmed AMTES's transition from being "technically functional" to "pedagogically usable."

AMTES has been deployed as a long-term auxiliary resource for extracurricular autonomous learning in the Diagnostics course. Despite the increasing prevalence of artificial intelligence (AI) tools [12,16,17], existing research predominantly focuses on technical feasibility, dialogue quality, and assessment validity [18-20], as well as learners' subjective experiences [21]. Consequently, high-quality evidence regarding the objective, long-term educational outcomes of AI history-taking tools in authentic, ecological educational settings remains scarce. A few studies attempting to evaluate educational effectiveness [22,23] typically use small-sample controlled trials with mandatory, prescribed-dose, short-term training. These highly controlled designs lack ecological validity, failing to reveal students' authentic self-regulated learning (SRL) patterns during routine, voluntary use, or to reflect the long-term impact of these behaviors on subsequent summative examinations.

Therefore, it is important to focus on SRL as a core mechanism, as it has been widely recognized as a key determinant of academic achievement and clinical competence in medical education [24]. However, most quantitative SRL studies still rely on self-report questionnaires (eg, the Motivated Strategies for Learning Questionnaire), which are vulnerable to social desirability and recall bias, and can only provide static, single-point-in-time snapshots. Yet, SRL is fundamentally a dynamic, cyclical process involving forethought, performance control, and self-reflection surrounding a specific task [25]. Many such instruments were initially developed for traditional classroom settings and may not adequately capture how learners self-regulate in blended or technology-enhanced environments [26]. In light of this, recent research increasingly advocates for the use of process data and event-based measurements [25], such as log data from intelligent tutoring systems and AI-supported platforms [27], to obtain real-time, fine-grained, and objective evidence of how learners implement SRL strategies in authentic learning contexts.

Crucially, research exploring whether students with different ability levels or practice behavior patterns derive varying degrees of benefit from AI tools remains extremely limited. Educational psychology frameworks, such as the aptitude-treatment interaction (ATI) theory [28], emphasize that the impact of instructional interventions is often not uniform but varies systematically with learner characteristics [29,30]. This critical gap highlights the urgent need to investigate the heterogeneity of effects in AI-supported history-taking training across different learner groups.

XSL•FO
RenderX

This study is uniquely situated within a high-baseline training context, where participants had been receiving high-quality SP training before the introduction of AMTES. Therefore, our primary objective is to determine the marginal contribution of AMTES to existing instruction. To address the endogeneity issues inherent in such opt-in, extracurricular tools (eg, omitted variable bias) [31,32], this study used a retrospective cohort design combined with propensity score matching (PSM) to provide robust and impartial causal estimates.

Building on the successful validation of AMTES and addressing the gaps in current research, this study aims to systematically address three incremental core questions, moving from average effects to process mechanisms and heterogeneity:

- First, within the context of existing SP-assisted instruction, does the voluntary, supplementary use of AMTES yield a significant average treatment effect on the treated (ATT) on students' practical skills?
- Second, among AMTES users, do different self-regulated practice behavior patterns (eg, low-intensity vs high-intensity) lead to different academic outcomes?
- Finally, does the effectiveness of different practice behavior patterns vary according to students' initial aptitude, suggesting the presence of an ATI effect?

By systematically investigating these questions, this study aims to provide causal evidence for the real-world integration of customized medical LLMs, offer practical insights for instructional design, and illuminate the path toward personalized and precision medical education.

Methods

Study Design, Setting, and Participants

We conducted a retrospective cohort study to evaluate the real-world impact of the AMTES on student performance in an undergraduate Diagnostics course at Shantou University Medical College in the 2024-2025 academic year. This study is reported in accordance with the Journal Article Reporting Standards for quantitative research [33]. Total population sampling was used. All undergraduate medical students enrolled in the course were included (280/478, 58.6% male and 198/478, 41.4% female; mean age 20.38, SD 0.72 years). Because this was a complete-course cohort and the exposure was determined retrospectively, no a priori sample size calculation was performed.

All students, regardless of subsequent AMTES use, received the standard curriculum prescribed by the medical school, which comprised regular didactic lectures and in-person practical training sessions with SPs accompanied by individualized feedback from clinical instructors. This SP-based training constituted the primary instructional foundation for the acquisition of history-taking skills for all students.

One month before the final practical skills examination (held in early July 2025), AMTES was introduced as a voluntary extracurricular supplement. Students were free to decide whether and how often to use the system for additional practice during this 1-month period. Based on their actual usage, we retrospectively categorized students into 2 cohorts—an AMTES user group (≥1 training sessions) and a nonuser group (0 sessions).

After the course concluded, AMTES system usage logs were aggregated and linked with students' course examination scores. This design enabled comparison of outcomes between AMTES users and nonusers and allowed us to estimate the marginal educational value of AMTES on top of an existing high-quality curriculum.

Procedures

System Implementation

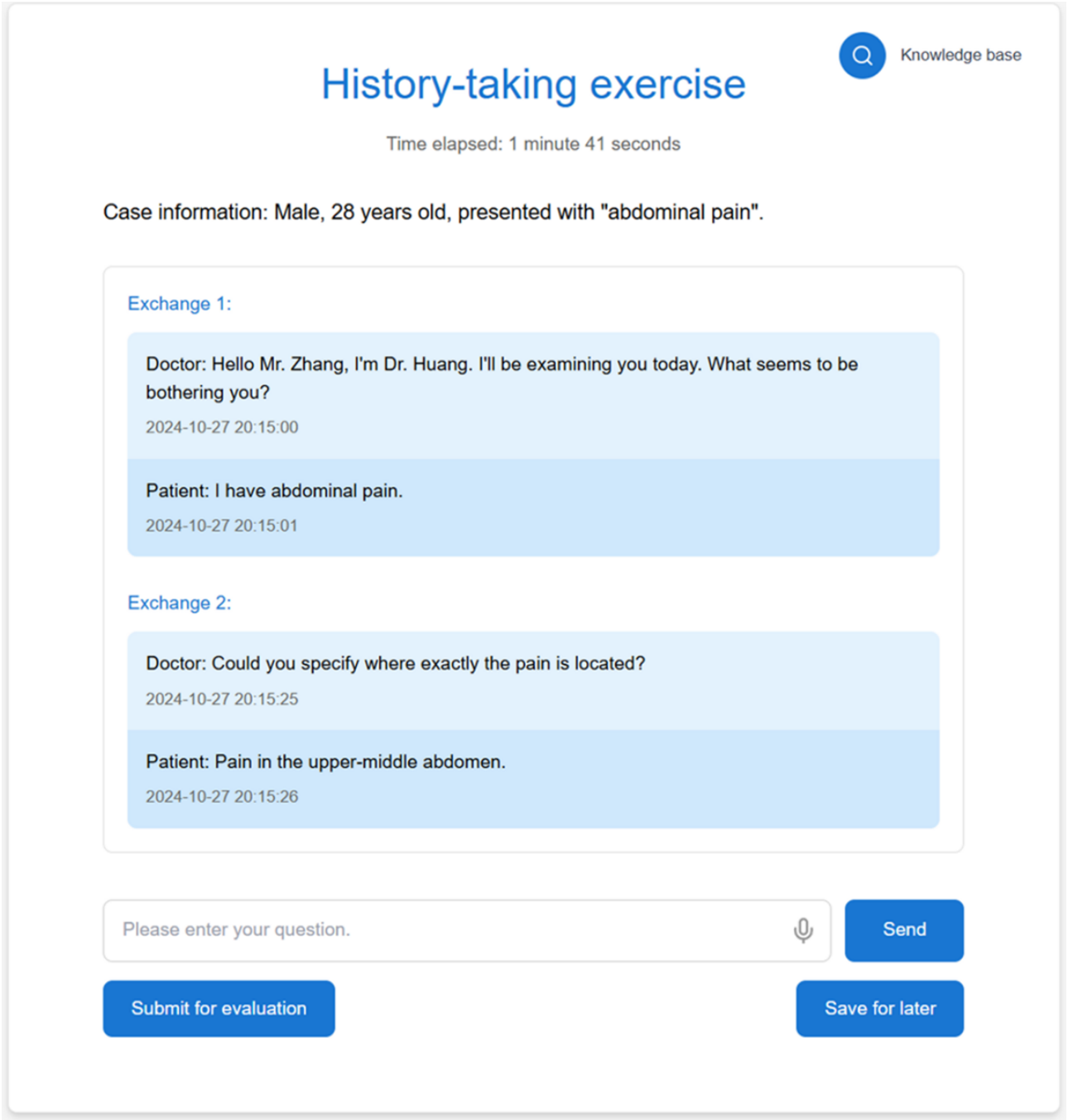
During the month preceding the final practical skills examination, AMTES was made available to all students in the Diagnostics course as a voluntary extracurricular resource. The system provides a web-based interface in which students conduct multiturn consultations with a VP using either typed or voice input (Figure 1). During each simulated consultation, AMTES generated patient responses in real time and recorded the full question-answer dialogue.

At the end of each session, students received an immediate structured feedback report summarizing their overall performance, subscores across key skill dimensions, missed critical items, and suggestions for improvement (Figure 2). These automated reports were intended to emulate an instructor's critique and provide timely guidance for self-reflection and adjustment.

AMTES offered 3 validated training cases of increasing difficulty, covering different chief complaints and organ systems, that are cough and sputum (simple), abdominal pain (moderate), and urinary frequency (complex). The underlying diagnoses and core symptoms of these training cases were designed not to overlap with those of the final practical skills examination case, thereby reducing the risk of direct teaching to the test. Students could use AMTES at any time and as often as they wished during the 1-month period. In practice, 205 (42.9%) students completed at least 1 AMTES session (user group), whereas 273 (57.1%) did not (nonuser group).

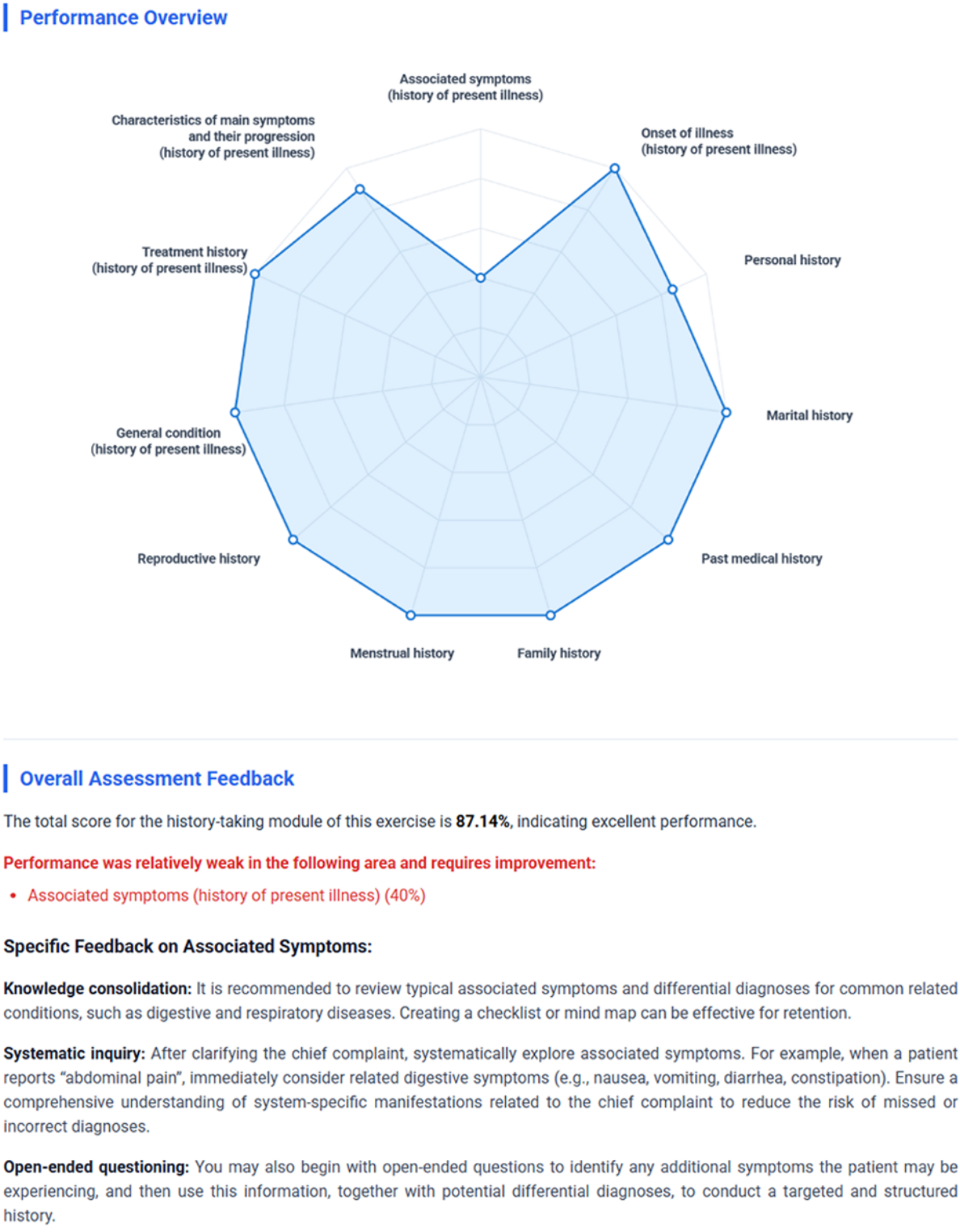
XSL•FO
RenderX

Figure 1. Screenshot of history-taking interface where students could conduct multiturn consultations.



XSL•FO
RenderX

Figure 2. Example of an automated assessment feedback report from the artificial intelligence–powered medical history-taking training and evaluation system.



Final Practical Skills Examination and Scoring

The summative assessment consisted of a 20-minute one-on-one history-taking consultation with an SP. An assessor-blinded design was implemented to ensure objectivity, whereby examiners were blinded to students' AMTES usage status. Furthermore, given the retrospective nature of this study, students were unaware at the time of examination that their AMTES usage data would be analyzed in subsequent research, thereby minimizing potential behavioral modifications due to research awareness (Hawthorne effect). All participating SPs underwent rigorous pre-examination training on the case script and performance baseline standards to ensure consistency and standardization. Examiners also received standardized training to ensure uniform application of the scoring criteria.

XSL•FO
RenderX

The scoring rubric was developed with reference to the National Clinical Practitioner Licensing Examination standards of China, with a maximum score of 70 points. The rubric covered 6 key domains—chief complaint, history of present illness, past medical history, personal history, marital and reproductive history, and family history. Points were allocated according to clinical importance, with each scoring item corresponding to a single, clearly defined evaluation criterion. To maximize objectivity and minimize examiner subjectivity, a dichotomous scoring method was used: students received full points for an item if the required information was successfully elicited, and zero points otherwise.

According to institutional academic regulations, performance grades were classified as excellent (≥85%), pass (60%-84%), and fail (<60%). Since the history-taking component in this study had a maximum raw score of 70 points, we applied an equivalence ratio of 0.7 to convert these institutional thresholds. Specifically, we defined ≥42 points (ie, 60% of the maximum score) as the "pass" threshold and ≥59.5 points (ie, 85% of the maximum score) as the "excellence" threshold.

Data Collection, Measures, and Covariates

Overview

All data used in this study were obtained from anonymized AMTES usage logs and official academic records, capturing both students' actual learning behaviors and their performance outcomes in an authentic educational context. We considered four types of variables, that is, baseline covariates, intervention variables, outcome variables, and practice behavior logs.

Baseline Covariates

We selected 5 baseline covariates as predictors for the propensity score model. These included demographic characteristics (age and sex) and 3 academic performance metrics obtained before the AMTES period—the current course midterm examination score, previous semester theory examination score, and previous semester practical skills examination score, representing students' baseline academic aptitude. These academic metrics were chosen because previous performance is a robust proxy for underlying attributes, such as motivation, self-discipline, and previous knowledge [34,35]. Additionally, age and sex were incorporated as key individual-difference variables highlighted in the unified theory of acceptance and use of technology 2 framework, which identifies them as influential factors in technology adoption and usage behavior [36].

Intervention Variable

A binary indicator of AMTES use during the 1-month window, defining a user group (≥1 training sessions) and a nonuser group (0 sessions).

Outcome Variables

The primary outcome was the total score (0-70 points) on the final medical history-taking skills examination. Secondary outcomes were 2 dichotomous indicators derived from this score, that is, excellence (score ≥59.5) and pass (score ≥42), based on institutional grading standards.

Practice Behavior Logs

For AMTES users, we extracted behavioral metrics from the system logs, including the total number of sessions, the number of unique cases practiced, session-level AMTES scores, and dialogue turns. These log-based indicators were used as behavioral proxies of students' SRL processes when engaging with AMTES.

Statistical Analysis

Overview

All statistical analyses were conducted using *R* (version 4.2.0; *R* Foundation for Statistical Computing) and Python (version 3.10; Python Software Foundation). A 2-sided significance level of α=.05 was used. Since all study data, including digital trace logs and academic records, were automatically captured and integrated, no missing data were identified for the baseline covariates or outcome variables analyzed in this cohort. Continuous variables were summarized as mean (SD) or median (IQR), and categorical variables as counts and percentages.

PSM Analysis

To mitigate selection bias arising from voluntary AMTES use, we applied PSM to construct comparable cohorts of users and nonusers. Propensity scores were estimated using a logistic regression model with AMTES use (yes or no) as the dependent variable, conditioned on the 5 aforementioned baseline covariates. We performed 1:1 nearest-neighbor matching without replacement, using a caliper of 0.2 times the SD of the logit of the propensity score. Additionally, exact matching on sex was imposed to ensure sex comparability within each matched pair. Covariate balance was assessed using standardized mean difference (SMD), with |SMD|<0.1 indicating adequate balance [37]. A Love plot was generated to visually inspect covariate balance.

Within the matched sample, we first examined the normality of within-pair differences in final practical skills examination scores using the Shapiro-Wilk test, which did not indicate substantial deviations from normality (Multimedia Appendix 1). On this basis, we estimated the ATT by comparing final practical skills examination scores between AMTES users and matched nonusers using paired *t* tests, reporting mean differences with 95% CIs. To assess the robustness of the findings to potential unmeasured confounding, we conducted a Rosenbaum sensitivity analysis [38] and reported the Γ value at which the ATT would no longer be statistically significant.

Practice Behavior Pattern Clustering

To identify distinct SRL-related practice patterns among AMTES users in the matched cohort (n=157), we performed an unsupervised clustering analysis based on log-derived features. From the practice logs, we initially extracted 11 candidate features across 4 conceptual dimensions—engagement and coverage, practice strategy, platform performance, and interaction dynamics. Before clustering, the positively skewed metric (total_sessions) was log-transformed, and all continuous variables were centered and scaled. To reduce redundancy, we examined pairwise Spearman rank correlations among candidate features. For highly correlated pairs ($|\rho|\geq 0.80$), we retained the

XSL•FO
RenderX

feature with greater interpretability and conceptual relevance. Variance inflation factors computed on the reduced feature set confirmed acceptable multicollinearity (all variance inflation factors<5). The final clustering used 5 retained features (Table 1). Full details of feature construction and selection appear in Multimedia Appendix 2.

We used the K-means algorithm to cluster students based on these features. The optimal number of clusters was determined to be 2 (K=2) using the silhouette coefficients. To ensure stability, K-means was run with multiple random initializations, which yielded consistent cluster assignments (Multimedia Appendix 3). Each student was subsequently assigned to 1 of 2 clusters for subsequent comparisons. Given the nonnormal distribution of log-derived performance metrics, we used the Mann-Whitney *U* test (2-sided) to compare cluster characteristics.

Table 1. Log-derived features used for clustering analysis among matched artificial intelligence–powered medical history-taking training and evaluation system users (n=157).

Feature[a]	Interpretation
Sessions_total[b]	Total number of AMTES[c] sessions completed
Unique_cases	Number of distinct cases practiced
Avg_score	Mean session score across all sessions
Avg_turns	Average number of dialogue turns per session
Turns_per_min	Overall dialogue turns per minute

[a]All features were standardized (*z* scores) before clustering.

[b]Skewed variable was transformed using a log1p transformation (log[x+1]).

[c]AMTES: artificial intelligence–powered medical history-taking training and evaluation system.

Exploratory Analysis of Heterogeneous Effects

To address our third research question, we conducted an exploratory analysis grounded in the ATI framework, examining whether the effectiveness of different AMTES practice patterns varied by baseline academic ability. Within the AMTES user group, we constructed a standardized composite baseline ability score by averaging *z* scores of midterm, previous theory, and previous practical skills examination scores. We then fitted 3 regression models with identical predictors—cluster assignment (practice pattern), the composite baseline ability score, and their interaction term. A linear regression model was used for the continuous final examination score, and 2 logistic regression models were used for the binary outcomes of pass (score ≥42 vs <42) and excellence (score ≥59.5 vs <59.5).

Interaction terms were tested using Wald tests. To aid interpretation and generate hypotheses for future research, we computed and plotted marginal effects of being in the high-intensity versus low-intensity group at the 25th, 50th, and 75th percentiles of the baseline ability score. For the linear model, marginal effects were expressed as differences in predicted final practical skills examination scores; for the logistic models, they were expressed as risk differences in pass and excellence probabilities, each with 95% CIs.

Ethical Considerations

This study received ethical approval from the Ethics Committee of Shantou University Medical College (approval SUMC-2025-028). All procedures were conducted in accordance with the principles of the Declaration of Helsinki and complied with relevant Chinese laws and institutional ethical standards. Before participation, all participants provided written informed consent. To ensure the privacy and confidentiality of participants, all data were anonymized before analysis and used solely for academic research purposes. No financial compensation was provided to participants for their involvement in this study. All images presented in this paper contain only simulated demonstration data and do not display any identifiable information of actual participants. No identifiable images of individual participants appear in the manuscript or supplementary materials.

Results

ATT of AMTES on Final Practical Skills Examination Scores

Baseline Characteristics

Overall, 478 undergraduate students were included in the initial cohort, comprising 205 AMTES users and 273 nonusers (Figure 3). Before matching, significant selection bias was evident. The user group demonstrated a stronger academic baseline (absolute SMD far exceeding the |0.1| threshold for all academic covariates) and differed in gender distribution (|SMD|=0.21) compared with the nonuser group. The propensity score model showed moderate discriminatory ability (area under the receiver operating characteristic curve=0.690; Multimedia Appendix 4).

After 1:1 nearest-neighbor matching was performed, 157 matched pairs (n=314) were retained (Figure 3). All covariates achieved excellent balance (|SMD|<0.1), and the PS model's area under the receiver operating characteristic curve was correspondingly reduced to a chance level of 0.521 (Figure 4; Table 2; Multimedia Appendix 4). These results indicated that PSM successfully controlled for observable confounding variables, creating comparable cohorts for evaluating the effect of AMTES.

XSL•FO
RenderX

Figure 3. Study flow diagram of participant selection and propensity score matching. AMTES: artificial intelligence–powered medical history-taking training and evaluation system; ATT: average treatment effect on the treated.

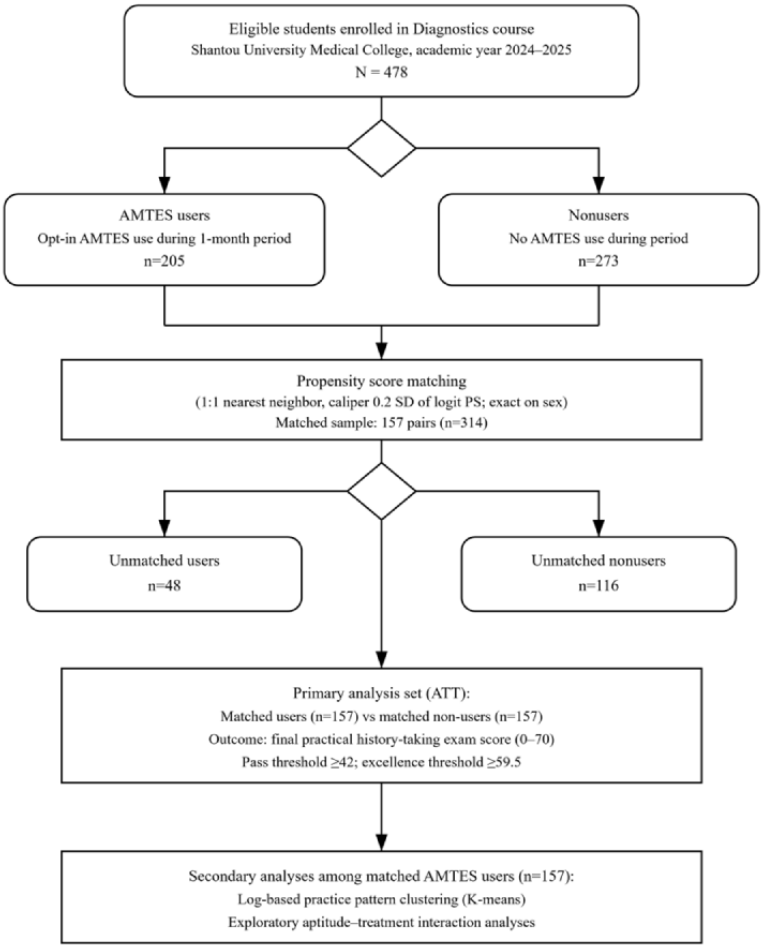
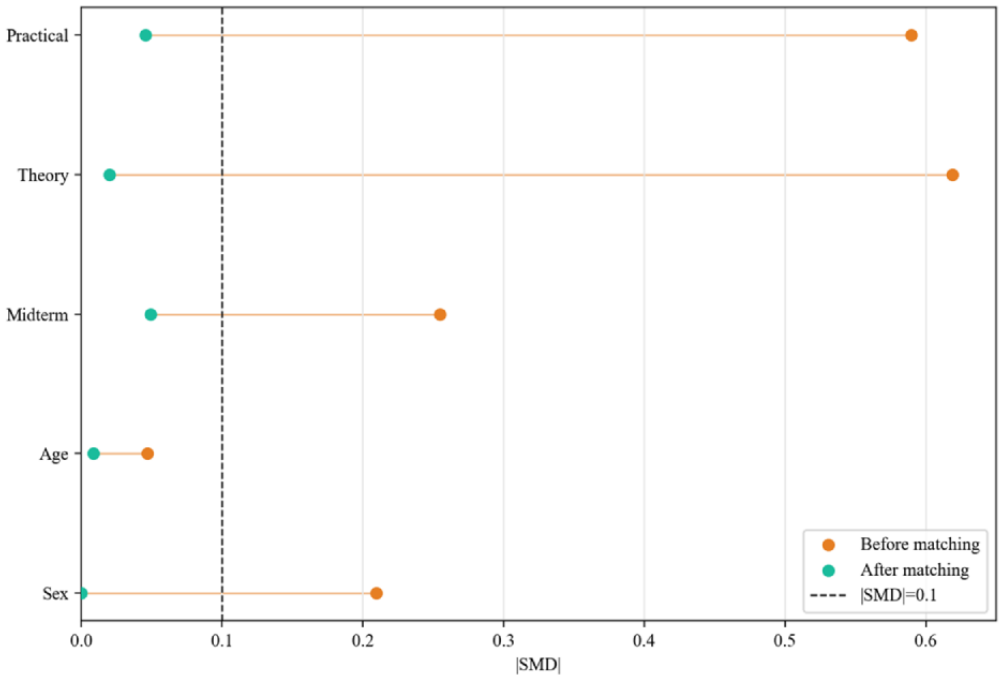


Figure 4. Love Plot of baseline covariates before and after propensity score matching (N=478). SMD: standardized mean difference.



XSL•FO
RenderX

Table 2. Comparison of covariate balance metrics before and after matching (N=478).

Metric	Before matching				After matching			
	Treated	Control	SMD[a,b]	AUROC[c]	Treated	Control	SMD[b]	AUROC
Total (N=478), n (%)	205 (42.9)	273 (57.1)	—[d]	—	157 (76.6)[e]	157 (57.5)[f]	—	—
Sex, n (%)[g]			0.21	—			0	—
Male	108 (52.7)	172 (63)			92 (58.6)	92 (58.6)		
Female	97 (47.3)	101 (37)			65 (41.4)	65 (41.4)		
Age (y), mean (SD)	20.380 (0.722)	20.414 (0.702)	0.047	—	20.408 (0.742)	20.401 (0.724)	0.009	—
Midterm examination, mean (SD)	26.361 (2.652)	25.571 (3.491)	0.255	—	26.121 (2.809)	26.248 (2.369)	0.049	—
Theory examination, mean (SD)	72.185 (11.472)	64.385 (13.636)	0.619	—	68.873 (10.906)	69.096 (11.345)	0.02	—
Practical skills examination, mean (SD)	76.487 (9.886)	70.083 (11.752)	0.590	—	73.604 (9.2)	74.036 (9.762)	0.045	—
Final examination, mean (SD)	54.311 (6.427)	50.738 (7.239)	—	—	53.788 (6.591)	51.701 (6.577)	—	—
Overall	—	—	—	0.69	—	—	—	0.521

[a]SMD: standardized mean difference.

[b]Covariates were considered well-balanced if the absolute standardized mean difference was <0.1.

[c]AUROC: area under the receiver operating characteristic curve.

[d]Not applicable.

[e]Retention was 76.6% for the treated group.

[f]Retention was 57.5% for the control group.

[g]For propensity score estimation, sex was coded as male=1 and female=0.

Final Practical Skills Examination Scores

Analysis of the 157 matched pairs revealed a significant, positive effect of AMTES on final practical skills examination scores. The user group achieved a higher mean score than the nonuser group (mean 53.788, SD 6.591 vs mean 51.701, SD 6.577). The effect was 2.09 points (95% CI 0.75-3.42; *P*=.002). Given a maximum score of 70, this was equivalent to an increase of approximately 3% of the total score, suggesting a practically meaningful impact.

A Rosenbaum sensitivity analysis showed that this conclusion was robust to unobserved confounding of weak strength. An unobserved confounder would need to increase the odds of a student using AMTES by a factor of approximately 1.23 to render the difference nonsignificant (*P*>.05).

Practice Behavior Clustering Results

K-means clustering (K=2) identified 2 distinct practice patterns among the 157 AMTES users based on a multidimensional construct of usage intensity, which incorporated practice frequency, case coverage, average session score, and interaction intensity in the matched cohort (Table 3). Cluster 1 (n=74, 47.1%) was characterized by lower practice frequency and narrower case coverage, referred to as a low-intensity group. In contrast, cluster 2 (n=83, 52.9%) demonstrated substantially higher activity across all dimensions—practice frequency, case coverage, and interaction intensity (all *P*<.001; Table 3), referred to as a high-intensity group. Average score differences were smaller but significant (*P*=.04). These patterns reflected differences in both practice quantity and quality.

While the 2 user groups demonstrated significantly different practice behaviors, the added effort did not yield a significant performance advantage. Both groups achieved high pass rates approaching 100% (high-intensity 94% vs low-intensity 97.3%), leaving little room for improvement at this threshold. The potential benefits of higher practice intensity were more evident in the excellence rate (25.3% vs 17.6%) and mean score (54.52, SD 6.4 vs 52.97, SD 6.7; Cohen *d*=0.24), although these differences remained nonsignificant (Table 4).

XSL•FO
RenderX

Table 3. Practice behavior characteristics of the two practice intensity clusters among matched artificial intelligence–powered medical history-taking training and evaluation system users (n=157).

Performance feature	Low-intensity group, mean (SD)	High-intensity group, mean (SD)	*P* value
Sessions_total	1.243 (0.463)	3.205 (1.166)	<.001
Unique_cases	1.135 (0.344)	2.590 (0.495)	<.001
Avg_score	43.677 (10.752)	47.603 (5.351)	.04
Avg_turns	67.029 (28.849)	89.379 (23.240)	<.001
Turns_per_min	2.684 (0.836)	3.462 (0.872)	<.001

Table 4. Comparison of final practical skills examination scores between high- and low-intensity practice groups among matched artificial intelligence–powered medical history-taking training and evaluation system users (n=157).

Outcome metric	Low-intensity cluster (n=74)	High-intensity cluster (n=83)	Risk difference, percentage points (95% CI)	Relative risk (95% CI)	Mean difference (95% CI)	Cohen *d* (95% CI)
Excellence rate (≥59.5)	n=13 (17.6%)	n=21 (25.3%)	7.7 (−5.0 to 20.5)	1.44 (0.78 to 2.67)	—[a]	—
Pass rate (≥42)	n=72 (97.3%)	n=78 (94%)	−3.3 (−9.6 to 3.0)	0.97 (0.90 to 1.03)	—	—
Total score (0-70)	Mean 52.97 (SD 6.7)	Mean 54.52 (SD 6.4)	—	—	1.6 (−0.5 to 3.6)	0.24 (−0.08 to 0.55)

[a]Not applicable.

Exploratory Analysis of Heterogeneous Effects: Moderating Role of Baseline Ability

Regression models examining ATI suggested that practice intensity effects may vary by baseline ability for excellence rate (β_3=1.461; *P*=.04) and final score (β_3=2.58; *P*=.07), but not for pass rate (*P*=.94).

Marginal effects suggested that higher-intensity practice was associated with larger benefits among students with higher baseline ability (Q3), including higher final examination scores (marginal effect +2.88 points, 95% CI 0.39-5.37; *P*=.02) and a higher probability of excellence (marginal effect +18.93 percentage points, 95% CI 1.32-36.54; *P*=.03; Table 5; Figure 5; Figure 6). At lower baseline ability quantiles (Q1-Q2), marginal effects were not statistically significant for any outcome. Pass rate showed no differences across quantiles, with both clusters maintaining high performance (>94%; Table 5; Figure 6). In summary, these exploratory findings suggested that higher practice intensity may be more beneficial for academically stronger students.

Table 5. Marginal effects of high vs low practice intensity by baseline ability quantile among matched artificial intelligence–powered medical history-taking training and evaluation system users (n=157).

Outcome and baseline quantile[a]	Baseline *z* score	Marginal effect (95% CI)	*P* value
Final score (points)			
Q1	−0.45	0.39 (−1.98 to 2.75)	.75
Q2	0.07	1.72 (−0.29 to 3.73)	.09
Q3	0.52	2.88 (0.39 to 5.37)	.02
Excellence rate (percentage points)			
Q1	−0.45	−5.32 (−19.04 to 8.40)	.45
Q2	0.07	4.79 (−8.62 to 18.21)	.48
Q3	0.52	18.93 (1.32 to 36.54)	.04
Pass rate (percentage points)			
Q1	−0.45	−4.4 (−12.21 to 3.41)	.27
Q2	0.07	−2.58 (−8.50 to 3.35)	.39
Q3	0.52	−1.59 (−6.72 to 3.53)	.54

[a]Q1, Q2, and Q3 represent baseline ability at 25th, 50th, and 75th percentiles, respectively.

XSL•FO
RenderX

Figure 5. Predicted final examination scores (0-70) by practice intensity and baseline ability index among matched artificial intelligence–powered medical history-taking training and evaluation system users (n=157). Baseline ability was defined as a composite z score: baseline_z=mean[z(practical), z(theory), z(midterm)]. For interpretability, baseline_z was mapped to an "Avg score" scale: Avg_score=mean(raw_avg)+baseline_z×SD(raw_avg), where raw_avg=(practical+theory+midterm)/3; quartile reference points: Q1=−0.45 (Avg_score=53.58 approximately), Q2=0.07 (56.62 approximately), Q3=0.52 (59.26 approximately); Rug marks show baseline ability distribution by cluster. avg: average.

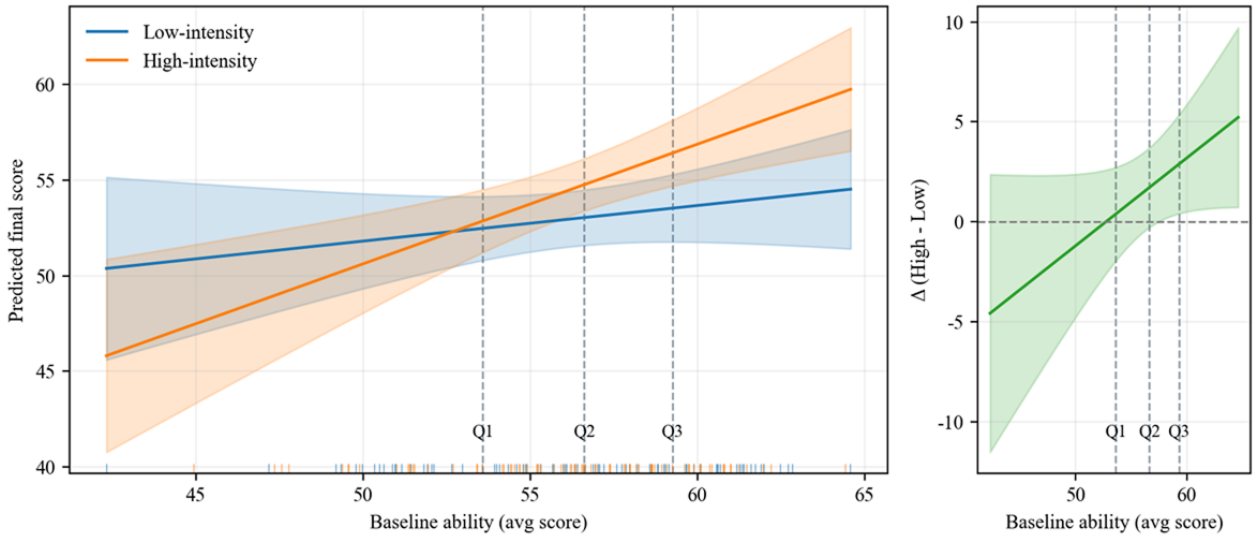
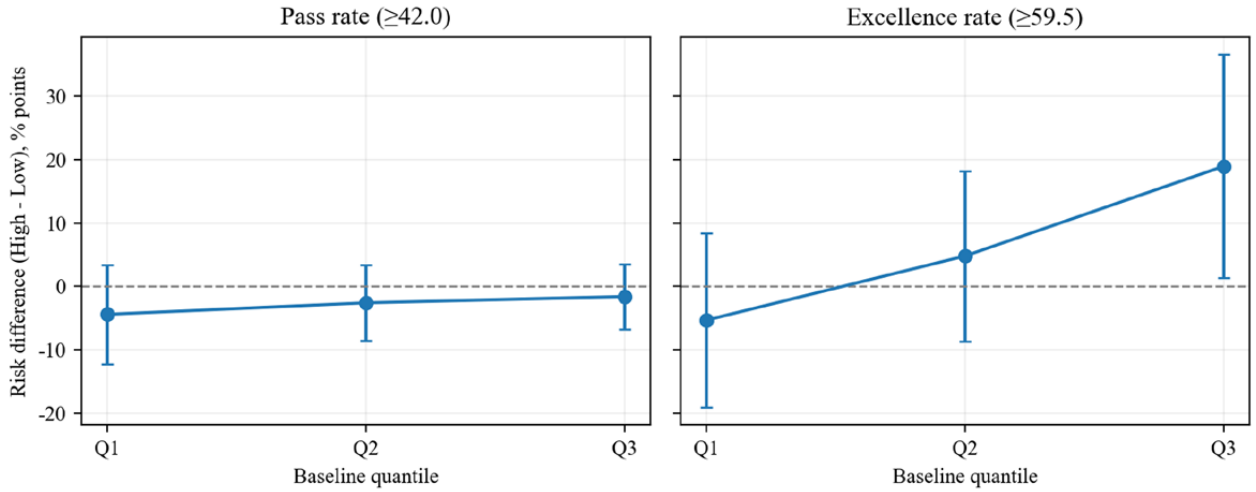


Figure 6. Risk difference by baseline ability quantile: pass (≥42) vs excellence (≥59.5). For pass rate, all point estimates were negative with CIs crossing zero, indicating no clear differences across baseline ability. For excellence rate, the Q1 estimate was negative and imprecise (CI crossed zero), whereas the Q3 estimate was positive and significant (CI excluded zero); sample sizes by baseline ability quartile were (1) for Q1: n=40 (low-intensity=21, high-intensity=19), (2) for Q2: n=78 (low-intensity=33, high-intensity=45), and (3) for Q3: n=39 (low-intensity=20, high-intensity=19).



Discussion

Principal Findings

This retrospective cohort study provides real-world evidence on the educational effectiveness of AI-supported medical history-taking training in routine curricula and offers insights relevant to instructional design and precision-oriented medical education. In propensity score–matched analyses, voluntary extracurricular AMTES use was associated with significantly higher scores on the final high-stakes history-taking examination compared with matched nonusers. Among users, digital trace-based clustering identified distinct low- and high-intensity practice patterns; however, more intensive engagement was not consistently associated with additional performance gains. Exploratory ATI analyses further suggested that baseline academic performance may moderate the incremental benefit of intensive engagement, indicating potential heterogeneity in AI-supported skills training.

Bridging the Gap: Real-World Effectiveness and Practical Value in a High-Baseline Context

This study's primary contribution is bridging the critical gap between the validated technical feasibility of an AI educational tool and its real-world educational effectiveness. Building upon our previous work that established the AMTES as a technically reliable system [15], this study addressed a more consequential question, that is, does voluntary use of this tool translate into tangible learning gains in an authentic educational context?

XSL•FO
RenderX

Through rigorous PSM, which aligns with recommended practices for causal inference in education [39,40], we observed a statistically and educationally meaningful improvement in final practical skills examination scores among AMTES users compared with matched nonusers. Sensitivity analyses suggested that this finding was robust to unobserved confounding of weak strength.

To our knowledge, this study is among the first to provide empirical evidence that AI-driven history-taking training can yield measurable improvements in high-stakes examination outcomes. Previous studies in this domain have largely focused on system feasibility [20,41]. Although few studies have begun to examine educational effectiveness, such evaluations have typically been conducted under tightly controlled experimental conditions. For instance, in a recent randomized controlled trial, Wang et al [23] compared GPT-based VPs with traditional teacher-facilitated role-playing and reported substantial improvements in structured history-taking scores (Cohen d=2.6 approximately) [23]. Our study extends this line of research by demonstrating the effectiveness of AMTES when voluntarily integrated into the routine curriculum, thereby achieving high ecological validity. Moreover, the training cases were intentionally designed to differ from the final examination case in both organ system and core symptoms, ruling out rote memorization or a narrow "teaching to the test" as plausible explanations. Instead, the observed gains are most consistent with a genuine enhancement of students' generalizable history-taking competence, a fundamental and transferable clinical skill that supports diagnostic reasoning and doctor-patient communication [5,41].

Crucially, the significance of this finding is amplified by the study's context. Before accessing AMTES, all students had already received a high-quality curriculum that included in-person SP training and direct feedback from clinical instructors. In this setting, demonstrating an incremental benefit of 2.09 points (approximately 3% of the 70-point total) is considerably more challenging than improving performance from a novice baseline. This supports previous findings that AI tools can still provide meaningful gains even within a high-quality, resource-intensive educational framework [27].

In practical terms, this improvement was achieved with favorable cost-effectiveness. AMTES functioned as a low-cost, scalable tool that provided flexible, on-demand practice without increasing faculty workload or consuming scheduled class time [15]. Given the negligible marginal cost of providing additional practice, the effect size represents a substantial educational return on investment. This is consistent with previous work showing that AI-driven virtual SPs can serve as a cost-efficient supplement to traditional SP training [42], reducing personnel and logistical costs while providing scalable and flexible practice opportunities that can be integrated seamlessly into the curriculum.

Unveiling Authentic SRL Through Digital Traces

Our analyses of AMTES usage logs for the second research question revealed 2 distinct practice profiles among users—a low-intensity group and a high-intensity group. Notably, these clusters differed not only in frequency of use but also in case coverage and interaction depth, suggesting combined differences in practice quantity and quality. Although the 2 groups differed in practice behaviors, these pronounced engagement differences did not translate into statistically significant differences in final practical skills examination scores. This finding warrants cautious interpretation, as the absence of statistical significance should not be construed as evidence that practice volume or depth of engagement is unimportant for skill acquisition. Rather, it suggests that the effectiveness of AMTES may not follow a straightforward dose-response relationship within the observed range of practice behaviors in this cohort.

SRL offers a useful framework for interpreting such practice patterns [24,26]. Rather than relying on self-reported measures, which may not capture SRL as it unfolds over time [25], we used objective process data from AMTES to characterize how students actually engaged with history-taking practice. Digital traces, such as practice frequency, temporal distribution, case coverage, repetition rate, and dialogue length, provided behavioral indicators of how learners planned, monitored, and adjusted their use of the system over time, yielding a higher-fidelity picture of their engagement in this educational context. The clustering approach enabled a more fine-grained description of SRL-related engagement patterns than a simple user-nonuser distinction.

The absence of a clear performance advantage for the high-intensity cluster is also consistent with previous work indicating that relationships between SRL and academic achievement are not uniformly strong or statistically significant [24]. Although several studies have reported positive links between SRL and exam performance or clinical skills [43-45], other longitudinal and clinical training studies have reported weak or inconsistent associations between specific SRL components and achievement outcomes, with some strategies showing no significant relationships [25,46]. In our setting, this may help explain why greater AMTES use did not yield markedly higher scores. AMTES provided structured opportunities for reflective, feedback-driven, deliberate practice [15]. Students who actively processed feedback and adapted their questioning strategies were likely able to translate even modest amounts of practice into meaningful gains, whereas undirected, repetitive sessions produced diminishing returns and limited additional benefit. This interpretation aligns with contemporary SRL theory, which emphasizes metacognitive monitoring, strategic planning, and iterative adjustment as core drivers of skill acquisition [26,47]. Additionally, opportunity costs may partly explain this pattern, as repeated practice within a small set of validated cases could displace other high-yield learning activities (eg, advanced case discussions or authentic clinical encounters) [48].

Heterogeneity of Effect: Exploratory ATI Patterns

In authentic, opt-in implementations, average dose-response relationships may appear weak or inconsistent when effects are heterogeneous and potentially nonlinear across learners [30]. Therefore, the nonsignificant difference between practice intensity groups does not preclude differential effects across learner subgroups—benefits for higher-ability students, alongside smaller or null effects for lower-ability students, could

XSL•FO
RenderX

aggregate to a near-zero average difference. To examine this possibility, we conducted an exploratory analysis guided by the ATI framework [49].

Our exploratory ATI analysis examined whether the benefits of intensive AMTES engagement varied by baseline academic ability. Among AMTES users, regression models suggested that the added value of high- versus low-intensity engagement may be moderated by baseline academic ability. The pattern appeared more evident for excellence-related outcomes and among higher-performing students, whereas pass rates were uniformly high with limited room for differentiation.

These heterogeneity findings should be interpreted as hypothesis-generating. However, the observed pattern is conceptually consistent with the ATI framework [28,30,49], which posits that the effects of an instructional treatment may depend systematically on learner characteristics. One possible explanation is that cognitive resources and instructional fit may shape how effectively learners act on structured feedback [49], such that stronger baseline competence may support greater gains [47]. Conversely, learners with weaker foundations may require additional instructional scaffolding to benefit fully from intensive practice [50].

Overall, these exploratory findings highlight the potential value of aligning AI-supported practice design (eg, case diversity, adaptive difficulty, and feedback scaffolding) with learners' baseline competence to optimize educational impact and instructional fit [48]. These considerations inform the practical implications discussed next, including design and instructional strategies for personalized, adaptive use of AMTES within routine curricula.

Implications for Teaching Practice: Toward SRL-Supportive and Precision Medical Education

Our findings suggest that AI systems, such as AMTES, should be evaluated not only as technical innovations, but also by their impact in authentic teaching contexts. When integrated into modern medical curricula, such tools need to be used in a learner-oriented way that supports SRL [51] and precision education [52].

From an SRL perspective, AMTES is most valuable when it supports students in planning, monitoring, and adjusting their learning [26]. As Lajoie and Gube [51] argued, technology-enhanced learning environments should scaffold SRL processes, such as planning, monitoring, and reflection, to cultivate adaptive expertise rather than being used solely for automated, routine practice. Instructors can therefore help students establish efficient usage patterns grounded in SRL by deliberately structuring AMTES practice around reflection and feedback integration. This approach may encourage more purposeful, feedback-informed practice.

At the same time, our exploratory findings suggest that AMTES may support learners across ability levels, but intensive engagement may yield greater gains among students with stronger baseline competence. This pattern aligns with the emerging framework of precision medical education, which emphasizes leveraging longitudinal learning analytics and data-driven (including AI-enabled) insights to inform personalized support and interventions tailored to individual learner needs [52]. Instructors may therefore support students by promoting efficient, SRL-aligned engagement strategies and encouraging practice plans that are responsive to individual learning needs.

Taken together, these implications point to a learner-centered approach to integrating AMTES into medical curricula. In this way, AI-based systems, such as AMTES, are positioned not as simple extra practice tools, but as supports for self-regulation and individualized learning trajectories that can complement existing instruction.

Limitations and Future Directions

Several limitations must be acknowledged. First, as an observational retrospective cohort study, residual confounding cannot be fully ruled out [53]. The Rosenbaum bounds sensitivity analysis indicated robustness only to unmeasured confounding of weak strength, and unobserved factors, such as intrinsic motivation, SRL skills [54], digital literacy [55], or time management could, in principle, still influence both AMTES use and examination performance. However, our matching model was based on a comprehensive set of covariates, including 2 demographic variables (age and sex) and 3 key baseline academic performance metrics (midterm, previous theory, and practical examination scores) collected before the AMTES usage period. These performance metrics are widely recognized as strong proxies for underlying traits, such as motivation, self-discipline, and study habits [34,35,56,57]. By adjusting for these performance indicators, we likely reduced a substantial proportion of confounding related to these latent characteristics, although the possibility of residual bias from other unmeasured factors cannot be completely excluded.

Second, the clustering of practice patterns and the interaction analysis were exploratory in nature. Interaction effects are inherently difficult to estimate precisely, and our heterogeneity analyses should be treated as exploratory and hypothesis-generating [29,30]. Future studies with larger samples, richer case libraries, and longer follow-up periods are warranted to further examine whether practice patterns yield differential benefits across learners with different baseline competencies.

Finally, this study was conducted at a single institution over 1 academic cycle, and it only assessed immediate examination outcomes, lacking follow-up on long-term effects. It remains unclear whether AMTES training can lead to long-term improvements in students' performance during clinical clerkships or subsequent examinations, or its deeper impact on motivation and SRL habits. Furthermore, the differential effects of varying numbers of training cases and the system's applicability across different cultures and educational systems remain to be validated. Generalizing the findings to other institutions and teaching environments requires caution.

Future research should use larger-scale designs (such as randomized controlled trials or expanded real-world studies), include follow-up on subsequent student performance and attitudinal changes, and extend validation to different grade levels, courses, and multiple centers. Such work will be

XSL•FO
RenderX

important for confirming the robustness of the present findings and for refining how AI-based history-taking trainers like AMTES can best be integrated into medical curricula worldwide.

Conclusion

Building upon previous technical validation, this study contributes real-world evidence from an authentic, high-baseline educational context in which AMTES was offered as a voluntary extracurricular supplement alongside the routine curriculum. Compared with many previous works focusing on technical feasibility or short-term controlled trials, this study found voluntary extracurricular AMTES use was associated with modest yet meaningful improvements in summative history-taking performance. Exploratory analyses suggested that the added value of more intensive engagement may be moderated by baseline academic ability, indicating that the benefits of AI-supported practice could vary across learner profiles. These findings support the scalability and resource efficiency of AI-enabled supplementary training and inform precision-oriented instructional design. Future studies should prospectively evaluate heterogeneous effects across diverse learners and test adaptive implementation strategies, including expanded case libraries, SRL-guided scaffolding, and ability-aligned progression when implementing AI-based training tools.

Acknowledgments

The authors declare the use of generative artificial intelligence (GenAI) in the research and writing process. According to the GAIDeT taxonomy (2025), the following tasks were delegated to GenAI tools under full human supervision, (1) proofreading and editing, and (2) translation. The GenAI tool used was Gemini 3 (Google). Responsibility for the final manuscript lies entirely with the authors.

GenAI tools are not listed as authors and do not bear responsibility for the final outcomes.

Funding

We acknowledge the support from the Key Project on Medical Education Research by the National Medical Education Development Center, Medical Education Branch of the Chinese Medical Association (2025A25); the 2025 Teaching Reform and Research Project of Shantou University Medical College (number 10); and the 2025 Teaching Reform and Research Project (AI Curriculum Special Program) of Shantou University Medical College (number 8). The funders had no involvement in the study design, data collection, analysis, interpretation, or the writing of the manuscript.

Data Availability

The datasets generated or analyzed during this study are available from the corresponding author on reasonable request.

Authors' Contributions

CS developed the methodology, secured funding, supervised the research team, and led project management. OX conceptualized and designed the study and led project management. WZ conceptualized and designed the study, developed the software system, contributed to data analysis, and managed the project. YL developed the research methodology, conducted the investigation, performed data analysis, and prepared the original manuscript draft. YZ developed analytical methodology, performed formal statistical analysis, and generated data visualizations. LW contributed to the conceptualization and design of the study and assisted with project management. XL contributed to the conceptual framework and study design. MY was responsible for data curation and contributed to data analysis.

YL, YZ, and WZ contributed equally to this work and are cofirst authors; OX and CS contributed equally to this work and are cocorresponding authors.

All authors reviewed and approved the final manuscript for publication.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Normality test for within-pair differences in final practical skills examination scores.
[PDF File (Adobe PDF File), 104 KB - mededu_v12i1e89367_app1.pdf]

Multimedia Appendix 2
Clustering features construction and selection.
[PDF File (Adobe PDF File), 209 KB - mededu_v12i1e89367_app2.pdf]

XSL•FO
RenderX

Multimedia Appendix 3
Determination of the optimal number of clusters (K).
[PDF File (Adobe PDF File), 112 KB - mededu_v12i1e89367_app3.pdf]

Multimedia Appendix 4
Propensity score distributions before and after matching.
[PDF File (Adobe PDF File), 190 KB - mededu_v12i1e89367_app4.pdf]

References

1. Alharbi L, Almoallim H. History-taking skills in rheumatology. In: Almoallim H, Cheikh M, editors. Skills in Rheumatology. Singapore: Springer; 2021:3-16.
2. Kantar A, Marchant JM, Song WJ, Shields MD, Chatziparasidis G, Zacharasiewicz A, et al. History taking as a diagnostic tool in children with chronic cough. Front Pediatr 2022;10:850912 [FREE Full text] [doi: 10.3389/fped.2022.850912] [Medline: 35498777]
3. Steinkellner C, Schlömmer C, Dünser M. [Medical history taking and clinical examination in emergency and intensive care medicine]. Med Klin Intensivmed Notfmed 2020;115(7):530-538 [FREE Full text] [doi: 10.1007/s00063-020-00731-x] [Medline: 32885280]
4. Meng X, Zhang M, Ma W, Cheng X, Yang X. A clinical medicine level test at Jinan University School of Medicine reveals the importance of training medical students in clinical history-taking. PeerJ 2023;11:e15052 [FREE Full text] [doi: 10.7717/peerj.15052] [Medline: 37009162]
5. Keifenheim KE, Teufel M, Ip J, Speiser N, Leehr EJ, Zipfel S, et al. Teaching history taking to medical students: a systematic review. BMC Med Educ 2015 Sep 28;15:159 [FREE Full text] [doi: 10.1186/s12909-015-0443-x] [Medline: 26415941]
6. Kolb DA. In: Englewood Cliffs NJ, editor. Experiential Learning: Experience as the Source of Learning and Development. New Jersey, U.S: Prentice-Hall; 1984:256.
7. Aranda JH, Monks SM. Roles and Responsibilities of the Standardized Patient Director in Medical Simulation. Treasure Island (FL): StatPearls. StatPearls Publishing Copyright © 2025, StatPearls Publishing LLC; 2025.
8. Or AJ, Sukumar S, Ritchie HE, Sarrafpour B. Using artificial intelligence chatbots to improve patient history taking in dental education (Pilot study). J Dent Educ 2024;88 Suppl 3(Suppl 3):1988-1990. [doi: 10.1002/jdd.13591] [Medline: 38783404]
9. Webb JJ. Proof of concept: Using ChatGPT to teach emergency physicians how to break bad news. Cureus 2023;15(5):e38755 [FREE Full text] [doi: 10.7759/cureus.38755] [Medline: 37303324]
10. Cook DA, Overgaard J, Pankratz VS, Del Fiol G, Aakre CA. Virtual patients using large language models: Scalable, contextualized simulation of clinician-patient dialogue with feedback. J Med Internet Res 2025;27:e68486 [FREE Full text] [doi: 10.2196/68486] [Medline: 39854611]
11. Wang C, Li S, Lin N, Zhang X, Han Y, Wang X, et al. Application of large language models in medical training evaluation-using ChatGPT as a standardized patient: Multimetric assessment. J Med Internet Res 2025;27:e59435 [FREE Full text] [doi: 10.2196/59435] [Medline: 39742453]
12. Ahmed T, Rabinowitz LG, Rodman A, Berzin TM. Generative artificial intelligence tools in gastroenterology training. Clin Gastroenterol Hepatol 2024;22(10):1975-1978. [doi: 10.1016/j.cgh.2024.05.050] [Medline: 39122232]
13. Choi HS, Song JY, Shin KH, Chang JH, Jang BS. Developing prompts from large language model for extracting clinical information from pathology and ultrasound reports in breast cancer. Radiat Oncol J 2023 Sep;41(3):209-216 [FREE Full text] [doi: 10.3857/roj.2023.00633] [Medline: 37793630]
14. Ning Y, Ong JCL, Cheng H, Wang H, Ting DSW, Tham YC, et al. How can artificial intelligence transform the training of medical students and physicians? Lancet Digit Health 2025;7(10):100900 [FREE Full text] [doi: 10.1016/j.landig.2025.100900] [Medline: 41047321]
15. Liu Y, Shi C, Wu L, Lin X, Chen X, Zhu Y, et al. Development and validation of a large language model-based system for medical history-taking training: Prospective multicase study on evaluation stability, Human-AI consistency, and transparency. JMIR Med Educ 2025;11:e73419 [FREE Full text] [doi: 10.2196/73419] [Medline: 40882613]
16. Cukurova M. The interplay of learning, analytics and artificial intelligence in education: A vision for hybrid intelligence. Brit J Educational Tech 2024;56(2):469-488. [doi: 10.1111/bjet.13514]
17. Waikel RL, Othman AA, Patel T, Ledgister Hanchard S, Hu P, Tekendo-Ngongang C, et al. Recognition of genetic conditions after learning with images created using generative artificial intelligence. JAMA Netw Open 2024;7(3):e242609 [FREE Full text] [doi: 10.1001/jamanetworkopen.2024.2609] [Medline: 38488790]
18. Yi Y, Kim KJ. The feasibility of using generative artificial intelligence for history taking in virtual patients. BMC Res Notes 2025;18(1):80 [FREE Full text] [doi: 10.1186/s13104-025-07157-8] [Medline: 39994780]
19. Huh CY, Lee J, Kim G, Jang Y, Ko H, Suh MJ, et al. Using large language models to simulate history taking: Implications for symptom-based medical education. Information 2025;16(8):653. [doi: 10.3390/info16080653]

XSL•FO
RenderX

20. Holderried F, Stegemann-Philipps C, Herschbach L, Moldt J, Nevins A, Griewatz J, et al. A generative pretrained transformer (GPT)-Powered Chatbot as a simulated patient to practice history taking: Prospective, mixed methods study. JMIR Med Educ 2024;10:e53961 [FREE Full text] [doi: 10.2196/53961] [Medline: 38227363]
21. Sheth U, Lo M, McCarthy J, Baath N, Last N, Guo E, et al. Understanding the role of large language model virtual patients in developing communication and clinical skills in undergraduate medical education. IME 2025;4(4):39. [doi: 10.3390/ime4040039]
22. Luo MJ, Bi S, Pang J, Liu L, Tsui CK, Lai Y, et al. A large language model digital patient system enhances ophthalmology history taking skills. NPJ Digit Med 2025;8(1):502 [FREE Full text] [doi: 10.1038/s41746-025-01841-6] [Medline: 40760042]
23. Wang Z, Fan T, Li M, Zhu N, Wang X. Feasibility study of using GPT for history-taking training in medical education: A randomized clinical trial. BMC Med Educ 2025;25(1):1030 [FREE Full text] [doi: 10.1186/s12909-025-07614-9] [Medline: 40640776]
24. Cho KK, Marjadi B, Langendyk V, Hu W. The self-regulated learning of medical students in the clinical environment - a scoping review. BMC Med Educ 2017;17(1):112 [FREE Full text] [doi: 10.1186/s12909-017-0956-6] [Medline: 28693468]
25. Zheng B, Zhang Y. Self-regulated learning: The effect on medical student learning outcomes in a flipped classroom environment. BMC Med Educ 2020;20(1):100. [doi: 10.1186/s12909-020-02023-6] [Medline: 32234040]
26. Ballouk R, Mansour V, Dalziel B, McDonald J, Hegazi I. Medical students' self-regulation of learning in a blended learning environment: A systematic scoping review. Med Educ Online 2022;27(1):2029336 [FREE Full text] [doi: 10.1080/10872981.2022.2029336] [Medline: 35086439]
27. Sun Y, Liu F. Real-world implementation of an AI learning tool-MetaGP-Edu in medical education: A multi-center cohort study. Computers & Education 2025;237:105388. [doi: 10.1016/j.compedu.2025.105388]
28. Cronbach LJ, Snow RE. Aptitudes and Instructional Methods: A Handbook for Research on Interactions. New York: Irvington Publishers: distributed by Halsted Press; 1977:534.
29. DeJong TL, Chen Q. Utility of a slopes difference test for probing longitudinal multilevel aptitude treatment interactions: A simulation. Front Psychol 2023;14:1156962 [FREE Full text] [doi: 10.3389/fpsyg.2023.1156962] [Medline: 37441330]
30. Preacher KJ, Sterba SK. Aptitude-by-Treatment interactions in research on educational interventions. Exceptional Children 2018;85(2):248-264. [doi: 10.1177/0014402918802803]
31. Ma YN, Ishak Z. Cognitive strategy use and self-regulation as mediators between personality and academic achievement among Chinese undergraduates in Malaysia. Jonus 2024;9(1):245-268. [doi: 10.24200/jonus.vol9iss1pp245-268]
32. Song Y, Zhang R, Li Z, Xin C, Ding N. Unveiling the link between self-regulated learning and academic success: A longitudinal study on Chinese medical students. Adv Health Sci Educ Theory Pract 2025;30(4):1181-1197. [doi: 10.1007/s10459-024-10400-9] [Medline: 39621181]
33. Appelbaum M, Cooper H, Kline RB, Mayo-Wilson E, Nezu AM, Rao SM. Journal article reporting standards for quantitative research in psychology: The APA publications and communications board task force report. Am Psychol 2018;73(1):3-25. [doi: 10.1037/amp0000191] [Medline: 29345484]
34. Richardson M, Abraham C, Bond R. Psychological correlates of university students' academic performance: A systematic review and meta-analysis. Psychol Bull 2012;138(2):353-387. [doi: 10.1037/a0026838] [Medline: 22352812]
35. Zimmerman BJ. Self-Efficacy: An essential motive to learn. Contemp Educ Psychol 2000;25(1):82-91. [doi: 10.1006/ceps.1999.1016] [Medline: 10620383]
36. Venkatesh V, Thong J, Xu X. Consumer acceptance and use of information technology: Extending the unified theory of acceptance and use of technology. MIS Quart 2012;36(1):157-158. [doi: 10.2307/41410412]
37. Shikino K, Nishizaki Y, Fukui S, Kataoka K, Yokokawa D, Shimizu T, et al. Resident physician recognition of tachypnea in clinical simulation videos in Japan: Cross-Sectional study. JMIR Med Educ 2025;11:e72640 [FREE Full text] [doi: 10.2196/72640] [Medline: 40743513]
38. Liu W, Kuramoto SJ, Stuart EA. An introduction to sensitivity analysis for unobserved confounding in nonexperimental prevention research. Prev Sci 2013;14(6):570-580 [FREE Full text] [doi: 10.1007/s11121-012-0339-5] [Medline: 23408282]
39. Tsikas SA, Afshar K, Fischer V. Does voluntary practice improve the outcome of an OSCE in undergraduate medical studies? A propensity score matching approach. PLoS One 2024;19(10):e0312387 [FREE Full text] [doi: 10.1371/journal.pone.0312387] [Medline: 39446696]
40. Powell MG, Hull DM, Beaujean AA. Propensity score matching for education data: Worked examples. The Journal of Experimental Education 2019;88(1):145-164. [doi: 10.1080/00220973.2018.1541850]
41. Holderried F, Stegemann-Philipps C, Herrmann-Werner A, Festl-Wietek T, Holderried M, Eickhoff C, et al. A language model-powered simulated patient with automated feedback for history taking: Prospective study. JMIR Med Educ 2024;10:e59213 [FREE Full text] [doi: 10.2196/59213] [Medline: 39150749]
42. Brügge E, Ricchizzi S, Arenbeck M, Keller MN, Schur L, Stummer W, et al. Large language models improve clinical decision making of medical students through patient simulation and structured feedback: A randomized controlled trial. BMC Med Educ 2024;24(1):1391 [FREE Full text] [doi: 10.1186/s12909-024-06399-7] [Medline: 39609823]
43. Song HS, Kalet AL, Plass JL. Assessing medical students' self-regulation as aptitude in computer-based learning. Adv Health Sci Educ Theory Pract 2011;16(1):97-107 [FREE Full text] [doi: 10.1007/s10459-010-9248-1] [Medline: 20872071]

XSL•FO
RenderX

44. Sobral DT. An appraisal of medical students' reflection-in-learning. Med Educ 2000;34(3):182-187. [doi: 10.1046/j.1365-2923.2000.00473.x] [Medline: 10733703]
45. Cleary TJ, Sandars J. Assessing self-regulatory processes during clinical skill performance: A pilot study. Med Teach 2011;33(7):e368-e374. [doi: 10.3109/0142159X.2011.577464] [Medline: 21696270]
46. Turan S, Konan A. Self-regulated learning strategies used in surgical clerkship and the relationship with clinical achievement. J Surg Educ 2012;69(2):218-225. [doi: 10.1016/j.jsurg.2011.09.003] [Medline: 22365869]
47. Maqsood Z, Sajjad M, Yasmin R. Effect of feedback-integrated reflection, on deep learning of undergraduate medical students in a clinical setting. BMC Med Educ 2025;25(1):66 [FREE Full text] [doi: 10.1186/s12909-025-06648-3] [Medline: 39810114]
48. Coates WC. Precision education - a call to action to transform medical education. Int J Emerg Med 2025;18(1):21 [FREE Full text] [doi: 10.1186/s12245-025-00819-1] [Medline: 39924527]
49. Sana F, Fenesi B. Working memory and instructional fit: Reintroducing aptitude-treatment interaction in education research. Behav Sci (Basel) 2025;15(6):765 [FREE Full text] [doi: 10.3390/bs15060765] [Medline: 40564547]
50. Faber TJE, Dankbaar MEW, van den Broek WW, Bruinink LJ, Hogeveen M, van Merriënboer JJG. Effects of adaptive scaffolding on performance, cognitive load and engagement in game-based learning: A randomized controlled trial. BMC Med Educ 2024;24(1):943 [FREE Full text] [doi: 10.1186/s12909-024-05698-3] [Medline: 39210381]
51. Lajoie SP, Gube M. Adaptive expertise in medical education: Accelerating learning trajectories by fostering self-regulated learning. Med Teach 2018;40(8):809-812. [doi: 10.1080/0142159X.2018.1485886] [Medline: 30033791]
52. Triola MM, Burk-Rafel J. Precision medical education. Acad Med 2023;98(7):775-781. [doi: 10.1097/ACM.0000000000005227] [Medline: 37027222]
53. Bikdeli B, Ross JS, Bukhari S, Jeffery MM, Lip GY, You SC, et al. Comparative effectiveness research using randomized trials and observational studies: Validity and feasibility considerations. Thromb Haemost 2025:00 [FREE Full text] [doi: 10.1055/a-2664-7887] [Medline: 40719151]
54. Broadbent J, Poon WL. Self-regulated learning strategies and academic achievement in online higher education learning environments: A systematic review. The Internet and Higher Education 2015;27:1-13. [doi: 10.1016/j.iheduc.2015.04.007]
55. Sandars J, Correia R, Dankbaar M, de Jong P, Goh PS, Hege I, et al. Twelve tips for rapidly migrating to online learning during the COVID-19 pandemic. MedEdPublish (2016) 2020;9:82 [FREE Full text] [doi: 10.15694/mep.2020.000082.1] [Medline: 38058947]
56. Mahdavi P, Valibeygi A, Moradi M, Sadeghi S. Relationship between achievement motivation, mental health and academic success in university students. Community Health Equity Res Policy 2023;43(3):311-317. [doi: 10.1177/0272684X211025932] [Medline: 34176355]
57. Puddey IB, Mercer A. Predicting academic outcomes in an Australian graduate entry medical programme. BMC Med Educ 2014;14:31 [FREE Full text] [doi: 10.1186/1472-6920-14-31] [Medline: 24528509]

Abbreviations

AI: artificial intelligence
AMTES: AI-powered Medical History-Taking Training and Evaluation System
ATI: aptitude-treatment interaction
ATT: average treatment effect on the Treated
LLM: large language model
PSM: propensity score matching
SMD: standardized mean difference
SP: standardized patient
SRL: self-regulated learning
VP: virtual patient

Edited by S Brini; submitted 11.Dec.2025; peer-reviewed by F Fukuzawa, L He; comments to author 13.Jan.2026; accepted 06.Feb.2026; published 24.Feb.2026.

Please cite as:
Liu Y, Zhu Y, Zhang W, Lu X, Wu L, Yue M, Xia O, Shi C
Real-World Impact and Educational Effectiveness of an AI-Powered Medical History-Taking System: Retrospective Propensity Score-Matched Cohort Study
JMIR Med Educ 2026;12:e89367
URL: https://mededu.jmir.org/2026/1/e89367
doi:10.2196/89367
PMID:

XSL•FO
RenderX

©Yang Liu, Yiying Zhu, Weishan Zhang, Xian Lu, Liping Wu, Minghui Yue, Oudong Xia, Chujun Shi. Originally published in JMIR Medical Education (https://mededu.jmir.org), 24.Feb.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Education, is properly cited. The complete bibliographic information, a link to the original publication on https://mededu.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

AI-generated Feedback Following Social Robotic Virtual Patient Interactions and Medical Student Performance: Nonrandomized Quasi-Experimental Study

Alexander Borg[1], MD, PhD; Jonathan Schiött[1], MSc; William Ivegren[1], MSc; Cidem Gentline[1], MD; Viking Huss[2], MD; Anna Margareta Hugelius[1], MD; Benjamin Jobs[1], MD; Mini Ruiz[3], MD, PhD; Samuel Edelbring[1,4,5], PhD; Carina Georg[6], RN, CCRN, PhD; Gabriel Skantze[7], PhD; Ioannis Parodis[1,8], MD, PhD

[1]Division of Rheumatology, Department of Medicine Solna, Karolinska Institutet, Karolinska University Hospital, and Center for Molecular Medicine (CMM), Solna, Sweden

[2]Division of Clinical Epidemiology, Department of Medicine Solna, Karolinska Institutet and Karolinska University Hospital, Solna, Sweden

[3]Department of Clinical Science, Intervention and Technology, Karolinska Institutet, Stockholm, Stockholm, Sweden

[4]Department of Educational Sciences and Arts, Faculty of Philosophy, Mälardalen University, Västerås, Västmanland, Sweden

[5]School of Health Sciences, Örebro University, Örebro, Örebro, Sweden

[6]Department of Neurobiology, Care Sciences and Society, Karolinska Institutet, Stockholm, Stockholm, Sweden

[7]Division of Speech Music and Hearing, Royal Institute of Technology (KTH), Stockholm, Stockholm, Sweden

[8]Department of Rheumatology, Faculty of Medicine and Health, Örebro University, Örebro, Örebro, Sweden

Corresponding Author:
Alexander Borg, MD, PhD
Division of Rheumatology
Department of Medicine Solna
Karolinska Institutet, Karolinska University Hospital, and Center for Molecular Medicine (CMM)
D2:01 Rheumatology, Karolinska University Hospital Solna
Solna, SE-171 76
Sweden
Phone: 46 737533351
Email: alexander.borg@ki.se

Abstract

Background: Virtual patients (VPs) demonstrate effectiveness in improving clinical reasoning skills; however, traditional VP platforms often lack individualized feedback mechanisms. Advances in large language models (LLMs) enable automated analysis of student-VP interactions, providing scalable feedback on clinical performance. While artificial intelligence (AI)–enhanced social robotic VP platforms show promise for clinical reasoning training, no studies have examined whether AI-generated feedback integrated in such platforms improves clinical performance in standardized assessments.

Objective: This study evaluated whether AI-generated postconsultation feedback integrated into social robotic VP interactions improves medical students' clinical performance, emphasizing medical history taking and communication.

Methods: A quasi-experimental study with 115 sixth-semester medical students (N=157, 73.2% of eligible students) was conducted at Karolinska Institutet, Stockholm, Sweden, during spring 2025. Students were allocated by hospital site to receive (n=61, 53%) or not receive (n=54, 46.9%) AI-generated feedback following interactions with a Social AI-Enhanced Robotic Interface. All students completed 9 VP cases; the intervention group received approximately 1 page of structured feedback after each VP case. The feedback system used multiple LLMs following a 2-stage algorithm: assessing student-VP dialogues using an assessment rubric, then generating structured feedback on history-taking performance. Both groups participated in case-specific follow-up seminars led by consultant rheumatologists following each VP encounter. Clinical performance was assessed through an 8-minute objective structured clinical examination (OSCE)-based evaluation, with a standardized patient portraying axial spondylarthritis, evaluated by a blinded consultant rheumatologist using a 10-point rubric across 5 domains: communication at consultation start, generic medical history, targeted medical history, diagnostics and management reasoning, and communication at consultation end.

XSL•FO
RenderX

Results: Students receiving AI-generated feedback achieved significantly higher total OSCE scores (mean 7.39, SD 0.86 vs mean 6.68, SD 1.04 points; mean difference 0.70; 95% CI 0.35-1.06; *P*<.001; Cohen *d*=0.74). Domain-specific analysis revealed significant improvement in generic medical history after Bonferroni correction (mean 2.46, SD 0.65 vs mean 2.03, SD 0.79 points; *P*=.004; *r*=0.27), while other domains showed no significant differences: communication at start (*P*=.13; *r*=0.14), targeted medical history taking (*P*=.60; *r*=0.05), diagnostics and management (*P*=.14; *r*=0.14), and communication at consultation end (*P*=.31; *r*=0.09). Pass rates were significantly higher in the feedback group (96.7% vs 79.6%; odds ratio 7.55, 95% CI 1.51-72.2; *P*=.006), with a number needed to assess of 6 students, that is, for every 6 students receiving feedback, 1 additional student passed the assessment.

Conclusions: AI-generated feedback following social robotic VP interactions significantly improved medical students' OSCE-based performance, particularly in generic medical history taking. These findings support integrating validated AI feedback systems as a supplement to expert-led teaching during VP simulations for clinical training and demonstrate the feasibility of scalable, automated feedback in medical education. The domain-specific improvements in generic medical history highlight the importance of targeted, competency-specific feedback design in VP platforms.

(JMIR Med Educ 2026;12:e90368) doi:10.2196/90368

KEYWORDS

virtual patients; medical education; clinical reasoning; large language models; social robotics; artificial intelligence; AI; educational technology

Introduction

Clinical reasoning (CR) represents a fundamental cognitive process that guides diagnostic and therapeutic decisions in clinical practice [1,2]. Virtual patients (VPs) have emerged as valuable educational tools for developing CR skills in medical education, providing safe environments for practicing patient interactions [3,4]. While VPs demonstrate effectiveness in improving CR skills, traditional VP platforms often lack sophisticated and individualized feedback mechanisms, potentially limiting skill transfer to real-world clinical encounters [5,6]. The integration of artificial intelligence (AI) within VP simulations addresses this limitation, with recent systematic reviews demonstrating positive effects on CR skills, particularly in case-specific domains such as data gathering and diagnostic accuracy [6,7].

Social robots are physically embodied, autonomous systems designed to interact with humans following social behaviors and communication norms [8]. Social robotics represents a promising frontier in simulation-based interactions, with educational applications showing enhanced engagement and learning outcomes across various disciplines [9]. However, no studies have examined the integration of AI-generated feedback within social robotic VPs or evaluated whether such feedback improves objectively measured clinical performance.

Diagnostic errors represent a significant challenge in health care, with inadequate history taking identified as a leading contributor, as medical history alone can often aid in establishing a correct diagnosis [10,11]. Scalable approaches to improving the quality of medical history–taking training therefore carry implications beyond educational outcomes, extending to actual patient benefit [12]. Feedback represents a cornerstone of effective medical education, particularly in developing CR skills [13,14]. Effective feedback should be specific, timely, and actionable, addressing both the content and process of CR [15]. Recent advances in large language models (LLMs) enable automated analysis of student-VP interactions, providing scalable feedback on medical history taking and diagnostic reasoning [16,17]. Studies using AI-generated feedback have demonstrated improved diagnostic accuracy and enhanced systematism during clinical problem solving [18].

We have developed a Social AI-Enhanced Robotic Interface (SARI) for VP encounters and demonstrated that this interface provides superior design characteristics for CR training compared with conventional computer-based VP platforms [19-22]. While technological advances such as AI-enhanced VPs show promise, ensuring the transfer of skills from simulated to real clinical encounters remains a critical challenge in VP education [4]. Objective structured clinical examinations (OSCEs) enable standardized assessment of application of skills in controlled settings and provide a structured bridge between simulated training environments and real-world clinical performance [23-25]. In our prior work, we examined the added value of AI-enhanced social robotic VPs compared with conventional computer-based VP simulations, focusing on medical students' self-reported perceptions [19-22]. However, these studies did not include objective performance measures, and the VP platforms lacked systematic feedback on students' CR processes.

While AI-enhanced social robotic VPs show promise for CR training, no studies have examined whether integrated AI-generated feedback translates into improved clinical performance as measured by means of standardized assessments. This study aimed to evaluate whether the integration of AI-generated postconsultation feedback within social robotic VP training enhances medical students' performance in CR skills with emphasis on history taking and communication, assessed through an OSCE-based evaluation with a standardized patient. We used a quasi-experimental design to test our hypothesis that students receiving AI-generated feedback following VP encounters would achieve higher OSCE scores compared with students completing the same VP training without such feedback. The findings are primarily relevant for medical educators, curriculum designers, and researchers developing AI-enhanced simulation-based training platforms.

XSL•FO
RenderX

Methods

Overview

We conducted a quasi-experimental study to evaluate whether AI-generated feedback following social robotic VP interaction improves medical students' clinical performance in medical history taking. The study was conducted at Karolinska Institutet, Stockholm, Sweden, during spring 2025, following the Transparent Reporting of Evaluations with Nonrandomized Designs reporting guidelines [26].

The study participants (N=115) completed an established virtual outpatient clinic educational activity within rheumatology, which involves interaction with 9 VP cases using the SARI [19-22] and the computer-based virtual interactive case system [27] for CR training. Students were allocated either to the intervention group (n=61, 53%), where AI-generated feedback was provided after each VP encounter with the SARI, or to the control group (n=54, 46.9%), where students completed identical cases without receiving AI-generated feedback. During a total of 6 weeks, students from 4 different hospital sites attended their week-long rheumatology clinical rotation, with each student based at Karolinska University Hospital Solna for the virtual outpatient clinic. The virtual outpatient clinic spanned one and a half days in total during this week. There was no crossover, and each student was exposed to either the intervention or control condition. Allocation was determined by hospital site to facilitate scheduling, which rotated weekly between intervention and control conditions to avoid potential selection bias owing to hospital site specifics. A flowchart of study participants and allocation is illustrated in Figure S1 in Multimedia Appendix 1.

Immediately following completion of all 9 VP cases at the end of the virtual outpatient clinic, students were evaluated for their performance in CR through an OSCE-based assessment, with emphasis placed on medical history taking and communication. The primary outcome was a comparison of total OSCE scores between the 2 groups, with secondary outcomes including domain-specific OSCE performance across 5 subdomains and pass or fail rates using a predetermined threshold. The OSCE-based rubric is presented in Figure S2 in Multimedia Appendix 1.

Study Participants

All sixth-semester medical students attending the virtual outpatient clinic during their rheumatology clinical rotation were eligible and invited to take part in the study. Of 157 eligible students, 115 (73.2%) were included in the study. Participation was voluntary, with students providing informed consent before enrollment. There were no formal inclusion criteria, instead all students in the sixth semester were asked to participate. Students were excluded if they were unable to attend the scheduled OSCE-like assessment or did not provide informed consent.

The virtual outpatient clinic includes interaction with 9 VP cases, followed by case-specific seminars with consultant rheumatologists. Students work in pairs or small groups of three, with one student taking the lead in each patient interaction, consistent with evidence supporting collaborative VP learning for CR training [28]. Before this educational activity, students receive instructions and training in structured medical history taking, covering both generic aspects and case-specific aspects with relevance for the discipline of rheumatology (Figure S3 in Multimedia Appendix 1). Students were based at 4 clinical sites in the Stockholm region, that is, Karolinska University Hospital Solna, Karolinska University Hospital Huddinge, Danderyd Hospital, and Södersjukhuset, but the virtual outpatient clinic took place at Karolinska University Hospital Solna for all students.

AI-Generated Feedback

We developed and validated an AI-powered feedback system integrated with the SARI. The system uses multiple LLMs from OpenAI to analyze transcribed student-VP dialogues and generate structured feedback on medical history–taking performance.

The feedback algorithm followed a 2-stage design, developed using historical SARI interaction data from previous semesters, and validated through consultant rheumatologist review. The evaluation framework covers communication skills; generic and rheumatology-specific aspects of medical history taking, including systematic symptom enquiry, pain characterization, and functional assessment; and selected domains of CR essential for diagnostic and management purposes.

The first stage of the feedback algorithm is an assessment model that evaluates student-VP dialogues using a predefined rubric developed in collaboration with consultant rheumatologists. The model was iteratively refined before the study and has been described previously [29]. Initial algorithm outputs were compared against expert assessments across 30 student-VP interaction dialogues that were graded twice by each rheumatologist. Final validation demonstrated a high degree of agreement between AI-generated assessment and assessment provided by consultant rheumatologists (81.8% accuracy, Cohen k=0.68) [29].

The second stage of the feedback algorithm involves generating a feedback output based on the stage-1 assessment. Students received approximately 1 page of structured written feedback immediately after completing each VP encounter with the SARI (but not the virtual interactive case system), across a total of 5 VP scenarios. This feedback focused on medical history taking within the context of rheumatology and included constructive comments with examples covering general history taking, specific symptom enquiries, and systematic assessment of the VPs. No feedback was provided on diagnostic accuracy or the appropriateness of proposed management steps. The 2-stage feedback algorithm, the stage 2 prompt used to generate feedback from the stage 1 assessment, and an example of AI-generated feedback are illustrated in Figures S4-S6 in Multimedia Appendix 1.

The Social AI-Enhanced Robotic Interface

The SARI is a VP modality concept developed and implemented by our research group and combines AI using LLMs with a social robot [19-22]. In this specific context, the SARI combines an LLM from OpenAI (GPT-4o-mini) [30] with a social robot from Furhat Robotics, which encompasses a back-projected

XSL•FO
RenderX

face on a robotic head and is capable of producing humanlike facial expressions, natural head movements, and lip-synchronized speech [31]. The principles of VP case development in the SARI have been described previously and include a prompt that combines a patient description with the last 10 dialogue turns to avoid repetition (Figure S7 in Multimedia Appendix 1). Improvements to the SARI are made between study periods following feedback from students participating in the rheumatology clinical placements and with the release of updated versions of LLMs. Refinements implemented before the investigation for this study included the use of a newer LLM (GPT 4o-mini), which also enabled removal of the previously required LED dialogue indicator. These improvements were important based on a previous qualitative study wherein students perceived the dialogue with the previous version of SARI somewhat mechanical at times [20]. The SARI configuration was frozen before data collection commenced and remained unchanged throughout the study.

VP cases are initiated with a patient introducing themselves to a medical student, describing why they are there. Afterward, the dialogue is free, and students explore the virtual environment until they feel that they can proceed to preliminary diagnostics and the suggestion of a management plan. Students engaged with the SARI through natural speech dialogue, with all interactions being audio-recorded and transcribed automatically within the system for the generation of AI-powered feedback. The transcribed dialogue was available to all students following interaction with the SARI. Along with the cases, students receive a standard laboratory test panel, including case-specific laboratory test results along with the corresponding reference values. All VP cases were in English to also accommodate visiting international exchange students who participate in clinical rotations in rheumatology at Karolinska University Hospital within the frame of separate educational activities or courses. However, all sixth-semester students who participated in this study were Swedish speaking, with most having Swedish as their mother tongue.

Data Collection

Immediately after completing the virtual outpatient clinic, participating students underwent an OSCE-like evaluation designed to assess performance in medical history taking and CR. The 8-minute evaluation involved interaction with a standardized patient, portrayed by 2 engineering coauthors of this study without medical training. They were instructed to portray a patient with axial spondylarthritis and received written case information in advance. Performance was assessed independently by 1 of 2 consultants in rheumatology using a standardized rubric, with the assessors being blinded to whether students had received AI-generated feedback as a part of the intervention or were allocated to the control group. Unlike the VP interactions during the virtual outpatient clinic, this OSCE-like evaluation was conducted in Swedish. Instructions for the assessment, provided to students and standardized patients, are presented in Figures S8 and S9 in Multimedia Appendix 1.

The assessment used a 10-point scale across 5 domains with differential weighting reflecting their relative importance during a clinical encounter: communication at consultation start (0-3 points), generic medical history (0-3.5 points), targeted medical history (0-1.5 points), diagnostics and management reasoning (0-1 point), and communication at consultation end (0-1 point). The rubric and domain weighting was developed and determined collaboratively with consultant rheumatologists based on established Entrustable Professional Activity frameworks [32], reflecting the emphasis on history-taking competencies in the virtual outpatient clinic. The rubric underwent iterative refinement through pilot testing with consultant rheumatologists who reviewed and provided feedback on the scoring criteria, domain definitions, and descriptions to ensure clarity and clinical relevance. A pass threshold of 6 out of 10 points was predetermined based on clinical competency standards and local Entrustable Professional Activity framework requirements (Figure S2 in Multimedia Appendix 1).

Statistical Analysis

Sample size was determined by the number of available students during clinical rotations in rheumatology, representing a convenience sample of all available participants. Baseline characteristics were compared between groups using 2-tailed independent samples *t* tests for continuous variables and Fisher exact tests for categorical variables. Normality assumptions were assessed using the Shapiro-Wilk test to determine the appropriate statistical approach for the comparisons between groups in relation to the primary and secondary study outcomes. The primary outcome (total OSCE score) was analyzed using an independent samples *t* test after confirming normal distribution. Secondary outcomes (domain-specific OSCE performance) were analyzed using Mann-Whitney *U* tests following confirmation of nonnormal distributions. To control for the increased risk of type 1 error when conducting multiple comparisons across the 5 OSCE domains and ensure that domain-specific findings represent robust effects rather than stochastic observations, Bonferroni correction was applied and yielded an adjusted significance level of a=.01 (0.05/5 comparisons). The Fisher exact test was used to compare pass or fail rates.

Results from independent *t* tests are presented as means and SDs, along with 95% CI for mean differences and Cohen *d* effect sizes. Results from Mann-Whitney *U* tests are presented as means and SDs, and effect size (*r*). Results from Fisher exact tests are presented as frequencies and the corresponding percentage, odds ratio, and 95% CI. The statistical analysis was performed using R (version 4.3.3; R Foundation for Statistical Computing). Differences yielding *P* values <.05 were considered statistically significant, except for secondary outcomes where the Bonferroni-adjusted threshold of *P*<.01 was applied.

Ethical Considerations

The study was reviewed and approved by the Swedish Ethical Review Authority (reference: 2024-05876-02). Students received both written and verbal information about the study and provided written informed consent before participation. Students could withdraw from the study at any time without providing a reason, and withdrawal had no impact on their educational outcomes. All data were handled in accordance with the General Data Protection Regulation and stored on secure servers at Karolinska

XSL•FO
RenderX

Institutet. Audio recordings, transcriptions from feedback generation, and assessments were pseudonymized before analysis. No compensation was provided to participating students.

Results

Study Participant Characteristics and Baseline Comparisons

All 115 study participants completed the virtual outpatient clinic and subsequent OSCE-based evaluation, yielding complete data for analysis in all cases. Of the 115 medical students, 61 (53%) were allocated to the AI-generated feedback group and 54 (46.9%) to the control group. Baseline characteristics were well balanced between groups (Table 1). The mean age was similar in the feedback group (mean 26, SD 5.9 years) and the control group (mean 26.8, SD 5.3 years; *P*=.49). Sex distribution yielded no statistically significant difference between the two groups; the feedback group had a numerically greater percentage of female students (35/61, 57.4%) compared with the control group (24/54, 44.4%; *P*=.17). Finally, students' hospital site distribution was balanced (Table 1).

Table 1. Baseline characteristics of the study participants (N=115).

Characteristics	Feedback group	Control group	*P* value
Age (y), mean (SD)	26.0 (5.9)	26.8 (5.3)	.49
Sex, n (%)			.17
Female	35 (57.4)	24 (44.4)	
Male	26 (42.6)	30 (55.6)	
Hospital site, n (%)			.85
KS[a] Huddinge or DS[b]	31 (50.8)	29 (53.7)	
KS Solna or SÖS[c]	30 (49.2)	25 (46.3)	

[a]KS: Karolinska University Hospital.

[b]DS: Danderyd Hospital.

[c]SÖS: Södersjukhuset.

Total OSCE Score Comparisons

Students who received AI-generated feedback achieved significantly higher total OSCE scores compared with the control group (mean 7.39, SD 0.86 vs mean 6.68, SD 1.04 points), representing a mean difference of 0.70 points (95% CI 0.35-1.06; t_{113}=3.96; *P*<.001). This difference corresponds to a medium to large effect size (Cohen *d*=0.74). Detailed results are presented in Table 2 and Figure 1. No statistically significant differences were noted when comparing results from the 2 assessors (Figure S10 in Multimedia Appendix 1).

Table 2. Comparison of objective structured clinical examination performance outcomes between students receiving artificial intelligence–generated feedback following virtual patient interaction with the Social AI-Enhanced Robotic Interface (SARI) and students who did not receive feedback.

Outcomes	Max points	Feedback group (n=61), mean (SD)	Control group (n=54), mean (SD)	Mean difference	Effect size	*P* value
Primary outcome[a]						
Total OSCE[b] score	10.0	7.39 (0.86)	6.68 (1.04)	0.70	*d*=0.74	*<.001* [c]
Secondary outcomes[d]						
Communication at start	3.0	2.76 (0.44)	2.62 (0.57)	0.14	*r*=0.14	.13
Generic medical history	3.5	2.46 (0.65)	2.03 (0.79)	0.43	*r*=0.27	*.004* [e]
Targeted medical history	1.5	1.13 (0.22)	1.15 (0.27)	−0.02	*r*=0.05	.60
Diagnostics and management	1.0	0.59 (0.28)	0.52 (0.25)	0.07	*r*=0.14	.14
Communication at end	1.0	0.44 (0.46)	0.36 (0.46)	0.08	*r*=0.09	.31

[a]Analyzed using the independent samples *t* test.

[b]OSCE: objective structured clinical examination.

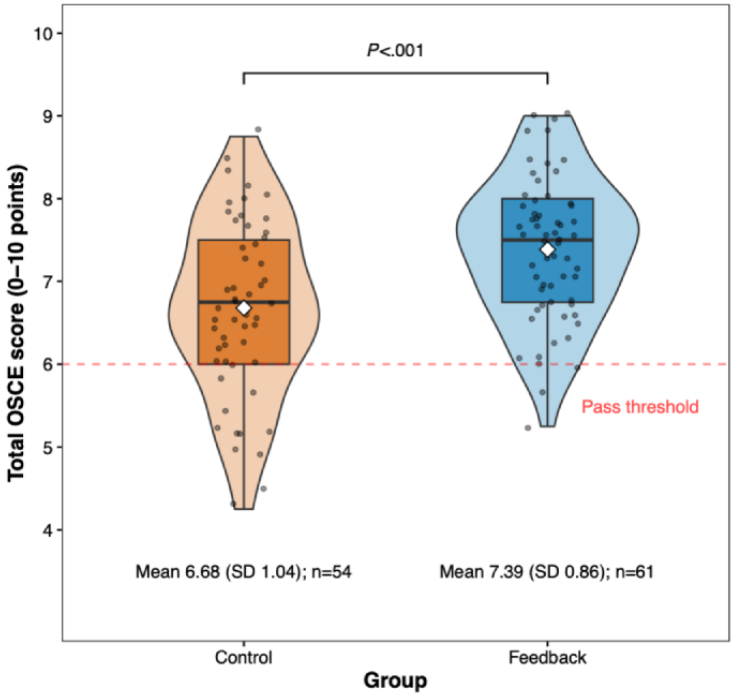
[c]Statistically significant.

[d]Analyzed using Mann-Whitney *U* tests.

[e]After Bonferroni correction (α=.01 for 5 comparisons).

XSL•FO
RenderX

Figure 1. Violin plots illustrating results from the independent samples *t* test comparing the total objective structured clinical examination (OSCE) score between students who received artificial intelligence (AI)–generated feedback following interaction with virtual patients in the Social AI-Enhanced Robotic Interface (SARI) compared with controls. Violin plots display score distributions with embedded boxplots showing the IQR. Individual student scores are shown as jittered points, with means marked by diamonds. The dashed horizontal red line indicates the pass threshold of 6 points.

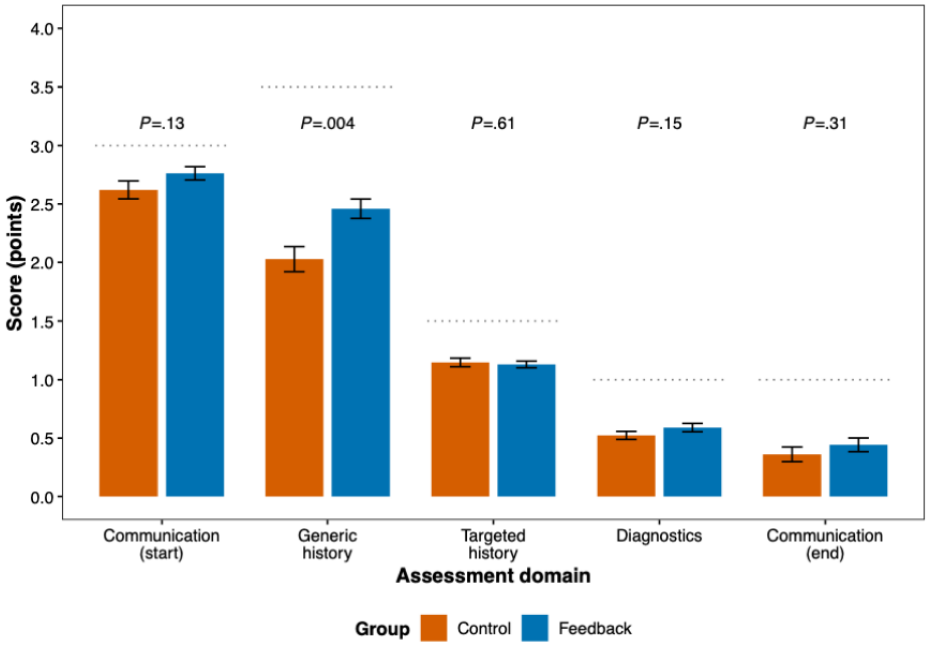


Individual OSCE Domain Comparisons

Domain-specific analysis using Mann-Whitney *U* tests revealed mixed results across the 5 domains of the OSCE-like assessment. Generic medical history was the only domain that demonstrated a statistically significant difference in favor of the integration of AI-generated feedback; this difference remained significant after applying Bonferroni correction for multiple comparisons (mean 2.46, SD 0.65 vs mean 2.03, SD 0.79 points; r=0.27; P=.004). No statistically significant differences were observed for communication at consultation start, targeted medical history taking, diagnostics and management reasoning, or communication at consultation end after correction for multiple testing. However, most subsets demonstrated a benefit in favor of the AI-generated feedback group. No statistically significant differences were noted between the assessors. Detailed results are presented in Table 2, Figure 2, and Figure S11 in Multimedia Appendix 1.

XSL•FO
RenderX

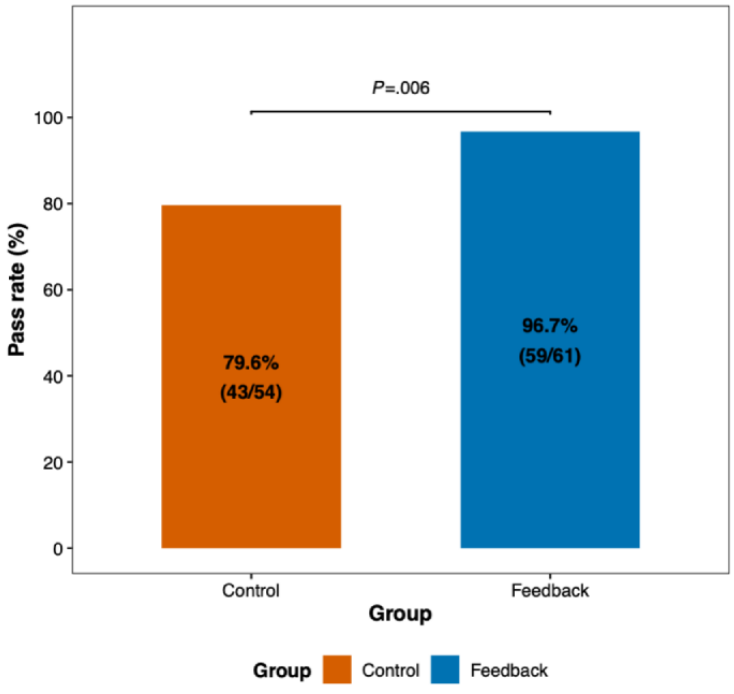
Figure 2. Domain-specific objective structured clinical examination performance across 5 domains of assessment: communication at consultation start, generic medical history, targeted medical history, diagnostics and management reasoning, and communication at consultation end. Bar plots display mean scores and SEs for students receiving artificial intelligence (AI)–generated feedback following interaction with virtual patients in the Social AI-Enhanced Robotic Interface (SARI) vs controls. Dotted horizontal lines indicate maximum scores for each domain. The *P* values from Mann-Whitney *U* tests are displayed above each domain.



Pass-Rate Comparisons

Using the predetermined pass threshold of 6 out of 10 points, students who had received AI-generated feedback demonstrated significantly higher pass rates compared with the control group (96.7% vs 79.6%; odds ratio 7.55, 95% CI 1.51-72.2; *P*=.006). This translates to an absolute risk reduction of 17.1 percentage points and a number needed to treat of 6. Detailed results are presented in Figure S12 in Multimedia Appendix 1 and Figure 3. No statistically significant differences in pass rates were noted between the 2 assessors (Figure S13 in Multimedia Appendix 1).

Figure 3. Bar plot illustrating objective structured clinical examination (OSCE) pass rates in students receiving artificial intelligence (AI)–generated feedback following interaction with virtual patients in the Social AI-Enhanced Robotic Interface (SARI) compared with controls. Numbers and percentages are shown within bars. Successful evaluation (pass score) was defined as a total OSCE score of ≥6 points.



XSL•FO
RenderX

Discussion

Principal Findings

This quasi-experimental study provides the first quantitative evidence that the addition of AI-generated feedback on top of expert-led seminars following social robotic VP interactions yields benefits regarding certain aspects of medical students' performance in subsequent OSCE-based assessments of CR, specifically pertaining to medical history–taking and communication skills. Students receiving AI-generated feedback achieved meaningfully higher total OSCE scores compared with those in the control group, representing a medium to large effect size, while having received feedback was shown to enhance pass rates. The domain-specific improvement in generic medical history suggests that targeted AI feedback effectively enhances systematic medical history–taking competencies, with effects translating from virtual environments to objective clinical assessments.

Our findings build on previous research from our group demonstrating the added value of AI-enhanced social robotic VPs for CR training in medical education [19-22]. Although those earlier studies relied on students' self-reported perceptions, this study evaluates whether the integration of AI-generated feedback within the SARI translates into measurable improvements in clinical performance compared with training without such feedback. The observed effect size compares favorably with systematic reviews, showing 58% to 72% of VP studies demonstrate positive effects on CR skill training, with particularly strong evidence when VPs include AI-generated feedback that targets specific CR components [6,7].

The factors for improvement likely relate to the immediacy and consistency of AI-generated feedback addressing systematic enquiries during training of medical history taking. Effective feedback principles emphasize the importance of being specific, timely, and actionable [13,14], with research demonstrating that immediate feedback enhances learning retention compared with delayed feedback [15]. As both groups received educational feedback through case-specific follow-up seminars led by consultant rheumatologists, the intervention evaluated the added value of AI-generated written feedback beyond this standard teaching. In addition to nonstructured feedback during these seminars, our AI-system provided additional standardized written feedback immediately following each VP case interaction, likely enhancing retention and skill transfer in the intervention group. While previous studies have shown positive results of integrating LLM-generated feedback in clinical dialogue and VP interaction [16,17], the integration within a social robotic VP platform (SARI) may have created favorable conditions for experiential learning, combining realistic patient interactions with analytical feedback to facilitate skill consolidation through reflective practice [28,33].

Improvements in generic medical history (35% of the total OSCE score), which includes past and current illnesses, lifestyle habits, medications, allergies, and family history, provide important insights into the mechanism of overall performance enhancement. This alignment between the content of the AI-generated feedback, which focused specifically on medical history taking, and the domain where benefit was observed strengthens the specificity of the intervention. Notably, the AI-generated feedback did not address diagnostic accuracy or management reasoning, which may explain why these domains did not show statistically significant differences. However, the sample size may also have limited the statistical power in comparative analyses to yield measurable benefit across domains. This study does not provide direct evidence for improvement in higher-order CR tasks such as differential diagnosis or hypothesis testing. According to a recent systematic review that investigated the role of VPs in providing feedback on CR skills, most feedback systems are designed to target information gathering, while other, more difficult-to-measure aspects of CR largely remain on the research agenda [6].

The domain weighting likely explains why students achieved significantly higher total OSCE scores despite improvements being concentrated to the generic medical history–taking domain. The VP activity design and the virtual outpatient clinic context, which place emphasis on medical history taking, likely create conditions where generic medical history–taking components are particularly responsive to structured feedback. This is in line with previous evidence showing that structured feedback can effectively improve communication skills in medical education [34].

An important consideration is that the AI-generated feedback in our modality is derived from the analysis of transcribed voice interactions between students and the SARI. This suggests that the observed benefits may extend to other voice-enabled VP platforms that facilitate natural conversational interactions. The integration of speech recognition and natural language processing technologies in VP systems may therefore represent an important component for effective AI-generated feedback implementation, warranting further research.

A known risk with LLM use in medical education is the potential for hallucination and factual errors [35]. Our 2-stage feedback algorithm was specifically designed to mitigate this risk. During the first stage, a structured assessment rubric is used to constrain the output of the LLM to predefined evaluation criteria, and during the second stage, feedback is generated based on this structured assessment rather than from unconstrained, noninstructed AI-generated content.

Limitations and Strengths

The quasi-experimental design with sequential allocation may have introduced selection bias, although baseline characteristics were well balanced between groups. Assessment was limited to 1 OSCE-based rheumatology evaluation, potentially limiting generalizability to other clinical settings and specialties. Although the coauthors portrayed standardized patients, the consultant rheumatologist assessors were blinded to intervention status, minimizing potential bias. Furthermore, there were no statistically significant differences in grading between the assessors, indicating that the standardized OSCE rubric and the clear instructions mitigated the risk of subjectivity for the intended purpose. The single-institution design and use of English for VP interactions and feedback while conducting assessments in Swedish may affect the external validity and represent methodological considerations for future research.

XSL•FO
RenderX

Although a statistically significant benefit was only observed in the generic medical history–taking domain, the AI-generated feedback group showed numerical advantages in most assessment domains. This pattern suggests that the study may have been underpowered to detect smaller effect sizes in secondary outcomes. Alternatively, a ceiling effect may have contributed, as both groups generally performed well across other areas of the evaluation rubric.

Subsequent studies should ideally conduct both VP training and assessment in the same language to eliminate potential confounding effects of language switching on clinical performance. This language alignment would provide more precise estimates of AI feedback effectiveness and better reflect real-world clinical education scenarios. Additionally, multicenter studies across different educational contexts and clinical specialties would enhance generalizability, while longitudinal assessment of skill retention and transfer to actual clinical practice would strengthen the evidence.

The study also has several important strengths, including the use of objective clinical performance measures, substantial sample size, and integration with established educational activities. This represents the first study demonstrating that AI-generated feedback in social robotic VPs translates to objectively measured improved medical history taking and CR performance in standardized OSCE-based evaluations.

Implications

These findings support investment in developing AI feedback systems within medical education curricula, specifically integrated in VP encounters. By systematically improving the quality of medical history–taking training, such systems could contribute to reducing diagnostic errors in clinical practice, where inadequate history taking has been identified as a leading contributor to such errors [10]. However, longitudinal studies following students into clinical practice are required to determine whether the observed benefit in OSCE performance translates to improved patient outcomes. The voice-based nature of student-VP interactions in our study suggests that AI-generated feedback systems may be particularly effective when integrated with conversational VP platforms that enable natural speech communication. This approach could apply across various VP technologies with voice interaction capabilities, potentially expanding beyond social robotic platforms.

Conclusions

This study demonstrates that AI-generated feedback following social robotic VP interactions yields significant benefit in medical students' CR performance, particularly in medical history taking, based on standardized OSCE-based assessments. The substantial effect size and greater pass rates among students who received AI-generated feedback provide evidence supporting the integration of AI feedback systems within social robotic platforms for enhancing medical history–taking skills in medical education. The observed differences across domains of the assessment may imply a potential value of domain-specific feedback approaches.

Acknowledgments

The authors would like to thank the medical students who participated in the study and the medical staff at the Division of Rheumatology at the Karolinska University Hospital, Stockholm, Sweden. The authors used ChatGPT (OpenAI) for language refinement and grammar checking in specific sections of the manuscript during the writing process. All scientific content, study design, data analysis, statistical interpretation, and conclusions were developed entirely by the authors. The authors take full responsibility for the accuracy, integrity, and scientific validity of all content in this manuscript.

Funding

This work was supported by grants from Region Stockholm ALF Pedagogy (FoUI-977096), Karolinska Institutet Pedagogical Project Funding (FoUI-964139), the Swedish Rheumatism Association (R-1013624), King Gustaf V's 80-year Foundation (FAI-2023-1055), Swedish Society of Medicine (SLS-974449), Nyckelfonden (OLL-1023269), Professor Nanna Svartz Foundation (2021-00436), Ulla and Roland Gustafsson Foundation (2024-43), Region Stockholm (FoUI-1004114), and Karolinska Institutet. The funding sources had no role in the design or conduct of the study; collection, management, analysis, or interpretation of the data; preparation, review, or approval of the manuscript; or decision to submit the manuscript for publication.

Data Availability

The anonymized datasets generated and analyzed during this study are available from the corresponding author upon reasonable request. Access to data will be granted following appropriate ethical review and data sharing agreements and will require completion of a data transfer agreement and approval from the Swedish Ethical Review Authority, as per Swedish data protection regulations and the European General Data Protection Regulation.

Authors' Contributions

Conceptualization: AB, MR, SE, C Georg, GS, IP
Data Curation: AB, JS, WI, C Gentline, VH, AMH, BJ, FE, IP
Formal analysis: AB, MR, SE, C Georg, IP

XSL•FO
RenderX

Funding acquisition: IP
Investigation: AB, JS, WI
Methodology: AB, MR, SE, C Georg, GS, IP
Project administration: AB, JS, WI, C Gentline, AMH, IP
Supervision: MR, SE, C Georg, GS, IP
Visualization: AB
Writing—original draft: AB (lead), IP (supporting)
Writing—review and editing: AB, MR, SE, JS, WI, C Georg, C Gentline, VH, AMH, BJ, FE, GS, IP

Conflicts of Interest

IP has received research funding and/or honoraria from Amgen, AstraZeneca, Aurinia, BMS, Elli Lilly, Gilead, GSK, Janssen, Novartis, Otsuka, and Roche. GS, who is a cofounder and chief scientist at Furhat Robotics, provided technical expertise regarding the robotic platform but was not involved in study design, data collection, data analysis, or interpretation of results. All other authors declare no other conflicts of interest.

Multimedia Appendix 1
Data on participant recruitment and allocation flowchart, objective structured clinical examination (OSCE)–like assessment rubric, structured history–taking guide with example questions, artificial intelligence (AI)–generated feedback algorithm and prompt configurations for the Social AI-enhanced Robotic Interface (SARI), example AI-generated feedback output, SARI interaction prompt example, translated instructions to students and actors for the OSCE-like assessment, violin plots comparing total and domain-specific OSCE scores between assessors, and tables comparing pass or fail frequencies by feedback condition and by assessor.
[DOCX File , 692 KB - mededu_v12i1e90368_app1.docx]

References

1. Eva KW. What every teacher needs to know about clinical reasoning. Med Educ 2005 Jan;39(1):98-106. [doi: 10.1111/j.1365-2929.2004.01972.x] [Medline: 15612906]
2. Norman G. Research in clinical reasoning: past history and current trends. Med Educ 2005 Apr;39(4):418-427. [doi: 10.1111/j.1365-2929.2005.02127.x] [Medline: 15813765]
3. Kononowicz AA, Woodham LA, Edelbring S, Stathakarou N, Davies D, Saxena N, et al. Virtual patient simulations in health professions education: systematic review and meta-analysis by the digital health education collaboration. J Med Internet Res 2019 Jul 02;21(7):e14676 [FREE Full text] [doi: 10.2196/14676] [Medline: 31267981]
4. Cook DA, Erwin PJ, Triola MM. Computerized virtual patients in health professions education: a systematic review and meta-analysis. Acad Med 2010 Oct;85(10):1589-1602. [doi: 10.1097/ACM.0b013e3181edfe13] [Medline: 20703150]
5. Hege I, Kononowicz AA, Adler M. A clinical reasoning tool for virtual patients: design-based research study. JMIR Med Educ 2017 Nov 02;3(2):e21 [FREE Full text] [doi: 10.2196/mededu.8100] [Medline: 29097355]
6. Jay R, Sandars J, Patel R, Leonardi-Bee J, Ackbarally Y, Bandyopadhyay S, et al. The use of virtual patients to provide feedback on clinical reasoning: a systematic review. Acad Med 2025 Feb 01;100(2):229-238. [doi: 10.1097/ACM.0000000000005908] [Medline: 39485118]
7. Plackett R, Kassianos AP, Mylan S, Kambouri M, Raine R, Sheringham J. The effectiveness of using virtual patient educational tools to improve medical students' clinical reasoning skills: a systematic review. BMC Med Educ 2022 May 13;22(1):365 [FREE Full text] [doi: 10.1186/s12909-022-03410-x] [Medline: 35550085]
8. Breazeal C. Designing Sociable Robots. Cambridge, MA: MIT Press; 2004.
9. Belpaeme T, Kennedy J, Ramachandran A, Scassellati B, Tanaka F. Social robots for education: a review. Sci Robot 2018 Aug 15;3(21):eaat5954. [doi: 10.1126/scirobotics.aat5954] [Medline: 33141719]
10. Hampton JR, Harrison MJ, Mitchell JR, Prichard JS, Seymour C. Relative contributions of history-taking, physical examination, and laboratory investigation to diagnosis and management of medical outpatients. Br Med J 1975 May 31;2(5969):486-489 [FREE Full text] [doi: 10.1136/bmj.2.5969.486] [Medline: 1148666]
11. Graber ML, Franklin N, Gordon R. Diagnostic error in internal medicine. Arch Intern Med 2005 Jul 11;165(13):1493-1499. [doi: 10.1001/archinte.165.13.1493] [Medline: 16009864]
12. Committee on Diagnostic Error in Health Care, Board on Health Care Services, Institute of Medicine, The National Academies of Sciences, Engineering, and Medicine, Balogh EP, Miller BT, et al. Improving Diagnosis in Health Care. Washington, DC: National Academies Press (US); 2015.
13. Hattie J, Timperley H. The power of feedback. Rev Educ Res 2007 Mar 01;77(1):81-112. [doi: 10.3102/003465430298487]
14. Boud D, Molloy E. Rethinking models of feedback for learning: the challenge of design. Assess Eval High Educ 2013;38(6):698-712. [doi: 10.1080/02602938.2012.691462]
15. van de Ridder JM, Stokking KM, McGaghie WC, ten Cate OT. What is feedback in clinical education? Med Educ 2008 Feb;42(2):189-197. [doi: 10.1111/j.1365-2923.2007.02973.x] [Medline: 18230092]

XSL•FO
RenderX

16. Cook DA, Overgaard J, Pankratz VS, Del Fiol G, Aakre CA. Virtual patients using large language models: scalable, contextualized simulation of clinician-patient dialogue with feedback. J Med Internet Res 2025 Apr 04;27:e68486 [FREE Full text] [doi: 10.2196/68486] [Medline: 39854611]
17. Holderried F, Stegemann-Philipps C, Herrmann-Werner A, Festl-Wietek T, Holderried M, Eickhoff C, et al. A language model-powered simulated patient with automated feedback for history taking: prospective study. JMIR Med Educ 2024 Aug 16;10:e59213 [FREE Full text] [doi: 10.2196/59213] [Medline: 39150749]
18. Furlan R, Gatti M, Menè R, Shiffer D, Marchiori C, Giaj Levra A, et al. A natural language processing-based virtual patient simulator and intelligent tutoring system for the clinical diagnostic process: simulator development and case study. JMIR Med Inform 2021 Apr 09;9(4):e24073 [FREE Full text] [doi: 10.2196/24073] [Medline: 33720840]
19. Borg A, Georg C, Jobs B, Huss V, Waldenlind K, Ruiz M, et al. Virtual patient simulations using social robotics combined with large language models for clinical reasoning training in medical education: mixed methods study. J Med Internet Res 2025 Mar 03;27:e63312 [FREE Full text] [doi: 10.2196/63312] [Medline: 40053778]
20. Borg A, Jobs B, Huss V, Gentline C, Espinosa F, Ruiz M, et al. Enhancing clinical reasoning skills for medical students: a qualitative comparison of LLM-powered social robotic versus computer-based virtual patients within rheumatology. Rheumatol Int 2024 Dec 16;44(12):3041-3051. [doi: 10.1007/s00296-024-05731-0] [Medline: 39412574]
21. Borg A, Parodis I, Skantze G. Creating virtual patients using robots and large language models: a preliminary study with medical students. In: Proceedings of the Companion of the 2024 ACM/IEEE International Conference on Human-Robot Interaction. 2024 Presented at: HRI '24; March 11-15, 2024; Boulder, CO. [doi: 10.1145/3610978.3640592]
22. Borg A, Schiött J, Ivegren W, Gentline C, Huss V, Hugelius A, et al. AI-enhanced social robotic versus computer-based virtual patients for clinical reasoning training in medical education: observational crossover cohort study. J Med Internet Res 2025 Nov 27;27:e82541 [FREE Full text] [doi: 10.2196/82541] [Medline: 41309100]
23. Harden RM. What is an OSCE? Med Teach 1988 Jul 03;10(1):19-22. [doi: 10.3109/01421598809019321] [Medline: 3221760]
24. Khan KZ, Gaunt K, Ramachandran S, Pushkar P. The objective structured clinical examination (OSCE): AMEE Guide no. 81. Part II: organisation and administration. Med Teach 2013 Aug 22;35(9):e1447-e1463. [doi: 10.3109/0142159x.2013.818635]
25. Daniel M, Rencic J, Durning SJ, Holmboe E, Santen SA, Lang V, et al. Clinical reasoning assessment methods: a scoping review and practical guidance. Acad Med 2019 Jun;94(6):902-912. [doi: 10.1097/ACM.0000000000002618] [Medline: 30720527]
26. TREND statement checklist. Centers for Disease Control and Prevention. 2004. URL: https://stacks.cdc.gov/view/cdc/149677 [accessed 2026-02-27]
27. Virtual Interactive Case (VIC): perioperative interactive education. Perioperative Interactive Education. URL: https://pie.med.utoronto.ca/VIC/index.htm [accessed 2024-04-17]
28. Edelbring S, Parodis I, Lundberg IE. Increasing reasoning awareness: video analysis of students' two-party virtual patient interactions. JMIR Med Educ 2018 Feb 27;4(1):e4 [FREE Full text] [doi: 10.2196/mededu.9137] [Medline: 29487043]
29. Schiött J, Ivegren W, Borg A, Parodis I, Skantze G. Using LLMs to grade clinical reasoning for medical students in virtual patient dialogues. In: Proceedings of the 26th Annual Meeting of the Special Interest Group on Discourse and Dialogue. 2025 Presented at: SIGDIAL 2025; August 25-27, 2025; Avignon, France URL: https://aclanthology.org/2025.sigdial-1.56/
30. Introducing ChatGPT. OpenAI. URL: https://openai.com/blog/chatgpt/ [accessed 2024-06-04]
31. Al Moubayed S, Beskow J, Skantze G, Granström B. Furhat: a back-projected human-like robot head for multiparty human-machine interaction. In: Esposito A, Esposito AM, Vinciarelli A, Hoffmann R, Müller VC, editors. Cognitive Behavioural Systems. Berlin, Germany: Springer; 2012.
32. ten Cate O. Entrustability of professional activities and competency-based training. Med Educ 2005 Dec;39(12):1176-1177. [doi: 10.1111/j.1365-2929.2005.02341.x] [Medline: 16313574]
33. Yardley S, Teunissen PW, Dornan T. Experiential learning: AMEE Guide No. 63. Med Teach 2012;34(2):e102-e115. [doi: 10.3109/0142159X.2012.650741] [Medline: 22289008]
34. Lai MM, Roberts N, Mohebbi M, Martin J. A randomised controlled trial of feedback to improve patient satisfaction and consultation skills in medical students. BMC Med Educ 2020 Aug 20;20(1):277 [FREE Full text] [doi: 10.1186/s12909-020-02171-9] [Medline: 32819352]
35. Thirunavukarasu AJ, Ting DS, Elangovan K, Gutierrez L, Tan TF, Ting DS. Large language models in medicine. Nat Med 2023 Aug 17;29(8):1930-1940. [doi: 10.1038/s41591-023-02448-8] [Medline: 37460753]

Abbreviations

AI: artificial intelligence
CR: clinical reasoning
LLM: large language model
OSCE: objective structured clinical examination
SARI: Social AI-Enhanced Robotic Interface

XSL•FO
RenderX

VP: virtual patient

Edited by B Lesselroth; submitted 26.Dec.2025; peer-reviewed by B Giglio, L Mosca; comments to author 10.Feb.2026; revised version received 15.Feb.2026; accepted 25.Feb.2026; published 25.Mar.2026.

Please cite as:
Borg A, Schiött J, Ivegren W, Gentline C, Huss V, Hugelius AM, Jobs B, Ruiz M, Edelbring S, Georg C, Skantze G, Parodis I
AI-generated Feedback Following Social Robotic Virtual Patient Interactions and Medical Student Performance: Nonrandomized Quasi-Experimental Study
JMIR Med Educ 2026;12:e90368
URL: https://mededu.jmir.org/2026/1/e90368
doi:10.2196/90368
PMID:

©Alexander Borg, Jonathan Schiött, William Ivegren, Cidem Gentline, Viking Huss, Anna Margareta Hugelius, Benjamin Jobs, Mini Ruiz, Samuel Edelbring, Carina Georg, Gabriel Skantze, Ioannis Parodis. Originally published in JMIR Medical Education (https://mededu.jmir.org), 25.Mar.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Education, is properly cited. The complete bibliographic information, a link to the original publication on https://mededu.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

User Acceptance of an AI-Powered Medical History–Taking Training System Among Undergraduate Medical Students: Mixed Methods Study

Yang Liu[1*], MM; Yiying Zhu[1*], MSc; Chujun Shi[1*], Prof Dr; Xian Lu[2], BSc; Liping Wu[1], Prof Dr; Minghui Yue[1], MPH; Xiaolin Hong[1], BMed; Oudong Xia[3*], Prof Dr, MD; Weishan Zhang[1*], MSc

[1]Medical Simulation Center, Shantou University Medical College, Shantou, China

[2]Department of Medical Physics and Informatics, Shantou University Medical College, Shantou, China

[3]Shantou University Medical College, Shantou, China

[*]these authors contributed equally

Corresponding Author:
Weishan Zhang, MSc
Medical Simulation Center
Shantou University Medical College
No. 22 Xinling Road
Shantou, 515041
China
Phone: 86 754 88900459
Email: zwsyyl@stu.edu.cn

Abstract

Background: AI-powered virtual patient systems provide medical students with repeatable practice environments for history-taking training. However, user acceptance of such systems and the experience dimensions associated with that acceptance lack mixed methods evidence.

Objective: This study aimed to (1) examine the associations of system experience and learning experience/intrinsic motivation with overall acceptance among undergraduate medical students using an AI-powered medical history-taking training and evaluation system (AMTES), (2) explore user experience patterns through open-ended questions, and (3) integrate quantitative and qualitative findings to inform system refinement and pedagogical implementation.

Methods: A cross-sectional convergent mixed methods design was used. A total of 66 undergraduate medical students at a Chinese medical college completed a postuse questionnaire after AMTES training. The primary outcome was an overall acceptance composite combining use intention, recommendation intention, and overall satisfaction. Associations with system experience and learning experience/intrinsic motivation were examined using linear regression with heteroscedasticity-consistent SEs (type 3) and bootstrap CIs. Sensitivity analyses included covariate adjustment, single-outcome models, a fractional logit model, and content-overlap sensitivity checks for the system-experience composite. Open-ended responses were analyzed using codebook-oriented thematic analysis and integrated with quantitative findings through a joint display.

Results: Both system experience and learning experience/intrinsic motivation were positively associated with overall acceptance (standardized β=.526; P<.001 and standardized β=.377; P=.002, respectively; R^2=0.662). Sensitivity analyses supported the robustness of both positive associations. Qualitative analysis showed that the most frequently nominated benefits clustered within the “Practice Accessibility and Feedback Support” theme, particularly self-directed practice (28/66, 42.4%) and immediate feedback (25/66, 37.9%). Refinement priorities clustered around 4 dialogue-quality and assessment dimensions, including semantic understanding, contextual consistency, conversational naturalness, and scoring logic, each cited by 21%-35% of respondents. The mixed methods joint display indicated contextual (same-source) alignment on dialogue and scoring concerns and suggested that perceived scoring-feedback discrepancies may be associated with lower student trust in the feedback function. This cross-dimensional pattern was not apparent from the quantitative model alone.

Conclusions: In this exploratory cohort of undergraduate medical students, both system interaction quality and perceived learning value were positively associated with overall acceptance of AMTES, with system interaction quality showing the stronger association within the study’s measurement specification. Dialogue coherence, semantic understanding, and scoring-feedback

XSL•FO
RenderX

alignment emerged as the most frequently nominated refinement priorities and are plausible candidate targets for improving acceptance-related perceptions. This study emphasizes implementation-level acceptance rather than solely technical reliability or educational effectiveness. Interaction-quality problems may be associated with less favorable acceptance even when learning value is recognized. These findings may inform system refinement priorities and the curricular integration of AI-powered history-taking training systems, while larger multicenter and longitudinal studies are needed to examine whether improvements in interaction quality translate into gains in learner acceptance and downstream training outcomes.

(JMIR Med Educ 2026;12:e101425) doi:10.2196/101425

KEYWORDS

artificial intelligence; Kirkpatrick framework; medical education; medical history-taking; mixed methods; user acceptance; virtual patient

Introduction

History-taking is a fundamental clinical skill, and its quality directly affects diagnostic accuracy [1,2] and subsequent clinical reasoning [1]. For medical students, mastering a complete history-taking process and developing fluent clinical interview skills are core tasks in the transition from classroom learning to clinical practice [3]. However, achieving this transition has become increasingly difficult, as the erosion of bedside teaching and the decline of core clinical skills have been recognized as growing challenges in medical education [4]. Simulation-based instruction and standardized patient (SP) encounters have been proposed as effective alternatives to address this gap, with evidence supporting their value in improving student confidence and clinical competence [5,6]; however, their systematic implementation remains constrained by time and resource demands, limiting opportunities for repeated practice and structured feedback [5].

Virtual patient (VP) systems have emerged as a promising approach to address these challenges, offering interactive and repeatable practice environments suited to clinical reasoning training [7]. Existing evidence supports their educational value: a systematic review of conversational VP interventions found consistent improvements in history-taking and clinical reasoning competencies across multiple studies [8], and a single-cohort intervention study demonstrated statistically significant gains in both self-reported competence and confidence, corroborated by improved objective structured clinical examination performance in medical and surgical history-taking [9]. Notably, that same systematic review found that student satisfaction tended to be higher when VP systems incorporated AI and natural language processing to enable more realistic conversational interaction [8], suggesting that AI-enhanced dialogue may be associated with more favorable learner experiences compared with scripted or menu-driven formats. Building on this trajectory, recent advances in large language models have enabled a new generation of AI-powered virtual patient (AI-VP) systems capable of open-ended natural language interaction and automated feedback [10-12]. However, empirical evidence specifically evaluating AI-VP systems in history-taking education remains limited and preliminary, and a recent scoping review of AI applications in medical education underscored both the promise of these technologies and the need for rigorous evaluation [13]. Sustainable adoption therefore depends not only on technical performance and educational effectiveness, but also on user acceptance [14,15].

From a training-evaluation perspective, Kirkpatrick's 4-level model provides a useful framework for considering how learners respond to educational interventions. Its first level, "Reaction," focuses on learners' immediate responses to training, including satisfaction and perceived value, and is sometimes considered an early indicator of subsequent learning engagement, though the assumption that Level 1 reactions predict higher-level outcomes has been critically examined in medical education [16,17]. Nevertheless, Level 1 evaluation remains a recognized starting point for understanding learner responses to novel educational technologies, particularly in early-stage implementations where downstream outcome data are not yet available. The Technology Acceptance Model (TAM) and related frameworks such as the Unified Theory of Acceptance and Use of Technology (UTAUT) emphasize that perceived usefulness, ease of use or effort expectancy, and related performance beliefs are central determinants of technology acceptance [18,19]. A recent systematic review of TAM in medical education suggested considering broader acceptance frameworks, including TAM2 extensions, for educational technologies in context [14]. Although mixed methods work on trust and acceptance of AI are emerging in medicine [15], research that combines quantitative associations with qualitative exploration of underlying experiential patterns in medical education remains limited. In this study, Kirkpatrick Level 1 provides the curricular-evaluation positioning, whereas TAM/UTAUT-style acceptance logic informs the operational acceptance outcome (use intention, recommendation intention, and overall satisfaction). We treated acceptance as an exploratory, theory-informed construct rather than as a formal test or validated adaptation of TAM or UTAUT, and the experience composites correspond only partially to established acceptance constructs, as detailed in the Methods.

Our team has previously conducted a series of studies on the self-developed AI-powered medical history-taking training and evaluation system (AMTES): study 1 demonstrated the consistency between AMTES automated scoring and human scoring for medical history-taking [20], and study 2 provided preliminary evidence for its educational effectiveness in a real-world teaching environment [21]. Although study 1 reported favorable overall user approval as a secondary finding, neither prior study was designed to explain the determinants of acceptance or identify user-experienced priorities for refinement.

XSL•FO
RenderX

Prior work on AI-powered history-taking training systems appears limited, particularly regarding studies that simultaneously examine which experience dimensions are associated with user acceptance and explore the experiential patterns underlying those relationships.

As the third study in this series, this study extends the evidence base to user acceptance, positioned as a Level 1–aligned evaluation of learner reactions after structured curricular use. A convergent mixed methods design was adopted: the quantitative strand estimated associations between 2 theory-informed, empirically constructed experience composites—system experience (interaction quality and ease of use) and learning experience/intrinsic motivation (perceived learning value)—and overall acceptance, while open-ended responses contextualized those associations and identified concrete refinement targets. This study aimed to (1) examine how these 2 dimensions were associated with overall acceptance, (2) explore user-reported experiential patterns underlying perceived benefits and refinement priorities, and (3) integrate quantitative and qualitative findings to generate evidence-informed candidate recommendations for system refinement and pedagogical implementation.

Methods

Study Design

This cross-sectional study used a convergent mixed methods design, with quantitative and qualitative data collected through the same postuse questionnaire. Reporting followed the STROBE (Strengthening the Reporting of Observational Studies in Epidemiology) statement [22] and the CHERRIES (Checklist for Reporting Results of Internet E-Surveys) [23], and the qualitative and integration components drew on established thematic analysis [24] and mixed methods guidance [25], where applicable.

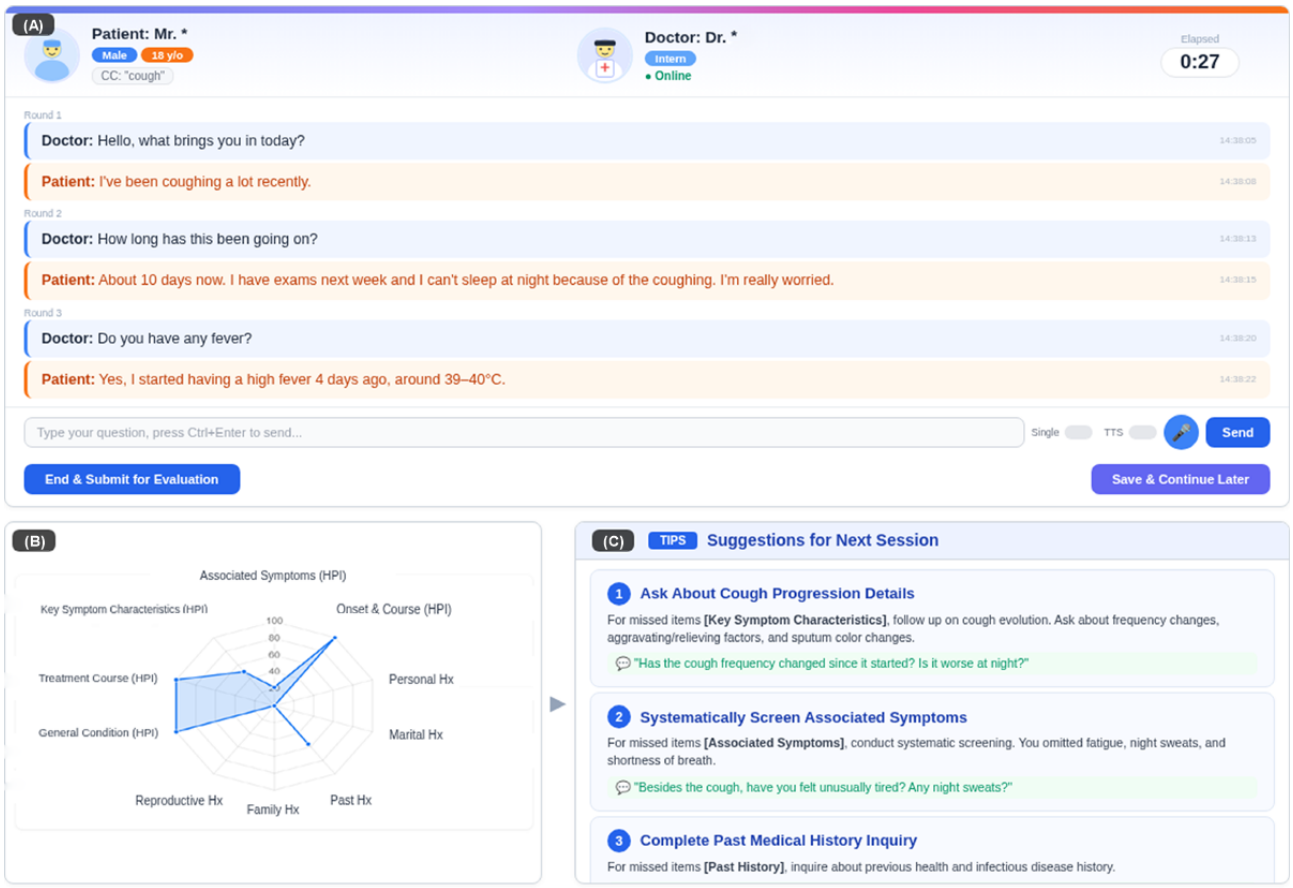
Study Setting and Participants

The study was conducted at Shantou University Medical College from October to November 2025. Students used a self-developed AMTES to complete history-taking training for 3 assigned cases through text or voice interactions with virtual patients. After each session, they received automated scores, session summaries, and personalized feedback, including feedback on omitted or insufficiently explored aspects of the patient's history and targeted suggestions for improvement. The AMTES platform was the same system evaluated in studies 1 and 2 [20,21], with the underlying large language model upgraded from DeepSeek-V2.5 to DeepSeek-V3 while the core training workflow and scoring logic remained unchanged.

Participants were all undergraduate medical students enrolled in the Diagnostics course, who completed a postuse questionnaire after using AMTES for history-taking training as part of the course curriculum. Practice-log data showed 272 completed sessions from 66 students across 3 cases. Most students (62/66, 93.9%) completed all 3 required cases, and 30/66 (45.5%) undertook additional voluntary practice. Each session involved a mean of 108.0 (SD 37.6) dialogue turns and lasted 31.1 (SD 18.0) minutes, yielding a cumulative mean exposure of 128.4 minutes and 445.2 dialogue turns per student. The questionnaire was collected after the practice period ended, so participants responded after repeated hands-on exposure to the system; however, the anonymous questionnaire contained no identifiers for linking individual survey responses with practice-log variables. An example interface and feedback display of AMTES is shown in Figure 1.

XSL•FO
RenderX

Figure 1. Example AI-powered medical history-taking training and evaluation system interface and feedback components available to students during and after a history-taking session. (A) The multiturn dialogue interface where students interact with a large language model–powered virtual patient through text or voice input. (B) A radar chart visualizing performance across multiple history-taking dimensions. (C) AI-generated personalized feedback identifying missed history items and providing specific suggested follow-up questions.



Recruitment, Sampling, and Eligibility Criteria

A whole-cohort sampling approach was used. Eligible students were all undergraduate medical students enrolled in the Diagnostics course during the study period who participated in the designated AMTES history-taking training. After the AMTES training and the summative history-taking examination had been completed and all course grades had been finalized and officially recorded, the study was introduced to the cohort during a subsequent regularly scheduled class session held several days later by a member of the research team who was also involved in course teaching. Students were informed of the study purpose and of the voluntary and anonymous nature of participation, and were told that the decision to participate or to decline would not affect their grades, course standing, or relationship with instructors. Because the questionnaire was anonymous and contained no identifiers, individual responses could not be linked to students; and because all grades had already been finalized and officially recorded, participation could not affect academic outcomes. Those who chose to participate completed the anonymous web-based questionnaire individually on their personal devices through a class messaging link during the same supervised class session; no reminders or incentives were used, and instructors did not monitor individual responses. Questionnaires with missing core outcome items or response patterns suggesting invalid completion (eg, identical responses across all conceptually distinct items) would have been excluded. None met these criteria. All 66 students had completed at least 1 assigned history-taking case, so every eligible student had substantive hands-on experience with AMTES on which to base their responses; the 4 students who did not complete all 3 assigned cases were retained because their completed sessions reflected meaningful engagement with the system. Thus, all 66 eligible students completed the questionnaire, yielding a 100% response rate.

Questionnaire Design and Development

The questionnaire consisted of a baseline characteristics section; item sets on system experience, learning experience/intrinsic motivation, and overall acceptance; and 2 open-ended questions (C1 and C2), included to capture perceived strengths and areas for improvement not fully reflected in the structured item sets. C1 asked students to describe the greatest advantage of the system or what helped them most (“您认为这个系统最大的优点或给您带来最大帮助的是什么?”), and C2 asked them to identify the one aspect of the system that most needed improvement (“您认为这个系统最需要改进的一个方面是什么?”). Both prompts were deliberately valence-leading—C1 toward perceived benefits and C2 toward a single improvement priority—so the resulting theme frequencies are interpreted as respondent-nominated salience rather than prevalence or importance (see the Qualitative Analysis section). The baseline characteristics section captured participants’ age, sex, previous history-taking training experience, and previous experience with

XSL•FO
RenderX

comparable AI systems. Age was entered numerically from the questionnaire categories, with the top category ("24 or above") coded as 24, and was used together with sex for sample description and covariate-adjusted sensitivity analyses. Items on previous experience were retained for descriptive purposes only.

The quantitative item sets were each informed by established English-language scales and adapted to maintain contextual relevance and reduce redundancy. Candidate system-experience items (Block A; 12 items entering exploratory factor analysis [EFA], 5-point Likert scale) combined System Usability Scale (SUS)–informed usability items [26,27] with author-developed dialogue-quality items designed to capture system usability, interaction fluency, and dialogue experience. Learning experience/intrinsic motivation items (Block B; 11 items, 7-point Likert scale) were informed by selected Intrinsic Motivation Inventory subscales [28] covering interest/enjoyment, perceived competence, effort/importance, pressure/tension, perceived choice, and value/usefulness. Overall acceptance was assessed using 3 items rated on a 5-point Likert scale, covering intention to use, intention to recommend, and overall satisfaction.

To clarify the theoretical positioning, the acceptance model was treated as an exploratory, theory-informed framework rather than a formal test or validated adaptation of TAM or UTAUT, and the experience composites correspond only partially to established acceptance constructs. System experience (SUS-informed usability items together with author-developed dialogue-quality items) corresponds partially to perceived ease of use and effort expectancy as well as to system- and interaction-quality perceptions; learning experience/intrinsic motivation (Intrinsic Motivation Inventory subscales together with perceived value and confidence) corresponds partially to perceived usefulness and performance expectancy as well as to intrinsic motivation; and overall acceptance (intention to use, intention to recommend, and satisfaction) corresponds to behavioral intention and to Kirkpatrick Level 1 reaction and satisfaction. Because both composites are multifaceted and were constructed empirically through EFA, these correspondences denote partial conceptual alignment rather than validated construct equivalence (the mapping is summarized in Multimedia Appendix 1); accordingly, TAM and UTAUT were used as sensitizing frameworks to organize acceptance-related concepts, not as confirmatory latent-variable models to be statistically tested in this sample. The UTAUT constructs of social influence and facilitating conditions were not modeled: in a single, mandatory curricular deployment, these were expected to show limited between-student variability and fell outside the exploratory Level 1 scope. We therefore treat them as boundaries of this framework rather than as irrelevant factors, and note them as priorities for future, more fully specified acceptance models.

All items were translated into Chinese using a forward translation, cross-checking, and iterative revision process to improve semantic equivalence and expression consistency. The questionnaire was presented in bilingual format (Chinese as primary, English original as reference) to facilitate local readability while preserving item traceability during administration.

Online Survey Administration

The questionnaire was administered as an anonymous closed web-based survey. Access was restricted to the eligible cohort and limited to 1 submission per respondent. The survey was distributed in person during a subsequent class session held several days after the summative history-taking examination had been completed and all grades had been finalized and officially recorded, ensuring that participation and responses could not affect academic outcomes. Although completion was voluntary, all core items were mandatory for those who chose to participate. Respondents could review and revise their answers before submission. Students accessed the survey individually on personal devices during a supervised on-site session, without course instructors observing responses in real time. The exported dataset was deidentified before analysis.

Data Processing and Scoring

Reverse-worded items were reverse-scored before analysis. All 5-point and 7-point Likert items were then linearly rescaled to a 0-1 range to facilitate comparability across scales. Composite scores were calculated as the mean of the retained items within each set. Overall acceptance was calculated as the mean of intention to use, intention to recommend, and overall satisfaction.

Statistical Analysis

Overview

All statistical analyses were performed using Python (version 3.12). All inferential tests were 2-sided, with the significance level set at α=.05. Upon data verification, the key variables used in this study had no missing values. Continuous variables were described using means and SDs, and categorical variables were reported as frequencies and percentages.

EFA

Factorability and Sampling Adequacy

Separate EFAs were conducted for Block A and Block B item sets to inform composite construction; the EFA was used for sample-specific data reduction to construct composites, not as formal scale validation. Factorability was assessed using Kaiser-Meyer-Olkin (KMO) measures of sampling adequacy [29] and Bartlett tests of sphericity [30]. Item evaluation considered item-level measure of sampling adequacy (MSA), factor loadings estimated using least-squares (minimum residual) factor extraction, and the consistency between item content and the intended construct. Items with MSA of <0.50 or factor loading of <0.40 were excluded from the final composite unless there was a strong conceptual rationale for retention. Factor retention for the final retained item sets was guided by parallel analysis [31,32]. Cronbach α was calculated for the final Block A and Block B item sets to assess internal consistency. Composite scores were calculated as the mean of the final retained items within each identified factor and used in subsequent regression analyses.

Supplementary Common-Method-Variance Check

Because quantitative data were from a single time point self-report questionnaire, Harman single-factor test was used

XSL•FO
RenderX

as a supplementary common method bias check for common method variance [33]. Given that the acceptance items may reflect variance shared with the experience items, the Harman test excluding acceptance items was prioritized in interpretation.

Regression Models

To examine the association between experience-related measures and overall acceptance, we fitted a linear regression model with the overall acceptance composite as the dependent variable, and the retained composite scores derived from the EFA-informed item sets as independent variables. Standardized regression coefficients and *P* values based on heteroscedasticity-consistent SEs (type 3) [34] were reported, along with percentile bootstrap 95% CIs based on 5000 replications. Multicollinearity was assessed using variance inflation factors. Unique and shared variance decomposition was conducted as supplementary analyses.

Several sensitivity analyses were conducted to assess the robustness and directional consistency of the primary findings, including adjustment for age and sex, single-outcome models using use intention, recommendation intention, and overall satisfaction as separate dependent variables, and a fractional logit model for the bounded outcome [35]. A post hoc content-overlap check was also conducted by excluding the most overtly evaluative Block A items from the system-experience composite; Block A items were classified by the research team as acceptance-proximal (ease of use, confidence in use, feedback and report helpfulness, consultation fluency, and response usefulness) versus lower-overlap (functional integration, virtual-patient realism, nonmechanical interaction, and clinical-scenario realism) according to their semantic proximity to the acceptance items, and the item-level basis for this classification is reported in Multimedia Appendix 2.

Qualitative Analysis

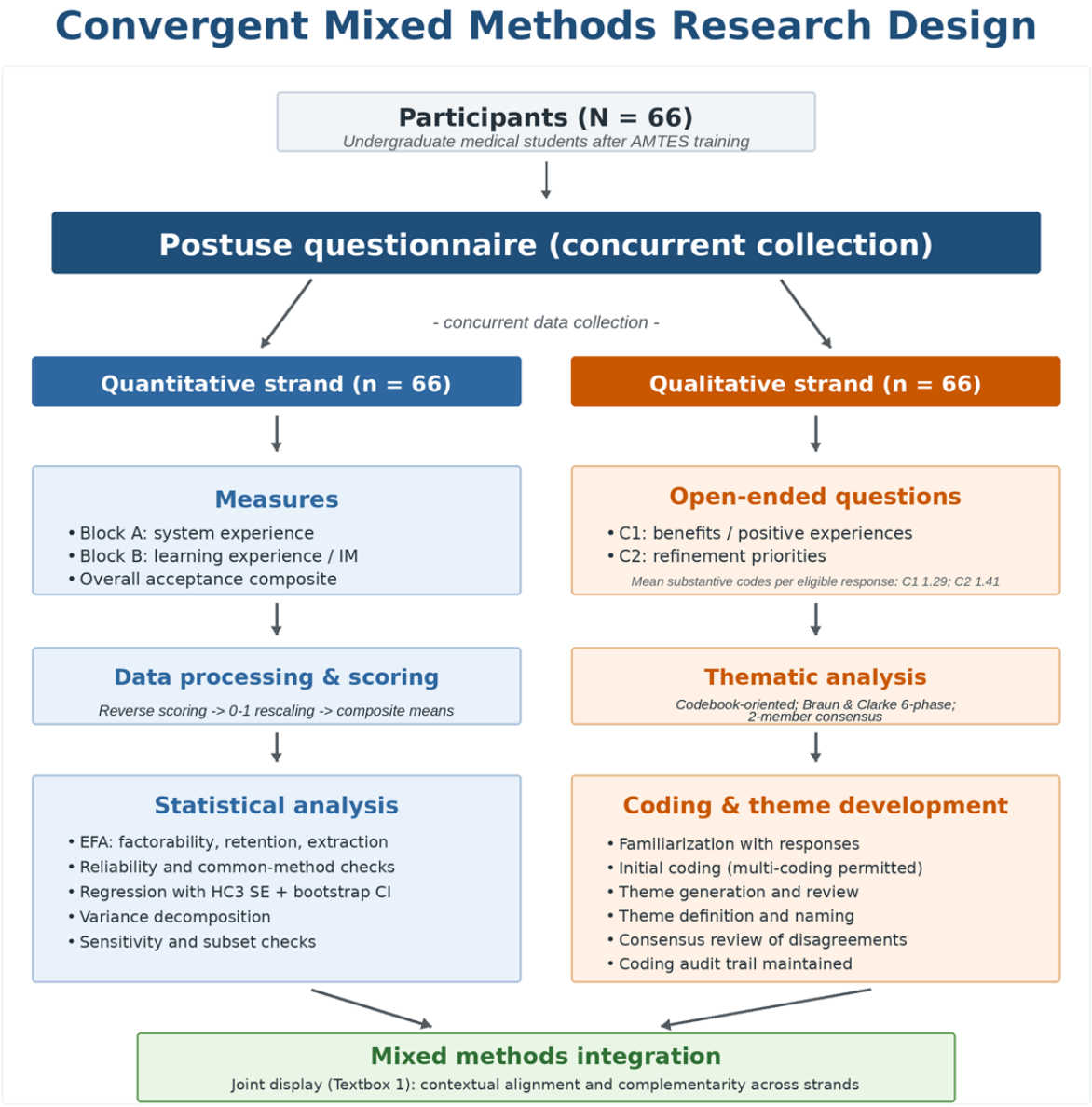
Text responses to C1 and C2 were analyzed using codebook-oriented thematic analysis adapted for brief qualitative survey responses, drawing on Braun and Clarke's [24] 6-phase approach as an organizing guide rather than as reflexive thematic analysis [36]. Because responses were brief written comments rather than interview transcripts, the analysis aimed to identify experiential patterns and optimization priorities rather than to develop deep process-oriented theory. Two AMTES research-team members jointly developed and refined the codebook and resolved coding disagreements through consensus; multicoding was permitted when a response described more than one distinct benefit or improvement direction; thus, although C2 asked respondents to nominate a single improvement priority, responses that raised more than one distinct actionable issue were multicoded for descriptive salience reporting rather than for prevalence estimation. Responses without an identifiable substantive theme were assigned a mutually exclusive "Vague/unclear" category. Translated quotes were verified by a bilingual researcher. Qualitative trustworthiness was supported by consensus-based codebook development, retention of verbatim representative quotes, transparent reporting of denominators and multicoding rules, and bilingual verification of translated quotes; because coding was consensus-based rather than independently blinded, interrater reliability statistics were not computed, and independent double-coding, member checking, and external auditing were not undertaken. Although reporting theme frequencies can imply a prevalence logic that Braun and Clarke [36] caution against for reflexive thematic analysis, transparent reporting of nomination frequencies can aid interpretation in codebook-oriented analysis of brief survey responses when the counts are framed as salience; accordingly, theme frequencies and all reported percentages should be interpreted as respondent-nominated salience rather than prevalence estimates or relative importance, especially because C1 and C2 were valence-leading prompts. We nonetheless acknowledge that frequency reporting in qualitative analysis should be interpreted cautiously and is intended here to enhance descriptive transparency for brief survey responses rather than to imply thematic importance.

Mixed Methods Integration

This study used a convergent integration approach [25,37] in which quantitative and qualitative strands were analyzed separately and integrated at the results level through a joint display to identify contextual alignment, complementarity, and discrepancies; the overall design is summarized in Figure 2.

XSL•FO
RenderX

Figure 2. Convergent mixed methods design of the study. Quantitative and qualitative data were collected through the same postuse questionnaire, analyzed separately, and integrated at the results level through a joint display. AMTES: AI-powered medical history-taking training and evaluation system; EFA: exploratory factor analysis; HC3: heteroscedasticity-consistent standard errors (type 3); IM: intrinsic motivation.



Ethical Considerations

This study received ethical approval from the Ethics Committee of Shantou University Medical College (SUMC-2024-079) and followed the principles of the Declaration of Helsinki. AMTES training was part of the course, but participation in the anonymous postuse questionnaire and permission to use responses for research and publication were voluntary and covered by separate written informed consent obtained after summative examination and grade recording. All data were deidentified before analysis and stored on password-protected institutional servers accessible only to the research team; full free-text responses were not publicly released to protect participant privacy in this single-institution cohort. No financial compensation was provided.

Results

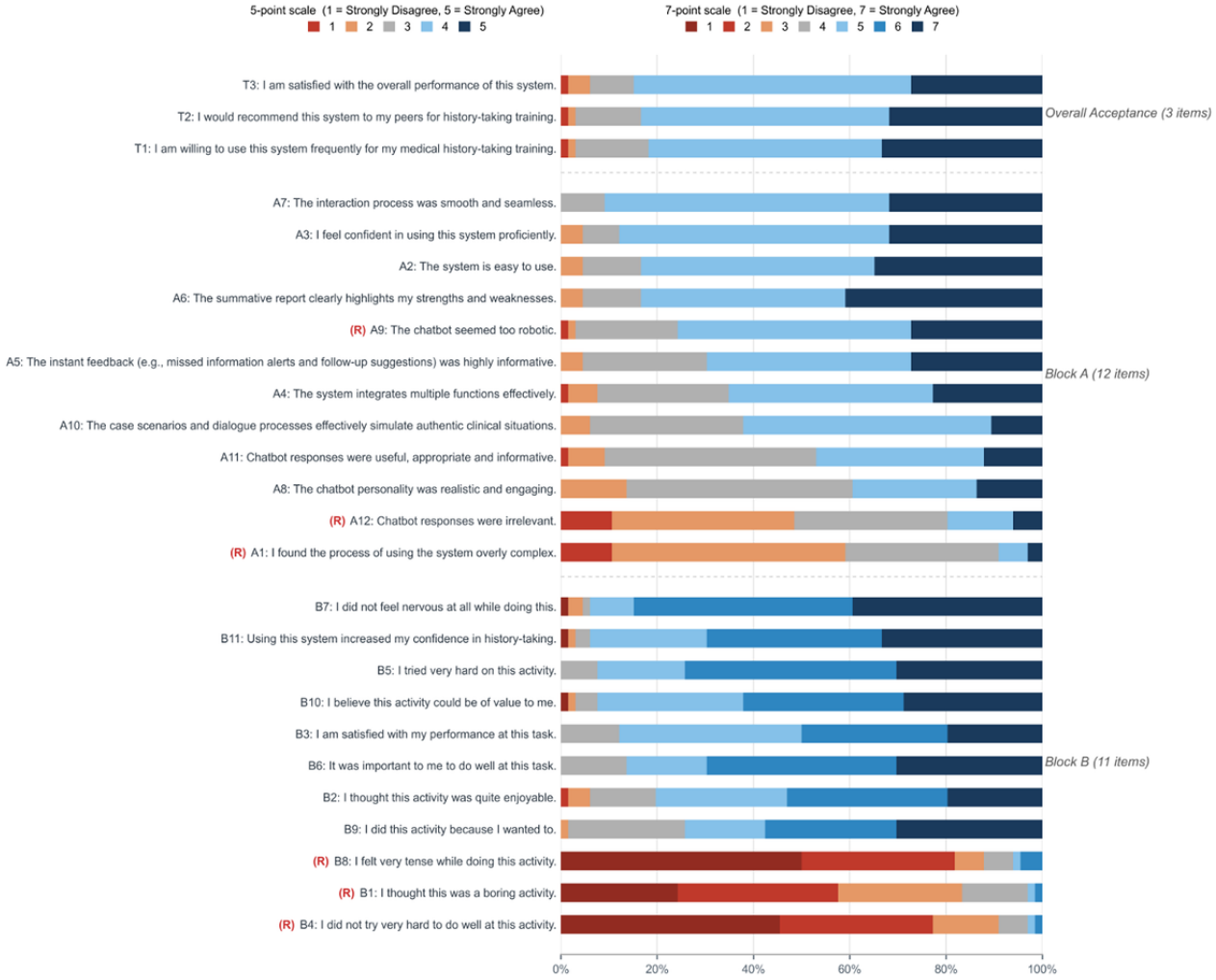
Participant Characteristics

All 66 undergraduate medical students enrolled in the Diagnostics course completed the questionnaire and were included in the analysis (response rate 100%). The mean age was 20.58 (SD 0.98) years; 33 (50%) were female, and 33 (50%) were male. Regarding previous experience, 22/66 (33.3%) students reported 1-2 previous history-taking practice occasions, 24/66 (36.4%) reported 3-4, and 20/66 (30.3%) reported 5 or more. For previous experience of using similar AI history-taking systems, 58/66 (87.9%) had never used one, 8/66 (12.1%) reported rare prior use, and none reported frequent prior use. Figure 3 presents item-level response distributions for overall acceptance, system experience, and learning experience/intrinsic

XSL•FO
RenderX

motivation items. Responses were generally concentrated toward the favorable end of the scales, whereas reverse-coded items showed comparatively lower endorsement.

Figure 3. Item-level response distributions for overall acceptance, system experience, and learning experience/intrinsic motivation items among undergraduate medical students (N=66). Items are displayed as originally worded. Stacked bars show the percentage selecting each response category. Across most items, responses were concentrated in the favorable range, while reverse-worded items showed the opposite pattern.



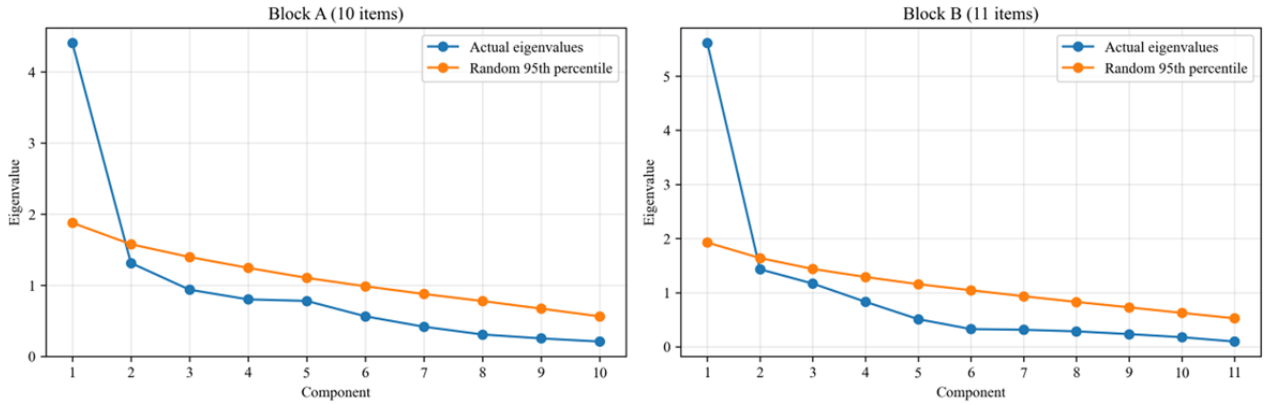
EFA Results

Factorability and Factor Retention

Both Block A and Block B demonstrated adequate factorability. Block A had an overall KMO of 0.745, with a significant Bartlett test of sphericity (χ^2_{66}=317.66; P<.001); Block B had an overall KMO of 0.835, with a similarly significant Bartlett test (χ^2_{55}=460.97; P<.001). These results indicated that the interitem correlation structures in both sets were suitable for EFA. After item-level screening, parallel analysis of the final retained item sets further suggested retaining 1 factor for both Block A and Block B (Figure 4), indicating that both modules primarily reflected a single dominant dimension in this sample (Multimedia Appendix 3). Because the participant-to-item ratio was modest (approximately 5-6:1) and parallel analysis of the unscreened 12-item Block A set initially suggested a possible 2-factor structure, the 1-factor solutions are best regarded as provisional, exploratory composites for this sample rather than as evidence of a validated unidimensional scale.

XSL•FO
RenderX

Figure 4. Parallel analysis results for factor retention in Block A and Block B (N=66). In both blocks, only the first observed eigenvalue exceeded the corresponding 95th percentile of the random eigenvalues, indicating a 1-factor solution.



Item Refinement and Internal Consistency

Block A initially included 12 candidate items. Two items were excluded during item screening because they did not meet the prespecified MSA and factor loading criteria, resulting in a final 10-item set. All 11 items in Block B were retained. The final item sets showed good internal consistency, with Cronbach α of 0.856 for Block A and 0.895 for Block B. The retained items were loaded onto a single factor within each block and were averaged to generate composite scores: A_total for system experience (Block A) and B_total for learning experience and intrinsic motivation (Block B), for use in subsequent regression analyses. Across the retained items, single-factor loadings ranged from 0.45 to 0.77 in Block A and from 0.42 to 0.84 in Block B, with corresponding communalities of approximately 0.20 to 0.60 and 0.17 to 0.70, respectively. Item-level MSA values, factor loadings, communalities, and retention decisions are reported in detail in Multimedia Appendix 3. Given the modest sample size, these composites should be interpreted as exploratory, and the factor structure may be subject to instability and overfitting; replication in larger samples is needed.

Supplementary Common-Method-Variance Check

Harman single-factor test showed that the first factor explained 38.3% of the variance after excluding acceptance items, and did not indicate a dominant single factor under this heuristic. Results were similar when all acceptance items were included, with the first factor explaining 41.7% of the variance. However, Harman single-factor test is an insensitive diagnostic, and residual common method variance cannot be excluded.

Regression Analyses

On the 0-1 rescaled composites, mean scores were 0.675 (SD 0.134) for A_total, 0.804 (SD 0.129) for B_total, and 0.771 (SD 0.194) for overall acceptance. The overall-acceptance composite showed high internal consistency (Cronbach α=0.938). Both the system-experience composite (A_total) and the learning-experience/intrinsic-motivation composite (B_total) were both significantly and positively associated with overall acceptance. A_total showed a stronger association (standardized β=.526, 95% CI 0.283-0.722; *P*<.001) than B_total (standardized β=.377, 95% CI 0.162-0.599; *P*=.002). The primary model explained 66.2% of the variance in overall acceptance (R^2=0.662; adjusted R^2=0.652; $F_{2,\ 63}$=61.80; *P*<.001). Multicollinearity was low, with a variance inflation factor of 1.61 for both predictors. Additional variance-decomposition analyses are reported in Multimedia Appendix 4.

Sensitivity analyses yielded results consistent with the primary model. Adjustment for age and sex did not materially change the main associations, and similar patterns were observed across alternative outcome specifications (use intention, recommendation intention, and overall satisfaction) and a fractional logit model for the bounded acceptance outcome. In the post hoc content-overlap check, the lower-overlap system-experience composite remained positively associated with overall acceptance when modeled with B_total, although its independent association was attenuated in a partitioned model after the higher-overlap items were entered simultaneously, indicating that semantic content overlap contributed to the magnitude of the primary system-experience association (Multimedia Appendix 2).

Qualitative Findings From Open-Ended Questions

Overview

Both open-ended questions yielded 66 responses. For C1 (perceived benefits), codebook-oriented thematic analysis with multicoding generated 2 main themes, 5 substantive subthemes, and 1 nonsubstantive category for vague responses. Multicoding was assigned to 18/66 (27.3%) C1 responses (18/65, 27.7% eligible substantive responses) and 18/66 (27.3%) C2 responses (18/59, 30.5% eligible substantive responses); theme percentages in Table 1 use the full response denominator (N=66), whereas the coding-density summaries reported in this paragraph use eligible substantive responses (excluding 1 vague C1 response and 7 nonspecific C2 responses). For C2 (refinement priorities), a multicoding approach was applied, generating 7 substantive themes and 1 nonsubstantive category; the 4 most-frequently-nominated themes (semantic understanding and intent recognition, contextual consistency across multiturn dialogue, conversational naturalness, and scoring accuracy and contextual scoring logic) were each cited by 21%-35% of all 66 respondents. Theme frequencies and representative quotes are summarized in Table 1.

XSL•FO
RenderX

Table 1. Summary of qualitative themes from open-ended questions (N=66)[a].

Theme and subtheme	Responses, n (%)	Representative quote
C1 (Benefits)		
Practice accessibility and feedback support		
Convenient and accessible self-directed practice	28 (42.4)	"It can be used anytime and anywhere, which is quite convenient" (P52)
Low-pressure safe practice space	7 (10.6)	"I can think slowly without rushing to ask questions" (P27)
Immediate feedback for self-assessment and gap identification	25 (37.9)	"I can review what was missed...clearly see where points were lost" (P43)
Competence development and clinical transition		
History-taking skill and process improvement	16 (24.2)	"Through repeated practice with this system, I became more familiar with the consultation process" (P58)
Clinical realism and confidence building	8 (12.1)	"First get familiar with history-taking, then move on to SP conversations, and finally to clinical practice" (P26)
Nonsubstantive		
Vague or aesthetic-only response	1 (1.5)	—[b]
C2 (Refinement priorities)		
Dialogue and interaction quality		
Semantic understanding and intent recognition	23 (34.8)	"Sometimes one needs to phrase the question very precisely for the system to understand" (P15)
Contextual consistency across multiturn dialogue	17 (25.8)	"When asking about the situation three months earlier without restating the time frame, the patient answered about the current episode instead" (P51)
Conversational naturalness	17 (25.8)	"I hope the AI patient can incorporate some normal conversational expressions and not appear so stiff" (P7)
Assessment logic		
Scoring accuracy and contextual scoring logic	14 (21.2)	"I had asked about certain points, but the system indicated I had not" (P35)
System and content resources		
System stability and performance	7 (10.6)	"Sometimes when many people are using it, you can't log in" (P5)
Feedback and report content	3 (4.5)	"The feedback could ideally provide a more complete case summary" (P55)
Case library sufficiency	2 (3)	"The case set is limited; we tend to practice the same few cases repeatedly" (P42)
Nonsubstantive		
Blank, generic, or no specific suggestion	7 (10.6)	"None"/"Nothing for now"

[a]For both C1 and C2, multicoding was permitted: a single response could be assigned to more than one substantive subtheme. C1 subtheme percentages can sum to more than 100% across the full substantive subtheme set because 18/66 (27.3%) of responses were assigned more than one subtheme; C2 subtheme percentages similarly can sum to more than 100% across themes. The "Vague/unclear" nonsubstantive category is mutually exclusive with substantive subthemes. Participant identifiers (P1-P66) correspond to the survey sequence number. Percentages are based on total responses (N=66). All quotes were translated from Chinese. C2 percentages represent the frequency of student-identified suggestions, not dissatisfaction rates.

[b]Not available.

C1: Perceived Benefits and Positive Experiences

Theme 1: Practice Accessibility and Feedback Support

This theme, the most commonly identified in C1, captured students' positive appraisal of AMTES as an accessible practice tool with helpful feedback.

Convenient and Accessible Self-Directed Practice

This was the most commonly identified benefit subtheme overall under multicoding (28/66, 42.4%), encompassing both anytime/anywhere convenience and the system's role as an accessible practice channel for students with limited bedside opportunities. Students emphasized that AMTES extended practice opportunities beyond formal class time. For example, "It can be used anytime and anywhere, which is quite convenient; it reduces the tedious preparation before practice" (P52).

XSL•FO
RenderX

Immediate Feedback for Self-Assessment and Gap Identification

This was the second most commonly identified benefit subtheme (25/66, 37.9%). Students valued immediate feedback for identifying omissions and supporting reflection, with one student noting: "After the consultation, I can review what was missed, and there are records so I can clearly see the scoring and areas where points were lost" (P43).

Low-Pressure Safe Practice Space

Some students (7/66, 10.6%) described the system as a less stressful environment than SP or real-patient encounters, allowing them to think through questions more calmly (eg, P27 and P53). This subtheme also includes 2 responses where students described the system as allowing a slower, more deliberate thinking pace.

Theme 2: Competence Development and Clinical Transition

This theme captured students' perceived value of AMTES for skill development and preparation for clinical encounters.

History-Taking Skill and Process Improvement

Some students (16/66, 24.2%) reported that repeated practice helped them become more familiar with the consultation structure and key questions before real history-taking. For example, "It helped me become more familiar with the consultation framework and key points before real history-taking" (P51).

Clinical Realism and Confidence Building

Students also linked AMTES to greater confidence in subsequent SP or clinical encounters, with some explicitly describing an "AI-to-SP-to-clinical" progression (8/66, 12.1%). For example, "You can first get familiar with history-taking, then move on to SP conversations, and finally to clinical practice" (P26).

One response (1/66, 1.5%) expressed a nonsubstantive aesthetic reaction without any identifiable benefit theme and was coded under the "Vague/unclear" nonsubstantive category.

C2: Student-Identified Refinement Priorities

Dialogue and Interaction Quality

Semantic Understanding and Intent Recognition

The most frequently cited refinement priority concerned the system's ability to recognize question intent and respond appropriately (23/66, 34.8%). Students noted that semantically varied or context-dependent questions were not always well understood, suggesting room for improvement in context-sensitive intent disambiguation, for example, "sometimes one needs to phrase the question very precisely for the system to understand" (P15).

Contextual Consistency Across Multiturn Dialogue

Some students (17/66, 25.8%) suggested room for improvement in maintaining state across turns, noting occasional inconsistencies between earlier and later parts of the same conversation. For example, "when asking about the situation three months earlier without restating the time frame, the patient answered about the current episode instead" (P51).

Conversational Naturalness

Few students (17/66, 25.8%) suggested incorporating more natural conversational style and softening the AI patient's response register, for example, "I hope the AI patient can incorporate some normal conversational expressions and not appear so stiff" (P7).

Assessment Logic

Scoring accuracy and contextual scoring logic (14/66, 21.2%): open-ended responses highlighted 2 opportunities to strengthen the alignment between student actions and system feedback: (1) more consistent recognition of history-taking content already covered by students, and (2) feedback that better reflects the clinical priorities of each case. As one student noted, "I had asked about certain points, but the system indicated I had not" (P35). These comments do not adjudicate scoring accuracy directly, but they are important because perceived mismatches may affect trust in automated feedback.

System and Content Resources

Three less frequent refinement priorities concerned system infrastructure and content: system stability and performance (7/66, 10.6%), mainly occasional access difficulty during peak use; feedback and report content (3/66, 4.5%), including requests for more comprehensive postsession summaries; and case library sufficiency (2/66, 3.0%), reflecting a desire for additional practice cases.

A total of 7 out of 66 (10.6%) responses provided no specific refinement suggestion; these included blank entries and brief generic remarks such as "None" or "Nothing for now."

Mixed Methods Integration: Joint Display

Textbox 1 presents a joint display integrating the quantitative regression findings with the qualitative themes from C1 and C2, organized by the 2 core experience dimensions, to identify contextual alignment, complementarity, and cross-dimensional relationships not apparent from either strand alone.

XSL•FO
RenderX

Textbox 1. Joint display integrating quantitative and qualitative findings. Quantitative associations are described in plain language, with exact coefficients and 95% CIs reported in the text. C1 parent-theme percentages are respondent-level deduplicated: theme 1 was nominated by 49/66 (74.2%) students and theme 2 by 22/66 (33.3%) at the parent-theme level; because multicoding was permitted and 6 responses were coded to subthemes under both themes, these percentages are not mutually exclusive. Subtheme frequencies are reported in Table 1.

System experience (A_total)

- Quantitative finding
  - Stronger positive association with overall acceptance; robust across sensitivity analyses.
- Qualitative: benefits (C1)
  - Practice accessibility and feedback support was the most frequently nominated benefit cluster, with convenient self-directed practice and immediate feedback as the 2 highest-frequency subthemes (Table 1).
- Qualitative: refinement priorities (C2)
  - Four dialogue-quality and assessment dimensions emerged as leading refinement priorities: semantic understanding and intent recognition, contextual consistency across multi-turn dialogue, conversational naturalness, and scoring accuracy and contextual scoring logic (Table 1).
- Integration assessment
  - Both strands show same-source alignment on system interaction as the more salient acceptance dimension. Qualitative data add 4 actionable refinement directions not derivable from the regression coefficient alone. Cross-strand analysis further suggests that perceived scoring-feedback discrepancies may be associated with reduced trust in the feedback function—a candidate explanatory pattern not apparent from the quantitative model alone.

Learning experience/intrinsic motivation (B_total)

- Quantitative finding
  - Significant positive association with overall acceptance; weaker than system experience but directionally consistent across sensitivity analyses.
- Qualitative: benefits (C1)
  - Competence development and clinical transition was the secondary benefit theme, with a subset of students spontaneously describing an AI-to-SP-to-clinical training progression (Table 1).
- Qualitative: refinement priorities (C2)
  - Learning-motivation dimensions were infrequently nominated as refinement priorities in C2, suggesting that learning value was not a salient refinement concern in this cohort.
- Integration assessment
  - Same-source alignment: the comparatively weaker quantitative association aligns with the lower qualitative salience of learning-related themes.
  - Complementarity: learning-related concerns were infrequently nominated; interaction-level issues were more frequent, suggesting interaction quality may have been the more salient acceptance-related concern.

The joint display revealed contextual alignment between the 2 strands on system interaction as the more salient acceptance dimension, while qualitative data contributed complementary explanatory depth. Most notably, it suggested a hypothesized pathway whereby perceived scoring-feedback discrepancies may be associated with reduced feedback trust, a candidate cross-dimensional pattern not captured by the quantitative model. Because both strands were drawn from the same postuse questionnaire, their agreement reflects same-source alignment rather than independent triangulation; methodological implications of this are addressed in the Discussion.

Discussion

Principal Findings

As a Level 1–aligned evaluation of learner reactions following structured curricular use of AMTES, this study addressed 3 prespecified objectives. First, both system experience and learning experience/intrinsic motivation were significantly and positively associated with overall acceptance, with system experience showing the stronger of the 2 associations—a pattern that held across multiple sensitivity analyses. Second, qualitative analysis identified immediate feedback for self-assessment and convenient, accessible practice as the most frequently nominated benefits, while semantic understanding and intent recognition, contextual consistency across multiturn dialogue, and

XSL•FO
RenderX

conversational naturalness emerged as the most frequently nominated refinement priorities. Third, integration of both strands through a joint display revealed contextual alignment between the quantitative and qualitative evidence and surfaced explanatory depth not apparent from the regression model alone. Most notably, this approach identified scoring-feedback alignment as a concrete optimization target and clarified how perceived assessment discrepancies may be associated with reduced trust in the feedback function. Unlike studies 1 and 2, which addressed technical reliability and educational effectiveness, respectively, this study addressed a distinct implementation question: which experience dimensions are associated with learner acceptance of AMTES in a structured curricular deployment, and where optimization efforts might be directed.

System Experience and Learning Experience as Correlates of Overall Acceptance

Both system experience and learning experience/intrinsic motivation showed significant positive associations with overall acceptance. The descriptive variance decomposition in Multimedia Appendix 4 attributed unique explained variance to each predictor and substantial shared variance; given the predictor correlation, the modest sample size, and the exploratory composite construction, these attributions should be treated only as sample-specific context rather than durable evidence of predictor importance.

The proportion of variance explained by the primary model (R^2=0.662) should be interpreted as an upper-bound, context-specific figure rather than as a precise effect size. Because the predictors and the outcome were measured by self-report within the same questionnaire, shared method variance is expected to inflate both the observed associations and the overall R^2, and partial semantic overlap between some system-experience items and the acceptance composite likely contributes further. Consistent with this, content-overlap sensitivity analyses showed that a lower-overlap system-experience composite remained positively associated with acceptance, whereas in a partitioned model its independent association was attenuated once the higher-overlap items were entered (Multimedia Appendix 2). The coefficients should therefore be interpreted as context-specific associational estimates that may partly reflect measurement overlap, rather than as evidence of causal influence. At the same time, the persistence of a positive association in the lower-overlap sensitivity analyses suggests that the observed relationship is unlikely to be entirely an artifact of content overlap. The observation that system experience showed the stronger association holds within the original measurement specification rather than as evidence of a durable construct hierarchy. Because the design was cross-sectional, all reported relationships are associational, and the proposed mechanisms are candidate hypotheses for future longitudinal or experimental testing.

Substantively, this pattern is consistent with technology-acceptance research, in which perceived usefulness and effort-related interaction quality are central to acceptance [14,18,19]. In AI-supported medical and medical-education settings, emerging work further links trust and acceptance not only to perceived usefulness but also to technology-related factors affecting confidence in system outputs, interaction processes, and feedback quality [10,15]. One plausible interpretation is that when dialogue fluency, contextual continuity, or recognition accuracy require learners to work around the system, perceived educational value alone may not be sufficient to secure strong acceptance. The substantial shared variance between the 2 predictors (Multimedia Appendix 4) further indicates that the 2 experience dimensions overlap to a degree that necessitates joint interpretation rather than isolated analysis.

Qualitative data collected from the same postuse questionnaire provided contextual elaboration of the quantitative pattern. Among student-identified refinement priorities, 3 closely related dialogue-quality concerns emerged as primary student priorities: semantic understanding, contextual consistency across multiturn dialogue, and conversational naturalness. The coherence deficits students described (eg, needing to repeat subject and time qualifiers in follow-up questions) and semantic-understanding limitations (apparent keyword-matching reliance) map directly onto constructs measured by the system-experience items; when students must repeatedly work around recognition limitations, their system-experience ratings may be less favorable.

These findings operate at different levels: study 1 [20] assessed script-level technical accuracy by comparing system responses against case scripts and expert ratings, whereas this study captured user-perceived interaction quality—including multiturn contextual coherence and phrasing flexibility. The apparent inconsistency is therefore not a true contradiction. A response can match the case script yet still feel mechanical, poorly connected across turns, or misaligned with students' scoring-feedback expectations.

A parallel pattern applies to scoring: study 1's intraclass correlation coefficient reflected agreement with expert evaluators [20], whereas C2 captured students' item-level perceptions of whether asked content was credited. C2 comments should therefore not be read as direct evidence that the system lacks expert-level scoring reliability; instead, they more likely reflect a combination of residual recognition mismatches, divergent student-expert interpretations of scoring criteria, and high learner expectations of near-perfect item-level agreement in this cohort. Prior medical and health care education literature indicates that learners judge assessment-generated feedback through perceived credibility, relevance, fairness, and alignment with educational purposes, and these judgments can influence whether feedback is used or discounted [38-40]. Because the questionnaire was administered anonymously, individual student comments about scoring or feedback accuracy could not be cross-checked against system logs at the item level. Even occasional discrepancies may disproportionately affect user trust in the feedback function [41]. Considered together, study 1 and this study provide complementary evidence: the former supports technical reliability, while the latter clarifies the mechanism by which user trust may still be challenged by perceived discrepancies in real educational use.

A further interpretive possibility is that interaction-quality concerns may spill over into perceptions of feedback credibility.

XSL•FO
RenderX

When students perceive that the content they asked about was not credited, the core benefit of immediate feedback may appear less trustworthy, even if the underlying scoring framework is broadly reliable. Prior research suggests that medical students' perceived trustworthiness of AI-generated feedback is not uniformly high, with ratings varying considerably across individuals depending on their prior attitudes toward AI [42]. This cross-theme relationship illustrates an explanatory contribution of qualitative analysis not captured by the 2 aggregated quantitative indicators. Specifically, it demonstrates how perceived assessment discrepancies may contribute to less favorable overall system-experience ratings.

The Role of Learning Experience and Intrinsic Motivation

The learning-experience/intrinsic-motivation composite showed a statistically significant positive association with overall acceptance that was directionally consistent across sensitivity analyses. Qualitative findings were contextually aligned with this pattern: a substantial portion of benefit-related responses reflected students' recognition of AMTES's value in history-taking skill development and clinical confidence building, broadly consistent with prior reports that virtual patient systems can support clinical skill learning [8,9,43] and learner confidence [9].

A small subset of students spontaneously described an AI-to-SP-to-clinical progression, suggesting that AMTES was perceived as one component of a broader clinical competence-development sequence rather than as a stand-alone tool. This narrative is compatible with Kirkpatrick's hierarchical logic, in which favorable Level 1 reactions may support later learning engagement and behavioral transfer, but these cross-sectional data do not test such downstream outcomes. Although based on this small subset of responses, the narrative suggests a candidate pedagogical implication: AMTES may be more clearly positioned within a structured pathway such as preclass preview, AI practice, SP assessment, and clinical clerkship.

Students' refinement suggestions concentrated on interaction quality rather than learning value. Because C2 asked respondents to nominate a single improvement priority, this pattern indicates nomination salience rather than direct evidence that learning value was already adequate.

Contextual Elaboration and Complementarity in Mixed Methods Integration

Because both strands were collected through the same postuse questionnaire, their agreement is best interpreted as same-source alignment plus qualitative elaboration rather than quasi-independent triangulation [25]. Integration nevertheless added explanatory depth beyond the quantitative model in 3 ways: qualitative data identified concrete refinement directions across 4 dialogue-quality and assessment dimensions; exposed cross-dimensional overlap whereby scoring-feedback concerns affected both perceived system function and learning value; and clarified how students situated AMTES within a broader AI-to-SP-to-clinical training sequence, a pedagogical expectation not apparent from Likert-scale data alone.

Comparison With Prior Work

The finding that system-interaction quality and perceived learning value were both positively associated with acceptance is broadly consistent with research on AI-driven educational tools. Prior virtual patient studies show educational value [7-9,43,44], while newer large language model–based systems underscore the importance of dialogue authenticity, contextual coherence, and feedback quality [10-12].

The contribution of this study lies less in reiterating that AI-driven history-taking systems can be educationally useful and more in clarifying which interaction-level factors are associated with user acceptance during actual curricular use. Specifically, acceptance judgments in this context appeared to reflect not only perceived learning value but also students' perceptions of whether the system recognized the intent of semantically varied follow-up questions, maintained coherence across turns, sustained a natural conversational style, and provided contextually appropriate scoring feedback. This adds implementation-level explanatory depth to a literature that has often emphasized overall effectiveness or satisfaction without identifying the interaction-level bottlenecks most relevant to routine adoption.

At the methodological level, the convergent mixed methods design with results-level integration through a joint display is consistent with mixed methods integration approaches in recent medical education guidance and broader methodological literature [25,37]. The joint display was particularly useful because it supported descriptive integration between the quantitative coefficient pattern and the qualitative emphasis on system-related benefits, while also helping identify refinement directions and cross-dimensional relationships not readily apparent from the quantitative analysis alone. This design may therefore provide a methodological example for future user evaluations of AI-driven educational technologies in clinical training.

Practical Implications

The recommendations below are grounded in this single-institution undergraduate curricular deployment of AMTES; their transferability to other AI-driven history-taking systems, learner levels, or institutional settings requires independent evaluation. Based on the integrated evidence, this study proposes the following candidate priorities for refinement and pedagogical implementation within the current cohort.

First priority: enhance dialogue interaction quality. Students' responses clustered around 3 closely related dialogue-quality dimensions: semantic understanding and intent recognition, contextual consistency across multiturn dialogue, and conversational naturalness. Addressing these issues may involve better intent recognition to reduce keyword dependence, stronger tracking of prior dialogue context, and more varied response phrasing to reduce the mechanical tone of simulated-patient responses. Because dialogue quality appears relevant to users' immediate experience in natural-language virtual patient systems, these improvements are hypothesized to have relevance beyond AMTES itself, pending independent replication.

XSL•FO
RenderX

Second priority: bridge the gap between automated assessment and student expectations through contextual scoring logic. Technically, better recognition of consultation content despite wording variations may reduce mismatches between what students ask and what the system credits; scoring rules should also be adapted to case metadata so that age-, sex-, and context-dependent items are evaluated only when clinically appropriate. Pedagogically, guided postsession debriefing, such as instructor-led review of AI-generated feedback, may help students understand the scoring logic and turn perceived discrepancies into learning opportunities.

Third priority: strengthen learning value through curricular integration and content expansion. AMTES may be particularly useful when embedded within a structured training sequence rather than used as a stand-alone activity. Positioning it within a pathway such as preclass preview, AI practice, SP assessment, and clinical clerkship may clarify its pedagogical role and strengthen perceived learning value, although this recommendation draws on a relatively small subset of student responses. Case-library expansion may also be considered as a secondary improvement to support repeated practice, although this was a low-salience request in this cohort.

Limitations

This study has several limitations. First, the sample was drawn from a single institution and a single course cohort, which limits the generalizability of the findings to other institutions, learner levels, and educational contexts. The modest sample size means that factor-analytic results should be treated as exploratory and sample-specific rather than as formal psychometric validation; replication in larger and more diverse samples is needed before the observed patterns can be interpreted with confidence. Given the developer/evaluator dual role disclosed in the Conflicts of Interest section, findings should be interpreted with this context in mind. Several authors were involved in developing, deploying, teaching with, and evaluating AMTES. To mitigate the associated bias, the questionnaire was administered anonymously and only after the summative examination and after all grades had been finalized and officially recorded, with no linkage to grades and with no instructor access to identifiable responses; multiple sensitivity analyses, including post hoc analyses clearly labeled as such, were reported transparently; and student-identified critical feedback was deliberately retained and reported transparently irrespective of valence. Nonetheless, the primary outcomes were self-reported, the qualitative coding was performed by research team members who were not blinded to the system's authorship and without an independent external auditor, and no independent third-party evaluation of AMTES was undertaken; independent external evaluation is an important direction for future work.

Second, quantitative data were collected via a self-report questionnaire at a single time point, introducing the possibility of social desirability bias and common method bias. Although procedural safeguards, including postexamination administration, explicit separation from course grading, and anonymized data handling, were implemented, the Harman single-factor test is an insensitive diagnostic that cannot rule out common method variance. Therefore, the observed associations and explained variance may be partly inflated by shared method variance, and future studies should incorporate procedural and statistical remedies such as temporal or source separation of predictors and outcomes or marker-variable techniques. Some Block A items also shared evaluative content with the acceptance composite, so the observed association between system experience and overall acceptance may partly reflect content overlap. Finally, acceptance reflected self-reported intention and satisfaction rather than observed adoption behavior, and the anonymous questionnaire format precluded person-level linkage between acceptance judgments and actual usage data. Several considerations support the plausibility of the 100% response rate and mitigate concern about coercion: the cohort was small and sampled as a whole; every student had already completed at least one history-taking case and therefore had direct experience of AMTES on which to base their responses; the survey was brief and administered in person during a single supervised session; and it was conducted only after the summative examination had ended and all grades had been finalized and officially recorded, several days later, so that participation could not affect academic outcomes. Participation was voluntary and anonymous, with no linkage to grades, and students were told that declining carried no academic consequence. We nonetheless acknowledge that the study was introduced by a member of the research team who was also involved in course teaching, and that a supervised classroom setting may carry some perceived expectation to participate; although the anonymous administration (which prevented any instructor from linking responses to individual students) and the postgrading timing were designed to minimize this, a residual perception of obligation cannot be entirely excluded.

Third, the qualitative component was constrained by the brevity of open-ended responses, which limited thematic depth and precluded saturation; future research incorporating semistructured interviews or focus groups could yield richer experiential narratives. Student-reported scoring discrepancies were not independently adjudicated against expert review at the item level, meaning the qualitative data could not distinguish actual system errors from misunderstanding of scoring criteria or expectation-based perceptions. As a cross-sectional study, this work cannot address longitudinal changes in acceptance or sustained real-world adoption; future linkage-enabled studies are needed to examine whether postuse acceptance predicts subsequent voluntary engagement and whether iterative system refinement translates into improved user experience over time.

Conclusions

To our knowledge, relatively few prior studies have used a convergent mixed methods design to examine both the experience dimensions associated with user acceptance and the underlying experiential patterns in AI-powered history-taking training systems. This study contributes to this small but growing literature. Unlike prior work that has primarily addressed technical reliability or educational effectiveness, this study extends the focus to implementation-level acceptance. The findings suggest that, in this cohort and within the study's measurement specification, system interaction quality showed a stronger association with acceptance than with perceived

XSL•FO
RenderX

learning value during structured curricular use of AMTES. This distinction carries practical implications for system developers, who may need to prioritize interaction-level refinements over feature expansion, and for educators, who may benefit from explicitly positioning such systems within structured training pathways.

From a broader perspective, these findings contribute to the growing evidence base on AI-driven educational technologies by suggesting a plausible mechanism linking natural-language interaction quality to user acceptance. Specifically, perceived coherence deficits and scoring-feedback discrepancies may be associated with reduced trust in the feedback function and, by extension, with less favorable overall acceptance even when educational value is recognized. The convergent mixed methods design, with results-level integration through a joint display, was particularly useful in surfacing this cross-dimensional relationship, which would have remained invisible to purely quantitative approaches. For the field more broadly, dialogue coherence, semantic flexibility, and scoring-feedback alignment may warrant attention as candidate optimization priorities in natural-language virtual patient systems, pending replication in diverse settings. Larger multicenter and longitudinal studies are needed to test the generalizability of these patterns and to examine whether iterative improvements in interaction quality translate into measurable and sustained gains in learner acceptance and downstream training outcomes.

Acknowledgments

The authors declare the use of generative AI (GenAI) in the research and writing process. According to the GAIDeT taxonomy (2025), the following tasks were delegated to GenAI tools under full human supervision: (1) proofreading and editing and (2) translation. The GenAI tool used was Claude Opus 4.7 (Anthropic). Responsibility for the final manuscript lies entirely with the authors.

GenAI tools are not listed as authors and do not bear responsibility for the final outcomes.

Data Availability

Deidentified quantitative survey data, case-level practice-log summary data, and the Python analysis scripts used for data cleaning and statistical analysis are available from the corresponding author on reasonable request. Because raw free-text responses from this single-institution cohort may contain contextual details that increase the risk of reidentification, those materials are not publicly released; a redacted version may be made available on reasonable request, subject to institutional and ethics requirements.

Funding

We acknowledge the support from the Key Project on Medical Education Research by the National Medical Education Development Center, Medical Education Branch of the Chinese Medical Association (2025A25); the 2025 Teaching Reform and Research Project of Shantou University Medical College (No. 10); the 2025 Teaching Reform and Research Project (AI Curriculum Special Program) of Shantou University Medical College (No. 8). The funders had no involvement in the study design, data collection, analysis, interpretation, or the writing of the manuscript.

Authors' Contributions

Conceptualization: WZ, OX, XL, LW
Methodology: WZ, OX, YL, YZ, CS, XL, LW
Formal analysis: WZ, YL, YZ, MY, XH
Investigation: YL
Data curation: MY, XH
Software: WZ
Visualization: YZ
Project administration: WZ, OX, LW
Writing–original draft: YL
Writing–review and editing: CS, WZ, OX, YL, YZ, XL, LW, MY, XH
YL, YZ, and CS contributed equally to this work and share co-first authorship. WZ and OX contributed equally to this work and share co-corresponding authorship. All authors reviewed and approved the final manuscript for publication.

Conflicts of Interest

No financial competing interests are declared. Several authors were involved in the development and/or institutional deployment of AMTES, which represents a relevant nonfinancial competing interest (developer/evaluator dual role).

Multimedia Appendix 1
Construct mapping between study composites and established acceptance frameworks.
[PDF File (Adobe PDF File), 172 KB - mededu_v12i1e101425_app1.pdf]

XSL•FO
RenderX

Multimedia Appendix 2
Sensitivity analyses.
[PDF File (Adobe PDF File), 273 KB - mededu_v12i1e101425_app2.pdf]

Multimedia Appendix 3
Item-level sampling adequacy, factor loadings, communalities, and item retention.
[PDF File (Adobe PDF File), 124 KB - mededu_v12i1e101425_app3.pdf]

Multimedia Appendix 4
Supplementary variance-decomposition analyses.
[PDF File (Adobe PDF File), 104 KB - mededu_v12i1e101425_app4.pdf]

References

1. Peterson MC, Holbrook JH, Von Hales D, Smith NL, Staker LV. Contributions of the history, physical examination, and laboratory investigation in making medical diagnoses. West J Med 1992;156(2):163-165 [FREE Full text] [Medline: 1536065]
2. Fukuzawa F, Yanagita Y, Yokokawa D, Uchida S, Yamashita S, Li Y, et al. Importance of patient history in artificial intelligence-assisted medical diagnosis: comparison study. JMIR Med Educ 2024;10:e52674 [FREE Full text] [doi: 10.2196/52674] [Medline: 38602313]
3. Keifenheim KE, Teufel M, Ip J, Speiser N, Leehr EJ, Zipfel S, et al. Teaching history taking to medical students: a systematic review. BMC Med Educ 2015;15:159 [FREE Full text] [doi: 10.1186/s12909-015-0443-x] [Medline: 26415941]
4. Faustinella F, Jacobs RJ. The decline of clinical skills: a challenge for medical schools. Int J Med Educ 2018;9:195-197 [FREE Full text] [doi: 10.5116/ijme.5b3f.9fb3] [Medline: 30007951]
5. Ziv A, Wolpe PR, Small SD, Glick S. Simulation-based medical education: an ethical imperative. Acad Med 2003;78(8):783-788. [doi: 10.1097/00001888-200308000-00006] [Medline: 12915366]
6. Flanagan O, Cummings K. Standardized patients in medical education: a review of the literature. Cureus 2023;15(7):e42027 [FREE Full text] [doi: 10.7759/cureus.42027] [Medline: 37593270]
7. Cook DA, Triola MM. Virtual patients: a critical literature review and proposed next steps. Med Educ 2009;43(4):303-311. [doi: 10.1111/j.1365-2923.2008.03286.x] [Medline: 19335571]
8. García-Torres D, Vicente Ripoll MA, Fernández Peris C, Mira Solves JJ. Enhancing clinical reasoning with virtual patients: a hybrid systematic review combining human reviewers and ChatGPT. Healthcare (Basel) 2024;12(22):2241 [FREE Full text] [doi: 10.3390/healthcare12222241] [Medline: 39595439]
9. Raafat N, Harbourne AD, Radia K, Woodman MJ, Swales C, Saunders KEA. Virtual patients improve history-taking competence and confidence in medical students. Med Teach 2024;46(5):682-688. [doi: 10.1080/0142159X.2023.2273782] [Medline: 38084413]
10. Holderried F, Stegemann-Philipps C, Herrmann-Werner A, Festl-Wietek T, Holderried M, Eickhoff C, et al. A language model-powered simulated patient with automated feedback for history taking: prospective study. JMIR Med Educ 2024;10:e59213 [FREE Full text] [doi: 10.2196/59213] [Medline: 39150749]
11. Cook DA, Overgaard J, Pankratz VS, Del Fiol G, Aakre CA. Virtual patients using large language models: scalable, contextualized simulation of clinician-patient dialogue with feedback. J Med Internet Res 2025;27:e68486 [FREE Full text] [doi: 10.2196/68486] [Medline: 39854611]
12. Wang C, Li S, Lin N, Zhang X, Han Y, Wang X, et al. Application of large language models in medical training evaluation-using ChatGPT as a standardized patient: multimetric assessment. J Med Internet Res 2025;27:e59435. [doi: 10.2196/59435] [Medline: 39742453]
13. Gordon M, Daniel M, Ajiboye A, Uraiby H, Xu NY, Bartlett R, et al. A scoping review of artificial intelligence in medical education: BEME guide no. 84. Med Teach 2024;46(4):446-470 [FREE Full text] [doi: 10.1080/0142159X.2024.2314198] [Medline: 38423127]
14. Lee JWY, Tan JY, Bello F. Technology acceptance model in medical education:systematic review. JMIR Med Educ 2025;11:e67873 [FREE Full text] [doi: 10.2196/67873] [Medline: 40668663]
15. Shevtsova D, Ahmed A, Boot IWA, Sanges C, Hudecek M, Jacobs JJL, et al. Trust in and acceptance of artificial intelligence applications in medicine: mixed methods study. JMIR Hum Factors 2024;11:e47031 [FREE Full text] [doi: 10.2196/47031] [Medline: 38231544]
16. Kusmiati M. A comprehensive evaluation in medical curriculum using the Kirkpatrick hierarchical approach: a review and update. MRAJ 2025;13(5). [doi: 10.18103/mra.v13i5.6557]
17. Yardley S, Dornan T. Kirkpatrick's levels and education 'evidence'. Med Educ 2012;46(1):97-106. [doi: 10.1111/j.1365-2923.2011.04076.x] [Medline: 22150201]
18. Davis FD. Perceived usefulness, perceived ease of use, and user acceptance of information technology. MIS Quarterly 1989;13(3):319-340. [doi: 10.2307/249008]

XSL•FO
RenderX

19. Venkatesh V, Morris MG, Davis GB, Davis FD. User acceptance of information technology: toward a unified view. MIS Quarterly 2003;27(3):425-478. [doi: 10.2307/30036540]
20. Liu Y, Shi C, Wu L, Lin X, Chen X, Zhu Y, et al. Development and validation of a large language model-based system for medical history-taking training: prospective multicase study on evaluation stability, human-AI consistency, and transparency. JMIR Med Educ 2025;11:e73419 [FREE Full text] [doi: 10.2196/73419] [Medline: 40882613]
21. Liu Y, Zhu Y, Zhang W, Lu X, Wu L, Yue M, et al. Real-world impact and educational effectiveness of an AI-powered medical history-taking system: retrospective propensity score-matched cohort study. JMIR Med Educ 2026;12:e89367 [FREE Full text] [doi: 10.2196/89367] [Medline: 41734000]
22. von Elm E, Altman DG, Egger M, Pocock SJ, Gøtzsche PC, Vandenbroucke JP, STROBE Initiative. The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies. Ann Intern Med 2007;147(8):573-577 [FREE Full text] [doi: 10.7326/0003-4819-147-8-200710160-00010] [Medline: 17938396]
23. Eysenbach G. Improving the quality of web surveys: the Checklist for Reporting Results of Internet E-Surveys (CHERRIES). J Med Internet Res 2004;6(3):e34 [FREE Full text] [doi: 10.2196/jmir.6.3.e34] [Medline: 15471760]
24. Braun V, Clarke V. Using thematic analysis in psychology. Qual Res Psychol 2008;3(2):77-101. [doi: 10.1191/1478088706qp063oa]
25. Fetters MD, Curry LA, Creswell JW. Achieving integration in mixed methods designs-principles and practices. Health Serv Res 2013;48(6 Pt 2):2134-2156 [FREE Full text] [doi: 10.1111/1475-6773.12117] [Medline: 24279835]
26. Brooke J. SUS: A 'quick and dirty' usability scale. In: Jordan PW, Thomas B, Weerdmeester BA, McClelland IL, editors. Usability Evaluation in Industry. London: Taylor & Francis; 1996:189-194.
27. Hertzum M. System Usability Scale: a meta-analysis of how SUS relates to workload, task time, and error rate. Int J Hum Comput Interact 2026:1-14. [doi: 10.1080/10447318.2026.2625260]
28. Ryan RM. Control and information in the intrapersonal sphere: an extension of cognitive evaluation theory. J Pers Soc Psychol 1982;43(3):450-461. [doi: 10.1037/0022-3514.43.3.450]
29. Kaiser HF. An index of factorial simplicity. Psychometrika 2025;39(1):31-36. [doi: 10.1007/BF02291575]
30. Bartlett MS. Tests of significance in factor analysis. Br J Stat Psychol 2011;3(2):77-85. [doi: 10.1111/j.2044-8317.1950.tb00285.x]
31. Horn JL. A rationale and test for the number of factors in factor analysis. Psychometrika 1965;30:179-185. [doi: 10.1007/BF02289447] [Medline: 14306381]
32. Fabrigar LR, Wegener DT, MacCallum RC, Strahan EJ. Evaluating the use of exploratory factor analysis in psychological research. Psychol Methods 1999;4(3):272-299. [doi: 10.1037/1082-989X.4.3.272]
33. Podsakoff PM, MacKenzie SB, Lee JY, Podsakoff NP. Common method biases in behavioral research: a critical review of the literature and recommended remedies. J Appl Psychol 2003;88(5):879-903. [doi: 10.1037/0021-9010.88.5.879] [Medline: 14516251]
34. Long JS, Ervin LH. Using heteroscedasticity consistent standard errors in the linear regression model. Am Stat 2000;54(3):217-224. [doi: 10.1080/00031305.2000.10474549]
35. Papke LE, Wooldridge JM. Econometric methods for fractional response variables with an application to 401(k) plan participation rates. J Appl Econ 1996;11(6):619-632. [doi: 10.1002/(sici)1099-1255(199611)11:6<619::aid-jae418>3.0.co;2-1]
36. Braun V, Clarke V. One size fits all? What counts as quality practice in (reflexive) thematic analysis? Qual Res Psychol 2020;18(3):328-352. [doi: 10.1080/14780887.2020.1769238]
37. Battista A, Torre D, Konopasky A. Essential concepts for effective mixed methods research in the health professions: AMEE Guide No. 173. Med Teach 2025;47(5):766-778. [doi: 10.1080/0142159X.2024.2401464] [Medline: 39312179]
38. Long S, Rodriguez C, St-Onge C, Tellier P, Torabi N, Young M. Factors affecting perceived credibility of assessment in medical education: a scoping review. Adv Health Sci Educ Theory Pract 2022;27(1):229-262. [doi: 10.1007/s10459-021-10071-w] [Medline: 34570298]
39. Ricci M, St-Onge C, Xiao J, Young M. Students as stakeholders in assessment: how students perceive the value of an assessment. Perspect Med Educ 2018;7(6):352-361 [FREE Full text] [doi: 10.1007/s40037-018-0480-3] [Medline: 30421333]
40. Dai CM, Bertram K, Chahine S. Feedback credibility in healthcare education: a systematic review and synthesis. Med Sci Educ 2021;31(2):923-933 [FREE Full text] [doi: 10.1007/s40670-020-01167-w] [Medline: 34457934]
41. Lee JD, See KA. Trust in automation: designing for appropriate reliance. Hum Factors 2004;46(1):50-80. [doi: 10.1518/hfes.46.1.50_30392] [Medline: 15151155]
42. Bauermann M, Rotthoff T, Hallmen T, Kunz M, André E, Schindler A. Medical students' perceptions of AI-based feedback and feedforward on communication skills in doctor-patient consultation - an acceptance study in a video-based simulation. Med Educ Online 2025;30(1):2592414 [FREE Full text] [doi: 10.1080/10872981.2025.2592414] [Medline: 41324408]
43. Kononowicz AA, Woodham LA, Edelbring S, Stathakarou N, Davies D, Saxena N, et al. Virtual patient simulations in health professions education: systematic review and meta-analysis by the digital health education collaboration. J Med Internet Res 2019;21(7):e14676 [FREE Full text] [doi: 10.2196/14676] [Medline: 31267981]
44. Consorti F, Mancuso R, Nocioni M, Piccolo A. Efficacy of virtual patients in medical education: a meta-analysis of randomized studies. Comput Educ 2012;59(3):1001-1008. [doi: 10.1016/j.compedu.2012.04.017]

XSL•FO
RenderX

Abbreviations

AI-VP: AI-powered virtual patient
AMTES: AI-powered medical history-taking training and evaluation system
CHERRIES: Checklist for Reporting Results of Internet E-Surveys
EFA: exploratory factor analysis
KMO: Kaiser-Meyer-Olkin
MSA: measure of sampling adequacy
SP: standardized patient
STROBE: Strengthening the Reporting of Observational Studies in Epidemiology
SUS: System Usability Scale
TAM: Technology Acceptance Model
UTAUT: Unified Theory of Acceptance and Use of Technology
VP: virtual patient

Edited by A Stone; submitted 15.May.2026; peer-reviewed by IT Isip-Tan, N Ismail; comments to author 02.Jun.2026; revised version received 06.Jun.2026; accepted 01.Jul.2026; published 21.Jul.2026.

Please cite as:
Liu Y, Zhu Y, Shi C, Lu X, Wu L, Yue M, Hong X, Xia O, Zhang W
User Acceptance of an AI-Powered Medical History–Taking Training System Among Undergraduate Medical Students: Mixed Methods Study
JMIR Med Educ 2026;12:e101425
URL: https://mededu.jmir.org/2026/1/e101425
doi:10.2196/101425
PMID:

©Yang Liu, Yiying Zhu, Chujun Shi, Xian Lu, Liping Wu, Minghui Yue, Xiaolin Hong, Oudong Xia, Weishan Zhang. Originally published in JMIR Medical Education (https://mededu.jmir.org), 21.Jul.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Education, is properly cited. The complete bibliographic information, a link to the original publication on https://mededu.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Effect of Large Language Model–Powered Virtual Standardized Patients on History-Taking Among Undergraduate Medical Students: Propensity-Matched Cohort Study

Yuchen He[1*], MM; Chen Chen[1,2,3*], MD; Rong Yin[4], MSc; Wuyang Zhang[1], MTI; Haoli Chang[5], MM; Wei Yang[6], MD; Fei Li[3,7], MD; Xinhua Li[1], MD; Zhuying Xia[5], MD; Xiaoyun Xie[8], MD; Jing Huang[8], MD; Qiuming Zeng[3,9,10], MD; Guang Yang[11], MD; Jing Wu[1,3,5,12], MD

[1]Clinical Skills Training Center, Xiangya Hospital Central South University, Changsha, Hunan, China

[2]Department of Nephrology, Xiangya Hospital Central South University, Changsha, Hunan, China

[3]National Clinical Research Center of Geriatric Disorders, Xiangya Hospital Central South University, Changsha, Hunan, China

[4]Teaching and Research Section of Diagnostics, Xiangya Hospital Central South University, Changsha, Hunan, China

[5]Department of Endocrinology, Xiangya Hospital Central South University, Changsha, Hunan, China

[6]Department of Respiratory Medicine, Xiangya Hospital Central South University, Changsha, Hunan, China

[7]Department of Geriatric Medicine, Xiangya Hospital Central South University, Changsha, Hunan, China

[8]Department of Rheumatology and Immunology, Xiangya Hospital Central South University, Changsha, Hunan, China

[9]Department of Neurology, Xiangya Hospital Central South University, Changsha, Hunan, China

[10]Clinical Research Center for Neuroimmune and Neuromuscular disorders, Xiangya Hospital Central South University, Changsha, Hunan, China

[11]Department of General Medicine, Xiangya Hospital Central South University, Changsha, Hunan, China

[12]Hunan Engineering Research Center for Obesity and its Metabolic Complications, Xiangya Hospital Central South University, Changsha, Hunan, China

[*]these authors contributed equally

Corresponding Author:
Jing Wu, MD
Department of Endocrinology
Xiangya Hospital Central South University
No.87, Xiangya road
Kaifu District
Changsha, Hunan, 410008
China
Phone: 86 13574120508
Email: wujing0731@csu.edu.cn

Abstract

Background: Medical history taking (MHT) is a foundational clinical competency for medical students; however, traditional training models using standardized patients face challenges such as resource constraints. Large language model–powered virtual standardized patients (LLM-VSPs) offer a safe, repeatable platform for self-directed practice with AI-automated feedback. Nevertheless, their effectiveness in authentic teaching environments and underlying learning mechanisms require further investigation.

Objective: This study aims to evaluate the impact of an LLM-VSP system as an extracurricular self-practice tool on undergraduate medical students' MHT performance in an authentic educational setting without disrupting routine instruction, and to further explore potential associations among practice behaviors, baseline proficiency, and intervention effects.

Methods: This prospective cohort study enrolled 168 third-year medical students. Based on voluntary participation, students were assigned to an intervention group (n=120, using LLM-VSP) or a control group (n=48, receiving routine instruction). Propensity score matching (PSM) balanced confounding factors, yielding 40 matched pairs. Baseline MHT performance was assessed via virtual patient examination after didactic instruction but before clinical practicum. The primary outcome was end-of-term MHT performance assessed at an Objective Structured Clinical Examination station with real standardized patients. The differences between groups were compared using independent samples *t* test, with robustness validated through multiple

XSL•FO
RenderX

linear regression, sensitivity analyses, and Rosenbaum bounds analyses. Exploratory analyses investigated the association between practice behaviors and scores, and observed benefit differences across baseline levels.

Results: After PSM, baseline characteristics were balanced (standardized mean difference <0.1). The intervention group exhibited higher total MHT scores than controls (mean 87.71, SD 7.29 vs mean 83.74, SD 8.06; mean difference 3.98, 95% CI 0.55-7.40 points; *P*=.02; with a medium effect size of Cohen *d*=0.52). Advantages were observed in content (*P*=.03) and communication skills (*P*=.04) subscores. Regression analysis confirmed robust intervention effects (*B*=3.924, 95% CI 1.90-5.94; *P*<.001; R^2=0.708), with sensitivity analyses supporting reliability. Exploratory analysis suggested that mere practice behavior metrics were not independent predictors of final scores, potentially being constrained by baseline proficiency, implying a "cognitive threshold" for effective AI-assisted training. Subgroup analysis indicated a trend of differential benefits: high baseline students appeared to show larger gains (matched sample: mean difference 5.93, 95% CI 2.17-9.70; overall sample: mean difference 7.04, 95% CI 3.84-10.25), whereas improvements in medium and low baseline students were relatively limited (matched sample: mean difference 2.94-2.99; overall sample: mean difference 1.29-1.64).

Conclusions: Introducing LLM-VSPs as a self-practice tool in diagnostics education may help improve undergraduate medical students' MHT performance. Preliminary evidence suggests a potential "cognitive threshold," implying students with solid theoretical foundations and higher baseline proficiency may better achieve skill transformation through AI-assisted autonomous practice. This provides a basis for future stratified teaching strategies and differentiated guidance for students of varying baseline levels.

(JMIR Med Educ 2026;12:e92486) doi:10.2196/92486

KEYWORDS

artificial intelligence; education; large language model; medical history taking; medical; practice behaviors; real-time feedback; undergraduate; virtual standardized patients

Introduction

Medical history taking (MHT) is an essential practical skill for clinicians in the diagnostic and treatment process [1], and is also an important learning objective of the *diagnostics* course. MHT, as a core practical clinical competency, requires repeated deliberate practice for its mastery [2]. Traditional MHT teaching predominantly relies on 2 primary pedagogical approaches: bedside teaching involving real patients and standardized patient (SP)–based teaching [3-5]. Yet the aforementioned instructional process remains plagued by an array of challenges and inherent limitations. Specifically, bedside teaching with real patients struggles to accommodate large-scale educational demands and carries the potential risk of exacerbating doctor-patient tensions; in contrast, SP-based teaching is confronted with issues such as substantial training costs, cumbersome management processes, difficulties in ensuring instructional standardization and homogeneity, and inadequacies in simulating authentic, complex clinical symptoms and signs [6-8].

Virtual standardized patients (VSPs) refer to virtual teaching tools constructed using digital technologies that can simulate the clinical characteristics of real patients and provide interactive feedback [9], used for the training and assessment of medical students' clinical diagnosis and treatment skills and communication abilities. VSPs can effectively address the limitations of SPs while exhibiting characteristics of repeatability and cost-effectiveness, providing new instructional modalities for medical education [10,11]. However, early VSPs exhibited rigid interaction and were difficult to simulate the natural language communication of real patients [12,13]. Large language model–powered virtual standardized patients (LLM-VSPs) can deliver more naturalistic and authentic interactive experiences. They can provide medical students with highly realistic clinical consultation scenarios and timely, individualized feedback [14,15], which endows them with great application potential in medical history collection, doctor-patient communication, and clinical reasoning training [16-19]. However, the quality of LLM prompts directly affects the output performance of VSPs [20]. Algorithmic biases and hallucinations in AI may lead to factually incorrect outputs [21], which poses challenges to the reliability of automated AI scoring. Moreover, unsupervised autonomous practice with LLM-VSPs may foster communication behaviors that violate clinical ethics. Prior research on companion chatbots has documented not only AI-initiated inappropriate behaviors but also the potential for users to direct offensive or dismissive language toward AI [22]. In educational contexts, such bidirectional misbehavior may erode rather than cultivate empathy, and risks reinforcing unprofessional communication habits.

Current research on the effectiveness of LLM-VSPs in diagnostic courses remains relatively scarce, particularly studies that conduct in-depth analysis of LLM-VSPs' usage behaviors and the corresponding learning outcomes among medical students, as well as exploration of differential benefits among students with different abilities. Accordingly, large-scale blind integration of this tool into teaching is premature. To clarify the real effectiveness and optimal application methods of LLM-VSPs, we conducted a small-scale pilot empirical exploration without interfering with the regular undergraduate teaching process. This study aims to conduct a prospective cohort study using voluntary recruitment combined with propensity score matching (PSM) to investigate 3 core research questions: first, whether incorporating LLM-VSPs into the auxiliary practice of MHT can improve medical students' proficiency in this skill; second, what the correlation is between students' practice behaviors during LLM-VSPs interactions and their subsequent learning outcomes; third, what the differential

XSL•FO
RenderX

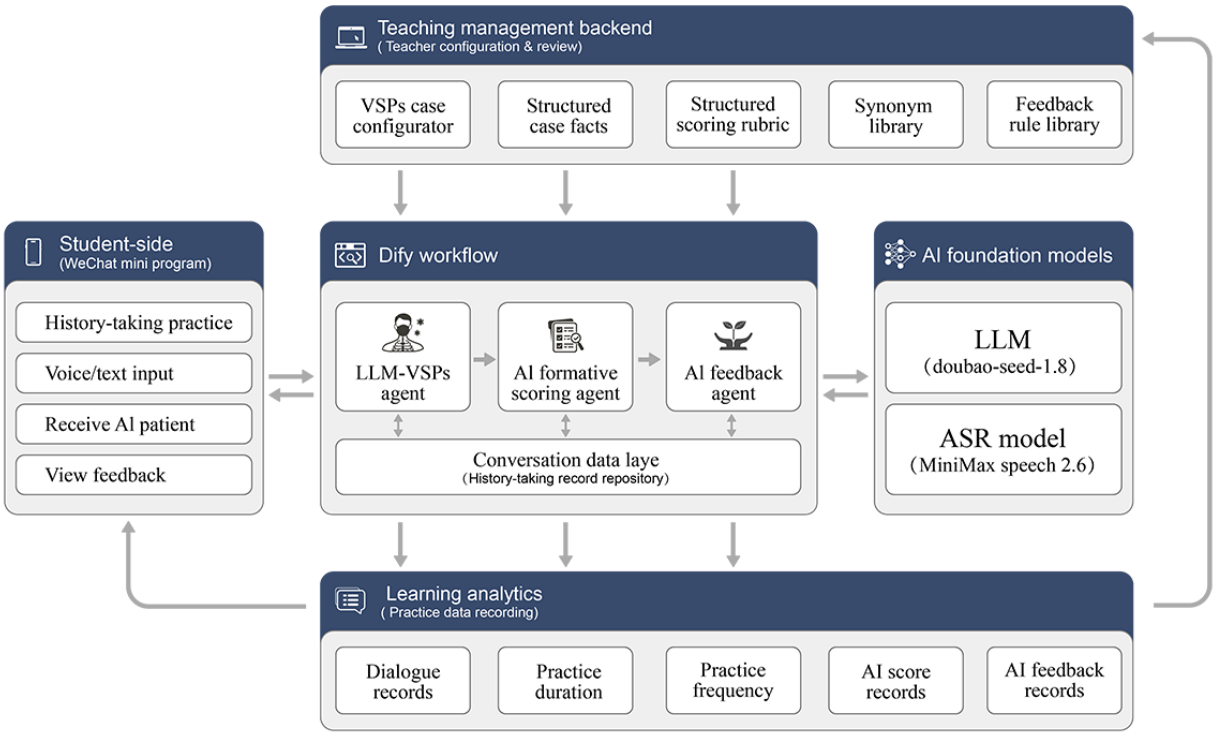
benefits are among student subgroups with different abilities from this intervention. The findings of this study are intended to provide empirical evidence and optimization directions for the future application of AI in MHT skill training.

Methods

Creation of LLM-VSPs

This study used the DoctorU teaching system for MHT practice. This system constructed a multiagent workflow including LLM-VSPs agent, AI formative scoring agent, and AI feedback agent (Figure 1). The LLM-VSPs agent (see Multimedia Appendix 1 for detailed prompts) was responsible for simulating patient roles. The AI formative scoring agent scored students' single MHT performance based on a structured scoring rubric. After the MHT session, the system input the dialogue records, structured case facts (Multimedia Appendix 1), and corresponding structured scoring rubrics into the agent, which assessed whether students covered key MHT content through semantic judgment based on preset scoring points, score values, and scoring conditions, and generated scores, missed items, and brief evaluations. For synonymous questioning methods or nonstandard expressions, the system allowed semantic matching through a synonym library to determine whether students substantially addressed corresponding scoring points. The AI feedback agent generated immediate formative feedback after practice, combining dialogue records and scoring results, covering completed content, missed important information, deficiencies in MHT logic, and suggestions for subsequent improvement based on the feedback rule library. The system-maintained context within a single MHT session to ensure coherence of patient responses; however, individualized dialogue memories were not shared between different students and different training sessions to avoid cross-student information contamination. Simultaneously, the backend database logs the complete practice data of all students to support longitudinal tracking of individual learning growth and cross-sectional analysis of interaction performance and learning processes across students and cases. All system parameters remained fixed during the study period, and specific model invocation parameter settings are shown in Table S1 in Multimedia Appendix 1.

Figure 1. Architecture and workflow of the DoctorU large language model–powered virtual standardized patients (LLM-VSPs) system. ASR: automatic speech recognition; LLM: large language model; VSP: virtual standardized patient.



In accordance with the teaching syllabus and curriculum schedule of the diagnostic course for students in the 5-year clinical medicine program, and with reference to the textbook *Diagnostics (10th Edition)* published by the People's Health Press, clinical cases were selected following a symptom-oriented approach. The criteria for case development were determined to primarily focus on newly admitted patients presenting with a single symptom, thereby ensuring uniformity in case difficulty levels. All cases did not involve real clinical patient data. Concurrently, a structured scoring rubric was developed, with a total score of 100 points. This rubric comprised 2 core dimensions: 80 points allocated to the completeness and accuracy of MHT content, and 20 points dedicated to the proficiency of MHT communication skills. Serving as a core functional module of the LLM-VSPs system, this rubric was designed to facilitate students' completeness and logicality of MHT content through a cycle of "practice-assessment-feedback-correction." Furthermore, given the existing constraints of AI systems in simulating nonverbal communication, fine-grained assessment of MHT interpersonal skills remained unfeasible [23]; therefore, the scoring weight leaned more heavily toward content items. A total of 13 cases

XSL•FO
RenderX

were developed: 11 for student practice and 2 for baseline assessment (see Multimedia Appendix 2 for detailed cases and the scoring rubrics). Overall, 8 diagnostic course instructors of the current semester were invited to conduct a comprehensive review and refinement of all LLM-VSP cases, to safeguard the accuracy and instructional applicability of the cases.

Participants

The participants of this study consisted of 276 third-year medical students enrolled in the *diagnostics* course at Xiangya Hospital of Central South University during the spring semester of the 2024-2025 academic year. These students were from 3 cohorts: the 2022 entry class of the 5-year clinical medicine program, the 2022 entry class of the 5-year stomatology program, and the 2022 entry class of the 5-year anesthesiology program. All participating students had no previous training experience in MHT skills prior to the study. The inclusion criteria for this study were defined as follows: (1) being enrolled in the *diagnostics* course during the current semester and (2) voluntarily participating in this study and providing informed consent. The exclusion criteria were established as follows: (1) failure to complete the baseline assessment of MHT skills at the start of the semester and (2) failure to complete the postintervention assessment of MHT skills at the end of the semester. Upon completion of the research, 108 students were excluded (107 students failed to complete the baseline test and 1 student failed to complete the final test), and finally 168 students were included in data analysis.

To determine student participation in the LLM-VSP practice sessions, a registration link was distributed to all students enrolled in this study. According to the principle of voluntary participation, all eligible students were assigned to one of the two groups: the experimental group (group A) received MHT training via LLM-VSPs–based practice, while the control group (group B) undertook conventional teaching for the same clinical skill training. However, according to the minimum effective practice criteria established in this study, among the students who had voluntarily registered for the LLM-VSPs–based practice group, 9 individuals failed to meet the aforementioned minimum practice threshold based on their practice records. They were deemed to have not effectively participated in the LLM-VSPs–based practice. According to the per-protocol analysis principle, the aforementioned students were reclassified into the control group. The final actual group allocation resulted in 120 students in the experimental group and 48 students in the control group (see Multimedia Appendix 3 for the detailed participant screening flowchart).

Intervention Measures

In this study, students in group A used LLM-VSPs to practice history taking. In alignment with the curriculum schedule and teaching progress of the *diagnostics* course, starting from the third week of the semester, each student in group A was assigned 1 LLM-VSP–based clinical case per week, resulting in a total of 11 cases distributed throughout the intervention period. They could complete the autonomous practice of LLM-VSPs for history taking during their spare time through a mobile app. There were no restrictions on the number of practices and the duration of each practice. After each practice, the system would immediately generate AI-based assessment feedback. Students in group B did not use LLM-VSPs for the practice. Both groups of students received conventional teaching on MHT skills, including sessions with SPs and bedside practical training, in accordance with the predefined curriculum schedule of the *diagnostics* course.

To ensure the authenticity and effectiveness of the LLM-VSPs intervention, we invited diagnostic course instructors of the current semester to conduct simulation tests. The results indicated that skillfully completing a single MHT session comprising basic self-introduction, chief complaint, history of present illness, past medical history, personal history, and family history required a minimum duration of 4 minutes. Furthermore, according to the LLM-VSPs structured scoring rubric, inquiring once for each module of the MHT content, including self-introduction, patient general information, chief complaint, history of present illness, and past medical history, would achieve a total score of 25 points. The research team unanimously agreed that practice sessions failing to meet the aforementioned criteria could not achieve the training objectives of structured MHT logic and content completeness and were considered ineffective practice. Therefore, this study established the minimum effective practice criteria as a single practice session duration of ≥4 minutes and an AI score of ≥25 points.

Data Collection

Baseline Characteristics

Demographic data, including sex and academic major, were retrieved from the university's student administration records. Sex was recorded as male or female. After completing the theoretical course of MHT and the practical training course of *Medical Interview and Communication Skills* at the start of the semester, 2 case-based MHT tests were administered to all participants via the DoctorU instructional system. The baseline score for each student was determined using the average of their AI-generated assessment scores from these 2 tests.

Outcome Measures

At the end of the semester, a postintervention assessment of MHT skills was administered to all participating students. The assessment was carried out in the format of Objective Structured Clinical Examination (OSCE) stations, with real SPs serving as simulated patients for the history-taking interaction. The assessment comprised 4 stations, each corresponding to a distinct clinical case scenario: coronary heart disease, pneumonia, urinary tract infection, and hyperthyroidism. Notably, none of these scenarios had been included in the LLM-VSP practice cases library prior to the assessment. All participants were randomly assigned by computer to one of the 4 assessment stations to complete the MHT assessment. Two standardized and trained examiners independently scored each candidate's performance. The final assessment score was the average of the scores assigned by the 2 examiners. The intraclass correlation coefficient (ICC) based on a 2-way random effects model (absolute agreement) was used to evaluate interrater reliability, which was calculated independently within each assessment case. The scoring rubric was provided by the Department of Diagnostic Medicine, with a total score of 100 (including 60

XSL•FO
RenderX

points for MHT content and 40 points for MHT skills). This rubric served as a summative assessment tool, aiming to comprehensively evaluate students' comprehensive MHT abilities after systematic diagnostic course learning; therefore, the scoring weights for MHT content and skills were set relatively balanced (see Multimedia Appendix 4 for the specific scoring rubrics).

LLM-VSPs Practice Data of Group A

All students' MHT practice data were recorded in detail via the backend of the DoctorU system. Unqualified practice data were excluded in accordance with the predefined minimum threshold for effective practice established earlier in the study. Ultimately, the LLM-VSPs–based MHT practice data of all students in group A were compiled, encompassing the following key metrics: the total number of effective practice sessions, the total number of effective practice cases, the average duration per practice session, the total duration of practice, the highest AI-generated assessment score, and the average AI-generated assessment score.

Statistics

Statistical analyses were performed using R software (version 4.4.2; R Foundation for Statistical Computing). PSM was implemented to mitigate selection bias arising from voluntary group allocation. Matching variables included sex, academic major, baseline test scores, and final test cases, with 1:1 nearest neighbor matching conducted between groups A and B using a caliper width of 0.1 SDs of the logit of the propensity score. Postmatching balance was evaluated through standardized mean differences (SMDs), where SMD<0.1 indicated adequate balance. Final MHT assessment score comparisons between matched groups used independent samples *t* test or *t'* test, and Cohen *d* effect size was calculated. Multivariate linear regression adjusted for covariates was used to confirm intervention effect robustness. To assess the potential impact of unmeasured confounding factors on the results, the Rosenbaum bounds (Γ) were calculated. To verify the robustness of the primary findings against protocol deviations, a sensitivity analysis was conducted following the intention-to-treat (ITT) principle. Specifically, the 9 students who failed to meet the practice threshold were retained in the intervention group according to their original allocation. PSM and subsequent comparative analyses (including independent samples *t* tests and multivariate linear regression) were repeated on this ITT sample to ensure the consistency of the conclusions.

An exploratory analysis was conducted within group A, using Pearson correlation analysis and partial correlation analysis (adjusted for baseline scores) to investigate the associations between practice behavior metrics and MHT scores, and a multivariate linear regression model was further constructed to explore the independent predictive effect of practice behaviors. The model included baseline scores as a covariate, and practice behavior metrics were selected based on the results of the preliminary correlation analyses.

To investigate the heterogeneity of intervention effects across baseline proficiency levels, a generalized linear model was constructed. This model incorporated an interaction term between group assignment and baseline scores (as a continuous variable), while adjusting for sex, academic major, and final test case. This analysis was conducted in both the matched and overall samples. Furthermore, students were divided into high-, medium-, and low-level subgroups based on the 33rd and 67th percentiles of baseline scores. An exploratory subgroup analysis was performed to describe the outcome distributions across these subgroups. The subgroup models were adjusted for final test cases (with the overall sample additionally adjusting for baseline scores), excluding sex and academic major. Results are presented as adjusted means with SEs.

All tests were 2-tailed with the significance threshold set at α=.05.

Ethical Considerations

The study was reviewed by the Clinical Medical Ethics Review Committee of Xiangya Hospital of Central South University. The ethics committee reviewed the study in its entirety and concluded that the research was eligible for exemption from formal ethical review, as it complied with Specific Applicable Situations 1 and 2 of Article 32 of the Measures for Ethical Review of Life Science and Medical Research Involving Humans (2023 edition) issued by the National Health Commission of the People's Republic of China [24]. An official ethical review comment letter was issued (ethics committee record number 2026081836).

Although the study was exempt from formal ethical review, comprehensive information regarding the research was fully provided, and written informed consent was secured from all participating students before enrollment.

All data were deidentified prior to analysis. No individual student data or identifiable information are disclosed in this manuscript.

Participants received no compensation.

Results

Baseline Characteristics of the Participants

Table 1 shows significant prematching differences between group A (n=120) and group B (n=48) in academic major distribution (*P*=.02) and baseline assessment scores (*P*=.01). Owing to computerized random station allocation, the distribution of final test cases was well-balanced between the two groups (*P*=.66). Following matching, 80 participants (40 in each group) were retained for final analysis. All SMDs between matched groups fell below 0.1, demonstrating balanced demographic characteristics (sex), academic majors, baseline test scores, and final test cases between the 2 groups (Figure 2).

XSL•FO
RenderX

Table 1. Baseline characteristics of participants before and after propensity score matching.

Variables	Before matching					After matching					SMD[a]
	Total (N=168)	Group A (n=120)	Group B (n=48)	Test statistic	*P* value	Total (n=80)	Group A (n=40)	Group B (n=40)	Test statistic	*P* value	
Sex , n (%)				0.54 (1)[b]	.46				0.06 (1)[b]	.81	
Male	61 (36.3)	41 (34.2)	20 (41.7)			26 (32.5)	14 (35.0)	12 (30.0)			0.050
Female	107 (63.7)	79 (65.8)	28 (58.3)			54 (67.5)	26 (65.0)	28 (70.0)			–0.050
Academic major , n (%)				7.74 (2)[b]	.02				—[c]	1.00	
Stomatology	39 (23.2)	32 (26.7)	7 (14.6)			16 (20.0)	9 (22.5)	7 (17.5)			0.050
Clinical medicine	110 (65.5)	71 (59.2)	39 (81.3)			59 (73.8)	28 (70.0)	31 (77.5)			–0.075
Anesthesiology	19 (11.3)	17 (14.2)	2 (4.2)			5 (6.3)	3 (7.5)	2 (5.0)			0.025
Baseline test score, mean (SD)	63.2 (11.2)	64.6 (10.8)	59.7 (11.5)	2.57 (166)[d]	.01	61.4 (10.3)	61.3 (9.2)	61.4 (11.4)	–0.04 (78)[d]	.97	0.008
Final test case , n (%)				1.59 (3)[b]	.66				0.20 (3)[b]	.98	
Pneumonia	48 (28.6)	34 (28.3)	14 (29.2)			23 (28.8)	12 (30.0)	11 (27.5)			0.025
CHD[e]	42 (25.0)	32 (26.7)	10 (20.8)			19 (23.8)	9 (22.5)	10 (25.0)			–0.025
Hyperthyroidism	36 (21.4)	23 (19.2)	13 (27.1)			19 (23.8)	10 (25.0)	9 (22.5)			0.025
UTI[f]	42 (25.0)	31 (25.8)	11 (22.9)			19 (23.8)	9 (22.5)	10 (25.0)			–0.025

[a]SMD: standardized mean difference. SMD value of less than 0.1 is widely recognized in clinical research, educational evaluation, and social science studies as a critical threshold for excellent baseline balance between 2 comparison groups.

[b]Chi-square test (*df*).

[c]For "academic major" after matching, the *P* value was calculated using the Fisher exact test, which does not produce a test statistic (indicated by "—").

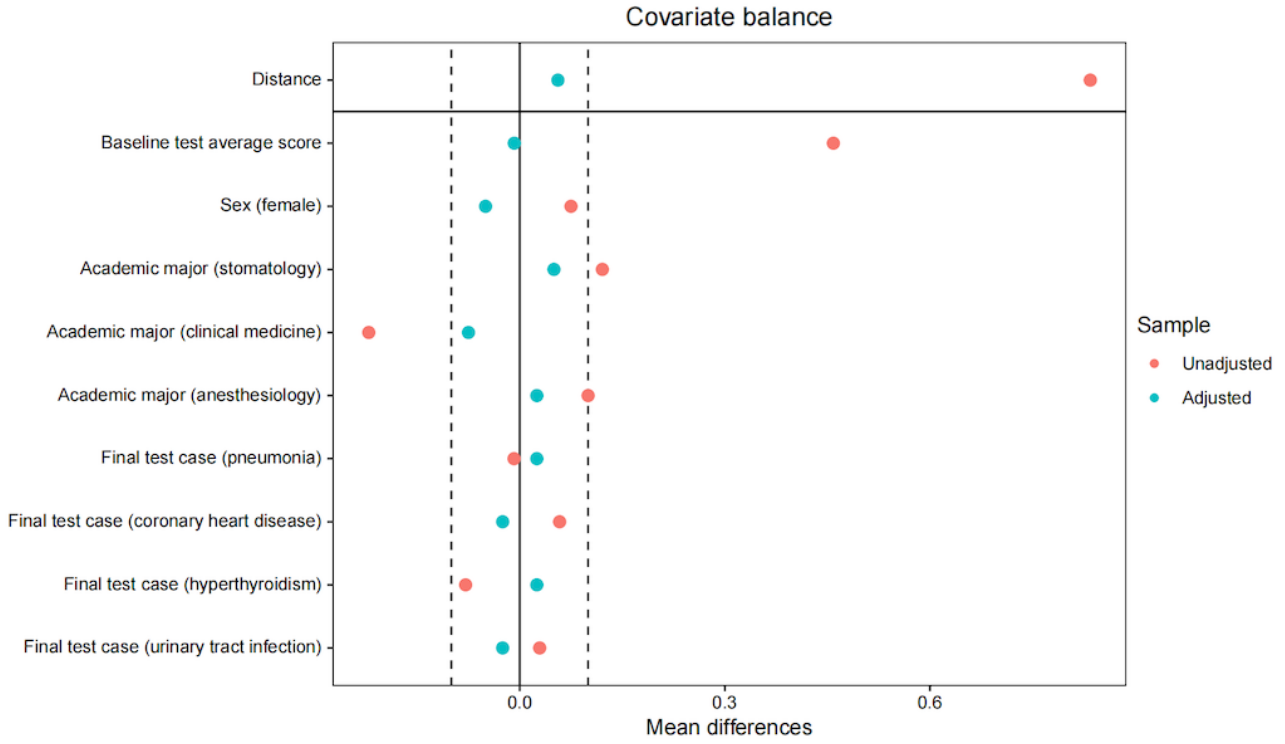
[d]*t* test (*df*).

[e]CHD: coronary heart disease.

[f]UTI: urinary tract infection.

XSL•FO
RenderX

Figure 2. Propensity score matching Love plot.



Comparison of the Final MHT Assessment Scores

Interrater reliability analysis of the final MHT assessment scores revealed ICCs across all 4 station cases ranging from 0.65 to 0.80 (Table S1 in Multimedia Appendix 5), indicating moderate to good agreement between examiners across the different stations. The postmatching comparison results of the final test scores for MHT between the 2 groups after intervention are presented in Table 2. The average total score of group A in MHT was 3.98 points higher than that of group B (mean 87.71, SD 7.29 vs mean 83.74, SD 8.06; *P*=.02; mean difference 3.98, 95% CI 0.55-7.40; Cohen *d*=0.52). Group A also demonstrated statistically significant advantages over group B in both the MHT content subscore (*P*=.03) and the MHT skill subscore of the assessment (*P*=.04).

Table 2. Comparison of the final medical history-taking (MHT) test scores between the 2 groups after propensity score matching.

Variables	Total (n=80), mean (SD)	Group A (n=40), mean (SD)	Group B (n=40), mean (SD)	Mean Difference (95% CI)	*t* test (*df*)	*P* value	Cohen *d* (95% CI)
Total score of MHT (100 points)	85.72 (7.90)	87.71 (7.29)	83.74 (8.06)	3.98 (0.55-7.40)	2.31 (78)	.02	0.52 (0.06-0.97)
Score of MHT content (60 points)	48.80 (5.93)	50.23 (5.65)	47.38 (5.93)	2.85 (0.27-5.43)	2.20 (78)	.03	0.49 (0.04-0.94)
Score of MHT skills (40 points)	36.92 (2.49)	37.49 (2.13)	36.36 (2.70)	1.13 (0.04-2.21)	2.07 (74.0)	.04	0.46 (0.01-0.91)

Robustness Test

Multivariate linear regression confirmed the robustness of the above results (Table 3). In the unadjusted model, group A demonstrated significantly higher total MHT scores than group B (*B*=3.975; *P*=.02). Following adjustment for covariates (baseline test scores, sex, academic major, and final test cases), the intervention effect remained statistically significant (*B*=3.924; *P*<.001). The model's coefficient of determination R^2 reached 0.708, indicating that the model explained the majority of the variance in final MHT assessment scores. In addition, to assess the potential threat of unmeasured confounding factors to the matching results, this study calculated the Rosenbaum bounds Γ=1.21 for this primary analysis sample.

XSL•FO
RenderX

Table 3. Multivariate linear regression analysis of the final medical history-taking test scores (n=80).

Model and variables	*B*[a] (95% CI)	SE	β[b]	*t* test (*df*)	*P* value	VIF[c] value
Model 1 (without adjusting for covariates)[d]						
Constant	83.738 (81.318 to 86.157)	1.215	N/A[e]	68.898 (78)	<.001	N/A
Group A (reference group=group B)	3.975 (0.553 to 7.397)	1.719	0.253	2.313 (78)	.02	N/A
Model 2 (adjusted for covariates)[f]						
Constant	84.919 (77.770-92.068)	3.585	N/A	23.686 (71)	<.001	N/A
Group A (reference group=group B)	3.924 (1.904 to 5.944)	1.013	0.250	3.873 (71)	<.001	1.007
Baseline test scores (points)	0.012 (–0.089 to 0.114)	0.051	0.016	0.245 (71)	.81	1.039
Female (reference group=male)	3.481 (1.200 to 5.761)	1.144	0.208	3.043 (71)	.003	1.065
Academic major (reference g roup= stomatology)						1.040
Clinical medicine	2.911 (0.233 to 5.589)	1.343	0.163	2.168 (71)	.03	
Anesthesiology	–0.528 (–5.196 to 4.140)	2.341	–0.016	–0.226 (71)	.82	
Final test case (reference group = pneumonia)						1.041
CHD[g]	–15.854 (–18.765 to –12.942)	1.460	–0.860	–10.856 (71)	<.001	
Hyperthyroidism	–4.289 (–7.174 to –1.405)	1.447	–0.233	–2.965 (71)	.004	
UTI[h]	–6.745 (–9.552 to –3.938)	1.408	–0.366	–4.792 (71)	<.001	

[a]Unstandardized coefficient.

[b]Standardized coefficient.

[c]VIF: variance inflation factor.

[d]Model parameters: R^2=0.064; Radj^2=0.052; $F_{1, 78}$=5.348; *P*=.02.

[e]N/A: not applicable.

[f]Model parameters: R^2=0.708; Radj^2=0.675; $F_{8,71}$=21.552; *P*<.001.

[g]CHD: coronary heart disease.

[h]UTI: urinary tract infection.

Sensitivity Analysis

According to the ITT analysis principle, the 9 students who were reassigned to the control group for failing to meet the practice threshold were reincluded in the intervention group (group A with 129 students and group B with 39 students), and PSM was reperformed to obtain 33 matched pairs. The analysis results are presented in Table S2 in Multimedia Appendix 5, revealing that group A's MHT total scores remained significantly higher than those of group B (mean difference 5.27, 95% CI 1.52-9.02; *P*=.01; Cohen *d*=0.70). The multivariate linear regression results after adjusting for covariates remained significant (*B*=5.190, 95% CI 2.505-7.875; *P*<.001). In addition, the Rosenbaum bounds Γ value for this sensitivity sample increased from 1.21 in the primary analysis to 1.42.

Exploratory Analysis of LLM-VSPs Practice Data and the Final MHT Test Scores.

Table S3 in Multimedia Appendix 5 summarizes the valid LLM-VSPs practice data for students in group A. On average, participants completed 6.25 valid practice sessions covering 5.21 unique clinical cases. Preliminary Pearson correlation analyses revealed that without adjustment for baseline scores, only total duration of practice (*r*=0.183; *P*=.046) and the average AI-generated assessment score (*r*=0.233; *P*=.01) exhibited weak positive linear associations with final MHT test scores. All other practice metrics failed to demonstrate statistically significant linear correlations with the outcome. To exclude baseline scores as an important confounding factor, we further conducted partial correlation analysis (Table S3 in Multimedia Appendix 5). The results showed that after controlling for baseline scores, the correlations of both the total duration of practice ($r_{partial}$=0.075; *P*=.42) and the average AI-generated assessment score ($r_{partial}$=0.097; *P*=.29) with final MHT test scores disappeared. Further multiple linear regression analysis confirmed (Table 4) that after adjusting for baseline, neither the total duration of practice (β=.058; *P*=.56) nor the average AI-generated assessment score (β=.096; *P*=.38) could independently predict final test scores, whereas baseline scores showed a marginal predictive trend (β=.209; *P*=.06). Although the overall regression model showed statistical significance ($F_{3, 116}$=3.947; *P*=.01), the adjusted coefficient of determination Radj^2=0.069 was at a low level.

XSL•FO
RenderX

Table 4. Multiple linear regression analysis predicting the final medical history-taking (MHT) test scores in group A (n=120).[a]

Variables	*B* (95% CI)	SE	β	*t* test (*df*)	*P* value
Constant	71.011 (60.515 to 81.508)	5.300	N/A[b]	13.400 (116)	<.001
Baseline test scores (points)	0.146 (–0.004 to 0.296)	0.076	0.209	1.922 (116)	.06
Total duration of LLM-VSPs[c]–based MHT practice sessions (minutes)[d]	0.006 (–0.015 to 0.027)	0.010	0.058	0.582 (116)	.56
Average AI-generated assessment score for LLM-VSPs–based MHT[d]	0.076 (–0.093 to 0.246)	0.086	0.096	0.891 (116)	.38

[a]Model parameters: R^2=0.093; Radj^2=0.069; $F_{3, 116}$=3.947; *P*=.01.

[b]N/A: not applicable.

[c]LLM-VSP: large language model–powered virtual standardized patient.

[d]The practice behavior metrics were selected as predictors based on their significant associations observed in the preliminary unadjusted correlation analyses.

Baseline Heterogeneity Tests and Subgroup Analyses

Results from the interaction term analysis regarding baseline heterogeneity suggested that the interaction between group assignment and baseline status did not reach statistical significance in either the matched or overall samples (matched sample: *P*=.63; overall sample: *P*=.20; see Tables S4 and S5 in Multimedia Appendix 5). However, the main effect of group assignment indicated that students in the LLM-VSP group had significantly higher final exam scores than those in the control group in both the matched sample (mean difference 3.92, SE 1.02, 95% CI 1.89-5.95; *P*<.001) and the overall sample (mean difference 3.61, SE 0.99, 95% CI 1.65-5.57; *P*<.001). Exploratory subgroup analyses (Table 5) indicated that the between-group mean difference in the high baseline subgroup appeared to be 5.93 (95% CI 2.17-9.70) in the matched sample and 7.04 (95% CI 3.84-10.25) in the overall sample. In contrast, the mean differences observed in the low and moderate baseline subgroups appeared relatively smaller (matched sample: 2.94-2.99; overall sample: 1.29-1.64).

XSL•FO
RenderX

Table 5. Exploratory descriptive analysis of the final medical history-taking test scores by baseline tertiles in overall unmatched and matched samples.[a]

Sample, subgroup, and group	Baseline, mean (SD)	Adjusted final, mean (SE; 95% CI)	Adjusted final (group A – group B), mean difference (95% CI)
Matched sample (n=80)			
Low tertile			2.99 (–1.80 to 7.77)
Group A (n=12)	49.73 (5.10)	87.33 (1.70; 83.80-90.87)	
Group B (n=14)	N/A[b]	84.35 (1.53; 81.16-87.53)	
Medium tertile			2.94 (–0.70 to 6.58)
Group A (n=14)	61.56 (3.15)	87.00 (1.20; 84.50-89.49)	
Group B (n=13)	N/A	84.05 (1.28; 81.41-86.70)	
High tertile			5.93 (2.17 to 9.70)
Group A (n=14)	72.42 (5.18)	88.14 (1.24; 85.56-90.72)	
Group B (n=13)	N/A	82.21 (1.26; 79.60-84.82)	
Overall unmatched sample (N=168)			
Low tertile			1.64 (–2.36 to 5.63)
Group A (n=33)	50.81 (6.28)	84.27 (1.29; 81.69-86.86)	
Group B (n=23)	N/A	82.64 (1.51; 79.61-85.66)	
Medium tertile			1.29 (–2.37 to 4.94)
Group A (n=42)	63.56 (2.60)	85.79 (0.88; 84.02-87.56)	
Group B (n=14)	N/A	84.50 (1.57; 81.34-87.66)	
High tertile			7.04 (3.84 to 10.25)
Group A (n=45)	75.27 (5.54)	88.41 (0.70; 87.02-89.81)	
Group B (n=11)	N/A	81.37 (1.44; 78.49-84.25)	

[a]Matched sample models adjusted for final test case; overall unmatched sample models adjusted for final test case and baseline score. Given that the interaction analysis showed test case explained the largest proportion of variance (partial η^2>0.43), whereas sex and major contributed minimally, these variables were not included in the models.

[b]N/A: not applicable.

Discussion

Overview

This study introduced LLM-VSPs as a self-directed practice tool for MHT in an authentic educational setting and evaluated their effectiveness through PSM and multidimensional robustness checks. The results indicated that the integration of LLM-VSPs effectively enhanced medical students' MHT competency. Nevertheless, exploratory analysis suggests that this benefit may not be uniformly distributed across student cohorts with varying baseline proficiencies. Specifically, while students across all proficiency levels appeared to benefit from the practice, the high baseline students with a robust theoretical knowledge foundation tended to show a greater magnitude of improvement, whereas the gains for students with medium to low baseline levels seemed relatively limited. These observations imply that the effectiveness of LLM-driven clinical skills training tools may be influenced by learners' prior knowledge levels, offering critical reference points for the practical deployment of such tools and the formulation of stratified teaching strategies.

Intervention Efficacy and Robustness

Analysis of the matched samples revealed that the introduction of LLM-VSPs had a significant positive effect on enhancing medical students' MHT skills (mean difference 3.98; *P*=.02). The positive effect remained robust (*B*=3.924; *P*<.001) even after further controlling for known confounders using a multiple linear regression model. For interventional research conducted in undergraduate teaching settings, randomized controlled trials (RCTs) represent the gold standard for causal inference. Nevertheless, random allocation of educational resources may undermine teaching homogeneity and educational equity. Therefore, this study adopted a prospective cohort design based on voluntary registration and PSM.

PSM [25,26] can simulate the randomization process of an RCT, reconstructing a new sample set with balanced baseline characteristics and thus establishing a methodological foundation for between-group comparisons. However, controlling for observed confounders is not equivalent to eliminating hidden bias [27]. A Rosenbaum bounds analysis yielded a Γ value of merely 1.21, suggesting that the conclusions remain somewhat sensitive to relatively minor hidden biases. Given the voluntary enrollment design, learning motivation is the most prominent

XSL•FO
RenderX

potential unobserved confounder [28]. This raises the question of whether the outstanding performance of the intervention group was influenced by stronger learning motivation rather than the actual effect of the LLM-VSPs. Results from the ITT sensitivity analysis provide tentative supportive evidence. By retaining 9 students who voluntarily enrolled but failed to meet the effective practice standards—implying relatively lower motivation—in the intervention group, we artificially diluted the group's motivational advantage. Instead of diminishing, the between-group difference further expanded (mean difference 3.98, 95% CI 0.55-7.40 vs mean difference 5.27, 95% CI 1.52-9.02), the effect size increased from 0.52 to 0.70, and the Rosenbaum bounds Γ value rose to 1.42. The results suggest that motivational bias likely did not inflate the intervention effect. However, we must interpret this finding cautiously. Since rematching altered the sample composition (from 40 to 33 pairs) and these 9 students might differ in attributes other than motivation, we cannot rely solely on this sensitivity analysis to definitively rule out hidden bias on the intervention effects. Nevertheless, these findings still provide supplementary support for the robustness of the LLM-VSPs intervention effects.

This finding aligns with the results of Luo et al [29] regarding virtual patients improving ophthalmology history-taking skills. LLM-VSPs provide medical students with a secure, repeatable, and accessible practice platform [30], allowing for repeated trial-and-error and consolidation of learning, thereby helping them improve their MHT competency. Furthermore, real human SPs possess greater clinical contextual authenticity and induce performance pressure; students may experience anxiety that impedes knowledge retrieval and expression. Repeated simulated interactions with LLM-VSPs can also enhance students' self-efficacy, enabling them to remain more composed when facing real SPs [31].

Dimensional Differences in Efficacy

A detailed analysis of the dimensional characteristics of the LLM-VSPs intervention effects revealed that the MHT score improvements facilitated by LLM-VSPs were primarily reflected in the content dimension, while the enhancement in the skills dimension was relatively limited. This finding aligns closely with the discrepancy in scoring weights between the formative assessment [32] via LLM-VSPs and the end-of-term summative assessment. Such a weighting discrepancy is not a contradiction in the study design but rather stems from the differing evaluation purposes and technical media of the two assessments. LLM-VSPs focus on helping beginners master the correct MHT content framework through a "practice-assessment-feedback-correction" cycle, with the scoring emphasizing the completeness and accuracy of the consultation content. Furthermore, constrained by the current limitations of LLM technology, LLM-VSPs struggle to simulate nonverbal communication elements during consultations, such as eye contact, gestures, and nodding. Students also exhibit challenges in conveying genuine affective reactions in their interactions with LLM-VSPs [33,34]. Consequently, the scoring and feedback primarily concentrate on the quantitative assessment of content completeness. In contrast, the end-of-term summative evaluation based on human SPs aims to comprehensively assess students' overall clinical competence, with more balanced weights assigned to both dimensions. This outcome clearly delineates the current applicability of LLM-VSPs as a training tool for MHT skills. While LLM-VSPs predominantly fulfill the training needs in the content dimension of MHT, they fall short in providing effective practice and precise feedback for interviewing techniques, including the fluency and accessibility of language, empathy, various nonverbal communications [35], and responsiveness to patients' emotions [36]. This finding resonates with recent research [23] conclusions regarding the application of conversational AI in medical education, which indicate that due to current technical bottlenecks in affective computing and multimodal interaction, AI systems are ill-equipped to handle the comprehensive training of communication skills. Nevertheless, with the rapid advancement of LLMs equipped with visual and auditory perception capabilities, future LLM-VSPs are expected to achieve breakthroughs in simulating nonverbal interactions [37], thereby enabling their application in more comprehensive interviewing skills practice.

Practice Behaviors and the Cognitive Threshold

Although the preceding analysis indicated the intervention effects of LLM-VSPs on medical students, the learning mechanisms underlying whether merely providing LLM-VSPs practice tools can lead to skill improvement require further investigation. To this end, this study analyzed the association between LLM-VSP practice behaviors and final MHT scores at a deeper level. We observed a thought-provoking phenomenon: before adjusting for baseline scores, the average AI score was weakly positively correlated with the final exam score, while practice frequency and duration showed no significant association, seemingly corroborating the notion that "practice quality matters more than practice quantity." However, after controlling for baseline scores, the significant correlation between the average AI score and the final MHT scores disappeared. Multiple linear regression further indicated that compared to LLM-VSP practice behavior indicators, baseline scores had a stronger predictive power for medical students' final MHT competence. This suggests that in the self-directed practice mode, merely increasing practice duration or frequency may not necessarily lead to competence improvement, and AI scores appear to be strongly associated with students' prior knowledge levels. Admittedly, these results involved multiple comparisons without strict statistical correction, merely identifying preliminary associations between practice behaviors, baseline scores, and final MHT competence.

Based on the above findings, we speculate that the self-directed use of AI-empowered practice tools may involve a "cognitive threshold." The acquisition of MHT skills is essentially the transformation from "declarative knowledge" to "procedural skills" [1,38]. Baseline assessments mainly reflect the stock of declarative knowledge at the "remembering and understanding" cognitive tier shortly after theoretical coursework, while the final examination demands students deploy procedural skills at the "application and analysis" level in authentic interpersonal clinical encounters. The transition from "knowing what to ask" to "being able to ask fluently" not only necessitates sustained simulated practice [39,40] but, more importantly, requires the prerequisite of mastering solid declarative knowledge.

XSL•FO
RenderX

While LLM-VSPs deliver real-time quantitative feedback that precisely identifies omissions in history-taking content [32] and facilitates deliberate practice [41], learners lacking a robust foundational theoretical framework may fail to thoroughly comprehend the diagnostic reasoning underpinning AI-generated feedback. Instead, they may merely compile inquiry items in a passive manner, confining their training to superficial mechanical repetition. This hinders efficient knowledge transfer and internalization [42,43]. In contrast, students with higher baseline proficiency have already mastered the theoretical principles of history-taking and possess the cognitive capacity to thoroughly interpret AI-generated feedback [44,45]. This allows them to leverage such feedback to rectify omissions in clinical inquiry and remedy knowledge deficits within a virtuous "practice-feedback-correction" cycle. This may be the prerequisite for AI-supported deliberate practice to take effect [41]: effective practice requires not only repetition but also the correction of errors during the process.

Differential Effects by Baseline Proficiency

To verify the above speculation, we further analyzed the potential heterogeneous effect of baseline ability on the LLM-VSP intervention. Strict interaction term tests showed that the interaction between baseline scores and group assignment did not reach statistical significance in either the matched or full samples, suggesting that LLM-VSPs appeared to demonstrate general effectiveness across different baseline levels without presenting significant statistical differences. However, given the limited sample size of this study, which might affect the statistical power of the interaction term analysis [46], we conducted an exploratory subgroup analysis to observe the data characteristics of final exam scores across different baseline levels in greater detail. Considering that PSM might lead to the loss of some high baseline samples [25], we analyzed both the full sample and the matched sample. The results revealed that while students in all baseline subgroups appeared to benefit from the use of LLM-VSPs, the high baseline subgroup showed a trend of larger between-group differences, whereas the differences in the medium and low baseline subgroups were relatively smaller. Moreover, this trend appeared more pronounced in the full sample compared to the matched sample. It should be cautiously noted that this result might be influenced by the small sample size of the control group in that subgroup (n=11). Additionally, the ICC values for several stations in this study were relatively low; such measurement error might weaken the statistical power to some extent, making it difficult to detect potential subtle skill improvements among medium and low baseline students. Nevertheless, it still aligns with our speculation: students in the high baseline group seemed to benefit more from the self-directed practice with LLM-VSPs.

Furthermore, if this trend of differential benefit holds true in larger samples, it offers a plausible explanation for why the overall effect size was only moderate. Given that students with medium or low baseline proficiency constitute the majority of the sample, their relatively limited benefit magnitudes may have diluted the larger benefits observed among the high baseline subgroup, thereby narrowing the overall mean difference. Admittedly, we must acknowledge that since the interaction term analysis did not reach statistical significance, the above heterogeneity analysis remains a trend observation, and future studies with larger samples and more rigorous designs are needed for confirmation.

Implications for Stratified Teaching

These findings offer important reference points for the application of LLM-VSPs in *diagnostics* education. As an after-class self-directed learning tool, LLM-VSPs can enhance medical students' MHT competency. However, maximizing the pedagogical benefits of this tool requires further reflection in teaching practice. Based on the trends observed in the aforementioned analysis, we speculate that merely increasing practice duration or frequency may not necessarily guarantee competence improvement, and the effectiveness of practice quality might be constrained by baseline ability. Therefore, paying attention to students' baseline levels and implementing stratified interventions may help optimize teaching outcomes. For students who already possess a solid theoretical foundation, self-directed use of LLM-VSPs can facilitate the transformation of knowledge into skills. Nevertheless, for learners with medium or low baseline proficiency, structured guided support may be required to replace independent exploration to maximize the intervention benefits delivered by LLM-VSPs. Specifically, using LLM-VSPs under teacher guidance could help these students consolidate their theoretical groundwork and interpret AI feedback, thereby achieving a transition from mechanical repetition to effective practice. However, further empirical research is required to verify whether this hybrid teaching model combining instructor guidance and LLM-VSPs practice can substantially improve MHT proficiency among students with medium and low baseline ability.

Furthermore, the significance of LLM-VSPs extends beyond their immediate educational outcomes to their potential in promoting digital equity in medical education. Even in the current era of widespread AI accessibility, underresourced grassroots medical schools often struggle with limited faculty and SP resources. In these settings, educators face challenges in transforming general-purpose LLMs into effective pedagogical tools [47] and lack the capacity to undertake the laborious training required for human SPs [48]. The standardized LLM-VSPs agent developed in this study is accessible via a WeChat miniprogram. It not only eliminates development and training costs for grassroots medical institutions but also enables students at these institutions to receive high-quality history-taking training comparable to that offered by top-tier medical schools using only a smartphone and internet access. However, the accessibility of a tool does not equate to the efficiency of learning. How to assist learners of varying proficiency levels to derive equally effective practice benefits may be the key to unlocking the potential of AI teaching tools and realizing high-quality educational equity.

Advantages and Limitations

This study validated the effectiveness of LLM-VSPs within an authentic diagnostic teaching context. It adopted PSM and multiple linear regression analysis methods to minimize confounding bias to the greatest extent, providing a methodological paradigm for causal inference in educational settings where RCTs were impractical. Simultaneously, this

XSL•FO
RenderX

study extended beyond simple verification of intervention efficacy; rather, it conducted an in-depth analysis combining various behavioral data from practice with baseline levels. It preliminarily revealed the "cognitive threshold" phenomenon during the self-directed use of LLM-VSPs and identified a trend of differential benefits across different groups, providing data for the subsequent implementation of refined teaching with LLM-VSPs.

Admittedly, this study has certain limitations. First, although we used various statistical methods to control for confounding factors, Rosenbaum bounds sensitivity analysis indicated that our results remained comparatively vulnerable to unmeasured confounders; future studies require more rigorous experimental designs to draw definitive causal inferences. Second, this study was conducted as a single-center investigation with a limited sample size, and PSM also led to a reduction in sample size, which may affect the statistical power and generalizability of the research conclusions. Third, the analyses concerning the association between practice behaviors and final history-taking competency, as well as baseline heterogeneity, were exploratory in nature. The "cognitive threshold" hypothesis and the trend of "greater benefits for high baseline students" proposed in this study were primarily based on trend observations and lack robust statistical support evidence; thus, their causal mechanisms remain to be clarified. In future research, multicenter, large-sample RCTs should be conducted to generate more robust evidence in support of the findings. In addition, the LLM-VSPs agent developed herein has undergone iterative testing and refinement throughout its development phase to secure accurate automated scoring. Nevertheless, follow-up validation studies are still needed to formally establish its measurement reliability.

Conclusion

Based on PSM analysis, this study has indicated that introducing LLM-VSPs as an after-class self-directed practice tool in *diagnostics* education may help improve medical students' MHT competency. Exploratory analysis further suggests a trend that the educational benefits of LLM-VSPs appear more pronounced among high baseline students with firm theoretical foundations, while the improvement for medium or low baseline students is relatively limited. This phenomenon implies that during the self-directed use of LLM-driven learning tools, learners' prior knowledge levels may constrain the efficiency of using instant AI feedback to achieve the transformation from declarative knowledge into procedural skills. Accordingly, we suggest that future teaching practices should consider students' differences in baseline competencies and explore teaching strategies for stratified interventions when deploying LLM-VSPs. High baseline students might attempt self-directed use, whereas for medium and low baseline students, it is necessary to explore a guided teaching model combining instructor guidance and LLM-VSPs. Future research using multicenter, large-sample RCTs is needed to further verify the aforementioned differences in benefits and the effectiveness of stratified teaching strategies.

Acknowledgments

The authors acknowledge the tireless educational efforts of diagnostics faculty in delivering core clinical competencies and participating students for their proactive engagement with this pedagogical innovation.

Funding

This work was supported by the Department of Education of Hunan Province (grant 202401000359) and Central South University (grants 2025jy024, 2025ALK023, 2024jy178, and 2024jy051-1).

Data Availability

The datasets generated or analyzed during this study are available from the corresponding author on reasonable request.

Authors' Contributions

Conceptualization: YH (lead), CC (equal)
Data curation: YH (lead), RY (equal), HC (supporting)
Formal analysis: YH (lead), RY (supporting)
Funding acquisition: JW
Investigation: YH (lead), CC (equal)
Methodology: YH (lead), CC (equal)
Project administration: JW (lead), YH (supporting), CC (supporting)
Resources: JW
Software: YH (lead), CC (equal), WY, FL, XL, ZX, XX, JH, QZ, GY (all supporting)
Supervision: JW
Validation: YH (lead), CC (equal)
Visualization: YH (lead), CC (equal)
Writing—original draft: YH
Writing—review and editing: YH (lead), JW (equal), CC (supporting), WZ (supporting)

XSL•FO
RenderX

Conflicts of Interest

None declared.

Multimedia Appendix 1
Development of the DoctorU teaching system for medical history-taking practice.
[DOCX File , 43 KB - mededu_v12i1e92486_app1.docx]

Multimedia Appendix 2
Scoring rubrics for virtual standardized patient cases.
[DOCX File , 230 KB - mededu_v12i1e92486_app2.docx]

Multimedia Appendix 3
The detailed participant screening flowchart.
[DOCX File , 57 KB - mededu_v12i1e92486_app3.docx]

Multimedia Appendix 4
Scoring rubrics for the final medical history-taking test.
[DOCX File , 32 KB - mededu_v12i1e92486_app4.docx]

Multimedia Appendix 5
Supplementary statistical analyses.
[DOCX File , 30 KB - mededu_v12i1e92486_app5.docx]

References

1. Keifenheim KE, Teufel M, Ip J, Speiser N, Leehr EJ, Zipfel S, et al. Teaching history taking to medical students: a systematic review. BMC Med Educ 2015;15:159 [FREE Full text] [doi: 10.1186/s12909-015-0443-x] [Medline: 26415941]
2. Elendu C, Amaechi DC, Okatta AU, Amaechi EC, Elendu TC, Ezeh CP, et al. The impact of simulation-based training in medical education: a review. Medicine (Baltimore) 2024;103(27):e38813 [FREE Full text] [doi: 10.1097/MD.0000000000038813] [Medline: 38968472]
3. May W, Park JH, Lee JP. A ten-year review of the literature on the use of standardized patients in teaching and learning: 1996-2005. Med Teach 2009;31(6):487-492. [doi: 10.1080/01421590802530898] [Medline: 19811163]
4. Barrows HS. An overview of the uses of standardized patients for teaching and evaluating clinical skills. AAMC. Acad Med 1993;68(6):443-51; discussion 451. [doi: 10.1097/00001888-199306000-00002] [Medline: 8507309]
5. Zerilli T, Fidler BD, Tendhar C. Assessing the impact of standardized patient encounters on students' medical history-taking skills in practice. Am J Pharm Educ 2023;87(4):ajpe8989 [FREE Full text] [doi: 10.5688/ajpe8989] [Medline: 36375843]
6. Cleland JA, Abe K, Rethans J. The use of simulated patients in medical education: AMEE Guide No 42. Med Teach 2009;31(6):477-486. [doi: 10.1080/01421590903002821] [Medline: 19811162]
7. Yang H, Xu Y, Liang N, Chen W, Yan X, Yang P, et al. Standardized patient methodology in mainland China: a nationwide survey. BMC Med Educ 2019;19(1):214 [FREE Full text] [doi: 10.1186/s12909-019-1630-y] [Medline: 31208408]
8. Aranda JH, Monks SM. Roles and Responsibilities of the Standardized Patient Director in Medical Simulation. Treasure Island (FL): StatPearls Publishing; 2026.
9. Zhang X, Zeng D, Wang X, Fu Y, Han Y, He M, et al. Analysis of virtual standardized patients for assessing clinical fundamental skills of medical students: a prospective study. BMC Med Educ 2024;24(1):981 [FREE Full text] [doi: 10.1186/s12909-024-05982-2] [Medline: 39256732]
10. Kononowicz AA, Woodham LA, Edelbring S, Stathakarou N, Davies D, Saxena N, et al. Virtual patient simulations in health professions education: systematic review and meta-analysis by the digital health education collaboration. J Med Internet Res 2019;21(7):e14676 [FREE Full text] [doi: 10.2196/14676] [Medline: 31267981]
11. Hamilton A, Molzahn A, McLemore K. The evolution from standardized to virtual patients in medical education. Cureus 2024;16(10):e71224. [doi: 10.7759/cureus.71224] [Medline: 39525234]
12. Ellaway R, Poulton T, Fors U, McGee JB, Albright S. Building a virtual patient commons. Med Teach 2008;30(2):170-174. [doi: 10.1080/01421590701874074] [Medline: 18464142]
13. Jiang J, Ye MZ, Kwok TT, Wong JYH. GenAI-supported virtual patients in health care education: systematic review. J Med Internet Res 2026;28:e82756 [FREE Full text] [doi: 10.2196/82756] [Medline: 42098926]
14. Wang C, Li S, Lin N, Zhang X, Han Y, Wang X, et al. Application of large language models in medical training evaluation-using ChatGPT as a standardized patient: multimetric assessment. J Med Internet Res 2025;27:e59435 [FREE Full text] [doi: 10.2196/59435] [Medline: 39742453]

XSL•FO
RenderX

15. Helms JT, Crouch TB. Virtual patients, real conversations: ChatGPT advanced voice mode for pain communication training. Med Teach 2026;48(3):357-359. [doi: 10.1080/0142159X.2025.2536149] [Medline: 40705500]
16. Dang HNN, Luong TV, Vo HTH, Nguyen LTK, Tran TN, Pham HTA, et al. Artificial intelligence-powered virtual standardized patients in teaching history-taking skills to medical students: A randomized controlled trial. BMC Med Educ 2026;26(1):984 [FREE Full text] [doi: 10.1186/s12909-026-09305-5] [Medline: 42063049]
17. Raafat N, Harbourne AD, Radia K, Woodman MJ, Swales C, Saunders KEA. Virtual patients improve history-taking competence and confidence in medical students. Med Teach 2024;46(5):682-688. [doi: 10.1080/0142159X.2023.2273782] [Medline: 38084413]
18. Xu L, Xu Q, Liu C, Chen B, Wang C. Virtual standardized patients for improving clinical thinking ability training in residents: Randomized controlled trial. JMIR Med Educ 2025;11:e73196 [FREE Full text] [doi: 10.2196/73196] [Medline: 41359953]
19. Sun N, Zhou X, Yang Z, Zhou Y, Ma R, Wang Z. Application of AI-based virtual standardized patients in physician-patient communication training: A study based on the SEGUE framework. Front Public Health 2026;14:1768518 [FREE Full text] [doi: 10.3389/fpubh.2026.1768518] [Medline: 41988587]
20. Meskó B. Prompt engineering as an important emerging skill for medical professionals: tutorial. J Med Internet Res 2023;25:e50638 [FREE Full text] [doi: 10.2196/50638] [Medline: 37792434]
21. Eysenbach G. The role of ChatGPT, generative language models, and artificial intelligence in medical education: a conversation with ChatGPT and a call for papers. JMIR Med Educ 2023;9:e46885 [FREE Full text] [doi: 10.2196/46885] [Medline: 36863937]
22. Namvarpour M, Pauwels H, Razi A. AI-induced sexual harassment: investigating contextual characteristics and user reactions of sexual harassment by a companion chatbot. Proc ACM Hum Comput Interact 2025;9(7):1-28. [doi: 10.1145/3757548]
23. Gilbert V, Philip P, Llorca P, Samalin L. Use of artificial intelligence-enhanced virtual patients in educational approaches to medical interview training: a systematic review. BMC Med Educ 2026;26(1):588 [FREE Full text] [doi: 10.1186/s12909-026-08804-9] [Medline: 41781962]
24. National Health Commission of the People's Republic of China MoE, Ministry of ScienceTechnology, National Administration of Traditional Chinese Medicine. Measures for ethical review of life science and medical research involving humans. NHC. 2023. URL: https://www.nhc.gov.cn/qjjys/c100016/202302/6b6e447b3edc4338856c9a652a85f44b.shtml [accessed 2026-08-14]
25. Reiffel JA. Propensity score matching: the 'devil is in the details' where more may be hidden than you know. Am J Med 2020;133(2):178-181. [doi: 10.1016/j.amjmed.2019.08.055] [Medline: 31618617]
26. Stuart EA. Matching methods for causal inference: a review and a look forward. Stat Sci 2010;25(1):1-21 [FREE Full text] [doi: 10.1214/09-STS313] [Medline: 20871802]
27. Austin PC. An introduction to propensity score methods for reducing the effects of confounding in observational studies. Multivariate Behav Res 2011;46(3):399-424 [FREE Full text] [doi: 10.1080/00273171.2011.568786] [Medline: 21818162]
28. Nabizadeh S, Hajian S, Sheikhan Z, Rafiei F. Prediction of academic achievement based on learning strategies and outcome expectations among medical students. BMC Med Educ 2019;19(1):99 [FREE Full text] [doi: 10.1186/s12909-019-1527-9] [Medline: 30953500]
29. Luo M, Bi S, Pang J, Liu L, Tsui C, Lai Y, et al. A large language model digital patient system enhances ophthalmology history taking skills. NPJ Digit Med 2025;8(1):502 [FREE Full text] [doi: 10.1038/s41746-025-01841-6] [Medline: 40760042]
30. Cook DA. Creating virtual patients using large language models: scalable, global, and low cost. Med Teach 2025;47(1):40-42. [doi: 10.1080/0142159X.2024.2376879] [Medline: 38992981]
31. Anton NE, Rendina MA, Hennings JM, Stambro R, Stanton-Maxey KJ, Stefanidis D. Association of medical students' stress and coping skills with simulation performance. Simul Healthc 2021;16(5):327-333. [doi: 10.1097/SIH.0000000000000511] [Medline: 33086369]
32. Lipnevich AA, Mattern K, Feddock C. Formative assessment and feedback in medical education: a practical guide: AMEE Guide No. 189. Med Teach 2026;48(6):921-940 [FREE Full text] [doi: 10.1080/0142159X.2025.2569623] [Medline: 41134874]
33. Gilbert A, Carnell S, Lok B, Miles A. Using virtual patients to support empathy training in health care education: an exploratory study. Simul Healthc 2024;19(3):151-157. [doi: 10.1097/SIH.0000000000000742] [Medline: 37639216]
34. Deladisma AM, Cohen M, Stevens A, Wagner P, Lok B, Bernard T, Association for Surgical Education. Do medical students respond empathetically to a virtual patient? Am J Surg 2007;193(6):756-760. [doi: 10.1016/j.amjsurg.2007.01.021] [Medline: 17512291]
35. Lorié Á, Reinero DA, Phillips M, Zhang L, Riess H. Culture and nonverbal expressions of empathy in clinical settings: a systematic review. Patient Educ Couns 2017;100(3):411-424. [doi: 10.1016/j.pec.2016.09.018] [Medline: 27693082]
36. Maicher KR, Zimmerman L, Wilcox B, Liston B, Cronau H, Macerollo A, et al. Using virtual standardized patients to accurately assess information gathering skills in medical students. Med Teach 2019;41(9):1053-1059. [doi: 10.1080/0142159X.2019.1616683] [Medline: 31230496]

XSL•FO
RenderX

37. Ning Y, Ong JCL, Cheng H, Wang H, Ting DSW, Tham YC, et al. How can artificial intelligence transform the training of medical students and physicians? Lancet Digit Health 2025;7(10):100900 [FREE Full text] [doi: 10.1016/j.landig.2025.100900] [Medline: 41047321]
38. Sawyer T, White M, Zaveri P, Chang T, Ades A, French H, et al. Learn, see, practice, prove, do, maintain: An evidence-based pedagogical framework for procedural skill training in medicine. Acad Med 2015;90(8):1025-1033. [doi: 10.1097/ACM.0000000000000734] [Medline: 25881645]
39. Hawkins N, Younan H, Fyfe M, Parekh R, McKeown A. Exploring why medical students still feel underprepared for clinical practice: a qualitative analysis of an authentic on-call simulation. BMC Med Educ 2021;21(1):165 [FREE Full text] [doi: 10.1186/s12909-021-02605-y] [Medline: 33731104]
40. Burgess A, van Diggele C, Roberts C, Mellis C. Tips for teaching procedural skills. BMC Med Educ 2020;20(Suppl 2):458 [FREE Full text] [doi: 10.1186/s12909-020-02284-1] [Medline: 33272273]
41. Abraham RM, Singaram VS. Using deliberate practice framework to assess the quality of feedback in undergraduate clinical skills training. BMC Med Educ 2019;19(1):105 [FREE Full text] [doi: 10.1186/s12909-019-1547-5] [Medline: 30975213]
42. Izquierdo-Condoy JS, Arias-Intriago M, Tello-De-la-Torre A, Busch F, Ortiz-Prado E. Generative artificial intelligence in medical education: enhancing critical thinking or undermining cognitive autonomy? J Med Internet Res 2025;27:e76340 [FREE Full text] [doi: 10.2196/76340] [Medline: 41183320]
43. Pham TD, Karunaratne N, Exintaris B, Liu D, Lay T, Yuriev E, et al. The impact of generative AI on health professional education: a systematic review in the context of student learning. Med Educ 2025;59(12):1280-1289. [doi: 10.1111/medu.15746] [Medline: 40533396]
44. Weidener L, Fischer M. Artificial intelligence teaching as part of medical education: Qualitative analysis of expert interviews. JMIR Med Educ 2023;9:e46428 [FREE Full text] [doi: 10.2196/46428] [Medline: 36946094]
45. Lenihan D. Three effective, efficient, and easily implementable ways to integrate A.I. into medical education. Cureus 2023;15(10):e47204 [FREE Full text] [doi: 10.7759/cureus.47204] [Medline: 37854479]
46. D'Alessandro A, Kim J, Adhikari S, Goff D, Bargagli-Stoffi FJ, Santacatterina M. Modern causal inference approaches to improve power for subgroup analysis in randomized controlled trials. Stat Med 2026;45(3-5):e70436. [doi: 10.1002/sim.70436] [Medline: 41700655]
47. Merkebu J, Samuel A. Humanizing AI training for health professions educators. Med Teach 2026;48(3):360-363. [doi: 10.1080/0142159X.2025.2522237] [Medline: 40632734]
48. Kirkman M, Simmonds S, Pook CJ, Haas-Heger T. Overcoming barriers for the implementation of simulation-based education within lower-resource settings: a medical student perspective. J Clin Nurs 2023;32(11-12):2941-2942. [doi: 10.1111/jocn.16063] [Medline: 34585456]

Abbreviations

ICC: intraclass correlation coefficient
ITT: intention-to-treat
LLM: large language model
LLM-VSP: large language model–powered virtual standardized patient
MHT: medical history taking
OSCE: Objective Structured Clinical Examination
PSM: propensity score matching
RCT: randomized controlled trial
SMD: standardized mean difference
SP: standardized patient
VSP: virtual standardized patient

Edited by A Stone; submitted 02.Feb.2026; peer-reviewed by MA Vicente, O Xia; comments to author 29.May.2026; revised version received 10.Aug.2026; accepted 11.Aug.2026; published 09.Sep.2026.

Please cite as:
He Y, Chen C, Yin R, Zhang W, Chang H, Yang W, Li F, Li X, Xia Z, Xie X, Huang J, Zeng Q, Yang G, Wu J
Effect of Large Language Model–Powered Virtual Standardized Patients on History-Taking Among Undergraduate Medical Students: Propensity-Matched Cohort Study
JMIR Med Educ 2026;12:e92486
URL: https://mededu.jmir.org/2026/1/e92486
doi:10.2196/92486
PMID:

XSL•FO
RenderX

©Yuchen He, Chen Chen, Rong Yin, Wuyang Zhang, Haoli Chang, Wei Yang, Fei Li, Xinhua Li, Zhuying Xia, Xiaoyun Xie, Jing Huang, Qiuming Zeng, Guang Yang, Jing Wu. Originally published in JMIR Medical Education (https://mededu.jmir.org), 09.Sep.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Education, is properly cited. The complete bibliographic information, a link to the original publication on https://mededu.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Perceived Ratings of AI-Generated and Expert-Reviewed vs Clinician-Developed Virtual Cardiology Patient Scenarios Among Medical Students: Randomized Controlled Study

Marios Georgios Bantidos[1,2*], MD; Eleni Dafli[1*], MD, PhD; Efstratios Karagiannidis[2], MD, PhD; Barbara Fyntanidou[2], MD, PhD; Sotiria Gilou[1], MSc, PhD; Athina Nasoufidou[3], MD; Panagiotis Bamidis[1], PhD

[1]Lab of Medical Physics and Digital Innovation, School of Medicine, Aristotle University of Thessaloniki, Thessaloniki, Central Macedonia, Greece

[2]Department of Emergency Medicine, AHEPA University General Hospital, School of Medicine, Aristotle University of Thessaloniki, Thessaloniki, Central Macedonia, Greece

[3]Second Department of Cardiology, Hippokration General Hospital, School of Medicine, Aristotle University of Thessaloniki, Thessaloniki, Central Macedonia, Greece

[*]these authors contributed equally

Corresponding Author:
Panagiotis Bamidis, PhD
Lab of Medical Physics and Digital Innovation
School of Medicine
Aristotle University of Thessaloniki
Building D, Entrance 8, 3rd Floor
University Campus
Thessaloniki, Central Macedonia, 54124
Greece
Phone: 30 2310999310
Email: pdbamidis@gmail.com

Abstract

Background: Virtual patients (VPs) are widely used in medical education, but creating high-quality cases requires clinical expertise, iterative revision, and alignment with educational objectives. Generative AI may support VP authoring, but direct comparisons between AI-generated and clinician-developed cases delivered within the same educational setting remain limited.

Objective: This study aimed to compare medical students' learner-perceived ratings of 1 AI-generated, expert-reviewed virtual cardiology patient scenario with those of 1 clinician-developed virtual cardiology patient scenario delivered through the same mobile platform.

Methods: A randomized, participant-blinded 2-arm controlled study was conducted in October 2025 during scheduled fifth-year emergency medicine course sessions. Eligible participants were fifth-year medical students who had completed the cardiology course. Students were assigned 1:1 to complete either a clinician-developed VP (CD-VP; n=64) or a ChatGPT-generated VP (GPT-VP; n=67). Both scenarios followed a shared acute coronary syndrome blueprint and were delivered through the same mobile-optimized app (mobile VPs app [MobiViP]) with identical interface structure and branching logic. The GPT-VP was generated using GPT-5 and refined through iterative follow-up prompting under expert review; direct manual rewriting was prohibited. After completion, students completed a mandatory 19-item questionnaire covering realism and validity, cognitive load, self-reported learning and reflection, and engagement and satisfaction. Analyses included reliability testing, 1-sample tests against a neutral midpoint, Mann-Whitney *U* tests, and Spearman correlations. Full scenario content and development details are provided as multimedia appendices.

Results: A total of 131 complete questionnaires were analyzed, with no item-level missing data. Mean subscale scores in the CD-VP group were 3.81 (SD 0.64) for realism and validity, 3.67 (SD 0.74) for cognitive load, 3.64 (SD 0.73) for self-reported learning and reflection, and 4.16 (SD 0.64) for engagement and satisfaction; corresponding GPT-VP scores were 4.11 (SD 0.56), 3.03 (SD 0.76), 3.86 (SD 0.77), and 4.28 (SD 0.58), respectively. Compared with the CD-VP, the GPT-VP received higher ratings for realism and validity (Mann-Whitney U=1503.0; P=.003) and lower cognitive load ratings, indicating a lower perceived burden (Mann-Whitney U=1070.5; P<.001). Self-reported learning and reflection showed a marginal difference between groups, with a higher rating for the GPT-VP (Mann-Whitney U=1719.0; P=.05), whereas engagement and satisfaction did not differ significantly

XSL•FO
RenderX

(Mann-Whitney U=1922.5; P=.31). Internal consistency was acceptable to good for most subscales, although internal consistency for the realism and validity subscale in the GPT-VP group was moderate (Cronbach α=0.67).

Conclusions: In this single-center randomized study, 1 AI-generated, expert-reviewed cardiology VP scenario was rated similarly to or more favorably than 1 clinician-developed scenario in selected learner-perceived domains. Future research should expand to objective learning, retention, development workload, and scalability outcomes.

(JMIR Med Educ 2026;12:e100839) doi:10.2196/100839

KEYWORDS

cardiology; ChatGPT; generative artificial intelligence; generative AI; medical education; randomized controlled trial; RCT; virtual patients; VPs

Introduction

Case-Based Learning

Case-based learning (CBL) is a core strategy in health professions education, linking biomedical knowledge with clinical reasoning in authentic contexts. Well-designed cases support diagnostic thinking, guideline-based management, and productive uncertainty, and CBL has been associated with improved learner engagement and performance compared with more didactic approaches [1,2]. However, high-quality cases require time, clinical expertise, iterative editing, and ongoing updates as evidence evolves—factors that often slow content refresh cycles [3].

Virtual Patients

Virtual patients (VPs), defined as "interactive computer simulations of real-life clinical scenarios for the purpose of health care and medical training, education or assessment," help scale experiential learning through branching pathways, immediate feedback, and multimedia use [4]. Meta-analyses suggest that VPs are at least noninferior to traditional methods in terms of knowledge outcomes and can strengthen clinical reasoning skills [5,6]. Recent work has also described practical implementation models for undergraduate programs [7]. Still, authoring coherent and realistic VP scenarios remains resource-intensive.

The Role of Generative AI

Generative AI, particularly large language models (LLMs), such as OpenAI's GPT models, including ChatGPT, may support educational content authoring by rapidly generating structured text, adapting material to the learner level, and responding to iterative prompting. The use of generative AI in medical education has expanded rapidly, alongside ongoing debate regarding reliability, transparency, and appropriate oversight [8,9]. Documented limitations, including hallucinations, fabricated references, and hypersensitivity to prompting, raise concerns about independent use in training contexts [10,11]. At the same time, improved performance on licensing-style evaluations (eg, GPT models on United States Medical Licensing Examination [USMLE]–aligned tasks) highlights both rapid progress and substantial variability across domains and item types [12-14]. For interactive clinical scenarios, these limitations are especially relevant because errors may affect not only isolated facts but also the sequence, prioritization, and consequences of clinical decisions. Comparative research on AI-generated educational materials remains mixed. Controlled studies of AI-generated multiple-choice questions (MCQs) and vignettes frequently recommend expert review and report variability in difficulty and factual accuracy [15]. Direct comparisons within the same medical domain and delivery platform remain limited [16].

Evaluation Domains

Evaluation of AI-generated VPs should prioritize the fidelity of the evolving clinical trajectory, guideline concordance, and the plausibility of branching pathways, including consequence-specific corrective ("rescue") routes. These requirements extend beyond factual recall and place substantial demands on automated scenario generation. Such demands are most apparent in acute care specialties, where decisions are sequential and errors compound rapidly. Cardiology provides a stringent test environment, requiring synchronization of multimodal data with time-critical decisions.

Study Objective

This study compared medical students' learner-perceived ratings of 1 AI-generated, expert-reviewed mobile cardiology VP scenario with those of 1 clinician-developed mobile cardiology VP scenario. The aim was to evaluate the acceptability of AI-assisted VP development within a controlled medical education setting.

Methods

Study Design and Setting

This 2-arm comparative study evaluated medical students' assessments of 2 cardiology VP scenarios: a clinician-developed VP (CD-VP) and a ChatGPT-generated VP (GPT-VP). The study was conducted in October 2025 during the fifth-year (undergraduate) emergency medicine course (fall semester). Participants were fifth-year medical students who had completed the cardiology course in the previous academic year. Students accessed the study on their mobile phones via a link distributed during the scheduled teaching session. Scenario allocation was performed using simple 1:1 randomization, with assignment generated at the time of access using a random number generator (randomizer app). Scenarios were presented without authorship labeling, and students were not informed that 1 scenario was AI generated. After providing written informed consent, students were instructed to complete the assigned VP scenario individually and then complete the postscenario questionnaire. No individualized tutoring or content-specific feedback from faculty was provided, and group work was not allowed.

XSL•FO
RenderX

Statistical analyses were performed by blinded independent statisticians, and the principal investigators did not have direct access to group-labeled results during analysis. The study is reported in accordance with the 2025 CONSORT (Consolidated Standards of Reporting Trials) statement [17]. Details are provided in Supplementary Appendix S1 in Multimedia Appendix 1. Because both interventions were web based, a completed CONSORT-EHEALTH (Consolidated Standards of Reporting Trials of Electronic and Mobile Health Applications and Online Telehealth) checklist is provided as Multimedia Appendix 2 [18].

Ethical Considerations

The study was conducted under the overall approval of the Bioethics and Deontology Committee of the Aristotle University of Thessaloniki (266491/2022). As this was an educational, nonclinical, minimal risk study conducted within the scope of the approved research program, a separate ethics approval was not required under institutional procedures. However, the study was reviewed under the institutional data protection and General Data Protection Regulation (GDPR) procedures of the Aristotle University of Thessaloniki (322/31-10-2024). The institutional determination confirmed that the study used anonymized postscenario learner-perceived data collected within scheduled undergraduate teaching and did not involve patients, patient-level clinical outcomes, or identifiable health data. All participants provided written informed consent before participation. Participation was voluntary, and participants received no compensation.

The study was not prospectively registered because it evaluated an educational intervention, did not involve patients or patient-level clinical outcomes, and was not a therapeutic, diagnostic, or health service intervention requiring registration under the applicable institutional and trial registration criteria.

VP Development and Delivery

To ensure structural and pedagogical comparability, both VP scenarios were developed based on a shared blueprint focused on acute coronary syndrome (ACS), covering identification, clinical progression, and management, and emphasizing high-yield clinical reasoning and learning objectives relevant to undergraduate medical training according to the 2023 European Society of Cardiology guidelines for the management of ACS [19]. Table 1 summarizes key design features of the CD-VP and GPT-VP.

The CD-VP was clinician developed by a 5-member team that included 2 senior professors of medicine, 1 in cardiology and 1 in emergency medicine. The same 2 senior clinical experts also reviewed the GPT-VP, providing shared oversight across both development pathways. The GPT-VP underwent 2 substantive expert review rounds before implementation. In the first round, the full scenario was reviewed for clinical sequence, decision nodes, answer options, automated feedback, rationales, and planned visual material. Review was not blinded to AI authorship because it was conducted as formative quality assurance of the AI-assisted scenario before student use rather than as an independent comparative rating exercise. Expert comments were reconciled through discussion, and final readiness was determined by consensus between the cardiology and emergency medicine experts after the second review round.

The GPT-VP was generated in September 2025 using OpenAI's ChatGPT (Plus subscription) with GPT-5, using both the standard mode and the extended reasoning "Thinking" setting (GPT-5 Thinking) under constrained prompting. The initial draft underwent structured expert review and targeted revision prior to implementation (Figure 1). Importantly, revisions were implemented through iterative follow-up prompting rather than direct manual modification of the scenario content. Review domains, representative modifications, and prompting workflow details are summarized in Supplementary Appendix S2 in Multimedia Appendix 1. The full CD-VP and GPT-VP scenario contents are provided in Multimedia Appendix 3. The complete chronological ChatGPT conversation history was not exported or retained as a verbatim prompt transcript; therefore, exact prompt-by-prompt reproduction is not possible. To improve transparency, Multimedia Appendix 1 reports the available development record, including the reconstructed initial prompt template, main follow-up prompt categories, mode use approach, 2 substantive expert-guided revision rounds, and criteria used to accept, revise, regenerate, or reject content.

Both scenarios were delivered via the mobile VPs app (MobiViP), a web-based platform optimized for mobile devices that supports branching decision pathways, integrated clinical media, and automated option-specific feedback. The app is available online [20]. The platform environment, user interface, and feedback architecture were standardized across study arms to ensure consistent delivery conditions [21].

XSL•FO
RenderX

Table 1. Structural, development, and delivery features of the 2 virtual patient scenarios[a].

Features	CD-VP[b]	GPT-VP[c]
Development pathway	Clinician developed	AI generated using ChatGPT (GPT-5), followed by expert review and iterative prompt-based revision
Development and/or review team	5-member development team: 1 medical student, 1 resident, 1 fellow, 1 cardiology professor, and 1 emergency medicine professor	2 expert reviewers: 1 cardiology professor and 1 emergency medicine professor, both of whom were senior members of the CD-VP development team
Approximate development calendar time	1 weekend day plus 2 additional afternoons, including team coordination and retrieval and selection of existing clinical media	Half a day, including prompt-based drafting, expert review, and final approval before implementation
Delivery platform	MobiViP[d]	MobiViP
Delivery format	Mobile-optimized branching VP[e] scenario	Mobile-optimized branching VP scenario
Core clinical domain	ACS[f] and post-MI[g] management	ACS and post-MI management
Number of decision points	16	16
Educational level	Undergraduate medical students	Undergraduate medical students
Initial presentation	STEMI[h]-focused acute chest pain presentation	STEMI-focused acute chest pain presentation
Main learning emphasis	ACS recognition, urgent reperfusion, catheterization laboratory decision-making and complication management, subsequent HFrEF[i] treatment, and device-therapy decisions	ACS recognition, antiplatelet and anticoagulant selection, rhythm and complication management, subsequent HFrEF treatment, second ACS management and treatment algorithm shift, and device-therapy decisions
Clinical media	Electrocardiograms, angiography, echocardiograms, and monitor-based clinical material	Scenario images, chest radiography, and monitor-based clinical material
Automated feedback structure	Option-specific	Option-specific with explanatory rationale
Completion time or time on task	No strict time limit within the practical limits of the scheduled teaching session (individual time on task was not prospectively recorded)	No strict time limit within the practical limits of the scheduled teaching session (individual time on task was not prospectively recorded)
Information volume	Not objectively quantified	Not objectively quantified
Full scenario content	Provided in Multimedia Appendix 3	Provided in Multimedia Appendix 3

[a]Case-specific differences in clinical trajectory, information volume, media, wording, and feedback style may have influenced learner ratings. Approximate development time was retrospectively estimated by the author team and was not prospectively time-logged; therefore, these values are descriptive and should not be interpreted as a formal workload or efficiency comparison.

[b]CD-VP: clinician-developed virtual patient.

[c]GPT-VP: ChatGPT-generated virtual patient.

[d]MobiViP: mobile virtual patients app.

[e]VP: virtual patient.

[f]ACS: acute coronary syndrome.

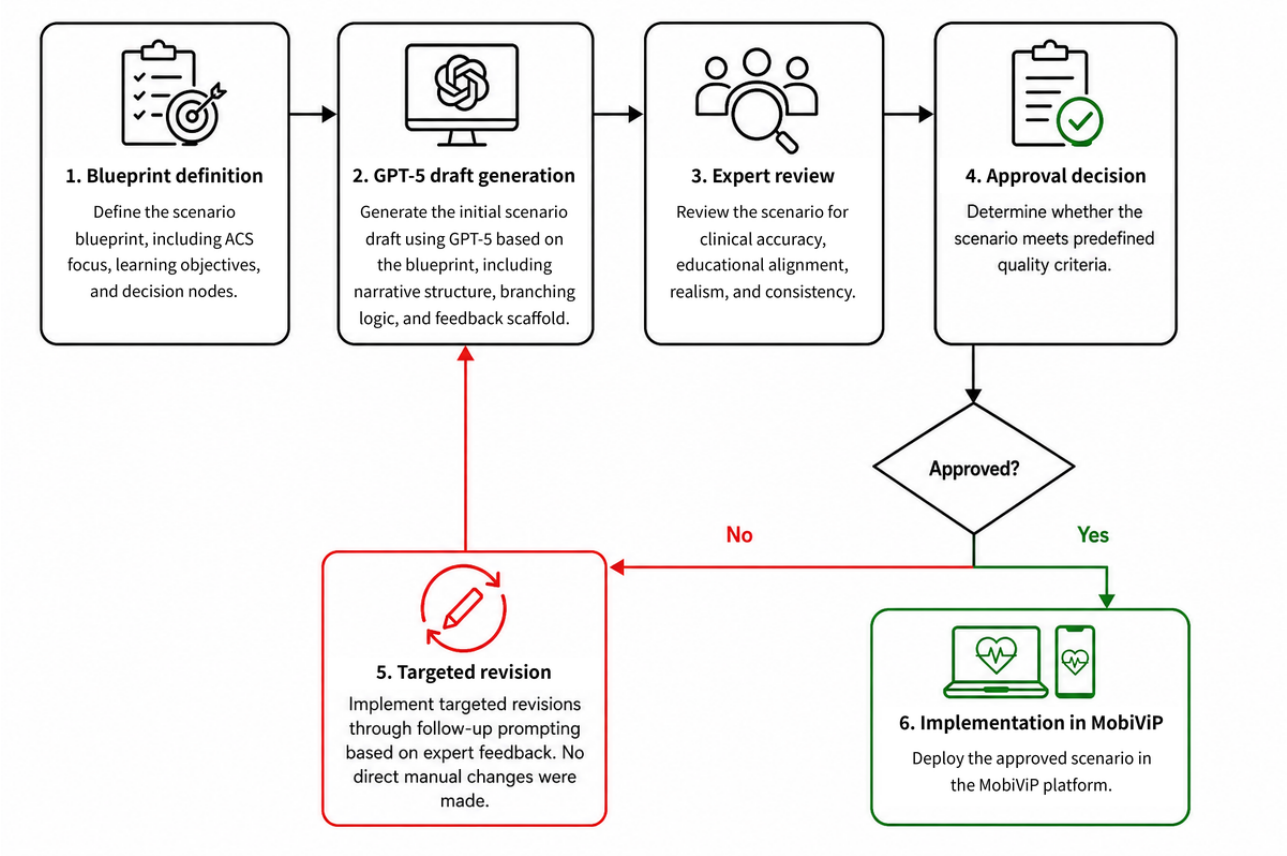
[g]MI: myocardial infarction.

[h]STEMI: ST-elevation myocardial infarction.

[i]HFrEF: heart failure with reduced ejection fraction.

XSL•FO
RenderX

Figure 1. Expert-supervised workflow for development of the ChatGPT-generated virtual patient (GPT-VP). ACS: acute coronary syndrome; MobiViP: mobile virtual patients app.



Measures and Outcomes

The structured postscenario questionnaire consisted of 5-point Likert scale items (1="strongly disagree" to 5="strongly agree") and was configured with mandatory responses for all items, resulting in no item-level missing data among submitted questionnaires. The full questionnaire is provided in Supplementary Appendix S3 in Multimedia Appendix 1. The instrument comprised 19 items grouped a priori into 4 subscales: (1) realism and validity (3 items), assessing perceived realism, internal coherence, and appropriateness of difficulty; (2) cognitive load (3 items), assessing perceived mental effort, time pressure, and cognitive demands (including sustained attention and recall of clinical knowledge); (3) self-reported learning and reflection (6 items), assessing perceived knowledge gains, identification of knowledge gaps, confidence in clinical decision-making, preparedness for similar cases, and motivation for further study; and (4) engagement and satisfaction (7 items), assessing engagement or immersion, perceived usefulness of feedback, overall satisfaction, and willingness to recommend or further integrate similar VPs into the curriculum. The same questionnaire structure and subscale composition were used for both scenarios to ensure comparability of evaluations. For each participant, subscale scores were calculated as the mean of the corresponding items. The questionnaire was developed for this study to capture learner perceptions rather than to serve as a fully validated measure of objective learning or clinical performance. Item grouping was defined a priori according to the intended constructs, and no claim of full psychometric validation is made.

The primary outcomes were between-scenario differences (CD-VP vs GPT-VP) in the 4 subscale scores. Secondary outcomes included comparisons of subscale scores against the neutral midpoint of the Likert scale (3) and exploratory analyses of within-scenario correlations between subscales.

Statistical Analysis

Descriptive statistics (sample size, mean values, SDs, and score ranges) were calculated for each subscale by scenario. Normality was assessed using the Shapiro-Wilk test. Internal consistency of each subscale was evaluated using Cronbach alpha, calculated separately for CD-VP and GPT-VP; scores ≥0.70 were considered acceptable.

To examine whether student ratings differed from the neutral midpoint (3), 1-sample tests were performed using Wilcoxon signed-rank tests when normality assumptions were violated; where normality was met (self-reported learning and reflection subscale in the CD-VP), a 1-sample, 2-tailed *t* test was applied. Between-scenario comparisons were performed using Mann-Whitney *U* tests. For interpretability, between-group differences were additionally expressed as a common language effect size (CLES; probability of superiority), derived from the Mann-Whitney *U* statistic and defined as the probability that a randomly selected GPT-VP participant would report a more favorable subscale score than a randomly selected CD-VP participant, with 95% CIs. For cognitive load, "more favorable" was defined as a lower score.

Within-scenario associations between questionnaire subscales were examined using Spearman rank-order correlation

XSL•FO
RenderX

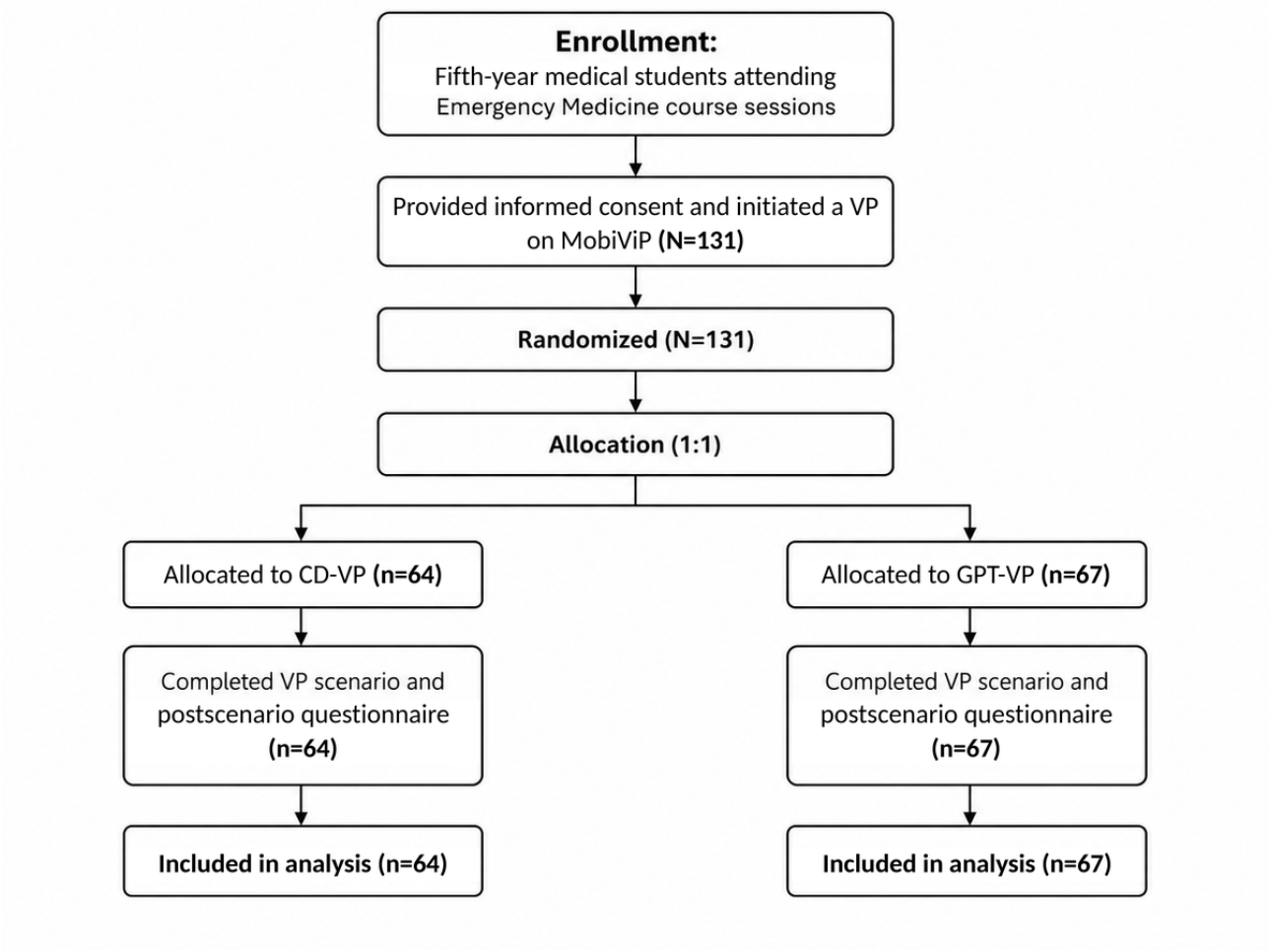
coefficients, calculated separately for CD-VP and GPT-VP. The following pairs were examined a priori: realism and validity with self-reported learning and reflection; realism and validity with engagement and satisfaction; self-reported learning and reflection with engagement and satisfaction; cognitive load with self-reported learning and reflection; and cognitive load with engagement and satisfaction.

All tests were 2-tailed with $P<.05$. Because this was an exploratory educational study with prespecified learner-perceived subscales, no formal adjustment for multiple comparisons was applied. Accordingly, statistically significant and marginal findings were interpreted cautiously. Analyses were conducted using SPSS (version 29; IBM Corp) and Python (version 3.13.5; Python Software Foundation).

Results

All submitted questionnaires were complete and included in the analysis; a total of 64 students completed the questionnaire for the CD-VP and 67 completed the questionnaire for the GPT-VP. Participant flow is shown in Figure 2.

Figure 2. Participant flow diagram. CD-VP: clinician-developed virtual patient; GPT-VP: ChatGPT-generated virtual patient.



Internal Consistency

Internal consistency for each questionnaire subscale is presented in Table 2. In the CD-VP scenario, Cronbach α indicated acceptable internal consistency for realism and validity (α=0.71) and cognitive load (α=0.70), and good internal consistency for self-reported learning and reflection (α=0.82) and engagement and satisfaction (α=0.87). In the GPT-VP scenario, internal consistency was moderate for realism and validity (α=0.67), acceptable for cognitive load (α=0.70), good for self-reported learning and reflection (α=0.85), and acceptable for engagement and satisfaction (α=0.77).

XSL•FO
RenderX

Table 2. Internal consistency (Cronbach α) of questionnaire subscales for the 2 virtual patient scenarios.

Subscales	Clinician-developed virtual patient		ChatGPT-generated virtual patient	
	Cronbach α	Interpretation[a]	Cronbach α	Interpretation[a]
Realism and validity	0.71	Acceptable	0.67	Moderate
Cognitive load	0.70	Acceptable	0.70	Acceptable
Self-reported learning and reflection	0.82	Good	0.85	Good
Engagement and satisfaction	0.87	Good	0.77	Acceptable

[a]Interpretation thresholds: α<0.70 was considered "moderate," α≥0.70 was considered "acceptable," and α≥0.80 was considered "good."

Descriptive Subscale Scores

Descriptive subscale statistics are presented in Table 3 and visualized in Figure 3. In the CD-VP scenario, mean scores were 3.81 (SD 0.64) for realism and validity, 3.67 (SD 0.74) for cognitive load, 3.64 (SD 0.73) for self-reported learning and reflection, and 4.16 (SD 0.64) for engagement and satisfaction. In the GPT-VP scenario, mean scores were 4.11 (SD 0.56) for realism and validity, 3.03 (SD 0.76) for cognitive load, 3.86 (SD 0.77) for self-reported learning and reflection, and 4.28 (SD 0.58) for engagement and satisfaction. Engagement and satisfaction showed the highest mean score in both scenarios.

Table 3. Descriptive statistics of questionnaire subscale scores by virtual patient scenario.

Subscales and scenarios	Score[a], mean (SD; range)
Realism and validity	
CD-VP[b]	3.81 (0.64; 2.00-5.00)
GPT-VP[c]	4.11 (0.56; 2.67-5.00)
Cognitive load	
CD-VP	3.67 (0.74; 1.67-5.00)
GPT-VP	3.03 (0.76; 1.67-5.00)
Self-reported learning and reflection	
CD-VP	3.64 (0.73; 1.50-5.00)
GPT-VP	3.86 (0.77; 1.67-5.00)
Engagement and satisfaction	
CD-VP	4.16 (0.64; 2.57-5.00)
GPT-VP	4.28 (0.58; 3.00-5.00)

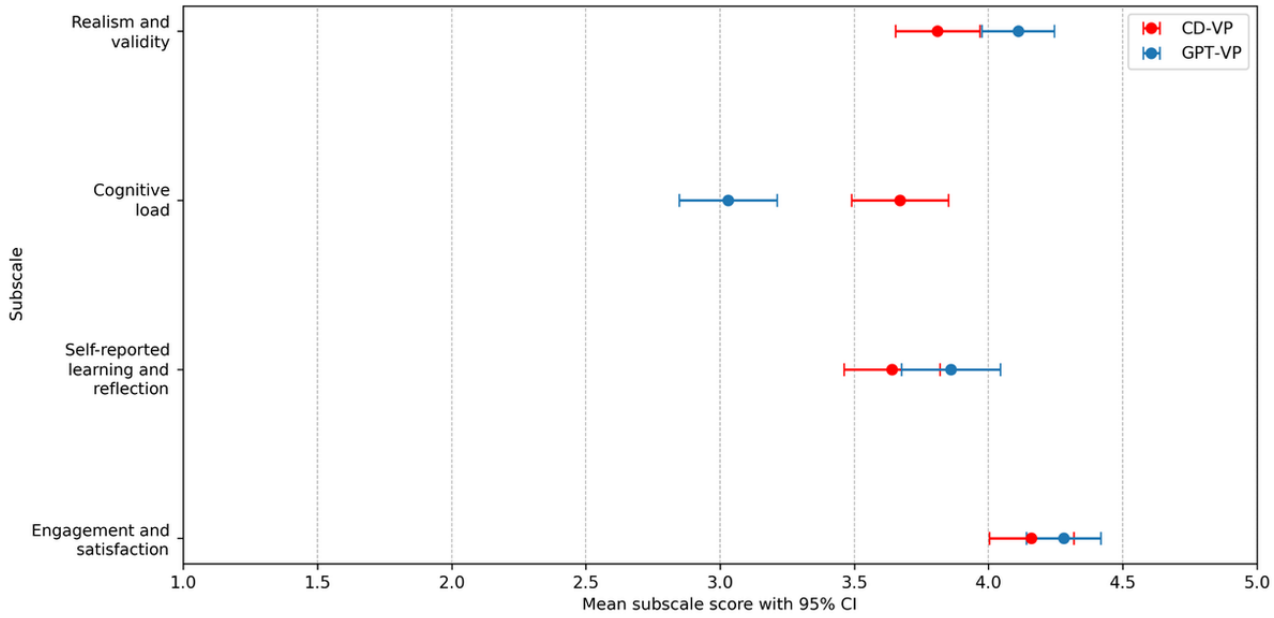
[a]All questionnaire items were rated on a 5-point Likert scale (1="strongly disagree" to 5="strongly agree").

[b]CD-VP: clinician-developed virtual patient, n=64.

[c]GPT-VP: ChatGPT-generated virtual patient, n=67.

XSL•FO
RenderX

Figure 3. Mean student ratings by subscale for clinician-developed virtual patient (CD-VP) and ChatGPT-generated virtual patient (GPT-VP). Points represent mean subscale scores (on a 5-point Likert scale), and whiskers indicate 95% CIs.



Comparisons Against the Neutral Midpoint

One-sample analyses examined whether subscale ratings differed from the neutral midpoint of the Likert scale (3). Because normality was violated for all subscales except self-reported learning and reflection in the CD-VP scenario, Wilcoxon signed-rank tests were used for most comparisons, and a 1-sample *t* test was applied where normality assumptions were met.

In the CD-VP scenario, ratings were significantly above the midpoint for realism and validity (Z=−6.192; P<.001), cognitive load (Z=−5.242; P<.001), and engagement and satisfaction (Z=−6.778; P<.001). Self-reported learning and reflection was analyzed using a 1-sample *t* test and was also significantly above the midpoint (t_{63}=6.920; P<.001), with a mean difference of 0.64 (95% CI 0.45-0.82).

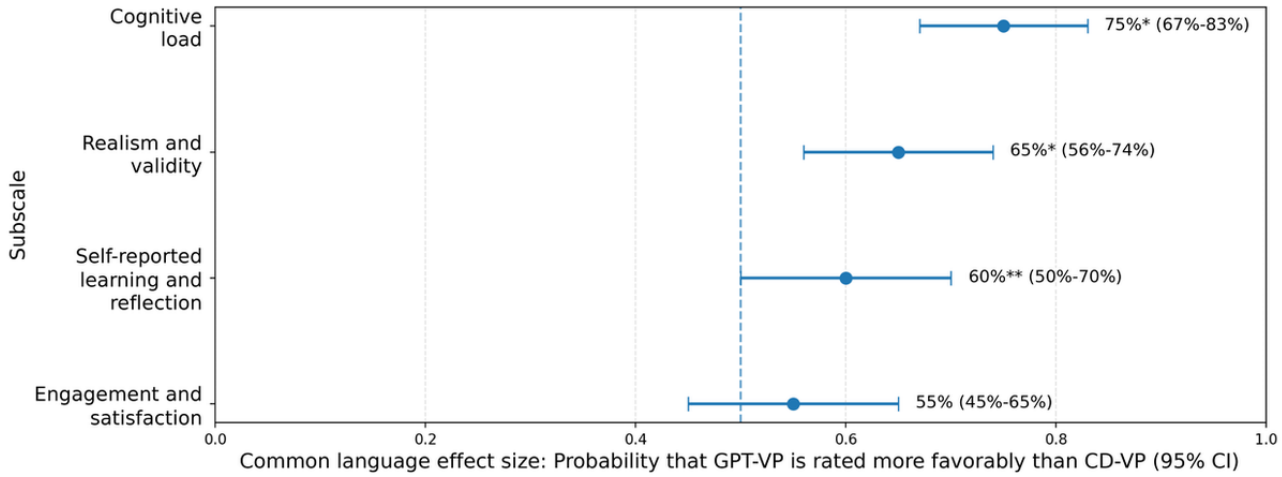
In the GPT-VP scenario, ratings were significantly above the midpoint for realism and validity (Z=−6.916; P<.001), self-reported learning and reflection (Z=−6.052; P<.001), and engagement and satisfaction (Z=−7.071; P<.001). Cognitive load did not differ from the midpoint (Z=−0.210; P=.83).

Between-Scenario Comparisons

Between-scenario differences in subscale scores (CD-VP vs GPT-VP) were assessed using Mann-Whitney *U* tests, as the assumption of normality was violated. Realism and validity differed significantly between scenarios (Mann-Whitney U=1503.0; P=.003), with higher mean ranks in the GPT-VP scenario than in the CD-VP scenario (75.57 vs 55.98). Cognitive load also differed significantly (Mann-Whitney U=1070.5; P<.001), with higher mean ranks in the CD-VP scenario than in the GPT-VP scenario (82.77 vs 49.98). Self-reported learning and reflection showed a marginal between-scenario difference between groups, with a higher rating for the GPT-VP (Mann-Whitney U=1719.0; P=.05), with higher mean ranks in the GPT-VP scenario than in the CD-VP scenario (72.34 vs 59.36). Engagement and satisfaction did not differ significantly between scenarios (Mann-Whitney U=1922.5; P=.31). To aid interpretation, these between-group differences are additionally presented as a CLES (probability of superiority) with 95% CIs (Figure 4).

XSL•FO
RenderX

Figure 4. Common language effect size for between-scenario differences (rank-based). Points represent the probability that a randomly selected participant in the ChatGPT-generated virtual patient (GPT-VP) group reports a more favorable rating (subscale score) than a randomly selected participant in the clinician-developed virtual patient (CD-VP) group; whiskers indicate 95% CIs. The vertical reference line at 0.5 indicates no difference between scenarios (equal chance); values >0.5 favor GPT-VP and values <0.5 favor CD-VP. For cognitive load, "more favorable" is defined as a lower score. **P*<.05; ***P*=.05 (marginally significant).



Within-Scenario Correlations

Within-scenario associations between questionnaire subscales were examined using Spearman rank-order correlation coefficients (ρ; Table 4) and are summarized visually in Figure 5 (CD-VP: lower triangle; GPT-VP: upper triangle). Correlation strength was interpreted using conventional thresholds (ρ≈0.10="weak," ρ≈0.30="moderate," and ρ≥0.50="strong").

In the CD-VP scenario, realism and validity showed strong positive correlations with self-reported learning and reflection (ρ=0.635; *P*<.001) and with engagement and satisfaction (ρ=0.622; *P*<.001). Self-reported learning and reflection was also strongly correlated with engagement and satisfaction (ρ=0.683; *P*<.001). Cognitive load showed a weak negative association with self-reported learning and reflection (ρ=−0.253; *P*=.04) and no significant association with engagement and satisfaction (ρ=−0.094; *P*=.46).

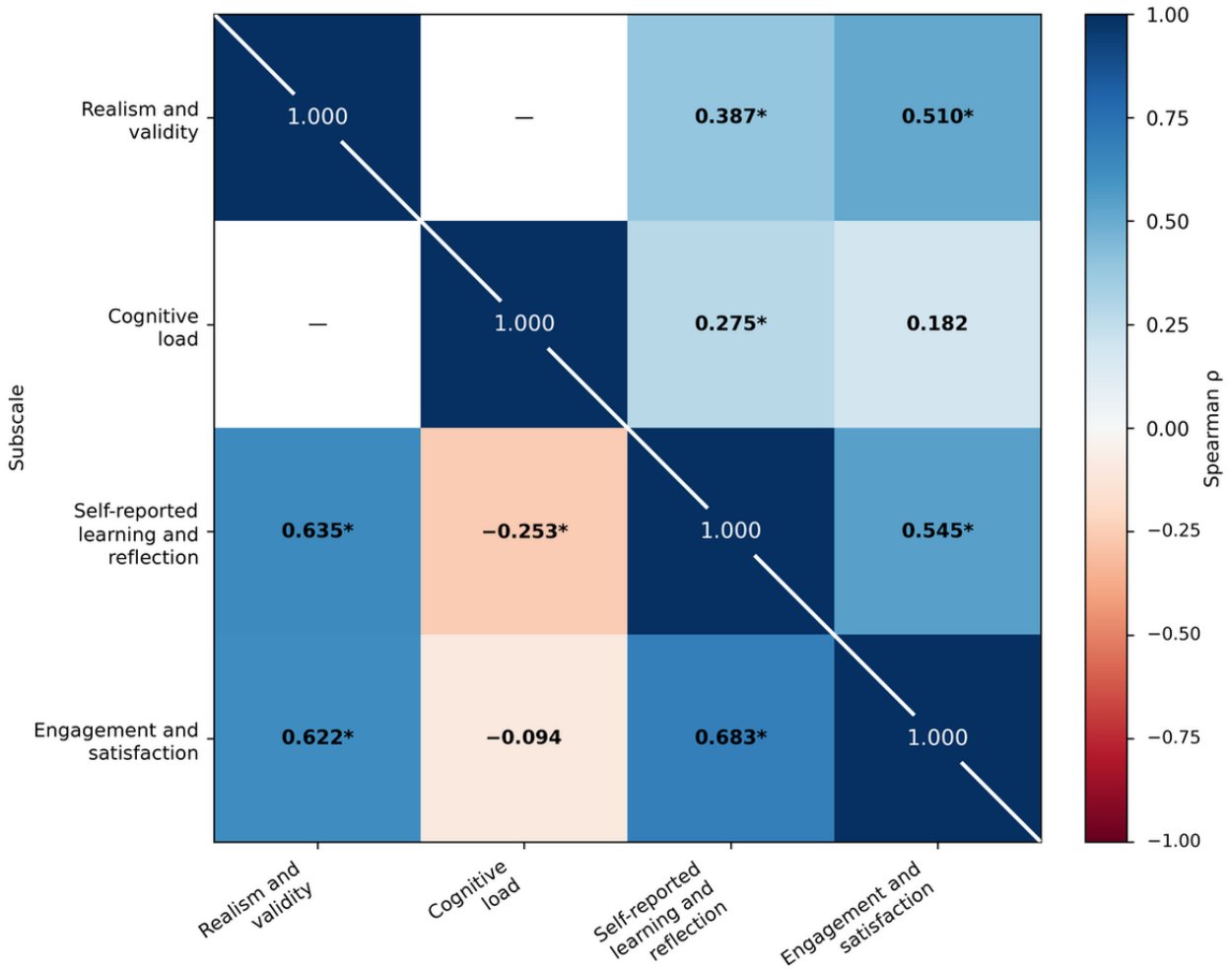
In the GPT-VP scenario, realism and validity correlated moderately with self-reported learning and reflection (ρ=0.387; *P*=.001) and strongly with engagement and satisfaction (ρ=0.510; *P*<.001). Self-reported learning and reflection was strongly correlated with engagement and satisfaction (ρ=0.545; *P*<.001). Cognitive load showed a weak positive association with self-reported learning and reflection (ρ=0.275; *P*=.02) and no significant association with engagement and satisfaction (ρ=0.182; *P*=.14).

Table 4. Spearman rank correlations between questionnaire subscales for the clinician-developed virtual patient (CD-VP) and ChatGPT-generated virtual patient (GPT-VP) scenarios.

Subscale pairs	CD-VP, Spearman ρ	*P* value	GPT-VP, Spearman ρ	*P* value
Realism and validity–self-reported learning and reflection	0.635	<.001	0.387	.001
Realism and validity–engagement and satisfaction	0.622	<.001	0.510	<.001
Self-reported learning and reflection–engagement and satisfaction	0.683	<.001	0.545	<.001
Cognitive load–self-reported learning and reflection	−0.253	.04	0.275	.02
Cognitive load–engagement and satisfaction	−0.094	.46	0.182	.14

XSL•FO
RenderX

Figure 5. Within-scenario correlations between questionnaire subscales (Spearman ρ). The lower triangle shows correlations for the clinician-developed virtual patient (CD-VP) and the upper triangle shows correlations for the ChatGPT-generated virtual patient (GPT-VP). Cell values are Spearman ρ, and color indicates direction and magnitude of the correlation (−1 to +1). "—" indicates correlations not tested or not reported in Table 4. *P<.05.



Discussion

Principal Findings

In this randomized, participant-blinded comparison of 1 CD-VP and 1 expert-reviewed GPT-VP, both scenarios were rated positively by medical students. The questionnaire demonstrated acceptable to good internal consistency, supporting the reliability of the 4 conceptual subscales under these exploratory conditions. Overall, the scenarios were broadly comparable, although the GPT-VP received more favorable ratings in selected domains, including higher realism and validity, lower cognitive load, and marginally higher ratings for self-reported learning and reflection. A graphical abstract summarizing the study is provided in Multimedia Appendix 4.

One plausible interpretation is not that AI-generated content is inherently superior, but that the AI-assisted workflow—standardized prompting followed by expert review—may have produced a more streamlined and internally consistent scenario. This may help explain why the GPT-VP was perceived as more realistic while also imposing less cognitive burden. It is also possible that the GPT-VP benefited from more textbook-aligned phrasing, which may have been more familiar to students and therefore perceived as clearer or more realistic than the more idiomatic clinical language sometimes used by experienced clinicians [22,23]. However, this interpretation should be considered cautiously, as students' prior clinical exposure may also have influenced these perceptions [24]. By contrast, the similarly high engagement and satisfaction ratings in both scenarios indicate that students found the VP format educationally valuable regardless of authorship.

Comparisons against the neutral midpoint reinforced this overall pattern, with both scenarios rated positively, although cognitive load differed from the midpoint only in the CD-VP, indicating that this case was experienced as more demanding. The within-scenario correlation patterns provide additional insight. In the CD-VP, higher cognitive load was weakly associated with lower perceived learning, whereas in the GPT-VP it showed a weak positive association with learning. This pattern indicates that cognitive demand may have been experienced differently across the 2 cases—more as a burden in the CD-VP and more as a manageable or productive challenge in the GPT-VP.

XSL•FO
RenderX

Comparison With Existing Literature

The present findings are broadly consistent with prior VP literature, which supports VPs as useful tools for engagement and applied learning in health professions education. In the Digital Health Education Collaboration systematic review and meta-analysis, Kononowicz et al [5] concluded that VPs were at least as effective as traditional education for knowledge outcomes and potentially more effective for skills outcomes, although the overall evidence base remained heterogeneous and of low-to-moderate certainty. Earlier European Commission-funded initiatives, including electronic Virtual Patients (eViP) and mEducator, systematically explored VP development, reuse, sharing, and implementation [25-27]. This work also highlighted the workload associated with authoring and adapting VP scenarios, which can limit adoption by individual clinical teachers and institutions.

In this context, generative AI represents a potentially useful approach. Most empirical work on generative AI in medical education has focused on more static educational outputs than full interactive VPs. In a randomized controlled experiment involving 74 medical students, Coşkun et al [16] found that ChatGPT-generated clinical vignettes did not differ significantly from human-written cases across 10 learner-rated statements, although the associated MCQ set was more variable, with only 6 of 15 items showing acceptable point-biserial discrimination. A similar pattern appears in the assessment literature. In the multinational prospective study by Cheung et al [28], 50 AI-generated and 50 human-authored graduate medical MCQs were rated by 5 independent international assessors. AI-generated questions were produced much faster (20 minutes 25 seconds vs 211 minutes 33 seconds), and overall quality was broadly comparable, although AI items scored significantly lower on relevance (mean 7.56, SD 0.94 vs mean 7.88, SD 0.52; P=.04) and showed wider variability. These studies suggest that generative AI can produce educationally usable material, but consistency and educational fit may still depend on review and refinement.

A particularly close comparator to the present work is the recent study by Szydlak et al [29] on ChatGPT-generated vs human-authored concept maps for clinical reasoning training with VPs. Using 10 diverse VP cases, the authors found that ChatGPT-generated maps contained significantly more concepts and connections (P<.001), while clinical content quality remained comparable to that of expert-generated maps. However, the AI maps were judged to lack abstraction, prioritization, and contextual alignment, and in some cases were considered to exceed the optimal cognitive load for learners.

The cognitive load pattern observed in the present study is also compatible with previous VP research showing that learner experience depends strongly on how a case is structured and embedded instructionally. In a study of 171 of 183 students, Marei et al [30] found that collaborative VP use after a lecture produced the highest instructional efficiency and the lowest intrinsic and extraneous cognitive load compared with independent VP formats. Fuoad et al [31] similarly reported, in 2 cohorts of 69 and 70 students, that collaborative VP use was associated with higher situational interest and lower extraneous cognitive load than lecture plus independent VP use.

Finally, adjacent simulation research in emergency and cardiovascular education supports the plausibility that well-designed digital cases can produce meaningful gains when evaluated objectively [32]. In a randomized controlled trial of VP simulation in undergraduate emergency medicine, Law et al [33] recruited 138 final-year students and found that the 4-week exposure group improved more than the 2-week group, with a mean score difference of 18 of 100 (95% CI 9.8-26.3). In immersive ACS virtual reality training, Einloft et al [34] studied 247 medical students and found significant gains in knowledge and clinical quality indicators (CQIs), with composite CQI scores improving from 0.47 before training to 0.76 to 0.80 at follow-up.

Strengths and Limitations

The study has several strengths. It was conducted within routine teaching, enhancing ecological validity [35]. In addition, both scenarios were delivered under the same general conditions and within a shared instructional blueprint, strengthening comparability between study arms. Students were unaware of scenario authorship, including the use of AI in 1 study arm, reducing expectancy bias. The study included a sample size comparable to or exceeding that of most single-center medical education randomized studies with complete response capture across all participants, eliminating missing data concerns.

Several limitations should nonetheless be acknowledged. First, the outcomes were based on student perceptions rather than objective measures of knowledge, clinical reasoning, retention, or performance. Second, although both scenarios shared the same platform, ACS-focused blueprint, delivery conditions, and number of decision points, the study compared only 1 CD-VP and 1 GPT-VP within a single clinical domain and educational context. Observed differences may therefore reflect case-specific features, including clinical difficulty, branching complexity, information volume, wording, feedback style, or visual material, rather than the development pathway alone [36]. This is particularly relevant for cognitive load: lower perceived burden in the GPT-VP may indicate clearer organization and reduced extraneous load, but it may also reflect lower complexity or a more familiar case structure. Third, although the questionnaire showed acceptable to good internal consistency for most subscales, it was developed for this study to capture learner perceptions and was not fully validated as a measure of overall scenario quality. Fourth, the study did not include independent structured educator ratings of final scenario quality and clinical accuracy after implementation. Fifth, the findings reflect expert-supervised AI-assisted VP development rather than unsupervised AI generation, because the GPT-VP was refined through expert review and iterative prompting. Sixth, although approximate development calendar time was retrospectively estimated, the study did not prospectively record person-hours, review and revision burden, prompting duration, or detailed platform analytics. Finally, the full chronological AI prompt history was not exported or retained as a verbatim transcript, limiting exact reproducibility of the development workflow. Future studies should address these limitations

XSL•FO
RenderX

through multiscenario comparisons, objective learning outcomes, log-based analytics, and formal evaluation of development workload.

Future Directions

Future studies should explore whether AI-assisted VP development translates into measurable educational benefits by assessing outcomes such as postintervention test performance, clinical reasoning in novel cases, and longer-term knowledge retention. Future work should also compare the practical efficiency of clinician-only and AI-assisted VP workflows, including revision burden and update cycles.

It will also be important to determine which VP design features are most effective, including branching density, information volume, feedback structure, and media use. Given the potential language barrier, future studies should evaluate whether the present workflow performs similarly across languages, as multilingual medical LLM performance has been shown to vary according to both input language and prompt language [37].

Finally, future research should examine whether, and how quickly, AI-assisted VP authoring can support the incorporation of evolving evidence from big data, digital health, personalized medicine, and precision medicine while maintaining appropriate educational value.

Conclusions

In this randomized, participant-blinded, controlled study of 2 mobile cardiology VP scenarios, both the CD-VP and GPT-VP cases were rated positively by fifth-year medical students. The AI-generated, expert-reviewed scenario received similarly favorable or more favorable learner-perceived ratings in selected domains. These findings provide initial controlled evidence supporting the acceptability of AI-assisted VP development within routine undergraduate medical education. Future research should investigate objective learning outcomes, longer-term retention measures, development workload data, and platform-based analytics.

Acknowledgments

The authors thank the fifth-year medical students who participated in this study and the faculty members who supported its implementation during the emergency medicine course. The authors also acknowledge the use of ChatGPT-5 (OpenAI; GPT-5, standard and Thinking modes, accessed via ChatGPT Plus in September 2025) for developing the GPT-VP scenario. No generative AI tools were used for language editing, figure generation, or plot development. All AI-generated outputs were reviewed, revised through iterative follow-up prompting, and approved by the authors, who take full responsibility for the final content.

Funding

No external funding was received for the present study.

Data Availability

Deidentified data may be made available by the corresponding author on reasonable request, subject to institutional and ethical requirements.

Generative AI Use

The authors declare that generative AI was not used in the preparation of this manuscript.

Authors' Contributions

Conceptualization: MGB, ED, PB
Data curation: SG, AN
Formal analysis: SG
Investigation: MGB, BF, EK, AN
Methodology: MGB, ED, SG
Project administration: MGB
Resources: BF, EK
Supervision: PB
Validation: PB
Visualization: MGB, SG
Writing—original draft: MGB, ED
Writing—review and editing: MGB, ED, BF, EK, SG, AN, PB

Conflicts of Interest

None declared.

XSL•FO
RenderX

Multimedia Appendix 1
Supplementary reporting material, AI prompting and expert revision workflow, and postscenario questionnaire.
[DOCX File , 38 KB - mededu_v12i1e100839_app1.docx]

Multimedia Appendix 2
CONSORT-eHEALTH checklist (V 1.6.1).
[PDF File (Adobe PDF File), 17 KB - mededu_v12i1e100839_app2.pdf]

Multimedia Appendix 3
Full virtual patient scenario materials, including the full clinician-developed and AI-generated and expert-reviewed scenarios.
[DOCX File , 9738 KB - mededu_v12i1e100839_app3.docx]

Multimedia Appendix 4
Graphical abstract.
[DOCX File , 918 KB - mededu_v12i1e100839_app4.docx]

References

1. Chéron M, Ademi M, Kraft F, Löffler-Stastka H. Case-based learning and multiple choice questioning methods favored by students. BMC Med Educ 2016 Feb 02;16:41 [FREE Full text] [doi: 10.1186/s12909-016-0564-x] [Medline: 26830675]
2. Maia D, Andrade R, Afonso J, Costa P, Valente C, Espregueira-Mendes J. Academic performance and perceptions of undergraduate medical students in case-based learning compared to other teaching strategies: a systematic review with meta-analysis. Educ Sci 2023 Feb 23;13(3):238. [doi: 10.3390/educsci13030238]
3. Donkin R, Yule H, Fyfe T. Online case-based learning in medical education: a scoping review. BMC Med Educ 2023 Aug 09;23(1):564 [FREE Full text] [doi: 10.1186/s12909-023-04520-w] [Medline: 37559108]
4. Ellaway R, Poulton T, Fors U, McGee JB, Albright S. Building a virtual patient commons. Med Teach 2008;30(2):170-174. [doi: 10.1080/01421590701874074] [Medline: 18464142]
5. Kononowicz AA, Woodham LA, Edelbring S, Stathakarou N, Davies D, Saxena N, et al. Virtual patient simulations in health professions education: systematic review and meta-analysis by the digital health education collaboration. J Med Internet Res 2019 Jul 02;21(7):e14676 [FREE Full text] [doi: 10.2196/14676] [Medline: 31267981]
6. Plackett R, Kassianos AP, Mylan S, Kambouri M, Raine R, Sheringham J. The effectiveness of using virtual patient educational tools to improve medical students' clinical reasoning skills: a systematic review. BMC Med Educ 2022 May 13;22(1):365 [FREE Full text] [doi: 10.1186/s12909-022-03410-x] [Medline: 35550085]
7. Dafli E, Fountoukidis I, Hatzisevastou-Loukidou C, Bamidis PD. Curricular integration of virtual patients: a unifying perspective of medical teachers and students. BMC Med Educ 2019 Nov 09;19(1):416 [FREE Full text] [doi: 10.1186/s12909-019-1849-7] [Medline: 31706296]
8. Fatima A, Shafique MA, Alam K, Fadlalla Ahmed TK, Mustafa MS. ChatGPT in medicine: a cross-disciplinary systematic review of ChatGPT's (artificial intelligence) role in research, clinical practice, education, and patient interaction. Medicine (Baltimore) 2024 Aug 09;103(32):e39250 [FREE Full text] [doi: 10.1097/MD.0000000000039250] [Medline: 39121303]
9. Xu X, Chen Y, Miao J. Opportunities, challenges, and future directions of large language models, including ChatGPT in medical education: a systematic scoping review. J Educ Eval Health Prof 2024;21:6 [FREE Full text] [doi: 10.3352/jeehp.2024.21.6] [Medline: 38486402]
10. Farquhar S, Kossen J, Kuhn L, Gal Y. Detecting hallucinations in large language models using semantic entropy. Nature 2024 Jun;630(8017):625-630 [FREE Full text] [doi: 10.1038/s41586-024-07421-0] [Medline: 38898292]
11. The Lancet Digital Health. ChatGPT: friend or foe? Lancet Digit Health 2023 Mar;5(3):e102 [FREE Full text] [doi: 10.1016/S2589-7500(23)00023-7] [Medline: 36754723]
12. Kung TH, Cheatham M, Medenilla A, Sillos C, De Leon L, Elepaño C, et al. Performance of ChatGPT on USMLE: potential for AI-assisted medical education using large language models. PLOS Digit Health 2023 Feb 09;2(2):e0000198 [FREE Full text] [doi: 10.1371/journal.pdig.0000198] [Medline: 36812645]
13. Bicknell BT, Butler D, Whalen S, Ricks J, Dixon CJ, Clark AB, et al. ChatGPT-4 Omni performance in USMLE disciplines and clinical skills: comparative analysis. JMIR Med Educ 2024 Nov 06;10:e63430 [FREE Full text] [doi: 10.2196/63430] [Medline: 39504445]
14. Liu M, Okuhara T, Chang X, Shirabe R, Nishiie Y, Okada H, et al. Performance of ChatGPT across different versions in medical licensing examinations worldwide: systematic review and meta-analysis. J Med Internet Res 2024 Jul 25;26:e60807 [FREE Full text] [doi: 10.2196/60807] [Medline: 39052324]
15. Law AK, So J, Lui CT, Choi YF, Cheung KH, Kei-ching Hung K, et al. AI versus human-generated multiple-choice questions for medical education: a cohort study in a high-stakes examination. BMC Med Educ 2025 Feb 08;25(1):208 [FREE Full text] [doi: 10.1186/s12909-025-06796-6] [Medline: 39923067]

XSL•FO
RenderX

16. Coşkun Ö, Kıyak YS, Budakoğlu I. ChatGPT to generate clinical vignettes for teaching and multiple-choice questions for assessment: a randomized controlled experiment. Med Teach 2025 Feb;47(2):268-274. [doi: 10.1080/0142159X.2024.2327477] [Medline: 38478902]
17. Hopewell S, Chan AW, Collins GS, Hróbjartsson A, Moher D, Schulz KF, et al. CONSORT 2025 statement: updated guideline for reporting randomised trials. Lancet 2025 May 03;405(10489):1633-1640. [doi: 10.1016/S0140-6736(25)00672-5] [Medline: 40245901]
18. Eysenbach G, CONSORT-EHEALTH Group. CONSORT-EHEALTH: improving and standardizing evaluation reports of web-based and mobile health interventions. J Med Internet Res 2011 Dec 31;13(4):e126 [FREE Full text] [doi: 10.2196/jmir.1923] [Medline: 22209829]
19. Byrne RA, Rossello X, Coughlan JJ, Barbato E, Berry C, Chieffo A, et al. 2023 ESC guidelines for the management of acute coronary syndromes. Eur Heart J 2023 Oct 12;44(38):3720-3826 [FREE Full text] [doi: 10.1093/eurheartj/ehad191] [Medline: 37622654]
20. MobiViP. URL: https://mvp.med.auth.gr/ [accessed 2026-09-25]
21. Dafli E, Nikolaidou M, Rodina-Theocharaki A, Antoniou P, Bamidis PD. Simulation medicine and scenario-based learning for emergency care: A guide for educators and trainers. Erasmus Pus. URL: https://erasmusplus.org.ua/en/projects/simulation-medicine-and-scenario-based-learning-for-emergency-care/ [accessed 2026-09-25]
22. Coombs RH, Chopra S, Schenk DR, Yutan E. Medical slang and its functions. Soc Sci Med 1993 Apr;36(8):987-998. [doi: 10.1016/0277-9536(93)90116-l] [Medline: 8475427]
23. Bennett N, Lockyer J, Mann K, Batty H, LaForet K, Rethans JJ, et al. Hidden curriculum in continuing medical education. J Contin Educ Health Prof 2004;24(3):145-152. [doi: 10.1002/chp.1340240305] [Medline: 15490546]
24. Haider N, Morjaria L, Sheth U, Al-Jabouri N, Sibbald M. MD student perceptions of ChatGPT for reflective writing feedback in undergraduate medical education. Int Med Educ 2025 Jul 23;4(3):27. [doi: 10.3390/ime4030027]
25. Poulton T, Balasubramaniam C. Virtual patients: a year of change. Med Teach 2011;33(11):933-937. [doi: 10.3109/0142159X.2011.613501] [Medline: 22022903]
26. Dafli E, Bamidis PD, Pappas C, Dombros N. A case study of repurposing medical educational material to three different content types. Bio Algorithms Med Syst 2010;6(11):99 [FREE Full text]
27. Antoniades A, Nicolaidou I, Spachos D, Mylläri J, Giordano D, Dafli E, et al. Medical content searching, retrieving, and sharing over the internet: lessons learned from the mEducator through a scenario-based evaluation. J Med Internet Res 2015 Oct 09;17(10):e229 [FREE Full text] [doi: 10.2196/jmir.3650] [Medline: 26453250]
28. Cheung BH, Lau GK, Wong GT, Lee EY, Kulkarni D, Seow CS, et al. ChatGPT versus human in generating medical graduate exam multiple choice questions—a multinational prospective study (Hong Kong S.A.R., Singapore, Ireland, and the United Kingdom). PLoS One 2023 Aug 29;18(8):e0290691 [FREE Full text] [doi: 10.1371/journal.pone.0290691] [Medline: 37643186]
29. Szydlak R, Kiyak YS, Hege I, Górski S, Linglart L, Shchudrova T, et al. ChatGPT versus human authors: a comparative study of concept maps for clinical reasoning training with virtual patients. Med Teach 2026 Apr;48(4):712-720 [FREE Full text] [doi: 10.1080/0142159X.2025.2583403] [Medline: 41223082]
30. Marei HF, Donkers J, Al-Eraky MM, Van Merrienboer JJ. Collaborative use of virtual patients after a lecture enhances learning with minimal investment of cognitive load. Med Teach 2019 Mar;41(3):332-339. [doi: 10.1080/0142159X.2018.1472372] [Medline: 29798713]
31. Fuoad SA, El-Sayed W, Marei H. Effect of different teaching/learning approaches using virtual patients on student's situational interest and cognitive load: a comparative study. BMC Med Educ 2022 Nov 07;22(1):763 [FREE Full text] [doi: 10.1186/s12909-022-03831-8] [Medline: 36344961]
32. Antoniou P, Dafli E, Giannakoulas G, Igimbayeva G, Visternichan O, Kyselov S, et al. Education of patients with atrial fibrillation and evaluation of the efficacy of a mobile virtual patient environment: protocol for a multicenter pseudorandomized controlled trial. JMIR Res Protoc 2024 Jan 23;13:e45946 [FREE Full text] [doi: 10.2196/45946] [Medline: 38261376]
33. Law AK, Hung KK, James KS, Kelly AM, Graham CA. Virtual patient simulation in undergraduate emergency medicine education during COVID-19: randomized controlled trial. Hong Kong J Emerg Med 2023 Jul 22;30(6):439-446. [doi: 10.1177/10249079231189376]
34. Einloft J, Russ P, Bedenbender S, Meyer HL, Morgenschweis ML, Ganser A, et al. Effectiveness of fully immersive virtual reality-based simulation training on objective knowledge acquisition in acute coronary syndrome/ST-elevation myocardial infarction emergency management: a pre-post-intervention study. Eur Heart J Digit Health 2025 Sep 4;7(1):ztaf094 [FREE Full text] [doi: 10.1093/ehjdh/ztaf094] [Medline: 41574044]
35. Rothwell PM. Factors that can affect the external validity of randomised controlled trials. PLoS Clin Trials 2006 May;1(1):e9 [FREE Full text] [doi: 10.1371/journal.pctr.0010009] [Medline: 16871331]
36. Shah K, Xu AY, Sharma Y, Daher M, McDonald C, Diebo BG, et al. Large language model prompting techniques for advancement in clinical medicine. J Clin Med 2024 Aug 28;13(17):5101 [FREE Full text] [doi: 10.3390/jcm13175101] [Medline: 39274316]

XSL•FO
RenderX

37. Strasser LM, Anschuetz W, Dennstädt F, Hastings J. Performance evaluation of large language models in multilingual medical multiple-choice questions: mixed methods study. JMIR Med Educ 2026 Mar 05;12:e81399 [FREE Full text] [doi: 10.2196/81399] [Medline: 41813244]

Abbreviations

ACS: acute coronary syndrome
CBL: case-based learning
CD-VP: clinician-developed virtual patient
CLES: common language effect size
CONSORT: Consolidated Standards of Reporting Trials
CONSORT-EHEALTH: Consolidated Standards of Reporting Trials of Electronic and Mobile Health Applications and Online Telehealth
CQI: clinical quality indicator
eViP: electronic Virtual Patients
GDPR: General Data Protection Regulation
GPT-VP: ChatGPT-generated virtual patient
LLM: large language model
MCQ: multiple-choice question
MI: myocardial infarction
MobiViP: mobile virtual patients app
USMLE: United States Medical Licensing Examination
VP: virtual patient

Edited by S Guinez-Molinos; submitted 30.May.2026; peer-reviewed by M Sibbald, Y Fan; comments to author 02.Jul.2026; revised version received 29.Jul.2026; accepted 03.Aug.2026; published 28.Oct.2026.

Please cite as:
Bantidos MG, Dafli E, Karagiannidis E, Fyntanidou B, Gilou S, Nasoufidou A, Bamidis P
Perceived Ratings of AI-Generated and Expert-Reviewed vs Clinician-Developed Virtual Cardiology Patient Scenarios Among Medical Students: Randomized Controlled Study
JMIR Med Educ 2026;12:e100839
URL: https://mededu.jmir.org/2026/1/e100839
doi:10.2196/100839
PMID:

©Marios Georgios Bantidos, Eleni Dafli, Efstratios Karagiannidis, Barbara Fyntanidou, Sotiria Gilou, Athina Nasoufidou, Panagiotis Bamidis. Originally published in JMIR Medical Education (https://mededu.jmir.org), 28.Oct.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Education, is properly cited. The complete bibliographic information, a link to the original publication on https://mededu.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Association of Preexamination AI Virtual-Patient Trainer Scores With Scores on a Local Single-Station History-Taking Objective Structured Clinical Examination: Observational Cohort Study

Weishan Zhang[1*], MSc; Yang Liu[1*], MM; Yiying Zhu[1], MSc; Xian Lu[2], BSc; Liping Wu[1], Prof Dr; Minghui Yue[1], MPH; Xiaolin Hong[1], BMed; Oudong Xia[3*], Prof Dr, MD; Chujun Shi[1*], Prof Dr

[1]Medical Simulation Center, Shantou University Medical College, Shantou, Guangdong, China

[2]Department of Medical Physics and Informatics, Shantou University Medical College, Shantou, China

[3]Shantou University Medical College, Shantou, China

[*]these authors contributed equally

Corresponding Author:
Chujun Shi, Prof Dr
Medical Simulation Center
Shantou University Medical College
Number 22 Xinling Road
Shantou, Guangdong, 515041
China
Phone: 86 754 88900459
Email: cjshi@stu.edu.cn

Abstract

Background: Large language model–based virtual-patient systems score history-taking practice, but whether scores correspond to independently scored examination performance remains untested. Existing evaluations emphasize internal platform metrics or simulated-patient fidelity, leaving criterion-related evidence limited.

Objective: This study tests whether mean preexamination AI-trainer practice scores were associated with performance on 1 local objective structured clinical examination (OSCE) history-taking station, whether the association depended on the complaint-matched training case, and whether a model including the trainer score retained predictive performance in held-out students.

Methods: At 1 medical school, second-year students received a required 6-case assignment on the AI-Powered Medical History-Taking Training and Evaluation System (AMTES) and completed an end-of-course OSCE. Each student completed 1 station rated by 1 examiner; interrater and cross-station reliability could not be estimated. The 4 OSCE diagnoses differed from the training diagnoses: 3 shared a presenting-complaint category with an AMTES case, whereas 1 OSCE case had no presenting-complaint match. Platform-to-OSCE record linkage was 474 out of 480 (98.8%); 473 students with analyzable preexamination practice formed the analytic sample. The primary predictor was the mean of normalized scores across practice records time-stamped before each student's examination-day cut-off. Analyses used Spearman correlations, OSCE-case fixed-effects regression, and repeated 5-fold cross-validation.

Results: OSCE scores (mean 81.5, SD 6.7) were associated with platform practice scores (Spearman ρ=0.388; P<.001). With OSCE-case fixed effects, each 1SD increase in the platform score corresponded to a 2.83-point increase in the OSCE score (95% CI 2.26-3.39). The association was observed in the AMTES-unmatched low back/flank pain stratum (Spearman ρ=0.370; n=108) and with a leave-matched-case-out score (Spearman ρ=0.398 vs 0.370 for the full mean among the same 339 students). Under a single cohort-wide cut-off at the first examination day, the association attenuated but persisted among the same 469 students (Spearman ρ decreased from 0.385 to 0.346; paired difference −0.040, 95% CI −0.065 to −0.018; regression coefficient decreased from 2.83 to 2.68 points per SD). In repeated 5-fold cross-validation (n=471), the platform-only model achieved Spearman ρ=0.374 and R^2=0.176. Adding the platform score to the 2 prior physical-examination course scores increased cross-validated R^2 from 0.141 to 0.236 (ΔR^2=0.096; 95% CI 0.053-0.142). Adding practice volume and case breadth did not improve performance (ΔR^2=−0.002).

XSL•FO
RenderX

Conclusions: By linking required curricular practice scores to the score on 1 independently scored local OSCE history-taking station, this observational study provides criterion-related evidence for interpreting AMTES scores. The association did not depend on including the complaint-matched training-case score. The scores also added predictive information beyond the 2 prior physical-examination course scores in internal cross-validation. These findings establish neither causal effects of practice nor within-student transfer, and residual confounding by prior ability cannot be excluded. They support further evaluation of AMTES analytics as a candidate low-stakes formative signal, with independent external validation required before operational use.

(JMIR Med Educ 2026;12:e109752) doi:10.2196/109752

KEYWORDS

artificial intelligence; large language models; virtual patients; medical history taking; objective structured clinical examination; criterion-related validity; educational measurement; learning analytics; medical education

Introduction

Eliciting an accurate clinical history is a foundational diagnostic skill, yet providing sufficient interactive practice and expert feedback at scale is resource-intensive. Large language model (LLM)–based virtual patients offer interactive and repeatable opportunities to practice history-taking [1,2]. More recent systems can also provide immediate, structured automated feedback [2] and rubric-based scoring with evidence-linked feedback [3], potentially expanding access to practice and feedback. These systems build on earlier computer-based virtual-patient research [4,5] and the broader technology-enhanced simulation literature [6]. Early LLM-based studies examined simulated-patient plausibility and usability [1], interrater agreement between automated and human evaluations of history-content coverage [2], and the stability and transparency of automated rubric-based scoring [3]. A recent scoping review similarly found that evaluations largely emphasized user experience, authenticity, and system performance, whereas objective learning outcomes were assessed infrequently [7]. More broadly, evidence that performance demonstrated in virtual-patient environments translates to independently assessed clinical performance remains limited and heterogeneous [5,7].

A platform may generate consistent practice scores, but such consistency provides evidence of reliability and does not, by itself, establish validity [8,9]. Contemporary validity theory defines validity in terms of the extent to which evidence and theory support the proposed interpretations and uses of scores, rather than as an inherent property of an assessment instrument [10-12]. Interpreting automated practice scores as indicators of history-taking performance therefore requires evidence that these scores relate as expected to a relevant external measure [8,12]. Argument-based approaches to validation provide practical guidance for organizing and evaluating such evidence in health professions education [11,13]. An independently administered objective structured clinical examination (OSCE) is a relevant external criterion for a history-taking trainer because OSCEs are widely used as standardized assessments of clinical performance [14-18]. In this study, the criterion is the score recorded at a single, platform-independent, assessor-blinded local OSCE history-taking station (platform independence is defined in the “OSCE Outcome” section). The study does not constitute multiinstitutional external validation.

Learning analytics are increasingly proposed for describing and predicting learner performance and supporting formative action, but their interpretation and use require appropriate validation [19,20]. Automated scoring also has precedents in medical education, including the automated scoring of constructed responses [21] and research on natural language processing–assisted scoring of patient notes in the former US Medical Licensing Examination Step 2 Clinical Skills examination [22]. Prior LLM-based virtual-patient research has also examined learner perceptions and feasibility [23], while early intervention studies have evaluated educational outcomes in OSCE preparation [24,25]. These studies addressed system performance, learner experience, rater agreement, or intervention effectiveness rather than whether automated scores accumulated during routine practice correspond to performance on a separately administered clinical assessment.

Previous studies of the AI-Powered Medical History-Taking Training and Evaluation System (AMTES) examined automated scoring stability and human-AI consistency [3], associations with summative educational outcomes under voluntary use [26], and learner acceptance [27]. However, they did not determine whether AMTES scores generated during routine practice correspond to performance on an independently administered clinical assessment. This study therefore addresses a distinct, noncausal measurement question in the context of required curricular use. To our knowledge, cohort-scale evidence linking automated trainer scores derived from practice sessions submitted before an examination to independently assessed OSCE performance remains scarce.

Whether this association holds matters for potential deployment. If AMTES performance is associated with the score recorded on a separately administered, local, single-station OSCE, platform analytics may provide a scalable, low-stakes formative signal to guide faculty review and targeted educational support. If not, the automated scores may primarily reflect platform-specific performance. Even if an association is observed, it may reflect shared prior ability or conscientious engagement in both settings, or rehearsal matched to the examination content rather than a broader history-taking performance signal. To address these alternative explanations, this study used a local OSCE aligned with established curriculum-blueprinting principles [16-18]. The OSCE was administered independently of AMTES, and examiners had no access to students’ platform records.

Accordingly, this study addressed 3 research questions (RQs):

XSL•FO
RenderX

RQ1 (criterion-related evidence): Are students' mean AMTES practice scores associated with the score recorded at a local, single-station, independently scored OSCE history-taking station?

RQ2 (dependence on the complaint-matched AMTES case): Does the association persist in the AMTES-unmatched presenting-complaint stratum and when the score contribution from the presenting-complaint-matched AMTES case is omitted from the platform score?

RQ3 (internal cross-validated prediction): How well does a model predict held-out students' OSCE scores using prior examination scores and AMTES measures derived from time stamp–restricted practice records, without assigned-case information?

Methods

Study Design

This observational study drew on the complete eligible curricular cohort within a course that included 2 required components: a 6-case AMTES assignment and an end-of-course OSCE. Practice records retained for construction of the primary predictor were time-stamped between May 10 and June 26, 2026. The OSCE was administered over 5 days (June 22-26, 2026), with different classes examined on different days. As the practice and examination periods overlapped at the cohort level, only practices time-stamped before each student's own OSCE cut-off, as defined below, were included in the primary predictor. Neither AMTES exposure nor practice dose was randomized. Study size was determined by the eligible curricular cohort rather than by an a priori power calculation. Reporting followed the STROBE (Strengthening the Reporting of Observational Studies in Epidemiology) guidelines [28]. Supplementary numeric results are provided in Multimedia Appendix 1, and extended methods and results are provided in Multimedia Appendix 2.

Setting, Participants, and the AI Trainer

Eligibility for the analytic cohort was defined by the curricular and assessment arrangements established before examination performance was observed. Eligible participants were second-year students in the 2024-entry Clinical Medicine program at Shantou University Medical College whose curriculum included both the 6-case AMTES assignment and the concurrently scheduled end-of-course OSCE. Some students taking the same history-taking course completed the AMTES assignment but were not scheduled to take the OSCE under their curricular programs' assessment arrangements; they were therefore ineligible for this analysis. One repeat-year student who returned only to resit the OSCE without participating in the corresponding AMTES course component was also ineligible. Eligibility was not determined by actual AMTES engagement, successful record linkage, or OSCE performance. Accordingly, whether a student had a retained scored record for each of the 6 assigned cases was treated as an exposure characteristic rather than as an eligibility or analytic-inclusion criterion.

AMTES, the LLM-based virtual-patient trainer described previously [3], was deployed as a required self-study in the students' first formal history-taking course. The assignment comprised 6 history-taking cases involving cough, fever, urinary frequency, chest pain, headache, and abdominal pain. Completing each assigned case once counted toward the course requirement; repeat practice was optional and earned no additional credit. After each practice session submitted for evaluation, the platform generated item-level rubric feedback, overall feedback, and an overall score. For the primary predictor, retained platform records were limited to evaluated practice records with valid scores that were time-stamped before the student-specific cut-off defined below.

AI Trainer Configuration During the Study Period

Virtual-patient dialogue was generated by DeepSeek-V4-flash (Hangzhou DeepSeek Artificial Intelligence Basic Technology Research Co Ltd). Practice scoring was fully automated and comprised item-level scoring by DeepSeek-V3.2 at temperature 0 using a shared scoring prompt template, followed by automated postprocessing of the item scores before the resulting score was returned to the student. The scores analyzed here are those received by students. The configured model identifiers and platform-level prompt templates were the same for all students and shared across the 6 training cases; each case supplied its own clinical content and scoring rubric. According to the platform configuration and deployment records, these model identifiers and templates remained unchanged during the study period, and AMTES version 4.0 was not redeployed. The AMTES workflow, multilevel verification procedures, and comprehensive history-taking scoring prompt are described in our previous system-development report and its multimedia appendices 1-3 [3]. The model configuration used for the present deployment is specified above.

OSCE Outcome

At the end of the course, students completed an OSCE history-taking station involving 1 of 4 presenting complaints: chest pain, cough/dyspnea, fever/rash, or low back/flank pain. Each student completed 1 station and received a single rating from the examiner assigned to that station; no performance was independently double-scored. Within each examination session, all 4 cases operated concurrently as parallel stations. Examination staff acting as guides directed waiting students to available stations without using student identity, class, prior performance, or preference to determine station assignment. No predetermined candidate-to-station sequence, formal random-allocation procedure, or assignment log was used. We therefore describe the process as availability-driven allocation by examination staff rather than as randomization.

Examiners were blinded specifically to AMTES data: they were neither provided with nor given access to individual students' AMTES practice records or platform-derived scores before or during scoring. As used in this paper, platform independence means that a different instrument was administered and scored outside the platform by assessors masked to platform data. It does not imply construct independence, because both measures target history-taking; statistical independence of the 2 scores; separation of personnel, as several authors developed or

XSL•FO
RenderX

deployed AMTES (conflicts of interest); or any claim about the reliability or generalizability of the OSCE score itself. The examination followed a structured OSCE format [14-16] and used standardized patients (SPs). Each local station consisted of a 20-minute, 1-on-1 history-taking consultation. Because of available examiner and SP staffing, assessment of the full eligible cohort was distributed across 5 days using parallel, case-specific stations. The local scoring rubric was developed with reference to China's National Medical Licensing Examination. It comprised case-specific history-content items and a 6D communication-skills and comprehensive-performance block (hereafter referred to as the *communication block*). The official recorded station score, used in all analyses, ranged from 0 to 100 points and comprised 70 points for history content and 30 points for the communication block. The total was calculated as the sum of the corresponding item-level scores. Before the examination, all SPs and examiners received common training intended to promote consistency in case portrayal and scoring. Such training and standardization are recognized quality-assurance safeguards for performance-based assessments [29].

Nine classes sat the examination, 1 per half-day session (4 morning and 5 afternoon) over 5 days; class and session therefore coincided, so the class-grouped cross-validation reported below is equivalently session-grouped. Sessions used 4-7 stations, depending on class size; when a case required more than 1 station in a session, each station was staffed by a different SP and scored by a different examiner. Thirteen SPs took part, each portraying 1 case (4 SPs portrayed the chest pain case, and 3 SPs portrayed each of the other cases) and participating in 4 or 5 of the 9 sessions. Twelve examiners scored the history-taking stations; they were assigned to station slots rather than to cases and therefore scored different cases across sessions. Examiner identity was not recorded with individual scores, so examiner-specific effects could not be estimated. The station deployment grid is provided in Table S7 in Multimedia Appendix 1, with further administration details provided in Multimedia Appendix 2.

The 4 OSCE cases involved diagnoses that were not represented among the 6 AMTES training cases: gastroesophageal reflux disease presenting with chest pain, pleural effusion presenting with cough/dyspnea, dengue fever presenting with fever/rash, and renal stone presenting with low back/flank pain. For the content-overlap analyses, 3 OSCE cases were matched to an AMTES case solely based on a prespecified presenting-complaint category: chest pain with the myocardial infarction training case, cough with the pneumonia training case, and fever with the liver abscess training case. The diagnoses differed within all 3 matched pairs. The renal-stone OSCE case, designated as low back/flank pain, had no corresponding case among the 6 assigned AMTES cases with the same prespecified presenting-complaint category and was therefore classified as the AMTES-unmatched presenting-complaint case. In this article, "AMTES unmatched" refers only to the absence of a presenting-complaint match among the assigned AMTES cases and does not imply that students had no prior exposure to or practice with that complaint in other settings. Within the matched pairs, the OSCE and AMTES case scripts also differed in features such as onset, precipitating factors, and illness evolution. Nevertheless, generic symptom-elicitation domains and complaint-level interviewing schemas could still overlap (Table S3 in Multimedia Appendix 1).

Linkage and Measures

OSCE records were linked to platform records using a crosswalk between student numbers and platform user IDs. Inclusion in the analytic sample required an OSCE history-taking score and at least one retained AMTES practice record time-stamped before the examination (Figure 1); this analytic requirement was distinct from curricular eligibility.

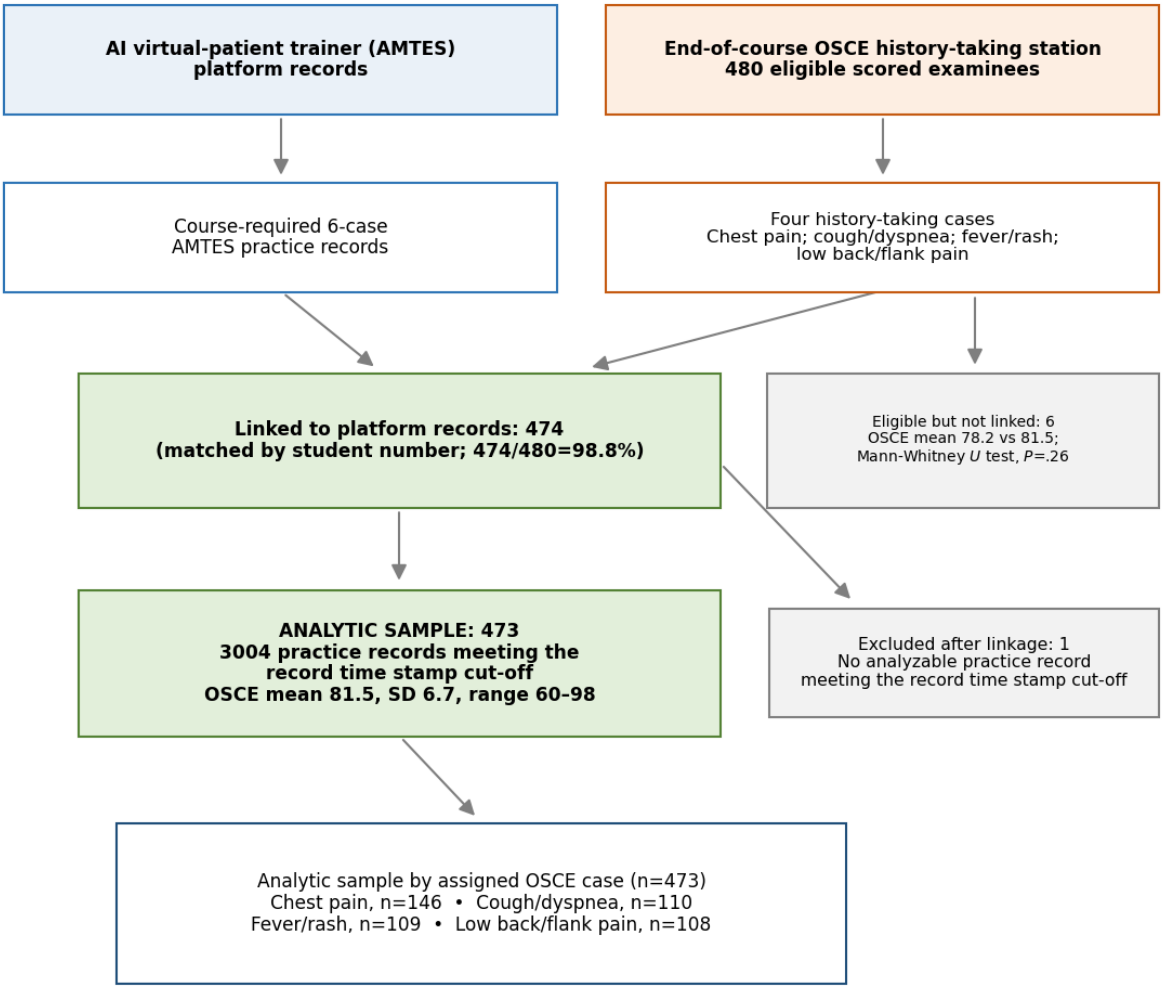
The primary predictor was each student's mean normalized score across retained practices, chosen to represent the student's average score across recorded practice encounters, with each encounter contributing equally. As this approach gives greater weight to cases practiced repeatedly, a case-balanced mean (calculated by averaging scores within each practiced case and then across practiced cases) was evaluated as a sensitivity analysis (see the "Results" section; also see Table S1 in Multimedia Appendix 1).

Each retained practice score was divided by the maximum score for its case, placing scores from cases with different rubric maxima on a common scale from 0 to 1. The case-specific maximum was 62 points for the cough case and 70 points for each of the other 5 AMTES cases. For descriptive reporting, cough scores were also expressed on an equivalent 70-point scale; this linear rescaling did not change their normalized values. Practice-record time stamps indicate when students submitted completed practices for evaluation. The normalized practice scores were then averaged using only records time-stamped before a conservative examination-day cut-off, defined as the earliest scheduled OSCE start time (8 AM) on each student's examination date. All analyzed practice scores are traceable to their original scoring records.

Secondary predictors included the leave-matched-case-out score, practice volume, and case breadth. Practice volume was defined as the number of retained practices, and case breadth as the number of distinct AMTES cases practiced. The leave-matched-case-out score was calculated by excluding practices from the AMTES case matched to the student's assigned OSCE case by presenting-complaint category and then averaging the remaining normalized practice scores. This score was defined only for students who had practiced both the complaint-matched AMTES case and at least one additional case. The measured baseline covariates were the 2 examinations from the preceding physical-examination course, which was a prerequisite for the history-taking course: its written theory examination (Clinical Basic Skills I; 0-100) and its practical OSCE examination (total score). Both examinations were taken in the preceding semester, before the practice window. Both assess physical examination rather than history-taking and are the only prior physical-examination course scores available for this cohort. They were used as baseline covariates in adjusted association models and as predictors in models without assigned OSCE-case information.

XSL•FO
RenderX

Figure 1. Participant flow and linkage of eligible objective structured clinical examination (OSCE) examinees to preexamination platform records (analytic sample, N=473). AMTES: AI-Powered Medical History-Taking Training and Evaluation System.



Student sex was obtained from the examination score sheets. All students belonged to the same curricular year group and took the course together. Date of birth was not available in the examination or platform records used to assemble the analytic dataset; age was therefore not reported.

The primary outcome was the official recorded OSCE station total, comprising history content and communication/comprehensive performance. We examined whether the content-focused AMTES practice score was associated with this broader local assessment of a history-taking encounter, without assuming that the 2 measures represented identical constructs. The history-content subscore, which more directly aligned with the AMTES rubric, was analyzed as a supporting outcome; the communication block provided a comparison with a less directly aligned component. The official total, unchanged, is the score used for progression, resit, and transcript decisions.

Statistical Analysis

Overview

Analyses first addressed the RQs, followed by sensitivity analyses, robustness checks, and additional design checks.

Primary Association (RQ1)

Spearman correlations quantified bivariate associations. Ordinary least squares regression with OSCE-case fixed effects was used to estimate semistandardized slopes. Predictors were *z* scored, whereas the outcome was retained on its original point scale. The fixed effects were included to account for average score differences across the 4 OSCE cases. They adjust only for between-case mean differences and do not, by themselves, account for examiner or station measurement error, which is addressed as a limitation below. For descriptive visualization, students were grouped by quartiles of the platform score, and OSCE scores were compared across quartiles using the Kruskal-Wallis test.

XSL•FO
RenderX

Dependence on the Complaint-Matched AMTES Case (RQ2)

Dependence was examined in 2 ways: (1) by estimating the association within the AMTES-unmatched low back/flank pain OSCE stratum and (2) by examining the association between the leave-matched-case-out platform score and OSCE performance, with the complaint-matched AMTES case score reported as a descriptive comparison. The leave-matched-case-out analysis was restricted to students who had practiced the complaint-matched AMTES case and at least one additional case. All comparisons between the leave-matched-case-out score and the overall mean were computed for the same students.

Internal Cross-Validated Prediction (RQ3)

Candidate predictors were the 2 prior physical-examination course scores and AMTES measures derived from records meeting the student-specific time stamp cut-off. The full model without assigned OSCE-case information included both prior examination scores, the platform score, practice volume, and case breadth. Out-of-sample performance was assessed using internal 5-fold cross-validation, repeated 200 times with a fixed seed. Model fitting and predictor standardization were performed within the training folds. The evaluated models included the platform score alone, both prior examinations alone, their combination, and the full model; OSCE-case fixed effects alone and in combination with the platform score were evaluated as conditional comparators. Cross-validated Spearman ρ, R^2, and mean absolute error (MAE) were calculated from pooled out-of-fold predictions within each repetition and averaged across repetitions. Models incorporating the baseline covariates were fitted in the subset of 471 students with data available for both examination scores. Bootstrap 95% CIs were calculated using 2000 paired student resamples. In each resample, performance was calculated separately from the out-of-fold predictions for each repetition and then averaged across repetitions, using the same resampled students for all model contrasts. Models and folds were held fixed; therefore, these intervals are conditional on the existing out-of-fold predictions and do not incorporate additional uncertainty from refitting, fold generation, or examination clustering (Multimedia Appendix 2).

Sensitivity, Robustness, and Post Hoc Analyses

Sensitivity and robustness checks addressed outcome definition; construct alignment, using the communication block as a less directly aligned comparison outcome; practice dose under fixed exposure; the analysis window; measured baseline physical-examination performance; case breadth; predictor weighting; within-case outcome scaling; case-assignment balance; and class- and examination-day–grouped validation. Post hoc checks explicitly addressed the influence of low scores; score construction and practice strategy; examination-day and case-by-day effects; recorded station-slot labels; and the platform score's incremental contribution beyond measured baseline and practice covariates. Full specifications are provided in Multimedia Appendix 2.

Additional Analyses

Additional analyses comprised a cohort-wide preexamination cut-off; a comparison of students included in and excluded from the leave-matched-case-out analysis; a common-scale comparison of the history-content and communication components; and sex-related analyses of differential performance, differential prediction, and out-of-fold subgroup calibration. The cohort-wide cut-off included only practices time-stamped before the first examination day (June 22, 2026). The primary association and cross-validation comparisons were repeated in the same 469 students, whereas RQ2 analyses were repeated in their respective defined-score samples and reported descriptively. The 2 dependent Spearman correlations were compared using a paired student bootstrap. Exploratory within-student discontinuity and practice-pattern checks were also conducted. All analyses are specified in Multimedia Appendix 2.

Analysis Population and Missing Data

No values were imputed; each model used records with complete data for its variables. Baseline-adjusted analyses were restricted to complete cases with both prior physical-examination course scores, and the unadjusted model was refitted in the same subset. OSCE scores of linked and unlinked eligible examinees were compared using the Mann-Whitney *U* test.

RQ1 was the single primary association, whereas RQ2 and RQ3 were supportive research-question analyses. Per-case, quartile, component, sensitivity, robustness, post hoc, and additional analyses (including the sex-related and cohort-wide-cut-off analyses) were not treated as independent confirmatory tests; therefore, no multiplicity adjustment was applied. Precision was summarized using 2-sided 95% CIs throughout. Regression (ordinary least squares) CIs were calculated using heteroskedasticity-consistent SEs (type 3; HC3), and CIs for Spearman correlations were calculated using the Fisher *z* transformation.

Analyses were conducted using Python 3.12 (pandas, SciPy, and statsmodels; Python Foundation). Analysis and figure-generation code and derived data are available as described in the "Data Availability" section.

OSCE Measurement Evidence and Scoring Sensitivity

The OSCE stations were aligned with the local course blueprint and used predefined case-specific scoring criteria, consistent with established OSCE blueprinting principles [16-18]. Assessor masking and common pre-OSCE training are described above; however, no formal cross-examiner calibration or independent double-rating data were available for this administration. Case-specific Cronbach α was calculated for the 6D communication block. For the history-content checklist, α was reported descriptively rather than treated as sufficient or primary evidence of score reliability, because internal consistency does not capture important rater- and station-related sources of measurement error in performance assessment [9]. In a scoring-rule sensitivity analysis, history-content items were dichotomized using full- and any-credit rules, and the platform-content association was recalculated. As no performance was independently rated by a second examiner

XSL•FO
RenderX

and each student completed only 1 station, interrater reliability and student-level cross-station generalizability could not be estimated in this study [9,30]. Detailed results are reported in Table S2 in Multimedia Appendix 1.

Ethical Considerations

This study received ethical approval from the Ethics Committee of Shantou University Medical College (SUMC-2026-017) and was conducted in accordance with the Declaration of Helsinki. The 6-case AMTES assignment was a required course component and contributed to coursework credit; students received no additional compensation for the research use of their records. Written informed consent covered the research use of platform learning records and their linkage to examination records, including OSCE scores. Before analysis, the linked data were pseudonymized and stored on password-protected institutional servers accessible only to authorized members of the research team.

Results

Sample, Linkage, and Participant Flow

Of 480 eligible examinees with scored OSCEs, 474 (98.8%) were linked to platform records. The analytic sample comprised 473 students with an OSCE history-taking score and analyzable preexamination practice records; 1 linked examinee had no analyzable preexamination practice record (Figure 1). These students contributed 3004 retained preexamination practices, with a median of 6 (IQR 6-7) practices across a median of 6 distinct cases (IQR 5-6 distinct cases; Table 1). Although 366 (77.4%) students had at least six retained practice records, 349 (73.8%) had at least one retained, validly scored record for each of the 6 assigned cases. Students who had not completed the assignment were retained because completion was treated as an exposure characteristic rather than an inclusion criterion (Table 1; Table S1 in Multimedia Appendix 1). OSCE scores had a mean of 81.5 (SD 6.7), a median of 82, and a range of 60-98. Case counts were 146 for chest pain, 110 for cough/dyspnea, 109 for fever/rash, and 108 for low back/flank pain (Table 1).

Mean OSCE scores differed across the 4 allocated case groups (case means 77.1-84.1; 1-way $F_{3,469}$=31.06; P<.001; case-only R^2=0.166). These between-case mean differences cannot be attributed uniquely to case difficulty; the OSCE-case fixed effects used in subsequent analyses adjust for the composite differences without decomposing them. All 9 classes were represented in every case. Students allocated to the 4 cases had similar mean AMTES practice scores (mean normalized scores 0.738, 0.729, 0.730, and 0.727; $F_{3,469}$=0.38; P=.77; R^2=0.002) and similar scores on both prior physical-examination course examinations (theory: case-group means 67.5-68.7; $F_{3,468}$=0.22; P=.88; practical: 36.9-37.5; $F_{3,468}$=0.29; P=.83; n=472 for each). These measured similarities provide no evidence of systematic assignment based on the observed student characteristics, but they do not establish formal randomization or balance in unmeasured characteristics.

Among the 6 unlinked eligible examinees, the mean OSCE score was 78.2, compared with 81.5 among the 474 linked eligible examinees (Mann-Whitney U test, P=.26). Given the small number of unlinked examinees, this comparison is imprecise.

XSL•FO
RenderX

Table 1. Participant, OSCE[a], and preexamination AMTES[b] practice characteristics (N=473).

Characteristics	Values
Participants, n	
Analytic sample	473
Sex, n (%)	
Male	307 (64.9)
Female	166 (35.1)
OSCE performance	
Total score, mean (SD)	81.5 (6.7)
Total score, median (range)	82 (60-98)
Assigned OSCE case, n	
Chest pain	146
Cough/dyspnea	110
Fever/rash	109
Low back/flank pain[c]	108
OSCE score, mean (SD)	
Chest pain	84.1 (6.3)
Cough/dyspnea	77.1 (5.7)
Fever/rash	80.9 (6.4)
Low back/flank pain[c]	83.2 (6.0)
Preexamination AMTES practice	
Mean normalized score, mean (SD)	0.73 (0.09)
Retained practices, median (IQR)	6 (6-7)
Distinct cases practiced, median (IQR)	6 (5-6)
Students with ≥6 retained practice records, n (%)	366 (77.4)
Students who practiced all 6 distinct assigned cases, n (%)	349 (73.8)

[a]OSCE: objective structured clinical examination.

[b]AMTES: AI-Powered Medical History-Taking Training and Evaluation System.

[c]Low back/flank pain was the AMTES-unmatched OSCE presenting-complaint stratum (ie, no assigned AMTES case shared its prespecified presenting-complaint category).

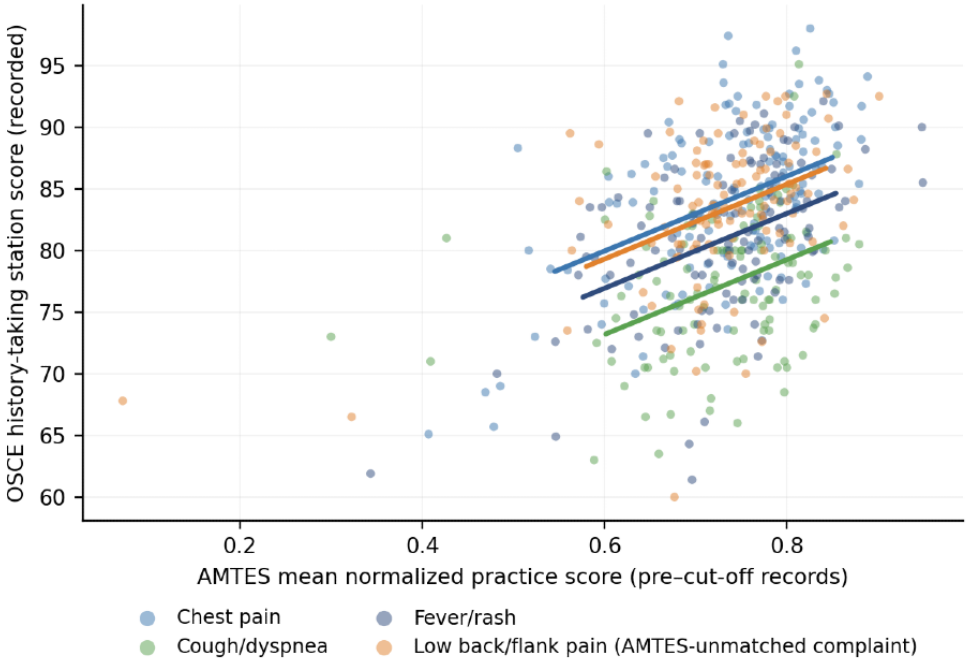
RQ1: AI Trainer Performance Is Associated With the Recorded OSCE Station Score

Mean platform practice performance was associated with the recorded OSCE history-taking station score (Spearman ρ=0.388; *P*<.001; N=473; Figure 2). A similar correlation was observed after OSCE scores were standardized within case (Spearman ρ=0.393; *P*<.001; Table S1 in Multimedia Appendix 1). In the OSCE-case-fixed-effects model, each 1SD increase in platform score was associated with a 2.83-point increase in OSCE score (95% CI 2.26-3.39; *P*<.001; Table 2). Model R^2 increased from 0.166 with case fixed effects alone to 0.344 after the platform score was added.

This estimate was equivalent to approximately 0.42 of the observed OSCE score SD (6.7 points). This was a between-student association; it does not imply that additional practice would raise an individual student's OSCE score by 2.83 points. Mean OSCE scores increased across ascending quartiles of platform performance, from 77.9 in Q1 to 81.1, 82.4, and 84.8 in Q2-Q4, respectively (Figure 3). OSCE score distributions differed across the 4 quartiles (Kruskal-Wallis *H*=60.46, *df*=3; *P*<.001). Practice volume and case breadth showed smaller bivariate correlations with OSCE score (Spearman ρ=0.213 and Spearman ρ=0.169, respectively; both *P*<.001). The case-balanced score yielded a Spearman ρ of 0.399 (95% CI 0.32-0.47) and an association of +2.92 OSCE points per SD (95% CI 2.34-3.49), compared with Spearman ρ=0.388 and +2.83 points for the practice-weighted score (Table S1 in Multimedia Appendix 1). Among the 471 students with both prior physical-examination course scores, adjustment for the 2 examinations reduced the platform coefficient in the case-fixed-effects model from +2.80 (95% CI 2.23-3.36) to +2.15 OSCE points per SD (95% CI 1.61-2.69; *P*<.001). A post hoc nested model that also included practice volume and case breadth yielded +2.16 points per SD (95% CI 1.60-2.72; *P*<.001; Table 2).

XSL•FO
RenderX

Figure 2. Mean normalized AI-trainer practice score versus recorded objective structured clinical examination (OSCE) history-taking station score (N=473). AMTES: AI-Powered Medical History-Taking Training and Evaluation System.



XSL•FO
RenderX

Table 2. Primary and supportive analyses of the association between AMTES[a] performance and the recorded OSCE[b] history-taking station score.

Analysis	n	Statistic	Estimate	95% CI	*P* value
Primary association					
Platform score and OSCE station total	473	Spearman ρ	0.388	0.31-0.46	<.001
OSCE-case fixed effects only	473	Model R^2	0.166	N/A[c]	N/A
OSCE-case fixed effects plus platform score	473	B/OSCE points per SD[d]	2.83	2.26-3.39	<.001
Model fit after adding the platform score	473	Model R^2	0.344	N/A	N/A
Cohort-wide preexamination cut-off sensitivity analysis[e]					
Platform score and OSCE station total, primary window → June 22, 2026, cut-off	469	Spearman ρ	0.385 → 0.346	difference −0.040 (−0.065 to −0.018)	N/A
OSCE-case fixed effects plus platform score, primary window → June 22, 2026, cut-off	469	B/OSCE points per SD[d]	2.83 → 2.68	2.11-3.26 (cut-off)	<.001
Dependence on the complaint-matched AMTES case					
AMTES-unmatched low back/flank pain stratum[f]	108	Spearman ρ	0.370	0.19-0.52	<.001
Leave-matched-case-out platform score[g]	339	Spearman ρ	0.398	0.30-0.48	<.001
Overall mean score in the same 339 students (same-sample comparator)	339	Spearman ρ	0.370	0.27-0.46	N/A
Internal repeated cross-validation[h]					
Platform score only	471	Spearman ρ, R^2, and MAE[i]	0.374/0.176/4.85	N/A	N/A
Full model without assigned OSCE-case information (platform score + 2 prior examinations + practice volume + case breadth)	471	Spearman ρ, R^2, and MAE	0.437/0.235/4.71	0.36-0.51/0.15-0.30/4.40-5.04	N/A
Calibration of the full model without assigned OSCE-case information	471	Calibration-in-the-large and slope	0.000 and 0.949	slope 0.80-1.10	N/A
Baseline and covariate adjustment					
Platform score, unadjusted (complete-case subset)	471	B/OSCE points per SD[d]	2.80	2.23-3.36	<.001
Platform score, adjusted for both prior physical-examination course scores	471	B/OSCE points per SD[d]	2.15	1.61-2.69	<.001
Platform score, adjusted for both examinations, practice volume, and case breadth	471	B/OSCE points per SD[d]	2.16	1.60-2.72	<.001
Component outcomes					
Difference in fully standardized β, history-content subscore minus communication block	473	Δβ	0.202	0.091-0.314	<.001

[a]AMTES: AI-Powered Medical History-Taking Training and Evaluation System.

[b]OSCE: objective structured clinical examination.

[c]N/A: not applicable.

[d]B coefficients represent the difference in OSCE score associated with a 1SD higher platform score. Models included OSCE-case fixed effects and used heteroskedasticity-consistent SEs (type 3) to calculate 95% CIs.

[e]This sensitivity analysis restricted practices to those time-stamped before June 22, 2026, and compared this cut-off with the primary window in the same 469 students. The paired difference in Spearman ρ was computed from unrounded correlations. Full results are in Table S4 in Multimedia Appendix 1.

[f]Low back/flank pain was the AMTES-unmatched OSCE stratum; no assigned AMTES case shared its prespecified presenting-complaint category.

[g]The leave-matched-case-out score excluded practices from the AMTES case matching the assigned OSCE presenting complaint; the comparator used

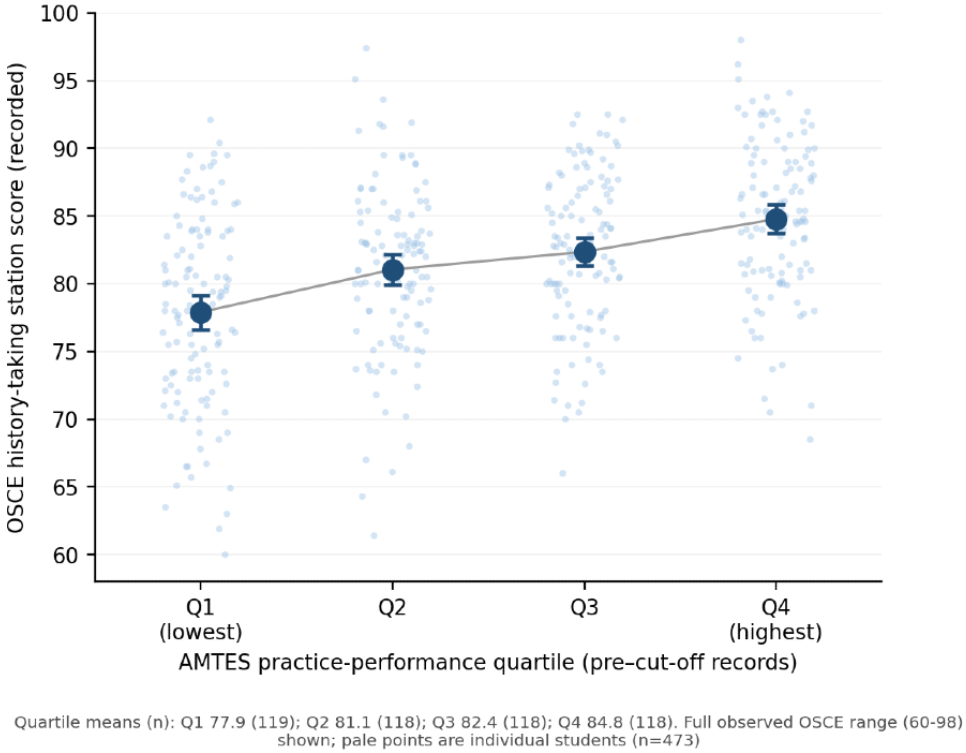
XSL•FO
RenderX

the overall mean score in the same students.

[h]Repeated 5-fold cross-validation used within-fold standardization and model fitting. Reported metrics are averages across repetitions from held-out predictions; models using prior examinations included 471 complete cases. Additional models are in Table S5 in Multimedia Appendix 1.

[i]MAE: mean absolute error.

Figure 3. Unadjusted objective structured clinical examination (OSCE) history-taking scores by quartile of AI-trainer practice performance. AMTES: AI-Powered Medical History-Taking Training and Evaluation System.



Component Outcomes: History-Content Subscore and Communication Block

With OSCE-case fixed effects, the platform score was associated with the history-content subscore at +2.30 points per SD (95% CI 1.83-2.78) and with the communication block at +0.53 points per SD (95% CI 0.27-0.78). The association with the communication block was weaker but not zero. On a common scale, the fully standardized coefficients were 0.391 (95% CI 0.310-0.472) for content and 0.189 (95% CI 0.098-0.279) for communication, a difference of 0.202 (95% CI 0.091-0.314; *P*<.001; Table 2). Standardization equates scale but not reliability. Moreover, the communication block is ceiling-prone and nearly uncorrelated with content (Spearman ρ=0.032). Thus, this difference reflects alignment with observed scores rather than relationships between true scores; no disattenuated estimates are reported. The association was also present when the content subscore replaced the station total as the criterion (Spearman ρ=0.370; 95% CI 0.289-0.445). The difference between the correlations for the total score and content subscore was small (Spearman Δρ=0.018; 95% CI −0.018 to 0.057), although the 2 outcomes shared 70 of the 100 station points (see Table S1 and Figure S1 in Multimedia Appendix 1).

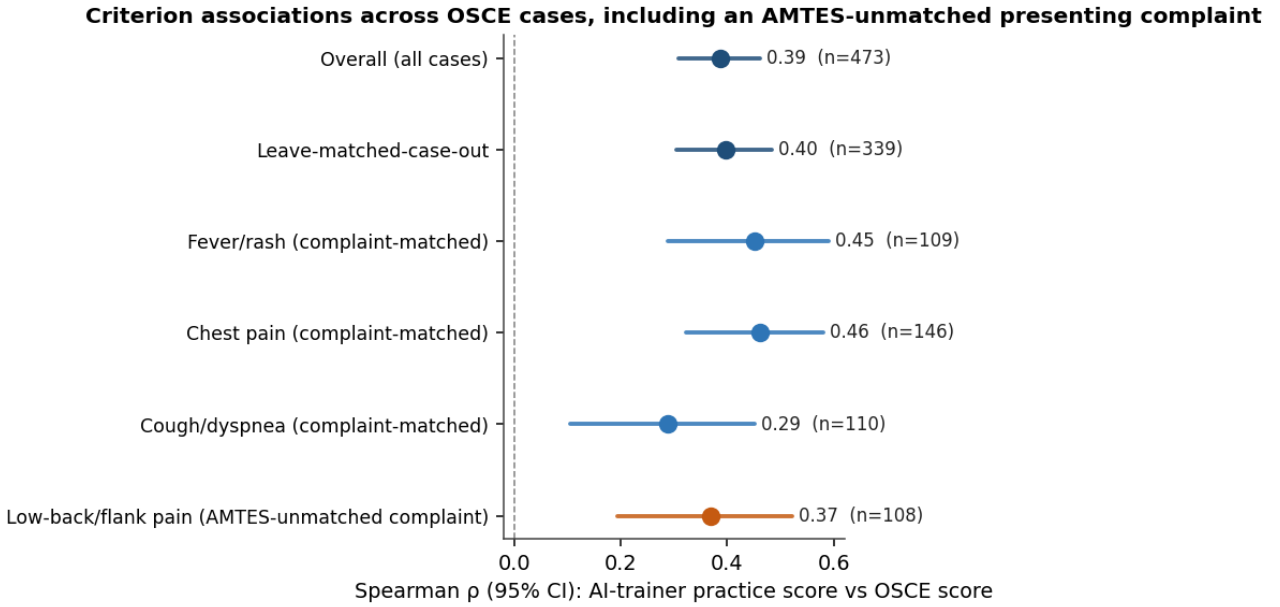
RQ2: The Association Does Not Depend on Including the Complaint-Matched AMTES-Case Score

Within the low back/flank pain OSCE stratum, whose presenting complaint had no match among the assigned AMTES cases, platform performance was associated with OSCE score (Spearman ρ=0.370; 95% CI 0.19-0.52; *P*<.001; n=108). Positive associations were also observed in the 3 presenting-complaint-matched strata: chest pain (Spearman ρ=0.462; n=146), cough/dyspnea (Spearman ρ=0.288; n=110), and fever/rash (Spearman ρ=0.452; n=109; Figure 4 and Table S1 in Multimedia Appendix 1).

Among the 341 students who had practiced the training case matching their OSCE presenting complaint, 339 had practiced at least one additional case and therefore had a defined leave-matched-case-out score. In these 339 students, the leave-matched-case-out score showed a numerically higher correlation with the OSCE station score than the overall mean score (Spearman ρ=0.398; 95% CI 0.30-0.48 vs Spearman ρ=0.370; 95% CI 0.27-0.46). The complaint-matched training-case score itself was less strongly correlated with the OSCE score (Spearman ρ=0.174; *P*=.001; n=341; descriptive comparison). However, the leave-matched-case-out score remained highly correlated with the full platform aggregate (Spearman ρ=0.917; n=339), indicating that this analysis primarily assesses whether the overall association depends on the matched case rather than providing independent evidence of generalization. Both analyses involve between-student comparisons and do not establish within-student transfer across clinical presentations.

XSL•FO
RenderX

Figure 4. Associations between AI-trainer scores and recorded objective structured clinical examination (OSCE) station scores overall, for leave-matched-case-out scores, and by assigned OSCE case. AMTES: AI-Powered Medical History-Taking Training and Evaluation System.



The 26 students in the complaint-matched strata without a defined leave-out score had fewer practices (median 2.5 vs 6), and lower platform scores (0.654 vs 0.739) than the 339 included students. Within the 365-student complaint-matched stratum, the platform-score-by-inclusion interaction was −0.70 points per SD (95% CI −2.96 to 1.55; *P*=.54), providing no evidence of a detectable difference in slope. However, these comparisons do not establish the direction or magnitude of any selection effect (see Table S6 in Multimedia Appendix 1; also see Multimedia Appendix 2).

RQ3: Internal Cross-Validated Performance Without Assigned OSCE-Case Information

Across 200 repetitions of 5-fold cross-validation among the 471 students with both prior examination scores, the platform score alone yielded held-out predictive performance for OSCE scores of Spearman ρ=0.374, R^2=0.176, and an MAE of 4.85 points. The full model without assigned OSCE-case information yielded Spearman ρ=0.437 (bootstrap 95% CI 0.36-0.51), R^2=0.235 (95% CI 0.15-0.30), and an MAE of 4.71 points (95% CI 4.40-5.04), with a calibration slope of 0.949 (95% CI 0.80-1.10).

Adding the platform score to the 2 prior examinations increased cross-validated R^2 from 0.141 to 0.236 (ΔR^2=0.096; 95% CI 0.053-0.142; computed at full precision). By contrast, adding practice volume and case breadth did not improve mean cross-validated performance in this cohort (ΔR^2=−0.002; 95% CI −0.012 to 0.008). Additional model comparisons, calibration details, and prediction-error summaries are reported in Table S5 in Multimedia Appendix 1. Metrics from models that also included the assigned OSCE case are conditional on that assignment (see Figures S2 and S3 in Multimedia Appendix 1; also see Multimedia Appendix 2).

Examination-Week Overlap: Cohort-Wide Preexamination Cut-Off

Applying a single cohort-wide cut-off before the first examination day (June 22, 2026) to all students excluded 391 of the 3004 (13.02%) practices and left 469 students with an available predictor. In these 469 students, the association attenuated modestly: Spearman ρ decreased from 0.385 to 0.346 (95% CI 0.26-0.42; paired difference −0.040, 95% CI −0.065 to −0.018), and the case-fixed-effects coefficient decreased from +2.83 to +2.68 points per SD (95% CI 2.11-3.26), while the quartile gradient remained monotonic. Both RQ2 analyses remained positive (AMTES-unmatched stratum: Spearman ρ=0.323; n=107; leave-matched-case-out analysis: Spearman ρ=0.375; n=321). The cut-off analysis does not identify the cause of the attenuation or exclude the possibility of unrecorded exchanges of examination information, which would leave no practice record. Exploratory timing and practice-pattern analyses are reported in Table S4 in Multimedia Appendix 1 and in Multimedia Appendix 2.

Sex-Related Differences in Scores and Prediction

Sex was recorded for all 473 students (307 male and 166 female). Women scored higher than men on the OSCE (83.1 vs 80.7 points; Cohen *d*=0.36; *P*<.001) and on the platform predictor (0.76 vs 0.72; Cohen *d*=0.50; *P*<.001). The association was present in both sexes (men: Spearman ρ=0.39; +2.65 OSCE points per SD, 95% CI 1.99-3.31; women: Spearman ρ=0.29; +2.84 points per SD, 95% CI 1.92-3.77). No detectable difference in slope was found (interaction +0.30 points per SD, 95% CI −0.88 to 1.49; *P*=.61), although the CI was wide. Adjusting the primary model for sex changed the platform coefficient from +2.83 to +2.71. At the same platform score and assigned OSCE case, women had a higher conditional mean OSCE score (+1.10 points, 95% CI 0.12-2.09; *P*=.03). This difference was concentrated in the history-content component

XSL•FO
RenderX

(+1.03 points, 95% CI 0.20-1.85), whereas the difference in the communication block was small (+0.07 points, 95% CI −0.41 to 0.56). Out-of-fold mean prediction errors from the full model without assigned OSCE-case information were +0.25 points in women and −0.13 points in men; the corresponding calibration slopes were 0.95 and 0.92. Adding sex did not improve prediction (ΔR^2=−0.003). Full results are reported in Table S8 in Multimedia Appendix 1, and the analysis specifications are provided in Multimedia Appendix 2.

Robustness Checks

Examination-day fixed effects (+2.79 points per SD, 95% CI 2.23-3.34), case-by-day cell fixed effects (+2.78 points per SD, 95% CI 2.23-3.33), and recorded station-slot-label fixed effects (+2.83 points per SD, 95% CI 2.25-3.40) produced estimates similar to those of the primary analysis. Similar results were obtained in grouped cross-validation, leaving out 1 class, 1 examination day, or 1 recorded station-slot label at a time. These checks assessed predictive stability across administrative conditions but did not estimate examiner-specific effects or establish homogeneity across days (see Table S1 in Multimedia Appendix 1; also see Multimedia Appendix 2).

Discussion

Principal Findings

This prospectively sequenced observational study examined whether preexamination AMTES performance was associated with the recorded score on an independently assessed OSCE history-taking station, whether the association depended on including the complaint-matched training-case score, and whether a model incorporating the platform score retained predictive performance in held-out students. Building on prior AMTES research on system development, scoring performance, learner acceptance, and educational effectiveness, this study extends the evaluation of routinely generated AMTES scores to their relationship with an independently administered OSCE under required curricular use. Platform performance was positively associated with the score recorded on a single, independently scored OSCE history-taking station after accounting for differences among the assigned OSCE cases. The association was also observed in the AMTES-unmatched presenting-complaint stratum and when the contribution of the presenting-complaint-matched training case was omitted from the platform score. The association attenuated but persisted when the predictor was reconstructed using only practices completed before the first examination day. Models incorporating the platform score retained predictive performance in internal cross-validation and added information beyond the 2 prior physical-examination course scores. The association also persisted after adjustment for those scores, practice volume, and case breadth. Together, these findings provide cohort-scale criterion-related evidence to inform the interpretation of AMTES scores in this curricular context.

Interpretation and Scope of the Evidence

The principal findings support a specific and bounded interpretation of AMTES scores under the current curricular deployment. They provide initial criterion-related evidence for interpreting routinely generated AMTES scores as a candidate indicator of between-student differences relevant to performance on a single, independently scored local OSCE history-taking station; they do not validate AMTES as a general measure of history-taking competence [10,31]. Score interpretations and uses require evidence specific to the proposed context and purpose [10,11,13,31]. The stronger observed association with history content than with the communication block on a common standardized scale is consistent with the content focus of the AMTES rubric. However, standardization addresses differences in scale but does not account for measurement error or the ceiling-prone distribution of communication scores.

The associations observed in the presenting-complaint stratum unmatched to AMTES and with the leave-matched-case-out score suggest that the overall association did not depend on including the complaint-matched AMTES case score. However, the leave-matched-case-out score remained highly correlated with the overall mean platform score, and these analyses compared different students rather than changes within the same students. Therefore, they do not establish within-student transfer of history-taking performance across clinical presentations.

After adjustment for both prior physical-examination course scores, the platform score remained positively associated with the recorded OSCE score, although the association was attenuated. This attenuation suggests that AMTES performance overlaps partly with measured prior physical-examination performance, while the remaining association suggests that the platform score provides information not captured by these prior assessments. However, because the earlier examinations are imperfect proxies for baseline history-taking ability, the residual association cannot be attributed specifically to learning through AMTES or fully separated from unmeasured learner characteristics. Accordingly, these findings are best interpreted as evidence of incremental predictive information within this cohort rather than as evidence of an independent educational effect of AMTES. Whether this additional information improves individual-level educational decisions remains untested. The absence of further average predictive gain from adding practice volume and case breadth should also be interpreted cautiously because both measures showed limited variation under the required 6-case assignment. This finding therefore does not establish that additional or broader practice lacks educational value.

The association persisted when the predictor was restricted to records submitted before the examination period, although it attenuated modestly. This attenuation could reflect fewer observations, the exclusion of later learning or examination-informed practice, or a combination of these mechanisms; the study design cannot distinguish among these explanations. Moreover, persistence of the association under the earlier cut-off does not exclude the possibility that students subsequently received examination information that affected their OSCE performance.

The association between platform and OSCE scores was observed in both sexes, with no detectable difference in slope. However, the interaction estimate was imprecise and does not establish equivalent performance. At the same platform score

XSL•FO
RenderX

and assigned OSCE case, women had a higher conditional mean OSCE score. The model without assigned OSCE-case information showed a smaller sex-related difference in mean out-of-fold prediction error; these results reflect different comparisons. The findings are therefore preliminary and do not establish fairness. Evaluation of subgroup performance, calibration, and the consequences of any score-based decisions will be needed before operational use [32,33].

Comparison With Prior Work

Prior studies of LLM-based virtual patients have primarily examined feasibility, simulated-patient fidelity, learner experience, and agreement between automated and human ratings [1,2]. Previous AMTES studies addressed platform scoring stability and human-AI consistency [3], educational effectiveness under voluntary extracurricular use [26], and learner acceptance [27]. Educational effectiveness and criterion-related evidence address complementary but distinct questions: the former concerns whether an intervention improves outcomes, whereas the latter concerns how the scores generated by the platform should be interpreted.

Intervention studies provide relevant neighboring evidence. A nonrandomized study reported higher preclerkship OSCE interview scores among students who participated in AI-simulated-patient training than among historical controls [25]. A recent quasi-experimental study similarly reported higher OSCE-based interview scores after AI-generated feedback following social-robotic virtual-patient interactions [34]. Both studies evaluated the effects of training or feedback interventions; neither examined whether automated scores accumulated during routine practice corresponded to performance on an independently scored clinical examination. A naturalistic study documented intensive voluntary use of an AI-based OSCE platform during national licensure examination preparation but did not link its routine automated scores to an independent examination outcome. The authors called for controlled or externally benchmarked studies of clinical competence and examination outcomes [35].

More broadly, the technology-enhanced simulation literature has synthesized validity evidence for simulation-based assessment [36], and a study in a conventional problem-based learning curriculum found positive correlations between formative-assessment scores and OSCE and progress-test scores [37]. However, this literature does not establish whether scores generated during cohort-scale curricular use of an LLM-based history-taking trainer correspond to performance on an independently scored OSCE station. This study helps address this gap by linking routinely generated AMTES scores to an external, assessor-blinded OSCE criterion under required curricular use.

Implications for Education and Future Research

The criterion-related and internally cross-validated evidence supports further evaluation of AI-trainer analytics as a candidate scalable, low-stakes formative signal [19,20]. Such analytics could complement, but should not replace, faculty-observed performance assessments within a program of assessment [38].

One potential human-in-the-loop workflow, not tested in this study, would use persistently low platform scores available before an individual student's examination as a low-stakes triage signal for faculty review, rather than as a basis for competence classification or other consequential assessment decisions. Faculty could examine item-level feedback and patterns across cases, confirm any concerns through direct observation or discussion with the student, and offer targeted support, such as focused questioning practice, supervised simulation, or additional feedback. The platform signal would inform faculty judgment; it would not diagnose competence, automatically trigger a consequential decision, or replace direct assessment. Before implementation, evidence would be needed that flagged learners benefit from the support offered and that acting on the signal improves subsequent performance. This study addresses neither question; moreover, the remaining prediction error and the absence of an evaluated decision rule limit immediate individual-level use.

These findings motivate 3 priorities for future research. First, the association between platform scores and the criterion should be prospectively evaluated in new cohorts, institutions, platform versions, and case sets. Second, candidate formative workflows should be evaluated to determine whether faculty review and targeted support initiated by the platform signal improve subsequent learning or clinical performance. Such evaluations should also examine differential performance, subgroup calibration, and potential bias across learner groups beyond the single characteristic examined here. Third, randomized studies of trainer exposure or practice dose are needed to determine whether platform practice itself improves independently assessed history-taking performance. Any alert or decision rule would require separate evidence supporting the specific interpretation and use of the scores [10,11].

Strengths

A key design strength was the course-required deployment within an eligibility-defined curricular cohort. This approach reduced dependence on voluntary platform adoption while preserving variation in preexamination practice behavior. Near-complete record linkage further supported cohort coverage.

The student-specific time stamp restriction and the separately administered OSCE were additional design features. The criterion was independent of the platform in 2 scenarios: it was obtained using a separate assessment instrument whose local rubric was referenced to China's National Medical Licensing Examination, and it was scored by examiners who had no access to students' platform records or scores. This reduces concern that the observed association arose solely from the platform's scoring process or from examiners being influenced by prior knowledge of platform performance. The association was also evaluated under routine curricular examination conditions involving multiple cases, classes, and examination days, rather than under research-optimized assessment conditions.

Under the availability-driven allocation process, examination staff assigned students to available stations, and students did not choose whether they entered the AMTES-unmatched presenting-complaint stratum. Although this process did not constitute formal randomization, it reduced the specific concern

XSL•FO
RenderX

that students selected their OSCE case according to their own preparation or preferences.

Limitations

Overview

This study has several limitations, and these are described in the following subsections.

Observational Design and Residual Confounding

As this was a prospectively sequenced observational study rather than a randomized intervention, residual confounding cannot be fully ruled out. Measured prior physical-examination performance was an imperfect proxy for baseline history-taking ability, and unmeasured differences in learner ability, motivation, or engagement may have contributed to both AMTES and OSCE performance.

OSCE Measurement and Administration

Each student completed only 1 OSCE station, and each performance received a single examiner rating. Consequently, interrater reliability and student-level cross-station generalizability could not be estimated in this study [9,30]. OSCE checklists may also incompletely represent increasing levels of clinical expertise [39]. Case fixed effects accounted for average score differences among the assigned cases but could not separate case, examiner, standardized-patient, station, and session effects. Although examiners were not given access to platform records, random rating error, systematic scoring differences, and possible effects of examiners' prior familiarity with students could not be excluded. Examiner-specific effects could not be estimated because examiner identity and formal cross-examiner calibration data were unavailable. Random rating error would reduce outcome reliability, potentially attenuating score correlations and reducing statistical precision [9], whereas systematic examiner-, case-, or session-related effects could bias the association in either direction. Disattenuated coefficients were therefore not estimated because the required reliability parameters were unavailable.

The recorded station-slot labels represented composite administrative conditions rather than verified examiner identifiers and therefore could not isolate examiner-specific effects. Sensitivity analyses using these labels, examination day, class, and case-by-day cells produced similar associations, but they cannot replace repeated ratings or a crossed station-by-rater design. As the OSCE was administered over 5 days, students examined later had more time to practice, and unrecorded exchange of examination information across days could not be excluded. With only 5 examination days and 9 classes, cluster-level inference was not used for the primary CIs because estimates based on so few clusters would have been unstable. The HC3 CIs therefore relied on student-level independence, whereas grouped cross-validation assessed predictive stability rather than correcting statistical inference for clustering.

Incomplete Record Linkage

A small number of eligible OSCE examinees could not be linked to platform records, and 1 linked examinee had no analyzable preexamination practice. The comparison of OSCE scores between linked and unlinked examinees was imprecise because the unlinked group was small. Residual linkage bias therefore cannot be excluded.

Limited Generalizability and Incomplete Measurement Evidence

The study was conducted at a single institution using 1 cohort, 1 platform version, and 1 OSCE case set. Internal cross-validation used held-out students from the same cohort and therefore does not establish generalizability across institutions, cohorts, platform versions, training cases, OSCE cases, raters, stations, or assessment occasions. Temporal external validation or multisite validation is required before predictive use. Predictor-side item-level scoring properties and item-level construct alignment were not reevaluated, and reproducibility across platform versions and case sets remains unknown. As the trainer's scoring behavior depends on the configured models, prompt templates, case content, and scoring rubrics, replication studies must specify these elements [3]. Sex was the only demographic characteristic available for subgroup analysis. Although no difference in the platform-score slope by sex was detected, the CI was wide and does not establish equivalent performance. Age was not recorded in the source data, and no other demographic or subgroup characteristic was available. Model performance across other learner groups, as well as the consequences of any proposed alert threshold, therefore requires evaluation before operational use.

Operational Decision Use Was Not Evaluated

The study did not develop or prospectively evaluate an alert, classification, or decision rule. No threshold was selected, and sensitivity, specificity, predictive values, alert burden, and the consequences of false-positive and false-negative classifications were not assessed. Individual-level prediction precision was also limited (Table S5 in Multimedia Appendix 1). The findings therefore do not support automated alerts, summative assessment, progression decisions, or other high-stakes uses.

Developer-Evaluator Dual Role

Several authors developed or deployed AMTES (see the "Conflicts of Interest" section), and the system was evaluated at their own institution and within their own curriculum, creating the potential for setting-specific optimization and favorable interpretation. Design features that limit, but do not eliminate, this risk are described in the "Conflicts of Interest" section. Independent external replication by investigators not involved in developing or deploying the platform is required before broader claims can be made.

Conclusions

This study provides cohort-scale criterion-related evidence linking routine AMTES practice scores to recorded scores on an independently scored local OSCE history-taking station under required curricular use. The association did not depend on including the complaint-matched training-case score and attenuated but persisted under a cohort-wide preexamination cut-off. AMTES scores also added predictive information beyond the 2 prior physical-examination course scores in internal cross-validation. These between-student findings do

XSL•FO
RenderX

not establish within-student transfer or causal effects of platform practice, and residual confounding by prior ability may remain.

These findings support further evaluation of AMTES analytics as a candidate low-stakes formative signal for faculty review and targeted support. Independent external validation and prospective evaluation of the proposed workflow are needed before operational use. Any faculty-facing signal should trigger further assessment and support rather than classification of competence, and should complement direct performance assessment.

Funding

We acknowledge support from the Key Project on Medical Education Research by the National Medical Education Development Center, Medical Education Branch of the Chinese Medical Association (grant 2025A25), and the 2025 Teaching Reform and Research Project (AI Curriculum Special Program) of Shantou University Medical College (Number 8). The funders had no involvement in the study design, data collection, analysis, interpretation, or writing of the manuscript.

Data Availability

Deidentified derived analytic data underlying all reported tables and figures, along with the Python analysis and figure-generation scripts, are available from the corresponding author (CS) on reasonable request. As raw platform interaction logs from this single-institution cohort may contain contextual details that increase the risk of reidentification, those materials are not publicly released. A redacted version may be made available on reasonable request, subject to institutional and ethics requirements.

Generative AI Use

The authors declare that generative AI was used in the creation of this manuscript. Generative AI (ChatGPT, Claude) was used for proofreading, editing, or language polishing; translation. The authors verified all outputs of the tool(s) used in the preparation of this manuscript and take full responsibility for the final content of the manuscript.

Authors' Contributions

CS developed the methodology, secured funding, supervised the research team, and led project management. OX conceptualized and designed the study and led project management. WZ conceptualized and designed the study, developed the software system, contributed to data analysis, and prepared the original manuscript draft. YL developed the research methodology, conducted the investigation, performed data analysis, and prepared the original manuscript draft. YZ developed the analytical methodology, performed formal statistical analysis, and generated data visualizations. LW contributed to the conceptualization and design of the study and assisted with project management. XH contributed to the investigation and data curation. XL contributed to the conceptual framework and study design. MY was responsible for data curation and contributed to data analysis. WZ and YL contributed equally to this work and are cofirst authors; CS and OX contributed equally to this work and are cocorresponding authors. All authors reviewed and approved the final manuscript for publication.

Conflicts of Interest

No financial competing interests are declared. Several authors were involved in the development or institutional deployment of AMTES, the platform evaluated in this study, which represents a relevant nonfinancial competing interest (developer/evaluator dual role). The individual roles were as follows: WZ developed the AMTES software system. WZ and YZ were responsible for its institutional deployment and day-to-day operation during the study period; neither took part in OSCE administration or scoring. YL and LW served as OSCE examiners in 3 and 2 of the 9 examination sessions, respectively; like all examiners, they did not know in advance which students they would examine and had no access to individual AMTES practice records or scores when scoring. YZ performed the formal statistical analysis of the linked dataset. The remaining authors (XL, MY, XH, OX, and CS) had no role in AMTES development, deployment, or operation, or in the OSCE. Design features that limit, but do not eliminate, the influence of this dual role are described in the “Methods” section: the OSCE used a separate instrument and was administered and scored outside AMTES by examiners with no access to platform records or scores; the outcome was the officially recorded station total; practice scoring was fully automated; and the linked data were pseudonymized before analysis. These features do not remove the need for independent external replication.

Multimedia Appendix 1
Additional data analysis.
[DOCX File , 326 KB - mededu_v12i1e109752_app1.docx]

Multimedia Appendix 2
Detailed study methods and results.

XSL•FO
RenderX

[DOCX File , 33 KB - mededu_v12i1e109752_app2.docx]

References

1. Holderried F, Stegemann–Philipps C, Herschbach L, Moldt JA, Nevins A, Griewatz J, et al. A Generative Pretrained Transformer (GPT)–powered chatbot as a simulated patient to practice history taking: prospective, mixed methods study. JMIR Med Educ 2024 Jan 16;10:e53961. [doi: 10.2196/53961] [Medline: 38227363]
2. Holderried F, Stegemann-Philipps C, Herrmann-Werner A, Festl-Wietek T, Holderried M, Eickhoff C, et al. A language model–powered simulated patient with automated feedback for history taking: prospective study. JMIR Med Educ 2024 Aug 16;10:e59213. [doi: 10.2196/59213] [Medline: 39150749]
3. Liu Y, Shi C, Wu L, Lin X, Chen X, Zhu Y, et al. Development and validation of a large language model–based system for medical history-taking training: prospective multicase study on evaluation stability, human-AI consistency, and transparency. JMIR Med Educ 2025 Aug 29;11:e73419. [doi: 10.2196/73419] [Medline: 40882613]
4. Cook DA, Triola MM. Virtual patients: a critical literature review and proposed next steps. Medical Education 2009 Mar 18;43(4):303-311. [doi: 10.1111/j.1365-2923.2008.03286.x] [Medline: 19335571]
5. Kononowicz AA, Woodham LA, Edelbring S, Stathakarou N, Davies D, Saxena N, et al. Virtual patient simulations in health professions education: systematic review and meta-analysis by the digital health education collaboration. J Med Internet Res 2019 Jul 02;21(7):e14676 [FREE Full text] [doi: 10.2196/14676] [Medline: 31267981]
6. Cook DA, Hatala R, Brydges R, Zendejas B, Szostek JH, Wang AT, et al. Technology-enhanced simulation for health professions education: a systematic review and meta-analysis. JAMA 2011 Sep 07;306(9):978-988. [doi: 10.1001/jama.2011.1234] [Medline: 21900138]
7. Zeng J, Qi W, Shen S, Liu X, Li S, Wang B, et al. Embracing the future of medical education with large language model-based virtual patients: scoping review. J Med Internet Res 2025 Nov 13;27:e79091 [FREE Full text] [doi: 10.2196/79091] [Medline: 41232097]
8. Downing SM. Validity: on meaningful interpretation of assessment data. Med Educ 2003 Sep;37(9):830-837. [doi: 10.1046/j.1365-2923.2003.01594.x] [Medline: 14506816]
9. Downing SM. Reliability: on the reproducibility of assessment data. Med Educ 2004 Sep;38(9):1006-1012. [doi: 10.1111/j.1365-2929.2004.01932.x] [Medline: 15327684]
10. Messick S. Validity of psychological assessment: validation of inferences from persons' responses and performances as scientific inquiry into score meaning. American Psychologist 1995 Sep;50(9):741-749. [doi: 10.1037/0003-066x.50.9.741]
11. Kane MT. Validating the interpretations and uses of test scores. J Educational Measurement 2013 Mar 14;50(1):1-73. [doi: 10.1111/jedm.12000]
12. Cook DA, Beckman TJ. Current concepts in validity and reliability for psychometric instruments: theory and application. Am J Med 2006 Feb;119(2):166.e7-166.e16. [doi: 10.1016/j.amjmed.2005.10.036] [Medline: 16443422]
13. Cook DA, Brydges R, Ginsburg S, Hatala R. A contemporary approach to validity arguments: a practical guide to Kane's framework. Med Educ 2015 Jun;49(6):560-575. [doi: 10.1111/medu.12678] [Medline: 25989405]
14. Harden RM, Gleeson FA. Assessment of clinical competence using an objective structured clinical examination (OSCE). Med Educ 1979 Jan;13(1):41-54. [Medline: 763183]
15. Harden RM, Stevenson M, Downie WW, Wilson GM. Assessment of clinical competence using objective structured examination. Br Med J 1975 Mar 22;1(5955):447-451 [FREE Full text] [doi: 10.1136/bmj.1.5955.447] [Medline: 1115966]
16. Khan KZ, Ramachandran S, Gaunt K, Pushkar P. The Objective Structured Clinical Examination (OSCE): AMEE Guide No. 81. Part I: an historical and theoretical perspective. Med Teach 2013 Sep;35(9):e1437-e1446. [doi: 10.3109/0142159X.2013.818634] [Medline: 23968323]
17. Pell G, Fuller R, Homer M, Roberts T, International Association for Medical Education. How to measure the quality of the OSCE: a review of metrics - AMEE guide no. 49. Med Teach 2010;32(10):802-811 [FREE Full text] [doi: 10.3109/0142159X.2010.507716] [Medline: 20854155]
18. Newble D. Techniques for measuring clinical competence: objective structured clinical examinations. Med Educ 2004 Mar;38(2):199-203. [doi: 10.1111/j.1365-2923.2004.01755.x] [Medline: 14871390]
19. Ellaway RH, Pusic MV, Galbraith RM, Cameron T. Developing the role of big data and analytics in health professional education. Med Teach 2014 Mar;36(3):216-222. [doi: 10.3109/0142159X.2014.874553] [Medline: 24491226]
20. Chan T, Sebok-Syer S, Thoma B, Wise A, Sherbino J, Pusic M. Learning analytics in medical education assessment: the past, the present, and the future. AEM Educ Train 2018 Apr;2(2):178-187 [FREE Full text] [doi: 10.1002/aet2.10087] [Medline: 30051086]
21. Gierl MJ, Latifi S, Lai H, Boulais AP, De Champlain A. Automated essay scoring and the future of educational assessment in medical education. Med Educ 2014 Oct;48(10):950-962. [doi: 10.1111/medu.12517] [Medline: 25200016]
22. Salt J, Harik P, Barone MA. Leveraging natural language processing: toward computer-assisted scoring of patient notes in the USMLE step 2 clinical skills exam. Acad Med 2019 Mar;94(3):314-316. [doi: 10.1097/ACM.0000000000002558] [Medline: 30540567]

XSL•FO
RenderX

23. Cross J, Kayalackakom T, Robinson RE, Vaughans A, Sebastian R, Hood R, et al. Assessing ChatGPT's capability as a new age standardized patient: qualitative study. JMIR Med Educ 2025 May 20;11:e63353 [FREE Full text] [doi: 10.2196/63353] [Medline: 40393017]
24. Lee HY, Kim J, Choi H, Bae H, Jeong A, Choi S, et al. Comparing AI chatbot simulation and peer role-play for OSCE preparation: a pilot randomized controlled trial. BMC Med Educ 2025 Nov 24;25(1):1755 [FREE Full text] [doi: 10.1186/s12909-025-08308-y] [Medline: 41286823]
25. Yamamoto A, Koda M, Ogawa H, Miyoshi T, Maeda Y, Otsuka F, et al. Enhancing medical interview skills through AI-simulated patient interactions: nonrandomized controlled trial. JMIR Med Educ 2024 Sep 23;10:e58753 [FREE Full text] [doi: 10.2196/58753] [Medline: 39312284]
26. Liu Y, Zhu Y, Zhang W, Lu X, Wu L, Yue M, et al. Real-world impact and educational effectiveness of an AI-powered medical history-taking system: retrospective propensity score-matched cohort study. JMIR Med Educ 2026 Mar 24;12:e89367 [FREE Full text] [doi: 10.2196/89367] [Medline: 41734000]
27. Liu Y, Zhu Y, Shi C, Lu X, Wu L, Yue M, et al. User acceptance of an AI-powered medical history-taking training system among undergraduate medical students: mixed methods study. JMIR Med Educ 2026 Jul 21;12:e101425 [FREE Full text] [doi: 10.2196/101425] [Medline: 42387330]
28. von Elm E, Altman DG, Egger M, Pocock SJ, Gøtzsche PC, Vandenbroucke JP, STROBE Initiative. The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies. Lancet 2007 Oct 20;370(9596):1453-1457 [FREE Full text] [doi: 10.1016/S0140-6736(07)61602-X] [Medline: 18064739]
29. Boulet JR, McKinley DW, Whelan GP, Hambleton RK. Quality assurance methods for performance-based assessments. Adv Health Sci Educ Theory Pract 2003;8(1):27-47. [doi: 10.1023/a:1022639521218] [Medline: 12652167]
30. Crossley J, Davies H, Humphris G, Jolly B. Generalisability: a key to unlock professional assessment. Med Educ 2002 Oct;36(10):972-978. [doi: 10.1046/j.1365-2923.2002.01320.x] [Medline: 12390466]
31. St-Onge C, Young M, Eva KW, Hodges B. Validity: one word with a plurality of meanings. Adv Health Sci Educ Theory Pract 2017 Oct;22(4):853-867. [doi: 10.1007/s10459-016-9716-3] [Medline: 27696103]
32. American Educational Research Association, American Psychological Association, National Council on Measurement in Education. Standards for Educational and Psychological Testing. Washington, DC: American Educational Research Association; 2014.
33. Zwick R. Fairness in educational measurement: theory and concepts. In: Cook LL, Pitoniak MJ, editors. Educational Measurement. 5th ed. New York, NY: Oxford University Press; 2025:383-438.
34. Borg A, Schiött J, Ivegren W, Gentline C, Huss V, Hugelius AM, et al. AI-generated feedback following social robotic virtual patient interactions and medical student performance: nonrandomized quasi-experimental study. JMIR Med Educ 2026 Mar 25;12:e90368 [FREE Full text] [doi: 10.2196/90368] [Medline: 41881044]
35. Song JW, Kim M, Lee YG, Cheong SW, Kim JJ, Cho J, et al. Naturalistic adoption and deliberate practice use of an AI-based OSCE platform during national licensure preparation. NPJ Digit Med 2026 Jul 15 [FREE Full text] [doi: 10.1038/s41746-026-03024-3] [Medline: 42458025]
36. Cook DA, Brydges R, Zendejas B, Hamstra SJ, Hatala R. Technology-enhanced simulation to assess health professionals: a systematic review of validity evidence, research methods, and reporting quality. Acad Med 2013 Jun;88(6):872-883. [doi: 10.1097/ACM.0b013e31828ffdcf] [Medline: 23619073]
37. Couto LB, Durand MT, Wolff ACD, Restini CBA, Faria M, Romão GS, et al. Formative assessment scores in tutorial sessions correlates with OSCE and progress testing scores in a PBL medical curriculum. Med Educ Online 2019 Dec;24(1):1560862 [FREE Full text] [doi: 10.1080/10872981.2018.1560862] [Medline: 31023185]
38. Van Der Vleuten CP. The assessment of professional competence: developments, research and practical implications. Adv Health Sci Educ Theory Pract 1996 Jan;1(1):41-67. [doi: 10.1007/BF00596229] [Medline: 24178994]
39. Hodges B, Regehr G, McNaughton N, Tiberius R, Hanson M. OSCE checklists do not capture increasing levels of expertise. Acad Med 1999 Oct;74(10):1129-1134. [doi: 10.1097/00001888-199910000-00017] [Medline: 10536636]

Abbreviations

AMTES: AI-Powered Medical History-Taking Training and Evaluation System
GAIDeT: Generative AI Delegation Taxonomy
GenAI: generative AI
HC3: heteroskedasticity-consistent SEs (type 3)
LLM: large language model
MAE: mean absolute error
OSCE: objective structured clinical examination
RQ: research question
SP: standardized patient
STROBE: Strengthening the Reporting of Observational Studies in Epidemiology

XSL•FO
RenderX

Edited by T de Azevedo Cardoso; submitted 16.Aug.2026; peer-reviewed by B Depboylu, Y Ohira, A Hudon; comments to author 09.Sep.2026; revised version received 22.Sep.2026; accepted 20.Oct.2026; published 03.Nov.2026.

Please cite as:
Zhang W, Liu Y, Zhu Y, Lu X, Wu L, Yue M, Hong X, Xia O, Shi C
Association of Preexamination AI Virtual-Patient Trainer Scores With Scores on a Local Single-Station History-Taking Objective Structured Clinical Examination: Observational Cohort Study
JMIR Med Educ 2026;12:e109752
URL: https://mededu.jmir.org/2026/1/e109752
doi:10.2196/109752
PMID:

©Weishan Zhang, Yang Liu, Yiying Zhu, Xian Lu, Liping Wu, Minghui Yue, Xiaolin Hong, Oudong Xia, Chujun Shi. Originally published in JMIR Medical Education (https://mededu.jmir.org), 03.Nov.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Education, is properly cited. The complete bibliographic information, a link to the original publication on https://mededu.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Communication Education in Prelicensure and Postlicensure Occupational Therapy: Scoping Review

Ryuichi Shimada[1,2], MS; Yukiko Sano[3], MS; Takeshi Sasaki[2], PhD; Astuko Tanimura[2], PhD

[1]Department of Occupational Therapy, Faculty of Health Sciences, Iryo Sosei University, 5-5-1 Chuodai Iino, Iwaki, Fukushima, Japan

[2]Department of Occupational Therapy, Graduate School of Human Health Sciences, Tokyo Metropolitan University, Arakawa-ku, Tokyo, Japan

[3]Comuro Co, Ltd, Visiting Nursing Station Ai, Fukushima, Fukushima, Japan

Corresponding Author:
Ryuichi Shimada, MS
Department of Occupational Therapy, Faculty of Health Sciences, Iryo Sosei University, 5-5-1 Chuodai Iino, Iwaki, Fukushima, Japan

Abstract

Background: Effective communication, therapeutic relationships, and interprofessional collaboration are essential to high-quality occupational therapy (OT) practice. However, communication education across prelicensure and postlicensure OT training has not been comprehensively mapped.

Objective: This review aims to identify research gaps and map communication education in prelicensure and postlicensure OT education according to content, delivery methods, assessment approaches, and outcomes.

Methods: Following Joanna Briggs Institute guidance, this review was reported using the Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews (PRISMA-ScR) and Preferred Reporting Items for Systematic Reviews and Meta-Analyses Literature Search Extension (PRISMA-S) frameworks. PubMed, CINAHL, ERIC, Web of Science, Ichushi-Web, and Open Access Theses and Dissertations were searched without publication year restrictions; the final search was conducted on May 25, 2026. Eligible sources were English- or Japanese-language publications that examined communication education involving OT students or practicing occupational therapists. The review protocol was developed a priori but was not registered. Two reviewers independently screened records identified through database and gray literature searches, whereas records identified through backward citation searching were screened by one reviewer and checked by the second reviewer. Data were charted for study characteristics, educational content, delivery methods, assessment approaches, and outcomes. Educational outcomes were categorized using the Kirkpatrick model. Formal critical appraisal was not conducted, and the findings were synthesized descriptively.

Results: Database and gray literature searches identified 10,243 records, and backward citation searching identified 9 additional records. After duplicate removal and screening, 55 reports corresponding to 54 studies were included. Of the 55 reports, 47 (85.5%) focused on prelicensure education and 8 (14.5%) on postlicensure education; prelicensure reports mainly involved OT students learning either within OT programs alone or with students from other health professions. Single-group pre-post studies were the most common design (n=22, 40%). Educational content was diverse and included interprofessional communication and teamwork, communication with specific client populations or conditions, model-based communication training, experiential communication skills training, and OT-specific therapeutic communication or therapeutic use of self. Outcome assessment relied mainly on learner-reported measures, including perceived competence, self-efficacy, self-reported communication skills, and attitudes. Objective or observer-rated assessments were less common. Across the 55 included reports, 15 (27.3%) reported reaction-level outcomes, 53 (96.4%) learning-level outcomes, and 5 (9.1%) behavior-level outcomes; no report assessed results-level outcomes, and reports could contribute to more than one level.

Conclusions: This review maps communication education across both prelicensure and postlicensure OT education in the English- and Japanese-language literature. The findings highlight the need for more explicit occupation-centered education, including occupational history taking, values-based goal setting, therapeutic relationship development, and therapeutic use of self. They may inform curriculum design, continuing professional development, and outcome measure selection. Interpretation should remain limited to the included literature, which was dominated by single-group pre-post designs and learner-reported outcomes.

(JMIR Med Educ 2026;12:e95909) doi:10.2196/95909

KEYWORDS

occupational therapy; communication; education; communication skills; scoping review

XSL•FO
RenderX

Introduction

High-quality occupational therapy (OT) depends on effective communication with clients, their families, and other professionals. The primary goal of OT is to support client participation in meaningful daily activities, thereby facilitating role and routine restoration and improving quality of life [1]. Achieving this goal requires occupational therapists to understand clients' values and life histories through client-centered communication and to integrate this information into goal setting and intervention planning. Therapeutic alliances grounded in empathic understanding [2], as well as effective interprofessional collaboration [3], are associated with improved clinical outcomes. Communication functions as a fundamental clinical tool in OT for promoting occupational participation. Therefore, OT education must develop essential professional competencies, including (1) eliciting occupational histories that reflect clients' values and life experiences [1], and (2) intentionally applying therapists' attitudes and engagement strategies to establish and sustain therapeutic relationships and support intervention (therapeutic use of self) [4]. The World Federation of Occupational Therapists (WFOT) Minimum Standards for the Education of Occupational Therapists 2026 identify elements of therapeutic interaction and communication skills, team-based learning and collaboration, and complex clinical reasoning as competencies for which online methods should be supplemented by in-person methods [5]. Although these standards define common competencies for accredited entry-level programs worldwide, they do not prescribe how communication education should be designed, delivered, or assessed. Curricula and teaching methods are shaped by national regulatory frameworks, entry-level degree structures, available educational resources, and cultural and linguistic contexts, which may contribute to variation in how communication education is implemented and reported across countries and settings. In Japan, the "Model Core Curriculum for Occupational Therapist Education" identifies communication competence, including relationship development through active listening and empathy, as a core competency [6]. Together, these international and national frameworks support the need to examine how communication education has been delivered and evaluated in OT education across different educational stages and contexts.

Within OT, theory-informed communication education has been described, including approaches based on the intentional relationship model, which conceptualizes the therapeutic use of self [4]. Recent qualitative work has further described therapeutic use of self as a relational pedagogy in OT education, reflecting growing attention to how communication and therapeutic relationships are taught [7]. However, substantial heterogeneity in educational methods and outcome measures has limited synthesis across studies. Furthermore, outcomes frequently emphasize subjective or indirect indicators, such as self-efficacy and knowledge, rather than observable communication behaviors or client outcomes. Consequently, evidence linking communication education to measurable clinical communication performance and client outcomes remains limited. Previous reviews in other health professions have reported that communication skills training and interprofessional education are associated with improvements in learner-reported outcomes and collaborative practice-related outcomes [8-10]. In OT education, related reviews have examined simulation-based education and coaching [11,12]. However, these bodies of literature do not specifically map how communication education in OT has been designed, delivered, and evaluated across both prelicensure and postlicensure education.

In this review, communication education refers to educational activities designed to develop occupational therapists' competencies in communicating with clients, families, and other professionals across prelicensure and postlicensure education. It includes education related to therapeutic relationships, eliciting clients' occupational histories and values, collaborative goal setting, communication with specific client populations or conditions, and interprofessional collaboration. Accordingly, this scoping review aims to map the content, delivery methods, assessment approaches, and outcomes of communication education in prelicensure and postlicensure OT education and to identify gaps in the existing literature. The review questions were as follows: (1) What study characteristics, educational content, and delivery methods have been reported for communication education in prelicensure and postlicensure OT education? and (2) What assessment approaches and outcomes have been reported?

Methods

Scoping Review Framework and Reporting Guidelines

A scoping review methodology was used to map relevant literature and summarize existing evidence. This approach is appropriate for examining the scope and characteristics of evidence within a field and identifying research gaps, consistent with the objectives of this study [13]. The methodology followed Joanna Briggs Institute guidance [14], and the findings were reported according to the Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews (PRISMA-ScR) framework [15]. The search strategy was reported with reference to the Preferred Reporting Items for Systematic Reviews and Meta-Analyses Literature Search Extension (PRISMA-S) framework [16]. The PRISMA-ScR checklist is included as Checklist 1. Formal critical appraisal of individual sources of evidence, including assessment of methodological quality or risk of bias, was not conducted.

Protocol and Registration

Because communication is a broad and multifaceted concept and is described using diverse terminology in the OT literature, preliminary searches were conducted to clarify relevant terminology and refine the eligibility criteria. A protocol was developed before the formal search and screening as an internal working document. It was not prospectively registered or publicly posted at project initiation. During manuscript revision, the internal protocol was retrospectively archived on the Open Science Framework [17]. The search strategy was subsequently expanded and updated from the protocol by adding education-focused and multidisciplinary databases and a gray literature source. No other major deviations from the a priori protocol occurred.

XSL•FO
RenderX

Eligibility Criteria

Eligibility criteria were formulated using the population-concept-context framework. The "population" comprised OT students and practicing occupational therapists, reflecting both prelicensure and postlicensure education. The "concept" was communication education, as defined in the Introduction section. Studies were eligible when communication was explicitly addressed as an educational objective, educational content, or outcome assessment. Eligible communication topics included, but were not limited to, interpersonal communication, therapeutic relationships, therapeutic use of self, interviewing, shared decision-making, motivational interviewing, patient- or client-centered communication, collaborative goal setting, communication with specific client populations or conditions, and interprofessional communication or teamwork. The "context" included English- and Japanese-language publications without publication year restrictions.

The inclusion criteria were as follows: (1) the studies examined an educational activity related to communication; (2) the participants included OT students or occupational therapists; (3) communication-related elements were explicitly described as an educational objective, educational content, or outcome assessment; and (4) communication-related data were sufficiently extractable for data charting. Studies involving interprofessional education or participants from multiple disciplines were included only when OT students or occupational therapists were included, and communication-related data relevant to the review questions could be extracted. Sources were excluded if they met any of the following criteria: (1) they did not examine communication education; (2) communication-related educational components or outcome data were not sufficiently distinguishable or extractable; (3) they did not include OT students or occupational therapists; (4) they focused only on the development or validation of an assessment measure; (5) they were review papers; or (6) they were books, conference abstracts, editorials, commentaries, or case studies. In addition, studies in which communication was only incidental to the educational activity and was not described as an educational objective, educational content, or outcome assessment were excluded.

Information Sources

PubMed (MEDLINE), CINAHL, ERIC, Web of Science, and Ichushi-Web (Japan Medical Abstracts Society) were searched. Open Access Theses and Dissertations (OATD) was also searched as a gray literature source. These sources were selected to cover biomedical and allied health literature (PubMed, CINAHL), multidisciplinary literature (Web of Science), education-focused literature (ERIC), and Japanese-language literature (Ichushi-Web), with OATD capturing gray literature in the form of theses and dissertations. Controlled vocabulary terms, where available, and free-text terms were combined and adapted for each database. No publication year restrictions were applied. The final search was conducted on May 25, 2026.

In addition to database searching, backward citation searching was conducted by manually screening the reference lists of included reports to identify additional potentially relevant sources. Forward citation searching was not conducted because it was not prespecified in the review protocol and could not be applied consistently to all included reports within the review timeline. Study registries were not searched, and authors or experts were not contacted to identify additional sources. The search strategy was developed and refined through discussion within the research team; however, formal peer review using the Peer Review of Electronic Search Strategies (PRESS) 2015 guideline was not conducted.

Search

Search strategies were developed using the population-concept-context framework. The search covered 3 core concepts: OT, communication, and education. Controlled vocabulary terms, such as MeSH terms, were used where available, and free-text terms were searched using database-specific field modifiers and syntax. The search strategy included terms related to OT, such as "occupational therapy" and "occupational therapist"; communication-related terms, such as "communication," "professional-patient relations," "interprofessional relations," "interview," "therapeutic relationship," "therapeutic use of self," "patient-centered communication," and "interprofessional communication"; and education-related terms, such as "education," "training," "curriculum," "learning," and "competency." No publication year filters were applied. The full search strategies for each database and information source, including database-specific syntax, limits, filters, search dates, and records retrieved, are provided in Multimedia Appendix 1.

Selection of Sources of Evidence

Retrieved records were compiled in Microsoft Excel, and duplicate records were removed by comparing bibliographic information, including titles, authors, publication years, and journal names. Two reviewers (the first and second authors) independently screened records identified through the bibliographic database searches and the OATD gray literature search, with disagreements resolved through discussion and consensus. Records identified through backward citation searching were screened by one reviewer and subsequently checked by the second reviewer; these records were not independently screened in duplicate. Interreviewer agreement before consensus was assessed separately at the title and abstract and full-text screening stages using percentage agreement and unweighted Cohen κ with 95% CIs, calculated using the vcd package (version 1.4 - 14) in R (version 4.6.1; R Foundation for Statistical Computing). The selection process was conducted in two stages. First, titles and abstracts were screened to identify potentially eligible sources. Second, full texts were assessed against the eligibility and exclusion criteria. For sources involving interprofessional education or broader educational programs, eligibility was determined by whether communication-related educational components or outcomes relevant to OT students or occupational therapists could be clearly identified and extracted.

Data Charting Process

A data charting form was developed in Microsoft Excel based on the review questions and eligibility criteria. The form was refined through initial extraction and team discussion to ensure

XSL•FO
RenderX

that the charted items aligned with the review questions. One reviewer (RS) extracted data from all included reports. Two additional reviewers (YS and TS) reviewed the completed data charting spreadsheet and summary tables and checked the corresponding original reports when clarification or verification was needed. Data were not independently extracted in duplicate. Discrepancies or questions were resolved through discussion and consensus. When reports involved interprofessional education or broader educational programs, the reviewers charted only communication-related educational components and outcomes that were relevant to the review questions.

Data Items

The charted data items included bibliographic information, country, publication year, study design, participant characteristics, education stage, educational aims, educational content and delivery methods, communication-related assessment approaches, communication-related outcome measures, and communication-related findings. Educational content categories and Kirkpatrick outcome levels were subsequently assigned during synthesis.

Critical Appraisal of Individual Sources of Evidence

Formal critical appraisal of individual sources of evidence was not conducted. This approach was consistent with the purpose of this scoping review, which was to map the extent, characteristics, and gaps in the published literature rather than to determine intervention effectiveness or the certainty of evidence. Reported outcomes were therefore interpreted descriptively and were not used to infer intervention effectiveness.

Synthesis of Results

Extracted data were summarized using descriptive statistics. Study designs were categorized as randomized controlled trials, nonrandomized controlled trials, single-group pre-post studies, mixed methods studies, qualitative studies, or practice reports/descriptive studies. Educational content categories were developed inductively from the charted data. Educational content and outcomes were initially coded by one reviewer and subsequently reviewed by the research team. Because independent duplicate coding was not conducted, interrater reliability statistics were not calculated. Categories were refined through discussion until consensus was reached. Outcome measures were categorized according to the primary information source or assessment approach into 3 categories: self-report measures, qualitative data sources, and objective or observer-rated assessments. Qualitative data sources included qualitative text analysis, open-ended questionnaires, interviews or focus groups, and qualitative analysis of video-recorded data. Educational outcomes were also classified using the Kirkpatrick model: level 1, reaction; level 2, learning; level 3, behavior; and level 4, results [18]. This model was used to classify outcomes from learners' perceptions and knowledge acquisition to behavioral change and client- or practice-level results. Analyses were conducted at the report level, covering 55 reports. Each report was assigned to a mutually exclusive study design category and one primary educational content category. A report could contribute to more than one Kirkpatrick outcome level when multiple levels were assessed, but it was counted only once within each level. We descriptively cross-tabulated educational content and study design by Kirkpatrick level and visualized these distributions as an evidence gap map and a heatmap; no inferential comparisons were performed. Both visualizations were generated using R (version 4.6.1).

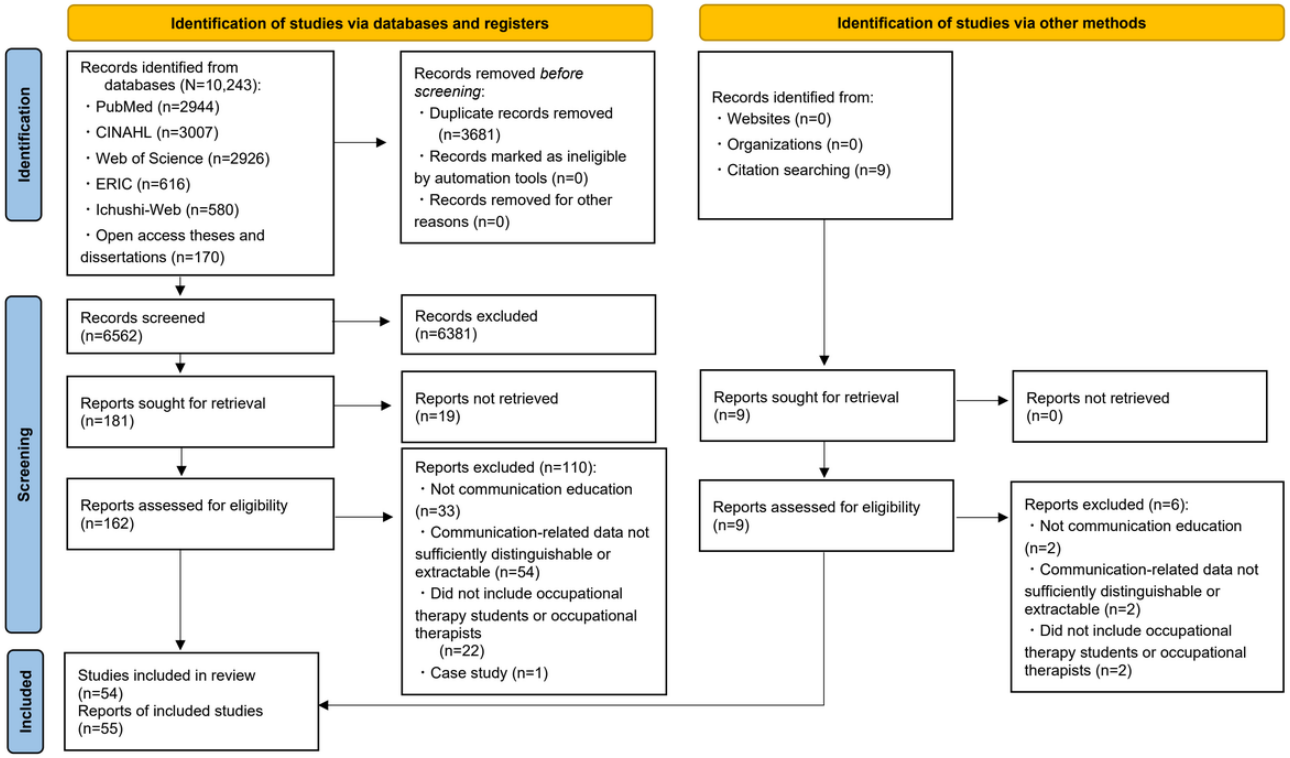
Results

Selection of Sources of Evidence

Database searches and gray literature yielded 10,243 records. After removing duplicates, 6562 records underwent title and abstract screening, resulting in the exclusion of 6381 records. Full texts were sought for 181 reports, of which 19 (10.5%) could not be retrieved. A total of 162 reports were assessed for eligibility, and 110 reports were excluded, yielding 52 reports from the database searches. At the title and abstract screening stage, the 2 reviewers agreed on 6419 of 6562 (97.8%) database-derived records, with an unweighted Cohen κ of 0.610 (95% CI 0.551 - 0.669). At the full-text screening stage, they agreed on 137 of 162 (84.6%) database-derived reports, with an unweighted Cohen κ of 0.601 (95% CI 0.461 - 0.742). Exclusion decisions predominated for both reviewers at the title and abstract screening stage (reviewer 1: 6412/6562, 97.7%; reviewer 2: 6335/6562, 96.5%) and the full-text screening stage (reviewer 1: 120/162, 74.1%; reviewer 2: 119/162, 73.5%). The corresponding expected chance agreement was 94.4% and 61.3%, respectively. All disagreements were resolved through discussion and consensus. Backward citation searching of the reference lists of included reports identified 9 additional records; all 9 full-text reports were retrieved and assessed for eligibility, of which 6 were excluded and 3 were included. Two included reports were based on the same study cohort, with one report examining longer-term outcomes in the same participants (Multimedia Appendix 2). Consequently, 55 reports corresponding to 54 studies were included in the final analysis (Figure 1). Additional details of the included studies are provided in Multimedia Appendix 2.

XSL•FO
RenderX

Figure 1. PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) 2020 flow diagram for the identification and selection of reports on communication education in prelicensure and postlicensure occupational therapy, with searches conducted through May 25, 2026. The flow diagram was adapted from the PRISMA 2020 for new systematic reviews. No automation tools were used to exclude records. ERIC: education-focused literature.



Characteristics of Sources of Evidence

The 54 included studies were represented by 55 reports. Table 1 summarizes the characteristics of the included reports. Most reports were published in 2020 or later (n=29, 52.7%) or between 2010 and 2019 (n=24, 43.6%). The largest number of reports came from the United States (n=22, 40%), followed by Australia (n=13, 23.6%). Most reports focused on prelicensure education. Studies involving OT students and students from other health professions were most common (n=28, 50.9%), followed by studies involving only OT students (n=20, 36.4%). Reports involving practicing occupational therapists were limited (n=7, 12.7%), including 2 reports involving only occupational therapists (n=2, 3.6%) and 5 involving occupational therapists and other professionals (n=5, 9.1%). Single-group pre-post studies were the most frequent design (n=22, 40.0%), followed by mixed methods studies (n=15, 27.3%). Controlled intervention studies accounted for 7 reports, including randomized controlled trials (n=5, 9.1%) and nonrandomized controlled trials (n=2, 3.6%).

XSL•FO
RenderX

Table . Characteristics of the 55 included reports (54 studies) on communication education in prelicensure and postlicensure occupational therapy (OT) education, 1997 - 2026[a].

Characteristics and category	Reports, n (%)
Publication year	
1997 - 2009	2 (3.6)
2010 - 2019	24 (43.6)
2020 - 2026	29 (52.7)
Country	
United States	22 (40)
Australia	13 (23.6)
Japan	6 (10.9)
United Kingdom	4 (7.3)
Taiwan	3 (5.5)
Norway	2 (3.6)
Canada	1 (1.8)
China	1 (1.8)
Ireland	1 (1.8)
Israel	1 (1.8)
Sweden	1 (1.8)
Participant group	
OT students only	20 (36.4)
OT students and other health profession students	28 (50.9)
Occupational therapists only	2 (3.6)
Occupational therapists and other professionals	5 (9.1)
Study design	
Single-group pre-post study	22 (40)
Mixed methods study	15 (27.3)
Descriptive survey or practice report	6 (10.9)
Randomized controlled trial	5 (9.1)
Qualitative study	5 (9.1)
Nonrandomized controlled trial	2 (3.6)

[a]Percentages were calculated using all included 55 reports as the denominator. One study (Country column) was represented by 2 reports; therefore, counts are reported at the report level.

Outcome Measures Used to Evaluate Communication Education

Table 2 summarizes the assessment approaches used to evaluate communication education. Overall, 106 assessment items were extracted. Self-report measures were the most frequently used approach, accounting for 53 (50%) items and primarily assessing perceived competence, confidence, self-efficacy, communication skills, and attitudes. Qualitative data sources accounted for 29 (27.4%) items and included open-ended questionnaires, focus groups, and interviews exploring learners' experiences and perceived changes after education. Objective or observer-rated assessments accounted for 24 (22.6%) items and included knowledge tests, objective structured clinical examinations, rubric-based ratings, video coding, and external ratings. Overall, evaluations of communication education were more often based on learner-reported outcomes than on objective, observer-rated, or performance-based assessments.

XSL•FO
RenderX

Table . Assessment approaches used across 55 reports representing 54 studies of communication education in prelicensure and postlicensure occupational therapy, 1997 - 2026 (N=106)[a].

Assessment approach	Description	Examples	Items, n (%)
Self-report measures	Learner-reported scales or questionnaires assessing perceived competence, confidence, or self-efficacy	SETUS[b], ICCAS-R[c], and original questionnaires	53 (50)
Qualitative data sources	Interview or text-based data used to explore learner experience, perceived changes after education	Open-ended questionnaires and focus groups	29 (27.4)
Objective or observer-rated assessments	Structured assessments of knowledge or communication performance rated by faculty, standardized patients, or external raters	Knowledge tests, OSCE[d], rubric-based ratings, video coding, and MITI[e]	24 (22.6)

[a]Counts represent assessment items, not reports. Each assessment item was assigned to one primary assessment approach. A single report could include multiple assessment items and could contribute to more than one assessment category. Percentages were calculated using the total number of extracted assessment items as the denominator (N=106).

[b]SETUS: self-efficacy for therapeutic use of self.

[c]ICCAS-R: Interprofessional collaborative competencies attainment survey (revised).

[d]OSCE: objective structured clinical examination.

[e]MITI: motivational interviewing treatment integrity.

Content, Delivery Methods, and Outcomes of Communication Education

Educational Content and Delivery Methods

Table 3 summarizes the educational content categories by education stage. Across the 55 reports, most reports focused on prelicensure education (n=47, 85.5%), whereas postlicensure education was reported less frequently (n=8, 14.5%). The most common categories were communication with specific client populations or conditions (n=14, 25.5%) [19-32] and interprofessional communication and teamwork (n=14, 25.5%) [33-46]. Communication with specific client populations or conditions included education focused on communication with people with aphasia, dementia, communication disorders, or cancer. Interprofessional communication and teamwork were limited to prelicensure education and included interprofessional education simulations, interprofessional cotreatment, and virtual patient learning. Technique- or model-based communication training was reported in 9 (16.4%) reports [47-55] and included approaches such as motivational interviewing, shared decision-making, and client-centered counseling. General experiential communication skills training was reported in 10 (18.2%) reports [56-65] and included broad communication skills training and real-patient conversation practice. OT-specific therapeutic communication and therapeutic use of self was reported in 8 (14.5%) reports [66-73], including education grounded in therapeutic use of self and therapeutic relationship development. Delivery methods varied across reports and included lectures, simulation-based learning, role-play or standardized patient encounters, real-patient encounters, discussions, reflective activities, and online or virtual learning. These methods were often combined within the same educational program. Details for each report are provided in Multimedia Appendix 2.

XSL•FO
RenderX

Table . Educational content categories across 55 reports representing 54 studies of communication education in prelicensure and postlicensure occupational therapy (OT), 1997 - 2026[a].

Educational content category	Definition	Examples	Prelicensure, n (%)	Postlicensure, n (%)	Total reports, n (%)
OT-specific therapeutic communication and therapeutic use of self	Education grounded in OT-specific concepts of therapeutic communication	IRM[b] or therapeutic use of self [66-73]	7 (12.7)	1 (1.8)	8 (14.5)
Communication with specific client populations or conditions	Education focused on communication strategies for specific client groups or conditions	Communication with people with aphasia, dementia, communication disorders, and cancer [19-32]	9 (16.4)	5 (9.1)	14 (25.5)
Interprofessional communication and teamwork	Education addressing communication within interprofessional collaboration	IPE[c] simulation and interprofessional cotreatment [33-46]	14 (25.5)	0 (0)	14 (25.5)
Technique- or model-based communication training	Education based on a named communication approach or model	Motivational interviewing and shared decision-making [47-55]	7 (12.7)	2 (3.6)	9 (16.4)
General experiential communication skills training	Broad experiential learning aimed at general communication	Communication skills training and real-patient conversation practice [56-65]	10 (18.2)	0 (0)	10 (18.2)
Total			47 (85.5)	8 (14.5)	55 (100)

[a]Percentages were calculated using all included reports as the denominator (N=55). Each report was assigned to one primary educational content category. Reference numbers in brackets indicate the reports in each category.

[b]IRM: intentional relationship model.

[c]IPE: interprofessional education.

Outcomes of Communication Education

Across the 55 reports, 53 (96.4%) reported level 2 (learning) outcomes, 15 (27.3%) level 1 (reaction) outcomes, 5 (9.1%) level 3 (behavior) outcomes, and none reported level 4 (results) outcomes. Percentages do not sum to 100% because reports could contribute to more than one Kirkpatrick level.

Figure 2 presents an evidence gap map of educational content categories and Kirkpatrick outcome levels. Learning-level outcomes were the most frequently reported across all educational content categories. These outcomes included changes in knowledge, confidence, self-efficacy, perceived communication skills, attitudes, and communication-related understanding. Reaction-level outcomes were reported less frequently and were mainly identified in general experiential communication skills training, interprofessional communication and teamwork, and communication with specific client populations or conditions. Behavior-level outcomes were limited. They were identified in technique- or model-based communication training, communication with specific client populations or conditions, interprofessional communication and teamwork, and general experiential communication skills training. None of the reports included in this review assessed results-level outcomes, such as client-level outcomes, occupational participation, quality of life, or broader practice-level effects. Overall, within the reports included in this review, communication education outcomes were concentrated at the learning level, with relatively limited evidence on behavioral change and no reports on results-level outcomes.

Figure 3 presents a heatmap of study designs and Kirkpatrick outcome levels. Learning-level outcomes were represented across all study designs and were most frequently reported in single-group pre-post studies (21 reports) and mixed methods studies (15 reports). Reaction-level outcomes were most frequent in descriptive/practice reports (6 reports) and were not reported in randomized or nonrandomized controlled trials. Behavior-level outcomes were reported in only 5 reports: 3 single-group pre-post studies, 1 mixed methods study, and 1 nonrandomized controlled trial. Behavior-level outcomes were not reported in randomized controlled trials, qualitative studies, or descriptive/practice reports, and no reports in any study design category assessed results-level outcomes.

XSL•FO
RenderX

Figure 2. Evidence gap map of educational content categories and Kirkpatrick outcome levels across 55 reports of prelicensure and postlicensure OT communication education, 1997 - 2026. Bubble sizes are scaled to the number of reports, as indicated by the size legend (1, 2, 5, and 10 reports); blank cells indicate zero reports. Circles indicate prelicensure education, and triangles indicate postlicensure education. Kirkpatrick level 1=reaction; level 2=learning; level 3=behavior; and level 4=results. None of the included reports assessed level 4 results-level outcomes. Each report was assigned to one primary educational content category; reports could contribute to more than one Kirkpatrick outcome level when multiple outcomes were assessed. Category labels were shortened for readability; full category names are provided in Table 3. OT: occupational therapy.

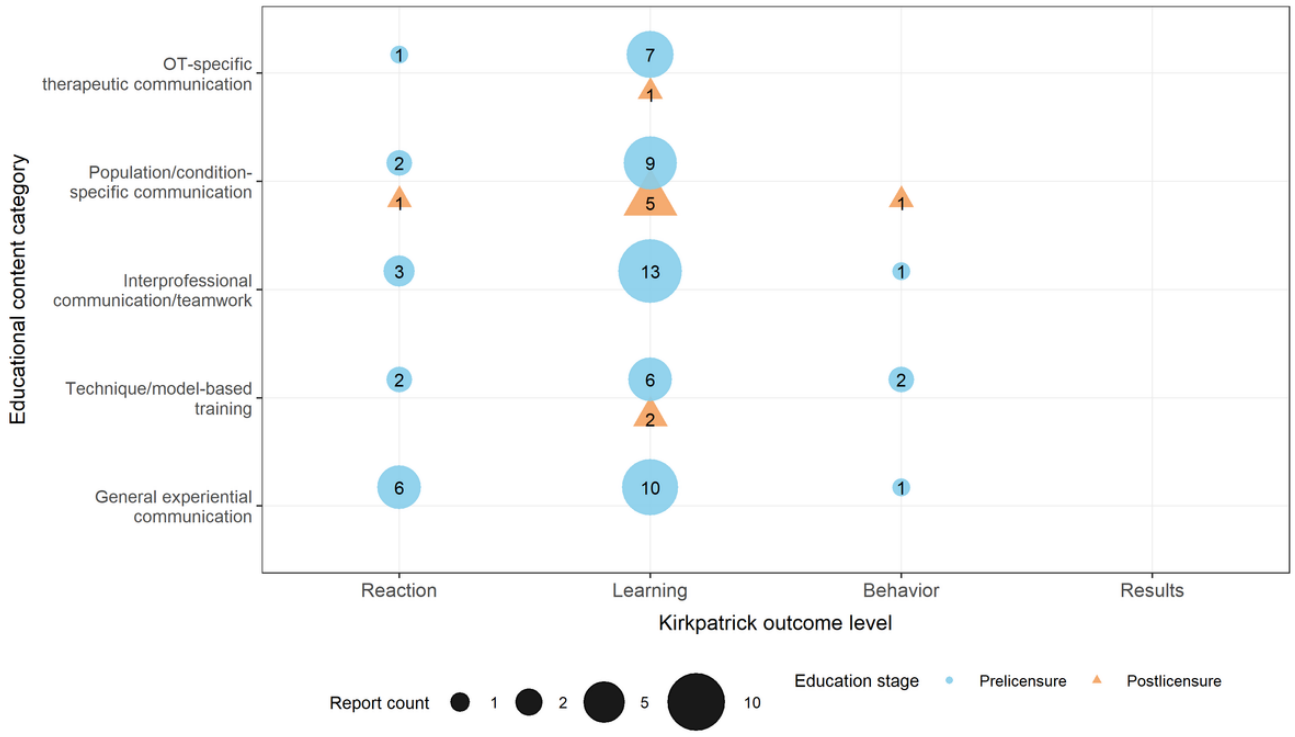
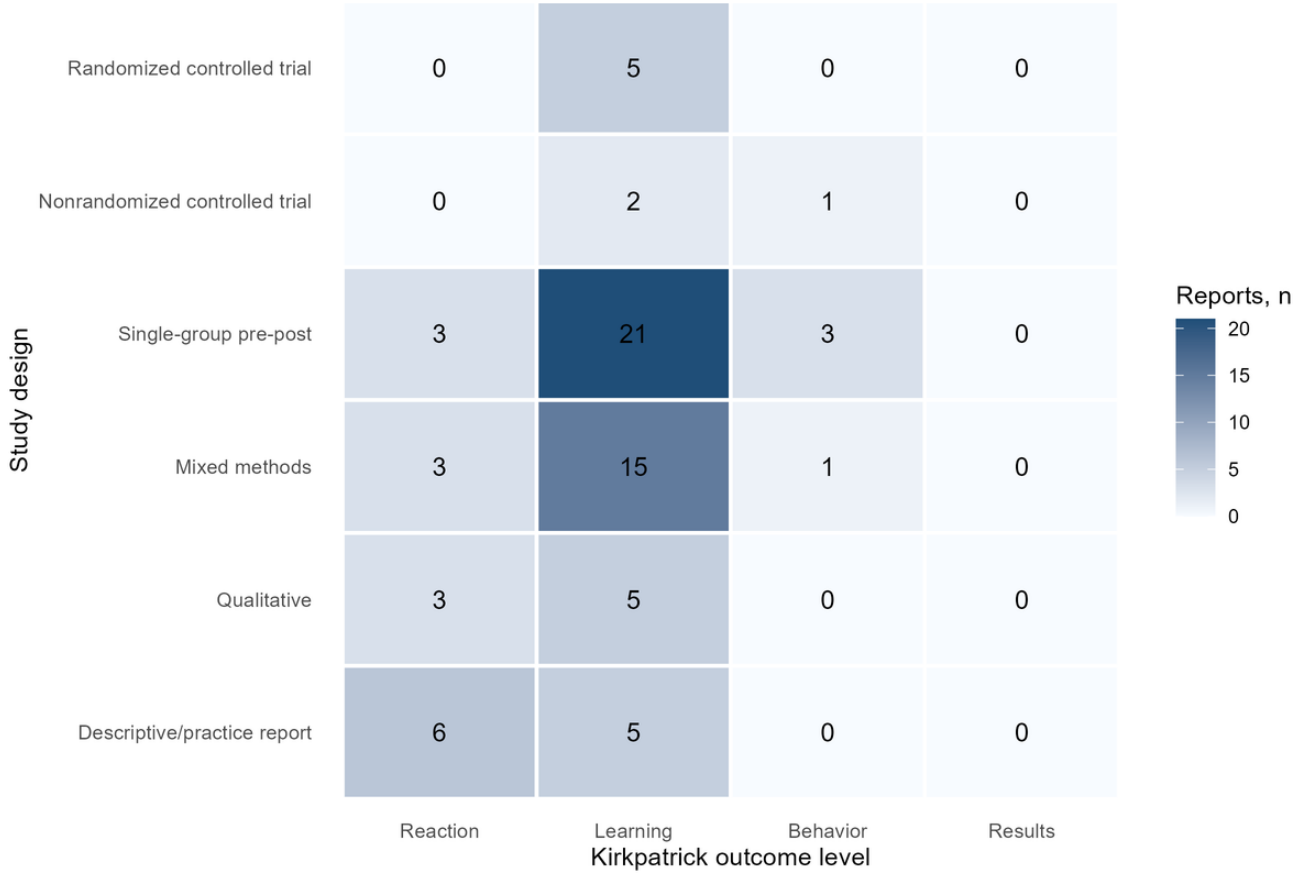


Figure 3. Heatmap of study designs and Kirkpatrick outcome levels across 55 reports of communication education in prelicensure and postlicensure occupational therapy, 1997 - 2026. Cell values and color intensity indicate the number of reports. Reports could contribute to more than one Kirkpatrick outcome level when multiple outcomes were assessed. Kirkpatrick level 1=reaction; level 2=learning; level 3=behavior; and level 4=results.



XSL•FO
RenderX

Discussion

Summary of Evidence

The included reports were concentrated mainly in prelicensure education and addressed diverse educational content and delivery methods. Outcome assessment relied primarily on learner-reported measures, whereas objective or observer-rated assessments were less common. Within the reports included in this review, Kirkpatrick outcomes were concentrated at the learning level, behavior-level outcomes were limited, and results-level outcomes were not assessed. A similar pattern was observed across study designs, with none of the randomized controlled trials assessing behavior- or results-level outcomes. Overall, these patterns highlight the limited evaluation of observable communication behavior and client- or practice-level outcomes.

Comparison With Existing Literature

The concentration of reports in prelicensure education is consistent with broader trends in health professions education, where communication skills, teamwork, and simulation-based learning are commonly introduced during professional preparation [8,10,11]. Previous reviews have reported improvements in learner-reported outcomes, including self-efficacy and confidence, after communication skills training [8], and have shown that simulation-based education is widely used in OT education [11]. The present review extends this literature by showing that, within OT, communication education has often been reported in interprofessional or condition-specific contexts rather than as a distinct occupation-centered educational domain. The prominence of interprofessional communication and teamwork also aligns with the collaborative nature of health and social care practice [3,10]. Interprofessional education has been examined as a strategy to prepare students and practitioners for collaborative practice [10]. In the included OT literature, communication was often embedded within teamwork, role clarification, and other collaborative learning activities. This positioning may provide meaningful opportunities for communication learning [10,11]. However, because communication was often embedded within broader educational activities in the included reports, OT-specific communication tasks may have been less visible.

The findings also correspond with OT literature emphasizing therapeutic relationships and therapeutic use of self as core components of practice [4,74]. These perspectives suggest that OT communication involves not only information exchange but also efforts to understand clients' values, life histories, occupational identities, and participation goals [1,4,74]. However, the included literature did not consistently make these occupation-centered communication processes explicit as educational targets or assessment outcomes. Alongside this limited specification of occupation-centered communication, the outcome patterns identified in this review are consistent with concerns in the wider communication education literature [75-77]. Previous reviews have noted limitations in outcome measurement, including reliance on learner-reported outcomes and limited assessment of behavioral transfer or client-level outcomes [75-77]. Similarly, among the reports included in this review, learning-level outcomes predominated, behavior-level outcomes were uncommon, and no results-level outcomes were reported. These findings caution against assuming that improvements in learner-reported outcomes translate directly into observable clinical performance or benefits for clients and practice [75-77]. The diversity of educational content and delivery methods identified in this review may appear inconsistent with the existence of entry-level OT programs guided by the WFOT minimum standards [5]. However, the WFOT standards define competencies to be achieved without standardizing specific teaching methods [5]. Programs may therefore adapt communication education to their regulatory, institutional, cultural, and linguistic contexts. In addition, communication education was often embedded within interprofessional or condition-specific learning activities rather than delivered as a distinct curricular unit, and reporting practices may differ across countries and journals. These explanations are hypotheses derived from the mapped literature and warrant comparative research on how communication education is structured across accredited programs.

Implications for OT Education and Practice

The findings suggest that communication may need to be made more explicit as a learning objective in OT education and in reports of educational interventions. Communication education was often embedded in interprofessional education, condition-specific training, and experiential learning activities such as simulation, role-play, standardized patient encounters, real-patient encounters, discussion, and reflection. These approaches may provide useful opportunities for experiential learning [10,11]. Their educational value may be enhanced when communication learning is grounded in OT-specific clinical situations and occupation-centered practice [1,4]. For OT, communication education may extend beyond general interpersonal skills to include communication tasks central to the OT process [1,4]. These include eliciting clients' occupational histories, understanding values and life contexts, negotiating occupation-based goals, developing therapeutic relationships, and using the self intentionally in intervention [1,4]. Interprofessional and condition-specific communication education are important, but the findings indicate a need to make occupation-centered communication more visible in curricula and educational reports. Making these communication processes explicit may help learners connect communication skills with core OT practices, including collaborative goal setting, therapeutic relationship building, and client engagement in occupation-based intervention [1,4]. The mapped educational content may also inform continuing professional development planning by helping educators identify communication competencies that require reinforcement or further development after entry to practice.

Educational programs may also use the assessment approaches mapped in this review to select outcome measures that align with their communication learning objectives [76]. Reflective self-assessment can support learners' awareness and confidence, but it may need to be complemented by objective or observer-rated approaches when the intended outcome is communication performance [76]. Examples include structured feedback, rubric-based ratings, standardized patient assessments,

XSL•FO
RenderX

objective structured clinical examinations, and video-based review or interaction analysis of communication behavior [76,78,79].

Implications for Future Research

Future research could examine whether communication education is associated with observable changes in communication behavior and whether these changes are sustained in clinical settings. Objective or observer-rated assessments, workplace-based evaluation, and longitudinal designs may help determine whether learners can apply communication skills beyond the educational setting [76,77]. Such studies should distinguish performance demonstrated in simulated or educational settings from communication applied in routine practice and should report the timing, clinical context, and raters used for assessment [77]. Further research is also needed on outcomes beyond the learning level. Future studies could examine whether communication education influences therapeutic relationship quality and client engagement and, where feasible, whether it is associated with occupational participation, quality of life, or broader practice-level outcomes. Evaluation approaches that combine multiple assessment methods and both quantitative and qualitative information may be useful for examining measurable behavioral change as well as learners' experiences of applying communication skills in practice [79]. Postlicensure communication education remains underrepresented in the published literature and warrants further investigation. Communication demands may become more complex after entry to practice, as practitioners report challenges related to therapeutic use of self and therapeutic relationships [74], yet the published literature has focused mainly on prelicensure education. Future studies could examine continuing professional development programs for practicing occupational therapists and clarify how advanced communication competencies, such as therapeutic use of self and communication in complex clinical situations, can be taught and evaluated.

Limitations

This review has some limitations. First, several aspects of the search process may have led to relevant sources being missed. The search was limited to English- and Japanese-language publications, the languages the review team could screen and chart reliably. In addition, 19 reports could not be retrieved for full-text assessment. Because the content and eligibility of these unavailable reports could not be assessed, their exclusion may have introduced availability-related selection bias, potentially affecting the mapped distribution of educational content, settings, and outcomes. Forward citation searching, searches of study registries, and contact with authors or experts were not undertaken because they were not prespecified in the review protocol and could not be completed within the review timeline. Formal PRESS review was considered during manuscript revision but could not be undertaken because an independent information specialist with appropriate expertise was not available within the revision period. Because backward citation searching identifies earlier references but not subsequent reports citing the included literature, the absence of forward citation searching may have resulted in relevant recent literature being missed. Furthermore, the included literature was concentrated in a small number of mainly English-speaking countries, and unpublished curricula and local continuing professional development programs may be underrepresented. This geographic concentration may limit the geographic and cultural representativeness of the mapped literature. Because exclusion decisions predominated, particularly during title and abstract screening, the κ estimates should be interpreted alongside percentage agreement. The resulting high expected chance agreement may have contributed to the discrepancy between the high percentage agreement and the moderate κ values. Second, although a review protocol was developed before the formal search and screening, it was not prospectively registered before the review began. The protocol was retrospectively archived on the Open Science Framework during manuscript revision [17], providing readers with access to the internal protocol. However, because public archiving occurred after completion of the review, it cannot independently verify that the methods were prespecified or clearly distinguish prospectively planned methods from decisions made during the review process. This limits the external verifiability of the review process. Third, no formal critical appraisal of individual sources was conducted because the purpose of this scoping review was to map the extent and characteristics of the existing literature rather than to evaluate the effectiveness of educational interventions. Consequently, the findings should be interpreted as a descriptive overview rather than evidence of educational effectiveness, and counts of assessment items and Kirkpatrick outcome levels should be interpreted as indicators of how outcomes were reported rather than as effect estimates. In addition, the boundaries of communication education were not always clearly defined in the included literature. Interprofessional education, teamwork, and client-centered practice sometimes incorporated communication-related elements without identifying communication as their primary educational focus. To improve conceptual consistency, we included only reports in which communication was explicitly described as an educational objective, educational content, or outcome assessment. Nevertheless, the included reports were heterogeneous in participants, educational content, delivery methods, and outcome measures, and some degree of interpretive judgment was unavoidable when determining eligibility and classifying educational content. Educational content and Kirkpatrick outcome levels were not independently coded by multiple reviewers. Therefore, interrater reliability was not assessed, and the classifications may have been influenced by reviewer judgment despite team review and consensus.

Conclusions

This scoping review maps communication education across both prelicensure and postlicensure OT education in the English- and Japanese-language literature, extending previous reviews that examined specific educational methods such as simulation-based education [11] and coaching [12]. The findings indicate that, among the reports included in this review, most focused on prelicensure education and learner-reported outcomes; fewer addressed postlicensure education or observable communication behavior, and none assessed client- or practice-level outcomes. These findings can inform

XSL•FO
RenderX

curriculum design, continuing professional development planning, and the selection of communication-related outcome measures in OT education. In particular, they suggest a need to make occupation-centered communication—such as eliciting occupational histories, negotiating occupation-based goals, developing therapeutic relationships, and applying the therapeutic use of self—more explicit in educational objectives and assessment. Future research could examine whether such education is associated with observable and sustained changes in communication behavior and with outcomes that are meaningful to clients and practice. The findings should be interpreted as a map of the published literature included in this review rather than as a comprehensive representation of communication education practices worldwide.

Acknowledgments

The authors thank members of the seminar at Tokyo Metropolitan University for their feedback on the manuscript structure and presentation. The manuscript was reviewed and edited by Editage [80] for English-language editing. All AI-assisted outputs, including quantitative checks, suggested references, interpretations, and wording, as well as all professionally edited text, were independently verified and revised by the authors, who take full responsibility for the final manuscript.

Generative AI Use

Using the GAIDeT Declaration Generator, the first author (RS) reports the use of ChatGPT (GPT-5.5 and GPT-5.6; OpenAI) under full human supervision for literature search and systematization, validation, text generation, proofreading and editing, summarizing text, formulation of conclusions, translation, reformatting, quality assessment, identification of limitations, recommendations, and publication support during manuscript drafting and revision in response to peer review.

Funding

This work was supported by JSPS KAKENHI (grant JP25K06473). The funder had no role in study design; data collection, analysis, and interpretation; manuscript preparation; or the decision to submit the manuscript for publication.

Authors' Contributions

RS conceptualized the review, conducted study screening, charted and synthesized the data, and drafted the manuscript. The search strategy was developed and refined through discussion among all authors. YS independently screened the records. YS and TS reviewed the charted data and summary tables, and TS reviewed the screening results. TS and AT contributed to discussions on classification and categorization, reviewed the manuscript structure and content, and provided recommendations. All authors approved the final version for publication and accept accountability for the integrity of the work.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Search strategies.
[DOCX File, 19 KB - mededu_v12i1e95909_app1.docx]

Multimedia Appendix 2
Details of the included studies.
[XLSX File, 27 KB - mededu_v12i1e95909_app2.xlsx]

Checklist 1
PRISMA-ScR checklist.
[DOCX File, 88 KB - mededu_v12i1e95909_app3.docx]

References

1. Occupational therapy practice framework: domain and process—fourth edition. Am J Occup Ther 2020 Aug 1;74(Supplement_2):7412410010p1-7412410010p87. [doi: 10.5014/ajot.2020.74S2001] [Medline: 34780625]
2. Hall AM, Ferreira PH, Maher CG, Latimer J, Ferreira ML. The influence of the therapist-patient relationship on treatment outcome in physical rehabilitation: a systematic review. Phys Ther 2010 Aug;90(8):1099-1110. [doi: 10.2522/ptj.20090245] [Medline: 20576715]
3. Reeves S, Pelone F, Harrison R, Goldman J, Zwarenstein M. Interprofessional collaboration to improve professional practice and healthcare outcomes. Cochrane Database Syst Rev 2017 Jun 22;6(6):CD000072. [doi: 10.1002/14651858.CD000072.pub3] [Medline: 28639262]

XSL•FO
RenderX

4. Taylor RR. The Intentional Relationship: Occupational Therapy and Use of Self: F.A. Davis Company; 2020.
5. Minimum standards for the education of occupational therapists. : World Federation of Occupational Therapists; 2026 URL: https://wfot.org/resources/new-minimum-standards-for-the-education-of-occupational-therapists-2026 [accessed 2026-07-03]
6. Occupational therapy education guidelines 2019 and model core curriculum for occupational therapist education [Report in Japanese]. : Japanese Association of Occupational Therapists (JAOT); 2019 URL: https://www.jaot.or.jp/document/dl/0b1dfaee826fa2be25b8b9bcedb3f55e [accessed 2026-07-06]
7. Lee Bunting K, Meyer B, Janzen K, Lieberman G, Willick K, Davis JA. Therapeutic-use-of-self as relational pedagogy in occupational therapy education: a qualitative description study. Can J Occup Ther 2025;92(2):102-112. [doi: 10.1177/00084174241293541] [Medline: 39497421]
8. Mata Á, de Azevedo KPM, Braga LP, et al. Training in communication skills for self-efficacy of health professionals: a systematic review. Hum Resour Health 2021 Mar 6;19(1):30. [doi: 10.1186/s12960-021-00574-3] [Medline: 33676515]
9. Derksen F, Bensing J, Lagro-Janssen A. Effectiveness of empathy in general practice: a systematic review. Br J Gen Pract 2013 Jan;63(606):e76-e84. [doi: 10.3399/bjgp13X660814] [Medline: 23336477]
10. Reeves S, Zwarenstein M, Goldman J, et al. Interprofessional education: effects on professional practice and healthcare outcomes. Cochrane Database Syst Rev 2008 Jan 23(1):CD002213. [doi: 10.1002/14651858.CD002213.pub2] [Medline: 18254002]
11. Grant T, Thomas Y, Gossman P, Berragan L. The use of simulation in occupational therapy education: a scoping review. Aust Occup Ther J 2021 Aug;68(4):345-356. [doi: 10.1111/1440-1630.12726] [Medline: 33751584]
12. Graham F, Kessler D, Nott M, Bernie C, Kanagasabai P, Barthow CA. A scoping review of coaching in occupational therapy: mapping methods, populations and outcomes. Aust Occup Ther J 2024 Dec;71(6):1106-1130. [doi: 10.1111/1440-1630.12991] [Medline: 39315515]
13. Peters MDJ, Godfrey CM, Khalil H, McInerney P, Parker D, Soares CB. Guidance for conducting systematic scoping reviews. Int J Evid Based Healthc 2015 Sep;13(3):141-146. [doi: 10.1097/XEB.0000000000000050] [Medline: 26134548]
14. Aromataris E, Lockwood C, Porritt K, Pilla B, Jordan Z, editors. JBI Manual for Evidence Synthesis: JBI; 2024. URL: https://jbi-global-wiki.refined.site/space/MANUAL [accessed 2026-10-16]
15. Tricco AC, Lillie E, Zarin W, et al. PRISMA extension for scoping reviews (PRISMA-ScR): checklist and explanation. Ann Intern Med 2018 Oct 2;169(7):467-473. [doi: 10.7326/M18-0850] [Medline: 30178033]
16. Rethlefsen ML, Kirtley S, Waffenschmidt S, et al. PRISMA-S: an extension to the PRISMA statement for reporting literature searches in systematic reviews. Syst Rev 2021 Jan 26;10(1):39. [doi: 10.1186/s13643-020-01542-z] [Medline: 33499930]
17. Communication education in occupational therapy: a protocol for a scoping review. OSF. URL: https://osf.io/buxnd/ [accessed 2026-10-16]
18. Rouse DN. Employing Kirkpatrick's evaluation framework to determine the effectiveness of health information management courses and programs. Perspect Health Inf Manag 2011 Apr 1;8(Spring):1c. [Medline: 21464860]
19. Potter AMK, Dwyer M, May M, et al. Changes in the self-efficacy and communication of nursing, occupational therapy, and speech-language pathology students participating in a simulated patient simulation. Clin Simul Nurs 2024 Oct;95:101602. [doi: 10.1016/j.ecns.2024.101602]
20. Ishizawa Y, Kudo C, Kodama Y. Understanding of dialects among first-year nursing and medical technology university students: learning after a class on communication using dialects in medical settings. J Jpn Soc Nurs Res 2024;47(2):263-272. [doi: 10.15065/jjsnr.20231210235]
21. Morrisby C, Bogle J, Dillon R, Reen C, Tanner G. Peer-led virtual learning: impact of dementia-specific communication training for occupational therapy students. Phys Occup Ther Geriatr 2023;41(1):10-30. [doi: 10.1080/02703181.2022.2065403]
22. Mach H, Baylor C, Burns M, Yorkston K. Training students from rehabilitation professions on communicating with patients with communication disorders. PM R 2022 Jan;14(1):58-67. [doi: 10.1002/pmrj.12580] [Medline: 33611858]
23. Lin ML, Harmon TG, Scronce G, Jacks A. Effects of an interdisciplinary communication partner training program on student learning. J Allied Health 2022;51(2):110-115. [Medline: 35640289]
24. Korc-Grodzicki B, Alici Y, Nelson C, et al. Addressing the quality of communication with older cancer patients with cognitive deficits: development of a communication skills training module. Palliat Support Care 2020 Aug;18(4):419-424. [doi: 10.1017/S1478951520000218] [Medline: 32295656]
25. Power E, Falkenberg K, Barnes S, Elbourn E, Attard M, Togher L. A pilot randomized controlled trial comparing online versus face-to-face delivery of an aphasia communication partner training program for student healthcare professionals. Int J Lang Commun Disord 2020 Nov;55(6):852-866. [doi: 10.1111/1460-6984.12556] [Medline: 32654395]
26. Cameron A, McPhail S, Hudson K, Fleming J, Lethlean J, Finch E. Telepractice communication partner training for health professionals: a randomised trial. J Commun Disord 2019;81:105914. [doi: 10.1016/j.jcomdis.2019.105914] [Medline: 31229734]
27. Nixon J, Gray L, Turner J, Bernard A, Scaife J, Cartmill B. Communicating actively responding empathically (CARE): comparison of communication training workshops for health professionals working in cancer care. J Cancer Educ 2020 Feb;35(1):54-60. [doi: 10.1007/s13187-018-1439-0] [Medline: 30386978]

XSL•FO
RenderX

28. Springfield E, Honnery M, Bennett S. Evaluation of a simulation clinic for improving occupational therapy students' perceptions of interaction with parents and infants. Br J Occup Ther 2018;81(1):51-58. [doi: 10.1177/0308022617736504]
29. Cameron A, McPhail SM, Hudson K, Fleming J, Lethlean J, Finch E. A pre–post intervention study investigating the confidence and knowledge of health professionals communicating with people with aphasia in a metropolitan hospital. Aphasiology 2017;31(3):359-374. [doi: 10.1080/02687038.2016.1225277]
30. Haracz K, Arrighi G, Joyce B. Simulated patients in a mental health occupational therapy course: a pilot study. Br J Occup Ther 2015;78(12):757-766. [doi: 10.1177/0308022614562792]
31. Alden J, Toth-Cohen S. Impact of an educational module on occupational therapists' use of elderspeak and attitudes toward older adults. Phys Occup Ther Geriatr 2015 Jan 2;33(1):1-16. [doi: 10.3109/02703181.2014.975884]
32. Cameron A, McPhail SM, Hudson K, Fleming J, Lethlean J, Finch E. Increasing the confidence and knowledge of occupational therapy and physiotherapy students when communicating with people with aphasia: a pre–post intervention study. Speech Lang Hear 2015;18(3):148-155. [doi: 10.1179/2050572814Y.0000000062]
33. Hanover M, Martinez C, Mohan G. Experiential pediatric co-treatment to enhance interprofessional competencies: a mixed-methods study. Indian J Phys Ther Res 2025;7(2):151-156. [doi: 10.4103/ijptr.ijptr_55_25]
34. Keck C, Lankford A, Pfeiffer D, Durfee A, Guilfoyle R. Speech-language pathology and occupational therapy students' perceptions of an interprofessional stroke workshop: a mixed methods analysis. Teach Learn Commun Sci Disord 2025;9(1). [doi: 10.61403/2689-6443.1349]
35. Pfeiffer D, Lankford A, Durfee A, Branson C. A mixed-methods examination of interprofessional preschool virtual simulations for speech-language pathology and occupational therapy graduate students. Teach Learn Commun Sci Disord 2025;9(3). [doi: 10.61403/2689-6443.1395]
36. Brentnall J, Rossiter L, Judd B, et al. Educational design insights for interprofessional immersive simulation to prepare allied health students for clinical placements. Adv Simul (Lond) 2024 Nov 27;9(1):45. [doi: 10.1186/s41077-024-00316-0] [Medline: 39605089]
37. Vanderzwan K, Stephens K, Preissner K, Kilroy S. Simulated interprofessional handoff among nursing and occupational therapy students: determining students' self-efficacy for collaborative practice. Clin Simul Nurs 2024 Dec;97:101637. [doi: 10.1016/j.ecns.2024.101637]
38. Copley J, Martin R, Dix C, et al. Fostering collaborative practice through interprofessional simulation for occupational therapy, physiotherapy, dietetics, and nursing students. J Interprof Care 2024;38(3):534-543. [doi: 10.1080/13561820.2024.2303499] [Medline: 38343271]
39. Kon H, Narita H, Fujita T, Fujita A, Nomura O, Tanaka M, et al. Educational effectiveness of Interprofessional Work workshops for medical students on practical skills and communication [Article in Japanese]. Health Sci Res 2024;15(1):15-21 [FREE Full text]
40. Shorten A, Cruz Walma DA, Bosworth P, et al. Development and implementation of a virtual "collaboratory" to foster interprofessional team-based learning using a novel faculty-student partnership. J Prof Nurs 2023;46:155-162. [doi: 10.1016/j.profnurs.2023.03.008] [Medline: 37188405]
41. Fleming A, Buckley C, Kamal S, et al. Development of a pilot interprofessional education workshop for healthcare students and assessment of interprofessional collaborative competency attainment. J Interprof Care 2023 Nov 2;37(6):954-963. [doi: 10.1080/13561820.2023.2202189] [Medline: 37161380]
42. Curro K, Shooman L, Foo S. The use of interprofessional education (IPE) to address collaboration for individualized education plans (IEPs): a retrospective study of occupational therapy, speech-language pathology, and special education students' perceptions. Teaching Learn Commun Sci Disord 2022;6(2). [doi: 10.30707/TLCSD6.2.1660595992.549661]
43. Björklund K, Silén C. Occupational therapy and physiotherapy students' communicative and collaborative learning in an interprofessional virtual setting. Scand J Occup Ther 2021 May;28(4):264-273. [doi: 10.1080/11038128.2020.1761448] [Medline: 32412813]
44. Nieuwoudt L, Hutchinson A, Nicholson P. Pre-registration nursing and occupational therapy students' experience of interprofessional simulation training designed to develop communication and team-work skills: a mixed methods study. Nurse Educ Pract 2021 May;53:103073. [doi: 10.1016/j.nepr.2021.103073] [Medline: 33991968]
45. Clauser J, Richardson BB, Odom-Maryon T, et al. Standardized patient simulation using SBIRT (screening, brief intervention, and referral for treatment) as a tool for interprofessional learning. MedEdPORTAL 2020 Sep 11;16:10955. [doi: 10.15766/mep_2374-8265.10955] [Medline: 32934979]
46. Sweigart LI, Umoren RA, Scott PJ, et al. Virtual TeamSTEPPS(®) simulations produce teamwork attitude changes among health professions students. J Nurs Educ 2016 Jan;55(1):31-35. [doi: 10.3928/01484834-20151214-08] [Medline: 26812380]
47. Fujimoto S, Ogawa T, Komukai K, Nakayama T. Effect of education on physical and occupational therapists' perceptions of clinical practice guidelines and shared decision making: a randomized controlled trial. J Phys Ther Sci 2022 Jun;34(6):445-453. [doi: 10.1589/jpts.34.445] [Medline: 35698556]
48. Liao F, Murphy D. Evaluation of therapeutic relationship skills training for mental health professionals: the therapeutic relationship enabling programme (TREP). MedEdPublish (2016) 2019;8:112. [doi: 10.15694/mep.2019.000112.1] [Medline: 38089364]

XSL•FO
RenderX

49. Norris M, Eva G, Fortune J, Frater T, Breckon J. Educating undergraduate occupational therapy and physiotherapy students in motivational interviewing: the student perspective. BMC Med Educ 2019 Apr 27;19(1):117. [doi: 10.1186/s12909-019-1560-8] [Medline: 31029114]
50. Fortune J, Breckon J, Norris M, Eva G, Frater T. Motivational interviewing training for physiotherapy and occupational therapy students: effect on confidence, knowledge and skills. Patient Educ Couns 2019 Apr;102(4):694-700. [doi: 10.1016/j.pec.2018.11.014] [Medline: 30482468]
51. Yamagata R, Ohno H, Tsuchiya K, Inoue K. Practice of social skills training for occupational therapy students [Article in Japanese]. Jpn Occup Ther Res 2018;37(3):310-315 [FREE Full text]
52. Schoo AM, Lawn S, Rudnik E, Litt JC. Teaching health science students foundation motivational interviewing skills: use of motivational interviewing treatment integrity and self-reflection to approach transformative learning. BMC Med Educ 2015 Dec 21;15(1):228. [doi: 10.1186/s12909-015-0512-1] [Medline: 26689193]
53. Wener PF, Bergen CO, Diamond-Burchuk LG, Yamamoto CM, Hosegood AE, Staley JD. Enhancing student occupational therapists' client-centred counselling skills. Can J Occup Ther 2015 Dec;82(5):307-315. [doi: 10.1177/0008417415577422] [Medline: 26590230]
54. Hoffmann TC, Bennett S, Tomsett C, Del Mar C. Brief training of student clinicians in shared decision making: a single-blind randomized controlled trial. J Gen Intern Med 2014 Jun;29(6):844-849. [doi: 10.1007/s11606-014-2765-5] [Medline: 24481686]
55. Jeffery B, Hicks C. The impact of counselling skills training on the interpersonal skills of undergraduate occupational therapy students. Br J Occup Ther 1997 Sep;60(9):395-400. [doi: 10.1177/030802269706000904]
56. Chen TT, Huang YJ, Chen CR, Hsu CW, Huang SL, Hsieh CL. Effects of feedback and reflection on communication skills training for occupational therapy students. Am J Occup Ther 2025 Jul 1;79(4):7904205060. [doi: 10.5014/ajot.2025.050793] [Medline: 40493860]
57. Pai HC, Lu YC, Lee S, Lee SY. The effect of comics and interprofessional community dialogue on health humanities, empathy, and communication among healthcare students. Nurse Educ Today 2025 Apr;147:106582. [doi: 10.1016/j.nedt.2025.106582] [Medline: 39837255]
58. Tal-Saban M, Zaguri-Vittenberg S, Weintraub N. Enhancing professional identity of first-year occupational therapy students with the community-academia student tutoring program. Br J Occup Ther 2024 Feb;87(2):106-113. [doi: 10.1177/03080226231198341] [Medline: 40336908]
59. Wang Z, Wu P, Li M, Chen X, Lin C. Perceived effectiveness of the person-environment-occupation innovative teaching model in rehabilitation therapy students. BMC Med Educ 2024 Oct 21;24(1):1180. [doi: 10.1186/s12909-024-06196-2] [Medline: 39434130]
60. Nicola-Richmond K, Richards K, Britt K. The impact of an authentic, simulated learning activity on student preparedness for work-integrated learning. Asia Pac J Coop Educ 2015;16(4):343-354 [FREE Full text]
61. Herge EA, Lorch A, Deangelis T, Vause-Earland T, Mollo K, Zapletal A. The standardized patient encounter: a dynamic educational approach to enhance students' clinical healthcare skills. J Allied Health 2013;42(4):229-235. [Medline: 24326920]
62. Knecht-Sabres LJ. Experiential learning in occupational therapy: can it enhance readiness for clinical practice? J Exp Educ 2013 Mar;36(1):22-36. [doi: 10.1177/1053825913481584]
63. Satake M, Togashi S. The educational effect of small group learning in human science: effectiveness for improving interpersonal and communication skills [Article in Japanese]. J Osaka Kawasaki Rehabil Univ 2012;6:109-114 [FREE Full text]
64. Knecht-Sabres LJ. The use of experiential learning in an occupational therapy program: can it foster skills for clinical practice? Occup Ther Health Care 2010 Oct;24(4):320-334. [doi: 10.3109/07380577.2010.514382] [Medline: 23898958]
65. Katagiri K, Ohori T. Significance of communication training for occupational therapy students through participation in a community-based functional training program and its influence on assessment fieldwork [Article in Japanese]. Hokkaido J Occup Ther 2009;25(3):90-99 [FREE Full text]
66. Taylor CJ, Brussa AK, Barlow KG. Intentional resiliency mentoring: a mixed methods study exploring a pilot mentorship program for occupational therapy students. Open J Occup Ther 2026;14(1):1-15. [doi: 10.15453/2168-6408.2334]
67. Selim C, Tsotsoros J, Brown C, Jeffries L, Ciro C. Impact of medical improvisation on therapeutic use of self: a randomized controlled waitlist study. Am J Occup Ther 2025 Jan 1;79(1):7901205120. [doi: 10.5014/ajot.2025.050785] [Medline: 39705089]
68. Woods L, Lewis-Kipkulei P, Xing K. Incorporating real-time simulated patient and instructor feedback to enhance simulation learning experience for occupational therapy students: a mixed methods study. Open J Occup Ther 2024;12(4):1-13. [doi: 10.15453/2168-6408.2250]
69. McIvor L, Karnes M. Role-play as an effective way to teach relationship building with telehealth. Open J Occup Ther 2019;7(2). [doi: 10.15453/2168-6408.1527]
70. Hussain RA, Carstensen T, Yazdani F, Ellingham B, Bonsaksen T. Short-term changes in occupational therapy students' self-efficacy for therapeutic use of self. Br J Occup Ther 2018 May;81(5):276-284. [doi: 10.1177/0308022617745007]

XSL•FO
RenderX

71. Schwank K, Carstensen T, Yazdani F, Bonsaksen T. The course of self-efficacy for therapeutic use of self in Norwegian occupational therapy students: a 10-month follow-up study. Occup Ther Int 2018;2018:2962747. [doi: 10.1155/2018/2962747] [Medline: 29805333]
72. Gorenberg M. Instructional insights: continuing professional education to enhance therapeutic relationships in occupational therapy. Occup Ther Health Care 2013 Oct;27(4):393-398. [doi: 10.3109/07380577.2013.834404] [Medline: 24050494]
73. Maloney SM, Griffith K. Occupational therapy students' development of therapeutic communication skills during a service-learning experience. Occup Ther Ment Health 2013 Jan;29(1):10-26. [doi: 10.1080/0164212X.2013.760288]
74. Taylor RR, Lee SW, Kielhofner G, Ketkar M. Therapeutic use of self: a nationwide survey of practitioners' attitudes and experiences. Am J Occup Ther 2009;63(2):198-207. [doi: 10.5014/ajot.63.2.198] [Medline: 19432058]
75. Uitterhoeve RJ, Bensing JM, Grol RP, Demulder PHM, VAN Achterberg T. The effect of communication skills training on patient outcomes in cancer care: a systematic review of the literature. Eur J Cancer Care (Engl) 2010 Jul;19(4):442-457. [doi: 10.1111/j.1365-2354.2009.01082.x] [Medline: 20030702]
76. Fischer F, Helmer S, Rogge A, et al. Outcomes and outcome measures used in evaluation of communication training in oncology—a systematic literature review, an expert workshop, and recommendations for future research. BMC Cancer 2019 Aug 14;19(1):808. [doi: 10.1186/s12885-019-6022-5] [Medline: 31412805]
77. van den Eertwegh V, van Dulmen S, van Dalen J, Scherpbier AJJA, van der Vleuten CPM. Learning in context: identifying gaps in research on the transfer of medical communication skills to the clinical workplace. Patient Educ Couns 2013 Feb;90(2):184-192. [doi: 10.1016/j.pec.2012.06.008] [Medline: 22796303]
78. Roter D, Larson S. The Roter interaction analysis system (RIAS): utility and flexibility for analysis of medical interactions. Patient Educ Couns 2002 Apr;46(4):243-251. [doi: 10.1016/s0738-3991(02)00012-5] [Medline: 11932123]
79. Norcini J, Anderson MB, Bollela V, et al. 2018 Consensus framework for good assessment. Med Teach 2018 Nov;40(11):1102-1109. [doi: 10.1080/0142159X.2018.1500016] [Medline: 30299187]
80. Editage [Article in Japanese]. URL: https://www.editage.jp/ [accessed 2026-10-16]

Abbreviations

OATD: Open Access Theses and Dissertations
OT: occupational therapy
PRESS: Peer Review of Electronic Search Strategies
PRISMA-S: Preferred Reporting Items for Systematic Reviews and Meta-Analyses Literature Search Extension
PRISMA-ScR: Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews
WFOT: World Federation of Occupational Therapists

Edited by S Brini; submitted 23.Mar.2026; peer-reviewed by A Torad, J Moreno-Chaparro; revised version received 08.Sep.2026; accepted 09.Sep.2026; published 13.Nov.2026.

Please cite as:
Shimada R, Sano Y, Sasaki T, Tanimura A
Communication Education in Prelicensure and Postlicensure Occupational Therapy: Scoping Review
JMIR Med Educ 2026;12:e95909
URL: https://mededu.jmir.org/2026/1/e95909
doi:10.2196/95909

© Ryuichi Shimada, Yukiko Sano, Takeshi Sasaki, Astuko Tanimura. Originally published in JMIR Medical Education (https://mededu.jmir.org), 13.Nov.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Education, is properly cited. The complete bibliographic information, a link to the original publication on https://mededu.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Generation of Case-Based Materials Using Large Language Models in Health Professions Education: Scoping Review

Zineb Zouakia[1,2,3], MD; Alan Szymczak[1,2,3], MD; Salomé Tordjman[1,2,3], MD; Emmanuel Logak[1,2,3], MD; Thomas Meyre[1,2,3], MD; Matthieu Resche-Rigon[3,4], MD, PhD; Alain Cariou[3,5], MD, PhD; Anita Burgun[1,2,3], MD, PhD; Rosy Tsopra[1,2,3,6], MD, PhD

[1]Clinical Bioinformatics Laboratory, Imagine Institute, Université Paris Cité, INSERM UMR1163, 24 Bd du Montparnasse, 75015, Paris, F-75006, France

[2]Department of Medical Informatics, Hôpital Européen Georges Pompidou, Hôpital Necker Enfants Malades, AP-HP, Paris, France

[3]Faculté de Santé, DigiHealth UPCité Team, Université Paris Cité, Paris, F-75006, France

[4]Biostatistics and Medical Information Department, Saint-Louis Hospital, Assistance Publique Hôpitaux de Paris, Paris, France

[5]Medical Intensive Care Unit, Cochin Hospital, Assistance Publique-Hôpitaux de Paris, Paris, France

[6]UPPERS, Université Paris Cité, US 007, Paris, F-75006, France

Corresponding Author:
Zineb Zouakia, MD
Clinical Bioinformatics Laboratory, Imagine Institute, Université Paris Cité, INSERM UMR1163, 24 Bd du Montparnasse, 75015, Paris, F-75006, France

Abstract

Background: Large language models (LLMs) are increasingly used to generate case-based materials for health professions education (HPE). Previous reviews have examined broad applications in HPE or focused on clinical cases, but none have comprehensively synthesized the generation and evaluation of diverse case-based materials.

Objective: This study aims to synthesize current evidence on LLM-generated case-based materials in HPE, focusing on generation approaches, evaluation methods, key findings, and reported challenges.

Methods: This scoping review followed the Joanna Briggs Institute methodology and PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews) reporting guidelines. Seven databases (PubMed, Embase, Scopus, Web of Science, CINAHL, ERIC, and IEEE Xplore) were searched on May 10, 2026, for studies published from January 1, 2017. Eligible studies were peer-reviewed original research articles, published in English, reporting the use of generative LLMs to create text-based case-based materials for HPE. Data were synthesized using descriptive statistics and narrative synthesis.

Results: Seventy-two studies were included. Most were published between 2024 and 2026 (67/72, 93%) and focused on medical education (59/72, 82%), with undergraduate students as the most frequently studied population. Case-based multiple-choice questions (MCQs; 28/72, 39%) and clinical cases or vignettes (22/72, 30%) were the most frequently generated materials. Generative Pretrained Transformer-4 (GPT-4) family models were the most commonly used (40/72, 55%), and generation predominantly relied on relatively simple prompting approaches, whereas more advanced strategies, such as few-shot prompting, prompt chaining, or use of external knowledge sources, were rarely adopted. Evaluation primarily relied on expert review, typically involving small panels, while student assessment (25/72, 35%) and psychometric analyses were less common. Overall, LLM-generated materials were frequently judged to be clinically plausible, structurally clear, and reasonably aligned with educational objectives, particularly after refinement. However, consistent limitations included factual inaccuracies, construction flaws, superficial clinical realism, pedagogical misalignment, suboptimal psychometric performance, and bias. No LLM consistently outperformed others, although methodological heterogeneity limits cross-study comparisons.

Conclusions: This review provides a comprehensive synthesis of the literature on LLM-generated case-based materials in HPE. Unlike previous reviews examining broad educational applications or focusing on clinical cases, it integrates evidence across a broad range of case-based materials while synthesizing generation methods, evaluation approaches, reported findings, and challenges. The evidence suggests that LLMs can support the development of case-based materials but still require expert oversight because of recurring concerns regarding clinical accuracy, educational quality, psychometric performance, clinical realism, and bias. By identifying methodological and evidence gaps, this review provides educators, curriculum developers, and researchers with a structured overview of this rapidly evolving field and may inform the development of more robust generation workflows, more standardized evaluation approaches, and the responsible use of LLM-generated case-based materials in HPE.

XSL•FO
RenderX

(JMIR Med Educ 2026;12:e94795) doi:10.2196/94795

KEYWORDS

large language models; generative AI; GenAI; health profession education; medical education; case-based learning; digital health; medical informatics

Introduction

Rationale

The emergence of large language models (LLMs) is substantially transforming health education by introducing novel teaching methods and educational practices. Their potential is considerable and extends across multiple levels, including curriculum design, resource generation, personalized learning support, assessment support, and assistance with academic writing and bibliographic management [1-7].

In health professions education (HPE), one key challenge relates to the development of case-based educational materials, which are widely used to promote clinical reasoning, knowledge integration, and decision-making in authentic contexts [8]. For the purpose of this review, case-based materials were defined as text-based educational resources that present realistic clinical situations requiring learners to apply knowledge, interpret information, and make decisions [9]. These materials include clinical vignettes or cases [10], which summarize patient situations to support reasoning and discussion; case-based questions [11], which combine clinical cases with assessment items, such as multiple-choice questions (MCQs), and are commonly used in high-stakes examinations; and simulation scenarios [12], which consist of structured scripts designed for use in simulated environments. Evidence indicates that case-based approaches lead to greater improvements in learner knowledge and performance compared with traditional lectures [13-16]. However, the development of high-quality case-based materials is resource-intensive: cases must be clinically coherent, aligned with learning outcomes, appropriately challenging, and frequently refined through expert review and psychometric testing [17-19]. These requirements can limit the size, diversity, and timely updating of case libraries and assessment banks, particularly where faculty time or specialty expertise is limited.

LLMs offer novel opportunities for the generation of case-based educational materials. Recent reviews have examined the broader use of LLMs in HPE [1-3]. Lin et al [1] synthesized applications across medical education, including educational resources, teaching methods, and assessment, and highlighted both opportunities and implementation challenges. Basil et al [3] mapped the broad integration of generative AI (GenAI) across HPE, identifying applications in curriculum design, content creation and delivery, personalized learning, assessment, automated grading, and feedback. Similarly, Astbury et al [2] reviewed the use of GenAI for teaching and assessment across health professions and emphasized its growing role in learning resource development and educational assessment. These reviews demonstrate the expanding role of LLMs in HPE but address case-based generation only as 1 component of a broader educational landscape. Notably, Basil et al [3] identified content creation as the predominant area of GenAI integration in HPE, highlighting the growing importance of this application. Given the central role of case-based materials in HPE, a more focused and detailed synthesis of their generation using LLMs is therefore warranted.

A more focused scoping review by Kaya et al [20] examined LLM-generated clinical cases and vignettes. However, its scope was limited to cases and vignettes intended to function as standalone cases and explicitly excluded case-based MCQs and other formats combining clinical scenarios with assessment activities. In addition, although prompting and evaluation approaches were reported, their methodological characterization remained relatively broad, focusing mainly on prompt reporting and single versus iterative prompting, while evaluation findings were synthesized primarily in terms of overall case quality and bias. A more comprehensive and granular synthesis encompassing a broader spectrum of case-based educational materials and providing greater detail on their generation and evaluation processes therefore remains needed. The current review addresses this gap by encompassing clinical cases and vignettes, case-based questions, and simulation scenarios, and by providing a detailed characterization of the models and reference materials used, prompting and refinement approaches, evaluation methods and criteria, and reported findings and challenges across multiple quality domains.

Objectives

The objective of this scoping review was to synthesize the current evidence on the use of LLMs to generate case-based materials in HPE. Specifically, the review sought to examine how these models were used to generate case-based educational materials, how the resulting materials were evaluated, and what findings and challenges were identified. To achieve this objective, the review addressed the following questions: (1) How are LLMs being used to generate case-based materials in HPE? (2) How are these LLM-generated case-based materials evaluated? (3) What key findings and challenges are reported in the existing literature?

Methods

Overview

This review was guided by the Joanna Briggs Institute methodology for scoping reviews and reported following the PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews) guidelines [21]. The completed PRISMA-ScR checklist is available in Checklist 1.

Eligibility Criteria

Types of Participants

Studies were eligible if participants (ie, learners, educators, or programs involved) were within HPE.

XSL•FO
RenderX

Concept

The concept of interest was the use of generative LLMs to create case-based educational materials. For consistency, the review focused exclusively on educational resources generated in textual format, such as clinical cases or vignettes, case-based questions, and simulation scenarios. Studies involving the generation of images, audio, or video were therefore excluded. Studies were also excluded if they focused on non-GenAI technologies (eg, rule-based chatbots, expert systems, predictive or classification models without text-generation capabilities), noneducational apps, or educational apps unrelated to case-based material generation.

Context

The context of interest was HPE, including medicine, nursing, dentistry, pharmacy, physician assistant or associate education, physiotherapy, occupational therapy, speech and language therapy, dietetics, nutrition, paramedicine, midwifery, and other allied health disciplines.

Types of Evidence Sources

Full-text, peer-reviewed original research articles were included to capture empirical evidence on the use of generative LLMs to create case-based educational materials in HPE. Conference papers were also included when sufficient methodological detail was provided. Studies published in English between January 1, 2017, and May 10, 2026, were eligible for inclusion. The start date was selected to capture literature published following the introduction of transformer-based architectures that underpin contemporary LLMs. Only studies available in English were included. The following publication types were excluded: review articles, study protocols, preprints, conference abstracts, commentaries, editorials, letters, book chapters, and nonscholarly sources (eg, blogs and media articles).

Information Sources

A comprehensive literature search was conducted in 7 electronic bibliographic databases: PubMed, Embase, Scopus, Web of Science, CINAHL, ERIC, and IEEE Xplore. Searches were conducted on May 10, 2026, and covered studies published from January 1, 2017, onwards. All retrieved records were imported into Rayyan (Rayyan Systems Inc), a web-based systematic review management tool, for screening and study selection [22]. Duplicate records were identified using Rayyan's automated duplicate detection function and subsequently verified through manual review by 1 author (ZZ).

Search

The search strategy was initially developed by reviewers EL and ZZ and subsequently refined by the review team. It combined controlled vocabulary terms adapted to the database (eg, MeSH and Emtree terms) with free-text keywords, related to 3 concepts: (1) GenAI and LLMs, (2) case-based educational materials, and (3) HPE. The search strategy was iteratively refined in PubMed, and search terms and syntax were subsequently adapted to the indexing and search functionalities of each database. An initial search was conducted on September 8, 2025, and updated on May 10, 2026, following expansion of the search strategy and database coverage. Corresponding authors were contacted by authors (ZZ and ST) when full-text articles could not be obtained through institutional access. The complete database-specific search strategies are provided in Multimedia Appendix 1. Reporting of the search strategy followed the PRISMA-S (Preferred Reporting Items for Systematic Reviews and Meta-Analyses Literature Search Extension) guidelines [23] (Checklist 2).

Selection of Sources of Evidence

Study selection was conducted using Rayyan in 2 stages: (1) title and abstract screening and (2) full-text review. At each stage, records were assessed independently by 2 reviewers (ZZ, AS, or ST) against the predefined eligibility criteria. Discrepancies were resolved through discussion and consensus. A third reviewer (RT) was consulted in 2 cases where consensus could not be reached.

Data Charting Process

A structured data charting form was developed in Microsoft Excel to extract relevant information from the included studies. The form was piloted through independent data extraction by 2 reviewers (ZZ and AS) on 3 randomly selected studies and subsequently refined to improve clarity, consistency, and completeness [24]. Data charting was then performed by 1 reviewer (ZZ, AS, or ST) and independently verified by a second reviewer (ZZ, AS, or ST) [24]. Any discrepancies were resolved through discussion and consensus.

Data Items

Data charted from the included studies [25-96] comprised bibliographic characteristics (authors, title, year, journal, and country); educational context (health profession and learner level); study characteristics (study type and objectives); characteristics of the LLM application (task performed by the LLM, model used, input data or reference materials, prompting strategy, prompt availability, and human oversight); characteristics of the generated case-based materials (type of material and sample size); evaluation approaches (evaluation methods and criteria); and study outcomes, including key findings and reported challenges or limitations. The complete data charting form is provided in Multimedia Appendix 2.

Synthesis of Results

Charted data were synthesized using descriptive statistics and narrative synthesis. Quantitative data were summarized using frequencies and percentages where appropriate. Findings were organized to address the review objectives and research questions, including (1) characteristics of the included studies [25-96], (2) technical and methodological approaches to LLM-based case generation, (3) evaluation methods and criteria used to assess generated materials, and (4) key findings and reported challenges. Results are presented using tables, narrative summaries, and visual evidence-mapping approaches, including a geographic distribution map and an evidence gap map, to facilitate comparison across studies and identify patterns and gaps in the literature.

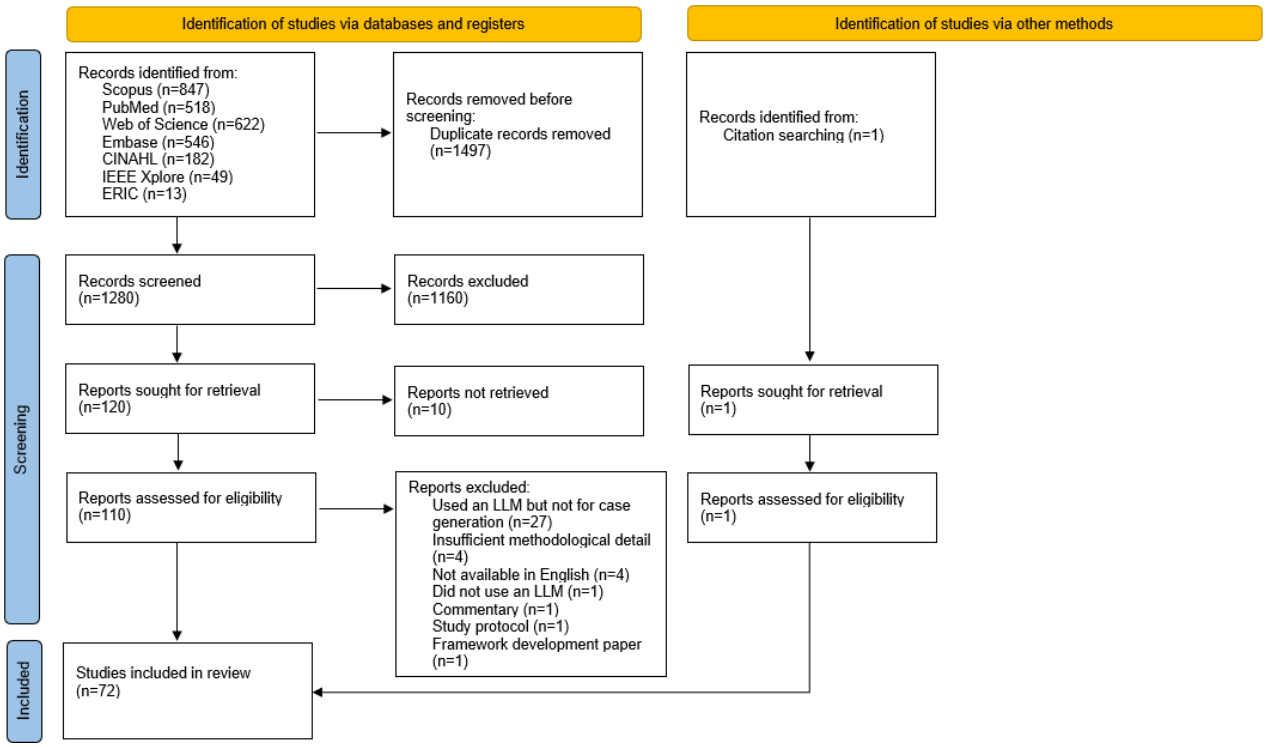
XSL•FO
RenderX

Results

Selection of Sources of Evidence

A total of 2777 records were identified through 7 databases and citation tracking. After removing 1497 duplicate records, 1280 unique records were screened by title and abstract, with 120 retained for full-text review. Of these, 110 were successfully retrieved and assessed for eligibility, leading to the exclusion of 39 full-text articles. Reasons for exclusion included the absence of LLM-based case generation (n=27), insufficient methodological detail (n=4), non-English publications (n=4), lack of LLM use (n=1), commentary articles (n=1), study protocols (n=1), and framework development papers (n=1). One additional study was identified through citation searching and met the inclusion criteria. Ultimately, a total of 72 studies [25-96] were included in the final synthesis (Figure 1). Detailed charted information for all included studies is provided in Multimedia Appendix 2.

Figure 1. PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) flow diagram showing the identification, screening, and inclusion of studies using LLMs to generate case-based materials for HPE. HPE: health professions education; LLM: large language model.



Characteristics of Sources of Evidence

Most studies were published in 2024 - 2025 (53/72, 74%), although an additional 14 studies [25-38] had already been published in 2026 by the beginning of May, highlighting the continued expansion of this field. The included studies most frequently originated from the United States (19/72, 26%) [27,35,44,48,53-55,57,58,60,62,63,74,81,83,87,88,90,95] and Türkiye (10/72, 14%) [28,38,42,47,49,65,72,73,76,93] (Table 1). Geographically, the evidence was concentrated in Asia (35/72, 49%) [26,28,30,31,33,36,38,41,42,45-47,49,50,52,56,65,67-73,76,80,82,84-86,89,91,93,94], and North America (27/72, 37%) [27,35,37,40,44,48,51,53-55,57,58,60-63,66,74,75,78,81,83,87,88,90,95,96], with fewer studies originating from Europe (10/72, 14%) [25,29,31,34,39,43,64,77,79,92], Africa (3/72, 4%) [31,32,43], and South America (1/72, 1%) [59], while no studies were identified from Oceania (Figure 2).

XSL•FO
RenderX

Table . Characteristics of included studies [25-96] examining the use of LLMs[a] for the generation of case-based materials for HPE[b] (N=72).

Characteristics and categories	Studies, n (%)	Article references
Year of publication		
2026	14 (19)	[25-38]
2025	32 (44)	[39-70]
2024	21 (29)	[71-91]
2023	5 (7)	[92-96]
Country		
United States	19 (26)	[27,35,44,48,53-55,57,58,60,62,63,74,81,83,87,88,90,95]
Türkiye	10 (14)	[28,38,42,47,49,65,72,73,76,93]
Canada	5 (7)	[37,61,66,75,96]
India	5 (7)	[26,45,82,84,94]
China	4 (5)	[30,50,52,69]
Kingdom of Bahrain	4 (5)	[31,70,85,91]
Saudi Arabia	4 (5)	[31,41,56,80]
United Kingdom	4 (5)	[31,39,43,77]
Egypt	3 (4)	[31,32,43]
Japan	3 (4)	[67,86,89]
Mexico	2 (3)	[51,78]
Republic of Korea	2 (3)	[33,68]
Spain	2 (3)	[29,34]
Other: Brazil, France, Greece, Grenada, Iran, Switzerland, Italy, Thailand, the Netherlands, United Arab Emirates	9 (12)	[25,36,40,46,59,64,71,79,92]
Study type		
Descriptive or explorative study	10 (14)	[34,59,78,82,84,88,92-95]
Evaluation study (no comparison)	30 (42)	[25,26,31,33,35,36,40,43,44,46,51,53-55,57,58,61,62,66,67,71,74,76,79,81,83,86,87,89,90]
Comparison study: LLM [a]versus Expert	15 (21)	[30,37,41,45,48,49,52,63,64,66,67,72,75,77,96]
Comparison study: LLM versus LLM	13 (18)	[27-29,38,39,42,50,56,60,70,80,85,91]
Comparison study: LLM versus LLM or other method versus Expert	4 (5)	[32,47,65,73]
Health profession		
Medicine	59 (82)	[25,28-31,34,36-45,47-54,56-58,61,63-66,69-86,88-96]
Nursing	7 (10)	[33,35,46,60,62,67,87]
Pharmacy	2 (3)	[32,55]
Other: dentistry, interprofessional health care, occupational therapy, physiotherapy, respiratory therapy, and paramedicine	5 (9)	[26,27,59,68,96]
Student level		
Undergraduate	42 (58)	[25,26,28,30,32,34,36,40,42,44,45,47,51,53,55,56,58,60,63,65,67,69,70,72,73,75,79,81,83,85,87,91,95]
Postgraduate	8 (11)	[44,48,63,66,70,79,88,89]
Continuing education	1 (1)	[71]
Not specified	22 (30)	[27,31,33-35,43,46,50,52,57,59,64,68,74,80,82,86,90,92-94,96]
Type of case-based materials		
Case-based MCQ[c]	28 (39)	[29,36,39,42,44,45,47-49,55,56,58,64-66,69,70,73,74,76,79-85,93]

XSL•FO
RenderX

Characteristics and categories	Studies, n (%)	Article references
Clinical vignette or case	22 (30)	[26,30,31,37,40,41,50,53,57,59,67,68,70-72,78,86,89-92,94]
Simulation scenario	18 (25)	[27,28,32-35,46,51,52,54,60,62,63,77,87,88,95,96]
SCT[d] or variant	4 (5)	[38,43,61,75]
OSCE[e], OSPE[f]	3 (4)	[25,70,85]
Short answer question	2 (3)	[70,85]

[a]LLMs: large language models.

[b]HPE: health professions education.

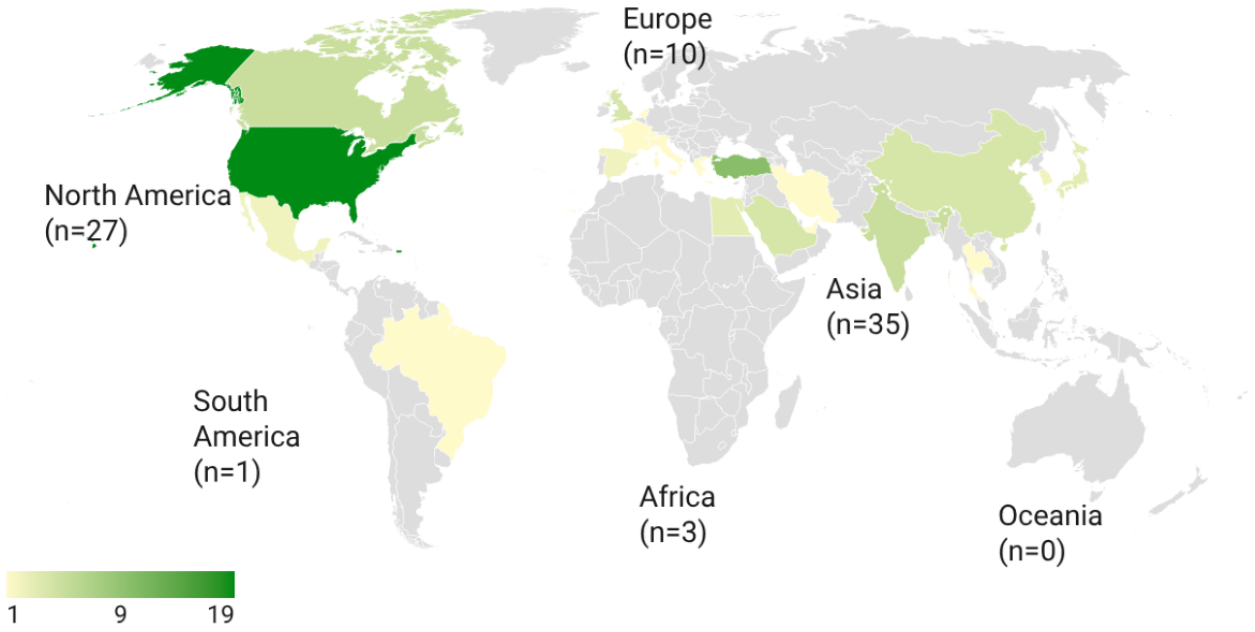
[c]MCQ: multiple-choice question.

[d]SCT: Script Concordance Test.

[e]OSCE: Objective Structured Clinical Examination.

[f]OSPE: Objective Structured Practical Examination.

Figure 2. Geographic distribution of studies using LLM to generate case-based educational materials in health professions education by continent. Studies involving multiple countries were counted in each relevant continent; therefore, continent totals exceed the total number of included studies.



The most frequent study design was LLM evaluation studies without comparison groups (30/72, 42%) [25,26,31,33,35,36,40,43,44,46,51,53-55,57,58,61,62,68,69,71,74,76,79,81,83,86,87,89,90]. Comparative approaches were also common and included comparisons between LLM-generated and expert-generated materials (15/72, 21%) [30,37,41,45,48,49,52,63,64,66,67,72,75,77,96], between different LLMs (13/72, 18%) [27-29,38,39,42,50,56,60,70,80,85,91], or between LLMs or other generation methods and expert-generated materials (4/72, 5%) [32,47,65,73]. Descriptive or exploratory studies accounted for 10/72 (14%) [34,59,78,82,84,88,92-95] of the included publications (Table 1).

Case-based MCQs were the most frequently generated materials (28/72, 39%) [29,36,39,42,44,45,47-49,55,56,58,64-66,69,70,73,74,76,79-85,93], followed by clinical cases or vignettes (22/72, 30%) [26,30,31,37,40,41,50,53,57,59,67,68,70-72,78,86,89-92,94] and simulation scenarios (18/72, 25%) [27,28,32-35,46,51,52,54,60,62,63,77,87,88,95,96]. Medical education was the predominant focus (59/72, 82%) [25,28-31,34,36-45,47-54,56-58,61,63-66,69-86,88-96], and undergraduate students were the most commonly targeted population (42/72, 58%) [25,26,28-30,32,36-42,45,47,49,51,53-56,58,60-62,65,67,69,70,72,73,75-79,81,83-85,87,91,95], although student level was not specified in nearly one-third of studies (Table 1). The evidence gap map (Figure 3) highlighted substantial variation across health professions. Medical education studies covered all material types, particularly case-based questions, whereas nursing studies (7/72, 10%) [33,35,46,60,62,67,87] were concentrated on simulation scenarios. Pharmacy studies (2/72, 3%) [32,55] generated case-based questions and simulation scenarios, while studies involving other health professions (5/72, 7%) [26,27,59,68,96] predominantly generated clinical cases or vignettes. Few studies

XSL•FO
RenderX

targeted postgraduate or continuing education learners, revealing important gaps in the current evidence base.

Figure 3. Evidence gap map of studies using large language models to generate case-based educational materials in HPE. Bubble plot summarizing the distribution of included studies [25-96] according to the type of generated case-based material and target health profession. Generated materials were grouped into 3 categories: case-based questions, including case-based MCQs, short-answer questions, script concordance tests (SCTs), and variants; cases or vignettes; and simulation scenarios, including OSCE and OSPE scenarios. Bubble size is proportional to the number of included studies within each category, while bubble color indicates student level. HPE: health professions education; MCQs: multiple-choice questions; OSCE: Objective Structured Clinical Examination; OSPE: Objective Structured Practical Examination.



Synthesis of Results

Use of LLMs for Case-Based Material Generation

Generative Pretrained Transformer-4 (GPT-4) (OpenAI) family models were the most frequently used (40/72, 55%) [25,26,28,29,32-34,36,37,39,42,43,45-49,51-56,60,61,63-66,69-71,73,77,80,81,84,86,89,90], followed by GPT-3.5 family models (21/72, 29%) [29,39,40,44,57,65,67,68,72,74-76,78-80,83,85,91,92,95,96]. Other commonly used models included Bard and Gemini (Google AI) (13/72, 18%) [27,28,32,38,39,41,42,47,56,58,65,80,91], GPT models with unspecified versions (9/72, 12%) [27,31,35,59,62,82,87,88,94], Claude (Anthropic) (7/72, 10%) [27,28,32,47,56,60,85], and DeepSeek (5/72, 7%) [28,30,32,56,70], while several other models were reported only occasionally.

Although several studies cited guidelines to follow, only a few provided these resources directly to the LLM (7/72, 10%) [27,50,56,65,80,85,91], and most studies relied on prompt-only generation (59/72, 82%) [26-32,34,35,37-40,42,44-46,49-54,56,58,60-66,68,69,71-88,90-96]. The majority of studies published their prompts in full or in part (68/72, 94%) [25-56,58-66,68-72,74-93,95,96], and all studies used basic instructional and contextual prompting. Among more advanced prompting techniques, output formatting or template-based prompting was the most common (57/72, 79%) [25,26,28-38,42-63,65,68-73,75-77,79-81,83,85,86,89,90,92,93,95,96], whereas role prompting (22/72, 30%) [25,31-33,38,39,42,43,47,51,52,59-61,66,69-71,75,77,86,89], prompt chaining (15/72, 21%) [25,33,40,41,48,52,57,59,60,62,70,74,86,92,96], and few-shot prompting (6/72, 8%) [25,33,47,48,59,89] were used less frequently. Refinement strategies were reported in several studies, most commonly prompt refinement (25/72, 35%) [27,29,32,33,36,37,40,43,45,58,60,62,63,66,68,69,72,78,79,83,85,88,92,95,96], followed by output refinement by experts or authors (16/72, 22%) [26,35-37,44-47,49,55,57,62,63,65,77,96] and LLM self-refinement (5/72, 7%) [33,52,57,72,86]. The quantity of generated materials varied. Most studies generated ≤10 outputs (29/72, 40%) [26,30,33,35,41,43,46,49,51,58,59,62,64,68,71,74-78,82-84,87,88,93-96] or between 11 and 100 outputs (25/72, 35%) [25,27,28,36-38,44,45,47,48,52,53,55,57,61,63,65,66,69,72,73,79,86,91,92], whereas 14 studies (19%) [29,32,34,39,40,42,50,54,56,80,81,85,89,90] generated more than 100 outputs (Table 2).

XSL•FO
RenderX

Table . LLM[a] generation approaches and technical characteristics for case-based material generation in health professions education (N=72).

Characteristics and categories	Studies, n (%)	Article references
Model used		
• GPT-4[b] family (OpenAI)	40 (55)	[25,26,28,29,32-34,36,37,39,42,43,45-47,51-53,55,56,61,63,66,69-71,73,77,78,81,84,86,89,90]
• GPT-3.5 family (OpenAI)	21 (29)	[29,39,40,44,57,65,67,68,72,74-76,78-80,83,85,91,92,95,96]
• Bard • Gemini (Google AI)	13 (18)	[27,28,32,38,39,41,42,47,56,58,65,80,91]
• GPT unspecified version (OpenAI)	9 (12)	[27,31,35,59,62,82,87,88,94]
• Claude (Anthropic)	7 (10)	[27,28,32,47,56,60,85]
• DeepSeek (DeepSeek AI)	5 (7)	[28,30,32,56,70]
• Bing • Copilot (Microsoft)	4 (5)	[27,39,60,80]
• GPT-5 family (OpenAI)	2 (3)	[36,38]
• Other (models used in ≤2 articles: Baichuan, Ernie, GPT o1 and o3, Grok, Mixtral, PaLM, Poe, Qwen)	7 (10)	[27,50,56,65,80,85,91]
Input data or reference material		
• None (prompt-only)	59 (82)	[26-32,34,35,37-40,42,44-46,49-54,56,58,60-66,68,69,71-88,90-96]
• Examples or prior cases	6 (8)	[25,33,41,47,59,89]
• Textbook chapters or guidelines	7 (10)	[25,36,43,48,55,57,70]
Prompt publication		
• Yes	62 (86)	[25-40,42-56,58,60-63,68,70,72,74-93,95,96]
• Partly (extracts)	6 (8)	[41,59,64,65,69,71]
• No	4 (5)	[57,67,73,94]
Prompting techniques		
Basic		
• Instructional	71 (99)	[25-66,68-96]
• Contextual	71 (99)	[25-66,68-96]
• Output format • Template	57 (79)	[25,26,28-38,42-63,65,68-73,75-77,79-81,83,85,86,89,90,92,93,95,96]
Advanced		
• Role	22 (30)	[25,31-33,38,39,42,43,47,51,52,59-61,66,69-71,75,77,86,89]
• Prompt chaining	15 (21)	[25,33,40,41,48,52,57,59,60,62,70,74,86,92,96]
• One-shot • Few-shot	6 (8)	[25,33,47,48,59,89]
Refinement process		
• Prompt refinement	25 (35)	[27,29,32,33,36,37,40,43-45,58,60,62,63,66,68,69,72,78,79,83,85,88,92,95,96]
• Output refinement	16 (22)	[26,35-37,44-47,49,55,57,62,63,65,77,96]

XSL•FO
RenderX

Characteristics and categories	Studies, n (%)	Article references
• Self-refinement (by the LLM)[a]	5 (7)	[33,52,57,72,86]
Sample size (materials generated)		
• ≤10 outputs	29 (45)	[26,30,33,35,41,43,46,49,51,58,59,62,64,68,71,74,78,82,84,87,88,93,96]
• 11 - 100 outputs	25 (35)	[25,27,28,36-38,44,45,47,48,52,53,55,57,61,63,65,66,69,72,73,79,86,91,92]
• >100 outputs	14 (19)	[29,32,34,39,40,42,50,54,56,80,81,85,89,90]
• Not specified	4 (5)	[31,60,67,70]

[a]LLM: large language model.

[b]GPT-4: Generative Pretrained Transformer-4.

Evaluation of LLM-Generated Case-Based Materials

Most studies used expert review as the primary evaluation method (56/72, 78%) [25-33,35-39,41-49,51-58,60,62,63,65-73,75-77,79-81,83,85-87,89,91,96], while fewer incorporated student assessment (25/72, 35%). Statistical analyses for bias testing were uncommon (4/72, 5%), and 10/72 studies (14%) were exploratory without a formal evaluation.

Regarding the volume of evaluated materials, 30% (22/72) of studies assessed ≤10 outputs, 35% (25/72) [26,28,30,32,36,42,44,45,47-49,51,52,55,61-67,72,76,77,79] evaluated 11 - 100 outputs, and 15% (11/72) [29,32,34,40,50,54,56,80,85,89,90] evaluated >100 outputs, while 7% (5/72) [31,57,60,67,70] did not report the number of materials evaluated. Among studies using expert review, most involved 10 or fewer experts (42/72, 58%) [25,27,29,31,33-35,37-41,43,46,50,53,54,56-60,68-71,73-75,78,80-96], whereas only 7/72 (10%) [63,71,73,77,86,87,96] engaged 11 - 100 experts and 1/72 (1%) [75] involved more than 100 experts; 22% (16/72) [34,40,50,59,61,64,74,78,82,84,88,90,92-95] did not include expert evaluation and 8% (6/72) [30,52,57,62,72,76] did not specify the number of experts involved. Student assessment most commonly involved 11 - 100 students (16/72, 22%) [26,28,36,42,44,48,51,55,61,63-66,72,76,79], while only a few studies included larger student samples (Table 3).

XSL•FO
RenderX

Table . Evaluation methods of LLM[a]-generated case-based materials for health profession education (N=72).

Category	Studies, n (%)	Article references
Evaluation method		
Expert review	56 (78)	[25-33,35-39,41-49,51-58,60,62,63,65-73,75-77,79-81,83,85-87,89,91,96]
Student assessment	25 (35)	[26,28,30,32,36,42,44,45,47-49,51,52,55,61-67,72,76,77,79]
Statistical analysis	4 (5)	[34,40,50,90]
No formal evaluation (exploratory or descriptive)	10 (14)	[59,74,78,82,84,88,92-95]
Sample size (materials evaluated)		
≤10 outputs	22 (30)	[26,30,33,35,41,43,46,49,51,58,59,62,64,68,71,73,75-77,83,87,96]
11 - 100 outputs	25 (35)	[25,27,28,36-39,42,44,45,47,48,52,53,55,61,63,65,66,69,72,79,81,86,91]
>100 outputs	11 (15)	[29,32,34,40,50,54,56,80,85,89,90]
Not specified	5 (7)	[31,57,60,67,70]
Number of experts involved in evaluation		
≤10 experts	42 (58)	[25-29,31-33,35-39,41-49,51,53-56,58,60,65-70,79-81,83,85,89,91]
11 - 100 experts	7 (10)	[63,71,73,77,86,87,96]
>100 experts	1 (1)	[75]
Not specified	6 (8)	[30,52,57,62,72,76]
No expert evaluation	16 (22)	[34,40,50,59,61,64,74,78,82,84,88,90,92-95]
Number of students involved in evaluation		
≤10 students	1 (1)	[67]
11 - 100 students	16 (22)	[26,28,36,42,44,48,51,55,61,63-66,72,76,79]
101 - 300 students	5 (7)	[30,32,45,47,77]
>300	1 (1)	[49]
Not specified	2 (3)	[52,62]
No student assessment	47 (72)	[25,27,29,31,33-35,37-41,43,46,50,53,54,56-60,68-71,73-75,78,80-96]

[a]LLM: large language model.

Across the 62 studies [25-58,60-73,75-77,79-81,83,85-87,89-91,96] that conducted an evaluation, metrics clustered into a few major groups, but their operationalization varied widely. In expert-based evaluations, the most frequently assessed dimensions were clinical accuracy (38/62, 61%) [25-29,32,36,37,39,41-46,49,51,53,54,56-58,62,69-71,73,76,80,81,83,85-87,89,91,96], educational alignment (31/62, 50%) [25-28,30-32,35-39,41-46,51,52,55,56,60,68,70,76,79-81,85,86], and clinical realism (27/62, 44%) [25,27,28,31-33,41-46,50,52,60,62,68-71,75,77,79,86,87], with notable variation in how each was defined (eg, free-text qualitative judgments vs structured Likert ratings). In student-based evaluations, psychometric analyses were the most common, particularly discrimination indices (11/62, 18%) [36,42,45,47-49,55,66,76,77,79] and difficulty indices (10/62, 16%) [36,42,45,48,49,55,65,76,77,79]. Student perceptions were also frequently evaluated through Likert-scale surveys (11/62, 18%) [26,28,30,32,51,52,61,63,64,67,72], whereas measures of student performance, such as correct response rates or pre-post comparisons, were less common. Other evaluation approaches included assessments of bias (8/62, 13%) [27,29,40,50,53,60,70,90], dataset validity (1/62, 2%) [34], and practical considerations such as time, cost, consistency, and workload associated with case generation (6/62, 10%) [30,39,47,67,77,96] (Table 4).

XSL•FO
RenderX

Table . Evaluation criteria of LLM[a]-generated case-based materials for health profession education, in studies reporting an evaluation (N=62)[b].

Criteria	Studies, n (%)	Articles and examples
Expert review		
Qualitative (Independent expert review, group consensus, and focus group discussion)		
Clinical accuracy	11 (18)	[42,44-46,51,62,66,86,87,91,96] [42]: "technical accuracy"; [44]: "clinical appropriateness"; [46]: "clinical accuracy," "alignment with current medical guidelines and best practices";[86,87,96]: free-text responses noting the inaccuracies
Educational alignment	8 (13)	[30,42,44-46,51,60,68] [46]: "educational effectiveness, relevance to nursing competencies, learning outcomes"
Clinical realism	7 (11)	[42,44-46,60,62,68] [42]:"clinical relevance"; [44]: "ethical realism, and contextual relevance"; [46]: "scenario realism"
Completeness, sufficiency of information	4 (6)	[60,66,68,91]
Bias	2 (3)	[27]: "allow evaluators to identify content perceived as biased or stereotypical of the patient populations"; [60]
Clarity, coherence, and language quality	1 (2)	[91]: "vague instructions for drug administration," "use of medical terms less likely understood by patient"
Format, construction quality	1 (2)	[66]: "quality of the stem construction, plausibility of the distractors"
Open-ended feedback	18 (29)	[25-28,31,35-37,43,45,46,55,60,69,75,86,87,96] [28]: "what do you think about using such scenarios in clinical education?"; [75]: "advantages and drawbacks of the presented SCTs[c]"; [87]: "what surprised you about the information ChatGPT presented?"
Quantitative (Likert scales, binary (yes, no))		
Clinical accuracy	28 (45)	[25-29,32,36,37,39,41,43,49,53,54,56-58,69-71,73,76,80,81,83,85,86,89]
Educational alignment	26 (42)	[25-28,31,32,35-39,41,43,45,52,55,56,68-70,76,79-81,85,86]
Format, construction quality	23 (37)	[25,28,29,31-33,35,37-39,49,56,58,69,70,73,75,76,79-81,83,85]
Completeness, sufficiency of information	22 (35)	[25,28,29,32,33,35,36,38,41,50,53,57,58,68-70,73,76,79,85,86,89]
Clinical realism	21 (34)	[25,27,28,31-33,37,38,41,43,50,52,68-71,75,77,79,86,87]
Clarity, coherence, language quality	21 (34)	[25,28,32,33,36,37,39,41,43,45,49,58,69-71,73,75,76,79,86,89]
Difficulty, challenge appropriateness	19 (31)	[28,29,37,38,41,52,56,58,63,69-71,73,75,77,80,81,83,86]
Engagement, reflection	11 (18)	[25,28,36,37,41,52,58,68-71]
Usability, exam readiness	10 (16)	[25,27,28,31,36,38,47,50,55,89]
Content validity	3 (5)	[33,39]: "Item Content Validity Index" [55]; "Content Validity Ratio"
Blinded attribution (human vs LLM)	2 (3)	[63,75]
Feasibility	2 (3)	[25,37]
Originality	2 (3)	[25,28]
Bias	2 (3)	[53]: "possibility of demographic bias"; [70]: "cultural and contextual sensitivity; diversity and representation; ethical and professional considerations"
Overall quality or rating	5 (8)	[25,26,28,53,96]

XSL•FO
RenderX

Criteria	Studies, n (%)	Articles and examples
Other	4 (6)	[31]: "human oversight," "importance of local adaptation"; [71]: "which group the cases were most suitable for: medical students, medical residents, health professionals, or all of the above"; [56,80]: "variety (does it repeat previous responses?)"
Student assessment		
Psychometric analysis		
Discrimination index	11 (18)	[36,42,45,47-49,55,66,76,77,79]
Difficulty index	10 (16)	[36,42,45,48,49,55,65,76,77,79]
Distractor effectiveness	4 (6)	[36,45,47,76]
Reliability	4 (6)	[42,79]: Cronbach α; [45,55]: KR-20 reliability coefficient
Other	1 (2)	[55]: "passing standard (Angoff method)," item response theory and Rasch analysis"
Student performance		
Performance tracking during intervention or pre-post intervention	3 (5)	[42,44,67]
Correct response rate	4 (6)	[47,48,55,66]
Group-level performance comparison (between student levels or specialties)	3 (5)	[42,48,79]
Student evaluation		
Likert scale survey	11 (18)	[26,28,30,32,51,52,61,63,64,67,72] [51]: "scenario construction (clarity, realism of the case), development or execution (interactivity and simulator responses), relevance (usefulness for clinical training, appropriate level of challenge)"; [52]: "student engagement"; [61]: "perceived difficulty"; [72]: "clarity, spelling, contradictions, enjoyment, clinical reasoning, keyword generation, research question suitability, overall liking, EBM learning support, need for correction"
Blinded attribution (Human vs LLM)	5 (8)	[28,48,52,61,63]
Other		
__[d]	6 (10)	[30,47,67,77,96]: time to produce cases; [39]: "consistency between chatbot outputs across the 3 researchers (exact, similar or different)," "assessment of bias or preference in the correct answer allocation"; [67]: "perceived burden of case creation"; [77]: "financial cost of drafting GPT-assisted vs standard cases"
Statistical analysis		
Bias analysis		

XSL•FO
RenderX

Criteria	Studies, n (%)	Articles and examples
—	5 (8)	[27]: descriptive statistics of sociodemographic data (gender, race, ethnicity, English proficiency, sexual orientation, marital status, employment, household income, source of payment for health services, and living situation); [29]: χ^2 tests to compare demographic and clinical content features across MCQs; [40,50,90]: epidemiological concordance (comparison of distributions against actual prevalence data); [40]: "compare changes in ethnoracial representation between March and May 2024 and differences between initial prompting and self-prompting"; [90]: "impact of geographic setting or using debiasing strategies"
Dataset validity		
—	1 (2)	[34]: distribution and relationships between variables of age, gender, occupation, reason for consultation, medical history, vital signs and diagnosis

[a]LLM: large language model.

[b]Studies could evaluate the same criterion using both qualitative and quantitative approaches. Consequently, a study may appear in more than 1 row for a given evaluation criterion. Counts reported in the narrative synthesis represent unique studies and therefore may differ from the sum of qualitative and quantitative counts shown in this table.

[c]SCTs: Script Concordance Tests.

[d]Not available.

Key Findings and Challenges Identified in the Included Studies

Clinical Accuracy

LLM-generated materials were often assessed as clinically accurate by experts [25,28,31,39,42,43,46,53,56,57,68,73,83,85,86,89], particularly for common conditions [54] and structured scenarios [25,43]. For example, Rao et al [53] reported a mean scientific accuracy score of 4.45/5, Saad et al [83] found that 80% of items met medical accuracy criteria, and Ananthanarayanan [57] reported medical accuracy scores ranging from 9.59 to 10.0 out of 10. At the same time, inaccuracies and missing or contradictory clinical information were frequently reported, including hallucinations [32,69], diagnostic inconsistencies and inappropriate investigations [46,63,69,86,87,89], pharmacotherapeutic errors and inaccurate treatment recommendations [32,46,58,62,76,83,89,91,92], and incomplete patient histories or missing clinical details [25-27,29,33,35,46,60,63,69,74,79,86,87,89,92,96]. These findings show that while LLMs can generate clinically acceptable drafts, expert validation and revision are still required to ensure accuracy.

Structural Quality

Structural quality, including clarity, language quality, coherence, and adherence to construction rules, was frequently examined. Several studies reported clear wording in LLM-generated materials [25,28,32,36,37,39,41,43,73,75,76]. Language quality and medical terminology accuracy in non-English outputs varied, with Turkish-language scenarios praised for their linguistic precision and understandability [28], whereas Japanese-language vignettes exhibited occasional issues with terminology use, grammatical correctness, and language consistency, including sporadic insertion of nontarget languages [86,89]. Reviewers also highlighted concise case descriptions and vignettes that included the information necessary to support learning [36,38,69,71]. However, some studies identified inconsistencies between case narratives and associated questions [69], excessive length despite instructions to be concise [56], unnecessary detail [55], and occasional coherence problems in medical contexts [28]. Recurring construction problems further limited direct reuse, including positional answer bias and variable item balance [39], multiple, ambiguous, or incorrect answers [32,36,48,49,66,79,85], repetitive or shallow scenarios [80], and pseudovignette construction [37,58,83].

Educational Alignment and Pedagogical Quality

Educational alignment was frequently assessed across studies. Many studies showed that LLM-generated materials could appropriately target intended competency levels and learning objectives [25-28,30-32,36-38,41,45,46,52,55,68-70,72,76,85,86]. For example, Arain et al [41] reported higher alignment with learning objectives for LLM-generated vignettes than for faculty-authored versions (mean 3.0 vs 1.25), and Takahashi et al [86] reported a mean "educational usefulness" score of 3.55/5 across 18 vignettes. However, challenges included inconsistent alignment across training stages [85], the need for adaptation to local curricula [31,35,76], some lack of alignment with course objectives [26,45], and pedagogical limitations such as the presence of leading formulations or cues [25,37,41,58,66,69,70,79], as well as a tendency toward low-complexity, recall-level items, insufficiently challenging students [37,41,55,58,66,69,83,85]. Iterative refinement approaches were associated with improved educational alignment comparable to expert-authored scenarios [52].

XSL•FO
RenderX

Clinical Realism

Clinical realism was variably achieved. Several studies found that LLM-generated materials were generally realistic, with experts and students rating them as plausible [28,31,38,46,51,68,69,77], although realism appeared to improve with more structured and iterative prompting strategies [25,33] and varied across LLMs [28,60]. Despite generally high ratings for narrative plausibility, one study [32] found that student satisfaction did not align with expert assessments of clinical accuracy and safety, suggesting that realistic narratives may obscure underlying inaccuracies. Common problems were also reported, including less realistic presenting illness timelines or physical examination data, vague terminology, and lack of patient-centered context [27,86]; unrealistic combinations (eg, mismatched names and ethnic backgrounds or unusual patient behavior [92]); and overly generic or stereotypical cases with limited cultural and linguistic relevance [44,45,62,69,75,87]. These limitations were further evidenced when several LLM-generated exam items required exclusion due to unrealistic scenarios [55]. Overall, superficial realism was common, but expert refinement was typically required to achieve realism.

Psychometric Analysis and Student Perception

Psychometric analyses showed mixed results. Difficulty indices were frequently acceptable for classroom use [42,48,49,76,77,79]. Comparisons with faculty-authored questions yielded mixed findings, with several studies reporting LLM-generated items to be easier [45,55], whereas Yapici Coşkun et al [65] found LLM-generated questions to be substantially more difficult and less balanced across difficulty levels. Discrimination indices were generally acceptable [36,47,48,77,79], but lower discrimination than faculty-written items was consistently observed in some settings [42,45,66,76]. Distractor effectiveness was frequently suboptimal, with high rates of nonfunctional distractors reported [36,45,47,76]. Student performance analyses showed comparable correct response rates between LLM- and human-authored items [47,48,66], improvements over time during interventions [42,44], and variable group-level performance across student levels and subspecialties [42,48,79]. Furthermore, student surveys generally indicated high engagement and favorable perceptions of LLM-generated materials [26,28,30,32,51,52,72], and blinded attribution evaluations reported difficulty distinguishing LLM- from human-authored materials [48,52,61,63,75].

Bias Analysis

Bias issues were explicitly evaluated in a subset of studies [27,40,50,90]. Large-scale analyses showed that generated materials differed substantially from real-world disease prevalence and frequently overrepresented stereotyped groups [90]. Similarly, initial prompting often produced highly stereotyped ethnoracial assignments, while self-prompting altered distributions without restoring epidemiological concordance [40]. Model-specific bias profiles were also observed, including systematic over- or underrepresentation of minority, low-income, and low-education groups across conditions [50]. Bias evaluation of LLM-generated geriatric simulation scenarios revealed limited representation of LGBTQIA + (Lesbian, Gay, Bisexual, Transgender, Queer or Questioning, Intersex, and Asexual) individuals, Indigenous populations, and social determinants of health, with most cases occurring in hospital or primary care office [27]. Other studies reporting on bias noted limited demographic diversity, stereotypical trait-occupation pairings, and culturally incomplete or oversimplified representations [29,53,62,78,92]. Notably, findings from Sridharan et al [70] suggested that incorporating explicit inclusivity criteria into prompt design may support more diverse and culturally responsive content. These studies highlighted a persistent risk of reinforcing social stereotypes and misrepresenting population prevalence in educational materials, potentially affecting clinical training and increasing automation bias.

Comparison of Model Performance

Multiple studies directly compared several LLMs [27,28,32,38,39,42,47,56,60,70,80,85,91], with results varying across models and studies. In case-based question generation tasks, Abouzeid et al [39] found that Gemini produced more consistent items, while ChatGPT-4 (OpenAI) generated greater variability in question content and structure. Robison et al [60] similarly reported distinct strengths, with ChatGPT excelling in detail-rich narratives and structured approaches, Claude demonstrating strong communication and diversity, and Copilot providing valuable educational resources but requiring greater refinement for accuracy and depth. In professionalism and ethical decision-making item generation, İş-Kara and Kıyak [38] found that Gemini 2.5 Pro achieved higher global suitability rates and checklist scores than ChatGPT-5 Thinking, producing more realistic vignettes and more consistently appropriate levels of ambiguity for learners. By contrast, Başaranoğlu et al [42] found no significant difference in technical accuracy between urology MCQs generated by ChatGPT-4 and Gemini 1.5. Also in urology, Szabó and Laein [56] assessed 13 models and found Claude-3.5 Sonnet achieved the highest scores for correctness, relevance, complexity, and variety; though all models showed systematic content imbalance through overrepresentation or omission of key concepts.

Comparisons involving clinical cases and simulation scenarios yielded similarly mixed findings. In gross anatomy, ChatGPT-4 achieved the highest scores, followed by Gemini and ChatGPT-3.5 variants, while Copilot (Microsoft AI) and PaLM 2 (Google) showed high repetition rates and ranked lowest [80]. A comparable pattern was observed in geriatric scenario generation, where ChatGPT achieved the highest overall quality scores, followed by Claude, whereas Grok received the lowest ratings because of incomplete outputs [27]. However, other studies favored different models. Farrag et al [32] found that Gemini achieved the highest overall success rate, Claude achieved the highest clinical accuracy scores, DeepSeek performed best in instructional design quality and clinical reasoning fidelity, whereas GPT-4o showed the lowest performance. Similarly, Öncü et al [28] reported that Claude-3.7 Sonnet produced the highest-quality Turkish-language scenarios, outperforming ChatGPT-4o, Gemini 1.5, and DeepSeek in measures of clarity, realism, educational value, and support for clinical reasoning.

XSL•FO
RenderX

Finally, a subset of studies focused on pharmacotherapy and likewise revealed complementary performance. Güvel et al [47] showed that Claude-3.5 Sonnet produced the highest proportion of error-free pharmacotherapy questions, whereas ChatGPT-4o generated the fewest unusable items. Sridharan and Sequeira [85,91] observed substantial overlap among generated materials, with ChatGPT-3.5 generally providing more detailed explanations and patient-oriented guidance, whereas all models required expert review because of omissions and inconsistencies. More recently, Sridharan and Sivaramakrishnan [70] found that both DeepSeek-V3 and ChatGPT-4.0 generated pedagogically sound materials, but DeepSeek produced more rubric-aligned, contextually rich, and assessment-ready content. Importantly, despite differences in rankings across studies, these findings showed that no model generated case-based materials without identifiable flaws across evaluated domains.

Discussion

Summary of Evidence

This scoping review showed that the use of LLMs to generate case-based materials in HPE has expanded rapidly since 2023, predominantly within undergraduate medical education and largely focused on the production of case-based MCQs and clinical vignettes. In the 72 studies [25-96] reviewed, LLMs were mostly used with simple prompts. More advanced strategies, such as few-shot prompting, prompt chaining, or using external knowledge sources, were rarely applied. Evaluation was usually based on small expert reviews, while psychometric testing and student-based validation were less common. Overall, LLM-generated materials were frequently judged to be clinically plausible, structurally clear, and reasonably aligned with educational objectives, particularly after refinement. However, consistent limitations emerged, including factual inaccuracies, construction flaws, superficial clinical realism, pedagogical misalignment with curricula or training level, suboptimal psychometric performance, and recurrent demographic and epidemiological bias. Finally, comparative studies reported varying performance across LLMs, with no model consistently outperforming others. However, methodological differences between studies limit conclusions about which models perform best.

Comparison With the Literature

Our findings are consistent with previous reviews examining the use of GenAI in HPE. Basil et al [3], in a broad scoping review of 241 studies, identified content creation as the predominant educational application of GenAI, followed by assessment, with medicine and nursing accounting for most published research. They also highlighted generally positive perceptions of GenAI alongside ongoing concerns regarding accuracy, ethical considerations, governance, and the need for more rigorous evaluation of educational outcomes. Similarly, Astbury et al [2] reviewed 23 studies focused on teaching and assessment and found that GenAI was primarily used for learning resource development, assessment, feedback, and simulation, with reported benefits including personalized learning, improved efficiency, and resource use, but persistent concerns regarding accuracy, inconsistent outputs, algorithmic bias, and academic integrity. Our findings reinforce these broader observations while providing a focused and comprehensive synthesis of one important educational application: the generation of case-based educational materials.

A recent review, Kaya et al [20], specifically examined LLM-generated cases. Our findings broadly aligned with their conclusion that LLM-generated cases represent a promising but evolving educational resource, supported by a heterogeneous evidence base. Similar to Kaya et al [20], we found that GPT-based models were the most frequently used LLMs, medicine was the predominant health profession represented, and expert evaluation remained the primary method for assessing generated materials, underscoring the continued need for human oversight before educational implementation. Consistent with their findings, we also observed that prompts were reported in most studies, although prompting approaches varied considerably. However, whereas they primarily described whether the prompt was reported and the use of single versus iterative prompting, our review extends these findings by providing a more granular characterization of generation approaches. In addition to reporting prompt use, we examined the use of reference materials, prompting techniques, and refinement processes, showing that generation predominantly relied on relatively simple prompting strategies with limited adoption of more advanced prompt engineering. Furthermore, while the authors described quality evaluation in broad categories, such as overall case quality and bias, our review also offers a more detailed synthesis of quality assessment by integrating both evaluation methods and findings across multiple quality domains, including clinical accuracy, structural quality, educational alignment, pedagogical quality, realism, psychometric performance, learner perceptions, and bias. This broader perspective provides a more comprehensive understanding of current practices, methodological limitations, and evidence gaps. Finally, our review substantially captures a wider spectrum by synthesizing more than 3 times as many studies, encompassing a broader range of case-based educational materials, including clinical cases, case-based MCQs, Objective Structured Clinical Examination (OSCE) or Objective Structured Practical Examination (OSPE) scenarios, simulation scenarios, script concordance tests (SCTs), and short-answer questions.

Future Directions

From Prompt-Only Generation Toward Robust Generation Pipelines

As shown in our review, most articles relied on relatively simple, prompt-based generation approaches and reported variable output quality that often required expert refinement. These findings underscore the need to move toward more robust generation pipelines. Future research should therefore prioritize structured pipelines that integrate more sophisticated technical strategies, including advanced prompting strategies, grounding in reliable reference sources, and iterative expert oversight.

More advanced prompting strategies, including prompt few-shot, chain-of-thought (CoT), template-based, self-refinement or iterative refinement, and agentic workflows, have the potential to improve the accuracy, clarity, and reliability of LLM outputs

XSL•FO
RenderX

[97-100]. For example, among included studies, Qingquan et al [52] reported that iterative refinement improved the quality of LLM-generated clinical scenarios compared with single-step prompting, with performance comparable to or exceeding human-developed cases. Another promising approach is retrieval-augmented generation (RAG). Introduced in 2020, RAG enables LLMs to generate responses based not only on pretrained knowledge, but also on information retrieved from external, trusted sources such as up-to-date clinical guidelines, textbooks, or medical literature [101,102]. A growing body of evidence has shown that RAG-based systems can enhance LLMs for health care-related tasks (eg, medical question answering, decision support, and summarization), particularly in terms of accuracy, relevance, and reliability [103-106]. Although RAG was rarely used in the studies included in our review, its potential was illustrated by Ananthanarayanan [57], who combined RAG with iterative AI-feedback refinement to generate diagnostic cases. Their pipeline retrieved relevant medical knowledge before generation and iteratively refined outputs using an AI critic, producing cases with excellent medical accuracy.

A simpler alternative to RAG is the use of customized GPTs [107], personalized versions of GPT, configured with precise instructions and domain-specific reference documents to perform targeted tasks. A few studies included in our review adopted this approach. Shultz et al [55] developed a custom GPT using pharmacotherapy textbook chapters and National Board of Medical Examiners (NBME) item-writing guidelines to generate pharmacy examination questions. Abouzeid and Sallam [43] developed a customized GPT for SCT generation by incorporating script theory, relevant literature, and expert-derived scoring keys, whereas Abouzeid et al [31] designed a “Problem-based learning (PBL) Case Builder” that integrates curricular parameters and structured pedagogical guidance to generate PBL cases. Similarly, Zouakia et al [25] used a customized GPT enriched with domain-specific reference documents to generate OSCE stations in digital health, demonstrating that combining a knowledge base with structured prompting improved the quality of generated content.

Finally, robust generation pipelines may benefit from iterative expert-in-the-loop workflows. Incorporating expert feedback across multiple refinement rounds could also improve LLM outputs, aligning with established medical item-development practices and reflecting the level of oversight required in high-stakes contexts [108]. Among included studies [25-96], Coşkun et al [72] described a workflow in which experts reviewed LLM-generated clinical vignettes, identified cases that were insufficiently specific or pedagogically inappropriate, and prompted the AI to regenerate revised versions addressing these deficiencies. Similarly, Lee and Shin [33] adopted a structured expert-in-the-loop approach in which prompts were iteratively refined through expert consultation, followed by separate expert evaluations of educational applicability and clinical representativeness, with each round informing subsequent prompt revisions.

Toward Standardized Evaluation Frameworks

With the rapid expansion of GenAI in health care, several reporting and methodological guidelines have been developed to improve transparency and reproducibility, eg, CONSORT-AI (Consolidated Standards of Reporting Trials-AI) [109] and SPIRIT-AI (Standard Protocol Items: Recommendations for Interventional Trials-AI) [110] for AI-based clinical trials; GAMER (Generative AI tools in Medical Research) [111] for GenAI use in medical research; STAGER (Standardized Testing and Assessment Guidelines for Evaluating Generative AI Reliability) [112] for benchmarking GenAI systems; or METRICS (Model, Evaluation, Timing, Range, Randomization, Individual factors, Count, and Specificity of prompts and language) [113] for GenAI studies in health care education. These frameworks primarily aimed to standardize study reporting and methodological description, rather than provide criteria for evaluating AI-generated educational content itself.

More recently, Olla et al [114] proposed the AI-Integrated Method for Simulation (AIMS), a dual-phase framework for evaluating AI chatbot simulations in nursing education. Unlike reporting guidelines, AIMS evaluates both prompt design (clinical and content credibility, educational alignment, persona clarity, inclusivity, transparency, and personalization) and chatbot performance (credibility and accuracy, user experience, feedback quality, user agency, equity and accessibility, transparency, learning outcomes, and interactivity). Although developed specifically for AI-powered simulation chatbots, this framework illustrates the growing recognition that structured, multidimensional evaluation is needed for AI-generated educational materials.

Consistent with this emerging literature, our review found that evaluations of LLM-generated case-based educational materials relied on heterogeneous, locally defined criteria, most often based on small expert panels, limiting comparability across studies and cumulative evidence building. Based on the recurring evaluation domains identified in our review, we propose that health professions educators and researchers use a minimum evaluation set for LLM-generated case-based educational materials that includes (1) clinical accuracy and alignment with current guidelines, (2) clinical realism, (3) educational alignment (with learning objectives and student level), (4) structural quality (adherence to construction rules, clarity, coherence, and language quality), (5) pedagogical quality (appropriate difficulty, length, and student engagement), and (6) safety considerations, including the identification of biased or potentially harmful content. For assessment applications, this minimum set should be complemented by psychometric analyses, as well as student performance analysis and perceptions.

From Feasibility Studies to Responsible Curriculum Integration

As shown in our review, most studies on LLM-generated case-based materials in HPE were limited to feasibility or proof-of-concept, focusing on output plausibility or short-term evaluations rather than sustained curriculum integration. This aligns with broader evidence summarized by Basil et al [3], who identified persistent barriers to GenAI implementation in HPE, such as infrastructure and access constraints, cost, training

XSL•FO
RenderX

gaps, ethical and legal concerns, and inconsistent output quality, factors that collectively slow down curriculum-level adoption. Addressing these challenges, Astbury et al [2] proposed a Generative AI in Health Professions Education (AI-HPE) checklist to guide responsible implementation, emphasizing institutional readiness, faculty and student preparedness, ethical governance, and ongoing quality assurance. At a more operational level, Kim et al [115] further argued that sustainable integration of AI-generated assessment items requires systematic preparation of faculty and examination bodies, including structured training pathways to develop AI and prompt literacy, clearly defined institutional workflows, shared prompt libraries aligned with local standards, and checklist-driven validation and calibration processes. These converging perspectives highlight that responsible curriculum integration requires not only evidence of validity and feasibility but also institutional readiness, faculty development, ongoing quality assurance, and structured human-AI workflows that maintain educator oversight.

Limitations

This review has limitations related to its scope and methods. First, although a comprehensive search strategy was conducted across 7 major databases, some studies may have been missed due to language limitations (ie, English) and exclusion of gray literature and other databases (eg, Cochrane Library). Moreover, given the rapid development of LLMs in education, recent studies may not have been included; therefore, these findings should be interpreted as a picture of the current evidence. Second, consistent with the exploratory approach of scoping reviews, the methodological quality (including study bias) of the studies was not formally assessed, which could limit conclusions regarding the strength of the evidence. Additionally, the included studies [25-96] were heterogeneous regarding their design, educational settings, generation approaches, and outcome measures, limiting direct comparisons and conclusions about which approaches may improve the quality, reliability, or educational usefulness of LLM-generated materials. Future systematic reviews and meta-analyses could investigate the comparative effectiveness of different generation and refinement approaches. Finally, the review focused only on text-based case materials to maintain clarity and comparability. Studies involving multimodal case generation, such as images or videos, were outside the scope of this review.

Conclusions

This scoping review provides a comprehensive synthesis of the rapidly evolving literature on the use of LLMs to generate case-based materials in HPE. Unlike previous reviews that examined GenAI across multiple educational apps or focused primarily on clinical cases, this review integrates evidence across a broad range of case-based materials while also synthesizing generation approaches, prompting strategies, evaluation methods, reported findings, and challenges. The mapped evidence suggests that LLMs have the potential to support the development of case-based materials, but current approaches continue to require expert oversight because of recurring concerns regarding clinical accuracy, educational quality, psychometric performance, clinical realism, and bias. The review also identifies important evidence gaps, including the predominance of studies in undergraduate medical education, limited evidence across other health professions, the reliance on small expert-based evaluations, limited adoption of advanced generation approaches, and the absence of standardized evaluation frameworks. By synthesizing the current evidence and identifying methodological and evidence gaps, this review provides educators, curriculum developers, and researchers with a structured overview of this rapidly evolving field and highlights priorities for future research. As evidence and LLMs continue to evolve, the findings of this review may inform the development of more robust generation workflows, more consistent evaluation approaches, and the responsible use of LLM-generated case-based materials in HPE.

Generative AI Use

The authors declare the use of generative AI (GenAI) in the research and writing process. According to the 2025 Generative AI Delegation Taxonomy (GAIDeT), the following tasks were delegated to GenAI tools under full human supervision: proofreading and editing, summarizing text, and reformatting. The GenAI tool used was ChatGPT-5 (OpenAI). Responsibility for the final manuscript lies entirely with the authors. GenAI tools are not listed as authors and do not bear responsibility for the final outcomes.

Funding

This work was supported by a French government grant managed by the Agence Nationale de la Recherche under the France 2030 program, reference DigiHealth Paris Cité ANR-23-CMAS-0034 (Coordinator: Dr Rosy TSOPRA).

Data Availability

All data generated or analyzed during this study are included in this published article and its supplementary materials.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Search Strategy.
[DOCX File, 28 KB - mededu_v12i1e94795_app1.docx]

XSL•FO
RenderX

Multimedia Appendix 2
Data charting table of included studies.
[XLSX File, 122 KB - mededu_v12i1e94795_app2.xlsx]

Checklist 1
PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews) checklist.
[DOCX File, 27 KB - mededu_v12i1e94795_app3.docx]

Checklist 2
PRISMA-S (Preferred Reporting Items for Systematic Reviews and Meta-Analyses Literature Search Extension) checklist.
[DOCX File, 17 KB - mededu_v12i1e94795_app4.docx]

References

1. Lin Y, Luo Z, Ye Z, et al. Applications, challenges, and prospects of generative artificial intelligence empowering medical education: scoping review. JMIR Med Educ 2025 Oct 23;11:e71125. [doi: 10.2196/71125] [Medline: 41128430]
2. Astbury H, Fortune E, Ponte CD, Lyons K, Shaw L. Generative artificial intelligence for teaching and assessment in health professions education: a scoping review. Nurse Educ Pract 2026 Feb;91:104697. [doi: 10.1016/j.nepr.2025.104697] [Medline: 41477967]
3. Basil M, Ahmed W, Hajeomar R, Strawbridge J, Lynch M, Mukhalalati B. A scoping review of the use of generative artificial intelligence tools in health profession education. BMC Med Educ 2026 Jan 23;26(1):291. [doi: 10.1186/s12909-025-08527-3] [Medline: 41578287]
4. Zhui L, Yhap N, Liping L, et al. Impact of large language models on medical education and teaching adaptations. JMIR Med Inform 2024 Jul 25;12:e55933. [doi: 10.2196/55933] [Medline: 39087590]
5. Zhang Y, Xie X, Xu Q. ChatGPT in medical education: bibliometric and visual analysis. JMIR Med Educ 2025 Oct 7;11:e72356. [doi: 10.2196/72356] [Medline: 41056572]
6. Zhao Z, Liu Z, Guo L, et al. Artificial intelligence for assessment and feedback in medical education: bibliometric mapping study and thematic evidence map. JMIR Med Educ 2026 Jul 2;12:e98949. [doi: 10.2196/98949] [Medline: 42392589]
7. Pham TD, Karunaratne N, Exintaris B, et al. The impact of generative AI on health professional education: A systematic review in the context of student learning. Med Educ 2025 Dec;59(12):1280-1289. [doi: 10.1111/medu.15746] [Medline: 40533396]
8. McLean SF. Case-based learning and its application in medical and health-care fields: a review of worldwide literature. J Med Educ Curric Dev 2016;3:JMECD.S20377. [doi: 10.4137/JMECD.S20377] [Medline: 29349306]
9. Thistlethwaite JE, Davies D, Ekeocha S, et al. The effectiveness of case-based learning in health professional education. A BEME systematic review: BEME Guide No. 23. Med Teach 2012;34(6):e421-e444. [doi: 10.3109/0142159X.2012.680939] [Medline: 22578051]
10. Piryani RM, Piryani S. Clinical vignette-based interactive discussion sessions: feedback from residents. Adv Med Educ Pract 2019;10:829-833. [doi: 10.2147/AMEP.S218157] [Medline: 31576187]
11. Al Shuraiqi S, Aal Abdulsalam A, Masters K, Zidoum H, AlZaabi A. Automatic generation of medical case-based multiple-choice questions (MCQs): a review of methodologies, applications, evaluation, and future directions. BDCC 2024 Oct 17;8(10):139. [doi: 10.3390/bdcc8100139]
12. Daneshfar M, Moonaghi HK. The impact of clinical simulation on bridging the theory-practice gap in nursing education: a systematic review. BMC Med Educ 2025 Aug 28;25(1):1216. [doi: 10.1186/s12909-025-07790-8] [Medline: 40877870]
13. Shi XY, Lu BR, Yin Q, Wang QW, Fang YQ, Sun ZG. Whether case-based teaching combined with the flipped classroom is more valuable than traditional lecture-based teaching methods in clinical medical education: a systematic review and meta-analysis. BMC Med Educ 2025 Jul 1;25(1):906. [doi: 10.1186/s12909-025-07465-4] [Medline: 40598010]
14. Varma B, Karuveettil V, Fernandez R, et al. Effectiveness of case-based learning in comparison to alternate learning methods on learning competencies and student satisfaction among healthcare professional students: a systematic review. J Educ Health Promot 2025;14(1):76. [doi: 10.4103/jehp.jehp_510_24] [Medline: 40144176]
15. Xiong X, Xu J, Luo M, et al. Efficacy of problem-based learning combined with case-based learning versus lecture-based learning in orthopedic education: a systematic review and meta-analysis. BMC Med Educ 2025 Oct 6;25(1):1357. [doi: 10.1186/s12909-025-07741-3] [Medline: 41053695]
16. Yu Z, Zhao Z, Chen X, et al. Effects of standardised patients (SP) combined with case-based learning (CBL) in Chinese clinical education: a systematic review and meta-analysis. BMJ Open 2025 Sep 3;15(9):e095705. [doi: 10.1136/bmjopen-2024-095705] [Medline: 40903082]
17. Zhu Y, Zhang J, Fei J, Fang H, Zhang Z. Problem-based learning and case-based learning in clinical practical teaching for gynecology residents: a narrative review. Adv Med Educ Pract 2025;16:1269-1279. [doi: 10.2147/AMEP.S534053] [Medline: 40726960]

XSL•FO
RenderX

18. Delany C, Illing J, McIlroy M, et al. Exploring the lived experience of faculty implementing case-based learning: a phenomenological study. BMC Med Educ 2025 Jul 1;25(1):892. [doi: 10.1186/s12909-025-07364-8] [Medline: 40597072]
19. Ali R, Afzal A, Shakil S, et al. Transforming education: case-based integrated learning development and implementation - a mixed methods study at a private medical college. J Adv Med Educ Prof 2025 Jan;13(1):49-58. [doi: 10.30476/jamp.2024.104285.2048] [Medline: 39906080]
20. Kaya AB, Emekli E, Kiyak YS. Mapping applications and outcomes of large-language-model-generated cases in health professions education: a scoping review. Rev Esp Edu Med 2026 Jan 2;7(1). [doi: 10.6018/edumed.691691]
21. Tricco AC, Lillie E, Zarin W, et al. PRISMA Extension for Scoping Reviews (PRISMA-ScR): Checklist and Explanation. Ann Intern Med 2018 Oct 2;169(7):467-473. [doi: 10.7326/M18-0850] [Medline: 30178033]
22. Ouzzani M, Hammady H, Fedorowicz Z, Elmagarmid A. Rayyan-a web and mobile app for systematic reviews. Syst Rev 2016 Dec 5;5(1):210. [doi: 10.1186/s13643-016-0384-4] [Medline: 27919275]
23. Rethlefsen ML, Kirtley S, Waffenschmidt S, et al. PRISMA-S: an extension to the PRISMA Statement for Reporting Literature Searches in Systematic Reviews. Syst Rev 2021 Jan 26;10(1):39. [doi: 10.1186/s13643-020-01542-z] [Medline: 33499930]
24. Pollock D, Peters MDJ, Tricco AC. Scoping reviews. In: Aromataris E, Lockwood C, Porritt K, Pilla B, Jordan Z, editors. JBI Manual for Evidence Synthesis: JBI; 2024. [doi: 10.46658/JBIMES-24-09]
25. Zouakia Z, Logak E, Szymczak A, Jais JP, Burgun A, Tsopra R. AI-driven objective structured clinical examination generation in digital health education: comparative analysis of three GPT-4o configurations. JMIR Med Educ 2026 Jan 15;12:e82116. [doi: 10.2196/82116] [Medline: 41539673]
26. Malpe Gopal S, Venkatesh Murthy S, B R S. Evaluation of a hybrid case-based learning and small-group teaching model using artificial intelligence-assisted case generation: an innovative dental education study. Adv Med Educ Pract 2026;17:591344. [doi: 10.2147/AMEP.S591344] [Medline: 42079984]
27. Ruggiano N, Sahoo S, Brashear A, et al. Evaluating AI-generated geriatric case studies for interprofessional education: systematic analysis across 5 platforms. JMIR Med Educ 2026 Jan 30;12:e83085. [doi: 10.2196/83085] [Medline: 41616316]
28. Öncü S, Torun F, Ülkü HH. Comparison of large language models for clinical scenario generation in medical education: a mixed-methods study. BMC Med Educ 2026 Feb 24;26(1):412. [doi: 10.1186/s12909-026-08821-8] [Medline: 41736016]
29. Rojo-Bofill LM, Ribes Jordán G, Monedero Carrasco L, et al. Impact of the large language model and prompt specificity in generating psychiatric multiple-choice questions. Educ Médica 2026 Mar;27(2):101161. [doi: 10.1016/j.edumed.2026.101161]
30. Yao Y, Geng J, Li W, Li S, Xu Y. Exploring full-cycle DeepSeek-assisted case-based learning in undergraduate radiology education: a respiratory system example. Acad Radiol 2026 Jul;33(7):2779-2789. [doi: 10.1016/j.acra.2026.03.026] [Medline: 41936502]
31. Abouzeid E, Zakaria I, Alshobaily M, Shehata MH, Atwa H. AI-assisted case builder: Making PBL more scalable and adaptable. Med Teach 2026 Feb 13:1-5. [doi: 10.1080/0142159X.2026.2623964] [Medline: 41689255]
32. Farrag AN, El-Zeiny A, Ali AM. Evaluating large language models for pharmacotherapy simulations: a mixed-methods study. NPJ Digit Med 2026 May 5;9(1):355. [doi: 10.1038/s41746-026-02626-1] [Medline: 42086710]
33. Lee M, Shin S. Development and validation of prompts for generating simulation scenarios in nursing education using generative artificial intelligence. Clin Simul Nurs 2026 Apr;113:101922. [doi: 10.1016/j.ecns.2026.101922]
34. Herrera Montano I, Góngora Alonso S, Martínez Licort R, et al. Anamnesio_bot: a chatbot prototype based on generative artificial intelligence for clinical anamnesis study and simulation. In: Rocha A, García Peñalvo F, Costa CJ, Gonçalves R, editors. Proc 20th Iber Conf Inf Syst Technol CISTI 2025: Springer Nature Switzerland; 2026:132-143. [doi: 10.1007/978-3-032-10731-2_13]
35. Wells-Beede E, Thai L, Seo JH, et al. Using AI to create simulation scenarios for a screening brief intervention and referral to treatment virtual reality simulation. Clin Simul Nurs 2026 Jan;110:101883. [doi: 10.1016/j.ecns.2025.101883]
36. Apornvirat S, Pisutpunya A, Atiroj N, Laohawetwanit T. From lecture slides to personalized assessments: ChatGPT-driven digestive pathology questions for targeted learning. Pathol Int 2026 Jan;76(1):e70083. [doi: 10.1111/pin.70083] [Medline: 41508759]
37. Morjaria L, Burns L, Gandhi B, et al. Can large language models generate high-quality short-answer assessments? A comparative study in undergraduate medical education. Appl Sci 2026 Mar 6;16(5):2535. [doi: 10.3390/app16052535]
38. İş-Kara T, Kıyak YS. Large language models in professionalism and ethical decision-making education: a transferable process for generating concordance of professional judgment items. Acad Med 2026 Oct 1;101(10):1248-1252. [doi: 10.1093/acamed/wvag147] [Medline: 42105211]
39. Abouzeid E, Wassef R, Jawwad A, Harris P. Chatbots' role in generating single best answer questions for undergraduate medical student assessment: comparative analysis. JMIR Med Educ 2025 May 30;11:e69521. [doi: 10.2196/69521] [Medline: 40446330]
40. Andrew K, Montalbano MJ. Through a glass darkly: perceptions of ethnoracial identity in artificial intelligence generated medical vignettes and images. Med Sci Educ 2025 Jun;35(3):1473-1488. [doi: 10.1007/s40670-025-02332-9] [Medline: 40625992]

XSL•FO
RenderX

41. Arain SA, Akhund SA, Barakzai MA, Meo SA. Transforming medical education: leveraging large language models to enhance PBL-a proof-of-concept study. Adv Physiol Educ 2025 Jun 1;49(2):398-404. [doi: 10.1152/advan.00209.2024] [Medline: 39918742]
42. Başaranoğlu M, Akbay E, Erdem E. AI-generated questions for urological competency assessment: a prospective educational study. BMC Med Educ 2025 Apr 25;25(1):611. [doi: 10.1186/s12909-025-07202-x] [Medline: 40281621]
43. Abouzeid E, Sallam MA. AI-assisted script concordance tests: enhancing feasibility with customized ChatGPT. Med Teach 2026 May;48(5):757-760. [doi: 10.1080/0142159X.2025.2533405] [Medline: 40705532]
44. Bies JJ, Reddy S, Botros M, et al. Impact of a three-week ChatGPT 3.5-generated case-based professionalism curriculum for internal medicine residents at a United States Medical Residency: a pilot study. Cureus 2025 Jun;17(6):e85401. [doi: 10.7759/cureus.85401] [Medline: 40621296]
45. Chauhan A, Khaliq F, Nayak KR. Title: assessing quality of scenario-based multiple-choice questions in physiology: faculty-generated vs. ChatGPT-generated questions among phase I medical students. Int J Artif Intell Educ 2025 Dec;35(4):2315-2344. [doi: 10.1007/s40593-025-00471-z]
46. Ghaffari F, Langarizadeh M, Nabovati E, Sabery M. Effectiveness of ChatGPT for clinical scenario generation: a qualitative study. Arch Acad Emerg Med 2025;13(1):e49. [doi: 10.22037/aaemj.v13i1.2690] [Medline: 40487903]
47. Güvel MC, Kıyak YS, Varan HD, Sezenöz B, Coşkun Ö, Uluoğlu C. Generative AI vs. human expertise: a comparative analysis of case-based rational pharmacotherapy question generation. Eur J Clin Pharmacol 2025 Jun;81(6):875-883. [doi: 10.1007/s00228-025-03838-2] [Medline: 40205076]
48. Jackson FI, Keller NA, Kouba I, Kouba W, Bracero LA, Blitz MJ. Large language model clinical vignettes and multiple-choice questions for postgraduate medical education. Acad Med 2025 Oct 1;100(10):1163-1166. [doi: 10.1097/ACM.0000000000006137] [Medline: 40550116]
49. Kıyak YS, Soylu A, Coşkun Ö, Budakoğlu I, Peker TV. Can ChatGPT generate acceptable case-based multiple-choice questions for medical school anatomy exams? A pilot study on item difficulty and discrimination. Clin Anat 2025 May;38(4):505-510. [doi: 10.1002/ca.24271] [Medline: 40129054]
50. Liu C, Zheng J, Liu Y, et al. Potential to perpetuate social biases in health care by Chinese large language models: a model evaluation study. Int J Equity Health 2025 Jul 15;24(1):206. [doi: 10.1186/s12939-025-02581-5] [Medline: 40665345]
51. Peralta Ramirez AA, Trujillo López S, Navarro Armendariz GA, De la Torre Othón SA, Sierra Cervantes MR, Medina Aguirre JA. Clinical simulation with ChatGpt: a revolution in medical education? J CME 2025;14(1):2525615. [doi: 10.1080/28338073.2025.2525615] [Medline: 40589612]
52. Qingquan T, Feng R, Bin Z, et al. Iteratively refined ChatGPT outperforms clinical mentors in generating high-quality interprofessional education clinical scenarios: a comparative study. BMC Med Educ 2025 Jun 5;25(1):845. [doi: 10.1186/s12909-025-07414-1] [Medline: 40474219]
53. Rao AS, Kim J, Mu A, et al. Synthetic medical education in dermatology leveraging generative artificial intelligence. NPJ Digit Med 2025 May 4;8(1):247. [doi: 10.1038/s41746-025-01650-x] [Medline: 40320492]
54. Scherr R, Spina A, Dao A, et al. Novel evaluation metric and quantified performance of ChatGPT-4 patient management simulations for early clinical education: experimental study. JMIR Form Res 2025 Feb 27;9:e66478. [doi: 10.2196/66478] [Medline: 40013991]
55. Shultz B, DiDomenico RJ, Goliak K, Mucksavage J. Exploratory Assessment of GPT-4's Effectiveness in Generating Valid Exam Items in Pharmacy Education. Am J Pharm Educ 2025 May;89(5):101405. [doi: 10.1016/j.ajpe.2025.101405] [Medline: 40246172]
56. Szabó A, Laein GD. Comparative evaluation of large language models performance in medical education using urinary system histology assessment. Sci Rep 2025 Aug 29;15(1):31933. [doi: 10.1038/s41598-025-17571-4] [Medline: 40883524]
57. Ananthanarayanan A. Generating medical diagnostic scenarios with LLM-based reinforcement learning feedback: dataset release and methodology. Presented at: 2025 IEEE Integrated STEM Education Conference (ISEC); Mar 15, 2025; Princeton, NJ p. 1-5. [doi: 10.1109/ISEC64801.2025.11147251]
58. Almasri W, Saad M, Mohiyeddini C. Evaluating Google Gemini's capability to generate NBME-standard pharmacology questions using a 16-criterion NBME rubric. Algorithms 2025 Sep 29;18(10):612. [doi: 10.3390/a18100612]
59. Couto De Azevedo M, Toledo Gomes M, Portela Da Fonseca C, Lima Pires L, Bittencourt Machado C. AI-Generated clinical case studies in physiotherapy: enhancing education through integrated Artificial Intelligence. : AHFE International; 2025 Presented at: 13th International Conference on Human Interaction & Emerging Technologies; Apr 22-24, 2025; Malaga, Spain. [doi: 10.54941/ahfe1005935]
60. Robison E, Cooney T, Schwaab T, Rahman S. Development of a prompt template to support simulation design: maximizing the potential of generative artificial intelligence. Clin Simul Nurs 2025 Nov;108:101822. [doi: 10.1016/j.ecns.2025.101822]
61. Hudon A, Phan V, Charlin B, Wittmer R. Teaching clinical reasoning in health care professions learners using AI-generated script concordance tests: mixed methods formative evaluation. JMIR Form Res 2025 Nov 20;9:e76618. [doi: 10.2196/76618] [Medline: 41264864]
62. Wyss LL, Hess RF. Developing simulation content using ChatGPT to teach nursing students culturally appropriate amish burn care. J Transcult Nurs 2026 Jan;37(1):169-176. [doi: 10.1177/10436596251375925] [Medline: 40977597]

XSL•FO
RenderX

63. Panni UY, Donald C, Blatnik JA, Williams M, Yu J, Wise PE. Evaluating the quality of AI-written scenarios for virtual oral surgical board preparatory examination. J Surg Educ 2025 Dec;82(12):103736. [doi: 10.1016/j.jsurg.2025.103736] [Medline: 41125020]
64. Karampinis E, Bozi Tzetzi DA, Pappa G, et al. Use of a large language model as a dermatology case narrator: exploring the dynamics of a chatbot as an educational tool in dermatology. JMIR Dermatol 2025 Sep 16;8:e72058. [doi: 10.2196/72058] [Medline: 40957071]
65. Yapici Coşkun Z, Kiyak YS, Coşkun Ö, Koca M, Budakoğlu İ, Özdemi˙r Ö. Artificial intelligence generated questions in medical education: how prompt design in different chatbots shapes assessment in obstetrics and gynecology? Forbes 2025 Dec 5;6(3):237-241. [doi: 10.4274/forbes.galenos.2025.01033]
66. Touma NJ, Patel R, Skinner T, Leveridge M. Artificial intelligence as a discriminator of competence in urological training: are we there? J Urol 2025 Apr;213(4):504-511. [doi: 10.1097/JU.0000000000004357] [Medline: 39653020]
67. Higashitsuji A, Otsuka T, Watanabe K. Impact of ChatGPT on case creation efficiency and learning quality in case-based learning for undergraduate nursing students. Teach Learn Nurs 2025 Jan;20(1):e159-e166. [doi: 10.1016/j.teln.2024.10.002]
68. Lee SA, Park JH. The usability of neurological occupational therapy case studies generated by ChatGPT. Health Care (Don Mills) 2025 Jun 4;13(11):1341. [doi: 10.3390/healthcare13111341]
69. Zhang Q, Huang Z, Huang Y, et al. Generative AI in medical education: feasibility and educational value of LLM-generated clinical cases with MCQs. BMC Med Educ 2025 Oct 27;25(1):1502. [doi: 10.1186/s12909-025-08085-8] [Medline: 41146115]
70. Sridharan K, Sivaramakrishnan G. Large language models as educational collaborators: developing non-conventional teaching aids in pharmacology & therapeutics. BMC Med Educ 2025 Oct 30;25(1):1525. [doi: 10.1186/s12909-025-08134-2] [Medline: 41168789]
71. Berbenyuk A, Powell L, Zary N. Feasibility and educational value of clinical cases generated using large language models. Stud Health Technol Inform 2024 Aug 22;316:1524-1528. [doi: 10.3233/SHTI240705] [Medline: 39176494]
72. Coşkun Ö, Kıyak YS, Budakoğlu I. ChatGPT to generate clinical vignettes for teaching and multiple-choice questions for assessment: a randomized controlled experiment. Med Teach 2025 Feb;47(2):268-274. [doi: 10.1080/0142159X.2024.2327477] [Medline: 38478902]
73. Emekli E, Karahan BN. Comparison of automatic item generation methods in the assessment of clinical reasoning skills. Rev Esp Edu Med 2024 Nov 25;6(1). [doi: 10.6018/edumed.637221]
74. Han Z, Battaglia F, Udaiyar A, Fooks A, Terlecky SR. An explorative assessment of ChatGPT as an aid in medical education: use it with caution. Med Teach 2024 May;46(5):657-664. [doi: 10.1080/0142159X.2023.2271159] [Medline: 37862566]
75. Hudon A, Kiepura B, Pelletier M, Phan V. Using ChatGPT in psychiatry to design script concordance tests in undergraduate medical education: mixed methods study. JMIR Med Educ 2024 Apr 4;10:e54067. [doi: 10.2196/54067] [Medline: 38596832]
76. Kıyak YS, Coşkun Ö, Budakoğlu I, Uluoğlu C. ChatGPT for generating multiple-choice questions: evidence on the use of artificial intelligence in automatic item generation for a rational pharmacotherapy exam. Eur J Clin Pharmacol 2024 May;80(5):729-735. [doi: 10.1007/s00228-024-03649-x] [Medline: 38353690]
77. Lam G, Shammoon Y, Coulson A, et al. Utility of large language models for creating clinical assessment items. Med Teach 2025 May;47(5):878-882. [doi: 10.1080/0142159X.2024.2382860] [Medline: 39186054]
78. Lopez M, Goh PS. Catering for the needs of diverse patient populations: using ChatGPT to design case-based learning scenarios. Med Sci Educ 2024 Apr;34(2):319-325. [doi: 10.1007/s40670-024-01975-4] [Medline: 38686143]
79. Lotto C, Sheppard SC, Anschuetz W, et al. ChatGPT generated otorhinolaryngology multiple-choice questions: quality, psychometric properties, and suitability for assessments. OTO Open 2024;8(3):e70018. [doi: 10.1002/oto2.70018] [Medline: 39328276]
80. Mavrych V, Ganguly P, Bolgova O. Using large language models (ChatGPT, Copilot, PaLM, Bard, and Gemini) in gross anatomy course: comparative analysis. Clin Anat 2025 Mar;38(2):200-210. [doi: 10.1002/ca.24244] [Medline: 39573871]
81. Patel D, Raut G, Zimlichman E, et al. Evaluating prompt engineering on GPT-3.5's performance in USMLE-style medical calculations and clinical scenarios generated by GPT-4. Sci Rep 2024 Jul 28;14(1):17341. [doi: 10.1038/s41598-024-66933-x] [Medline: 39069520]
82. Chabhadiya PR, Hirapara R, Keshwala A, et al. Potential of large language model chat Generative Pre-trained Transformer in constructing multiple choice questions on pharmacology of diabetes. Asian J Pharm Clin Res 2024:94-96. [doi: 10.22159/ajpcr.2024v17i8.51308]
83. Saad M, Almasri W, Hye T, Roni M, Mohiyeddini C. Analysis of ChatGPT-3.5's potential in generating NBME-standard pharmacology questions: what can be improved? Algorithms 2024 Oct;17(10):469. [doi: 10.3390/a17100469]
84. Saluja S, Tigga SR. Capabilities and limitations of ChatGPT in anatomy education: an interaction with ChatGPT. Cureus 2024 Sep;16(9):e69000. [doi: 10.7759/cureus.69000] [Medline: 39385914]
85. Sridharan K, Sequeira RP. Artificial intelligence and medical education: application in classroom instruction and student assessment using a pharmacology & therapeutics case study. BMC Med Educ 2024 Apr 22;24(1):431. [doi: 10.1186/s12909-024-05365-7] [Medline: 38649959]
86. Takahashi H, Shikino K, Kondo T, et al. Educational utility of clinical vignettes generated in Japanese by ChatGPT-4: mixed methods study. JMIR Med Educ 2024 Aug 13;10:e59133. [doi: 10.2196/59133] [Medline: 39137031]

XSL•FO
RenderX

87. Vaughn J, Ford SH, Scott M, Jones C, Lewinski A. Enhancing healthcare education: leveraging ChatGPT for innovative simulation scenarios. Clin Simul Nurs 2024 Feb;87:101487. [doi: 10.1016/j.ecns.2023.101487]
88. Wilson T, Lee KKW. How I do it: using ChatGPT to create clinical scenarios for surgical learning and oral boards preparation. J Surg Educ 2024 Nov;81(11):1623-1631. [doi: 10.1016/j.jsurg.2024.07.030] [Medline: 39276707]
89. Yanagita Y, Yokokawa D, Uchida S, Li Y, Uehara T, Ikusaka M. Can AI-generated clinical vignettes in Japanese be used medically and linguistically? J Gen Intern Med 2024 Dec;39(16):3282-3289. [doi: 10.1007/s11606-024-09031-y] [Medline: 39313665]
90. Zack T, Lehman E, Suzgun M, et al. Assessing the potential of GPT-4 to perpetuate racial and gender biases in health care: a model evaluation study. Lancet Digit Health 2024 Jan;6(1):e12-e22. [doi: 10.1016/S2589-7500(23)00225-X] [Medline: 38123252]
91. Sridharan K, Sequeira RP. Evaluation of artificial intelligence-generated drug therapy communication skill competencies in medical education. Br J Clin Pharmacol 2025 Aug;91(8):2168-2175. [doi: 10.1111/bcp.16144] [Medline: 38953544]
92. Bakkum MJ, Hartjes MG, Piët JD, et al. Using artificial intelligence to create diverse and inclusive medical case vignettes for education. Br J Clin Pharmacol 2024 Mar;90(3):640-648. [doi: 10.1111/bcp.15977] [Medline: 38016816]
93. Kıyak YS. A ChatGPT prompt for writing case-based multiple-choice questions. Rev Esp Edu Med 2023 Oct 20;4(3). [doi: 10.6018/edumed.587451]
94. Mondal H, Marndi G, Behera JK, Mondal S. ChatGPT for teachers: practical examples for utilizing artificial intelligence for educational purposes. Indian J Vasc Endovasc Surg 2023;10(3):200-205. [doi: 10.4103/ijves.ijves_37_23]
95. Scherr R, Halaseh FF, Spina A, Andalib S, Rivera R. ChatGPT interactive medical simulations for early clinical education: case study. JMIR Med Educ 2023 Nov 10;9:e49877. [doi: 10.2196/49877] [Medline: 37948112]
96. Violato E, Corbett C, Rose B, Rauschning B, Witschen B. The effectiveness and efficiency of using ChatGPT for writing health care simulations. Int J Healthc Simul 2023 Nov 1. [doi: 10.54531/wjgb5594]
97. Wang L, Chen X, Deng X, et al. Prompt engineering in consistency and reliability with the evidence-based guideline for LLMs. NPJ Digit Med 2024 Feb 20;7(1):41. [doi: 10.1038/s41746-024-01029-4] [Medline: 38378899]
98. Mudrik A, Nadkarni GN, Efros O, Soffer S, Klang E. Prompt engineering in large language models for patient education: a systematic review. Health Informatics. Preprint posted online on 2025. [doi: 10.1101/2025.03.28.25324834]
99. Al Shuraiqi S, AlZaabi A, Aal Abdulsalam A. Prompt engineering strategies for generating medical case-based MCQs with large language models: a multi-model comparative study. MAKE 2026;8(2):41. [doi: 10.3390/make8020041]
100. Weerathunge T, Jayalal S, Wijayasiriwardhane K. Optimizing response consistency of large language models in medical education through prompt engineering. Presented at: 2025 5th International Conference on Advanced Research in Computing (ICARC); Feb 19-20, 2025; Belihuloya, Sri Lanka p. 1-6. [doi: 10.1109/ICARC64760.2025.10963313]
101. Neha F, Bhati D, Shukla DK. Retrieval-Augmented Generation (RAG) in healthcare: a comprehensive review. AI 2025 Sep 11;6(9):226. [doi: 10.3390/ai6090226]
102. Amugongo LM, Mascheroni P, Brooks S, Doering S, Seidel J. Retrieval augmented generation for large language models in healthcare: a systematic review. PLOS Digit Health 2025 Jun;4(6):e0000877. [doi: 10.1371/journal.pdig.0000877] [Medline: 40498738]
103. Ke YH, Jin L, Elangovan K, et al. Retrieval augmented generation for 10 large language models and its generalizability in assessing medical fitness. NPJ Digit Med 2025 Apr 5;8(1):187. [doi: 10.1038/s41746-025-01519-z] [Medline: 40185842]
104. Liu S, McCoy AB, Wright A. Improving large language model applications in biomedicine with retrieval-augmented generation: a systematic review, meta-analysis, and clinical development guidelines. J Am Med Inform Assoc 2025 Apr 1;32(4):605-615. [doi: 10.1093/jamia/ocaf008] [Medline: 39812777]
105. Lewis M, Thio S, Roberts A, et al. Grounding large language models in clinical evidence: a retrieval-augmented generation system for querying UK NICE clinical guidelines. arXiv. Preprint posted online on Oct 3, 2025. [doi: 10.48550/arXiv.2510.02967]
106. Xiong G, Jin Q, Lu Z, Zhang A. Benchmarking retrieval-augmented generation for medicine. arXiv. Preprint posted online on Feb 23, 2024. [doi: 10.18653/v1/2024.findings-acl.372]
107. Creating and editing GPTs. OpenAI. URL: https://help.openai.com/en/articles/8554397-creating-a-gpt [accessed 2026-02-16]
108. Rudolph MJ, Daugherty KK, Ray ME, Shuford VP, Lebovitz L, DiVall MV. Best practices related to examination item construction and post-hoc review. Am J Pharm Educ 2019 Sep;83(7):7204. [doi: 10.5688/ajpe7204] [Medline: 31619832]
109. Liu X, Cruz Rivera S, Moher D, Calvert MJ, Denniston AK, SPIRIT-AI and CONSORT-AI Working Group. Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI extension. Nat Med 2020 Sep;26(9):1364-1374. [doi: 10.1038/s41591-020-1034-x] [Medline: 32908283]
110. Cruz Rivera S, Liu X, Chan AW, et al. Guidelines for clinical trial protocols for interventions involving artificial intelligence: the SPIRIT-AI extension. Nat Med 2020 Sep;26(9):1351-1363. [doi: 10.1038/s41591-020-1037-7] [Medline: 32908284]
111. Luo X, Tham YC, Giuffrè M, et al. Reporting guideline for the use of generative artificial intelligence tools in MEdical research: the GAMER statement. BMJ EBM 2025 Dec;30(6):390-400. [doi: 10.1136/bmjebm-2025-113825]
112. Chen J, Zhu L, Mou W, et al. STAGER checklist: standardized testing and assessment guidelines for evaluating generative artificial intelligence reliability. IMetaOmics 2024 Sep;1(1):e7. [doi: 10.1002/imo2.7] [Medline: 41675544]

XSL•FO
RenderX

113. Sallam M, Barakat M, Sallam M. A preliminary checklist (METRICS) to standardize the design and reporting of studies on generative artificial intelligence-based models in health care education and practice: development study involving a literature review. Interact J Med Res 2024 Feb 15;13:e54704. [doi: 10.2196/54704] [Medline: 38276872]
114. Olla P, Wodwaski N, Long T. Beyond the bot: a dual-phase framework for evaluating AI chatbot simulations in nursing education. Nurs Rep 2025 Jul 31;15(8):280. [doi: 10.3390/nursrep15080280] [Medline: 40863667]
115. Kim B, Kang J, Kim MY, Ahn J. Prompt engineering for single-best-answer multiple-choice questions in licensing examinations: a narrative review with a case study involving the Korean medical licensing examination. J Educ Eval Health Prof 2025;22:34. [doi: 10.3352/jeehp.2025.22.34] [Medline: 41140266]

Abbreviations

AIMS: AI-Integrated Method for Simulation
CONSORT-AI: Consolidated Standards of Reporting Trials-AI
CoT: chain-of-thought
GAMER: Generative AI tools in Medical Research
GenAI: Generative AI
GPT: Generative Pretrained Transformer
HPE: Health Professions Education
LLM: large language model
LQBTQIA+: Lesbian, Gay, Bisexual, Transgender, Queer or Questioning, Intersex, and Asexual
MCQ: multiple-choice question
METRICS: Model, Evaluation, Timing, Range, Randomization, Individual factors, Count, and Specificity of prompts and language
NBME: National Board of Medical Examiners
OSCE: Objective Structured Clinical Examination
OSPE: Objective Structured Practical Examination
PBL: problem-based learning
PRISMA-S: Preferred Reporting Items for Systematic Reviews and Meta-Analyses Literature Search Extension
PRISMA-ScR: Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews
RAG: retrieval-augmented generation
SCT: Script Concordance Test
SPIRIT-AI: Standard Protocol Items: Recommendations for Interventional Trials-AI
STAGER: Standardized Testing and Assessment Guidelines for Evaluating Generative AI Reliability

Edited by S Brini; submitted 08.Mar.2026; peer-reviewed by AB Kaya, G Shankar Ganesh, H Takahashi; revised version received 12.Aug.2026; accepted 16.Aug.2026; published 03.Nov.2026.

Please cite as:
Zouakia Z, Szymczak A, Tordjman S, Logak E, Meyre T, Resche-Rigon M, Cariou A, Burgun A, Tsopra R
Generation of Case-Based Materials Using Large Language Models in Health Professions Education: Scoping Review
JMIR Med Educ 2026;12:e94795
URL: https://mededu.jmir.org/2026/1/e94795
doi:10.2196/94795

© Zineb Zouakia, Alan Szymczak, Salomé Tordjman, Emmanuel Logak, Thomas Meyre, Matthieu Resche-Rigon, Alain Cariou, Anita Burgun, Rosy Tsopra. Originally published in JMIR Medical Education (https://mededu.jmir.org), 3.Nov.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Education, is properly cited. The complete bibliographic information, a link to the original publication on https://mededu.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Interventions to Enhance COVID-19 Pandemic Health Literacy in Health Professionals: Systematic Review

Claudia Hasenpusch; Paula Kuper; Eva-Maria Skiba; Antonia Anabella Sprenger; Uwe Matterne; Lena Kannengießer; Esther Ivanka Grätsch; Jaekyung Shin; Mengtong Li; Christian Joachim Apfelbacher

Institute of Social Medicine and Health Systems Research, Medical Faculty, Otto von Guericke University Magdeburg, Leipziger Straße 44, Magdeburg, Germany

Corresponding Author:
Claudia Hasenpusch
Institute of Social Medicine and Health Systems Research, Medical Faculty, Otto von Guericke University Magdeburg, Leipziger Straße 44, Magdeburg, Germany

Abstract

Background: The COVID-19 pandemic has placed a significant burden on health professionals (HPs). They face higher infection risks due to the nature of their work environment and patient care responsibilities. Their ability to access and apply reliable COVID-19 information affects their own preventive behavior and that of those around them. In this context, health literacy (HL) has become increasingly important. Despite extensive research, information to foster COVID-19–related HL in HPs remains limited.

Objective: This systematic review aimed to identify, appraise, and synthesize intervention studies on the effectiveness of COVID-19–related HL interventions in HPs.

Methods: Five electronic databases (eg, PubMed (MEDLINE), Embase), six clinical trials registries (eg, ISRCTN registry), one preprint server (MEDRXIV), published conference proceedings, and five gray literature databases (eg, opengrey.eu, ProQuest) were searched in May 2022 and updated in August 2025. Reference lists of included studies were screened manually. Two reviewers independently screened titles, abstracts, and full-texts according to eligibility criteria and extracted data; disagreements were resolved by discussion or consultation with a third reviewer. We included randomized controlled trials (RCTs), nonrandomized studies of interventions, and uncontrolled before-and-after studies evaluating the effectiveness of any COVID-19–related HL intervention. Primary outcomes include COVID-19–related HL, its four facets (access, understand, appraise, and apply COVID-19 information), and indicators (eg, COVID-19–related knowledge), assessed at postintervention and follow-up. When studies were sufficiently similar, random-effects meta-analyses were performed; otherwise, a narrative synthesis was provided. Risk of bias was assessed using validated tools based on study design, and the overall certainty of the evidence was evaluated by the GRADE (Grading of Recommendations, Assessment, Development, and Evaluation) approach.

Results: We included 15 RCTs (2034 participants), 4 nonrandomized studies of interventions (291 participants), 74 uncontrolled before-and-after studies (327,298 participants), 5 ongoing studies, and 1 study with awaiting classification. Interventions targeted a broad range of health occupational groups. Intervention type, delivery mode, methods, settings, and comparator varied widely. No outcome measure explicitly referred to an HL model. Most studies aimed to enhance COVID-19–related knowledge and skills, and had a high risk of bias. COVID-19-related interventions may increase knowledge of vaccines (standardized mean difference 1.00; 95% CI 0.33 to 1.67, I^2=24%), and the infection prevention control skills, such as donning and doffing of personal protective equipment (standardized mean difference 1.95; 95% CI 1.82 to 3.09, I^2=46%), but the evidence remains very uncertain.

Conclusions: COVID-19–related HL interventions may promote HP's short-term competencies in infection control. However, the evidence remains uncertain, primarily due to the low quality of studies, characterized by a high risk of bias. Interventions specifically designed to enhance the full COVID-19 HL operationalized by its four facets are lacking. High-quality RCTs with sufficient statistical power, grounded in HL theoretical principles, are needed to achieve precise understanding.

Trial Registration: PROSPERO CRD42022301810; https://www.crd.york.ac.uk/PROSPERO/view/CRD42022301810

(JMIR Med Educ 2026;12:e70400) doi:10.2196/70400

KEYWORDS

pandemics; SARS-CoV-2; infection control; comparative effectiveness research; health literacy; health education

XSL•FO
RenderX

Introduction

The occurrence of novel infectious diseases, such as COVID-19, requires a rapid implementation of prevention measures to contain the spread of the virus. Infection control measures to protect oneself and others are linked to behavior changes of individuals [1]. In the face of constantly changing recommendations and infection control measures in the early phase of the pandemic, individuals needed to recognize recommended measures such as social distancing, wearing a face mask, washing and disinfecting their hands regularly, testing for SARS-CoV-2, or getting vaccinated, and had to be capable of carrying them out [2-4]. For example, in March 2020, using a face mask was not considered an effective preventive measure for the prevention of potential transmission of SARS-CoV-2, while in June 2020, the World Health Organization (WHO) had modified their recommendation based on new evidence of its effectiveness [5]. Attainment of habituation of preventative behavior is important to consider as well if one wishes the behavior to be sustained [1].

The pandemic has not only posed unprecedented challenges for the general population, but health professionals (HPs) in particular faced a high burden, including occupational exposure, increased workload demands, rapidly changing infection-prevention requirements in patient care, and considerable psychological strain [6-8]. HPs, including physicians, nurses, midwives, physiotherapists, and other HPs [9], have a heightened risk of infection, especially during patient care and disposal of biomedical infectious waste. The risk for infection in this occupational population is 35% higher than in professions without infection-relevant working conditions [10], given the direct exposure to individuals who may be asymptomatic carriers of infectious diseases such as COVID-19. Besides their daily demands, HPs need to be cautious to prevent the spread of infections to noninfected patients, colleagues, and even their own social environment, simultaneously [7,11,12]. This requires real-time adjustments, that is, adapting to changes in governmental policies, recommendations, and adhering to guidance and clinical protocols immediately in order to ensure that care is provided adequately and safely. In addition, HPs deal with ethical responsibilities and moral obligations [7,13,14] when deciding to spend time in environments where the risk of infection is heightened. This dilemma, the responsibility to provide high-quality care to patients while simultaneously safeguarding themselves from potential harm, was intensified during the pandemic, when the demand for health care services was overwhelming [7]. Infection disease prevention competencies, such as proper hand hygiene, surface disinfection, use of personal protective equipment (PPE), biomedical waste management, and infection prevention and control practice training, are crucial in mitigating the infection risks for HPs and those around them [5,6,15].

Implementing personal protective measures in one's everyday life depends on the ability to access, understand, appraise, and apply pandemic-relevant recommendations and information [16,17]. This ability can be defined as COVID-19–related health literacy (HL) [16,18,19]. In general, HL refers to the motivation, knowledge, and competencies of individuals (access, understand, appraise, and apply health information) to make decisions related to their health [20]. Various factors influence the health-related competencies of individuals, for instance, situational (eg, media use, social support, family, and peer influences) and personal determinants (eg, age, gender, and socioeconomic status) [20]. As the onset of the COVID-19 pandemic, and followed by its transition into an endemic, HL in the context of communicable diseases, particularly of COVID-19, has gained substantial importance [16,21,22]. It has been suggested that adequate HL may not be as prevalent among populations as needed to successfully navigate the increasing amount of both reliable information and misinformation during the onset of the pandemic, referred to as "infodemic" [21,23-26]. The *Health Literacy Survey COVID-19* found that during the first wave of the COVID-19 outbreak in Germany (spring 2020), 50.4% (n=1037) of the study population had "insufficient" COVID-19–related HL [17,22]. These findings are comparable to the results of previous studies on general HL [22,27]. However, the proportion of participants with low HL decreased by 14.9% at the second time of measurement (December 2020) [22]. Okan et al [22] assumed this can be explained by the fact that, since the beginning of the COVID-19 pandemic, information about SARS-CoV-2 and COVID-19 has been provided continuously and extensively on different information channels. Further, synthesized evidence has shown an association between HL and (nonpandemic) infectious disease prevention behavior [28]. This association may be even more relevant during a pandemic.

HL may be improved by providing valid and comprehensible information, effective communication strategies, and structured education [29]. Interventions are either tailored to the specific needs of at-risk individuals (eg, patients or predisposed individuals) or the general population [29]. HL is a crucial intervention target, as it may counteract the widespread dis- and misinformation during a pandemic [17,22,30].

There is evidence that HL can be improved by interventions. A systematic review of school-based educational interventions found higher competence in the critical appraisal of health claims shortly after intervention (eg, knowledge and skills) [31]. Another systematic review concluded that group and individual interventions in both primary health care and community settings may support sustained change in HL [32]. Further systematic reviews reported that interventions aiming to improve HPs' mental HL [33] and eHealth literacy [34] were effective.

However, interventions aiming to improve HL should be conceptually distinguished from broader health education interventions, infection prevention, or behavioral skills training. While health education programs often focus on providing knowledge, teaching procedural skills, or promoting specific behaviors, HL interventions more specifically aim to strengthen individuals' competencies to access, understand, appraise, and apply health information in order to make informed health decisions [35]. Although health education strategies are often used to foster HL competencies, these approaches are not conceptually synonymous and should therefore be differentiated when defining and interpreting intervention studies.

XSL•FO
RenderX

While a considerable amount of research has been undertaken on HL during the COVID-19 pandemic [17,22,35-37], evidence on interventions specifically aiming to enhance COVID-19–related HL among HPs remains limited [7,36]. Therefore, the purpose of this systematic review was to identify, appraise, and synthesize studies examining the effectiveness of interventions aiming to improve COVID-19–related HL, its dimensions (access, understand, appraise, and apply), or related competency indicators among HPs.

Methods

Study Design

The systematic review was reported according to the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) 2020 Statement and Checklist (Checklist 1) and abstract checklist (Checklist 2) [38]. It was registered at PROSPERO (CRD42022301810).

Ethics Approval and Consent to Participate

This study is a systematic review of the literature. Ethical approval was not necessary.

Search Strategy

We searched in five electronic databases (PubMed [MEDLINE], Embase, Web of Science [Science Citation Index & Social Sciences Citation Index], PsycInfo, and Epistemonikos Database). To identify ongoing studies, we also searched five clinical trial registries (Cochrane Central Register of Controlled Trials, Cochrane COVID-19 Study Register, ISRCTN registry, Australian New Zealand Clinical Trials Registry, and EU Clinical Trials Register). Additionally, we screened preprint servers (medRxiv), published conference proceedings, and gray literature (Base, opengrey.eu, Social Science Open Access, Centers for Disease Control and Prevention WONDER, and ProQuest). We manually searched the reference lists of included studies and contacted corresponding authors to obtain unpublished study reports. The initial search was performed in May 2022, followed by a first and second update search in December 2023 and August 2025. Only studies published after the onset of COVID-19 (December 2019) were considered; a filter was applied accordingly. The search strategy was not restricted in terms of the population of interest or language. The review question was operationalized based on the PICO (Population-Intervention-Comparator-Outcome) framework, which was extended by the study design [39]. The search strategies are presented in detail in the additional file Multimedia Appendix 1 [40-42].

Eligibility Criteria

Population

According to the international classification of health workers provided by the WHO (2019) [9], we included HPs, representing an occupational population, who (1) provide preventive, curative, rehabilitative, and promotional health services based on extensive theoretical and practical knowledge in diagnosing and treating diseases (eg, doctors and nurses), (2) perform technical and practical tasks and assist (eg, medical and pathology laboratory technicians, pharmaceutical technicians, community health workers, and ambulance workers), (3) provide direct personal care in health care and residential settings, assist with health care procedures, and perform routine tasks to support health services (eg, health care assistants and home-based personal care workers), (4) provide health management and support (eg, biomedical engineers, medical physicists, clinical psychologists, social workers, medical secretaries, ambulance drivers, and other professional, technical, administrative, and support staff), and (5) participate in the health labor market by providing health services, including medical interns and trainees offering clinical services as part of their education. Additionally, we also included studies on students of various health care–related fields (eg, medicine, nursing, paramedics, and health sciences).

Interventions

Any type of intervention was eligible if aimed at enhancing COVID-19 pandemic–related HL or at least one of the four HL components, or a related competency conceptually linked to these components. Moreover, educational or training interventions were eligible when their content or intended outcomes could be meaningfully aligned with the underlying HL framework. Therefore, we also included interventions addressing COVID-19–related knowledge, skills, and competencies as HL proxies, here referred to as indicators. Eligible intervention approaches included, for instance, digital interventions (eg, self-help tools delivered via apps and online training sessions via video), group-based or individual interventions, and educational interventions (eg, lectures and workshops) delivered by scientists, educators, or HPs. We excluded experimental interventions such as manipulations examining cognitive processes. Further, for intervention studies that target behavior changes without a discernible link to HL-related competencies, or mental health outcomes, mental HL was judged as not eligible. Interventions, which contain exposure to other epidemic or pandemic infectious disease outbreaks (eg, SARS-CoV-1 and MERS [Middle East respiratory syndrome]) or chronic infectious diseases (eg, tuberculosis and hepatitis) were also excluded.

Outcomes

This systematic review refers to the integrated model of HL by Sørensen et al [20]. We adapted the definition of HL in the context of COVID-19 as "knowledge and competences to access, understand, appraise, and apply COVID-19 pandemic–related health information in order to make judgments and take decisions in everyday life concerning COVID-19 infection prevention" [17,20]. Based on this definition, the primary outcomes were COVID-19–related HL as well as components and indicators (eg, knowledge and skills), assessed using either objective or subjective instruments.

As validated instruments specifically measuring COVID-19–related HL among HPs were largely unavailable during the early phase of the pandemic [43], we considered observable indicators reflecting HL competencies. These indicators were not regarded as equivalent to the full construct of HL, but as proxy measures for specific components within the underlying framework. Within the integrated HL model, knowledge and related competencies are required for processing

XSL•FO
RenderX

and using health information. Knowledge tests were therefore considered proxy indicators of the "understand" component, while performance-based infection prevention skills were considered indicators of "apply" for the purpose of this review. Performance or achievement indices and COVID-19–specific knowledge tests were eligible when they could be conceptually linked to objectively assessed HL-related competencies. Subjective HL comprised self-reported judgments about how well individuals perceive themselves to know and be able to access, understand, appraise, and apply COVID-19–specific health information.

With regard to "knowledge" as an HL indicator, we distinguished between general and specific COVID-19 knowledge. General COVID-19 knowledge covered a broad range of basic themes, including, but not limited to, disease symptoms, routes of transmission, incubation periods, and infection prevention and control measures. Specific COVID-19 knowledge referred to more focused areas, such as vaccination, handling of PPE, infection management strategies, mechanisms of virus transmission, or detailed infection prevention and control measures.

Further, skills were considered as HL-related indicators only when they reflected the application of COVID-19–specific infection prevention knowledge rather than purely technical clinical procedures. In this review, the indicator "skill" referred to the ability or capacity of protecting oneself and/or others (eg, patients, colleagues, and family members) from a SARS-CoV-2 infection through correctly performing recommended prevention control measures such as donning and doffing of PPE, hand hygiene, surface disinfection, wearing a mask, managing biomedical waste, and conducting nasopharyngeal swabs.

The secondary outcomes were classified and conceptually assigned to the underlying model [20]: prevalence of infection prevention behavior and adherence to recommended infection prevention measures (health behavior), attitudes toward recommended infection prevention measures and perceived confidence and self-efficacy (determinants), usage and frequency of consultation of trusted sources of information (access), and perceived or objective health status, health-related quality of life, and frequency of health care usage (health outcomes). Behavioral outcomes were not considered as direct measures of the "apply" facet but were interpreted as behavioral outcomes potentially influenced by HL competencies.

Study Types

We included randomized controlled trials (RCTs, eg, individual and cluster RCTs), nonrandomized studies of interventions (NRSIs; eg, quasi-randomized and controlled trials), and uncontrolled before-and-after studies defined as one-group prepost designs in which outcomes are measured before and after an intervention in the absence of a comparison group [44]. Studies that did not report an interventional design or a before-and-after comparison were excluded.

Data Collection and Analysis

Selection of Studies

All identified references were deduplicated after being uploaded into supportive software (EPPI-Reviewer [45] and Rayyan [46]). Each title and abstract was screened by two independent reviewers according to predefined eligibility criteria. Full-text papers of potentially relevant references were reviewed by the same reviewers for inclusion. Any discrepancies were solved by discussion or by consultation with a third reviewer until consensus was reached. Studies that did not meet our eligibility criteria were excluded.

When multiple records referred to the same study population (eg, study protocols with subsequent completed publications or abstracts with full-text reports), we included only the most complete and relevant version.

Data Extraction and Management

We designed a data extraction form in Microsoft Excel, with predefined items that were pilot tested based on 5 included studies. At least two reviewers performed data extraction. Hereby, one reviewer extracted relevant data. A second reviewer checked the data extraction table for accuracy. Disagreements were resolved by discussion or through arbitration by a third reviewer when consensus was not reached. In case of insufficient data, we contacted the authors of the studies and requested further information. We extracted in detail the following information presented in Table 1.

Table . Overview of extracted data for analysis.

Study characteristics	Authors, journal, year of publication, geographical location/country, study design, study duration (year/period of study implementation), theoretical foundation, funding sources, disclosure of conflict of interest
Intervention characteristics	• Type of intervention, intervention duration, frequency, response rate, follow-up period, and delivery mode of the intervention
Participant characteristics	• Age, sex, number of participants recruited/allocated/evaluated, health profession, setting (eg, hospital and community setting)
Study outcomes	• Primary outcome categories: COVID-19–related health literacy, at least one facet (access, understand, appraise, or apply) or indicator (eg, knowledge and skills) • Secondary outcomes categories: Prevalence of, adherence to, attitudes toward infection prevention behavior, consultation of trusted sources of information, confidence and self-efficacy, health status, and health-related quality of life

XSL•FO
RenderX

Assessment of Risk of Bias Within Included Studies

Risk of bias was assessed by groups of two independent reviewers (CH, AAS, PK, EIG, ML, and JS) using the Cochrane RoB 2 Tool (Cochrane Risk of Bias 2 Tool) [47] for RCTs, the ROBINS-I (Risk of Bias in Nonrandomized Studies of Interventions) [48] for NRSIs tool and the EPHPP-QAT (Quality Assessment Tool for Quantitative Studies) [49] for uncontrolled before-and-after studies. The tools consist of several domains, which are shown in Table 2. Eligible studies were appraised for each domain of the respective risk of bias tool. Based on the overall rating, the level of risk of bias for each domain was judged for the RoB 2 tool ("low risk," "some concerns," and "high risk"), for ROBINS-I ("low risk," "moderate risk," "serious risk," "critical risk," and "no information") and for EPHHP-QAT tool ("strong," "moderate," and "weak") [47-49]. For the overall judgment of risk of bias using the Cochrane RoB 2 Tool, an algorithm generated a proposed assessment for each domain based on the highest score on an individual domain. The overall judgment for outcome assessments with ROBINS-I was based on a mapping of domain-level ratings. Both overall judgments are informed by the responses to the signaling questions. The overall score for each domain of the EPHPP-QAT was used to determine the global rating for uncontrolled before-and-after studies, with scores ranging from 1.00 to 1.50 categorized as "weak," 1.51 to 2.50 as "moderate," and up to 3.00 as "strong" [49].

Table . Risk of bias assessment domains of the tools used in the systematic review.

Tools and domains	Items (n)
Cochrane RoB 2 tool[a], Sterne et al [47], domains (n=5)	
Bias arising from the randomization process	3
Bias due to deviations from intended interventions	7
Bias due to missing outcome data	4
Bias in measurement of the outcome	5
Bias in selection of the reported result	3
ROBINS–I[b], Sterne et al [48], domains (n=7)	
Bias due to confounding	8
Bias in selection of participants into the study	5
Bias in classification of interventions	3
Bias due to deviations from intended interventions	6
Bias due to missing data	5
Bias in measurement of outcomes	4
Bias in selection of the reported result	3
EPHPP-QAT[c], Armijo-Olivo et al [49], domains (n=6)[d]	
Selection bias	2
Study design	4
Confounders	2
Blinding	2
Data collection	2
Withdrawals/dropouts	2

[a]Cochrane RoB 2 Tool: Cochrane Risk of Bias 2 Tool.

[b]ROBINS–I: Risk of Bias in Nonrandomized Studies of Interventions.

[c]EPHPP-QAT: Quality Assessment Tool for Quantitative Studies.

[d]The EPHPP tool comprises eight domains, with the risk of bias being assessed based on six of these [49].

Certainty of Evidence

The overall certainty of the evidence by outcome was determined according to the GRADE (Grading of Recommendations, Assessment, Development, and Evaluation) [50] approach by two independent reviewers (CH and EMS). The quality of the body of evidence was assessed based on four levels of certainty—"high," "moderate," "low," and "very low" [50,51]. Considering the study design (RCTs or observational studies, eg, before-and-after studies), the quality of the eligible studies was either downgraded referring to five reasons (risk of bias, inconsistency, indirectness, imprecision, and publication bias) or upgraded based on three reasons (large effect, dose response, and all plausible residual confounding) [50]. If any discrepancies between the reviewer judgments occurred, a third reviewer (PK) was consulted.

XSL•FO
RenderX

Data Synthesis

Study characteristics and results were narratively summarized and presented in a summary of findings table. When at least two studies were judged sufficiently similar in terms of intervention, outcome measurement, and time point, a meta-analysis was performed [42]. We calculated Hedges *g* for each comparison to correct for small sample bias. If means and SDs were not reported, we converted other statistics such as medians and IQRs or odds ratios to Hedges *g* where possible [42,52,53]. To pool effect sizes, we used random effects meta-analysis with the inverse variance method, as we expected the true intervention effects to differ due to considerable remaining between-study heterogeneity. If a study contributed more than one effect size, a three-level random effects meta-analysis model was used, with effect sizes nested in studies [54]. The heterogeneity variance τ2 was estimated using the Restricted Maximum-Likelihood Estimator [55]. We used the Hartung-Knapp-Sidik-Jonkman method for calculating test statistics and CIs to account for uncertainty in the heterogeneity estimation, where appropriate, given the number of studies and the extent of between-study heterogeneity [56]. Statistical significance was defined at a 2-sided *P* value of less than .05. The heterogeneity variance τ2 and the I2 statistic were reported along with their respective CIs as measures of heterogeneity. I2 statistics ranging from 0% to 40% were interpreted as "might not be important," from 30% to 60% as "moderate," from 50% to 90% as "substantial," and from 75% to 100% as "considerable" [42]. For values falling in overlapping ranges (eg, 55% or 80%), we interpreted them in context and used them for broad orientation rather than strict categorization. In addition, we also considered τ^2 as an absolute measure of heterogeneity to support our assessment of heterogeneity and calculated 95% prediction intervals (PIs) indicating the range in which effects of future studies are expected based on present evidence using the Nagashima method [57]. However, for analyses including only a single study, PIs were not estimated as between-study heterogeneity could not be assessed [57]. If heterogeneity was moderate to high and the analysis included more than five studies, subgroup analysis using study- or result-level variables such as risk of bias was conducted to explore heterogeneity. If the analysis included ten or more comparisons, funnel plots and the Egger test were applied to assess potential publication bias [58]. The results were visualized in forest plots where studies were ordered by weight in ascending order. Analysis was conducted in R Studio (version 2024.04.1 for Windows; R Foundation) using the packages *meta* [59] and *metafor* [60].

Deviation From Study Protocol

During the review process, several deviations from the original protocol were introduced in response to characteristics of the identified literature. All deviations were implemented before the final data synthesis. First, the original protocol did not restrict the target population, as we initially assumed that only a limited number of eligible studies would be available. During the review process, it became evident that a considerable proportion of potentially relevant studies targeted various populations. To ensure conceptual comparability of PICO-criteria, the population was restricted to HPs.

Second, the protocol initially did not specify language restrictions to maximize the sensitivity of the search. As a substantial number of studies originated from different linguistic contexts, a language restriction was introduced to ensure accurate assessment of eligibility and data extraction. This decision was based on the research team's present proficiency in English, French, and German. Spanish studies were translated using the artificial intelligence translation tool DeepL by DeepL SE, and ChatGPT by OpenAI, followed by an expert verification.

Third, although gray literature and preprints were searched, only completed studies were included in the present analysis to ensure methodological rigor and robustness of the findings. Ongoing studies were documented for potential inclusion in future updates.

Fourth, the protocol did not specify an approach for assessing uncontrolled before-and-after studies. Therefore, these studies were evaluated using the EPHPP-QAT, ensuring a standardized assessment of methodological quality.

Fifth, during data extraction, we observed discrepancies between reported intervention content and measured outcomes, and combined multiple outcomes in one single measure. To adequately capture these measures and provide a more comprehensive evaluation of the interventions, studies reporting composite outcomes were considered for inclusion into the review.

Finally, while the protocol prespecified detailed procedures for addressing heterogeneity and conducting meta-analyses, substantial variability across interventions and outcome measures required a more cautious analytical approach. Therefore, the meta-analyses were conducted only when studies were deemed sufficiently comparable and were interpreted as exploratory. These listed deviations did not affect the overall conclusions of this review.

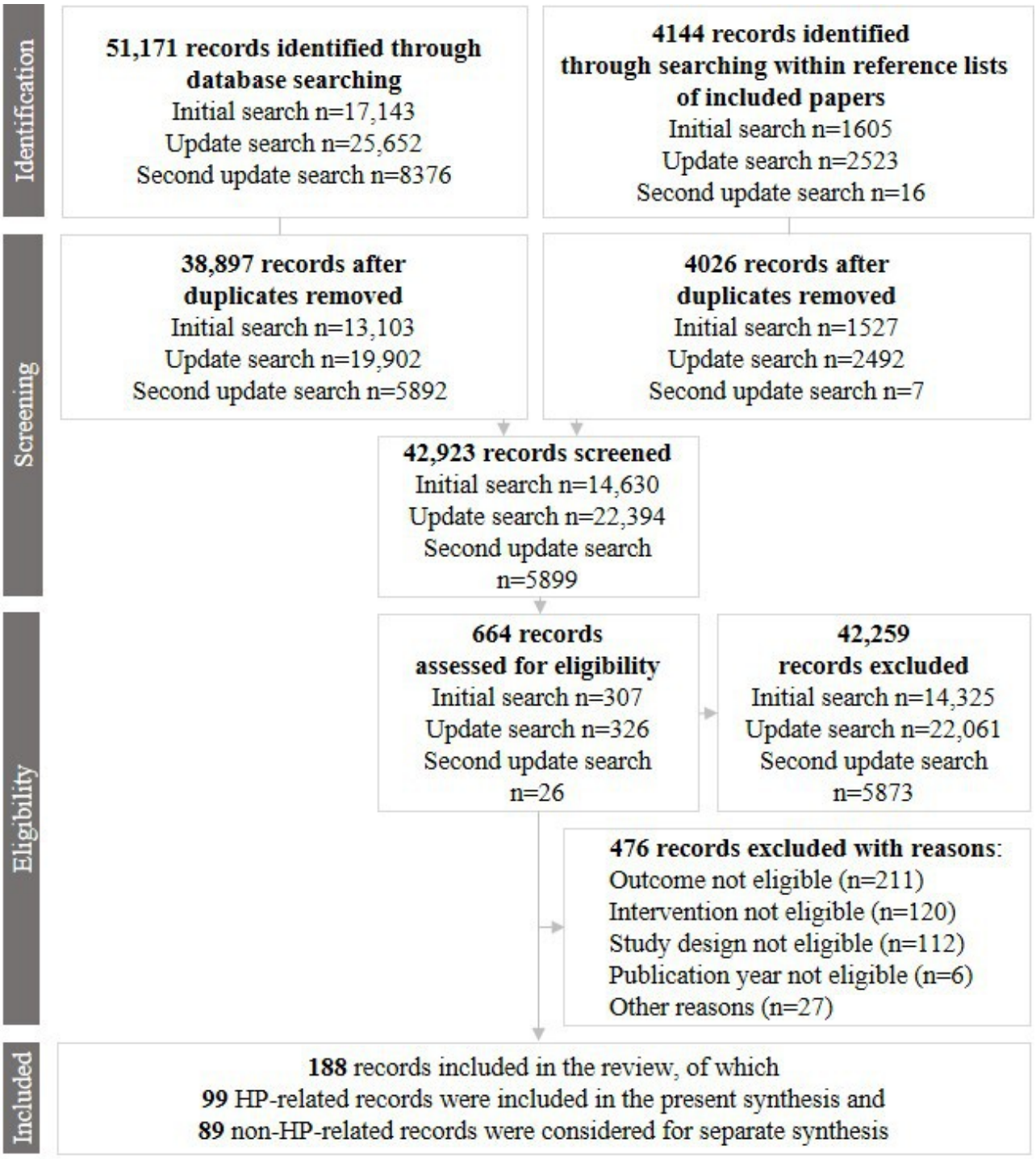
Results

Overview

The results of the study search and selection process are shown in Figure 1. The search yielded a total of 51,171 papers, of which 188 studies were deemed relevant. Among these, 89 studies addressed the target population of non-HPs, while 99 focused on HPs. Given the considerable number of potentially relevant studies, the present systematic review primarily focused on the population of HPs. The results about other populations will be presented in a separate publication. For future updates, we identified ongoing studies for potential inclusion; however, only published studies have been included in the present analysis to report the outcomes. In the additional file Multimedia Appendix 2 [61-188], Table S1 shows the excluded studies with reasons provided.

XSL•FO
RenderX

Figure 1. Flowchart diagram according to PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) statement 2020 [38]. Other reasons: study protocols and study registries excluded due to already included full-text publication, not eligible context or population. HP: health professional.



Study Characteristics

Of the included studies, there were 15 RCTs [61-74], 4 NRSIs [75-78], 74 uncontrolled before-and-after studies [79-138,188], 5 ongoing studies [139-143], and 1 study as awaiting classification [144]. The references were published between 2020 and 2025. Most studies (96/99 studies, 97%) were published in English and 3 studies were published in Spanish. In Multimedia Appendix 2, the Tables S2-S4 show the summarized characteristics of included studies.

Population

In total, 329,623 HPs were enrolled in interventional studies, ranging from 18 to 234,643 participants per intervention. Most participants were enrolled in uncontrolled before-and-after studies (n=327,298), whereas substantially fewer participants were included in RCTs (n=2034) and NRSIs (n=291). Most interventions were directed toward a broad range of mixed health occupational groups, accounting for 43% of studies (40/93 studies, 298,220 participants). Interventions targeted particularly at students from various disciplines, including medicine, nursing, dentistry, or pharmacy, were examined in 29% of studies (27/93

XSL•FO
RenderX

studies, 4388 participants). Five of the 93 (5%) studies were conducted in both HPs and non-HPs (9068 participants). Table 3 shows the distribution of study populations by study design in more detail. RCTs were predominantly conducted among students and specialist medical practitioners, whereas uncontrolled before-and-after studies more frequently included mixed HP groups, nursing professionals, pharmacists, midwives, and students. Therefore, the distribution of professional groups differed across study designs.

The average age of participants was 29.9 (SD 8.8) years, as reported in 37% (34/93 studies) of studies. Further, 64% of the participants were female (65 studies). Most studies were conducted in Asia (48/93 studies, 52%), North America (17/93 studies, 18%), and Europe (15/93 studies, 16%). Overall, of the 93 included studies 39 (42%) took place in high-income countries, 27 (29%) in lower-middle countries, 26 (28%) in upper-middle-income countries, and 3 (3%) in low-income countries.

XSL•FO
RenderX

Table . Study population of included studies (N=93) by study designs.

Study population	Participants, n total	Studies, k total	RCT[a], k (n participants)	NRSI[b], k (n participants)	UC[c], k (n participants)
Health professionals[d] (n=7451, k=12)					
Generalist medical practitioners	5886	1	—[e]	—	1 (5886)
Specialist medical practitioners	126	2	2 (126)	—	—
Nursing professionals	950	7	1 (344)	—	6 (606)
Pharmacists	279	1	—	—	1 (279)
Midwives	210	1	—	—	1 (210)
Health associate professionals (n=237, k=4)					
Health care laboratory professionals	52	1	—	—	1 (52)
Environmental, occupational health, and hygiene professionals	77	1	—	—	1 (77)
Community health workers	108	2	1 (18)	—	1 (90)
Students (n=4388, k=27)					
Medical students	571	7	1 (29)	1 (129)	5 (413)
Nursing students	1537	9	—	1 (50)	8 (1487)
Paramedics students	203	2	2 (203)	—	—
Dental students	118	1	—	—	1 (118)
Pharmacy students	85	1	—	—	1 (85)
Respiratory therapist students	37	1	—	—	1 (37)
Mixed student disciplines	1837	6	2 (444)	—	4 (1393)
Mixed health professional groups (n=317,297, k=50)					
Various health professional occupation groups	298,220	40	5 (581)	2 (112)	33 (297,527)
Students and health professionals	10,009	5	1 (21)	—	4 (9988)
Health professionals and nonhealth professionals	9068	5	—	—	5 (9068)
Health professional occupation not specified (n=250, k=1)					
	—	1	1 (250)	—	—

[a]RCT: randomized controlled trial.

[b]NRSI: nonrandomized study of interventions.

[c]UC: uncontrolled before-and-after studies.

[d]Categorization in accordance with the World Health Organization classification of health workers [9].

[e]Not applicable.

Interventions and Comparators

The type of intervention, delivery mode, and methods, settings, and comparator varied widely across the included studies in Table S2 in Multimedia Appendix 2. We identified three predominant intervention types: (1) educational interventions, (2) training programs, and (3) blended intervention approaches. Overall, 61 of the 93 (66%) studies applied educational interventions with traditional teaching of theoretical knowledge. Six (6%) studies provided training programs comprising demonstrations and the performance of practical, hands-on skills

XSL•FO
RenderX

(eg, donning and doffing of PPE). Within 26 of the 93 (28%) studies, blended interventions were conducted, which combined both theoretical and practical approaches. While educational interventions were the most common intervention type across all study designs, they were predominantly evaluated in uncontrolled studies. Training programs were mainly evaluated in controlled study designs, whereas blended approaches were evaluated in both controlled and uncontrolled studies (Table 4). In controlled studies, comparators most commonly consisted of educational approaches delivered in formats similar to the intervention (Table 5), indicating that many studies compared alternative educational strategies rather than using a passive control group.

XSL•FO
RenderX

Table . Intervention characteristics by study designs (N=93).

Intervention characteristics	Studies, k total (%)	RCT[a], k (%)	NRSI[b], k (%)	UC[c], k (%)
Type of interventions				
Educational interventions	61 (66)	9 (10)	2 (2)	50 (54)
Training programs	6 (6)	2 (2)	—[d]	4 (4)
Blended intervention approaches	26 (28)	4 (4)	2 (2)	20 (22)
Guidance				
Self-guided	16 (17)	5 (5)	1 (1)	10 (11)
Provider-guided	43 (46)	6 (6)	1 (1)	36 (39)
Self- and provider-guided	32 (34)	4 (4)	2 (2)	26 (28)
Delivery mode				
Group	50 (54)	9 (10)	2 (2)	39 (42)
One-to-one	25 (27)	7 (8)	1 (1)	17 (18)
In-person	29 (31)	3 (3)	—	26 (28)
Online distance learning	43 (46)	6 (6)	1 (1)	36 (39)
Blended learning (ie, online and in-person combined)	11 (12)	5 (5)	—	6 (6)
Technologically enhanced in-person learning	6 (6)	2 (2)	2 (2)	2 (2)
Information on delivery mode not available	7 (8)	—	—	7 (8)
Delivery method				
Technology-enhanced methods				
e-learning (through digital resources and platforms)	20 (22)	4 (4)	—	16 (17)
Virtual reality simulation	3 (3)	—	2 (2)	1 (1)
Interactive methods (eg, workshops)	3 (3)	1 (1)	—	2 (2)
Lecture-based methods	10 (11)	—	—	10 (11)
Self-directed methods	5 (5)	3 (3)	—	2 (2)
Practice-based methods (eg, demonstrations and role play)	5 (5)	—	1 (1)	4 (4)
Gamified methods				
Simulation games	1 (1)	—	1 (1)	—
Educational games	2 (2)	—	—	2 (2)
Blended methods	51 (55)	13 (14)	—	38 (41)
Information on delivery method not available	6 (6)	—	—	6 (6)

[a]RCT: randomized controlled trial.

[b]NRSI: nonrandomized study of interventions.

[c]UC: uncontrolled before-and-after studies.

[d]Not applicable.

XSL•FO
RenderX

Table . Comparator characteristics by study designs (N=93).

Type of comparator	Studies, k total (%)	RCT[a], k (%)	NRSI[b], k (%)	UC[c], k (%)
Educational intervention	13 (14)	10 (11)	3 (3)	—[d]
Training programs	1 (1)	1 (1)	—	—
Blended intervention approaches	2 (2)	2 (2)	—	—
No intervention	1 (1)	—	1 (1)	—
Placebo intervention	1 (1)	1 (1)	—	—
Usual care	1 (1)	1 (1)	—	—
Waitlist	2 (2)	2 (2)	—	—
No control group	74 (80)	—	—	74 (80)

[a]RCT: randomized controlled trial.

[b]NRSI: nonrandomized study of interventions.

[c]UC: uncontrolled before-and-after studies.

[d]Not applicable.

HPs (27/93 studies, 29%), health educators (10/93 studies, 11%), trained instructors (13/93 studies, 14%), and researchers (18/93 studies, 19%) delivered the interventions. Universities (23/93 studies, 25%) and hospitals (33/93 studies, 35%) were the most prevalent settings. The duration of interventions lasted from 10 minutes to 19 weeks. Most of the interventions were provided in groups (50/93 studies, 54%), followed by one-to-one sessions (25/93 studies, 27%). Online distance learning (43/93 studies, 46%) was more prevalent than in-person teaching (29/93 studies, 31%) and hybrid learning (11/93 studies, 12%).

Underlying Frameworks for Guiding Interventional Studies

Among 93 studies, 41 (44%) reported a theoretical framework or model for guiding processes such as development, implementation, and evaluation of the intervention. Hereby, 6 studies (6%) referred to learning models (Peyton's approach [64], Miller's pyramid-based training [127], cognitive learning theory [91,100,103], and simulation-based mastery learning [116]) and 2 (2%) applied health behavior change models (eg, health belief model [66] and theory of change model [112]). Further, two (2%) studies were based on Kirkpatrick's evaluation model [110,145]. Health manuals, COVID-19 protocols, and clinical information (eg, WHO protocols and Infection Prevention Control guidelines) were applied in 27 (29%) studies. While 2 (2%) of the included studies incorporated HL as a component within the curriculum [188] or as an underlying concept for an online learning platform [121], no study explicitly referred to an HL model or provided an established definition.

Outcomes Measurement

The vast majority of studies reported primary outcomes related to changes in COVID-19 knowledge (75/93 studies, 81%). Of the 93 studies, 20 (22%) assessed general knowledge of COVID-19, while 43 (46%) focused on specific knowledge outcomes, primarily concerning infection prevention and control measures (19 studies, 20%), PPE performance (8 studies, 9%), COVID-19 vaccine (9 studies, 10%), SARS-CoV-2 transmission (5 studies, 5%), and infection management (4 studies, 4%). In 21 (23%) studies [62-64,67-69,72,73,75,78,84,91,92,97,106,116,122,127,129,146,147] changes of COVID-19 infection prevention skills were targeted, with enhancement of PPE performance being the most frequently addressed outcome (13 studies, 14%).

In total, 38 of the 93 (41%) studies [62,65,66,74,76,78,80,83,84,86,87,90,93,97,101,104,105,108,111,112,119,123,130-132,135,146-155,188] reported secondary outcomes such as confidence and self-efficacy (19 studies, 20%), compliance toward and prevalence of infection prevention measures behavior (6 studies, 7%), attitudes (11 studies, 12%), and composite outcomes (21 studies, 23%). One study [121] developed an online learning platform aimed at fostering HL and focused in particular on knowledge, attitudes, and practices. Another study [62] reported adverse events related to motion sickness caused by the use of virtual reality simulation. Among 21 studies (23%) a discrepancy was observed between the reported outcomes and the items actually used to measure the intended construct.

Most studies (65/93 studies, 70%) measured outcomes at short-term follow-up, defined as immediately following the intervention for up to four weeks from the program's start. Of 93 studies, 8 (9%) reported medium-term follow-up, ranging up to six months. Additionally, 9 (10%) studies determined outcomes at a second follow-up ranging from one week to eight months after completing the intervention. In 19 (20%) studies, the time point of the postintervention outcome assessment was not clearly reported.

The assessment differed across all 93 studies. The content of the assessment tools was based on hospital internal protocols [68,72,73,83], national and international guidelines, such as those provided by the WHO [84,91,100,146,156], the Centers for Disease Control and Prevention [91,116,134], or the Ministry of Health [84,153]. Further, 60 studies (65%) applied self-developed instruments, created for the study. Additionally, 26 (28%) studies reported information on the development of the assessment tool. Among 37 (40%) studies using a validated tool for outcome measurement, 11 (12%) studies did not provide

XSL•FO
RenderX

details regarding the validation process and measurement properties. Most studies (82/93 studies, 88%) used participants' reported measures with Likert scales, multiple-choice, "true"/"false" or "yes"/"no" response options. In 17 of the 93 (18%) studies [62-64,69,73,75,83,84,90-92,97,106,116,122,127,129], outcomes were assessed through performance-based evaluations conducted by an observer (eg, teacher, instructor, or member of the study team) using a checklist. In one study [188], the assessment tool included a HL-related item specifically.

Overall, the included studies indicated substantial conceptual overlap between COVID-19–related HL and educational or skills-based infection control interventions. Most studies focus on knowledge and procedural competencies, corresponding to the "understand" and "apply" facets, rather than explicitly addressing the full construct. By contrast, no study explicitly assessed COVID-19–related HL as an overall construct or examined changes in "access" and "critical appraisal" of health information (Table 6).

XSL•FO
RenderX

Table . Outcomes conceptually assigned to levels of the underlying health literacy model by study design (N=93).

Components and studies, n (%)	References
HL-related components	
Access	—[a]
Understand	
General COVID-19–related knowledge	
1 (1) RCT[b]	[66]
1 (1) NRSIs[c]	[77]
18 (19) UCs[d]	[84,89,92,93,95,102,103,110,112-115,117,137,146,152,156,188]
Specific COVID-19–related knowledge (eg, on vaccination, PPE[e] and IPC[f] measures)	
7 (8) RCTs	[61,64,70-73,149]
1 (1) NRSI	[78]
35 (38) UCs	[81-83,85,87,88,90,94-100,104,106-108,110,118,122,124-126,130,131,133,134,136,138,145,147,150,157,158]
Composite knowledge-related outcomes	
1 (1) NRSI	[76]
11 (12) UCs	[80,83,105,111,123,130,131,151,153-155]
Critically appraise	—
Apply	
General COVID-19–related infection prevention performance skills	
1 (1) NRSI	[78]
6 (6) UCs	[84,92,106,122,129,147]
PPE performance skills (donning and/or doffing)	
8 (9) RCTs	[63,64,67-69,71-73]
1 (1) NRSI	[75]
4 (4) UCs	[91,97,116,127]
Hand disinfection	
2 (2) RCTs	[62,64]
Nasopharyngeal swab performance	
1 (1) RCT	[62]
Composite skills-related outcomes	
1 (1) UC	[154]
HL-related factors	
Attitude	
2 (2) RCTs	[74,149]
1 (1) NRSI	[147]
8 (9) UCs	[84,92,93,109,112,146,150,152]
Confidence/self-efficacy	
7 (8) RCTs	[64-67,69-71]
1 (1) NRSI	[78]
11 (12) UCs	[83,87,96,97,101,104,105,108,119,128,188]
Prevention behavior	
Prevalence of COVID-19 IPC measures behavior	
1 RCT (1)	[66]

XSL•FO
RenderX

Components and studies, n (%)	References
2 (2) UCs	[79,152]
Compliance to IPC measures	
1 (1) UC	[90]
Adherence to IPC measures	
2 (2) UC	[93,112]

[a]Not applicable.

[b]RCT: randomized controlled trial.

[c]NRSI: nonrandomized study of interventions.

[d]UC: uncontrolled before-and-after studies.

[e]PPE: personal protective equipment.

[f]IPC: infection prevention control.

Risk of Bias Assessment

An overview of the risk of bias assessments of included studies is presented in detail in the additional file Multimedia Appendix 3 [61-138,145-158,188]. Among primary outcomes of RCTs (*k*=15) assessed with the Cochrane RoB 2 Tool, 14 of 26 (54%) compared results of primary outcomes were evaluated as "some concerns." Primary outcomes of 6 of 26 (23%) results had "low" risk of bias, while another 6 of 26 (23%) results were rated at "high" risk of bias. Most results (16/26 studies, 62%) were rated at high risk of bias regarding the selection of the reported results due to a lack of prespecified statistical analysis plans or study protocols. Some studies [61,63] did not adequately handle missing outcome data.

Based on the ROBINS-I assessments for NRSIs, all six evaluated outcomes of 4 included studies were judged as having "serious" risk of bias. Pertaining to the risk of bias of uncontrolled before-and-after studies using EPHPP-QAT, 73 of 74 (99%) studies were rated as being "weak," and 1 study as "moderate." Most uncontrolled studies used a nonprobability purposive or convenience sampling technique, resulting in a higher risk of selection bias. Neither did they report information on the blinding of participants nor outcome assessors. All NRSI and uncontrolled before-and-after studies are prone to confounding (eg, preexisting COVID-19 knowledge, interest in gaining more knowledge, and socioeconomic status), but no study adjusted or controlled for potential confounding.

Overview of Findings Pertaining to the Effectiveness of Interventions

We performed meta-analyses for changes of COVID-19–related knowledge and skills. Hereby, we restricted meta-analyses to RCTs and primary outcomes due to high heterogeneity across interventions and outcomes in both NRSIs and uncontrolled before-and-after studies and due to inconsistencies in secondary outcome definitions in RCTs. The results are presented in the summary of findings table, along with GRADE certainty of evidence ratings (Table 7). The summary of findings tables for all included studies, separated by study design, are shown in the additional file Multimedia Appendix 4 [61-138,145-158,188,189].

XSL•FO
RenderX

Table . Summary of findings. Interventions for enhancing COVID-19–related health literacy (HL) in health professionals.[a]

Outcomes	Relative effect (95% CI)	Participants, n (studies)	Certainty of the evidence (GRADE[b])	Comments
Primary outcomes				
COVID-19–related knowledge				
General knowledge (F1)	—[c]	181 (2 nonrandomized studies) Buyego et al [76], Hu et al [77]	—[d,e,f]	Meta-analysis and the GRADE domain inconsistency were not applicable because metrics of outcomes differed widely across NRSIs[g]
Vaccine knowledge (F1)	SMD[h] 1.00 (0.33; 1.67)	155 (1 RCT[i]) Alotaibi et al [61]	very low[e,j,k]	The included study contains three different intervention groups, on which the meta-analysis and the GRADE assessment were based
PPE[l] knowledge (F1)	SMD 0.21 (−0.09, 0.51)	565 (4 RCTs) Currat et al [64], Suppan et al [70], Suppan et al [71], Xie et al [73]	low[d,e]	Among 4 included studies, 1 study [70] reported outcomes in a metric that was not convertible; therefore, the meta-analysis was based on only three studies: Currat et al [64], Suppan et al [71], and Xie et al [73]
COVID-19–related infection prevention skills				
PPE performance doffing (F1)	SMD 0.69 (−0.95; 1.78)	154 (4 RCTs) Birrenbach et al [62], Christensen et al [63], Currat et al [64], Manggala et al [68]	very low[m,n,o]	—
PPE performing doffing (F2)	SMD 0.61 (−0.07; 1.29)	93 (2 RCTs) Birrenbach et al [62], Currat et al [64]	very low[p,q,m]	—
PPE performing donning (F1)	SMD 0.28 (−0.48; 1.04)	61 (2 RCTs) Christensen et al [63], Manggala et al [68]	very low[m,r]	—
PPE performing donning and doffing (F1)	SMD 1.95 (0.82; 3.09)	274 (3 RCTs) Li et al [67], Rueda-Medina et al [69], Xie et al [73]	very low[m,r,s]	—
Hand disinfection (F1)	—	93 (2 RCTs) Birrenbach et al [62], Currat et al [64]	—[p,q,d]	Meta-analysis and the GRADE domain inconsistency were not applicable because metrics of outcomes differed widely across RCTs
Hand disinfection (F2)	—	93 (2 RCTs) Birrenbach et al [62], Currat et al [64]	—[p,q,d]	Meta-analysis and the GRADE domain inconsistency were not applicable because metrics of outcomes differed widely across RCTs
Secondary outcomes				
Perceived confidence in the ability to use PPE	—	553 (4 RCTs) Currat et al [64], Li et al [67], Suppan et al [70], Suppan et al [71]	very low[r,d,t]	Meta-analysis and the GRADE domain inconsistency were not applicable because metrics of outcomes differed widely across RCTs.

[a]Population: health professionals of various professions; setting: various health care settings; intervention: various interventions; comparison: various comparators.

[b]GRADE: Grading of Recommendations, Assessment, Development, and Evaluation.

[c]Not applicable.

[d]As the sample size is very small in these studies, we rated down the quality of evidence by one level for imprecision.

XSL•FO
RenderX

[e]The proportion of information from studies at high risk is sufficient to affect the interpretation of the results; crucial limitation for one criterion, or some limitations for multiple criteria, sufficient to lower confidence in the estimated effect.

[f]As the metrics of the results differ widely across the studies, we have downgraded the quality of certainty by one level.

[g]NRSI: nonrandomized study of interventions.

[h]SMD: standardized mean difference.

[i]RCT: randomized controlled trial.

[j]A study sample based on one study and that covers only a subgroup of the population of interest.

[k]We rated down the quality of evidence by one level due to the very small sample size.

[l]PPE: personal protective equipment.

[m]As the sample size is very small in both studies, and the CI includes appreciable benefit and no effect, we rated down the quality of evidence by two levels for imprecision.

[n]Most information from results is at low risk or has some concerns, but 1 study was at high risk of bias (missing data domain). Potential limitations are likely to lower confidence in the estimated effect.

[o]Minor CI overlap. Wide variance of point estimates. I^2 implies moderate to substantial heterogeneity. Heterogeneity cannot be explained.

[p]Most information from results is at low risk of bias or which were rated as some concerns (selection bias and performance bias). Plausible bias that raises some doubt about the results. Potential limitations are likely to lower confidence in the estimated effect.

[q]Very few studies that address only a subgroup of the population of interest (eg, medical students) compared according to the underlying World Health Organization classification of health workers.

[r]The proportion of information from studies from both at high risk and with some concerns is sufficient to affect the interpretation of the results. A crucial limitation for one criterion, or some limitations for multiple criteria, is sufficient to lower confidence in the estimated effect.

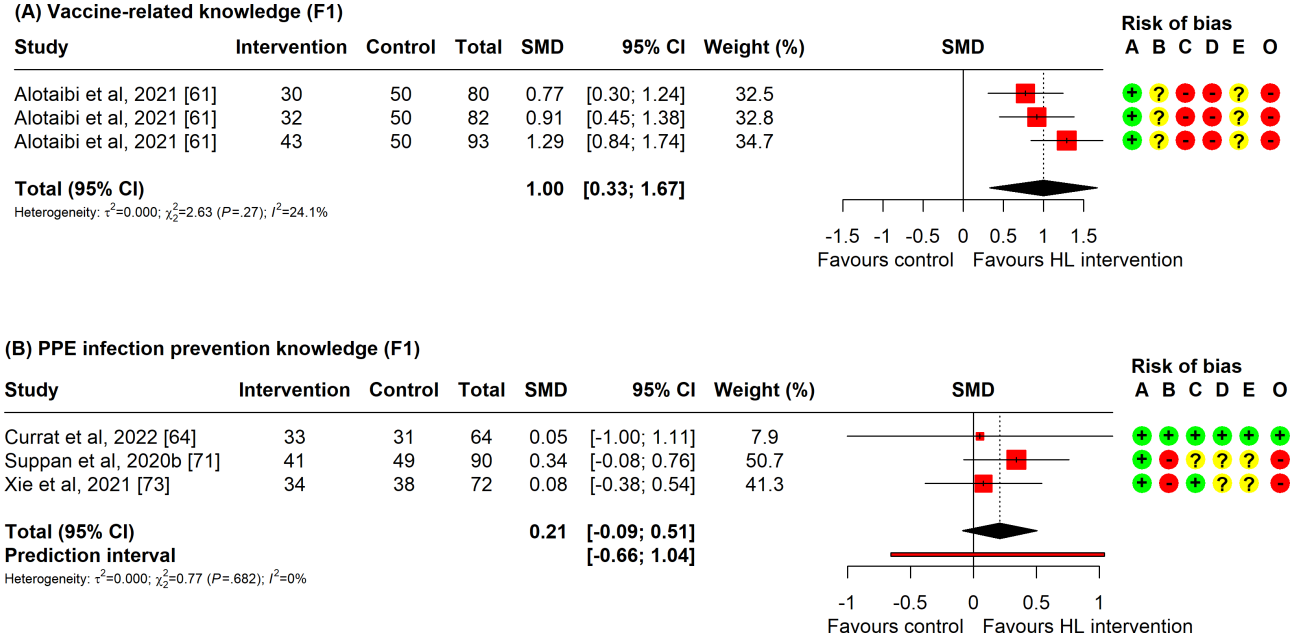
[s]CI overlap. Wide variance of point estimates. I^2 implies moderate to substantial heterogeneity. Heterogeneity cannot be explained.

[t]As the effects differ widely across the studies, we have downgraded the quality of certainty by one level.

Effectiveness of Interventions in Enhancing COVID-19–Related Knowledge

Data from four RCTs were considered for the meta-analysis on the outcome COVID-19–related knowledge, grouped into knowledge on vaccine and PPE (Figure 2). COVID-19-relevant HL interventions may increase vaccine knowledge at postintervention (standardized mean difference [SMD] 1.00; 95% CI 0.33 to 1.67, I^2=24%), but the evidence is very uncertain about the effect. Further, the evidence is uncertain and suggests that interventions may not increase PPE knowledge at postintervention (SMD 0.21; 95% CI −0.09 to 0.51, I^2=0%).

Figure 2. COVID-19–related knowledge. SMD: standardized mean difference.



Effectiveness of Interventions in Enhancing COVID-19–Related Infection Prevention and Control Skills

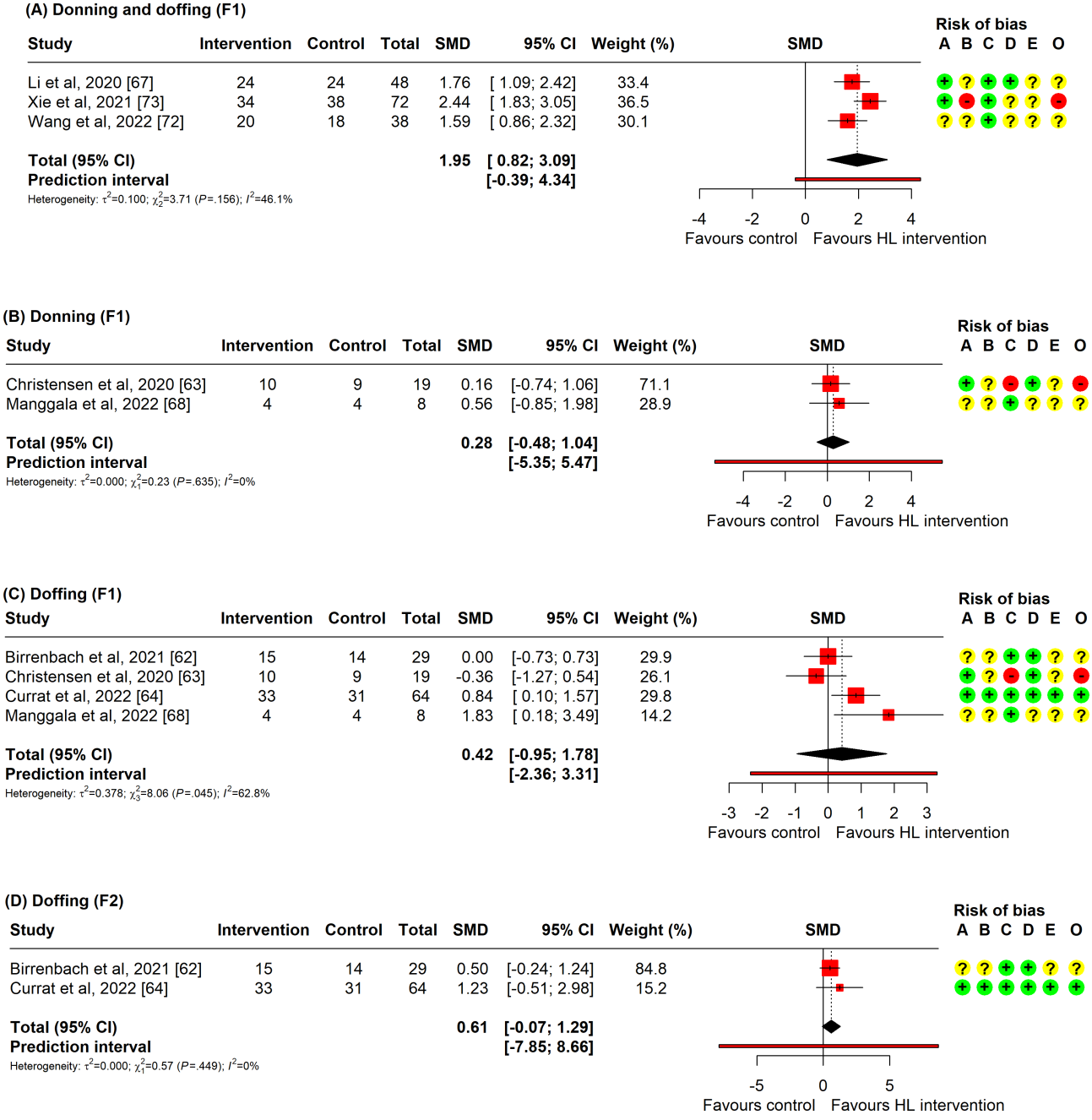
Of the 15 RCTs included, 7 RCTs contributed data to the meta-analysis (Figure 3). They reported on COVID-19–related infection prevention and control performance and were categorized into competencies in PPE donning, doffing, or both. The HL-relevant interventions may increase postintervention donning and doffing skills, though the evidence remains very uncertain (SMD 1.95; 95% CI 0.82 to 3.09, I^2=46.1%). The evidence is very uncertain about the effect of interventions on doffing skills at postintervention (SMD 0.42; 95% CI −0.95 to 1.78, I^2=62.8%) and at follow-up (SMD 0.61; 95% CI −0.07 to

XSL•FO
RenderX

1.29, I^2=0%), as well as on donning skills at postintervention (SMD 0.28; 95% CI −0.48 to 1.04, I^2=0%).

In addition to the findings of the meta-analysis, a narrative description of the 74 uncontrolled before-and-after studies is provided in Multimedia Appendix 5 [94,107,112].

Figure 3. COVID-19–related PPE donning and doffing. PPE: personal protective equipment; SMD: standardized mean difference.



Discussion

Principal Findings

This systematic review aimed to identify and critically evaluate evidence pertaining to interventions that focus on enhancing COVID-19–related HL in HPs. Meta-analysis revealed favorable changes in COVID-19–related infection prevention and control skills after the intervention, particularly in the performance of PPE donning and doffing, which are essential competencies for reducing the risk of health care–associated infections. These findings are similar to those of Saunders et al [190], who systematically investigated the impact of HL training in HP students.

However, the pooled effect sizes must be interpreted with caution. Although some effect sizes were large, they were derived from a small number of studies with limited sample sizes, which increases the risk of imprecision and unstable effect estimates. Small studies are also more susceptible to bias and may overestimate intervention effects. This is reflected in the GRADE assessment, which indicated low to very low certainty of evidence due to risk of bias, inconsistency, and imprecision. Further, many interventions targeted only a subgroup of the population of interest, for example, medical students. Whether

XSL•FO
RenderX

the reported results can be applied to the population of HPs at large remains unclear [9]. Thus, the identified body of evidence can largely be characterized by high inconsistency. While the pooled effect estimate suggested a statistically significant effect for the outcomes COVID-19 vaccine-related knowledge and competencies in PPE donning and doffing at postintervention, the PI included nonsignificant and both potentially beneficial and harmful effects, indicating considerable variability and limiting the generalizability of effects across intervention approaches, HP occupational populations, settings, and outcomes. This finding is consistent with the results of similar reviews [190,191].

Moreover, the small sample sizes in most of the included studies may have led to a failure to detect meaningful differences due to insufficient statistical power. They might be attributed to the high workload of HPs, which was exacerbated during the onset of the pandemic, and associated with time constraints, limited personnel resources, and high rates of illness among infected staff, perceived exhaustion, and pandemic fatigue [6,8]. Many studies used nonprobability purposive or convenience sampling techniques, which are associated with a higher risk of bias. This might also be linked to the ad hoc nature of the interventions, which were implemented in response to the pandemic. Further, the high risk of bias in most studies weakens confidence in the findings. NRSIs are prone to confounding. For instance, participants may already possess a higher level of knowledge about COVID-19 due to extensive media coverage or a general interest in acquiring more information. In uncontrolled before-and-after studies, we cannot with certainty attribute observed effects to the intervention, as there is no control group.

None of the included studies were based on a theoretical framework of HL, aligning with findings of previous studies [7,37,191]. It appears that COVID-19 HL has more often been addressed in cross-sectional than in interventional studies [7,37]. Baumeister et al [191] concluded that the integrated HL model [20] might not be as familiar with the model as desired among researchers. More than half of the studies were conducted in Asia and North America, regions where the model might be less known, as it was developed in Europe [191,192].

Further, the identified body of evidence revealed conceptual overlap between COVID-19–related HL and educational or skills-based interventions in the context of COVID-19 prevention and infection control. Most included studies were not designed as explicit HL interventions. Instead, COVID-19–related HL was frequently operationalized indirectly through knowledge- and skills-related indicators rather than validated HL instruments. These indicators were relevant to the present review as they could be mapped onto specific HL facets, particularly “understand” and “apply.” However, they should not be interpreted as equivalent to the full construct of COVID-19–related HL. None of the studies examined all four facets of HL, and the facets relating to “access” and “critical appraisal” were not explicitly addressed. The available evidence should therefore be understood as HL-relevant, but only partly HL-specific. Nevertheless, the theoretical foundation remains essential for systematically developing, comparing, and refining interventions [193]. Most of the studies investigated socio-cognitive outcomes combined, such as knowledge, attitudes, and practical skill acquisition, which align more closely with the Knowledge, Attitudes, and Practices framework [194].

It should also be noted that the intended outcomes were primarily assessed subjectively through self-report. Self-assessments of participants pertained to self-reported judgments about how well a person perceives themselves to know COVID-19–specific health information. This often leads to over- or underestimation of their own knowledge or confidence, and can also evoke social desirability bias, which might affect confidence in the results [195]. Miller [196] highlighted the significant impact of measurement moderators on the effect sizes of HL interventions. Their findings indicated that effect sizes, and consequently the observed improvements, were greater when HL was assessed subjectively rather than through objective measures by performance or achievement indices [196].

Moreover, the constructs used in the instruments were often merged, making it unclear which specific outcome was being measured, potentially leading to over- or underestimation of the results within the study. In 21 studies, a mismatch between the reported intervention content and the measured outcomes was observed. This finding is also consistent with the results of Walters et al [193]. Therefore, uncertainty regarding composite outcomes cannot be ruled out.

In more than half of the studies (60 studies), new questionnaires were developed; however, 26 studies reported the development process. Limited information regarding the development process and the origin of the items was observed. These findings compare with those of Grepmeier et al [7]. The lack of detailed information could be attributed to the need for implementing and conducting interventions rapidly, given the urgency to quickly develop and foster pandemic-related competencies and knowledge. Instruments for assessing COVID-19–related HL in respective countries may not yet be widely available or accessible in various languages. Currently, only a few instruments address COVID-19 HL, such as the validated tools developed in Europe for the general population [17,197] and specifically for HPs, such as the “COVID-19 related health literacy in health professionals” [43].

Regarding the delivery mode, most of the interventions were conducted through online distance learning, likely due to the circumstances and contact restrictions during the pandemic. Nevertheless, this demonstrates how quickly the implementation of interventions was adapted to rapidly changing external conditions.

Although the results indicate that COVID-19 HL interventions are beneficial for selected outcomes, individuals with high levels of self-reported and observable skills may still encounter real-world challenges in applying these competencies. This can occur, for example, in unfamiliar environments or when interacting with individuals they find intimidating, such as those in hierarchically higher-ranking positions [35]. Whether HPs are able to apply what they have learned over a longer period and translate it into actual behavior cannot be confirmed based on the available data. Consistent with similar studies [190,191], the outcomes were primarily measured immediately or shortly

XSL•FO
RenderX

after the intervention was conducted. No long-term follow-ups, extending up to one year, were reported. This may be due to the urgency and relevance of rapidly providing interventional effects to the scientific community. As a result, it remains unclear to what extent COVID-19–related HL interventions have long-term sustainability.

Limitations and Strengths

This review has several limitations. Substantial heterogeneity in interventions, comparators, settings, HP occupations, and outcomes limited the comparability between studies. Therefore, we were not able to state the effect of the interventions separated by delivery mode (eg, in-person and online distance learning), and delivery methods (eg, interactive and self-directed), intervention types, or specific occupational HP groups (eg, nurses, physicians, and medical students). Given the limited comparability across studies and substantial heterogeneity, the meta-analyses were conducted in an exploratory manner. Accordingly, the observed effects should be interpreted in light of the high risk of bias and low to very low certainty of evidence according to GRADE, and therefore interpreted with caution and regarded as preliminary evidence rather than definitive conclusions.

Moreover, for certain outcomes, a meta-analysis was not considered appropriate due to substantial differences in outcome metrics across studies. Consequently, the GRADE assessment could only be applied when meta-analytical results were available, and the domain of inconsistency could not be evaluated for these outcomes. The findings for these outcomes were reported narratively. We encountered challenges in contextualizing the results and the effect sizes within the existing literature, as there are currently only a few comparable reviews that report a meta-analysis based on HL interventions [190,191,198,199].

Further, most included studies did not explicitly aim to enhance the full construct of COVID-19–related HL. Instead, most studies assessed knowledge- and skills-based outcomes that were interpreted in this review as HL-related indicators. As a result, the review captures an evidence base that is closely related to HL, but not consistently grounded in explicit HL theory or measurement. In terms of the meta-analysis, it was not possible to determine publication bias due to the small number of studies included (≤10 studies). Additionally, a considerable proportion of effect sizes required conversion, and in certain cases, the assumptions underlying conversion could only be partly verified due to limited data availability in the primary studies.

A further limitation is the use of Microsoft Excel for data extraction. Although this program is commonly used for extracting data, it lacks the automated error-checking and audit functions of specialized systematic review software. To minimize potential error, data extraction was performed by two reviewers; one reviewer extracted relevant data, and the second reviewer checked the data extraction table for accuracy.

While these limitations should be considered, the present review also offers notable strengths. To our knowledge, this is the first study systematically synthesizing and summarizing evidence on interventions aiming to increase COVID-19–related HL in HPs. We ensured a comprehensive synthesis approach by searching 15 databases and registries, and including NRSIs and uncontrolled before-and-after studies in addition to RCT evidence, while closely adhering to guidelines for systematic reviews [42]. We followed the integrated model by Sørensen et al [20] as a framework for grouping and interpreting outcomes and intervention approaches. Further, we used validated assessment tools for rating risk of bias [47-49] and conducting synthesis in light of the certainty of the evidence by using GRADE [50]. The interpretation and reporting of the meta-analytic findings throughout this review were aligned with the corresponding GRADE assessments.

Implications

Our findings imply a need for further measures to be taken. Future intervention studies should be based on established HL concepts as a theoretical framework for designing and implementing interventions to promote COVID-19–related HL as an essential key competency in HPs on the individual level. Existing models, such as the conceptualization of HL by Sørensen et al [20], the Medical Research Council framework [200], or the Optimizing HL and Access approach [201], offer valuable foundations for the systematic development and adaptation of HL interventions. Further, as Nutbeam and Lloyd [35] concluded, a consolidated standard curriculum for enhancing HL among HPs would be essential to adapt and tailor methods in response to crises such as the pandemic. A toolbox of established targeted educational and practice-based approaches for communicable diseases might guide researchers and clinicians in developing and implementing HL interventions rapidly in acute situations. For instance, the facet “understand” was frequently assessed through performance tests or self-reported questionnaires in order to capture participants’ COVID-19 knowledge. A more effective method for assessing whether participants have understood the content of a training program might be using the widely applied communication technique “teach-back,” where participants repeat in their own words what they have understood [35]. The “apply” facet may be better evaluated through observations conducted by a trained reviewer, using a predeveloped checklist (reported in 17 studies) compared to the often used self-assessments.

Moreover, the use of validated instruments available in various languages to evaluate changes in COVID-19–related HL is emphasized. Further, to monitor changes over time and capture long-term effects, future trials should incorporate repeated follow-up assessments after the intervention for at least 12 months after the intervention.

This review focused on the individual HL of HPs. The representation of professional HL in the context of the pandemic was not within the scope of this work. Beyond individual skills, clinicians with higher COVID-19–related HL, and thus a deeper understanding of COVID-19 and its related guidelines, may be better positioned to make informed decisions regarding patient management. This includes recognizing symptoms, knowing when to increase care, advising on appropriate treatments, answering patient questions effectively, and providing clear instructions on infection prevention and control measures. This

XSL•FO
RenderX

is especially critical in environments in which misinformation and frequently changing guidelines can create uncertainty for both providers and patients [30,202]. Moreover, HPs are among others the primary sources of information for patients and the public [30,203]. Therefore, the level of COVID-19–related HL among HPs may also influence their surroundings, including the behaviors and attitudes of patients.

In this context, professional HL plays a crucial role, as it extends beyond individual competencies. Professional HL refers to the ability to communicate health information to patients in an understandable manner and to act on this information [204]. Pandemic-related HL interventions should not only focus on enhancing individual skills but also address professional HL, as HPs play a key role in promoting HL within the general population. Their transferable skills can significantly contribute to improving public HL [35,204].

HL-targeted interventions may empower individuals to enhance their competencies regarding infection control measures, enabling them to navigate the vast amount of pandemic-related information more effectively. This is particularly important as the pandemic transitions into an endemic phase, where ongoing community transmission necessitates sustained health education efforts. Given the ongoing presence of SARS-CoV-2 and the anticipated infection rates due to seasonal factors [205], we assume that the evidence landscape from COVID-19–related HL interventions continues to evolve. The increased number of records identified through database searches between the initial and updated searches (Figure 1) supports this assumption. Therefore, we aim to continuously identify and integrate new evidence and thus transform this systematic review into a living systematic review.

Conclusions

A total of 15 studies investigated the effect of interventions relevant to COVID-19–related HL among HPs. However, the identified evidence base was heterogeneous in terms of intervention content, delivery, study populations, and outcomes, limiting the comparability and the synthesis of effects. Most included studies evaluated short-term educational or skills-based infection control interventions and operationalized HL indirectly through knowledge- and skills-related indicators rather than explicit, theory-based HL measures.

Overall, the identified interventions might support HPs' short-term competencies relevant to infection prevention and control measures. Although some interventions showed large effects on specific outcomes, the evidence remains uncertain due to small sample sizes, high risk of bias, heterogeneity, imprecision, and limited theoretical grounding. In the absence of long-term follow-up data, the sustainability of observed effects remains unclear. The available evidence is closely related to HL, but not fully representative of explicitly HL-based interventions. Interventions explicitly designed to enhance the full construct of COVID-19–related HL remain largely absent. However, high-quality RCTs with adequate statistical power are needed to advance the current understanding of how interventions can strengthen HL in the context of communicable diseases, including COVID-19.

Acknowledgments

We used the machine learning artificial intelligence Rayyan for conducting the title, abstract, and full-text screening. For correcting grammar and spelling, we used ChatGPT and DeepL. After using these tools, the authors reviewed and edited the content as needed and take full responsibility for the content of this publication.

Funding

This systematic review is supported by the German Research Foundation (DFG, Deutsche Forschungsgemeinschaft), partially funded as subprojects of DFG Research Group FOR 2959 (HELICAP), AP 253/3-1. The HELICAP research group (project number 409800133) is an association of leading scientists at six locations in Germany. In addition to the Otto von Guericke University Magdeburg, the University of Education Freiburg, the University Medical Centre Mainz, the Brandenburg Medical School, the Hannover Medical School, and the University of Regensburg are involved in the interdisciplinary HELICAP research group. The members of the HELICAP steering committee are Prof Dr Christian Apfelbacher, Prof Dr Eva Maria Bitzer, Prof Dr Barbara Fillenberg, Prof Dr Christine Holmberg, Dr Julia von Sommoggy, Dr Jonas Lander, and PD Dr Susanne Brandstetter. The funder had no involvement in this study's design, data collection, analysis, interpretation, or the writing of this paper.

Authors' Contributions

CJA and UM conceived this study. CJA, UM, and CH developed the search strategy, CH performed the literature search, and CH, UM, AS, EMS, LK, EIG, and ML screened titles and abstracts. CH, AS, EMS, LK, EIG, and ML conducted full-text screening, and CH, AS, EIG, ML, and JS extracted the data. CH, PK, AS, EIG, JS, and ML conducted risk of bias assessments. PK and AS conducted the meta-analyses. CH, AS, EMS and PK wrote and edited the manuscript draft. All authors reviewed the paper for important intellectual content. All authors read and approved this paper.

Conflicts of Interest

None declared.

XSL•FO
RenderX

Multimedia Appendix 1
Search strategies.
[PDF File, 113 KB - mededu_v12i1e70400_app1.pdf]

Multimedia Appendix 2
Characteristics of studies.
[PDF File, 946 KB - mededu_v12i1e70400_app2.pdf]

Multimedia Appendix 3
Risk of bias assessments.
[PDF File, 335 KB - mededu_v12i1e70400_app3.pdf]

Multimedia Appendix 4
Summary of findings tables.
[PDF File, 588 KB - mededu_v12i1e70400_app4.pdf]

Multimedia Appendix 5
Narrative analyses of primary outcomes in uncontrolled before-and-after studies.
[PDF File, 89 KB - mededu_v12i1e70400_app5.pdf]

Checklist 1
PRISMA Checklist.
[PDF File, 67 KB - mededu_v12i1e70400_app6.pdf]

Checklist 2
PRISMA Checklist Abstract.
[PDF File, 41 KB - mededu_v12i1e70400_app7.pdf]

References

1. West R, Michie S, Rubin GJ, Amlôt R. Applying principles of behaviour change to reduce SARS-CoV-2 transmission. Nat Hum Behav 2020 May;4(5):451-459. [doi: 10.1038/s41562-020-0887-9] [Medline: 32377018]
2. Ghio D, Lawes-Wickwar S, Tang MY, et al. What influences people's responses to public health messages for managing risks and preventing infectious diseases? A rapid systematic review of the evidence and recommendations. BMJ Open 2021 Nov 11;11(11):e048750. [doi: 10.1136/bmjopen-2021-048750] [Medline: 34764167]
3. Recommendations for national SARS-cov-2 testing strategies and diagnostic capacities: interim guidance. World Health Organization. 2021 Jun 25. URL: https://apps.who.int/iris/bitstream/handle/10665/342002/WHO-2019-nCoV-lab-testing-2021.1-eng.pdf?sequence=1&isAllowed=y [accessed 2026-6-18]
4. Bundesministerium für Gesundheit, Robert Koch-Institut. Aktualisierung der Nationalen Teststrategie und Neuverkündung der Verordnung zum Anspruchauf Testung in Bezug auf einen direkten Erregernachweis des Coronavirus SARS-CoV-2 (Coronavirus-Testverordnung – TestV) [Update to the National Testing Strategy and Re-promulgation of the Regulation on the Right to Testing for Direct Detection of the SARS-CoV-2 Coronavirus (Coronavirus Testing Regulation – TestV)]. Epidemiologisches Bull 2020 Oct 22;43:3-6. [doi: 10.25646/7189]
5. Chu DK, Duda S, Solo K, Yaacoub S, Schunemann H. Physical distancing, face masks, and eye protection to prevent person-to-person transmission of SARS-CoV-2 and COVID-19: a systematic review and meta-analysis. J Vasc Surg 2020 Oct;72(4):1500. [doi: 10.1016/j.jvs.2020.07.040]
6. Gordon A, Lyons C, Rao S, Finoli L. Health care workers' challenges in the care of a COVID-19 patient. Crit Care Nurs Q 2020;43(4):400-406. [doi: 10.1097/CNQ.0000000000000325] [Medline: 32833776]
7. Grepmeier EM, Pawellek M, Curbach J, et al. Health literacy in health professionals two years into the COVID-19 pandemic: results from a scoping review. JMIR Med Educ 2022 Oct 17;8(4):e39023. [doi: 10.2196/39023] [Medline: 36179148]
8. Gualano MR, Sinigaglia T, Lo Moro G, et al. The burden of burnout among healthcare professionals of intensive care units and emergency departments during the COVID-19 pandemic: a systematic review. Int J Environ Res Public Health 2021 Aug 2;18(15):8172. [doi: 10.3390/ijerph18158172] [Medline: 34360465]
9. Classifying health workers: mapping occupations to the international standard classification. World Health Organization. 2019 Jul 31. URL: https://www.who.int/publications/m/item/classifying-health-workers [accessed 2026-06-18]
10. Bauer A, Grienberger K, Matthes B, Jucknewitz R, Schramm A. Berufe und COVID-19-Pandemie: wie hoch ist das berufsspezifische Ansteckungsrisiko? [occupations and the COVID-19 pandemic: how high is the occupation-specific risk

XSL•FO
RenderX

of infection?] [article in german]. : IAB-Forum; 2021 Sep 21 URL: https://www.iab-forum.de/berufe-und-covid-19-pandemie-wie-hoch-ist-das-berufsspezifische-ansteckungsrisiko/ [accessed 2026-06-30]
11. Buselli R, Corsi M, Baldanzi S, et al. Professional quality of life and mental health outcomes among health care workers exposed to Sars-Cov-2 (COVID-19). Int J Environ Res Public Health 2020 Aug 26;17(17):6180. [doi: 10.3390/ijerph17176180] [Medline: 32858810]
12. Thompson AK, Faith K, Gibson JL, Upshur REG. Pandemic influenza preparedness: an ethical framework to guide decision-making. BMC Med Ethics 2006 Dec 4;7(1):E12. [doi: 10.1186/1472-6939-7-12] [Medline: 17144926]
13. McConnell D. Balancing the duty to treat with the duty to family in the context of the COVID-19 pandemic. J Med Ethics 2020 Jun;46(6):360-363. [doi: 10.1136/medethics-2020-106250]
14. Ulrich CM, Rushton CH, Grady C. Nurses confronting the coronavirus: challenges met and lessons learned to date. Nurs Outlook 2020;68(6):838-844. [doi: 10.1016/j.outlook.2020.08.018] [Medline: 33097227]
15. Tian C, Lovrics O, Vaisman A, et al. Risk factors and protective measures for healthcare worker infection during highly infectious viral respiratory epidemics: a systematic review and meta-analysis. Infect Control Hosp Epidemiol 2022 May;43(5):639-650. [doi: 10.1017/ice.2021.18] [Medline: 33487203]
16. Paakkari L, Okan O. COVID-19: health literacy is an underestimated problem. Lancet Public Health 2020 May;5(5):e249-e250. [doi: 10.1016/S2468-2667(20)30086-4] [Medline: 32302535]
17. Okan O, Bollweg TM, Berens EM, Hurrelmann K, Bauer U, Schaeffer D. Coronavirus-related health literacy: a cross-sectional study in adults during the COVID-19 infodemic in Germany. Int J Environ Res Public Health 2020 Jul 30;17(15):5503. [doi: 10.3390/ijerph17155503] [Medline: 32751484]
18. Abel T, McQueen D. Critical health literacy and the COVID-19 crisis. Health Promot Int 2020 Dec 1;35(6):1612-1613. [doi: 10.1093/heapro/daaa040] [Medline: 32239213]
19. Sentell T, Vamos S, Okan O. Interdisciplinary perspectives on health literacy research around the world: more important than ever in a time of COVID-19. Int J Environ Res Public Health 2020 Apr 26;17(9):3010. [doi: 10.3390/ijerph17093010] [Medline: 32357457]
20. Sørensen K, Van den Broucke S, Fullam J, et al. Health literacy and public health: a systematic review and integration of definitions and models. BMC Public Health 2012 Jan 25;12:80. [doi: 10.1186/1471-2458-12-80] [Medline: 22276600]
21. Okan O, de Sombre S, Hurrelmann K, et al. Gesundheitskompetenz der Bevölkerung im Umgang mit der Coronavirus-Pandemie [Article in German]. Monitor Versorgungsforsch 2020:40-45 [FREE Full text]
22. Okan O, Bollweg TM, Bauer U, Hurrelmann K, Janner C, Schaeffer D. Trendstudie zur coronaspezifischen Gesundheitskompetenz: Ergebnisse der zweiten Erhebung der HLS-COVID-19 Studie [Book in German]. : Universität Bielefeld, Interdisziplinäres Zentrum für Gesundheitskompetenzforschung; 2021. [doi: 10.4119/unibi/2950307]
23. Paasche-Orlow MK, Parker RM, Gazmararian JA, Nielsen-Bohlman LT, Rudd RR. The prevalence of limited health literacy. J Gen Intern Med 2005 Feb;20(2):175-184. [doi: 10.1111/j.1525-1497.2005.40245.x] [Medline: 15836552]
24. Wagner CV, Knight K, Steptoe A, Wardle J. Functional health literacy and health-promoting behaviour in a national sample of British adults. J Epidemiol Community Health 2007 Dec;61(12):1086-1090. [doi: 10.1136/jech.2006.053967]
25. Eichler K, Wieser S, Brügger U. The costs of limited health literacy: a systematic review. Int J Public Health 2009;54(5):313-324. [doi: 10.1007/s00038-009-0058-2] [Medline: 19644651]
26. Schaeffer D, Berens EM, Vogt D. Health literacy in the German population. Dtsch Arztebl Int 2017 Jan 27;114(4):53-60. [doi: 10.3238/arztebl.2017.0053] [Medline: 28211318]
27. Schaeffer D, Berens EM, Gille S, et al. Gesundheitskompetenz der Bevölkerung in Deutschland – vor und während der Corona Pandemie: Ergebnisse des HLS-GER 2 [Book in German].: Universität Bielefeld, Interdisziplinäres Zentrum für Gesundheitskompetenzforschung; 2021. [doi: 10.4119/unibi/2950305]
28. Castro-Sánchez E, Alexander CM, Atchison C, et al. Evaluation of a personal protective equipment support programme for staff during the COVID-19 pandemic in London. J Hosp Infect 2021 Mar;109:68-77. [doi: 10.1016/j.jhin.2020.12.004] [Medline: 33307145]
29. Nutbeam D, McGill B, Premkumar P. Improving health literacy in community populations: a review of progress. Health Promot Int 2018 Oct 1;33(5):901-911. [doi: 10.1093/heapro/dax015] [Medline: 28369557]
30. Zarocostas J. How to fight an infodemic. Lancet 2020 Feb 29;395(10225):676. [doi: 10.1016/S0140-6736(20)30461-X] [Medline: 32113495]
31. Nordheim LV, Gundersen MW, Espehaug B, Guttersrud Ø, Flottorp S. Effects of school-based educational interventions for enhancing adolescents abilities in critical appraisal of health claims: a systematic review. PLoS One 2016;11(8):e0161485. [doi: 10.1371/journal.pone.0161485] [Medline: 27557129]
32. Taggart J, Williams A, Dennis S, et al. A systematic review of interventions in primary care to improve health literacy for chronic disease behavioral risk factors. BMC Fam Pract 2012 Dec;13(1):49. [doi: 10.1186/1471-2296-13-49]
33. Cheung J, Chan CY, Cheng HY. The effectiveness of interventions on improving the mental health literacy of health care professionals in general hospitals: a systematic review of randomized controlled trials. J Am Psychiatr Nurses Assoc 2024;30(3):465-479. [doi: 10.1177/10783903231194579] [Medline: 37615199]
34. Brørs G, Larsen MH, Hølvold LB, Wahl AK. eHealth literacy among hospital health care providers: a systematic review. BMC Health Serv Res 2023;23(1):1144. [doi: 10.1186/s12913-023-10103-8]

XSL•FO
RenderX

35. Nutbeam D, Lloyd JE. Understanding and responding to health literacy as a social determinant of health. Annu Rev Public Health 2021 Apr 1;42(159-173):159-173. [doi: 10.1146/annurev-publhealth-090419-102529] [Medline: 33035427]
36. Matterne U, Egger N, Tempes J, et al. Health literacy in the general population in the context of epidemic or pandemic coronavirus outbreak situations: rapid scoping review. Patient Educ Couns 2021 Feb;104(2):223-234. [doi: 10.1016/j.pec.2020.10.012] [Medline: 33109429]
37. Seng JJB, Yeam CT, Huang CW, Tan NC, Low LL. Pandemic-related health literacy: a systematic review of literature in COVID-19, SARS and MERS pandemics. Singapore Med J 2025 May 1;66(5):244-255. [doi: 10.4103/singaporemedj.SMJ-2021-026] [Medline: 37459004]
38. Page MJ, McKenzie JE, Bossuyt PM, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ 2021 Mar 29;372:n71. [doi: 10.1136/bmj.n71] [Medline: 33782057]
39. PICO search strategy tips & examples. Critical Appraisal Skills Programme. 2024 Jun 26. URL: https://casp-uk.net/news/pico-search-strategy/ [accessed 2026-06-18]
40. Avau B, Van Remoortel H, De Buck E. Translation and validation of PubMed and Embase search filters for identification of systematic reviews, intervention studies, and observational studies in the field of first aid. J Med Libr Assoc 2021 Oct 1;109(4):599-608. [doi: 10.5195/jmla.2021.1219] [Medline: 34858089]
41. Hausner E, Metzendorf MI, Richter B, Lotz F, Waffenschmidt S. Study filters for non-randomized studies of interventions consistently lacked sensitivity upon external validation. BMC Med Res Methodol 2018 Dec 18;18(1):171. [doi: 10.1186/s12874-018-0625-4] [Medline: 30563471]
42. Higgins JPT, Thomas J, Chandler J, Cumpston M, Li T, Page MJ. In: Welch VA, editor. Cochrane Handbook for Systematic Reviews of Interventions Version 65 (Updated August 2024): Cochrane; 2024. URL: https://www.cochrane.org/authors/handbooks-and-manuals/handbook [accessed 2026-06-18]
43. Hiltrop K, Hiebel N, Geiser F, et al. Measuring COVID-19 related health literacy in healthcare professionals-psychometric evaluation of the HL-COV-HP instrument. Int J Environ Res Public Health 2021 Nov 14;18(22):11959. [doi: 10.3390/ijerph182211959] [Medline: 34831720]
44. Sterne JAC, Hernán MA, McAleenan A, Reeves BC, Higgins JPT. Assessing risk of bias in a non-randomized study. In: Higgins JPT, Thomas J, Chandler J, Cumpston M, Li T, Page MJ, et al, editors. Cochrane Handbook for Systematic Reviews of Interventions Version 65. 2024. URL: https://www.cochrane.org/authors/handbooks-and-manuals/handbook/current/chapter-25 [accessed 2026-06-18]
45. Thomas J, Graziosi S, Brunton J, et al. EPPI-reviewer: advanced software for systematic reviews, maps and evidence synthesis. : EPPI Centre, UCL Social Research Institute, University College; 2023 URL: https://eppi.ioe.ac.uk/cms/er4 [accessed 2026-06-30]
46. Ouzzani M, Hammady H, Fedorowicz Z, Elmagarmid A. Rayyan-a web and mobile app for systematic reviews. Syst Rev 2016 Dec 5;5(1):210. [doi: 10.1186/s13643-016-0384-4] [Medline: 27919275]
47. Sterne JAC, Savović J, Page MJ, et al. RoB 2: a revised tool for assessing risk of bias in randomised trials. BMJ 2019;366:l4898. [doi: 10.1136/bmj.l4898]
48. Sterne JA, Hernán MA, Reeves BC, et al. ROBINS-I: a tool for assessing risk of bias in non-randomised studies of interventions. BMJ 2016;355:i4919. [doi: 10.1136/bmj.i4919]
49. Armijo-Olivo S, Stiles CR, Hagen NA, Biondo PD, Cummings GG. Assessment of study quality for systematic reviews: a comparison of the Cochrane Collaboration Risk of Bias Tool and the Effective Public Health Practice Project Quality Assessment Tool: methodological research. J Eval Clin Pract 2012 Feb;18(1):12-18. [doi: 10.1111/j.1365-2753.2010.01516.x] [Medline: 20698919]
50. Balshem H, Helfand M, Schünemann HJ, et al. GRADE guidelines: 3. rating the quality of evidence. J Clin Epidemiol 2011 Apr;64(4):401-406. [doi: 10.1016/j.jclinepi.2010.07.015]
51. Schünemann HJ, Brennan S, Akl EA, et al. The development methods of official GRADE articles and requirements for claiming the use of GRADE – a statement by the GRADE guidance group. J Clin Epidemiol 2023 Jul;159:79-84. [doi: 10.1016/j.jclinepi.2023.05.010] [Medline: 37211327]
52. Wan X, Wang W, Liu J, Tong T. Estimating the sample mean and standard deviation from the sample size, median, range and/or interquartile range. BMC Med Res Methodol 2014 Dec 19;14:135. [doi: 10.1186/1471-2288-14-135] [Medline: 25524443]
53. Hedges LV, Olkin I. Statistical Methods for Meta-Analysis: Elsevier; 1985.
54. Van den Noortgate W, López-López JA, Marín-Martínez F, Sánchez-Meca J. Three-level meta-analysis of dependent effect sizes. Behav Res 2013 Jun;45(2):576-594. [doi: 10.3758/s13428-012-0261-6]
55. Viechtbauer W. Bias and efficiency of meta-analytic variance estimators in the random-effects model. J Educ Behav Stat 2005 Sep;30(3):261-293. [doi: 10.3102/10769986030003261]
56. IntHout J, Ioannidis JPA, Borm GF. The Hartung-Knapp-Sidik-Jonkman method for random effects meta-analysis is straightforward and considerably outperforms the standard DerSimonian-Laird method. BMC Med Res Methodol 2014 Feb 18;14(25):25. [doi: 10.1186/1471-2288-14-25] [Medline: 24548571]
57. Nagashima K, Noma H, Furukawa TA. Prediction intervals for random-effects meta-analysis: a confidence distribution approach. Stat Methods Med Res 2019 Jun;28(6):1689-1702. [doi: 10.1177/0962280218773520] [Medline: 29745296]

XSL•FO
RenderX

58. Egger M, Davey Smith G, Schneider M, Minder C. Bias in meta-analysis detected by a simple, graphical test. BMJ 1997 Sep 13;315(7109):629-634. [doi: 10.1136/bmj.315.7109.629] [Medline: 9310563]
59. Balduzzi S, Rücker G, Schwarzer G. How to perform a meta-analysis with R: a practical tutorial. Evid Based Mental Health 2019 Nov;22(4):153-160. [doi: 10.1136/ebmental-2019-300117]
60. Viechtbauer W. Conducting meta-analyses in R with the metafor package. J Stat Softw 2010;36(3):1-48. [doi: 10.18637/jss.v036.i03]
61. Alotaibi N, Al-Sayegh N, Nadar M, Shayea A, Allafi A, Almari M. Investigation of health science students' knowledge regarding healthy lifestyle promotion during the spread of COVID-19 pandemic: a randomized controlled trial. Front Public Health 2021;9:774678. [doi: 10.3389/fpubh.2021.774678] [Medline: 34820354]
62. Birrenbach T, Zbinden J, Papagiannakis G, et al. Effectiveness and utility of virtual reality simulation as an educational tool for safe performance of COVID-19 diagnostics: prospective, randomized pilot trial. JMIR Serious Games 2021 Oct 8;9(4):e29586. [doi: 10.2196/29586] [Medline: 34623315]
63. Christensen L, Rasmussen CS, Benfield T, Franc JM. A randomized trial of instructor-led training versus video lesson in training health care providers in proper donning and doffing of personal protective equipment. Disaster Med Public Health Prep 2020 Aug;14(4):514-520. [doi: 10.1017/dmp.2020.56] [Medline: 32223776]
64. Currat L, Suppan M, Gartner BA, et al. Impact of face-to-face teaching in addition to electronic learning on personal protective equipment doffing proficiency in student paramedics: randomized controlled trial. Int J Environ Res Public Health 2022 Mar 5;19(5):3077. [doi: 10.3390/ijerph19053077] [Medline: 35270768]
65. Jafree SR, Zakar R, Rafiq N, et al. WhatsApp-delivered intervention for continued learning for nurses in Pakistan during the COVID-19 pandemic: results of a randomized-controlled trial. Front Public Health 2022;10:739761. [doi: 10.3389/fpubh.2022.739761] [Medline: 35242728]
66. Jeihooni AK, Namdari A, Kashfi SM, Kamyab A, Harsini PA, Rakhshani T. Effects of an educational intervention based on the health belief model on COVID-19 preventive behaviors among health personnel in Abadan, Iran. J Public Health (Berl) 2025 Apr;33(4):769-779. [doi: 10.1007/s10389-023-02053-6]
67. Li Y, Wang Y, Li Y, et al. Comparison of repeated video display vs combined video display and live demonstration as training methods to healthcare providers for donning and doffing personal protective equipment: a randomized controlled trial. Risk Manage Healthcare Policy 2020;13:2325-2335. [doi: 10.2147/RMHP.S267514] [Medline: 33154684]
68. Manggala SK, Tantri AR, Sugiarto A, Sianipar IR, Prasetyono TOH. In situ simulation training for a better interprofessional team performance in transferring critically ill patients with COVID-19: a prospective randomised control trial. Postgrad Med J 2022 Aug;98(1162):617-621. [doi: 10.1136/postgradmedj-2021-141426] [Medline: 35101969]
69. Rueda-Medina B, Aguilar-Ferrándiz ME, Esteban-Burgos AA, et al. Impact of non-face-to-face teaching with passive training on personal protective equipment use in health science students: a randomized controlled trial. Int J Environ Res Public Health 2022 Oct 10;19(19):12981. [doi: 10.3390/ijerph191912981] [Medline: 36232282]
70. Suppan L, Abbas M, Stuby L, et al. Effect of an e-learning module on personal protective equipment proficiency among prehospital personnel: web-based randomized controlled trial. J Med Internet Res 2020 Aug 21;22(8):e21265. [doi: 10.2196/21265] [Medline: 32747329]
71. Suppan L, Stuby L, Gartner B, et al. Impact of an e-learning module on personal protective equipment knowledge in student paramedics: a randomized controlled trial. Antimicrob Resist Infect Control 2020 Nov 10;9(1):185. [doi: 10.1186/s13756-020-00849-9] [Medline: 33168097]
72. Wang X, Zhou Y, Song Z, Wang Y, Chen X, Zhang D. Practical COVID-19 prevention training for obstetrics and gynecology residents based on the conceive–design–implement–operate framework. Front Public Health 2022 Mar 3;10:808084. [doi: 10.3389/fpubh.2022.808084]
73. Xie M, Zhou Q, Kang Y, et al. The skill training of resident anesthesiologists during the outbreak of COVID-19. Research Square. Preprint posted online on Jan 19, 2021. [doi: 10.21203/rs.3.rs-144144/v1]
74. Rakhshani T, Dolatkhah SM, Kashfi SM, Khani Jeihooni A. The effect of a self-learned virtual learning package on knowledge, attitude, and self-care behaviors of COVID-19 in people referred to health and treatment centers. BMC Public Health 2024;24(1):1710. [doi: 10.1186/s12889-024-19233-y]
75. Ansari A, Urooj U, Waseem M, Ihtasham A. Video based learning vs instructor led training for optimizing personal protective equipment use to prevent COVID-19 infection - a comparative study. J Pak Med Assoc 2022 May;72(5):807-810. [doi: 10.47391/JPMA.3359] [Medline: 35713035]
76. Buyego P, Katwesigye E, Kebirungi G, et al. Feasibility of virtual reality based training for optimising COVID-19 case handling in Uganda. BMC Med Educ 2022 Apr 13;22(1):274. [doi: 10.1186/s12909-022-03294-x] [Medline: 35418070]
77. Hu H, Xiao YY, Li H. The effectiveness of a serious game versus online lectures for improving medical students' coronavirus disease 2019 knowledge. Games Health J 2021 Apr;10(2):139-144. [doi: 10.1089/g4h.2020.0140] [Medline: 33434099]
78. Yu M, Yang MR. Effectiveness and utility of virtual reality infection control simulation for children with COVID-19: quasi-experimental study. JMIR Serious Games 2022 May 27;10(2):e36707. [doi: 10.2196/36707] [Medline: 35575818]
79. Abbas K, Nawaz SMA, Amin N, et al. A web-based health education module and its impact on the preventive practices of health-care workers during the COVID-19 pandemic. Health Educ Res 2020 Oct 1;35(5):353-361. [doi: 10.1093/her/cyaa034] [Medline: 32951026]

XSL•FO
RenderX

80. Altillo BSA, Gray M, Avashia SB, et al. Global health on the front lines: an innovative medical student elective combining education and service during the COVID-19 pandemic. BMC Med Educ 2021 Mar 27;21(1):186. [doi: 10.1186/s12909-021-02616-9] [Medline: 33773585]
81. Aqel O, Alqadheeb B, Felix M, et al. Cultivating COVID-19 vaccine confidence in pharmacy professionals. Pharmacy (Basel) 2023 Mar 7;11(2):50. [doi: 10.3390/pharmacy11020050] [Medline: 36961028]
82. Aujee DS, Aghamkar J, Yangad S, Salvi R. Assessment of educational intervention on knowledge regarding World Health Organization COVID-19 protocols during delivery among staff nurses working in labour room of selected hospitals. J Pharm Negat Results 2022;13:4717-4721. [doi: 10.47750/pnr.2022.13.S08.613]
83. Bakhsh A, Asiri R, Alotaibi H, Alsaeedi R, Shahbar R, Boker A. Rapid cycle training for non-critical care physicians to meet intensive care unit staff shortage at an academic training center in a developing country during the COVID-19 pandemic. BMC Med Educ 2023 Jul 5;23(1):493. [doi: 10.1186/s12909-023-04478-9] [Medline: 37403115]
84. Bayomi R, Taha N. Effect of teaching guidelines on knowledge, attitudes, and practices regarding COVID19 among the first year nursing students. Assiut Sci Nurs J 2021 Jun 1;9(25):38-47. [doi: 10.21608/asnj.2021.72355.1156]
85. Bechini A, Vannacci A, Salvati C, et al. Knowledge and training of Italian students in healthcare settings on COVID-19 vaccines and vaccination strategies, one year after the immunization campaign. J Prev Med Hyg 2023 Aug 1;64(2):E152-E160. [doi: 10.15167/2421-4248/jpmh2023.64.2.2934] [Medline: 37654864]
86. Bieri J, Tuor C, Nendaz M, et al. Implementation of a student-teacher–based blended curriculum for the training of medical students for nasopharyngeal swab and intramuscular injection: mixed methods pre-post and satisfaction surveys. JMIR Med Educ 2023 Mar 2;9:e38870. [doi: 10.2196/38870] [Medline: 36862500]
87. Blake H, Fecowycz A, Starbuck H, Jones W. COVID-19 vaccine education (CoVE) for health and care workers to facilitate global promotion of the COVID-19 vaccines. Int J Environ Res Public Health 2022 Jan 7;19(2):653. [doi: 10.3390/ijerph19020653] [Medline: 35055476]
88. Boccalini S, Vannacci A, Crescioli G, et al. Knowledge of university students in health care settings on vaccines and vaccinations strategies: impact evaluation of a specific educational training course during the COVID-19 pandemic period in Italy. Vaccines (Basel) 2022 Jul 6;10(7):1085. [doi: 10.3390/vaccines10071085] [Medline: 35891250]
89. Calik A, Cakmak B, Kapucu S, Inkaya B. The effectiveness of serious games designed for infection prevention and promotion of safe behaviors of senior nursing students during the COVID-19 pandemic. Am J Infect Control 2022 Dec;50(12):1360-1367. [doi: 10.1016/j.ajic.2022.02.025] [Medline: 35231565]
90. Clay J, Morton K, Franz D, Jaqua E, Nguyen V. Quality improvement for outpatient COVID-19 infection control. Cureus 2021 Jul;13(7):e16373. [doi: 10.7759/cureus.16373] [Medline: 34408928]
91. Díaz-Guio DA, Ricardo-Zapata A, Ospina-Velez J, Gómez-Candamil G, Mora-Martinez S, Rodriguez-Morales AJ. Cognitive load and performance of health care professionals in donning and doffing PPE before and after a simulation-based educational intervention and its implications during the COVID-19 pandemic for biosafety. Infez Med 2020 Jun 1;28(suppl 1):111-117. [Medline: 32532947]
92. Elasrag G, Elsabagh NE, Abdelmonem AF, Ahmed A. Impact of educational intervention on nurses' knowledge, practice and attitude related prevention measures of COVID 19. Indian J Forensic Med Toxicol 2021;15(3):2939-2948. [doi: 10.37506/ijfmt.v15i3.15751]
93. Etebarian A, Khoramian Tusi S, Momeni Z, Hejazi K. Impact of educational intervention regarding COVID-19 on knowledge, attitude, and practice of students before dental school re-opening. BMC Oral Health 2023 Mar 18;23(1):156. [doi: 10.1186/s12903-023-02845-y] [Medline: 36934228]
94. Findyartini A, Greviana N, Hanum C, et al. Supporting newly graduated medical doctors in managing COVID-19: an evaluation of a massive open online course in a limited-resource setting. PLoS One 2021;16(9):e0257039. [doi: 10.1371/journal.pone.0257039] [Medline: 34506524]
95. Fuentes GM, Carbajales León EB, Carbajales León AI. Educative intervention about COVID-19 in the medicine students from Joaquín de Agüero policlinic [Article in Spanish]. Rev Cienc Méd Habana 2020;27(3):366-376 [FREE Full text]
96. Girard H, Bosshard W, Krief H, Büla CJ. Effectiveness of information sessions about COVID-19 vaccines in healthcare professionals working in geriatrics. Gerontol Geriatr Med 2022;8:23337214221115235. [doi: 10.1177/23337214221115235] [Medline: 35911950]
97. Greaves SW, Alter SM, Ahmed RA, et al. A Simulation-based PPE orientation training curriculum for novice physicians. Infect Prev Pract 2023 Mar;5(1):100265. [doi: 10.1016/j.infpip.2022.100265] [Medline: 36536774]
98. Halemani K, Cheema M, Khatun S, et al. An effectiveness of training program on COVID-19 among healthcare students: a cross section study. IJRPS 2020;11(SPL1):1250-1254. [doi: 10.26452/ijrps.v11iSPL1.3613]
99. Han B, Zang F, Liu J, et al. Effect analysis of "four-step" training and assessment tool in the prevention and control of COVID-19. Infect Drug Resist 2022;15:1247-1257. [doi: 10.2147/IDR.S346559] [Medline: 35360372]
100. Hwang WJ, Lee J. Effectiveness of the infectious disease (COVID-19) simulation module program on nursing students: disaster nursing scenarios. J Korean Acad Nurs 2021 Dec;51(6):648-660. [doi: 10.4040/jkan.21164] [Medline: 35023855]
101. Instrum RS, Koch RW, Rocha T, et al. Improving nasopharyngeal swab technique via simulation for frontline workers. Laryngoscope 2023 Jan;133(1):38-42. [doi: 10.1002/lary.30034] [Medline: 35102548]

XSL•FO
RenderX

102. Kasapoğlu ES, Yıldız Y, Saldamlı A, Karaçetin F. The effect of COVID-19 patient care and emergency response interprofessional training on COVID-19 knowledge, perception, behavior and readiness for care. Work 2023;75(3):767-778. [doi: 10.3233/WOR-220227] [Medline: 36710711]
103. Khari S, Pazokian M, Salimi Akin Abadi A, Zarmehrparirouy M, Ahmadvand Y. The effect of e-learning program for COVID-19 patient care on the knowledge of nursing students: a quasi-experimental study. SAGE Open Nurs 2022;8:23779608221124421. [doi: 10.1177/23779608221124421] [Medline: 36147727]
104. Kobayashi D, Mami K, Fujishiro S, et al. Online training of COVID-19 infection prevention and control for healthcare workers in psychiatric institutes. BMC Psychiatry 2023 May 9;23(1):325. [doi: 10.1186/s12888-023-04826-5] [Medline: 37161463]
105. Kufel WD, Blaine BE, Avery LM. Pharmacy students' knowledge and confidence of COVID-19 following an interactive didactic class. J Am Coll Clin Pharm 2022 Oct;5(10):1082-1087. [doi: 10.1002/jac5.1678] [Medline: 35942356]
106. Li Z, Cheng J, Zhou T, Wang S, Huang S, Wang H. Evaluating a nurse training program in the emergency surgery department based on the Kirkpatrick's model and clinical demand during the COVID-19 pandemic. Telemed J E Health 2020 Aug;26(8):985-991. [doi: 10.1089/tmj.2020.0089] [Medline: 32357086]
107. Macht L, Worlitzsch D, Braijoshri N, et al. COVID-19: development and implementation of a video-conference-based educational concept to improve the hygiene skills of health and nursing professionals in the Republic of Kosovo. GMS Hyg Infect Control 2022;17:Doc09. [doi: 10.3205/dgkh000412] [Medline: 35707230]
108. Mark ME, LoSavio P, Husain I, Papagiannopoulos P, Batra PS, Tajudeen BA. Effect of implementing simulation education on health care worker comfort with nasopharyngeal swabbing for COVID-19. Otolaryngol Head Neck Surg 2020 Aug;163(2):271-274. [doi: 10.1177/0194599820933168] [Medline: 32482155]
109. Mektrirat R, Sathanawongs A, Tiwananthagorn S, et al. Achieving interprofessional education on collaborative problem-solving for COVID-19 using project-based approach. Int J Infect Dis 2022 Mar;116:S67. [doi: 10.1016/j.ijid.2021.12.158]
110. Mohamed Y, Hezeri P, Kama H, et al. Evaluation of an online training program on COVID-19 for health workers in Papua New Guinea. Trop Med Infect Dis 2023 Jun 19;8(6):327. [doi: 10.3390/tropicalmed8060327] [Medline: 37368745]
111. Naz F, Ohri P, Sharma A, Spandana BS, Gupta K. Impact of training on awareness of COVID-19 among the health care workers in a tertiary care hospital of Dehradun. Indian J Community Health 2022;34(1):20-25. [doi: 10.47203/IJCH.2022.v34i01.005]
112. Odusanya OO, Adeniran A, Bakare OQ, et al. Building capacity of primary health care workers and clients on COVID-19: results from a web-based training. PLoS One 2022;17(10):e0274750. [doi: 10.1371/journal.pone.0274750] [Medline: 36206255]
113. Otu A, Okuzu O, Effa E, et al. Training health workers at scale in Nigeria to fight COVID-19 using the InStrat COVID-19 tutorial app: an e-health interventional study. Ther Adv Infect 2021 Jan;8:20499361211040704. [doi: 10.1177/20499361211040704]
114. Otu A, Okuzu O, Ebenso B, et al. Introduction of mobile health tools to support COVID-19 training and surveillance in Ogun State Nigeria. Front Sustain Cities 2021;3:638278. [doi: 10.3389/frsc.2021.638278]
115. Perera N, Haldane V, Ratnapalan S, et al. Implementation of a coronavirus disease 2019 infection prevention and control training program in a low-middle income country. JBI Evid Implement 2022 Sep 1;20(3):228-235. [doi: 10.1097/XEB.0000000000000307] [Medline: 35170481]
116. Pokrajac N, Schertzer K, Poffenberger CM, et al. Mastery learning ensures correct personal protective equipment use in simulated clinical encounters of COVID-19. West J Emerg Med 2020 Jul 21;21(5):1089-1094. [doi: 10.5811/westjem.2020.6.48132] [Medline: 32970559]
117. Puga RR, Cardoso AL, Rodríguez OL. Educational intervention in medical students on post-COVID-19 disabling sequelae from the public health subject [Article in Spanish]. Rev Cuba Reumatol 2023;25(3) [FREE Full text]
118. Qasmi SA, Standley C, Mohsin S, Sarwar S, Malik L, Aziz F. Effectiveness of international virtual training on biorisk management in the context of COVID-19. Front Public Health 2022;10:888097. [doi: 10.3389/fpubh.2022.888097] [Medline: 36339241]
119. Roberts EN, Smithing RT, Tucker P. Measuring the impact of a COVID-19 continuing education program. J Am Assoc Nurse Pract 2022 Jun 1;34(6):835-843. [doi: 10.1097/JXX.0000000000000715] [Medline: 35353068]
120. Roberts KJ, Zumstein KK, Lamphere TR, et al. Improving students' knowledge and skills through a tele-ICU clinical rotation. Respir Care 2022 Jul;67(7):789-794. [doi: 10.4187/respcare.09896] [Medline: 35640997]
121. Rosas-Magallanes C, Basto-Abreu A, Barrientos-Gutiérrez T, et al. CLIMSS online platform as a health literacy tool during the health crisis of COVID-19. Salud Publica Mex 2022;64(3):320-327. [doi: 10.21149/13103] [Medline: 36130387]
122. Saati AA, Alkalash SH. Promotion of knowledge, attitude, and practice among medical undergraduates regarding infection control measures during COVID-19 pandemic. Front Public Health 2022;10:932465. [doi: 10.3389/fpubh.2022.932465] [Medline: 36187651]
123. Salehi R, de Young S, Asamoah A, et al. Evaluation of a continuing professional development strategy on COVID-19 for 10 000 health workers in Ghana: a two-pronged approach. Hum Resour Health 2023 Mar 6;21(1):18. [doi: 10.1186/s12960-023-00804-w] [Medline: 36879262]

XSL•FO
RenderX

124. Sangwan J, Lathwal S, Lohan K, Adlakha N, Mane P, Gole S. Impact of training on knowledge, attitude and perceived barriers for compliance regarding use of protective equipment kit among frontline healthcare workers during COVID-19 pandemic. J Clin Diagn Res 2022;16(1):JC17-JC21. [doi: 10.7860/JCDR/2022/50845.15913]
125. Shahrin L, Parvin I, Sarmin M, et al. In-person training on COVID-19 case management and infection prevention and control: evaluation of healthcare professionals in Bangladesh. PLoS One 2022;17(10):e0273809. [doi: 10.1371/journal.pone.0273809] [Medline: 36201257]
126. Sharma R, Mohanty A, Singh V, et al. Effectiveness of video-based online training for health care workers to prevent COVID-19 infection: an experience at a tertiary care level institute, Uttarakhand, India. Cureus 2021 May 1;13(5):e14785. [doi: 10.7759/cureus.14785] [Medline: 34094747]
127. Smith CR, Vasilopoulos T, Frantz AM, et al. Staying proper with your personal protective equipment: how to don and doff. J Clin Anesth 2023 Jun;86:111057. [doi: 10.1016/j.jclinane.2023.111057] [Medline: 36696834]
128. Ta'an WF, Al-Hammouri MM, Al-Faouri I, Suliman MM. The effectiveness of COPA-based training program on the infection-control competencies of newly hired healthcare professionals. Teach Learn Nurs 2023 Jan;18(1):160-165. [doi: 10.1016/j.teln.2022.06.009] [Medline: 36778202]
129. Tan W, Ye Y, Yang Y, et al. Whole-process emergency training of personal protective equipment helps healthcare workers against COVID-19. J Occup Environ Med 2020 Jun;62(6):420-423. [doi: 10.1097/JOM.0000000000001877] [Medline: 32469194]
130. Thakre SS, Jadhao AR, Dhoble MA, Dass R, Thakre SB, Somani A. Evaluation of effectiveness of COVID-19 training and assessment of anxiety among nurses of a tertiary health care centre during the corona virus pandemic- an experimental study. J Clin Diagn Res 2020;14(11):LC34-LC37. [doi: 10.7860/JCDR/2020/45464.14301]
131. Thakre SS, Thakre SB, Jadhao A, Dass R, Dhoble MA, Tiwari PN. Evaluation of effectiveness of COVID-19 training of tertiary health care workers. Int J Community Med Public Health 2020;7(7):2635. [doi: 10.18203/2394-6040.ijcmph20202989]
132. Tsiouris F, Hartsough K, Poimboeuf M, et al. Rapid scale-up of COVID-19 training for frontline health workers in 11 African countries. Hum Resour Health 2022;20(1):43. [doi: 10.1186/s12960-022-00739-8]
133. Wang SH, Yimer G, Bisesi M, et al. Rapid virtual training and field deployment for COVID-19 surveillance officers: experiences from Ethiopia. Pan Afr Med J 2022;43:23. [doi: 10.11604/pamj.2022.43.23.28787] [Medline: 36451715]
134. Zafar N, Jamal Z, Mujeeb Khan M. Preparedness of the healthcare personnel against the coronavirus disease 2019 (COVID-19) outbreak: an audit cycle. Front Public Health 2020;8:502. [doi: 10.3389/fpubh.2020.00502] [Medline: 33042944]
135. Zhao J, Rozenberg D, Kaul R, et al. The positive impact of a telemedicine education program on healthcare workers during the COVID-19 pandemic in Ontario, Canada. Ann Fam Med 2022 Apr 1;20(20 Suppl 1):3260. [doi: 10.1370/afm.20.s1.3260] [Medline: 36706423]
136. Brito-Brito PR, Fernández-Gutiérrez D, Martínez-Alberto CE, Sáez-Rodríguez MJ, Núñez-Marrero J, García-Hernández AM. Use of the nursing outcomes classification (NOC) to measure perceived knowledge about the control of SARS-CoV-2 infection: the impact of a training program in primary healthcare professionals. Int J Nurs Knowl 2022 Oct;33(4):259-269. [doi: 10.1111/2047-3095.12356] [Medline: 34939361]
137. Kharel R, Baird J, Vaishnav H, et al. Development and assessment of novel virtual COVID-19 trainer-of trainers course implemented by an academic–humanitarian partnership. Global Health Action 2022 Dec 31;15(1):2010391. [doi: 10.1080/16549716.2021.2010391] [Medline: 35006037]
138. Singh V, Supehia S, Gupta PK, et al. Effectiveness of video modules in infection control trainings during COVID-19 pandemic: a quasi-experimental study in tertiary care institute. J Educ Health Promot 2021;10(1):183. [doi: 10.4103/jehp.jehp_1009_20] [Medline: 34250117]
139. Stengel S. Piloting interventions (case conferences, training network) for adaptive, cross-sectoral health care long-/post-COVID in baden-württemberg - SEVEN - PCS württemberg - SEVEN - PCS - transfer. Cochrane Library. URL: https://www.cochranelibrary.com/central/doi/10.1002/central/CN-02604635/full [accessed 2026-06-15]
140. Leung A. A blended gaming COVID-19 training system (BGCTS) with WHO guidelines for staff in residential care homes. Cochrane Library. URL: https://www.cochranelibrary.com/central/doi/10.1002/central/CN-02249570/full [accessed 2026-06-15]
141. Johnston J. Effectiveness of interactive voice response for COVID-19 vaccination training in the Democratic Republic of the Congo. Cochrane Library. URL: https://www.cochranelibrary.com/central/doi/10.1002/central/CN-02352592/full [accessed 2026-06-15]
142. Yang L. Education program of hand hygiene for nursing students. Cochrane Library. URL: https://www.cochranelibrary.com/central/doi/10.1002/central/CN-02564780/full [accessed 2026-06-15]
143. Pala SC. Development and efficiency evaluation of a respiratory disease pandemics preparedness training program for pre-hospital emergency healthcare workers. 2022 Jan 14. URL: https://clinicaltrials.gov/study/NCT05192967 [accessed 2026-06-18]
144. Hanson B, Welch L, Frese M, Gallo C. 558. Innovative CME tools in the teaching of evolving strategies in the management and prevention of COVID-19. Open Forum Infect Dis 2020 Dec 31;7(Supplement_1):S344-S344. [doi: 10.1093/ofid/ofaa439.752]

XSL•FO
RenderX

145. Espinoza-Castro B, Encina V, Garrido MA, et al. Online learning for crisis response: evaluating reach and perceived knowledge gains from the MOOC “Infection, Prevention, and Control of Acute Respiratory Infections for Healthcare Workers in Low- and Middle-Income Countries (IPC MOOC)”. BMC Med Educ 2025 Aug 7;25(1):1150. [doi: 10.1186/s12909-025-07661-2] [Medline: 40775697]
146. Gupta S, Goswami B, Madhu SV. Effectiveness of video-based educational intervention on knowledge, attitude, and practice (KAP) of COVID-19 health care workers: lesson for future pandemic preparedness. Ann Natl Acad Med Sci 2023;59:233-236. [doi: 10.1055/s-0043-1772218]
147. Ahmed NH, Tosson MM, Badia T. Effect of educational program on maternity nurses’ knowledge, attitude and practice of preventive measures towards COVID-19. Assiut Sci Nurs J 2022 Feb;10(28):77-90. [doi: 10.21608/asnj.2022.116049.1301]
148. Bohara A, Thapa S, Yilmaz SK, McBee SH. An impact evaluation of COVID-19 training program: knowledge and awareness of public health professionals of Province Five, Nepal. : University of New Mexico; 2021 URL: https://digitalrepository.unm.edu/nsc_research/77 [accessed 2026-06-18]
149. Amiri B, Khajavian N, Rahmani R, Bilandi RR. Comparing the impact of multimedia and educational brochures on knowledge, attitude and work performance of healthcare about COVID-19 management in pregnancy, childbirth, and breastfeeding. Iranian Red Crescent Med J 2023;25(11). [doi: 10.32592/ircmj.2023.25.11.2498]
150. Fadel EA, Alshawish E, El-Shaboury RHR, Khalil DE, Mahmoud FZ, El-Feshawy NI. Effect of implementing virtual educational sessions on nursing students’ knowledge, attitude and hesitancy regarding COVID-19 vaccination. Inquiry 2025;62:469580251339114. [doi: 10.1177/00469580251339114] [Medline: 40468759]
151. Garcia KS, Rodriguez A, Gonzalez Z, et al. Pretest-post-test evaluation with lay midwives in remote Guatemala after educational activities about COVID-19. RRH 2024;24(3):8387. [doi: 10.22605/RRH8387]
152. Nassar AAH, Al Serouri AA, Al-Shahethi AH, Almoayed KA. Effectiveness of training on health care workers’ knowledge, attitude and practice regarding COVID-19 infection prevention and control, Yemen, 2021. BMC Health Serv Res 2024 Nov 15;24(1):1411. [doi: 10.1186/s12913-024-11927-8] [Medline: 39548437]
153. Sabandüzen H, Kavaklı Ö. Evaluation of the effectiveness of the training on “home care of COVID-19 positive/suspicious patients” given to nursing students: a quasi-experimental study. J Educ Health Promot 2024;13(1):250. [doi: 10.4103/jehp.jehp_1574_23] [Medline: 39309990]
154. Said AR. Knowledge and practices of nurses regarding corona virus (COVID-19): an educational intervention. MLU 2021 Mar 12;21(2):36-47. [doi: 10.37506/mlu.v21i2.2642]
155. Strehlow MC, Johnston JS, Aluri KZ, et al. Evaluation of a massive open online course for just-in-time training of healthcare workers. Front Public Health 2024;12:1395931. [doi: 10.3389/fpubh.2024.1395931]
156. McConnell H, Duncan D, Stark P, et al. Enhancing COVID-19 knowledge among nursing students: a quantitative study of a digital serious game intervention. Healthcare (Basel) 2024 May 23;12(11):1066. [doi: 10.3390/healthcare12111066] [Medline: 38891141]
157. Lalitha ND, Bhadauria US, Agarwal D, et al. Comparing the effectiveness of two educational methods for oral health management in COVID-19 pandemic among dental professionals. Przegl Epidemiol 2024;78(1):90-93. [doi: 10.32394/pe.77.48]
158. Kaufman J, Overmars I, Fong J, et al. Training health workers and community influencers to be vaccine champions: a mixed-methods RE-AIM evaluation. BMJ Glob Health 2024 Sep 8;9(9):e015433. [doi: 10.1136/bmjgh-2024-015433] [Medline: 39251236]
159. Kennedy RS, Lane NE, Berbaum KS, Lilienthal MG. Simulator sickness questionnaire: an enhanced method for quantifying simulator sickness. Int J Aviat Psychol 1993 Jul;3(3):203-220. [doi: 10.1207/s15327108ijap0303_3]
160. Salway RJ, Williams T, Londono C, Roblin P, Koenig K, Arquilla B. Comparing training techniques in personal protective equipment use. Prehospital Disaster Med 2020 Aug;35(4):364-371. [doi: 10.1017/S1049023X20000564] [Medline: 32390583]
161. Choi J. Nurse’s knowledge, attitude and use of personal protective equipment related to acute respiratory infections [Master’s thesis]. 2016.
162. Kwon M. Knowledge, perception, and performance related to personal protective equipment among nurses in specialized infectious disease hospitals [Master’s thesis]. 2021.
163. Ayres H. Factors related to motivation to learn and motivation to transfer learning in a nursing population [Dissertation]. 2005.
164. Mohammed A. Knowledge, attitudes, and practices towards COVID-19 among health care workers in primary health care units Dar El Salam, Suhag, Egypt. Sohag Med J 2021 Jan 1;25(1):50-58. [doi: 10.21608/smj.2020.47286.1209]
165. Nwagbara UI, Osual EC, Chireshe R, et al. Knowledge, attitude, perception, and preventative practices towards COVID-19 in sub-Saharan Africa: a scoping review. PLOS ONE 2021;16(4):e0249853. [doi: 10.1371/journal.pone.0249853] [Medline: 33872330]
166. Mahajan S, Kohli HS, Gupta KL, Prasad N, Agarwal SK, COVID-19 Working Group of Indian Society of Nephrology. Infection prevention and control guidelines for COVID. Indian J Nephrol 2020;30(3):185-187. [doi: 10.4103/ijn.IJN_165_20] [Medline: 33013068]
167. Rational use of personal protective equipment for coronavirus disease (COVID-19) and considerations during severe shortages: interim guidance. World Health Organization. 2020 Dec 23. URL: https://www.who.int/publications/i/item/

XSL•FO
RenderX

rational-use-of-personal-protective-equipment-for-coronavirus-disease-(covid-19)-and-considerations-during-severe-shortages [accessed 2026-06-30]
168. Naseef HA, Al-Shami NA, Hadba LSA, et al. Knowledge, attitudes, and practices about coronavirus disease (COVID-19) among Birzeit University students: a cross-sectional study. J Public Health (Berl) 2022 Sep;30(9):2139-2152. [doi: 10.1007/s10389-021-01665-0]
169. Wharrad H, Morales R, Windle R, Bradley C. A toolkit for a multilayered, cross institutional evaluation strategy. In: Luca J, Weippl E, editors. : Association for the Advancement of Computing in Education (AACE); 2026 Jun 17 Presented at: ED-MEDIA 2008--World Conference on Educational Multimedia, Hypermedia & Telecommunications; Jun 30 to Jul 4, 2008; Vienna, Austria p. 4921-4925 URL: https://www.learntechlib.org/primary/p/29054 [accessed 2026-06-17]
170. Moorhead S, Swanson E, Johnson M, Maas M. Nursing Outcome Classification (NOC): Measurement of Health Outcomes, 6th edition: Elsevier Health Sciences; 2018.
171. Gharpure R, Hunter CM, Schnall AH, et al. Knowledge and practices regarding safe household cleaning and disinfection for COVID-19 prevention - United States, May 2020. MMWR Morb Mortal Wkly Rep 2020 Jun 12;69(23):705-709. [doi: 10.15585/mmwr.mm6923e2] [Medline: 32525852]
172. McIntosh K, Hirsch MS, Bloom A. COVID-19: epidemiology, virology, and prevention. : UpToDate; 2021 URL: http://hoiyhoctphcm.org.vn/wp-content/uploads/2021/08/B001-COVID19-Epidemiology-virology-and-prevention.pdf [accessed 2026-06-17]
173. Salman M, Mustafa Z, Asif N, et al. Knowledge, attitude and preventive practices related to COVID-19 among health professionals of Punjab province of Pakistan. J Infect Dev Countries 2020 Jul 31;14(7):707-712. [doi: 10.3855/jidc.12878] [Medline: 32794458]
174. Martin LR, Petrie KJ. Understanding the dimensions of anti-vaccination attitudes: the Vaccination Attitudes Examination (VAX) Scale. Ann Behav Med 2017 Oct;51(5):652-660. [doi: 10.1007/s12160-017-9888-y] [Medline: 28255934]
175. Park YM. The development and effects of a simulation-based disaster nursing education program for nursing students using standardized patients [Dissertation]. : Seoul: Kyung Hee University; 2020.
176. COVID-19 strategic preparedness and response plan: operational planning guideline. : World Health Organization; 2021 Mar 3 URL: https://www.who.int/publications/m/item/covid-19-strategic-preparedness-and-response-plan-operational-planning-guideline [accessed 2026-06-17]
177. Heininger SK, Baumgartner M, Zehner F, et al. Measuring hygiene competence: the picture-based situational judgement test HygiKo. BMC Med Educ 2021 Jul 30;21(1):410. [doi: 10.1186/s12909-021-02829-y] [Medline: 34330263]
178. Coronavirus disease (COVID-19) advice for the public: mythbusters. World Health Organization. 2022 Jan 19. URL: https://www.who.int/emergencies/diseases/novel-coronavirus-2019/advice-for-public/myth-busters [accessed 2026-06-17]
179. Alrubaiee GG, Al-Qalah TAH, Al-Aawar MSA. Knowledge, attitudes, anxiety, and preventive behaviours towards COVID-19 among health care providers in Yemen: an online cross-sectional survey. BMC Public Health 2020 Oct 13;20(1):1541. [doi: 10.1186/s12889-020-09644-y] [Medline: 33050896]
180. Ye Y, Shi P, Gui Y, et al. Point-of-care training program on COVID-19 infection prevention and control for pediatric healthcare workers: a multicenter, cross-sectional questionnaire survey in Shanghai, China. Transl Pediatr 2021 Jan;10(1):44-53. [doi: 10.21037/tp-20-247] [Medline: 33633936]
181. Huynh G, Nguyen TH, Tran V, Vo K, Vo V, Pham L. Knowledge and attitude toward COVID-19 among healthcare workers at District 2 Hospital, Ho Chi Minh City. Asian Pac J Trop Med 2020;13(6):260. [doi: 10.4103/1995-7645.280396]
182. Tadesse DB, Gebrewahd GT, Demoz GT. Knowledge, attitude, practice and psychological response toward COVID-19 among nurses during the COVID-19 outbreak in northern Ethiopia, 2020. New Microbes New Infect 2020 Nov;38:100787. [doi: 10.1016/j.nmni.2020.100787] [Medline: 33072339]
183. Honarvar B, Lankarani KB, Kharmandar A, et al. Knowledge, attitudes, risk perceptions, and practices of adults toward COVID-19: a population and field-based study from Iran. Int J Public Health 2020 Jul;65(6):731-739. [doi: 10.1007/s00038-020-01406-2] [Medline: 32583009]
184. Paul A, Sikdar D, Hossain MM, et al. Knowledge, attitudes, and practices toward the novel coronavirus among Bangladeshis: implications for mitigation measures. PLoS ONE 2020;15(9):e0238492. [doi: 10.1371/journal.pone.0238492] [Medline: 32877449]
185. Moore DE, Green JS, Gallis HA. Achieving desired results and improved outcomes: integrating planning and assessment throughout learning activities. J Contin Educ Health Prof 2009;29(1):1-15. [doi: 10.1002/chp.20001] [Medline: 19288562]
186. Tomas ME, Kundrapu S, Thota P, et al. Contamination of health care personnel during removal of personal protective equipment. JAMA Intern Med 2015 Dec;175(12):1904-1910. [doi: 10.1001/jamainternmed.2015.4535] [Medline: 26457544]
187. Healthcare professional preparedness checklist for transport and arrival of patients with confirmed or possible COVID-19. : Centers for Disease Control and Prevention (U.S; 2020 Feb 21 URL: https://stacks.cdc.gov/view/cdc/85557 [accessed 2026-06-17]
188. Wu TY, Hoffman JL, Chow CM, Hartl B. Training community health navigators in the public health workforce to respond during the COVID-19 pandemic. J Public Health (Berl) 2024 Feb;32(2):287-294. [doi: 10.1007/s10389-022-01812-1]
189. Drevon D, Fursa SR, Malcolm AL. Intercoder reliability and validity of WebPlotDigitizer in extracting graphed data. Behav Modif 2017 Mar;41(2):323-339. [doi: 10.1177/0145445516673998] [Medline: 27760807]

XSL•FO
RenderX

190. Saunders C, Palesy D, Lewis J. Systematic review and conceptual framework for health literacy training in health professions education. Health Prof Educ 2019 Mar;5(1):13-29. [doi: 10.1016/j.hpe.2018.03.003]
191. Baumeister A, Aldin A, Chakraverty D, et al. Interventions for improving health literacy in migrants. Cochrane Database Syst Rev 2023 Nov 14;11(11):CD013303. [doi: 10.1002/14651858.CD013303.pub2] [Medline: 37963101]
192. Cudjoe J, Delva S, Cajita M, Han HR. Empirically tested health literacy frameworks. Health Lit Res Pract 2020 Feb 11;4(1):e22-e44. [doi: 10.3928/24748307-20191025-01] [Medline: 32053206]
193. Walters R, Leslie SJ, Polson R, Cusack T, Gorely T. Establishing the efficacy of interventions to improve health literacy and health behaviours: a systematic review. BMC Public Health 2020 Jun 30;20(1):1040. [doi: 10.1186/s12889-020-08991-0] [Medline: 32605608]
194. Papasteri C, Letzner RD, Pascal S. Pandemic KAP framework for behavioral responses: initial development from lockdown data. Curr Psychol 2024 Jul;43(26):22767-22779. [doi: 10.1007/s12144-024-05670-w]
195. Lorini C, Lastrucci V, Paolini D, Bonaccorsi G, Florence Health Literacy Research Group. Measuring health literacy combining performance-based and self-assessed measures: the roles of age, educational level and financial resources in predicting health literacy skills. A cross-sectional study conducted in Florence (Italy). BMJ Open 2020 Oct 5;10(10):e035987. [doi: 10.1136/bmjopen-2019-035987] [Medline: 33020080]
196. Miller TA. Health literacy and adherence to medical treatment in chronic and acute illness: a meta-analysis. Patient Educ Couns 2016 Jul;99(7):1079-1086. [doi: 10.1016/j.pec.2016.01.020] [Medline: 26899632]
197. Falcón M, Rodríguez-Blázquez C, Fernández-Gutiérrez M, Romay-Barja M, Bas-Sarmiento P, Forjaz MJ. Measuring COVID-19 health literacy: validation of the COVID-19 HL questionnaire in Spain. Health Qual Life Outcomes 2022;20(1):138. [doi: 10.1186/s12955-022-02050-5]
198. Dong Q, Liu T, Liu R, Yang H, Liu C. Effectiveness of digital health literacy interventions in older adults: single-arm meta-analysis. J Med Internet Res 2023 Jun 28;25:e48166. [doi: 10.2196/48166] [Medline: 37379077]
199. Kim SH, Lee A. Health‐literacy‐sensitive diabetes self‐management interventions: a systematic review and meta‐analysis. Worldviews Evid Based Nurs 2016 Aug;13(4):324-333. [doi: 10.1111/wvn.12157] [Medline: 27104337]
200. Shahsavari H, Matourypour P, Ghiyasvandian S, Nejad MRG. Medical Research Council framework for development and evaluation of complex interventions: a comprehensive guidance. J Educ Health Promot 2020;9:88. [doi: 10.4103/jehp.jehp_649_19] [Medline: 32509896]
201. Osborne RH, Elmer S, Hawkins M, Cheng C. The Ophelia Manual. : Swinburne University of Technology; 2021 URL: https://healthliteracydevelopment.com/wp-content/uploads/2024/09/Ophelia-Manual-Osborne-Elmer-Hawkins-Cheng-et-al-2022.pdf [accessed 2026-06-30]
202. Naeem SB, Bhatti R. The COVID‐19 'infodemic': a new front for information professionals. Health Info Libraries J 2020 Sep;37(3):233-239. [doi: 10.1111/hir.12311]
203. Merati-Fashi F, Dalvandi A, Yekta ZP. Health information seeking and its achievements in patients with chronic disease. J Nurse Pract 2022 Apr;18(4):411-416. [doi: 10.1016/j.nurpra.2021.12.024]
204. Schaeffer D, Haarmann A, Griese L. Professionelle Gesundheitskompetenz ausgewählter Gesundheitsprofessionen in Deutschland. Ergebnisse des HLS-PROF-GER [Article in German]. : Hertie School/Universität Bielefeld/Stiftung Gesundheitswissen; 2023 URL: https://www.nap-gesundheitskompetenz.de/app/download/8463076363/Ergebnisbericht_HLS-PROF-GER.pdf?t=1697105483 [accessed 2026-06-18]
205. ARE-Wochenbericht KW 40. : Robert Koch Institute; 2024. [doi: 10.25646/12870]

Abbreviations

Cochrane RoB 2 Tool: Cochrane Risk of Bias 2 Tool
EPHPP-QAT: Quality Assessment Tool for Quantitative Studies
GRADE : Grading of Recommendations, Assessment, Development, and Evaluation
HL: health literacy
HP: health professional
MERS: Middle East respiratory syndrome
NRSI: nonrandomized study of interventions
PI: prediction interval
PICO: Population-Intervention-Comparator-Outcome
PPE: personal protective equipment
PRISMA : Preferred Reporting Items for Systematic Reviews and Meta-Analyses
RCT: randomized controlled trial
ROBINS-I: Risk of Bias in Nonrandomized Studies of Interventions
SMD: standardized mean difference
WHO: World Health Organization

XSL•FO
RenderX

Edited by A Stone, TDA Cardoso; submitted 22.Dec.2024; peer-reviewed by JJB Seng, L Stuby, S Markham; revised version received 05.May.2026; accepted 21.May.2026; published 10.Jul.2026.

Please cite as:
Hasenpusch C, Kuper P, Skiba EM, Sprenger AA, Matterne U, Kannengießer L, Grätsch EI, Shin J, Li M, Apfelbacher CJ
Interventions to Enhance COVID-19 Pandemic Health Literacy in Health Professionals: Systematic Review
JMIR Med Educ 2026;12:e70400
URL: https://mededu.jmir.org/2026/1/e70400
doi:10.2196/70400

© Claudia Hasenpusch, Paula Kuper, Eva-Maria Skiba, Antonia Anabella Sprenger, Uwe Matterne, Lena Kannengießer, Esther Ivanka Grätsch, Jaekyung Shin, Mengtong Li, Christian Joachim Apfelbacher. Originally published in JMIR Medical Education (https://mededu.jmir.org), 10.Jul.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Education, is properly cited. The complete bibliographic information, a link to the original publication on https://mededu.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Video Recording of Patient-Clinician Interactions in Health Education: Scoping Review

Mariana Morgado[1], PhD; Yago Leira[2], PhD; Tiago Leitão[1], MSc; José João Mendes[1], PhD; Luís Proença[1], PhD

[1]Egas Moniz Center for Interdisciplinary Research (CiiEM), Egas Moniz School of Health & Science, Campus Universitário s/n, Quinta da Granja, Almada, Portugal

[2]Periodontology Unit, Faculty of Medicine and Odontology, Universidade de Santiago de Compostela, Santiago de Compostela, Spain

Corresponding Author:
Mariana Morgado, PhD
Egas Moniz Center for Interdisciplinary Research (CiiEM), Egas Moniz School of Health & Science, Campus Universitário s/n, Quinta da Granja, Almada, Portugal

Abstract

Background: Video recordings of patient-clinician interactions (PCIs) have become an integral tool in health professions education, with authentic opportunities to enhance clinical communication and decision-making skills. Despite their educational value, ethical and logistical challenges regarding their use remain underexplored.

Objective: This scoping review aims to map the use of video-recorded PCIs in health education, categorizing types and educational purposes, examining applications across disciplines, and identifying ethical and logistical issues.

Methods: Comprehensive searches were conducted in MEDLINE, Embase, and ERIC (Education Resources Information Center) databases, supplemented by gray literature, with the last search performed on October 30, 2024. Two independent reviewers screened and extracted data on study design, intervention characteristics, educational objectives, and barriers to implementation. Given the scoping review nature, risk of bias assessment was not conducted. Data synthesis involved descriptive analysis and thematic categorization of findings.

Results: A total of 69 studies met inclusion criteria, predominantly from medical (n=28, 41%), dental (n=15, 22%), and nursing (n=8, 12%) education. Publication frequency increased over the past decade. Studies used varied video types, including authentic consultations, simulated encounters, and procedural recordings. Main educational aims focused on clinical skills, communication, professionalism, and reflective practice. Ethical reporting was inconsistent, with variable adherence to informed consent, patient privacy, and confidentiality standards. Common logistical barriers included limitations in technical infrastructure, data management, and curriculum integration.

Conclusions: Video-recorded PCIs offer valuable educational resources but face challenges in ethical standardization and logistical feasibility. Future research should prioritize the development of ethical frameworks and standardized protocols, together with the evaluation of long-term educational and psychological outcomes, to support the responsible and scalable integration of PCI recordings into health education curricula.

Trial Registration: OSF Registries 8w3m6; https://osf.io/8w3m6/overview

(JMIR Med Educ 2026;12:e70324) doi:10.2196/70324

KEYWORDS

video recordings; patient-clinician interaction; medical education; clinical skills; clinician communication; communication skills; health professions education

Introduction

Rationale

Video-based learning is increasingly recognized as a transformative approach in health professions education, supported by several key educational theories. According to constructivist theory, learners actively build new knowledge by integrating multimedia content with prior experience, a process that is facilitated by the contextual richness of video recordings [1,2]. Cognitive load theory also supports video-based learning, demonstrating that well-designed video materials optimize cognitive processing, reduce extraneous mental effort, and enhance comprehension and retention through dual-channel presentation (visual and auditory) [3-5]. Experiential learning theory emphasizes the value of authentic, concrete experiences, such as those provided by video-recorded patient-clinician interactions (PCIs), which enable learners to observe, reflect, conceptualize, and experiment in a cycle that deepens practical understanding [6-8]. Social learning theory highlights the importance of observation and modeling, with video enabling learners to witness expert behaviors and communication

XSL•FO
RenderX

strategies in realistic settings [9,10]. Finally, active learning principles are supported by using video as a springboard for discussion, reflection, and collaborative analysis, all of which are linked to improved learner engagement and knowledge retention [11,12].

Since the 1960s, there has been a significant evolution in the integration of video recordings into health education, initially used to assess and improve the technical and interpersonal skills of junior doctors [13]. Technological advances have broadened the scope of video applications, from procedural demonstrations to the recording of real or simulated PCIs [14]. These interactions now serve as valuable resources for developing clinical, communication, and reflective skills [15]. Studies have shown that watching video content before practical exercises improves performance, reduces anxiety, and enhances attitudes, particularly in fields such as nursing and medicine [16-18].

Despite the educational benefits of video recording in clinical education, its adoption raises important ethical and logistical challenges [19,20]. Critical to responsible implementation are ensuring informed consent, maintaining confidentiality, and adhering to legal frameworks such as HIPAA (Health Insurance Portability and Accountability Act) [21,22]. Logistical barriers, including technical setup, scheduling, and resource allocation, also impact the feasibility of integrating video into curricula [21-23]. While most patients and clinicians recognize the educational value of video recording, concerns about privacy and increased scrutiny persist [22,23].

In light of the increasing prevalence and complexity of video-based approaches, there is an evident requirement for a systematic mapping of the characteristics, educational outcomes, and practical challenges associated with the video recording of PCIs in health education.

Objectives

This scoping review aims to systematically map the use of video recordings of PCIs in health education. More specifically, the review will examine the types and educational purposes of these recordings, compare their effectiveness with that of traditional teaching methods, evaluate their impact on learner outcomes such as clinical competence and communication skills, and identify key ethical and logistical considerations relating to their implementation.

Methods

Study Design

We conducted a scoping review in accordance with the Joanna Briggs Institute Reviewer Manual [24,25]. The scoping review is reported according to the PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses Extension for Scoping Reviews) checklist [26] (Checklist 1). The final protocol was registered on the Open Science Framework [27] on October 31, 2024.

Eligibility Criteria

The eligibility criteria for this scoping review were established a priori in order to comprehensively address the research questions. Studies were included if they assessed the usefulness of video recordings of PCIs as an educational tool in health education. These criteria focused on the types, characteristics, and educational outcomes of these video recordings, as well as the ethical and logistical considerations involved in their use. In order to ensure the relevance of the findings to educational contexts, studies were required to include learners in health education, including medical, nursing, dental, and allied health students of all educational levels, who used video recordings of PCIs as part of their training. The video recordings could include full-length consultations, specific clinical procedures, or role-playing scenarios designed to replicate real-world clinical interactions. Other forms of educational media, such as written case studies or noninteractive video content, were excluded to maintain the focus on interactive and video-based learning. All comparators were eligible for inclusion, encompassing studies that compared video-based learning with traditional teaching methods (eg, lectures or textbooks) or with different video formats (eg, live encounters vs simulated scenarios). Noncomparative studies were also considered, providing a comprehensive understanding of the impact of video-based learning. To capture a wide range of evidence, all types of publications were considered, including peer-reviewed articles, conference proceedings, editorials, and book chapters, and no restrictions were placed on the time period or educational setting (eg, universities, colleges, and vocational training programs) to allow for an exploration of trends and developments in the use of video recordings over time. Conversely, studies that did not incorporate learners in health education or that did not use video recordings as a teaching instrument were excluded in order to maintain the focus on the educational application of video-based learning. Studies published in languages other than English were excluded to ensure consistency in data extraction and analysis. Furthermore, unpublished studies and non–peer-reviewed literature were excluded to ensure the credibility and reliability of the findings.

Information Sources

A comprehensive literature search was conducted in MEDLINE, Embase, and ERIC (Education Resources Information Center) to identify potentially relevant studies on the use of video recordings of PCIs in health education. The final date for the retrieval of articles was October 30, 2024, thereby ensuring that the information included in this review is current.

In addition to the electronic database searches, a comprehensive gray literature search was conducted in accordance with the CADTH (Canadian Agency for Drugs and Technologies in Health) Grey Matters checklist (Checklist 2). This entailed an investigation of a range of gray literature sources and relevant websites with a view to identifying unpublished or difficult-to-locate studies related to video-based learning. Furthermore, the reference lists of pertinent reviews were examined to ascertain any additional studies that may have been overlooked in the database searches. This approach facilitated a more comprehensive collection of information, encompassing both published literature and challenging-to-locate documents.

The final search results were exported into Zotero (Digital Scholar), a reference management software. The presence of duplicates was identified through the use of Zotero's automated

XSL•FO
RenderX

duplicate detection function, which uses a comparison of citation fields such as title, authors, publication year, and journal. A manual review was conducted when necessary in order to ensure that all duplicates were accurately identified and excluded prior to the screening process. This multifaceted search strategy was designed to ensure a comprehensive understanding of the existing literature on the use of video as an educational tool in health education.

Search Strategy

To facilitate replication, the comprehensive search strategy used, including specific database terms, filters, and a systematic approach to the gray literature, is described below. The primary search strategy, conducted entirely in MEDLINE, is presented in Multimedia Appendix 1, with parallel structures applied to additional databases including Embase and ERIC. This search strategy was developed by one researcher, peer-reviewed by a more experienced member of the research team, and finalized on October 30, 2024.

Selection of Sources of Evidence

The selection of sources for this scoping review was designed to enable a comprehensive mapping of the use and impact of video recordings of PCIs in health education. To ensure consistency, a calibration exercise was initially conducted. This involved all reviewers independently screening a set of 20 publications with the aim of ensuring a consistent interpretation of the inclusion criteria. This process revealed ambiguities, which were subsequently addressed through group discussions and resulted in the formulation of specific amendments to the screening and data extraction manual.

To facilitate the selection process and ensure consistency, a bespoke screening form was created in Google Sheets. The form included specific questions aligned with the PICO (population, intervention, comparison, and outcomes) framework, which were designed to elicit information regarding the type of video recording (eg, consultations, clinical procedures, and role-playing scenarios), its educational purpose (eg, teaching clinical or communication skills), and how it compares with traditional teaching methods. The form was then trialed by multiple reviewers using sample citations and full-text articles, enabling the questions and layout to be refined to ensure optimal clarity and ease of use. The use of Google Sheets facilitated data management and the efficient tracking of selection decisions.

In the formal screening phase, each title, abstract, and full-text publication identified in the search was evaluated independently by two reviewers (MM and TL), thereby ensuring unbiased study selection. The implementation of an independent duplicate screening process served to minimize the risk of relevant studies being omitted. Disagreements between the reviewers were resolved through consensus discussions. In the event of persistent disagreement, a third reviewer was consulted to make the final decision. This structured and standardized approach guaranteed that our scoping review was conducted systematically, thereby yielding a reliable set of sources for the evaluation of the role of video recordings in enhancing educational outcomes in health education.

Data Charting Process

A structured data charting form was developed for the purpose of data extraction, with the objective of capturing detailed information from each included source. The charting form was designed with the specific intention of collecting data that aligned with the review's PICO framework and objectives. It included headers such as screening status, author, year, origin, type of article, study design, aims/purpose, study population, area of study, sample size, methods, intervention type, duration, comparator, outcome measures, and key findings that relate to research questions. The structured approach enabled the systematic capture of data pertaining to the characteristics of the studies, the details of the interventions, and the outcomes.

The initial charting form was developed by two reviewers in collaboration, with items selected based on their relevance to the study objectives. The form was trialed on a sample set of studies, with each reviewer independently charting the data and discussing the results to identify areas requiring elucidation. As a result of this iterative pilot testing, minor adjustments were made to the form in order to ensure clarity and comprehensive coverage. All modifications were duly documented.

In the phase of comprehensive data charting, each study was independently charted by two reviewers using Google Sheets. This approach ensured the organized management of data and enabled the tracking of updates in real time. Any discrepancies in the charting process were resolved through consensus-based discussions. In the event of an unresolved issue, a third reviewer was consulted to provide a final decision. In instances where essential data required clarification, the study investigators were directly contacted.

The structured approach to both selection and data charting guaranteed a comprehensive and reliable dataset, which constituted a robust basis for the analysis of the role of video recordings in health education. Furthermore, it permitted an in-depth examination of the impact of these tools on the educational outcomes of health learners.

Data Items

The following key characteristics were abstracted from the included articles: author and year of publication, origin of the study, type of article, research design, and data type. Furthermore, additional variables were included, such as the aims and purpose of the research, the characteristics of the student population (eg, year of study, study cycle, and area of study), and details of the intervention (eg, type, duration, and comparator). Additionally, data were collected on the sample size, outcome variables, and instruments used to measure these outcomes. The data items were designed to capture both qualitative and quantitative information, where applicable. For instance, outcome variables and instruments required reviewer judgment based on study reporting. The complete data chart is available in Multimedia Appendix 2, and the detailed definitions of each data item are provided in Multimedia Appendix 3.

Synthesis of Results

To synthesize the range of evidence included in the scoping review, the studies were grouped based on the type of

XSL•FO
RenderX

intervention, which included PCIs, simulated PCIs, and other forms of intervention. For each group, key characteristics were summarized, including the types of settings, populations, and study designs, along with the measures used and the broad findings. Moreover, the studies were classified according to the comparator groups, which included the following: no intervention/control group, traditional methods, simulated PCIs, PCIs, static video recordings of simulated PCIs, self-guided learning, and various other methods.

Studies were also categorized according to the research design and data type used. The research designs included quasi-experimental, observational, descriptive, randomized controlled trial, phenomenological, experimental, comparative, case study, cross-sectional, and nonexperimental. With regard to the data types, the studies were classified as follows: those using a mixed methods approach, those using quantitative methods, and those using qualitative methods.

In addition, the studies were classified according to article type, comprising original research papers and review articles, including literature reviews, scoping reviews, and systematic reviews. Moreover, the studies were classified according to the population degree. The distribution included the following categories: undergraduate, postgraduate, undergraduate and postgraduate, master's, not applicable, not stated, undergraduate and master's, and doctorate.

The evidence was further synthesized by categorizing it into key areas of study, including medical education, dental education, nursing education, interprofessional education in health care, occupational therapy education, speech-language therapy education, veterinary medical education, pharmacy education, nutrition and dietetics education, and physiotherapy education.

We synthesized the collated data by using descriptive statistics (frequencies and proportions). We used Microsoft Excel and SPSS software (version 26.0; IBM Corp) to analyze the data.

The results were presented in both narrative form and tabular format. The former provided an overview of the studies within each category and subcategory, while the latter highlighted the general outcome variables and instruments used. Moreover, visual representations, such as diagrams, were included to facilitate the illustration of the distribution of studies across the various categories.

The full synthesis and comprehensive categorization of all evidence are presented in tabular and visual form, facilitating a more nuanced comprehension of the trends, gaps, and overarching findings derived from the reviewed studies.

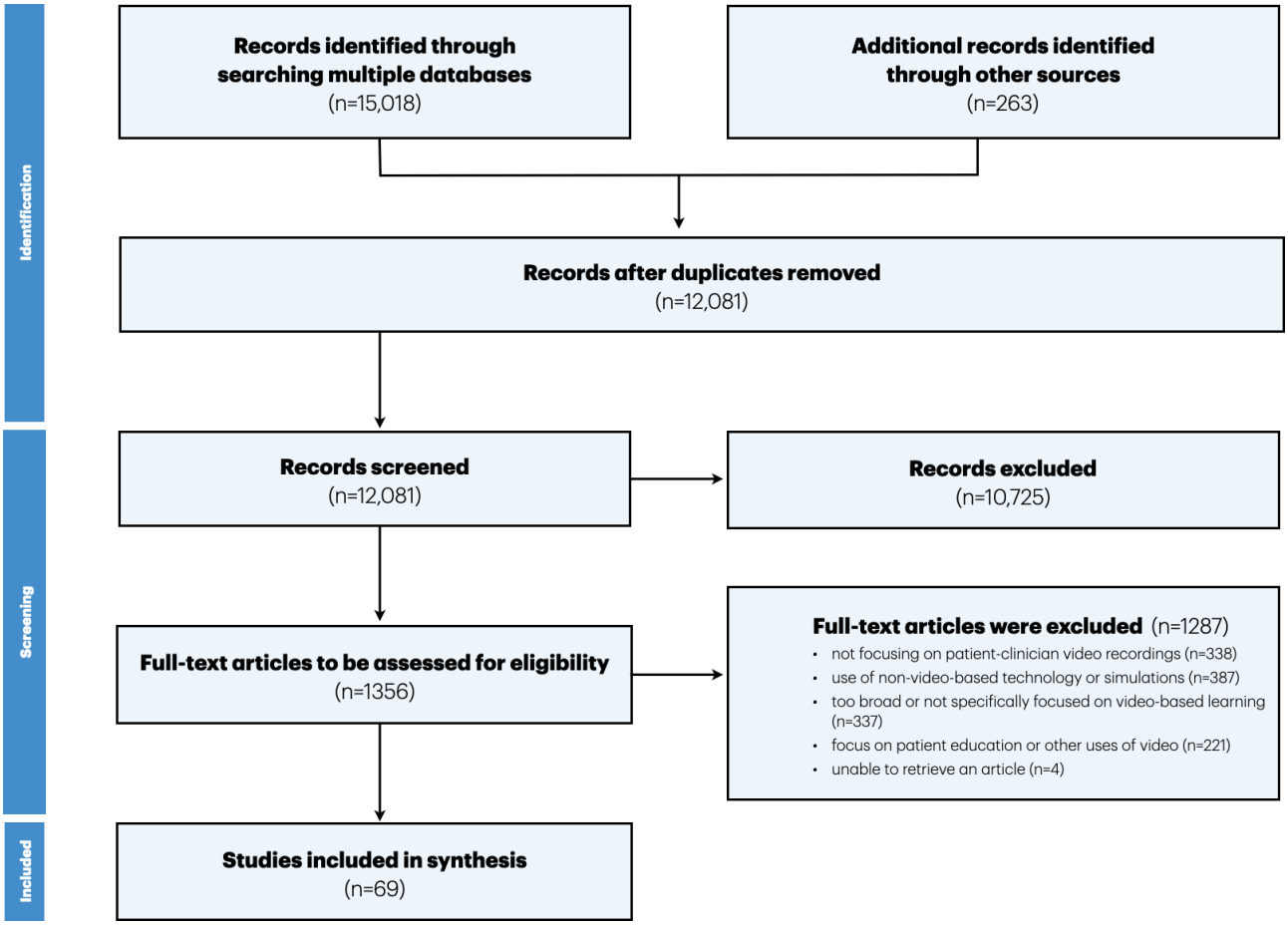
Results

Selection of Sources of Evidence

The search strategy yielded a total of 15,281 records, of which 12,081 were screened after the removal of duplicates. Following a thorough examination of the full texts, 69 studies were identified as suitable candidates for inclusion in the scoping review [15,20,28-94]. The exclusion of patient-clinician video recordings, the use of non–video-based or irrelevant technologies, and noneducational contexts were identified as the most common reasons for exclusion. The comprehensive selection process is delineated in Figure 1, which details the identification, screening, eligibility, and inclusion outcomes.

XSL•FO
RenderX

Figure 1. PRISMA flow diagram showing inclusion and exclusion strategy [26].



Characteristics of Sources Evidence

Multimedia Appendix 2 summarizes each individual source of evidence's objectives, target population, research setting, research design, data collection methods, outcome measures, and primary findings. The references for each study are provided in the list of sources alongside the relevant details.

Synthesis of Results

Characteristics of the Included Documents

The 69 included studies demonstrate a clear increase in research interest in video-based learning over time, with the majority of these studies having been published between 2011 and 2020 (n=41, 59%) (Table 1). This trend underscores the growing recognition of video-based learning as a valuable tool in health education.

XSL•FO
RenderX

Table . Characteristics of the included studies by publication year, article type, research design, and data type (n=69).

Characteristics	Studies, n (%)
Publication year (n=69)	
Before 2000	3 (4)
2001 - 2010	14 (20)
2011 - 2020	41 (59)
Since 2021	11 (16)
Data type (n=62)	
Mixed methods	29 (47)
Quantitative	17 (27)
Qualitative	16 (26)
Type of article and research design (n=69)	
Original research article	
Quasi-experimental design	23 (33)
Observational design	15 (22)
Descriptive design	5 (7)
Randomized controlled trial	5 (7)
Phenomenological design	4 (6)
Experimental design	4 (6)
Comparative design	2 (3)
Case study design	2 (3)
Cross-sectional design	1 (1)
Nonexperimental design	1 (1)
Review article	
Literature review	4 (6)
Scoping review	2 (3)
Systematic review	1 (1)
Place of publication (n=69)	
North America	
United States	33 (48)
Canada	3 (4)
Europe	
United Kingdom	5 (7)
Sweden	5 (7)
Denmark	2 (3)
Netherlands	1 (1)
Norway	1 (1)
Spain	1 (1)
Ireland	1 (1)
Germany	1 (1)
Asia	
South Korea	3 (4)
Turkey	2 (3)
China	2 (3)

XSL•FO
RenderX

Characteristics	Studies, n (%)
Saudi Arabia	1 (1)
Japan	1 (1)
India	1 (1)
Oceania	
Australia	4 (6)
Africa	
South Africa	2 (3)

In terms of geography, the largest proportion of studies originated from North America (n=36, 52%), with the United States contributing 33 (48%) of these. The other regions represented were Europe (n=17, 25%), Asia (n=10, 15%), Oceania (n=4, 6%), and Africa (n=2, 3%). This distribution illustrated the global adoption of this educational approach (Table 1).

A significant proportion of the included studies were original research articles (n=62, 90%). The most prevalent research designs were quasi-experimental (n=23, 33%) and observational (n=15, 22%), reflecting a focus on applied educational interventions. Methodologically, the most prevalent approach was that of mixed methods studies (29/62, 47%), followed by quantitative (17/62, 27%) and qualitative (16/62, 26%) designs. This distribution reflects increasing recognition of the importance of combining quantitative outcomes with contextual, learner-centered perspectives when evaluating video-based learning. The extensive use of original datasets, in conjunction with the emphasis on the cultivation of clinical and communication skills, rather than merely theoretical knowledge, further reinforces the applied, practice-oriented character of video-based learning research. For a comprehensive overview of the breakdowns of research designs and data types, please refer to Table 1.

Educational Contexts and Learner Populations

The majority of studies were conducted in the field of medical education (n=28, 41%), followed by dental education (n=15, 22%) and nursing education (n=8, 12%). Regarding the level of learners, the overwhelming focus was on undergraduate students, who were the subject of 50 studies (73%) and accounted for 4648 participants, or 83% of the total sample (n=5685). Please refer to Table 2 for details on the distribution by health field and learner level.

XSL•FO
RenderX

Table . Distribution of included studies and sample sizes by health discipline and education level.

	Studies (N=69), n (%)	Sample size (N=5685), n (%)
Health field of study		
Medical education	28 (41)	2170 (38)
Dental education	15 (22)	1298 (23)
Nursing education	8 (12)	317 (6)
Interprofessional education (IPE) in health care	4 (6)	903 (16)
Occupational therapy education	4 (6)	294 (5)
Speech-language therapy education	3 (4)	75 (1)
Veterinary medical education	3 (4)	600 (11)
Pharmacy education	2 (3)	8 (0)
Nutrition and dietetics education	1 (1)	12 (0)
Physiotherapy education	1 (1)	8 (0)
Educational level		
Undergraduate	50 (72)	4648 (82)
Postgraduate	5 (7)	97 (2)
Undergraduate and postgraduate	3 (4)	515 (9)
Master	3 (4)	114 (2)
Undergraduate and master's	1 (1)	14 (0)
Doctorate	1 (1)	24 (0)
Not stated	3 (4)	181 (3)
Not applicable	3 (4)	92 (2)

Types and Characteristics of Interventions and Comparators

Intervention types were grouped into 3 main categories: PCIs, simulated PCIs, and other methods of video-based education (Table 3). One study contributed two intervention types to the total of 70 instances across 69 studies.

The most frequently used intervention across the included studies was video recordings of PCIs, either in their raw form or for review and reflection purposes. This category accounted for 39 (56%) out of the 70 recorded intervention types. The second-largest intervention category was simulated PCIs, present in 22 (31%) out of 70 intervention types.

The comparators varied widely. The most common approach was to use a control or no-intervention group, which was used in 32 (46%) out of 69 studies. Traditional teaching methods, such as lectures and textbooks, were used as comparators in 26 (38%) out of 69 studies (Table 4). One study contributed two comparator types to the total of 70 instances across 69 studies.

XSL•FO
RenderX

Table . Intervention and subintervention types of the included studies (n=69). One study reported dual interventions (n=70 across 69 studies; see Multimedia Appendix 2).

Intervention/subintervention type	Studies, n (%)
Patient-clinician interactions	
Video recording of patient-clinician interactions	24 (34)
Review of video recording of patient-clinician interactions	14 (20)
First-person video recording of patient-clinician interactions	1 (1)
Simulated patient-clinician interactions	
Video recording of simulated patient-clinician interactions	15 (21)
Review of video recording of simulated patient-clinician interactions	6 (9)
First-person video recording of simulated patient-clinician interactions	1 (1)
Other	
Video case-based learning	4 (6)
Video-based assessment	3 (4)
Various video-based education methods	2 (3)

Table . Comparator types of the included studies (n=69). One study reported dual comparators (n=70 across 69 studies; see Multimedia Appendix 2).

Comparator type	Studies, n (%)
No intervention/control group	32 (46)
Traditional methods	26 (37)
Simulated PCIs[a]	5 (7)
Patient-clinician interactions	3 (4)
Static video recordings of simulated PCIs	2 (3)
Self-guided learning	1 (1)
Various	1 (1)

[a]PCIs: patient-clinician interactions.

The average duration of video-based interventions across all studies was approximately 20.37 minutes. Those involving PCIs exhibited considerable variation, with an overall mean duration of 19.44 minutes. Video recordings of PCIs were the longest on average (24.53 minutes), followed by reviews of these recordings (23.29 minutes). By contrast, simulated interactions tended to be shorter, particularly reviews of simulated recordings (5.67 minutes) and first-person perspectives (10.50 minutes) (Table 5).

Table . Mean duration of different types of PCI[a] interventions.

Intervention/subintervention type	Mean duration (min)
PCIs	
Video recording of PCIs	24.53
Review of video recording of PCIs	23.29
First-person video recording of PCIs	10.50
Simulated PCIs	
Video recording of simulated PCIs	22.22
Review of video recording of simulated PCIs	5.67
First-person video recording of simulated PCIs	10.50

[a]PCI: patient-clinician interaction.

XSL•FO
RenderX

Outcome Categories and Subcategories

A total of 226 outcome variables were extracted and categorized within a structured framework of domains, subcategories, and specific variables (Figure 2; Multimedia Appendices 4 and 5).

The category most frequently assessed was clinical competence, accounting for 46 (20%) out of 226 outcomes. Within this domain, clinical skills comprised the majority (29/226, 13%), with a particular focus on clinical reasoning, diagnostic accuracy, technical skills, and procedural application. A further 17 (8%) out of 226 outcomes focused on clinical performance, including metrics such as performance assessments, efficiency, and reliability.

Communication skills were the second-largest outcome domain, with 44 (19%) out of 226 outcomes. This category was subdivided into verbal and nonverbal communication (23/226, 10%), patient interaction and empathy (12/226, 5%), and communication-focused reflection (9/226, 4%). Common variables included communication behaviors, nonverbal awareness, and patient-centeredness.

The third most prominent category was engagement and the learning experience. This category comprised 44 (19%) out of 226 outcomes and was divided into 4 subcategories: learner satisfaction, perceptions, motivation, and interaction with the educational environment. Learner engagement alone was reported in 22 (10%) out of 226 outcomes, highlighting the importance of video-based learning in encouraging active participation and boosting learners' confidence.

Further outcome domains included critical thinking and reasoning (23/226, 10%), which focused on knowledge retention, problem identification and self-reflection, and feedback and evaluation (23/226, 10%), which encompassed feedback quality, feedback from standardized patients, and self-assessment metrics.

Emotional and psychological outcomes were documented in 17 (8%) out of 226, addressing emotional responses, confidence, affective learning, and personal development. Less frequently reported categories included educational process outcomes (n=14, 6%), such as the perceived effectiveness of learning tools, as well as outcomes related to teamwork and interprofessional learning (n=8, 4%), and outcomes related to the patient-centered clinical context (n=7, 3%).

XSL•FO
RenderX

Figure 2. Radial hierarchical bar chart illustrating the distribution of outcome variables across educational domains. Bar length corresponds to the frequency of outcomes within each category.



Assessment Instruments and Measurement Tools

A total of 101 instruments were extracted and categorized within a structured framework of domains, subcategories, and specific instruments. The detailed list of specific instruments and their classifications can be found in Multimedia Appendix 6.

The most frequently used instruments were questionnaires and surveys (34/101, 34%), predominantly general, with a smaller proportion addressing specific topics including validated tools such as the Care Plan Evaluation [31], the Compassionate Care Interactions Questionnaire [53], the Froehlich Communication Survey [31], and the Interprofessional Collaborative Competencies Attainment Survey (ICCAS) [31]. Observation and analysis tools accounted for 20 out of 101 instruments (20%), including conversation and thematic analysis, microanalysis, and structured observation checklists. Interviews and focus groups were used in 12 out of 101 instruments (12%), primarily in the form of focus group discussions and semistructured interviews. Technical and performance

XSL•FO
RenderX

assessments were documented in 12 out of 101 instruments (12%), predominantly as "objective and structured assessments" (n=8, 8%), such as the Bariatric Objective Structured Assessment of Technical Skill (BOSATS) and the Objective Structured Assessment of Technical Skill (OSATS) [28,62]. Performance metrics were documented in 4 out of 101 instruments (4%), focusing on clinical documentation and diagnostic accuracy. For a comprehensive list of instruments and their characteristics, please refer to Table 6.

Table . Categorization of instruments of the included studies by category and subcategory (n=101).

Category/subcategory	Studies, n (%)
Questionnaires and surveys	
General surveys	28 (28)
Specific topic-based questionnaires	6 (6)
Observations and analysis tools	
Conversation and thematic analysis	8 (8)
Transcription and coding	6 (6)
Microanalysis and critical thinking	3 (3)
Observation tools	3 (3)
Interviews and focus groups	
Focus group discussions	6 (6)
Semistructured interviews	5 (5)
Video elicitation interviews	1 (1)
Technical and performance assessments	
Objective and structured assessments	8 (8)
Performance metrics	4 (4)
Peer and self-assessment	
Self-assessments	7 (7)
Peer and faculty assessments	2 (2)
Feedback and satisfaction tools	
Feedback mechanisms	6 (6)
Peer-led approaches	1 (1)
Grading and rubrics	
Grading tools	3 (3)
Data analytics	
Activity logs and analytics	2 (2)
Specialized instruments and scales	
Specialized questionnaires and scales	5 (5)

Ethical Implications and Logistical Challenges

Ethical Implications and Risk Mapping

Out of the 69 studies included in this scoping review, only 34 (49%) explicitly addressed ethical considerations, which were grouped into 6 primary domains, as summarized in Table 7. The detailed list of ethical considerations can be found in Multimedia Appendix 7.

XSL•FO
RenderX

Table . Categorization of ethical considerations addressed in the included studies (n=69).

Ethical considerations	Definition	Studies, n (%)
Informed consent	Process for obtaining explicit permission from patients before recording their interactions with clinicians	33 (48)
Confidentiality and privacy	Ensure the protection of patient information through secure handling and anonymization	33 (48)
Ethical use of recordings	Ensure that recorded patient-clinician interactions are used only for their intended educational or research purposes, with clear guidelines and appropriate consent for any secondary use	9 (13)
Emotional impact	Recognizing and mitigating the psychological effects that recording clinical interactions may have on patients and clinicians, such as stress, discomfort, or altered behavior	8 (12)
Regulatory compliance	Adhere to legal and institutional standards, such as privacy laws and ethical guidelines, to ensure the proper handling, storage, and use of recordings in clinical and educational settings	6 (9)

Informed consent and *confidentiality and privacy* were reported in 33 (48%) out of 69 studies each, reflecting a moderate emphasis on obtaining explicit permission and protecting data. The *ethical use of recordings* was discussed in 9 out of 69 studies (13%), focusing on restricting use to approved educational or research purposes. Of the 69 studies, 8 (12%) considered the emotional impact of recording on participants, and 6 (9%) mentioned compliance with regulations and institutional policies.

Ethical considerations were categorized into 3 risk levels based on prevalence: low risk (addressed in more than 75% of relevant studies), moderate risk (addressed in 50% - 75% of studies), and high risk (addressed in less than 50% of studies), as illustrated in Figure 3. Informed consent and confidentiality and privacy were classified as low risk relative to studies that addressed ethical issues but as high risk in absolute terms, indicating that these considerations were omitted in over half of the total sample (Table 8). Other considerations, including emotional impact, ethical use of recordings, and regulatory compliance, were classified as high risk under both relative and absolute criteria.

Figure 3. Flowchart of risk mapping for ethical considerations in the included studies.

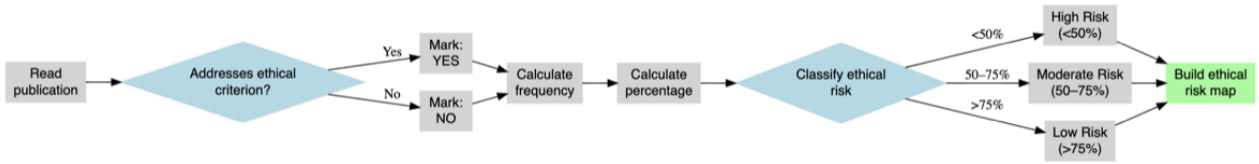


Table . Risk map of ethical considerations in the included studies (n=69).

Ethical considerations	Risk level (relative)	Relative %	Risk level (absolute)	Absolute %
Informed consent	Low risk	100	High risk	48
Confidentiality and privacy	Low risk	100	High risk	48
Emotional impact	High risk	24	High risk	12
Ethical use of recordings	High risk	27	High risk	13
Regulatory compliance	High risk	18	High risk	9

Logistical Challenges

A total of 69 studies were included in this scoping review, of which 33 (48%) addressed logistical challenges. These challenges were categorized into 5 core domains, as detailed in Table 9. A comprehensive list of these challenges can be found in Multimedia Appendix 8.

As shown in Table 9, the most frequently reported challenges across studies were technical setup (33/69, 48%) and scheduling coordination (30/69, 43%). Additionally, less frequently reported

XSL•FO
RenderX

challenges included resource allocation (14/69, 20%), training for participants (8/69, 12%), and feedback mechanisms (6/69, 9%). These findings underscore the need to address these key areas to optimize implementation processes, as technical setup and scheduling coordination are critical factors that impact study outcomes.

Table . Categorization of logistical considerations addressed in the included studies (n=69).

Logistical consideration	Definition	Studies, n (%)
Technical setup	Ensure that recording equipment is properly set up and functioning, including video/audio recording technology and connectivity.	33 (48)
Scheduling coordination	Coordinate patient, clinician, and technical staff schedules to ensure recording sessions are timely and uninterrupted.	30 (43)
Resource allocation	Manage the allocation of technical resources, including equipment, time, and teacher support for recording sessions.	14 (20)
Training for participants	Provide appropriate training for staff, clinicians, and students in the use of recording technology and protocols.	8 (12)
Feedback mechanisms	Organize feedback sessions and ensure appropriate processes are in place to provide reflective feedback to participants.	6 (9)

Discussion

This scoping review systematically mapped the use of video recordings of PCIs in health education, with a focus on the following areas: types of recordings, educational purposes, comparative effectiveness, impact on learner outcomes, and key ethical and logistical considerations.

Principal Findings

This scoping review identified 69 studies examining the use of video recordings of PCIs for health education purposes. The majority of these studies focused on medical education (28/69, 41%), followed by dental education (15/69, 22%) and nursing education (8/69, 12%). This highlights the widespread emphasis on clinical skill development across health disciplines. Of the 69 included studies, 41 (59%) were published between 2011 and 2020, reflecting a growing interest in video-based pedagogical tools. The analyzed interventions included real and simulated patient-clinician encounters, with real clinical interactions predominating.

Video-based learning was generally associated with improvements in key educational outcomes, particularly in the domains of clinical competence, communication skills, and learner engagement. Where available, comparisons with traditional instructional methods suggested that video recordings may provide advantages or complementary benefits for experiential and reflective learning. However, the heterogeneity of study designs, outcome measurements, and assessment tools limited the comparability of the findings, preventing definitive conclusions from being drawn regarding their overall efficacy of video-based learning.

Ethical considerations, particularly informed consent and confidentiality, were addressed inconsistently across studies, mapping these as high-risk areas. Logistical challenges, including technological limitations and scheduling constraints, were also commonly reported. Despite these limitations, video recordings of PCIs were consistently perceived by educators and learners alike as valuable, versatile, and pedagogically effective tools in the context of contemporary health education.

Comparison to Prior Work

The findings of this scoping review indicate that video recordings of PCIs are increasingly incorporated into health professions education and are widely regarded as valuable pedagogical tools for supporting the development of core clinical competencies. The reviewed literature highlights the use of PCI recordings across a variety of educational objectives, including enabling learners to observe complex, real-world clinical interactions; fostering reflective practice; and providing structured opportunities for formative feedback [95,96]. These uses are indicative of a more general pedagogical shift toward learner-centered, active educational strategies, which are grounded in well-established theoretical frameworks such as experiential learning, social learning, and constructivism [3,97,98].

PCI recordings facilitate the experiential learning theory model by enabling learners to engage with authentic clinical scenarios in a safe and repeatable format, promoting both observation and reflection prior to active clinical engagement [99,100]. In a similar vein, Bandura's social learning theory emphasizes the significance of modeling and observational learning in the acquisition of behavior [98]. PCI recordings facilitate the observation of expert communication strategies, diagnostic reasoning, and interpersonal conduct in action, thereby providing a reference point for the learner's own professional practice [20].

The extant literature also positions PCI recordings as an effective bridge between theoretical instruction and clinical application. This function is underpinned by constructivist learning theory, which posits that learners build knowledge through contextual engagement and integration of new information into existing

XSL•FO
RenderX

cognitive frameworks [101,102]. PCI recordings situate learning within the complexity of real clinical contexts, offering nuanced representations of uncertainty, interpersonal dynamics, and decision-making. Furthermore, the principles of cognitive load theory provide additional support for the use of video as a tool to scaffold complex learning [100,102]. The provision of multimodal content, encompassing visual, auditory, and contextual elements, is a salient feature of PCI recordings. This multimodal approach has been demonstrated to facilitate the management of both intrinsic and extraneous cognitive load, a phenomenon that is particularly pronounced when these recordings are used in conjunction with instructional scaffolding or expert narration [12,103,104].

It is important to note that PCI recordings also support the development of metacognitive skills and professional identity formation. The observation of actual clinicians engaged in complex patient interactions serves to encourage learners to engage in reflection on their future roles, values, and behaviors as health professionals [105,106]. This approach is in alignment with the emerging educational goals that place a premium not only on competence, but also on ethical reasoning, self-awareness, and identity development [107-110]. Consequently, the use of PCI recordings transcends the delivery of content, thereby supporting the comprehensive professional development of learners.

Strengths and Limitations

This scoping review is strengthened by its comprehensive, methodologically rigorous design following Joanna Briggs Institute guidelines. A comprehensive search strategy encompassing multiple databases (ie, MEDLINE, Embase, and ERIC) and gray literature was used to enhance the scope of the study and mitigate publication bias. The implementation of dual independent screening and data extraction processes served to ensure the consistency of the review and to minimize the impact of reviewer bias.

This study focuses specifically on video recordings of authentic or simulated PCIs, rather than generic video tools. This approach allows for a detailed examination of unique pedagogical, ethical, and implementation issues. The incorporation of studies from diverse health professions has been demonstrated to enhance the generalizability of research findings.

It is important to note several limitations of the study. First, substantial heterogeneity in study design, interventions, and outcomes limited the comparability of the studies, and thus prevented meta-analysis. The paucity of standardized, validated outcome measures, coupled with the dearth of direct comparisons with traditional methods, serves to impede the ability to draw definitive conclusions regarding the effectiveness of the latter. Second, the majority of studies assessed short-term outcomes exclusively, thereby leaving the question of long-term impacts on clinical performance and patient care unanswered. Third, ethical and logistical considerations were inconsistently reported, reflecting variable adherence to institutional and regulatory frameworks (eg, HIPAA), which complicates establishing best practices. Fourth, practical barriers such as limited access to recording, privacy concerns, faculty training gaps, and administrative burdens were frequently mentioned but insufficiently detailed. Finally, the restriction of the review to English-language publications may have resulted in the exclusion of pertinent non-English studies, and some unpublished or nonindexed evidence may have been overlooked.

Future Directions

It is recommended that future research efforts concentrate on the development and widespread adoption of standardized, validated outcome measures. The objective of this recommendation is to enable rigorous evaluation and meta-analytic synthesis of PCI-based educational interventions. The necessity for well-designed longitudinal and controlled studies is evident in order to elucidate the sustained effects of PCI recordings on clinical competence, behavioral change, and patient-centered outcomes.

There is a need for improvement in the consistency of reporting and adherence to ethical standards, including informed consent and data protection protocols. The establishment of formal institutional policies is necessary to ensure regulatory compliance and to safeguard the rights of participants. In order to achieve this, a coordinated effort is required among educators, clinicians, institutional review boards, and legal authorities to develop ethical frameworks that are scalable and practical. These frameworks must balance educational efficacy with legal and moral obligations. Empirical evaluation of these frameworks' impact on feasibility, learner outcomes, and patient perceptions is warranted to inform implementation best practices.

In order to address logistical barriers, there is a necessity for institutional commitment to resource allocation, the development of scalable technological solutions, comprehensive faculty development, and streamlined administrative processes. The identification and dissemination of effective implementation strategies will be vital for the successful integration of PCI recordings into health professions curricula.

Furthermore, the expansion of the research scope to encompass a range of health professions, educational settings, and non-English language publications is expected to enhance the external validity and global applicability of findings.

Collectively, these directions are essential to advance the evidence base and optimize the integration of PCI recordings within health professions education.

Conclusions

This scoping review systematically maps the current landscape of video recordings of PCIs in health education. A synthesis of 69 studies shows that video-based methodologies are being adopted more widely in medical, dental, and nursing curricula. There is robust evidence that these methodologies are more effective than traditional methods in improving clinical skills, communication, and learner engagement.

However, despite these educational benefits, the review identifies critical and persistent gaps in ethical and logistical domains. Informed consent and confidentiality were addressed inconsistently, highlighting the need for standardized ethical protocols and rigorous safeguarding of patient data. Logistical barriers, including technical setup, scheduling, and resource

XSL•FO
RenderX

allocation, remain significant impediments to widespread implementation.

A notable finding is the marked heterogeneity in outcome measures and assessment instruments, which limits comparability and synthesis across studies. This emphasizes the urgent need for the development and adoption of standardized, validated outcome sets and measurement tools in future research.

Taken together, these findings emphasize the transformative potential and complex challenges of integrating video recordings into health education. Responsible and effective use of this technology depends on establishing robust ethical frameworks, standardized methodologies, and institutional commitment to best practices. Prioritizing these issues will be crucial for maximizing the educational impact and ensuring the sustainability of video-based learning in health professions education.

Funding

This work was supported by FCT (Fundação para a Ciência e a Tecnologia, I.P.) by project reference UID/4585/2025.

Data Availability

The data that support the findings of this study are available from the corresponding author upon reasonable request. In addition, datasets and supplementary materials are provided as multimedia appendices with this article.

Authors' Contributions

LP and MM conceptualized the study. MM and LP curated the data. Formal analysis was performed by MM, TL, and YL. Funding acquisition was managed by JJM. Investigation was conducted by MM and TL. Methodology was developed by MM, LP, and JJM. Project administration was carried out by MM and LP. Resources were provided by MM and JJM. Software was managed by MM and LP. Supervision was provided by LP and JJM. Validation was performed by LP and JJM. Visualization was completed by MM, YL, and TL. MM and LP wrote the original draft. MM, YL, LP, JJM, and TL reviewed and edited the manuscript. All authors have read and approved the final version.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Search strategy.
[DOCX File, 29 KB - mededu_v12i1e70324_app1.docx]

Multimedia Appendix 2
Summary of extracted data.
[DOCX File, 76 KB - mededu_v12i1e70324_app2.docx]

Multimedia Appendix 3
Definitions of data items.
[DOCX File, 18 KB - mededu_v12i1e70324_app3.docx]

Multimedia Appendix 4
Categorization of outcome variables of the included studies by author and year, category, subcategory, general outcome variable, and specific outcome variable.
[DOCX File, 45 KB - mededu_v12i1e70324_app4.docx]

Multimedia Appendix 5
Categorization of outcome variables of the included studies by category, subcategory, and general outcome variable.
[DOCX File, 35 KB - mededu_v12i1e70324_app5.docx]

Multimedia Appendix 6
Categorization of specific instruments of the included studies by author and year, category, and subcategory.
[DOCX File, 25 KB - mededu_v12i1e70324_app6.docx]

Multimedia Appendix 7
Categorization of ethical considerations addressed in the included studies.

XSL•FO
RenderX

[DOCX File, 19 KB - mededu_v12i1e70324_app7.docx]

Multimedia Appendix 8
Categorization of logistical considerations addressed in the included studies.
[DOCX File, 18 KB - mededu_v12i1e70324_app8.docx]

Checklist 1
PRISMA-ScR checklist.
[DOCX File, 119 KB - mededu_v12i1e70324_app9.docx]

Checklist 2
CADTH Grey Matters checklist.
[DOC File, 1148 KB - mededu_v12i1e70324_app10.doc]

References

1. Bull PH. Cognitive constructivist theory of multimedia: designing teacher-made interactive digital. Creat Educ 2013;04(9):614-619. [doi: 10.4236/ce.2013.49088]
2. Kaufman DM. Teaching and learning in medical education: how theory can inform practice. In: Swanwick T, Forrest K, O'Brien BC, editors. Understanding Medical Education: Evidence, Theory, and Practice: Wiley; 2018:37-69. [doi: 10.1002/9781119373780]
3. Dong H, Lio J, Sherer R, Jiang I. Some learning theories for medical educators. Med Sci Educ 2021 Mar 22;31(3):1157-1172. [doi: 10.1007/s40670-021-01270-6] [Medline: 34457959]
4. Mutlu-Bayraktar D, Cosgun V, Altan T. Cognitive load in multimedia learning environments: a systematic review. Comput Educ 2019 Nov;141:103618. [doi: 10.1016/j.compedu.2019.103618]
5. Sweller J. Implications of cognitive load theory for multimedia learning. In: Mayer R, editor. The Cambridge Handbook of Multimedia Learning: Cambridge University Press; 2005:19-30. [doi: 10.1017/CBO9780511816819.003]
6. Harris IB. Conceptions and theories of learning for workplace education. In: Extraordinary Learning in the Workplace: Springer Nature; 2011:39-62. [doi: 10.1007/978-94-007-0271-4_3]
7. Morris TH. Experiential learning – a systematic review and revision of Kolb's model. Interact Learn Environ 2020 Nov 16;28(8):1064-1077. [doi: 10.1080/10494820.2019.1570279]
8. Kolb DA, Boyatzis RE, Mainemelis C. Experiential learning theory: previous research and new directions. In: Sternberg RJ, Zhang L, editors. Perspectives on Thinking, Learning, and Cognitive Styles: Routledge; 2015:227-248. [doi: 10.4324/9781410605986]
9. Mann KV. Theoretical perspectives in medical education: past experience and future possibilities. Med Educ 2011 Jan;45(1):60-68. [doi: 10.1111/j.1365-2923.2010.03757.x] [Medline: 21155869]
10. Stanley MJ, Banks S, Matthew W, Brown S. Operationalization of Bandura's social learning theory to guide interprofessional simulation. J Nurs Educ Pract 2020;10(10):61. [doi: 10.5430/jnep.v10n10p61]
11. Chauvin SW. Applying educational theory to simulation-based training and assessment in surgery. Surg Clin North Am 2015 Aug;95(4):695-715. [doi: 10.1016/j.suc.2015.04.006] [Medline: 26210964]
12. Brame CJ. Effective educational videos: principles and guidelines for maximizing student learning from video content. CBE Life Sci Educ 2017 Oct 13;15(4):es6. [doi: 10.1187/cbe.16-03-0125] [Medline: 27789532]
13. Butler DJ. A review of published guidance for video recording in medical education. Fam Syst Health 2018 Mar;36(1):4-16. [doi: 10.1037/fsh0000328] [Medline: 29369649]
14. Chowdhury PN, Vaish A, Puri B, Vaishya R. Medical education technology: past, present and future. Apollo Med 2024 Dec;21(4):374-380. [doi: 10.1177/09760016241256202]
15. Roland D, Coats TJ, Matheson D. Towards a conceptual framework demonstrating the effectiveness of audiovisual patient descriptions (patient video cases): a review of the current literature. BMC Med Educ 2012 Dec 21;12(1):125. [doi: 10.1186/1472-6920-12-125] [Medline: 23256787]
16. Salina L, Ruffinengo C, Garrino L, et al. Effectiveness of an educational video as an instrument to refresh and reinforce the learning of a nursing technique: a randomized controlled trial. Perspect Med Educ 2012 May;1(2):67-75. [doi: 10.1007/s40037-012-0013-4] [Medline: 23316461]
17. Glosser LD, Lombardi CV, Hopper WA, et al. Impact of educational instruction on medical student performance in simulation patient. Int J Med Educ 2022 Jun 23;13:158-170. [doi: 10.5116/ijme.62a5.96bf] [Medline: 35752175]
18. Clerkin R, Patton D, Moore Z, Nugent L, Avsar P, O'Connor T. What is the impact of video as a teaching method on achieving psychomotor skills in nursing? A systematic review and meta-analysis. Nurse Educ Today 2022 Apr;111:105280. [doi: 10.1016/j.nedt.2022.105280] [Medline: 35139443]
19. O'Reilly M, Parker N, Hutchby I. Ongoing processes of managing consent: the empirical ethics of using video-recording in clinical practice and research. Clin Ethics 2011 Dec 5;6(4):179-185. [doi: 10.1258/ce.2011.011040]

XSL•FO
RenderX

20. Henry SG, White AEC, Magnan EM, et al. Making the most of video recorded clinical encounters: optimizing impact and productivity through interdisciplinary teamwork. Patient Educ Couns 2020 Oct;103(10):2178-2184. [doi: 10.1016/j.pec.2020.06.005] [Medline: 32576422]
21. Gabrielli M, Valera L, Barrientos M. Audio and panoramic video recording in the operating room: legal and ethical perspectives. J Med Ethics 2022 Jan 7;47(12):798-802. [doi: 10.1136/medethics-2019-106056]
22. Lloyd A, Lowe DJ, Edgar S, Caesar D, Dewar A, Clegg GR. Video recording in the emergency department: a pathway to success. Emerg Med J 2017 Sep;34(9):628-630. [doi: 10.1136/emermed-2017-206731] [Medline: 28659303]
23. Brogaard L, Uldbjerg N. Filming for auditing of real-life emergency teams: a systematic review. BMJ Open Qual 2019 Dec 6;8(4):e000588. [doi: 10.1136/bmjoq-2018-000588] [Medline: 31909207]
24. Peters MDJ, Godfrey CM, Khalil H, McInerney P, Parker D, Soares CB. Guidance for conducting systematic scoping reviews. Int J Evid Based Healthc 2015 Sep;13(3):141-146. [doi: 10.1097/XEB.0000000000000050] [Medline: 26134548]
25. Aromataris E, Munn Z, editors. Scoping reviews. In: JBI Manual for Evidence Synthesis: JBI; 2020:407-453. [doi: 10.46658/JBIMES-20-12]
26. Tricco AC, Lillie E, Zarin W, et al. PRISMA extension for scoping reviews (PRISMA-ScR): checklist and explanation. Ann Intern Med 2018 Oct 2;169(7):467-473. [doi: 10.7326/M18-0850] [Medline: 30178033]
27. Morgado M. Video recordings of patient-clinician interactions in health education. : OSF; 2024 Oct 31. [doi: 10.17605/OSF.IO/8W3M6]
28. Ahmet A, Gamze K, Rustem M, Sezen KA. Is video-based education an effective method in surgical education? A systematic review. J Surg Educ 2018;75(5):1150-1158. [doi: 10.1016/j.jsurg.2018.01.014] [Medline: 29449162]
29. Alsalamah YS, Alharbi MS, Labani S, et al. Video-based simulation among Saudi undergraduate nursing students during COVID-19: a qualitative study. Cureus 2023 Nov 10;15(11):e48605. [doi: 10.7759/cureus.48605] [Medline: 38084198]
30. Balslev T, de Grave WS, Muijtjens AMM, Scherpbier A. Comparison of text and video cases in a postgraduate problem-based learning format. Med Educ 2005 Nov;39(11):1086-1092. [doi: 10.1111/j.1365-2929.2005.02314.x] [Medline: 16262803]
31. Batteson TJ, Chirica MG, Fox KS, et al. Utilizing video simulation in interprofessional education to teach IPEC competencies and social determinants of health: a universal IPE teaching tool. J Interprof Educ Pract 2023 Jun;31:100617. [doi: 10.1016/j.xjep.2023.100617]
32. Bessette H, Kapanen AI, Medgyesi N, Yuen J. Learner perceptions of a video recording activity for patient consultation training in a pharmacist-led primary care clinic located in an academic institution. Innov Pharm 2021 Jun 29;12(3):3. [doi: 10.24926/iip.v12i3.3632]
33. Bethea DP, Smith N, Allison LK, et al. Live standardized patient scenario improves attitudes toward and readiness for interprofessional education in occupational therapy and physical therapy students. J Allied Health 2019;48(2):81-87. [Medline: 31167008]
34. Botelho M. The communal consultation: video-recorded student consultations. Med Educ 2016 May;50(5):573-574. [doi: 10.1111/medu.13047] [Medline: 27072459]
35. Bowles K, Sehgal HS, Santelmann L, Pham EP, Kohli R. Dental student to patient communication analysis: a pilot study. Health Commun 2020 Jan;35(1):10-17. [doi: 10.1080/10410236.2018.1536953] [Medline: 30358413]
36. Chan LK, Patil NG, Chen JY, Lam JCM, Lau CS, Ip MSM. Advantages of video trigger in problem-based learning. Med Teach 2010;32(9):760-765. [doi: 10.3109/01421591003686260] [Medline: 20795807]
37. Chi DL, Pickrell JE, Riedy CA. Student learning outcomes associated with video vs. paper cases in a public health dentistry course. J Dent Educ 2014 Jan;78(1):24-30. [Medline: 24385521]
38. Courteille O, Josephson A, Larsson LO. Interpersonal behaviors and socioemotional interaction of medical students in a virtual clinical encounter. BMC Med Educ 2014 Apr 1;14:64. [doi: 10.1186/1472-6920-14-64] [Medline: 24685070]
39. Davies BR, Leung AN, Dunne SM, Dillon J, Blum IR. Bespoke video vignettes – an approach to enhancing reflective learning developed by dental undergraduates and their clinical teachers. Eur J Dental Educ 2017 Feb;21(1):33-36. [doi: 10.1111/eje.12175]
40. Edrees HY, Ohlin J, Ahlquist M, Tessma MK, Zary N. Patient demonstration videos in predoctoral endodontic education: aspects perceived as beneficial by students. J Dent Educ 2015 Aug;79(8):928-933. [doi: 10.1002/J.0022-0337.2015.79.8.TB05983.X] [Medline: 26246531]
41. Farnan JM, O'Leary KJ, Didwania A, et al. Promoting professionalism via a video-based educational workshop for academic hospitalists and housestaff. J Hosp Med 2013 Jul;8(7):386-389. [doi: 10.1002/jhm.2056] [Medline: 23780912]
42. Fero LJ, O'Donnell JM, Zullo TG, et al. Critical thinking skills in nursing students: comparison of simulation-based performance with metrics. J Adv Nurs 2010 Oct;66(10):2182-2193. [doi: 10.1111/j.1365-2648.2010.05385.x] [Medline: 20636471]
43. Flood M, Ryan B, Conroy A, Strawbridge J. Video killed the...the development of video-based case studies for more patient-centred, authentic assessment of postgraduate pharmacy students. In: EDULEARN19 Proceedings: IATED Digital Library; 2019:7419. [doi: 10.21125/edulearn.2019.1775]
44. Forbes H, Bucknall TK, Hutchinson AM. Piloting the feasibility of head-mounted video technology to augment student feedback during simulated clinical decision-making: an observational design pilot study. Nurse Educ Today 2016 Apr;39:116-121. [doi: 10.1016/j.nedt.2016.01.012] [Medline: 27006042]

XSL•FO
RenderX

45. Giles AK, Carson NE, Breland HL, Coker-Bolt P, Bowman PJ. Use of simulated patients and reflective video analysis to assess occupational therapy students' preparedness for fieldwork. Am J Occup Ther 2014;68(Suppl 2):S57-S66. [doi: 10.5014/ajot.2014.685S03] [Medline: 25397940]
46. Hafen M, Drake AAS, Rush BR, Nelson SC. Using authentic client interactions in communication skills training: predictors of proficiency. J Vet Med Educ 2013;40(4):318-326. [doi: 10.3138/jvme.0113-019R] [Medline: 24113724]
47. Hafen M Jr, Siqueira Drake AA, Rush BR, Sibley DS. Engaging students: using video clips of authentic client interactions in pre-clinical veterinary medical education. J Vet Med Educ 2015;42(3):252-258. [doi: 10.3138/jvme.0614-059R1] [Medline: 26075626]
48. Lundeborg Hammarström I, Samuelsson C. Speech and language interventions for stroke-induced aphasia. J Interact Res Commun Disord 2021 Mar 30;10(1):47-66. [doi: 10.1558/jircd.19317]
49. Hammoud MM, Morgan HK, Edwards ME, Lyon JA, White C. Is video review of patient encounters an effective tool for medical student learning? A review of the literature. Adv Med Educ Pract 2012 Mar;3:19-30. [doi: 10.2147/AMEP.S20219] [Medline: 23761999]
50. Henry SG, Fetters MD. Video elicitation interviews: a qualitative research method for investigating physician-patient interactions. Ann Fam Med 2012;10(2):118-125. [doi: 10.1370/afm.1339] [Medline: 22412003]
51. Janda MS, Mattheos N, Nattestad A, et al. Simulation of patient encounters using a virtual patient in periodontology instruction of dental students: design, usability, and learning effect in history‐taking skills. Eur J Dent Educ 2004 Aug;8(3):111-119. [doi: 10.1111/j.1600-0579.2004.00339.x]
52. Ju M, Hwang J, Kim J, Kang J. The effect of recorded video monitoring on students' self reflection of patient‐physician interaction. Korean Med Educ Rev 2017 Jun;19(2):83-89. [doi: 10.17496/kmer.2017.19.2.83]
53. Kalish R, Dawiskiba M, Sung Y, Blanco M. Raising medical student awareness of compassionate care through reflection of annotated videotapes of clinical encounters. Educ Health 2011;24(3):490. [doi: 10.4103/1357-6283.101427]
54. Kalwitzki M. Self‐reported changes in clinical behaviour by undergraduate dental students after video‐based teaching in paediatric dentistry. Eur J Dental Education 2005 Aug;9(3):108-114 [FREE Full text] [doi: 10.1111/j.1600-0579.2005.00372.x]
55. Kamin C, O'Sullivan P, Deterding R, Younger M. A comparison of critical thinking in groups of third-year medical students in text, video, and virtual PBL case modalities. Acad Med 2003 Feb;78(2):204-211. [doi: 10.1097/00001888-200302000-00018] [Medline: 12584102]
56. Lee WS, Hwang JY, Lim JE, Suh SY, Park KH, Sung NJ. The effect of videotaping students' interviews with patients for interview skill education. Korean J Fam Med 2013 Mar;34(2):90-97. [doi: 10.4082/kjfm.2013.34.2.90] [Medline: 23560207]
57. Leeds FS, Sommer EM, Andrasik WJ, Atwa KM, Crawford TN. A patient-narrative video approach to teaching fibromyalgia. J Med Educ Curric Dev 2020 Jul 31;7. [doi: 10.1177/2382120520947068] [Medline: 32821851]
58. De Leng B, Dolmans D, Van de Wiel M, Muijtjens A, Van Der Vleuten C. How video cases should be used as authentic stimuli in problem‐based medical education. Med Educ 2007 Feb;41(2):181-188. [doi: 10.1111/j.1365-2929.2006.02671.x]
59. Leone TA, Rich W, Finer NN. Educational perspectives: addressing the complexities of medical education with video. Neoreviews 2006 Dec 1;7(12):e597-e601. [doi: 10.1542/neo.7-12-e597]
60. Lewis A, Moore C, Nang C. Using video of student-client interactions to engage students in reflection and peer review. J Univ Teach Learn Pract 2015 Sep 1;12(4). [doi: 10.53761/1.12.4.7]
61. Malon M, Cortes D, Greisen GO. Medical students' assessment of pediatric patients – teaching and evaluation using video cases. BMC Med Educ 2014 Nov 13;14:241. [doi: 10.1186/s12909-014-0241-x] [Medline: 25391289]
62. McQueen S, McKinnon V, VanderBeek L, McCarthy C, Sonnadara R. Video-based assessment in surgical education: a scoping review. J Surg Educ 2019;76(6):1645-1654. [doi: 10.1016/j.jsurg.2019.05.013] [Medline: 31175065]
63. Miller CJ, Metz MJ. Can clinical scenario videos improve dental students' perceptions of the basic sciences and ability to apply content knowledge? J Dent Educ 2015 Dec;79(12):1452-1460. [Medline: 26632300]
64. Minardi HA, Ritter S. Recording skills practice on videotape can enhance learning – a comparative study between nurse lecturers and nursing students. J Adv Nurs 1999 Jun;29(6):1318-1325. [doi: 10.1046/j.1365-2648.1999.01017.x] [Medline: 10354225]
65. Muench J, Sanchez D, Garvin R. A review of video review: new processes for the 21st century. Int J Psychiatry Med 2013;45(4):413-422. [doi: 10.2190/PM.45.4.k] [Medline: 24261274]
66. Murphy LF, Stav WB. The impact of online video cases on clinical reasoning in occupational therapy education: a quantitative analysis. Open J Occup Ther 2018 Jul;6(3). [doi: 10.15453/2168-6408.1494]
67. Nilsen S, Baerheim A. Feedback on video recorded consultations in medical teaching: why students loathe and love it – a focus-group based qualitative study. BMC Med Educ 2005 Jul 19;5:28. [doi: 10.1186/1472-6920-5-28] [Medline: 16029509]
68. Nissen K, Chipchase L, Conroy T, Farrer O. The impact of using an authentic patient video on health professional students' attitudes toward interprofessional and person-centered care. J Interprof Care 2024;38(4):722-728. [doi: 10.1080/13561820.2024.2334959] [Medline: 38600788]
69. Noverati N, R Naro G, J Fischer R, M Thompson B. Using video and virtual patients in problem-based learning: a scoping review. Med Sci Educ 2020 Dec;30(4):1685-1691. [doi: 10.1007/s40670-020-01108-7] [Medline: 34457832]

XSL•FO
RenderX

70. Nunohara K, Imafuku R, Saiki T, et al. How does video case-based learning influence clinical decision-making by midwifery students? An exploratory study. BMC Med Educ 2020 Mar 6;20(1):67. [doi: 10.1186/s12909-020-1969-0] [Medline: 32143611]
71. Nyström A, Pålsson Y, Hofsten A, Häggström E. Nursing students' experiences of being video-recorded during examination in a fictive emergency care situation. Int J Nurs Pract 2014 Oct;20(5):540-548. [doi: 10.1111/ijn.12199] [Medline: 24118615]
72. Omar H, Archer L, Schroeder K, Robinson B. Web‐based standardized patient encounters for preclinical dental students during COVID‐19 pandemic. J Dent Educ 2021 Jun;85(S1):1167-1169. [doi: 10.1002/jdd.12401]
73. Oosthuizen H. Speech therapy students' perceptions of authentic video cases in a theory module on child language disorders. S Afr J Commun Disord 2019 Jan 22;66(1):e1-e5. [doi: 10.4102/sajcd.v66i1.602] [Medline: 30843410]
74. Parlak Özer Z, Balaman U. An investigation into student dietitians' professional development through video-mediated communication training on patient counselling. Health Commun 2025 Apr;40(4):574-584. [doi: 10.1080/10410236.2024.2355440] [Medline: 38773766]
75. Quinn S, Herron D, Menzies R, et al. The video interaction guidance approach applied to teaching communication skills in dentistry. Eur J Dent Educ 2016 May;20(2):94-101. [doi: 10.1111/eje.12146] [Medline: 25891196]
76. Raja K. Use of videos as a medium of clinical teaching in undergraduate physiotherapy students in India – a randomized single blinded controlled pilot study. Indian J Physiother Occup Ther 2008;2(2):15-19 [FREE Full text]
77. Reher V, Rehbein G, Reher P, et al. Integrating video recording and self-reflection to enhance communication skills training for dental students. In: Van Toi V, Le T, Ngo H, Nguyen TH, editors. 7th International Conference on the Development of Biomedical Engineering in Vietnam (BME7): Springer; 2020:715-719. [doi: 10.1007/978-981-13-5859-3_120]
78. Roberts JK, Sullivan M, Atwater S, et al. Use of virtual interactive patient encounters to prepare first-year medical students for clinical practice. Acad Med 2023 Oct 1;98(10):1146-1153. [doi: 10.1097/ACM.0000000000005286] [Medline: 37257029]
79. Rodríguez-Bailón M, Fernández-Solano AJ, Merchán-Baeza JA, Vidaña-Moya L. From clinical practice to the classroom. Advantages and disadvantages of video and paper cases on the motivation and clinical reasoning of occupational therapy students. Int J Environ Res Public Health 2021 Sep 14;18(18):9671. [doi: 10.3390/ijerph18189671] [Medline: 34574609]
80. Roland D, Balslev T. Use of patient video cases in medical education. Arch Dis Child Educ Pract Ed 2015 Aug;100(4):210-214. [doi: 10.1136/archdischild-2014-308030] [Medline: 25972599]
81. Basu Roy R, McMahon GT. Video-based cases disrupt deep critical thinking in problem-based learning. Med Educ 2012 Apr;46(4):426-435. [doi: 10.1111/j.1365-2923.2011.04197.x] [Medline: 22429179]
82. Sanders C, Kleinert HL, Boyd SE, Herren C, Theiss L, Mink J. Virtual patient instruction for dental students: can it improve dental care access for persons with special needs? Spec Care Dentist 2008;28(5):205-213. [doi: 10.1111/j.1754-4505.2008.00038.x] [Medline: 18782198]
83. Sanderson TR, Kearney RC, Kissell D, Salisbury J. Evaluating student self-assessment through video-recorded patient simulations. J Dent Hyg 2016 Aug;90(4):257-262. [Medline: 27551147]
84. Scheidt PC, Lazoritz S, Ebbeling WL, Figelman AR, Moessner HF, Singer JE. Evaluation of system providing feedback to students on videotaped patient encounters. J Med Educ 1986 Jul;61(7):585-590. [doi: 10.1097/00001888-198607000-00006]
85. Schwartz B, Bohay R. Can patients help teach professionalism and empathy to dental students? Adding patient videos to a lecture course. J Dent Educ 2012 Feb;76(2):174-184. [Medline: 22319082]
86. Temple KL, Giles AK. Using a video case study as a learning tool to increase OT student confidence before fieldwork. Am J Occup Ther 2022 Jul 1;76(Suppl 1). [doi: 10.5014/ajot.2022.76S1-PO69]
87. Terasaki MRT, Morgan CO, Elias L. Medical student interactions with cancer patients: evaluation with videotaped interviews. Med Pediatr Oncol 1984;12(1):38-42. [doi: 10.1002/mpo.2950120110] [Medline: 6700539]
88. Thomas RM, Moore LP, Urquhart BB, et al. Use of simulated telenursing with standardized patients to enhance prelicensure nursing education. Nurse Educ 2023;48(6):E191-E195. [doi: 10.1097/NNE.0000000000001410] [Medline: 37000901]
89. Tully J, Dameff C, Kaib S, Moffitt M. Recording medical students' encounters with standardized patients using Google Glass: providing end-of-life clinical education. Acad Med 2015 Mar;90(3):314-316. [doi: 10.1097/ACM.0000000000000620] [Medline: 25551855]
90. Vessey JA, Huss K. Using standardized patients in advanced practice nursing education. J Prof Nurs 2002;18(1):29-35. [doi: 10.1053/jpnu.2002.30898] [Medline: 11859491]
91. White JG, Krüger C, Snyman WD. Development and implementation of communication skills in dentistry: an example from South Africa. Eur J Dent Educ 2008 Feb;12(1):29-34. [doi: 10.1111/j.1600-0579.2007.00488.x] [Medline: 18257762]
92. Yoon BY, Choi I, Choi S, et al. Using standardized patients versus video cases for representing clinical problems in problem-based learning. Korean J Med Educ 2016 Jun;28(2):169-178. [doi: 10.3946/kjme.2016.24] [Medline: 26923094]
93. Zahl DA, Schrader SM, Edwards PC. Student perspectives on using egocentric video recorded by smart glasses to assess communicative and clinical skills with standardised patients. Eur J Dent Educ 2018 May;22(2):73-79. [doi: 10.1111/eje.12217] [Medline: 27380732]
94. Hafen M, Rush BR, Nelson SC. Utilizing filmed authentic student-client interactions as a communication teaching tool. J Vet Med Educ 2009;36(4):429-435. [doi: 10.3138/jvme.36.4.429] [Medline: 20054082]
95. Huang A, Ryan A, Bearman M, Molloy E. The influence of clinical practice video on learning conversations in postgraduate medical education: a scoping review. FoHPE 2024;25(3):26-43. [doi: 10.11157/fohpe.v25i3.695]

XSL•FO
RenderX

96. van der Vleuten CPM, Schuwirth LWT. Assessing professional competence: from methods to programmes. Med Educ 2005 Mar;39(3):309-317. [doi: 10.1111/j.1365-2929.2005.02094.x] [Medline: 15733167]
97. Yardley S, Teunissen PW, Dornan T. Experiential learning: transforming theory into practice. Med Teach 2012;34(2):161-164. [doi: 10.3109/0142159X.2012.643264] [Medline: 22288996]
98. Bandura A. Social cognitive theory. In: Van Lange PAM, Kruglanski AW, Higgins ET, editors. Handbook of Theories of Social Psychology: SAGE Publications; 2012:349-374. [doi: 10.4135/9781446249215.n18]
99. Choi JJ, Gribben J, Lin M, Abramson EL, Aizer J. Using an experiential learning model to teach clinical reasoning theory and cognitive bias: an evaluation of a first-year medical student curriculum. Med Educ Online 2023 Dec;28(1):2153782. [doi: 10.1080/10872981.2022.2153782] [Medline: 36454201]
100. Kinney A, Nordon-Craft A, Kardong-Edgren S, Kumar A, Thompson A. Digital recordings of a clinical encounter facilitate reflection in physical therapy students and clinicians. Front Med 2025;11(11). [doi: 10.3389/fmed.2024.1516724]
101. Seif GA, Brown D. Video-recorded simulated patient interactions: can they help develop clinical and communication skills in today's learning environment? J Allied Health 2013;42(2):e37-e44. [Medline: 23752242]
102. Niederriter J, Hovland C, Hazelett S, et al. Using the constructivist/active learning theoretical framework to develop and test a simulation-based interprofessional geriatric training curriculum. J Interprof Educ Pract 2020 Jun;19(11):100322. [doi: 10.1016/j.xjep.2020.100322]
103. Mayer RE. Multimedia Learning, 2nd edition: Cambridge University Press; 2009. [doi: 10.1017/CBO9780511811678]
104. van Merriënboer JJG, Sweller J. Cognitive load theory in health professional education: design principles and strategies. Med Educ 2010 Jan;44(1):85-93. [doi: 10.1111/j.1365-2923.2009.03498.x] [Medline: 20078759]
105. Wald HS. Professional identity (trans)formation in medical education: reflection, relationship, resilience. Acad Med 2015 Jun;90(6):701-706. [doi: 10.1097/ACM.0000000000000731] [Medline: 25881651]
106. Sandars J. The use of reflection in medical education: AMEE Guide No. 44. Med Teach 2009 Aug;31(8):685-695. [doi: 10.1080/01421590903050374] [Medline: 19811204]
107. Cruess SL, Cruess SR, Boudreau JD, Snell L, Steinert Y. Reframing medical education to support professional identity formation. Acad Med 2014 Nov;89(11):1446-1451. [doi: 10.1097/ACM.0000000000000427] [Medline: 25054423]
108. Field JC, Cowpe JG, Walmsley AD. The Graduating European Dentist: a new undergraduate curriculum framework. Eur J Dent Educ 2017 Dec;21 Suppl 1(Suppl 1):2-10. [doi: 10.1111/eje.12307] [Medline: 29205782]
109. Field J, Vital S, Dixon J, Murphy D, Davies J. The Graduating European Dentist Curriculum framework: a 7 - year review. Eur J Dent Educ 2025 Feb;29(1):155-161. [doi: 10.1111/eje.13058] [Medline: 39563643]
110. Hodges BD, Ginsburg S, Cruess R, et al. Assessment of professionalism: recommendations from the Ottawa 2010 Conference. Med Teach 2011;33(5):354-363. [doi: 10.3109/0142159X.2011.577300] [Medline: 21517683]

Abbreviations

BOSATS: Bariatric Objective Structured Assessment of Technical Skill
CADTH: Canadian Agency for Drugs and Technologies in Health
ERIC: Education Resources Information Center
HIPAA: Health Insurance Portability and Accountability Act
ICCAS: Interprofessional Collaborative Competencies Attainment Survey
OSATS: Objective Structured Assessment of Technical Skill
PCI: patient-clinician interaction
PICO: population, intervention, comparison, and outcomes
PRISMA-ScR: Preferred Reporting Items for Systematic Reviews and Meta-Analyses Extension for Scoping Reviews

Edited by T Leung; submitted 19.Dec.2024; peer-reviewed by P Shilo, T Gladman; revised version received 26.Jan.2026; accepted 31.Jan.2026; published 13.Jul.2026.

Please cite as:
Morgado M, Leira Y, Leitão T, Mendes JJ, Proença L
Video Recording of Patient-Clinician Interactions in Health Education: Scoping Review
JMIR Med Educ 2026;12:e70324
URL: https://mededu.jmir.org/2026/1/e70324
doi:10.2196/70324

© Mariana Morgado, Yago Leira, Tiago Leitão, José João Mendes, Luís Proença. Originally published in JMIR Medical Education (https://mededu.jmir.org), 13.Jul.2026. This is an open-access article distributed under the terms of the Creative Commons

XSL•FO
RenderX

Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Education, is properly cited. The complete bibliographic information, a link to the original publication on https://mededu.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

AI for Clinical Competency Assessment: Scoping Review of Methods and Applications

Wenjia (Stella) Zhang[1,2], BA; Benjamin B Daniels[1,2], BA, MSc; Carol Mita[3], MLiS; Hoang Nguyen[4], MSc, PhD; David B Duong[2,5], MPH, MD

[1]Department of Global Health and Population, T. H. Chan School of Public Health, Harvard University, 677 Huntington Ave, Boston, MA, United States

[2]Program in Global Primary Health Care, Harvard Medical School, Harvard University, Boston, MA, United States

[3]Francis A. Countway Library of Medicine, Harvard Medical School, Harvard University, Boston, MA, United States

[4]Beth Israel Deaconess Medical Center (BIDMC), Partnership for Health Advancement in Vietnam (HAIVN), Hanoi, Vietnam

[5]Division of Global Health Equity, Brigham and Women's Hospital, Boston, MA, United States

Corresponding Author:
Wenjia (Stella) Zhang, BA
Department of Global Health and Population, T. H. Chan School of Public Health, Harvard University, 677 Huntington Ave, Boston, MA, United States

Abstract

Background: Strengthening the global health workforce is central to achieving universal health coverage, but health systems cannot improve what they cannot measure. Valid and scalable assessment of clinical competency is essential for monitoring workforce readiness and ensuring that expanded service coverage translates into high-quality care. Traditional standardized patients, however, remain resource-intensive, difficult to scale, and vulnerable to evaluator-related bias. Recent advances in AI have enabled AI-led simulated standardized patients (SSPs) that may address these limitations.

Objective: Although existing reviews have examined AI-SSPs broadly within medical education, their use as clinical competency assessment systems remains less well characterized. This scoping review aimed to address this gap by systematically mapping the scope, design features, and validation approaches of AI-SSP tools for clinical competency assessment.

Methods: We registered the protocol prospectively with the Open Science Framework and conducted a scoping review following the Joanna Briggs Institute Manual for Evidence Synthesis. We searched MEDLINE, Embase, CINAHL Complete, Education Source, Web of Science Core Collection, Inspec, ERIC, and Cochrane CENTRAL, supplemented by manual searching. Eligible primary studies involved an AI-based SSP interaction that yielded a quantitative score of respondents' clinical competency. Two reviewers independently screened records and resolved conflicts through discussion. Data were charted on study characteristics and populations, frontend platform and interface features, backend AI architectures, performance scoring mechanisms, and tool evaluation methods and outcomes.

Results: Of 5987 database records and 4 identified through other methods, 21 studies describing 20 unique AI-SSP systems were included. Studies were published between 2008 and 2026, with 15 (71%) published in 2024 or later. Most studies were conducted in high-income settings and involved medical students. Systems shifted over time from rule-based virtual patient to large language model–enabled conversational platforms after 2022. History-taking was the most frequently assessed competency, and checklist coverage was the most common scoring approach. Among studies comparing AI-generated scores with human ratings, findings were mixed: some reported moderate to strong agreement, while others showed score inflation and inconsistent performance. Most studies focused on feasibility; fewer evaluated large-scale use or sustained real-world implementation.

Conclusions: This review is innovative in shifting attention from AI-SSPs as educational simulations to AI-SSPs as emerging infrastructures for clinical competency assessment. Unlike prior reviews of virtual patients or AI in medical education, this review maps not only system design and learner interaction but also which competencies are assessed, how scores are generated, and how the assessment evidence has been validated. AI-SSPs could make clinical competency assessment more scalable and potentially more standardized, although more evidence is still needed to determine whether they achieve validity and reliability across diverse learner or practitioner groups.

(JMIR Med Educ 2026;12:e92826) doi:10.2196/92826

KEYWORDS

artificial intelligence; large language models; clinical competency assessment; health care quality; medical education; AI; LLM

XSL•FO
RenderX

Introduction

Universal health coverage (UHC) cannot be realized without adequate human resources for health—a recognition that has shaped the UHC agenda since its inception and integration into the Sustainable Development Goals [1]. Achieving UHC depends not only on expanding service coverage but also on ensuring that care is delivered safely, effectively, and consistently. Despite decades of investment, with an average of more than 70% of health system expenditures spent on human resources, the global health workforce continues to face persistent shortages and wide heterogeneity in skills and practice quality, especially in resource-constrained settings [2-4]. These workforce quality gaps have direct and severe consequences, contributing to preventable deaths and compromised patient trust in health systems. Globally, more than 3 million deaths occur annually due to unsafe care, with estimates suggesting that in low- and middle-income countries, as many as 4 in 100 people die from unsafe care [5]. Tackling health workforce challenges, as well as improving the efficiency of health systems worldwide, requires valid, timely, and scalable measures of clinical competencies.

Clinical competency assessments are commonly used in 2 main contexts: medical education, where they support learning and progression across undergraduate, graduate, and continuing medical education; and health system monitoring, where they help track the skills of practicing providers. Written knowledge tests, such as multiple-choice questions, offer one scalable and standardized approach to assessment. However, because they are inherently closed-ended, they often fail to capture the complexity and interpersonal dimensions of real clinical encounters [6]. To address this, performance-based assessments exist to offer a more authentic approach that evaluates skills more comprehensively, including history-taking, counseling, clinical reasoning, and professionalism. The “gold standard” performance-based assessment is the Objective Structured Clinical Examination (OSCE), in which learners interact with actors who role-play as standardized patients (SPs) and are scored using checklists or global rating scales by an observer on the side [7]. Yet, OSCE programs are widely recognized as resource-intensive, requiring SP recruitment and training, examiner time, facilities, and coordination, all creating barriers to frequent deployment, especially at scale or in low- and middle-income countries [8,9]. Large-scale health workforce assessments, such as the World Bank’s decade-long Service Delivery Indicators surveys, further illustrate both the value and difficulty of generating representative measures of provider readiness using clinical vignettes [9]. These efforts can cost millions of dollars and require complex field operations, which are especially difficult in the reduced funding environment for global health when governments and international organizations face more limited capacity for financing.

Besides their resource intensity, another major limitation of OSCEs is that the standardization of cases does not fully eliminate subjectivity in evaluation. Even when clinical scenarios, SP scripts, checklists, and rating forms are designed to create comparable testing conditions, learner scores may still be shaped by evaluator-related bias and other sources of variation. Recent literature has emphasized that fairness in medical education assessment requires greater attention to the sociocultural processes through which judgments are produced [10]. SP clinical scenarios may encode gendered or stereotyped assumptions, while the observation and the checklist completion components may be affected by the social identities of both learners and assessors [11,12]. Evaluator effects are also well documented, including halo effects, contrast effects, and systematic differences in stringency or leniency across examiners, with examiner stringency alone accounting for up to 29% of score variation in some OSCE settings [13,14]. Thus, the need for scalable assessment is inseparable from the need for fairer and more consistent assessment systems.

Recent advances in AI create an opportunity to develop conversational, case-based simulations that mimic human SP encounters, potentially improving scalability and reducing the risk of evaluator bias. AI, including large language models (LLMs), has the potential to engage learners or practitioners in naturalistic clinical encounters at scale, responding dynamically to history questions, laboratory test requests, and more based on prespecified case scenarios. Several reviews have examined adjacent areas, but none fully address the assessment-focused scope of this review. Earlier reviews of virtual patients in health professions education established that digital patient simulations can support knowledge acquisition and skills development, but these reviews largely evaluated virtual patients as educational interventions rather than as systems for competency assessment [15,16]. More recent reviews have focused on LLM-based virtual patients, describing their growing use for medical training [17-20]. Broader reviews of AI and generative AI in medical education have also summarized applications across curriculum design, feedback, and educational support while highlighting ongoing concerns about validity, implementation, and evidence quality [21-23]. However, these reviews do not distinguish whether the AI system functions as the patient, the assessor, the feedback generator, or all three; nor do they systematically characterize what clinical competencies are assessed, how learner performance is scored, whether scoring is human-generated or AI-generated, and what evidence supports reliability, validity, or implementation readiness. This gap is important because AI–standardized simulated patStandardized Simulated Patients (SSPs) are not only educational simulations but also may become infrastructures for more scalable, routine, and fairer competency assessment. A dedicated scoping review is therefore needed to map the designs, assessed domains, and scoring and feedback approaches of AI-SSP systems for clinical competency assessment.

Given these gaps, this scoping review aimed to map the existing literature on AI-enabled SSP systems for clinical competency assessment. Specifically, we sought to characterize (1) the study characteristics, populations, and clinical contexts in which AI-SSPs have been evaluated; (2) the system architectures, model types, platforms, and interaction designs used to simulate SP encounters; (3) the clinical competencies assessed; (4) the quantitative scoring and qualitative feedback approaches used, including whether scores were generated by humans, AI, or hybrid methods; and (5) the maturity of available evidence, including feasibility, usability, scoring accuracy, learning

XSL•FO
RenderX

outcomes, and implementation evidence. By synthesizing these dimensions, this review identifies how AI-SSPs are currently being used for competency assessment and what further validation, design, and implementation research are needed.

Methods

Protocol and Registration

This scoping review was conducted according to the Joanna Briggs Institute Manual for Evidence with special attention to Chapter 10, Scoping Reviews [24,25]. The reporting of this review is aligned with the PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses Extension for Scoping Reviews) checklist [26] and the PRISMA-S (Preferred Reporting Items for Systematic Reviews and Meta-Analyses Literature Search Extension) checklist (Checklist 1) [27]. Our protocol was registered prospectively with the Open Science Framework on June 10, 2025 [28].

Eligibility Criteria

Studies were eligible for inclusion if they described an AI-based SSP system with a clinical competency assessment component. For this review, "assessment" was defined as the production of a quantitative score that could support comparison either within a learner over time or across learners. The score did not need to be AI-generated; studies were eligible if the simulation itself was AI-based and the resulting performance was scored by any method, including by AI or by human raters. Eligible systems also had to be implemented or evaluated as a functioning tool, prototype, or product. Purely conceptual frameworks, design proposals, and systematic reviews were excluded.

Inclusion criteria were as follows:

- Primary research source: Reported original empirical data.
- AI-based SSP: The interaction between the user and the SSP must be AI-based.
- Health professions context: Focused on learners or practitioners in medicine, nursing, pharmacy, or other health professions.
- Clinical competency focus: Involves clinical skills such as history-taking, clinical reasoning, diagnosis, counseling, and so on.
- Assessment component: Produced a quantitative competency score, either as part of the AI-SSP platform itself or assigned post hoc by human raters based on the AI-SSP encounter. Platforms used for formative assessments ("assessment for learning") and summative assessments ("assessment of learning") are both eligible.
- Any scoring source: The score could be generated by AI or human raters reviewing AI-simulated encounters.
- Functioning system: Described a functioning tool, prototype, platform, or product.

Exclusion criteria were as follows:

- Nonprimary source: Systematic review, editorial, commentary, opinion piece, or protocol without results.
- No AI-based SSP interaction: The SSP encounter was not AI enabled, for example, video-based or enacted by human SPs.
- No competency score: Unstructured, narrative feedback only, which does not allow for comparability across learners.
- No clinical competency focus: Did not assess clinical or health professional competencies.
- No functioning system: Described only a conceptual model, design proposal, or system architecture without evaluating a functioning tool or prototype.

Information Sources

A comprehensive literature search was conducted by searching MEDLINE (National Library of Medicine, OVID), Embase (Elsevier, embase.com), Cumulative Index to Nursing and Allied Health Literature (CINAHL Complete, EBSCOhost), Education Source (EBSCOhost), and Web of Science Core Collection (Clarivate), covering the period from their inception to an initial search date of June 12, 2025. This search was updated on May 7, 2026, to expand the search vocabulary, with the addition of Inspec (Institution of Engineering and Technology, EBSCOhost), a technology-oriented database to identify articles from the engineering literature, and a second education database, the Education Resources Information Center (ERIC, EBSCOhost), as well as the Cochrane Central Register of Controlled Trials (Cochrane Library, Wiley) to identify unpublished trials. Each database was searched individually on the specified platform.

The databases searched included meeting abstracts and proceedings, dissertations, and reports, in addition to articles published in the scholarly literature. A supplementary manual search was conducted by reviewing reference lists of relevant review articles retrieved by the database searches and by conducting targeted searches in Google Scholar using keywords. No online or print resources were purposefully browsed, and no additional studies or data were sought from authors or experts.

Search

The search strategies were designed and carried out by a health sciences librarian (CM) based on discussion with the research team and predetermined eligibility criteria. No substantial portions of strategies from previously published reviews were used. The search terms encompassed three concepts: (1) patient simulations of clinical case–based vignettes, (2) AI and LLMs, and (3) clinical competency assessment. Controlled subject vocabulary (ie, MeSH, Emtree, and subject headings from CINAHL, Education Source, ERIC, and Inspec) was included when available and appropriate. No language, search filters, or publication date limits were applied. The searches did not undergo peer review. Search reporting was guided by PRISMA-S recommendations. The exact search strategies from each database are provided in Multimedia Appendix 1.

Selection of Sources of Evidence

Records from all database searches were imported into Covidence systematic review software (Veritas Health Innovation Ltd), where duplicates were removed prior to screening. Using Covidence, 2 reviewers (BD and WZ) independently screened titles and abstracts from the search results. Conflicts were resolved through discussion. Full texts

XSL•FO
RenderX

were then retrieved and reviewed by the same 2 reviewers independently. Any discrepancies between reviewers were discussed among reviewers and resolved by discussion. A formal critical appraisal of the included studies was not conducted, as the goal was to map the extent of the literature.

Data Charting

A structured data charting template was developed in Covidence to extract information from the included studies. The template captured study characteristics, study population, SSP system features, clinical scenario design, learner interaction modalities, feedback and scoring mechanisms, backend model architecture, and product evaluation methods and results. The charting form was developed iteratively and pilot tested on an initial set of included studies by the 2 independent reviewers (BD and WZ).

Data Items

Extracted variables included general study information (study ID, title, authors, country, and relevant context on why the tool was developed); study population (target learner or practitioner group, health profession, training level, total number of respondents, and any reported demographic or clinical experience information); SSP system characteristics (tool name, frontend interface, online or offline support, and compatible devices); clinical scenario design (number and type of clinical scenarios and how they were developed or validated); scenario display features (whether the SSP was displayed using text, avatar, images, or other modalities); user interaction (respondent mode of input and available clinical actions); quantitative scoring (the type of scores used and the logic by which they were derived); qualitative feedback (the type of qualitative feedback and the logic by which they were generated); backend AI model (response generation model of SSP interaction, any prompting strategy for LLM-based systems, knowledge graphs for rule-based systems, and so on); language localization; and product evaluation (methods and results used to evaluate scoring accuracy and reliability and respondent experience and perceptions). Any unclear or missing information was clearly indicated as “not reported” with potential reasons specified.

Synthesis of Results

Extracted data were first reviewed and synthesized through discussions between the 2 reviewers to reach consensus. Thematic analysis was then conducted using the completed data charting spreadsheet. Each extracted variable was tabulated and compared across studies, with themes developed inductively from observed patterns and summarized using frequency counts and narrative synthesis where appropriate.

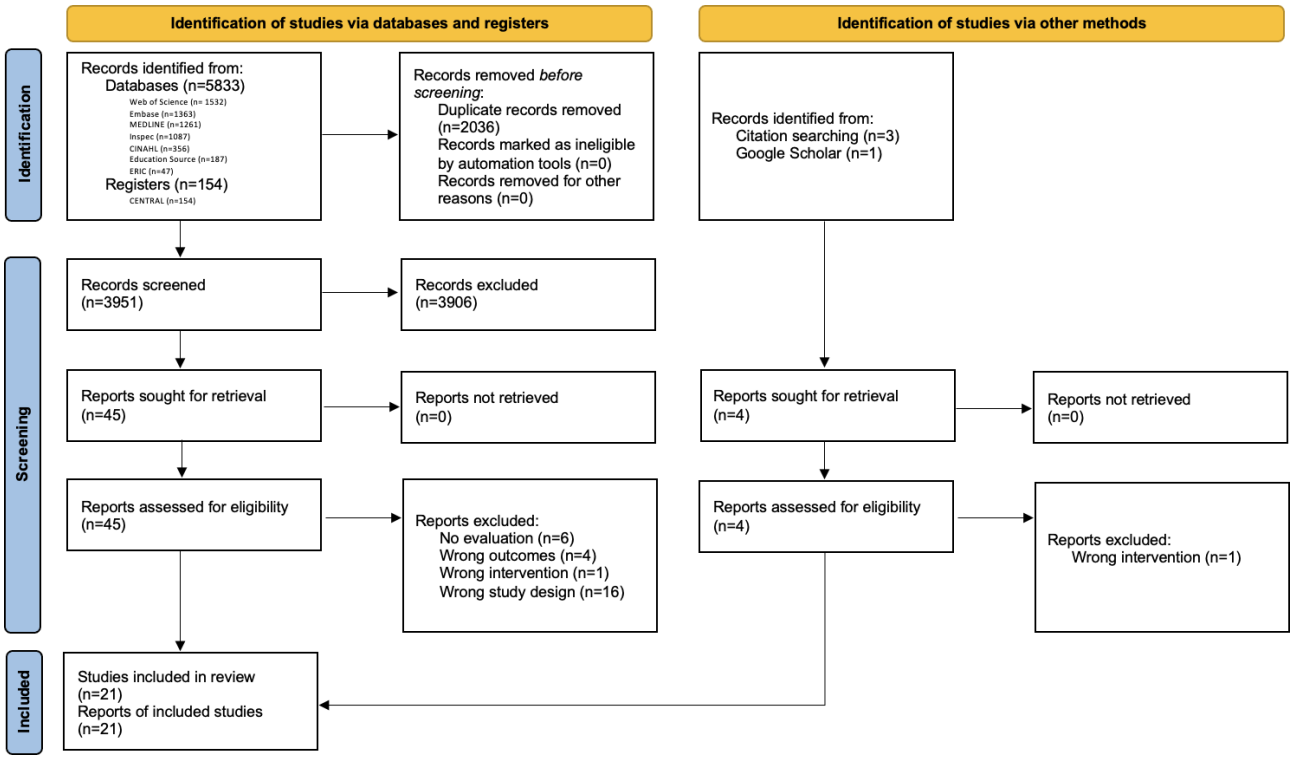
Results

Study Selection

A total of 5987 records were first identified through database searches and imported into Covidence. After removing 2036 (34%) duplicates, 3951 (66%) unique records were available for title and abstract screening. Two reviewers independently screened all titles and abstracts against the predetermined eligibility criteria, which focused on AI-based SSP systems with competency assessment components. Conflicts were resolved through discussion. At this stage, 3906 (98%) records were excluded because they did not meet the inclusion criteria. The full texts of 45 (1%) reports were then retrieved and assessed independently by the same 2 reviewers. Of these, 27 (60%) reports were excluded, with the most common reasons for exclusion including not involving an AI-based interaction with SP, no quantitative competency score produced, or not an empirical primary study. Four additional records were then identified through citation searching and Google Scholar, of which 3 (75%) studies were included. A final total of 21 studies were included in the final scoping review. The PRISMA flow diagram (Figure 1) provides further details on the screening process and reasons for exclusion.

XSL•FO
RenderX

Figure 1. PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) flow diagram of studies selected.29 The diagram summarizes the number of records identified through database and supplementary searches, records removed before screening, records screened by title and abstract, full-text reports assessed for eligibility, studies excluded with reasons, and final studies included in the scoping review of AI-enabled simulated standardized patient for competency assessment in health professions.



Study Characteristics

Table 1 summarizes the characteristics of the 21 studies included in this scoping review. The included studies were published between 2008 and 2026, with most studies published in recent years. Fifteen (71%) of the 21 studies were published in 2024 or later, reflecting the growth of AI-enabled SSP systems following the broader availability of LLMs. Earlier studies primarily evaluated rule-based or knowledge-based virtual patient systems, while all studies published after 2022 examined LLM-based SSPs.

XSL•FO
RenderX

Table . Study characteristics and populations.[a]

Study ID	Study title	Country	Study type	Study population	Sample size
Courteille et al, 2008 [30]	The use of a virtual patient case in an OSCE[b]-based exam - A pilot study	Sweden	Single-arm pilot feasibility evaluation	Fourth-year medical students in a 20-week clinical surgery course	N=110 participated in the ISP[c] station (of 118 enrolled); 68 (62%) returned the ISP questionnaire; 47 video observed (p.e70-e72)
Oliven et al, 2011 [31]	Implementation of a web-based interactive virtual patient case simulation as a training and assessment tool for medical students	Israel	Prospective randomized crossover noninferiority study	Medical students at the start of clinical year in an introductory clinical course	N=262 across 3 y: 2008, n=87; 2009, n=76; and 2010, n=99. All took both the conventional human OSCE and the VP-OSCE[d] in a randomized crossover design (p.236, Table 1).
Bond et al, 2019 [32]	Virtual standardized patient simulation: case development and pilot application to high-value care	USA	Single-arm pilot feasibility evaluation	First-year internal medicine and medicine-pediatrics residents	N=14 first-year residents; single arm (no control); 14 learners×8 cases=112 transcripts analyzed.
Maicher et al, 2019 [33]	Using virtual standardized patients to accurately assess information gathering skills in medical students	USA	Single-arm observational study with scoring validation substudy	First-year medical students early in medical interviewing curriculum	N=102 first-year medical students (single arm, observational). Scoring accuracy substudy: 20 randomly selected dialogues rated by 3 human raters+1 computer rater (ChatScript) (p.1054; p.1055).
Yang et al, 2019 [34]	MCRDR[e] knowledge-based 3D dialogue simulation in clinical training and assessment	Australia	Proof-of-concept technical evaluation with user survey	Medical students and medical tutors	N=25 total: 20 medical students+5 medical tutors (students split into baseline n=10/explore n=10 for the interactive-efficiency test) (p.17).
Lee et al, 2022 [35]	Development of a chatbot to train physiotherapy students in clinical questioning and reasoning	Singapore	Development or proof-of-concept study	Year 1 physiotherapy students before clinical placement	Not reported—development or proof-of-concept paper; experimental evaluation explicitly stated as “ongoing,” no participant N given (p.477 abstract; p.480 sec.VI).
Brugge et al, 2024 [36]	Large language models improve clinical decision making of medical students through patient simulation and structured feedback: a randomized controlled trial	Germany	Randomized controlled trial	Medical students without clinical rotations; mostly early-semester learners	N=21 final (feedback: n=10; and control: n=11). 29 agreed; 5 dropped pre-randomization; 3 excluded post-randomization (technical data-save failures) (p.5).
Holderried et al, 2024 - e53961 [37]	A generative pretrained transformer (GPT)-powered chatbot as a simulated patient to practice history-taking: prospective, mixed methods study	Germany	Prospective mixed-methods single-arm feasibility study	Medical students and 2 midwifery students at a skills-refreshing event	N=28; single arm, no control. All 28 completed chatbot sessions and the Chatbot Usability Questionnaire (p.7).

XSL•FO
RenderX

Study ID	Study title	Country	Study type	Study population	Sample size
Holderried et al, 2024 - e59213 [38]	A language model-powered simulated patient with automated feedback for history-taking: prospective study	Germany	Prospective validation study of automated feedback	Medical students, mostly third-year, during skills training course break	N=106 participated and completed≥1 conversation. 106 conversations, 1894 analyzable Q-A pairs after excluding 26 (p.5 - 6).
Li et al, 2024 [39]	Leveraging large language model as simulated patients for clinical education	China	Technical system evaluation or human-LLM[f] scoring validation	System or framework evaluators: nonmedical evaluators, clinical medicine students, senior physicians; no learner trial	SP quality eval: number of expert evaluators not reported (engaged 6 LLMs×8 cases, up to 20 rounds). Assessment validation: 80 dialogue records or 8 cases scored by 1 human expert. LLM-as-VD[g]: 9 human evaluators (3+3+3)×5 repeats×8 cases (p.3 sec.2.2; p.5 - 6 sec.2.3 - 2.4).
Chiu et al, 2025 [40]	Exploration of the role of ChatGPT in teaching communication skills for medical students: a pilot study	USA (inferred from author affiliations)	Single-arm pilot pre-post study	Second-year medical students	N=12; single-arm pre-post design; no separate arms (p.1872 Activity; p.1875 Multimedia Appendix 1)
Gorcilov et al, 2025 [41]	Large language models deliver interactive learning cases accurately and with rapid appropriate feedback	Australia (primary author); USA (co-authors)	Conference abstract; small pilot validation	Two junior physicians and 1 medical student acting as case users or investigators	N=3 learners (2 junior physicians+1 medical student); 5 cardiology cases; 250 LLM responses total. Pilot or proof-of-concept (p.S497-S498 Results).
Haut et al, 2025—AI Standardized Patient arXiv [42]	AI standardized patient improves human conversations in advanced cancer care	USA	Randomized controlled trial	Health professional students and practitioners: medical, PA[h], nursing, residents, clinicians	N=51 (SOPHIE[m]: n=26; and control: n=25). Debrief survey completers 22/26 SOPHIE, 21/25 control; UI/UX[i] survey 23/26 SOPHIE; poststudy interview 46/51; conversation reviews rated 506/510 (p.4 sec.2.1, Table 1).
Haut 2025—USIHC dissertation [43]	USIHC: understand, simulate, and improve human compassion in serious illness communication	USA	Dissertation; system development and randomized controlled trial	Medical, nursing, and PA trainees plus practicing clinicians	N=51 (SOPHIE intervention: n=26; control: n=25). 102 SP[k] conversations recorded; 506/510 conversation ratings collected (p.72 sec.5.2.2; p.80 sec.5.4.1).
Liu et al, 2025 [44]	Development and validation of a large language model-based system for medical history-taking training: prospective multicase study on evaluation stability, human-AI consistency, and transparency	China	Prospective feasibility study	Third-year medical students with theoretical history-taking instruction but limited SP experience	N=31, single arm (no randomization or learner comparison groups; all used the fully processed AMTES[l] system).

XSL•FO
RenderX

Study ID	Study title	Country	Study type	Study population	Sample size
Qian et al, 2025 [45]	Use of large language models for rapid quantitative feedback in case-based learning: a pilot study	USA; Australia (multisite)	Pilot expert-comparison validation study	Medical student investigators using neurological cases	N=4 student investigators (p.1169).
Zheng et al, 2025 [46]	LLM-as-a-fuzzy-judge: fine-tuning large language models as a clinical evaluation judge with fuzzy logic	USA	Technical model validation or computational evaluation	Medical students	2303 conversation logs ->~2303 individual student utterances as data points; annotated by 7 expert judges. Split: 1611 train/231 val/461 test (p.5 sec.3.2; p.7 sec.3.5).
Zhou et al, 2025 [47]	Use of artificial intelligence to identify student behaviors associated with improved diagnostic accuracy of pulmonary diseases	USA (inferred from author affiliations)	Conference abstract; exploratory observational transcript analysis	Second-year medical students in pulmonary curriculum	602 unique student encounter transcripts; student headcount not reported—only encounter count n=602 (p.1).
Chen, 2026 [48]	AI-enhanced virtual reality simulation for nursing students' empathy: automated scoring and interrater reliability in a randomized controlled study	Taiwan	Randomized controlled mixed methods study	Nursing students in obstetric nursing education	80 nursing students (experimental group: n=40; control group: n=40) (p.1 Abstract; p.2 sec.2.3; p.4 sec.3.1)
Hicke et al, 2026 [49]	MedSimAI: simulation and formative feedback generation to enhance deliberate practice in medical education	USA	Multisite pilot or quasi-experimental mixed methods evaluation	Medical students across 3 US institutions	410 unique learners; 1024 completed encounters. Inst A: 275 students/620 cases; Inst B: 104 students/362 cases; Inst C: 31 students/42 cases. Surveys: Inst B baseline n=85, exit n=19. Validation corpus: 104 OSCE transcripts (Inst C) rescored (p.6 - 7 Table 1, sec.5.1).
Reynolds et al, 2026 [50]	AI versus human consensus: correlation of rubric-based scoring in pharmacy counseling	USA	Conference abstract; AI-human scoring correlation validation	Pharmacy students in a pharmaceutics class	Not reported (number of students or transcripts not stated in the conference abstract).

[a]Studies varied widely in design, ranging from single-arm pilot and feasibility evaluations to randomized controlled trials, validation studies, and technical model evaluations. Sample sizes also varied substantially, from small expert- or student-based pilots to larger multisite or multiyear deployments.

[b]OSCE: objective structured clinical examination.

[c]ISP: interactive simulated patient.

[d]VP-OSCE: Virtual Patient–Objective Structured Clinical Examination.

[e]MCRDR: Multiple Classification Ripple Down Rules.

[f]LLM: large language model.

[g]VD: virtual doctor.

[h]PA: physician assistant.

[i]UI/UX: user interface/user experience.

[j]SOPHIE: Standardized Online Patient for Healthcare Interaction Education.

[k]SP: standardized patient.

[l]AMTES: Artificial Intelligence-Powered Medical History-Taking Training and Evaluation System.

The included studies were conducted across a range of geographic settings, with the largest number from the United States or involving US-based study sites or author affiliations [32,33,40-43,45-47,49,50]. Other countries represented included Germany [36-38], China [39,44], Sweden [30], Israel [31], Australia [34,41,45], Singapore [35], and Taiwan [48]. Overall,

XSL•FO
RenderX

the evidence base was concentrated in high-income countries, suggesting that the generalizability of findings across diverse medical educational contexts remains uncertain.

Study designs varied substantially, but most reflected early-stage development, feasibility testing, or validation rather than large-scale implementation research. Most studies conducted single-arm pilot feasibility testing [30,32,33,37,38,40,41,44,45,47,50]; others were framed as proof-of-concept, technical, or computational evaluations [34,39,46], while a limited number of studies used randomized controlled trials, crossover, or quasi-experimental designs [31,36,42,43,48,49]. Only 2 (10%) studies provided evidence beyond initial development or short-term evaluation [31,49]. Overall, the evidence base remained concentrated in pilot, feasibility, proof-of-concept, and validation work, with fewer controlled evaluations and large-scale implementation studies.

Study Populations

Medical students across all academic years were the most frequently cited population, including students at the start of clinical training [31], first-year students in medical interviewing curricula [33], early-semester students without clinical rotation experience [36], second-year students [40], third-year students in history-taking training but with limited SP experience [44], and fourth-year medical students in surgery [30]. Several studies extended beyond medical students to include residents, junior physicians, practicing clinicians, or mixed groups of health professional trainees and practitioners [32,41-43]. Fewer studies examined other health professions learners, including physiotherapy students [35], midwifery students [37,38], nursing students [48], pharmacy students [50], and physician assistant trainees [42,43]. One technical evaluation study did not involve a conventional learner trial and instead evaluated AI-SSP quality and scoring using a combination of nonmedical evaluators and senior physicians. Overall, the included studies primarily centered on medical education rather than clinicians in practice, with more limited but emerging representation from nonmedical health professions.

Sample sizes and units of analysis varied widely across studies. Among studies reporting learner-level participation, sample sizes ranged from very small pilot studies with 3 to 4 participants [41,45] to larger evaluations involving more than 100 learners [30,31,38,49], and in one case, 410 unique learners across 3 institutions [49]. Several studies reported encounter-, transcript-, utterance-, or dialogue-level data rather than only learner-level sample sizes, including 112 transcripts [32], 80 dialogue records [39], 602 student encounter transcripts [47], 1024 completed encounters [49], and more than 2300 conversation logs or students' utterances [46]. Some studies did not clearly report learner sample size or were development focused without a completed learner evaluation [35,50].

AI-SSP System Architecture and Interaction Design

Systems Overview and Intended Use

As summarized in Table 2, the included studies evaluated a diverse set of AI-enabled SSP systems, ranging from dialogue simulation platforms to immersive virtual reality environments. Several studies involved formally named platforms, such as USC Standard Patient [32], CureFun [39], SOPHIE [42,43], AMTES [44], Turing Interactive Case Learning [45], 2-Sigma [46,47], VR-AI-ECS [48], and MedSimAI [49], while others described the intervention more generally as a "virtual patient," "simulated patient chatbot," or "AI-supported online platform" [31,33,35-38,40,41,50]. Almost all studies justified their systems by pointing to the same underlying problem: traditional SPs are resource intensive, costly, logistically difficult, and hard to scale. The systems varied in format and implementation, but most were designed to create repeatable clinical encounters in which learners could practice, receive feedback, or be assessed in a standardized environment. Overall, 20 unique AI-SSP systems were identified across 19 unique study teams, with 1 system evaluated by 2 studies [42,43] and 1 study team evaluating 2 variations of a system [37,38].

XSL•FO
RenderX

Table . AI-enabled simulated standardized patient (AI-SSP) architecture and interaction design.[a]

Study ID	AI-SSP system name	Primary purpose	Access and device compatibility	System architecture	Clinical scenarios	Scenario development	SSP display features	Respondent mode of input	Available clinical actions
Courteille et al, 2008 [30]	No proprietary name reported, referred to as "ISP (Interactive Simulation of Patients)"; developed at VP-Lab, Karolinska Institutet (p.e67, p.e76)	Assessment	Offline; computer	Rule-based or NLP	1 colorectal cancer surgical case	Adapted from existing ISP case; redesigned or validated by surgeons and senior clinicians	Avatar; images	Typed free text+structured actions	History-taking; visual examination; laboratory tests; clinical examination; and other
Oliven et al, 2011 [31]	No proprietary name reported, referred to as "Web-Based Virtual Patient (VP) System/VP-OSCE" (p.233)	Education + assessment	Online; computer	Rule-based or NLP	25 complaint prototypes with multiple cases	Faculty-authored prototype cases; keyword lexicon expanded through use	Text or web case interface	Typed free text+structured actions	History-taking, visual examination, laboratory tests, and clinical examination
Bond et al, 2019 [32]	USC Standard Patient (virtual standardized patient platform; open-source, standardpatient.org; USC Institute for Creative Technologies, 2014 - 2017) (p.242 Platform).	Education + assessment	Online; computer	Rule-based or NLP	8 high-value care cases across 4 complaints	Faculty or resident-authored templates; iterative NLP mapping refinement	Avatar; text	Typed free text	History-taking, visual examination, chart review, laboratory tests, clinical examination; other: differential diagnosis and treatment
Maicher et al, 2019 [33]	No proprietary name reported, referred to as "Virtual Standardized Patient (VSP) system" (p.1053 Abstract; p.1054).	Education	Online or web; and computer	Rule-based or NLP (ChatScript)	15 VSP cases; study analyzed 1 low-back-pain case	Faculty-defined history elements using Kalamazoo checklist	3D avatar	Typed free text	History-taking

XSL•FO
RenderX

Study ID	AI-SSP system name	Primary purpose	Access and device compatibility	System architecture	Clinical scenarios	Scenario development	SSP display features	Respondent mode of input	Available clinical actions
Yang et al, 2019 [34]	No proprietary name reported, referred to as "MCRDR Knowledge-Based 3D Dialogue Simulation System ("3D dialogue simulation system")" (p.1 - 2).	Education + assessment	Offline; computer	Rule or knowledge-based	GI[b] or respiratory cases; ≥14 conditions	Tutor-authored from clinical samples or databases; MCRDR[c] rules	3D avatar; text	Typed free text	History-taking, visual examination, laboratory tests, and clinical examination
Lee et al, 2022 [35]	No proprietary name reported, referred to as "Chatbot Virtual Patient (VP)/speech-enabled AI chatbot" (p.477 - 478)	Education	Online; computer or speech-enabled web platform	Rule-based or NLP (DialogFlow)	1 musculoskeletal elbow-pain case	Student-team script supervised by faculty; iteratively refined	Text; speech-enabled chatbot	Speech	History-taking
Brugge et al, 2024 [36]	No proprietary name reported, referred to as "chatbot"	Education	Online; computer	LLM (GPT-3.5)	4 neurological or neurosurgical emergency cases	Contextual prompts; limited case details in prompt	Text	Typed free text	History-taking; other: differential diagnosis
Holderried et al, 2024 [37]	No proprietary name reported, referred to as "chatbot"	Education	Online; computer	LLM (GPT-3.5)	1 T2DM[d]-like history-taking case	Expert-authored illness script loaded into prompt	Text	Typed free text	History-taking
Holderried et al, 2024 [38]	No proprietary name reported, referred to as "chatbot"	Education	Online; computer	LLM (GPT-4)	1 reused history-taking illness script	Expert-authored illness script; feedback categories derived from script	Text	Typed free text	History-taking
Li et al 2024 [39]	CureFun	Education + assessment	Online; computer	LLM+RAG or knowledge graph (multiple LLMs)	8 Chinese SP[e] cases across GI, endocrine, and pulmonary conditions	Scripts from an industry dataset cleaned and converted into case graphs	Text	Typed free text	History-taking, laboratory tests, and clinical examination
Chiu et al, 2025 [40]	No custom-built application, used ChatGPT (version 3.5) via the public web interface (p.1871 Abstract; p.1872 Activity)	Education	Online; computer	LLM (ChatGPT-3.5)	1 breaking-bad-news diabetes scenario	Research team scenario; student-authored prompt; SPIKES framework	Text	Typed free text	Other: breaking bad news

XSL•FO
RenderX

Study ID	AI-SSP system name	Primary purpose	Access and device compatibility	System architecture	Clinical scenarios	Scenario development	SSP display features	Respondent mode of input	Available clinical actions
Gorcilov et al, 2025 [41]	No proprietary name reported, referred to as "an online platform using OpenAI's GPT-4o" (p.S497 Method).	Education	Online; computer	LLM (GPT-4o)	5 cardiology cases	Investigator-developed cases; details limited	Text	Typed free text	History-taking; laboratory tests; and clinical examination
Haut et al, 2025 [42]	SOPHIE (Standardized Online Patient for Healthcare Interaction Education)	Education	Online; computer	Hybrid LLM+classifier (GPT-3.5-turbo)	3 advanced-cancer SIC[f] cases	Palliative or oncology expert-authored; adapted from validated SIP[g]/VOICE[h] cases	Avatar; voice	Speech	Other: SIC, guide through different treatment decisions
Haut, 2025 [43]	SOPHIE (Standardized Online Patient for Healthcare Interaction Education)	Education	Online; computer	Hybrid LLM+classifier (GPT-3.5-turbo)	Stage IV cancer communication modules+3 RCT[i] cases	URMC[j] faculty and validated SIC frameworks; actors or SP training	Avatar; voice	Speech	Other: SIC
Liu et al, 2025 [44]	AMTES (Artificial Intelligence-Powered Medical History-Taking Training and Evaluation System)	Education + assessment	Online; electronic device	LLM (DeepSeek-V2.5; Qwen-Max tested)	3 history-taking cases: respiratory, urinary, and GI	Senior clinical experts; aligned to licensing syllabus; and iteratively refined	Text	Typed free text	History-taking
Qian et al, 2025 [45]	Turing interactive case learning (platform name in Figure 2, p.1170); powered by GPT-4o.	Assessment	Online; computer	LLM (GPT-4o)	5 neurology cases	Faculty or expert-scored cases; development details limited	Text	Typed free text	History-taking; laboratory tests, clinical examination, and other: preliminary diagnosis
Zheng et al, 2025 [46]	LLM-as-a-Fuzzy-Judge (underlying clinical simulation platform: "2-Sigma") (p.1; p.4 sec.3.1)	Assessment	Online platform; device not specified	LLM or fine-tuned judge (multiple backbones)	Wide range of 2-Sigma clinical dialogue scenarios	2-Sigma platform scenarios; and details limited	Text	Typed free text	History-taking; laboratory tests; clinical examination; and other
Zhou et al, 2025 [47]	2-Sigma platform (AI-based; University of Cincinnati College of Medicine) (p.1)	Assessment	Online platform; device not specified	Generative AI or multiagent (model not reported)	3 pulmonary cases: OSA[k], COPD[l], and PE[m]	Predefined 2nd-year curriculum framework; and details limited	Text	Typed free text	History-taking; and clinical examination

XSL•FO
RenderX

Study ID	AI-SSP system name	Primary purpose	Access and device compatibility	System architecture	Clinical scenarios	Scenario development	SSP display features	Respondent mode of input	Available clinical actions
Chen, 2026 [48]	Virtual Reality Integrated Artificial Intelligence Empathic Communication Simulation (VR-AI-ECS)	Education	VR[n]-based; device not fully specified	AI dialogue engine (model not reported)	Obstetric empathy scenarios across childbirth care	Expert-informed prompts; grounded in empathy and experiential-learning models	VR; voice; and role-switching	Speech+embodied VR interaction	Other: users alternate between the roles of a pregnant woman and a nurse to experience both perspectives of maternity care
Hicke et al, 2026 [49]	MedSimAI	Education + assessment	Online; computer or voice supported	LLM (GPT-4o; OpenAI Realtime API)	Instructor-configurable diverse cases	Multi-institution SME[o] co-design; structured templates or checklists	Text; and voice option	Typed or speech input	History-taking
Reynolds et al, 2026 [50]	No proprietary name, referred to as "investigational board-approved automated AI grading agent" (p.S1809).	Assessment	Text-based chatbot; and device not specified	Specific AI or LLM; model not reported	Pharmacy counseling cases	Preauthored cases; development details not reported	Text	Typed free text	Other: drug counseling

[a]Across studies, systems varied from early rule-based or traditional natural language processing (NLP) virtual patient platforms to more recent large language model (LLM)–based, retrieval-augmented generation (RAG)–supported, hybrid LLM classifier, and virtual reality systems.

[b]GI: gastrointestinal.

[c]MCRDR: Multiple Classification Ripple Down Rules.

[d]T2DM: type 2 diabetes mellitus.

[e]SP: standardized patient.

[f]SIC: serious illness communication.

[g]SIP: Study of Individualized Patient Care.

[h]VOICE: Values and Options in Cancer Care.

[i]RCT: randomized controlled trial.

[j]URMC: University of Rochester Medical Center.

[k]OSA: obstructive sleep apnea.

[l]COPD: chronic obstructive pulmonary disease.

[m]PE: pulmonary embolism.

[n]VR: virtual reality.

[o]SME: subject matter expert.

Across these 20 unique AI-SSP systems, their primary intended use fell into 3 broad categories. The largest group, 9 (45%) of 20 systems, framed AI-SSPs primarily as deliberate practice or educational tools, emphasizing safe, repeatable, and low-stress environments for learners to practice history-taking, communication, empathy, counseling, or case-based clinical reasoning [33,35-38,40-43,48]. While they do not list assessment as a primary purpose, they were included in this review because they all produce some form of quantitative score, and their functionality demonstrates strong potential for direct adaptation to assessment purposes. Six (30%) of 20 systems were positioned as hybrid training and assessment tools, combining SP interaction with structured feedback and scoring [31,32,34,39,44,49]. The remaining 5 systems focused explicitly on assessment using AI-SSP systems to evaluate user performance. Taken together, the literature suggests that AI-SSPs, even the ones with scoring or assessment components, have most often been introduced as formative learning environments. However, there exists an increasing subset of systems that explicitly emphasize assessment purposes, including automated scoring, structured feedback, and performance analytics.

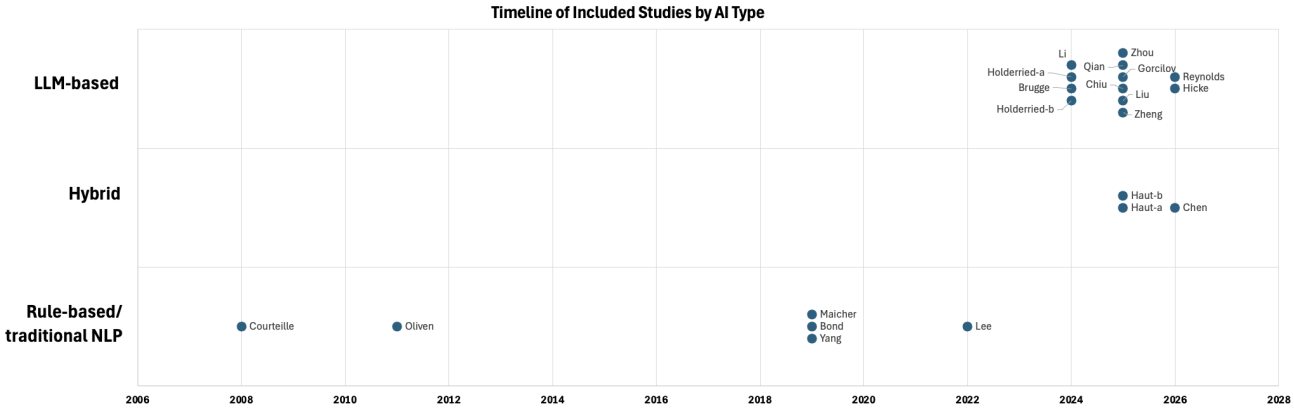
Backend Architecture and AI Model

Backend architectures varied substantially across studies and showed a clear temporal shift in model type, as summarized in Figure 2. Earlier AI-SSP systems, particularly those published before 2022, relied on rule-based natural language processing,

XSL•FO
RenderX

scripted dialogue structures, decision trees, keyword matching, or intent classification approaches. These systems used preauthored cases and expected learner inputs to be matched against predefined system response pathways, making the interaction more predictable but less conversationally fluid. For example, Maicher et al [33] used approximately 2500 ChatScript rules to support the virtual patient dialogue, Yang et al [34] developed a knowledge-based 3D dialogue system using classification rules, and Lee et al [35] used DialogFlow intents to classify learner questions and return scripted patient responses. These earlier architectures emphasized control, standardization, and reproducibility but were constrained by the range of prespecified utterances, intents, and response rules built into the system.

Figure 2. Timeline of included studies by AI model type. This figure maps the included studies by publication year and underlying AI model type, showing the field's transition from earlier rule-based or traditional natural language processing (NLP) systems to more recent large language model (LLM)–based and hybrid AI approaches.



After 2022, all included studies either partially or fully relied on LLMs to power patient dialogue, scoring, or feedback generation. In contrast to rule-based systems, LLMs are AI models trained on large text datasets to interpret and generate natural language, which allows them to support more open-ended, contextually responsive SSP dialogue. Specifically, 2 (10%) studies used the public ChatGPT interface directly, representing a relatively low-barrier implementation model in which learners interacted with an existing chatbot rather than a fully custom AI-SSP platform [36,40]. In a study by Brugge et al, contextual prompts were entered into ChatGPT right before participants arrived at the computer, while Chiu et al instructed students to write their own prompt specifying that ChatGPT should act as a patient before they began the dialogue.

The majority of the newer studies, however, embedded LLMs within more controlled custom platforms or structured prompt pipelines rather than using an open-ended public chatbot interface. These systems typically combined a system prompt with case information, behavioral instructions, and specific constraints to define how the simulated patient should respond. For example, Holderried et al [37,38] used a 2-part prompt with behavioral instructions and illness scripts, Haut/SOPHIE [42,43] used serious illness communication scenarios with target skills such as agenda setting and responding to emotion, and Hicke et al/MedSimAi [49] used instructor-provided templates to build system prompts and separate evaluation prompts with JSON output. Several other studies also separated the simulated patient prompt from the scoring or feedback prompt, indicating that the same LLM was not simply prompted once to "be a patient" but was instead used through multiple task-specific prompt layers [37,38,40,42-44,49].

A smaller subset of studies incorporated additional backend structures to constrain LLM output, improve consistency, and reduce hallucinations. These included retrieval-augmented generation, structured case graphs, templates, or guardrails that limited the information available to the model and specified how the virtual patient should behave. For example, Li et al [39] used a structured case graph so that the model could retrieve relevant patient facts before generating a response, Haut/SOPHIE [42,43] aligned dialogue with target communication frameworks, and Hicke et al [49] used templates specifying patient demographics, history, communication style, emotional state, and vocabulary. These approaches suggest that newer AI-SSP systems are mostly "LLM wrapped" rather than fully open-ended. The model generates naturalistic conversation, but the architecture has constraints in place for what information it can use, how the patient should respond, and what learning or assessment goals the interaction is intended to support.

Clinical Case Scenarios and Their Development

The number of simulated cases ranged from 1 to 25 across studies. Most used a relatively small set of cases rather than a large clinical case bank [30,33,35,38,40]. The clinical content spanned multiple domains, including gastrointestinal and metabolic conditions, cardiopulmonary and respiratory presentations, neurology and neurosurgery, musculoskeletal complaints, oncology and cancer communication, obstetric and nursing empathy scenarios, and pharmacy counseling.

Earlier systems often used cases to assess biomedical clinical reasoning and diagnostics [30-32,34,36,45,47]. Later studies increasingly expanded case content to include communication-oriented and interpersonal competencies, including empathy, breaking bad news, counseling, professionalism, serious illness communication, and patient-centered communication behaviors [40,43,46,48-50]. Across studies, cases typically included patient demographics, chief complaint, symptom history, and expected learner tasks, with some systems also incorporating physical examination

XSL•FO
RenderX

findings, investigation results, differential diagnosis, treatment decisions, emotional context, and grading rubrics.

Case development was mostly expert-led or faculty-led rather than AI-generated. Fourteen (67%) studies reported cases that were authored, adapted, or reviewed by clinical experts, faculty, residents, or other domain specialists [30-33,35,37,38,40-45,49]. Human expertise informed scenario writing, checklist or rubric design, natural language processing rules, and other chatbot response logic. Only 1 (5%) study explicitly described cases as being developed by an LLM using contextual prompts [36]. Overall, case development remained largely a human-authored process, with AI more commonly used as an interaction and/or scoring mechanism rather than as a generator of clinical case content.

Platform, Access, and Frontend Interactions

Frontend design ranged from simple text-based chat interfaces to avatar-based immersive virtual environments. Fourteen (67%) of 21 studies used primarily text- or chat-based interfaces, making this the most common display format overall [31,35-41,44-47,49,50]. These systems presented the simulated patient through written dialogue and allowed learners to type questions in a conversational format. Text-based interfaces were especially common among recent LLM-enabled systems, suggesting that newer AI-SSPs prioritized conversational fluidity over visual embodiment.

Seven (33%) studies incorporated avatars, patient images, or other visual representations to add realism to the simulated encounter [30,32-34,42,43,48]. These systems were common before the wide adoption of LLMs and often included patient portraits, animated avatars, or graphical interfaces for selecting clinical actions beyond conversations. One (5%) study in particular used an even more immersive avatar-based virtual reality (VR) environment, in which learners wore vests and a VR headset to interact with the AI-SSP in a 3D simulated clinical setting [48]. Overall, earlier systems tended to be more visually elaborate but less conversationally flexible, whereas newer LLM-enabled systems were visually simpler but allowed more open-ended dialogue.

Most systems were delivered through online or computer-based platforms, although access details were inconsistently reported. Three systems were hosted offline, including earlier virtual patient systems [30,34,48]. Two studies used smartphone or tablet-based devices [33,44]. One study used a VR headset to support embodied interaction in an immersive environment [48]. The included studies suggest that AI-SSP frontend design has evolved from visually rich, platform-specific simulation environments toward more lightweight, chat-based interfaces that can be deployed through web, mobile, or existing LLM platforms.

Clinical Competency Assessment

Clinical Skills Assessed

Table 3 summarizes the clinical competencies each included study assessed. Across the 21 studies, the most commonly assessed domain was history-taking or clinical inquiry, which was assessed in 11 (52%) studies. Diagnostic accuracy was assessed in 8 (38%) studies, and communication, empathy, professionalism, or counseling was also assessed in 8 (38%) studies. Only 4 (19%) studies assessed clinical examination or laboratory test selection, and only 2 (10%) evaluated disease management and treatment planning skills. Temporally, the focus of assessment also shifted from more structured clinical reasoning tasks, such as diagnosis and information gathering, toward more interactional competencies, including communication, empathy, counseling, and professionalism, largely mirroring the broader transition from rule-based virtual patient systems to more fluid, conversational LLM-enabled SSPs.

XSL•FO
RenderX

Table . AI-enabled simulated standardized patient competency assessment approaches.[a]

Study ID	Competency domains assessed	Quantitative scoring approach	Score derivation or metric	Score accuracy or reliability evaluation	Qualitative feedback
Courteille et al, 2008 [30]	History-taking; examination or laboratory selection; diagnosis	Checklist coverage+binary diagnosis	Percentage of required history questions or laboratories ordered; correct diagnosis; and completion time versus expert checklist.	Not formally assessed for automated scoring accuracy.	No automatic in-system feedback; and paper case summary provided post-session.
Oliven et al, 2011 [31]	Clinical skills or OSCE[b]; history, examination, and laboratories or imaging	Checklist coverage	Percentage of mandatory items identified or asked; item difficulty; discrimination; and Cronbach α computed.	Internal consistency: VP-OSCE[c] α=0.82 - 0.89 versus human OSCE α=0.65 - 0.74; Pearson *r*=0.68 - 0.71 with human OSCE.	Practice-mode gap feedback: asked versus missed or incorrect mandatory items; suppressed in examination mode.
Bond et al, 2019 [32]	High-value care; history, examination, diagnosis, testing, treatment	Weighted checklist+domain scores	NLP[d] maps questions to prelabeled content; must ask=3 points, nice to ask=1 point; diagnosis, treatment, or testing scores.	Faculty versus platform rankings Spearman ρ mean=0.80; human-human ICC[e] approximately 0.84 - 0.85; direct human-machine IRR[f] not computable.	Real-time pie chart plus end-of-case credit labels, content areas, transcript, and avatar-response appropriateness.
Maicher et al, 2019 [33]	Information gathering or history-taking	Checklist coverage	Essential history elements covered/21×100; subscale percentage for HOPI[g], PMH[h], FH[i], and SH[j].	ChatScript versus 3 faculty: no overall score difference (*P*=.07); all-rater agreement 82.6%; ChatScript categorization approximately 87%.	Automated session report: covered versus missed essential elements by history category plus expert model approach.
Yang et al, 2019 [34]	Medical knowledge; procedural communication; and investigation or diagnosis	Multicomponent competency scores	Right or wrong quiz; missed mandatory inquiry topics; and investigations or diagnoses compared with case possibilities.	Expert ground-truth overlap 75% - 88.9%; interactive accuracy 83.4% - 88.6% versus baseline 70.6% - 77.8%; CNN[k] accuracy 0.85.	On-screen unasked-topic overlay; final student versus correct or possible investigation and diagnosis sets with detail view.
Lee et al, 2022 [35]	Physiotherapy clinical questioning or reasoning; empathy	Penalty or bonus score	Score out of 100; penalties for missed questions, red or yellow flags, consent, difficult questions; and empathy bonus.	No human agreement validation reported; intent mismatches manually corrected during testing.	Automated report after each attempt: missed or scored inputs with topic labels and example phrasings.
Brugge et al, 2024 [36]	Clinical decision-making or history-taking	Rubric or Likert scale	ChatGPT scores CRI-HTI[l]: 8 items across 3 domains, each 1 - 5.	Human-human ICC=0.924 for transcript ratings; no direct AI-human agreement reported.	Feedback arm received 2-sentence justification per CRI-HTI item plus 3 personalized improvement suggestions.
Holderried et al, 2024 [37]	History-taking	Checklist coverage (completed by humans)	No in-tool quantitative competency score; manual post hoc QAP[m] or script coverage analysis and CUQ[n] usability score.	Manual analysis only; no kappa or ICC reported.	Unclear r or limited: UI[o] showed feedback button, but feedback content or logic not described.
Holderried et al, 2024 [38]	History-taking completeness	Checklist or category coverage	GPT-4 returns yes or no JSON for 45 illness script categories; parsed to compute coverage %.	GPT-4 versus human rater Cohen κ=0.832 overall; several low-prevalence or overlapping categories <0.6.	Immediate GPT-4 structured narrative report: what was done well and missed items by history category.

XSL•FO
RenderX

Study ID	Competency domains assessed	Quantitative scoring approach	Score derivation or metric	Score accuracy or reliability evaluation	Qualitative feedback
Li et al, 2024 [39]	Clinical inquiry	Weighted checklist coverage	Aspect items (0.3)+specific information items (0.7); Llama-13B ensemble voting; normalized 0 - 1.	Versus expert grading: Spearman r=0.655 - 0.954; Pearson r=0.765 - 0.943 across cases.	LLM[p] writer generates examination report with explanatory advice from score results and conversation history.
Chiu et al, 2025 [40]	Breaking bad news; communication	Rubric score	ChatGPT scores SPIKES categories: 6 domains×5 points=30; with and without rubric.	ChatGPT overscored versus faculty (means 27.92/24.08 versus 16.79); significant differences; prompt or rubric dependent.	ChatGPT narrative feedback by SPIKES category: strengths, areas for improvement, and overall summary.
Gorcilov et al, 2025 [41]	Case evaluation and management plan	Numerical LLM score	LLM scored evaluation and plan per candidate; and rubric or maximum not reported.	Versus experts: Pearson r=0.88/0.76 for history, examination, or investigations; 0.69 and 0.71 for assessment and plan.	LLM-generated narrative feedback; and prompting and feedback logic not fully reported.
Haut et al, 2025 [42]	Serious illness communication; 3E[q] skills	Classifier or process metrics+human outcome score	Real-time 3E skill labels and process metrics; RCT[r] outcome separately human rated on 0 - 1 scale.	Skill classifier accuracy=72%, F_1=51%; and human outcome-rubric ICC=0.882.	Post-module GPT suggestions using transcript, 3E annotations, ACT[s] context, and examples; and detailed PDF with excerpts and highlights.
Haut, 2025 [43]	Serious illness communication; 3E skills	Classifier or process metrics+human-rated rubric	Hybrid rule+ BERT[t] 3E tags drive module progression; 18-item human-rated QSUM[u] outcome normalized 0 - 1.	BERT classifier accuracy=72%, F_1=51%; human-human ICC=0.882 across 4 reviewers.	Feedback page: what done well, missed opportunities, actionable suggestions, color-coded transcript, and PDF report.
Liu et al, 2025 [44]	History-taking across multicase scenarios	Item-level checklist score	66/59/67 case items worth 0.5 - 2 points; total 70; LLM scoring with decomposition and verification.	AI-human ICC=0.978/0.923/0.972 by case; item consistency 95.75% - 97.13%; cross-model consistency high.	Structured transparent report: dialogue, total score, category completeness, scored or missed items, citations, and rationale.
Qian et al, 2025 [45]	Case-based clinical reasoning	Key-point coverage	LLM given case scoring criteria; score=percentage of key points elicited; max 5 - 8 key points per case.	Versus 4 experts: ICC=0.78 - 0.82; Pearson r=0.67 - 0.90; LLM initially about 12 pp higher, improved after 10 pp calibration.	No automatic qualitative feedback to learners reported.
Zheng et al, 2025 [46]	Clinical communication	Fuzzy categorical labels	Per-utterance labels for professionalism, medical relevance, ethical behavior, and contextual distraction with confidence.	Held-out test versus human majority vote: accuracy or weighted F_1-score of approximately 0.82 - 0.84; hybrid model outperformed baselines.	No learner-facing narrative feedback; outputs are fuzzy evaluation labels.
Zhou et al, 2025 [47]	Diagnostic accuracy in pulmonary disease	Diagnostic accuracy	Diagnostic accuracy reported post hoc as research outcome; not described as learner-facing automatic scoring.	Observational analysis only; no interrater or AI-human scoring agreement reported.	No automatic qualitative feedback reported; targeted feedback noted as future direction.

XSL•FO
RenderX

Study ID	Competency domains assessed	Quantitative scoring approach	Score derivation or metric	Score accuracy or reliability evaluation	Qualitative feedback
Chen, 2026 [48]	Empathic obstetric nursing communication	Rubric score	ECP-RS[v]: 5 empathy domains rated 0 - 2; total /10 using NLP and speech or paraverbal analysis.	AI versus educator average ICC=0.81 single-measure, 0.89 average-measure; human ICCs 0.82 - 0.88.	Personalized feedback by empathy domain: strengths, improvements, strategies for language, tone, pacing, and context.
Hicke et al, 2026 [49]	Clinical interviewing; history-taking; and communication	Hybrid rubric+checklist	MIRS[w] items rated 1 - 5 with anchored prompts; checklist completion present, absent, or unknown with excerpts.	External OSCE validation: exact accuracy 32.5%, off-by-one 64.1%, thresholded accuracy 87.0%.	Post-encounter itemized strengths or gaps with evidence quotes, suggestions, checklist feedback, and Learning Hub scaffolds.
Reynolds et al, 2026 [50]	Pharmacy counseling competencies	Rubric or ordinal scores	AI grading agent scores SME[x]-developed counseling rubric categories; categorical or ordinal scores.	AI versus human consensus Pearson *r* varied: −0.037 to 0.717; weaker for inference-heavy categories.	No automatic qualitative feedback described; rubric scores only.

[a]For each study, the table reports the competency domains assessed, quantitative scoring approach, score derivation or metric, evaluation of scoring accuracy or reliability, and whether qualitative feedback was provided.

[b]OSCE: objective structured clinical examination.

[c]VP-OSCE: virtual patient–objective structured clinical examination.

[d]NLP: natural language processing.

[e]ICC: intraclass correlation coefficient.

[f]IRR: interrater reliability.

[g]HOPI: history of present illness.

[h]PMH: past medical history.

[i]FH: family history.

[j]SH: social history.

[k]CNN: convolutional neural network.

[l]CRI-HTI: clinical reasoning indicators for history-taking.

[m]QAP: question-answer pair.

[n]CUQ: Chatbot Usability Questionnaire.

[o]UI: user interface.

[p]LLM: large language model.

[q]3E: empathize, be explicit, and empower.

[r]RCT: randomized controlled trial.

[s]ACT: Advanced Communication Training.

[t]BERT: bidirectional encoder representations from transformers.

[u]QSUM: Quality of Communication questionnaire summary score.

[v]ECP-RS: Empathic Communication and Perspective-Taking Rating Scale.

[w]MIRS: Medical Interview Rating Scale.

[x]SME: subject matter expert.

Quantitative Scoring Approaches and Accuracy Validation

Although all included studies generated some form of learner performance measure, the source of scoring varied across studies. In 3 (14%) systems, the AI-SSP functioned primarily as the simulated patient, while performance was scored by human faculty, raters, or coders after the encounter using transcripts [36,42,43,48]. In other studies, scoring was automated by the AI-SSP system itself [38,39,44-46,49,50]. Several studies used both scoring approaches, and AI-generated scores were compared directly with human-led results [32,40,44,45,48-50].

Checklist-based scoring was the most common strategy. Ten (48%) studies used this approach where they translated expert-defined case content into required history-taking questions and clinical actions, then scored performance as item coverage, percent completion, or weighted checklist scores [30-33,37-39,44,45,49]. Nine (43%) studies used rubric-based or Likert scale scoring, particularly for communication-intensive competencies, such as empathy, professionalism, counseling, clarity, and ethical appropriateness [35,36,40,42,43,46,48-50]. A smaller number of studies used hybrid scoring formats, combining checklist completion with rubric ratings,

XSL•FO
RenderX

classifier-generated process metrics, quizzes, diagnosis accuracy scores, or missed mandatory topic counts [34,42,43,49].

Where scores were AI generated, their accuracy and reliability were evaluated unevenly across the literature. Several studies did not formally validate automated scoring against a human benchmark [30,35,37,47], while others relied on internal consistency, classifier performance, or post hoc learning outcomes. Among studies comparing AI-generated scores with human ratings, the results were generally promising but variable. Correlation-based validation showed moderate to strong associations in Bond et al (Spearman ρ=0.80) [32], Li et al (Spearman r=0.66 - 0.95 and Pearson r=0.77 - 0.94) [39], Gorcilov et al (Pearson r=0.69 - 0.88) [41], and Qian et al (Pearson r=0.67 - 0.90) [45]. However, Reynolds et al [50] reported wider variability across checklist categories, with correlations ranging from negative values to 0.72.

In addition to correlation tests, agreement and reliability analyses similarly showed mixed but overall positive findings. Holderried et al [38] reported strong GPT-4 versus human agreement, with Cohen κ of 0.832 across 106 conversations and 45 categories [38]. Liu et al [44] reported high human-AI agreement across 3 cases, with intraclass correlation coefficients (ICCs) of 0.92 to 0.98 and item-level consistency of 95.75% to 97.13% [44]. Qian et al [45] reported ICCs of 0.78 to 0.82 after adjusting for consistent differences, while Chen [48] reported an AI versus educator average ICC of 0.81, comparable to human-human educator ICCs of 0.82 to 0.88. Classification-based evaluations were more variable: Haut [42,43] reported bidirectional encoder representations from transformers classifier accuracy of 72% and an F_1-score of 51%, Zheng et al [46] reported accuracy of 0.82 to 0.84 and a weighted F_1-score of 0.82 to 0.84, and Hicke et al [49] reported exact accuracy of 32.5%, off-by-one accuracy of 64.1%, and thresholded accuracy of 87.0%.

Several studies also identified important AI-based scoring limitations. Chiu et al [40] found that ChatGPT scores were substantially higher than faculty scores and sensitive to prompt design, while Qian et al [45] reported that LLMs were consistently around 12 percentage points higher than expert scores before calibration. Reynolds found that AI-human correlations varied widely by rubric category, including negative correlations for certain counseling-related items [50]. Although human-human reliability was sometimes reported as a benchmark [32,36,42,43,48], it was not consistently used to determine whether automated scoring achieved human-level performance [32,36,42,43,48]. Overall, while several systems demonstrated moderate to strong alignment with human ratings, validation approaches remained inconsistent, and evidence for reliably automated scoring was still limited.

Qualitative Feedback Approaches

Qualitative feedback was common across studies but varied substantially in specificity and actionability. Earlier systems often provided checklist- or gap-based feedback, identifying missed questions, omitted clinical actions, or incomplete case elements [31-35,40,49]. Newer LLM-enabled systems increasingly generated narrative feedback that included strengths, weaknesses, transcript-based examples, and personalized improvement suggestions [36,38,40-43,48]. Several studies used hybrid feedback approaches that combined checklist completion and narrative comments [39,42,43,48,49]. Overall, feedback evolved from relatively simple “what was missed” summaries toward more individualized coaching focused on communication, clinical reasoning, and next steps for improvement.

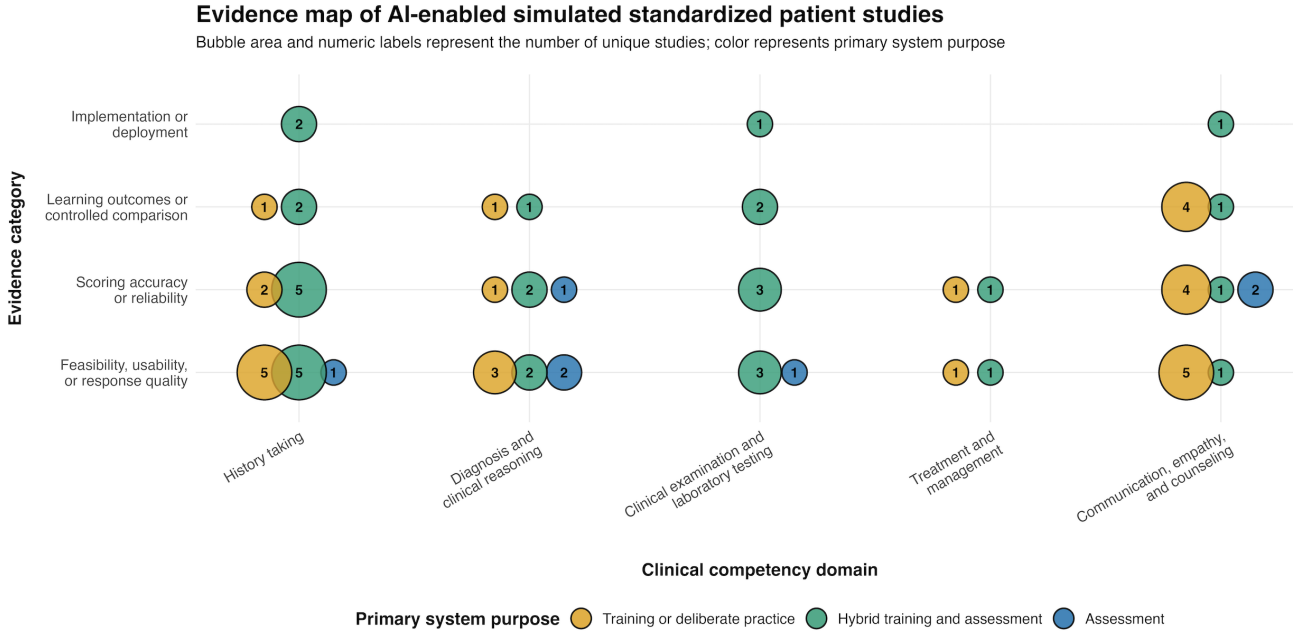
AI-SSP Product Evaluation and Evidence Maturity

Beyond scoring accuracy, studies also evaluated AI-SSP systems using a range of product-level outcomes, including learner experience, AI-SSP response quality, and learning outcomes. Several studies reported high learner satisfaction, perceived usefulness, ease of navigation, or willingness to recommend the system [30,32-34,37,38,40,42-44,49]. Response quality and realism were also commonly evaluated, assessing whether AI-SSP responses were plausible, realistic, or aligned with expected case content [30,32,38,39,41-43]. Evidence for downstream skills improvement was more limited, with only 6 (6/21, 29%) studies assessing this outcome [36,40,42,43,48,49].

As illustrated in Figure 3, the available evidence was concentrated primarily in early-stage product evaluation rather than learning outcomes, controlled comparisons, or sustained implementation. Feasibility, usability, or response quality was the most frequently assessed evidence category, reported in 18 (86%) studies, followed closely by scoring accuracy or reliability in 16 (76%) studies. In contrast, only 8 (38%) studies evaluated learning outcomes or incorporated a controlled comparison, and this evidence was unevenly distributed across competency domains, with the strongest concentration in communication-related skills and history-taking. Evidence was particularly limited for treatment and management and for clinical examination or laboratory testing, suggesting that the literature has not yet established whether AI-SSPs can support assessment across the full clinical workflow. Only 2 (10%) studies reported evidence of implementation or deployment, indicating that most systems remain at the pilot, proof-of-concept, or validation stage. Studies of training-oriented systems contributed mainly to feasibility, usability, and early learning evidence, whereas studies of hybrid training and assessment systems demonstrated the broadest evidence profile, accounting for all identified implementation evidence and a substantial share of scoring validation evidence. Assessment-focused systems were more narrowly concentrated in scoring accuracy and reliability, with comparatively little evidence on learning effects or real-world adoption. Overall, the evidence gap map shows that the field has made meaningful progress in demonstrating technical feasibility and initial scoring accuracy and reliability, but stronger evidence is still needed on educational effectiveness, sustained use, scalability, integration into high-stakes assessment, and performance across the full range of clinical competency domains.

XSL•FO
RenderX

Figure 3. Evidence gap map of AI-enabled simulated standardized patient studies. The figure maps the distribution of included studies across 5 clinical competency domains—history-taking; diagnosis and clinical reasoning; clinical examination and laboratory testing; treatment and management; and communication, empathy, and counseling—and 4 categories of evidence maturity: feasibility, usability, or response quality; scoring accuracy or reliability; learning outcomes or controlled comparison; and implementation or deployment. Each bubble represents the number of unique studies at a given intersection of competency domain, evidence category, and primary system purpose. Bubble area is proportional to the number of studies, and the numeric label inside each bubble gives the exact count. Bubble color indicates the primary purpose of the system: training or deliberate practice, hybrid training and assessment, or assessment. The unit represented is the study rather than the unique AI-SSP system; therefore, separate publications evaluating the same system are counted separately. Studies could also contribute to more than one competency domain and more than one evidence category. Accordingly, counts are not mutually exclusive and should not be summed across cells.



Discussion

This scoping review aimed to map how AI-enabled SSP systems have been used for clinical competency assessment, with attention to system architecture, interaction design, assessed clinical competencies, scoring and feedback approaches, and evidence maturity. Across 21 studies and 20 unique AI-SSP systems, we found that the field is expanding rapidly, with 15 (71%) studies published in 2024 or later and a clear temporal shift from earlier rule-based virtual patient systems to LLM-enabled conversational platforms after 2022. The evidence base remained concentrated in high-income academic settings, focused primarily on medical students, and dominated by pilot, feasibility, and proof-of-concept studies rather than large-scale implementation research. Although all included studies produced some form of competency-related performance measure, AI-SSPs were most often positioned as formative learning tools rather than formal assessment systems, with 9 systems focused primarily on deliberate practice, 6 combining training and assessment, and 5 explicitly assessment-oriented. Across studies, history-taking was the most commonly assessed competency, while laboratory tests, diagnosis, treatment planning, communication, empathy, and professionalism were assessed less consistently. Quantitative scoring and qualitative feedback were common but heterogeneous: checklist-based scoring was most frequent, rubric-based scoring was used mainly for communication-intensive competencies, and validation of automated scoring remained inconsistent. Overall, our findings suggest that AI-SSPs are increasingly capable of supporting scalable and repeatable competency assessment, but the current literature is more mature in demonstrating feasibility and usability than in establishing validity and implementation readiness.

Previous research on virtual patients has shown that digital patient simulations can give learners consistent clinical cases to practice with, helping them build medical knowledge, clinical reasoning skills, and confidence [15,16]. However, these systems have historically functioned primarily as educational simulations rather than scalable competency assessment infrastructures. Similarly, recent reviews of LLM-based virtual patients describe a newer wave of systems focused on conversational realism, history-taking, and medical training, reflecting broader enthusiasm for generative AI in health professions education [17-19]. Our findings build on this literature by showing that AI-SSPs are beginning to move beyond simulation alone toward competency measurement but remain in a transitional stage. The shift from rule-based virtual patient systems to LLM-enabled platforms has improved conversational fluency, responsiveness, and perceived realism, but most systems still rely on human-authored cases, structured prompts, and rubrics, which represent important guardrails against hallucinations. Current AI-SSPs are not fully agentic assessors of clinical competence, but semistructured assessment environments in which generative AI is layered onto human-defined clinical content and scoring logic.

The concentration of assessed competencies also has important implications for how AI-SSPs should be interpreted as assessment tools. Consistent with the broader virtual patient literature, the included systems most often assessed history-taking, diagnosis, and communication-related skills, domains that map naturally onto text-based or conversational

XSL•FO
RenderX

interaction [16,17,51]. However, fewer studies assessed physical examination, laboratory or investigation selection, and management planning. This pattern suggests that current AI-SSPs are strongest for evaluating how learners gather information and communicate with patients but less mature for assessing the full clinical workflow from examination to diagnostic interpretation and treatment decision-making. For formative education, this may be sufficient and even valuable. For competency assessment, however, the narrow domain coverage raises concerns about how a learner may perform well in an AI-SSP conversation without demonstrating the broader set of skills required for safe and effective clinical care.

In terms of automated scoring of performances, our review showed that checklist-based scoring was most common, reflecting the need to convert complex encounters into observable items. Rubric-based scoring was used mainly for communication-intensive domains, where performance depended not only on whether an item was mentioned but also on how the learner expressed empathy, professionalism, clarity, or ethical reasoning. Several included studies reported moderate to strong agreement between AI-generated and human or expert scores, suggesting that automated scoring may reduce faculty burden and enable more frequent formative assessment. However, there is also evidence of consistent score inflation, prompt sensitivity, and variable performance across rubric categories in other studies. This pattern mirrors findings from the comparison of AI and human evaluation of the same video-recorded OSCE performances by Tekin et al [52], in which AI models consistently assigned higher scores than human evaluators across clinical skills, including a mean score of 28.23 versus 25.25 for intramuscular injection. Similar tendencies appeared in our included studies: Chiu et al [40] found that ChatGPT scores were substantially higher than faculty scores, and Qian et al [45] reported that LLM scores were approximately 12 percentage points higher than expert scores before calibration. These findings suggest that AI scoring may be scalable and internally consistent, but not necessarily equivalent to human judgment. Additionally, while several studies reported high correlations or ICCs between AI and human raters, evidence regarding construct validity, fairness, and suitability for high-stakes summative competency assessment remains limited. Stronger validation should therefore include transparent scoring logic, calibration against expert benchmarks, comparison with human-human reliability, and domain-specific evaluation rather than reliance on aggregate correlations alone.

The overall evidence maturity of AI-SSPs remains limited compared to their technical promise. This pattern is consistent with broader reviews of digital simulation and AI in health professions education, which have found that usability, satisfaction, and short-term learning outcomes are more commonly evaluated than large-scale trials, implementation, or real-world clinical impact [22]. In this review, most studies demonstrated that AI-SSPs were feasible, acceptable, realistic, or capable of producing structured scores and feedback, but fewer examined whether they improved clinical competency, could be integrated into existing curricula, or performed reliably across various types of users, especially in low-resource contexts. As promising prototypes do not automatically translate into scalable assessment infrastructure, future studies should move beyond product-level feasibility toward larger-scale validation, assess longitudinal learning outcomes, and test in diverse geographic, socioeconomic, and health professional contexts.

A further implementation concern is the need for AI safety protections when AI-SSPs are used for assessment, especially for future high-stakes summative assessments beyond formative ones. As these systems may collect detailed learner transcripts, performance scores, and interaction logs, institutions will need clear data governance policies covering consent, storage, and secondary use of assessment data [53]. Technical guardrails are also needed to prevent unsafe or inappropriate patient responses, protect confidential case content, and ensure that automated feedback does not provide misleading clinical guidance [54]. These governance and safety issues are central to whether AI-SSPs can be trusted as assessment tools. Furthermore, assessment-oriented AI-SSPs also create risks related to new ways of cheating, plagiarism, and the authenticity of performance, particularly if learners can use external AI tools, reuse shared prompts, or access case materials outside the assessment setting. Future systems should therefore include procedural safeguards such as secure testing environments, controlled access to cases and rubrics, and identity verification when appropriate.

This review has several limitations. First, it synthesizes heterogeneous study designs and tool implementations, many of which were small-scale pilots or descriptive prototypes rather than fully validated systems. Second, the rapid evolution of LLM technologies also means that some findings may become outdated as models and deployment practices change. Third, most included studies originated in high-income settings, limiting generalizability to low- and middle-income health systems where SSPs may offer the greatest potential value. At the same time, this review has important strengths. It is, to our knowledge, the first to systematically characterize AI-led SSPs for clinical competency “assessment.” We explored frontend design, backend architecture, and evaluation methods, offering a holistic view of how these tools are currently constructed and assessed. By identifying common architectural patterns and persistent gaps, this review provides a conceptual framework for future SSP development and highlights key design and validation priorities to advance scalable, reliable clinical competency assessment.

In conclusion, AI-enabled SSPs represent an important emerging direction for clinical competency assessment, but the field remains in an early stage of development. This review shows that recent advances in LLMs have accelerated a shift from rule-based virtual patient systems toward more conversational and flexible assessment environments. At the same time, most systems continue to rely on human-authored cases, structured prompts, checklists, rubrics, and other guardrails, underscoring that AI-SSPs are not autonomous assessors of clinical competence. The current evidence base suggests strongest readiness for formative assessment of history-taking, communication, and selected clinical reasoning skills, while broader clinical workflows, including clinical examination,

XSL•FO
RenderX

laboratory tests, treatment planning, and management, remain less consistently evaluated. For AI-SSPs to move from promising prototypes to trustworthy assessment infrastructure, future studies must establish stronger evidence of reliability, validity, fairness, and implementation feasibility across diverse learners, professions, and health system contexts. With rigorous validation and responsible governance, AI-SSPs could help expand access to more frequent, standardized, and scalable competency assessments, particularly in settings where traditional SP-based assessments are difficult to sustain.

Acknowledgments

The authors declare the use of generative AI (Gen-AI) in the research and writing process. According to the GAIDeT (Generative AI Delegation Taxonomy; 2025), the following tasks were delegated to Gen-AI tools under full human supervision: validation, proofreading, and editing. The Gen-AI tools used were GPT-5.5 and Grammarly. Responsibility for the final manuscript lies entirely with the authors. Gen-AI tools are not listed as authors and do not bear responsibility for the final outcomes.

Funding

The authors declared no financial support was received for this work.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Complete database search strategies. Detailed search strategies, including databases and platforms searched, search dates, keywords, controlled vocabulary terms, Boolean operators, and limits applied.
[DOCX File, 23 KB - mededu_v12i1e92826_app1.docx]

Checklist 1
PRISMA-S checklist.
[PDF File, 115 KB - mededu_v12i1e92826_app2.pdf]

References

1. Ghebreyesus TA. All roads lead to universal health coverage. Lancet Glob Health 2017 Sep;5(9):e839-e840. [doi: 10.1016/S2214-109X(17)30295-4] [Medline: 28728920]
2. Ranson MK, Chopra M, Atkins S, Dal Poz MR, Bennett S. Priorities for research into human resources for health in low- and middle-income countries. Bull World Health Organ 2010 Jun;88(6):435-443. [doi: 10.2471/BLT.09.066290] [Medline: 20539857]
3. Chen L, Evans T, Anand S, et al. Human resources for health: overcoming the crisis. Lancet 2004;364(9449):1984-1990. [doi: 10.1016/S0140-6736(04)17482-5] [Medline: 15567015]
4. Daniels B, Das J, Gatti R, Yi Chang A. Caseloads and competence in sub-saharan africa: a fundamental reassessment of the human resources crisis in primary health care. : World Bank; 2025 URL: https://hdl.handle.net/10986/42721 [doi: 10.1596/1813-9450-11043]
5. Slawomirski L, Klazinga N. The economics of patient safety: from analysis to action. : OECD Publishing; 2022 Aug. [doi: 10.1787/761f2da8-en]
6. Matet A, Fournel L, Gaillard F, et al. Impact of integrating objective structured clinical examination into academic student assessment: large-scale experience in a French medical school. PLoS ONE 2021;16(1):e0245439. [doi: 10.1371/journal.pone.0245439] [Medline: 33444375]
7. Hijazi M, Downing SM. Objective structured clinical examinations as an assessment method in residency training: practical considerations. Ann Saudi Med 2008;28(3):192-199. [doi: 10.5144/0256-4947.2008.192] [Medline: 18500173]
8. Behera P, Gopi G, Siddaiah A, Sridhar PR, Patro BK, Subba SH. Challenges of objective structured clinical examination as a tool in medical assessment. Indian Journal of Community and Family Medicine 2022;8(2):156-160. [doi: 10.4103/ijcfm.ijcfm_12_22]
9. Conner R, Gatti R, Andrews K, Avitabile C, Sharma J, Chang AY. The quality of health and education systems across Africa: evidence from a decade of service delivery indicators surveys. World Bank Group. 2021 Nov. URL: https://documents.worldbank.org/en/publication/documents-reports/documentdetail/380481637326015488 [accessed 2026-02-12]
10. Hauer KE, Park YS, Bullock JL, Tekian A. "My assessments are biased!" measurement and sociocultural approaches to achieve fairness in assessment in medical education. Acad Med 2023 Aug 1;98(8S):S16-S27. [doi: 10.1097/ACM.0000000000005245] [Medline: 37094278]

XSL•FO
RenderX

11. Arsever S, Broers B, Cerutti B, Wiesner J, Dao MD. A gender biased hidden curriculum of clinical vignettes in undergraduate medical training. Patient Educ Couns 2023 Nov;116:107934. [doi: 10.1016/j.pec.2023.107934] [Medline: 37595505]
12. Ismaeel R, Pusic L, Gottlieb M, Chan TM, Oyedokun TO, Thoma B. Bias in observed assessments in medical education: a scoping review. Acad Med 2024 Dec 1;99(12):1438-1450. [doi: 10.1097/ACM.0000000000005794] [Medline: 38924499]
13. Chong L, Taylor S, Haywood M, Adelstein BA, Shulruf B. The sights and insights of examiners in objective structured clinical examinations. J Educ Eval Health Prof 2017;14:34. [doi: 10.3352/jeehp.2017.14.34] [Medline: 29278906]
14. Schleicher I, Leitner K, Juenger J, et al. Examiner effect on the objective structured clinical exam - a study at five medical schools. BMC Med Educ 2017 Apr 24;17(1):71. [doi: 10.1186/s12909-017-0908-1] [Medline: 28438196]
15. Cook DA, Erwin PJ, Triola MM. Computerized virtual patients in health professions education: a systematic review and meta-analysis. Acad Med 2010 Oct;85(10):1589-1602. [doi: 10.1097/ACM.0b013e3181edfe13] [Medline: 20703150]
16. Kononowicz AA, Woodham LA, Edelbring S, et al. Virtual patient simulations in health professions education: systematic review and meta-analysis by the digital health education collaboration. J Med Internet Res 2019 Jul 2;21(7):e14676. [doi: 10.2196/14676] [Medline: 31267981]
17. Li D, Lebai Lutfi S. Large language model-based virtual patient systems for history-taking in medical education: comprehensive systematic review. JMIR Med Inform 2026 Jan 2;14:e79039. [doi: 10.2196/79039] [Medline: 41481915]
18. Zeng J, Qi W, Shen S, et al. Embracing the future of medical education with large language model-based virtual patients: scoping review. J Med Internet Res 2025 Nov 13;27:e79091. [doi: 10.2196/79091] [Medline: 41232097]
19. Yu H, Zhou J, Li L, et al. Simulated patient systems powered by large language model-based AI agents offer potential for transforming medical education. Commun Med (Lond) 2025 Dec 19;6(1):27. [doi: 10.1038/s43856-025-01283-x] [Medline: 41420084]
20. Gandhi B, Morjaria L, Illamperuma I, Nadesan P, Arora A, Sibbald M. Large language model-based virtual patients for simulated clinical learning: a scoping review. AI in Medicine 2026 Mar;1(1):7. [doi: 10.3390/aimed1010007]
21. Lin Y, Luo Z, Ye Z, et al. Applications, challenges, and prospects of generative artificial intelligence empowering medical education: scoping review. JMIR Med Educ 2025 Oct 23;11:e71125. [doi: 10.2196/71125] [Medline: 41128430]
22. Shaw K, Henning MA, Webster CS. Artificial intelligence in medical education: a scoping review of the evidence for efficacy and future directions. Med Sci Educ 2025 Jun;35(3):1803-1816. [doi: 10.1007/s40670-025-02373-0] [Medline: 40625971]
23. Feigerlova E, Hani H, Hothersall-Davies E. A systematic review of the impact of artificial intelligence on educational outcomes in health professions education. BMC Med Educ 2025 Jan 27;25(1):129. [doi: 10.1186/s12909-025-06719-5] [Medline: 39871336]
24. JBI Manual for Evidence Synthesis. URL: https://synthesismanual.jbi.global [accessed 2026-06-07]
25. Peters MDJ, Marnie C, Tricco AC, et al. Updated methodological guidance for the conduct of scoping reviews. JBI Evid Synth 2020 Oct;18(10):2119-2126. [doi: 10.11124/JBIES-20-00167] [Medline: 33038124]
26. Tricco AC, Lillie E, Zarin W, et al. PRISMA Extension for Scoping Reviews (PRISMA-ScR): checklist and explanation. Ann Intern Med 2018 Oct 2;169(7):467-473. [doi: 10.7326/M18-0850] [Medline: 30178033]
27. Rethlefsen ML, Kirtley S, Waffenschmidt S, et al. PRISMA-S: an extension to the PRISMA statement for reporting literature searches in systematic reviews. J Med Libr Assoc 2021 Apr 1;109(2):174-200. [doi: 10.5195/jmla.2021.962] [Medline: 34285662]
28. A scoping review of AI-assisted vignettes for clinical competency assessment. OSF. URL: https://osf.io/bu5en/ [accessed 2026-07-27]
29. Page MJ, et al. BMJ 2021;372:71. [doi: 10.1136/bmj.n71]
30. Courteille O, Bergin R, Stockeld D, Ponzer S, Fors U. The use of a virtual patient case in an OSCE-based exam--a pilot study. Med Teach 2008;30(3):e66-e76. [doi: 10.1080/01421590801910216] [Medline: 18484444]
31. Oliven A, Nave R, Gilad D, Barch A. Implementation of a web-based interactive virtual patient case simulation as a training and assessment tool for medical students. Stud Health Technol Inform 2011;169:233-237. [Medline: 21893748]
32. Bond WF, Lynch TJ, Mischler MJ, et al. Virtual standardized patient simulation: case development and pilot application to high-value care. Simul Healthc 2019 Aug;14(4):241-250. [doi: 10.1097/SIH.0000000000000373] [Medline: 31116172]
33. Maicher KR, Zimmerman L, Wilcox B, et al. Using virtual standardized patients to accurately assess information gathering skills in medical students. Med Teach 2019 Sep;41(9):1053-1059. [doi: 10.1080/0142159X.2019.1616683] [Medline: 31230496]
34. Yang W, Hebert D, Kim S, Kang B. MCRDR knowledge-based 3D dialogue simulation in clinical training and assessment. J Med Syst 2019 May 23;43(7):200. [doi: 10.1007/s10916-019-1262-0] [Medline: 31123826]
35. Lee C, Low MYH, Soon BTC, Lu L, Lee H, Patel NS. Development of a chatbot to train physiotherapy students in clinical questioning and reasoning. Presented at: 2022 IEEE International Conference on Teaching, Assessment and Learning for Engineering (TALE); Dec 4-7, 2022; Hung Hom, Hong Kong p. 477-481. [doi: 10.1109/TALE54877.2022.00084]
36. Brügge E, Ricchizzi S, Arenbeck M, et al. Large language models improve clinical decision making of medical students through patient simulation and structured feedback: a randomized controlled trial. BMC Med Educ 2024 Nov 28;24(1):1391. [doi: 10.1186/s12909-024-06399-7] [Medline: 39609823]

XSL•FO
RenderX

37. Holderried F, Stegemann-Philipps C, Herschbach L, et al. A generative pretrained transformer (GPT)-powered Chatbot as a simulated patient to practice history taking: prospective, mixed methods study. JMIR Med Educ 2024 Jan 16;10:e53961. [doi: 10.2196/53961] [Medline: 38227363]
38. Holderried F, Stegemann-Philipps C, Herrmann-Werner A, et al. A language model-powered simulated patient with automated feedback for history taking: prospective study. JMIR Med Educ 2024 Aug 16;10:e59213. [doi: 10.2196/59213] [Medline: 39150749]
39. Li Y, Zeng C, Zhong J, Zhang R, Zhang M, Zou L. Leveraging large language model as simulated patients for clinical education. arXiv. Preprint posted online on Apr 13, 2024. [doi: 10.48550/arXiv.2404.13066]
40. Chiu J, Castro B, Ballard I, et al. Exploration of the role of ChatGPT in teaching communication skills for medical students: a pilot study. Med Sci Educ 2025 Aug;35(4):1871-1882. [doi: 10.1007/s40670-025-02394-9] [Medline: 41112920]
41. Gorcilov J, Lim M, Leslie A, et al. Large language models deliver interactive learning cases accurately and with rapid appropriate feedback. Heart Lung Circ 2025;34:S497-S498. [doi: 10.1016/j.hlc.2025.06.625]
42. Haut K, Hasan M, Carroll T, Epstein R, Sen T, Hoque E. AI standardized patient improves human conversations in advanced cancer care. arXiv. Preprint posted online on May 5, 2025 URL: http://arxiv.org/abs/2505.02694 [accessed 2026-08-04]
43. Haut K. USIHC: understand, simulate, and improve human compassion in serious illness communication (SIC) [Dissertation]. : University of Rochester; 2025 URL: http://hdl.handle.net/1802/38511 [accessed 2026-08-04]
44. Liu Y, Shi C, Wu L, et al. Development and validation of a large language model-based system for medical history-taking training: prospective multicase study on evaluation stability, human-AI consistency, and transparency. JMIR Med Educ 2025 Aug 29;11:e73419. [doi: 10.2196/73419] [Medline: 40882613]
45. Qian C, Gao C, Park SO, et al. Use of large language models for rapid quantitative feedback in case-based learning: a pilot study. Med Sci Educ 2025 Jun;35(3):1169-1171. [doi: 10.1007/s40670-025-02343-6] [Medline: 40625928]
46. Zheng W, Turner L, Kropczynski J, Ozer M, Nguyen T, Halse S. LLM-as-a-fuzzy-judge: fine-tuning large language models as a clinical evaluation judge with fuzzy logic. arXiv. Preprint posted online on Jun 12, 2025. [doi: 10.48550/arXiv.2506.11221]
47. Zhou CY, Kelleher M, Overla S, Weber D, Turner L. Use of artificial intelligence to identify student behaviors associated with improved diagnostic accuracy of pulmonary diseases. Chest 2025 Oct;168(4):A4483. [doi: 10.1016/j.chest.2025.07.2506]
48. Chen PJ. AI-enhanced virtual reality simulation for nursing students' empathy: automated scoring and inter-rater reliability in a randomised controlled study. Nurse Educ Today 2026 Apr;159:106968. [doi: 10.1016/j.nedt.2025.106968] [Medline: 41500151]
49. Hicke Y, et al. MedSimAI: simulation and formative feedback generation to enhance deliberate practice in medical education. arXiv. Preprint posted online on Mar 1, 2026. [doi: 10.48550/arXiv.2503.05793]
50. Reynolds EC, Benton E, Lu J, Schnable C, Southwood R. AI versus human consensus: correlation of rubric-based scoring in pharmacy counseling. Presented at: 2025 ASHP Midyear Clinical Meeting; Dec 7-10, 2025. [doi: 10.1093/ajhp/zxaf258]
51. García-Torres D, Vicente Ripoll MA, Fernández Peris C, Mira Solves JJ. Enhancing clinical reasoning with virtual patients: a hybrid systematic review combining human reviewers and ChatGPT. Healthcare (Basel) 2024 Nov 11;12(22):2241. [doi: 10.3390/healthcare12222241] [Medline: 39595439]
52. Tekin M, Yurdal MO, Toraman Ç, Korkmaz G, Uysal İ. Is AI the future of evaluation in medical education?? AI vs. human evaluation in objective structured clinical examination. BMC Med Educ 2025 May 1;25(1):641. [doi: 10.1186/s12909-025-07241-4] [Medline: 40312328]
53. Janssen M, Brous P, Estevez E, Barbosa LS, Janowski T. Data governance: organizing data for trustworthy artificial Intelligence. Gov Inf Q 2020 Jul;37(3):101493. [doi: 10.1016/j.giq.2020.101493]
54. Zavaleta-Monestel E, Arguedas-Chacón S, Mora-Jiménez J, Millán González R. When compliance is not safety: the regulatory blind spot in ai companion chatbots. Cureus 2026 Mar;18(3):e105902. [doi: 10.7759/cureus.105902] [Medline: 42037948]

Abbreviations

AI: artificial intelligence
GAIDeT: Generative AI Delegation Taxonomy
Gen-AI: generative artificial intelligence
ICC: intraclass correlation coefficient
LLM: large language model
OSCE: objective structured clinical examination
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
PRISMA-S: Preferred Reporting Items for Systematic Reviews and Meta-Analyses Literature Search Extension
PRISMA-ScR: Preferred Reporting Items for Systematic Reviews and Meta-Analyses Extension for Scoping Reviews
SP: standardized patient
SSP: simulated standardized patient

XSL•FO
RenderX

UHC: Universal Health Coverage
VR: virtual reality

Edited by S Brini; submitted 03.Feb.2026; peer-reviewed by D Lungeanu, S Mohanadas, S Ismaile, S Brownie, W Xu; revised version received 29.Jun.2026; accepted 02.Jul.2026; published 14.Aug.2026.

Please cite as:
Zhang W(, Daniels BB, Mita C, Nguyen H, Duong DB
AI for Clinical Competency Assessment: Scoping Review of Methods and Applications
JMIR Med Educ 2026;12:e92826
URL: https://mededu.jmir.org/2026/1/e92826
doi:10.2196/92826

© Wenjia (Stella) Zhang, Benjamin Daniels, Carol Mita, Hoang Nguyen, David B Duong. Originally published in JMIR Medical Education (https://mededu.jmir.org), 14.Aug.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Education, is properly cited. The complete bibliographic information, a link to the original publication on https://mededu.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Application of AI-Generated Content in Medical Education: Systematic Review of the Impact on Critical Thinking Abilities of Medical Students

Jinlei Li[1,2*], MBBS; Fen Ai[2*], MM; Jueyan Wang[1,2], MBBS; Bingxin Cheng[1,2], MBBS; Yu Li[1,2], MM; Zhen Chen[2], MD

[1]School of Medicine, Jianghan University, Wuhan, China

[2]Department of Emergency Medicine, The Central Hospital of Wuhan, Tongji Medical College, Huazhong University of Science and Technology, 26 Shengli Street, Jiang'an District, Wuhan, China

[*]these authors contributed equally

Corresponding Author:
Zhen Chen, MD
Department of Emergency Medicine, The Central Hospital of Wuhan, Tongji Medical College, Huazhong University of Science and Technology, 26 Shengli Street, Jiang'an District, Wuhan, China

Abstract

Background: With the rapid development of artificial intelligence technology, artificial intelligence–generated content (AIGC) is increasingly widely applied in the field of medical education. Large language models, such as ChatGPT, are a prominent type of AIGC technology. Critical thinking is a core ability in medical education, but the impact of AIGC technology on the critical thinking ability of medical students remains unclear. Medical students are at a crucial stage in cultivating critical thinking, and the intervention of AIGC technology may have a profound impact on this process.

Objective: This study aims to systematically review the impact of AIGC technology on the complex mechanisms affecting medical students' critical thinking abilities and build a corresponding strategic framework. The findings are intended to provide theoretical support and practical guidance for applying AIGC in medical education.

Methods: This study followed 2020 PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) guidelines, with the retrieval scope limited to English studies published between November 2022 and June 2025. Through the PubMed database, combined with the search methods of subject terms and free words, relevant studies involving the impact of AIGC on the critical thinking of medical students were screened for using keywords such as "AIGC," "medical students," and "critical thinking." Two independent reviewers screened and evaluated the literature, and ultimately conducted a qualitative analysis based on the common themes extracted from the literature.

Results: AIGC technology in medical education is 2-fold. First, AIGC's powerful information capabilities provide abundant learning resources and efficient tools. This accelerates knowledge acquisition and broadens learning scope. Second, overreliance on AIGC may lead to mental inertia, weaken critical thinking skills, and cause academic integrity issues among students. Research has found that strategies such as customized AIGC tools, virtual standardized patients, new models of resource integration, and proactive assessment of AI limitations can effectively make up for the deficiencies of AIGC in cultivating high-level critical thinking, helping medical students maintain and enhance their critical thinking and problem-solving abilities.

Conclusions: AIGC technology application in medical education needs to carefully weigh the pros and cons. By optimizing the design and usage of AIGC tools and combining them with the guidance and supervision of educators, they can be transformed into powerful tools for promoting the development of critical thinking among medical students. Future research should further expand the scope of study, optimize research methods, pay attention to individual differences, track long-term effects, and deeply explore the influence of ethical and cultural factors to more comprehensively assess the application potential and challenges of AIGC technology in medical education.

(JMIR Med Educ 2026;12:e79939) doi:10.2196/79939

KEYWORDS

artificial intelligence–generated content; AIGC; medical students; large language models; LLM; ChatGPT; critical thinking

XSL•FO
RenderX

Introduction

Critical thinking is defined as a deliberate and self-regulatory cognitive process that involves the examination and evaluation of information to arrive at evidence-based conclusions. Within the medical domain, critical thinking is primarily manifested in 4 key areas: medical decision-making, ethical considerations, evidence-based medical practice, and scientific research capabilities. In the context of medical decision-making, health care professionals are required to synthesize multidimensional evidence, develop comprehensive diagnostic and treatment strategies, and ensure that their decisions are both scientifically sound and effective. From an ethical and safety standpoint, practitioners must recognize and rectify cognitive biases, avoid logical fallacies, and uphold standards of medical safety and ethical integrity. In the realm of evidence-based medicine, physicians are tasked with critically appraising research findings and discerning high-quality evidence to enhance clinical decision-making. Regarding scientific research competencies, medical professionals must possess strong logical reasoning, hypothesis testing skills, data analysis capabilities, and the ability to interpret results, all of which are essential for advancing medical research and clinical practice [1,2].

For medical students, the development of critical thinking skills is of paramount importance. This group is at a pivotal juncture in their journey to becoming physicians. This period represents an optimal time for fostering critical thinking abilities. Students who lack these skills may become passive recipients of knowledge. They will likely struggle to navigate the complexities, uncertainties, and unpredictabilities inherent in clinical practice [3]. Consequently, medical education should prioritize the cultivation of critical thinking, equipping students with the skills necessary to actively engage in problem-solving, analysis, and critical evaluation, thereby establishing a robust foundation for their future careers in medicine.

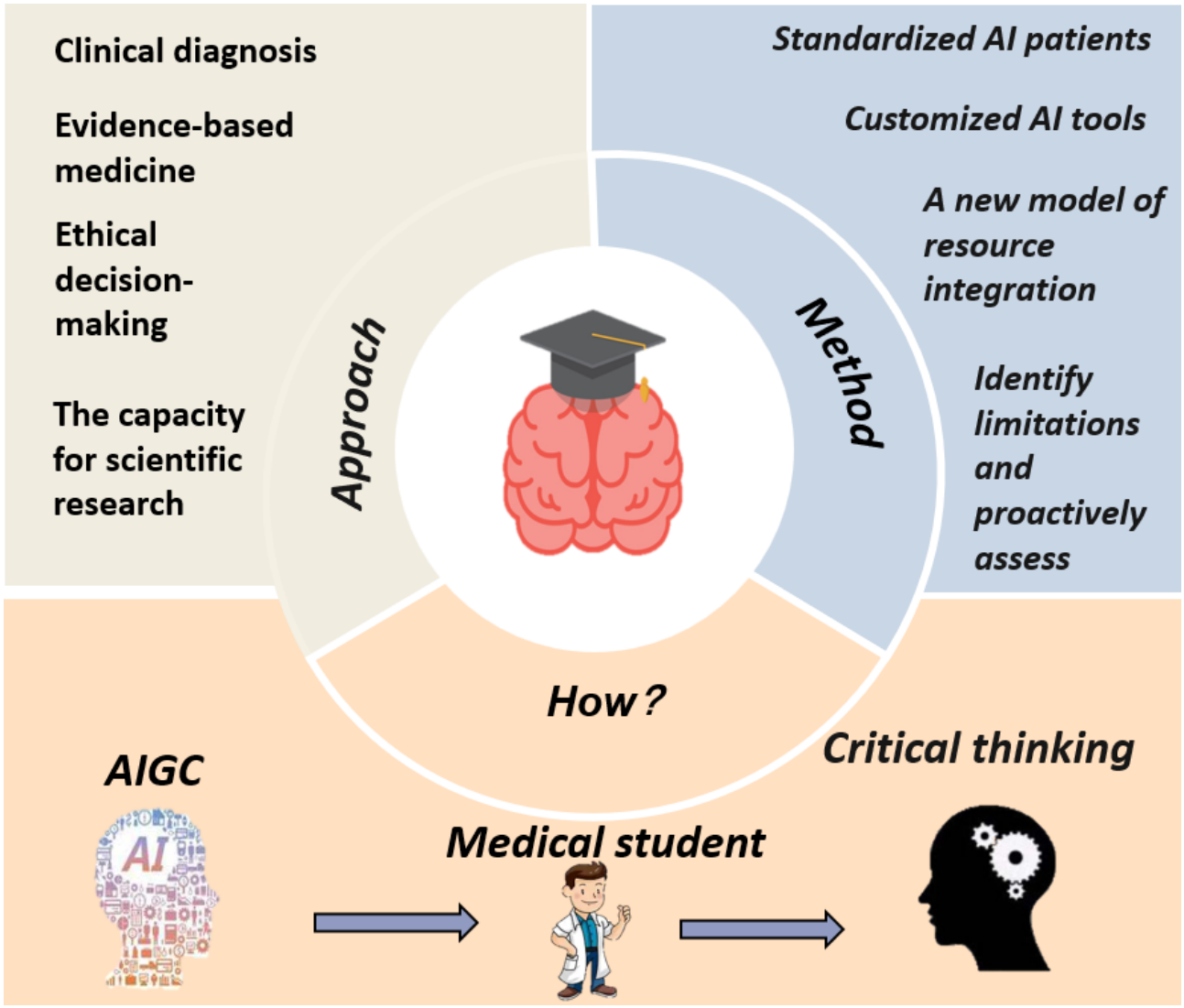
Artificial intelligence (AI)–generated content (AIGC) encompasses the application of generative AI technology to autonomously produce multimodal content, including text, images, and audio. A prominent category of AIGC tools is large language models (LLMs), which are designed to understand and generate humanlike text. ChatGPT is a widely used example of an LLM. By the year 2025, medical schools across the globe have progressively incorporated AIGC tools into their curricula. For instance, ChatGPT has demonstrated performance comparable to that of human candidates on the United States Medical Licensing Examination [4]. While AIGC can enhance educational efficiency, it raises significant concerns about its impact on critical thinking skills. Seetharaman et al [5] note that contemporary medical education often overemphasizes objective assessments while neglecting subjective competencies, such as communication and critical thinking. In this context, although AIGC tools (eg, LLMs such as ChatGPT) can simulate doctor-patient interactions, their use risks undermining the development of critical thinking in medical students.

This study seeks to elucidate the intricate mechanisms through which AIGC technology influences the critical thinking abilities of medical students, using a systematic literature review methodology. As illustrated in Figure 1, our proposed framework delineates a dual-pathway model wherein AIGC both supports and potentially hinders critical thinking development through various cognitive and educational mediators. The investigation encompasses a comprehensive analysis of global empirical studies on AIGC, with a particular focus on tools such as ChatGPT, from its public release in November 2022 through June 2025. Notably, this research introduces the novel “dual-path action model” and concentrates on 4 fundamental dimensions: clinical diagnostic reasoning, evidence-based medical practice, ethical decision-making, and scientific research innovation. Figure 2 provides a conceptual map of how AIGC applications map onto these 4 critical thinking dimensions, offering a visual summary of our analytical approach. By thoroughly examining the differentiated effects of AIGC applications on critical thinking within each dimension, the study aims to establish a robust theoretical foundation to inform targeted interventions and guide the development of educational policies.

In light of the pivotal transition occurring in medical education, this study seeks to address a fundamental question: how can the critical thinking skills of medical students be safeguarded and enhanced while simultaneously leveraging the capabilities of AIGC technology? The findings of this inquiry will have significant implications for the quality of future medical professionals and will also influence the preservation of the humanistic aspects of medicine amid the ongoing technological revolution.

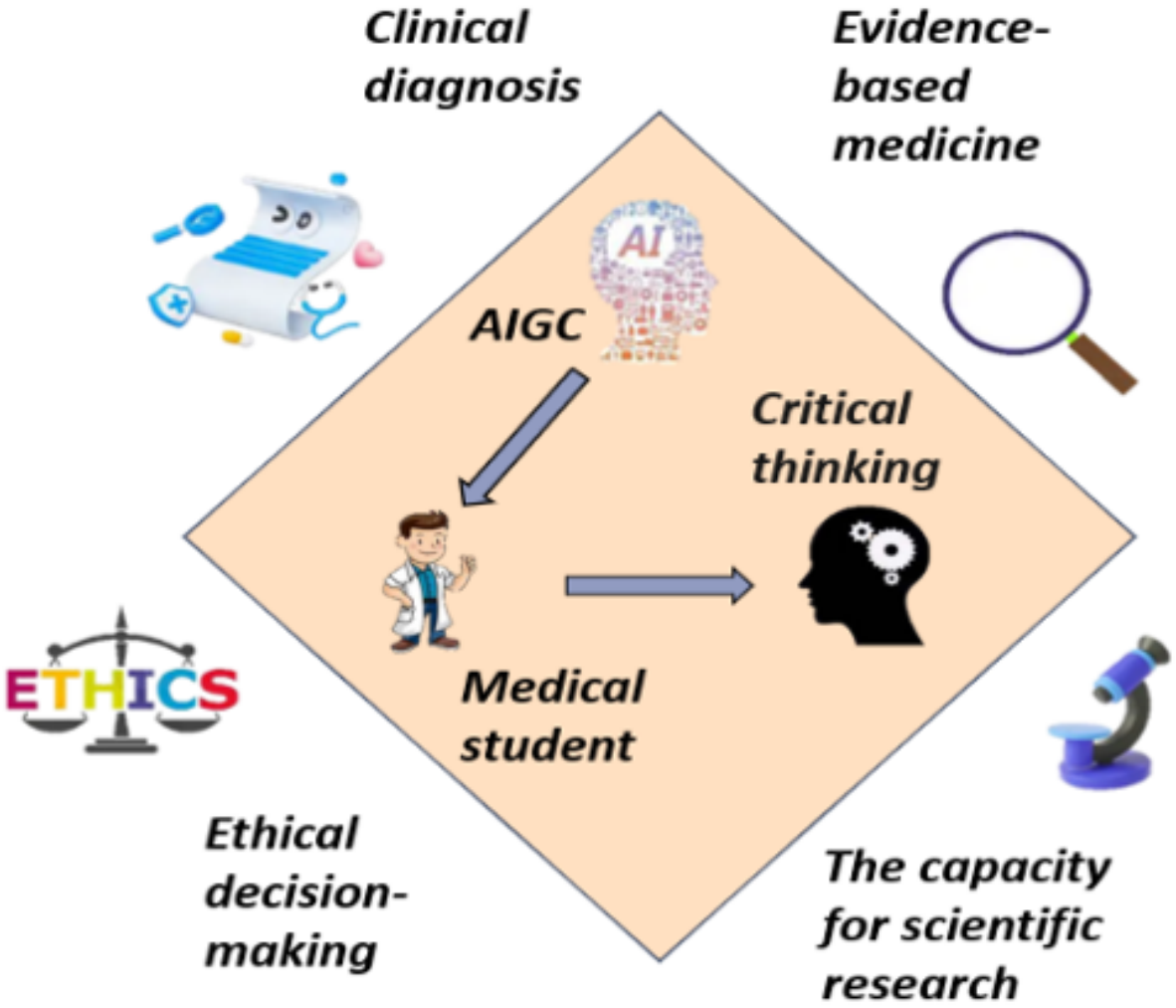
XSL•FO
RenderX

Figure 1. Complex mechanism of artificial intelligence (AI)–generated content (AIGC) on the critical thinking ability of medical students.



XSL•FO
RenderX

Figure 2. Application of artificial intelligence (AI)–generated content (AIGC) to the critical thinking of medical students.



Methods

PRISMA Flowchart-Based Literature Screening and Analysis

This systematic review aims to methodically identify, evaluate, and synthesize existing literature to elucidate the specific effects of AIGC on the critical thinking abilities of medical students. The review was conducted in strict accordance with the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) 2020 guidelines and used a systematic methodology to assess the influence of AIGC on medical students' critical thinking (Checklist 1). Additionally, the review protocol was registered with PROSPERO under the registration number CRD420251116204.

Retrieval Strategy and Information Sources

This study used a combined search strategy utilizing both controlled vocabulary (subject terms) and free-text terms (refer to Table 1 for specifics) to perform a comprehensive literature search within the PubMed database. The search focused on keywords including AIGC and its prominent applications such as LLMs, medical students, and critical thinking. To ensure the inclusion of the most current and relevant research, the search was restricted to studies published subsequent to the introduction of ChatGPT, thereby capturing the latest advancements in this domain.

XSL•FO
RenderX

Table . PubMed search strategy for identifying studies.

Search	Query
1	(“Medical Student” [All Fields] OR “Medical Students” [All Fields] OR “Student, Medical”[All Fields])
2	(“Generative Artificial Intelligence” [All Fields] OR “Artificial Intelligence, Generative” [All Fields] OR “Chatbot” [All Fields] OR “Chatbots” [All Fields] OR “Chat-GPT” [All Fields] OR “Chat GPT” [All Fields] OR “ChatGPT” [All Fields] OR “ChatGPTs” [All Fields])
3	(Critical Thinking OR Thinking, Critical OR Thinking Skills OR Thinking Skill OR Thought OR Thoughts)
4	1 AND 2 AND 3
5	4 Limits:y_5[Filter]

Research Screening and Inclusion and Exclusion Criteria

The inclusion criteria are relevant studies involving the impact of AIGC on the critical thinking of medical students. The exclusion criterion is that the full text of the literature cannot be obtained or it is not published in English.

Risk of Bias Assessment and Data Extraction

Considering the variations in research design types, this study used a tiered quality assessment approach. Specifically, the Cochrane RoB 2.0 tool was used to assess the risk of bias in randomized controlled trials. For nonrandomized intervention studies, the ROBINS-I framework was applied to conduct a systematic evaluation. For other study designs, including observational, qualitative, or mixed methods research, the JBI series of checklists—endorsed by international consensus and tailored to specific design types—was selected to ensure methodological rigor and comparability of evidence quality.

The Main Results of Literature Retrieval

Two reviewers (JW and ZC) independently screened the literature. First, they performed a preliminary screening of titles and abstracts to exclude obviously irrelevant studies. Then, they conducted a full-text review of the selected studies. This review focused on the relevance of AIGC to the development of critical thinking skills in medical students. Throughout the review, a comprehensive appraisal was conducted regarding the methodological rigor, the reliability of the findings, and the validity of the conclusions presented. Key themes pertaining to the application of AIGC in enhancing critical thinking among medical students were identified and extracted. In cases of disagreement between the 2 reviewers during the screening process, consensus was achieved through detailed discussion to maintain the objectivity and accuracy of the selection outcomes. A qualitative analysis was subsequently performed based on the recurring themes identified in the literature, with the objective of providing a robust and evidence-based foundation to inform the policy development and practical implementation of AI technologies within the academic medical domain.

Results

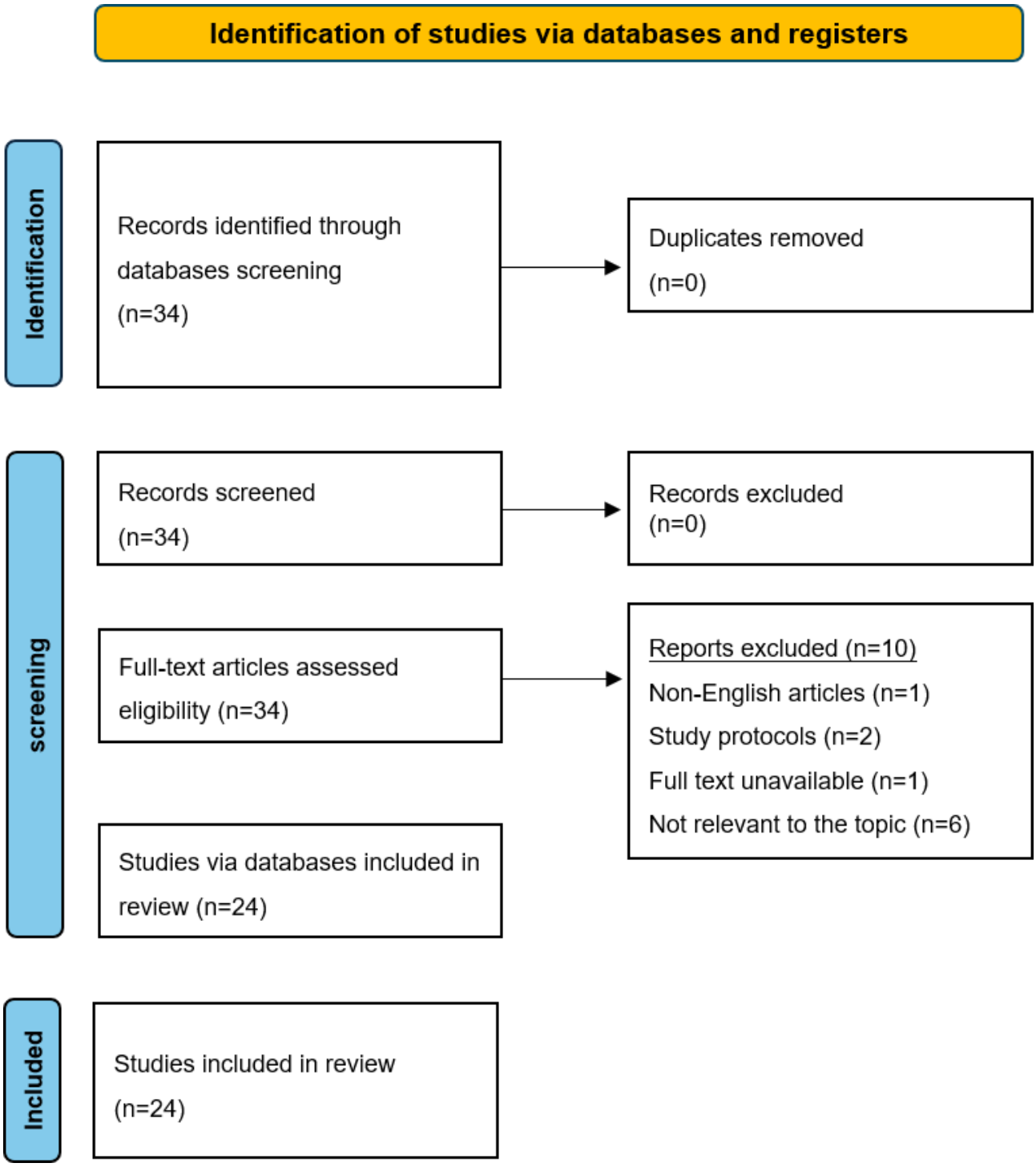
Overview

The literature selection process is summarized in the PRISMA flow diagram (Figure 3). The initial database search yielded 34 records. After removing 0 duplicates, 34 records underwent title and abstract screening. Following this, 34 full-text papers were assessed for eligibility, resulting in the final inclusion of 24 studies that met all criteria. The primary reasons for exclusion at the full-text stage were that the studies did not focus on medical students or lacked empirical data on critical thinking.

Table 2 shows the risk assessment and quality appraisal of the included studies, and Table 3 shows the key takeaways.

XSL•FO
RenderX

Figure 3. Flowchart of the system review. This flowchart shows the screening process of the included and excluded studies and marks the reasons for exclusion.



XSL•FO
RenderX

Table . Quality appraisal and risk of bias summary of the included studies.

Author and year	Type of research	Evaluation tool	Quality assessment by author JW	Quality assessment by author ZC	Rank
Ahmad et al [6] (2024)	Narrative review	SANRA[a]	6/12	7/12	Moderate quality
Jeyaraman et al [7] (2023)	Opinion review	SANRA	9/12	10/12	High quality
Prober and Desai [8] (2023)	Commentary	SANRA	10/12	9/12	High quality
Magalhães Araujo et al [9] (2024)	Mixed methods study	MMAT[b]	6/7	6/7	High quality
Xu et al [10] (2024)	Review article	SANRA	9/12	10/12	High quality
Madrid et al [11] (2025)	Nonrandomized interventional study	ROBINS-I[c]	Moderate risk	Moderate risk	Moderate quality
Kaliterna et al [12] (2024)	Cross-sectional studies	JBI	6/8	6/8	High quality
Abdullah et al [13] (2024)	Cross-sectional studies	JBI	6/8	5/8	Moderate quality
Moskovic and Rozani [14] (2025)	Mixed methods study	MMAT	6/7	6/7	High quality
Almazrou et al [15] (2024)	Cross-sectional studies	JBI	6/8	7/8	High quality
Abouammoh et al [16] (2025)	Qualitative study	JBI	8/10	8/10	High quality
Jiang et al [17] (2024)	RCT[d]	Cochrane RoB 1.0	Low risk	Low risk	High quality
Öncü et al [18] (2025)	Nonrandomized interventional study	ROBINS-I	High risk	High risk	Low quality
Shalong et al [19] (2024)	RCT	Cochrane RoB[e] 1.0	Low risk	Low risk	High quality
Reinke et al [20] (2025)	Observational study	JBI	Moderate quality	Moderate quality	Moderate quality
Elabd et al [21] (2025)	Tutorial	JBI	High quality	High quality	High quality
Wang et al [22] (2025)	Qualitative study	JBI	9/10	9/10	High quality
Gong et al [23] (2024)	Nonrandomized interventional study	ROBINS-I	Moderate risk	Moderate risk	Moderate risk
Montagna et al [24] (2025)	RCT	Cochrane RoB-1	Moderate risk	Moderate risk	Moderate risk
Mondal [25] (2025)	Cross-sectional studies	JBI	Moderate quality	Moderate quality	Moderate quality
Laupichler et al [26] (2024)	Nonrandomized interventional study	ROBINS-I	Moderate risk	Moderate risk	Moderate risk
Thomae et al [27] (2024)	Cross-sectional studies	JBI	Moderate quality	Moderate quality	Moderate quality
Mondal [28] (2025)	Expert opinion	JBI	High quality	High quality	High quality
Wu et al [29] (2024)	Expert opinion	JBI	High quality	High quality	High quality

[a]SANRA: Scale for the Assessment of Narrative Review Articles.

[b]MMAT: Mixed Methods Appraisal Tool.

[c]ROBINS-I: Risk of Bias in Non-Randomized Studies of Interventions.

[d]RCT: randomized controlled trial.

[e]RoB: Risk of Bias.

XSL•FO
RenderX

Table . Key takeaways of core studies.

Author and year	Key takeaway
Ahmad et al [6] (2024)	The dual impact of AI[a] on medical education and clinical practice was explored, with particular attention to potential risks in cognitive development of medical students.
Jeyaraman et al [7] (2023)	The potential and risk of ChatGPT in medicine, education, and scientific research are systematically reviewed, and the core viewpoint of "complementarity rather than substitution" is put forward.
Prober and Desai [8] (2023)	In view of the serious shortage of doctors in the United States, 3 major reform directions are proposed. Appeal to pay attention to the cultivation of critical thinking.
Magalhães Araujo et al [9] (2024)	Application of ChatGPT in medical informatics education.
Xu et al [10] (2024)	The potential impact of AI, especially generative AI (such as GPT-4), on medical education, residency training, and continuing education is discussed.
Madrid et al [11] (2025)	The application potential of LLM[b] in medical field is discussed, especially the performance of GPT series model in German medical license examination.
Kaliterna et al [12] (2024)	A study on AI's ability to generate ethical dilemma essays: comparing students' writing with AI-generated personal ethical dilemma reflection essays.
Abdullah et al [13] (2024)	Investigating ethical issues of ChatGPT in medical education.
Moskovic and Rozani [14] (2025)	The hybrid approach reveals students' real needs and fears for AI, beyond theoretical speculation.
Almazrou et al [15] (2024)	A cross-sectional survey of medical students' perceptions of ChatGPT on critical thinking skills.
Abouammoh et al [16] (2025)	This paper discusses the early cognition, experience, and potential challenge of ChatGPT among middle school teachers and students in medical education.
Jiang et al [17] (2024)	A randomized controlled trial was designed to explore the effect of chatbot-based standardized patient combined with case-based learning in colorectal cancer teaching.
Öncü et al [18] (2025)	This paper discusses the application of ChatGPT-4o as virtual standardized patient (SP) in medical education, focusing on the evaluation of clinical case management ability of interns.
Shalong et al [19] (2024)	Randomized controlled trial to assess the impact of LearnGuide on medical students to support autonomous learning among medical students.
Reinke et al [20] (2025)	The integration path of ChatGPT in medical informatics education was explored.
Elabd et al [21] (2025)	Using AI tools to generate the potential value of personalized multimodal memory in medical education, this paper explores the integration path of ChatGPT in medical informatics education.
Wang et al [22] (2025)	This paper discusses the willingness of Chinese medical students to use ChatGPT in programming courses and its influencing factors.
Gong et al [23] (2024)	The potential application of 2 large language models in gastroenterology was discussed.
Montagna et al [24] (2025)	Comparison of 3 types of clinical decision support systems in medical student case solving.
Mondal [25] (2025)	To investigate the use of question-answering resources by medical students in class. To compare the performance of 3 types of clinical decision support systems in medical students' case solving.
Laupichler et al [26] (2024)	Double-blind control design to compare the quality difference between ChatGPT-3.5 generated questions and artificial questions.

XSL•FO
RenderX

Author and year	Key takeaway
Thomae et al [27] (2024)	Three different modes of integration of ChatGPT in medical curricula were systematically explored, and a comprehensive hybrid approach assessment was conducted.
Mondal [28] (2025)	This paper discusses the opportunities and challenges of large language models (such as ChatGPT) in medical education and proposes a specific ethical framework for their use.
Wu et al [29] (2024)	The potential, application scenarios, and challenges of ChatGPT in medical education were discussed.

[a]AI: artificial intelligence.

[b]LLM: large language model.

Dual Impact of AIGC on Medical Students' Critical Thinking

By conducting a systematic literature review and synthesis, this study identifies a dual impact of AIGC on the critical thinking abilities of medical students. The subsequent analysis will comprehensively examine both the positive and negative effects of AIGC on medical students' critical thinking, with particular emphasis on 4 key dimensions: clinical diagnostic reasoning, evidence-based medical practice, ethical decision-making, and scientific research innovation. The objective is to offer theoretical insights and a practical framework to enhance the effective utilization of AIGC tools by medical students.

AIGC to the Critical Thinking Skills of Medical Students

Phenomenon of AI Syndrome

Ahmad et al [6] introduced the notion of "AI syndrome," cautioning that reliance on AIGC may result in "progressive cognitive decline" among medical students. The authors argue that students tend to favor the immediate responses offered by AIGC, which diminishes their capacity for independent thought and undermines the development of critical clinical reasoning skills. While these short-term efficiencies may be appealing, they could ultimately contribute to a deterioration of cognitive abilities, thereby jeopardizing clinical judgment and patient safety in the long run.

Mental Inertia

In 2023, Jeyaraman et al [7] highlighted that while AIGC demonstrates proficiency in memory-related tasks, it exhibits significant shortcomings in deep reasoning, challenging established knowledge and mitigating unwarranted self-confidence. These deficiencies may inadvertently promote a lack of diligence among students, leading them to neglect critical thinking. Concurrently, Prober and Desai [8] advocated for a transformation in medical education, suggesting a shift from prioritizing students with strong memorization skills to fostering essential competencies, such as critical thinking and empathy. This recommendation is underscored by the fact that AI has successfully passed the United States Medical Licensing Examination, indicating a diminishing emphasis on rote memorization. Subsequently, a mixed methods study conducted by Magalhães Araujo et al [9] in 2024 revealed that approximately 20% of students exhibited signs of cognitive inertia, relying heavily on cue words, thereby underscoring the potential adverse effects of AIGC on students' independent thinking capabilities. In the same year, Xu et al [10] explicitly cautioned against the dangers of excessive dependence on AIGC, warning that it could foster "intellectual laziness." They noted that the authoritative presentation of information generated by AIGC might lead students to accept such information uncritically, thereby undermining their critical thinking skills. In 2025, Madrid et al [11] further corroborated the limitations of AIGC regarding deep reasoning and the challenge of knowledge in their analysis of the German medical licensing examinations, emphasizing its detrimental effects on students' cognitive engagement.

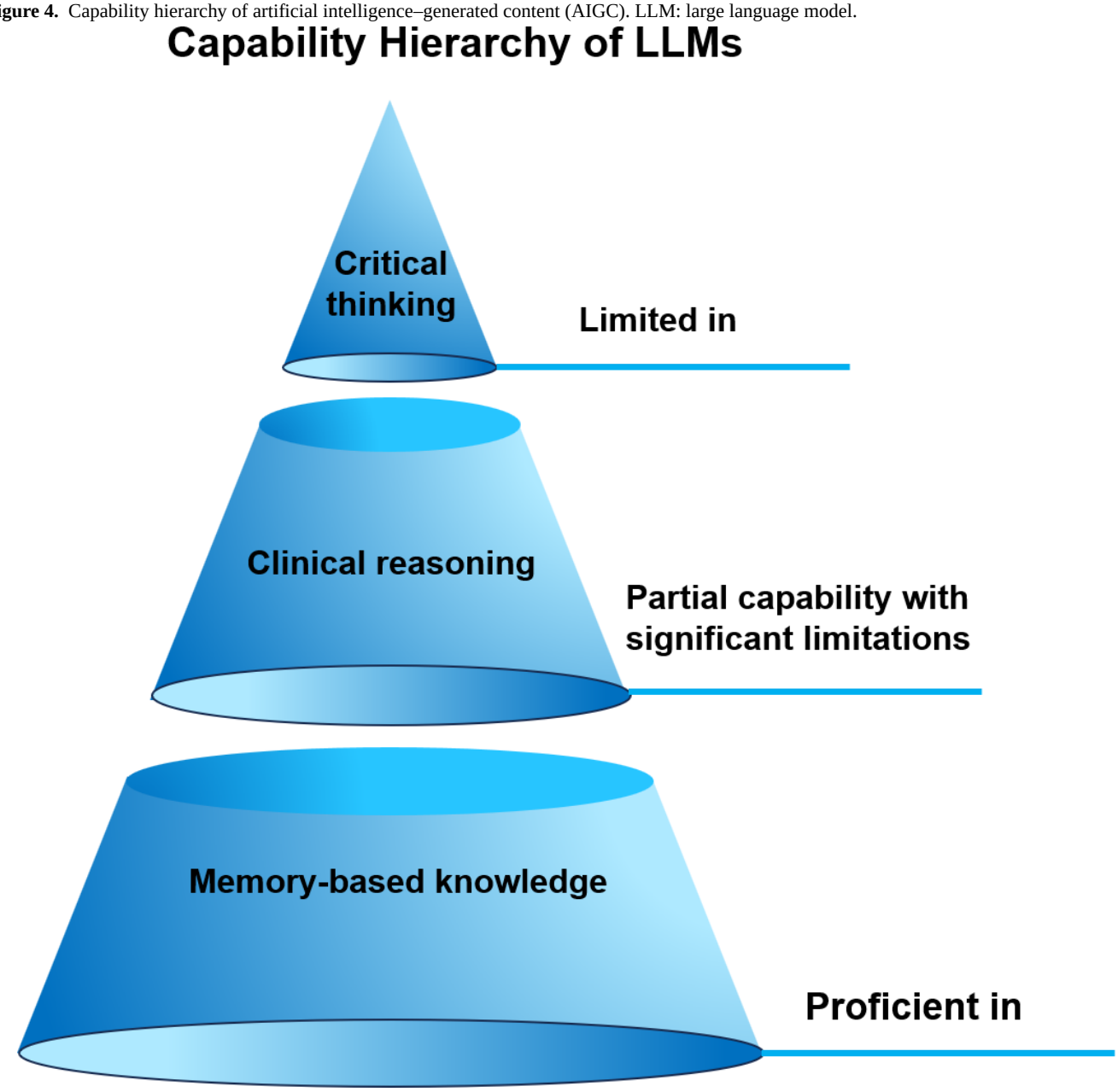
Academic Integrity

In 2023, Jeyaraman et al [7] were pioneers in identifying the potential challenges associated with AIGC within the realm of medical education, advocating for the creation of detection tools to mitigate the risks of AIGC misuse. Following this, a cross-sectional study conducted by Kaliterna et al [12] in 2024 further elucidated the gravity of the issue, revealing that as many as 34% of reflective papers authored by medical students were likely generated or altered by AI. This occurrence not only underscores the significant threat posed to academic integrity by the misuse of AIGC but also indicates a potential detrimental impact on students' independent thinking and writing skills.

These investigations illustrate that while AIGC offers notable advantages in the retrieval and application of memory-based knowledge, thereby providing efficient and convenient support for medical education, it also presents considerable limitations in fostering higher-order critical thinking. As visualized in Figure 4, AIGC's capabilities are predominantly situated at the lower levels of the cognitive hierarchy (eg, information retrieval and knowledge comprehension), with diminishing efficacy at higher levels requiring analysis, evaluation, and creation. This hierarchy effectively illustrates the inherent limitations of current AIGC tools in cultivating advanced critical thinking skills. Addressing these challenges necessitates a collaborative effort among educators, students, and policymakers to ensure that the integration of AIGC in medical education enhances, rather than impedes, the comprehensive development of students through appropriate guidance and effective regulation.

XSL•FO
RenderX

Figure 4. Capability hierarchy of artificial intelligence–generated content (AIGC). LLM: large language model.



Potential of AIGC in Fostering Critical Thinking

Student Perspectives

Numerous surveys indicate a nuanced and often contradictory perspective regarding AIGC. In a 2024 survey conducted by Abdullah et al [13], it was reported that 69% of medical students acknowledged the potential advantages of AIGC in fostering critical thinking, while 31% expressed concerns regarding its potential to induce mental inertia. A subsequent study by Moskovich and Rozani [14] in 2025 revealed a prevalent apprehension among medical students that AIGC may undermine critical thinking skills and promote complacency. Notably, a 2023 investigation by Almazrou et al [15] indicated that as many as 92.6% of medical students recognized the prospective benefits of AIGC in enhancing critical thinking. Furthermore, research by Abouammoh et al [16] suggested that students exhibited a more pragmatic outlook compared to educators, who tended to highlight "critical thinking deficits." These findings imply that while medical students generally perceive AIGC as a valuable tool for improving critical thinking, they simultaneously harbor concerns about the potential for mental inertia, thereby illustrating the dual opportunities and challenges presented by the integration of AIGC in medical education.

Standardized AI Patients

In recent years, the utilization of virtual standardized patients (SP) has emerged as an innovative tool in medical education, garnering increasing attention for its potential to enhance the critical thinking skills of medical students. A study conducted by Jiang et al [17] in 2024 revealed that the integration of virtual SP training with case-based learning significantly enhances students' systematic and evidence-based thinking abilities. However, the study noted that the short-term effects on higher-order critical thinking skills were not as pronounced, indicating that a more profound integration of virtual SP training is necessary to effectively foster the development of critical

XSL•FO
RenderX

thinking among medical students. Furthermore, in 2025, Öncü et al [18] highlighted the considerable advantages of AIGC, such as ChatGPT-4o, in terms of cost-effectiveness, accessibility, and standardization as a virtual standardized patient. This advancement presents new opportunities for the optimization and promotion of virtual SP training.

Nevertheless, from the perspective of critical thinking development, further investigation is required to explore how the benefits of AIGC can address the limitations of virtual SP training in enhancing higher-order thinking skills.

Customized AIGC Tools

In 2024, Shalong et al [19] conducted a randomized controlled trial that revealed the efficacy of a tailored GPT AI tool, which offers structured guidance rather than direct responses, in enhancing students' scores on the Cornell Critical Thinking Test over a 12-week period and beyond. This finding indicates that the development of critical thinking skills among medical students can be effectively facilitated through the thoughtful design of AIGC guidance tools. In 2025, Reinke et al [20] introduced a framework for "guided use," which positions AIGC as a critical thinking "partner" rather than a mere "substitute." Additionally, Elabd et al [21] employed AIGC, such as ChatGPT and DALL-E 3, to create personalized multimodal memory methods tailored to various learning styles, while emphasizing the importance of treating these tools as supplementary resources to mitigate the risk of over-reliance. Furthermore, Wang et al [22] advocated for the development of open-ended tasks, such as clinical data analysis, and the establishment of case libraries showcasing AIGC applications (including examples of both accurate and inaccurate uses) within medical education, particularly in programming courses, to cultivate students' critical utilization skills. Collectively, these studies and recommendations underscore the potential of customized AIGC tools in medical education to enhance critical thinking abilities and accommodate diverse student needs through personalized and varied learning approaches. However, the effectiveness of these tools is contingent upon their design and positioning to avert potential dependency risks. In contrast, Gong et al [23] examined the clinical applicability of customized LLMs in gastroenterology and found that the generic GPT-4o, equipped with retrieval-augmented generation capabilities, performed at an expert level, significantly surpassed the performance of customized models.

New Model of Resource Integration

Montagna et al [24] conducted a systematic comparison of the efficacy of 3 distinct types of clinical decision support systems: traditional clinical practice guidelines, online knowledge bases (such as UpToDate), and LLMs (ChatGPT) in the context of medical student case resolution. Their findings indicated that while ChatGPT provided the quickest responses, clinical practice guidelines demonstrated superior accuracy. Concurrently, Mondal [25] explored the potential influence of a novel model of resource integration on the critical thinking abilities of medical students. Their survey revealed that medical students frequently utilize AI, particularly LLMs, to rapidly construct knowledge frameworks, subsequently validating detailed information through authoritative databases, such as PubMed. This hybrid approach, termed "LLM (Large Language Model) + Search Engine + Authoritative Database," not only facilitates the efficient integration of multisource information but also fosters the development of students' "multiresource integration ability." This capability is essential for addressing potential obsolescence or inaccuracies in AIGC-generated information, while also empowering students to maintain critical thinking when confronted with diverse information sources, thereby enhancing their skills in identifying and verifying the accuracy and reliability of information. This innovative model offers medical students a more holistic educational pathway, ultimately aiding in the enhancement of their critical thinking and decision-making competencies within complex informational contexts.

Identify Limitations and Proactively Assess

Laupichler et al [26] conducted a comparative analysis of the quality of questions generated by ChatGPT-3.5 versus those created artificially, revealing that AI-generated questions were more susceptible to being "guessed," while artificially constructed questions were more effective in accurately assessing students' true capabilities. Thomae et al [27] further illustrated the benefits of engaging students in the design of solutions and subsequent comparisons with AI-generated alternatives within the same academic year, particularly in the context of a "placebo control condition" unit. This comparative exercise not only enables students to visually discern the strengths and weaknesses of AI but also promotes critical thinking through facilitated discussions. Additionally, Mondal [28] emphasized the necessity for students to actively articulate their evaluations of AI-generated information and engage in discourse regarding its potential inaccuracies and biases. This practice is not only fundamental to the development of critical thinking skills but also essential for navigating the limitations of AI technology. Collectively, these studies underscore that through the processes of comparison, evaluation, and discussion of AI-generated content, medical students can enhance their critical thinking abilities, thereby enabling them to apply AI technology more judiciously in their future medical practices.

In summary, AIGC functions simultaneously as an enhancer and a potential inhibitor of critical thinking among medical students. When utilized appropriately, AIGC can evolve into an effective "practice partner." However, improper use may lead to adverse outcomes, such as AIGC dependency syndrome, cognitive stagnation, and a deterioration of academic integrity. Addressing these challenges requires an integrated approach comprising standardized AIGC patient simulations, personalized instructional guidance, multifaceted evaluation methods, and proactive assessment strategies.

Synopsis of Capabilities, Quick Check Applications, and Critical Thinking Hazards for Medical Students

This study systematically synthesized the dispersed findings from the 24 reviewed papers and developed a "risk-return matrix" (Table 4). The matrix is structured with the typical cognitive processes of medical students utilizing generative AIGC represented along the vertical axis: data input (dimensionality), information retrieval (application), standardization compliance (standardization), knowledge

XSL•FO
RenderX

integration (knowledge), and deep clinical reasoning (deep thinking). The horizontal axis delineates the appealing "quick-check" functionalities available at each stage, including second-level differential diagnosis, automated extraction of key literature points, and collaborative draft decision-making between physicians and patients.

Table . Application of artificial intelligence–generated content (AIGC) to the critical thinking of medical students.

Dimensionality	Clinical diagnosis	Evidence-based medicine	Ethical decision-making	The capacity for scientific research
Application scenarios	• Enter virtual patient symptoms, practice history acquisition, and differential diagnosis • AI[a] inquiry synchronous prompt differential diagnosis, reduce missed diagnosis	• Quick search of the latest literature and treatment guidelines • Quick search assisted retrieval of textbook standardization knowledge • Quick Check Scholar AI plugin to extract key points from paper points	• Assisting doctor-patient joint decision making	• Assisted thesis writing (AI contribution needs to be marked transparently)
Advantages of rapid inspection technology	• Rapid detection reduces the likelihood of missed diagnosis	• Rapid inspection provides the latest evidence-based evidence • Rapid Search to improve knowledge retrieval efficiency	• A proposal for a standardized ethical decision framework for rapid inspection	• Quick reference generation (accuracy needs to be verified)
Rapid detection of potential risks	• Rapid inspection—students rely too much on AI "quick solutions," weakening deep thinking ability • Rapid detection—lack of practical experience in differential diagnosis, resulting in insufficient emergency response capability • Quick detection—AI generates error messages (eg, "hallucination" problem) • Quick check—unprofessional questions lead to misleading answers	• Quick check—lack of references to trace sources • Quick check—answers may be oversimplified or knowledge fragmented • Quick check—authoritative intonation covers errors	• Quick checks risk oversimplification; they promote solutions that are morally idealistic but practically unworkable • Rapid detection—social biases (race, gender) in training data are reinforced • Quick check—Patient Privacy breach (HIPAA[b] or GDPR[c] violations)	• Quick detection—generating false or erroneous references • Quick detection—promoting plagiarism • Quick check—marginal medical views are overwritten by mainstream content
Quick check deep challenge	• Rapid examination—LLM[d] fails to mimic human doctors' comprehensive reasoning abilities • Quick examination—unprofessional questioning by students leads to misdiagnosis risk	• Quick check—relying on open access databases, information may not be comprehensive • Challenge: Nonexperts' difficulty in verifying fast-generated outputs	• Rapid inspection—difficulty in clarifying AI's responsibility boundaries in medicine • Fast check data entry compliance issues	• Fast detection—initial data bias is reinforced by loop learning • Quick inspection-expert cross-validation offsets efficiency gains

[a]AI: artificial intelligence.

[b]HIPAA: Health Insurance Portability and Accountability Act.

[c]GDPR: General Data Protection Regulation.

[d]LLM: large language model.

Adjacent to each function on the right side, the matrix presents the "rapid-detection" outcomes derived from the systematic review. This comparative framework enables readers to promptly discern the specific contexts in which AIGC effectively enhances efficiency, as well as identify the stages where it may inadvertently undermine the critical thinking and clinical judgment skills of medical students.

XSL•FO
RenderX

Table 4 delineates the specific application scenarios of AIGC across various dimensions, including data input, information retrieval, knowledge integration, and deep reasoning. Examples encompass virtual patient symptom input, rapid literature retrieval, auxiliary differential diagnosis, and support in academic writing. Concurrently, the table highlights potential risks associated with these applications, such as students' overreliance on AI-generated quick responses, which may contribute to cognitive decline, diminished capacity for deep analytical thinking, and challenges arising from errors or biases inherent in AIGC-produced content. The discussion underscores the necessity for the judicious use of AIGC within medical education, urging educators and learners to balance efficiency gains with vigilance regarding the possible adverse effects on the cultivation of medical students' critical thinking skills and professional competencies.

Discussion

Principal Findings

The integration of AIGC technology into medical education exhibits a dual nature. As outlined in Table 4, which summarizes the 4 dimensions of critical thinking as applied by AIGC to medical students, this technology offers substantial benefits. On the one hand, AIGC enhances the educational experience by providing an abundance of learning resources and efficient tools, leveraging its advanced capabilities in information generation and processing. This facilitates a more rapid acquisition of knowledge and expands the scope of learning for medical students. Conversely, excessive reliance on AIGC may lead to cognitive inertia, a decline in critical thinking skills, and potential issues related to academic integrity. Nevertheless, through the strategic optimization of AIGC tool design and implementation, alongside appropriate guidance and oversight from educators, AIGC can serve as a valuable asset in fostering critical thinking among medical students. The research indicates that approaches such as the development of customized AIGC tools, the use of virtual standardized patients, innovative models for resource integration, and proactive assessments of AIGC limitations can effectively address the deficiencies of AIGC in cultivating advanced critical thinking skills. These strategies can assist medical students in sustaining their critical thinking abilities throughout the learning process, thereby strengthening their analytical and problem-solving capabilities.

Limitations

While we made every effort to comprehensively retrieve relevant literature using the PRISMA framework, our research predominantly centered on ChatGPT's applications. Future research should expand to other generative AI models to holistically evaluate their potential and differences in medical education. Currently, there is a scarcity of comparative studies on different models within the academic community, which restricts the in-depth analysis of their similarities and differences. To address this, future research could employ cross-model comparative experiments to explore each model's unique advantages in cultivating critical thinking.

The study systematically reviewed literature to examine how AIGC technology affects medical students' critical thinking and proposed coping strategy frameworks. However, several limitations might affect the findings' comprehensiveness and the conclusions' generalizability. The literature search was limited to studies published between November 2022 and June 2025. Although this captures recent AIGC advancements in medical education, it may miss earlier valuable research and post-June 2025 studies that could reveal significant trends.

Despite evaluation by 2 independent reviewers to reduce bias, some subjectivity may still exist in assessing studies' relevance and quality. While disagreements were resolved through discussion, they could still somewhat influence the objectivity. The reviewed studies showed marked variability in design, sample size, participant characteristics, and assessment tools. Although this diversity enriches the research, it also causes result heterogeneity, compromising conclusion consistency and generalizability.

As AIGC technology rapidly evolves, the referenced literature and case studies may not fully reflect the latest developments. While AIGC excels in knowledge generation and information retrieval, concerns remain about the accuracy and reliability of its content. AI models might be limited by training data biases, algorithmic limitations, and language comprehension challenges, risking the generation of incorrect or misleading information. This could negatively impact medical students' critical thinking, yet research on this issue is still inadequate.

Moreover, this study mainly focused on AIGC's short-term effects on medical students' critical thinking. As critical thinking development is a long-term process, further research is needed to investigate AIGC's long-term effects, such as its potential to cause cognitive inertia or dependency. Additionally, due to potential methodological limitations in some included studies (eg, small sample sizes and lack of control groups), the study's conclusions require further verification based on higher-quality evidence.

Outlook

The future design of AIGC tools should prioritize customization and personalization tailored to the learning styles, knowledge levels, and educational needs of medical students. This approach entails the provision of targeted learning resources and guidance methods. By utilizing intelligent recommendation systems and adaptive learning platforms, AIGC can assist students in enhancing their critical thinking skills. Medical students are encouraged to implement a hybrid learning strategy that integrates LLMs, search engines, and authoritative databases. This strategy fosters the ability to critically evaluate diverse information sources and synthesize multiple resources effectively. Such competencies not only improve students' information literacy but also bolster their decision-making capabilities in complex informational contexts. Furthermore, medical education should emphasize the development of AI literacy among students, enabling them to critically assess the advantages and limitations of AIGC technology and use AI tools judiciously in clinical practice. Through education in AI ethics and critical thinking, students can cultivate appropriate technical concepts and professional attributes.

XSL•FO
RenderX

Funding

In 2024, funding was provided by the Higher Education Steering Committee of the Ministry of Education for the project titled "Evidence-Based Translational Medicine," under project number B_YXC2024-01-01_10. In 2025, the Health Commission of Hubei Province funded the project named "AIGC-Driven Statistical Thinking to Create an Integrated Professional Postgraduate Training Model for Scientific Research-Clinical Transformation," with project number HBJG-250041.

Conflicts of Interest

None declared.

Checklist 1
PRISMA-LSR checklist.
[DOCX File, 27 KB - mededu_v12i1e79939_app1.docx]

References

1. Sharples JM, Oxman AD, Mahtani KR, et al. Critical thinking in healthcare and education. BMJ 2017 May 16;357:j2234. [doi: 10.1136/bmj.j2234] [Medline: 28512135]
2. Pavlov CS, Kovalevskaya VI, Varganova DL, et al. Critical thinking in medical education [Article in Russian]. Cardiovasc Ther Prev 2023;22(2S):3566. [doi: 10.15829/1728-8800-2023-3566]
3. Silva AOVD, Carvalho ALRFD, Vieira RM, Pinto CMCB. Clinical supervision strategies, learning, and critical thinking of nursing students. Rev Bras Enferm 2023;76(4):e20220691. [doi: 10.1590/0034-7167-2022-0691] [Medline: 37820150]
4. Gencer G, Gencer K. A comparative analysis of ChatGPT and medical faculty graduates in medical specialization exams: uncovering the potential of artificial intelligence in medical education. Cureus 2024 Aug;16(8):e66517. [doi: 10.7759/cureus.66517] [Medline: 39246999]
5. Seetharaman R. Revolutionizing medical education: can ChatGPT boost subjective learning and expression? J Med Syst 2023 May 9;47(1):61. [doi: 10.1007/s10916-023-01957-w] [Medline: 37160568]
6. Ahmad O, Maliha H, Ahmed I. AI syndrome: an intellectual asset for students or a progressive cognitive decline. Asian J Psychiatr 2024 Apr;94:103969. [doi: 10.1016/j.ajp.2024.103969] [Medline: 38387116]
7. Jeyaraman M, Ramasubramanian S, Balaji S, Jeyaraman N, Nallakumarasamy A, Sharma S. ChatGPT in action: harnessing artificial intelligence potential and addressing ethical challenges in medicine, education, and scientific research. World J Methodol 2023 Sep 20;13(4):170-178. [doi: 10.5662/wjm.v13.i4.170] [Medline: 37771867]
8. Prober CG, Desai SV. Medical school admissions: focusing on producing a physician workforce that addresses the needs of the United States. Acad Med 2023 Sep 1;98(9):983-986. [doi: 10.1097/ACM.0000000000005262] [Medline: 37130009]
9. Magalhães Araujo S, Cruz-Correia R. Incorporating ChatGPT in medical informatics education: mixed methods study on student perceptions and experiential integration proposals. JMIR Med Educ 2024 Mar 20;10:e51151. [doi: 10.2196/51151] [Medline: 38506920]
10. Xu Y, Jiang Z, Ting DSW, et al. Medical education and physician training in the era of artificial intelligence. Singapore Med J 2024 Mar 1;65(3):159-166. [doi: 10.4103/singaporemedj.SMJ-2023-203] [Medline: 38527300]
11. Madrid J, Diehl P, Selig M, et al. Performance of plug-in augmented ChatGPT and its ability to quantify uncertainty: simulation study on the German medical board examination. JMIR Med Educ 2025 Mar 21;11:e58375. [doi: 10.2196/58375] [Medline: 40116759]
12. Kaliterna M, Žuljević MF, Ursić L, Krka J, Duplančić D. Testing the capacity of Bard and ChatGPT for writing essays on ethical dilemmas: a cross-sectional study. Sci Rep 2024 Oct 30;14(1):26046. [doi: 10.1038/s41598-024-77576-3] [Medline: 39472544]
13. Abdullah HMA, Naeem NIK, Malkana GA, Javed M, Shahzaib MM. Exploring the ethical implications of ChatGPT in medical education: privacy, accuracy, and professional integrity in a cross-sectional survey. Cureus 2024 Dec;16(12):e75895. [doi: 10.7759/cureus.75895] [Medline: 39822443]
14. Moskovich L, Rozani V. Health profession students' perceptions of ChatGPT in healthcare and education: insights from a mixed-methods study. BMC Med Educ 2025 Jan 21;25(1):98. [doi: 10.1186/s12909-025-06702-0] [Medline: 39833868]
15. Almazrou S, Alanezi F, Almutairi SA, et al. Enhancing medical students critical thinking skills through ChatGPT: an empirical study with medical students. Nutr Health 2025 Sep;31(3):1023-1033. [doi: 10.1177/02601060241273627] [Medline: 39150341]
16. Abouammoh N, Alhasan K, Aljamaan F, et al. Perceptions and earliest experiences of medical students and faculty with ChatGPT in medical education: qualitative study. JMIR Med Educ 2025 Feb 20;11:e63400. [doi: 10.2196/63400] [Medline: 39977012]

XSL•FO
RenderX

17. Jiang Y, Fu X, Wang J, et al. Enhancing medical education with chatbots: a randomized controlled trial on standardized patients for colorectal cancer. BMC Med Educ 2024 Dec 20;24(1):1511. [doi: 10.1186/s12909-024-06530-8] [Medline: 39707245]
18. Öncü S, Torun F, Ülkü HH. AI-powered standardised patients: evaluating ChatGPT-4o's impact on clinical case management in intern physicians. BMC Med Educ 2025 Feb 20;25(1):278. [doi: 10.1186/s12909-025-06877-6] [Medline: 39979969]
19. Shalong W, Yi Z, Bin Z, et al. Enhancing self-directed learning with custom GPT AI facilitation among medical students: a randomized controlled trial. Med Teach 2025 Jul;47(7):1126-1133. [doi: 10.1080/0142159X.2024.2413023] [Medline: 39425996]
20. Reinke NB, Parkinson AL, Kafer GR. A tutorial activity for students to experience generative artificial intelligence: students' perceptions and actions. Adv Physiol Educ 2025 Jun 1;49(2):461-470. [doi: 10.1152/advan.00245.2024] [Medline: 40105755]
21. Elabd N, Rahman ZM, Abu Alinnin SI, Jahan S, Campos LA, Baltatu OC. Designing personalized multimodal mnemonics with AI: a medical student's implementation tutorial. JMIR Med Educ 2025 May 8;11:e67926. [doi: 10.2196/67926] [Medline: 40341190]
22. Wang C, Xiao C, Zhang X, et al. Exploring medical students' intention to use of ChatGPT from a programming course: a grounded theory study in China. BMC Med Educ 2025 Feb 8;25(1):209. [doi: 10.1186/s12909-025-06807-6] [Medline: 39923098]
23. Gong EJ, Bang CS, Lee JJ, et al. The potential clinical utility of the customized large language model in gastroenterology: a pilot study. Bioengineering (Basel) 2024 Dec 24;12(1):1. [doi: 10.3390/bioengineering12010001] [Medline: 39851275]
24. Montagna M, Chiabrando F, De Lorenzo R, Rovere Querini P, Medical Students. Impact of clinical decision support systems on medical students' case-solving performance: comparison study with a focus group. JMIR Med Educ 2025 Mar 18;11:e55709. [doi: 10.2196/55709] [Medline: 40101183]
25. Mondal H. Evolving resource use for self-directed learning in physiology among first-year medical students in a classroom setting. Adv Physiol Educ 2025 Jun 1;49(2):394-397. [doi: 10.1152/advan.00236.2024] [Medline: 39993016]
26. Laupichler MC, Rother JF, Grunwald Kadow IC, Ahmadi S, Raupach T. Large language models in medical education: comparing ChatGPT- to human-generated exam questions. Acad Med 2024 May 1;99(5):508-512. [doi: 10.1097/ACM.0000000000005626] [Medline: 38166323]
27. Thomae AV, Witt CM, Barth J. Integration of ChatGPT into a course for medical students: explorative study on teaching scenarios, students' perception, and applications. JMIR Med Educ 2024 Aug 22;10:e50545. [doi: 10.2196/50545] [Medline: 39177012]
28. Mondal H. Ethical engagement with artificial intelligence in medical education. Adv Physiol Educ 2025 Mar 1;49(1):163-165. [doi: 10.1152/advan.00188.2024] [Medline: 39749996]
29. Wu Y, Zheng Y, Feng B, Yang Y, Kang K, Zhao A. Embracing ChatGPT for medical education: exploring its impact on doctors and medical students. JMIR Med Educ 2024 Apr 10;10:e52483. [doi: 10.2196/52483] [Medline: 38598263]

Abbreviations

AI: artificial intelligence
AIGC: artificial intelligence–generated content
LLM: large language model
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
SP: standardized patient

Edited by T Gladman; submitted 01.Jul.2025; peer-reviewed by A Mohebbi, JW Browning, K Yusif-zada; revised version received 08.Dec.2025; accepted 14.Dec.2025; published 27.Feb.2026.

Please cite as:
Li J, Ai F, Wang J, Cheng B, Li Y, Chen Z
Application of AI-Generated Content in Medical Education: Systematic Review of the Impact on Critical Thinking Abilities of Medical Students
JMIR Med Educ 2026;12:e79939
URL: https://mededu.jmir.org/2026/1/e79939
doi:10.2196/79939

© Jinlei Li, Fen Ai, Jueyan Wang, Bingxin Cheng, Yu Li, Zhen Chen. Originally published in JMIR Medical Education (https://mededu.jmir.org), 27.Feb.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Education, is properly cited. The complete bibliographic

XSL•FO
RenderX

information, a link to the original publication on https://mededu.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Educational Formats and Content Domains of Interprofessional Education for Licensed Rehabilitation Professionals: Scoping Review

Kohei Ikeda[1*], PhD; Takao Kaneko[2*], MS; Someka Hijikuro[3*], BS; Natsuki Inoue[3*], MS; Takuto Nakamura[1*], PhD; Taisei Takeda[3*], MS; Junya Uchida[4*], BS; Hirofumi Nagayama[1*], PhD

[1]Division of Occupational Therapy Program, School of Rehabilitation, Faculty of Health and Social Service, Kanagawa University of Human Services, 1-10-1 Heiseicho, Yokosuka, Japan

[2]Department of Rehabilitation, Yamagata Prefectural Central Hospital, Yamagata, Japan

[3]Department of Rehabilitation Therapy, Saiseikai Higashikanagawa Rehabilitation Hospital, Yokohama, Japan

[4]Dalla Lana School of Public School, University of Toronto, Toronto, ON, Canada

[*]all authors contributed equally

Corresponding Author:
Kohei Ikeda, PhD
Division of Occupational Therapy Program, School of Rehabilitation, Faculty of Health and Social Service, Kanagawa University of Human Services, 1-10-1 Heiseicho, Yokosuka, Japan

Abstract

Background: Interprofessional education (IPE) is a key strategy for enhancing collaboration and patient safety. While evidence for student populations is abundant, studies focusing on licensed physical therapists (PTs), occupational therapists (OTs), and speech-language pathologists (SLPs) remain limited. In contemporary rehabilitation practice, continuous IPE is increasingly important to address professional burnout and the growing complexity of patient needs.

Objective: This scoping review aimed to systematically map and synthesize the educational formats, content domains, and reported outcomes of IPE programs specifically targeting licensed PTs, OTs, and SLPs.

Methods: Following Joanna Briggs Institute and PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews) guidelines, we searched the PubMed, Web of Science, Cumulative Index to Nursing and Allied Health Literature, and Educational Resources Information Center databases through December 31, 2025. The eligibility criteria were based on the population, concept, and context framework, including peer-reviewed, English-language studies of licensed PTs, OTs, and SLPs (population) participating in structured IPE interventions (concept) within clinical or community settings (context). Studies focusing solely on students or prelicensure trainees were excluded. Following the screening of 3234 records by independent pairs of reviewers, 9 studies were ultimately selected for inclusion. Methodological quality was appraised using Joanna Briggs Institute critical appraisal checklists and the Mixed Methods Appraisal Tool. Data were synthesized using an evidence gap map to visualize research density across domains relative to established competency frameworks.

Results: A total of 9 studies from Australia, the United States, Canada, and the Philippines were included, with sample sizes ranging from 8 to 197. Most used single-group pre-post or mixed methods designs; notably, no randomized controlled trials were identified. Methodological quality was generally high, though limited by the lack of control groups. Systematic mapping identified 7 educational formats, with lectures and discussions being the most dominant across all competency domains. Primary content domains included communication and role clarification. Specific successful interventions included pharmacist-led medication safety workshops and the Kawa model for team building. While participants reported immediate improvements in role understanding and collaborative confidence, simulation-based training showed inconsistent effects on long-term clinical behavior. A substantial evidence gap was identified in experiential learning approaches targeting collaborative leadership.

Conclusions: This scoping review adds a novel perspective by focusing exclusively on licensed rehabilitation professionals (PTs, OTs, and SLPs), highlighting learning needs distinct from those of prelicensure students. It brings to the field a clearer understanding of a potential “leadership gap” and the current overreliance on didactic methods for experienced clinicians. Real-world implications suggest the need for health care institutions to transition toward systematic, practice-integrated IPE models that incorporate objective behavioral assessments. By addressing identified gaps in collaborative leadership and team functioning through longitudinal programs, health care institutions may contribute to more resilient team cultures, ultimately improving patient safety and the quality of rehabilitation care.

Trial Registration: Open Science Framework 10.17605/osf.io/25utv; https://osf.io/25utv/overview

XSL•FO
RenderX

(JMIR Med Educ 2026;12:e76189) doi:10.2196/76189

KEYWORDS

educational content domains; interprofessional education; licensed rehabilitation professionals; educational formats; scoping review

Introduction

Patient safety remains a global priority in health care systems worldwide. The World Health Organization estimates that approximately 1 in 10 patients receiving medical care experiences an adverse event, more than half of which are considered preventable [1]. Communication failures among health care professionals, as well as between professionals and patients or families, are consistently identified as a major contributing factor, accounting for nearly 70% of adverse events [1,2]. Consequently, improving interprofessional communication and collaboration has been recognized as a key strategy for enhancing patient safety and quality of care [2-4].

Interprofessional education (IPE) has been promoted as an effective educational approach to address these challenges. The World Health Organization [1] defines IPE as occasions when "two or more professions learn about, from, and with each other to enable effective collaboration and improve health outcomes." A growing body of literature demonstrates that IPE can improve communication skills, mutual understanding of professional roles, collaborative attitudes, and teamwork among health care students and trainees [5-8]. IPE also deepens understanding of the expertise and responsibilities of other health professionals [5,6,9]. In addition, previous studies have reported positive downstream effects of IPE on clinical practice, including improved teamwork in intensive care settings, reduced fall risk in patients, and more efficient patient transfers [7].

Within rehabilitation settings, collaboration among physical therapists (PTs), occupational therapists (OTs), and speech-language pathologists (SLPs) is particularly critical, as patient care often requires coordinated, goal-oriented interventions across disciplines. Interprofessional collaboration (IPC) in rehabilitation has been associated with improved functional outcomes, enhanced patient safety, and greater continuity of care [10-12]. Importantly, IPE for licensed rehabilitation professionals differs from student-based IPE, in that it emphasizes experiential learning grounded in clinical practice, professional identity, and role negotiation within real-world care environments [13]. Recent literature further emphasizes that while prelicensure IPE focuses on fundamental attitudinal changes—such as fostering mutual respect and role awareness among students—postlicensure IPE for practicing clinicians is a vital tool for addressing contemporary challenges, such as the increase in clinician burnout and the rising complexity of geriatric and neurological rehabilitation [14,15]. Furthermore, updated international frameworks, such as the 2023 interprofessional education collaborative core competencies, now explicitly highlight the need for practice-integrated learning to foster a resilient team culture and improve staff retention in high-pressure clinical environments [16,17]. Despite its recognized importance, evidence suggests that structured IPC remains insufficiently implemented in rehabilitation practice. A qualitative study conducted in an Irish rehabilitation hospital reported that many teams identifying as interdisciplinary lacked regular opportunities for shared goal-setting and reflective dialog [18]. Similarly, a Norwegian national survey revealed substantial variability in team communication and coordination, indicating that consistent interprofessional practices are not yet well established across rehabilitation settings [19]. The need for a more robust evidence base is underscored by the current global shortage of rehabilitation professionals, which demands more efficient, collaborative delivery models to maintain care quality. These findings highlight the persistent gap between the conceptual ideal of interprofessional teamwork and its practical realization in clinical rehabilitation.

Although interest in IPE has increased over the past decade, most existing research has focused on prelicensure health care students. Evidence regarding IPE for licensed rehabilitation professionals remains limited and fragmented. Previous reviews by Spaulding et al [8] and Shuyi et al [7] identified only a small number of studies targeting practicing clinicians and did not comprehensively examine how such programs are structured, what educational content they include, or how outcomes are evaluated. Furthermore, few reviews have used integrative visualization tools, such as evidence gap maps, to pinpoint specific intersections between educational formats and competency domains in this specific professional cohort. Given the diversity of professional roles, clinical settings, and educational approaches within rehabilitation, a broad mapping of existing evidence is warranted.

Therefore, this scoping review aimed to systematically map and synthesize the existing literature on IPE for licensed PTs, OTs, and SLPs. Specifically, this review sought to address the following research questions:

1. What are the structures (eg, target group, format, frequency, or duration) of IPE programs for licensed PTs, OTs, and SLPs?
2. What educational content (eg, themes, skills, or learning objectives) is addressed in these programs?
3. How are these structures and contents evaluated, and what impacts have been reported?

By clarifying current practices and evidence gaps, this review seeks to inform the design of future IPE initiatives and contribute to strengthening IPC in rehabilitation settings.

Methods

Design

This scoping review was conducted in accordance with the methodological framework originally proposed by Arksey and O'Malley [20] and subsequently refined by Levac et al and the Joanna Briggs Institute (JBI) [21,22]. This approach is

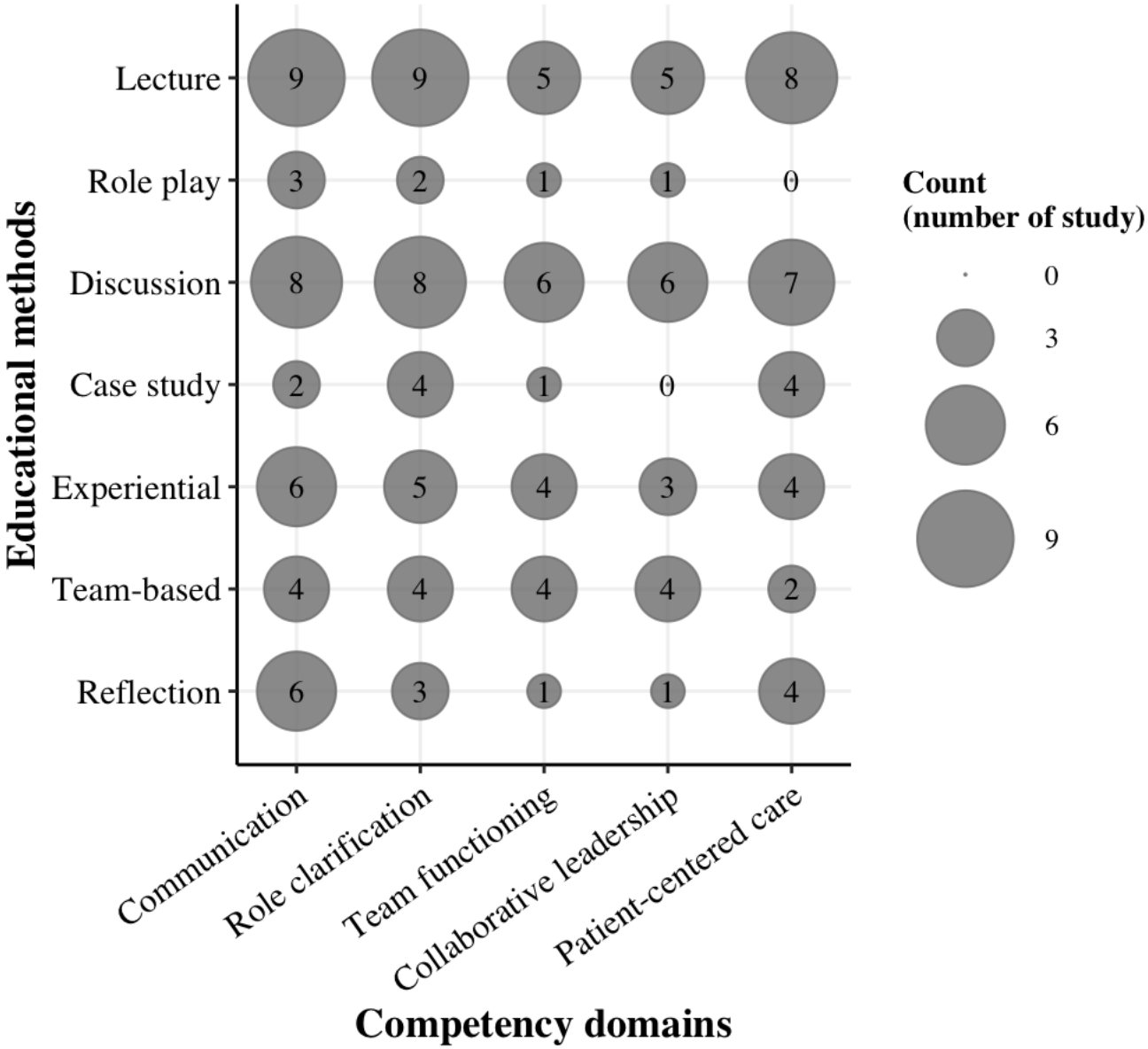
XSL•FO
RenderX

well-suited for mapping the extent, range, and nature of evidence across a broad topic area; identifying research gaps; and clarifying key concepts. Given the exploratory aim of this study—to examine the implementation and educational content of IPE for licensed PTs, OTs, and SLPs—a scoping review was deemed the most appropriate method.

The reporting of this review followed the PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews) guidelines [23,24] (Checklist 1; 2). In addition, the search strategy and reporting of the literature search process adhered to the PRISMA-S (Preferred Reporting Items for Systematic Reviews and Meta-Analyses Search Reporting Extension) guidelines to ensure transparency, reproducibility, and comprehensive documentation of all search components, including databases searched, search strategies, and limits applied [25] (Checklist 3).

The review protocol was prospectively registered on the Open Science Framework (DOI: 10.17605/OSF.IO/G5HQN). Furthermore, a preprint of this review has been made publicly available (DOI: 10.2196/preprints.76189) to enhance transparency, promote early dissemination, and allow for public scrutiny of the methodology. There were minor deviations from the original study protocol registered on Open Science Framework. Specifically, while the protocol initially planned to present findings primarily through descriptive tables, we expanded the visualization to include an evidence gap map (Figure 1) and thematic clustering of educational content (Table 1). These additions were made to provide a more intuitive and comprehensive overview of the research landscape and to better highlight identified gaps in the literature.

Figure 1. Evidence gap map of interprofessional education learning formats and competency domains.



XSL•FO
RenderX

Table . Mapping of interprofessional education competency domains, outcome categories, and evaluation measures across included studies.

Author	Competency domains					Outcome categories				Measurement instruments				
	Communication	Role clarification	Team Functioning	Collaborative leadership	Patient-centered care	Knowledge	Attitudes	Confidence	Practice change	ATHCTS[a]	ISVS[b]	TSS[c]	IPC[d]	Other measures[e]
Eskola et al [26]	✓[f]	✓	✓	✓	✓		✓		✓					✓
Day et al [27]	✓	✓	✓	✓	✓		✓	✓	✓		✓			
Cohen et al [28]	✓	✓	✓	✓	✓		✓		✓	✓		✓		
Zhang et al [29]	✓	✓			✓	✓	✓	✓	✓	✓				
Ober & Lape [30]	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓		✓		
Scuderi & Pain [31]	✓	✓			✓	✓		✓			✓			
Sy [32]	✓	✓	✓		✓		✓			✓			✓	
Phillips et al [33]	✓	✓	✓	✓	✓	✓	✓	✓	✓		✓			
Kenaszchuk et al [34]	✓	✓	✓	✓	✓	✓	✓							✓

[a]ATHCTS: Attitudes Towards Health Care Teams Scale.

[b]ISVS: Interprofessional Socialization and Valuing Scale.

[c]TSS: Team Skills Scale.

[d]IPC Scale: Interprofessional Collaboration Scale.

[e]Consultative and Relational Empathy, Interpersonal Reactivity Index, confidence-based marking, self-assessment tools used Likert scales to assess knowledge, confidence, attitudes, perceived practice changes, etc.

[f]Indicates that the domain or outcome was addressed in the study.

Study Eligibility (Inclusion and Exclusion Criteria)

To ensure a systematic selection process, inclusion and exclusion criteria were structured according to the population, concept, and context framework [22]. The population, concept, and context framework is the recommended standard for scoping reviews to clearly define the boundaries of the inquiry beyond the traditional population, intervention, comparison, outcome format, which is often too restrictive for mapping broad educational interventions. This framework was selected to effectively capture the specific professional cohorts (population), the diverse pedagogical structures of IPE (concept), and the varied clinical environments of rehabilitation (context).

Participants (Population)

This review focused on licensed rehabilitation professionals—specifically PTs, OTs, and SLPs—working in rehabilitation, health care, or community-based settings. These professionals play critical roles in improving patient function and quality of life, and their collaboration is vital for preventing adverse events such as falls and enhancing care quality [10-12]. PTs primarily address physical rehabilitation, OTs focus on daily living activities, and SLPs manage communication and swallowing functions. Their complementary expertise contributes to comprehensive patient care and supports the effectiveness of IPE initiatives [7,11]. Studies involving health care students, prelicensure trainees, or nonprofessional caregivers were excluded, as the objective of this review was to analyze IPE among qualified practitioners.

Concept

Eligible studies were those investigating IPE programs involving at least 1 licensed PT, OT, or SLP. IPE was defined as a structured educational intervention delivered through various methods—such as lectures, workshops, simulations, or on-site training—aimed at improving IPC and communication. Studies were excluded if they lacked a clear description of the IPE content, implementation frequency, or reported outcomes.

XSL•FO
RenderX

Context

Eligible studies were conducted in clinical practice environments, including rehabilitation facilities, hospitals, and community-based care settings. The context was restricted to real-world clinical or community practice environments; studies conducted exclusively in academic or simulation-only settings without relevance to licensed clinical practice were excluded.

Study Design and Publication Type

Eligible study designs included intervention studies (eg, pre-post, quasi-experimental, and randomized controlled trials [RCTs]) and observational studies. Publications were original, peer-reviewed research articles written in English. Descriptive reports, theoretical papers, standalone qualitative studies, dissertations, conference proceedings, and letters were excluded.

Rationale for Excluding Qualitative Studies

Although scoping reviews typically encompass a variety of study designs, this review prioritized studies allowing statistical comparison and synthesis of quantifiable outcomes. To maintain a consistent analytical framework focused on measurable effects of IPE, qualitative studies were excluded. Nonetheless, the complementary value of qualitative research is acknowledged for future synthesis.

Search Strategy

A comprehensive literature search was conducted using four databases: PubMed, Web of Science, CINAHL (Cumulative Index to Nursing and Allied Health Literature), and Educational Resources Information Center —covering the period from the inception of each database up to December 31, 2025. These databases were selected due to their comprehensive coverage of research related to IPE and health care professionals, with a particular focus on rehabilitation, nursing, and educational fields. Due to institutional restrictions, access to the Scopus database, which covers a broad range of scientific literature, was unavailable during the review period. PubMed served as the interface to access Medical Literature Analysis and Retrieval System Online-indexed records; Medical Literature Analysis and Retrieval System Online was not searched separately.

The initial search strategy was developed in PubMed using both controlled vocabulary (Medical Subject Headings terms) and free-text keywords, including “interprofessional education,” “rehabilitation,” “physical therapy,” “occupational therapy,” and “speech-language pathology.” Boolean operators (AND, OR) were used to refine the results. This PubMed strategy was then adapted for each database to reflect differences in indexing systems and search syntax—for example, Medical Subject Headings terms were replaced with CINAHL, headings in CINAHL, and equivalent keywords were applied in Web of Science and Educational Resources Information Center using appropriate truncation symbols and Boolean operators. The complete search strings for all databases are provided in Multimedia Appendix 1 to ensure transparency and reproducibility. Reference lists of included articles were also manually screened to identify additional relevant studies. To ensure that the review reflects the most recent scholarly contributions in the field, the initial search was updated on January 3, 2026, by rerunning the exact search strings across all 4 databases. This update extended the search period to include all relevant records published through December 31, 2025.

In accordance with PRISMA-S guidelines, further methodological details are disclosed to ensure transparency. Contacting study authors for additional data was not performed as the published reports provided sufficient information, and no additional methods, such as searching clinical trial registries, were used given the review’s focus on peer-reviewed literature. While a formal external peer review of the search strategy was not conducted, the strategy was developed and validated through internal pilot testing and consensus among the research team.

Study Selection

All retrieved records were imported into Rayyan software (Qatar Computing Research Institute), and duplicates were automatically removed. Title and abstract screening were conducted independently by pairs of reviewers (HN and JU, TN and TK, NI and TT, SH and KI). Each reviewer’s decisions were blinded to their respective partner within the pair to ensure unbiased screening. Any discrepancies were resolved through consultation with a third reviewer. The same independent and blinded process was applied during full-text screening. The selection process is illustrated in the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) flow diagram.

Data Extraction and Analysis

A structured data-extraction form was pilot-tested. A total of 2 or more reviewers independently extracted data regarding study design, setting, participant characteristics, IPE format, frequency, outcome measures, and key findings. Methodological quality was appraised using the appropriate JBI critical appraisal checklists [35] and the Mixed Methods Appraisal Tool [36] for descriptive purposes (Multimedia Appendices 2 and 3).

To enhance clarity, educational formats were categorized according to the Centre for the Advancement of Interprofessional Education framework [37], and content was classified using the framework proposed by Reeves et al [38]. An evidence gap map (Figure 1) was generated to visualize the density of research across these categories.

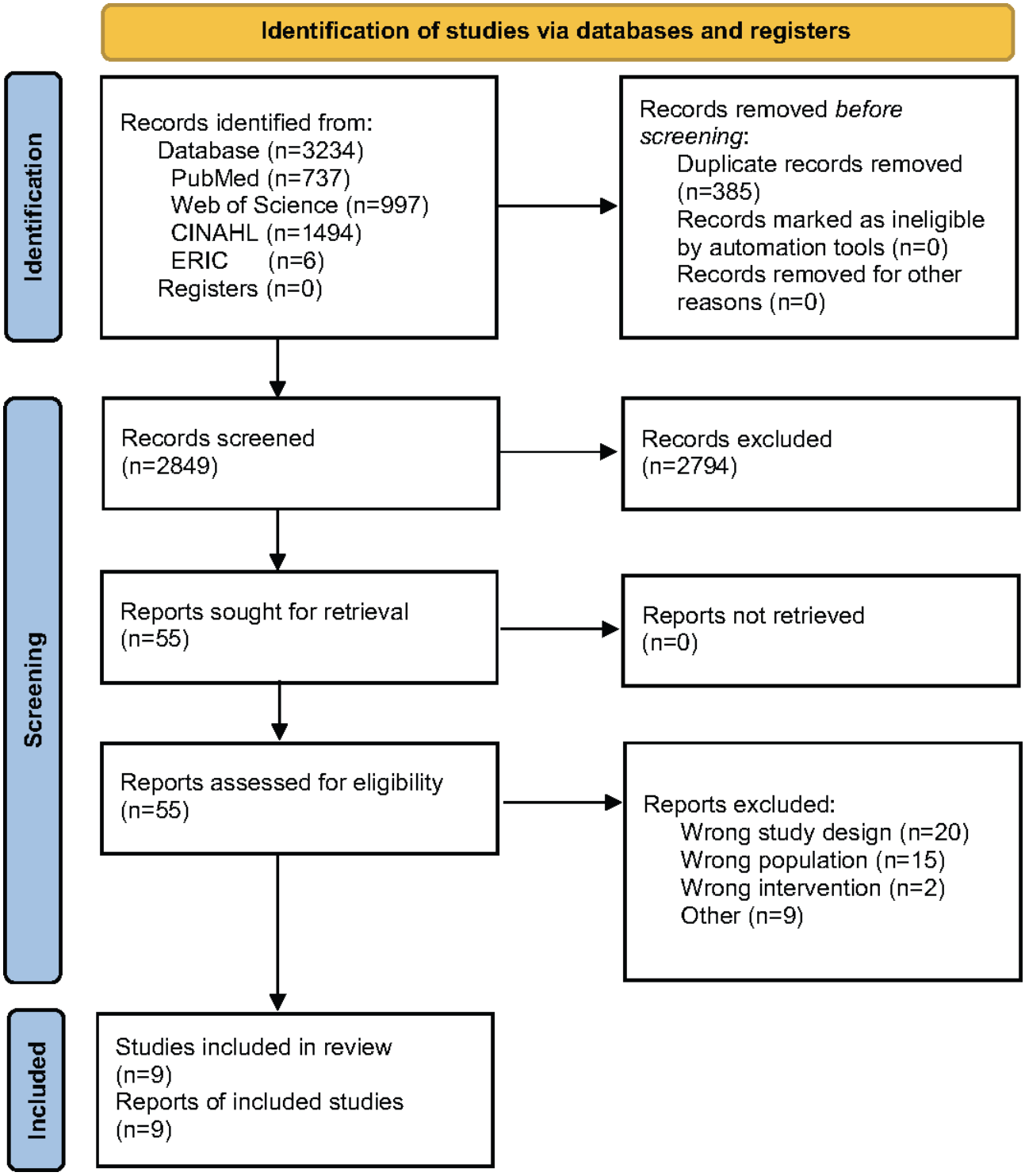
Results

Study Selection

A total of 3234 articles were identified through database searches. After removing 385 duplicates, 2849 titles and abstracts were screened, resulting in the exclusion of 2794 articles. Full-text review of the remaining 55 articles led to the inclusion of 9 studies that met the eligibility criteria. One additional study was identified through the updated search conducted up to December 31, 2025. The list of excluded studies and their reasons for exclusion is provided in Multimedia Appendix 4. The study selection process is summarized in the PRISMA flow diagram (Figure 2).

XSL•FO
RenderX

Figure 2. PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) flow diagram. ERIC: Educational Resources Information Center; CINAHL: Cumulative Index to Nursing and Allied Health Literature.



Study Characteristics

Of the 9 included studies, 1 used a cross-sectional design [32], 5 used single-group pre-post comparisons [28,29,31,33,34], and 3 adopted mixed methods combining pre-post and qualitative approaches [26,27,30] (Table 2). Notably, 1 pre-post study followed a quasi-experimental design [29]. Despite the rigorous selection process, no RCTs were identified.

XSL•FO
RenderX

Table . Characteristics of included studies.

Author, year, country	Study design	Setting	Population	Professions	Key findings
Eskola et al, 2024 [26], United States	Mixed methods pre-post	Veterans affairs hospital and clinics	71 clinicians	Nurses, providers (MD[a], NP[b], PA[c], etc), SWs[d], OTs[e], and pharmacists	Significant improvements in listening, understanding patient concerns, and showing compassion (P<.05). Increased perceived value of peer observation and reflection. Improved ability to work within interprofessional teams.
Day et al, 2022 [27], Australia	Mixed methods process evaluation	9 organizations across Adelaide and Sydney (aged care, disability, and navigation services)	9 professionals (3 aged care, 3 disability, and 3 navigation)	Health services professionals, and OTs	High feasibility and completion; ISVS[f]-21 scores improved; NoMAD[g] showed increased confidence but reduced perceived integration; qualitative data revealed enhanced understanding, motivation, and team-based implementation; sustainability supported by structured planning and peer learning.
Cohen et al, 2021 [28], United States	Quasi-experimental pre-post	5 regions in the United States (New York City, Phoenix, Chicago, North Carolina, and Washington DC)	236 practitioners	MD, RN[h], NP, PA, OT, PT[i], SLP[j], MT[k], and SW	Significant improvement in TSS[l] scores overall (median 3.76 [IQR 3.4-4.1] to 4.35 [IQR 3.9-4.7]; P<.001); all professions improved except the PA group (n=3); 58% rated themselves "very likely" to change practice, and this group showed greater TSS gains (P=.03); TSS change correlated with ATHCTS[m] (ρ=0.42), confidence with PD[n] (ρ=0.18-0.21), team knowledge (ρ=0.23-0.29), and new information gained (ρ=0.26-0.31).
Zhang et al 2021 [29], Australia	Single-group pre-post	Pediatric hospital and health service in New South Wales	30 interpreters or translators	Interpreters and translators (various languages and cultural backgrounds)	Significant improvement in knowledge and confidence (pre→post); practice improved post but declined at the 2-month follow-up (2MPost); attitudes remained high throughout; less experienced ITs showed greater knowledge gains.

XSL•FO
RenderX

Author, year, country	Study design	Setting	Population	Professions	Key findings
Zhang et al, 2021 [29], Australia	Quasi-experimental	Pediatric hospital and health service	49 SPs[o] (27 training, 22 control)	SLPs	Training group showed significant improvements in knowledge, confidence, and practice (pre→post, pre→2MPost); practice declined at 2MPost; attitudes mostly unchanged except A1; training rated highly useful and informative.
Zhang et al, 2021 [29], Australia	Quasi-experimental	Pediatric hospital and health service	24 clinicians (16 training and 8 control)	Allied health (OT, PT, psychology, etc) and nursing	Training group showed significant improvements in Knowledge, confidence, and practice (pre→post, pre→2MPost); practice declined at 2MPost; attitudes unchanged; training rated useful, but application to practice was modest.
Ober & Lape, 2019 [30], United States	Mixed methods pre-post	1 suburban acute care hospital in Pennsylvania	8 rehabilitation professionals	PTs, OTs, and SLPs	Quantitative: All 7 items improved; familiarity with the Kawa model increased most (+3.13); team collaboration scores improved to mean 4.75 (SD not reported). Qualitative: Postsurvey responses showed deeper recognition of individual differences, expanded definitions of professionalism, and more creative collaboration strategies. The Kawa model facilitated shared language and mutual understanding.
Scuderi & Pain, 2019 [31], Australia	Observational (pre-post)	Acute care hospital	33 professionals (28 junior and 5 senior)	PTs, dieticians, SLPs, and OTs	Significant improvement in knowledge and confidence on 5/7 items; 93% satisfaction; CBM[p] effective in promoting safe decision-making.
Sy, 2017 [32], Philippines	Cross-sectional survey	National (online survey of licensed professionals in the Philippines)	189 total (OT: 100, PT: 56, and SLP: 33)	OTs, PTs, and SLPs	Prior IPE[q] experience associated with more positive IPC[r] attitudes; lecture and CBR[s] case studies linked to higher AI[t] scores.

XSL•FO
RenderX

Author, year, country	Study design	Setting	Population	Professions	Key findings
Phillips et al, 2016 [33], Australia	Quasi-experimental (pre-post + 3 mo follow-up)	45 workshops across 7 states/territories via Medicare locals	837 participants (645 enrolled in evaluation)	GPs[u], nurses, mental health professionals, OT, PT, pharmacists, podiatrists, and social workers	Significant improvements in knowledge and confidence; increased use of psychological strategies (especially GPs); 77% of planned practice changes achieved; increased network ties to psychologists, psychiatrists, VVCS[v]; attitudes toward interprofessional working improved.
Kenaszchuk et al, 2011 [34], Canada	Single-group pre-post (uncontrolled intervention)	Community hospital near Toronto	154 health professionals (physicians, nurses, and allied health)	Physicians, nurses, allied health (OT, PT, SLP, pharmacists, social workers, and dietitians)	Simulation training led to attitude improvements among those with lower initial IPC scores; shared leadership attitudes improved over time; reliability of measures varied by profession.

[a]MD: doctor of medicine.
[b]NP: nurse practitioner.
[c]PA: physician assistant.
[d]SW: social worker.
[e]OT: occupational therapist.
[f]ISVS: Interprofessional Socialization and Valuing Scale.
[g]NoMAD: Normalization Measure Development questionnaire.
[h]RN: registered nurse.
[i]PT: physical therapist.
[j]SLP: speech-language pathologist.
[k]MT: medical technologist.
[l]TSS: Team Skills Scale.
[m]ATHCTS: Attitudes Towards Health Care Teams Scale.
[n]PD: Parkinson disease.
[o]SP: standardized patient.
[p]CBM: confidence-based marking.
[q]IPE: interprofessional education.
[r]IPC: interprofessional collaboration.
[s]CBR: community-based rehabilitation.
[t]AI: artificial intelligence.
[u]GP: general practitioner.
[v]VVCS: Virtual Village Case Study.

Participants included a wide range of health professionals, such as OTs, PTs, SLPs, physicians, nurses, pharmacists, social workers, psychologists, dietitians, mental health specialists, podiatrists, prosthetists, and interpreters. Sample sizes ranged from 8 to 197. Training facilitators included clinicians and academic educators with formal teaching experience [28,30,31]. Recruitment strategies varied globally, with significant representation from Australia [27,29,33], the United States [26,30], Canada [34], and the Philippines [32].

Implementation and Content

The educational formats and content domains of the IPE programs are summarized in Table 3 and visualized in our evidence gap map (Figure 1). Systematic mapping identified 7 educational formats. Didactic methods, specifically lectures (n=36 mentions across studies) and discussions (n=35), were the most dominant and were consistently applied across all competency domains. In contrast, experiential formats such as role play (n=7) and case studies (n=11) were notably less frequent.

XSL•FO
RenderX

Table . Characteristics of IPE[a] formats.

Author	IPE format	Frequency	Trainer background
Eskola et al [26]	6.5-hour workshop (lecture, actor-based simulation, and applied improvisation)	Single-day session	Palliative care physicians and an experienced medical educator
Day et al [27]	6-module online education + webinars + mentoring + team-based implementation planning	Weekly sessions over 5 weeks	Delivered by university faculty with simulation teaching experience; professions included OT[b], PT[c], and nursing
Cohen et al [28]	ATTP-IPE[d] training program (single session)	Once	Delivered by interprofessional faculty team; included MDs[e], RNs[f], and educators with IPE teaching background
Zhang et al [29]	120-minute face-to-face training (presentation, handouts, role play, and discussion)	Not applicable (retrospective survey)	Delivered by bilingual clinical educators with teaching experience; professions not specified
Ober & Lape [30]	5-week team-building intervention using Kawa model in individual and group sessions	One-time training per participant	Delivered by OT faculty with expertise in Kawa model and team facilitation
Scuderi & Pain [31]	2-hour pharmacist-led workshop with didactic + case study + CBM[g]-based quiz	Single session per participant	Delivered by clinical pharmacists with teaching experience
Sy [32]	Retrospective self-report of prior IPE experience; classified into 8 TL[h] strategies (eg, case discussion, small group discussion, and lecture)	1 session per participant (12 workshops total)	Not applicable; no training delivered in study
Phillips et al [33]	6-hour or 2×3-hour workshop with didactic content, facilitator guide, and commitment-to-change	6-month engagement (mean 8.3 hours education)	Delivered by trained facilitators including GPs[i], nurses, and mental health professionals; all had prior IPE facilitation experience
Kenaszchuk et al [34]	1-day interprofessional simulation workshop with didactic + experiential learning + debrief	Single session	Delivered by trained facilitators from participating professions; simulation experts with IPC[j] background

[a]IPE: interprofessional education.

[b]OT: occupational therapist.

[c]PT: physical therapist.

[d]ATTP-IPE: allied team training for Parkinson interprofessional education.

[e]MD: doctor of medicine.

[f]RN: registered nurse.

[g]CBM: confidence-based marking.

[h]TL: teaching-learning.

[i]GP: general practitioner.

[j]IPC: interprofessional collaboration.

Regarding educational content, communication (n=38) and role clarification (n=35) were the most frequently addressed competencies. The evidence gap map (Figure 1) highlights a high density of research focusing on these foundational competencies via traditional didactic methods. Conversely, collaborative leadership (n=20) was the least addressed domain, representing a significant gap in the current IPE landscape, particularly regarding its implementation through experiential learning.

Outcome Measures

Outcome evaluation focused primarily on short-term educational effects, such as changes in knowledge, attitudes, and confidence. A variety of validated instruments were used, including the Attitudes Toward Health Care Teams Scale [39], the Interprofessional Socialization and Valuing Scale [40], and the Team Skills Scale [41]. Several studies supplemented these with study-specific questionnaires or confidence-based marking to assess decision-making safety [29,33]. Long-term evaluations of behavioral changes in clinical practice or patient-level outcomes were rarely reported across the included studies.

XSL•FO
RenderX

Main Findings

Overall, participants reported positive attitudes toward IPE. Lecture and case study formats were particularly effective in enhancing understanding of professional roles and medication management [30,31]. The Allied Team Training for Parkinson program demonstrated significant gains in perceived team skills and interprofessional understanding [28].

Temporal analysis showed that structured workshops yielded immediate gains in knowledge and confidence [29,33]. However, some inconsistent effects were noted; for instance, simulation-based training did not result in long-term attitudinal improvements [34]. Furthermore, community-based strategies occasionally faced challenges such as group tensions and mismatched professional expectations [27]. These findings underscore that while foundational knowledge is successfully transmitted through current IPE models, sustained behavioral change remains a complex challenge for licensed professionals.

Discussion

This scoping review synthesized the implementation formats, content domains, and reported outcomes of IPE initiatives targeting licensed PTs, OTs, and SLPs, addressing our primary research questions regarding how these programs are structured and evaluated. We identified a clear predominance of didactic formats, such as lectures and discussions, which were most frequently applied to foundational competency domains, including communication and role clarification (Figure 1). While these initiatives generally improved participants' understanding of professional roles and collaborative skills, the evidence suggests a notable scarcity of experiential or simulation-based approaches, particularly in complex domains such as collaborative leadership. These results suggest that although the "what" and "how" of IPE are well-documented for foundational skills, more advanced competencies remain insufficiently addressed, suggesting that the field has yet to fully embrace pedagogical methods that challenge the entrenched professional silos of established practitioners.

The findings of this scoping review extend the existing IPE literature by clarifying how educational formats and content domains are distributed within programs targeting licensed rehabilitation professionals. Previous reviews of IPE have predominantly focused on prelicensure learners and consistently reported positive effects of didactic teaching, simulation, and structured reflection on attitudes toward teamwork and role understanding [5-8]. In contrast, our review demonstrates that, although similar didactic formats dominate postlicensure IPE for practicing PTs, OTs, and SLPs, their reported effects are largely confined to short-term gains in knowledge, confidence, and role clarification. This divergence suggests that educational strategies effective for students may not directly translate to experienced clinicians, whose learning needs are shaped by established professional identities, ingrained clinical routines, and organizational constraints [42,43]. By mapping these patterns across competency domains, this review provides empirical support for calls to differentiate postlicensure IPE from student-focused models rather than assuming continuity across career stages.

From a methodological perspective, interpretation of these findings must be situated within the methodological characteristics and heterogeneity of the included studies. Although overall methodological quality was judged to be acceptable using JBI and Mixed Methods Appraisal Tool criteria, most studies used single-group pre-post or mixed methods designs, and no RCTs were identified. Outcome assessment relied predominantly on self-reported measures of attitudes, confidence, or perceived competence, with generally short follow-up periods. Such designs limit causal inference and are vulnerable to social desirability and recall biases, potentially overstating educational impact [44-46]. Moreover, substantial heterogeneity was observed across studies with respect to participant composition, clinical context, and IPE intensity, ranging from brief single-session workshops to multiweek interventions. While this heterogeneity reflects the diversity of real-world rehabilitation practice, it also constrains the comparability of outcomes and highlights the current lack of robust evidence evaluating sustained behavioral or organizational change in postlicensure IPE.

One of the most salient findings of this review is the relative paucity of IPE interventions explicitly targeting collaborative leadership, despite its recognition as a core competency in contemporary interprofessional frameworks [16,17]. Existing theories of workplace learning and continuing professional development emphasize that leadership and team coordination skills are most effectively cultivated through experiential, context-embedded learning rather than through didactic instruction alone [47,48]. However, our evidence gap mapping indicates that leadership-related competencies are rarely addressed using such approaches in rehabilitation-focused IPE. This misalignment may partly explain why improvements in attitudes or confidence do not consistently translate into sustained changes in clinical behavior [38,49]. By making this gap explicit, the present review contributes to the field by reframing leadership not as an ancillary outcome of communication training, but as a distinct educational target requiring intentional design and evaluation in postlicensure IPE.

Several limitations warrant consideration. The search strategy was restricted to 4 electronic databases and focused on English-language publications, which introduces a risk of publication bias and may limit the generalizability of the findings across different linguistic or cultural contexts. Due to technical constraints, certain comprehensive databases like Scopus were not included, potentially causing the review to miss relevant studies. Methodologically, the evidence base remains limited to a small number of studies, with a notable absence of RCTs, thereby constraining our ability to draw strong causal inferences. Most included studies relied heavily on self-reported outcomes and standardized questionnaires, which are susceptible to social desirability and recall biases. Such measures often capture "perceived" rather than "actual" change in competence and may not fully reflect objective changes in clinical behavior or team dynamics. Furthermore, the exclusion of standalone qualitative studies may have limited our depth of insight into the contextual influences and implementation processes that shape IPE in practice. Finally, as illustrated in our evidence gap map (Figure 1), the effectiveness of modalities

XSL•FO
RenderX

such as simulation remains highly dependent on their alignment with specific interprofessional objectives rather than general instructional goals.

Future research should prioritize the development and evaluation of practice-integrated IPE models that move beyond short-term educational outcomes. Longitudinal and comparative study designs are needed to examine whether experiential learning strategies—such as team-based quality improvement initiatives, facilitated reflective practice, or shared leadership interventions—can produce sustained behavioral change among licensed rehabilitation professionals. Attention should be given to objective indicators of team functioning and decision-making in complex clinical contexts, rather than relying solely on self-reported measures. Additionally, future studies should explicitly account for professional mix, organizational context, and baseline team maturity, as these factors likely moderate IPE effectiveness. Addressing these gaps will be essential for generating more robust evidence base capable of informing institutional investment in IPE as a strategy for improving patient safety, workforce resilience, and the quality of rehabilitation care.

In conclusion, this scoping review adds a novel perspective by focusing exclusively on licensed rehabilitation professionals, distinguishing it from the vast body of existing literature centered on prelicensure students. It brings to the field a clearer understanding of the potential "leadership gap" and the current overreliance on didactic methods for experienced clinicians. The real-world implications suggest that, to move beyond short-term attitudinal change, professional-level IPE must transition toward systematic, practice-integrated models that address identified research gaps in leadership and team functioning. Real-world implementation should prioritize facilitator preparation and institutional support as critical components for sustainability. By fostering enduring improvements in health care team performance through longitudinal programs and objective behavioral measures, health care institutions can ultimately contribute to improvements in patient safety and the quality of rehabilitation care.

Acknowledgments

During the preparation of this manuscript, the authors used a generative artificial intelligence (AI) tool (Gemini, Google) under full human supervision for proofreading and editing and for reformatting to improve linguistic clarity and ensure adherence to the journal's structural requirements, in accordance with the GAIDeT taxonomy (2025). All AI-assisted content was reviewed, revised, and approved by the authors, who take full responsibility for the final manuscript. Generative AI tools are not listed as authors and bear no responsibility for the content of this publication. This research was supported by Japan Society for the Promotion of Science KAKENHI (grant 24K02671).

Funding

The authors declare no competing financial interests. The funding agency had no role in the design of the study; the collection, analysis, and interpretation of data; the writing of the manuscript; or the decision to submit the manuscript for publication.

Data Availability

The data analyzed during this scoping review consist of published articles retrieved through systematic database searches. Full texts cannot be shared due to copyright restrictions. However, the complete search strategy and list of included studies are available in the supplementary materials. Further details are available from the corresponding author upon reasonable request.

Authors' Contributions

KI and HN contributed to the conceptualization of the study. The methodology was developed by KI, HN, TK, and TN. Literature search and screening, as well as data extraction and analysis, were carried out by KI, TK, SH, NI, TN, TT, JU, and HN. The original draft of the manuscript was written by KI, and all authors, including TK, SH, NI, TN, TT, JU, and HN, contributed to the review and editing process. The study was supervised by HN. All authors have read and approved the final manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Search strategy.
[DOCX File, 15 KB - mededu_v12i1e76189_app1.docx]

Multimedia Appendix 2
Quality appraisal template.
[DOCX File, 14 KB - mededu_v12i1e76189_app2.docx]

Multimedia Appendix 3

XSL•FO
RenderX

Quality assessment of studies.
[DOCX File, 16 KB - mededu_v12i1e76189_app3.docx]

Multimedia Appendix 4
List of excluded studies.
[DOCX File, 21 KB - mededu_v12i1e76189_app4.docx]

Checklist 1
PRISMA 2020 checklist.
[DOCX File, 270 KB - mededu_v12i1e76189_app5.docx]

Checklist 2
PRISMA-ScR checklist.
[DOCX File, 96 KB - mededu_v12i1e76189_app6.docx]

Checklist 3
PRISMA-S checklist.
[DOCX File, 17 KB - mededu_v12i1e76189_app7.docx]

References

1. Patient safety. World Health Organization. 2023. URL: https://www.who.int/news-room/fact-sheets/detail/patient-safety [accessed 2024-09-05]
2. Rosen MA, DiazGranados D, Dietz AS, et al. Teamwork in healthcare: key discoveries enabling safer, high-quality care. Am Psychol 2018;73(4):433-450. [doi: 10.1037/amp0000298] [Medline: 29792459]
3. Ndoro S. Effective multidisciplinary working: the key to high-quality care. Br J Nurs 2014;23(13):724-727. [doi: 10.12968/bjon.2014.23.13.724] [Medline: 25072333]
4. Stewart MA. Stuck in the middle: the impact of collaborative interprofessional communication on patient expectations. Shoulder Elbow 2018 Jan;10(1):66-72. [doi: 10.1177/1758573217735325] [Medline: 29276540]
5. Marion-Martins AD, Pinho DLM. Interprofessional simulation effects for healthcare students: a systematic review and meta-analysis. Nurse Educ Today 2020 Nov;94:104568. [doi: 10.1016/j.nedt.2020.104568] [Medline: 32932058]
6. Saragih ID, Arna Uly Tarihoran DET, Sharma S, Chou FH. A systematic review and meta-analysis of outcomes of interprofessional education for healthcare students from seven countries. Nurse Educ Pract 2023 Aug;71:103683. [doi: 10.1016/j.nepr.2023.103683] [Medline: 37433234]
7. Shuyi AT, Zikki LYT, Mei Qi A, Koh Siew Lin S. Effectiveness of interprofessional education for medical and nursing professionals and students on interprofessional educational outcomes: a systematic review. Nurse Educ Pract 2024 Jan;74:103864. [doi: 10.1016/j.nepr.2023.103864] [Medline: 38101092]
8. Spaulding EM, Marvel FA, Jacob E, et al. Interprofessional education and collaboration among healthcare students and professionals: a systematic review and call for action. J Interprof Care 2021;35(4):612-621. [doi: 10.1080/13561820.2019.1697214] [Medline: 31865823]
9. Wang Z, Feng F, Gao S, Yang J. A systematic meta-analysis of the effect of interprofessional education on health professions students' attitudes. J Dent Educ 2019 Dec;83(12):1361-1369. [doi: 10.21815/JDE.019.147] [Medline: 31548305]
10. McDonald CL, Fuhs AK, Kartin D. An examination of current stroke rehabilitation practice in Peru: implications for interprofessional education. J Interprof Care 2018 May;32(3):329-338. [doi: 10.1080/13561820.2018.1425290] [Medline: 29364739]
11. Strasser DC, Falconer JA, Stevens AB, et al. Team training and stroke rehabilitation outcomes: a cluster randomized trial. Arch Phys Med Rehabil 2008 Jan;89(1):10-15. [doi: 10.1016/j.apmr.2007.08.127] [Medline: 18164324]
12. Weaver SJ, Rosen MA, DiazGranados D, et al. Does teamwork improve performance in the operating room? A multilevel evaluation. Jt Comm J Qual Patient Saf 2010 Mar;36(3):133-142. [doi: 10.1016/s1553-7250(10)36022-3] [Medline: 20235415]
13. Harden RM, Crosby JR, Davis MH, Friedman M. AMEE Guide No. 14: outcome-based education: part 5-from competency to meta-competency: a model for the specification of learning outcomes. Med Teach 1999;21(6):546-552. [doi: 10.1080/01421599978951] [Medline: 21281173]
14. Jiang Y, Cai Y, Zhang X, Wang C. Interprofessional education interventions for healthcare professionals to improve patient safety: a scoping review. Med Educ Online 2024 Dec 31;29(1):2391631. [doi: 10.1080/10872981.2024.2391631] [Medline: 39188239]
15. van der Weerd LP, Hurkmans J, Reinders JJ, et al. Facilitators and barriers to interprofessional learning, collaboration, and identity development during interprofessional internship in rehabilitation care: a qualitative study. J Interprof Care 2025;39(5):736-749. [doi: 10.1080/13561820.2024.2447259] [Medline: 39812316]

XSL•FO
RenderX

16. Council on Academic Accreditation (CAA) in Audiology and Speech-Language Pathology. Standards for accreditation of graduate education programs in audiology and speech-language pathology. : American Speech-Language-Hearing Association (ASHA); 2023 URL: https://caa.asha.org/siteassets/files/accreditation-standards-for-graduate-programs.pdf [accessed 2025-01-09]
17. IPEC core competencies for interprofessional collaborative practice: version 3. : Interprofessional Education Collaborative (IPEC); 2023 URL: https://www.ipecollaborative.org/assets/core-competencies/IPEC_Core_Competencies_Version_3_2023.pdf [accessed 2025-01-09]
18. Christophers L, Torok Z, Trayer A, Hong GCC, Carroll A. Interdisciplinary teamworking in rehabilitation: experiences of change initiators in a national rehabilitation hospital. BMC Health Serv Res 2025 May 7;25(1):651. [doi: 10.1186/s12913-025-12795-6] [Medline: 40329284]
19. Hustoft M, Biringer E, Gjesdal S, Aβmus J, Hetlevik Ø. Relational coordination in interprofessional teams and its effect on patient-reported benefit and continuity of care: a prospective cohort study from rehabilitation centres in Western Norway. BMC Health Serv Res 2018 Sep 17;18(1):719. [doi: 10.1186/s12913-018-3536-5] [Medline: 30223847]
20. Arksey H, O'Malley L. Scoping studies: towards a methodological framework. Int J Soc Res Methodol 2005 Feb;8(1):19-32. [doi: 10.1080/1364557032000119616]
21. Khalil H, Peters M, Godfrey CM, McInerney P, Soares CB, Parker D. An evidence-based approach to scoping reviews. Worldviews Evid Based Nurs 2016 Apr;13(2):118-123. [doi: 10.1111/wvn.12144] [Medline: 26821833]
22. Levac D, Colquhoun H, O'Brien KK. Scoping studies: advancing the methodology. Implement Sci 2010 Sep 20;5:69. [doi: 10.1186/1748-5908-5-69] [Medline: 20854677]
23. Peters MDJ, Marnie C, Tricco AC, et al. Updated methodological guidance for the conduct of scoping reviews. JBI Evid Synth 2020 Oct;18(10):2119-2126. [doi: 10.11124/JBIES-20-00167] [Medline: 33038124]
24. Tricco AC, Lillie E, Zarin W, et al. PRISMA Extension for Scoping Reviews (PRISMA-ScR): checklist and explanation. Ann Intern Med 2018 Oct 2;169(7):467-473. [doi: 10.7326/M18-0850] [Medline: 30178033]
25. Rethlefsen ML, Kirtley S, Waffenschmidt S, et al. PRISMA-S: an extension to the PRISMA statement for reporting literature searches in systematic reviews. Syst Rev 2021 Jan 26;10(1):39. [doi: 10.1186/s13643-020-01542-z] [Medline: 33499930]
26. Eskola L, Silverman E, Rogers S, Zelenski A. Right-sizing interprofessional team training for serious-illness communication: a strength-based approach. PEC Innov 2024;4:100267. [doi: 10.1016/j.pecinn.2024.100267] [Medline: 38414869]
27. Day S, Couzner L, Laver KE, Withall A, Draper B, Cations M. Cross-sector learning collaboratives can improve post-diagnosis care integration for people with young onset dementia. Health Soc Care Community 2022 Nov;30(6):e6135-e6144. [doi: 10.1111/hsc.14051] [Medline: 36177663]
28. Cohen EV, Luce V, González MJ, Hagestuen R, González-Ramos G, Cohen HW. Improving team skills of health care practitioners: effects of an interprofessional education program. J Contin Educ Health Prof 2021 Jul 1;41(3):185-194. [doi: 10.1097/CEH.0000000000000358] [Medline: 33929352]
29. Zhang CX, Crawford E, Marshall J, Bernard A, Walker-Smith K. Developing interprofessional collaboration between clinicians, interpreters, and translators in healthcare settings: outcomes from face-to-face training. J Interprof Care 2021;35(4):521-531. [doi: 10.1080/13561820.2020.1786360] [Medline: 32693645]
30. Ober JL, Lape JE. Cultivating acute care rehabilitation team collaboration using the Kawa model. Internet J Allied Health Sci Pract 2019;17(3). [doi: 10.46743/1540-580X/2019.1827]
31. Scuderi C, Pain T. Using confidence assessment to explore physiotherapists, dieticians, speech pathologists and occupational therapists knowledge of medicines. J Interprof Care 2019;33(5):593-597. [doi: 10.1080/13561820.2018.1551860] [Medline: 30486689]
32. Sy MP. Filipino therapists' experiences and attitudes of interprofessional education and collaboration: a cross-sectional survey. J Interprof Care 2017 Nov;31(6):761-770. [doi: 10.1080/13561820.2017.1359509] [Medline: 29043925]
33. Phillips CB, Hall S, Irving M. Impact of interprofessional education about psychological and medical comorbidities on practitioners' knowledge and collaborative practice: mixed method evaluation of a national program. BMC Health Serv Res 2016 Sep 2;16(1):465. [doi: 10.1186/s12913-016-1720-z] [Medline: 27590686]
34. Kenaszchuk C, MacMillan K, van Soeren M, Reeves S. Interprofessional simulated learning: short-term associations between simulation and interprofessional collaboration. BMC Med 2011 Mar 28;9(29):29. [doi: 10.1186/1741-7015-9-29] [Medline: 21443779]
35. Critical appraisal tools. Joanna Briggs Institute (JBI). 2025. URL: https://jbi.global/critical-appraisal-tools [accessed 2025-11-11]
36. Hong QN, Fàbregues S, Bartlett G, et al. The Mixed Methods Appraisal Tool (MMAT) version 2018 for information professionals and researchers. Educ Inf 2018;34(4):285-291. [doi: 10.3233/EFI-180221]
37. Barr H, Gray R, Helme M, Low H, Reeves S, editors. Interprofessional education guidelines 2016. : Centre for the Advancement of Interprofessional Education (CAIPE); 2016 URL: https://www.caipe.org/wp-content/uploads/download-files/2025/08/caipe-interprofessional-education-guidelines-2016-1.pdf [accessed 2026-02-13]
38. Reeves S, Perrier L, Goldman J, Freeth D, Zwarenstein M. Interprofessional education: effects on professional practice and healthcare outcomes (update). Cochrane Database Syst Rev 2013 Mar 28;2013(3):CD002213. [doi: 10.1002/14651858.CD002213.pub3] [Medline: 23543515]

XSL•FO
RenderX

39. Heinemann GD, Schmitt MH, Farrell MP, Brallier SA. Development of an Attitudes Toward Health Care Teams Scale. Eval Health Prof 1999 Mar;22(1):123-142. [doi: 10.1177/01632789922034202] [Medline: 10350960]
40. King G, Orchard C, Khalili H, Avery L. Refinement of the Interprofessional Socialization and Valuing Scale (ISVS-21) and development of 9-item equivalent versions. J Contin Educ Health Prof 2016;36(3):171-177. [doi: 10.1097/CEH.0000000000000082] [Medline: 27583993]
41. Grymonpre R, Van Ineveld C(, Nelson M, et al. See it—do it—learn it: learning interprofessional collaboration in the clinical context. JRIPE 2010;1(2):127-144. [doi: 10.22230/jripe.2010v1n2a13]
42. Best S, Williams S. Professional identity in interprofessional teams: findings from a scoping review. J Interprof Care 2019;33(2):170-181. [doi: 10.1080/13561820.2018.1536040] [Medline: 30335534]
43. Paradis E, Whitehead CR. Louder than words: power and conflict in interprofessional education articles, 1954-2013. Med Educ 2015 Apr;49(4):399-407. [doi: 10.1111/medu.12668] [Medline: 25800300]
44. McFadyen AK, Webster VS, Maclaren WM, O'neill MA. Interprofessional attitudes and perceptions: results from a longitudinal controlled trial of pre-registration health and social care students in Scotland. J Interprof Care 2010 Sep;24(5):549-564. [doi: 10.3109/13561820903520369] [Medline: 20218778]
45. Colthart I, Bagnall G, Evans A, et al. The effectiveness of self-assessment on the identification of learner needs, learner activity, and impact on clinical practice: BEME guide no. 10. Med Teach 2008;30(2):124-145. [doi: 10.1080/01421590701881699] [Medline: 18464136]
46. Davis DA, Mazmanian PE, Fordis M, Van Harrison R, Thorpe KE, Perrier L. Accuracy of physician self-assessment compared with observed measures of competence: a systematic review. JAMA 2006 Sep 6;296(9):1094-1102. [doi: 10.1001/jama.296.9.1094] [Medline: 16954489]
47. Sonnino RE. Health care leadership development and training: progress and pitfalls. J Healthc Leadersh 2016;8:19-29. [doi: 10.2147/JHL.S68068] [Medline: 29355187]
48. O'Brien BC, Battista A. Situated learning theory in health professions education research: a scoping review. Adv Health Sci Educ Theory Pract 2020 May;25(2):483-509. [doi: 10.1007/s10459-019-09900-w] [Medline: 31230163]
49. Grimshaw JM, Thomas RE, MacLennan G, et al. Effectiveness and efficiency of guideline dissemination and implementation strategies. Health Technol Assess 2004 Feb;8(6). [doi: 10.3310/hta8060] [Medline: 14960256]

Abbreviations

CINAHL: Cumulative Index to Nursing and Allied Health Literature
IPC: interprofessional collaboration
IPE: interprofessional education
JBI: Joanna Briggs Institute
OT: occupational therapist
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
PRISMA-S: Preferred Reporting Items for Systematic Reviews and Meta-Analyses Search Reporting Extension
PRISMA-ScR: Preferred Reporting Items for Systematic Reviews and Meta-Analyses Extension for Scoping Reviews
PT: physical therapist
RCT: randomized controlled trial
SLP: speech-language pathologist

Edited by S Brini; submitted 18.Apr.2025; peer-reviewed by D Ajayi, S Javorszky, VP Jover; revised version received 23.Jan.2026; accepted 23.Jan.2026; published 04.Mar.2026.

Please cite as:
Ikeda K, Kaneko T, Hijikuro S, Inoue N, Nakamura T, Takeda T, Uchida J, Nagayama H
Educational Formats and Content Domains of Interprofessional Education for Licensed Rehabilitation Professionals: Scoping Review
JMIR Med Educ 2026;12:e76189
URL: https://mededu.jmir.org/2026/1/e76189
doi:10.2196/76189

© Kohei Ikeda, Takao Kaneko, Someka Hijikuro, Natsuki Inoue, Takuto Nakamura, Taisei Takeda, Junya Uchida, Hirofumi Nagayama. Originally published in JMIR Medical Education (https://mededu.jmir.org), 4.Mar.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR

XSL•FO
RenderX

Medical Education, is properly cited. The complete bibliographic information, a link to the original publication on https://mededu.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Retuning the Premedical Compass in American Programs Worldwide: Scoping Review

Michelle E Chouairy[1]; Laetitia I Bou Samra[1], BS; Christelle I Zeidan[1]; Ola El Zein[2], PhD; Bilal R Kaafarani[1,3], Prof Dr

[1]Medical Research Volunteer Program (MRVP), American University of Beirut, Beirut, Lebanon

[2]Clinical Research Institute, Faculty of Medicine, American University of Beirut, Beirut, Lebanon

[3]Department of Chemistry, American University of Beirut, PO Box 11-0236, Riad El Solh, Beirut, Lebanon

Corresponding Author:
Bilal R Kaafarani, Prof Dr
Medical Research Volunteer Program (MRVP), American University of Beirut, Beirut, Lebanon

Abstract

Background: Premedical education provides the academic foundation for entry into medical school, yet requirements differ widely across institutions and have shifted substantially over the past 2 decades. Recent Medical College Admission Test reforms and growing calls for more comprehensive and interdisciplinary premedical training have triggered curricular changes fragmented throughout the literature. Such manifestations make it particularly difficult for institutions and students to interpret the direction, rationale, and effectiveness of these changes.

Objective: This scoping review aims to map the literature in order to identify key reforms in premedical course requirements and classify these changes into different types: addition of courses, removal of courses, redesign of courses, and the proposition of a new curriculum. This review also evaluates the motives, rationales, and outcomes of these changes for the sake of building a unified foundation for institutions, stakeholders, and students aspiring to delve into medical education.

Methods: We systematically searched 4 databases up to July 4, 2025. We included studies focusing on American undergraduate premedical curricula and reporting changes to course or curriculum requirements. Studies discussing non-American programs, medical or postmedical curricula, or requirements beyond courses were excluded. Two reviewers independently screened studies at the title, abstract, and full-text levels, with conflicts resolved by consensus among all team members. Data extraction was also performed in duplicates. The Mixed Methods Appraisal Tool (McGill University) was used to assess the risk of bias. Because of its heterogeneity, the data were synthesized narratively.

Results: We included 70 studies, which ranged across mixed methods, quantitative, qualitative, and descriptive designs. Three recurring drivers of reform emerged: conforming to Medical College Admission Test revisions, educating and forming well-prepared aspiring physicians, and transitioning to competency-based education. From the 70 studies, 36 highlighted adding courses, 25 described revising existing courses, 5 discussed course deletions, and 10 proposed new curricula. Outcomes prevalently showed improved preparedness for medical school, positive student feedback, and enhanced academic performance, though some results were mixed.

Conclusions: To our knowledge, there is no comprehensive review on how the premedical curriculum has evolved over time. Thus, this scoping review, encompassing all these changes, will be a unified framework that may help both institutions and students become aware of the premedical course requirements on a broader scale and determine which reforms are most effective. Since 2000, premedical education has undergone a shift from a narrow science checklist toward a more comprehensive outcomes-oriented preparation, whereby a balanced approach that merges scientific core courses with social sciences and humanities seems to be the most promising for preparing premedical students for medical school and contributing to their professional identity formation. Future work should investigate to what extent these changes to the premedical course requirements have been implemented.

(JMIR Med Educ 2026;12:e85002) doi:10.2196/85002

KEYWORDS

American premedical curriculum; premedical education; revisiting premedical requirements; medical school preparedness; competency-based premedical curriculum; Medical College Admission Test; MCAT revision

XSL•FO
RenderX

Introduction

Rationale

Many students aspiring to become doctors choose to follow the American program whether at an institution in the United States or abroad. This American-style pathway is widely adopted internationally, yet what constitutes "adequate" premedical preparation is not standardized across institutions. This program is characterized by a structured premedical education track, an academic foundation that ensures students enter medical school prepared and ready. The latter includes the completion of an undergraduate degree alongside essential premedical course requirements in general biology, chemistry, physics, English, and other subjects necessary to prepare students for the Medical College Admission Test (MCAT), a standardized multiple-choice exam that assesses students' readiness for admission to medical school.

Historically, premedical education was shaped by the 1910 Flexner Report, which highlighted the necessity of establishing a premedical curriculum centered on natural sciences like physics, biology, and chemistry [1]. At that time, these courses were considered tools to "weed out" less-prepared medical school applicants. However, recent critiques and contemporary reports, such as those from the Howard Hughes Medical Institute–Association of American Medical Colleges, have emerged and called for greater emphasis on liberal arts, humanities, and ethics in premedical education [2]. Over the years, several revisions to the MCAT have been attempted, with the most recent occurring in 2015 [3]. Pertinent to that is the shift in the framework from a rigid course "checklist" toward competencies and outcomes-oriented preparation.

Indeed, in 2015, the Association of American Medical Colleges revised the MCAT and included psychology and sociology critical analysis sections, shifting the premedical focus from natural sciences to broader critical thinking skills [4]. Importantly, these changes in the MCAT altered the narrative of scientific preparation by shedding light on biochemistry and research methods, whereas the previous exam focused on discrete knowledge within chemistry and biology [5]. The new version required students to combine these disciplines in order to solve complex problems that serve to mirror the interdisciplinary nature of modern medical curricula. Additionally, the substitution of "verbal reasoning" with critical analysis and reasoning skills further expanded the scope of premedical inquiry into the humanities and social sciences, while prioritizing the ability to synthesize diverse information over simple reading comprehension skills. Consequently, this pathway forced undergraduate institutions to move away from "premedical checklists" in light of curricula that instigate scientific inquiry and interdisciplinary reasoning [5]. Ultimately, this redesign explicitly broadened the framework for assessing readiness by integrating behavioral and social sciences and emphasizing reasoning skills. Nevertheless, these changes created a disconnect between what is tested on the MCAT, what is required for admission by different medical schools, and what is taught in undergraduate institutions. This misalignment may be amplified by the persistent heterogeneity in prerequisite structures across medical schools.

Changes in the MCAT have led institutions, educators, physicians, and even students to question the content, purpose, and structure of their undergraduate curricula that differs from one institution to another. The undergraduate curriculum of the American program must be fixed, and various measures and reforms are taking place in many universities around the world. Growing critiques are found in the literature, as rethinking premedical education might not be evolving in the right direction. The literature is filled with studies discussing how different institutions are reevaluating their curricula and changing their premedical requirements. However, the literature available is fragmented. In fact, competency-based education has increasingly influenced how readiness could be connotatively defined, with newer Association of American Medical Colleges frameworks emphasizing broader competencies, such as thinking or reasoning, interpersonal skills, and professional development alongside the scope of scientific foundations [4]. Ultimately, recent scholarships have also reevaluated the educational value of the humanities and broader interdisciplinary preparation for professional identity formation (PIF) in premedical students, preparing them for "thinking, acting, and feeling like physicians." Indeed, PIF refers to the developmental process through which students internalize the values, ethical considerations, behaviors, and norms of the profession they aspire to practice [6].

Objectives

Motivated by the need to understand the nature and impact of recent modifications to premedical course requirements across the literature, this scoping review may serve as a guide for educators, curriculum designers, medical schools, and students themselves to identify the best combination of courses needed to fulfill the academic requirements for medical education. The aim of this scoping review is to answer the following question: How have premedical course requirements in American undergraduate programs changed since the year 2000, and what are the motives and outcomes associated with these modifications?

Through an intricate screening process of the available literature, this scoping review provides an evidence-based, unbiased synthesis of the most effective changes to the premedical curriculum according to their outcomes and impact on students' admission to medical school and future academic performance. This scoping review aims to classify the changes in the literature into 4 types: the addition of courses to fulfill a certain need [1]; the removal of courses that do not affect students' preparation for medical school [2,7]; the redesign of courses to tailor them more to the premedical students' needs [8]; and the proposition of a new curriculum that no longer includes traditional premedical course requirements, such as a new major designed specifically for premedical student preparation [2,9-11]. Each of these changes' drivers, outcomes, and rationales will also be reviewed and examined to build a unified synthesis of the available literature. By consolidating this fragmented evidence base, our findings are directed to support decision-making for

XSL•FO
RenderX

institutions seeking to modernize premedical requirements and for students navigating variable prerequisites across programs.

Methods

Protocol and Registration

This scoping review critically maps and synthesizes the literature on premedical requirements in American programs worldwide. We registered the protocol for this scoping review on the Open Science Framework, adhering to the PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses Extension for Scoping Reviews) guidelines (Checklist 1) [12].

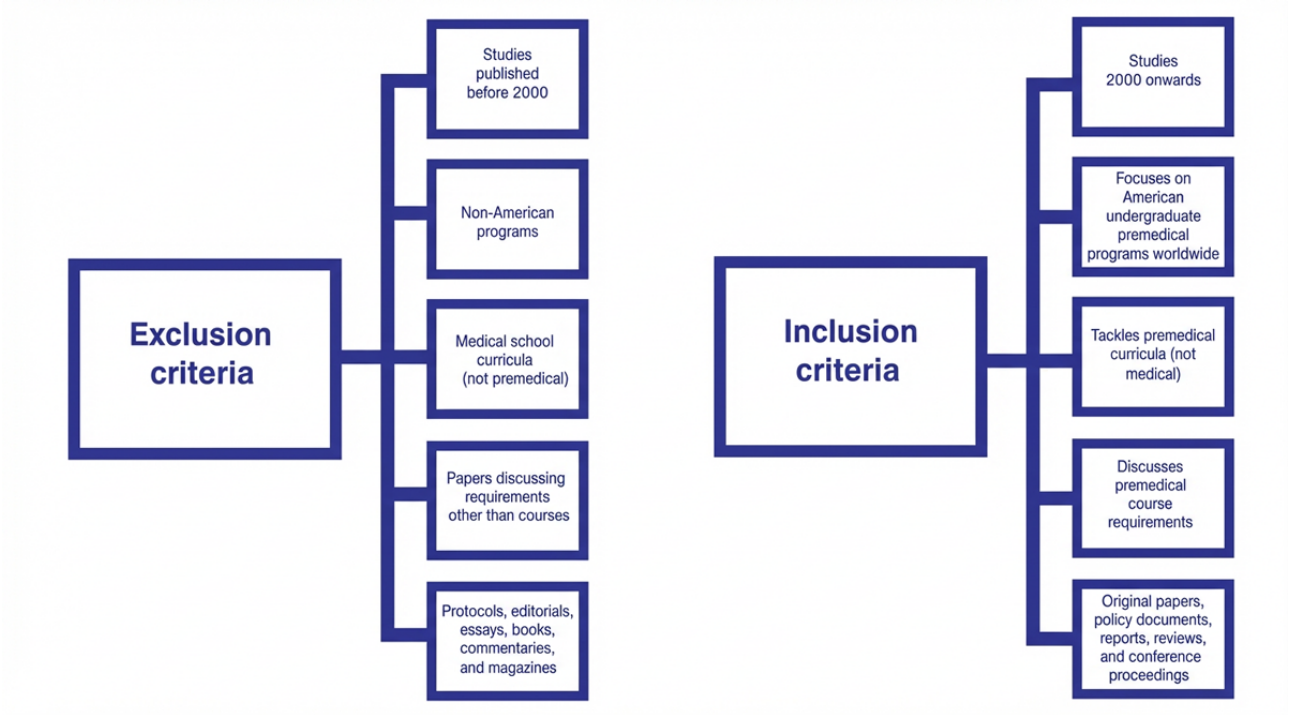
Search Strategy and Information Sources

We individually searched the following 4 electronic databases on December 16, 2024, and then reran the search up until July 4, 2025, in accordance with the PRISMA-S (Preferred Reporting Items for Systematic Reviews and Meta-Analyses Literature Search Extension) guidelines (Checklist 2), without language limitations: MEDLINE (using Ovid), Embase (using Elsevier), ERIC, and Education Research Complete (using EBSCO) [13]. The search strategy included both controlled vocabularies (eg, Medical Subject Headings [MeSH] and Emtree) and keywords such as premedical requirements, premedical education, prerequisite courses for medical school, premedical students, school admission criteria, premedical acceptance, premedical curriculum, premedical courses, prerequisite, precondition, and applying for medical school. This review did not involve searching study registries, applying predefined search filters, or conducting a formal peer review of the search strategy. The search strategies were not adapted from other reviews; they were developed by the director of evidence synthesis (OZ) at the American University of Beirut (AUB). The full search strategies for MEDLINE, Embase, ERIC, and Education Research Complete are provided in Multimedia Appendices 1-4. As for gray literature, both backward and forward citation tracking of all relevant studies was performed. The Howard Hughes Medical Institute–Association of American Medical Colleges reports were purposefully searched, but no additional data were sought by contacting authors, experts, manufacturers, or others.

Eligibility Criteria

The studies were selected based on a specific set of inclusion and exclusion criteria, as shown in Figure 1. Studies were included if they were: (1) published from 2000 onwards; (2) focused on American undergraduate premedical programs worldwide; (3) tackled premedical curricula (not medical); (4) discussed premedical course requirements; and (5) identified as original studies, policy documents, reports, reviews, or conference proceedings.

Figure 1. Exclusion and inclusion criteria.



Studies were excluded if they were: (1) published before 2000; (2) tackled non-American programs; (3) discussed medical and not premedical curricula; (4) focused on requirements other than courses, such as extracurriculars, internships, and research; or (5) were identified as a protocol, editorial, erratum, essay, book, commentary, or magazine article.

Selection of Sources of Evidence

All studies were imported into Covidence (Veritas Health Innovation Ltd), an online software platform designed to improve the efficiency and experience of creating and maintaining scoping reviews [14]. The platform was also used to detect and remove duplicates.

XSL•FO
RenderX

Data Charting Process

Pairs of researchers (MEC, LIB, and CIZ) independently performed title, abstract, and full-text screening in duplicates, and after calibration exercises, reached a 95% agreement rate among reviewers. Disagreements among reviewers were resolved through discussions among all team members. Duplicate entries were removed manually and by Covidence before screening. The reasons for excluding studies at the full-text level were documented and shown in the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) flowchart.

Data Extraction Process

Pairs of researchers (MEC, LIB, and CIZ) conducted data extraction independently and in duplicates using a standardized table developed specifically for this review and piloted on a small subset of included studies. For each study, we abstracted the following variables: author, year, location, institution, type of change (add, remove, revise, or new curriculum), motives, outcomes, and themes of each study, which were compiled into a Microsoft Excel table. A unified and comprehensive table was synthesized after resolving any discrepancies in the extracted information through consensus (Multimedia Appendix 5).

Critical Appraisal

The Mixed Methods Appraisal Tool (MMAT; McGill University) was used for quality and risk of bias assessment. This method allows for the appraisal and description of methodological quality across 3 domains: mixed, qualitative, and quantitative (partitioned into 3 subdomains: randomized controlled, nonrandomized, and descriptive) [15]. The risk of bias assessment was completed, and the findings were compiled in an Excel spreadsheet (Multimedia Appendix 6).

Synthesis of Results

The data from the included studies were synthesized narratively. Studies were grouped according to 3 key themes related to the addition, removal, or modification of a course. Outcomes were categorized into 4 types: positive feedback about the course, positive impact and improvement for future career development or medical school, mixed outcomes, or new insights about the course.

Ethical Considerations

Ethical approval was not required as this study was a scoping review.

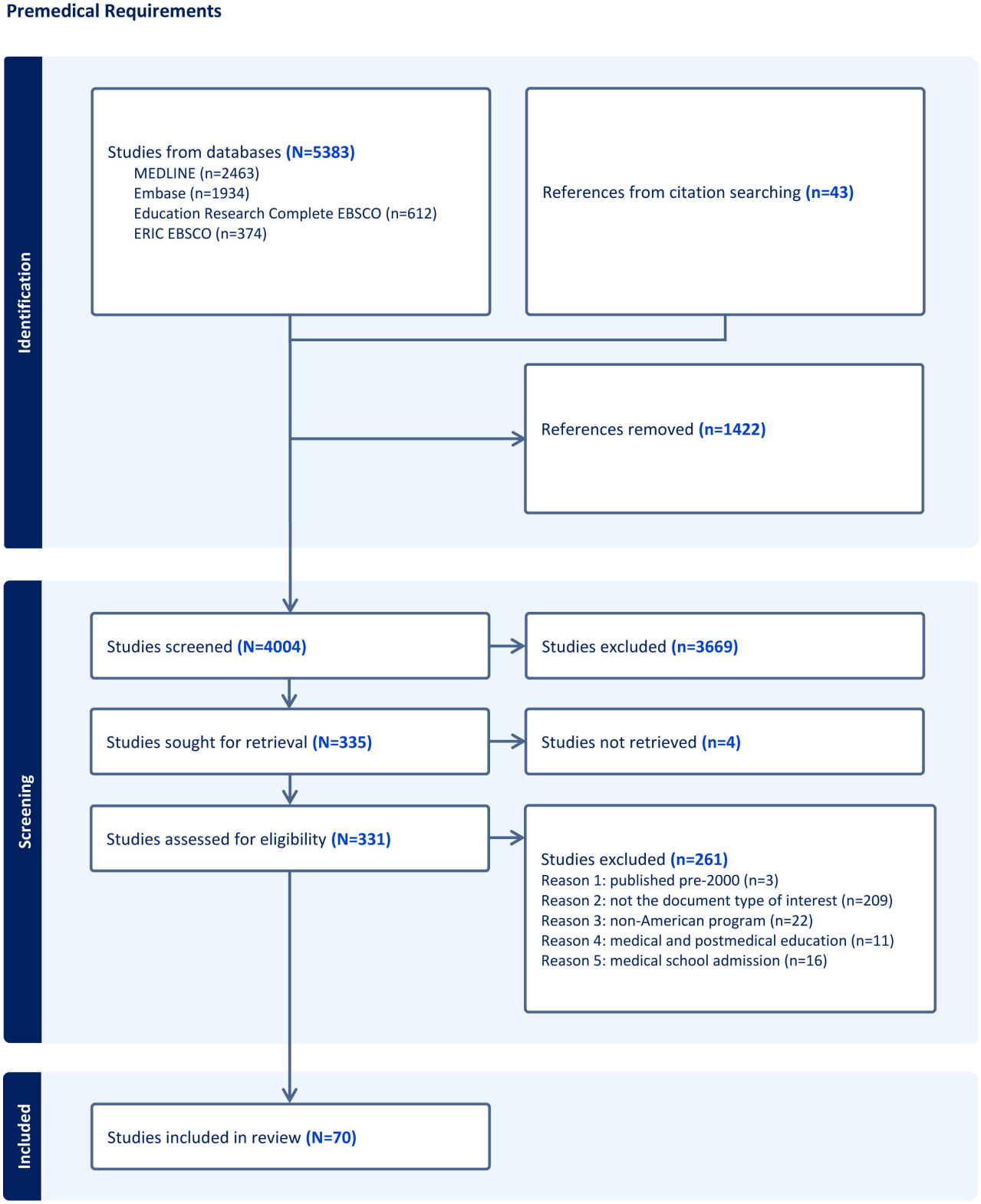
Results

Selection of Sources of Evidence

The search strategy and citation tracking retrieved 4004 studies after duplicate removal. We conducted title and abstract screening, leaving 335 studies for full-text screening. Following full-text screening, 261 studies were excluded for not meeting the selection criteria defined by specific reasons. Three references were excluded under reason 1 (published pre-2000), 209 references were excluded under reason 2 (not the document type of interest), 22 under reason 3 (non-American programs), 11 under reason 4 (medical and postmedical education), and 16 under reason 5 (medical school admission). Four studies lacked accessible full texts and were subsequently excluded at this stage. Accordingly, 70 studies were included in data extraction, analysis, and quality assessment (Figure 2).

XSL•FO
RenderX

Figure 2. PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) flowchart.



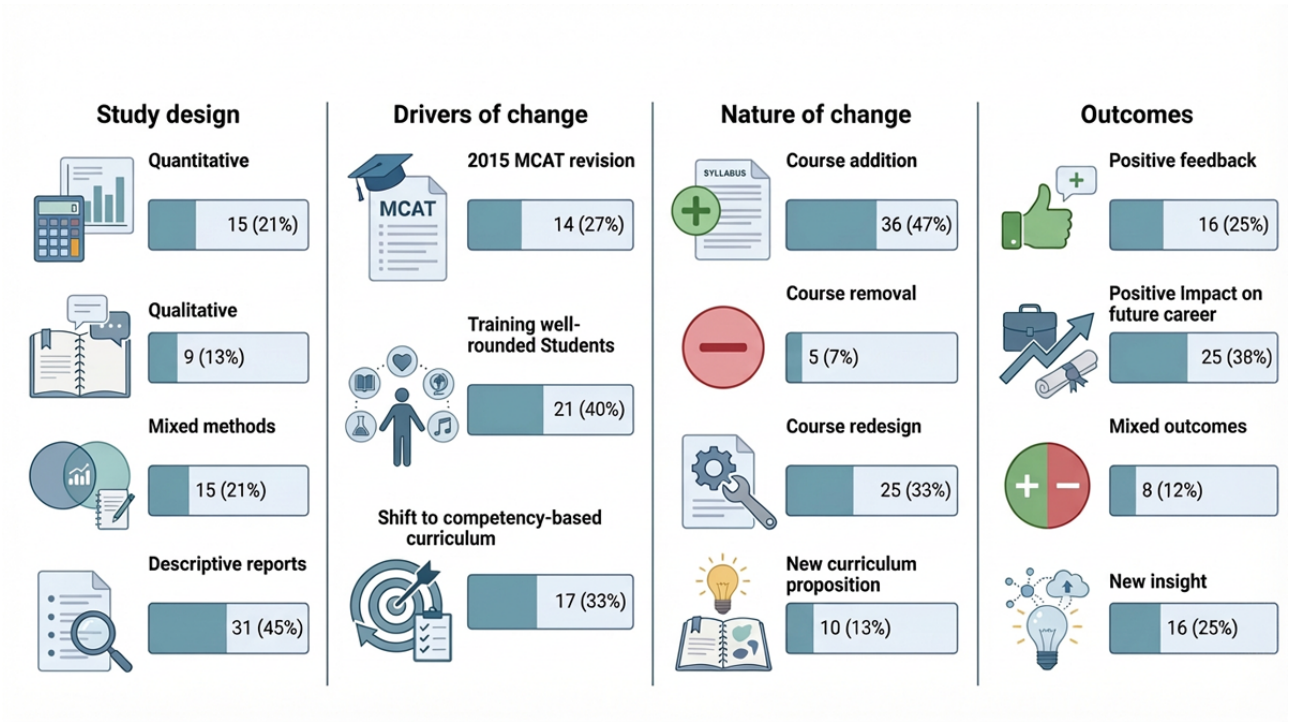
covidence

XSL•FO
RenderX

Characteristics of Sources of Evidence

The included studies were published from 2000 onwards, predominantly in the United States, with occasional exceptions (1 in Australia [16] and 1 in Qatar [17]). Study types comprised 15 mixed methods studies, 15 quantitative studies, 9 qualitative studies, and 31 descriptive reports that were not eligible for MMAT risk of bias assessment. All institutions involved in these studies were universities following the American undergraduate premedical track (Figure 3).

Figure 3. Summary characteristics of studies reviewed (N=70). Percentages are calculated based on the total number of studies included. Counts may not sum to 70 in all categories due to multiple categorizations or rounding. The figure was generated by FigureLabs. MCAT: Medical College Admission Test.



Critical Appraisal Within Sources of Evidence

The remaining 39 studies that qualified as quantitative, qualitative, or mixed methods studies were appraised in Excel using MMAT [15]. Following the risk of bias assessment, 3 studies that did not effectively explain the relevance of their outcomes to course or curriculum reforms were excluded [18-20]. Quality scores ranged from meeting none of the 5 criteria (0) to meeting all 5 criteria (5). The risk of bias assessment results are provided in Multimedia Appendix 6. Each study's design, outcomes, changes to courses, and drivers of change are reported in Multimedia Appendix 5.

Synthesis of Results

Main Drivers of Change

Across the 67 remaining studies, 3 recurring drivers of curriculum reform or course modification were identified. First, the demand for medical schools to accept well-rounded students (n=21, 40%; Figure 3) drove the integration of courses emphasizing empathy [21], patient-physician interactions [22], and general education, such as global health [23]. Second, a shift from course-based to competency-based education (n=17, 33%) encouraged institutions to implement interdisciplinary classes [24] and specialized courses focusing on technical [25] and interpersonal skills [22]. Finally, a significant driver was the 2015 MCAT revision (n=14, 27%; Figure 3), which introduced behavioral and social sciences as well as critical reasoning sections. This motivated institutional reforms, as every college sought to prepare students to succeed in the standardized exam indispensable for continuing a career in medicine. Universities responded by introducing courses in reflective writing [26], psychology [1,10,17,27,28], sociology [7,29,30], and anthropology [3,27].

Main Natures of Change

All the reforms resulting from these convergent institutional motives clustered into 4 categories, thus characterizing the nature of the changes made. Particularly, 36 (47%) studies (Figure 3) focused on adding new courses to the premedical curriculum, in which the most frequent additions were courses in anatomy [31-35], psychology [1,10,17,27,28], statistics [1,27], histology [31,33], and mathematics [2,36]. On the other hand, other studies introduced specialized topics such as neuroimmune pharmacology [37], point-of-care ultrasound [25], childhood obesity [38], microbiology [39], evolution [40-42], and antibiotic resistance [43]. Simultaneously, several studies also emphasized making humanities-based subjects such as global health [27], medical anthropology [3,27], and sociology [17,27-29] required for premedical students to better prepare them for medical school (Multimedia Appendix 5).

From the 67 studies, 25 (33%; Figure 3) addressed redesigning, revising, or replacing existing courses to align with the shift toward competency-based tracks. For example, physics [1,44-50], sociology [17,28,30], biochemistry [40,51-53], and

XSL•FO
RenderX

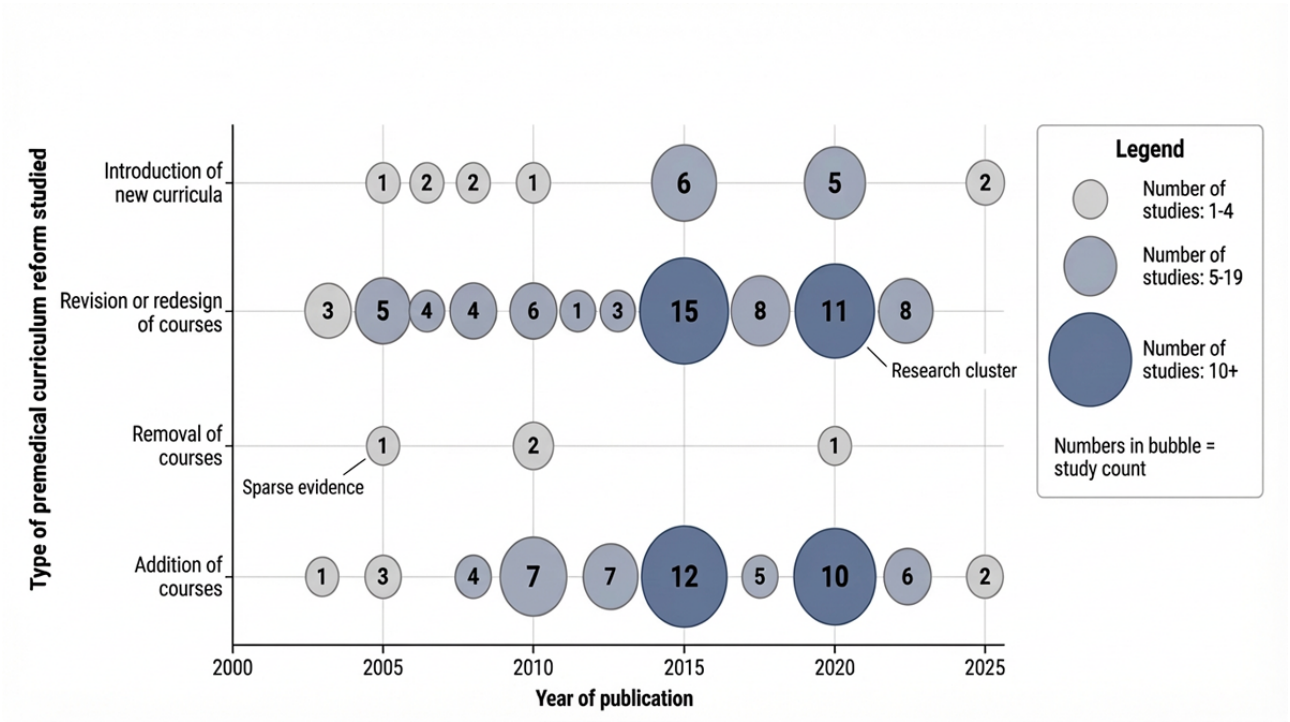
biology [27,40,53,54] courses were most often targeted for redesign because their content was perceived as diverging from their primary purpose: ensuring student readiness for medical school admission. Some institutions proposed redesigning courses by integrating laboratory components (cadaver [55] and dissection anatomy and physiology labs [56,57]) or changing course duration (intensive general chemistry in 1 semester [58]; Multimedia Appendix 5).

Furthermore, 10 (13%) studies (Figure 3) proposed new curricula or competency-based sequences [2,9-11] to help students succeed by building interdisciplinary connections and interpersonal skills. These studies demonstrated that some students could thrive without a traditional science background, as seen in programs such as the Humanities and Medicine Program (HuMed) [59,60]. Particularly, some institutions also designed new majors in biology [61,62] and biochemistry [8] specifically tailored for premedical students and aspiring physicians. Essentially, these programs aimed to achieve a more flexible and comprehensive curriculum that would better fulfill students' needs and cultivate diverse skills (Multimedia Appendix 5). This curriculum included courses tailored to meet premedical students' needs while preparing them for medical school.

It is also critical to note that only 5 out of 67 (7%) studies (Figure 3) examined removing courses from the undergraduate premedical track. For instance, chemistry (organic [2,7] and general [58]) and calculus [7], traditionally viewed as "weed-out" courses by medical school admissions committees, were the main targets. These reports argued that these courses did not benefit students; instead, they harmed them by adding stress to an already competitive premedical journey (Multimedia Appendix 5).

Moreover, an evidence gap map was generated to summarize the data collected from these references and to visualize the distribution of these studies clearly (Figure 4). The map showcases the pattern of changes to premedical requirements worldwide from 2000 onwards. It aids in identifying areas of strong evidence, such as research clusters of studies examining course redesign and additions, particularly after 2010. Additionally, it highlights domains like the removal of courses and the proposition of a new curriculum, where research remains limited and future investigation is needed to further support these findings. Finally, this map shows a temporal increase in evidence availability after 2010, with the highest concentration of studies occurring between 2015 and 2020 (Figure 4).

Figure 4. Evidence gap map of studies examining premedical curriculum reforms in American programs worldwide (2000 - 2025). Each bubble represents the number of studies in a given year and reform category; bubble size and embedded numbers reflect the study count. The figure was generated by FigureLabs.



Main Outcomes

Indeed, a significant pattern of outcomes was retrieved across the 67 analyzed studies, in which a large number emphasized the positive impact of these changes on future career development and medical school preparedness (n=25, 38%; Figure 3) by comparing grade point average [22,32], rate of acceptance [59], step 1 scores [59], as well as future performance in medical school [31,32,60,63]. Additionally, the incorporation of anatomy courses into the premedical curriculum was proven to be a very good predictor of improvement at the start of the medical school journey (Multimedia Appendix 5) [31,32]. Consequently, well-designed premedical courses can significantly support career development and guide students toward a successful path.

Another apparent outcome encountered was essentially positive feedback regarding the course (n=16, 25%; Figure 3), especially in studies that underpinned the interdisciplinary approaches. In

XSL•FO
RenderX

fact, students expressed that integrating different subjects, such as biology and mathematics, was more interactive [24]. Moreover, additional positive feedback was solicited from students who were taught anatomy and physiology through active, hands-on approaches (Multimedia Appendix 5) [55,56]. Additionally, students valued the integration of humanities into the undergraduate premedical years and had a positive attitude regarding this addition to the premedical curriculum [26]. In all the aforementioned cases, the changes served to increase student satisfaction and motivation.

To continue, several studies have proposed new insights into courses and curricula (n=16, 25%; Figure 3), specifically when measurable outcomes were limited. The perceived versus actual use of mathematics in medicine instigates a gap between what students believe about the use of mathematics and what practicing physicians actually report, thereby offering a completely new rationale for curricular reform [36]. Simultaneously, a study by Watkins et al [54] provided valuable lessons about aligning physics curricula with the practices of biomedical science (Multimedia Appendix 5).

Finally, it is important to note that some studies reported negative outcomes as well. Particularly, mixed findings (n=8, 12%; Figure 3) were observed, where benefits were achieved in certain contexts but not in others. For example, after taking undergraduate-level anatomy, undergraduate premedical students concluded that, while they valued anatomy courses, the reform did not lead to measurable, quantified improvements in future academic performance [34]. Essentially, while reforms may fuel engagement and self-perceived readiness, they do not always yield significant changes in traditional academic outcomes such as grades or standardized test scores, and ultimately, medical school preparedness and acceptance (Multimedia Appendix 5) [7,16,35,39,60,64,65].

Discussion

Summary of Evidence

This scoping review maps and synthesizes evidence from 70 studies published from 2000 onwards that discuss reforms to premedical course requirements in American programs worldwide. Our scoping review screens the literature to identify these reforms, classify the types of changes, and evaluate the motives, rationales, and outcomes of these changes for the sake of building a unified foundation for institutions, stakeholders, and students aspiring to delve into medical education. Our findings underscore the most common drivers of curricular change pertaining to the 2015 MCAT revision, the increasing importance of preparing well-trained aspiring physicians, as well as the shift toward competency-based education. Particularly, institutions opted to add new courses such as, but not limited to, anatomy [31-35], psychology [1,10,17,27,28], statistics [1,27], and humanities [17,27-29]; revise traditional ones to integrate clinical or laboratory components; and, in some cases, dilute some long-standing courses such as organic chemistry [2,7] and remove others like calculus [7]. Across these interventions, the prevailing outcome was a perceived or, at times, measurable improvement in student preparedness for medical school.

Interpretations

In fact, the 2015 MCAT revision played a significant role in reshaping the plight of US premedical requirements. These revisions to the MCAT drew scholarly attention to premedical curriculum reforms during this period, explaining the concentration of studies occurring between 2015 and 2020 (Figure 4). The 2015 New MCAT retuned the requirements in terms of preparation for competencies in biochemistry, behavioral and social sciences, and critical analysis, thus nudging programs from legacy checklist models toward competency-oriented design. Indeed, recent reports framed the exam as both a signal of evolving expectations that institutions could use to justify curricular and advising changes [1,28,66]. The 2015 MCAT brought about a shift toward evidence-informed, competency-focused preparation, with the strongest momentum coming from local pilots reporting concrete learning advantages [47,52,67]. Nonetheless, although the MCAT has endured a broader competency profile, misalignment persists between what is tested and what is required. For example, sociology and psychology remain “recommended” rather than required courses at many institutions, leaving students to discover these expectations later through advising or even peers [29]. To illustrate, recent US commentaries also note that applicants can be “MCAT-ready but list-deficient” (or the reverse), depending on the target school, which is essentially an artifact of decentralized governance and historical precedent [1,66,68]. The downstream effect is avoidable inequity: students at campuses with strong prehealth advising or clear catalogs are anticipated to front-load MCAT-tested domains, while others may take the exam without formal exposure to key content [29]. Indeed, evidence from redesigned science sequences (eg, life-science physics and compact general chemistry) suggests that fit-for-purpose courses can close this gap while preserving rigor, but it is undeniable that adoption remains uneven [9,47,58,67].

Furthermore, while it is clear that a paradigm shift in premedical curricula is essential, many programs are reluctant to embrace these changes due to multiple obstacles. Some of the crucial challenges are clustered around legacy course loyalty, departmental identity, coordination costs, and ambiguity regarding equivalencies. Indeed, faculty and staff report concerns about reduced rigor in relation to challenging traditional requirements [59,66]. It is also important to note attempts to integrate premedical preparation with medical curricula, which have reported incentive misalignments, assessment mismatches, and governance friction across undergraduate and professional schools [31]. Finally, anchoring social or behavioral foundations in the early years (with clear MCAT linkages) is essential to bridge the gap between what students learn and how they are assessed by medical school admissions committees [29,35]. To embrace these changes, coordination between medical schools and undergraduate colleges is essential to ensure the implemented changes align with the competencies they look for in potential applicants.

Comparison With Existing Literature

Our results are consistent with previous reports that have questioned the adequacy of traditional premedical curricula

XSL•FO
RenderX

dominated by natural sciences [29,31,41,68]. On one hand, the pattern of reforms unraveled in this review mirrors broader shifts in medical education, shedding light on the integration of the humanities, ethics, and social sciences to cultivate empathy, cultural competence, and critical reasoning [26,38]. On the other hand, reforms were aligned with physics, chemistry, and biology within biomedical contexts, resonating with the literature and stressing the need for disciplinary authenticity—that is, aligning traditional foundational science knowledge with medical applications using interdisciplinary approaches [24,25,43,63]. Consequently, these trends together reflect a shift from content rigid coursework to competency-based curricula that integrate scientific knowledge and concrete applications. This substantiates the hypothesis that targeted course reform can foster not only cognitive readiness but also PIF at the undergraduate stage [21,50]. Hence, premedical students applying to medical school will have undergone a developmental process through which they internalize the values, roles, and responsibilities of the medical profession [6,21,50]. Collectively, these reforms suggest that integrating scientific coursework with humanistic competencies better prepares students for the MCAT exam, medical school admission, and professional practice. Competency-based education emphasizes that professional identity encompasses not only theoretical scientific knowledge but also attributes such as empathy, integrity, resilience, beneficence, and proper patient-centered care [6]. Consequently, retuning the premedical compass expands the concept of readiness beyond academic excellence aligning with the evolution of the MCAT toward assessing social, behavioral, and critical thinking domains.

Strengths

This review has valuable strengths. It is essentially the first, to our knowledge, to comprehensively examine course-level reform in premedical education across 2 and a half decades. The inclusion of both American and international factions following the American program greatly enhances its representativeness. Furthermore, the large number of studies that were included, combined with the systematic methodology used to find these studies, provides a strong foundation for synthesizing fragmented literature.

Limitations

However, limitations must be acknowledged; considerable heterogeneity in study designs, outcomes, and reporting limited quantitative synthesis. A significant portion of the studies was not eligible for MMAT critical appraisal; many relied on self-reported student feedback or descriptive accounts without rigorous statistical evaluation, thereby raising concerns about response and publication bias [21,25,26,29]. In addition, studies addressing the removal of courses are scarce over the entire 25-year period, and evidence appears to be very limited. Furthermore, research on the proposition of new curricula remains relatively limited compared to other types of changes (Figure 4). This indicates that further investigation is needed to fully determine which courses need to be removed and whether proposing new curricula is a good approach to best prepare premedical students for medical school.

Implications

Regardless of these caveats, the findings hold tangible implications. For students, greater transparency in premedical requirements afforded by institutions to benefit students could reduce uncertainty and promote earlier academic planning [65,68]. For educators and institutions, greater transparency and better implementation of changes based on evidence could improve the quality of education provided to these premedical students [18,41,45]. At a policy level, medical schools and accrediting bodies should consider harmonizing admission requirements with both MCAT content and undergraduate course offerings to reduce misalignment and student burden [20,27,60]. Practically, longitudinal research points toward understanding the essentiality of the link between undergraduate course exposure to medical school and career outcomes [31]. Essentially, more evidence is needed to quantify to what extent the changes in these courses were implemented to provide students with a comprehensive foundation for medical school. That is why future research should prioritize prospective, longitudinal studies that measure both academic and nonacademic outcomes, including well-being and patient-related competencies [69-75].

Conclusions

In conclusion, our review encompasses all the changes fragmented in the literature and accentuates that targeted reforms to premedical course requirements can enhance preparedness for medical school. However, it is critical to note that their impact is uneven and context dependent. The latter necessitates a rather balanced approach, integrating rigorous scientific foundations with the humanities and social sciences, which appears most promising for shaping the next generation of physicians prepared both academically and professionally. This review is innovative because it moves beyond isolated institutional descriptions and commentary by consolidating 70 sources into a single evidence map of premedical course–requirement reform since 2000 across American programs worldwide, explicitly linking reform types to their stated drivers and outcomes. Specifically, what this review brings to the field is a transferable and dynamic framework that helps interpret a fragmented literature and highlights persistent misalignment between the discrepancy in what is being tested, taught, and required—a pertinent issue with equity consequences when access to advising and coursework varies across institutions. In reality, these findings can inform curriculum committees and program leaders designing updated premedical tracks, support advising strategies that anticipate heterogeneous medical school requirements, and provide admissions stakeholders with a clearer understanding of how prerequisite policies interact with competency-focused readiness, thus helping to reduce avoidable gaps in preparation for students across diverse educational contexts. Therefore, essential future work aligning prerequisites to 2015 MCAT competencies, coupled with advising clarity and outcome reporting to understand the extent to which these course requirements have been implemented, would significantly offer a pathway to equity, efficiency, and preparedness without sacrificing scientific rigor.

XSL•FO
RenderX

Acknowledgments

The authors thank the staff members of the Saab Medical Library at the American University of Beirut for their assistance. All authors declared that they had insufficient funding to support open access publication of this manuscript, including from affiliated organizations or institutions, funding agencies, or other organizations. JMIR Publications provided article processing fee support for the publication of this article. The authors declare the use of generative artificial intelligence (GAI) to produce Figures 3 and 4, and the table of contents image. According to the GAIDeT (Generative AI Delegation Taxonomy; 2025), the following task was delegated to GAI tools under full human supervision: visualization. The GAI tool used was FigureLabs. Responsibility for the final manuscript lies entirely with the authors. GAI tools are not listed as authors and do not bear responsibility for the final outcomes.

Funding

No external financial support or grants were received from any public, commercial, or not-for-profit entities for the research, authorship, or publication of this manuscript.

Data Availability

All data used for the study have been included in the manuscript and multimedia appendices.

Authors' Contributions

BRK and OZ contributed equally as corresponding authors to this manuscript. BRK and OZ designed the study. OZ conducted the literature search. MEC, LIBS, and CIZ reviewed the titles, abstracts, and full-text screenings. All authors participated in regular meetings to interpret the studies and contribute to manuscript writing. All authors were involved in drafting the manuscript and have read and approved the final version of the manuscript. All authors have agreed to be accountable for all aspects of the work.

Conflicts of Interest

None declared.

Multimedia Appendix 1
The full MEDLINE search terms used in the review.
[DOC File, 28 KB - mededu_v12i1e85002_app1.doc]

Multimedia Appendix 2
The full Embase search terms used in the review.
[DOCX File, 12 KB - mededu_v12i1e85002_app2.docx]

Multimedia Appendix 3
The full ERIC search terms used in the review.
[DOCX File, 13 KB - mededu_v12i1e85002_app3.docx]

Multimedia Appendix 4
The full Education Research Complete search terms used in the review.
[DOCX File, 12 KB - mededu_v12i1e85002_app4.docx]

Multimedia Appendix 5
Comprehensive data extraction table.
[PDF File, 225 KB - mededu_v12i1e85002_app5.pdf]

Multimedia Appendix 6
Mixed Methods Appraisal Tool (MMAT; McGill University) table.
[PDF File, 118 KB - mededu_v12i1e85002_app6.pdf]

Checklist 1
PRISMA-ScR checklist.
[DOCX File, 86 KB - mededu_v12i1e85002_app7.docx]

XSL•FO
RenderX

Checklist 2
PRISMA-S checklist
[DOCX File, 17 KB - mededu_v12i1e85002_app8.docx]

References

1. Chambers DA, Cohen RL, Girotti J. A century of premedical education. Perspect Biol Med 2011;54(1):17-23. [doi: 10.1353/pbm.2011.0011] [Medline: 21399379]
2. Alpern RJ, Belitsky R, Long S. Competencies in premedical and medical education: the AAMC-HHMI report. Perspect Biol Med 2011;54(1):30-35. [doi: 10.1353/pbm.2011.0001] [Medline: 21399381]
3. Stodola TJ, Bruna SP. Medical anthropology courses and concepts tested on the mcat: a content analysis of 40 U.S. course syllabi. J Med Educ Curric Dev 2021;8:23821205211010812. [doi: 10.1177/23821205211010812] [Medline: 33997288]
4. Kaplan RM, Satterfield JM, Kington RS. Building a better physician—the case for the new MCAT. N Engl J Med 2012 Apr 5;366(14):1265-1268. [doi: 10.1056/NEJMp1113274] [Medline: 22475589]
5. Hanson JT, Busche K, Elks ML, et al. The validity of MCAT scores in predicting students' performance and progress in medical school: results from a multisite study. Acad Med 2022 Sep 1;97(9):1374-1384. [doi: 10.1097/ACM.0000000000004754] [Medline: 35612915]
6. Merlo G, Ryu H, Harris TB, Coverdale J. MPRO: A professionalism curriculum to enhance the professional identity formation of university premedical students. Med Educ Online 2021 Dec;26(1):1886224. [doi: 10.1080/10872981.2021.1886224] [Medline: 33606590]
7. Wershof Schwartz A, Abramson JS, Wojnowich I, Accordino R, Ronan EJ, Rifkin MR. Evaluating the impact of the humanities in medical education. Mt Sinai J Med 2009 Aug;76(4):372-380. [doi: 10.1002/msj.20126] [Medline: 19642151]
8. Fried C, Madar S, Donley C. The Biomedical Humanities program: merging humanities and science in a premedical curriculum at Hiram College. Acad Med 2003 Oct;78(10):993-996. [doi: 10.1097/00001888-200310000-00010] [Medline: 14534095]
9. Schnoebelen C, Towns MH, Chmielewski J, Hrycyna CA. Design and evaluation of a one-semester general chemistry course for undergraduate life science majors. J Chem Educ 2018;95(5):734-740. [doi: 10.1021/acs.jchemed.7b00869]
10. Fishbein RH. Origins of modern premedical education. Acad Med 2001 May;76(5):425-429. [doi: 10.1097/00001888-200105000-00009] [Medline: 11346515]
11. Wasserman JA, Lopez AA, Knickerbocker R, et al. What premedical students need to succeed: updated premed competencies for entering medical students. Acad Med 2025 Dec 1;100(12):1452-1458. [doi: 10.1097/ACM.0000000000006065] [Medline: 40249857]
12. Tricco AC, Lillie E, Zarin W, et al. PRISMA extension for scoping reviews (PRISMA-ScR): checklist and explanation. Ann Intern Med 2018 Oct 2;169(7):467-473. [doi: 10.7326/M18-0850] [Medline: 30178033]
13. Rethlefsen ML, Kirtley S, Waffenschmidt S, et al. PRISMA-S: an extension to the PRISMA statement for reporting literature searches in systematic reviews. Syst Rev 2021 Jan 26;10(1):39. [doi: 10.1186/s13643-020-01542-z] [Medline: 33499930]
14. Babineau J. Product review: Covidence (systematic review software). J Can Health Libr Assoc 2014;35(2):68. [doi: 10.5596/c14-016]
15. Hong QN, Fàbregues S, Bartlett G, et al. The Mixed Methods Appraisal Tool (MMAT) version 2018 for information professionals and researchers. Education Inf 2018;34(4):285-291. [doi: 10.3233/EFI-180221]
16. Moro C, McLean M, Phelps C. Doing justice when incorporating sustainability into pre-medical curricula. Med Educ 2025 Jun;59(6):658-659. [doi: 10.1111/medu.15653] [Medline: 40022590]
17. del Pozo PR, Fins JJ. The globalization of education in medical ethics and humanities: evolving pedagogy at Weill Cornell Medical College in Qatar. Acad Med 2005 Feb;80(2):135-140. [doi: 10.1097/00001888-200502000-00005] [Medline: 15671316]
18. Stevens C, Witkow MR. Training scientific thinking skills: evidence from an MCAT2015-aligned classroom module. Teach Psychol 2014;41(2):115-121. [doi: 10.1177/0098628314530341]
19. Stowe RL, Cooper MM. Practicing what we preach: assessing "Critical Thinking" in organic chemistry. J Chem Educ 2017 Dec 12;94(12):1852-1859. [doi: 10.1021/acs.jchemed.7b00335]
20. Barr DA, Matsui J. "Turning point" for minority pre-meds: the effect of early undergraduate experience in the sciences on aspirations to enter medical school of minority students at UC Berkeley and Stanford University. : Center for Studies in Higher Education, UC Berkeley; 2008 URL: https://cshe.berkeley.edu/sites/default/files/publications/2008th1.pdf [accessed 2026-03-25]
21. Vongsachang H, Oshima S, Nguyen C, Gaulocher S, Wang NE. Undergraduate premedical student perceptions of an emergency department–based social needs screening program. Med Sci Educ 2020 Jun;30(2):673-677. [doi: 10.1007/s40670-020-00936-x] [Medline: 34457723]
22. Wilson JW, Baer RD, Villalona S. Patient shadowing: a useful research method, teaching tool, and approach to student professional development for premedical undergraduates. Acad Med 2019 Nov;94(11):1722-1727. [doi: 10.1097/ACM.0000000000002882] [Medline: 31335819]

XSL•FO
RenderX

23. Kolachalama VB. Machine learning and pre-medical education. Artif Intell Med 2022 Jul;129:102313. [doi: 10.1016/j.artmed.2022.102313] [Medline: 35659392]
24. Ryan N, Wallace D. Applications of Calculus to Biology and Medicine: Case Studies from Lake Victoria: World Scientific Publishing; 2017.
25. Smalley CM, Browne V, Kaplan B, Russ B, Wilson J, Lewiss RE. Early innovative immersion: a course for pre-medical professions students using point-of-care ultrasound.. J Ultrasound Med 2016 Dec;35(12):2681-2686. [doi: 10.7863/ultra.16.01050]
26. Bracken RC, Major A, Paul A, Ostherr K. Reflective writing about near-peer blogs: a novel method for introducing the medical humanities in premedical education. J Med Humanit 2021 Dec;42(4):535-569. [doi: 10.1007/s10912-021-09693-3] [Medline: 33871756]
27. Ferren AS, Paris DC. Faculty leadership in premedical education. Peer Rev 2012;14(4):18-20 [FREE Full text]
28. Mitchell K, Lewis RS, Satterfield J, Hong BA. The new Medical College Admission Test: implications for teaching psychology. Am Psychol 2016;71(2):125-135. [doi: 10.1037/a0039975] [Medline: 26866988]
29. Olsen LD. “It’s on the MCAT for a Reason”: premedical students and the perceived utility of sociology. Teach Sociol 2016;44(2):72-83. [doi: 10.1177/0092055X15624744]
30. Lotspeich-Yadao MC, Trevino-Boissel K, Hoekstra C. Integrating experiential learning in introduction to sociology: cultivating the sociological imagination in premedicine students. Teach Sociol 2025 Apr;53(2):110-125. [doi: 10.1177/0092055X241268787]
31. Forester JP, McWhorter DL, Cole MS. The relationship between premedical coursework in gross anatomy and histology and medical school performance in gross anatomy and histology. Clin Anat 2002 Mar;15(2):160-164. [doi: 10.1002/ca.1114] [Medline: 11877798]
32. Kondrashov P, McDaniel DJ, Jordan RM. Premedical anatomy experience and student performance in medical gross anatomy. Clin Anat 2017 Apr;30(3):303-311. [doi: 10.1002/ca.22846] [Medline: 28192872]
33. Miller SA, Perrotti W, Silverthorn DU, Dalley AF, Rarey KE. From college to clinic: reasoning over memorization is key for understanding anatomy. Anat Rec 2002 Apr 15;269(2):69-80. [doi: 10.1002/ar.10071] [Medline: 12001213]
34. Louro MD, Meegan G, Rudin LR, Granatosky MC, Thompson NE. Students with prior anatomy experience start out stronger in medical school gross anatomy. Anat Sci Educ 2024 Oct;17(7):1406-1416. [doi: 10.1002/ase.2495] [Medline: 39075733]
35. Thomas R, Yancey T, Skidmore C, et al. The effects of pre-medical anatomy and clinical experiences on medical school anatomy-related academic performance. Med Sci Educ 2021;31(6):1839-1849. [doi: 10.1007/s40670-021-01372-1] [Medline: 34956700]
36. Savic M, Martin P. The perceived versus actual use of mathematics in medicine according to pre-medicine students and practicing physicians. Teach Math Appl 2017;37(4):192-201. [doi: 10.1093/teamat/hrx011]
37. Buch SJ. Neuroimmune pharmacology as an emerging curriculum for pre-medical students. J Neuroimmune Pharmacol 2011 Mar;6(1):68-70. [doi: 10.1007/s11481-010-9229-4] [Medline: 20607431]
38. Cooke NK, Wilson KL, Bissett EE, Goodell LS. A qualitative assessment of undergraduate pre‐healthcare student views of childhood obesity. FASEB J 2011 Apr;25(S1). [doi: 10.1096/fasebj.25.1_supplement.99.6]
39. Ramos RL, Guercio E, Martinez LR. Pre-medical preparation in microbiology among applicants and matriculants in osteopathic medical school in the United States. J Microbiol Biol Educ 2017;18(3):18.3.61. [doi: 10.1128/jmbe.v18i3.1372] [Medline: 29854054]
40. Moltzau Anderson J, Horn F. (Re-) Defining evolutionary medicine. Ecol Evol 2020 Oct;10(20):10930-10936. [doi: 10.1002/ece3.6825] [Medline: 33144938]
41. Antolin MF, Jenkins KP, Bergstrom CT, et al. Evolution and medicine in undergraduate education: a prescription for all biology students. Evolution 2012 Jun;66(6):1991-2006. [doi: 10.1111/j.1558-5646.2011.01552.x] [Medline: 22671563]
42. Nesse RM, Bergstrom CT, Ellison PT, et al. Evolution in health and medicine Sackler colloquium: making evolutionary biology a basic science for medicine. Proc Natl Acad Sci U S A 2010 Jan 26;107 Suppl 1(Suppl 1):1800-1807. [doi: 10.1073/pnas.0906224106] [Medline: 19918069]
43. Da Silva Carvalho C. Knowledge and perception of antibiotic resistance and stewardship among pre-health and agriculture undergraduate students. J Microbiol Biol Educ 2024 Dec 12;25(3):e0006924. [doi: 10.1128/jmbe.00069-24] [Medline: 39291990]
44. Cahn SB, Mochrie SGJ. Biologic: gene circuits and feedback in an introductory physics sequence for biology and premedical students. Am J Phys 2014 May;82(5):412-421. [doi: 10.1119/1.4866290]
45. Mochrie SGJ. Vision and change in introductory physics for the life sciences. arXiv. Preprint posted online on Oct 2, 2015. [doi: 10.48550/arXiv.1510.00753]
46. Crouch CH, Heller K. Introductory physics in biological context: an approach to improve introductory physics for life science students. Am J Phys 2014 May 1;82(5):378-386. [doi: 10.1119/1.4870079]
47. Redish EF, Bauer C, Carleton KL, et al. NEXUS/Physics: An interdisciplinary repurposing of physics for biologists. Am J Phys 2014 May 1;82(5):368-377. [doi: 10.1119/1.4870386]
48. Hilborn RC. Physics and the revised Medical College Admission Test. Am J Phys 2014 May 1;82(5):428-433. [doi: 10.1119/1.4862138]

XSL•FO
RenderX

49. Kortemeyer G. The challenge of teaching introductory physics to premedical students. Phys Teach 2007 Dec 1;45(9):552-557. [doi: 10.1119/1.2809149]
50. Mylott E, Kutschera E, Dunlap JC, Christensen W, Widenhorn R. Using biomedically relevant multimedia content in an introductory physics course for life science and pre-health students. J Sci Educ Technol 2016 Apr;25(2):222-231. [doi: 10.1007/s10956-015-9588-y]
51. Stover S. Integrating case studies into a pre-medical curriculum. Bioscene 2020 May;46:23-26 [FREE Full text]
52. Bolduc KC, McCollough SL, Stoeckman AK. From classroom to clinic: biochemistry lab for pre-health majors. Biochem Mol Biol Educ 2023 Jan;51(1):10-14. [doi: 10.1002/bmb.21675] [Medline: 36102284]
53. Brenner C. Changes in chemistry and biochemistry education: creative responses to medical college admissions test revisions in the age of the genome. Biochem Mol Biol Educ 2013;41(1):1-4. [doi: 10.1002/bmb.20653] [Medline: 23281187]
54. Watkins J, Coffey JE, Redish EF, Cooke TJ. Disciplinary authenticity: enriching the reforms of introductory physics courses for life-science students. Phys Rev ST Phys Educ Res 2012 Apr;8(1):010112. [doi: 10.1103/PhysRevSTPER.8.010112]
55. Afghani B, Manalang J, Salazar A, et al. Should cadaver lab be introduced to premedical students in an effort to promote their interest towards health careers? . 2013 Jan(1). [doi: 10.2310/JIM.0b013e31827d3ac9]
56. Harris DM, Bellew C, Cheng ZJ, Cendán JC, Kibble JD. High-fidelity patient simulators to expose undergraduate students to the clinical relevance of physiology concepts. Adv Physiol Educ 2014 Dec;38(4):372-375. [doi: 10.1152/advan.00063.2014] [Medline: 25434023]
57. Hyatt JPK, Hurst SA. Novel undergraduate physiology laboratory using a human patient simulator. Med Educ 2010 May;44(5):523-523. [doi: 10.1111/j.1365-2923.2010.03651.x] [Medline: 20374473]
58. Kennerly WW, Frederick KA, Sheppard K. General chemistry in just one semester for all majors. J Chem Educ 2020 May 12;97(5):1295-1302. [doi: 10.1021/acs.jchemed.9b00898]
59. Muller D, Kase N. Challenging traditional premedical requirements as predictors of success in medical school: the Mount Sinai School of Medicine Humanities and Medicine Program. Acad Med 2010 Aug;85(8):1378-1383. [doi: 10.1097/ACM.0b013e3181dbf22a] [Medline: 20671464]
60. Muller D. FlexMed: a nontraditional admissions program at Icahn School of Medicine at Mount Sinai. Virtual Mentor 2014 Aug 1;16(8):614-617. [doi: 10.1001/virtualmentor.2014.16.08.medu2-1408] [Medline: 25140682]
61. Goodman BE, Koster KL, Swanson DL. The development and implementation of a new medical biology major including physiology. Adv Physiol Educ 2015 Jun;39(2):67-75. [doi: 10.1152/advan.00010.2015] [Medline: 26031721]
62. Baldi B, Utts J. What your future doctor should know about statistics: must-include topics for introductory undergraduate biostatistics. Am Stat 2015;69(3):231-240. [doi: 10.1080/00031305.2015.1048903]
63. Tran J, Meller L, Le V, Tam J, Nicholas A. Behavioral assessment of soft skill development in a highly structured pre-health biology course for undergraduates. J Microbiol Biol Educ 2024 Aug 29;25(2):e0019223. [doi: 10.1128/jmbe.00192-23] [Medline: 38860778]
64. Peterson CA, Tucker RP. Undergraduate coursework in anatomy as a predictor of performance: comparison between students taking a medical gross anatomy course of average length and a course shortened by curriculum reform. Clin Anat 2005 Oct;18(7):540-547. [doi: 10.1002/ca.20154] [Medline: 16015636]
65. Kadavakollu S, Moullet Z, Yoshida M, Qureshi M, Graneto J, Boyanovsky B. Undergraduate pre-requisite coursework: six important tips for pre-medical students considering osteopathic medical school in the USA. Int J Osteopath Med 2021 Dec;42:34-42. [doi: 10.1016/j.ijosm.2021.10.009]
66. Haburchak DR, Mitchell BC, Boomer CJ. Quixotic medicine: physical and economic laws perilously disregarded in health care and medical education. Acad Med 2008 Dec;83(12):1140-1145. [doi: 10.1097/ACM.0b013e31818c65c0] [Medline: 19202481]
67. Young SC, Colabroy KL, Baar MR. Comparable educational benefits in half the time: an alternating organic chemistry laboratory sequence targeting prehealth students. J Chem Educ 2016 Dec 13;93(12):2004-2011. [doi: 10.1021/acs.jchemed.6b00254]
68. Malvitz M, Khan N, Morgenstern LB. College preparation for a medical career in the United States. PLoS One 2024;19(2):e0298203. [doi: 10.1371/journal.pone.0298203] [Medline: 38349896]
69. Charon R. Reading, writing, and doctoring: literature and medicine. Am J Med Sci 2000 May;319(5):285-291. [doi: 10.1097/00000441-200005000-00004] [Medline: 10830551]
70. Dixson L, Pomales B, Hashemzadeh M, Hashemzadeh M. Is organic chemistry helpful for basic understanding of disease and medical education? J Chem Educ 2022 Feb 8;99(2):688-693. [doi: 10.1021/acs.jchemed.1c00772]
71. Gishen K, Ovadia S, Arzillo S, Avashia Y, Thaller SR. The current format and ongoing advances of medical education in the United States. J Craniofac Surg 2014 Jan;25(1):35-38. [doi: 10.1097/01.scs.0000436737.67665.ab] [Medline: 24406555]
72. Harder B. Are MOOCs the future of medical education? BMJ 2013 Apr 26;346(apr26 2):f2666. [doi: 10.1136/bmj.f2666] [Medline: 23624666]
73. Sainfort F, Jacko JA, Cohen K, Rosman A, Vieweg J. Disciplinary intersection of medicine and business: a novel population health management premedical pathway for medicine and other health care professions. Popul Health Manag 2022 Oct;25(5):616-624. [doi: 10.1089/pop.2022.0090] [Medline: 35763303]

XSL•FO
RenderX

74. Silverthorn DU, Galey WR, Dalley AE, Bond JS. Medical and premedical scientific competencies: opportunity and challenge. FASEB J 2009 Apr;23. [doi: 10.1096/fasebj.23.1_supplement.LB167]
75. Young AE. Foreword. Obstet Gynecol 2015 Oct;126 Suppl 4:i. [doi: 10.1097/AOG.0000000000001045] [Medline: 26375559]

Abbreviations

MCAT: Medical College Admission Test
MeSH: Medical Subject Headings
MMAT: Mixed Methods Appraisal Tool
PIF: professional identity formation
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
PRISMA-S: Preferred Reporting Items for Systematic Reviews and Meta-Analyses Literature Search Extension
PRISMA-ScR: Preferred Reporting Items for Systematic Reviews and Meta-Analyses Extension for Scoping Reviews

Edited by S Brini, T Leung; submitted 29.Sep.2025; peer-reviewed by E Rincon, Z Dimassi; revised version received 12.Mar.2026; accepted 12.Mar.2026; published 09.Apr.2026.

Please cite as:
Chouairy ME, Bou Samra LI, Zeidan CI, El Zein O, Kaafarani BR
Retuning the Premedical Compass in American Programs Worldwide: Scoping Review
JMIR Med Educ 2026;12:e85002
URL: https://mededu.jmir.org/2026/1/e85002
doi:10.2196/85002

© Michelle E Chouairy, Laetitia I Bou Samra, Christelle I Zeidan, Ola El Zein, Bilal R Kaafarani. Originally published in JMIR Medical Education (https://mededu.jmir.org), 9.Apr.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Education, is properly cited. The complete bibliographic information, a link to the original publication on https://mededu.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Use of Digital Technology for Developing Communication Skills in Undergraduate and Postgraduate Medical Education: Scoping Review

Princella Seripenah[1], MBBS, MPH; Heidi Emery[1], MBBS, MPH; Bakula Patel[2], MBChB; Edward Tyrrell[1], BMBS, BMedSci, MSc; Julie Carson[1], BSc, BMBS, PhD; Jo Leonardi-Bee[3,4], BSc, MSc, PhD; Catrin Evans[4], BSc, MA, PhD; Emma Wilson[3,4], MA, MPH, PhD; Jaspal Taggar[1], MBChB, MSc, PhD

[1]Primary Care Education Unit, Centre for Academic Primary Care, School of Medicine, University of Nottingham, Room C37, C Floor, The Medical School, Queens Medical Centre, Nottingham, United Kingdom

[2]Education Centre, School of Medicine, University of Nottingham, Nottingham, United Kingdom

[3]Centre for Public Health and Epidemiology, School of Medicine, University of Nottingham, Nottingham, United Kingdom

[4]Centre for Evidence Based Healthcare, Faculty of Medicine and Health Sciences, University of Nottingham, Nottingham, United Kingdom

Corresponding Author:
Jaspal Taggar, MBChB, MSc, PhD
Primary Care Education Unit, Centre for Academic Primary Care, School of Medicine, University of Nottingham, Room C37, C Floor, The Medical School, Queens Medical Centre, Nottingham, United Kingdom

Abstract

Background: Effective doctor-patient communication is fundamental to safe, high-quality health care and is a core competency across undergraduate and postgraduate medical education. Communication skills training (CST) has traditionally relied on workforce-intensive methods such as role-play and standardized patient encounters, which face increasing pressure from rising student numbers, constrained faculty capacity, and growing clinical workloads. Digital technologies offer scalable, flexible alternatives, yet the extent, educational design, and strength of evidence supporting digital CST remain unclear.

Objective: This study aimed to comprehensively map the digital technologies used for CST in undergraduate and postgraduate medical education, examine the reported outcomes in the context of educational theory, and identify gaps relevant for future research and clinical practice.

Methods: This scoping review followed Joanna Briggs Institute (JBI) methodology and is reported in accordance with PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses Extension for Scoping Reviews) guidelines. Four electronic databases (Medline, Embase, CINAHL, ERIC) were searched from inception to January 5, 2026. Eligible studies examined any digital technology used to support active, 2-way CST for undergraduate or postgraduate medical learners. Passive learning approaches were excluded. Data were synthesized descriptively. To support structured interpretation of heterogeneous outcomes, interventions were mapped to Kolb's experiential learning cycle to examine learning processes and to Kirkpatrick's evaluation model to assess depth of educational and translational impact.

Results: A total of 11,179 records were identified, of which 121 studies met the inclusion criteria. Most studies were published within the past decade (92/121, 76%) and were conducted in North America and Europe (93/121, 76.9%), with 58.7% (71/121) of studies focusing on undergraduate learners. Recording-based tools (51/121, 41.8%), live stream platforms (33/121, 27%), and virtual patient simulators (32/121, 26.2%) were the most used digital technologies. General communication and history taking was the most frequent topic taught. Only 28.1% (34/121) of studies used validated objective outcome measures. Educationally, digital interventions overwhelmingly supported early stages of experiential learning (120/121, 99.2%), with almost no progression to abstract conceptualization or active experimentation. Outcome evaluation was similarly limited in depth; most studies assessed outcomes at Kirkpatrick Levels 1 and 2. Few studies evaluated behavior change in clinical practice (6/121, 5%) or patient-level outcomes (1/121, 0.8%). A small but growing subset of studies incorporated artificial intelligence, primarily within virtual patient simulators, showing promising but methodologically limited evidence.

Conclusions: Although digital CST interventions show promise for supporting early-stage learning outcomes, the evidence is constrained by weak study designs, inconsistent use of validated measures, and minimal real-world evaluation. Current technologies support only initial phases of experiential learning, with no evidence of progression to competency development or translation into improved patient care. For educators investing in digital CST, these technologies should be integrated thoughtfully within broader curricula rather than treated as standalone solutions, accompanied by evaluation extending to clinical outcomes. Future research that prioritizes robust comparative designs evaluating whether digital training meaningfully improves clinical communication and patient care is warranted.

XSL•FO
RenderX

(JMIR Med Educ 2026;12:e87012) doi:10.2196/87012

KEYWORDS

digital technology; communication skills; medical education; undergraduate training; postgraduate training; doctor–patient communication; virtual simulation; video-based learning; educational technology; scoping review.

Introduction

Background

Effective doctor-patient communication is fundamental to the delivery of safe and high-quality care. Strong communication enhances patient outcomes by building trust, collaborative decision-making, improved treatment adherence, and better self-care [1]. This importance is reflected in empirical evidence showing that poorer physician communication is associated with higher risks of patient nonadherence, while communication skills training (CST) is linked to meaningful improvements in adherence and other patient-relevant outcomes [2]. Furthermore, interpersonal communication is a core graduate competency for medical training programs globally, and the need to develop and refine communication skills throughout a medical career is recognized as a critical professional standard [3,4]. However, many communication skills are not acquired spontaneously during medical training and require training for their development [5].

Methods of teaching communication skills during medical education have traditionally relied on didactic lectures, role-playing, and standardized patient interactions facilitated by more senior clinicians [6]. These are workforce-intensive methods, particularly when personalized, specific feedback is provided, which has been shown to be effective in improving skill acquisition [6]. Pressures on effective communication skills teaching are further compounded by increasing medical student numbers [7], global shortage of health professionals [8,9], and rising clinical workloads among supervisors [10].

Digital technologies have therefore been increasingly adopted in medical education as a means of addressing these structural constraints. Unlike traditional communication skills teaching, which depends on synchronous, staff-intensive delivery, digital approaches such as virtual patient (VP) simulators and online modules require substantial upfront development but can subsequently be reused at scale, offer flexible and repeated access, and enable learners to practice a wider range of communication scenarios than is feasible in conventional formats [11-13]. These approaches support self-directed learning and allow simulation of interactions that are difficult to stage with standardized patients, including communication involving rare or complex clinical presentations. As a result, digital technologies are increasingly viewed as a means of expanding training capacity and equity while reducing reliance on faculty-intensive delivery and preserving patient safety [11].

There has been a rapid acceleration in the use of digital technology within medical education since 2018, particularly following the shift toward virtual learning environments during the COVID-19 pandemic [14,15], though its application specifically for teaching and learning communication skills remains less well evidenced. Emerging digital modalities, including immersive virtual reality and artificial intelligence (AI)–supported interventions, are increasingly being explored for CST, reflecting a widening technological landscape beyond conventional online or recording-based approaches [16]. Previous reviews examining digital approaches to CST in medical education have tended to adopt narrow foci. For example, Kyaw et al [11] restricted inclusion to randomized controlled trials and focused primarily on knowledge and skills outcomes without examining learning processes. Similarly, Fernández-Alcántara et al [17] concentrated exclusively on virtual simulation tools. Reviews of emerging technologies such as AI [18] have similarly focused on specific modalities rather than examining outcomes across the broader landscape of digital approaches. Of particular importance in the context of clinical education, no previous research has systematically examined how digital interventions align with established educational learning theories or assessed whether evaluation extends beyond educational settings to measure impact on clinical practice.

Educational theory provides a structured lens for interpreting heterogeneous outcomes across diverse digital interventions, enabling appraisal not only of whether learning occurs but how learning is supported and whether it progresses beyond individual knowledge acquisition toward behavioral and practice-level impact [19]. Without such theoretical grounding, the evidence base remains fragmented by technology type rather than organized around learning mechanisms and translational relevance. This scoping review is informed by two well-established theoretical frameworks in medical education that have not previously been integrated within CST. Kolb's experiential learning cycle [20] is particularly relevant to CST as it captures how learners develop communication competencies through concrete experience, reflection, conceptualization, and experimentation. Kirkpatrick's evaluation model [21] complements this by providing a framework to assess the depth and real-world impact of interventions, from learner reactions through to practice-level outcomes. Together, these frameworks provide a dual-lens approach that bridges individual learning processes with clinical translation.

Research Aim

This scoping review mapped digital technologies used to develop clinical communication competencies in undergraduate and postgraduate medical education and examined reported outcomes through a dual lens of pedagogical theories for clinical education, notably experiential learning and translational impact.

Research Objectives

Our overarching aim was achieved by the following objectives:

1. To characterize digital technologies used for developing communication skills in undergraduate and postgraduate medical education, including technology modalities, their

XSL•FO
RenderX

educational context, learner outcomes, and how their use has evolved over time
2. To examine how digital communication skills interventions align with experiential learning processes and the depth of educational impact, as mapped using Kolb's experiential learning cycle and Kirkpatrick's evaluation model
3. To identify evidence gaps in the design, evaluation, and reporting of digital communication skills interventions in relation to behavioral change and clinical practice outcomes

Methods

Study Design

This scoping review was conducted using the Joanna Briggs Institute (JBI) methodological framework [22] and reported adhering to the PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses Extension for Scoping Reviews) [23]. The PRISMA-ScR checklist is included in Checklist 1. This protocol was prospectively registered with Open Science Framework [24].

Research Question

We defined our research question as "What is known in the existing literature about the use of digital technologies for CST in undergraduate and postgraduate medical education, and to what extent do reported outcomes reflect experiential learning processes and extend beyond individual learning toward behavioral change and clinical practice-level impact?"

Eligibility Criteria

Studies were included if they met the criteria outlined in the following sections.

Participants

Participants were learners participating in medical education. For the purposes of this review, we referred to learners as either undergraduates or postgraduates.

Undergraduates were defined as medical students at the preregistration level (ie, individuals enrolled in a medical degree program who have not yet obtained a license or registration to practice).

Postgraduates were defined as medical doctors at the postregistration level, including those in internship, residency, specialty training, or other forms of postgraduate education and continuing professional development.

Concept

The concept involved the delivery of CST through digital and technology-enhanced methods. Communication skills included content (what is said), process (how the communication is done), and perceptual (underlying thoughts, emotions, and clinical reasoning) skills as described by Kurtz et al [25]. For this review, we further defined communication skills as the capacity for at least 2-way, dynamic interaction between the learner and a patient (or patient's representative) to convey health care information between the parties involved. Passive interactions, such as watching a premade video, were excluded.

For the purposes of this review, a patient was defined as either a real human patient, a trained actor (for example, a standardized patient), or a digital or simulated representation designed to support the communication of health care information. Digital technologies were defined as digital systems, tools, or platform devices that use electronic data processing to support the delivery, practice, or assessment of communication skills through active learner engagement, including interactive, simulated, or feedback-enabled educational activities. Studies describing only the assessment of communication skills (eg, in Objective Structured Clinical Examination situations) were excluded.

Context

The context included any setting where communication skills were taught, including academic institutions, clinical environments, and synchronous or asynchronous simulation-based scenarios. The interactions were either real-world or role-play encounters.

Evidence Types

All primary study designs were included to gain comprehensive evidence of the topic. We excluded secondary sources such as systematic reviews, scoping reviews, and meta-analyses, as well as nonresearch publications including study protocols, expert opinion, discussion papers, letters, comments, editorials, and book chapters. Conference abstracts without accompanying full-text publications were excluded due to insufficient detail for data extraction. Primary studies cited within relevant reviews were extracted and assessed for eligibility.

Search Strategy

An initial scoping search in Medline (OVID) identified relevant index terms and keywords, which informed the development of a comprehensive search strategy. The strategy incorporated both keywords and controlled vocabulary across three core concept areas: medical education (undergraduate and postgraduate), digital technologies (VPs, simulation, online learning, AI, virtual reality), and communication skills, combined using Boolean operators. The search strategy was adapted from a previously published scoping review on digital education and communication skills [11] and refined to align with this review's inclusion criteria. During development, the strategy was reviewed in consultation with a research librarian and refined following feedback from the review team, including methodological experts. The final strategy was tested iteratively in Medline and adapted across all selected databases (see Multimedia Appendix 1). Reporting of the search strategy follows the PRISMA-S (Preferred Reporting Items for Systematic Reviews and Meta-Analyses Extension for Literature Searches) checklist [26] (Checklist 2).

The electronic databases Medline (via Ovid), Embase (via Ovid), Educational Resource Information Center (ERIC; via EBSCO), and Cumulative Index to Nursing and Allied Health Literature (CINAHL; via EBSCO) were searched from inception to January 5, 2026. Searches were initially conducted in November 2024, re-run in August 2025, and subsequently updated on January 5, 2026, using the same search strategy to identify newly published studies. No changes were made to the search terms

XSL•FO
RenderX

or eligibility criteria across the three search runs. Reference lists of included articles and relevant reviews were hand-searched for additional studies.

We imposed no language restrictions on any of the searches. Non-English texts were translated using DeepL Translator software [27]. A standardized quality assessment tool was not used in this review, as the primary aim was to map existing research rather than conduct a detailed critical assessment of individual studies.

Selection of Sources of Evidence

Identified studies were uploaded to Covidence [28] and automatically deduplicated. All titles and abstracts were screened independently by two reviewers (HE, PS, or ET), with discrepancies resolved by group discussion with a fourth reviewer (JT). Full texts of potentially eligible studies were screened independently by two reviewers (HE, PS, JC, or BP) with discrepancies resolved through discussion with the whole research group.

Data Charting Process

Data on study characteristics, learner population and training context, digital technology modality and intervention design features, communication domains addressed, outcome measures and effect direction, use of educational theory, and authors' conclusions were extracted into a piloted spreadsheet (Multimedia Appendix 2). Data from 10% of included studies were extracted independently by two reviewers (HE, PS, ET, or BP), with findings compared and discussed with the research group. After a high level of agreement was achieved, remaining studies were extracted by one reviewer (HE or PS).

Synthesis of Results

A descriptive analysis was conducted and reported using frequencies and percentages. Studies were categorized by their primary mode of delivery and interaction for developing communication. To support structured synthesis of heterogeneous outcomes, we applied two complementary frameworks:

1. *Kolb's experiential learning cycle* [20]: Interventions were assessed to identify which stages of experiential learning were explicitly supported (where interventions supported more than one stage of the cycle, all relevant stages were recorded):
   - Stage 1 (concrete experience): direct patient/scenario interaction
   - Stage 2 (reflective observation): guided reflection on experiences
   - Stage 3 (abstract conceptualization): synthesis of principles from experiences
   - Stage 4 (active experimentation): application of learning in new contexts
2. *Kirkpatrick's evaluation model* [21]: Study outcomes were mapped to increasing levels of evaluation:
   - Level 1: learner satisfaction or reactions only
   - Level 2: improvement in knowledge or communication skills acquisition
   - Level 3: behavior change in clinical practice
   - Level 4: patient or organizational outcomes

Kirkpatrick's evaluation model was selected because it is widely recognized in medical education research and enables comparison of outcome depth across diverse intervention types and study designs [29].

The use of these two frameworks allowed consistent appraisal of how interventions were pedagogically structured to support learning processes as well as the depth of outcome assessment. As neither framework was explicitly referenced in included studies, classifications were assigned by authors. Two reviewers (PS and HE) independently conducted all framework assignments, with discrepancies resolved through discussion. This approach enabled systematic comparison across technologies and identification of patterns and evidence gaps in learning processes and outcome depth.

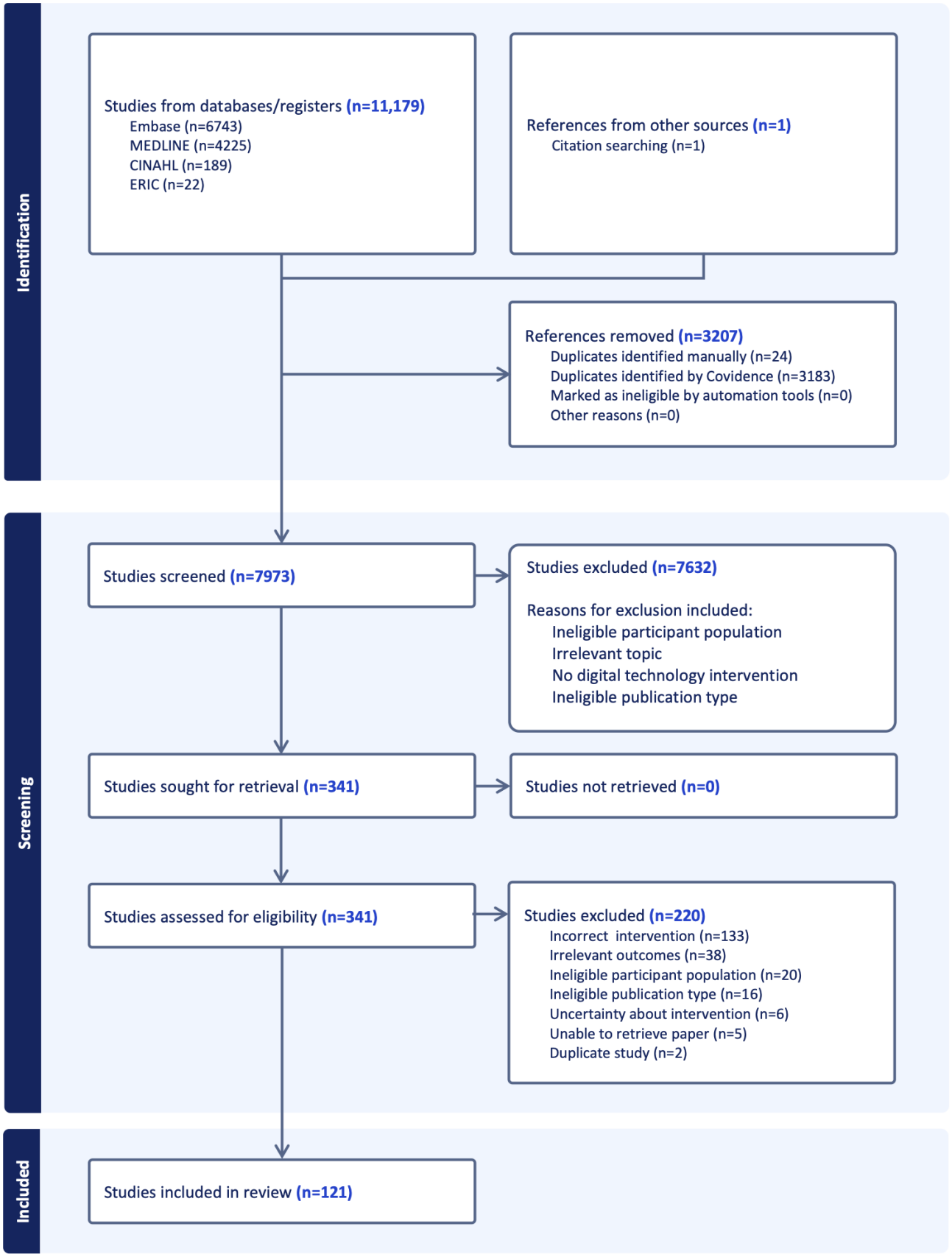
Results

Selection of Sources of Evidence

A total of 11,179 records were identified through database searches, and 1 study was identified via citation searching. Following the removal of 3207 duplicate records, 7973 records were screened. Of these, 341 full-text articles were sought for retrieval, with 336 successfully obtained. After full-text review, 220 studies were excluded based on the predefined eligibility criteria. The reasons for exclusion were incorrect intervention (n=133), irrelevant outcomes (n=38), ineligible participant population (n=19), unsuitable article format (n=16), uncertainty regarding the intervention (n=6), inability to retrieve paper (n=5), and duplicate study (n=1). A total of 121 studies were included in the final synthesis, as detailed in the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) flow diagram (Figure 1). A summary of all included studies is provided in Multimedia Appendix 3 [30-150].

XSL•FO
RenderX

Figure 1. PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) flow diagram showing the identification, screening, eligibility, and inclusion of studies using digital technology for developing communication skills in undergraduate and postgraduate medical education.



Characteristics of Sources of Evidence

Table 1 summarizes key characteristics of the included studies. The earliest studies were published prior to 2000, with relatively low publication activity until the mid-2010s. Most studies were published between 2015 and 2025 (92/121, 76.0%). Publication output peaked in 2025 (24/121), with high publication activity

XSL•FO
RenderX

also noted in 2020, 2021, and 2023. Geographically, most studies originated from North America (48/121, 39.7%) and Europe (45/121, 37.2%), with more limited representation from Asia (16/121, 13.2%), Oceania (7/121, 5.8%), South America (4/121, 3.3%), and Africa (1/121, 0.8%; Figures 2 and 3).

Table . Characteristics of studies included within the scoping review examining digital communication skills training in medical education (n=121).

Domain and feature	Values, n (%)
Study design	
Quasi-experimental	51 (42.1)
Cross-sectional study	29 (24)
Nonrandomized comparative study	10 (8.3)
Mixed methods	8 (6.6)
Qualitative research	2 (1.7)
Descriptive case study	1 (0.8)
Randomized controlled trial	20 (16.5)
Location (by region)	
North America	48 (39.7)
Europe	45 (37.2)
Asia	16 (13.2)
Oceania	7 (5.8)
South America	4 (3.3)
Africa	1 (0.8)
Number of students	
0 - 50	60 (49.6)
51 - 100	28 (231)
>100	31 (25.6)
Not reported	2 (1.7)
Learner level	
Undergraduate	71 (58.7)
Postgraduate	45 (37.2)
Mixed	5 (4.1)
Communication skills focus	
General communication and history taking	62 (51.2)
Difficult conversations (BBN[a], GOC[b], DOME[c])	27 (22.3)
Communication with defined patient groups	15 (12.4)
Shared decision-making and consent	7 (5.8)
Telecommunication/remote consultations	5 (4.1)
Patient education and information giving	5 (4.1)

[a]BBN: breaking bad news.

[b]GOC: goals of care.

[c]DOME: disclosure of medical error.

XSL•FO
RenderX

Figure 2. Distribution of studies examining digital communication skills training in medical education by year of publication, illustrating yearly publication trends in digital communication skills training research.

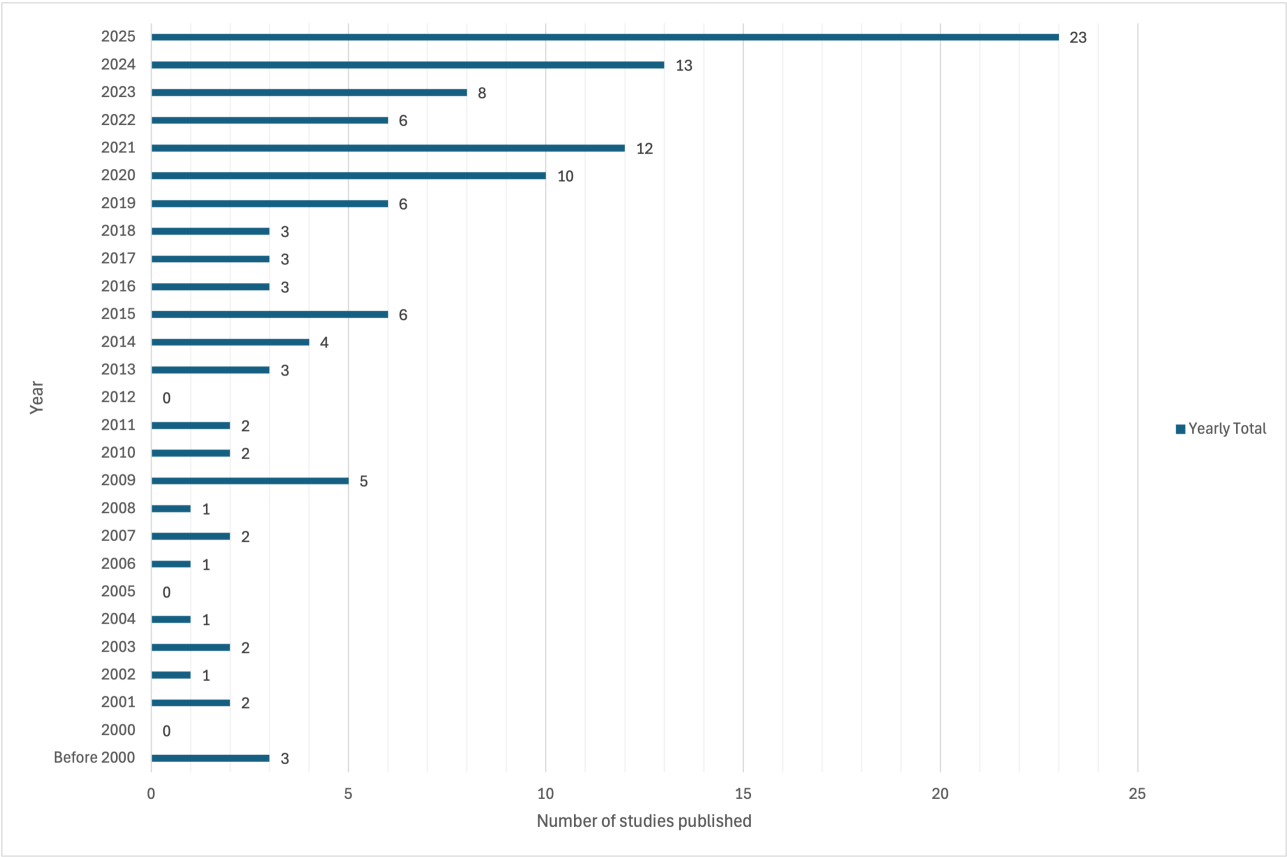
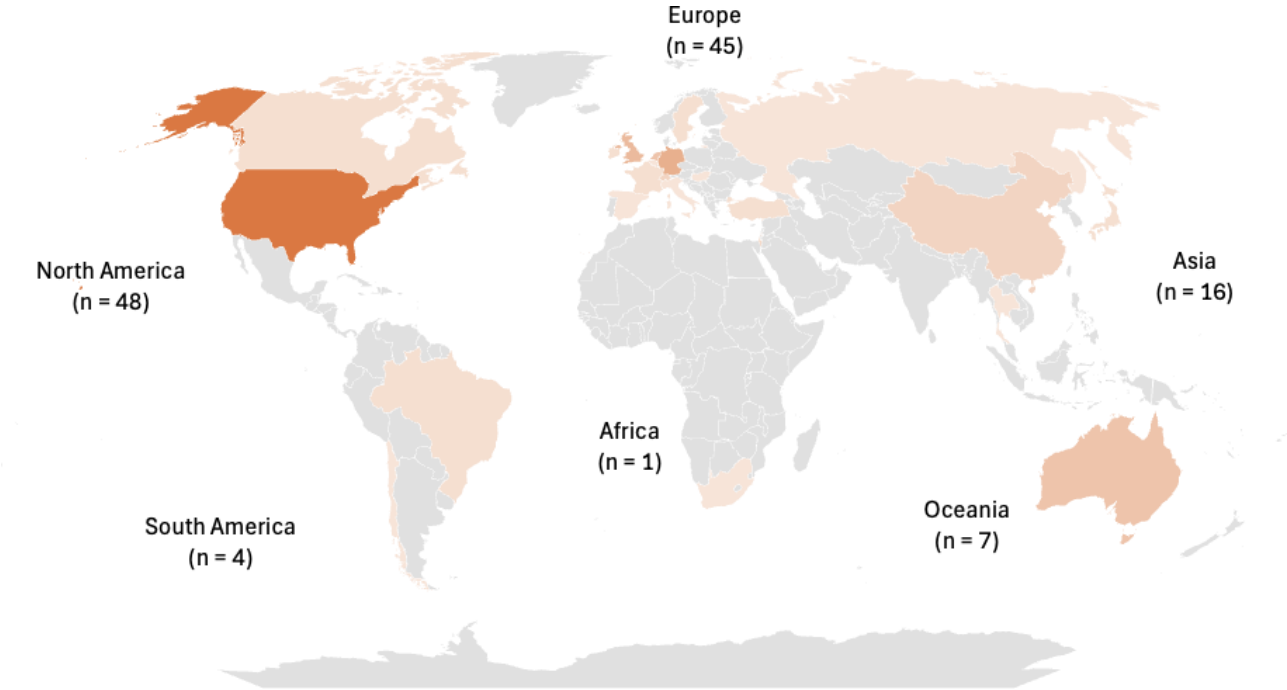


Figure 3. Geographic distribution of studies examining digital communication skills training in medical education by global region.



Quasi-experimental designs were most common (51/121, 42.1%), followed by cross-sectional studies (29/121, 24%) and randomized controlled trials (20/121, 16.5%), with few using cohort, case study, or qualitative methods. Almost one-half of the studies (60/121, 49.6%) involved fewer than 50 students, while the remainder were split between those with 51 - 100 students (28/122, 23.1%) and those with more than 100 (31/121, 25.6%). Most studies focused on undergraduate learners (71/121, 58.7%), with a smaller proportion targeting postgraduates (45/121, 37.2%), and a few included mixed groups (5/121, 4.1%).

XSL•FO
RenderX

In terms of CST, general communication and history taking was the most taught (62/121, 51.2%), followed by difficult conversations, such as breaking bad news, goals of care, or disclosure of medical errors (27/121, 22.3%). A smaller proportion of studies focused on communication with defined patient groups (15/121, 12.4%), including adolescents, children, patients with hearing loss or on dialysis, and individuals receiving psychiatric care. Other areas addressed less frequently were shared decision-making and consent (7/121, 5.8%), telecommunication or remote consultations (5/121, 4.1%), and patient education and information giving (5/121, 4.1%).

Synthesis of Results

Digital Technologies Used

Overview

Using an inductive approach, 3 main approaches to digital communication training were identified: recording-based tools where audio or video was used for teaching, with reflection and feedback [30-80]; live stream platforms that supported synchronous interaction between learners, educators, or simulated patients [81-113]; and VP simulators encompassing both scripted systems and those driven by AI. “Other” digital tools were investigated by 5 further studies [146-150].

Recording-based tools (51/121, 41.8%) mostly used video recording (42/121, 34.7%), with fewer using audio recording (3/121, 2.5%) or the recording functions of videoconferencing platforms (6/121, 5.0%). These typically involved capturing simulated or real consultations for later review, often followed by structured feedback and reflection.

Live stream platforms (33/121, 27.3%) were classified separately and referred exclusively to real-time videoconferencing without recording, typically involving simulated patient encounters, role-play sessions, or workshops with immediate feedback.

VP simulators (32/121, 26.4%) used computer-based or AI-enabled VP to replicate clinical interactions, sometimes in immersive or branching scenarios that allowed repeated practice.

A further 5 studies (5/121, 4.1%) evaluated other digital tools, including AI-assisted case vignettes, patient portals, translation apps, general communication software, and specialty-specific video case platforms.

Recording-Based Digital Tools

The most used approach across all learner groups were recording-based tools [30-80]. These were used fairly equally with postgraduates (26/51, 51%) and undergraduate learners (24/51, 47.1%), demonstrating applicability across training levels, though mixed learner groups were rarely studied (1/51, 2%). Study content was heavily concentrated on general communication and history taking (29/51, 56.9%), followed by difficult conversations such as breaking bad news (11/51, 21.6%). Very few studies addressed communication with defined patient groups (6/51, 11.8%), shared decision-making and consent (4/51, 7.8%), or patient education and information giving (1/51, 2%). The use of these technologies was predominantly beneficial (33/51, 64.7%) [32,34-36,38,41,43-45,48-53,55,56,60-62,64-68,71-73,75-77,79,80], with reported improvements noted in learner performance, reflective capacity, and confidence. However, a notable minority of studies were inconclusive (8/51, 15.7%) [30,33,39,42,54,57,59,63], reported mixed findings (5/51, 9.8%) [31,46,47,74,78], or showed no effect (5/51, 9.8%) [37,40,58,69,70].

Live Stream Platforms

Live stream–based interventions were evaluated in 33 studies [81-113]. These were used more frequently with undergraduate learners (20/33, 60.6%) than postgraduates (12/33, 36.4%), with mixed learner groups rarely studied (1/33, 3%). Study content showed greater diversity: general communication and history taking: 11/33, 33.3%; difficult conversations: 8/33, 24.2%; telecommunication or remote consultations: 5/33, 15.2%; communication with defined population groups: 5/33, 15.2%; shared decision-making and consent: 2/33, 6.1%; and patient education and information giving: 2/33, 6.1%. The use of live-streaming platforms was highly beneficial, with more than two-thirds of studies (23/33, 64.9%) demonstrating positive effects including increased communication confidence [82,85,89,95,101-103,105,108,110], improved self-perceived competence [97,98,100,106,109], high learner satisfaction [86-88,90], and, in some cases, objective gains in communication performance [83,84,91,102,104]. However, a notable proportion reported mixed findings (7/33, 21.2%) [81,92-94,96,111,112], while few studies were inconclusive (1/33, 3.0%) [99], were equivalent to traditional teaching methods (1/33, 3.0%) [113], or demonstrated no measurable effect (1/33, 3.0%) [107].

VP Simulators

VP simulators were evaluated in 32 studies [114-145]. These were used predominantly with undergraduate learners (24/32, 75%), with fewer at the postgraduate level (5/32, 15.6%) or involving mixed learner groups (3/32, 9.4%). Study content returned a concentrated focus similar to recording-based tools, most often addressing general communication and history taking (19/32, 59.4%), followed by difficult conversations (7/32, 21.9%), with minimal attention to communication with defined patient groups (3/32, 9.4%), patient education (2/32, 6.3%), and shared decision-making and consent (1/32, 3.1%). Reported outcomes were generally positive; nearly two-thirds of studies (21/32, 65.6%) found use of VP simulators beneficial [114,119,120,124-127,129,130,132,134-141,143-145], including improvements in communication performance, empathy, and confidence. Notably, 3 of the 32 studies (9.4%) [133,142,151] reported outcomes equivalent to traditional methods. The remaining studies were inconclusive (4/32, 2.5%) [115,116,118,123], reported mixed findings (2/32, 6.3%) [122,131], or showed no effect (2/32, 6.3%) [117,128].

Other Approaches

Tools that did not align with the 3 main categories were evaluated in 5 studies [146-150]. These included AI-assisted case vignettes (1/5, 20%) [146], patient portals for drafting messages (1/5, 20%) [147], speech-to-speech translation apps (1/5, 20%) [148], general communication software (1/5, 20%) [149], and specialty-specific video case platforms (1/5, 20%) [150]. Learners comprised both undergraduates (3/5, 60%) and

XSL•FO
RenderX

postgraduates (2/5, 40%). The majority demonstrated improvements (3/5, 60%) in communication practice and confidence [146,147,150]; 1 study was inconclusive (1/5, 20%) [148], and another reported mixed effects (1/5, 20%) [149].

Trends in the Use of Digital Technologies for CST

Across all learner levels, recording-based approaches were present from the earliest period captured within this review (before 2000) and remained the dominant approach until approximately 2020. Live-streaming platforms expanded markedly after 2020, likely accelerated by the COVID-19 pandemic, with publication numbers surging from sporadic use to 6 studies in 2021 and remaining elevated at 7 studies in 2025. Most strikingly, VP simulators demonstrated explosive recent growth, particularly in 2024 with 14 published studies, representing a dramatic shift in the field's technological focus (Figure S4 in Multimedia Appendix 4).

These trends varied by learner level. In undergraduate education, the surge in VP simulators was particularly pronounced, with 10 studies published in 2024 alone, while live streaming also showed sustained growth from 2021 onward, plateauing at 3 studies annually by 2025 (Figure S5 in Multimedia Appendix 4). In postgraduate education, the pattern differed notably: Recording-based approaches dominated until the mid-2010s before declining to near absence by 2025, while live streaming showed steady growth after 2020, peaking at 4 studies in 2021 - 2022 and 2025 (Figure S6 in Multimedia Appendix 4). VP simulators remained less prominent at the postgraduate level compared with undergraduate settings, showing only gradual uptake in recent years without the dramatic spike observed in undergraduate training.

Training Assessment Tools Reported

Assessment tools for measuring impact of digital technology on CST varied widely across studies (see Table 2 and Figure S7 in Multimedia Appendix 5). Outcome measures were categorized as either objective or self-reported and further classified by validation status (validated, nonvalidated, or unknown). Classification was based on the type and clarity of evidence regarding communication skill development.

Table . Assessment tools used by studies examining digital communication skills training in medical education to evaluate outcomes, stratified by learner level and validation status.

Learner level and assessment type	Validated, n (%)[a]	Nonvalidated, n (%)[a]	Unknown, n (%)[a]
Undergraduate (n=71)			
Objective (n=39)	16 (41)	16 (41)	7 (17.9)
Self-reported (n=32)	0 (0)	32 (100)	0 (0)
Postgraduate (n=45)			
Objective (n=31)	16 (51.6)	7 (22.6)	8 (25.8)
Self-reported (n=14)	0 (0)	13 (92.9)	1 (7.1)
Mixed (n=5)			
Objective (n=2)	2 (100)	0 (0)	0 (0)
Self-reported (n=3)	1 (33.3)	2 (66.7)	0 (0)

[a]Percentages represent the proportion of each validation category within the respective assessment type (objective or self-reported) for each learner level.

Assessment patterns varied by learner level. Among undergraduate learners, objective measures were used slightly more frequently (39/71, 54.9%) than self-reported assessments (32/71, 45.1%). In contrast, postgraduate studies showed a clearer preference for objective assessment (31/45, 68.9%) compared with self-reported tools (14/45, 31.1%). Across all studies, a significant validation gap emerged: Most studies used nonvalidated assessment tools (71/121, 58.7%), while fewer used validated measures (35/121, 28.9%) and a minority having unknown validation status (16/121, 13.2%).

Only 34 of the 121 studies (28.1%) assessed outcomes with a validated objective measure. The objective measures commonly included global rating scales, structured checklists, and simulated patient assessments. A subset of studies used validated frameworks, such as the Calgary-Cambridge Observation Guide, Communication Assessment Tool, SEGUE framework, SPIKES protocol, and the Jefferson Scale of Physician Empathy. Several studies also incorporated qualitative methods such as thematic analysis of interviews or open-ended feedback responses.

Reported Impact of Digital Technology in Relation to Presence of Comparator Group

Study outcomes were grouped into 5 categories: beneficial, mixed, equivalent, inconclusive, or no effect. Studies were classified as beneficial if they included either objective measures (eg, assessor ratings, checklist scores) or self-reported reflections that indicated improvement or effective demonstration of communication skills. Studies were classified as mixed if they reported improvement in certain aspects of communication skills while showing no improvement or even decline in others. Studies were considered equivalent if they compared a digital intervention to another method (eg, traditional teaching or an alternative digital tool) and found no significant difference in communication skill outcomes between groups. Studies were considered inconclusive if they lacked sufficient outcome data to assess impact on communication skills. Finally, studies were labeled no effect if they explicitly reported no improvement in communication performance postintervention.

XSL•FO
RenderX

Most studies demonstrating positive effects of digital technology for CST (80/121) lacked robust methodological designs. Nearly three-quarters (60/81, 74.1%) used either within-group designs (39/80, 48.8%) or had no comparator at all (20/80, 25%), limiting causal inference. Fewer studies (21/80, 26.2%) included an external comparison group, most often comparing against traditional teaching (11/80, 13.8%), followed by no-intervention controls (4/80, 5%), alternative digital tools (3/80, 3.8%), and other nontraditional teaching methods (3/80, 3.8%).

The use of validated assessment tools further weakened the evidence base: Only 26 of the 80 positive studies (32.5%) used validated measures. Among these 26 validated studies, one-half used within-group designs (13/26, 50%), with smaller proportions comparing against traditional teaching (5/26, 19.2%), using no comparator (3/26, 11.5%), no-intervention controls (2/26, 7.7%), other nontraditional teaching (2/26, 7.7%), and alternative digital tools (1/26, 3.8%). Overall, the predominantly positive findings in this field arise largely from study designs with limited methodological rigor and inconsistent use of validated measures, substantially reducing confidence in causal claims about technology effectiveness (Table 3).

Table . Comparator groups used in studies examining digital communication skills training in medical education that report beneficial effects of digital communication skills training.

Type of comparator	Total beneficial studies (n=81), n (%)[a]	Beneficial studies using validated measures (n=26), n (%)[b]
No comparator	20 (25)	3 (11.5)
Within-group	39 (48.8)	13 (50)
Traditional teaching	11 (13.8)	5 (19.2)
Alternative digital tool	3 (3.8)	1 (3.8)
Other nontraditional teaching	3 (3.8)	2 (7.7)
No-intervention control	4 (5)	2 (7.7)

[a]Percentages were calculated using the total number of studies reporting beneficial outcomes as the denominator (n=80).

[b]Percentages were calculated using the total number of beneficial studies using validated outcome measures as the denominator (n=26).

Stage of Kolb's Experiential Learning Cycle

Studies predominantly situated digital technologies within the early phases of Kolb's experiential learning cycle [20] with 120 of 121 studies (99.2%) focusing solely on Stage 1 (Concrete Experience), Stage 2 (Reflective Observation), or both. Nearly one-half of studies (54/121, 44.6%) integrated both stages, typically through combinations of simulated encounters and structured feedback, making this the most common approach. Among studies focusing on a single stage, Stage 1 alone was addressed in 28.9% of studies (35/121), while Stage 2 alone was addressed in 25.6% (31/121).

Progression beyond these initial phases was virtually absent: No studies aligned primarily with Stage 3 (Abstract Conceptualization: 0/121, 0%), and only 1 study extended to Stage 4 (Active Experimentation: 1/121, 0.8%). Figure S8 in Multimedia Appendix 6 shows the distribution of studies across Kolb's learning stages by technology type, demonstrating that all digital modalities concentrated overwhelmingly on Stages 1 and 2.

Outcomes Mapped to Kirkpatrick's Training Evaluation Model

All included studies reported outcomes that could be classified using Kirkpatrick's 4-level model [21]. The vast majority focused on classroom or simulation-level outcomes: A significant number of studies (91/121, 75.2%) were evaluated at Level 2, measuring changes in learners' knowledge, skills, or observed performance in simulated environments, while nearly all others (23/121, 19%) assessed Level 1 outcomes such as learner satisfaction, confidence, or perceived usefulness. Combined, 94.2% of studies (114/121) measured outcomes within educational settings rather than clinical practice.

In contrast, evaluation of real-world impact was minimal. Only 6 of the 121 studies (5%) assessed behavior change in clinical practice at Kirkpatrick's Level 3, and 1 study (1/121, 0.8%) examined patient-level outcomes at Level 4. This concentration in Levels 1 and 2 indicates that, although there is substantial evidence for the educational effectiveness of digital technologies in controlled settings, the field has generated almost no evidence regarding their impact on clinical practice or patient outcomes. Figure S9 in Multimedia Appendix 7 presents the distribution of studies across technology types and Kirkpatrick levels, illustrating the concentration at Levels 1 and 2 with minimal progression to behavior change (Level 3) or patient outcomes (Level 4).

AI in CST

A small but emerging subset of studies (15/121, 12.4%) [50,116,127,130,131,135-138,140,142-144,150,152] explicitly incorporated AI into CST, spanning undergraduate (11/15, 73.3%), postgraduate (3/15, 20%), and mixed learner populations (1/15, 6.7%). AI applications were concentrated predominantly within VP simulators (12/15, 80%) [116,127,130,131,135-138,140,142-144], with smaller numbers categorized under other digital interventions (2/15, 13.3%) [146,150] and video-recording approaches (1/15, 6.7%) [50]. These AI applications primarily simulated patient interactions and delivered automated feedback, supporting the development of skills such as history taking, information gathering, interview techniques, nonverbal behavior, and empathy.

XSL•FO
RenderX

Reported outcomes were largely positive, with 12 of the 15 studies (80%) demonstrating beneficial effects on communication skills, a higher success rate than digital technologies overall. However, the remaining studies showed more limited evidence: 1 reported mixed findings [131], 1 was inconclusive [116], and 1 other found outcomes equivalent to traditional actor-based training [142]. Notably, the study with mixed findings [131], which relied on self-reported outcomes, highlighted limitations in AI realism and did not provide clear evidence of impact, suggesting that the effectiveness of AI may depend on both implementation quality and measurement approach. The distribution of evidence across technology types, communication domains, learner levels, Kolb's learning stages, and Kirkpatrick's evaluation levels is presented in the evidence and gap map (Multimedia Appendix 7), which visually highlights areas of concentrated research and gaps in the current evidence base.

Discussion

Summary of Principal Findings

This review mapped the digital technologies used for CST and examined reported outcomes through the lens of experiential learning theory and translational evaluation frameworks. The evidence base has expanded substantially over the past decade, with notable acceleration following the COVID-19 pandemic, though research remains concentrated in North America and Europe and predominantly involves undergraduate learners. Three principal technology categories dominated the field: recording-based approaches, live-streaming platforms, and VP simulators. Application of Kolb's experiential learning cycle [20] and Kirkpatrick's evaluation model [21] revealed that current digital interventions overwhelmingly support early-stage learning processes and are evaluated almost exclusively within educational settings, with minimal evidence of progression to deeper competency development or translation into clinical practice and patient outcomes.

Comparison of Findings and Interpretation Within the Context of Literature

The field of digital CST has undergone substantial transformation in recent years. Live-streaming platforms expanded markedly after 2020, likely accelerated by the shift to virtual learning environments during the COVID-19 pandemic, while VP simulators demonstrated the most striking recent growth. These trends mirror broader patterns observed in the adoption of digital tools in medical education [15] and in the expansion of virtual simulation for CST [17]. Despite this technological diversification, recording-based approaches that enable structured feedback and reflection continue to be widely used, demonstrating their continued relevance in CST.

Study content across technologies varied in ways that have implications for curriculum design. Recording-based tools and VP simulators were heavily focused on general communication and history taking, with more limited attention to areas such as shared decision-making, patient education, or communication with specific patient populations. Live stream platforms, by contrast, demonstrated greater diversity across communication domains, including telecommunication and remote consultation skills. This uneven exploration suggests that some modalities may be more readily adapted to diverse communication scenarios than others, though the reasons for these differences remain unclear and warrant further investigation.

The evidence base, although growing, reveals substantial methodological limitations that constrain confidence in reported outcomes. The vast majority of studies demonstrating beneficial effects relied on designs with limited capacity for causal inference, predominantly using within-group comparisons or lacking comparators entirely. Validation of outcome measures was similarly inconsistent, with most studies using nonvalidated tools. These patterns indicate that, although digital technologies show promise, the strength of evidence supporting their effectiveness remains weak [153]. Notwithstanding, live stream platforms showed both the highest rate of beneficial outcomes and greater diversity in the communication domains addressed, suggesting this modality may offer particular flexibility in supporting varied learning objectives [154,155]. Although most VP simulator studies demonstrated beneficial outcomes, a small number of studies reported outcomes equivalent to traditional methods, raising questions about when and for whom VPs offer advantages over established approaches. These patterns suggest that effectiveness may depend not only on the technology itself but also on how it is implemented, for what learning objectives, and with which learner populations [156].

Application of educational theory frameworks revealed two striking gaps in current research. First, digital technologies are almost exclusively used to support the initial phases of experiential learning (ie, concrete experience and reflective observation), with virtually no progression to abstract conceptualization or active experimentation. This concentration suggests that current technologies facilitate earlier learning processes for CST, stopping short of the deeper conceptual understanding and experimental application that characterize advanced skill development [20]. Second, outcome evaluation overwhelmingly focused on classroom and simulation-level measures, with minimal assessment of whether learning gains translate to actual clinical practice or patient outcomes. This near absence of real-world evaluation represents a critical knowledge gap of whether the educational benefits observed in controlled settings yield meaningful improvements in patient care [1,2].

The findings of this review are consistent with previous observations that research on CST in medical education is dominated by descriptive and quasi-experimental designs, with relatively few robust comparative studies [157]. Our analysis adds new depth by demonstrating that most studies reporting beneficial outcomes relied on within-group designs or lacked comparators entirely and validated outcome measures were used in few positive studies. This methodological profile reveals an evidence base that remains largely exploratory rather than definitive, a pattern that was particularly evident during the rapid post-COVID-19 adoption of digital tools, when educational continuity was prioritized over rigorous evaluation [15]. The heterogeneity in study designs and outcome measures continues to hamper synthesis of effectiveness claims across technologies [11].

XSL•FO
RenderX

AI-enabled CST represented a small proportion of the included studies, with most applications embedded within VP simulators and focused on automating patient interactions or feedback. Although many of these studies reported beneficial effects, few used validated communication outcome measures, and evidence was largely derived from noncomparative designs and proximal learning outcomes. Recent literature [158] has similarly noted that, although interest in AI-supported communication skills education is increasing, research to date has predominantly focused on feasibility and instructional potential, with limited attention to the psychometric validation of AI-specific assessment approaches. Together, these findings suggest that AI-supported CST remains at an early stage of evidence development, requiring more rigorous and theory-aligned evaluation to support meaningful comparison and wider implementation.

A recent meta-analysis of 3 randomized controlled trials [11] assessed digital technology to be as effective as traditional teaching for developing communication skills, though low-quality evidence limited confidence in the finding. Our review supported this conclusion but extends the evidence base by including a wider range of study designs and technologies and confirmed that many reports of benefit arise from weaker methodological approaches. The finding that a minority of studies included external comparators means that most claims of benefit rest on uncontrolled observations, providing weak grounds for judging whether digital approaches outperform, match, or fall short of established methods. Digital technologies therefore have potential in this field, but stronger study designs and robust outcomes, also advocated by Gilligan et al [6], are required to achieve meaningful conclusions and would enhance confidence for attributing gains directly to the interventions. Additionally, our study has illuminated the positioning of technologies for CST within educational pedagogy and translation into clinical outcomes. Evidence of enhancement was most apparent in formative learning contexts. Learners frequently reported increased confidence, satisfaction, and perceived competence, and studies using video recording or conferencing highlighted the value of repeated practice, self-reflection, and timely feedback. These features may offer advantages in certain learning domains, especially where opportunities for repeated practice, asynchronous reflection, or scalable feedback are limited in traditional classroom-based approaches, which typically emphasize role-play and face-to-face interactions [6,159].

Cook and Ellaway [19] previously proposed an evaluation framework to aid synthesis of technology-enhanced learning that focused on the implementation process. Alternatively, placing intervention assessment within education theory, as in this review, can aid comparison while additionally providing insight into the learning and teaching process, thus informing future curricula development. However, the concentration of outcomes at Kirkpatrick Levels 1 and 2 (satisfaction and skill demonstration in educational settings) [21] with minimal assessment of behavior change in practice or patient outcomes raises a fundamental question about translational impact. The persistent emphasis on proximal learning outcomes may reflect pragmatic constraints, as behavior change and patient-level outcomes are more difficult and costly to measure, but it leaves largely unanswered whether communication skills developed through digital training transfer to clinical practice and ultimately improve patient care. This gap between educational and clinical outcomes has been noted in wider educational research, which often prioritizes measures that are easier to capture [160]. However, the near-total absence of practice-level and patient-level evaluation in this field is particularly concerning given the fundamental premise that CST should ultimately enhance patient care. The application of Kirkpatrick's framework in this review makes this gap explicit: Demonstrating that learners can perform skills in simulated settings does not establish that they will apply those skills effectively in real clinical encounters or that such application improves patient outcomes.

Strengths and Limitations

To our knowledge, this is the first broad and comprehensive review of digital technologies for CST in both undergraduate and postgraduate medical education. Our systematic approach, guided by predefined eligibility criteria and mapped to established educational frameworks, enabled synthesis across diverse study designs and outcomes. This approach not only provided a structured view of where and how digital technologies are being applied within CST but also revealed patterns in how learning is being conceptualized and measured across different educational contexts.

Several limitations should be noted. This was a scoping review; therefore, a formal risk of bias or quality assessment was not undertaken. However, the extracted data clearly indicate that many studies were of low methodological quality, often relying on small samples, weak and noncomparative designs, and unvalidated outcome measures. The predominance of such designs limits the strength of causal claims that can be made about technology effectiveness. Across the included studies, no digital intervention was associated with a clear deterioration in communication skills. Where benefits were not observed, findings were most often neutral or mixed rather than negative. Although this may suggest that digital approaches are unlikely to cause harm, it may represent publication bias, outcome measurement limitations, and the methodological weaknesses evident across most included studies, which lack the rigor needed to detect potential adverse effects. In addition, rapid advances in technology mean that some emerging tools may not yet be captured in the published literature. Gray literature was not searched, and, although such sources often lack the detailed outcome reporting necessary for mapping to theoretical frameworks, this exclusion may have resulted in some relevant evidence being missed. However, the impact of this is likely mitigated by the large number of studies included in our review. Most included studies were conducted in North America or Europe, potentially limiting the generalizability of findings to developing countries, as infrastructure, personal cost, and institutional constraints shape how digital technologies are used in medical training globally [161].

Although Kolb's experiential learning cycle provided a useful framework for examining which learning processes were targeted by digital CST interventions, we acknowledge that

XSL•FO
RenderX

contemporary cognitive and neuroscience research suggests learning may occur nonlinearly, with stages overlapping or occurring simultaneously rather than sequentially [162-164]. In this review, all stages of Kolb's cycle supported by each intervention were recorded, and we recorded multiple stages of learning when appropriate. More than one-half of the included studies supported both concrete experience and reflective observation, reflecting common pedagogical designs that combine simulated encounters with structured reflection. Despite allowing for overlap, very few studies extended to later stages of experiential learning, with almost no interventions supporting abstract conceptualization or active experimentation. This pattern, whether viewed through a sequential or nonlinear lens, indicates that digital technologies as currently designed and evaluated support initial skill acquisition and reflection but do not facilitate progression toward the conceptual understanding and experimental application needed for mastery.

Conclusions and Recommendations

Digital CST interventions showed promise for supporting early-stage learner outcomes, such as improved confidence, satisfaction, and knowledge or skill acquisition. However, there was little evidence that these technologies supported progression to longer-term competency development in either undergraduate or postgraduate learners, and the translation of digitally acquired skills into improved patient communication or care remains largely unevaluated. Although these technologies offer scalable, flexible training opportunities, current evidence does not establish their effectiveness relative to traditional methods or their impact on clinical practice. Digital CST should be integrated thoughtfully within broader curricula and accompanied by rigorous evaluation extending beyond educational settings to assess impact on patient outcomes. Several priorities emerge for future research and practice. First, greater methodological rigor in evaluation is essential, including comparative designs that establish whether digital approaches match, exceed, or fall short of traditional training, and consistent use of validated outcome measures to ensure robust assessment. Second, research must extend beyond educational outcomes to assess whether learning gains translate to clinical practice and patient care to determine whether learning gains translate beyond educational settings. Third, investigation is needed into how digital technologies might support progression through complete learning cycles, moving beyond initial skill acquisition and reflection to support conceptual understanding and experimental application in varied contexts [162-164]. Finally, as AI-enabled approaches continue to emerge in CST, there is opportunity to design them not only as training tools but also as technologies that scaffold competency development and translation into improved patient outcomes.

Acknowledgments

The authors declare that no generative artificial intelligence (AI) tools were used in the research or writing process. According to the GAIDeT taxonomy (2025), no tasks were delegated to generative AI tools. Responsibility for the final manuscript lies entirely with the authors. Generative AI tools are not listed as authors and do not bear responsibility for the final outcomes. Declaration submitted by PS (on behalf of all authors).

Funding

This research did not receive specific external funding. The work was supported through the authors' affiliation with the University of Nottingham.

Data Availability

The datasets generated or analyzed during this study are available from the corresponding author on reasonable request.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Search strategy.
[DOCX File, 31 KB - mededu_v12i1e87012_app1.docx]

Multimedia Appendix 2
Blank data extraction form.
[XLSX File, 14 KB - mededu_v12i1e87012_app2.xlsx]

Multimedia Appendix 3
Summary table of data extracted from included studies by category of technology.
[DOCX File, 69 KB - mededu_v12i1e87012_app3.docx]

Multimedia Appendix 4

XSL•FO
RenderX

Distribution of studies examining digital communication skills training in medical education across learner groups by technology type and year of publication.
[PDF File, 592 KB - mededu_v12i1e87012_app4.pdf]

Multimedia Appendix 5
Validation status of assessment tools used in studies examining digital communication skills training in medical education, stratified by learner level.
[PNG File, 234 KB - mededu_v12i1e87012_app5.png]

Multimedia Appendix 6
Evidence gap map showing distribution of included studies across Kolb's experiential learning cycle and Kirkpatrick's evaluation model by technology type.
[PDF File, 221 KB - mededu_v12i1e87012_app6.pdf]

Multimedia Appendix 7
Evidence gap map of studies examining digital communication skills training in medical education.
[PDF File, 425 KB - mededu_v12i1e87012_app7.pdf]

Checklist 1
PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses Extension for Scoping Reviews) checklist.
[DOCX File, 112 KB - mededu_v12i1e87012_app8.docx]

Checklist 2
PRISMA-S (Preferred Reporting Items for Systematic Reviews and Meta-Analyses Extension for Literature Searches) checklist.
[DOCX File, 25 KB - mededu_v12i1e87012_app9.docx]

References

1. Coulter A, Collins A. Making shared decision-making a reality: no decision about me, without me. The King's Fund. 2011. URL: www.kingsfund.org.uk [accessed 2026-04-03]
2. Campos CFC, Olivo CR, Martins MDA, Tempski PZ. Physicians' attention to patients' communication cues can improve patient satisfaction with care and perception of physicians' empathy. Clinics (Sao Paulo) 2024;79:100377. [doi: 10.1016/j.clinsp.2024.100377] [Medline: 38703716]
3. Good medical practice. General Medical Council. 2018. URL: https://www.gmc-uk.org/professional-standards/the-professional-standards/good-medical-practice [accessed 2026-04-03]
4. Good medical practice: a code of conduct for doctors in Australia. Australian Medical Council. URL: https://www.amc.org.au/images/Final_Code.pdf [accessed 2026-04-03]
5. Aspegren K, Lønberg-Madsen P. Which basic communication skills in medicine are learnt spontaneously and which need to be taught and trained? Med Teach 2005 Sep;27(6):539-543. [doi: 10.1080/01421590500136501] [Medline: 16199362]
6. Gilligan C, Powell M, Lynagh MC, et al. Interventions for improving medical students' interpersonal communication in medical consultations. Cochrane Database Syst Rev 2021 Feb 8;2(2):CD012418. [doi: 10.1002/14651858.CD012418.pub2] [Medline: 33559127]
7. Rigby PG, Gururaja RP. World medical schools: the sum also rises. JRSM Open 2017 Jun;8(6):2054270417698631. [doi: 10.1177/2054270417698631] [Medline: 28620505]
8. Global strategy on human resources for health: workforce 2030. World Health Organization. 2020. URL: https://www.who.int/publications/i/item/9789241511131 [accessed 2026-04-03]
9. Pressures in general practice. British Medical Association. 2026. URL: https://www.bma.org.uk/advice-and-support/nhs-delivery-and-workforce/pressures/pressures-in-general-practice-data-analysis [accessed 2026-04-03]
10. Rotenstein LS, Brown R, Sinsky C, Linzer M. The association of work overload with burnout and Intent to leave the job across the healthcare workforce during COVID-19. J Gen Intern Med 2023 Jun;38(8):1920-1927. [doi: 10.1007/s11606-023-08153-z] [Medline: 36959522]
11. Kyaw BM, Posadzki P, Paddock S, Car J, Campbell J, Tudor Car L. Effectiveness of digital education on communication skills among medical students: systematic review and meta-analysis by the digital health education collaboration. J Med Internet Res 2019 Aug 27;21(8):e12967. [doi: 10.2196/12967] [Medline: 31456579]
12. McGee RG, Wark S, Mwangi F, et al. Digital learning of clinical skills and its impact on medical students' academic performance: a systematic review. BMC Med Educ 2024 Dec 18;24(1):1477. [doi: 10.1186/s12909-024-06471-2] [Medline: 39696150]

XSL•FO
RenderX

13. Kyaw BM, Posadzki P, Dunleavy G, et al. Offline digital education for medical students: systematic review and meta-analysis by the digital health education collaboration. J Med Internet Res 2019 Mar 25;21(3):e13165. [doi: 10.2196/13165] [Medline: 30907731]
14. Ahmady S, Kallestrup P, Sadoughi MM, et al. Distance learning strategies in medical education during COVID-19: a systematic review. J Educ Health Promot 2021;10(1):421. [doi: 10.4103/jehp.jehp_318_21] [Medline: 35071627]
15. Gordon M, Daniel M, Ajiboye A, et al. A scoping review of artificial intelligence in medical education: BEME Guide No. 84. Med Teach 2024 Apr 2;46(4):446-470. [doi: 10.1080/0142159X.2024.2314198]
16. Liaw SY, Ooi SW, Rusli KDB, Lau TC, Tam WWS, Chua WL. Nurse-physician communication team training in virtual reality versus live simulations: randomized controlled trial on team communication and teamwork attitudes. J Med Internet Res 2020 Apr 8;22(4):e17279. [doi: 10.2196/17279] [Medline: 32267235]
17. Fernández-Alcántara M, Escribano S, Juliá-Sanchis R, et al. Virtual simulation tools for communication skills training in health care professionals: literature review. JMIR Med Educ 2025 May 6;11:e63082. [doi: 10.2196/63082] [Medline: 40327882]
18. Stamer T, Steinhäuser J, Flägel K. Artificial intelligence supporting the training of communication skills in the education of health care professions: scoping review. J Med Internet Res 2023 Jun 19;25:e43311. [doi: 10.2196/43311] [Medline: 37335593]
19. Cook DA, Ellaway RH. Evaluating technology-enhanced learning: a comprehensive framework. Med Teach 2015;37(10):961-970. [doi: 10.3109/0142159X.2015.1009024] [Medline: 25782599]
20. Kolb D. Experiential Learning: Experience as the Source of Learning and Development, 2nd edition: Pearson FT Press; 2015.
21. Kirkpatrick JD, Kirkpatrick WK. Kirkpatrick's Four Levels of Training Evaluation: Association for Talent Development; 2016.
22. Pollock D, Peters MDJ, Khalil H, et al. Recommendations for the extraction, analysis, and presentation of results in scoping reviews. JBI Evid Synth 2023 Mar 1;21(3):520-532. [doi: 10.11124/JBIES-22-00123] [Medline: 36081365]
23. Tricco AC, Lillie E, Zarin W, et al. PRISMA Extension for Scoping Reviews (PRISMA-ScR): checklist and explanation. Ann Intern Med 2018 Oct 2;169(7):467-473. [doi: 10.7326/M18-0850] [Medline: 30178033]
24. Seripenah P, Emery H, Tyrrell E, Carson J, Leonardi-Bee J, Evans C, et al. The Use of Digital Technology For Developing Communication Skills in Undergraduate and Post-Graduate Medical Education: A Scoping Review Protocol.: Open Science Framework; 2024. URL: https://osf.io/gmfe3/overview [accessed 2026-04-03]
25. Kurtz S, Silverman J, Draper J, Dalen J, Platt FW. Teaching and Learning Communication Skills in Medicine, 2nd edition: CRC Press; 2017. [doi: 10.1201/9781315378398]
26. Rethlefsen ML, Kirtley S, Waffenschmidt S, et al. PRISMA-S: an extension to the PRISMA Statement for Reporting Literature Searches in Systematic Reviews. Syst Rev 2021 Jan 26;10(1):39. [doi: 10.1186/s13643-020-01542-z] [Medline: 33499930]
27. DeepL Translate. DeepL. URL: https://www.deepl.com/en/translator [accessed 2026-04-03]
28. Covidence. URL: https://www.covidence.org [accessed 2026-04-03]
29. Bewley WL, O'Neil HF. Evaluation of medical simulations. Mil Med 2013 Oct;178(10 Suppl):64-75. [doi: 10.7205/MILMED-D-13-00255] [Medline: 24084307]
30. Beyth Y, Hardoff D, Rom E, Ziv A. A simulated patient-based program for training gynecologists in communication with adolescent girls presenting with gynecological problems. J Pediatr Adolesc Gynecol 2009 Apr;22(2):79-84. [doi: 10.1016/j.jpag.2007.11.002] [Medline: 19345912]
31. Bonnaud-Antignac A, Campion L, Pottier P, Supiot S. Videotaped simulated interviews to improve medical students' skills in disclosing a diagnosis of cancer. Psychooncology 2010 Sep;19(9):975-981. [doi: 10.1002/pon.1649] [Medline: 19918865]
32. Bos-van den Hoek DW, van Laarhoven HWM, Ali R, et al. Blended online learning for oncologists to improve skills in shared decision making about palliative chemotherapy: a pre-posttest evaluation. Support Care Cancer 2023 Feb 23;31(3):184. [doi: 10.1007/s00520-023-07625-6] [Medline: 36820944]
33. Bußenius L, Kadmon M, Berberat PO, Harendza S. Evaluating the Global Rating scale's psychometric properties to assess communication skills of undergraduate medical students in video-recorded simulated patient encounters. Patient Educ Couns 2022 Mar;105(3):750-755. [doi: 10.1016/j.pec.2021.06.001] [Medline: 34112546]
34. Cals JWL, Scheppers NAM, Hopstaken RM, et al. Evidence based management of acute bronchitis; sustained competence of enhanced communication skills acquisition in general practice. Patient Educ Couns 2007 Nov;68(3):270-278. [doi: 10.1016/j.pec.2007.06.014] [Medline: 17714907]
35. Carrard V, Bourquin C, Stiefel F, Schmid Mast M, Berney A. Undergraduate training in breaking bad news: a continuation study exploring the patient perspective. Psychooncology 2020 Feb;29(2):398-405. [doi: 10.1002/pon.5276] [Medline: 31702843]
36. Denizon Arranz S, Blanco Canseco JM, Pouplana Malagarriga MM, et al. Multi-source evaluation of an educational program aimed at medical students for interviewing/taking the clinical history using standardized patients. GMS J Med Educ 2021;38(2):Doc40. [doi: 10.3205/zma001436] [Medline: 33763525]

XSL•FO
RenderX

37. Dohms MC, Collares CF, Tibério IC. Video-based feedback using real consultations for a formative assessment in communication skills. BMC Med Educ 2020 Feb 24;20(1):57. [doi: 10.1186/s12909-020-1955-6] [Medline: 32093719]
38. Dowling S, Rouse M, Farrell J, Gannon L, Sullivan G, Cussen K. The acceptability, feasibility and educational impact of a new tool for formative assessment of the consultation performance of specialty registrars in an Irish general practice training scheme. Educ Prim Care 2007 Jan;18(6):724-735. [doi: 10.1080/14739879.2007.11493611]
39. Evens S, Curtis P. Using patient-simulators to teach telephone communication skills to health professionals. J Med Educ 1983 Nov;58(11):894-898. [doi: 10.1097/00001888-198311000-00008] [Medline: 6631928]
40. Farnill D, Todisco J, Hayes SC, Bartlett D. Videotaped interviewing of non-English speakers: training for medical students with volunteer clients. Med Educ 1997 Mar;31(2):87-93. [doi: 10.1111/j.1365-2923.1997.tb02464.x] [Medline: 9231107]
41. Fischbeck S, Hardt J, Malkewitz C, Petrowski K. Evaluation of a digitized physician-patient-communication course evaluated by preclinical medical students: a replacement for classroom education? GMS J Med Educ 2020;37(7):Doc85. [doi: 10.3205/zma001378] [Medline: 33364364]
42. Freytag J, Chu J, Hysong SJ, et al. Acceptability and feasibility of video-based coaching to enhance clinicians' communication skills with patients. BMC Med Educ 2022 Feb 8;22(1):85. [doi: 10.1186/s12909-021-02976-2] [Medline: 35135521]
43. Hardoff D, Gefen A, Sagi D, Ziv A. Training physicians toward a dignifying approach in adolescents' health care: a promising simulation-based medical education program. Isr Med Assoc J 2016 Aug;18(8):484-488. [Medline: 28471581]
44. Harnof S, Hadani M, Ziv A, Berkenstadt H. Simulation-based interpersonal communication skills training for neurosurgical residents. Isr Med Assoc J 2013 Sep;15(9):489-492. [Medline: 24340839]
45. Harendza S, Bußenius L, Gärtner J, Heuser M, Ahles J, Prediger S. "Fit for the finals" - project report on a telemedical training with simulated patients, peers, and assessors for the licensing exam. GMS J Med Educ 2023;40(2):Doc17. [doi: 10.3205/zma001599] [Medline: 37361248]
46. Hulsman RL, Harmsen AB, Fabriek M. Reflective teaching of medical communication skills with DiViDU: assessing the level of student reflection on recorded consultations with simulated patients. Patient Educ Couns 2009 Feb;74(2):142-149. [doi: 10.1016/j.pec.2008.10.009] [Medline: 19062232]
47. Hulsman RL, van der Vloodt J. Self-evaluation and peer-feedback of medical students' communication skills using a web-based video annotation system. Exploring content and specificity. Patient Educ Couns 2015 Mar;98(3):356-363. [doi: 10.1016/j.pec.2014.11.007] [Medline: 25433967]
48. Kaltman S, Talisman N, Pennestri S, Syverson E, Arthur P, Vovides Y. Using technology to enhance teaching of patient-centered interviewing for early medical students. Simul Healthc 2018 Jun;13(3):188-194. [doi: 10.1097/SIH.0000000000000304] [Medline: 29771814]
49. Knowles C, Kinchington F, Erwin J, Peters B. A randomised controlled trial of the effectiveness of combining video role play with traditional methods of delivering undergraduate medical education. Sex Transm Infect 2001 Oct;77(5):376-380. [doi: 10.1136/sti.77.5.376] [Medline: 11588286]
50. Kobayashi M, Katayama M, Hayashi T, et al. Effect of multimodal comprehensive communication skills training with video analysis by artificial intelligence for physicians on acute geriatric care: a mixed-methods study. BMJ Open 2023 Mar 3;13(3):e065477. [doi: 10.1136/bmjopen-2022-065477] [Medline: 36868602]
51. Liu C, Lim RL, McCabe KL, Taylor S, Calvo RA. A web-based telehealth training platform incorporating automated nonverbal behavior feedback for teaching communication skills to medical students: a randomized crossover study. J Med Internet Res 2016 Sep 12;18(9):e246. [doi: 10.2196/jmir.6299] [Medline: 27619564]
52. Makaricheva EV, Mutigullina AA. Application of situated and simulation-based learning technologies in the formation of communication skills in medical students. Cardiovasc Ther Prev 2024;23(3S):4216 [FREE Full text] [doi: 10.15829/1728-8800-2024-4216]
53. Mauksch L, Farber S, Greer HT. Design, dissemination, and evaluation of an advanced communication elective at seven U.S. medical schools. Acad Med 2013 Jun;88(6):843-851. [doi: 10.1097/ACM.0b013e31828fd5ed] [Medline: 23633673]
54. Mohos A, Mester L, Barabás K, Nagyvári P, Kelemen O. Doctor-patient communication training with simulated patient during the coronavirus pandemic. Orv Hetil 2020 Aug;161(33):1355-1362. [doi: 10.1556/650.2020.31930] [Medline: 32749235]
55. Moulton CA, Tabak D, Kneebone R, Nestel D, MacRae H, LeBlanc VR. Teaching communication skills using the integrated procedural performance instrument (IPPI): a randomized controlled trial. Am J Surg 2009 Jan;197(1):113-118. [doi: 10.1016/j.amjsurg.2008.09.006] [Medline: 19101253]
56. Müller E, Diesing A, Rosahl A, Scholl I, Härter M, Buchholz A. Evaluation of a shared decision-making communication skills training for physicians treating patients with asthma: a mixed methods study using simulated patients. BMC Health Serv Res 2019 Aug 30;19(1):612. [doi: 10.1186/s12913-019-4445-y] [Medline: 31470856]
57. Naik N, Greenwald P, Hsu H, Harvey K, Clark S, Sharma R. "Web-side manner": a simulation-based, telemedicine communication curriculum. Ann Emerg Med 2018;72(4 (Supplement):S105. [doi: 10.1111/acem.13433] [Medline: 622358806]
58. Noordman J, Post B, van Dartel AAM, Slits JMA, Olde Hartman TC. Training residents in patient-centred communication and empathy: evaluation from patients, observers and residents. BMC Med Educ 2019 May 2;19(1):128. [doi: 10.1186/s12909-019-1555-5] [Medline: 31046756]

XSL•FO
RenderX

59. Noordman J, Verhaak P, van Dulmen S. Web-enabled video-feedback: a method to reflect on the communication skills of experienced physicians. Patient Educ Couns 2011 Mar;82(3):335-340. [doi: 10.1016/j.pec.2010.11.003] [Medline: 21146342]
60. Öhlmann H, Icenhour A, Elsenbruch S, Benson S. "Powerful placebo": a teaching and learning concept addressing placebo and nocebo effects in competency-based communication training. GMS J Med Educ 2024;41(4):Doc38. [doi: 10.3205/zma001693] [Medline: 39415814]
61. Ozcakar N, Mevsim V, Guldal D, et al. Is the use of videotape recording superior to verbal feedback alone in the teaching of clinical skills? BMC Public Health 2009 Dec 19;9:474. [doi: 10.1186/1471-2458-9-474] [Medline: 20021688]
62. Perron NJ, Louis-Simonet M, Cerutti B, Pfarrwaller E, Sommer J, Nendaz M. Feedback based on videotaped consultations or immediately after direct observation: which is more effective. Praxis 2015;104 [FREE Full text] [doi: 10.1024/1661-8157/a001994] [Medline: 71976969]
63. Pless A, Hari R, Brem B, Woermamm U, Schnabel KP. Using self and peer video annotations of simulated patient encounters in communication training to facilitate the reflection of communication skills: an implementation study. GMS J Med Educ 2021;38(3):Doc55. [doi: 10.3205/zma001451] [Medline: 33824891]
64. Ravitz P, Lancee WJ, Lawson A, et al. Improving physician-patient communication through coaching of simulated encounters. Acad Psychiatry 2013 Mar 1;37(2):87-93. [doi: 10.1176/appi.ap.11070138] [Medline: 23475235]
65. Roter DL, Larson S, Shinitzky H, et al. Use of an innovative video feedback technique to enhance communication skills training. Med Educ 2004 Feb;38(2):145-157. [doi: 10.1111/j.1365-2923.2004.01754.x] [Medline: 14871385]
66. Roter DL, Hall JA, Kern DE, Barker LR, Cole KA, Roca RP. Improving physicians' interviewing skills and reducing patients' emotional distress. A randomized clinical trial. Arch Intern Med 1995 Sep 25;155(17):1877-1884. [Medline: 7677554]
67. Ruesseler M, Sterz J, Bender B, Hoefer S, Walcher F. The effect of video-assisted oral feedback versus oral feedback on surgical communicative competences in undergraduate training. Eur J Trauma Emerg Surg 2017 Aug;43(4):461-466. [doi: 10.1007/s00068-016-0734-x] [Medline: 27730252]
68. Scardovi A, Rucci P, Gask L, et al. Improving psychiatric interview skills of established GPs: evaluation of a group training course in Italy. Fam Pract 2003 Aug;20(4):363-369. [doi: 10.1093/fampra/cmg404] [Medline: 12876103]
69. Setubal MSV, Antonio M, Amaral EM, Boulet J. Improving perinatology residents' skills in breaking bad news: a randomized intervention study. Rev Bras Ginecol Obstet 2018 Mar;40(3):137-146. [doi: 10.1055/s-0037-1621741] [Medline: 29342495]
70. Slort W, Blankenstein AH, Schweitzer BPM, Deliens L, van der Horst HE. Effectiveness of the "availability, current issues and anticipation" (ACA) training programme for general practice trainees on communication with palliative care patients: a controlled trial. Patient Educ Couns 2014 Apr;95(1):83-90. [doi: 10.1016/j.pec.2013.12.012] [Medline: 24412094]
71. Smith MM, Secunda KE, Cohen ER, Wayne DB, Vermylen JH, Wood GJ. Clinical experience is not a proxy for competence: comparing fellow and medical student performance in a breaking bad news simulation-based mastery learning curriculum. Am J Hosp Palliat Care 2023 Apr;40(4):423-430. [doi: 10.1177/10499091221106176] [Medline: 35641315]
72. Smith PEM, Fuller GN, Kinnersley P, Brigley S, Elwyn G. Using simulated consultations to develop communications skills for neurology trainees. Eur J Neurol 2002 Jan;9(1):83-87. [doi: 10.1046/j.1468-1331.2002.00339.x] [Medline: 11784381]
73. Supiot S, Bonnaud-Antignac A. Using simulated interviews to teach junior medical students to disclose the diagnosis of cancer. J Cancer Educ 2008;23(2):102-107. [doi: 10.1080/08858190701849437] [Medline: 18569245]
74. Trent ME, Butz A, Serwint J, Gauda E. Teaching adolescent health & cultural communication through simulation. Journal of Adolescent Health 2015 Feb;56(2):S44. [doi: 10.1016/j.jadohealth.2014.10.089]
75. Van Rossem I, Devroey D, De Paepe K, et al. A training game for students considering family medicine: an educational project report. JMedLife 2019 Oct;12(4):411-418. [doi: 10.25122/jml-2019-0056]
76. White AA, King AM, D'Addario AE, et al. Crowdsourced feedback to improve resident physician error disclosure skills: a randomized clinical trial. JAMA Netw Open 2024 Aug 1;7(8):e2425923. [doi: 10.1001/jamanetworkopen.2024.25923] [Medline: 39110461]
77. Yuan YY, Scott S, Van Horn N, Oke O, Okada P. Objective evaluation of a simulation course for residents in the pediatric emergency medicine department: breaking bad news. Cureus 2019 Jan 16;11(1):e3903. [doi: 10.7759/cureus.3903] [Medline: 30911458]
78. Zick A, Granieri M, Makoul G. First-year medical students' assessment of their own communication skills: a video-based, open-ended approach. Patient Educ Couns 2007 Oct;68(2):161-166. [doi: 10.1016/j.pec.2007.05.018] [Medline: 17640843]
79. Levine O, Myers J, Jyothi Kumar S, et al. Testing the ABCs of serious illness program for oncology trainees: a feasibility trial comparing different learning formats for a virtual communication curriculum. Palliat Med Rep 2025;6(1):436-445. [doi: 10.1177/26892820251376359] [Medline: 41132314]
80. Xiao Y, Lian G, Zhang J, et al. Efficacy of a smart glass-enhanced training programme for core doctor-patient communication skills among radiology residents in China. Eur Radiol Exp 2025 Sep 19;9(1):92. [doi: 10.1186/s41747-025-00630-w] [Medline: 40973881]
81. Abraham R. Turning constraints into opportunities: online delivery of communication skills simulation sessions to undergraduate medical students during the COVID-19 pandemic. pie 2021;39(4):57-71 [FREE Full text] [doi: 10.18820/2519593X/pie.v39.i4.5]

XSL•FO
RenderX

82. Afonso N, Kelekar A, Alangaden A. “I have a cough”: an interactive virtual respiratory case-based module. MedEdPORTAL 2020 Dec 17;16(101714390):11058. [doi: 10.15766/mep_2374-8265.11058] [Medline: 33365392]
83. Aluce LM, Cooper JJ, Emlet LL, et al. Bringing competency-based communication training to scale: a multi-institutional virtual simulation-based mastery learning curriculum for emergency medicine residents. Med Teach 2025 Mar;47(3):505-512. [doi: 10.1080/0142159X.2024.2345267] [Medline: 38803304]
84. Bittner A, Bittner J, Jonietz A, Dybowski C, Harendza S. Translating medical documents improves students’ communication skills in simulated physician-patient encounters. BMC Med Educ 2016 Feb 27;16(1):72. [doi: 10.1186/s12909-016-0594-4] [Medline: 26920138]
85. Booth E, McFetridge K, Ferguson E, Paton C. Teaching undergraduate medical students virtual consultation skills: a mixed-methods interventional before-and-after study. BMJ Open 2022 Jun 16;12(6):e055235. [doi: 10.1136/bmjopen-2021-055235] [Medline: 35710246]
86. Bramstedt KA, Prang M, Dave S, Shin PNH, Savy A, Fatica RA. Telemedicine as an ethics teaching tool for medical students within the nephrology curriculum. Prog Transplant 2014 Sep;24(3):294-297. [doi: 10.7182/pit2014289] [Medline: 25193732]
87. Clever SL, Novack DH, Cohen DG, Levinson W. Evaluating surgeons’ informed decision making skills: pilot test using a videoconferenced standardised patient. Med Educ 2003 Dec;37(12):1094-1099. [doi: 10.1046/j.1365-2923.2003.01713.x] [Medline: 14984115]
88. Daetwyler CJ, Cohen DG, Gracely E, Novack DH. eLearning to enhance physician patient communication: a pilot test of “doc.com” and “WebEncounter” in teaching bad news delivery. Med Teach 2010;32(9):e381-e390. [doi: 10.3109/0142159X.2010.495759] [Medline: 20795797]
89. Deming J, Horecki P, Brustad R, et al. A virtual communication workshop to increase confidence using telehealth modalities. WMJ 2024 May;123(2):124-126. [Medline: 38718241]
90. Godoy-Pozo J, Illesca Pretty M, Vidal Villa A, et al. Remote simulation with simulated patient: Initial clinical learning experience in medical students. Rev Med Chil 2023 Nov;151(11):1446-1455. [doi: 10.4067/s0034-98872023001101446] [Medline: 39270111]
91. Gür D, Offergeld C, Fabry G, Wünsch A. Communication training in otorhinolaryngology education: comparison of an online and a classroom-based training course. HNO 2024 May;72(5):334-340. [doi: 10.1007/s00106-024-01475-2] [Medline: 38597969]
92. Hayes JR, Ark T, Ruffalo L, Mumm B, Nowik J. “Mastering the difficult conversation” communications course: utilizing OSCEs and workshops to prepare learners for residency. PRiMER 2025;9(101726396):36. [doi: 10.22454/PRiMER.2025.631398] [Medline: 40727862]
93. Heller RE, Phillips Z, Wilhite JA, Sartori D, Zabar S, Hayes R. What happens in the Zoom?: a workplace-based assessment of residents’ post-discharge skills in a telemedicine visit with a standardized patient. J Gen Intern Med 2023;38(Supplement 2). [doi: 10.1007/s11606-023-08226-z] [Medline: 641879833]
94. Holmes RJ, Fischer J, Lowe J, Schell JO, Farouk SS, Sparks MA. Implementation and assessment of virtual standardized patient sessions to teach communication skills to nephrology fellows during COVID-19. J Am Soc Nephrol 2020;31(448). [doi: 10.1681/asn.2019111188] [Medline: 633703814]
95. Iammeechai W, Srikulmontri T, Siritongtaworn P, Ratta-Apha W. Attitudes and confidence in communication skills of fourth-year medical students after online small group discussion and peer role-play: a survey study. Acad Psychiatry 2025 Apr;49(2):136-141. [doi: 10.1007/s40596-024-01998-4] [Medline: 38918286]
96. Jones S, McNeil M, Rothenberger SD, Jeong K, Nikiforova T. Training internal medicine residents to perform telemedicine visits: a novel skill-based curriculum. MedEdPORTAL 2025;21(101714390):11540. [doi: 10.15766/mep_2374-8265.11540] [Medline: 40636467]
97. Khawand-Azoulai M, Kavensky E, Sanchez J, et al. An authentic learning experience for medical students on conducting a family meeting. Am J Hosp Palliat Care 2025 Sep;42(9):882-888. [doi: 10.1177/10499091241286089] [Medline: 39301669]
98. Knie K, Schwarz L, Frehle C, Schulte H, Taetz-Harrer A, Kiessling C. To Zoom or not to Zoom - the training of communicative competencies in times of COVID 19 at Witten/Herdecke University illustrated by the example of “sharing information”. GMS J Med Educ 2020;37(7):Doc83. [doi: 10.3205/zma001376] [Medline: 33364362]
99. Lenes A, Klasen M, Adelt A, et al. Crisis as a chance. A digital training of social competencies with simulated persons at the Medical Faculty of RWTH Aachen, due to the lack of attendance teaching in the SARS-Cov-2 pandemic. GMS J Med Educ 2020;37(7):Doc82. [doi: 10.3205/zma001375] [Medline: 33364361]
100. Mack MC, Claxton RN, Arnold RM, Maurer SH. Peds OncoTalk: a curriculum to teach communication skills to pediatric hematology/oncology fellows. Pediatr Blood Cancer 2025 Nov;72(11):e31912. [doi: 10.1002/pbc.31912] [Medline: 40692208]
101. Newcomb AB, Duval M, Bachman SL, Mohess D, Dort J, Kapadia MR. Building rapport and earning the surgical patient’s trust in the era of social distancing: teaching patient-centered communication during video conference encounters to medical students. J Surg Educ 2021;78(1):336-341. [doi: 10.1016/j.jsurg.2020.06.018] [Medline: 32709566]

XSL•FO
RenderX

102. Newcomb AB, Appelbaum RD, Kapadia M, et al. Implementation of a skills-based virtual communication curriculum for medical students interested in surgery. Global Surg Educ 2022;1(1):48. [doi: 10.1007/s44186-022-00054-9] [Medline: 38013712]
103. Pang JH, Finlay E, Fortner S, Pickett B, Wang ML. Teaching effective informed consent communication skills in the virtual surgical clerkship. J Am Coll Surg 2021 Jul;233(1):64-72. [doi: 10.1016/j.jamcollsurg.2021.04.026] [Medline: 34015451]
104. Phillips Z, Wong L, Crotty K, et al. Implementing an experiential telehealth training and needs assessment for residents and faculty at a Veterans Affairs primary care clinic. J Grad Med Educ 2023 Aug;15(4):456-462. [doi: 10.4300/JGME-D-22-00868.1] [Medline: 37637347]
105. Pozo P, Landino MC, Maga JM, et al. Situational awareness in telehealth: a virtual standardized patient case for transitioning preclinical to clinical medical students. MedEdPORTAL 2025;21(101714390):11517. [doi: 10.15766/mep_2374-8265.11517] [Medline: 40225042]
106. Rasalam R, Bandaranaike S. Virtual WIL clinics in medicine: overcoming the COVID-19 challenge. Int J Work Integr Learn 2020;21(5):573-585 [FREE Full text]
107. Rivet EB, Feldman M, Khandelwal S, et al. Adapting compassionate conversations for virtual mediated communication. J Surg Educ 2023 Sep;80(9):1296-1301. [doi: 10.1016/j.jsurg.2023.06.012] [Medline: 37423804]
108. Ruddock K, Herbert K, Neil C, Gajree N, Dempsey K. Immersive psychiatry simulation: a novel course for medical student training. BJPsych Open 2021 Jun;7(S1):S153-S154 [FREE Full text] [doi: 10.1192/bjo.2021.430]
109. Sasnal M, Miller-Kuhlmann R, Merrell SB, et al. Feasibility and acceptability of virtually coaching residents on communication skills: a pilot study. BMC Med Educ 2021 Sep 29;21(1):513. [doi: 10.1186/s12909-021-02936-w] [Medline: 34583691]
110. Taylor AD, Connolly J, Pearce C. A rural doctor's telehealth training program during the COVID-19 pandemic. Rural Remote Health 2024 Feb;24(1):8032. [doi: 10.22605/RRH8032] [Medline: 38296265]
111. Tsui AL, Chau SWH. Small-group, online, actor-as-instructor clinical interview training: a single-blind, randomised controlled study. East Asian Arch Psychiatry 2024 Dec;34(4):134-140. [doi: 10.12809/eaap2451] [Medline: 39743488]
112. Yudkowsky R, Valdes W, Raja S, Kiser R. Assessing residents' telehealth communication skills using standardised patients. Med Educ 2011 Nov;45(11):1155. [doi: 10.1111/j.1365-2923.2011.04104.x] [Medline: 21936864]
113. Geng W, Cao J, Hu J, et al. "Face-to-face" is not superior to "face-to-screen": comparing effects of online and offline communication skills course in postgraduate medical students. Front Med (Lausanne) 2025;12(1685789):1685789. [doi: 10.3389/fmed.2025.1685789] [Medline: 41458493]
114. Andrade AD, Bagri A, Zaw K, Roos BA, Ruiz JG. Avatar-mediated training in the delivery of bad news in a virtual world. J Palliat Med 2010 Dec;13(12):1415-1419. [doi: 10.1089/jpm.2010.0108] [Medline: 21091407]
115. Bearman M, Cesnik B. Comparing student attitudes to different models of the same virtual patient. Stud Health Technol Inform 2001;84(Pt 2):1004-1008. [Medline: 11604882]
116. Borg A, Jobs B, Huss V, et al. Enhancing clinical reasoning skills for medical students: a qualitative comparison of LLM-powered social robotic versus computer-based virtual patients within rheumatology. Rheumatol Int 2024 Dec;44(12):3041-3051. [doi: 10.1007/s00296-024-05731-0] [Medline: 39412574]
117. Bruen C, Kreiter C, Wade V, Pawlikowska T. Investigating a self-scoring interview simulation for learning and assessment in the medical consultation. Adv Med Educ Pract 2017;8:353-358. [doi: 10.2147/AMEP.S128321] [Medline: 28603434]
118. Carrard V, Bourquin C, Orsini S, Schmid Mast M, Berney A. Virtual patient simulation in breaking bad news training for medical students. Patient Educ Couns 2020 Jul;103(7):1435-1438. [doi: 10.1016/j.pec.2020.01.019] [Medline: 32019697]
119. Courteille O, Josephson A, Larsson LO. Interpersonal behaviors and socioemotional interaction of medical students in a virtual clinical encounter. BMC Med Educ 2014 Apr 1;14(1):64. [doi: 10.1186/1472-6920-14-64] [Medline: 24685070]
120. Detering K, Silvester W, Corke C, et al. Teaching general practitioners and doctors-in-training to discuss advance care planning: evaluation of a brief multimodality education programme. BMJ Support Palliat Care 2014 Sep;4(3):313-321. [doi: 10.1136/bmjspcare-2013-000450] [Medline: 24844586]
121. Deladisma AM, Cohen M, Stevens A, et al. Do medical students respond empathetically to a virtual patient? Am J Surg 2007 Jun;193(6):756-760. [doi: 10.1016/j.amjsurg.2007.01.021] [Medline: 17512291]
122. Foster A, Chaudhary N, Murphy J, Lok B, Waller J, Buckley PF. The use of simulation to teach suicide risk assessment to health profession trainees-rationale, methodology, and a proof of concept demonstration with a virtual patient. Acad Psychiatry 2015 Dec;39(6):620-629. [doi: 10.1007/s40596-014-0185-9] [Medline: 25026950]
123. Frey-Vogel AS, Ching K, Dzara K, Mallory L. The acceptability of avatar patients for teaching and assessing pediatric residents in communicating medical ambiguity. J Grad Med Educ 2022 Dec;14(6):696-703. [doi: 10.4300/JGME-D-22-00088.1] [Medline: 36591423]
124. Jacklin S, Maskrey N, Chapman S. Shared decision-making with a virtual patient in medical education: mixed methods evaluation study. JMIR Med Educ 2021 Jun 10;7(2):e22745. [doi: 10.2196/22745] [Medline: 34110299]
125. Kleinsmith A, Rivera-Gutierrez D, Finney G, Cendan J, Lok B. Understanding empathy training with virtual patients. Comput Human Behav 2015 Nov 1;52(8510313):151-158. [doi: 10.1016/j.chb.2015.05.033] [Medline: 26166942]

XSL•FO
RenderX

126. Kron FW, Fetters MD, Scerbo MW, et al. Using a computer simulation for teaching communication skills: a blinded multisite mixed methods randomized controlled trial. Patient Educ Couns 2017 Apr;100(4):748-759. [doi: 10.1016/j.pec.2016.10.024] [Medline: 27939846]
127. McCarrick CA, McEntee PD, Boland PA, et al. A randomized controlled trial of a deep language learning model-based simulation tool for undergraduate medical students in surgery. J Surg Educ 2025 Sep;82(9):103629. [doi: 10.1016/j.jsurg.2025.103629] [Medline: 40729832]
128. McCarthy DM, Formella KT, Ou EZ, et al. There's an app for that: teaching residents to communicate diagnostic uncertainty through a mobile gaming application. Patient Educ Couns 2022 Jun;105(6):1463-1469. [doi: 10.1016/j.pec.2021.09.038] [Medline: 34674922]
129. Mool A, Schmid J, Johnston T, et al. Using generative AI to simulate patient history-taking in a problem-based learning tutorial: a mixed-methods study. Tech Know Learn 2026. [doi: 10.1007/s10758-025-09929-4]
130. Mukadam A, Suresh S, Jacobs C. Beyond traditional simulation: an exploratory study on the effectiveness and acceptability of ChatGPT-4o advanced voice mode for communication skills practice among medical students. Cureus 2025 May;17(5):e84381. [doi: 10.7759/cureus.84381] [Medline: 40535400]
131. Poulose P. Evaluating the role of AI-simulated patients compared with peer-to-peer learning models in the enhancement of medical education: is it beyond theoretical functionality? Future Healthcare Journal 2025 Jun;12(2):100398. [doi: 10.1016/j.fhj.2025.100398]
132. Raafat N, Harbourne AD, Radia K, Woodman MJ, Swales C, Saunders KEA. Virtual patients improve history-taking competence and confidence in medical students. Med Teach 2024 May;46(5):682-688. [doi: 10.1080/0142159X.2023.2273782] [Medline: 38084413]
133. Sezer B, Sezer TA. Teaching communication skills with technology: creating a virtual patient for medical students. AJET 2019;35(5):183-198. [doi: 10.14742/ajet.4957]
134. Thompson CM, Bishop MJ, Dillard TC, et al. Healing health care disparities: development and pilot testing of a virtual reality implicit bias training module for physicians in the context of Black maternal health. Health Commun 2025 Mar;40(3):445-456. [doi: 10.1080/10410236.2024.2347000] [Medline: 38711251]
135. Wang Z, Fan TT, Li ML, Zhu NJ, Wang XC. Feasibility study of using GPT for history-taking training in medical education: a randomized clinical trial. BMC Med Educ 2025 Jul 10;25(1):1030. [doi: 10.1186/s12909-025-07614-9] [Medline: 40640776]
136. Yamamoto A, Koda M, Ogawa H, et al. Enhancing medical interview skills through AI-simulated patient interactions: nonrandomized controlled trial. JMIR Med Educ 2024 Sep 23;10:e58753. [doi: 10.2196/58753] [Medline: 39312284]
137. Chiu J, Castro B, Ballard I, et al. Exploration of the role of ChatGPT in teaching communication skills for medical students: a pilot study. Med Sci Educ 2025 Aug;35(4):1871-1882. [doi: 10.1007/s40670-025-02394-9] [Medline: 41112920]
138. Comulada WS, Ganz PA, Huang YM, et al. A pilot test of an AI voice-driven simulation with feedback for medical students to practice discussing diagnostic mammogram results With patients. Cureus 2025 Oct;17(10):e95606. [doi: 10.7759/cureus.95606] [Medline: 41322766]
139. Dávidovics A, Dávidovics K, Hillebrand P, Rendeki S, Németh T, Dávidovics A. Virtual patient simulation to enhance medical students' clinical communication and decision-making skills: a pilot study. BMC Med Educ 2025 Dec 30;26(1):171. [doi: 10.1186/s12909-025-08507-7] [Medline: 41469878]
140. Herschbach L, Festl-Wietek T, Stegemann-Philipps C, et al. Evaluation of an AI-based chatbot providing real-time feedback in communication training for mental health care professionals: proof-of-concept observational study. J Med Internet Res 2025 Nov 28;27:e82818. [doi: 10.2196/82818] [Medline: 41314643]
141. Jadoon M, Naushad K, Aman F, Khan Y, Durrani M, Ali S. Virtual patients for communication skills training: a mixed methods evaluation. Cureus 2025 Aug;17(8):e91005. [doi: 10.7759/cureus.91005] [Medline: 41018381]
142. Lee HY, Kim J, Choi H, et al. Comparing AI chatbot simulation and peer role-play for OSCE preparation: a pilot randomized controlled trial. BMC Med Educ 2025 Nov 24;25(1):1755. [doi: 10.1186/s12909-025-08308-y] [Medline: 41286823]
143. Suárez-García RX, Chavez-Castañeda Q, Orrico-Pérez R, et al. DIALOGUE: a generative AI-based pre-post simulation study to enhance diagnostic communication in medical students through virtual type 2 diabetes scenarios. Eur J Investig Health Psychol Educ 2025 Aug 7;15(8):152. [doi: 10.3390/ejihpe15080152] [Medline: 40863274]
144. Tyrrell EG, Sandhu SK, Berry K, et al. Web-based AI-driven virtual patient simulator versus actor-based simulation for teaching consultation skills: multicenter randomized crossover study. JMIR Form Res 2025 Nov 20;9:e71667. [doi: 10.2196/71667] [Medline: 41264856]
145. Young DG, Herrmann LE, Adeyanju O, et al. Remote VR supports medical students' communication. Clin Teach 2025 Dec;22(6):41236383. [doi: 10.1111/tct.70251] [Medline: 41236383]
146. Ba H, Zhang L, Yi Z. Enhancing clinical skills in pediatric trainees: a comparative study of ChatGPT-assisted and traditional teaching methods. BMC Med Educ 2024 May 22;24(1):558. [doi: 10.1186/s12909-024-05565-1] [Medline: 38778332]
147. Cheloff AZ, Johnson GM, Joseph NP, Fernandez L, Cluett JL, Kriegel GR, et al. Engaging medical students in communication with primary care patients through the patient portal: lessons during COVID-19. J Gen Intern Med 2021;36(SUPPL 1). [doi: 10.1007/s11606-021-06830-5] [Medline: 635796828]

XSL•FO
RenderX

148. Herrmann-Werner A, Loda T, Zipfel S, Holderried M, Holderried F, Erschens R. Evaluation of a language translation app in an undergraduate medical communication course: proof-of-concept and usability study. JMIR Mhealth Uhealth 2021 Dec 2;9(12):e31559. [doi: 10.2196/31559] [Medline: 34860678]
149. Sun C, Zou J, Zhao L, et al. New doctor-patient communication learning software to help interns succeed in communication skills. BMC Med Educ 2020 Jan 8;20(1):8. [doi: 10.1186/s12909-019-1917-z] [Medline: 31914987]
150. White AA, King AM, D'Addario AE, et al. Effects of practicing with and obtaining crowdsourced feedback from the video-based communication assessment app on resident physicians' adverse event communication skills: pre-post trial. JMIR Med Educ 2022 Oct 3;8(4):e40758. [doi: 10.2196/40758] [Medline: 36190751]
151. Dickerson R, Johnsen K, Raij A, et al. Virtual patients: assessment of synthesized versus recorded speech. Stud Health Technol Inform 2006;119:114-119. [Medline: 16404028]
152. Cimini A. 15th National Congress of the Italian Association of Nuclear Medicine and Molecular Imaging (AIMN). Clin Transl Imaging 2022;10(S1):1-111. [doi: 10.1007/s40336-022-00492-x]
153. Sanson-Fisher R, Hobden B, Waller A, Dodd N, Boyd L. Methodological quality of teaching communication skills to undergraduate medical students: a mapping review. BMC Med Educ 2018 Jun 27;18(1):151. [doi: 10.1186/s12909-018-1265-4] [Medline: 29945587]
154. Liao F, Murphy D, Wu JC, Chen CY, Chang CC, Tsai PF. How technology-enhanced experiential e-learning can facilitate the development of person-centred communication skills online for health-care students: a qualitative study. BMC Med Educ 2022 Jan 25;22(1):60. [doi: 10.1186/s12909-022-03127-x] [Medline: 35078482]
155. Car J, Ong QC, Erlikh Fox T, et al. The digital health competencies in medical education framework: an international consensus statement based on a Delphi study. JAMA Netw Open 2025 Jan 2;8(1):e2453131. [doi: 10.1001/jamanetworkopen.2024.53131] [Medline: 39888625]
156. Kelly S, Smyth E, Murphy P, Pawlikowska T. A scoping review: virtual patients for communication skills in medical undergraduates. BMC Med Educ 2022 Jun 3;22(1):429. [doi: 10.1186/s12909-022-03474-9] [Medline: 35659213]
157. Bylund CL, Vasquez TS, Peterson EB, et al. Effect of experiential communication skills education on graduate medical education trainees' communication behaviors: a systematic review. Acad Med 2022 Dec 1;97(12):1854-1866. [doi: 10.1097/ACM.0000000000004883] [Medline: 35857395]
158. Dorrestein L, Ritter C, De Mol Z, et al. Validity evidence for communication skills assessment in health professions education: a scoping review. BMJ Open 2025 Sep 5;15(9):e096799. [doi: 10.1136/bmjopen-2024-096799] [Medline: 40912699]
159. Tan XH, Foo MA, Lim SLH, et al. Teaching and assessing communication skills in the postgraduate medical setting: a systematic scoping review. BMC Med Educ 2021 Sep 9;21(1):483. [doi: 10.1186/s12909-021-02892-5] [Medline: 34503497]
160. Blackmore A, Kasfiki EV, Purva M. Simulation-based education to improve communication skills: a systematic review and identification of current best practice. BMJ Simul Technol Enhanc Learn 2018;4(4):159-164. [doi: 10.1136/bmjstel-2017-000220] [Medline: 35519010]
161. Gutiérrez Hernández A, Pompa Mansilla M, Vadillo Bueno G, Sánchez-Mendiola M. Digital learning in motion: exploring mobile device use in Mexican residents. J Grad Med Educ 2025 Dec;17(6):705-712. [doi: 10.4300/JGME-D-24-01024.1] [Medline: 41415999]
162. Jarvis P. Towards a Comprehensive Theory of Human Learning: Routledge; 2012. [doi: 10.4324/9780203001677]
163. Illeris K. Transformative Learning and Identity 2013. [doi: 10.4324/9780203795286]
164. Dayan E, Cohen LG. Neuroplasticity subserving motor skill learning. Neuron 2011 Nov 3;72(3):443-454. [doi: 10.1016/j.neuron.2011.10.008] [Medline: 22078504]

Abbreviations

AI: artificial intelligence
CINAHL: Cumulative Index to Nursing and Allied Health Literature
CST: communication skills training
ERIC: Educational Resource Information Center
JBI: Joanna Briggs Institute
PRISMA-S: Preferred Reporting Items for Systematic Reviews and Meta-Analyses Extension for Literature Searches
PRISMA-ScR: Preferred Reporting Items for Systematic Reviews and Meta-Analyses Extension for Scoping Reviews
VP: virtual patient

XSL•FO
RenderX

Edited by S Brini; submitted 03.Nov.2025; peer-reviewed by R Patel, R Knox; revised version received 17.Mar.2026; accepted 17.Mar.2026; published 20.Apr.2026.

Please cite as:
Seripenah P, Emery H, Patel B, Tyrrell E, Carson J, Leonardi-Bee J, Evans C, Wilson E, Taggar J
Use of Digital Technology for Developing Communication Skills in Undergraduate and Postgraduate Medical Education: Scoping Review
JMIR Med Educ 2026;12:e87012
URL: https://mededu.jmir.org/2026/1/e87012
doi:10.2196/87012

© Princella Seripenah, Heidi Emery, Bakula Patel, Edward Tyrrell, Julie Carson, Jo Leonardi-Bee, Catrin Evans, Emma Wilson, Jaspal Taggar. Originally published in JMIR Medical Education (https://mededu.jmir.org), 20.Apr.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Education, is properly cited. The complete bibliographic information, a link to the original publication on https://mededu.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

The Effectiveness of Artificial Intelligence in Undergraduate Health Professions Education: Systematic Review and Meta-Analysis of Randomized Controlled Trials

Nai Ming Lai[1], MBBS, MRCP, MRCPCH; Yin Sear Lim[1], MBBS, MRCPCH; Min Thein Win[1], MMedSc, MD; Prabal Bhargava[1], MD, DNB (Ophthal); Paraidathathu Thomas[2], PhD; Qi Chwen Ong[3], MBBS, MPH

[1]School of Medicine, Faculty of Health and Medical Sciences, Taylor's University, Subang Jaya, Selangor, Malaysia

[2]Faculty of Health and Medical Sciences, Taylor's University, Subang Jaya, Selangor, Malaysia

[3]Division of Cardiovascular Medicine, Radcliffe Department of Medicine, John Radcliffe Hospital, University of Oxford, Level 6, West Wing, John Radcliffe Hospital, Headington, Oxford, England, United Kingdom

Corresponding Author:
Qi Chwen Ong, MBBS, MPH
Division of Cardiovascular Medicine, Radcliffe Department of Medicine, John Radcliffe Hospital, University of Oxford, Level 6, West Wing, John Radcliffe Hospital, Headington, Oxford, England, United Kingdom

Abstract

Background: Health professions education faces increasing challenges from rising health care complexity, pedagogical shifts, and constrained curricular space, and rapidly expanding knowledge and technological advances. While artificial intelligence (AI) shows promise for transforming health professions education, evidence of its effectiveness remains unclear.

Objective: This study synthesized evidence from randomized controlled trials (RCTs) on the effectiveness of AI in undergraduate health professions education.

Methods: We included RCTs, randomized crossover trials, and cluster RCTs comparing AI against standard educational interventions at the undergraduate level. We excluded quasi-experimental studies and those without clear AI components. We searched PubMed, Cochrane, Embase, Educational Resources Information Center, and Web of Science up to January 26, 2026. Outcomes were categorized according by Kirkpatrick levels; risk of bias was assessed using the Risk Of Bias Instrument for Use in Systematic Reviews for Randomised Controlled Trials tool; random-effects meta-analysis was conducted in RevMan (Cochrane); and certainty of evidence was rated using the Grading of Recommendations, Assessment, Development, and Evaluation approach. AI interventions were subcategorized by technology type and educational functions, yielding 13 subcategories.

Results: Of 39,783 records identified, 66 RCTs (N=4911 participants; 2020-2026) were included. Subcategorized analyses across 7 outcome domains yielded 48 comparisons. Most studies had high risk of bias, mainly due to poor allocation concealment and blinding, and certainty of evidence ranged from low to very low. Large language model (LLM)–based personalized learning aids comprised the largest evidence base and showed positive effects for satisfaction (standardized mean difference [SMD] 0.93, 95% CI 0.40-1.46; 7 studies; 430 participants; I^2=74%), confidence (SMD 0.91, 95% CI 0.54-1.29; 7 studies; 609 participants; I^2=64%), and theoretical knowledge (SMD 0.53, 95% CI 0.13-0.94; 12 studies; 955 participants; I^2=86%), all with very low certainty. Other AI subtypes, including LLM content generators, natural language processing (NLP) chatbots, and non-LLM adaptive learning platforms, showed generally favorable point estimates but substantial heterogeneity and wide CIs, often included no effect. Prediction intervals frequently crossed the null, indicating uncertainty across educational setting. No studies assessed Kirkpatrick levels 3 or 4.

Conclusions: This review synthesized RCT evidence on AI in undergraduate health professions education by technology type and function, incorporating evidence certainty. Despite the large number of included studies, evidence remains insufficient to inform educational practice. Some AI interventions may improve some learning outcomes, but effects are inconsistent and not reliably reproducible. High risk of bias, heterogeneity, imprecision, and absence of higher-level outcomes limit conclusions. AI applications should therefore be used cautiously and on a trial basis.

Trial Registration: PROSPERO CRD42021243832; https://www.crd.york.ac.uk/PROSPERO/view/CRD42021243832

(JMIR Med Educ 2026;12:e88933) doi:10.2196/88933

KEYWORDS

health professions education; undergraduate; artificial intelligence; meta-analysis; educational resource; large languge model

XSL•FO
RenderX

Introduction

Background

Health professions education is rapidly expanding, with an estimated growth rate of 13.3% per year in the global health education market from 2024, reaching US $264 billion by 2030, according to a market analysis report [1]. This growth reflects rising health care complexity, expanding knowledge bases, increasing pedagogical demands, evolving learner expectations, and technological capabilities [2,3]. Traditional health professions education approaches struggle to keep pace with these changes because of challenges in incorporating innovations within the constraints of curricular space [4] and resistance from educators and stakeholders [5]. Educators face substantial pressure to deliver comprehensive curricular content while ensuring understanding and practical competency in a personalized, student-directed learning paradigm [6]. The challenge is compounded by the lack of strong evidence linking the adoption of any educational approach to consistent improvements in a demonstrated chain of relevant outcomes, from learning perception and knowledge to real-life practice and eventual health outcomes [7,8]. The COVID-19 pandemic accelerated the adoption of digital technologies in health professions education but also highlighted their limitations [9]. Artificial intelligence (AI), which has been applied in health professions education for over 2 decades [10], is now seen as a potentially powerful option to address these challenges given its recent rapid advancements.

A widely accepted definition of AI is the "science of making machines do things that would require intelligence if done by humans" [11]. AI in its many forms, including machine learning (ML), deep learning (DL), and artificial neural networks, has been used to analyze data patterns and make predictions with efficiency and consistency unmatched by humans [12]. Generative AI, a form of DL, has rapidly gained traction across a wide range of fields, including education, following the introduction of ChatGPT (OpenAI), a large language model (LLM), in November 2022 [13].

In health professions education, AI has been used in teaching-learning and assessment for over 2 decades [10]. AI-powered systems for teaching-learning include intelligent tutoring systems [14], adaptive learning platforms [15], and surgical training simulators such as the virtual operative assistant [16]. In assessment, AI has been used to automate student scoring and predict future performance in health disciplines [17]. To this end, LLMs have emerged as particularly influential, supporting clinical learning note compilation and summarization, virtual patient simulation, and personalized mentoring while facilitating self-directed learning and writing assistance [18]. While some studies have reported impressive performance of LLMs in undergraduate and postgraduate medical examinations [19], others have shown considerable gaps between LLMs and human experts in clinical diagnostic abilities [20]. The findings raise concerns about the appropriate use and limitations of LLMs in contemporary medical education [21].

Rationale

Despite extensive AI use in health professions education, fundamental questions about effectiveness remain unclear. While one recent meta-analysis examined early studies on the use of generative AI in medical education [22], no prior review has comprehensively synthesized randomized controlled trial (RCT) evidence across all forms of AI, including ML, DL, and generative AI, specifically in undergraduate health professions education, which differs from postgraduate training in pedagogical approaches [23]. Published reviews on AI in health professions education are mostly scoping or narrative reviews that are useful in providing guidance on the extent, facilitators, and barriers of AI. However, they do not provide quantitative effect estimates, which are essential for educators and researchers who seek to understand the magnitude of benefits or harm from AI applications, particularly for undergraduate education. The surge in AI publications following ChatGPT's November 2022 introduction [13] has created a critical mass of RCT evidence that is rapidly expanding, enabling meaningful meta-analysis.

Objectives

This systematic review and meta-analysis aimed to synthesize evidence on the effects of AI educational applications, compared with standard teaching methods, on learning outcomes among undergraduate health professions students.

Methods

Eligibility Criteria

We included RCTs, randomized crossover, and cluster-randomized trials that enrolled undergraduate students of health sciences, including medicine, pharmacy, nursing, and allied health. We accepted studies conducted in any health training setting, including clinical teaching that involved patient engagement, skills training, classroom learning in the form of problem-based, task-based, or case-based learning, written assignments, as well as any form of assessment across all periods of follow-up. We excluded studies that evaluated postgraduate or residency training to limit the scope of our review, given the sufficient number of studies published at the undergraduate level. We imposed no restrictions on publication year, language, or report format (abstract or full paper).

In terms of intervention, we accepted studies that explicitly reported the use of AI, which included ML, DL, or generative AI, chiefly LLMs, used for teaching and assessment purposes, either as the main or supplementary tool. We only included studies that evaluated robotics and virtual or augmented reality with a clear description of AI involvement, as these applications may or may not involve an operating AI component [24]. We classified AI interventions as generative AI if they were explicitly described as using LLMs or other generative architectures capable of producing novel text, images, or dialogue in response to user input (eg, ChatGPT, GPT-4–based tutors (OpenAI), and LLM-driven virtual patients). We classified all remaining AI interventions as nongenerative AI, including ML and DL applications that analyze or classify data without generating novel output (eg, imaging diagnostic aids, automated

XSL•FO
RenderX

scoring systems, and rule-based chatbots). The comparison included conventional methods of teaching and assessment involving direct human face-to-face input, as well as the use of various teaching-learning technologies except AI.

The review outcomes included any form of learning gain, evaluated using various scales in accordance with the levels depicted in the Kirkpatrick Model of Evaluation, a framework developed by Kirkpatrick in 1959 [25] and that has since been widely used in evaluating all forms of training. The list of outcomes is detailed under the subsequent subheading of Data Items. We included studies that fulfilled the criteria in terms of population and intervention, regardless of whether the study reported relevant outcomes or produced suitable outcome data for our meta-analysis.

Information Sources

We searched MEDLINE (PubMed), the Cochrane Central Register of Controlled Trials (which included records from PubMed, Embase, CINAHL, and trial registers, including the World Health Organization (WHO) International Trials Registry Platform and ClinicalTrials.gov), Embase, Web of Science (Ovid), and Educational Resources Information Center for published studies and trial register records up to January 26, 2026. We searched the reference lists of relevant reviews for additional studies but did not identify any. We did not search conferences or other online sources or contact authors or personnel for additional studies.

Search Strategies

We report the search strategies following the PRISMA-S (Preferred Reporting Items for Systematic Reviews and Meta-Analyses–Search extension) guideline [26]. The search strategies were developed by an author (NML) who has extensive systematic review experience, and were reviewed by another author (QCO) with systematic review experience (refer to Part 2 in Multimedia Appendix 1 for search strategies). We searched without applying any preset search filter or any language and publication type restriction.

Data Collection Process

Two authors (NML and YSL) independently screened titles and abstracts for shortlisting and evaluated shortlisted papers in full text to determine eligibility, after deduplication using EndNote (version 19; Clarivate). One author (NML) used Claude Sonnet (Anthropic) [27], an LLM application, to perform preliminary data extraction, including population, intervention, comparison, and outcomes, and verified accuracy against the full texts. We provided the LLM with structured prompts that included the overall context, aided by the background information (background and methodology of the current review) as well as a data extraction spreadsheet template with desired headings, assigning the LLM a role as an expert in systematic review methodology, and instructing the platform to read and extract stipulated data types in a step-by-step manner, pointing out the location of the data source to facilitate human verification (examples of prompts are available in Part 11 in Multimedia Appendix 1), and manually transcribed the verified data into an Excel (Microsoft) spreadsheet, while a second author (QCO) cross-checked the extracted data by NML against the full texts. We assessed multiple publications of the same study via the setting and characteristics as enumerated above and only selected the reference with the most complete description of the study as the primary reference. We resolved disagreements by discussion leading to a consensus, with referral to the third author (MTW) as required.

Data Items

We extracted participant characteristics, including setting (field of health professions education, eg, medicine, nursing, and dentistry), specific subject or skills evaluated (eg, clinical skills, ophthalmology, radiograph interpretation, and denture mounting), the region where the study was conducted, intervention technologies (eg, LLM or non-LLM and specific technology used, including the main purpose of teaching and learning vs assessment), frequency and duration of the intervention, assessment and follow-up period, outcomes reported, and funding sources.

We categorized the outcomes, based on the Kirkpatrick Model of Training Evaluation into perceptionor satisfaction and self-efficacy or confidence (level 1), knowledge gain (theoretical knowledge, clinical skills, practical skills [including competence and task efficiency], and generic or personal skills) (level 2), behavioral change assessed from real-life practice (level 3), and improvement in outcomes at the level of the recipient, namely, patient health outcomes (level 4) [28].

Assessment of Missing Data

If we found a significant dropout rate (>20%), we would judge the study at high risk of bias in terms of missing outcome data. We did not contact any author to request further information, as we did not consider any missing data to be critical for our meta-analysis.

Study Risk-of-Bias Assessment

Two authors (NML and QCO) independently assessed risk of bias using the ROBUST-RCT (Risk Of Bias Instrument for Use in Systematic Reviews for Randomised Controlled Trials) tool, developed by Wang et al [29], with adaptation in the domain of blinding from “health providers” to “education providers.” The tool consisted of 6 core domains (random sequence generation, allocation concealment, blinding of participants, blinding of educational providers, blinding of outcome assessors, and outcome data not included in the analysis), with 8 optional domains that we did not assess. Detailed guidance, including suggested rules and a decision table of the tool, is freely available in the additional material of the paper by Wang et al [29].

Risk of bias was assessed separately for subjective and objective outcomes, particularly for items related to blinding of education providers. For subjective outcomes, we split blinding of providers (Item 4) into 2 subdomains. Item 4a (perception and satisfaction) was rated probably low risk, as anonymous self-reporting by adult learners is unlikely to be systematically influenced by unblinded providers in the absence of evidence of coercion or deliberate influence on ratings. Item 4b (confidence and self-perceived competence) was rated probably high risk by default, as unblinded providers can plausibly

XSL•FO
RenderX

influence learners' self-appraisals through differential feedback, encouragement, and validation beyond the educational content delivered; exceptions were made for programs that were largely self-directed or used highly standardized provider-learner interactions. For objective outcomes, Item 4 was similarly rated probably high risk by default, as differential input from unblinded providers can affect actual competence even when outcomes are objectively scored, with exceptions for standardized or self-directed programs.

In ROBUST-RCT, there is no explicit rule provided in assigning the overall risk-of-bias status. Consequently, we described the risk of bias per study per domain and evaluated the overall degree of concern regarding risk of bias in our certainty of evidence rating by judging the proportion of high-risk domains in the body of evidence, as detailed in Part 1 in Multimedia Appendix 1. For example, if the majority of the included studies for the outcome assessed had a single high-risk domain (with blinding as a whole considered as 1 domain for this purpose), we would consider the body of evidence to have serious concerns in terms of risk of bias, and downgrade the certainty of the evidence by 1 level due to study limitations. However, if the majority of the included studies had 2 or more high-risk domains, we would consider the body of evidence to have very serious concerns in terms of risk of bias and downgrade the certainty of the evidence by 2 levels due to study limitations.

Effect Measures

For continuous outcomes, we used standardized mean difference (SMD) to pool the study results, as each included study reported its outcome data using different scales. We used Cohen benchmarks as a descriptive reference (SMD <0.2 small, 0.2 - 0.5 small to medium, >0.5 - 0.8 medium to large, and >0.8 large) [30], while recognizing that these thresholds were not developed for educational intervention contexts and should not be interpreted as definitive indices of educational importance [31]. Accordingly, we contextualized effect sizes using a default minimally important difference of SMD 0.5, consistent with the empirically derived half-standard-deviation rule [32], applying this threshold in our interpretation of synthesis results. For dichotomous outcomes, we reported the results using relative risk (RR).

Synthesis Methods

Studies with suitable numerical outcome data, either reported in text, table or derivable from figures were eligible for synthesis; those presenting results only narratively were excluded. We derived missing SD by multiplying the SE with the square root of the sample size in the corresponding group. For studies that reported their results in median and IQR, we approximated the median as the mean and obtained an estimate of the SD by dividing the IQR by 1.35, as recommended by Hozo et al [33]. Examples of studies from which we derived outcome data are available in Part 1 in Multimedia Appendix 1. For multiarm studies, we selected the relevant AI and control arms. In studies with multiple control groups, we selected the group representing the current standard educational intervention. We ensured each study appeared only once per analysis to avoid counting participants multiple times. For randomized crossover studies, we planned to include outcome data from the first period only, before the crossover, following one of the approaches suggested in the Cochrane Handbook [34]. If first-period data were not available, we would accept the data as reported by the authors and conduct sensitivity analysis to assess the impact of these crossover studies on the pooled results.

We tabulated major components of the intervention and comparison in each study in the characteristics of included studies table (Part 5 in Multimedia Appendix 1). We performed separate comparisons for each subcategory of AI application and divided the outcomes according to the levels of the Kirkpatrick Model of Training Evaluation [25], with further division for different subdomains based on the characteristics of the outcomes assessed, for example, self-efficacy and attitude (level 1) and theoretical knowledge, clinical, and practical skills (level 2), following discussion among the review authors.

If studies reported multiple outcomes in the same domain, we exercised our judgment in selecting the most suitable outcomes for our meta-analysis following a discussion among the review authors. For example, we used a combined score in preference to scores for individual components, and if no combined scores were reported, we selected one component that was most relevant to the domain; and if there were multiple components of equal relevance, we manually derived the combined mean and SD following the formula recommended in Chapter 10 of the Cochrane Handbook [35].

We presented our meta-analysis results graphically using forest plots. However, because of the large number of meta-analyses performed, we presented the synthesis results numerically in Table 1 and kept all the analysis details, including the forest plots in Multimedia Appendices 1 and 2.

XSL•FO
RenderX

Table . The full analysis results for each comparison between artificial intelligence (AI) subcategories and controlto with prediction intervals where applicable and certainty of evidence.

Outcome and specific comparison	Studies (participants)	I^2	Results (point estimate with 95% CI)	95% prediction interval	Certainty of evidence (GRADE[a])
Kirkpatrick level 1: satisfaction or motivation					
LLM[b] content generator vs control	5 (509)	96%	SMD[c] 0.65 (−0.73 to 2.04)	−2.67 to 3.98	Very low
LLM personalized learning aid vs control	7 (430)	74%	SMD 0.93 (0.40 to 1.46)	−0.40 to 2.26	Very low
LLM virtual patient vs control	3 (127)	64%	SMD 0.69 (−0.83 to 2.21)	−1.86 to 3.24	Very low
LLM content generator + LLM virtual patient + LLM personalized learning aid vs control	1 (88)	N/A	SMD −0.02 (−0.44 to 0.40)	N/A[h]	Very low
LLM-integrated curriculum vs control	1 (96)	N/A	SMD 1.31 (0.87 to 1.76)	N/A	Very low
Non-LLM AI[e]-moderated adaptive learning platform vs control	2 (143)	0%	SMD 0.55 (−1.00 to 2.11)	−1.00 to 2.11	Very low
NLP[f] rule-based chatbot vs control	2 (146)	74%	SMD 0.74 (−3.65 to 5.13)	−6.18 to 7.66	Very low
NLP rule-based chatbot + rule based virtual patient vs control	1 (61)	N/A	SMD 0.17 (−0.33 to 0.67)	N/A	Very low
NLP rule-based virtual patient vs control	1 (79)	N/A	SMD 0.75 (0.29 to 1.20)	N/A	Low
Kirkpatrick level 1: self-efficacy or confidence					
LLM personalized learning aid vs control	7 (609)	64%	SMD 0.91 (0.54 to 1.29)	0.05 to 1.77	Very low
LLM virtual patient vs control	2 (100)	91%	SMD 1.36 (−8.09 to 10.81)	−14.49 to 17.20	Very low
LLM-integrated curriculum vs control	1 (96)	N/A	SMD 1.23 (0.79 to 1.67)	N/A	Very low
Non-LLM AI procedure assistant vs control	1 (40)	N/A	SMD 0.00 (−0.62 to 0.62)	N/A	Very low
Non-LLM AI moderated adaptive learning platform vs control	1 (40)	N/A	SMD 2.45 (1.61 to 3.28)	N/A	Low
NLP rule-based chatbot vs control	2 (146)	43%	SMD 0.87 (−2.11 to 3.86)	−3.21 to 4.96	Very low
Non-LLM AI-VR[g] virtual doctor vs control	1 (64)	N/A	SMD −0.68 (−1.18 to−0.17)	N/A	Very low
AI procedure assistant + adaptive learning platform vs control	1 (20)	N/A	SMD 0.55 (−0.34 to 1.45)	N/A	Very low
AI procedure assistant (non-LLM) vs control (proportion of participants confident in echocardiography view)	1 (43)	N/A	RR 1.26 (0.45 to 3.50)	N/A	Very low

XSL•FO
RenderX

Outcome and specific comparison	Studies (participants)	I^2	Results (point estimate with 95% CI)	95% prediction interval	Certainty of evidence (GRADE[a])
Kirkpatrick level 2: theoretical knowledge score					
LLM content generator vs control	3 (359)	93%	SMD 0.99 (−1.04 to 3.01)	−2.96 to 4.93	Very low
LLM gamification tool vs control	1 (48)	N/A	SMD 0.79 (0.20 to 1.38)	N/A	Low
LLM personalized learning aid vs control	12 (955)	86%	SMD 0.53 (0.13 to 0.94)	−0.81 to 1.88	Very low
Non-LLM AI moderated adaptive learning platform vs control	1 (40)	N/A	SMD 0.68 (0.04 to 1.32)	N/A	Low
NLP rule-based chatbot vs control	3 (530)	97%	SMD 1.06 (−2.19 to 4.32)	−5.28 to 7.41	Very low
Non-LLM AI imaging diagnostic aid vs control	2 (69)	78%	SMD 1.26 (−6.23 to 8.74)	−10.73 to 13.24	Very low
Non-LLM AI-VR virtual doctor vs control	1 (64)	N/A	SMD 0.67 (0.16 to 1.17)	N/A	Very low
Proportion with grade A or B: non-LLM AI gamification tool vs control	1 (73)	N/A	RR 1.33 (1.01 to 1.74)	N/A	Very low
Kirkpatrick level 2: clinical skills					
LLM content generator vs control	2 (295)	97%	SMD 0.52 (−8.66 to 9.69)	−15.22 to 16.25	Very low
LLM personalized learning aid vs control	9 (609)	83%	SMD 0.49 (0.00 to 0.97)	−0.93 to 1.90	Very low
LLM virtual patient vs control	1 (56)	N/A	SMD 2.53 (1.82 to 3.25)	N/A	Low
LLM content generator + LLM virtual patient + LLM personalized learning aid vs control	1 (88)	N/A	SMD −0.11 (−0.53 to 0.31)	N/A	Very low
LLM virtual patient + LLM personalized learning aid vs control	3 (124)	95%	SMD 1.82 (−3.63 to 7.27)	−8.61 to 12.24	Very low
Non-LLM AI imaging diagnostic aid vs control	4 (176)	62%	SMD 0.43 (−0.41 to 1.27)	−1.13 to 1.99	Very low
Non-LLM AI moderated adaptive learning platform vs control	2 (139)	98%	SMD 0.59 (−19.98 to 21.17)	−34.78 to 35.97	Very low
NLP rule-based virtual patient vs control	1 (79)	N/A	SMD 2.03 (1.49 to 2.58)	N/A	Moderate
NLP rule-based chatbot + virtual patient vs control	1 (61)	N/A	SMD 0.24 (−0.26 to 0.74)	N/A	Very low
AI-VR virtual doctor vs control	1 (64)	N/A	SMD 0.21 (−0.28 to 0.71)	N/A	Very low
AI procedure assistant + AI-moderated adaptive learning platform vs control	1 (20)	N/A	SMD 0.80 (−0.12 to 1.72)	N/A	Very low

XSL•FO
RenderX

Outcome and specific comparison	Studies (participants)	I^2	Results (point estimate with 95% CI)	95% prediction interval	Certainty of evidence (GRADE[a])
Kirkpatrick level 2: practical skills					
LLM personalized learning aid vs control	1 (187)	N/A	SMD 0.67 (0.37 to 0.96)	N/A	Low
Non-LLM AI procedure assistant vs control	6 (305)	92%	SMD 0.18 (−0.97 to 1.34)	−2.77 to 3.13	Low
Kirkpatrick level 2: task efficiency (time taken to perform tasks)					
LLM personalized learning aid vs control	2 (100)	47%	SMD −0.15 (−4.24 to 3.95)	−5.99 to 5.69	Very low
Non-LLM AI imaging diagnostic aid vs control	1 (40)	N/A	SMD 2.70 (1.82 to 3.58)	N/A	Very low
Non-LLM AI procedure assistant vs control	2 (52)	88%	SMD −1.26 (−14.65 to 12.12)	−23.49 to 20.96	Very low
Non-LLM AI procedure assistant + AI-moderated adaptive learning platform vs control	1 (20)	N/A	SMD −0.87 (−1.80 to 0.05)	N/A	Very low
Kirkpatrick level 2: generic or personal skills					
LLM personalized learning aid vs control	1 (101)	N/A	SMD 0.45 (0.05 to 0.84)	N/A	Low
LLM virtual patient vs control	1 (27)	N/A	SMD 0.00 (−1.06 to 1.06)	N/A	Very low
LLM-integrated curriculum vs control	1 (96)	N/A	SMD 0.60 (0.19 to 1.01)	N/A	Moderate
Non-LLM AI communication analysis vs control	1 (25)	N/A	SMD 1.85 (0.88 to 2.81)	N/A	Low
AI-VR virtual doctor vs control	1 (64)	N/A	SMD 0.31 (−0.18 to 0.81)	N/A	Low

[a]GRADE: Grading of Recommendations, Assessment, Development, and Evaluation.

[b]LLM: large language model.

[c]SMD: standardized mean difference.

[d]Not applicable.

[e]AI: artificial intelligence.

[f]NLP: natural language processing.

[g]AI-VR: artificial intelligence–virtual reality.

We performed random-effects meta-analysis via the inverse-variance method using Review Manager (RevMan) online software (Cochrane), with between-study variance estimated using the restricted maximum likelihood method. We presented all point estimates with their 95% CIs, generated using the Hartung-Knapp-Sidik-Jonkman method, as simulation studies suggest this approach provides more appropriate coverage probabilities, particularly when heterogeneity is substantial [36,37]. To identify and quantify heterogeneity, we visually inspected the forest plots and used the I^2 statistic (detailed below in the "Assessment of Heterogeneity" subheading), with generation of a 95% prediction interval (PI) as an indication of the likely range of effects in individual studies [35].

Assessment of Heterogeneity

We evaluated heterogeneity statistically and explored plausible explanations in terms of educational characteristics of the studies [38]. We used the I^2 statistic with a cutoff of 50% to indicate substantial heterogeneity [38]. If substantial heterogeneity was found, we explored educational characteristics via subgroup analyses in terms of the field or program evaluated, the region where the study was conducted, the category of AI intervention delivered (divided broadly into LLM and non-LLM), the major purpose of intervention (teaching-learning or assessment), and the frequency of intervention (single or multiple sessions across a period of time). As part of heterogeneity exploration, we generated a 95% PI, which indicated the expected range of true

XSL•FO
RenderX

effects in similar individual studies [39]. AI subcategories and outcome types were not part of the heterogeneity exploration, as we had already separated the analyses according to these subcategories, as well as grouped the outcomes according to Kirkpatrick levels.

Sensitivity Analysis

We planned to conduct sensitivity analyses by restricting analyses to studies at low or probably low risk of bias for each domain where at least 5 studies were available in each risk stratum, consistent with Cochrane guidance [35]. However, sensitivity analysis was only feasible based on 2 risk-of-bias domains and for only 1 analysis, as detailed below in the Results section.

Publication and Reporting Bias Assessment

We created funnel plots to screen for small-study effects, a possible reason for publication bias, for outcomes in which there were ≥10 studies. If a funnel plot showed significant asymmetry, we would downgrade the certainty of the evidence based on a strong suspicion of publication bias from small-study effects [40]. However, in this review, the substantial heterogeneity observed across outcomes rendered funnel plot interpretation unreliable, especially when asymmetry is present [41]. Nonetheless, the single funnel plot generated in our review did not show significant asymmetry. We did not specifically assess the risk of bias from missing results in a meta-analysis (ie, reporting biases), using recently introduced risk-of-bias tools, such as the RoB-ME tool developed by Page et al [42].

Certainty of Evidence Rating

Two authors (NML and QCO) independently assessed the certainty of the evidence for 7 major outcomes for each comparison, namely, self-efficacy or confidence, attitude or satisfaction, theoretical knowledge, clinical skills, practical skills, task efficiency, and generic or personal skills using the Grading of Recommendations, Assessment, Development, and Evaluation approach [43]. We considered evidence from RCTs as high certainty to begin with, downgrading 1 level for serious (or 2 levels for very serious) limitations based upon 5 considerations: risk of bias, inconsistency across studies, indirectness of the evidence, imprecision of estimates, and publication bias (refer to Part 1 in Multimedia Appendix 1 for details). We used GRADEpro GDT (The GRADE Working Group) to create the summary of findings table (Part 3 in Multimedia Appendix 1).

Ethical Considerations

This work is a systematic review and meta-analysis of previously published studies; therefore, no ethics approval or consent to participate was required. The review was registered in PROSPERO (CRD42021243832). We conducted the review following Cochrane methods [38] and reported according to the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) 2020 guidelines [44] (Checklist 1), with particular reference to the PRISMA 2020 expanded checklist [45] in the structure of our paper. Additional details on our methods, including changes from the protocol, are described in Part 1 in Multimedia Appendix 1.

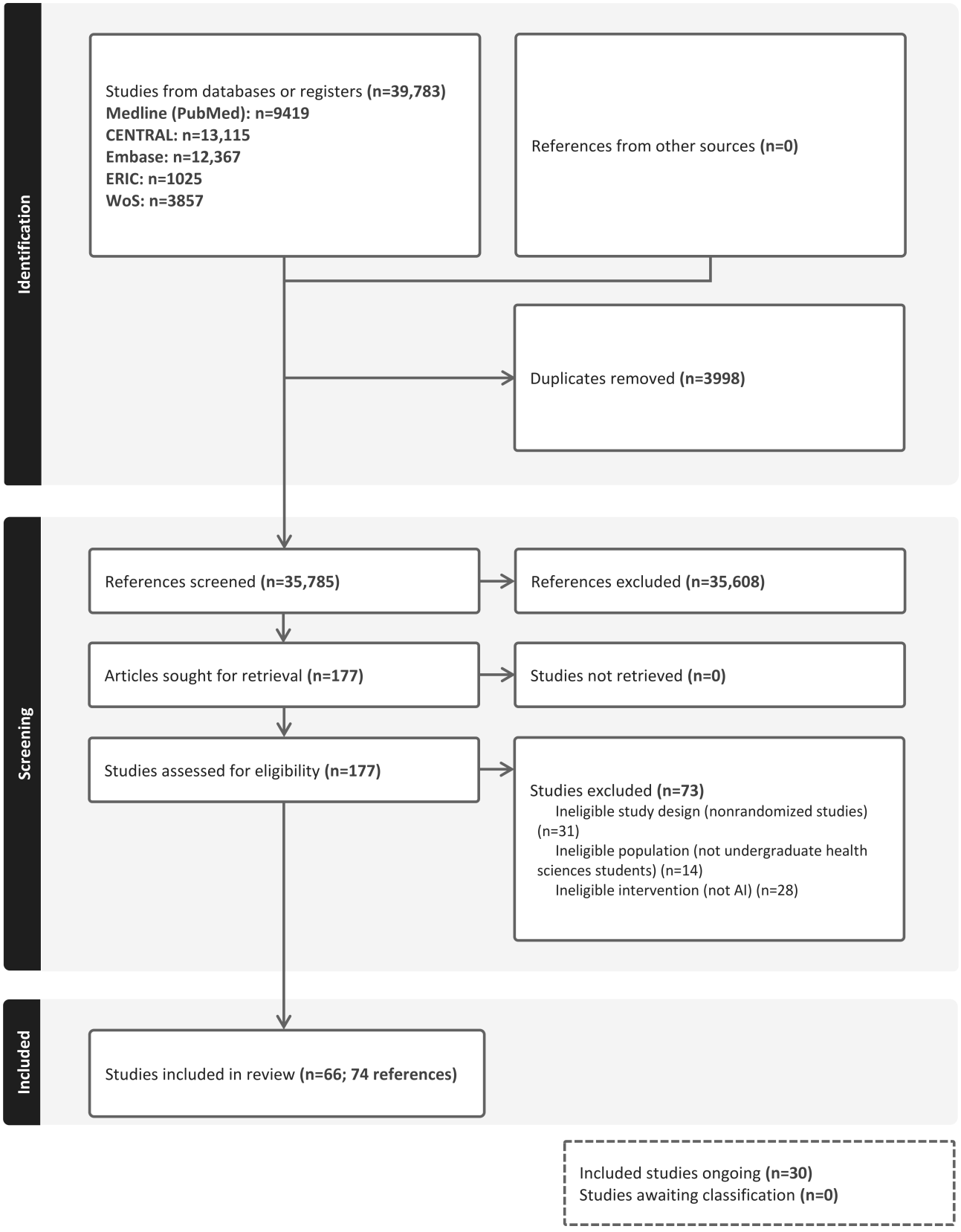
Results

Study Selection

The initial search identified 39,783 records, with 35,785 records remaining after deduplication. The vast majority of the initial records (n=35,608) were clearly nonrelevant from an inspection of the title or abstract, and these were rejected outright. We shortlisted 177 papers that appeared relevant and assessed their full texts in depth, excluding an additional 73 papers due to mismatch in study design (n=31, 42.5% papers), population (not undergraduate students; n=14, 19.2% papers), or intervention (not AI; n=28, 38.4% papers) (refer to Figure 1 for reasons for exclusion and Part 4 in Multimedia Appendix 1 for the full list). We also identified 30 relevant ongoing studies. Finally, we included 74 records that described 66 studies in our analysis, as 3 studies had multiple records, including 3 with 2 records [46-48], 1 with 3 records [49], and 1 with 4 records [50]. The flow diagram of the studies from the initial search to meta-analysis is shown in Figure 1.

XSL•FO
RenderX

Figure 1. PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) flow diagram depicting the process from initial screening to final inclusion. AI: artificial intelligence.



Study Characteristics

The characteristics of included studies are summarized as follows (refer to Part 5 in Multimedia Appendix 1 for a detailed tabulation).

Design and Setting

Out of 66 included studies (N=4911 participants) published between 2020 and 2026 [46-49,51-116], 58 (87.9%) were parallel RCTs, 7 (10.6%) were randomized crossover trials, and 1 (1.5%) was a cluster-randomized trial. Most studies were

XSL•FO
RenderX

conducted in China (n=14, 21.2%), followed by Turkey (n=11, 16.7%), the United States (n=5, 7.6%), Canada (n=4, 6.1%), Germany and Taiwan (n=3 each, 4.5%), and Italy, South Korea, Hong Kong, Morocco, Singapore and the United Kingdom (n=2 each, 3%), with a single study conducted in each of Colombia, Denmark, France, India, Ireland, Israel, Japan, New Zealand, Norway, Pakistan, Spain, and Thailand, with 2 multicenter studies involving different countries [66,84]. The number of participants per study ranged from 16 [47,79] to 426 [97], with 63 [82,102] as the median sample size.

Population and Educational Field

Over half of the included studies enrolled students from medicine (n=35, 53%), followed by nursing (n=15, 22.7%), dentistry (n=9, 13.6%), physiotherapy, pharmacy, and health sciences (n=2 each, 3%), and optometry (n=1, 1.5%). Within each program, there was a wide range of disciplines assessed. In medicine, these included preclinical sciences (physiology, anatomy, pathology, pharmacology, biochemistry, immunology, microbiology, and histology), as well as clinical specialties such as internal medicine, cardiorespiratory medicine, gastroenterology, pulmonology, obstetrics and gynecology, ophthalmology, urology, radiology, neurosurgery, surgery, and emergency and critical care medicine; generic skills such as clinical reasoning, evidence-based medicine, and clinical decision-making were also represented. In nursing, subfields included surgical nursing, critical care nursing, sepsis care, clinical skills, communication skills, medical terminology, and patient education. In dentistry, areas covered included dental radiology, endodontics, restorative and operative skills, and prosthodontics. Physiotherapy studies addressed clinical reasoning and chronic low back pain rehabilitation, while health sciences studies covered infection control and chronic disease management. Specific practical skills assessed across fields included simulated brain tumor resection, cardiac ultrasonography, and dental access cavity preparation.

Intervention

The studies used AI mainly as teaching-learning enhancement tools (n=54, 81.8%), as an assessment tool only (n=1, 1.5%) [104], or both (11, 16.7%). LLM applications were used in 38 (57.6%) studies, primarily featuring ChatGPT (versions 3.5 and 4.0; Open AI), alongside Gemini (Google), Copilot (Microsoft), Perplexity (Perplexity AI, Inc), Claude (Anthropic), and specialized chatbots. The remaining 28 (42.4%) studies used non-LLM AI systems composed of specific algorithms such as convolutional neural networks, linear support vector machines, NLP rule-based architectures, AI-powered virtual doctor or diagnostic systems, or intelligent monitoring and assessment systems, without any explicit use of LLM.

The AI applications were further categorized according to their predominant functions as described in the study. Among LLM studies, the applications included personalized learning aids (n=20), clinical content generators (n=6), virtual patients (n=6), combined virtual patient and personalized learning aid (n=3), a multifunction LLM combining content generation, virtual patient simulation, and personalized feedback (n=1), a gamification tool (n=1), and 1 study described as a general "LLM-integrated curriculum."

Among non-LLM studies, applications included AI procedure assistant (n=7), AI imaging diagnostic aid (n=7), NLP rule-based chatbot (n=4), and AI-moderated adaptive learning platform (n=4), as well as an artificial intelligence–virtual reality (AI-VR) virtual doctor, AI communication analysis system, AI gamification tool, rule-based virtual patient, a combined NLP rule-based chatbot and virtual patient, and a combined procedure assistant with adaptive learning platform (n=1 each). These applications aimed to assist in achieving competencies from foundational knowledge to complex clinical decision-making. A working description of each AI subcategory is available in Table 2. This categorization recognizes potential overlap between subcategories, as some studies incorporate multiple AI functionalities or serve overlapping educational purposes.

XSL•FO
RenderX

Table . Working description of artificial intelligence (AI) subcategories for the purpose of classification in this review.

AI[a] subcategory	Description
LLM[b]-based interventions	
LLM content generator	LLM-based applications generating educational materials (clinical vignettes, questions, explanatory text, case presentations, and learning guides).
LLM gamification tool	LLM-powered applications incorporating game-based elements (escape rooms and interactive challenges) to enhance learning engagement.
LLM personalized learning aid	LLM-based conversational tutors providing on-demand responses, explanations, postsimulation debriefing, and individualized feedback.
LLM virtual patient	LLM applications simulating patient encounters for clinical history-taking and communication skills practice.
LLM-integrated curriculum	Structured program with LLMs systematically embedded throughout curriculum delivery without specifying any component in learning.
Non-LLM AI interventions	
AI imaging diagnostic aid	AI systems using CNNs[c] or machine learning to analyze medical images for diagnostic support, lesion detection, or educational annotation.
AI gamification tool	AI-powered applications incorporating game design elements using specialized algorithms for interactive skill-based learning.
AI procedure assistant	AI systems providing real-time guidance, quality assessment, or feedback during practical skills training using machine learning classifiers or computer vision.
AI-moderated adaptive learning platform	Intelligent tutoring systems using algorithms (Deep Q-Networks and sentiment analysis) to personalize learning pathways and optimize content delivery.
AI-VR[d] virtual doctor	AI-powered virtual physician characters in VR simulations generating dynamic clinical responses through AI algorithms.
Natural language processing (NLP) rule-based chatbot	NLP chatbots using structured decision trees and preprogrammed dialogue patterns to deliver content and provide feedback.
Rule-based virtual patient	Virtual patient systems using decision tree algorithms to simulate patient responses through predetermined pathways.
AI communication analysis	AI systems using computer vision or behavioral analysis to assess and provide feedback on learner communication skills.

[a]AI: artificial intelligence.

[b]LLM: large language model.

[c]CNN: convolutional neural network.

[d]VR: virtual reality.

Study durations varied considerably, from single sessions of under 3 hours (n=24, 36.4%), short-term from 1 day to 3 weeks (n=19, 28.8%), medium-term from 1 to 2 months (n=13, 19.7%), and long-term from 2 to 6 months (n=5, 7.6%), with 5 studies not specifying duration.

Comparison

Over half of the studies (n=39, 59.1%) reported the control group as having received standard methods of learning as implemented in the existing curriculum, including student presentations, lectures, web-based tools, standard assessment procedures with existing technologies led by human tutors, and human-based simulation methods such as standardized patients, peer role-play, and real patient-based training. In 18 (27.3%) studies, authors reported that participants in the control group received the same educational content without AI tools. Four studies [53,72,79,100] included 2 control groups that used different sources of evidence: 1 from standard institutional resources and the other from external resources via online searches. In these studies, we selected the group that used standard institutional resources as the control group for our analysis.

Outcome Assessment

According to the Kirkpatrick Model of Evaluation [28], all studies measured outcomes either in level 1 (reaction) or level 2 (knowledge), and none reported outcomes in level 3 (real-life practice) or level 4 (health outcomes). The outcomes reported are summarized as follows:

- Level 1: 36 (54.5%) studies reported attitude, perception, or satisfaction, and 21 (31.8%) studies reported self-efficacy or confidence.

XSL•FO
RenderX

- Level 2: 26 (39.4%) studies reported theoretical knowledge, 30 (45.5%) studies reported clinical skills, 8 (12.1%) studies reported practical skills, 6 (9.1%) studies reported task efficiency, and 5 (7.6%) studies reported generic or personal skills.

Most outcomes were measured immediately after the intervention. Five studies [56,62,67,86,90] in addition reported outcomes at intermediate time points, ranging from 1 week to 3 months after the intervention.

Funding

Nine (13.6%) studies received funding from government-linked agencies, including national research councils and government ministries. Eighteen (27.2%) studies were funded by universities, hospitals, or dedicated research institutes. Three studies received contributions from commercial companies: 2 received in-kind contributions in the form of AI platform or device access [55,82], and 1 received direct financial sponsorship from a commercial AI company [98]. Seventeen (25.8%) studies declared no funding, while 19 (28.8%) did not provide any funding statement.

Risk of Bias in Studies

The full risk-of-bias assessment is shown in Part 6 in Multimedia Appendix 1. There are significant methodological limitations across the included studies, particularly in allocation concealment and blinding. Random sequence generation was adequate in 44 (66.7%) studies, but allocation concealment was adequate in only 15 (22.7%) studies. Nonblinding of participants resulted in high or probably high risk of bias from differential participant expectations in 55 (83.3%) studies.

Blinding of education providers was similarly not achieved in all but 4 studies. The risk implications, however, differed by outcome type. For participants' perception or satisfaction (assessed in 47 studies), nonblinding of providers was judged probably low risk in 40 (85.1%) studies, on the basis that anonymous self-reported evaluations of a teaching program by adult learners were unlikely to be materially influenced by unblinded providers. For self-efficacy or confidence (assessed in 24 studies), the majority (n=20, 83.3%) were judged probably high risk, as provider awareness of group allocation could plausibly influence student confidence through differential feedback, encouragement, and validation beyond the educational content itself; this concern did not apply to the 4 remaining studies, which used largely self-directed learning formats [47,68,70,80]. For objective outcomes, 40 (66.7%) studies were judged probably high risk and 20 (33.3%) probably low risk, the latter reflecting programs that were largely self-directed or sufficiently standardized to limit meaningful provider influence.

While none of the assessors for subjective, self-reported outcomes were blinded, blinding of outcome assessors was achieved in 29 out of 60 (48.3%) studies that evaluated objective outcomes, but almost all studies were judged to have low or probably low risk of bias because of clear documentation of objective assessment criteria, with only 2 studies [110,113] judged as probably high risk due to a lack of description of the scoring criteria despite reporting numerical checklists. In terms of missing outcome data, most studies (n=59, 89.4%) demonstrated low risk of bias.

Results of Individual Studies

Overall, 60 studies (90.9%; n=4506 participants) contributed suitable data for meta-analysis. The large number of AI subcategories resulted in numerous comparisons, most comprising few studies with small sample sizes. The results of individual studies are presented in the forest plots in Multimedia Appendix 2. None of the 7 randomized crossover trials provided separate data for the first period, and so we included data as reported in the paper where applicable.

Results of Synthesis

A full display of the synthesis results is shown numerically in Table 1 and graphically in Multimedia Appendix 2. Substantial heterogeneity was present across almost all pooled estimates and was not adequately explained by subgroup analyses based on field and region of study, LLM vs non-LLM, teaching-learning vs assessment, and single vs multiple sessions (Multimedia Appendix 2). Consequently, the certainty of evidence was rated low or very low for most comparisons (Table 1). Ninety-five percent PIs, reported for all comparisons with 2 or more studies, were wide throughout and almost universally included the null value, indicating highly uncertain effects of AI interventions in any new study setting.

The results are summarized below, organized by Kirkpatrick level.

Kirkpatrick Level 1: Participant Perception or Satisfaction

Twenty-three (1679 participants) studies assessed participant perception or satisfaction across 9 AI subtype comparisons (Table 1).

Among LLM-based comparisons, LLM personalized learning aids appeared to show an important positive effect (7 studies, n=430; SMD 0.93, 95% CI 0.40 to 1.46; I^2=74%; 95% PI −0.40 to 2.26; Figure 2 [48,70,80,84,113,115,116]), whereas LLM content generators (5 studies, n=509; SMD 0.65, 95% CI −0.73 to 2.04; I^2=96%; 95% PI −2.67 to 3.98; Figure 3 [49,51,61,62,100]) and LLM virtual patients (3 studies, n=127; SMD 0.69, 95% CI −0.83 to 2.21; I^2=64%, 95% PI −1.86 to 3.24) showed no clear difference, with substantial heterogeneity and wide PIs spanning the null. The remaining comparisons each involved 1 or 2 studies and were imprecise (Tables 1 and 2).

XSL•FO
RenderX

Figure 2. Forest plot for perception or satisfaction (large language model [LLM] personalized learning aid vs control) [48,70,80,84,113,115,116].

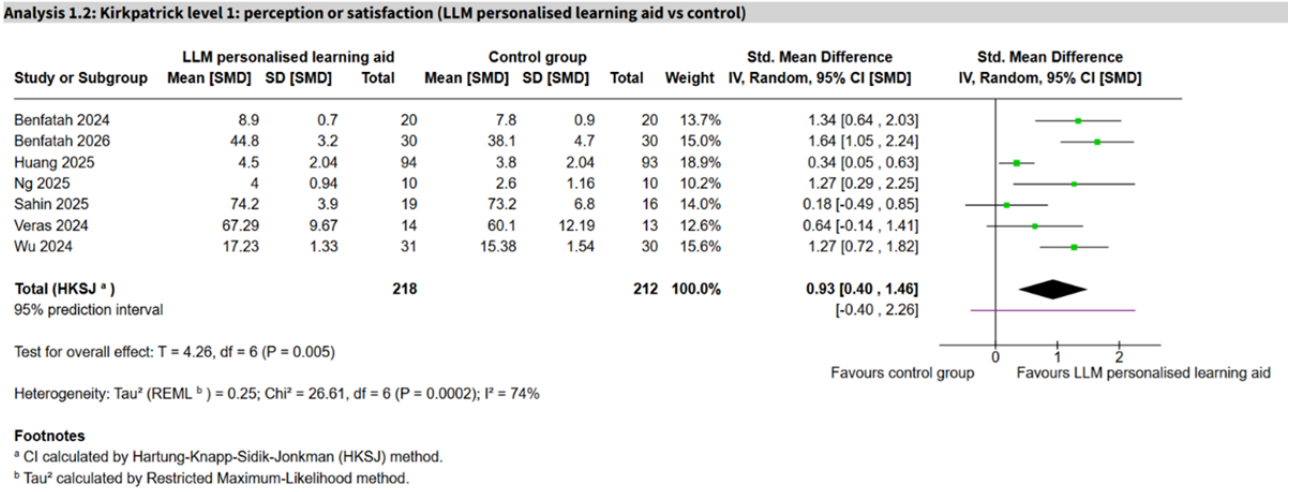
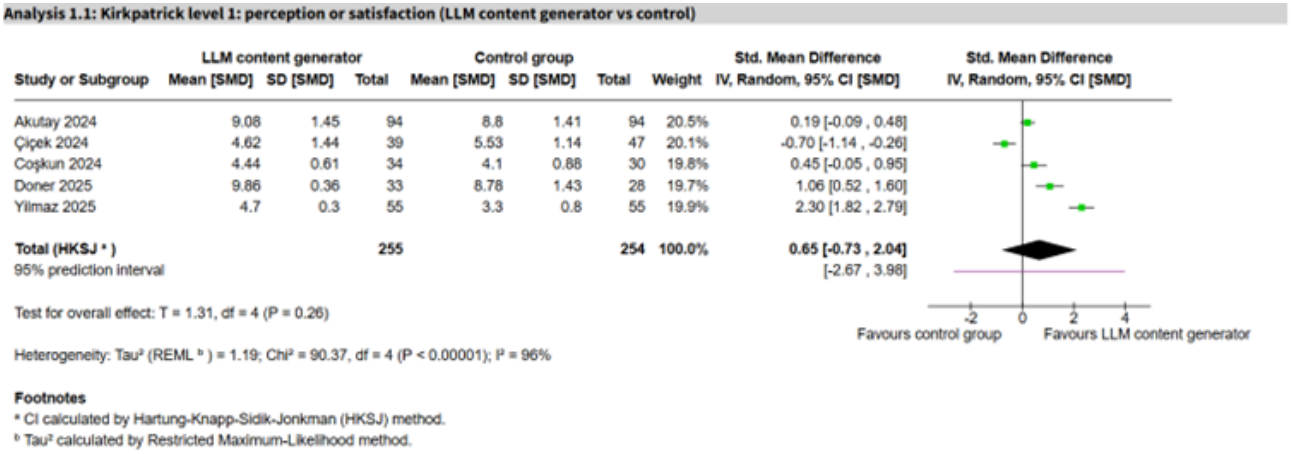


Figure 3. Forest plot for perception or satisfaction (large language model [LLM] content generator vs control) [49,51,61,62,100].



Self-Efficacy or Confidence

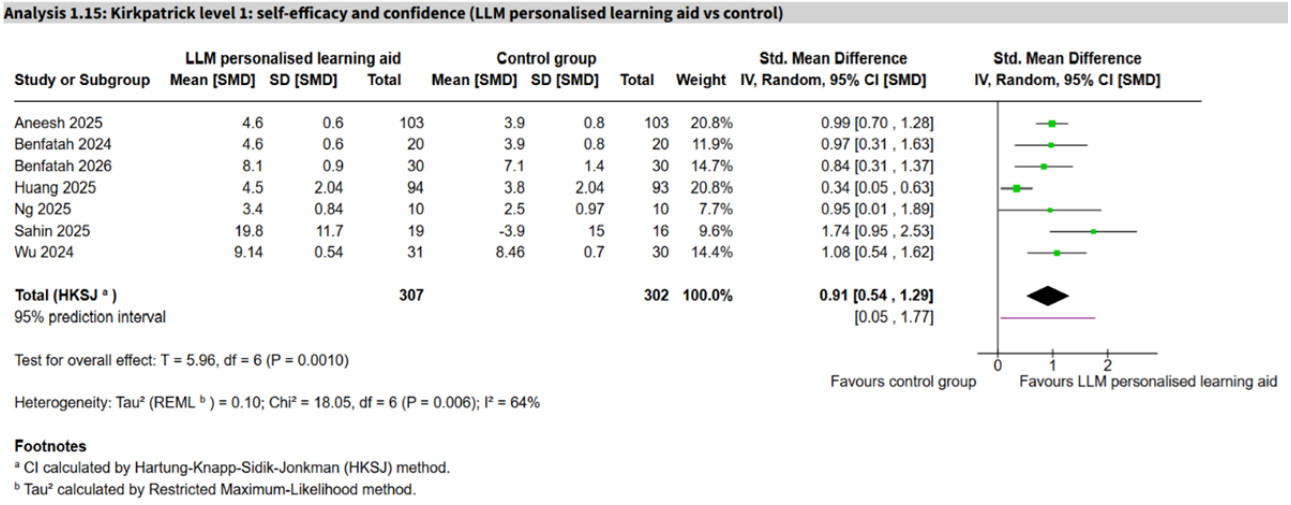
Sixteen studies (1115 participants) assessed self-efficacy or confidence across 9 AI subtype comparisons, and 1 additional study reported a dichotomous outcome (Table 1).

The largest comparison was LLM personalized learning aids, which appeared to show an important positive effect (7 studies, n=609; SMD 0.91, 95% CI 0.54 to 1.29; I^2=64%, 95% PI 0.05 to 1.77; very low certainty; Figure 4 [53,70,80,84,113,115,116]).

All other comparisons were based on 1 or 2 studies with no clear differences between groups. One notable single-study finding was a non-LLM AI-VR virtual doctor that showed reduced self-efficacy compared with control (SMD −0.68, 95% CI −1.18 to −0.17; very low certainty). A separate single study reported the proportion of participants confident in echocardiography views using a non-LLM AI procedure assistant, with no clear difference (RR 1.26, 95% CI 0.45 to 3.50; very low certainty).

XSL•FO
RenderX

Figure 4. Forest plot for self-efficacy or confidence (large language model [LLM] personalized learning aid vs control) [53,70,80,84,113,115,116].

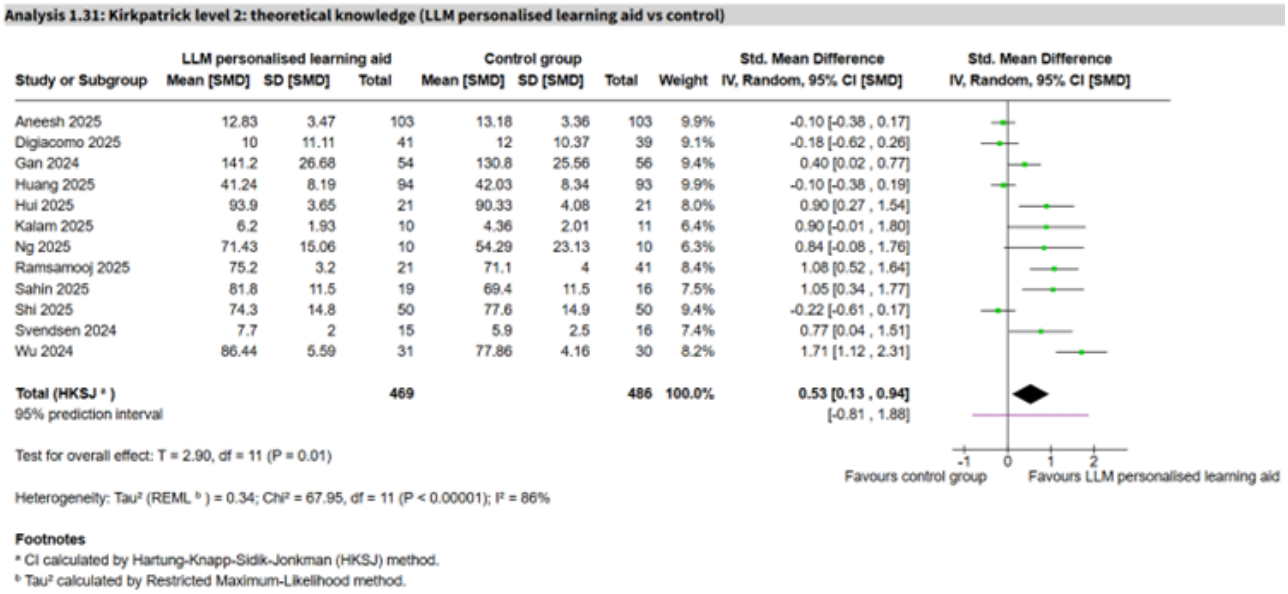


Kirkpatrick Level 2: Theoretical Knowledge

Twenty-four studies (2126 participants) assessed theoretical knowledge across 7 AI subtype comparisons, and 1 additional study reported a dichotomous outcome (Table 1).

The most informative subtype comparison was LLM personalized learning aids, which appeared to show an important positive effect but with substantial heterogeneity (12 studies, n=955; SMD 0.53, 95% CI 0.13 to 0.94; I^2=86%, 95% PI −0.81 to 1.88; very low certainty; Figure 5 [53,63,67,70-72,80,82,84,86,90,113]; publication bias not suggested from funnel plot, Part 8 in Multimedia Appendix 1). LLM content generators (3 studies, n=359; SMD 0.99, 95% CI −1.04 to 3.01; I^2=80%, 95% PI −2.96 to 4.93) and NLP rule-based chatbots (3 studies, n=530; SMD 1.06, 95% CI −2.19 to 4.32; I^2=65%, 95% PI −5.28 to 7.41) showed imprecise results with substantial heterogeneity and very wide PIs, all of very low certainty. Single-study comparisons of LLM gamification, non-LLM AI-moderated adaptive learning, and AI-VR virtual doctor each appeared to favor AI, while a non-LLM AI imaging diagnostic aid comparison (2 studies) was inconclusive (Table 1). One small study reported the proportion achieving grade A or B with a non-LLM AI gamification tool (RR 1.33, 95% CI 1.01 to 1.74; very low certainty).

Figure 5. Forest plot for theoretical knowledge (large language model [LLM] personalized learning aid vs control) [53,63,67,70-72,80,82,84,86,90,113].



Clinical Skills

Twenty-six studies (n=1711) assessed clinical skills across 11 AI subtype comparisons (Table 1). The largest comparisons were LLM personalized learning aids (9 studies, n=609; SMD 0.49, 95% CI 0.00 to 0.97; I^2=83%, 95% PI −0.93 to 1.90; very low certainty; Figure 6 [71,81,84,86,91,113,115,116]) and non-LLM imaging diagnostic aid (4 studies, n=176; SMD 0.43, 95% CI −0.41 to 1.27; I^2=62%, 95% PI −1.13 to 1.99; very low certainty; Figure 7 [56,60,85,99]), which showed no clear differences between groups and substantial heterogeneity. A single study of a NLP rule-based virtual patient showed a large effect (SMD 2.03, 95% CI 1.49 to 2.58; moderate certainty), and a single LLM virtual patient study similarly showed a large effect (SMD 2.53, 95% CI 1.82 to 3.25; low certainty). The rest of the comparisons were inconclusive.

XSL•FO
RenderX

Figure 6. Forest plot for clinical skills (large language model [LLM] personalized learning aid vs control) [71,81,84,86,91,113,115,116].

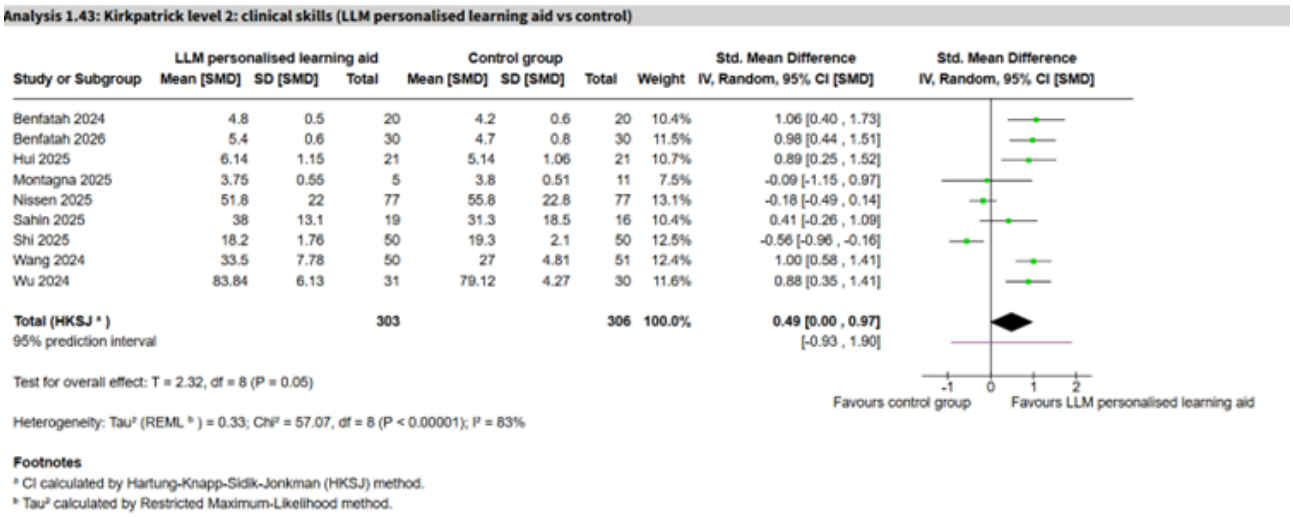
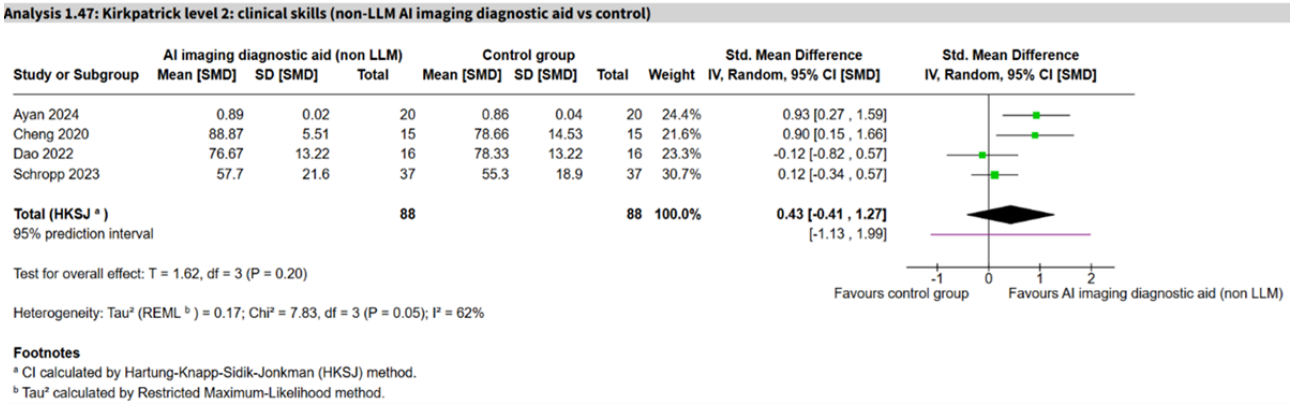


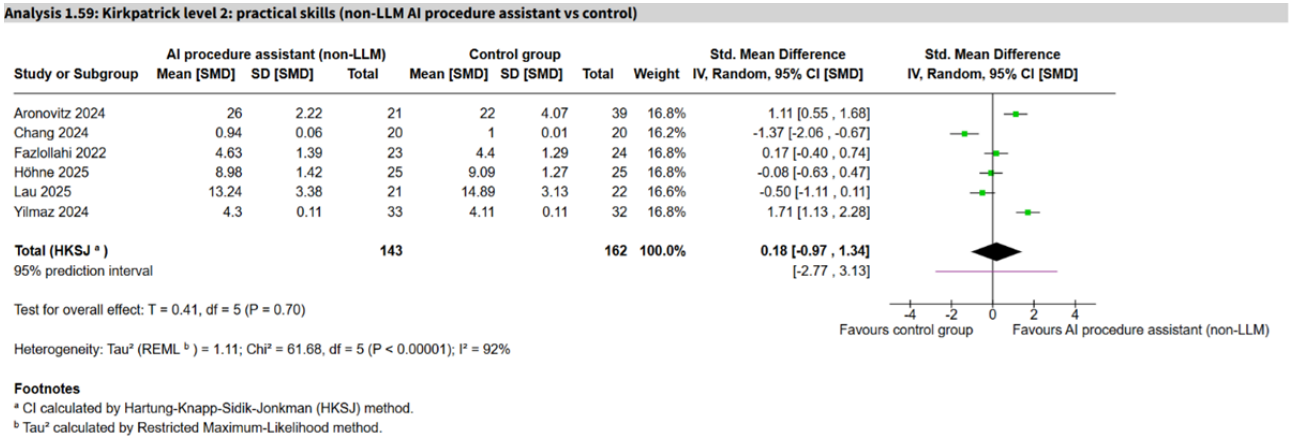
Figure 7. Forest plot for clinical skills (non–large language model [LLM] imaging diagnostic aid vs control) [56,60,85,99].



Practical Skills

Seven studies (492 participants) assessed practical skills (Table 1). One LLM personalized learning aid study (187 participants) appeared to show an important improvement (SMD 0.67, 95% CI 0.37 to 0.96; low certainty), whereas 6 studies of non-LLM AI procedure assistants (=305 participants) showed no clear difference but substantial heterogeneity (SMD 0.18, 95% CI −0.97 to 1.34; I^2=92%, 95% PI −2.77 to 3.13; low certainty; Figure 8 [49,50,55,58,69,105]).

Figure 8. Forest plot for practical skills (non–large language model [LLM] artificial intelligence [AI] procedure assistant vs control) [49,50,55,58,69,105].



Task Efficiency

Six studies (212 participants) measured task efficiency, defined as time taken to complete tasks, across 4 subtype comparisons (Table 1). A single study of a non-LLM AI imaging diagnostic aid showed markedly increased time required to complete task (SMD 2.70, 95% CI 1.82 to 3.58), whereas other comparisons showed no clear differences, all with very low certainty.

Generic or Personal Skills

Five studies (313 participants) assessed generic or personal skills across 5 single-study comparisons (Table 1). A single

XSL•FO
RenderX

LLM-integrated curriculum study appeared to show an important improvement (SMD 0.60, 95% CI 0.19-1.01; moderate certainty), as did a single non-LLM AI communication analysis study (SMD 1.85, 95% CI 0.88 to 2.81; low certainty). A single LLM personalized learning aid study showed a small improvement (SMD 0.45, 95% CI 0.05 to 0.84; low certainty). The remaining comparisons showed no clear differences.

Investigation of Heterogeneity

A detailed report on the results of our heterogeneity exploration is available in Part 7 in Multimedia Appendix 1. In summary, there was a very high degree of heterogeneity across almost all outcomes (I^2: 86% - 94%) with no consistent studies or group of studies identified as the chief contributors. Subgroup analyses (by field and region of study, LLM vs non-LLM, predominant function, and single vs multiple sessions), conducted for all outcomes with sufficient studies (Multimedia Appendix 2), revealed no consistent explanation of heterogeneity, as none of these factors consistently moderated the direction or magnitude of the pooled estimates.

Sensitivity Analysis

Sensitivity analysis was only feasible for the domains of random sequence generation and allocation concealment, and for the analysis of theoretical knowledge (comparing LLM personalized learning aid against control), as this was the only analysis with at least 5 studies in more than 1 risk stratum. In this analysis, after removing 7 studies with high risk of bias in allocation concealment, the pooled estimate of 5 remaining studies changed from showing a significant improvement in knowledge to no significant difference between groups. This might indicate that low-risk studies tended to show more conservative estimates, or reflected reduced statistical power from the smaller number of remaining studies. We could not perform sensitivity analysis to exclude crossover trials because in none of the analyses where these trials were included, there were sufficient studies in each stratum to enable a meaningful analysis.

Certainty of Evidence

Certainty of evidence was low to very low for most outcomes, due mainly to study limitations (high risk of bias across multiple domains), inconsistency (high degree of heterogeneity with wide 95% PIs), and imprecision (wide 95% CIs crossing multiple effect thresholds; for details refer to Part 3 in Multimedia Appendix 1).

Discussion

Principal Findings

This review offers the first comprehensive RCT synthesis of AI in undergraduate health professions education, with granular subcategorization by technology type and educational function across diverse global settings. However, the wide diversity of AI subcategories produced numerous specific analyses, most containing few studies with imprecise, very-low-certainty estimates, leaving the body of evidence far from sufficient to inform educational practice, despite the overall large number of included RCTs.

Based predominantly on very low-certainty evidence, the findings suggest that certain AI intervention types may improve specific educational outcomes, but the pattern of benefit is inconsistent. LLM-based personalized learning aids, with comparatively the largest volume of evidence across outcomes, show a possible positive effect in perception, confidence, and knowledge, but the PIs of these outcomes were wide and mostly crossing the null, indicating substantial uncertainty about the likely effect in future educational contexts.

Across all other AI subcategories, point estimates were generally positive but CIs and PIs were wide and typically crossed the null, precluding firm conclusions. No studies assessed Kirkpatrick Level 3 (real-world practice change) or Level 4 (health outcomes), where evidence would be most compelling for educational policy.

The subcategorized analyses also clarify which comparisons remain informationally sparse. Several AI subtypes, particularly multicomponent interventions and novel tool combinations, contributed only single-study estimates, highlighting where future research is most needed.

Comparison With Prior Work

Our findings build on and extend previous syntheses in this field. Before this work, approximately 50 reviews examined AI in health education (Part 9 in Multimedia Appendix 1), but the majority are scoping or narrative reviews, and most express cautious optimism without quantitative evidence from RCTs. Two meta-analyses evaluating LLMs in national licensing examinations reported a wide performance range and concluded that further model development and evaluation are needed before LLMs can be recommended as primary teaching tools [117,118]. A scoping review of AI across medical education similarly concluded that AI applications show promise but require rigorous evaluation [10].

Most directly comparable to our work is a recent meta-analysis of 11 RCTs (786 participants) examining generative AI in medical education, which reported no clear difference in overall knowledge acquisition but improved practical skills and satisfaction [22]. Our review extends this in several important respects: a more recent search (January 2026 vs January 2025), broader scope (all AI applications and health disciplines, not only generative AI in medicine), larger evidence base (60 RCTs, 4506 participants in meta-analysis), systematic subcategorization of AI intervention types, outcome classification using Kirkpatrick framework, risk of bias assessment using ROBUST-RCT, and, crucially, Grading of Recommendations, Assessment, Development, and Evaluation certainty-of-evidence ratings, absent from all previously published reviews of this topic. Without certainty ratings, positive point estimates risk being interpreted as practice-ready evidence; yet, in almost every comparison, certainty is low or very low, and PIs indicate that effects may not reliably reproduce across settings.

Limitations

We acknowledge several limitations of the evidence gathered and the review process. In terms of the evidence gathered, despite comprehensive and up-to-date searches, several limitations require acknowledgment. First, even after

XSL•FO
RenderX

subcategorization, substantial residual heterogeneity remained within several comparisons (I^2: 64% - 98%), which was not adequately explained by the study characteristics reported. This heterogeneity likely reflects the inherently context-dependent nature of educational interventions, where learner characteristics, instructor expertise, institutional culture, implementation fidelity, and assessment approaches interact in ways that are rarely reported in sufficient detail [119,120]. Rather than invalidating the findings, this suggests the effectiveness of AI tools may be context-sensitive, shaped by local factors rarely reported in sufficient detail. The consistent direction of effect (most studies favoring AI) provides a preliminary signal of benefit, but the magnitude of any effect in a given setting remains uncertain.

Second, granular subcategorization, while educationally more informative, resulted in many comparisons based on only 1 or 2 studies, which are too imprecise to provide actionable guidance for practice.

Next, we included 7 crossover trials [47,48,53,66,74,80,81], which may have additional concerns in their study design that we have not adequately addressed, because the newly established ROBUST-RCT tool has not been expanded to cover crossover design [29]. Additionally, none of these studies provided data separately for the first period, as desired. However, given that most included studies were already rated at high risk of bias for allocation concealment and blinding and the evidence certainty was very low for most outcomes, this limitation is unlikely to substantially alter the overall certainty assessments, but cannot be dismissed.

In terms of the review process, despite screening nearly 40,000 records and shortlisting 177 potentially eligible records, we may have missed studies evaluating AI-enabled tools (such as robotic or VR-based systems) not described as AI interventions in their reports. Our risk of bias assessment and decisions on which outcomes to include for meta-analysis, although transparent, involved some subjective judgment and might have been influenced by personal biases. Pooling continuous outcomes using SMD across heterogeneous instruments, particularly for subjective outcomes such as satisfaction and self-efficacy, is another limitation of this review; however, we incorporated an SMD of 0.5 as a default minimally important difference and PIs to improve interpretability in educational practice. Finally, our subcategorization of specific AI applications was made for the purpose of the current review, and our decision to assign one study to one subcategory or another, although largely reliant on the information provided by the study, involved subjective judgment. However, we hope this taxonomy provides a working framework for future syntheses as the evidence base grows.

Conclusions and Implications

Early evidence from RCTs suggests that the effects of AI applications are neither uniform across AI types nor consistent across settings. Although certain AI intervention types used in undergraduate health professions education appeared to improve some educational outcomes, particularly satisfaction and theoretical knowledge with LLM-based personalized learning aids, the findings are highly heterogeneous with low- to very low–certainty evidence and are far from sufficient in informing educational practice. No studies have yet assessed real-world practice change or health-related outcomes, which are the levels of evidence most relevant to educational policy. We recommend that AI applications in undergraduate health education continue to be used on a trial basis. Future RCTs should be adequately powered and use robust allocation concealment, such as central randomization or online systems with real-time allocation, with transparent documentation. Standardized outcome measures should be used to minimize performance and detection bias, with clear reporting of implementation details to support meaningful subgroup and heterogeneity analyses. Studies should also begin to assess longer-term Kirkpatrick Level 3 and 4 outcomes, which reflect real-world behavior change and patient health outcomes, as these would particularly strengthen the evidence base for educational guideline adoption.

Acknowledgments

We used Claude Sonnet 4.5 (Anthropic), a large language model [27], for preliminary data extraction including study characteristics in terms of population, intervention, comparison, outcomes, and methodologies, as well as final proofreading of the manuscript draft, and used Elicit AI (Ought) [121] on a trial basis to identify relevant papers for our background and discussion. Examples of prompts for each purpose mentioned, as well as human verification and the decision process are available in Part 10 in Multimedia Appendix 1. However, the human authors performed final verification, finalized the extracted data, selected the relevant citations to be included in the write-up, and wrote this manuscript.

Funding

No external financial support or grants were received for this work.

Data Availability

The datasets used and analyzed during this study, including the complete data extraction forms, risk-of-bias assessments, forest plots, and RevMan analysis files, are available from the corresponding author on reasonable request. The PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) 2020 checklist and detailed search strategies are provided in Checklist 1 and Multimedia Appendix 1, respectively.

XSL•FO
RenderX

Authors' Contributions

NML conceptualized the review, drafted the protocol including methodology, performed the search, extracted data with assistance from Claude Sonnet, performed risk-of-bias assessment, curated the data for meta-analysis and certainty-of-evidence rating, drafted the initial manuscript, and critically reviewed and revised the manuscript. QCO performed data extraction, checked data accuracy, performed risk-of-bias assessment and certainty-of-evidence rating, and critically reviewed and revised the manuscript. MTW and PB drafted part of the first draft, critically reviewed and revised the manuscript. YSL checked data accuracy, critically reviewed and revised the manuscript. PT critically reviewed and revised the manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Compilation of parts 1-11.
[DOCX File, 1084 KB - mededu_v12i1e88933_app1.docx]

Multimedia Appendix 2
Full analyses with forest plots.
[PDF File, 323 KB - mededu_v12i1e88933_app2.pdf]

Checklist 1
PRISMA 2020 checklist.
[DOCX File, 30 KB - mededu_v12i1e88933_app3.docx]

Checklist
PRISMA-S checklist.
[DOCX File, 17 KB - mededu_v12i1e88933_app4.docx]

References

1. Healthcare education market (2025-2033) size, share & trends analysis report by providers (continuing medical education providers, educational platforms, learning management systems, universities & academic centers), by application, by delivery mode, by end use, by region, and segment forecasts. : Grand View Research; 2024 URL: https://www.grandviewresearch.com/industry-analysis/healthcare-education-market-report [accessed 2025-08-18]
2. Alotaibi N, Wilson CB, Traynor M. Enhancing digital readiness and capability in healthcare: a systematic review of interventions, barriers, and facilitators. BMC Health Serv Res 2025 Apr 4;25(1):500. [doi: 10.1186/s12913-025-12663-3] [Medline: 40186200]
3. Grainger R, Liu Q, Gladman T. Learning technology in health professions education: realising an (un)imagined future. Med Educ 2024 Jan;58(1):36-46. [doi: 10.1111/medu.15185] [Medline: 37555302]
4. Slavin S, D'Eon MF. Overcrowded curriculum is an impediment to change (Part A). Can Med Educ J 2021 Sep;12(4):1-6. [doi: 10.36834/cmej.73532] [Medline: 34567300]
5. Talwar S, Dhir A, Islam N, Kaur P, Almusharraf A. Resistance of multiple stakeholders to e-health innovations: integration of fundamental insights and guiding research paths. J Bus Res 2023 Nov;166:114135. [doi: 10.1016/j.jbusres.2023.114135]
6. Caverzagie KJ, Nousiainen MT, Ferguson PC, et al. Overarching challenges to the implementation of competency-based medical education. Med Teach 2017 Jun;39(6):588-593. [doi: 10.1080/0142159X.2017.1315075] [Medline: 28598747]
7. Kumaravel B, Hearn JH, Jahangiri L, Pollard R, Stocker CJ, Nunan D. A systematic review and taxonomy of tools for evaluating evidence-based medicine teaching in medical education. Syst Rev 2020 Apr 24;9(1):91. [doi: 10.1186/s13643-020-01311-y] [Medline: 32331530]
8. McCluskey A, Lovarini M. Providing education on evidence-based practice improved knowledge but did not change behaviour: a before and after study. BMC Med Educ 2005 Dec 19;5:40. [doi: 10.1186/1472-6920-5-40] [Medline: 16364181]
9. Salehi S, Ballen CJ, Bolander Laksov K, et al. Global perspectives of the impact of the COVID-19 pandemic on learning science in higher education. PLoS One 2023;18(12):e0294821. [doi: 10.1371/journal.pone.0294821] [Medline: 38060473]
10. Gordon M, Daniel M, Ajiboye A, et al. A scoping review of artificial intelligence in medical education: BEME Guide No. 84. Med Teach 2024 Apr;46(4):446-470. [doi: 10.1080/0142159X.2024.2314198] [Medline: 38423127]
11. McCarthy J, Minsky ML, Rochester N, Shannon CE. A proposal for the Dartmouth summer research project on artificial intelligence: August 31, 1955. AI Mag 1955 Aug 31;27(4):12. [doi: 10.1609/aimag.v27i4.1904]
12. Ali O, Abdelbaki W, Shrestha A, Elbasi E, Alryalat MAA, Dwivedi YK. A systematic literature review of artificial intelligence in the healthcare sector: benefits, challenges, methodologies, and functionalities. J Innov Knowl 2023 Jan;8(1):100333. [doi: 10.1016/j.jik.2023.100333]

XSL•FO
RenderX

13. Introducing ChatGPT. OpenAI. 2022. URL: https://openai.com/blog/chatgpt [accessed 2025-08-18]
14. Mukherjee M, Le J, Chow YW. Generative AI-enhanced intelligent tutoring system for graduate cybersecurity programs. Future Internet 2025;17(4):154. [doi: 10.3390/fi17040154]
15. Caudell TP, Summers KL, Holten J 4th, et al. Virtual patient simulator for distributed collaborative medical education. Anat Rec B New Anat 2003 Jan;270(1):23-29. [doi: 10.1002/ar.b.10007] [Medline: 12526063]
16. Mirchi N, Bissonnette V, Yilmaz R, Ledwos N, Winkler-Schwartz A, Del Maestro RF. The Virtual Operative Assistant: an explainable artificial intelligence tool for simulation-based training in surgery and medicine. PLoS ONE 2020;15(2):e0229596. [doi: 10.1371/journal.pone.0229596] [Medline: 32106247]
17. Mastour H, Dehghani T, Moradi E, Eslami S. Explainable artificial intelligence for predicting medical students' performance in comprehensive assessments. Sci Rep 2025 Jul 3;15(1):23752. [doi: 10.1038/s41598-025-07460-1] [Medline: 40610576]
18. Vrdoljak J, Boban Z, Vilović M, Kumrić M, Božić J. A review of large language models in medical education, clinical decision support, and healthcare administration. Healthcare (Basel) 2025 Mar 10;13(6):603. [doi: 10.3390/healthcare13060603] [Medline: 40150453]
19. Waldock WJ, Zhang J, Guni A, Nabeel A, Darzi A, Ashrafian H. The accuracy and capability of artificial intelligence solutions in health care examinations and certificates: systematic review and meta-analysis. J Med Internet Res 2024 Nov 5;26:e56532. [doi: 10.2196/56532] [Medline: 39499913]
20. Shan G, Chen X, Wang C, et al. Comparing diagnostic accuracy of clinical professionals and large language models: systematic review and meta-analysis. JMIR Med Inform 2025 Apr 25;13:e64963. [doi: 10.2196/64963] [Medline: 40279517]
21. Lucas HC, Upperman JS, Robinson JR. A systematic review of large language models and their implications in medical education. Med Educ 2024 Nov;58(11):1276-1285. [doi: 10.1111/medu.15402] [Medline: 38639098]
22. Li J, Yin K, Wang Y, Jiang X, Chen D. Effectiveness of generative artificial intelligence-based teaching versus traditional teaching methods in medical education: a meta-analysis of randomized controlled trials. BMC Med Educ 2025;25(1):1175. [doi: 10.1186/s12909-025-07750-2]
23. Samarakoon L, Fernando T, Rodrigo C, Rajapakse S. Learning styles and approaches to learning among medical undergraduates and postgraduates. BMC Med Educ 2013 Dec;13(1). [doi: 10.1186/1472-6920-13-42]
24. Sharma AD. Virtual reality: robotic improved surgical precision using AI techniques. In: Goyal LS, Rehman A, editors. Artificial Intelligence and Internet of Things: CRC Press; 2021. URL: https://www.taylorfrancis.com/chapters/edit/10.1201/9781003097204-16/virtual-reality-aditi-sharma-aman-dureja [accessed 2026-04-30]
25. Alliger GM, Janak EA. Kirkpatrick's levels of training criteria: thirty years later. Pers Psychol 1989 Jun;42(2):331-342. [doi: 10.1111/j.1744-6570.1989.tb00661.x]
26. Rethlefsen ML, Kirtley S, Waffenschmidt S, et al. PRISMA-S: an extension to the PRISMA Statement for Reporting Literature Searches in Systematic Reviews. Syst Rev 2021 Jan 26;10(1):39. [doi: 10.1186/s13643-020-01542-z] [Medline: 33499930]
27. Homepage. Claude. 2025. URL: https://claude.ai [accessed 2025-08-21]
28. Johnston S, Coyer FM, Nash R. Kirkpatrick's evaluation of simulation and debriefing in health care education: a systematic review. J Nurs Educ 2018 Jul 1;57(7):393-398. [doi: 10.3928/01484834-20180618-03] [Medline: 29958308]
29. Wang Y, Keitz S, Briel M, et al. Development of ROBUST-RCT: risk of bias instrument for use in systematic reviews-for randomised controlled trials. BMJ 2025 Mar 25;388:e081199. [doi: 10.1136/bmj-2024-081199] [Medline: 40132800]
30. Cohen J. Statistical Power Analysis for the Behavioral Sciences, 2nd edition: Routledge; 1988. [doi: 10.4324/9780203771587]
31. Kraft MA. Interpreting effect sizes of education interventions. Educ Res 2020 May;49(4):241-253. [doi: 10.3102/0013189X20912798]
32. Norman GR, Sloan JA, Wyrwich KW. Interpretation of changes in health-related quality of life: the remarkable universality of half a standard deviation. Med Care 2003 May;41(5):582-592. [doi: 10.1097/01.MLR.0000062554.74615.4C] [Medline: 12719681]
33. Hozo SP, Djulbegovic B, Hozo I. Estimating the mean and variance from the median, range, and the size of a sample. BMC Med Res Methodol 2005 Apr 20;5(1):13. [doi: 10.1186/1471-2288-5-13] [Medline: 15840177]
34. Higgins JP. Including variants on randomized trials. In: Higgins JP, Chandler J, Cumpston M, editors. Cochrane Handbook for Systematic Reviews of Interventions Version 65: Cochrane; 2024. URL: https://www.cochrane.org/authors/handbooks-and-manuals/handbook/current/chapter-23 [accessed 2026-04-30]
35. Altman DG, McKenzie JE, Veroniki AA. Analysing data and undertaking meta-analyses. In: Higgins JP, Chandler J, Cumpston M, editors. Cochrane Handbook for Systematic Reviews of Interventions: Cochrane. URL: https://www.cochrane.org/authors/handbooks-and-manuals/handbook/current/chapter-10 [accessed 2026-04-30]
36. Langan D, Higgins JPT, Jackson D, et al. A comparison of heterogeneity variance estimators in simulated random-effects meta-analyses. Res Synth Methods 2019 Mar;10(1):83-98. [doi: 10.1002/jrsm.1316] [Medline: 30067315]
37. IntHout J, Ioannidis JPA, Borm GF. The Hartung-Knapp-Sidik-Jonkman method for random effects meta-analysis is straightforward and considerably outperforms the standard DerSimonian-Laird method. BMC Med Res Methodol 2014 Feb 18;14(1):25. [doi: 10.1186/1471-2288-14-25] [Medline: 24548571]
38. Higgins JPT, Thomas J, Chandler J, Cumpston M, Li T, Page MJ, et al. Cochrane Handbook for Systematic Reviews of Interventions: The Cochrane Collaboration; 2022. URL: https://training.cochrane.org/handbook [accessed 2026-04-30]

XSL•FO
RenderX

39. IntHout J, Ioannidis JPA, Rovers MM, Goeman JJ. Plea for routinely presenting prediction intervals in meta-analysis. BMJ Open 2016 Jul 12;6(7):e010247. [doi: 10.1136/bmjopen-2015-010247] [Medline: 27406637]
40. Sterne JA, Egger M, Moher D, Boutron I. Addressing reporting biases. In: Higgins JP, Chandler J, Cumpston MS, editors. Cochrane Handbook for Systematic Reviews of Interventions: The Cochrane Collaboration; 2017. URL: https://onlinelibrary.wiley.com/doi/10.1002/9780470712184.ch10 [accessed 2026-04-30]
41. Sterne JAC, Sutton AJ, Ioannidis JPA, et al. Recommendations for examining and interpreting funnel plot asymmetry in meta-analyses of randomised controlled trials. BMJ 2011 Jul 22;343(jul22 1):d4002. [doi: 10.1136/bmj.d4002] [Medline: 21784880]
42. Page MJ, Sterne JAC, Boutron I, et al. ROB-ME: a tool for assessing risk of bias due to missing evidence in systematic reviews with meta-analysis. BMJ 2023 Nov 20;383:e076754. [doi: 10.1136/bmj-2023-076754] [Medline: 37984978]
43. Schünemann H, Brożek J, Guyatt G, Oxman A, editors. Handbook for Grading the Quality of Evidence and the Strength of Recommendations Using the GRADE Approach: Group GW; 2013. URL: https://gdt.gradepro.org/app/handbook/handbook.html [accessed 2026-04-30]
44. Page MJ, McKenzie JE, Bossuyt PM, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. Syst Rev 2021 Mar 29;10(1):89. [doi: 10.1186/s13643-021-01626-4] [Medline: 33781348]
45. Page MJ, McKenzie JE, Bossuyt PM, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ 2021 Mar 29;372:n71. [doi: 10.1136/bmj.n71] [Medline: 33782057]
46. Liaw SY, Tan JZ, Bin Rusli KD, et al. Artificial intelligence versus human-controlled doctor in virtual reality simulation for sepsis team training: randomized controlled study. J Med Internet Res 2023 Jul 26;25:e47748. [doi: 10.2196/47748] [Medline: 37494112]
47. Ting PW, Wolffsohn JS. Artificial intelligence-driven patient history and symptoms combined with slit-lamp eye simulation for enhancing the clinical training of students. Clin Exp Optom 2026 Apr;109(3):434-443. [doi: 10.1080/08164622.2025.2544809] [Medline: 40975895]
48. Veras M, Dyer JO, Shannon H, et al. A mixed methods crossover randomized controlled trial exploring the experiences, perceptions, and usability of artificial intelligence (ChatGPT) in health sciences education. Digit Health 2024;10:20552076241298485. [doi: 10.1177/20552076241298485] [Medline: 39917060]
49. Yilmaz R, Bakhaidar M, Alsayegh A, et al. Real-time multifaceted artificial intelligence vs in-person instruction in teaching surgical technical skills: a randomized controlled trial. Sci Rep 2024 Jul 2;14(1):15130. [doi: 10.1038/s41598-024-65716-8] [Medline: 38956112]
50. Fazlollahi AM, Bakhaidar M, Alsayegh A, et al. Effect of artificial intelligence tutoring vs expert instruction on learning simulated surgical skills among medical students. JAMA Netw Open 2022 Feb 1;5(2):e2149008. [doi: 10.1001/jamanetworkopen.2021.49008] [Medline: 35191972]
51. Akutay S, Yüceler Kaçmaz H, Kahraman H. The effect of artificial intelligence supported case analysis on nursing students' case management performance and satisfaction: a randomized controlled trial. Nurse Educ Pract 2024 Oct;80:104142. [doi: 10.1016/j.nepr.2024.104142] [Medline: 39299058]
52. Ali M, Rehman S, Cheema E. Impact of artificial intelligence on the academic performance and test anxiety of pharmacy students in objective structured clinical examination: a randomized controlled trial. Int J Clin Pharm 2025 Aug;47(4):1034-1041. [doi: 10.1007/s11096-025-01876-5] [Medline: 39903358]
53. Aneesh KV, Mohanan S, Jose S, Sajla K, Indulekha C, Sukumaran S, et al. Effectiveness of generative AI versus traditional resources for self-directed learning in physiology among MBBS students: a comparative interventional study. Int J Med Public Health 2025;15(3). [doi: 10.70034/ijmedph.2025.3.385] [Medline: 40830476]
54. Arkan B, Dallı Ö, Varol B. The impact of ChatGPT training in the nursing process on nursing students' problem-solving skills, attitudes towards artificial intelligence, competency, and satisfaction levels: single-blind randomized controlled study. Nurse Educ Today 2025 Sep;152:106765. [doi: 10.1016/j.nedt.2025.106765] [Medline: 40334550]
55. Aronovitz N, Hazan I, Jedwab R, et al. The effect of real-time EF automatic tool on cardiac ultrasound performance among medical students. PLoS ONE 2024;19(3):e0299461. [doi: 10.1371/journal.pone.0299461] [Medline: 38547257]
56. Ayan E, Bayraktar Y, Çelik Ç, Ayhan B. Dental student application of artificial intelligence technology in detecting proximal caries lesions. J Dent Educ 2024 Apr;88(4):490-500. [doi: 10.1002/jdd.13437] [Medline: 38200405]
57. Brügge E, Ricchizzi S, Arenbeck M, et al. Large language models improve clinical decision making of medical students through patient simulation and structured feedback: a randomized controlled trial. BMC Med Educ 2024 Nov 28;24(1):1391. [doi: 10.1186/s12909-024-06399-7] [Medline: 39609823]
58. Chang J, Bliss L, Angelov N, Glick A. Artificial intelligence-assisted full-mouth radiograph mounting in dental education. J Dent Educ 2024 Jul;88(7):933-939. [doi: 10.1002/jdd.13524] [Medline: 38545660]
59. Chen Y. Evaluation of the impact of AI-driven personalized learning platform on medical students' learning performance. Front Med 2025;12(1610012). [doi: 10.3389/fmed.2025.1610012]
60. Cheng CT, Chen CC, Fu CY, et al. Artificial intelligence-based education assists medical students' interpretation of hip fracture. Insights Imaging 2020 Nov 23;11(1):119. [doi: 10.1186/s13244-020-00932-0] [Medline: 33226480]
61. Çiçek FE, Ülker M, Özer M, Kıyak YS. ChatGPT versus expert feedback on clinical reasoning questions and their effect on learning: a randomized controlled trial. Postgrad Med J 2025 Apr 22;101(1195):458-463. [doi: 10.1093/postmj/qgae170]

XSL•FO
RenderX

62. Coşkun Ö, Kıyak YS, Budakoğlu I. ChatGPT to generate clinical vignettes for teaching and multiple-choice questions for assessment: a randomized controlled experiment. Med Teach 2025 Feb;47(2):268-274. [doi: 10.1080/0142159X.2024.2327477] [Medline: 38478902]
63. Digiacomo A, Orsini A, Cicchetti R, et al. Chatgpt vs traditional pedagogy: a comparative study in urological learning. World J Urol 2025 May 8;43(1):286. [doi: 10.1007/s00345-025-05654-w] [Medline: 40338279]
64. Fazlollahi A, Bakhaidar M, Alsayegh A, et al. 510 artificial intelligence tutoring compared with expert instruction in neurosurgical simulation training: a randomized controlled trial. Clin Neurosurg 2022;68(Supplement_1):128-129. [doi: 10.1227/NEU.0000000000001880_510]
65. Ferrer-Peña R, Di-Bonaventura S, Pérez-González A, Lerín-Calvo A. Feasibility of a randomized controlled trial of large AI-based linguistic models for clinical reasoning training of physical therapy students: pilot randomized parallel-group study. JMIR Form Res 2025 Jul 23;9:e66126. [doi: 10.2196/66126] [Medline: 40702721]
66. Fung TCJ, Chan SL, Lam CFM, et al. Effects of generative artificial intelligence (GenAI) patient simulation on perceived clinical competency among global nursing undergraduates: a cross-over randomised controlled trial. BMC Nurs 2025 Jul 17;24(1):934. [doi: 10.1186/s12912-025-03492-0] [Medline: 40676632]
67. Gan W, Ouyang J, Li H, et al. Integrating ChatGPT in orthopedic education for medical undergraduates: randomized controlled trial. J Med Internet Res 2024 Aug 20;26:e57037. [doi: 10.2196/57037] [Medline: 39163598]
68. Han JW, Park J, Lee H. Development and effects of a chatbot education program for self-directed learning in nursing students. BMC Med Educ 2025 Jun 2;25(1):825. [doi: 10.1186/s12909-025-07316-2] [Medline: 40457281]
69. Höhne E, Bauer E, Bauer C, et al. A comparative bicentric study on ultrasound education for students: app- and AI-supported learning versus traditional hands-on instruction (AI-teach study). Acad Radiol 2025 Aug;32(8):4930-4938. [doi: 10.1016/j.acra.2025.04.024] [Medline: 40300993]
70. Huang S, Wen C, Bai X, et al. Exploring the application capability of ChatGPT as an instructor in skills education for dental medical students: randomized controlled trial. J Med Internet Res 2025;27:e68538. [doi: 10.2196/68538]
71. Hui Z, Zewu Z, Jiao H, Yu C. Application of ChatGPT-assisted problem-based learning teaching method in clinical medical education. BMC Med Educ 2025;25(1):CN. [doi: 10.1186/s12909-024-06321-1]
72. Kalam KA, Masoud FD, Muntaser A, Ranga R, Geng X, Goyal M. ChatGPT as a learning tool for medical students: results from a randomized controlled trial. Cureus 2025 Jun;17(6):e85767. [doi: 10.7759/cureus.85767] [Medline: 40656250]
73. Kestel S, Çalık A, Kuş M. The effect of chatbot-supported instruction on nursing students' history-taking questioning skills and stress level: a randomized controlled study. J Prof Nurs 2025;60:93-100. [doi: 10.1016/j.profnurs.2025.07.004] [Medline: 40915772]
74. Li J, Ouyang J, Liu J, et al. Artificial intelligence-based online platform assists blood cell morphology learning: a mixed-methods sequential explanatory designed research. Med Teach 2023 Jun;45(6):596-603. [doi: 10.1080/0142159X.2023.2190483] [Medline: 36971649]
75. Lyu X, Dong L, Fan Z, et al. Artificial intelligence-based graded training of pulmonary nodules for junior radiology residents and medical imaging students. BMC Med Educ 2024;24(1):CN. [doi: 10.1186/s12909-024-05723-5]
76. Mahrous A, Botsko DL, Elgreatly A, Tsujimoto A, Qian F, Schneider GB. The use of artificial intelligence and game-based learning in removable partial denture design: a comparative study. J Dent Educ 2023 Aug;87(8):1188-1199. [doi: 10.1002/jdd.13225] [Medline: 37186466]
77. McCarrick CA, McEntee PD, Boland PA, et al. A randomized controlled trial of a deep language learning model-based simulation tool for undergraduate medical students in surgery. J Surg Educ 2025 Sep;82(9):103629. [doi: 10.1016/j.jsurg.2025.103629] [Medline: 40729832]
78. Molu B. Improving nursing students' learning outcomes in neonatal resuscitation: a quasi-experimental study comparing ai-assisted care plan learning with traditional instruction. J Eval Clin Pract 2025 Feb;31(1):e14286. [doi: 10.1111/jep.14286] [Medline: 39733257]
79. Montagna M, Chiabrando F, De Lorenzo R, Rovere Querini P, Medical Students. Impact of clinical decision support systems on medical students' case-solving performance: comparison study with a focus group. JMIR Med Educ 2025 Mar 18;11:e55709. [doi: 10.2196/55709] [Medline: 40101183]
80. Ng ISH, Siu A, Han CSJ, et al. Evaluating a custom chatbot in undergraduate medical education: randomised crossover mixed-methods evaluation of performance, utility, and perceptions. Behav Sci (Basel) 2025 Sep 19;15(9):1284. [doi: 10.3390/bs15091284] [Medline: 41009314]
81. Nissen L, Rother JF, Heinemann M, Reimer LM, Jonas S, Raupach T. A randomised cross-over trial assessing the impact of AI-generated individual feedback on written online assignments for medical students. Med Teach 2025 Sep;47(9):1544-1550. [doi: 10.1080/0142159X.2025.2451870] [Medline: 39831699]
82. Ramsamooj A, Ibrahim SM, Gerriets VA, Cusick JK, Ramsamooj R. Artificial intelligence versus traditional learning in a medical school setting. Cureus 2025;17(6):CN. [doi: 10.7759/cureus.85262]
83. Saatçi G, Korkut S, Ünsal A. The effect of the use of artificial intelligence in the preparation of patient education materials by nursing students on the understandability, actionability and quality of the material: a randomized controlled trial. Nurse Educ Pract 2024 Nov;81:104186. [doi: 10.1016/j.nepr.2024.104186] [Medline: 39520840]

XSL•FO
RenderX

84. Sahin GE, Bayram GA, Sierra AS, et al. Effects of artificial intelligence based physiotherapy educational approach in developing clinical reasoning skills: a randomized controlled trial. BMC Med Educ 2025 Oct 9;25(1):1378. [doi: 10.1186/s12909-025-07926-w] [Medline: 41068907]
85. Schropp L, Sørensen APS, Devlin H, Matzen LH. Use of artificial intelligence software in dental education: a study on assisted proximal caries assessment in bitewing radiographs. Eur J Dent Educ 2024 May;28(2):490-496. [doi: 10.1111/eje.12973] [Medline: 37961027]
86. Kejingyun S, Mingjun R. Randomized controlled study on the impact of problem-based learning combined with large language models on critical thinking skills in nursing students. Nurse Educ 2025;50(4):216-220. [doi: 10.1097/NNE.0000000000001879] [Medline: 40261697]
87. Shin H, De Gagne JC, Kim SS, Hong M. The impact of artificial intelligence-assisted learning on nursing students' ethical decision-making and clinical reasoning in pediatric care. Comput Inform Nurs 2024 Oct 1;42(10):704-711. [doi: 10.1097/CIN.0000000000001177] [Medline: 39152099]
88. Song D, Zhang P, Zhu Y, et al. Effects of generative artificial intelligence on higher-order thinking skills and artificial intelligence literacy in nursing undergraduates: a quasi-experimental study. Nurse Educ Pract 2025 Oct;88:104549. [doi: 10.1016/j.nepr.2025.104549] [Medline: 40915080]
89. Staples G, Webster C, Baraza W. SP3.03 artificially intelligent simulated patients in undergraduate surgical teaching: a single blinded randomized controlled study. Br J Surg 2025 Aug 28;112(Supplement_13):CN. [doi: 10.1093/bjs/znaf166.037]
90. Svendsen K, Askar M, Umer D, Halvorsen KH. Short-term learning effect of ChatGPT on pharmacy students' learning. Explor Res Clin Soc Pharm 2024 Sep;15:100478. [doi: 10.1016/j.rcsop.2024.100478] [Medline: 39139501]
91. Shalong W, Yi Z, Bin Z, et al. Enhancing self-directed learning with custom GPT AI facilitation among medical students: a randomized controlled trial. Med Teach 2025 Jul;47(7):1126-1133. [doi: 10.1080/0142159X.2024.2413023] [Medline: 39425996]
92. Wang YF, Hsu MH, Yue-Feng Wang M. Gamified mobile learning: an escape room chatbot to enhance medical terminology acquisition. Health Educ J 2025 Mar;84(2):174-188. [doi: 10.1177/00178969241313442]
93. Wang Z, Fan TT, Li ML, Zhu NJ, Wang XC. Feasibility study of using GPT for history-taking training in medical education: a randomized clinical trial. BMC Med Educ 2025;25(1):CN. [doi: 10.1186/s12909-025-07614-9]
94. Yeo CT, Ungi T, U-Thainual P, Lasso A, McGraw RC, Fichtinger G. The effect of augmented reality training on percutaneous needle placement in spinal facet joint injections. IEEE Trans Biomed Eng 2011 Jul;58(7):2031-2037. [doi: 10.1109/TBME.2011.2132131] [Medline: 21435970]
95. Gokkurt Yilmaz BN, Ozbey F, Yilmaz BE. Effect of artificial intelligence-assisted personalized feedback on radiographic diagnostic performance of dental students: a controlled study. BMC Med Educ 2025 Oct 10;25(1):1403. [doi: 10.1186/s12909-025-07875-4] [Medline: 41074130]
96. Zeng J, Sun K, Qin P, Liu S. Enhancing ophthalmology students' awareness of retinitis pigmentosa: assessing the efficacy of ChatGPT in AI-assisted teaching of rare diseases-a quasi-experimental study. Front Med (Lausanne) 2025;12:1534294. [doi: 10.3389/fmed.2025.1534294] [Medline: 40171502]
97. Al Kahf S, Roux B, Clerc S, et al. Chatbot-based serious games: a useful tool for training medical students? A randomized controlled trial. PLoS ONE 2023;18(3):e0278673. [doi: 10.1371/journal.pone.0278673] [Medline: 36913346]
98. Castano-Villegas N, Llano I, Villa MC, Zea J, Velásquez L. Clinical med students' validation of arkangel AI are their responses any better when supported by the AI? SSRN. Preprint posted online on 2025. [doi: 10.2139/ssrn.5679000]
99. Dao L, Harmouch SS, Chin A, et al. Effect of an artificial intelligence chest x-ray disease prediction system on the radiological education of medical students: a pilot study. medRxiv. Preprint posted online on 2022. [doi: 10.1101/2022.07.11.22277278]
100. Döner A, Ceyhan Ö, Taşci S. Effects of artificial intelligence-supported case analysis method on nursing students' clinical competencies: mixed design research. Nurse Educ 2026;51(3):E199-E205. [doi: 10.1097/NNE.0000000000002086] [Medline: 41441838]
101. Fang Q, Reynaldi R, Araminta AS, et al. Artificial intelligence (AI)-driven dental education: exploring the role of chatbots in a clinical learning environment. J Prosthet Dent 2025 Oct;134(4):1296-1303. [doi: 10.1016/j.prosdent.2024.03.038] [Medline: 38644064]
102. Huang Y, Xu BB, Wang XY, Luo YC, Teng MM, Weng X. Implementation and evaluation of an optimized surgical clerkship teaching model utilizing ChatGPT. BMC Med Educ 2024;24(1):CN. [doi: 10.1186/s12909-024-06575-9]
103. Jiang Y, Fu X, Wang J, et al. Enhancing medical education with chatbots: a randomized controlled trial on standardized patients for colorectal cancer. BMC Med Educ 2024 Dec 20;24(1):1511. [doi: 10.1186/s12909-024-06530-8] [Medline: 39707245]
104. Kobayashi M, Iwamoto M, Une S, Kurazume R, Nakazawa A, Honda M. Simulated communication skills training program effects using augmented reality with real‐time feedback: a randomized control study. Alzheimer's & Dementia 2022 Dec;18(S8) [FREE Full text] [doi: 10.1002/alz.062055]
105. Lau YH, Acharyya S, Wee CWL, et al. Effectiveness of traditional, artificial intelligence-assisted, and virtual reality training modalities for focused cardiac ultrasound skill acquisition: a randomised controlled study. Ultrasound J 2025 Nov 21;17(1):61. [doi: 10.1186/s13089-025-00469-7] [Medline: 41269495]

XSL•FO
RenderX

106. Lee HY, Kim J, Choi H, et al. Comparing AI chatbot simulation and peer role-play for OSCE preparation: a pilot randomized controlled trial. BMC Med Educ 2025 Nov 24;25(1):1755. [doi: 10.1186/s12909-025-08308-y] [Medline: 41286823]
107. Louie YLJ, Lin CTL, Fong BYR, Yung KK, See C, Poon LCY, et al. AI vs conventional learning of clinical skills in O&G - A comparison. BJOG 2025;132(18 - 9):CN. [doi: 10.1111/1471-0528.18215]
108. Luo MJ, Bi S, Pang J, et al. A large language model digital patient system enhances ophthalmology history taking skills. NPJ Digit Med 2025 Aug 4;8(1):502. [doi: 10.1038/s41746-025-01841-6] [Medline: 40760042]
109. Meng J, Chen J, Lai Y, Wang N, Ho CF, Tartari E, et al. AI-based augmented reality for hand hygiene: preliminary CRCT results in health sciences students. Antimicrob Resist Infect Control 2025:CN. [doi: 10.1186/s13756-025-01610-w]
110. Simsek-Cetinkaya S, Cakir SK. Evaluation of the effectiveness of artificial intelligence assisted interactive screen-based simulation in breast self-examination: an innovative approach in nursing students. Nurse Educ Today 2023 Aug;127:105857. [doi: 10.1016/j.nedt.2023.105857] [Medline: 37253303]
111. Usta SN, Silva E, Tekkanat H, Keskin C. Virtual reality haptic simulators and mobile applications-potential AI-enhanced tools for improving clinical endodontic training: a randomized controlled trial. J Endod 2026 Mar;52(3):337-344. [doi: 10.1016/j.joen.2025.11.005] [Medline: 41242655]
112. Vannaprathip N, Haddawy P, Schultheis H, Suebnukarn S. SDMentor: a virtual reality-based intelligent tutoring system for surgical decision making in dentistry. Artif Intell Med 2025 Apr;162:103092. [doi: 10.1016/j.artmed.2025.103092] [Medline: 40015210]
113. Wu C, Chen L, Han M, Li Z, Yang N, Yu C. Application of ChatGPT-based blended medical teaching in clinical education of hepatobiliary surgery. Med Teach 2025 Mar;47(3):445-449. [doi: 10.1080/0142159X.2024.2339412] [Medline: 38614458]
114. Wu D, Xiang Y, Wu X, et al. Artificial intelligence-tutoring problem-based learning in ophthalmology clerkship. Ann Transl Med 2020 Jun;8(11):700. [doi: 10.21037/atm.2019.12.15] [Medline: 32617320]
115. Benfatah M, Elazizi I, Lamiri A, et al. AI-assisted prebriefing to enhance simulation readiness in nursing education. Teach Learn Nurs 2026 Jan;21(1):e57-e63. [doi: 10.1016/j.teln.2025.07.030]
116. Benfatah M, Marfak A, Saad E, Hilali A, Nejjari C, Youlyouz-Marfak I. Assessing the efficacy of ChatGPT as a virtual patient in nursing simulation training: a study on nursing students' experience. Teach Learn Nurs 2024 Jul;19(3):e486-e493. [doi: 10.1016/j.teln.2024.02.005]
117. Jin HK, Lee HE, Kim E. Performance of ChatGPT-3.5 and GPT-4 in national licensing examinations for medicine, pharmacy, dentistry, and nursing: a systematic review and meta-analysis. BMC Med Educ 2024 Sep 16;24(1):1013. [doi: 10.1186/s12909-024-05944-8] [Medline: 39285377]
118. Liu M, Okuhara T, Chang X, et al. Performance of ChatGPT across different versions in medical licensing examinations worldwide: systematic review and meta-analysis. J Med Internet Res 2024 Jul 25;26:e60807. [doi: 10.2196/60807] [Medline: 39052324]
119. Murray E. Challenges in educational research. Med Educ 2002 Feb;36(2):110-112. [doi: 10.1046/j.1365-2923.2002.01155.x] [Medline: 11869436]
120. Riding R. Individual differences and educational performance. Educ Psychol (Lond) 2005 Dec;25(6):659-672. [doi: 10.1080/01443410500344712]
121. The AI research assistant. Elicit. 2025. URL: https://elicit.com [accessed 2025-08-20]

Abbreviations

AI: artificial intelligence
AI-VR: artificial intelligence–virtual reality
DL: deep learning
LLM: large language model
ML: machine learning
PI: prediction interval
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
PRISMA-S: Preferred Reporting Items for Systematic Reviews and Meta-Analyses–Search extension
RCT: randomized controlled trial
ROBUST-RCT: Risk Of Bias Instrument for Use in Systematic Reviews for Randomised Controlled Trials
RR: relative risk
SMD: standardized mean difference
WHO: World Health Organization

XSL•FO
RenderX

Edited by S Brini, T Leung; submitted 04.Dec.2025; peer-reviewed by P Shilo, Y Hu; revised version received 30.Mar.2026; accepted 30.Mar.2026; published 05.May.2026.

Please cite as:
Lai NM, Lim YS, Win MT, Bhargava P, Thomas P, Ong QC
The Effectiveness of Artificial Intelligence in Undergraduate Health Professions Education: Systematic Review and Meta-Analysis of Randomized Controlled Trials
JMIR Med Educ 2026;12:e88933
URL: https://mededu.jmir.org/2026/1/e88933
doi:10.2196/88933

© Nai Ming Lai, Yin Sear Lim, Min Thein Win, Prabal Bhargava, Paraidathathu Thomas, Qi Chwen Ong. Originally published in JMIR Medical Education (https://mededu.jmir.org), 5.May.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Education, is properly cited. The complete bibliographic information, a link to the original publication on https://mededu.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Outcomes of Digital Training for Community Health Workers in Low- and Middle-Income Countries: Scoping Review

Tapiwa A Tembo[1], MSc; Nora Ellen Rosenberg[2], PhD; Firaol Ayele[1], MSc; Saeed Ahmed[1,3], MD; Linda-Gail Bekker[4], MD, PhD

[1]Baylor College of Medicine Children's Foundation Malawi, Off Mzimba Road, P/Bag B-397, Lilongwe, Malawi

[2]Department of Health Behavior, Gillings School of Global Public Health, University of North Carolina at Chapel Hill, Chapel Hill, NC, United States

[3]Department of Pediatrics, Baylor College of Medicine, Houston, TX, United States

[4]The Desmond Tutu HIV Centre, Institute of Infectious Disease and Molecular Medicine, Faculty of Health Sciences, University of Cape Town, Cape Town, South Africa

Corresponding Author:
Tapiwa A Tembo, MSc
Baylor College of Medicine Children's Foundation Malawi, Off Mzimba Road, P/Bag B-397, Lilongwe, Malawi

Abstract

Background: Community health workers (CHWs) play an important role in delivering essential health services in low- and middle-income countries (LMICs). Training CHWs using digital approaches is on the rise. Although scoping and systematic reviews of digital training have been conducted for medical professionals in high-income countries (HICs), none have been conducted with lay professionals in LMICs, a population with different considerations.

Objective: This review describes the characteristics of digital training for CHWs and identifies their impact on health services outcomes in LMICs.

Methods: A scoping review approach based on Arksey and O'Malley's guiding principles was used to retrieve, review, and analyze existing literature. We searched 10 foremost databases using keywords and Medical Subject Headings terms for CHWs, LMICs, and digital learning to identify primary, peer-reviewed studies published up to and including November 26, 2024. An updated search of studies in all the databases was conducted on January 12, 2026, by the research team. No registries were searched. Articles that focused on the provision of digital or blended learning training for CHWs working in LMICs in any disease domain evaluating a learning, implementation, or clinical outcome met the eligibility criteria. Two reviewers (TAT and FA) screened the articles at the title and abstract levels and at full-text review. Study details, study designs, training attributes, technology and CHW descriptions, and outcomes were abstracted using a data-charting form. Descriptive analysis was conducted of the population, training characteristics, and reported outcomes. The PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) guidelines for reporting scoping reviews were used.

Results: A total of 892 articles were retrieved and screened for eligibility, of which 18 original articles met the inclusion criteria. Most (n=13) were conducted in Asia. Most (n=15) used nonrandomized study designs. The most common attributes included synchronous (n=8), accessible in the community (n=14), use of smartphones (n=6), and accessible online (n=9). The majority reported learning outcomes (n=14), about half reported implementation outcomes (n=10), and only one reported clinical outcomes (n=1). The learning outcomes focused on knowledge gained and were mostly positive. The implementation outcomes included CHW's acceptability and feasibility to use the digital training approach. The clinical outcome was effectiveness.

Conclusions: We found few evaluations of digital training for CHWs in LMICs, in spite of a proliferation of such trainings. Digital trainings had a broad range of attributes. Many evaluations had knowledge, acceptability, and feasibility outcomes. However, other learning outcomes (eg, attitudes and skills), implementation outcomes (eg, appropriateness and fidelity), and clinical outcomes were rare. Most lacked experimental designs. Although the existing evidence suggests that digital training can impact knowledge in lay health workers in LMICs, more rigorous studies with a broader range of outcomes are needed.

(JMIR Med Educ 2026;12:e82772) doi:10.2196/82772

KEYWORDS

community health worker; digital training: blended learning; eLearning; LMIC; lay health worker

XSL•FO
RenderX

Introduction

Low- and middle-income countries (LMICs) face a shortage of professional health care workers (HCWs) [1]. To overcome this gap, lay HCWs or community health workers (CHWs) have been recruited and trained to carry out an increasing array of tasks [2]. CHWs are laypeople working within their own community, performing functions related to health care delivery and health promotion, but have not received formal professional or paraprofessional certificates or degrees [3]. CHWs are often trained for specific tasks such as HIV testing, disease screening, or provision of immunization and have been recognized as critical role-players within the primary health care setting [4], where task shifting to CHWs has been shown to be a cost-effective method of service delivery scale-up [5,6]. When provided with the correct resources, training, and support, CHWs have improved accessibility to basic health services, resulting in better health outcomes [7,8].

The World Health Organization (WHO) recommends that CHWs receive regular training and supervision to fulfill their roles successfully [9]. To increase the competency of CHWs in health care provision, there is a need for more effective, higher-quality, and easily accessible health training [10]. The design of these training approaches would ideally also minimize the burden on an already strained health care system by limiting the time CHWs spend away from workstations while attending training required to improve service delivery.

The use of digital technology in CHW education may help overcome these challenges. In line with this, the increasing availability of technology could address the shortcomings of in-service training provision to CHWs. Digital training may provide more accessible, standardized, relevant, timely, and affordable solutions [11,12]. Moreover, digital training may provide flexibility that would allow participants to balance their tasks and learning endeavors effectively [13,14]. Given the widespread availability of technology devices, digital training by itself or in combination with face-to-face classroom training (ie, blended learning) has been widely used for medical education in a variety of settings [5,9,15-17]. These training approaches have facilitated improvements in knowledge, skills, overall competence, and clinical performance of health professionals across various health care settings, such as clinicians, nurses, and public health practitioners [18,19]. In addition, the COVID-19 pandemic necessitated and reinforced the practice of digital training for health workers, including lay cadres [14,20]. However, regardless of the benefits associated with digital training, the evidence surrounding these trainings for CHWs has not yet been synthesized.

Given that there are over 8 million CHWs in LMICs [21,22] globally, ensuring that they all receive optimal and appropriate training efficiently and cost-effectively is a challenge but also a critical priority. Digital solutions may present a viable option given that digital training can be effective in high-income countries (HICs) with professional cadres. However, evidence on the impact of digital training on capacity-building for CHWs on factors including knowledge, skills enhancement, and improved health care needs to be explored and understood. It is important to know whether these solutions can be applied to nonprofessional cadres and LMIC settings where there are limited digital skills and infrastructures [23,24], and whether they will be found to be acceptable, feasible, and effective. To develop and implement successful digital training programs for CHWs in resource-constrained settings, a greater understanding is needed of the training attributes and the extent to which these training modalities enhance CHW capacity. The Kirkpatrick evaluation framework articulates four levels of training outcomes: (1) participant reaction to the training program experience; (2) learning outcomes; (3) participant's change in behavior; and (4) impact on the clinical setting [25]. We undertook this scoping review with this framework in mind. The overall purpose of this scoping review was to synthesize the evidence from the literature surrounding digital and blended learning among lay HCWs in LMICs to highlight areas for future research and implementation. Specifically, we sought to (1) characterize populations using these digital tools by geography and cadres, (2) describe the nature and attributes of blended and digital learning tools, (3) organize types of training outcomes examined by the levels defined in the Kirkpatrick framework, and (4) examine preliminary learning, implementation, and clinical outcomes.

Methods

Overview

We conducted a scoping review in accordance with the PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews) guidelines [26] on the provision of digital and blended learning training for CHWs in LMICs. Our scoping review followed systematic and transparent research steps guided by the framework described by Arksey and O'Malley [27] and updated by Levac et al [28], to characterize the attributes and nature of digital and blended learning training for CHWs in LMICs, as well as the outcomes considered. An internal protocol was developed for the review to define the inclusion and exclusion criteria and the review methods before data extraction.

Search Strategy

The Cochrane Library and PROSPERO (International Prospective Register of Systematic Reviews) were searched to identify available or ongoing scoping and systematic reviews pertaining to the provision of digital or blended training for CHWs in LMICs. No previous or ongoing relevant reviews were identified.

We then designed a comprehensive search strategy to identify all relevant studies according to the PRISMA-S (PRISMA-Search Extension) guideline [29]. We did not use published search filters. There were no search strategies that were adapted from other studies. A publication date limit was not set. The initial search was conducted on November 26, 2024. Twenty relevant search terms for "Community Health Workers" and 21 search terms for "digital learning" and "blended learning" were developed. These were combined with the World Bank Group 2022 list of LMICs using Boolean operators to develop a master search query. Where appropriate, each index-linked Medical Subject Headings term was expanded to contain all

XSL•FO
RenderX

relevant subheadings. In addition, synonyms were searched for each key term, along with wildcards and truncation for free-text words. The following databases were searched to identify primary, peer-reviewed studies published up to and including January 12, 2026, including Cumulated Index in Nursing and Allied Health Literature (CINAHL), Cochrane, Embase, Education Resources Information Center (ERIC), Global Health, Google Scholar, PsycINFO, PubMed, Scopus, and Social Science Research Network (SSRN). The search terms were adjusted accordingly for each database. An updated search of studies, rerunning the searches, was conducted on January 12, 2026, by the research team. Cited references were examined for any relevant studies by browsing reference lists. Study registries were not searched. A full record of the conducted search for each database is provided in the online supplementary material (Table S1 in Multimedia Appendix 1).

Eligibility Criteria

Studies were included in the review if the population comprised primary participants who were CHWs [3], the concept involved CHWs being trained using digital or blended learning modalities, and the context included the following: (1) the training occurred in a country defined as an LMIC according to the World Bank Group 2022 classification of economies [30]; (2) the primary data were collected; (3) primary focus of the training addressed a communicable or noncommunicable disease domain; and (4) the article reported a learning, implementation, or clinical outcome. Digital learning was defined as a practice of learning using technologies to deliver educational content and training programs [31,32]. Blended learning was an education approach that combines face-to-face and digital learning approaches [33]. The training outcomes needed either to have been evaluated within the same group through pre-post study design or with a control group through randomized or nonrandomized controlled trials.

The scoping review did not restrict based on study design and included both qualitative and quantitative outcomes. Studies had to qualify as an original, full-text research study to be considered for inclusion in the review. Review articles, commentaries, letters, policy briefs, protocols, needs assessments, and conference abstracts were excluded.

Outcomes

We assessed learning, implementation, or clinical outcomes and noted the studies that reported each outcome. The learning outcomes included knowledge, attitudes, and behaviors. “Knowledge” referred to information acquired through the training. Attitudes focused on confidence to perform tasks. Behaviors focused on whether trainees used what they learned. The implementation outcomes included acceptability, appropriateness, feasibility, and fidelity. Acceptability was defined as the CHW perception that the training approach was agreeable or satisfactory. Appropriateness was the CHW’s perceived fit of the training. Feasibility was defined as the extent to which the training was successfully carried out in the setting. Fidelity was defined as the degree to which the training was delivered as intended. Clinical outcomes included how well the training improved patients’ uptake in clinical practice [34].

Selection of Sources of Evidence

All articles identified via database searching and other methods, including citation searching were exported into EndNote and imported into Covidence review software, and duplicate references were removed. The screening process was performed according to PRISMA-ScR. Titles and abstracts of all the articles identified in the search were screened by 2 authors (TAT and FA) to determine whether they would be considered relevant for a full-text review. For title and abstract screening, interreviewer reliability with Cohen κ was 0.84. Based on Landis and Koch’s [35] threshold values reference, a κ of $\geq$0.81 is considered “almost perfect agreement.” The full text of all the articles identified as potentially relevant was then retrieved and reviewed in full against the inclusion and exclusion criteria by both reviewers. All articles that did not meet the eligibility criteria were excluded from the review database, and reasons for exclusion were recorded. At all stages of the review, discrepancies between the authors were resolved via discussion. Where appropriate, the authors of individual papers were contacted for further information. References and other sources were reviewed to identify any other articles for full-text review. A critical appraisal of individual sources of evidence was not performed.

Data Charting, Extraction, and Synthesis

Data were independently extracted and tabulated by 2 authors (TAT and FA) using a data charting form in a Microsoft Excel spreadsheet that was approved by the research team. The data charting form was used to extract information regarding the population characteristics, training attributes, and identified reported outcomes. The use of a data charting form or table was recommended by Arksey and O’Malley [27] and Levac et al [28]. The data extraction form was piloted by 2 authors on 5 studies prior to use to ensure that all necessary data were captured appropriately. Where necessary, the corresponding authors for relevant studies were contacted via email to clarify aspects of their work prior to final extraction.

The bibliographic data extracted included the first author, title, year of publication, and country in which the study took place. CHW descriptions included CHW cadre name, number of CHWs trained, sex, age, education, disease domain, employment type (full-time or part-time), duty station (community or facility-based), remuneration (presence or absence), and employer. Training features included type of training, modality, synchronicity, venue, device type, availability online or offline, technology medium, duration, number of sessions, and pedagogical approaches. Outcomes included measurements, results, and limitations. The study design information was extracted and categorized into cross-sectional cohort, longitudinal cohort, quasi-experimental, and randomized controlled trials. The analytic approaches were classified into qualitative, quantitative, and mixed methods.

Once the data had been transferred into the spreadsheet, one author (TAT) reviewed the information and organized it into one of the following categories: (1) learning outcomes, (2) implementation outcomes, and (3) clinical outcomes. The descriptions of the included studies were analyzed and organized in tabular format, accompanied by a narrative summary.

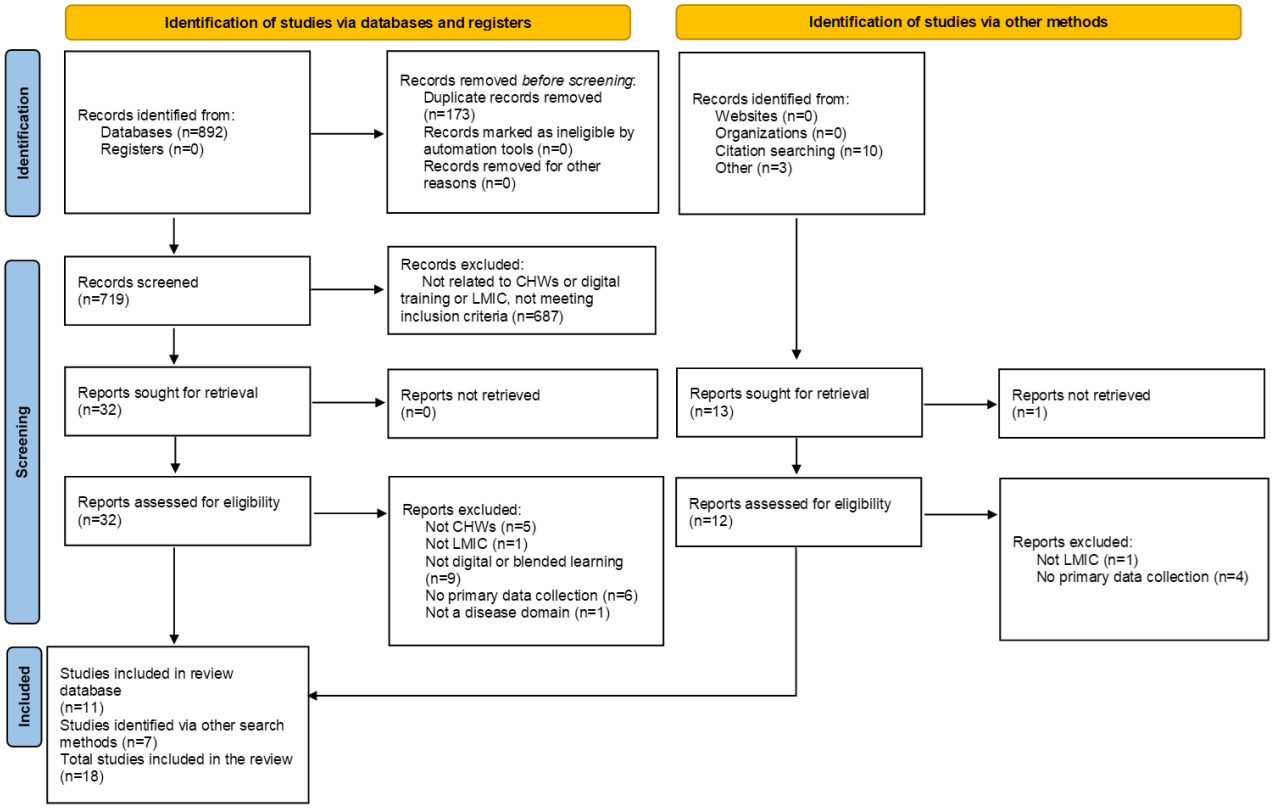
XSL•FO
RenderX

Results

Search Results

The initial search of the 10 databases yielded 892 articles (see online Table S2 in Multimedia Appendix 1). After the exclusion of 173 duplicates, 719 papers were identified for initial screening. After title and abstract screening, 32 studies were identified for full-text review. An additional 13 articles were selected from reference lists and other sources, totaling 45 articles for full-text review. After the full text review, 18 studies [36-53] were identified for data extraction, and 27 were excluded because they did not meet inclusion criteria. Reasons for exclusion at full-text screening can be found in the PRISMA flowchart (Figure 1).

Figure 1. PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) flow diagram. CHW: community health worker; LMIC: low- and middle-income country.

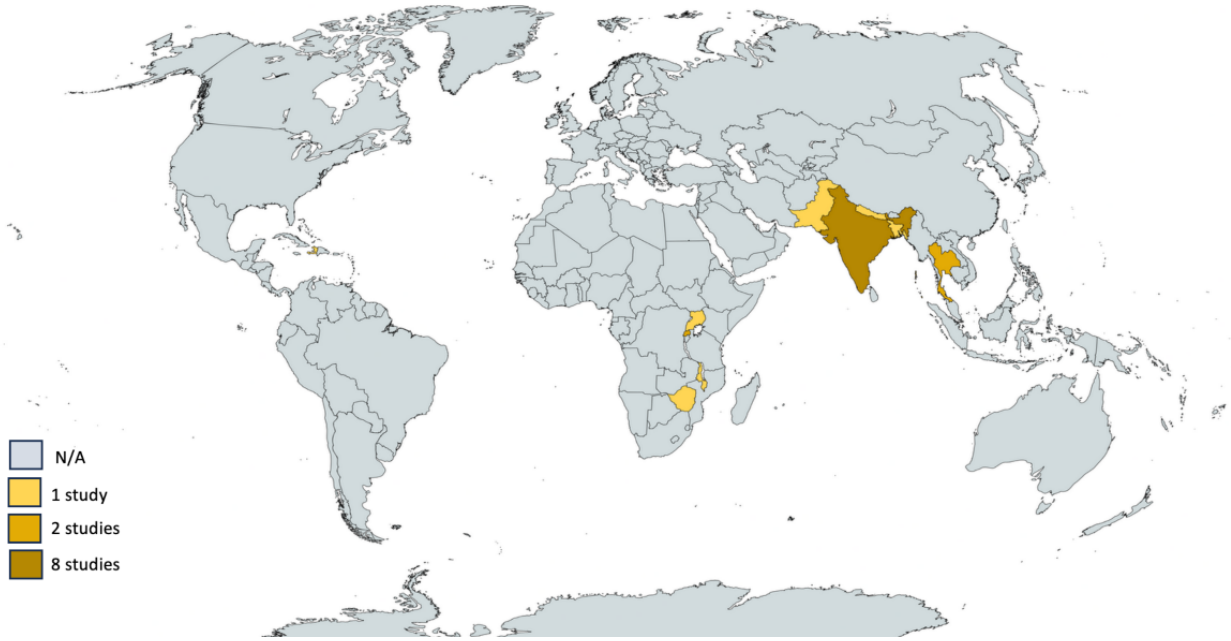


Characteristics of Included Studies, CHWs, and Training

All 18 [36-53] studies in this review were published from 2009 to 2024. A majority of studies were conducted in Asia (n=13) [37-43,45,47,48,51-53], especially in India (n=8) [37,39,41-43,51-53], a few were conducted in Africa (n=5) [36,44,46,49,50], and one study was conducted in the Caribbean [37]. None were conducted in Latin America, Eastern or Southern Europe, or Oceania (Figure 2). The countries where the studies were undertaken are summarized in Table 1.

XSL•FO
RenderX

Figure 2. Map of countries where studies included in the scoping review were conducted.



XSL•FO
RenderX

Table . Details of studies included in the review.

First author, year	Title	Country	Community health workers name	Number of community health workers	Community health worker sex	Community health worker age	Community health worker education	Disease or domain of focus	Full-time or part-time	Community- or facility-based	Paid or incentivized or not paid	Government or nongovernmental organization
Bertman et al (2019) [36]	Health worker text messaging for blended learning, peer support, and mentoring in pediatric and adolescent HIV/AIDS care: a case study in Zimbabwe	Zimbabwe	Primary counselors (PCs)	293	Male and female	≥25 years	Secondary school education	HIV	Full-time	Both	Paid	Government
Khan et al (2019) [37]	An electronic-based curriculum to train acute care providers in rural Haiti and India	Haiti and India	Acute care providers	Haiti: n=6; India: n=55	Haiti: male and female; India: female	Haiti: mean age 24 years; India: mean age 39 years	Haiti: high school diploma; India: primary education	General acute conditions	Full-time	Community	Not reported	Not reported
Kharel et al (2022) [38]	Training program for female community volunteers to combat COVID 19 in rural Nepal	Nepal	Female community health volunteers (FCHV)	183	Female	Not reported	Not reported	COVID-19	Part-time	Community	Paid	Government
Lakshminarayanan et al (2020) [39]	Delivery of perinatal mental health services by training lay counselors using digital platforms	India	Lay counselors	23	Female	Not reported	Not reported	Perinatal mental health	Part-time	Community	Not reported	Not reported

XSL•FO
RenderX

First author, year	Title	Country	Community health workers name	Number of community health workers	Community health worker sex	Community health worker age	Community health worker education	Disease or domain of focus	Full-time or part-time	Community- or facility-based	Paid or incentivized or not paid	Government or nongovernmental organization
Limaye et al (2019) [40]	Enhancing the knowledge and behaviors of fieldworkers to promote family planning and maternal, newborn, and child health in Bangladesh through a digital health training package: results from a pilot study	Bangladesh	Community health workers (CHWs; Field workers)	Pre: 306 and post: 265	Female	Mean age pre: 35 (SD 12) years and mean age post: 34 (SD 12) years	Not reported	Maternal, newborn, and child health and family planning	Full-time	Community	Not reported	Government
Muke et al (2019) [41]	Acceptability and feasibility of digital technology for training community health workers to deliver brief psychological treatment for depression in rural India	India	Accredited social health activist (ASHA)	32	Female	24 - 45 years	Minimum education level of grade 8	Depression	Part-time	Community	Paid	Government

XSL•FO
RenderX

First author, year	Title	Country	Community health workers name	Number of community health workers	Community health worker sex	Community health worker age	Community health worker education	Disease or domain of focus	Full-time or part-time	Community- or facility-based	Paid or incentivized or not paid	Government or nongovernmental organization
Muke et al (2020) [42]	Digital training for nonspecialist health workers to deliver a brief psychological treatment for depression in primary care in India: findings from a randomized pilot study	India	ASHA	45	Female	≥18 years	Minimum education level of 8th standard	Mental health	Part-time	Community	Paid	Government
Nedungadi et al (2019) [43]	Rural health in digital India: interactive simulations for community health workers	India	CHWs	23	Female	≥18 years	Minimum education level of grade 8	Communicable and noncommunicable diseases and nutrition deficiency	Part-time	Community	Paid	Government
O'Donovan et al (2018) [44]	The use of low-cost Android tablets to train community health workers in Mukono, Uganda, in the recognition, treatment and prevention of pneumonia in children under five: a pilot randomized controlled trial	Uganda	CHWs	163	Male and female	Mean age control group 44.6 (SD 12.5) years; intervention group 43.7 (SD 10.3) years	9 years of education	Pneumonia	Part-time	Community	Not paid	Government

XSL•FO
RenderX

First author, year	Title	Country	Community health workers name	Number of community health workers	Community health worker sex	Community health worker age	Community health worker education	Disease or domain of focus	Full-time or part-time	Community- or facility-based	Paid or incentivized or not paid	Government or nongovernmental organization
Rahman et al (2019) [45]	Using technology to scale-up training and supervision of community health workers in the psychosocial management of perinatal depression: a non-inferiority, randomized controlled trial	Pakistan	CHWs	80	Female	Mean age control group 35 (SD 8) years; intervention group 36 (SD 7) years	Not reported	Perinatal depression	Full-time	Community	Paid	Government
Sangwa et al (2024) [46]	Using eLearning to improve and retain the knowledge of community health workers in maternal and neonatal health in Rwanda: a cohort study	Rwanda	CHWs	36	Female	≥25 years	Completed primary education	Maternal and newborn health	Part-time	Community	Paid	Government
Sanchaicharnpong et al (2009) [47]	Process and outcome evaluation of a diabetes prevention education program for community health care workers in Thailand	Thailand	Community health care workers (CHCW)	69	Male and female	25 - 54 years	Diploma level to bachelor's	Diabetes	Full-time	Facility-based	Paid	Government

XSL•FO
RenderX

First author, year	Title	Country	Community health workers name	Number of community health workers	Community health worker sex	Community health worker age	Community health worker education	Disease or domain of focus	Full-time or part-time	Community- or facility-based	Paid or incentivized or not paid	Government or non-governmental organization
Sanchaisuriyong and Hanning (2012) [48]	Diabetes prevention education program for community health care workers in Thailand	Thailand	Community health care workers (CHCW)	69	Male and female	25 - 54 years	Diploma level to bachelor's	Diabetes	Full-time	Facility-based	Paid	Government
Tembo et al (2021) [49]	Pilot-testing a blended learning package for health care workers to improve index testing services in Southern Malawi: an implementation science study	Malawi	CHWs and HIV diagnostic assistants	12	Male and female	20 - 42 years	Completed secondary education	HIV	Full-time	Facility-based	Paid	Non-governmental organization
Willems et al (2021) [50]	Co-creation and evaluation of nationwide remote training service for mental health education of community health workers in Rwanda	Rwanda	CHWs	51,858	Male and Female	20 - 50 years	Not reported	Mental health	Part-time	Community	Paid	Government

XSL•FO
RenderX

First author, year	Title	Country	Community health workers name	Number of community health workers	Community health worker sex	Community health worker age	Community health worker education	Disease or domain of focus	Full-time or part-time	Community- or facility-based	Paid or incentivized or not paid	Government or nongovernmental organization
Yadav et al (2017) [51]	Sangoshthi: empowering community health workers through peer learning in rural India	India	ASHA	40	Female	26 - 50 years	Minimum education level of grade 10	Maternal and child health	Part-time	Community	Paid	Government
Yadav D (2017) [52]	Low-cost mobile learning solutions for community health workers	India	ASHA	40	Female	Not reported	Not reported	Maternal and child health	Part-time	Community	Paid	Government
Yadav et al (2019) [53]	LEAP: scaffolding collaborative learning of community health workers in India	India	ASHA	120	Female	25 - 45 years	Minimum education level of grade 8	Maternal and child health	Part-time	Community	Paid	Government

Six different CHW terms were identified across the 18 studies, with variations being noted between studies in terms of CHW sex, type of employment contract, duty station, remuneration, and employer. The terms “community health worker” (n=9) [40,43-50] and “accredited social health activist” (n=5) [41,42,51-53] were commonly used. “Lay counselor,” “primary counselor,” “female community health volunteers,” and “acute care provider” were used in one study each. The majority of studies reported employing female CHWs only (n=11) [37-43,45,51-53] and the others employed males and females. The age of CHWs varied from 18 to 54 years. Most CHWs worked part-time (n=10) [38,39,41-44,46,50-53]; were community-based (n=14) [37-46,50-53]; were paid a monthly salary, incentive, or honorarium (n=14) [36,38,41-43,45-53]; and were employed by the government (n=15; Table 1) [36,38,40-44,46-48,50-53].

A total of 4 study designs were identified. Studies were cross-sectional (n=9) [36,37,39-41,43,47,48,50], cohort (n=1) [46], quasi-experimental (n=5) [38,49,51-53], and randomized controlled trials (RCTs, n=3) [42,44,45]. With regard to the analytic approaches applied, most studies were quantitative (n=10) [37-40,44-49] and mixed methods (n=6) [42,43,50-53]. Two used qualitative methods only [36,41].

The training modalities described in the studies were digital learning (n=14) [37-46,50-53] and blended learning (n=4) [36,47-49]. Some training approaches made use of tablets or smartphones (n=6) [36,37,42,44-46]. This was followed by computers (n=5) [40,43,47-49] and basic or feature phones (n=1) [50]. There were some studies that reported the use of multiple technologies (n=5) [39,41,51-53]. The training approaches were all implemented by researchers (n=18) and supported by government (n=2) [46,50] and nongovernmental organizations (n=1) [49]. The training formats were synchronous (n=8) [38,41-43,49,51-53], asynchronous (n=5) [37,40,44,46,50], and both (synchronous and asynchronous) (n=5) [36,39,45,47,48].

The training venue, duration, focus, and pedagogical approaches varied between studies. The majority of the training occurred in the community (n=14) [37-46,50-53] compared to a classroom (n=3) [47-49] or both (n=1) [36]. The trainings focused on the following disease areas: COVID-19 prevention, testing, and management (n=1) [38]; diabetes prevention (n=2) [47,48]; general conditions care or management (n=2) [37,43]; HIV testing services (n=2) [36,49]; maternal and child health (n=5) [40,46,51-53], mental health or depression screening and counseling (n=5) [39,41,42,45,50]; and pneumonia recognition,

XSL•FO
RenderX

treatment, and prevention (n=1) [44]. Most of the training was accessible online (n=8) [36,38,39,41-43,46,53] rather than offline (n=6) [37,40,44,45,49,50] or both (n=4) [47,48,51,52]. The duration ranged from a minimum of one day to a maximum of 8 months. They lasted from one to 48 hours. The main pedagogical approaches employed were individual learning (n=5) [39-41,44,50] and group discussion (n=4) [47,48,51,52]. Other approaches included practicing scenarios, modeling, and receiving feedback to reinforce information [36-38,42,43,45,49] (Table 2). Full details of the training are summarized in Table 3

XSL•FO
RenderX

Table . Digital and blended learning training characteristics.

Study	Type of training	Training modality	Synchronicity (or both)	Health facility or conference or community	Device	Technology medium	Method of training (online vs offline)	Training accessible offline (yes or no)	Time spent in training	Training duration	Number of training sessions or modules	Pedagogical approaches
Bertman et al [36]	In-service	Blended learning	Both	Facility and classroom	Tablet and mobile phone	Videos	Online	Not reported	Not reported	7 weeks	Not reported	Learning, group discussion, and feedback
Khan et al [37]	Preservice	Digital training	Asynchronous	Community	Tablet	Videos	Haiti: online and offline; India: offline	Yes	Not reported	Haiti: 8 months ; India: 4 months	Haiti: 39 modules; India: 14 modules	Learning, group discussion, and practice
Kharel et al [38]	In-service	Digital training	Synchronous	Community	Not reported	Videos	Online	No	4 hours	1 day	3 modules	Learning, modeling, and discussion
Lakshminarayanan et al [39]	In-service	Digital training	Both	Community	Mobile phone, tablets, and basic computers	Videos	Online	no	20 - 25 h	1 month	10 sessions	Learning
Limaye et al [40]	In-service	Digital training	Asynchronous	Community	Computer	Videos	Offline	Yes	Not reported	4 months	8 courses	Learning
Muke et al [41]	In-service	Digital training	Synchronous	Conference or classroom	Tablets, mobile phone, and laptops	Videos	Online	No	2 - 3 hours	1 day	2 modules	Learning
Muke et al [42]	In-service	Digital training	Synchronous	Community	Mobile phone	Videos	Online	No	48 hours	30 days	16 modules	Learning and modeling
Nedungadi et al [43]	In-service	Digital training	Synchronous	Conference	Computer	Videos	Online	No	Not reported	1 day	Not reported	Learning and practicing
O'Donovan et al [44]	In-service	Digital training	Asynchronous	Community	Tablet	Videos	Offline	Yes	Not reported	5 days	4 sessions	Learning
Rahman et al [45]	In-service	Digital training	Both	Community	Tablet	Videos	Offline	No	20 hours	5 days	Not reported	Learning, modeling, and group discussion
Sangwa et al [46]	In-service	Digital training	Asynchronous	Community	Mobile phone	Videos	Online	No	Not reported	4 weeks	4 sessions	Learning
Sanchaicenpong et al [47]	In-service	Blended learning	Both	Conference and community	Computer	Videos	Online and offline	No	20 - 24 h	4 months	8 modules	Learning and group discussion

XSL•FO
RenderX

Study	Type of training	Training modality	Synchronicity (or both)	Health facility or conference or community	Device	Technology medium	Method of training (online vs offline)	Training accessible offline (yes or no)	Time spent in training	Training duration	Number of training sessions or modules	Pedagogical approaches
Sanchaisuriyong and Hanning [48]	In-service	Blended learning	Both	Conference and community	Computer	Videos	Online and offline	No	Not reported	4 months	8 modules	Learning and group discussion
Tembo et al [49]	In-service	Blended learning	Synchronous	Conference	Computer	Videos	Offline	Yes	16 hours	2 days	Not reported	Learning, modeling, practicing, and feedback
Willems et al [50]	In-service	Digital training	Asynchronous	Community	Feature phone	Audios	Offline	Yes	40 minutes	4 weeks	8 modules	Learning
Yadav et al [51]	In-service	Digital training	Synchronous	Community	Mobile phone and feature phone	Audios	Online and offline	No	18 hours	22 days	12 sessions	Learning and group discussion
Yadav [52]	In-service	Digital training	Synchronous	Community	Mobile phone and feature phone	Audios	Online and offline	No	18 hours	22 days	12 sessions	Learning and group discussion
Yadav et al [53]	In-service	Digital training	synchronous	Community	Mobile phone and feature phone	Audios	Online	No	7.5 - 10 h	6 weeks	10 sessions	Group discussion

XSL•FO
RenderX

Table . Panel of the population, training description, and outcomes.

Studies	Training description and outcomes
Bertman et al [36]	• Who: a total of 293 primary counselors from 233 health facilities in Zimbabwe were trained in the Ministry of Health HIV testing guidelines for children and adolescents. • What: the 7-week blended training included tablet-based self-study, a one-and-a-half-day classroom session at the beginning of the course, and a 4-hour classroom session at the end. The self-study component included videos, podcasts, case studies, quizzes, and self-reflection questions. These activities were complemented by the SMS text messaging component, including a whole-group discussion forum as well as 2-person partner discussions. The course was facilitated by 3 trainers who presented content, managed logistics, and interacted with the WhatsApp (Meta) groups. The training incorporated group discussions and peer support through WhatsApp, where participants could share complex cases, seek advice, and engage in problem-solving. • Outcome: the SMS text messaging–based peer support and learning platform was well-received by participants, with high levels of engagement in case discussions while ensuring anonymity. Feedback indicated that the WhatsApp groups helped identify knowledge gaps and provided meaningful learning opportunities. The platform facilitated access to supervisors and experts, enriching the learning experience and fostering collaboration.
Khan et al [37]	• Who: a total of 6 previously untrained lay individuals in Haiti and 55 from India were trained as acute care providers (ACP) for general acute medical conditions. • What: trainees participated in electronic, asynchronous training using videos preloaded on tablets focusing on the diagnosis, treatment, and triage of 30 acute care chief complaints specific to these communities. The training included self-study modules, weekly discussions via video conferences, and multiple assessments. In Haiti, trainees completed 39 modules over 8 months, while in India, they completed 14 modules over 4 months. • Outcome: ACP demonstrated significant improvements in knowledge in diagnosing acute conditions. The training program was feasible to implement, with nearly all completing the training and passing the posttest.
Kharel et al [38]	• Who: a total of 183 female community health volunteers (FCHVs) in rural Nepal were trained to support the COVID-19 response. • What: the 3-module digital training program aimed to equip FCHVs with practical knowledge and skills to combat the COVID-19 pandemic. The 4-hour training included interactive didactic modules, case-based learning, discussions, and video simulations. Training materials were translated into Nepali and delivered by a lead facilitator fluent in the language. • Outcome: the training led to significant improvements in knowledge among FCHVs. The program demonstrated that virtual training is feasible in low-resource settings and can effectively enhance the capacity of community health workers.
Lakshminarayanan et al [39]	• Who: a total of 23 lay counselors were trained to understand, identify, and respond to perinatal mental health services in India. • What: the digital training had 10 sessions delivered over a period of 1 month, each session lasted 2-2.5 h. Training was delivered via videos, videoconferencing and instant messaging, file transfer, screen sharing via mobile phones, tablets, and basic computers. • Outcome: trained counselors were satisfied and improved knowledge to deliver mental health counseling.

XSL•FO
RenderX

Studies	Training description and outcomes
Limaye et al [40]	• Who: a total of 306 frontline workers (FWs) in rural Bangladesh were trained to promote family planning and maternal, newborn, and child health. • What: a digital training package was implemented for a 4-month duration using 8 courses accessible offline via netbooks (basic computers). Training incorporated multimedia elements, interactive learning, videos, and text messages to reinforce key concepts. • Outcome: significant improvements in mean knowledge scores across family planning and maternal, newborn, and child health topics were observed. Measurable changes in counseling behaviors were reported.
Muke et al [41]	• Who: a total of 32 accredited social health activists (ASHAs) in rural India were trained to treat depression by delivering the Healthy Activity Program (HAP). • What: training was conducted using videos, PowerPoints (Microsoft Corp) with voice-over narration using tablets, mobile phones, and laptops. Training was for 2 - 3 h. • Outcome: ASHAs found the training acceptable for learning new content and feasible to use.
Muke et al [42]	• Who: a total of 45 ASHAs participated in the study in Madhya Pradesh, India, and were trained to treat depression by delivering the Healthy Activity Program (HAP). • What: ASHAs randomized to the digital training received 48 hours of training over 30 days using a mobile phone through the Moodle Learning Management System. The training sessions consisted of videos, role-play videos, PowerPoint presentations, and reading materials. • Outcome: the training improved knowledge and skills in delivering the HAP for depression.
Nedungadi et al [43]	• Who: a total of 23 community health workers (CHWs) trained in the management of communicable and noncommunicable diseases and nutritional deficiency in rural India. • What: the SwastyaSIM is a digital training accessible online via a smartphone or computer designed to improve health literacy among CHWs by offering interactive simulations for medical training, including diagnostic tests. • Outcome: SwastyaSIM significantly improved the knowledge of CHWs, particularly in infection control practices.
O'Donovan et al [44]	• Who: a total of 163 community health workers in Mukono, Uganda, participated in a pilot randomized controlled trial. • What: CHWs participated in a digital training to recognize, treat, and prevent pneumonia in accordance with integrated Community Case Management (iCCM) using tablets preloaded with videos in the local dialect (Luganda) with English subtitles for 5 days. • Outcome: training showed improvements in test scores, in terms of knowledge acquisition post training.
Rahman et al [45]	• Who: a total of 80 lady health workers (LHWs) in Pakistan were trained in the psychosocial management of perinatal depression. • What: LHWs participated in the 20-hour digital training consisting of videos, role-play videos, sharing experiences, and problem-solving strategies over 5 days. • Outcome: LHW competency improved after the training.
Sangwa et al [46]	

XSL•FO
RenderX

Studies	Training description and outcomes
	• Who: a total of 36 CHWs in Rutsiro (rural) and Ngoma (periurban) districts of Rwanda were trained in maternal and neonatal health. • What: the training was delivered via the Ministry of Health's (MOH) eLearning platform, based on Moodle (Moodle Pty Ltd) through a smartphone over 4 weeks. • Outcome: postintervention and 6-month follow-up tests showed a significant increase in knowledge. Older CHWs performed better in follow-up tests, while all experience levels demonstrated consistent improvement, highlighting the effectiveness of eLearning across diverse learner groups.
Sranacharoenpong et al [47]	• Who: a total of 69 community health care workers (CHCWs) in Thailand were trained in diabetes prevention education training. • What: the training integrated traditional and e-learning approaches, using 8 in-classroom sessions and 8 online learning sessions for a 4-month duration. Each session was for 2.5 - 3 h. Learning materials included videos, lecture notes, PowerPoint presentations, and monthly newsletters. A facilitator coordinated discussions and provided virtual support during the training period. • Outcome: postintervention knowledge scores significantly improved. Participants demonstrated enhanced skills in measuring health metrics and understanding dietary recommendations.
Sranacharoenpong and Hanning [48]	• Who: a total of 69 community health care workers (CHCWs) were trained in a tailored diabetes prevention education program in Chiang Mai, Thailand. • What: the intervention involved 8 group classes and 8 self-directed e-learning sessions using videos via a computer over 4 months. Each session was for 2.5-3 h. • Outcome: knowledge improvement was significant, and CHCWs felt confident in applying the knowledge and believed they could teach at-risk populations about diabetes prevention.
Tembo et al [49]	• Who: a total of 12 health care workers (HCWs) from 6 facilities in Southern Malawi were trained to deliver Malawi's index case testing services. • What: the HCWs were trained using a blended learning package (combining digital and face-to-face approaches) for 16 hours in 2 days to improve Malawi's HIV index case testing services. • Outcome: the blended learning package resulted in significant improvements in HCWs' fidelity to the index testing protocol. The package was perceived to be acceptable by HCWs and improved a number of meaningful clinical indicators, including index clients counseled, contacts elicited, and contacts who reported for HIV testing.
Willems et al [50]	• Who: over 51,000 CHWs across Rwanda participated in a nationwide remote training service (RTS) focused on mental health education. • What: the RTS used Interactive Voice Response (IVR) technology to deliver audio-based training modules to CHWs via simple feature phones. Over 4 weeks, 2 modules lasting 5 minutes each were released weekly, totaling 8 modules covering topics such as common mental disorders, posttraumatic stress disorder (PTSD), depression, drug abuse, and epilepsy. • Outcome: knowledge of mental health topics significantly improved after the training. CHWs' awareness and self-confidence in identifying and referring patients with mental health conditions were enhanced.
Yadav et al [51]	

XSL•FO
RenderX

Studies	Training description and outcomes
	• Who: a total of 40 ASHAs in India participated in a maternal and child health care training. • What: the digital training included 12 sessions covering 10 topics over a period of 22 days. The content was adapted to local contexts and recorded in an audio dialog format. ASHAs accessed training via feature phones, while trainers used smartphones with internet connectivity to conduct the training. • Outcome: the ASHAs showed a significant improvement in knowledge postintervention compared to the control group.
Yadav [52]	• Who: a total of 40 community health workers in India participated in the training program. • What: the digital training consists of facilitators using a smartphone scheduling to call participants through their feature phones. The facilitator delivered the training using audios and needed an internet connection to conduct the training while the participant needs to be where there is a telecommunication network to participate in the training. • Outcome: the training program significantly improved ASHAs' knowledge on maternal and child health. The training allowed for interaction and discussion with experts and peer learning, fostering active engagement.
Yadav et al [53]	• Who: a total of 120 ASHAs in India were trained in maternal and child health through a peer-led educational intervention. • What: ASHAs were provided with prerecorded audio learning materials in the native language (Hindi) and were expected to listen to 3 sessions via feature phones and discuss as a small group discussion. The duration of a single session was 15 minutes. Sessions were scheduled in the afternoons with flexibility to reschedule. Group facilitators (ASHA with a smartphone) played a key role in leading discussions and promoting engagement. • Outcome: the intervention resulted in significant knowledge gains and ASHAs reported feeling empowered through peer discussions and knowledge sharing.

Training Outcomes

Fourteen studies reported learning outcomes (Table 4). The learning outcomes reported in the studies were knowledge, attitude, and behaviors. Assessment of knowledge was made mainly through pretraining and posttraining tests (n=13) [37-40,42,44-48,50-53]. All 13 studies reported improved knowledge outcomes after the training. Two studies reported improved attitude. One study from India that implemented both synchronous and asynchronous learning approaches reported enhanced attitude (confidence) in delivering maternal mental health services [39] and another from Rwanda demonstrated asynchronous learning improvements in providing support to patients with mental health conditions in the community [50]. One study from Bangladesh reported that asynchronous learning improved behavior in counseling couples on all available contraceptive options and child spacing [40].

The reported outcomes varied across the studies. Most studies presented one outcome from one category. Table 5 summarizes the available literature, notes gaps of digital training outcomes where little or no research has been conducted, and highlights areas for future research. Ten studies reported implementation outcomes [36,37,41-44,49-53]. Seven studies used synchronous learning [41-43,49,51-53], 3 asynchronous learning [37,44,50] and 2 studies implemented both approaches [36,39]. The implementation outcomes reported were acceptability, appropriateness, feasibility, and fidelity. Acceptability included perceived effectiveness, perception of remote training, usability, and satisfaction (n=10) [36,39,41-44,49-52]. Acceptability was assessed posttraining mainly through questionnaires, written qualitative feedback, or in-depth interviews. Three studies reported that participants were satisfied with the training [39,47,50] and the other 7 reported that the training was agreeable. Two studies reported on appropriateness and found the training suitable for capacity building for CHWs [41,50].

XSL•FO
RenderX

Table . Digital and blended learning training, implementation, and clinical outcomes.

Study	Analytic approach	Knowledge	Attitude	Behaviors	Acceptability	Appropriateness	Feasibility	Fidelity	Effectiveness
Bertman et al [36]	Qualitative	No	No	No	Yes	No	No	No	No
Khan et al [37]	Quantitative	Yes	No	No	No	No	Yes	No	No
Kharel et al [38]	Quantitative	Yes	No	No	No	No	No	No	No
Lakshminarayanan et al [39]	Quantitative	Yes	Somewhat	No	Somewhat	No	Somewhat	No	No
Limaye et al [40]	Quantitative	Yes	No	Yes	No	No	No	No	No
Muke et al [41]	Qualitative	No	No	No	Yes	Yes	Yes	No	No
Muke et al [42]	Mixed	Yes	No	No	Yes	No	Yes	No	No
Nedungadi et al [43]	Mixed	No	No	No	Yes	No	No	No	No
O'Donovan et al [44]	Quantitative	Yes	No	No	Yes	No	No	No	No
Rahman et al [45]	Quantitative	Yes	No	No	No	No	No	No	No
Sangwa et al [46]	Quantitative	Yes	No	No	No	No	No	No	No
Sranacharoenpong et al [47]	Quantitative	Yes	No	No	No	No	No	No	No
Sranacharoenpong and Hanning [48]	Quantitative	Yes	No	No	No	No	No	No	No
Tembo et al [49]	Quantitative	No	No	No	Yes	No	No	Yes	Yes
Willems et al [50]	Mixed	Yes	Yes	No	Yes	Yes	No	Yes	No
Yadav et al [51]	Mixed	Yes	No	No	Yes	No	Yes	No	No
Yadav [52]	Mixed	Yes	No	No	Yes	No	Yes	No	No
Yadav et al [53]	Mixed	Yes	No	No	No	No	Yes	No	No

Less than half of the studies reported on the feasibility of participating in digital training, and all found the training was successfully carried out (feasible) [37,39,41,42,51-53]. Feasibility assessment mostly involved posttraining surveys or qualitative interviews or feedback or reviewing training logs. This included responses to questions about digital device familiarity [41] to the review of log-in attempts and the number of hours spent in the training system [51]. Some studies reported technical challenges that affected training feasibility. Three studies reported poor internet connectivity challenges, including lag or slow internet [38,41,42]. Four studies reported poor cellular network infrastructure [36,51-53]. One mentioned other technical glitches [39] and one discussed power cuts hindering charging tablets [36].

One study from Rwanda reported high fidelity to the asynchronous training schedule [50]. Another study from Malawi found synchronous learning improved fidelity in providing counseling services to patients as intended based on a CHW checklist [49]. This study also reported improved effectiveness on index case testing indicators, such as sexual contacts elicited [49]. This was the only clinical outcome. Full details of the outcomes for individual studies are summarized in Table 3.

XSL•FO
RenderX

Table . Training outcomes of the included studies.

Study	Learning outcomes			Implementation outcomes							Clinical outcomes				
	Knowledge	Attitudes	Skills	Behavior	Acceptability	Feasibility	Fidelity	Appropriateness	Sustainability	Uptake	Effectiveness	Efficiency	Safety	Equity	Patient satisfaction
Berman et al [36]					✓										
Sinchaipong et al [47]	✓														
Sinchaipong and Hanning [48]	✓														
Tembo et al [49]					✓		✓				✓				
Khan et al [37]	✓					✓									
Kharel et al [38]	✓														
Lakshmi-narayanan et al [39][39]	✓														
Limaye et al [40]	✓			✓											
Muke et al [41]					✓	✓		✓							
Muke et al [42][42]	✓				✓	✓									
Nedungadi et al [43]					✓										
O'Donovan et al [44]	✓				✓										
Rahman et al [45]	✓														
Sangwa et al [46]	✓														
Willems et al [50]	✓	✓			✓	✓	✓								

XSL•FO
RenderX

Study	Learning outcomes			Implementation outcomes							Clinical outcomes				
	Knowledge	Attitudes	Skills	Behavior	Acceptability	Feasibility	Fidelity	Appropriateness	Sustainability	Uptake	Effectiveness	Efficiency	Safety	Equity	Patient satisfaction
Yadav et al [51]	✓				✓	✓									
Yadav [52]	✓				✓	✓									
Yadav et al [53]	✓					✓									

Discussion

Principal Findings

This scoping review brings together studies and reports of the characteristics and outcomes of digital training for CHWs in LMICs. Despite a rigorous search across multiple databases and broad inclusion criteria, we identified only 18 studies, which highlights the dearth of robust evaluations of digital training in LMICs. Yet, the review indicated that despite technological and infrastructural challenges, training CHWs using digital technologies was acceptable and feasible and led to knowledge acquisition. This body of evidence was largely descriptive, with few studies using experimental or quasi-experimental designs. Most reported only on learning outcomes, with very few reporting on implementation fidelity or clinical outcomes as a result of CHW training. Additionally, the majority of studies in this review only measured Kirkpatrick [25] Levels 1 and 2. A few measured Level 3 behaviors, and only one measured Level 4 results. This highlights the need for additional research evaluating additional level 3 and 4 outcomes.

This is the first review of digital training among CHWs in LMIC. Other reviews have focused on digital training primarily for professional health workers, mostly from HICs [15,16]. These reviews have similarly found relatively few studies, mostly descriptive in nature, with only knowledge outcomes evaluated. A scoping review that focused on digital training for rural professional cadres [16] identified few nonexperimental studies (n=5) and focused mainly on knowledge acquisition. Similarly, a systematic review that evaluated digital learning for medical education in LMICs [54] found two-thirds of the studies identified were from upper-middle-income countries, and only 4 were RCTs. Additionally, a review of digital learning for professional HCWs [15] found a few studies (n=14) conducted in LMICs; only 7 were RCTs and mainly reported knowledge outcomes.

Most of the evidence gathered in our scoping review showed that the digital or blended learning training resulted in significantly improved knowledge of CHWs after the training, compared to before, and is consistent with other learning modalities [55,56]. Similarly, our results are consistent with other studies in medical education reporting on the effectiveness of digital or blended learning in enhancing knowledge acquisition among certified health professionals in both LMICs [57,58] and HICs [59-61]. This supports the finding that digital training accounts for the knowledge needed to perform tasks across disease entities and health cadres in both high- and lower-economic settings. Furthermore, a systematic review comparing the effectiveness of blended learning to traditional learning for professional or certified HCWs [17] concluded that blended learning demonstrated consistently positive effects on knowledge outcomes when compared with traditional learning in medical education. Our work extends these findings to CHWs' knowledge to perform delegated assignments.

A few studies included in this scoping review addressed acceptability and appropriateness [41,50]. Feedback from CHWs suggested that they found digital or blended training acceptable for learning new information and appropriate as a mode of training. This has also been reported in other studies where digital training was acceptable among CHWs in diverse settings for data collection training [62]. Acceptability and appropriateness are important factors for training completion and skills application.

Overall, digital training was feasible, though approaches differed considerably. This is consistent with other studies with professional cadres [15,63]. For example, studies in Kenya [14] and Nigeria [57] found digital training for HCWs was feasible at scale. These studies highlight the potential feasibility of implementing digital training interventions in LMICs. Digital training programs in this review included the use of various technologies, including tablets, mobile phones, or feature phones. Some training approaches did not require internet or physical infrastructure, which may be lacking or unreliable in low-resource settings. For example, CHWs could access the training on a feature phone in the community without needing internet access or being in centralized locations [50]. The use of mobile devices made it possible for synchronous or asynchronous learning models to occur. Both models appear to be of benefit. Asynchronous learning offered CHWs the opportunity to participate in the training in a flexible manner at a time and place that suited them without missing any material due to other demanding personal or work-related responsibilities. Synchronous learning offered CHWs the opportunity to participate in group discussions, ask questions in real-time, and practice and receive feedback. Furthermore, synchronous sessions provided the trainees with access to technical support periodically. With the increased use of smartphones in LMICs, digital training could successfully be implemented in LMICs [64].

XSL•FO
RenderX

This review also underlined that using digital training programs in LMICs will require technological challenges to be overcome. Technical challenges, including slow or poor internet and cellular networks, were frequently cited as barriers to training completion. For example, Muke et al [42] reported that due to the challenge of poor internet connectivity, some participants were unable to access all content from the training modules because videos did not play. Preloading training material could eliminate reliance on connectivity and may ensure availability offline [65]. In addition, using feature phones demonstrated the possibility of mitigating network challenges [50]. The design of the digital training program is imperative. Simpler platforms may be easier to use and scale, and good accessibility may provide efficiencies that enable greater impact of the training.

Policy and Practice Recommendations

Given the global shortage of professional health workers, recruiting and training CHWs is one solution to overcome this gap in LMICs. The integration of digital training programs into national health strategies could play a pivotal role in achieving high-quality health care outcomes. LMICs should consider investing in digital infrastructure, including technology devices and literacy for lay cadres, to enhance accessibility to training and scalability of digital solutions. Tailoring and co-designing digital training programs with CHWs to ensure relevance and practicality will encourage engagement and sustainability of the digital tools. These recommendations support the WHO's suggestion of using digital training to supplement health workers' continuous capacity-building efforts [66].

Limitations

Our review had important limitations. Some studies did not provide enough details on training characteristics and outcomes. There was a potential to give greater emphasis to the studies that provided more detailed information when using narrative synthesis in data analysis. To overcome this, we contacted all authors for more information and were able to obtain some additional data that were not published in the articles. Additionally, we extracted data according to a predesigned structure and had 2 authors (TAT and FA) extract data independently.

Conclusion

CHWs play an essential role in providing health care and improving health outcomes in LMIC environments. Training must continue to be a core component of improving CHW skills. In LMIC environments that often have lower technological literacy and infrastructure, we sought to understand what role digital training could play. Digital trainings had a broad range of attributes. Few evaluations of digital training for CHWs in LMICs were identified in this review, in spite of a proliferation of such trainings. Furthermore, most evaluations lacked experimental designs. The existing evidence suggests that digital training can impact knowledge in lay health workers in LMICs; more rigorous studies with a broader range of outcomes are needed. Such additional work is needed to ensure lay health workers are well-capacitated to deliver critical health services.

Acknowledgments

We acknowledge and thank Tingathe Outreach Programme, and Baylor College of Medicine Children's Foundation Malawi. We would also like to thank Dr Diane Nguyen for the orientation on conducting reviews. Generative artificial intelligence was not used in the research and writing process.

Funding

No external funding for the review. TAT was supported by the Fogarty International Center of the National Institutes of Health under Award Number D43 TW010060. NER was supported by R01 MH124526.

Data Availability

The data supporting the findings of this scoping review are available within the article and/or its supplementary materials.

Authors' Contributions

TAT, NER, and LGB conceptualized the study. TAT and FA performed the data extraction. TAT analyzed the data and interpreted the results under the guidance of NER and LGB. TAT drafted the manuscript and all subsequent drafts. All authors critically reviewed, edited, and approved the final manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Database search terms and results.
[DOCX File, 30 KB - mededu_v12i1e82772_app1.docx]

Checklist 1
PRISMA checklist.

XSL•FO
RenderX

[DOCX File, 27 KB - mededu_v12i1e82772_app2.docx]

References

1. Darzi A, Evans T. The global shortage of health workers-an opportunity to transform care. Lancet 2016 Nov 26;388(10060):2576-2577. [doi: 10.1016/S0140-6736(16)32235-8] [Medline: 27894651]
2. Perry HB, Zulliger R, Rogers MM. Community health workers in low-, middle-, and high-income countries: an overview of their history, recent evolution, and current effectiveness. Annu Rev Public Health 2014;35:399-421. [doi: 10.1146/annurev-publhealth-032013-182354] [Medline: 24387091]
3. Lewin SA, Dick J, Pond P, et al. Lay health workers in primary and community health care. Cochrane Database Syst Rev 2005 Jan 25;2005(1):CD004015. [doi: 10.1002/14651858.CD004015.pub2] [Medline: 15674924]
4. Winters N, Langer L, Geniets A. Scoping review assessing the evidence used to support the adoption of mobile health (mHealth) technologies for the education and training of community health workers (CHWs) in low-income and middle-income countries. BMJ Open 2018 Jul 30;8(7):e019827. [doi: 10.1136/bmjopen-2017-019827] [Medline: 30061430]
5. O'Donovan J, O'Donovan C, Kuhn I, Sachs SE, Winters N. Ongoing training of community health workers in low-income andmiddle-income countries: a systematic scoping review of the literature. BMJ Open 2018 Apr 28;8(4):e021467. [doi: 10.1136/bmjopen-2017-021467] [Medline: 29705769]
6. Winters N, Langer L, Nduku P, et al. Using mobile technologies to support the training of community health workers in low-income and middle-income countries: mapping the evidence. BMJ Glob Health 2019;4(4):e001421. [doi: 10.1136/bmjgh-2019-001421] [Medline: 31413872]
7. Zulu JM, Silumbwe A, Munakampe M, et al. A scoping review of the roles, challenges, and strategies for enhancing the performance of community health workers in the response against COVID-19 in low- and middle-income countries. BMC Prim Care 2025 May 14;26(1):163. [doi: 10.1186/s12875-025-02853-7] [Medline: 40369415]
8. Kim MH, Ahmed S, Buck WC, et al. The Tingathe programme: a pilot intervention using community health workers to create a continuum of care in the prevention of mother to child transmission of HIV (PMTCT) cascade of services in Malawi. J Int AIDS Soc 2012 Jul;15(S2):17389. [doi: 10.7448/IAS.15.4.17389] [Medline: 22789644]
9. Lehmann U, Sanders D. Policy brief: community health workers: what do we know about them? the state of the evidence on programmes, activities, costs and impact on health outcomes of using community health workers. evidence and information for policy. : World Health Organization; 2007.
10. Sultan MA, Miller E, Tikkanen RS, et al. Competency-based education and training for community health workers: a scoping review. BMC Health Serv Res 2025 Feb 17;25(1):263. [doi: 10.1186/s12913-025-12217-7] [Medline: 39962470]
11. Tumlinson K, Jaff D, Stilwell B, Onyango DO, Leonard KL. Reforming medical education admission and training in low- and middle-income countries: who gets admitted and why it matters. Hum Resour Health 2019 Dec 2;17(1):91. [doi: 10.1186/s12960-019-0426-9] [Medline: 31791358]
12. Hippe DS, Umoren RA, McGee A, Bucher SL, Bresnahan BW. A targeted systematic review of cost analyses for implementation of simulation-based education in healthcare. SAGE Open Med 2020;8:2050312120913451. [doi: 10.1177/2050312120913451] [Medline: 32231781]
13. Millimouno TM, Delamou A, Kourouma K, et al. Outcomes of blended learning for capacity strengthening of health professionals in Guinea. BMC Med Educ 2021 Jul 28;21(1):406. [doi: 10.1186/s12909-021-02847-w] [Medline: 34320967]
14. Allott H, Smith A, White S, et al. Improving capacity for advanced training in obstetric surgery: evaluation of a blended learning approach. BMC Med Educ 2025 Jan 17;25(1):80. [doi: 10.1186/s12909-025-06660-7] [Medline: 39819576]
15. Aryee GFB, Amoadu M, Obeng P, et al. Effectiveness of eLearning programme for capacity building of healthcare professionals: a systematic review. Hum Resour Health 2024 Sep 2;22(1):60. [doi: 10.1186/s12960-024-00924-x] [Medline: 39223555]
16. Woods L, Martin P, Khor J, Guthrie L, Sullivan C. The right care in the right place: a scoping review of digital health education and training for rural healthcare workers. BMC Health Serv Res 2024 Sep 2;24(1):1011. [doi: 10.1186/s12913-024-11313-4] [Medline: 39223581]
17. Vallée A, Blacher J, Cariou A, Sorbets E. Blended learning compared to traditional learning in medical education: systematic review and meta-analysis. J Med Internet Res 2020 Aug 10;22(8):e16504. [doi: 10.2196/16504] [Medline: 32773378]
18. Wang W, Zhang HB, Liu JM, et al. Variations, effectiveness and its associated factors of a nationwide web-based hypertension management training project in China: insights from a government-led campaign for 1.2 million lay health workers. J Geriatr Cardiol 2024 Jul 28;21(7):733-750. [doi: 10.26599/1671-5411.2024.07.010] [Medline: 39183951]
19. Ladur AN, Egere U, Ravit M, et al. A blended learning approach for capacity strengthening to improve the quality of integrated HIV, TB, and malaria services during antenatal and postnatal care in LMICs: a feasibility study. BMC Med Educ 2025 Jan 8;25(1):35. [doi: 10.1186/s12909-024-06633-2] [Medline: 39780210]
20. Boutros P, Kassem N, Nieder J, et al. Education and training adaptations for health workers during the COVID-19 pandemic: a scoping review of lessons learned and innovations. Healthcare (Basel) 2023 Nov 4;11(21):2902. [doi: 10.3390/healthcare11212902] [Medline: 37958046]

XSL•FO
RenderX

21. Hodgins S, Kok M, Musoke D, et al. Community health workers at the dawn of a new era: 1. Introduction: tensions confronting large-scale CHW programmes. Health Res Policy Syst 2021 Oct 12;19(Suppl 3):109. [doi: 10.1186/s12961-021-00752-8] [Medline: 34641886]
22. Health for the people: national community health worker programs from afghanistan to zimbabwe. ChwcentralOrg. 2021. URL: https://chwcentral.org/wp-content/uploads/2021/11/Health_for_the_People_Natl_Case%20Studies_Oct2021.pdf [accessed 2026-04-28]
23. Yan Liu RA, Stemmler H. Digital adoption: accelerating postpandemic, yet a widening divide. : Digital Progress and Trends Report; 2023 URL: https://openknowledge.worldbank.org/server/api/core/bitstreams/b6125358-fb1c-47bb-a6e7-4896b2152904/content [accessed 2026-04-28]
24. Development data group income level/low- and middle-income countries. World Bank. 2023. URL: https://data.worldbank.org/income-level/low-and-middle-income [accessed 2026-04-28]
25. Yardley S, Dornan T. Kirkpatrick's levels and education "evidence". Med Educ 2012 Jan;46(1):97-106. [doi: 10.1111/j.1365-2923.2011.04076.x] [Medline: 22150201]
26. Tricco AC, Lillie E, Zarin W, et al. PRISMA extension for scoping reviews (PRISMA-ScR): checklist and explanation. Ann Intern Med 2018 Oct 2;169(7):467-473. [doi: 10.7326/M18-0850] [Medline: 30178033]
27. Arksey H, O'Malley L. Scoping studies: towards a methodological framework. Int J Soc Res Methodol 2005 Feb;8(1):19-32. [doi: 10.1080/1364557032000119616]
28. Levac D, Colquhoun H, O'Brien KK. Scoping studies: advancing the methodology. Implement Sci 2010 Sep 20;5(1):69. [doi: 10.1186/1748-5908-5-69] [Medline: 20854677]
29. Rethlefsen ML, Kirtley S, Waffenschmidt S, et al. PRISMA-S: an extension to the PRISMA statement for reporting literature searches in systematic reviews. Syst Rev 2021 Jan 26;10(1):39. [doi: 10.1186/s13643-020-01542-z] [Medline: 33499930]
30. World bank country and lending groups. World Bank. 2022. URL: https://datahelpdesk.worldbank.org/knowledgebase/articles/906519-world-bank-country-and-lending-groups [accessed 2026-04-28]
31. Car J, Carlstedt-Duke J, Tudor Car L, et al. Digital education in health professions: the need for overarching evidence synthesis. J Med Internet Res 2019 Feb 14;21(2):e12913. [doi: 10.2196/12913] [Medline: 30762583]
32. Martinengo L, Yeo NJY, Tang ZQ, Markandran KD, Kyaw BM, Tudor Car L. Digital education for the management of chronic wounds in health care professionals: protocol for a systematic review by the digital health education collaboration. JMIR Res Protoc 2019 Mar 25;8(3):e12488. [doi: 10.2196/12488] [Medline: 30907743]
33. Kim KJ, Bonk CJ, Oh E. The present and future state of blended learning in workplace learning settings in the United States. Perf Improv 2008 Sep;47(8):5-16. [doi: 10.1002/pfi.20018]
34. Proctor E, Silmere H, Raghavan R, et al. Outcomes for implementation research: conceptual distinctions, measurement challenges, and research agenda. Adm Policy Ment Health 2011 Mar;38(2):65-76. [doi: 10.1007/s10488-010-0319-7] [Medline: 20957426]
35. Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics 1977 Mar;33(1):159-174. [doi: 10.2307/2529310] [Medline: 843571]
36. Bertman V, Petracca F, Makunike-Chikwinya B, et al. Health worker text messaging for blended learning, peer support, and mentoring in pediatric and adolescent HIV/AIDS care: a case study in Zimbabwe. Hum Resour Health 2019 Jun 7;17(1):41. [doi: 10.1186/s12960-019-0364-6] [Medline: 31174543]
37. Khan A, Sebok-Syer SS, Linstadt H, et al. An electronic-based curriculum to train acute care providers in rural Haiti and India. J Grad Med Educ 2019 Aug;11(4 Suppl):152-157. [doi: 10.4300/JGME-D-18-01019] [Medline: 31428273]
38. Kharel R, Regmi SP, Lin T, Levine AC, Aluisio AR. Training program for female community volunteers to combat COVID 19 in rural Nepal. Glob Health Action 2022 Dec 31;15(1):2134425. [doi: 10.1080/16549716.2022.2134425] [Medline: 36369910]
39. Lakshminarayanan M, Kathuria N, Mehra S. Delivery of perinatal mental health services by training lay counselors using digital platforms. Asian J Psychiatr 2020 Dec;54:102277. [doi: 10.1016/j.ajp.2020.102277] [Medline: 32682298]
40. J Limaye R, Ballard Sara A, Ahmed N, et al. Enhancing the knowledge and behaviors of fieldworkers to promote family planning and maternal, newborn, and child health in Bangladesh through a digital health training package: results from a pilot study. Int Q Community Health Educ 2020 Jan;40(2):143-149. [doi: 10.1177/0272684X19861866] [Medline: 31274369]
41. Muke SS, Shrivastava RD, Mitchell L, et al. Acceptability and feasibility of digital technology for training community health workers to deliver brief psychological treatment for depression in rural India. Asian J Psychiatr 2019 Oct;45:99-106. [doi: 10.1016/j.ajp.2019.09.006] [Medline: 31557716]
42. Muke SS, Tugnawat D, Joshi U, et al. Digital training for non-specialist health workers to deliver a brief psychological treatment for depression in primary care in India: findings from a randomized pilot study. Int J Environ Res Public Health 2020 Sep 1;17(17):6368. [doi: 10.3390/ijerph17176368] [Medline: 32883018]
43. Nedungadi P, Jinachandran R, Mohan A, Raman R. Rural health in digital india: interactive simulations for community health workers. Presented at: 2019 IEEE Tenth International Conference on Technology for Education (T4E); Jul 26-31, 2019; Goa, India p. 86-89. [doi: 10.1109/T4E.2019.00-45]

XSL•FO
RenderX

44. O'Donovan J, Kabali K, Taylor C, et al. The use of low-cost android tablets to train community health workers in Mukono, Uganda, in the recognition, treatment and prevention of pneumonia in children under five: a pilot randomised controlled trial. Hum Resour Health 2018 Sep 19;16(1):49. [doi: 10.1186/s12960-018-0315-7] [Medline: 30231894]
45. Rahman A, Akhtar P, Hamdani SU, et al. Using technology to scale-up training and supervision of community health workers in the psychosocial management of perinatal depression: a non-inferiority, randomized controlled trial. Glob Ment Health 2019;6:e8. [doi: 10.1017/gmh.2019.7]
46. Sangwa Y, Ndaruhutse V, Radeny S, et al. Using eLearning to improve and retain the knowledge of community health workers in maternal and neonatal health in Rwanda: a cohort study. Public Health Chall 2024 Jun;3(2):e174. [doi: 10.1002/puh2.174] [Medline: 40496154]
47. Sranacharoenpong K, Hanning RM, Sirichakwal PP, Chittchang U. Process and outcome evaluation of a diabetes prevention education program for community healthcare workers in Thailand. Educ Health (Abingdon) 2009 Dec;22(3):335. [Medline: 20029769]
48. Sranacharoenpong K, Hanning RM. Diabetes prevention education program for community health care workers in Thailand. J Community Health 2012 Jun;37(3):610-618. [doi: 10.1007/s10900-011-9491-2] [Medline: 21971628]
49. Tembo TA, Simon KR, Kim MH, et al. Pilot-testing a blended learning package for health care workers to improve index testing services in Southern Malawi: an implementation science study. J Acquir Immune Defic Syndr 2021 Dec 15;88(5):470-476. [doi: 10.1097/QAI.0000000000002796] [Medline: 34483296]
50. Willems A, Iyamuremye JD, Misage CN, Smith-Swintosky V, Kayiteshonga Y. Co-creation and evaluation of nationwide remote training service for mental health education of community health workers in Rwanda. Front Public Health 2021;9:632793. [doi: 10.3389/fpubh.2021.632793] [Medline: 34504821]
51. Yadav D, Singh P, Montague K, Kumar V, Sood D, Balaam M, et al. Sangoshthi: empowering community health workers through peer learning in rural india. Presented at: Conference on Human Factors in Computing Systems 2017; May 6-11, 2017. [doi: 10.1145/3038912.3052624]
52. Yadav D. Low-Cost Mobile Learning Solutions for Community Health Workers. 2017 Presented at: the 26th International Conference; Apr 3-7, 2017; Perth, Australia p. 729-734 URL: http://dl.acm.org/citation.cfm?doid=3041021 [accessed 2026-04-28] [doi: 10.1145/3041021.3053377]
53. Yadav D, Bhandari A, Singh P. L. Scaffolding collaborative learning of community health workers in india. Proc ACM Hum Comput Interact 2019:1-27. [doi: 10.1145/3359271]
54. Barteit S, Guzek D, Jahn A, Bärnighausen T, Jorge MM, Neuhann F. Evaluation of e-learning for medical education in low- and middle-income countries: a systematic review. Comput Educ 2020 Feb;145:103726. [doi: 10.1016/j.compedu.2019.103726] [Medline: 32565611]
55. Holt CL, Tagai EK, Santos SLZ, et al. Web-based versus in-person methods for training lay community health advisors to implement health promotion workshops: participant outcomes from a cluster-randomized trial. Transl Behav Med 2019 Jul 16;9(4):573-582. [doi: 10.1093/tbm/iby065] [Medline: 29955889]
56. Abdel-All M, Putica B, Praveen D, Abimbola S, Joshi R. Effectiveness of community health worker training programmes for cardiovascular disease management in low-income and middle-income countries: a systematic review. BMJ Open 2017 Nov 3;7(11):e015529. [doi: 10.1136/bmjopen-2016-015529] [Medline: 29101131]
57. Hicks JP, Allsop MJ, Akaba GO, et al. Acceptability and potential effectiveness of eHealth tools for training primary health workers from Nigeria at scale: mixed methods, uncontrolled before-and-after study. JMIR Mhealth Uhealth 2021 Sep 16;9(9):e24182. [doi: 10.2196/24182] [Medline: 34528891]
58. Little KM, Nwala AA, Demise E, et al. Use of a hybrid digital training approach for hormonal IUD providers in Nigeria: results from a mixed method study. BMC Health Serv Res 2023 Nov 29;23(1):1316. [doi: 10.1186/s12913-023-10211-5] [Medline: 38031098]
59. O'Connell J, Shafran R, Pote H. A randomized controlled trial evaluating the effectiveness of face-to-face and digital training in improving child mental health literacy rates in frontline pediatric hospital staff. Front Psychiatry 2020;11:570125. [doi: 10.3389/fpsyt.2020.570125] [Medline: 33643077]
60. Strehlow MC, Johnston JS, Aluri KZ, et al. Evaluation of a massive open online course for just-in-time training of healthcare workers. Front Public Health 2024;12:1395931. [doi: 10.3389/fpubh.2024.1395931] [Medline: 39411488]
61. Hurley DA, Keogh A, Mc Ardle D, et al. Evaluation of an E-Learning training program to support implementation of a group-based, theory-driven, self-management intervention for osteoarthritis and low-back pain: pre-post study. J Med Internet Res 2019 Mar 7;21(3):e11123. [doi: 10.2196/11123] [Medline: 30843863]
62. Greuel M, Sy F, Bärnighausen T, et al. Community health worker use of smart devices for health promotion: scoping review. JMIR Mhealth Uhealth 2023 Feb 22;11:e42023. [doi: 10.2196/42023] [Medline: 36811947]
63. Clair V, Musau A, Mutiso V, et al. Blended-eLearning improves alcohol use care in Kenya: pragmatic randomized control trial results and parallel qualitative study implications. Int J Ment Health Addiction 2022 Dec;20(6):3410-3437. [doi: 10.1007/s11469-022-00841-x]
64. Long LA, Pariyo G, Kallander K. Digital technologies for health workforce development in low- and middle-income countries: a scoping review. Glob Health Sci Pract 2018 Oct 10;6(Suppl 1):S41-S48. [doi: 10.9745/GHSP-D-18-00167] [Medline: 30305338]

XSL•FO
RenderX

65. Reisach U, Weilemann M. Organisational aspects and benchmarking of e-learning initiatives: a case study with South African community health workers. Glob Health Promot 2016 Jun;23(2):57-66. [doi: 10.1177/1757975914567178] [Medline: 25733133]
66. World Health Organization. WHO Guideline: Recommendations on Digital Interventions for Health System Strengthening: World Health Organization; 2019:1-124.

Abbreviations

CHW: community health worker
CINAHL: Cumulated Index in Nursing and Allied Health Literature
ERIC: Education Resources Information Center
HCW: health care worker
HIC: high-income country
LMIC: low- and middle-income country
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
PRISMA-S: Preferred Reporting Items for Systematic Reviews and Meta-Analyses–Search Extension
PRISMA-ScR: Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews
PROSPERO: International Prospective Register of Systematic Reviews
RCT: randomized controlled trial
SSRN: Social Science Research Network
WHO: World Health Organization

Edited by S Brini; submitted 22.Aug.2025; peer-reviewed by RSG Mahmoud, S Ganesh, V Gopichandran; revised version received 12.Mar.2026; accepted 17.Mar.2026; published 19.May.2026.

Please cite as:
Tembo TA, Rosenberg NE, Ayele F, Ahmed S, Bekker LG
Outcomes of Digital Training for Community Health Workers in Low- and Middle-Income Countries: Scoping Review
JMIR Med Educ 2026;12:e82772
URL: https://mededu.jmir.org/2026/1/e82772
doi:10.2196/82772

© Tapiwa Tembo, Nora Ellen Rosenberg, Firaol Ayele, Saeed Ahmed, Linda-Gail Bekker. Originally published in JMIR Medical Education (https://mededu.jmir.org), 19.May.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Education, is properly cited. The complete bibliographic information, a link to the original publication on https://mededu.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Augmented Reality in Surgical Training: Systematic Review of Its Impact on Technical Performance in Surgical Trainees

Mahmoud El Ashry[1*], MBChB; Ahmed El Ashry[1*], MBChB, MSc; Hamza Khalique[1], MBChB; Yahye Abdalle[1], MBChB; Thomas Yeung[2], MBBCh

[1]Bristol Medical School, University of Bristol, Beacon House, Queens Road, Bristol, United Kingdom

[2]ENT/Otolaryngology Department, University Hospitals of North Midlands NHS Trust, Stoke-on-Trent, United Kingdom

[*]these authors contributed equally

Corresponding Author:
Mahmoud El Ashry, MBChB
Bristol Medical School, University of Bristol, Beacon House, Queens Road, Bristol, United Kingdom

Related Article:

This is a corrected version. See correction statement: https://mededu.jmir.org/2026/1/e106576

Abstract

Background: Surgical training has changed over the past decade. Augmented reality (AR) has become one of the more talked-about developments within that space. At its core, AR works by placing digital information over the real-world environment. This gives trainees guidance and spatial cues during a procedure as they perform it. What remains uncertain is whether AR moves the needle on technical skill development in trainees. The studies that address this directly are few, and the ones that do exist rarely speak to each other in any meaningful way. Outcome measures shift from paper to paper, the hardware studied spans a wide range of maturity, and methodological consistency is hard to find.

Objective: This systematic review assesses the impact of AR on the objective technical skills of surgical trainees when compared with traditional methods.

Methods: We searched PubMed, MEDLINE, Embase, IEEE Xplore, Scopus, and Web of Science for studies published between January 1, 2020, and September 15, 2025. From 4799 initial records, 1417 remained after deduplication. Of these, 101 underwent detailed abstract review and 29 were assessed in full text. Eleven studies met the inclusion criteria. Two reviewers (MEA and YA) independently screened all records, with a third senior reviewer (TY) resolving disagreements. We performed a narrative synthesis following SWiM (Synthesis Without Meta-Analysis) guidelines across 5 thematic domains to account for study heterogeneity.

Results: The final analysis included 11 studies (347 participants across 7 specialties) published between 2021 and 2025. These included 9 randomized controlled trials and 2 prospective cohort studies. The studies used platforms such as the Microsoft HoloLens (1 and 2), Magic Leap One, and Vuzix M300XL. Of the 11 studies, 9 reported improvements in one or more objective technical metrics. Key findings included consistent error reduction (5/5 studies), faster learning curves (4/11 studies), and lower cognitive workload (3/11 studies). Notably, an “expertise reversal” effect was observed, where AR provided substantial benefits to novices but diminishing returns for experienced surgeons.

Conclusions: AR significantly improves technical performance for surgical novices, particularly in tasks involving complex visuospatial reasoning. AR is an effective tool in surgical education. Future research should focus on multicenter trials to evaluate long-term skill retention and cost-effectiveness in clinical practice.

(JMIR Med Educ 2026;12:e71572) doi:10.2196/71572

KEYWORDS

augmented reality; surgical education; systematic review; head-mounted display; simulation; PRISMA; telestration; skill acquisition

Introduction

Surgical training has transformed over the past two decades. The erosion of the traditional apprenticeship model, driven by reduced working hours, patient safety concerns, and the expansion of minimally invasive techniques, has created an urgent need for high-fidelity, reproducible alternatives to case-based learning [1,2]. The consequences of this shift are well documented: surgical residents are entering independent practice with fewer operative experiences than their

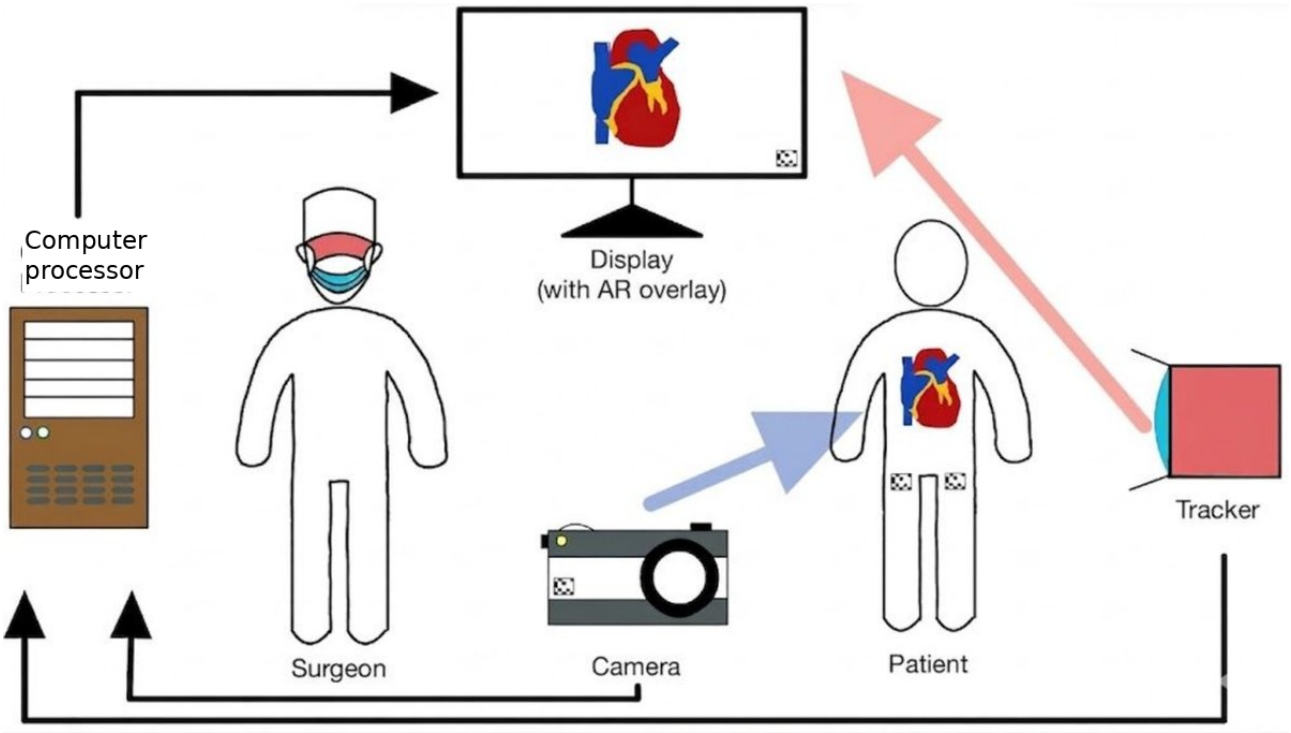
XSL•FO
RenderX

predecessors, and the early learning curve, particularly in complex and high-stakes procedures, carries measurable risk to patients [3,4]. Thus, simulation-based surgical education has emerged as a necessary complement to the operating room, and the last decade has witnessed extraordinary investment in the development and evaluation of digital training modalities [5,6].

Virtual reality (VR) and augmented reality (AR) have particularly attracted attention among users. Both AR and VR rely on immersive technology to create a virtual surgical learning environment for students. However, they differ fundamentally in their relationship to the physical world. VR relies solely on a synthetic environment, keeping the student separate from real-world stimuli, and has demonstrated efficacy in improving laparoscopic and robotic surgical skills across several procedural domains [7,8]. AR, by contrast, overlays computer-generated information, such as anatomical structures, navigational guides, procedural annotations, or expert telestration, directly onto the user's real-world field of view through optical head-mounted displays (HMDs), smart glasses, or screen-based systems [9,10].

This technical workflow, as illustrated in Figure 1, relies on a closed-loop system where physical data is captured, processed, and reprojected as a digital overlay in real time. These AR modalities maintain contact with physical instruments, simulators, and task environments while simultaneously providing contextually embedded guidance. The relative merits of VR and AR in surgical training remain an active area of investigation; the two modalities address different cognitive and technical challenges and are most productively viewed as complementary rather than competitive [11]. Mixed reality (MR), exemplified by platforms such as the Microsoft HoloLens, extends AR by enabling dynamic interaction between virtual and physical objects through spatial mapping, though the boundary between AR and MR in the surgical training literature remains inconsistently defined. While the boundary between AR and MR is often inconsistently defined in broader literature, this review strictly defines AR as the unidirectional overlay of digital data onto a physical field. We explicitly exclude bidirectional, spatially mapped "mixed reality" interactions to isolate the cognitive impact of the digital overlay itself.

Figure 1. Schematic diagram illustrating the basic principles of augmented reality as applied in a surgical context: a camera captures the physical operative environment; a computer processor generates a digital overlay (eg, anatomical structures, procedural guidance, or expert telestration); the overlay is projected onto a display visible to the surgeon, while a tracker provides positional feedback for accurate registration of virtual content to the physical world. AR: augmented reality.



The theoretical basis for AR's educational value is well established. Cognitive load theory posits that learners have finite working memory capacity and that effective instruction should minimize extraneous cognitive load while maximizing germane load, the mental effort directed toward schema formation [12]. Conventional verbal instruction in minimally invasive surgery is particularly demanding, requiring trainees to mentally translate auditory descriptions into spatial awareness of the operative field in real time. AR telestration and anatomical overlay systems bypass this cognitive translation step by providing visual guidance precisely colocated with the task being performed, theoretically reducing extraneous load and accelerating the formation of procedural schemas [13]. The expertise reversal effect predicts that while AR provides essential scaffolding for novices (maximizing germane load), these same overlays may function as extraneous cognitive load for experts who have already developed robust internal schemas, potentially hindering rather than helping performance [14].

Since 2020, published evidence for AR usage has expanded significantly, but despite this, existing systematic reviews have significant limitations: they frequently analyze AR, VR, and MR simultaneously or separate them with loose boundaries, and sometimes include older prototype systems. Studies also

XSL•FO
RenderX

rely predominantly on subjective or process-based outcomes rather than objective technical performance metrics [15,16]. No contemporary synthesis has focused exclusively on AR's impact on trainee technical performance using validated, objective outcome measures across the most recent generation of AR technology. This study aims to achieve this while also maintaining strict boundaries surrounding the inclusion of solely AR-based projects for review.

This systematic review addresses that gap. Using a prespecified PICO (population, intervention, comparison, outcome) framework: surgical trainees (population), AR-based training interventions (intervention), traditional surgical teaching methods (comparison), and objective measures of technical performance (outcome), we synthesize contemporary evidence from 2020 to 2025 to answer the question: does augmented reality, when used as an adjunct to or replacement for traditional surgical training, improve objective technical performance in surgical novices? By confining our scope to studies with defined AR interventions, controlled comparators, and objective outcome measures, this review provides a focused, methodologically rigorous assessment of AR's current standing as a surgical training modality.

Methods

Study Design and Registration

This systematic review included a literature search and a write-up that were both carried out with respect to the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) 2020 statement [17]. The completed PRISMA 2020 expanded checklist is provided as Checklist 1. A narrative synthesis was performed in accordance with the Synthesis Without Meta-Analysis (SWiM) reporting guideline [18]. This type of analysis was chosen as significant clinical and statistical heterogeneity existed across included studies making statistical pooling and group analysis of studies unrealistic. The approach is consistent with the methodological guidance of the *Cochrane Handbook for Systematic Reviews of Interventions* [19]. Searches were additionally reported in accordance with the PRISMA-S extension for the reporting of literature searches in systematic reviews [20]. The review was not prospectively registered. However, an a priori protocol specifying the research question, eligibility criteria, databases, search strategy, data extraction variables, and planned synthesis approach was developed and followed throughout.

Eligibility Criteria

Eligible studies for analysis had to be original, peer-reviewed articles published in English, with a publication window between January 1, 2020, and September 15, 2025. Participants could be trainees at any stage of their training. This meant if participants of studies were medical students, residents, and fellows, the study was appropriate for inclusion.

On the intervention side, a clearly described augmented reality component was required. For the purposes of this review, AR was understood as technology that overlays digital content directly onto a person's view of the real world. Since AR, VR, and MR tend to appear together frequently in the literature, screening by YA and MEA was agreed beforehand to be meticulous and deliberate in only including studies that clearly focused on independently analyzing AR only. Each study also needed to include a comparator condition, whether that entailed traditional instruction, conventional operative guidance, or a freehand approach. At least one objective measure of technical performance or skill acquisition had to be reported (Table 1).

Table . Inclusion and exclusion criteria for study eligibility.

Category	Inclusion criteria	Exclusion criteria
Publication	Peer-reviewed original research; published after January 1, 2020; English language	Reviews, editorials, letters, conference abstracts; published before 2020; non-English
Participants	Human surgical or procedural trainees (medical students, residents, and fellows) at any training stage	Expert surgeons performing clinical procedures without a trainee component; nonmedical participants
Intervention	Clearly described AR[a] component overlaying digital information onto real-world view; outcomes attributable to AR isolable from VR[b]/MR[c]	AR inseparable from VR or MR; purely passive AR use without guidance or training function; technology validation without trainee performance outcomes
Comparator	Traditional instruction, verbal guidance, conventional teaching, or freehand technique	No comparator or control condition present
Outcomes	At least one objective measure of technical performance (eg, accuracy, error count, validated skill score, and procedure time)	Outcomes entirely subjective (questionnaire only); usability/feasibility data only; no measurable performance data reported

[a]AR: augmented reality.

[b]VR: virtual reality.

[c]MR: mixed reality.

Several categories of studies were excluded from the outset. Work published before 2020 was not considered, nor were non-original outputs such as reviews, editorials, or conference abstracts. Studies conducted exclusively with expert surgeons performing live clinical procedures fell outside the scope unless a formal training element was present. Where AR could not be disentangled from VR or MR, or where all reported outcomes

XSL•FO
RenderX

were subjective in nature, studies were similarly excluded—as were those lacking any comparator.

The 2020 cutoff was a deliberate methodological choice rather than an arbitrary date. AR hardware and software underwent substantial development around 2018 - 2019, and a good deal of earlier work was conducted using prototype or near-prototype systems that bear little resemblance to the tools in use today. Including that literature risked drawing conclusions that would not generalize meaningfully to contemporary training contexts, so it was excluded on those grounds. The 2020 cutoff marks the transition from prototype-based research to the use of enterprise-grade, high-fidelity hardware (eg, Microsoft HoloLens 2 and Magic Leap One). Including earlier data from low-resolution prototype systems would introduce technological bias and yield conclusions that do not generalize to contemporary surgical training environments.

Information Sources

Six electronic databases were searched: PubMed (MEDLINE), Ovid MEDLINE, Embase, IEEE Xplore, Scopus, and Web of Science. The selection was deliberate rather than exhaustive for its own sake. AR in surgical training sits at the intersection of clinical medicine and engineering, and no single database captures that breadth adequately—so the combination was chosen to reflect it. PubMed, Ovid MEDLINE, and Embase covered the biomedical literature, with Embase included specifically because its indexing patterns differ enough from PubMed to reduce the risk of missing relevant work, consistent with *Cochrane Handbook* recommendations on database selection [19]. IEEE Xplore addressed the engineering and technology side of the literature, where much of the platform development and human-computer interaction research is published. Scopus and Web of Science brought broader multidisciplinary coverage across both domains.

Google Scholar was not included in the primary database search. Reasons for this included the search base's nontransparent indexing algorithm, inclusion of non–peer-reviewed sources, and absence of a reproducible search interface. These issues would result in the search falling short of the reproducibility standards required by PRISMA and the *Cochrane Handbook*, which we aimed for [19,21]. The search across all 6 databases was executed on April 24, 2026, with results filtered to the eligibility window of January 1, 2020, through September 15, 2025. Reference lists of relevant systematic reviews identified during screening were then hand-searched to capture any eligible work that database searching alone might have missed.

Search Strategy

Our search strings were drafted by combining Medical Subject Headings (MeSH) terms with free-text keywords, referring to the PICO framework. We also reflected on guidance from Chapter 4 of the *Cochrane Handbook* [19]. There were 3 domains in particular that formed the backbone of the search strategy: the technology itself (augmented reality), the population and setting (surgical or procedural trainees), and the outcome domain (technical performance or skill acquisition). Boolean operators (AND, OR) were then used to combine terms within search strings. Truncation with wildcards was applied where database syntax allowed. The full PubMed search string is provided in Textbox 1.

Textbox 1. Full PubMed search string.

("Augmented Reality"[MeSH] OR "augmented reality"[tiab] OR "mixed reality"[tiab] OR "head-mounted display"[tiab] OR "HMD"[tiab] OR "heads-up display"[tiab] OR "HUD"[tiab] OR "holographic"[tiab] OR "telestration"[tiab] OR "HoloLens"[tiab] OR "Magic Leap"[tiab] OR "smart glasses"[tiab] OR "optical see-through"[tiab] OR "AR-assisted"[tiab] OR "AR-guided"[tiab] OR "AR-enhanced"[tiab]) AND ("Education, Medical"[MeSH] OR "Education, Medical, Graduate"[MeSH] OR "Clinical Competence"[MeSH] OR "Simulation Training"[MeSH] OR "Internship and Residency"[MeSH] OR "surgical train*"[tiab] OR "surgical educat*"[tiab] OR "surgical skill*"[tiab] OR "surgical simulat*"[tiab] OR "procedural train*"[tiab] OR "procedural skill*"[tiab] OR "resident*"[tiab] OR "novice*"[tiab] OR "trainee*"[tiab] OR "medical student*"[tiab] OR "laparoscopic train*"[tiab] OR "minimally invasive train*"[tiab] OR "neurosurgery train*"[tiab]) AND ("learning curve"[tiab] OR "technical performance"[tiab] OR "skill acquisition"[tiab] OR "psychomotor"[tiab] OR "accuracy"[tiab] OR "proficiency"[tiab] OR "competency"[tiab] OR "OSATS"[tiab] OR "GOALS"[tiab] OR "error rate"[tiab] OR "performance score"[tiab] OR "procedure time"[tiab]) AND ("2020/01/01"[PDat]:"2025/09/15"[PDat])

This search strategy was adapted for the syntax and controlled vocabulary of each database. For IEEE Xplore, MeSH terms were replaced with IEEE Thesaurus terms if appropriate. For Scopus and Web of Science, equivalent field tags (TITLE-ABS-KEY) were used with the same conceptual terms. For Ovid MEDLINE and Embase, the Ovid MeSH explode function was used to capture all relevant subheadings. Full search strategies for all 6 databases are provided in Multimedia Appendix 1, reported in line with the PRISMA-S guideline [20].

A data restriction of January 1, 2020, to September 15, 2025, was applied across all databases. No language filter was applied at the search stage itself. We felt that by restricting the language at that point, we risked inadvertently suppressing relevant records before they could be assessed. Non-English records that did not come through were excluded at screening. Similarly, no publication type or study design filters were applied during the initial search. This was on the basis that this runs the risk of missing eligible studies that could be indexed or tagged inconsistently across databases.

Study Selection

All records retrieved from the 6 databases were imported into Rayyan (Qatar Computing Research Institute) for deduplication and screening. Rayyan automatically identified potential duplicates, which were manually verified and removed, leaving 1417 unique records for screening. Study selection was then conducted in 3 sequential stages by two independent reviewers (MEA and YA). In the first stage, titles of all 1417 unique records were screened against the prespecified eligibility criteria, reducing the pool to 101 records. In the second stage, the 101 records were reviewed through a detailed assessment of their abstracts against the full eligibility criteria to determine if they could pass onto the next stage. In the third stage, full texts of the 29 records were retrieved and assessed independently for final inclusion. Records for which full text could not be obtained

XSL•FO
RenderX

via institutional access or interlibrary loan were pursued via direct author contact before being counted as inaccessible. Discrepancies at all stages were resolved through discussion and consensus between the two reviewers. If consensus could not be reached, a third senior reviewer (TY) acted as adjudicator. The reasons for exclusion at each stage are documented and reported in the PRISMA flow diagram.

Data Collection Process and Items

Data extraction was carried out independently by two reviewers (MEA and YA) using a prestandardized form built in Microsoft Excel. Before the main extraction began, the form was piloted on two studies and adjusted where wording was ambiguous, or fields needed refinement. This was a small but useful step that avoided inconsistencies surfacing later in the process.

The form captured 7 categories of information: study identification details (first author, year, country, and journal); study design; participant characteristics, including specialty, training level, total sample size, and group allocation; details of the AR intervention, covering the specific platform and device used, what the AR content actually consisted of, and how it was delivered; comparator characteristics; the task or procedural scenario; and all reported outcomes measures including primary and secondary endpoints alongside their associated statistical results, including means (SDs), *P* values, and effect sizes if reported.

Responsibility for extraction was shared equally between MEA and YA. If the two reviewers disagreed on whether a data point should be included, they discussed this together with adjudication by a third reviewer (TY) if no consensus could be reached.

Risk of Bias Assessment

Risk of bias was assessed independently by two reviewers (MEA and YA) using a tool selection determined by the study design. Randomized controlled trials (RCTs) were assessed using the Cochrane Risk of Bias 2 (RoB 2) tool [22]. Nonrandomized prospective cohort studies were assessed using the Risk of Bias in Nonrandomized Studies of Interventions (ROBINS-I) tool [23]. Crossover randomized trials were assessed using the RoB 2 tool with the crossover extension, as crossover designs remain RCTs and ROBINS-I is not applicable to them. Disagreements between reviewers were resolved by discussion; unresolved disagreements were adjudicated by the third reviewer (TY). Risk-of-bias findings are presented narratively in the Results section and considered in the interpretation of the overall body of evidence.

Synthesis Methods

Given the substantial clinical heterogeneity across included studies in terms of AR platforms used, participant training levels, surgical specialties, comparator conditions, and outcome measures used, statistical pooling (meta-analysis) was not appropriate and was not performed. No standardized effect sizes (eg, Cohen *d* and standardized mean difference) were reported by included studies in a sufficiently consistent form to enable pooling; where individual studies reported effect sizes, these are noted in the narrative synthesis. Findings were synthesized narratively following the SWiM reporting guideline [18]. Studies were grouped according to five prespecified thematic outcome domains derived during the protocol stage: (1) technical accuracy and procedural performance, (2) error reduction and procedural safety, (3) learning trajectory and skill acquisition, (4) cognitive load and gaze efficiency, and (5) operational efficiency and procedure time. Within each domain, consistency and heterogeneity of findings across studies were assessed narratively, and the direction and magnitude of effects were described in relation to methodological quality.

Results

Study Selection

The systematic search returned 4596 records across the 6 databases (Scopus: 1640; Web of Science: 1039; IEEE Xplore: 790; Embase: 638; Ovid MEDLINE: 438; PubMed: 51). A further 203 records came from reference list checking and gray literature searching, bringing the total to 4799. Rayyan automatically identified 3382 potential duplicates, which were manually verified and removed, leaving 1417 unique records for screening.

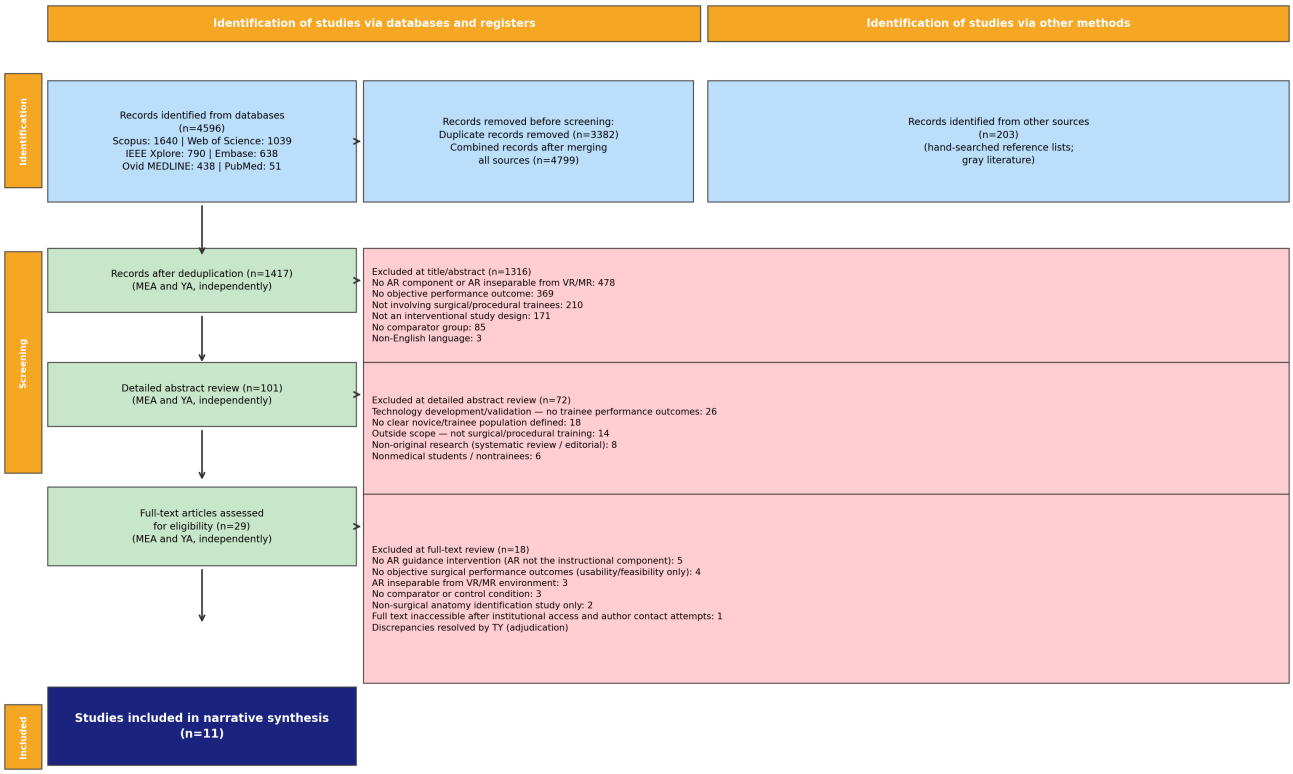
MEA and YA independently screened all 1417 records by title, excluding 1316. The most common reasons were: no AR component or AR not separable from VR/MR (n≈478), no objective performance outcome (n≈369), not involving surgical or procedural trainees (n≈210), not an interventional study design (n≈171), no comparator group (n≈85), and non-English language (n≈3). Disagreements were resolved by consensus. The remaining 101 records proceeded to a detailed abstract review.

Both reviewers independently assessed all 101 abstracts, excluding 72. The main reasons were technology development or validation studies with no trainee performance outcomes (n=26), no clearly defined novice or trainee population (n=18), outside scope (n=14), non-original research formats including systematic reviews and editorials (n=8), and nontrainee participants (n=6). Where the two reviewers could not reach agreement through discussion, TY adjudicated. Twenty-nine records were then retrieved for full-text assessment.

At full-text review, again conducted independently by MEA and YA, 18 articles were excluded. Reasons were: AR present in the setting but not functioning as the instructional component (n=5); usability or feasibility assessment only, with no objective performance outcomes (n=4); AR inseparable from a VR or MR environment (n=3); no comparator or control condition (n=3); anatomy identification study without any surgical skill training component (n=2); and full text inaccessible despite institutional access attempts, an interlibrary loan request, and direct contact with the authors (n=1). Disagreements were again resolved by discussion, with TY available for adjudication where needed. Eleven studies met all prespecified eligibility criteria and were included in the final narrative synthesis. The full selection process is shown in the PRISMA 2020 flow diagram (Figure 2).

XSL•FO
RenderX

Figure 2. PRISMA 2020 flow diagram illustrating the study selection process. Records were identified across 6 databases and additional sources. Sequential screening at title/abstract, detailed abstract review, and full-text stages progressively narrowed the pool from 4799 records to 11 included studies. AR: augmented reality; MR: mixed reality; VR: virtual reality.



Study Characteristics

The 11 included studies span 2021 to 2025 and together enrolled 347 participants. Nine were RCTs—5 parallel-group [3,5,6,8,9], 3 crossover [1,4,7], and 1 three-arm [11], with the remaining 2 being prospective cohort comparisons [2,10]. Sample sizes ranged from 8 (Liu et al [2]) to 60 (Wild et al [7]). Across all studies, participants had no or very limited prior experience in the procedure being trained. Seven specialties featured in the included work: minimally invasive and laparoscopic surgery (n=4), open and basic surgical skills (n=2), neurosurgery (n=2), neurovascular and cerebrovascular surgery (n=1), otology (n=1), and spine surgery (n=1). Full characteristics of each included study are provided in Table 2.

XSL•FO
RenderX

Table . Characteristics of included studies (n=11), covering 7 surgical specialties, published between 2021 and 2025, encompassing 347 participants with novice or limited prior procedural experience.

Author (year)	Specialty/procedure	n	Participants	Design	AR[a] platform	Comparator	Primary outcome
Wolf et al (2021) [1]	ECMO[b] cannulation	21	Medical students (Y3-Y4)	Crossover RCT[c]	HoloLens 2 step-by-step AR guide	Conventional SOP[d] (paper/video)	Error count; UEQ[e]
Liu et al (2024) [2]	Neurosurgery: MCA[f] aneurysm localization	8	Neurosurgery residents (PGY1-4; <5 aneurysm cases)	Prospective cohort	Magic Leap One + Brainlab HUD[g] (Zeiss Kinevo 900)	2D imaging review only	Aneurysm localization deviation (mm)
Cizmic et al (2023) [3]	Laparoscopic cholecystectomy (ex vivo × 10)	40	Medical students (Y3-Y6)	Parallel RCT	iSurgeon AR telestration	Verbal guidance only	Cumulative GOALS[h]/OSATS[i]; complications; CVS[j] achievement
Felinska et al (2023) [4]	Laparoscopic basic tasks + ex vivo cholecystectomy	40	MIS[k]-naive medical students	Crossover RCT	iSurgeon + Pupil Core eye-tracking	Verbal guidance only	Gaze latency, errors, OSATS, NASA-TLX[l]
Lopes et al (2022) [5]	Basic open surgical skills: suturing (x 5 types)	20	Medical students (no prior suturing experience)	Parallel RCT (evaluator-blinded)	Vuzix M300XL smart glasses + remote telestration	Traditional on-site teaching	Independent performance time; mOSATS score
Van Gestel et al (2021) [6]	Neurosurgery: EVD[m] placement (phantom)	16	Medical students (no prior EVD experience)	Parallel RCT (pre/post training)	HoloLens 1 + IR[n] inside-out tracking	Freehand technique	Mean target error (mm); mKS[o] grade
Wild et al (2022) [7]	Laparoscopic basic skills + ex vivo cholecystectomy	60	Laparoscopic novices (medical students Y3-Y6)	Crossover RCT	iSurgeon AR telestration	Verbal guidance only	Total training time; GOALS/OSATS; complications; NASA-TLX
Hadida Barzilai et al (2025) [8]	Otology: mastoidectomy drilling (3D-printed model)	21	Medical students (clinical clerkship; no prior temporal bone drilling)	Parallel RCT	HoloLens 2 + D2P QR-code registration	Anatomy review + instructional video + dissection manual	Modified Welling Scale (mWS, /25)
Nagayo et al (2022) [9]	Open surgery: subcuticular interrupted suturing	38	Medical students (suturing novices)	Parallel RCT (evaluator-blinded)	HoloLens 2 self-training (3D expert procedure replication)	2D instructional video	Global rating; task-specific suturing scores
Kong et al (2025) [10]	Spine surgery: pedicle screw placement (L2 lumbar model)	4 surgeons / 80 screws	1 experienced surgeon + 1 novice per group	Prospective cohort	HoloLens 2 + Vuforia 3D registration	Freehand technique	Linear deviation (mm); angular deviation; Gertzbein-Robbins accuracy
Dodier et al (2024) [11]	Neurosurgery: intracranial aneurysm clipping (perfused phantom)	9 residents	Neurosurgery residents (PGY1-6; no prior clipping as lead surgeon)	3-arm RCT	HoloLens 1 holographic AR clipping simulation (SOFA[p])	No interim training or video review only	Occlusion rate (Raymond-Roy class 1); clipping attempts; wrist tremor

[a]AR: augmented reality.

[b]ECMO: extracorporeal membrane oxygenation.

[c]RCT: randomized controlled trial.

[d]SOP: standard operating procedure.

[e]UEQ: User Experience Questionnaire.

[f]MCA: middle cerebral artery.

XSL•FO
RenderX

[g]HUD: heads-up display.
[h]GOALS: Global Operative Assessment of Laparoscopic Skills.
[i]OSATS: Objective Structured Assessment of Technical Skills.
[j]CVS: critical view of safety.
[k]MIS: minimally invasive surgery.
[l]NASA-TLX: NASA Task Load Index.
[m]EVD: external ventricular drain.
[n]IR: infrared.
[o]mKS: Modified Kakarla Scale.
[p]SOFA: Simulation Open Framework Architecture.

AR Technology Platforms

The Microsoft HoloLens (versions 1 and 2) was the most commonly used AR platform, applied in 5 of the 11 included studies [1,6,8,9,11]. The HoloLens is an optical see-through HMD capable of projecting holographic content into the user's visual field while preserving contact with the physical environment. Three studies used the iSurgeon telestration system, a laparoscopic screen-based AR device that projects a real-time feed of the instructor's hand gestures onto the operative monitor [3,4,7]. The Magic Leap One combined with the Brainlab Mixed Reality Viewer and intraoperative heads-up display was used in one neurosurgical study [2]. The Vuzix M300XL smart glasses were evaluated in one suturing study [5]. Kong et al [10] used the HoloLens 2 in combination with custom surgical guides and the Vuforia 3D registration software for spinal navigation, and Dodier et al [11] used the HoloLens 1 to deliver holographic finite-element simulation of aneurysm clipping.

Risk of Bias

Risk of bias assessment findings are summarized narratively below. Among the 9 RCTs, 6 were assessed as having some concerns regarding randomization or blinding processes [1,3,4,7,9,11]; full blinding of participants and instructors to group allocation is inherently unfeasible in AR training studies, representing a structural limitation of all trials in this field. The remaining 3 RCTs [5,6,8] were assessed as low risk across all domains. The 2 nonrandomized studies [2,10] were assessed using ROBINS-I and rated as moderate risk, reflecting their small sample sizes and lack of formal randomization, though both did use internal controls. Outcome assessment blinding was reported in 4 studies [5,6,8,9], which goes some way toward reducing detection bias. None of the included studies reported any long-term follow-up or skill retention data—a gap that runs consistently across the entire evidence base.

Narrative Synthesis of Outcomes

Domain 1: Technical Accuracy and Procedural Performance

Of the 11 included studies, 6 provided direct evidence of AR improving objective technical accuracy. The most pronounced effects were observed in procedural tasks with a strong visuospatial component. Van Gestel et al [6] demonstrated that untrained medical students using AR guidance for external ventricular drain (EVD) placement on a phantom model achieved a mean target error of 11.9 mm, compared with 19.9 mm for the untrained freehand group (P=.003). Critically, untrained AR-guided performance matched that of trained freehand performers, indicating that AR effectively compressed the procedural learning curve. The quality of EVD placement was also significantly superior in the AR group (59.4% vs 25% Modified Kakarla Scale grade 1, P=.005) [6].

Hadida Barzilai et al [8] reported significantly superior overall mastoidectomy performance in the AR group (median Modified Welling Scale 19.5/25) compared with controls (12/25; P=.001), with significant advantages on 6 of 8 subscales including mastoidectomy margin definition, sinodural angle, and tegmen exposure. The AR group's score exceeded published novice benchmark values.

Liu et al [2] demonstrated that AR-assisted visuospatial training significantly reduced aneurysm localization deviation among neurosurgical residents, from 8.1 mm at AR Test 1 to 2.7 mm at AR Test 2 (P<.001). Crucially, this improvement was retained in the final test conducted without any AR assistance (AR group: 2.1 mm vs control: 5.9 mm; P<.001), confirming durable skill acquisition rather than performance scaffolding alone.

Kong et al [10] demonstrated AR's equalizing effect on novice-expert performance disparity in pedicle screw placement: overall AR accuracy 95% versus 77.5% freehand (P<.05 for both linear and angular deviation). In contrast, Nagayo et al [9] found no significant difference between AR and video-based self-training in suturing skill improvement (global rating: P=.54; task specific: P=.91), and Lopes et al [5] similarly found no significant difference in mOSATS scores, although the telestration group performed tasks significantly faster when working independently (1393 s vs 1679 s; P=.04). These findings represent noninferiority rather than inferiority of AR.

Domain 2: Error Reduction and Procedural Safety

AR was consistently associated with reduced procedural errors across all 5 studies that measured this outcome. Felinska et al [4] demonstrated the most dramatic reduction, with AR-instructed trainees making a mean of 0.18 errors per task compared with 1.94 for the verbal instruction group (P<.01; ηp^2=0.92), representing a tenfold reduction in error rate. Complementary eye-tracking data revealed the mechanism: AR reduced gaze latency from 2.04 to 0.21 seconds (P<.01; ηp^2=0.95), confirming that AR telestration functions by directing trainees' visual attention to operationally relevant structures more rapidly and precisely than verbal instruction [4].

Wolf et al [1] found that AR-based extracorporeal membrane oxygenation cannulation instructions resulted in a 66% reduction

XSL•FO
RenderX

in knowledge-related errors for the more complex second procedure (18 vs 53 errors; $P<.05$), while handling errors were unchanged, suggesting that AR's error-reducing effect is specifically mediated by improved information accessibility and cognitive offloading. Cizmic et al [3] reported that the iSurgeon group incurred significantly fewer total complications and achieved the critical view of safety in 79.5% of procedures compared with only 41.4% in the verbal guidance group ($P\leq.001$). Wild et al [7] similarly reported a significant reduction in complication rates with AR telestration (13.3% vs 40%; $P=.02$).

Domain 3: Learning Trajectory and Skill Acquisition

Four studies provided explicit evidence regarding AR's effect on the learning trajectory. Liu et al [2] showed a steep decline in localization deviation across successive AR test blocks while the control groups remained relatively flat, indicating an accelerated learning curve rather than a simple one-time performance advantage. Cizmic et al [3] provided longitudinal evidence across 10 cholecystectomy sessions, demonstrating that the AR telestration group maintained consistently higher GOALS (Global Operative Assessment of Laparoscopic Skills) and OSATS (Objective Structured Assessment of Technical Skills) scores from the first session onwards, with the performance gap not narrowing over time. Van Gestel et al [6] finding that untrained AR performers matched trained freehand performers is perhaps the most striking demonstration of learning curve compression in this review. Dodier et al [11] found that only the video-plus-AR cohort achieved a statistically significant improvement in aneurysm occlusion rate between the first and final sessions (67% to 93%; $P=.05$), demonstrating that AR adds value beyond video review alone for complex microsurgical skill acquisition.

Domain 4: Cognitive Load and Gaze Efficiency

Three studies measured cognitive load using validated instruments. Felinska et al [4] reported significantly lower NASA Task Load Index scores during basic laparoscopic tasks with AR telestration compared with verbal instruction (mean 50 [SD 21] vs mean 56 [SD 22]; $P<.01$), alongside a lower objective blink rate. Wild et al [7] reported that participants found AR training significantly less mentally demanding (mean 33.3 [SD 14.8] vs mean 48.9 [SD 14.3]; $P<.001$) and less physically demanding (mean 35.1 [SD 13.8] vs mean 38.1 [SD 13.3]; $P=.002$). Liu et al [2] noted that the AR group took significantly longer to complete tasks in early test phases ($P=.003$), attributed to the additional cognitive processing required to colocate spatial AR hologram information with physical understanding, an effect that decreased as participants became more familiar with the AR system.

Domain 5: Operational Efficiency and Procedure Time

The effect of AR on procedure and training time was mixed across studies. Wild et al [7] reported the most pronounced efficiency gain: total laparoscopic training time was reduced by 29.8% with AR telestration (mean 1163 [SD 275] vs mean 1658 [SD 375] seconds; $P<.001$). Lopes et al [5] reported significantly faster independent suture completion in the AR group (1393 vs 1679 seconds; $P=.04$). For more complex procedures, no significant differences in total operative time were observed (mean 79.6 [SD 25.7] vs mean 84.5 [SD 33.2] minutes; $P=.09$), suggesting that AR's efficiency benefits are most readily detectable in discrete, structured tasks (Table 3).

XSL•FO
RenderX

Table . Summary of primary outcomes and key quantitative findings across included studies (n=11), covering participants enrolled in surgical and procedural training programs across 7 specialties, 2021 - 2025.

Author (year)	Primary outcome	AR[a] result	Control result	*P* value	Key interpretation
Wolf et al (2021) [1]	Error count (procedure 2)	18 knowledge errors	53 knowledge errors	<.05	66% reduction in knowledge errors; handling errors unchanged—AR reduces cognitive errors specifically
Liu et al (2024) [2]	Localization deviation (mm)	Mean 2.7 (SD 1.0) mm (AR test 2); 2.1 mm (final test without AR)	Mean 5.8 (SD 4.1) mm; 5.9 mm (final test)	.01; <.001	AR accelerates visuospatial learning curve; improvement retained in final unassisted test
Cizmic et al (2023) [3]	GOALS[b]; OSATS[c]; CVS[d]	GOALS 17.3; OSATS 50.8; CVS 79.5%	GOALS 16.0; OSATS 41.2; CVS 41.4%	<.001 (all)	AR telestration maintains higher performance trajectory across 10 LCs[e]; CVS achievement nearly doubled
Felinska et al (2023) [4]	Error count; gaze latency; OSATS	0.18 errors; 0.21 s gaze latency	1.94 errors; 2.04 s gaze latency	<.01 (all)	Tenfold error reduction; gaze guidance mechanism confirmed by eye-tracking; reduced NASA-TLX[f]
Lopes et al (2022) [5]	Independent task time; mOSATS	1393 s total; mOSATS trend higher	1679 s total	.04 (time)	AR telestration produces faster independent performance; quality comparable—viable alternative to on-site teaching
Van Gestel et al (2021) [6]	Mean target error (mm); mKS[g] grade 1	11.9 mm; 59.4% grade 1	19.9 mm; 25% grade 1	.003; .005	AR eliminates procedural learning curve for EVD[h] placement; untrained AR matches trained freehand
Wild et al (2022) [7]	Training time; GOALS/OSATS; complications	1163 s; GOALS 21; OSATS 67; 13.3% complications	1658 s; GOALS 18; OSATS 61; 40% complications	<.001; .007; .015; .020	29.8% training time; significant quality and safety improvement; reduced NASA-TLX
Hadida Barzilai et al (2025) [8]	Modified Welling Scale (/25)	19.5/25 (median)	12/25 (median)	.001	7.5-point advantage; AR group exceeds published novice benchmarks; 6/8 subscales significant
Nagayo et al (2022) [9]	Global rating (GR); task-specific (TS) scores	GR 16.03; TS 15.03 (posttest)	GR 15.5; TS 15.11 (posttest)	.54; .91 (NS[i])	Noninferior to video self-training; AR rated more useful for 3D instrument motion (*P*=.02)
Kong et al (2025) [10]	Linear deviation; angular deviation; Gertzbein-Robbins accuracy	Novice: 1.73 mm / 2.87 degrees; 90% accuracy	Novice: 5.25 mm / 7.15 degrees; 70% accuracy	<.05 (all)	AR equalizes novice-expert performance gap; overall accuracy 95% vs 77.5% freehand
Dodier et al (2024) [11]	Aneurysm occlusion rate (Raymond-Roy class 1)	67%-93% (video + AR cohort)	67%-73% (video only); stable (control)	.046	Only video + AR cohort achieved significant occlusion improvement; AR adds value beyond video alone

[a]AR: augmented reality.

[b]GOALS: Global Operative Assessment of Laparoscopic Skills.

XSL•FO
RenderX

[c]OSATS: Objective Structured Assessment of Technical Skills.
[d]CVS: critical view of safety.
[e]LC: laparoscopic cholecystectomy.
[f]NASA-TLX: NASA Task Load Index.
[g]mKS: Modified Kakarla Scale.
[h]EVD: external ventricular drain.
[i]NS: not significant.

Discussion

Principal Findings

This systematic review synthesizes evidence from 11 contemporary studies (2021 - 2025) evaluating AR's impact on the objective technical performance of surgical trainees. The principal finding is that AR demonstrates a consistent, measurable positive effect on technical performance, most strongly in domains requiring visuospatial reasoning, spatial anatomical understanding, and procedural accuracy. Of the 11 studies, 9 reported at least one significant improvement in an objective technical performance metric. The 2 studies that did not demonstrate AR superiority [5,9] nonetheless showed noninferiority, with AR-trained groups performing equivalently to comparators on skill quality metrics while achieving time advantages in independent performance. The absence of superiority in lower-stakes basic skills tasks is not evidence of inefficacy; it may reflect a ceiling effect in tasks where traditional instruction is already adequate for novice performance.

Across the body of evidence, the most robust performance advantages were observed in tasks with a strong visuospatial or spatial navigation component: EVD placement [6], aneurysm localization [2], mastoidectomy drilling [8], and pedicle screw placement [10]. This pattern is theoretically coherent: AR's capacity to render 3D anatomical structures in the trainee's visual field directly addresses a fundamental cognitive challenge in procedural surgery: the mental reconstruction of volumetric anatomy from 2D imaging data.

Interpretation and Comparison With Prior Literature

The findings of this review are consistent with, and substantially extend, the conclusions of prior systematic reviews. Abu Halimah et al [15] and Xiong et al [16] identified broad potential for AR in surgical skills training, but their reviews included older studies with heterogeneous definitions of AR and outcomes. By restricting our scope to post-2020 studies with objective outcomes and clear AR definitions, we provide a more precise assessment applicable to current training environments. Importantly, this review positions AR and VR as complementary rather than competitive modalities, a distinction emphasized in the literature [11,24].

The observed expertise reversal pattern, AR providing the greatest benefit to novices with diminishing returns at higher levels of proficiency, is consistent with predictions from both cognitive load theory [12] and the expertise reversal effect [14]. In Kong et al [10], AR navigation essentially equalized the novice-expert performance gap. This pattern has direct implications for curriculum design: AR-assisted training may be most efficiently used during the early stages of procedural learning, with progressive withdrawal of AR guidance as competency develops, a strategy consistent with the scaffolding framework in educational theory [25,26].

The gaze-guidance mechanism elucidated by Felinska et al [4] provides the most direct experimental evidence for the cognitive mechanism underpinning AR's training benefit. By demonstrating that AR telestration reduced gaze latency tenfold and aligned trainee gaze with expert gaze, this study demonstrates that AR's error-reducing effect is mediated by directing visual attention to operationally relevant anatomical regions more efficiently than verbal instruction.

Dodier et al [11] finding that only the combined video-plus-AR cohort achieved a significant improvement in aneurysm occlusion rate (67% to 93%; P=.05) is particularly noteworthy. The holographic AR clipping simulation allowed residents to test different clipping strategies on the exact same patient-specific anatomy as the physical phantom, a form of deliberate practice that is impossible with traditional simulation.

Limitations

Several limitations of the constituent studies and of this review must be acknowledged. The most fundamental limitation is heterogeneity: the 11 included studies span 7 surgical specialties, use 6 distinct AR platforms, and measure outcomes using a wide variety of instruments, precluding statistical synthesis. Sample sizes were consistently small (range: 8 - 60 participants), limiting statistical power. Publication bias cannot be excluded.

None of the 11 included studies reported long-term follow-up of skill retention or assessed the transfer of AR-trained skills to the real operating room or clinical environment. This is perhaps the most significant gap in the current evidence base. Device-related limitations were noted across several studies, including physical discomfort associated with prolonged HMD use [6] and interface familiarization time across HoloLens-based studies. The risk of bias assessment found some concerns in 6 of 11 RCTs, predominantly related to allocation concealment and blinding, which is an inherent structural limitation of AR training trials rather than a correctable methodological weakness. The review was not prospectively registered, which is acknowledged as a limitation.

One study for which full text could not be obtained despite institutional access, interlibrary loan request, and direct author contact was excluded; this represents 1 of 29 full-text articles reviewed (3.4%) and, given the consistency of findings across the 11 included studies, is unlikely to materially alter the direction of the conclusions.

XSL•FO
RenderX

Future Research Directions

The current evidence points toward several concrete priorities for future work. Most pressing is the need for adequately powered multicenter randomized trials across the more promising AR platforms and specialty domains, using standardized outcome measures that would actually allow findings to be compared across studies—something the current literature makes difficult. Alongside this, longitudinal studies with skill-retention assessments at 3, 6, and 12 months posttraining are needed to establish whether the performance gains associated with AR hold over time or fade once the technology is removed. Transfer studies examining whether AR-trained skills translate meaningfully into clinical performance are a logical next step that the field has yet to take seriously. Finally, cost-effectiveness analyses will matter enormously for any health system considering curriculum-level adoption—the training benefit needs to be weighed against the real costs of hardware, software, and implementation, and that work has not yet been done.

Conclusions

This review found consistent evidence that AR improves technical performance in surgical novices—reduced procedural errors, better accuracy, faster progression along the learning curve, and lower cognitive load, particularly in tasks with high visuospatial demands. The expertise reversal pattern that emerged across multiple studies is worth taking seriously: AR appears to deliver its greatest benefit during the early, high-error phase of skill acquisition, with returns diminishing as experience accumulates. That finding has practical implications for how AR should be used—not as a permanent scaffold, but as a targeted intervention in early training, with guidance progressively withdrawn as competency develops.

What this review cannot claim is that the evidence is mature. Sample sizes are small, platforms vary enormously, outcome measures are inconsistent, and no study has yet examined whether skills are retained or transferred to real clinical settings. AR shows genuine promise as an adjunct within structured surgical curricula—not a replacement for expert mentorship or traditional teaching, but something that adds real value when used thoughtfully alongside them. Turning that promise into confident implementation guidance will require the kind of rigorous, large-scale, longitudinal work that the field has not yet produced.

Acknowledgments

The authors thank the editorial team and peer reviewers at *JMIR Medical Education* for their rigorous and constructive engagement with this manuscript across multiple review rounds.

Several generative AI tools were used at specific stages of this manuscript's preparation. Claude (Anthropic) and ChatGPT (OpenAI) were used (1) to assist with the verification and optimization of database search string syntax during the development of the search strategy (all search terms, conceptual domains, and PICO components were defined by the human authors; AI was used to check syntactic compatibility with database field codes and Boolean operators); (2) to support calculations and structured analysis of objective outcome measures extracted from included studies (all primary data were extracted independently by human reviewers; AI assisted in organizing and cross-checking numerical results for consistency); (3) for grammar, punctuation, and language editing of the final manuscript draft; and (4) for drafting of the Acknowledgements section. No AI tool was used to independently generate, fabricate, or interpret scientific findings, formulate conclusions, conduct screening or eligibility assessment, or make any intellectual judgment regarding the inclusion or exclusion of studies. All scientific content, data synthesis, risk-of-bias assessments, interpretations, and conclusions are the sole responsibility of the human authors.

Funding

No external funding was received for this study. No funder was involved in any aspect of the study design, data collection, analysis, interpretation, or the writing of the manuscript.

Data Availability

This systematic review analyzed data reported in the published studies cited in the reference list. No new primary data were generated. The data extraction spreadsheet and screening records are available from the corresponding author upon reasonable request.

Authors' Contributions

MEA conceived the review, developed the protocol, conducted database searches, performed title/abstract and full-text screening (first reviewer), conducted data extraction and risk of bias assessment, and drafted the manuscript. YA performed independent title/abstract and full-text screening (second reviewer) and independent data extraction and risk of bias assessment. TY adjudicated screening and data extraction discrepancies. AEA and HK contributed to manuscript review and revision. All authors reviewed and approved the final manuscript.

XSL•FO
RenderX

Conflicts of Interest

None declared.

Multimedia Appendix 1
Full electronic search strategies.
[DOCX File, 23 KB - mededu_v12i1e71572_app1.docx]

Checklist 1
PRISMA 2020 checklist.
[DOCX File, 27 KB - mededu_v12i1e71572_app2.docx]

References

1. Wolf J, Wolfer V, Halbe M, Maisano F, Lohmeyer Q, Meboldt M. Comparing the effectiveness of augmented reality-based and conventional instructions during single ECMO cannulation training. Int J Comput Assist Radiol Surg 2021 Jul;16(7):1171-1180. [doi: 10.1007/s11548-021-02408-y] [Medline: 34023976]
2. Liu X, Xiao W, Yang Y, Yan Y, Liang F. Augmented reality technology shortens aneurysm surgery learning curve for residents. Comput Assist Surg (Abingdon) 2024 Dec;29(1):2311940. [doi: 10.1080/24699322.2024.2311940] [Medline: 38315080]
3. Cizmic A, Müller F, Wise PA, et al. Telestration with augmented reality improves the performance of the first ten ex vivo porcine laparoscopic cholecystectomies: a randomized controlled study. Surg Endosc 2023 Oct;37(10):7839-7848. [doi: 10.1007/s00464-023-10360-y] [Medline: 37612445]
4. Felinska EA, Fuchs TE, Kogkas A, et al. Telestration with augmented reality improves surgical performance through gaze guidance. Surg Endosc 2023 May;37(5):3557-3566. [doi: 10.1007/s00464-022-09859-7] [Medline: 36609924]
5. Neves Lopes V, Dantas I, Barbosa JP, Barbosa J. Telestration in the teaching of basic surgical skills: a randomized trial. J Surg Educ 2022;79(4):1031-1042. [doi: 10.1016/j.jsurg.2022.02.013] [Medline: 35331681]
6. Van Gestel F, Frantz T, Vannerom C, et al. The effect of augmented reality on the accuracy and learning curve of external ventricular drain placement. Neurosurg Focus 2021 Aug;51(2):E8. [doi: 10.3171/2021.5.FOCUS21215] [Medline: 34333479]
7. Wild C, Lang F, Gerhäuser AS, et al. Telestration with augmented reality for visual presentation of intraoperative target structures in minimally invasive surgery: a randomized controlled study. Surg Endosc 2022 Oct;36(10):7453-7461. [doi: 10.1007/s00464-022-09158-1] [Medline: 35266048]
8. Hadida Barzilai D, Tejman-Yarden S, Yogev D, et al. Augmented reality-guided mastoidectomy simulation: a randomized controlled trial assessing surgical proficiency. Laryngoscope 2025 Feb;135(2):894-900. [doi: 10.1002/lary.31791] [Medline: 39315469]
9. Nagayo Y, Saito T, Oyama H. Augmented reality self-training system for suturing in open surgery: a randomized controlled trial. Int J Surg 2022 Jun;102:106650. [doi: 10.1016/j.ijsu.2022.106650] [Medline: 35525415]
10. Kong H, Wang S, Zhang C, Chen Z. Augmented reality navigation using surgical guides versus conventional techniques in pedicle screw placement. J Shanghai Jiaotong Univ (Sci) 2025 Feb;30(1):10-17. [doi: 10.1007/s12204-023-2689-5]
11. Dodier P, Civilla L, Mallouhi A, et al. An evaluation of physical and augmented patient-specific intracranial aneurysm simulators on microsurgical clipping performance and skills: a randomized controlled study. Neurosurg Focus 2024 Jan;56(1):E9. [doi: 10.3171/2023.10.FOCUS23640] [Medline: 38163349]
12. Sweller J. Cognitive load during problem solving: effects on learning. Cogn Sci 1988 Apr;12(2):257-285. [doi: 10.1207/s15516709cog1202_4]
13. Barsom EZ, Graafland M, Schijven MP. Systematic review on the effectiveness of augmented reality applications in medical training. Surg Endosc 2016 Oct;30(10):4174-4183. [doi: 10.1007/s00464-016-4800-6] [Medline: 26905573]
14. Sweller J, Ayres P, Kalyuga S. The expertise reversal effect. In: Cognitive Load Theory: Springer; 2011:155-170. [doi: 10.1007/978-1-4419-8126-4_12]
15. Abu Halimah J, Mojiri ME, Ali AA, et al. Assessing the impact of augmented reality on surgical skills training for medical students: a systematic review. Cureus 2024 Oct 10;16. [doi: 10.7759/cureus.71221]
16. Xiong J, Dai X, Zhang Y, Liu X, Zhou X. Augmented reality for basic skills training in laparoscopic surgery: a systematic review and meta-analysis. Surg Endosc 2025 Jan;39(1):307-318. [doi: 10.1007/s00464-024-11387-5] [Medline: 39532736]
17. Page MJ, McKenzie JE, Bossuyt PM, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ 2021 Mar 29;372:n71. [doi: 10.1136/bmj.n71] [Medline: 33782057]
18. Campbell M, McKenzie JE, Sowden A, et al. Synthesis Without Meta-Analysis (SWiM) in systematic reviews: reporting guideline. BMJ 2020 Jan 16;368:l6890. [doi: 10.1136/bmj.l6890] [Medline: 31948937]
19. Lefebvre C, Glanville J, Briscoe S, et al. Searching for and selecting studies. In: Cochrane Handbook for Systematic Reviews of Interventions: Cochrane; 2019:67-107. [doi: 10.1002/9781119536604.ch4]
20. Rethlefsen ML, Kirtley S, Waffenschmidt S, et al. PRISMA-S: an extension to the PRISMA statement for reporting literature searches in systematic reviews. Syst Rev 2021 Jan 26;10(1):39. [doi: 10.1186/s13643-020-01542-z] [Medline: 33499930]

XSL•FO
RenderX

21. Haddaway NR, Collins AM, Coughlin D, Kirk S. The role of Google Scholar in evidence reviews and its applicability to grey literature searching. PLoS ONE 2015;10(9):e0138237. [doi: 10.1371/journal.pone.0138237] [Medline: 26379270]
22. Sterne JAC, Savović J, Page MJ, et al. RoB 2: a revised tool for assessing risk of bias in randomised trials. BMJ 2019 Aug 28;366:l4898. [doi: 10.1136/bmj.l4898] [Medline: 31462531]
23. Sterne JA, Hernán MA, Reeves BC, et al. ROBINS-I: a tool for assessing risk of bias in non-randomised studies of interventions. BMJ 2016 Oct 12;355:i4919. [doi: 10.1136/bmj.i4919] [Medline: 27733354]
24. McKnight RR, Pean CA, Buck JS, Hwang JS, Hsu JR, Pierrie SN. Virtual reality and augmented reality-translating surgical training into surgical technique. Curr Rev Musculoskelet Med 2020 Dec;13(6):663-674. [doi: 10.1007/s12178-020-09667-3] [Medline: 32779019]
25. Wood D, Bruner JS, Ross G. The role of tutoring in problem solving. J Child Psychol Psychiatry 1976 Apr;17(2):89-100. [doi: 10.1111/j.1469-7610.1976.tb00381.x] [Medline: 932126]
26. Aggarwal R, Mytton OT, Derbrew M, et al. Training and simulation for patient safety. BMJ Qual Safety 2010;19(Suppl 2):i34-i43. [doi: 10.1136/qshc.2009.038562] [Medline: 20693215]

Abbreviations

AR: augmented reality
EVD: external ventricular drain
GOALS: Global Operative Assessment of Laparoscopic Skills
HMD: head-mounted display
MeSH: Medical Subject Headings
MR: mixed reality
NASA-TLX: NASA Task Load Index
OSATS: Objective Structured Assessment of Technical Skills
PICO: population, intervention, comparison, outcome
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
RCT: randomized controlled trial
RoB 2: Cochrane Risk of Bias 2 Tool
ROBINS-I: Risk of Bias in Nonrandomized Studies of Interventions
SWiM: Synthesis Without Meta-Analysis
VR: virtual reality

Edited by A Stone; submitted 21.Jan.2025; peer-reviewed by B Niroomand, D Demirel, F Agresta; revised version received 01.May.2026; accepted 04.May.2026; published 03.Jun.2026.

Please cite as:
El Ashry M, El Ashry A, Khalique H, Abdalle Y, Yeung T
Augmented Reality in Surgical Training: Systematic Review of Its Impact on Technical Performance in Surgical Trainees
JMIR Med Educ 2026;12:e71572
URL: https://mededu.jmir.org/2026/1/e71572
doi:10.2196/71572

© Mahmoud El Ashry, Ahmed El Ashry, Hamza Khalique, Yahye Abdalle, Thomas Yeung. Originally published in JMIR Medical Education (https://mededu.jmir.org), 3.Jun.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Education, is properly cited. The complete bibliographic information, a link to the original publication on https://mededu.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Review

Main Recommendations for Developing Education and Awareness Strategies for Rare Diseases: Scoping Review

Diego Bettiol Yamada[1], PhD; Tatiana Takahasi Komoto[1], PhD; Vinícius Costa Lima[1,2], PhD; Julio Souza[3], PhD; Michele de Souza Seixas[1], BSc; Victor Cassão[1], MSc; Victor Evangelista de Faria Ferraz[1], MD, PhD; Bibiana Mello de Oliveira[4], MD, PhD; Têmis Maria Félix[5,6], MD, PhD; Domingos Alves[1,6], PhD

[1]Ribeirao Preto Medical School, University of Sao Paulo, Ribeirao Preto, Brazil

[2]RISE-Health, Faculty of Medicine, University of Porto, Porto, Portugal

[3]Institute of Engineering, Polytechnic of Porto, Porto, Portugal

[4]Federal University of Health Sciences of Porto Alegre, Porto Alegre, Brazil

[5]Medical Genetics Service, Hospital de Clinicas de Porto Alegre, Porto Alegre, Brazil

[6]Brazilian National Institute on Rare Diseases, Porto Alegre, Brazil

Corresponding Author:
Diego Bettiol Yamada, PhD
Ribeirao Preto Medical School, University of Sao Paulo
Av. Bandeirantes, 3900
Ribeirao Preto, 14049-900
Brazil
Phone: 55 16 3602 2433
Email: diego.yamada@alumni.usp.br

Abstract

Background: According to the World Health Organization, education and awareness are essential components of public health promotion strategies. In the context of rare diseases (RDs), these actions are particularly critical because of persistent stigma, fragmented knowledge, and the frequent absence of consolidated clinical and organizational protocols. These gaps often result in inappropriate referrals, inefficient care pathways, unnecessary procedures, and delays in diagnosis, negatively affecting health outcomes and quality of life.

Objective: This study aimed to identify and systematize the main recommendations for health education and awareness in the field of RDs, supporting the development of health care programs, public policies, and strategic initiatives.

Methods: We formulated the research question using the Population, Concept, and Context framework. This scoping review followed the PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses Extension for Scoping Reviews) guidelines to ensure methodological transparency. Eligible records included peer-reviewed research articles of any design and official documents published in Portuguese, English, or Spanish, with no time restrictions. Records that did not address the research question, lacked sufficient rigor, or focused exclusively on specific subgroups of RDs were excluded. Searches were performed in PubMed/MEDLINE, Scopus, Embase, Web of Science, as well as gray literature. Study selection and data extraction were conducted by the research team, with disagreements resolved and the included sources reviewed by an RDs expert. Data were thematically categorized by consensus, and descriptive statistics were used to summarize findings.

Results: A total of 58 sources of evidence were included. Among the identified recommendations related to education and awareness, most sources focused on professional education and training (49/58, 84.4%), followed by public policies and intersectoral integration (36/58, 62%), education and awareness for the general population (28/58, 48.2%), digital technologies (27/58, 46.5%), emotional support and experience sharing (20/58, 34.4%), and awareness events and dates (8/58, 13.7%). Percentages exceed 100% because individual sources could report multiple recommendations. Overall, the literature emphasizes integrating RDs content into educational initiatives and strengthening professional competencies, intersectoral collaboration, digital technologies, and broader awareness strategies.

Conclusions: This scoping review systematically mapped and organized recommendations from diverse sources of evidence on strategies for health education and awareness related to RDs. It synthesizes heterogeneous evidence using a structured approach to provide a comprehensive overview of strategies in this field, consolidating dispersed knowledge into a coherent body of evidence. The findings may inform improvements in health services, as well as professional and managerial practices, and

XSL•FO
RenderX

initiatives aimed at supporting patients, families, and advocacy groups involved in RDs, with potential implications for strengthening diagnostic processes, referral coordination, and more equitable access to information and care.

(JMIR Med Educ 2026;12:e79027) doi:10.2196/79027

KEYWORDS

educational strategies; health awareness; health education; public health recommendations; rare diseases

Introduction

Rationale

According to the World Health Organization (WHO) recommendations, education and awareness are fundamental processes for the success of public health promotion strategies [1]. In the context of rare diseases (RDs), these actions become even more necessary and relevant because of the persistence of stigma and discrimination associated with these conditions. This scenario stems from several factors, including gaps in structured and disseminated knowledge about RDs, the absence of consolidated clinical and organizational protocols within health care institutions, and the lack of coordinated frameworks capable of supporting adaptable educational initiatives [2-4]. Without shared educational approaches, public health messaging and professional training efforts often remain fragmented, limiting the dissemination of reliable information among stakeholders involved in RD care [3,4].

RDs are conditions with low prevalence in a given country or region. However, there is no global consensus on their definition, as each jurisdiction classifies them according to its own criteria. In the United States, a disease is considered rare when it affects fewer than 200,000 people or approximately 1 person per 1500 inhabitants [5]. In the European Union, a condition is classified as rare when it affects fewer than 1 person per 2000 individuals [6]. In Brazil, the Unified Health System (Portuguese: Sistema Único de Saúde) adopts the WHO criterion, which considers a disease rare if it affects up to 1.3 individuals per 2000 people [7]. Currently, there are more than 7000 known distinct RDs worldwide [8]. Although each condition individually affects a limited number of people, collectively these diseases represent a substantial public health challenge, with a global prevalence rate estimated at 5.9%, corresponding to nearly 300 million affected individuals worldwide [9,10].

The heterogeneity of RDs makes the establishment of standardized clinical protocols and care approaches particularly challenging. In addition, the lack of precise and up-to-date data regarding the incidence and prevalence of these conditions compromises health care planning and resource allocation [11]. In response to these limitations, several research and development networks focused on RDs have been established in regions such as the European Union [12], the United States [13], Brazil [3], Latin America [14], and Asia [15]. However, despite the institutionalization of these initiatives, education and awareness efforts related to RDs are not always transparent, well-structured, or effective in practice [16]. This gap in educational practices challenges health care providers, caregivers, and patients, who frequently face misinformation or limited access to reliable information, especially in regions with weaker health care systems or digital infrastructure [16,17].

Although health curricula are already overloaded, the development of engaging and innovative educational strategies for RDs remains necessary [18]. The absence of comprehensive informational and educational frameworks directly contributes to delays in diagnosis and treatment initiation within health care systems worldwide. In the United States, for instance, the average time required to diagnose a RD can reach up to 7 years, a period commonly referred to as a “diagnostic odyssey” [19,20]. This interval may be even longer in developing countries [21]. Beyond delayed diagnosis, the high informational and managerial complexity associated with RDs also contributes to referral errors, inefficient patient flows, unnecessary procedures, and inadequate care coordination. Considering that most RDs are chronic and debilitating conditions, this scenario imposes a significant burden on patients, families, health care institutions, and society as a whole [3,22,23].

In light of these considerations, effective educational and awareness strategies must be designed based on a robust understanding of existing recommendations, evidence-based practices, community experiences, and digital approaches capable of improving knowledge dissemination and promoting equity in access to care [9,16,18]. However, despite the growing recognition of the importance of education and awareness in the RD field, the available evidence remains fragmented across different disciplines, stakeholders, and health care contexts. Educational initiatives may involve health care professionals, patients, families, advocacy organizations, policymakers, and educators. This complexity highlights the need for a broad and exploratory synthesis capable of systematically mapping the literature and identifying the main recommendations regarding education and awareness in RDs.

Objective

Thus, this study aims to conduct a scoping review to identify and systematize the main recommendations for health education and awareness in the context of RDs, supporting the development of health programs, strategies, and public policies in this area.

Methods

Overview

As previously mentioned, a scoping review was conducted to investigate the state of the art regarding the main global recommendations for health education and awareness in the domain of RDs. The Population, Concept, and Context (PCC) methodology was adopted to define the research question. Initially proposed by the Joanna Briggs Institute, the PCC

XSL•FO
RenderX

framework is particularly robust for reviews aiming to explore and map the state of the art in the literature on a given topic, identifying key concepts, knowledge gaps, and domain-specific demands [24].

It is especially useful for exploratory themes and broad investigations. The adoption of this methodology, therefore, enables a systematic approach to conducting the stages of a scoping review, providing a solid foundation on which the study is built and contributing to more consistent and relevant results [24]. Accordingly, the guiding question of the scoping review was established and is described as follows: "What are the main recommendations on health education and awareness within the context of RDs?"

After defining the research question, the PCC framework guided the development of the search strategy. To ensure clear, systematic, and transparent reporting, this scoping review followed the PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses Extension for Scoping Reviews); the completed checklist is provided in Multimedia Appendix 1. The selection of sources of evidence was illustrated using the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) 2020 flow diagram, which can be adapted for scoping reviews and applied in conjunction with this framework. These tools are widely recognized for providing a structured approach to documenting all stages of a review, including the identification, screening, assessment of eligibility, and selection of sources, to answer the research guiding question [25,26].

The PRISMA-ScR is especially useful for documenting the aforementioned steps in a structured manner. It also serves as a support tool for research teams, reviewers, and evaluators involved in the review process. From a scientific standpoint, its use is important because it enables the reproducibility of the process by other researchers, helping to avoid redundant work while maintaining methodological rigor. For these reasons, it is considered a scientific best practice, aligned with international guidelines for conducting and reporting reviews [22,25].

In the following sections, we describe the inclusion and exclusion criteria, databases used, search strategy, study selection process, PRISMA 2020 flow diagram, and procedures for extracting knowledge from the selected sources of evidence. These procedures were designed to address the research question and guide the subsequent stages of the study.

Protocol and Registration

A review protocol was not registered for this scoping review. The study is supported by the São Paulo Research Foundation (Fundação de Amparo à Pesquisa do Estado de São Paulo) and was approved by the Research Ethics Committee of the Hospital das Clínicas da Faculdade de Medicina de Ribeirão Preto da Universidade de São Paulo (HCFMRP-USP; protocol number CAAE 82571424.4.0000.5440).

Eligibility Criteria

The inclusion and exclusion criteria were defined to ensure relevance, appropriateness, and precision in alignment with the investigation's objectives. These criteria considered the types of studies, languages, publication periods, and the evaluation of the content of each source. The specific criteria are outlined in detail in Textbox 1.

Textbox 1. Inclusion and exclusion criteria.

Inclusion criteria

- Research articles, including full papers, short communications, and peer-reviewed conference papers
- All study designs were eligible for inclusion, including qualitative, quantitative, mixed methods studies, and reviews
- Official government documents, such as policies, programs, action plans, and technical reports
- Sources published in Portuguese, English, and Spanish
- No restrictions were applied regarding the date of publication
- Sources were included if they addressed the research question and explicitly reported on relevant aspects of health education, health awareness, or literacy in the context of rare diseases (RDs)

Exclusion criteria

- Publications that did not meet minimum methodological standards to support the research question (eg, lacking defined methods or sources)
- Editorials, opinion pieces, and content from social media
- Sources published in languages other than those listed in the inclusion criteria
- Sources that mention relevant terms without providing substantive description or analysis of these topics
- Sources focused exclusively on specific clinical procedures, an individual RD, or specific groups of RDs, without addressing broader educational or awareness-related aspects

Given the interdisciplinary and multifaceted nature of education and awareness in the field of RDs, the inclusion criteria encompassed a broad range of study types, including qualitative, quantitative, mixed methods research, reviews, and official government documents. The inclusion of sources in English, Spanish, and Portuguese was intended to capture diverse regional perspectives, particularly from countries with established RD frameworks in the Americas and Europe, while

XSL•FO
RenderX

acknowledging the central role of English in international scientific communication.

No date restrictions were applied to allow the mapping of historical and recent developments in the field. Exclusion criteria were applied to ensure conceptual and methodological consistency by filtering out documents lacking analytical depth, empirical grounding, or alignment with the scope of the review. Publications focusing narrowly on clinical aspects without addressing educational or awareness-related dimensions were excluded to maintain focus on the broader systemic and communicative strategies relevant to the objectives of the review.

Information Sources

The databases were selected based on their scope and relevance to the research topic, literature coverage, accessibility, robustness, and the quality of the indexed studies. Therefore, the databases defined for the search were PubMed/MEDLINE, Scopus, Embase, and Web of Science.

The selection of these databases aims to ensure adequate coverage of the central theme of the review and the necessary geographic scope, while also considering the types of documents indexed, the curation of records, scientific relevance, reproducibility of results, and methodological rigor associated with validated evidence. Thus, we built and applied the search strategy within these databases to identify the studies composing the scoping review. The initial search was conducted on May 12, 2025. The search strategy was updated iteratively during the review process and was last rerun on March 15, 2026, across all databases.

In addition to bibliographic databases, gray literature was searched across governmental and institutional websites (eg, WHO, the European Commission, the US Food and Drug Administration, Rare Diseases International, and Asia-Pacific Economic Cooperation resources), as well as through Google Scholar as a supplementary web search tool. Records identified through this approach were considered eligible if they met the predefined eligibility criteria and addressed the research question. These online resources were manually browsed and searched to identify relevant sources. Citation searching was also performed by screening the reference lists of included studies. No additional studies were sought by contacting authors or experts.

Search

An initial limited search was conducted in PubMed/MEDLINE and Embase to identify relevant articles. Text words appearing in the titles, abstracts, and indexed descriptors of the retrieved studies were analyzed to inform the development of a comprehensive search strategy. Using the refined terms, a full search was subsequently performed in PubMed/MEDLINE, Embase, Scopus, and Web of Science. The reference lists of included studies were manually screened to identify additional relevant sources of evidence.

The search strategy was also informed by the Medical Subject Headings (MeSH) vocabulary [27], which was consulted to identify relevant concepts and synonyms. Controlled vocabulary, free-text terms, and field modifiers were combined using Boolean operators (AND/OR), with search strategies adapted to the specific characteristics of each database in accordance with methodological guidelines for systematic and scoping reviews [24].

PubMed/MEDLINE supports comprehensive free-text searching across multiple fields and uses Automatic Term Mapping to map search terms to MeSH, synonyms, and other relevant terms when applicable. Given its central role in indexing biomedical and public health literature, a high-sensitivity search strategy was implemented in PubMed/MEDLINE. The search was conducted using the “All Fields” option in combination with the species: humans filter to maximize retrieval, consistent with the exploratory nature of scoping review methodology.

In contrast, databases with different indexing structures required structured field restrictions to optimize retrieval precision and minimize nonrelevant records. Accordingly, searches in Scopus were limited to TITLE-ABS-KEY fields, while in Web of Science the Topic (TS) field was used, encompassing titles, abstracts, and author keywords. In Embase, database-specific syntax was applied, combining controlled vocabulary (Emtree terms with explosion, where appropriate) and free-text terms restricted to title, abstract, and keyword fields (ti, ab, and kw).

These adaptations reflect inherent differences in indexing systems and search functionalities across databases. The search was not restricted by publication date or study design to ensure comprehensive and exploratory coverage of the available evidence, consistent with the objectives of a scoping review. Language restrictions (English, Spanish, and Portuguese) were applied where available using database filters and were also considered during the eligibility assessment, reflecting the language proficiency of the review team, ensuring accurate interpretation of the included evidence, and aligning with the scope of the review as defined by the eligibility criteria. A filter for studies involving humans was applied where available.

The full search strategies for all databases are provided in Multimedia Appendix 2. Additionally, the PRISMA-S (Preferred Reporting Items for Systematic Reviews and Meta-Analyses Literature Search Extension) checklist [28] is provided in Multimedia Appendix 3 to enhance transparency and completeness in reporting the search strategy and information sources.

Selection of Sources of Evidence

After applying the search strategy across the 4 selected databases, all records were imported into Rayyan (Rayyan Systems Inc), where potential duplicates were flagged using its automatic deduplication function, and all duplicate removals were confirmed through manual review before exclusion. Titles and abstracts were then screened according to predefined inclusion and exclusion criteria. Full texts of potentially eligible studies were subsequently assessed for eligibility. Screening was conducted by 2 independent reviewers, and any conflicts regarding study eligibility were resolved by a third reviewer. All selected sources were reviewed by an expert in the field of RDs. The search results and sources of evidence selection process were reported in accordance with the PRISMA 2020

XSL•FO
RenderX

flow diagram [25], allowing the presentation of the number of sources of evidence included at each stage of the review. The reasons for excluding each article were duly documented.

Data Charting Process

Data were charted from the included sources of evidence using a structured spreadsheet developed by the research team. Data charting was performed by 1 reviewer and verified by a second reviewer. Discrepancies were resolved through discussion. No attempts were made to contact study authors for additional information.

Data Items

The extracted data included authors, title, year of publication, origin/geographic context, study purposes, methods, and key findings for each included source of evidence. The complete dataset is provided in Multimedia Appendix 4.

Critical Appraisal of Individual Sources of Evidence

A formal critical appraisal of the included sources of evidence was not conducted, as the objective of this scoping review was to map and synthesize the available literature rather than assess methodological quality.

Synthesis of Results

The findings from the included sources of evidence were organized into thematic categories based on similarities identified during the data extraction process and through consensus among the authors. Each category consisted of specific elements reported across the included sources. Descriptive statistics were used to summarize the frequency and proportion of outcome measures, allowing comparison of patterns and convergence of information across different health care contexts.

This approach facilitated the synthesis and organization of information through tabulation. Based on these categories, we calculated the proportions of sources of evidence that indicated convergence with the established recommendations. The organized data were then analyzed and interpreted, considering that each selected source of evidence could have mentioned elements across multiple categories. The resulting thematic categories and their respective components are presented in Table 1.

Table 1. Thematic categories and their components identified in the included sources of evidence.

Category	Key points
Professional education and training in RD[a]	• Knowledge gaps among health care professionals regarding RDs • Lack of continuous education and technical training • Inclusion of RD topics in health care curricula within undergraduate and postgraduate programs • Organization of workshops, seminars, courses, and training sessions based on up-to-date evidence • Implementation of continuing education initiatives • Encouragement of the involvement of subject-matter experts in the educational process to enhance academic and professional training
Use of digital technologies and online platforms	• Implementation of digital health functionalities in health care processes related to RDs • Development and maintenance of digital platforms as repositories of validated evidence for consultation • Strategic use of social media for communication, reducing information inequality, and promoting health and awareness • Strengthening records in health information systems
Public policies, government actions, and cross-sectoral integration	• Development of effective public policies for RDs • Government investment and funding for health education programs • Integration between the health and education sectors • Development of strategies at regional, international, and global levels • Creation of research networks and standardization of procedures • Use and improvement of management and health indicators
Education and awareness for the general population	• Promotion of educational initiatives for patients, families, and associations • Dissemination of validated public health information to the general population • Distribution of content through the internet and printed informational materials • Use of clear, accessible, and transparent language
Emotional support and sharing of experiences	• Promotion of emotional support among professionals, patients, families, and associations • Sharing of experiences among professionals, patients, families, and associations as a form of empowerment
Awareness events and commemorative dates	• Organization and implementation of public events and campaigns • Establishment of specific dates to raise awareness for the RDs cause • Awareness-raising actions for social engagement

[a]RD: rare disease.

XSL•FO
RenderX

Results

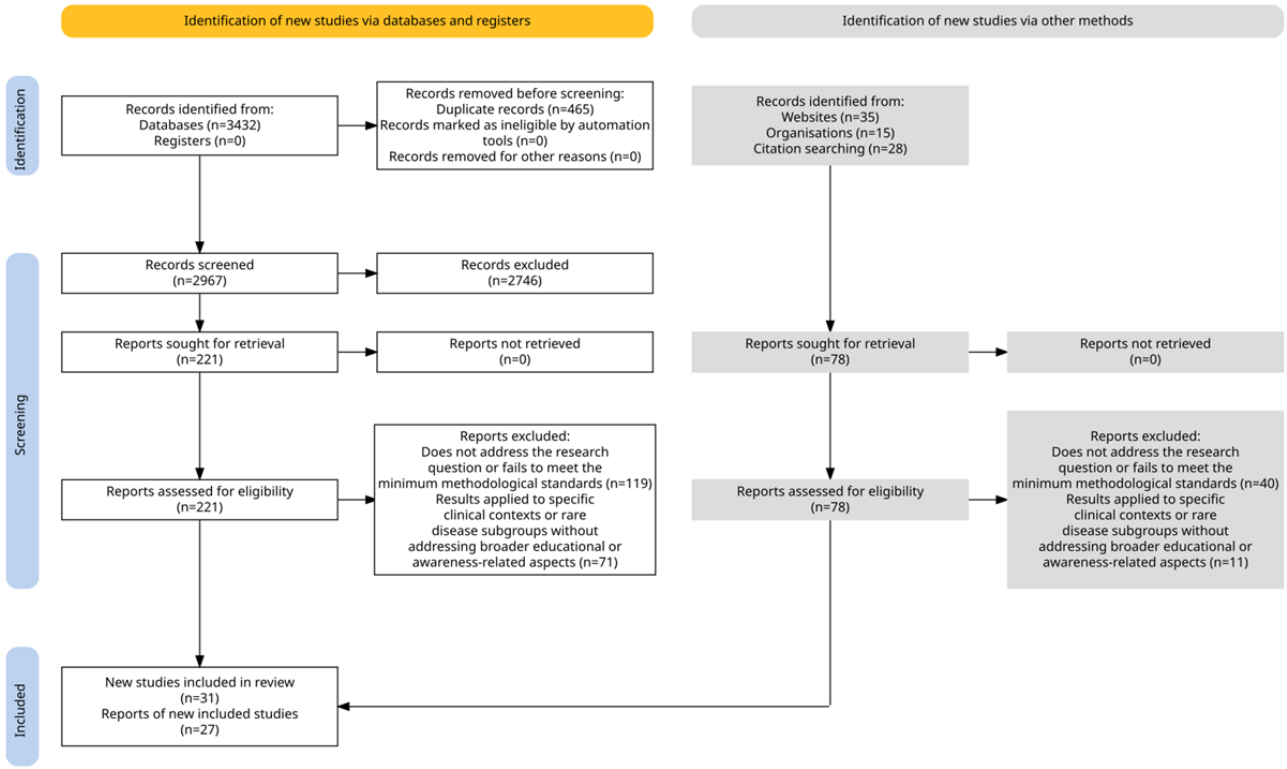
Selection of Sources of Evidence

Initially, a total of 3432 records were identified across 4 selected databases (PubMed/MEDLINE, n=1348; Embase, n=1643; Scopus, n=351; Web of Science, n=90). After removing 465 duplicates, a total of 2967 records remained for title and abstract screening. Of these, 221 full-text articles were assessed for eligibility. Following full-text review, 119 studies were excluded for not addressing the research question or for lacking the scientific rigor required to meet the objectives of this investigation, and 71 were excluded for being applied to specific clinical contexts or specific RD subgroups without addressing broader educational or awareness-related aspects. Hence, 31 studies were included in the review from the databases.

Regarding gray literature, 50 records were initially identified, including 35 from websites and 15 from documents produced by institutional organizations. Additionally, citation searching identified 28 records. In total, 78 records were identified through other methods. All records were screened and assessed for eligibility according to the predefined criteria.

Among these 78 records, 40 records were excluded because they did not address the research question or failed to meet the minimum methodological standards, and 11 were excluded because their results applied to specific clinical contexts or RD subgroups without addressing broader educational or awareness-related aspects. As a result, 27 sources of evidence identified through other methods were included. In total, 58 sources of evidence were included in the scoping review (Multimedia Appendix 4 provides the complete list of included sources). The selection process is illustrated in Figure 1.

Figure 1. PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) 2020 flow diagram.



Characteristics of Sources of Evidence

After selecting the sources of evidence included in the scoping review, relevant data were extracted to address the research question and systematically organize the main health education and awareness recommendations within the scope of RDs. The general characteristics of the included sources of evidence are presented in Multimedia Appendix 4.

Critical Appraisal Within Sources of Evidence

No results of critical appraisal are reported because no formal critical appraisal of the included sources of evidence was conducted.

Results of Individual Sources of Evidence

The detailed outcomes reported by each included source of evidence are presented in Multimedia Appendix 4, which summarizes the main results extracted from the included records.

Synthesis of Results

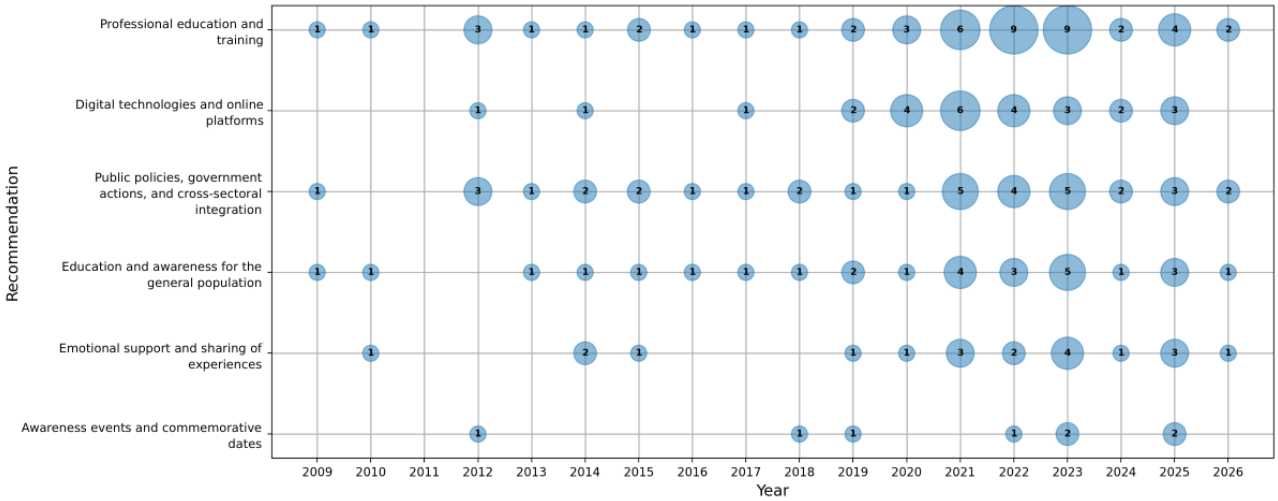
Of the 58 selected sources of evidence, and considering that a single record could report information relevant to more than 1 category, professional education and training in RDs was the most frequently reported recommendation (49/58, 84.4%), followed by public policies, government actions, and cross-sectoral integration (36/58, 62%), education and awareness for the general population (28/58, 48.2%), the use of digital technologies and online platforms (27/58, 46.5%), emotional support and experience sharing (20/58, 34.4%), and awareness

XSL•FO
RenderX

events and commemorative dates (8/58, 13.7%). Percentages are not mutually exclusive.

As presented in Figure 2, there is a clear increase in the frequency of recommendations over time, particularly in the categories of professional education and training and digital technologies and online platforms, especially after 2019. Public policies, government actions, and cross-sectoral integration also show a consistent presence across the years, with moderate growth in recent periods. Recommendations related to education and awareness for the general population appear with a more irregular distribution, while those related to emotional support and experience sharing and awareness events and commemorative dates occur less frequently compared with the other recommendations.

Figure 2. Temporal distribution of recommendations identified in the included sources of evidence.



Regarding the geographic distribution of this body of work, approximately 67.2% (39/58) of the sources of evidence were conducted at the country level. Examples include sources conducted in Poland [29-32], Germany [33-35], Australia [36-40], United States [41-43], Belgium [44], Canada [45], China [46-49], Brazil [50,51], Spain [52,53], Tanzania [54], Mexico [55], Bahrain [56], Bulgaria [57], Iran [58], India [59], and Kazakhstan [60]. Other sources addressed broader geographical regions, including Europe [61-67], Latin America [68], Asia-Pacific [69], Southern Africa [70], North America [71], as well as globally oriented records [16,72-78]. A comprehensive view of the geographic distribution of the included sources is presented in Table 2.

XSL•FO
RenderX

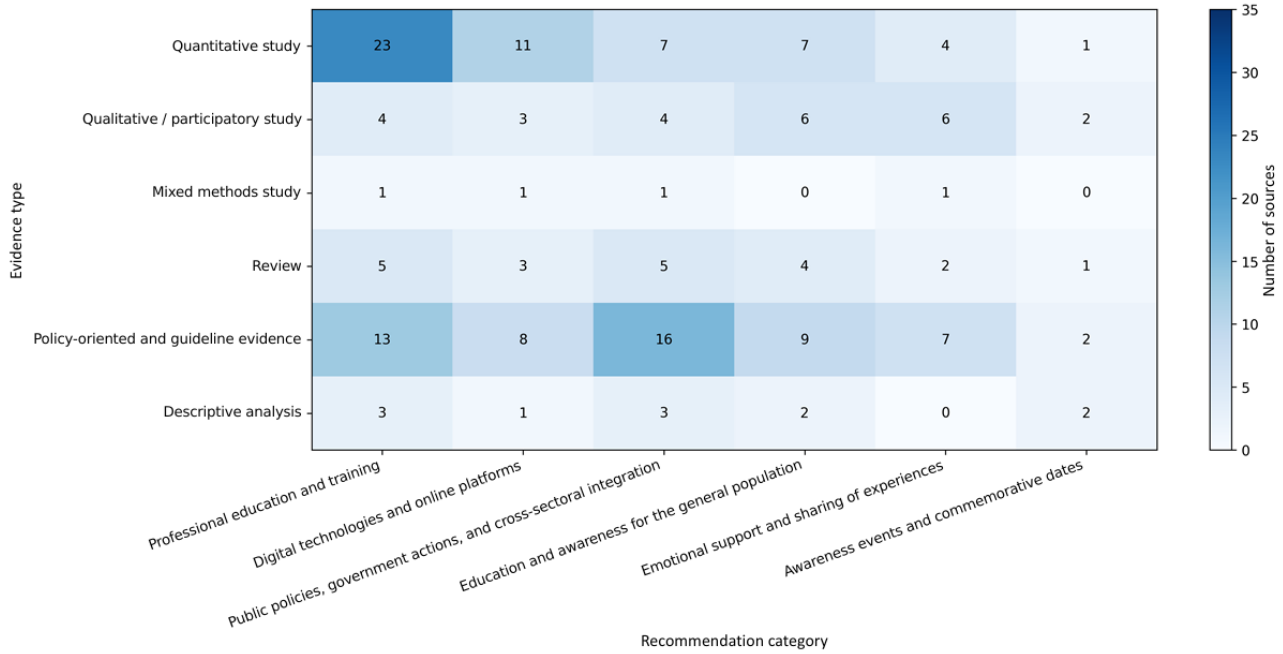
Table 2. Distribution of sources of evidence by country and region.

Country/region	Sources, n
Global	8
Multicountry (Europe)	7
Poland	5
Australia	5
China	4
Germany	3
United States	3
Brazil	2
Spain	2
Turkey	2
Peru	2
United Kingdom	1
Kazakhstan	1
Tanzania	1
Belgium	1
Mexico	1
Bahrain	1
Multicountry (Latin America)	1
Multicountry (Asia-Pacific)	1
Multicountry (Southern Africa)	1
Bulgaria	1
Canada	1
Iran	1
India	1
Serbia	1
Multicountry (North America)	1

A higher concentration was observed in globally oriented and multicountry sources in Europe, while individual countries such as Poland, Australia, China, Germany, and the United States contributed substantially. Conversely, the presence of numerous countries and regions represented by only 1 or 2 sources suggests a fragmented geographic distribution of the evidence. Additionally, to facilitate synthesis and interpretation of the findings, study designs were aggregated into broader evidence-type categories, and their distribution across recommendation categories is presented in Figure 3.

XSL•FO
RenderX

Figure 3. Distribution of recommendation categories across evidence types.



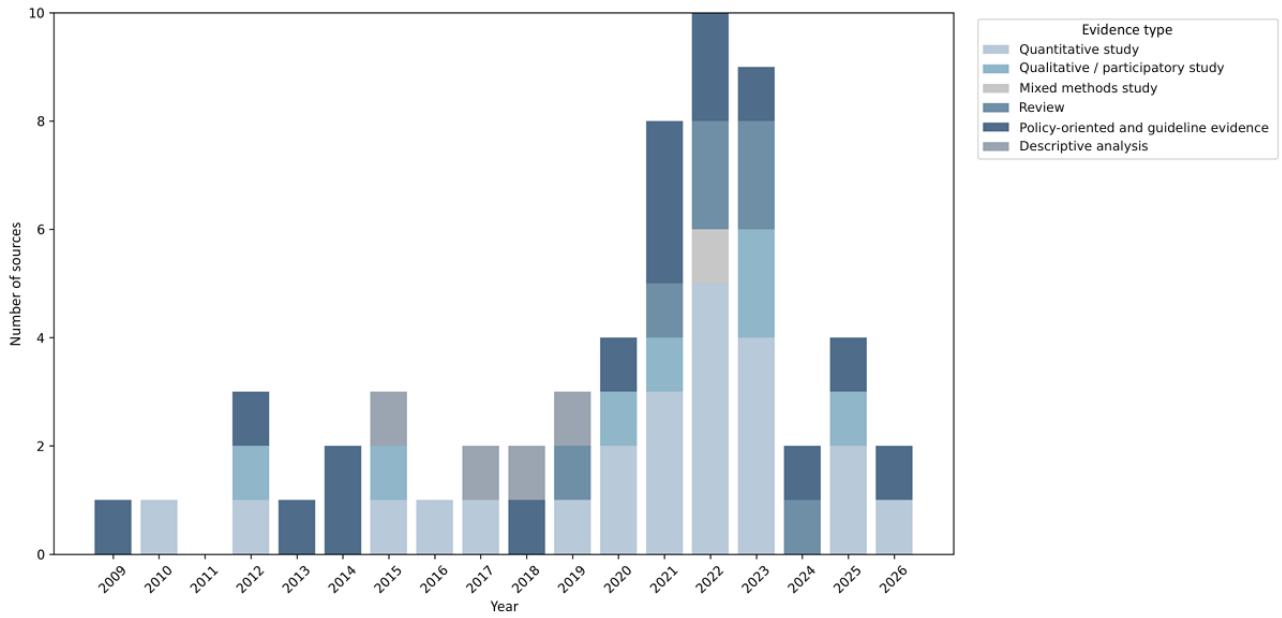
Quantitative studies were frequently represented across several recommendation categories, particularly in professional education and training, which showed the highest number of quantitative sources and demonstrated relatively broad methodological diversity. These findings suggest that professional education and training are among the most consolidated areas within the literature and that the training and capacity building of health care professionals have been widely prioritized as a central strategy in the context of RDs.

Recommendations related to public policies, government actions, and cross-sectoral integration were strongly associated with policy-oriented and guideline evidence, indicating substantial involvement of institutional and governmental actors in shaping recommendations in this area. Recommendations related to emotional support and experience sharing showed a meaningful contribution from qualitative and participatory studies, reflecting the relevance of patient-centered and experience-based approaches to understanding the social and emotional dimensions associated with RDs. Policy-oriented and guideline evidence, however, was also frequently represented within this category.

A generally low representation of mixed methods studies was observed across all recommendation categories, suggesting limited integration between methodological approaches and highlighting a potential gap in the field. Awareness events and commemorative dates represented the least frequently reported category across nearly all evidence types, which may indicate comparatively lower attention to campaign-based strategies within the literature. Meanwhile, recommendations related to digital technologies and online platforms demonstrated a heterogeneous methodological distribution, encompassing quantitative, qualitative, review-based, as well as policy-oriented and guideline evidence, suggesting growing multidisciplinary interest in the use of digital tools to support education and awareness strategies for RDs. Overall, the coexistence of diverse types of evidence reinforces the multidimensional and interdisciplinary nature of health education and awareness strategies in the context of RDs. The temporal distribution of evidence types across the included sources is illustrated in Figure 4.

XSL•FO
RenderX

Figure 4. Temporal evolution of evidence types identified in the included sources of evidence.



A progressive increase in evidence production over time was observed, particularly from 2019 onward, with a peak between 2021 and 2023. Quantitative studies represented the predominant evidence type throughout most of the temporal series, especially in the most recent period. Policy-oriented and guideline evidence, as well as qualitative and participatory studies, became more frequently represented after 2020. Overall, the findings indicate a gradual diversification of evidence types over time, with recent years showing broader methodological diversity and greater coexistence of quantitative, qualitative, review-based, and policy-oriented and guideline evidence.

Concerning the thematic content of the recommendations, many sources highlighted a lack of structured knowledge about RDs among health care professionals and students across various roles, as well as other informational gaps related to insufficient training, capacity building, and continuing education opportunities for these professionals [33,35,40,50,56,58,59,79,80]. It is therefore recommended that curricular content on RDs be strengthened and expanded within undergraduate and postgraduate programs [29,30,32,43,46,50,55,58,71].

In addition, continuing education initiatives, including targeted training and capacity-building courses for practicing professionals, should be implemented, and workshops grounded in up-to-date scientific evidence and clinical guidelines should also be regularly offered to support continuous knowledge development [16,18,36,60,62,65,69,70,76]. Furthermore, the involvement of subject-matter experts in this educational process is encouraged to enhance academic and professional training [57,61].

Another issue raised was the need to leverage digital health functionalities to support RD processes, as well as the development, maintenance, and use of digital platforms and websites as repositories of validated evidence and knowledge for consultation purposes, in addition to strengthening records in health information systems, while internet communication channels and social media were also highlighted as active tools that can be used by health agencies and councils to address gaps in knowledge and access to information and to raise awareness about the topic [18,34,36-38,53,65,68,79].

The recommendations also emphasize the development of effective public health policies, increased government funding, closer integration between the health and education sectors, and coordinated regional, international, and global strategies [18,36,37,46,47,54,64,67,74]. Additionally, they highlight the creation of research networks and committees, the standardization of procedures, the improvement of indicators and system records, and the definition of best practices in management, organization, and innovation [51,54,55,61,62,69,72,74,81].

The need for education and awareness initiatives on RDs for the general population was also mentioned, emphasizing the dissemination of validated and relevant information to patients, families, and associations through educational actions and programs, including making such content available and distributing educational and informational materials in different formats in a clear and accessible way [35,42,46,63-65,67,74,82].

Emotional and psychological support, as well as the sharing of experiences among health care professionals, patients, families, and associations, were also identified as important factors in this scenario [16,34,35,40,42,45,51,53,75]. In addition, the organization of awareness events and the establishment of specific dates were highlighted as strategies to strengthen the RDs agenda and draw the attention of society to this public health issue [16,18,38,49,53,55,56,75].

Discussion

Summary of Evidence

The main findings of this scoping review indicate that recommendations for developing educational and awareness strategies in the context of RDs focus on professional education

XSL•FO
RenderX

and training; use of digital technologies and online platforms; public policies, government actions, and cross-sectoral integration; education and awareness for the general population; emotional support and sharing of experiences; and awareness events and commemorative dates. In line with the objective of identifying and systematizing the main recommendations for health education and awareness in the context of RDs, the evidence suggests that improving knowledge dissemination and awareness in this field requires integrated approaches that combine clinical education, accessible information systems, institutional coordination, and patient-centered support strategies. Collectively, these dimensions reinforce that education and awareness in RDs extend beyond information delivery alone and depend on coordinated actions involving health care systems, governments, educational institutions, patient organizations, and society.

The evidence synthesized in this scoping review indicates that education and awareness processes related to RDs remain fragmented and insufficiently structured in many health care settings, creating significant barriers to diagnosis, management, and continuity of care. Several included sources highlighted persistent knowledge gaps among health care professionals and medical students, particularly regarding the recognition, diagnosis, and management of these conditions [29,30,32,35,43,44,46-48,50,52,56-60,62,71,80,83,84]. These findings reflect broader structural limitations within professional training programs, where RD-related content is often underrepresented in undergraduate and continuing education curricula. Consequently, physicians, nurses, and students frequently report low confidence in their ability to identify and manage RDs, compromising timely diagnosis and coordinated care [29,50,52,60]. These findings support previous recommendations emphasizing the need for systematic educational reforms and continuous professional development initiatives specifically tailored to RDs [16,36,37,45,50,54,64,68,70,75].

Within this context, our findings also highlight the growing importance of digital technologies and online knowledge platforms as tools to support education, diagnostic processes, and access to information related to RDs [18,37,40,44,79]. The lack of accessible and reliable information remains one of the major barriers to early and accurate diagnosis [16,55,63,68,69,79], and digital resources have increasingly emerged as valuable tools to address this gap. Structured platforms and ontologies such as the Human Phenotype Ontology, the Orphanet Rare Disease Ontology, and PubCaseFinder exemplify this movement by enabling standardized phenotypic descriptions, facilitating symptom-based searches, and supporting diagnostic processes through evidence-based systems [85-87]. Beyond their clinical applications, these resources have progressively been incorporated into educational settings, helping health care professionals and students improve the structured description and interpretation of phenotypes and supporting diagnostic investigation [88-90].

The use of online educational resources also appears to strengthen information dissemination and patient support. Sources included in this scoping review emphasize the relevance of websites, digital repositories, and online educational materials as accessible resources for patients, caregivers, and professionals [34,38,48,77]. In many situations, online information platforms initially represent the primary source of knowledge for patients and families before they establish contact with specialists or support organizations [34]. These findings are consistent with broader global trends in digital health and reinforce the importance of developing accessible, validated, and user-centered information systems capable of supporting both educational processes and clinical decision-making [91].

Another important dimension identified in this study involves the role of public policies, institutional coordination, and cross-sectoral collaboration in strengthening RD education and awareness strategies. The included sources consistently emphasize the importance of improving patient registries, fostering collaborative networks, funding research, and standardizing clinical and managerial processes [40,51,61-63,66,69]. These actions are aligned with international recommendations proposed by the WHO, which advocate for comprehensive and integrated policies intended to improve access to care, monitor health care indicators, and strengthen national and regional health systems for RDs [92]. In addition, national and international initiatives have highlighted the importance of communication strategies that promote accessible and inclusive health information tailored to different audiences, including patients, families, health care professionals, and the general population [34,35,41,65].

The findings further suggest that awareness and educational efforts in the context of RDs should not be restricted to clinical or institutional dimensions alone. Emotional and psychological support emerged as essential components for improving quality of life and strengthening care pathways for patients and families affected by RDs [18,35,40,53,73,82]. In this context, experience sharing among patients, caregivers, and health care professionals plays a fundamental role in building support networks, reducing social isolation, and fostering resilience. Social media platforms such as YouTube, Facebook, and Instagram have increasingly become important spaces for information exchange, peer support, and patient engagement, particularly for individuals facing prolonged diagnostic journeys and limited access to specialized services [93-95].

The integration of patient support groups and participatory educational strategies also appears to strengthen patient empowerment and improve the relevance of communication initiatives. Previous studies emphasize that health care professionals can contribute to these processes by moderating discussions, supporting the dissemination of reliable information, and facilitating the referral of newly diagnosed patients to support communities [96,97]. Furthermore, participatory approaches involving patients and caregivers in the codevelopment of educational resources and awareness strategies may enhance communication effectiveness and ensure that educational initiatives better reflect the lived experiences and practical needs of individuals affected by RDs [45]. These findings reinforce the importance of adopting collaborative and patient-centered models in RD education and awareness initiatives.

XSL•FO
RenderX

Finally, this scoping review highlights the relevance of awareness events, campaigns, and commemorative dates in increasing the visibility of RDs and promoting broader societal engagement [18,37,38,49,54,55]. Public awareness actions, including initiatives such as Rare Disease Day, represent important opportunities to mobilize policymakers, health care institutions, educational organizations, patient associations, and the general public around the challenges associated with RDs. In parallel, collaboration between health and education systems may contribute to the development of more sustainable and coordinated awareness strategies, capable of integrating scientific knowledge, professional training, public engagement, and patient advocacy efforts [18,37].

Limitations

Among the limitations of this study, the variability in the territorial scope of the included sources may limit the generalizability of the findings to settings with different income levels or less structured health care systems. In addition, differences were identified in study designs, data sources, methodological approaches, and outcome measures (Multimedia Appendix 4 provides full details). These factors may influence the interpretation and comparability of findings and should therefore be considered when applying the results of this scoping review.

Nevertheless, many of the principles identified, such as the need for intersectoral policies, government actions, professional education and capacity building, emotional support and the sharing of experiences, and increased education and awareness among the general population, are consistent with global guidelines, including those proposed by the WHO for health care management [16,91]. However, the feasibility and implementation of some of the identified recommendation categories may vary across resource and income settings and may require contextual adaptation according to regional and institutional characteristics. In particular, the use of digital technologies and online platforms depends on the availability of appropriate infrastructure and may require adaptation in resource-constrained settings [2,16,46,91].

Similarly, integration across referral networks, specialized centers, and cross-sectoral systems may pose additional implementation challenges [98]. Professional education and training are widely recognized as key strategies across different contexts but may vary in effectiveness depending on how well they are designed and delivered to meet the needs of different target audiences [29,44,46,50,77]. In addition, recommendations related to public awareness, emotional support, experience sharing, and awareness events are widely reported and implemented in a variety of contexts [34,41,45,49,65,66,68].

Additionally, although a formal assessment of methodological quality among the included sources is not standard practice in scoping reviews, we acknowledge that the robustness of the included evidence varies across sources. In this context, future research could build upon the evidence mapped in this study by conducting approaches such as systematic reviews that critically assess the methodological quality of included sources, using appraisal tools such as the Critical Appraisal Skills Programme [99] or A Measurement Tool to Assess Systematic Reviews-2 [100].

The findings presented in this study may therefore serve as a preliminary foundation for further exploration in this field. Furthermore, the language restriction applied during the selection process may have limited access to relevant sources published in excluded languages, particularly those originating from Eastern regions or from countries underrepresented in international databases. However, as much of the available evidence is commonly disseminated in English and indexed in major databases, the impact of this limitation may be partially mitigated.

Despite these limitations, the findings of this study contribute to expanding the current body of evidence on recommendations for educational and awareness strategies related to RDs, underscoring the relevance of integrated, multisectoral approaches. By highlighting the intersection between health, education, communication, policy, and society, this review reinforces the need to engage diverse social actors in the design and implementation of effective and context-sensitive initiatives. Ultimately, strengthening such collaborative efforts may support improved knowledge dissemination, earlier recognition, more inclusive care pathways, and more informed health planning and decision-making in the complex context of RDs.

Conclusions

This scoping review mapped and synthesized existing recommendations related to the development of health education and awareness strategies for RDs. The findings contribute to understanding how these strategies have been addressed in the literature, highlighting important gaps and variations across different contexts. They may support evidence-informed decision-making and public health planning, particularly regarding persistent bottlenecks in diagnostic processes and care coordination. Given the complexity of RD care, the scarcity of available knowledge associated with these conditions, and their substantial societal impact, strengthening health education and awareness strategies may contribute to broader public health efforts aimed at improving access to information, continuity of care, and health care management in the context of RDs.

Future research should explore the effectiveness and implementation of specific health education and awareness strategies for RDs across different contexts. More focused systematic reviews may also assess the impact of particular interventions, building upon the body of evidence identified in this scoping review while critically evaluating the methodological quality of included sources. Overall, by consolidating dispersed and heterogeneous evidence into a coherent body of knowledge, this scoping review contributes to advancing the field and may inform improvements in health services, professional and managerial practices, and initiatives aimed at supporting patients, families, and advocacy groups involved in RDs.

XSL•FO
RenderX

Acknowledgments

The authors declare that generative artificial intelligence (AI) was used in the research and writing process. In accordance with the Generative AI Disclosure Taxonomy (GAIDeT) (2025), the following tasks were delegated to generative AI tools under full human supervision: data curation and organization, visualization, proofreading and editing, text summarization, translation, and reformatting. The generative AI tool used was ChatGPT-5.5 (OpenAI). Responsibility for the final manuscript rests entirely with the authors. Generative AI tools are not listed as authors and bear no responsibility for the final outcomes. This declaration is submitted under collective responsibility.

Data Availability

Data generated or analyzed during this study are included in this published article and its supplementary information files. Additional datasets generated or analyzed during this study are available from the corresponding author on reasonable request.

Funding

This study is supported by the São Paulo Research Foundation (Fundação de Amparo à Pesquisa do Estado de São Paulo [FAPESP]) as part of the Public Policy Research Program (grant number 2023/10203-8, managed by DA). DBY is also supported by FAPESP (grant number 2024/06751-2).
The funder had no involvement in the study design, data collection, analysis, interpretation, or writing of the manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
PRISMA-ScR checklist.
[DOCX File , 86 KB - mededu_v12i1e79027_app1.docx]

Multimedia Appendix 2
Full search strategies for databases.
[DOCX File , 1635 KB - mededu_v12i1e79027_app2.docx]

Multimedia Appendix 3
PRISMA-S checklist.
[DOCX File , 17 KB - mededu_v12i1e79027_app3.docx]

Multimedia Appendix 4
The general characteristics of the included sources.
[XLSX File (Microsoft Excel File), 24 KB - mededu_v12i1e79027_app4.xlsx]

References

1. Health education: theoretical concepts, effective strategies and core competencies: a foundation document to guide capacity development of health educators. World Health Organization Regional Office for the Eastern Mediterranean. 2012. URL: https://applications.emro.who.int/dsaf/emrpub_2012_en_1362.pdf [accessed 2026-02-18]
2. Alves D, Yamada DB, Bernardi FA, Carvalho I, Filho MEC, Neiva MB. Mapping, infrastructure, and data analysis for the Brazilian network of rare diseases: protocol for the RARASnet observational cohort study. JMIR Res Protoc 2021;10(1):e24826 [FREE Full text] [doi: 10.2196/24826] [Medline: 33480849]
3. Félix TM, de Oliveira BM, Artifon M, Carvalho I, Bernardi FA, Schwartz IVD. Epidemiology of rare diseases in Brazil: protocol of the Brazilian rare diseases network (RARAS-BRDN). Orphanet J Rare Dis 2022;17(1):84 [FREE Full text] [doi: 10.1186/s13023-022-02254-4] [Medline: 35209917]
4. Helderman R, Macica CM, Weinstein A, Feinn R, Doyle M. Addressing challenges in diagnosis and management of rare disease through interprofessional education. Rare 2024;2:100044. [doi: 10.1016/j.rare.2024.100044]
5. Rare Diseases Act of 2002. United States Congress. URL: https://www.congress.gov/congressional-report/107th-congress/house-report/543 [accessed 2025-04-24]
6. Moliner A, Waligora J. The European Union policy in the field of rare diseases. Adv Exp Med Biol 2017;1031:561-587. [doi: 10.1007/978-3-319-67144-4_30] [Medline: 29214592]
7. Política nacional de atenção integral às pessoas com doenças raras: Portaria GM/MS nº 199, de 30 de janeiro de 2014. Ministério da Saúde, Brazil. 2014. URL: https://bvsms.saude.gov.br/bvs/saudelegis/gm/2014/prt0199_30_01_2014.html [accessed 2023-07-11]

XSL•FO
RenderX

8. Rare diseases. World Health Organization. 2023. URL: https://www.who.int/standards/classifications/frequently-asked-questions/rare-diseases [accessed 2023-07-11]
9. Find information on medicines. EURORDIS Rare Diseases Europe. 2023. URL: https://www.eurordis.org/information-support/find-information-on-treatments/ [accessed 2025-06-05]
10. Nguengang Wakap S, Lambert DM, Olry A, Rodwell C, Gueydan C, Lanneau V. Estimating cumulative point prevalence of rare diseases: analysis of the orphanet database. Eur J Hum Genet 2020;28(2):165-173. [doi: 10.1038/s41431-019-0508-0] [Medline: 31527858]
11. Valdez R, Ouyang L, Bolen J. Public health and rare diseases: oxymoron no more. Prev Chronic Dis 2016;13:E05 [FREE Full text] [doi: 10.5888/pcd13.150491] [Medline: 26766846]
12. European Reference Networks. European Commission. 2025. URL: https://health.ec.europa.eu/rare-diseases-and-european-reference-networks/european-reference-networks_en [accessed 2025-05-14]
13. Rare Diseases Clinical Research Network. URL: https://www.rarediseasesnetwork.org/ [accessed 2025-05-14]
14. Gonzaga-Jauregui C, Salazar C, MacDonald J, Reichardt JK, Groft SC. ERCAL, a regional initiative for rare diseases in Latin America and the Caribbean. Rare Dis Orphan Drugs J 2024;3(1):6. [doi: 10.20517/rdodj.2023.48]
15. Rare disease. Asia Pacific Alliance of Rare Disease Organisations (APARDO). URL: https://www.apardo.org/ [accessed 2025-05-14]
16. Rare diseases: a global health priority for equity and inclusion. World Health Organization. 2025. URL: https://apps.who.int/gb/ebwha/pdf_files/EB156/B156_CONF2-en.pdf [accessed 2026-02-18]
17. Pelentsov LJ, Laws TA, Esterman AJ. The supportive care needs of parents caring for a child with a rare disease: a scoping review. Disabil Health J 2015;8(4):475-491 [FREE Full text] [doi: 10.1016/j.dhjo.2015.03.009] [Medline: 25959710]
18. Tumiene B, Peters H, Melegh B, Peterlin B, Utkus A, Fatkulina N. Rare disease education in Europe and beyond: time to act. Orphanet J Rare Dis 2022;17(1):441 [FREE Full text] [doi: 10.1186/s13023-022-02527-y] [Medline: 36536417]
19. Rare disease impact report: insights from patients and the medical community. Shire Human Genetic Technologies. 2013. URL: https://globalgenes.org/wp-content/uploads/2013/04/ShireReport-1.pdf [accessed 2025-06-05]
20. Bauskis A, Strange C, Molster C, Fisher C. The diagnostic odyssey: insights from parents of children living with an undiagnosed condition. Orphanet J Rare Dis 2022;17(1):233 [FREE Full text] [doi: 10.1186/s13023-022-02358-x] [Medline: 35717227]
21. Taruscio D, Salvatore M, Lumaka A, Carta C, Cellai LL, Ferrari G. Undiagnosed diseases: needs and opportunities in 20 countries participating in the undiagnosed diseases network international. Front Public Health 2023;11:1079601 [FREE Full text] [doi: 10.3389/fpubh.2023.1079601] [Medline: 36935719]
22. Bernardi FA, Mello de Oliveira B, Bettiol Yamada D, Artifon M, Schmidt AM, Machado Scheibe V. The minimum data set for rare diseases: systematic review. J Med Internet Res 2023;25:e44641 [FREE Full text] [doi: 10.2196/44641] [Medline: 37498666]
23. Adams DR, van Karnebeek CDM, Agulló SB, Faùndes V, Jamuar SS, Lynch SA. Addressing diagnostic gaps and priorities of the global rare diseases community: recommendations from the IRDiRC diagnostics scientific committee. Eur J Med Genet 2024;70:104951 [FREE Full text] [doi: 10.1016/j.ejmg.2024.104951] [Medline: 38848991]
24. Aromataris E, Munn Z. JBI Manual for Evidence Synthesis. Adelaide: JBI (Joanna Briggs Institute); 2024.
25. Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ 2021;372:n71 [FREE Full text] [doi: 10.1136/bmj.n71] [Medline: 33782057]
26. Tricco AC, Lillie E, Zarin W, O'Brien KK, Colquhoun H, Levac D. PRISMA extension for scoping reviews (PRISMA-ScR): checklist and explanation. Ann Intern Med 2018;169(7):467-473 [FREE Full text] [doi: 10.7326/M18-0850] [Medline: 30178033]
27. Medical subject headings (MeSH) home page. National Library of Medicine. 2025. URL: https://www.nlm.nih.gov/mesh/meshhome.html [accessed 2025-04-24]
28. Rethlefsen ML, Kirtley S, Waffenschmidt S, Ayala AP, Moher D, Page MJ. PRISMA-S: an extension to the PRISMA statement for reporting literature searches in systematic reviews. J Med Libr Assoc 2021;109(2):174-200 [FREE Full text] [doi: 10.5195/jmla.2021.962] [Medline: 34285662]
29. Walkowiak D, Domaradzki J. Needs assessment study of rare diseases education for nurses and nursing students in Poland. Orphanet J Rare Dis 2020;15(1):167 [FREE Full text] [doi: 10.1186/s13023-020-01432-6] [Medline: 32600383]
30. Walkowiak D, Domaradzki J. Are rare diseases overlooked by medical education? Awareness of rare diseases among physicians in Poland: an explanatory study. Orphanet J Rare Dis 2021;16(1):400 [FREE Full text] [doi: 10.1186/s13023-021-02023-9] [Medline: 34583737]
31. Kope G, Podolec C. Establishing a curriculum on rare diseases for medical students. J Rare Cardiovasc Dis 2015;2(3):74-76 [FREE Full text] [doi: 10.20418/jrcd.vol2no3.194]
32. Jonas K, Waligóra M, Hołda M, Sulicka-Grodzicka J, Strach M, Podolec P, et al. Knowledge of rare diseases among health care students – the effect of targeted education. Przegl Epidemiol 2017;71(1):80-89 [FREE Full text] [Medline: 28742309]
33. Kessel M, Hannemann-Weber H, Kratzer J. Innovative work behavior in healthcare: the benefit of operational guidelines in the treatment of rare diseases. Health Policy 2012;105(2-3):146-153 [FREE Full text] [doi: 10.1016/j.healthpol.2012.02.010] [Medline: 22405486]

XSL•FO
RenderX

34. Litzkendorf S, Frank M, Babac A, Rosenfeldt D, Schauer F, Hartz T. Use and importance of different information sources among patients with rare diseases and their relatives over time: a qualitative study. BMC Public Health 2020;20(1):860 [FREE Full text] [doi: 10.1186/s12889-020-08926-9] [Medline: 32503483]
35. von Gizycki R. Contribution of rare disease patient organisations to medical education. Orphanet J Rare Dis 2010;5(Suppl 1):O25. [doi: 10.1186/1750-1172-5-s1-o25]
36. Bhattacharya K, Millis N, Jaffe A, Zurynski Y. Rare diseases research and policy in Australia: on the journey to equitable care. J Paediatr Child Health 2021;57(6):778-781. [doi: 10.1111/jpc.15507] [Medline: 33861492]
37. Saggu H, Jones C, Lewis A, Baynam G. mEDUrare: supporting integrated care for rare diseases by better connecting health and education through policy. Yale J Biol Med 2021;94(4):693-702 [FREE Full text] [Medline: 34970108]
38. Molster C, Youngs L, Hammond E, Dawkins H, National Rare Diseases Coordinating Committee, National Rare Diseases Working Group. Key outcomes from stakeholder workshops at a symposium to inform the development of an Australian national plan for rare diseases. Orphanet J Rare Dis 2012;7:50 [FREE Full text] [doi: 10.1186/1750-1172-7-50] [Medline: 22883422]
39. National strategic action plan for rare diseases. Australian Government Department of Health and Aged Care. 2020. URL: https://www.health.gov.au/resources/publications/national-strategic-action-plan-for-rare-diseases [accessed 2026-02-18]
40. National recommendations for rare disease health care. Rare Disease Awareness, Education, Support and Training (RArEST) Project. 2024. URL: https://rarevoices.org.au/wp-content/uploads/2026/04/National-Recommendations-for-Rare-Disease-Health-Care_03.26.pdf [accessed 2026-04-26]
41. Plackowski EF, Bogart KR. "If not me, who?": Awareness- and self-advocacy-related experiences of adults with diverse rare disorders. Qual Health Res 2023;33(1-2):63-80. [doi: 10.1177/10497323221135974] [Medline: 36444970]
42. Munro M, Cook AM, Bogart KR. An inductive qualitative content analysis of stigma experienced by people with rare diseases. Psychol Health 2022;37(8):948-963. [doi: 10.1080/08870446.2021.1912344] [Medline: 33886382]
43. Morgenthau A, Margus C, Mackley MP, Miller AP. Rare disease education outside of the classroom and clinic: evaluation of the RARE compassion program for undergraduate medical students. Genes (Basel) 2022;13(10):1707 [FREE Full text] [doi: 10.3390/genes13101707] [Medline: 36292592]
44. Vandeborne L, van Overbeeke E, Dooms M, De Beleyr B, Huys I. Information needs of physicians regarding the diagnosis of rare diseases: a questionnaire-based study in Belgium. Orphanet J Rare Dis 2019;14(1):99 [FREE Full text] [doi: 10.1186/s13023-019-1075-8] [Medline: 31054581]
45. Quintal A, Carreau I, Grenier A, Hébert C, Yergeau C, Berthiaume Y, et al. An ethics action plan for rare disease care: participatory action research approach. J Particip Med 2023;15:e46607 [FREE Full text] [doi: 10.2196/46607] [Medline: 37995128]
46. Chung CCY, Hong Kong Genome Project, Chu ATW, Chung BHY. Rare disease emerging as a global public health priority. Front Public Health 2022;10:1028545 [FREE Full text] [doi: 10.3389/fpubh.2022.1028545] [Medline: 36339196]
47. Li X, Zhang X, Zhang S, Lu Z, Zhang J, Zhou J. Rare disease awareness and perspectives of physicians in China: a questionnaire-based study. Orphanet J Rare Dis 2021;16(1):171 [FREE Full text] [doi: 10.1186/s13023-021-01788-3] [Medline: 33849615]
48. Zhou L, Xu J, Yang J. Poor education and urgent information need for emergency physicians about rare diseases in China. Orphanet J Rare Dis 2022;17(1):211 [FREE Full text] [doi: 10.1186/s13023-022-02354-1] [Medline: 35619153]
49. Huang R, Wei Y, Hu J, Kong F, He J, Yang Y. The progress of, challenges faced by, and future of rare disease patient organizations in China. Intractable Rare Dis Res 2019;8(2):158-160 [FREE Full text] [doi: 10.5582/irdr.2019.01069] [Medline: 31218170]
50. Melo DG, de Paula PK, de Araujo Rodrigues S, da Silva de Avó LR, Germano CMR, Demarzo MMP. Genetics in primary health care and the national policy on comprehensive care for people with rare diseases in Brazil: opportunities and challenges for professional education. J Community Genet 2015;6(3):231-240 [FREE Full text] [doi: 10.1007/s12687-015-0224-6] [Medline: 25893505]
51. Diretrizes para atenção integral às pessoas com doenças raras no Sistema Único de Saúde (SUS). Ministério da Saúde, Brazil. 2014. URL: https://bvsms.saude.gov.br/bvs/publicacoes/diretrizes_atencao_integral_pessoa_doencas_raras_SUS.pdf [accessed 2026-02-18]
52. Ramalle-Gómara E, Domínguez-Garrido E, Gómez-Eguílaz M, Marzo-Sola ME, Ramón-Trapero JL, Gil-de-Gómez J. Education and information needs for physicians about rare diseases in Spain. Orphanet J Rare Dis 2020;15(1):18 [FREE Full text] [doi: 10.1186/s13023-019-1285-0] [Medline: 31952528]
53. Hernández-Rodríguez J, Martínez-Valle F, Acebes X, Alerany C, Antón J, Calvo G. Identification of strengths and weaknesses of the healthcare system for persons living with rare diseases in Catalonia (Spain), and recommendations to improve its comprehensive attention: the "acERca las enfermedades raras" project. Orphanet J Rare Dis 2025;20(1):42 [FREE Full text] [doi: 10.1186/s13023-024-03518-x] [Medline: 39875900]
54. Kaywanga F, Alimohamed MZ, David AB, Maeda D, Mbarak S, Mavura T. Rare diseases in Tanzania: a national call for action to address policy and urgent needs of individuals with rare diseases. Orphanet J Rare Dis 2022;17(1):343 [FREE Full text] [doi: 10.1186/s13023-022-02498-0] [Medline: 36064429]

XSL•FO
RenderX

55. Rivera-Silva G, Treviño-de la Fuente F, Treviño-Alanís MG. [Rare diseases in Mexico]. Rev Med Inst Mex Seguro Soc 2018;56(3):214-215 [FREE Full text] [Medline: 30365280]
56. Sinan I, Mihdawi M, Farahat AR, Fida M. Knowledge and awareness of rare diseases among healthcare professionals in the Kingdom of Bahrain. Cureus 2023;15(10):e47676 [FREE Full text] [doi: 10.7759/cureus.47676] [Medline: 38022232]
57. Hristova-Atanasova E, Iskrov G, Atanasov I, Genc A, Stefanov R. What is the awareness of rare diseases among medical students? A survey in Bulgaria. Orphanet J Rare Dis 2023;18(1):213 [FREE Full text] [doi: 10.1186/s13023-023-02820-4] [Medline: 37491304]
58. Jahanshahi R, Nasirzadeh A, Farzan M, Domaradzki J, Jouybari L, Sanagoo A. Iranian future healthcare professionals' knowledge and opinions about rare diseases: cross-sectional study. Orphanet J Rare Dis 2022;17(1):366 [FREE Full text] [doi: 10.1186/s13023-022-02458-8] [Medline: 36175905]
59. Vashishta L, Bapat P, Bhattacharya Y, Choudhury MC, Chirmule N, D'Costa S. A survey of awareness of diagnosis and treatment of rare diseases among healthcare professionals and researchers in India. J Biosci 2023;48:37 [FREE Full text] [Medline: 37795705]
60. Walkowiak D, Bokayeva K, Miraleyeva A, Domaradzki J. The awareness of rare diseases among medical students and practicing physicians in the Republic of Kazakhstan. An exploratory study. Front Public Health 2022;10:872648 [FREE Full text] [doi: 10.3389/fpubh.2022.872648] [Medline: 35462837]
61. Taruscio D, Gentile AE, Evangelista T, Frazzica RG, Bushby K, Montserrat AM. Centres of expertise and European reference networks: key issues in the field of rare diseases. The EUCERD Recommendations. Blood Transfus 2014;12 Suppl 3(Suppl 3):s621-S625 [FREE Full text] [doi: 10.2450/2014.0026-14s] [Medline: 24922304]
62. Taruscio D, Gentile AE, De Santis M, Ferrelli RM, Posada de la Paz M, Hens M. EUROPLAN: a project to support the development of national plans on rare diseases in Europe. Public Health Genomics 2013;16(6):278-287. [doi: 10.1159/000355932] [Medline: 24503588]
63. Council recommendation of 8 June 2009 on an action in the field of rare diseases. Council of the European Union. 2009. URL: https://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:C:2009:151:0007:0010:EN:PDF [accessed 2026-02-12]
64. Syed AM, Camp R, Mischorr-Boch C, Houÿez F, Aro AR. Policy recommendations for rare disease centres of expertise. Eval Program Plann 2015;52:78-84. [doi: 10.1016/j.evalprogplan.2015.03.006] [Medline: 25935363]
65. De Santis M, Hervas C, Weinman A, Bosi G, Bottarelli V. Patient empowerment of people living with rare diseases. Its contribution to sustainable and resilient healthcare systems. Ann Ist Super Sanita 2019;55(3):283-291 [FREE Full text] [doi: 10.4415/ANN_19_03_15] [Medline: 31553324]
66. Recommendations from the Rare 2030 Foresight Study: the future of rare diseases starts today. EURORDIS - Rare Diseases Europe. 2026. URL: https://download2.eurordis.org/rare2030/Rare2030_recommendations.pdf [accessed 2026-02-18]
67. Lynn S, Hedley V, Atalaia A, Evangelista T, Bushby K, EUCERD Joint Action. How the EUCERD joint action supported initiatives on rare diseases. Eur J Med Genet 2017;60(3):185-189. [doi: 10.1016/j.ejmg.2017.01.002] [Medline: 28087401]
68. Wainstock D, Katz A. Advancing rare disease policy in Latin America: a call to action. Lancet Reg Health Am 2023;18:100434 [FREE Full text] [doi: 10.1016/j.lana.2023.100434] [Medline: 36844013]
69. Global rare disease policy network. APEC Action Plan on Rare Diseases. URL: https://www.rarediseasepolicy.org/apec/english [accessed 2026-02-12]
70. David A, Alimohamed MZ, Modern G, Mbarak S, Mad-hy H, Özcan B. The plight of rare diseases in Southern Africa: health and social services policy recommendations. Future Rare Dis 2025;6(1). [doi: 10.1080/23995270.2025.2607961]
71. Ayyappan V, Gonzalez E, Pichette É, Spahic H, Guzman S. Spotlighting the zebras: a role for medical students in shaping rare disease care. J Vasc Anomalies 2022;3(1):e034. [doi: 10.1097/jova.0000000000000034]
72. Gittus M, Chong J, Sutton A, Ong ACM, Fotheringham J. Barriers and facilitators to the implementation of guidelines in rare diseases: a systematic review. Orphanet J Rare Dis 2023;18(1):140 [FREE Full text] [doi: 10.1186/s13023-023-02667-9] [Medline: 37286999]
73. Yabumoto M, Miller E, Rao A, Tabor HK, Ormond KE, Halley MC. Perspectives of rare disease social media group participants on engaging with genetic counselors: mixed methods study. J Med Internet Res 2022;24(12):e42084 [FREE Full text] [doi: 10.2196/42084] [Medline: 36542454]
74. Groft SC, Posada M, Taruscio D. Progress, challenges and global approaches to rare diseases. Acta Paediatr 2021;110(10):2711-2716. [doi: 10.1111/apa.15974] [Medline: 34105798]
75. Adachi T, El-Hattab AW, Jain R, Nogales Crespo KA, Quirland Lazo CI, Scarpa M. Enhancing equitable access to rare disease diagnosis and treatment around the world: a review of evidence, policies, and challenges. Int J Environ Res Public Health 2023;20(6):4732 [FREE Full text] [doi: 10.3390/ijerph20064732] [Medline: 36981643]
76. van Karnebeek CD, O'Donnell-Luria A, Baynam G, Baudot A, Groza T, Jans JJM. Leaving no patient behind! Expert recommendation in the use of innovative technologies for diagnosing rare diseases. Orphanet J Rare Dis 2024;19(1):357 [FREE Full text] [doi: 10.1186/s13023-024-03361-0] [Medline: 39334316]
77. Rohani-Montez SC, Bomberger J, Zhang C, Cohen J, McKay L, Evans WR. Educational needs in diagnosing rare diseases: a multinational, multispecialty clinician survey. Genet Med Open 2023;1(1):100808 [FREE Full text] [doi: 10.1016/j.gimo.2023.100808] [Medline: 39669242]

XSL•FO
RenderX

78. Forman J, Taruscio D, Llera VA, Barrera LA, Coté TR, Edfjäll C. The need for worldwide policy and action plans for rare diseases. Acta Paediatr 2012;101(8):805-807 [FREE Full text] [doi: 10.1111/j.1651-2227.2012.02705.x] [Medline: 22519914]
79. Durmus S, Yucesan E, Aktug S, Utz B, Caglayan AO, Gencpinar P. Management of rare and undiagnosed diseases: insights from researchers and healthcare professionals in Türkiye. Front Public Health 2024;12:1501942 [FREE Full text] [doi: 10.3389/fpubh.2024.1501942] [Medline: 39911789]
80. Domaradzki J, Walkowiak D. Knowledge and attitudes of future healthcare professionals toward rare diseases. Front Genet 2021;12:639610 [FREE Full text] [doi: 10.3389/fgene.2021.639610] [Medline: 34122502]
81. Abarca-Barriga HH, Alvariño Tello R, Laso-Salazar MC, La Serna-Infantes J, Castro Mujica MDC. Knowledge and attitudes toward rare diseases in Lima, Peru: a cross-sectional study on public awareness of healthcare. Future Rare Dis 2026;6(1):2621388. [doi: 10.1080/23995270.2026.2621388]
82. Gunes D, Karaca M, Durmus A, Ak B, Aktay Ayaz N, Altınel ZU. Challenges in the clinical management of rare diseases and center-based multidisciplinary approach to creating solutions. Eur J Pediatr 2025;184(5):281. [doi: 10.1007/s00431-025-06101-z] [Medline: 40186762]
83. Medić B, Divac N, Stopić N, Savić-Vujović K, Glišić A, Cerovac N. The attitudes of medical students towards rare diseases: a cross-sectional study. Vojnosanit Pregl 2016;73(8):703-713. [doi: 10.2298/VSP150326094M] [Medline: 29328568]
84. Flores A, Burgos S, Abarca-Barriga H. Knowledge level of medical students and physicians about rare diseases in Lima, Peru. Intractable Rare Dis Res 2022;11(4):180-188 [FREE Full text] [doi: 10.5582/irdr.2022.01079] [Medline: 36457581]
85. Gargano MA, Matentzoglu N, Coleman B, Addo-Lartey EB, Anagnostopoulos AV, Anderton J. The human phenotype ontology in 2024: phenotypes around the world. Nucleic Acids Res 2024;52(D1):D1333-D1346 [FREE Full text] [doi: 10.1093/nar/gkad1005] [Medline: 37953324]
86. Orphanet Rare Disease Ontology (ORDO). Institut national de la santé et de la recherche médicale (INSERM). URL: https://www.orphadata.com/ordo/ [accessed 2025-06-05]
87. PubCaseFinder: search system for rare genetic diseases. Database Center for Life Science. URL: https://pubcasefinder.dbcls.jp/ [accessed 2025-06-05]
88. Sergouniotis PI, Maxime E, Leroux D, Olry A, Thompson R, Rath A. An ontological foundation for ocular phenotypes and rare eye diseases. Orphanet J Rare Dis 2019;14(1):8 [FREE Full text] [doi: 10.1186/s13023-018-0980-6] [Medline: 30626441]
89. Köhler S, Gargano M, Matentzoglu N, Carmody LC, Lewis-Smith D, Vasilevsky NA. The human phenotype ontology in 2021. Nucleic Acids Res 2021;49(D1):D1207-D1217 [FREE Full text] [doi: 10.1093/nar/gkaa1043] [Medline: 33264411]
90. Fujiwara T, Shin JM, Yamaguchi A. Advances in the development of PubCaseFinder, including the new application programming interface and matching algorithm. Hum Mutat 2022;43(6):734-742 [FREE Full text] [doi: 10.1002/humu.24341] [Medline: 35143083]
91. Global strategy on digital health 2020-2025. World Health Organization. 2021. URL: https://www.who.int/publications/i/item/9789240020924 [accessed 2025-06-05]
92. World health statistics 2021: monitoring health for the SDGs, sustainable development goals. World Health Organization. URL: https://digitallibrary.un.org/record/3935247 [accessed 2025-06-05]
93. Davis DZ, Calitz W. Finding healthcare support in online communities: an exploration of the evolution and efficacy of virtual support groups. JVWR 2014;7(3) [FREE Full text]
94. Morgan G, Subbiah V. Rare disease research powered by empowered patients: solving the zebra puzzle through social media. Cancer 2023;129(14):2128-2131 [FREE Full text] [doi: 10.1002/cncr.34765] [Medline: 37011025]
95. Social media: research and data. Pew Research Center. URL: https://www.pewresearch.org/topic/internet-technology/technology-policy-issues/social-media/ [accessed 2026-02-12]
96. Ashtari S, Taylor A. Patients with rare diseases and the power of online support groups: implications for the medical community. JMIR Form Res 2023;7:e41610 [FREE Full text] [doi: 10.2196/41610] [Medline: 37707878]
97. Deuitch NT, Beckman E, Halley MC, Young JL, Reuter CM, Kohler J. "Doctors can read about it, they can know about it, but they've never lived with it": how parents use social media throughout the diagnostic odyssey. J Genet Couns 2021;30(6):1707-1718 [FREE Full text] [doi: 10.1002/jgc4.1438] [Medline: 34096130]
98. Sun L, Booth A, Sworn K. Adaptability, scalability and sustainability (ASaS) of complex health interventions: a systematic review of theories, models and frameworks. Implement Sci 2024;19(1):52 [FREE Full text] [doi: 10.1186/s13012-024-01375-7] [Medline: 39020399]
99. CASP checklists. Critical Appraisal Skills Programme (CASP). URL: https://casp-uk.net/casp-tools-checklists/ [accessed 2025-06-05]
100. Shea BJ, Reeves BC, Wells G, Thuku M, Hamel C, Moran J. AMSTAR 2: a critical appraisal tool for systematic reviews that include randomised or non-randomised studies of healthcare interventions, or both. BMJ 2017;358:j4008 [FREE Full text] [doi: 10.1136/bmj.j4008] [Medline: 28935701]

Abbreviations

MeSH: Medical Subject Headings
PCC: Population, Concept, and Context

XSL•FO
RenderX

PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
PRISMA-S: Preferred Reporting Items for Systematic Reviews and Meta-Analyses Literature Search Extension
PRISMA-ScR: Preferred Reporting Items for Systematic Reviews and Meta-Analyses Extension for Scoping Reviews
RD: rare disease
WHO: World Health Organization

Edited by S Zelko, S Brini; submitted 13.Jun.2025; peer-reviewed by K Kostadinov, N Mofolo; comments to author 05.Nov.2025; accepted 15.May.2026; published 12.Jun.2026.

Please cite as:
Yamada DB, Komoto TT, Lima VC, Souza J, Seixas MDS, Cassão V, Ferraz VEDF, de Oliveira BM, Félix TM, Alves D
Main Recommendations for Developing Education and Awareness Strategies for Rare Diseases: Scoping Review
JMIR Med Educ 2026;12:e79027
URL: https://mededu.jmir.org/2026/1/e79027
doi:10.2196/79027
PMID:

©Diego Bettiol Yamada, Tatiana Takahasi Komoto, Vinícius Costa Lima, Julio Souza, Michele de Souza Seixas, Victor Cassão, Victor Evangelista de Faria Ferraz, Bibiana Mello de Oliveira, Têmis Maria Félix, Domingos Alves. Originally published in JMIR Medical Education (https://mededu.jmir.org), 12.Jun.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Education, is properly cited. The complete bibliographic information, a link to the original publication on https://mededu.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Review

Toolkits to Support Health Care Providers in Health Behavior Change for Patients With Chronic Conditions: Scoping Review

Jillian Scandiffio[1,2], MSc; Robert Simpson[1,3], MD, PhD; Ashvene Sureshkumar[1,2], PhD; Jerry Tang[4], BSc; Paloma Fleming[5], BMedSci; Dina Gaid[2], PhD; Harzaan Gnanakaran[6]; Marina Wasilewski[1,7], PhD; Nicole Woods[8,9], PhD; Sarah J Donkers[10], PhD; Sarah Munce[1,9,11], PhD

[1]Rehabilitation Sciences Institute, University of Toronto, Toronto, ON, Canada
[2]KITE Research Institute, Toronto Rehabilitation Institute, University Health Network, Toronto, ON, Canada
[3]St. Michael's Hospital, Unity Health Toronto, Toronto, ON, Canada
[4]Faculty of Medicine, University of Ottawa, Ottawa, ON, Canada
[5]Schulich School of Medicine & Dentistry, Western University, London, ON, Canada
[6]Faculty of Science, McMaster University, Hamilton, ON, Canada
[7]St. John's Rehab Research Program, Sunnybrook Health Science Centre, Toronto, ON, Canada
[8]The Institute for Education Research, University Health Network, Toronto, ON, Canada
[9]Institute of Health Policy, Management, and Evaluation, University of Toronto, Toronto, ON, Canada
[10]School of Rehabilitation Science, University of Saskatchewan, Toronto, SK, Canada
[11]Bloorview Research Institute, Holland Bloorview Kids Rehabilitation Hospital, Toronto, ON, Canada

Corresponding Author:
Jillian Scandiffio, MSc
Rehabilitation Sciences Institute
University of Toronto
500 University Ave
Toronto, ON, M5G 1V7
Canada
Phone: 1 4163604000 ext 40047
Email: jillian.scandiffio@mail.utoronto.ca

Abstract

Background: Health care providers (HCPs) are key agents in supporting health behavior (eg, smoking and physical activity) change in patients. Toolkits have previously been used to guide HCPs in care and have shown effectiveness in improving clinical outcomes. However, there is a lack of understanding of how toolkits should be developed to best support HCPs.

Objective: This scoping review aimed to map the existing literature on the theories, models, frameworks, approaches, and methods that have been used in the development of toolkits to support HCPs in health behavior change.

Methods: Five databases (MEDLINE, Embase, CINAHL, Web of Science, and Scopus) were searched from inception to August 2026. Original research articles published in peer-reviewed academic journals in English from any geographic area describing the development of a toolkit to support HCPs in health behavior change for patients with chronic conditions were included. Qualitative data were analyzed using conventional content analysis and quantitative data were examined in tabular terms, with a simple descriptive synthesis of the findings. Barriers and facilitators identified from studies were deductively coded using the Theoretical Domains Framework (TDF) and reported according to TDF domains identified in studies.

Results: Nineteen articles met inclusion criteria. The articles included toolkits addressing several health behaviors, including physical activity (n=16), nutrition (n=6), sedentary behavior (n=2), sleep (n=2), and weight management more broadly (n=2). Toolkit formats (eg, use of videos and written components) varied greatly between studies. All studies included knowledge users in the development of the toolkit. Use of a theory, model, or framework was described in 14 studies, and barriers and facilitators to HCP promotion of health behavior change were examined in 14 studies. The most frequently reported TDF domain was environmental context and resources, with time as the most frequently cited barrier. Toolkit development typically consisted of an exploratory phase to determine knowledge users' needs and wants, toolkit development, dissemination, and evaluation.

XSL•FO
RenderX

Conclusions: Toolkits varied greatly in format and development process but generally involved iterative development based on feedback from knowledge users. However, no toolkits targeted smoking, alcohol use, or recreational drug use, which warrant further study. To assist future investigators looking to develop a toolkit for health behavior change, we present a definition of a toolkit and a series of recommendations for toolkit design and development. Although previous reviews have mapped the existence of various toolkits, they have not examined how these have been developed. This review provides a novel description of this development and a guide for the development of future toolkits to increase uptake of toolkits and reduce the knowledge-to-practice gap. Additionally, this review defines the term "toolkit" and identifies its minimum components, which can be used by researchers to classify and compare toolkits in the future.

Trial Registration: OSF Registries osf.io/bstmy; https://osf.io/bstmy/overview

(JMIR Med Educ 2026;12:e88653) doi:10.2196/88653

KEYWORDS

toolkit; health behavior; development; design; knowledge translation; scoping review

Introduction

Rationale

Lifestyle behaviors (eg, smoking, physical activity, and nutrition) are indicators of population health more broadly [1-4] and are associated with a range of health outcomes, including comorbidity (eg, cardiovascular disease, diabetes, and cancer) [5,6], health service use [7-9], and quality of life [10]. These behaviors are multifactorial and include a series of behaviors addressing both physical and psychosocial health [11]. Health care providers (HCPs) are key agents in supporting health behavior change in a variety of clinical populations [12-14]. However, there is evidence to suggest that HCPs' ability to address health behaviors in day-to-day care is hindered by multiple external factors, including, but not limited to, lack of confidence in counseling patients about behavior change, lack of time to discuss behavior change, and health system gaps [15].

It is important to develop user-friendly, evidence-based tools that support HCPs in promoting health behavior change in their patients. Toolkits are collections of adaptable resources to inform and provide guidance that are typically intended to bridge the evidence-to-practice gap, support implementation and adoption, and/or support sustained use of the intended change [16]. The definition of what encompasses a "toolkit" is somewhat heterogeneous, as toolkits range greatly in modality, formatting, and content and may include individual documents or combinations of documents such as timelines, educational materials, and assessment instruments [16,17]. Toolkits can be used by themselves as a knowledge mobilization strategy or can be part of a larger multistrategy intervention [17]. Toolkits have shown effectiveness in improving clinical outcomes (eg, reduced patient falls, decreased hospital length of stay, and improved diabetes management) in a range of patient populations (eg, cancer, diabetes, and depression) [17-19]. The wide range of toolkit designs suggests that, before a toolkit is developed and implemented, there should be a clear understanding of how it should be developed. However, there has been limited evidence on how toolkits have been developed in this context. An evidence-informed approach to toolkit design may improve toolkit engagement, uptake, and effectiveness.

While scoping reviews have been conducted on toolkits in health care [17,20], there is a lack of information regarding the design and development of these toolkits. When referring to design, this includes considerations of what the toolkit should look like and its components, including how the modality of the toolkit will be used and understood by end users. When referring to development, this includes considerations of the process of creating the toolkit, including stages such as identifying knowledge gaps, adapting the toolkit to local needs, and implementing the toolkit. In short, when we refer to design, we are referring to the what (the toolkit) and when we refer to development, we are referring to the how (the process of creating the toolkit). Additionally, there have been recommendations for development of a toolkit in areas outside of health behavior change (eg, quality improvement) [21], but there is a lack of information regarding best practices for the development of toolkits to support HCPs in health behavior change. Further, there is a lack of information regarding best practices for the development of toolkits to support HCPs in health behavior change. For example, a recent review examined toolkits for physical activity adherence in patients with chronic conditions and noted that the process of toolkit development was not specified in any of the included studies [22]. This includes a lack of information about theories, models, or frameworks (TMFs) that guide toolkit development. It is important to consider how and whether TMFs have been used in toolkit development, as understanding the theoretical approaches used can enhance understanding of why implementation succeeds or fails [23]. Further, using TMFs is crucial for enabling knowledge translation and implementation of interventions into practice [24]. Thus, there is a gap in understanding best practices for the development of a toolkit to support HCPs in addressing health behavior change, which this review aimed to address. By gaining a deeper understanding of the optimal approaches to toolkit development, we can better design future toolkits for HCPs. The findings from this review may act as a resource for investigators looking to develop future toolkits for HCPs in health behavior change and more broadly.

Objective

The objective of this scoping review was to map the existing literature on the TMFs, approaches, and methods used in the development of toolkits to support HCPs in health behavior change.

XSL•FO
RenderX

Methods

Protocol and Registration

This review was undertaken in accordance with the Joanna Briggs Institute (JBI) methodological guidance for conducting scoping reviews [25]. The protocol was registered with the Open Science Framework on November 22, 2024, and published [26]. There were no deviations from the registered protocol. The reporting of the review follows the PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses Extension for Scoping Reviews) guidelines (Multimedia Appendix 1) [27].

Eligibility Criteria

Overview

Studies were included in the review based on population, concept, and context (PCC) criteria [25].

Types of Participants

When considering population, studies were included that involved a toolkit targeted to HCPs for changing the health behaviour of patients with chronic conditions. Chronic conditions were defined as "conditions that last 1 year or more and require ongoing medical attention or limit activities of daily living or both" [28]. HCPs needed to be identified as end users or the target audience for the toolkit for the study to be included. If a toolkit was meant for use by another end user (eg, patients) and did not note that HCPs were also intended end users, the study was not included.

Concept

The conceptual focus was on the approach to developing toolkits to support HCPs in addressing lifestyle-related health behavior change among their patients with chronic conditions. Health behaviors were categorized according to the categories reported by Statistics Canada [1], including negative health behaviors (smoking or tobacco use, recreational drug use, and alcohol use), movement behaviors (eg, physical activity and sedentary behavior), sleep, and nutrition. Chronic conditions were based on those identified by Statistics Canada and the Public Health Agency of Canada [29,30], with additional chronic conditions included when relevant (eg, a rare chronic neurological condition not explicitly mentioned by the organizations but falling within a broader category of "neurological condition"). Studies were included if they (1) described the development of a toolkit to support HCPs in health behavior change and (2) had a primary focus on the toolkit development approach. Studies in which authors described their intervention as a "tool" or a related term (eg, toolkit and toolbox) were included, but studies where the intervention was not referred to using these terms (eg, package and materials) were excluded.

Context

Included studies were limited to those reporting original research that were published in English in peer-reviewed academic journals.

Types of Evidence Sources

All study designs were considered, provided that they reported primary research data and described the approaches for developing a toolkit.

Information Sources

The search is reported according to the PRISMA-S (PRISMA Search Reporting Extension) [31]. The search strategy was developed in collaboration with an academic librarian and a Peer Review of Electronic Search Strategies (PRESS) [32] was conducted by a second academic librarian. The search was first conducted in MEDLINE (Ovid) and then was translated to the following additional databases: Embase (Ovid), CINAHL (EBSCOhost), Web of Science (Clarivate), and Scopus. Each database was searched separately. The reference lists of all included sources of evidence were screened for additional studies. No additional online or print sources were purposefully searched (eg, table of contents, websites). We searched each database from inception onward, with the initial search conducted in November 2024 and the search was rerun in August 2026. No limits were used in the search. No additional studies or data were sought by contacting others (eg, authors). The search strategy was neither adapted nor reused from other literature reviews. Full search details are available in Multimedia Appendix 2.

Selection of Sources of Evidence

Search results were imported, deduplicated, and screened in Covidence (Veritas Health Innovation). Each title/abstract was reviewed against the inclusion criteria for eligibility by 2 independent reviewers (JS, AS, JT, PF, DG, or HG), with the full texts of potentially eligible studies reviewed in by 2 independent reviewers (JS, AS, JT, PF, DG, or HG) against inclusion criteria. If disagreements occurred, consensus was reached through discussion, with a third reviewer (SM) available for arbitration. To ensure reliability and consistency in screening, a pilot review of 100 of the titles was screened by reviewers to establish interrater reliability, with a Cohen κ of 0.8 as the threshold raters needed to meet, indicating substantial agreement [33].

Data Charting Process

Data were extracted from included articles by 2 independent reviewers using a data extraction tool developed by the research team, guided by the JBI template (Multimedia Appendix 3) [34]. The data extraction tool was developed and piloted at the protocol stage and trialed in the first few studies by 2 independent reviewers to ensure that all relevant results were extracted. Any disagreements that arose between reviewers were resolved through discussion, or by consultation with an additional senior reviewer.

Data Items

Data charted included authors, year of publication, origin/country of publication, country of toolkit development, study aims/purpose, HCP population (eg, nurses and allied health professionals), patient population of interest, PROGRESS+ considerations [35] in toolkit tailoring (eg, gender and race/ethnicity), health behavior, toolkit aim, toolkit

XSL•FO
RenderX

componentry, toolkit development, evaluation process, TMFs applied, methods, approaches, knowledge users involved, outcomes, and authors' recommendations/lessons learned. For this review, a theory is defined as "a set of analytical principles or statements designed to structure our observation, understanding, and explanation of the world," a model is defined as a simplification or representation of a specific aspect of a phenomenon, and a framework as an explanation of a phenomenon through a structure, overview, outline, system, or plan of descriptive categories and the relationships between them [23].

Synthesis of Results

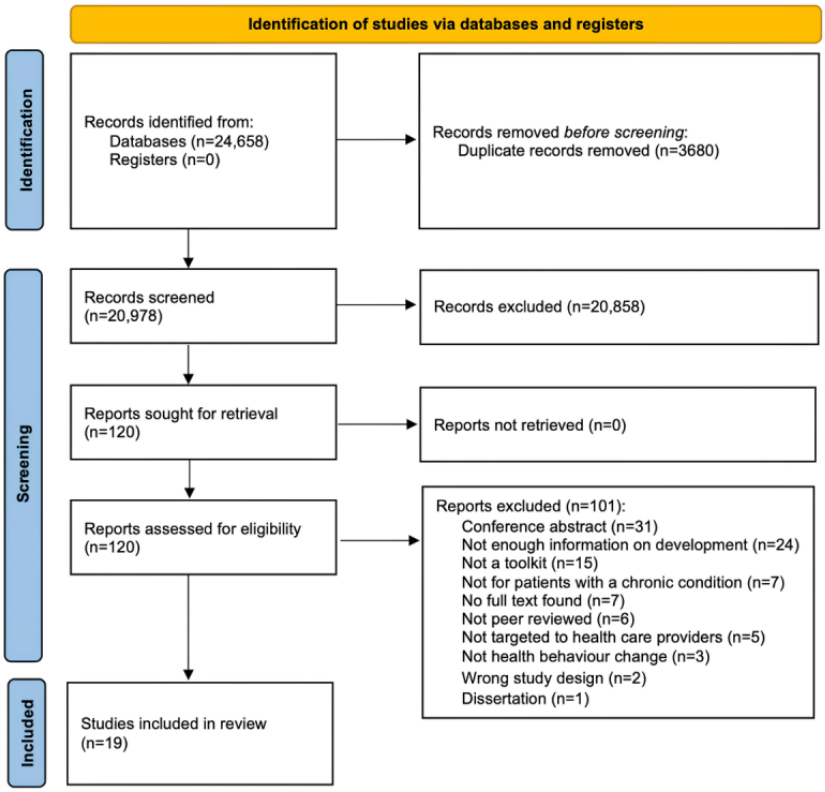
Conventional content analysis [36] was undertaken for qualitative data, with a simple descriptive synthesis of the findings. Quantitative data were examined in tabular terms and reported descriptively. Barriers and facilitators identified from studies were deductively coded using the Theoretical Domains Framework (TDF) [37] and reported according to the TDF domains identified in the studies. This framework was selected because it considers behavior change theories and was originally developed to identify influences on HCP behavior for implementing evidence-based recommendations [38]. A series of recommendations was developed regarding toolkit design and development based on the barriers and facilitators and recommendations described in included studies. The most commonly reported barriers, facilitators, and recommendations across studies were presented to an integrated knowledge translation panel, including people with chronic conditions, implementation scientists, HCPs, and researchers with expertise in health behaviors. Final recommendations for design and development were developed through iterative discussions with panel members, where members agreed on recommendations that should be included and provided feedback on their wording.

Results

Selection of Sources of Evidence

Our database search identified 24,658 studies for possible inclusion (Figure 1). Following deduplication, a total of 20,978 studies were screened for inclusion at the title and abstract stage, and a subsequent 120 full texts were examined. After screening, 19 studies were included in the scoping review [39-57].

Figure 1. PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) flow diagram of included studies.



Characteristics of Sources of Evidence

Included studies were published between 2007 and 2025 (Multimedia Appendix 4). Studies took place in the following countries: Canada (n=7) [39-42,45,47,48], Australia (n=5) [42,49,51,52,57], the United Kingdom (n=4) [43,44,50,54], Netherlands (n=2) [46,56], and the United States (n=2) [53,55]. It should be noted that one of these studies took place in both Canada and Australia [42]. Approximately half of the studies (n=10) [40-43,45,47,48,51,54,56] used mixed or multiple methods and approximately half (n=8) [39,44,46,49,50,52,53,55] used qualitative methods for toolkit development; no studies used only quantitative methods.

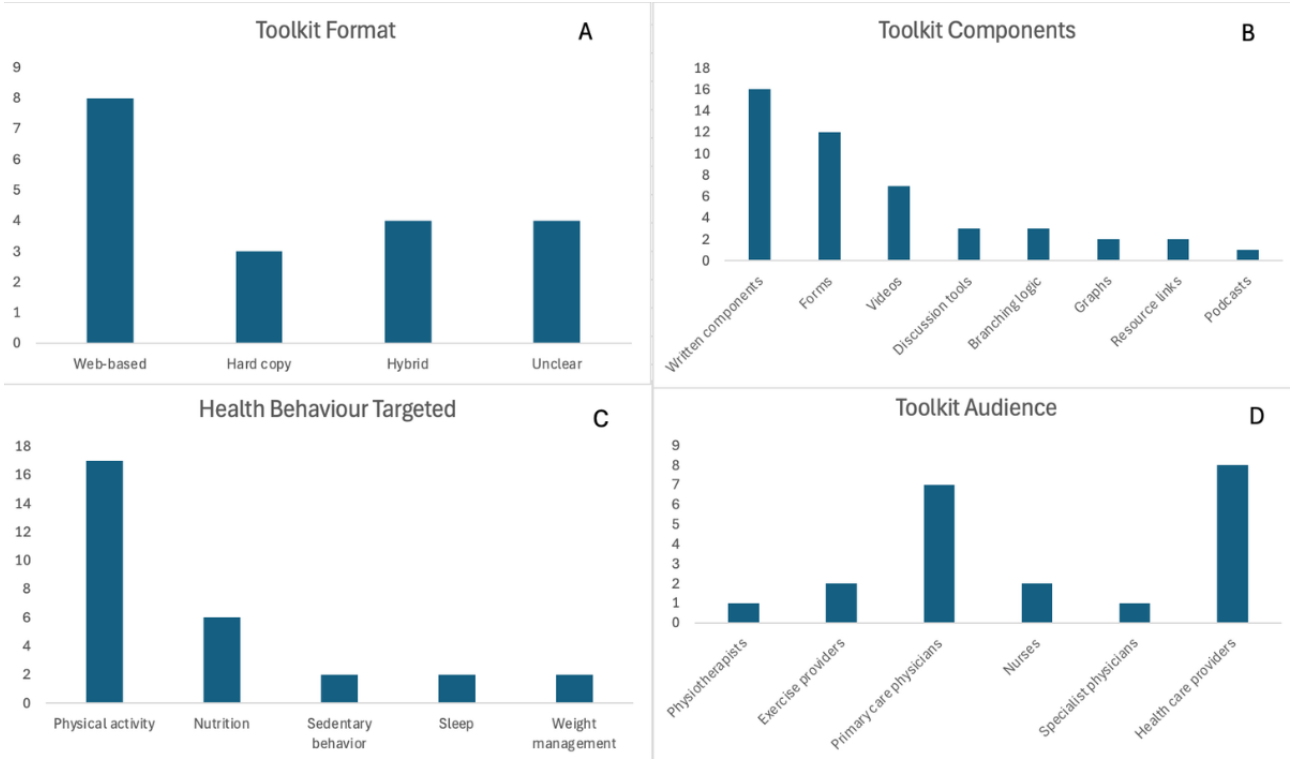
Characteristics of Included Toolkits

When classifying the toolkit, 13 studies used the term "toolkit" when referring to their intervention [39-42,45,47-51,53,54,57], 1 study used the term "toolbox" [52], and 5 studies referred to

XSL•FO
RenderX

it only as a "tool" [43,44,46,55,56]. Only one study provided a citation for the concept of a toolkit [57]. Toolkit modalities varied, including web-based (n=8) [42,45,48-51,55,56], hard-copy (n=3) [44,53,54], and hybrid (n=4) [41,46,47,57] formats. Toolkit modality was unclear or not specified in 4 studies [39,40,43,52]. The toolkits were used in combination with other activities (eg, training program) in 4 studies [53-56] and was the sole intervention in 15 studies [39-52,57]. Selected toolkit characteristics are visually represented in Figure 2.

Figure 2. Frequency of the reported characteristics of included toolkits, including (A) toolkit format, (B) toolkit components, (C) the health behavior targeted by the toolkit, and (D) the toolkit's target audience.



Toolkit components varied greatly between studies. Five studies [39,41,49,50,52] used only one type of component (eg, written information only or forms only), while the other 14 toolkits [40,42-48,51,53-57] were multimodal. Written components such as summaries of evidence were included in 16 toolkits [40,42-55,57]. Further, 12 toolkits included forms for patients and/or HCPs to complete together (eg, SMART forms) [39-44,48,53-57]. Additional toolkit components included videos (n=7) [42,45-48,51,54], discussion tools (eg, conversation dice; n=3) [46,51,53], branching logic to make recommendations on care (n=3) [51,55,57], graphs (n=2) [40,53], links to resources (n=2) [43,56], and podcasts (n=1) [42].

The health behaviors targeted by the toolkit included physical activity/exercise (n=17) [40-52,54-57], nutrition (n=6) [40,41,49-51,55], sedentary behavior (n=2) [41,56], sleep (n=2) [40,41], and unspecified weight management (n=2) [39,53]. Approximately half of the toolkits addressed only one health behavior (physical activity) [42-48,52,54,57] and the other half addressed multiple health behaviors [39-41,49-51,53,55,56]. The HCP audience for the toolkits included physiotherapists (n=1) [46], exercise providers (n=2) [42,45], primary care physicians (n=7) [40,49,50,53-56], nurses (n=2) [50,56], specialist physicians (n=1; cancer care workers) [43], and HCPs more broadly (n=8) [39,41,44,47,48,51,52,57]. The toolkits were targeted to a range of patient populations, including obesity (n=6) [39-41,49,53,55], neurological conditions (n=4) [44,45,52,57], cancer (n=2) [42,43], diabetes/prediabetes (n=2) [50,54], physical disability (n=2) [46,47], depression (n=1) [48], osteoarthritis (n=1) [51], and patients with both type 2 diabetes and chronic obstructive pulmonary disease (n=1) [56].

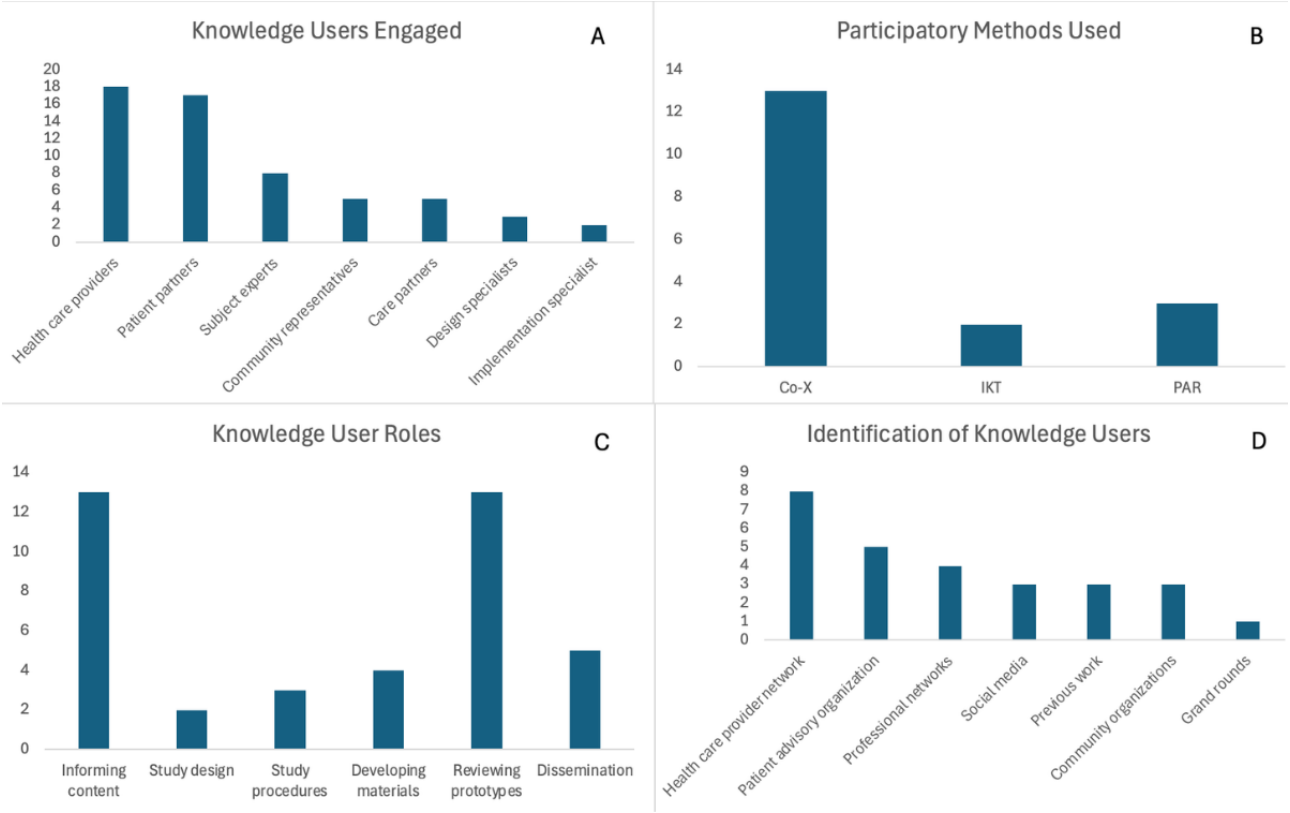
Toolkit Development Process

All studies reported including knowledge users in the development process. Most studies included HCPs (n=18) [39-51,53-57] and patients (n=17) [39,41-52,54-57], with additional knowledge users including subject experts/researchers (n=8) [44-48,50,51,56], community representatives (n=5) [45-49], care partners (n=5) [44-46,53,57], toolkit design specialists (n=3) [39,46,51], and implementation specialists (n=2) [47,48]. Participatory methods were reported in most studies (n=16) [39-48,50-52,54,56,57], including Co-X approaches (ie, co-design, co-development, co-creation, and co-production) methods (n=13) [39-46,50-52,54,57], integrated knowledge translation (n=2) [48,57], and participatory action research (n=3) [45-47]. Roles of knowledge users engaged in participatory methods included informing toolkit content or layout (eg, recommending topics the toolkit should cover and suggesting toolkit presentation; n=13) [39-47,50-52,54], study design (n=2) [41,45], assisting in study procedures (eg, recruitment and analyzing data; n=3) [45,50,56], developing toolkit materials (n=4) [39-41,47], reviewing toolkit prototypes (n=13) [39-48,51,52,57], and dissemination (n=5) [41,43,47,48,52]. Knowledge users were identified from HCP networks (n=8) [40,42-44,49,53,54,57], patient advisory

XSL•FO
RenderX

organizations (n=5) [44,45,51-53], professional networks (n=4) [49,51,56,57], social media (n=3) [42,51,57], previous work (n=3) [51,55,57], community organizations/clinics (n=3) [45,52,57], and HCP grand rounds (n=1) [49]. Figure 3 provides a visual representation of the frequency of reported toolkit development processes.

Figure 3. Frequency of the reported toolkit development process, including (A) knowledge users engaged, (B) participatory methods used in toolkit development, (C) roles of knowledge users engaged in participatory methods, and (D) how knowledge users included in participatory roles were identified. IKT: integrated knowledge translation; PAR: participatory action research.

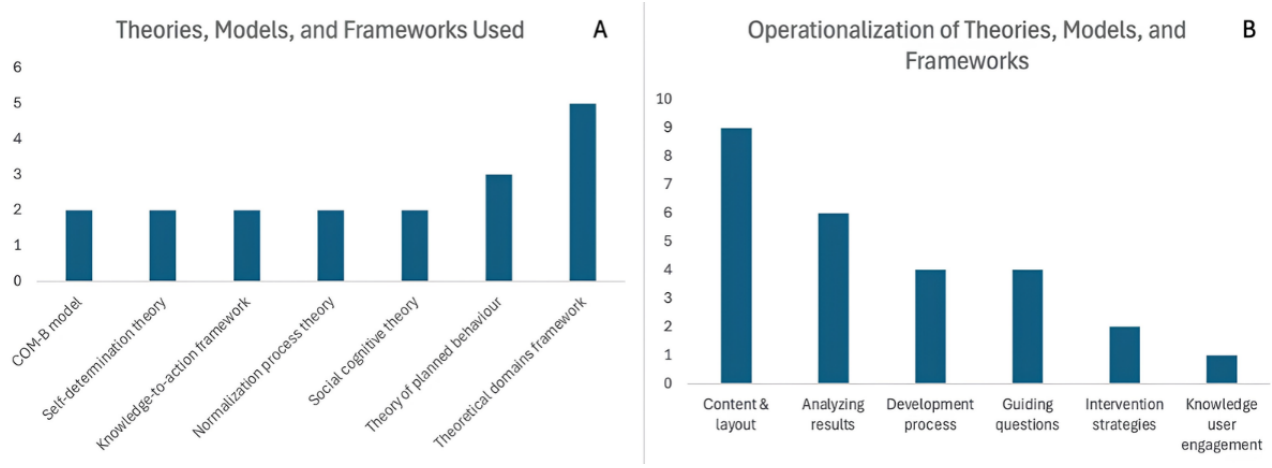


Theories, Models, and Frameworks

Use of TMFs in toolkit development was described in 15 studies [40-45,48-52,54-57]. These TMFs were from a range of disciplines, including education, psychology, and implementation science. TMFs used in multiple studies were the TDF [37] (n=5; [42,43,48,54,57]), Theory of Planned Behavior [58] (n=3; [54-56]), Knowledge-to-Action framework [59] (n=2; [45,49]), Normalization Process Theory [60] (n=2; [44,49]), Social Cognitive Theory [61] (n=2; [41,54]), the COM-B (capability, opportunity, and motivation–behavior) model [62] (n=2; [43,57]), and Self-Determination Theory [63] (n=2; [51,56]). More than one TMF was used in 11 studies [41,43,45,48-51,54-57], while 4 used only one TMF [40,42,44,52]. Operationalization of TMFs included informing toolkit content and layout (n=9) [40,41,48,50-52,54-56], analyzing results (n=6) [43,45,48,49,54,57], guiding the toolkit development process (n=4) [45,49,54,55], guiding questions asked to knowledge users (n=4) [42,44,54,57], informing intervention strategies (n=2) [45,48], and guiding knowledge user engagement (n=1) [45]. These findings can be seen in graphical format in Figure 4.

XSL•FO
RenderX

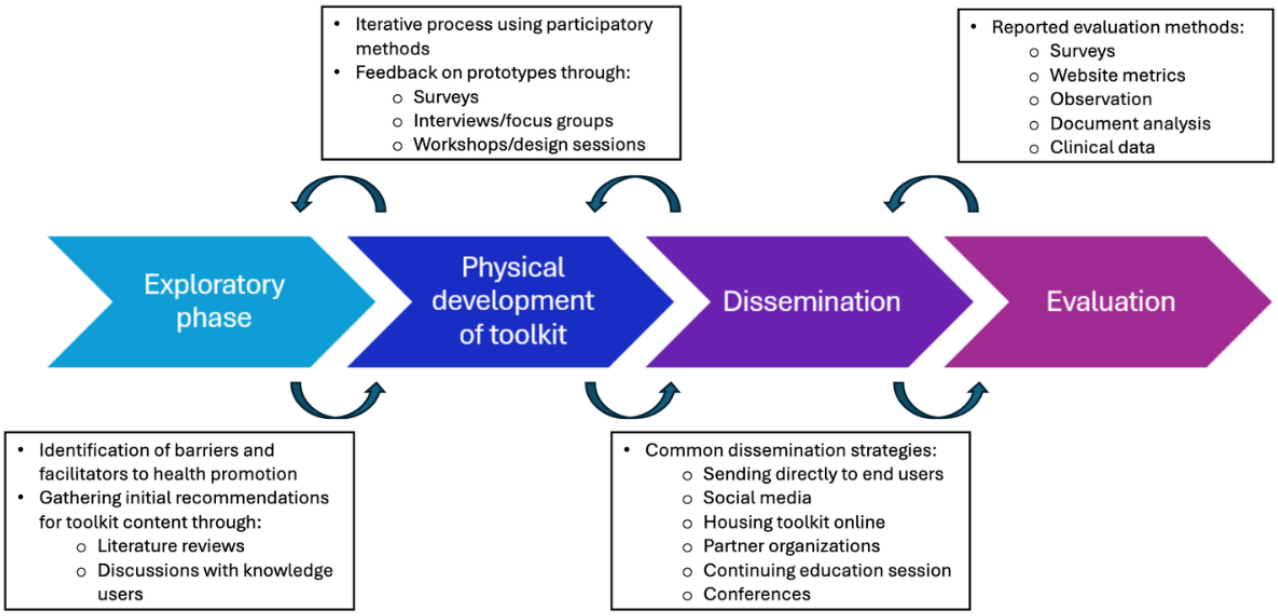
Figure 4. Frequency of the reported (A) theories, models, and frameworks used in multiple articles, and (B) how theories, models, and frameworks were operationalized. COM-B: capability, opportunity, and motivation–behavior.



Phases of Development

Studies included 2-5 development phases. Development typically took the following phases: understanding the problem and doing preliminary work, toolkit development, dissemination, and evaluation. These phases, along with a description of some activities that occurred during each phase, are shown in Figure 5 and described further below.

Figure 5. Phases of toolkit development seen in included studies, including information on activities seen during each phase.



During the exploratory stage, studies often used literature reviews and/or discussions with knowledge users to identify barriers and facilitators to health promotion behaviors and needs for an intervention to address a gap. In addition to gathering barriers and facilitators, this preliminary work also included literature reviews and/or discussions with knowledge users to gather recommendations for toolkit content (eg, what topics should be included, how the toolkit should be presented). Once this information was gathered, teams typically moved into the next phase of developing the toolkit itself.

The toolkit development phase was often an iterative process rather than a linear or a one-time process. Iterative practices often included discussions with knowledge users to get feedback; the number of times that these users were engaged throughout the process varied between studies. Methods of engaging knowledge users for feedback included surveys [40,41,43,45,47,51,54,57], interviews/focus groups [39-57], and workshops/design sessions [42,45,51,52,54,57]. Almost all studies (n=18) gathered recommendations from knowledge users and each of these studies incorporated these recommendations into the intervention [39-47,49-57].

Recommendations from knowledge users during the iterative design stage often included suggestions regarding language (ie, simplifying language and avoiding stigmatized terminology/using gentler language) [41,45,53,55,57], presentation (eg, readable font and use of color) [41,43,45,47,49,51,52,57], content additions (eg, increased information on exercise recommendations for specific groups) [40,43,45-47,49-51,53,54,57], and key messages [40,43,47,50]. Further, knowledge users provided feedback on structure [47,50-52,54,57], name suggestions [42,53], target audiences [45,52,56], dissemination strategies [47,50], and evaluation

XSL•FO
RenderX

strategies [47]. We have provided a series of suggestions for the design of future toolkits based on commonly reported recommendations by study authors (Table 1). All studies incorporated recommendations from knowledge users, where appropriate, during iterative development of the toolkit.

Table 1. Recommendations for the design of future toolkits to support health care providers in health behavior change for patients with chronic conditions.

Recommendation	Rationale	Supporting evidence
Toolkits should include simple messages and avoid too much information (eg, including a summary sheet)	Time was cited as a frequent barrier to HCP[a] promotion, so tools that are easy to reference and quick to use are needed	[40,42,47,49-51,57]
Toolkits should use lay terms and avoid jargon	Toolkits need to be easily accessible to and not confusing for end users, who may not be aware of technical terms	[41,45,55,57]
Toolkits should be visually appealing and use color, inclusive font, and images	Knowledge users highlighted the importance of making the toolkit visually appealing and easy to read to engage users	[41,43,45,47,49,52,57]
Toolkits should include practical examples and be easy for use	Including practical examples allows users to better understand and actively use the toolkit in practice	[39,40,42,45,47,53,57]
Toolkit information should be relevant to the patient population and tailored to their specific needs	End users recommended tailoring toolkit information to the patient population they are working with so that they can better provide recommendations based on the patients' unique needs	[40,43,45,49,50,56,57]
Avoid sensitive terminology and harsh wording	Some terms that the research teams used were identified as charged/offensive by knowledge users who reviewed findings	[41,53,55]

[a]HCP: health care provider.

Dissemination strategies were described in 10 studies; however, 3 of these studies described dissemination for piloting the toolkit and not for full rollout of the toolkit. Common dissemination strategies included sending directly to end users (n=4) [41,46,54,64], social media/blog posting (n=3) [41,45,48], hosting the toolkit online (n=3) [45,48,51], sharing with partner organizations/community representatives (n=3) [45,47,52], continuing education sessions (n=2) [45,53], and presenting at conferences (n=2) [45,52].

Toolkit evaluation methods were described in 12 studies [39,40,42-44,46-48,50,51,54,55], with results described in 11 studies (one study described future evaluation initiatives) [39,40,42-44,47,48,50,51,54,55]. One study evaluated knowledge user engagement but did not evaluate the toolkit itself [45]. Common evaluation methods included surveys (n=7) [40,42,43,47,48,51,54], interviews (n=7) [39,44,46,47,50,54,55], website metrics (n=2) [47,48], observation (n=1) [39], document analysis (n=1) [39], and clinical data (n=1) [44]. No TMFs were explicitly stated as guiding the evaluation procedures. The evaluation outcomes most commonly examined included toolkit usability (n=5) [42,43,49,54,55], acceptability (n=4) [49,51,53,54], use of the toolkit in practice (n=3) [44,48,50], user experience with the toolkit (n=3) [39,46,55], feasibility of toolkit use (n=3) [49,54,56], whether recommendations from knowledge users were adequately incorporated (n=2) [47,48], usefulness of the toolkit (n=2) [40,42], change in knowledge (n=2) [47,50], and increased skills (n=2) [42,50].

Barriers and Facilitators Identified

Fifteen studies examined barriers and/or facilitators to promoting health behaviors among patients [39-49,51,54,56,57], of which 12 examined these in HCPs [39-41,44,46-49,51,54,56,57]. Barriers corresponding to TDF domains included environmental context and resources (n=9) [42-44,47-49,54,56,57], skills (n=5) [40,41,45,48,57], beliefs about consequences (n=4) [44,48,54,56], social/professional role/identity (n=3) [51,54,56], beliefs about capabilities (n=3) [46,51,56], optimism (n=2) [47,54], knowledge (n=2) [39,57], memory (n=1) [54], and social influences (n=1) [54]. Facilitators corresponding to TDF domains included environmental context and resources (n=4) [40,41,48,57], knowledge (n=4) [41,47,54,57], skills (n=4) [39,41,54,57], beliefs about capabilities (n=2) [41,54], social influences (n=2) [47,48], intentions (n=1) [54], memory (n=1) [39], optimism (n=2) [54,57], and social/professional role/identity (n=2) [44,57]. The most frequently reported barrier was time (n=7) [43,44,47,49,54,56,57], including a lack of time to prescribe and tools that take a long time to use. A description of the domains that were cited in this study along with an example for each domain can be found in Table 2.

XSL•FO
RenderX

Table 2. Examples of barriers and facilitators mapped to the Theoretical Domains Framework.

Domain	Definition[a]	Example
Beliefs about capabilities	Acceptance of the truth, reality, or validity about an ability, talent, or facility that a person can put to constructive use	Perceived lack of capability to provide health behavior advice
Beliefs about consequences	Acceptance of the truth, reality, or validity about outcomes of a behavior in a given situation	Belief that prescribing exercise will improve patient health
Environmental context and resources	Any circumstance of a person's situation or environment that discourages or encourages the development of skills and abilities, independence, social competence, and adaptive behavior	Time to prescribe health behaviors
Intentions	A conscious decision to perform a behavior or a resolve to act in a certain way	Decision to address health behaviors in future consultations with patients
Knowledge	An awareness of the existence of something	Lack of knowledge about the target patient population
Memory, attention, and decision processes	The ability to retain information, focus selectively on aspects of the environment and choose between 2 or more alternatives	Difficulties recalling certain elements when prescribing health behaviors
Optimism	The confidence that things will happen for the best or that desired goals will be attained	Confidence that the practice will lead to beneficial changes in patient behavior
Skills	An ability or proficiency acquired through practice	Need for training on how to prescribe for the patient population
Social influences	Those interpersonal processes that can cause individuals to change their thoughts, feelings, or behaviors	Advocacy by respected health care champions
Social/professional role/identity	A coherent set of behaviors and displayed personal qualities of an individual in a social or work setting	Belief that exercise prescription aligned with their role

[a]Definitions for each domain were extracted verbatim from Atkins et al [38].

Toolkit Development

Based on our findings and the recommendations/lessons learned reported by authors of included studies, we have developed a series of recommendations for the development process of toolkits to support HCPs in health behavior change that future toolkit developers may consider (Table 3). This provides a novel contribution, as a previous review noted that there is a lack of information regarding the development of toolkits to change health behaviors in people with chronic conditions [22]. First, in line with how we have operationalized toolkits, we recommend that knowledge users should be engaged from start to finish, as several authors noted that having knowledge user feedback and support was critical for the design and rollout of the toolkit. Next, we recommend that toolkit development should be underpinned by theory, as this can provide guidance for implementers on how to change behaviors. Third, we recommend that while there should be systematic planning for toolkit development, there should be flexibility for changes and iteration as needed, including incorporating feedback into future versions of the toolkit for sustainability of toolkit use. Fourth, we note that developers may consider using virtual communication tools during the toolkit development process. Using virtual communication tools was noted by authors as effective for allowing team members from a wide range of areas to be engaged in toolkit development quickly and without having to travel for meetings. Finally, we suggest that implementers may consider providing additional training opportunities for HCPs to support toolkit use. This may include information on how to use the toolkit and/or information on how to promote health behavior change in the patient population.

XSL•FO
RenderX

Table 3. Recommendations for the development process for future toolkits to support health care providers in health behavior change for patients with chronic conditions.

Recommendations	Rationale	Supporting evidence
Knowledge users should be engaged from start to finish	Authors noted that engaging a range of knowledge users was a strength and made the toolkit more applicable to their audience	[39,43-51,54,55]
Toolkit development should be underpinned by behavior change theory	Having a behavior change theory underpinning toolkit development can help guide developers in the design and development of the toolkit	[43,54,55]
Toolkits should be developed using systematic planning with room for flexibility and iterative processes	Allowing space for iterative processes allows for feedback to be incorporated to develop a toolkit best suited to end users	[39,48,52,55,56]
Using virtual communication tools (eg, Zoom) may be beneficial for conducting toolkit development	Using virtual communication methods could overcome barriers like transportation time and geography to engage team members from a range of areas	[45,49,52]
Additional training should be provided to health care providers to support toolkit use	Supplementing the toolkit with additional training (eg, workshops) could help end users feel more knowledgeable and confident using the toolkit	[41,50,54]

Toolkit Definition

Based on our findings, we propose minimum components required for an intervention to be considered a toolkit. Once these components are met, we recognize that toolkits are flexible in what they can include. Developers should consider the needs of their end users and aims of the toolkit when designing the toolkit. Toolkits at a minimum should:

1. Include materials aimed at improving end user' ability to conduct the intended behavior (eg, increased health behavior prescription). These may include a variety of materials including educational materials, forms to assist in conducting intended behaviors, etc.
2. Include more than one resource. Toolkits may include different modes of resources (eg, videos and written materials) or they may include one type of resource (eg, written sections), so long as there are multiple resources (eg, for a toolkit containing only written sections, multiple written sections should be included).
3. Be designed with end users' involvement and needs in mind. Resources should be intentionally designed to address the recommendations and barriers/facilitators of end users.

Discussion

Summary of Evidence

This scoping review is the first to examine the TMFs, approaches, and methods used in the development of toolkits to support HCPs in changing the health behaviors of their patients with chronic conditions. This review adds to the literature regarding development of toolkits by presenting information on the "toolkit" concept and offering practical recommendations that can be used by future toolkit developers. We included 19 articles in our review that described toolkit development processes. Toolkits were used for a range of HCP audiences, patient populations, and health behaviors. Included studies on toolkit development were from a range of countries. The number of studies increased over the 18-year publication period, with approximately half published since 2020. While the exact reasoning for this growth in toolkit popularity is unknown, it notably coincides with knowledge mobilization plans being increasingly required in grant applications [65,66].

Reported toolkit development phases typically involved initial exploratory work, iterative toolkit development, dissemination, and evaluation. Exploratory work typically included using literature reviews and/or discussions with knowledge users to identify the need for a toolkit and recommendations for toolkit content to address their needs. These recommendations were collected through surveys, interviews/focus groups, and workshops/design sessions. A previous review similarly noted that these are common methods for informing the development of interventions for HCPs [20]. All studies included knowledge users in the process, with a majority reporting the use of participatory methods. Knowledge users were identified from a range of areas and often leveraged current professional networks and wider networks to recruit a range of knowledge users. This finding is similar to a previous review on the methods for designing interventions to change health care professionals' behaviors, which found that a majority of studies included user engagement when designing the intervention [20]. While several studies reported using participatory methods to engage knowledge users, only one study reported using a TMF to guide knowledge user engagement. Future studies should consider the use of TMFs to guide engagement of knowledge users in toolkit development to ensure equitable, effective, and acceptable engagement. Additionally, it was unclear in several studies how knowledge users were involved in the participatory component of development, especially regarding the role of knowledge users in study procedures. Future studies using co-approaches should systematically delineate the specific roles assumed by knowledge users across all phases of the research cycle, indicating whether their involvement was confined to discrete tasks (eg, toolkit design) or integrated throughout the project, thereby enhancing transparency and evidencing the authenticity of the co-approach. Only one study [57] provided a citation for the term "toolkit". Toolkits were often multimodal and toolkit components (eg, videos and forms) varied greatly between studies. Thus, there is a need for a clearer definition

XSL•FO
RenderX

and conceptualization of a toolkit. This is especially important in implementation research, as the use of consistent terminology and a clear, shared conceptualization of the intervention being examined (ie, a toolkit) is fundamental to studying, comparing, and improving such interventions [67].

In addition to presenting recommendations for toolkit design and development, our review presented a definition of a toolkit. By defining the minimum components of a toolkit and noting recommendations for how toolkits should be developed, future developers can better consider why they are choosing a toolkit as an intervention strategy and how they are developing the toolkit. As there was a previous lack of description and formal recommendations on methods for developing toolkits in this area [22], this can help future developers guide how toolkits should be developed to best suit the needs of their end users and ensure that they use toolkits in a sustainable and effective way. Future research should consider the design of the toolkit in addition to its development, as multiple studies described involving knowledge users to get feedback on the barriers and facilitators to toolkit use or prescribing health behaviors but largely did not describe theories guiding toolkit design (eg, applying learning theory to toolkit design to ensure end users can best learn from toolkit use). In summary, toolkits should be both designed and developed in a systematic way that is informed by TMFs.

Limitations

This review was not without limitations. First, articles were limited to those written in English, so toolkits that were developed and reported on in largely non-English speaking countries may have been missed. Further, this review excluded gray literature, which may have resulted in exclusion of literature related to the development of toolkits in nonacademic settings. Finally, the generalizability of findings for health behavior change must be considered. No included toolkits were aimed at reducing smoking, alcohol use, or recreational drug use. Additionally, obesity and neurological conditions made up most of the chronic conditions examined in this review. Thus, as toolkits included generally clustered around exercise behaviors and obesity, the generalizability of our findings to populations and toolkits to other populations and health behaviors may be limited. Further, future research may consider a broader definition of health behaviors, including those related to seeking health services, such as those in the taxonomy of Nudelemen and Shiloh [11]. Future research should examine the methods used in the development of toolkits for these areas.

Conclusions

Our review demonstrates that toolkits used to support HCPs in health behavior change for patients with chronic conditions vary considerably in their components and audiences. Inclusion of knowledge users was found to be an effective component of toolkit development. Based on our findings, we developed a series of recommendations that future teams can consider for the design and the development process of toolkits. Future research should report on the development and design of toolkits for HCPs that are aimed at reducing smoking, excessive alcohol use, and recreational drug use in patients with chronic conditions. This review has important implications for knowledge translation initiatives, as it defines a “toolkit” to allow for better comparisons and contextualization of toolkits in the future. Additionally, the development of recommendations for both design and development of toolkits allows for toolkits that are more likely to be effectively used and close the knowledge-to-practice gap.

Funding

This work was supported by a MS Canada Discovery Grant (1257781). JS is supported by a Canadian Institute of Health Research (CIHR) Doctoral Award and KITE-Toronto Rehab’s TD Graduate Scholarship. The funder was not involved in the study design, data collection, analysis, interpretation, or the writing of the manuscript

Generative AI Use

No generative AI tools were used at any stage in the preparation of this manuscript.

Authors' Contributions

Conceptualization: JS, NW, MW, SJD, RS, SM
Data curation: AS, JT, PF, DG, HG, JS
Formal analysis: JS, SJD, RS, SM
Investigation: AS, JT, PF, DG, HG
Writing – original draft: JS
Writing – review & editing: JS, NW, MW, SJD, RS, SM, AS, JT, PF, DG, HG
All authors read and approved the final manuscript.

Conflicts of Interest

SM is the Editor-in-Chief of *JMIR Rehabilitation and Assistive Technologies*. The other authors report no conflicts of interest.

Multimedia Appendix 1
PRISMA-ScR checklist.

XSL•FO
RenderX

[PDF File (Adobe PDF File), 153 KB - mededu_v12i1e88653_app1.pdf]

Multimedia Appendix 2
Search strategies.
[DOCX File , 47 KB - mededu_v12i1e88653_app2.docx]

Multimedia Appendix 3
Data extraction template.
[PDF File (Adobe PDF File), 97 KB - mededu_v12i1e88653_app3.pdf]

Multimedia Appendix 4
Main findings for each included study.
[DOCX File , 29 KB - mededu_v12i1e88653_app4.docx]

References

1. Health Behaviours. Statistics Canada. 2016. URL: https://www150.statcan.gc.ca/n1/pub/82-229-x/2009001/deter/int3-eng.htm [accessed 2026-09-22]
2. Seo B, Nan H, Monahan PO, Duszynski TJ, Thompson WR, Zollinger TW, et al. Association between US residents' health behavior and good health status at the city level. Transl J Am Coll Sports Med 2024;9(2). [doi: 10.1249/tjx.0000000000000258]
3. Mozaffarian D, Afshin A, Benowitz NL, Bittner V, Daniels SR, Franch HA, American Heart Association Council on EpidemiologyPrevention, Council on Nutrition, Physical ActivityMetabolism, Council on Clinical Cardiology, Council on Cardiovascular Disease in the Young, Council on the Kidney in Cardiovasc. Population approaches to improve diet, physical activity, and smoking habits: a scientific statement from the American Heart Association. Circulation 2012 Sep 18;126(12):1514-1563 [FREE Full text] [doi: 10.1161/CIR.0b013e318260a20b] [Medline: 22907934]
4. Short SE, Mollborn S. Social determinants and health behaviors: conceptual frames and empirical advances. Curr Opin Psychol 2015;5:78-84 [FREE Full text] [doi: 10.1016/j.copsyc.2015.05.002] [Medline: 26213711]
5. Spring B, Moller AC, Coons MJ. Multiple health behaviours: overview and implications. J Public Health (Oxf) 2012;34 Suppl 1(Suppl 1):3-10 [FREE Full text] [doi: 10.1093/pubmed/fdr111] [Medline: 22363028]
6. Giannakou K, Kyprianidou M, Chrysostomou S, Christophi CA. Editorial: the associations of lifestyle factors and behaviors with multimorbidity. Front Public Health 2023;11:1227381 [FREE Full text] [doi: 10.3389/fpubh.2023.1227381] [Medline: 37441654]
7. Tran B, Falster MO, Douglas K, Blyth F, Jorm LR. Health behaviours and potentially preventable hospitalisation: a prospective study of older Australian adults. PLoS One 2014;9(4):e93111 [FREE Full text] [doi: 10.1371/journal.pone.0093111] [Medline: 24691471]
8. Foundation health measures: determinants of health. US Department of Health and Human Services. 2017. URL: https://www.cdc.gov/nchs/healthy_people/hp2020/foundation-health-measures.htm#footnotes [accessed 2026-10-06]
9. Manuel DG, Bennett C, Perez R, Wilton AS, Rohit Dass A, Laporte A, et al. Burden of health behaviours and socioeconomic position on health care expenditure in Ontario. F1000Res 2019;8:303 [FREE Full text] [doi: 10.12688/f1000research.18205.2] [Medline: 31723417]
10. Espinosa-Salas S, Gonzalez-Arias M. Behavior Modification for Lifestyle Improvement. Treasure Island, FL: StatPearls Publishing; 2025.
11. Nudelman G, Shiloh S. Mapping health behaviors: constructing and validating a common-sense taxonomy of health behaviors. Soc Sci Med 2015;146:1-10. [doi: 10.1016/j.socscimed.2015.10.004] [Medline: 26473449]
12. Cunningham C, O'Sullivan R. Healthcare professionals' application and integration of physical activity in routine practice with older adults: a qualitative study. Int J Environ Res Public Health 2021;18(21):11222 [FREE Full text] [doi: 10.3390/ijerph182111222] [Medline: 34769742]
13. Pipe AL, Evans W, Papadakis S. Smoking cessation: health system challenges and opportunities. Tob Control 2022;31(2):340-347. [doi: 10.1136/tobaccocontrol-2021-056575] [Medline: 35241609]
14. Kraef C, Wood B, von Philipsborn P, Singh S, Peterson SS, Kallestrup P. Primary health care and nutrition. Bull World Health Organ 2020;98(12):886-893. [doi: 10.2471/blt.20.251413]
15. Vallis M, Lee-Baggley D, Sampalli T, Ryer A, Ryan-Carson S, Kumanan K, et al. Equipping providers with principles, knowledge and skills to successfully integrate behaviour change counselling into practice: a primary healthcare framework. Public Health 2018;154:70-78. [doi: 10.1016/j.puhe.2017.10.022] [Medline: 29216495]
16. Thoele K, Ferren M, Moffat L, Keen A, Newhouse R. Development and use of a toolkit to facilitate implementation of an evidence-based intervention: a descriptive case study. Implement Sci Commun 2020;1(1):86 [FREE Full text] [doi: 10.1186/s43058-020-00081-x] [Medline: 33043301]

XSL•FO
RenderX

17. Yamada J, Shorkey A, Barwick M, Widger K, Stevens BJ. The effectiveness of toolkits as knowledge translation strategies for integrating evidence into clinical care: a systematic review. BMJ Open 2015;5(4):e006808-e006808 [FREE Full text] [doi: 10.1136/bmjopen-2014-006808] [Medline: 25869686]
18. Dykes PC, Duckworth M, Cunningham S, Dubois S, Driscoll M, Feliciano Z, et al. Pilot testing fall TIPS (tailoring interventions for patient safety): a patient-centered fall prevention toolkit. Jt Comm J Qual Patient Saf 2017;43(8):403-413. [doi: 10.1016/j.jcjq.2017.05.002] [Medline: 28738986]
19. Helmle KE, Chacko S, Chan T, Drake A, Edwards AL, Moore GE, et al. Knowledge translation to optimize adult inpatient glycemic management with basal bolus insulin therapy and improve patient outcomes. Can J Diabetes 2018;42(5):505-513. [doi: 10.1016/j.jcjd.2017.12.010] [Medline: 29555341]
20. Colquhoun HL, Squires JE, Kolehmainen N, Fraser C, Grimshaw JM. Methods for designing interventions to change healthcare professionals' behaviour: a systematic review. Implement Sci 2017;12(1):30 [FREE Full text] [doi: 10.1186/s13012-017-0560-5] [Medline: 28259168]
21. Hempel S, Miake-Lye I, Brega AG, Buckhold F, Hassell S, Nowalk MP, et al. Quality improvement toolkits: recommendations for development. Am J Med Qual 2019;34(6):538-544. [doi: 10.1177/1062860618822102] [Medline: 30675797]
22. Levy T, Huxley K, Vuu S, Lewis LK. Physical activity for people with chronic conditions: a systematic review of toolkits to promote adherence. Disabil Rehabil 2024;46(22):5199-5212 [FREE Full text] [doi: 10.1080/09638288.2023.2296525] [Medline: 38131330]
23. Nilsen P. Making sense of implementation theories, models and frameworks. Implement Sci 2015;10(1):53 [FREE Full text] [doi: 10.1186/s13012-015-0242-0] [Medline: 25895742]
24. Damschroder LJ. Clarity out of chaos: use of theory in implementation research. Psychiatry Res 2020;283:112461 [FREE Full text] [doi: 10.1016/j.psychres.2019.06.036] [Medline: 31257020]
25. Peters M, Godfrey CM, Mcinerney P, Munn Z, Trico A, Khalil H. Chapter 10: scoping reviews. In: JBI Manual for Evidence Synthesis. Adelaide, Australia: JBI Manual for Evidence Synthesis; 2024.
26. Scandiffio J, Sureshkumar A, Feng G, Gaid D, Wasilewski M, Woods N, et al. The approaches, theories, models, frameworks, and methods in designing toolkits to support healthcare providers in health behaviour change: a scoping review protocol. PLoS One 2026;21(6):e0349867 [FREE Full text] [doi: 10.1371/journal.pone.0349867] [Medline: 42224302]
27. Tricco AC, Lillie E, Zarin W, O'Brien KK, Colquhoun H, Levac D, et al. PRISMA extension for scoping reviews (PRISMA-ScR): checklist and explanation. Ann Intern Med 2018;169(7):467-473 [FREE Full text] [doi: 10.7326/M18-0850] [Medline: 30178033]
28. About chronic diseases. Centers for Disease Control and Prevention. 2024. URL: https://www.cdc.gov/chronic-disease/about/index.html [accessed 2026-09-22]
29. Health of Canadians-health outcomes. Statistics Canada. 2025. URL: https://www150.statcan.gc.ca/n1/pub/82-570-x/2024001/section2-eng.htm [accessed 2026-09-22]
30. Public Health Agency of Canada. Chronic Diseases. 2025. URL: https://www.canada.ca/en/public-health/services/chronic-diseases.html [accessed 2026-09-22]
31. Rethlefsen ML, Kirtley S, Waffenschmidt S, Ayala AP, Moher D, Page MJ, PRISMA-S Group. PRISMA-S: an extension to the PRISMA statement for reporting literature searches in systematic reviews. Syst Rev 2021;10(1):39 [FREE Full text] [doi: 10.1186/s13643-020-01542-z] [Medline: 33499930]
32. McGowan J, Sampson M, Salzwedel DM, Cogo E, Foerster V, Lefebvre C. PRESS peer review of electronic search strategies: 2015 guideline statement. J Clin Epidemiol 2016;75:40-46 [FREE Full text] [doi: 10.1016/j.jclinepi.2016.01.021] [Medline: 27005575]
33. McHugh ML. Interrater reliability: the kappa statistic. Biochem Med 2012;22(3):276-282. [doi: 10.11613/bm.2012.031]
34. Peters M, Godfrey CM, Mcinerney P, Baldini SC, Khalil H, Parker D. Scoping reviews. In: JBI Manual for Evidence Synthesis. Australia: JBI; 2020.
35. O'Neill J, Tabish H, Welch V, Petticrew M, Pottie K, Clarke M, et al. Applying an equity lens to interventions: using PROGRESS ensures consideration of socially stratifying factors to illuminate inequities in health. J Clin Epidemiol 2014;67(1):56-64. [doi: 10.1016/j.jclinepi.2013.08.005] [Medline: 24189091]
36. Hsieh HF, Shannon SE. Three approaches to qualitative content analysis. Qual Health Res 2005;15(9):1277-1288. [doi: 10.1177/1049732305276687] [Medline: 16204405]
37. Cane J, O'Connor D, Michie S. Validation of the theoretical domains framework for use in behaviour change and implementation research. Implement Sci 2012;7:37 [FREE Full text] [doi: 10.1186/1748-5908-7-37] [Medline: 22530986]
38. Atkins L, Francis J, Islam R, O'Connor D, Patey A, Ivers N, et al. A guide to using the theoretical domains framework of behaviour change to investigate implementation problems. Implement Sci 2017;12(1):77 [FREE Full text] [doi: 10.1186/s13012-017-0605-9] [Medline: 28637486]
39. Noël G, Luig T, Heatherington M, Campbell-Scherer D. Developing tools to support patients and healthcare providers when in conversation about obesity. Inf Des J 2019;24(2):131-150. [doi: 10.1075/idj.00004.noe]
40. Osunlana AM, Asselin J, Anderson R, Ogunleye AA, Cave A, Sharma AM, et al. 5As team obesity intervention in primary care: development and evaluation of shared decision-making weight management tools. Clin Obes 2015 Aug;5(4):219-225 [FREE Full text] [doi: 10.1111/cob.12105] [Medline: 26129630]

XSL•FO
RenderX

41. Kebbe M, Perez A, Buchholz A, McHugh TF, Scott SD, Richard C, et al. Conversation cards for adolescents©: a patient-centered communication and behavior change tool for adolescents with obesity and health care providers. J Commun Healthc 2020;13(2):79-88. [doi: 10.1080/17538068.2020.1765126]
42. Dennett A, Tang C, Osadnik C. The cancer exercise toolkit: giving exercise professionals online tools to bridge the evidence-practice gap. Asia-Pac J Clin Oncol 2022;18(S3):53-114. [doi: 10.1111/ajco.13868]
43. Gale N, Jones U, Rees T, Hicks A, Davies J, Holliday S, et al. A cancer personalised activity and lifestyle tool (CAN-PAL): a codesign study with patients and healthcare professionals. J Clin Nurs 2024;33(2):572-579. [doi: 10.1111/jocn.16931] [Medline: 38062580]
44. Jones U, Hamana K, O'Hara F, Busse M. The development of PAT-HD: a co-designed tool to promote physical activity in people with huntington's disease. Health Expect 2021;24(2):638-647 [FREE Full text] [doi: 10.1111/hex.13210] [Medline: 33580995]
45. Middleton LE, Pelletier CA, Koch M, Norman R, Dupuis S, Astell A, et al. The dementia-inclusive choices in exercise project: using participatory action research to improve physical activity supports for persons with dementia. Dementia (London) 2023;22(8):1651-1676. [doi: 10.1177/14713012231197144] [Medline: 37715689]
46. Bolster EAM, van Gessel C, Welten M, Hermsen S, van der Lugt R, Kotte E, et al. Using a co-design approach to create tools to facilitate physical activity in children with physical disabilities. Front Rehabil Sci 2021;2:707612 [FREE Full text] [doi: 10.3389/fresc.2021.707612] [Medline: 36188842]
47. Arbour-Nicitopoulos KP, Martin Ginis KA, Latimer-Cheung AE, Bourne C, Campbell D, Cappe S, et al. Development of an evidence-informed leisure time physical activity resource for adults with spinal cord injury: the SCI Get Fit Toolkit. Spinal Cord 2013;51(6):491-500. [doi: 10.1038/sc.2013.7] [Medline: 23608809]
48. Glowacki K, Arbour-Nicitopoulos K, Burrows M, Chesick L, Heinemann L, Irving S, et al. It's more than just a referral: development of an evidence-informed exercise and depression toolkit. Ment Health Phys Act 2019;17:100297. [doi: 10.1016/j.mhpa.2019.100297]
49. Sturgiss EA, Douglas K. A collaborative process for developing a weight management toolkit for general practitioners in Australia-an intervention development study using the knowledge to action framework. Pilot Feasibility Stud 2016;2(1):20 [FREE Full text] [doi: 10.1186/s40814-016-0060-4] [Medline: 27965840]
50. Evans PH, Greaves C, Winder R, Fearn-Smith J, Campbell JL. Development of an educational 'toolkit' for health professionals and their patients with prediabetes: the WAKEUP study (ways of addressing knowledge education and understanding in pre-diabetes). Diabet Med 2007;24(7):770-777. [doi: 10.1111/j.1464-5491.2007.02130.x] [Medline: 17403125]
51. Goff AJ, De Oliveira Silva D, Ezzat AM, Crossley KM, Pazzinatto MF, Barton CJ. Co-design of the web-based 'My Knee' education and self-management toolkit for people with knee osteoarthritis. Digit Health 2023;9:20552076231163810 [FREE Full text] [doi: 10.1177/20552076231163810] [Medline: 37009308]
52. Carr JJ, Lalara J, Lalara G, Lalara G, Daniels B, Clough AR, et al. Staying strong toolbox: co-design of a physical activity and lifestyle program for aboriginal families with Machado-Joseph disease in the top end of Australia. PLoS One 2021;16(2):e0244311 [FREE Full text] [doi: 10.1371/journal.pone.0244311] [Medline: 33544709]
53. Chock GYK, Kerr NA. A report on the development of the Hawai'i pediatric weight management toolkit. Hawaii Med J 2011;70(7 Suppl 1):49-51 [FREE Full text] [Medline: 21886295]
54. Avery L, Charman SJ, Taylor L, Flynn D, Mosely K, Speight J, et al. Systematic development of a theory-informed multifaceted behavioural intervention to increase physical activity of adults with type 2 diabetes in routine primary care: movement as medicine for type 2 diabetes. Implement Sci 2016;11(1):99 [FREE Full text] [doi: 10.1186/s13012-016-0459-6] [Medline: 27430648]
55. Mateo KF, Berner NB, Ricci NL, Seekaew P, Sikerwar S, Tenner C, et al. Development of a 5As-based technology-assisted weight management intervention for veterans in primary care. BMC Health Serv Res 2018;18(1):47 [FREE Full text] [doi: 10.1186/s12913-018-2834-2] [Medline: 29378584]
56. Verwey R, van der Weegen S, Spreeuwenberg M, Tange H, van der Weijden T, de Witte L. Upgrading physical activity counselling in primary care in the netherlands. Health Promot Int 2016;31(2):344-354 [FREE Full text] [doi: 10.1093/heapro/dau107] [Medline: 25539787]
57. Levy T, Vuu S, Huxley K, Chamberlain S, Lewis L, Lynch E. Co-designing the "Adherence Counselling Toolkit (ACT now)" to promote exercise adherence of survivors of stroke. Disabil Rehabil 2025;47(24):6443-6456 [FREE Full text] [doi: 10.1080/09638288.2025.2483467] [Medline: 40165052]
58. Ajzen I. Perceived behavioral control, self‐efficacy, locus of control, and the theory of planned behavior. J Appl Soc Psychol 2006;32(4):665-683. [doi: 10.1111/j.1559-1816.2002.tb00236.x]
59. Graham ID, Logan J, Harrison MB, Straus SE, Tetroe J, Caswell W, et al. Lost in knowledge translation: time for a map? J Contin Educ Health Prof 2006;26(1):13-24. [doi: 10.1002/chp.47] [Medline: 16557505]
60. May C, Rapley T, Mair FS, Treweek S, Murray E. New technologies, and business processes in healthcare are complex and demanding. Normalization Process Theory. 2021. URL: https://normalization-process-theory.northumbria.ac.uk/ [accessed 2026-09-22]
61. Bandura A. Social foundations of thought and action. In: A Social Cognitive Theory. Englewood Cliffs, NJ: Prentice Hall; 1986:523-582.

XSL•FO
RenderX

62. Michie S, Van Stralen MM, West R. The behaviour change wheel: a new method for characterising and designing behaviour change interventions. Implement Sci 2011;6:42 [FREE Full text] [doi: 10.1186/1748-5908-6-42] [Medline: 21513547]
63. Markland D, Ryan RM, Tobin VJ, Rollnick S. Motivational interviewing and self–determination theory. J Soc Clin Psychol 2005;24(6):811-831. [doi: 10.1521/jscp.2005.24.6.811]
64. Jones D, Drewery R, Windle K, Humphrey S, de Paiva AF. Dementia prevention and the GP's role: a qualitative interview study. Br J Gen Pract 2023;74(741):e242-e249. [doi: 10.3399/bjgp.2023.0103]
65. Knowledge mobilization at CIHR. Canadian Institute of Health Research. 2026. URL: https://cihr-irsc.gc.ca/e/29529.html [accessed 2026-09-22]
66. Guidelines for effective knowledge mobilization. Social Sciences and Humanities Research Council. 2025. URL: https://sshrc-crsh.canada.ca/en/funding/policies-regulations-and-guidelines/guidelines-effective-knowledge-mobilization.aspx [accessed 2026-09-22]
67. Proctor E, Silmere H, Raghavan R, Hovmand P, Aarons G, Bunger A, et al. Outcomes for implementation research: conceptual distinctions, measurement challenges, and research agenda. Adm Policy Ment Health 2011;38(2):65-76 [FREE Full text] [doi: 10.1007/s10488-010-0319-7] [Medline: 20957426]

Abbreviations

COM-B: capability, opportunity, and motivation–behavior
HCP: health care provider
JBI: Joanna Briggs Institute
PCC: population, concept, and context
PRISMA-S: Preferred Reporting Items for Systematic Reviews and Meta-Analyses Search Reporting Extension
PRISMA-ScR: Preferred Reporting Items for Systematic Reviews and Meta-Analyses Extension for Scoping Reviews
TDF: Theoretical Domains Framework
TMF: theories, models, and framework

Edited by S Brini; submitted 28.Nov.2025; peer-reviewed by E Neter, Y Wei; comments to author 20.Jun.2026; revised version received 07.Jul.2026; accepted 17.Sep.2026; published 26.Oct.2026.

Please cite as:
Scandiffio J, Simpson R, Sureshkumar A, Tang J, Fleming P, Gaid D, Gnanakaran H, Wasilewski M, Woods N, Donkers SJ, Munce S
Toolkits to Support Health Care Providers in Health Behavior Change for Patients With Chronic Conditions: Scoping Review
JMIR Med Educ 2026;12:e88653
URL: https://mededu.jmir.org/2026/1/e88653
doi:10.2196/88653
PMID:

©Jillian Scandiffio, Robert Simpson, Ashvene Sureshkumar, Jerry Tang, Paloma Fleming, Dina Gaid, Harzaan Gnanakaran, Marina Wasilewski, Nicole Woods, Sarah J Donkers, Sarah Munce. Originally published in JMIR Medical Education (https://mededu.jmir.org), 26.Oct.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Education, is properly cited. The complete bibliographic information, a link to the original publication on https://mededu.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

AI-Mediated Assessment of Continuing Medical Education: The Case-based Learning Intelligence Credit System (CLICS) Framework

Somkiat Wattanasirichaigoon[1,2,3], MD

[1]Center for Continuing Medical Education, Medical Council of Thailand, The Royal Golden Jubilee Building, 2nd Floor Tiwanon Road Talat Khwan, Mueang Nonthaburi District, Nonthaburi, Thailand

[2]Dhurakij Pundit University, Bangkok, Thailand

[3]Division of Information Technology Management, Faculty of Engineering, Mahidol University, Nakhon Pathom, Thailand

Corresponding Author:
Somkiat Wattanasirichaigoon, MD
Center for Continuing Medical Education, Medical Council of Thailand, The Royal Golden Jubilee Building, 2nd Floor Tiwanon Road Talat Khwan, Mueang Nonthaburi District, Nonthaburi, Thailand

Abstract

Continuing medical education (CME) and continuing professional development (CPD) systems have traditionally relied on time-based credit allocation, using participation duration as a proxy for professional learning. Although administratively simple and scalable, this model does not reliably demonstrate whether physicians have engaged in meaningful learning, improved clinical reasoning, or critically appraised evidence. The emergence of generative AI creates an opportunity to rethink how physician learning is documented, assessed, and credited. This viewpoint proposes the Case-based Learning Intelligence Credit System (CLICS), a conceptual framework for translating AI-mediated clinical learning interactions into auditable evidence of reasoning-related engagement that could support CME/CPD credit. CLICS introduces the professional learning episode (PLE) as the basic unit of creditable learning: a coherent AI-mediated interaction demonstrating a clinically meaningful problem, reasoning development through iterative inquiry, contextual or evidentiary integration, and reflective synthesis. PLEs are evaluated using the proposed Practice Intelligence Score-7 (PIS-7) rubric, subject to human calibration and oversight; the rubric assesses observable reasoning behavior within the episode rather than the AI's answer, and qualifying PLEs may be translated into CME/CPD credit through threshold-based, human-auditable conversion rules. CLICS is not intended to replace traditional CME but to extend it as an optional, evidence-generating pathway for personalized, practice-embedded professional development. Its implementation requires iterative validation, stratified human audit, privacy-by-design architecture, antigaming controls, bias monitoring, and professional oversight. If validated, CLICS may enable identification of domain-specific areas for improvement and support personalized, adaptive learning pathways.

(JMIR Med Educ 2026;12:e99520) doi:10.2196/99520

KEYWORDS

continuing medical education; continuing professional development; generative AI; clinical reasoning; medical education; physician learning; large language models; learning analytics; AI assessment

Introduction

Continuing medical education (CME) and continuing professional development (CPD) systems commonly translate attendance, completion, or duration into professional learning credit [1-3]. Although practical and scalable, these participation metrics leave the quality of reasoning and evidence appraisal during learning largely unobserved.

Conversational generative AI creates a new opportunity, because physicians increasingly use it to explore clinical questions, compare alternatives, and reflect on decisions through iterative interactions embedded in clinical practice [4-6].

This viewpoint proposes the Case-based Learning Intelligence Credit System (CLICS), a framework for translating AI-mediated learning traces into auditable CME/CPD evidence. CLICS makes 3 conceptual shifts: from time-based toward evidence-informed CME; from AI as a learning assistant alone toward AI-assisted evaluation of learning traces, subject to human calibration and oversight; and from isolated learning activities toward practice-embedded professional learning infrastructure.

With CLICS, we do not propose that AI independently certify physician competence or that token count, conversation length, or prompt complexity determine credit. Instead, we distinguish superficial AI use from qualifying learning interactions through the professional learning episode (PLE), defined below. Table

XSL•FO
RenderX

1 contrasts conventional CME/CPD with CLICS to show how the proposed framework may complement established systems by recognizing reasoning-centered learning that participation metrics do not capture.

Table . A conceptual comparison between traditional continuing medical education (CME)/continuing professional development (CPD) and the Case-based Learning Intelligence Credit System (CLICS) framework, showing shifts from participation to reasoning-related evidence.

Dimension	Traditional CME/CPD (established paradigm)	CLICS (proposed paradigm)	Interpretive implication
Foundational assumption	Learning is approximated by time spent in accredited activities	Learning is inferred from observable, reasoning-related engagement within PLEs[a]	Shifts the proxy of learning from duration to observable evidence
Unit of analysis	Educational event (eg, lecture, module, workshop)	PLE (AI-mediated reasoning interaction)	Reframes learning as episodic reasoning rather than attendance-based
Measurement target	Participation and completion	Reasoning-related interaction behavior, inquiry, and reflection	Moves from exposure metrics to reasoning-behavior metrics
Nature of evidence	Indirect (attendance as proxy)	Direct evidence of interaction behavior; inferential evidence of reasoning	Enables partial observation of reasoning-related behaviors
Assessment approach	Often absent or knowledge-based (eg, quizzes)	Multidomain assessment of reasoning-related behavior (PIS-7[b])	Introduces a structured framework for evaluating reasoning behavior
Temporal context of learning	Scheduled and externally organized	Embedded within real-time clinical problem solving	Aligns CME with practice-based learning moments
Degree of personalization	Limited and content-driven	Potentially high; physician-driven and context-specific	Supports individualized learning trajectories
Feedback structure	Minimal or delayed	Immediate, domain-specific (eg, reasoning depth, critical appraisal)	Enables formative, actionable feedback loops
Auditability	Administrative verification (attendance logs)	Transcript-based audit with human calibration	Expands audit from compliance to reasoning-related evidence
Vulnerability to superficial completion	Possible (passive attendance)	Designed to be mitigated by PLE criteria and reflective synthesis requirements	Introduces structural resistance to low-effort credit accumulation
Governance model	Established regulatory frameworks	Emerging governance (audit, bias control, privacy, AI oversight)	Necessitates new regulatory and ethical architectures
Position within CME ecosystem	Foundational and established	Complementary and exploratory	Supports coexistence rather than replacement

[a]PLE: professional learning episode.

[b]PIS-7: Practice Intelligence Score-7.

Limitations of Time-Based CME

Traditional CME systems have played an important role in supporting lifelong learning, and systematic reviews show that CME can influence physician performance and, in some cases, patient outcomes, particularly when learning is interactive and linked to clinical practice [1,2]. However, CME effectiveness is highly variable and depends on how learning is designed, delivered, and assessed [2,7]. Furthermore, time-based credit allocation has 3 fundamental limitations. First, it relies on participation as a proxy for learning: attendance can be verified administratively, but it does not reflect whether meaningful cognitive engagement has occurred; assessment systems should be grounded in meaningful interpretation of performance rather than exposure time alone [3,8,9]. Second, many CME activities are weakly aligned with real-world clinical reasoning—delivered in structured, preplanned formats that do not reflect the complexity and contextual variability of practice—even though reasoning, defined as the process of collecting information, generating hypotheses, and evaluating evidence, develops through iterative problem-solving and exposure to uncertainty rather than passive information acquisition [10-15]. Third, time-based systems offer limited personalization. Physicians differ in specialty, experience, and practice environment yet typically receive standardized content, even though tailored, context-specific feedback is known to enhance retention and reasoning performance [3,7,15,16].

CLICS addresses these limitations by shifting the focus from learning exposure to learning evidence—asking not how long a physician participated but what observable evidence of reasoning-related engagement is present—consistent with performance-informed assessment principles that emphasize observable evidence of learning rather than participation time alone [8,17].

XSL•FO
RenderX

The CLICS Framework

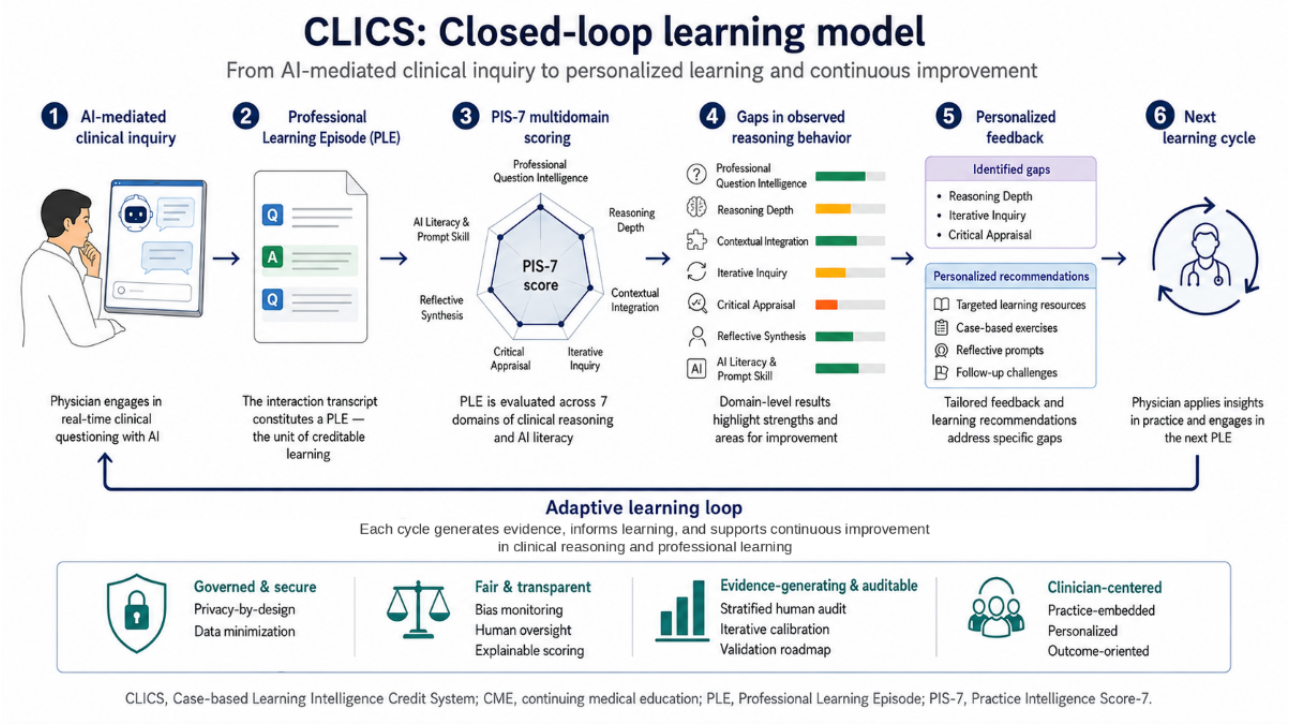
CLICS integrates AI-mediated interaction, learning-trace capture, the Practice Intelligence Score-7 (PIS-7), human calibration, and credit conversion within a coordinated CME/CPD infrastructure. Although the underlying CLICS architecture is designed to be applicable across regulated professions requiring CPD, this viewpoint focuses specifically on physician CME, reflecting its current development for physician CME and planned phase 0 feasibility and reliability/agreement evaluation. Profession-specific adaptation, calibration, and governance would be required before extension beyond this scope.

At the foundation of CLICS is the AI-mediated learning trace: the sequence of physician-AI exchanges generated during a problem-centered inquiry [4-6,18-20]. Such traces provide observable, though indirect, evidence of reasoning behavior and must be interpreted cautiously because linguistic output does not fully represent underlying cognition [10,14].

CLICS defines the PLE as its fundamental unit of creditable learning: a coherent, AI-mediated interaction demonstrating (1) a clinically meaningful problem, (2) reasoning development through iterative inquiry, (3) integration of contextual or evidentiary information, and (4) reflective synthesis. Interactions limited to factual lookup, repetitive prompting, or passive copying of AI output do not qualify, a distinction essential to prevent superficial AI use from being misread as meaningful learning.

PLEs are evaluated by an AI-based evaluator that is conceptually distinct from the AI system generating educational content. Its purpose is to assess how the physician engages with the interaction rather than to judge the correctness of the AI's answer; the PIS-7 rubric used for this assessment is described below [8,9]. Moreover, only qualifying PLEs are eligible for proposed CME/CPD credit. Figure 1 summarizes the closed-loop workflow from inquiry to evaluation, feedback, and subsequent learning.

Figure 1. Closed-loop learning model of the Case-based Learning Intelligence Credit System (CLICS). AI-mediated clinical inquiries generate professional learning episodes (PLEs), which are evaluated using the proposed Practice Intelligence Score-7 (PIS-7) framework.



PLEs and PIS-7 Scoring

PLE components are grounded in the established perspectives that meaningful medical learning involves problem-solving, reasoning, contextualization, and reflection rather than passive exposure [13-15]. Problem framing supports hypothesis generation [10,11]; iterative reasoning supports hypothesis testing and uncertainty management [11,12]; contextual integration incorporates patient, evidence, and system constraints [12,17]; and reflective synthesis supports deeper learning and knowledge transfer [14,15,21]. All 4 components must be present for an interaction to qualify as a PLE.

To operationalize the evaluation of PLEs, CLICS uses the PIS-7 framework, a multidomain rubric designed to assess physician reasoning behavior within a PLE. The 7 domains and their conceptual definitions are summarized in Table 2. Each PIS-7 domain is scored from 0 to 5, and the scores are converted into a weighted composite ranging from 0 to 100:

$$PIS\text{-}7=\sum i=17(Di5\times Wi)$$

where D is the domain score (0 - 5), W is the domain weight expressed in percentage points, and ΣW =100. For example, a score of 4/5 in a 20% domain contributes 16 points.

XSL•FO
RenderX

Table . Practice Intelligence Score-7 (PIS-7) domains and proposed weighting for evaluating the quality of professional learning episodes.

Domain	Construct assessed	Indicators of high-quality engagement	Weight (%)
Professional question intelligence	Ability to frame clinically meaningful, context-aware questions	Defines relevant clinical problems with appropriate clinical context	15
Reasoning depth	Depth of clinical reasoning and uncertainty management	Demonstrates differential reasoning, prioritization, and risk-benefit analysis	20
Contextual integration	Integration of patient, evidence, and system factors	Incorporates patient factors, guidelines, and local practice constraints	15
Iterative inquiry	Refinement of reasoning through follow-up questions	Uses iterative questioning to clarify uncertainty and explore alternatives	15
Critical appraisal	Evaluation of AI-generated information	Requests evidence, challenges unsupported claims, and recognizes limitations	15
Reflective synthesis	Consolidation of learning into actionable insight	Summarizes learning and articulates implications for clinical practice	10
AI literacy and prompt skill	Responsible and effective use of AI tools	Provides context, verifies outputs, and maintains clinical responsibility	10
Total			100

The weighting scheme assigns the highest weight to reasoning depth (20%), with the other 6 domains weighted at 10% to 15% each. This author-derived asymmetry reflects the conceptual centrality of diagnostic and therapeutic reasoning to physician decision-making [10-12]. The weights have not undergone formal consensus-based elicitation (eg, Delphi) and remain provisional pending empirical validation. After ethics approval and consent, a planned calibration workshop will use synthetic anchor PLEs to train 20 to 25 expert raters on the PIS-7 behavioral anchors before blinded scoring of study PLEs. Human-AI agreement and human interrater reliability will then be assessed using prespecified intraclass correlation coefficient (ICC) analyses to refine the scoring rubric and evaluator configuration rather than to rederive the domain weights.

The composite PIS-7 score summarizes observable reasoning behavior within a PLE rather than physician competence. Outputs should include domain-level explanations, transcript-based justification, and confidence indicators to support transparent human interpretation [8,9]. The PIS-7 is not intended to rank physicians or assign labels of intelligence; its formative purpose is to identify episode-specific areas for improvement and support personalized learning [9,15].

A preliminary credit conversion model may classify PLEs into noncreditable (<40), basic (40–59), meaningful (60–74), advanced (75–89), and very high (≥90) engagement tiers, each associated with a proposed base credit unit of 0, 0.25, 0.5, 0.75, and 1, respectively—consistent with the 0.25 - 1 base credit unit range specified in the associated patent claims—with final CPD credit equal to this unit multiplied by a jurisdiction-specific scale factor. Tier labels refer only to engagement within a PLE. These thresholds are conceptual pending empirical calibration; Multimedia Appendix 1 provides 3 illustrative, hypothetical worked examples spanning a nonqualifying interaction and lower- and higher-scoring qualifying PLEs, with domain-level scores, rationale, and proposed credit equivalence. Table 3 summarizes the conceptual workflow from interaction screening to feedback.

XSL•FO
RenderX

Table . Conceptual screening, scoring, and credit workflow in the Case-based Learning Intelligence Credit System (CLICS).

Stage	Operational question or action	Output
PLE[a] eligibility gating	Does the interaction demonstrate a clinically meaningful problem, iterative reasoning, contextual or evidentiary integration, and reflective synthesis? Single factual lookups, repetitive prompting, or passive copying do not qualify.	Qualifying PLE, or no PIS-7[b] score and no credit
PIS-7 scoring	Score 7 reasoning-related domains from 0 to 5 using the proposed weights.	Domain profile and weighted 0 to 100 composite
Blinded human reference and calibration	For phase 0, use a primary set of at least 150 eligible PLEs, with ≥3 independent expert ratings per PLE and AI scores hidden until human scoring is final; if >150 PLEs are available, select 150 by score-stratified random sampling.	Human reference score, human-AI agreement, interrater reliability, and poststudy refinement of scoring anchors
Conceptual credit conversion	Map the composite score to an author-derived engagement tier and base credit unit only as a proposed future conversion model pending empirical validation and regulatory approval.	Illustrative base credit unit (0 - 1) × jurisdiction-specific scale factor; no actual continuing medical education/continuing professional development credit in phase 0
Feedback and next cycle	Use domain-level results to identify episode-specific areas for improvement.	Personalized feedback and subsequent PLEs within the closed learning loop

[a]PLE: professional learning episode.

[b]PIS-7: Practice Intelligence Score-7.

Validation Roadmap

A framework that translates AI-mediated learning interactions into CME/CPD credit requires systematic validation of content validity, scoring reliability, construct validity, and implementation validity. Content validity concerns whether the PLE construct and PIS-7 domains adequately represent the intended reasoning-related learning behaviors [8,9]; scoring reliability requires comparison of AI-generated scores with human expert ratings using measures such as the ICC, weighted kappa, and Bland-Altman analysis [22-24]. A recent comparison of ChatGPT (OpenAI) and faculty scoring of formative medical education assessments reported 67% overall exact agreement [25], reinforcing the need for human calibration. Construct validity must establish that higher scores reflect deeper reasoning rather than verbosity [10,11], while implementation validity concerns feasibility, scalability, and auditability.

In the phase 0 pilot study, we plan to enroll up to 75 licensed physicians, each contributing up to 3 PLEs (maximum 225 PLEs), together with an expert-rater pool of 20 to 25 physicians or medical educators. The primary reliability analysis requires a minimum evaluable set of 150 eligible PLEs, each independently scored by at least 3 blinded expert raters; the mean of the assigned human ratings (minimum 3) will serve as the human reference score. Phase 0 is designed to evaluate workflow feasibility and preliminary PIS-7 scoring reliability/agreement; study scores will not be used to award actual CME/CPD credit, affect licensure, or make decisions about physician competence.

If more than 150 eligible PLEs are available, the primary set of 150 will be selected by score-stratified random sampling across low-, middle-, and high-score ranges using locked AI scores before human review; if fewer than 150 are available, feasibility and agreement will be reported as exploratory without claiming validation. Primary human-AI agreement will compare the locked AI composite with the mean human reference using a prespecified absolute agreement ICC with 95% CI, while human interrater reliability will be reported separately; weighted kappa and Bland-Altman analysis will provide secondary agreement checks.

Scoring discrepancies and rubric-clarity feedback will inform subsequent refinement, but the primary evaluator configuration will remain locked throughout the primary analysis. If a critical safety or security change becomes necessary, affected scoring will be paused and the change will be documented under formal change control, with pre- and postchange data separated. These safeguards will preserve the distinction between developmental calibration and post hoc adjustment to observed results.

Governance, Ethics, and Risk Control

CLICS requires governance for transparency, accountability, fairness, data protection, and human oversight. It should function as an educational assessment system rather than a surveillance mechanism: physicians should know what data are collected and how scores are generated, while expert educators and CME regulatory bodies define thresholds, review audits, and adjudicate disputes [8,9,26-29].

The phase 0 pilot study protocol was submitted to the Central Research Ethics Committee (CREC Thailand), under the supervision of the Foundation for Human Research Promotion in Thailand, on June 24, 2026 (case number CREC(S)24.06.69_01). The CREC reviewed the protocol on August 17, 2026, and requested revision for approval; the revised phase 0 protocol (version 4.1) was resubmitted on August 26, 2026, and remains under review at the time of this revision. The study was prospectively registered with the Thai Clinical Trials Registry on August 1, 2026 (TCTR20260801001), following registry submission on July

XSL•FO
RenderX

23, 2026. No participant or expert-rater scoring data collection has begun, and no such data will be collected before CREC approval is obtained.

For phase 0, research PLEs will use hypothetical or general clinical-learning problems without intentional collection of identifiable patient data, medical records, images, or patient outcomes. Research transcripts will be processed within CCME Version 8.0, the digital infrastructure/platform developed and operated by the Center for Continuing Medical Education (CCME), Medical Council of Thailand, and hosted on Thailand's Government Data Center and Cloud Service using a prespecified, locally deployed, open-weight, primary PIS-7 evaluator; transcripts will not be transmitted to external commercial AI APIs or used to train or fine-tune external models. Before the first participant PLE, the evaluator model artifact/build, prompt and rubric versions, inference configuration, and relevant run-time environment will be locked and versioned to support reproducibility and auditability.

AI-mediated transcripts may contain patient-adjacent information and indirect identifiers, creating reidentification risk. CLICS should therefore incorporate data minimization, pseudonymization, secure storage, role-based access, and audit logging [27,28], with General Data Protection Regulation (GDPR)–comparable safeguards; under the GDPR, pseudonymized data that remain attributable to an identifiable person are still personal data [30]. Participation should initially be voluntary and restricted to educational use. AI-based evaluators may also be influenced by linguistic style independent of reasoning quality [26]. Generative AI may produce plausible but incorrect information or sycophantically agree with flawed premises [31-35]. Platform-level safeguards are designed to encourage challenges to unsupported premises, explicit uncertainty, and evidence traceability; these measures aim to mitigate, not eliminate, hallucinations and confirmation bias and require empirical validation.

The framework should also include safeguards against gaming, including autonomous or delegated AI agents generating a PLE without the physician's direct engagement. In any future operational implementation in which qualifying PLEs are converted to CME/CPD credit, reauthentication at the point of credit attribution could strengthen identity assurance. Reauthentication confirms the identity of the individual claiming credit but cannot, by itself, establish that every preceding reasoning step was personally generated by that individual. Additional safeguards may therefore include session-continuity and behavioral-consistency checks, similarity and anomalous-pattern analyses, and credit caps. The reliable technical detection of agent-delegated interactions has not yet been validated; this remains an unresolved governance challenge rather than a solved capability.

Discussion and Policy Implications

CLICS may encourage a shift from participation-based metrics toward observable evidence of reasoning-related engagement, aligning with broader movements toward performance-informed and evidence-oriented assessment while remaining distinct from formal competency assessment [8,17]. This framing enables domain-specific feedback and individualized learning pathways [14,15,21], while the AI literacy domain reflects the growing integration of AI into clinical workflows [4-6,18-20].

CLICS differs in 3 specific architectural respects from simpler structured reflective CME approaches, such as asking a physician, "What did you learn from this interaction?" First, a standardized Socratic rescue mode, implemented through a response integrity middleware pipeline, can activate when, after 4 physician-AI exchanges, the interaction remains nonqualifying or off-topic. Rather than simply supplying an answer, it returns targeted, nonleading Socratic questions to redirect the interaction toward a qualifying learning process. In phase 0, this mechanism is a standardized learning scaffold rather than an experimental arm, and its activation is analyzed only as an exploratory process metric. Second, once an interaction qualifies as a PLE, the evidence-based response protocol is designed to require traceable literature citations and explicit evidence-level or uncertainty markers for factual claims, providing an auditable evidentiary check that unstructured reflection does not offer. Third, the PIS-7 scores 7 distinct reasoning domains rather than a single undifferentiated judgment, enabling longitudinal, domain-level tracking of reasoning-related behavior and cohort-level analysis—a level of granularity that free-text reflection is not designed to provide.

These structural distinctions provide plausible mechanisms by which CLICS may capture signals unavailable from simpler reflective approaches. This remains a theoretical argument, not a demonstrated finding: CLICS has not yet been shown to be empirically superior to structured reflective CME, and direct comparative evaluation is identified as a priority for future research.

Because AI-mediated learning tools have developed largely outside traditional educational oversight structures, CME/CPD accreditors and regulators are increasingly relevant to quality assurance for AI-mediated learning. This is consistent with guidance from the Accreditation Council for Continuing Medical Education (ACCME), including its January 2026 guidance on responsible AI use in accredited continuing education, its April 2026 alert clarifying that responsibility for learner-facing AI-generated content remains with the accredited provider rather than shifting to a vendor, and the recent literature on governance of generative AI in medical education [36-38]. The response integrity middleware pipeline and evidence-based response protocol function as procedural and evidence-traceability safeguards within a single interaction—enforcing citation and evidence-level marking—but do not by themselves constitute the content-validation infrastructure described in ACCME guidance, which includes predeployment validation against defined clinical scenarios, a defined clinical oversight structure with authority to intervene, ongoing monitoring and revalidation, and vendor accountability sufficient for provider oversight. Commercial bias that is subtle and consistent with evidence is difficult to detect at the level of a single response and is distinct from physician reasoning quality, which is what the PIS-7 measures. Phase 0 intentionally prioritizes reproducibility and research data protection by using a single locked, locally deployed, open-weight, primary evaluator rather than randomizing across external providers. In future operational

XSL•FO
RenderX

implementations that incorporate multiple AI providers or knowledge pathways, aggregated PIS-7 and evidence-marker data could support the investigation of systematic models or vendor-specific patterns as 1 supplementary input to independent content validation and accreditor oversight.

A potential strength of CLICS is its alignment with practice-embedded learning: physicians often learn while addressing real problems involving uncertainty [10,13]. CLICS is also designed to mitigate epistemic homogenization by structuring AI-mediated learning around inquiry rather than answer conformity. Beyond the phase 0 evaluation setting, its modular architecture could support multiple models or knowledge pathways to expose learners to alternative interpretations and competing hypotheses; phase 0 intentionally uses a single, locked primary evaluator to maximize reproducibility. In the pilot, the design seeks to support epistemic pluralism through open-ended problems, Socratic counterquestioning, requests for disconfirming evidence, and PIS-7 domains—including iterative inquiry, critical appraisal, contextual integration, and reflective synthesis—which reward reasoned challenge and reflection rather than agreement with a predetermined answer. Repeated PLEs may also provide longitudinal signals relevant to metacognition, including recognition of uncertainty, consideration of alternatives, and self-monitoring of reasoning. These mechanisms aim to support evidence-constrained epistemic pluralism, although their effectiveness in preventing homogenization or improving metacognition remains to be empirically validated.

CLICS has important limitations. Learning traces remain a partial, indirect view of cognition [10,11], and implementation feasibility has not yet been established. The residual evaluator and AI-content risks described above require empirical testing. These limitations support phased validation and cautious interpretation.

From a policy perspective, CLICS is best understood as a complementary pathway that could eventually allow physicians to obtain a portion of CME credit through validated PLEs while continuing established educational activities. Earlier efforts to move beyond time-based participation include the American Nurses Credentialing Center outcome-based continuing education model and competency-based continuing professional development frameworks described in Canada [39,40]. Experience with competency-based assessment in Canada has also highlighted the administrative burden that can accompany frequent assessment and documentation [41]. We do not assume that the use of CLICS will remove such implementation challenges by default. Its plausible points of difference are structural: credit derives directly from a machine-readable transcript rather than a separately authored document, scoring criteria are operationalized as explicit behavioral anchors, and the architecture integrates with AI-mediated learning that physicians already use. Furthermore, because it can engage with novel, nonroutine problems as they arise, CLICS is designed to support CPD at the point of practice in real time rather than only through retrospective reflection on resolved cases. Whether automated capture sufficiently reduces administrative burden and assessor variability to improve sustained adoption remains an empirical question for phase 0 and subsequent phases, not a conclusion asserted here in advance of evidence.

Conclusions

We propose CLICS, a complementary pathway for recognizing AI-mediated, practice-embedded learning through the PLE and PIS-7. The PIS-7 provides an episode-specific summary of observable reasoning behavior, not a definitive measure of physician competence or clinical performance. If validated, CLICS may support adaptive, evidence-informed professional development by linking feedback and iterative learning to auditable, reasoning-related evidence. Its value will depend on empirical validation, human oversight, and governance addressing content integrity, privacy, fairness, and auditability.

Acknowledgments

The author acknowledges the global community of medical educators, clinicians, and regulators working to define responsible approaches to AI in medical education and continuing professional development.

The author used generative AI (ChatGPT, OpenAI) for text generation, editing/language polishing and proofreading, and literature search and summarization during preparation of the original manuscript. During preparation of the revision, the author used both ChatGPT (OpenAI) and Claude (Anthropic) to assist with the analysis of editor and reviewer comments, literature identification and summarization, drafting and restructuring of selected passages, condensation of text, and language refinement. The underlying conceptual architecture of both the Case-based Learning Intelligence Credit System (CLICS) and Practice Intelligence Score-7 (PIS-7) was developed by the author. The author independently reviewed and verified all AI-assisted content, references, interpretations, and final wording and retains full responsibility for the manuscript and its conclusions. Generative AI was not used to conduct participant-level data analysis, generate study code, or determine the research design or selection of research methods.

Funding

Preparation of this manuscript received no external funding. The phase 0 pilot study described herein is funded by the Center for Continuing Medical Education, the Medical Council of Thailand.

XSL•FO
RenderX

Data Availability

No participant-level or expert-rater scoring datasets were generated or analyzed for this viewpoint. The article presents a conceptual framework and proposed validation strategy. Any future sharing of deidentified transcripts, scoring rubrics, evaluator prompts, analytic code, or aggregate datasets will remain subject to ethics, privacy, and institutional governance requirements.

Conflicts of Interest

The corresponding author is the named inventor on a Thai patent application and an international patent application covering the Case-based Learning Intelligence Credit System (CLICS) architecture, both filed May 27, 2026, and on associated Thai trademark applications for CLICS and the Practice Intelligence Score-7 (PIS-7), filed May 25, 2026. Background intellectual property is held by MD Products and Services Co Ltd, a company with which the corresponding author is affiliated. The corresponding author also serves as director of the Center for Continuing Medical Education (CCME), the Medical Council of Thailand, which is funding the planned phase 0 pilot study. The CCME executive committee approved proceeding with the phase 0 pilot study under the director's leadership before any operational implementation of CLICS; the corresponding author was not present and did not participate in the meeting at which this decision was made. These relationships are disclosed for transparency and will be kept current in subsequent empirical, pilot, or implementation studies arising from this work.

Multimedia Appendix 1
Illustrative worked examples of Practice Intelligence Score-7 (PIS-7) scoring and continuing medical education (CME) credit conversion.
[PDF File, 164 KB - mededu_v12i1e99520_app1.pdf]

References

1. Davis DA, Thomson MA, Oxman AD, Haynes RB. Changing physician performance. A systematic review of the effect of continuing medical education strategies. JAMA 1995 Sep 6;274(9):700-705. [doi: 10.1001/jama.274.9.700] [Medline: 7650822]
2. Cervero RM, Gaines JK. The impact of CME on physician performance and patient health outcomes: an updated synthesis of systematic reviews. J Contin Educ Health Prof 2015;35(2):131-138. [doi: 10.1002/chp.21290] [Medline: 26115113]
3. Moore DE Jr, Green JS, Gallis HA. Achieving desired results and improved outcomes: integrating planning and assessment throughout learning activities. J Contin Educ Health Prof 2009;29(1):1-15. [doi: 10.1002/chp.20001] [Medline: 19288562]
4. Wartman SA, Combs CD. Medical education must move from the information age to the age of artificial intelligence. Acad Med 2018 Aug;93(8):1107-1109. [doi: 10.1097/ACM.0000000000002044] [Medline: 29095704]
5. Masters K. Artificial intelligence in medical education. Med Teach 2019 Sep;41(9):976-980. [doi: 10.1080/0142159X.2019.1595557] [Medline: 31007106]
6. Chan KS, Zary N. Applications and challenges of implementing artificial intelligence in medical education: integrative review. JMIR Med Educ 2019 Jun 15;5(1):e13930. [doi: 10.2196/13930] [Medline: 31199295]
7. Cook DA, Levinson AJ, Garside S, Dupras DM, Erwin PJ, Montori VM. Internet-based learning in the health professions: a meta-analysis. JAMA 2008 Sep 10;300(10):1181-1196. [doi: 10.1001/jama.300.10.1181] [Medline: 18780847]
8. Cook DA, Hatala R. Validation of educational assessments: a primer for simulation and beyond. Adv Simul (Lond) 2016;1:31. [doi: 10.1186/s41077-016-0033-y] [Medline: 29450000]
9. Downing SM. Validity: on meaningful interpretation of assessment data. Med Educ 2003 Sep;37(9):830-837. [doi: 10.1046/j.1365-2923.2003.01594.x] [Medline: 14506816]
10. Eva KW. What every teacher needs to know about clinical reasoning. Med Educ 2005 Jan;39(1):98-106. [doi: 10.1111/j.1365-2929.2004.01972.x] [Medline: 15612906]
11. Norman G. Research in clinical reasoning: past history and current trends. Med Educ 2005 Apr;39(4):418-427. [doi: 10.1111/j.1365-2929.2005.02127.x] [Medline: 15813765]
12. Gruppen LD. Clinical reasoning: defining it, teaching it, assessing it, studying it. West J Emerg Med 2017 Jan;18(1):4-7. [doi: 10.5811/westjem.2016.11.33191] [Medline: 28115999]
13. Bowen JL. Educational strategies to promote clinical diagnostic reasoning. N Engl J Med 2006 Nov 23;355(21):2217-2225. [doi: 10.1056/NEJMra054782] [Medline: 17124019]
14. Mamede S, Schmidt HG. The structure of reflective practice in medicine. Med Educ 2004 Dec;38(12):1302-1308. [doi: 10.1111/j.1365-2929.2004.01917.x] [Medline: 15566542]
15. Sandars J. The use of reflection in medical education: AMEE Guide No. 44. Med Teach 2009 Aug;31(8):685-695. [doi: 10.1080/01421590903050374] [Medline: 19811204]
16. Ericsson KA. Deliberate practice and acquisition of expert performance: a general overview. Acad Emerg Med 2008 Nov;15(11):988-994. [doi: 10.1111/j.1553-2712.2008.00227.x] [Medline: 18778378]
17. Epstein RM, Hundert EM. Defining and assessing professional competence. JAMA 2002 Jan 9;287(2):226-235. [doi: 10.1001/jama.287.2.226] [Medline: 11779266]

XSL•FO
RenderX

18. Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. Nat Med 2019 Jan;25(1):44-56. [doi: 10.1038/s41591-018-0300-7] [Medline: 30617339]
19. Rajpurkar P, Chen E, Banerjee O, Topol EJ. AI in health and medicine. Nat Med 2022 Jan;28(1):31-38. [doi: 10.1038/s41591-021-01614-0] [Medline: 35058619]
20. Lin Y, Luo Z, Ye Z, et al. Applications, challenges, and prospects of generative artificial intelligence empowering medical education: scoping review. JMIR Med Educ 2025 Oct 23;11:e71125. [doi: 10.2196/71125] [Medline: 41128430]
21. Schön DA. The Reflective Practitioner: How Professionals Think in Action: Basic Books; 1983.
22. Koo TK, Li MY. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. J Chiropr Med 2016 Jun;15(2):155-163. [doi: 10.1016/j.jcm.2016.02.012] [Medline: 27330520]
23. McHugh ML. Interrater reliability: the kappa statistic. Biochem Med (Zagreb) 2012;22(3):276-282. [Medline: 23092060]
24. Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. Lancet 1986 Feb 8;1(8476):307-310. [Medline: 2868172]
25. Sreedhar R, Chang L, Gangopadhyaya A, et al. Comparing scoring consistency of large language models with faculty for formative assessments in medical education. J Gen Intern Med 2025 Jan;40(1):127-134. [doi: 10.1007/s11606-024-09050-9] [Medline: 39402411]
26. Char DS, Shah NH, Magnus D. Implementing machine learning in health care - addressing ethical challenges. N Engl J Med 2018 Mar 15;378(11):981-983. [doi: 10.1056/NEJMp1714229] [Medline: 29539284]
27. Ethics and governance of artificial intelligence for health: WHO guidance. World Health Organization.: World Health Organization; 2021. URL: https://www.who.int/publications/i/item/9789240029200 [accessed 2026-08-23]
28. Regulatory considerations on artificial intelligence for health. World Health Organization.: World Health Organization; 2023. URL: https://www.who.int/publications/i/item/9789240078871 [accessed 2026-08-23]
29. Recommendation on the Ethics of Artificial Intelligence. United Nations Educational, Scientific and Cultural Organization.: UNESCO; 2021. URL: https://www.unesco.org/en/legal-affairs/recommendation-ethics-artificial-intelligence [accessed 2026-08-23]
30. Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation) (Text with EEA relevance). EUR-Lex. 2016. URL: https://eur-lex.europa.eu/eli/reg/2016/679/oj/ [accessed 2026-08-23]
31. Sallam M. ChatGPT utility in healthcare education, research, and practice: systematic review on the promising perspectives and valid concerns. Healthcare (Basel) 2023 Mar 19;11(6):887. [doi: 10.3390/healthcare11060887] [Medline: 36981544]
32. Kung TH, Cheatham M, Medenilla A, et al. Performance of ChatGPT on USMLE: potential for AI-assisted medical education using large language models. PLOS Digit Health 2023 Feb;2(2):e0000198. [doi: 10.1371/journal.pdig.0000198] [Medline: 36812645]
33. Thirunavukarasu AJ, Ting DSJ, Elangovan K, Gutierrez L, Tan TF, Ting DSW. Large language models in medicine. Nat Med 2023 Aug;29(8):1930-1940. [doi: 10.1038/s41591-023-02448-8] [Medline: 37460753]
34. Kanjee Z, Crowe B, Rodman A. Accuracy of a generative artificial intelligence model in a complex diagnostic challenge. JAMA 2023 Jul 3;330(1):78-80. [doi: 10.1001/jama.2023.8288] [Medline: 37318797]
35. Chen S, Gao M, Sasse K, et al. When helpfulness backfires: LLMs and the risk of false medical information due to sycophantic behavior. NPJ Digit Med 2025 Oct 17;8(1):605. [doi: 10.1038/s41746-025-02008-z] [Medline: 41107408]
36. Guidance on the responsible use of artificial intelligence (AI) in accredited continuing education (CE). Accreditation Council for Continuing Medical Education.: ACCME; 2026. URL: https://accme.org/wp-content/uploads/2026/01/1098_20260130_Guidance_on_Artificial_Intelligence_in_Accredited_CE_ACCME.pdf [accessed 2026-08-23]
37. Urgent alert on the use of AI in accredited CE. Accreditation Council for Continuing Medical Education.: ACCME; 2026. URL: https://accme.org/news/urgent-alert-on-the-use-of-ai-in-accredited-ce/ [accessed 2026-08-23]
38. Tran M, Balasooriya C, Jonnagaddala J, et al. Situating governance and regulatory concerns for generative artificial intelligence and large language models in medical education. NPJ Digit Med 2025 May 27;8(1):315. [doi: 10.1038/s41746-025-01721-z] [Medline: 40425695]
39. Graebe J. Continuing professional development: utilizing competency-based education and the American Nurses Credentialing Center outcome-based continuing education model©. J Contin Educ Nurs 2019 Mar 1;50(3):100-102. [doi: 10.3928/00220124-20190218-02] [Medline: 30835317]
40. Campbell C, Silver I, Sherbino J, Cate OT, Holmboe ES. Competency-based continuing professional development. Med Teach 2010;32(8):657-662. [doi: 10.3109/0142159X.2010.500708] [Medline: 20662577]
41. Cheung K, Rogoza C, Chung AD, Kwan BYM. Analyzing the administrative burden of competency based medical education. Can Assoc Radiol J 2022 May;73(2):299-304. [doi: 10.1177/08465371211038963] [Medline: 34449283]

Abbreviations

ACCME: Accreditation Council for Continuing Medical Education
CCME: Center for Continuing Medical Education

XSL•FO
RenderX

CLICS: Case-based Learning Intelligence Credit System
CME: continuing medical education
CPD: continuing professional development
ICC: intraclass correlation coefficient
PIS-7: Practice Intelligence Score-7
PLE: professional learning episode

Edited by TDA Cardoso; submitted 26.Apr.2026; peer-reviewed by G McMahon, R Cervero, U Khan; revised version received 29.Aug.2026; accepted 31.Aug.2026; published 30.Sep.2026.

Please cite as:
Wattanasirichaigoon S
AI-Mediated Assessment of Continuing Medical Education: The Case-based Learning Intelligence Credit System (CLICS) Framework
JMIR Med Educ 2026;12:e99520
URL: https://mededu.jmir.org/2026/1/e99520
doi:10.2196/99520

© Somkiat Wattanasirichaigoon. Originally published in JMIR Medical Education (https://mededu.jmir.org), 30.Sep.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Education, is properly cited. The complete bibliographic information, a link to the original publication on https://mededu.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

The New Gravity: How Generative AI Remapped Innovation and What Academia Must Do Next

Hossein Akbarialiabad[1,2*], MD, MSc, HMBA; Ayman Grada[3], MD, MS; Sancy A Leachman[1], MD, PhD; David Rhew[4], MD; Alireza Haghighi[5,6,7*], MD, DPhil

[1]Department of Dermatology, Oregon Health & Science University, Portland, OR, United States

[2]Faculty of Medicine and Health, University of New South Wales (UNSW Sydney), Sydney, NSW, Australia

[3]Department of Dermatology, Case Western Reserve University School of Medicine, Cleveland, OH, United States

[4]Healthcare, Microsoft, Redmond, WA, United States

[5]Harvard Medical School, 77 Avenue Louis Pasteur, Boston, MA, United States

[6]Broad Institute of MIT and Harvard, Cambridge, MA, United States

[7]Department of Medicine, Brigham and Women's Hospital, Harvard Medical School, Boston, MA, United States

[*]these authors contributed equally

Corresponding Author:
Alireza Haghighi, MD, DPhil
Harvard Medical School, 77 Avenue Louis Pasteur, Boston, MA, United States

Abstract

Artificial intelligence is rapidly reshaping clinical research, diagnostics, health system operations, and medical education. Frontier generative AI development is increasingly concentrated in private industry, whose computational infrastructure, proprietary datasets, and capital exceed those of most academic institutions. Although not designed exclusively for medicine, these systems increasingly influence clinical applications. This shift raises a question for academic medical centers (AMCs): What role remains when leading AI systems emerge outside universities? This viewpoint examines industry's growing role in generative AI, clarifies AMCs' responsibilities, and proposes a framework for AI governance and literacy across undergraduate, graduate, and continuing medical education (CME). Rather than compete with industry on computational scale, AMCs should reinforce functions for which AMCs have distinctive strengths. These include (1) mechanistic inquiry linking algorithmic outputs to biological and pathophysiologic understanding; (2) interdisciplinary synthesis integrating clinical, ethical, legal, and sociotechnical perspectives; (3) independent and methodologically rigorous validation of clinical AI tools; and (4) institutional stewardship of ethical governance for patient populations that are often underrepresented in commercial datasets. As nonprofit institutions training clinicians, AMCs have a distinct role in safeguarding transparency, fairness, and accountability in AI-enabled health care. We position this 4-role framework to refine AI governance and translational medicine approaches, clarifying which responsibilities AMCs should retain while collaborating with industry. The approach includes longitudinal AI literacy training across undergraduate medical education, graduate medical education, and CME, standardized pathways for evaluating clinical AI, stronger conflict-of-interest safeguards, and complementary partnerships with industry and regulators. We also outline educational assessment methods and safety-monitoring processes, including competency-based assessment across undergraduate medical education, graduate medical education, and CME, postdeployment monitoring, and clear escalation triggers. Embedding verification habits and governance into training and practice can help ensure that generative AI supports clinical reasoning while preserving trust, equity, and scientific rigor. This viewpoint addresses academic leaders, curriculum committees, residency directors, faculty developers, and accreditors. Safe clinical AI requires sustained academic leadership.

(JMIR Med Educ 2026;12:e89741) doi:10.2196/89741

KEYWORDS

education; innovation; AI; generative AI; academic medical centers; industry

The Changing Landscape of AI Innovation

In 2023, industry produced 51 notable machine learning models, compared with 15 from academia; industry-academia collaborations produced a further 21 models [1]. Although these models are not necessarily designed for medicine, this imbalance reflects where computational power, proprietary data, and capital for frontier development are now concentrated [2]. The term "frontier" in this context refers to general-purpose foundation models trained at scale with broad, cross-domain capabilities rather than task-specific clinical algorithms [3]. Their architectural innovations increasingly shape downstream clinical tools through fine-tuning, API integration, and embedded workflow systems [4]. This concentration has only deepened

XSL•FO
RenderX

since then. By 2024, close to 9 in 10 notable AI models originated from industry, and by 2025, the share exceeded 90%, even as academia remained the leading institutional source of highly cited AI research [5,6]. The split is therefore not simply that industry builds more; it is that the 2 sectors are pulling apart along different axes, with industry concentrating on scale and academia concentrating on the kind of scrutiny that scale does not buy.

As a result, the concentration of frontier development in private laboratories indirectly shapes the technical substrate upon which many clinical AI applications are built [7]. For academic medical centers (AMCs), this does not mean disengagement from innovation, but rather a reorientation toward evaluating and governing the clinical instantiations of these broader technological advances [8].

At the same time, the higher education landscape is undergoing its own structural strain. Data from the Federal Reserve Bank of New York document underemployment among recent US college graduates working in jobs that typically do not require a college degree [9]. While not specific to medicine, this trend reflects a broader misalignment between traditional degree pathways and rapidly evolving labor-market expectations in technology-intensive fields [10-12]. For medical education, the implication is not job displacement, but curricular relevance: training programs must ensure that physicians graduate with competencies that remain durable in environments increasingly shaped by algorithmic systems [11,13]. Accreditation bodies increasingly emphasize competency-based education; AI literacy must become part of that competency architecture [14].

Together, these forces pose a profound challenge for AMCs [15]. As the locus of innovation shifts toward industry and the educational pipeline struggles to keep pace with technological change, AMCs may risk marginalization if institutional adaptation lags behind technological transformation [10].

For AMCs, the priority is not to match the computational scale or speed of industry, but to focus on the scientific, educational, and public service functions for which they are structurally best suited (Table 1) [16]. For example, the 2024 AI Index estimated training compute costs of US $78 million for GPT-4 and US $191 million for Gemini Ultra [17]. Analyses suggest that a substantial proportion of newly trained AI PhDs now enter private-sector laboratories [18]. Academic institutions care for diverse patient populations, including Medicaid beneficiaries, Medicare recipients with complex conditions, and newly arrived immigrants. AMCs should evaluate whether clinical AI performs equitably across the populations they serve [19,20]. These asymmetries underscore the need for independent academic oversight to prevent AI deployment from exacerbating existing health inequities.

Effective oversight of clinical AI will require close coordination between academic investigators and national regulatory bodies [21,22]. Academic institutions can generate independent evidence, characterize model limitations, and inform federal standards. Regulators retain statutory enforcement responsibilities, while AMCs remain responsible for institutional oversight of procurement and clinical use [23]. Many industry actors have also strengthened internal validation and compliance infrastructures; however, independent academic evaluation remains essential to mitigate structural conflicts of interest and to preserve public credibility [24,25].

The central question, therefore, is how academic medicine can fulfill its mission when the most powerful AI systems are developed outside university walls. The answer does not lie in replicating industry scale, but in strengthening roles for which AMCs have distinctive strengths: mechanistic inquiry, transparent validation, ethical stewardship, and workforce training. Industry-academia research collaborations can support knowledge cocreation, although differences in priorities and time horizons require explicit management [26]. It is important to recognize that industry is not a monolith; it includes multinational technology corporations, venture-backed startups, nonprofit research laboratories, and public-private partnerships, each operating under distinct governance structures and accountability mechanisms. These developments are constructive and should be welcomed. However, structural differences remain: commercial entities are ultimately accountable to investors and product timelines, whereas academic institutions operate under nonprofit missions centered on patient welfare, education, and public accountability. These differing incentive structures underscore why independent academic evaluation and governance remain essential complements to industry innovation rather than redundant efforts.

This viewpoint seeks to (1) examine the structural shift in generative AI development toward private industry; (2) articulate the distinct and enduring roles of AMCs in ensuring scientific rigor, ethical governance, and educational integrity; and (3) outline a concrete implementation framework for integrating AI literacy and oversight across undergraduate medical education (UME), graduate medical education (GME), and continuing professional development. The intended audience includes academic medical leaders, curriculum designers, residency program directors, faculty development committees, and accreditation stakeholders navigating the integration of AI into clinical training.

XSL•FO
RenderX

Table . The 4 distinctive roles of academic medical centers (AMCs) in clinical AI.

Role	Core function	What to remember (one-liner)	Evaluation anchor/example
1. Foundational mechanistic inquiry	Wet-lab and translational studies that determine whether AI-suggested biomarkers or therapeutic targets reflect true pathophysiologic mechanisms	AMCs explain *why* models work or fail biologically	Experimental validation of AI-generated hypotheses through bench and translational research
2. Interdisciplinary synthesis	Integrates clinical medicine, biomedicine, computer and social sciences, ethics, and law	Evaluates not only what AI *can* do but what it *should* do and when	Multifaculty AI governance councils within AMCs
3. Independent validation and assurance	Serves as an impartial evaluator of bias, reproducibility, and generalizability in clinical AI tools	Provides staged safety and performance assessment across the AI life cycle	Clinical AI validation pipelines modeled on drug trial frameworks
4. Institutional governance and trust stewardship	Defines procurement standards, privacy safeguards, and equity frameworks	Preserves the public's trust in responsible AI care	Dedicated AI safety and governance boards with authority analogous to IRBs[a] or P&T[b] committees, responsible for auditing model performance, procurement decisions, and clinical integration

[a]IRB: institutional review board.

[b]P&T: pharmacy and therapeutics.

Engaging Credible Counterpositions

Several considerations qualify a simple division between academia and industry. Industry-academia collaborations can support knowledge cocreation [26]. The examples below illustrate ways to combine academic expertise with industry resources. Such arrangements should be assessed on their methods, governance, and safeguards for independent evaluation.

These developments illustrate that the landscape is neither static nor binary. However, even in collaborative or hybrid models, independent academic oversight remains structurally important to preserve methodological neutrality, mitigate conflicts of interest, and maintain public trust. In other words, the question is not whether AMCs should collaborate, but where they must retain independence to evaluate and govern.

Several collaborations illustrate how academic evaluation and industry-scale deployment can be combined without collapsing independence. In radiology, an academic AI development and evaluation laboratory has paired with a large clinical network to translate academic validation methods into life cycle monitoring tools meant for use well beyond a single site, combining research-grade evaluation with deployment experience across thousands of clinical locations [27]. A national specialty society, working with a commercial data partner, has launched an assessment laboratory that evaluates cardiovascular risk algorithms against safety and efficacy standards before health systems adopt them [28]. The Dana-Farber Cancer Institute implemented GPT4DFCI for operations, administration, and research, with direct clinical care explicitly excluded [29]. Further, precompetitive research consortia in the mold of the MIT-IBM Watson AI Lab have pooled academic and corporate resources for shared, publishable work [30]. These arrangements are productive, and the argument here is not that academic medicine should stand apart from them. Academic partners can strengthen these collaborations by providing evaluation protected from sponsor influence through transparent funding, conflict-of-interest safeguards, and the right to publish unfavorable findings.

Steps Forward: Four Academic Functions That Do Not Require Frontier-Scale Compute

Before turning to the 4 roles, it is worth being precise about why these functions resist substitution, since the claim that they are distinctively academic has to be argued rather than asserted. Industry and hybrid bodies have not been idle on governance. The Coalition for Health AI, a multistakeholder group spanning health systems, technology companies, patient advocates, and academic centers, has issued consensus best-practice guides for trustworthy AI, and its work with the Joint Commission produced sector guidance on responsible use [31]. Specialty societies have begun to establish assessment laboratories of their own, and most large vendors now maintain internal validation and audit functions. These efforts have value; the question is whether they can substitute for institutionally independent oversight. The trajectory of the Coalition's flagship proposal is instructive. Fierce Healthcare reported that its proposed national network of AI assurance laboratories was discontinued as the Coalition shifted its focus toward AI governance and postdeployment monitoring [32]. The design and sustainability of assurance arrangements require attention alongside evaluation methods. Institutional independence is central to credible evaluation, but it is not guaranteed by nonprofit status. AMCs must protect independence through transparent funding arrangements, conflict-of-interest controls, access to data, and the right to publish unfavorable findings. Commercial organizations may face investor expectations and product deadlines, whereas AMCs also carry obligations to patients, trainees, and the public. AMCs should preserve the

XSL•FO
RenderX

ability to publish null or unfavorable findings without sponsor interference.

The 4-role framework that follows is best understood as an extension of an existing model rather than a wholly new one. The translational science spectrum, conventionally described as the T0 through T4 phases that carry a discovery from basic mechanism to population-level effectiveness, already gives medicine a vocabulary for staged, rigorous evaluation, and recent work has begun to map AI and machine learning development onto it, placing exploratory model work at the early stages and clinical-grade tools at T2 and beyond [33]. The roles proposed here track that spectrum: mechanistic inquiry corresponds to the early, explanatory end; independent validation to the clinical testing phases; and governance and trust stewardship to the implementation and population phases, with interdisciplinary synthesis running across all of them as a connective function. This framework also sits alongside, rather than in competition with, the process-oriented governance frameworks now in circulation: the National Institute of Standards and Technology (NIST) AI Risk Management Framework, the World Health Organization's guidance on AI for health, the Coalition for Health AI blueprints, and the various health care AI governance maturity models. Those frameworks specify what activities a responsible actor should perform; what they largely leave open is which actor is structurally suited to perform which activity. The contribution of the 4-role framework is to make that actor-specific argument explicit for academic medicine: it recasts the translational spectrum not as a research pipeline but as a distribution of institutional responsibilities, and identifies the subset that cannot be delegated without losing the independence that gives it meaning.

First, AMCs must champion foundational mechanistic inquiry, addressing a limitation of many predictive AI systems: associations do not, by themselves, establish causal mechanisms [34]. Although industry focuses on commercially viable applications, academia can explore foundational questions: when an AI model predicts disease progression, does it reflect true biological pathways or statistical artifacts [35]? Understanding why a prediction works is crucial for advancing beyond pattern recognition to genuine medical insight and determining when a tool can be safely generalized. This translational bridge between computational prediction and biological mechanism remains a domain where academic laboratories possess deep methodological expertise.

The importance of mechanistic grounding is particularly evident in areas such as oncology, radiology, and predictive analytics for chronic disease [36]. An algorithm may identify imaging patterns associated with malignancy or electronic health record variables predictive of sepsis; however, without understanding whether these signals correspond to underlying biological pathways, causal drivers, or merely correlational artifacts embedded in the training data, clinical generalization becomes precarious [37]. History offers cautionary examples of models that performed well in development cohorts but failed when underlying workflows or demographic distributions shifted. Academic environments are uniquely positioned to interrogate these discrepancies through prospective trials, translational validation studies, and integration with laboratory-based investigation. By linking model outputs to mechanistic hypotheses and testing them across diverse populations, AMCs can move clinical AI from statistical association to scientifically interpretable insight [35].

Fulfilling this responsibility represents a distinctive contribution to academic leadership. Commercial developers can also fund mechanistic work, although research priorities may differ across organizations and projects. Academic investigators can make mechanistic explanation a primary scientific objective, drawing on institutional access to basic bench science, tissue repositories, and physiologic models. The well-documented case of a pneumonia-detection model that had partly learned to recognize hospital-specific image features correlated with disease prevalence serves as a cautionary reminder that statistical performance and clinical understanding can diverge quietly [37]. Exposing such hazardous divergences requires independent evaluators with a mandate and safeguards supporting independent scrutiny.

Second, AMCs are uniquely positioned to foster the interdisciplinary synthesis required to evaluate AI as a sociotechnical system [38]. Integrating expertise from medicine, computer science, ethics, law, and sociology is essential for assessing the real-world impact of AI on patient outcomes, clinical workflows, and health equity [39]. This holistic perspective is a core academic strength that can complement industry expertise in development and deployment. Evaluating AI as a sociotechnical system requires attention not only to model accuracy but also to workflow integration, liability implications, clinician cognitive load, and patient perception [40]. For example, a hypothetical AI triage tool that modestly improves diagnostic speed may simultaneously alter physician-patient communication patterns or shift responsibility boundaries in ways that are not captured by performance metrics alone. Academic institutions, with embedded departments of ethics, law, public health, and social science, are structurally equipped to examine these second-order effects. This interdisciplinary synthesis allows AMCs to assess not only whether an AI system works, but also how it reshapes clinical culture and institutional accountability [41].

Third, AMCs must strengthen their role as transparent evaluators of clinical AI [42]. Although many validation studies today are supported by industry, a legitimate source of concern regarding conflicts of interest, academic institutions remain essential for rigorous, methodologically independent assessment [43,44]. To strengthen methodological independence, we recommend explicit institutional firewalls, standardized reporting, and funding structures that reduce sponsor influence on study design, analysis, and publication. To improve transparency and comparability across studies, AMC-led evaluations should follow reporting guidelines appropriate to the study type: TRIPOD+AI (Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis—AI) for prediction-model studies [45], CONSORT-AI (Consolidated Standards of Reporting Trials–AI) for clinical trial reports [46] or SPIRIT-AI (Standard Protocol Items: Recommendations for Interventional Trials–AI) for clinical trial protocols [47]. Evaluation pipelines should include (1) external validation across settings and populations, (2) stress-testing for dataset

XSL•FO
RenderX

shift and performance drift, and (3) explicit reporting of failure modes and subgroup performance because model behavior frequently changes across time, sites, and workflows [48]. Embedding these processes within institutional governance structures, similar in authority to institutional review boards (IRBs) or pharmacy and therapeutics committees, can help normalize AI oversight as a routine component of clinical safety rather than an ad hoc research activity [49]. Independence matters because otherwise evaluation becomes vulnerable to subtle distortions, including outcome measures chosen to flatter, comparator groups selected for contrast, and unfavorable subgroup performance left unpublished.

In practical terms, this may involve the establishment of dedicated Clinical AI Oversight Committees with multidisciplinary representation, including clinicians, data scientists, ethicists, legal scholars, and patient advocates [50]. Such bodies could evaluate procurement decisions, require standardized performance reporting, mandate periodic postdeployment audits, and ensure mechanisms for clinician override and incident reporting [51]. Importantly, governance should extend beyond initial approval to continuous monitoring, recognizing that model performance may drift as data distributions evolve [48]. Embedding AI oversight within institutional quality and safety infrastructures reframes algorithmic evaluation as an extension of patient safety rather than a purely technical exercise [52].

Finally, although academic medicine has its own history of ethical lapses, including research misconduct, conflicts of interest, and the exploitation of trainees, its nonprofit mission, public accountability, and responsibility for training future clinicians position AMCs as essential stewards of ethical AI oversight. This role does not assume moral superiority but recognizes that AMCs operate under public-facing obligations that differ structurally from those of commercial entities. Operationalizing this stewardship will require transparent governance frameworks, clearer conflict-of-interest policies, and institutional mechanisms that prioritize patient welfare over commercial incentives. This includes establishing standards for AI procurement, protecting patient privacy, ensuring human oversight in clinical decision-making, and addressing the algorithmic biases that can perpetuate health disparities.

Re-Educating Clinicians for the Age of AI

To realize this vision, AMCs must adopt a concrete, 5-step action blueprint (Textbox 1). As illustrated in Textbox 1, this framework requires (1) longitudinal curriculum reform across UME, GME, and continuing medical education (CME); (2) standardized model evaluation pipelines for continuous clinical auditing; (3) precompetitive joint labs to develop shared benchmark datasets; (4) stackable microcredentials tied directly to clinical pathways; and (5) faculty upskilling combined with formal governance charters. These 5 steps reinforce one another, creating a longitudinal model of AI stewardship embedded within academic medicine. The framework also requires new partnership models that move beyond traditional sponsored research to create deeply integrated joint labs and data-sharing consortia, enabling academic researchers to work with industry-scale resources on precompetitive, public-interest challenges [30]. At the undergraduate level, AI literacy should emphasize 3 competencies: understanding model architecture at a conceptual level, recognizing common failure modes (including hallucination and bias amplification), and developing structured verification habits [53]. Assessment formats such as case-based critique of AI-generated clinical reasoning, simulated documentation review, and explicit "human-in-the-loop" reflection exercises can reinforce these competencies without transforming medical students into data scientists. Importantly, AI education should not be framed as technical supplementation but as an extension of professional formation [54]. In UME, this may include integrating AI literacy into preclinical foundational science courses and structured reflection on AI-assisted reasoning during clerkships. In GME, specialty-specific modules can address algorithmic triage, automated documentation, and human-override decision-making in supervised clinical environments. Faculty development programs should establish baseline AI governance competencies for attending physicians responsible for supervision and evaluation.

These 5 components address complementary implementation needs. Each is matched to a specific failure mode that recurs when academic centers approach clinical AI in an ad hoc manner. Curriculum reform addresses a workforce that cannot reliably appraise the tools it is given. Standardized evaluation pipelines address deployment of models that were never tested outside their development settings. Precompetitive joint labs address resource asymmetry that otherwise leaves academic evaluators without representative data. Microcredentials address the mismatch between multiyear curricular cycles and software that changes in months. Finally, comprehensive faculty upskilling paired with formal institutional governance charters addresses the absence of centralized institutional authority required to act on what evaluation reveals. These components are interdependent rather than strictly sequential. Governance authority and faculty development should begin early, while curriculum reform, evaluation pipelines, joint laboratories, and microcredentials mature iteratively. An AI-literate workforce supports rigorous clinical validation, and shared benchmark data can strengthen evaluation across institutions. Targeted microcredentials maintain workforce competencies as models and testing pipelines evolve. A formal governance charter connects these activities to a durable institutional safety function.

A practical way to prevent the blueprint from remaining aspirational is to define the minimum evidence and governance artifacts that a system must produce to earn routine clinical use. At a minimum, every deployed clinical AI tool should have a short "AI label" or model card adapted to clinical practice, stating intended use, population, training and validation cohorts, known failure modes, and required clinician oversight. That label should be paired with a local implementation dossier that specifies workflow integration points, override expectations, and a monitoring plan with explicit triggers for reassessment. These triggers can be simple: performance drift beyond a prespecified margin, sustained increases in override rates, documented safety events, or consistent subgroup disparities. Importantly, prespecified triggers linked to accountable action

XSL•FO
RenderX

convert postdeployment monitoring from passive observation into an active safety process. This is also where AMCs can be distinctive without requiring computational scale. They can set a standard that procurement is conditional, not final: clinical AI tools are adopted provisionally, audited as they are used, and can be restricted or removed when monitoring reveals unacceptable performance. Over time, the institutional habit that matters most is not purchasing better tools, but making deimplementation possible.

Textbox 1. A 5-step blueprint for responsible academic medical center (AMC) action in the era of generative AI.

1. Longitudinal curriculum reform

- Integrate AI literacy and adversarial model evaluation throughout undergraduate medical education, graduate medical education, and continuing medical education, emphasizing clinical reasoning alongside AI outputs

2. Standardized model evaluation pipelines

- Implement institution-wide audit procedures for bias, reproducibility, external validity, and clinical safety. Publish results openly

3. Precompetitive joint labs and shared benchmark datasets

- Develop transparent partnerships where shared benchmarks and evaluation protocols (not proprietary products) are the outcomes

4. Stackable microcredentials tied to clinical pathways

- Issue AMC-based certifications such as “Clinical AI Safety,” “Responsible Large Language Model Use in Care,” and “AI-Augmented Clinical Workflows,” integrated into residency and fellowship training

5. Faculty upskilling and governance charters

- Create annual faculty certification in AI system interpretation and ethical oversight, and publish institutional AI governance charters defining decision-making authority and accountability

The barriers to implementing this model are predictable and should be named explicitly. First is resourcing. Sustained evaluation and monitoring require staff time, data access, and analytic capacity that many AMCs do not yet fund as a core safety function. Second is authority. Oversight groups that can only recommend but cannot condition procurement or require revalidation tend to become symbolic. Third is fragmentation. If evaluation sits in isolated departments rather than in quality and safety operations, results are harder to translate into institutional action. Fourth is faculty capability. Education reforms fail when supervisors are expected to assess competencies they do not themselves hold. These barriers point to realistic mitigation strategies. AMCs can start with a narrow portfolio of high-stakes systems, treat oversight as part of the patient safety infrastructure, and adopt a governance charter that defines procurement conditions, audit intervals, and escalation pathways. Faculty upskilling should be treated as a supervisory competence requirement, not as optional enrichment. Finally, AMCs can share the burden through consortia. Shared benchmark datasets, shared evaluation protocols, and pooled approaches to postdeployment monitoring reduce duplication and improve external validity without requiring each institution to build an assurance laboratory from scratch.

Progress can be made legible through a small number of measures: the proportion of supervising faculty who have completed formal AI safety training, the share of clinically deployed models that have passed an institutional evaluation, completion of an independent assessment before routine clinical deployment, followed by timely reassessment after material changes, and the existence of a chartered oversight body with documented authority over procurement and postdeployment audits. A major constraint is faculty capacity, since there are still few clinician-educators equipped to teach this material. Other barriers include the administrative load of standing up new committees, competition for protected curricular time, and the risk, already visible in early institutional experience, that an AI governance committee without a clearly bounded scope becomes overwhelmed and slow rather than authoritative [55].

Just as evidence-based medicine reshaped curricula in the late twentieth century, AI literacy must be embedded in clinical reasoning, ethics, and patient communication [56]. The objective is not automation dependency but reflective practice: clinicians who understand both the power and limitations of algorithmic systems and can articulate their use transparently to patients [57]. Generative AI represents a similar inflection point: clinicians must learn to interrogate algorithmic reasoning, recognize when outputs conflict with clinical context, and communicate uncertainty effectively to patients [58]. Embedding these competencies within professional identity formation can help ensure that AI becomes an instrument of augmentation rather than substitution [54,58]. The long-term legitimacy of AI-assisted care will depend, in part, on whether clinicians can articulate the rationale for AI-informed decisions and maintain patient trust in shared decision-making processes [59].

The proposed educational content can be made explicit across the training continuum. Drawing on the principles for AI in medical education issued by the Association of American Medical Colleges and on competency frameworks proposed in recent reviews and Delphi work [60-62], the relevant content can be organized into 4 domains that hold across the training continuum: foundational understanding of how models are built and where their outputs come from; practical and applied skills in using and supervising AI tools within real workflows; critical

XSL•FO
RenderX

appraisal and verification, including recognition of failure modes such as hallucination, automation bias, and bias amplification; and ethical, legal, and communication competence, covering accountability, consent, equity, and how to convey algorithmic uncertainty to patients. These domains do not change from one stage of training to the next; what changes is their depth and the setting in which they are taught and assessed. Table 2 sets out the progression across UME, GME, and CME, pairing each stage with an assessment approach so that AI literacy is examined rather than merely delivered.

Table . A structured AI literacy competency outline across the medical education continuum.

Competency domain	Undergraduate medical education	Graduate medical education	Continuing medical education	Proposed assessment approach
Foundational understanding of AI systems	Conceptual model architecture; data provenance; what model training does and does not mean	Model types relevant to the specialty and the assumptions built into them	Briefings on new classes of models as they enter routine practice	Concept checks; brief structured written or oral items
Practical and applied use	Supervised exposure to AI-assisted reasoning during clerkships	Supervised use of triage, documentation, and decision-support tools	Procurement literacy and safe local adoption in one's own setting	Observed use with documented override decisions
Critical appraisal and verification	Recognizing hallucination, automation bias, and bias amplification; structured verification habits	Appraising validation evidence; detecting performance drift in local data	Reading model cards and AI reporting standards (TRIPOD+AI[a], CONSORT-AI[b])	Case-based critique of AI-generated clinical reasoning
Ethics, equity, law, and communication	Accountability, consent, and equity; communicating algorithmic uncertainty to patients	Documentation standards; escalating suspected model error; awareness of liability	Institutional governance roles; participation in postdeployment audit	Reflection exercises; portfolio-based or audit-based assessment

[a]TRIPOD+AI: Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis—AI.

[b]CONSORT-AI: Consolidated Standards of Reporting Trials—AI.

Two design features of this structural architecture are worth explicit emphasis. First, competency-based assessment is treated as integral rather than optional. Any clinical competency that is taught but never examined tends not to survive contact with an already saturated curriculum. Consequently, each educational stage is tied to a proposed assessment approach: case-based critiques of AI-generated clinical reasoning, structured "human-in-the-loop" exercises in UME, supervised use with documented override decisions during GME, and formal audit-style or portfolio assessment in CME. Second, implementation is purposefully incremental and relies on existing structures to avoid the administrative resistance of creating standalone courses. AI literacy integrates into UME through existing foundational science modules and clerkship coursework rather than as a standalone course; GME can integrate algorithmic triage, automated documentation, and clinical override decision-making into specialty-specific teaching and into morbidity and mortality or quality-improvement conferences. Finally, CME can leverage modular, stackable microcredentials that practicing clinicians complete without leaving practice. Faculty development remains the critical rate-limiting step; AMCs must reframe AI governance proficiency as a baseline supervisory competence requirement for attending physicians rather than an optional faculty enrichment exercise, given that trainees cannot be validly assessed by supervisors who lack these core competencies themselves.

To effectively engage with an AI-enhanced world, faculty across disciplines must become digitally fluent, not only in using generative AI tools but also in understanding their architectures, limitations, and ethical implications [7]. In GME, specialty-specific integration becomes critical. Dermatology provides one example of a specialty in which AI image classifiers have demonstrated diagnostic potential [63]. Training must therefore include explicit discussion of when algorithmic outputs should be overridden, how to document AI-assisted decisions transparently, and how to escalate concerns about model error [52,58]. Embedding AI discussion within morbidity and mortality conferences or quality-improvement rounds may provide a natural institutional mechanism for cultivating a culture of oversight [64,65].

Academic institutions in both high-income countries and low- and middle-income countries (LMICs) face these pressures to embrace AI, though with distinct challenges. For LMICs, public-private partnerships and AI-focused centers of excellence may provide important pathways to meaningful participation in the global innovation ecosystem [7,66,67]. AI may support personalized learning in higher education [68]. By responding to learners' prior knowledge, performance, and progress, AI tools may support tailored content, adaptive assessments, and timely feedback [69]. For CME, modular microcredentials in clinical AI safety, procurement literacy, and bias evaluation may offer a scalable pathway for practicing physicians to remain current. Such approaches acknowledge that AI systems will evolve more rapidly than traditional curricular cycles, requiring ongoing professional adaptation rather than one-time exposure during training [70]. Institutional metrics may help translate these educational reforms into measurable progress. Examples include tracking faculty completion of formal AI safety training, integrating AI oversight modules into residency accreditation reviews, and establishing cross-disciplinary committees to periodically reassess deployed systems. Without such structural reinforcement, AI education risks remaining aspirational rather than operational [71].

XSL•FO
RenderX

The shape of this work is not the same everywhere, and the difference between high-income and lower-income settings is worth making concrete. In high-income systems, access to tools is rarely the binding constraint. Dense electronic health records, an active regulatory apparatus, the evolving device pathway of the US Food and Drug Administration, the European Union AI Act, and a well-capitalized vendor market mean that the academic role is weighted toward validating and governing an already-crowded field of deployed products. Conversely, in many LMICs, the binding constraints sit earlier in the translational life cycle: highly fragmented, noninteroperable data systems, limited computational and connectivity infrastructure, and a shortage of trained clinician-informaticians. Policy and digital-health initiatives can support implementation, but pilot projects may face barriers to sustained integration into local health systems. Technically successful AI pilots targeting conditions such as tuberculosis, HIV, or maternal conditions may not be sustained after external funding ends without integration into local health systems and budgets [23]. For academic centers within these resource-constrained settings, institutional priorities must focus on capacity building, establishing regional centers of excellence, and anchoring regional, South-South, and North-South training partnerships to build local data science expertise, rather than on policing a saturated tool market. The 4 foundational academic roles are broadly applicable, with priorities and implementation determined locally.

Conclusions

If AMCs fail to adapt, the trajectory of AI-enabled clinical practice will increasingly be shaped by proprietary systems whose underlying assumptions, biases, and optimization objectives remain largely invisible to clinicians. To prevent the erosion of public trust and ensure that algorithmic advances serve patient-centered goals, AMCs must actively assert their oversight role rather than withdraw from it. Over the next decade, the defining measure of academic medicine will not be its access to computational scale but its capacity to institutionalize responsible AI integration. By embedding rigorous validation pathways and prioritizing AI literacy for both faculty and trainees, institutions can shape national standards rather than reactively adapt to externally defined norms.

Ultimately, framing the future of medical AI as a competition between academia and industry is a false dichotomy. A more productive, complementary relationship is essential for responsible innovation (Table 3). This transformation mirrors the historical industrialization of pharmaceutical development: academic medicine did not attempt to compete with industry's manufacturing capabilities; instead, it focused on understanding disease mechanisms, conducting independent clinical trials, and advocating for patient safety. By adopting a similar complementary posture today, AMCs can help shape the next era of health care, helping ensure that generative AI augments clinical reasoning while preserving the foundational human elements of medical practice.

Table . Complementary roles of industry and academia in medical AI innovation[a].

Dimension	Industry focus	Academic medicine focus
Primary driver	Commercial ROI[b] and scale	Scientific understanding and public health
Core strength	Rapid development and computational scale	Mechanistic inquiry, validation, and clinical integration
Innovation type	Pattern recognition and accuracy	Causal understanding, biological mechanisms, and safety
Data assets	Proprietary datasets and user feedback	Clinical cohorts, longitudinal outcomes, and diverse populations
Governance role	Regulatory compliance, risk management, and product liability	Patient safety standards, equity assessment, transparency, and independent evaluation
Success metrics	Accuracy, revenue, and adoption	Clinical utility, reproducibility, and fairness
Workforce role	Product training	Critical evaluation and clinical judgment
Public trust	Product reliability	Independent verification and bias detection
Shared responsibilities	Collaboration in joint laboratories, data-sharing partnerships, and regulatory engagement	Codevelopment of standards, shared evaluation pipelines, and partnership-driven oversight

[a]The table summarizes illustrative emphases rather than exclusive capabilities or responsibilities.

[b]ROI: return on investment.

Acknowledgments

Based on the Generative AI Delegation Taxonomy (GAIDeT) criteria provided, Claude (Anthropic) was used to assist with proofreading and language editing of the manuscript, including the wording/layout of Textbox 1.

XSL•FO
RenderX

Funding

The authors received no financial support for this work.

Authors' Contributions

Conceptualization: HA, AH
Writing – original draft: HA, AG
Writing – review & editing: HA, AG, SAL, DR, AH
All authors reviewed and approved the final manuscript.

Conflicts of Interest

DR is employed by Microsoft, SAL is employed by Skinbit, and AG is employed by AbbVie. Microsoft, Skinbit, and AbbVie had no role in the conception, drafting, revision, or the decision to submit this manuscript. AH is the cofounder of NucleicAI.

References

1. Industry continues to dominate frontier AI research. Stanford Institute for Human-Centered Artificial Intelligence (HAI), Stanford University. URL: https://hai.stanford.edu/ai-index/2024-ai-index-report/research-and-development [accessed 2025-11-09]
2. Ahmed N, Wahed M, Thompson NC. The growing influence of industry in AI research. Science 2023 Mar 3;379(6635):884-886. [doi: 10.1126/science.ade2420] [Medline: 36862769]
3. Singh Y, Unal O, Ali F, et al. Large-scale foundation models for radiological image analysis: clinical applications, technical challenges, and future directions. J Imaging Inform Med 2026 Jan 13;2026:1-11. [doi: 10.1007/s10278-025-01823-w] [Medline: 41530420]
4. Gvajaia N, Kutchava L, Alavidze L, Yandamuri SP, Pestvenidze E, Kupradze V. Artificial intelligence in the medical field: diagnostic capabilities of GPT-4 in comparison with physicians. J Med Artif Intell 2025;8:20-20. [doi: 10.21037/jmai-24-276]
5. Artificial intelligence index report 2025. : Stanford Institute for Human-Centered Artificial Intelligence (HAI), Stanford University; 2025 URL: https://hai.stanford.edu/assets/files/hai_ai_index_report_2025.pdf [accessed 2026-08-05]
6. Artificial intelligence index report 2026. : Stanford Institute for Human-Centered Artificial Intelligence (HAI), Stanford University; 2026 URL: https://hai.stanford.edu/assets/files/ai_index_report_2026.pdf [accessed 2026-08-05]
7. Ethics and governance of artificial intelligence for health. : World Health Organization; 2021 URL: https://iris.who.int/server/api/core/bitstreams/f780d926-4ae3-42ce-a6d6-e898a5562621/content [accessed 2026-08-05]
8. Timilsina M, Buosi S, Razzaq MA, Haque R, Judge C, Curry E. Harmonizing foundation models in healthcare: a comprehensive survey of their roles, relationships, and impact in artificial intelligence's advancing terrain. Comput Biol Med 2025 May;189:109925. [doi: 10.1016/j.compbiomed.2025.109925] [Medline: 40081208]
9. The labor market for recent college graduates. Federal Reserve Bank of New York. URL: https://www.newyorkfed.org/research/college-labor-market [accessed 2025-11-09]
10. Future of jobs report 2023. : World Economic Forum; 2023 URL: https://www3.weforum.org/docs/WEF_Future_of_Jobs_2023.pdf [accessed 2026-08-05]
11. Miguez-Pinto JP, Garcia-Rosa B, Maggitti-Bezerril M, et al. The medical student of the future: redefining competencies in a transformative era. Front Med (Lausanne) 2025;12:1593685. [doi: 10.3389/fmed.2025.1593685] [Medline: 40589962]
12. Singh V, Singh G, Sharma G. The future of the healthcare workforce in the age of automation. In: Driving Global Health and Sustainable Development Goals With Smart Technology: IGI Global Scientific Publishing; 2024:453-472. [doi: 10.4018/979-8-3373-0240-9.ch019]
13. Simou E. The growing importance of soft skills in medical education in the AI era: balancing humanistic care and artificial intelligence. Int Med Educ 2025;4(4):50. [doi: 10.3390/ime4040050]
14. Common program requirements. Accreditation Council for Graduate Medical Education. 2022. URL: https://www.acgme.org/programs-and-institutions/programs/common-program-requirements/ [accessed 2026-08-05]
15. The Topol Review. NHS Topol Review. 2019. URL: https://topol.digitalacademy.nhs.uk/the-topol-review [accessed 2026-02-15]
16. Angus DC, Khera R, Lieu T, et al. AI, health, and health care today and tomorrow: the JAMA summit report on artificial intelligence. JAMA 2025 Nov 11;334(18):1650-1664. [doi: 10.1001/jama.2025.18490] [Medline: 41082366]
17. Artificial intelligence index report 2024. : Stanford Institute for Human-Centered Artificial Intelligence (HAI), Stanford University; 2024 URL: https://hai-production.s3.amazonaws.com/files/hai_ai-index-report-2024-smaller2.pdf [accessed 2026-08-05]
18. Ahmed N, Thompson NC. What should be done about the growing influence of industry in AI research? Brookings Institution. 2023. URL: https://www.brookings.edu/articles/what-should-be-done-about-the-growing-influence-of-industry-in-ai-research/ [accessed 2026-08-05]
19. Agyemang CO, Boakye RA, Gyamfi G. Data-driven Medicaid reform to advance equitable healthcare access for immigrant populations across US states. Int J Res Publ Rev 2025;6(5):252-273. [doi: 10.55248/gengpi.6.0525.1882]

XSL•FO
RenderX

20. Santos PMG, Narayan A, Hong AS, et al. Landscape of emergency Medicaid and health care coverage for undocumented immigrants in the US. JAMA Intern Med 2025 Jul 1;185(7):866-873. [doi: 10.1001/jamainternmed.2025.0604] [Medline: 40354040]
21. Warraich HJ, Tazbaz T, Califf RM. FDA perspective on the regulation of artificial intelligence in health care and biomedicine. JAMA 2025 Jan 21;333(3):241-247. [doi: 10.1001/jama.2024.21451] [Medline: 39405330]
22. Goktas P, Grzybowski A. Shaping the future of healthcare: ethical clinical challenges and pathways to trustworthy AI. J Clin Med 2025 Feb 27;14(5):1605. [doi: 10.3390/jcm14051605] [Medline: 40095575]
23. Joseph J. From pilot to policy: why AI health interventions fail to scale in developing countries. Front Digit Health 2026;8:1699005. [doi: 10.3389/fdgth.2026.1699005] [Medline: 41694175]
24. Jordana J, Triviño‐Salazar JC. Turning vulnerability into strength: how independent regulatory agencies enhance accountability and build stakeholder trust. Regul Gov 2026 Jul;20(3):832-846. [doi: 10.1111/rego.70053]
25. Ranjitsingh LM, Rao TVS. Establish legal and regulatory standards for the testing and validation of AI systems to ensure their reliability and safety in operational environments. Int J Syst Assur Eng Manag 2025 Oct;16(10):3338-3353. [doi: 10.1007/s13198-025-02865-7]
26. Marijan D, Sen S. Industry–academia research collaboration and knowledge co-creation: patterns and anti-patterns. ACM Trans Softw Eng Methodol 2022 Jul 31;31(3):1-52. [doi: 10.1145/3494519]
27. Stanford radiology AI development and evaluation lab announces strategic partnership with radiology partners to advance AI safety and reliability in medical imaging. Stanford Medicine. 2026 Jan 28. URL: https://med.stanford.edu/radiology/news/2026-news/stanford-radiology-ai-development-and-evaluation-lab-announces-s.html [accessed 2026-05-08]
28. The American Heart Association AI TORCH, powered by Dandelion Health: testing outcomes and research in clinical healthcare. American Heart Association. URL: https://www.heart.org/en/professional/quality-improvement/american-heart-association-ai-torch [accessed 2026-09-30]
29. Private and secure generative AI tool supports operations and research at Dana-Farber. Dana-Farber Cancer Institute. 2024 Mar 28. URL: https://www.dana-farber.org/newsroom/news-releases/2024/private-and-secure-generative-ai-tool-supports-operations-and-research-at-dana-farber [accessed 2026-09-30]
30. Amini L, Chen CH, Cox D, Oliva A, Torralba A. Experiences and insights for collaborative industry–academic research in artificial intelligence. AI Mag 2020 Mar;41(1):70-81. [doi: 10.1609/aimag.v41i1.5201]
31. Joint Commission and Coalition for Health AI (CHAI) release initial guidance to support responsible AI adoption across U.S. health systems. Joint Commission. 2025. URL: https://www.jointcommission.org/en-us/knowledge-library/news/2025-09-jc-and-chai-release-initial-guidance-to-support-responsible-ai-adoption [accessed 2026-05-08]
32. Beavins E. CHAI's AI oversight ambitions falter with scrapped AI labs. Fierce Healthcare. 2026 Feb 19. URL: https://www.fiercehealthcare.com/ai-and-machine-learning/inside-chais-failed-assurance-labs [accessed 2026-08-05]
33. Aliferis C, Simon G. The development process and lifecycle of clinical grade and other safety and performance-sensitive AI/ML models. In: Artificial Intelligence and Machine Learning in Health Care and Medical Sciences: Best Practices and Pitfalls: Springer; 2024:289-340. [doi: 10.1007/978-3-031-39355-6_6]
34. Pearl J, Mackenzie D. The Book of Why: The New Science of Cause and Effect: Basic Books; 2018.
35. Rudin C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nat Mach Intell 2019 May;1(5):206-215. [doi: 10.1038/s42256-019-0048-x] [Medline: 35603010]
36. Kelly CJ, Karthikesalingam A, Suleyman M, Corrado G, King D. Key challenges for delivering clinical impact with artificial intelligence. BMC Med 2019 Oct 29;17(1):195. [doi: 10.1186/s12916-019-1426-2] [Medline: 31665002]
37. Zech JR, Badgeley MA, Liu M, Costa AB, Titano JJ, Oermann EK. Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: a cross-sectional study. PLoS Med 2018 Nov;15(11):e1002683. [doi: 10.1371/journal.pmed.1002683] [Medline: 30399157]
38. Greenhalgh T, Wherton J, Papoutsi C, et al. Beyond adoption: a new framework for theorizing and evaluating nonadoption, abandonment, and challenges to the scale-up, spread, and sustainability of health and care technologies. J Med Internet Res 2017 Nov 1;19(11):e367. [doi: 10.2196/jmir.8775] [Medline: 29092808]
39. Oduro S, Kneese T. AI governance needs sociotechnical expertise: why the humanities and social sciences are critical to government efforts. : Data & Society; 2024 URL: https://datasociety.net/wp-content/uploads/2024/05/DS_AI_Governance_Policy_Brief.pdf [accessed 2026-08-05]
40. Sendak MP, D'Arcy J, Kashyap S, et al. A path for translation of machine learning products into healthcare delivery. EMJ Innov 2020 Jan. [doi: 10.33590/emjinnov/19-00172]
41. Vayena E, Blasimme A, Cohen IG. Machine learning in medicine: addressing ethical challenges. PLoS Med 2018 Nov;15(11):e1002689. [doi: 10.1371/journal.pmed.1002689] [Medline: 30399149]
42. Beam AL, Kohane IS. Big data and machine learning in health care. JAMA 2018 Apr 3;319(13):1317-1318. [doi: 10.1001/jama.2017.18391] [Medline: 29532063]
43. Redman BK. Conflicts of interest in biomedical, social, and environmental aspects of health. In: Business Ethics in the Healthcare Industry: Springer; 2026:495-510. [doi: 10.1007/978-3-032-07649-6_33]
44. Bauer D, Orchard DA, Day PG, Tunzi M, Satin DJ. A literature review of non-financial conflicts of interest in healthcare research and publication. BMC Med Ethics 2025 May 15;26(1):61. [doi: 10.1186/s12910-025-01221-5] [Medline: 40369522]

XSL•FO
RenderX

45. Collins GS, Moons KGM, Dhiman P, et al. TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. BMJ 2024 Apr 16;385:e078378. [doi: 10.1136/bmj-2023-078378] [Medline: 38626948]
46. Liu X, Cruz Rivera S, Moher D, Calvert MJ, Denniston AK, SPIRIT-AI and CONSORT-AI Working Group. Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI extension. Lancet Digit Health 2020 Oct;2(10):e537-e548. [doi: 10.1016/S2589-7500(20)30218-1] [Medline: 33328048]
47. Cruz Rivera S, Liu X, Chan AW, et al. Guidelines for clinical trial protocols for interventions involving artificial intelligence: the SPIRIT-AI extension. Nat Med 2020 Sep;26(9):1351-1363. [doi: 10.1038/s41591-020-1037-7] [Medline: 32908284]
48. Otles E, Oh J, Li B, et al. Mind the performance gap: examining dataset shift during prospective validation. Presented at: Proceedings of the 6th Machine Learning for Healthcare Conference; Aug 6-7, 2021 URL: https://proceedings.mlr.press/v149/otles21a.html [accessed 2026-08-05]
49. Reddy S, Allan S, Coghlan S, Cooper P. A governance model for the application of AI in health care. J Am Med Inform Assoc 2020 Mar 1;27(3):491-497. [doi: 10.1093/jamia/ocz192] [Medline: 31682262]
50. Meskó B, Topol EJ. The imperative for regulatory oversight of large language models (or generative AI) in healthcare. NPJ Digit Med 2023 Jul 6;6(1):120. [doi: 10.1038/s41746-023-00873-0] [Medline: 37414860]
51. Wells BJ, Nguyen HM, McWilliams A, et al. A practical framework for appropriate implementation and review of artificial intelligence (FAIR-AI) in healthcare. NPJ Digit Med 2025 Aug 11;8(1):514. [doi: 10.1038/s41746-025-01900-y] [Medline: 40790350]
52. Shortliffe EH, Sepúlveda MJ. Clinical decision support in the era of artificial intelligence. JAMA 2018 Dec 4;320(21):2199-2200. [doi: 10.1001/jama.2018.17163] [Medline: 30398550]
53. Davidson AE, Jose A, Shickel B, et al. Building AI competence in the healthcare workforce with the AI for clinical care workshop: a Bridge2AI for clinical CHoRUS project. J Clin Transl Sci 2025;9(1):e243. [doi: 10.1017/cts.2025.10156] [Medline: 41395159]
54. Wartman SA, Combs CD. Reimagining medical education in the age of AI. AMA J Ethics 2019 Feb 1;21(2):E146-E152. [doi: 10.1001/amajethics.2019.146] [Medline: 30794124]
55. Kim JY, Hasan A, Kueper J, et al. Establishing organizational AI governance in healthcare: a case study in Canada. NPJ Digit Med 2025 Aug 15;8(1):522. [doi: 10.1038/s41746-025-01909-3] [Medline: 40817280]
56. Greenhalgh T, Howick J, Maskrey N, Evidence Based Medicine Renaissance Group. Evidence based medicine: a movement in crisis? BMJ 2014 Jun 13;348:g3725. [doi: 10.1136/bmj.g3725] [Medline: 24927763]
57. Char DS, Shah NH, Magnus D. Implementing machine learning in health care—addressing ethical challenges. N Engl J Med 2018 Mar 15;378(11):981-983. [doi: 10.1056/NEJMp1714229] [Medline: 29539284]
58. Amann J, Blasimme A, Vayena E, Frey D, Madai VI, Precise4Q consortium. Explainability for artificial intelligence in healthcare: a multidisciplinary perspective. BMC Med Inform Decis Mak 2020 Nov 30;20(1):310. [doi: 10.1186/s12911-020-01332-6] [Medline: 33256715]
59. Grote T, Berens P. On the ethics of algorithmic decision-making in healthcare. J Med Ethics 2020 Mar;46(3):205-211. [doi: 10.1136/medethics-2019-105586] [Medline: 31748206]
60. Principles for the responsible use of artificial intelligence in and for medical education. AAMC (Association of American Medical Colleges). 2025. URL: https://www.aamc.org/about-us/mission-areas/medical-education/principles-ai-use [accessed 2026-05-08]
61. Ma Y, Song Y, Balch JA, et al. Promoting AI competencies for medical students: a scoping review on frameworks, programs, and tools. arXiv. Preprint posted online on Jul 10, 2024. [doi: 10.48550/arXiv.2407.18939]
62. Singla R, Pupic N, Ghaffarizadeh SA, et al. Developing a Canadian artificial intelligence medical curriculum using a Delphi study. NPJ Digit Med 2024 Nov 18;7(1):323. [doi: 10.1038/s41746-024-01307-1] [Medline: 39557985]
63. Esteva A, Kuprel B, Novoa RA, et al. Dermatologist-level classification of skin cancer with deep neural networks. Nature 2017 Feb 2;542(7639):115-118. [doi: 10.1038/nature21056] [Medline: 28117445]
64. Bates DW, Levine D, Syrowatka A, et al. The potential of artificial intelligence to improve patient safety: a scoping review. NPJ Digit Med 2021 Mar 19;4(1):54. [doi: 10.1038/s41746-021-00423-6] [Medline: 33742085]
65. Pronovost PJ, Berenholtz SM, Needham DM. Translating evidence into practice: a model for large scale knowledge translation. BMJ 2008 Oct 6;337:a1714. [doi: 10.1136/bmj.a1714] [Medline: 18838424]
66. Akbarialiabad H, Sewankambo NK. Centres of excellence in AI for global health equity—a strategic vision for LMICs. Nature 2024 Jan;625(7995):450. [doi: 10.1038/d41586-024-00113-9] [Medline: 38228792]
67. Akbarialiabad H, Sadeghian N, Haghighat S, et al. The utility of generative AI in advancing global health. NEJM AI 2025 Feb 27;2(3):AIp2400875. [doi: 10.1056/AIp2400875]
68. Holmes W, Bialik M, Fadel C. Artificial Intelligence in Education: Promises and Implications for Teaching and Learning: Center for Curriculum Redesign; 2019.
69. Merino-Campos C. The impact of artificial intelligence on personalized learning in higher education: a systematic review. Trends High Educ 2025;4(2):17. [doi: 10.3390/higheredu4020017]

XSL•FO
RenderX

70. Hunt VM, Sprehe LK, de Lomba WC, et al. What the AI era doctor should know: a scoping review of proposed artificial intelligence competencies for medical education. NPJ Digit Med 2026 May 12. [doi: 10.1038/s41746-026-02761-9] [Medline: 42120544]
71. Monlezun DJ, Omutoko L, Oduor P, et al. Digitalization of health care in low- and middle-income countries. Bull World Health Organ 2025 Feb 1;103(2):148-154. [doi: 10.2471/BLT.24.291643] [Medline: 39882493]

Abbreviations

AMC: academic medical center
CME: continuing medical education
CONSORT-AI: Consolidated Standards of Reporting Trials—AI
GME: graduate medical education
LMICs: low- and middle-income countries
NIST: National Institute of Standards and Technology
SPIRIT-AI: Standard Protocol Items: Recommendations for Interventional Trials—AI
TRIPOD+AI: Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis—AI
UME: undergraduate medical education

Edited by S Brini; submitted 16.Dec.2025; peer-reviewed by LV Niekerk, MBR Mahmoud; revised version received 19.May.2026; accepted 08.Jun.2026; published 19.Oct.2026.

Please cite as:
Akbarialiabad H, Grada A, Leachman SA, Rhew D, Haghighi A
The New Gravity: How Generative AI Remapped Innovation and What Academia Must Do Next
JMIR Med Educ 2026;12:e89741
URL: https://mededu.jmir.org/2026/1/e89741
doi:10.2196/89741

© Hossein Akbarialiabad, Ayman Grada, Sancy A Leachman, David Rhew, Alireza Haghighi. Originally published in JMIR Medical Education (https://mededu.jmir.org), 19.Oct.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Education, is properly cited. The complete bibliographic information, a link to the original publication on https://mededu.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Integrating Large Language Models and Metaverse-Related Technologies in Urology Education: Viewpoint

Peizhe Li[1*], MMed; Huijie Gong[1*], MMed; Heyang Liu[1], MMed; Siyu Han[1], MD; Yiqin Lu[2*], MD; Yongji Yan[1*], PhD

[1]Department of Urology, Dongzhimen Hospital, Beijing University of Chinese Medicine, No. 5, Haiyuncang, Dongcheng District, Beijing, China

[2]Department of Gynecology, Dongzhimen Hospital, Beijing University of Chinese Medicine, Beijing, China

[*]these authors contributed equally

Corresponding Author:
Yongji Yan, PhD
Department of Urology, Dongzhimen Hospital, Beijing University of Chinese Medicine, No. 5, Haiyuncang, Dongcheng District, Beijing, China

Abstract

Urology education must prepare learners to integrate anatomical visualization, diagnostic reasoning, technical performance, communication, and ethical judgment, often under restricted working hours and limited access to high-fidelity cases. This viewpoint argues that large language models (LLMs) and metaverse-related technologies should be integrated as complementary, supervised educational tools rather than treated as autonomous replacements for educators or clinical experience. LLMs can support cognitive apprenticeship, documentation practice, assessment, and communication simulations, whereas extended reality, cloud computing, and virtual digital humans can support spatial, procedural, and remote learning. Our aim is to articulate 3 key messages: first, the educational value of these technologies depends on alignment with defined urological competencies; second, implementation should proceed in stages from governed data infrastructure to intelligent interaction and immersive multimodal training; and third, educational effectiveness must be established through prospective evaluations of transfer to clinical performance, not through technical novelty alone. We emphasize the need for rigorous randomized controlled trials, identify requirements for haptic and biomechanical fidelity, and discuss hallucination, bias, privacy, accountability, and deskilling. The responsible future of intelligent urology education will therefore require collaboration among learners, educators, clinicians, engineers, ethicists, institutions, and professional bodies.

(JMIR Med Educ 2026;12:e93235) doi:10.2196/93235

KEYWORDS

large language model; metaverse; artificial intelligence; AI; medical education; urology education

Introduction

Educational Requirements in Urology

Training a urologist takes more than textbook recall. Residents need a working grasp of pelvic and retroperitoneal anatomy, have to read imaging, endoscopic findings, and pathology together, and must carry out precise operations. The work is also intimate and often distressing. Trainees have to counsel patients through oncological uncertainty, discuss loss of sexual or reproductive function, and manage stomas or urinary incontinence with tact.

However, clinical service alone rarely provides adequate, deliberate practice for these varied competencies. Duty-hour restrictions, patient safety obligations, and ethical limits on novice practice substantially narrow hands-on opportunities in the operating room and outpatient clinic [1]. The move toward advanced laparoscopy and robotic platforms adds a steep learning curve: fine bimanual control and spatial translation are hard to learn safely at the bedside [2]. The problem is clearest in uncommon or sensitive situations, such as penile cancer counseling, talks about sexual function, urinary diversion, fertility preservation, and disclosure of an uncertain cancer diagnosis [3].

Aim and Viewpoint of This Article

This article is a viewpoint based on published work on AI, large language models (LLMs), virtual patients, extended reality (XR), robotic surgery, simulation, and medical education, together with our experience as urological and medical educators. We discuss 3 key points. First, LLMs and metaverse-related technologies cover different but linked parts of training; LLMs can support cognitive and communicative learning, while XR and related systems can support spatial and procedural learning [4]. Second, these tools should be brought in through a staged, competency-based, and human-supervised framework. Third, claims of educational benefit should rest on validated measures and evidence of transfer to clinical performance, not on realism, usability, or model accuracy alone.

XSL•FO
RenderX

Cognitive and Communicative Learning With LLMs

General-Purpose and Specialized LLMs

General-purpose LLMs can explain concepts, generate questions, summarize literature, role-play clinical conversations, and provide feedback on written reasoning [5,6]. Their flexibility makes them useful for early exploratory learning [6,7]. However, a general-purpose model may lack reliable knowledge of local workflows, specialty terminology, guideline updates, or the clinical context required for urological decision-making. It may also produce fluent but unsupported answers.

A specialized or retrieval-augmented model can be limited to curated guidelines, institutional protocols, checked teaching materials, and properly deidentified cases [8]. Fine-tuning or retrieval augmentation may improve specialty relevance, traceability, and consistency, but neither approach guarantees correctness [9]. A tuned model can reproduce errors or biases present in its source data, and a retrieval system can return an outdated, incomplete, or poorly matched document. Accordingly, educators should compare general-purpose and specialized models against an expert-annotated reference set before deployment and should display source provenance, uncertainty, and the date of guideline retrieval whenever feasible [8,10].

Uses in Assessment and Clinical Reasoning

LLMs may assist, but should not independently determine, several components of urological assessment. Potential applications include generating formative questions; evaluating the completeness of a history or operative note; identifying whether a learner considered relevant differential diagnoses; providing structured feedback on documentation; and analyzing simulated consultations for missed questions, unclear explanations, or inappropriate language [11-13]. They may also support oral examinations by presenting sequential information, such as a patient with gross hematuria whose history, urinalysis, imaging, cystoscopy, and pathology findings are released progressively.

In urological oncology, an educational system should teach diagnostic algorithms rather than merely provide final answers. For example, a hematuria case may require risk assessment, urine testing, imaging selection, cystoscopic evaluation, and interpretation of pathology. A prostate cancer case may require integration of prostate-specific antigen, digital rectal examination, multiparametric magnetic resonance imaging, targeted or systematic biopsy, grade group, staging, and patient preferences. A renal mass case may require consideration of imaging phenotype, renal function, tumor complexity, metastatic evaluation, and the trade-off between oncological control and renal preservation. These pathways should be represented as branching decisions with explicit uncertainty and should be linked to expert-verified guidelines and contemporary diagnostic literature [14-16].

The appropriate educational role of an LLM is therefore to make reasoning visible and discussable. It can ask why a learner selected a test, identify missing information, compare plausible options, and prompt reflection. It should not silently replace the learner's reasoning or present one recommendation as universally correct. Representative LLMs illustrate the range of possible applications, from clinical knowledge reasoning and electronic health record processing to virtual patient simulation and guideline-based tutoring (Table 1).

Table . Representative large language models and their potential applications in urology education.

Models	Developer/researcher	Key features	Potential applications in urology education
Med-PaLM 2	Google Research [9]	Clinical knowledge reasoning and performance on USMLE-style[a] evaluations	Generates complex decision-making questions and serves as a virtual examiner for advanced resident training
GatorTron	University of Florida [17]	Trained on large volumes of EHRs[b] and designed to process clinical narrative language	Simulates urological clinic dialogues and medical records for interview and documentation training
AIPatient	Yu et al [18]	EHR-based knowledge graphs and retrieval-augmented generation for virtual patient interviewing	Provides high-fidelity practice for breaking bad news and preoperative communication, with performance assessment
BioMedLM	Stanford University and MosaicML [19]	Biomedical literature comprehension and generation	Supports literature review and research skills development on urological topics
UroBot	Hetz et al [20]	Urology question answering enhanced with EAU[c] guidelines	Provides guideline-based tutoring and knowledge tracing

[a]USMLE: United States Medical Licensing Examination.

[b]EHR: electronic health record.

[c]EAU: European Association of Urology.

XSL•FO
RenderX

Hallucination, Empathy, and Human Supervision

LLM errors are not limited to isolated factual mistakes. They may include input-related errors caused by ambiguous, incomplete, or incorrectly transcribed data; general knowledge errors in which the model produces an inaccurate medical statement; retrieval-related errors caused by irrelevant or outdated sources; reasoning errors in which correct facts are combined incorrectly; and communication errors in which an emotionally inappropriate response is phrased fluently [10]. In a sensitive consultation, an apparently empathic sentence may still be clinically misleading, culturally inappropriate, or insufficiently responsive to the patient's distress.

These limits matter most when an LLM is used to simulate a patient with cancer or sexual health concerns. The system should be restricted to teaching, clearly labeled as simulated, watched by an educator, and blocked from unsupervised clinical advice. Feedback to trainees should cover not only facts, but also empathy, privacy, shared decision-making, and psychosocial needs. The instructor still has to review the scenario, correct errors, and lead the debrief [21-24].

Immersive, Spatial, and Procedural Learning

XR and Surgical Simulation

The educational value of virtual reality, augmented reality, and mixed reality lies in repeated practice, controlled variation of difficulty, spatial visualization, and the possibility of objective data capture [4,25]. Virtual reality may support laparoscopic, endoscopic, and robotic skills, while augmented reality and mixed reality can overlay patient-specific anatomy or imaging information onto a physical environment. Studies of urological simulation have shown gains in selected technical tasks, including transurethral procedures, robotic needle passing, and endoscopic spatial orientation [26-30]. These findings support XR as a training method. However, XR does not, by itself, show shorter operative time, fewer complications, better oncological results, or lasting transfer to independent clinical work.

Haptic and Biomechanical Fidelity

A credible surgical simulator must address more than visual realism. Laparoscopic and robotic skills also depend on force perception, instrument-tissue interaction, tissue deformation, traction, cutting resistance, needle passage, and recognition of abnormal structures. Urological tissues also differ in elasticity, vascularity, fragility, and response to manipulation. A simulator that provides visually appealing anatomy, but unrealistic resistance, may teach inappropriate force application [31].

Technical needs include real-time tissue deformation models, collision detection, instrument friction, variable tissue stiffness, bleeding and smoke, camera-control limits, and latency low enough for force feedback. Haptic systems should be calibrated against expert-defined task requirements and, where possible, against physical or instrumented reference data. Higher fidelity is not always better. The level of fidelity should match the skill being taught, the learner's stage, safety, cost, and evidence of benefit. Future validation studies should report haptic fidelity apart from visual fidelity and test whether each one improves measured performance.

Remote Learning and Virtual Digital Humans

Cloud computing can take on heavy rendering, storage, analytics, and remote access. Remote and virtual surgical teaching may widen access to expert instruction, especially where faculty or equipment are scarce [32-35]. Virtual digital humans can provide an embodied interface for LLM-based patients, mentors, or examiners, combining dialogue with gestures, facial expressions, anatomical models, and scenario-specific behavior [36,37]. Their educational benefit will depend on the accuracy of their clinical content, the appropriateness of their affective responses, the quality of their multimodal interaction, and the extent to which learners perceive the scenarios as educationally authentic.

A further concern is that highly repeatable and technically perfect AI-generated cases may be valued less than real cases containing ambiguity, workflow interruptions, competing priorities, and ethically difficult decisions. Case systems should therefore include controlled variability and realistic clinical dilemmas rather than presenting idealized pathways only [38].

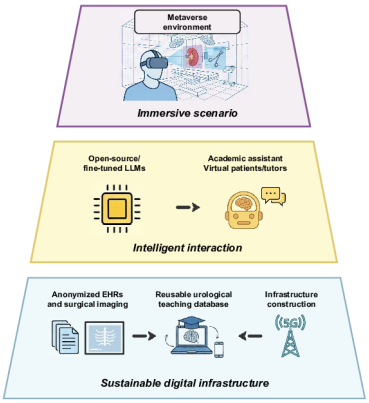
A Staged Framework With a Clinical Example

Data and Governance Foundation

We propose a 3-stage implementation framework that moves from a governed data and infrastructure foundation to intelligent interaction and then immersive training. The framework is summarized in Figure 1. The first stage is the creation of a governed teaching-data infrastructure. Relevant materials may include deidentified histories, laboratory results, imaging, endoscopy, pathology, operative videos, and assessment rubrics. Data should be minimized, access controlled, documented, and reviewed for representativeness. Educational datasets should not be assembled from electronic health records without appropriate institutional approval, deidentification, data use agreements, and safeguards against reidentification [8,10,17,39].

XSL•FO
RenderX

Figure 1. A framework for the synergistic integration of large language models and the metaverse in urology education. EHR: electronic health record; LLM: large language model.



Intelligent Interaction

The second stage is the development of specialty-aware agents that retrieve approved content, present cases progressively, ask targeted questions, and provide explainable feedback. Agents should record the information used to generate an answer, distinguish evidence from inference, and escalate uncertain or unsafe outputs to a human educator. General-purpose LLMs may be used for low-risk brainstorming or language practice, whereas higher-risk diagnostic and assessment tasks should use constrained systems validated against expert standards.

Immersive Multimodal Training

The third stage places these agents in XR environments. Consider a simulated patient with visible hematuria. The learner first takes a history, looks for risk factors and red flags, explains the initial workup, reviews imaging, interprets cystoscopic findings, and discusses the pathology report. The LLM-based patient can show anxiety and ask questions, while the XR environment shows anatomy and procedural steps. The educator can then review the diagnostic sequence, communication, time use, unnecessary tests, and technical performance. A similar prostate cancer case may combine magnetic resonance imaging interpretation, biopsy counseling, staging, treatment discussion, and concerns about sexual function. The aim is not a perfect virtual world; it is to bring cognitive, technical, and humanistic skills together in a setting that is safe and can be assessed [4,26,37].

Proposed Evaluation and Implementation Safeguards

Prospective Randomized Controlled Trials

Rigorous, well-designed randomized controlled trials are essential to establish whether LLM- and metaverse-supported educational tools improve urology training beyond conventional approaches. Future studies should prioritize clinically meaningful and validated outcomes rather than technological novelty, learner satisfaction, or simulator scores alone. Relevant domains include technical performance, diagnostic reasoning, communication, retention, and, critically, transfer of learning to supervised clinical practice. Outcome selection should reflect the specific competency targeted by the intervention and should be informed by established work on diagnostic complexity and objective assessment in XR [31]. Confidence in such evidence will also depend on sufficiently independent assessment and appropriate consideration of variation in learner experience and training context. Equally important, evaluation should examine the durability, generalizability, and potential unintended educational effects of these tools. Evidence of improved performance within a simulated environment, although valuable, should not be considered sufficient to demonstrate clinical competence or justify broad educational implementation.

Ethical, Logistical, and Equity Considerations

LLM and metaverse systems can be computationally expensive, which raises costs and may widen gaps between institutions [40,41]. Sensitive medical data require governance, encryption, auditing, retention limits, and clear responsibility for errors. Biased training data may produce unequal performance across sex, age, ethnicity, language, disability, or socioeconomic groups

XSL•FO
RenderX

[39]. Systems should therefore undergo subgroup validation and continuous monitoring.

Overreliance may also reduce independent reasoning or encourage deskilling [24]. Curricula should require learners to make an initial independent assessment before viewing AI suggestions, explain disagreements with the system, and complete periodic no-AI assessments. Human educators should approve cases, supervise use, evaluate performance, and remain accountable for educational decisions.

Conclusions

Key Messages

LLMs and metaverse-related technologies may strengthen urology education when their roles are clearly differentiated and deliberately integrated. LLMs are most useful for cognitive apprenticeship, documentation, formative assessment, and supervised communication practice. XR, haptics, cloud platforms, and virtual digital humans are most useful for spatial, procedural, and remote learning. Neither technological fluency nor visual realism is equivalent to competence.

Future Directions

The field should move from speculative demonstrations to competency-based implementation and prospective validation. The next generation of studies should report transparent model behavior, haptic and biomechanical fidelity, subgroup performance, educational authenticity, retention, and transfer to clinical practice. A staged, equitable, and human-supervised approach offers a more defensible path toward intelligent urology education than the assumption that automation or immersion will inevitably improve training.

Acknowledgments

YY and YL are co-corresponding authors on this work, and YL can be reached by email at k_lyq@sina.com. During the preparation of this work, the author(s) used ChatGPT 5.6 Sol (OpenAI) and DeepSeek V4 Pro solely for language polishing and improvement of readability. Generative AI (Google Gemini 3.1) was used solely to assist with figure design and layout. It was not used for data analysis, interpretation, or initial manuscript writing. The authors reviewed and approved the final figure. After using these tools/services, the author(s) reviewed and edited the content as needed and take full responsibility for the content of the published article. ChatGPT 5.6 Sol (OpenAI) was used in the preparation of the revised manuscript for language polishing, structural revision, and assistance in drafting responses to reviewer comments. The authors reviewed, revised, and verified the manuscript and remain responsible for its content; no AI tool is an author, and no unpublished patient-identifiable information was entered into a generative AI system.

Funding

This work was supported by the Educational Science Research Project of Beijing University of Chinese Medicine (number XJZ24003) and General Program Research Project of Beijing Higher Education Association (number MS2024379).

Data Availability

Not applicable.

Authors' Contributions

Conceptualization: PL, HG, YY
Data curation: PL, HG, SH
Funding acquisition: YY, HG
Writing – original draft: PL, HG
Writing – review & editing: YL, YY
All authors reviewed the results and approved the final version of the manuscript.

Conflicts of Interest

None declared.

References

1. Hori K, Abe T, Abe N, et al. Gap analysis between trainees' subjective competencies and the competencies expected by instructors in urology: a need assessment survey in Japan. Int J Urol 2024 Jun;31(6):653-661. [doi: 10.1111/iju.15430] [Medline: 38366737]
2. Hughes T, Rai B, Madaan S, Chedgy E, Somani B. The availability, cost, limitations, learning curve and future of robotic systems in urology and prostate cancer surgery. J Clin Med 2023 Mar 15;12(6):2268. [doi: 10.3390/jcm12062268] [Medline: 36983269]
3. Bhat GS, Nelivigi G, Barude V, Shastry A. Sexuality in surgically treated carcinoma penis patients and their partners. Indian J Surg 2018 Feb;80(1):19-23. [doi: 10.1007/s12262-016-1543-5] [Medline: 29581680]

XSL•FO
RenderX

4. Checcucci E, Veccia A, Puliatti S, et al. Metaverse in surgery - origins and future potential. Nat Rev Urol 2026 Jan;23(1):50-63. [doi: 10.1038/s41585-024-00941-4] [Medline: 39349948]
5. Leng L. Challenge, integration, and change: ChatGPT and future anatomical education. Med Educ Online 2024 Dec 31;29(1):2304973. [doi: 10.1080/10872981.2024.2304973] [Medline: 38217884]
6. Miao Y, Wen J, Luo Y, Li J. MedARC: adaptive multi-agent refinement and collaboration for enhanced medical reasoning in large language models. Int J Med Inform 2026 Feb;206:106136. [doi: 10.1016/j.ijmedinf.2025.106136] [Medline: 41109093]
7. Almansour M, Soliman M, Aldekhyyel R, Binkheder S, Temsah MH, Malki KH. An academic viewpoint (2025) on the integration of generative artificial intelligence in medical education: transforming learning and practices. Cureus 2025 Mar 25;17(3):e81145. [doi: 10.7759/cureus.81145] [Medline: 40276436]
8. Javid M, Reddiboina M, Bhandari M. Emergence of artificial generative intelligence and its potential impact on urology. Can J Urol 2023 Aug;30(4):11588-11598. [Medline: 37633285]
9. Singhal K, Tu T, Gottweis J, et al. Toward expert-level medical question answering with large language models. Nat Med 2025 Mar;31(3):943-950. [doi: 10.1038/s41591-024-03423-7] [Medline: 39779926]
10. Thirunavukarasu AJ, Ting DSJ, Elangovan K, Gutierrez L, Tan TF, Ting DSW. Large language models in medicine. Nat Med 2023 Aug;29(8):1930-1940. [doi: 10.1038/s41591-023-02448-8] [Medline: 37460753]
11. Wen Y, Li X, Zeng F, Lei J, Chen S. Application of medical record quality control system based on artificial intelligence [Article in Chinese]. J Sichuan Univ (Med Sci) 2023 Nov 20;54(6):1263-1268. [doi: 10.12182/20231160206] [Medline: 38162053]
12. Mukadam A, Suresh S, Jacobs C. Beyond traditional simulation: an exploratory study on the effectiveness and acceptability of ChatGPT-4o advanced voice mode for communication skills practice among medical students. Cureus 2025 May 19;17(5):e84381. [doi: 10.7759/cureus.84381] [Medline: 40535400]
13. Karampinis E, Bozi Tzetzi DA, Pappa G, et al. Use of a large language model as a dermatology case narrator: exploring the dynamics of a chatbot as an educational tool in dermatology. JMIR Dermatol 2025 Sep 16;8:e72058. [doi: 10.2196/72058] [Medline: 40957071]
14. Germanà E, Pepe L, Pizzimenti C, et al. Programmed cell death ligand 1 (PD-L1) immunohistochemical expression in advanced urothelial bladder carcinoma: an updated review with clinical and pathological implications. Int J Mol Sci 2024 Jun 19;25(12):6750. [doi: 10.3390/ijms25126750] [Medline: 38928456]
15. Pepe P, Pepe L, Fiorentino V, Curduman M, Pennisi M, Fraggetta F. PSMA PET/CT accuracy in diagnosing prostate cancer nodes metastases. In Vivo 2024;38(6):2880-2885. [doi: 10.21873/invivo.13769] [Medline: 39477430]
16. Fiorentino V, Pepe L, Pizzimenti C, et al. PD-L1 expression in prostate cancer and Gleason Grade Group: is there any relationship? Findings from a multi-institutional cohort. Pathol Res Pract 2025 May;269:155916. [doi: 10.1016/j.prp.2025.155916] [Medline: 40107012]
17. Yang X, Chen A, PourNejatian N, et al. A large language model for electronic health records. NPJ Digit Med 2022 Dec 26;5(1):194. [doi: 10.1038/s41746-022-00742-2] [Medline: 36572766]
18. Yu H, Zhou J, Li L, et al. Simulated patient systems powered by large language model-based AI agents offer potential for transforming medical education. Commun Med (Lond) 2026;6(1):27. [doi: 10.1038/s43856-025-01283-x] [Medline: 41420084]
19. Bolton E, Venigalla A, Yasunaga M, et al. BioMedLM: a 2.7B parameter language model trained on biomedical text. arXiv. Preprint posted online on Mar 27, 2024. [doi: 10.48550/arXiv.2403.18421]
20. Hetz MJ, Carl N, Haggenmüller S, et al. Superhuman performance on urology board questions using an explainable language model enhanced with European Association of Urology guidelines. ESMO Real World Data Digit Oncol 2024 Dec;6:100078. [doi: 10.1016/j.esmorw.2024.100078] [Medline: 41646097]
21. Elhilali A, Ngo ASH, Reichenpfader D, Denecke K. Large language model-based patient simulation to foster communication skills in health care professionals: user-centered development and usability study. JMIR Med Educ 2025 Dec 12;11:e81271. [doi: 10.2196/81271] [Medline: 41385781]
22. Yi Y, Kim KJ. The feasibility of using generative artificial intelligence for history taking in virtual patients. BMC Res Notes 2025 Feb 24;18(1):80. [doi: 10.1186/s13104-025-07157-8] [Medline: 39994780]
23. Cook DA. Creating virtual patients using large language models: scalable, global, and low cost. Med Teach 2025 Jan;47(1):40-42. [doi: 10.1080/0142159X.2024.2376879] [Medline: 38992981]
24. Patino GA, Amiel JM, Brown M, Lypson ML, Chan TM. The promise and perils of artificial intelligence in health professions education practice and scholarship. Acad Med 2024 May 1;99(5):477-481. [doi: 10.1097/ACM.0000000000005636] [Medline: 38266214]
25. Yeung AWK, Tosevska A, Klager E, et al. Virtual and augmented reality applications in medicine: analysis of the scientific literature. J Med Internet Res 2021 Feb 10;23(2):e25499. [doi: 10.2196/25499] [Medline: 33565986]
26. Frankiewicz M, Vetterlein MW, Matuszewski M, Young Academic Urologists (YAU) Trauma and Reconstructive Urology Working Group. VR, reconstructive urology and the future of surgery education. Nat Rev Urol 2023 Jun;20(6):325-326. [doi: 10.1038/s41585-022-00722-x] [Medline: 36604520]

XSL•FO
RenderX

27. Tokas T, Ortner G, Peteinaris A, et al. Simulation training in transurethral resection/laser vaporization of the prostate; evidence from a systematic review by the European Section of Uro-Technology. World J Urol 2022 May;40(5):1091-1110. [doi: 10.1007/s00345-021-03886-0] [Medline: 34800135]
28. Malpani A, Vedula SS, Lin HC, Hager GD, Taylor RH. Effect of real-time virtual reality-based teaching cues on learning needle passing for robot-assisted minimally invasive surgery: a randomized controlled trial. Int J Comput Assist Radiol Surg 2020 Jul;15(7):1187-1194. [doi: 10.1007/s11548-020-02156-5] [Medline: 32385598]
29. Kallidonis P, Peteinaris A, Ortner G, et al. Transurethral resection of bladder tumor: a systematic review of simulator-based training courses and curricula. Asian J Urol 2024 Jan;11(1):1-9. [doi: 10.1016/j.ajur.2022.08.005] [Medline: 38312823]
30. Talyshinskii A, Hameed BZ, Zhanbyrbekuly U, et al. Role of virtual reality in improving the spatial perception of the kidney during flexible ureteroscopy: a feasibility study using virtual reality simulators and 3D models. Curr Urol 2024 Mar;18(1):24-29. [doi: 10.1097/CU9.0000000000000207] [Medline: 38505158]
31. Qi Z, Corr F, Grimm D, Nimsky C, Bopp MHA. Extended reality-based head-mounted displays for surgical education: a ten-year systematic review. Bioengineering (Basel) 2024 Jul 23;11(8):741. [doi: 10.3390/bioengineering11080741] [Medline: 39199699]
32. Patel V, Saikali S, Kavoussi L, et al. Best practices in telesurgery: framework and recommendations from the society of robotic surgery (SRS) for safe and effective implementation. J Robot Surg 2025 Jul 11;19(1):370. [doi: 10.1007/s11701-025-02523-x] [Medline: 40643832]
33. From Rome to Beijing - remote robotic surgery. Surgery International. 2024 Jun 26. URL: https://surgery.international/from-rome-to-beijing-remote-robotic-surgery/ [accessed 2026-10-09]
34. Anyinkeng AB, Girma SM, Maurice T, JohnPaul E, Hiwot T, Awad AK. The role of remote and virtual surgical training in expanding cardiothoracic surgical capacity in low-resource regions. BMC Surg 2025 Aug 25;25(1):393. [doi: 10.1186/s12893-025-03142-x] [Medline: 40855426]
35. Chen C, Wenli Z, Yue W, Xue S, Qidong L, Kun S. Research on the construction of a medical education metaverse based on cloudytization technique [Article in Chinese]. China Med Equip 2025;22(1):138-142. [doi: 10.3969/j.issn.1672-8270.2025.01.025]
36. Qin H, Lei J. The application of virtual reality technology in advertising communication. In: 2019 International Conference on Virtual Reality and Intelligent Systems (ICVRIS): IEEE:73-76. [doi: 10.1109/ICVRIS.2019.00027]
37. Qing L, Yong L. Design and path planning of the standard system of educational virtual digital human [Article in Chinese]. Mod Educ Technol 2024 Jul;34:70-80 [FREE Full text]
38. Vernez SL, Huynh V, Osann K, Okhunov Z, Landman J, Clayman RV. C-SATS: assessing surgical skills among urology residency applicants. J Endourol 2017 Apr;31(S1):S95-S100. [doi: 10.1089/end.2016.0569] [Medline: 27633332]
39. Obermeyer Z, Powers B, Vogeli C, Mullainathan S. Dissecting racial bias in an algorithm used to manage the health of populations. Science 2019 Oct 25;366(6464):447-453. [doi: 10.1126/science.aax2342] [Medline: 31649194]
40. Lee LH, Braud T, Zhou PY, et al. All one needs to know about metaverse: a complete survey on technological singularity, virtual ecosystem, and research agenda. Found Trends Hum-Comput Interact 2024 Nov 6;18(2-3):100-337. [doi: 10.1561/1100000095]
41. Gozum IEA. How can the principle of justice as fairness help address the digital divide in AI and healthcare? J Public Health (Oxf) 2025 Mar;47(1):e187-e188. [doi: 10.1093/pubmed/fdae163] [Medline: 39077847]

Abbreviations

LLM: large language model
XR: extended reality

Edited by TDA Cardoso; submitted 10.Feb.2026; peer-reviewed by E Karampinis, T Düzcü, V Fiorentino; revised version received 23.Sep.2026; accepted 23.Sep.2026; published 19.Oct.2026.

Please cite as:
Li P, Gong H, Liu H, Han S, Lu Y, Yan Y
Integrating Large Language Models and Metaverse-Related Technologies in Urology Education: Viewpoint
JMIR Med Educ 2026;12:e93235
URL: https://mededu.jmir.org/2026/1/e93235
doi:10.2196/93235

© Peizhe Li, Huijie Gong, Heyang Liu, Siyu Han, Yiqin Lu, Yongji Yan. Originally published in JMIR Medical Education (https://mededu.jmir.org), 19.Oct.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction

XSL•FO
RenderX

in any medium, provided the original work, first published in JMIR Medical Education, is properly cited. The complete bibliographic information, a link to the original publication on https://mededu.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

AI-Mediated Relational Competence in Medical Education: Construct Rationale and Curriculum Maturity Framework

Robert J McGrath, MS, PhD

Department of Health Management and Policy, College of Health and Human Services, University of New Hampshire, 4 Library Lane, Durham, NH, United States

Corresponding Author:
Robert J McGrath, MS, PhD
Department of Health Management and Policy, College of Health and Human Services, University of New Hampshire, 4 Library Lane, Durham, NH, United States

Abstract

AI is quickly becoming a foundational component of many clinical workflows. New tools are continually evolving to enable a host of functionalities, from ambient documentation to enhanced decision support, assistive diagnostic tools, risk prediction, patient triage, and patient communication. As a result, medical education has begun to address issues of AI literacy and the ethics of its use, including some work on competency frameworks, yet comparatively little attention has been given to the relational competencies clinicians require when AI increasingly mediates the patient encounter. Existing AI competency frameworks emphasize technical literacy, evaluation skills, and sometimes ethical awareness, yet the literature suggests that AI tools alter the structure of the patient-clinician relationship itself, raising questions about trust, transparency, empathic presence, and accountable human judgment. The empathy education literature has established that empathic communication is both teachable and assessable, and the AI competency literature has established that clinicians require new knowledge and skills for AI-enabled practice. These 2 bodies of work, however, remain insufficiently integrated. Medical education currently lacks a coherent competency framework for the relational work clinicians must perform when AI tools shape clinical information, communication, documentation, diagnosis, risk prediction, or treatment recommendations. This viewpoint proposes AI-mediated relational competence as a distinct construct for medical education, operationalized through 6 observable subcompetencies: empathic presence, trustworthiness, transparency in AI use, relational judgment, accountable human decision-making, and equity-aware trust-building. A complementary 5-level curriculum maturity model is introduced for evaluating whether medical schools move beyond curricular mention to systematic, competency-based implementation. Together, the construct and maturity framework are intended to provide the conceptual foundation for subsequent empirical work, including online curriculum assessment, national survey research, expert consensus validation, and competency-based assessment instrument development.

(JMIR Med Educ 2026;12:e105459) doi:10.2196/105459

KEYWORDS

medical education; empathy; trust; competency-based medical education; patient-clinician relationship; curriculum assessment; ambient AI scribe; shared decision-making; AI

Introduction

AI-Mediated Care Requires Relational Competence

AI is no longer a peripheral feature of clinical care. As of the US Food and Drug Administration's June 2026 update of its AI-Enabled Medical Device List, which covers marketing authorizations through March 2026, more than 1500 AI-enabled medical devices had been authorized for use in the United States [1]. Ambient AI scribes, generative language models, and predictive decision-support systems have rapidly entered routine practice, with other reviews documenting AI's accelerating integration across a breadth of clinical workflows [2-4]. A 2025 evidence synthesis of digital scribes confirmed that these tools have moved from pilot deployments into mainstream documentation workflows at many major health systems within a span of months rather than years [5]. Not surprisingly, the pace of AI integration has quickly outstripped the pace at which medical education curricula have managed to adapt to it.

Responses to AI to date have come primarily through the development of AI literacy and competency frameworks. Scoping reviews and stakeholder analyses have mapped the technical, evaluative, and ethical competencies clinicians require for AI-enabled practice [6,7]. One review identified core AI competency areas to include AI fundamentals, data analysis skills, ethical and legal considerations, and the ability to humanize patient care through effective communication so that the time AI returns to clinicians can be reinvested in compassionate engagement rather than absorbed by other tasks [8].

The pace of curricular uptake has been substantial. Among US medical schools granting Doctors of Medicine and Osteopathic

XSL•FO
RenderX

Medicine degrees that responded to the 2024 AAMC (Association of American Medical Colleges) and AACOM (American Association of Colleges of Osteopathic Medicine) Curriculum SCOPE (Structure, Content, Organization, Process, and Evaluation) Survey (182 of 208 schools, 88% response rate), the proportion incorporating AI into their curricula increased from 53% in 2023 to 77% in 2024 [9]. Recent scoping reviews of AI integration across US health professions curricula and of AI in undergraduate clinical-skills education show technical literacy, ethics, clinical applications, and tool use but do not isolate the relational competencies (empathic presence, patient trust, transparency, and accountable judgment) that AI-mediated care requires [10,11]. Additionally, this rapid integration is occurring in the absence of an established competency framework. It is worth noting that the AAMC's *AI Competencies Across the Learning Continuum* had, by mid-2026, reached a second public draft, with the final report targeted for fall 2026 [12]. Its publicly stated scope centers on foundational AI concepts, the ethical, legal, and professional implications of AI-enabled care, and the capacity to use, evaluate, and collaborate around AI tools across the learning continuum. These are necessary foundations. They are not, however, organized around the relational work of the AI-mediated encounter itself. These include the preservation of empathic presence, the building of patient trust, and the exercise of accountable human judgment when an algorithm is present at the point of care. The construct proposed here is not positioned as an alternative to the emerging national framework but as the relational complement it underweights.

A recurring theme across this literature is that AI competence is not reducible to AI literacy. Clinicians and patients alike express concerns that AI tools may dehumanize the clinical encounter or may threaten shared decision-making, thereby eroding the relational trust that underpins effective care [13-18]. In a widely cited 2018 *Journal of the American Medical Association* (*JAMA*) Viewpoint, Verghese et al [13] warned that AI must always be guided by physicians and that humanism and empathic care must remain the organizing center of clinical work and not be displaced by algorithmic outputs. Others have identified foundational issues such as consent challenges, the diffusion of responsibility to the AI, and underlying latent biases that could permeate algorithmic training data as ongoing barriers to the effective implementation of machine learning in health care [14], although, conceivably, these will continue to improve technologically over time. Others have called for a more refined and nuanced conceptual understanding of trust as it applies to AI in health care, noting that existing frameworks conflate technical reliability with relational trustworthiness [15]. One study of practitioners further noted that many physicians anticipate AI tools will disrupt the therapeutic relationship in ways that are both positive, by creating efficiencies that allow for relationship-building, and negative, by failing to capture the empathic and nonverbal dimensions central to health and mental health care [16].

The central educational question is, therefore, not only whether learners can effectively use AI tools but also whether they can use them in ways that remain empathic, transparent, trustworthy, and accountable.

Empathy Education Provides a Foundation but Not the Full Answer

Empathy is among the most well-established teachable domains in medical education. A landmark systematic review published in *Academic Medicine* established that empathy can be enhanced through targeted educational interventions such as communication skills training, standardized patient encounters, reflective writing, patient narratives, early clinical exposure, mindfulness practice, and humanities-based teaching more generally [19]. Subsequent systematic reviews have replicated and extended these findings, suggesting that empathy and compassion remain malleable across the undergraduate medical curriculum when interventions are sustained, multimodal, and assessed [20,21]. However, some studies have suggested that medical student empathy can erode during training because of emotional exhaustion and the dissonance between formal teaching and observed practice [22].

While important, much of the foundational empathy literature predates the widespread integration of generative AI in clinical systems. The standardized patient encounters and communication skills curricula on which empathy education has historically relied did not anticipate a clinical environment in which a third agent can both influence and even structure the encounter [13,16,23,24]. A recent patient survey of attitudes toward ambient AI scribes found that patients are generally open to the technology but emphasize that patient education and explicit consent are critical to maintaining trust during AI-mediated encounters [25]. A separate time-motion study showed measurable changes in clinician eye contact and consultation flow when ambient scribes were introduced, underscoring that the dynamics of the encounter itself shift when AI is present [23].

The new educational question is therefore not whether empathy can be taught, but how empathic presence is preserved and expressed during the encounter and how it is assessed when AI tools are in use. Existing empathy frameworks provide a useful foundation, but they require modification to account for the AI-mediated context.

Trustworthiness Is Distinct From Empathy

Trust is frequently treated as a downstream consequence of empathy in the clinical communication literature, but contemporary work on AI proliferation in health care suggests that trustworthiness ought to be treated as an analytically distinct educational construct. More recently, trust has been defined as a multidimensional concept encompassing competence, reliability, transparency, accountability, fairness, and respect for patient agency, none of which is reducible to empathic communication alone [15]. A second review examined how explainable AI features can both increase and decrease clinicians' trust, depending on how explanations are framed and whether they support or undermine clinician judgment [26]. London [27] has argued that opaque decisions are more common in medicine than critics often acknowledge and that the demand for algorithmic explainability must be balanced against the moral responsibilities clinicians hold for the decisions they make and explain to patients. Together, these contributions sharpen rather than resolve the educational challenge. Clinicians

XSL•FO
RenderX

must learn to communicate in a trustworthy manner even when the underlying algorithmic reasoning is somewhat or wholly opaque.

Patient-facing research reinforces this distinction. A recent *JAMA Network Open* study of US adults reported that overall trust in health systems to use AI responsibly is low, and that this distrust is unevenly distributed across demographic groups, signaling a need for focused organizational trustworthiness and improved communication in addition to empathic care [18]. Patient focus group research further found that patients view AI as a supportive tool but that patients prefer human oversight of automated tools as they are sensitive to the reliability of information delivered by AI and expect their clinician to remain an accountable interpreter of AI [17]. What is unclear, however, is whether these views will continue as AI increasingly pervades patients' everyday lives.

The clinical implication is that AI-mediated care introduces a new set of questions that empathic communication alone cannot resolve. Patients may reasonably wonder where care recommendations are coming from, be it the clinician, the AI, or perhaps a third-party vendor via algorithmic determinations. They may question how an algorithm trained on a nonrepresentative population applies to them. They also may ask what the clinician would have done in the absence of the AI [14,28-32]. Students therefore need explicit preparation in the trustworthiness and accountable use of AI-supported recommendations, skills that overlap with but are not identical to empathic communication.

Proposed Construct: AI-Mediated Relational Competence

AI-mediated relational competence is proposed here as the emerging clinician's ability to preserve empathy and patient trust, improve care transparency and shared decision-making, and exercise equity awareness and accountable human judgment when AI tools have the potential to influence all the elements of the care encounter and the resulting treatment recommendations. The construct integrates existing scholarship on empathy education [19-22], trust in health care AI [14,15,18,26-31], AI competencies for clinicians [6-8,14,33], clinical communication, and competency-based medical education [34-37].

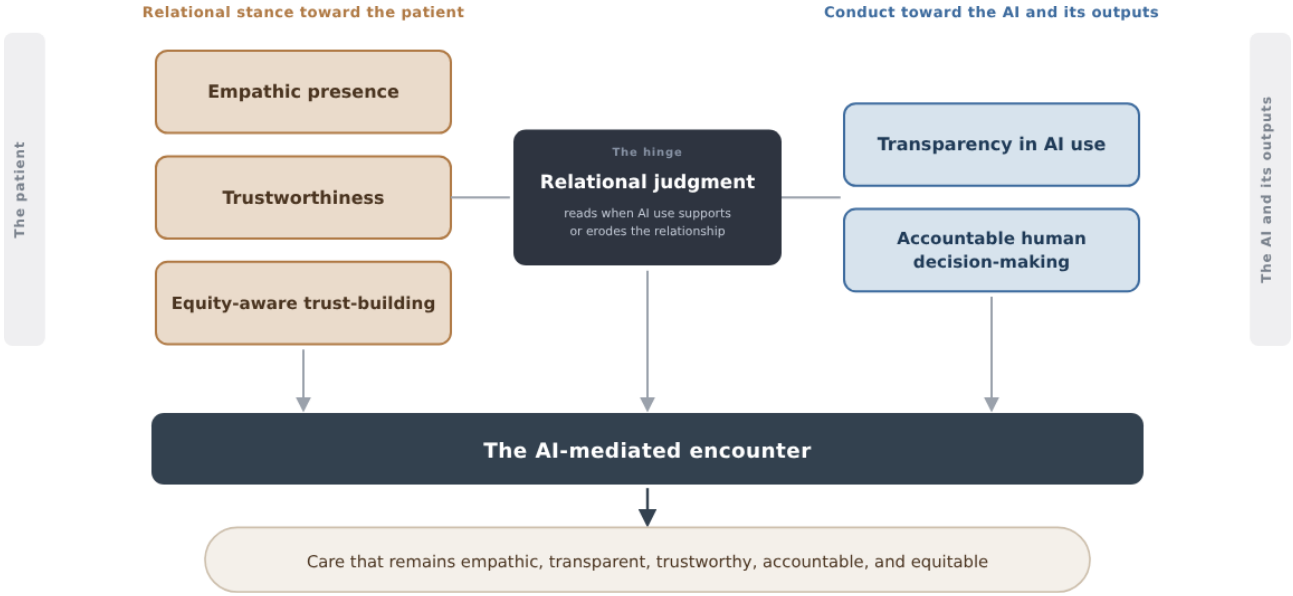
Three features distinguish this construct from existing AI literacy frameworks. First, it is *relational rather than technical*. Where AI literacy frameworks emphasize what the clinician must know about the algorithm, relational competence emphasizes what the clinician must do with that information on behalf of the patient [13,14]. Second, it is *assessable rather than aspirational*. The construct is operationalized through observable learner behaviors that can be taught, practiced, evaluated, and remediated within established competency-based assessments including objective structured clinical examinations (OSCEs), standardized patient encounters, direct observation, and/or portfolio assessment [35,37]. Third, it is *integrative rather than additive*. The construct is not proposed as a new soft skill bolted onto existing curricula. It is a clinical performance domain for technology-enabled practice that maps onto the 8 competency domains synthesized in the Physician Competency Reference Set, in particular, professionalism, interpersonal and communication skills, and patient care [36] domains, and it aligns with the AAMC competency framework currently in development [12].

A recent framework proposed a three-lever approach to concurrent reform of medical AI education, emphasizing that curricular, assessment, and regulatory mechanisms must move together [33]. The construct proposed here is intended to provide the conceptual content that such reform efforts can teach and assess.

The proposed 6 subcompetencies are related but not redundant, and the relationships among them are themselves part of the construct. They can be read along 2 dimensions (Figure 1). The first concerns the relational stance the clinician sustains with the patient, expressed through empathic presence, demonstrated trustworthiness, and equity-aware trust-building. The second concerns the clinician's conduct toward the AI and its outputs within the encounter, expressed through transparency in AI use, relational judgment, and accountable human decision-making. Relational judgment functions as the hinge between the 2: the capacity to read when the tool is supporting or eroding the relationship and to act accordingly. Because the construct describes relational work, some conceptual overlap among the subcompetencies is expected rather than a defect; the encounter is a single performance, and these are facets of it rather than independent modules.

XSL•FO
RenderX

Figure 1. AI-mediated relational competence: a conceptual model of 6 observable subcompetencies across 2 dimensions, bridged by relational judgment.



What justifies treating these facets as distinct assessable targets is that they dissociate in practice. A learner may disclose that AI informed the recommendation (transparency) while uncritically deferring to its output (a failure of accountable decision-making). A learner may communicate warmly and attentively (empathic presence) while overstating the authority of an algorithm (a failure of trustworthiness). Or a learner may explain a tool clearly to one patient while dismissing the well-founded skepticism of another (a failure of equity-aware trust-building). Because these behaviors can be present or absent independently of one another, each can be separately observed, taught, and remediated in a way that a single global rating of "relational skill" would obscure. The construct is therefore neither a relabeling of empathy nor a relabeling of AI ethics. Rather, it specifies the distinct and separately assessable behaviors through which a clinician keeps an AI-mediated encounter empathic, transparent, trustworthy, accountable, and equitable.

Construct Boundaries and Relationship to Existing Competencies

The relational dimensions of clinical care were well described long before AI entered the encounter. Competency frameworks have consistently named interpersonal and communication skills, professionalism, and patient care as distinct domains of physician performance [36]. Empathy education has established that the capacities underlying those domains are teachable and assessable [19-21], and the competency-based movement has supplied the assessment architecture through which they are evaluated [34,35]. What that literature has not required is a single organizing term. "Relational competence" is used here as an integrative label for capacities that existing frameworks distribute across several domains. It is proposed in this viewpoint rather than adopted from established usage, and the claim advanced is not that these capacities are new but that AI mediation changes the conditions under which they must be enacted and assessed.

The proposed construct is, therefore, best understood as a performance layer rather than a replacement for adjacent domains, each of which answers a different question. AI literacy concerns what the clinician knows about the tool: its inputs, limits, and failure modes. AI ethics concerns the normative questions the tool raises, including fairness, privacy, bias, and the just distribution of benefit and risk. Empathy education concerns the clinician's recognition of and response to the patient's experience. Professionalism concerns role fidelity and the standards of conduct to which the clinician is accountable. Calls for a data-informed health workforce sit alongside these as a further technical competency [38]. AI-mediated relational competence names none of these. It names the observable clinical behaviors through which all of them are enacted when AI participates in a specific encounter with a specific patient. Some conceptual overlap with its parent domains is expected, since the encounter is a single performance rather than a set of independent modules.

This raises a reasonable question about whether earlier clinical technologies required the same construct. Electronic health records, rule-based decision support, and telemedicine each altered the encounter, shifting clinician attention toward the screen and generating its own digital-health competency frameworks [39]. Some of these systems incorporated earlier forms of AI well before the recent widespread adoption of generative models. The distinction proposed here is therefore better understood as a matter of degree than of kind, turning on the extent of a system's participation in the encounter rather than its mere presence. Earlier systems, AI-enabled or not, were principal systems in which the clinician queried a record or a rule set and remained the author of the note and thereby the source of the words the patient heard. Contemporary generative systems are more integrative and proactive. They generate the clinical record rather than store it [40]; they deliver recommendations already formed and carrying implied authority [14,27]; and they can produce patient-facing language that reads as empathic. Each shift relocates something that previously belonged to the clinician. It is that relocation, rather than screen time or workflow disruption, that the construct is designed to

XSL•FO
RenderX

address. The technologies principally at issue here are therefore ambient documentation tools, generative language models, and predictive or diagnostic decision-support systems, rather than digital tools generally.

Observable Subcompetencies

Competency-based medical education requires observable behaviors rather than aspirational values [34,35]. Accordingly, this proposed construct is divided into 6 subcompetencies, each defined in terms of learner behaviors that can be assessed in clinical or simulated settings and evaluated against explicit performance criteria. Table 1 provides the full construct justification matrix. The subcompetencies are empathic presence, trustworthiness, transparency in AI use, relational judgment, accountable human decision-making, and equity-aware trust-building. Curricular maturity is treated separately as an institution-level dimension.

Table . Construct justification matrix: Subcompetency definition, anchoring, observable learner behaviors, and illustrative survey or curriculum-review questions.

Construct	Definition	Literature anchors	Observable learner behaviors	Survey/curriculum questions
Empathic presence	Capacity to recognize, communicate understanding of, and to respond to a patient's experience while using digital or AI tools	Empathy is teachable and assessable [19-21,41]; AI shifts the conditions of empathic expression [13,16,23]	Responds to emotion; maintains patient-centered attention; explains screen or AI use; and integrates patient narrative into decisions	Is empathy explicitly taught, practiced, and evaluated in AI-mediated scenarios?
Trustworthiness	Demonstrated capacity to act and communicate in ways that justify patient confidence	Trust is distinct from empathy and multidimensional [15,18,26,28]	Explains clinician role in interpreting AI; acknowledges uncertainty; avoids overstating AI authority; invites patient concern; respects patient agency in decisions; applies consistent standards across patients; and sustains reliability across encounters	Are trust-building and trustworthiness explicit learning objectives, competencies, or assessment domains?
Transparency in AI use	Ability to explain when and how AI informs care in patient-understandable language	AI competency frameworks emphasize safe, ethical use [6-8,14]; patients demand explicit disclosure [17,25]; and explainability remains contested [26,27]	Explains AI-supported recommendations; distinguishes AI input from clinician judgment; and communicates uncertainty and limitations	Do curricula teach disclosure, explanation, and communication of AI-supported decisions?
Relational judgment	Ability to judge when AI use supports or undermines patient-centered care	AI may create distance, confusion, mistrust, overreliance, or inequity [13,16,42]	Recognizes when AI may undermine trust or agency; adapts the encounter; and prioritizes patient concern over tool use	Do students practice AI-mediated cases involving concern, disagreement, bias, or uncertainty?
Accountable human decision-making	Ability to maintain professional responsibility for recommendations shaped by AI	AI may diffuse responsibility [14,26,30]; CBME[a] should preserve clinician accountability [34-36]	Verifies AI outputs; explains that AI informs but does not replace clinical responsibility; documents reasoning	Do curricula teach accountability for AI-supported recommendations?
Equity-aware trust-building	Ability to recognize how bias, inequity, prior harm, or historical mistrust may shape patient responses to AI-mediated care	Bias and fairness are central to AI ethics [2,14,42] and trust varies across demographic groups [18]	Names bias risks; adapts explanations to patient concerns; avoids dismissing mistrust; and supports shared decision-making	Does AI, trust, and empathy teaching include equity, bias, and vulnerability?
Curricular maturity	Institutional degree to which domains are defined, taught, reinforced, assessed, remediated, and tied to competency expectations	Curricular mention is not competency formation [35,39] and maturity models support empirical assessment	Institution-level construct and not individual learner behavior	Distinguish symbolic inclusion from systematic and competency-based implementation.

[a]CBME: competency-based medical education.

Empathic presence is the capacity to recognize, communicate understanding of, and respond to a patient's experience while concurrently using digital or AI tools. It extends established empathy education [19-21] into AI-mediated contexts in which attention and conversational rhythm are altered by the presence of ambient scribes and other screen-mediated workflows [13,16,23]. Validated empathy measurement instruments, such as the Jefferson Scale of Physician Empathy [41], provide the psychometric foundations for assessing empathic orientation in medical learners but were developed prior to widespread AI integration and rely on self-report rather than observed behavior,

XSL•FO
RenderX

thus offering a necessary but insufficient starting point for assessing empathic presence in AI-mediated encounters.

A further distinction is necessary as generative systems become capable of producing patient-facing language. Large language models can generate text that reads as warm, attentive, and responsive. Controlled work in medical education has shown that structured practice with role-prompted conversational agents can improve learners' perceived communication competence across emotionally demanding scenarios [43]. Empathic presence, however, is not reducible to the production of sympathetic phrasing. It requires attention to what is occurring in front of the clinician. This requires sensitivity to context that the model cannot observe and to the professional accountability that attaches to a person rather than a system. AI-generated language may support communication and, in some settings, may improve it, but the clinician remains responsible for interpreting the patient's situation, judging what the moment requires, and delivering care within a relationship for which someone can be held accountable. Assessment of empathic presence in AI-mediated encounters should therefore target the learner's attentiveness and judgment rather than the fluency of the language produced.

Trustworthiness is the demonstrated capacity to act and communicate in ways that justify patient confidence. It draws on the expert consensus work, establishing trust as a multidimensional construct distinct from empathy [15], and on patient research, demonstrating that transparency, accountability, and respect for patient agency are central to patient trust in AI-enabled care [17,18,25].

Transparency in AI use is the ability to explain when and how AI informs care in patient-understandable language. This subcompetency reflects both the patient's need for clear disclosure [17,25] and the clinician's challenge of explaining AI-supported recommendations where the underlying logic may be opaque [26,27]. Conceptual work on AI-supported shared decision-making argues that effective shared decision-making in AI-enabled care requires clinically meaningful reasoning beyond just technical transparency [44].

Relational judgment is the ability to determine when AI use supports or undermines patient-centered care. AI tools have been shown to create distance between clinicians and patients, encourage an overreliance on AI outputs, and potentially amplify inequity [13,16,42]. Clinical learners must develop the skills to recognize these effects when they are present and then adapt the encounter accordingly.

Accountable human decision-making is the ability to maintain professional responsibility for recommendations shaped by AI. The literature on AI in clinical decision support has frequently identified the risk of responsibility diffusion when algorithmic recommendations enter the workflow [14,26,30]. Competency-based medical education must preserve clinician accountability, requiring learners to verify AI outputs and then convey that AI informs but does not replace clinical judgment. They must then document their reasoning in a way that maintains the clinician as the accountable decision-maker [35,36].

Equity-aware trust-building is the ability to recognize how bias, inequity, prior harm, or historical mistrust may shape patient responses to AI-mediated care. Bias and fairness concerns are central to AI ethics [2,14,42], and these concerns further shape relational legitimacy. Patients from communities with prior experience of medical bias may be more skeptical of algorithmic recommendations and of being distanced from their clinicians, and learners must be prepared to acknowledge and engage with that skepticism rather than dismiss it. The previously cited *JAMA* survey also documented that trust in health systems' use of AI varies systematically across demographic groups, thus reinforcing the centrality of equity to relational trust-building [18].

Because trustworthiness, transparency in AI use, and accountable human decision-making all concern the clinician's conduct in relation to the tool, their assessment boundaries warrant statement. Transparency concerns disclosure and whether the patient is told that AI informed their care in terms they can understand and use. Accountable human decision-making concerns verification and ownership and whether the clinician checks the output and retains responsibility for the resulting recommendation. Trustworthiness is broader than either and is assessed at the level of the encounter as a whole, namely whether the clinician's conduct justifies the patient's confidence, including respect for patient agency, consistency of care across patients, reliability over time, and candor about the limits of what the tool and the clinician can know. The distinction is observable in practice. A learner may disclose AI use fully and still not be trustworthy, for instance, by disclosing in terms that discourage questions, and a learner may verify an output diligently while conveying a false impression of certainty. Table 2 sets out illustrative competent and concerning behaviors for each subcompetency to support assessment at these boundaries. As with the maturity model and assessment blueprint described here, these behavioral anchors are proposed rather than empirically validated and require confirmation before use in high-stakes assessment or institutional benchmarking.

XSL•FO
RenderX

Table . Competent and concerning behaviors for each subcompetency.

Subcompetency	Competent behavior	Concerning behavior
Empathic presence	Maintains attention to the patient while using the tool; responds to emotional cues; and explains AI use without disrupting rapport	Fixates on the screen or device; misses or overrides emotional cues; and allows documentation to set the pace of the encounter
Trustworthiness	Acknowledges uncertainty; avoids overstating AI authority; respects patient agency in decisions; applies consistent standards across patients; and sustains reliability across encounters	Presents AI output as definitive or unquestionable; dismisses patient preferences; and varies candor according to the patient
Transparency in AI use	Explains in patient-understandable language when and how AI informed care and distinguishes AI input from clinician judgment	Conceals AI involvement; relies on technical jargon that forecloses questions; and discloses in terms the patient cannot act on
Relational judgment	Recognizes when AI use is undermining comfort, agency, or disclosure; pauses or adapts tool use accordingly	Continues tool use despite visible patient discomfort and prioritizes workflow over the patient's evident concern
Accountable human decision-making	Verifies AI output against clinical findings; frames the recommendation as clinical judgment; and documents reasoning	Defers to the algorithm without examination; attributes the decision to the system; and cannot articulate a rationale independent of the tool
Equity-aware trust-building	Validates concerns about bias; explains known limitations and safeguards; and adapts the explanation to the patient's stated concern	Dismisses mistrust as irrational or uninformed; applies a uniform explanation regardless of concern; and treats skepticism as nonadherence

Three clinical vignettes illustrating the construct in practice, offered as proposed rather than empirically validated teaching examples, are provided in Multimedia Appendix 1: an ambient documentation tool that is running when a patient hesitates to disclose that cost has prevented them from taking a medication as prescribed; a patient who asks whether an AI-generated low-risk assessment reflects the tool's judgment or the clinician's; and a patient who asks whether a risk model was developed using data from patients like them. Each vignette identifies the subcompetencies it engages, contrasts a competent response with a plausible failure, and indicates where the scenario could be assessed.

Each subcompetency defined here can further be mapped to established teaching and assessment modalities [24,35,37]. Recent work has demonstrated the feasibility of AI-simulated patient interactions for training and assessing clinical interview skills [24], suggesting that AI-supported assessment infrastructure is actively developing in parallel with the educational challenges that AI itself creates.

Curricular Maturity

The key empirical question raised by this construct is not whether medical schools *mention* trust, empathy, or even AI in their curricular outlines. Nearly all do in some form, whether in mission language, course and syllabus catalogs, or in stand-alone AI literacy sessions [7]. The more consequential question is whether schools *systematically* teach, reinforce, and assess these competencies.

This distinction is well-recognized in competency-based medical education scholarship. The transition from time-based to outcomes-based training has highlighted persistent gaps between what curricula declare and what learners can demonstrably do [34,35,37]. A recent study on the design and implementation of a digital competence framework for health professionals emphasized that competence assessment must be tied to defined frameworks and tested in real professional contexts rather than inferred from curricular inclusion [39]. Table 3 presents a 5-level curriculum maturity model for AI-mediated relational competence, ranging from *implicit or absent* (level 1) through *introduced*, *integrated*, and *assessed* to *competency-based* (level 5), with the latter defined by linkage to feedback, remediation, progression, and entrustment decisions.

XSL•FO
RenderX

Table . Curriculum maturity levels for AI-mediated relational competence: proposed level, label, and definitions.

Level	Label	Description
1	Implicit or absent	Trust, empathy, and AI-mediated relational competence are assumed, modeled informally, or not visible in curriculum documentation.
2	Introduced	Topics are named in isolated sessions, learning objectives, or professionalism language, without longitudinal reinforcement.
3	Integrated	Topics are reinforced across multiple courses, clerkships, phases, or required learning experiences.
4	Assessed	Learners are evaluated using rubrics, OSCEs[a], standardized patients, direct observation, portfolios, or patient feedback.
5	Competency-based	Performance is linked to feedback, remediation, progression, graduation expectations, or entrustment decisions.

[a]OSCEs: Objective Structured Clinical Examinations.

The maturity model is proposed rather than validated. Its levels are derived from established distinctions in competency-based medical education between curricular mention, longitudinal reinforcement, formal assessment, and entrustment [34,35], but they have not yet been subjected to consensus or psychometric evaluation, and they should not be used for high-stakes curriculum evaluation or institutional comparison on the strength of this viewpoint alone. A validation program would proceed in stages: expert consensus review to confirm that the levels are exhaustive, mutually exclusive, and correctly ordered; cognitive interviewing with curriculum leaders to establish that the level descriptors are interpreted consistently across institutions; pilot coding of publicly available curriculum documents with interrater reliability testing to determine whether independent coders assign the same level to the same program; and systematic collection of validity evidence linking assigned levels to observable learner performance [37]. Until that evidence exists, the model is best used formatively as a shared vocabulary for institutional self-assessment and curricular planning rather than as a benchmarking instrument.

This framework is intended to serve 2 empirical purposes. First, it can support online curriculum assessment, in which publicly available curriculum materials are coded against the maturity levels to produce a national baseline picture of where US medical schools currently stand. Second, it can inform more systematic national survey research of medical and clinical programs. Comparing what schools publicly document with what curriculum leaders report is itself a useful empirical question.

Implications for Survey Development and Curriculum Assessment

A construct framework is a methodological precondition for both online curriculum assessment and national survey research. Without an explicit construct, coders cannot reliably distinguish symbolic from systematic inclusion, and survey items cannot be mapped to a coherent underlying competency model. The framework proposed here is designed to support both.

For online curriculum assessment, the construct provides a coding rubric that can be applied to publicly available materials and to the AAMC Curriculum Inventory. Each subcompetency from Table 1 generates a set of indicators (keywords, learning objectives, assessment modalities, and course types) that coders can search for and classify against the maturity levels of Table 3. This methodology mirrors published approaches to curricular content analysis in medical education.

For national survey development, the construct provides item domains corresponding to each subcompetency and maturity level. Items can ask not only whether a topic is taught but also whether it is reinforced longitudinally, assessed with explicit criteria, reflected upon for remediation, and linked to progression decisions.

Table 4 sets out an illustrative assessment blueprint, indicating how the subcompetencies might be distributed across settings and methods already available in most programs, from preclinical standardized patient encounters through workplace-based observation, simulation, and portfolio reflection. The blueprint is a worked example rather than a validated instrument. It has not been piloted; the mapping of subcompetencies to settings is one defensible arrangement among several; and the validation requirements stated for the maturity model apply to it equally. Its purpose is to show that the construct can be assessed using infrastructure that programs already operate, rather than to prescribe how any particular program should do so.

XSL•FO
RenderX

Table . Illustrative assessment blueprint for AI-mediated relational competence.

Setting and scenario	Subcompetencies assessed	Assessment method
Preclinical OSCE[a]: the learner introduces an ambient documentation tool at the start of a sensitive history.	Empathic presence; transparency in AI use; and relational judgment	Standardized patient checklist with faculty global rating
Clerkship workplace-based observation: the learner discusses an AI-generated risk estimate with a patient who questions it.	Trustworthiness and accountable human decision-making	Direct observation with narrative feedback
Simulation: an AI tool produces a biased or incorrect recommendation that the learner must recognize and address.	Relational judgment; equity-aware trust-building; and accountable human decision-making	Structured debrief with rubric-based rating
Portfolio: the learner reflects on an encounter in which a patient expressed mistrust of AI-supported care.	Equity-aware trust-building and trustworthiness	Reflection rubric with faculty feedback

[a]OSCE: Objective Structured Clinical Examinations.

The competent and concerning behaviors in Table 2 correspond to the 6 subcompetencies defined above and are intended to complement, rather than duplicate, the maturity levels in Table 3. Table 3 describes how deeply a curriculum addresses the construct, while Table 2 describes what competent and concerning performance looks like within any given encounter.

Table 4 builds on both preceding tables by mapping the same subcompetencies onto concrete assessment settings and methods already available in most training programs.

Faculty Development and the Hidden Curriculum

Curricular maturity describes what an institution teaches. It does not describe what learners observe. The values and behaviors transmitted through routine clinical practice, rather than through formal instruction, are recognized influences on professional formation and plausible contributors to the erosion of empathy documented during training [22]. AI-mediated relational competence is exposed to the same dynamic. A learner taught to disclose AI use, verify AI outputs, and address patients' concerns about algorithmic fairness will discount that teaching if supervising clinicians do not follow suit and reinforce these behaviors.

Faculty development is therefore a condition of the construct's viability rather than an implementation detail. Supervising clinicians require preparation to model the behaviors specified in Table 1 by disclosing AI use in patient-understandable terms, seeking patient agreement where documentation tools are in use, verifying AI outputs before acting on them, recognizing automation bias in their own reasoning, and engaging concerns about fairness and reliability with the seriousness afforded other clinical questions. Evidence from AI-generated educational documentation is instructive here. Mischaracterizations and hallucinations occurred at low but nontrivial rates and were largely corrected when faculty reviewed the output before submission [40]. The findings illustrate the potential value and feasibility of faculty verification in that setting. An institution could rate highly on formal curricular provision while its clinical environment models the opposite, and self-assessment against Table 3 should be interpreted with that limitation in view.

Conclusion

AI does not render trust and empathy obsolete in medical education; rather, it raises the importance of both as foundational human inputs to the care encounter. As AI tools enter routine clinical workflows, the relational skills clinicians perform with patients become more challenging and also more critical to assess. The existing AI competency literature has identified the technical, evaluative, and ethical skills clinicians need [6-8,14,33], and the existing empathy education literature has similarly established that relational skills can be taught and assessed [19-22,41]. What has been missing is a competency framework that bridges these 2 bodies of work and operationalizes the relational competencies required for AI-mediated practice.

The construct proposed here and the accompanying curriculum maturity model are offered as the bridge. The framework shifts the question from *whether* trust and empathy are mentioned in medical curricula to *whether* they are taught, practiced, assessed, and required as observable performance in AI-enabled care. The rationale for this construct is intended to provide a conceptual foundation for subsequent empirical work.

Generative AI Use

The author used a generative AI assistant (Claude Sonnet 5.5, Anthropic) to support drafting and editing of this manuscript, including prose arrangement, reference-verification support, and preparation of the conceptual figure. The author reviewed and verified all content including all citations, claims, and elements of the proposed framework. The author is solely accountable for the final manuscript.

XSL•FO
RenderX

Funding

No external funding supported this work.

Authors' Contributions

The author conceived the construct and the curriculum maturity framework, conducted the literature synthesis, and drafted, revised, and approved the final manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Clinical vignettes illustrating AI-mediated relational competence.
[DOCX File, 17 KB - mededu_v12i1e105459_app1.docx]

References

1. Artificial intelligence-enabled medical devices. US Food and Drug Administration. 2026. URL: https://www.fda.gov/medical-devices/software-medical-device-samd/artificial-intelligence-enabled-medical-devices [accessed 2026-10-07]
2. Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. Nat Med 2019 Jan;25(1):44-56. [doi: 10.1038/s41591-018-0300-7] [Medline: 30617339]
3. Rajkomar A, Dean J, Kohane I. Machine learning in medicine. N Engl J Med 2019 Apr 4;380(14):1347-1358. [doi: 10.1056/NEJMra1814259] [Medline: 30943338]
4. Wachter RM, Brynjolfsson E. Will generative artificial intelligence deliver on its promise in health care? JAMA 2024 Jan 2;331(1):65-69. [doi: 10.1001/jama.2023.25054] [Medline: 38032660]
5. Kanaparthy NS, Villuendas-Rey Y, Bakare T, et al. Real-world evidence synthesis of digital scribes using ambient listening and generative artificial intelligence for clinician documentation workflows: rapid review. JMIR AI 2025 Oct 10;4:e76743. [doi: 10.2196/76743] [Medline: 41071988]
6. Garvey KV, Thomas Craig KJ, Russell R, Novak LL, Moore D, Miller BM. Considering clinician competencies for the implementation of artificial intelligence–based tools in health care: findings from a scoping review. JMIR Med Inform 2022 Nov 16;10(11):e37478. [doi: 10.2196/37478] [Medline: 36318697]
7. Tolentino R, Baradaran A, Gore G, Pluye P, Abbasgholizadeh-Rahimi S. Curriculum frameworks and educational programs in AI for medical students, residents, and practicing physicians: scoping review. JMIR Med Educ 2024 Jul 18;10:e54793. [doi: 10.2196/54793] [Medline: 39023999]
8. Gazquez-Garcia J, Sánchez-Bocanegra CL, Sevillano JL. AI in the health sector: systematic review of key skills for future health professionals. JMIR Med Educ 2025 Feb 5;11:e58161. [doi: 10.2196/58161] [Medline: 39912237]
9. Artificial intelligence curricula in U.S. and Canadian medical schools. : Association of American Medical Colleges (AAMC); 2025 URL: https://www.aamc.org/media/84666/download [accessed 2026-10-07]
10. Taheri R, Foroughi M, Parang K. Artificial intelligence integration in U.S. healthcare professional degree programs: a rapid scoping review. Res Sq. Preprint posted online on Jun 19, 2026. [doi: 10.21203/rs.3.rs-9819693/v1]
11. Thind BS, Javidi D, Schwartz LM. Artificial intelligence in undergraduate medical education clinical skills curricula: a scoping review of implementations since 2022. Front Digit Health 2026;8:1830254. [doi: 10.3389/fdgth.2026.1830254] [Medline: 42359442]
12. Artificial intelligence competencies across the learning continuum. AAMC. URL: https://www.aamc.org/about-us/medical-education/ai-competencies [accessed 2026-10-07]
13. Verghese A, Shah NH, Harrington RA. What this computer needs is a physician: humanism and artificial intelligence. JAMA 2018 Jan 2;319(1):19-20. [doi: 10.1001/jama.2017.19198] [Medline: 29261830]
14. Char DS, Shah NH, Magnus D. Implementing machine learning in health care - addressing ethical challenges. N Engl J Med 2018 Mar 15;378(11):981-983. [doi: 10.1056/NEJMp1714229] [Medline: 29539284]
15. Starke G, Gille F, Termine A, et al. Finding consensus on trust in AI in health care: recommendations from a panel of international experts. J Med Internet Res 2025 Feb 19;27:e56306. [doi: 10.2196/56306] [Medline: 39969962]
16. Weir IB, Stroud AM, Stout JJ, et al. Physician perspectives on the impact of artificial intelligence on the therapeutic relationship in mental health care: qualitative study. JMIR Ment Health 2025 Dec 30;12:e81970. [doi: 10.2196/81970] [Medline: 41468531]
17. Foresman G, Biro J, Tran A, et al. Patient perspectives on artificial intelligence in health care: focus group study for diagnostic communication and tool implementation. J Particip Med 2025 Jul 24;17:e69564. [doi: 10.2196/69564] [Medline: 40705399]
18. Nong P, Platt J. Patients' trust in health systems to use artificial intelligence. JAMA Netw Open 2025 Feb 3;8(2):e2460628. [doi: 10.1001/jamanetworkopen.2024.60628] [Medline: 39951270]

XSL•FO
RenderX

19. Batt-Rawden SA, Chisolm MS, Anton B, Flickinger TE. Teaching empathy to medical students: an updated, systematic review. Acad Med 2013 Aug;88(8):1171-1177. [doi: 10.1097/ACM.0b013e318299f3e3] [Medline: 23807099]
20. Menezes P, Guraya SY, Guraya SS. A systematic review of educational interventions and their impact on empathy and compassion of undergraduate medical students. Front Med (Lausanne) 2021;8:758377. [doi: 10.3389/fmed.2021.758377] [Medline: 34820397]
21. Boshra M, Lee A, Kim I, Malek-Adamian E, Yau M, LaDonna KA. When patients teach students empathy: a systematic review of interventions for promoting medical student empathy. Can Med Educ J 2022 Nov;13(6):46-56. [doi: 10.36834/cmej.73058] [Medline: 36440084]
22. Howick J, Dudko M, Feng SN, et al. Why might medical student empathy change throughout medical school? A systematic review and thematic synthesis of qualitative studies. BMC Med Educ 2023 Apr 24;23(1):270. [doi: 10.1186/s12909-023-04165-9] [Medline: 37088814]
23. Tan JYE, Rafi IBM, Sng GGR, et al. Impact of an ambient AI scribe among clinicians and patients: real-world prospective observational time-motion study. JMIR Med Inform 2026 Mar 31;14:e85580. [doi: 10.2196/85580] [Medline: 41915701]
24. Yamamoto A, Koda M, Ogawa H, et al. Enhancing medical interview skills through AI-simulated patient interactions: nonrandomized controlled trial. JMIR Med Educ 2024 Sep 23;10:e58753. [doi: 10.2196/58753] [Medline: 39312284]
25. Leiserowitz G, Mansfield J, MacDonald S, Jost M. Patient attitudes toward ambient voice technology: preimplementation patient survey in an academic medical center. JMIR Med Inform 2025 Nov 27;13:e77901. [doi: 10.2196/77901] [Medline: 41308194]
26. Rosenbacke R, Melhus Å, McKee M, Stuckler D. How explainable artificial intelligence can increase or decrease clinicians' trust in AI applications in health care: systematic review. JMIR AI 2024 Oct 30;3:e53207. [doi: 10.2196/53207] [Medline: 39476365]
27. London AJ. Artificial intelligence and black-box medical decisions: accuracy versus explainability. Hastings Cent Rep 2019 Jan;49(1):15-21. [doi: 10.1002/hast.973] [Medline: 30790315]
28. Kerasidou A. Artificial intelligence and the ongoing need for empathy, compassion and trust in healthcare. Bull World Health Organ 2020 Apr 1;98(4):245-250. [doi: 10.2471/BLT.19.237198] [Medline: 32284647]
29. Nagy M, Sisk B. How will artificial intelligence affect patient-clinician relationships? AMA J Ethics 2020 May 1;22(5):E395-E400. [doi: 10.1001/amajethics.2020.395] [Medline: 32449655]
30. Sauerbrei A, Kerasidou A, Lucivero F, Hallowell N. The impact of artificial intelligence on the person-centred, doctor-patient relationship: some problems and solutions. BMC Med Inform Decis Mak 2023 Apr 20;23(1):73. [doi: 10.1186/s12911-023-02162-y] [Medline: 37081503]
31. Buhr E, Onder O, Rudra P, Ursin F. Trust and artificial intelligence in the doctor-patient relationship: epistemological preconditions and reliability gaps. Ethics Inf Technol 2025;27(4). [doi: 10.1007/s10676-025-09867-w]
32. Faria V, Goturi N, Dynak A, et al. Triadic relations in healthcare: surveying physicians' perspectives on generative AI integration and its role on empathy, the placebo effect and patient care. Front Psychol 2025;16:1612215. [doi: 10.3389/fpsyg.2025.1612215] [Medline: 41312268]
33. Kelly SL, Forgie M, Walker MC. Bridging the implementation gap in medical AI education: 3-lever framework for concurrent reform. JMIR Med Educ 2026 Apr 28;12:e88778. [doi: 10.2196/88778] [Medline: 42048663]
34. Frank JR, Snell LS, Cate OT, et al. Competency-based medical education: theory to practice. Med Teach 2010;32(8):638-645. [doi: 10.3109/0142159X.2010.501190] [Medline: 20662574]
35. Holmboe ES, Sherbino J, Long DM, Swing SR, Frank JR. The role of assessment in competency-based medical education. Med Teach 2010;32(8):676-682. [doi: 10.3109/0142159X.2010.500704] [Medline: 20662580]
36. Englander R, Cameron T, Ballard AJ, Dodge J, Bull J, Aschenbrener CA. Toward a common taxonomy of competency domains for the health professions and competencies for physicians. Acad Med 2013 Aug;88(8):1088-1094. [doi: 10.1097/ACM.0b013e31829a3b2b] [Medline: 23807109]
37. Cook DA, Beckman TJ. Current concepts in validity and reliability for psychometric instruments: theory and application. Am J Med 2006 Feb;119(2). [doi: 10.1016/j.amjmed.2005.10.036] [Medline: 16443422]
38. Doll J, Anzalone AJ, Clarke M, Cooper K, Polich A, Siedlik J. A call for a health data-informed workforce among clinicians. JMIR Med Educ 2024 Jun 17;10:e52290. [doi: 10.2196/52290] [Medline: 38889091]
39. Saigí-Rubió F, Romeu T, Hernández Encuentra E, Guitert M, Andrés E, Reixach E. Design, implementation, and analysis of an assessment and accreditation model to evaluate a digital competence framework for health professionals: mixed methods study. JMIR Med Educ 2024 Oct 17;10:e53462. [doi: 10.2196/53462] [Medline: 39418092]
40. Talwalkar JS, Chartash D, Zhang L, et al. Ambient AI scribes to create educational feedback notes for medical students: randomized trial. JMIR Med Educ 2026 May 28;12:e89996. [doi: 10.2196/89996] [Medline: 42208058]
41. Hojat M, Mangione S, Nasca TJ, et al. The Jefferson Scale of Physician Empathy: development and preliminary psychometric data. Educ Psychol Meas 2001 Apr;61(2):349-365. [doi: 10.1177/00131640121971158]
42. Hou J, Zhang Z, Cheng X, Wang W. Patient concerns regarding artificial intelligence applications in health care: systematic review and meta-synthesis based on social ecological theory. J Med Internet Res 2026 Apr 28;28:e85663. [doi: 10.2196/85663] [Medline: 42048645]

XSL•FO
RenderX

43. Müller Y, Lemberg H, Gellert P, Zoellick JC. Training empathetic communication skills in medical students with a role-prompted GPT-4o chatbot: quasi-experimental intervention study. JMIR Med Educ 2026 Jun 12;12:e88092. [doi: 10.2196/88092] [Medline: 42284570]
44. As'ad M, Faran N, Joharji H. AI-supported shared decision-making (AI-SDM): conceptual framework. JMIR AI 2025 Aug 7;4:e75866. [doi: 10.2196/75866] [Medline: 40773762]

Abbreviations

AACOM: American Association of Colleges of Osteopathic Medicine
AAMC: Association of American Medical Colleges
***JAMA*:** *Journal of the American Medical Association*
OSCE: objective structured clinical examination
SCOPE: Structure, Content, Organization, Process, and Evaluation

Edited by M Montagna; submitted 25.Jun.2026; peer-reviewed by M Ikuno, S Ghazi; revised version received 24.Sep.2026; accepted 25.Sep.2026; published 27.Oct.2026.

Please cite as:
McGrath RJ
AI-Mediated Relational Competence in Medical Education: Construct Rationale and Curriculum Maturity Framework
JMIR Med Educ 2026;12:e105459
URL: https://mededu.jmir.org/2026/1/e105459
doi:10.2196/105459

© Robert J McGrath. Originally published in JMIR Medical Education (https://mededu.jmir.org), 27.Oct.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Education, is properly cited. The complete bibliographic information, a link to the original publication on https://mededu.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Artificial Intelligence as a Core Public Health Competency: Proposal and Perspectives in Education and Research

Laura Brunelli[1,2], MD, SPH, PhD; Federico Fonda[1,2], RN, MSc; Silvio Brusaferro[1,2], MD, SPH, Prof Dr Med

[1]Department of Medicine, University of Udine, via Colugna 50, Udine, Friuli Venezia Giulia, Italy

[2]Accreditation, Quality and Clinical Risk Unit, Azienda Sanitaria Universitaria Friuli Centrale, Udine, Italy

Corresponding Author:
Laura Brunelli, MD, SPH, PhD
Department of Medicine, University of Udine, via Colugna 50, Udine, Friuli Venezia Giulia, Italy

Abstract

The continuous development of AI presents unprecedented opportunities for public health. This development prompts educators and researchers to consider how AI can be applied to and integrated into a rapidly evolving global landscape in which AI is reshaping health systems. In turn, this generates an urgent need for structured proposals that bridge theory and practice. In this viewpoint, we present perspectives and proposals on AI applications in public health. We operationalize AI literacy as a core public health competency, defined through knowledge, skills, and attitudes and scaled across professional roles. For education, we describe how AI can support learners, teachers, and organizations through personalized learning experiences, dynamic scenarios, immersive simulations, content preparation, and communication strategies. For research, we analyze how AI can be embedded as a transversal enabler supporting the World Health Organization's essential public health functions. We emphasize that AI should enhance, rather than replace, human capabilities and propose that AI literacy should be recognized as a core public health competency. We suggest that harnessing AI for public health education and research is, therefore, less a technological challenge than a collective responsibility shared by educators, researchers, and institutions.

(JMIR Med Educ 2026;12:e92246) doi:10.2196/92246

KEYWORDS

core competencies; public health; artificial intelligence; AI; education; research

Introduction

The continuous development of AI technologies, defined as tools capable of imitating certain functions of human intelligence [1], prompts educators and researchers to consider how AI can be applied and integrated into public health education and research. AI is already being embedded in the daily work of public health, from outbreak detection to the design of health promotion campaigns [2], and growing evidence and debate on the promise of AI highlight the unprecedented opportunities these technologies offer, while also revealing significant challenges. In this context, educators and researchers navigate a complex landscape and have the opportunity to use AI technologies thoughtfully and strategically, aiming to realize their full potential while acknowledging their inherent limitations in both public health education and research. Addressing the implications of AI in public health education and research arises from the rapidly evolving global landscape in which AI is fundamentally reshaping health systems across diverse countries, with growing international implications. A potential gap between technological advancement and educational and research readiness [2-5] creates an urgent need for structured proposals at the international level that bridge theory and practice. This requires reflecting on perspectives and implications for public health education and research to ensure that current and future public health professionals are equipped with the technical, ethical, and critical thinking skills required to participate in and navigate these changes. Our central argument is that the response to this gap is not simply more technology, but a competency; AI literacy should be formally recognized as a core public health competency that is defined, taught, governed, and evaluated with the same rigor applied to other core disciplines such as epidemiology or health systems science [3,6]. The aim of this viewpoint is therefore threefold: (1) to define AI literacy as a core public health competency in terms of the knowledge, skills, and attitudes it comprises; (2) to examine how this competency can be developed in education and applied in research across learners, teachers, and organizations; and (3) to provide an appropriate framework for embedding and governing this competency at scale.

Defining AI Literacy as a Core Public Health Competency

The concept of competency, following the competency-based education tradition, is defined as an observable ability that integrates knowledge, skills, attitudes, and values and can be developed and assessed along a continuum of expertise [7]. AI literacy is the ability to understand, use, monitor, and critically evaluate AI applications (including their limitations and ethical

XSL•FO
RenderX

implications) within a given professional context or scenario [3]. AI literacy is related to, but broader than, general digital literacy; digital literacy concerns the use of digital tools and information, whereas AI literacy adds the capacity to reason about systems that are probabilistic, data dependent, opaque, and capable of generating plausible but incorrect outputs.

Operationally, AI literacy as a public health competency comprises three interdependent domains: (1) knowledge, (2) skills, and (3) attitudes. The knowledge domain includes the fundamentals of AI and machine learning, data literacy, and an understanding of the regulatory and ethical frameworks that govern AI in health. The skills domain includes the practical ability to interact directly or indirectly with AI systems (eg, formulating and iteratively refining prompts), to integrate AI outputs into public health tasks, and to critically appraise those outputs for accuracy, bias, and relevance. The attitudes domain includes ethical and reflective judgment, which involves recognizing when not to use AI, retaining accountability for decisions, and sustaining professional skepticism toward automated outputs. These domains develop along a continuum, from foundational awareness expected of all professionals, through applied competence for those who routinely use AI in practice, and to advanced expertise for those who design, validate, or govern AI systems. In fact, we argue that this competency should be tailored to the specific role(s) in the public health workforce.

We therefore distinguish 3 cumulative tiers that mirror the tiered logic of established public health workforce competency frameworks (eg, the Council on Linkages Core Competencies for Public Health Professionals, which separates frontline, program management, and senior leadership responsibilities [8]). At the operational level (user level; eg, frontline and program support roles), every professional and trainee should be a competent, critical user of AI tools in day-to-day work (eg, querying data, drafting and appraising health communication materials, and recognizing when an output is unreliable or inappropriate). At the applied level (service and research level; eg, program management and supervisory roles), those who select, adapt, implement, or evaluate AI applications need deeper methodological and data literacy (model validity, algorithmic bias, data provenance, and the appropriateness of a tool for a given population) so that they can judge not merely how to use AI, but whether and when to adopt it. At the strategic level (macro and stewardship level; eg, senior management and executive leadership roles), public health leaders and policymakers who set strategy for networks, regions, or national systems need competencies in governance, procurement, ethics, equity, and regulation to allocate resources, safeguard accountability and transparency, and steer system-level adoption in the public interest. These tiers are cumulative rather than mutually exclusive, as strategic leaders remain users.

Defining the competency also clarifies how its acquisition should be evaluated: per role and per scenario, with programs that should align assessment with established educational models (eg, Kirkpatrick levels, which distinguish reaction, learning, behavior, and results [9]). As a result, the field can move from teaching about AI to demonstrating changes in public health practice and, ultimately, in population health outcomes.

Acknowledging AI literacy as a core public health competency carries several implications, for both medical education and research. We discuss these implications in the following sections.

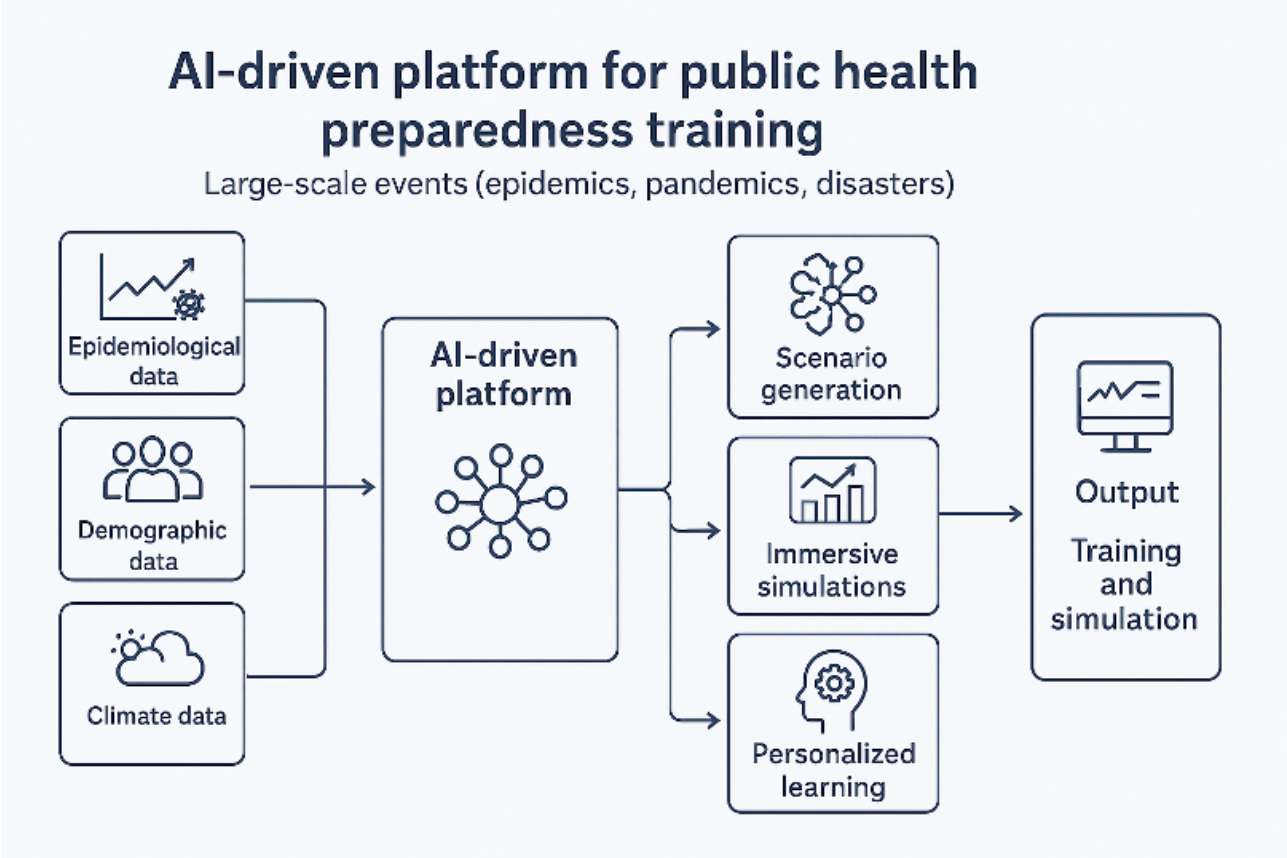
Implications for Education

According to the United Nations Educational, Scientific and Cultural Organization (UNESCO), possible applications of AI in education include solutions designed for learners, teachers, and organizations. These applications may also involve support for the management and delivery of education, learning and assessment processes, enhancement of teaching methods, and continuing education [10]. In public health, the concept of “learners” extends beyond students enrolled in traditional academic programs to include health care professionals engaged in continuing medical education, as well as patients and communities. With the contribution of AI, learners may benefit from a tailored educational approach rather than a one-size-fits-all model, be supported in identifying personal learning gaps and needs, and experience personalized assistance [8-11]. For example, in public health preparedness, tools aimed at learners include those capable of a wide range of functions [4] as shown in Figure 1, such as the following:

- Generation of realistic and dynamic scenarios (ie, data-driven simulations and adaptive models with varying levels of complexity based on the learner characteristics)
- Training through immersive simulations (ie, serious games and virtual reality, interaction with AI agents for rehearsal)
- Predictive analysis and evaluation of decisions made (ie, automatic feedback, impact prediction supporting trial-and-error learning in a safe context)
- Personalization of the learning experience (ie, adaptive training based on the user’s level of knowledge and the preferences and needs of learners regarding the educational materials used)
- Analysis of individual and group performance (ie, simulation with multiple stakeholder perspectives and interaction)
- Support for interdisciplinary collaboration (ie, multiactor scenarios, supply chain, and logistics simulation)

XSL•FO
RenderX

Figure 1. AI-driven platform for public health preparedness training: an example. Source: Image generated with OpenAI ChatGPT (GPT-5) and DALL·E, 2025.



Throughout this section we distinguish, in line with the competency defined above, between applications supported by established evidence, those with early or limited evidence, and those that remain speculative, since conflating the 3 types of applications risks overstating the maturity of the field. Adaptive, personalized learning and feedback are currently the best-evidenced applications; immersive simulations and virtual reality rehearsal are promising but supported mainly by early, short-term evidence, and predictive "impact scoring" of learner decisions remains largely speculative. Figure 1 provides an illustrative example of such integration. It depicts an AI-driven platform for public health preparedness training in which epidemiological, demographic, and environmental data inform the generation of dynamic outbreak scenarios, the delivery of immersive simulations, and the personalization of individual learning pathways, with learner performance subsequently feeding back into further training. In doing so, the figure operationalizes the operational (user) tier of the competency within a realistic preparedness setting for large-scale events and illustrates how otherwise discrete learner-facing functions can be coordinated into a single, coherent training system rather than deployed in isolation. For example, a concrete learning goal at this tier is that every public health trainee should be able to use a generative AI tool to draft and then critically correct a piece of health communication content, identifying errors and biases.

Similarly to learners, the concept of "teachers" in public health extends beyond traditional academic faculties to include public health practitioners and professionals involved in primary prevention, health promotion, and community-based education. In this context, AI can support teachers and public health educators across all settings, from academia to primary prevention and health promotion. For example, AI could assist with information selection and visual representation when preparing teaching materials, help to better understand learners' learning processes and outcomes to inform decisions on educational strategies and interventions, perform monotonous tasks such as marking exams, and provide simple feedback about lesson and exam schedules [12,13]. AI can also help teachers and organizations effectively communicate health promotion and prevention content, adapting it to the abilities and characteristics of the target population, and considering linguistic [5], cultural, literacy, and health literacy differences [14], as well as any disabilities [4]. Following an optimistic vision of AI implementation for public health [15], when used for good, AI could also help counteract infodemics (defined by the World Health Organization [WHO] as an overabundance of information, including false or misleading information, during a health emergency). AI-based screening methods could be used to check online big data for fake news, with machine learning or artificial neural networks searching social media for recurrent words, phrases, images, or other problematic patterns and flagging content automatically for human review, rather than replacing editorial and public health judgment [16].

Organizations have a transversal role, bridging learners and teachers and supporting public health education at multiple levels. They can facilitate and provide access to AI technologies for both groups, promoting effective teaching and learning.

XSL•FO
RenderX

Organizations may also benefit from AI in administrative tasks, such as enrollment forecasting, resource allocation optimization, and automation of school operations [4]. Moreover, organizations are called to act as stewards of responsible AI use, establishing policies, monitoring implementation, and ensuring that AI applications align with ethical and legal standards, thereby safeguarding collective educational processes and public health delivery. In practice, this stewardship role means issuing acceptable use and transparency guidance for teachers and students, investing in faculty development so that teachers themselves reach applied competence, defining assessment safeguards that preserve academic integrity, creating governance structures with clear accountability, and addressing inequities in access to tools and infrastructure.

Implications for Research

In public health research, the systematic use of big data, the Internet of Things, and Industry 4.0 products is already becoming part of everyday life for researchers, with AI applications being developed in an increasingly wide range of areas (eg, development of new drugs and therapies, identification of biomarkers for population screening, identification and prediction of individual and community risk factors, etc). AI software already enables studies that can personalize responses by integrating knowledge derived from population data (real or simulated), the exposome, and "omic" sciences [15,17,18]. Moreover, AI can support researchers in project management and data acquisition, such as using augmented reality tools (eg, portable devices [19]), social listening, and data mining to find patterns, trends, correlations, anomalies, and features in datasets [16] or on social media [2]. This can aid behavioral epidemiology and deep learning for the One Health approach. AI support extends through to data analysis (eg, augmented reality tools, satellite image analysis for environmental hygiene, medicine in conflict zones or natural disasters), and the identification, implementation, and scientific validation of possible responses.

We suggest that a structural priority for public health research is to assess how, and to what extent, AI can be integrated into public health systems in relation to its contribution to the essential public health functions (EPHFs) of the WHO [6]. Framing AI as an enabling tool for the EPHFs provides a system-level analytical framework, allowing researchers to link technological innovation to surveillance, preparedness, prevention, health promotion, governance, workforce development, research, and equity. When strategically implemented and empirically evaluated, AI has the potential to support several, if not all, of the 12 EPHFs. While AI is not a standalone solution, it can act as a transversal enabler that strengthens public health capacities across the entire policy and service continuum (Table 1). In fact, the EPHFs are the set of fundamental activities (spanning surveillance, emergency management, stewardship, financing, health protection, disease prevention, health promotion, community engagement, workforce development, quality and equity, research, and access to health products [6]) that every health system must perform to protect and improve population health. The WHO defined them precisely to give countries a common, globally legitimate reference for assessing and strengthening public health capacity. Because they are already used for workforce planning and are explicitly cross-sectoral, they allow innovation to be linked to system-level needs rather than to isolated tools. Framing AI as a transversal enabler of the EPHFs therefore gives researchers an analytical lens for prioritizing and evaluating AI against population health functions rather than technological novelty. Moreover, because one of these functions is workforce development (EPHF 9), the same map that shows where AI can strengthen surveillance or prevention also identifies the competencies the workforce must acquire to deliver those functions responsibly, tying the research agenda directly to education.

XSL•FO
RenderX

Table . Essential public health functions (EPHFs): examples of AI applications and main associated challenges.

Essential public health function	Examples of AI application	Main associated challenges[a]
EPHF 1: Public health surveillance and monitoring	Support for data surveillance and monitoring, including natural language processing for interpreting medical records and social media/web searches/contents monitoring to detect early outbreak signals (social listening)	Privacy and data protection
EPHF 2: Public health emergency and management	Generation of dynamic interpretative scenarios to support public health emergency management decisions, predictive analytics	Overreliance on automated predictions
EPHF 3: Public health stewardship	Automation of administrative operations, support complex decision-making	Unclear accountability and responsibility for decisions
EPHF 4: Multisectoral planning, financing, and management for public health	Optimization of allocation of financial and human resources for resource forecasting	Algorithmic bias in resource allocation
EPHF 5: Health protection	Environmental hygiene, forecasting environmental risks	Data quality and representativeness
EPHF 6: Disease prevention and early detection	Accelerate the identification and prediction of individual and community risk factors	Algorithmic bias and risk of stigmatization
EPHF 7: Health promotion	Chatbots to provide real-time information (eg, healthy lifestyles, symptoms, and vaccines) and personalization of health messages, helping to contrast infodemics with fake news screening	Accuracy of automated outputs and misinformation
EPHF 8: Community engagement and social participation	Facilitation for understanding complex medical language, customization of messages and adaptation for communities for greater inclusivity	Digital exclusion and inequitable access
EPHF 9: Public health workforce development	Personalized training for professionals, assisted identification of individual learning gaps, immersive simulations and scenarios	De-skilling and overreliance on automated systems
EPHF 10: Health service quality and equity	Assisted technologies to ensure quality and equity to all populations, targeting individual needs	Equity gaps and bias
EPHF 11: Public health research, evaluation, and knowledge	Support in data acquisition and complex analysis, accelerate the development and implementation of quality evaluation projects	Transparency and reproducibility (“black box”)
EPHF 12: Access and utilization of health products, supplies, equipment, and technologies	Assisted utilization of complex technological infrastructures, assisted supply chain and forecasting of population needs	Data interoperability and accountability

[a]The challenges listed are illustrative and not exhaustive; most are cross-cutting and apply to varying degrees, across all functions. Notable cross-cutting challenges include privacy and data protection, algorithmic bias and discrimination/stigmatization, lack of transparency and explainability (black box), unclear accountability, overreliance on automated systems, digital exclusion and inequitable access, misuse or misinterpretation of predictive outputs, and the overarching need for ethical governance.

Issues and Ethical Implications

We identify several unresolved issues. In research, it is not yet clear to what extent AI can, or should, be credited as a coauthor of publications [20], or act as a semiautonomous investigator (eg, in systematic reviews [21]) especially when population data are involved. Beyond authorship, the main challenges include the overall protection of data (security, quality, accuracy, availability, and privacy); the environmental sustainability of systematic AI use, given its energy and carbon footprint [22-24]; the social and professional impact of these technologies (possible inequalities arising from uneven access to and use of digital tools [14,25]), the dehumanization and intrusiveness of AI-mediated teaching and learning [10], and the de-skilling of professionals, especially in the early stages of their careers [26]; and the potential for malicious use through misinformation and disinformation. In addition to these, there are risks inherent in the technology itself: AI is prone to hallucination and inaccuracy (returning plausible but false information), is not free from algorithmic bias (arising from unrepresentative data or design choices), and often behaves as a “black box” whose reasoning cannot be fully traced. A subtler issue is the information asymmetry emerging between generations, which can invert the usual direction of expertise; handled constructively, it becomes an opportunity for teachers and more experienced colleagues to manage uncertainty together with younger ones [26].

These concerns could become clearer, and easier to act on, when organized into four categories: (1) governance, (2) equity, (3) pedagogical integrity, and (4) professional practice, whose weight differs across the research and education domains. In research, governance is chiefly about authorship, autonomous

XSL•FO
RenderX

review agents, and the lawful use of data; in education, it is about acceptable use policies, assessment integrity, and transparency regarding when and how AI has been used. This governance should be anchored in the growing body of responsible AI guidance, from the WHO's work on the ethics and governance of AI for health to academic frameworks such as that of Char and colleagues [27] on implementing machine learning in health care. Equity, in turn, is sharpened by well-documented algorithmic bias; previous research has shown, for example, that a widely used clinical algorithm underestimated the needs of certain patient groups because it used cost as a proxy for illness [28], a risk to both research validity and the fairness of any service or educational tool built on biased data. Pedagogical integrity (dehumanization, overreliance, de-skilling) is chiefly an education domain concern, whereas professional practice risks, such as misinformation and AI's energy and carbon footprint, span both. Seen this way, the information asymmetry noted above is not merely a difficulty to be endured but part of the competency: managing it transparently and ethically is a professional skill to be taught.

Conclusions and Future Directions

AI application in both education and research in public health can follow two strategic approaches, as reported by Abdulnour and colleagues [26]: (1) the centaur approach, in which the user strategically divides tasks between themselves and the AI, allocating responsibilities according to the strengths and capabilities of each; or (2) the cyborg approach, where the user works with AI iteratively by prompting, correcting, and requesting justifications, then refining the output jointly with the AI. Regardless of the chosen approach, as stated in the UNESCO guidelines for policymakers, we emphasize that AI should be used to enhance, rather than replace, human capabilities, shifting the paradigm from AI to augmented intelligence [10], where human skills are empowered by the specific capabilities provided by up-to-date technologies. All these developments and implementations must consider the regulatory framework surrounding AI at both international and national levels. WHO documents have outlined the main principles to be followed (ie, "Ethics and Governance of Artificial Intelligence for Health" [29]) and have further analyzed large language models [30] and benchmarking among AI systems supporting health care [31]. For example, the same discussion has taken place at the European level, with the release of the AI Act [32], and at the Italian national level as well [33]. In addition, guidelines for the evaluation of studies conducted with the aid of AI have been published [34]. The final goal should not be to replace public health professionals, teachers, or researchers, but to provide an additional tool for professionals who are already trained and capable of identifying the problem, formulating the research or teaching question, categorizing the data, developing the algorithm, deciding how the pieces of the puzzle fit together, and drawing conclusions based on the data [10]. In fact, according to the Moravec paradox, even if AI may be better at pattern discovery and statistical reasoning, it is poor at self-directed learning, applying common sense, and making value judgments [35]. This principle of augmentation has a precedent worth naming. Friedman's [36] "fundamental theorem" of biomedical informatics holds that a person working in partnership with an information resource should perform better than that person unaided. We propose an analogous fundamental principle for AI in public health: an AI system is valuable only insofar as a competent public health professional supported by it achieves better population health outcomes than the same professional working without it.

The pace and shape of responsible AI governance are not uniform across the world, and the reasons are structural rather than incidental. Where health systems are largely publicly funded and health is treated as a public good, the surrounding incentives tend to favor precaution, equity, and shared standards, and governance often seeks to guide innovation from the outset. Where the development and deployment of AI are driven more strongly by market competition, governance tends to evolve alongside (and at times behind) innovation, reflecting different priorities such as speed, scalability, and user choice. Keeping in mind that public health is inherently public good, each orientation carries its own risks, from overcaution and delayed adoption on one side to fragmentation and uneven safeguards on the other. The challenge is to find the best possible balance between values and constraints.

We therefore suggest several priority actions in public health. Educators should define AI literacy learning outcomes for each learner group and assess them beyond satisfaction, using established models, and embed critical appraisal exercises that make the limitations of AI explicit. Institutions and organizations should issue acceptable use and transparency guidance, invest in faculty development, protect assessment integrity, and address inequities in access. Researchers should prioritize studies that evaluate AI's real-world contribution to specific EPHFs, with explicit attention to bias and equity, and develop the governance and reporting standards needed for AI's evolving role in the research process itself. Most importantly, we believe that it is important to provide cross-disciplinary, continuous, and systematic AI training for students, postgraduates, health care professionals, and researchers to improve individual levels of AI literacy [26,37], embedding it as a core public health competency, comparable to epidemiological methods or health systems knowledge [6]. Such training should include fundamentals of AI and machine learning, data literacy, prompt engineering and interaction with generative AI, limitations and risks of AI, practical applications in public health, ethics, transparency, governance, practical skills, and tools. It is also essential to promote and develop research in various fields of application, test its impact and real effectiveness in practice, and govern its implementation in traditional health care sectors including health promotion, disease protection, and prevention, with a view to personalization for individuals and communities. At the end, we suggest that harnessing AI for public health education and research is, therefore, less a technological challenge than a collective responsibility shared by educators, researchers, and institutions. In the near future, the question will no longer be whether AI will shape public health practice, but how to engage with it critically, ethically, and competently. Ultimately, the value of AI in public health will be determined

XSL•FO
RenderX

not by the advancements and sophistication of algorithms, but by the wisdom with which humans choose to deploy them.

Acknowledgments

Figure 1 was created using the generative AI tool DALL-E by OpenAI ChatGPT (GPT-5) and DALL·E, 2025. Responsibility for the final manuscript lies entirely with the authors. Generative AI tools are not listed as authors and do not bear responsibility for the final outcomes.

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Conflicts of Interest

None declared.

References

1. World Commission on the Ethics of the Scientific Knowledge and Technology. Preliminary study on the ethics of artificial intelligence. : UNESCO; 2019 Feb 26 URL: https://unesdoc.unesco.org/ark:/48223/pf0000367823 [accessed 2026-09-21]
2. Panteli D, Adib K, Buttigieg S, et al. Artificial intelligence in public health: promises, challenges, and an agenda for policy makers and public health institutions. Lancet Public Health 2025 May;10(5):e428-e432. [doi: 10.1016/S2468-2667(25)00036-2] [Medline: 40031938]
3. Semi MMA, Das S, Utsho MR, et al. Artificial intelligence in public health education: a scoping review of workforce competency development. Health Sci Rep 2026 Mar 9;9(3):e72066. [doi: 10.1002/hsr2.72066] [Medline: 41816647]
4. Wang J, Li J. Artificial intelligence empowering public health education: prospects and challenges. Front Public Health 2024 Jul 3;12:1389026. [doi: 10.3389/fpubh.2024.1389026] [Medline: 39022411]
5. Naumova EN. Artificial intelligence and data analytics competencies for public health professionals. J Public Health Policy 2024 Sep;45(3):407-412. [doi: 10.1057/s41271-024-00499-w] [Medline: 38926467]
6. Defining essential public health functions and services to strengthen national workforce capacity. : World Health Organization; 2024 May 25 URL: https://www.who.int/publications/i/item/9789240091436 [accessed 2026-09-21]
7. Frank JR, Snell LS, Cate OT, et al. Competency-based medical education: theory to practice. Med Teach 2010;32(8):638-645. [doi: 10.3109/0142159X.2010.501190] [Medline: 20662574]
8. The Council on Linkages Between Academia and Public Health Practice. Core competencies for public health professionals. PHF. 2021 Oct. URL: https://phf.org/programs/core-competencies-for-public-health-professionals/ [accessed 2026-09-21]
9. Yardley S, Dornan T. Kirkpatrick's levels and education "evidence". Med Educ 2012 Jan;46(1):97-106. [doi: 10.1111/j.1365-2923.2011.04076.x] [Medline: 22150201]
10. Miao F, Holmes W, Huang R, Zhang H. AI and Education: Guidance for Policy-Makers: UNESCO; 2021. [doi: 10.54675/PCSP7350]
11. Nagi F, Salih R, Alzubaidi M, et al. Applications of artificial intelligence (AI) in medical education: a scoping review. Stud Health Technol Inform 2023 Jun 29;305:648-651. [doi: 10.3233/SHTI230581] [Medline: 37387115]
12. Chan KS, Zary N. Applications and challenges of implementing artificial intelligence in medical education: integrative review. JMIR Med Educ 2019 Jun 15;5(1):e13930. [doi: 10.2196/13930] [Medline: 31199295]
13. Pizzolla I, Aro R, Duez P, De Lièvre B, Briganti G. Integrating artificial intelligence into medical education: lessons learned from a Belgian initiative. J Interact Learn Res 2023;34(2):401-424. [doi: 10.70725/265312mqoebd]
14. Uddin J, Feng C, Xu J. Health communication on the internet: promoting public health and exploring disparities in the generative AI era. J Med Internet Res 2025 Mar 6;27:e66032. [doi: 10.2196/66032] [Medline: 40053755]
15. Venturini P, Faria PL, Cordeiro JV. AI and omics technologies in biobanking: applications and challenges for public health. Public Health 2025 Jun;243:105726. [doi: 10.1016/j.puhe.2025.105726] [Medline: 40315692]
16. Benke K, Benke G. Artificial intelligence and big data in public health. Int J Environ Res Public Health 2018 Dec 10;15(12):2796. [doi: 10.3390/ijerph15122796] [Medline: 30544648]
17. Argentieri MA, Amin N, Nevado-Holgado AJ, et al. Integrating the environmental and genetic architectures of aging and mortality. Nat Med 2025 Mar;31(3):1016-1025. [doi: 10.1038/s41591-024-03483-9] [Medline: 39972219]
18. Noorbakhsh-Sabet N, Zand R, Zhang Y, Abedi V. Artificial intelligence transforms the future of health care. Am J Med 2019 Jul;132(7):795-801. [doi: 10.1016/j.amjmed.2019.01.017] [Medline: 30710543]
19. Experience the power of assistive technology with OrCam's AI devices. OrCam. 2025. URL: https://www.orcam.com/en-us/home [accessed 2026-09-21]
20. Jungwirth D, Haluza D. Artificial intelligence and public health: an exploratory study. Int J Environ Res Public Health 2023 Mar 3;20(5):4541. [doi: 10.3390/ijerph20054541] [Medline: 36901550]

XSL•FO
RenderX

21. İlgen B, Pilic A, Harder T, Hattab G. Pre-training to identify immunization-related entities from systematic reviews. In: NLPIR '23: Proceedings of the 2023 7th International Conference on Natural Language Processing and Information Retrieval: Association for Computing Machinery; 2024:234-239. [doi: 10.1145/3639233.3639355]
22. Chen S. How much energy will AI really consume? the good, the bad and the unknown. Nature 2025 Mar;639(8053):22-24. [doi: 10.1038/d41586-025-00616-z] [Medline: 40044891]
23. Ding Z, Wang J, Song Y, et al. Tracking the carbon footprint of global generative artificial intelligence. Innovation (Camb) 2025 May 5;6(5):100866. [doi: 10.1016/j.xinn.2025.100866] [Medline: 40432776]
24. Wolf RM, Abramoff MD, Channa R, Tava C, Clarida W, Lehmann HP. Potential reduction in healthcare carbon footprint by autonomous artificial intelligence. NPJ Digit Med 2022 May 12;5(1):62. [doi: 10.1038/s41746-022-00605-w] [Medline: 35551275]
25. Miao F, Holmes W. Guidance for Generative AI in Education and Research: UNESCO; 2023. [doi: 10.54675/EWZM9535]
26. Abdulnour REE, Gin B, Boscardin CK. Educational strategies for clinical supervision of artificial intelligence use. N Engl J Med 2025 Aug 21;393(8):786-797. [doi: 10.1056/NEJMra2503232] [Medline: 40834302]
27. Char DS, Shah NH, Magnus D. Implementing machine learning in health care - addressing ethical challenges. N Engl J Med 2018 Mar 15;378(11):981-983. [doi: 10.1056/NEJMp1714229] [Medline: 29539284]
28. Obermeyer Z, Powers B, Vogeli C, Mullainathan S. Dissecting racial bias in an algorithm used to manage the health of populations. Science 2019 Oct 25;366(6464):447-453. [doi: 10.1126/science.aax2342] [Medline: 31649194]
29. Ethics and governance of artificial intelligence for health: WHO guidance. : World Health Organization; 2021 Jun 28 URL: https://www.who.int/publications/i/item/9789240029200 [accessed 2026-09-21]
30. Ethics and governance of artificial intelligence for health: guidance on large multi-modal models. : World Health Organization; 2025 Mar 25 URL: https://www.who.int/publications/i/item/9789240084759 [accessed 2026-09-21]
31. Shaping ethics, regulation and standardization in AI for health - ITU-WHO Focus Group on AI for Health. : International Telecommunication Union; 2025 URL: http://handle.itu.int/11.1002/pub/82693bf5-en [accessed 2026-09-21]
32. European Union. Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence and amending Regulations (EC) no 300/2008, (EU) no 167/2013, (EU) no 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (artificial intelligence act) (text with EEA relevance). EUR-Lex. 2024 Jul 12. URL: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32024R1689 [accessed 2026-09-21]
33. Italian Republic. Disposizioni e deleghe al Governo in materia di intelligenza artificiale. (25G00143) (GU Serie Generale n.223 del 25-09-2025) [Article in Italian]. Gazzetta Ufficiale. 2025. URL: https://www.gazzettaufficiale.it/eli/id/2025/09/25/25G00143/sg
34. Rajpurkar P, Chen E, Banerjee O, Topol EJ. AI in health and medicine. Nat Med 2022 Jan;28(1):31-38. [doi: 10.1038/s41591-021-01614-0] [Medline: 35058619]
35. Arora A. Moravec's paradox and the fear of job automation in health care. Lancet 2023 Jul 15;402(10397):180-181. [doi: 10.1016/S0140-6736(23)01129-7] [Medline: 37453744]
36. Friedman CP. A "fundamental theorem" of biomedical informatics. J Am Med Inform Assoc 2009 Mar;16(2):169-170. [doi: 10.1197/jamia.M3092] [Medline: 19074294]
37. Wang X, He X, Wei J, Liu J, Li Y, Liu X. Application of artificial intelligence to the public health education. Front Public Health 2023 Jan 10;10:1087174. [doi: 10.3389/fpubh.2022.1087174] [Medline: 36703852]

Abbreviations

EPHF: essential public health function
UNESCO: United Nations Educational, Scientific and Cultural Organization
WHO: World Health Organization

Edited by B Lesselroth; submitted 27.Jan.2026; peer-reviewed by D Dinh, D Winterton, W Lu; revised version received 03.Jul.2026; accepted 12.Aug.2026; published 30.Sep.2026.

Please cite as:
Brunelli L, Fonda F, Brusaferro S
Artificial Intelligence as a Core Public Health Competency: Proposal and Perspectives in Education and Research
JMIR Med Educ 2026;12:e92246
URL: https://mededu.jmir.org/2026/1/e92246
doi:10.2196/92246

XSL•FO
RenderX

© Laura Brunelli, Federico Fonda, Silvio Brusaferro. Originally published in JMIR Medical Education (https://mededu.jmir.org), 30.Sep.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Education, is properly cited. The complete bibliographic information, a link to the original publication on https://mededu.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

From Anatomy to Innovation: Embedding Patient-Centered Design in Medical Education

Louis John Koizia[1], MBBS; Pieter Vandekerckhove[2], PhD; Steven Howard[3], PhD; Bettina Maisch[4], PhD; Benjamin Howell Lole Harris[1,5], PhD

[1]Cutrale Perioperative and Ageing Group, Imperial College London, Sir Michael Uren Hub, 86 Wood Lane, White City Campus, London, United Kingdom

[2]Delft Centre for Entrepreneurship, Delft University of Technology, Delft, The Netherlands

[3]Department of Health Services Administration, School of Health Professions, University of Alabama at Birmingham, Birmingham, AL, United States

[4]Department of Applied Sciences and Mechatronics, Munich University of Applied Sciences, Munich, Germany

[5]Department of Oncology, University of Oxford, Oxford, United Kingdom

Corresponding Author:
Louis John Koizia, MBBS
Cutrale Perioperative and Ageing Group, Imperial College London, Sir Michael Uren Hub, 86 Wood Lane, White City Campus, London, United Kingdom

Abstract

Medical education has traditionally focused on mastering biomedical knowledge, clinical skills, and professionalism. Yet it remains poorly equipped to prepare doctors for the complex, rapidly changing health systems they will work in. Despite repeated calls for transformative learning, progress in teaching system innovation has been limited. Current curricula produce clinicians adept at treating disease but often underprepared to address system failures or contribute to organizational change. Embedding patient-centered innovation within training may help address this gap. This approach integrates design thinking, value-based health care, and co-design with patients to cultivate the metacompetence required to engage with system change. Emerging examples suggest that innovation education is most effective when integrated within existing activities such as quality improvement initiatives, multidisciplinary collaboration, and protected time for improvement work. Teaching these principles early may help normalize innovation as an integral component of clinical practice rather than a peripheral activity. Lessons from the successful integration of communication skills and ethics into medical curricula demonstrate that such cultural shifts are possible. Practical approaches may include incorporating design-thinking exercises, patient co-creation activities, and innovation projects alongside existing quality improvement and leadership programs. Reframing innovation as part of professional responsibility may empower clinicians to act as catalysts for system improvement. From anatomy to innovation, medical education must continue to evolve, preparing doctors not only to treat patients but also to engage with the systems that shape health care delivery.

(JMIR Med Educ 2026;12:e86230) doi:10.2196/86230

KEYWORDS

educational innovation; value-based health care; reform; curriculum; transformative learning

Introduction

Medical education has traditionally focused on equipping students with a strong foundation in biological sciences, clinical skills, and professional values essential for safe and compassionate care [1]. Curricula emphasize anatomy, physiology, pathology, and pharmacology, alongside integrative competencies such as communication, ethics, and teamwork. Within medical education scholarship, these competencies are commonly conceptualized through competency-based medical education frameworks and theories of experiential learning, which emphasize the progressive development of knowledge, skills, and professional identity through practice and reflection. However, a persistent mismatch remains between the competencies taught and the evolving needs of health systems. In response, many programs have introduced transformative learning elements, including leadership and digital health. Yet, from the student perspective, curricula still lack sufficient emphasis on digital health and practical experience with emerging technologies [1-4]. Despite improvements, the gaps identified by Frenk et al [1] in 2010 persist. Future clinicians are not adequately prepared to address systemic challenges such as aging populations, multimorbidity, digital disruption, and the demand for value-based care.

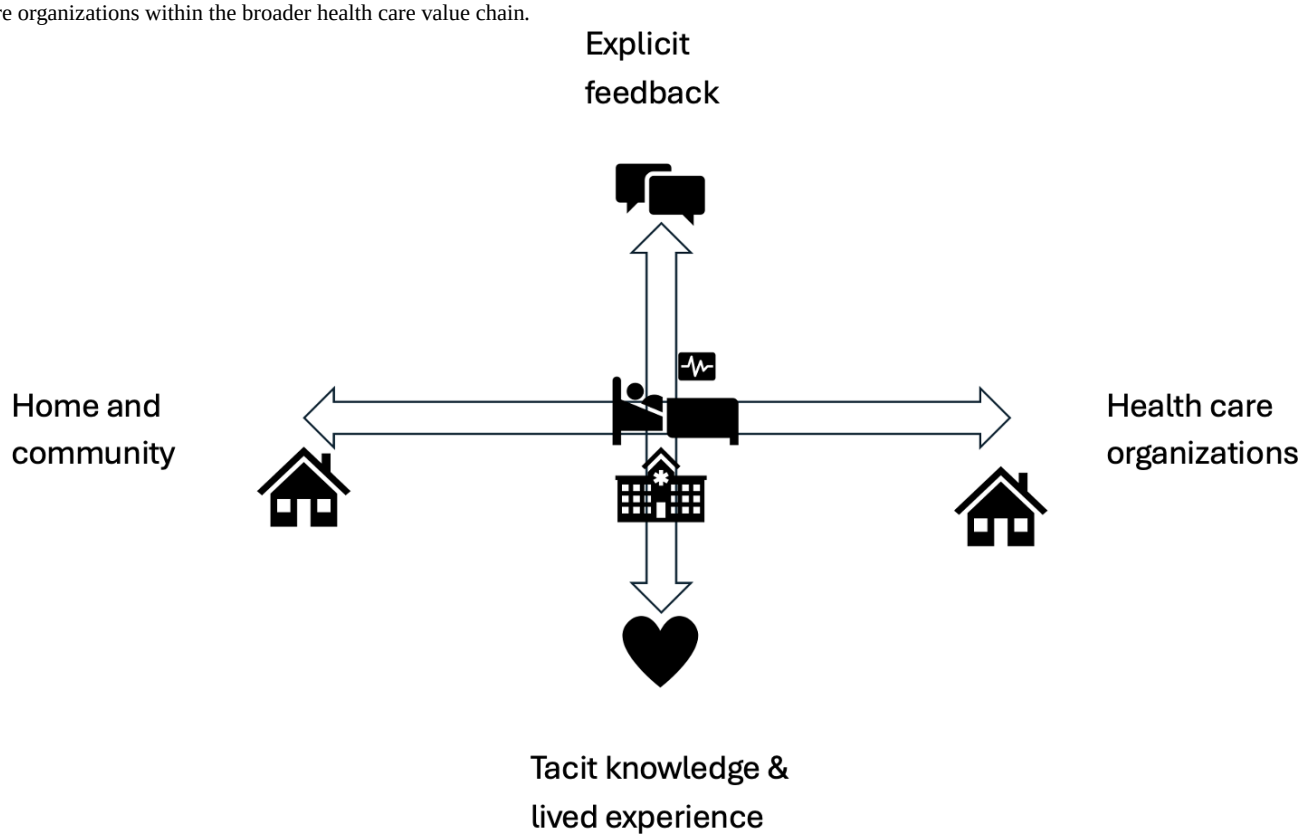
Graduates are trained to diagnose and treat disease but often lack the skills to identify and address system-level failures. Without the skills to innovate, clinicians are sidelined in system transformation efforts, which are typically led by policymakers and managers. We argue for a renewed focus on competencies

XSL•FO
RenderX

that empower clinicians to drive health system innovation and enhance patient-centered care.

In this article, we propose a conceptual framework for integrating patient-centered innovation into medical education. The framework distinguishes between 2 complementary dimensions of system improvement: vertical integration, which focuses on deep engagement with stakeholder experience and tacit knowledge, and horizontal integration, which focuses on positioning innovations across the broader patient journey and health care value chain. By articulating these dimensions, the framework aims to provide a conceptual structure for understanding how innovation competencies may be integrated into medical education and applied within complex health care systems (Figure 1).

Figure 1. Conceptual framework for patient-centered innovation in medical education. The vertical dimension represents the depth of stakeholder understanding, ranging from explicit feedback to deeper emotional and tacit knowledge uncovered through participatory design approaches. The horizontal dimension represents the integration of innovations across the patient journey, linking home environments, community services, and health care organizations within the broader health care value chain.



The Systemic Gap in Medical Training

Many medical schools remain anchored to a 20th-century model centered on biomedical knowledge, clinical competence, and professional ethics. While these pillars are essential, they are insufficient in today's complex health care landscape. Scholars in health professions education increasingly highlight the need for clinicians to develop capabilities that extend beyond individual patient encounters toward system awareness, interdisciplinary collaboration, and adaptive problem solving. Importantly, not all medical students aspire to traditional clinical roles. A recent Finnish survey revealed that approximately one-third of medical students consider product development a career path [5].

Health care systems face workforce shortages, widening health inequities, and rapid technological change. To navigate this disruption, students need metacompetence, defined as the ability to develop new competencies in response to evolving challenges and to adapt knowledge and skills across changing professional contexts [5,6]. This includes understanding how to lead change across organizational and technological boundaries and how to collaborate across disciplines. Students must understand the competencies required to implement change successfully and how innovations become embedded within complex interactions between people, organizations, and technologies.

This metacompetence requires the ability to apply skills from entrepreneurship, innovation, leadership, technology, and human-centered design. While some programs have begun to introduce these areas, they often lack integration [7]. Globally, health entrepreneurship and innovation education focused on addressing complex system challenges remain underdeveloped and universities struggle to embed it into medical curricula [8].

Conceptual Framework for Patient-Centered Innovation in Medical Education

To improve care at a systemic level, metacompetence must be cultivated. The conceptual framework proposed in this paper (Figure 1) illustrates how innovation competencies can be understood through 2 interacting dimensions: depth of stakeholder understanding (vertical integration) and alignment

XSL•FO
RenderX

across health care pathways and organizations (horizontal integration).

This perspective aligns with conceptualizations of health care as a complex adaptive system, in which outcomes emerge from interactions between individuals, organizations, and technologies rather than from isolated interventions. Within such systems, clinicians require capabilities that extend beyond individual clinical decision-making toward understanding relationships, feedback loops and system dynamics. The proposed framework therefore situates innovation competencies within the broader context of system-aware clinical practice.

Patient-centered innovation requires skills in design thinking, stakeholder co-creation, and value-based health care [9]. In this article, patient-centered innovation refers to the development and implementation of health care solutions that are co-designed with stakeholders and aligned with patient needs, experiences, and outcomes across the care pathway. These intersecting disciplines must be introduced in ways that highlight their relevance to clinical practice. Medical students should therefore learn to communicate and collaborate with engineers, managers, financial experts, and data scientists.

Educational theory provides useful insights into how such competencies may be developed. Experiential and design-based learning approaches emphasize iterative cycles of problem identification, experimentation, and reflection, making them particularly suited to innovation education within complex health care environments.

Increasingly, health care innovation is driven by advances in genomics, data science, and precision medicine, which are reshaping how diseases are understood and managed. Studies linking genetic variants to key disease phenotypes illustrate the growing importance of integrating large-scale biological data into clinical reasoning and innovation in care delivery [10-12].

This also demands a shift in mindset toward entrepreneurship, experimentation, and collaboration. These developments raise an important question for medical education: whether and how innovation-related competencies should be incorporated into clinical training.

Why Innovation Belongs in the Curriculum

Innovation in health care is frequently equated with technology. However, true patient-centered innovation aligns system outcomes with what matters most to patients. Several frameworks can guide this integration:

- Design thinking, which is an iterative approach to problem-solving [13]
- Value-based health care, which focuses on patient value rather than service volume [14]
- Experience-based and generative co-design, which positions diverse stakeholders, including care professionals and patients as active partners in service development [9]

Building on these perspectives, the proposed framework emphasizes 2 complementary dimensions: vertical integration of stakeholder insight and horizontal integration across the health care value chain (Figure 1) [15].

Clinicians as Catalysts for Change

Clinicians are not merely stakeholders in innovation, they are potential catalysts. They shape patient experiences, influence adoption patterns, and guide resource use. Clinicians in leadership roles act as gatekeepers, yet face barriers such as time constraints, hierarchical cultures, misaligned incentives, and fear of failure.

Importantly, clinicians can play a transformative role in design-led innovation, particularly through participatory design and generative co-design approaches [16]. These methods build on design thinking principles and emphasize deep collaboration with stakeholders, including patients, carers, and frontline staff. Participatory design involves users in the early stages of problem definition and solution development, while generative co-design enables stakeholders to articulate latent needs and envision future possibilities through creative exploration.

Understanding patient perspectives is essential in such processes. Research exploring public perceptions demonstrates how attitudes toward emerging health technologies vary across populations [17] and highlights the importance of engaging diverse communities in the development of health care innovations.

Although clinicians are often involved in such projects, they are rarely educated about these methodologies during their training. This lack of exposure limits their ability to lead or fully engage in innovation processes. Embedding education on participatory and generative co-design into medical curricula would empower clinicians to become proactive agents of change, capable of driving systemic improvements from within.

Unless curricula explicitly prepare students to recognize opportunities and navigate these challenges, innovation will remain peripheral. Embedding innovation within quality improvement, patient co-design, and multidisciplinary collaboration is essential. In addition, ring-fencing protected time for engagement can normalize innovation as a core clinical responsibility.

Lessons From History and Practice

Medical education has successfully integrated new domains before. Communication, once considered peripheral, is now embedded across curricula and assessed in exams. Assessment methods have also evolved, with structured question formats increasingly used to evaluate the application of clinical knowledge in modern medical training [18]. Similarly, medical ethics, professionalism, and interprofessional teamwork have moved from the margins to the mainstream [19]. These shifts reflect a growing recognition that excellence in medicine requires more than scientific knowledge.

However, medical education must also be prepared to change gears, even when the broader health system evolves slowly. Curricula must equip learners not only to adapt to innovation but to lead it. Increasingly, programs are incorporating health and technology perspectives. For example, Finland has made notable progress in digital health education, as illustrated by a

XSL•FO
RenderX

recent study tracking medical student competencies and attitudes between 2016 and 2022 [5,19]. In the Netherlands, Erasmus Medical Centre's *Arts 2030* initiative focuses more on interprofessional skills and the integration of technology.

Despite these advances, attention to patient-centered innovation within medical education remains limited. While some US programs have introduced entrepreneurship and innovation into medical training, these efforts often fall short of teaching students how to integrate technology in a human-centered way that truly benefits patients across horizontal and vertical dimensions.

To truly prepare future clinicians, medical education must go beyond technical literacy and foster the ability to co-create solutions with patients, embed value-driven principles, and lead systemic change.

Practical Steps for Integration

Embedding patient-centered innovation need not displace core sciences but should complement them. Rather than introducing entirely new curricular content, innovation competencies may be integrated within existing educational activities such as quality improvement projects, interdisciplinary learning, and experiential problem-solving exercises. For example, students could apply design thinking principles within existing quality improvement modules, use the horizontal and vertical pathway framework when analyzing clinical cases, or examine patient care pathways as part of health care delivery teaching. Clinical placements could also incorporate reflective exercises in which students identify system-level barriers to care and propose potential improvements.

On a policy level, greater attention to patient-centered innovation by medical colleges, professional associations, and accreditation bodies could facilitate the integration of these competencies into training programs. Practical approaches may include:

- Integration of design thinking and value-based care principles into existing case-based teaching
- Patient co-creation exercises incorporated into curriculum design or service improvement projects
- Innovation projects aligned with research or quality improvement initiatives addressing real health care system challenges
- Postgraduate reinforcement through leadership training, quality improvement programs, and specialist education

Global perspectives should also be considered when implementing innovation education. Health care systems differ widely in resources, workforce structures, and educational priorities. In low- and middle-income countries, resource constraints may limit access to digital technologies or dedicated innovation programs, but locally driven problem-solving and frugal innovation may offer valuable educational opportunities. Cultural contexts may also influence how clinicians, patients, and communities engage in co-design processes. Educational approaches therefore need to remain adaptable, allowing institutions to integrate innovation competencies in ways that reflect local health care delivery models, cultural expectations, and available resources. Implementation will also require attention to practical considerations such as faculty development, alignment with accreditation standards, and the availability of institutional resources. Recognizing these factors may help support integration within existing curricula and mitigate resistance from traditional medical education stakeholders.

Importantly, the aim of such training is not to transform all physicians into entrepreneurs or technology developers. Rather, it is to cultivate innovation literacy: the ability to understand system challenges, collaborate across disciplines, and contribute meaningfully to the development and implementation of patient-centered solutions.

While empirical evidence linking innovation education directly to patient outcomes remains limited, emerging examples suggest potential benefits. Several institutions have introduced innovation and design-oriented programs within medical education, including hackathons, interdisciplinary design courses, and health innovation tracks, which have been associated with increased student engagement in quality improvement, digital health, and health system innovation activities. However, robust evaluation frameworks are still developing. Future research should focus on measurable outcomes such as students' ability to analyze care pathways, collaborate across disciplines, and implement system improvements. Longitudinal studies following graduates exposed to innovation-oriented curricula, as well as comparative studies between traditional and innovation-enhanced training programs, will be important to better understand their impact on health care delivery and patient outcomes.

Conclusion

Health systems face an uncertain future. Aging populations and chronic disease are intensifying pressures. Clinicians must be equipped not only to treat disease but also to redesign systems of care. This article proposes a conceptual framework for patient-centered innovation education and highlights practical ways such competencies may be integrated into existing medical curricula through quality improvement activities, interdisciplinary learning, and clinical reflection. Embedding these competencies into curricula will not turn every doctor into an entrepreneur, but it may produce clinicians who are better equipped to engage with system improvement and multidisciplinary innovation.

From anatomy to innovation, medical education must evolve. Integrating patient-centered design and value-driven care into undergraduate and postgraduate training may represent an important step toward preparing clinicians to engage with system-level challenges in modern health care.

Funding

BHLH would like to thank the Orthogeriatric Research Fund for its support of this paper.

XSL•FO
RenderX

Conflicts of Interest

None declared.

References

1. Frenk J, Chen L, Bhutta ZA, et al. Health professionals for a new century: transforming education to strengthen health systems in an interdependent world. Lancet 2010 Dec 4;376(9756):1923-1958. [doi: 10.1016/S0140-6736(10)61854-5] [Medline: 21112623]
2. Das S, Ahmed N, Al Rahbi I, et al. Medical and biomedical students' perspective on digital health and its integration in medical curricula: recent and future views. Int J Environ Res Public Health 2025 Jul 30;22(8):1193. [doi: 10.3390/ijerph22081193] [Medline: 40869779]
3. Gillissen A, Kochanek T, Zupanic M, Ehlers J. Medical students' perceptions towards digitization and artificial intelligence: a mixed-methods study. Healthcare (Basel) 2022 Apr 13;10(4):723. [doi: 10.3390/healthcare10040723] [Medline: 35455898]
4. Utukuri M, D'souza F, Deighton A, et al. Digital health: a neglected part of health curricula? Future Healthc J 2022 Mar;9(1):18-20. [doi: 10.7861/fhj.2021-0102] [Medline: 35372764]
5. Veikkolainen P, Tuovinen T, Kulmala P, et al. The evolution of medical student competencies and attitudes in digital health between 2016 and 2022: comparative cross-sectional study. JMIR Med Educ 2025 Jul 31;11:e67423. [doi: 10.2196/67423] [Medline: 40743520]
6. Asakura K, Occhiuto K, Tarshis S, Dubrowski A. Designing and conducting healthcare simulations: contributions from social work. Cureus 2021 Jul;13(7):e16193. [doi: 10.7759/cureus.16193] [Medline: 34367797]
7. Niccum BA, Sarker A, Wolf SJ, Trowbridge MJ. Innovation and entrepreneurship programs in US medical education: a landscape review and thematic analysis. Med Educ Online 2017;22(1):1360722. [doi: 10.1080/10872981.2017.1360722] [Medline: 28789602]
8. Preiksaitis C, Dayton JR, Kabeer R, Bunney G, Boukhman M. Teaching principles of medical innovation and entrepreneurship through hackathons: case study and qualitative analysis. JMIR Med Educ 2023 Feb 24;9:e43916. [doi: 10.2196/43916] [Medline: 36826988]
9. Vandekerckhove P, Timmermans J, de Bont A, de Mul M. Diversity in stakeholder groups in generative co-design for digital health: assembly procedure and preliminary assessment. JMIR Hum Factors 2023 Feb 14;10:e38350. [doi: 10.2196/38350] [Medline: 36787170]
10. Di Giovannantonio M, Harris BH, Zhang P, et al. Heritable genetic variants in key cancer genes link cancer risk with anthropometric traits. J Med Genet 2021 Jun;58(6):392-399. [doi: 10.1136/jmedgenet-2019-106799] [Medline: 32591342]
11. Harris BHL, Di Giovannantonio M, Zhang P, et al. New role of fat-free mass in cancer risk linked with genetic predisposition. Sci Rep 2024 Mar 27;14(1):7270. [doi: 10.1038/s41598-024-54291-7] [Medline: 38538606]
12. Koizia LJ, Giovannantonio MD, Zhang P, Fertleman MB, Lole Harris BH. Genome-wide study of the UK Biobank highlights the importance of the homeobox-C gene cluster in hip fracture risk. Geriatr Orthop Surg Rehabil 2025;16:21514593251336568. [doi: 10.1177/21514593251336568] [Medline: 40292382]
13. Deitte LA, Omary RA. The power of design thinking in medical education. Acad Radiol 2019 Oct;26(10):1417-1420. [doi: 10.1016/j.acra.2019.02.012] [Medline: 30867087]
14. Teisberg E, Wallace S, O'Hara S. Defining and implementing value-based health care: a strategic framework. Acad Med 2020 May;95(5):682-685. [doi: 10.1097/ACM.0000000000003122] [Medline: 31833857]
15. Vandekerckhove P, Harris BHL, Koizia LJ, Handa A, Brainard C, Howard S. Why we need a patient-centered innovation renaissance: a horizontal and vertical integration of knowledge to transform care pathways. Patient Exp J 2025;12(2):9-13. [doi: 10.35680/2372-0247.2023]
16. Bird M, McGillion M, Chambers EM, et al. A generative co-design framework for healthcare innovation: development and application of an end-user engagement framework. Res Involv Engagem 2021 Mar 1;7(1):12. [doi: 10.1186/s40900-021-00252-7] [Medline: 33648588]
17. Harris BHL, McCabe C, Shafique H, et al. Diversity of thought: public perceptions of genetic testing across ethnic groups in the UK. J Hum Genet 2024 Jan;69(1):19-25. [doi: 10.1038/s10038-023-01199-1] [Medline: 37907557]
18. Walsh JL, Harris BHL, Smith PE. Single best answer question-writing tips for clinicians. Postgrad Med J 2017 Feb;93(1096):76-81. [doi: 10.1136/postgradmedj-2015-133893] [Medline: 27371033]
19. Ferreira-Padilla G, Ferrández-Antón T, Lolas-Stepke F, Almeida-Cabrera R, Brunet J, Bosch-Barrera J. Ethics competences in the undergraduate medical education curriculum: the Spanish experience. Croat Med J 2016 Oct 31;57(5):493-503. [doi: 10.3325/cmj.2016.57.493] [Medline: 27815940]

XSL•FO
RenderX

Edited by A Stone; submitted 24.Oct.2025; peer-reviewed by H Thongdaeng, TH Champney; revised version received 07.Mar.2026; accepted 27.Mar.2026; published 20.Apr.2026.

Please cite as:
Koizia LJ, Vandekerckhove P, Howard S, Maisch B, Harris BHL
From Anatomy to Innovation: Embedding Patient-Centered Design in Medical Education
JMIR Med Educ 2026;12:e86230
URL: https://mededu.jmir.org/2026/1/e86230
doi:10.2196/86230

© Louis John Koizia, Pieter Vandekerckhove, Steven Howard, Bettina Maisch, Benjamin Howell Lole Harris. Originally published in JMIR Medical Education (https://mededu.jmir.org), 20.Apr.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Education, is properly cited. The complete bibliographic information, a link to the original publication on https://mededu.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Bridging the Implementation Gap in Medical AI Education: 3-Lever Framework for Concurrent Reform

Sherrie L Kelly[1], BSc, MSc, PhD; Melissa Forgie[2], MSc, MD, FRCAM (Dublin); Mark C Walker[1,2,3], MSc, MD, FRCSC, MHCM

[1]Clinical Epidemiology Program, Ottawa Hospital Research Institute, 501 Smyth Road, Ottawa, ON, Canada

[2]Faculty of Medicine, University of Ottawa, Ottawa, ON, Canada

[3]Maternal Fetal Medicine, General Campus, Ottawa Hospital, Ottawa, ON, Canada

Corresponding Author:
Sherrie L Kelly, BSc, MSc, PhD
Clinical Epidemiology Program, Ottawa Hospital Research Institute, 501 Smyth Road, Ottawa, ON, Canada

Abstract

The rapid integration of artificial intelligence (AI) into clinical practice necessitates urgent restructuring of medical education and physician assessment to ensure that future physicians are proficient and responsible users of AI tools. Despite the existence of core AI competencies, the current state of AI education in Canadian undergraduate medical education is highly inconsistent and disjointed, and available data indicate that most medical students receive minimal to no formal AI training even as they anticipate that AI will profoundly shape their future careers. National policy, specifically the Pan-Canadian AI for Health Guiding Principles, has advanced the agenda by calling for AI literacy among health professionals and emphasizing core values such as equity, robust data practices, and Indigenous-led data governance. However, these principles offer limited practical guidance on the educational and regulatory mechanisms required for effective implementation. We contend that this critical implementation deficit arises from a traditional, sequential reform model in which faculty development, curriculum change, and regulatory updates occur in isolation. This slow, siloed approach is fundamentally inadequate for addressing AI's inherent speed, opacity, and significant equity risks. To overcome this challenge, we propose a 3-lever concurrent implementation framework that provides a conceptual lens to address the interdependencies among faculty development, curriculum, and regulation. This model posits that AI competencies transition from abstract requirements to practical application only when 3 levers—clinician-educator capacity, digitally enabled learning environments, and regulatory and assessment reform—are activated simultaneously and in alignment. This Viewpoint extends existing AI competency frameworks by theorizing AI curriculum implementation as a problem of concurrent lever activation and by outlining minimum concurrent actions for deans and regulators that can be adapted to competency-based medical education systems. Although illustrated with Canadian examples, the framework is designed to be transferable beyond Canada and offers a testable, licensure-level blueprint for embedding AI competence in medical education.

(JMIR Med Educ 2026;12:e88778) doi:10.2196/88778

KEYWORDS

artificial intelligence; AI; medical undergraduate education; curriculum; accreditation; health equity

Introduction

Artificial intelligence (AI)—machine-based systems that support diagnosis, risk stratification, workflow, or documentation—is now woven into clinical practice; however, undergraduate medical curricula reflect a pre-AI era. We hypothesize that unless 3 system levers—human capacity, learning environments, and regulation—are activated together, AI competencies will remain optional and inequitable. This framework synthesizes existing evidence and provides a theoretical foundation for addressing the AI implementation gap. While the framework is grounded in current knowledge, future research should focus on evaluating its practical application and assessing its effectiveness across different medical education systems, including institutional settings, regulatory frameworks, and levels of training.

Singla et al [1] have defined core AI learning elements for Canadian undergraduate medical education (UGME). Similarly, reviews and conceptual syntheses argue that AI and data science literacy should be core competencies, organized around domains such as ethics, application, communication, and quality improvement. Most efforts have been small elective pilots with limited system-level uptake [2-5]. Our 3-lever framework extends this foundation by explaining why familiar levers—faculty development, learning environments, and regulatory mandates—have not yet moved in concert for AI and by theorizing how their interaction can shift AI competence from an aspirational concept to an assessable requirement for licensure.

XSL•FO
RenderX

Across world regions, competency-based medical education systems face an AI implementation gap: competency frameworks exist; however, clinical training continues to lag [2,3,5]. Unlike static knowledge domains, AI's velocity, opacity, and inherent risk of inequity render the conventional sequential activation of these levers inadequate and unsafe. Our contribution is to frame this implementation problem as a theory of concurrent lever activation. We argue that AI competence will remain optional and inequitable as long as faculty development, learning environments, and regulatory instruments move sequentially and in isolation. Existing AI competency lists and frameworks primarily define *what* learners should know about AI, whereas most digital health and AI education implementation reports focus on a single lever, for example, introducing new courses, local pilots, or technical infrastructure without concurrent regulatory change [2-5]. In contrast, generic calls for "AI literacy" often stop at high-level advocacy without specifying which system levers must change together. Therefore, our 3-lever framework is deliberately positioned as an implementation theory that complements content-focused competency lists and single-lever curricular reports by modeling how faculty, learning environments, and regulatory instruments must be activated concurrently for AI competence to become a licensure-level requirement rather than an optional enrichment.

The 3-lever framework differs from previous AI curriculum work in 2 ways. First, it explains recurring failure modes when levers move alone, for example, faculty-led initiatives that leave students experimenting unsupervised with powerful tools or regulatory reforms that assess AI knowledge abstractly without supporting practice in digital environments. Second, it generates operational prescriptions and predictions about how specific combinations of lever activation are likely to shape learner exposure, assessment stakes, and equity outcomes.

For example, if institutions activate only lever 2 (curriculum) by introducing AI electives without lever 1 (faculty development) or lever 3 (assessment), AI exposure will remain elective, champion-dependent, and vulnerable to rapid obsolescence. If regulators move only on lever 3 by adding AI items to licensure examinations without corresponding curricular and digital infrastructure, learners will memorize AI terminology without practicing AI-mediated clinical reasoning. If health systems adopt clinical AI tools without educational levers, students will first encounter AI through unsupervised clinical use rather than in safe sandboxes. Therefore, the framework functions as a predictive implementation model for AI in medical education. Specifically, it generates testable predictions that can be examined empirically, including the following: if only lever 2 (curriculum and learning environments) is activated, AI exposure will remain elective, uneven, and heavily dependent on local champions; if levers 1 and 2 move without Lever 3 (regulation and assessment), AI teaching and digital sandboxes will expand but remain low stakes and easily displaced when curricular time is constrained; if lever 3 moves without levers 1 and 2, high-stakes examinations will incentivize rote memorization of AI terminology rather than supervised practice in AI-mediated clinical reasoning; and if all 3 levers are activated concurrently, AI competence is more likely to become a licensure-level expectation with observable performance in digital clinical environments, including competencies related to equity and data governance.

As summarized in Figure 1 [1], lever 1 strengthens human infrastructure by supporting AI- literate clinician-educators. Lever 2 redesigns the curriculum and learning environments, including digital infrastructure such as sandboxes and simulated electronic health records (EHRs). Lever 3 aligns regulatory and assessment frameworks—accreditation standards, Medical Council of Canada (MCC) objectives, licensure examinations, and entrustable professional activities (EPAs)—to make AI literacy a nonnegotiable requirement. Throughout, reforms must be operationalized through algorithmic justice and Indigenous data sovereignty—not as add-ons but as primary lenses for safety [6-8]. Although we focus on Canadian UGME, similar competency- based systems internationally face parallel challenges, and our framework is applicable across the continuum, including in postgraduate medical education and continuing professional development (postgraduate medical education alignment is illustrated in Table 1). It can also be adapted for decentralized systems, such as those in Europe, by accounting for regional variability in regulatory structures and educational priorities while maintaining its core principles.

XSL•FO
RenderX

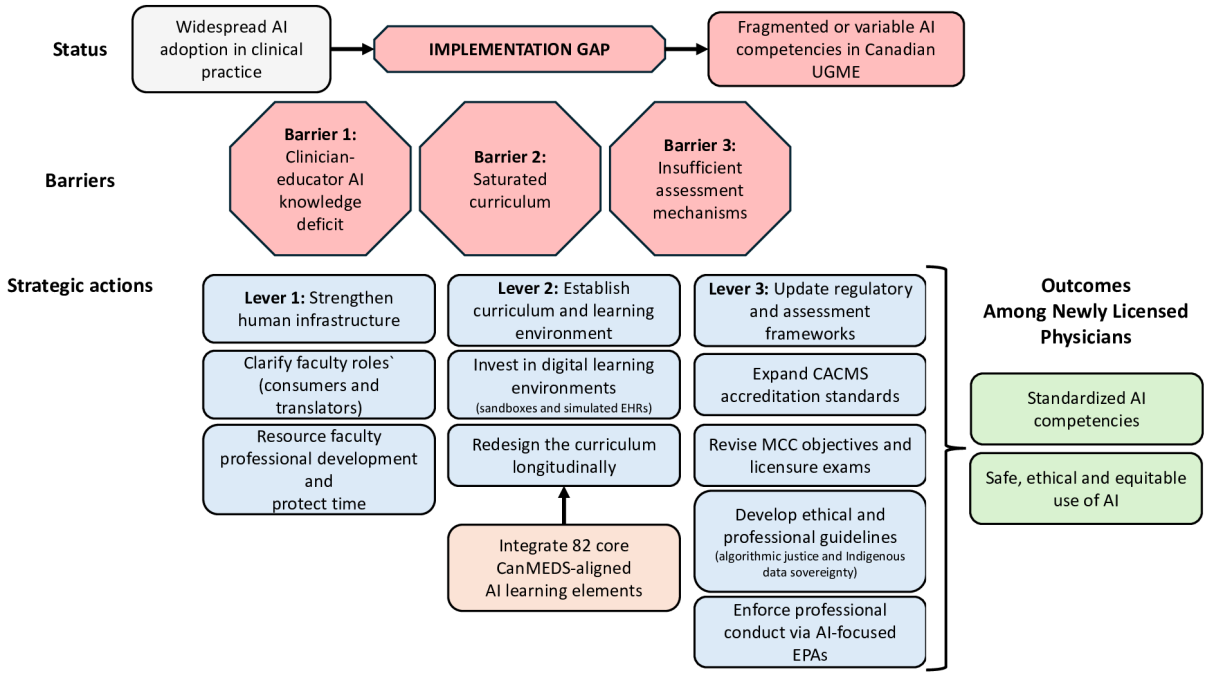
Table . Regulatory and assessment alignment for artificial intelligence (AI) competence across the Canadian and US undergraduate medical education (UGME)–postgraduate medical education (PGME) continuum.

Phase	Core AI competency focus	Lever 1: faculty capacity (who)	Lever 2: learning environment (how)	Lever 3: regulation and assessment (what)
UGME (foundation)	AI literacy, safety, and ethics with focus on understanding bias, data provenance, and critical appraisal of AI-generated information	Designate clinical AI leads for curriculum mapping and longitudinal faculty mentoring	Integrate AI training longitudinally into core clerkship and ethics sessions and practice in shared digital sandboxes	• Canadian alignment: Committee on Accreditation of Canadian Medical Schools standards and Medical Council of Canada objectives (digital health focus) • US alignment: Liaison Committee on Medical Education standards and United States Medical Licensing Examination step 1 or 2
PGME (application)	Safe, supervised AI, and oversight with focus on specialty-specific applications, identifying system failures, and QI[a] related to AI	Embedding of AI competency into academic half-day and QI projects with dedication of AI-focused faculty mentors in subspecialties	Use of AI-enabled electronic health records in workplace-based assessment, with mandatory debriefs on automation bias cases	• Canadian alignment: Royal College of Physicians and Surgeons of Canada or College of Family Physicians of Canada EPAs[b] • US alignment: Accreditation Council for Graduate Medical Education milestones and specialty-specific EPAs

[a]QI: quality improvement.

[b]EPA: entrustable professional activity.

Figure 1. Multisystem framework for bridging the artificial intelligence (AI) competency gap in undergraduate medical education (UGME). The 82 CanMEDS (Canadian Medical Education Directives for Specialists)–aligned AI learning elements referred to in the figure were developed by Singla et al [1]. CACMS: Committee on Accreditation of Canadian Medical Schools; EHR: electronic health record; EPA: entrustable professional activity; MCC: Medical Council of Canada.



XSL•FO
RenderX

Policy to Practice

Our Viewpoint draws on 4 strands of literature and policy. First, the expert-based curriculum developed by Singla et al [1] provides detailed CanMEDS (Canadian Medical Education Directives for Specialists)–aligned AI learning objectives for UGME, specifying what students should know. However, it is largely agnostic about system- level levers—accreditation, licensure, and national assessment—that would make these competencies mandatory.

Second, recent reviews and a conceptual paper argue that AI and data science literacy should be core training outcomes and find that most AI curricula remain fragmented and elective [2-5]. Few are anchored in high- stakes assessment or accreditation, leaving exposure variable and dependent on local champions.

Third, empirical studies of Canadian learners confirm this patchwork reality. A mixed methods study across Canadian medical schools found that most students (85%) reported no formal AI education despite anticipating that AI would significantly affect their future practice [9]. A national survey of health professional students found similar patterns: most predicted AI would shape their careers yet felt unprepared and unsure how to respond [10].

Fourth, the AI for Health (AI4H) Guiding Principles, endorsed by Canadian health ministers, set a policy context that stresses AI literacy, equity, and Indigenous- led data governance and calls for robust oversight across health systems [6]. However, AI4H does not articulate specific educational reforms or regulatory mechanisms, instead assuming that health professionals will be able to understand, critique, and oversee AI systems without specifying how or where that competence will be cultivated and verified.

Taken together, these strands show that Canada has detailed AI competencies for UGME, evidence that learners are not consistently achieving them, and a policy framework that demands AI- literate, equity-focused physicians—but no coherent implementation strategy linking competencies, curricula, faculty development, infrastructure, accreditation, and assessment. Existing work has described what to teach and how AI should behave but lacks a theory explaining why AI competencies have not become routine and which system levers must move together to change that. While this framework addresses these gaps, its successful implementation will require careful navigation of political and economic barriers, including stakeholder alignment, resource allocation, and competing institutional priorities. The 3-lever framework provides such a theory by linking content frameworks (eg, the curriculum developed by Singla et al [1]), learner surveys, and policy principles (eg, AI4H) to specific combinations of faculty development, learning environments, and regulatory instruments and by making testable predictions about how different patterns of lever activation will entrench or narrow the current implementation gap.

In this Viewpoint, we synthesize national policy, competency frameworks, and empirical work with Canadian learners to propose a 3-lever concurrent implementation model for AI in UGME. We show how different patterns of lever activation predict whether AI competence remains optional or becomes a licensure-level expectation and outline concrete regulatory and curricular actions that can be adapted to competency-based medical education systems internationally.

Lever 1: Human Infrastructure

Lever 1 focuses on who can teach and supervise responsible AI use: the human infrastructure of clinician-educators. Most clinicians were trained in a pre-AI era and report limited confidence in teaching, supervising, or evaluating AI-mediated clinical reasoning [11].

Faculty development in clinical AI should be treated as foundational infrastructure rather than optional enrichment, given its necessity for safe clinical oversight. This includes equipping educators with key competencies such as critically appraising AI models, interpreting performance metrics, identifying biases, and supervising learners in the use of AI tools within clinical contexts. Without AI-literate educators, even well-designed curricula will be delivered inconsistently, and learners will likely rely on unsupervised experimentation with powerful tools. Development efforts should be longitudinal and rapid-cycle, focusing less on turning clinicians into data scientists and more on oversight skills that they already apply to other technologies: understanding how AI models are developed and validated, interpreting performance metrics, recognizing when models are used outside their intended population or context, and supervising learners who use AI-enabled tools [2-5].

Resourcing faculty development is essential. Without protected time and institutional recognition, programs risk becoming perfunctory. Institutions should leverage funding for teaching innovation; allocate dedicated time to clinical AI leads; and recognize AI-focused teaching, mentoring, and curriculum reform in promotion and merit processes. While some Canadian faculties have begun appointing clinicians with expertise in AI and education as “clinical AI leads,” this role remains aspirational in many institutions and requires support to become widespread. These roles typically focus on faculty mentoring, curriculum design, and serving as liaisons with data governance structures.

Operationally, the 3-lever framework implies that schools should treat AI-focused faculty development as minimum safety infrastructure rather than optional enrichment. At a minimum, each school should designate a clinical AI lead with protected time, embed AI oversight skills into existing faculty development programs for clinical teachers, and track whether every clerkship discipline has at least 1 educator capable of supervising AI-mediated clinical reasoning.

A common counterargument is that increasingly user- friendly AI interfaces will allow clinicians to use them safely without extensive training [12]. This view underestimates both the sociotechnical complexity of clinical AI and the evidence from safety science. Studies on automation bias—the tendency to overtrust automated recommendations—show persistent risk

XSL•FO
RenderX

of overreliance on algorithms, underrecognition of new forms of error, and diffusion of responsibility when decision support systems are introduced [7,12,13]. A study on AI- driven decision-making similarly cautions that apparent fluency can mask shallow understanding [14]. Without faculty who can teach these hazards and model algorithmic skepticism—a stance of critical engagement with AI outputs that prioritizes identifying and mitigating bias and equity harms—AI tools may be used uncritically [7,12].

Lever 2: Curriculum and Learning Environments

Lever 2 focuses on where and how learners encounter AI: curricula and learning environments. Singla et al [1] recommend that AI learning be woven longitudinally throughout UGME rather than confined to stand- alone electives, a view echoed by global reviews that find that piecemeal, elective-based approaches reinforce inequities in exposure and rarely shape assessment or accreditation [2-5].

Early in training, students can be introduced to core AI and data science concepts—such as how bias enters through data collection, labeling, and outcome selection—through case-based discussions and clinical examples. These cases must be designed to interrogate the ethical and sovereignty implications of data used for training AI systems [6,8]. Later, as students move into clinical skills courses and clerkship, AI-enabled tools should appear in realistic scenarios, for example, decision support suggesting a differential diagnosis or a risk score influencing triage. Cases should require learners to reconcile AI outputs with examination findings, patient narratives, and contextual information so that AI becomes part of routine clinical reasoning pedagogy rather than an isolated topic.

Generative AI (GenAI) further adds complexity. Large language models now perform at or near passing thresholds on United States Medical Licensing Examination (USMLE)– style assessments, raising questions about what such tests measure [14,15]. Curricula should address GenAI both as content and as context: students should learn responsible use (disclosure, verification, and professional boundaries), and assessment formats that are easily outsourced to GenAI (eg, unsupervised take- home essays) will require redesign to reflect individual competence and uphold academic integrity [16,17]. This connects lever 2 to lever 3 because GenAI forces reconsideration of both teaching strategies and assessment design.

Digital Learning Environments

Even a well-designed curriculum will fall short if learners never interact meaningfully with AI tools. We suggest that faculties invest in secure digital “sandboxes” and simulated EHRs that embed AI decision support tools. In a sandbox, students can explore how models behave with different inputs, observe performance changes when data distributions shift, and experiment with prompting strategies for GenAI using deidentified or synthetic data in a test environment [10]. This allows observation rather than merely reading about phenomena such as model drift, bias, and overfitting. For example, a second-year student might work through a simulated emergency department case in which entering different vital signs, comorbidities, and social history variables changes an AI-generated sepsis risk score, and the student would then be required to document why they accept or override the recommendation before debriefing with a faculty supervisor. Such concrete exercises make abstract concepts such as distribution shift, calibration, and automation bias visible and discussable.

Simulated EHRs can extend this experience to realistic workflows. Learners can navigate simulated electronic charts in which AI- driven tools surface alerts, risk scores, or suggested orders, practicing when to accept, override, or ignore AI outputs and how to explain those decisions to patients and colleagues.

From a 3-lever perspective, digital sandboxes and simulated EHRs are not aspirational extras but mechanisms that make AI competence observable. They translate abstract competencies into tasks that can be practiced and assessed, such as overriding an unsafe AI recommendation or explaining an AI-generated risk score to a patient. Without these environments, curricular objectives remain largely theoretical and cannot meaningfully support the regulatory changes proposed under lever 3.

Building this infrastructure is resource intensive. Relying solely on local resources risks creating a digital divide between medical schools. We propose leveraging existing national digital health partnerships to coordinate shared, cloud-based sandboxes. The long-term cost of graduating physicians who learn AI through unsupervised, unsafe experimentation on patients is likely higher than the investment in shared training environments. Shared infrastructure ensures equitable access without duplicating costs at every institution. We recognize the constraints facing underresourced schools; therefore, a balanced implementation must also support agile, low-resource curricular integration models in parallel with national infrastructure development.

Lever 3: Regulatory and Assessment Frameworks

Lever 3 focuses on what makes AI competence nonoptional: the regulatory and assessment frameworks that shape institutional behavior. Modernizing faculty development and curricula will have limited impact if AI competence remains optional from the standpoint of accreditation and licensure. Coordinating action across accrediting bodies, licensing boards, and Indigenous communities presents significant challenges, including jurisdictional variability, misaligned incentives, and resistance to change. However, these barriers can be addressed through clear mandates from national bodies, alignment of incentives (eg, funding for compliance), and inclusive governance structures that prioritize equity and Indigenous leadership.

As seen in other areas such as patient safety and quality improvement, accreditation standards, national objectives, licensure examinations, and EPAs have historically driven change in Canada. For example, when national bodies strengthened patient safety and quality improvement expectations within accreditation standards and objective

XSL•FO
RenderX

structured clinical examination (OSCE) blueprints, many schools rapidly introduced new safety curricula, checklists, and simulation cases to avoid accreditation and examination risk; AI-related objectives are likely to exert similar system-level effects once they are explicitly blueprinted into high-stakes assessments. AI literacy, including its equity and data governance dimensions, will likely need to follow a similar path. Although we illustrate lever 3 using Canadian (Canadian Accreditation Council for Medical Schools and MCC) and US (Liaison Committee on Medical Education and USMLE) structures, the underlying principle—making AI competence a licensure-level expectation rather than an optional elective—applies to any competency-based medical education system.

At present, AI-related competencies are largely implicit or absent from these frameworks. Accreditation standards from the Committee on Accreditation of Canadian Medical Schools and postgraduate bodies emphasize broad domains such as patient safety, professionalism, and use of health information technologies, which are relevant to AI but do not explicitly mandate AI competencies. We suggest that accrediting bodies articulate expectations that graduates can interpret and critically appraise AI tools, understand their equity and data sovereignty implications, and function safely in AI- enabled clinical environments [18]. Standards should focus on outcomes and principles rather than prescribing detailed curricula, for example, requiring evidence of a longitudinal AI learning thread mapped to national competencies, AI- related faculty development, and explicit attention to AI4H principles and Indigenous data sovereignty within ethics and professionalism curricula while leaving design to local context [1,4,6].

Licensure and certification examinations are equally important. The MCC objectives (revised 2022) explicitly address general competence in digital health, including AI and machine learning, but a detailed assessment framework is still needed. As the Council is in the process of modernizing its digital infrastructure, it has an opportunity to incorporate AI- related content into examination objectives and instruments [19]. We suggest that the MCC objectives be revised so that graduates are expected to interpret AI outputs and performance metrics in a clinical context, recognize bias and equity implications, and communicate AI-related benefits and limitations to patients. These expectations should be assessed through written questions that require candidates to critique AI- generated outputs or identify unsafe uses and through OSCE stations in which candidates navigate conflicts between AI recommendations and clinical judgment or obtain informed consent for AI- mediated care [7,14]. EPAs can then embed AI competence into workplace- based assessment by translating broad competencies into observable units of work [20].

Concretely, the framework predicts that once AI-related objectives are tied to blueprinted examination content and OSCE stations that explicitly require candidates to critique AI outputs, schools will reallocate curricular time and faculty development resources toward AI competence to avoid accreditation and licensure risk. Conversely, if AI remains absent from high-stakes assessment, even well-designed curricula and sandboxes will remain vulnerable to marginalization when time and resources are constrained.

AI evolves faster than assessment cycles, so examining bodies are strongly encouraged to adopt agile, rapid-cycle blueprint reviews. This challenge is not unique to Canada. The 3-lever framework applies to any competency-based system; for example, aligning lever 3 with USMLE content outlines would drive lever 2 resource allocation in US medical schools. Aligning AI competencies with existing high-stakes regulatory instruments—whether the MCC, the Liaison Committee on Medical Education, or the Accreditation Council for Graduate Medical Education—is more effective than creating entirely new governance structures.

Ethics and Indigenous Data Sovereignty

Technical literacy alone is insufficient for safe clinical AI. Much of the early AI education literature emphasizes how models work and how they might improve efficiency or accuracy [2,3,5], with less attention given to how AI systems can reproduce inequities or to the rights of Indigenous Peoples over data that concern them. In Canada, AI4H explicitly names equity, Indigenous data sovereignty, and frameworks such as OCAP (ownership, control, access, and possession) and the CARE (collective benefit, authority to control, responsibility, and ethics) principles as core requirements for health-related AI [6,8].

We propose organizing ethical and professional teaching about AI around 2 interlocking concepts: algorithmic justice and Indigenous data sovereignty. Algorithmic justice focuses on how seemingly neutral technical decisions—such as which outcomes to predict, what data sources to include, or which proxy variables to use—can embed historical and structural inequities into AI systems [10]. For example, students can examine how pulse oximetry, dermatology image datasets underpinning diagnostic algorithms, or historical estimated glomerular filtration rate equations have systematically misclassified risk in racialized populations and can practice explaining these failures to patients and interprofessional teams. Students should interrogate AI case studies through an equity lens, asking: Who is represented in the training data? Which populations are missing? How does performance vary across groups? What types of harms might fall on those least represented in the training data? [6,7] These exercises move discussions of fairness from abstract principles to concrete professional responsibilities and should be integrated into clinical reasoning, ethics, and quality improvement components of the curriculum, rather than confined to stand-alone lectures.

Indigenous data sovereignty extends these questions by asserting that Indigenous nations have the right to govern the collection, ownership, use, and sharing of data about their peoples, lands, and knowledge [8]. In a Canadian context, this includes instances in which administrative or research datasets have been used to develop AI models about Indigenous communities without respecting community-defined governance processes or data sharing agreements. This is directly relevant to AI systems trained on health, genomic, or geospatial data that include Indigenous communities. Canadian guidance emphasizes

XSL•FO
RenderX

that Indigenous data governance frameworks should inform health system practices and that AI principles must respect Indigenous governance [6,8]. Therefore, UGME should include Indigenous- led teaching on data sovereignty and AI, co-designed with Indigenous organizations and aligned with reconciliation commitments, and graduates should be able to recognize when a proposed AI deployment lacks explicit Indigenous data governance agreements and escalate such concerns appropriately.

Algorithmic justice and Indigenous data sovereignty are central to all 3 levers. Faculty development (lever 1) should equip educators to teach about bias, inequity, and data governance. Curricular design and digital environments (lever 2) should embed simulations that surface these issues, rather than presenting only "average" patients. Regulatory and assessment frameworks (lever 3) should explicitly reference equity and data sovereignty expectations, so that graduates are assessed not only on technical understanding but also on ethical and community-responsive use of AI. Concretely, programs can blueprint at least 1 written examination question and 1 OSCE or simulated case in which learners must recognize and articulate an algorithmic bias or a breach of an Indigenous data governance agreement and propose an alternative course of action. Assessing these scenarios signals that graduation and licensure depend not only on technical understanding but also on ethical and community-responsive use of AI.

Conclusions

Frameworks such as the 82 essential AI competencies for Canadian UGME [1] and policy documents such as AI4H [6] set expectations for literacy, safety, equity, and Indigenous data sovereignty. The central question for academic medicine is how to move from these frameworks and principles to practice so that every graduating physician can use AI responsibly.

We have argued that closing this implementation gap requires coordinated action across 3 levers—human infrastructure, curriculum and learning environments, and regulatory systems. Textbox 1 and Table 1 translate this framework into a set of minimum concurrent actions and illustrative alignments across training phases and jurisdictions. This framework builds on existing evidence and offers a structured approach to embedding AI competencies into medical education. While grounded in evidence, its implementation across diverse settings will offer opportunities for refinement and optimization.

AI's urgency and interconnectedness favor applying these 3 levers concurrently rather than sequentially. However, their application must account for the complexity of nested systems with competing priorities, requiring adaptive and context-sensitive approaches. Faculties should invest in AI-literate clinician-educators; redesign curricula to integrate AI into clinical reasoning, ethics, and systems- based practice, taught in partnership with Indigenous communities and equity experts and practiced within digital sandboxes and simulated EHRs [1,4,5]; and collaborate with accrediting bodies, the MCC, and specialty colleges to embed AI competencies into accreditation standards, examination objectives, and EPAs, with mechanisms for rapid-cycle adjustment.

Textbox 1. Minimum concurrent actions for deans and regulators to embed artificial intelligence (AI) competence in undergraduate medical examination.

Lever 1: human infrastructure

- Appoint at least 1 clinical AI lead or a small team (eg, 0.1 - 0.2 full-time equivalent each) at every medical school to coordinate AI teaching and supervision
- Integrate a short AI oversight module (eg, 3 - 4 sessions) into existing faculty development programs, focusing on supervising AI-mediated clinical reasoning
- Ensure, at the program level, that each core clerkship has access to at least 1 AI-literate supervisor for learners

Lever 2: curriculum and learning environment

- Map the 82 CanMEDS (Canadian Medical Education Directives for Specialists)–aligned AI learning elements onto existing courses and clinical experiences to identify overlaps and gaps
- Introduce at least 1 shared digital sandbox or simulated electronic health record exercise in both preclerkship and clerkship phases in which students must work with AI outputs
- Embed longitudinal case discussions in which students reconcile AI recommendations with physical examination findings, social history, and patient preferences

Lever 3: regulatory and assessment frameworks

- Add explicit AI-related enabling language to accreditation self-study templates and program evaluation reports
- Blueprint a defined proportion (eg, 5% - 10%) of written examination content and objective structured clinical examination stations for AI literacy, equity, and data governance objectives
- Develop or adapt at least 1 entrustable professional activity that requires safe, supervised use of AI tools in clinical decision-making as a condition of graduation or licensure

XSL•FO
RenderX

Future work should evaluate how different models of faculty development, simulation design, and regulatory reform influence AI oversight skills, susceptibility to automation bias, and downstream outcomes such as quality of care and equity and should examine how Indigenous-led data governance frameworks can be authentically incorporated into medical education and how these levers can be adapted for diverse resource settings and regulatory architectures [6,8]. If Canadian medical education fails to act, physicians risk graduating underprepared for AI-mediated practice, with consequences for patient safety, equity, and public trust. Conversely, timely and coordinated reform—grounded in algorithmic justice and Indigenous data sovereignty and operationalized through the 3 levers we propose—can help ensure that AI becomes a well-governed tool that supports safer, more ethical, and more equitable care across Canada and beyond.

Funding

The authors declare that no financial support was received for this work.

Data Availability

Data sharing is not applicable to this paper as no datasets were generated or analyzed during this study.

Authors' Contributions

SLK and MCW conceived the idea for the viewpoint. SLK conducted the review and analysis and drafted the initial manuscript. MF and MCW provided expert policy context and critically reviewed the manuscript. All authors read and approved the final manuscript.

Conflicts of Interest

None declared.

References

1. Singla R, Pupic N, Ghaffarizadeh SA, et al. Developing a Canadian artificial intelligence medical curriculum using a Delphi study. NPJ Digit Med 2024 Nov 18;7(1):323. [doi: 10.1038/s41746-024-01307-1] [Medline: 39557985]
2. Chan KS, Zary N. Applications and challenges of implementing artificial intelligence in medical education: integrative review. JMIR Med Educ 2019 Jun 15;5(1):e13930. [doi: 10.2196/13930] [Medline: 31199295]
3. Ghaffarizadeh A, Pupic N, Hu R. Artificial intelligence education for medical students: a systematic review. CMBES Proc 2023;45:1-6 [FREE Full text]
4. Seth P, Hueppchen N, Miller SD, et al. Data science as a core competency in undergraduate medical education in the age of artificial intelligence in health care. JMIR Med Educ 2023 Jul 11;9:e46344. [doi: 10.2196/46344] [Medline: 37432728]
5. Simoni J, Urtubia-Fernandez J, Mengual E, et al. Artificial intelligence in undergraduate medical education: an updated scoping review. BMC Med Educ 2025 Nov 17;25(1):1609. [doi: 10.1186/s12909-025-08188-2] [Medline: 41250109]
6. Pan-Canadian AI for health (AI4H) guiding principles. Health Canada. 2025. URL: https://www.canada.ca/en/health-canada/corporate/transparency/health-agreements/pan-canadian-ai-guiding-principles.html [accessed 2026-03-24]
7. Challen R, Denny J, Pitt M, Gompels L, Edwards T, Tsaneva-Atanasova K. Artificial intelligence, bias and clinical safety. BMJ Qual Saf 2019 Mar;28(3):231-237. [doi: 10.1136/bmjqs-2018-008370] [Medline: 30636200]
8. Indigenous data sovereignty (DDN3-A11). Canada School of Public Service (CSPS). 2025. URL: https://www.csps-efpc.gc.ca/tools/articles/indigenous-data-sovereignty-eng.aspx [accessed 2026-04-08]
9. Pucchio A, Rathagirishnan R, Caton N, et al. Exploration of exposure to artificial intelligence in undergraduate medical education: a Canadian cross-sectional mixed-methods study. BMC Med Educ 2022 Nov 28;22(1):815. [doi: 10.1186/s12909-022-03896-5] [Medline: 36443720]
10. Teng M, Singla R, Yau O, et al. Health care students' perspectives on artificial intelligence: countrywide survey in Canada. JMIR Med Educ 2022 Jan 31;8(1):e33390. [doi: 10.2196/33390] [Medline: 35099397]
11. Task force report on artificial intelligence and emerging digital technologies. Royal College of Physicians and Surgeons of Canada. 2020. URL: https://www.royalcollege.ca/content/dam/document/membership-and-advocacy/2020-task-force-report-on-ai-and-emerging-digital-technologies-e.pdf [accessed 2025-12-02]
12. Khera R, Simon MA, Ross JS. Automation bias and assistive AI: risk of harm from AI-driven clinical decision support. JAMA 2023 Dec 19;330(23):2255-2257. [doi: 10.1001/jama.2023.22557] [Medline: 38112824]
13. Goddard K, Roudsari A, Wyatt JC. Automation bias: a systematic review of frequency, effect mediators, and mitigators. J Am Med Inform Assoc 2012;19(1):121-127. [doi: 10.1136/amiajnl-2011-000089] [Medline: 21685142]
14. Gilson A, Safranek CW, Huang T, et al. How does ChatGPT perform on the United States Medical Licensing Examination (USMLE)? The implications of large language models for medical education and knowledge assessment. JMIR Med Educ 2023 Feb 8;9:e45312. [doi: 10.2196/45312] [Medline: 36753318]

XSL•FO
RenderX

15. Mbakwe AB, Lourentzou I, Celi LA, Mechanic OJ, Dagan A. ChatGPT passing USMLE shines a spotlight on the flaws of medical education. PLOS Digit Health 2023 Feb;2(2):e0000205. [doi: 10.1371/journal.pdig.0000205] [Medline: 36812618]
16. Ichikawa T, Olsen E, Vinod A, et al. Generative artificial intelligence in medical education-policies and training at US osteopathic medical schools: descriptive cross-sectional survey. JMIR Med Educ 2025 Feb 11;11:e58766. [doi: 10.2196/58766] [Medline: 39934984]
17. Kang J, Ahn J. Technologies, opportunities, challenges, and future directions for integrating generative artificial intelligence into medical education: a narrative review. Ewha Med J 2025 Oct;48(4):e53. [doi: 10.12771/emj.2025.00787] [Medline: 41035210]
18. Rethinking technology. Medical Council of Canada. 2024 Nov 27. URL: https://mcc.ca/news/rethinking-technology/ [accessed 2025-12-02]
19. MCC examination objectives. Medical Council of Canada. URL: https://mcc.ca/objectives/ [accessed 2025-12-02]
20. Ten Cate O. Nuts and bolts of entrustable professional activities. J Grad Med Educ 2013 Mar;5(1):157-158. [doi: 10.4300/JGME-D-12-00380.1] [Medline: 24404246]

Abbreviations

AI: artificial intelligence
AI4H: Artificial Intelligence for Health
CanMEDS: Canadian Medical Education Directives for Specialists
CARE: collective benefit, authority to control, responsibility, and ethics
EHR: electronic health record
EPA: entrustable professional activity
GenAI: generative artificial intelligence
MCC: Medical Council of Canada
OCAP: ownership, control, access, and possession
OSCE: objective structured clinical examination
UGME: undergraduate medical education
USMLE: United States Medical Licensing Examination

Edited by J Gentges; submitted 02.Dec.2025; peer-reviewed by D Dellacorte, Z Liu; revised version received 21.Feb.2026; accepted 26.Feb.2026; published 28.Apr.2026.

Please cite as:
Kelly SL, Forgie M, Walker MC
Bridging the Implementation Gap in Medical AI Education: 3-Lever Framework for Concurrent Reform
JMIR Med Educ 2026;12:e88778
URL: https://mededu.jmir.org/2026/1/e88778
doi:10.2196/88778

© Sherrie L Kelly, Melissa Forgie, Mark C Walker. Originally published in JMIR Medical Education (https://mededu.jmir.org), 28.Apr.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Education, is properly cited. The complete bibliographic information, a link to the original publication on https://mededu.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

From Classroom to Operating Room: Addressing Shortcomings in Gender Affirmation Surgical Education

Samuel G Robinson[1], BS; Alireza Hamidian Jahromi[2], MD, MRCS

[1]Lewis Katz School of Medicine, Philadelphia, PA, United States

[2]Division of Plastic and Reconstructive Surgery, Temple University Medical Center, 3401 N Broad St, Philadelphia, PA, United States

Corresponding Author:
Alireza Hamidian Jahromi, MD, MRCS
Division of Plastic and Reconstructive Surgery, Temple University Medical Center, 3401 N Broad St, Philadelphia, PA, United States

Abstract

The prevalence of American youth identifying as transgender has doubled in the past five years, emphasizing the crucial role of gender affirmation surgery (GAS) in treating gender dysphoria. However, the current health care infrastructure faces challenges in meeting the escalating demand for GAS interventions. Transgender and gender-diverse patients encounter barriers such as travel and extended waitlists for specialized surgeons. This viewpoint highlights the insufficient exposure to GAS in medical education, spanning from medical students to attending surgeons, and examines the uneven distribution of GAS practitioners across specialties. The absence of formal training and board certification compounds the issue, urging a comprehensive reevaluation of medical education to ensure quality care for the expanding transgender and gender-diverse population.

(JMIR Med Educ 2026;12:e71843) doi:10.2196/71843

KEYWORDS

medical training; education; transgender; residency; plastic surgery

The percentage of American youth (aged 13‐17 years) identifying as transgender has doubled in the past five years, from 0.7% in 2017 to 1.43% in 2022 [1,2]. With gender affirmation surgery (GAS) serving a vital role in the treatment of gender dysphoria, we must ask if our current infrastructure is prepared to meet the increasing demand for GAS interventions, both in terms of clinical management and for the training of the next generation of gender affirmation–trained surgeons.

It is well-documented that transgender and gender-diverse patients face increased barriers to health care compared to cisgender patients. Patients are often forced to travel out of state to access treatment from a gender affirmation–trained surgeon [3]. Even after locating a gender affirmation–trained surgeon, patients commonly experience prolonged waitlist periods, reflecting broader surgical capacity constraints as well as the limited number of surgeons offering procedure-specific, gender-affirming care [4]. This discrepancy is expected to increase as the transgender and gender-diverse population grows. The medical education system must react to the evolving needs of this already marginalized patient population by working to expand the number of GAS practitioners. Unfortunately, it is the authors' opinion that the lack of adequate exposure to GAS spans all levels of medical education and training, including medical students, residents, fellows, and attending surgeons from different surgical disciplines.

The current pool of specialties providing gender-affirming surgical care includes plastic surgeons, urologists, otolaryngologists, general surgeons, surgical oncologists, gynecologists, and oral maxillofacial surgeons, with plastic surgeons making up the vast majority of GAS practitioners [4]. It is estimated that a total of 660 physicians in the United States offer at least one GAS procedure [4]. When compared to the United States' 1.6 million transgender individuals, there is one gender-affirming surgeon per 2400 transgender patients, who may rely on surgical intervention for the treatment of gender dysphoria. Unlike population-based ratios used in other surgical specialties, demand for GAS is concentrated within a subset of patients and limited by procedure specificity, such that population-level ratios likely underestimate unmet surgical need [5-7]. We have to bear in mind that each transgender and gender-diverse individual may be interested in GAS in more than one area of his/her/their body. Additionally, some of the surgical procedures may require multiple surgical steps. For instance, masculinization bottom surgery may involve metoidioplasty with or without urethral lengthening, suprapubic catheter placement, hysterectomy, oophorectomy, phalloplasty, glansplasty, scrotoplasty, monsplasty, and the insertion of testicular and/or penile implants. To make matters worse, the geographic distribution of plastic surgeons and transgender and gender-diverse individuals do not align. Plastic surgeons are more often located in zip codes with higher household incomes, while transgender and gender-diverse people are more commonly located in areas with lower household incomes [8,9].

XSL•FO
RenderX

Further identifying surgeons with adequate experience in GAS is extremely difficult, especially for patients with lower health literacy. While not currently in development, the formation of a board-certifying body in GAS could provide both practitioners and patients the tools needed to increase access to care. Board certification is associated with improved patient outcomes [10] and would serve as a means for transgender and gender-diverse patients to evaluate whether a surgeon or procedure is a safe and smart option. There remains no formal board certification for surgeons who practice GAS. This is in part due to the lack of formal training available for surgeons interested in the specialty.

A cross-sectional study from 2022 identified only 7 GAS fellowships available for plastic surgeons in the United States [11]. Each fellowship lasts 12 months and offers an inconsistent mix of training in feminizing or masculinizing chest and genital surgeries and facial feminization, with some only offering one of the three surgical categories. The presence of designated GAS fellowships reflects the increasing demand for gender affirmation–trained surgeons. However, with 525 plastic surgeons offering GAS procedures in the United States [4], it is safe to assume the vast majority of contemporary gender-affirming surgeons are not fellowship trained. This fact alone leads one to question if residency programs are offering sufficient education on the subject to warrant practitioners the means to include GAS in their practice.

In contrast to the United States, where no formal certification or credentialing pathway is required for surgeons to perform GAS, several countries have national or regional approval mechanisms that regulate who is permitted to provide GAS. In the United Kingdom, surgeons must be credentialed through National Health Service (NHS)–commissioned gender services and operate within designated centers [12]. These models function as de facto training and certification pathways. However, by limiting GAS to a small number of designated centers, these centralized systems may constrain surgical capacity and geographic access. Hybrid approaches to education, such as expanding structured GAS training within US residency programs, may better balance quality and availability, although a direct international comparison remains challenging.

The Accreditation Council of Graduate Medical Education (ACGME) designates the education components required for a plastic surgery residency program to be accredited. A resident must complete at least 1150 surgical cases in order to graduate [13]. While many of the required cases involve surgical technique similar to those used in GAS procedures, none of the 1150 cases specifically refer to GAS interventions. For most surgeons performing GAS today, the unregulated exposure they have received in residency is the extent of their training in this specialty. ACGME is encouraged to expand their required plastic surgery resident competencies to include GAS-specific operations so that the growing transgender and gender-diverse patient population receives quality care even when a practitioner is not trained in a GAS fellowship. Importantly, high-quality exposure to gender-affirming surgery during residency does not require fellowship-trained faculty, but rather structured operative involvement, multidisciplinary care, and longitudinal patient follow-up [14,15]. With this said, the issue of inadequate education on the topic originates far before residency training begins.

The Liaison Committee on Medical Education (LCME) requires all MD-accredited medical schools to include curricular content on health care disparities including gender identity [16]. Despite this requirement, only 26.1% of the 4070 surveyed medical students feel comfortable discussing GAS with transgender and gender-diverse patients [17]. To what extent gender-diverse care is included in today's medical school curriculum is largely unknown. Medical schools vary significantly in the manner in which they mandate or voluntarily offer gender-based education, with some requiring students to attend multiple lecture series on the topic and some simply allowing students to seek out the knowledge if they please. A survey encompassing students from 175 medical schools found that 37% of respondents attended institutions where mandatory preclinical instruction was provided on "LGBT patient care" [17]. With medical schools choosing to disregard the importance of gender-conscious care in their curriculum, the transgender and gender-diverse patient population relies on students seeking out the knowledge on their own. The medical education system originates in the preclinical classrooms of medical schools, where students will carry forward the principles they are taught into their careers as resident physicians, clinical fellows, and eventually attendings. The educational system's disregard for gender-focused learning in medical school is materializing as consequences shouldered by transgender and gender-diverse patients. This marginalized population is desperately seeking out gender-affirming surgeons who do not have the time, resources, or formal training needed to meet the needs of the growing transgender and gender-diverse community.

Funding

The authors declared no financial support was received for this work.

Conflicts of Interest

None declared.

References

1. Herman JL, Flores AR, O'Neill KK. How many adults and youth identify as transgender in the United States? : The Williams Institute, UCLA School of Law; 2022 URL: https://williamsinstitute.law.ucla.edu/wp-content/uploads/Trans-Pop-Update-Jun-2022.pdf [accessed 2026-05-12]

XSL•FO
RenderX

2. Herman JL, Flores AR, Brown TNT, Wilson BDM, Conron KJ. Age of individuals who identify as transgender in the United States. : The Williams Institute, UCLA School of Law; 2017 URL: https://williamsinstitute.law.ucla.edu/wp-content/uploads/Age-Trans-Individuals-Jan-2017.pdf [accessed 2026-05-12]
3. Puckett JA, Cleary P, Rossman K, Newcomb ME, Mustanski B. Barriers to gender-affirming care for transgender and gender nonconforming individuals. Sex Res Social Policy 2018 Mar;15(1):48-59. [doi: 10.1007/s13178-017-0295-8] [Medline: 29527241]
4. Cohen W, Maisner RS, Mansukhani PA, Keith J. Barriers to finding a gender affirming surgeon. Aesthetic Plast Surg 2020 Dec;44(6):2300-2307. [doi: 10.1007/s00266-020-01883-z] [Medline: 32696165]
5. Wright JD, Chen L, Suzuki Y, Matsuo K, Hershman DL. National estimates of gender-affirming surgery in the US. JAMA Netw Open 2023 Aug 1;6(8):e2330348. [doi: 10.1001/jamanetworkopen.2023.30348] [Medline: 37610753]
6. Borah L, Zebib L, Sanders HM, Lane M, Stroumsa D, Chung KC. State restrictions and geographic access to gender-affirming care for transgender youth. JAMA 2023 Jul 25;330(4):375-378. [doi: 10.1001/jama.2023.11299] [Medline: 37490093]
7. Robinson SG, Mandel A, Nicosia J, Siegel J, Hamidian Jahromi A. Racial disparity in gender affirming surgery: a comparative study on plastic surgeon social media use. Plast Reconstr Surg Glob Open 2023 May;11(5):e5009. [doi: 10.1097/GOX.0000000000005009] [Medline: 37197009]
8. Bauder AR, Sarik JR, Butler PD, et al. Geographic variation in access to plastic surgeons. Ann Plast Surg 2016 Feb;76(2):238-243. [doi: 10.1097/SAP.0000000000000651] [Medline: 26545221]
9. Crissman HP, Berger MB, Graham LF, Dalton VK. Transgender demographics: a household probability sample of US adults, 2014. Am J Public Health 2017 Feb;107(2):213-215. [doi: 10.2105/AJPH.2016.303571] [Medline: 27997239]
10. Sharp LK, Bashook PG, Lipsky MS, Horowitz SD, Miller SH. Specialty board certification and clinical outcomes: the missing link. Acad Med 2002 Jun;77(6):534-542. [doi: 10.1097/00001888-200206000-00011] [Medline: 12063199]
11. Finkelstein E, Ha M, Ngaage LM, et al. An overview of gender-affirming surgical fellowships in the United States. Plast Reconstr Surg Glob Open 2022;10(10S):149-150. [doi: 10.1097/01.GOX.0000899108.03250.1d]
12. Service specification: gender identity services for adults (surgical interventions).: NHS England; 2019. URL: https://www.england.nhs.uk/publication/service-specification-gender-identity-services-for-adults-surgical-interventions/ [accessed 2023-01-09]
13. Review Committee for Plastic Surgery. Defined category minimums for plastic surgery. Accreditation Council for Graduate Medical Education (ACGME). 2023. URL: https://www.acgme.org/globalassets/pfassets/programresources/operative_minimums_effective_07012014.pdf [accessed 2026-05-12]
14. Khouri AN, Haley C, MacEachern M, Morrison SD. Current concepts in gender-affirming surgery postgraduate training. Indian J Plast Surg 2022 Apr;55(2):129-138. [doi: 10.1055/s-0042-1751021] [Medline: 36017402]
15. Morrison SD, Dy GW, Chong HJ, et al. Transgender-related education in plastic surgery and urology residency programs. J Grad Med Educ 2017 Apr;9(2):178-183. [doi: 10.4300/JGME-D-16-00417.1] [Medline: 28439350]
16. Functions and structure of a medical school: standards for accreditation of medical education programs leading to the M.D. degree. Liaison Committee on Medical Education. 2023. URL: https://lcme.org/standards-and-rules/ [accessed 2026-05-12]
17. White W, Brenman S, Paradis E, et al. Lesbian, gay, bisexual, and transgender patient care: medical students' preparedness and comfort. Teach Learn Med 2015;27(3):254-263. [doi: 10.1080/10401334.2015.1044656] [Medline: 26158327]

Abbreviations

ACGME: Accreditation Council of Graduate Medical Education
GAS: gender affirmation surgery
LCME: Liaison Committee on Medical Education
NHS: National Health Service

Edited by A Stone; submitted 27.Jan.2025; peer-reviewed by A Edalatpour, S Jung, T Neiman; revised version received 02.Feb.2026; accepted 04.Feb.2026; published 04.Jun.2026.

Please cite as:
Robinson SG, Hamidian Jahromi A
From Classroom to Operating Room: Addressing Shortcomings in Gender Affirmation Surgical Education
JMIR Med Educ 2026;12:e71843
URL: https://mededu.jmir.org/2026/1/e71843
doi:10.2196/71843

© Samuel G Robinson, Alireza Hamidian Jahromi. Originally published in JMIR Medical Education (https://mededu.jmir.org), 4.Jun.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License

XSL•FO
RenderX

(https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Education, is properly cited. The complete bibliographic information, a link to the original publication on https://mededu.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

AI-Assisted Formative Assessment in Clinical Education: From Algorithms to Agency

Quang Thanh Nguyen[1,2,3*], MD, MSc; Thuy Minh Ha[4*], MD, MHPE; Tina Mai[5], MD

[1]College of Health Sciences, VinUniversity, Da Ton, Gia Lam, Hanoi, Hanoi, Vietnam

[2]Department of Surgery, Vietnam National Children's Hospital, Hanoi, Hanoi, Vietnam

[3]Royal College of Surgeons in Ireland and University College Dublin Malaysia Campus, George Town, Penang, Malaysia

[4]Center for Medical and Health Sciences Education, University of Auckland, Auckland, New Zealand

[5]School of Medicine, University of California, San Francisco, San Francisco, CA, United States

[*]these authors contributed equally

Corresponding Author:
Quang Thanh Nguyen, MD, MSc
College of Health Sciences, VinUniversity, Da Ton, Gia Lam, Hanoi, Hanoi, Vietnam

Abstract

Artificial intelligence (AI) is rapidly reshaping clinical education by embedding assessment and feedback into everyday learning activities. Medical students can now use machine learning dashboards, generative AI, large language models, and emerging agentic systems to practice clinical reasoning, communication, and procedural skills while receiving individualized feedback within seconds. However, the availability of more data and more feedback does not necessarily produce better learning. This Viewpoint is intended for clinical educators, assessment leaders, curriculum committees, faculty developers, and institutional leaders who must decide how AI should be used in formative activities without reducing education to automated scoring. AI-assisted formative assessment is defined in this paper as the intentional use of AI tools to generate, organize, and support interpretation of performance information for learning rather than grading. Its distinctive contribution lies in the scale, adaptivity, conversational simulation, pattern detection, and possible autonomy of AI systems. However, AI outputs become formative only when learners and educators interpret them critically, judge their trustworthiness, and translate them into a small number of focused follow-on learning actions. This paper synthesizes the current evidence base while noting that much of it remains early, heterogeneous, and concentrated in short-term or single-setting studies. It examines key risks, including hallucination, automation bias, epistemic overtrust, hidden curricular effects, and broader concerns related to professional identity, power asymmetries, data privacy, and inequitable access. It also presents context-specific implementation examples for preclinical case-based learning, communication and objective structured clinical examination preparation, procedural skill laboratories, clerkship learning, and programmatic assessment portfolios, together with practical implications for faculty development, institutional governance, and phased local implementation. As a Viewpoint rather than an empirical study or systematic review, the framework and examples should be interpreted as evidence-informed design propositions that require local evaluation and validation. Overall, the value of AI-assisted formative assessment depends less on the volume of AI-generated feedback than on educational designs that preserve learner agency, professional judgment, and human accountability.

(JMIR Med Educ 2026;12:e93710) doi:10.2196/93710

KEYWORDS

artificial intelligence; AI; formative assessment; feedback; medical education; clinical education

Introduction

Formative assessment has always depended on more than the delivery of comments or scores. It requires learners to understand the gap between current and desired performance, receive information that guides improvement, and have opportunities to act on that information [1,2]. In clinical education, these conditions have been difficult to achieve because observation by experienced clinicians is limited, learning encounters are unpredictable, and faculty time is scarce. Artificial intelligence (AI) now changes these conditions by making feedback available during routine learning activities, often at a scale and speed that human educators cannot match.

Recent reviews show rapid expansion of AI applications in medical education, including adaptive learning, automated assessment, simulation, virtual patients, clinical reasoning support, and curriculum reform [3-5]. The scale of this change is significant. A medical student studying late at night can now log into an AI platform, run complex clinical scenarios, and receive individualized feedback that highlights specific areas for improvement within seconds. Unlike a textbook, the system interacts. Unlike a busy attending physician, it has continuous

XSL•FO
RenderX

availability. Services that once required coordinated faculty schedules are now available on demand.

These developments extend beyond efficiency. Technology is evolving from static multiple-choice engines into dynamic partners. Machine learning algorithms and large language models (LLMs) now generate high-fidelity virtual patients, analyze procedural performance, and synthesize vast stores of medical knowledge in real time [3,4]. Generative AI has further accelerated these trends by enhancing scalability and enabling more personalized learning experiences [4]. These capabilities create new opportunities for safe, deliberate practice. They also create a pedagogical paradox: learners and faculty may receive more performance data while having less clarity about what to trust, what actions to follow, and who remains accountable for educational decisions. Most existing reports catalog tools and attitudes or describe broad opportunities and risks yet offer limited guidance for clinical teachers on how to turn AI outputs into meaningful assessment for learning [3-5].

Therefore, the gap addressed in this Viewpoint is not whether AI can generate feedback but how this information can be curated into assessment for learning. We define AI-assisted formative assessment, distinguish relevant types of AI tools, summarize the evidence base and its limitations, analyze ethical and practical risks, clarify the role of AI-assisted formative assessment within programmatic assessment, and propose actionable implementation suggestions for specific curriculum contexts. This paper is intended for clinical educators, assessment leaders, curriculum committees, faculty developers, and institutional leaders in health professions education. Its practical concern is how medical schools and clinical training programs should design AI-assisted formative activities so that AI-generated information supports learning, professional judgment, and responsible clinical practice. As a Viewpoint rather than an empirical study or systematic review, this paper offers an evidence-informed conceptual framework and implementation heuristics, not validated protocols or original outcome data.

What Counts as AI-Assisted Formative Assessment?

The term “AI” is often used broadly in medical education, but this lack of precision can obscure important differences between tools [6]. These distinctions matter because different AI systems create different educational possibilities, risks, and responsibilities [7]. The updated Organisation for Economic Co-operation and Development definition describes an AI system as a machine-based system that can infer from inputs how to generate outputs such as predictions, content, recommendations, or decisions with varying degrees of autonomy and adaptiveness [8]. Within clinical education, AI-assisted formative assessment may involve several types of AI, but these systems should not be treated as interchangeable [9].

Some systems are primarily analytic or predictive, processing learner performance data to generate scores, classifications, patterns, or recommendations. Examples include adaptive learning platforms that identify knowledge gaps, simulation dashboards that quantify procedural efficiency or safety, and automated systems that flag recurrent errors. These precision education tools may uncover nuanced learner patterns and support personalization at scale [9,10]. However, analytic systems generally lack the open-ended interactivity of generative models. Although they may reveal patterns that are difficult for human observers to detect, they usually do not engage learners in open-ended dialogue or independently plan learning activities.

Other systems, particularly generative AI and LLMs, produce text, dialogue, explanations, simulated patient responses, or feedback narratives. Their educational value lies not only in the speed of feedback generation but also in their interactivity. Learners can ask follow-up questions, test alternative reasoning pathways, rehearse clinical conversations, and compare different approaches to a case [4,11]. At the same time, these systems can generate plausible but unsupported explanations. A recent scoping review identified hallucinations and factually incorrect medical content as recurring risks in medical education applications [12]. Therefore, the use of AI systems in formative assessment requires careful verification, interpretation, and evaluative judgment, including the ability to critically assess both the AI output and one’s own clinical reasoning [13,14].

Emerging agentic AI systems represent another category. These systems may pursue goals across multiple steps, select tasks, call external tools, retrieve resources, and personalize learning pathways with greater autonomy. They may eventually support tutoring or coaching functions, although this remains more prospective than established in medical education [15]. However, their use in assessment requires stronger safeguards and clear governance structures because they may influence not only the content of feedback but also what learners practice, what evidence is collected, and which learning priorities are emphasized [6,7].

Building on formative assessment scholarship, AI-assisted formative assessment is conceptualized in this paper as the intentional use of AI tools to generate, organize, and support interpretation of performance information for the purpose of improving learning rather than grading [16,17]. Therefore, it is not synonymous with automated feedback, automated scoring, or AI-generated comments. It becomes formative only when learners and educators use AI-generated information to understand performance and decide what to do next [2,17]. The same AI output could be formative if used in a coaching conversation, summative if used to determine progression, or harmful if treated as an unquestioned verdict [18].

The conceptual contribution of AI-assisted formative assessment is that AI changes the ecology of formative assessment. Traditional formative assessment often depends on episodic human observation and scheduled feedback encounters, whereas AI-assisted formative assessment may enable more continuous, learner-initiated, and responsive cycles of practice and feedback across repeated attempts [9,19]. For example, a student may rehearse a difficult communication scenario in an AI-supported simulation, receive immediate feedback, revise the approach, and repeat the encounter before meeting with a faculty coach. Similarly, a procedural simulator may identify repeated patterns

XSL•FO
RenderX

in performance that can inform subsequent deliberate practice. These functions are meaningful only when they support rather than replace learner judgment [13]. Medical students' experiences of feedback suggest that interpretation and uptake are central to whether feedback becomes educationally useful [20]. Therefore, AI-assisted formative assessment should be understood as a strategy for developing feedback agency, meaning the learner's capacity to interpret feedback, judge its trustworthiness, and translate it into action, rather than simply as a way of increasing feedback volume [1,21].

This emphasis on interpretation, judgment, and action also helps distinguish AI-assisted formative assessment from adjacent frameworks. AI-assisted formative assessment sits alongside but is not identical to learning analytics, AI-supported feedback, and programmatic assessment. Learning analytics mainly emphasizes detecting and visualizing patterns in learner data [9,10]. AI-supported feedback refers more broadly to AI-generated comments, scores, or suggestions that may or may not be used formatively [4,22]. Programmatic assessment is a broader system in which evidence is collected across multiple occasions, methods, and contexts, with low-stakes information used to support learning and aggregated evidence used for higher-stakes decisions [23]. AI-assisted formative assessment contributes a narrower pedagogic lens by focusing on how AI-generated information is interpreted and translated into focused subsequent actions for learning. Accordingly, AI-generated feedback from a virtual patient, simulation dashboard, or LLM-based reasoning exercise should usually function as one low-stakes source of information within a broader assessment system [7,10,14,23].

Evidence Base and Current Limitations

The evidence supporting AI-assisted formative assessment is promising but still uneven. Broad reviews document increasing use of AI across medical education, including applications in teaching, assessment, simulation, clinical reasoning, and learning analytics [5,22,24]. These syntheses are valuable for mapping the field, but most are scoping, narrative, or integrative reviews rather than comparative evaluations, so they identify domains of use more readily than they establish educational effectiveness. Reviews and curriculum-focused studies on generative AI similarly describe potential benefits for personalized learning, content generation, simulated dialogue, and feedback while emphasizing challenges related to accuracy, bias, transparency, privacy, academic integrity, and educational governance [4,25].

Several empirical studies illustrate the potential of AI-assisted formative assessment. Stronger empirical evidence currently comes from bounded, single-setting studies, including simulation-based evaluations and at least one randomized controlled trial [26-28]. In surgical simulation, the Virtual Operative Assistant used explainable machine learning to translate raw simulator data into feedback aligned with expert performance benchmarks, supporting more targeted practice in technical skill training [26]. In undergraduate medical education, an LLM-based patient simulation and structured feedback intervention was evaluated in a randomized controlled trial and reported short-term support for clinical decision-making practice among medical students [27]. AI-driven personalized learning platforms have also been studied for their effects on medical students' learning performance, engagement, and self-directed learning [28].

However, existing literature remains limited and early in development. Many studies are of short-term scope; tool specific; and focused on usability, learner satisfaction, or immediate test outcomes rather than durable changes in clinical reasoning, transfer to workplace performance, patient-centered communication, or professional identity formation. Scoping work on AI-driven virtual patients for communication skill training suggests that these tools may benefit learners, but the evidence of effectiveness remains limited and requires stronger educational design, theory-informed evaluation, and stakeholder involvement [29]. Outcome measures are also heterogeneous, making cross-study comparison difficult. Therefore, claims that AI-assisted formative assessment may support feedback agency, self-regulated learning, clinical reasoning, or broader assessment system change should be interpreted as plausible mechanisms or design hypotheses informed by prior literature rather than outcomes demonstrated in this manuscript. Accordingly, the framework and examples offered below are intended as design propositions for local testing, not as validated protocols or implementation recipes.

Ethical and Practical Risks Requiring Design Controls

The main risks of AI-assisted formative assessment are not limited to technical inaccuracy. They involve learner trust in feedback, institutional use of learner data, and the potential for AI systems to reshape educational relationships. These risks are particularly important because formative assessment depends on psychological safety, learner agency, and trust between learners and educators.

One concern is epistemic overtrust. AI-generated feedback may appear authoritative because it is fluent, immediate, personalized, or numerically precise. Therefore, learners may accept AI outputs even when the underlying reasoning is incomplete, biased, or unsupported [4,6,30]. This risk is especially relevant for LLM-based feedback, where plausible explanations may not be reliably grounded in clinical evidence. Therefore, AI-assisted formative assessment activities should require learners to verify AI claims against clinical guidelines, local protocols, patient context, and faculty judgment. Verification should not be treated as an optional technical skill but as part of clinical reasoning and professional formation.

A second concern is automation bias and metric gaming. Learners may begin to optimize their performance for what the AI system measures rather than for what clinical competence requires. A procedural simulator may reward speed or economy of movement while insufficiently capturing tissue handling, situational awareness, communication, or teamwork. A virtual patient platform may reward diagnostic completeness while failing to adequately assess empathy, cultural humility, shared decision-making, or nonverbal communication. If educators do not explicitly address these limitations, AI-assisted formative

XSL•FO
RenderX

assessment may create a hidden curriculum in which machine-measurable performance is mistaken for professional competence.

A third concern involves professional identity and power. If AI feedback is positioned as more objective than human judgment, learners may experience faculty coaching as secondary or feel that their development is being monitored by systems they cannot question. This is particularly problematic when AI dashboards are visible to supervisors, when learner data are stored without clear boundaries, or when AI-generated evidence is discussed in assessment meetings without opportunities for learner explanation. Research on AI-based clinical decision support suggests that trust and professional identity are shaped by whether users can understand, contest, and contextualize AI outputs [30,31]. In educational settings, these dynamics suggest plausible downstream risks for professional identity, supervisory relationships, and assessment culture even though such effects have not yet been demonstrated directly in AI-assisted formative assessment studies [14,30,31].

A fourth concern is institutional inequity and data governance. Students with access to paid AI tools, stronger digital literacy, better language proficiency, or more supportive faculty may benefit more from AI-assisted formative assessment than their peers. Institutions also need explicit rules for patient data, learner data, vendor access, data retention, algorithmic bias monitoring, and the boundary between low-stakes coaching and high-stakes progression decisions [6,32,33]. Without such structures, AI-assisted formative assessment may widen existing inequities, compromise privacy, or undermine trust in assessment systems.

These risks are not an argument against AI-assisted formative assessment. Rather, they define the minimum safeguards for responsible use: governance, human review, equity, and explicit limits on higher-stakes use [6,7,31-33]. The next section translates these safeguards into practical design and implementation considerations.

Designing and Implementing AI-Assisted Formative Assessment in Specific Curriculum Contexts

Implementation Design and Expectations

Implementation of AI-assisted formative assessment should begin with the curriculum problem rather than with the AI tool. In practical terms, educators should first identify the learning outcome, then define the learner task, specify what the AI system will contribute, clarify the human educator's role, and require a concrete learner action at the end of the activity. This design logic is important because feedback becomes formative only when learners can compare their current performance with expected standards, identify the gap, and use that information to improve subsequent performance [2,34-36]. It is also consistent with the broader argument that AI implementation in health professions education should be guided by educational purpose, human oversight, and governance rather than by technology adoption alone [6,7,25].

Learner Routines and Action Commitments

Across curriculum contexts, a useful learner routine is to attempt, compare, critique, and act. The learner first completes the task independently so that their own reasoning, performance, or communication strategy is visible before AI feedback is introduced [35,37]. The learner then compares the response with AI-generated feedback and critiques both sources of information, which supports feedback literacy and evaluative judgment [1,13,21]. The final step is to identify one or two specific changes for the next case, encounter, or simulation because feedback has limited educational value unless it closes the loop and changes subsequent learning behavior [14,17,37].

Preparing Learners for Critical Use of AI

This learner routine requires explicit preparation. Students should be taught how to formulate prompts, judge the plausibility of AI-generated feedback, verify sources, recognize hallucinations, manage uncertainty, and translate feedback into learning plans. This preparation is particularly important for LLMs because they can produce fluent and confident responses even when outputs are incomplete, inaccurate, or unsupported [4,11]. One recent scoping review identified hallucinations and medically incorrect content as important risks in undergraduate medical education applications, underscoring the need for verification routines in educational use [12]. Learners should also be taught boundaries for acceptable AI use, especially that identifiable patient information should not be entered into unapproved systems because privacy, data security, and secondary data use are central governance risks [6,33,38].

Faculty Calibration and Debriefing

Faculty preparation should be built into the same curriculum design process rather than treated as a separate technical training exercise. Educators need sufficient AI literacy to select tools that fit learning outcomes, recognize limitations in AI outputs, and help learners interpret feedback in relation to clinical reasoning and professional values [9,25,32]. Recent work on AI competencies in medical education emphasizes skills in AI literacy, ethical awareness, data use, critical appraisal, and human oversight [39,40]. Therefore, faculty development should focus on calibration and debriefing: educators can examine an AI-generated critique of a learner's performance; identify what is accurate, what is missing, and what may be misleading; and then convert the output into coaching questions [6,7,15,41,42]. This preserves the relational function of feedback because effective feedback depends not only on information delivery but also on dialogue, interpretation, and learner uptake [20,21]. It also keeps faculty in an interpretive and coaching role rather than a supervisory role over automated scoring.

Institutional Implementation Model

Institutional preparation should be embedded before AI-assisted formative assessment is scaled across a curriculum. Institutions should specify which AI tools are approved, what data may be entered, who can access learner outputs, and how long learner data are retained. They should also clarify whether vendors may use submitted data for model training and how AI-generated evidence may be used in assessment conversations [6,7,33]. A pragmatic model is to begin with low-stakes pilots in one or

XSL•FO
RenderX

two activities, prespecify local evaluation indicators, review unintended consequences; and scale only after governance, faculty capacity, and equity safeguards are in place [6,7,25,33]. Feasibility will also depend on faculty workload for calibration and debriefing, local technical support, availability of approved tools, and willingness of faculty and learners to adopt new workflows. These constraints may justify slower scale-up or narrower pilots even when the educational rationale is strong. The boundary between formative and summative use should be explicit because programmatic assessment requires careful interpretation of evidence across multiple occasions, methods, and contexts rather than reliance on a single score or data source [16,23]. Learners should be able to question AI-generated feedback and request human review when AI-derived evidence is discussed in assessment meetings because trust in AI-supported systems depends on whether users can understand, contest, and contextualize AI outputs [14,30,31].

Equity and Scale-Up Safeguards

Equity should be treated as part of implementation design rather than as an afterthought. If AI-assisted formative assessment becomes part of the formal curriculum, institutions should provide access to approved tools instead of leaving students to purchase or select AI systems individually because unequal access to AI tools and digital infrastructure may widen educational disparities [10,32]. Institutions should also monitor whether AI feedback performs differently across learner characteristics such as language background, disability, gender, race, socioeconomic status, prior digital experience, or familiarity with prompt-based tools [7,10]. Bias monitoring and impact assessment are recommended components of responsible AI risk management [33]. Curriculum committees should also examine whether AI-assisted formative assessment activities remain aligned with learning outcomes, whether learners experience them as supportive rather than surveillant, and whether faculty have enough time to debrief outputs meaningfully because AI systems perceived as opaque or controlling may threaten trust and professional identity [10,30,31].

Focused Action Rather Than Data Accumulation

Across contexts, AI-assisted formative assessment should end with one or two specific changes for the next task or encounter rather than a long list of AI-generated observations [17,34,37]. Table 1 illustrates how this principle can be operationalized across different curriculum settings.

Finally, AI-assisted formative assessment should be evaluated based on educational outcomes rather than technology adoption metrics alone. Counts of prompts submitted, comments generated, or dashboard views do not necessarily reflect learning [9]. Before scaling, local pilots should prespecify a small evaluation set, such as learner participation, quality of action plans, change between repeated attempts, quality of dialogue between faculty and learners, learner trust, faculty time burden, equity of access, and unintended effects on assessment culture. This phased piloting approach is consistent with recent road map work emphasizing trust building, equity, and staged implementation [19]. These indicators align more closely with feedback literacy and self-regulated learning than simple measures of technology use [1].

XSL•FO
RenderX

Table . Context-specific implementation of artificial intelligence (AI)–assisted formative assessment (AFA).

Curriculum context	Curriculum problem	AFA activity and learner action	Faculty role	Institutional safeguard
Preclinical case-based diagnostic reasoning	Students may receive limited individualized feedback on diagnostic reasoning before clinical rotations.	Learners generate their own differential diagnosis and management plan, then ask an approved LLM[a] for alternatives. They classify each suggestion as accepted, modified, rejected, or uncertain and justify the decision using guidelines or local protocols.	Debrief how the AI output affected reasoning, uncertainty, and premature closure.	Use approved tools, exclude identifiable patient data, and require source checking against trusted clinical guidance.
Communication skills and OSCE[b] preparation	Students need repeated low-stakes practice in history taking, explanation, empathy, and shared decision-making.	Learners practice with an AI virtual patient, review the transcript, and identify one communication strength and one behavior to improve.	Use selected transcripts or reflections to coach empathy, uncertainty, cultural sensitivity, and patient-centered language.	Do not let AI ratings replace faculty, standardized patient, or patient ratings, especially for nonverbal behavior, professionalism, and relational care.
Procedural or surgical skill laboratory	Students need repeated practice and timely feedback on technical performance without compromising patient safety.	Simulation software or AI analytics identify efficiency, instrument path, safety, or error patterns. Learners select one metric for deliberate practice and repeat the task.	Confirm that the metric is clinically meaningful and prevent speed from being rewarded over safety.	Validate metrics locally against expert performance and patient safety priorities before using them in assessment conversations.
Clerkship and workplace-based learning	Workplace feedback may be fragmented, delayed, or difficult for learners to synthesize across supervisors and encounters.	Learners use an approved AI tool to organize reflections, compare management plans with local guidelines, or summarize recurring workplace-based assessment themes. They bring one improvement goal to coaching.	Review the learner's interpretation of AI summaries rather than forwarding unexamined outputs to committees.	Prohibit identifiable patient data in public tools and define access, consent, and storage rules for AI-generated summaries.
Programmatic assessment portfolio	Learners and coaches may struggle to identify developmental patterns across multiple low-stakes data points.	AI helps visualize patterns across practice attempts, reflections, OSCEs, and workplace-based assessment comments. Learners use these patterns to prepare a short improvement plan.	Triangulate AI-generated patterns with workplace observations, OSCE performance, portfolio evidence, and learner explanation.	Require transparent criteria, human review, and an appeal process for any higher-stakes use.

[a]LLM: large language model.

[b]OSCE: objective structured clinical examination.

Limitations

This Viewpoint is conceptual and synthetic rather than empirical. It does not report original data, follow a systematic review methodology, or test the proposed framework or implementation examples. The framework and Table 1 are intended as design heuristics to support local planning, not as validated protocols. Therefore, claims about possible effects on feedback agency, self-regulated learning, clinical reasoning, or workplace performance should be interpreted as evidence-informed possibilities rather than demonstrated outcomes in this paper. In addition, institutional feasibility will vary according to local resources, learner level, assessment culture, faculty capacity, governance, and regulatory context. For that reason, local pilots should include predefined outcomes, equity monitoring, and mechanisms for human review before wider scale-up.

Conclusions

AI-assisted formative assessment offers a meaningful opportunity to strengthen assessment for learning in clinical education, but its value depends on careful human oversight. AI can deliver immediate, scalable, and personalized feedback, yet these tools should complement rather than replace educators' professional judgment. Therefore, the framework proposed in this paper should be read as an evidence-informed guide for design and local testing rather than as proof of educational effectiveness. The future of AI-assisted formative assessment should be judged not by the sophistication of algorithms or the volume of data produced but by whether these systems help learners understand their decisions; refine their reasoning; and act on focused feedback in ways that preserve agency, professional values, and human accountability.

XSL•FO
RenderX

Acknowledgments

During revision of this manuscript, ChatGPT (OpenAI) was used to support grammar and language editing. No artificial intelligence tool was used to generate data or perform analyses. The authors reviewed, edited, and take full responsibility for the final content.

Funding

The authors declared no financial support was received for this work.

Data Availability

All data generated or analyzed during this study are included in this published article.

Authors' Contributions

QTN contributed to conceptualization, writing—original draft, project administration, and writing—review and editing. TM contributed to conceptual refinement, validation, and writing—review and editing. TMH contributed to conceptualization, validation, and writing—review and editing. All authors read and approved the final manuscript for submission.

Conflicts of Interest

None declared.

References

1. Nicol DJ, Macfarlane‐Dick D. Formative assessment and self‐regulated learning: a model and seven principles of good feedback practice. Stud High Educ 2006;31(2):199-218. [doi: 10.1080/03075070600572090]
2. Rushton A. Formative assessment: a key to deep learning? Med Teach 2005 Sep;27(6):509-513. [doi: 10.1080/01421590500129159] [Medline: 16199357]
3. Zarei M, Eftekhari Mamaghani H, Abbasi A, Hosseini MS. Application of artificial intelligence in medical education: a review of benefits, challenges, and solutions. Med Clín Práctica 2024;7(2):100422. [doi: 10.1016/j.mcpsp.2023.100422]
4. Preiksaitis C, Rose C. Opportunities, challenges, and future directions of generative artificial intelligence in medical education: scoping review. JMIR Med Educ 2023 Oct 20;9:e48785. [doi: 10.2196/48785] [Medline: 37862079]
5. Rincón EH, Jimenez D, Aguilar LA, Flórez JM, Tapia Á, Peñuela CL. Mapping the use of artificial intelligence in medical education: a scoping review. BMC Med Educ 2025 Apr 12;25(1):526. [doi: 10.1186/s12909-025-07089-8] [Medline: 40221725]
6. Masters K. Ethical use of artificial intelligence in health professions education: AMEE guide no. 158. Med Teach 2023 Jun;45(6):574-584. [doi: 10.1080/0142159X.2023.2186203] [Medline: 36912253]
7. Franco D'Souza R, Mathew M, Mishra V, Surapaneni KM. Twelve tips for addressing ethical concerns in the implementation of artificial intelligence in medical education. Med Educ Online 2024 Dec 31;29(1):2330250. [doi: 10.1080/10872981.2024.2330250] [Medline: 38566608]
8. Explanatory memorandum on the updated OECD definition of an AI system. Organisation for Economic Co-operation and Development. 2024. URL: https://www.oecd.org/en/publications/explanatory-memorandum-on-the-updated-oecd-definition-of-an-ai-system_623da898-en.html [accessed 2026-05-28]
9. Boscardin CK, Abdulnour RE, Gin BC. Macy Foundation innovation report part I: current landscape of artificial intelligence in medical education. Acad Med 2025 Sep 1;100(9S Suppl 1):S15-S21. [doi: 10.1097/ACM.0000000000006107] [Medline: 40456178]
10. Turner L, Hashimoto DA, Vasisht S, Schaye V. Demystifying AI: current state and future role in medical education assessment. Acad Med 2024 Apr 1;99(4S Suppl 1):S42-S47. [doi: 10.1097/ACM.0000000000005598] [Medline: 38166201]
11. Thirunavukarasu AJ, Ting DS, Elangovan K, Gutierrez L, Tan TF, Ting DS. Large language models in medicine. Nat Med 2023 Aug;29(8):1930-1940. [doi: 10.1038/s41591-023-02448-8] [Medline: 37460753]
12. Uzoechina G, Osajiuba T. Use of large language models in undergraduate medical education a scoping review. Discov Educ 2026;5:266. [doi: 10.1007/s44217-026-01239-w]
13. Bearman M, Tai J, Dawson P, Boud D, Ajjawi R. Developing evaluative judgement for a time of generative artificial intelligence. Assess Eval High Educ 2024;49(6):893-905. [doi: 10.1080/02602938.2024.2335321]
14. Roveta A, Castello LM, Massarino C, Francese A, Ugo F, Maconi A. Artificial intelligence in medical education: a narrative review on implementation, evaluation, and methodological challenges. AI 2025;6(9):227. [doi: 10.3390/ai6090227]
15. Mehta N, Nielsen C, Zack A, Christensen T, Isaacson JH. Creating custom GPTs for faculty development: an example using the Johari Window and Crucial Conversation frameworks for providing feedback to struggling students. Med Teach 2025 Aug;47(8):1266-1268. [doi: 10.1080/0142159X.2024.2445043] [Medline: 39787039]
16. Konopasek L, Norcini J, Krupat E. Focusing on the formative: building an assessment system aimed at student growth and development. Acad Med 2016 Nov;91(11):1492-1497. [doi: 10.1097/ACM.0000000000001171] [Medline: 27028028]

XSL•FO
RenderX

17. Ninomiya S. The possibilities and limitations of assessment for learning: exploring the theory of formative assessment and the notion of “closing the learning gap”. Educ Stud Japan 2016;10:79-91. [doi: 10.7571/esjkyoiku.10.79]
18. Brand PL, Jaarsma AD, van der Vleuten CP. Driving lesson or driving test?: A metaphor to help faculty separate feedback from assessment. Perspect Med Educ 2021 Jan;10(1):50-56. [doi: 10.1007/s40037-020-00617-w] [Medline: 32902828]
19. Rezaei-Zadeh M, Cerbin-Koczorowska M. A research roadmap for AI opportunities in student assessment for medical education. BMC Med Educ 2025 Nov 26;25(1):1653. [doi: 10.1186/s12909-025-08078-7] [Medline: 41299574]
20. Spooner M, Larkin J, Liew SC, Jaafar MH, McConkey S, Pawlikowska T. “Tell me what is 'better'!” How medical students experience feedback, through the lens of self-regulatory learning. BMC Med Educ 2023 Nov 22;23(1):895. [doi: 10.1186/s12909-023-04842-9] [Medline: 37993832]
21. Carless D, Boud D. The development of student feedback literacy: enabling uptake of feedback. Assess Eval High Educ 2018;43(8):1315-1325. [doi: 10.1080/02602938.2018.1463354]
22. Chan KS, Zary N. Applications and challenges of implementing artificial intelligence in medical education: integrative review. JMIR Med Educ 2019 Jun 15;5(1):e13930. [doi: 10.2196/13930] [Medline: 31199295]
23. van der Vleuten CP, Schuwirth LW, Driessen EW, et al. A model for programmatic assessment fit for purpose. Med Teach 2012;34(3):205-214. [doi: 10.3109/0142159X.2012.652239] [Medline: 22364452]
24. Gordon M, Daniel M, Ajiboye A, et al. A scoping review of artificial intelligence in medical education: BEME guide no. 84. Med Teach 2024 Apr;46(4):446-470. [doi: 10.1080/0142159X.2024.2314198] [Medline: 38423127]
25. Shimizu I, Kasai H, Shikino K, et al. Developing medical education curriculum reform strategies to address the impact of generative AI: qualitative study. JMIR Med Educ 2023 Nov 30;9:e53466. [doi: 10.2196/53466] [Medline: 38032695]
26. Mirchi N, Bissonnette V, Yilmaz R, Ledwos N, Winkler-Schwartz A, Del Maestro RF. The Virtual Operative Assistant: an explainable artificial intelligence tool for simulation-based training in surgery and medicine. PLoS One 2020;15(2):e0229596. [doi: 10.1371/journal.pone.0229596] [Medline: 32106247]
27. Brügge E, Ricchizzi S, Arenbeck M, et al. Large language models improve clinical decision making of medical students through patient simulation and structured feedback: a randomized controlled trial. BMC Med Educ 2024 Nov 28;24(1):1391. [doi: 10.1186/s12909-024-06399-7] [Medline: 39609823]
28. Chen Y. Evaluation of the impact of AI-driven personalized learning platform on medical students’ learning performance. Front Med (Lausanne) 2025;12:1610012. [doi: 10.3389/fmed.2025.1610012] [Medline: 41020237]
29. Bowers P, Graydon K, Ryan T, Lau JH, Tomlin D. Artificial intelligence-driven virtual patients for communication skill development in healthcare students: a scoping review. Australas J Educ Technol 2024;40(3):39-57. [doi: 10.14742/ajet.9307]
30. Tun HM, Rahman HA, Naing L, Malik OA. Trust in artificial intelligence-based clinical decision support systems among health care workers: systematic review. J Med Internet Res 2025 Jul 29;27:e69678. [doi: 10.2196/69678] [Medline: 40772775]
31. Ackerhans S, Wehkamp K, Petzina R, Dumitrescu D, Schultz C. Perceived trust and professional identity threat in AI-based clinical decision support systems: scenario-based experimental study on AI process design features. JMIR Form Res 2025 Mar 26;9:e64266. [doi: 10.2196/64266] [Medline: 40138691]
32. Pohn B, Mehnen L, Fitzek S, Choi KE, Braun RJ, Hatamikia S. Integrating artificial intelligence into pre-clinical medical education: challenges, opportunities, and recommendations. Front Educ 2025;10:1570389. [doi: 10.3389/feduc.2025.1570389]
33. Artificial intelligence risk management framework: generative artificial intelligence profile. : National Institute of Standards and Technology; 2024 URL: https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.600-1.pdf [accessed 2026-05-28]
34. Hattie J, Timperley H. The power of feedback. Rev Educ Res 2007 Mar;77(1):81-112. [doi: 10.3102/003465430298487]
35. Sadler DR. Formative assessment and the design of instructional systems. Instr Sci 1989 Jun;18:119-144. [doi: 10.1007/BF00117714]
36. Black P, Wiliam D. Assessment and classroom learning. Assess Educ Princ Policy Pract 1998 Mar;5(1):7-74. [doi: 10.1080/0969595980050102]
37. Ramani S. Reflections on feedback: closing the loop. Med Teach 2016;38(2):206-207. [doi: 10.3109/0142159X.2015.1044950] [Medline: 26030380]
38. Busch F, Adams LC, Bressem KK. Biomedical ethical aspects towards the implementation of artificial intelligence in medical education. Med Sci Educ 2023;33(4):1007-1012. [doi: 10.1007/s40670-023-01815-x] [Medline: 37546190]
39. Çalışkan SA, Demir K, Karaca O. Artificial intelligence in medical education curriculum: an e-Delphi study for competencies. PLoS One 2022;17(7):e0271872. [doi: 10.1371/journal.pone.0271872] [Medline: 35862401]
40. Gazquez-Garcia J, Sánchez-Bocanegra CL, Sevillano JL. AI in the health sector: systematic review of key skills for future health professionals. JMIR Med Educ 2025 Feb 5;11:e58161. [doi: 10.2196/58161] [Medline: 39912237]
41. Gardner A, Bilyeu C, Bazurto L, Michalsen K, Carlson C. Getting by with a little help from avatar friends: use of AI-powered avatars to implement feedback training for faculty. Med Teach 2026 Jun;48(6):941-944. [doi: 10.1080/0142159X.2025.2534073] [Medline: 40674545]
42. Hopfenbeck TN, Zhang Z, Sun SZ, Robertson P, McGrane JA. Challenges and opportunities for classroom-based formative assessment and AI: a perspective article. Front Educ 2023;8:1270700. [doi: 10.3389/feduc.2023.1270700]

XSL•FO
RenderX

Abbreviations

AI: artificial intelligence
LLM: large language model

Edited by A Stone; submitted 18.Feb.2026; peer-reviewed by J Wilkinson, JR Jaber; revised version received 08.May.2026; accepted 08.May.2026; published 04.Jun.2026.

Please cite as:
Nguyen QT, Ha TM, Mai T
AI-Assisted Formative Assessment in Clinical Education: From Algorithms to Agency
JMIR Med Educ 2026;12:e93710
URL: https://mededu.jmir.org/2026/1/e93710
doi:10.2196/93710

© Quang Thanh Nguyen, Thuy Minh Ha, Tina Mai. Originally published in JMIR Medical Education (https://mededu.jmir.org), 4.Jun.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Education, is properly cited. The complete bibliographic information, a link to the original publication on https://mededu.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

AI in UK Medical Education: A Framework for Curriculum Reform

Aditya Gaur[1], MBBS; Joecelyn Kirani Tan[2], BSc; Medha Sridhar Rao[3], MB BCh, BAO; Muhtasim Fuad[2], MBChB; Taha Bhatti[2], MBChB; Hareesha Rishab Bharadwaj[2], MBChB; Khabab Abbasher Hussien Mohamed Ahmed[4], MBBS

[1]Yeovil District Hospital, Somerset NHS Foundation Trust, Yeovil, United Kingdom

[2]Faculty of Biology, Medicine, and Health, University of Manchester, Oxford Rd, Manchester, United Kingdom

[3]School of Medicine, Dentistry and Biomedical Sciences, Queen's University Belfast, Belfast, United Kingdom

[4]Faculty of Medicine, University of Khartoum, Khartoum, Sudan

Corresponding Author:
Joecelyn Kirani Tan, BSc
Faculty of Biology, Medicine, and Health, University of Manchester, Oxford Rd, Manchester, United Kingdom

Abstract

Artificial intelligence (AI) is increasingly transforming health care through improvements in diagnosis, predictive analytics, and workflow optimization. However, there remains a significant gap in AI training within UK medical education, leaving future clinicians underprepared for AI-driven health care environments. This viewpoint paper investigated global best practices for AI integration into medical education and proposes a structured framework for embedding AI into the UK medical curriculum. It aimed to assess current attitudes, highlight existing knowledge gaps, and recommend practical implementation strategies. An analysis of international case studies (eg, Stanford University, the University of Toronto, and Chinese University of Hong Kong) was conducted alongside a review of teaching methodologies, stakeholder perspectives, and UK-based surveys to identify core competencies and challenges in AI education. Effective integration strategies include the use of AI-powered simulations, interdisciplinary collaboration, elective modules, and faculty training. Major barriers include lack of AI-literate educators, insufficient ethical training, and limited infrastructure. Knowledge gaps persist among students and faculty in areas such as algorithmic bias, AI ethics, and clinical decision-making. To meet the demands of modern health care, the UK medical curriculum must adopt comprehensive AI training. This includes practical exposure, ethical awareness, and stakeholder engagement. Proactive reform will ensure that graduates are equipped to critically and ethically apply AI tools in clinical practice.

(JMIR Med Educ 2026;12:e81953) doi:10.2196/81953

KEYWORDS

artificial intelligence; AI; medical education; curriculum reform; artificial intelligence literacy; AI literacy; UK health care; medical students; interdisciplinary training; digital health; artificial intelligence ethics; AI ethics; simulation-based learning

Introduction

Artificial intelligence (AI) is increasingly embedded in diagnostic, administrative, and clinical decision support systems across health care settings [1,2]. From radiological image analysis to predictive risk stratification and generative documentation tools, AI technologies are reshaping how clinical information is processed and applied. As these systems become more integrated into routine practice, medical education must evolve to ensure that graduates can critically appraise, interpret, and responsibly use AI-enabled tools in patient care [3,4].

Despite AI's rapid expansion in health care, structured AI education remains limited within UK medical training programs. A recent national survey reported that 88% (432/484) of UK medical students recognized the importance of AI in health care yet only a minority had received formal instruction, highlighting a significant educational gap [5]. This shortfall extends into postgraduate training: in another UK-based survey, 92% (193/210) of trainees felt that AI training in their current curricula was insufficient, and 81% (170/210) of trainees emphasized the necessity of formal AI education [6]. These findings suggest a growing disconnect between technological advancement and educational provision.

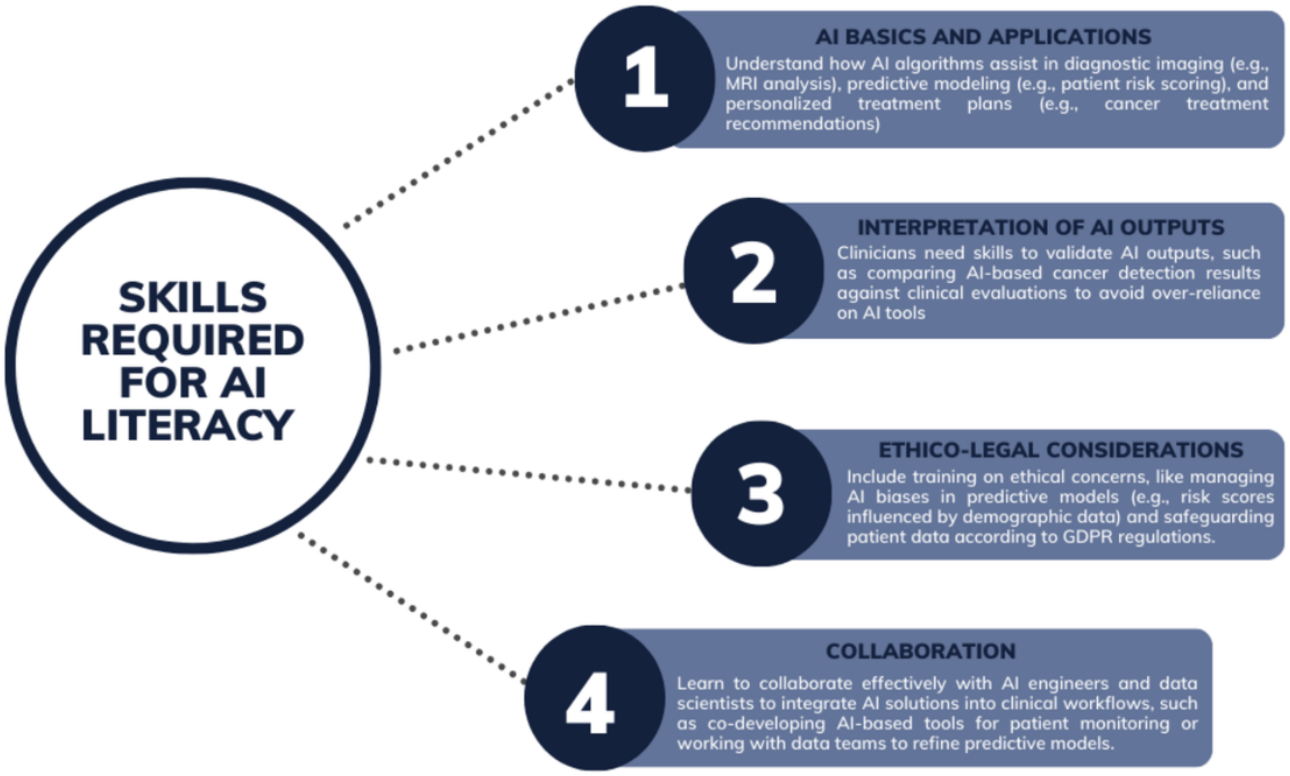
Addressing this gap requires more than informal exposure to digital tools. It demands the development of AI literacy, defined in this viewpoint paper as the foundational knowledge, critical appraisal skills, and ethical understanding required to interpret AI outputs, recognize their limitations, communicate their implications to patients, and maintain professional accountability when using AI-supported systems. Therefore, AI literacy encompasses not only technical awareness of concepts such as machine learning and data bias but also the professional judgment necessary to integrate AI safely into clinical reasoning.

Research consistently highlights the importance of equipping future clinicians with competencies that enable safe and effective engagement with AI technologies. These competencies, outlined in Figure 1, include foundational technical knowledge, ethical awareness, critical evaluation skills, and applied clinical

XSL•FO
RenderX

judgment [7]. However, while numerous publications have discussed AI applications in health care and emerging educational initiatives, few have provided a structured, UK-specific synthesis that integrates international models, regulatory considerations, and competency-based curriculum reform into a unified framework. Existing literature often examines individual teaching interventions or isolated institutional initiatives without systematically analyzing their transferability to centrally regulated systems such as the United Kingdom.

Figure 1. Essential competencies for equipping health care providers for an artificial intelligence (AI)–driven future. GDPR: General Data Protection Regulation; MRI: magnetic resonance imaging.



This viewpoint paper addresses that gap by synthesizing current evidence and proposing a structured, context-sensitive framework tailored to the UK medical education system. By evaluating existing educational models, regulatory requirements, and learner perspectives, this viewpoint paper aimed to present a practical road map for developing an AI-literate medical workforce capable of navigating future technological advancements in health care.

Search Strategy

The review performed in this viewpoint paper was conducted as a structured narrative synthesis aimed at identifying existing approaches to AI integration in undergraduate medical education and evaluating their relevance to the UK context. A defined literature selection process with explicit inclusion and exclusion criteria was applied to enhance methodological transparency and rigor.

Literature searches were conducted across multiple major databases, including PubMed and MEDLINE, Embase, the Cochrane Library, and Scopus. Searches were limited to full-text articles published in English between 2015 and 2025 to reflect contemporary developments in AI technologies. A combination of keywords and Boolean operators was used, including "artificial intelligence," "AI," "machine learning," "deep learning," "medical education," "curriculum," "competency," and "medical students." The reference lists of recent review articles were manually screened to identify additional relevant studies.

Studies were included if they (1) examined AI applications within undergraduate or postgraduate medical education; (2) described curricular design, teaching methodology, competency development, or student attitudes; or (3) reported empirical findings, implementation strategies, or institutional case examples. Studies were excluded if they were (1) conference abstracts or proceedings, (2) editorials, (3) case reports unrelated to educational structure, or (4) unpublished. This approach prioritized peer-reviewed evidence describing implemented or evaluated educational initiatives.

International case studies were selected from included studies where institutions demonstrated structured curricular implementation of AI beyond isolated workshops, with clear description of teaching models and institutional support. These examples were synthesized to represent distinct integration models (elective-based, postgraduate specialist and longitudinally embedded approaches) rather than designate a single definitive "best practice." For the purposes of this viewpoint paper, best practice was defined as initiatives demonstrating interdisciplinary collaboration, sustained curricular integration, ethical governance consideration, and scalability within institutional constraints. UK-based surveys were identified using additional search terms related to UK

XSL•FO
RenderX

medical students and AI attitudes, and findings were synthesized narratively to identify recurring gaps in preparedness and training demand. Proposed competency domains were derived through thematic identification of commonly reported educational priorities across included studies and were subsequently mapped conceptually to the outcome framework of the General Medical Council (GMC) to ensure regulatory alignment.

The following sections synthesize findings from the existing literature regarding current gaps, teaching methodologies, barriers, and international models of AI integration in medical education.

Current Gaps in AI Literacy

For AI to be effectively incorporated into health care education, existing gaps in AI literacy and the scope of AI applications must be addressed. The main gaps stem from a lack of access to AI tools designed for educational use, limited awareness of AI capabilities, fears surrounding AI-generated academic content, inaccuracies in AI outputs, and a shortage of knowledgeable educators. Additionally, there is currently no ethical framework from the GMC or the National Health Service (NHS) guiding AI use in UK medical education. While Health Education England (HEE) developed the Digital, Artificial Intelligence, and Robotics Technologies in Education program [8], it is not widely implemented in UK medical schools. However, the Artificial Intelligence and Digital Healthcare Technologies Capability framework developed by HEE reinforces the national goal of enhancing AI capabilities among health care professionals [8].

Understanding the attitudes of health care professionals and students toward AI is instrumental in enhancing its use in medical practice. Tools such as ChatGPT can aid education by generating questions, flash cards, and writing assistance, enabling students and educators to focus more on learning. However, there is increasing concern about the misuse of AI for academic writing, where students may bypass critical thinking by generating entire essays [9]. This threatens academic integrity, prompting institutions to balance the benefits of AI with safeguards against misuse.

To integrate AI effectively, its role in health care must be aligned with literacy objectives in curricula. Applications such as machine learning for imaging diagnostics, medical administration, drug discovery, and digital twins [10] illustrate AI's scope.

Despite these knowledge gaps, medical students generally express optimism about AI's potential to enhance patient outcomes, streamline clinical workflows, and reduce diagnostic errors [11]. Nevertheless, skepticism remains regarding the reliability and transparency of AI systems, with concerns over the risk of overreliance on AI and its potential to perpetuate biases [12].

Faculty perceptions of AI integration appear more mixed than student attitudes. Some faculty members have expressed concern about the potential for AI to replace aspects of traditional medical roles, whereas students are more likely to focus on whether their education adequately prepares them for practice in an AI-enabled health care system [13]. UK survey evidence supports this distinction: while students overwhelmingly recognize the future importance of AI in medicine, they report low levels of confidence in using AI tools in clinical contexts [5].

In contrast, there is currently limited UK-specific empirical evidence examining faculty preparedness or institutional infrastructure gaps. However, international literature suggests that successful integration of AI into medical education depends heavily on educator capability and institutional capacity. A recent global review (January 2026) reported that students across regions consistently favored interactive, practice-based learning over traditional didactic lectures and concluded that meaningful integration requires sustained investment in faculty development and structured curricula addressing both technical and ethical competencies [14].

Similarly, Rani et al [15] found that, although 91% (n=273) of the faculty members surveyed agreed that AI has the potential to enhance medical education, only 12% (n=36) described themselves as "very familiar" with AI technologies. This discrepancy highlights a recurrent international theme: conceptual support for AI integration does not necessarily translate into educator readiness.

Taken together, while UK-specific data on faculty capability and infrastructure remain limited, converging international evidence indicates that AI-literate educators, structured faculty development programs, and accessible technological infrastructure are likely prerequisites for sustainable curriculum reform.

Teaching Approaches in AI Education

Overview

Medical education continues to evolve with diverse teaching methodologies aimed at equipping students with the theoretical knowledge and practical skills essential for patient care. As AI becomes increasingly integrated into health care, new approaches are required to ensure that future clinicians are proficient in AI applications. This section explores key teaching methodologies and highlights the existing knowledge gaps and attitudes in preparing medical students for AI-driven clinical environments.

Hands-On Laboratories

Hands-on laboratories remain a critical component of medical education, providing students with direct experience in anatomical dissection, surgical techniques, and diagnostic procedures. These laboratories enhance learning through experiential engagement, reinforcing conceptual understanding while developing essential clinical skills such as dexterity, coordination, and procedural proficiency [16,17]. However, hands-on training requires substantial resources, including specialized facilities, equipment, and skilled instructors, making it expensive and logistically demanding. Additionally, while these laboratories provide valuable technical training, they do not expose students to the AI-driven tools increasingly used in diagnostics and decision-making [18].

XSL•FO
RenderX

AI Simulation

AI-powered simulation refers to digitally mediated training environments that use AI algorithms such as machine learning, natural language processing, and adaptive analytics to generate dynamic, responsive clinical scenarios that evolve in real time based on learner input. Unlike traditional static simulations, AI-powered systems can modify patient presentations, diagnostic data, and treatment responses according to a student's decisions, thereby creating an interactive and personalized learning experience.

AI-powered simulations have become an increasingly valuable tool in medical training, offering students an adaptive environment for clinical decision-making. These systems replicate patient interactions, diagnostic challenges, and treatment planning scenarios, enabling learners to practice clinical reasoning without real-world consequences. AI-driven simulations are particularly beneficial for rare or complex medical cases to which exposure during routine placements may be limited. By providing immediate, data-informed feedback and iterative practice opportunities, such platforms can support the refinement of diagnostic judgment and therapeutic planning skills [19,20].

Despite these advantages, the development and maintenance of high-quality AI simulations remain costly and technically demanding. Infrastructure requirements, data governance considerations, and the need for continual algorithm validation present significant institutional challenges. Moreover, while simulations can enhance learning, they cannot fully replicate the unpredictability, emotional nuance, and contextual complexity of real-world patient care. Therefore, there is a risk that overreliance on simulated environments may limit exposure to authentic clinical uncertainty, reinforcing the need for AI simulation to complement rather than replace supervised clinical experience [21].

Case-Based Learning

Case-based learning is widely used in medical education to encourage students to apply theoretical knowledge to real or hypothetical patient scenarios. This method promotes critical thinking, diagnostic reasoning, and problem-solving skills, facilitating a deeper understanding of clinical practice [22,23]. In the context of AI education, case studies can be adapted to include AI-assisted diagnostics and treatment recommendations, allowing students to critically assess machine-generated insights. However, the effectiveness of this approach depends on the quality and relevance of the cases presented, as well as the expertise of educators in guiding discussions. Some students may struggle with the complexity of AI-integrated cases, leading to varied learning outcomes [1].

Online Learning Platforms

With the rise of digital education platforms, online learning has become an increasingly important component of medical training. Online resources, including recorded lectures, interactive modules, and virtual discussions, provide students with flexible, self-paced learning opportunities. This approach is particularly useful for foundational AI education, allowing students to gain exposure to key concepts such as machine learning and natural language processing [11,24]. However, a major limitation of online learning is the lack of direct, hands-on engagement, which is crucial for developing clinical and AI-related competencies. Additionally, online learning requires strong self-discipline and motivation, and it may not fully replace the collaborative and interactive aspects of in-person education [21].

Knowledge Gaps in AI Competency

As AI technologies become increasingly integrated into health care, it is essential to assess the knowledge gaps and attitudes of medical students and faculty regarding AI. Research suggests that many medical students and faculty possess only a superficial understanding of AI and its applications in clinical practice [25,26]. Fundamental concepts such as machine learning, data analytics, and algorithmic bias remain largely unfamiliar [26].

Beyond technical knowledge, there is also a limited understanding of the ethical and legal implications of AI in health care, including concerns about data privacy, informed consent, and algorithmic transparency [27]. Uncertainty persists regarding AI's role in clinical decision-making and patient outcomes, particularly in recognizing its strengths and limitations and determining when AI-generated insights should be integrated into patient care [28].

Barriers to Integration

Despite the potential benefits of AI integration in medical education, several challenges hinder its implementation. One of the primary obstacles is the lack of a structured AI curriculum in most UK medical schools. Medical education traditionally prioritizes biological sciences, human anatomy, physiology, and clinical procedural skills, with limited emphasis on AI-driven domains such as machine learning and data science. Therefore, introducing AI comprehensively into medical training would require curriculum restructuring, protected teaching time, additional resources, and specialized faculty expertise—elements that remain limited within many UK institutions [8].

Ethical considerations pose further challenges, particularly in relation to patient autonomy, data privacy, algorithmic bias, and accountability in AI-assisted decision-making. Maintaining patient confidentiality is a core principle of medical practice, yet AI systems often rely on large-scale data processing that may not fully align with existing ethical and legal standards. Medical students must be equipped to navigate these complexities; however, AI ethics training remains inconsistently embedded in UK medical curricula [29,30]. The integration of structured ethics education into AI training is further constrained by already dense curricula, potentially leaving graduates underprepared to critically evaluate AI's implications for patient rights and health care equity.

These concerns extend beyond education alone. The 2025 FUTURE-AI consortium, an international consensus involving 117 experts across 50 countries and published in *The BMJ*, emphasized that "trustworthy AI" in health care requires technical robustness, transparency, validation, and traceability throughout the system life cycle [31]. Such standards imply the need for secure data environments, governance frameworks,

XSL•FO
RenderX

and systems capable of auditing algorithmic outputs. Many current medical school infrastructures are not configured to support this level of technical integration, underscoring the gap between aspirational AI literacy and the operational requirements of trustworthy clinical AI deployment.

Another barrier is the shortage of AI-proficient faculty. Effective AI education requires instructors with expertise spanning both clinical medicine and AI applications, yet relatively few clinicians possess formal training in data science or machine learning. Recruiting or retraining faculty capable of delivering AI instruction is both time intensive and costly [29,30]. In addition, unfamiliarity with AI concepts may contribute to hesitancy among existing educators, potentially slowing curriculum reform.

These workforce limitations are compounded by financial constraints. The development of AI-focused programs requires sustained investment in computational infrastructure, secure data storage, specialized software, and ongoing system maintenance. For many UK medical schools operating within restricted budgets, allocating long-term funding for such initiatives is challenging [8].

Taken together, these barriers extend beyond short-term logistical concerns. They reflect a broader structural misalignment: effective AI education demands interdisciplinary expertise, governance mechanisms, and technological capacity that traditional medical faculties were not originally designed to support. Without strategic investment, regulatory clarity, and cross-department collaboration, isolated training efforts are unlikely to achieve scalable and sustainable integration.

International Models of AI Integration in Medical Education

International approaches to AI integration in medical education can be synthesized into three broad curricular models: (1) elective-specialist pathways, (2) postgraduate fellowship expansion, and (3) fully embedded longitudinal curricula. These models differ in scope, scalability, and structural demands, offering distinct implications for centrally regulated systems such as the United Kingdom.

Elective-Specialist Pathways

Elective-based models introduce AI education through optional modules or symposia that allow interested students to develop specialized competencies. For example, Stanford University incorporates AI through dedicated electives and its Artificial Intelligence in Medicine and Imaging Symposium, which emphasizes experiential learning and practical application in areas such as medical imaging [32].

This model is flexible and relatively low risk as it does not require wholesale curriculum restructuring. However, its scalability is limited. Elective pathways attract students already predisposed toward technology-focused careers, potentially reinforcing inequities in AI literacy across the wider student body [32]. While effective in cultivating future clinician innovators, this model may not ensure baseline AI competency among all graduates.

For systems such as the United Kingdom, where curricular time is tightly regulated, elective integration may represent a feasible entry point. However, its reliance on student self-selection raises concerns regarding universal preparedness.

Postgraduate Fellowship Expansion

A second model focuses on postgraduate specialization through structured fellowships. At Mayo Clinic, AI-focused fellowships target trainees in areas such as diagnostic imaging and personalized medicine, promoting advanced technical expertise and continuous professional development [32].

This approach strengthens workforce capacity by developing clinician leaders capable of driving AI innovation within health care systems. It is particularly scalable in decentralized systems where institutional autonomy enables specialized program development.

However, as a postgraduate strategy, it does not address foundational AI literacy at the undergraduate level. In centralized systems such as the NHS, fellowship-based expansion could enhance specialist expertise but would not alone ensure system-wide competency among newly qualified physicians.

Fully Embedded Longitudinal Curricula

The most comprehensive model integrates AI across undergraduate and postgraduate training in a longitudinal manner. At the University of Toronto, AI training is embedded within existing curricula through interdisciplinary collaboration with computer science departments [33]. Similarly, the Chinese University of Hong Kong (CUHK) has implemented AI-powered clinical learning tools, invested in digital laboratory infrastructure, and prioritized faculty development [34].

These models emphasize interdisciplinary learning, ethical reasoning, and sustained exposure to AI applications throughout training. Compared to elective or fellowship approaches, embedded curricula promote equity by ensuring that all students develop baseline AI competency [33,34].

However, this model is structurally demanding. It requires institutional autonomy, protected curriculum time, technological infrastructure, and significant faculty development. In highly centralized and tightly regulated systems, implementation may necessitate coordinated national reform rather than isolated institutional initiatives.

Structural Constraints and System-Level Implications

While prior literature has primarily documented innovative AI initiatives within individual institutions, less attention has been paid to how such models translate to centrally regulated and publicly funded health care systems. This viewpoint paper moves beyond descriptive comparison to critically evaluate the structural assumptions underlying international models and identify gaps in scalability, equity, and regulatory alignment. By situating AI curriculum reform within the governance and funding realities of UK medical education, this paper advances

XSL•FO
RenderX

a system-level framework tailored to a nationally standardized training environment.

International examples illustrate 3 broad models of AI integration: elective-based exposure (eg, Stanford University), postgraduate fellowship pathways (eg, Mayo Clinic), and fully embedded longitudinal curricula (eg, University of Toronto and CUHK) [32-34]. While these differ in intensity, they share enabling conditions: institutional autonomy, protected curricular time, and interdisciplinary faculty capacity. The United Kingdom differs structurally in 2 critical respects.

First, medical education is centrally regulated by the GMC. Therefore, any substantive curricular reform must align with nationally defined graduate outcomes. Unlike more decentralized systems, particularly in the United States, UK institutions cannot independently implement large-scale curriculum redesign without regulatory alignment. This regulatory structure promotes standardization but may slow rapid innovation.

Second, the NHS operates as a publicly funded, centrally coordinated health care system. While this offers opportunities for national alignment, it also constrains financial flexibility. AI-driven educational initiatives require investment in digital infrastructure, simulation platforms, secure data environments, and faculty training. For many UK medical schools operating within fixed budgets and increasing financial pressure, allocating new resources without external funding support may be challenging.

These structural realities suggest that comprehensive, fully embedded AI curricula, as seen in some international institutions, may not be immediately feasible in the United Kingdom without national funding strategies or cross-institutional resource sharing.

Cost and Implementation Challenges

Beyond high-level structural considerations, practical implementation raises additional challenges.

AI education requires more than curricular time; it demands computational infrastructure, secure data storage, software licensing, and ongoing system maintenance. These recurring costs extend beyond initial development and must be sustained in the long term. In publicly funded institutions facing competing priorities, opportunity costs become a significant consideration: expanding AI content may require the displacement of existing material or additional teaching hours. The NHS Long Term Workforce Plan [35] committed £2.4 billion (US $3.3 billion) to additional training places, but the British Medical Association has noted that funding for simulated learning and AI-ready infrastructure is unevenly distributed, risking a digital divide between older analog hospitals and new digital flagship trusts [36].

Faculty capacity presents an equally significant barrier. AI education requires interdisciplinary expertise spanning data science, ethics, informatics, and clinical medicine—competencies not traditionally embedded within medical faculties. Recruiting new staff or retraining existing faculty entails both financial cost and protected development time. Without clear funding pathways or national coordination, implementation risks are uneven across institutions, potentially exacerbating interschool disparities.

Moreover, variation in curricular structure across UK medical schools introduces logistical complexity. Integration must be sufficiently flexible to accommodate problem-based, system-based, and traditional models of teaching without compromising consistency in learning outcomes.

Taken together, these considerations reinforce the need for incremental and scalable reform rather than immediate wholesale integration.

Depth, Equity, and Governance

A continuum emerges across international approaches. Elective models maximize flexibility but risk uneven competency distribution. Fellowship pathways cultivate advanced expertise but do not ensure universal literacy. Embedded models promote equity and standardization but require substantial structural investment.

For the United Kingdom, this continuum highlights a central policy tension: balancing depth of expertise with equitable baseline literacy under constrained resources. If AI training remains elective, disparities in confidence and engagement, including gender-based differences reported in European cohorts, may widen [37]. Conversely, universal embedding without adequate faculty preparation risks superficial implementation.

Ethical governance further complicates integration. Institutions must establish clear policies distinguishing legitimate AI-assisted learning from academic misconduct as generative tools become increasingly accessible [38]. Without explicit guidance, ambiguity may undermine professional standards and public trust. Therefore, embedding ethical reasoning and responsible AI use within curricula is not optional but foundational.

The NHS HEE AI Lab provides a relevant example of centrally coordinated workforce development in digital health. Its model demonstrates how national infrastructure can support scalable AI literacy and structured training pathways across the health care system [18]. However, translating such initiatives into undergraduate medical curricula would require careful adaptation, sustainable funding mechanisms, and explicit alignment with the GMC competency framework.

Regulatory expectations further reinforce the urgency of curricular reform. The *Good Medical Practice* update from the GMC emphasizes that physicians should remain professionally accountable for all clinical decisions, including those informed by AI systems [39]. This position is reflected in professional attitudes: a 2024 survey conducted by the Alan Turing Institute supported by the GMC found that, although many physicians report using AI tools, 99% (814/822) would prioritize their own clinical judgment over an AI recommendation in cases of conflict [40].

Despite growing enthusiasm for AI integration, several gaps remain insufficiently addressed in the existing literature. First,

XSL•FO
RenderX

few studies examine long-term sustainability or cost-effectiveness beyond initial implementation. Second, there is limited evaluation of competency outcomes, particularly whether exposure translates into measurable clinical decision-making proficiency. Third, the regulatory implications of AI competency standards in nationally governed systems remain underexplored. Finally, most initiatives focus on technologically advanced institutions, leaving questions about equity of access across diverse educational settings. These unresolved issues are particularly salient in the UK context and underscore the need for a coordinated, evaluative, and economically conscious integration strategy.

A Phased Framework for UK Integration

Overview

In light of these feasibility, cost, regulatory, and workforce considerations, a phased approach is most appropriate for the UK context. Reform should begin with a comprehensive needs assessment to establish baseline attitudes, knowledge gaps, institutional readiness, and available faculty capacity. Survey evidence from UK medical students demonstrates strong recognition (432/484, 88%) of AI's future importance alongside low levels of confidence in applying AI tools, underscoring the need for structured intervention rather than informal exposure [5]. As of 2025, a total of 88% (916/1041) of UK undergraduate students report using generative AI tools (up from 53% in 2024), yet only 36% (375/1041) report receiving formal institutional training [41]. Therefore, initial assessments must evaluate not only learner demand but also financial viability, digital infrastructure capacity, and alignment with existing program outcomes defined by the GMC.

Mapping AI Competencies to GMC Outcomes

To ensure regulatory coherence, AI competencies should be explicitly mapped to the GMC's outcomes for graduates domains: (1) professional values and behaviors: understanding ethical AI use, data governance, algorithmic bias, and responsible integration into patient care; (2) professional skills: critically appraising AI-generated outputs, communicating AI-informed decisions to patients, and maintaining clinical accountability when AI recommendations conflict with professional judgment; and (3) professional knowledge: foundational literacy in machine learning principles, data limitations, validation processes, and real-world clinical applications.

Rather than positioning AI as a discrete subject, embedding competencies within these existing domains reduces curricular displacement while reinforcing that AI literacy is integral to safe, modern clinical practice.

Curriculum Design and Module Structure

Within a pilot module or integrated teaching block, illustrative learning objectives may include (1) describing the basic principles underlying machine learning and clinical decision support systems, (2) critically evaluating the strengths and limitations of AI tools in imaging or diagnostic contexts, (3) identifying ethical risks associated with algorithmic bias and data misuse, and (4) demonstrating appropriate professional judgment when AI outputs conflict with clinical reasoning.

A feasible structure could involve a blended model combining (1) introductory lectures (foundational concepts), (2) case-based seminars integrating AI into clinical scenarios, (3) simulation-based exercises using AI-assisted diagnostic tools, and (4) interdisciplinary workshops codelivered with data science faculty.

Such integration mirrors approaches observed in international embedded longitudinal models, including those at the University of Toronto and CUHK [33,34], while adapting them to the structural constraints of UK programs.

Assessment Strategies

Assessment should align with existing formats to maintain feasibility. Potential approaches include (1) incorporating AI critical appraisal questions into written examinations, (2) objective structured clinical examination stations requiring interpretation of AI-assisted outputs, (3) reflective essays on ethical AI dilemmas, and (4) workplace-based assessments evaluating appropriate AI integration in simulated scenarios.

Embedding AI evaluation within established assessment structures avoids the need for parallel examination systems while reinforcing accountability and patient safety principles.

Faculty Development and Infrastructure Requirements

Faculty development is central to sustainable implementation. Institutions may adopt a tiered approach: (1) foundational AI literacy workshops for all teaching staff, (2) advanced training for designated AI curriculum leads, and (3) cross-department collaboration with computer science or informatics departments.

Minimum infrastructure requirements include secure digital platforms for simulation exercises, access to validated AI demonstration tools, and data governance protocols consistent with NHS standards. National coordination through NHS-linked digital education initiatives may reduce duplication and promote equitable access across institutions.

Progressive Integration

Early pilot programs such as student-selected components or elective modules should serve as evaluation platforms to quantify resource requirements, faculty workload implications, and cost sustainability. As institutional capacity develops, competencies can be progressively embedded within core modules such as radiology, diagnostics, ethics, and clinical decision-making.

Continuous evaluation should extend beyond assessment modalities to clearly defined outcome indicators. Institutions should establish baseline AI literacy and faculty readiness metrics prior to implementation followed by structured pre- and postintervention comparisons. Suggested indicators include validated AI literacy or digital health competency scales (where available), performance in AI-integrated objective structured clinical examination stations, case-based ethics assessments examining bias recognition, learner confidence surveys, and faculty self-reported preparedness following development programs. A phased evaluation timeline—baseline (before

XSL•FO
RenderX

implementation), pilot year assessment, and longitudinal review at 2 to 3 years—would enable quantification of competency attainment, cost-effectiveness, and sustainability. Defining these metrics explicitly ensures that AI integration generates evaluable evidence rather than remaining a conceptual reform initiative. The phased framework for AI integration in UK undergraduate medical education is summarized in Table 1.

XSL•FO
RenderX

Table . Phased framework for artificial intelligence (AI) integration in UK undergraduate medical education.

Phase	Key components	Practical strategies	Evaluation metrics	References
Needs assessment and institutional readiness	Establish baseline learner knowledge, attitudes, faculty capacity, and infrastructure	• Survey AI literacy and confidence among students • Assess faculty readiness and expertise • Review digital infrastructure and financial feasibility • Align with GMC[a] outcomes for graduates	• Baseline AI literacy scores • Student confidence surveys • Faculty preparedness metrics • Infrastructure audit	[5,41]
Competency mapping to GMC outcomes	Establish baseline learner knowledge, attitudes, faculty capacity, and infrastructure	• Map AI competencies to professional values and behaviors (ethics, bias, and governance), professional skills (AI appraisal, communication, and accountability), and professional knowledge (ML[b] principles, validation, and limitations)	• Explicit curriculum mapping • Regulatory alignment review	GMC outcomes for graduates
Curriculum design and delivery	Embed AI into existing modules rather than as a stand-alone subject	• Introductory lectures (foundations) • Case-based seminars • AI-assisted simulation exercises • Interdisciplinary workshops with data science faculty	• Module participation rates • Learner feedback • Objective knowledge assessments	[33,34]
Assessment integration	Align AI evaluation with existing assessment structures	• AI critical appraisal questions in written exams • AI-integrated OSCE[c] stations • Reflective essays on ethics and bias • Workplace-based assessments	• Performance in AI-specific exam items • OSCE scoring outcomes • Ethics case analysis performance	[7]
Faculty development and infrastructure	Develop sustainable teaching capacity	• Tiered AI literacy workshops • Advanced training for curriculum leads • Collaboration with informatics departments • Secure simulation platforms and validated tools	• Faculty training completion rates • Faculty confidence surveys • Infrastructure compliance with NHS[d] standards	[18]
Progressive integration and longitudinal evaluation	Scale from pilot modules to core curriculum	• Begin with electives or student-selected components • Expand into radiology, diagnostics, ethics, and clinical reasoning • Establish phased evaluation timeline (baseline, year 1 pilot, and 2- or 3-year review)	• Pretest-posttest AI literacy comparisons • AI-integrated OSCE performance trends • Learner confidence shifts • Faculty readiness improvements • Cost-effectiveness analysis	[5,41]

[a]GMC: General Medical Council.

XSL•FO
RenderX

[b]ML: machine learning.

[c]OSCE: objective structured clinical examination.

[d]NHS: National Health Service.

Conclusions

AI is no longer a speculative addition to health care but an operational reality shaping diagnosis, decision-making, and system design. However, UK medical education has not evolved at a comparable pace. This viewpoint paper demonstrates that, while international institutions have piloted diverse models of AI integration, sustainable reform within the United Kingdom requires a regulatory-aligned, economically conscious, and evaluative approach. Developing AI literacy must extend beyond technical familiarity to encompass ethical reasoning, critical appraisal, and professional accountability in AI-supported clinical environments. A phased, competency-mapped framework grounded in GMC standards, supported by faculty development, and reinforced through measurable outcome indicators offers a pragmatic pathway forward. Without structured integration, disparities in preparedness may widen, and graduates risk entering practice underequipped for an AI-enabled NHS. Therefore, proactive, coordinated curriculum reform is not simply an educational innovation but a professional imperative to safeguard patient care, uphold public trust, and ensure that future physicians remain critically engaged leaders in an increasingly intelligent health care system.

Funding

This study received no funding or financial support.

Conflicts of Interest

None declared.

References

1. Buch VH, Ahmed I, Maruthappu M. Artificial intelligence in medicine: current trends and future possibilities. Br J Gen Pract 2018 Mar;68(668):143-144. [doi: 10.3399/bjgp18X695213] [Medline: 29472224]
2. Jiang F, Jiang Y, Zhi H, et al. Artificial intelligence in healthcare: past, present and future. Stroke Vasc Neurol 2017;2(4):230-243. [doi: 10.1136/svn-2017-000101] [Medline: 29507784]
3. Secinaro S, Calandra D, Secinaro A, Muthurangu V, Biancone P. The role of artificial intelligence in healthcare: a structured literature review. BMC Med Inform Decis Mak 2021 Apr 10;21(1):125. [doi: 10.1186/s12911-021-01488-9] [Medline: 33836752]
4. Davenport T, Kalakota R. The potential for artificial intelligence in healthcare. Future Healthc J 2019 Jun;6(2):94-98. [doi: 10.7861/futurehosp.6-2-94] [Medline: 31363513]
5. Sit C, Srinivasan R, Amlani A, et al. Attitudes and perceptions of UK medical students towards artificial intelligence and radiology: a multicentre survey. Insights Imaging 2020 Feb 5;11(1):14. [doi: 10.1186/s13244-019-0830-7] [Medline: 32025951]
6. Banerjee M, Chiew D, Patel KT, et al. The impact of artificial intelligence on clinical education: perceptions of postgraduate trainee doctors in London (UK) and recommendations for trainers. BMC Med Educ 2021 Aug 14;21(1):429. [doi: 10.1186/s12909-021-02870-x] [Medline: 34391424]
7. Lee J, Wu AS, Li D, Kulasegaram KM. Artificial intelligence in undergraduate medical education: a scoping review. Acad Med 2021 Nov 1;96(11S):S62-S70. [doi: 10.1097/ACM.0000000000004291] [Medline: 34348374]
8. Artificial intelligence (AI) and digital healthcare technologies capability framework. NHS England. URL: https://digital-transformation.hee.nhs.uk/support-for-organisations/research-and-publications/dart-ed/horizon-scanning/ai-and-digital-healthcare-technologies [accessed 2026-05-12]
9. Doyal AS, Sender D, Nanda M, Serrano RA. ChatGPT and artificial intelligence in medical writing: concerns and ethical considerations. Cureus 2023 Aug;15(8):e43292. [doi: 10.7759/cureus.43292] [Medline: 37692694]
10. Vallée A. Digital twin for healthcare systems. Front Digit Health 2023;5:1253050. [doi: 10.3389/fdgth.2023.1253050] [Medline: 37744683]
11. Sassis L, Kefala-Karli P, Sassi M, Zervides C. Exploring medical students' and faculty's perception on artificial intelligence and robotics. A questionnaire survey. Artif Intell Med Sci 2021;2(1-2):76-84. [doi: 10.2991/jaims.d.210617.002]
12. Moldt JA, Festl-Wietek T, Madany Mamlouk A, Nieselt K, Fuhl W, Herrmann-Werner A. Chatbots for future docs: exploring medical students' attitudes and knowledge towards artificial intelligence and medical chatbots. Med Educ Online 2023 Dec;28(1):2182659. [doi: 10.1080/10872981.2023.2182659] [Medline: 36855245]
13. Al Hadithy ZA, Al Lawati A, Al-Zadjali R, Al Sinawi H. Knowledge, attitudes, and perceptions of artificial intelligence in healthcare among medical students at Sultan Qaboos University. Cureus 2023 Sep;15(9):e44887. [doi: 10.7759/cureus.44887] [Medline: 37814766]

XSL•FO
RenderX

14. Ren Y, Wang Y, Dong S, et al. Application of artificial intelligence in medical education: a systematic and narrative review of pedagogical potential and ethical implications. Adv Med Educ Pract 2026;17:567190. [doi: 10.2147/AMEP.S567190] [Medline: 41973159]
15. Rani S, Kumari A, Ekka SC, Chakraborty R. Perception of medical students and faculty regarding the use of artificial intelligence (AI) in medical education: a cross-sectional study. Cureus 2025 Jan;17(1):e77514. [doi: 10.7759/cureus.77514] [Medline: 39958022]
16. Sapci AH, Sapci HA. Artificial intelligence education and tools for medical and health informatics students: systematic review. JMIR Med Educ 2020 Jun 30;6(1):e19285. [doi: 10.2196/19285] [Medline: 32602844]
17. Khater AS, Zaaqoq AA, Wahdan MM, Ashry S. Knowledge and attitude of Ain Shams University medical students towards artificial intelligence and its application in medical education and practice. Educ Res Innov J 2023 Jul;3(10):29-42. [doi: 10.21608/erji.2023.306718]
18. Wood EA, Ange BL, Miller DD. Are we ready to integrate artificial intelligence literacy into medical school curriculum: students and faculty survey. J Med Educ Curric Dev 2021;8:23821205211024078. [doi: 10.1177/23821205211024078] [Medline: 34250242]
19. Kimmerle J, Timm J, Festl-Wietek T, Cress U, Herrmann-Werner A. Medical students' attitudes toward AI in medicine and their expectations for medical education. J Med Educ Curric Dev 2023;10:23821205231219346. [doi: 10.1177/23821205231219346] [Medline: 38075443]
20. Korniichuk OY, Bambyzov LM, Kosenko VM, Spaska AM, Tsekhmister YV. Application of the case study method in medical education. Int J Learn Teach Educ Res 2021;20(7):175-191. [doi: 10.26803/ijlter.20.7.10]
21. Napier J, Khogali SE. Developing and recognising undergraduate medical educators in a UK medical school - a case study. MedEdPublish (2016) 2019;8:155. [doi: 10.15694/mep.2019.000155.1] [Medline: 38089396]
22. Biri SK, Kumar S, Panigrahi M, Mondal S, Behera JK, Mondal H. Assessing the utilization of large language models in medical education: insights from undergraduate medical students. Cureus 2023 Oct;15(10):e47468. [doi: 10.7759/cureus.47468] [Medline: 38021810]
23. Jha N, Shankar PR, Al-Betar MA, Mukhia R, Hada K, Palaian S. Undergraduate medical students' and interns' knowledge and perception of artificial intelligence in medicine. Adv Med Educ Pract 2022;13:927-937. [doi: 10.2147/AMEP.S368519] [Medline: 36039185]
24. Chan KS, Zary N. Applications and challenges of implementing artificial intelligence in medical education: integrative review. JMIR Med Educ 2019 Jun 15;5(1):e13930. [doi: 10.2196/13930] [Medline: 31199295]
25. Liu DS, Sawyer J, Luna A, et al. Perceptions of US medical students on artificial intelligence in medicine: mixed methods survey study. JMIR Med Educ 2022 Oct 21;8(4):e38325. [doi: 10.2196/38325] [Medline: 36269641]
26. Teng M, Singla R, Yau O, et al. Health care students' perspectives on artificial intelligence: countrywide survey in Canada. JMIR Med Educ 2022 Jan 31;8(1):e33390. [doi: 10.2196/33390] [Medline: 35099397]
27. Ahmed Z, Bhinder KK, Tariq A, et al. Knowledge, attitude, and practice of artificial intelligence among doctors and medical students in Pakistan: a cross-sectional online survey. Ann Med Surg (Lond) 2022;76:103493. [doi: 10.1016/j.amsu.2022.103493] [Medline: 35308436]
28. Kwak Y, Ahn JW, Seo YH. Influence of AI ethics awareness, attitude, anxiety, and self-efficacy on nursing students' behavioral intentions. BMC Nurs 2022 Sep 30;21(1):267. [doi: 10.1186/s12912-022-01048-0] [Medline: 36180902]
29. Azer SA, Guerrero AP. The challenges imposed by artificial intelligence: are we ready in medical education? BMC Med Educ 2023 Sep 19;23(1):680. [doi: 10.1186/s12909-023-04660-z] [Medline: 37726741]
30. Graham A. AI in medical education: a bright new frontier. BMJ 2024 Jun 21;385:q1334. [doi: 10.1136/bmj.q1334] [Medline: 38906539]
31. Lekadir K, Frangi AF, Porras AR, et al. FUTURE-AI: international consensus guideline for trustworthy and deployable artificial intelligence in healthcare. BMJ 2025 Feb 5;388:e081554. [doi: 10.1136/bmj-2024-081554] [Medline: 39909534]
32. Paranjape K, Schinkel M, Nannan Panday R, Car J, Nanayakkara P. Introducing artificial intelligence training in medical education. JMIR Med Educ 2019 Dec 3;5(2):e16048. [doi: 10.2196/16048] [Medline: 31793895]
33. Charow R, Jeyakumar T, Younus S, et al. Artificial intelligence education programs for health care professionals: scoping review. JMIR Med Educ 2021 Dec 13;7(4):e31043. [doi: 10.2196/31043] [Medline: 34898458]
34. CUHK receives HK$30 million donation from Mr Li Ka-shing to further enhance artificial intelligence in medical education. Chinese University of Hong Kong. 2023. URL: https://www.med.cuhk.edu.hk/press-releases/cuhk-receives-hkd30-million-donation-from-mr-li-ka-shing-to-further-enhance-artificial-intelligence-in-medical-education [accessed 2026-05-12]
35. NHS Long Term Workforce Plan fact sheet. NHS England. 2023. URL: https://healthmedia.blog.gov.uk/2023/06/30/nhs-long-term-workforce-plan-fact-sheet/ [accessed 2026-05-12]
36. Tonkin T. Tech adoption poses risk to NHS. British Medical Association. 2025. URL: https://www.bma.org.uk/news-and-opinion/tech-adoption-poses-risk-to-nhs [accessed 2026-05-12]
37. Laupichler MC, Aster A, Meyerheim M, Raupach T, Mergen M. Medical students' AI literacy and attitudes towards AI: a cross-sectional two-center study using pre-validated assessment instruments. BMC Med Educ 2024 Apr 10;24(1):401. [doi: 10.1186/s12909-024-05400-7] [Medline: 38600457]

XSL•FO
RenderX

38. Boscardin CK, Gin B, Golde PB, Hauer KE. ChatGPT and generative artificial intelligence for medical education: potential impact and opportunity. Acad Med 2024 Jan 1;99(1):22-27. [doi: 10.1097/ACM.0000000000005439] [Medline: 37651677]
39. Artificial intelligence and innovative technologies. General Medical Council. URL: https://www.gmc-uk.org/professional-standards/learning-materials/artificial-intelligence-and-innovative-technologies [accessed 2026-05-12]
40. One in four UK doctors are using artificial intelligence. Alan Turing Institute. 2024. URL: https://www.turing.ac.uk/sites/default/files/2024-10/public_services_gmc_report_v3.pdf [accessed 2026-05-12]
41. Freeman J. Student generative AI survey 2025. : Higher Education Policy Institute; 2025 URL: https://www.hepi.ac.uk/reports/student-generative-ai-survey-2025/ [accessed 2026-05-12]

Abbreviations

AI: artificial intelligence
CUHK: Chinese University of Hong Kong
GMC: General Medical Council
HEE: Health Education England
NHS: National Health Service

Edited by S Guinez-Molinos; submitted 06.Aug.2025; peer-reviewed by F Fukuzawa, H Kasai, L Ramnath, S Ito; revised version received 04.Mar.2026; accepted 02.Apr.2026; published 12.Jun.2026.

Please cite as:
Gaur A, Kirani Tan J, Sridhar Rao M, Fuad M, Bhatti T, Rishab Bharadwaj H, Mohamed Ahmed KAH
AI in UK Medical Education: A Framework for Curriculum Reform
JMIR Med Educ 2026;12:e81953
URL: https://mededu.jmir.org/2026/1/e81953
doi:10.2196/81953

© Aditya Gaur,Joecelyn Kirani Tan, Medha Sridhar Rao, Muhtasim Fuad, Taha Bhatti, Hareesha Rishab Bharadwaj, Khabab Abbasher Hussien Mohamed Ahmed. Originally published in JMIR Medical Education (https://mededu.jmir.org), 12.Jun.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Education, is properly cited. The complete bibliographic information, a link to the original publication on https://mededu.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

The AI Colleague: Reimagining Interprofessional Collaboration and Education in Medicine

Chisung Yuh[1], DDS; Chul-Hyun Cho[2*], MD, PhD; Young-Mee Lee[1*], MD, PhD, MSED

[1]Department of Medical Education, Korea University College of Medicine, Seoul, Republic of Korea

[2]Department of Psychiatry, Biomedical Informatics, and Medical Education, Korea University College of Medicine, 73 Goryeodae-ro, Seongbuk-gu, Seoul, Republic of Korea

[*]these authors contributed equally

Corresponding Author:
Chul-Hyun Cho, MD, PhD
Department of Psychiatry, Biomedical Informatics, and Medical Education, Korea University College of Medicine, 73 Goryeodae-ro, Seongbuk-gu, Seoul, Republic of Korea

Abstract

AI is no longer confined to optional decision support; it is becoming a routine presence in clinical workflows, shaping diagnostic hypotheses, triage priorities, risk estimates, and documentation. Yet most educational responses still treat AI as a tool operated by an individual clinician. This framing underestimates how AI reshapes the actual unit of practice: the interprofessional team. We propose the concept of AI-expanded interprofessional collaboration (AI-IPC), in which AI systems function as consequential participants in team cognition—not as moral agents, but as sources of recommendations, uncertainty, and constraints that reorganize communication, authority gradients, and accountability. Building on interprofessional education (IPE) theory, situated learning, and distributed cognition, we argue that "AI literacy" alone is insufficient; learners must be trained to coordinate human-AI-human collaboration in realistic clinical settings. We outline a pragmatic, theory-aligned approach for AI-expanded interprofessional education (AI-IPE): clarifying boundary conditions for AI participation, mapping AI-specific subcompetencies onto established IPE frameworks, and evaluating performance at the level of team behaviors rather than knowledge recall. We present the Interprofessional Education Collaborative (IPEC)–aligned evaluation scaffold with concrete learning activities and assessment approaches, and we outline how the approach can be tailored across undergraduate, postgraduate, and continuing education settings. We further emphasize that implementation requires digital infrastructure, institutional governance, and educator capacity that bridges AI, clinical workflow design, and IPE facilitation. Finally, we address the deliberate use of the "AI colleague" metaphor—not to anthropomorphize AI, but to make the relational and coordinative demands of AI integration visible and teachable. The pedagogical aim is calibrated, critical engagement: teams that verify, question, and, when warranted, override AI contributions rather than defer to them. AI will not replace interprofessional collaboration; it will change what collaboration requires. Education should make that change explicit, rehearsable, and assessable.

(JMIR Med Educ 2026;12:e94034) doi:10.2196/94034

KEYWORDS

AI; interprofessional education; interprofessional collaboration; medical education; competency-based education

Introduction: The Unit of Practice Is Shifting

In contemporary clinical settings, AI is increasingly present at the point where decisions are made: in radiology workflows, risk prediction and early warning dashboards, medication safety alerts, ambient documentation systems that draft clinical notes from recorded encounters [1], and emerging "agentic" systems that can coordinate tasks across clinical and operational domains [2-5]. Even when AI is not autonomous, its outputs can become the scaffolding around which teams organize attention, deliberation, and documentation.

Medical education has responded with calls for AI literacy, data science training, and ethical competencies [6-8]. These are necessary—but they are also incomplete. Many educational proposals implicitly assume that the unit of practice is a single clinician operating a tool. In reality, the unit of practice in modern health care is often the team. Interprofessional collaboration (IPC) is not merely a social ideal; it is a cognitive and organizational strategy for managing complexity, uncertainty, and accountability in patient care [9,10]. The World Health Organization defines collaborative practice as occurring when "multiple health workers from different professional backgrounds work together with patients, families, carers, and communities to deliver the highest quality of care" [11]. In everyday medicine, IPC takes concrete forms: a ward round in which physicians, nurses, and pharmacists reconcile a medication plan; a tumor board in which radiologists, pathologists, surgeons, and oncologists negotiate a staging

XSL•FO
RenderX

decision; a rapid response team in which nursing observations trigger escalation and shared decision-making. Each of these settings is now a plausible site of AI participation—an early warning score prompting the escalation, a draft imaging report anchoring the tumor board discussion, an interaction alert reshaping the medication review.

Recent literature has begun to map the intersection of AI and interprofessional education (IPE). A mixed methods scoping review identified 15 studies exploring AI applications aimed at fostering interprofessional teamwork, including virtual reality simulations, clinical decision support systems, and machine learning algorithms [12]. Similarly, simulation-based approaches leveraging big data analytics and AI have been proposed to enhance IPC training [13]. However, most of this work addresses AI as a tool for improving existing IPE or describes what AI applications currently exist. What remains underexplored is a conceptual and practical framework for how IPE itself must be redesigned when AI becomes a routine participant in clinical team deliberations—and how such redesigned IPE can be assessed.

This viewpoint addresses that gap. Starting from a simple observation—that AI is beginning to reconfigure collaboration—we propose a framework for AI-expanded IPE (AI-IPE) that focuses on the educational design and evaluation of team competencies under AI-integrated conditions. Our contribution is not to catalog AI applications for IPE, but to articulate what teams must learn to do differently when AI outputs enter everyday clinical workflows, and how that learning can be structured, practiced, and assessed. We write primarily for health professions educators, curriculum and assessment leaders, and directors of IPE programs; clinical supervisors, institutional leaders responsible for AI governance, and health professions education researchers are secondary audiences with clear stakes in the argument. Two commitments frame everything that follows. The first concerns our aim: we want to cultivate disciplined skepticism as a collaborative habit, so that teams verify AI outputs, question them openly, and override them when the evidence warrants, rather than defer to them. The second concerns what we do and do not claim about AI: we do not suggest that AI systems think, understand, or bear responsibility. The educational problem is precisely that systems without understanding now produce outputs that teams must treat as consequential.

Guiding Educational Frameworks

Because the argument draws on several educational traditions, we introduce them at the outset. Situated learning theory holds that competence develops through legitimate peripheral participation in authentic activity: people learn to collaborate by collaborating, within communities of practice rather than in lecture halls [14,15]. Distributed cognition extends the unit of analysis beyond the individual mind: cognitive work in clinical settings is distributed across people, artifacts, and environments, so a risk score or a generated report is not merely information but a structural participant in how a team thinks [16,17].

Competency-based medical education (CBME), the organizing paradigm of contemporary training, completes the frame: it defines outcomes as observable behaviors and asks what learners can be entrusted to do in real practice [18,19]. For interprofessional competencies specifically, the Interprofessional Education Collaborative (IPEC) Core Competencies (version 3) provide the most widely used anchor, organizing collaboration into values and ethics, roles and responsibilities, interprofessional communication, and teams and teamwork [20]. These perspectives matter here because each is stressed differently when AI enters the room: situated learning tells us where AI collaboration must be learned, distributed cognition tells us why AI outputs reshape team reasoning, and CBME and IPEC tell us how the resulting competencies can be specified and assessed. We return to each throughout.

From “Tool” to “Colleague”: A Deliberate Conceptual Choice

Why insist on the language of “colleague”? Because it forces education to confront a neglected dimension of AI integration: interaction. In many clinical deployments, AI is not just an instrument that produces a number; it enters conversations—explicitly or implicitly—as something that can be cited, deferred to, or overridden. It can shift authority gradients (“the algorithm says...”), narrow the hypothesis space, and alter the timing of action.

Consider a concrete example. Machine learning–based sepsis early warning systems are now in routine clinical use. The TREWS (Targeted Real-Time Early Warning System) system, implemented across 5 hospitals in the Johns Hopkins Health System, generates real-time sepsis alerts that clinicians review and confirm [21], and the COMPOSER deep learning model, deployed in emergency departments at UC San Diego Health, continuously monitors more than 150 patient variables and issues a nurse-facing alert when sepsis risk is elevated [22]. In both deployments, the alert does not simply produce a number; it triggers a communication sequence—the nurse evaluates the alert, contacts the physician, and together they decide whether to initiate a treatment bundle. Prospective evaluations have reported lower mortality and faster treatment when this human-in-the-loop workflow functions well [21,22]. Yet parallel evidence shows that alert fatigue, mistrust, and poorly calibrated reliance can undermine outcomes [23]. These failure modes are not technical—they are social-cognitive: they emerge in teams, under time pressure, with uneven expertise, and in the presence of accountability demands.

A fair question is why AI deserves special educational attention when clinical teams already work amid infusion pumps, physiological monitors, and rule-based decision support systems that interpret data and issue recommendations. We see continuity rather than rupture: automation bias and alarm fatigue were documented long before modern machine learning [24]. Three features of contemporary AI nonetheless change the collaborative stakes. First, these systems are adaptive and often opaque; their behavior shifts with training data and context in ways that resist the simple mental models clinicians build for conventional devices. Second, generative systems produce fluent narratives—draft reports, summaries, and replies—that enter team communication in the same register as human speech

XSL•FO
RenderX

rather than as a beep or a threshold flag. Third, AI now spans the entire workflow, from triage to documentation, so its outputs cross professional boundaries instead of remaining within a single device-operator dyad. We agree that regulatory clarity—ensuring that AI tools are appropriately categorized and governed as medical devices—is a necessary complement to education [25,26]; our claim is that education must additionally prepare teams for the interactional consequences that follow even when such governance is sound.

We are aware that the "colleague" label invites criticism. AI is not a moral agent, does not hold professional licensure, and cannot bear accountability. Anthropomorphism can obscure power asymmetries and deflect responsibility. There is also a subtler hazard, which deserves emphasis at the outset rather than as a closing footnote: clinicians already tend to experience fluent AI systems as thinking interlocutors when they are in fact probabilistic models that do not understand the propositions they generate. Any pedagogy that borrows the language of collegiality must build in this corrective at the same time. We use the term deliberately, not to grant AI personhood, but as a pedagogical provocation: if education treats AI as merely another stethoscope, it will fail to prepare learners for the relational complexity that AI actually introduces. The "colleague" metaphor is a heuristic that makes the coordinative demands of AI integration visible and discussable in educational settings. Throughout this paper, we use the more precise term "consequential participant" to maintain conceptual rigor while retaining the pedagogical utility of the metaphor. The metaphor earns its keep only if it makes AI easier to question, not easier to trust.

From the perspective of distributed cognition theory, this reframing is well grounded. As noted earlier, cognition is spread across people, artifacts, and environments rather than being sealed inside individual minds [16,17]. In clinical settings, AI outputs function as cognitive artifacts that shape how teams allocate attention, interpret uncertainty, and coordinate action. When an AI system generates a risk score that restructures team communication, it is participating in the distributed cognitive system—regardless of whether we call it a colleague or a tool.

Automation bias—overreliance on computer suggestions even when incorrect—has been documented extensively in clinical decision support, and it can be amplified when AI outputs carry perceived objectivity or institutional endorsement [24,27]. Equally, unreflective rejection of AI can occur when trust is poorly calibrated, leading to missed opportunities for safety and efficiency. Both failure modes are social-cognitive challenges that demand collaborative—not merely individual—educational responses. IPE already possesses relevant tools for this problem. The TeamSTEPPS (Team Strategies and Tools to Enhance Performance and Patient Safety) framework, developed by the Agency for Healthcare Research and Quality (AHRQ), provides training in structured mutual challenge; its "two-challenge rule" obliges any team member to voice a safety concern at least twice and to escalate if the concern is not acknowledged [28]. AI-expanded interprofessional collaboration (AI-IPC) extends exactly this repertoire to a nonhuman participant: teams need norms that make questioning an algorithmic output as routine—and as psychologically safe—as questioning a senior colleague's plan. Framed this way, the competency at stake is not to treat AI as "another professional" but to ensure that its suggestions receive the same disciplined scrutiny that well-functioning teams apply to any consequential claim.

Defining AI-IPC

AI-IPC is not intended as a metaphor. It is a structural description of how team-based care changes when AI outputs are embedded in workflows. We define AI-IPC as follows: a mode of interprofessional practice in which AI systems function as consequential participants in team deliberation by generating recommendations, risk estimates, or synthesized narratives that materially influence attention, reasoning, and documentation. Professional accountability remains entirely human; AI is a participant in the sense that it contributes inputs that teams must negotiate.

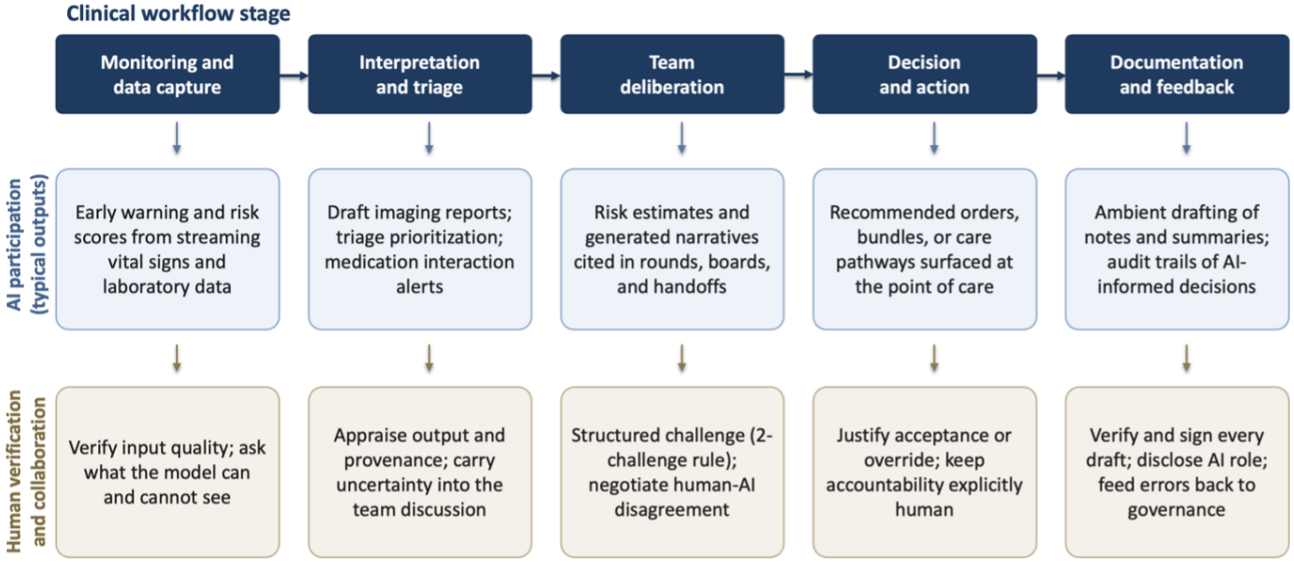
AI becomes a meaningful participant when 3 boundary conditions are satisfied:

- Operational integration—AI outputs are routinely available at the point of care (eg, triage dashboards, imaging report generation, and ambient note drafting).
- Override governance—there are explicit norms or policies for accepting, contesting, or overriding AI recommendations.
- Traceable accountability—teams must document and justify decisions, including when they depart from AI output.

These conditions are increasingly common as multimodal generative systems mature and as "agentic AI teammates" are envisioned for coordinated task execution [4,29]. The question for education is not whether AI will be present, but whether clinicians will be prepared to collaborate safely and transparently in AI-IPC environments. Figure 1 locates these boundary conditions in a generic clinical workflow: it marks where AI outputs typically enter team activity—from monitoring and triage through deliberation, decision, and documentation—and pairs each entry point with the human verification and collaboration practices that education must make rehearsable.

XSL•FO
RenderX

Figure 1. Points of AI participation across a generic clinical workflow, and the corresponding human verification and collaboration practices that AI-expanded interprofessional education (AI-IPE) must rehearse. Accountability remains human at every stage: AI contributes inputs that teams verify, question, and—when warranted—override.



Why AI Literacy Is Necessary but Not Sufficient

As used here, AI literacy denotes the cluster of knowledge and skills that allows health professionals to understand what AI systems can and cannot do, interpret their outputs, and recognize their ethical and legal implications—including data provenance, algorithmic bias, and privacy [6-8]. AI literacy typically focuses on understanding model capabilities, limitations, and ethical issues. These are essential foundations. However, they do not specify how teams should work when AI outputs collide with human judgment, when uncertainty must be communicated across disciplines, or when documentation must assign responsibility for an AI-influenced decision. Such moments are not rare edge cases—they are the routine friction points of AI integration. Ambient documentation is a case in point: generated drafts are fluent and mostly accurate, which is precisely why they demand line-by-line verification before signature. Knowing in the abstract that errors occur does not settle who verifies, who signs, and how a team recovers when a fabricated detail is discovered after the fact [1].

Consider a common team scenario: a nurse reports an AI-based early warning score suggesting deterioration, while a resident believes the patient is stable. A pharmacist raises concerns about a medication recommendation surfaced by an AI-driven interaction checker. A senior physician asks whether the AI output is explainable enough to justify an invasive intervention. Each question is simultaneously clinical and collaborative: it requires negotiation across roles, explicit communication of uncertainty, and a clear account of who decides and who documents. These are IPE competencies—but with an AI-specific twist.

In other words, the educational target is not "how to use AI," but "how to collaborate when AI is in the room." Part of that competence is provenance awareness: registering that a given artifact is AI-crafted and adjusting practice accordingly. Provenance matters for at least 4 reasons. It signals how an output was produced—statistical association rather than clinical reasoning—and therefore how it is likely to fail. It determines what verification is owed before the output informs action. It shapes documentation and disclosure, including what patients are told about how a recommendation arose. And it calibrates the epistemic weight a team should grant the output in discussion, especially when it conflicts with human judgment. A team that cannot tell—or does not ask—whether a narrative was machine-generated can perform none of these adjustments.

Situated Learning and Distributed Cognition: Making AI-IPC Teachably Real

IPE succeeds when it is not purely declarative but practiced in contexts that resemble real work. As outlined earlier, situated learning locates competence in social participation and authentic activity [14,15]. IPE, at its best, is a situated apprenticeship in collaboration.

AI-IPC strengthens the case for situated approaches. AI outputs are context-sensitive: the same recommendation can be interpreted differently depending on role, workflow, patient context, and institutional norms. Moreover, AI introduces new "objects" of coordination—probabilistic risk scores, generated narratives, uncertainty estimates, and provenance questions—that must be discussed in real time. These are not learned effectively through lectures alone. Generative systems can themselves supply raw material for such training: large language models can draft clinical cases with realistic parameters, although comparative work in dermatology education found that AI-generated cases, while rich in descriptive detail, lacked some of the clinical depth and dilemma quality of instructor-crafted ones—a useful reminder that educational artifacts generated by AI require the same critical appraisal as clinical outputs [30].

Distributed cognition theory further illuminates why simulation is essential. Because cognitive work in AI-IPC environments is distributed across clinicians, AI artifacts, and documentation

XSL•FO
RenderX

systems, effective team performance requires rehearsing the entire sociotechnical system—not merely individual AI skills [17]. Learners must practice interpreting AI outputs while simultaneously communicating with team members, managing uncertainty, and documenting decisions—activities that only emerge in realistic interprofessional settings.

Thus, we argue for AI-IPE: interprofessional education that explicitly incorporates AI as a structured participant in the learning environment so that learners develop collaborative fluency under AI-integrated conditions.

AI-IPE: Extending, Not Replacing, Established Frameworks

AI-IPE does not replace what we already know about IPE. Instead, it should extend established frameworks with AI-specific subcompetencies. The IPEC Core Competencies (version 3), introduced above, remain the natural anchor: values and ethics, roles and responsibilities, interprofessional communication, and teams and teamwork [20]. Our contribution is to make explicit how each domain is reshaped when AI outputs enter into everyday team deliberations.

For example, values and ethics now include competence in recognizing algorithmic bias, understanding transparency limits, and protecting patient autonomy in AI-informed decisions [26,31]. Roles and responsibilities require explicit agreements about who interprets AI outputs, who validates them, and how accountability is documented. Interprofessional communication must incorporate communication of probabilistic uncertainty and provenance [32]. Teams and teamwork must manage disagreement—among professionals and between professionals and AI outputs—without defaulting to either algorithmic deference or reflexive dismissal.

CBME further supports this direction, as introduced earlier, because AI-IPE competencies are observable behaviors: critical appraisal of AI outputs, justification of override decisions, and transparent documentation [18,19]. The educational question becomes: what can learners be entrusted to do—safely, ethically, and collaboratively—when AI is part of the workflow?

Table 1 contrasts traditional IPE with AI-IPE across 5 dimensions, emphasizing that the shift is not merely technical but relational.

Table . Traditional interprofessional education (IPE) and AI-Expanded IPE (AI-IPE) compared across 5 dimensions.

Dimension	Traditional IPE	AI-IPE	Educational implication
Team composition	Human professionals only	Human professionals+AI systems producing consequential inputs	Teach human-AI-human collaboration
Core coordination problem	Role clarity and communication across professions	Role clarity+AI interpretation, trust calibration, and override governance	Train teams to negotiate AI output
Uncertainty management	Clinical uncertainty shared across humans	Clinical uncertainty+algorithmic uncertainty and provenance	Practice communicating probabilistic outputs
Accountability	Human accountability, professional norms	Human accountability with AI-influenced decisions and audit trails	Train documentation and justification practices
Learning setting	Simulation of interprofessional scenarios	Simulation including AI-generated recommendations, conflicts, and error modes	Situated rehearsal of AI-IPC[a]

[a]IPC: interprofessional collaboration.

Tailoring AI-IPE Across the Learner Continuum

AI-IPE cannot be one-size-fits-all, because awareness of and familiarity with AI—and the authority to act on its outputs—differ systematically across the educational continuum. At the undergraduate (prelicensure) stage, learners are still forming their professional identities and arrive with heterogeneous levels of AI exposure, much of it from consumer tools rather than clinical systems. Here, the emphasis should be on foundational AI literacy, a shared interprofessional vocabulary for AI outputs, and low-stakes simulation in which questioning an algorithm is practiced before questioning carries real consequences. Mixed-profession groups are valuable early, precisely because norms of cross-professional challenge are easiest to establish before hierarchies harden.

In postgraduate training, learners hold real decision authority within specialty-specific workflows, and AI-IPE should correspondingly move into workplace-based rehearsal of override decisions, escalation pathways, and documentation practices, exercised in the learner's actual clinical context, with supervisors assessing entrustment under AI-integrated conditions. For continuing professional development, the audience is intact teams with entrenched routines and highly variable AI familiarity—including inverted expertise gradients in which junior members understand the AI better than their seniors. Short, team-based rehearsals embedded in existing quality and safety structures (eg, morbidity and mortality conferences or periodic simulation refreshers) fit this group better than standalone coursework. Facilitation should be planned as deliberately as content: pairing an experienced IPE facilitator with a clinical informatician covers both the group dynamics and the technical substance, and the composition of each learner group—homogeneous or mixed across professions and career stages—should be an explicit design decision rather than an accident of scheduling.

XSL•FO
RenderX

From Concept to Assessment: A Proposed, IPEC-Aligned Evaluation Scaffold

Because AI-IPE targets team behaviors, evaluation must focus on performance in collaboration—not solely on knowledge. We therefore propose (not prescribe) an evaluation scaffold aligned with IPEC domains and CBME principles, designed to assess learning and behavior levels (Kirkpatrick levels 2 - 3; Table 2) [9,33]. Textbox 1 complements the scaffold with brief cases and reflective questions that educators can use to seed discussion at any of these levels.

Table . A proposed Interprofessional Education Collaborative (IPEC)–aligned evaluation framework for AI-expanded interprofessional education (AI-IPE)[a].

IPEC domain	AI subcompetency	Learning activity	Assessment approach	Level
Values and ethics	Recognize bias, transparency limits, and patient autonomy issues in AI-informed decisions	AI-assisted triage discussion with fairness challenge (eg, biased risk score across demographics)	Bias-identification rubric+structured reflection	Learning
Roles and responsibilities	Clarify accountability for interpreting, validating, and documenting AI output	Interprofessional simulation where AI recommendation conflicts with clinician judgment (eg, sepsis alert vs clinical assessment)	Justification checklist+documentation audit	Learning or behavior
Interprofessional communication	Communicate probabilistic AI output (uncertainty/provenance) to team and patient	Standardized patient encounter explaining AI-informed plan (eg, explaining AI-generated risk score to patient and family)	OSCE[b]-style communication rating+debrief transcript review	Learning or behavior
Teams and teamwork	Manage disagreement constructively; calibrate trust and coordinate overrides	Multidisciplinary case conference with AI-generated report and embedded error mode (eg, hallucinated findings)	Team performance rubric+override-decision log review	Learning or behavior

[a]This table is proposed as an illustrative scaffold. Institutions should adapt it to local contexts, available AI tools, and governance requirements. Activities labeled "Learning or Behavior" assess Kirkpatrick level 2 when conducted as simulation, and level 3 only when they involve direct observation of behavior in live clinical practice (eg, workplace-based observation, override-decision audits, and multisource feedback).

[b]OSCE: objective structured clinical examination.

Textbox 1. Illustrative AI-IPC cases with reflective questions for educators.

Case 1: The contested alert. At 2 AM, an early warning system flags a postoperative patient for probable sepsis. The bedside nurse finds the patient sleeping comfortably; the covering resident, examining the patient, is reassured and inclined to dismiss the alert. The nurse, following unit policy, asks the resident to state a rationale for the override so it can be documented.

Reflective questions: What would count as an adequate justification to override this alert? Who documents the override, and where? If the patient deteriorates 3 hours later, how will the team's earlier reasoning be reconstructed—and defended?

Case 2: The fluent draft. An ambient documentation system produces a discharge summary that reads well but attributes a normal neurological examination that no one performed. The intern notices the discrepancy only because a physiotherapist queries the mobility plan that was built on it.

Reflective questions: Who is responsible for verifying and signing AI-drafted documentation? How should the team repair the record—and the plan—once a fabricated detail has propagated? What should the patient be told?

Case 3: The skewed score. During discharge planning rounds, a nurse notices that a readmission risk score consistently rates patients from one demographic group as lower risk, with the effect that follow-up resources are being allocated away from them.

Reflective questions: How should the nurse raise this concern, and to whom? What does a 2-challenge escalation—TeamSTEPPS's [Team Strategies and Tools to Enhance Performance and Patient Safety] expectation that a safety concern be voiced at least twice before escalating—look like when the "colleague" being challenged is an algorithm, and the humans in the room did not build it? At what point does an educational observation become a governance report?

XSL•FO
RenderX

What Makes AI-IPE Feasible: Infrastructure, Governance, and Educator Capacity

A concept only matters educationally if institutions can actually implement and assess it. AI-IPE, therefore, depends on 3 enabling conditions that are often underestimated.

The first condition is digital infrastructure. AI-IPE requires learning environments where AI outputs are not hypothetical but operationally similar to what teams will encounter. This can include AI-enabled simulation cases (eg, generated imaging reports or risk dashboards), structured error-mode scenarios (eg, biased outputs or hallucinated narratives), and digital capture of decisions [34,35]. Assessment also requires digitized tools that can record override decisions, documentation patterns, and team interactions (eg, audit trails, rubric platforms, or annotated transcripts). Without such tools, evaluation collapses into self-reporting and loses credibility.

The second condition is institutional governance. Trustworthy AI guidance emphasizes transparency, safety, and accountability as prerequisites for clinical deployment [25,26]. In education, these translate into teachable rules: documentation norms for AI-informed decisions, escalation pathways when AI output is contested, and structured communication with patients about the role of AI in their care. Clinical AI reporting frameworks (eg, CONSORT-AI [Consolidated Standards of Reporting Trials–AI Extension], SPIRIT-AI [Standard Protocol Items: Recommendations for Interventional Trials–AI Extension], and DECIDE-AI [Developmental and Exploratory Clinical Investigations of Decision support systems driven by AI]) provide models for how transparency and accountability can be operationalized [36,37]. Training without governance risks normalizing unsafe practices; governance without training risks creating policies that are ignored in practice.

The third condition is educator capacity. AI-IPE cannot be delivered by AI experts alone, nor by IPE facilitators without AI literacy. Educators must hold an integrated understanding of (1) AI capabilities and limitations, (2) IPE or IPC theory and facilitation, (3) clinical workflow realities, and (4) digital assessment methodologies. Faculty development should, therefore, be an explicit component of AI-IPE implementation [7,8]. In practice, institutions may need interdisciplinary educator teams (clinical faculty, medical education specialists, informaticians, and quality or safety leaders) to co-design AI-IPE cases and assessments. If educators cannot model how to interrogate AI output, negotiate disagreement, and document accountability, learners will be trained to perform either deference or skepticism—rather than calibrated, collaborative judgment.

Practical Next Steps for Implementation

For institutions persuaded by the argument so far, the practical question is where to begin. A staged sequence keeps the work tractable and lets faculty capacity grow alongside it.

1. Map the AI already present. Audit where AI outputs actually enter local workflows—triage scores, imaging drafts, interaction alerts, ambient notes—and identify the 2 or 3 that most often cross professional boundaries.
2. Convene an interdisciplinary design team. Pair IPE facilitators with clinical informaticians, frontline clinicians from each profession, and quality and safety leads; none of these groups can design AI-IPE alone.
3. Retrofit one existing IPE scenario. Rather than building a new curriculum, add a realistic AI artifact—with an embedded error mode—to a scenario the institution already runs, and script the decision points it creates.
4. Pilot with an explicit debriefing protocol. Run the scenario with mixed-professional groups and debrief on trust calibration, challenging behavior, and documentation—not only on clinical correctness.
5. Attach the assessment early. Apply an adapted version of the Table 2 scaffold, with digital capture of override decisions and team communication, so that evaluation data accumulate from the first iteration.
6. Report through governance channels and iterate. Feed findings back to the institutional AI governance structure, align teaching with local policies, and expand into additional scenarios and learner groups as faculty capacity grows.

Implications for Patients

The case for AI-IPE is ultimately a case about patient care. Collaborative practice, by the definition adopted above, includes patients, families, and carers—not only professionals [11]. When teams rehearse uncertainty communication and provenance disclosure, the intended beneficiary is the patient who must decide whether to accept a plan that an algorithm helped to shape. Focus-group research indicates that patients bring specific and reasonable concerns to AI-informed care—safety, threats to choice, cost, data-source bias, and data security—and that their acceptance is contingent on seeing these concerns addressed [38]. A team trained to explain what an AI contribution was, and was not, is better positioned to support informed consent and shared decision-making than a team that either conceals the algorithm or hides behind it. We treat the fuller patient perspective—including whether AI-IPE measurably improves patient understanding and trust—as an empirical question for the research agenda below.

Limitations and Research Agenda

This viewpoint starts from a conceptual claim: AI reshapes the unit of collaborative practice and therefore reshapes IPE. This claim should be open to empirical scrutiny. Future work should examine whether AI-IPE improves team performance in domains such as the quality of override decisions, the clarity of uncertainty communication, documentation transparency, and patient understanding. Designs might include simulation-based controlled studies, longitudinal workplace assessments, and implementation evaluations aligned with reporting frameworks for clinical AI (eg, CONSORT-AI, SPIRIT-AI, and DECIDE-AI) [36,37]: syntheses of the resulting evidence should in turn follow emerging standards for AI systematic reviews

XSL•FO
RenderX

[39]. Because the present argument is conceptual and narrative in nature, a natural next step is a domain-specific implementation study. Sepsis response teams and oncology multidisciplinary conferences are attractive first testbeds: AI tools are already consequential in both, and the team decision points are well demarcated. Such a study would operationalize the Table 2 scaffold in a single clinical domain, with prespecified team-level outcomes before any claim of generality is made.

A further limitation is heterogeneity: AI systems differ, workflows differ, and cultures of collaboration differ. AI-IPE should therefore be conceptualized as a framework for local adaptation rather than a single standardized curriculum. Additionally, our proposed evaluation scaffold (Table 2) is illustrative; empirical validation across institutions and professional contexts is needed before it can serve as a prescriptive assessment standard.

Finally, we acknowledge that the "AI colleague" metaphor has limits, as we have emphasized from the outset. It risks conveying agency, intentionality, or accountability where none exists. The value of the metaphor lies precisely in its provocation: by framing AI as a colleague, we make visible the coordination demands that would otherwise remain hidden under the label of "tool." The purpose is to prompt educational action, not to make ontological claims about AI.

Conclusion: Education Should Make Collaboration With AI Teachable

AI will not replace interprofessional collaboration; it will change what collaboration requires. When AI outputs become consequential inputs into team cognition, the collaborative work of health care teams shifts: communication, trust calibration, disagreement management, and documentation become AI-conditioned practices. Treating AI as merely a tool is educationally insufficient because it hides the interactive, relational nature of AI integration.

AI-IPC and AI-IPE provide a language for making this shift visible, teachable, and testable. Grounded in situated learning and distributed cognition and aligned with established IPE frameworks, AI-IPE asks learners to rehearse human-AI-human collaboration under realistic conditions and to demonstrate observable team behaviors—verification, structured challenge, justified override, and transparent documentation—especially in moments of uncertainty and disagreement. Realizing this vision requires more than enthusiasm for AI: it requires infrastructure, digitized assessment, institutional governance, and educators equipped to bridge AI, workflows, and interprofessional pedagogy. If we can teach teams to work well with each other, we can also teach them to work well with AI—provided that we are explicit about what that work entails.

Acknowledgments

During manuscript preparation, the authors made limited use of a generative AI tool (Claude Opus 4.7 and 4.8) to support language refinement and structural editing. The authors reviewed and edited all content and take full responsibility for the final manuscript.

Funding

This study was supported by the National Research Foundation (NRF) of Korea (grant NRF-2021R1A5A8032895, RS-2026 - 25523612, and RS-2026 - 25471696). The funders had no role in study design, data collection, analysis, interpretation, manuscript preparation, or the decision to submit the paper for publication.

Data Availability

Data sharing is not applicable to this paper as no datasets were generated or analyzed during this work.

Authors' Contributions

CHC and YML are the corresponding authors responsible for the manuscript submission, peer review, publication process, and authorship details, and can be reached via email at david0203@gmail.com and ymleehj@korea.ac.kr, respectively. CY and CHC led conceptualization and drafted the manuscript. CHC and YML contributed to conceptual development and critically revised the manuscript for important intellectual content. All authors approved the final version and take responsibility for the decision to submit for publication.

Conflicts of Interest

None declared.

References

1. Kanaparthy NS, Villuendas-Rey Y, Bakare T, et al. Real-world evidence synthesis of digital scribes using ambient listening and generative artificial intelligence for clinician documentation workflows: rapid review. JMIR AI 2025 Oct 10;4:e76743. [doi: 10.2196/76743] [Medline: 41071988]
2. Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. Nat Med 2019 Jan;25(1):44-56. [doi: 10.1038/s41591-018-0300-7] [Medline: 30617339]

XSL•FO
RenderX

3. Rajkomar A, Dean J, Kohane I. Machine learning in medicine. N Engl J Med 2019 Apr 4;380(14):1347-1358. [doi: 10.1056/NEJMra1814259] [Medline: 30943338]
4. Zou J, Topol EJ. The rise of agentic AI teammates in medicine. Lancet 2025 Feb 8;405(10477):457. [doi: 10.1016/S0140-6736(25)00202-8] [Medline: 39922663]
5. Moritz M, Topol E, Rajpurkar P. Coordinated AI agents for advancing healthcare. Nat Biomed Eng 2025 Apr;9(4):432-438. [doi: 10.1038/s41551-025-01363-2] [Medline: 40169759]
6. Wartman SA, Combs CD. Medical education must move from the information age to the age of artificial intelligence. Acad Med 2018 Aug;93(8):1107-1109. [doi: 10.1097/ACM.0000000000002044] [Medline: 29095704]
7. Russell RG, Lovett Novak L, Patel M, et al. Competencies for the use of artificial intelligence-based tools by health care professionals. Acad Med 2023 Mar 1;98(3):348-356. [doi: 10.1097/ACM.0000000000004963] [Medline: 36731054]
8. Masters K. Ethical use of artificial intelligence in health professions education: AMEE Guide No. 158. Med Teach 2023 Jun;45(6):574-584. [doi: 10.1080/0142159X.2023.2186203] [Medline: 36912253]
9. Reeves S, Perrier L, Goldman J, Freeth D, Zwarenstein M. Interprofessional education: effects on professional practice and healthcare outcomes (update). Cochrane Database Syst Rev 2013 Mar 28;2013(3):CD002213. [doi: 10.1002/14651858.CD002213.pub3] [Medline: 23543515]
10. Thistlethwaite J. Interprofessional education: a review of context, learning and the research agenda. Med Educ 2012 Jan;46(1):58-70. [doi: 10.1111/j.1365-2923.2011.04143.x] [Medline: 22150197]
11. Framework for action on interprofessional education & collaborative practice. : World Health Organization; 2010 URL: https://iris.who.int/server/api/core/bitstreams/d743ea4e-8c14-493c-b3a5-1a0022e6ce54/content [accessed 2026-07-08]
12. Hu L, Argus G, Pineda RC, MacAskill W, Martin P. The role of artificial intelligence in enhancing interprofessional education and collaborative practice: a mixed methods scoping review. J Interprof Care 2026;40(2):390-399. [doi: 10.1080/13561820.2025.2576241] [Medline: 41347741]
13. Guraya SY. Transforming simulation in healthcare to enhance interprofessional collaboration leveraging big data analytics and artificial intelligence. BMC Med Educ 2024 Aug 28;24(1):941. [doi: 10.1186/s12909-024-05916-y] [Medline: 39198809]
14. Lave J, Wenger E. Situated Learning: Legitimate Peripheral Participation: Cambridge University Press; 1991.
15. Wenger E. Communities of Practice: Learning, Meaning, and Identity: Cambridge University Press; 1998.
16. Hutchins E. Cognition in the Wild: MIT Press; 1995.
17. Boyle JG, Walters MR, Jamieson S, Durning SJ. Distributed cognition: theoretical insights and practical applications to health professions education: AMEE Guide No. 159. Med Teach 2023 Dec;45(12):1323-1333. [doi: 10.1080/0142159X.2023.2190479] [Medline: 37043405]
18. Frank JR, Snell LS, Cate OT, et al. Competency-based medical education: theory to practice. Med Teach 2010;32(8):638-645. [doi: 10.3109/0142159X.2010.501190] [Medline: 20662574]
19. ten Cate O. Entrustability of professional activities and competency-based training. Med Educ 2005 Dec;39(12):1176-1177. [doi: 10.1111/j.1365-2929.2005.02341.x] [Medline: 16313574]
20. IPEC core competencies for interprofessional collaborative practice: version 3. : Interprofessional Education Collaborative; 2023 URL: https://www.ipecollaborative.org/assets/core-competencies/IPEC_Core_Competencies_Version_3_2023.pdf [accessed 2026-07-08]
21. Adams R, Henry KE, Sridharan A, et al. Prospective, multi-site study of patient outcomes after implementation of the TREWS machine learning-based early warning system for sepsis. Nat Med 2022 Jul;28(7):1455-1460. [doi: 10.1038/s41591-022-01894-0] [Medline: 35864252]
22. Boussina A, Shashikumar SP, Malhotra A, et al. Impact of a deep learning sepsis prediction model on quality of care and survival. NPJ Digit Med 2024 Jan 23;7(1):14. [doi: 10.1038/s41746-023-00986-6] [Medline: 38263386]
23. Wong A, Otles E, Donnelly JP, et al. External validation of a widely implemented proprietary sepsis prediction model in hospitalized patients. JAMA Intern Med 2021 Aug 1;181(8):1065-1070. [doi: 10.1001/jamainternmed.2021.2626] [Medline: 34152373]
24. Parasuraman R, Riley V. Humans and automation: use, misuse, disuse, abuse. Hum Factors 1997 Jun;39(2):230-253. [doi: 10.1518/001872097778543886]
25. Artificial intelligence risk management framework (AI RMF 1.0). : National Institute of Standards and Technology (NIST); 2023 URL: https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.100-1.pdf [accessed 2026-07-08]
26. Ethics and governance of artificial intelligence for health: guidance on large multi-modal models. : World Health Organization; 2024 URL: https://iris.who.int/server/api/core/bitstreams/e9e62c65-6045-481e-bd04-20e206bc5039/content [accessed 2026-07-08]
27. Goddard K, Roudsari A, Wyatt JC. Automation bias: a systematic review of frequency, effect mediators, and mitigators. J Am Med Inform Assoc 2012;19(1):121-127. [doi: 10.1136/amiajnl-2011-000089] [Medline: 21685142]
28. Clancy CM, Tornberg DN. TeamSTEPPS: assuring optimal teamwork in clinical settings. Am J Med Qual 2007;22(3):214-217. [doi: 10.1177/1062860607300616] [Medline: 17485563]
29. Rao VM, Hla M, Moor M, et al. Multimodal generative AI for medical image interpretation. Nature 2025 Mar;639(8056):888-896. [doi: 10.1038/s41586-025-08675-y] [Medline: 40140592]

XSL•FO
RenderX

30. Karampinis E, Bozi Tzetzi DA, Pappa G, et al. Use of a large language model as a dermatology case narrator: exploring the dynamics of a chatbot as an educational tool in dermatology. JMIR Dermatol 2025 Sep 16;8:e72058. [doi: 10.2196/72058] [Medline: 40957071]
31. Obermeyer Z, Powers B, Vogeli C, Mullainathan S. Dissecting racial bias in an algorithm used to manage the health of populations. Science 2019 Oct 25;366(6464):447-453. [doi: 10.1126/science.aax2342] [Medline: 31649194]
32. Amann J, Blasimme A, Vayena E, Frey D, Madai VI, Precise4Q consortium. Explainability for artificial intelligence in healthcare: a multidisciplinary perspective. BMC Med Inform Decis Mak 2020 Nov 30;20(1):310. [doi: 10.1186/s12911-020-01332-6] [Medline: 33256715]
33. Kirkpatrick DL, Kirkpatrick JD. Evaluating Training Programs: The Four Levels, 3rd edition: Berrett-Koehler Publishers; 2006.
34. Cook DA, Hatala R, Brydges R, et al. Technology-enhanced simulation for health professions education: a systematic review and meta-analysis. JAMA 2011 Sep 7;306(9):978-988. [doi: 10.1001/jama.2011.1234] [Medline: 21900138]
35. McGaghie WC, Issenberg SB, Cohen ER, Barsuk JH, Wayne DB. Does simulation-based medical education with deliberate practice yield better results than traditional clinical education? A meta-analytic comparative review of the evidence. Acad Med 2011 Jun;86(6):706-711. [doi: 10.1097/ACM.0b013e318217e119] [Medline: 21512370]
36. Liu X, Cruz Rivera S, Moher D, Calvert MJ, Denniston AK, SPIRIT-AI and CONSORT-AI Working Group. Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI extension. Lancet Digit Health 2020 Oct;2(10):e537-e548. [doi: 10.1016/S2589-7500(20)30218-1] [Medline: 33328048]
37. Vasey B, Nagendran M, Campbell B, et al. Reporting guideline for the early stage clinical evaluation of decision support systems driven by artificial intelligence: DECIDE-AI. BMJ 2022 May 18;377:e070904. [doi: 10.1136/bmj-2022-070904] [Medline: 35584845]
38. Richardson JP, Smith C, Curtis S, et al. Patient apprehensions about the use of artificial intelligence in healthcare. NPJ Digit Med 2021 Sep 21;4(1):140. [doi: 10.1038/s41746-021-00509-1] [Medline: 34548621]
39. Cacciamani GE, Chu TN, Sanford DI, et al. PRISMA AI reporting guidelines for systematic reviews and meta-analyses on AI in healthcare. Nat Med 2023 Jan;29(1):14-15. [doi: 10.1038/s41591-022-02139-w] [Medline: 36646804]

Abbreviations

AHRQ: Agency for Healthcare Research and Quality
AI-IPC: AI-expanded interprofessional collaboration
AI-IPE: AI-expanded interprofessional education
CBME: competency-based medical education
CONSORT-AI: Consolidated Standards of Reporting Trials–AI Extension
DECIDE-AI: Developmental and Exploratory Clinical Investigations of Decision Support Systems Driven by AI
IPC: interprofessional collaboration
IPE: interprofessional education
IPEC: Interprofessional Education Collaborative
SPIRIT-AI: Standard Protocol Items: Recommendations for Interventional Trials–AI Extension
TeamSTEPPS: Team Strategies and Tools to Enhance Performance and Patient Safety
TREWS: Targeted Real-Time Early Warning System

Edited by A Stone; submitted 27.Feb.2026; peer-reviewed by E Karampinis, J Kim, K Blondon, L Lianov, M Zubairi; revised version received 17.Jul.2026; accepted 29.Jul.2026; published 14.Aug.2026.

Please cite as:
Yuh C, Cho CH, Lee YM
The AI Colleague: Reimagining Interprofessional Collaboration and Education in Medicine
JMIR Med Educ 2026;12:e94034
URL: https://mededu.jmir.org/2026/1/e94034
doi:10.2196/94034

© Chisung Yuh, Chul-Hyun Cho, Young-Mee Lee. Originally published in JMIR Medical Education (https://mededu.jmir.org), 14.Aug.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Education, is properly cited. The complete bibliographic information, a link to the original publication on https://mededu.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Beyond Pattern Recognition: Call for Functionally Aware AI for Anatomical Illustration

Agata Maria Kawalec-Rutkowska, MD, PhD; Marian Simka, MD, PhD

Department of Anatomy, Institute of Medical Sciences, University of Opole, Oleska 48, Opole, Opole Voivodeship, Poland

Corresponding Author:
Marian Simka, MD, PhD
Department of Anatomy, Institute of Medical Sciences, University of Opole, Oleska 48, Opole, Opole Voivodeship, Poland

Abstract

This paper critically assesses the role of generative AI in anatomical illustration, identifying fundamental barriers that currently preclude AI from replacing human medical illustrators. Despite the promise of unprecedented efficiency, contemporary models exhibit persistent anatomical inaccuracies and "hallucinations" of nonexistent structures—flaws stemming from statistical pattern-matching rather than genuine anatomical understanding. These systems further lack pedagogical intent, clinical context, and the capacity for deliberate visual judgment, while raising unresolved ethical and copyright concerns regarding training data. Although a specialized AI for this purpose is theoretically feasible, its development as a standalone goal remains economically nonviable given the niche nature of the profession. Rather than replacing human illustrators, AI's future role will be augmentative, with the requisite anatomical intelligence likely emerging as a byproduct of broader advances in clinical applications such as surgical planning and personalized medicine. For AI-generated imagery to become educationally and clinically reliable, it will require rigorous human supervision, curated gold standard datasets, and a foundation of genuine anatomical comprehension.

(JMIR Med Educ 2026;12:e92700) doi:10.2196/92700

KEYWORDS

artificial intelligence; AI; anatomical illustration; medical education; spatial reasoning deficit; intellectual property

Introduction

Since the pioneering work of Renaissance masters like Leonardo da Vinci and Andreas Vesalius, anatomical illustration has been a cornerstone of medical science and education. Da Vinci's meticulous, fluid sketches and Vesalius' groundbreaking, systematic representations in *De Humani Corporis Fabrica* did more than just document the human form; they transformed our understanding of it, establishing a visual language that bridged empirical observation with pedagogical clarity. For centuries, this tradition of hand-rendered accuracy has been indispensable, allowing generations of students to visualize the intricate architecture of the body [1,2].

Today, AI emerges as a potential new frontier, promising to generate intricate anatomical models with unprecedented speed and scale. The seamless integration of AI-powered anatomical illustrations would significantly facilitate medical education at both the undergraduate and postgraduate levels. In particular, the ability to present students with high-quality, error-free drawings that accurately depict anatomical variants would greatly enhance instruction across many specialties [3,4]. Contemporary textbooks and atlases typically illustrate only "standard" anatomy; to showcase clinically relevant variants, educators must search for them in scattered articles—a time-consuming process often complicated by authorship restrictions. Moreover, these resources often underrepresent diverse populations, with illustrations disproportionately reflecting a narrow demographic—most commonly young adult males of a single ethnic background—thereby limiting learners' preparedness for treating patients of different sexes, ethnicities, and age groups [5,6]. Similarly, existing anatomical illustrations usually present a structure from only one or at most a few perspectives, making it difficult to augment clinical reasoning with visuals that fully capture spatial relationships. Such illustrations—those that emphasize anatomical variability and spatial relationships—are not easily accessible today. The high cost and time required to create such educational tools on demand, if a human illustrator were the author, put them out of reach for the majority of medical universities. AI-powered illustration tools could overcome these limitations by generating variant-rich, multiperspective, and demographically inclusive imagery on demand, thereby bridging the gap between anatomical knowledge and clinical application. However, this promise is tempered by significant barriers.

This paper presents an expert viewpoint, based primarily on the authors' practical experiences with AI-generated anatomical artworks in undergraduate anatomical courses, as well as on current published evidence regarding this topic. To inform this perspective, we conducted a literature search and performed a critical analysis of the possible sources of AI incapabilities in this domain. In addition, we went beyond formal analysis by drawing on our personal experiences in teaching anatomy—specifically, guiding students to transform textbook

XSL•FO
RenderX

and atlas knowledge into clinical reasoning [7]. This paper will therefore examine the critical problems encountered by current AI applications in generating anatomical illustrations and assess the feasibility of overcoming these challenges in the future.

Current Limitations of AI in Anatomical Illustration

Problems Related to Current Generative AI Capabilities

Despite the promising potential of AI, its application in anatomical illustration is fraught with significant challenges that currently limit its reliability for precise medical and educational use. Current generative AI systems are fundamentally unsuited to replace human medical illustrators in creating educational anatomical images due to a core disconnect: they operate on statistical patterns, not conceptual understanding. Whether future systems might overcome this limitation remains an open question. Models like DALL-E and Midjourney are sophisticated pattern-matching systems that associate text prompts with visual data, but they possess no knowledge of functional anatomy, 3D spatial relationships, or physiological logic. In essence, current AI models generate a visually compelling collage of patterns, rather than a coherent, functional anatomical system [8-13]. As demonstrated by Moin et al [8] in their assessment of corneal transplant illustrations and by Davis et al [9] in a neurosurgical context, AI models frequently generate images with critical errors. These can range from the misplacement of structures and incorrect spatial relationships to the fabrication of entirely nonexistent anatomical features, a phenomenon known as "hallucination." For instance, Kumar et al [10] highlighted how AI struggled to accurately depict the nuanced pathophysiology of conditions like Horner syndrome, producing visually compelling but scientifically incorrect images. This inherent unpredictability poses a direct threat to the foundational principle of accuracy in medical science [11]. Importantly, a hallucinated blood vessel or an incorrectly positioned nerve is not a stylistic choice. It is a critical failure that can mislead a student or misinform a surgeon [11,14-19]. This AI's inability to create accurate and educational images stems from a problem more fundamental than a simple lack of training data. Critically, medical illustration is not merely the photographic reproduction of an anatomical specimen. This distinction is evident when comparing illustrated anatomical atlases to photographic ones; the former are often superior for education, even at the cost of photorealistic detail. In a masterful medical illustration, such as those by Frank H Netter, the artist makes deliberate pedagogical choices: key structures are emphasized and clarified, while less relevant details are subdued or sketched. This hierarchy of information, designed to guide the viewer's eye and understanding, is the core of its educational value [20]. This essential, goal-oriented design currently cannot be achieved without the illustrator's deep, firsthand experience derived from real contact with the human body—both in the dissection lab and the operating room—forging a cognitive link between form, function, and clinical relevance. Current AI models, trained on pixels and patterns, still lack the clinical judgment and intentionality to make these decisions reliably. They cannot yet discern what is important to teach without expert guidance.

While current generative AI is unreliable for anatomical illustration, the creation of a specialized, accurate tool is not beyond its future capabilities. The significant barriers hindering AI's adoption in anatomical illustration stand in sharp contrast to its rapid and transformative integration into other creative and technical fields. In domains like marketing, conceptual art, and graphic design, AI image generators have become powerful, widely accepted tools. In these areas, an AI-generated image is valued for its aesthetic impact and its ability to rapidly visualize a mood or a concept. This conceptual flexibility allows human artists to use AI as a "copilot" for brainstorming and rapid prototyping, dramatically speeding up workflows [21-25].

This stands in direct opposition to the nonnegotiable demands of medical education. As emphasized by Zoltie [26] and Adams and Erolin [27], a human medical illustrator does not merely replicate anatomy but curates, emphasizes, and simplifies complex information to guide a learner's focus and enhance comprehension. While generative AI tools lack intrinsic intentionality, they are nonetheless capable of producing goal-directed outputs. A similar concern arose in natural language generation, where such goal-directed capabilities were initially thought to be beyond AI—an assumption later proven incorrect. In the case of anatomical art, similarly pedagogically structured outputs are achievable. Thus, the current limitations of AI-generated anatomical illustrations appear to stem less from an inherent inability to exercise judgment and more from the absence of sufficiently grounded anatomical, spatial, and clinical data with which to train these models. This deficiency is exacerbated by the nature of traditional medical knowledge itself. Current anatomical textbooks and atlases, at their core, provide 1D (textual) or 2D (illustrational, cross-sectional) information [7]. Feeding this existing corpus to a standard AI in its current form would be inefficient, as it would risk perpetuating the same 2D pattern-matching limitations, making it difficult to infer the true spatial relationships and functional dynamics of living anatomy. Of note, even the most accurate and high-quality AI-generated illustrations cannot fully replace education based on cadaver dissections and the study of cadaveric prosection specimens, which remain essential for developing 3D spatial understanding and tactile appreciation of the human body [28,29].

Problems Related to Ethical and Intellectual Property Issues

Beyond factual errors, significant ethical and legal dilemmas remain unresolved. The training of generative AI models relies on vast datasets of existing images, often scraped from the internet without the explicit consent of the original creators. As noted by Cornwall et al [30] and Jaillant and Aske [31], this raises profound questions about copyright infringement and the intellectual property rights of medical illustrators whose work is used without attribution or compensation. Furthermore, the "black box" nature of many AI systems, as discussed by Chen et al [14], creates a problem of explainability. If an AI generates an erroneous illustration, it is often impossible to trace the

XSL•FO
RenderX

logical pathway that led to the mistake, undermining scientific accountability [32].

Similarly, the materials used to train AI systems must meet rigorous ethical standards: body donors should provide explicit written consent for the use of their bodies in AI training, and the same principle should apply to medical images drawn from hospital datasets such as computed tomography (CT) and magnetic resonance imaging (MRI) scans [33]. Furthermore, existing published medical illustrations must be protected from unauthorized use for AI training; alternatively, their authors or copyright holders should receive appropriate financial compensation for such use [32].

A Path Toward AI-Driven Anatomical Illustration

Problems to Overcome

Significant technical and logistical hurdles should be overcome to move beyond general-purpose models. The most promising solution lies in developing highly specialized "expert AIs" trained not on the chaotic internet, but on meticulously curated, gold standard datasets. These datasets would comprise verified sources like textbook illustrations (eg, Netter's illustrations), labeled cadaver photos, and annotated 3D models, effectively teaching the AI a prior example of correct anatomy and eliminating its current tendency to create inaccurate averages. A practical implementation of this is fine-tuning existing powerful models on this specialized data, steering their inherent capability for rendering and style toward anatomical precision. Even more powerful could be a hybrid approach, where AI would be integrated with a rule-based 3D anatomy atlas. In this model, the software would generate a perfectly accurate 3D structure, and the AI's role would be limited to applying requested styles, angles, and lighting—bypassing its weakness in 3D reasoning and leveraging its strength in visualization. This development cycle could be refined through reinforcement learning from human feedback, where expert illustrators and surgeons would continuously rate outputs, training the model to prioritize accuracy. These challenges are not trivial. They involve the immense cost and effort of dataset curation, navigating copyright issues, and the computational expense of training. Yet, this is not merely theoretical. Research institutions and medical technology startups are already actively fine-tuning models on biomedical data—including 3D images, which are similarly structured to anatomical illustrations. Such attempts are proliferating across clinical medicine, with numerous examples where AI is being adapted to tackle complex diagnostic and planning tasks. Current research is actively bridging the gap between AI's pattern-matching prowess and the nuanced demands of clinical medicine [34-38]. One pertinent example concerns management of acute stroke, specifically the assessment of the penumbra—the brain tissue at risk of irreversible ischemic damage. While algorithm-based software, such as the rapid processing of perfusion and diffusion (RAPID), remains the gold standard in stroke management [39], a number of companies are now developing AI-driven tools for this purpose. Although these attempts are still suboptimal compared to established solutions, significant progress is being made in this area. Of note, financial barriers are likely of minor significance here, as existing software for stroke assessment is already quite costly, and the end users—hospitals—are eager to pay for solutions that can improve patient outcomes and streamline critical care workflows [40-42].

AI-Driven Anatomical Illustration as Part of a Bigger Task

While the prospect of an AI capable of generating accurate anatomical illustrations is not beyond the realm of possibility, the standalone economic case for developing such a tool is weak. The profession of medical illustration is highly specialized and niche, meaning that the market for a tool designed solely for this purpose is likely too small to justify the immense cost and effort required to create a specialized, gold standard dataset and model. The necessary capabilities for anatomical illustration are far more likely to be acquired by AI as a byproduct of solving larger, more financially viable medical challenges. We are already witnessing this foundational work in other domains. In radiology [43-46] and epidemiology [47-49], AI excels at finding hidden patterns in vast datasets—whether tracking disease spread or identifying a tumor in a CT scan—tasks where its pattern-recognition prowess offers a clear advantage. Yet, many other medical specializations remain beyond AI's current reach precisely because they require a synthesis of multidimensional knowledge. Current AI systems cannot yet replicate this integration, in part because they lack a fundamental, 3D understanding of the human body.

A successful approach would require a paradigm shift, not merely more data. Even the ambitious inclusion of massive, real-world patient data from national-scale medical repositories, as recently attempted in China [50], may yield diminishing returns. Such approaches primarily offer more data, not deeper comprehension. The true breakthrough will require moving beyond mere data correlation to building a foundational, causal model of the human body. This necessitates an AI that does not just process information but actively "learns" human anatomy in 4D—understanding the dynamic, 3D relationships between structures and their integrated function in health and disease. Without this core anatomical intelligence, AI's diagnostic and therapeutic insights will remain reactive, confined to patterns of the past. To excel in these areas, an AI would need to be built upon a foundational, interactive 3D model of human anatomy—a digital twin that understands biomechanics, physiological pathways, and spatial relationships. It would learn from dynamic sources, like surgical videos, real-time sensor data, and 3D medical scans, moving beyond static textbooks and atlases. Creating such a model is a monumental task, but one with a compelling economic and clinical rationale that a niche illustration tool would lack. Once this robust, functionally aware anatomical intelligence exists, the ability to generate correct and educational illustrations could emerge as a natural byproduct. Consequently, the future of AI in anatomical art is inextricably linked to its evolution in core clinical reasoning.

Conclusions

Despite its promise, AI's role as an anatomical illustrator is currently impeded by fundamental obstacles: persistent

XSL•FO
RenderX

anatomical inaccuracies and hallucinations, unresolved ethical and copyright issues surrounding training data, and a lack of genuine 3D and pedagogical understanding—a deficiency rooted in the fact that current AI tools lack intrinsic intentionality, operating on statistical patterns rather than deliberate educational or spatial reasoning. Overcoming these barriers will require a deliberate shift from general-purpose models toward specialized, expert-driven systems trained on curated, gold standard datasets, potentially integrated with rule-based 3D anatomy platforms. In the future, AI is unlikely to replace human medical illustrators but will instead serve as a sophisticated tool that augments their expertise. Such capabilities will likely emerge as a byproduct of AI's broader evolution in clinical domains like surgical planning and personalized medicine, where robust anatomical understanding is a prerequisite.

Acknowledgments

During the preparation of this work, the authors used the DeepSeek-V3 AI tool as a writing editor, particularly for refining scientific language and structuring the findings into a cohesive narrative in order to enhance the paper's clarity. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the content of the published article.

Funding

The authors received funding for this study from the University of Opole.

Conflicts of Interest

None declared.

References

1. Kemp M. Style and non-style in anatomical illustration: from Renaissance humanism to Henry Gray. J Anat 2010 Feb;216(2):192-208. [doi: 10.1111/j.1469-7580.2009.01181.x] [Medline: 20447244]
2. Netter FM, Friedlaender GE. Frank H. Netter MD and a brief history of medical illustration. Clin Orthop Relat Res 2014 Jan 17;472(3):812-819. [doi: 10.1007/s11999-013-3459-8] [Medline: 24435714]
3. Sallam M. ChatGPT utility in healthcare education, research, and practice: systematic review on the promising perspectives and valid concerns. Healthcare (Basel) 2023 Mar 19;11(6):887. [doi: 10.3390/healthcare11060887] [Medline: 36981544]
4. Kim TW. Application of artificial intelligence chatbots, including ChatGPT, in education, scholarly work, programming, and content generation and its prospects: a narrative review. J Educ Eval Health Prof 2023;20:38. [doi: 10.3352/jeehp.2023.20.38] [Medline: 38148495]
5. Beresheim A, Zepeda D, Pharel M, Soy T, Wilson AB, Ferrigno C. Anatomy's missing faces: an assessment of representation gaps in atlas and textbook imagery. Anat Sci Educ 2024 Jul;17(5):1055-1070. [doi: 10.1002/ase.2432] [Medline: 38695348]
6. Kimmorley AL, Reynolds MS, Gregory LS. Anatomical variation is the norm: a novel curriculum framework. Anat Sci Educ 2024 Jul;17(5):1038-1054. [doi: 10.1002/ase.2428] [Medline: 38654399]
7. Simka M. Teaching anatomy of the lower extremity veins: educational challenges. Vasc Invest Ther 2025 Jan;8(1):1-6. [doi: 10.4103/vit.vit_34_24]
8. Moin KA, Nasir AA, Petroff DJ, et al. Assessment of generative artificial intelligence (AI) models in creating medical illustrations for various corneal transplant procedures. Cureus 2024 Aug 26;16(8):e67833. [doi: 10.7759/cureus.67833] [Medline: 39328681]
9. Davis P, Napole A, Reddy C, et al. Systematic evaluation of AI-based text-to-image models for generating medical illustrations in neurosurgery: a multi-stage comparative study. Clin Neurol Neurosurg 2025 Oct;257:109039. [doi: 10.1016/j.clineuro.2025.109039] [Medline: 40614319]
10. Kumar A, Burr P, Young TM. Using AI text-to-image generation to create novel illustrations for medical education: current limitations as illustrated by hypothyroidism and Horner syndrome. JMIR Med Educ 2024 Feb 22;10:e52155. [doi: 10.2196/52155] [Medline: 38386400]
11. Buzzaccarini G, Degliuomini RS, Borin M, et al. The promise and pitfalls of AI-generated anatomical images: evaluating Midjourney for aesthetic surgery applications. Aesth Plast Surg 2024 May;48(9):1874-1883. [doi: 10.1007/s00266-023-03826-w] [Medline: 38238569]
12. Jong S, Dihan QA, Khodeiry MM, Alzein A, Scelfo C, Elhusseiny AM. Evaluating text-to-image generation in pediatric ophthalmology. J Pediatr Ophthalmol Strabismus 2026 May;63(3):186-192. [doi: 10.3928/01913913-20250724-03] [Medline: 41001819]
13. Haider SA, Prabha S, Gomez-Cabello CA, et al. A validity analysis of text-to-image generative artificial intelligence models for craniofacial anatomy illustration. J Clin Med 2025 Mar 21;14(7):2136. [doi: 10.3390/jcm14072136] [Medline: 40217587]
14. Chen H, Gomez C, Huang CM, Unberath M. Explainable medical imaging AI needs human-centered design: guidelines and evidence from a systematic review. NPJ Digit Med 2022 Oct 19;5(1):156. [doi: 10.1038/s41746-022-00699-2] [Medline: 36261476]

XSL•FO
RenderX

15. Eldesoqui M, Albadawi EA, AlQumaizi KI, Radwan MNM, Ebrahim HA, Elsaid MAE. Assessing the anatomical accuracy of AI-generated medical illustrations: a comparative study of text-to-image generator tools in anatomy education. Clin Anat 2025 Sep;38(6):712-717. [doi: 10.1002/ca.70002] [Medline: 40635255]
16. Kim C, Gadgil SU, DeGrave AJ, et al. Transparent medical image AI via an image-text foundation model grounded in medical literature. Nat Med 2024 Apr;30(4):1154-1165. [doi: 10.1038/s41591-024-02887-x] [Medline: 38627560]
17. Negahdaripour M, Keshavarzi MH, Amini M. Enhancing medical education with AI-generated visual art: benefits and challenges. Iran J Med Sci 2025 Feb 1;50(2):129-131. [doi: 10.30476/ijms.2024.104198.3777] [Medline: 40026297]
18. Noel GPJC. Evaluating AI-powered text-to-image generators for anatomical illustration: a comparative study. Anat Sci Educ 2024 Jul;17(5):979-983. [doi: 10.1002/ase.2336] [Medline: 37694692]
19. Paslı B, Yücedağ Gündoğdu H. Text to image generators for anatomical illustrations: potential and limitations. Surg Radiol Anat 2025 Sep 27;47(1):215. [doi: 10.1007/s00276-025-03699-5] [Medline: 41015577]
20. Creze M, Ghosh SK. Illustrator-anatomist-clinician triad based pedagogical model in anatomy education: an auxiliary resource amidst lack of hands-on human dissection. Anat Sci Educ 2025 Apr;18(4):386-405. [doi: 10.1002/ase.70018] [Medline: 40077959]
21. Dwivedi YK, Kshetri N, Hughes L, et al. Opinion paper: “so what if ChatGPT wrote it?” multidisciplinary perspectives on opportunities, challenges and implications of generative conversational AI for research, practice and policy. Int J Inf Manage 2023 Aug;71:102642. [doi: 10.1016/j.ijinfomgt.2023.102642]
22. Ooi KB, Tan GWH, Al-Emran M, et al. The potential of generative artificial intelligence across disciplines: perspectives and future directions. J Comput Inf Syst 2025 Jan 2;65(1):76-107. [doi: 10.1080/08874417.2023.2261010]
23. Heigl R. Generative artificial intelligence in creative contexts: a systematic review and future research agenda. Manag Rev Q 2026 Feb;76(1):955-992. [doi: 10.1007/s11301-025-00494-9]
24. Cillo P, Rubera G. Generative AI in innovation and marketing processes: a roadmap of research opportunities. J Acad Mark Sci 2025 May;53(3):684-701. [doi: 10.1007/s11747-024-01044-7]
25. Yu WF. AI as a co-creator and a design material: transforming the design process. Des Stud 2025 Mar;97:101303. [doi: 10.1016/j.destud.2025.101303]
26. Zoltie T. Artificial intelligence in medical photography and illustration: a tool but not a replacement. J Vis Commun Med 2024 Jul;47(3):75-76. [doi: 10.1080/17453054.2025.2473205] [Medline: 40035631]
27. Adams EM, Erolin C. The devil is in the details: developing a modern methodology for detailed medical illustrations. J Vis Commun Med 2021 Jul;44(3):97-116. [doi: 10.1080/17453054.2021.1921566] [Medline: 33988480]
28. Cornwall J, Hildebrandt S, Champney TH, Goodman K. Ethical concerns surrounding artificial intelligence in anatomy education: should AI human body simulations replace donors in the dissection room? Anat Sci Educ 2024 Jul;17(5):937-943. [doi: 10.1002/ase.2335] [Medline: 37750493]
29. Cortese K, Falletta P. Anatomy at the threshold: teaching the human body in a hybrid age. Anat Sci Educ 2025 Nov;18(11):1301-1309. [doi: 10.1002/ase.70121] [Medline: 40916065]
30. Cornwall J, White R, Pennefather P, et al. Legal and ethical considerations around the use of existing illustrations to generate new illustrations in the anatomical sciences. Anat Sci Educ 2025 Mar;18(3):289-300. [doi: 10.1002/ase.70002] [Medline: 39925144]
31. Jaillant L, Aske K. AI and medical images: addressing ethical challenges to provide responsible access to historical medical illustrations. Digit Humanit Q 2024 Jul 20;18(3). [doi: 10.63744/qmk2g3ptxa93]
32. Skulmowski A, Engel-Hermann P. The ethics of erroneous AI-generated scientific figures. Ethics Inf Technol 2025 Apr;27(2):31. [doi: 10.1007/s10676-025-09835-4]
33. Cornwall J, Hildebrandt S, Champney TH, Billings B, Schmitt B, Winkelmann A. IFAA recommendations for the ethical use of anatomical images. Anat Sci Educ 2024 Jan;17(1):7-10. [doi: 10.1002/ase.2349] [Medline: 37919841]
34. Burns ML, Chen SY, Tsai CA, et al. Generative AI costs in large healthcare systems, an example in revenue cycle. NPJ Digit Med 2025 Sep 30;8(1):579. [doi: 10.1038/s41746-025-01971-x] [Medline: 41028226]
35. Tu T, Schaekermann M, Palepu A, et al. Towards conversational diagnostic artificial intelligence. Nature 2025 Jun;642(8067):442-450. [doi: 10.1038/s41586-025-08866-7] [Medline: 40205050]
36. Hulsen T. Explainable artificial intelligence (XAI): concepts and challenges in healthcare. AI 2023;4(3):652-666. [doi: 10.3390/ai4030034]
37. Chen YM, Hsiao TH, Lin CH, Fann YC. Unlocking precision medicine: clinical applications of integrating health records, genetics, and immunology through artificial intelligence. J Biomed Sci 2025 Feb 7;32(1):16. [doi: 10.1186/s12929-024-01110-w] [Medline: 39915780]
38. Chen E, Prakash S, Janapa Reddi V, Kim D, Rajpurkar P. A framework for integrating artificial intelligence for clinical care with continuous therapeutic monitoring. Nat Biomed Eng 2025 Apr;9(4):445-454. [doi: 10.1038/s41551-023-01115-0] [Medline: 37932379]
39. Zhang Y, Song S, Li Z, Huang B, Geng Y, Zhang L. The application of software “rapid processing of perfusion and diffusion” in acute ischemic stroke. Brain Sci 2022 Oct 27;12(11):1451. [doi: 10.3390/brainsci12111451] [Medline: 36358379]

XSL•FO
RenderX

40. Alwood BT, Meyer DM, Ionita C, et al. Multicenter comparison using two AI stroke CT perfusion software packages for determining thrombectomy eligibility. J Stroke Cerebrovasc Dis 2024 Jul;33(7):107750. [doi: 10.1016/j.jstrokecerebrovasdis.2024.107750] [Medline: 38703875]
41. Alqadi MM, Morales Vidal SG. Artificial intelligence in vascular neurology: applications, challenges, and a review of AI tools for stroke imaging, clinical decision making, and outcome prediction models. Curr Neurol Neurosci Rep 2025 May 9;25(1):34. [doi: 10.1007/s11910-025-01422-w] [Medline: 40343612]
42. Slater LA, Ravintharan N, Goergen S, et al. RapidAI compared with human readers of acute stroke imaging for detection of intracranial vessel occlusion. Stroke Vasc Interv Neurol 2024 Jan 23;4(2):e001145. [doi: 10.1161/SVIN.123.001145] [Medline: 41583584]
43. Katal S, York B, Gholamrezanezhad A. AI in radiology: from promise to practice - a guide to effective integration. Eur J Radiol 2024 Dec;181:111798. [doi: 10.1016/j.ejrad.2024.111798] [Medline: 39471551]
44. Kotter E, D'Antonoli TA, Cuocolo R, et al. Guiding AI in radiology: ESR's recommendations for effective implementation of the European AI Act. Insights Imaging 2025 Feb 13;16(1):33. [doi: 10.1186/s13244-025-01905-x] [Medline: 39948192]
45. Obuchowicz R, Lasek J, Wodziński M, Piórkowski A, Strzelecki M, Nurzynska K. Artificial intelligence-empowered radiology-current status and critical review. Diagnostics (Basel) 2025 Jan 24;15(3):282. [doi: 10.3390/diagnostics15030282] [Medline: 39941212]
46. Flory MN, Napel S, Tsai EB. Artificial intelligence in radiology: opportunities and challenges. Semin Ultrasound CT MR 2024 Apr;45(2):152-160. [doi: 10.1053/j.sult.2024.02.004] [Medline: 38403128]
47. Cox LA Jr. An AI assistant to help review and improve causal reasoning in epidemiological documents. Glob Epidemiol 2023 Dec 16;7:100130. [doi: 10.1016/j.gloepi.2023.100130] [Medline: 38188038]
48. Chen S, Yu J, Chamouni S, Wang Y, Li Y. Integrating machine learning and artificial intelligence in life-course epidemiology: pathways to innovative public health solutions. BMC Med 2024 Sep 2;22(1):354. [doi: 10.1186/s12916-024-03566-x] [Medline: 39218895]
49. Gawande MS, Zade N, Kumar P, Gundewar S, Weerarathna IN, Verma P. The role of artificial intelligence in pandemic responses: from epidemiological modeling to vaccine development. Mol Biomed 2025 Jan 3;6(1):1. [doi: 10.1186/s43556-024-00238-3] [Medline: 39747786]
50. Basu M. A Chinese AI tool can manage chronic disease - could it revolutionize health care? Nature 2025 Sep 19. [doi: 10.1038/d41586-025-02362-8] [Medline: 40973751]

Abbreviations

CT: computed tomography
MRI: magnetic resonance imaging
RAPID: rapid processing of perfusion and diffusion

Edited by M Montagna; submitted 02.Feb.2026; peer-reviewed by S Hildebrandt, Y Yu; revised version received 23.Jun.2026; accepted 03.Jul.2026; published 20.Aug.2026.

Please cite as:
Kawalec-Rutkowska AM, Simka M
Beyond Pattern Recognition: Call for Functionally Aware AI for Anatomical Illustration
JMIR Med Educ 2026;12:e92700
URL: https://mededu.jmir.org/2026/1/e92700
doi:10.2196/92700

© Agata Maria Kawalec-Rutkowska, Marian Simka. Originally published in JMIR Medical Education (https://mededu.jmir.org), 20.Aug.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Education, is properly cited. The complete bibliographic information, a link to the original publication on https://mededu.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Promise, Mimicry, and Surveillance: Responsibly Integrating Artificial Intelligence With Socratic Inquiry in Medical Education

José Daniel Sánchez, MD

Facultad de Ciencias de la Salud y Bienestar Humano, Universidad Tecnológica Indoamérica, Machala y Sabanilla, Quito, Pichincha Province, Ecuador

Corresponding Author:
José Daniel Sánchez, MD
Facultad de Ciencias de la Salud y Bienestar Humano, Universidad Tecnológica Indoamérica, Machala y Sabanilla, Quito, Pichincha Province, Ecuador

Abstract

Large language models are increasingly proposed as a way to deliver Socratic dialogue at scale in medical education, offering personalized, always-available, and lower-stakes inquiry that human educators cannot logistically provide. In this viewpoint, I argue that this promise is plausible but unproven and that 2 failure modes deserve more attention than they currently receive. The first is the mimicry trap: a system that fluently generates Socratic-sounding questions can appear to cultivate reasoning while functioning as interactive content delivery. I position this as an educational instance of the ELIZA effect and of the proxy-outcome problem; separating conversational mimicry from genuine metacognitive gain remains a central unresolved evaluation problem. The second is what I term the Panopticon Paradox: the data collection that makes AI tutoring effective may erode the psychological safety that Socratic inquiry requires, pushing learners toward performative rather than authentic engagement, in a manner analogous to gaming behaviors documented in intelligent tutoring systems. I present this second construct as a testable causal model with specified mediators, moderators, and falsifiable predictions rather than as an established finding. Because the evidence base is dominated by proof-of-concept tools, cross-sectional surveys, and short-term evaluations, I argue that AI is complementary rather than a replacement technology, and I propose a 3-pillar framework (governance, curriculum, and faculty development) tied explicitly to the 2 failure modes. I distinguish 4 modalities of delegation and argue that ethical deliberation, emotionally complex communication, and ambiguous clinical judgment cannot presently be recommended for autonomous or primary AI delivery.

(JMIR Med Educ 2026;12:e102958) doi:10.2196/102958

KEYWORDS

AI; medical education; Socratic method; clinical reasoning; large language models; learning analytics; educational ethics

Introduction: The Allure of an Untiring Socrates

Medical education rests on a standing tension: an expanding knowledge base must be transmitted while the critical thinking required for clinical practice is cultivated [1]. The Socratic method has long been the pedagogy of choice for the second task, promoting reasoning through structured, question-driven dialogue that mirrors the analytic uncertainty of clinical work [2,3]. However, authentic Socratic teaching is time-intensive, hard to scale beyond one-on-one interaction, and chronically vulnerable to degrading into “pimping”: rapid-fire, evaluative questioning that humiliates rather than teaches [2].

Against this structural backdrop, large language models (LLMs) are attractive precisely because they appear to relieve the tension between scale and pedagogical quality. They sustain extended dialogue, adapt to a learner’s responses, and do so on demand and in private, at scales human educators cannot achieve [4,5]. Where infrastructure and licensing permit, the marginal cost per additional learner is low relative to faculty time, although acquisition, integration, and oversight costs are not negligible and access is unevenly distributed across health systems [6]. The pitch writes itself: an untiring, patient Socrates for every learner, available at any hour, in multiple languages, with no enrollment limit. Some of this promise has pedagogical substance. Well-designed AI dialogue can reduce evaluation anxiety and provide immediate feedback on reasoning, and learners in several studies report a lower-stakes space in which to reason aloud [7,8]—though whether a given system is experienced as nonjudgmental depends on its design, on how the institution uses its outputs, and on what learners believe happens to their data.

My aim is not to dismiss the promise or to call for abandoning AI in medical education. It is to argue that deployment is outpacing evidence [6,9] and that 2 specific failure modes—the mimicry trap and the Panopticon Paradox—are underexamined precisely because the promise is attractive and the resource pressure is acute. Naming these problems and situating them within existing scholarly traditions is a precondition for responsible integration [10]. This argument is directed at medical educators, curriculum designers, and institutional leaders who decide how and under what safeguards AI tutoring enters medical curricula.

XSL•FO
RenderX

A word on the evidence base: this is a viewpoint, not a systematic review, and the literature was assembled narratively and selectively. I searched PubMed, Scopus, and Google Scholar using terms combining AI, LLMs, and generative AI with medical education, the Socratic method, clinical reasoning, learning analytics, and educational technology governance; the last search was performed in December 2025, with targeted updates through March 2026. I prioritized records published from 2019 onward, supplemented by older sources from which the underlying constructs originated. Selection was purposive, so the evidence cited should be read as illustrative rather than exhaustive; systematic syntheses of this literature are available elsewhere [6,9].

What Authentic Socratic Teaching Requires

Before asking whether AI can emulate Socratic teaching, it is worth being precise about the target because much of the optimism in the field rests on emulating the wrong one. Genuine Socratic dialogue is constructivist: a shared inquiry in which the instructor is a facilitator rather than an information source, using sequenced, probing questions to surface and examine a learner's underlying beliefs and reasoning processes [2,3]. It is not question-asking; it is a deliberate strategy to shift intellectual agency from teacher to learner. It targets analysis, evaluation, and synthesis, and it functions as a metacognitive tool, helping learners recognize the limits of their knowledge and develop self-directed inquiry habits [3]. The psychological contract is explicit: the teacher creates a space in which uncertainty is expected and productive, and questions are invitations to think rather than traps to expose ignorance [2,11].

Because the current evidence bears unevenly on the outcomes claimed for AI tutoring, 4 constructs must be kept distinct throughout. Clinical reasoning is the situated capacity to generate, prioritize, and revise diagnostic and management hypotheses in authentic cases under uncertainty; it is typically assessed through script concordance, key-features testing, or workplace-based observation. Critical thinking is the more general higher-order analysis and evaluation of claims and evidence, usually assessed with domain-general instruments. Metacognition is the capacity to monitor and regulate one's own thinking, including the calibration of confidence to accuracy. Self-directed learning is the autonomous identification of learning needs and the pursuit of resources to meet them. These constructs are correlated but not interchangeable, and the distinction is consequential: a trial that improves self-directed learning scores has not thereby demonstrated an effect on clinical reasoning [7].

The corrupted twin of Socratic teaching, "pimping," superficially resembles it but inverts nearly every element. Where Socratic questions are process-oriented, open-ended, and lead toward synthesis, pimping questions are closed, often deliberately obscure, asked before an audience, and serve to expose ignorance and assert hierarchy [2]. The distinction is psychological rather than stylistic: authentic dialogue creates productive discomfort inside a psychologically safe space, whereas pimping creates fear, shame, and defensive withdrawal. Psychological safety here is not a loose metaphor but a measurable construct, the shared belief that the setting is safe for interpersonal risk-taking such as admitting error, asking for help, or raising a doubt, for which validated scales exist [11]. The prevalence of pimping is not merely a failure of individual character; it is a rational adaptation to structural constraints, namely large cohorts, chronic time pressure, and the simultaneous mandate to teach and to evaluate summatively [2,12]. Given these pressures, the realistic alternative to AI tutoring is often not ideal one-to-one human tutoring but a lecture hall or a pimping session, and that comparison is the strongest argument for AI in medical education [4,5]. Whether AI fills the gap pedagogically is the question to which I now turn.

The Mimicry Trap: Fluency Is Not Pedagogy

A consequential risk in the current literature is to treat the generation of Socratic-sounding questions as equivalent to Socratic teaching; several widely discussed systems are characterized by the dialogue they can produce rather than by any measured effect on reasoning [13,14]. An LLM can produce an adaptive, grammatically sophisticated, and well-ordered sequence of follow-up questions without any of this contributing to the learner's metacognition or reasoning development; a dialogue can be fluent, contextually responsive, and pedagogically inert at the same time [15]. I call this failure mode the mimicry trap: the risk that a system fluent enough to sound Socratic will be mistaken for Socratic teaching, and that measures of dialogue quality—comprehensibility, coherence, adaptivity, and learner satisfaction—will be mistaken for measures of learning.

The construct is not novel, and I do not claim it as such. It is an educational instance of what has been called the ELIZA effect: the tendency to attribute understanding and intent to a system that merely produces human-like output. Weizenbaum's original demonstration showed that users attributed comprehension and even therapeutic concern to a pattern-matching script with no model of meaning whatsoever [16], an attribution later named and analyzed as a general property of human interaction with symbolic systems [17]. It is also a variant of the proxy-outcome problem: when performance on an easily quantified indicator is substituted for the construct of interest, the indicator degrades as a measure precisely because it has become a target [18,19]. In AI-supported Socratic education, dialogue quality is the proxy and durable reasoning is the construct. Distinguishing genuine, durable gains in metacognition and clinical reasoning from conversational mimicry is, I contend, a central and still largely unresolved evaluation problem in this field [15].

The current evidence does not resolve it. Recent systematic reviews and meta-analyses of AI-supported problem-based and case-based learning report encouraging but heterogeneous effects [20,21], while reviews of AI-enabled tools stress that rigorous, comparative, and longitudinal data remain scarce [9]. A systematic review addressing AI-enabled tools and clinical reasoning in undergraduate nursing education concluded that the available studies are few, heterogeneous in design and

XSL•FO
RenderX

outcome measurement, and methodologically limited [22]. The strongest individual signals are real but narrow. A randomized controlled trial of custom GPT facilitation improved self-directed learning scores and assessed critical thinking with a domain-general instrument but did not measure clinical reasoning in authentic cases [7]. A quasi-experimental Socratic AI platform improved self-efficacy but yielded no statistically significant change in critical thinking at a single site over a short interval [8]. ChatGPT-assisted problem-based learning raised theoretical knowledge and mini-clinical evaluation exercise scores in one controlled cohort of urology interns [23]. Named implementations are proliferating—generative teaching assistants [24], Socratic voice assistants [25], customized tutors such as AnatomyGPT and Anatbuddy [26,27], teacher-AI-student models [28], and case companions built explicitly on Socratic dialogue [29]—but these are predominantly pilots, perception surveys, and proof-of-concept systems. Some frequently discussed demonstrations are not peer reviewed: an adaptive Socratic tutor built on clinical pathological conference cases offers an instructive architecture but exists as a preprint without an outcome trial [13], and a blueprint for AI patient simulation using the Socratic method describes the mechanism rather than the measured effect [14]. Cross-sectional studies show high student readiness and a clear training need but cannot establish causation [30,31].

Altogether, these studies justify cautious optimism about engagement, satisfaction, and self-efficacy and considerably less about durable clinical reasoning. Transferable diagnostic reasoning is not entirely unmeasured, but the existing evidence is limited in volume, heterogeneous in how the construct is operationalized, and methodologically weak, with few longitudinal or multisite comparisons [9,22]. Conflating proximal with distal outcomes is how a genuinely promising tool gets oversold, and it is exactly what the mimicry trap predicts.

The Panopticon Paradox: When Datafication Undermines Inquiry

The Paradox Stated

Suppose the mimicry problem was solved and AI tutoring did reliably foster reasoning. A second tension would remain. The capability that makes AI tutoring powerful—continuous logging of learner behavior, fine-grained tracking of comprehension, recording of wrong answers and conceptual stumbles, algorithmic analysis of query patterns—is also the mechanism by which it may undermine the psychological safety that Socratic inquiry requires. I call this the Panopticon Paradox, borrowing Foucault's account of the prison in which inmates, unable to know when they are observed, internalize the possibility of observation, and modify their behavior toward compliance [32]. I offer it as a hypothesis to be tested, not as a validated construct.

The pedagogical promise of AI tutoring is that it dissolves the pimping problem by providing a private space in which learners can fail safely [7,8]. However, the personalization that makes such tutoring effective requires pervasive data collection: a tutor that adapts to evolving misconceptions must store those misconceptions; one that scaffolds must retain a record of what the learner struggled with; one that paces adaptively must log every interaction with timestamps and outcomes [24,33]. These data are genuinely valuable for pedagogy. They are also sensitive because they reveal not only knowledge gaps but also episodic confusion, trial-and-error reasoning, and dead ends that learners may reasonably wish to have forgotten once resolved [33]. If learners come to believe that their practice space feeds institutional models, contributing to grades, permanent profiles, or opaque ranking systems, the incentive structure changes. The deliberate admission of not knowing the essential first move in Socratic dialogue becomes risky [11]. The learner shifts from a participant in inquiry to an actor in a performance.

This prediction is not speculative in its behavioral component. Learners in monitored digital learning environments are well documented to develop strategies that satisfy the system rather than the learning objective, a family of behaviors characterized in the intelligent tutoring literature as gaming the system [34]. Scholarship on learning analytics has similarly argued that continuous predictive modeling reframes students as data points rather than agents, with consequences for vulnerability and consent [33], and qualitative work on AI in higher education documents both dependency and ambivalence among educators asked to adopt these systems [35]. What has not been tested, to my knowledge, is the specific pedagogical claim: that perceived surveillance in an AI tutor degrades reasoning outcomes through a reduction in psychological safety.

Four Constructs That Are Routinely Conflated: Data Security, Confidentiality, Psychological Privacy, and Freedom from Institutional Assessment

Discussions of "privacy" in educational AI collapse 4 distinct constructs, and the paradox described here depends on the difference between them. Data security is the technical protection of stored and transmitted data against unauthorized access [36]. Confidentiality is the set of rules governing who may legitimately access identifiable data and for what purpose [36,37]. Psychological privacy is the learner's subjective sense of an unobserved space in which thought can be provisional; it is a perception, and it can be absent even when security and confidentiality are impeccable [11,33]. Freedom from institutional assessment is a structural property of the system: whether the record of a learning interaction can, in principle, be used in a summative judgment [33,38]. The pedagogical mechanism runs through the last 2, not the first 2. A system can be encrypted, compliant, and access-controlled while still being experienced as a place where mistakes count, and it is that experience, not the cryptography, that determines whether a learner admits confusion [11].

A Testable Model: Causal Sequence

Stated as a causal sequence, the paradox runs: data collection intensity and its visibility to learners increase perceived surveillance [32,33]; perceived surveillance reduces psychological safety [11]; reduced psychological safety reduces authentic inquiry behaviors—help-seeking, admission of uncertainty, exploratory and speculative questioning, tolerance

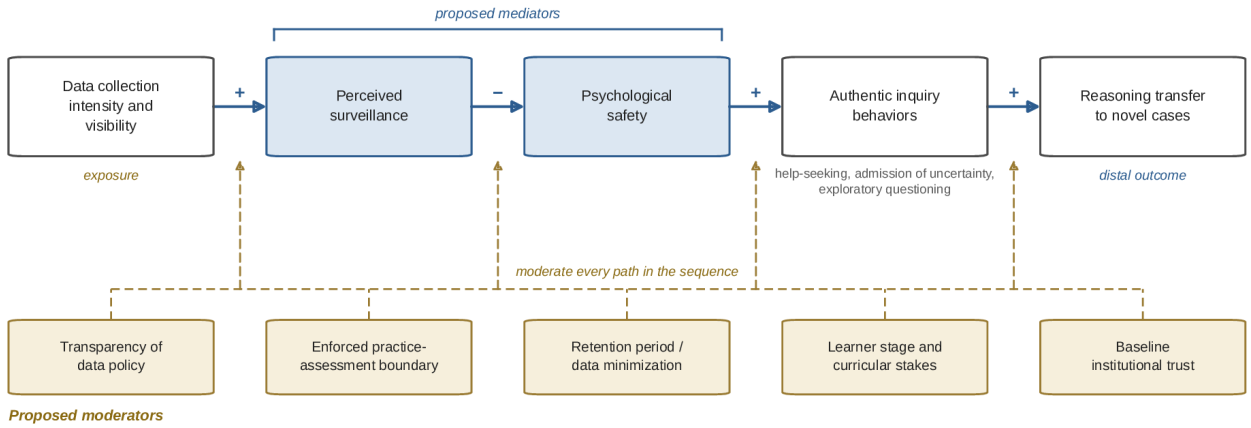
XSL•FO
RenderX

of being wrong in the record [11,34]; and reduced authentic inquiry reduces reasoning transfer to novel cases [3]. Perceived surveillance and psychological safety are the proposed mediators; the model predicts that any effect of tracking on reasoning transfer is fully or partly mediated through them, and that an intervention that reduces perceived surveillance without changing the underlying data flows will still improve outcomes.

Each component is measurable with existing instruments or straightforward adaptations: data collection intensity by system configuration and log volume; perceived surveillance by self-report scales adapted from workplace monitoring research; psychological safety by established team and learning-climate scales [11]; authentic inquiry behaviors by log-derived indicators of help-seeking, hedged or exploratory phrasing, and question initiation, together with gaming-detection approaches developed for intelligent tutoring systems [34]; and reasoning transfer by performance on novel, clinically authentic cases not covered in AI-tutored practice. The proposed moderators are the transparency of the data policy, the presence or absence of an enforced practice-assessment boundary, the retention period, learner stage and the stakes of the surrounding curriculum, and baseline institutional trust.

The model yields falsifiable predictions. First, learners randomized to identical AI tutors with visible tracking will admit uncertainty less often and produce fewer exploratory questions than learners in an untracked condition, with a corresponding difference in transfer performance [11,34]. Second, this difference will attenuate when a practice-assessment boundary is credibly communicated and enforced [33,38]. Third, mediation analysis will show that the effect of tracking on transfer runs through perceived surveillance and psychological safety rather than through engagement volume. Fourth, the effect will be larger in high-stakes curricular contexts and among learners with lower institutional trust [33]. If a well-powered study finds no difference in uncertainty admission between tracked and untracked conditions, the model as stated is wrong, and the governance recommendations that follow from it lose their empirical warrant. Figure 1 summarizes the proposed sequence, its mediators, and its moderators.

Figure 1. Proposed causal sequence of the Panopticon Paradox, with mediators, moderators, and the direction of predicted effects. The model is advanced as a falsifiable hypothesis; the predictions it generates are stated in the text. Signs indicate the predicted direction of effect. The model predicts full or partial mediation: any effect of tracking on reasoning transfer should operate through perceived surveillance and psychological safety rather than through interaction volume.



Why Data Minimization Is a Pedagogical Constraint

None of the available mitigations are complete. Transparent disclosure with opt-in consent, a clean separation between untracked practice and tracked assessment, and governance requiring deletion of individual interaction data each addresses part of the problem while leaving the core tension intact [33,38]. The practical lesson is that data minimization should be justified pedagogically, not only legally. In the United States, the Family Educational Rights and Privacy Act (FERPA) governs education records and grants specific rights over them [37]; the European Union's General Data Protection Regulation imposes a distinct, and in several respects, stricter regime, including an explicit data minimization principle [36]. These regimes are not equivalent, and an institution outside the United States cannot treat FERPA compliance as a benchmark. The pedagogical rationale, by contrast, is jurisdiction-independent: collecting less data reduces perceived surveillance, which is the mechanism through which the paradox operates. An institution in a jurisdiction with no applicable statute has the same pedagogical reason to minimize collection as one bound by the strictest.

Where AI Should Not Be the Primary Modality: A Concrete Case

A serious position on integration must specify not only where AI helps but also where it should not be trusted as a primary modality. Based on current evidence, AI-based Socratic tutoring is most defensible for low-stakes conceptual rehearsal: foundational-science questioning, adaptive case prompting, and formative practice in which errors are inexpensive and human review remains available [4,6,39]. It is least defensible in domains that hinge on human judgment and relationship: ethical deliberation, emotionally complex communication such as breaking bad news, ambiguous clinical judgment under uncertainty, and patient-centered nuance [40,41].

XSL•FO
RenderX

Two limitations underwrite this position, and both require careful statements. The first is factual reliability. LLMs are optimized for linguistic plausibility rather than truth and can produce confident, fluent, and incorrect guidance. Direct evaluations in medical domains bear this out with useful specificity: expert raters scored a general-purpose model's factual accuracy on thoracic anatomy queries significantly below that of a retrieval-augmented tutor constrained to a curated knowledge base [27], and the systematic human evaluation of medical question answering has found that model outputs, even from domain-adapted systems, contain content that raters judged potentially harmful at nontrivial rates [42]. Models have also been shown to reproduce discredited race-based clinical claims, which matter both for accuracy and for equity [43].

The second is empathy, and here, the evidence is less straightforward than the usual claim allows. Blinded raters have judged chatbot responses to patient questions as more empathic than physician responses in a public-forum comparison [44], which should caution against asserting a simple empathy deficit. What that study demonstrates is that text can be rated as empathic; it does not demonstrate that a system holds a model of another person's difficulty, sustains a longitudinal relationship, or bears responsibility for it. The relevant limitation is therefore not the production of empathic language but the absence of the relational and moral substrate that clinical empathy denotes [40,41]. In domains where that substrate is the object of instruction, AI is a poor primary teacher regardless of how warm its prose is. "AI should complement, not replace, the educator" is a familiar slogan; it becomes meaningful only when made this specific.

A concrete case holds the distinction in view. Consider a learner rehearsing a conversation in which a patient with newly diagnosed metastatic disease asks whether treatment is worth pursuing. An LLM can generate a plausible patient turn, respond to the learner's phrasing, and offer feedback that reads as thoughtful [44]. What it cannot do is register that the learner's voice changed, that the pause before the answer was the significant event, or that the learner deflected into prognostic statistics because the emotional content was intolerable; nor can it hold the learner accountable in the way a supervisor who will see them on the ward tomorrow can [40,41]. The same tool used differently is defensible: generating the scenario, providing the learner unlimited low-stakes repetitions, and then handing the transcript to a faculty member for debriefing places the AI where its capabilities lie and keeps the interpretive and relational work with a human [39]. The question is never whether AI touches a domain, but which function within that domain it is asked to perform and who reviews the result.

A 3-Pillar Framework for Responsible Integration: Governance, Curriculum, and Faculty Development

If the goal is human-AI symbiosis rather than substitution, good intentions are insufficient. I propose 3 interdependent pillars, each with specific controls, accountable actors, and evaluation measures [10,39]. Triadic models of technology adoption are not novel; the contribution here is to tie each pillar to the 2 failure modes above rather than applying generic governance templates to a nongeneric tool.

Governance addresses the Panopticon Paradox. Institutions should move beyond point-of-purchase vendor agreements to sustained multidisciplinary oversight by convening bodies that include educators, technical experts able to explain model behavior, data-governance specialists, legal counsel, and learner representatives [10,33]. Before acquisition, institutions should adopt responsible-use principles such as those articulated by the Association of American Medical Colleges [10] and require the documented assessment of the specific educational gap being addressed; a data inventory specifying what is collected, how long it is retained, with whom it is shared, and when it is deleted [36]; system limitations and the thresholds at which a tool is withdrawn; a cost-benefit analysis that accounts for opportunity cost and the value of displaced educator time; and named actors accountable for ongoing audit. Two structural decisions follow from the model in Figure 1. Data minimization means committing as a policy to collect only what the system needs to function, not comprehensive behavioral data for secondary analytics [36,38]. Practice-assessment separation means enforcing a categorical boundary between formative interaction data and data usable for grades, competency judgments, or certification and communicating that boundary to learners in terms they find credible, since it is perceived surveillance, not the data architecture alone, that the model identifies as the operative mediator [11,33].

The curriculum counters the mimicry trap. AI literacy should be a declared, teachable competency taught from the beginning of training: learners should understand the capabilities and limitations of dialogue systems, their propensity for confident factual error, and the mechanisms by which they can appear more competent than they are [10,42]. Learners should also be taught the ELIZA effect and the proxy-outcome problem explicitly, so that they can evaluate claims about AI tutoring, including their own impressions of it [16,18,19]. Implementation should be phased by stakes and cognitive level: Socratic AI tools are well suited to foundational, lower-stakes contexts in preclinical years and early rotations and should shift to a supplementary role, embedded in human-led teaching and followed by faculty debriefing, as stakes rise [39]. Most importantly, assessment must target what matters rather than what is easy to capture. Programs should assess reasoning transfer, that is, application to novel and clinically authentic cases not covered in AI-tutored practice, together with metacognitive development [3,20]. These are harder to measure than satisfaction, which is precisely why the field defaults to satisfaction [9,21].

Faculty development enables both. Educators need systematic upskilling in the critical appraisal of AI tools, in the verification of AI-generated content, and in a shifted pedagogical role that moves from content delivery toward facilitation and debriefing. This shift is not trivial: it requires faculty to model the intellectual humility and ethical reasoning that AI cannot supply and to bridge between AI-mediated practice and authentic clinical reasoning. Recent guidance on clinical supervision of AI use maps directly onto the educational setting, where the

XSL•FO
RenderX

supervisor ensures the tool is used appropriately, interprets outputs correctly, and escalates to human judgment when uncertainty demands it [39]. Research on faculty perceptions consistently finds interest paired with low confidence and limited training [12,45], which makes this pillar the most frequently underfunded and the most likely point of failure.

It is worth distinguishing 4 modalities of AI use, each with a different risk profile [10,39]. Autonomous use, in which AI determines judgments without human oversight, is the highest-risk category and should be avoided. AI as a primary instructional modality, in which learners receive core instruction from AI with human engagement supplementary, is of lower risk but remains problematic for high-stakes reasoning and ethical deliberation [42,43]. Supervised AI-supported practice, in which learners work with AI dialogue under faculty oversight, is substantially of lower risk and aligns with the human-centered model advocated here [39]. Simulated scenarios followed by structured human debriefing are of lowest risk and perhaps of highest value because they preserve the faculty-learner relationship and the interpretive work in which reasoning develops [40,41]. The framework supports the third and fourth modalities, the second cautiously and only in foundational contexts, and the first not at all.

A word is owed to institutions for which this framework may read as written for someone else. The governance apparatus described above presupposes administrative capacity that many schools in low-income and middle-income settings do not have, and the resource pressure that makes AI tutoring attractive is most acute precisely where that capacity is thinnest [6,9]. Two points are as follows. First, the 2 structural decisions that matter most for the Panopticon Paradox, data minimization and practice-assessment separation, are policy choices rather than expensive infrastructure; a school can adopt both in a single faculty-council decision and communicate them to learners in a paragraph [36,38]. Second, dependence on vendors headquartered in other jurisdictions raises questions about where learner data reside and under whose law, which cannot be resolved by pointing to the vendor's compliance with a foreign statute [36,37]. Institutions in these settings have the strongest interest in minimal collection and the weakest leverage to negotiate it.

Limitations

This viewpoint has limitations that bear on how its claims should be weighed. The literature it draws on was selected purposively rather than systematically, and I may have favored work consonant with the critical framing advanced here; no protocol, dual screening, or gray-literature search was undertaken, and none of the studies cited were appraised with a formal risk-of-bias instrument [6]. Both central constructs are conceptual proposals rather than validated models: the mimicry trap reframes an existing problem in measurement [18,19], and the Panopticon Paradox, though stated here in testable form, has not been tested. The framework I propose is offered as a reasoned structure for institutional decision-making, not as an intervention with demonstrated effects on learning outcomes.

The relative strengths and weaknesses that justify this complementary division of labor are summarized in Table 1. Because the claims in such comparisons are frequently asserted rather than demonstrated, each row identifies whether the AI-side claim rests on direct empirical evidence or is a reasoned expectation awaiting test.

XSL•FO
RenderX

Table . Attributes of human-led vs AI-driven Socratic teaching: a context-dependent heuristic, not a meta-analysis or a systematic comparison of effect sizes[a].

Attribute	Human-led Socratic teaching	AI-driven Socratic dialogue	Basis of the AI-side claim
Scalability	Intrinsically limited to small groups or one-to-one dialogue [2]	Scales to large cohorts; marginal cost per additional learner is low once infrastructure and licensing are in place [5]	Reasoned expectation; cost claims are context-dependent and rarely audited against displaced faculty time
Availability	Time-dependent and schedule-dependent [2]	Continuous and on demand where connectivity, licensing, and devices permit [4]	Empirical as a system property; conditional on access, which is unevenly distributed
Personalization	High in dyadic settings; limited at scale [3]	Adapts to learner responses and prior performance within a session, as described in system reports and narrative overviews [24,29,45]	Reported capability rather than direct evidence; adaptation within a Socratic tutoring system has been described rather than directly evaluated, and its effect on learning outcomes is untested
Consistency	Varies with instructor state, expertise, and bias [12]	Configurable and reproducible across sessions and cohorts [14]	Reasoned expectation; consistency is not neutrality, and models carry biases of their own [43]
Psychological safety	Variable: can be high, or humiliating where it degrades into pimping [2]	Potentially high, contingent on data minimization, an enforced practice-assessment boundary, and what learners believe about data use	Hypothesis, not demonstrated; this is the mechanism that the Panopticon Paradox predicts can fail
Empathy and authenticity	Present; reads emotional and social cues and carries relational responsibility [40]	Can generate text that blinded raters judge as highly empathic, without a relational or moral substrate [44]	Empirical for rated text quality; the limitation concerns relationship and accountability, not prose [40,41]
Content accuracy	Generally high when guided by expert knowledge [6]	Variable and prone to plausible-sounding factual error; accuracy improves when the model is constrained to a curated knowledge base [27,42]	Empirical, from the direct expert rating of model outputs in medical domains
Reasoning transfer	The declared goal of the method; assessed through workplace-based observation [3]	Not established; few longitudinal or comparative studies report transfer outcomes [9,22]	Evidence gap rather than a claim in either direction

[a]Entries reflect tendencies reported across the cited literature; variation within each modality is considerable and depends on implementation quality, learner characteristics, and institutional context. The last column states whether the AI-side claim rests on evidence directly demonstrated by the cited source or is a reasoned expectation awaiting test.

A Research Agenda

The priorities follow from the analysis. First, longitudinal, multi-institutional studies that track clinical reasoning over academic years rather than weeks and that determine whether Socratic AI dialogue produces durable gains in diagnostic reasoning and transfer to novel problems are needed [9,22]. Comparative trials should disaggregate the components of AI tutoring by design, separating conversational fluency from reasoning development; without that separation, the field will converge on dialogue quality as a proxy for pedagogical effect [18,19], which is the mimicry trap operating at the level of an entire literature.

Second, equity and algorithmic-bias audits are needed, with attention to generalization across diverse learner populations and low-income and middle-income educational contexts. Most current evidence derives from high-income, English-language, well-resourced schools [6,9]. This is not only an ethical problem but also a scientific one: models trained predominantly on English text and North American norms may perform differently for learners with other linguistic, cultural, and epistemological backgrounds and have been shown to reproduce biased clinical content [43].

Third, and most directly tied to the Panopticon Paradox, the causal model in Figure 1 should be tested. Designs comparing AI-tutored learning in tracked vs untracked conditions, with the measurement of perceived surveillance, psychological safety, help-seeking, uncertainty admission, and reasoning transfer, would either support the mediation pathway or falsify it [11,34]. Governance models—data minimization, practice-assessment separation, and transparency to learners—should be tested as moderators rather than assumed to work [33,38].

Finally, this program should engage philosophers, ethicists, and learners themselves, not only cognitive scientists and educationalists [33,38]. What is the relationship between feeling safe and being safe in a datafied learning environment? What do learners and institutions owe each other in terms of authentic participation and privacy? These are not questions that trials alone can answer.

XSL•FO
RenderX

Conclusions

The case for AI in Socratic medical education is strong enough to take seriously and weak enough to demand restraint. The honest reading of the current evidence base is that AI shows promise in improving engagement, satisfaction, and self-efficacy [7,8,20,21]; that it plausibly complements educators in overstretched systems [4,5]; and that it has not been convincingly shown to build durable, transferable clinical reasoning that persists over longitudinal follow-up [9,22]. The existing studies are few, heterogeneous, and methodologically limited, and the most cited demonstrations are frequently pilots or preprints rather than outcome trials [13,14]. That is not an argument for abandonment. It is an argument for proportionality between what is claimed and what has been shown.

The 2 failure modes I have emphasized are not exotic risks. The mimicry trap is the ordinary consequence of measuring what is easy to measure: a system that produces fluent, adaptive, Socratic-sounding dialogue will generate excellent proxy metrics whether or not it changes how a learner reasons about a patient, and every incentive in procurement, publication, and program evaluation rewards the proxy [18,19]. The Panopticon Paradox is the ordinary consequence of building personalization on observation: the same logs that let a tutor adapt are also the logs that tell a learner their confusion is on record [32,33]. Neither failure requires bad faith. Both are structural, which is why both must be addressed structurally [10,39].

I have tried to state the second construct in a form that can be shown to be wrong. If learners in tracked and untracked conditions admit uncertainty at the same rate, if psychological safety does not mediate the relationship between tracking and reasoning transfer [11], or if a credible practice-assessment boundary makes no difference to either of these, then the Panopticon Paradox is a compelling metaphor with no empirical content, and the governance recommendations that rest on it should be revised accordingly. I would regard such a finding as a contribution because the alternative, deploying surveillance-intensive tutoring at scale on the assumption that observation is pedagogically neutral, is a much more expensive way to find out.

What follows practically is modest and specific. Institutions considering Socratic AI tutoring should ask what educational gap the tool addresses, what data it collects and for how long, whether formative interaction data are structurally walled off from summative judgment, whether learners believe that wall exists, and how the program will detect the difference between fluent dialogue and improved reasoning [10,39]. Educators should treat AI dialogue as a rehearsal space subordinate to human debriefing, not as a substitute teacher. Researchers should measure transfer, not satisfaction [9,22]. AI in Socratic medical education is a beginning, not an end point, and the discipline it demands is the same discipline the Socratic method itself teaches: a willingness to keep asking what we actually know [2,3].

Acknowledgments

The author dedicates this work to the health sciences faculty who continue to teach with responsibility and vocation in demanding circumstances and who use new tools wisely rather than allowing them to displace the essence of teaching. The author would like to express special thanks to "El Bus"—a committed group of faculty members dedicated to providing quality teaching in medical education—for their unwavering support, collegiality, and shared commitment to educational excellence. Their work embodies the humanistic values that guide this research. In accordance with the GAIDET taxonomy for disclosing generative AI involvement in scholarly work, the author declares the following. Generative AI and AI-assisted tools, including ChatGPT (OpenAI), were used for literature discovery (identification of candidate references), for language editing and stylistic revision of author-written text, and for formatting assistance. Generative AI was not used to generate data, conduct analysis, produce original scientific claims, or draft conclusions. All sources were retrieved, read, and verified against the original publications by the author. All conceptual content, argument, interpretation, and final wording are the author's own, and the author takes full responsibility for the integrity and accuracy of the work.

Funding

This work received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Authors' Contributions

This study is solely the work of the author and no external reviewers, collaborators, or third parties contributed to the development of the manuscript before submission. The author takes full responsibility for all content, including the conceptual framework, argumentation, interpretation, data verification, referencing, and final wording. The author also confirms that all sources cited have been retrieved, read, and verified against the original publications.

Conflicts of Interest

None declared.

References

XSL•FO
RenderX

1. Chan KS, Zary N. Applications and challenges of implementing artificial intelligence in medical education: integrative review. JMIR Med Educ 2019 Jun 15;5(1):e13930. [doi: 10.2196/13930] [Medline: 31199295]
2. Oyler DR, Romanelli F. The fact of ignorance: revisiting the Socratic method as a tool for teaching critical thinking. Am J Pharm Educ 2014 Sep 15;78(7):144. [doi: 10.5688/ajpe787144] [Medline: 25258449]
3. Ho YR, Chen BY, Li CM. Thinking more wisely: using the Socratic method to develop critical thinking skills amongst healthcare students. BMC Med Educ 2023 Mar 20;23(1):173. [doi: 10.1186/s12909-023-04134-2] [Medline: 36941597]
4. Preiksaitis C, Rose C. Opportunities, challenges, and future directions of generative artificial intelligence in medical education: scoping review. JMIR Med Educ 2023 Oct 20;9:e48785. [doi: 10.2196/48785] [Medline: 37862079]
5. Khakpaki A. Advancements in artificial intelligence transforming medical education: a comprehensive overview. Med Educ Online 2025 Dec;30(1):2542807. [doi: 10.1080/10872981.2025.2542807] [Medline: 40798935]
6. Gordon M, Daniel M, Ajiboye A, et al. A scoping review of artificial intelligence in medical education: BEME Guide No. 84. Med Teach 2024 Apr;46(4):446-470. [doi: 10.1080/0142159X.2024.2314198] [Medline: 38423127]
7. Shalong W, Yi Z, Bin Z, et al. Enhancing self-directed learning with custom GPT AI facilitation among medical students: a randomized controlled trial. Med Teach 2025 Jul;47(7):1126-1133. [doi: 10.1080/0142159X.2024.2413023] [Medline: 39425996]
8. Li Y, Yorke J, Li J, et al. An innovative Socratic method-based artificial intelligence platform for healthcare education: a quasi-experimental study. Nurse Educ Pract 2026 Mar;92:104770. [doi: 10.1016/j.nepr.2026.104770] [Medline: 41690156]
9. Simoni J, Urtubia-Fernandez J, Mengual E, et al. Artificial intelligence in undergraduate medical education: an updated scoping review. BMC Med Educ 2025 Nov 17;25(1):1609. [doi: 10.1186/s12909-025-08188-2] [Medline: 41250109]
10. Principles for the responsible use of artificial intelligence in and for medical education. Association of American Medical Colleges (AAMC). 2023. URL: https://www.aamc.org/about-us/mission-areas/medical-education/principles-ai-use [accessed 2026-08-22]
11. Edmondson A. Psychological safety and learning behavior in work teams. Adm Sci Q 1999 Jun;44(2):350-383. [doi: 10.2307/2666999]
12. Salih SM. Perceptions of faculty and students about use of artificial intelligence in medical education: a qualitative study. Cureus 2024 Apr;16(4):e57605. [doi: 10.7759/cureus.57605] [Medline: 38707183]
13. Golchini N, Passalacqua E, Vaughn L, Abdulnour REE, Zack T, Finlayson S. Socratic AI: an adaptive tutor for clinical case based learning. medRxiv. Preprint posted online on Jun 23, 2025. [doi: 10.1101/2025.06.22.25329661]
14. Malhotra A, Buller M, Modi K, Pajazetovic K, Wijesinghe DS. Blueprint for constructing an AI-based patient simulation to enhance the integration of foundational and clinical sciences in didactic immunology in a US Doctor of Pharmacy program: a step-by-step prompt engineering and coding toolkit. Pharmacy (Basel) 2025 Mar 1;13(2):36. [doi: 10.3390/pharmacy13020036] [Medline: 40126309]
15. Izquierdo-Condoy JS, Arias-Intriago M, Tello-De-la-Torre A, Busch F, Ortiz-Prado E. Generative artificial intelligence in medical education: enhancing critical thinking or undermining cognitive autonomy? J Med Internet Res 2025 Nov 3;27:e76340. [doi: 10.2196/76340] [Medline: 41183320]
16. Weizenbaum J. ELIZA—a computer program for the study of natural language communication between man and machine. Commun ACM 1966 Jan;9(1):36-45. [doi: 10.1145/365153.365168]
17. Hofstadter DR, Fluid Analogies Research Group. Fluid Concepts and Creative Analogies: Computer Models of the Fundamental Mechanisms of Thought: Basic Books; 1995.
18. Campbell DT. Assessing the impact of planned social change. Eval Program Plann 1979 Jan;2(1):67-90. [doi: 10.1016/0149-7189(79)90048-X]
19. Strathern M. 'Improving ratings': audit in the British university system. Eur Rev 1997 Jul;5(3):305-321. [doi: 10.1002/(SICI)1234-981X(199707)5:3<305::AID-EURO184>3.0.CO;2-4]
20. Wei H, Dai Y, Yuan K, et al. AI-powered problem- and case-based learning in medical and dental education: a systematic review and meta-analysis. Int Dent J 2025 Aug;75(4):100858. [doi: 10.1016/j.identj.2025.100858] [Medline: 40578029]
21. Pham TD, Karunaratne N, Exintaris B, et al. The impact of generative AI on health professional education: a systematic review in the context of student learning. Med Educ 2025 Dec;59(12):1280-1289. [doi: 10.1111/medu.15746] [Medline: 40533396]
22. Jelloul H, Arhoun S, Rmili M, Nafidi Y. Artificial intelligence-enabled educational tools and clinical reasoning in undergraduate nursing education: a systematic review. J Pioneering Med Sci 2026 Mar 5;15(2):192-203. [doi: 10.47310/jpms2026150226]
23. Hui Z, Zewu Z, Jiao H, Yu C. Application of ChatGPT-assisted problem-based learning teaching method in clinical medical education. BMC Med Educ 2025 Jan 11;25(1):50. [doi: 10.1186/s12909-024-06321-1] [Medline: 39799356]
24. Thesen T, Park SH. A generative AI teaching assistant for personalized learning in medical education. NPJ Digit Med 2025 Nov 4;8(1):627. [doi: 10.1038/s41746-025-02022-1] [Medline: 41188616]
25. Li TP, Slocum S, Sahoo A, et al. Socratic Artificial Intelligence Learning (SAIL): the role of a virtual voice assistant in learning orthopedic knowledge. J Surg Educ 2024 Nov;81(11):1655-1666. [doi: 10.1016/j.jsurg.2024.08.006] [Medline: 39288509]

XSL•FO
RenderX

26. Collins BR, Black EW, Rarey KE. Introducing AnatomyGPT: a customized artificial intelligence application for anatomical sciences education. Clin Anat 2024 Sep;37(6):661-669. [doi: 10.1002/ca.24178] [Medline: 38721869]
27. Arun G, Perumal V, Urias F, et al. ChatGPT versus a customized AI chatbot (Anatbuddy) for anatomy education: a comparative pilot study. Anat Sci Educ 2024 Oct;17(7):1396-1405. [doi: 10.1002/ase.2502] [Medline: 39169464]
28. Wang L, Xu Y, Zhang M, Bai R, Xie T. Teaching innovation in a pharmacy course: integration of “questioning-training of comprehensive knowledge application” and a "teacher-AI-student interaction model". BMC Med Educ 2025 Jul 1;25(1):964. [doi: 10.1186/s12909-025-07549-1] [Medline: 40596980]
29. Chastain AP, Roman C, Bogenschutz KM. Student perceptions of a custom artificial intelligence clinical case companion. J Physician Assist Educ 2026 Mar 1;37(1):111-115. [doi: 10.1097/JPA.0000000000000697] [Medline: 40928352]
30. Rezazadeh H, Mahani AM, Salajegheh M. Insights into the future: assessing medical students’ artificial intelligence readiness—a cross-sectional study at Kerman University of Medical Sciences (2022). Health Sci Rep 2025 May;8(5):e70870. [doi: 10.1002/hsr2.70870] [Medline: 40432697]
31. Al-Thani SN, Anjum S, Bhutta ZA, et al. Comparative performance of ChatGPT, Gemini, and final-year emergency medicine clerkship students in answering multiple-choice questions: implications for the use of AI in medical education. Int J Emerg Med 2025 Aug 7;18(1):146. [doi: 10.1186/s12245-025-00949-6] [Medline: 40775272]
32. Foucault M. Discipline and Punish: The Birth of the Prison: Pantheon Books; 1977.
33. Prinsloo P, Slade S. Student vulnerability, agency, and learning analytics: an exploration. J Learn Anal 2016;3(1):159-182. [doi: 10.18608/jla.2016.31.10]
34. Baker RS, Corbett AT, Koedinger KR. Detecting student misuse of intelligent tutoring systems. In: Lester JC, Vicari RM, Paraguacu F, editors. Intelligent Tutoring Systems 7th International Conference, ITS 2004, Maceió, Alagoas, Brazil, August 30 - September 3, 2004, Proceedings: Springer; 2004:531-540. [doi: 10.1007/978-3-540-30139-4_50]
35. Alhur AA, Khlaif ZN, Hamamra B, Hussein E. Paradox of AI in higher education: qualitative inquiry into AI dependency among educators in Palestine. JMIR Med Educ 2025 Sep 15;11:e74947. [doi: 10.2196/74947] [Medline: 40953310]
36. Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation) (Text with EEA relevance). : European Union; 2016 URL: https://eur-lex.europa.eu/eli/reg/2016/679/oj [accessed 2026-08-26]
37. U.S. Department of Education, Student Privacy Policy Office. FERPA. Protecting Student Privacy. URL: https://studentprivacy.ed.gov/ferpa [accessed 2026-08-22]
38. Slade S, Prinsloo P. Learning analytics: ethical issues and dilemmas. Am Behav Sci 2013;57(10):1510-1529. [doi: 10.1177/0002764213479366]
39. Abdulnour REE, Gin B, Boscardin CK. Educational strategies for clinical supervision of artificial intelligence use. N Engl J Med 2025 Aug 21;393(8):786-797. [doi: 10.1056/NEJMra2503232] [Medline: 40834302]
40. Silver JK, Dodurgali MR, Gavini N. Artificial intelligence in medical education and mentoring in rehabilitation medicine. Am J Phys Med Rehabil 2024 Nov 1;103(11):1039-1044. [doi: 10.1097/PHM.0000000000002604] [Medline: 39016292]
41. Weidener L, Fischer M. Artificial intelligence teaching as part of medical education: qualitative analysis of expert interviews. JMIR Med Educ 2023 Apr 24;9:e46428. [doi: 10.2196/46428] [Medline: 36946094]
42. Singhal K, Azizi S, Tu T, et al. Large language models encode clinical knowledge. Nature 2023 Aug;620(7972):172-180. [doi: 10.1038/s41586-023-06291-2] [Medline: 37438534]
43. Omiye JA, Lester JC, Spichak S, Rotemberg V, Daneshjou R. Large language models propagate race-based medicine. NPJ Digit Med 2023 Oct 20;6(1):195. [doi: 10.1038/s41746-023-00939-z] [Medline: 37864012]
44. Ayers JW, Poliak A, Dredze M, et al. Comparing physician and artificial intelligence chatbot responses to patient questions posted to a public social media forum. JAMA Intern Med 2023 Jun 1;183(6):589-596. [doi: 10.1001/jamainternmed.2023.1838] [Medline: 37115527]
45. Narayanan S, Ramakrishnan R, Durairaj E, Das A. Artificial intelligence revolutionizing the field of medical education. Cureus 2023 Nov;15(11):e49604. [doi: 10.7759/cureus.49604] [Medline: 38161821]

Abbreviations

FERPA: Family Educational Rights and Privacy Act
LLM: large language model

XSL•FO
RenderX

Edited by S Brini; submitted 29.May.2026; peer-reviewed by FGH Yap, Y Hu; revised version received 31.Jul.2026; accepted 03.Aug.2026; published 10.Sep.2026.

Please cite as:
Sánchez JD
Promise, Mimicry, and Surveillance: Responsibly Integrating Artificial Intelligence With Socratic Inquiry in Medical Education
JMIR Med Educ 2026;12:e102958
URL: https://mededu.jmir.org/2026/1/e102958
doi:10.2196/102958

© José Daniel Sánchez. Originally published in JMIR Medical Education (https://mededu.jmir.org), 10.Sep.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Education, is properly cited. The complete bibliographic information, a link to the original publication on https://mededu.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Artificial Intelligence in Medical Education: Transformative Potential, Current Applications, and Future Implications

Juan S Izquierdo-Condoy[1], MSc, MD; Marlon Arias-Intriago[1], MD; Laura Montero Corrales[2], PhD; Esteban Ortiz-Prado[1], MSc, MPH, MD, PhD

[1]One Health Research Group, Universidad de Las Américas, Via Nayon S/N, Quito, Ecuador

[2]Escuela de Comunicación, Universidad Latina de Costa Rica, San Jose, Costa Rica

Corresponding Author:
Esteban Ortiz-Prado, MSc, MPH, MD, PhD
One Health Research Group, Universidad de Las Américas, Via Nayon S/N, Quito, Ecuador

Abstract

Artificial intelligence (AI) is increasingly influencing medical education by enabling adaptive learning, AI-assisted assessment, and scalable instructional tools. Natural language processing, machine learning, and generative large language models offer innovative ways to support teaching and learning, yet their integration raises ethical, pedagogical, and infrastructural challenges. This viewpoint article aims to examine the current applications, benefits, and challenges of AI in medical education and propose strategies for responsible and effective integration. AI tools such as chatbots, virtual patients, and intelligent tutoring systems enhance personalized and immersive learning. Automated grading and predictive analytics support efficient evaluations, while AI-assisted writing tools streamline content creation. Despite these advances, concerns persist around data privacy, algorithmic bias, unequal access, and diminished critical thinking. Key solutions include AI literacy training, data oversight, equitable infrastructure, and curriculum reform. The FACETS framework offers 6 dimensions (ie, form, application, context, instructional mode, technology, and the SAMR [substitution, augmentation, modification, redefinition model]) to evaluate AI integration effectively. AI offers substantial opportunities to transform medical education, but its adoption must be ethical, equitable, and pedagogically grounded. Strategic frameworks such as FACETS, combined with institutional governance and cross-sector collaboration, are essential to guide implementation so that AI enhances learning outcomes while preserving the humanistic foundations of medical practice.

(JMIR Med Educ 2026;12:e77127) doi:10.2196/77127

KEYWORDS

artificial intelligence; medical education; machine learning; adaptive learning systems; future implications

Introduction

Artificial intelligence (AI) emerged in the mid-20th century, particularly after the Dartmouth Conference in 1956, as an interdisciplinary field integrating computer science, mathematics, logic, and cognitive science. The first published study on AI in medical education dates back to 1992, although AI has existed much longer than this. Its goal was to simulate human cognitive processes and endow computer systems with human-like abilities, such as reasoning, learning, and decision-making [1]. In recent decades, AI applications have expanded significantly, finding roles in fields such as programming, statistics, preschool education, and university disciplines such as medicine and medical education. However, it was not until the launch of ChatGPT (OpenAI) in 2022 that AI became exponentially popularized, taking on a prominent role in many domains of knowledge. This growth has been driven by technological advances and increasing expectations fueled by globalization and the vast flow of information in today's interconnected world [2,3].

In the field of sciences, including medicine, biomedicine, and related disciplines, the use of AI extends far beyond the traditional teaching-learning process. It now plays a fundamental role in the entire professional life cycle of medical practitioners. This includes preparation, continuous education, and the acquisition of knowledge. The rapid growth of scientific literature and the constant influx of new information push physicians to seek knowledge from diverse sources, aiming to stay as current as possible. These sources go beyond traditional books or scientific articles, incorporating videos, interactive platforms, virtual classes, and other innovative tools that facilitate learning and professional development [1,2].

AI supports both learning and teaching, leveraging technologies such as natural language processing (NLP), machine learning (ML), and generative pretrained transformer architectures. These technologies enable the creation of innovative educational strategies, revolutionary teaching methods, curriculum design, content development, and the evaluation of various academic processes. This makes AI an indispensable tool for medical

XSL•FO
RenderX

education, facilitating both the transmission of knowledge and the continuous improvement of educational systems [4].

The dynamics have changed so rapidly that teachers, educators, students, and trainees are using AI, yet very few have been adequately trained for this purpose. Scientific literature is growing at an unprecedented pace, and while the integration of AI into medical education offers significant benefits, it also presents substantial challenges. One of the most profound critiques involves ethical considerations, the risk of undermining academic integrity, and concerns about students' reliance on AI for assignments, raising critical questions among educators. As Pineda-de-Alcázar [5] suggests, these technological advances raise fundamental questions about the nature of communication between humans and machines, as AI moves closer to interacting with us in ways that mimic human thought and emotional complexity. Within this context, we explore the applicability and evidence for using AI in biomedical teaching practices.

Against this backdrop, this manuscript addresses the following research question: How are distinct AI modalities being deployed in medical education across learning, assessment, administration, and academic content creation, and what ethically grounded strategies best support their responsible integration? Accordingly, our objective is to examine the applications, benefits, and challenges of AI in medical education across these domains and propose strategies for responsible and effective integration.

Current Applications of AI in Medical Education

Overview

AI is rapidly transforming medical education beyond the use of conversational agents such as ChatGPT. Its current impact can be categorized into four main areas: interactive learning tools, intelligent assessment systems, administrative and logistical support, and content creation.

Interactive Learning Tools

AI-driven virtual assistants and chatbots—such as ChatGPT—facilitate dialogue-based learning, allowing students to practice history taking and enhance communication skills through simulated patient encounters [6]. Moreover, immersive technologies, including virtual reality, intelligent tutoring systems (ITS), medical robotics, and augmented reality, provide hands-on training experiences. These tools promote the development of clinical reasoning, surgical technique, and medical imaging interpretation while delivering personalized feedback and adaptive learning pathways [7,8].

Intelligent Assessment Systems

AI enhances evaluation processes by enabling automated grading and reducing evaluator bias. NLP can analyze narrative feedback to identify patterns related to competency, professionalism, and potential learner risk [9]. Innovations such as virtual Objective Structured Clinical Examinations (OSCEs) and automated assessments of clinical case presentations streamline performance evaluations [10-13]. In addition, ML contributes to the creation and validation of examination items, although human oversight remains essential to maintain content integrity [14,15].

Administrative and Logistical Support

In educational administration settings, AI automates the documentation of clinical experiences. For example, NLP applications can map trainee documentation to core competencies with high accuracy (92% - 97%), significantly reducing clerical workload [16-19]. Furthermore, predictive modeling aids in optimizing residency selection and procedural tracking, enhancing the efficiency of academic programs [20,21].

Content Creation and Academic Writing

AI also supports academic writing by assisting in the drafting of structured clinical notes, research manuscripts, and literature summaries. These tools enhance clarity, reduce writing time, and facilitate effective scholarly communication [22,23] (Table 1).

XSL•FO
RenderX

Table . Synthesis of artificial intelligence (AI) applications in medical education.

Category and AI application	Description
Virtual assistants and chatbots	
Interactive dialogue systems	Chatbots enable dialogue-based learning, supporting communication, and history-taking practice.
Generative conversational AI	Tools such as ChatGPT simulate realistic patient interactions to develop clinical communication skills.
Simulated medical cases	NLP[a]-driven simulations help assess and improve students' clinical reasoning and decision-making.
Learning tools, simulation, and VR[b]	
Virtual surgical assistants	AI evaluates surgical performance in simulations, offering feedback for skill refinement.
Virtual patient avatars	2D/3D avatars simulate clinical scenarios to train students in consultations and emergency care.
Virtual patient simulators and ITS[c]	Adaptive platforms offer real-time feedback to strengthen reasoning and procedural skills.
Medical robots	Used for pharmacologic simulations and clinical knowledge assessments in competency-based training.
Robot-assisted surgical training	Enables safe, simulated practice of complex surgical procedures.
Augmented reality in simulations	Augmented reality tools provide interactive training for interpreting imaging (eg, x-rays, CT[d] scans).
AI-integrated electronic learning platforms	Combine AI tutoring with online and in-person education for personalized learning paths.
NLP for literature processing	NLP tools assist in summarizing and interpreting large volumes of scientific literature.
Intelligent assessment systems	
Unbiased candidate selection	AI reduces demographic bias in residency selection processes.
Test creation	AI generates multiple-choice questions and clinical scenarios for examinations and curriculum.
Automated grading	Systems provide instant feedback on assessments while requiring human oversight.
Content creation and writing support	
AI-assisted academic writing	Tools assist in drafting structured notes and summaries to enhance academic communication.
AI in scientific research	Supports manuscript drafting and editing, with attention to avoiding fabricated content.

[a]NLP: natural language processing.

[b]VR: virtual reality.

[c]ITS: intelligent tutoring system.

[d]CT: computed tomography.

Benefits of AI in Medical Education

AI technologies are reshaping medical education by enabling personalized learning, enhancing clinical training, and supporting academic content development. While these advancements offer considerable promise, they also raise significant risks and ethical challenges. Responsible integration of AI requires careful attention to its implications, along with institutional commitment to adapting curricula and policies.

AI-driven technologies provide highly personalized learning experiences that can improve academic performance and clinical competence. One major advantage is the ability to tailor educational content and assessments to the individual learner's needs, fostering more efficient knowledge acquisition and deeper conceptual understanding [24]. Furthermore, AI-powered simulations—including virtual surgical assistants, patient avatars, and procedural simulators—offer safe, immersive environments where students can refine their diagnostic and technical skills without risk to patients [25].

Exposure to AI also prepares students for its growing role in clinical practice. Familiarity with decision support tools enhances future physicians' readiness to integrate AI into

XSL•FO
RenderX

diagnosis, treatment planning, and patient care [26]. From an instructional perspective, generative AI accelerates the development of educational content by facilitating the creation of multiple-choice questions, clinical vignettes, and tailored teaching modules [23,27,28]. Notably, AI-generated resources have the potential to reduce educational disparities by enhancing access to high-quality materials in underserved regions [23,27].

Risks and Ethical Considerations

Risks and Concerns

Despite its benefits, the incorporation of AI into medical training poses several concerns. One is the potential impact on career decisions. Some students may be deterred from entering fields such as radiology due to perceived reductions in future demand, as AI systems become increasingly proficient at image interpretation [29]. Another concern is the limited integration of AI-related content in existing medical curricula, leaving students unprepared to engage with AI-driven health care systems [30]. Additionally, there is the risk of overreliance on AI tools, which may undermine the development of critical hands-on clinical skills, diagnostic reasoning, and human-centered decision-making [31].

Ethical Considerations

The ethical use of AI in medical education is a growing focus of scholarly and institutional discourse. A key consideration is the preservation of human judgment. While AI can support decision-making, it must not displace the clinician's responsibility for critical thinking, empathy, and ethical deliberation [32]. Moreover, transparency and accountability are essential. Medical students must be taught to recognize biases, uphold data privacy, and ensure fairness in AI-mediated outcomes. Equitable access is also critical; AI integration must address the digital divide to ensure that learners across different regions and socioeconomic backgrounds benefit equally [29,33,34].

In medical education, AI use may compromise several ethical principles: academic integrity (fabrication or inaccuracy of references), justice (potential biases), nonmaleficence (overreliance on algorithmic outputs by students), and confidentiality (risks of training data extraction). These concerns are supported by empirical findings. For example, in a set of 115 references generated with ChatGPT-3.5 (OpenAI), 47% (n/N) were fabricated, 46% (n/N) inaccurate, and 7% (n/N) authentic and correct [35]; in another analysis of 636 citations, 55% (n/N) were fabricated with GPT-3.5 and 18% with GPT-4 (OpenAI), alongside errors in "real" citations [36]. Editorial practice has likewise documented high similarity indices with plagiarism risk and failures to detect misuse in medical-scientific articles—including AI-generated images with biological errors and the publication of stock large language model (LLM) phrases (eg, "I do not have access to real-time information...") in papers that passed peer review [37]. Beyond GenAI, computer vision used for instructional purposes (eg, medical imaging or intraoperative video) exhibits performance degradation due to dataset shift, dependence on spurious signals across hospitals or equipment, and vulnerability to adversarial attacks, necessitating external validation, audits, and technical safeguards when incorporated into educational activities [38,39]. From a privacy standpoint, LLMs have demonstrated memorization of training data and the potential for verbatim disclosure, reinforcing the need to avoid uploading protected health information (PHI) and to adopt institutional controls (data policies, privacy-preserving environments, and prior review before use in classrooms or OSCEs) [40,41].

In alignment with the United Nations Educational, Scientific and Cultural Organization (UNESCO) Recommendation on the Ethics of Artificial Intelligence (2021) and the World Health Organization Guidance on Ethics and Governance of AI for Health (2021), we translate principles into practice through five tightly scoped, actionable axes [42,43]: (1) governance and accountability: establish an oversight committee, maintain a public registry of AI systems, and require predeployment Algorithmic Impact Assessments and Data Protection Impact Assessments; (2) human oversight in high-stakes uses: mandate dual human+AI scoring, structured debriefs, and explicit override criteria grounded in clinical or educational judgment; (3) transparency: disclose in syllabi which tools are used and for what purposes and limits, provide model or data cards, and require student disclosure of AI assistance; (4) privacy-preserving data governance: prohibit uploading PHI to public models; enforce data minimization and deidentification, role-based access controls, secure logging, and retention limits; and prefer on-prem and virtual private cloud solutions when processing student data; and (5) equity, safety, and robustness: conduct subgroup bias testing and disparate impact monitoring, pursue cross-site validation where feasible, and set predefined error thresholds with rollback plans. To operationalize these pillars, we propose a staged pathway that begins with pilots in controlled sandbox environments with Algorithmic Impact Assessments or Data Protection Impact Assessments and a clear evaluation plan, then scales gradually with bias and robustness audits and cross-site performance monitoring, and culminates in institutional integration with periodic audits, update protocols, and public reporting. Together, these steps aim to improve learning while safeguarding autonomy, justice, and trust.

Curriculum and Teaching Strategies for AI

AI in Curricular Development

To ensure responsible and effective use of AI, medical education should incorporate foundational and applied competencies. Core knowledge areas should include data science, ML, algorithm design, statistics, and basic coding principles [44,45].

Curricular modules should also address clinical applications of AI in domains such as surgical training, radiology interpretation, ophthalmologic analysis, and hematologic diagnostics [46-49]. Ethical and legal education remains vital, with an emphasis on transparency, privacy, and professional accountability [34,50]. Notably, one publication has even proposed establishing a dedicated specialty in Medical Data Sciences to address the growing demand for AI expertise in clinical settings [51].

XSL•FO
RenderX

Teaching Strategies

Educators are actively exploring pedagogical strategies to integrate AI into undergraduate medical training. AI-enhanced curriculum design, such as the use of chatbots and generative platforms, enables the creation of dynamic learning experiences tailored to student progression and clinical reasoning levels [28]. Similarly, adaptive assessment systems powered by AI can identify individual student weaknesses and provide targeted feedback, thereby optimizing skill development and educational outcomes [23,27].

Implementation Path to AI in Medical Education: Adoption Trajectory, Challenges, and Solutions

The integration of AI into medical education can be meaningfully interpreted through Oberg's theory of culture shock, which comprises four progressive phases: honeymoon, frustration, adaptation, and acceptance [52]. This framework provides a valuable lens to examine the emotional, cognitive, and institutional transitions associated with AI adoption in medical education.

Adoption Trajectory

The Honeymoon Phase: Enthusiasm and Idealism

The initial response to AI is often marked by enthusiasm and idealism. During this phase, perspective articles and expert commentaries emphasize AI's transformative potential, exploring novel applications and offering conceptual frameworks for implementation. However, these early contributions frequently understate practical limitations, ethical dilemmas, and the systemic challenges associated with integration [52].

The Frustration Phase: Uncertainty and Resistance

As implementation progresses, institutions often encounter a phase of skepticism and resistance. Concerns may stem from limited technical knowledge, fears of professional displacement, or mistrust in AI's reliability. Faculty unfamiliar with digital tools may experience anxiety or a sense of obsolescence. Despite its discomfort, this phase is critical, as it catalyzes essential discussions on ethics, equity, and the risks of uncritical adoption [52].

The Adaptation Phase: Pragmatic Implementation

Over time, a more balanced perspective emerges, with institutions adopting AI tools in measured and context-sensitive ways. In this phase, AI is viewed not as a panacea but as a complement to existing pedagogical practices. Implementation strategies begin to emphasize design thinking, iterative refinement, and context-specific use cases. Nonetheless, many approaches remain limited in scope and lack a longitudinal vision [52].

The Acceptance Phase: Thoughtful Integration and Leadership

In the final phase, AI becomes a normalized and integrated component of the educational ecosystem. Educators demonstrate fluency in AI applications and embed them into curricula, assessments, and research activities with ethical awareness and pedagogical intent. Institutions at this stage offer scalable models characterized by innovation, critical engagement, and a commitment to preserving the humanistic core of medical education [52].

Persistent Challenges in AI Integration

Despite increasing interest and promising use cases, several challenges continue to hinder the effective integration of AI in medical education. Ethical and legal concerns include risks to data privacy, opaque accountability for AI-supported decisions, and the potential erosion of humanistic principles in teaching and care; AI use must remain aligned with the ethical foundations of medical practice and must respect the clinician-patient relationship [53]. Unreliable or low-quality outputs from generative models, such as ChatGPT and similar tools, may perpetuate outdated, biased, or inaccurate information when these systems are trained on flawed data; inaccuracies in AI-generated educational content can compromise learner safety and clinical competence, underscoring the need for rigorous data curation and active human oversight [1,54]. Scalability and equity barriers arise from limited digital infrastructure and high implementation costs, particularly in low- and middle-income countries, where disparities restrict equitable access to AI-driven tools and may reinforce existing educational inequities [54]. Finally, curricular displacement is a concern: overdependence on AI may inadvertently marginalize foundational clinical reasoning, communication, and hands-on skills if these technologies are introduced without explicit pedagogical safeguards and thoughtful alignment with existing curricula [51].

Proposed Solutions for Effective and Ethical AI Adoption

To harness the benefits while mitigating potential risks, a comprehensive strategy is essential—one that encompasses ethical considerations, pedagogical advancements, and technological inclusivity. Several complementary approaches can facilitate the responsible adoption of AI in medical education. Strengthening ethical frameworks is critical to ensure transparency, fairness, and respect for patient-centered values throughout AI development and deployment [53]. Enhancing data quality and human oversight is equally important: training AI systems on peer-reviewed data and subjecting outputs to expert review can reduce misinformation and support safe implementation [1,54]. Promoting equitable access involves designing adaptable, low-cost solutions tailored to diverse educational contexts and resource environments, particularly in underserved regions [54]. Reforming curricula to include AI literacy prepares future physicians to engage critically with algorithmic tools, with core competencies that encompass data science, algorithmic reasoning, ethical discernment, and digital fluency [51]. Supporting longitudinal research on the long-term educational and clinical impacts of AI integration will provide the evidence base needed for informed decision-making and continuous improvement [54].

XSL•FO
RenderX

These strategies are further detailed in Table 2, which aligns specific challenges with corresponding solutions to guide effective AI integration in medical education.

Table . Strategic challenges and solutions for the integration of artificial intelligence (AI) in medical education.

Challenge	Proposed solution
Ethical and legal risks	Strengthen ethical oversight and align AI use with core humanistic values [53]
Unreliable AI-generated content	Ensure high-quality data sources and implement human oversight [54]
Limited scalability and access	Develop accessible, adaptable tools for global use [54]
Curriculum imbalance	Integrate AI literacy into core training without displacing clinical fundamentals [45,51]
Need for evidence-based integration	Support rigorous research on educational impact and long-term outcomes [54]

Future Directions and Strategic Frameworks for AI in Medical Education

FACETS as a Unifying Lens for Implementation

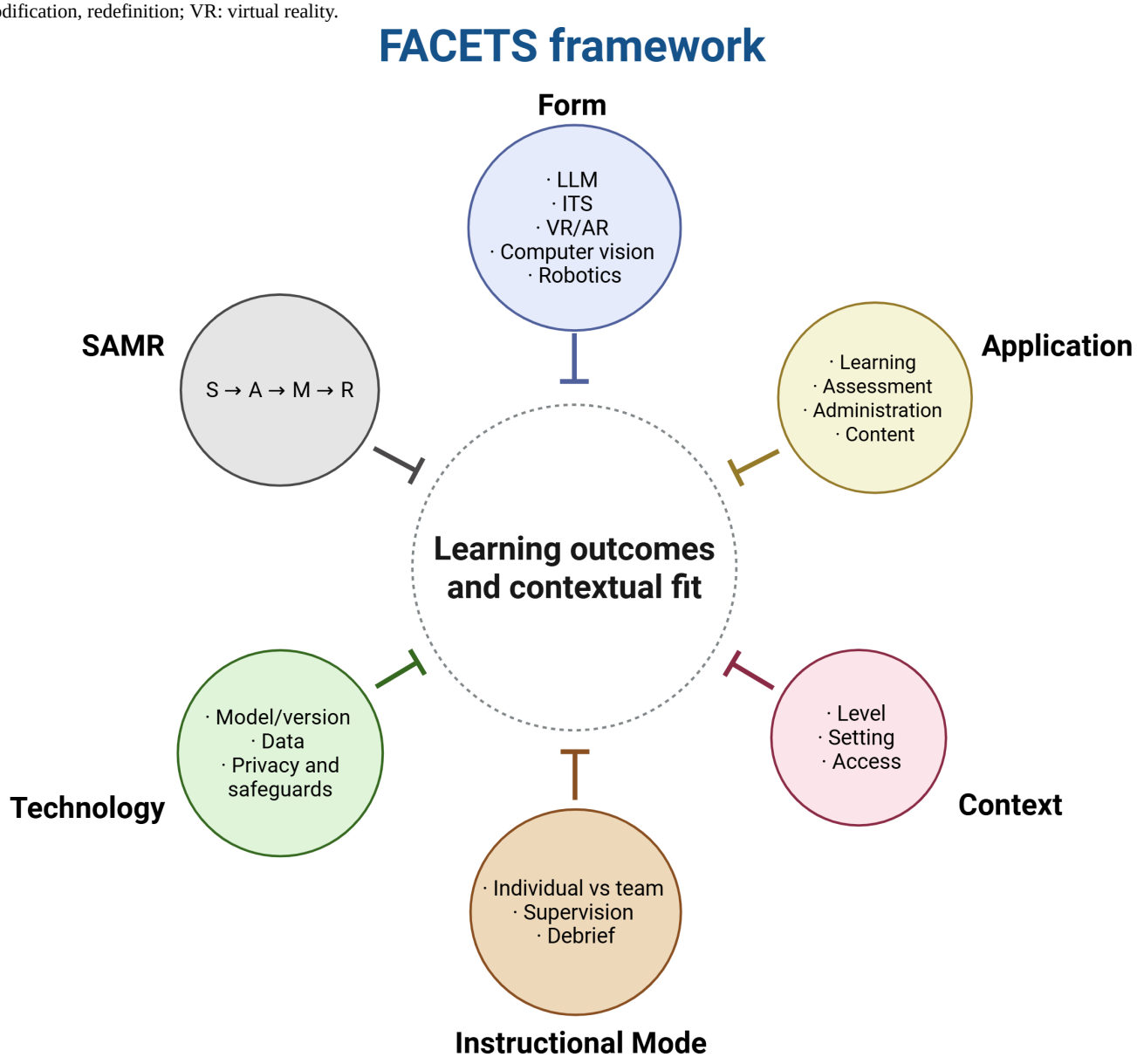
In the current context, recent studies typically address isolated components of AI in medical education, such as specific teaching technologies or environments. However, to advance this field, a more cohesive framework is required that facilitates replication, innovation, and a comprehensive understanding of the educational role of AI [2]. To this end, the FACETS framework has been proposed as a guiding model for future research and implementation in medical education settings. By examining six key dimensions—form, application, context, instructional mode, technology, and the SAMR (substitution, augmentation, modification, redefinition) model—this framework provides a structured approach for assessing how AI tools align with educational objectives (Figure 1). This multifaceted analysis ensures that AI implementations are pedagogically sound and contextually appropriate [2].

The tangible applicability of FACETS can be appreciated by retrospectively mapping published interventions (Table 3). Holderried et al [55] evaluated a GPT-4–based simulated patient that provided immediate, structured feedback for history taking among third-year medical students; they observed more than 99% medically plausible exchanges and near-perfect agreement with a human rater (κ=0.832), although 8 of 45 feedback categories showed lower concordance, underscoring the need for human oversight. Similarly, Luordo et al [56] used GPT-4 to grade OSCE reports from 96 students, finding high concordance with human graders (intraclass correlation coefficient=0.77 for single measures; 0.91 for average measures), systematically stricter AI scoring (AI mean 8.88, SD 2.96 vs experts' mean 12.39, SD 3.22; mean difference −3.51 points), and substantially shorter grading times (~24 min vs 2 - 4 h). Taken together, although the original studies did not explicitly use FACETS, this mapping supports the utility of the FACETS framework as an analytic lens to describe, compare, and guide future implementations of AI in medical education, particularly in undergraduate contexts.

Beyond the specific cases mapped earlier, FACETS can also be used to anticipate and assess emerging applications of AI in medical education—ranging from ITS and personalized learning platforms to chatbots (eg, ChatGPT) and intraoperative video analysis—by aligning each use case with pedagogical intent, context, and level of transformation [2].

XSL•FO
RenderX

Figure 1. FACETS framework for integrating AI into medical education. This schematic depicts the FACETS framework with six dimensions—form, application, context, instructional mode, technology, and SAMR—arrayed around a central objective: aligning interventions with learning outcomes and contextual fit for the integration of AI in medical education. AI: artificial intelligence; AR: augmented reality; CV: computer vision; GPT: generative pretrained transformer; HITL: human-in-the-loop; ITS: intelligent tutoring system; LLM: large language model; ML: machine learning; NLP: natural language processing; OSCE: Objective Structured Clinical Examination; PHI: protected health information; SAMR: substitution, augmentation, modification, redefinition; VR: virtual reality.



XSL•FO
RenderX

Table . Form, application, context, instructional mode, technology, SAMR[c] (FACETS) mapping of 2 real-world artificial intelligence (AI) interventions in medical education.

Dimension (FACETS)	Case 1—prospective study; 106 conversations (1894 question-answer pairs) [55]	Case 2—observational study; 96 students; AI versus human graders [56]
Form	LLM[a] chatbot simulating a patient with structured feedback	Automated scoring system for OSCE[b] reports (LLM)
Application	History-taking practice+evaluation of coverage of critical items	Formative or summative grading of reports using an institutional checklist
Context	Third-year students, European medical school, individual practice	Hospital teaching unit; 96 students in a real OSCE
Instructional mode	AI-guided practice with comparison against a human rater (for debrief)	AI-assisted assessment with expert benchmarking (human-in-the-loop)
Technology	GPT-4; analysis of response plausibility and category-based feedback quality	Batch GPT-4 pipeline; comparison with human graders; time logging
SAMR (level of transformation)	Augmentation → modification (replaces SP+adds rubric-aligned feedback or analytics)	Augmentation → modification (from manual scoring to automated with traceability or speed)

[a]LLM: large language model.

[b]OSCE: objective structured clinical examination.

[c]SAMR: substitution, augmentation, modification, redefinition.

Emerging Applications and Integrative Synthesis for Decision-Making

As AI becomes more deeply embedded in health care delivery, its incorporation into medical education continues to represent a critical yet uneven frontier. Although AI tools have shown promise in diagnosis, therapeutic planning, and operational efficiency, educational adoption has lagged—both in scale and in the rigor of implementation and evaluation. Key barriers include heterogeneous technological infrastructure—particularly in resource-limited settings—and limited AI literacy among students and faculty [57,58]. These challenges are compounded by the fact that many medical students, despite their interest in AI's potential, express anxiety about downstream labor market implications [59,60].

Beyond the case-based exemplars mapped with FACETS, diverse emerging applications in medical education share several salient features: they prioritize deliberate practice with adaptive feedback and competency-aligned performance traceability; they operate predominantly at the augmentation and modification levels of the SAMR model (with few achieving true redefinition); and they rely on high-quality data (interaction logs, rubrics, audio-video) that demand psychometric validation, bias and drift monitoring, and robust privacy safeguards. The evidentiary base remains heterogeneous, with single-site pilots and intermediate metrics predominating over longitudinal outcomes. The human factor remains essential—explicit instructional design, structured debriefing, and human-in-the-loop approaches are needed to calibrate criteria and prevent overreliance. Absent targeted faculty development and institutional processes (including data governance, audits, and reporting standards), these solutions risk amplifying equity gaps.

To operationalize these principles and translate them into programmatic, monitorable curricular decisions, it is helpful to note that these common features are especially evident in intelligent tutoring systems (ITSs), clinical chatbots, AI-assisted assessment, personalized learning platforms, virtual reality and AI simulators, anatomy tools, and intraoperative video analytics (Table 4). Building on that foundation, cross-cutting applications that enable and govern implementation across workstreams assume particular importance: (1) multimodal adaptive analytics, which fuse text, audio-video, and interaction traces to personalize pacing and content while flagging early risk at the learner or course level [61]; (2) learning analytics dashboards, which integrate competencies, assessment evidence, and progression signals to guide just-in-time feedback, remediation, and coaching for students and instructors [62]; and (3) educational clinical decision support, implemented in deidentified, guideline-constrained sandbox environments to teach evidence retrieval, uncertainty appraisal, and human-AI teaming in diagnostic and therapeutic planning [63,64] (Table 4).

Effective adoption of these cross-cutting layers requires the very institutional capacities and safeguards they presuppose. Curricula must evolve to incorporate AI training that fosters both technical competence and ethical responsibility [2]. This entails not only understanding AI tools but also recognizing their limitations, biases, and societal implications. Cross-sector collaboration among educational institutions, technology developers, and researchers is crucial to establishing universal digital literacy standards that equip learners and faculty to critically evaluate AI-generated content and apply such tools responsibly in clinical and educational settings [4].

XSL•FO
RenderX

Table . Emerging artificial intelligence (AI) applications in medical education.

Application	Description
Intelligent tutoring systems	AI-powered platforms enhancing decision-making and clinical reasoning [65]
AI-assisted learner assessment	Automated grading using NLP[a] and semantic analysis for case summaries [66]
Chatbots (eg, ChatGPT)	Teaching clinical management, USMLE[b] prep, communication skills, USMLE prep, communication skills [67,68]
Personalized learning platforms	Tailored learning paths and feedback for individualized student development [69]
Surgical simulations (VR[c]/AI)	Virtual reality tools for surgical skill training and evaluation [70]
Enhanced anatomy education	AI-assisted tools for deeper learning and retention in anatomy [71]
AI in admissions	Tools supporting application reviews and personal statement development [72]
AI-generated art	Improving patient education through visual storytelling [73]
Intraoperative video analysis	Teaching competency-based assessments via machine learning [74]
Multimodal adaptive analytics	Fuses text, audio-video, interaction data to personalize pacing and flag early risk [61]
Learning analytics dashboards	Aggregates competencies and assessment traces for just-in-time feedback and monitoring [62]
AI-supported CDS[d] for education	Guideline-constrained sandbox CDS to teach evidence retrieval, uncertainty, and human-AI teaming [63]

[a]NLP: natural language processing.

[b]USMLE: United States Medical Licensing Examination.

[c]VR: virtual reality.

[d]CDS: clinical decision support.

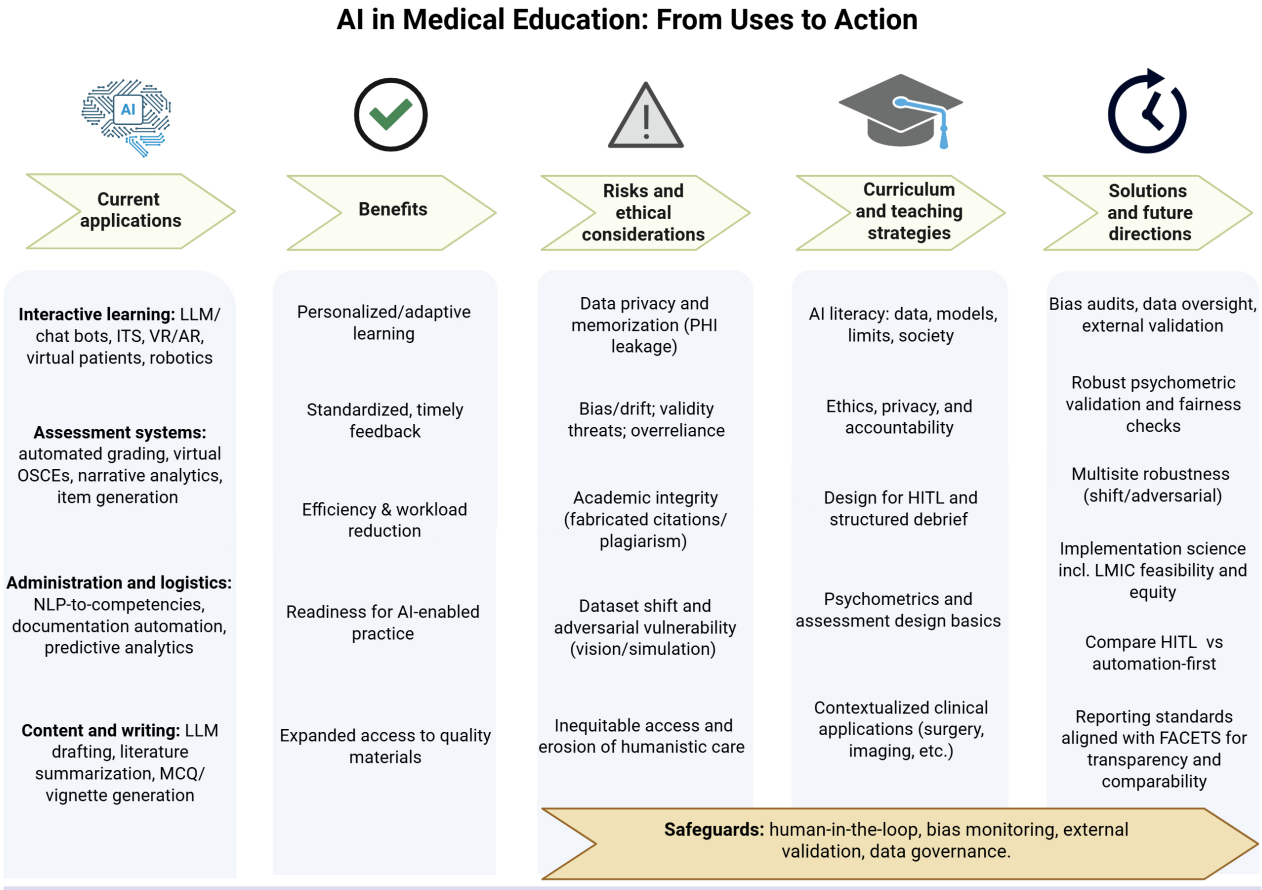
AI-powered platforms are expanding personalized learning by tailoring educational content to individual needs. Technologies such as virtual patients, augmented reality simulations, and mobile platforms offer dynamic, interactive experiences that increase engagement and broaden access to high-quality education, particularly in resource-limited regions. Nonetheless, the risk of depersonalizing medical practice—together with substantial concerns about data privacy—underscores the need for rigorous ethical and regulatory frameworks that engage all stakeholders in the educational process [75].

Given AI's expanding role in health care, prohibiting its use in education is neither practical nor beneficial. Institutions should instead establish comprehensive guidelines to ensure that AI-generated content remains reliable, relevant, and ethically sound. A structured, responsible approach—grounded in robust educational infrastructure, interdisciplinary collaboration, and consistent regulatory oversight—has the potential to improve learning outcomes and, ultimately, strengthen patient care.

As a synthesis of the manuscript's arguments, we recommend a concise, actionable set of practices: (1) report every AI pilot with a FACETS-aligned template to ensure transparency and comparability; (2) maintain human oversight for high-impact uses (eg, dual scoring with override), reserving automation-first approaches for low-risk contexts with active monitoring; (3) conduct predeployment audits for bias, drift, and adversarial robustness and, where appropriate, cross-site validation before scaling; (4) operate privacy-preserving workflows (no PHI in public LLMs, institutional sandboxes, data minimization or deidentification, role-based access, secure logging and retention); (5) deploy learning analytics dashboards tied to competency frameworks with ongoing psychometric monitoring; and (6) evaluate equity and feasibility using implementation science designs, including in resource-limited settings with disparate impact monitoring. This roadmap is summarized in Figure 2.

XSL•FO
RenderX

Figure 2. Schematic overview of AI applications in medical education. This figure synthesizes a practical pathway for AI uses in medical education. From left to right, it maps current applications—interactive learning, assessment systems, administration or logistics, and content or writing—to the benefits they enable; the risks and ethical considerations they entail; and the curriculum and teaching strategies required for responsible use. The rightmost column summarizes solutions and future directions: bias audits and data oversight; robust psychometrics and fairness checks; multisite robustness testing; implementation science (including LMIC feasibility and equity); comparative studies of HITL versus automation-first approaches; and reporting standards aligned with FACETS to support transparency and comparability. The bottom banner highlights cross-cutting safeguards—human-in-the-loop processes, bias monitoring, external validation, and data governance—that should accompany all deployments. AI: artificial intelligence; AR: augmented reality; FACETS: form, application, context, instructional mode, technology, and SAMR (substitution-augmentation-modification-redefinition); HITL: human-in-the-loop; ITS: intelligent tutoring system; LLM: large language model; LMIC: low- and middle-income country; MCQ: multiple-choice question; NLP: natural language processing; OSCE: Objective Structured Clinical Examination; PHI: protected health information; VR: virtual reality.



Conclusions

AI is redefining the landscape of medical education by offering innovative tools that enhance learning, assessment, and administrative processes. AI-driven platforms, such as ITS, virtual simulations, and generative models, provide personalized and interactive educational experiences, improving knowledge acquisition and clinical skills while streamlining administrative tasks to increase efficiency and reduce educator workload. These advancements hold immense potential to revolutionize medical education further. By fostering interdisciplinary collaboration, investing in robust infrastructure, and prioritizing ethical considerations, stakeholders can harness AI's capabilities to enhance learning outcomes and ultimately improve patient care.

At the same time, integrating AI into medical education presents several challenges. Ethical concerns such as data privacy, algorithmic bias, and the potential erosion of humanistic care must be carefully addressed. Technological disparities, particularly in low-resource settings, also hinder equitable AI adoption. To meet these challenges, medical curricula must evolve to include AI training that promotes both technical competence and ethical awareness, encompassing knowledge of AI tools, their limitations, and their societal implications. Cross-sector collaboration among educational institutions, technology developers, and researchers is essential to establishing universal digital literacy standards. These standards should equip students and educators to critically assess AI-generated content and use these tools responsibly in both educational and clinical settings.

Looking ahead, priorities include longitudinal studies linking AI-enabled instruction to validated learning outcomes and real-world performance; rigorous psychometric validation of AI-assisted assessments with routine bias audits; multisite robustness testing—addressing dataset shift and adversarial risk—especially for vision and simulation; implementation science work in diverse, including low-resource, settings on feasibility, cost-effectiveness, and equity; comparative evaluations of human-in-the-loop versus automation-first designs and their impact on critical thinking and professional identity; and FACETS-aligned reporting standards to strengthen transparency, reproducibility, and cross-study comparability. With ethically grounded design and cross-sector collaboration,

XSL•FO
RenderX

AI can enhance learning while preserving the humanistic foundations of medical practice.

Data Availability

Data sharing is not applicable to this article as no datasets were generated or analyzed during this study.

Conflicts of Interest

None declared.

References

1. Mir MM, Mir GM, Raina NT, et al. Application of artificial intelligence in medical education: current scenario and future perspectives. J Adv Med Educ Prof 2023 Jul;11(3):133-140. [doi: 10.30476/JAMP.2023.98655.1803] [Medline: 37469385]
2. Gordon M, Daniel M, Ajiboye A, et al. A scoping review of artificial intelligence in medical education: BEME Guide No. 84. Med Teach 2024 Apr;46(4):446-470. [doi: 10.1080/0142159X.2024.2314198] [Medline: 38423127]
3. Rahman MA, Victoros E, Ernest J, Davis R, Shanjana Y, Islam MR. Impact of artificial intelligence (AI) technology in healthcare sector: a critical evaluation of both sides of the coin. Clin Pathol 2024;17:2632010X241226887. [doi: 10.1177/2632010X241226887] [Medline: 38264676]
4. Haleem A, Javaid M, Qadri MA, Suman R. Understanding the role of digital technologies in education: a review. Sustain Oper Comput 2022;3:275-285. [doi: 10.1016/j.susoc.2022.05.004]
5. Pineda-de-Alcázar MY. Inteligencia Artificial y Modelos de Comunicación [Article in Spanish]. Razón y Palabra 2017 Dec 14;21(4_99):332-346 [FREE Full text]
6. Hale J, Alexander S, Wright ST, Gilliland K. Generative AI in undergraduate medical education: a rapid review. J Med Educ Curric Dev 2024 Jan;11:23821205241266697. [doi: 10.1177/23821205241266697]
7. Riddle EW, Kewalramani D, Narayan M, Jones DB. Surgical simulation: virtual reality to artificial intelligence. Curr Probl Surg 2024 Nov;61(11):101625. [doi: 10.1016/j.cpsurg.2024.101625] [Medline: 39477664]
8. Shahrezaei A, Sohani M, Taherkhani S, Zarghami SY. The impact of surgical simulation and training technologies on general surgery education. BMC Med Educ 2024 Nov 13;24(1):1297. [doi: 10.1186/s12909-024-06299-w] [Medline: 39538209]
9. Booth GJ, Ross B, Cronin WA, et al. Competency-based assessments: leveraging artificial intelligence to predict subcompetency content. Acad Med 2023 Apr 1;98(4):497-504. [doi: 10.1097/ACM.0000000000005115] [Medline: 36477379]
10. King AJ, Kahn JM, Brant EB, Cooper GF, Mowery DL. Initial development of an automated platform for assessing trainee performance on case presentations. ATS Sch 2022 Dec;3(4):548-560. [doi: 10.34197/ats-scholar.2022-0010OC] [Medline: 36726701]
11. Hamdy H, Sreedharan J, Rotgans JI, et al. Virtual clinical encounter examination (VICEE): a novel approach for assessing medical students' non-psychomotor clinical competency. Med Teach 2021 Oct;43(10):1203-1209. [doi: 10.1080/0142159X.2021.1935828] [Medline: 34130589]
12. Pereira DSM, Falcão F, Nunes A, Santos N, Costa P, Pêgo JM. Designing and building OSCEBot ® for virtual OSCE - performance evaluation. Med Educ Online 2023 Dec;28(1):2228550. [doi: 10.1080/10872981.2023.2228550] [Medline: 37347808]
13. Merritt C, Glisson M, Dewan M, Klein M, Zackoff M. Implementation and evaluation of an artificial intelligence driven simulation to improve resident communication with primary care providers. Acad Pediatr 2022 Apr;22(3):503-505. [doi: 10.1016/j.acap.2021.12.013] [Medline: 34923145]
14. Agarwal M, Sharma P, Goswami A. Analysing the applicability of ChatGPT, Bard, and Bing to generate reasoning-based multiple-choice questions in medical physiology. Cureus 2023 Jun;15(6):e40977. [doi: 10.7759/cureus.40977] [Medline: 37519497]
15. Ayub I, Hamann D, Hamann CR, Davis MJ. Exploring the potential and limitations of chat generative pre-trained transformer (ChatGPT) in generating board-style dermatology questions: a qualitative analysis. Cureus 2023 Aug;15(8):e43717. [doi: 10.7759/cureus.43717] [Medline: 37638266]
16. Bond WF, Zhou J, Bhat S, et al. Automated patient note grading: examining scoring reliability and feasibility. Acad Med 2023 Nov 1;98(11S):S90-S97. [doi: 10.1097/ACM.0000000000005357] [Medline: 37983401]
17. Cianciolo AT, LaVoie N, Parker J. Machine scoring of medical students' written clinical reasoning: initial validity evidence. Acad Med 2021 Jul 1;96(7):1026-1035. [doi: 10.1097/ACM.0000000000004010] [Medline: 33637657]
18. Wang M, Sun Z, Jia M, et al. Intelligent virtual case learning system based on real medical records and natural language processing. BMC Med Inform Decis Mak 2022 Mar 4;22(1):60. [doi: 10.1186/s12911-022-01797-7] [Medline: 35246134]
19. Woo CW, Evens MW, Freedman R, et al. An intelligent tutoring system that generates a natural language dialogue using dynamic multi-level planning. Artif Intell Med 2006 Sep;38(1):25-46. [doi: 10.1016/j.artmed.2005.10.004] [Medline: 16352427]

XSL•FO
RenderX

20. Gupta N, Khatri K, Malik Y, et al. Exploring prospects, hurdles, and road ahead for generative artificial intelligence in orthopedic education and training. BMC Med Educ 2024 Dec 28;24(1):1544. [doi: 10.1186/s12909-024-06592-8] [Medline: 39732679]
21. Grévisse C. LLM-based automatic short answer grading in undergraduate medical education. BMC Med Educ 2024 Sep 27;24(1):1060. [doi: 10.1186/s12909-024-06026-5] [Medline: 39334087]
22. Khalifa M, Albadawy M. Using artificial intelligence in academic writing and research: an essential productivity tool. Computer Methods and Programs in Biomedicine Update 2024;5:100145. [doi: 10.1016/j.cmpbup.2024.100145]
23. Dwivedi YK, Kshetri N, Hughes L, et al. Opinion Paper: "So what if ChatGPT wrote it?" Multidisciplinary perspectives on opportunities, challenges and implications of generative conversational AI for research, practice and policy. Int J Inf Manage 2023 Aug;71:102642. [doi: 10.1016/j.ijinfomgt.2023.102642]
24. Narayanan S, Ramakrishnan R, Durairaj E, Das A. Artificial intelligence revolutionizing the field of medical education. Cureus 2023 Nov;15(11):e49604. [doi: 10.7759/cureus.49604] [Medline: 38161821]
25. Spatscheck N, Schaschek M, Winkelmann A. The effects of generative AI's human-like competencies on clinical decision-making. J Decis Syst 2024:1-39. [doi: 10.1080/12460125.2024.2430731]
26. Bansal M, Jindal A. Artificial intelligence in healthcare: should it be included in the medical curriculum? A students' perspective. Natl Med J India 2022;35(1):56-58. [doi: 10.25259/NMJI_208_20] [Medline: 36039630]
27. Kamalov F, Santandreu Calonge D, Gurrib I. New era of artificial intelligence in education: towards a sustainable multifaceted revolution. Sustainability 2023 Jan;15(16):12451. [doi: 10.3390/su151612451]
28. Arévalo JA, Cordero MQ. ChatGPT: la creación automática de contenidos con Inteligencia Artificial y su impacto en la comunicación académica y educativa [Article in Spanish]. Desiderata 2023(22):136-142 [FREE Full text]
29. Alamer A. Medical students' perspectives on artificial intelligence in radiology: the current understanding and impact on radiology as a future specialty choice. Curr Med Imaging Rev 2023 Jul;19(8):921-930. [doi: 10.2174/1573405618666220907111422]
30. Civaner MM, Uncu Y, Bulut F, Chalil EG, Tatli A. Artificial intelligence in medical education: a cross-sectional needs assessment. BMC Med Educ 2022 Nov 9;22(1):772. [doi: 10.1186/s12909-022-03852-3] [Medline: 36352431]
31. van de Ridder JMM, Shoja MM, Rajput V. Finding the place of ChatGPT in medical education. Acad Med 2023 Aug 1;98(8):867. [doi: 10.1097/ACM.0000000000005254] [Medline: 37162206]
32. Izquierdo-Condoy JS, Arias-Intriago M, Tello-De-la-Torre A, Busch F, Ortiz-Prado E. Generative artificial intelligence in medical education: enhancing critical thinking or undermining cognitive autonomy? J Med Internet Res 2025 Nov 3;27(1):e76340. [doi: 10.2196/76340] [Medline: 41183320]
33. Ejaz H, McGrath H, Wong BL, Guise A, Vercauteren T, Shapey J. Artificial intelligence and medical education: a global mixed-methods study of medical students' perspectives. Digit Health 2022;8:20552076221089099. [doi: 10.1177/20552076221089099] [Medline: 35521511]
34. Zhong JY, Fischer NL. Commentary: the desire of medical students to integrate artificial intelligence into medical education: an opinion article. Front Digit Health 2023;5:1151390. [doi: 10.3389/fdgth.2023.1151390] [Medline: 37090065]
35. Bhattacharyya M, Miller VM, Bhattacharyya D, Miller LE. High rates of fabricated and inaccurate references in ChatGPT-generated medical content. Cureus 2023 May;15(5):e39238. [doi: 10.7759/cureus.39238] [Medline: 37337480]
36. Walters WH, Wilder EI. Fabrication and errors in the bibliographic citations generated by ChatGPT. Sci Rep 2023 Sep 7;13(1):14045. [doi: 10.1038/s41598-023-41032-5] [Medline: 37679503]
37. Izquierdo-Condoy JS, Vásconez-González J, Ortiz-Prado E. "AI et al." The perils of overreliance on artificial intelligence by authors in scientific research. Clinical eHealth 2024 Dec;7:133-135. [doi: 10.1016/j.ceh.2024.09.001]
38. Zech JR, Badgeley MA, Liu M, Costa AB, Titano JJ, Oermann EK. Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: a cross-sectional study. PLoS Med 2018 Nov;15(11):e1002683. [doi: 10.1371/journal.pmed.1002683] [Medline: 30399157]
39. D'Amour A, Heller K, Moldovan D, et al. Underspecification presents challenges for credibility in modern machine learning. J Mach Learn Res 2022;23(226):1-61 [FREE Full text]
40. Hayes J, Swanberg M, Chaudhari H, et al. Measuring memorization in language models via probabilistic extraction. In: Chiruzzo L, Ritter A, Wang L, editors. 2025 Presented at: Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers); Apr 29 to May 4, 2025; Albuquerque, New Mexico p. 9266-9291 URL: https://aclanthology.org/volumes/2025.naacl-long/ [doi: 10.18653/v1/2025.naacl-long.469]
41. Carlini N, Tramer F, Wallace E, et al. Extracting training data from large language models. 2021 Presented at: Proceedings of the 30th USENIX Security Symposium(USENIX Security 21); Aug 11-13, 2021 URL: https://www.usenix.org/system/files/sec21-carlini-extracting.pdf [accessed 2026-01-30]
42. Recommendation on the ethics of artificial intelligence. : United Nations Educational, Scientific and Cultural Organization; 2022 URL: https://digitallibrary.un.org/record/4062376?v=pdf [accessed 2026-01-30]
43. Ethics and governance of artificial intelligence for health. : World Health Organization; 2021 URL: https://iris.who.int/server/api/core/bitstreams/f780d926-4ae3-42ce-a6d6-e898a5562621/content [accessed 2026-02-06]

XSL•FO
RenderX

44. Nagy M, Radakovich N, Nazha A. Why machine learning should be taught in medical schools. Med Sci Educ 2022 Apr;32(2):529-532. [doi: 10.1007/s40670-022-01502-3] [Medline: 35528308]
45. Ngo B, Nguyen D, vanSonnenberg E. The cases for and against artificial intelligence in the medical school curriculum. Radiol Artif Intell 2022 Sep;4(5):e220074. [doi: 10.1148/ryai.220074] [Medline: 36204540]
46. Chai SY, Hayat A, Flaherty GT. Integrating artificial intelligence into haematology training and practice: opportunities, threats and proposed solutions. Br J Haematol 2022 Sep;198(5):807-811. [doi: 10.1111/bjh.18343] [Medline: 35781249]
47. Ward TM, Mascagni P, Madani A, Padoy N, Perretta S, Hashimoto DA. Surgical data science and artificial intelligence for surgical education. J Surg Oncol 2021 Aug;124(2):221-230. [doi: 10.1002/jso.26496] [Medline: 34245578]
48. Valikodath NG, Cole E, Ting DSW, et al. Impact of artificial intelligence on medical education in ophthalmology. Transl Vis Sci Technol 2021 Jun 1;10(7):14. [doi: 10.1167/tvst.10.7.14] [Medline: 34125146]
49. Tejani AS, Elhalawani H, Moy L, Kohli M, Kahn Jr CE. Artificial intelligence and radiology education. Radiol Artif Intell 2023 Jan;5(1):e220084. [doi: 10.1148/ryai.220084] [Medline: 36721409]
50. Busch F, Adams LC, Bressem KK. Biomedical ethical aspects towards the implementation of artificial intelligence in medical education. Med Sci Educ 2023 Aug;33(4):1007-1012. [doi: 10.1007/s40670-023-01815-x] [Medline: 37546190]
51. Cussat-Blanc S, Castets-Renard C, Monsarrat P. Doctors in medical data sciences: a new curriculum. Int J Environ Res Public Health 2022 Dec 30;20(1):675. [doi: 10.3390/ijerph20010675] [Medline: 36612994]
52. Zhou Y, Jindal-Snape D, Topping K, Todman J. Theoretical models of culture shock and adaptation in international students in higher education. Stud High Educ 2008 Feb;33(1):63-75. [doi: 10.1080/03075070701794833]
53. Mennella C, Maniscalco U, De Pietro G, Esposito M. Ethical and regulatory challenges of AI technologies in healthcare: a narrative review. Heliyon 2024 Feb 29;10(4):e26297. [doi: 10.1016/j.heliyon.2024.e26297] [Medline: 38384518]
54. Abdelwanis M, Alarafati HK, Tammam MMS, Simsekler MCE. Exploring the risks of automation bias in healthcare artificial intelligence applications: a Bowtie analysis. Journal of Safety Science and Resilience 2024 Dec;5(4):460-469. [doi: 10.1016/j.jnlssr.2024.06.001]
55. Holderried F, Stegemann-Philipps C, Herrmann-Werner A, et al. A language model-powered simulated patient with automated feedback for history taking: prospective study. JMIR Med Educ 2024 Aug 16;10:e59213. [doi: 10.2196/59213] [Medline: 39150749]
56. Luordo D, Torres Arrese M, Tristán Calvo C, et al. Application of artificial intelligence as an aid for the correction of the objective structured clinical examination (OSCE). Appl Sci 2025 Jan;15(3):1153. [doi: 10.3390/app15031153]
57. Izquierdo-Condoy JS, Arias-Intriago M, Nati-Castillo HA, et al. Exploring smartphone use and its applicability in academic training of medical students in Latin America: a multicenter cross-sectional study. BMC Med Educ 2024 Nov 30;24(1):1401. [doi: 10.1186/s12909-024-06334-w] [Medline: 39616324]
58. Ali H, ul Mustafa A, Aysan AF. Global adoption of generative AI: what matters most? J Econ Technol 2025 Nov;3:166-176. [doi: 10.1016/j.ject.2024.10.002]
59. Sami A, Tanveer F, Sajwani K, et al. Medical students' attitudes toward AI in education: perception, effectiveness, and its credibility. BMC Med Educ 2025 Jan 17;25(1):82. [doi: 10.1186/s12909-025-06704-y] [Medline: 39833834]
60. Busch F, Hoffmann L, Truhn D, et al. Global cross-sectional student survey on AI in medical, dental, and veterinary education and practice at 192 faculties. BMC Med Educ 2024 Sep 28;24(1):1066. [doi: 10.1186/s12909-024-06035-4] [Medline: 39342231]
61. Guerrero-Sosa JDT, Romero FP, Menéndez-Domínguez VH, Serrano-Guerrero J, Montoro-Montarroso A, Olivas JA. A comprehensive review of multimodal analysis in education. Appl Sci (Basel) 2025 Jan;15(11):5896. [doi: 10.3390/app15115896]
62. Masiello I, Mohseni Z(, Palma F, Nordmark S, Augustsson H, Rundquist R. A current overview of the use of learning analytics dashboards. Educ Sci 2024 Jan;14(1):82. [doi: 10.3390/educsci14010082]
63. Elhaddad M, Hamam S. AI-driven clinical decision support systems: an ongoing pursuit of potential. Cureus 2024 Apr;16(4):e57728. [doi: 10.7759/cureus.57728] [Medline: 38711724]
64. Soares A, Afshar M, Moesel C, et al. Playing in the clinical decision support sandbox: tools and training for all. JAMIA Open 2023 Jul;6(2):ooad038. [doi: 10.1093/jamiaopen/ooad038] [Medline: 37351012]
65. Lamti S, El Malhi M, Sekhsoukh R, Kerzazi N. Intelligent tutoring system for medical students: opportunities and challenges. 2024 Presented at: International Conference on Smart Medical, IoT & Artificial Intelligence (ICSMAI 2024); Apr 18-20, 2024. [doi: 10.1007/978-3-031-66850-0_27] [Medline: 38617415]
66. Salt J, Harik P, Barone MA. Leveraging natural language processing: toward computer-assisted scoring of patient notes in the USMLE step 2 clinical skills exam. Acad Med 2019 Mar;94(3):314-316. [doi: 10.1097/ACM.0000000000002558] [Medline: 30540567]
67. Karabacak M, Ozkara BB, Margetis K, Wintermark M, Bisdas S. The advent of generative language models in medical education. JMIR Med Educ 2023 Jun 6;9:e48163. [doi: 10.2196/48163] [Medline: 37279048]
68. Koga S. The potential of ChatGPT in medical education: focusing on USMLE preparation. Ann Biomed Eng 2023 Oct;51(10):2123-2124. [doi: 10.1007/s10439-023-03253-7] [Medline: 37248408]
69. Abd-Alrazaq A, AlSaad R, Alhuwail D, et al. Large language models in medical education: opportunities, challenges, and future directions. JMIR Med Educ 2023 Jun 1;9:e48291. [doi: 10.2196/48291] [Medline: 37261894]

XSL•FO
RenderX

70. Rogers MP, DeSantis AJ, Janjua H, Barry TM, Kuo PC. The future surgical training paradigm: virtual reality and machine learning in surgical education. Surgery 2021 May;169(5):1250-1252. [doi: 10.1016/j.surg.2020.09.040] [Medline: 33280858]
71. Abdellatif H, Al Mushaiqri M, Albalushi H, Al-Zaabi AA, Roychoudhury S, Das S. Teaching, learning and assessing anatomy with artificial intelligence: the road to a better future. Int J Environ Res Public Health 2022 Oct 31;19(21):14209. [doi: 10.3390/ijerph192114209] [Medline: 36361089]
72. Hashimoto DA, Johnson KB. The use of artificial intelligence tools to prepare medical school applications. Acad Med 2023 Sep 1;98(9):978-982. [doi: 10.1097/ACM.0000000000005309] [Medline: 37369073]
73. Huston JC, Kaminski N. A picture worth a thousand words, created with one sentence: using artificial intelligence-created art to enhance medical education. ATS Sch 2023 Jun;4(2):145-151. [doi: 10.34197/ats-scholar.2022-0141PS] [Medline: 37533539]
74. Kawka M, Gall TMH, Fang C, Liu R, Jiao LR. Intraoperative video analysis and machine learning models will change the future of surgical training. Intell Surg 2022 Jan;1:13-15. [doi: 10.1016/j.isurg.2021.03.001]
75. Maleki Varnosfaderani S, Forouzanfar M. The role of AI in hospitals and clinics: transforming healthcare in the 21st century. Bioengineering (Basel) 2024 Mar 29;11(4):337. [doi: 10.3390/bioengineering11040337] [Medline: 38671759]

Abbreviations

AI: artificial intelligence
FACETS: form, application, context, instructional mode, technology, SAMR (framework components)
ITS: intelligent tutoring system
LLM: large language model
ML: machine learning
NLP: natural language processing
OSCE: Objective Structured Clinical Examination
PHI: protected health information
SAMR: substitution, augmentation, modification, redefinition

Edited by A Hasan Sapci, A Stone, T Leung; submitted 07.May.2025; peer-reviewed by S Mohanadas, SM Shaffi; accepted 14.Dec.2025; published 17.Feb.2026.

Please cite as:
Izquierdo-Condoy JS, Arias-Intriago M, Montero Corrales L, Ortiz-Prado E
Artificial Intelligence in Medical Education: Transformative Potential, Current Applications, and Future Implications
JMIR Med Educ 2026;12:e77127
URL: https://mededu.jmir.org/2026/1/e77127
doi:10.2196/77127

© Juan S Izquierdo-Condoy, Marlon Arias-Intriago, Laura Montero Corrales, Esteban Ortiz-Prado. Originally published in JMIR Medical Education (https://mededu.jmir.org), 17.Feb.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Education, is properly cited. The complete bibliographic information, a link to the original publication on https://mededu.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

From Realism to Learner Engagement: Rethinking Fidelity in Simulation-Based Education

Julien Pico[1,2], MD; Jean-Noel Evain[2], MD, PhD; Christina Aron[2], MD; Gilles Martin[2], PhD; Ilian Cruz-Panesso[3], PhD; Leonida-Mihai Georgescu[2], MD; Issam Tanoubi[2,4], MD

[1]Centre Hospitalier Universitaire de Montpellier, Montpellier, Occitanie, France

[2]Centre d'apprentissage des attitudes et habiletés cliniques (CAAHC), Université de Montréal, Pavillon Roger-Gaudry, 2900, boul. Édouard-Montpetit, 8e étage, Montréal, QC, Canada

[3]Direction de l'enseignement et de l'Académie CHUM (DEAC), Montréal, QC, Canada

[4]Centre de Recherche en Pédagogie de la Santé (CRPS), Faculté de Médecine, Université de Montréal, Montréal, QC, Canada

Corresponding Author:
Julien Pico, MD
Centre Hospitalier Universitaire de Montpellier, Montpellier, Occitanie, France

Abstract

Simulation has become an essential pedagogical tool in health professions education, traditionally valued for its ability to approximate clinical practice. Higher simulation fidelity is often assumed to directly enhance learner engagement and improve educational outcomes; however, this assumption oversimplifies a complex relationship. Fidelity is multidimensional, encompassing physical, emotional, and contextual dimensions, as well as qualitative and quantitative considerations, each influencing learners' perception of realism in distinct ways. Engagement is shaped not only by these dimensions of fidelity but also by intrinsic factors such as motivation, prior experience, stress, and emotional resilience, and by extrinsic factors including instructional design, facilitation, debriefing, and psychological safety. A central mediator in this process is the fiction contract, an implicit agreement that enables learners to suspend disbelief and engage authentically despite inherent limitations in realism. Technological sophistication alone does not necessarily translate into greater educational impact. Rather, fidelity should be intentionally aligned with learning objectives: advanced patient simulators may support procedural training, standardized patients may enhance communication skills, and task trainers may optimize focused psychomotor practice. This viewpoint advocates for a goal-oriented, multimodal approach that redefines high-fidelity simulation not as the pursuit of maximum realism, but as the deliberate alignment of fidelity with pedagogy to optimize learner engagement and educational effectiveness.

(JMIR Med Educ 2026;12:e84684) doi:10.2196/84684

KEYWORDS

fidelity; fiction contract; learning; simulation training; medical education

Simulation is a realistic representation of clinical practice in health care education, guiding learners to the apex of Miller's pyramid of competence[1]. In the context of health professions education, simulation plays a pivotal role in developing practical competence in clinical environments, specifically the acquisition of skills in real conditions and practical knowledge. The primary educational goal is to engage learners in Kolb's experiential learning cycle [2]. However, this goal often assumes that student engagement with simulation is directly and linearly linked to its fidelity [3]. But how valid is this assumption, and along which dimensions [4]? Simulation-based education in health professions has long been influenced by the assumption that higher fidelity directly enhances learning. However, foundational studies have questioned this linear relationship. Norman et al [5] demonstrated that the level of simulation fidelity had minimal impact on the transfer of learning, emphasizing instead the role of instructional design. Building on this, Hamstra et al [6] introduced the idea of functional task alignment, suggesting that fidelity should be defined relative to educational goals rather than technical sophistication. Petrosoniak et al [7] further argued that simulation design must serve both learner needs and broader health care system objectives and Dieckmann et al [8] emphasized the complex and socially constructed nature of realism in simulation-based education. These perspectives provide the theoretical backdrop for our reevaluation of fidelity and its relationship to learner engagement. This position, while increasingly supported in the literature, remains debated and calls for a more nuanced, evidence-informed understanding of when and how fidelity contributes to learning. In this essay, we revisit the concept of fidelity in simulation, analyze its multiple components, and evaluate their potential impact on student engagement. We will then suggest a redefinition of high-fidelity simulation aligned with educational objectives and highlight the significance of the fiction contract as a precondition for any simulation activity.

XSL•FO
RenderX

Qualitative and Quantitative Fidelity

Simulation fidelity encompasses two key dimensions—qualitative and quantitative—which are used here as a heuristic lens rather than as a comprehensive or exclusive framework [9,10]. Other frameworks describe fidelity and ecological validity across broader sociotechnical dimensions, including tasks, technologies, environments, systems, and professional. Qualitative fidelity refers to the degree to which simulated elements authentically mirror real-world counterparts. For instance, incorporating equipment used in clinical practice, such as an actual anesthesia machine, into an operating room simulation enhances this qualitative dimension. In contrast, quantitative fidelity pertains to the number and richness of elements included in the simulation activity compared to real life. For example, a patient's fall in a hospital room with a suitcase full of clothes, an overturned table, and shattered glass on the floor increases quantitative fidelity. However, increasing the number of elements within a scenario does not inherently enhance educational fidelity and may introduce unnecessary cognitive load if not aligned with learning objectives.

A higher level of fidelity allows learners to become immediately immersed in a real-world context, reducing the need for imagination and fostering stronger situational awareness. Conversely, lower fidelity requires learners to engage their imagination continuously, which can disrupt natural responses and divert cognitive resources away from patient care. This additional cognitive load, resulting from efforts to compensate for missing realism, may ultimately impair performance and reduce learning effectiveness [11,12]. These dimensions of fidelity further interact with qualitative and quantitative considerations and must be interpreted relative to the intended educational goals.

The Dimensions of Fidelity in Simulation-Based Medical Education

Three major dimensions of fidelity are frequently described in the literature. (1) Physical fidelity refers to the degree of visual and functional resemblance between simulated elements and their real-world counterparts. (2) Emotional fidelity pertains to elements, often sensory or contextual, that evoke emotional responses in participants. (3) Contextual fidelity relates to the progression of a situation as it would unfold in real clinical settings. Each of these types of fidelity influences the learner's perception of realism [13]. While the impact of physical resemblance may seem obvious, the effects of the other forms of fidelity are less straightforward, as some studies suggest that simulation can elicit emotional and physiological responses comparable to real clinical situations despite limitations in physical realism [14].

Emotional fidelity enhances realism by eliciting the same emotions in the learner that would arise in a real situation [15]. For example, an actor portraying a distraught and tearful father during the resuscitation of his newborn could significantly heighten the realism of the simulation activity. These emotional responses are crucial for several reasons. First, they may create cognitive load similar to real-life stress, forcing learners to manage their emotions, regulate their reactions, and remain focused on patient care. Without this exposure in simulation, learners may face such emotions for the first time in real clinical situations, potentially resulting in suboptimal performance [16]. Emotionally charged situations also promote memory retention [17], a desirable outcome in education. However, this type of fidelity could have adverse effects if the learner has previously experienced a similar emotional situation or if there is emotional overload, potentially leading to "freezing" or memory suppression [18].

Contextual fidelity is equally critical from an educational perspective. The timing of the simulation activity progression must respect the real-life temporal context. Accelerating or slowing down events during the activity, such as receiving the results of a diagnostic test in 2 minutes when it should take 20 or allowing the patient's condition to deteriorate slowly to give participants time to find the "solution," could detract from experiential learning. The decisions made by the learner are tailored to the designed activity, which might not be applicable to a real clinical setting or might even be wrong. Contextual fidelity is even more critical in the context of in situ simulation. It is essential for uncovering latent safety threats within the clinical environment [19].

Relationship Between Simulation Fidelity and Learner Engagement in the Simulation Activity

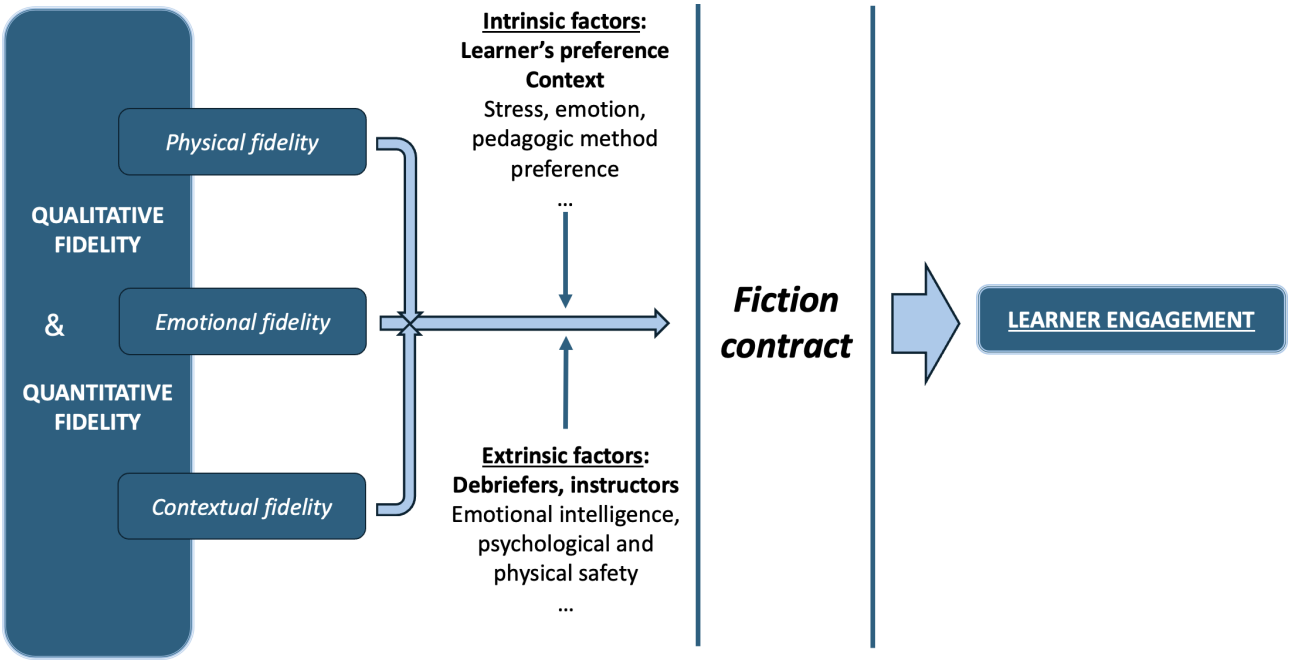
It is commonly assumed that greater fidelity leads to stronger learner engagement, as participants perceive the activity as more realistic and therefore more pedagogically useful [20]. However, this assumption is not consistently supported, as studies comparing simulators of differing fidelity have shown similar learner performance and decision-making outcomes, even when using less technologically advanced simulation modalities [21,22]. Other internal factors may modulate this engagement, such as the learner's motivation, performance-related stress, and evaluation anxiety; the emotional intelligence of the instructor and the debriefer; and the clinical complexity of the simulation activity.

In this context, the importance of the *fiction contract* becomes paramount [3,23,24]. The fiction contract is an implicit, verbally acknowledged agreement between the instructor and the learners. The fiction contract acts as a mediator between fidelity and learner engagement by enabling learners to suspend disbelief and engage authentically despite inevitable limitations in simulation realism. During the prebriefing, the instructor is encouraged to clarify the limitations of the simulation, describe the team's efforts to enhance fidelity, and explain that the realism of the situation also stems from the authenticity of the learners' responses. By explicitly framing expectations and acknowledging these constraints, the fiction contract allows learners to focus on the learning task rather than on imperfections in realism. This approach encourages participants to "buy into" the activity, treating the simulated situation as if it were real. The ultimate goal is to leverage simulation as a

XSL•FO
RenderX

pedagogical tool at the "show how" stage of Miller's pyramid, thereby fostering engagement in the experiential learning cycle through simulated practice (Figure 1).

Figure 1. Modulating factors of learner engagement during medical simulation.



Goal-Oriented Multimodal Simulation Fidelity

Our proposal to redefine high-fidelity simulation as the alignment of simulation features with educational objectives is consistent with existing literature that challenges the traditional focus on realism. Norman et al [5] and Hamstra et al [6] both caution against overemphasizing fidelity, highlighting instead the importance of instructional relevance and cognitive congruence. Petrosoniak et al [7] emphasize a systems-level view, advocating for simulation design that is responsive to functional demands rather than esthetic realism. These frameworks collectively support our goal-oriented approach, reinforcing the idea that the effectiveness of a simulation lies not in its resemblance to reality, but in how well it facilitates the intended learning outcomes [25,26]. Simulation modalities must be intentionally selected based on their alignment with specific learning goals rather than technological sophistication alone [27,28].

The inherent limitations of fidelity and the near impossibility of making a simulation fully resemble real life compel instructors to align simulation design with educational objectives. For example, in a simulation focused on a critical clinical situation requiring urgent intervention, the emphasis should be on the progression of the patient's clinical deterioration, the simulated patient's vital signs, and the dynamic clinical response to treatment. Conversely, in a simulation aimed at managing the emotional response of a patient or family members discovering the death of a loved one should prioritize the performance of the actors portraying the simulated family [29]. In the first case, the use of an advanced patient simulator is essential, whereas actors are more appropriate in the second activity. In the same vein, a task trainer will be most useful when acquiring the skill of vaginal examination [30]. Therefore, fidelity should be tailored to the learning goal, rather than relying on the use of an advanced patient simulator. A highly technological simulation is not necessarily synonymous with high fidelity and may, in fact, lead to resource overconsumption [31] with limited educational benefit. To maximize impact, simulation educators must keep learning objectives at the center, carefully considering the most effective simulation method—be it actors, advanced patient simulators, or task trainers—and aligning the fidelity to meet that objective [32-34].

Our argument, that higher fidelity is not inherently superior and may even be counterproductive when misaligned with learning goals, is consistent with a growing body of literature. Studies have demonstrated that the educational impact of simulation does not correlate directly with its realism, but rather with how well the simulation supports cognitive and procedural learning; these studies advocate for simulation design that serves specific system and learner needs rather than aiming for maximal fidelity [5-7].

However, we acknowledge that this perspective is not without debate. Some educators argue that high-fidelity simulations can enhance psychological immersion or better prepare learners for complex clinical environments. While this may be true in specific contexts, such as team-based crisis management or emotional preparedness, we suggest that these benefits arise not from fidelity per se, but from how the design elements align with targeted competencies. In fact, pursuing high levels of realism without clear pedagogical justification can lead to resource inefficiencies or even cognitive overload, detracting from the intended learning outcomes.

Rather than taking a position against fidelity, our goal is to promote intentionality in its use. By considering fidelity as a flexible, goal-oriented construct rather than a fixed technological

XSL•FO
RenderX

standard, educators can make more informed decisions that balance effectiveness, efficiency, and learner experience.

Conclusion

Simulation fidelity is multidimensional, encompassing physical, emotional, and contextual dimensions that shape learners' perception of realism. However, fidelity alone does not ensure effective learning. Rather than maximizing realism, educators should redefine high-fidelity simulation as the deliberate alignment of fidelity with pedagogy, treating fidelity not as a fixed standard but as a flexible design tool that serves pedagogical intent above all.

Acknowledgments

JP acknowledges the French Society of Anesthesiology and Intensive Care for its support through the award of a mobility grant that funded a simulation fellowship at the University of Montreal.

Funding

No external financial support or grants were received from any public, commercial, or not-for-profit entities for the research, authorship, or publication of this article.

Conflicts of Interest

None declared.

References

1. Ten Cate O, Carraccio C, Damodaran A, et al. Entrustment decision making: extending Miller's pyramid. Acad Med 2021 Feb 1;96(2):199-204. [doi: 10.1097/ACM.0000000000003800] [Medline: 33060399]
2. Bergsteiner H, Avery GC. The twin-cycle experiential learning model: reconceptualising Kolb's theory. Studies in Continuing Education 2014 Sep 2;36(3):257-274. [doi: 10.1080/0158037X.2014.904782]
3. Fenzi G, Alemán-Jiménez C, Cayuela-Fuentes PS, Segura-López G, Leal-Costa C, Díaz-Agea JL. The expository phase of debriefing in clinical simulation: a qualitative study. BMC Nurs 2025 Apr 30;24(1):476. [doi: 10.1186/s12912-025-03067-z] [Medline: 40307824]
4. Lioce L, Lopreiato J. Healthcare simulation dictionary. In: Healthcare Simulation Dictionary, 3rd edition: Agency for Healthcare Research and Quality; 2024. URL: https://www.ssih.org/sites/default/files/2025-03/Healthcare-Simulation-Dictionary-3.pdf [accessed 2026-02-19]
5. Norman G, Dore K, Grierson L. The minimal relationship between simulation fidelity and transfer of learning. Med Educ 2012 Jul;46(7):636-647. [doi: 10.1111/j.1365-2923.2012.04243.x] [Medline: 22616789]
6. Hamstra SJ, Brydges R, Hatala R, Zendejas B, Cook DA. Reconsidering fidelity in simulation-based training. Acad Med 2014 Mar;89(3):387-392. [doi: 10.1097/ACM.0000000000000130] [Medline: 24448038]
7. Petrosoniak A, Brydges R, Nemoy L, Campbell DM. Adapting form to function: can simulation serve our healthcare system and educational needs? Adv Simul (Lond) 2018;3:8. [doi: 10.1186/s41077-018-0067-4] [Medline: 29942658]
8. Dieckmann P, Gaba D, Rall M. Deepening the theoretical foundations of patient simulation as social practice. Simul Healthc 2007;2(3):183-193. [doi: 10.1097/SIH.0b013e3180f637f5] [Medline: 19088622]
9. Lewis R, Strachan A, Smith MM. Is high fidelity simulation the most effective method for the development of non-technical skills in nursing? A review of the current evidence. Open Nurs J 2012;6:82-89. [doi: 10.2174/1874434601206010082] [Medline: 22893783]
10. Rudolph JW, Simon R, Raemer DB. Which reality matters? Questions on the path to high engagement in healthcare simulation. Simul Healthc 2007;2(3):161-163. [doi: 10.1097/SIH.0b013e31813d1035] [Medline: 19088618]
11. Pawar S, Jacques T, Deshpande K, Pusapati R, Meguerdichian MJ. Evaluation of cognitive load and emotional states during multidisciplinary critical care simulation sessions. BMJ Simul Technol Enhanc Learn 2018 Apr;4(2):87-91. [doi: 10.1136/bmjstel-2017-000225] [Medline: 29670763]
12. Dhaif F, Paparoidamis G, Sideris M, et al. The role of anxiety in simulation-based dexterity and overall performance: does it really matter? J Invest Surg 2019 Mar;32(2):164-169. [doi: 10.1080/08941939.2017.1387624] [Medline: 29286827]
13. Newton J, Smith A. Developing emotional preparedness and mental resilience through high-fidelity simulation: a "bridge too far" for institutions teaching major trauma management and mass-casualty medicine? BMC Med Educ 2024 May 15;24(1):544. [doi: 10.1186/s12909-024-05526-8] [Medline: 38750542]
14. Peek R, Moore L, Arnold R. Psychophysiological fidelity: a comparative study of stress responses to real and simulated clinical emergencies. Med Educ 2023 Dec;57(12):1248-1256. [doi: 10.1111/medu.15155] [Medline: 37392166]
15. Daglius Dias R, Scalabrini Neto A. Stress levels during emergency care: a comparison between reality and simulated scenarios. J Crit Care 2016 Jun;33:8-13. [doi: 10.1016/j.jcrc.2016.02.010] [Medline: 26987261]

XSL•FO
RenderX

16. Frerk C, Mitchell VS, McNarry AF, et al. Difficult Airway Society 2015 guidelines for management of unanticipated difficult intubation in adults. Br J Anaesth 2015 Dec;115(6):827-848. [doi: 10.1093/bja/aev371] [Medline: 26556848]
17. Tran SHN, Fernandes MA. Drawing enhances memory for emotional words. Can J Exp Psychol 2023 Dec;77(4):296-307. [doi: 10.1037/cep0000319] [Medline: 37917424]
18. Sattar R, Lawton R, Janes G, et al. A systematic review of workplace triggers of emotions in the healthcare environment, the emotions experienced, and the impact on patient safety. BMC Health Serv Res 2024 May 9;24(1):603. [doi: 10.1186/s12913-024-11011-1] [Medline: 38720302]
19. Dadiz R, Bender J, Robin B. Simulation-based operations testing in new neonatal healthcare environments. Semin Perinatol 2023 Nov;47(7):151828. [doi: 10.1016/j.semperi.2023.151828] [Medline: 37775365]
20. Liang JN, Ciampa M, Kobylarz F, et al. Impact of live tissue training on provider confidence for operative trauma management. Mil Med 2025 Feb 27;190(3-4):e784-e789. [doi: 10.1093/milmed/usae403] [Medline: 39172654]
21. Roach VA, Traina DJ, Hananel DM, Speich JR, Norfleet JE, Sweet RM. Endotracheal intubation performance in novices: comparing the training effect resulting from two airway management training simulators. Mil Med 2024 Aug 19;189(Suppl 3):431-438. [doi: 10.1093/milmed/usae148] [Medline: 39160866]
22. Duarte HMS, Castanheira JA, Pereira ASF, Pragosa Â, Santos ETP, Dixe MDA. Comparative study between high-fidelity simulation and medium-fidelity simulation in decision-making of nursing students: experimental study. Rev Lat Am Enfermagem 2024;32:e4269. [doi: 10.1590/1518-8345.6847.4269] [Medline: 39140563]
23. Rudolph JW, Raemer DB, Simon R. Establishing a safe container for learning in simulation: the role of the presimulation briefing. Simul Healthc 2014 Dec;9(6):339-349. [doi: 10.1097/SIH.0000000000000047] [Medline: 25188485]
24. Hughes PG, Hughes KE. Briefing Prior to Simulation Activity: StatPearls; 2024. URL: https://www.ncbi.nlm.nih.gov/books/NBK545234/ [accessed 2026-02-19]
25. Tun JK, Alinier G, Tang J, Kneebone RL. Redefining simulation fidelity for healthcare education. Simul Gaming 2015 Apr;46(2):159-174. [doi: 10.1177/1046878115576103]
26. Alinier G, Oriot D. Simulation-based education: deceiving learners with good intent. Adv Simul (Lond) 2022 Mar 18;7(1):8. [doi: 10.1186/s41077-022-00206-3] [Medline: 35303963]
27. Sahakian G, Buléon C, Alinier G. Chapter 14 - educational foundations of instructional design applied to simulation-based education. In: Chiniara G, editor. Clinical Simulation, 2nd edition: Academic Press; 2019:185-206. [doi: 10.1016/B978-0-12-815657-5.00014-0]
28. Watts PI, McDermott DS, Alinier G, et al. Healthcare simulation standards of best practice simulation design. Clin Simul Nurs 2021 Sep;58:14-21. [doi: 10.1016/j.ecns.2021.08.009]
29. Lateef F, Peckler B, Saindon E, et al. The art of sim-making: what to learn from film-making. J Emerg Trauma Shock 2022;15(1):3-11. [doi: 10.4103/jets.jets_153_21] [Medline: 35431474]
30. Hill E, Moreau A, Martin L, Papin-Groseil J. A prospective study of the acquisition of vaginal examination skills using simulation. Int J Gynaecol Obstet 2025 Feb;168(2):650-655. [doi: 10.1002/ijgo.15874] [Medline: 39161282]
31. Inam H, Arshad H, Waqar U, Ahmad Zaidi SM, Ahmed S. Training in cardiothoracic surgery: the role for simulation. J Pak Med Assoc 2024 Jul;74(7):1316-1320. [doi: 10.47391/JPMA.10049] [Medline: 39028061]
32. Nevins NA, Roche R, Dailey SF, et al. Using full dive virtual reality to operationalize warfighter resilience: from proof of concept and usability of hardware and software to upcoming integrated psychological skills training. Mil Med 2024 Aug 19;189(Suppl 3):480-488. [doi: 10.1093/milmed/usae158] [Medline: 39160849]
33. Roche AF, Redmond T, Zilani G, Healy V, Condron CM. Developing a high fidelity, low cost simulation model for retrosigmoid craniotomy and microvascular decompression of the trigeminal nerve. Br J Neurosurg 2025 Dec;39(6):828-831. [doi: 10.1080/02688697.2024.2391858] [Medline: 39155267]
34. Kaddoura R, Faraji H, Otaki F, et al. High-fidelity simulation versus case-based tutorial sessions for teaching pharmacology: convergent mixed methods research investigating undergraduate medical students' performance and perception. PLoS One 2024;19(8):e0302609. [doi: 10.1371/journal.pone.0302609] [Medline: 39150900]

Edited by B Lesselroth; submitted 23.Sep.2025; peer-reviewed by F Ismail, G Alinier, R Brydges; revised version received 15.Jan.2026; accepted 02.Feb.2026; published 23.Feb.2026.

Please cite as:
Pico J, Evain JN, Aron C, Martin G, Cruz-Panesso I, Georgescu LM, Tanoubi I
From Realism to Learner Engagement: Rethinking Fidelity in Simulation-Based Education
JMIR Med Educ 2026;12:e84684
URL: https://mededu.jmir.org/2026/1/e84684
doi:10.2196/84684

XSL•FO
RenderX

© Julien Pico, Jean-Noel Evain, Christina Aron, Gilles Martin, Ilian Cruz-Panesso, Leonida-Mihai Georgescu, Issam Tanoubi. Originally published in JMIR Medical Education (https://mededu.jmir.org), 23.Feb.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Education, is properly cited. The complete bibliographic information, a link to the original publication on https://mededu.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Voices of Innovation: Reflective Report on Integrating Artificial Intelligence–Simulated Mental Health Patient Scenarios Into Undergraduate Nursing Education in the United Arab Emirates

Amina Ahmad[1], MSc; Janisha Kavumpurath[1], PhD; Raheesa Kader[1], MSc; Muna Altamimi[1], PhD; Monia El Hajj[1], BSc; Fatma Refaat Ahmed Ahmed[1,2], PhD; Muhammad Arsyad Subu[1], PhD; Taliaa Yafei[3], MSc; Hind Rashed Ali[1], BSc; Idil Saleh[1], MScN; Nabeel Al-Yateem[1], PhD

[1]Department of Nursing, College of Health Sciences, University of Sharjah, University City Road, P.O. Box 27272, Sharjah, University City, United Arab Emirates

[2]Department of Critical Care and Emergency Nursing, Critical Care & Emergency Nursing, Faculty of Nursing, Alexandria University, Alexandria, Egypt

[3]Faculty of Nursing, Ras al-Khaimah Medical and Health Sciences University, Ras al-Khaimah, United Arab Emirates

Corresponding Author:
Amina Ahmad, MSc
Department of Nursing, College of Health Sciences, University of Sharjah, University City Road, P.O. Box 27272, Sharjah, University City, United Arab Emirates

Abstract

Limited clinical placements for mental health courses in the United Arab Emirates have made it difficult to provide consistent experiential learning for undergraduate nursing students. As a result, nurse educators are considering technology-enabled learning approaches to deliver clinical skills training. This Viewpoint presents a reflective, theory-informed account of the first-year integration of an artificial intelligence (AI)–enabled, voice-interactive simulated patient into an undergraduate mental health nursing practicum. Grounded in Kolb's experiential learning cycle and aligned with established simulation best practices, the initiative was designed to support therapeutic communication, psychiatric assessment, and clinical reasoning through structured prebriefing, immersive interaction, and guided debriefing. The paper describes the educational rationale, scenario development, implementation processes, and contextual challenges encountered during real-world deployment across university and clinical environments. AI-supported simulations offered a standardized and psychologically safe context for students to engage with complex psychiatric scenarios, particularly when direct patient interaction is constrained. We discuss operational insights related to technical reliability, environmental requirements, faculty preparation, and assessment integration alongside considerations for scalability and sustainability in resource-limited settings. While AI-supported objective structured clinical examinations have been incorporated to support assessment consistency, formal psychometric validation and outcome comparisons have not been undertaken at this stage. By sharing lessons learned from early implementation, this Viewpoint contributes practical insights for nursing educators facing similar structural constraints. AI-enabled simulation is presented as a strategic complement to, rather than a replacement for, traditional clinical placements, with future empirical research needed to evaluate educational outcomes and long-term impact.

(JMIR Med Educ 2026;12:e78161) doi:10.2196/78161

KEYWORDS

mental health nursing; simulation training; artificial intelligence; AI; nursing education; clinical skills; experiential learning

Introduction and Background

The United Arab Emirates (UAE) admitted its first cohort of bachelor's-level nursing students in 1998. At that time, enrollment was small, and the limited number of hospitals could accommodate clinical training needs [1]. Since then, rapid population growth to approximately 10 million—largely driven by expatriates—together with major investment in health services has been matched by a sharp expansion in nursing education. More than 15 nursing training institutions are now operating in the country [1]. Although the number of approved clinical training sites has increased, it has not kept pace with the growth in programs [1]. As a result, many institutions have turned to simulation-based training due to competition for clinical placements.

This challenge is not unique to the UAE. Globally, nursing programs continue to face limits in clinical capacity due to workforce shortages, rising student enrollment, and increasing patient acuity. Recent literature also suggests that relying solely on clinical simulation to address these pressures is not

XSL•FO
RenderX

sustainable, particularly as both programs and student numbers continue to grow [2-4].

At the same time, simulation technologies have advanced. In addition to low- and medium-fidelity simulators, newer options such as virtual reality (VR) and artificial intelligence (AI)–powered conversational agents have broadened how simulation can be used in nursing education [5,6]. For these technologies to have maximum educational value, however, they should be integrated as a structured learning opportunity rather than treated as an "add-on." In practice, this means aligning their use with Kolb's experiential learning cycle: concrete experience, reflective observation, abstract conceptualization, and active experimentation.

Psychiatric nursing education presents additional challenges that are less common in other areas of health care education. Limited psychiatric facilities, complex ethical issues, student vulnerability, and the risk of exposure to sensitive cases reduce opportunities for repeated practice of key interventions [7,8]. Evidence indicates that simulation in psychiatric education can offer a controlled environment where learners can practice communication, emotional regulation, and clinical decision-making skills when real clinical exposure is limited [7,9,10].

These constraints are particularly evident in the UAE, where psychiatric clinical placements are relatively few. Providing every student with high-quality, guided exposure in mental health settings is therefore difficult. To reduce placement pressure while maintaining—and potentially improving—learning quality, we pilot tested a technology-based simulation approach. Specifically, we used high-immersion VR environments combined with AI-based simulated patients to support complex psychiatric interactions at scale.

Early evidence suggests that AI-simulated patients can improve realism, standardization, and scalability, offering psychologically safe opportunities for deliberate practice in communication and clinical reasoning [11,12]. During periods of system strain, such as the COVID-19 pandemic, AI applications aligned with Industry 4.0 have also been used to support health care work processes by strengthening service delivery infrastructure rather than replacing professional judgment [13,14].

This paper describes our first-year experience embedding an AI voice interaction system—designed to simulate a patient with a mental disorder—into a mental health nursing practicum course. Using a reflective approach grounded in relevant theory, we examine the implementation processes and challenges associated with integrating AI simulation technologies into mental health nursing education, particularly within structural limits related to clinical placement capacity.

The Indispensable Role of Clinical Training

Practical experience is the foundation of nursing education. It allows students to translate theory into patient care while building psychomotor, cognitive, and interpersonal skills and strengthening professional identity, confidence, and competence [15,16]. However, deliberate practice is often limited by institutional inefficiencies, poorly aligned programs, and insufficient supervision. These constraints can reduce student satisfaction and slow skills development [17,18].

These pressures are most pronounced in mental health nursing. With only 3 dedicated psychiatric facilities across the country, students may face stigma, inconsistent staff attitudes, and high stress, which can contribute to anxiety or even avoidance of the specialty despite the need for well-prepared practitioners [19,20]. Simulation can help bridge this gap by offering a controlled environment—using manikins and standardized patients—where learners can practice safely, reduce anxiety, and build competence. However, traditional simulation is often resource intensive [21,22].

Newer immersive approaches may extend practice opportunities further. VR can simulate complex and dynamic psychiatric scenarios, while AI-powered chatbots can support adaptive, realistic therapeutic conversations. Together, these tools can expand access to intentional practice and reflection, particularly where clinical exposure is limited [23-25].

The UAE Context

In the UAE, particularly in the Northern Emirates, the number of inpatient psychiatric beds has traditionally been limited. Prior to 2020, nursing students were placed on rotation in only 2 specialist psychiatric hospitals—both of which already shared a limited number of spots with medical and allied health programs. Just as the demand for these placements became acute in the aftermath of the COVID-19 pandemic—due to a marked rise in postpandemic nursing enrollments [25]—the number of available placements plummeted. At the University of Sharjah, only 35% of students enrolled in the 2023 - 2024 practicum in mental health nursing were able to secure a traditional placement.

Outpatient psychiatric clinics offered little relief. As some of these clinics are part of for-profit multispecialty hospitals, patients and families often refuse to cooperate with student interviews to avoid the stigma associated with being labeled "psychiatric." This has put student cohorts at risk of not meeting course objectives related to conducting mental status examinations (MSEs) or practicing de-escalation skills—skills they may only perform occasionally, if at all.

The combination of (1) a chronic shortage of psychiatric training facilities, (2) the diversion of such facilities during the pandemic, and (3) an increasing number of students has created a pressing need for an effective and scalable simulation option. Fortunately, the current state of VR simulation and AI-driven simulated patients has reached a level of maturity that allows for the realistic replication of complex psychiatric scenarios.

This reflection, therefore, assesses the introduction of an AI-based VR simulation program within an undergraduate mental health nursing practicum course. The aim is to evaluate whether immersive, intelligent simulation can bridge the clinical exposure gap, ensure that students meet identified competencies,

XSL•FO
RenderX

and establish a sustainable framework for teaching mental health nursing in developing contexts such as the UAE.

The Experience: Development of the AI-Driven Solution

Simulation has proven to be of enormous benefit to the nursing industry. The maturity and rapid evolution of AI technologies, combined with the severe shortage of mental health training placements, were the primary drivers behind the introduction of AI-driven simulated patients that can interact with students through voice.

Prior to implementation, an intensive preimplementation workshop was conducted over 1 week with all participants. This workshop reinforced core concepts in mental health care, including models of therapeutic communication, MSE principles, patient education strategies, risk assessment within MSEs, and the application of the nursing process to mental health scenarios. In parallel, participants attended theoretical classes twice a week focused on mental health nursing concepts. This convergence of activities during the preimplementation phase ensured that students were adequately equipped with foundational knowledge, thereby supporting improved critical thinking and readiness to engage with AI-based simulated patient interactions.

The next step involved the creation of realistic psychiatric patient scenarios. Using diagnoses and symptoms identified in the *Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition, Text Revision* (*DSM-5-TR*) reference guide [26], scenarios were developed to reflect common mental health conditions such as schizophrenia, bipolar I disorder, and major depressive disorder with psychotic features.

Each patient scenario included the following elements: (1) in-depth information about the patient's history, (2) detailed accounts of past mental and emotional circumstances, and (3) clinical information, including pharmacological management.

An example of the schizophrenia patient simulation is presented in Textbox 1.

To enhance content validity, these scenarios were developed in collaboration with mental health faculty members and university professors with practical experience as psychiatric mental health nurses.

Textbox 1. Simulation scenario.

Age and gender: 24-year-old female

Presenting concern: "They're watching me through the ceiling vents. I can hear them talking about me."

Background

Ms. L, a 24-year-old university student, was brought to the emergency department by her mother following several weeks of increasingly unusual behavior. Her mother reports that Ms. L has become socially withdrawn, staying isolated in her room, skipping university classes, and speaking less. In the last two weeks, she began covering her bedroom vents with tape and voiced fears that people were monitoring her. She refuses food prepared by family, citing contamination concerns. She is occasionally heard speaking or laughing when alone. During the assessment, she avoids eye contact, speaks slowly, and appears distracted by internal stimuli. She denies suicidal ideation but displays significant mistrust. She has no prior psychiatric history or substance use, though a maternal uncle was diagnosed with schizophrenia.

Medications

- Olanzapine 10 mg at bedtime (initiated)
- Lorazepam 1 mg (as needed) for agitation
- (Note: initially resistant to treatment; currently partially adherent under supervision)

Nonverbal cues

- Frequently looks around the room suspiciously
- Avoids direct eye contact
- Flat affect
- Occasionally murmurs to herself
- Limited verbal output

Assessment and learning objectives

- Apply therapeutic communication techniques to establish rapport
- Conduct a comprehensive mental status examination
- Assess safety risks and patient needs based on behavioral cues
- Provide psychoeducation on medication adherence and symptom management
- Identify the *Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition, Text Revision*, diagnostic features of schizophrenia based on the patient scenario

XSL•FO
RenderX

A set of simulation prompts was also designed to promote authentic responses during student interactions. Each prompt instructed the AI to fully embody the psychiatric patient's persona, displaying appropriate emotional expressions, speech patterns, and symptoms consistent with the diagnosed condition. Furthermore, the prompts included descriptions of nonverbal behaviors—such as sighing, looking away, leaning forward, or avoiding eye contact—prior to verbal responses. Accurately interpreting nonverbal cues is a critical component of therapeutic communication.

The AI simulation prompt is illustrated in Textbox 2.

Textbox 2. Standardized simulation prompt.

"You are going to role-play a psychiatric patient for a clinical nursing simulation. I will give you a case scenario, and you must respond strictly in character as the patient described.

Use realistic voice tone, emotion, and language that reflect the symptoms and emotional state of the patient.

Begin by giving a brief behavioral description before each verbal response—for example: 'sighs and looks away before speaking' or 'fidgets with hands, then answers softly'.

These nonverbal behaviors are important for the student to assess. Do not speak unless the student initiates the conversation.

Respond naturally to the student's questions, but remain within the emotional and behavioral context of the case.

You must not break character. Speak only in English."

Implementation and Integration Into Clinical Education

Students were scheduled for structured AI simulation activities as part of their clinical practicum hours. They were provided with minimal background information prior to participation—mirroring actual psychiatric practice, where building rapport and using effective communication strategies are essential to obtain a comprehensive patient history.

Before each structured simulation activity, a prebriefing session was conducted to orient students to key aspects related to learning objectives, psychological safety, and appropriate communication strategies, particularly in relation to risk and professionalism. During the simulation, students engaged in conversations with AI-simulated patients to conduct comprehensive MSEs; deliver psychoeducation on medication adherence; and perform risk assessments for suicidal ideation, self-harm, and aggression. On the basis of their assessments, students developed individualized patient care plans.

Faculty members from the school observed these simulation activities either live or via recordings. Each simulation was followed by a debriefing session, which facilitated reflective learning and enabled students to critically evaluate their clinical decision-making processes and develop more effective communication strategies.

While the initial plan was to conduct these AI simulations solely at the university, a need arose to integrate them into students' hospital rotations. In outpatient settings, particularly within private hospitals, students often encountered barriers to direct patient interaction. The stigma surrounding mental illness led some patients to decline student participation in interviews, even though these hospitals were not strictly mental health facilities. Many patients also sought services discreetly, further limiting opportunities for students to participate in consultations. These factors posed challenges to the development of students' therapeutic communication skills.

To address these issues and maintain continuity in learning, AI simulations began to be used within hospital environments. However, logistical challenges emerged. Unreliable Wi-Fi connections often disrupted simulation sessions, leading to delays or misinterpretations by the AI tool. Finding suitable, private locations for simulation within hospitals was also difficult, and background noise occasionally interfered with the AI's speech recognition capabilities.

Despite these operational challenges, the implementation of AI simulations in medical training remained largely effective. Students gained valuable experience in therapeutic communication, psychiatric assessment, risk evaluation, and critical thinking—skills they might not have been able to develop otherwise.

Extension Into Assessment: AI-Supported Objective Structured Clinical Examinations

On the basis of the success achieved through AI-powered simulated patient interactions during clinical training, the simulated activity was incorporated into the assessment framework through the use of objective structured clinical examinations (OSCEs). To ensure a fair platform for all examinees, students participated in a set of stations designed to assess various skills in the domain of psychiatric nursing competence.

Each participant was randomly assigned a case scenario from a pool of prepared cases. Each scenario contained core clinical information typically found in a hospital patient file, including the presenting problem, relevant history, psychiatric diagnosis, current medications, and risk factors. Students were given 5 minutes of preparation time before beginning the clinical activity based on the case.

The first station assessed therapeutic communication skills. Students engaged in a live voice interaction exercise with the AI-simulated patient, applying key concepts such as empathy, active listening, validation, and rapport building. Performance was evaluated using a structured rubric.

XSL•FO
RenderX

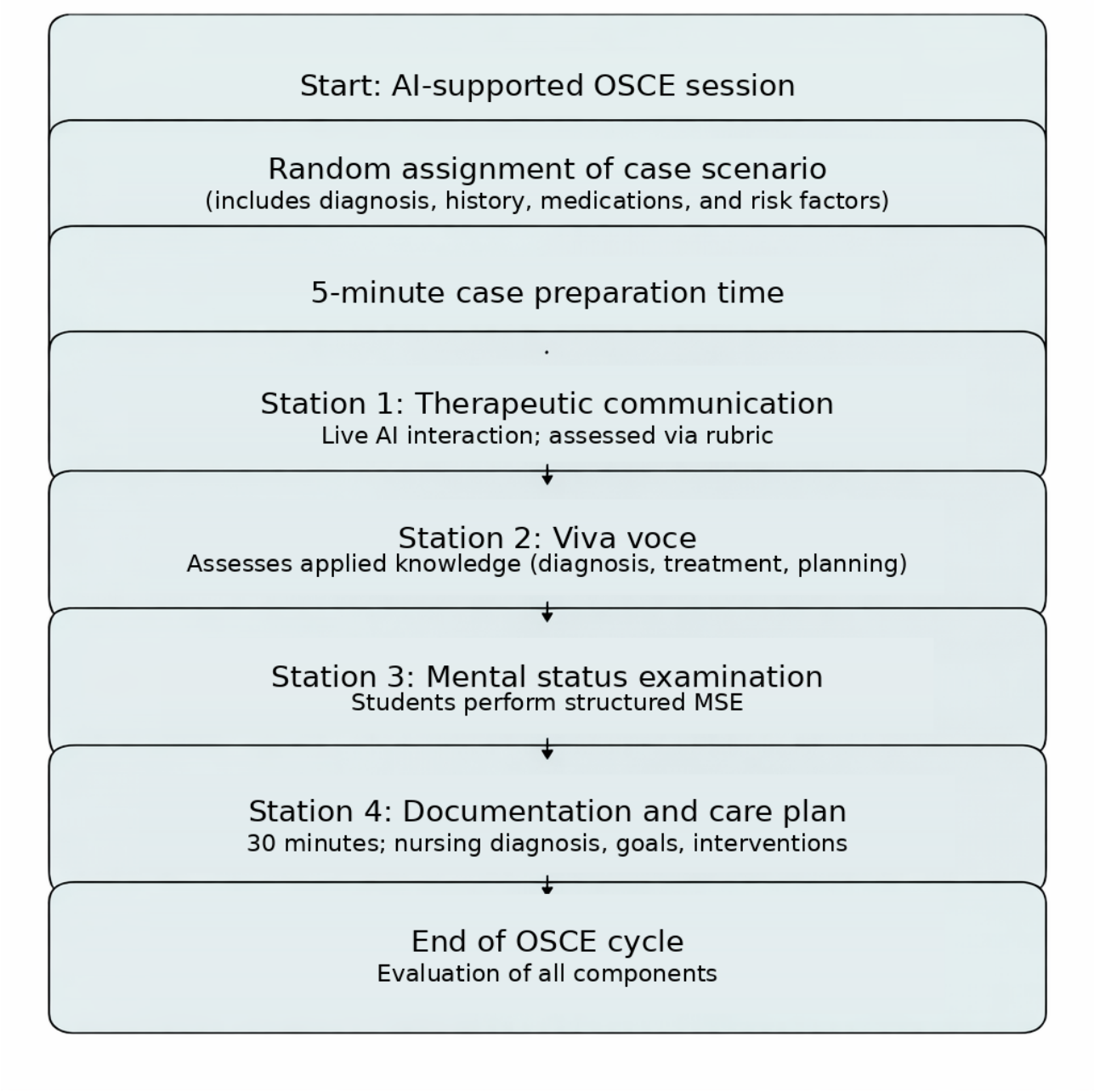
The second station was a viva voce session in which students' applied knowledge of the case was evaluated. They were asked questions to assess their understanding of the clinical presentation, differential diagnosis, psychopharmacological approach, and evidence-based nursing management.

The third station involved the demonstration of a full MSE. Students conducted a guided MSE on the AI-simulated patient through active engagement in a voice-only conversation. Although the interaction was audio based, the prompts were designed to include brief behavioral cues (eg, "fidgets with hands" and "avoids eye contact") preceding verbal responses. This allowed students to evaluate domains such as appearance, behavior, speech, mood, affect, thought process, cognition, insight, and judgment—drawing on both content and contextual cues. This setup encouraged the use of clinical judgment under realistic conditions.

Following the simulation stations, students were given 30 minutes to record their assessment findings and develop a corresponding nursing care plan. This component measured their ability to synthesize information from the simulation and formulate an appropriate care plan.

The inclusion of AI-led simulated patients throughout the OSCE ensured consistency in the assessment process and eliminated variability commonly associated with live patients or inconsistencies in standardized patient performances. The structural layout of the OSCE is illustrated in Figure 1.

Figure 1. Flowchart of artificial intelligence (AI)–integrated objective structured clinical examinations (OSCE) structure in mental health nursing education. MSE: mental status examination.



Although the AI-supported OSCEs were intended to enhance standardization and authenticity in student assessment, the

XSL•FO
RenderX

current implementation did not include formal validation, reliability testing, or outcome comparisons. These factors should be addressed in future studies.

Opportunities Created by AI Integration

The integration of AI simulation provided a range of learning opportunities that extended beyond simply addressing the clinical placement gap. From an instructional design perspective, the simulation offered a balanced opportunity for students to engage with complex mental health scenarios—particularly important in contexts where hospital placements are dependent on availability. It allowed for consistent exposure to essential mental health nursing skills.

The model also introduced flexibility in curriculum delivery. The use of AI simulations enabled the incremental development and deployment of patient scenarios, allowing faculty to adapt to changes in enrollment without placing additional strain on clinical partners. This is particularly valuable in environments where expanding clinical capacity is not feasible, even as student numbers increase.

Another key learning affordance was the provision of immediate feedback. Faculty members were able to observe students during simulation activities and conduct debriefing sessions focused on clinical decision-making, communication strategies, and the application of professional judgment.

Importantly, the simulations fostered an environment of psychological safety in which students could develop their therapeutic communication and psychiatric assessment skills. Some students reported feeling anxious about interacting with clients with mental illness during their initial clinical exposures. Practicing these interactions in a controlled, simulated setting helped them build familiarity and confidence.

Beyond mental health education, this strategy shows potential for broader application across other areas of nursing education, including physical assessment and adult medical-surgical nursing. Incorporating AI-enabled voice interaction into existing simulation systems could further enhance the learning experience.

These observations highlight the educational benefits of AI simulation but do not evaluate its overall effectiveness, underscoring the need for future research and formal evaluation.

Challenges and Lessons Learned

Although the AI-based simulation model provided valuable learning opportunities, several challenges emerged during its application, many of which informed improvements in later versions.

Reliability was identified as a fundamental requirement. Stable Wi-Fi connectivity was essential for uninterrupted speech communication; however, occasional network failures occurred. These disruptions, particularly during periods of high use, suggested the potential need for multiple subscription services to ensure consistent access.

Environmental factors also influenced the quality of the simulations. Conducting sessions in sound-absorbing environments proved critical, as background noise in real-world settings sometimes interfered with speech recognition, leading to inconsistent or inaccurate responses. This highlights the importance of designated simulation spaces to maintain high-quality interactions.

Limitations in nonverbal communication were another noted issue. While verbal communication was effectively simulated, the lack of dynamic facial expressions and gestures reduced the realism of certain interactions. Although prompts included descriptions of nonverbal cues before verbal responses, future enhancements could incorporate AI avatars or graphical interfaces to improve the simulation of nonverbal behaviors.

Faculty readiness was also a key consideration. Successful facilitation required educators not only to be proficient with the technology but also to manage essential instructional tasks, such as conducting prebriefs, monitoring learner engagement, leading debriefings, and aligning the simulation with learning outcomes.

In light of these challenges and limitations, AI-enhanced simulation should be viewed as a *complementary strategy* in clinical training rather than a substitute for traditional clinical experiences. Its value lies in augmenting, not replacing, real-world learning.

Reflection and Future Directions

The inclusion of ChatGPT Voice in the Mental Health Nursing Practicum course at the University of Sharjah was a major milestone in applying innovation to nursing education. The innovation sought to fill the gap that had been experienced in psychiatric clinical placements by applying AI to design a dynamic virtual environment to better link theory and practice.

Looking ahead, there are a number of opportunities that can be identified to further improve the educational quality of such a simulation-based system. Increasing the number of clinical simulations available that relate to a broader set of psychiatric disorders can potentially create a more inclusive learning environment. The addition of algorithms related to theories of emotional intelligence can also increase the realism of emotionally based patient reactions, thereby enhancing the level of therapeutic involvement that can take place during a student-patient interaction.

Future studies might also examine student satisfaction and learning results using AI-supported simulation activities. These studies would offer significant insights into the efficacy and continuing improvement of such educational instruments.

Notably, the new approach aligns with the international standards that promote the convergence of technology and creativity in medical education programs [27,28]. Furthermore, the approach aligns with the UAE Vision 2021 and Centennial 2071 strategies in innovation and the convergence of AI in health-related education programs in the UAE [29]. In fact, it aligns with the Strategic Plan 2024 - 2030 of the University of Sharjah [30].

XSL•FO
RenderX

Conclusions

With unprecedented challenges experienced in sourcing clinical placement opportunities for nurses, the ChatGPT Voice Interaction simulation embodies the potential for AI simulation to offer a sustainable approach to mental health nurse education.

Through its structured implementation based on theories, the approach enabled opportunities for the engagement of skills in therapeutic communication, psychiatric assessment, and planning for a safe and fair learning context.

This project shows how new technology could be conceptualized to work well with existing approaches to education, especially where access to clinical training facilities is limited. The developing nature of technology means that well-structured uptake of AI simulation learning could help to ensure education continuity by preparing nursing graduates to handle the growing complexity of modern-day health care.

Acknowledgments

The authors declare the use of generative artificial intelligence (GenAI) in the research and writing process. According to the Generative Artificial Intelligence Delegation Taxonomy (GAIDeT; 2025), the following tasks were delegated to GenAI tools under full human supervision: proofreading and editing. The GenAI tool used was ChatGPT's Plus and Pro subscription versions, with voice interaction capability. Responsibility for the final manuscript lies entirely with the authors. The GenAI tools are not listed as authors and do not bear responsibility for the final outcomes.

Funding

This study was internally funded by the College of Health Sciences for simulation and educational purposes. The funder had no role in the study design; data collection, analysis, or interpretation; manuscript preparation; or the decision to submit the manuscript for publication.

Authors' Contributions

Conceptualization: AA, MAS, NA-Y
Investigation: AA, HRA, IS, JK, MA, MMEH, RK, TY
Methodology: AA, FRAA, MAS, NA-Y
Project administration: AA
Resources: MA, TY
Supervision: AA, MAS, NA-Y
Validation: AA, FRAA, JK
Visualization: IS, MMEH
Writing – original draft: AA
Writing – review and editing: AA, FRAA, HRA, IS, JK, MA, MAS, MMEH, NA-Y, RK, TY

Conflicts of Interest

None declared.

References

1. Al-Yateem N, Almarzouqi A, Dias JM, Saifan A, Timmins F. Nursing in the United Arab Emirates: current challenges and opportunities. J Nurs Manag 2021 Mar;29(2):109-112. [doi: 10.1111/jonm.12984] [Medline: 32100891]
2. Daneshfar M, Moonaghi HK. The impact of clinical simulation on bridging the theory-practice gap in nursing education: a systematic review. BMC Med Educ 2025 Aug 28;25(1):1216. [doi: 10.1186/s12909-025-07790-8] [Medline: 40877870]
3. Jallad ST. Effectiveness of simulation-based education on educational practices of communication skills, satisfaction, and self-confidence among undergraduate nursing students. Creat Nurs 2025 May;31(2):135-143. [doi: 10.1177/10784535241301115] [Medline: 39639530]
4. Kindi ZA, Jabri WA, Zadjali MA, Muliira J. Utilizing simulation in community health nursing education: a scoping review of current trends and applications. BMC Nurs 2025 Aug 12;24(1):1055. [doi: 10.1186/s12912-025-03718-1] [Medline: 40796842]
5. Lampropoulos G, Fernández-Arias P, Del Bosque A, Vergara D. Augmented reality in health education: transforming nursing, healthcare, and medical education and training. Nurs Rep 2025 Aug 8;15(8):289. [doi: 10.3390/nursrep15080289] [Medline: 40863676]
6. Chang W, Lin CC, Crilly J, Lee HL, Chen LC, Han CY. Virtual reality simulation for undergraduate nursing students for care of patients with infectious diseases: mixed methods study. JMIR Med Educ 2025 Feb 11;11:e64780. [doi: 10.2196/64780] [Medline: 39933166]

XSL•FO
RenderX

7. Alexander L, Coyte B, Evans A, Dickson J, Guinea S, Foster K. The educational value of mental health simulation in undergraduate bachelor of nursing education: an integrative review. Clin Simul Nurs 2023 Nov;84:101459. [doi: 10.1016/j.ecns.2023.101459]
8. Furuheim ACH, Lindenskov L. Virtual reality simulation as a learning activity for nursing students in mental health practice. Nurse Educ Pract 2025 Feb;83:104272. [doi: 10.1016/j.nepr.2025.104272] [Medline: 39892253]
9. Short PR, Pope WS, Wang CH, Renfroe K, Wright L, Chidume T. Enhancing mental health nursing education through virtual reality simulation: a focus on bipolar disorder. Arch Psychiatr Nurs 2025 Oct;58:151967. [doi: 10.1016/j.apnu.2025.151967] [Medline: 40921524]
10. Cho MK, Kim MY. The effect of virtual reality simulation on nursing students' communication skills: a systematic review and meta-analysis. Front Psychiatry 2024 Jul;15:1351123. [doi: 10.3389/fpsyt.2024.1351123] [Medline: 39035600]
11. Liaw SY, Tan JZ, Lim S, et al. Artificial intelligence in virtual reality simulation for interprofessional communication training: mixed method study. Nurse Educ Today 2023 Mar;122:105718. [doi: 10.1016/j.nedt.2023.105718] [Medline: 36669304]
12. Sengul T, Sarıköse S. Enhancing learning outcomes through AI-driven simulation in nursing education: a systematic review. Clin Simul Nurs 2025 Sep;106:101797. [doi: 10.1016/j.ecns.2025.101797]
13. Sood SK, Rawat KS, Kumar D. A visual review of artificial intelligence and Industry 4.0 in healthcare. Comput Electr Eng 2022 Jul;101:107948. [doi: 10.1016/j.compeleceng.2022.107948] [Medline: 35495094]
14. Sood SK, Rawat KS, Sharma G. Role of enabling technologies in soft tissue engineering: a systematic literature review. IEEE Eng Manag Rev 2022;50(4):155-169. [doi: 10.1109/EMR.2022.3195923]
15. Anarado AN, Agu GU, Nwonu EI. Factors hindering clinical training of students in selected nursing educational institutions in Southeastern Nigeria. Nurse Educ Today 2016 May;40:140-145. [doi: 10.1016/j.nedt.2016.02.022] [Medline: 27125164]
16. Ama Amoo S, Innocentia Ebu Enyan N. Clinical learning experiences of nursing and midwifery students; a descriptive cross-sectional study. Int J Afr Nurs Sci 2022;17:100457. [doi: 10.1016/j.ijans.2022.100457]
17. Khoshbakht-Pishkhani M, Javadi-Pashaki N, Esfandi NA, Koodakani MB, Maroufizadeh S, Madani AH. The effect of educational application in nursing internship clinical training on cognitive and functional skills and students' satisfaction. BMC Nurs 2024 Jun 5;23(1):381. [doi: 10.1186/s12912-024-01954-5] [Medline: 38840192]
18. Jasemi M, Whitehead B, Habibzadeh H, Zabihi RE, Rezaie SA. Challenges in the clinical education of the nursing profession in Iran: a qualitative study. Nurse Educ Today 2018 Aug;67:21-26. [doi: 10.1016/j.nedt.2018.04.010] [Medline: 29727826]
19. Cutts M, Hitchcock KH. Mental health nursing students and generic proficiencies: how educators can support assessment. Br J Nurs 2024 Nov 7;33(20):956-962. [doi: 10.12968/bjon.2024.0117] [Medline: 39506228]
20. Nair M, Moss N, Bashir A, et al. Mental health trends among medical students. Proc (Bayl Univ Med Cent) 2023 Mar 16;36(3):408-410. [doi: 10.1080/08998280.2023.2187207] [Medline: 37091765]
21. Sheykhmohammadi N, Zampunieris D, Stammet P, Grévisse C, Reis S. Developing automated software for improving feedback analysis and self-debriefing in high-fidelity medical simulation. Presented at: 2024 IEEE 18th International Conference on Application of Information and Communication Technologies (AICT); Sep 25-27, 2024. [doi: 10.1109/AICT61888.2024.10740406]
22. Spence AD, Derbyshire S, Walsh IK, Murray JM. Does video feedback analysis improve CPR performance in phase 5 medical students? BMC Med Educ 2016 Aug 12;16(1):203. [doi: 10.1186/s12909-016-0726-x] [Medline: 27519273]
23. Vogelsang L, Wright S, Risling T, et al. Exploring the use of immersive virtual reality in nursing education: a scoping review. Clin Simul Nurs 2024 Dec;97:101648. [doi: 10.1016/j.ecns.2024.101648]
24. Yeung JWY, Ho KHM, Cheung J, Tsang JTY, Chong DKY. Artificial intelligence-based technology in communication training in nursing education: a scoping review. J Prof Nurs 2025;59:40-50. [doi: 10.1016/j.profnurs.2025.04.011] [Medline: 40659433]
25. Henshall C, Davey Z, Merriman C, et al. The impact of the COVID-19 pandemic on nursing students' navigation of their nursing programmes and experiences of resilience. A qualitative study. Int J Nurs Stud Adv 2023 Jul 28;5:100146. [doi: 10.1016/j.ijnsa.2023.100146] [Medline: 38746594]
26. Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition, Text Revision (DSM-5-TR): American Psychiatric Association; 2022. URL: https://doi/book/10.1176/appi.books.9780890425787 [accessed 2026-02-24]
27. Sznajde K, Naughton D, Kar A, et al. Fostering dialogues in global health education: a graduate and undergraduate approach. In: Winchester M, Knapp C, BeLue R, editors. Global Health Collaboration: Springer; 2018:21-31. [doi: 10.1007/978-3-319-77685-9_3]
28. Chapman HJ, Veras-Estévez BA. Strengthening global health content in health professions curricula. Health Educ J 2020 Nov;79(7):839-850. [doi: 10.1177/0017896920915231]
29. Frache G, Tombras GS, Nistazakis HE, Thompson N. Pedagogical approaches to 21st century learning: a model to prepare learners for 21st century competencies and skills in engineering. Presented at: 2019 IEEE Global Engineering Education Conference (EDUCON); Apr 8-11, 2019. [doi: 10.1109/EDUCON.2019.8725214]
30. University of Sharjah EDGE strategy. University of Sharjah. 2024. URL: https://www.sharjah.ac.ae/-/media/project/uos/sites/uos/discover-uos/strategydocument-en.pdf [accessed 2025-05-27]

XSL•FO
RenderX

Abbreviations

AI: artificial intelligence
***DSM-5-TR*:** *Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition, Text Revision*
MSE: mental status examination
OSCE: objective structured clinical examination
UAE: United Arab Emirates
VR: virtual reality

Edited by A Stone; submitted 28.May.2025; peer-reviewed by T Pummanee, KS Rawat, L Paul; revised version received 11.Jan.2026; accepted 22.Jan.2026; published 12.Mar.2026.

Please cite as:
Ahmad A, Kavumpurath J, Kader R, Altamimi M, El Hajj M, Ahmed FRA, Subu MA, Yafei T, Ali HR, Saleh I, Al-Yateem N
Voices of Innovation: Reflective Report on Integrating Artificial Intelligence–Simulated Mental Health Patient Scenarios Into Undergraduate Nursing Education in the United Arab Emirates
JMIR Med Educ 2026;12:e78161
URL: https://mededu.jmir.org/2026/1/e78161
doi:10.2196/78161

© Amina Ahmad, Janisha Kavumpurath, Raheesa Kader, Muna Altamimi, Monia El Hajj, Fatma Refaat Ahmed Ahmed, Muhammad Arsyad Subu, Taliaa Yafei, Hind Rashed Ali, Idil Saleh, Nabeel Al-Yateem. Originally published in JMIR Medical Education (https://mededu.jmir.org), 12.Mar.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Education, is properly cited. The complete bibliographic information, a link to the original publication on https://mededu.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Dopamine, Distraction, and Disruption: Perspectives on How Technology and Generation Z Are Reshaping Medical Education

Bruno B Andrade[1,2], MD, PhD; Mariana Araújo Pereira[1,2], PhD; Katia Avena[1,2], PhD

[1]Laboratório de Pesquisa Clínica e Translacional, Fundação Oswaldo Cruz, Rua Waldemar Falcão, Salvador, Brazil

[2]Instituto de Pesquisa Clínica e Translacional (IPCT), Medicina Zarns, Salvador, Brazil

Corresponding Author:
Bruno B Andrade, MD, PhD
Laboratório de Pesquisa Clínica e Translacional, Fundação Oswaldo Cruz, Rua Waldemar Falcão, Salvador, Brazil

Abstract

This viewpoint reflects on how Generation Z (born between 1995 and 2009), shaped by constant digital engagement, a growing awareness of mental health, and a dopamine-driven environment, is transforming medical education and practice. We explore, from a reflective and interdisciplinary perspective, how the defining characteristics of Generation Z, such as their familiarity with technology, demand for emotional safety, and resistance to traditional hierarchies, might reshape the ways we teach, learn, and practice medicine. Drawing on neuroscience, psychology, sociology, and the medical education literature, this viewpoint emphasizes the need to move beyond knowledge transmission and foster self-regulation, critical thinking, and ethical judgment. We call for a deliberate and compassionate adaptation of medical education to cultivate the skills required for a profession increasingly practiced in a context of overstimulation and complexity.

(JMIR Med Educ 2026;12:e80791) doi:10.2196/80791

KEYWORDS

medical education; undergraduate medical education; generational differences; generation divide; digital natives; dopamine; Generation Z

Introduction: From Generational Transition to Digital Learning Design

Background

Medicine has always been a mirror held up to society. Its rhythms, hierarchies, and even its blind spots are shaped by the values, anxieties, and aspirations of those who practice it [1]. Each generation of physicians leaves behind discoveries and treatments, in addition to cultural imprints, unique ways of thinking, diagnosing, and caring. The Hippocratic humility of the ancients, the anatomical fervor of the Renaissance, the clinical detachment of the 20th century, and the evidence-based revolution of the 2000s each emerged from broader generational shifts in how the world is understood.

These generational traits shape medical practice and interact with the challenges of their time. Social, economic, and technological transformations simultaneously affect both the epidemiological profile of populations and how professionals position themselves in relation to care [2]. How a generation understands the body, disease, and the physician-patient relationship is closely linked to the historical conditions in which it operates [3]. Thus, understanding generational tensions and the contexts that produce them is essential for interpreting the current dilemmas faced by educational institutions and health systems.

We are currently living through rapid demographic and epidemiological transitions [4] that demand constant adaptation from educational institutions. However, beyond these challenges, a new factor deserves attention: the arrival of digital natives in medical training. Raised in environments of high digital exposure, frequent stimulation, digital saturation, and dopamine overload, these students, members of Generation Z (Gen Z), may significantly influence both medical education and practice, bringing potential benefits as well as challenges [5,6]. This process has implications that go beyond classrooms and clinics, with possible repercussions for widely recognized professional values in medicine, such as empathy, ethics, and self-control.

This historical context and the challenges of our time underline the need to reflect on the impact new generations may have on how medicine is taught, learned, and practiced. The entry of Gen Z (born between 1995 and 2009) [6-8] into medical education is not merely a generational renewal; it is a cultural, cognitive, and technological transition that challenges traditional models of training and professional conduct [9].

Understanding this transition is crucial for anticipating trends, adjusting curricula, and rethinking the core values of the medical profession, including the need to redefine the competencies expected of future physicians in light of emerging technologies and societal changes [10]. Against this backdrop, this critical viewpoint explores how the characteristics of Gen Z may influence medical education and practice, both in promising

XSL•FO
RenderX

and challenging ways. We take digital learning design as our guiding thread and discuss how the neurocognitive and sociocultural characteristics of Gen Z translate into specific demands for medical education.

This viewpoint does not aim to present quantitative data or causal generalizations but rather to reflect on emerging trends in medical education, considering social, cultural, and technological transformations. To this end, it adopts an interpretive approach based on a narrative review of recent literature, published institutional reports, and teaching experiences in different academic contexts. Observations on aspects such as punctuality, emotional regulation, and resilience are used as illustrative examples of formative tensions rather than as population-level estimates. Whenever possible, these interpretations are anchored in peer-reviewed studies; when anecdotal evidence or local reports are used, their limited scope is made explicit. The purpose, therefore, is to generate hypotheses and offer pedagogical directions for future inquiry, situating the debate within a reflective and interdisciplinary perspective.

Who Is Gen Z? Between Constant Stimulation and Digital Learning Design

According to generational cohort theory, individuals can be grouped by similar behaviors, values, and perceptions shaped by shared life experiences and historical events. Today, 6 living generations are typically recognized: the Silent Generation (or "traditionalists"; born between 1925 and 1945), baby boomers (1946 and 1964), Generation X (1965 and 1979), Generation Y (or millennials; 1980 and 1994), Gen Z (1995 and 2009), and Generation Alpha (2010 and 2024) [6-8,11].

Gen Z now represents a significant portion of the global population, accounting for around 22% of people worldwide and already 27% of the global workforce [11]. This generation grew up amid profound digital, economic, and social transformations [10]. Projections suggest that over their lifetime, they will hold an average of 18 jobs, switch between 6 different careers, and live in approximately 15 different residences, pointing to a reality defined by mobility, adaptability, and multiple reinventions in both personal and professional life [11]. These dynamics reinforce the importance of understanding their motivations, expectations, and challenges, which can particularly impact them within health education and training.

Much has been written about the distinctiveness of Gen Z. As the first true digital natives [8,12], they have not simply acquired technology; they were born into its omnipresence. While expanded connectivity brings clear advantages, growing up in a hyperconnected environment has brought its own challenges. Whereas previous generations fought for access to knowledge, mentors, and opportunities, Gen Z's struggle is for focus, stillness, and meaning.

Dopamine Nation [13] gives voice to this shift. In a world engineered for the overconsumption of pleasure, the brain's reward system becomes dysregulated. Pleasure leads to pain. Novelty dulls attention. Meaning drowns in noise. This is not simply a sociological curiosity; it is a neurobiological condition [13,14]. In medical practice, such overstimulation can impair core skills like active listening, progressive clinical reasoning, and tolerance for uncertainty, all fundamental to ethical and effective care [15].

At first glance, the traits of this generation appear promising. Gen Z is radically open about mental health, deeply collaborative, socially engaged, and skeptical of inherited hierarchies [14]. They seek to heal, but on their own terms. They reject the stoic pride that previous generations placed on sleep deprivation, emotional suppression, and boundless sacrifice. They speak, with surprising fluency, the language of burnout, boundaries, and balance [16]. This translates into an explicit search for training environments that value emotional safety, horizontality in relationships, and integration between personal and professional life, especially in the face of rapid digital transformation and intergenerational gaps in the use and understanding of technology [11,13,14]. These tensions are not mere contradictions but dispositions that coexist within the same habitus, reflecting the social and institutional pressures that shape the formative experience [17].

Considering this profile, medical education is already feeling the pressure. The old passive learning model—lecture-heavy, memorization-focused, and obedience-based [18,19]—has proven to be less aligned with current preferences. Evidence suggests that Gen Z students tend to prefer more interactive and flexible methods. This generation seeks connection, not coercion [12,20]. This reorientation of expectations resonates with current proposals for medical education reform, which advocate for personalized learning pathways, technological fluency, and humanistic training to prepare students for dynamic health care systems [10].

However, such a shift in expectations also brings with it inherent tensions, particularly when the culture of immediacy collides with the slow, effortful nature of learning, especially within the clinical training required for medical mastery. A generation conditioned by instant feedback and low friction may resist the slow, iterative, and often painful path to mastery [21]. Medical training involves prolonged processes of competence acquisition and, as a rule, does not provide instant gratification. The depth of clinical judgment, the nuance of human physiology, and the texture of uncertainty do not reveal themselves to the impatient or easily bored [22].

These characteristics, particularly strong digital exposure, the need for predictability, and the valuing of rapid returns, directly impact how digital learning experiences should be organized. In practice, this means offering short and clearly delimited units, making objectives explicit, signaling content complexity, providing support to manage difficulties, and ensuring formative feedback that preserves conceptual depth without fostering dependence on immediate answers.

Redefining the Clinic: Human Connection and Digital Overidentification

We recognize that generational categories are useful as a starting point, but they also have limitations when taken too rigidly. For this reason, we combine the cohort perspective with contributions that broaden the understanding of how social and

XSL•FO
RenderX

historical contexts shape educational trajectories. The notion of generational consciousness and distinct groups within the same generation helps situate collective experiences in their time [23].

The concept of habitus, associated with forms of capital and social fields, highlights how dispositions and expectations vary according to cultural and institutional conditions [24]. The idea of the risk society further illustrates how global uncertainties and pressures influence professional choices and educational patterns [25]. Similarly, the humanities and critical social sciences call attention to the limits of approaches centered exclusively on biology. Analyses of neurochemical identities [26], the notion of neuroplasticity as also social [27], and the role of the brain in contemporary definitions of identity [28] reinforce the importance of integrating sociocultural dimensions into neurobiological explanations. Thus, we use both generational and neuroscientific perspectives as complementary lenses, always in dialogue with the social, cultural, and institutional contexts that shape medical education.

These evolving expectations around training and wellness inevitably spill over into clinical environments, where differing generational views on professionalism begin to reshape physician-patient and teacher-student relationships [3,19,21]. Age can influence whether an individual considers a particular behavior professional or unprofessional in clinical settings [19]. The values and traits of each generation may shape how they define and perceive medical professionalism [19,21].

Within this context, the clinic will also shift under the influence of this new generation. Gen Z physicians may stand out in their ability to forge human connections, especially with patients whose struggles mirror their own: anxiety, substance use, technology-related burnout, and identity crises [22]. They may redefine the physician-patient relationship by flattening hierarchies, enhancing transparency, and more seamlessly integrating mental and physical care.

However, this approach risks collapsing professional distance, blurring the line between empathy and overidentification. There is virtue in vulnerability, but medicine must also remain a space where clarity, composure, and decisiveness prevail, especially in times of crisis [29,30]. Beyond the emphasis on individual virtue, it is useful to recognize empathy as part of the "emotional labor" institutionally required [31] and as a repertoire that can be curricularized, measured, and even commodified in certain arrangements [31]. Feminist perspectives on care remind us that empathic practices are shaped by relations of power, the distribution of responsibilities, and the material conditions of work [32]. This implies that educating for empathy requires attention to the structures that sustain it, rather than treating it solely as a psychological attribute of the individual.

Striking a balance between empathy and firmness is particularly challenging in high-pressure contexts, such as emergency situations or difficult conversations, where the physician's emotional stability is part of the care process [33,34]. These recommendations resonate with the field of Medical Humanities, which, for decades, has promoted practices of narrative reflection, acceptance of failure, and tolerance of ambiguity as foundations of professionalism and ethical care [35]. These challenges are not limited to patient interactions; they arise before and during the training of physicians themselves. Within classrooms and learning environments, students' behavior reflects deep tensions between contemporary expectations and the traditional ethos of medicine [24].

Institutional reports describe changes in patterns of punctuality, attendance, and responsiveness to feedback, including warnings about behavior considered unprofessional conduct, low tolerance for criticism, and a growing need to address issues of "etiquette" in clinical and educational settings, ranging from dress codes to cellphone use during clinical rounds [21,30].

Among many Gen Z medical students, failure is not merely avoided; it is seen as unacceptable. Anything short of perfection is perceived as invalid. The demand for immediate results and flawless performance reflects habits shaped by short-form content and frictionless interfaces [36]. If a learning experience is not fast, intuitive, or immediately rewarding, it risks being dismissed altogether. By repositioning "failure" as a pedagogical component, this approach aligns with the "medical humanities" literature, which understands it as a resource for learning, professional identity, and patient safety. However, this outlook undermines the iterative, uncertain, and often uncomfortable process required for meaningful learning, not only in medicine but in any complex skill.

In an age of accelerated media and superficial metrics of influence, the appearance of knowledge can overshadow the substance of learning. Popularity, reach, and digital fluency may take precedence over reflective thinking and epistemic humility. These episodes reveal a disconnect between Gen Z's aspirations for autonomy and well-being and the historical values, norms of professional decorum, and aphorisms that have shaped medical practice for millennia [21,30,37].

Among the most critical consequences of this disconnect is the progressive decline in empathy throughout medical training [38-40], a trend that may be intensified by the psychological distress and emotional exhaustion experienced by many Gen Z students [11,40]. Preserving empathy requires more than idealism; it demands intentionality [41]. Empathy can be cultivated through curricular interventions that promote deep listening, perspective-taking, and the strengthening of interpersonal communication—skills directly linked to clinical performance and patient satisfaction [21,39,41]. Given the risk of empathic decline, especially in emotionally and institutionally stressful contexts common to Gen Z, medical schools must foster psychologically supportive environments and pedagogical practices grounded in structured reflection [40]. Cultivating the affective domain, therefore, requires more than attitudinal content; it calls for safe spaces that enable emotional insight, interpersonal development, and the reinforcement of ethical, compassionate care [11,21].

In medical training, the cognitive and affective domains move in tandem. How a student feels about the world profoundly shapes how they interpret and act within it. The construction of clinical reasoning depends on logic and on the emotional regulation that sustains listening, attention, and discernment amid uncertainty [34]. Dopamine dysregulation is not limited to the patient population; it may also shape physicians

XSL•FO
RenderX

themselves. In a profession that demands presence, endurance, and long-term thinking, constant digital stimuli and fragmented attention can erode clinical focus [42].

Although Gen Z is highly familiar with digital technologies and multitasking, these qualities present substantial challenges in medical education [13,41]. The temptations to turn to phones, multitask, or escape discomfort are habits that may compromise patient safety by increasing the risk of missed clinical signs, rushed interpretations, and shallow thinking. Many students in this generation struggle to distinguish fact from opinion [43], weakening their ability to develop critical thinking and undermining clinical reasoning [13,43]. Gen Z is marked by paradox: idealistic yet anxious, collaborative yet individualistic, engaged yet frequently overwhelmed [13]. They often feel lost among an overload of data and information and require active guidance to identify reliable sources and interpret reality more reflectively [13,43].

Medical education must begin to address this not as a moral failing, but as a neurological and cultural condition. Training must extend beyond foundational biomedical knowledge to include practices that promote self-regulation, attention hygiene, and recovery of cognitive depth [21]. This approach is not a luxury; it is now a clinical imperative. This involves, for example, incorporating practices such as mindfulness, self-care strategies, and digital education into the curriculum, preparing students to recognize and manage their own distraction triggers [44]. Embedding structured reflection and self-assessment within digital learning paths can also mitigate cognitive biases and strengthen essential diagnostic skills, such as hypothesis generation and decision-making under uncertainty [43].

The same factors that influence empathy, communication, and professionalism in face-to-face clinical settings also emerge in digital learning environments, whether in the conduct of simulated consultations, the appropriate use of mobile devices, or the handling of sensitive content in online modules. For this reason, e-learning and simulation experiences need to include clear guidance on professional conduct, opportunities for emotional preparation, and spaces for structured reflection.

Teaching Medicine in the Age of Overstimulation: Technology and Design

Overview

Medical education has undergone several significant transformations over time. The first of these occurred in the early 20th century, when a robust, science-based curriculum was created that helped establish the medical profession [18,19]. Then, in the second half of the 20th century, innovations such as problem-based learning emerged, aiming to better connect theoretical knowledge to clinical practice [18]. Nowadays, we are experiencing a third major transformation, driven both by a new generation of university students and the ever-evolving needs of health systems.

This reform should not be limited to expanding content or adopting new teaching methodologies. Instead, it requires a comprehensive shift in the definition of core competencies, integrating interdisciplinary knowledge, leadership development, and digital literacy from the undergraduate level [10]. It must focus on developing critical, flexible leaders who are capable of innovating and addressing the complex challenges of modern health care [6]. In line with medical humanities, curricula that normalize formative error, cultivate reflection, and train tolerance for ambiguity tend to sustain attention, empathy, and clinical judgment in contexts of high complexity.

Addressing the challenges of Gen Z requires embedding digital innovation into curricular design. Adaptive learning platforms and learning analytics can personalize content delivery, helping students cope with cognitive overload by tailoring feedback and pacing to individual progress [36,43,45]. Digital portfolios and e-assessment systems, increasingly integrated into competency-based curricula, encourage structured reflection and longitudinal tracking of competencies, reinforcing professional identity development [46]. Simulation-based environments, ranging from high-fidelity clinical scenarios to virtual and augmented reality, provide safe spaces for repeated practice, error tolerance, and the cultivation of resilience [4,45,47], and can strengthen clinical reasoning when coupled with structured debriefings, though they require consistent pedagogical integration and face barriers of cost and adoption [48]. Furthermore, incorporating design principles such as interactivity, multimodality, and user-centered navigation ensures that digital tools meet Gen Z's expectations for immediacy and visual engagement [14,36]. When combined with mentorship and psychological support systems, these approaches position technology not as a source of distraction but as a vehicle for meaningful learning and professional growth [49,50].

Technology-mediated learning environments offer concrete opportunities to translate the cultural and psychological traits of Gen Z into educational practice. E-learning platforms and adaptive online modules can provide self-paced learning and continuous formative feedback, helping students manage cognitive overload through tailored content and individualized pacing [43,45]. Virtual patient encounters, increasingly embedded in digital curricula, foster clinical reasoning and diagnostic decision-making in interactive and safe contexts that mirror real-world complexity [51,52]. Simulation-based training, whether through high-fidelity scenarios or virtual and augmented reality, allows for repeated practice, error tolerance, and the cultivation of resilience without compromising patient safety [45]. Artificial intelligence (AI)-assisted learning tools, including intelligent tutoring systems and automated feedback algorithms, further expand the potential for personalized learning pathways and real-time adaptive feedback, supporting self-regulated learning while requiring safeguards against over-reliance [5,14,36,53]. Together, these innovations exemplify how educational design can align with the expectations of Gen Z learners while preparing them for the complexities of modern health care.

In this context, it is crucial to understand who these students are. Today, most medical school entrants worldwide belong to Gen Z, with the majority being 24 years or younger [54,55]. With their digital fluency and strong sense of social urgency, they can act as catalysts for this change. Their emotional

XSL•FO
RenderX

intelligence, ethical sensitivity, and resistance to traditional hierarchies make them ideal candidates for a more transformative type of education—one that goes beyond the technical to include leadership, adaptability, and a new way of caring that considers local contexts and leverages global resources [6,13,21,56].

Gen Z prefers quiet learning environments with clearly defined objectives, opportunities for small group mentoring, and greater integration of technology into classrooms [11]. They also demonstrate increased sensitivity to emotionally charged content, including topics related to abuse or death [11,57]. This has driven a growing demand for the use of content warnings [57] and the creation of safe spaces within medical curricula, challenging the long-standing tradition of exposing students to difficult topics without prior emotional preparation [11,57].

In addition, Gen Z's engagement with health is strongly shaped by technology and social media. They tend to prefer visual content, such as short videos, over traditional texts [13,14,46]. Their expectations emphasize speed, convenience, and the availability of digital interactions [56]. This generation also favors approaches that integrate emotional and physical well-being [14,21].

Another noteworthy shift is their evolving perception of volunteerism. Gen Z tends to reject traditional charity-based models and instead prefers technological and entrepreneurial solutions to social problems [11]. This may reduce participation in conventional student-run clinics, but it creates space for more innovative social impact initiatives to be integrated into the curriculum.

Among the interventions we consider most promising for addressing the identified challenges, three stand out as high priorities: (1) the use of simulation environments with structured debriefing, capable of strengthening clinical reasoning and patient safety [58]; (2) the use of digital portfolios and longitudinal formative feedback, which support the development of professional identity and student self-regulation; and (3) the adoption of adaptive platforms and learning analytics, which help mitigate cognitive overload by adjusting content and pacing to individual needs [36]. Complementarily, practices of attentional hygiene, such as mindfulness and digital literacy, as well as explicit guidelines for professionalism in virtual settings, provide support for the development of self-regulation and the preservation of ethical care. Taken together, these strategies outline a practical path to align technological innovation, cognitive depth, and humanistic formation, responding in an integrated way to the specific demands of Gen Z in medical education.

Beyond generational dispositions, these dynamics are co-produced by institutional choices. Curricular structures, the organization of formative work, and, above all, assessment regimes shape what students prioritize [59]. Systems based on superficial metrics tend to reinforce immediacy and perfectionism, whereas programmatic assessment and entrustable professional activities (EPAs) [60], with frequent low-stakes feedback cycles, promote formative learning and tolerance for error [42]. In addition, the hidden curriculum and organizational culture communicate norms and expectations that can amplify or mitigate these tensions [61], and the quality of feedback cultures depends on institutional design and supervisory practices. Finally, organizational factors are implicated in student well-being and burnout, indicating that interventions must combine individual support with structural change.

Competency Development in Digital Ecosystems

Medical schools of the future will need to move beyond simply integrating science and care or adopting competency-based curricula [35]. There is a growing demand for individualized learning processes and immersive clinical experiences that prepare students to work in increasingly complex health care systems [17].

Faced with Gen Z's presence in academic settings, schools must develop new technologies and tools to manage the learning environment, as well as new assessment methods to measure competency acquisition [17]. Many institutions are already responding to these challenges by offering immersive experiences within health systems, allowing students to engage with patient-centered care and collaborative practice [35]. Schools are also implementing educational trajectories based on EPAs [60], using portfolios for guided self-assessment and personalized goal-setting based on performance data [5,62].

Moreover, EPAs can be understood not only as assessment instruments but also as structural interventions capable of mitigating some of the challenges described throughout this article. By clearly defining performance expectations in progressively entrusted clinical activities, EPAs reduce the anxiety related to ambiguity and fear of failure, which are common among Gen Z students [42]. The model of supervised progression, based on levels of trust, provides continuous and calibrated feedback, reconciling this generation's need for rapid returns with the preservation of formative depth [63]. In addition, the longitudinal nature of EPAs supports the integration of cognitive, affective, and professional competencies, promoting both clinical reasoning and the development of professional identity within digital learning ecosystems [64].

These transformations must also be understood in dialogue with established debates in medical education. The literature on professional identity formation shows that becoming a physician is a gradual process of socialization that involves both ideals and contradictions experienced in daily practice [65,66]. At the same time, some of these tensions are expressed in the so-called hidden curriculum, a set of implicit norms and messages that often reinforce hierarchies and practices that diverge from the formal curriculum [61]. Finally, our argument also aligns with the sociology of medical knowledge, which recognizes medicine as a cultural and historical construction, shaped by regimes of truth and forms of legitimizing knowledge [67].

We are living in a time of unprecedented informational access and, paradoxically, growing uncertainty about what truly matters. For Gen Z, navigating this overload requires more than search skills; it requires the ability to filter, discern, and transform data into clinical wisdom. Medical education must rise to this challenge, fostering competencies that go beyond technical knowledge to embrace critical thinking, adaptability,

XSL•FO
RenderX

and ethical judgment [4,12,15,17,18,22,68]. This aligns with the emerging vision of a medical student profile capable of navigating AI-based systems, leading innovations, and promoting global health [20].

We would be wrong to romanticize previous generations. The stoicism of the past often came at the cost of silence, loneliness, and unprocessed trauma. However, we must also avoid the pendulum swing toward fragility. Medicine is hard; it asks of us the ability to delay pleasure, hold ambiguity, and face suffering without paralysis. The literature on dopaminergic dysregulation suggests possible difficulties in these tasks [69]. Nevertheless, the solution is not nostalgia for a lost generation; it is the deliberate cultivation of new skills for a new world.

Cultivating New Skills for a New World

As this viewpoint has argued, the arrival of Gen Z signals profound changes in medical education, requiring adaptations in curricula, infrastructure, psychological support systems, and pedagogical approaches to meet their unique characteristics.

Perhaps Gen Z, with its hunger for authenticity, can lead us there if we, the educators and leaders, are willing to challenge them gently but firmly. It is recommended to model attention instead of distraction, to provide structured mentorship rather than demand performance, and to promote intellectual discipline alongside emotional openness to dialogue. If we dare to teach them how to be bored again, how to wait, and how to listen deeply—not just to patients, but to themselves.

A generation shaped by constant digital stimulation is entering the most human of professions. The impact of this process will be significant and will depend on how curricula and learning environments are adapted to respond to these generational characteristics. Among the possibilities are educational design strategies that support structured and responsible learning: adaptive modules that adjust the pace of study to individual needs; dashboards that allow students to monitor their progress; electronic portfolios that document acquired competencies; simulation environments, including virtual or augmented reality, that provide safe and repeated practice; AI tools designed to guide learning in a calibrated manner [5,36]; and institutional policies for digital well-being, such as device-free periods or reduced notifications, which help protect attention and reduce cognitive overload.

By integrating evidence on neurocognition and Gen Z preferences with principles of digital design, we propose a framework that combines personalization with depth, efficiency with reflection, and technology with professionalism. The effectiveness of these strategies should be tracked through learning metrics (retention, transfer, and diagnostic reasoning), well-being indicators, and measures of professionalism in digital environments.

Future Directions: Specific and Testable Questions

Overview

Building on the interpretive lens of this viewpoint, future research should examine how the generational characteristics of Gen Z interact with curricular structures, assessment systems, and digital learning environments. To advance beyond reflection and provide an empirical foundation for innovation in medical education, we propose the following specific, testable questions:

- Does attentional self-regulation training improve learning and clinical reasoning in Gen Z learners?
- Do adaptive e-learning modules reduce perceived cognitive load for Gen Z students compared to previous cohorts?
- Can structured reflection mitigate the longitudinal decline in empathy or professionalism among Gen Z learners?
- Does AI-assisted tutoring improve diagnostic calibration in Gen Z learners without fostering over-reliance?
- Do portfolios with frequent formative feedback strengthen professional identity and self-regulation among Gen Z learners?
- Do coaching and structured mentoring programs reduce burnout and attrition among Gen Z medical students?

Taken together, these questions outline an empirical agenda that moves beyond broad generalizations about generational traits. They encourage rigorous testing of how technology-mediated interventions, assessment innovations, and institutional supports can shape the learning experiences of Gen Z in medical education, while also offering transferable insights for future cohorts.

Conclusions

Gen Z is a generation marked by familiarity with technology, a strong emphasis on emotional safety, and lower adherence to traditional hierarchies. Their presence in medical education calls for deliberate adaptations in curricula, student support, and the design of digital learning experiences. By aligning technological interventions with pedagogical principles and consistent assessment strategies, it is possible to combine personalization with depth, fostering sustained attention, critical thinking, and ethical judgment in contexts of high stimulation and complexity.

Funding

The viewpoint was supported by an intramural research program of the Oswaldo Cruz Foundation (Fiocruz), Brazil. BBA is a senior scientist from the Brazilian National Council for Scientific and Technological Development (CNPq 1D). KA is a postdoctoral researcher at Gonçalo Moniz Institute, Oswaldo Cruz Foundation. The funding source played no role in this viewpoint, data collection, analysis or interpretation, or the writing of the manuscript. The researchers retained complete independence in the conduct of this viewpoint.

XSL•FO
RenderX

Authors' Contributions

Conceptualization: BBA, KA
Data curation: BBA
Funding acquisition: BBA
Methodology: BBA, KA
Project administration: BBA
Supervision: BBA
Writing – original draft: BBA, KA
Writing – review & editing: BBA, MAP, KA

Conflicts of Interest

None declared.

References

1. Wright SM, Kern DE, Kolodner K, Howard DM, Brancati FL. Attributes of excellent attending-physician role models. N Engl J Med 1998 Dec 31;339(27):1986-1993. [doi: 10.1056/NEJM199812313392706]
2. Rosa RS, Souza DAD, Silva de Oliveira Boery RN, et al. (Re) thinking the professional training in health before the epidemiological transition [Article in Portuguese]. Revista Salud Pública 2022;24(6):1-9. [doi: 10.15446/rsap.v24n6.75823]
3. Cushing A. History of the doctor–patient relationship. In: Clinical Communication in Medicine: John Wiley & Sons, Ltd; 2015:6-20. [doi: 10.1002/9781118728130]
4. Yeung AWK, Parvanov ED, Hribersek M, et al. Digital teaching in medical education: scientific literature landscape review. JMIR Med Educ 2022 Feb 9;8(1):e32747. [doi: 10.2196/32747] [Medline: 35138260]
5. Nykiel-Bailey S, Burrows K, Szafarowicz BE, Moquin R. Faculty perceptions on the roles of mentoring, advising, and coaching in an anesthesiology residency program: mixed methods study. JMIR Med Educ 2025 Jan 21;11:e60255. [doi: 10.2196/60255] [Medline: 39864951]
6. Frenk J, Chen L, Bhutta ZA, et al. Health professionals for a new century: transforming education to strengthen health systems in an interdependent world. The Lancet 2010 Dec;376(9756):1923-1958. [doi: 10.1016/S0140-6736(10)61854-5]
7. Twenge JM, Joiner TE, Rogers ML, Martin GN. Increases in depressive symptoms, suicide-related outcomes, and suicide rates among U.S. adolescents after 2010 and links to increased new media screen time. Clin Psychol Sci 2018 Jan;6(1):3-17. [doi: 10.1177/2167702617723376]
8. Dimock M. Defining generations: where Millennials end and Generation Z begins. Pew Research Center. 2019 Jan 17. URL: https://www.pewresearch.org/short-reads/2019/01/17/where-millennials-end-and-generation-z-begins/ [accessed 2025-04-02]
9. Defining our six generations. Pew. 2019 Feb 8. URL: https://www.pew.org/en/research-and-analysis/data-visualizations/2019/defining-our-six-generations [accessed 2025-07-02]
10. Miguez-Pinto JP, Garcia-Rosa B, Maggitti-Bezerril M, et al. The medical student of the future: redefining competencies in a transformative era. Front Med (Lausanne) 2025;12:1593685. [doi: 10.3389/fmed.2025.1593685] [Medline: 40589962]
11. Plochocki JH. Several ways Generation Z may shape the medical school landscape. J Med Educ Curric Dev 2019;6:2382120519884325. [doi: 10.1177/2382120519884325] [Medline: 31701014]
12. Gen Z and Gen Alpha infographic update. McCrindle. URL: https://mccrindle.com.au/article/topic/generation-z/gen-z-and-gen-alpha-infographic-update/ [accessed 2025-07-02]
13. Mohr KAJ, Mohr ES. Understanding Generation Z students to promote a contemporary learning environment. J Empow Teach Excel 2017;1(1):8494. [doi: 10.15142/T3M05T]
14. Cecconi C, Adams R, Cardone A, et al. Generational differences in healthcare: the role of technology in the path forward. Front Public Health 2025;13:1546317. [doi: 10.3389/fpubh.2025.1546317] [Medline: 40078753]
15. Lembke A. Dopamine Nation: Finding Balance in the Age of Indulgence: Dutton; 2021.
16. Rotenstein LS, Torre M, Ramos MA, et al. Prevalence of burnout among physicians: a systematic review. JAMA 2018 Sep 18;320(11):1131-1150. [doi: 10.1001/jama.2018.12777] [Medline: 30326495]
17. Bourdieu P. The Logic of Practice: Stanford University Press; 1990.
18. Impact of the Flexner's report on health professional education. World J Adv Res Rev 2023;19(3):939-941. [doi: 10.30574/wjarr.2023.19.3.1888]
19. Flexner A. Medical education in the United States and Canada. Science 1910 Jul 8;32(810):41-50. [doi: 10.1126/science.32.810.41]
20. McCrindle M, Wolfinger E. The ABC of XYZ: Understanding the Global Generations: University of New South Wales Press; 2009.
21. Eckleberry-Hunt J, Lick D, Hunt R. Is medical education ready for generation Z? J Grad Med Educ 2018 Aug;10(4):378-381. [doi: 10.4300/JGME-D-18-00466.1] [Medline: 30154963]

XSL•FO
RenderX

22. Dyrbye LN, Shanafelt TD. Physician burnout: a potential threat to successful health care reform. JAMA 2011 May 18;305(19):2009-2010. [doi: 10.1001/jama.2011.652] [Medline: 21586718]
23. Mannheim K. Essays on the Sociology of Knowledge: Routledge & Kegan Paul; 1952.
24. Bourdieu P. Distinction: A Social Critique of the Judgement of Taste: Harvard University Press; 1987.
25. Beck U. Risk Society: Towards a New Modernity: Sage Publications Ltd; 1992.
26. Rose N. The Politics of Life Itself: Biomedicine, Power, and Subjectivity in the Twenty-First Century: Princeton University Press; 2007.
27. Catherine M. What Should We Do with Our Brain: Fordham University Press; 2008.
28. Ortega F, Vidal F. Neurocultures: Glimpses into an Expanding Universe: Peter Lang Pub Inc; 2011.
29. Halpern J. What is clinical empathy? J Gen Intern Med 2003 Aug;18(8):670-674. [doi: 10.1046/j.1525-1497.2003.21017.x] [Medline: 12911651]
30. Papadimos TJ. Hippocrates, the first aphorism, and decorum: a personal view of caution and guidance regarding Generation Z students. Med Teach 2025 Feb;47(2):366-367. [doi: 10.1080/0142159X.2024.2331051] [Medline: 38564808]
31. Segal JZ. Health and the Rhetoric of Medicine: Southern Illinois University Press; 2008.
32. Tronto J. Moral Boundaries: A Political Argument for an Ethic of Care: Routledge; 1993.
33. Alam R, Cheraghi-Sohi S, Panagioti M, Esmail A, Campbell S, Panagopoulou E. Managing diagnostic uncertainty in primary care: a systematic critical review. BMC Fam Pract 2017 Aug 7;18(1):79. [doi: 10.1186/s12875-017-0650-0] [Medline: 28784088]
34. Politi MC, Légaré F. Physicians' reactions to uncertainty in the context of shared decision making. Patient Educ Couns 2010 Aug;80(2):155-157. [doi: 10.1016/j.pec.2009.10.030] [Medline: 19948387]
35. Wear D, Zarconi J. Can compassion be taught? Let's ask our students. J Gen Intern Med 2008 Jul;23(7):948-953. [doi: 10.1007/s11606-007-0501-0] [Medline: 18612722]
36. Monzon N, Hays FA. Leveraging generative artificial intelligence to improve motivation and retrieval in higher education learners. JMIR Med Educ 2025 Mar 11;11:e59210. [doi: 10.2196/59210] [Medline: 40068170]
37. Antoniou GA, Antoniou SA, Georgiadis GS, Antoniou AI. A contemporary perspective of the first aphorism of Hippocrates. J Vasc Surg 2012 Sep;56(3):866-868. [doi: 10.1016/j.jvs.2012.05.002] [Medline: 22917047]
38. Dinoff A, Lynch S, Hameed AS, Koestler J, Ferrando SJ, Klepacz L. When did the empathy die?: Examining the correlation between length of medical training and level of empathy. Med Sci Educ 2023 Apr;33(2):489-497. [doi: 10.1007/s40670-023-01768-1] [Medline: 37251206]
39. Neumann M, Edelhäuser F, Tauschel D, et al. Empathy decline and its reasons: a systematic review of studies with medical students and residents. Acad Med 2011 Aug;86(8):996-1009. [doi: 10.1097/ACM.0b013e318221e615] [Medline: 21670661]
40. Dedivitis RA, de Matos LL, de Castro MAF, de Castro AAF, Giaxa RR, Tempski PZ. Empathy and spirituality in medical students and residents [Article in Portuguese]. Rev Bras Educ Med 2024;48(4):e104. [doi: 10.1590/1981-5271v48.4-2023-0042]
41. Hojat M, Gonnella JS, Nasca TJ, Mangione S, Vergare M, Magee M. Physician empathy: definition, components, measurement, and relationship to gender and specialty. Am J Psychiatry 2002 Sep;159(9):1563-1569. [doi: 10.1176/appi.ajp.159.9.1563] [Medline: 12202278]
42. ten Cate O. Nuts and bolts of entrustable professional activities. J Grad Med Educ 2013;5(1):157-158. [doi: 10.4300/JGME-D-12-00380.1]
43. Supper P, Urban D, Acker I, et al. A concept for adapting medical education to the next generations via three-staged digital peer teaching key feature cases. Wien Med Wochenschr 2023 Apr;173(5-6):108-114. [doi: 10.1007/s10354-022-00990-7] [Medline: 36542219]
44. Wilmer HH, Sherman LE, Chein JM. Smartphones and cognition: a review of research exploring the links between mobile technology habits and cognitive functioning. Front Psychol 2017;8:605. [doi: 10.3389/fpsyg.2017.00605] [Medline: 28487665]
45. McGaghie WC, Issenberg SB, Petrusa ER, Scalese RJ. A critical review of simulation-based medical education research: 2003-2009. Med Educ 2010 Jan;44(1):50-63. [doi: 10.1111/j.1365-2923.2009.03547.x] [Medline: 20078756]
46. Vizcaya-Moreno MF, Pérez-Cañaveras RM. Social media used and teaching methods preferred by Generation Z students in the nursing clinical learning environment: a cross-sectional research study. Int J Environ Res Public Health 2020 Nov 9;17(21):8267. [doi: 10.3390/ijerph17218267] [Medline: 33182337]
47. Sousa PD, Gazineu TR, Filho RLL, Avena KDM, Quintanilha LF. Realistic simulation as a teaching strategy in medical undergraduation: a systematic review [Article in Portuguese]. Sci Med (Porto Alegre) 2022;32:1-11. [doi: 10.15448/1980-6108.2022.1.42717]
48. Lie SS, Helle N, Sletteland NV, Vikman MD, Bonsaksen T. Implementation of virtual reality in health professions education: scoping review. JMIR Med Educ 2023 Jan 24;9(1):e41589. [doi: 10.2196/41589] [Medline: 36692934]
49. Quintanilha LF, Avena KDM, Portilho EML, Pereira MA, Nazar AN, Andrade BB. Scientific mentoring in undergraduate medical school: impacts on student satisfaction, engagement and production. Rev Bras Educ Med 2023;47(1):e20. [doi: 10.1590/1981-5271v47.1-20220074.ing]

XSL•FO
RenderX

50. Henry-Noel N, Bishop M, Gwede CK, Petkova E, Szumacher E. Mentorship in medicine and other health professions. J Cancer Educ 2019 Aug;34(4):629-637. [doi: 10.1007/s13187-018-1360-6] [Medline: 29691796]
51. Kononowicz AA, Zary N, Edelbring S, Corral J, Hege I. Virtual patients—What are we talking about? A framework to classify the meanings of the term in healthcare education. BMC Med Educ 2015 Feb 1;15(1):11. [doi: 10.1186/s12909-015-0296-3] [Medline: 25638167]
52. Borg A, Georg C, Jobs B, et al. Virtual patient simulations using social robotics combined with large language models for clinical reasoning training in medical education: mixed methods study. J Med Internet Res 2025 Mar 3;27:e63312. [doi: 10.2196/63312] [Medline: 40053778]
53. Chardonnens S. Adapting educational practices for Generation Z: integrating metacognitive strategies and artificial intelligence. Front Educ 2025;10:1504726. [doi: 10.3389/feduc.2025.1504726]
54. Scheffer M. Medical demography in Brazil 2025 [Article in Portuguese]. : Ministry of Health of Brazil; 2025 URL: https://bvsms.saude.gov.br/bvs/publicacoes/demografia_medica_brasil_2025.pdf [accessed 2025-05-26]
55. Canadian medical education statistics. : Association of Faculties of Medicine of Canada; 2020 URL: https://www.afmc.ca/wp-content/uploads/2022/10/CMES2020-Complete_EN.pdf? [accessed 2026-03-05]
56. Shamoon M. "Future-Proofing" communication skills for future physician learners. J Grad Med Educ 2019 Feb;11(1):106-106. [doi: 10.4300/JGME-D-18-00858.1] [Medline: 30805109]
57. Fisher J, Bennett J, Atkinson A, Errington L. Trigger warnings in medical student education: a scoping review. Clin Teach 2025 Feb;22(1):e13826. [doi: 10.1111/tct.13826] [Medline: 39496276]
58. Liu SJ, Chang YZ, Hwu LJ, Ku MS. Educational videos to reduce parental rejection of pediatric cardiac catheterization during the COVID-19 pandemic. Healthcare (Basel) 2023 May 11;11(10):1395. [doi: 10.3390/healthcare11101395] [Medline: 37239682]
59. van der Vleuten CPM, Schuwirth LWT, Driessen EW, et al. A model for programmatic assessment fit for purpose. Med Teach 2012;34(3):205-214. [doi: 10.3109/0142159X.2012.652239] [Medline: 22364452]
60. Hennus MP, Jarrett JB, Taylor DR, Ten Cate O. Twelve tips to develop entrustable professional activities. Med Teach 2023 Jul;45(7):701-707. [doi: 10.1080/0142159X.2023.2197137] [Medline: 37027517]
61. Lempp H, Seale C. The hidden curriculum in undergraduate medical education: qualitative study of medical students' perceptions of teaching. BMJ 2004 Oct 2;329(7469):770-773. [doi: 10.1136/bmj.329.7469.770] [Medline: 15459051]
62. Skochelak SE, Stack SJ. Creating the medical schools of the future. Acad Med 2017 Jan;92(1):16-19. [doi: 10.1097/ACM.0000000000001160] [Medline: 27008357]
63. Shaughnessy AF, Sparks J, Cohen-Osher M, Goodell KH, Sawin GL, Gravel J. Entrustable professional activities in family medicine. J Grad Med Educ 2013 Mar;5(1):112-118. [doi: 10.4300/JGME-D-12-00034.1] [Medline: 24404237]
64. Englander R, Flynn T, Call S, et al. Toward defining the foundation of the MD degree: core entrustable professional activities for entering residency. Acad Med 2016 Oct;91(10):1352-1358. [doi: 10.1097/ACM.0000000000001204] [Medline: 27097053]
65. Jarvis-Selinger S, Pratt DD, Regehr G. Competency is not enough: integrating identity formation into the medical education discourse. Acad Med 2012 Sep;87(9):1185-1190. [doi: 10.1097/ACM.0b013e3182604968] [Medline: 22836834]
66. Cruess RL, Cruess SR, Boudreau JD, Snell L, Steinert Y. Reframing medical education to support professional identity formation. Acad Med 2014 Nov;89(11):1446-1451. [doi: 10.1097/ACM.0000000000000427] [Medline: 25054423]
67. Bloom SW. Structure and ideology in medical education: an analysis of resistance to change. J Health Soc Behav 1988 Dec;29(4):294-306. [doi: 10.2307/2136864] [Medline: 3253321]
68. Volkow ND, Koob GF, McLellan AT. Neurobiologic advances from the brain disease model of addiction. N Engl J Med 2016 Jan 28;374(4):363-371. [doi: 10.1056/NEJMra1511480] [Medline: 26816013]
69. Grace AA. Dysregulation of the dopamine system in the pathophysiology of schizophrenia and depression. Nat Rev Neurosci 2016 Aug;17(8):524-532. [doi: 10.1038/nrn.2016.57] [Medline: 27256556]

Abbreviations

AI: artificial intelligence
EPA: entrustable professional activity
Z: Generation Z

Edited by T Leung; submitted 16.Jul.2025; peer-reviewed by B Tuohy, S Palo, TN Leavelle; revised version received 29.Sep.2025; accepted 31.Dec.2025; published 18.Mar.2026.

Please cite as:
Andrade BB, Araújo Pereira M, Avena K
Dopamine, Distraction, and Disruption: Perspectives on How Technology and Generation Z Are Reshaping Medical Education
JMIR Med Educ 2026;12:e80791
URL: https://mededu.jmir.org/2026/1/e80791
doi:10.2196/80791

XSL•FO
RenderX

© Bruno B Andrade, Mariana Araújo Pereira, Katia Avena. Originally published in JMIR Medical Education (https://mededu.jmir.org), 18.Mar.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Education, is properly cited. The complete bibliographic information, a link to the original publication on https://mededu.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Digital Choice Architecture in Medical Education: Applying Behavioral Economics to Online Learning Environments

Victoria Ekstrom[1,2], MBBS, MRCP, Msc

[1]Department of Gastroenterology and Hepatology, Singapore General Hospital, 20 College Road, Singapore, Singapore

[2]Medical Humanities Institute, SingHealth Duke-NUS Academic Medical Centre, Singapore, Singapore

Corresponding Author:
Victoria Ekstrom, MBBS, MRCP, Msc
Department of Gastroenterology and Hepatology, Singapore General Hospital, 20 College Road, Singapore, Singapore

Abstract

Health care has widely adopted behavioral economics to influence clinical practice, with documented success using defaults and social comparison feedback in electronic health records. However, online medical education, now the dominant modality for continuing professional development, remains designed on assumptions of rational learning that behavioral science has disproven in clinical contexts. This viewpoint examines the paradox of applying sophisticated behavioral insights to clinical work while designing digital learning environments as if learners are immune to cognitive limitations. We propose digital choice architecture for medical education: intentional integration of behavioral design principles into learning management systems and online platforms. Drawing from clinical nudge units and implementation science, we demonstrate how defaults, social norms, and commitment devices can be systematically applied to digital continuing education. As medical education becomes increasingly technology-mediated, behavioral science provides the theoretical foundation and practical tools for designing online learning environments that align with how clinicians actually make decisions.

(JMIR Med Educ 2026;12:e86497) doi:10.2196/86497

KEYWORDS

medical education; behavioral economics; digital learning; learning management systems; LMS; choice architecture; learning analytics; nudge theory; implementation science

Introduction

Despite a global continuing medical education (CME) market projected to reach US $13.52 billion by 2028, continuing professional development (CPD) demonstrates remarkably limited effectiveness. A Cochrane systematic review of 215 studies with over 28,000 health care professionals found that educational meetings produce only a 6% median improvement in practice compliance and 3% improvement in patient outcomes [1]. Over 80% of CME activities rely on passive learning methods despite evidence showing that didactic presentations have little or no beneficial effect [2,3]. Health care organizations invest billions in educational technology while accepting single-digit effectiveness rates that would be unacceptable in any other clinical intervention.

The failure stems from flawed design assumptions. Current digital CPD platforms operate on an "information deficit model"—the assumption that providing knowledge changes behavior—which behavioral science has comprehensively debunked [2]. New information follows the Ebbinghaus forgetting curve, with 70% forgotten within 24 hours [3]. Physicians demonstrate poor self-assessment accuracy, with systematic reviews finding no correlation between self-assessed and measured competence in 65% of comparisons [4]. This exposes the paradox: we design learning platforms as if clinicians are perfectly rational decision-makers while deploying behavioral interventions in clinical systems because we know they are not.

Digital platforms now dominate CPD. In 2022, accredited providers reported 230,000 total CME activities—a 13% increase—with online modules representing the most common format [5]. However, despite this technological transformation, most digital medical education assumes that clinicians will engage systematically, complete modules through intrinsic motivation, and independently translate knowledge into practice change.

Behavioral economics challenges this assumption. In clinical settings, health care organizations have proven that physician behavior responds predictably to choice architecture. When Penn Medicine redesigned electronic health record (EHR) defaults to favor generic medications, prescribing rates increased from 75% to 98% without additional education [6]. Health care nudge units have demonstrated that defaults, social comparison feedback, and environmental cues systematically shape clinical decisions [7]. These interventions succeed because clinicians, despite extensive training, exhibit status quo bias, present bias, decision fatigue, and susceptibility to social norms [8,9].

XSL•FO
RenderX

Decision fatigue alone has measurable effects: physicians prescribe antibiotics more frequently as the day progresses [10].

This paradox demands resolution. As medical education investment accelerates toward digital delivery, designing platforms on disproven assumptions wastes resources and perpetuates the knowing-doing gap. The behavioral science infrastructure that transformed clinical practice—nudge units, implementation scientists, and A/B testing—remains largely absent from educational technology development. Closing this gap requires systematic integration of behavioral economics into digital learning architecture.

This viewpoint argues that digital medical education should evolve toward behaviorally informed design, creating online learning architectures that work with human cognitive realities. We propose comprehensive integration of behavioral economics principles into digital CPD, offering both theoretical foundations and practical implementation frameworks. It is important to note that while these principles have demonstrated effectiveness in clinical contexts, their educational applications remain emerging and context-dependent, requiring systematic evaluation across diverse learning environments.

Terminology and Conceptual Clarity

Because behavioral economics terminology varies across disciplines, we establish the following definitions:

- Digital choice architecture refers to the systematic design of decision environments in digital learning platforms, encompassing defaults, interface structures, information displays, and feedback mechanisms that influence learner behavior while preserving choice. This term describes the overall design philosophy.
- Choice architecture represents the broader concept from behavioral economics, referring to how choices are organized and presented across any context [11].
- Behavioral design describes the practical application of behavioral economics and psychology principles to create systems supporting desired behaviors. We use this term when discussing the design process itself.
- Behavioral economics–informed design specifically indicates design decisions drawn from behavioral economics research, emphasizing the evidence-based grounding recommendations.
- Nudges are specific interventions that alter behavior predictably without forbidding options or significantly changing incentives [11]. Examples include defaults, social comparison feedback, and timely reminders.

These terms share conceptual overlap but maintain distinct emphases. Throughout this paper, we apply them consistently, distinguishing between broad design philosophy (choice architecture), practical implementation (behavioral design), specific interventions (nudges), and evidence-based approaches (behavioral economics–informed design).

Theoretical Foundations: Core Behavioral Economics Concepts

Behavioral economics integrates psychology and economics to understand how people actually make decisions—often deviating predictably from rational actor models. For digital medical education, these insights may explain why well-designed platforms achieve dramatically different learning outcomes than behaviorally naive alternatives, though direct empirical evidence in educational contexts remains limited.

Choice Architecture and Nudges

Choice architecture—defined by Thaler and Sunstein [11] as "organizing the context in which people make decisions"—recognizes that no choice presentation is neutral. Every interface element, default setting, and information display influences behavior. A comprehensive meta-analysis examining over 200 studies with 440+ effect sizes (N=2,148,439) found that choice architecture interventions promote behavior change with Cohen d=0.43, a small-to-medium effect practically significant for population-level interventions [8].

For digital learning platforms, choice architecture operates through every design decision. Whether modules appear in alphabetical order or by relevance to practice gaps shapes engagement. Whether completion tracking displays absolute progress ("3 of 10 modules") or social comparison ("You've completed fewer modules than 67% of peers") influences motivation through different behavioral mechanisms. The question is whether design reflects evidence or intuition.

Defaults and status quo bias leverage the tendency to stick with current states through multiple mechanisms: loss aversion, cognitive effort minimization, and implied endorsement [11]. Johnson and Goldstein [12] found that the organ donation consent rates were approximately twice as high in opt-out versus opt-in countries. Platforms that auto-enroll clinicians in modules addressing documented practice gaps leverage defaults while preserving opt-out capability.

Present bias and temporal discounting explain persistent procrastination. Clinicians understand CPD improves patient care, but immediate opportunity costs outweigh psychologically distant benefits. Laibson's [13] quasi-hyperbolic discounting model demonstrates that people systematically prefer smaller immediate rewards over larger delayed rewards. On the basis of these findings, we propose that digital platforms may counteract this through temporal optimization—delivering microlearning at the point of clinical relevance, which minimizes the gap between learning effort and application reward.

Decision fatigue and cognitive depletion have measurable effects on clinical judgment. As noted, physicians prescribe antibiotics more frequently as the day progresses [10]. Judicial rulings decline from 65% favorable at session start to near zero before breaks, returning to baseline after rest [14]. For digital learning, this suggests that platforms should minimize unnecessary choices that deplete cognitive resources rather than exhausting capacity on navigation decisions.

XSL•FO
RenderX

Social norms and social comparison powerfully influence behavior. Social comparison feedback showing how behavior compares to peers consistently drives behavior change across clinical contexts. Digital dashboards displaying peer prescribing comparisons reduce inappropriate antibiotic use [15]. On the basis of the available evidence, we propose that learning dashboards could leverage similar mechanisms, though whether clinical effects translate to educational outcomes requires systematic investigation. Studies of collaborative learning modules show that embedding social cues could enhance motivation and learning clarity [16], suggesting that digital platforms could achieve similar benefits through discussion forums and peer-review features integrated into the learning management systems (LMSs).

Integration With Complementary Theories

Behavioral economics does not operate in isolation. Cognitive load theory (CLT) provides mechanistic insights into working memory limitations—working memory can process only 4 to 7 information chunks simultaneously [17,18]. CLT distinguishes between intrinsic (inherent task difficulty), extraneous (imposed by suboptimal presentation), and germane (devoted to learning) cognitive loads. Choice architecture principles directly reduce extraneous cognitive load, freeing capacity for germane processing. The AMEE (Association for Medical Education in Europe) Guide No. 86 establishes CLT's "particular relevance to medical education because many professional activities require simultaneous integration of multiple knowledge sets" [19].

Self-determination theory identifies 3 psychological needs essential for intrinsic motivation: autonomy, competence, and relatedness [20]. Williams et al [21] demonstrated that "autonomously motivated learning leads to better educational outcomes." This creates productive tension with nudging approaches, as defaults can reduce perceived locus of causality. However, autonomy-preserving designs can reconcile these tensions through transparent nudges that explain rationale, self-nudges where learners choose their own interventions, and "boosts" that build competence rather than merely steering behavior. Hertwig and Grüne-Yanoff [17] distinguish boosting—which aims to "foster people's competence to make their own choices"—from nudging, arguing that boosts create lasting cognitive "capital stock" and necessarily preserve autonomy.

These complementary frameworks converge on similar design principles. Behavioral economics explains why interventions work (cognitive biases and heuristics), CLT specifies how much information learners can process, and self-determination theory identifies what motivates sustained engagement.

Digital Transformation and Clinical Lessons

The shift from classroom to technology-mediated education has fundamentally altered how clinicians access professional development. LMSs track every interaction—login times, completion rates, assessment attempts, and module abandonment points. Mobile apps deliver microlearning during clinical workflows. Artificial intelligence algorithms personalize content sequences. This digital infrastructure generates unprecedented behavioral data while creating new decision points where choice architecture can be applied. Every interface element implicitly shapes learning behavior. The question is not whether digital platforms incorporate choice architecture, but whether they do so intentionally and based on evidence.

Clinical decision support systems demonstrate what behaviorally informed digital design can achieve. EHRs now routinely incorporate default order sets favoring evidence-based practices. Digital dashboards provide peer comparison feedback on prescribing patterns. Predictive analytics trigger just-in-time reminders when clinical decisions deviate from best practice. A systematic review found robust evidence that digital defaults and social proof could meaningfully change a physician's behavior across multiple domains [21]. Notably, these effects persist among experienced practitioners—behavioral tendencies are not eliminated by expertise. Digital choice architecture in clinical systems represents the convergence of behavioral science with technological capability, creating decision environments that systematically support better choices at scale.

Behavioral Design Gap in Digital Medical Education

Applications of behavioral economics to online medical education exist but remain limited. Emerging evidence demonstrates both promise and important limitations.

Gurley et al [22] found that email reminders increased residency assessment completion rates, though the intervention required repeated prompts and the effect sizes were modest. More comprehensive interventions showed mixed results at scale. Kizilcec et al's [23] landmark study tested 5 behavioral interventions across 269,169 students in 247 massive open online courses. Interventions showing effectiveness at a small scale (planning prompts: +29% completion for committed students) demonstrated minimal impact when implemented broadly. Only value-relevance interventions showed modest improvements for disadvantaged students in courses with existing achievement gaps. This large-scale evidence reveals critical constraints: behavioral interventions may require targeting and personalization, and what works in controlled trials may not translate to population-level deployment.

More promising results emerge from targeted interventions. A 2023 study combining learning analytics with behavioral nudges found personalized reminders significantly increased student resource clicks (planning courses: 415 to 778 clicks, $P<.001$; education courses: 304-446 clicks, $P<.001$) [24]. Research on tailored digital nudges matched to students' motivation profiles showed improved engagement when interventions aligned with learner characteristics [25].

Medical education–specific applications remain rare. Van Gaalen et al's [26] systematic review found that gamification in health professional education showed no negative outcomes, though most studies were descriptive with few using theory or well-defined control groups. Kerfoot et al's [27] high-quality randomized controlled trial with 141 - medical students found

XSL•FO
RenderX

that gamified and spaced learning improved hypertension management knowledge and produced modest reductions in days to reach target blood pressure—demonstrating transfer to patient outcomes.

Recent evidence from anatomy education demonstrated how behavioral principles could be successfully embedded. Specialized study modules (SSMs) incorporating intentional pathway design, collaborative learning structures, and clear learning objectives produced significant improvements in student motivation, teamwork, and clarity of learning goals [16,28]. These findings parallel mechanisms that digital choice architecture seeks to operationalize: both approaches rely on well-designed pathways that make desired behaviors easier, clearer, and more rewarding. Just as SSMs improved learning through structured guidance, digital platforms can leverage defaults, social nudges, and collaborative cues to streamline learning decisions and reduce friction. The observed heterogeneity in learning strategies across educational contexts [16] reinforces the need for adaptive nudges and personalized digital pathways accommodating diverse learner preferences.

Systematic reviews note that education-related studies represent only 4% of choice architecture interventions tested across all domains [29], with learning analytics in medical education "still in infancy" with "systematic applications limited" [30]. Platforms rarely use predictive analytics to identify at-risk learners or deploy targeted nudges. Research demonstrates that only half of people who form intentions actually take action [31], highlighting the critical intention-behavior gap. The majority of accredited CPD activities do not target clinical behavior change [32], with behavior change techniques appearing in only a subset of CME [32].

Considering typical design patterns in online CPD platforms, the default pathway after completing a module is returning to the dashboard—no implementation planning, no commitment to practice change, and no follow-up accountability. Courses are structured as optional selections requiring active search rather than defaults based on practice gaps. Progress tracking emphasizes completion percentages without peer comparison. Educational content is delivered months before clinical application, despite evidence that temporal distance undermines behavior change. The field that transformed clinical practice through digital behavior change has barely touched the digital systems used to educate clinicians.

A Framework for Digital Choice Architecture in Online Medical Education

Digital choice architecture for medical education means intentionally designing online learning environments using behavioral principles proven effective in clinical contexts. We propose a comprehensive framework integrating established implementation science models with practical design principles. It is critical to emphasize that educational applications of these principles remain emerging, requiring context-specific adaptation and rigorous evaluation.

Established Implementation Frameworks

Multiple frameworks can guide the systematic application of behavioral insights to educational platforms. The UK Behavioral Insights Team's EAST (easy, attractive, social, and timely) framework provides the following accessible design heuristics [33] as : Make it Easy (reduce effort through defaults, simplified messages, and minimal "hassle factors"), Make it Attractive (use visual design, immediate rewards, and salient feedback), Make it Social (leverage peer comparison, network effects, and commitment to others), and Make it Timely (deliver prompts when learners are most receptive).

The COM-B (capability, opportunity, motivation, and behavior) model provides diagnostic capacity by conceptualizing behavior as arising from the interaction of 3 core components: capability (physical and psychological), opportunity (environmental and social enablers), and motivation (reflective and automatic processes) [34]. This diagnostic approach links to the behavior change wheel's 9 intervention functions, enabling systematic selection of appropriate behavioral strategies based on identified barriers. The theoretical domains framework extends analysis across 14 granular domains, including knowledge, skills, professional identity, and beliefs about capabilities [35]. The Behavioral Insights Team's 4-step method provides a cyclical process: define the outcome, understand the context, build your intervention, and test or learn or adapt [33].

Stepwise Design Principles for LMS Platforms

Translating frameworks into practice requires systematic steps.

Step 1: Audit Current Choice Architecture

Before implementing behavioral design, organizations must document existing implicit architecture. This includes mapping all default settings. For example, what happens when learners complete a module? What enrollment pathways exist? Organizations should identify every decision point where learners make choices, catalog implicit assumptions embedded in platform design, and establish baseline behavioral metrics including completion rates, time-to-completion, drop-off points, and peak engagement times. This audit reveals where behavioral insights can add value while providing baseline data for measuring improvement.

Step 2: Apply Behavioral Design Principles

Defaults leverage status quo bias. Auto-enrollment in required modules based on practice type or documented knowledge gaps preserves choice while making evidence-based learning the path of least resistance. Research demonstrates that providing clear, guided pathways—similar to defaults—improves motivation and learning clarity while reducing cognitive burden [16,28]. Evidence-based pathway before selection prioritizes high-impact content. Postmodule defaults should route learners to implementation planning tools rather than simply returning to catalogs. Assessment scheduling can prepopulate optimal spacing intervals. Each default preserves learner autonomy through transparent opt-out.

Friction reduction addresses that every additional click creates abandonment opportunity. One-click enrollment minimizes barriers. Single sign-on eliminates authentication friction.

XSL•FO
RenderX

Streamlined navigation reduces unnecessary steps. Form simplification eliminates nonessential fields. These micro-optimizations cumulatively reduce extraneous cognitive load.

Social comparison and peer feedback leverage descriptive norms without public shaming. Learning dashboards can display "You have completed 60% of required CPD; 80% of peers in your specialty have completed more." Anonymized leaderboards show rankings for voluntary activities. Specialty-specific benchmarks compare engagement to relevant peer groups. Completion milestones highlight social norms: "Join the 73% of cardiologists who completed this module." Peer-supported learning structures enhance motivation and learning clarity [16], suggesting that platforms can achieve benefits through discussion forums, peer-review features, and collaborative assignments integrated into LMSs.

Timely prompts and momentum nudges address present bias. Deadline reminders sent 3 days before, 1 day before, and on the day counteract procrastination. Momentum messages like "You're 80% through this module—finish now?" capitalize on progress. Peak engagement timing delivers notifications at moments when individual users are most likely to engage, based on their historical behavior. Calendar integration schedules protected learning time directly into clinicians' calendars.

Personalization based on behavioral data increases relevance while reducing cognitive load. Role-based recommendations curate content by specialty. Practice-gap-informed suggestions link learning to quality metrics. Adaptive difficulty adjusts complexity based on assessment performance. Temporal optimization delivers notifications when individual learners historically show the highest engagement.

Step 3: Implement With Transparency

Ethical behavioral design requires disclosure and opt-out mechanisms. Platform documentation should clearly explain that the system uses behavioral design principles. Opt-out mechanisms must allow users to disable social comparison features, modify notification frequency, or change default pathways. Transparent nudges make intent reconstructable—for example, "This module was recommended based on your practice gaps in diabetes management." All behavioral design choices should be documented and reviewed through appropriate oversight mechanisms. Research shows transparent nudges remain effective [36,37].

Step 4: Measure and Iterate

Continuous measurement enables data-driven improvement. Organizations should track engagement metrics (logins, completion rates, time-on-platform, and drop-off points), learning outcomes (assessment scores and knowledge retention), and practice change where feasible (quality metrics and chart audits). A/B testing allows systematic comparison of behavioral interventions. User feedback through surveys and qualitative interviews provides insight into learner experience. On the basis of these data, organizations should refine interventions quarterly, abandoning approaches that do not work and scaling those that do.

Digital Commitment Devices and Implementation Intentions

The intention-behavior gap represents a critical barrier. Research demonstrates that only half of the people who form intentions actually take action [31]. Digital platforms can incorporate commitment mechanisms grounded in implementation intention research, which shows medium-to-large effect sizes (Cohen d=0.65) when individuals create specific if-then plans [38]. Practical applications include implementation planning prompts after module completion requiring learners to specify "When will you apply this? In what clinical situation?" Calendar commitments directly schedule follow-up learning or application time blocks. Automated accountability check-ins follow up, asking "Did you apply what you learned? What barriers did you encounter?"

Implementation Roadmap for Health Care Organizations

Health care organizations can follow these phased approaches.

Phase 1: Assessment and Planning (Months 1-3)

Phase 1 includes the following:

1. Learning analytics specialists conduct comprehensive choice architecture audits documenting all platform defaults, decision points, and implicit assumptions.
2. Learner surveys identify motivation barriers and preferences while behavioral analytics reveal completion patterns and peak engagement times.
3. Deliverables include current state reports, priority intervention lists ranked by impact-effort ratio, baseline metrics dashboards, and stakeholder engagement plans. This requires a learning analytics specialist, behavioral scientist consultant (0.2 - 0.5 full-time equivalents), platform administrator, end-user clinicians, and an instructional designer.

Phase 2: Pilot Interventions (Months 4-9)

Phase 2 consists of the following:

1. Implement 3 to 5 behavioral design changes in a limited scope, typically targeting a single department.
2. Conduct A/B testing where feasible. Measurement includes quantitative metrics (completion rates, time-to-completion, and assessment performance) and qualitative methods (user feedback and focus groups). Example pilots: changing default postmodule action to "create implementation plan," adding peer comparison feedback to learner dashboards, implementing smart reminders based on individual engagement patterns, testing auto-enrollment versus opt-in, or adding commitment devices.
3. Track completion rates, time-to-completion, assessment performance, user satisfaction, opt-out rates, and qualitative feedback themes.

Phase 3: Scale and Sustain (Months 10+)

Phase 3 consists of the following:

1. Roll out successful interventions system-wide.

XSL•FO
RenderX

2. Build ongoing A/B testing infrastructure.
3. Train staff in behavioral design principles.
4. Establish governance structures for ethical oversight.
5. Integrate behavioral metrics into routine quality improvement cycles.
6. Share results through publications and presentations.
7. Sustainability requires behavioral science expertise (0.5 - 1.0 FTE), learning analytics infrastructure, regular review cycles, ethical oversight committees, and continuous professional development.

Resource Requirements and Common Pitfalls

Resource requirements scale with organizational size. Small organizations serving fewer than 5000 learners typically invest US $50,000 to US $100,000 annually, including part-time behavioral science consultation, analytics platform subscriptions, and platform customization. Medium organizations serving 5000 to 50,000 learners typically invest US $200,000 to US $400,000 annually, including full-time behavioral science expertise, dedicated analytics platforms, testing infrastructure, and platform development. Large organizations serving over 50,000 learners typically invest US $500,000 to US $1 million or more annually, including behavioral science teams, advanced analytics platforms, dedicated nudge unit infrastructure, and platform development teams. These figures represent ongoing investments necessary for effective educational delivery in the digital age.

Organizations should avoid common implementation pitfalls. Overcomplication from starting with too many simultaneous interventions makes isolating effects impossible; organizations should begin with 2 to 3 high-priority changes, demonstrate value, and then expand. Metrics myopia from optimizing purely for completion rates without monitoring quality or satisfaction can lead to gaming behaviors; use balanced scorecards instead. Transparency failures from implementing behavioral interventions without disclosure violate ethical principles and risk backlash; always explain design rationale up front. Ethical shortcuts from rushing implementation without ethical review create institutional risks; establish oversight early. Siloed work with behavioral science operating independently from instructional design creates a fragmented user experience; integrate from the start. Technology determinism assumes that building features automatically change behavior without attention to psychology; focus on behavioral science first and technology second. Insufficient testing by scaling interventions before demonstrating effectiveness wastes resources; pilot rigorously before scaling. Ignoring heterogeneity by assuming all learners respond identically misses personalization opportunities; conduct segment analyses and adapt interventions accordingly.

Research Agenda

Translating behavioral economics from clinical to educational digital systems requires systematic investigation. Key priorities include conducting comprehensive audits of existing online CPD platforms to document current choice architecture, cataloging defaults and behavioral assumptions. Randomized controlled trials should compare traditional digital learning designs with behaviorally optimized alternatives, measuring not just completion but practice change outcomes. Implementation frameworks need development to provide practical guidance for educational institutions.

The key research questions are as follows:

1. Do behavioral interventions remain effective in clinical EHRs transfer to educational LMS contexts with similar effect sizes?
2. How does professional identity moderate responses to digital educational nudges?
3. What is the optimal balance between algorithmic personalization and learner autonomy?
4. How do clinicians respond to transparent disclosure that platforms incorporate behavioral design?

Technical implementation requires interdisciplinary collaboration [39]. Educational technologists must partner with behavioral scientists to translate abstract principles into specific interface elements. Data scientists need frameworks for analyzing learning analytics through a behavioral lens. Software developers require training in ethical behavioral design to avoid dark patterns that maximize compliance at the expense of learning quality or user autonomy.

Ethical Implications of Behavioral Design in Digital Learning

Applying behavioral economics to digital education raises legitimate concerns about manipulation and autonomy. Well-designed nudges can preserve freedom of choice while supporting goal achievement [40]. The distinction between facilitation and manipulation lies in transparency, alignment with learner objectives, and preservation of autonomy.

The Hansen and Jespersen Framework

Hansen and Jespersen [36] provide essential ethical distinctions through a 2D typology. Nudges vary along cognitive processing (type 1 automatic vs type 2 reflective) and transparency (transparent vs nontransparent intent and mechanism). Transparent type 2 nudges engage reflective processing while making intent and mechanism clear—these "do not comprise the unethical aspect of manipulating or secretly influencing individuals' behavior" and represent the gold standard [37]. For example, "This module was recommended based on your documented practice gaps in diabetes management. View criteria and opt out if not relevant." Transparent type 1 nudges operate through automatic processing but are disclosed. For example, "We've pre-selected the most commonly chosen learning path for your specialty. You can customize at any time." Nontransparent nudges raise manipulation concerns and are generally impermissible in educational contexts.

Supportive Versus Coercive Design

Supportive nudges preserve choice through easy, clear opt-out mechanisms. They maintain transparent intent, align with learners' stated goals, build capability alongside behavioral steering, and respect autonomy. In contrast, coercive manipulation obscures alternatives or makes opt-out difficult. It uses hidden intent, optimizes for platform engagement metrics

XSL•FO
RenderX

rather than learner welfare, and uses dark patterns such as fake urgency, hidden costs, forced continuity, or obstruction.

Research decisively shows that transparent nudges are not less effective than nontransparent alternatives [36,37]. This evidence suggests that transparency serves as an "ethical filter" without sacrificing outcomes, meaning platforms can—and should—openly explain their behavioral design principles while maintaining effectiveness.

Ethical Decision Framework: 5-Step Process

Organizations implementing digital choice architecture should apply the following framework before deploying any intervention:

1. Is the nudge transparent? Can learners reconstruct the intent?
2. Does it engage reflective processing (type 2)? Prefer interventions that prompt conscious consideration.
3. Is opt-out easy and clear? Can learners disable the intervention in 3 clicks or fewer?
4. Does it align with learner objectives? Is the intervention designed to help learners achieve their stated goals?
5. Does it build competence alongside behavior change? Does the intervention include educational components explaining why the behavior is beneficial?

If an intervention fails any criteria, it should be redesigned or rejected.

Dark Patterns: Explicitly Prohibited

Educational platforms must avoid deceptive design practices including fake scarcity ("Only 3 spots left!" when capacity is unlimited), hidden costs, forced continuity, obstruction (making cancelation unreasonably difficult), misdirection, nagging (excessive notifications), and sneaking (adding items without consent). These practices violate trust, undermine autonomy, and have no place in professional education.

The Future of Behaviorally Informed Digital Medical Education

The integration of behavioral economics into digital medical education represents a paradigm shift from information-deficit models to evidence-based decision support. As artificial intelligence and adaptive learning technologies mature, behavioral design principles will become increasingly critical. Machine learning algorithms can personalize nudges based on individual response patterns, learning analytics can identify optimal timing for interventions, and natural language processing can generate personalized feedback. However, technological sophistication must be matched by ethical sophistication. The more powerful these tools become, the more important transparency and oversight become.

Looking forward, digital choice architecture may extend beyond traditional CPD to include just-in-time clinical decision support that simultaneously educates and guides practice, simulation-based training environments that use defaults and social comparison to accelerate skill acquisition, and longitudinal professional development pathways that adapt to career stages and practice evolution. The convergence of behavioral science, educational technology, and implementation science offers unprecedented opportunity to close the knowing-doing gap that has plagued medical education for decades.

Limitations and Future Research

This viewpoint has important limitations. First, we extrapolate from clinical behavioral interventions to educational contexts without comprehensive evidence that effect sizes transfer. Educational effects of behavioral design principles are inferred from clinical contexts and remain context-dependent, requiring systematic empirical validation across diverse learning environments before widespread adoption. The Kizilcec et al's [23] study demonstrates that small-scale educational successes may not replicate at the population level, suggesting caution in assuming that clinical behavioral economics directly translates to educational platforms. Second, we focus primarily on CPD rather than undergraduate or postgraduate training, where different considerations may apply. Third, ethical frameworks proposed require empirical validation—we need evidence on how transparent nudges affect trust, engagement, and learning outcomes specifically in medical education contexts.

Fourth, there exists a risk of overoptimization or technocratic control, where excessive focus on measurable behavioral metrics (completion rates and click-through rates) may overshadow harder-to-quantify educational outcomes such as critical thinking, professional identity formation, or intrinsic motivation. Behavioral design could inadvertently create systems that optimize for compliance rather than meaningful learning.

Fifth, resistance from clinicians who value professional autonomy represents a significant implementation barrier. Physicians may perceive behavioral interventions—even transparent ones—as patronizing or as threats to professional independence. This resistance may be particularly strong among senior clinicians with established practice patterns who view educational decisions as matters of professional judgment rather than targets for behavioral modification.

Sixth, institutional constraints pose substantial challenges. These include cost barriers (significant upfront investment in behavioral science expertise, analytics infrastructure, and platform redesign), governance complexities (establishing ethical oversight, determining acceptable intervention boundaries, and managing stakeholder concerns), and legacy system limitations (many organizations operate on established LMS platforms with limited customization capabilities, vendor dependencies, and lengthy procurement cycles that impede rapid iteration). Health care organizations must navigate these practical constraints while attempting implementation.

Future research should prioritize large-scale randomized trials comparing behaviorally optimized platforms to traditional designs across multiple outcomes (completion, learning, satisfaction, and practice change). Studies should examine heterogeneity in treatment effects—Which clinician subgroups respond most strongly to which interventions? Implementation science research should document barriers and facilitators to

XSL•FO
RenderX

adopting behavioral design in health care organizations. Ethical research should examine learner perceptions of nudges, optimal disclosure methods, and long-term effects on professional development autonomy.

Conclusions

The question facing medical education is not whether to design learning environments but whether to design them well or badly. Every digital interface embodies choice architecture—defaults, menu structures, notification timing, and progress displays—that shapes learning behavior. Currently, most platforms incorporate architecture based on intuition rather than evidence. This represents a missed opportunity of staggering proportions.

Health care organizations should conduct choice architecture audits of existing CPD platforms, identifying defaults and decision points; pilot behavioral interventions with rigorous measurement before scaling; and partner with behavioral scientists during platform design, not as an afterthought. Academic medical centers should deploy implementation science frameworks to systematically translate behavioral insights into educational practice. Accreditation bodies should incentivize behavioral design through quality metrics that reward practice change, not just completion rates. Vendors should partner with behavioral scientists from the design stage, building behavioral principles into product architecture. Professional societies should develop best practice guidelines for ethical behavioral design in medical education, facilitate sharing through communities of practice, and advocate for research funding.

The evidence base exists. The tools are available. The imperative is urgent. The path is clear. The question is whether we will take it.

Acknowledgments

The author thanks colleagues at the SingHealth Duke-NUS Medical Humanities Institute for their feedback on earlier drafts.

Authors' Contributions

VE is the sole author and conceived, researched, and wrote this paper.

Conflicts of Interest

None declared.

References

1. Forsetlund L, O'Brien MA, Forsén L, et al. Continuing education meetings and workshops: effects on professional practice and healthcare outcomes. Cochrane Database Syst Rev 2021 Sep 15;9(9):CD003030. [doi: 10.1002/14651858.CD003030.pub3] [Medline: 34523128]
2. Cabana MD, Rand CS, Powe NR, et al. Why don't physicians follow clinical practice guidelines? A framework for improvement. JAMA 1999 Oct 20;282(15):1458-1465. [doi: 10.1001/jama.282.15.1458] [Medline: 10535437]
3. Murre JMJ, Dros J. Replication and analysis of Ebbinghaus' forgetting curve. PLoS One 2015;10(7):e0120644. [doi: 10.1371/journal.pone.0120644] [Medline: 26148023]
4. Davis DA, Mazmanian PE, Fordis M, Van Harrison R, Thorpe KE, Perrier L. Accuracy of physician self-assessment compared with observed measures of competence: a systematic review. JAMA 2006 Sep 6;296(9):1094-1102. [doi: 10.1001/jama.296.9.1094] [Medline: 16954489]
5. ACCME data report: renewal and growth in accredited continuing education – 2022. : Accreditation Council for Continuing Medical Education; 2022 URL: https://accme.org/wp-content/uploads/2024/05/2022-ACCME-Annual-Data-Report-1003_20230808.pdf?utm_source=chatgpt.com [accessed 2026-01-24]
6. Patel MS, Day SC, Halpern SD, et al. Generic medication prescription rates after health system-wide redesign of default options within the electronic health record. JAMA Intern Med 2016 Jun 1;176(6):847-848. [doi: 10.1001/jamainternmed.2016.1691] [Medline: 27159011]
7. Patel MS, Volpp KG, Asch DA. Nudge units to improve the delivery of health care. N Engl J Med 2018 Jan 18;378(3):214-216. [doi: 10.1056/NEJMp1712984] [Medline: 29342387]
8. Mertens S, Herberz M, Hahnel UJJ, Brosch T. The effectiveness of nudging: a meta-analysis of choice architecture interventions across behavioral domains. Proc Natl Acad Sci U S A 2022 Jan 4;119(1):e2107346118. [doi: 10.1073/pnas.2107346118] [Medline: 34983836]
9. Samuelson W, Zeckhauser R. Status quo bias in decision making. J Risk Uncertainty 1988 Mar;1(1):7-59. [doi: 10.1007/BF00055564]
10. Linder JA, Doctor JN, Friedberg MW, et al. Time of day and the decision to prescribe antibiotics. JAMA Intern Med 2014 Dec;174(12):2029-2031. [doi: 10.1001/jamainternmed.2014.5225] [Medline: 25286067]
11. Thaler RH, Sunstein CR. Nudge: Improving Decisions About Health, Wealth, and Happiness: Yale University Press; 2008. URL: https://books.google.co.in/books/about/Nudge.html?id=dSJQn8egXvUC [accessed 2026-01-28]

XSL•FO
RenderX

12. Johnson EJ, Goldstein DG. Medicine. Do defaults save lives? Science 2003 Nov 21;302(5649):1338-1339. [doi: 10.1126/science.1091721] [Medline: 14631022]
13. Laibson D. Golden eggs and hyperbolic discounting. Q J Econ 1997 May 1;112(2):443-478. [doi: 10.1162/003355397555253]
14. Danziger S, Levav J, Avnaim-Pesso L. Extraneous factors in judicial decisions. Proc Natl Acad Sci U S A 2011 Apr 26;108(17):6889-6892. [doi: 10.1073/pnas.1018033108] [Medline: 21482790]
15. Meeker D, Linder JA, Fox CR, et al. Effect of behavioral interventions on inappropriate antibiotic prescribing among primary care practices: a randomized clinical trial. JAMA 2016 Feb 9;315(6):562-570. [doi: 10.1001/jama.2016.0275] [Medline: 26864410]
16. Ogut E, Yildirim FB, Senol Y, Senol AU. Comprehensive evaluation of the educational impact and effectiveness of specialized study modules in cross-sectional anatomy: a study on student engagement and learning outcomes. BMC Med Educ 2025 Apr 10;25(1):514. [doi: 10.1186/s12909-025-07050-9] [Medline: 40211255]
17. Hertwig R, Grüne-Yanoff T. Nudging and boosting: steering or empowering good decisions. Perspect Psychol Sci 2017 Nov;12(6):973-986. [doi: 10.1177/1745691617702496] [Medline: 28792862]
18. Blumenthal-Barby JS, Krieger H. Cognitive biases and heuristics in medical decision making: a critical review using a systematic search strategy. Med Decis Making 2015 May;35(4):539-557. [doi: 10.1177/0272989X14547740] [Medline: 25145577]
19. Young JQ, Van Merrienboer J, Durning S, Ten Cate O. Cognitive load theory: Implications for medical education: AMEE Guide No. 86. Med Teach 2014 May;36(5):371-384. [doi: 10.3109/0142159X.2014.889290] [Medline: 24593808]
20. Ryan RM, Deci EL. Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being. Am Psychol 2000;55(1):68-78. [doi: 10.1037//0003-066x.55.1.68] [Medline: 11392867]
21. Williams GC, Saizow RB, Ryan RM. The importance of self-determination theory for medical education. Acad Med 1999 Sep;74(9):992-995. [doi: 10.1097/00001888-199909000-00010] [Medline: 10498090]
22. Gurley A, Jenkins C, Nguyen T, Woodall A, An J. Nudge theory: effectiveness in increasing emergency department faculty completion of residency assessments. West J Emerg Med 2024 Jan;25(1):40-42. [doi: 10.5811/westjem.57721] [Medline: 38205983]
23. Kizilcec RF, Reich J, Yeomans M, et al. Scaling up behavioral science interventions in online education. Proc Natl Acad Sci U S A 2020 Jun 30;117(26):14900-14905. [doi: 10.1073/pnas.1921417117] [Medline: 32541050]
24. Brown A, Basson M, Axelsen M, Redmond P, Lawrence J. Empirical evidence to support a nudge intervention for increasing online engagement in higher education. Educ Sci 2023;13(2):145. [doi: 10.3390/educsci13020145]
25. Plak S, Cornelisz I, Meeter M, van Klaveren C. Early warning systems for more effective student counselling in higher education: evidence from a Dutch field experiment. High Educ Q 2022 Jan;76(1):131-152. [doi: 10.1111/hequ.12298]
26. van Gaalen AEJ, Brouwer J, Schönrock-Adema J, Bouwkamp-Timmer T, Jaarsma ADC, Georgiadis JR. Gamification of health professions education: a systematic review. Adv in Health Sci Educ 2021 May;26(2):683-711. [doi: 10.1007/s10459-020-10000-3]
27. Kerfoot BP, Baker H, Pangaro L, et al. An online spaced-education game to teach and assess medical students: a multi-institutional prospective trial. Acad Med 2012 Oct;87(10):1443-1449. [doi: 10.1097/ACM.0b013e318267743a] [Medline: 22914524]
28. Bhatti F, Mowforth O, Butler M, et al. Meeting the shared goals of a student-selected component: pilot evaluation of a collaborative systematic review. JMIR Med Educ 2023 Mar 15;9:e39210. [doi: 10.2196/39210] [Medline: 36920459]
29. Szaszi B, Palinkas A, Palfi B, Szollosi A, Aczel B. A systematic scoping review of the choice architecture movement: toward understanding when and why nudges work. J Behav Decis Mak 2018 Jul;31(3):355-366. [doi: 10.1002/bdm.2035]
30. Bojic I, Mammadova M, Ang CS, et al. Empowering health care education through learning analytics: in-depth scoping review. J Med Internet Res 2023 May 17;25:e41671. [doi: 10.2196/41671] [Medline: 37195746]
31. Gollwitzer PM, Sheeran P. Implementation intentions and goal achievement: a meta‐analysis of effects and processes. Adv Exp Soc Psychol 2006;38:69-119. [doi: 10.1016/S0065-2601(06)38002-1]
32. Légaré F, Freitas A, Thompson-Leduc P, et al. The majority of accredited continuing professional development activities do not target clinical behavior change. Acad Med 2015 Feb;90(2):197-202. [doi: 10.1097/ACM.0000000000000543] [Medline: 25354076]
33. EAST: four simple ways to apply behavioural insights. : The Behavioural Insights Team; 2014 URL: https://www.behaviourchange.net/docs/bit-publication-east_fa_web.pdf [accessed 2026-01-24]
34. Michie S, van Stralen MM, West R. The behaviour change wheel: a new method for characterising and designing behaviour change interventions. Implement Sci 2011 Apr 23;6:42. [doi: 10.1186/1748-5908-6-42] [Medline: 21513547]
35. Atkins L, Francis J, Islam R, et al. A guide to using the theoretical domains framework of behaviour change to investigate implementation problems. Implement Sci 2017 Jun 21;12(1):77. [doi: 10.1186/s13012-017-0605-9] [Medline: 28637486]
36. Hansen PG, Jespersen AM. Nudge and the manipulation of choice: a framework for the responsible use of the nudge approach to behaviour change in public policy. Eur J Risk Regul 2013;4(1):3-28. [doi: 10.1017/S1867299X00002762]
37. Gollwitzer PM. Implementation intentions: strong effects of simple plans. Am Psychol 1999;54(7):493-503. [doi: 10.1037/0003-066X.54.7.493]

XSL•FO
RenderX

38. Bieleke M, Keller L, Gollwitzer PM. If-then planning. Eur Rev Soc Psychol 2021 Jan 2;32(1):88-122. [doi: 10.1080/10463283.2020.1808936]
39. Hodson N, Powell BJ, Nilsen P, Beidas RS. How can a behavioral economics lens contribute to implementation science? Implement Sci 2024 Apr 26;19(1):33. [doi: 10.1186/s13012-024-01362-y] [Medline: 38671508]
40. Sunstein CR. The ethics of nudging. Yale J Regul 2015;32(2):413-450 [FREE Full text]

Abbreviations

AMEE: Association for Medical Education in Europe
CLT: cognitive load theory
CME: continuing medical education
COM-B: capability, opportunity, motivation, and behavior
CPD: continuing professional development
EAST: easy, attractive, social, and timely
EHR: electronic health record
LMS: learning management system

Edited by J Gentges; submitted 25.Oct.2025; peer-reviewed by E Ogut, O Ng; revised version received 07.Jan.2026; accepted 14.Jan.2026; published 06.Feb.2026.

Please cite as:
Ekstrom V
Digital Choice Architecture in Medical Education: Applying Behavioral Economics to Online Learning Environments
JMIR Med Educ 2026;12:e86497
URL: https://mededu.jmir.org/2026/1/e86497
doi:10.2196/86497

© Victoria Ekstrom. Originally published in JMIR Medical Education (https://mededu.jmir.org), 6.Feb.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Education, is properly cited. The complete bibliographic information, a link to the original publication on https://mededu.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Implementing Large Language Models to Support Misconception-Based Collaborative Learning in Health Care Education

Brandon C J Cheah[1,2]; Shefaly Shorey[3], RN, RM, PhD; Jun Hong Ch'ng[1], BSc, PhD; Chee Wah Tan[1], BSc, PhD

[1]Department of Microbiology and Immunology, Yong Loo Lin School of Medicine, National University of Singapore, 5 Science Drive 2, Singapore, Singapore

[2]Department of Pharmacy and Pharmaceutical Sciences, National University of Singapore, Singapore, Singapore

[3]Alice Lee Centre of Nursing Studies, Yong Loo Lin School of Medicine, National University of Singapore, Singapore, Singapore

Corresponding Author:
Chee Wah Tan, BSc, PhD
Department of Microbiology and Immunology, Yong Loo Lin School of Medicine, National University of Singapore, 5 Science Drive 2, Singapore, Singapore

Abstract

This paper proposes a framework for leveraging large language models (LLMs) to generate misconceptions as a tool for collaborative learning in health care education. While misconceptions—particularly those generated by AI—are often viewed as detrimental to learning, we present an alternative perspective: that LLM-generated misconceptions, when addressed through structured peer discussion, can promote conceptual change and critical thinking. The paper outlines use cases across health care disciplines, including both clinical and basic science contexts, and a practical 10-step guidance for educators to implement the framework. It also highlights the need for medium- to long-term research to evaluate the impact of LLM-supported learning on student outcomes. This framework may support health care educators globally in integrating emerging AI technologies into their teaching, regardless of the disciplinary focus.

(JMIR Med Educ 2026;12:e81875) doi:10.2196/81875

KEYWORDS

large language model; refutation text; collaborative learning; health care education; misconception-based learning; LLM

Introduction

In an age where the availability of information is the norm and not the exception, misinformation from a range of information sources—including social media—can dominate the information landscape. Potential misinformation exists in the medical world, given the increasing rates of knowledge dissemination and the presence of potentially misleading information sources.

Medical practitioners and health professionals are not exempt from its effects, with the consequences having wide-reaching implications. An example of a high-profile case where physicians generated harm due to misinformation includes the physician-authored Great Barrington Declaration during the COVID-19 pandemic, which may have potentially increased resistance to lockdown policies in the United States and Europe [1]. Other examples include a Houston-based physician tweeting in 2022 to her account of more than 50,000 followers that there were no medically valid reasons for vaccination at the time [2], or physician-promoted anti-diet movements supporting the consumption of discretionary foods [3]. Medical training must therefore explicitly cultivate the capacity to detect and dismantle misinformation.

To address misconceptions stemming from misinformation, examples of current approaches by medical educators include manual assessment of student answers in examinations [4] and in-class polling [5] to identify misconceptions for use in misconception-based learning. However, these methods assume that (1) various student cohorts over time share similar misconceptions, without potential idiosyncratic variability in thought patterns; (2) educators have a breadth of experience with multiple cohorts; and (3) educators have the time capacity to generate multiple misconceptions in-depth. These assumptions may not hold true in practice, ultimately limiting the scalability of traditional misconception-based learning to educator experience.

The advent of ChatGPT and other large language models (LLMs) fueled by artificial intelligence (AI) presents opportunities to augment misconception-based learning. LLMs have demonstrated potential to personalize learning experiences, offer instant feedback and enhance student engagement [6]. Within misconception-based learning, LLMs offer the potential to both assist educators in generating misconceptions and students in identifying inaccurate thought patterns through feedback. However, limited research exists on LLM-assisted misconception-based learning in health care education.

XSL•FO
RenderX

Therefore, key questions arise: (1) how do LLMs fit within the process of misconception-based learning; and (2) how can LLMs be applied in the context of misconception-based learning in health care education? To address the research gap, this paper serves as a theoretical foundation and a conceptual framework with potential cases to apply LLMs in health care education.

Theoretical Premise of Misconception-Based Learning

Misinformation, and later entrenched misconceptions, can result from cognitive biases [7] and the lack of critical thinking [8]. It is important to note that scholarly debate exists on the use of the word misconception, particularly surrounding its definition and appropriateness of use [9]. Here misconceptions are referred to in the lens of the Piagetian constructivist theory, which sees misconceptions as a part of active constructions of students' mental models [10]. Misconceptions, in this context, are not deficiencies in knowledge but rather a gap between the student's perceptions and scientifically accurate information. To be more effective, educators could elucidate and challenge misconceptions early in the process of knowledge acquisition.

One of the foundations of health care professions education lies in conceptual acquisition. Conceptual acquisition could range from logic-derived processes (ie, mechanics of inspiration and breathing in physiology) to memory-based internalization (ie, eukaryotic transcription factors in molecular biology). Misconceptions would thus be the students' prior beliefs and constructed mental models that contradict accepted scientific concepts [11]. In health care professions, not addressing misconceptions originating from education may have ramifications on clinical reasoning and patient safety [12].

Figure 1 presents a simplified model of the process surrounding scientific conceptual acquisition. The assumed educational goal is the possession of student-learned knowledge supported by factual accuracy (output). The student is first given an input of target knowledge either through didactic or autodidactic methods. Given that the target knowledge input in health sciences can be highly conceptual, which has little to no attribution to sensorimotor information [13], the student is likely to begin the process of mental abstraction that includes simplification and consolidation of knowledge (labeled (A) in Figure 1). Conceptual acquisition could occur in a mechanistic process, with the assembly of smaller and cognitively plausible concepts as a bootstrap for developing larger concepts [14].

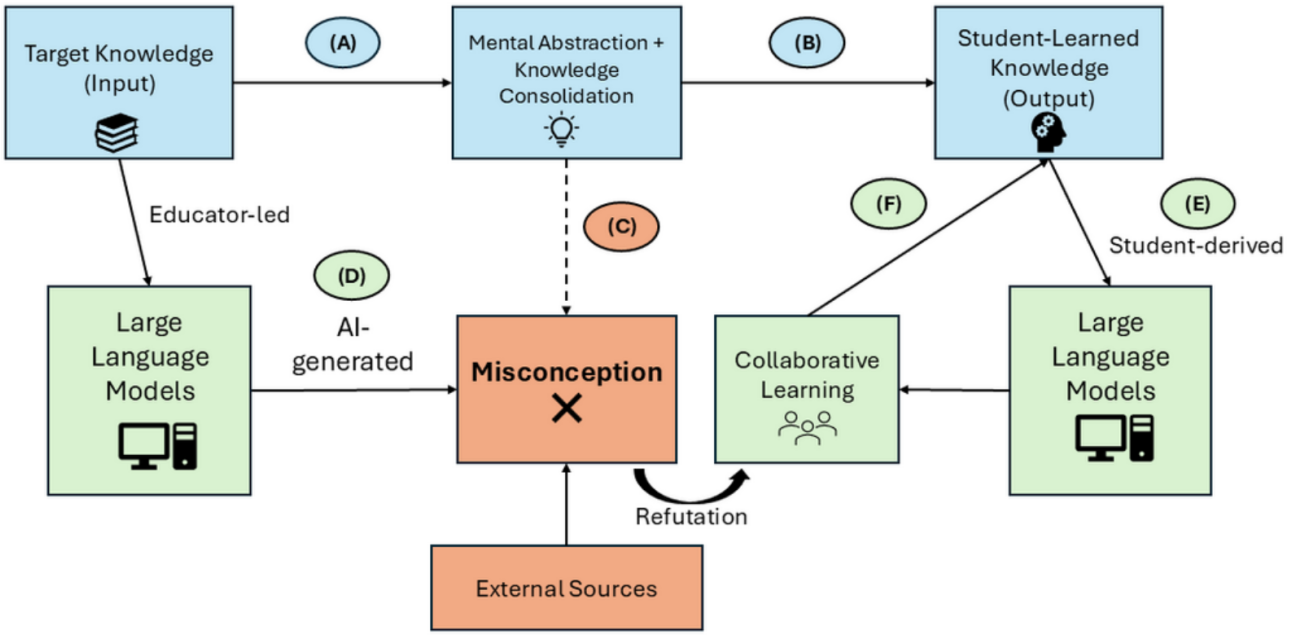
Ideally, the iterative process of conceptual acquisition is factually accurate, with the student utilizing prior learned knowledge to acquire new concepts, a core component of critical thinking (labeled (B) in Figure 1). However, a student is likely to develop misconceptions due to a continuously evolving, plausible intuitive interpretation of the concept according to lived experiences (denoted by the dotted arrow labeled (C) in Figure 1) [15]. Misconceptions can interfere or inhibit with the further process of mental abstraction, given that the student is likely to form a false belief that the concept learned is accurate, resulting in a persistent misconception.

One of the ways to manage misconceptions arising from the process of scientific conceptual acquisition is to pre-emptively expose students to misconceptions in the classroom and systematically debunk such misconceptions. A method to which misconception-based learning can be applied in practice is using refutation text. Refutation text is a misconception-centric pedagogical method first rooted in theories of conceptual accommodation in science [16]. Conceptual accommodation is the internalization of new concepts upon facing limitations in utilizing their existing conceptions to solve problems [17]. Conceptual accommodation, therefore, is the goal of refutation text and misconception-based learning.

Three components work together to form the refutation text pedagogical method: (1) stating common but inaccurate knowledge held by the student; (2) explicit indication of the incorrect aspects of said knowledge; and (3) stating factually correct information with supporting explanation and applications [18]. The advantages of refutation text over traditional text or didactic instruction include explicit mention of assumed prior incorrect beliefs, and increased propensity for conceptual acquisition even for students with low prior topic knowledge [19]. Given that students may find it challenging to state their own misconceptions held, the educator would be the primary source of providing the refutation text. Misconception exposure and the scientific method have already been in practice within the National University of Singapore common curriculum [20].

XSL•FO
RenderX

Figure 1. A framework for LLM-augmented misconception-based collaborative learning. LLM: large language model.



Integrating LLMs: An LLM-Augmented Misconception-Based Collaborative Learning Approach

There are two potential areas for LLM augmentation: educator-led or student-derived inputs of misconceptions into LLMs (denoted by the parts labeled (D) and (E), respectively, in Figure 1). In the educator-led case, LLMs can be used to simulate common student misconceptions. Educators can input a defined set of assessment-relevant concepts and prompt the model to generate plausible errors, providing a basis for targeted instruction. Techniques in prompt engineering include few-shot prompting [21], which is the provision of additional examples of desired tasks by the user in follow-up prompts. Given that LLMs predict appropriate outputs based on patterns in training data, providing structured follow-up prompts such as through few-shot prompting can help refine the scope and relevance of these generated misconceptions. For more complex cognitive processes, such as clinical reasoning, advanced strategies like chain-of-thought prompting—which simulates step-by-step heuristic thinking—can be employed to trace and identify potential misconceptions within the reasoning pathways of clinicians-in-training [21].

In the student-derived case, LLMs can drive deeper engagement between students and the course material. Prior student examination answers, or reference texts for the concepts, could be supplied to an LLM along with a pre-defined prompt to role-play as a junior trainee. The students could then begin a Socratic-style conversation with the LLM, identifying misconceptions based on the LLM's questions. Afterwards, the students could collaboratively correct the misconception and explain the correct understanding to the LLM. Such dialogic engagement could encourage metacognitive awareness, sustain curiosity due to the personalized nature of responses, and foster deeper conceptual understanding.

LLM-generated misconceptions complement student misconceptions revealed through assessment responses. Importantly, however, LLM-generated misconceptions may not be universally applicable across all student populations or educational contexts. As such, the lecturer's role includes adapting LLM-generated content to suit specific curricular needs, positioning the LLM as a quasi-teaching assistant in the generation of discussion prompts.

Overall, this framework in Figure 1 outlines the approach of LLM-augmented, misconception-based collaborative learning (CL). In addition to saving educator time, reducing reliance on educator experience, and identifying idiosyncratic student misconceptions, LLMs can also drive deeper engagement between students and the learning material. Educators could then shift their focus toward fostering higher order thinking skills aligned with the upper levels of Bloom's Taxonomy, though the extent of this shift will depend on the course's learning objectives and institutional priorities. Additionally, student-led presentation on evaluating misconceptions generated by AI can integrate into existing pedagogical approaches on the flipped classroom, encouraging student autonomy. Ultimately, the flexibility in this framework's extent of implementation potentially allows for a variety of use cases.

The Utility of CL in LLM-Assisted Misconception-Based Learning

Refuting misconceptions generated by LLMs could be achieved in CL settings, such as tutorials or workshops. CL emphasizes consensus building and collective problem-solving within a peer group, in contrast to individualistic or competitive approaches. By shifting learners from passive to active modes of engagement, CL promotes critical thinking, deeper conceptual understanding, and the development of social support networks [22].

XSL•FO
RenderX

In the context of addressing misconceptions, CL provides a structured environment for students to identify, articulate, and refute conceptual errors. A three-step approach can be used to scaffold this process: (1) students first work individually to pinpoint and justify any conceptual flaws within a statement; (2) small groups are then allocated time to discuss and contrast these ideas with factually accurate knowledge, refining their reasoning through peer interaction; and (3) groups present their consensus and rationale to the larger class, facilitating the synthesis and reinforcement of accurate understanding. After the group presentation, educators could identify strengths and weaknesses in the student groups' arguments. Additionally, LLMs could provide supplemental feedback alongside educators' comments on the group's quality of arguments. CL can thus have positive feedback on student-learned knowledge and conceptual retention (denoted by the part labeled (F) in Figure 1).

CL can present a risk of unmoderated reinforcement of misconceptions in LLM-supported environments through distractor information. Distractor information generated by LLMs adds unnecessary extraneous cognitive load to a student's working memory. This generates a quasi-redundancy effect that involves processing unnecessary or misleading information, possibly reducing learning efficiency [23].

However, the presence of CL here can be a crucial first line "guardrail" to the accuracy of knowledge. Through peer discussion, learners can identify and correct hallucination errors, making CL an essential rather than peripheral component of the learning process. CL has shown to be resource efficient in digitally enabled classroom settings when addressing misconceptions [24]. Educators could then review the understanding of the student group, acting as a second "guardrail" to address problematic misconceptions. Educators could further explore initiatives to scaffold understanding through diagnosis of student understanding and reinforcing understanding within CL groups with learning checkpoints [25].

Operational Considerations and Obstacles for Implementation

A key challenge in applying LLMs to misconception-based CL lies in data quality. LLMs learn from data patterns, and their performance depends critically on the quality and reliability of training materials [26]. Low-quality or inappropriate data can result in hallucinations or the generation of misconceptions that misrepresent core concepts. To mitigate this, domain-specific LLM architectures could be trained to draw exclusively from vetted, purpose-built educational resources such as notes, textbooks, or slides. A cost-efficient alternative for developing domain-specific LLMs is to input relevant educational resources into general-purpose models, such as GPT-4o, prior to prompt generation—though this approach may yield less domain-specific or contextually relevant outputs.

Providing high-quality data for LLM training is especially vital in health care education, where complex competencies such as clinical reasoning and clinical history taking have significant patient safety implications. In these contexts, custom LLM pipelines that utilize high-quality data—including pre-existing manual diagnostic frameworks currently in use by clinicians—are required. However, given present concerns on data security and ethical use, such data must be rigorously vetted for content, suitability, and compliance with privacy regulations before it is used to support learning and assessment. Privacy breaches in AI-assisted learning environments could compromise patient confidentiality and institutional trust in patient data security. Any adoption of LLMs in medical curricula must therefore follow the same rigor as clinical data governance.

Additionally, given that there is an absence of gold standards for clinical reasoning pipelines, it may be challenging in practice to develop an AI model specifically for clinical reasoning. Such pipelines of reasoning could be obtained from experienced clinicians for database construction, which would lead to greater quality of input data for LLMs to generate misconceptions for use in education.

While LLMs can generate customized feedback, pitfalls of LLMs in education include its potential to mislead students with false answers under the pretense of conversational and easy-to-understand responses [27]. However, a learning opportunity exists amidst the discourse, that is, how might we, as educators (1) utilize such misconceptions as a teaching opportunity, and (2) demonstrate the value of critical thinking and source evaluation in the process? A shift in the paradigm of conceptual acquisition can be achieved by seeing LLMs not as a source of harm, but as a pedagogical opportunity. This is pertinent in the health care education context, as hallucinated misinformation generated by LLMs may reinforce clinically unsafe assumptions if students are not trained to critically verify references.

Recent advances in ChatGPT-4o functionality include a deep research mode, which attempts to provide source references for the content it generates. However, because many publications remain behind paywalls and inaccessible to LLMs, the tool may still produce inaccurate citations or "hallucinated" references despite this feature. This limitation presents an opportunity for educators to help students critically evaluate the credibility of retrieved materials, distinguishing peer-reviewed from non-peer-reviewed sources and identifying instances of fabricated references.

Use Cases for LLM-Assisted Misconception Generation

Educators seeking to use LLMs such as ChatGPT for generating misconceptions can employ role-play within a single prompt, followed by targeted clarification prompts to leverage the few-shot learning capabilities of LLMs. Table 1 illustrates example use cases across disciplines including pharmacy, nursing, medicine, molecular biology, and microbiology, highlighting the potential application of this approach to both the practice-oriented and basic science domains of health care education. Table 1 only displays the initial prompt and excludes any subsequent prompts for elaboration based on the individual educator's expertise and use needs.

XSL•FO
RenderX

Table . Example use cases generated by GPT-4o for large language model (LLM)–assisted misconception-based collaborative learning.

Discipline	Concept(s)	Prompt	Learning goal	ChatGPT-4o output
Pharmacy	Hepatic extraction ratio	I am an educator within the pharmacy department of my university. I would like to teach the concept of hepatic extraction ratio. Come up with likely misconceptions that students may possess.	• Differentiate between pharmacokinetic parameters of hepatic extraction ratio and hepatic clearance.	• Misconception about hepatic extraction ratio: "The hepatic extraction ratio is the same as hepatic clearance." • Why it's wrong: Hepatic clearance depends on both hepatic blood flow and extraction ratio, while the hepatic extraction ratio (HER) is only the fraction of drug removed from the blood during a single pass through the liver. So, clearance can vary even if HER is constant, and vice versa.

XSL•FO
RenderX

Discipline	Concept(s)	Prompt	Learning goal	ChatGPT-4o output

Nursing	High-flow oxygen therapy	I am a nurse educator in my university's school of medicine. I am trying to convey that not all patients benefit from high-flow oxygen therapy, and that post-operation patients should be given fluid intake. Come up with likely misconceptions that students may have about this concept.	• Apply clinical reasoning to determine when high-flow oxygen is indicated or contraindicated.
			• Explain the effects of the lack of fluid intake post-surgery.
				• Common misconception about high-flow oxygen therapy: "All patients with low SpO_2[a] need high-flow oxygen."
				  • Why it's wrong: Not all hypoxic patients benefit from aggressive oxygen therapy (eg, patients with COPD[b] at risk for hypercapnia). Reinforce understanding by highlighting clinical scenarios where high-flow oxygen is beneficial (eg, severe hypoxia or ARDS[c]) versus when it can be detrimental (eg, certain COPD exacerbations).
				• Common misconception about fluid intake in post-operation patients: "Patients should remain NPO[d] for as long as possible post-surgery."
				  • Why it's wrong: Prolonged NPO status can increase the risk of ileus, dehydration, and delays in recovery. Early feeding (as tolerated) is now recommended in many ERAS (enhanced recovery after surgery) protocols.

Medicine	Dengue immunity and waning antibody titres	I am a clinician educator in my university hospital. I am delivering a lecture to a group of medical students. I would like to convey the concept of waning immunity and the fact that having dengue once does not guarantee long-term immunity. Provide some common student misconceptions.	• Explain serotype-specific immunity and the effect of waning antibody titres in dengue reinfection.	

XSL•FO
RenderX

Discipline	Concept(s)	Prompt	Learning goal	ChatGPT-4o output

				• Common misconception on dengue immunity: "Having dengue once guarantees long-term immunity against it."
				  • Why it's wrong: Dengue is caused by four distinct serotypes (DENV1, DENV2, DENV3, and DENV4). Infection with one serotype only provides long-term immunity to that specific serotype. It does not confer complete protection against the others. Vaccines must balance immunity across all four serotypes.
				• Common misconception on dengue immunity: "Once infected, dengue antibodies stay strong forever."
				  • Why it's wrong: Antibody levels wane over time, reducing neutralization and potentially allowing more severe disease upon reinfection.

Molecular Biology	DNA replication	I am teaching an introductory biology class to freshmen in Pharmacy. I am trying to convey the concept that DNA replication begins at multiple points of origin and DNA Polymerase uses dNTPs[e] for DNA replication. Come up with likely misconceptions that students may have about these two concepts.	• Explain that DNA replication begins at multiple points. • Explain how dNTPs provide energy for phosphodiester bond formation during DNA synthesis.	

XSL•FO
RenderX

Discipline	Concept(s)	Prompt	Learning goal	ChatGPT-4o output

				• Common misconception about DNA replication at multiple points (in eukaryotes): "Replication starts at only one point, like in bacteria."
				  • Why it's wrong: Students often generalize what they learn from prokaryotic replication (which starts at a single origin) to eukaryotic systems.
				• Common misconception about dNTPs in DNA replication: "The dNTPs are added in their final form directly into DNA."
				  • Why it's wrong: Students may not understand that the energy for bond formation comes from the triphosphate tail of the dNTP, which is hydrolyzed during incorporation.

Microbiology	Viral spillover effects in humans, and viral evolution	I am an educator within my university's microbiology and immunology department. I am delivering a lecture on SARS-CoV-2 spillover effects and viral evolution. Generate a list of plausible student misconceptions that I can debunk and explain.	• Explain how evolutionary pressures might shape viral transmissibility without assuming linear directionality.
			• Describe viral mutation as a process of random mutation.
				• Common misconception about SARS-CoV-2 spillover: "Viruses always evolve to become weaker over time."
				  • Why it's wrong: Evolution is not directional; some variants evolve for higher transmissibility, some for immune escape, and others may evolve increased or decreased virulence.
				• Common misconception about viral evolution: "Viruses evolve intentionally to bypass immunity or treatments."
				  • Why it's wrong: Evolution is a process of random mutations and natural selection — there's no 'intent' behind it.

XSL•FO
RenderX

[a]SpO2: oxygen saturation.

[b]COPD: chronic obstructive pulmonary disease.

[c]ARDS: acute respiratory distress syndrome.

[d]NPO: nil per os (nothing by mouth).

[e]dNTP: deoxynucleotide triphosphate.

Of interest are the inherent assumptions made by ChatGPT about a student's prior knowledge in the subject. For example, in the molecular biology misconception on DNA replication (Table 1), ChatGPT assumes that students in an introductory molecular biology class can contrast DNA replication between prokaryotes and eukaryotes, despite the prompt only indicating misconceptions on multiple points of DNA replication. Such an assumption may not hold true in an average cohort. This may indicate a form of metacognitive hindsight bias intrinsic within GPT-related architectures. Educators could then include clarification prompts on the existing knowledge base of the target audience and modify AI-generated misconceptions to suit teaching needs.

In addition, educators can draw students' attention to instances of hasty generalization within LLM-generated misconceptions such as in Table 1. For example, absolute qualifiers such as "all" (nursing prompt), "guarantees" (medical prompt), or "always" (microbiology prompt) imply universal causation with no exceptions. By explicitly highlighting these terms, educators can introduce counterexamples that challenge such overgeneralizations, providing a method for debunking misconceptions.

Finally, educators with prior experience of running past editions of the course in previous years could provide students' examination answers as an input for LLM analysis, after anonymizing student answer scripts. This could clarify the scope and customize the misconceptions generated to a specific educator's need. However, given excess specific input, the LLM runs a risk of factoring in irrelevant information, potentially leading to erroneous outputs. Educators could potentially mitigate this by limiting excessive irrelevant information and vetting the input contents to prevent overloading of the LLM's context window.

A 10-Step Guide for Implementing LLM-Augmented Misconception-Based CL

For educators who intend on implementing the framework in Figure 1 in practice, the following is a suggested 10-step guide with the concept of serotype-specific immunity in Table 1 used as an example.

Educator-Led Steps

Step 1: Setup

Input lecture slides and/or reference textbook chapters into the LLM. Include specifics of your role (eg, clinician educator) and audience (eg, medical students) in the prompt.

In this step, educators should input lecture slides and/or reference textbook chapters into the LLM. Include specifics of your role (eg, clinician educator) and audience (eg, medical students) in the prompt.

Step 2: Review

Educators review the misconceptions provided. LLMs such as ChatGPT-4o may occasionally provide misconceptions based on content outside the scope of coverage.

Step 3: Tailor

Educators tailor the subsequent prompts to vary the length of misconceptions provided by the LLM. For example, if the concepts covered pertain to the nature of serotype-specific immunity and its relationship to long-term immunity such as antibody-dependent enhancement, a prompt instructing the LLM to integrate these concepts into a single misconception statement may be necessary. The integration of multiple related concepts into one misconception can test students' attention to detail.

Student-Derived Steps

Step 4: Setup

Upload anonymized student examination answers and lecture notes into an institutional pre-trained LLM chatbot.

Step 5: Role-play

Prompt the chatbot to: (1) assume the role of a junior medical student, (2) engage in a Socratic-style conversational dialogue, (3) ask questions individually and reply to responses with a follow-up question, and (4) provide some misconceptions within its train of thought based on the uploaded student examination answers.

Alternatively, if institutional chatbots are not available, ChatGPT-4o may be used. Only input lecture notes and/or reference textbook chapters and not student examination answers would be used in this case, as sending past student examination answers to current students may potentially infringe upon data privacy.

Step 6: Group

Depending on the cohort size and availability of educators, split the students into groups of 3‐5 to discuss the generated misconceptions for CL purposes. CL should preferably be implemented after students have gone through the content at least once, either through pre-reading or classroom instruction.

Step 7: Present

Student groups should conduct short presentations to identify and explain the LLM-generated misconceptions.

Step 8: Converse

Instruct the students to access the prompted institutional chatbot or ChatGPT-4o and initiate the conversation.

XSL•FO
RenderX

Step 9: Collaborate

Allow the students to work collaboratively in groups to address the questions posed by the LLM and "teach" it the concepts. Students could download the chat history or transcripts after the conclusion of the "lesson."

Step 10: Assess

Students' misconceptions could be assessed afterwards by reviewing their responses within the downloaded chat transcripts.

Limitations and Directions for AI-Assisted Pedagogy in Health Care Education Studies

It should be noted that only 2 quantitative studies [28,29] have been conducted on misconception-based reasoning and the wider misconception-based learning literature within the context of health care education. More than half of these studies are focused on the fields of Newtonian physics, genetically modified organisms, and evolution [30]. While misconception-based learning was determined to be robust across disciplines independent of specific text design [31], generalizing misconception-based learning's efficacy from other fields to health care education may not be fully applicable. More studies are required to examine the impact of this pedagogical method on student outcomes in health care education.

A common methodological critique of this pedagogical approach is its vulnerability to pre-test priming effects. In typical designs, students complete a pre-test before exposure to the instructional intervention and a post-test afterward. However, evidence suggests that notifying students about an impending test may activate prior knowledge and trigger conceptual change, making it challenging to distinguish between the effects of pre-test priming and those of LLM-assisted, misconception-based CL. To address this, researchers could track individual student responses and apply statistical controls—such as paired *t* tests or ANCOVA analyses—to adjust for pretest influences and better isolate the intervention's true impact.

CL applied in LLM-generated misconception generation assumes that there is a form of homogeneity in conceptual understanding within and between groups. In a group comprised only of weaker students, false explanations of the LLM-generated misconceptions may be negatively reinforced instead. While ensuring the distribution of students with varying backgrounds and degrees of conceptual understanding within groups may be ideal to fully capitalize on the benefits of CL, such methods are difficult to implement in practice and are a key limitation of this pedagogical method.

Given the limited number of studies present to examine AI-assisted pedagogy in health care education, multiple avenues for follow-up research studies exist. Few studies have assessed cognitive attitudes and student perspectives on CL using AI [32]. Importantly, the critical nature of such studies stems from the increasingly interconnected working environment. Studying cognitive attitudes on CL using AI from students' perspectives could better associate AI technology implementation in the classroom with student outcomes.

Additionally, there has been a lack of medium- to long-term research on student outcomes using AI [33]. The prevalence of AI technologies and its impact on the educational setting have far-reaching implications for changing pedagogy and student learning. However, since the rise of LLMs to prominence in 2023, studies have largely been cross-sectional, capturing only a snapshot of AI's influence at a specific point in its rapid development.

Cross-sectional studies are likely only able to capture a snapshot of AI's impact on student learning at a specific point of time in AI's development. Such studies often overlook how a student's learning trajectory evolves alongside advances in AI itself. For example, an educational experience supported by GPT-3 would likely differ substantially from one shaped by GPT-4o, which introduced capabilities such as image generation. Longitudinal, mixed-methods studies are needed to track these evolving dynamics, providing deeper insights into both shifts in student perceptions and longer-term learning outcomes as AI technologies continue to evolve. Such an approach could also assist to identify key factors that support or hinder the long-term implementation of AI in teaching practice.

Yet, it should be noted that the fast-evolving nature of LLMs—which at the time of writing include Gemini 2.5 and the upcoming GPT-5 by OpenAI—could impact the reproducibility of generated misconceptions. With larger token sizes and greater architecture complexity, prompts provided to newer models may not produce the same outputs. Future educational research involving LLMs could ensure the reporting of the specific model version used, sampling conditions, and prompts used for an effort towards greater reproducibility. While the stochastic nature of LLMs carries a risk of low reproducibility of outputs with the same prompt, the role of educators to mitigate such a risk could be to verify and vet the LLM-generated outputs. Thus, LLM-generated misconceptions should serve to complement curriculum design, and the educator should not be removed from the process.

Additionally, greater data privacy risk comes with the evolving dynamics of new LLM models. The extent to which newer models retain data inputs by users through larger context windows are not well understood. Therefore, such risks could be mitigated by educators through anonymization and the removal of original file metadata. On the end of educational technology developers, methods to protect data security such as adversarial attack testing or encryption specific to LLM modalities could be explored [34].

Finally, explicit indications of the misconception and subsequent explanations by educators may not always work if the misconception is entrenched in the students' mental models. Greater emphasis needs to be placed on understanding the root of the falsehood, perhaps through analysis of past lived experiences of students that led to this network of beliefs, to facilitate effective conceptual change. Addressing misconceptions in education would therefore likely not be a one-size-fits-all solution. Supplementing refutation-based learning with other educational approaches such as concept

XSL•FO
RenderX

maps and audiovisual representation may improve coherence of concept representation [29].

Conclusion – LLM's Potential for Augmenting Misconception-Based Learning in Health Care Education

Misconceptions, which could stem from naive misinformation, have implications on critical thinking and student learning. Yet, when leveraged intentionally—for example, through LLM-augmented, misconception-based CL—they can catalyze transformative shifts in teaching practice. This approach not only deepens students' understanding of subject matter but also provides a valuable teaching opportunity for exploring both the benefits and limitations of LLMs.

Additionally, the framework serves as a springboard for future empirical research, generating several potentially testable hypotheses. For example, does pre-bunking with LLM-generated misconceptions elicit stronger propensity for conceptual accommodation as compared to educator-generated refutation text? Does CL around LLM-generated misconceptions improve reasoning accuracy and quality compared to individual learning? Does LLM-generated misconceptions change classroom dynamics for CL? These are important questions for identifying and testing the value of LLMs in misconception-based CL.

However, the true value of LLMs in misconception-based learning exists not to replace educator involvement and expertise. Rather, its purpose is to augment student learning with personalized feedback, synthesize "edge-case" concepts across topics, and reduce the time taken for educators by supplementing misconceptions. The educator's role is still important in context-setting, the facilitation of CL, and critical review of the AI output. Ultimately, the proposed framework positions LLMs not as an endpoint for teaching and learning, but to foster self-directed, critical, and reflective thinking.

Acknowledgments

We used the generative AI tool ChatGPT by OpenAI to generate examples of use cases in Table 1. Additionally, ChatGPT was used to refine the grammatical structure of the main text to enhance clarity and readability. We would like to thank Dr Ronald Boyd Anderson (National University of Singapore) for his kind comments to improve the paper on the concepts surrounding LLMs and reproducibility.

Funding

This work was funded by the National University of Singapore Startup Grant (NUHSRO/2023/018/Startup/10).

Data Availability

No original data were generated or collected for the purpose of this paper.

Authors' Contributions

Conceptualization: BCJC (lead), JHC (supporting), CWT (supporting)
Investigation: BCJC
Methodology: BCJC
Evidence curation: BCJC
Writing – Original draft: BCJC
Writing – Review and editing: BCJC (lead), JHC (supporting), CWT (supporting), SS (supporting)
Supervision: JHC (equal), CWT (equal), SS (supporting)
Funding acquisition: CWT

Conflicts of Interest

None declared.

References

1. Tandar CE, Lin JC, Stanford FC. Combating medical misinformation and rebuilding trust in the USA. Lancet Digit Health 2024 Nov;6(11):e773-e774. [doi: 10.1016/S2589-7500(24)00197-3] [Medline: 39379267]
2. Rubin R. When physicians spread unscientific information about COVID-19. JAMA 2022 Mar 8;327(10):904. [doi: 10.1001/jama.2022.1083]
3. Denniss E, Lindberg R. Social media and the spread of misinformation: infectious and a threat to public health. Health Promot Int 2025 Mar 5;40(2):daaf023. [doi: 10.1093/heapro/daaf023] [Medline: 40159949]
4. Olde Bekkink M, Donders A, Kooloos JG, de Waal RMW, Ruiter DJ. Uncovering students' misconceptions by assessment of their written questions. BMC Med Educ 2016 Aug 24;16(1):221. [doi: 10.1186/s12909-016-0739-5] [Medline: 27558400]
5. Bordes SJ, Manyevitch R, Huntley JD, Li Y, Murray IVJ. Medical student misconceptions in cardiovascular physiology. Adv Physiol Educ 2021 Jun 1;45(2):241-249. [doi: 10.1152/advan.00220.2020] [Medline: 33825522]

XSL•FO
RenderX

6. Al-Zahrani AM, Alasmari TM. Exploring the impact of artificial intelligence on higher education: the dynamics of ethical, social, and educational implications. Humanit Soc Sci Commun 2024 Dec 1;11(1). [doi: 10.1057/s41599-024-03432-4]
7. Britt MA, Rouet JF, Blaum D, Millis K. A reasoned approach to dealing with fake news. Policy Insights Behav Brain Sci 2019 Mar;6(1):94-101. [doi: 10.1177/2372732218814855]
8. List JA, Ramirez LM, Seither J, Unda J, Vallejo BH. Critical thinking and misinformation vulnerability: experimental evidence from Colombia. PNAS Nexus 2024 Oct;3(10):pgae361. [doi: 10.1093/pnasnexus/pgae361] [Medline: 39411082]
9. Leonard MJ, Kalinowski ST, Andrews TC. Misconceptions yesterday, today, and tomorrow. LSE 2014 Jun;13(2):179-186. [doi: 10.1187/cbe.13-12-0244]
10. Rowell JA, Dawson CJ, Lyndon H. Changing misconceptions: a challenge to science educators. Int J Sci Educ 1990 Apr;12(2):167-175. [doi: 10.1080/0950069900120205]
11. Boshuizen HPA, Marambe KN. Misconceptions in medicine, their origin and development in education and working life. Int J Educ Res 2020;100:101536. [doi: 10.1016/j.ijer.2020.101536]
12. Södervik I, Mikkilä-Erdmann M, Chi MTH. Conceptual change challenges in medicine during professional development. Int J Educ Res 2019;98:159-170. [doi: 10.1016/j.ijer.2019.07.003]
13. Yee E. Abstraction and concepts: when, how, where, what and why? Lang Cogn Neurosci 2019 Nov 26;34(10):1257-1265. [doi: 10.1080/23273798.2019.1660797]
14. diSessa AA. Toward an epistemology of physics. Cogn Instr 1993 Apr;10(2-3):105-225. [doi: 10.1080/07370008.1985.9649008]
15. Vosniadou S, Ioannides C. From conceptual development to science education: a psychological point of view. Int J Sci Educ 1998 Dec;20(10):1213-1230. [doi: 10.1080/0950069980201004]
16. Posner GJ, Strike KA, Hewson PW, Gertzog WA. Accommodation of a scientific conception: toward a theory of conceptual change. Sci Educ 1982 Apr;66(2):211-227. [doi: 10.1002/sce.3730660207]
17. Nussbaum J, Novick S. Alternative frameworks, conceptual conflict and accommodation: toward a principled teaching strategy. Instr Sci 1982 Dec;11(3):183-200. [doi: 10.1007/BF00414279]
18. Danielson RW, Jacobson NG, Patall EA, et al. The effectiveness of refutation text in confronting scientific misconceptions: a meta-analysis. Educ Psychol 2025 Jan 2;60(1):23-47. [doi: 10.1080/00461520.2024.2365628]
19. Hynd CR. Refutational texts and the change process. Int J Educ Res 2001 Jan;35(7-8):699-714. [doi: 10.1016/S0883-0355(02)00010-1]
20. National University of Singapore, Faculty of Science. Stay curious, question everything: ‘Citizen scientists’ explore how and why science works. 2024. URL: https://www.science.nus.edu.sg/blog/2024/04/innovative-educators-assoc-prof-ryan-bettens/ [accessed 2025-07-09]
21. Chen B, Zhang Z, Langrené N, Zhu S. Unleashing the potential of prompt engineering for large language models. Patterns (N Y) 2025 Jun 13;6(6):101260. [doi: 10.1016/j.patter.2025.101260] [Medline: 40575123]
22. Laal M, Ghodsi SM. Benefits of collaborative learning. Procedia - Social and Behavioral Sciences 2012;31:486-490. [doi: 10.1016/j.sbspro.2011.12.091]
23. Sweller J. Cognition in education. In: Ross BH, Mestre JP, editors. Psychol Learn Motiv: Elsevier; 2011:313 URL: https://shop.elsevier.com/books/cognition-in-education/mestre/978-0-12-387691-1 [accessed 2026-01-05]
24. Jansen M, Kim Cavdar I. Adapting team-based learning for medical education: a case study with scalable and resource-efficient implementation. MedSciEduc ;35(2):883-892. [doi: 10.1007/s40670-024-02246-y]
25. van de Pol J, Mercer N, Volman M. Scaffolding student understanding in small-group work: students’ uptake of teacher support in subsequent small-group interaction. Journal of the Learning Sciences 2019 Mar 15;28(2):206-239. [doi: 10.1080/10508406.2018.1522258]
26. Cheah BCJ, Vicente CR, Chan KR. Machine learning and artificial intelligence for infectious disease surveillance, diagnosis, and prognosis. Viruses 2025 Jun 23;17(7):882. [doi: 10.3390/v17070882] [Medline: 40733500]
27. Li R, Wu T. Delving into the practical applications and pitfalls of large language models in medical education: narrative review. Adv Med Educ Pract 2025;16:625-636. [doi: 10.2147/AMEP.S497020] [Medline: 40271151]
28. Holford D, Mäki KO, Karlsson LC, Lewandowsky S, Gould VC, Soveri A. A randomized controlled trial of empathetic refutational learning with health care professionals. BMC Public Health 2025 Feb 12;25(1):583. [doi: 10.1186/s12889-025-21787-4] [Medline: 39939961]
29. Versteeg M, van Loon MH, Wijnen-Meijer M, Steendijk P. Refuting misconceptions in medical physiology. BMC Med Educ 2020 Aug 5;20(1):250. [doi: 10.1186/s12909-020-02166-6] [Medline: 32758215]
30. Zengilowski A, Schuetze BA, Nash BL, Schallert DL. A critical review of the refutation text literature: methodological confounds, theoretical problems, and possible solutions. Educ Psychol 2021 Jul 3;56(3):175-195. [doi: 10.1080/00461520.2020.1861948]
31. Schroeder NL, Kucera AC. Refutation text facilitates learning: a meta-analysis of between-subjects experiments. Educ Psychol Rev 2022;34(2):957-987. [doi: 10.1007/s10648-021-09656-z] [Medline: 35095236]
32. Yu JH, Chauhan D, Iqbal RA, Yeoh E. Mapping academic perspectives on AI in education: trends, challenges, and sentiments in educational research (2018–2024). Education Tech Research Dev 2025 Feb;73(1):199-227. [doi: 10.1007/s11423-024-10425-2]

XSL•FO
RenderX

33. Adiguzel T, Kaya MH, Cansu FK. Revolutionizing education with AI: exploring the transformative potential of ChatGPT. CONT ED TECHNOLOGY 2023;15(3):ep429. [doi: 10.30935/cedtech/13152]
34. Yan B, Li K, Xu M, et al. On protecting the data privacy of large language models (llms): a survey. 2024 Mar 14 Presented at: 2024 International Conference on Meta Computing (ICMC). [doi: 10.1109/ICMC60390.2024.00008]

Abbreviations

AI: artificial intelligence
CL: collaborative learning
LLM: large language model
MBC: misconception-based collaborative

Edited by S Tsuei; submitted 09.Aug.2025; peer-reviewed by A Alhur, JWY Lee, MNH Mohamad SaidnAff; revised version received 05.Nov.2025; accepted 06.Nov.2025; published 16.Jan.2026.

Please cite as:
Cheah BCJ, Shorey S, Ch'ng JH, Tan CW
Implementing Large Language Models to Support Misconception-Based Collaborative Learning in Health Care Education
JMIR Med Educ 2026;12:e81875
URL: https://mededu.jmir.org/2026/1/e81875
doi:10.2196/81875

© Brandon C J Cheah, Shefaly Shorey, Jun Hong Ch'ng, Chee Wah Tan. Originally published in JMIR Medical Education (https://mededu.jmir.org), 16.Jan.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Education, is properly cited. The complete bibliographic information, a link to the original publication on https://mededu.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Viewpoint

Preparing Physicians for Tomorrow's Practice: Contemporary Practice Competencies as a Framework for Technology-Driven Change

Joseph E Safdieh[1], MD; June Chan[1], MBBS; Domenick Falcone[1], PhD; Erika L Abramson[1], MD

Weill Cornell Medicine, New York, NY, United States

Corresponding Author:
Joseph E Safdieh, MD
Weill Cornell Medicine
1300 York Ave C118
New York, NY, 10065
Phone: 1 2127463113
Email: jos9046@med.cornell.edu

Abstract

Medical education has long relied on stable, high-level program objectives to articulate the outcomes of undergraduate medical training. These objectives have served an essential role in defining professional identity, guiding curricular design, and ensuring accountability. However, the pace of contemporary clinical change increasingly exceeds the capacity of static curricular structures to adapt. Expectations for how physicians practice, communicate, and exercise judgment, particularly in technology-mediated care environments, are evolving faster than traditional curricular governance mechanisms can accommodate. Recent accreditation developments, including proposed changes to the organization of required content domains, signal a broader inflection point: medical schools are being challenged to move beyond additive, topic-based approaches to curriculum design toward more durable and adaptive frameworks for preparing physicians for practice. This viewpoint argues that many of the tensions currently facing medical education reflect an overreliance on problem lists and topical coverage models that were designed for a slower era of clinical evolution. As new priorities emerge, ranging from artificial intelligence and data-driven decision support to prevention, climate health, and changing patient expectations, curricula risk becoming increasingly dense, fragmented, and difficult to govern. Simply elevating each new priority to the level of a program objective undermines the coherence and longevity of objective frameworks and politicizes curricular decision-making. To address this challenge, we propose contemporary practice competencies (CPCs) as a complementary, middle-layer construct positioned between medical educational program objectives and course- or clerkship-level objectives. CPCs are dynamic, revisable competencies designed to translate stable educational aims into forward-facing clinical capabilities, particularly in domains where technology and data-driven tools are reshaping clinical reasoning. Using artificial intelligence as a primary test case, we illustrate how CPCs enable longitudinal, assessable development of emerging practice capabilities without repeatedly restructuring high-level objectives. By providing a transparent and adaptable mechanism for curricular evolution, CPCs offer a practical approach to aligning medical education with the accelerating realities of contemporary clinical practice.

(JMIR Med Educ 2026;12:e90637) doi:10.2196/90637

KEYWORDS

medical education; artificial intelligence; competencies; undergraduate; curriculum; contemporary practice competencies

Overview

Medical school curricula have always evolved in response to forces beyond the lecture hall, such as emerging diseases, shifting demographics, scientific discoveries, and emerging therapeutics. What has changed is the pace of clinical evolution. Expectations for how physicians practice, communicate, and earn trust are shifting faster than the mechanisms used to guide curricular change, which often remain reliant on static content categories, infrequently revised objectives, and additive approaches to new priorities. Recent accreditation developments, including the removal of "societal problems" as a discrete curricular category by the Liaison Committee on Medical Education (LCME) [1], mark an important inflection point, challenging medical schools to move beyond continually expanding catalogs of required topics designed to react to problems.

XSL•FO
RenderX

One useful way to interpret this moment is as a shift from a problem-list paradigm to a practice-readiness paradigm. Historically, curricular requirements have been periodically expressed through topical expectations, what might be called "lists of importance." Many schools responded by creating explicit coverage maps, adding sessions, and, at times, promoting specific content areas into high-level educational program objectives. Over time, this approach becomes vulnerable to curricular inflation. New priorities compete for permanent placement, the curriculum grows denser, more fragmented, and less teachable. Educational leaders and learners recognize that overcrowded curriculum undermines deep learning, invites superficial coverage, and can compound stress and burnout [2].

The elimination or reorganization of content expectations previously grouped under "societal problems" offers an opportunity to recalibrate. This is not an argument that social and structural drivers of health no longer matter but a question of framing. A "societal problems" construct is inherently reactive, implying that curricula should follow emerging pathologies. Contemporary practice increasingly demands something different: the ability to anticipate, prevent, and redesign care. The aim of this viewpoint is to propose a conceptual framework, contemporary practice competencies (CPCs), to support the integration of evolving domains of clinical practice into medical education while preserving the stability of high-level program objectives.

This tension is particularly evident in areas now central to contemporary practice but awkward to incorporate as discrete topics, including artificial intelligence (AI) in clinical decision-making [3], climate change, nutrition in the prevention and management of chronic disease, and communication in an era of erosion of public trust in science. Among these, AI offers a particularly clear illustration of why traditional curricular structures struggle to keep pace with practice evolution.

Medical Education Program Objectives Should Remain Thematic, Not Topical

LCME-accredited medical schools already have a stable set of Medical Education Program Objectives (EPOs) that are intentionally broad, thematic, and durable. EPOs describe the physician the program aims to produce. Their strength lies in their longevity. Their limitation is that they can become too abstract to reliably drive curricular design when practice shifts rapidly.

This is where schools face a recurring temptation: if AI is important, make it an EPO; if nutrition is important, make it an EPO; if prevention is important, make it an EPO. But high-level program objectives are meant to describe enduring capacities, not the shifting topical priorities of a particular decade. Once objectives become topical, governance becomes politicized as every emerging priority becomes a candidate for canonization, and every specialty's "pet project" risks being framed as universally required. The result is not curricular coherence; it is curricular lobbying.

As a recent example, the integration of nutrition into medical education highlights this challenge. Recent consensus efforts have identified a broad set of nutrition competencies spanning foundational knowledge, clinical assessment, patient communication, public health, and interprofessional care. These competencies cut across multiple domains of physician performance and are expected to be developed longitudinally across training rather than confined to a single course or objective. As a result, nutrition resists clean alignment with any single EPO. Mapping it to one domain risks fragmentation, while distributing it across many risks dilution and loss of accountability. This illustrates a broader structural limitation: some contemporary practice capabilities are inherently cross-cutting and longitudinal and therefore difficult to operationalize within static, high-level objective frameworks alone [4].

A Missing Middle Layer: CPCs

We propose a complementary construct: CPCs, a dynamic, revisable set of competencies designed to sit between enduring EPOs and course- and clerkship-level objectives. CPCs are not replacements for EPOs. They function as an implementation layer, translating stable, thematic outcomes into forward-facing clinical capabilities that must be developed intentionally over time. Conceptually, CPCs serve as an organizing framework for evolving clinical practice priorities within established EPOs.

CPCs are defined by 4 design principles. First, they focus on capabilities physicians need in current and emerging clinical practice, rather than on static content areas. Second, they are linked to existing EPOs, ensuring that new domains are integrated without altering the overall structure of the objective framework. Third, CPCs are developed across the curriculum over time, rather than introduced as isolated topics or single sessions. Finally, CPCs are assessed through observable clinical work, using existing approaches such as entrustable activities and workplace-based assessments rather than relying solely on content coverage.

CPCs align with the logic of entrustable professional activities (EPAs), which were introduced to bridge the gap between abstract competency domains and the real work of clinical practice [5]. While EPAs are often framed around discrete units of clinical work that can be observed and assessed in context, the underlying insight is broader: learners and faculty need a practical, observable language for readiness that can be mapped back to higher-order outcomes. However, EPAs are generally stable and task-based and are not designed to accommodate rapidly evolving, cross-cutting domains of practice that must be integrated longitudinally across the curriculum. CPCs extend this logic by providing a dynamic, intermediate layer that translates enduring program objectives into evolving, cross-cutting capabilities (Table 1). In this way, CPCs provide a structured mechanism for incorporating domains such as nutrition, AI, and climate health across disciplines without elevating each to a distinct EPO.

XSL•FO
RenderX

Table 1. Conceptual relationship between educational program objectives, contemporary practice competencies, and entrustable professional activities, highlighting how CPCs function as an intermediate layer linking stable program objectives to observable clinical work.

Level	Purpose	Characteristics	Example
Education program objectives (EPOs)	Define the physician the program aims to produce	Broad, thematic, stable over time	Perform both a focused and comprehensive history and physical examination, develop diagnostic hypotheses, order and evaluate diagnostic tests, and formulate an appropriate plan of care.
Contemporary practice competencies	Translate EPOs into evolving, cross-cutting capabilities	Dynamic, revisable, longitudinal, linked to multiple EPOs	Integrate artificial intelligence–supported clinical information into diagnostic reasoning and patient-centered care planning across clinical contexts.
Entrustable professional activities	Define units of observable clinical work	Task-based, context-specific, assessable in practice	Prioritize a differential diagnosis following a clinical encounter.

A key distinction is that EPAs are designed to organize clinical work, not to define evolving domains of practice. While EPAs can incorporate new elements such as the use of AI within a diagnostic encounter, they do not provide a mechanism for identifying, updating, and longitudinally integrating new capabilities across the curriculum as practice evolves. In contrast, CPCs are explicitly designed to define these evolving domains and to guide their development over time. In this way, CPCs complement EPAs rather than duplicate them: CPCs define what capabilities are needed, while EPAs define how those capabilities are enacted in specific clinical tasks.

For example, in the domain of AI, a CPC may define the expectation that learners develop the ability to integrate AI-supported clinical information into diagnostic reasoning and patient-centered care across clinical contexts. This is not a single observable task, but a longitudinal capability that informs multiple EPAs, such as prioritizing a differential diagnosis or developing a management plan. In this way, CPCs define what should be developed across the curriculum, while EPAs and associated assessment approaches define how that capability is demonstrated and evaluated in specific clinical contexts.

Why AI Is an Ideal Test Case

AI highlights the limitations of topic-based curricular design and the advantages of a CPC framework. If elevated to an EPO, AI would immediately compete with genomics, climate health, nutrition, and other emerging priorities, destabilizing the objective layer. More fundamentally, AI evolves too rapidly to be governed as a static content domain. Models change, tools are updated, and clinical applications expand faster than curricular approval cycles can accommodate.

Framed as a CPC, AI becomes a longitudinal capability rather than a discrete topic. Learners can be expected to develop the ability to critically evaluate AI-supported clinical recommendations, recognize limitations and sources of bias, integrate algorithmic outputs into clinical reasoning, communicate uncertainty with patients, and maintain professional judgment in technology-mediated environments. These capabilities can be scaffolded across the curriculum and assessed through observable clinical work rather than episodic exposure to rapidly outdated content. As an example of implementation, AI-related competencies aligned with this framework have been developed and approved through institutional curricular governance and are provided in Multimedia Appendix 1.

Nutrition provides a useful parallel example [4]. When framed as a CPC rather than an EPO, nutrition functions as a longitudinal capability integrated into clinical reasoning, patient counseling, interprofessional care, and prevention across the curriculum. For example, at our institution, nutrition-related competencies are addressed across multiple phases of training, including foundational science, clinical skills instruction, and clerkship experiences, rather than being confined to a single course or objective. This distributed approach reflects the inherently cross-cutting nature of nutrition and illustrates the challenge of aligning it with any single EPO. Together, AI and nutrition demonstrate how CPCs allow schools to address fast-evolving practice domains without repeatedly restructuring high-level program objectives.

Implications for Curricular Governance and Accreditation

CPCs are intended to simplify, not expand, curricular structures. They are designed to be limited in number, periodically reviewed, and integrated across existing curricular structures, thereby reducing the need for additive content and minimizing fragmentation. In practice, they would be developed and periodically updated by existing curricular leadership bodies, such as curriculum committees or educational oversight groups, using the same governance processes already in place for curriculum review. Rather than creating new content blocks or standalone requirements, CPCs would be mapped across existing courses and clerkships, guiding where and how emerging domains are addressed longitudinally. Importantly, CPCs do not bypass existing governance structures. Because they are explicitly linked to established EPOs, they remain anchored to the core aims of the program and subject to the same oversight processes. This allows institutions to adapt to evolving domains of practice while maintaining alignment with accreditation expectations and overall curricular coherence.

CPCs differ from routine content review in that they define explicit, program-level expectations for emerging domains of practice. While content review may introduce new topics into individual courses, it does not ensure coordinated, longitudinal development across the curriculum. CPCs provide a structured approach to integrating these domains over time, making

XSL•FO
RenderX

expectations visible, assessable, and aligned with program objectives. For example, a CPC in AI would not require a new course, but would inform how AI-related capabilities are integrated into clinical skills training, diagnostic reasoning sessions, and clerkship experiences over time. In this way, CPCs organize existing curricular elements more coherently rather than adding new ones.

This approach reduces the need for repeated modification of EPOs while providing a clear and adaptable mechanism for incorporating emerging domains of practice. For accreditation, CPCs offer a transparent way to demonstrate how programs respond to evolving clinical expectations without relying on continual expansion of high-level objectives or fragmented topic-based additions.

Conclusion

The retirement or reorganization of "societal problems" should not be treated as a bureaucratic deletion. It can be interpreted as an opening to modernize how medical education defines and governs competencies. In a world where the pace of practice evolution is accelerating, curricula cannot be managed solely through reactive topic lists or through repeated expansion of high-level program objectives [6].

CPCs offer one such approach, translating enduring educational aims into the evolving realities of clinical work. They invite medical education to move from fixing yesterday's problems to preparing for tomorrow's practice intentionally, transparently, and with governance structures designed to keep pace. These challenges are especially acute in technology-mediated care environments.

Funding

No external financial support or grants were received from any public, commercial, or not-for-profit entities for the research, authorship, or publication of this article.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Weill Cornell Medicine contemporary practice competencies for artificial intelligence.
[XLSX File (Microsoft Excel File), 13 KB - mededu_v12i1e90637_app1.xlsx]

References

1. Proposed changes to standard 7. Liaison Committee on Medical Education. URL: https://lcme.org/lcme-strategic-vision/ [accessed 2025-12-19]
2. Slavin S, D'Eon MF. Overcrowded curriculum is an impediment to change (Part A). Can Med Educ J 2021 Sep;12(4):1-6 [FREE Full text] [doi: 10.36834/cmej.73532] [Medline: 34567300]
3. Artificial intelligence framework for medical educators. Association of American Medical Colleges. URL: https://www.aamc.org/about-us/mission-areas/medical-education/advancing-ai-resource-collection/artificial-intelligence-competencies-medical-educators [accessed 2025-12-18]
4. Eisenberg DM, Cole A, Maile EJ, Salt M, Armstrong E, Broad Leib E, et al. Proposed nutrition competencies for medical students and physician trainees: a consensus statement. JAMA Netw Open 2024 Sep 03;7(9):e2435425 [FREE Full text] [doi: 10.1001/jamanetworkopen.2024.35425] [Medline: 39348126]
5. Cate OT. A primer on entrustable professional activities. Korean J Med Educ 2018 Mar;30(1):1-10 [FREE Full text] [doi: 10.3946/kjme.2018.76] [Medline: 29510603]
6. The CanMEDS Framework. Royal College of Physicians and Surgeons of Canada. URL: https://www.royalcollege.ca/en/standards-and-accreditation/canmeds.html [accessed 2025-12-20]

Abbreviations

AI: artificial intelligence
CPCs: contemporary practice competencies
EPA: entrustable professional activity
EPO: education program objective
LCME: Liaison Committee on Medical Education

XSL•FO
RenderX

Edited by B Lesselroth; submitted 31.Dec.2025; peer-reviewed by K Wilson-Mitchell, C Janse van Vuuren, H Vundavalli; comments to author 27.Mar.2026; revised version received 29.Mar.2026; accepted 07.Apr.2026; published 17.Apr.2026.

Please cite as:
Safdieh JE, Chan J, Falcone D, Abramson EL
Preparing Physicians for Tomorrow's Practice: Contemporary Practice Competencies as a Framework for Technology-Driven Change
JMIR Med Educ 2026;12:e90637
URL: https://mededu.jmir.org/2026/1/e90637
doi:10.2196/90637
PMID:41945634

©Joseph E Safdieh, June Chan, Domenick Falcone, Erika L Abramson. Originally published in JMIR Medical Education (https://mededu.jmir.org), 17.Apr.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Education, is properly cited. The complete bibliographic information, a link to the original publication on https://mededu.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Viewpoint

Is This Still Medical Education? The Educational Use of Nude Pediatric Clinical Photographs: Bioethical Reassessment

Iwona Bujek[1,2], PhD

[1]Poznan University of Medical Sciences, Poznan, Greater Poland, Poland

[2]GGZ Voor Elkaar, Mental Health Care Institution, Utrecht, North Holland, The Netherlands

Corresponding Author:
Iwona Bujek, PhD
Poznan University of Medical Sciences
Bukowska 70
Poznan, Greater Poland, 3515 BW
Poland
Phone: 48 61 854 60
Email: iwona.bujek@gmail.com

Abstract

Clinical photography can convey phenotypic, anatomical, and diagnostic information that may be difficult to replace with text. This also applies to highly sensitive pediatric clinical photographs when the scope of patient exposure is necessary to achieve a specific educational objective. This viewpoint addresses a more specific problem: "How should the educational function of a highly sensitive pediatric clinical photograph be evaluated when the condition identified with the photograph is substantively described elsewhere in the educational material?" The analysis is grounded in a documented longitudinal observation of 28 editions and reprints of a single academic textbook lineage published between 2005 and 2019, across which the same image-condition configuration persisted. At the point of image presentation, the depicted clinical conditions were explicitly identified, while the accompanying text stated that they were described elsewhere in the textbook, where their substantive clinical descriptions were located in accordance with the textbook's declared curricular structure. On this basis, I propose educational disconnection to describe a situation in which the image-specific educational contribution at the point of presentation is difficult to reconstruct from the observable relationship among image location, the clinical condition identified with the image, the condition's declared curricular location, the location of its substantive clinical description, and the destination indicated in the text accompanying the photograph. The central distinction is that identification of a clinical condition, identification of where that condition is substantively described, and identification of the educational contribution of its photograph are not the same. A separate question therefore remains: "What additional educational information does the image contribute here?" The longitudinal dimension extends this analysis through the concept of educational value decay, referring to the possibility that the relationship between an image and the context-specific justification for its presence may weaken over time, without assuming that educational value automatically declines with the age or repeated reproduction of the image. In response, I propose an Educational Reassessment Framework—a practical, event-triggered process that proceeds from identifying the image's educational function and its relationship to the curricular structure, through assessment of its image-specific educational contribution and exposure-function proportionality, to the consideration of equivalent alternatives, the continuing validity of its justification, and the conditions of responsible continued use. Reassessment is not intended to automatically remove historical, repeatedly used, or highly sensitive photographs. Responsible reuse of an educationally valuable image may remain justified and may reduce the need to photograph additional patients, but educational value does not substitute for appropriate authorization or other conditions of responsible use. Its purpose is to determine whether continued use is supported by an identifiable, current, and context-specific educational contribution, whether the scope of patient information disclosed is necessary and proportionate to that contribution, and whether the conditions supporting use remain appropriate.

(JMIR Med Educ 2026;12:e104023) doi:10.2196/104023

KEYWORDS

medical education; clinical photography; pediatric patients; bioethics; educational value; educational accountability; educational necessity; visual ethics; clinical images; medical publishing

XSL•FO
RenderX

Introduction

Background

Clinical photography plays an important role in medical education. It can enable learners to recognize phenotypic features, anatomical changes, manifestations of disease, distributions of clinical features, and treatment outcomes that textual description alone may not convey with equal precision [1-6]. In certain circumstances, photographs depicting nudity, intimate anatomy, or other highly sensitive bodily features may likewise serve an important educational function that is difficult to replace [3,7-10]. If a particular anatomical feature, phenotype, or distribution of clinical findings is relevant to an educational objective, cropping or obscuring a clinically relevant area may reduce the educational information available from the image [1,3,11-14]. This viewpoint therefore does not challenge the use of such photographs as educational tools. The scope of patient exposure should, however, remain related to the information the image is expected to convey. Whole-body context may be necessary to assess phenotype, anatomical proportions, distribution of clinical features, or relationships among clinical characteristics [7,9,10]. The question is therefore not simply, "Could less be shown?" but rather "What scope of the image is necessary to convey the information required by the specific educational objective?"

A highly sensitive pediatric clinical photograph is a clinical photograph created during childhood that visually discloses nudity, intimate anatomy, sex-related anatomical characteristics, or other bodily features that reveal particularly sensitive information about the child.

This is a descriptive rather than normative category. It characterizes the nature and scope of the visual information disclosed and does not, by itself, imply that the photograph lacks educational value, is inappropriate, or should not be used. Its educational justification depends separately on the function the image serves and on whether the scope of the information disclosed is necessary and proportionate to that function.

This question should be distinguished from identifiability. Cropping, exposure minimization, and deidentification address related but nonidentical problems [1,15-17]. Removing or obscuring the face does not necessarily eliminate identification based on body shape, distinctive anatomical features, scars, birthmarks, a rare phenotype, or contextual information [1,15-18]. At the same time, restricting an image solely to reduce identifiability may remove clinically relevant information [1,3,15]. The scope of an image should therefore be evaluated both in terms of the information it reveals about the patient and the information it contributes to the educational objective, while maintaining appropriate conditions of consent, privacy, and protection against identification [1,15,19-22].

Repeated use of the same photograph does not in itself imply loss of educational value. Particularly for rare clinical conditions, a high-quality image may remain a useful educational resource for many years, and responsible reuse may reduce the need to photograph additional patients [1-3,23]. This potential advantage of reuse, however, does not substitute for appropriate authorization of the continued use of the existing image. Avoiding unnecessary photography of additional patients should not be achieved by concentrating the cumulative burden of repeated exposure on an individual who did not authorize that use. Responsible reuse should therefore protect both patients who might otherwise be newly photographed and the person whose existing image is repeatedly used. Continued use nevertheless requires two questions to be distinguished: whether the image continues to provide an identifiable educational contribution and whether its use remains supported by appropriate conditions of use [1,15,19-23].

These issues are particularly important for highly sensitive pediatric photographs, where potential educational benefit coexists with extensive disclosure of information about a person who, when the photograph was taken, had limited or no capacity to determine its future use [7-9,24-26].

The literature on clinical photography has focused primarily on consent, privacy, confidentiality, identifiability, deidentification, digital dissemination, patient perspectives on the acceptability and use of medical photographs, and protection of children and other groups requiring heightened safeguards [1,7,8,15,19-22,27-30].

These issues remain fundamental, but they do not exhaust a separate educational question: "What is the relationship between the scope of information disclosed by a highly sensitive image and its specific educational contribution at the point at which it is presented?"

This question arose from a documented longitudinal observation of a single academic textbook lineage. Across 28 documented editions and reprints published between 2005 and 2019, the same highly sensitive pediatric clinical photographs remained within a recurring image-condition configuration. At the point of image presentation, the depicted clinical conditions were explicitly identified, but were not identified in the textbook's table of contents as components of the chapter containing the photographs. Instead, the accompanying text stated that the respective conditions were described elsewhere in the textbook, where their substantive clinical descriptions were located in accordance with the textbook's declared curricular structure.

This produced an observable asymmetry: the declared curricular location of the condition, the location of its substantive clinical description, and the destination indicated in the text accompanying the photograph converged elsewhere, while the highly sensitive pediatric photograph remained at a different location. Its educational contribution at the point of presentation therefore cannot be derived solely from identification of the depicted condition, the location of its substantive clinical description, or the accompanying reference to that other location.

The central distinction of this viewpoint is that identification of the depicted clinical condition, identification of where that condition is substantively described, and identification of the educational contribution of its photograph are analytically distinct elements. A photograph may identify what condition is depicted; the textbook structure may indicate where that condition is addressed; and the accompanying text may state that the condition is described elsewhere. A separate question

XSL•FO
RenderX

remains: “What additional educational information does the photograph itself contribute at the point of its presentation?”

This analysis neither assumes that learners use chapters in a particular sequence nor reconstructs the intentions, learning objectives, or educational reasoning of the author, editor, or publisher. Physical separation between a photograph and the substantive clinical description of the depicted condition is not inherently problematic; cross-referencing may be an appropriate component of educational design [11,12,31-35], and an image may serve a distinct educational function outside the location in which the condition is substantively described.

I refer to educational disconnection as a situation in which the image-specific educational contribution at the point of presentation remains difficult to reconstruct from the observable relationship among the photographs, the condition identified with it, the condition’s declared curricular location, the location of its substantive clinical description, and the destination indicated in the accompanying text. The term does not imply inappropriate use or absence of educational value; it signals that the image’s context-specific educational contribution requires clarification. Only after this contribution has been identified can exposure-function proportionality be meaningfully assessed.

The longitudinal dimension adds a second question: “If the same photograph remains within the same image-condition configuration across successive editions and reprints, does the context-specific justification for its presence at that particular location remain current?” I refer to the possibility that the relationship between an image and its context-specific educational justification may weaken over time as educational value decays. This does not assume that time or repetition automatically diminishes educational value.

The continuing validity of educational justification is distinct from the basis for continued use. An image may retain educational value while the conditions supporting later reuse require reassessment; conversely, an appropriate basis for use does not itself establish the educational necessity of the image at a particular location [1,15,19-23].

On this basis, this viewpoint proposes an Educational Reassessment Framework—a practical, event-triggered process that begins by asking what educational contribution the image provides at its point of presentation and how its location relates to the observable structure of the material. It then assesses image-specific educational contribution, exposure-function proportionality, available alternatives, continuing validity of justification, and conditions of responsible continued use. The educational question developed here is not format-specific and may arise across printed and digital textbooks, multimedia teaching materials, online educational resources, and undergraduate, postgraduate, or continuing professional education [36-39].

Longitudinal Observation Underlying the Viewpoint

This viewpoint is neither a systematic literature review nor an empirical study designed to estimate the prevalence of the phenomenon under consideration in medical education. Its starting point is a documented longitudinal observation of specific educational material, used to identify and conceptualize an educational problem rather than to infer its prevalence.

The reconstruction began after an adult recognized, in 2023, clinical photographs taken of her during childhood that remained available in educational material. Using the name of the physician who had provided treatment during childhood as an identifying reference point, the earlier trajectory of image use was reconstructed from books and textbook publications, conference materials, ebooks, PDFs, presentations, and other available educational and digital resources. The broader reconstruction identified the same pediatric clinical photographs across more than 40 publications and educational or digital resources; this wider long-term reuse and governance trajectory has been reported previously [40].

For the purposes of the present viewpoint, the analysis was narrowed to a single academic textbook lineage , providing a bounded context in which the relationship between the photographs, the clinical conditions identified with them, and the location of their substantive clinical descriptions could be followed longitudinally. Twenty-eight documented editions and reprints containing the analyzed photograph were identified within this textbook lineage between 2005 and 2019. Continued curricular use of existing editions remains documented in 2026.

The more than 40 publications and educational or digital resources establish the wider dissemination trajectory, whereas the 28 documented editions and reprints constitute the bounded longitudinal sequence analyzed here. The present analysis therefore concerns the persistence of an observable image-condition configuration within this textbook lineage.

Across the analyzed sequence, several highly sensitive pediatric clinical photographs were presented with explicit identification of the depicted clinical conditions. Those conditions were not identified in the textbook’s table of contents as components of the chapter containing the photographs. Instead, the accompanying text stated that the respective conditions were described elsewhere in the textbook. Examination of those locations confirmed that their substantive clinical descriptions were provided there, consistent with the textbook’s declared curricular structure.

The observable image-condition configuration comprised five elements, which were (1) the clinical photograph ; (2) the clinical condition identified at the point of image presentation ; (3) the condition’s declared curricular location within the textbook structure; (4) the location of the condition’s substantive clinical description; and (5) the destination indicated in the text accompanying the photograph.

These elements did not converge on the same location.The condition was identified beside the photograph; its substantive clinical description and the destination indicated in the accompanying text were located elsewhere, consistent with its declared curricular location, while the highly sensitive pediatric clinical photograph remained at a different location.

This configuration persisted across the 28 editions and reprints identified between 2005 and 2019 . The central object of observation was therefore not the number of reproductions itself,

XSL•FO
RenderX

but the persistence of this observable relationship across successive editions and reprints.

This observation neither permits determination of the intentions, learning objectives, or educational reasoning of the authors, editors, or publisher nor establishes that the photograph lacked educational value. Physical separation between a clinical photograph and the substantive description of the depicted condition is not itself evidence of educational inadequacy; cross-referencing may be an appropriate component of educational design [11,12,32-35]. The observation establishes only the documented relationship among the photograph, the identified condition, the textbook structure, the location of the substantive clinical description, and the accompanying reference.

The reconstruction was subsequently extended beyond the 2005-2019 publication sequence to distinguish later events in the educational life cycle of the material: discovery and formal objection by the person depicted, contemporary inquiry into documentation supporting continued use, institutional reassessment, modification of access to existing educational material, and continued curricular use of existing editions in 2026. These events are analytically distinct from publication or reproduction because they concern material already in circulation.

Following discovery of continued circulation in 2023, the person depicted raised formal objections to further use during 2023-2025. The person depicted, who is also the author of this viewpoint, reports that she did not provide consent for the educational publication or repeated reuse of these childhood photographs. Separately, in 2025, an institutional inquiry established that the publisher could not produce the requested consent documentation concerning the photographed persons and described reliance on contractual declarations made by authors. These are analytically distinct observations: the first concerns the reported absence of authorization by the person depicted; the second concerns the inability of the publisher, during contemporary inquiry, to produce documentation establishing the basis and scope of authorization for continued use. Both constitute information relevant to contemporary reassessment.

Institutional responses were not uniform. At Pomeranian Medical University in Szczecin, identified editions and reprints were withdrawn from the university library collection and catalogue, with 146 copies withdrawn in total, and relevant recommended-reading lists were subsequently revised. At the same time, existing editions from the analyzed textbook lineage remained formally incorporated into medical and health-professions curricula in other educational settings in 2026.

These later events distinguish publication persistence from curricular persistence and reassessment. Continued designation of an existing textbook as basic, mandatory, recommended, or supplementary literature does not constitute a new publication or reproduction; it demonstrates that previously published material may remain part of contemporary education after the publication sequence has ended. The longitudinal sequence and subsequent reassessment events are summarized in Table 1.

A complete item-level reconstruction of the 28 documented editions and reprints from 2005 to 2019, together with the subsequent objection, authorization inquiry, institutional response, and curricular persistence, is provided in Tables S1 and S2 in Multimedia Appendix 1.

This longitudinal reconstruction provides the observational basis for the concepts of educational disconnection and educational value decay developed in the sections “The Problem of Educational Disconnection” and “Educational Value Decay and the Persistence of Clinical Images”; it does not itself establish absence or loss of educational value. Material changes in the information available to those responsible for continued use may provide a reason for reassessment without predetermining its outcome.

The observation is problem-generating and concept-generating rather than representative. It establishes the longitudinal basis for the question examined in this viewpoint but does not establish the prevalence of educational disconnection beyond the analyzed material.

XSL•FO
RenderX

Table 1. Persistence of the image-condition configuration across 28 documented editions and reprints and subsequent reassessment events, 2005-2026.

Year/period	Documented event	Status of the clinical image or material	Analytical significance
2005	First edition of the analyzed textbook lineage	Highly sensitive pediatric clinical photograph included	The depicted clinical condition was identified beside the photograph. The condition was not identified in the table of contents as part of the chapter containing the photograph; the accompanying text stated that the condition was described elsewhere, and its substantive clinical description was located at that other curricular location.
2005-2006	Subsequent reprints of the first edition	Same photograph reproduced	The same documented image-condition configuration was retained.
2006-2010	Subsequent documented editions and reprints	Same photograph retained	The observable relationship among image location, condition identification, declared curricular location, substantive clinical description, and the destination indicated in the accompanying text persisted.
2015	Second edition	Same photograph retained in a new edition	The same documented image-condition configuration was retained in the new edition.
2015-2019	Subsequent reprints of the second edition	Same photograph reproduced	The same documented image-condition configuration persisted across subsequent reprints.
2019	Last identified reprint within the analyzed 28-item textbook sequence	Same photograph remained present	The documented image-condition configuration remained present at the end of the 2005-2019 longitudinal sequence.
2023-2025	Discovery of continued circulation followed by formal objection by the person depicted	The person depicted became aware of continued use and objected to further use of the photograph	New information entered the educational life cycle of the image, creating a documented external trigger for reassessment; this was not a new publication event.
2025	Institutional inquiry into documentation supporting continued use	The publisher stated that it did not possess the requested consents concerning the photographed persons and described reliance on contractual declarations made by authors	The publisher was unable to produce the requested consent documentation during contemporary inquiry. This constituted new information relevant to reassessment of the conditions supporting continued use.
2025	Institutional reassessment and withdrawal at Pomeranian Medical University in Szczecin	Identified editions and reprints were withdrawn from the university library collection and catalogue; 146 copies were withdrawn in total	Documented example of an external reassessment trigger followed by institutional review and modification of access to existing educational material.
2026	Review and amendment of teaching syllabi at Pomeranian Medical University in Szczecin	Identified textbooks were removed from relevant recommended-reading lists	Documented example of reassessment leading to modification of formal curricular recommendations.
2026	Continued curricular use of existing editions in other educational settings	Existing editions remained designated as basic, mandatory, recommended, or supplementary literature according to the respective institution	Previously published material remained incorporated into contemporary medical and health-professions education after the objection; this represents curricular persistence rather than a new publication or reproduction event.

Author Positionality

The author holds a doctorate in medical and health sciences and is a psychologist, psychotherapist, psychotraumatologist, educator, clinician, and researcher whose work includes patient rights, trauma, medical ethics, and health care governance. At the same time, she is the person depicted in the pediatric clinical photographs whose long-term educational use provided the starting point for the questions developed in this viewpoint. This dual position—as both a researcher and the person depicted in the analyzed photographs—made it possible to identify not only the continued presence of highly sensitive images in educational material, but also their relationship to the curricular structure and substantive clinical teaching. This perspective contributed to the formulation of this paper's central question: "What additional educational function does a highly sensitive image serve at the point of its presentation when the textbook's declared curricular structure, substantive teaching of the depicted condition, and accompanying cross-reference direct the reader elsewhere?"

The analysis does not seek to determine intentions or evaluate the decisions of specific authors, editors, publishers, or institutions. It is limited to observable relationships within the educational material, the scope of patient information disclosed, and the identifiable educational function of the image.

XSL•FO
RenderX

The Problem of Educational Disconnection

Educational disconnection refers to a situation in which a highly sensitive clinical image remains present in educational material, but its image-specific educational contribution at the point of presentation is difficult to reconstruct from the observable relationship among the image, the clinical condition identified with it, the condition's declared curricular location, the location of its substantive clinical description, and the destination indicated in the text accompanying the photograph .

Physical separation of an image from the substantive clinical description of the depicted condition does not in itself constitute educational disconnection. Cross-referencing between chapters may be an appropriate component of educational design, and a photograph may serve a distinct educational function outside the location in which the depicted condition is substantively described [11,12,32-35] . The defining issue is therefore not simply the physical "distance" between an image and related clinical content. It arises when the condition is identified beside the photograph, while the accompanying text states that the condition is described elsewhere and its substantive clinical description is located at that other curricular location, yet the image-specific educational contribution at the point of presentation remains difficult to identify .

The distinction is important because identification of the depicted condition, identification of where that condition is substantively described, and identification of the educational contribution of its photograph are analytically distinct . A photograph may identify what condition is depicted; the textbook's declared curricular structure may indicate where that condition is addressed; and the accompanying text may explicitly state that the condition is described elsewhere. None of these elements alone establishes what additional educational information the photograph itself contributes at the location in which it is presented .

The relevant analytical question is therefore not merely whether the photograph depicts a clinically identifiable condition, but "What additional educational information does this particular photograph contribute here?" Educational disconnection does not mean that the photograph lacks educational value, that its location is inappropriate, or that continued use is unjustified. Nor does it permit reconstruction of the intentions, learning objectives, or educational reasoning of the author, editor, or publisher. It is an epistemic and decision-making signal indicating that the image's context-specific educational contribution requires clarification.

If that contribution can be identified—for example, recognition of a phenotype, distribution of clinical features, body proportions, diagnostic differentiation, or an anatomical relationship relevant to the educational context—the photograph may remain well justified even when the substantive clinical description of the depicted condition is located elsewhere [1,3,7-10] . In such a case, the apparent disconnection may be resolved by identifying the distinct educational contribution made by the image at its actual point of presentation.

If, however, the image-specific educational contribution at that location cannot be identified , reassessment of the image's function, location, or continued use is warranted. Reassessment does not predetermine removal. Its purpose is to determine whether a context-specific educational justification can be established.

Only after the image's educational contribution has been identified can exposure - function proportionality be meaningfully assessed. Nudity, intimate anatomy, or full-body presentation may be educationally necessary when these features convey information required for recognition of phenotype, anatomical relationships, distribution of clinical findings, or another specified educational objective [1,3,7-10] . Conversely, where additional exposure contributes no information necessary to the identified function, its inclusion requires separate justification. The relevant standard is therefore not minimum exposure in the abstract, but exposure proportional to the educational information that the image contributes at the point of presentation.

Educational disconnection is therefore neither a stand-alone educational theory nor an empirically validated measurement category. It is a conceptual tool introduced in this viewpoint for identifying situations in which the relationship among the clinical photograph, the condition identified with it, the condition's declared curricular location, the location of its substantive clinical description, the destination indicated in the accompanying text, the image-specific educational contribution, and the scope of patient exposure requires clarification and, where necessary, reassessment.

Educational Value Decay and the Persistence of Clinical Images

Educational disconnection is primarily contextual: it concerns whether the educational contribution of an image can be identified at the point of presentation in relation to the observable image - condition configuration . Longitudinal persistence introduces a different question: " If the same photograph remains within the same image–condition configuration across successive editions and reprints, does the context-specific justification for its presence at that particular location remain current over time?"

Neither the passage of time nor repeated reproduction demonstrates loss of educational value. A photograph depicting a characteristic phenotype, anatomical feature, or distribution of clinical findings may remain educationally valuable for many years [1-3,23]. The persistence of the image is therefore not itself the problem; the relevant question concerns the persistence of its context-specific educational justification.

Educational value depends on the relationship among what the image shows, what educational information it provides, where and how it is presented, and the educational objective it serves [11-13,31-35,41]. The photograph itself may remain unchanged while its educational context changes through revised objectives, relocation or redistribution, alternative representations, or a different technological or distribution environment.

In this viewpoint, educational value decay refers to the possibility that the relationship between a clinical image and

XSL•FO
RenderX

the context-specific educational justification for its presence at a particular location may weaken over time, even while the image itself continues to be reproduced or used.

The term does not mean that an older photograph is less educationally valuable or that repeated reproduction progressively diminishes its value. It describes a possible divergence between persistence of the image and persistence of the justification for its particular educational use . These trajectories may remain fully aligned, in which case continued use may remain well justified.

Educational value decay becomes a question for reassessment when a material change affects the image's educational function, location, available alternatives, or conditions of continued use, including information relevant to the basis or scope of authorization for that use . The concept therefore identifies a possibility requiring evaluation, not a conclusion that educational value has actually declined.

The distinction between the two constructs can be expressed through two questions: (1) educational disconnection: "What additional educational information does this image contribute here, when the condition identified with it is substantively described elsewhere in the material?" (2) educational value decay: "Does the context-specific justification for the presence of this image here remain current at the point of subsequent use?"

The first question is primarily contextual ; the second is temporal .

For highly sensitive pediatric photographs, this temporal distinction has particular significance because the information disclosed by the image remains embedded in the photograph while the educational and technological context surrounding its use may change [1,7,8,15,17-20,42] . Reassessment therefore asks whether the scope of exposure continues to correspond to the image-specific educational contribution at the point of current use , rather than assuming that sensitivity, age, or repetition makes continued use inappropriate.

Accordingly, educational value decay is not an empirically validated measurement category or evidence that educational value necessarily diminishes over time. It is a conceptual tool introduced in this viewpoint for identifying when longitudinal persistence creates a reason to reassess whether an image's context-specific educational justification remains current.

A Practical Educational Reassessment Framework for Highly Sensitive Pediatric Clinical Images

The Educational Reassessment Framework translates the preceding concepts into a practical, sequential decision-making process. It begins not with the age, number of reproductions, or sensitivity of the photograph, but by asking what educational contribution the image provides at the specific point at which it appears . The framework proceeds from identifying what the photograph depicts and what educational information it is expected to provide, through assessment of its relationship to the observable image - condition configuration and its image-specific educational contribution at that location, to exposure - function proportionality , equivalent alternatives, continuing validity of justification, conditions of responsible use, and a documented decision.

The framework is neither intended to classify photographs automatically as educational or noneducational, nor does it presume that historical, repeatedly used, or highly sensitive images should be withdrawn. Its purpose is to make the reasoning supporting continued use identifiable, current, context-specific, and reconstructable .

When Should Reassessment Be Triggered?

Educational disconnection is one trigger for reassessment. It arises when the photograph and the clinical condition identified with it can be observed, but the image-specific educational contribution at the point of presentation remains difficult to reconstruct from the relationship among image location, condition identification, the condition's declared curricular location, the location of its substantive clinical description, and the destination indicated in the accompanying text .

Reassessment may also be triggered by a material change in the educational life cycle of the image , such as a new edition, substantial content revision, relocation of the image or change in its educational context, change in audience or mode of distribution, migration to a digital environment, emergence of an equally effective but less intrusive alternative, or inability to reconstruct the previous justification for use [1-3,15-20,23,42-44] .

New material information may likewise provide a trigger, including a documented objection or request to discontinue use by the person depicted, newly identified information concerning the conditions under which the image was obtained or authorized, or inability during contemporary review to reconstruct or produce relevant authorization documentation [1,7-10,15,19-24,45] .

These events do not predetermine the outcome of reassessment. They alter the information or context relevant to the decision and therefore create a reason to reconsider whether the existing justification remains applicable. Not every technical reproduction requires repetition of the entire process. Reassessment is relevant when a decision concerning continued use is made or when a material change affects the context in which the image is expected to serve its educational function.

Who Should Perform the Reassessment?

Responsibility should be located as close as practicable to the decision concerning use of the image. Depending on the educational setting, this may be the author, content editor, educational team, publisher, or another entity responsible for approving or continuing its use.

Multiple parallel assessments are unnecessary. What is required is an identifiable point of responsibility capable of linking the image's educational function, its location within the educational material, the scope of patient information disclosed, the conditions of use, and the resulting decision.

The framework comprises 8 sequential stages (Table 2). The first 3 stages establish what condition the image depicts, where that condition is located and substantively described within the

XSL•FO
RenderX

material, and what additional educational information the image contributes at its actual point of presentation . Only after these questions have been addressed can exposure-function proportionality, equivalent alternatives, continuing validity of justification, and conditions of responsible use be meaningfully assessed.

Table 2. Educational reassessment framework: from clinical identification to context-specific educational justification.

Assessment stage	Decision question	Assessment criterion/possible action	Responsible entity
1. Educational function	What educational information is this image expected to provide?	Identify the specific information or educational function provided by the image itself, such as recognition of a phenotype, distribution of clinical findings, an anatomical relationship, diagnostic differentiation, or treatment outcome. Identification of the depicted clinical condition alone does not establish the educational function of the image [5,11-14,32-35,41,46].	Author or content editor
2. Image-condition configuration	Where does the material locate and substantively describe the condition identified with the photograph, and how does the image location relate to those locations?	Examine image location, condition identification, declared curricular location, location of substantive clinical description, and the destination indicated in the accompanying text. Separation among these elements is not inherently problematic; the purpose of this stage is to establish the observable configuration before inferring an educational function.	Author, content editor, or educational team
3. Contribution at the point of presentation	What additional educational information does this particular image contribute here?	Identify the image-specific educational contribution at the point of presentation. Where the accompanying text states that the depicted condition is described elsewhere, determine what educational information the photograph contributes at its present location beyond identification of the condition. If a context-specific contribution can be identified, the image location may remain justified. If it remains difficult to identify, *educational disconnection* signals the need for reassessment of function, location, or continued use.	Author, content editor, or educational team
4. Exposure-function proportionality	Is the scope of patient information disclosed necessary to fulfill the function identified in stages 1-3?	Retain full-body context, intimate anatomy, or other highly sensitive elements when they are necessary to convey the specified educational information. Where part of the exposure contributes no information necessary to the identified objective, consider limiting it without removing clinically relevant educational information [1,3,7-10,15].	Author or content editor
5. Equivalent alternative	Can the same context-specific information be conveyed equally effectively through a less intrusive representation?	Compare alternatives against the same identified educational function, rather than solely against degree of exposure. If a less intrusive representation preserves equivalent educational information, it should be considered. If it does not, continued use of the existing image may remain educationally justified, subject to the separate assessment of the conditions of responsible use in stage 7 [1,3,4,11-15,31,46].	Author, editor, or educational team
6. Temporal validity of justification	Does the context-specific justification for the presence of this image here remain current at the next material point of use?	Consider whether changes in image function, material structure, location, educational objectives, audience, available alternatives, or distribution environment affect the previous justification. Prior publication alone does not establish that the same context-specific justification remains current. This stage operationalizes the temporal question associated with *educational value decay*.	Entity responsible for continued use
7. Conditions of responsible use	Are the current conditions of image use, including the basis and scope of authorization, appropriate for the intended use?	Independently assess the basis for use, consent or assent where applicable, the documented scope of authorization in relation to the current use, documented objection or withdrawal request where relevant, privacy, identifiability, the availability and reconstructability of relevant authorization documentation, and the distribution environment [1,7-10,15-24,42-45,47-49]. Prior publication or repeated use alone should not be treated as evidence that the current use falls within the scope of an earlier authorization. Educational value does not substitute for appropriate conditions of use, and an appropriate basis for use does not itself establish educational necessity at a particular location.	Publisher or other entity responsible for use
8. Documented decision	Can it later be reconstructed why the image was retained, modified, relocated, replaced, or withdrawn?	Maintain a concise record of the decision sequence: educational function→image-condition configuration→contribution here→exposure-function proportionality→alternatives→temporal validity→conditions of use→decision.	Entity approving use

The framework should be understood as a sequence rather than a collection of independent criteria . In particular, neither can exposure-function proportionality nor the equivalence of an alternative be meaningfully assessed until it has first been established what condition the image depicts, where that condition is located and substantively described within the material, and what additional educational information the photograph contributes at its point of presentation .

XSL•FO
RenderX

This ordering is important. Asking whether a photograph reveals "too much" before establishing what information it needs to convey reverses the analytical sequence. Similarly, asking whether a less intrusive alternative is equivalent requires first identifying the educational function against which equivalence is to be assessed .

A discrepancy between image location and the location in which the depicted condition is substantively described does not predetermine the outcome of reassessment. An image positioned elsewhere may serve a distinct and well-justified educational function—for example, comparison, phenotype recognition, diagnostic differentiation, or illustration of a cross-cutting clinical feature [11,12,32,34,35,41,46] . The framework requires only that the function be identifiable and linked to the point at which the image is presented .

The outcome of reassessment is therefore not predetermined. If a photograph serves an identifiable educational function at the point of presentation, is meaningfully integrated into the educational material, and requires its full scope of exposure to convey that information, and if the conditions of responsible use—including the basis and scope of authorization—remain appropriate , reassessment may confirm continued use of the image without modification . In other circumstances, reassessment may support improved integration with the surrounding content, modification of location or caption, reduction of exposure, use of an equally effective alternative, or discontinuation of further use.

The framework therefore neither asks whether a highly sensitive photograph should remain because it has been used before nor whether it should be removed because it is old, repeatedly reproduced, or highly sensitive . It asks whether its continued use can be supported by a current, identifiable, and context-specific educational justification, together with appropriate conditions of responsible use, including an appropriate basis and scope of authorization for the intended use .

Discussion

Overview

The analysis suggests that evaluation of highly sensitive pediatric clinical photographs should begin with the relationship between the observable image - condition configuration and the educational contribution of the photograph at its point of presentation . Identification of the depicted clinical condition, identification of where that condition is substantively described, and identification of the educational contribution of its photograph are analytically distinct. The relevant question is what additional educational information the photograph itself contributes at the point of its actual presentation .

This is the function of educational disconnection as a conceptual signal. It does not imply that a photograph is unnecessary, inappropriate, or devoid of educational value. Rather, reassessment serves to clarify the image-specific educational contribution , not to confirm a predetermined assumption of inappropriate use.

Proportionality Follows Identification of Educational Function

The scope of exposure cannot be meaningfully assessed until the image's educational function has been identified. Nudity, intimate anatomy, full-body presentation, or other highly sensitive bodily features may be necessary to convey clinically relevant information [1,3,7-10] . Minimization should therefore not mean simply "show as little as possible," but disclosing the scope of patient information necessary to preserve the information required by the identified educational function .

The resulting principle is that the scope of exposure can be assessed as proportionate only in relation to an identified educational contribution.

If full exposure provides information necessary to achieve the identified educational objective, retaining it may be justified [1,3,7-10] . Conversely, if additional exposure contributes no information necessary to that function, its inclusion requires separate justification. The relevant comparison is between the patient information disclosed and the educational information contributed by that disclosure .

Reuse Is Not a Problem in Itself

Reuse is not an indicator of inappropriate use or loss of educational value. A high-quality photograph, particularly one depicting a rare or difficult-to-document condition, may retain substantial educational value for many years, and responsible reuse may reduce the need to photograph additional patients [1-3,23] .

This potential ethical advantage of reuse, however, depends on the conditions under which the existing image continues to be used. Avoiding unnecessary photography of additional patients should not be achieved by concentrating the cumulative burden of repeated exposure on a single individual who did not authorize that use. Responsible reuse should therefore protect both patients who might otherwise be newly photographed and the person whose existing image is repeatedly reproduced.

This distinction becomes particularly important when the same image persists across successive editions, formats, distribution environments, or educational settings over extended periods. In such circumstances, an individual may effectively become an enduring visual representation of a clinical condition. Where that representational role was not authorized by the person depicted, the fact that continued reuse may reduce the need to photograph additional patients does not, by itself, provide an ethical justification for continued exposure.

The significance of longitudinality is therefore not the number of reproductions itself. Educational value decay asks whether the context-specific justification for the image's presence remains current . Repeated use becomes relevant to reassessment when a material trigger changes the educational context, the conditions of responsible use, or the information available to the entity responsible for continued use. The longitudinal question is therefore not "Has this image been used too many times?" but "Does the justification for this particular use remain current?"

XSL•FO
RenderX

Consent and Educational Function Are Distinct Dimensions of Assessment

Consent, assent, privacy, and protection against identification remain fundamental conditions of responsible clinical-image use [1,7-10,15,19-22,24,45,47-50] , but they answer a different question from educational function.

An image may have substantial educational value while the conditions supporting a particular use require reassessment. Conversely, an appropriate basis for use does not establish why a particular photograph is educationally necessary at a particular location. Educational justification and authorization should therefore be evaluated as related but analytically distinct dimensions.

Historical consent, where documented, should not automatically be treated as establishing authorization for every subsequent form of use. The existence of consent, its documented scope, and the specific later use are distinct questions. Previous lifecycle-governance analysis has identified the importance of durable consent documentation and verification checkpoints before republication, digitization, or educational redistribution [40] . Changes in educational context, audience, distribution environment, or mode of dissemination may therefore require consideration of whether a subsequent use remains appropriately supported [1,7-10,15,19-22,24,45] .

For pediatric photographs, this assessment may additionally require consideration of assent where developmentally applicable and later preferences or objections expressed by the person depicted. Prior publication or repeated use alone should not be treated as evidence that every later reproduction, redistribution, or educational use falls within the scope of an earlier authorization.

The potential educational or ethical advantage of reusing an existing image—including avoidance of unnecessary photography of additional patients—does not substitute for establishing that the continued use of that particular image remains appropriately authorized. Educational value, rarity, image quality, or previous use cannot by themselves answer the separate question of whether the person depicted authorized the relevant use.

Two questions should therefore remain distinct: “Can we use this image?” and “Why do we need this particular image, here, and with this scope of exposure?”

The first concerns conditions of responsible use ; the second concerns educational function, image-specific contribution, and exposure - function proportionality . The Educational Reassessment Framework does not replace requirements concerning consent, assent, privacy, authorization, or patient protection; it adds a distinct level of educational justification .

In digital environments, protection against identification must also extend beyond the face. Distinctive bodily features, scars, birthmarks, rare phenotypes, or contextual information may enable recognition, while copying, searching, screenshotting, and secondary dissemination may alter the scale and persistence of exposure [1,15-18,42-44] . Digital redistribution may therefore change relevant conditions of use even when the underlying photograph remains unchanged.

The Significance of Educational Architecture

The proposed analysis extends beyond the conventional image-text relationship to the relationship between the photograph, the clinical condition identified with it, and the observable structure of the educational material . Educational design literature supports attention to relationships among instructional content, learning objectives, organization, and the integration of educational elements [11-13,31-35,51] .

This does not require linear educational materials or placement of every photograph alongside the substantive clinical description of the depicted condition. An image may be educationally valuable precisely because it appears elsewhere—for example, as a comparison, an example of phenotype, an element of differential diagnosis, or an illustration of a cross-cutting topic [11,12,32,34,35,41,46] . Such a context-specific function should, however, be identifiable.

The issue is therefore not “an image appears in another chapter,” but whether its presence at that location provides an identifiable additional educational contribution . An accompanying reference may indicate where the depicted condition is substantively described without establishing what the photograph contributes at its actual location.

Scope of the Proposed Argument

This viewpoint supports a conceptual rather than empirical generalization . Its contribution is to distinguish four questions: (1) “What clinical condition is identified with the image?” (2) “Where is that condition substantively described within the educational material?” (3) “What additional educational information does the image contribute at the point of its actual presentation?” and (4) “Is the scope of patient information disclosed necessary for that specific contribution?”

These questions correspond respectively to clinical identification, image - condition configuration, image-specific educational contribution, and exposure - function proportionality .

The broader implication is that continued presence, continued educational justification, and authorization for continued use are related but distinct questions . A photograph may remain educationally appropriate in its original form and location when its current educational contribution can be identified, and the scope of information disclosed is proportionate to that contribution; continued use additionally requires that the applicable conditions of responsible use, including the basis and scope of authorization, remain appropriately supported.

Limitations

This viewpoint has several limitations that define the scope of the proposed argument.

First, its starting point is a documented longitudinal observation of a single academic textbook lineage , rather than a representative sample of medical textbooks or a systematic review of educational resources. The persistence of the analyzed image - condition configuration across 28 documented editions

XSL•FO
RenderX

and reprints between 2005 and 2019 provides a longitudinal basis for formulating a conceptual problem but does not allow estimation of the prevalence of educational disconnection in medical education. The present argument therefore supports conceptual rather than empirical generalization.

Second, the analysis concerns the observable architecture of educational material , rather than direct learning outcomes. Neither were experiments involving learners conducted, nor were learning outcomes compared with and without the analyzed photographs. The analysis therefore cannot establish whether a photograph possessed a particular degree of educational value; it can establish only the observable relationship between the image, the condition identified with it, its location within the material, and its image-specific educational contribution.

Third, the complete process of educational design and use cannot be reconstructed from the final educational product. Authors, editors, or publishers may have had educational objectives or reasons for image placement that are not visible to the reader, while learners may use chapters in different sequences, consult other sections, or receive additional context from educators [11,12,31-35] . The analysis is therefore limited to the observable image - condition configuration and does not infer undocumented intentions, learning objectives, or educational reasoning.

Fourth, educational disconnection is a conceptual construct rather than an empirically validated measurement category. Further operationalization is required to determine whether independent researchers or educators can reliably identify and assess it. The construct should therefore be understood as a conceptual and decision-making signal rather than a diagnostic classification of educational material.

Fifth, educational value decay is a temporal extension of the proposed analysis rather than a claim that educational value declines over time. Persistence of the analyzed image - condition configuration across the longitudinal sequence cannot itself establish that an earlier educational justification weakened or became outdated. Determining whether that justification remains current requires assessment at subsequent points of use. Persistence of educational justification does not establish that the conditions authorizing continued use remain appropriate; educational justification and authorization are assessed as distinct dimensions in the proposed framework.

Sixth, the Educational Reassessment Framework is a proposed normative and practical decision-making tool rather than a validated measurement instrument. Its sequential logic requires further evaluation for feasibility, interuser agreement, reproducibility, and usefulness across specialties and educational settings.

Seventh, the analysis focuses on highly sensitive pediatric clinical photographs , for which the relationship between educational function and scope of patient exposure has particular significance. Its broader applicability to other illustrations used in medical education requires further study.

Finally, the author occupies a dual position as both researcher and the person depicted in the photographs that initiated the analysis. This position contributed to identifying the relevant relationships within the educational material but may also influence conceptualization of the problem. The analysis therefore distinguishes observable properties of the educational material from inferences concerning intentions, motives, learning objectives, or undocumented decision-making processes. The proposed constructs and framework should be examined by independent researchers and educators using materials beyond the case from which they were developed.

These limitations define a direct research agenda. A next step should be a systematic empirical study of a larger corpus of educational materials , independently coding image location, condition identification, declared curricular location, substantive clinical description, cross-reference destination, image-specific educational contribution, and scope of patient exposure . Such research could estimate the prevalence and forms of educational disconnection, test its operationalizability and interrater reliability, examine whether educational value decay can be operationalized longitudinally, and evaluate the feasibility and reproducibility of the Educational Reassessment Framework across educational settings.

Conclusions

Clinical photography can be an indispensable tool in medical education, including when it depicts nudity, intimate anatomy, or other highly sensitive bodily features of a child. Neither a high degree of exposure nor repeated use of an image determines, in itself, its educational value or the justification for its continued use.

The central distinction proposed in this viewpoint is that identification of the depicted clinical condition, identification of where that condition is substantively described, and identification of the educational contribution of the photograph are distinct elements . A separate question therefore remains: “ What additional educational information does this photograph contribute at the point of its actual presentation?”

Educational disconnection identifies situations in which this image-specific contribution is difficult to reconstruct from the observable image - condition configuration . It does not imply that the photograph lacks educational value or that its use or location is inappropriate; rather, it signals the need to clarify what the image contributes here and why it is needed at this particular location .

When the same image - condition configuration persists across successive editions and reprints, educational value decay adds a temporal question: “Does the context-specific justification for the image’s presence remain current?” The concept identifies the possibility that persistence of the photograph and persistence of its context-specific educational justification may cease to remain aligned , without assuming that educational value necessarily declines over time or through repeated reproduction.

The Educational Reassessment Framework translates these questions into a practical decision-making process: “ What clinical condition is identified with the image?”→“Where is that condition substantively described within the material?”→“What additional educational information does the image contribute here?”→“Why is this particular image needed

XSL•FO
RenderX

at this particular location?" Only then can the proportionality of patient exposure, equivalent alternatives, continuing validity of justification, and conditions of responsible use, including the basis and scope of authorization , be meaningfully assessed.

Reassessment is not a mechanism for the automatic removal of historical, repeatedly used, or highly sensitive clinical photographs. Responsible reuse of a high-quality image may remain educationally valuable and may reduce the need to photograph additional patients. However, educational value and the potential benefits of reuse do not substitute for appropriate authorization of the continued use of the existing image. Reassessment may therefore confirm continued use, including the full scope of exposure, when supported by both an identifiable, current, and context-specific educational contribution and appropriate conditions of responsible use .

The final question is therefore not simply, "Can this image remain?" It is "What additional educational information does this image contribute here, why is this particular image needed at this particular location, is the scope of patient information it discloses necessary and proportionate to that specific educational contribution, and are the conditions supporting its continued use appropriately established?"

Acknowledgments

Generative AI (ChatGPT, OpenAI) was used solely for language editing and to improve the clarity and readability of this paper. The author reviewed and verified the final text and takes full responsibility for the accuracy, originality, integrity, references, and citations.

Funding

The author received no external financial support for this work. All costs associated with the preparation of the manuscript were covered by the author from personal funds.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Detailed longitudinal reconstruction of 28 editions and reprints of a single academic textbook lineage and subsequent reassessment and curricular persistence, 2005-2026.
[DOCX File , 34 KB - mededu_v12i1e104023_app1.docx]

References

1. Shinkai K, Bruckner A, Robinson J. Best practices for sharing images in clinical care, research, and education-protecting patient privacy. JAMA Dermatol 2023;159(7):695-697. [doi: 10.1001/jamadermatol.2023.1080] [Medline: 37163261]
2. Harting MT, DeWees JM, Vela KM, Khirallah RT. Medical photography: current technology, evolving issues and legal perspectives. Int J Clin Pract 2015;69(4):401-409. [doi: 10.1111/ijcp.12627] [Medline: 25708063]
3. Kornhaber R, Betihavas V, Baber R. Ethical implications of digital images for teaching and learning purposes: an integrative review. J Multidiscip Healthcare 2015;8:299-305 [FREE Full text] [doi: 10.2147/JMDH.S84488] [Medline: 26089681]
4. Grunwald T, Corsbie-Massay C. Guidelines for cognitively efficient multimedia learning tools: educational strategies, cognitive load, and interface design. Acad Med 2006;81(3):213-223. [doi: 10.1097/00001888-200603000-00003] [Medline: 16501261]
5. Biggs JB, Tang CSK, Kennedy G. Teaching for Quality Learning at University. United Kingdom: Open University Press; 2011.
6. Benlidayi IC. Medical Photography for Scientific Publishing (MedPhotoPubl) in rheumatology and rehabilitation. J Korean Med Sci 2026;41(16):e189. [doi: 10.3346/jkms.2026.41.e189]
7. Kennedy A. Capturing the medical gaze regulating medical photography of minors with intersex variations. Med Law Rev 2025;33(4):fwaf045 [FREE Full text] [doi: 10.1093/medlaw/fwaf045] [Medline: 41412576]
8. Devakumar D, Brotherton H, Halbert J, Clarke A, Prost A, Hall J. Taking ethical photos of children for medical and research purposes in low-resource settings: an exploratory qualitative study. BMC Med Ethics 2013;14:27 [FREE Full text] [doi: 10.1186/1472-6939-14-27] [Medline: 23835013]
9. Creighton S, Alderson J, Brown S, Minto CL. Medical photography: ethics, consent and the intersex patient. BJU Int 2002;89(1):67-71. [doi: 10.1046/j.1464-4096.2001.01809.x] [Medline: 11849163]
10. Stikkelbroeck NM, Beerendonk CC, Willemsen WN, Schreuders-Bais CA, Feitz WF, Rieu PN, et al. The long term outcome of feminizing genital surgery for congenital adrenal hyperplasia: anatomical, functional and cosmetic outcomes, psychosexual development, and satisfaction in adult female patients. J Pediatr Adolesc Gynecol 2003;16(5):289-296. [doi: 10.1016/s1083-3188(03)00155-4] [Medline: 14597017]
11. Mayer RE. Multimedia learning. Psychol Learn Motiv 2002;41:85-139. [doi: 10.1016/s0079-7421(02)80005-6]

XSL•FO
RenderX

12. Mayer RE. The promise of multimedia learning: using the same instructional design methods across different media. Learn Instr 2003;13(2):125-139. [doi: 10.1016/s0959-4752(02)00016-6]
13. van Merriënboer JJG, Sweller J. Cognitive load theory in health professional education: design principles and strategies. Med Educ 2010;44(1):85-93. [doi: 10.1111/j.1365-2923.2009.03498.x] [Medline: 20078759]
14. Young JQ, Van Merrienboer J, Durning S, Ten Cate O. Cognitive load theory: implications for medical education: AMEE Guide No. 86. Med Teach 2014;36(5):371-384. [doi: 10.3109/0142159X.2014.889290] [Medline: 24593808]
15. Salvador T, Gu L, Hay JL, Kurtansky NR, Masterson-Creber R, Halpern AC, et al. Consent and identifiability for patient images in research, education, and image-based artificial intelligence. JAMA Dermatol 2024;160(4):470-472. [doi: 10.1001/jamadermatol.2024.0084] [Medline: 38477909]
16. Roberts E, Troiano C, Spiegel J. Standardization of guidelines for patient photograph deidentification. Ann Plast Surg 2016;76(6):611-614. [doi: 10.1097/SAP.0000000000000817] [Medline: 27015333]
17. Slobogin C. Hidden in plain sight: the covering of patients' eyes and a microethics of medical photography. Med Humanit 2025;50(4):770-778. [doi: 10.1136/medhum-2024-012894] [Medline: 39122258]
18. Marshall Z, Bhattacharjee M, Wang M, Cadri A, James H, Asghari S, et al. Finding medical photographs of patients online: randomized, cross-sectional study. J Med Internet Res 2024;26:e55352 [FREE Full text] [doi: 10.2196/55352] [Medline: 38913416]
19. Roguljić M, Šimunović D, Buljan I, Žuljević MF, Turić A, Marušić A. Publishing identifiable patient photographs in the digital age: focus group study of patients, doctors, and medical students. J Med Internet Res 2025;27:e59970 [FREE Full text] [doi: 10.2196/59970] [Medline: 40053737]
20. Roguljić M, Šimunović D, Peričić TP, Viđak M, Utrobičić A, Marušić M, et al. Publishing identifiable patient photographs in scientific journals: scoping review of policies and practices. J Med Internet Res 2022;24(8):e37594 [FREE Full text] [doi: 10.2196/37594] [Medline: 36044262]
21. Taiwo T, Obiakor B, McClung S, Shinkai K. Informed consent practices for publication of patient images in dermatology journals. JMIR Dermatol 2025;8:e60795 [FREE Full text] [doi: 10.2196/60795] [Medline: 40249653]
22. Nanda JK, Marchetti MA. Consent and deidentification of patient images in dermatology journals: observational study. JMIR Dermatol 2022;5(3):e37398 [FREE Full text] [doi: 10.2196/37398] [Medline: 36777646]
23. Chandawarkar R, Nadkarni P. Safe clinical photography: best practice guidelines for risk management and mitigation. Arch Plast Surg 2021;48(3):295-304 [FREE Full text] [doi: 10.5999/aps.2021.00262] [Medline: 34024075]
24. Cunniff C, Byrne J, Hudgins L, Moeschler J, Olney A, Pauli R. Informed consent for medical photographs. Genet Med 2000;2(6):353-355. [doi: 10.1097/00125817-200011000-00010]
25. General comment No. 25 (2021) on children's rights in relation to the digital environment. United Nations. 2021. URL: https://www.ohchr.org/en/documents/general-comments-and-recommendations/general-comment-no-25-2021-childrens-rights-relation [accessed 2026-09-04]
26. Convention on the rights of the child. United Nations. 1991. URL: https://www.ohchr.org/en/instruments-mechanisms/instruments/convention-rights-child [accessed 2026-09-04]
27. Hsieh C, Yun D, Bhatia A, Hsu J, Ruiz de Luzuriaga AM. Patient perception on the usage of smartphones for medical photography and for reference in dermatology. Dermatol Surg 2015;41(1):149-154. [doi: 10.1097/DSS.0000000000000213] [Medline: 25533160]
28. Cheung A, Al-Ausi M, Hathorn I, Hyam J, Jaye P. Patients' attitudes toward medical photography in the emergency department. Emerg Med J 2005;22(8):609 [FREE Full text] [doi: 10.1136/emj.2004.019893] [Medline: 16046785]
29. Lau CK, Schumacher HH, Irwin MS. Patients' perception of medical photography. J Plast Reconstr Aesthetic Surg 2010;63(6):e507-e511. [doi: 10.1016/j.bjps.2009.11.005] [Medline: 19959408]
30. Adeyemo W, Mofikoya B, Akadiri O, James O, Fashina A. Acceptance and perception of Nigerian patients to medical photography. Dev World Bioethics 2013;13(3):105-110. [doi: 10.1111/j.1471-8847.2012.00328.x] [Medline: 22708667]
31. Mayer RE, Moreno R. Nine ways to reduce cognitive load in multimedia learning. Educ Psychol 2003;38(1):43-52. [doi: 10.1207/S15326985EP3801_6]
32. Biggs J. Enhancing teaching through constructive alignment. High Educ 1996;32(3):347-364. [doi: 10.1007/bf00138871]
33. Harden R. Learning outcomes and instructional objectives: is there a difference? Med Teach 2002;24(2):151-155. [doi: 10.1080/0142159022020687] [Medline: 12098434]
34. Prideaux D. Curriculum design. BMJ 2003;326(7383):268-270 [FREE Full text] [doi: 10.1136/bmj.326.7383.268] [Medline: 12560283]
35. Thomas PA, Kern DE, Hughes MT, Tackett SA, Chen BY. Curriculum Development for Medical Education: A Six-Step Approach. Baltimore, MD: Johns Hopkins University Press; 2022.
36. Cook DA. The research we still are not doing: an agenda for the study of computer-based learning. Acad Med 2005;80(6):541-548. [doi: 10.1097/00001888-200506000-00005] [Medline: 15917356]
37. Cook DA, Levinson AJ, Garside S, Dupras DM, Erwin PJ, Montori VM. Internet-based learning in the health professions: a meta-analysis. JAMA 2008;300(10):1181-1196. [doi: 10.1001/jama.300.10.1181] [Medline: 18780847]
38. Cook DA, Levinson AJ, Garside S. Time and learning efficiency in internet-based learning: a systematic review and meta-analysis. Adv Health Sci Educ 2010;15(5):755-770. [doi: 10.1007/s10459-010-9231-x] [Medline: 20467807]

XSL•FO
RenderX

39. Ruiz JG, Mintzer MJ, Leipzig RM. The impact of e-learning in medical education. Acad Med 2006;81(3):207-212. [doi: 10.1097/00001888-200603000-00002] [Medline: 16501260]
40. Bujek I. Pediatric clinical images without consent: a governance gap in the long-term reuse of health data in digital health ecosystems. J Med Internet Res 2026;28:e94952 [FREE Full text] [doi: 10.2196/94952] [Medline: 42224673]
41. Norman G. Teaching basic science to optimize transfer. Med Teach 2009;31(9):807-811. [doi: 10.1080/01421590903049814] [Medline: 19811185]
42. Bhattacharjee M, Bagheri E, Asaad L, Marshall Z. Patient photographs and Google Images: an AI ethics case study. Int Rev Inf Ethics 2024;34(1). [doi: 10.29173/irie529]
43. Clunie DA, Flanders A, Taylor A, Erickson B, Bialecki B, Brundage D, et al. Report of the medical image de-identification (MIDI) task group--best practices and recommendations. arXiv Preprint posted online on Mar 16, 2025. [doi: 10.48550/arXiv.2303.10473]
44. El HM, Cetin D, Stamenkovic P, Huber N, Zünd F. VerA: versatile anonymization applicable to clinical facial photographs. : IEEE; 2025 Presented at: IEEE/CVF Winter Conference on Applications of Computer Vision (WACV); February 26, 2025; Tucson, AZ p. 127-138. [doi: 10.1109/wacv61041.2025.00023]
45. Hacard F, Maruani A, Delaplace M, Caille A, Machet L, Lorette G, et al. Patients' acceptance of medical photography in a French adult and paediatric dermatology department: a questionnaire survey. Br J Dermatol 2013;169(2):298-305. [doi: 10.1111/bjd.12345] [Medline: 23551168]
46. Kellman PJ, Krasne S. Accelerating expertise: perceptual and adaptive learning technology in medical learning. Med Teach 2018;40(8):797-802 [FREE Full text] [doi: 10.1080/0142159X.2018.1484897] [Medline: 30091650]
47. Kunde L, McMeniman E, Parker M. Clinical photography in dermatology: ethical and medico-legal considerations in the age of digital and smartphone technology. Australas J Dermatol 2013;54(3):192-197. [doi: 10.1111/ajd.12063] [Medline: 23713892]
48. Lakdawala N, Fontanella D, Grant-Kels JM. Ethical considerations in dermatologic photography. Clin Dermatol 2012;30(5):486-491. [doi: 10.1016/j.clindermatol.2011.06.017] [Medline: 22902218]
49. Berle I. Clinical photography and patient rights: the need for orthopraxy. J Med Ethics 2008;34(2):89-92. [doi: 10.1136/jme.2006.019166] [Medline: 18234945]
50. Rimoin L, Haberle S, DeLong Aspey L, Grant-Kels JM, Stoff B. Informed consent, use, and storage of digital photography among Mohs surgeons in the United States. Dermatol Surg 2016;42(3):305-309. [doi: 10.1097/DSS.0000000000000634] [Medline: 26863597]
51. Çeken B, Taşkın N. Multimedia learning principles in different learning environments: a systematic review. Smart Learn. Environ 2022;9(1):19. [doi: 10.1186/s40561-022-00200-2]

Edited by B Lesselroth; submitted 08.Jun.2026; peer-reviewed by K Shinkai, R Fatteh; comments to author 07.Aug.2026; revised version received 14.Aug.2026; accepted 31.Aug.2026; published 16.Sep.2026.

Please cite as:
Bujek I
Is This Still Medical Education? The Educational Use of Nude Pediatric Clinical Photographs: Bioethical Reassessment
JMIR Med Educ 2026;12:e104023
URL: https://mededu.jmir.org/2026/1/e104023
doi:10.2196/104023
PMID:

©Iwona Bujek. Originally published in JMIR Medical Education (https://mededu.jmir.org), 16.Sep.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Education, is properly cited. The complete bibliographic information, a link to the original publication on https://mededu.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Viewpoint

The AWARE (AI Use, Why, Attachment, Reality and Risk, and Effect on Functioning) Framework for Assessing Patients' Use of AI in Mental Health Care

Alexandre Hudon[1,2,3,4,5], BEng, MD, MSc(ed), PhD; Venkat Bhat[6,7], MSc, MD; John-Jose Nunez[8,9], MSc, MD; Vincent Paquin[10,11], MD, PhD; Amylie Malouin-Lachance[1,12], MD; François Lespérance[1,13,14], MD

[1]Department of Psychiatry and Addictology, Faculty of Medicine, Université de Montréal, Montreal, QC, Canada
[2]Forensic psychiatry services, Department of Psychiatry, Institut universitaire en santé mentale de Montréal, Montreal, QC, Canada
[3]Centre de recherche de l'Institut universitaire en santé mentale de Montréal, Montreal, QC, Canada
[4]Department of Psychiatry, Institut national de psychiatrie légale Philippe-Pinel, Montreal, QC, Canada
[5]Centre de recherche en pédagogie de la santé, Université de Montréal, Montreal, QC, Canada
[6]Department of Psychiatry, University of Toronto, Toronto, ON, Canada
[7]Department of Psychiatry, AI for Mental Health, St. Michael's Hospital, Toronto, ON, Canada
[8]Department of Psychiatry, University of British Columbia, Vancouver, BC, Canada
[9]Department of Computer Science, University of British Columbia, Vancouver, BC, Canada
[10]Department of Psychiatry, McGill University, Montreal, QC, Canada
[11]Lady Davis Institute for Medical Research, Jewish General Hospital, Montreal, QC, Canada
[12]Department of Psychiatry, Faculté de médecine, Université Laval, Québec, QC, Canada
[13]Département de psychiatrie, Centre Hospitalier de l'Université de Montréal, Montreal, QC, Canada
[14]Centre de recherche du Centre hospitalier de l'Université de Montréal, Montreal, QC, Canada

Corresponding Author:
Alexandre Hudon, BEng, MD, MSc(ed), PhD
Department of Psychiatry and Addictology
Faculty of Medicine
Université de Montréal
2900 Bd Edouard-Montpetit
Montreal, QC, H3T 1J4
Canada
Phone: 1 5149954842
Email: alexandre.hudon.1@umontreal.ca

Abstract

Conversational AI has rapidly become integrated into everyday life, with many individuals using large language model–based systems, such as ChatGPT, Gemini, and Claude, for education, productivity, health information, decision-making, emotional support, and companionship. As patients increasingly incorporate conversational AI into their cognitive, emotional, and social lives, these interactions may influence coping strategies, treatment engagement, clinical decision-making, and mental health outcomes. However, although medical education has increasingly emphasized AI literacy and the responsible use of generative AI by clinicians, comparatively little attention has been given to preparing health care professionals to assess patients' use of conversational AI during routine clinical encounters. This Viewpoint introduces the AWARE (AI use, why, attachment, reality and risk, and effect on functioning) framework, a practical educational framework designed to help mental health professionals systematically assess patients' use of conversational AI. Rather than functioning as a diagnostic instrument or psychometric scale, AWARE provides a structured approach to psychiatric interviewing across 5 clinically relevant domains: AI use, why, attachment, reality and risk, and effect on functioning. Together, these domains guide clinicians in exploring patterns of AI use, the motivations underlying engagement, the emotional significance of patient-AI interactions, potential influences on reality testing and clinical risk, and the overall impact of AI on psychological well-being, daily functioning, relationships, and recovery. The framework is intended to complement existing psychiatric interviewing practices by supporting comprehensive history taking, clinical reasoning, risk assessment, documentation, and learner education. We discuss the rationale for routinely asking patients about AI use, review emerging evidence regarding both the potential benefits and risks of conversational AI, and describe how AWARE may be

XSL•FO
RenderX

incorporated into undergraduate, postgraduate, and continuing professional education through simulation, objective structured clinical examinations, workplace-based assessment, and clinical supervision. We also outline priorities for future research, including validation, implementation, educational evaluation, and cross-cultural adaptation. As conversational AI becomes increasingly embedded within patients' daily lives, clinicians require practical approaches to understanding its role in mental health and health care. The AWARE framework offers a structured educational starting point for integrating assessment of patient-AI interactions into routine psychiatric interviewing while supporting patient-centered, evidence-informed clinical practice and health professions education.

(JMIR Med Educ 2026;12:e108766) doi:10.2196/108766

KEYWORDS

artificial intelligence; AI; conversational artificial intelligence; large language models; ChatGPT; psychiatry; mental health; psychiatric interview; clinical assessment; medical education; digital mental health

Introduction

The Rapid Integration of AI Into Everyday Life

The public release of large language model (LLM)–based systems, such as ChatGPT (OpenAI), Gemini (Alphabet Inc), and Claude (Anthropic PBC), has changed how people interact with AI. Unlike earlier narrowly defined applications, conversational AI supports natural language dialogue across education, problem-solving, decision-making, and personal concerns and has rapidly become integrated into everyday life [1,2].

Conversational AI nevertheless encompasses heterogeneous technologies that should not be treated as interchangeable. General-purpose LLM assistants are designed for broad conversational use, whereas purpose-built mental health agents such as Woebot (Woebot Labs Inc) and Wysa (Touchkin Eservices Pvt Ltd) deliver structured psychological or well-being interventions, and companion chatbots such as Replika (Luca Inc) emphasize sustained relational interaction. Evidence from 1 category should therefore not be assumed to generalize to another. In this Viewpoint, we distinguish these technologies when interpreting evidence while using "conversational AI" as an umbrella term for systems that interact with users through dialogue.

Use of conversational AI has also expanded beyond productivity and education [1,3] to emotional support, companionship, self-reflection, health information, and personal decision-making. In a survey of 1006 university students using Replika, participants described the companion chatbot in overlapping roles, including friend, therapist, and intellectual mirror, and a small subgroup reported that interactions had interrupted suicidal thinking [4]. These findings are specific to a companion-oriented system and should not be generalized to general-purpose LLMs. Surveys of ChatGPT users similarly indicate growing use for health-related questions and health care decision-making, with perceived trustworthiness influencing reliance on AI-generated recommendations [5].

These patterns are clinically relevant because patients may discuss symptoms with AI before seeing a clinician, seek emotional support during distress, or incorporate AI-generated information into their understanding of mental health and treatment. Understanding how patients use and interpret these systems is therefore becoming increasingly relevant to contemporary psychiatric assessment.

AI Represents a Fundamentally Different Digital Exposure

The adoption of conversational AI represents a fundamentally new type of digital exposure. Earlier technologies, including search engines and social media, primarily provided access to information or facilitated interactions among people. Although these technologies can influence beliefs, emotions, and behavior, they do not themselves function as conversational partners. In contrast, LLM-based conversational agents engage users directly in dialogue, generating personalized responses that adapt to language, preferences, and conversational context. Their ability to sustain coherent, context-sensitive exchanges creates a qualitatively different interaction in which AI becomes an active conversational participant.

Several features distinguish conversational AI from earlier digital technologies. Rather than relying on keywords or predefined commands, these systems support natural language interaction and iterative multiturn dialogue in which exchanges evolve dynamically. Their personalization and continuous availability further make interactions more intuitive and responsive. They can also generate responses that users may experience as empathic, supportive, validating, or emotionally responsive [6-9]. Such perceived relational qualities should not be equated with human empathy or reciprocal interpersonal understanding; rather, they reflect the capacity of conversational systems to generate context-sensitive language that can be experienced as socially and emotionally meaningful.

Evidence for perceived therapeutic alliance comes primarily from purpose-built mental health conversational agents rather than contemporary general-purpose LLMs. Users of Wysa, for example, reported working alliance scores comparable with those reported in previous studies of human-delivered and internet-based psychotherapy [6]. Systematic reviews similarly suggest that relational factors may contribute to engagement and user experience, although the included technologies are heterogeneous and findings should not be extrapolated directly to general-purpose LLM systems [10].

Experimental studies further suggest that meaningful relational processes can emerge when interactions are perceived as chatbot- or AI-mediated. Ho et al [7] found broadly similar effects of emotional self-disclosure when participants believed

XSL•FO
RenderX

they were conversing with a chatbot or another person. Barger [8] found that participants could report a working alliance with a coach they believed to be AI; however, the study used a Wizard-of-Oz design in which responses were actually generated by professional human coaches. This finding therefore informs the role of perceived agent identity rather than demonstrating that an autonomous AI system can independently establish such an alliance. In another experimental study, chatbot emotional disclosure was associated with greater perceived intimacy, user satisfaction, and intention to reuse the service [9]. Together, these findings indicate that relational experiences can emerge in interactions perceived as AI-mediated, while underscoring that evidence from experimental and purpose-built systems should not be generalized directly to contemporary general-purpose LLMs.

From a psychiatric perspective, this matters because conversational AI can simulate interpersonal behaviors traditionally associated with human relationships, including empathy, validation, encouragement, and reassurance. Although these interactions should not be assumed to involve reciprocal human-like understanding, users may nevertheless experience them as socially or emotionally meaningful, with implications for beliefs, emotions, coping, and decision-making. Conversational AI may therefore be understood not simply as another digital technology but as a novel relational environment. Earlier digital platforms have long supported parasocial and other psychologically meaningful interactions, but conversational AI introduces a more direct, adaptive, and personalized form of engagement whose effects on psychological functioning are only beginning to be understood.

Emerging Psychiatric Implications

Conversational AI presents both opportunities and challenges for mental health care. Purpose-built conversational agents can provide psychoeducation, cognitive behavioral therapy (CBT)–based exercises, symptom monitoring, and coping strategies in accessible and scalable formats [10-14]. Evaluations of systems such as Woebot and Wysa suggest potential benefits for depressive and anxiety symptoms, acceptability, and engagement [11-13], supporting their possible role as adjuncts to care, particularly when access to professional services is limited.

The same features that make conversational AI accessible and engaging may also create clinically relevant risks. These systems can sustain personalized, emotionally meaningful interactions that encourage repeated disclosure and may become sources of reassurance, companionship, validation, or therapeutic guidance [6-10]. Case reports describe reinforcement of delusional beliefs and unsafe health-related decision-making [15,16], while other studies raise concerns about emotional reliance, dependency, and displacement of human relationships [4,8,17]. General-purpose LLMs also vary in their responses to suicidal crises, with some responses diverging from established prevention principles [18,19]. Evidence regarding violence-related interactions remains limited.

The evidence base remains preliminary and heterogeneous, consisting largely of feasibility studies, observational research, cross-sectional surveys, and case reports [10,20]. Conversational AI should therefore not be regarded as inherently beneficial or harmful. Rather, its clinical significance is likely to depend on the patient's motivations, pattern of use, vulnerabilities, social context, and the ways AI interacts with symptoms, coping, interpersonal functioning, and treatment engagement.

The Educational Gap

Although medical education has rapidly embraced generative AI as a tool for teaching, learning, and clinical decision support, considerably less attention has been devoted to preparing clinicians to assess patients' own AI use during routine clinical encounters [21,22]. Existing educational initiatives primarily focus on AI literacy, ethical use of AI, prompt engineering, and the incorporation of AI into medical curricula rather than its implications for psychiatric assessment or patient care [21,22]. Consequently, clinicians may be well prepared to use AI themselves while remaining unfamiliar with how patients incorporate conversational AI into their emotional, cognitive, and social lives.

This educational gap is particularly evident in psychiatric interviewing. Existing psychiatric assessment frameworks and clinical guidance increasingly encourage clinicians to explore domains such as substance use, digital behaviors, internet and social media use, interpersonal relationships, and coping strategies because each may influence diagnosis, risk assessment, and treatment planning [20,23]. However, these frameworks were largely developed before the widespread adoption of conversational agents and provide limited practical guidance for assessing patients' interactions with generative AI. Although emerging publications have begun to address this issue, no widely adopted or standardized approach has yet been incorporated into routine psychiatric assessment. Consequently, clinicians may overlook clinically relevant information regarding the frequency, purpose, emotional significance, and influence of AI interactions despite their growing role in patients' everyday lives.

Why Clinicians Should Ask About AI

As conversational AI becomes integrated into everyday life, patients may use it not only for information but also for emotional support, symptom interpretation, and health-related decision-making. This creates an important parallel with clinicians, who increasingly encounter generative AI for learning, information synthesis, and clinical decision support [5,21,22]. Patients may similarly use these systems to interpret symptoms, consider possible diagnoses, prepare for clinical encounters, or weigh treatment decisions. Clinicians therefore need practical ways to explore not simply whether patients use AI, but how AI-generated information and interactions have entered their understanding of health, relationships, and care.

Objective

The objective of this Viewpoint is to operationalize emerging recommendations that clinicians assess patients' use of conversational AI by introducing the AWARE (AI use, why, attachment, reality and risk, and effect on functioning) framework, a practical educational tool that organizes this assessment into 5 clinically relevant domains and provides structured interview questions for their exploration. AWARE

XSL•FO
RenderX

organizes this inquiry across 5 clinically relevant domains: *AI use*, *why*, *attachment*, *reality and risk*, and *effect on functioning*. Although primarily intended for education and training, the framework may also serve as a flexible interviewing guide when a patient's AI use is disclosed or judged clinically relevant. It is not intended as a screening instrument, diagnostic tool, risk-prediction instrument, or standardized clinical protocol. Rather than assuming that AI use is either beneficial or harmful, AWARE encourages clinicians and learners to consider how conversational AI may interact with an individual patient's mental health, functioning, relationships, and care.

Although the acronym AWARE has previously been used by Lombardo et al [24] for a framework addressing clinical decision-making in crisis resolution and home treatment teams, the present framework is conceptually and operationally distinct and focuses specifically on assessing patients' interactions with conversational AI. Recent work has begun to explicitly recommend that clinicians ask patients about AI use and evaluate potentially risky patterns of engagement [25]. AWARE builds on these recommendations by providing a mnemonic structure and concrete interview questions that operationalize such assessment in clinical practice.

AI Is Becoming Part of Patients' Social Environments

Patients increasingly use conversational AI for purposes that extend beyond information retrieval. Depending on individual needs and circumstances, AI may function as a teacher for learning new concepts, a coach for improving performance or productivity, a source of health information, a decision aid for navigating complex choices, or a conversational partner for emotional support and self-reflection [1-5]. Some individuals describe AI as a confidant with whom they share experiences they would hesitate to disclose to others. In contrast, others report using conversational agents in ways that resemble psychotherapy or companionship [4-10].

The relational affordances of conversational systems vary substantially by design. Companion agents may be optimized for sustained relational engagement, purpose-built mental health chatbots may structure interactions around therapeutic or well-being content, and contemporary general-purpose LLMs can generate highly personalized, context-sensitive dialogue across unrestricted topics. Despite these technological differences, users may experience interactions with each category as emotionally meaningful. The strength, nature, and clinical implications of these experiences should therefore be evaluated at the level of the specific system and pattern of use rather than generalized across conversational AI technologies.

From a psychiatric perspective, AI should therefore be viewed not only as a technology but also as a novel relational environment that may shape patients' thoughts, emotions, and behaviors. Understanding how an individual patient incorporates AI into daily life provides important contextual information that may not otherwise emerge during routine psychiatric assessment.

AI May Influence Clinical Presentation

Conversational AI may interact with existing psychiatric symptoms and vulnerabilities in both adaptive and maladaptive ways. It may provide psychoeducation, coping support, or emotional expression, particularly where access to care is limited [10-14]. Conversely, emerging reports suggest that AI may become involved in reassurance seeking, social substitution, delusional belief systems, treatment decisions, or crisis-related interactions [4,15,17-19]. These observations should not be interpreted as establishing AI as an independent cause of psychopathology. Rather, AI use represents a contextual factor whose clinical significance depends on the patient's vulnerabilities, motivations, social circumstances, and pattern of use.

AI Use Should Be Considered During Psychiatric Assessment

Psychiatric interviewing has continually evolved to reflect changes in society and technology. Clinicians now routinely ask about substance use, internet behaviors, social media use, sleep, sexual health, and other lifestyle factors because these domains are important for patients and can guide diagnosis, risk assessment, and treatment planning. These questions are not asked because they presume a pathological behavior, but because they help clinicians understand patients within the broader context of their lives.

Conversational AI increasingly warrants similar consideration. A patient who occasionally uses ChatGPT to summarize lecture notes presents different clinical considerations than someone who spends several hours each day seeking emotional reassurance from an AI companion or who incorporates AI-generated responses into fixed beliefs about themselves or the world. Without asking about AI use, clinicians may overlook important influences on emotional well-being, interpersonal functioning, treatment adherence, and clinical risk.

Routine assessment of AI use should therefore be viewed as an extension of comprehensive psychiatric history taking rather than the creation of an entirely new assessment domain. The objective is neither to pathologize AI use nor to assume that it is beneficial, but rather to understand how conversational AI influences an individual patient's psychological functioning and clinical care. As with other components of the psychiatric interview, the depth of AI-related assessment should be guided by clinical relevance, presenting concerns, and available interview time rather than applied uniformly to every patient. To facilitate this process, we propose the AWARE framework as a flexible, structured approach that can be incorporated into routine mental health encounters when clinically indicated.

Conceptual Development of the AWARE Framework

AWARE was developed as an author-proposed conceptual interviewing framework rather than through a formal qualitative, Delphi, or psychometric development process. The 5 domains were organized through a narrative synthesis of the clinical issues identified in the literature reviewed above and according to their distinct functions within psychiatric interviewing. *AI use* establishes the nature and extent of exposure; *why* examines the motivations and needs underlying use; *attachment* addresses the relational and emotional significance of interactions with conversational AI; *reality and risk* examines whether these interactions influence beliefs, judgment, treatment decisions,

XSL•FO
RenderX

or safety; and *effect on functioning* evaluates their overall consequences for daily life, relationships, treatment, and recovery.

The 5 AWARE domains are intended as mnemonic domains of inquiry rather than mutually exclusive psychological constructs. Some overlap is therefore expected. For example, motivation for AI use may relate to attachment, trust may be relevant to both relational significance and appraisal of AI-generated information, and functional consequences may arise from patterns identified in several earlier domains. This redundancy is intentional insofar as AWARE is designed to support clinical interviewing rather than construct measurement. The framework should therefore not be interpreted as proposing 5 orthogonal or psychometrically distinct dimensions of patient-AI interaction.

Attachment was retained as a distinct domain because an emotionally meaningful relationship with conversational AI may be clinically relevant even in the absence of impaired reality testing, overt risk, or functional deterioration. Conversely, reality testing and risk were grouped within a single domain for clinical parsimony because both represent situations in which AI-generated interactions may warrant escalation from descriptive history taking to established psychiatric assessment of judgment, safety, and clinical risk. Their combination does not imply that impaired reality testing and risk are equivalent or necessarily co-occur.

Accordingly, the present 5-domain structure should be understood as a theoretically and clinically informed heuristic proposed for further evaluation, rather than as an empirically established taxonomy of patient-AI interactions. Formal content-validation work involving clinicians, patients, and other relevant stakeholders will be required to determine whether these domains are sufficiently comprehensive, distinct, and clinically useful.

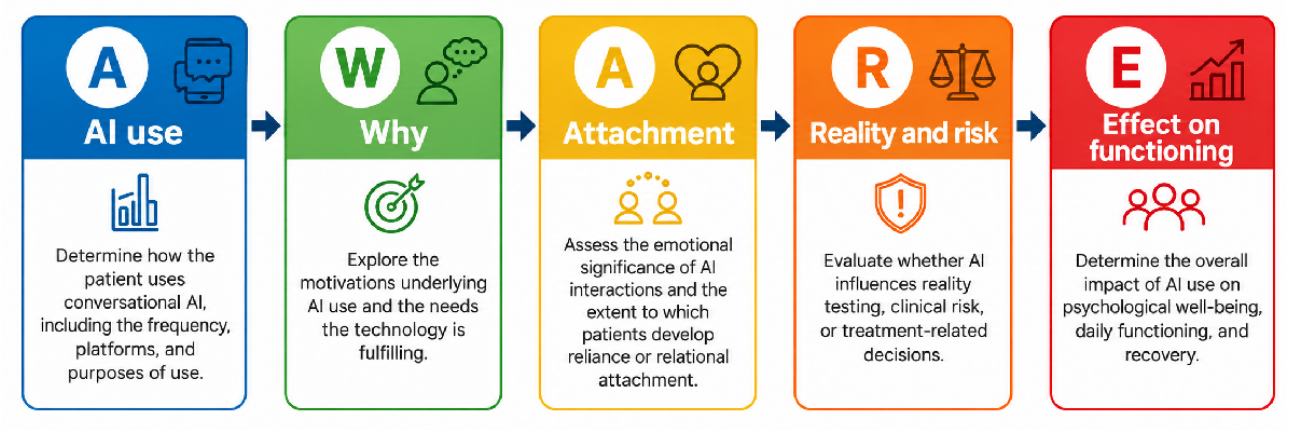
The AWARE Framework

Overview

The AWARE framework is an author-proposed conceptual interviewing framework and educational tool designed to help mental health professionals systematically explore patients' interactions with conversational AI during clinical encounters. Rather than determining whether AI use is inherently beneficial or harmful, AWARE guides clinicians through 5 domains that together characterize the patient's pattern of use, motivations, emotional relationship with AI, potential safety concerns, and overall impact on functioning.

The framework is intended to complement (not replace) existing psychiatric interviewing techniques by providing a practical structure for exploring an increasingly common aspect of patients' digital lives. Although the domains are presented sequentially, clinicians may adapt their order according to the clinical context. An overview of the framework is shown in Figure 1.

Figure 1. The AWARE (AI use, why, attachment, reality and risk, and effect on functioning) framework: a structured approach to assessing patients' use of conversational AI.



AI Use

Educational Objective

The educational objective for the *AI use* domain is to determine the patient's exposure to conversational AI, including which platforms they use, how frequently and in what contexts they engage with them, the modalities through which they interact, and whether sensitive personal, health-related, or third-party information is shared.

Why It Matters

The nature of information shared with conversational AI may also be clinically relevant even when the pattern of use is otherwise adaptive. Patients may enter sensitive psychiatric symptoms, trauma histories, medication information, personal identifiers, medical documents, images, or information concerning family members and other third parties. Data handling, retention, secondary use, and privacy protections vary across platforms and may not be fully understood by users. A clinical AI history can therefore include a brief exploration of sensitive data disclosure without assuming that such disclosure reflects problematic AI use.

Suggested Questions

Questions suggested for the *AI use* domain are as follows:

XSL•FO
RenderX

- What types of personal or health information do you share with AI?
- Have you ever uploaded medical records, reports, images, or other health documents?
- Do your conversations ever include identifiable or sensitive information about other people?
- How do you decide what information you are comfortable sharing with an AI system?

Clinicians should characterize frequency of use, duration, and intensity. Occasional educational use may have little clinical significance, whereas prolonged daily interactions or exclusive reliance on AI for emotional support may warrant further exploration. Exposure itself should not be viewed as problematic but rather as contextual information that guides subsequent assessment.

Why

Educational Objective

The educational objective for the *why* domain is to understand the motivations underlying AI use.

Why It Matters

The reason patients engage with AI is often more clinically informative than the amount of use itself. Similar patterns of use may have very different clinical implications depending on whether AI is being used for education, productivity, emotional support, companionship, or coping with psychological distress. Motivation helps clinicians understand the role AI occupies in relation to the patient's broader psychological and social needs.

Suggested Questions

Questions suggested for the *why* domain are as follows:

- What do you find most helpful about AI?
- Can you tell me about the situations in which you choose to talk with AI or another person?
- What role, if any, does AI play when you are dealing with stress or difficult emotions?
- What needs does AI meet that other resources do not?

Clinicians should differentiate between motivations such as education, work productivity, creativity, health information, therapy or emotional support, coping, and companionship. Educational and productivity uses generally require little further exploration. Uses centered on emotional regulation, loneliness, reassurance seeking, companionship, or interpersonal needs may warrant further contextual exploration, particularly when they are associated with distress, unwanted displacement of other supports, or functional consequences.

Attachment

Educational Objective

The educational objective for the *attachment* domain is to assess the emotional relationship between the patient and conversational AI.

Why It Matters

One of the defining characteristics of conversational AI is its capacity to support interactions that users may experience as relationally meaningful. Several distinct constructs may help characterize these interactions, including perceived therapeutic alliance, anthropomorphism, trust, companionship, parasocial interaction, emotional reliance, and attachment-related experiences. Although these constructs may overlap, they describe different aspects of patient-AI interaction and should not be treated as equivalent. Clinically, the purpose of this domain is therefore not to classify a patient along a single attachment trajectory, but to understand the emotional and relational significance that AI has acquired for that individual.

Although AI systems do not possess genuine emotions or reciprocal relationships, patients may nevertheless experience trust, comfort, companionship, or emotional security during repeated interactions.

Suggested Questions

Questions suggested for the *attachment* domain are as follows:

- How would you describe your relationship with AI?
- What is it like when you are unable to use AI?
- How would you describe your experience of being understood when interacting with AI?
- Can you describe how you decide when to trust AI and when to trust other people?
- Are there situations where you had personal conversations with AI? What happened?

Clinical Interpretation

The relational significance of conversational AI should not be reduced to a single continuum or assumed to progress through fixed stages. Different phenomena may be clinically relevant, including perceived therapeutic alliance, anthropomorphism, trust, companionship, emotional reliance, and dependence. These constructs overlap conceptually but are not interchangeable and may occur independently. Clinicians should therefore explore the nature, intensity, and functional significance of the patient's relationship with AI without assuming that preference for AI necessarily reflects pathology or predicts increasing dependence.

Reality and Risk

Educational Objective

The educational objective for the *reality and risk* domain is to identify whether AI interactions raise concerns requiring more focused assessment of (1) reality appraisal and judgment or (2) established areas of psychiatric risk and safety.

Why It Matters

The domain of *reality and risk* combines 2 clinically related but conceptually distinct areas of inquiry. They are grouped for mnemonic and interviewing purposes rather than because they represent a single psychological construct. The first concerns whether AI-generated content has become involved in the patient's appraisal of reality, beliefs, or important decisions. The second concerns whether AI interactions have influenced established areas of psychiatric safety, such as suicidal behavior, self-harm, violence, treatment decisions, or unsafe health behaviors. In either case, positive findings should prompt

XSL•FO
RenderX

assessment using established clinical approaches rather than interpretation through AWARE alone.

Reality Appraisal and Judgment

Questions suggested for the *reality and risk* domain under reality appraisal and judgment are as follows:

- How, if at all, have your conversations with AI influenced the way you think about yourself?
- How, if at all, have your conversations with AI influenced important decisions in your life?
- How would you compare your experience of being understood by AI with your experience of being understood by other people?
- Can you tell me about any situations in which AI responded differently to your beliefs or experiences than other people did?
- Have you ever used AI to understand what is real and what is not?
- How do you decide whether the information or advice provided by AI is accurate or trustworthy?

Clinicians should explore whether AI has become incorporated into self-identity, fixed beliefs, grandiosity, persecutory ideas, and delusional systems. Most patients will answer negatively. However, when positive responses occur, they should be interpreted within the broader clinical context rather than attributed solely to AI.

AI Involvement in Risk and Safety

AI may also become relevant during assessment of suicidal ideation, self-harm, violence, treatment refusal, medication discontinuation, and unsafe health behaviors. It is to be noted that AWARE does not assess suicide, self-harm, or violence risk. Questions in this domain are intended only to determine whether conversational AI has entered the context of an established area of psychiatric risk, for example, whether a patient has discussed suicidal thoughts with AI, sought crisis support from it, or reported that AI influenced a safety-related decision. Any concern regarding suicide, self-harm, violence, or other acute risk requires assessment using established clinical approaches independent of AWARE.

Questions suggested for the *reality and risk* domain under AI involvement in risk and safety are as follows:

- Can you tell me about any conversations you have had with AI about thoughts of suicide or self-harm?
- Can you tell me about any times you turned to AI during a crisis and what happened?
- What role, if any, has AI played in decisions about your treatment or medications?
- How, if at all, has AI influenced your decisions about seeking professional help?

Positive responses should prompt standard psychiatric risk assessment. AI should be understood as 1 component of a patient's environment that may either mitigate or exacerbate existing vulnerabilities rather than as an independent cause of harmful behavior.

Effect on Functioning

Educational Objective

The educational objective for the *effect on functioning* domain is to evaluate the overall impact of AI use on the patient's functioning and quality of life.

Why It Matters

The clinical significance of AI use depends on its effects on everyday functioning. Many patients may experience substantial benefits, whereas others may encounter adverse consequences. Assessment should therefore remain balanced and avoid assuming either benefit or harm.

Suggested Questions

Questions suggested for the *effect on functioning* domain are as follows:

- How, if at all, has using AI influenced your daily routine?
- How, if at all, has using AI influenced your work or school performance?
- How, if at all, has using AI influenced your relationships?
- How, if at all, has using AI changed your sleep?
- How, if at all, has using AI influenced your mental health treatment?

Clinicians should evaluate effects on sleep, occupation, education, relationships, daily functioning, and quality of life. Assessment should identify both positive and negative outcomes. AI may improve coping, learning, and access to support while simultaneously contributing to excessive use, interpersonal withdrawal, reassurance seeking, or reduced engagement with treatment. Clinical interpretation should therefore emphasize the overall balance of benefits and harms within the context of the individual patient's circumstances.

The suggested AWARE interview guide is presented in Table 1.

XSL•FO
RenderX

Table 1. The AWARE (AI use, why, attachment, reality and risk, and effect on functioning) interview guide.

Domains	Why assess?	Example questions	Potential concerns
AI use	Characterize exposure, context, and information sharing	Which AI tools do you use? How often? What types of personal or health information do you share?	Sensitive data disclosure, sharing of third-party information, privacy concerns, or intensive or exclusive use
Why	Understand motivations	What do you use AI for? Why do you prefer it?	Emotional coping, reassurance seeking, or social replacement
Attachment	Evaluate emotional relationship	Do you feel understood by AI? Do you rely on it emotionally?	Emotional reliance, dependence, anthropomorphism, or substitution for human relationships
Reality and risk	Assess influence on beliefs and safety	Has AI changed your beliefs or treatment decisions? Have you discussed suicide or self-harm with AI?	Delusion reinforcement, unsafe advice, treatment refusal, or crisis management
Effect on functioning	Evaluate clinical significance	Has AI affected your sleep, work, relationships, or recovery?	Functional impairment, treatment refusal, social withdrawal, or meaningful improvements in coping and functioning

Documentation

When AI use emerges as clinically relevant, documentation may briefly capture the pattern of use, its psychological significance, and any identified benefits, concerns, or implications for care. AWARE may provide learners with a structure for organizing such documentation without prescribing a standardized documentation requirement. Documentation should describe the pattern of AI use, its psychological significance, and any identified risks or benefits.

A statement such as “uses ChatGPT” provides little clinically useful information. A more useful example would be as follows: “Patient uses ChatGPT approximately 1–2 hours daily for emotional support and problem solving. Reports feeling understood by the chatbot but does not describe preferring AI over interpersonal relationships. Patient reports no AI-related changes in important beliefs or treatment decisions and denies using AI for support related to suicidal or self-harm thoughts. Reports improved coping with anxiety and no perceived impairment in occupational, social, or academic functioning.”

This useful documentation describes the reported role of AI in the patient’s life; it should not be interpreted as documenting the absence of suicide risk, impaired reality testing, or other psychopathology. When clinically indicated, these areas require separate assessment and documentation using established psychiatric methods.

Educational Implementation

AWARE is intended primarily as an educational framework across the continuum of health professions education. Its educational objective is to help learners recognize when a patients’ use of conversational AI may be clinically relevant and provide a structured approach to exploring it. The framework does not imply that every patient requires comprehensive AI-related questioning; rather, learners should develop the judgment to determine when brief inquiry is sufficient and when more detailed exploration is warranted.

At the undergraduate level, AWARE may introduce AI history taking alongside established psychosocial components of the clinical interview. In postgraduate psychiatry training, it may support more advanced formulation by encouraging learners to consider how AI interacts with symptoms, coping, relationships, treatment engagement, and safety. The same principles may be adapted for psychologists, nurses, social workers, family physicians, and practicing clinicians according to professional scope and clinical context (Table 2).

Table 2. Educational applications of the AWARE (AI use, why, attachment, reality and risk, and effect on functioning) framework.

Learner groups	Educational goals	Teaching methods
Medical students	Introduce AI history taking	Lectures, case-based learning, and simulation
Psychiatry residents	Advanced formulation and risk assessment	Clinical supervision, objective structured clinical examinations, and workplace-based assessment
Psychologists	AI and therapeutic processes	Case discussion and reflective practice
Nurses	Routine mental health assessment	Simulation and bedside teaching
Social workers	Social context and support systems	Case-based learning and supervision
Family physicians	Screening and referral	CPD[a] workshops and workplace learning
Practicing clinicians	Updating interviewing skills	CPD

[a]CPD: continuing professional development.

Educational strategies may include case-based learning, simulation, objective structured clinical examinations (OSCEs), clinical supervision, and workplace-based assessment. For example, standardized patients could describe AI use for

XSL•FO
RenderX

emotional support, health-related decision-making, or during psychological crises, allowing learners to practice identifying when additional clinical assessment is warranted. In workplace settings, existing approaches, such as the mini–clinical evaluation exercise (mini-CEX) or case-based discussion, could incorporate the ability to obtain an AI history, interpret its relevance, and document findings appropriately. These applications concern the teaching and assessment of interviewing skills and should not be interpreted as evidence that AWARE itself has been clinically validated.

Because conversational technologies continue to evolve, AWARE is intended as a flexible educational scaffold rather than a fixed checklist or competency standard.

Future Directions

AWARE should be regarded as an educational interviewing framework rather than a diagnostic or psychometric instrument. It has not undergone formal development or validation, and its 5 domains should therefore be considered provisional. Initial research should establish content validity through structured input from clinicians, patients, and other stakeholders and determine whether the domains adequately capture clinically relevant patient-AI interactions. Studies should subsequently evaluate feasibility, acceptability, educational utility, and whether AWARE improves the completeness or clinical relevance of AI-related history taking.

Only if these initial evaluations are supportive should broader implementation or development of a standardized assessment instrument be considered. Future work should also examine cross-cultural applicability and whether emerging issues, including privacy, accessibility, and changing forms of patient-AI interaction, require modification of the framework.

Conclusions

Conversational AI is rapidly becoming part of patients' cognitive, emotional, and social lives. As individuals increasingly use AI for education, decision-making, emotional support, companionship, and mental health–related concerns, understanding these interactions is becoming an important component of contemporary psychiatric assessment. Despite this shift, clinicians currently receive little guidance on how to systematically explore patients' AI use during routine clinical encounters.

The AWARE framework provides a structured educational approach to addressing this gap. By guiding clinicians through 5 domains (*AI use*, *why*, *attachment*, *reality and risk*, and *effect on functioning*), the framework supports comprehensive history taking, clinical reasoning, risk assessment, documentation, and learner education without assuming that AI use is inherently beneficial or harmful. Instead, it encourages individualized assessment that considers how conversational AI interacts with each patient's symptoms, coping strategies, relationships, and overall functioning.

Although AWARE has not yet been formally validated and should not be considered a diagnostic, screening, or risk assessment instrument, it offers an educational starting point for preparing clinicians to recognize and explore clinically relevant patient-AI interactions. Its use in clinical practice should remain flexible and guided by presenting concerns rather than applied as a universal or comprehensive assessment. As conversational AI becomes increasingly embedded in everyday life, empirical evaluation will be needed to determine whether structured approaches such as AWARE improve education, interviewing, and patient care.

Funding

This study was funded indirectly by La Fondation de l'Institut universitaire en santé mentale de Montréal and the operating funds of l'Institut de valorisation des données (IVADO). AH is also funded by a Clinical Research Scholars J1 grant from the Fonds de recherche du Québec [26].

Authors' Contributions

AH was responsible for the conceptualization of the manuscript. All authors (AH, VB, VP, JJN, AM-L, and FL) contributed to the writing, critical review, and editing of the manuscript and approved the final version for publication.

Conflicts of Interest

VB is supported by an academic scholar award from the University of Toronto, Department of Psychiatry, and has received research funding from the Canadian Institutes of Health Research, Brain & Behavior Foundation, Ontario Ministry of Health Innovation Funds, Royal College of Physicians and Surgeons of Canada, Department of National Defense (government of Canada), New Frontiers in Research Fund, Associated Medical Services Inc Healthcare, American Foundation for Suicide Prevention, Roche Canada, Novartis, and Eisai. AM-L has received research funding from the Canadian Institutes of Health Research. All other authors declare no other conflicts of interest.

References

1. Brynjolfsson E, Li D, Raymond L. Generative AI at work. Q J Econ 2025 May;140(2):889-942. [doi: 10.1093/qje/qjae044]
2. Tao W, Yang J, Qu X. Utilization of, perceptions on, and intention to use AI chatbots among medical students in China: national cross-sectional study. JMIR Med Educ 2024 Oct 28;10:e57132 [FREE Full text] [doi: 10.2196/57132] [Medline: 39466038]

XSL•FO
RenderX

3. Guirguis M, Fotsing S, Fevry J, Landry C, Bouchard-Lamothe D, Lacroix J, et al. Artificial intelligence in health professions education: qualitative study of student experiences. J Med Internet Res 2026 Apr 02;28:e82432 [FREE Full text] [doi: 10.2196/82432] [Medline: 41926317]
4. Maples B, Cerit M, Vishwanath A, Pea R. Loneliness and suicide mitigation for students using GPT3-enabled chatbots. Npj Ment Health Res 2024 Jan 22;3(1):4 [FREE Full text] [doi: 10.1038/s44184-023-00047-6] [Medline: 38609517]
5. Choudhury A, Elkefi S, Tounsi A. Exploring factors influencing user perspective of ChatGPT as a technology that assists in healthcare decision making: a cross sectional survey study. PLoS One 2024 Mar 8;19(3):e0296151 [FREE Full text] [doi: 10.1371/journal.pone.0296151] [Medline: 38457373]
6. Beatty C, Malik T, Meheli S, Sinha C. Evaluating the therapeutic alliance with a free-text CBT conversational agent (Wysa): a mixed-methods study. Front Digit Health 2022 Apr 11;4:847991 [FREE Full text] [doi: 10.3389/fdgth.2022.847991] [Medline: 35480848]
7. Ho A, Hancock J, Miner AS. Psychological, relational, and emotional effects of self-disclosure after conversations with a chatbot. J Commun 2018 Aug;68(4):712-733 [FREE Full text] [doi: 10.1093/joc/jqy026] [Medline: 30100620]
8. Barger AS. Artificial intelligence vs. human coaches: examining the development of working alliance in a single session. Front Psychol 2025 Apr 15;15:1364054 [FREE Full text] [doi: 10.3389/fpsyg.2024.1364054] [Medline: 40313368]
9. Park G, Chung J, Lee S. Effect of AI chatbot emotional disclosure on user satisfaction and reuse intention for mental health counseling: a serial mediation model. Curr Psychol (Forthcoming) 2022 Nov 09:1-11 [FREE Full text] [doi: 10.1007/s12144-022-03932-z] [Medline: 36406852]
10. Li H, Zhang R, Lee YC, Kraut RE, Mohr DC. Systematic review and meta-analysis of AI-based conversational agents for promoting mental health and well-being. NPJ Digit Med 2023 Dec 19;6(1):236 [FREE Full text] [doi: 10.1038/s41746-023-00979-5] [Medline: 38114588]
11. Fitzpatrick KK, Darcy A, Vierhile M. Delivering cognitive behavior therapy to young adults with symptoms of depression and anxiety using a fully automated conversational agent (Woebot): a randomized controlled trial. JMIR Ment Health 2017 Jun 06;4(2):e19 [FREE Full text] [doi: 10.2196/mental.7785] [Medline: 28588005]
12. Inkster B, Sarda S, Subramanian V. An empathy-driven, conversational artificial intelligence agent (Wysa) for digital mental well-being: real-world data evaluation mixed-methods study. JMIR Mhealth Uhealth 2018 Nov 23;6(11):e12106 [FREE Full text] [doi: 10.2196/12106] [Medline: 30470676]
13. Grové C. Co-developing a mental health and wellbeing chatbot with and for young people. Front Psychiatry 2020 Feb 1;11:606041 [FREE Full text] [doi: 10.3389/fpsyt.2020.606041] [Medline: 33597898]
14. Mauriello ML, Tantivasadakarn N, Mora-Mendoza MA, Lincoln ET, Hon G, Nowruzi P, et al. A suite of mobile conversational agents for daily stress management (Popbots): mixed methods exploratory study. JMIR Form Res 2021 Sep 14;5(9):e25294 [FREE Full text] [doi: 10.2196/25294] [Medline: 34519655]
15. Pierre JM, Gaeta B, Raghavan G, Sarma KV. "You're not crazy": a case of new-onset AI-associated psychosis. Innov Clin Neurosci 2025 Dec 01;22(10-12):11-13 [FREE Full text] [Medline: 41635747]
16. Chung VH, Bernier P, Hudon A. Mass media narratives of psychiatric adverse events associated with generative AI chatbots: rapid scoping review. JMIR Ment Health 2026 Mar 30;13:e93040 [FREE Full text] [doi: 10.2196/93040] [Medline: 41911018]
17. Uğuz Ö, Doğan S, Güngör D. Loneliness and AI chatbot dependence in young adults: the mediating role of perceived social support. BMC Psychol 2026 Jun 09;14(1):1184 [FREE Full text] [doi: 10.1186/s40359-026-04916-5] [Medline: 42260619]
18. Campbell LO, Babb K, Lambie GW, Hayes BG. An examination of generative AI response to suicide inquires: content analysis. JMIR Ment Health 2025 Aug 14;12:e73623 [FREE Full text] [doi: 10.2196/73623] [Medline: 40811811]
19. McBain RK, Cantor JH, Zhang LA, Baker O, Zhang F, Burnett A, et al. Evaluation of alignment between large language models and expert clinicians in suicide risk assessment. Psychiatr Serv 2025 Nov 01;76(11):944-950. [doi: 10.1176/appi.ps.20250086] [Medline: 41174947]
20. Vaidyam AN, Wisniewski H, Halamka JD, Kashavan MS, Torous JB. Chatbots and conversational agents in mental health: a review of the psychiatric landscape. Can J Psychiatry 2019 Jul;64(7):456-464 [FREE Full text] [doi: 10.1177/0706743719828977] [Medline: 30897957]
21. Lee QY, Chen M, Ong CW, Ho CS. The role of generative artificial intelligence in psychiatric education- a scoping review. BMC Med Educ 2025 Mar 25;25(1):438 [FREE Full text] [doi: 10.1186/s12909-025-07026-9] [Medline: 40133891]
22. Rincón EH, Jimenez D, Aguilar LA, Flórez JM, Tapia Á, Peñuela CL. Mapping the use of artificial intelligence in medical education: a scoping review. BMC Med Educ 2025 Apr 12;25(1):526 [FREE Full text] [doi: 10.1186/s12909-025-07089-8] [Medline: 40221725]
23. Sheehan DV, Lecrubier Y, Sheehan KH, Amorim P, Janavs J, Weiller E, et al. The Mini-International Neuropsychiatric Interview (M.I.N.I.): the development and validation of a structured diagnostic psychiatric interview for DSM-IV and ICD-10. J Clin Psychiatry 1998;59 Suppl 20:22-33;quiz 34. [Medline: 9881538]
24. Lombardo C, Santos M, Van Bortel T, Croos R, Arensman E, Kar Ray M. Decision-making in crisis resolution and home treatment teams: the AWARE framework. BJPsych Bull 2019 Apr;43(2):61-66 [FREE Full text] [doi: 10.1192/bjb.2018.94] [Medline: 30451131]

XSL•FO
RenderX

25. Kahane K, Torous J. Psychiatry, social media, artificial intelligence: a call for engagement. J Psychiatr Pract 2026 Jul 01;32(4):209-211. [doi: 10.1097/PRA.0000000000000939] [Medline: 42475380]
26. Hudon A. Artificial intelligence in forensic psychiatry, clinical practice, and the digital world: risk prediction, judicial pedagogy, and digital therapeutic alliance [Article in French]. University of Montreal. 2026. URL: https://repertoire.frq.gouv.qc.ca/offres/ [accessed 2026-09-15]

Abbreviations

AWARE: AI use, why, attachment, reality and risk, and effect on functioning
CBT: cognitive behavioral therapy
LLM: large language model
mini-CEX: mini–clinical evaluation exercise
OSCE: objective structured clinical examination

Edited by T de Azevedo Cardoso; submitted 04.Aug.2026; peer-reviewed by S Wojcik, P Bagla; comments to author 31.Aug.2026; revised version received 08.Sep.2026; accepted 09.Sep.2026; published 16.Sep.2026.

Please cite as:
Hudon A, Bhat V, Nunez JJ, Paquin V, Malouin-Lachance A, Lespérance F
The AWARE (AI Use, Why, Attachment, Reality and Risk, and Effect on Functioning) Framework for Assessing Patients' Use of AI in Mental Health Care
JMIR Med Educ 2026;12:e108766
URL: https://mededu.jmir.org/2026/1/e108766
doi:10.2196/108766
PMID:

©Alexandre Hudon, Venkat Bhat, John-Jose Nunez, Vincent Paquin, Amylie Malouin-Lachance, François Lespérance. Originally published in JMIR Medical Education (https://mededu.jmir.org), 16.Sep.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Education, is properly cited. The complete bibliographic information, a link to the original publication on https://mededu.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Viewpoint

Generative Language Models in Medical Education: From Advent to Entrustment

Shiv Patil[1], MS; Mert Karabacak[2], MD; Sotirios Bisdas[3], MD, MSc, PhD; Konstantinos Margetis[4], MD, PhD

[1]Sidney Kimmel Medical College, Thomas Jefferson University, Philadelphia, PA, United States

[2]Department of Neurosurgery, Loma Linda University, Loma Linda, CA, United States

[3]Department of Translational Neuroscience and Stroke, Institute of Neurology, University College London, London, England, United Kingdom

[4]Department of Neurosurgery, Mount Sinai Hospital, New York, NY, United States

Corresponding Author:
Konstantinos Margetis, MD, PhD
Department of Neurosurgery
Mount Sinai Hospital
8th Floor Annenberg Building, 1468 Madison Ave
New York, NY, 10029
United States
Phone: 1 (212) 241 2377
Email: konstantinos.margetis@mountsinai.org

Abstract

Early discussions of generative language models in medical education emphasized their promise for simulation, digital patients, individualized feedback, learner assessment, health information dissemination, research support, and translation, while also warning about bias, privacy, academic integrity, misinformation, legal ambiguity, and unequal access. Since that first wave, generative AI has moved from novelty to routine exposure for learners, educators, researchers, and institutions. Medical education therefore needs a more mature framework than a catalog of opportunities and risks. This viewpoint argues that the next phase should be organized around educational entrustment: determining which functions can be delegated to AI systems, under what conditions, with what human supervision, and with what evidence of benefit. Building on recent proposals to apply entrustment to AI in health professions education, we operationalize the concept into a graduated, function-level model that specifies which educational functions may be delegated; at what stakes; and with what oversight, assessment, and governance. We classify use cases by educational stakes and AI autonomy and outline implications for assessment redesign, curriculum development, faculty capability, cognitive autonomy, equity, and institutional governance. The central challenge is whether medical schools can integrate these tools in ways that preserve clinical reasoning, professional identity, accountability, and fairness. The next generation of research should move beyond model performance on examinations and evaluate how AI changes learning, judgment, behavior, and patient care.

(JMIR Med Educ 2026;12:e105347) doi:10.2196/105347

KEYWORDS

generative artificial intelligence; medical education; large language models; assessment; AI literacy; governance; curriculum

Introduction: After the Advent

In 2023, generative language models entered medical education at an unprecedented pace. Our earlier viewpoint described a technology that could generate realistic patient scenarios, create digital patients, personalize feedback, assist with evaluation, improve access to health information, support research, and reduce language barriers [1]. It also warned that these benefits were inseparable from risks related to accuracy, bias, privacy, legal responsibility, academic dishonesty, transparency, and the digital divide. That framing remains useful because early enthusiasm often outpaced local policy and evidence. The field has since shifted from initial exploration toward operational decisions about which educational responsibilities can be entrusted to generative AI (GenAI), under what conditions, and with what forms of human oversight.

Several developments explain this shift. First, large language models (LLMs) have demonstrated rapidly improving performance on medical knowledge tasks, including licensing-style examinations and long-form medical question answering [2,3]. Second, generative systems have expanded from text to multimodal inputs and outputs. Models can

XSL•FO
RenderX

increasingly combine language, images, audio, structured data, and conversational interfaces, expanding their relevance to radiology, pathology, physical examination teaching, clinical documentation, communication skills, and simulation [4,5]. Third, medical schools are moving from informal experimentation toward institutional policy, curriculum design, and assessment reform. Reviews and guides now converge on the need for AI literacy, educator development, ethical safeguards, and evaluation within authentic learning environments [6-8]. Fourth, systems have begun to act rather than only respond. Agentic models that retrieve information, draft documentation, and execute multistep tasks move AI from a source of text toward an active participant in clinical and educational workflows, sharpening questions of oversight and accountability.

This viewpoint proposes that "Generative AI in Medical Education 2.0" should be framed as the transition from advent to entrustment. Entrustment is familiar to medical education through competency-based education and entrustable professional activities (EPAs), where responsibility is granted progressively based on ability, context, risk, and supervision [9-11]. A similar logic can help educators decide how to use AI. The relevant question is whether a defined educational function can be assigned to an AI system in a defined context with appropriate oversight.

This translation has already begun. Gin et al [12] proposed repurposing entrustment to appraise the trustworthiness of AI tools across the characteristics of ability, integrity, and benevolence. A parallel reappraisal of the Association of American Medical Colleges' core EPAs has proposed emerging activities suited to an AI-rich environment [13]. We build on this work but shift the unit of analysis from the trustworthiness of a tool to the delegation of a defined educational function. Our contribution is operational: a graduated model that pairs each level of delegation with explicit boundaries on stakes, supervision, assessment, and governance.

Why Entrustment Fits the AI Moment

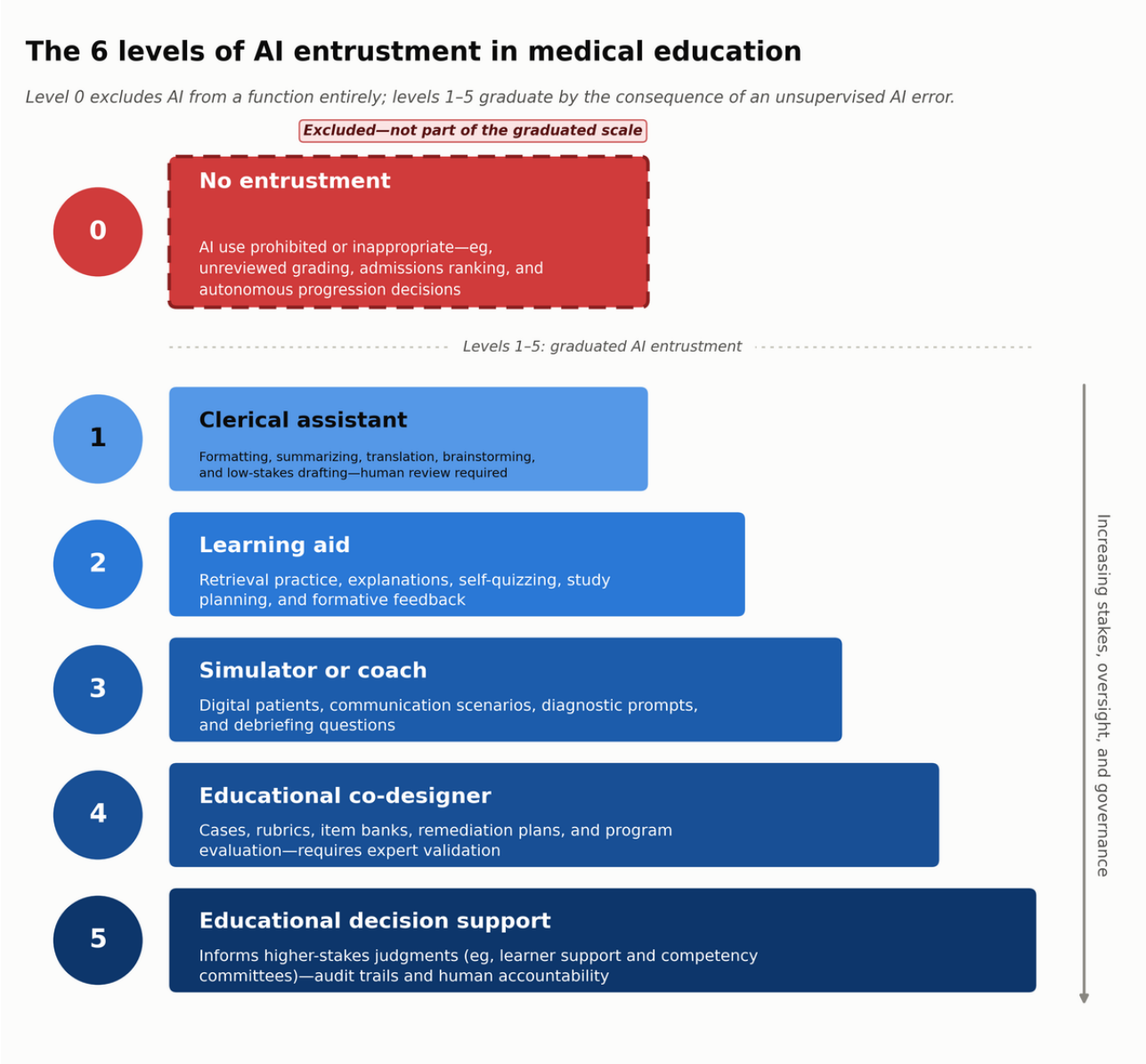
Medical education is structured around graduated responsibility. Learners begin by observing; progress to performing tasks under direct supervision; and then advance to indirect supervision and, ultimately, independent practice. This framework recognizes that competence is contextual: a student may be trusted to take a history in one clinical situation and require close supervision in another. The same logic should apply to AI systems. A tool may be appropriate for generating alternative explanations of a concept and inappropriate for making a progression decision about a learner. A chatbot may be useful for low-stakes history-taking practice and unsafe as the sole source of feedback on professional communication. A model may pass examination questions and still fail when confronted with local protocols, ambiguous patient narratives, incomplete information, or hidden bias. This graduated logic mirrors established learning theories: cognitive apprenticeship and scaffolding describe how expert support is calibrated to a learner's current ability, the gradual release of responsibility describes how instructors progressively transfer cognitive work from teacher to learner, and distributed cognition and epistemic agency describe how thinking can be distributed across humans and tools without necessarily diminishing the learner's own agency. Entrustment in the AI era can therefore be understood not only as a governance mechanism but also as an application of these established learning theories to a new class of cognitive tools.

The analogy between students and AI tools has an important limitation that must be recognized. Entrusting a trainee is developmental: trust grows because the learner learns from supervision and remains morally accountable for outcomes. An AI system does neither. It does not learn from a supervisor's correction, carries no accountability, and may regress without warning when a model is updated. Two distinct judgments therefore sit behind any AI-enabled activity: entrustment of the learner who uses AI, which remains a judgment about a developing professional, and entrustment of the AI to perform a function, which is a judgment about a tool whose behavior is contingent and must be reverified over time. The levels mentioned below describe the second judgment; the first remains governed by existing competency frameworks.

An entrustment lens also avoids 2 common errors. The first is technological determinism, a term used here to describe the assumption that a technology's capabilities or momentum make adoption inevitable. The second is blanket restriction, the assumption that uncertainty justifies institutional paralysis. Medical education needs structured adoption with explicit boundaries. We therefore propose 6 levels of AI entrustment in medical education. These levels are ordered by the consequence of an unsupervised error rather than by technical sophistication. A function sits higher when an unreviewed AI error would more directly damage a learner's progression, a fair assessment, or a patient. Figure 1 summarizes these 6 levels and their defining features.

XSL•FO
RenderX

Figure 1. The 6 levels of AI entrustment.



Level 0 is no entrustment. Level 0 is an exclusion category that sits outside the graduated sequence. It identifies functions for which AI entrustment is currently inappropriate, including autonomous or unreviewed decisions about admissions, grading, professionalism, or progression. Levels 1 to 5 describe permitted forms of AI participation with increasing consequence and governance requirements. Educational stakes, AI autonomy, and data sensitivity determine the oversight required for each use.

Level 1 is AI as a clerical assistant. AI helps with formatting, summarizing, translation, brainstorming, and drafting low-stakes educational material. Human review is required, and sensitive data should be excluded unless the tool is institutionally approved.

Level 2 is AI as a learning aid. AI supports retrieval practice, explanations, self-quizzing, study planning, and formative feedback. Learners are taught to verify outputs and document meaningful AI assistance.

Level 3 is AI as a simulator or coach. AI generates digital patients, communication scenarios, diagnostic prompts, or debriefing questions. Faculty define the scenario, review outputs, and monitor whether the tool supports reasoning and communication.

Level 4 is AI as an educational co-designer. AI helps educators build cases, rubrics, item banks, remediation plans, clerkship materials, and program evaluation summaries. Outputs require expert validation, version control, bias review, and documentation.

Level 5 is AI as an educational decision support. AI informs higher-stakes judgments, such as identifying learners who may need support or summarizing multiple data sources for a competency committee. This level requires local validation, transparency, audit trails, human accountability, and appeal mechanisms. Autonomous educational decision-making should remain outside routine use at present.

XSL•FO
RenderX

Figure 1 presents the 6 levels of AI entrustment described above, ordered by the consequence of an unsupervised AI error rather than by technical sophistication. The figure can help curriculum leaders and institutional committees quickly locate a proposed use case and identify the oversight it requires. Level 0 is an exclusion category.

Use Cases by Stakes and Maturity

The first wave of literature cataloged many possible uses of GenAI. The second wave should distinguish between use cases that are ready for supervised implementation, promising yet evidence-limited, and governance-intensive. Ready uses include generating practice questions, creating alternative explanations, simplifying patient education materials, translating educational content, drafting case stems, suggesting feedback language, and helping faculty overcome blank-page barriers. These functions are usually low stakes, easy to review, and aligned with existing educator judgment. These distinctions also interact with the stage of training. In the nonclinical years, most ready and promising uses are low stakes by nature because errors are easily checked against a still-developing knowledge base; in the clinical years, the same categories of tools intersect directly with patient care, documentation, and time-pressured decision-making, which raises the practical stakes of the supervision described below, particularly for ambient documentation and clinically embedded decision support.

Promising uses include AI standardized patients, adaptive tutoring, individualized remediation, simulation debriefing, communication coaching, and clinical reasoning scaffolds. These applications could transform learning because they create repeated opportunities for practice, feedback, and reflection. Early studies suggest that GenAI can support engagement through multimodal narratives and can be explored for medical interview training, although many studies remain small and context specific [14,15]. These uses deserve careful expansion through multi-institutional trials and direct comparison with existing educational approaches. These promising uses are also well suited to remote and geographically distributed learning, where AI standardized patients, simulation debriefing, and adaptive tutoring can extend a form of faculty presence to learners at affiliated clinical sites or in distance-based coursework, provided the same expectations for supervision, local validation, and equitable access described throughout this viewpoint are maintained.

Governance-intensive uses include summative grading, automated assessment of professionalism, admissions screening, progression recommendations, and any application involving sensitive student or patient data. They now also include agentic tools that take actions across systems and ambient documentation tools that draft clinical notes during real encounters, where the educational question is not only whether the output is accurate but whether the learner still performs the underlying cognitive work. These uses combine high educational stakes with risks of opacity, bias, data leakage, and contested accountability, meaning uncertainty about who owns and is responsible for an AI-mediated educational product or decision when it is inaccurate, harmful, inequitable, or disputed. Institutions should treat them as requiring formal review, similar to educational research protocols or clinical AI implementations.

From Demonstrations to Educational Workflows

A persistent problem in the AI literature is the gap between demonstration and implementation. Showing that a model can answer a question, draft a case, or imitate a patient is useful; however, educational value depends on workflow. A medical school adopting AI for simulation should specify who writes the scenario, who validates the clinical content, how learner data are stored, how bias is monitored, how students are oriented, and what faculty do when AI produces unsafe or misleading responses. A clerkship using AI for feedback should decide whether the feedback is private practice feedback, formative faculty-reviewed feedback, or evidence that enters the learner record. These distinctions matter because the same technical tool can have very different educational and ethical meanings depending on where it sits in the program.

Implementation should begin with the educational problem. Educators should ask the following: What learning gap are we addressing? What human activity is AI meant to augment? What evidence would count as improvement? What harms are plausible? and Who remains accountable? Health professions education has already emphasized the need for educators to understand AI as part of their professional role [16]. The next step is to translate that understanding into local operating procedures. Each AI-enabled educational activity should have a named owner, a defined learner population, a statement of permitted data, a review process for generated content, and a plan for evaluating outcomes. Tool version, date of use, and local modifications should be documented because model behavior can change over time.

A practical implementation sequence would include 5 steps. First, classify the use case by stakes, autonomy, and data sensitivity. Second, choose an approved tool and specify privacy boundaries. Third, conduct local validation with faculty and learners before full deployment. Fourth, monitor outputs, learner experience, and equity effects during implementation, and, for tools that take actions, log those actions for review. Fifth, revise or retire the activity when evidence, model behavior, or curricular priorities change. This sequence would make GenAI adoption more consistent with the quality improvement culture already familiar to medical schools.

Programs should also treat GenAI as an evolving educational intervention rather than a fixed resource. A printed handout can remain stable for years, but model behavior, vendor policies, training data, and user interfaces may change with little warning. Local validation therefore cannot be a one-time event. Schools should document the date, model, settings, prompt templates, review process, and intended use for each AI-enabled activity. This documentation would make it easier to reproduce educational materials; investigate errors; and explain decisions to learners, faculty, accreditors, and patients.

XSL•FO
RenderX

Assessment After GenAI

Assessment is the most urgent domain for redesign. Early debate often framed GenAI as a plagiarism problem. That framing is too narrow. GenAI challenges assessment validity because it changes what written products represent. A polished essay may reflect a learner's understanding, the learner's ability to frame the task through prompts, the model's capabilities, and the learner's decisions about verification and revision. Prompting can itself be intellectual work when it involves problem framing, strategic questioning, iterative critique, and synthesis. Ownership of the final product remains with the learner. The learner is accountable for accuracy, fairness, citation practices, and clinical appropriateness regardless of the extent of AI assistance. Detection tools cannot carry the full burden of academic integrity. Assessment systems should specify when AI use is prohibited, permitted, expected, or required.

AI-independent assessment is appropriate when the educational goal is to verify unaided competence. Examples include observed clinical encounters, oral examinations, in-person reasoning exercises, procedural skills assessments, closed-resource tests, and direct workplace-based assessments. These formats remain essential because physicians must still act under uncertainty, communicate with patients, and make accountable judgments.

AI-assisted assessment is appropriate when AI use resembles real professional practice. Learners may use AI to draft, revise, translate, or structure work, provided they disclose the tool, describe the nature of assistance, and submit evidence of their reasoning process. Such assessments should evaluate verification, critique, and revision rather than final prose alone.

AI-integrated assessment is appropriate when the learning objective is safe and compatible with AI use. Learners might compare AI-generated differential diagnoses, identify hallucinations, check outputs against guidelines, rewrite biased patient education materials, explain how AI altered their thinking, or counsel a simulated patient who arrives having already consulted an AI tool. The Association for Medical Education in Europe Guide on AI in health professions education assessment emphasizes that AI affects assessment types, competencies, ethics, faculty development, and acknowledgment practices [8]. A pilot study also showed that AI performance varies substantially by assessment type, performing strongly in rule-based and reflective tasks while struggling with technical accuracy and contextual application in some health information management tasks [17]. This reinforces a practical conclusion that the assessment design should make reasoning visible.

Curriculum and Faculty Development

AI literacy is becoming a core professional competency. A curriculum focused only on prompt technique would be insufficient. Learners need conceptual, ethical, epistemic, and practical competence. They should understand how LLMs generate outputs, why hallucinations occur, how bias enters training and deployment, how privacy can be compromised, and why fluent language can obscure weak reasoning. They should also learn how to verify outputs, cite or acknowledge AI assistance, recognize automation bias, and preserve accountability. They should also learn to work with patients who arrive already informed (or misinformed) by AI tools, correcting errors without dismissing the patient's engagement. Managing the AI-informed patient is plausibly an emerging entrustable activity in its own right.

Medical students have expressed interest in AI, while also reporting variable preparedness and concern about ethics, curriculum gaps, and future clinical roles [18]. Reviews of AI curricula emphasize that medical education needs structured frameworks, progressive sequencing, and integration with clinical contexts rather than isolated technical electives [19,20]. Practical AI literacy courses for early medical students and calls for AI ethics training provide starting points, but curriculum design should extend across undergraduate, graduate, and continuing medical education [21,22]. Curricula will also need validated measures. AI literacy and susceptibility to automation bias are currently assessed inconsistently, and the field lacks agreed-upon instruments for certifying that learners can use AI safely. Identifying and closing this measurement gap is itself a curricular priority.

Faculty development is equally important. Educators require enough understanding to supervise AI use responsibly. Faculty need training in tool selection, prompt design, output evaluation, bias recognition, privacy rules, assessment redesign, and learner coaching. They also need time and institutional support. Without faculty development, AI integration may default to uneven local experimentation, with wide variation in quality and access.

Cognitive Autonomy and Professional Identity

A mature sequel to the first wave of GenAI literature must address cognitive autonomy. Medical education is partly a process of productive struggle. Learners develop clinical judgment by generating hypotheses; being wrong; receiving feedback; revising mental models; and integrating biomedical, psychosocial, and contextual information. If AI systems prematurely remove that struggle, learners may become efficient producers of polished answers while developing weaker habits of inquiry.

The risk extends beyond student cheating to cognitive outsourcing. A learner who repeatedly asks AI to generate differential diagnoses may gradually lose the habit of constructing them independently. The clearest current example is ambient clinical documentation. Students increasingly rotate through clinics where an AI scribe drafts the note in real time. The convenience is real, but synthesizing a patient's story into an assessment and plan is precisely the reasoning work that builds clinical judgment. A learner who never drafts the note may never practice the synthesis. A faculty member who repeatedly asks AI to draft feedback may become less attentive to the learner's actual performance. A program that relies on AI-generated dashboards may confuse data aggregation with educational judgment. Recent work on AI dependency among educators found perceived risks to skills, pedagogy, motivation,

XSL•FO
RenderX

ethics, collaboration, and creativity, highlighting the need to study overreliance at the faculty level as well as the learner level [23]. Reviews of GenAI and professionalism similarly emphasize that technological fluency must be paired with empathy, integrity, accountability, and humanistic practice [24]. The cognitive work that should remain the learner's own is therefore specific rather than diffuse: generating an initial differential from first principles, synthesizing an undifferentiated encounter into an assessment and plan, weighing conflicting or incomplete evidence, and forming a working hypothesis under uncertainty before consulting any external aid. An AI-generated differential or note draft is a representation the learner can review and critique; it is not evidence that the learner has performed the reasoning that representation implies, and assessment should be designed to tell the two apart.

At the same time, AI may strengthen cognition when used deliberately. It can serve as a debate partner, generate counterarguments, expose learners to alternative explanations, and support diagnostic reflection. Studies of diagnostic reasoning suggest that AI tools may influence clinician reasoning and that prompting models to display reasoning can make outputs more interpretable [25,26]. The educational goal should be hybrid intelligence: learners who can think independently, use AI strategically, recognize when AI is misleading, and remain accountable for final judgments.

Governance and Equity

Institutional governance should move beyond academic honesty statements. A recent audit of AI-related documents across US medical schools found that many documents emphasized academic integrity and plagiarism, while audit mechanisms, technical infrastructure, and long-term planning were less commonly addressed [27]. Medical schools should develop policies that define approved tools, prohibited uses, disclosure expectations, privacy safeguards, data retention rules, procurement standards, bias monitoring, assessment categories, incident reporting, and periodic review. Textbox 1 summarizes minimal institutional commitments for generative AI in medical education. These commitments should be written in plain language, revisited regularly, and aligned across undergraduate medical education, graduate medical education, faculty development, and institutional compliance. A policy that focuses only on plagiarism leaves too many educational and safety questions unanswered.

Textbox 1. Minimum institutional commitments for generative AI in medical education.

Six operational questions every medical school should be able to answer before scaling AI use:

1. Which tools are approved for learners and faculty?
2. What data may be entered, and what data are prohibited?
3. When must AI assistance be disclosed?
4. Which assessments are AI-independent, AI-assisted, or AI-integrated?
5. Who reviews AI-generated educational content before learners see it?
6. How will the school monitor errors, bias, learner overreliance, access gaps, and model changes?

Approving a tool for institutional use should weigh, at minimum, the vendor's data privacy and retention practices, evidence of accuracy and bias testing relevant to the intended educational use, transparency about model versioning and training data, compatibility with institutional authentication and data security requirements, cost and licensing implications for equitable access, and a designated process for periodic rereview as the tool or its underlying model changes.

Governance should also address equity. The earlier "digital divide" now includes an "AI divide." Students and schools vary in access to paid models, high-quality institutional tools, secure platforms, faculty expertise, broadband, language support, and protected time for training. Learners in low-resource settings may face restricted access to the most capable systems while also being more exposed to unvetted free tools. Equity also includes disability access, language performance, cultural representation in cases, and fair treatment of learners whose writing style is flagged as suspicious. An institution that permits AI use without ensuring equitable access may inadvertently advantage learners with greater resources.

Policy language should be role-specific. Learners need clear expectations for disclosure and verification. Faculty need guidance for creating materials, providing feedback, and documenting AI assistance. Administrators need procurement criteria, data protection rules, and processes for reviewing vendors. Competency committees need boundaries around how AI-generated summaries may be used. Role-specific policy reduces ambiguity and makes responsible AI use easier to teach, supervise, and audit. Successful implementation also depends on stakeholders beyond faculty and learners. Information technology and data security personnel are needed to vet vendor data-handling practices, configure institutional access controls, and monitor for data leakage, while instructional designers and clinical informaticists help translate approved tools into workflow-embedded activities; governance structures should include these roles formally rather than treating AI oversight as an academic affairs function alone.

Research Agenda

The next research agenda should prioritize educational outcomes. The literature already contains many studies asking whether AI can answer examination questions. Medical education now needs studies asking whether AI improves durable learning, diagnostic calibration, communication, feedback quality, clinical performance, and patient outcomes.

XSL•FO
RenderX

A systematic review of GenAI in health professional education found that student use aligns with several learning activities, while evidence remains limited by short duration, small samples, and underexplored collaborative learning [28]. The field needs longitudinal, multicenter, theory-informed studies. The field is also moving past the point where novelty alone justifies attention; the priority now is whether documented learning gains persist and transfer to practice.

Key research questions include the following: Which AI tutoring designs improve retention and transfer? How does AI feedback compare with faculty feedback across clinical skills? Does AI-assisted simulation improve real patient communication? Which forms of AI disclosure support integrity without stigmatizing appropriate use? What learner groups are helped or harmed? How do AI tools affect workload, burnout, creativity, and teaching quality among faculty? What governance structures reduce risk while supporting innovation? How should institutions evaluate model updates that alter performance over time? What level of autonomy can agentic AI safely assume in educational tasks, and how should action-taking tools be evaluated differently from advisory ones? and Which validated instruments reliably measure AI literacy, automation bias, and appropriate reliance, and how do these correlate with clinical performance?

The AI Entrustment Matrix

Table 1 presents a conceptual matrix with 2 axes: educational stakes, from low to high, and AI autonomy, from assistant to decision support. The matrix can help curriculum leaders decide when routine disclosure is sufficient and when formal governance is required.

Table 1. A conceptual matrix for evaluating AI entrustment in the educational environment.

	Low educational stakes	High educational stakes
Low AI autonomy	Accept with disclosure and ordinary faculty review. Examples: brainstorming, formatting, translation, and practice questions	Use with validation, documentation, and accountable human review. Examples: rubric drafting, case design, and feedback support
High AI autonomy	Permit for practice environments with monitoring for overreliance. Examples: AI standardized patients and adaptive tutoring	Restrict or formally govern. Examples: progression recommendations, automated grading, and admissions screening

Conclusion: From Adoption to Accountability

GenAI has already entered medical education. The task now is to shape its role deliberately. The first phase established that these tools could generate plausible educational content, simulate conversations, support writing, and assist learning. The second phase must determine how, when, and under what supervision they should be used. Entrustment offers a language that medical educators already understand: responsibility should be granted progressively, contextually, and with evidence.

Medical schools should develop AI policies that are practical, transparent, and revisable; assessments that make reasoning visible; curricula that teach AI literacy as part of professionalism; and research programs that evaluate learning rather than novelty alone. GenAI should be judged by its effects on learners, educators, institutions, and ultimately patients. The goal is accountable integration: using AI to expand educational possibility while preserving the human judgment at the center of medicine. Although this framework is presented for medical education, the underlying questions, namely, which cognitive functions to delegate, to what degree, and with what oversight, are not unique to medicine. Educators across science, technology, engineering, and mathematics (STEM) disciplines and other professional fields built around graduated responsibility, including engineering, nursing, and teacher education, face comparable decisions about entrusting AI with defined instructional functions and may find this entrustment framework a useful starting point for their own context.

Acknowledgments

The authors used generative AI tools (Claude Sonnet 5 [Anthropic] and ChatGPT 5.6 [OpenAI]) to assist with literature search, language editing, and drafting portions of this manuscript. All AI-assisted output was reviewed, fact-checked, and edited by the authors, who take full responsibility for the accuracy, originality, and scientific integrity of the final manuscript.

Funding

The authors declared that no financial support was received for this work.

Authors' Contributions

SP, MK, SB, and KM conceptualized this viewpoint. SP and KM conducted the literature synthesis and prepared the original draft. SP and KM designed the figure. MK, SB, and KM provided critical revisions and supervised the intellectual content throughout development. All authors reviewed, edited, and approved the final manuscript.

XSL•FO
RenderX

Conflicts of Interest

None declared.

References

1. Karabacak M, Ozkara BB, Margetis K, Wintermark M, Bisdas S. The advent of generative language models in medical education. JMIR Med Educ 2023 Jun 06;9:e48163 [FREE Full text] [doi: 10.2196/48163] [Medline: 37279048]
2. Kung TH, Cheatham M, Medenilla A, Sillos C, De Leon L, Elepaño C, et al. Performance of ChatGPT on USMLE: potential for AI-assisted medical education using large language models. PLOS Digit Health 2023 Feb 9;2(2):e0000198 [FREE Full text] [doi: 10.1371/journal.pdig.0000198] [Medline: 36812645]
3. Singhal K, Tu T, Gottweis J, Sayres R, Wulczyn E, Amin M, et al. Toward expert-level medical question answering with large language models. Nat Med 2025 Mar;31(3):943-950. [doi: 10.1038/s41591-024-03423-7] [Medline: 39779926]
4. Gupta P, Zhang Z, Song M, Michalowski M, Hu X, Stiglic G, et al. Rapid review: growing usage of multimodal large language models in healthcare. J Biomed Inform 2025 Sep;169:104875. [doi: 10.1016/j.jbi.2025.104875] [Medline: 40754135]
5. Buess L, Keicher M, Navab N, Maier A, Tayebi Arasteh S. From large language models to multimodal AI: a scoping review on the potential of generative AI in medicine. Biomed Eng Lett 2025 Aug 22;15(5):845-863. [doi: 10.1007/s13534-025-00497-1] [Medline: 40917142]
6. Boscardin CK, Gin B, Golde PB, Hauer KE. ChatGPT and generative artificial intelligence for medical education: potential impact and opportunity. Acad Med 2024 Jan 01;99(1):22-27 [FREE Full text] [doi: 10.1097/ACM.0000000000005439] [Medline: 37651677]
7. Gordon M, Daniel M, Ajiboye A, Uraiby H, Xu NY, Bartlett R, et al. A scoping review of artificial intelligence in medical education: BEME guide no. 84. Med Teach 2024 Apr;46(4):446-470. [doi: 10.1080/0142159x.2024.2314198]
8. Masters K, MacNeil H, Benjamin J, Carver T, Nemethy K, Valanci-Aroesty S, et al. Artificial intelligence in health professions education assessment: AMEE guide no. 178. Med Teach 2025 Sep;47(9):1410-1424. [doi: 10.1080/0142159X.2024.2445037] [Medline: 39787028]
9. Ten Cate O. Competency-based education, entrustable professional activities, and the power of language. J Grad Med Educ 2013 Mar;5(1):6-7 [FREE Full text] [doi: 10.4300/JGME-D-12-00381.1] [Medline: 24404218]
10. Ten Cate O, Taylor DR. The recommended description of an entrustable professional activity: AMEE guide no. 140. Med Teach 2021 Oct;43(10):1106-1114 [FREE Full text] [doi: 10.1080/0142159X.2020.1838465] [Medline: 33167763]
11. Cate OT. A primer on entrustable professional activities. Korean J Med Educ 2018 Mar;30(1):1-10 [FREE Full text] [doi: 10.3946/kjme.2018.76] [Medline: 29510603]
12. Gin BC, O'Sullivan PS, Hauer KE, Abdulnour R, Mackenzie M, Ten Cate O, et al. Entrustment and EPAs for artificial intelligence (AI): a framework to safeguard the use of AI in health professions education. Acad Med 2025 Mar 01;100(3):264-272 [FREE Full text] [doi: 10.1097/ACM.0000000000005930] [Medline: 39761533]
13. Jacobs SM, Lundy NN, Issenberg SB, Chandran L. Reimagining core entrustable professional activities for undergraduate medical education in the era of artificial intelligence. JMIR Med Educ 2023 Dec 19;9:e50903 [FREE Full text] [doi: 10.2196/50903] [Medline: 38052721]
14. Hirosawa T, Yokose M, Sakamoto T, Harada Y, Tokumasu K, Mizuta K, et al. Utility of generative artificial intelligence for Japanese medical interview training: randomized crossover pilot study. JMIR Med Educ 2025 Aug 01;11:e77332 [FREE Full text] [doi: 10.2196/77332] [Medline: 40749190]
15. Bland T. Enhancing medical student engagement through cinematic clinical narratives: multimodal generative AI-based mixed methods study. JMIR Med Educ 2025 Jan 06;11:e63865 [FREE Full text] [doi: 10.2196/63865] [Medline: 39791333]
16. Lomis K, Jeffries P, Palatta A, Sage M, Sheikh J, Sheperis C, et al. Artificial intelligence for health professions educators. NAM Perspect 2021 Sep 8;2021:10.31478/202109a [FREE Full text] [doi: 10.31478/202109a] [Medline: 34901780]
17. Wani TA, Liem M, Prasad N, Robinson K, Nexhip A, Tassos M, et al. Susceptibility of assessment types to AI-generated content in digital health and health information management education: quasi-experimental pilot study. JMIR Med Educ 2026 Mar 30;12:e82988 [FREE Full text] [doi: 10.2196/82988] [Medline: 41911020]
18. Ejaz H, McGrath H, Wong BL, Guise A, Vercauteren T, Shapey J. Artificial intelligence and medical education: a global mixed-methods study of medical students' perspectives. Digit Health 2022 May 02;8:20552076221089099 [FREE Full text] [doi: 10.1177/20552076221089099] [Medline: 35521511]
19. Kim DH, Kang YJ, Lee YM. Twelve tips for developing and implementing AI curriculum for undergraduate medical education. Med Educ Online 2025 Dec 31;30(1):2585637 [FREE Full text] [doi: 10.1080/10872981.2025.2585637] [Medline: 41327942]
20. Tolentino R, Baradaran A, Gore G, Pluye P, Abbasgholizadeh-Rahimi S. Curriculum frameworks and educational programs in AI for medical students, residents, and practicing physicians: scoping review. JMIR Med Educ 2024 Jul 18;10:e54793 [FREE Full text] [doi: 10.2196/54793] [Medline: 39023999]
21. Levingston H, Anderson MC, Roni MA. From theory to practice: artificial intelligence (AI) literacy course for first-year medical students. Cureus 2024 Oct 02;16(10):e70706. [doi: 10.7759/cureus.70706] [Medline: 39493023]

XSL•FO
RenderX

22. Katznelson G, Gerke S. The need for health AI ethics in medical school education. Adv Health Sci Educ Theory Pract 2021 Oct;26(4):1447-1458. [doi: 10.1007/s10459-021-10040-3] [Medline: 33655433]
23. Alhur AA, Khlaif ZN, Hamamra B, Hussein E. Paradox of AI in higher education: qualitative inquiry into AI dependency among educators in Palestine. JMIR Med Educ 2025 Sep 15;11:e74947 [FREE Full text] [doi: 10.2196/74947] [Medline: 40953310]
24. Komasawa N, Yokohira M. Generative artificial intelligence (AI) in medical education: a narrative review of the challenges and possibilities for future professionalism. Cureus 2025 Jun 18;17(6):e86316. [doi: 10.7759/cureus.86316] [Medline: 40688936]
25. Goh E, Gallo R, Hom J, Strong E, Weng Y, Kerman H, et al. Large language model influence on diagnostic reasoning: a randomized clinical trial. JAMA Netw Open 2024 Oct 01;7(10):e2440969 [FREE Full text] [doi: 10.1001/jamanetworkopen.2024.40969] [Medline: 39466245]
26. Savage T, Nayak A, Gallo R, Rangan E, Chen JH. Diagnostic reasoning prompts reveal the potential for large language model interpretability in medicine. NPJ Digit Med 2024 Jan 24;7(1):20 [FREE Full text] [doi: 10.1038/s41746-024-01010-1] [Medline: 38267608]
27. Rush E, Byram JN, Garnett CN, DeVaul N, Smith L, Checchi M, et al. An audit of AI-related documents across U.S. medical schools: a framework-based qualitative content analysis. Med Teach 2026 Mar;48(3):493-505. [doi: 10.1080/0142159X.2025.2564869] [Medline: 41021415]
28. Pham TD, Karunaratne N, Exintaris B, Liu D, Lay T, Yuriev E, et al. The impact of generative AI on health professional education: a systematic review in the context of student learning. Med Educ 2025 Dec;59(12):1280-1289. [doi: 10.1111/medu.15746] [Medline: 40533396]

Abbreviations

EPA: entrustable professional activity
GenAI: generative AI
LLM: large language model
STEM: science, technology, engineering, and mathematics

Edited by A Stone; submitted 23.Jun.2026; peer-reviewed by DJ Carrejo, N Mahboobani; comments to author 31.Jul.2026; revised version received 01.Sep.2026; accepted 02.Sep.2026; published 30.Sep.2026.

Please cite as:
Patil S, Karabacak M, Bisdas S, Margetis K
Generative Language Models in Medical Education: From Advent to Entrustment
JMIR Med Educ 2026;12:e105347
URL: https://mededu.jmir.org/2026/1/e105347
doi:10.2196/105347
PMID:

©Shiv Patil, Mert Karabacak, Sotirios Bisdas, Konstantinos Margetis. Originally published in JMIR Medical Education (https://mededu.jmir.org), 30.Sep.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Education, is properly cited. The complete bibliographic information, a link to the original publication on https://mededu.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Viewpoint

From Competence to Expertise: The Role of Meaningful Experience in Physician Training

Nicole M Dubosh[1], MD; Alexander A Iyer[2], MD; Richard M Schwartzstein[3], MD

[1]Department of Emergency Medicine, Harvard Medical School, Beth Israel Deaconess Medical Center, Boston, MA, United States

[2]Department of Internal Medicine, University of California, San Francisco, San Francisco, CA, United States

[3]Department of Medicine, Harvard Medical School, Beth Israel Deaconess Medical Center, Boston, MA, United States

Corresponding Author:
Nicole M Dubosh, MD
Department of Emergency Medicine
Harvard Medical School
Beth Israel Deaconess Medical Center
One Deaconess Road
Rosenberg 2
Boston, MA, 02215
United States
Phone: 1 617 754 2450
Email: ndubosh@bidmc.harvard.edu

Abstract

The fundamental goal of medical education is to develop physicians who can independently provide safe and effective patient care. This requires learners to acquire not only knowledge but also meaningful experience to progress along the continuum from novice to expert. In the last several decades, however, medical education has undergone numerous transformations that have resulted in decreased clinical experience for medical students and residents alike. These transformations include the increasing complexity of health care delivery in the inpatient and outpatient settings; duty hour restrictions; and the explosion of digital technologies, including the electronic health record and AI. Additionally, the introduction and broad acceptance of competency-based medical education, which is focused on demonstrable behaviors rather than the time spent in training, risks exacerbating this loss of meaningful clinical experience if implemented without attention to trainees' need for clinical volume, variation, and redundancy. While these changes in the number and intensity of patient encounters in educational programs are multifactorial in origin, when combined, they jeopardize the acquisition of experience, the development of expertise, and the resulting quality of patient care. In this Viewpoint, our aim is to discuss the importance of meaningful experience in the development of expertise; highlight threats to its achievement resulting from the intersection of 3 transformations affecting medical training (the increased complexity of 21st-century health care systems, expansion of the electronic medical record and AI, and flawed understanding of the full implications of competency-based medical education as presently implemented); and present solutions for medical educators, academic institutions, and accrediting bodies. Our views are based on our perspectives and observations, as well as insights from the literature.

(JMIR Med Educ 2026;12:e106585) doi:10.2196/106585

KEYWORDS

experience; expertise; competency-based medical education; digital technologies; artificial intelligence; AI; clinical reasoning; graduate medical education; deskilling

The Importance of Meaningful Experience

On January 15, 2009, shortly after takeoff, Captain Chelsey "Sully" Sullenberger's aircraft struck a flock of geese, damaging both engines. With minutes to act, Sully ditched the aircraft onto the Hudson River, ensuring the safety of all 155 passengers and crew members and avoiding what could have been a fatal catastrophe. What allowed him to accomplish this life-saving feat?

Deliberative practice across a range of conditions allows experts such as Sully to react appropriately in high-stakes situations, particularly when new or unforeseen circumstances arise. Aviation training reflects the importance of experience in the development of expertise. Training reforms since the Colgan

XSL•FO
RenderX

Air Flight 3407 crash require that pilot school graduates log 1500 hours as copilots under a range of weather and mechanical conditions and have dedicated meaningful experience in various domains before captaining commercial airliners. It is one thing to demonstrate competence in the knowledge and skills of the core procedures of a pilot under routine conditions; it is something else to be able to expand upon those under more challenging circumstances. This extensive preparation, meaningful experience, and time are deemed necessary to independently fly a plane when lives are at stake.

While it is a different profession with different skill sets and limitations, medicine is not dissimilar from aviation. Physicians, like pilots, must develop the expertise necessary to make life-dependent decisions in a variety of circumstances—a process requiring the acquisition of both knowledge *and* experience. We intuitively recognize the value of experience when we request advice from a senior colleague in managing a difficult presentation of a disease or when we choose to refer a high-stakes surgical case to a surgeon who has performed that surgery 1000 times rather than on 10 occasions; it is not that they are not both competent, but basic competence may not suffice under less ordinary conditions. Historically, medical education has provided this experience for trainees through time-based training programs that included extensive patient encounters with graduated responsibility [1]. Many have criticized this traditional educational model for its perceived shortcomings: challenging work-life balance for trainees, concerns about medical errors, and variable assessment practices in graduate medical education (GME) [2-4].

In part because of these concerns, medical education has been systematically reducing the number and, at times, the continuity of meaningful patient care experiences for medical students and residents. In undergraduate medical education, there has been a steady decline in clinical activities during the fourth year and a movement toward accelerated 3-year programs [5,6]. While the Liaison Committee on Medical Education mandates that medical schools establish a tracking system to monitor clinical or simulated experiences for a student, such as exposure to a patient with chest pain, abdominal pain, or shortness of breath, it does not define the quantity or extent of these encounters [7]. Furthermore, it does not specify whether the focus is on the symptom vs the specific disease state, and there is no required number of patients for any given disease. The student is required to obtain a history, demonstrate a physical examination, generate a differential diagnosis, and participate in the care of a single patient with that symptom to satisfy the competency.

In GME, multiple factors have jeopardized the acquisition of clinical experience: the increased complexity of health care delivery, crowded hospitals with boarding in emergency departments, duty hour restrictions, the proliferation of digital health care technologies, and the oftentimes flawed or inadequate implementation of competency-based medical education (CBME). In this Viewpoint, our aim is to discuss the relationship between experience and expertise; describe how modern medical training environments may unintentionally limit the kinds of experience that foster adaptive expertise based on our observations and support from evidence in the literature; and pose solutions for educators, academic institutions, and accrediting bodies.

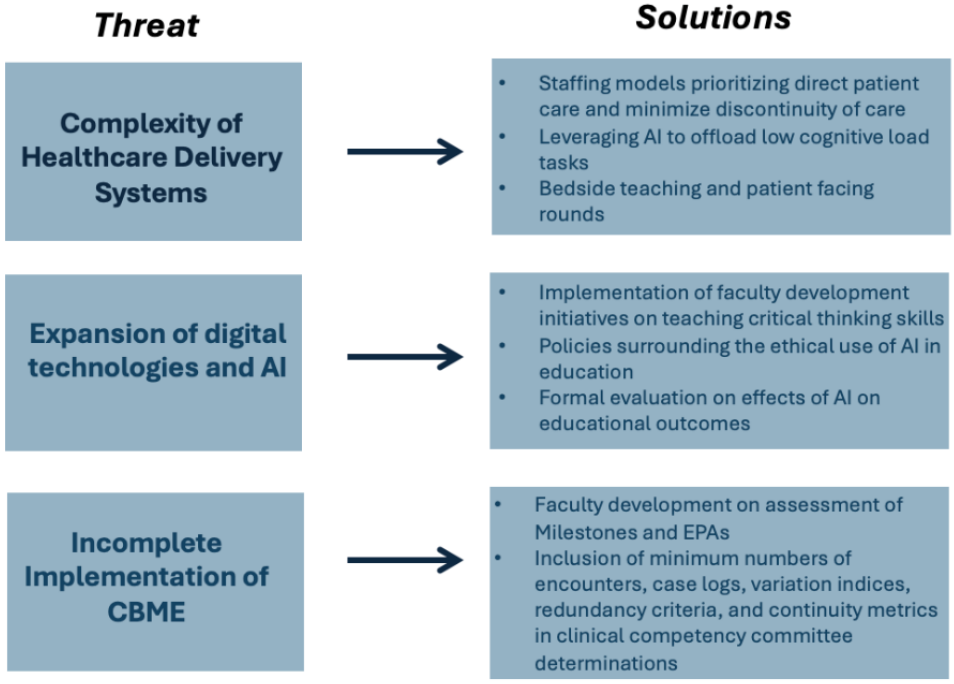
Experience and Expertise

What is expertise, and why is it important? In his model of adult skill acquisition, Dreyfus [8] outlines 5 stages: novice, advanced beginner, competent, proficient, and expert. These stages differ in the degree of intuition demonstrated by a practitioner. Less experienced learners tend to depend on rules and guidelines and illness scripts, while experts rely on strong foundational knowledge supplemented by intuition honed through extensive experience (eg, there is something about this story that just does not feel right for diagnosis X). This experience allows experts to recognize subtle nuances in a case and devise situationally appropriate responses to unfamiliar, ambiguous, and unstructured problems.

In their classic work on cognitive metaphors, Mylopoulos and Regehr [9] outline the construct of an adaptive expert, an individual who has significant clinical experience and can solve a problem that may appear unique by using a store of foundational knowledge. The ability to sense and appreciate that there is something different and then reason to a new solution characterizes the adaptive expert. This requires substantial clinical volume with variation and redundancy with respect to disease states during training. With the goal of developing physicians who can independently provide effective patient care and continue on the trajectory toward becoming an adaptive expert, medical educators should examine several recent changes that threaten medical trainees' acquisition of experience: the increased challenges associated with 21st-century health care, expansion of the electronic health record (EHR) and its effect on independent reasoning by students and trainees, introduction and adoption of AI, and impact of the implementation of CBME as the primary means of assessment and determination that one is ready for the next "level" or completion of training (Figure 1).

XSL•FO
RenderX

Figure 1. Threats to meaningful experience in the development of expertise. CBME: competency-based medical education; EPA: entrustable professional activity.



Increasing Complexity of Health Care Systems

Threat

The first threat to the cultivation of experience relates to the current health care system and its effect on direct patient care in training programs. The increasing complexity of health care delivery, including crowded emergency departments, shorter patient lengths of stay, and staffing shortages, has forced trainees to devote substantial time to nonclinical tasks such as documentation, billing, insurance navigation, scheduling, and multiple handoffs. These factors erode a sense of clinical continuity [10], limiting the full picture of the patient's disease and treatment.

The increased reliance and time spent engaging with the EHR also threatens time spent at the bedside and engaging in direct patient care. While the EHR certainly allows better centralization of medical records and standardization of care, its complexity comes at the cost of significant user time. A study in the ambulatory setting at one academic program conducted in 2017 to 2018 found that residents spent an average of 45.6 minutes in the EHR per patient, with 13.5% of that time spent after hours. While the residents showed improvement in efficiency with the EHR, there was no significant increase in patients seen per day or time spent in the EHR after hours [11]. A prospective, multicenter investigation of first-year internal medicine interns on clinical inpatient shifts using time-motion observations conducted in 2016 found that in a 24-hour period, a mean of only 3.0 hours was spent in direct patient care [12]. Concurrently, duty hour restrictions have resulted in fewer patient encounters across multiple specialties [13,14]. At our medical center and other academic centers with whom we have communicated informally, an internal medicine intern today admits a small fraction of the patients evaluated in the past. Over the course of 12 months at our institution, an intern in medicine admits fewer than one-quarter the number of patients admitted by an intern 30 years ago (internal data, 2026). While these data remain preliminary, institution specific, and not yet generalizable, the drop we observe warrants further investigation. There is growing concern as well about the clinical competence of surgical residents, who are not participating in a sufficient number of operative cases and/or having sufficient autonomy to be able to perform independently [15,16]. In a national survey in general surgery, program directors reported that trainees were less prepared than 10 years prior, citing duty hours and increased supervision among factors diminishing autonomy [17].

The impact of suboptimal experience has coincided with declining board examination pass rates across multiple specialties, including internal medicine and emergency medicine [18,19]. In emergency medicine, the reason for this decline is likely multifactorial, including the recent proliferation of new programs affiliated with for-profit institutions as well as the decrease in new patient volume due to the COVID-19 pandemic and boarding [20]. The decline in internal medicine board examination pass rates coincided with the COVID-19 pandemic, during which trainee experience and encounters were limited, and has since plateaued. While in both specialties these correlations are multifactorial and associated with pandemic-era and program-level factors, the educational effects of reduced patient exposure warrant further investigation.

Furthermore, when one adds the discontinuity of care associated with increased handoffs due to restricted duty hours, concerns about patient safety arise [21,22]. In 2 national cluster randomized trials of surgery and internal medicine residents, however, flexible hours were found to be noninferior with respect to death or serious complications compared to standard

XSL•FO
RenderX

or restricted work hours [23,24]. Furthermore, program directors in the internal medicine trial were less likely to report dissatisfaction with elements such as intern ownership of patient care, time for reflection, real-time feedback from attendings, frequency of handoffs, teaching opportunities, and continuity of care in the flexible hour model. The amount of time trainees spent in direct patient care and educational activities, however, did not significantly differ between the groups [25]. We recognize the well-intentioned desire to improve the wellness of trainees and concern for fatigue-related errors, and the implementation of duty hour restrictions certainly puts these aims at the forefront. Whether its benefits support the resulting effects on education, continuity, and time spent on patient care remains unclear. It is one thing to be deemed competent based on interactions with a 25-year-old patient with asthma and quite another to subsequently be confronted with a 75-year-old patient with asthma and a history of congestive heart failure and coronary artery disease. Nevertheless, reduced clinical experience, if not compensated for in other ways, threatens the development of true competence, let alone expertise in patient care. In the end, the recognition of one's shortcomings upon graduating from a residency program in which adequate clinical experience was lacking may compromise one's sense of accomplishment, which is essential for personal well-being [26].

Proposed Solutions

With the increasing complexity of the health care system, program directors and educational leaders need to advocate for the prioritization of staffing models that minimize trainee time spent performing nonclinical duties such as patient scheduling and similar administrative tasks that do not require physician oversight. This may allow for more opportunities for patient contact and clinical learning experiences during the workday. Furthermore, with the tremendous expansion in AI capabilities, educators can also leverage its benefits for certain time-consuming tasks that do not require the cognitive abilities of physicians or threaten the acquisition of critical thinking skills. There is evidence to show that using AI for medication reconciliation, for example, can decrease the administrative burden of clinicians [27] and, thus, allow more time for patient-facing clinical experience.

In light of duty hour limitations, institutions and frontline educators should reprioritize patient exposure as they structure teaching rounds and workflow, ensuring that time in the clinical space is patient facing. One strategy for this includes reincorporating bedside rounds, which have seen a significant decrease [28], to allow for more patient-facing interactions and exposure to clinical competencies such as physical examination techniques and effective communication skills [29]. Bedside rounds have been shown to be preferred by patients [30], and bedside teaching has had a positive effect on education, including the ability of preceptors to provide feedback [31]. Re-examination of staffing models that introduce discontinuity of care through handoffs, such as admission-only services, may also be warranted.

Expansion of the EHR and AI

Threat

A second threat to trainees' experience involves the expansion of the EHR and digital technologies, including AI. Increasing reliance on templated notes and "copy/paste" practices designed to maximize hospital revenue have made it easier to go through the motions of patient care without engaging in novel clinical reasoning [32]. We recognize that protocols embedded in the EHR can improve care, particularly for routine procedures and practices. Nevertheless, it is critical for trainees to gain experience thinking through clinical problems without relying excessively on these automated management aids. This will allow them to recognize when deviation from protocols is needed.

Similarly, AI has quickly become ubiquitous in clinical care. While AI can certainly improve efficiency for low–cognitive load, time-consuming tasks as noted above and serve as a beneficial learning resource, medical students and residents are using large language models across a variety of high–cognitive load tasks such as development of differential diagnosis, management plans, and documentation [33,34]. Lack of safeguards around the use of AI puts clinicians at risk of deskilling (degradation of established competencies in already trained clinicians), misskilling (acquisition of incorrect clinical reasoning patterns through uncritical adoption of erroneous or biased AI outputs), and never-skilling (failure to develop the foundational reasoning competencies during formative training). The effect of AI on deskilling or, potentially worse, never-skilling in the learner population remains theoretical and warrants investigation [35]. Without ever having developed these clinical reasoning skills, students and trainees risk not being able to recognize hallucinations, omissions, and faulty logic in AI output, nor will they be able to develop adequate prompts to maximize the output of AI, which could compromise patient care. The reduction in clinical experience and the confidence to use well-developed clinical and reasoning skills only worsens this problem. It is one thing to use AI as a tool to check on the completeness of one's assessment and/or treatment plan. It is another, however, to be so dependent on AI that a physician cannot function without it.

Proposed Solutions

Given the rapidly expanding digital landscape and unlimited access to generative AI tools, educators must recognize potential negative downstream effects on their learners. Training programs must prioritize faculty development initiatives focused on teaching critical thinking skills to trainees in the age of the rapidly expanding EHR, readily available protocols, and explosion of AI tools for answering clinical questions. Frameworks for this have been developed, including DEFT-AI (diagnosis, evidence feedback, teaching, and recommendation for AI use) [36]. Requiring students and residents to truly write their notes (as opposed to templates and carrying forward the prior notes) and present without reading directly from a prior note may improve assessment of their reasoning skills and should also be considered [37]. Institutions should develop and implement policies surrounding the use of AI in students and

XSL•FO
RenderX

trainees and ensure AI-free portions of the program [35] so clinical reasoning skills can be cultivated. The Association of American Medical Colleges has published principles for the use of AI in medical education, outlining the need for maintaining a human-centered focus, the need for ethical considerations, and the need for evaluation of how AI tools affect educational outcomes [38]. Well-designed, prospective studies investigating the efficacy and educational effect of this are needed.

Incomplete Implementation of CBME

Threat

A third change bearing scrutiny is the flawed or incomplete implementation of CBME in many training programs. CBME, if not implemented appropriately, can undermine trainees' acquisition of experience. Theoretical justifications for CBME are compelling and include the desire for clear learning outcomes, personalized training paths, and clear societal accountability. CBME correctly rejects the assumption that time spent in training alone guarantees readiness to provide appropriate care, relying on the development of "competency" as defined by tools such as milestones or entrustable professional activities (EPAs) [39-42]. CBME aims to ensure that all graduates possess the necessary knowledge, skills, and behaviors as demonstrated through measurable outcomes, ultimately to improve patient care and accountability to society [42]. CBME as designed, however, often varies from CBME as commonly enacted in many programs. For instance, outcome markers such as the EPAs include the ability to perform a physical examination but do not necessarily confirm that the learner can detect abnormalities or nuances that come with time and meaningful experience. With limited numbers of patient encounters and reliance on a CBME rubric alone, the learner may never have exposure to a patient with particular anomalies or one with multiple problems concurrently.

Despite its conceptual appeal, empirical evidence supporting CBME's efficacy remains mixed. Over two-thirds of published articles about CBME are commentaries rather than empirical research [43], and a recent review found only inconsistent empirical support for CBME's foundational assumptions [44]. A recent retrospective analysis of patient outcomes after endovascular aneurysm repair found that lower milestone ratings of vascular surgeons during their surgical training were associated with major complications in early surgical practice [45]. Similarly, in primary care, a retrospective study of practicing physicians who trained from 2015 to 2019 found that those who scored higher on professionalism and interpersonal skills and communication milestones in the previous 6 months of residency received better patient experience ratings in their first year of unsupervised practice on the Consumer Assessment of Healthcare Providers and Systems Clinician and Group Survey [46]. On the other hand, a retrospective analysis of a cohort of over 6000 hospitalists who completed training in 2016 to 2018 and were treating Medicare fee-for-service beneficiaries found that board examination scores correlated with lower 7-day patient mortality and readmission rates. Their milestone ratings, however, showed no clear association with 7-day mortality and readmission rates, 30-day mortality and readmission rates, length of stay, and subspecialist consultation frequency [47]. It is important to note that these studies were all correlational and not designed to test all formative or readiness uses of milestones and that the milestones were designed to describe the progressive level of performance of a trainee. Nevertheless, the data raise concerns about the present use of CBME. Prospective investigations are needed to determine whether the milestone ratings alone or combined with both the quantitative and qualitative types of experience that trainees have can truly gauge readiness for independent practices and patient outcomes.

While CBME *in principle* emphasizes robust competency assessment across a range of contexts, CBME *in practice* often devolves into resource-intensive administrative exercises in which clinical competency committees ensure that trainees receive the necessary approval to advance at the expected pace [48]. EPA levels are designed to be determined by observable behaviors and provide a forward-looking assessment of generic clinical skills; it is unclear that they predict the individual's readiness to respond to a more complex and unknown instance in the future. They do not account for the entirety of the actual clinical environment and real-world practice [49,50]. Additionally, the underlying rationale behind "entrustability" parallels the framework by Dreyfus [8] of novice to expert [40], further necessitating the need to consider experience as a critical factor on this scale. Competence is a necessary but not sufficient goal for graduates of a GME program.

Furthermore, the assumption that meeting the competencies, as defined by these instruments as a listing of the basic skills of a physician, guarantees competence in a novel situation is misguided; if a medical resident has never seen a patient with status asthmaticus or acute pulmonary edema, can they be deemed competent upon graduation? If a general surgical resident has never performed a hernia repair, are they competent? While case logs are required in surgery and time requirements spent in different venues are defined in nonprocedural fields, the role of the trainee caring for those patients and the exact disease states are not specified. To be sure, no trainee will have exposure to every disease or possible procedure in their training, but a competency-based model that does not explicitly incorporate clinical experience may obscure the loss of training in core areas. The implementation of CBME alone cannot substitute for a system that ensures sufficient trainee experience across relevant practice domains.

Proposed Solutions

With CBME now standard across training programs, clerkship and program directors need to ensure that determinations of competence truly represent a learner's stage on the path to expertise. Dedicated training on how these designations should be made along with accounting for experience may provide more accurate predictive assessments of abilities. For example, GME programs can revamp clinical competency committees to ensure that milestone ratings promote learner development and readiness for unsupervised practice [51] with consideration of case logs and exposure to certain disease states and presentations. This could include defining minimum numbers of encounters or procedures, a case mix or variation index,

XSL•FO
RenderX

redundancy criteria, continuity metrics, encounter definitions, and review intervals. In 2025, the Review Committee for Emergency Medicine recommended for the first time a minimum number of emergency department patient encounters for completion of residency. It also advocated for extending training to 48 months across all programs to better accommodate today's health care and educational environment [52]. While the latter recommendation comes with significant costs, including financial considerations, equity issues in reduced training environments, and trainee opportunity costs, that must be weighed against its potential benefits, it is an albeit controversial possible solution to address insufficient experience. We encourage other medical specialties and accrediting bodies to follow suit in explicitly defining numbers and ranges of various clinical experiences.

Conclusions

Educators, institutions, and accrediting bodies must remain steadfast in their commitment to ensuring that educational methods and assessment practices acknowledge *experience* as a fundamental pillar of expertise. The increased complexity of the health care delivery system and GME regulations resulting in fewer patient encounters, diminished trainee autonomy, the proliferation of digital health care technologies and AI, and overreliance on CBME alone threaten authentic experience. It is important to note that our supporting evidence is drawn largely from select US specialties, and we caution extrapolation to other training structures where health care, duty hours, and assessment models differ. Regardless of system, however, we should not compromise the education of future physicians or patient care by neglecting the importance of meaningful experience.

Funding

The authors declared no financial support was received for this work.

Generative AI Use

The authors declare that generative AI was not used in the preparation of this manuscript.

Authors' Contributions

NMD was responsible for conceptualization, writing—original draft, writing—review and editing, and project administration. AAI was responsible for writing—review and editing. RMS was responsible for supervision, conceptualization, and writing—review and editing.

Conflicts of Interest

None declared.

References

1. Lucey CR, Thibault GE, Ten Cate O. Competency-based, time-variable education in the health professions: crossroads. Acad Med 2018 Mar;93(3S):S1-S5. [doi: 10.1097/ACM.0000000000002080] [Medline: 29485479]
2. Landrigan CP, Rothschild JM, Cronin JW, Kaushal R, Burdick E, Katz JT, et al. Effect of reducing interns' work hours on serious medical errors in intensive care units. N Engl J Med 2004 Oct 28;351(18):1838-1848. [doi: 10.1056/NEJMoa041406] [Medline: 15509817]
3. Goitein L, Shanafelt TD, Wipf JE, Slatore CG, Back AL. The effects of work-hour limitations on resident well-being, patient care, and education in an internal medicine residency program. Arch Intern Med 2005 Dec;165(22):2601-2606. [doi: 10.1001/archinte.165.22.2601] [Medline: 16344417]
4. Goldhamer ME, Pusic MV, Nadel ES, Co JP, Weinstein DF. Promotion in place: a model for competency-based, time-variable graduate medical education. Acad Med 2024 May 01;99(5):518-523 [FREE Full text] [doi: 10.1097/ACM.0000000000005652] [Medline: 38285547]
5. Walling A, Merando A. The fourth year of medical education: a literature review. Acad Med 2010 Nov;85(11):1698-1704. [doi: 10.1097/ACM.0b013e3181f52dc6] [Medline: 20881826]
6. Schwartz CC, Ajjarapu AS, Stamy CD, Schwinn DA. Comprehensive history of 3-year and accelerated US medical school programs: a century in review. Med Educ Online 2018 Dec;23(1):1530557 [FREE Full text] [doi: 10.1080/10872981.2018.1530557] [Medline: 30376794]
7. Liaison Committee on Medical Education. Functions and structure of a medical school: standards for accreditation of medical education programs leading to the MD degree. Association of American Medical Colleges and American Medical Association. 2022. URL: https://www.umassmed.edu/globalassets/office-of-undergraduate-medical-education-media/faculty-handbook-2022/lcme-standards---march-2022.pdf [accessed 2026-05-25]
8. Dreyfus SE. The five-stage model of adult skill acquisition. Bull Sci Technol Soc 2004 Jun;24(3):177-181. [doi: 10.1177/0270467604264992]

XSL•FO
RenderX

9. Mylopoulos M, Regehr G. Cognitive metaphors of expertise and knowledge: prospects and limitations for medical education. Med Educ 2007 Dec;41(12):1159-1165. [doi: 10.1111/j.1365-2923.2007.02912.x] [Medline: 17986193]
10. Warm EJ, Desai SS, Bowen JL. Navigating the discontinuity crisis in medical education. N Engl J Med 2025 Jun 26;392(24):2447-2457. [doi: 10.1056/NEJMra2408679] [Medline: 40561530]
11. Holmgren AJ, Lindeman B, Ford EW. Resident physician experience and duration of electronic health record use. Appl Clin Inform 2021 Aug;12(4):721-728 [FREE Full text] [doi: 10.1055/s-0041-1732403] [Medline: 34348409]
12. Chaiyachati KH, Shea JA, Asch DA, Liu M, Bellini LM, Dine CJ, et al. Assessment of inpatient time allocation among first-year internal medicine residents using time-motion observations. JAMA Intern Med 2019 Jun 01;179(6):760-767 [FREE Full text] [doi: 10.1001/jamainternmed.2019.0095] [Medline: 30985861]
13. Pape HC, Pfeifer R. Restricted duty hours for surgeons and impact on residents quality of life, education, and patient care: a literature review. Patient Saf Surg 2009 Feb 20;3(1):3 [FREE Full text] [doi: 10.1186/1754-9493-3-3] [Medline: 19232105]
14. Vucicevic D, Mookadam F, Webb BJ, Labonte HR, Cha SS, Blair JE. The impact of 2011 ACGME duty hour restrictions on internal medicine resident workload and education. Adv Health Sci Educ Theory Pract 2015 Mar;20(1):193-203. [doi: 10.1007/s10459-014-9525-5] [Medline: 24916955]
15. Mullen MG, Salerno EP, Michaels AD, Hedrick TL, Sohn MW, Smith PW, et al. Declining operative experience for junior-level residents: is this an unintended consequence of minimally invasive surgery? J Surg Educ 2016;73(4):609-615 [FREE Full text] [doi: 10.1016/j.jsurg.2016.02.010] [Medline: 27066854]
16. Fonseca AL, Reddy V, Longo WE, Gusberg RJ. Graduating general surgery resident operative confidence: perspective from a national survey. J Surg Res 2014 Aug;190(2):419-428. [doi: 10.1016/j.jss.2014.05.014] [Medline: 24908164]
17. Damewood RB, Blair PG, Park YS, Lupi LK, Newman RW, Sachdeva AK. "Taking training to the next level": the American College of Surgeons committee on residency training survey. J Surg Educ 2017;74(6):e95-105. [doi: 10.1016/j.jsurg.2017.07.008] [Medline: 28781132]
18. Initial certification pass rates: certification exam ultimate pass rates 2008-2025. American Board of Internal Medicine. URL: https://www.abim.org/media/5hhbskg2/certification-pass-rates.pdf [accessed 2026-11-01]
19. Application and examination activity certification. American Board of Emergency Medicine. URL: https://www.abem.org/wp-content/uploads/2025/02/Exam-App-Info-Updated021325.pdf [accessed 2025-10-21]
20. Reisdorff EJ, Keim SM, Gorgas DL, White SR, Kendall JL, Ruff KC, et al. Declining performance on American Board of Emergency Medicine written examinations. AEM Educ Train 2025 Oct 24;9(5):e70105. [doi: 10.1002/aet2.70105] [Medline: 41141361]
21. Peets A, Ayas NT. Restricting resident work hours: the good, the bad, and the ugly. Crit Care Med 2012 Mar;40(3):960-966. [doi: 10.1097/CCM.0b013e3182413bc5] [Medline: 22297631]
22. Fabricant PD, Dy CJ, Dare DM, Bostrom MP. A narrative review of surgical resident duty hour limits: where do we go from here? J Grad Med Educ 2013 Mar;5(1):19-24 [FREE Full text] [doi: 10.4300/JGME-D-12-00081.1] [Medline: 24404221]
23. Bilimoria KY, Chung JW, Hedges LV, Dahlke AR, Love R, Cohen ME, et al. National cluster-randomized trial of duty-hour flexibility in surgical training. N Engl J Med 2016 Feb 25;374(8):713-727. [doi: 10.1056/NEJMoa1515724] [Medline: 26836220]
24. Silber JH, Bellini LM, Shea JA, Desai SV, Dinges DF, Basner M, et al. Patient safety outcomes under flexible and standard resident duty-hour rules. N Engl J Med 2019 Mar 07;380(10):905-914 [FREE Full text] [doi: 10.1056/NEJMoa1810642] [Medline: 30855740]
25. Desai SV, Asch DA, Bellini LM, Chaiyachati KH, Liu M, Sternberg AL, et al. Education outcomes in a duty-hour flexibility trial in internal medicine. N Engl J Med 2018 Apr 19;378(16):1494-1508 [FREE Full text] [doi: 10.1056/NEJMoa1800965] [Medline: 29557719]
26. Seligman ME. Flourish: A Visionary New Understanding of Happiness and Well-Being. New York, NY: Free Press; 2011.
27. Tabja Bortesi JP, Becerra MP, Ranisau J, Wen B, Nadesan P, Devereaux PJ, et al. AI-based automation for medication reconciliation: scoping review. J Med Internet Res 2026 May 11;28:e86760 [FREE Full text] [doi: 10.2196/86760] [Medline: 42114154]
28. Peters M, Ten Cate O. Bedside teaching in medical education: a literature review. Perspect Med Educ 2014 Apr;3(2):76-88 [FREE Full text] [doi: 10.1007/s40037-013-0083-y] [Medline: 24049043]
29. Garibaldi BT, Russell SW. Strategies to reinvigorate the bedside clinical encounter. N Engl J Med 2025 Nov 27;393(21):2142-2150. [doi: 10.1056/NEJMra2500226] [Medline: 41223363]
30. Gonzalo JD, Chuang CH, Huang G, Smith C. The return of bedside rounds: an educational intervention. J Gen Intern Med 2010 Aug;25(8):792-798 [FREE Full text] [doi: 10.1007/s11606-010-1344-7] [Medline: 20386997]
31. Rubisch HP, Blaschke AL, Berberat PO, Gartmeier M. Making student mistakes didactically useful in bedside teaching: relevance of learning content, teacher and patient presence. BMC Med Educ 2026 Jul 28;26(1):1216 [FREE Full text] [doi: 10.1186/s12909-026-09978-y] [Medline: 42522036]
32. Richardson KM, Cristiano JA, Schafer KR, Shen E, Burns CA. Writing is thinking: implementation and evaluation of an internal medicine residency clinical reasoning and documentation curriculum. Med Sci Educ 2022 Jun 13;32(4):773-777 [FREE Full text] [doi: 10.1007/s40670-022-01570-5] [Medline: 36035531]

XSL•FO
RenderX

33. Fried AJ, Dorn SD, Leland WJ, Mullen E, Williams DM, Zaas AK, et al. Large language models in internal medicine residency: current use and attitudes among internal medicine residents. Discov Artif Intell 2024 Oct 24;4:70. [doi: 10.1007/s44163-024-00173-w]
34. Patterson J, Kovacs M, Lees C. Ambient artificial intelligence scribes: a pilot survey of perspectives on the utility and documentation burden in palliative medicine. Healthcare (Basel) 2025 Aug 26;13(17):2118 [FREE Full text] [doi: 10.3390/healthcare13172118] [Medline: 40941470]
35. Ke Y, Jin L, Ong JC, Thirunavukarasu AJ, Car J, Cheung CY, et al. AI-induced never-skilling in medical education. Nat Med 2026 Jun;32(6):1997-2006. [doi: 10.1038/s41591-026-04438-y] [Medline: 42174254]
36. Abdulnour RE, Gin B, Boscardin CK. Educational strategies for clinical supervision of artificial intelligence use. N Engl J Med 2025 Aug 21;393(8):786-797. [doi: 10.1056/NEJMra2503232] [Medline: 40834302]
37. Bowker D, Torti J, Goldszmidt M. Documentation as composing: how medical students and residents use writing to think and learn. Adv Health Sci Educ Theory Pract 2023 May;28(2):453-475 [FREE Full text] [doi: 10.1007/s10459-022-10167-x] [Medline: 36319807]
38. Principles for the responsible use of artificial intelligence in and for medical education. Association of American Medical Colleges. URL: https://www.aamc.org/about-us/mission-areas/medical-education/principles-ai-use [accessed 2026-09-08]
39. Frank JR, Snell LS, Cate OT, Holmboe ES, Carraccio C, Swing SR, et al. Competency-based medical education: theory to practice. Med Teach 2010;32(8):638-645. [doi: 10.3109/0142159X.2010.501190] [Medline: 20662574]
40. ten Cate O, Burch VC, Chen HC, Chou FC, Hennus MP. Entrustable Professional Activities and Entrustment Decision-Making in Health Professions Education. London, UK: Ubiquity Press; 2024.
41. What is competency-based medical education? Society for Academic Emergency Medicine. URL: https://www.saem.org/education/saem-online-academic-resources/saem-competency-based-medical-education-primer/what-is-competency-based-medical-education [accessed 2026-05-11]
42. Cooper D, Holmboe ES. Competency-based medical education at the front lines of patient care. N Engl J Med 2025 Jul 24;393(4):376-388. [doi: 10.1056/NEJMra2411880] [Medline: 40700689]
43. Hamza DM, Hauer KE, Oswald A, van Melle E, Ladak Z, Zuna I, et al. Making sense of competency-based medical education (CBME) literary conversations: a BEME scoping review: BEME guide no. 78. Med Teach 2023 Aug;45(8):802-815. [doi: 10.1080/0142159X.2023.2168525] [Medline: 36668992]
44. Brydges R, Boyd VA, Tavares W, Ginsburg S, Kuper A, Anderson M, et al. Assumptions about competency-based medical education and the state of the underlying evidence: a critical narrative review. Acad Med 2021 Feb 01;96(2):296-306. [doi: 10.1097/ACM.0000000000003781] [Medline: 33031117]
45. Smith BK, Yamazaki K, Tekian A, Brooke BS, Mitchell EL, Park YS, et al. Accreditation council for graduate medical education milestone training ratings and surgeons' early outcomes. JAMA Surg 2024 May 01;159(5):546-552 [FREE Full text] [doi: 10.1001/jamasurg.2024.0040] [Medline: 38477914]
46. Chen JX, Mylod D, Jing Y, Kerschner M, Trock B, Meth S, et al. Trainee-physician milestones ratings and patient experience surveys in early unsupervised practice. JAMA Netw Open 2025 Oct 01;8(10):e2536380 [FREE Full text] [doi: 10.1001/jamanetworkopen.2025.36380] [Medline: 41060650]
47. Gray BM, Vandergrift JL, Stevens JP, Lipner RS, McDonald FS, Landon BE. Associations of internal medicine residency milestone ratings and certification examination scores with patient outcomes. JAMA 2024 Jul 23;332(4):300-309. [doi: 10.1001/jama.2024.5268] [Medline: 38709542]
48. Borges MC, Kennedy MM, Kropf CW, Caldwell MD, Huang RD, Kim GJ, et al. Cost and challenges of clinical competency committees and milestones assessment. J Med Educ Curric Dev 2025 May 25;12:23821205251336279 [FREE Full text] [doi: 10.1177/23821205251336279] [Medline: 40432835]
49. ten Cate O. Entrustability of professional activities and competency-based training. Med Educ 2005 Dec;39(12):1176-1177. [doi: 10.1111/j.1365-2929.2005.02341.x] [Medline: 16313574]
50. Ten Cate O, Taylor DR. The recommended description of an entrustable professional activity: AMEE guide no. 140. Med Teach 2021 Oct;43(10):1106-1114 [FREE Full text] [doi: 10.1080/0142159X.2020.1838465] [Medline: 33167763]
51. Goldhamer ME, Martinez-Lage M, Black-Schaffer WS, Huang JT, Co JP, Weinstein DF, et al. Reimagining the clinical competency committee to enhance education and prepare for competency-based time-variable advancement. J Gen Intern Med 2022 Jul;37(9):2280-2290 [FREE Full text] [doi: 10.1007/s11606-022-07515-3] [Medline: 35445932]
52. Regan L, McGee D, Davis F, Murano T. Building the future curriculum for emergency medicine residency training. J Grad Med Educ 2025 Apr;17(2):248-253 [FREE Full text] [doi: 10.4300/JGME-D-25-00101.1] [Medline: 40417083]

Abbreviations

CBME: competency-based medical education
EHR: electronic health record
EPA: entrustable professional activity
GME: graduate medical education

XSL•FO
RenderX

Edited by A Stone; submitted 09.Jul.2026; peer-reviewed by JX Chen, M Hany; comments to author 31.Jul.2026; revised version received 25.Oct.2026; accepted 27.Oct.2026; published 09.Nov.2026.

Please cite as:
Dubosh NM, Iyer AA, Schwartzstein RM
From Competence to Expertise: The Role of Meaningful Experience in Physician Training
JMIR Med Educ 2026;12:e106585
URL: https://mededu.jmir.org/2026/1/e106585
doi:10.2196/106585
PMID:

©Nicole M Dubosh, Alexander A Iyer, Richard M Schwartzstein. Originally published in JMIR Medical Education (https://mededu.jmir.org), 09.Nov.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Education, is properly cited. The complete bibliographic information, a link to the original publication on https://mededu.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Viewpoint

Defining and Exploring Experience in Health Education

Louis-Baptiste Jaunay[1,2,3], MD, MSc, PhD; Emmanuel Triby[1], PhD; Thierry Pelaccia[1,4], MD, PhD

[1]Interuniversity Laboratory of Education and Communication Sciences (LISEC, EA 2310), Faculty of Educational Sciences, University of Strasbourg, Strasbourg, France

[2]Department of General Practice, Faculty of Medicine, Université Paris Cité, Paris, Île-de-France, France

[3]Unit for Education and Research in Health (UPPERS, US 007), Faculty of Health, Université Paris Cité, Paris, Île-de-France, France

[4]Centre for Training and Research in Health Sciences Education (CFRPS), Faculty of Medicine, University of Strasbourg, Strasbourg, France

Corresponding Author:
Louis-Baptiste Jaunay, MD, MSc, PhD
Department of General Practice
Faculty of Medicine
Université Paris Cité
Site Cochin, 24, rue du Faubourg Saint-Jacques
Paris, Île-de-France, 75014
France
Phone: 33 1 44 41 23 63
Fax: 33 1 44 41 23 64
Email: louis-baptiste.jaunay@u-paris.fr

Abstract

The term “experience” is widely used in health education literature. However, it has not been explicitly defined in this field, which could lead to multiple interpretations by researchers, teachers, and health students. Philosophical works offer clear definitions of “experience,” which can provide useful theoretical insights for its appropriation in research, practice, and teaching. Notably, pragmatist, rationalist, and phenomenological philosophers John Dewey, Georges Canguilhem, and Maurice Merleau-Ponty define experience as the significance that emerges during the interaction between a subject and its environment. They distinguish between episodic (making an experience) and cumulative (acquiring experience) dimensions. The article further explores the implications of this definition in health education, particularly in relation to experiential learning, simulation-based education, and assessment. This conceptualization has also paved the way for a method of exploring experience designed to facilitate awareness, known as the “explicitation interview.” Studies using this method have yielded interesting results in understanding the activities of students, health care professionals, and patients in real-life situations, particularly regarding the influence of context on reasoning and decision-making. The explicitation interview can also be used as a training tool through the supervision of past situations. Practical guidance for its implementation in educational settings is provided. Overall, highlighting individuals’ meaning-making processes opens new perspectives for research, teaching, and assessment in health professions education.

(JMIR Med Educ 2026;12:e85039) doi:10.2196/85039

KEYWORDS

experience; explicitation interview; medical students; learning; interview skills

Introduction

The term “experience” is used extensively by students, teachers, patients, and researchers in the field of health professions education (HPE). It can refer to clinical experience (eg, “early clinical experience” [1] or “experienced clinician” [2]). In other cases, it refers to learning experiences, as something induced by educators [3]. In addition, the term is used to describe a learning approach, experience-based learning, in contrast to classroom-based learning [4]. In phenomenological research, it is also mobilized in studies investigating the lived experience of professionals [5], students [6], or patients [7].

The term is so overused that it is rarely defined, as if its meaning were self-evident, and its various applications warrant further examination. We argue that existing uses of the concept in HPE often oscillate between a broad, underspecified notion and narrow operationalizations that risk reducing its scope.

This lack of clarity creates practical challenges for educators. It affects how experience is elicited and analyzed, as well as how debriefings are conducted and how learning environments

XSL•FO
RenderX

are designed and assessed. In the development of competency frameworks, early stages of in situ analysis of practitioners' experience for competency mapping are often overlooked in favor of expert opinion. During debriefing or feedback, attention may focus primarily on observable performance or expressed opinions, while overlooking how learners perceived, acted, and interpreted the situation. When designing learning environments, educators may aim to improve learners' experience (eg, by structuring simulation scenarios to reduce stress and uncertainty) without clearly identifying which dimensions are involved (eg, meaning-making processes). In evaluating learning environments, "experience" can be reduced to questionnaires that measure a set of predetermined parameters, which are disconnected from real-life situations and focus more on students' opinions and satisfaction [8]. These issues are particularly salient in contemporary HPE, where learning increasingly takes place in artificial environments, such as simulation [9]. This raises practical questions about the nature of the experiences afforded and their relationship to real clinical situations, learning objectives, and assessment practices.

In this context, our primary contribution is conceptual: we propose a theoretically grounded definition of experience that aims to clarify its use in HPE. Building on this clarification, we then examine its methodological implications and the perspectives it opens up for debriefing in learning and clinical settings.

Advancements stemming from constructivist research suggest viewing experience not as context-independent information that can be easily measured and transferred but rather as an active construction by each individual based on their prior experiences and interactions with others [10]. Consequently, it became crucial to describe individuals' experience from their point of view, that is, through a phenomenological approach [11]. In phenomenological research, experience is mainly explored through interviews [12]. However, the literature does not provide consensual practical guidelines for conducting such interviews [13]. Most studies rely on in-depth semistructured interviews that focus on participants' perceptions and meaning-making [12,13]. However, this approach is prone to memory reconstruction. As a result, the data collected often reflect general and decontextualized judgments disconnected from the specific situation under study [14]. This limitation arises from considering experience solely in its passive dimension, whereas perceptions and meaning are inextricably linked to the subject's activity: sensations mark interruptions in the course of action, inciting meaning-making [15]. In phenomenological terms, prereflective consciousness—that is, the part of our experience that is lived without being recognized or without being immediately accessible to consciousness and verbal description—must be seen as consciousness in action, requiring guidance to become aware [16]. This observation led to the development of an interview method called the explicitation interview (EI). The EI explores experience by focusing on the interviewee's action-perception-meaning triad in lived situations. It aims to produce a detailed description of the lived experience, using action as the guiding thread [17]. It is particularly well aligned with the aims of phenomenological inquiry [18].

We first present a definition of experience based on the work conducted in 3 major philosophical currents that have studied this concept in the 20th century. Then, we attempt to engage the reader in experiencing this definition through a literary illustration. Following this, we discuss the connections this definition has with its developments in the fields of education and, particularly, HPE. Next, we introduce the EI, describing the techniques it uses to focus on activity, along with the results it has yielded in the field of HPE. Finally, we propose an example of an interview guide to be tested with students.

Definition of Experience

Pragmatist, phenomenological, and rationalist philosophers, including John Dewey, Maurice Merleau-Ponty, and Georges Canguilhem, have devoted part of their work to the conceptualization of experience. John Dewey (1859-1952) focused his work on a philosophy of experience grounded in action and the relationship between organism and environment, which he also extended to education through the idea of "learning by doing." Maurice Merleau-Ponty (1908-1961) emphasized the embodied and situated nature of experience. Georges Canguilhem (1904-1995), working at the intersection of philosophy and medicine, underscored how individuals actively construct meaning and norms in relation to their milieu, offering a perspective that resonates with patient-centered care. Although emanating from different traditions, they share a constructivist vision of experience built on the Darwinist understanding of the dynamic relationship between living beings and their environment [19].

Their conceptualizations of experience share a common ground in defining it as the meaning that emerges between the action and the perception of subjects during the interaction that links them to their environment [20-22]. In this definition, "subjects" refers to the agents who give meaning, and "environment" refers to the subjects' world, both external (whether inert or social, ie, embodied by other subjects) and internal (embodied through sensations and emotions). A continuity of experience emerges in the course of interaction, which means that every episodic experience both takes up something from those that have gone before and modifies in some way the quality of those that come after [23].

This definition is consistent with the uses of the word "experience" in academic, nonacademic, and philosophical discourses [24].

An Illustration

In the following text, we develop these theoretical considerations by illustrating them with a passage from Proust's [25] novel *Swann's Way*, segmented into 3 quotations:

> *The year before, at an evening party, Swann had heard a piece of music played on the piano and violin. At first, he had appreciated only the material quality of the sounds which those instruments secreted.*
>
> *But at a given moment, without being able to distinguish any clear outline, or to give a name to what was pleasing him, suddenly enraptured, he had*

XSL•FO
RenderX

> *tried to collect, to treasure in his memory the phrase or harmony—he knew not which—that had just been played, and had opened and expanded his soul, just as the fragrance of certain roses, wafted upon the moist air of evening, has the power of dilating our nostrils [25].*

In this situation (ie, the singular contextual whole composed of Swann, the evening party, and their interaction), Swann lives a meaningful episodic experience. The relationship between Swann and the piece of music evolves over time, moving from continuity (in the first paragraph) to disruption (in the next).

This first quotation allows a specification of each author's specific contribution to the concept of experience. Merleau-Ponty [20] emphasized the role of the body in linking perception and action: the perceptions of Swann arise from his embodied activity, and vice versa. This idea was later developed in cognitive science through the concept of enaction, which views cognition as emerging from interactions with the environment [26]. According to enaction, perception consists of action guided by perception, and cognitive structures emerge from recurrent sensorimotor patterns that enable such guidance [26]. Canguilhem [21] showed that the milieu is shaped by the subject's activity, norms, and values: the quality emerging in Swann's experience arouses his interest and modifies both his perception and his engagement. Dewey [22] further articulated that the meaning constructed by the subject is not only related to its purpose but also to the logic that emerges in the relationship between the subject's actions and their consequences. He formulated a theory to describe the emergence of all forms of logic from experience, called the theory of inquiry [27]. Inquiry, in the sense of Dewey, begins with this perception of indeterminacy that Swann experiences in relation to the musical phrase, which leads him to a reflective effort [25]:

> *And so, hardly had the delicious sensation, which Swann had experienced, died away, before his memory had furnished him with an immediate transcript, summary, it is true, and provisional, but one on which he had kept his eyes fixed while the playing continued, so effectively that, when the same impression suddenly returned, it was no longer uncapturable.*

Experience is accumulated: the episodic experience Swann has enacted and undergone influences the experiences that follow. This accumulated experience constitutes Swann's singular culture, the memory of which he mobilizes in the course of his present experience [25]:

> *But when he was once more at home, he needed it, he was like a man into whose life a woman, whom he has seen for a moment passing by, has brought a new form of beauty, which strengthens and enlarges his own power of perception* [...]
>
> *Indeed this passion for a phrase of music seemed, in the first few months, to be bringing into Swann's life the possibility of a sort of rejuvenation.*

The change in Swann is meaningful because it is connected with the return wave of consequences that flow from it. The modification of the course of Swann's activity is reflected in a change made in him, loaded with significance: he learned something [22]. Language can help Swann put this experience and learning into a narrative, objectifying and sharing them [28].

This literary illustration can be directly related to educational practice in HPE, as it exemplifies how meaning emerges from the unfolding of action and perception in a situation. Making this dynamic accessible is central to educational practices, such as debriefing, where educators aim to understand not only what learners achieved but also how they experienced the situation.

Developments in Educational Sciences

The theoretical foundations presented so far emphasize that experience arises through the sensemaking activities of individuals in relation to their environment over time. They have been extended by research in educational science, which has explored this interaction in school and professional contexts.

The link between experience and learning has been studied under the term "experiential learning," following the work of Dewey [23] and his advocacy for educational systems that prepare individuals for richer future experiences. Experiential learning has developed in contrast to a "traditional" learning approach that focuses on the transmission of theoretical knowledge before it is put into practice by the learner [29]. For instance, traditional methods often require students to memorize the structure of a medical interview before meeting a patient, whereas experiential learning favors the use of practical situations. Among the significant authors of experiential learning, Knowles [30], Kolb [31], Coleman [29], and Jarvis [32] each developed specific theoretical aspects in line with the definition of experience presented here. Schematically, Knowles [30] developed the importance of anchoring learning in the learner's accumulated experience, and Kolb [31] modeled experiential learning as a cycle between reflection and experimentation, highlighting the importance of the action-perception-meaning coupling in learning. Coleman [29] detailed the nature of the relationship between experiential learning and the transmission of symbolic information, while Jarvis [32] emphasized the social dimension of experience in meaning-making. Jarvis defined learning as a continuous process of making sense of everyday experiences through interactions with time, space, society, and their multiple relationships [32].

These developments have given rise to many studies in the field of HPE, sometimes referring to simulation [33], gamification [34], or the clinical environment [4], which can be understood in the light of our definition of experience.

The notion of interaction between the subject and the environment prompts an examination of the triggers and feedback provided by learning environments, particularly artificially constructed settings such as simulations or gamification. It also emphasizes the importance of engaging learners in active learning environments where their actions have a perceptible impact [35].

The importance of the meaning that participants give to their activity in the constitution of the experience raises questions

XSL•FO
RenderX

about the values conveyed by the educational context. As Coleman [29] has pointed out, simulations or games can transform the traditional teaching of theoretical knowledge into engaging learning experiences, as they artificially create meaning through the scenario and the objectives of the activity. Rather than reproducing the full complexity of authentic clinical situations—including the challenges of care and the richness of patients' accumulated experience—these environments are designed to provide safe, structured, and standardized conditions for learning, practice, and assessment [9]. This leads to 3 practical implications. First, during scenario design, instructors can intentionally highlight specific dimensions of experience, for example, diagnostic uncertainty, communication in difficult interactions, or attention management under pressure. Second, during observation and assessment, educators may complement performance-based evaluation by exploring how learners reason within the situation (eg, what they notice, how they prioritize, and what goals guide their actions). Third, during debriefing, explicitly discussing how the scenario shaped the experience (eg, through predefined roles or objectives) may help learners understand the limits and affordances of the simulation and support transfer to clinical practice [36].

Whether in a clinical or simulated setting, this focus on meaning also invites inquiry into the clinical sensemaking activities of students [37]: How do students discern salient situational elements? What meaning do they give to their relationship with the patient and care context? How do they gain further insight into the situation? What meaning do they give to their own actions? What tensions emerge between their narrative and those of others? These questions lead the educator to clarify the significance of the values to be communicated, particularly that of health, as a goal of care and HPE. Canguilhem [38] developed an experiential conception of health as the individual's feeling of being able to interact with their environment in their own way [19]. Similar to the contrast between experiential learning and traditional education based on the transmission of theoretical knowledge, this definition of health emerged in opposition to the tendency inherited from physiology to assess individuals' health according to externally established quantitative norms.

The principle of continuity of experience offers a way of examining the relationship between educational contexts and learning:

- Does learners' prior experience enable them to interact with this context? This question raises the issue of assessing the resources that learners can mobilize and the difficulties posed by the situation in which they are placed, so that supportive actions, such as debriefing, can be tailored accordingly.
- Will the experience they have in this context enable them to engage in future experiences in a more enriched way? This question of the transfer of learning also leads to an examination of the often-misleading correspondence between artificial learning environments and the unapparent complexity of real work situations. This examination can be discussed with students during debriefing [36].

Verbalization is a valuable approach for exploring past activities from the subjects' perspective [26]. However, the subject's verbalization of his or her experience is not a given: it requires awareness [16]. It is very difficult, for example, to describe precisely the succession of actions and perceptions involved in reading this article: what rhythm I am reading at, what ideas come to me as I read, when and how my attention drifts to other ancillary tasks, how the atmosphere of the place I am in influences my reading, and so on. Our most immediate experience, the one we live here and now, is also the most foreign to us [16]. Thus, the "talk-aloud" method, which involves asking the interviewee to verbalize their experience while performing the task, does not allow a detailed account while maintaining a certain coherence in the execution of the task they are attempting to comment on [14]. Among retrospective methods, the "critical incident technique" and "depth interviewing" are subject to problems of constructed recollection [14]. Indeed, when a subject tries to describe their experience, they often formulate general theoretical knowledge, explanations, and justifications [16]. It is therefore useful to explore a subject's experience by accompanying them toward a speaking position in which they relive the past situation [39]. The EI has been developed specifically to address this challenge [40].

A Method for Exploring Experience: the EI

The EI is a method formalized by Vermersch [17] in the 1990s to help subjects become aware of and verbalize their experience. It can be used to gather information about what is not observable in the interviewee's activity (particularly his or her logic) and/or what is no longer observable [41]. It can also be used for pedagogical purposes, enabling the interviewee to reflect on his or her own way of acting. Sometimes, it is combined with traces of the subject's activity, such as audiovisual recordings [42]. It is based on a specific way of relating to the subject's experience, which aims to avoid the usual explanatory speaking position, in which the interviewee is turned toward the other with the intention of communicating but in which the description of lived experience is altered by generalizations and reconstructions. The EI, by contrast, is designed to guide individuals toward a mode of embodied, introspective expression, to reconnect with the lived experience being explored [43]. It explores real-life situations, precisely situated in space and time. However, the type of memory targeted by this act of recall, which can be described as embodied memory, is inhibited by conscious effort to remember. A fundamental aspect of the method is to elicit effortless recall, a process of letting memories emerge spontaneously [17]. It is therefore preferable to allow the interviewee to choose the situation to remember by using phrasing such as "take your time...to let a memory...come to you..." rather than prompting an effort to recall with "remember precisely..." Additionally, because this type of memory is intimate and involves exploring information that the subject has not yet consciously accessed, an explicit prior communication agreement is required ("...if you agree," along with careful attention to the nonverbal dimension of the response) for ethical and functional reasons. Verbalization is promoted through reformulations and specific open-ended questions, empty of content, formulated in the present tense and focused on

XSL•FO
RenderX

describing the articulation between perceptions and actions, as well as the goals being pursued (eg, "What do you perceive?," "How do you go about it?," and "What goal are you aiming for by doing this?") [17]. Meanwhile, questions related to the explanation or causality of the situation (eg, "Why did you...?" and "Can you explain...?") are avoided, as they tend to generate interpretations, justifications, and comments that are disconnected from the actual experience. Certain observable clues indicate whether the interviewee is reconnecting with past experience (unhooking the gaze, slowing the pace of speech, congruence of verbal and nonverbal communication, use of "I" verbalization in the present tense, and descriptive vocabulary specific to the situation and its sensory connotations). The progression of the activity is explored step-by-step and can also be broken down into smaller steps by asking the following: "And when you do this, how do you go about it?" The moments of apparent inaction are explored through EI-specific questioning, such as "When you're not doing anything, what are you doing?" or "When you're not doing anything, what do you pay attention to?" This type of EI-specific question also helps overcome certain impasses linked to denials from the interviewee, such as "I don't remember" or "I don't know." These denials are to be considered not as missing information but as judgments that the interviewee makes about the state of their knowledge. In this regard, it can be fruitful to ask follow-up questions, such as "When you don't remember, what do you remember?" or "How do you know you don't know?" [17].

The EI can be used at several distinct yet complementary stages of research in HPE. First, by enabling verbalization of experience, it allows for the exploration and analysis of actors' practices (including students, professionals, and patients), facilitating an understanding of their underlying logic and contributing to the development of competency frameworks [44]. Second, as it helps actors become aware of their experience and express it in words, it can be used during debriefings and thus contribute to learning and the development of reflexivity. Third, because it explores their interaction with their environment, it facilitates the design of artificial learning environments (eg, simulations) that more accurately replicate the resources, error sources, and obstacles present in natural settings. Fourth, the exploration of learners' experience in the learning environment can lead to a better understanding of how the latter is likely to enable the transfer of learning to the clinical setting and adapt it accordingly.

Empirical Contributions of the EI in Health and HPE

The EI has led to interesting discoveries in the fields of health and HPE, both in research and in training.

In the study of caregivers, it revealed elements of their experience that are not immediately apparent or spontaneously accessible. Analysis of the first impressions of caregivers during an encounter with a patient revealed the influence of external (ambience and patient's attitude) and internal (bodily sensations and emotions) perceptions on the caregiver's generation of hypotheses (about the patient's internal state and intentions and about the relationship) and decision-making [45]. It also provided directions for understanding the actions and perceptions involved in clinical attention [46] and building a therapeutic alliance [47]. The exploration of the experiences of physicians working in medical emergency call centers has highlighted experiential knowledge, such as the physician's attentional dynamics when collaborating with both the team and the patient, as well as the crucial role played by the patient in the coconstruction of a shared context [48]. Among expert emergency physicians, the use of EI coupled with video recordings from a subjective perspective (via camera glasses) revealed the importance of intuitive reasoning and emotions in clinical reasoning [49]. With expert optometrists, it highlighted the importance given to the patient's point of view and planning in therapeutic reasoning [50].

In health care training, it has been used to explore the development of reasoning between preclinical and early clinical training of medical students, particularly the emergence of scripts [51]. It has also been used to support the development of emotional skills among nursing students by exploring significant caregiving situations [52]. An EI conducted during the training of childcare assistants regarding situations considered successful allowed them to become aware of their knowledge, the lessons learned, and the challenges faced, enabling them to implement adjustment actions [53]. The use of EI with students in family therapy training appears to contribute to the development of their professionalism, while supervisors enhance their relational skills [54]. Among expert nurse trainers skilled in technical procedures but uncertain about their teaching abilities, discussing past situations has enabled them to recognize their procedural knowledge, articulate it, and consider evolving their teaching methods [55].

The EI has also been used to study the experience of patients. It revealed the existence of prodromal symptoms in patients with epilepsy, in contradiction with the medical literature, according to which seizures are sudden and cannot be anticipated by the patient (right down to the etymology of the word "epilepsy," which means "to surprise") [56]. It has been used to explore the experience of breathing in a patient with pulmonary fibrosis [7], the impact of a widespread chronic pain condition in the bodily experience of people with fibromyalgia [57], the implicit thoughts and perceptions involved in emotional eating [58], older age challenges and measures to safeguard their existential survival and sustainability [59], and the experience of performing spirometry in patients with chronic obstructive pulmonary disease [60]. A study of patients' experiences of health problems during primary care consultations showed how patients identify and articulate their concerns, generate and compare hypotheses, and make decisions about how to address them, including whether to schedule an appointment with a general practitioner, as well as how they interpret the physicians' messages [61]. The study also emphasized the role that physicians can play in recognizing and supporting patients' inquiries.

Practical Implementation of the EI

To consider implementing the EI in your training practices, we invite you to try out the questioning style of this method using

XSL•FO
RenderX

the examples presented below. Please note that this is a written overview of the interview method, which does not include the techniques necessary to help stabilize the interviewee's attention on the situation in question.

The following guided exercise invites you to practice the questioning style of the EI by revisiting a recent teaching situation, so that you can experience its effects firsthand as a reader.

Sit down comfortably, take a deep breath, and let yourself be reminded of a recent situation in which you successfully completed a training session with students…

Take your time...

When you have identified the situation, read the following questions one by one, giving yourself time for the memories to return: (1) Where are you at that moment? (2) What's going on around you? (3) When you are there, what do you start with? (4) How do you go about it? (5) When you do that, what is your goal? (6) What happens to you when you do that? (7) Finally, in all of this, what is important to you?

This approach can be translated into practical strategies that are readily applicable in everyday teaching. For instance, educators can invite learners to revisit situations that were meaningful to them through either oral reflection or written formats, such as portfolios, to support the articulation of experiential knowledge.

It can also be integrated into routine teaching, even in time-constrained contexts. During bedside teaching, supervisors can introduce focused questions (eg, "What was your first impression of this patient?" "What came to mind when he says that?" and "What led you to respond in this way?") to help the learner make explicit how they experienced the clinical situation. In simulation debriefing, this perspective can complement established models by inviting learners to describe how they perceived the situation as it unfolded (eg, "At that moment, what did you notice first?" "When you notice that, where do you start?" and "What are you aiming for by doing that?"), thereby linking action, perception, and clinical reasoning. In the assessment of direct observations or indirect supervision, short elicitation sequences can be introduced—for instance, asking the learner to briefly describe a key moment of the encounter and what guided their action—before moving to formative and summative feedback. These brief prompts could help make learners' underlying interpretive frames more explicit, thereby enabling educators to provide more targeted feedback. They may also enable supervisors to better grasp the course of events during indirect supervision.

However, exploring experience through the EI also presents several limitations that warrant consideration. First, directing the interviewee's attention to specific moments of their past activity may require changes in habitual practices (eg, avoiding "why" questions) [17]. Second, the re-experiencing of past situations is shaped by the relationship established between the interviewer and the interviewee, as well as their shared understanding of the interview's aims [17]. Third, it depends on the interviewee's capacity to reaccess their lived experience, a skill that may vary between individuals and can be developed through practice [16]. More broadly, the phenomenological framework underlying the EI emphasizes the situated construction of meaning, which may not be compatible with assessment goals that aim for standardization and efficiency.

In this respect, the EI should be understood in relation to other qualitative approaches that provide access to different dimensions of experience. The think-aloud method provides access to concurrent verbalizations during task performance, offering insights into cognitive processes as they unfold [62]. Semistructured interviews, widely used in HPE research, provide insight into participants' representations and interpretations of past experiences [63]. Stimulated recall methods, based on video or audio recordings, enable participants to comment on specific moments of their activity, anchoring both the interviewee's and interviewer's reflections in observable data, and can be combined with elicitation techniques [42,64]. As a complement to these approaches, the EI offers a method that specifically aims to access the unfolding of action, perception, and meaning, including dimensions of experience that are not immediately available through spontaneous verbalization.

Conclusions

Despite the common and diverse uses of the term "experience," a unifying definition can be formulated as the meaning that emerges between action and perception in the interaction between individuals and their environment. The principle of continuity of experience allows us to distinguish between 2 main uses of the term: episodic (to experience) and cumulative (acquiring experience).

On the basis of this clarification, we have identified implications for research and education. We argued that studying experience requires approaches capable of accessing the dynamics of action, perception, and meaning and highlighted the potential contribution of the EI in this regard. At the pedagogical level, attention to learners' experience can enrich existing practices in teaching and assessment by making explicit the processes through which learners interpret and engage with situations.

The practical examples proposed in this article illustrate how this perspective can be implemented in routine educational settings, including those characterized by strong time constraints. By incorporating brief, targeted, and well-formulated prompts, educators can access otherwise implicit dimensions of learning, thereby fostering action-oriented reflection.

Beyond these pedagogical implications, taking into account the diversity of embodied sensemaking processes among professionals, students, and patients could enrich the development of more practical competency frameworks. We therefore call for the development of empirically grounded research to examine how experience can be elicited, analyzed, and integrated into educational practice. This includes examining the feasibility of experience-oriented approaches in routine teaching, their articulation with existing assessment frameworks, and their impact on learners' development over time.

In this perspective, the EI—and, more broadly, the principles that underpin it—offers a promising avenue, depending on the educational and research context, not only as a research method but also as a practical resource for supervision and feedback.

XSL•FO
RenderX

By rendering learners' meaning-making processes more accessible, it can contribute to the development of more coherent, reflective, and learner-centered approaches to HPE.

Acknowledgments

The authors thank Professor Elizabeth Macintyre for comments that improved the manuscript.

Funding

The authors declared that no financial support was received for this work.

Generative AI Use

The authors would like to acknowledge the professional translator who provided the initial English translation of the manuscript. ChatGPT (GPT-6, free version; OpenAI) was subsequently used to assist with final language editing and refinements. The authors reviewed all suggestions and take full responsibility for the final content.

Authors' Contributions

All authors jointly discussed the content of the article. LBJ wrote the first version of the manuscript, which was critically reviewed by the other authors. All authors approved the final version of the manuscript.

Conflicts of Interest

None declared.

References

1. Helmich E, Bolhuis S, Laan R, Koopmans R. Early clinical experience: do students learn what we expect? Med Educ 2011 Jul;45(7):731-740. [doi: 10.1111/j.1365-2923.2011.03932.x] [Medline: 21649706]
2. Zubairi MS, Lindsay S, Parker K, Kawamura A. Building and participating in a simulation: exploring a continuing education intervention designed to foster reflective practice among experienced clinicians. J Contin Educ Health Prof 2016;36(2):127-132. [doi: 10.1097/CEH.0000000000000065] [Medline: 27262157]
3. Leggat PA. Learning experiences in medical education. Med Teach 2009 Jul 03;22(3):288-292. [doi: 10.1080/01421590050006278]
4. Yardley S, Teunissen PW, Dornan T. Experiential learning: AMEE guide no. 63. Med Teach 2012;34(2):e102-e115. [doi: 10.3109/0142159X.2012.650741] [Medline: 22289008]
5. Whitehead PR. The lived experience of physicians dealing with patient death. BMJ Support Palliat Care 2014 Sep;4(3):271-276 [FREE Full text] [doi: 10.1136/bmjspcare-2012-000326] [Medline: 24644159]
6. Porteous DJ, Machin A. The lived experience of first year undergraduate student nurses: a hermeneutic phenomenological study. Nurse Educ Today 2018 Jan;60:56-61. [doi: 10.1016/j.nedt.2017.09.017] [Medline: 29045912]
7. Ladores S. Applying Petitmengin's explicitation interview method to elicit the lived experience of breathing upon waking by an individual with cystic fibrosis. J Patient Exp 2020 Dec;7(6):856-868 [FREE Full text] [doi: 10.1177/2374373520956740] [Medline: 33457512]
8. Mukhalalati B, Yakti O, Elshami S. A scoping review of the questionnaires used for the assessment of the perception of undergraduate students of the learning environment in healthcare professions education programs. Adv Health Sci Educ Theory Pract 2024 Sep;29(4):1501-1538. [doi: 10.1007/s10459-024-10319-1] [Medline: 38683300]
9. Elendu C, Amaechi DC, Okatta AU, Amaechi EC, Elendu TC, Ezeh CP, et al. The impact of simulation-based training in medical education: a review. Medicine (Baltimore) 2024 Jul 05;103(27):e38813 [FREE Full text] [doi: 10.1097/MD.0000000000038813] [Medline: 38968472]
10. Taylor DC, Hamdy H. Adult learning theories: implications for learning and teaching in medical education: AMEE guide no. 83. Med Teach 2013 Nov;35(11):e1561-e1572. [doi: 10.3109/0142159x.2013.828153] [Medline: 24004029]
11. Neubauer BE, Witkop CT, Varpio L. How phenomenology can help us learn from the experiences of others. Perspect Med Educ 2019 Apr;8(2):90-97 [FREE Full text] [doi: 10.1007/s40037-019-0509-2] [Medline: 30953335]
12. Tavakol M, Sandars J. Twelve tips for using phenomenology as a qualitative research approach in health professions education. Med Teach 2025 Sep;47(9):1441-1446 [FREE Full text] [doi: 10.1080/0142159X.2025.2478871] [Medline: 40089895]
13. Starks H, Trinidad SB. Choose your method: a comparison of phenomenology, discourse analysis, and grounded theory. Qual Health Res 2007 Dec;17(10):1372-1380. [doi: 10.1177/1049732307307031] [Medline: 18000076]
14. Light A. Adding method to meaning: a technique for exploring peoples' experience with technology. Behav Inf Technol 2006;25(2):175-187. [doi: 10.1080/01449290500331172]

XSL•FO
RenderX

15. Dewey J. Reconstruction in Philosophy. Mineola, NY: Dover Publications; 1921.
16. Petitmengin C. Describing one's subjective experience in the second person: an interview method for the science of consciousness. Phenomenol Cogn Sci 2006 Nov 29;5(3-4):229-269. [doi: 10.1007/s11097-006-9022-2]
17. Vermersch P. Describing the practice of introspection. J Conscious Stud 2009 Jan;16(10-12):20-57 [FREE Full text]
18. Høffding S, Martiny K. Framing a phenomenological interview: what, why and how. Phenomenol Cogn Sci 2016;15(4):539-564. [doi: 10.1007/s11097-015-9433-z]
19. Jaunay LB, Pelaccia T, Triby E. Enraciner la formation dans l'expérience des soignants et des soignés [Rooting training in the experience of caregivers and those cared for: the seeds of a transformation] [Article in French]. Éduc Perm 2024;241(4):22-34. [doi: 10.3917/edpe.241.0022]
20. Merleau-Ponty M. The Structure of Behavior. Oxford, England: Beacon Press; 1963.
21. Canguilhem G. Knowledge of Life. New York, NY: Fordham University Press; 2022.
22. Dewey J. Democracy and Education: An Introduction to the Philosophy of Education. New York, NY: Macmillan; 1916.
23. Dewey J. Experience and Education. New York, NY: Macmillan; 1938.
24. Sytsma J, Fischer E. 'Experience', ordinary and philosophical: a corpus study. Synthese 2023 Jun 08;201(6):210. [doi: 10.1007/s11229-023-04190-7]
25. Proust M. In Search of Lost Time: Volume 1: The Way by Swann's. New York, NY: Penguin Classics; 2004.
26. Varela FJ, Thompson E, Rosch E. The Embodied Mind: Cognitive Science and Human Experience. Cambridge, MA: MIT Press; 1991.
27. Dewey J. Logic: The Theory of Inquiry. New York, NY: Henry Holt and Company; 1938.
28. Merleau-Ponty M. Phenomenology of Perception. London, UK: Routledge; 2011.
29. Coleman JS. Experiential learning and information assimilation: toward an appropriate mix. J Exp Educ 1979 May 01;2(1):6-9. [doi: 10.1177/105382597900200102]
30. Knowles MS. The Adult Learner: A Neglected Species. 3rd Edition. Houston, TX: Gulf Publishing; 1984.
31. Kolb DA. Experiential Learning: Experience as the Source of Learning and Development. Englewood Cliffs, NJ: Prentice-Hall; 1984.
32. Jarvis P. Adult Learning in the Social Context. London, UK: Croom Helm; 1987.
33. Poore JA, Cullen DL, Schaar GL. Simulation-based interprofessional education guided by Kolb's experiential learning theory. Clin Simul Nurs 2014 May;10(5):e241-e247. [doi: 10.1016/j.ecns.2014.01.004]
34. Krishnamurthy K, Selvaraj N, Gupta P, Cyriac B, Dhurairaj P, Abdullah A, et al. Benefits of gamification in medical education. Clin Anat 2022 Sep;35(6):795-807. [doi: 10.1002/ca.23916] [Medline: 35637557]
35. Lebert A, Vilarroya Ó. The links between experiential learning and 4E cognition. Ann N Y Acad Sci 2024 Nov;1541(1):37-52. [doi: 10.1111/nyas.15238] [Medline: 39383206]
36. Rivière E, Jaffrelot M, Jouquan J, Chiniara G. Debriefing for the transfer of learning: the importance of context. Acad Med 2019 Jun;94(6):796-803. [doi: 10.1097/ACM.0000000000002612] [Medline: 30681450]
37. Koufidis C, Manninen K, Nieminen J, Wohlin M, Silén C. Clinical sensemaking: advancing a conceptual learning model of clinical reasoning. Med Educ 2024 Dec;58(12):1515-1527. [doi: 10.1111/medu.15461] [Medline: 38899766]
38. Canguilhem G. On the Normal and the Pathological. Dordrecht, The Netherlands: Springer; 1978.
39. Depraz N, Varela FJ, Vermersch P. On Becoming Aware: A Pragmatics of Experiencing. Amsterdam, The Netherlands: John Benjamins Publishing Company; 2003.
40. Froese T, Gould C, Barrett A. Re-viewing from within: a commentary on first- and second-person methods in the science of consciousness. Constr Found 2011 Mar;6(2):254-269 [FREE Full text]
41. Hunter AR, Morgan K, Harris K, Santos S, Mouchet A. Developing the use of explicitation techniques in team games as a pedagogical tool for coaching practice. Int J Sports Sci Coach 2023 Mar 06;18(4):1025-1035. [doi: 10.1177/17479541231159384]
42. Cahour B, Salembier P, Zouinar M. Analyzing lived experience of activity. Trav Hum 2016;79(3):259-284. [doi: 10.3917/th.793.0259]
43. Vermersch P. Introspection as practice. J Conscious Stud 1999 Jan;6(2-3):17-42 [FREE Full text]
44. Denis J, Tocquet M, Guillemette F, Hendrick S. Therapeutic processes in clinical interventions: a view of qualitative methodological approaches. Qual Rep 2022 Feb 18;27(2):578-590. [doi: 10.46743/2160-3715/2022.3620]
45. Ollagnier-Beldame M, Cazemajou A. Intersubjectivity in first encounters between healthcare practitioners and patients: micro-phenomenology as a way to study lived experience. Humanist Psychol 2019;47(4):404-425. [doi: 10.1037/hum0000133]
46. Katz E. Seeing clinical attention anew: results from an exploratory study of attention in clinical practice. J Soc Work Pract 2015 Mar 25;29(4):413-427. [doi: 10.1080/02650533.2015.1025374]
47. Coupé C, Ollagnier-Beldame M. Toward a phenomenology of taking care. Int J Qual Stud Health Well-being 2022 Dec;17(1):2045671 [FREE Full text] [doi: 10.1080/17482631.2022.2045671] [Medline: 35291910]
48. Mouchet A, Demeslay J, Bertrand C. Coopérer en régulation médicale au Samu centre 15 : l'expérience subjective des médecins urgentistes [Cooperating in medical regulation at the SAMU (Emergency Medical Service) center 15: the subjective experience of emergency physicians] [Article in French]. Trav Hum 2017;80(2):133-160. [doi: 10.3917/th.802.0133]

XSL•FO
RenderX

49. Pelaccia T, Tardif J, Triby E, Charlin B. A novel approach to study medical decision making in the clinical setting: the "own-point-of-view" perspective. Acad Emerg Med 2017 Jul;24(7):785-795. [doi: 10.1111/acem.13209] [Medline: 28449293]
50. Faucher C, Tardif J, Chamberland M. Optometrists' clinical reasoning made explicit: a qualitative study. Optom Vis Sci 2012 Dec;89(12):1774-1784. [doi: 10.1097/OPX.0b013e3182776002] [Medline: 23160443]
51. Bourget A. Explicitation du raisonnement clinique : méthodologie novatrice menant à l'identification de deux étapes de développement au cours des 12 mois entourant la fin de la formation préclinique et le début de la formation clinique d'étudiants en médecine [Explicitation of clinical reasoning: innovative methodology leading to the identification of two developmental stages during the 12 months surrounding the end of preclinical training and the beginning of clinical training of medical students] [Article in French]. Rech Qual 2013;32(2):320-345. [doi: 10.7202/1084633ar]
52. Donnaint, Marchand C, Gagnayre R. Formalisation d'une technique pédagogique favorisant le développement de la pratique réflexive et des compétences émotionnelles chez des étudiants en soins infirmiers [Formalization of a training method to develop reflective practice and emotional skills in nursing students] [Article in French]. Rech Soins Infirm 2015;123(4):66-76. [doi: 10.3917/rsi.123.0066]
53. Metté C, Besançon AM, Gagnayre R, Marchand C. Avantages et limites d'une technique d'aide à l'explicitation dans la formation des auxiliaires de puériculture: étude exploratoire [Advantages and limitations of a technique to aid in explicitation in the training of childcare assistants: an exploratory study] [Article in French]. Pedagog Med 2008 Aug 01;7(2):91-100. [doi: 10.1051/pmed:2006019]
54. Denis J, Winkopp C, Meriaux M. La (méta)morphose du soi du thérapeute: devenir professionnel grâce à l'explicitation dans l'espace de supervision [The (meta)morphosis of the therapist's self: becoming professional through explicitation in the supervision space] [Article in French]. TransForm (Lille) 2022 Oct 12(24):34-46 [FREE Full text]
55. Maurel M. The explicitation interview: examples and applications. J Conscious Stud 2009;16(10-12):58-89 [FREE Full text]
56. Petitmengin C, Navarro V, Quyen ML. Anticipating seizure: pre-reflective experience at the center of neuro-phenomenology. Conscious Cogn 2007 Sep;16(3):746-764. [doi: 10.1016/j.concog.2007.05.006] [Medline: 17590351]
57. Valenzuela-Moguillansky C. Pain and body awareness: an exploration of the bodily experience of persons suffering from fibromyalgia. Constr Found 2013;8(3):339-350 [FREE Full text]
58. Shireen H, Castelli S, Legault M, Dor-Ziderman Y, Milad J, Knäuper B. Phenomenological support for escape theory: a qualitative study using explicitation interviews with emotional eaters. J Eat Disord 2022 Nov 21;10(1):174 [FREE Full text] [doi: 10.1186/s40337-022-00690-y] [Medline: 36411446]
59. Bedin MG, Capelli F, Droz-Mendelzweig M. Explicitation interview: a method for collecting first-person perspectives on living alone in old age. Int J Qual Methods 2019 Jul 30;18:1609406919865840. [doi: 10.1177/1609406919865840]
60. Bremond M, Berthelot A, Plantier L, Breton H, Pautrat M. Micro-phenomenological approach to explore the patient experience during an initial spirometry examination to diagnose COPD in general practice in France. BMJ Open 2021 Jul 02;11(7):e049121 [FREE Full text] [doi: 10.1136/bmjopen-2021-049121] [Medline: 34215612]
61. Jaunay LB, Ivanoff AS, Philippot C, Triby E, Pelaccia T. How patients' inquiry shapes meaning in primary care consultations: a qualitative study using own-point-of-view video-stimulated recall. J Eval Clin Pract 2026 Oct;32(7):e70653. [doi: 10.1111/jep.70653] [Medline: 42900758]
62. Noushad B, Van Gerven PW, de Bruin AB. Twelve tips for applying the think-aloud method to capture cognitive processes. Med Teach 2024 Jul;46(7):892-897. [doi: 10.1080/0142159x.2023.2289847] [Medline: 38071621]
63. Kahlke R, O'Brien BC, Varpio L. Qualitative research interviews for health professions education: AMEE guide no. 185. Med Teach 2026 Apr;48(4):528-541. [doi: 10.1080/0142159X.2025.2536697] [Medline: 40772485]
64. van Braak M, de Groot E, Veen M, Welink L, Giroldi E. Eliciting tacit knowledge: the potential of a reflective approach to video-stimulated interviewing. Perspect Med Educ 2018 Dec;7(6):386-393 [FREE Full text] [doi: 10.1007/s40037-018-0487-9] [Medline: 30446951]

Abbreviations

EI: explicitation interview
HPE: health professions education

XSL•FO
RenderX

Edited by B Lesselroth; submitted 02.Oct.2025; peer-reviewed by CBT Rietmeijer, R Sanz-Ponce, H Majra, D Aranza; comments to author 10.Mar.2026; revised version received 16.Apr.2026; accepted 17.May.2026; published 13.Nov.2026.

Please cite as:
Jaunay LB, Triby E, Pelaccia T
Defining and Exploring Experience in Health Education
JMIR Med Educ 2026;12:e85039
URL: https://mededu.jmir.org/2026/1/e85039
doi:10.2196/85039
PMID:

©Louis-Baptiste Jaunay, Emmanuel Triby, Thierry Pelaccia. Originally published in JMIR Medical Education (https://mededu.jmir.org), 13.Nov.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Education, is properly cited. The complete bibliographic information, a link to the original publication on https://mededu.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Addressing the Engagement Blind Spot: 12-Tip Tutorial to Make Reasoning Audible in Health Professions Education

Inthrani Raja Indran[1*], PhD; Dujeepa D Samarasekera[2*], FAMS, FRCP (Edin)

[1]Department of Pharmacology, Yong Loo Lin School of Medicine, National University of Singapore, MD3, 16 Medical Drive, Singapore, Singapore
[2]Centre for Medical Education, Yong Loo Lin School of Medicine, National University of Singapore, Singapore, Singapore
[*]all authors contributed equally

Corresponding Author:
Inthrani Raja Indran, PhD
Department of Pharmacology, Yong Loo Lin School of Medicine, National University of Singapore, MD3, 16 Medical Drive, Singapore, Singapore

Abstract

Student engagement reforms have increased active learning in higher education. However, visible participation does not always require students to rehearse the spoken, interactive forms of reasoning central to professional practice, such as explaining one's thinking, responding to challenges, building on others' ideas, and working through uncertainty in real time. We describe this gap as an engagement blind spot. We define audible reasoning as the deliberate externalization of disciplinary or professional reasoning through spoken interaction so that its assumptions, evidence, uncertainty, and justification can be examined, challenged, revised, and developed collaboratively. It is particularly relevant in health professions education and also applies to laboratory-based and scientific disciplines where students must communicate their reasoning under conditions of uncertainty, scrutiny, and collaboration. Digital, hybrid, asynchronous, and AI-enabled learning environments sharpen this blind spot. Shared documents, online quizzes, discussion boards, and generative AI tools can support participation and access; however, they can also allow students to engage without speaking or to produce polished answers without making the underlying reasoning visible. In this context, educators need to ask not only whether students are engaged but also what behaviors they are rehearsing and whether their reasoning is available for examination and development. This tutorial offers 12 practical, evidence-informed tips developed iteratively from recurring challenges encountered in teaching, curriculum design, and faculty development and refined against selected literature. The tips are organized around 4 clusters covering the design of activities that require audible reasoning, the elicitation of reasoning in practice, the development of reasoning through listening and response, and the preservation of audible reasoning in digital and AI-enhanced environments. Across these 4 clusters, we suggest small but intentional shifts that help educators move from silent, constructive tasks toward audible, dialogic, and professionally relevant participation. These include designing tasks around uncertainty, creating low-stakes opportunities for spoken reasoning, using digital tools to surface and organize contributions for further discussion, structuring active listening and peer feedback, reviewing group dialogue formatively, and using conversational AI as a rehearsal partner rather than an authority. These strategies require attention to psychological safety, cultural responsiveness, accessibility, equity, privacy, data governance, and ongoing monitoring. Audible reasoning is not synonymous with speaking more, volunteering more frequently, or sounding fluent. Designing for audible reasoning is not simply an engagement strategy. It is a way of helping students practice the communicative and collaborative behaviors needed to explain, defend, question, and revise human judgment in clinical, scientific, and other professional contexts.

(JMIR Med Educ 2026;12:e104702) doi:10.2196/104702

KEYWORDS

student engagement; collaborative learning; clinical reasoning; educational technology; artificial intelligence

Introduction

Over the past two decades, active learning reforms have moved many higher education classrooms away from passive listening and toward more participatory formats, such as in-class polling and online quizzes. Yet a blind spot remains. Many implementations have been designed primarily to deepen individual learning, but they do not consistently require students to rehearse reasoning aloud, although this is central to professional practice. Such rehearsal includes explaining one's thinking, responding to challenge, building on others' ideas, and working through uncertainty in real time. We refer to this practice as audible reasoning.

This argument is not intended to dismiss the many dialogic pedagogies already used in health professions education. Case-based learning, team-based learning, problem-based learning, clinical case discussions, bedside teaching, ward rounds, and simulation all create valuable opportunities for

XSL•FO
RenderX

interaction and shared sense-making [1]. Yet the presence of such pedagogies does not guarantee that spoken reasoning is consistently rehearsed, made audible, or critically interrogated. Learners may still divide labor, defer to more confident peers, rely heavily on written or AI-generated outputs, or move too quickly toward consensus without articulating the assumptions and judgments shaping their conclusions. The issue, then, is not whether dialogue exists at all, but whether educational designs consistently require learners to think aloud in ways that make their reasoning open to challenge, refinement, and feedback.

This distinction carries particular weight in health professions education, where clinical reasoning is a core capability, although its expression differs across professions and settings [1]. The distinction between participation and reasoning has become more consequential for 2 reasons. First, digital, hybrid, and asynchronous formats have reduced the need for students to speak, allowing them to participate without having to make their reasoning audible. Second, generative AI can now produce polished summaries, explanations, and arguments on demand. A completed answer may therefore provide limited evidence of the reasoning through which it was produced. In technology-enhanced professional environments, information may be readily available. However, learners must still interpret its meaning, define the problem, identify assumptions, determine what matters in context, and justify their position when questioned. AI may help generate or compare possible solutions, but students must still learn not to outsource problem definition itself. This matters not only for professional preparation but also for learning itself.

From a sociocultural perspective, language is not merely a vehicle for expressing thought, but a means through which thought is formed [2]. This tutorial builds on established work in dialogic pedagogy and exploratory talk [3], adapting these principles to higher education, particularly health professions education, including assessment and digital and AI-mediated settings. The ICAP framework distinguishes Interactive, Constructive, Active, and Passive modes of engagement [4]. Constructive, active, and passive activities may occur silently and individually, whereas interactive engagement is inherently dialogic. Many classrooms now move students beyond passive learning but still stop short of requiring them to explain, test, and revise their reasoning with others. Likewise, the Making Thinking Visible movement has helped educators externalize student thinking [5]. However, in higher education, this often takes the form of written artifacts such as reflections, essays, or discussion posts. What may still be missing is deliberate design for the spoken, interactive work of making reasoning available for examination and development.

These behaviors must also be understood in a cultural context. Norms around speaking, questioning, and challenging views vary across educational cultures [6]. In hierarchical settings, restraint may reflect respect rather than disengagement. Designing for audible reasoning should not impose a single dominant style of classroom talk. Instead, it should create predictable, psychologically safe, culturally responsive, and accessible opportunities for learners to practice critical inquiry, professional communication, and reasoned judgment across diverse contexts.

We define audible reasoning as the deliberate externalization of disciplinary or professional reasoning through spoken interaction so that its assumptions, evidence, uncertainty, and justification can be examined, challenged, revised, and developed collaboratively. Audible reasoning is not synonymous with speaking more, volunteering more frequently, or sounding fluent. It draws on related traditions such as making thinking visible and exploratory talk but focuses specifically on designing learning environments in which the grounds for professional judgment become available to others for collective scrutiny and development. Although the construct centers on spoken interaction, this tutorial also considers selected text-based and AI-mediated activities where they preserve or rehearse the same dialogic practices. The question for educators is therefore not only, “Are my students engaged?” but also, “What behaviors are my students rehearsing in my classroom?”

This tutorial presents 12 practical, evidence-informed tips organized into 4 clusters that move from establishing the conditions for audible reasoning to eliciting reasoning in practice, building on the reasoning of others, and sustaining these practices in digital and AI-enabled settings. This tutorial aims to help health professions educators recognize where reasoning remains unarticulated or insufficiently examined and to redesign teaching, feedback, and assessment so that professional reasoning can be heard, challenged, and refined. Textbox 1 lists the learning objectives.

Textbox 1. Learning objectives.

1. Explain why visible participation does not necessarily indicate that students are articulating or developing their reasoning.
2. Redesign selected silent or individual learning activities as collaborative tasks that require students to verbalize, examine, and refine their reasoning.
3. Apply strategies that promote psychological safety, reduce anxiety, and support equitable participation in spoken reasoning.
4. Use generative AI for low-stakes conversational practice under educator oversight, rather than as an authoritative source.

Methods

The 12 tips presented in this tutorial were developed iteratively from recurring challenges encountered in the authors’ teaching, curriculum design, and faculty development work. They were refined against the selected literature cited throughout. The clusters follow the sequence of educational design decisions educators make, beginning with designing for audible reasoning, then eliciting audible reasoning in practice, then developing reasoning through listening and response, and finally sustaining audible reasoning in digital and AI-enhanced environments. The tips are evidence-informed rather than the product of a

XSL•FO
RenderX

systematic review or formal consensus process. Table 1 summarizes the 4 guiding clusters and the 12 associated tips.

Table . Overview of the 4 clusters for making reasoning audible.

Cluster	Illustrative shifts
Design for audible reasoning	Establish expectations and design tasks that require learners to explain, question, justify, and revise reasoning.
Eliciting audible reasoning in practice	Use facilitation strategies in teaching and workplace learning to elicit learners' reasoning rather than accepting brief or polished answers.
Develop reasoning through listening and response	Structure opportunities for learners to listen actively, question assumptions, build on others' ideas, and give feedback that advances collective reasoning.
Sustain audible reasoning in digital and AI-enhanced environments	Design digital and AI-mediated activities that preserve dialogue, require uptake of others' contributions, and use technology to support rehearsal and reflection rather than replace reasoning.

Results

Cluster 1: Design for Audible Reasoning

This cluster addresses the first question: Have I designed for audible reasoning by making the expected behaviors explicit and creating tasks that require students to explain, question, justify, and revise ideas with others?

Tip 1: Do My Teaching Strategies Require Students to Reason With One Another?

A useful first step is to determine whether a learning activity requires learners to reason with one another, that is, interactive engagement, rather than constructing responses individually, that is, constructive engagement. Many reforms stall at the constructive level, where students may be working hard writing reflections, building slides, solving equations, or preparing answers, but they do so silently or through a division of labor. These activities may deepen individual learning, but they do not necessarily require students to cocreate, defend, or revise ideas with peers. This aligns with research on dialogic teaching and Thinking Together, which suggests that productive collaboration depends on designing tasks in which talk is necessary rather than merely incidental [3].

While critics correctly note that internal cognitive engagement can be difficult to assess simply by observing silent students [7], interactive engagement is different because it is inherently social and observable. A silent classroom may contain effortful learning, but it is not yet rehearsing audible, collaborative reasoning. Educators can use the ICAP framework [4] to audit session plans by asking, "Must learners co-generate, explain, defend, or revise ideas with others to complete this task, or could they complete it independently or by dividing the work?" Identifying where reasoning remains private is the first step toward designing for audible interaction.

The distinction is not merely theoretical. In a quasi-experimental study, Wiggins et al [8] compared constructive and interactive versions of repeated classroom activities on topics such as protein translation and gene regulation in 2 back-to-back introductory biology sections of approximately 350 students each. The activities involved small-group work and addressed the same content. However, the interactive versions incorporated structured peer teaching and synthesis, requiring learners to exchange and integrate information, whereas the constructive versions did not require substantive interaction. The interactive condition produced a modest but statistically detectable improvement in learning. The authors cautioned that the gain was modest and that studies without repeated measures would likely be underpowered to detect it. The findings do not suggest that every spoken activity is superior to constructive individual work. Rather, they illustrate why the quality and necessity of interaction matter when the intended outcome includes collaborative reasoning.

Barrier and Solution: Educators may feel they lack the time to audit entire curricula or create space for interactive dialogue within already crowded sessions. Start small by reviewing 1 or 2 core sessions per module and identifying an activity that students can currently complete independently or by dividing the work without engaging with one another's reasoning. Rather than adding a new activity, redesign or replace this task so that successful completion requires students to reason collaboratively, such as comparing interpretations, explaining the grounds for a decision, responding to an alternative view, or revising a position after discussion.

Tip 2: Establish Audible Reasoning as an Explicit Course Expectation

Professional communication behaviors are frequently listed as broad graduate attributes, but they often remain less visible in the daily fabric of the classroom. Educators can establish audible reasoning as an explicit expectation from the outset. This means telling students not only that they are expected to participate but also what kind of participation matters. Students should be expected to explain the grounds for a judgment, respond to challenges or alternative perspectives, question ideas respectfully, and revise their positions in light of better evidence or arguments.

These expectations should be communicated clearly using context-specific language, together with a brief explanation of why interactive methods are being used, as understanding the rationale can reduce student resistance to active learning [9]. Educators should make explicit that the emphasis is on explaining the basis for a judgment and engaging with alternative perspectives, rather than simply speaking more or producing polished answers. The form this takes will vary by

XSL•FO
RenderX

context. Students might be asked to critique the methods and limitations of research papers, compare possible management approaches, or justify why they favor one interpretation over another.

What distinguishes audible reasoning from generic participation is that the expectation targets the quality of reasoning rather than airtime. The operative instruction is not simply to contribute but to make the reasoning or grounds for judgment available for others to examine, question, and build on. Students should also understand that this does not require them to speak frequently, respond immediately, or display polished fluency. In hierarchical or culturally reticent settings, designing for audible reasoning may involve predictable prompts, individual preparation, paired rehearsal, and structured roles before students contribute in larger groups.

Where audible reasoning contributes to assessment, educators should publish criteria that focus on the use of evidence, the clarity of justification, responsiveness to other perspectives, the acknowledgment of uncertainty, and the revision of judgment. Speaking frequency, confidence, or verbal fluency should not be used as proxies for *reasoning quality*. Self-reported confidence has been shown to correlate only weakly with objectively assessed competence among health professions students [10]. Students should therefore become familiar with these expectations through repeated low-stakes practice before applying them in summative assessment. This explicitness signals to students that speaking is not merely optional participation but a required course competency (Textbox 2).

Textbox 2. Tip 2—In practice.

At the first session, name the expectation explicitly and explain its rationale. The statement can be included in the syllabus and repeated before early interactive activities so that students encounter it as a consistent course norm rather than a one-off instruction: We are not aiming for perfect or polished answers, or for everyone to speak frequently. What matters is that when you contribute, you make your reasoning audible and explain how you arrived at your judgment, what evidence you are drawing on, where uncertainty remains, and what might change your view. You will often have time to think or rehearse with a partner first. Confidence or verbal fluency is not a substitute for the quality of your reasoning.

Barrier and Solution: Students may view course expectations as mere suggestions if they are not reflected in the design of learning activities or assessments. Address this by linking audible reasoning to relevant professional standards, framing dialogue as essential career preparation, and intentionally building in recurring opportunities to practice these behaviors during class. Even when not graded directly, consistent use of prompts, structured roles, and formative feedback signals that audible reasoning is part of how learning occurs. Where audible reasoning contributes to assessment, provide explicit criteria and opportunities for low-stakes rehearsal in advance.

Tip 3: Design Tasks Anchored in Uncertainty, Not Just Factual Recall

Disciplinary dialogue can feel superficial when students are asked to discuss facts. When the answer is clear, students may default to cumulative talk, in which they agree uncritically with one another, or to disputational talk, in which the emphasis shifts from collective understanding to defending individual positions [11]. To foster more genuine dialogue, educators can design prompts or scenarios that encourage exploratory talk by presenting situations in which facts alone are insufficient to determine the best course of action. For instance, a case in which 2 reasonable management options carry different risks or a dataset that 2 valid methods would interpret differently gives students something they must reason through aloud rather than simply recall. When the pathway to a solution is unclear, students are prompted to reason aloud, examine alternatives, and navigate ambiguity together. In this form of exploratory talk [11], students collaboratively examine, challenge, and constructively build on one another's ideas. In controlled and replication studies, groups taught the ground rules of exploratory talk outperformed untaught control groups on problem-solving tasks, and these gains transferred to individual reasoning tests rather than remaining confined to the group [12].

However, designing for uncertainty requires students to feel safe enough to be uncertain. Educators also need to resist the urge to resolve discomfort too quickly because students need time to think, sit with ambiguity, and test partially formed ideas before a polished answer is supplied [13]. Students often default to silence because they fear being wrong in front of peers, particularly when psychological safety has not been deliberately cultivated. Educators can reduce this fear by reassuring students that their reasoning need not be complete or correct before it is shared and examined.

Educators can also model the provisional reasoning they hope students will adopt. When teachers present only polished, linear solutions, students may infer that competent reasoning is immediate and certain. Thinking aloud through a novel problem, naming assumptions, identifying what further information is needed, demonstrating how to seek that information, or revising a position in response to a counterpoint can normalize uncertainty and support psychological safety [14].

This approach aligns with the modeling component of cognitive apprenticeship as applied to clinical reasoning education, in which educators deliberately externalize otherwise tacit reasoning processes, including points at which their reasoning remains uncertain or incomplete [15]. Educators can make use of both planned and spontaneous opportunities. When faced with a question they cannot answer immediately, they can acknowledge the uncertainty, explain how they would investigate it, and return later to confirm or revise their position. They can also design incomplete cases, competing interpretations, or deliberately ambiguous steps that allow students to hear what information-seeking, self-correction, and revision sound like in their discipline.

As exploring uncertainty takes time, educators should structure activities to allow meaningful exploration while preserving a clear path to closure. A guiding provisional synthesis, supported by appropriate resources or next steps, can help students

XSL•FO
RenderX

distinguish what is well supported, what remains uncertain, and how outstanding questions might be pursued, reducing the risk that misconceptions persist. The aim is not uncertainty for its own sake but to help students make reasoned judgments based on the available evidence while recognizing what remains uncertain.

Barrier and Solution: Students accustomed to clear right-or-wrong answers may initially feel frustrated by ambiguous tasks, and both students and educators may fear losing credibility when they appear uncertain in public. Address this by framing "navigating uncertainty" as the exercise's primary objective before it begins and by treating errors as normal moments in the reasoning process. When a student offers an incomplete or incorrect response, acknowledge the contribution, identify what is useful in it, and model how to refine the idea rather than simply marking it wrong. This helps students see that professional reasoning is not error-free performance but the ability to recognize, correct, and build on emerging understanding.

Cluster 2: Eliciting Reasoning Aloud

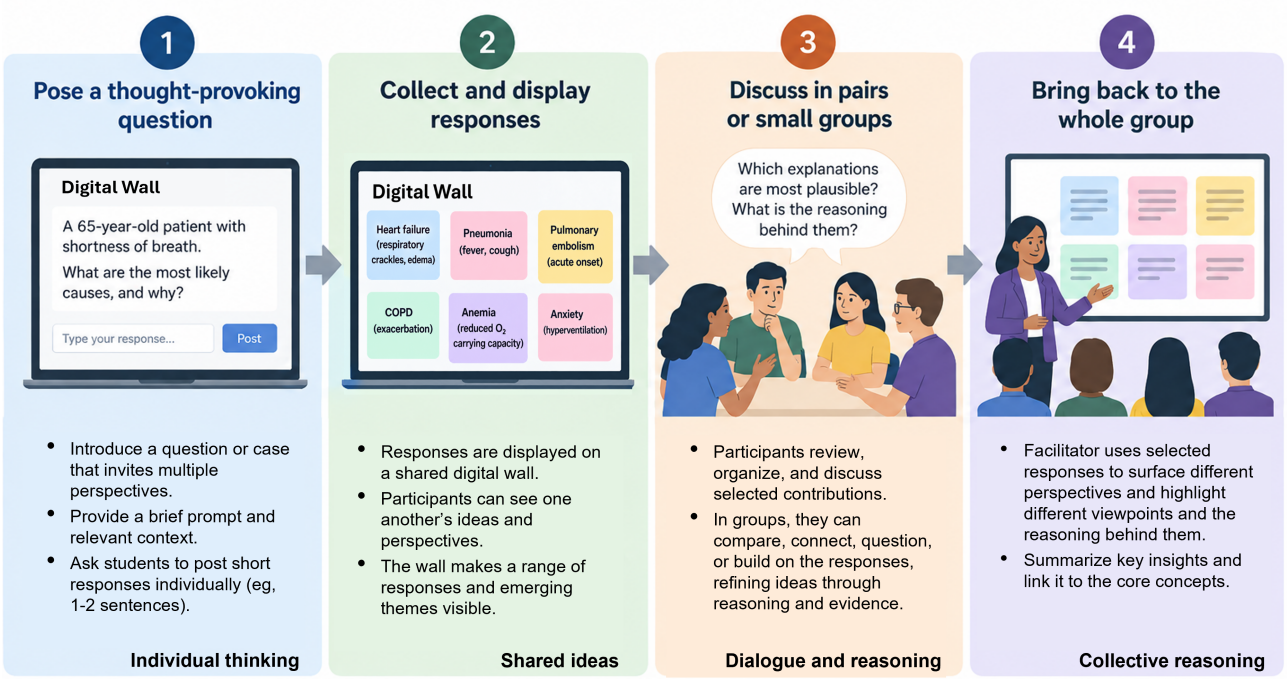
This cluster addresses the second question: Am I eliciting students' reasoning in the classroom by drawing out their thinking rather than simply accepting polished answers?

Tip 4: Step Away From the Podium and Move Into the Room to Elicit Reasoning Across the Class

Large lectures can easily reinforce a content delivery dynamic in which the educator speaks from the front, and students respond only when directly invited. To make thinking audible at scale, educators can shift from delivering content to facilitating reasoning by redistributing attention away from the educator and across the class. Where the physical setting permits, this may begin by stepping away from the podium and moving into the room. When a student answers a question, resist the urge to validate it immediately. Instead, invite responses from elsewhere in the room: "Who interprets this data differently?" or "How does your proposed mechanism build on what your classmate just said?" The aim is to make one student's answer available for others to examine, extend, or challenge, rather than allowing it to return immediately to the educator for evaluation. In this way, the educator's movement helps redistribute attention across the room and prompts students to listen actively and reason collaboratively, even in a large lecture hall.

Digital tools can complement this process by making students' contributions available for comparison, organization, and further discussion, rather than merely collecting answers. One example is Talkwall [16], a microblogging tool through which students submit brief contributions individually or in groups to a shared feed. Selected contributions can then be moved to a shared wall and spatially organized, helping recurring themes, contrasting perspectives, and relationships among ideas become visible. In teaching health professions, a comparable digital wall might be used to organize possible diagnoses, differentials, competing interpretations of a dataset, or arguments for different management options. Figure 1 illustrates one way a shared digital wall can surface different responses to a question and then return them to spoken, collaborative examination.

Figure 1. Using digital walls to seed spoken dialogue. Digital walls can make students' initial ideas visible and provide a shared starting point for discussion when educators structure how students engage with and build on one another's contributions. COPD: chronic obstructive pulmonary disease; O_2: oxygen.



Live polling followed by paired explanations and repolling can similarly show whether students reconsider their positions after discussion. At the same time, shared annotation tools can support the comparison of interpretations of a clinical image, research figure, or case. In each instance, digital contributions should seed and structure spoken dialogue rather than replace it [3].

XSL•FO
RenderX

The same principle applies to synchronous online teaching, where a videoconference space replaces the physical room. Here, an educator cannot move through the room but can achieve a similar redistribution of attention by naming students to extend or challenge a peer's point, routing a contribution to a breakout group for discussion before it returns to the whole class, or using the chat as a staging space from which spoken responses are then invited. The aim is unchanged. One student's reasoning should remain open for others to examine rather than be closed by the educator's evaluation.

Barrier and Solution: Students often find it intimidating to speak across a large lecture hall. This can be mitigated using a brief think-pair-share activity [17], allowing students to rehearse their ideas with a peer before sharing them with the wider room. When movement around the room is constrained by room design, class size, or accessibility considerations, educators can use roaming microphones, wireless slide clickers that preserve freedom of movement, student facilitators, paired spokespersons, or selected digital responses as prompts for discussion. The pairing step is what makes this approach workable at scale: experimental work shows that frequent, warmly framed invitations to speak can increase the number of students who participate voluntarily, with participation growing over the course of a semester [18]. It is worth being candid, however, that the "share" phase is not risk-free. Some students may still feel anxious when asked to speak in front of the class, especially if they have not volunteered. Think-pair-share should therefore be treated as a scaffold that lowers the threshold for participation, not a guarantee to remove anxiety. We return to this issue later when considering the adverse effects of audible reasoning.

Tip 5: Transform Practical and Laboratory Procedures Into Troubleshooting Dialogues

Practical and laboratory sessions are often cognitively demanding, leading students to follow protocols procedurally rather than reason aloud. Collaborative troubleshooting is frequently absent, although recent work shows that structured dialogue in laboratories can enhance conceptual understanding and engagement [19]. Introduce structured "stop-and-talk" checkpoints at critical junctures or at the end of the class to review the rationale for the protocol steps or investigate anomalous data. For instance, before leaving the class, require students to discuss the session's core rationale with their peers and verbally demonstrate their understanding to the facilitator. If a group cannot reach consensus, they can be asked to articulate their differing interpretations openly so that uncertainty becomes a resource for discussion rather than something to hide. This shifts the laboratory from a space of procedural completion to one in which students rehearse real-time reasoning, troubleshooting, and collaborative problem-solving (Textbox 3).

Textbox 3. Tip 5—In practice.

Insert 1 "stop-and-talk" checkpoint before students leave. Ask each pair or group to answer 3 questions aloud to the facilitator:

"What are you doing at this step, why are you doing it, and how would you interpret a result that differs from what you expected?"

If students in the group disagree about the interpretation, ask them to articulate both explanations and identify what additional information, observation, or test would help distinguish between them.

Used this way, the checkpoint can replace a redundant procedural step rather than simply adding another activity to the session, while allowing the facilitator to probe understanding that may not be apparent from a written report.

Barrier and Solution: Faculty may find it difficult to add discussion checkpoints to already tight, time-pressured laboratory schedules. One way to address this is to replace a redundant procedural step with a single high-impact verbal checkpoint, so that dialogue replaces rather than adds to students' cognitive load. In its simplest form, this checkpoint functions as a brief, low-stakes teaching strategy: students explain what they are doing, why they are doing it, and how they would interpret an unexpected result.

Where appropriate, such checkpoints could also be developed into a partial alternative to the traditional written laboratory report. In this more formal assessment role, the oral format offers an important advantage: the examiner can probe students' responses, which may reveal dimensions of understanding that a static written report does not capture. Oral assessment does, however, carry known risks of reliability and bias, particularly under high-stakes or poorly structured conditions [20]. If verbal checkpoints are used for grading, they should be structured with common prompts and explicit criteria and include appropriate examiner calibration and student accommodations to ensure fairness and interexaminer consistency.

Tip 6: Move Students From Scripted Delivery Toward Conversational Explanation

When students deliver presentations, they may rely heavily on scripted notes or read directly from text-heavy slides, stifling spontaneous, dialogic reasoning. To develop the habits of an agile professional, educators should help presenters move from memorized scripts toward explaining ideas conversationally, requiring them to internalize core concepts and to use cue cards strictly as brief prompts. To ensure students prioritize this practice over mere content delivery, educators should include clarity and coherence of explanation, responsiveness to questions, adherence to time limits, appropriate pacing, audience engagement, and equitable sharing of questions and answers in the grading rubric [21]. These communication behaviors should be assessed separately from the quality and justification of the underlying reasoning, as outlined in tip 2. Sharing this rationale up front helps students understand that time management and collective team defense are core professional skills, preventing a dynamic in which one confident student dominates the dialogue.

Expecting the team to pause and field questions organically transforms a static delivery into a genuinely interactive defense.

XSL•FO
RenderX

This kind of spoken presentation is not merely a classroom exercise but a recognized vehicle for clinical learning.

In a recent mixed methods study, medical students described preparing and delivering oral case presentations at academic conferences as strengthening their later clinical training and professional development, which suggests that practicing conversational, defensible explanation can carry value well beyond the immediate task (Textbox 4) [22].

Textbox 4. Tip 6—In practice.

Treat the presentation as a conversation, not a script. Pause at a key decision point and probe the reasoning directly:

"You've told me the plan. Now, tell me why. What in the patient's history or data pushed you towards this and away from the next most likely alternative? What are you still uncertain about, and what finding or piece of information would change your mind?"

Clinical reasoning frameworks can provide useful scaffolds by helping students organize information and develop a systematic approach to case presentation. As students progress, these frameworks may become more integrated into how they synthesize the case, rather than being followed as a rigid sequence. At the same time, educators should continue to probe the reasoning underlying key decisions so that students develop both an organized approach to presenting the case and the ability to explain and defend their clinical judgment.

Barrier and Solution: Removing scripts can significantly increase presenter anxiety among some students, particularly for nonnative speakers of the language of instruction. Mitigate this by affording reasonable flexibility where needed, such as allowing slightly more detailed prompts or additional structure, while maintaining the broader goal of developing more conversational explanation. Scaffold the experience throughout the course. Early on, students may benefit from structured formats, cue cards, or informal seated group updates to help organize their presentation, but these scaffolds should not postpone conversational probing of their reasoning. From the outset, educators can ask students to explain why they favored a particular interpretation, what evidence supports it, what uncertainty remains, or what information would change their position. The nature and degree of scaffolding can be tailored to the individual student. As students progress, these supports can become more flexibly incorporated into how they synthesize and present a case. The key question is what behavior the student is rehearsing. Rather than repeatedly practicing scripted delivery, students can be gradually supported to explain, adapt, and respond in real time, behaviors they will need in professional settings.

Cluster 3: Building on the Reasoning of Others

This cluster addresses the third question: Are students building on one another, listening and responding rather than delivering parallel monologues?

Tip 7: Cultivate Active Listening That Seeks to Understand, Not Just Respond

Student presentations can easily leave the audience passive, with students tuning out until it is their own turn to speak. Educators can rebalance this dynamic by treating the audience as active participants in a professional exchange rather than as observers. As established in cluster 1, audible reasoning needs to be explicitly expected. This tip extends that expectation to the listener: reasoning only becomes dialogic when peers listen closely enough to understand, question, and build on what has been said.

Active listening is too often assumed to be innate or passive, when it is better understood as an active, teachable professional practice [23]. It draws on the same capacities whether a clinician is listening to a patient, a trainee to a senior, or a student to a peer presenting. In medicine, listening supports diagnosis and the therapeutic relationship [24]. The classroom is therefore an important rehearsal space for a competency that professional practice constantly demands.

Begin by making attentive listening visible. Establish explicit rules of engagement for peer presentations, including specific periods when students close their laptops and silence their devices to direct their attention toward the speaker. Students who require devices for accessibility should continue to be accommodated. This signals that giving a colleague undivided attention is professional behavior, not merely a courtesy. However, attention alone is not enough. The aim is to listen to understand before replying. This means following a peer's reasoning, noticing the inferences one is already forming, and holding one's assumptions in check rather than projecting one's own meaning onto what was said. Only then can a listener respond with curiosity rather than waiting to rebut, correct, or display knowledge. Active listening can also be designed into the task itself by making each student's contribution dependent on understanding what has already been said.

This matters because building on another person's reasoning depends on first understanding it. A question asked to clarify, probe assumptions, offer another perspective, or identify a gap can extend the presenter's thinking. In contrast, a question asked only to challenge or showcase knowledge may close dialogue. Inattentive listening is not merely a missed opportunity; it can derail discussion or lead the group away from the reasoning that needs attention. Educators can make this expectation concrete by assigning rotating listening roles, such as clarifier, evidence seeker, assumption checker, or connector. Each role directs attention to a different aspect of the presenter's reasoning and gives the student a defined contribution to make after the presentation. Alternatively, each audience member could prepare 1 question that clarifies meaning, 1 comment that connects the talk to an earlier idea, and 1 point that respectfully tests an assumption. These moves align with the response practices through which listening becomes visible in group talk, including revoicing, integrating, interpreting, and critically extending what others have said [25]. In doing so, the audience shifts from passive listeners to active participants who practice the professional behaviors of attending, interpreting, questioning, and building on others' reasoning (Textbox 5).

XSL•FO
RenderX

Textbox 5. Tip 7—In practice.

In a preclinical class, the educator can portray a patient with a predetermined history and initially present with only a single symptom, such as headache. Further information is revealed only in response to appropriate questions. Students take turns assuming the physician's role and ask 1 history-taking question at a time as a microphone is passed around the room. Each student must listen to what has already been asked and elicited, avoid unnecessary repetition, and decide what still needs to be explored. At the end, another student summarizes the history, identifies salient positive and negative findings, and notes any important gaps. The activity makes listening observable through the relevance of each successive question and the accuracy and completeness of the final synthesis.

Barrier and Solution: Students may request access to their laptops to take notes or access course materials. In interactive sessions where students are expected to listen and respond, explain that limiting device use during these periods supports real-time listening, understanding, and engagement rather than simply receiving information. Provide paper for brief notes if needed or ask presenters to circulate a 1-page summary before the session so that the audience can focus less on transcribing content and more on the reasoning presented. Make reasonable exceptions when devices are required for accessibility or other learning needs.

Tip 8: Break the "Parallel Monologue" in Digital Forums

Asynchronous discussion forums frequently devolve into parallel monologues, with students posting isolated thoughts to meet a quota rather than engaging in genuine exchange [26]. Just as active listening in the room turns an audience into participants, the same disposition is needed online: reading peers' contributions to genuinely understand them, weighing those views with curiosity and openness, and offering one's own in response rather than in isolation. Sustained critical inquiry in text-based forums depends on precisely this kind of engagement, in which participants connect and build on one another's contributions rather than posting in parallel [27]. To foster this, educators might reconsider the common practice of asking students to post original responses to a static prompt.

Instead, forums can be designed around synthesis and critical review, requiring students not only to articulate their own perspectives but also to engage directly with the substance of their peers' contributions, creating a more interactive and evolving conversation. Educators can also bring selected forum posts into class, using contrasting interpretations or unresolved questions to prompt discussion of assumptions, evidence, and possible revisions. Discussion forums are a setting in which the engagement blind spot is especially apparent because text-based tools make it structurally easy to contribute without reasoning with anyone. Therefore, the same design intent that animates the spoken classroom, making thinking available for others to examine, question, and build on, must be carried deliberately into asynchronous space rather than assumed to follow from participation alone (Textbox 6).

Textbox 6. Tip 8—In practice.

Replace a generic "post and reply" instruction with one that requires students to engage directly with a peer's reasoning:

"Share your response to the case questions. Then choose 1 peer's contribution and refer to a specific point they have made. Explain 1 point you agree with and why, identify 1 claim you find uncertain or under-supported, and pose 1 question that could move the reasoning forward."

Where feasible, bring 2 contrasting discussion threads into the next live session and ask the class to examine where the students' reasoning diverged and why.

Barrier and Solution: In discussion forums, when asked to respond to peers, students may default to a brief, polite agreement before launching into a new, unrelated monologue. Overcome this by providing explicit instructional guidance on how to engage. Prompt them to identify which elements of their peer's post they agree with, which they find uncertain or challenging, and which require clarification or evidence, and to explain why. Furthermore, explicitly reiterate the ground rules for respectful, nonjudgmental exchange, reminding students that their peers' perspectives may be shaped by diverse life, educational, and clinical experiences.

Tip 9: Scaffold Peer-to-Peer Feedback Using Structured Templates and Feedback Prompts

Effective peer feedback is a professional skill that requires students to identify strengths, recognize areas for improvement, and communicate both respectfully and constructively. In the context of audible reasoning, it also gives students repeated practice in listening to another person's contribution, interpreting its value, and offering a response that helps the person improve. It can be enhanced through structured scaffolds [28]. Yet many students struggle to do this well because they may lack the vocabulary, confidence, or a sense of psychological safety to express their judgments clearly. Peer feedback can be scaffolded in the earlier years of higher education, particularly through group projects, using structured templates, cue words, sentence stems, and guided prompts that help students focus on specific examples, explain the impact of behavior, and suggest practical next steps. The aim is to move feedback away from vague praise or vague criticism and toward a disciplined form of professional dialogue in which students identify a specific behavior, explain its impact, and suggest how it might be strengthened.

The process can be further strengthened when recipients of the feedback evaluate its quality using a simple matrix that considers whether it was constructive, respectful in tone, and clear and actionable in content. This makes feedback a 2-way learning process in which students practice giving useful feedback while also learning to assess whether the feedback they receive is specific enough to support improvement. In a group setting, each student evaluates the feedback directed to them, followed by a brief group debrief on common strengths and areas for improvement. Faculty may sample the matrices to guide the

XSL•FO
RenderX

process rather than grade every exchange. As students progress through the years, these scaffolds can be gradually faded as constructive peer dialogue becomes more fluent and internalized (Textbox 7).

Textbox 7. Tip 9—In practice.

Provide students with a format that separates the strength, the issue, and the actionable next step, and require 1 of each:

"One strength you demonstrated was inviting quieter team members into the discussion [specific behavior], which helped the group hear a wider range of perspectives [positive outcome/impact]."

"One area to strengthen is explaining your reasoning more explicitly when you disagree [behavior to improve], as this would make your contribution clearer and easier for others to engage with [desired outcome/impact]."

"A helpful next step might be to state your rationale before offering an alternative view [specific suggestion]."

Then ask the recipient to rate the feedback they received on 3 dimensions. Was it constructive, respectful in tone, and specific enough to act on? Close with a brief group debrief on common patterns. This allows students to rehearse both giving useful feedback and judging the quality of feedback they receive, rather than simply exchanging vague praise or criticism.

Barrier and Solution: Students may avoid honest critique for fear of harming relationships, leading to vague or overly softened feedback. This can be addressed by structuring both the giving and receiving of feedback. Requiring 1 strength and 1 area for improvement encourages balanced commentary. In this way, peer feedback becomes a safe and structured practice for developing professional judgment, communication, and reflective capacity.

Cluster 4: Sustaining Audible Reasoning in an AI-Enhanced Era

This cluster addresses the fourth question: Does the practice hold up in technology- and AI-enabled settings, where reasoning most easily falls silent? Tip 10 focuses on preserving spoken reasoning in technology-mediated participation, tip 11 focuses on reviewing dialogue after it has occurred, and tip 12 focuses on using conversational AI to rehearse future interactions.

Tip 10: Set Ground Rules for Spoken Participation in Technology-Mediated Spaces

Technologies that lower the barrier to participation, such as shared documents, discussion boards, and typed chat, can support access and broaden participation. However, they can also make it structurally easy for students to engage without ever speaking. Silent typing during in-person group work or prolonged silence in online sessions is not simply a sign of poor motivation; it is often a predictable outcome of how participation has been designed. Educators can respond by establishing clear structural expectations for spoken interaction. Even where digital tools are used to surface and organize contributions, as discussed in tip 4, educators should decide which stages of the activity require students to negotiate, justify, or synthesize ideas aloud.

For example, in in-person group work, the first 5 minutes of a collaborative task might be designated as entirely "tech-free," allowing only brief handwritten notes, with appropriate accommodation for students who require devices for accessibility. Students must first brainstorm, negotiate, and justify ideas aloud before shifting to shared documents or AI tools. Once an initial position has been articulated, these technologies can be used to search for evidence, compare alternatives, or develop the group's synthesis.

In online settings, educators can assign explicit interaction roles to ensure that spoken contributions are distributed rather than left to a few confident voices. The point is not to reject technology but to decide deliberately when it supports reasoning and when it quietly replaces spoken interaction (Textbox 8).

Textbox 8. Tip 10—In practice.

Sequence the use of technology into the lesson structure:

Divide a collaborative task into spoken and digital phases and make the sequence explicit to students:

"For the first five minutes, discuss the problem without using digital or AI tools. Brainstorm, compare your views, and explain your reasoning aloud. Then use the shared document or AI tools to search for evidence and compare alternatives. Finally, return to your discussion and explain what your group retained, changed, or rejected, and why."

Where students require devices for accessibility or other learning needs, appropriate accommodation should be maintained. The purpose is not to exclude technology, but to ensure that it supports the group's reasoning rather than replacing the opportunity to reason together.

Barrier and Solution: Students may cite screen fatigue, privacy concerns, or discomfort with being visibly present online. This can be addressed by using a professional compromise where we allow flexibility during lower-interaction portions of the session, while clearly identifying selected moments where verbal and, where feasible, visible participation is expected. This respects student well-being while still rehearsing the norms of contemporary professional communication.

Tip 11: Use Recordings and AI-Assisted Analysis to Reflect on Participation and Dialogue

Once reasoning has been made audible, students can benefit from looking back at how their group dialogue unfolded. Short recordings or transcripts of group discussions can be used as reflective artifacts. With appropriate consent and data safeguards, educators may ask students, or an institutionally approved AI tool, to review a discussion for whether the group demonstrated patterns of collaborative reasoning, such as

XSL•FO
RenderX

inviting contributions, clarifying or revoicing ideas, integrating different viewpoints, extending the discussion through questions, and revising positions in response to others [29]. The review can also consider whether these practices were distributed sufficiently to support collective reasoning. These prompts draw on response practices that make listening visible in group talk, from acknowledging and revoicing to integrating and critically extending what others have said [25].

Reviewing group interaction in this way connects directly to the active listening in cluster 3 and lets students see whether they are building on one another's reasoning or merely taking turns speaking. For example, a group might review a brief excerpt of their interaction transcript to identify 1 instance in which members developed an idea collaboratively, note 1 missed opportunity to clarify or integrate a contribution, and agree on 1 practice to strengthen in its next discussion.

The purpose should be formative rather than punitive. AI may help surface observable patterns in listening and response, but these outputs should be treated as prompts for reflection rather than definitive judgments of competence. Educators should avoid reducing participation to simplistic metrics such as the number of turns, speaking time, or frequency of keywords because these may reward confidence or verbosity rather than meaningful contribution. AI-generated interpretations may also overlook context, tone, nonverbal communication, or the disciplinary significance of a contribution. They should therefore be checked against the recording or transcript and interpreted alongside group self-reflection, peer feedback, and educator judgment.

Barrier and Solution: Recording discussions raises legitimate concerns about privacy, consent, and the possibility that students will perform for the recording rather than participate naturally. Address this by clearly explaining the educational purpose, obtaining appropriate consent, minimizing identifiable data, using only institutionally approved tools that comply with appropriate data safeguards, keeping the focus formative, and deleting recordings and transcripts once their educational purpose is served. Do not upload identifiable material to publicly available AI tools. Framing the review as a shared, low-stakes reflection on how the group reasons together, rather than as surveillance of individual performance, may help preserve the psychological safety on which honest participation depends.

Tip 12: Use Conversational AI to Expand Opportunities for Rehearsing Audible Reasoning

Generative AI is often discussed as a tool for producing notes, summaries, and written responses, but one promising educational value may lie in creating additional opportunities for spoken rehearsal. In health professions education, and in adjacent scientific or laboratory-based disciplines, students must learn not only to arrive at a plausible answer but also to explain their reasoning clearly, respond to challenges, defend their decisions, and adapt their positions in real time. Conversational AI tools can support this by serving as low-stakes rehearsal partners beyond the limited hours of classroom time. While the previous tip used AI to review reasoning that had already taken place, this tip looks ahead, allowing students to rehearse reasoning before they have to perform it in class, clinical, or professional settings.

Educators may design simple, class-specific conversational AI activities that allow students to rehearse the forms of spoken reasoning and professional dialogue required in their discipline. Students might use these tools to simulate a presentation audience, respond to follow-up questions, defend weak points in an argument, or receive provisional formative feedback on clarity, structure, pacing, coherence, and communicative style. AI-generated questions and feedback may nevertheless be inaccurate, biased, overly confident, or poorly aligned with disciplinary expectations. Students should therefore be advised to verify feedback against appropriate evidence and professional standards and treat it as a prompt for reflection rather than an assessment of competence. In clinical or scientific contexts, such tools might also simulate skeptical peers, patients, examiners, or interdisciplinary colleagues. Used in this way, conversational AI does not replace authentic human dialogue but expands access to structured rehearsal for the audible, interactive skills that professional practice demands.

Early evidence is encouraging but should be read with care. In a mixed methods study, medical students at 3 universities used a conversational AI system as a simulated patient to rehearse demanding encounters, such as breaking bad news and counseling anxious patients. They reported gains in confidence and perceived usefulness [30]. In a study of nurse practitioner students, an AI-driven virtual simulation was used to practice oral case presentations and receive structured formative feedback on how well students communicated their clinical reasoning and demonstrated clinical judgment [31]. In language education, where the evidence base is larger, a controlled mixed methods study found that AI conversation partners can improve speaking performance while reducing speaking anxiety [32]. The important caveat is that most of these outcomes are self-reported confidence or second-language proficiency rather than demonstrated gains in clinical or scientific reasoning. As noted in tip 2, confidence that outpaces competence is itself a recognized hazard. Conversational AI is therefore best positioned as a rehearsal space that lowers the cost of practice, with judgments of competence reserved for human assessment.

Barrier and Solution: Students may overrely on AI-generated feedback or treat these exchanges as scripts to memorize rather than opportunities for genuine reasoning. This can be addressed by positioning AI as a rehearsal partner rather than an authority. Ask students to reflect on which suggestions they accepted, rejected, or modified, and why, and to identify feedback that was inaccurate, unsupported, biased, or inconsistent with disciplinary expectations. AI-assisted rehearsal should remain formative and complement, rather than displace, practice with peers, educators, patients, and other professionals. Final judgment should remain grounded in educator feedback, peer discussion, disciplinary expectations, and observed performance with real people. Where exchanges involve identifiable, sensitive, or clinical information, the privacy and data governance safeguards outlined in tip 11, tip 12, and the "Discussion" section should apply.

XSL•FO
RenderX

Discussion

Compelling students to externalize reasoning carries risks of its own. The most important of these is the psychological safety on which audible reasoning depends. Requiring students to speak when they have not volunteered can raise rather than lower anxiety, driven by a fear of negative evaluation, and this burden is not evenly distributed [33]. Students from underrepresented groups and nonnative speakers of the language of instruction may be disproportionately affected when fluency under pressure becomes the medium of participation. A poorly designed push toward audible reasoning, therefore, risks amplifying inequities and conflating verbal fluency with the quality of reasoning. These risks are mitigable rather than disqualifying.

The safeguards described throughout this tutorial can reduce these risks. These include giving students time to rehearse in pairs before whole-class sharing, using warm, predictable invitations rather than "gotcha" questions, and helping them reformulate a half-formed answer before sharing it with the wider class. When active learning is implemented with attention to inclusion, it can narrow achievement gaps rather than widen them [34]. The design choices advocated throughout this tutorial are therefore not optional refinements. Ground rules, psychological safety, low-stakes rehearsal, and structured participation are the conditions that make audible reasoning safe and equitable.

Where strategies involve recording discussion or routing student reasoning through AI tools, the practical safeguards set out in tips 11 and 12 should be observed. Such recordings and transcripts may contain sensitive student information and should be managed in accordance with relevant institutional policies and legal requirements. A particular concern in any setting is whether a third-party tool can access, store, or reuse student data. Educators should therefore use institutionally approved tools where possible, minimize the amount of identifiable data collected, keep the purpose formative, and delete recordings or transcripts once their educational purpose has been served.

Educators can strengthen implementation by monitoring how these strategies work in their own contexts, including unintended effects such as heightened anxiety or inequitable participation, as well as the quality of reasoning and patterns of participation. They should also examine whether students explain the grounds for their judgments, acknowledge uncertainty, respond to alternative interpretations, revise their positions, and invite or build on others' contributions. The aim is not simply to make students speak more but to create conditions in which more students can practice reasoning aloud, listen to one another, and build judgment together.

Taken together, the 4 clusters offer educators a compact self-check they can apply to almost any teaching encounter: have I set the conditions for reasoning aloud, am I eliciting that reasoning in the room, are students building on one another's contributions, and does the practice hold up in technology- and AI-mediated settings? Used deliberately, these moves bridge the gap between individual classroom participation and the collaborative competence that real-world clinical and scientific practice demands. Verbal communication should not be treated as a generic soft skill appended to disciplinary learning but as a fundamental mechanism through which complex thought is formed, tested, and refined. This is what the era of generative AI clarifies rather than diminishes. As machines increasingly assist with producing answers, the distinctive work of education becomes teaching students to voice, defend, and revise human judgment in dialogue with others. Designing for audible reasoning is therefore not simply an engagement strategy, but a way of preparing graduates to participate in the collaborative reasoning that clinical, scientific, and other professional contexts require.

Acknowledgments

The authors declare the use of generative AI (GenAI) during manuscript preparation and revision. According to the GAIDeT (Generative AI Delegation Taxonomy), the following tasks were delegated to GAI tools under full human supervision: proofreading, editing, and reformatting. The GAI tool used was ChatGPT (OpenAI; GPT-5.5). Responsibility for the final manuscript lies entirely with the authors. All AI-assisted outputs were critically reviewed, revised, and verified by the authors. GAI tools are not listed as authors and do not bear responsibility for the final outcomes. The authors gratefully acknowledge the Department of Pharmacology, Yong Loo Lin School of Medicine, National University of Singapore, for providing funding support for the publication of this work.

Funding

The authors gratefully acknowledge the Department of Pharmacology, Yong Loo Lin School ofMedicine, National University of Singapore, for providing funding support for the publication ofthis work.

Conflicts of Interest

None declared.

References

1. Elvén M, Welin E, Wiegleb Edström D, et al. Clinical reasoning curricula in health professions education: a scoping review. J Med Educ Curric Dev 2023;10:23821205231209093. [doi: 10.1177/23821205231209093] [Medline: 37900617]

XSL•FO
RenderX

2. Vygotsky LS. In: Cole M, John-Steiner V, Scribner S, Souberman E, editors. Mind in Society: The Development of Higher Psychological Processes: Harvard University Press; 1978. [doi: 10.2307/j.ctvjf9vz4]
3. Mercer N, Hennessy S, Warwick P. Dialogue, thinking together and digital technology in the classroom: some educational implications of a continuing line of inquiry. Int J Educ Res 2019;97:187-199. [doi: 10.1016/j.ijer.2017.08.007]
4. Chi MTH, Wylie R. The ICAP framework: linking cognitive engagement to active learning outcomes. Educ Psychol 2014 Oct 2;49(4):219-243. [doi: 10.1080/00461520.2014.965823]
5. Ritchhart R, Church M. The Power of Making Thinking Visible: Practices to Engage and Empower All Learners: Jossey-Bass; 2020.
6. Frambach JM, Driessen EW, Beh P, van der Vleuten CPM. Quiet or questioning? Students' discussion behaviors in student-centered education across cultures. Studies in Higher Education 2014 Jul 3;39(6):1001-1021. [doi: 10.1080/03075079.2012.754865]
7. Thurn CM, Edelsbrunner PA, Berkowitz M, Deiglmayr A, Schalk L. Questioning central assumptions of the ICAP framework. NPJ Sci Learn 2023 Nov 15;8(1):49. [doi: 10.1038/s41539-023-00197-4] [Medline: 37968570]
8. Wiggins BL, Eddy SL, Grunspan DZ, Crowe AJ. The ICAP active learning framework predicts the learning gains observed in intensely active classroom experiences. AERA Open 2017 May 19;3(2). [doi: 10.1177/2332858417708567]
9. Andrews M, Prince M, Finelli C, Graham M, Borrego M, Husman J. Explanation and facilitation strategies reduce student resistance to active learning. Coll Teach 2022 Oct 2;70(4):530-540. [doi: 10.1080/87567555.2021.1987183]
10. Lai NM, Teng CL. Self-perceived competence correlates poorly with objectively measured competence in evidence based medicine among medical students. BMC Med Educ 2011 May 28;11:25. [doi: 10.1186/1472-6920-11-25] [Medline: 21619672]
11. Mercer N. Sociocultural discourse analysis: analysing classroom talk as a social mode of thinking. Journal of Applied Linguistics 2004 May;1(2):137-168. [doi: 10.1558/japl.2004.1.2.137]
12. Wegerif R, Mercer N, Dawes L. From social interaction to individual reasoning: an empirical investigation of a possible socio-cultural model of cognitive development. Learn Instr 1999 Dec;9(6):493-516. [doi: 10.1016/S0959-4752(99)00013-4]
13. Kumagai AK. Discomfort, doubt, and the edge of learning. Acad Med 2022 May 1;97(5):649-654. [doi: 10.1097/ACM.0000000000004588] [Medline: 35020611]
14. Dong C, Altshuler L, Ban N, et al. Psychological safety in health professions education: insights and strategies from a global community of practice. Front Med (Lausanne) 2025 Jan 24;11:1508992-January. [doi: 10.3389/fmed.2024.1508992] [Medline: 39927270]
15. Jagannath AD, Dreicer JJ, Penner JC, Dhaliwal G. The cognitive apprenticeship: advancing reasoning education by thinking aloud. Diagnosis (Berl) 2023 Feb 1;10(1):9-12. [doi: 10.1515/dx-2022-0043] [Medline: 36450097]
16. Smørdal O, Rasmussen I, Major L. Supporting classroom dialogue through developing the Talkwall microblogging tool: considering emerging concepts that bridge theory, practice, and design. NJDL 2021 Aug 16;16(2):50-64. [doi: 10.18261/issn.1891-943x-2021-02-02]
17. Cooper KM, Schinske JN, Tanner KD. Reconsidering the share of a Think-Pair-Share: emerging limitations, alternatives, and opportunities for research. CBE Life Sci Educ 2021 Mar;20(1):fe1. [doi: 10.1187/cbe.20-08-0200] [Medline: 33444105]
18. Dallimore EJ, Hertenstein JH, Platt MB. Impact of cold-calling on student voluntary participation. J Manag Educ 2013 Jun;37(3):305-341. [doi: 10.1177/1052562912446067]
19. Seery MK, Agustian HY, Christiansen FV, Gammelgaard B, Malm RH. Incorporating dialogue in laboratory teaching. Anal Bioanal Chem 2024 Nov;416(28):6165-6170. [doi: 10.1007/s00216-024-05569-2] [Medline: 39352472]
20. Nallaya S, Gentili S, Weeks S, Baldock K. The validity, reliability, academic integrity and integration of oral assessments in higher education: a systematic review. Issues Educ Res 2024;34(2):629-646 [FREE Full text]
21. Pérez-Higueras JJ, Hidalgo Arroquia JJ, Gancedo-Caravia L. Rubric for peer evaluation of oral presentations: use and perceptions among experienced and non-experienced students. J Dent Educ 2025 Aug;89(8):1211-1220. [doi: 10.1002/jdd.13831] [Medline: 39825488]
22. Kasai H, Imafuku R, Shikino K, et al. Learning through oral case presentations and impact on medical students' clinical training and career development: a mixed-method study. Clin Teach 2025 Dec;22(6):e70242. [doi: 10.1111/tct.70242] [Medline: 41146431]
23. Fortune D. It's time: evidence-based strategies to improve student listening in the college classroom. Coll Teach 2025:1-5. [doi: 10.1080/87567555.2025.2488841]
24. Boudreau JD, Cassell E, Fuks A. Preparing medical students to become attentive listeners. Med Teach 2009 Jan;31(1):22-29. [doi: 10.1080/01421590802350776] [Medline: 19140065]
25. Edwards-Groves C, Davidson C. Noticing the multidimensionality of active listening in a dialogic classroom. AJLL 2020 Feb;43(1):83-94. [doi: 10.1007/BF03652045]
26. Galikyan I, Admiraal W. Students' engagement in asynchronous online discussion: the relationship between cognitive presence, learner prominence, and academic performance. Internet High Educ 2019 Oct;43:100692. [doi: 10.1016/j.iheduc.2019.100692]
27. Garrison DR, Anderson T, Archer W. Critical inquiry in a text-based environment: computer conferencing in higher education. Internet High Educ 1999 Mar;2(2-3):87-105. [doi: 10.1016/S1096-7516(00)00016-6]

XSL•FO
RenderX

28. Lahza HF, Demartini G, Noroozi O, Gašević D, Sadiq S, Khosravi H. Enhancing peer feedback provision through user interface scaffolding: a comparative examination of scripting and self-monitoring techniques. Comput Educ 2025 Jun;230:105260. [doi: 10.1016/j.compedu.2025.105260]
29. Chi MTH, Menekse M. Dialogue patterns in peer collaboration that promote learning. In: Resnick LB, Asterhan CSC, Clarke SN, editors. Socializing Intelligence Through Academic Talk and Dialogue: American Educational Research Association; 2015:263-274. [doi: 10.3102/978-0-935302-43-1_21]
30. Mukadam A, Suresh S, Jacobs C. Beyond traditional simulation: an exploratory study on the effectiveness and acceptability of ChatGPT-4o Advanced Voice Mode for communication skills practice among medical students. Cureus 2025 May;17(5):e84381. [doi: 10.7759/cureus.84381] [Medline: 40535400]
31. Anthamatten A, Holt JE, Pfieffer ML. Developing clinical competence through case presentations with artificial intelligence-driven simulation. J Nurse Pract 2025 Jul;21(7):105415. [doi: 10.1016/j.nurpra.2025.105415]
32. Ding D, Muhyiddin B Yusof A. Investigating the role of AI-powered conversation bots in enhancing L2 speaking skills and reducing speaking anxiety: a mixed methods study. Humanit Soc Sci Commun 2025;12(1):1223. [doi: 10.1057/s41599-025-05550-z]
33. Cooper KM, Downing VR, Brownell SE. The influence of active learning practices on student anxiety in large-enrollment college science classrooms. Int J STEM Educ 2018;5(1):23. [doi: 10.1186/s40594-018-0123-6] [Medline: 30631713]
34. Theobald EJ, Hill MJ, Tran E, et al. Active learning narrows achievement gaps for underrepresented students in undergraduate science, technology, engineering, and math. Proc Natl Acad Sci USA 2020 Mar 24;117(12):6476-6483. [doi: 10.1073/pnas.1916903117]

Abbreviations

GAIDeT: Generative AI Delegation Taxonomy
GenAI: generative AI

Edited by B Lesselroth; submitted 15.Jun.2026; peer-reviewed by K Sugimori, T Clement; revised version received 14.Sep.2026; accepted 15.Sep.2026; published 20.Oct.2026.

Please cite as:
Indran IR, Samarasekera DD
Addressing the Engagement Blind Spot: 12-Tip Tutorial to Make Reasoning Audible in Health Professions Education
JMIR Med Educ 2026;12:e104702
URL: https://mededu.jmir.org/2026/1/e104702
doi:10.2196/104702

© Inthrani Raja Indran, Dujeepa D Samarasekera. Originally published in JMIR Medical Education (https://mededu.jmir.org), 20.Oct.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Education, is properly cited. The complete bibliographic information, a link to the original publication on https://mededu.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Co-Designing and Evaluating a 1-Day Quality Improvement Workshop for Medical Students and Resident Physicians: Tutorial on Applying Kern's Curriculum Development Framework

Amanda Ling Jie Yee[1], MBChB; Mariam Idrissi[1], MBChB; Sam Sherratt-Mayhew[1], MBChB; Charles Page[1], MBChB; Maiar Elhariry[2,3], MBChB; Punith Kempegowda[2,4,5], MD, PhD

[1]Birmingham Medical School, University of Birmingham, Birmingham, United Kingdom

[2]College of Medicine and Health, Applied Health Sciences, School of Health Sciences, University of Birmingham, Birmingham, United Kingdom

[3]Sandwell and West Birmingham NHS Foundation Trust, Birmingham, United Kingdom

[4]Department of Diabetes and Endocrinology, University Hospitals Birmingham NHS Foundation Trust, Birmingham, United Kingdom

[5]Murray Learning Centre, Edgbaston, Birmingham, United Kingdom

Corresponding Author:
Punith Kempegowda, MD, PhD
College of Medicine and Health, Applied Health Sciences, School of Health Sciences, University of Birmingham, Birmingham, United Kingdom

Abstract

Background: Despite the importance of quality improvement in advancing patient care and safety, there is limited literature describing structured, practical, and co-designed quality improvement education.

Objective: This study aimed to (1) describe how learner co-design was operationalized within Kern's 6-step curriculum development framework to develop a quality improvement workshop for medical students and resident physicians, (2) evaluate preworkshop and postworkshop changes in learners' self-reported understanding of and confidence in quality improvement, and (3) explore participants' attitudes toward quality improvement and their perceptions of the workshop's relevance to future practice.

Methods: Using Kern's 6-step curriculum development model, informed by Kolb's Experiential Learning Theory, we co-designed a 1-day quality improvement workshop with medical students and resident physicians. To address objective 1, the workshop development process was guided by a literature review and a targeted needs assessment. To address objective 2, we used a mixed methods pre-post educational evaluation design. The workshop incorporated expert-led lectures, small-group project design exercises, and peer presentations addressing audit methodology, ethical considerations, and practical implementation. Preworkshop and postworkshop surveys assessed changes in participants' self-reported understanding of quality improvement concepts, confidence, and attitudes using 10-point Likert scales. Quantitative data were analyzed using the Wilcoxon matched-pairs signed-rank and Fisher exact tests. Semistructured interviews explored participants' experiences and helped to explain their quantitative responses. Interview transcripts were analyzed using thematic analysis.

Results: Findings from the literature review and targeted needs assessment identified gaps in practical quality improvement education related to project design, implementation, and ethical considerations, which informed workshop co-design. In total, 31 learners attended the workshop, and 77.4% (24/31) completed preworkshop and postworkshop surveys. There was a significant improvement in participants' understanding of the Plan-Do-Study-Act cycle (preworkshop median score 2.0, IQR 1.0 - 2.8 vs postworkshop median score 4.0, IQR 4.0 - 5.0; $P<.001$). Confidence in engaging in quality improvement projects improved significantly (preworkshop median score 4.5, IQR 2.3 - 7.0 vs postworkshop median score 7.5, IQR 6.3 - 8.0; $P=.004$). Self-reported knowledge of additional methodologies, including Six Sigma, Lean, and root cause analysis, also improved significantly. Participants rated the workshop highly (median score 9.5 out of 10). Qualitative findings indicated that participants perceived improved capability in project planning, greater ethical awareness, and stronger motivation to apply learning in clinical practice. These findings reflect self-reported learning experiences rather than objectively verified skill development.

Conclusions: Learner co-design was successfully integrated within Kern's curriculum development framework to develop a practical quality improvement workshop informed by identified learner needs. Participation in the workshop was associated with improved self-reported understanding, confidence, and positive perceptions of relevance and usefulness. Future research should examine longer-term outcomes and evaluate adaptation across broader educational settings.

(JMIR Med Educ 2026;12:e83657) doi:10.2196/83657

XSL•FO
RenderX

KEYWORDS

quality improvement; medical education; curriculum development; experiential learning; mixed methods study

Introduction

Quality improvement (QI) is pivotal in enhancing patient safety, outcomes, and overall health care efficacy [1]. QI is defined as a structured, collaborative, and inclusive approach for sustained transformation and continuous improvement in health care [2]. Medical audits and quality improvement projects (QIPs) are the most common approaches for operationalizing QI principles in clinical practice. Medical audits systematically evaluate health care delivery against established standardized guidelines within a clinical setting [3], whereas QIPs focus on identifying areas for improvement and implementing targeted interventions to enhance the quality of care [4].

Integrating QI principles into medical education curricula is increasingly recognized as essential [5]. The General Medical Council mandates that medical graduates understand and apply QI principles in clinical practice [6]. Furthermore, participation in QI activities forms part of the annual professional revalidation process [7], through which licensed physicians demonstrate that they remain fit to practice in the United Kingdom. Despite its recognized importance, significant gaps remain in the integration of QI into medical education, with limited literature on the impact of QI education on learners [8].

Involving end users in curriculum development improves learning outcomes, motivation, and satisfaction [9,10]. Although examples of co-designed curricula in medical education are increasing, there is a paucity of co-designed QI education. To address this paucity, this study had three objectives: (1) to describe how learner co-design was operationalized within Kern's 6-step curriculum development framework to develop a QI workshop for medical students and resident physicians; (2) to evaluate preworkshop and postworkshop changes in learners' self-reported understanding of and confidence in QI; and (3) to explore participants' attitudes toward QI and their perceptions of the workshop's relevance to future practice.

Methods

Overview

The study was conducted between April 2024 and July 2024 at a UK university that serves approximately 2000 medical students annually. A mixed methods pre-post educational intervention design was used to evaluate a co-designed QI educational workshop [11]. The workshop was developed and delivered by a group of medical students working under the mentorship of a consultant clinician with expertise in QI. These individuals constituted the organizing committee responsible for designing and implementing the educational intervention.

We adopted Kern's 6-step approach to curriculum development, informed by principles from Kolb's Experiential Learning Theory, to guide the co-design and development of the workshop [12,13]. Each step is described in detail later to show how learner co-design informed the development process. To measure the impact of the workshop, we used a pre-post educational evaluation design with surveys and interviews. Quantitative data were used to assess changes in participants' self-reported understanding, confidence, and attitudes, whereas qualitative data were used to explore how participants experienced the workshop and why particular elements were perceived as useful or challenging. The 2 data strands were collected sequentially and integrated at the interpretation stage using a convergent mixed methods approach, whereby the quantitative findings showed what changed after the workshop, and the qualitative findings helped explain why those changes may have occurred.

Participants included medical students from the University of Birmingham Medical School and resident physicians from the University Hospitals Birmingham National Health Service Foundation Trust who voluntarily registered to attend the workshop. Recruitment used a convenience sampling approach [14]. The event was advertised through university mailing lists, social media platforms, and institutional group messaging channels. Snowball sampling also occurred [15], as potential participants shared information about the workshop within their peer networks to encourage attendance.

Ethical Considerations

This project underwent institutional review at the time of submission and was approved to proceed without the requirement for full ethics committee review, as it was deemed not to raise substantial ethical concerns (ERN_2545-May2024). Informed consent was obtained electronically from participants before completion of the preworkshop and postworkshop surveys. Verbal consent was also confirmed at the start of each interview recording. To protect privacy and confidentiality, data were pseudoanonymized with participant ID codes and stored securely in password-protected cloud storage accessible only to the research team. No participant compensation was provided, and participation was entirely voluntary.

Problem Identification and General Needs Assessment

A targeted literature search was conducted by an author (CP) to identify previously reported needs for QI education in medical training. The search strategy is summarized in Multimedia Appendix 1. The search identified 134 records, of which 8 (6) were deemed relevant following screening [8,16-22]. The literature shows that students often lack understanding of QI and that targeted educational interventions can improve learners' knowledge of and attitudes toward QI participation [8,16,17,19,21].

The organizing committee considered several potential approaches to addressing these educational gaps. Rather than proposing formal curriculum reform, which may require substantial institutional resources, the group identified student-led workshops as a feasible and scalable strategy for introducing QI principles to learners [22]. However, the group also recognized potential limitations associated with peer-led education, including the possibility that participants may provide more favorable feedback when educational sessions are

XSL•FO
RenderX

delivered by peers. To mitigate this risk, all workshop content was reviewed by a consultant clinician with expertise in QI and clinical audit prior to delivery to ensure scientific accuracy and educational quality.

Targeted Needs Assessment

Individuals who registered for the workshop were invited to complete a preworkshop survey (Multimedia Appendix 2). Participants provided informed consent prior to completing the survey through an electronic participant information and consent form. The survey included 10-point Likert-scale items assessing participants' baseline knowledge, confidence, and attitudes toward QI activities, as well as their prior experience with audits and QIPs, and their expectations for the workshop. Open-ended questions were included to allow participants to elaborate on their responses and provide additional insights into their experiences and perceived learning needs. Responses to open-ended questions were analyzed thematically to identify common themes informing the development of the workshop curriculum (Multimedia Appendix 3).

Participants who completed the preworkshop survey were invited to participate in a semistructured interview to explore their experiences and perceptions of QI in greater depth (Multimedia Appendix 4). The interview guide was specifically developed by the research team to ensure that key topics were consistently addressed across participants. Interviews were conducted either via Zoom (Zoom Communications Inc) or in person, depending on participants' accessibility. Audio recordings were obtained with participant consent prior to the start of each interview. Among the 27 participants who completed the preworkshop survey, 12 (44.4%) agreed to participate in the interview (n=10, 83.3% medical students and n=2, 16.7% resident physicians). Interviews lasted approximately 20 minutes. All recordings were transcribed verbatim while ensuring accuracy. Transcripts were reviewed and analyzed independently by 2 study authors (ALJY and MI) using thematic analysis [23]. Data were stored securely in password-protected institutional cloud storage accessible only to members of the research team.

Two researchers collaboratively developed a coding framework after familiarizing themselves with the interview transcripts. Discrepancies in coding were discussed and resolved through consensus to ensure consistency. Qualitative analysis was conducted using NVivo (version 14; Lumivero LLC). Intercoder reliability was assessed using the Cohen kappa coefficient [24]. Codes were inductively grouped into broader themes, which were subsequently reviewed and refined.

Goals and Objectives

On the basis of the findings from the literature review and targeted needs assessment, the program goal and workshop learning objectives were developed to ensure that the final curriculum directly addressed the integration of learner co-design into the workshop's development. The overall program goal was to equip early-career health care learners with foundational knowledge, skills, and confidence to engage in QI and medical audit activities in clinical practice.

The workshop was designed to achieve the following learning objectives: (1) to improve participants' understanding of medical audits and QIPs, (2) to increase participants' confidence in engaging with QI activities, and (3) to strengthen participants' awareness of practical, ethical, and implementation issues relevant to QI work.

These outcomes were evaluated using self-reported preworkshop and postworkshop survey responses and qualitative interviews, which together addressed the study's second objective.

Educational Strategies

The workshop curriculum was designed using Kolb's Experiential Learning Theory [13], which conceptualizes learning as a cyclical process involving 4 stages: concrete experience, reflective observation, abstract conceptualization, and active experimentation. Educational activities were structured to approximate these stages within the constraints of a 1-day workshop. The workshop agenda can be found in Multimedia Appendix 5.

Implementation

Concrete Experience: Engaging With Practical QI Scenarios

Participants first engaged in small-group activities designed to simulate the early stages of QI project development. These sessions were facilitated by medical students with prior experience in conducting QIPs. Activities included identifying clinical problems, brainstorming potential QIP ideas, developing SMART (specific, measurable, achievable, relevant, and time-bound) project objectives, designing interventions using Plan-Do-Study-Act (PDSA) cycles, and considering potential data collection strategies and ethical considerations.

These activities provided participants with opportunities to engage with QI concepts through practical problem-solving rather than passive learning.

Reflective Observation: Feedback and Peer Learning

Following the group activities, participants engaged in structured reflection sessions facilitated by workshop leaders. Participants reviewed their proposed project ideas, discussed potential implementation barriers, and reflected on their decision-making processes.

The workshop also included presentations of completed audits and QIPs delivered by senior medical students and resident physicians. These presentations were followed by "question-and-answer" and "expert feedback" sessions from faculty members.

These activities supported reflective observation by encouraging participants to critically examine both their own proposed ideas and real-world QI implementation examples.

Abstract Conceptualization: Building Theoretical Understanding

Expert-led lectures were delivered by clinicians with expertise in QI, clinical governance, and medical education. These sessions introduced the theoretical foundations of QI methodologies and contextualized the practical activities.

XSL•FO
RenderX

Lecture topics included definitions and distinctions between audits, QIPs, and research; examples of successful QI initiatives in clinical practice; regulatory and institutional structures supporting QI activities; and ethical and practical considerations when conducting QIPs.

These sessions supported the abstract conceptualization stage of Kolb's learning cycle by linking participants' practical experiences to a broader theoretical framework.

Active Experimentation: Applying Learning in Clinical Practice

The final stage of Kolb's experiential learning cycle involves applying newly acquired knowledge in real-world settings. Although full implementation of QI initiatives extends beyond the workshop itself, participants were encouraged to initiate or participate in audits and QIPs within their clinical training environments following the workshop.

Participants were also invited to present their QI work at subsequent workshops, thereby supporting continued engagement with QI initiatives and extending the experiential learning process beyond the initial educational intervention.

Evaluation and Feedback

Participants were invited to complete a postworkshop survey (Multimedia Appendix 6) using the same 10-point Likert-scale measures as the preworkshop survey to evaluate changes in self-reported understanding, confidence, and attitudes toward QI. Participants also rated individual workshop components using a 5-point scale and provided an overall workshop rating using a 10-point Likert scale. Participants who completed the survey were invited to participate in follow-up interviews to explore their experiences in greater depth (Multimedia Appendix 7). Of the 31 participants who completed the postworkshop survey, 15 (48.4%) agreed to be interviewed. The interview adopted a semistructured format similar to the preworkshop interviews and lasted approximately 15 minutes.

Quantitative survey data were analyzed using Prism (version 10.3.0; GraphPad Software LLC). Participants who completed both preworkshop and postworkshop surveys were included in the analysis. Descriptive statistics for nonparametric continuous data are presented as median (IQR), while categorical data are summarized as n (%). Ordinal data derived from Likert scales were described with medians and IQRs, and the significance of the differences was assessed using the Wilcoxon matched-pairs signed-rank test. The Kruskal-Wallis test was used to identify the most efficacious aspects of the workshop, given the nonparametric nature of the data. The Fisher exact test was used to calculate *P* values for categorical data. Statistical significance was set at $P<.05$. Responses to the open-ended questions and postevent interviews were analyzed using the same approach as for the preworkshop survey (Multimedia Appendix 8).

Results

Participant Characteristics

In total, 31 delegates attended the event (n=29, 93.5% medical students and n=2, 6.5% resident physicians; n=24, 88.9% female). Attendees represented a range of ethnic backgrounds (Asian: n=17, 63%, White: n=6, 22.2%, and Black: n=4, 14.8%). Overall, 87.1% (27/31) participants completed the preworkshop survey (median age 21, IQR 21-22 years), and 77.4% (24/31) participants completed both preworkshop and postworkshop surveys.

Preworkshop Needs Assessment Findings

Overview

Intercoder reliability was substantial, with a Cohen κ coefficient of 0.70. A summary of the coding framework and resulting themes is provided in Multimedia Appendix 9. Three primary thematic domains were identified.

Education and Skills Development in Audits and QIPs

Although most participants had some exposure to audits and QIPs during medical school, there was a consensus on the inadequacy of the practical training. The PDSA cycle was the most recognized QI tool. Participants expressed confidence in data collection but acknowledged limited knowledge of designing a QIP and had limited skills in data analysis:

> *We did it this year as part of a 4th year curriculum, and I feel like...we weren't taught a lot about what it actually was by the medical school. Even though we did a project on it. I still think a lot of people don't really understand why it's important and why we need to, as HCPs [health care professionals], get involved in doing it.* [Participant 004]

Challenges and Solutions in Audits and QIPs

Key challenges included resource constraints, a lack of incentives, difficulty generating novel project ideas, and the time-consuming nature of the audit process. Participants proposed dedicated roles for QI, improved support systems, and the importance of acknowledging errors and seeking feedback. Table 1 shows the highlighted challenges and the corresponding examples.

XSL•FO
RenderX

Table . Challenges in conducting audits and quality improvement projects.

Challenges	Examples
Lack of incentive or motivation	"...most of the time quality improvement projects are carried out by doctors themselves taking the time out...hard to find an incentive to do that..." [Participants 003]
Lack of ideas	"...coming with like a unique idea of a QIP and trying to do something that hasn't been done before." [Participants 002]
Tedious process	"...when you first start, when you have your data collection tool...it doesn't accurately represent what you want to get out of it...have to use a small proportion of the patients...go back and amend..." [Participants 010]
Mistake acknowledgment	"...to acknowledge that mistakes have been made...take over the ego aspect..." [Participants 005]
Funding	"...if you have to bring something new into the world or into the hospital, then it becomes challenging." [Participants 001]

Patient Safety, Confidentiality, and Ethical Considerations

Participants highlighted the importance of acting in patients' best interests, anonymizing data, and adhering to strict confidentiality protocols. Ethical considerations included avoiding bias, maintaining neutrality, and respecting patient autonomy. However, there was limited experience or expertise on how to achieve these:

> *...So abiding by like the trust confidentiality...Not taking any confidential information with you outside the trust, anonymizing the data to the best that you can...and patient safety. I guess it depends on what you're doing. But if you're changing like procedures and stuff, making sure that it's safe to do so, and like getting permission from supervisors and talking to people that have a lot more expertise about if there are any patient concerns and stuff with whatever you're trying to propose.* [Participant 008]

Quantitative Postworkshop Findings

Knowledge and Skills

Understanding of QIPs improved significantly (preworkshop median score 4.5, IQR 3.0-6.0 vs postworkshop median score 7.0, IQR 7.0-8.0; $P<.001$) as did confidence in participating in QIPs (preworkshop median score 4.5, IQR 2.3-7.0 vs postworkshop median score 7.5, IQR 6.3-8.0; $P=.004$; Figure 1). Self-reported understanding of various QI methodologies also improved significantly (Table 2).

XSL•FO
RenderX

Figure 1. Changes in participants' confidence scores for participating in a quality improvement project (QIP). (A) Confidence in participating in a QIP and (B) understanding of QIPs before and after the workshop. • *P*<.05, **P*<.01, ***P*<.001. Horizontal lines are medians with 95% CIs.

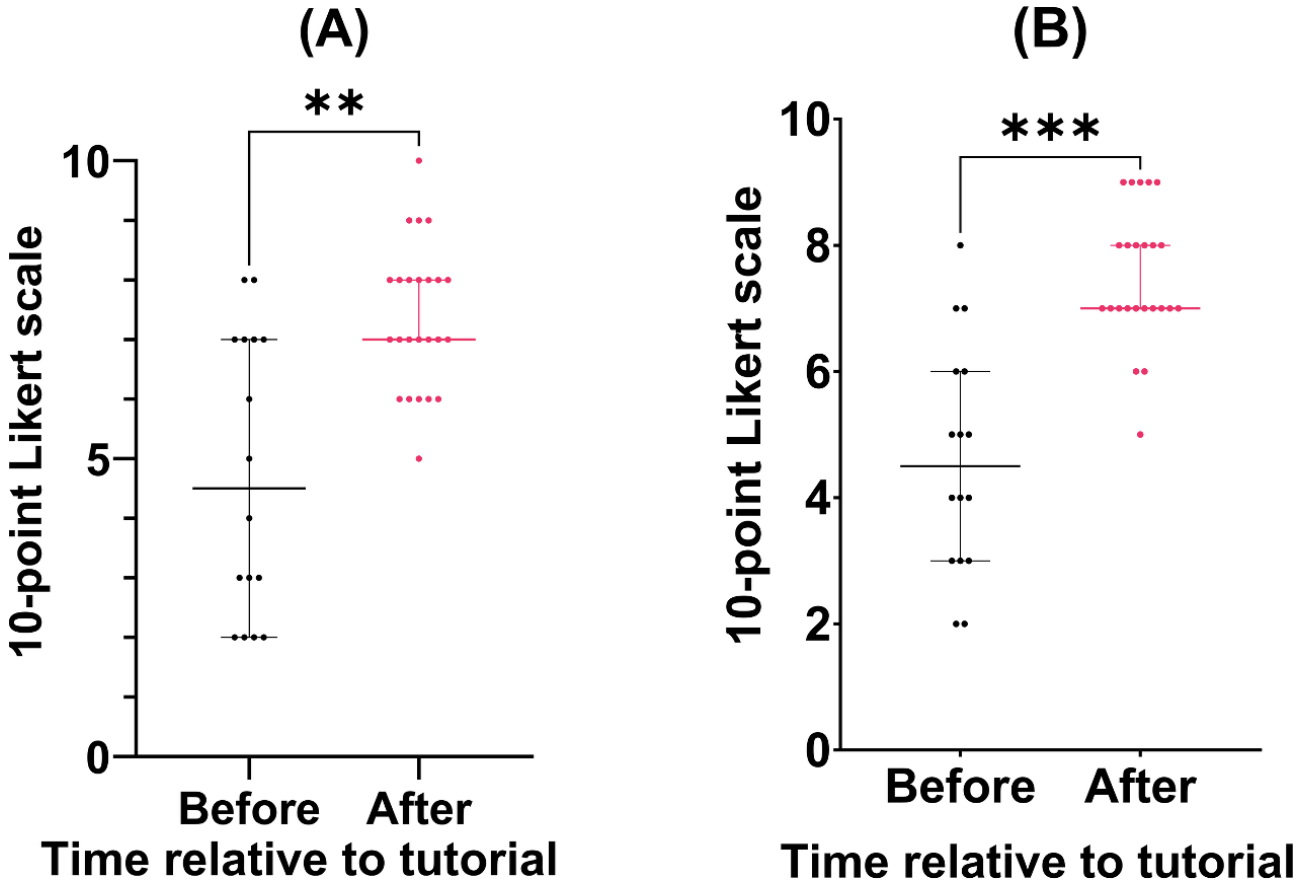


Table . Change in understanding of specific quality improvement (QI) methodologies.

QI methodologies	Pre-event score, median (IQR)	Postevent score, median (IQR)	*P* value
Medical audits	3.0 (2.0-4.0)	4.0 (4.0-5.0)	<.001
Plan-Do-Study-Act cycle	2.0 (1.0-2.8)	4.0 (4.0-5.0)	<.001
Six sigma	2.0 (1.0-2.0)	3.0 (2.0-4.0)	<.001
Model of improvement	2.0 (1.0-2.0)	3.0 (2.3-4.0)	.001
Lean methodology	1.5 (1.0-2.0)	3.0 (2.0-3.0)	.002
Total quality management	1.5 (1.0-2.0)	3.0 (2.0-3.0)	.007
Root cause analysis	2.0 (2.0-3.0)	3.0 (2.0-3.8)	.03

Attitudes Toward QI

Participants felt that QI education is essential for improving patient outcomes and reported a high intention to apply the skills learned in the workshop. Barriers to implementing QIPs included limited knowledge, lack of support, time constraints, and insufficient resources.

Workshop Evaluation

Overall experience was rated highly (median score 9.5 out of 10). All components (lectures, hands-on activities, group discussions, and materials) were well received, with no significant differences in ratings. Participants emphasized the value of lectures, workshops, and oral presentations. Suggestions for improvement included broader participant inclusion, more frequent workshops, wider resource dissemination, and more expert speakers.

Postworkshop Thematic Findings

Overview

Intercoder reliability was substantial, with a Cohen κ coefficient of 0.61. Four primary domains emerged from the postworkshop interviews. All codes and themes were summarized in Multimedia Appendix 10.

Positive Evaluation of the Workshop

Participants reported a substantially improved understanding of audits, QIPs, and implementation research. They valued the relevance of the workshop content, its delivery, and its alignment with their learning needs:

XSL•FO
RenderX

> *Overall, it was quite a positive experience, and I think the content felt very relevant. There was a lot of actionable stuff that was taught there, I think, and stuff I can use in the future to do audits and QIPs.* [Participant 002]

Knowledge and Skill Acquisition

Participants reported acquiring practical skills such as idea generation, objective setting, task allocation, and data analysis. They also gained a clearer understanding of the distinctions between audits and QIPs:

> *...made clear what is audit and what is QIP, and how it is different from the implementation research. I think now, I have a clearer idea of what is audit and what is QIP.* [Participant 001]

Recommendations for Future Participation and Access to QI Opportunities

Participants suggested improvements for future workshops, including broader participant inclusion, improved topic selection, and better signposting to QI opportunities and resources:

> *...like a dedicated sort of network, networking or signposting for opportunities out there, and how to get opportunities.* [Participant 003]

Ethical Awareness and Considerations

Participants highlighted the importance of patient confidentiality and informed consent in audits and QIPs. They also noted the complexity of ethics approval processes:

> *I was not aware of the ethical committees early on, but attending the conference helped me understand how projects are approved through the ethics committee. They also shared their experiences of the process, including that approval can take a couple of months. Overall, it gave me a better insight into these ethical issues.* [Participant 001]

Discussion

Principal Results

Co-design within a structured curriculum development model is feasible and may help produce a workshop that learners perceive as relevant, practical, and effective in building confidence. To our knowledge, this is among the few reported learner–co-designed QI workshops that integrate Kolb's Experiential Learning Theory [13] into Kern's curriculum development framework [12]. The workshop was associated with improved participants' self-reported understanding and confidence, suggesting greater perceived readiness to engage in QIPs. Learner involvement in the design process may have helped ensure that the workshop addressed real-world learning needs, improved relevance, and promoted active engagement, which may be less evident in traditional didactic approaches.

Although experiential learning is fully realized through implementation of QI initiatives in real clinical environments, educational workshops can approximate key elements of this cycle through small-group activities, structured project design exercises, and facilitated reflection. In this workshop, participants engaged in simulated QIP planning and peer discussions, allowing them to apply theoretical concepts and reflect on potential implementation challenges before undertaking real-world QI initiatives.

Our findings are consistent with the broader literature synthesized by Peiris-John et al [8], which suggests that structured QI teaching can improve learner-level outcomes such as self-reported knowledge, confidence, and engagement. The workshop was similar to the interventions identified in that review because it combined multiple educational strategies, including didactic teaching, experiential learning, and small-group discussion. However, this study differed in that it evaluated a 1-day workshop that was co-designed with learners within Kern's curriculum development framework and complemented by a qualitative evaluation of participants' experiences [12].

Kern's model was chosen for its practical [12] and stepwise approach to curriculum development, which aligns well with medical education, particularly when an intervention must progress from needs assessment to implementation and evaluation within a clearly structured and reportable framework. Its flexibility also supports refinement of educational programs in response to identified learner needs and feedback, and it has been widely applied in medical education [25]. Alternative approaches were considered but were less suitable for the aims of this study. For example, backward design is useful for aligning learning outcomes with assessment [26], but it places less emphasis on needs assessment and stakeholder-informed problem definition, both of which were central to the co-design process. Design-based research could also have been used because it supports iterative refinement in authentic settings [27]; however, it typically requires multiple cycles of testing and redesign and was therefore less feasible for a single workshop delivered within a limited time frame. Similarly, intervention development approaches primarily grounded in implementation science may have been useful for later scale-up [28], but they were less well suited to reporting the educational curriculum development process itself.

A convergent mixed methods design was chosen because it enabled triangulation of quantitative and qualitative findings, providing complementary insights into the educational intervention [29]. In this study, the quantitative component identified what changed after the workshop, particularly in participants' self-reported understanding, confidence, and attitudes, whereas the qualitative component helped explain why these changes may have occurred by exploring participants' experiences of the workshop. Alternative mixed methods approaches, including an explanatory sequential design, would typically require quantitative data collection and analysis to be completed before qualitative follow-up to explain specific findings, which was less practical within the time frame of a single workshop evaluation. An exploratory sequential design was also less appropriate, as this is generally used when qualitative findings are intended to inform the subsequent development of a quantitative instrument or intervention, which was not the purpose of this study.

XSL•FO
RenderX

This structured curriculum development process, based on Kern's framework [12], supports transparency and reproducibility, which is a strength of the study. The workshop was co-designed with learners, which helped ensure that the educational content addressed identified learner needs and was perceived as relevant to practice. The mixed methods design strengthens the evaluation by combining quantitative and qualitative data, allowing the study to show not only what changed after the workshop, but also why participants perceived the intervention as useful and applicable. Finally, the workshop provides a practical, resource-efficient educational model that can be adapted to similar training contexts.

Limitations

However, although the institution serves a large medical student population, this study evaluated a single voluntarily attended workshop using convenience and snowball sampling. As such, the sample was not intended to be statistically representative of the wider student population. These sampling approaches also limit the transferability of the findings to other settings and learner groups. A further limitation is that the evaluation primarily assessed self-reported understanding, confidence, and attitudes rather than objectively measuring knowledge acquisition or practical QI skills. The workshop placed greater emphasis on some QI approaches, particularly PDSA cycles, which meant that other methodologies were explored less extensively. In addition, longer-term follow-up examining whether participants subsequently implemented QI projects was not conducted due to logistical constraints and the limited study time frame.

Future research should explore the scalability of this model across multiple medical schools, assess its long-term impact on learners' QI engagement in practice, and incorporate perspectives from a wider range of stakeholders, including faculty and institutional leaders. Future studies should also include more objective educational outcomes, such as preintervention and postintervention knowledge tests, case-based assessments, observed structured scenarios, rubric-based evaluations, or workplace-based assessments. In addition, future work could examine whether longitudinal one-to-one mentorship models support sustained engagement and facilitate the development of tangible scholarly outputs such as conference posters or oral presentations.

Conclusions

The learner co-designed a QI workshop that improved understanding of QI concepts and increased self-reported confidence in engaging in QI activities, with positive perceptions of the workshop's relevance and usefulness. Future research is needed to examine longer-term outcomes and assess whether this co-designed model can be adapted effectively across broader educational settings.

Acknowledgments

The authors thank all medical students and health care professionals who participated in this study. The authors thank Dr Muhammad Ali Karamat for sharing his expert opinion during the conference. They also thank Dr Aspasia Manta for reviewing the paper to ensure methodological rigor. The authors declare that they used the generative artificial intelligence tool ChatGPT (GPT-5.4; OpenAI) for manuscript editing.

Funding

This study received no specific funding but was supported in kind by the Midlands Patient Safety Research Collaboration and the National Institute for Health and Care Research (NIHR) Race, Equity, and Diversity in Careers Incubator.

Data Availability

The quantitative datasets generated or analyzed during this study are available from the corresponding author on reasonable request. The raw qualitative interview transcripts generated or analyzed during this study are not publicly available as they contain potentially identifying information, but deidentified excerpts supporting the findings are available from the corresponding author on reasonable request.

Authors' Contributions

Conceptualization: PK (lead), ALJY (equal)
Data curation: ALJY (qualitative lead), SS-M (quantitative lead), MI (supporting), ME (supporting), PK (supporting)
Formal analysis: ALJY (qualitative lead), SS-M (quantitative lead), MI (supporting), ME (supporting), PK (supporting)
Investigation: ALJY (lead), MI (equal)
Methodology: PK (lead), ALJY (equal), ME (supporting)
Project administration: ALJY (lead), PK (equal), MI (supporting), SS-M (supporting), CP (supporting)
Resources: ALJY (lead), MI (supporting), SS-M (supporting), CP (supporting), PK (supporting)
Supervision: ALJY (lead), PK (equal), MI (supporting), SS-M (supporting), CP (supporting), ME (supporting)
Validation: ALJY (lead), PK (equal), MI (supporting), SS-M (supporting), CP (supporting)
Visualization: ALJY (lead), SS-M (equal), CP (supporting), PK (supporting)
Writing—original draft: ALJY (lead), MI (equal), SS-M (supporting), CP (supporting), ME (supporting)
Writing—review and editing: PK (lead), ALJY (equal), MI (supporting), SS-M (supporting), CP (supporting), ME (supporting)

XSL•FO
RenderX

Conflicts of Interest

None declared.

Multimedia Appendix 1
Search strategy for problem identification as part of Kern's 6-step approach to curriculum development.
[DOCX File, 17 KB - mededu_v12i1e83657_app1.docx]

Multimedia Appendix 2
Early-career Physicians & Investigators Collaboration 2024 preworkshop survey.
[DOCX File, 722 KB - mededu_v12i1e83657_app2.docx]

Multimedia Appendix 3
Themes and codes developed through thematic analysis of the preworkshop survey.
[DOCX File, 19 KB - mededu_v12i1e83657_app3.docx]

Multimedia Appendix 4
Guide for preworkshop interview.
[DOCX File, 18 KB - mededu_v12i1e83657_app4.docx]

Multimedia Appendix 5
Program of events.
[PDF File, 385 KB - mededu_v12i1e83657_app5.pdf]

Multimedia Appendix 6
Postworkshop survey.
[DOCX File, 385 KB - mededu_v12i1e83657_app6.docx]

Multimedia Appendix 7
Postworkshop interview guide.
[DOCX File, 17 KB - mededu_v12i1e83657_app7.docx]

Multimedia Appendix 8
Themes and codes developed through thematic analysis of the postworkshop survey.
[DOCX File, 29 KB - mededu_v12i1e83657_app8.docx]

Multimedia Appendix 9
Themes and domains developed through thematic analysis of the preworkshop interviews.
[DOCX File, 22 KB - mededu_v12i1e83657_app9.docx]

Multimedia Appendix 10
Themes and domains developed through thematic analysis of the postworkshop interviews.
[DOCX File, 19 KB - mededu_v12i1e83657_app10.docx]

References

1. Batalden PB, Davidoff F. What is "quality improvement" and how can it transform healthcare? Qual Saf Health Care 2007 Feb;16(1):2-3. [doi: 10.1136/qshc.2006.022046] [Medline: 17301192]
2. Quality improvement support. Healthcare Quality Improvement Partnership. URL: https://www.hqip.org.uk/services/quality-improvement [accessed 2026-03-11]
3. Clinical audit. NHS England. URL: https://www.england.nhs.uk/clinaudit/ [accessed 2026-05-16]
4. Adams D. Quality improvement; part 1: introduction and overview. BJA Educ 2018 Mar;18(3):89-94. [doi: 10.1016/j.bjae.2017.12.002] [Medline: 33456816]
5. Wong BM, Levinson W, Shojania KG. Quality improvement in medical education: current state and future directions. Med Educ 2012 Jan;46(1):107-119. [doi: 10.1111/j.1365-2923.2011.04154.x]
6. Outcomes for graduates summary. General Medical Council. URL: https://www.gmc-uk.org/education/standards-guidance-and-curricula/standards-and-outcomes/outcomes-for-graduates/outcomes-for-graduates [accessed 2026-03-11]

XSL•FO
RenderX

7. Guidance on supporting information for revalidation guide. General Medical Council. 2024. URL: https://www.gmc-uk.org/registration-and-licensing/managing-your-registration/revalidation/guidance-on-supporting-information-for-revalidation [accessed 2026-03-11]
8. Peiris-John R, Selak V, Robb G, et al. The state of quality improvement teaching in medical schools: a systematic review. J Surg Educ 2020;77(4):889-904. [doi: 10.1016/j.jsurg.2020.01.003] [Medline: 32057742]
9. Geraghty JR, Young AN, Berkel TD, et al. Empowering medical students as agents of curricular change: a value-added approach to student engagement in medical education. Perspect Med Educ 2020 Feb;9(1):60-65. [doi: 10.1007/s40037-019-00547-2] [Medline: 31823304]
10. Rae VI, Smith SE, Hopkins SR, Tallentire VR. From corners to community: exploring medical students' sense of belonging through co-creation in clinical learning. BMC Med Educ 2024 Apr 30;24(1):474. [doi: 10.1186/s12909-024-05413-2] [Medline: 38689267]
11. Smith-Lickess SK, Woodhead T, Burhouse A, Vasilakis C. Study design and protocol for a comprehensive evaluation of a UK massive open online course (MOOC) on quality improvement in healthcare. BMJ Open 2019 Dec;9(12):e031973. [doi: 10.1136/bmjopen-2019-031973] [Medline: 31874877]
12. Thomas PA, Kern DE, Hughes MT, Tackett SA, Chen BY. Curriculum Development for Medical Education: A Six-Step Approach, 4th edition: Johns Hopkins University Press; 2022.
13. Kolb DA. Experiential Learning: Experience as the Source of Learning and Development: Pearson Education; 2015.
14. Andrade C. The inconvenient truth about convenience and purposive samples. Indian J Psychol Med 2021 Jan;43(1):86-88. [doi: 10.1177/0253717620977000] [Medline: 34349313]
15. Parker C, Scott S, Geddes A. Snowball Sampling: SAGE Publications; 2019.
16. Dysinger WS, Pappas JM. A fourth-year medical school rotation in quality, patient safety, and population medicine. Am J Prev Med 2011 Oct;41(4 Suppl 3):S200-S205. [doi: 10.1016/j.amepre.2011.06.011] [Medline: 21961665]
17. Levitt DS, Hauer KE, Poncelet A, Mookherjee S. An innovative quality improvement curriculum for third-year medical students. Med Educ Online 2012;17. [doi: 10.3402/meo.v17i0.18391] [Medline: 22611330]
18. Tartaglia KM, Walker C. Effectiveness of a quality improvement curriculum for medical students. Med Educ Online 2015;20:27133. [doi: 10.3402/meo.v20.27133] [Medline: 25960052]
19. Maddalena V, Pendergast A, McGrath G. Quality improvement in curriculum development. Leadersh Health Serv (Bradf Engl) 2018 Oct 1;31(4):409-412. [doi: 10.1108/LHS-09-2017-0053] [Medline: 30234451]
20. Shah KP, Goyal S, Ramachandran V, et al. Efficacy of quality improvement and patient safety workshops for students: a pilot study. BMC Med Educ 2020 Apr 23;20(1):126. [doi: 10.1186/s12909-020-1982-3] [Medline: 32326951]
21. Aredo JV, Ding JB, Lai CH, et al. Implementation and evaluation of an elective quality improvement curriculum for preclinical students: a prospective controlled study. BMC Med Educ 2023 Jan 26;23(1):66. [doi: 10.1186/s12909-023-04047-0] [Medline: 36703204]
22. Robertson C, Smith K, McGowan A, Sammut T, Boyle JG. New approach to undergraduate quality improvement. Clin Teach 2023 Apr;20(2):e13568. [doi: 10.1111/tct.13568] [Medline: 36859750]
23. Ahmed SK, Mohammed RA, Nashwan AJ, et al. Using thematic analysis in qualitative research. Int J Hum Comput Interact 2025 Aug;6:100198. [doi: 10.1016/j.glmedi.2025.100198]
24. McHugh ML. Interrater reliability: the kappa statistic. Biochem Med (Zagreb) 2012;22(3):276-282. [Medline: 23092060]
25. Zimmermann K, Holzinger IB, Ganassi L, et al. Inter-professional in-situ simulated team and resuscitation training for patient safety: description and impact of a programmatic approach. BMC Med Educ 2015 Oct 29;15:189. [doi: 10.1186/s12909-015-0472-5] [Medline: 26511721]
26. Jensen JL, Bailey EG, Kummer TA, Weber KS. Using backward design in education research: a research methods essay. J Microbiol Biol Educ 2017;18(3):18.3.50. [doi: 10.1128/jmbe.v18i3.1367] [Medline: 29854045]
27. Scott EE, Wenderoth MP, Doherty JH. Design-based research: a methodology to extend and enrich biology education research. CBE Life Sci Educ 2020 Sep;19(3):es11. [doi: 10.1187/cbe.19-11-0245] [Medline: 32870086]
28. Mielke J, Brunkert T, Zúñiga F, Simon M, Zullig LL, De Geest S. Methodological approaches to study context in intervention implementation studies: an evidence gap map. BMC Med Res Methodol 2022 Dec 14;22(1):320. [doi: 10.1186/s12874-022-01772-w] [Medline: 36517765]
29. Fetters MD, Curry LA, Creswell JW. Achieving integration in mixed methods designs-principles and practices. Health Serv Res 2013 Dec;48(6 Pt 2):2134-2156. [doi: 10.1111/1475-6773.12117] [Medline: 24279835]

Abbreviations

PDSA: Plan-Do-Study-Act
QI: quality improvement
QIP: quality improvement project
SMART: specific, measurable, achievable, relevant, and time-bound

XSL•FO
RenderX

Edited by A Stone; submitted 11.Sep.2025; peer-reviewed by T Mesbah, T Anstett; revised version received 02.May.2026; accepted 04.May.2026; published 17.Jun.2026.

Please cite as:
Ling Jie Yee A, Idrissi M, Sherratt-Mayhew S, Page C, Elhariry M, Kempegowda P
Co-Designing and Evaluating a 1-Day Quality Improvement Workshop for Medical Students and Resident Physicians: Tutorial on Applying Kern's Curriculum Development Framework
JMIR Med Educ 2026;12:e83657
URL: https://mededu.jmir.org/2026/1/e83657
doi:10.2196/83657

© Amanda Ling Jie Yee, Mariam Idrissi, Sam Sherratt-Mayhew, Charles Page, Maiar Elhariry, Punith Kempegowda. Originally published in JMIR Medical Education (https://mededu.jmir.org), 17.Jun.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Education, is properly cited. The complete bibliographic information, a link to the original publication on https://mededu.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Empowering Educators to Build AI Chatbots in Health Professions Education: Tutorial on a No-Code Design Workflow

Charlene Xing Le Goh[1*]; Eugene Wang[1*]; Juanita S M Kong[2], BSc; Judy C G Sng[2*], BSc (Hons), MMSc, PhD

[1]Yong Loo Lin School of Medicine, National University of Singapore, Singapore, Singapore

[2]Department of Pharmacology, Yong Loo Lin School of Medicine, National University of Singapore, 16 Medical Drive MD3 Level 4, Singapore, Singapore

[*]these authors contributed equally

Corresponding Author:
Judy C G Sng, BSc (Hons), MMSc, PhD
Department of Pharmacology, Yong Loo Lin School of Medicine, National University of Singapore, 16 Medical Drive MD3 Level 4, Singapore, Singapore

Abstract

AI is entering clinical practice faster than health professions curricula can teach it, leaving many educators eager to use AI-based teaching tools but unsure of how to build them. Generative AI chatbots—configured as simulated patients, clinical coaches, or formative assessment partners—offer scalable, interactive practice without any programming, yet most educators lack a structured method for designing and deploying them well. This tutorial provides that method: a practical, platform-agnostic workflow for building no-code AI chatbots using widely available large language model platforms. The workflow is organized in five sequential sections that follow the arc of a design project: (1) defining the educational purpose, learner group, and persona; (2) configuring behavior through the system prompt, graduated information disclosure, and structured feedback; (3) refining the learner experience through communication-style calibration, voice interaction, and curated knowledge documents; (4) adding realism and testing, including AI avatar generation and rigorous pilot-testing; and (5) embedding the tool within the curriculum and governing its use ethically. Each section pairs concrete, copy-ready design steps with the reasoning behind them, drawing on the technological pedagogical content knowledge framework and on established learning mechanisms—deliberate practice, self-regulated learning, formative feedback, and simulation-based learning—so that design choices are pedagogically grounded rather than merely technical. Throughout, 2 locally developed initiatives, the Virtual Integrated Patient and the Depression Avatars project, serve as illustrative implementation examples that motivated specific design decisions. These are presented as feasibility and acceptability experiences, not as evidence of educational effectiveness. The tutorial also addresses when a chatbot is not the right tool, the principal risks (hallucination, automation bias, data privacy exposure, and bias in generated personas), and a practical governance checklist for safe deployment. Although the examples are clinical, the workflow is discipline-agnostic and transferable across higher education. No-code AI chatbots are a feasible, accessible way for educators to build interactive learning tools; rigorous, multi-institutional evaluation using validated instruments remains the essential next step.

(JMIR Med Educ 2026;12:e103399) doi:10.2196/103399

KEYWORDS

artificial intelligence; AI; chatbot; no code; health professions education; virtual patient; prompt engineering; medical education; tutorial; generative AI

Introduction

The integration of AI into health care has outpaced the formal training offered in nursing and medical education programs [1]. At the same time, surveys of health professions students reveal strong enthusiasm for AI-assisted learning, with most believing that AI tools will contribute to improved patient outcomes, new research breakthroughs, and greater efficiency in clinical practice (TJ Wong, unpublished data, January 2025). This paradox—students eager for AI-enhanced education within systems slow to adopt it—creates both an imperative and an opportunity for individual educators to act.

One accessible entry point is the AI chatbot. Unlike purpose-built clinical simulation platforms that require institutional infrastructure and technical teams, modern large language model (LLM)–powered chatbots can be configured by educators themselves using no-code interfaces. These tools allow for the construction of interactive learning experiences (simulated patients, clinical coaches, and formative assessment partners) without a single line of programming. The educator's role shifts from passive consumer to active designer, a transition that aligns with emerging frameworks for the evolving role of teachers in an AI-enabled environment [2].

XSL•FO
RenderX

This tutorial is scoped specifically to generative AI (GenAI) platforms—tools powered by LLMs with a configurable instruction or system prompt field (see Table 1 for representative examples). Rule-based or menu-driven chatbot builders, which operate on predefined decision trees, require different design approaches and are outside the scope of this tutorial.

Table . Representative no-code tools for educators (as of 2025). This list is representative, not exhaustive. Platform features, pricing tiers, and regional availability change frequently. Educators are advised to verify current terms of service, data sharing policies, and institutional licensing agreements before deployment.

Category and tool	Free tier	Key feature	Educator notes
LLM[a] chatbot platforms			
ChatGPT (OpenAI)	Yes (limited)	Custom GPT[b] builder; voice mode; knowledge upload	Custom GPT builder requires a paid plan; most widely used in education worldwide
Microsoft Copilot	Yes (institutional)	Integrated with Microsoft 365; voice input; enterprise privacy	Free via many institutional Microsoft licenses; strong data governance options
Google Gemini	Yes (limited)	Multimodal input; Google Workspace integration	Custom persona Gems available on paid plans
Claude (Anthropic)	Yes (limited)	Large context window; strong instruction following	Projects feature enables persona definition and document grounding
DeepSeek	Yes	Strong reasoning; accessible where other platforms are restricted	Open-source model; data privacy policies vary by deployment context
Avatar and video tools			
D-ID	Trial only	Text-to-avatar video; emotion and movement controls	Used in the NUS[c] Depression Avatars pilot; photorealistic output; credit-based pricing
HeyGen	Trial only	Photo and video cloning; voice cloning; multilingual translation	Strong multilingual support; widely adopted for educational video production
Synthesia	Trial only	Studio-quality avatars; slide integration; team sharing	Well suited for institutional-scale educational video production
Voice-specific tools			
ElevenLabs	Yes (limited)	High-fidelity voice synthesis and voice cloning	Can generate culturally appropriate patient voices for avatar scripts
NotebookLM (Google)	Yes	Audio overview mode from uploaded documents	Useful for converting course materials into spoken summaries for learners

[a]LLM: large language model.

[b]GPT: generative pretrained transformer.

[c]NUS: National University of Singapore.

While the examples throughout are drawn from health professions education—where patient simulation, clinical communication, and diagnostic reasoning present distinctive design challenges—the underlying workflow is discipline agnostic. Educators in law, engineering, social work, and the humanities will find the core principles equally applicable, substituting clinical scenarios with discipline-appropriate cases.

The workflow is presented in 5 sequential sections summarized in Table 2. Each section addresses a distinct phase of the design and deployment process, from initial planning to ethical stewardship of the deployed tool. A complete step-by-step build guide covering the text-based, voice-enabled, and avatar workflows described here, including ready-to-copy prompts and platform-specific instructions, is provided in Multimedia Appendix 1.

XSL•FO
RenderX

Table . The no-code chatbot design workflow: 5 sections at a glance. Each row states a general design principle; items marked as "Example" are illustrations drawn from health professions education. The principles are discipline agnostic, and the specific attributes named (eg, a patient's presenting concern) apply only to the persona type illustrated—educators in other fields should substitute the attributes relevant to their own context.

Section number	Section name	Focus	Design decisions covered
1	Before you build	Defining the purpose, learner, and persona	• Define the specific learning outcomes, target learner group, and stage of training. Apply the TPACK[a] framework to align technology, pedagogy, and content before opening any platform. • Specify the chatbot's identity in detail. Richly specified personas produce consistent, believable interactions; vague ones produce generic outputs. The attributes worth specifying depend on the persona type. Example (simulated patient): demographics, presenting concern, personality, cultural context, and communication style; a coaching, assessor, or nonclinical persona would specify a different set of attributes.
2	Configuring the chatbot	System prompts, disclosure, and feedback	• Craft the system prompt (variously labeled instructions, system prompt, or custom instructions across platforms) to specify the role, interaction rules, prohibited behaviors, and tone. This is the educator's most consequential technical decision. • Decide how much the chatbot should volunteer vs withhold. In cases in which the persona simulates an information source, configure graduated disclosure so that details emerge incrementally and only in response to direct questions rather than all at once. Example: a simulated patient reveals symptoms only when asked, mirroring an authentic clinical encounter; a simulated client or stakeholder behaves analogously. • Build in structured feedback from the outset, anchored to learning objectives. Use a consistent framework (eg, "What went well / What to improve / Try again") and include an end-of-session reflective summary.

XSL•FO
RenderX

Section number	Section name	Focus	Design decisions covered
3	Refining the experience	Learner calibration, voice, and knowledge	• Calibrate vocabulary, scaffolding level, response length, and cultural register to match the target learner. Voice-enabled chatbots require additional brevity and single-question turns. • Use voice mode deliberately: it increases ecological validity for communication skill training but demands specific design constraints and institutional access verification. • Upload concise, structured knowledge documents (guidelines, rubrics, and case files) to ground responses in curriculum-specific content rather than generic model outputs.
4	Adding realism and testing	Avatars, multimodal elements, and pilot-testing	• Incorporate AI-generated avatar videos to add visual and emotional authenticity. Generate personas from descriptive text prompts; expect and budget for multiple iterations to achieve satisfactory facial expression and demographic representation. • Pilot-test rigorously before deployment: subject matter expert review for content accuracy, then learner testing under realistic conditions. Refine failure modes (overdisclosure, excessive length, and persona drift) by adjusting the system prompt.
5	Embedding and sustaining the tool	Curriculum integration and ethics	• Situate the chatbot within a planned instructional sequence—after foundational teaching and before or alongside supervised real-world encounters—and map explicitly to program-level outcomes (the graduate competencies the overall curriculum is designed to achieve). • Engage proactively with ethical responsibilities: transparency with learners, data privacy, algorithmic bias, overreliance on AI outputs, and ongoing monitoring after platform updates. Apply principles of fairness, accountability, and harm prevention as a standing framework.

[a]TPACK: technological pedagogical content knowledge.

This paper is a tutorial, not an evaluation study: its contribution is a practical, theoretically informed design workflow. Two locally developed initiatives, the Virtual Integrated Patient (VIP) and the Depression Avatars project, are drawn on throughout as illustrative implementation examples that motivated specific design decisions. They are presented as feasibility and acceptability experiences, not as evidence of educational

XSL•FO
RenderX

effectiveness; the empirical claims should be read as illustrative rather than confirmatory.

Section 1: Before You Build—Defining the Purpose, Learner, and Persona

Defining Your Educational Purpose

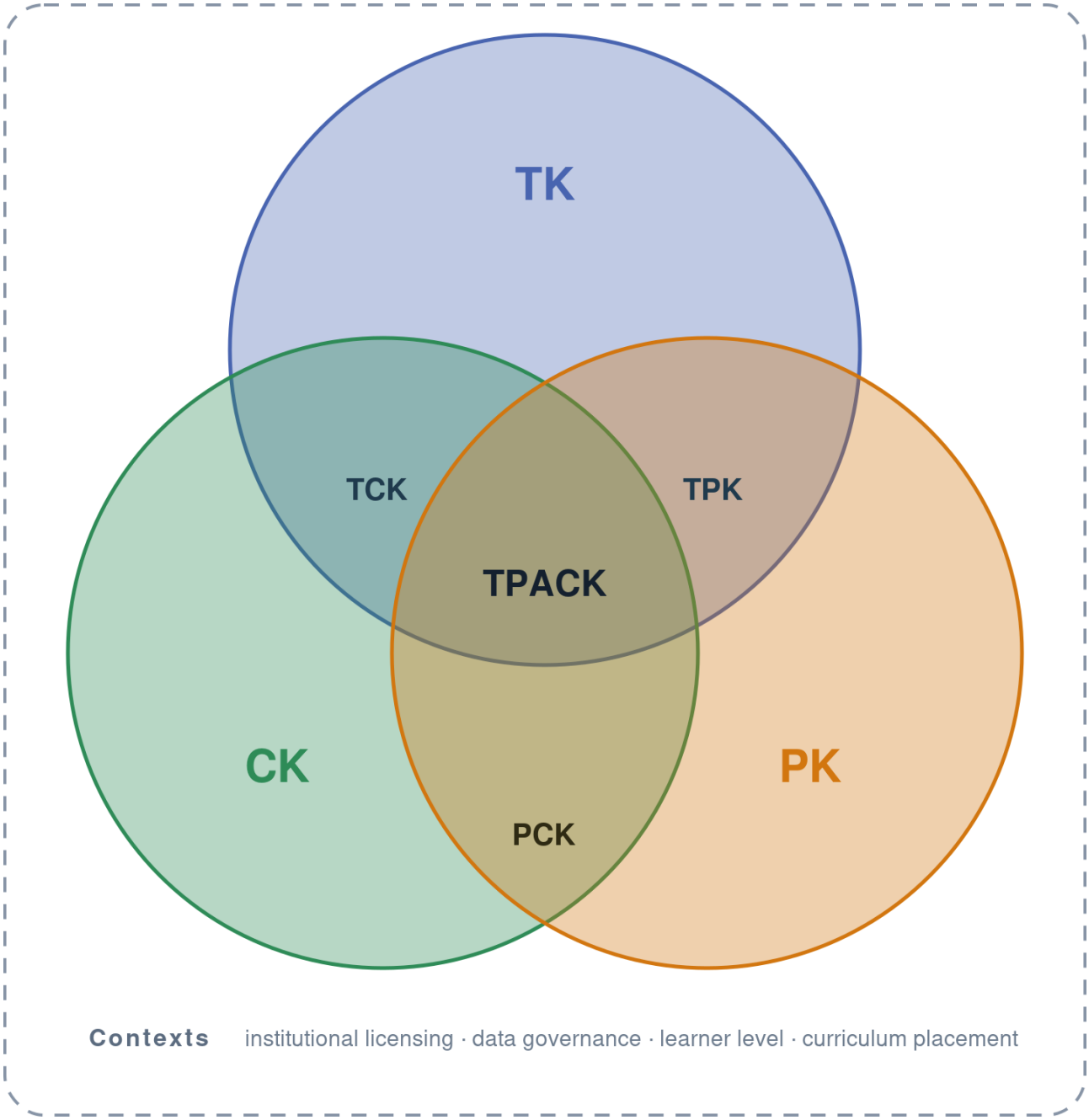
The most consequential decision in chatbot design is made before any tool is opened: clearly defining what the chatbot is meant to achieve. Health educators should begin by articulating the specific learning outcomes the chatbot will address, the learner group it will serve, and the stage of learning at which it will be deployed (preclinical, clinical, or postgraduate).

Chatbots serve distinct educational functions (formative assessment, deliberate practice, reflective debriefing, or knowledge retrieval), and the design choices for each differ substantially. A chatbot intended to simulate a psychiatric history-taking encounter requires a very different configuration from one designed to quiz students on pharmacological mechanisms. Conflating these purposes in a single tool frequently results in a chatbot that serves neither well. The same principle applies across disciplines: a law educator designing a client interview simulation faces analogous choices about scope and function.

The technological pedagogical content knowledge (TPACK) framework offers a useful heuristic at this stage [3]. Effective technology integration requires alignment among technological capability, pedagogical approach, and disciplinary content. Before building, the educator should be able to articulate clearly how the chatbot's technological features support the chosen pedagogical strategy and reinforce the specific content domain (Figure 1).

XSL•FO
RenderX

Figure 1. Application of the technological pedagogical content knowledge (TPACK) framework to no-code chatbot design.



While the TPACK framework explains what an educator must align, it does not by itself explain why the specific design choices in this workflow promote learning. Three established mechanisms provide that rationale and recur throughout the sections that follow. First, deliberate practice holds that competence develops through repeated, goal-directed effort on focused tasks with immediate feedback and progressive challenge [4]; the graduated disclosure and rehearsal properties of a well-designed chatbot (section 2) create exactly such focused, repeatable practice with consequences that are pedagogically informative rather than professionally harmful. Second, self-regulated learning describes how learners set goals, monitor their own performance, and adjust their approach [5]; structured, criterion-referenced feedback (section 2) supports this self-monitoring far more effectively than generic encouragement, which is why feedback is anchored to explicit objectives rather than offered impressionistically. Third, the simulation-based learning literature shows that technology-enhanced simulation improves knowledge and skill outcomes when it is integrated into a curriculum and used to supplement—not replace—authentic experience [6]; this is the basis for situating the chatbot within a planned instructional sequence (section 5) rather than deploying it in isolation. These mechanisms, together with the formative feedback evidence already noted, are what make the workflow pedagogically grounded rather than merely technically convenient.

Defining the Chatbot's Persona

Once the educational purpose is established, the chatbot requires a persona: a clearly specified identity that governs how it presents itself and communicates with learners. In health professions education, this most commonly takes the form of a

XSL•FO
RenderX

simulated patient, a clinical tutor, a coaching companion, or an assessment examiner. In other disciplines, the persona might be a simulated client, a historical figure, a debating opponent, or a community stakeholder.

Persona specification should be detailed and explicit. For a virtual patient, this includes demographic information (age, gender, cultural background, occupation, and living situation), presenting concern, relevant medical and social history, personality characteristics, and communication style. In the Depression Avatars project, patient profiles were meticulously developed by psychiatric experts to ensure accurate representation of conditions, including adjustment disorder, dysthymia, major depressive disorder, panic disorder, and somatic symptom disorder, and scripts were localized to reflect Singapore's cultural and health care context. This level of specificity produced clinically believable interactions and is directly traceable to the project's positive learner evaluations.

The level of detail in persona design directly affects the authenticity and consistency of the chatbot's responses. Vague personas produce generic outputs; richly specified ones produce interactions that challenge learners to think with contextual nuance regardless of the platform used.

Section 2: Configuring the Chatbot—System Prompts, Disclosure, and Feedback

Writing the System Prompt

In no-code LLM platforms, the system prompt is the primary mechanism through which an educator shapes chatbot behavior. It functions as standing instructions governing every learner interaction. This field is labeled differently across platforms (*Instructions* in ChatGPT's custom GPT builder, *System prompt* in Claude Projects, and *Custom instructions* in Microsoft Copilot), but the function is identical: it tells the model who it is, what it knows, how it should behave, and what it must not do. Despite its centrality, it is frequently underestimated by first-time designers.

A robust system prompt for a clinical simulation chatbot should address at minimum (1) the role the chatbot will adopt; (2) the specific knowledge base it should draw upon; (3) interaction conventions, such as one question at a time and no unsolicited disclosures; (4) what the chatbot should not do; and (5) tone and language register.

An example instruction set for a simulated psychiatric patient might be as follows: "You will act as a patient based on the provided script. Your responses should be realistic, casual, and natural. Avoid volunteering information unless asked. Do not reveal pain or mood scale ratings unless explicitly requested. Use emotional expressions appropriate to the patient's condition" [7]. Such specificity constrains the model toward clinically plausible, educationally productive responses and minimizes the risk of inappropriate information being offered spontaneously.

Designing Graduated Information Disclosure

A common failure mode in novice chatbot design is configuring the system to behave as a cooperative, fully disclosing informant—a pattern that bears no resemblance to real clinical encounters and offers little educational challenge. In practice, patients disclose symptoms and history gradually, often require prompting, and do not volunteer information simply because a clinician is present.

Instructions should specify that the chatbot should reveal information incrementally, respond only to the question asked, and require the learner to exercise appropriate clinical communication techniques to elicit further details. Some patient personas may be appropriately designed as hesitant, anxious, or evasive, reflecting the genuine heterogeneity of patients encountered in practice. This principle applies equally beyond health care: a simulated legal client should not immediately disclose all relevant facts; an engineering stakeholder should not pre-empt a student's needs analysis.

This design principle has direct pedagogical value: it rewards learners who ask open-ended, empathic, and systematically thorough questions and flags, through its silence, the omissions made by those who do not. The chatbot functions as a low-stakes rehearsal environment in which the consequences of incomplete questioning are pedagogically informative rather than professionally harmful [6].

Building in Structured Feedback

Formative feedback is one of the strongest drivers of learning in health professions education [8], and a chatbot without a feedback function is a missed pedagogical opportunity. Feedback should be built into the instruction set from the outset rather than added as an afterthought.

Structured feedback frameworks such as the "what went well / what could be improved / try again" format provide a consistent and actionable response to learner performance. Critically, the feedback mechanism should be anchored to the specific learning objectives defined in section 1. If the objective is to assess the completeness of a psychiatric history, the chatbot's feedback should explicitly address the components of a psychiatric history, not offer generic encouragement. This principle holds in any discipline: feedback should be criterion referenced, not impressionistic.

Effective feedback instructions might be as follows: "After each learner response, provide structured feedback identifying one specific strength, one area for improvement, and an invitation to attempt the response again. Do not reveal the model answer unless the learner explicitly requests it after two unsuccessful attempts. Use a supportive, non-judgmental tone and normalise mistakes as part of learning." An end-of-session summary in which the chatbot identifies patterns across the full interaction further enhances the reflective value of the encounter. A recording demonstrating this structured feedback mechanism in practice is provided in Multimedia Appendix 2.

XSL•FO
RenderX

Section 3: Refining the Experience—Learner Calibration, Voice, and Knowledge

Calibrating Communication Style for Your Learner Group

Health professions learners span a wide range of experience levels, from preclinical students encountering patient communication for the first time to postgraduate trainees refining advanced clinical reasoning. A chatbot calibrated for a preclinical medical student using accessible language, offering frequent encouragement, and providing scaffolding may be inappropriate for a postgraduate trainee who requires diagnostic challenge. The same calibration imperative applies in any educational context.

Communication style calibration includes adjusting vocabulary, response length, degree of scaffolding, tolerance for incorrect responses before providing guidance, and the nature of prompts used to advance the interaction. For voice-enabled chatbots in particular, brevity is essential: voice interactions cannot replicate the visual scanning afforded by text, and lengthy responses overwhelm rather than inform.

Cultural calibration is equally important. Health communication norms vary significantly across cultures and health care systems. The Depression Avatars project deliberately localized patient scripts and speech patterns to reflect Singapore's cultural and linguistic context. Educators should attend carefully to linguistic register and cultural assumptions embedded in patient personas before deploying to learners from different backgrounds.

Using the Voice Function Deliberately

Many contemporary GenAI platforms offer voice interaction capability, enabling learners to speak to the chatbot and receive spoken responses. This substantially increases the ecological validity of clinical simulation, particularly for communication skill training, by requiring learners to engage in spoken, real-time dialogue rather than compose written text. In disciplines in which oral communication is central to professional practice, voice mode offers an authenticity that text cannot replicate.

Voice mode demands specific design constraints: responses must be short and conversational, lists and structured summaries are disorienting when spoken aloud, and conversation starters should be provided to reduce cognitive load at the start of an interaction. Educators must also verify institutional access: microphone permissions, browser compatibility, and network reliability vary significantly across learning environments and should be tested before voice mode is designated a required activity. An illustrative audio recording of a learner interacting with a voice-enabled chatbot is available as Multimedia Appendix 3.

Supplementing With Curated Knowledge Documents

Most no-code GenAI platforms allow educators to upload supplementary documents (course materials, diagnostic criteria, clinical guidelines, rubrics, or patient case files) as a knowledge base from which the chatbot draws during interactions. This grounds the chatbot's responses in curriculum-specific content rather than generic model outputs and enables alignment between chatbot feedback and formal assessment criteria.

Documents uploaded to the knowledge base should be concise, clearly structured, and directly relevant to the chatbot's purpose. Uploading large, unstructured documents may degrade response quality. Educators should test chatbot outputs following each knowledge upload to verify that the intended material is appropriately influencing responses.

Section 4: Adding Realism and Testing—Avatars, Multimodal Elements, and Pilot-Testing

Adding Multimodal Elements to Increase Realism

Text-based chatbots are effective but limited in their ability to simulate the visual and emotional dimensions of professional encounters. Incorporating AI-generated avatar videos (photorealistic visual representations animated to match a generated script) can significantly enhance the perceived authenticity of simulated interactions. In the Depression Avatars project, patient scripts generated using GPT-4 were combined with avatar videos produced using the D-ID software, with facial expressions and speech synchronized to portray patients presenting with low mood and chest tightness. Student responses were positive, with most rating the avatar-based learning as more educationally useful than traditional formats [9]. A sample AI avatar patient video produced using these tools is available as Multimedia Appendix 4.

Educators can generate a visual patient persona from a descriptive text prompt alone. A prompt such as "A 50-year-old woman sitting in a clinic chair, wearing plain clothing, appearing slightly unkempt, hands folded in her lap, visibly anxious" can generate a photorealistic avatar that is then animated with the chatbot's script. This approach combines the interactivity of the conversational chatbot with the visual richness of a video-based case study without specialist technical knowledge.

Educators should budget for multiple iterations before reaching a satisfactory output. The Depression Avatars project required several rounds of prompt refinement to achieve acceptable facial synchronization and emotional expressivity. Emotional expression and body language remain current limitations of avatar technology, but the investment in iterative refinement yields material improvement in the learning experience's perceived authenticity.

Pilot-Testing Rigorously Before Deployment

A chatbot that has not been tested before deployment risks exposing learners to factually inaccurate information, inappropriate responses, or unreliable interactions—all of which damage learner trust and educational outcomes. Predeployment testing is not optional regardless of the platform or discipline.

Pilot-testing should include structured review by a subject matter expert assessing content accuracy across a range of likely learner inputs. In the Depression Avatars project, psychiatric experts

XSL•FO
RenderX

reviewed AI-generated patient profiles and scripts before student exposure to ensure clinical accuracy and cultural appropriateness. Expert review should be followed by testing with representative learners under realistic conditions, including voice mode if applicable.

Common failure modes to anticipate and test for include excessive unprompted disclosure of information, inaccurate disciplinary content, feedback that is too vague or poorly anchored to learning objectives, and persona inconsistency across extended interactions. Most failure modes can be addressed by adjusting the system prompt without platform-level changes. Simple additions such as "Do not volunteer information the learner has not asked for" or "Limit all responses to two to three sentences unless the learner explicitly asks for more" meaningfully improve output quality.

Section 5: Embedding and Sustaining the Tool—Curriculum Integration and Ethics

Situating the Chatbot Within a Broader Curriculum Framework

A persistent risk of educational technology adoption is deployment in isolation from the broader curriculum, reducing pedagogical impact and limiting transfer of learning. AI chatbots are most effective when situated within a planned instructional sequence that includes prelearning, supervised practical exposure, and structured reflection whether in a medical school, a law faculty, or a teacher education program.

Chatbots designed for deliberate practice function best when used after foundational content has been introduced through lectures or small-group teaching and before or alongside supervised real-world encounters. They are not a substitute for authentic professional experience; they function as a preparatory rehearsal environment that builds procedural fluency and reduces anxiety prior to higher-stakes settings, much as flight simulators prepare commercial pilots without replacing actual flight hours [4]. In the Depression Avatars project, avatar-based simulations were explicitly designed to supplement, not replace, psychiatric teaching and real patient contact.

Assessment integration strengthens the chatbot's educational value further. Where chatbots support formative assessment, learner performance data (completeness of questioning, quality of reasoning, and responsiveness to feedback) may inform subsequent teaching priorities. Educators should map chatbot-based activities to program outcomes and establish a plan for monitoring learner progress over time.

Engaging Proactively With Ethical Responsibilities

The deployment of AI chatbots in health professions education raises ethical responsibilities that must be addressed explicitly rather than deferred. These include data privacy, academic integrity, algorithmic bias, transparency with learners about AI involvement, and the risk of overreliance on AI-generated information. These concerns are not unique to health education; they are systemic issues wherever AI tools are deployed in learning environments.

Learners should be clearly informed that they are interacting with an AI tool, that its outputs may contain inaccuracies, and that professional decision-making must always be grounded in verified evidence and supervised practice. Safety messaging should be built into the system prompt directly: instructions such as "If uncertain, acknowledge uncertainty and direct the learner to consult a trusted clinical reference" reduce the risk of the chatbot being used as an authoritative source for consequential decisions.

Educators should attend to demographic and cultural biases embedded in the underlying LLM that may produce personas or scenarios that are unrepresentative or culturally insensitive. Ongoing monitoring of chatbot outputs is essential: AI systems evolve with platform updates, and a chatbot that performed appropriately at launch may require re-evaluation following changes to the underlying model. The principles of fairness, transparency, accountability, privacy, harm prevention, and trust building should serve as a standing framework for all AI-enhanced educational design [10].

When Not to Use a Chatbot and Key Risks

No-code chatbots are not appropriate for every educational task, and being explicit about their limits is part of responsible design. They are poorly suited to high-stakes summative assessment, where the lack of reproducibility, auditability, and validated scoring makes them unfair as a basis for consequential decisions. They are a weak substitute for any competency whose core is authentic human interaction (breaking bad news, disclosing error, or responding to a distressed or angry patient) because current models do not reliably reproduce emotional variability, nonverbal cues, or sustained interpersonal tension; learners in the VIP evaluation made exactly this point (J. Sng, unpublished data, 2026). They should not be used in any workflow that requires entering patient-identifiable or otherwise confidential data into a consumer platform, and they are inadvisable where no faculty oversight, monitoring, or governance is available. Finally, chatbots carry particular risk for novice learners who lack the baseline knowledge to detect model error, where confidently stated inaccuracies may reinforce misconceptions rather than correct them.

The principal risks follow from these limits. Generative models can hallucinate plausible but incorrect clinical content; they can foster automation bias and overreliance that erodes independent reasoning; they raise data privacy exposure when learners paste real material; they can encode demographic or cultural bias into generated personas; and unequal access to paid tiers or institutional licenses can create equity gaps between learners. Several of these risks are not yet well quantified in health professions education, which is itself a reason to frame current tools as promising rather than proven and monitor them in use.

A Practical Governance Checklist and the Conditions for Sustainable Use

General appeals to "privacy" and "transparency" are difficult to act on, so educators benefit from a concrete governance

XSL•FO
RenderX

checklist. Before deployment, a minimum standard is to (1) define what data must never be entered into a public platform (real patient identifiers, uploaded clinical records, and any institution confidential material) and state this rule to learners in writing; (2) verify the platform's data retention and training settings and prefer institutionally licensed tools with enterprise privacy terms where available; (3) assign a named owner responsible for monitoring outputs, retesting after model updates, and retiring the tool if quality degrades; (4) establish accountability for AI-generated error—the supervising educator, not the model, remains responsible for what learners are exposed to; and (5) train learners explicitly to interrogate AI outputs rather than accept them.

This last point reframes chatbot use as an opportunity rather than only a hazard. A central aim of deployment should be to move learners from uncritical acceptance toward verification and meta-cognitive engagement, treating the chatbot as a fallible interlocutor to be questioned, cross-checked against trusted references, and corrected [11]. Cultivating this critical AI literacy is arguably as valuable as the domain practice the chatbot provides, and it is a competency that learners will carry into clinical environments increasingly saturated with AI tools.

Sustainability, finally, depends on conditions beyond any individual educator's enthusiasm. No-code tools lower the technical barrier, but they do not remove the need for educator training, institutional licensing and platform access, data governance approval, curriculum alignment, recognition of the workload involved in building and maintaining tools, and ongoing quality assurance; designing and supervising AI-mediated learning redistributes rather than simply reduces educators' cognitive labor, with implications for professional judgment and reflective practice [12]. Where these enabling conditions are absent, even well-designed chatbots tend to remain isolated pilots rather than durable curriculum components. Educators are therefore advised to secure institutional support early, particularly given the high-stakes nature of professional preparation in health disciplines.

Illustrative Implementation Examples

The workflow presented in this tutorial was developed through iterative experience building and evaluating 2 chatbot-based tools at the Yong Loo Lin School of Medicine, National University of Singapore.

The first, the VIP, is a free-text conversational platform in which an LLM is configured to simulate patients presenting with a range of clinical conditions for history-taking practice. Qualitative interviews with medical students (n=8) identified VIP as a flexible, self-directed tool valued for applying knowledge in simulated patient scenarios. Students noted limited emotional realism and the absence of visual cues as constraints on the depth of the experience and consistently emphasized that VIP functioned best as a supplement to, not a replacement for, traditional teaching and real patient contact (J. Sng, unpublished data, 2026).

The second initiative, the Depression Avatars project, sought to address these limitations by combining GPT-4–generated clinical scripts with photorealistic AI avatar videos produced using the D-ID software. Patient profiles spanning low mood (adjustment disorder, dysthymia, and major depressive disorder) and chest tightness (adjustment disorder, panic disorder, and somatic symptom disorder) presentations were developed by psychiatric experts, localized to Singapore's clinical context, and presented to medical students as diagnostic exercises. Early feedback from medical educators and students suggested that the approach was engaging and acceptable, and a student pilot to gather quantitative and qualitative feedback was subsequently conducted; formal outcome data are not yet available [9]. These observations are reported as feasibility and acceptability experiences, not as evidence of learning gains.

The findings from both initiatives directly informed the 5-section workflow: VIP's evaluation shaped the sections on graduated disclosure and curriculum supplementation; the Depression Avatars project informed the sections on persona specificity, cultural localization, avatar iteration, and expert-led pilot-testing. Rigorous evaluation using controlled study designs, validated outcome measures, and multi-institutional replication remains necessary before definitive conclusions about learning efficacy can be drawn.

Conclusions

No-code AI chatbots represent a genuine opportunity to extend the reach and quality of experiential learning in health professions education, particularly where patient access is limited and contact time is constrained. The 5-section workflow presented in this paper, from defining the purpose and persona to configuring the interaction and refining for the learner to adding realism, testing rigorously, and sustaining the tool ethically, enables educators to move from intention to deployed chatbot with pedagogical rigor regardless of technical background.

While the primary audience is health professions educators, the workflow is transferable across higher education wherever interactive, scenario-based learning is valued. The tools listed in Table 1 will change; the design decisions summarized in Table 2 will not. As the AI landscape continues to evolve, sound instructional design remains the most durable foundation for effective chatbot development.

Funding

The authors gratefully acknowledge the financial support provided by the Wong Hock Boon Society for CXLG and EW at the Yong Loo Lin School of Medicine, National University of Singapore.

XSL•FO
RenderX

Conflicts of Interest

None declared.

Multimedia Appendix 1
Step-by-step educator guide: building text-based chatbots, voice-enabled chatbots, and AI avatar patient videos without coding. Includes system prompt templates, platform selector table, troubleshooting guides, and expert review checklist.
[DOCX File, 113 KB - mededu_v12i1e103399_app1.docx]

Multimedia Appendix 2
Audio recording of a medical student practicing psychiatric history taking using a voice-enabled custom generative pretrained transformer chatbot with structured AI-generated feedback. The recording illustrates the “what went well / what to improve / try again” feedback mechanism described in section 2 of this guide. Student consent was obtained prior to recording. Identifying information has been removed.
[MP4 File, 30833 KB - mededu_v12i1e103399_app2.mp4]

Multimedia Appendix 3
Audio recording of a medical student practicing psychiatric history taking using a voice-enabled custom generative pretrained transformer chatbot. Student consent was obtained prior to recording. Identifying information has been removed.
[MP4 File, 9954 KB - mededu_v12i1e103399_app3.mp4]

Multimedia Appendix 4
Sample AI-generated avatar video produced using HeyGen depicting a simulated psychiatric patient presenting with low mood in a clinical consultation setting. The avatar was generated from a text-based patient profile and a clinician-reviewed script written using a large language model. The “AI” watermark (lower left) indicates AI-generated content and should be disclosed to learners prior to viewing.
[MP4 File, 6795 KB - mededu_v12i1e103399_app4.mp4]

References

1. Ng FY, Thirunavukarasu AJ, Cheng H, et al. Artificial intelligence education: an evidence-based medicine approach for consumers, translators, and developers. Cell Rep Med 2023 Oct 17;4(10):101230. [doi: 10.1016/j.xcrm.2023.101230] [Medline: 37852174]
2. Martin K. The evolving role of educators in the age of AI. Katie L Martin. 2023. URL: https://katielmartin.com/2023/06/20/the-evolving-role-of-educators-in-the-age-of-ai/ [accessed 2026-07-25]
3. Mishra P, Koehler MJ. Technological pedagogical content knowledge: a framework for teacher knowledge. Teach Coll Rec 2006 Jun;108(6):1017-1054. [doi: 10.1177/016146810610800610]
4. Ericsson KA. Deliberate practice and the acquisition and maintenance of expert performance in medicine and related domains. Acad Med 2004 Oct;79(10 Suppl):S70-S81. [doi: 10.1097/00001888-200410001-00022] [Medline: 15383395]
5. Zimmerman BJ. Becoming a self-regulated learner: an overview. Theory Pract 2002;41(2):64-70. [doi: 10.1207/s15430421tip4102_2]
6. Cook DA, Hatala R, Brydges R, et al. Technology-enhanced simulation for health professions education: a systematic review and meta-analysis. JAMA 2011 Sep 7;306(9):978-988. [doi: 10.1001/jama.2011.1234] [Medline: 21900138]
7. Wang E. MindMate: an artificial intelligence (AI) chatbot for enhancing patient history taking and diagnosis practice in psychiatric education. Presented at: Asia Pacific Medical Education Conference (APMEC); Jan 13-18, 2025 URL: https://medicine.nus.edu.sg/cenmed/sites/apmec2025/documents/Free_Communications_5_YS.pdf [accessed 2026-08-11]
8. Hattie J, Timperley H. The power of feedback. Rev Educ Res 2007;77(1):81-112. [doi: 10.3102/003465430298487]
9. Goh CX, Wang E, SNG JC. The usage of generative artificial intelligence (AI) avatars for psychiatric medical education. In: Higher Education Conference in Singapore (HECS) 2024: National University of Singapore; 2024. URL: https://blog.nus.edu.sg/hecs/hecs2024-jcgsng-et-al/ [accessed 2026-08-11]
10. AI competency framework for students. UNESCO. 2024. URL: https://www.unesco.org/en/articles/ai-competency-framework-students [accessed 2026-08-07]
11. Rind IA, Bhatti J, Chellappan K. From blind trust to critical inquiry: epistemic beliefs and student engagement with ChatGPT in higher education. Jurnal Kejuruteraan 2026;38(1):01-14. [doi: 10.17576/jkukm-2026-38(1)-01]
12. Rind IA. Conceptualizing the impact of AI on teacher knowledge and expertise: a cognitive load perspective. Educ Sci 2026;16(1):57. [doi: 10.3390/educsci16010057]

Abbreviations

GenAI: generative AI

XSL•FO
RenderX

TPACK: technological pedagogical content knowledge
VIP: Virtual Integrated Patient

Edited by S Tsuei; submitted 02.Jun.2026; peer-reviewed by B Cynthia, IA Rind; revised version received 27.Jul.2026; accepted 27.Jul.2026; published 14.Aug.2026.

Please cite as:
Goh CXL, Wang E, Kong JSM, Sng JCG
Empowering Educators to Build AI Chatbots in Health Professions Education: Tutorial on a No-Code Design Workflow
JMIR Med Educ 2026;12:e103399
URL: https://mededu.jmir.org/2026/1/e103399
doi:10.2196/103399

© Charlene Xing Le Goh, Eugene Wang, Juanita S M Kong, Judy C G Sng. Originally published in JMIR Medical Education (https://mededu.jmir.org), 14.Aug.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Education, is properly cited. The complete bibliographic information, a link to the original publication on https://mededu.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Enhancing Pathology Knowledge for Biomedical Scientists: Pilot Cohort Study of a Comprehensive Self-Paced Tutorial for Bridging Fundamental Concepts With Translational Applications

Jennifer Sells[1], EdD; Leticia C Clemente[2,3], MD; David P Farris[4], MSIS; Mary L Sizemore[1], PhD; Ignacio I Wistuba[2,3], MD; Rama Soundararajan[2,3*], PhD; Maria G Raso[2,3*], MD

[1]Department of Education, Development & Innovation, The University of Texas MD Anderson Cancer Center, Houston, TX, United States

[2]Department of Translational Molecular Pathology, The University of Texas MD Anderson Cancer Center, 2130 W Holcombe Blvd., Houston, TX, United States

[3]Interdisciplinary Translational Education and Research Training (ITERT) Core, The University of Texas MD Anderson Cancer Center, Houston, Texas, United States

[4]Research Medical Library, The University of Texas MD Anderson Cancer Center, Houston, TX, United States

[*]these authors contributed equally

Corresponding Author:
Maria G Raso, MD
Department of Translational Molecular Pathology, The University of Texas MD Anderson Cancer Center, 2130 W Holcombe Blvd., Houston, TX, United States

Abstract

Background: Knowledge of pathology is integral to numerous health care and research disciplines, besides routine laboratory-medicine operations. Despite its importance, there remains a significant gap in structured educational programs designed to provide a fundamental understanding of the pathologic basis of disease for an interdisciplinary body of research and health care professionals. This lack of generalized pathology training may contribute to a broader knowledge deficit in its translational applications.

Objective: To address this unmet educational need, we developed a distinctive pathology training course tailored to cater to a broad array of scientists in basic and translational science.

Methods: Overall, 29 participants from the University of Texas MD Anderson Cancer Center and the Texas Medical Center completed the "Fundamental Pathology for Basic Scientists: A Self-Paced Online Certificate Course" pilot. The course, comprising 24 instructional modules, was delivered over the span of 6 months in an online, asynchronous, and readily-accessible CANVAS learning management format. The educational objectives for each module were organized around organ- and organ-system-specific content, aiming to (1) incorporate the pathological basis of diseases, (2) distinguish clinicopathologic correlations, and (3) recognize the impact of molecular- and translational pathology and associated state-of-the-art diagnostic technology. Participants were voluntarily surveyed both before and after the course to assess its effectiveness.

Results: The preentry survey revealed that participants represented 16 academic departments (with widely varied training backgrounds, career stages, and job descriptions within the biomedical research enterprise), with 76% (22/29) reporting no prior basic pathology training. The exit survey highlighted positive feedback on the course structure, instructors, delivery method, and lecture content. Overall, 100% (29/29) of learners completed the objective knowledge checks for each of the 24 modules. At least 89% (26/29) of learners concurred that each of the 24 modules had delivered its learning objectives. A total of 94% (27/29) of learners concurred that the course enriched their understanding of basic pathology. Additionally, 79% (23/29) of participants reported applying or being able to apply what they had learned to their current roles, and 94% (15/16) of postcourse survey respondents confirmed longer-term benefits 1 year after program completion (increased confidence in addressing projects involving pathology).

Conclusions: The initial outcomes of this pilot course suggest that the asynchronous delivery method was well-suited to participants' workloads, and that the content was valuable for basic- and translational-scientists from varied training backgrounds seeking to deepen their understanding of fundamental pathology in the context of biomedical research. This work establishes a model for scalable, technology-enabled pathology training adaptable to diverse learners within health care research centers.

(JMIR Med Educ 2026;12:e84903) doi:10.2196/84903

XSL•FO
RenderX

KEYWORDS

pathology education; continuing professional development; online learning; biomedical research training; translational medicine

Introduction

In biomedical research, it is essential for basic science scholars, graduate students, and postdoctoral fellows to not only master their research-specific disciplines, but also stay updated with advancements in interdisciplinary related fields that impact their mainstream research focus. One such interdisciplinary area is pathology, which plays a critical role in bridging the gap between basic research and its clinical applications. A broad understanding of this field is an indispensable asset for basic scientists and health care professionals alike. Insights into the pathological basis of disease are vital for uncovering mechanisms of disease progression and identifying potential therapeutic targets in the research laboratory. Similarly, professionals in laboratory medicine rely on pathology to diagnose and monitor diseases, interpret lab results, and contribute effectively to patient care. Although the need to build or update knowledge and skills in pathology is eminent for workers in health care research centers, many unfortunately either find themselves distanced from foundational pathology concepts over time or lack exposure to the latest developments in translational pathology and emerging technologies [1-3]. In addition to these challenges, there are a litany of unmet educational needs and preexisting or unsolved issues plaguing training programs in pathology, spanning decades from the 1980s to the present day: (1) limited resources for training research scientists in pathology, balancing emphasis on molecular vs general pathology topics, (2) lack of attraction to pathology as a dedicated career track for postgraduate students, and (3) attrition of trained pathologists (limited availability of qualified instructors) [4-8]. There is also a significant gap in the availability of local pathology courses designed for a broader range of professionals who have not pursued "formal" pathology training or work outside of traditional academic pathways, yet require a solid foundation in this field to enhance their everyday work. However, well-recognized training program scholars highlight that regardless of the job type, 70% of learning that directly impacts job performance occurs *on* the job, suggesting the dire need for locally accessible pathology learning courses catering to the career needs of an expansive learner population in health care research centers [9].

To directly address this educational gap, we engaged in the intentional development and delivery of a self-paced learning course to serve a broad breadth of learners in the health care workplace, ranging from graduate students and postdoctoral fellows in translational research to research assistants, technicians, basic scientists, histotechnologists, laboratory managers, quality assurance specialists, and data specialists, as well as research faculty. The "Fundamental Pathology for Basic Scientists" course was developed as part of the Interdisciplinary Translational Education and Research Training (ITERT) Core at the University of Texas (UT) MD Anderson Cancer Center (MD Anderson), a platform dedicated to advancing the educational and professional development of the next generation of leaders in translational research [10,11], in partnership with the Department of Education Practice & Workforce Advancement.

The curriculum was tailored to provide participants with a comprehensive awareness of traditional pathology concepts (Robbins & Kumar Basic Pathology, Eleventh Edition, the international gold-standard text) while integrating instruction on modern innovations in molecular pathology and methodologies that are pivotal to new-age translational research (instructors sharing relevant pathology research using state-of-the-art methodologies and instrumentation). The main purpose of this course was to offer a structured opportunity for participants to deepen their understanding of pathology in a research context, revisit core pathology principles, acquire updated knowledge in translational pathology, and, importantly, engage with interdisciplinary workforce professionals and network with a constellation of trained pathology mentors and instructors. This study describes the initial training outcomes of this pilot educational course.

Methods

Course Objectives

The primary goals for this course were to (1) address the nonavailability of workplace training in fundamental pathology (that is traditionally limited to degree-awarding academic programs) by creating a nondegree certificate course open to learners from different disciplines in biomedical research, (2) provide a structured learning rubric for understanding the pathological basis of disease in an organ-organ system format, (3) design flexible self-paced and asynchronous learning modules to serve full-time employees in the health care research enterprise, and (4) develop a support network (peer- and tiered-mentoring) for nonpathology health care professionals who indirectly interface with the pathology-domain.

Course Participants

Recruitment targeted researchers, staff scientists, and health professionals from across the Texas Medical Center in Houston, Texas, interested in completing a structured course on fundamental pathology for their professional development. The 29 individuals who completed all modules of the course encompassed 10 different professional roles across 16 different academic department affiliations enrolled in the course. The largest single-career stage represented was among 8 out of 29 participants (28%) being research postdoctoral fellows (Table S1 in Multimedia Appendix 1). Regarding the institutions of affiliation, 26 out of 29 participants (90%) were from MD Anderson, and 3 out of 29 participants (10%) were from UT Health Houston.

A learner analysis was conducted to gain an understanding of the participants' training backgrounds and learning objectives. This information was captured in a pre-course survey administered before participants began coursework. Following best practices for conducting a learner analysis, 5 essential

XSL•FO
RenderX

concepts were adhered to, in accordance with recommendations by Stefaniak [12]:

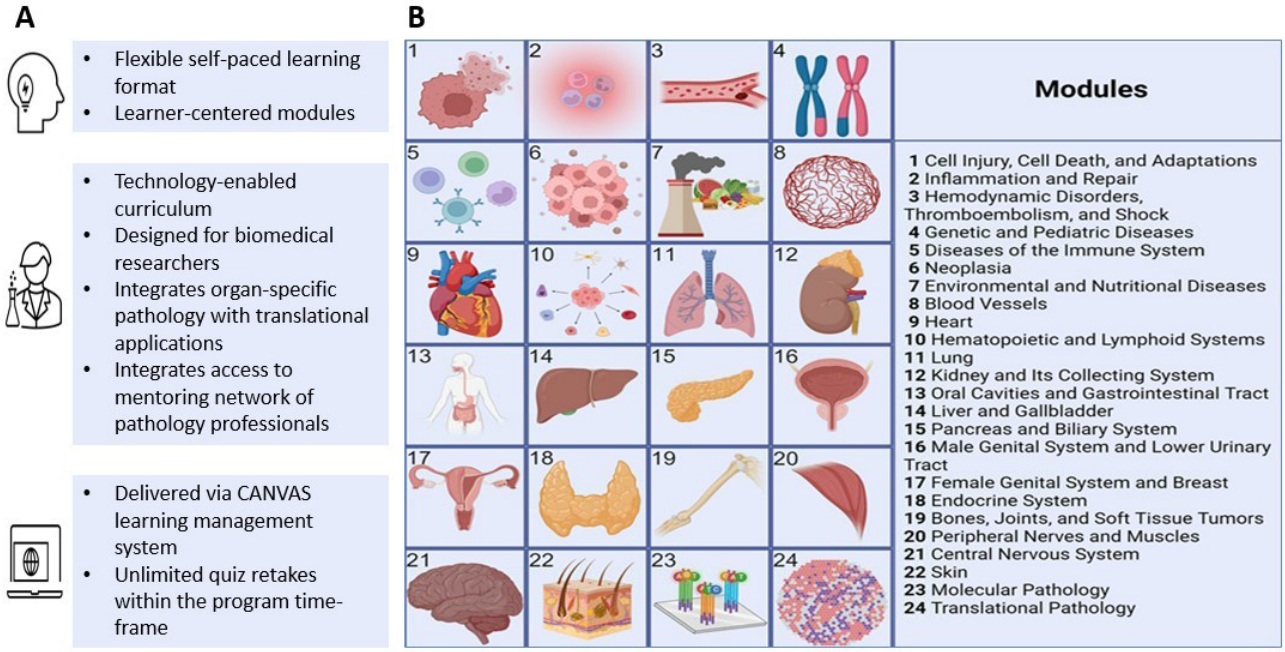
1. General characteristics
2. Prior pathology knowledge and prerequisite skills of participants
3. Predispositions pertaining to the topic of pathology
4. Attitude and motivation toward training
5. Characteristics of the group as a whole

This early analysis allowed for a comprehensive snapshot of learner demographics in terms of academic career stage, background experience in pathology, and instructional modules needing content tailoring.

Course Design, Curriculum, and Delivery

Three key factors were taken into consideration when designing this course: (1) the broad target audience of basic scientists from various departments and academic career stages, with diverse training backgrounds and work roles, (2) the need for participants to complete this course alongside their full-time or regular work duties, and (3) course accessibility for participants from across the Texas Medical Center in Houston, Texas. The CANVAS Learning Management System was selected for the asynchronous delivery of course materials (Figure 1A), which consisted of lecture videos, presentation files, supplementary reading material (research publications, techniques, and method aids), and quizzes.

Figure 1. Fundamental Pathology for Basic Scientists: Course structure highlights (A) Curriculum overview (B). Shown in B are the 24 instructional modules for this course. Image created in BioRender (Sells, J., 2025 [13]), licensed by authors with Academic account.



A total of 24 learning modules were organized, each focusing on topics of interconnecting physiopathology contents and clinicopathologic correlations with specific organ systems (Figure 1B). Traditional pathology topics were based on Robbins & Kumar Basic Pathology, Eleventh Edition, the international gold-standard text for pathological disease [14], and information from this book guided the selection of content for the majority of the modules (Modules 1 - 22). While we endeavored to adhere to the sequencing of chapters as in this well-established teaching resource that is recognized in the field, each lecture, however, was tailored to our specific audience with varying levels of pathology knowledge, and the course also included 2 additional modules covering molecular pathology and translational pathology, not addressed in [14]. Module 23 on molecular pathology covered the definition of molecular pathology, molecular basis of disease, molecular pathology tools, diagnostic molecular pathology, and tumor molecular profiling. Module 24 on translational pathology covered introduction to translational pathology, -omics research, genomics, metabolomics, and other omics for target identification and validation, biomarkers in oncology, genetics and molecular biomarkers, and artificial intelligence to improve diagnostic and prognostic efficacy. The inclusion of dedicated modules in molecular pathology and translational pathology was not intended as an additive novelty in isolation, but as an explicit curricular bridge connecting traditional histopathology with contemporary research methodologies that are central to modern biomedical research but typically absent from foundational pathology training.

Furthermore, we incorporated Elsevier's Evolve Resources for the Instructor-Led Course Robbins & Kumar Basic Pathology, Eleventh Edition [15], which allowed us to tailor the content and learning experience to the unique needs of each participant, ensuring broad accessibility and learner engagement. Additionally, instructors enhanced each module by sharing relevant recent literature to reveal the incorporation and use of modern methodologies and instrumentation for pathology work.

This 24-module pilot course was offered over a span of 6 months. One module was released sequentially in CANVAS each week and included an estimated 1 hour of recorded video instruction on the topic, supplemented with relevant reading material from current literature. Participants were allowed to complete the module at their convenience, which provided

XSL•FO
RenderX

considerable flexibility in accommodating their coursework into their ongoing work schedules and job responsibilities. Each module was created and taught by subject-matter experts from MD Anderson, and lecturers offered virtual office hours after every 4 modules to promote inquiry and interaction and follow-up on any additional learning resources the participants needed.

The professional roles of the instructors spanned the range from early-career assistant professors to associate professors, research scientists, and clinical pathology fellows to research postdoctoral fellows. All instructors were members of either the Pathology or Translational Molecular Pathology departments at MD Anderson and had received formal pathology training in the United States or internationally. Lecturers constructed a multiple-choice style quiz of 10 questions for each module, and learners were allowed multiple attempts to complete the quiz with a benchmark score of 80% for passing each module before they could progress to the next module. The 29 participants completed all 24 modules and were issued a certificate of completion at the end of the 6-month pilot course.

Course Evaluation

Three surveys were issued to obtain pertinent insight. All surveys were conducted using the institutional Qualtrics management software. The first survey was a pre-entry questionnaire administered before the course, intended to collect background information from participants on departmental affiliation, prior training in basic pathology, and application of pathology reports or basic pathology in their current jobs or roles.

The second survey for evaluation was the exit survey, administered to each participant after completing all 24 course modules en route to gaining their course certificate. Aligned with a prominent, evidence-based practice for evaluating training programs, the exit survey was framed according to the Kirkpatrick Four-Level Evaluation Model [9,16] that has persisted as a gold-standard to the present day. The survey was subdivided to flow through key evaluation points on course structure, lecture delivery, module content, learning and knowledge application, and concluded with a free-response section for additional feedback. Each of the 4 Kirkpatrick levels (ie, 1: reaction, 2: learning, 3: behavior, and 4: results) were sequentially featured (Table 1).

Table . Topics surveyed for course evaluation and their alignment with the Kirkpatrick model of training program evaluation.

Relevance to Kirkpatrick level	Topics surveyed
1 Reaction	Duration of the course, learning objectives of each module and organization of course content, use of the virtual office hours, delivery format, course balance with work duties, format, and delivery of quizzes
2 Learning	Pathology-specific knowledge and skills gained, quizzes, confidence in applying their newly learned skills
3 Behavior	Relevance to their current jobs or roles and/or career-development goals, enhancement of current work duties
4 Results	Specialized pathology course interest, long-term value

The majority of the exit survey items were strategically geared toward Kirkpatrick Levels 1 and 2, as these are areas of immediate and high interest to the launch of a new training program. This evaluation was essential to fine-tune and tailor the modules in real-time for optimizing learning and instructional design tailored to meet our primary goal: to provide a baseline interface where nonpathologists whose everyday work eventually impacts pathology could gain pathology awareness. Level 3 (behavior) and Level 4 (results) metrics were captured in an additional survey, 1-year post-program completion. In addition to these participant perception-based metrics, checkpoint quizzes embedded at the end of each module served to objectively measure the knowledge gain - participants had to score 80% or better in order to proceed to the next module.

We also assessed our evaluation against a well-established interdisciplinary, problem-centered approach to using educational technology for learning design and instructional systems development [17], the Spector approach. Spector posed 5 levels of evaluation on the application of educational technology for instructional design when constructing a course or program in any discipline. Applying the Spector scale to the evaluation of this pilot course, our key metrics are shown in Table 2.

XSL•FO
RenderX

Table . Course evaluation criteria breakdown in alignment with the Spector scale for instructional program evaluation.

Spector level	Criteria for good learning and instruction	Supporting course statistic
1	Learning goals and objectives are met (formative and summative)	Course requirement ensured a minimum of 80% passing score per module [all 29 of 29 participants (100%) passed each of the 24 modules to receive a certificate of completion]; at least 26 out of 29 participants (90%) confirmed that the course modules had met their stated learning objectives
2	Learning setting is useful, and content aligned with stated objectives	At least 26 out of 29 participants (90%) confirmed that the course delivery format was favorable; at least 26 out of 29 participants (90%) confirmed the course duration was appropriate and that the distribution of instructional content in the models was favorable and aligned with the stated objectives
3	Overall course experience is satisfactory	27 out of 29 participants (94%) confirmed they would recommend this course to a colleague
4	Program is sustainable	Future direction to probe (long-term course iteration evaluation is necessary to determine course sustainability).
5	No harm is done	At least 15 out of 16 post-course survey respondents (94%) concurred that their self-knowledge in fundamental pathology has increased because of participating in this course. Future direction to probe long-term technological impact (planned 3-year post-course participant follow-up). A second cohort will allow for more in-depth Level 5 determinations to be made.

The final survey was conducted 1 year after program-completion, to assess long-term impact of the course on participants' career and professional development, teaching abilities, and sustained confidence in basic pathology knowledge and skills.

Ethical Considerations

This study was reviewed and approved by the Institutional Review Board (IRB) at the UT MD Anderson Cancer Center (OHRP IRB Registration Number: IRB00000121; Reference ID# 2024 - 1381).

Results

Precourse Survey

Overall, 28 out of 29 participants (97%) who completed the course had previous biology training, and 21 out of the 29 participants (72%) had prior histology knowledge. Interestingly, while 23 out of 29 participants (79%) either used pathology reports or basic pathology concepts in daily work activities, only 24% of participants (7 out of 29) had previously received formal pathology training. Trainees (graduate students and postdoctoral fellows) only accounted for 9 out of 29 participants (31%), while 20 out of 29 participants (69%) were professionals in formal work capacities within the academic centers. A more detailed profile of the learners is provided in Table S1 in Multimedia Appendix 1.

Course Exit Survey (Immediate or Near-term Training Impact)

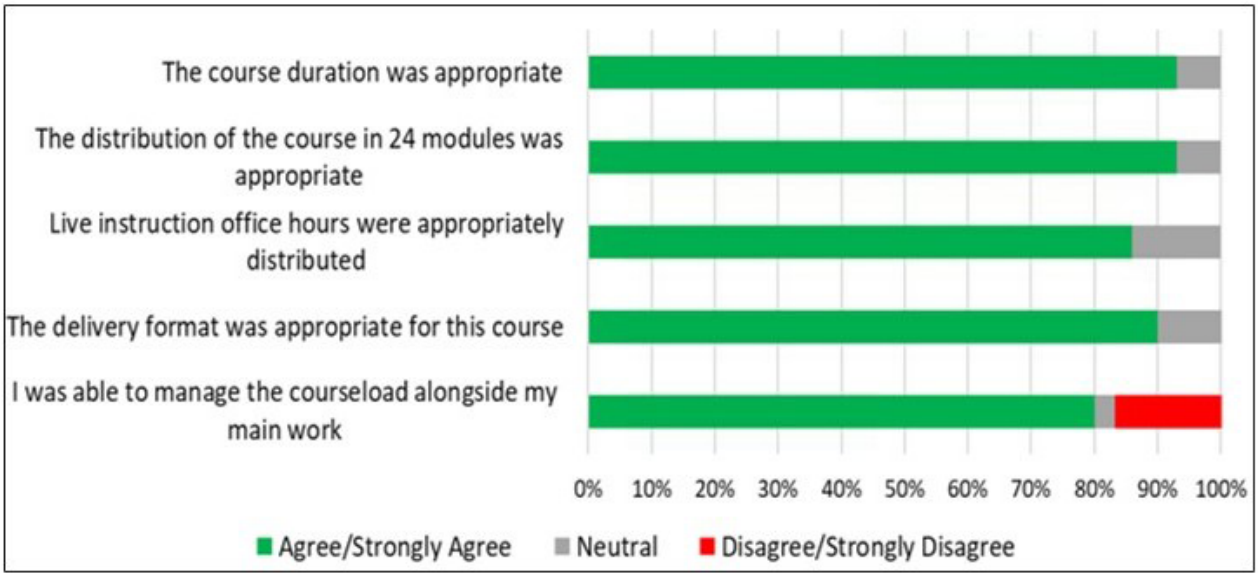
All 29 course participants who completed the course completed the exit survey as well. A 5-point Likert scale was used to assess participants' opinions on course structure, course content, and course lectures (Figure 2). For all questions, the scale anchors used were as follows: 1- strongly disagree, 2- disagree, 3- neutral, 4- agree, and 5- strongly agree. Since this was a small pilot study, the Likert anchors of agree or strongly agree were grouped together to enable ready visualization of positive responses, as they were the overwhelming majority of responses. Similarly, disagree or strongly disagree were grouped together to easily visualize negative responses since there were so few. At least 27 out of 29 participants (93%) concurred (agreed or strongly agreed) that the course duration and distribution of the course in 24 instructional modules was appropriate. When questioned about live instruction office hours, at least 25 out of 29 participants (86%) concurred that it was adequate and appropriately distributed. At least 26 out of 29 participants (90%) considered the course delivery format using lecture videos via CANVAS to be favorable, and at least 23 out of 29 participants (79%) reported being able to effectively manage the course requirements alongside their primary work.

The second section of the questionnaire surveyed participants' opinions on course content. For each module, 2 statements were explored: "I feel that the learning modules met the stated learning objectives," (*learning objectives met*) and "The lecturers of the course effectively covered the learning objectives in their lecture videos" (lecturers and videos covered objectives

XSL•FO
RenderX

effectively). The detailed assessments are shown in Table S2 in Multimedia Appendix 1. Any of the negative Likert responses received were attributed to just one individual's response of dissatisfaction within a given module. Overall, the feedback for both queries was strongly positive, with at least 89% of participants (at least 26 out of 29) concurring (agreeing or strongly agreeing) favorably with both statements for each of the 24 modules, underscoring the high quality of the instructional content, the superior teaching abilities of the instructors, and the strong alignment of the presented content with the stated learning objectives (Figure 3).

Figure 2. Fundamental Pathology for Basic Scientists: Data on participants' perceptions of course structure. A Likert scale assessment of agreement with five statements evaluating the course duration, distribution of modules, delivery format, live instruction availability, and participant ability to manage course load (n=29).



XSL•FO
RenderX

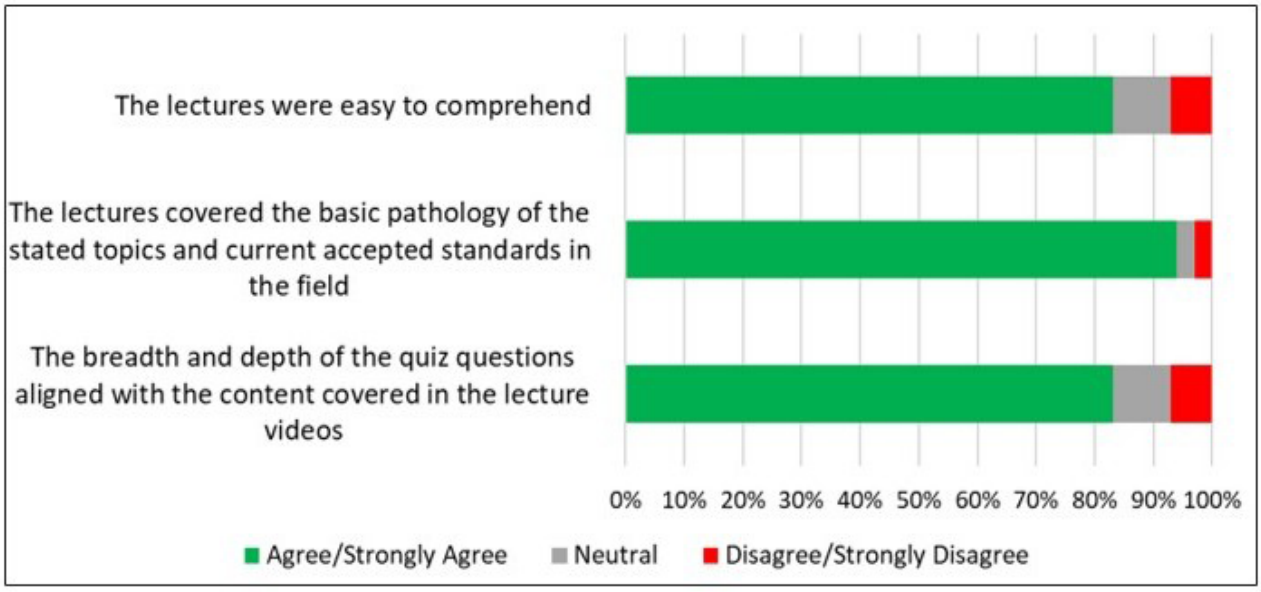
Figure 3. Fundamental Pathology for Basic Scientists: Data on participants' perceptions of course effectiveness. A Likert scale assessment of agreement with two statements evaluating achievement of learning objectives and effectiveness of instruction for each individual course module. At least 89% of learners agreed or strongly agreed that the lecturers and videos covered the stated objectives effectively and that the learning objectives were met in each of the 24 modules (n=29).



The third section of the exit survey focused on statements related to comprehension and standards of the course lectures, as well as the ability to address the quiz questions at module completion (Figure 4). In total, 83% of participants (24 out of 29) concurred (agreed or strongly agreed) that the lectures were easy to comprehend, and 27 out of 29 participants (94%) concurred that the lectures covered the basic pathology of the stated topics and matched currently accepted standards in the field. Eighty-three percent (24 out of 29) concurred (agreed or strongly agreed) that the breadth and depth of the quiz questions aligned with the content covered in the lecture videos.

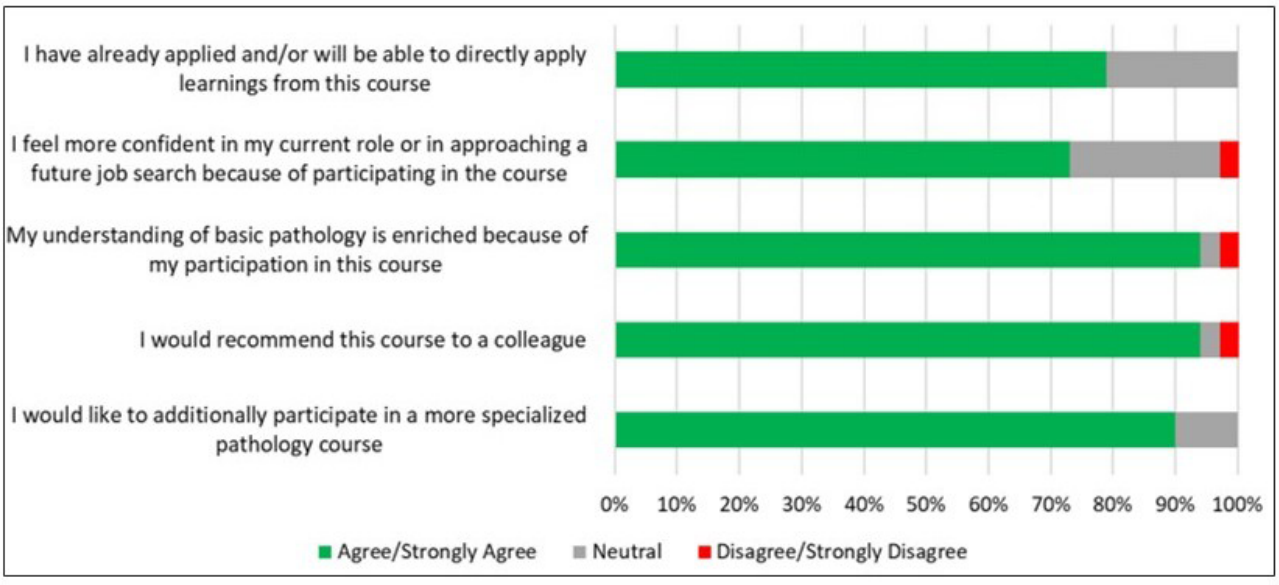
XSL•FO
RenderX

Figure 4. Fundamental Pathology for Basic Scientists: Data on participants' perceptions of course lectures. A Likert scale assessment of agreement with three statements evaluating course comprehension, standards and adequacy, and alignment of course assessment quizzes with instructional content (n=29).



The final section of the exit survey assessed the overall impact of the course on participants' careers and the potential application of what they learned in their current roles (Figure 5). Seventy-nine percent of participants (23 out of 29) concurred (agreed or strongly agreed) that they have either already applied or will be able to directly apply the content learned from this course to their current roles. Twenty-one out of 29 participants (72%) concurred (agreed or strongly agreed) that participation in this course made them feel more confident either in their current role or in approaching future job searches along their career trajectory. Twenty-seven out of 29 participants (94%) concurred (agreed or strongly agreed) that participation in the course directly enhanced their basic pathology knowledge. Further, 27 out of 29 participants (94%) also concurred that they would recommend this course to a colleague. Notably, 26 out of 29 participants (90%) concurred that they would like to return to participate in a more specialized pathology course (advanced topics in pathology, such as neuropathology, advanced tissue imaging and analysis techniques, animal disease model pathology, histology staining and processing techniques, histology interpretation, hematopathology, and cytopathology were suggested as content for future specialized course options), highlighting the positive impact of the pilot training course.

Figure 5. Fundamental Pathology for Basic Scientists: Data on participants' perceptions of near-term impact of course on professional development and career support. A Likert scale assessment of agreement with five statements evaluating course applicability to current career stage, confidence in job search, knowledge and skill development, and sustained interest in returning for an advanced course in pathology (n=29).



We also received free response feedback from 66% of participants (19 out of 29), regarding potential opportunities for course improvements in future iterations. Suggestions included the provision of additional relevant journal articles, the creation

XSL•FO
RenderX

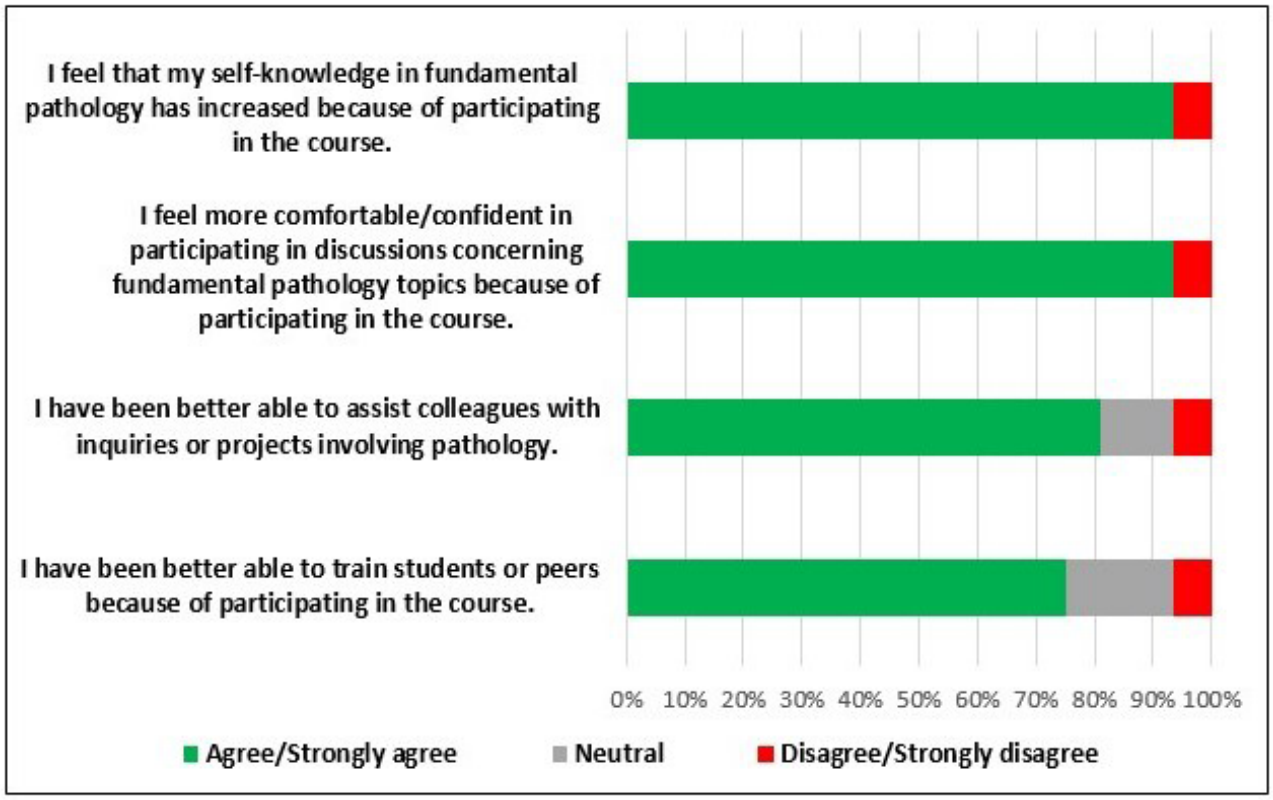
of a discussion board for learners to engage in conversations around pathology-related discussion prompts, providing more detailed explanations for images, varying the types of questions in quizzes, and collaborative activities among learners.

Postcourse Survey (Long-term Impact on Career and Professional Development)

A 5-point Likert scale was applied, with questions using the following scale anchors: 1- strongly disagree, 2- disagree, 3- neutral, 4- agree, and 5- strongly agree. Similar to the end-of-course survey, considering this is a small pilot study, the agree or strongly agree responses were combined into a single metric for ease of visualization of positive versus negative responses.The disagree or strongly disagree responses were similarly grouped together. Of the 29 participants who graduated from the pilot course, 55% (16 out of 29) completed postcourse survey responses a year after program graduation. This attrition in long-term follow-up is acknowledged; however, given that extended postprogram follow-up is a common hurdle in the field [18-20], this is an expected limitation. Nevertheless, obtaining feedback from over half of the program graduates was very valuable and will inform the design of the upcoming second course iteration.

Overall, 50% of the respondents (8 out of 16) had had a job change within the past year, and 81% (13 out of 16 respondents) concurred that their current jobs or roles involved pathology (data not shown). Moreover, 15 out of 16 respondents (94%) concurred (agreed or strongly agreed) that they felt more confident in participating in discussions concerning fundamental pathology topics at work (Figure 6). Fifteen out of 16 respondents (94%) also concurred that their knowledge in fundamental pathology has increased because of participating in this course (Figure 6). Feedback in the form of free responses indicated that the course helped learners identify and differentiate the distinct cell types within tissues, better understand mechanisms of resistance, better understand research results, quality assurance, better understand microscopic histology techniques, immunohistochemistry data, disease processes at a cellular level, as well as better understand normal tissue structure that has helped in recognizing tissue abnormalities. Learners also stated that this course made them more comfortable with pathology, made it easier to connect scientific concepts to real-world applications, and helped open career opportunities.

Figure 6. Fundamental Pathology for Basic Scientists: Data on longer-term (One-year post-program completion) impact of course on participants' pathology knowledge and teaching skills. A Likert scale assessment of agreement with four statements evaluating longer-term use of learned pathology knowledge and skills (n=16).



Extending to knowledge applications on the job, 13 out of 16 respondents (81%) concurred that they were better able to assist colleagues with inquiries or projects involving pathology and 12 out of 16 respondents (75%) agreed or strongly agreed that they were even better able to train students or peers because of participating in this course (Figure 6). Free response feedback also illuminated more avenues for course improvement, suggesting the need for better clarity in image descriptions within the video lectures.

XSL•FO
RenderX

Discussion

Principal Findings

The lack of accessible, structured learning opportunities in fundamental pathology creates a significant knowledge deficit that can limit career advancement for basic and translational scientists within the health care research enterprise. Furthermore, the rapid pace of advancements in biomedical research emphasizes the need for continuing education in pathology to ensure that professionals stay abreast of the latest developments. It is also imperative for such workplace training initiatives to embrace and integrate digital tools and virtual learning to solve access issues in the setting of the post-COVID-19 pandemic [21]. As the first iteration of the Fundamental Pathology for Basic Scientists course, the focus of this pilot study was on determining the value of the instructional design and program delivery to fill this training gap. Main findings yielded evaluative data on categories of course structure, course content, course effectiveness, course lectures, near-term impact of the course on professional development, and long-term impact on pathology knowledge applications. Line items within each category were crafted to obtain critical appraisal on overall course design and personal user aspects such as confidence in using the knowledge gained, as well as the actual application of learned skills on the job. Participant responses reflected strong endorsement of this course in terms of user value and use.

Notably, the course was developed using local institutional resources. It was able to draw participants from 16 different academic departments, representing 10 distinct job roles within the biomedical research enterprise at Texas Medical Center, ranging from early-career scientists to more established professionals. The majority of participants (22 out of 29 participants, 76%) lacked formal training in pathology. (Table S1 in Multimedia Appendix 1). The wide-ranging diversity of enrollees underscores the broad relevance of pathology knowledge across various career stages and disciplines. It also reflects the interdisciplinary nature of modern biomedical research, where a shared understanding of pathology can facilitate more effective collaboration and innovation. The positive feedback from participants, along with the heterogeneous mixture of departments and career stages represented, highlights the unmet demand for pathology education tailored to the needs of working professionals within the workplace setting. This demand suggests that similar courses could be highly beneficial if made more widely available, particularly in online, self-paced formats that accommodate the busy schedules of various professionals. Our study confirms and extends previous reports in the literature [22,23], showing that readily-accessible (such as locally-curated) educational resources would help close the knowledge gap in pathology and empower a broad audience of scientists and health care professionals to apply this critical education in their respective domains, and also facilitate advanced and interdisciplinary career opportunities within broader biomedical research disciplines (eg, teaching) (Figures 5 and 6).

The course's comprehensive curriculum, covering an extensive range of basic pathology topics, was one of its key strengths, with the benefits continuing to grow for participants even after a year of course completion (Figures 3 and 6). The observation that at least 26 out of 29 participants (89%) uniformly concurred that each of the 24 modules had met its stated learning objectives confirms the appropriateness of the instructional content. The evaluation method for course content primarily involved quizzes integrated into each learning module. These quizzes served as both an assessment tool and a reinforcement strategy, ensuring that participants engaged with and retained key concepts. This approach aligns with educational research that supports the use of frequent, low-stakes quizzes to enhance knowledge retention and promote active learning [24,25]. Participants responded positively to this form of assessment, appreciating the immediate feedback and the alignment of quizzes with the course content.

The observation that 93% (14 out of 16) of postcourse survey respondents felt this course significantly enhanced their knowledge and skills and further made them more confident in approaching projects involving pathology or even engaging in pathology-related discussion (Figure 6) demonstrates long-term impact of this course, and also its practical value for application across a wide spectrum of biomedical branches. Further, this highlights the course's ability to meet the diverse needs of a varied audience and reinforces its importance as a vital and readily-accessible resource for professionals seeking to deepen their understanding of pathology in the context of their work. One limitation of our evaluation is that we were able to receive the long-term survey responses from 55% (16 out of 29) of program graduates. While this attrition in contacting program alumni is indeed well described in the literature and is not uncommon [18-20], we anticipate being better positioned to evaluate and mitigate potential response bias, as the program expands with increased participant numbers in future iterations.

Opportunities for Program Enhancement in Future Iterations

The course, delivered over 6 months with 24 weekly modules, was designed to be flexible through an asynchronous format, allowing participants to integrate learning with their ongoing professional duties. While the course was effective in meeting its educational objectives, feedback indicated that managing the course workload alongside professional responsibilities remained a significant challenge for at least 6 out of the 29 participants (20%), suggesting that the course structure could benefit from further refinement. Several strategies could improve the course experience in future iterations: Extending the overall course timeline to 8 or 10 months would provide participants with more time to complete each module, making it easier to balance their educational pursuits with their professional duties. Modifying the module release schedule, by releasing modules biweekly or in clusters, instead of weekly, would offer participants greater time and flexibility in managing their schedules. Additionally, incorporating optional review periods or breaks between certain modules could serve as catch-up periods, reducing the pressure of keeping up with a weekly release schedule. Providing time management resources or strategies, such as workshops or tips on integrating learning with daily routines, would further help participants balance the course with their work. Regular feedback from participants could also be used to make real-time adjustments to the schedule

XSL•FO
RenderX

or workload, ensuring the course remains manageable for all participants.

Some participants suggested incorporating more diverse question types, such as case-based questions or scenario analysis, to better simulate real-world applications of pathology knowledge. Future iterations of the course will expand the quiz format to include a broader range of question types and consider adaptive quizzes that adjust in difficulty based on participants' responses.

Further, while the course's overall curriculum was well received, some participants expressed a desire for additional specialized topics, particularly in emerging areas of translational pathology. This could include advanced diagnostic techniques or deeper explorations of subspecialties like neuropathology or hematopathology. We view this as feasible because our course instructors are active professionals within the Texas Medical Center community who participate in continuing pathology education, pathology conferences, and/or meetings to keep up-to-date with the rapidly evolving landscape in pathology methods and research. There were also suggestions for more interactive elements, such as case-study discussions or group projects, to enhance engagement and deepen exploration of the material. Incorporating occasional synchronous sessions, despite the overall asynchronous format, could also provide a more dynamic learning experience. Finally, although virtual group office hours were offered, additional virtual availability with instructors on a one-on-one basis at the request of the participants seemed necessary to address inquiries or other ongoing course needs. Future iterations will seek to incorporate these suggestions and improvements into the curriculum.

Conclusions

The Fundamental Pathology for Basic Scientists course represents a local institution-based immersive learning tool to effectively address a critical educational gap in pathology for basic- and translational scientists. Emphasis was placed on learners acquiring a better sense of self-learning on pathology language and concepts as opposed to objectively measured cognitive competencies alone. Because participant makeup was explicitly nonpathologists, the building of their practical self-perceptions is a critical first step. While the pilot course was successful in achieving its training goals, multiple opportunities for course enhancement were identified from participant feedback. Future iterations of the course will consider extending the timeline, offering more specialized content, enhancing the image descriptions, and increasing interactive elements to enhance the learning experience. Regularly soliciting participant and instructor feedback will be crucial in continuously improving the course and ensuring it remains a valuable resource for the next generation of basic scientists and health care professionals.

Acknowledgments

We thank our course instructors for their generous service to education and training: Mohammad Al-Attar, Claudio Arrechedera, Harsh Batra, Elizve Barrientos Toro, Daniela Duenas, Shaimaa Elzamly, Sharia Hernandez, Caddie Laberiano, Debora Ledesma, Idania Lubo, Wei Lu, Rossana Lazcano, Mario Marquez Piubelli, Luisa Maren Solis Soto, Max Molina Ayala, Karen Nahmod, Oluwaseyi Olayinka, Karina Pinao, Ziba Rahbar, Maria E. Salvatierra, Leticia Hamana Zorrilla. We thank Ms. Ebonie Hatfield for program coordination and Dr. Kareen Chin for coordination with the Department of Pathology.

Funding

This pilot study was supported by the Department of Translational Molecular Pathology- and ITERT educational funds at the University of Texas MD Anderson Cancer Center (IIW, RS).

Data Availability

The (deidentified) cohort dataset used and analyzed during the current study is available from the corresponding author upon reasonable request for noncommercial academic purposes.

Authors' Contributions

JS and LCC wrote the primary manuscript and analyzed the data. RS and MGR conceptualized the course and curriculum. JS, LCC, IW, RS, and MGR designed the course structure. JS and LCC prepared and implemented the surveys. JS, LCC, DPF, MLS, IIW, RS, and MGR performed data review, proofreading, writing, and editing the manuscript. All authors have read and agreed to the publishable version of the manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Supplementary tables detailing the pre-entry survey outcomes and participants' evaluation of course content.
[DOCX File, 31 KB - mededu_v12i1e84903_app1.docx]

References

XSL•FO
RenderX

1. Burton JL. Teaching pathology to medical undergraduates. Curr Diagn Pathol 2005 Oct;11(5):308-316. [doi: 10.1016/j.cdip.2005.05.009]
2. Feinstein AR. Basic biomedical science and the destruction of the pathophysiologic bridge from bench to bedside. Am J Med 1999 Nov;107(5):461-467. [doi: 10.1016/s0002-9343(99)00264-8] [Medline: 10569301]
3. National Research Council Committee on the Evaluation of the Lucille P. Markey Charitable Trust Programs in Biomedical Sciences Bridging the Bed-Bench Gap: Contributions of the Markey Trust: National Academies Press; 2004. [doi: 10.17226/10920]
4. Anderson JR. Pathology and the undergraduate curriculum. J Pathol 1988 May;155(1):6-8. [doi: 10.1002/path.1711550104] [Medline: 3379518]
5. Fenderson BA. Strategies for teaching pathology to graduate students and allied health professionals. Hum Pathol 2005 Feb;36(2):146-153. [doi: 10.1016/j.humpath.2004.09.022] [Medline: 15754291]
6. Kemp JD, Bowser R, American Society for Investigative Pathology workshop. PhD graduate training programs in pathology: 2003 report from the American Society for Investigative Pathology workshop on pathology graduate education. Hum Pathol 2004 Jul;35(7):785-789. [doi: 10.1016/j.humpath.2004.03.014] [Medline: 15257540]
7. McCloskey CB, Johnson K, Brissette M, et al. Factors influencing US allopathic medical students to choose pathology as a specialty. Acad Pathol 2020;7:2374289520951924. [doi: 10.1177/2374289520951924] [Medline: 33110939]
8. Smith SM, Liauw D, Dupee D, Barbieri AL, Olson K, Parkash V. Burnout and disengagement in pathology: a prepandemic survey of pathologists and laboratory professionals. Arch Pathol Lab Med 2023 Jul 1;147(7):808-816. [doi: 10.5858/arpa.2022-0073-OA] [Medline: 36191345]
9. Kirkpatrick J, Kirkpatrick W. Kirkpatrick's Four Levels of Training Evaluation: ATD Press; 2016.
10. Hatfield ER, Shaw H, Farris DP, et al. Professional Alliance for Learning and Support (PALS): a pilot career development educational program for administrative support personnel in healthcare research. ACEQI 2025;1(1). [doi: 10.52519/ACEQI.25.1.1.a7]
11. Perez-Oquendo M, Romano G, Farris DP, et al. A structured curriculum supporting biomedical trainees' transition into independent academic positions and early career success. BMC Med Educ 2024 Apr 8;24(1):379. [doi: 10.1186/s12909-024-05370-w] [Medline: 38589919]
12. Stefaniak J. Advanced Instructional Design Techniques: Theories and Strategies for Complex Learning: Routledge; 2024. [doi: 10.4324/9781003287049]
13. Fundamental pathology for basic scientists: course learning modules. BioRender. URL: https://app.biorender.com/citation/68c4706ae1f0a76059a8792e [accessed 2026-05-20]
14. Kumar V, Abbas AK, Aster JC, Deyrup AT, Ae D, Robbins RSL. Kumar Basic Pathology, 11th edition: Elsevier; 2023. [doi: 10.33029/9704-6467-0-KRB-2022-1-1036]
15. Kumar V, Abbas A, Aster J, Deyrup A, Das A. Evolve Resources for Robbins & Kumar Basic Pathology: Elsevier; 2023. [doi: 10.33029/9704-6467-0-KRB-2022-1-1036]
16. Kirkpatrick D. Evaluating Training Programs: The Four Levels: Berrett-Koehler; 1994.
17. Spector J. Foundations of Educational Technology: Integrative Approaches and Interdisciplinary Perspectives: Routledge; 2016.
18. Chepp V, Baker C, Kostiha S, Smith JD. The Molecular Medicine PhD program alumni perceptions of career preparedness. PLoS ONE 2022;17(11):e0275996. [doi: 10.1371/journal.pone.0275996] [Medline: 36395255]
19. Elshami S, Awaisu A, Abidi R, et al. Examining pharmacy alumni's perceptions of job satisfaction, achievements, and preparedness: a mixed-methods study. Am J Pharm Educ 2023 May;87(5):100059. [doi: 10.1016/j.ajpe.2023.100059] [Medline: 37288694]
20. Leider JP, Rockwood TH, Mastrud H, Beebe TJ. Engaging public health alumni in the tracking of career trends: results from a large-scale experiment on survey fielding mode. Public Health Rep 2024;139(2):255-262. [doi: 10.1177/00333549231168640] [Medline: 37129371]
21. Ramseyer T, Pantanowitz L, Hassell L. Accelerating pathology education through virtual learning to address global health needs. In: Ordóñez de Pablos P, Zhang X, editors. Accelerating Strategic Changes for Digital Transformation in the Healthcare Industry: Academic Press; 2023:53-72. [doi: 10.1016/B978-0-443-15299-3.00013-0]
22. Lorenz RG, Karcher DS, Gautreaux MD, Limson M, Zander DS. The pathology workforce and clinical licensure: the role of the PhD clinical laboratorian in the United States. Acad Pathol 2018;5:2374289518775948. [doi: 10.1177/2374289518775948] [Medline: 29888322]
23. Straseski JA. Postdoctoral professional fellowships in laboratory medicine. EJIFCC 2013 Apr;24(1):21-29. [Medline: 27683435]
24. Lovett M, Bridges M, DiPietro M, Ambrose S, Norman M. How Learning Works: Eight Research-Based Principles for Smart Teaching, 2nd edition: Jossey-Bass; 2023.
25. Nilson L, Goodson LA. Online Teaching at Its Best: Merging Instructional Design with Teaching and Learning Research, 2nd edition: Jossey-Bass; 2021.

XSL•FO
RenderX

Abbreviations

UT: University of Texas

Edited by J Gentges; submitted 26.Sep.2025; peer-reviewed by C Pruss, J Dunleavey, MN Shalaby; revised version received 16.Mar.2026; accepted 31.Mar.2026; published 21.May.2026.

Please cite as:
Sells J, Clemente LC, Farris DP, Sizemore ML, Wistuba II, Soundararajan R, Raso MG
Enhancing Pathology Knowledge for Biomedical Scientists: Pilot Cohort Study of a Comprehensive Self-Paced Tutorial for Bridging Fundamental Concepts With Translational Applications
JMIR Med Educ 2026;12:e84903
URL: https://mededu.jmir.org/2026/1/e84903
doi:10.2196/84903

© Jennifer Sells, Leticia C Clemente, David P Farris, Mary L Sizemore, Ignacio I Wistuba, Rama Soundararajan, Maria G Raso. Originally published in JMIR Medical Education (https://mededu.jmir.org), 21.May.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Education, is properly cited. The complete bibliographic information, a link to the original publication on https://mededu.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Twelve Tips for Recognizing and Addressing the Adverse Effects of Medical Education Interventions: Tutorial

Tadayuki Hashimoto[1,2], MPH, MD; Mohammad Adrian Hasdianda[1], MCS, MMSc, MD; Kanapa Kornsawad[3], MD; Shunsuke Kosugi[4], MD; Makoto Kikukawa[4], MMEd, MD, PhD

[1]Department of Emergency Medicine, Brigham and Women's Hospital, 75 Francis St, Boston, MA, United States

[2]Osaka Medical and Pharmaceutical University, Takatsuki, Osaka, Japan

[3]The University of Texas Health Science Center at San Antonio, San Antonio, TX, United States

[4]Department of Medical Education, Faculty of Medical Sciences, Kyushu University, Fukuoka, Japan

Corresponding Author:
Tadayuki Hashimoto, MPH, MD
Department of Emergency Medicine, Brigham and Women's Hospital, 75 Francis St, Boston, MA, United States

Abstract

"First, do no harm" is a fundamental principle in health care, and clinical researchers carefully monitor adverse drug reactions to ensure patient safety. However, educational researchers and clinical educators rarely apply the same level of scrutiny to potential adverse effects arising from their own interventions. This reflects a persistent misconception that educational interventions are inherently harmless, an assumption that warrants critical examination. In this tutorial, we highlight the underrecognized concept of adverse effects in medical education by introducing 12 representative educational adverse effects and offering corresponding tips for mitigating them. These include the Dunning-Kruger effect, in which increased confidence does not align with actual competence; the undermining effect, whereby external rewards reduce intrinsic motivation; spoon feeding that stunts independent learning; cognitive overload resulting from excessive information delivery; patient dehumanization when education prioritizes technical proficiency over empathy; critiques of outcome-based medical education that may overemphasize measurable competencies at the expense of holistic professional development; and the expertise reversal effect, in which instructional strategies beneficial for novices become counterproductive as expertise grows. Additional adverse effects include compromised psychological safety despite formal safeguards, authority and confirmation biases that reinforce outdated practices, developer bias in intervention evaluation, the Hawthorne effect influencing observed behavior, and concerns that overreliance on generative artificial intelligence may hinder the development of critical thinking and metacognitive skills. To better understand the nature of these adverse effects, we categorize them into three overarching domains. Cognitive and psychological adverse effects occur within the learner. Structural and cultural adverse effects result from features of the educational environment. Methodological and evaluative adverse effects arise from how educational interventions are designed and assessed. While these domains overlap, they provide a practical framework for identifying how well-intended educational strategies may lead to harm. Some may argue that these phenomena represent unintended consequences rather than adverse effects. However, the term unintended consequence presumes that sufficient effort was made to anticipate and manage possible effects, an assumption that may not always hold in medical education. We argue that educators and educational researchers should explicitly recognize adverse effects, critically evaluate educational interventions, and adopt mitigation strategies with a level of rigor comparable to that applied in clinical research to better protect learners and improve the quality of medical education.

(JMIR Med Educ 2026;12:e89647) doi:10.2196/89647

KEYWORDS

medical education; educational interventions; adverse effects; unintended consequence; educational harm

Introduction

"First, do no harm" is widely recognized as a fundamental principle in health care [1]. All health care professionals are taught to uphold and practice this principle. In line with it, clinical researchers carefully design and conduct rigorous clinical trials, particularly when introducing new medications. Among these considerations, monitoring for adverse drug reactions is given the highest priority to ensure patient safety [2].

However, do educational researchers and even clinical educators apply the same level of scrutiny to potential adverse effects in their own interventions? The assumption that educational interventions are inherently harmless is a misconception. While often overlooked, certain educational strategies may result in educational adverse effects, an issue that remains underexplored in medical education research. Similar concerns have recently

XSL•FO
RenderX

been increasingly recognized in related fields, including psychosocial [3,4] and public health interventions [5].

In this review, we will highlight the underrecognized concept of adverse effects in medical education. To help prevent not only adverse outcomes in educational research but also those that may occur in day-to-day teaching and negatively affect learners, we introduce 12 representative educational adverse effects and offer 12 corresponding tips for mitigating them.

Tip 1: Confidence Hit the Wall of Ignorance (Dunning-Kruger Effect)

Providing information and training to beginners can enhance their confidence; however, this confidence may not align with actual competence, potentially leading to poor patient outcomes. This mismatch, known as the Dunning-Kruger effect, refers to a cognitive bias whereby individuals with limited expertise overestimate their abilities [6]. For instance, in a study on simulation-based training for joint injections among primary care physicians, those who reported higher confidence after training paradoxically performed worse in real clinical settings. (r=−0.253; P=.02) [7]. Such overconfidence can impair patient care and hinder proper skill development [8-10].

To mitigate this, educators can begin addressing this issue by acknowledging the presence of overconfidence and focusing on enhancing learners' metacognitive skills, enabling them to better evaluate their own competence [8]. Researchers, meanwhile, should refrain from relying solely on self-reported confidence as an outcome measure, a practice still common in educational studies [11,12].

Tip 2: Incentives Killed the Curiosity (Undermining Effect)

External rewards intended to motivate learners can paradoxically diminish their intrinsic motivation, a phenomenon known as the undermining effect [13]. This occurs when external incentives—such as grades or performance-based rewards—overshadow learners' internal drive to understand or improve. A meta-analysis investigating the impact of rewards on learner motivation found that engagement-, completion-, and performance-contingent rewards significantly reduced intrinsic motivation in free-choice learning tasks (effect sizes –0.40, –0.36, and –0.28, respectively) [14]. In medical education, a realist synthesis examining the impact of assessment on health care learners' motivation revealed that many evaluation systems encouraged surface-level learning and negatively impacted psychological well-being [15].

One effective way to encourage deeper clinical reasoning is for educators to move beyond standard multiple-choice questions and incorporate patient-centered tasks that require explanation and engagement [15]. These promote deeper engagement and align assessment with meaningful clinical competence. Researchers should also be cautious of the Hawthorne effect, where temporary motivation surges during the study period may obscure the intervention's true long-term impact [16].

Tip 3: Teach Learners to Learn, and You Feed Them for a Lifetime (Spoon Feeding)

When learners become accustomed to passively receiving one-way information, so-called spoon feeding, their intellectual and psychological growth can be stunted [17]. Teachers who deliver content without actively involving students risk undermining both independent learning and creativity [18]. Criticisms of this passive approach have persisted in medical education for over a century [19]. The solution is well-known: a shift toward active and self-directed learning [20]. Recent meta-analytic evidence supports this pedagogical transition. For instance, flipped classroom models have been shown to significantly improve both test scores (standardized mean difference 0.481, 95% CI 0.214-0.748) and skills performance (standardized mean difference 0.660, 95% CI 0.312-1.008) compared to traditional lectures [21].

However, the true challenge lies not in recognizing the problem, but in implementing meaningful cultural and institutional change [22,23]. At the individual level, educators should adopt active learning strategies within their own teaching contexts. At the institutional level, such as in medical schools, researchers should investigate how best to scale these pedagogical shifts—particularly by examining why adaptive learning models like problem-based learning, despite their theoretical appeal, have struggled in certain settings [24].

Tip 4: More Is Too Much (Cognitive Overload)

Educators often assume that more teaching leads to more learning. However, excessive delivery of information can overwhelm learners, sometimes resulting in lower retention than if less had been taught. This phenomenon is explained by cognitive load theory, which holds that working memory has a limited capacity to process information at any given time [25]. For example, in one study, intensive care unit attendings gave lectures with three key take-home messages. When 367 residents were later surveyed, two-thirds could recall only one message [26]. This illustrates the limits of cognitive processing under high load.

To mitigate this, educators should embrace the "less is more" principle in medical education, focusing not merely on the quantity of information but on managing the cognitive load imposed on learners [27]. Researchers, in turn, should explore strategies to optimize working memory use and reduce extraneous load in educational interventions [25].

Tip 5: The Patient Who Wasn't There (Patient Dehumanization)

Patient dehumanization occurs when learners begin to perceive patients not as individuals with emotions and unique needs, but merely as clinical cases—or worse, as objects to be examined and managed. This shift can stem from multiple factors,

XSL•FO
RenderX

including stress, rigid hierarchies, and an educational focus that prioritizes technical proficiency over empathy [28]. Although simulation-based training is valuable for skill acquisition [29], it may inadvertently contribute to this issue. In one study comparing two groups of medical students—one trained primarily through simulation and the other through bedside clinical clerkships—those in the simulation group scored significantly lower on measures of compassion (P=.02) [30].

To address this, educators must acknowledge the power dynamics inherent in clinical settings and remind learners that patients are not objective entities but sentient individuals with emotional experiences. This aligns with the concept of the iPatient—a term popularized by Dr Verghese [31]—which highlights how increasing reliance on electronic health records can shift attention away from real patients toward their digital representations. Researchers should critically evaluate educational interventions, particularly simulation-based training, for potential adverse effects such as dehumanization. This is especially important given that even existing reporting guidelines for simulation research rarely address this humanistic dimension [32].

Tip 6: In Objectivity We Trust (Critiques of Outcome-Based Medical Education)

Science prioritizes objectivity to ensure reproducibility. A prominent example of this in medical education is outcome-based medical education (OBME), which began to gain traction around the year 2000 and has since been implemented globally. This trend has mirrored the concept of entrustable professional activities as its most advanced form [33]. However, the emphasis on objectively measuring learner competence has also drawn criticism. Specifically, the rigid focus on predefined outcomes can oversimplify the nuanced, context-dependent skills required for clinical practice, reducing them to narrowly defined competencies [34]. For example, surgical safety checklists, widely adopted to enhance patient safety, emphasize standardized tasks like the “sign-out” phase. One study found that, following the implementation of a surgical safety program that included an educational intervention, sign-out completion improved from 79.3% to 94.5% (P<.0001), yet clinical incidents rose from 0.13% to 0.25% [35]. This raises concerns that an overemphasis on measurable behaviors may obscure more complex, less visible dimensions of patient safety.

OBME may also constrain educators’ creativity by limiting flexibility in teaching [36]. Rather than focusing solely on quantifiable outcomes, educators should preserve space to address essential but less easily measured aspects of training. From a research standpoint, there remains limited evidence demonstrating OBME’s long-term effectiveness. While funding and logistical constraints often hinder the assessment of patient-level outcomes, efforts should be made to evaluate the broader impact of OBME on clinical care [34].

Tip 7: A Farewell to Scaffolds (Expertise Reversal Effect)

As learners gain expertise, their cognitive load decreases, enabling them to benefit more from less structured, problem solving–based learning environments. Instructional methods that were once helpful for novices may become less effective—or even counterproductive—for more advanced learners. This phenomenon is known as the expertise reversal effect [37]. Multiple studies have demonstrated this effect. For example, novices benefited from additional explanatory text accompanying animations, while experts performed better without it [38]. Similarly, presenting correct examples before erroneous ones reduced mental load and improved learning outcomes for novices but had little impact on expert learners [39].

To address this, educators should not avoid categorizing learners based solely on their year or generation [40]; instead, they should assess learners’ cognitive developmental stage, such as those described in the RIME (reporter to interpreter to manager to educator) model [41]. Future research should explore adaptive learning systems that tailor instruction to learners’ expertise in real time. For example, intelligent tutoring systems and simulation platforms could adjust case complexity or the level of guidance based on a trainee’s performance and prior knowledge. Evaluating the effectiveness of such personalized, just-in-time instructional strategies could help translate the principles of the expertise reversal effect into broader educational practice.

Tip 8: Let Me Be Wrong (Psychological Safety)

Psychological safety is increasingly recognized as a foundational element of effective learning in medical education. It refers to the belief that one can take interpersonal risks—such as offering questions, admitting uncertainty, or sharing ideas—without fear of embarrassment, punishment, or negative judgment [42,43]. However, even when educators explicitly assure students that these will not impact their evaluations, learners may still tailor their entries to highlight only performance in “safe” or favorable contexts, fearing judgment nonetheless [42,43]. For example, in one study, students were asked to submit reflective portfolios, and they selectively recorded experiences where they performed well, rather than those that truly challenged them [44]. This demonstrates how psychological safety can be compromised despite well-intentioned safeguards. Another example of this paradox can be seen in a study where an asynchronous learning platform was developed to allow residents to reflect at their own convenience. Although residents reported high satisfaction with the perceived educational value, the ability to post during late-night hours unintentionally increased their workload, with many reflections being submitted after midnight [45].

To mitigate this, educators must actively cultivate psychologically safe learning environments, acknowledging that, due to hierarchical structures in clinical education, the primary responsibility lies with supervisors [42,43]. Researchers

XSL•FO
RenderX

should also prioritize the development and evaluation of interventions aimed at enhancing psychological safety, as prospective studies in this area remain relatively limited [46].

Tip 9: Old Ways Never Die (Authority Bias)

As educators, we must recognize that learners are particularly vulnerable to cognitive biases such as authority bias—the tendency to unquestioningly accept information from perceived experts [47]—and confirmation bias—the tendency to disregard evidence that contradicts pre-existing beliefs [48,49]. When trusted educators convey outdated or incorrect knowledge, these biases can reinforce misinformation and hinder evidence-based practice [50]. For example, the concepts behind initiatives such as "Things We Do for No Reason" and the campaign of "Choosing Wisely" illustrate how certain clinical practices persist despite lacking evidence, often due to tradition and uncritical deference to authority [51,52].

To mitigate these biases, educators need to go beyond delivering static content and prioritize the development of learners' critical appraisal skills and information literacy [53]. Researchers likewise bear responsibility and should avoid perpetuating discredited concepts, such as the widely debunked theory of learning styles [54].

Tip 10: Developer's Strangelove (Experimenter Effect and IKEA Effect)

Educational interventions and curricula are often developed by the same individuals who evaluate them. This dual role can introduce unconscious biases, such as the experimenter effect—where researcher expectations unintentionally influence outcomes [55]—and the IKEA effect—the tendency to overvalue one's own creations [56]. Similar concerns have been raised in pharmaceutical research, where criticism arises when drug developers lead their own clinical trials [57]. Yet in educational research, it is still common for developers to assess the effectiveness of their own interventions. In such cases, researchers' expectations can shape both the design and interpretation of outcomes, mirroring biases observed in behavioral studies [58].

To address these, educators and researchers should exercise caution when evaluating their own educational products. The involvement of independent evaluators can enhance the objectivity of their intervention [29]. In educational intervention research, adhering to reporting guidelines such as GREET (Guidelines for Reporting Educational Intervention Trials) [59] and relevant educational reporting guidelines [60] can not only ensure transparency in study design and reporting but also minimize potential bias introduced by developers.

Tip 11: Learners Are Watching You (Hawthorne Effect)

Direct observation is widely regarded as one of the most effective assessment tools in medical education [61]. Training programs and specialties have now defined required competencies, subcompetencies, developmental milestones, performance levels, and entrustable professional activities that can be directly observed and assessed [33]. However, the act of observation itself may influence behavior, a phenomenon known as the Hawthorne effect [16,62]. For example, when unannounced standardized patients were used to assess health care teams, some patient safety measures were found to be alarmingly deficient: only 15% of patients were screened for depression using the 2-item Patient Health Questionnaire, and 5% of clinicians performed hand hygiene before examinations [63].

Educators should be aware of the potential influence of observation and consider incorporating unscheduled direct observations or unannounced standardized patients as part of a broader, multimethod evaluation strategy. Researchers, in turn, should recognize that although behavioral change—as described in Kirkpatrick's third level of evaluation—is valuable, the ultimate goal remains to measure outcomes. Given the limited empirical evidence for the Hawthorne effect, further investigation into its actual impact on educational evaluations may still be warranted [64].

Tip 12: Do Learners Dream of Thinking Skills? (Concerns of Generative Artificial Intelligence)

The advancement of artificial intelligence (AI) has been remarkable and is undeniably reshaping the educational landscape. However, concerns have emerged that overreliance on generative AI (GenAI) may reduce meaningful learning opportunities. A 2024 systematic review suggested that excessive dependence on GenAI may hinder the development of key cognitive skills, including critical thinking, creativity, problem-solving, self-regulation, and metacognition [65]. Moreover, 83% of faculty members expressed concern that GenAI could reduce or eliminate essential human interactions in feedback and instructional processes [66].

To address these issues, educators should use GenAI with a clear understanding of its limitations, particularly about factual inaccuracies, lack of medical expertise, ethical considerations, and financial costs [67]. Recent frameworks emphasize the importance of clinical supervision strategies when learners use AI tools [68]. These include establishing clear guidelines for appropriate AI use, implementing structured supervision processes to monitor AI-assisted clinical decisions, and fostering reflective practices that help learners critically evaluate AI-generated outputs rather than accepting them uncritically. Researchers should prioritize investigations into the ethical implications of GenAI in medical education and its potential long-term effects on learner development.

Conclusion

To better understand the nature of adverse effects in medical education, we categorized the 12 identified examples into three overarching domains. Cognitive and psychological adverse effects occur within the learner, including overconfidence,

XSL•FO
RenderX

reduced intrinsic motivation, cognitive overload, loss of autonomy, compromised psychological safety, and diminished learning from overly structured support as expertise grows. Structural and cultural adverse effects result from features of the educational environment, such as rigid outcome-based frameworks, authority bias, practices that may dehumanize patients or hinder empathy, and overreliance on GenAI tools. Methodological and evaluative adverse effects arise from how educational interventions are designed or assessed, including biases introduced by developer-led evaluations and the Hawthorne effect. While these categories may overlap, they offer a practical framework to identify how educational strategies—even well-intended ones—may lead to unintended harm.

Some may argue that these 12 points are not truly adverse effects but rather unintended consequences. However, we would like to challenge whether the term unintended consequence is sufficient or even appropriate in this context. The notion of an unintended consequence implies that there was a deliberate and thoughtful attempt to intend an outcome in the first place. In other words, it presumes that sufficient effort was made to anticipate and manage possible effects, something that may not always be the case in medical education. We, educators and educational researchers, need these efforts more.

Funding

The authors declared that no financial support was received for this work.

Authors' Contributions

TH conceived the concept of this tutorial and drafted the manuscript. All authors participated in discussions regarding the content, provided critical feedback, and reviewed the manuscript. All authors read and approved the final manuscript.

Conflicts of Interest

None declared.

References

1. Inman T. Foundation for a New Theory and Practice of Medicine: John Churchill; 1860. URL: https://archive.org/details/foundationforan00inmagoog [accessed 2026-11-13]
2. Guideline for Good Clinical Practice E6(R3). International Council for Harmonisation of Technical Requirements for Pharmaceuticals for Human Use. 2025 Jan 6. URL: https://database.ich.org/sites/default/files/ICH_E6%28R3%29_Step4_FinalGuideline_2025_0106.pdf [accessed 2025-03-31]
3. Linden M, Schermuly-Haupt ML. Definition, assessment and rate of psychotherapy side effects. World Psychiatry 2014 Oct;13(3):306-309. [doi: 10.1002/wps.20153] [Medline: 25273304]
4. Boettcher J, Rozental A, Andersson G, Carlbring P. Side effects in internet-based interventions for social anxiety disorder. Internet Interv 2014 Mar;1(1):3-11. [doi: 10.1016/j.invent.2014.02.002]
5. Lorenc T, Oliver K. Adverse effects of public health interventions: a conceptual framework. J Epidemiol Community Health 2014 Mar;68(3):288-290. [doi: 10.1136/jech-2013-203118] [Medline: 24336237]
6. Kruger J, Dunning D. Unskilled and unaware of it: how difficulties in recognizing one's own incompetence lead to inflated self-assessments. J Pers Soc Psychol 1999;77(6):1121-1134. [doi: 10.1037/0022-3514.77.6.1121] [Medline: 10626367]
7. Leopold SS, Morgan HD, Kadel NJ, Gardner GC, Schaad DC, Wolf FM. Impact of educational intervention on confidence and competence in the performance of a simple surgical task. J Bone Joint Surg Am 2005 May;87(5):1031-1037. [doi: 10.2106/00004623-200505000-00014] [Medline: 15866966]
8. Gaeta TJ, Reisdorff E, Barton M, et al. The Dunning Kruger effect in resident predicted and actual performance on the American Board of Emergency Medicine in-training examination. J Am Coll Emerg Physicians Open 2024 Oct 24;5(5):e13305. [doi: 10.1002/emp2.13305] [Medline: 39463809]
9. Rahmani M. Medical trainees and the Dunning-Kruger effect: when they don't know what they don't know. J Grad Med Educ 2020 Oct;12(5):532-534. [doi: 10.4300/JGME-D-20-00134.1] [Medline: 33149817]
10. Knof H, Berndt M, Shiozawa T. Prevalence of Dunning-Kruger effect in first semester medical students: a correlational study of self-assessment and actual academic performance. BMC Med Educ 2024 Oct 24;24(1):1210. [doi: 10.1186/s12909-024-06121-7] [Medline: 39449011]
11. Kirkpatrick DL, Kirkpatrick JD. Evaluating Training Programs: The Four Levels, 3rd edition: Berrett-Koehler Publishers; 2006.
12. Yardley S, Dornan T. Kirkpatrick's levels and education "evidence". Med Educ 2012 Jan;46(1):97-106. [doi: 10.1111/j.1365-2923.2011.04076.x] [Medline: 22150201]
13. Lepper MR, Greene D, Nisbett RE. Undermining children's intrinsic interest with extrinsic reward: a test of the "overjustification" hypothesis. J Pers Soc Psychol 1973;28(1):129-137. [doi: 10.1037/h0035519]
14. Deci EL, Koestner R, Ryan RM. A meta-analytic review of experiments examining the effects of extrinsic rewards on intrinsic motivation. Psychol Bull 1999 Nov;125(6):627-668. [doi: 10.1037/0033-2909.125.6.627] [Medline: 10589297]

XSL•FO
RenderX

15. Kusurkar RA, Orsini C, Somra S, et al. The effect of assessments on student motivation for learning and its outcomes in health professions education: a review and realist synthesis. Acad Med 2023 Sep 1;98(9):1083-1092. [doi: 10.1097/ACM.0000000000005263] [Medline: 37146237]
16. Parsons HM. What happened at Hawthorne?: New evidence suggests the Hawthorne effect resulted from operant reinforcement contingencies. Science 1974 Mar 8;183(4128):922-932. [doi: 10.1126/science.183.4128.922] [Medline: 17756742]
17. Abu Samah SA, Jusoff K, Silong AD. Does spoon-feeding impede independent learning? Can Soc Sci 2009;5(3):82-90. [doi: 10.3968/j.css.1923669720090503.010]
18. Etchison C. Literature without lectures: an approach that works. Education Resources Information Center (ERIC). 1988. URL: https://eric.ed.gov/?id=ED300830 [accessed 2025-03-31]
19. Bois EF, Reznikoff P. The clinical clerkship in medicine: details of a method of instruction. JAMA 1926;87(8):542-545. [doi: 10.1001/jama.1926.02680080008003]
20. McKay J, Kember D. Spoon feeding leads to regurgitation: a better diet can result in more digestible learning outcomes. High Educ Res Dev 1997 Apr;16(1):55-67. [doi: 10.1080/0729436970160105]
21. Shi XY, Yin Q, Wang QW, et al. Is the flipped classroom more effective than the traditional classroom in clinical medical education: a systematic review and meta-analysis. Front Educ 2025 Jan 5;9:1485540. [doi: 10.3389/feduc.2024.1485540]
22. Nothnagle M, Anandarajah G, Goldman RE, Reis S. Struggling to be self-directed: residents' paradoxical beliefs about learning. Acad Med 2011 Dec;86(12):1539-1544. [doi: 10.1097/ACM.0b013e3182359476] [Medline: 22030764]
23. Charokar K, Dulloo P. Self-directed learning theory to practice: a footstep towards the path of being a life-long learner. J Adv Med Educ Prof 2022 Jul;10(3):135-144. [doi: 10.30476/JAMP.2022.94833.1609] [Medline: 35910513]
24. Hu X, Li J, Wang X, et al. Medical education challenges in Mainland China: an analysis of the application of problem-based learning. Med Teach 2025 Apr;47(4):713-728. [doi: 10.1080/0142159X.2024.2369238] [Medline: 39073879]
25. Young JQ, Van Merrienboer J, Durning S, Ten Cate O. Cognitive load theory: implications for medical education: AMEE Guide No. 86. Med Teach 2014 May;36(5):371-384. [doi: 10.3109/0142159X.2014.889290] [Medline: 24593808]
26. Lautrette A, Schwebel C, Gruson D, Talbot RW, Timsit JF, Souweine B. Transfer of take-home messages in graduate ICU education. Intensive Care Med 2011 Aug;37(8):1323-1330. [doi: 10.1007/s00134-011-2256-7] [Medline: 21660536]
27. Grady D, Redberg RF. Less is more: how less health care can result in better health. Arch Intern Med 2010 May 10;170(9):749-750. [doi: 10.1001/archinternmed.2010.90] [Medline: 20458080]
28. Amen TB, Nolen LT. Patient dehumanization in medical education: reflections on Mr. E. Acad Med 2021 Nov 1;96(11):1506. [doi: 10.1097/ACM.0000000000004356] [Medline: 34380928]
29. Cook DA, Hatala R, Brydges R, et al. Technology-enhanced simulation for health professions education: a systematic review and meta-analysis. JAMA 2011 Sep 7;306(9):978-988. [doi: 10.1001/jama.2011.1234] [Medline: 21900138]
30. Ebm C, Sarti R, Panico P, Pagliotta M, Vinci V, Oldani S. Enhancing compassion in medical education - a comparative study of the efficacy of clinical clerkships versus simulation-based training methodologies. BMC Med Educ 2025 Feb 4;25(1):181. [doi: 10.1186/s12909-025-06687-w] [Medline: 39905468]
31. Verghese A. A doctor's touch. TED. 2011 Jul. URL: https://www.ted.com/talks/abraham_verghese_a_doctor_s_touch [accessed 2025-03-31]
32. Cheng A, Kessler D, Mackinnon R, et al. Reporting guidelines for health care simulation research: extensions to the CONSORT and STROBE statements. Adv Simul (Lond) 2016 Jul 25;1:25. [doi: 10.1186/s41077-016-0025-y] [Medline: 29449994]
33. ten Cate O. Entrustability of professional activities and competency-based training. Med Educ 2005 Dec;39(12):1176-1177. [doi: 10.1111/j.1365-2929.2005.02341.x] [Medline: 16313574]
34. Morcke AM, Dornan T, Eika B. Outcome (competency) based education: an exploration of its origins, theoretical basis, and empirical evidence. Adv Health Sci Educ Theory Pract 2013 Oct;18(4):851-863. [doi: 10.1007/s10459-012-9405-9] [Medline: 22987194]
35. Gillespie BM, Harbeck EL, Lavin J, et al. Evaluation of a patient safety programme on surgical safety checklist compliance: a prospective longitudinal study. BMJ Open Qual 2018 Jul 12;7(3):e000362. [doi: 10.1136/bmjoq-2018-000362] [Medline: 30057963]
36. Harden RM. Developments in outcome-based education. Med Teach 2002 Mar;24(2):117-120. [doi: 10.1080/01421590220120669] [Medline: 12098427]
37. Kalyuga S, Ayres P, Chandler P, Sweller J. The expertise reversal effect. Educ Psychol 2003 Jun 8;38(1):23-31. [doi: 10.1207/S15326985EP3801_4]
38. Rey GD, Buchwald F. The expertise reversal effect: cognitive load and motivational explanations. J Exp Psychol Appl 2011 Mar;17(1):33-48. [doi: 10.1037/a0022243] [Medline: 21443379]
39. Wesenberg L, Krieglstein F, Jansen S, Rey GD, Beege M, Schneider S. The influence of the order and congruency of correct and erroneous worked examples on learning and (meta-)cognitive load. Front Psychol 2022 Oct 26;13:1032003. [doi: 10.3389/fpsyg.2022.1032003] [Medline: 36389584]

XSL•FO
RenderX

40. Bracken RC, Fredrickson ME, Fredrickson LA, Appleman M. Generational situatedness: challenging generational stereotypes in health professions education. Med Teach 2023 Apr;45(4):380-387. [doi: 10.1080/0142159X.2022.2135428] [Medline: 36306344]
41. Pangaro L. A new vocabulary and other innovations for improving descriptive in-training evaluations. Acad Med 1999 Nov;74(11):1203-1207. [doi: 10.1097/00001888-199911000-00012] [Medline: 10587681]
42. Nembhard IM, Edmondson AC. Making it safe: the effects of leader inclusiveness and professional status on psychological safety and improvement efforts in health care teams. J Organ Behavior 2006 Nov;27(7):941-966. [doi: 10.1002/job.413]
43. Caverzagie KJ, Goldenberg MG, Hall JM. Psychology and learning: the role of the clinical learning environment. Med Teach 2019 Apr;41(4):375-379. [doi: 10.1080/0142159X.2019.1567910] [Medline: 30761927]
44. Oudkerk Pool A, Jaarsma ADC, Driessen EW, Govaerts MJB. Student perspectives on competency-based portfolios: does a portfolio reflect their competence development? Perspect Med Educ 2020 Jun;9(3):166-172. [doi: 10.1007/s40037-020-00571-7] [Medline: 32274650]
45. Hashimoto T, Ariyoshi S, Ariyoshi T, et al. When is “convenient time” for residents?: a trial of Resident Study Log as asynchronous learning tool for residents. Med Educ Online 2025 Dec;30(1):2461579. [doi: 10.1080/10872981.2025.2461579] [Medline: 39893551]
46. McClintock AH, Fainstad T, Blau K, Jauregui J. Psychological safety in medical education: a scoping review and synthesis of the literature. Med Teach 2023 Nov;45(11):1290-1299. [doi: 10.1080/0142159X.2023.2216863] [Medline: 37266963]
47. Milgram S. Behavioral study of obedience. J Abnorm Psychol 1963 Oct;67:371-378. [doi: 10.1037/h0040525] [Medline: 14049516]
48. Wason PC. On the failure to eliminate hypotheses in a conceptual task. Q J Exp Psychol (Hove) 1960 Jul;12(3):129-140. [doi: 10.1080/17470216008416717]
49. Nickerson RS. Confirmation bias: a ubiquitous phenomenon in many guises. Rev Gen Psychol 1998 Jun;2(2):175-220. [doi: 10.1037/1089-2680.2.2.175]
50. Garcia-Arch J, Barberia I, Rodríguez-Ferreiro J, Fuentemilla L. Authority brings responsibility: feedback from experts promotes an overweighting of health-related pseudoscientific beliefs. Int J Environ Res Public Health 2022 Nov 17;19(22):15154. [doi: 10.3390/ijerph192215154] [Medline: 36429874]
51. Good Stewardship Working Group. The “top 5” lists in primary care: meeting the responsibility of professionalism. Arch Intern Med 2011 Aug 8;171(15):1385-1390. [doi: 10.1001/archinternmed.2011.231] [Medline: 21606090]
52. Feldman LS. Choosing Wisely(®): things we do for no reason. J Hosp Med 2015 Oct;10(10):696. [doi: 10.1002/jhm.2425] [Medline: 26375491]
53. Kendeou P, Johnson V. The nature of misinformation in education. Curr Opin Psychol 2024 Feb;55:101734. [doi: 10.1016/j.copsyc.2023.101734] [Medline: 38029642]
54. Grospietsch F, Mayer J. Pre-service science teachers’ neuroscience literacy: neuromyths and a professional understanding of learning and memory. Front Hum Neurosci 2019 Feb 14;13:20. [doi: 10.3389/fnhum.2019.00020] [Medline: 30890924]
55. Rosenthal R, Fode KL. The effect of experimenter bias on the performance of the albino rat. Syst Res 1963;8(3):183-189. [doi: 10.1002/bs.3830080302]
56. Norton MI, Mochon D, Ariely D. The IKEA effect: when labor leads to love. J Consum Psychol 2012 Jul;22(3):453-460. [doi: 10.1016/j.jcps.2011.08.002]
57. Schott G, Pachl H, Limbach U, Gundert-Remy U, Ludwig WD, Lieb K. The financing of drug trials by pharmaceutical companies and its consequences. Part 1: a qualitative, systematic review of the literature on possible influences on the findings, protocols, and quality of drug trials. Dtsch Arztebl Int 2010 Apr;107(16):279-285. [doi: 10.3238/arztebl.2010.0279] [Medline: 20467553]
58. Klein O, Doyen S, Leys C, et al. Low hopes, high expectations: expectancy effects and the replicability of behavioral experiments. Perspect Psychol Sci 2012 Nov;7(6):572-584. [doi: 10.1177/1745691612463704] [Medline: 26168114]
59. Phillips AC, Lewis LK, McEvoy MP, et al. Development and validation of the guideline for reporting evidence-based practice educational interventions and teaching (GREET). BMC Med Educ 2016 Sep 6;16(1):237. [doi: 10.1186/s12909-016-0759-1] [Medline: 27599967]
60. Upsher R, Dommett E, Carlisle S, et al. Improving reporting standards in quantitative educational intervention research: introducing the CLOSER and CIDER checklists. J New Approaches Educ Res 2025;14(1):2. [doi: 10.1007/s44322-024-00022-9] [Medline: 39916891]
61. Kogan JR, Hatala R, Hauer KE, Holmboe E. Guidelines: the do’s, don’ts and don’t knows of direct observation of clinical skills in medical education. Perspect Med Educ 2017 Oct;6(5):286-305. [doi: 10.1007/s40037-017-0376-7] [Medline: 28956293]
62. McCambridge J, Witton J, Elbourne DR. Systematic review of the Hawthorne effect: new concepts are needed to study research participation effects. J Clin Epidemiol 2014 Mar;67(3):267-277. [doi: 10.1016/j.jclinepi.2013.08.015] [Medline: 24275499]
63. Zabar S, Hanley K, Stevens D, et al. Unannounced standardized patients: a promising method of assessing patient-centered care in your health care system. BMC Health Serv Res 2014 Apr 5;14:157. [doi: 10.1186/1472-6963-14-157] [Medline: 24708683]

XSL•FO
RenderX

64. Paradis E, Sutkin G. Beyond a good story: from Hawthorne Effect to reactivity in health professions education research. Med Educ 2017 Jan;51(1):31-39. [doi: 10.1111/medu.13122] [Medline: 27580703]
65. Zhai C, Wibowo S, Li LD. The effects of over-reliance on AI dialogue systems on students' cognitive abilities: a systematic review. Smart Learn Environ 2024;11(28). [doi: 10.1186/s40561-024-00316-7]
66. Cervantes J, Smith B, Ramadoss T, D'Amario V, Shoja MM, Rajput V. Decoding medical educators' perceptions on generative artificial intelligence in medical education. J Investig Med 2024 Oct;72(7):633-639. [doi: 10.1177/10815589241257215] [Medline: 38785310]
67. Indran IR, Paranthaman P, Gupta N, Mustafa N. Twelve tips to leverage AI for efficient and effective medical question generation: a guide for educators using Chat GPT. Med Teach 2024 Aug;46(8):1021-1026. [doi: 10.1080/0142159X.2023.2294703] [Medline: 38146711]
68. Abdulnour REE, Gin B, Boscardin CK. Educational strategies for clinical supervision of artificial intelligence use. N Engl J Med 2025 Aug 21;393(8):786-797. [doi: 10.1056/NEJMra2503232] [Medline: 40834302]

Abbreviations

AI: artificial intelligence
GenAI: generative artificial intelligence
GREET: Guidelines for Reporting Educational Intervention Trials
OBME: outcome-based medical education
RIME: reporter to interpreter to manager to educator

Edited by W Jerjes; submitted 15.Dec.2025; peer-reviewed by C Bell, CP Fu, KH Miller; revised version received 01.Apr.2026; accepted 01.Apr.2026; published 22.May.2026.

Please cite as:
Hashimoto T, Hasdianda MA, Kornsawad K, Kosugi S, Kikukawa M
Twelve Tips for Recognizing and Addressing the Adverse Effects of Medical Education Interventions: Tutorial
JMIR Med Educ 2026;12:e89647
URL: https://mededu.jmir.org/2026/1/e89647
doi:10.2196/89647

© Tadayuki Hashimoto, Mohammad Adrian Hasdianda, Kanapa Kornsawad, Shunsuke Kosugi, Makoto Kikukawa. Originally published in JMIR Medical Education (https://mededu.jmir.org), 22.May.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Education, is properly cited. The complete bibliographic information, a link to the original publication on https://mededu.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Trust Analysis Canvas for Teaching in the Field of Digital Public Health and Medicine: Tutorial

Federica Zavattaro[1,2], MSc; Clara-Maria Barth[3], MSc; Caroline Brall[4], PhD; Viktor von Wyl[1,2,5,6], PhD; Felix Gille[1,2], PhD

[1]Digital Society Initiative, University of Zurich, Rämistrasse 69, Zurich, Switzerland

[2]Institute for Implementation Science in Health care (IfIS), University of Zurich, Zurich, Switzerland

[3]Department of Informatics, University of Zurich, Zurich, Switzerland

[4]Ethics and Policy Lab, Multidisciplinary Center for Infectious Diseases & Institute of Philosophy, University of Bern, Bern, Switzerland

[5]Epidemiology, Biostatistics & Prevention Institute (EBPI), University of Zurich, Zurich, Switzerland

[6]Swiss School of Public Health, Zurich, Switzerland

Corresponding Author:
Federica Zavattaro, MSc
Digital Society Initiative, University of Zurich, Rämistrasse 69, Zurich, Switzerland

Abstract

Trust is increasingly recognized as a cornerstone for the successful implementation of digital public health initiatives, from mobile apps to the use of artificial intelligence in medicine, yet it remains underrepresented in educational curricula. In the course of our research and teaching activities in the field of trust in digital public health and medicine, we identified a gap in existing educational resources that aimed at supporting students in conducting structured trust analyses. Digitalization introduces new complexities into trust relationships, as interactions become increasingly mediated by digital tools. Preparing future professionals, therefore, demands fostering a critical understanding of how trust operates within digital systems, especially in the health sector. To address this gap, we developed and tested the first Trust Analysis Canvas for Teaching (TACT), a tool designed to guide students in conducting trust analyses of case studies in digital public health and medicine. Grounded in conceptual research on trust in health systems and health data sharing, we (1) developed the canvas content and reviewed it with two trust researchers; (2) tested and iteratively refined the tool with 23 students (3 BSc, 14 MSc, and 6 PhD) from diverse disciplines and academic levels through in-person and online focus groups at the universities of Zurich and Bern; (3) collaborated with a graphic designer to optimize its visual layout; and (4) translated the final canvas into French, Italian, German, and Spanish to ensure accessibility across disciplines, academic levels, and languages while maintaining a clear and engaging visual design. This paper introduces TACT, a canvas comprising 16 guiding questions organized around 6 core dimensions, designed to enable students from diverse disciplinary backgrounds and academic levels to engage with the complex concept of trust in a structured and guided manner, thereby addressing the identified gap in the current curricula. We outline the development process and provide a practical, step-by-step tutorial demonstrating its application through a written trust analysis of a digital health case study, supported by references to relevant literature.

(JMIR Med Educ 2026;12:e79709) doi:10.2196/79709

KEYWORDS

trust; teaching; canvas; education, medical; digital health; medicine

Introduction

Defined as “a bet about the future contingent actions of others” [1], trust has gained increasing momentum in digital health research and health policy making [2-4], where it is recognized as a central component of the relationship between the public and the health care system, enabling public participation and legitimizing health care activities [5]. Trust also plays a critical role in digital public health, understood as the integration of digital technologies into public health to enhance population health outcomes and efficiency of services [6], as well as in medicine and beyond. It shapes the design and use of medical applications [7], the adoption of electronic health records [8], vaccination uptake [9], compliance with nonpharmaceutical public health interventions [10], the use of e-government services [11], e-banking platforms [12], and artificial intelligence (AI) systems [13-15]. Across and beyond the health care sector, higher levels of trust are associated with greater social cohesion, lower system costs, improved service efficiency, and overall societal prosperity. Conversely, low levels of trust can undermine system stability, affecting sectors ranging from health care to government [5].

Trust operates within a context of vulnerability, in which the trustor relinquishes some control, lowers certain defenses, and

XSL•FO
RenderX

exposes themselves to potential harm [16,17]. This is particularly relevant in digital public health and medicine, where digitalization enables the processing and sharing of large volumes of data, increasing both the complexity of trust relationships and the exposure of sensitive information to risks. In the context of health data—classified as sensitive personal data under the General Data Protection Regulation [18]—trust entails the expectation that institutions, organizations, and individuals responsible for collecting, storing, sharing, and processing health data will act with integrity, reliability, competence, and in the public's best interest [5].

Preparing future professionals for these environments, therefore, requires more than technical expertise: students should also develop a critical understanding of how trust operates within digital systems. Despite its central importance, the current status of trust education in relevant university programs remains limited. In medicine, public health, and digital health curricula across bachelor's (BSc), master's (MSc), and doctoral levels (PhD), at least in our setting in Switzerland [19], there is no structured or systematic approach to teaching trust as an analytical concept, nor guidance on applying trust frameworks to real-world cases. While existing courses may touch on related dimensions such as ethics, they rarely provide dedicated space on the concept of trust itself. The limited availability of teaching materials in the literature that translate trust theory into practical, pedagogical formats suggests that this gap extends beyond our local context. In response, and in recognition of the growing societal significance of trust, we integrated applied case studies on trust into university teaching within the Digital Society Initiative PhD Excellence Program at the University of Zurich (UZH) [20]. A robust trust analysis requires a clear conceptual understanding of trust, as well as insight into the mechanisms that foster or undermine it in specific contexts. Relying solely on trust theories can be challenging for students. In disciplines such as ethics, the canvas format has proven effective in simplifying complex concepts. Defined as a "visual tool designed to guide people through the process of using a methodology or framework" [21], the canvas format has shown value in teaching and in fostering problem-solving and critical thinking [22,23]. To date, we have used ethics-focused tools that share some conceptual overlap with trust [21,24-26]. While these tools have proven useful for teaching and for guiding students through an analysis process, they were designed for ethical analysis specifically and therefore serve a different purpose.

In response to the lack of teaching aids on trust, this paper introduces the first Trust Analysis Canvas for Teaching (TACT), a tool designed to guide students in conducting trust analyses of case studies in digital public health and medicine. The objective was to develop a canvas that equips students from diverse disciplinary backgrounds and academic levels with skills relevant to their future careers by analyzing key societal challenges—ranging from the erosion of institutional trust to the strengthening of interpersonal relationships, including the doctor-patient relationship, and by enabling them to grasp high-level concepts of trust, navigate the complexities of trust analysis, and apply these insights effectively to real-world case studies in digital public health and medicine.

Methods

Development of TACT

The development of TACT followed a design science research approach, in which knowledge is generated through the iterative building and evaluation of purposeful artifacts [27]. We followed the 6 design science research activities between May 2024 and October 2024, progressing from the initial conceptualization of the canvas to the translation of its final version into multiple languages. Each activity is described in the corresponding sections of this paper.

Activity 1: Problem Identification and Motivation

As outlined in the introduction, we identified a clear gap in existing teaching resources on trust and articulated the motivation for developing a structured tool to support trust analysis in educational settings.

Activity 2: Definition of Objectives

As stated in the aims of the study, the objectives of TACT were to develop a tool that would enable students to conduct structured trust analyses of case studies in digital public health and medicine.

Activity 3: Design and Development

FZ and FG developed the first version of TACT by integrating the "public trust in health data sharing" framework [28] and the "public trust in the health care system" framework [29] with theories that conceptualize trust as a relational, context-dependent, and inherently complex concept [1,5,30,31]. The conceptualization process was iterative: key trust dimensions from the selected frameworks were systematically reviewed, assessed against established trust theories, and translated into guiding questions for practical application in teaching contexts. Although the dimensions were initially derived from health-related frameworks, care was taken throughout the process to ensure they remained discipline-independent, allowing for their application across a broad range of case studies beyond the digital and public health domains. Once a preliminary set of 7 dimensions for conducting a trust analysis had been identified, FZ and FG tested the draft canvas internally using a fictitious case study, which led to an initial refinement. The canvas was then reviewed in writing by senior trust researchers, purposively selected for their expertise, and recruited via email by FG, to assess the conceptual robustness of the content. Their feedback informed further revisions, resulting in a refined second version (Multimedia Appendix 1) that was used for subsequent student testing and development.

Activities 4 and 5: Demonstration and Evaluation

Following the pilot testing within the research group, FZ and FG conducted 3 English-language focus groups involving BSc, MSc, and PhD students from different disciplinary backgrounds. Focus groups were chosen to enable in-depth exploration of participants' perceptions, capture interactive discussions on how TACT was interpreted and applied, and generate richer insights than individual interviews for its evaluation [32]. Reporting of the qualitative study followed the COREQ

XSL•FO
RenderX

(Consolidated Criteria for Reporting Qualitative Research) guidelines (Checklist 1). PhD students from the UZH were purposively sampled based on their research areas; MSc students from the AI in Medicine program at the University of Bern participated as part of a course on Ethical and Legal issues in AI; and BSc students were recruited via UZH mailing lists. The canvas was tested both in person (with MSc students) using its printed version and online (with BSc and PhD students) using its digital version. Following each focus group, it was refined to incorporate participant feedback. The MSc cohort engaged with the canvas through a guided learning approach [33], while the BSc and PhD cohorts used a problem-based learning approach [34]; both methods were used in conjunction with digital public health and medicine case studies. The case study method was chosen for its capacity to promote critical thinking, conceptual integration, and collaboration [35]. The guided learning approach provides a structured learning experience through a strong conceptual foundation delivered by the lecturer [33]. The MSc focus group was held in person at the University of Bern (duration: 1 h 45 min). Students first received an introduction to the research objectives and an in-depth explanation of the concept of trust, then worked in small groups (2 - 3 participants) to apply the canvas to a digital health case study (Multimedia Appendix 2), and participated in a debriefing plenary discussion. Two online focus groups (45 min each) were conducted via Microsoft Teams (version 4.19.82.0) with BSc and PhD students, using a problem-based learning approach, which is recognized for fostering deep conceptual understanding [34]. Students were given a brief digital public health case study and asked to engage with the canvas to conduct a trust analysis without prior theoretical instruction on trust. The case studies were tailored in complexity to suit each academic level (Multimedia Appendices 3 and 4). After 15 minutes of individual analysis, plenary discussions were held during which lecturers introduced the theoretical foundations of trust, facilitated the interpretation of the canvas, and gathered feedback and suggestions for improvement. FZ took notes during the focus group on the feedback received. The demonstration and evaluation phase concluded after focus groups with students from 3 academic levels were completed within the available timeframe. The consistency of feedback across groups indicated that further data collection was unlikely to yield additional value, giving us confidence that the key points requiring refinement had been adequately captured [36]. CMB, with expertise in graphic design, refined the canvas layout after each focus group based on student feedback, enhancing its visual intuitiveness and ensuring an engaging, user-friendly design. To overcome linguistic barriers and enhance both usability and applicability across countries, the canvas was translated from English into French, Italian, German, and Spanish. Each translation was carried out by 2 native speakers per language, all fluent in English.

Activity 6: Communication and Dissemination

Communication and dissemination of the problem, the artifact, and evidence of its use to researchers and practitioners are achieved through the publication of this tutorial.

Ethical Considerations

According to the Data Protection/Ethics Self-Assessment Tool of the University of Zurich's Digital Society Initiative (2025-266615), the study does not fall under the Human Research Act and does not require submission to the Cantonal Ethics Committee. All participants provided written informed consent prior to participating in the focus groups, and no personal data were collected or processed during the study. Students received a CHF 20 (US $ 25.99) voucher as a token of appreciation. They were neither taught, supervised, evaluated, nor graded by FZ and FG and were not enrolled in any of their courses, ensuring the absence of any dependency or hierarchical relationship between researchers and participants.

Results

Study Participants

We gathered feedback from 2 trust researchers and 23 students at BSc, MSc, and PhD levels from diverse disciplinary backgrounds, through focus groups conducted both online and in-person, using both guided and problem-based learning approaches (Table 1). Students applied TACT to case studies in digital public health and medicine (Multimedia Appendices 2-4).

Table . Focus group participants (N=23 students).

Focus group	Participants, n (%)	Academic level	Discipline	Method
1	3 (13)	BSc students (University of Zurich)	Art, computational linguistics, and business administration	Online (guided learning approach)
2	14 (60.9)	MSc students (University of Bern)	Artificial intelligence in medicine	In-person (problem-based learning approach)
3	6 (26.1)	PhD students (University of Zurich)	Communication science, education, epidemiology, computer science, mathematics, and law	Online (guided learning approach)

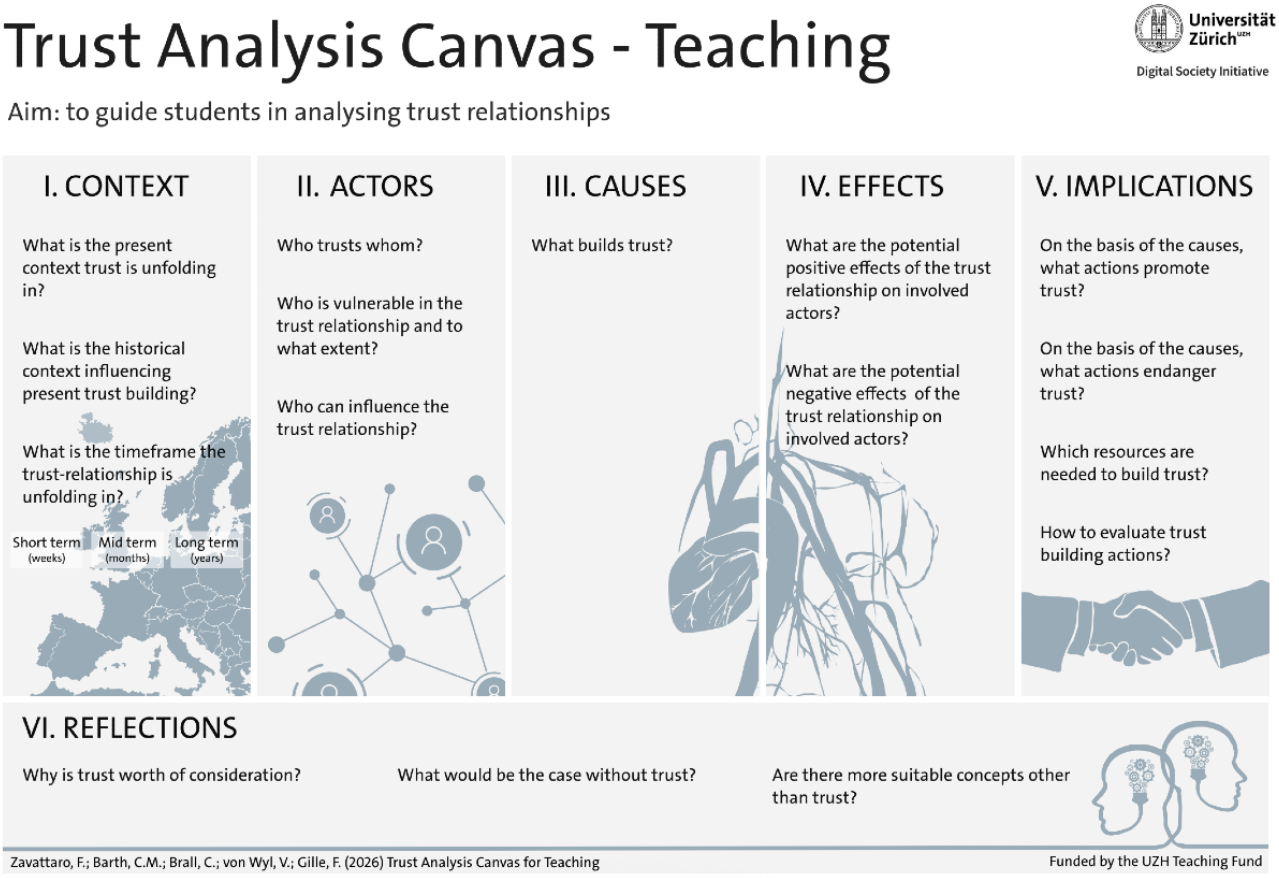
TACT

Feedback from trust researchers and students primarily concerned unclear phrasing and the layout of the tool. A comparison between the initial and final versions of the canvas highlights 5 key revisions (Multimedia Appendix 5). During testing, students reported that TACT effectively supported their understanding of trust under both guided and problem-based learning approaches, as it provided a logical structure for analyzing the various dimensions of trust within the case study.

XSL•FO
RenderX

The final TACT canvas adopts a linear structure comprising 6 dimensions, broken down into a total of 16 guiding questions that deconstruct the overarching concept of trust into smaller, manageable components (Figure 1). French, German, Italian, and Spanish versions of TACT are provided in Multimedia Appendices 6-9, with the downloadable high-quality English version of TACT provided in Multimedia Appendix 10.

Figure 1. Trust Analysis Canvas for Teaching (TACT).



The first dimension, *Context*, reflects the context-specific nature of trust and prompts reflection on the historical and present circumstances shaping the trust relationship. This involves examining the trustor's past experiences, current perceptions, and expectations of the trustee and the temporal frame in which the trust relationship unfolds. The relational nature of trust is addressed in the second dimension, *Actors*, which maps the key stakeholders involved in the trust dynamic—namely, the trustor, the trustee, and any additional actors who may influence trust. This dimension also prompts reflection on potentially vulnerable parties within the trust relationship, highlighting the need for tailored trust-building strategies. The third and fourth dimensions, *Causes* and *Effects*, are informed by the causal and effect dimensions identified in the "public trust in the health care system framework" [29]. These dimensions prompt analysis of the principles that influence the development of trust, as well as the potential positive and negative effects resulting from it. The fifth dimension, *Implications*, highlights that trust does not arise spontaneously; it requires continuous and deliberate efforts to be built and maintained over time. Once the causes and effects of trust are identified, it is necessary to determine the actions needed to promote these trust-building principles, as well as those that could potentially undermine trust, in order to actively safeguard it. This includes allocating the necessary resources and identifying appropriate evaluation methods to assess the effectiveness of the trust-building measures. Finally, *Reflections* prompts students to reflect critically on whether trust is indeed the most appropriate concept to apply in the given case study. This includes evaluating the potential consequences of a lack of trust and, if relevant, considering alternative concepts—such as confidence, reliance, or faith—that may offer a more suitable analytical lens.

Tutorial on TACT

Below is an example of the application of TACT, presented as a written trust analysis of the case study used with MSc students and supported by references to relevant literature. This serves three main purposes as follows: (1) to show the practical application of TACT, (2) to provide readers with an explanation of each dimension and the guiding questions, and (3) to offer lecturers academic resources for teaching trust. The trust analysis below is not intended to be exhaustive or definitive; the case discussion reflects selected classroom exchanges and illustrative reasoning rather than a comprehensive assessment. Its purpose is to demonstrate how the canvas can be applied in practice.

Conducting a Trust Analysis of a Case Study by Using TACT

The case study was sourced from the Digital Society Initiative Strategy Lab 2022, *Level "Now." Artificial Intelligence in*

XSL•FO
RenderX

Medicine—Case Study: "Diagnosis" [37], and it was slightly adapted by FZ and FG for conciseness.

Background

Depression is a major health challenge, impacting up to 20% of the population and potentially leading to suicide (10% - 15%). Diagnosis can be difficult due to similarities with other psychological conditions such as low mood or sadness. Innovative behavioral observation methods aim to identify severe depression or high-risk individuals to prevent suicide. Social network interactions increasingly shape our daily lives, revealing insights into mental health through information consumption, likes, shares, and connections.

DeDe is an app developed by an interdisciplinary consortium of researchers and approved by the cantonal data protection authority, capable of reliably identifying individuals at risk of severe depression based on their social media behavior. Social media providers can integrate DeDe as an option for users, with evaluations kept confidential from the providers. Health insurance may compensate providers for offering DeDe, as early depression diagnosis reduces health care costs, with DeDe financed by insurance premiums.

Case

Alice, prone to depression, discovers DeDe through her primary care physician and downloads the app. During "low" periods, she turns to news about the uncertain world, exacerbating her mood. DeDe notices this and suggests lighter news options, sometimes effective. However, as Alice isolates herself and her mood worsens, DeDe alerts her to seek help from her brother, Peter, listed in the app.

Despite efforts, Alice often ignores recommendations and sinks deeper into depression, as evidenced by her lack of physical activity tracked by DeDe. When her psychiatrist, Sabine, receives an alert, she intervenes, convincing Alice to seek immediate outpatient psychiatric care.

Through therapy and medication, Alice's depression is managed, averting a possible suicide.

1. Context

Trust is context-specific, shaped by past experiences, present perceptions, and future anticipations toward a benefit [5].

What Is the Present Context Trust Is Unfolding In?

Public trust develops in the public sphere through open public discourse on current perceptions of a system's trustworthiness [5]. It is therefore essential to analyze the present context—whether social, economic, and political—in which trust is unfolding. This can be achieved by assessing news narratives, expert opinions, policies, and public perceptions of what is currently being evaluated [38].

Case Study

The present context (Switzerland, 2024) is characterized by the ongoing digital transformation of healthcare, alongside growing societal awareness of the importance of mental health.

What Is the Historical Context Influencing Present Trust Building?

Familiarity and shared past experiences are considered significant determinants of present public trust [5]. The close relationship between past experiences and trust is described by a trust culture arising from "the collective and shared experiences of societal members over time," defining trust culture as "a product of history" [1]. Conferring trust requires consideration of all previous experiences, as trust can only be placed in a familiar world with a reliable background [30]. By collecting news articles, expert opinions, and policy documents from the past, it is possible to trace the evolution of the discourse and identify events that may have positively or negatively influenced current public trust levels [38].

Case Study

Past national scandals involving digital health apps (eg, the MyVaccination scandal in 2021) have shaped public narratives and perceptions on health information sharing, increasing skepticism. There are concerns about potential insurance adjustments based on a mental health diagnosis.

What Is the Timeframe the Trust-Relationship Is Unfolding In? Short-Term (Wks), Mid-Term (Mo), and Long-Term (Y)

Time plays a crucial role in trusting relationships, as trust cannot be rushed [5]. Establishing the timeframe for a trust relationship is essential for effective planning. The duration influences the dynamics of trust development, the factors that determine trustworthiness, and the resilience of the trust relationship over time. Short-term trust often relies heavily on immediate actions and clear communication, whereas long-term trust requires the development of deeper relational bonds. Analyzing the timeframe provides a more nuanced understanding of the strategies required to maintain or restore it.

Case Study

Trust begins when Alice opens up to her doctor and shares her struggles. Following the establishment of trust with her psychiatrist, Alice follows her recommendation, downloads the app, uses its features, and responds to alerts prompting timely interventions (short term, weeks). Sustained trust in the long term (years) among Alice, Sabine, and Peter is crucial, as DeDe supports Alice's ongoing mental health management while maintaining trust in its privacy safeguards, effectiveness, and integration into healthcare systems.

2. Actors

Who Trusts Whom?

Trust is a relational concept where A trusts B to do or not do something. The relational nature of trust means it is influenced by the actions and behaviors of the involved actors [5]. Mapping "who trusts whom" is of fundamental importance to ensure a clear representation of the trust network and to highlight existing dependencies.

Case Study

Alice trusts her psychiatrist, Sabine, to recommend the DeDe app as an appropriate tool for managing her care. She also trusts

XSL•FO
RenderX

her brother Peter to provide support and intervene if she is in distress when alerted by the app. Sabine trusts that the DeDe app was developed in a way that accurately monitors Alice's mental health, detects any deterioration, and notifies her when intervention is necessary. Sabine also trusts Alice to use the app responsibly and Peter to respond appropriately when needed.

Who Is Vulnerable in the Trust Relationship and to What Extent?

Trust arises in contexts characterized by complexity, risk, and vulnerability [5]. In certain circumstances, specific groups may be more vulnerable than others within trust relationships. For instance, individuals with low digital literacy might struggle to engage effectively with the digital health intervention [39].

It is therefore essential to identify who is vulnerable in this context as this helps determine whether trust is the appropriate analytical concept to apply.

Case Study

Alice is particularly vulnerable as, trusting her doctor's advice, she has downloaded the DeDe app to receive effective and timely recommendations. However, this requires her to share highly sensitive personal information, exposing her to potential risks related to data privacy and misuse.

Who Can Influence the Trust Relationship?

Trust is a relational construct that requires at least one other party in whom trust is placed, though it often involves multiple actors [5]. It is crucial not only to map "who trusts whom" but also to identify who can influence the trust relationship. These influencing factors may include individuals, institutions, or broader dynamics that shape the trust relationship.

Case Study

Cantonal data protection authority: By approving DeDe, the cantonal data protection authority reassures users about data security and privacy, fostering institutional trust. As a health care professional, Sabine influences trust through her endorsement and use of DeDe. Peter's role as a trusted contact can influence Alice's use of the app. If he responds positively to alerts, this reinforces Alice's trust in DeDe as a supportive tool.

3. Causes: What Builds Trust?

Familiarity, the presence of an active regulatory system, anonymization of personal data, offering autonomous choices, truthful and honest communication about health system activities, protecting privacy, demonstrating the capacity to fulfill the entrusted action, ensuring data security, certainty about the future, anticipation of a net benefit as an outcome of the trust relationship, and sufficient time for those handling the entrusted data are among the trust-building principles identified in "public trust in the health system" [29] and "public trust in health data sharing" frameworks [28]. These principles are broadly valid within the realms of health data sharing and health system initiatives. However, the specific principles that "build trust" in a given case study must be carefully selected and justified based on the context.

Case Study

Through communication, Sabine provides detailed information on the app's functionalities, as well as the associated risks and benefits of using it, enabling Alice to make an informed decision. Robust data security ensures that Alice feels safe sharing sensitive personal information, addressing concerns about privacy and fostering confidence in the app. The perceived personal benefit of the app to Alice's well-being, such as mood improvement and mental health support, incentivizes her to download DeDe and use the app. All of the above may be necessary to overcome skepticism toward digital health solutions involving the sharing of health data, particularly in light of the MyVaccination scandal.

4. Effects

What Are the Potential Positive Effects of the Trust Relationship on Involved Actors?

Health system legitimization and public acceptance are the results of a strong public-system trust relationship. Moreover, trust reduces uncertainty, streamlines decision-making processes, and promotes long-term partnerships. Additionally, trust may lead to greater efficiency, as it lowers the need for constant oversight and can result in mutual benefits such as shared learning, innovation, and the achievement of common goals [5].

Case Study

The use of the DeDe app facilitates timely intervention, leading to improved mental health, greater stability, and better management of Alice's depression. Trust fosters stronger relationships between Alice, her psychiatrist Sabine, and Peter, enabling coordinated support for her well-being.

What Are the Potential Negative Effects of the Trust Relationship on Involved Actors?

While trust can be beneficial, it may also carry potential risks and negative effects. Over-reliance on trust can make actors vulnerable to breaches or exploitation, particularly when trust is misplaced or taken for granted. This may lead to significant consequences such as reduced vigilance or a lack of accountability. Furthermore, if trust is broken, the resulting loss can damage relationships [5,40].

Case Study

Overreliance on the DeDe app may lead to a decline in direct personal interactions between Alice, her psychiatrist Sabine, and her brother Peter, potentially weakening the human aspect of care and support. If the app fails to identify severe episodes or provide inaccurate recommendations, Alice's condition could worsen, undermining her confidence in digital tools and health care systems.

5. Implications

Trust does not emerge by itself; it requires continuous and targeted efforts to be built and sustained over time. Once the principles that contribute to "building trust" have been identified, it is essential to determine the actions needed to promote these trust-building principles, as well as those that could potentially undermine trust, to actively safeguard it. To achieve this, it is

XSL•FO
RenderX

also necessary to allocate the required resources and identify appropriate evaluation methods to measure whether the trust-building actions have been successful.

On the Basis of the Causes, What Actions Promote Trust?

For instance, if we consider the principle of an "active regulatory system", the presence of legislation, policies, and laws can be seen as trust-promoting activities. Similarly, in the case of the "communication" principle, establishing a targeted communication campaign serves as a viable trust-building activity [28].

Case Study

Providing detailed information about the app's functionalities, as well as its risks and benefits, enables Alice to make an informed decision and fosters trust in the health care provider and confidence in the app. Implementing robust data security measures addresses concerns about privacy, ensuring Alice feels safe and confident sharing sensitive personal information.

On the Basis of the Causes, What Actions Endanger Trust?

On the contrary, a lack of effective communication can undermine trust, as it prevents stakeholders from accessing essential information about an initiative. Information is the lifeblood of trust; therefore, the absence of clear communication and information-sharing jeopardizes trust-building efforts [5].

Case Study

If Sabine fails to provide detailed information about the app's functionalities, as well as its risks and benefits, Alice may feel uncertain and could decide not to download the app. Insufficient protection of sensitive personal information can raise concerns about privacy breaches, undermining Alice's confidence in the app. If the app fails to deliver the expected benefits, her confidence in the app and trust in the health care professional may diminish.

Which Resources Are Needed to Build Trust?

Based on the causes of trust, a variety of resources are required for implementation. These include time, financial investment, and human resources, among others.

Case Study

Sufficient funding is needed to ensure regular app updates, data security measures, ethical evaluations, and to develop educational materials, outreach programs, and communication strategies.

How to Evaluate Trust-Building Actions?

Research indicates that the evaluation and measurement of trust remain challenging tasks, requiring methodological creativity [41]. Nevertheless, evaluation, including robust measurement, is critical to understanding a health system's capacity to build and sustain trust. A common method for collecting trust-related data is the use of survey instruments [42] complemented by qualitative indicators (eg, user testimonials) or by monitoring the expected effects of trust, such as participation or engagement with a digital solution.

Case Study

Indicators of user engagement, such as the frequency with which users interact with the app, respond to alerts, or act on its recommendations, may serve as proxies for trust.

6. Reflections

On the basis of the analysis conducted in points 1 to 5, students are asked to reflect on whether trust is ultimately the appropriate concept to apply to the given case study. They should evaluate the consequences of a lack of trust and, if necessary, propose alternative concepts that may be more suitable (eg, confidence, reliance, or faith).

Why Is Trust Worthy of Consideration?

Case Study

Trust fosters collaboration between users (Alice), family members (Peter), and professionals (Sabine), ensuring that the support network functions effectively, while confidence in data security and privacy facilitates Alice's use of the app.

What Would Be the Case Without Trust?

Case Study

Without trust, Alice might not have followed her doctor's recommendation, downloaded the app, or shared her data, thereby reducing the feasibility and effectiveness of early detection and intervention and increasing the risk of crisis escalation or suicide.

Are There More Suitable Concepts Other Than Trust?

Case Study

Trust is a relevant concept in the relationship between Alice, Sabine, and Peter. When it comes to downloading and using the app, the concept of confidence may be more accurate, as Alice needs to be confident that DeDe will perform as expected, keep her data secure, and maintain privacy.

Discussion

To our knowledge, TACT is the first canvas specifically designed to support students across disciplines and academic levels in analyzing trust-related issues in digital public health and medicine, by bridging theory and practice, and facilitating the application of conceptual knowledge to real-world scenarios. Feedback collected from students indicates that TACT met its core objective: to provide a structured and accessible framework that enables students to understand complex concepts of trust, simplify trust analysis, and apply these insights to practical case studies—thereby enhancing both teaching practices and student learning outcomes. Its adaptability to different pedagogical approaches (guided and problem-based learning) and learning environments (in-person and online), as well as its suitability for use in both digital and printed formats, further reinforces its value as a teaching tool. TACT offers educators a structured framework for teaching trust and provides students with a practical, engaging tool for conducting trust analyses of digital public health and medicine case studies.

XSL•FO
RenderX

During testing, students reported that TACT effectively fostered an understanding of trust across both teaching approaches by logically guiding them in considering different aspects of trust. When applied within the guided learning approach, TACT offered a well-informed and theoretically grounded framework, enabling students to engage with the canvas based on prior conceptual input. However, this approach may risk fostering a more passive learning experience, as students tend to depend on predelivered content rather than actively constructing their own understanding. Conversely, the problem-based learning approach encourages critical thinking and problem-solving skills, potentially leading to deeper engagement through self-directed exploration. However, its effectiveness relies on a carefully designed and time-intensive debriefing process to clarify conceptual ambiguities and consolidate learning. These observations align with Kirschner et al's [33] argument that guided instruction becomes less necessary when learners possess sufficient prior knowledge to engage with material independently. Consequently, problem-based learning may be more suitable for advanced students, such as MSc and PhD candidates, who possess stronger critical and independent thinking skills. In contrast, undergraduate students may benefit more from guided learning, which offers additional structure and foundational knowledge to support their learning process.

While the use of canvases has been shown to facilitate interactive and participatory learning, fostering active problem-solving, critical thinking, deeper understanding, and long-term retention of course material [23], literature highlights the need to address potential challenges related to content accessibility and readability. Students with dyslexia, for instance, may struggle with complex text structures, underscoring the importance of clear, structured, and visually accessible materials [43]. Similarly, non-native English speakers may encounter difficulties with academic texts that use advanced vocabulary or idiomatic expressions, potentially hindering their learning process [44]. Additionally, variations in academic preparedness across students further complicate this issue, as learners from diverse educational backgrounds may have different levels of experience with critical reading and text analysis, necessitating materials that accommodate a broad spectrum of reading abilities [23]. To enhance clarity, accessibility, and usability, TACT was structured as linearly as possible, minimizing lengthy text and prioritizing plain English. It was tested with both native and non-native English speakers across academic levels, from undergraduate to PhD students, ensuring that its content and structure were clear, logical, and intuitive. Additionally, the color palette was selected to be suitable for color-blind users, and icons were incorporated as complementary visual aids rather than essential informative elements to reduce the risk of misinterpretation. While these refinements improve accessibility, lecturers may still need to make ad hoc adjustments to accommodate individual student needs.

The use of case studies in education aligns with constructivist theory and follows the tradition of active and experiential learning, supporting students in developing critical thinking and decision-making skills alongside domain-specific knowledge [45-47]. While TACT is grounded in the authors' conceptual research on trust in digital health and policy and has thus far been tested only on digital public-health case studies, the deliberate effort to ensure its dimensions are discipline-independent, combined with the canvas's flexibility in case selection, supports its broader relevance and adaptability. As TACT could guide the analysis of trust in case studies from diverse domains, it offers lecturers the flexibility to tailor its use by selecting cases aligned with their students' learning objectives. Students may likewise apply TACT to their own research projects, thereby enhancing engagement and learning impact [48]. We encourage future research to test the applicability of TACT beyond case studies on digital public health and medicine, and in other geographical regions to further demonstrate its transferability and adaptability. Moreover, assessing the durability of students' conceptual understanding of trust and their ability to apply it across different contexts over time represents a valid direction for future research.

The primary goal of TACT is not to provide definitive answers but to stimulate critical thinking and discussion on trust. To achieve this, we deliberately avoided checklists or predefined answers, encouraging students to explore multiple perspectives and identify the most appropriate answers for their specific case studies. Lecturers can use the literature provided to deepen their understanding of trust and its underlying principles. The flexibility of TACT extends across disciplines, academic levels, and teaching formats. It can be used for both high-level and in-depth trust analyses, ranging from a quick exploratory tool that prompts consideration of various trust-related aspects in a case study to a systematic framework for comprehensive trust analysis, where each of the 6 dimensions can be explored in dedicated lectures, allowing for an in-depth study of the theoretical foundations. The decision on how to integrate TACT into teaching remains at the discretion of lecturers, who can adapt its use to align with students' curricula and academic needs.

Acknowledgments

The authors thank Georg Starke (Technical University of Munich, Germany) and Benjamin Ho (Vassar College, the United States) for reviewing the initial version of the canvas and proposing improvements to strengthen its conceptual robustness. Special thanks to Giulia Mina and Sara Messina (Italian), Giulia Oliveri and Mathilde Le Vu (French), Judit Martinez Moreno and Javier Munoz Laguna (Spanish), and André Gille and Katrin Gille (German) for their support in translating Trust Analysis Canvas for Teaching. The authors are grateful to all study participants for their valuable feedback, which helped refine the canvas, and to the University of Zurich Teaching Fund for funding this project. The authors attest that there was no use of generative artificial intelligence technology in the generation of text, figures, or other informational content of this manuscript.

XSL•FO
RenderX

Funding

This project was supported by the University of Zurich Teaching Fund. In addition, FZ and FG receive funding from the Digital Society Initiative at the University of Zurich. FG also receives support from the Swiss Academy of Engineering Sciences and the World Health Organization. FG is concurrently employed at the Federal Chancellery of Switzerland; however, the views expressed in this paper are solely those of the author and do not represent the position of the Federal Chancellery.

Data Availability

Anonymized feedback from the focus group on Trust Analysis Canvas for Teaching is available upon request.

Authors' Contributions

Conceptualization: FZ, FG
Data curation: FZ, FG
Formal analysis: FZ, FG
Funding acquisition: FZ, FG
Investigation: FZ, FG, CB
Methodology: FZ, FG, CB
Project administration: FZ, FG
Resources: FZ, FG
Supervision: FG, VVW
Validation: FG, VVW
Visualization: CMB
Writing – original draft: FZ
Writing – review & editing: CMB, CB, VVW, FG

Conflicts of Interest

None declared.

Multimedia Appendix 1
Initial version of the canvas reviewed by senior researchers.
[DOCX File, 133 KB - mededu_v12i1e79709_app1.docx]

Multimedia Appendix 2
Adapted AI in medicine “Diagnosis” case study used in MSc focus group (Digital Society Initiative Strategy Lab 2022).
[DOCX File, 19 KB - mededu_v12i1e79709_app2.docx]

Multimedia Appendix 3
COVID-19–based case study developed by authors and used in BSc online focus group.
[DOCX File, 18 KB - mededu_v12i1e79709_app3.docx]

Multimedia Appendix 4
Case study developed by authors based on recent Swiss health policy developments and used in PhD online focus group.
[DOCX File, 18 KB - mededu_v12i1e79709_app4.docx]

Multimedia Appendix 5
Comparison of initial and final canvas versions highlighting five key changes.
[DOCX File, 18 KB - mededu_v12i1e79709_app5.docx]

Multimedia Appendix 6
Trust analysis canvas–teaching (French).
[DOCX File, 846 KB - mededu_v12i1e79709_app6.docx]

Multimedia Appendix 7
Trust analysis canvas–teaching (German).
[DOCX File, 635 KB - mededu_v12i1e79709_app7.docx]

Multimedia Appendix 8

XSL•FO
RenderX

Trust analysis canvas–teaching (Italian).
[DOCX File, 849 KB - mededu_v12i1e79709_app8.docx]

Multimedia Appendix 9
Trust analysis canvas–teaching (Spanish).
[DOCX File, 896 KB - mededu_v12i1e79709_app9.docx]

Multimedia Appendix 10
Trust analysis canvas–teaching (TACT) (English).
[PDF File, 1184 KB - mededu_v12i1e79709_app10.pdf]

Checklist 1
COREQ checklist.
[PDF File, 818 KB - mededu_v12i1e79709_app11.pdf]

References

1. Sztompka P. Trust: A Sociological Theory, 1st edition: Cambridge University Press; 1999.
2. McKee M, Greenley R, Permanand G, European Observatory on Health Systems and Policies. Trust: the foundation of health systems. : World Health Organization; 2023 URL: https://iris.who.int/server/api/core/bitstreams/9ff3032f-c584-4901-93d4-d7b3e026ff38/content [accessed 2026-01-17]
3. Trust and public policy: how better governance can help rebuild public trust. : Organisation for Economic Co-Operation and Development; 2017 URL: https://www.oecd.org/content/dam/oecd/en/publications/reports/2017/03/trust-and-public-policy_g1g74ea6/9789264268920-en.pdf [accessed 2026-01-17]
4. Gille F, Brall C. Public trust: caught between hype and need. Int J Public Health 2020 Apr;65(3):233-234. [doi: 10.1007/s00038-020-01343-0] [Medline: 32114636]
5. Gille F. What Is Public Trust in the Health System?: Insights into Health Data Use: Policy Press; 2023. [doi: 10.51952/9781447367352]
6. Iyamu I, Xu AXT, Gómez-Ramírez O, et al. Defining digital public health and the role of digitization, digitalization, and digital transformation: scoping review. JMIR Public Health Surveill 2021 Nov 26;7(11):e30399. [doi: 10.2196/30399] [Medline: 34842555]
7. van Haasteren A, Gille F, Fadda M, Vayena E. Development of the mHealth app trustworthiness checklist. Digit Health 2019;5:2055207619886463. [doi: 10.1177/2055207619886463] [Medline: 31803490]
8. Papadopoulos K, von Wyl V, Gille F. What is public trust in national electronic health record systems? A scoping review of qualitative research studies from 1995 to 2021. Digit Health 2024;10:20552076241228024. [doi: 10.1177/20552076241228024] [Medline: 38288130]
9. Deml MJ, Buhl A, Huber BM, Burton-Jeangros C, Tarr PE. Trust, affect, and choice in parents' vaccination decision-making and health-care provider selection in Switzerland. Sociol Health Illn 2022 Jan;44(1):41-58. [doi: 10.1111/1467-9566.13388] [Medline: 34747500]
10. Siegrist M, Bearth A. Worldviews, trust, and risk perceptions shape public acceptance of COVID-19 public health measures. Proc Natl Acad Sci U S A 2021 Jun 15;118(24):e2100411118. [doi: 10.1073/pnas.2100411118] [Medline: 34045360]
11. Mpinganjira M. Use of e-government services: the role of trust. Int J Emerg Markets 2015 Sep 21;10(4):622-633. [doi: 10.1108/IJoEM-12-2013-0151]
12. Yap KB, Wong DH, Loh C, Bak R. Offline and online banking—Where to draw the line when building trust in e‐banking? Int J Bank Mark 2010 Feb 2;28(1):27-46. [doi: 10.1108/02652321011013571]
13. Starke G, Gille F, Termine A, et al. Finding consensus on trust in AI in health care: recommendations from a panel of international experts. J Med Internet Res 2025 Feb 19;27:e56306. [doi: 10.2196/56306] [Medline: 39969962]
14. Cross J, Kayalackakom T, Robinson RE, et al. Assessing ChatGPT's capability as a new age standardized patient: qualitative study. JMIR Med Educ 2025 May 20;11:e63353. [doi: 10.2196/63353] [Medline: 40393017]
15. Abouammoh N, Alhasan K, Aljamaan F, et al. Perceptions and earliest experiences of medical students and faculty with ChatGPT in medical education: qualitative study. JMIR Med Educ 2025 Feb 20;11:e63400. [doi: 10.2196/63400] [Medline: 39977012]
16. Mackenzie C. Vulnerability, insecurity and the pathologies of trust and distrust. Int J Philos Stud 2020 Oct 19;28(5):624-643. [doi: 10.1080/09672559.2020.1846985]
17. Brennan J. Trust as a test for unethical persuasive design. Philos Technol 2021 Dec;34(4):767-783. [doi: 10.1007/s13347-020-00431-6]
18. Regulation (EU) 2016/679 of the European parliament and of the council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing directive

XSL•FO
RenderX

95/46/EC (General Data Protection Regulation). : European Union; 2016 URL: https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32016R0679 [accessed 2026-01-17]
19. Academic programs. UZH - Universität Zürich. URL: https://www.uzh.ch/en/studies/programs.html [accessed 2026-01-17]
20. Excellence program (PhD). UZH - Universität Zürich. URL: https://www.dsi.uzh.ch/en/education/excellence-program.html [accessed 2026-01-17]
21. Hardebolle C, Macko V, Ramachandran V, Holzer A, Jermann P. Digital ethics canvas: a guide for ethical risk assessment and mitigation in the digital domain. Presented at: Proceedings of the 51st Annual Conference of the European Society for Engineering Education (SEFI 2023); Sep 11-14, 2023 URL: https://zenodo.org/records/17022248 [accessed 2026-01-17]
22. Wood EJ. Problem-based learning: exploiting knowledge of how people learn to promote effective learning. Biosci Educ 2004 May;3(1):1-12. [doi: 10.3108/beej.2004.03000006]
23. Oudat Q, Othman M. Embracing digital learning: benefits and challenges of using canvas in education. J Nurs Educ Pract 2024;14(10):39. [doi: 10.5430/jnep.v14n10p39]
24. Ferretti A. Ethics and governance of big data in health research and digital health applications [Dissertation]. : ETHZurich: Research Collection; 2021 URL: https://www.research-collection.ethz.ch/entities/publication/d5075693-1fbb-4586-adeb-97c4d4cc52a3 [accessed 2026-01-17]
25. Franzke AS, Muis I, Schäfer MT. Data Ethics Decision Aid (DEDA): a dialogical framework for ethical inquiry of AI and data projects in the Netherlands. Ethics Inf Technol 2021 Sep;23(3):551-567. [doi: 10.1007/s10676-020-09577-5]
26. The data ethics canvas. Open Data Institute. 2024. URL: https://theodi.org/insights/tools/the-data-ethics-canvas-2021/ [accessed 2026-01-17]
27. Peffers K, Tuunanen T, Rothenberger MA, Chatterjee S. A design science research methodology for information systems research. J Manage Inf Syst 2007 Dec;24(3):45-77. [doi: 10.2753/MIS0742-1222240302]
28. Zavattaro F, von Wyl V, Gille F. Examining the inclusion of trust and trust-building principles in European Union, Italian, French, and Swiss health data sharing legislations: a framework analysis. Milbank Q 2024 Dec;102(4):973-1003. [doi: 10.1111/1468-0009.12722] [Medline: 39629669]
29. Gille F, Smith S, Mays N. What is public trust in the healthcare system? A new conceptual framework developed from qualitative data in England. Soc Theory Health 2021 Mar;19(1):1-20. [doi: 10.1057/s41285-020-00129-x]
30. Luhmann N. Trust and Power: Two Works: Wiley; 1979.
31. Hardin R. Trust and Trustworthiness: Russell Sage Foundation; 2002.
32. Pope C, Mays N. Qualitative Research in Health Care, 4th edition: John Wiley & Sons; 2020. [doi: 10.1002/9781119410867]
33. Kirschner PA, Sweller J, Clark RE. Why minimal guidance during instruction does not work: an analysis of the failure of constructivist, discovery, problem-based, experiential, and inquiry-based teaching. Educ Psychol 2006 Jun;41(2):75-86. [doi: 10.1207/s15326985ep4102_1]
34. Duch BJ, Groh SE, Allen DE. The Power of Problem-Based Learning: A Practical 'How to' for Teaching Undergraduate Courses in Any Discipline, 1st edition: Stylus Publishing; 2001.
35. Widiandari LA, Redhana IW. Case studies in improving students' learning outcomes, is it effective? J Phys Conf Ser 2021 Mar 1;1806(1):012186. [doi: 10.1088/1742-6596/1806/1/012186]
36. Green J, Thorogood N, Melendez-Torres GJ. Qualitative Methods for Health Research, 5th edition: Sage Publications; 2025.
37. Christen M, Biller-Andorno N, Hastings J, Iqbal JD, Kölzer V, Krauthammer M, et al. Artificial intelligence in medicine—case study «diagnosis». : University of Zurich - Digital Society Initiative; 2022 URL: https://dsi.uzh.ch/dam/jcr:71d9fe5c-7e1a-426d-bf3a-97c2174a0753/StrategyLab2022_Case-Diagnosis_en.pdf [accessed 2026-01-17]
38. Daniore P, Zavattaro F, Gille F. Public trust in a Swiss health data space. : Universität Zürich: Digital Society Initiative; 2024 URL: https://health.dsi.uzh.ch/app/uploads/2024/03/Daniore-Zavattaro-Gille-2024-Public-Trust-in-a-Swiss-Health-Data-Space.pdf [accessed 2026-01-17]
39. Arias López MDP, Ong BA, Borrat Frigola X, et al. Digital literacy as a new determinant of health: a scoping review. PLOS Digit Health 2023 Oct;2(10):e0000279. [doi: 10.1371/journal.pdig.0000279] [Medline: 37824584]
40. Krot K, Rudawska I. The role of trust in doctor-patient relationship: qualitative evaluation of online feedback from Polish patients. Econ Sociol 2016 Sep;9(3):76-88. [doi: 10.14254/2071-789X.2016/9-3/7]
41. Taylor LA, Nong P, Platt J. Fifty years of trust research in health care: a synthetic review. Milbank Q 2023 Mar;101(1):126-178. [doi: 10.1111/1468-0009.12598] [Medline: 36689251]
42. Ozawa S, Sripad P. How do you measure trust in the health system? A systematic review of the literature. Soc Sci Med 2013 Aug;91:10-14. [doi: 10.1016/j.socscimed.2013.05.005] [Medline: 23849233]
43. McCarthy JE, Swierenga SJ. What we know about dyslexia and web accessibility: a research review. Univ Access Inf Soc 2010 Jun;9(2):147-152. [doi: 10.1007/s10209-009-0160-5]
44. Nguyen CH, Nguyen NX. Rethinking (non)nativeness among English-speaking teachers in Vietnam. Reflections 2023;30(2):574-589. [doi: 10.61508/refl.v30i2.267528]
45. Auster ER, Wylie KK. Creating active learning in the classroom: a systematic approach. J Manag Educ 2006 Apr;30(2):333-353. [doi: 10.1177/1052562905283346]

XSL•FO
RenderX

46. Kim S, Phillips WR, Pinsky L, Brock D, Phillips K, Keary J. A conceptual framework for developing teaching cases: a review and synthesis of the literature across disciplines. Med Educ 2006 Sep;40(9):867-876. [doi: 10.1111/j.1365-2929.2006.02544.x] [Medline: 16925637]
47. Cox S. Learning and teaching guides: case studies for active learning. AdvanceHE. 2014. URL: https://www.advance-he.ac.uk/knowledge-hub/learning-and-teaching-guides-case-studies-active-learning [accessed 2026-01-17]
48. Toogood C. Supporting students to engage with case studies: a model of engagement principles. Educ Rev 2025 Jun 7;77(4):1125-1139. [doi: 10.1080/00131911.2023.2281227]

Abbreviations

AI: artificial intelligence
COREQ: Consolidated Criteria for Reporting Qualitative Research
TACT: Trust Analysis Canvas for Teaching
UZH: University of Zurich

Edited by T Leung; submitted 26.Jun.2025; peer-reviewed by C Weerth, M Gong, S Meister; revised version received 15.Dec.2025; accepted 30.Dec.2025; published 17.Feb.2026.

Please cite as:
Zavattaro F, Barth CM, Brall C, von Wyl V, Gille F
Trust Analysis Canvas for Teaching in the Field of Digital Public Health and Medicine: Tutorial
JMIR Med Educ 2026;12:e79709
URL: https://mededu.jmir.org/2026/1/e79709
doi:10.2196/79709

© Federica Zavattaro, Clara-Maria Barth, Caroline Brall, Viktor von Wyl, Felix Gille. Originally published in JMIR Medical Education (https://mededu.jmir.org), 17.Feb.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Education, is properly cited. The complete bibliographic information, a link to the original publication on https://mededu.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Tutorial

Developing and Integrating Virtual Reality Courses in Medical Education: Tutorial and Implementation Guideline Informed by Best Practices From the National Project “medical tr.AI.ning”

Marvin Mergen[1], MD; Anna Junga[2], MD; Henriette Schulze[2]; Philipp Bozdere[3]; Pascal Kockwelp[4]; Leon Pielage[4]; Corbin Sassen[5]; Tina Glückselig[3]; Michael Schmitz[5], Dr; Kathrin Ungru[6], Prof Dr; Norbert Graf[1], Prof Dr Med; Benjamin Risse[4], Prof Dr; Marcel Meyerheim[1]

[1]Department of Pediatric Oncology and Hematology, Faculty of Medicine, Saarland University, Homburg, Germany

[2]Institute for Education and Student Affairs, University of Münster, Münster, Germany

[3]Institute for Society and Digital Media, Münster School of Design, FH Münster, Münster, Germany

[4]Institute for Geoinformatics, University of Münster, Münster, Germany

[5]xm:lab - experimental media lab, Hochschule der Bildenden Künste Saar, Saarbrücken, Germany

[6]Department of Electrical Engineering and Computer Science, Laboratory of Visual Computing, FH Münster, Münster, Germany

Corresponding Author:
Marvin Mergen, MD
Department of Pediatric Oncology and Hematology
Faculty of Medicine
Saarland University
Kirrberger Strasse 100
Homburg, 66421
Germany
Phone: 49 684116 28411
Email: kontakt@dr-mergen.de

Abstract

The increasing adoption of virtual reality (VR) in medical education offers substantial opportunities for immersive, practice-oriented training that complements traditional teaching methods. In particular, VR enables repeated, risk-free exposure to complex clinical scenarios and supports the development of clinical reasoning, communication skills, and procedural competence. However, implementing VR-based courses remains challenging due to high development costs, technical complexity, and the need for close interdisciplinary collaboration. This tutorial presents key insights and best practices from the medical tr.AI.ning project, a 3-year interdisciplinary initiative funded by the German Federal Ministry of Education and Research. The project’s objective was to develop an artificial intelligence (AI)-supported, VR-based training platform that allows medical students to practice clinical decision-making in immersive, interactive scenarios. The paper is structured as a tutorial and offers recommendations for planning, developing, and integrating VR courses into medical curricula. Each recommendation is illustrated with concrete examples from our project, serving as a practical blueprint to guide educators and developers in applying these guidelines in their own contexts. Successful implementation of a VR project in medical education requires strategic planning and collaboration, starting with a thorough identification of curricular gaps that VR can address and a clear justification of its added educational value. An interdisciplinary consortium that combines expertise from medical didactics experts, computer science, and design is essential to ensure the development of high-quality, pedagogically sound simulations and intuitive user interfaces. Key factors for success include defining specific learning objectives aligned with competency-based frameworks; iterative development with continuous feedback from medical experts, educators, and students; and structured pilot testing with systematic collection of quantitative and qualitative data to assess usability, immersion, and learning outcomes. Early engagement and walkthroughs with end users help identify practical challenges and inform iterative improvements. A dedicated authoring tool within the project allows medical teachers to create and adapt VR scenarios without prior technical experience, supporting the scalability and sustainability of the approach. Effective project management frameworks facilitate collaboration, clear task allocation, and adaptive progress throughout development. Additionally, considerations for hardware selection, technical infrastructure, and sustainable dissemination strategies, including open-access publications, project websites, and professional networking, are crucial to ensure long-term viability and broad adoption across institutions. By combining a tutorial format with practical, step-by-step recommendations, this article

XSL•FO
RenderX

provides a comprehensive guide for educators and developers on implementing immersive, AI-supported VR courses to enhance medical education. It highlights key lessons learned in interdisciplinary collaboration, iterative testing, systematic evaluation, and alignment with educational objectives, thereby facilitating the effective, evidence-based, and sustainable integration of VR into medical curricula across diverse institutions.

(JMIR Med Educ 2026;12:e80976) doi:10.2196/80976

KEYWORDS

virtual reality; project management; artificial intelligence; interdisciplinary collaboration; tutorial

Introduction

Medical education is increasingly adopting digital learning methods, driven by advancements in computer science, particularly in artificial intelligence (AI). As part of this shift, virtual reality (VR) is emerging as a valuable training tool for both aspiring and practicing health care professionals [1]. Initially, VR was used predominantly for training surgical interventions and procedures [1-3], but its applications are expanding rapidly across a range of other use cases. The motivation of medical educators to invest in and learn to use this new technology derives mainly from distinct advantages, such as the ability to repeat medical procedures in a safe, immersive environment as often as needed, without risk to real patients or users themselves [1,3]. At the same time, VR still encounters certain challenges, including technical limitations in terms of haptic feedback [4,5] and issues related to cybersickness [6,7]. However, the growing interest in using VR for medical education is driving ongoing innovations that may help overcome these current limitations. Additionally, VR's expanding role suggests potential for extended didactic approaches, further supporting its integration into curricula.

This paper is intentionally written as a tutorial informed by the 3-year, multiinstitutional project medical tr.AI.ning [8] (Germany; project start: December 2021). It synthesizes practical implementation lessons (planning, development, integration, and sustainability) rather than reporting a controlled effectiveness trial. While we refer to evaluation activities conducted during the project (eg, pilot courses and feasibility studies), the purpose of these activities in this tutorial is to inform implementation decisions, not to claim that VR is superior to other teaching formats.

How to Use This Tutorial

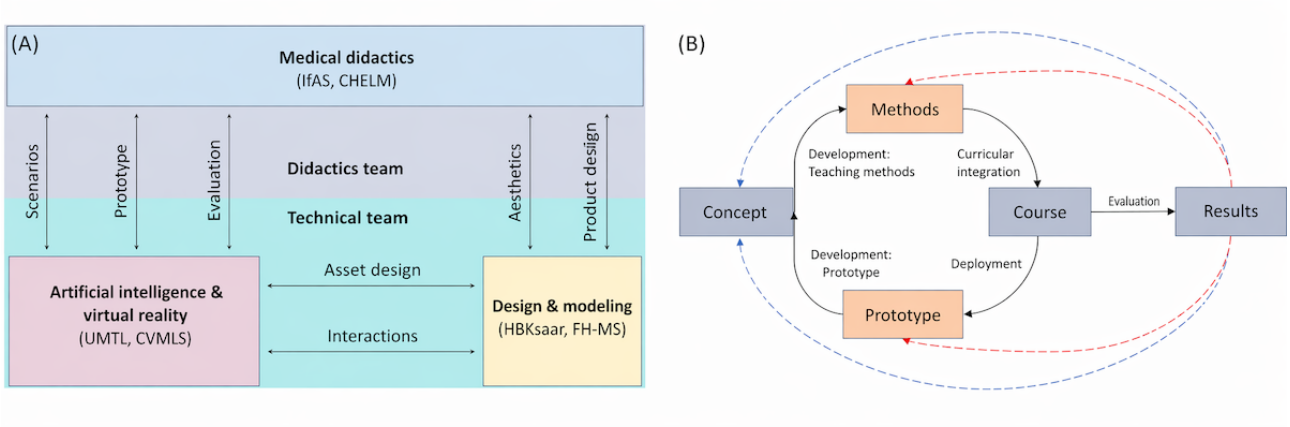
The guidance is organized into 4 phases that readers can follow sequentially or consult as needed:

- Phase 1: How to start (needs assessment, consortium setup, goals, resources).
- Phase 2: How to organize (communication, project management, infrastructure).
- Phase 3: How to build (learning design, technical foundation, AI options, interdisciplinary workflows).
- Phase 4: How to integrate (iterative testing, curricular embedding, dissemination, sustainability).

The medical tr.AI.ning project's primary goal is to create an AI-supported, VR-based training platform that enables medical students to practice clinical decision-making in customizable, immersive virtual scenarios. Additionally, a key project component is the development of an authoring tool that provides medical teachers with a visual interface, allowing them to easily create and adapt new training scenarios tailored to specific educational needs. The project draws on contributions from the University of Mu¨nster (Institut für Ausbildung und Studienangelegenheiten [IfAS] and Computer Vision & Machine Learning Systems [CVMLS]), Saarland University (Coordination Center Homburg for Education & Learning in Medicine [CHELM] and Ubiquitous Media Technologies Lab [UMTL]), the University of Applied Sciences Münster (FHMS; Department of Electrical Engineering and Computer Science [ETI], Münster School of Design [MSD], and Institut Gesellschaft und Digitales [GUD]), and the Saarbrücken Academy of Fine Arts (xm:lab and Hochschule der Bildenden Künste Saar [HBKsaar]). Responsibilities and competencies are broadly divided into 2 core teams (see Figure 1). The medical didactics team (IfAS and CHELM) is responsible for developing teaching methods and strategies, providing the medical foundation necessary for programming and designing scenarios in collaboration with medical specialists, and deploying the developed content into medical curricula. The design and technical team (UMTL, CVMLS, FHMS, xm:lab, and HBKsaar) handles platform implementation, including design, modeling, animations, and the creation of a unified interaction concept for VR and the authoring tool. While tasks are clearly defined, close interdisciplinary collaboration is essential throughout the project life cycle.

XSL•FO
RenderX

Figure 1. (A) Consortium structure and broadly defined responsibilities within the medical tr.AI.ning project; (B) Prototype life cycles. CHELM: Coordination Center Homburg for Education & Learning in Medicine; CVMLS: Computer Vision & Machine Learning Systems; FHMS: University of Applied Sciences Münster; HBKsaar: Hochschule der Bildenden Künste Saar; IfAS: Institut für Ausbildung und Studienangelegenheiten; UMTL: Ubiquitous Media Technologies Lab.



In this paper, we synthesize key insights from this 3-year project to enhance transparency in university-led research initiatives and to derive structured, practice-oriented recommendations.

Existing publications on VR in medical education predominantly focus on single educational interventions, technical feasibility, usability, or learning outcomes, often within narrowly defined curricular contexts [9-11]. While several reviews and framework papers discuss general design considerations for immersive learning, comprehensive end-to-end guidance on how to systematically plan, implement, integrate, and sustain VR courses within real-world medical curricula remains scarce [1,12,13].

In particular, operational aspects such as consortium building, project management structures, interdisciplinary coordination, and long-term curricular integration are rarely addressed in a consolidated manner. This tutorial addresses this gap by providing a step-by-step, implementation-oriented guideline informed by a national, multiinstitutional project. Drawing on 3 years of iterative development, deployment, and evaluation across multiple medical faculties, this paper offers practical, reproducible guidance covering the full life cycle of VR course development—from needs assessment and team setup to technical implementation, curricular integration, and dissemination. These recommendations are grounded in practical experience and collaborative processes across multiple institutions.

Methodologically, the guidance is informed by structured workshops with medical educators, iterative usability testing, pilot studies with students, and formative and summative evaluations conducted throughout the project.

In this tutorial, we present general recommendations for developing and integrating VR courses in medical education, and we complement each recommendation with concrete examples from the medical tr.AI.ning project. These examples serve as a practical blueprint, illustrating how the guidelines can be applied in a real-world VR development context.

Feedback from medical experts, teachers, and learners was systematically incorporated into successive project phases and directly shaped the procedural steps outlined in this article. The resulting tutorial, therefore, reflects practice-based, empirically informed recommendations, providing both conceptual guidance and operational examples for future implementations.

Phase 1: How to Start

Overview

This phase focuses on the initial conceptual stage of VR course development in medical education. It outlines how to identify educational needs, define realistic project goals, and establish the interdisciplinary foundations required for a successful VR project.

Defining Gap: Identifying Gaps in Medical Education That Benefit From VR

Identifying curricular gaps before integrating VR into medical education is crucial to ensure that the technology is applied strategically, targeting specific deficiencies in the current curriculum that are difficult or impossible to address with traditional learning methods. In practice, we recommend starting with a detailed review of existing learning objectives and mapping them to potential VR scenarios. This process involves defining a clear rationale for using VR, ensuring that it is selected for its distinct advantages over other teaching methods rather than being adopted purely for novelty.

VR offers immersive, interactive experiences that can simulate complex medical scenarios, such as high-risk procedures or examinations involving intimate areas, which we specifically addressed in the medical tr.AI.ning project. These scenarios are ideal candidates for VR, as they allow students to gain essential skills in a controlled, risk-free environment, which would be difficult or impossible to recreate with traditional methods such as lectures, manikins, real-life settings, or other forms of standardized simulations [14].

In addition, the customizable nature of VR simulations allows users to edit existing training scenarios or create new ones to align with various teaching specializations. This flexibility was a key reason for developing a dedicated authoring tool in our project.

XSL•FO
RenderX

Once the rationale is established, VR should be applied in a well-considered manner that aligns with overarching educational goals. The technology should enhance the learning experience rather than merely replace existing methods without adding benefits. Given that VR requires significant resources for development and implementation, its use must be justified by specific educational needs. The goal is not to replace patient contact or training with actors or simulated patients, but to supplement the curriculum with additional learning scenarios that fill existing educational gaps.

In the medical tr.AI.ning project, we therefore focused on the performance of a whole-body skin cancer screening, a procedure that is challenging for many medical students to practice in real-life settings. In particular, the examination of intimate body areas may be associated with embarrassment for both students and patients, which can limit opportunities for repeated, hands-on training. In addition, VR allows students to repeatedly practice the direct visual differentiation of potentially malignant lesions from benign skin findings under standardized conditions [14]. In real clinical settings, the availability of suitable patients (eg, with melanoma) is neither predictable nor equally distributed across teaching sessions. By embedding representative cases in VR, all students can be exposed to the same educational content and have equal opportunities to practice the required examination and decision-making steps.

Integrating AI

While the use of VR itself offers many benefits, additional value can be provided by AI technology. Real-time speech recognition and synthesis enable more natural, fluid conversations and eliminate the need for students to navigate predefined questions and commands, which might compromise the effectiveness of training in clinical decision-making. AI can enable dynamic, context-aware interactions with virtual agents, supporting adaptive, natural language responses that reflect realistic clinical dialogues [15]. This capability mirrors the unpredictability of real-world scenarios while providing a safe environment for students to practice and refine critical soft skills, such as empathy, active listening, and effective communication.

Furthermore, AI can be used to autonomously generate data protection-compliant clinical images [16]. This not only substantially increases the quantity and diversity of clinical images and their placement within the course, but also enhances the challenge for students to make accurate decisions.

Building the Consortium

Identifying and Recruiting Team Members (Medical Professionals/Didactics, IT Experts, 3D Designers, and Interaction Designers)

The success of an educational VR project heavily relies on assembling a carefully selected, multidisciplinary consortium. Securing expertise in medical didactics, computer science, and design is crucial to developing a comprehensive and impactful learning environment. When forming such a consortium, it is important to choose members with proven expertise in these domains, a strong record of collaborative work, and a shared commitment to educational innovation. Additionally, expertise in AI can substantially enhance the capabilities and outcomes of VR development. It is also important to define how medical specialists will be involved in the development and evaluation process. By ensuring these competencies are well represented, the project establishes a robust foundation for success.

Budget constraints may necessitate adjustments in project setup, such as reducing staffing levels. This can be addressed by focusing on specific goals and scaling back staffing in disciplines with lower workloads. To compensate, leveraging tools, libraries, and off-the-shelf solutions—such as standard user interface (UI) elements, 3D asset libraries, and VR world builders—can streamline development processes. While it is often inadvisable to entirely forgo a discipline, maintaining at least one senior team member to oversee and guide the effective use of these tools helps ensure quality and consistency.

Furthermore, the rapid evolution of generative AI offers new opportunities for innovative workflows in VR development. This advancement has the potential to substantially transform existing processes. Team members should stay actively informed about advancements in AI to identify and effectively integrate these developments into the project.

Defining Roles and Responsibilities

The inherent size and interdisciplinarity of a consortium developing VR simulations for medical education require both a project manager and a project coordinator. The project coordinator fulfills a support role, primarily focusing on administrative tasks such as scheduling, reporting, and coordinating communication among team members. The project coordinator ensures that day-to-day activities align with the project's objectives. By contrast, the project manager holds a leadership role, overseeing the entire project life cycle, including planning, resource allocation, risk management, and strategic decision-making. While the project manager adopts a high-level perspective and makes key decisions, the project coordinator focuses on operational aspects to ensure smooth execution.

Additionally, responsibilities within the teams at each location should be coordinated, ideally through team leaders who work closely on the respective topics.

From our experience, clearly defining responsibilities is especially important in interdisciplinary teams. Splitting the coordination role into technical and didactical (ie, nontechnical) components ensures that all needs and project goals are addressed. These coordinators must maintain constant communication and collaboratively address challenges, escalating issues to the project manager when necessary.

The technical coordinator is responsible for ensuring the technical implementation of the VR project, including the integration of hardware, software, and additional modules such as an authoring tool. This role involves monitoring the build process; diagnosing and resolving technical issues; managing updates and merge requests (ie, developers' proposals to integrate source code changes); and maintaining performance and compatibility standards for the VR simulations.

The didactical coordinator ensures that the educational content and learning objectives are correctly integrated into the VR simulations. Collaborating closely with educators and medical

XSL•FO
RenderX

experts, they align the simulations with current medical curricula, such as the NKLM 2.0 [17], and broader educational frameworks such as competency-based medical education. This role ensures that the simulations are pedagogically sound and meet the intended educational outcomes.

To streamline administrative tasks, an overall coordinator position can be beneficial. In our project, administrative responsibilities were also managed by the didactical coordinator.

Coordinating the technical aspects and workflows of such a multidisciplinary project across multiple sites (in our case, UMTL, CVMLS, FHMS, HBKsaar) is complex. In the medical tr.AI.ning project, the role of a central technical coordinator, as described above, was not planned in the granted project outline. Instead, each site operated under the leadership of a team leader who managed an interdisciplinary team and closely collaborated with the other sites. However, the absence of a central technical coordinator hindered cross-site collaboration and reduced overall project efficiency. It is recommended to establish a central technical coordinator to ensure efficient collaboration between sites and to guarantee the quality of the technical implementation, thereby achieving project goals more effectively. Table 1 suggests the distribution of roles and responsibilities as a RACI (responsible, accountable, consulted, and informed) matrix.

Table 1. The RACI[a] matrix used to clarify responsibilities in the multisite virtual reality development consortium (medical tr.AI.ning, Germany; 2021-2024). Roles include project manager, didactic lead, technical lead, medical experts, and student assistants involved in course delivery and evaluation[b].

Components	Project manager	Didactic lead	Technical lead	Medical expert	Student assistants
Needs assessment	A	R	C	R	I
Learning design	C	A/R	C	R	I
Virtual reality development	I	C	A/R	C	R
Technical testing	I	C	A/R	C	R
Curricular integration	R	A	C	R	I
Evaluation	R	A	C	C	R

[a]RACI: responsible, accountable, consulted, and informed.

[b]We recommend reviewing the RACI at each milestone to prevent role drift across disciplines and sites.

Defining Project Goals and Management

Choosing a Project Name

All consortium members should have the opportunity to propose ideas for a suitable project name. After gathering suggestions, potential candidates can be discussed within the team, culminating in a final vote and confirmation by the project manager. In our case, the name needed to represent both medical education and the integration of AI technology. As a result, we chose medical tr.AI.ning, emphasizing the acronym “AI” to highlight its dual focus and delivering a memorable, descriptive name.

Setting Project Goals and Milestones

Project goals and milestones must be clearly defined to ensure structured progress. Goals must align with the intended educational outcomes, such as improving procedural skills or enhancing clinical decision-making. Based on our experience, this step should precede detailed scenario design to ensure alignment between educational objectives and technical implementation.

Milestones should track key phases, including content development, technical implementation, user testing, and iterative evaluation cycles. Each milestone should be specific, measurable, achievable, relevant, and time-bound (SMART), including tasks such as completing scenario designs, integrating feedback from medical experts, and conducting beta testing. Regular reviews are essential to adjust for technical challenges or evolving educational requirements.

In the medical tr.AI.ning project, we identified 4 milestones (Textbox 1).

XSL•FO
RenderX

Textbox 1. Milestones of the medical tr.AI.ning project.

1. Developing a robust virtual reality foundation and authoring tool

Creating realistic medical scenarios with a focus on artificial intelligence (AI)–supported visual elements; designing the core functions of the virtual reality (VR) environment and the authoring tool.

2. Designing user navigation and smart interaction features

Implementing navigation functions within the VR environment, a clear visual design, and intelligent interaction features.

3. Enabling natural language and gesture-based interactions

Leveraging AI-based speech interaction for input and output; additionally, developing an optional gesture-based interface.

4. Integration and evaluation

Integrating all developed components into a comprehensive, user-friendly VR application and conducting thorough evaluations; finalizing and using the authoring tool to create additional VR scenarios.

Developing a Project Timeline

The project timeline should follow a phased approach with clear, realistic deadlines. This typically begins with the discovery phase, which includes goal setting, stakeholder alignment, and requirements gathering. Next, the design phase should focus on content creation, scenario mapping, and visual asset development, followed by the development phase, which covers programming, VR integration, and interactivity design. The testing phase involves feedback from medical experts and beta testing [18].

Finally, the deployment phase ensures that the scenarios are refined and ready for educational use. Periodic reviews should be scheduled to assess progress and make adjustments as needed. After the VR scenarios are integrated into the curriculum, feedback from students should be collected to optimize both usability and content.

Budgeting and Allocation of Resources

Key Cost Components and Considerations

Budgeting and resource allocation should be based on well-defined objectives, with careful attention to relevant cost categories. Key cost components include personnel expenses for medical (didactical) experts, designers, and developers, which often constitute the largest share. Other substantial costs include licenses for software tools and assets (eg, VR development platforms, 3D modeling software, communication, and data storage tools). Hardware costs for head-mounted displays (HMDs), powerful computers, and necessary accessories (eg, cables, tripods, hygiene equipment) also represent a substantial portion of the budget.

Additionally, providing technical support to manage ongoing maintenance and supervise related courses is essential. In the medical tr.AI.ning project, this was done by research assistants. For dissemination, costs related to presenting the project at national and international conferences, publishing in open-access venues, and ensuring public accessibility (eg, through a dedicated website) should be included in the calculations.

Cross-departmental collaboration plays a critical role in ensuring that financial resources are appropriately aligned with project milestones. Regular reviews should be scheduled to adjust for evolving needs and unforeseen expenses, ensuring that the project remains on budget and achieves its objectives.

Securing Financial Support: Identifying Appropriate Project Calls and Tenders

Securing financial support is essential to account for the interdisciplinary nature of the consortium and the associated costs. Researchers should actively explore grant opportunities and funding programs focused on technology in education, health care innovation, and digital transformation. Priority should be given to funding calls that emphasize interdisciplinary research; medical training innovations; and VR, augmented reality, or extended reality technology. Funding sources can include national and international project calls, European Union programs such as Horizon Europe, or regional grants targeting medical technology. Partnerships with private-sector companies specializing in educational medical hardware or software should also be considered to supplement public funding and potentially facilitate subsequent commercialization.

Additionally, it is crucial to designate a team member or external consultant to take responsibility for securing funding for follow-up projects early in the development cycle. This ensures continuity of the research and development pipeline, allowing the consortium to build on its initial achievements and scale the project's impact. Early identification of future funding opportunities and alignment with upcoming calls or partnerships can help avoid interruptions in development and support the long-term sustainability of the initiative.

The medical tr.AI.ning project was funded through the program "Artificial Intelligence in Higher Education" by the German Federal Ministry of Education and Research for a 3-year period (grant 16DHBKI077). The project was successfully submitted in the subcategory "national cooperative research project."

Minimum Viable Requirements

Based on our experience, a minimal viable setup for implementing VR courses in medical education includes (1) a core team consisting of at least 1 didactical coordinator, 1 clinical content expert, and 1 technical lead; (2) standalone VR hardware with institutional device management capabilities; (3) a dedicated physical space that allows safe VR use; and (4) basic evaluation instruments assessing usability, acceptance, and perceived learning benefit.

XSL•FO
RenderX

Key Actions (Phase 1)

- Define the curricular gap and justify why VR adds value compared with existing methods.
- Translate the gap into competency-based learning objectives (eg, NKLM 2.0).
- Build the consortium and assign roles.
- Define SMART milestones and a phased timeline, including pilot testing and curricular integration.
- Decide on the minimum viable setup (hardware, room, staffing, hygiene, evaluation instruments).

Phase 2: How to Organize

Overview

This phase addresses the organizational and managerial aspects of VR project implementation. It provides practical guidance on project management structures, task allocation, interdisciplinary communication, and technical infrastructure planning.

Communication and Data Storage

A key aspect of successful implementation is communication. Therefore, the respective platforms should be defined early, along with rules for notifications. Additionally, an efficient meeting schedule should be established, including agendas and time for discussing difficult topics.

Hybrid and remote team setups introduce unique challenges in maintaining communication, collaboration, and cohesion. Therefore, establishing clear rules for asynchronous and synchronous communication is vital to minimizing delays and misunderstandings. Team members should have access to well-defined workflows and tools tailored for remote use.

A minimal viable setup includes a shared cloud platform, a messaging tool for asynchronous communication, and a calendar system for scheduling regular review meetings. However, the abundance of available communication tools, such as email, communication and collaboration platforms (eg, Microsoft Teams, Slack, Mattermost), as well as video conferencing tools, can often lead to confusion if not properly managed. In hybrid or remote setups, balancing synchronous (eg, video calls) and asynchronous (eg, collaboration platforms) communication becomes essential to accommodate team members in different locations. Informal virtual spaces, such as casual chat channels, can help build team rapport and compensate for the lack of face-to-face interaction. In our project, we opted to use email for formal communications and important updates.

Meetings were scheduled, and calendar invites were shared via Microsoft Outlook, reducing the need for frequent reminder emails. Online meetings were conducted via the video conferencing tool Zoom (Zoom Communications, Inc), using a permanent access link that ensured easy retrieval and accessibility for all project members at any time.

For informal and real-time discussion, we used the collaboration platform Mattermost (Mattermost Inc), which was hosted by the Center for Information Technology, University of Münster. Several topic-specific channels were defined to keep conversations organized and prevent information overload. A reasonable allocation of members to channels helped prevent flooding with undirected information and kept discussions focused on relevant themes. However, the number of communication channels should be limited to avoid overwhelming team members with multiple sources of information.

A fixed meeting cadence supports transparency and alignment in interdisciplinary VR projects. Based on our experience, we recommend a 2-tier structure:

- Weekly working meeting (60-90 minutes): All research associates review progress, clarify blockers, and prioritize tasks. Meetings may alternate between development-focused discussions and milestone planning.
- Monthly consortium meeting:
  Project leads and working group representatives review overall progress, address strategic decisions, and confirm next steps.

Clear agendas and concise written minutes should document decisions and assigned responsibilities. The meeting structure should be adapted as the project transitions from exploration to implementation and curricular integration phases.

To share and store all relevant data, a common cloud solution should be used that complies with national data protection regulations. This is particularly important in the medical context when dealing with sensitive data (eg, patient or study data) used or collected during the project. It is also essential that all parties have equal access to the cloud and that its capacity is sufficient for software-based projects. In our case, we used the university cloud for North Rhine-Westphalia, “Sciebo” (version 5.3.1.14360; University of Münster, Münster, Germany).

Moreover, it is advisable to use a designated agile project management framework, for example, Scrum [19], and tools such as Kanban [20] to provide an overview of current tasks and ongoing work. The iterative approach of agile software development, with short development cycles, facilitates continuous feedback and improvement, which is especially beneficial in complex, interdisciplinary projects. Kanban serves as a visual workflow management method designed to optimize efficiency and productivity by tracking tasks on a board divided into columns, typically representing stages of a process.

In our project, we used the Boards function of the Mattermost platform as a Kanban tool for meeting agendas and protocols, as well as the ticketing function of GitLab (GitLab Inc) for specific work assignments in software development. In addition, we partially used selected methods of Scrum and agile software development for the creation of the VR simulation.

While short iteration cycles enabled early evaluation and adaptation of VR prototypes, the complexity of VR development often exceeded the constraints of short sprints. Therefore, we did not prioritize all aspects of the Scrum framework, although, in retrospect, stricter use of this method could have helped streamline some steps of the development process.

XSL•FO
RenderX

Lessons Learned and Implications for Future Implementations

If the project were to be implemented again, several structural decisions would be adjusted based on our experience. In particular, establishing a central technical coordinator from the outset would be prioritized to improve cross-site synchronization, reduce redundancy, and streamline technical decision-making.

In addition, a more consistent application of agile project management practices, including clearer sprint definitions, responsibilities, and review cycles, could further enhance development efficiency and transparency.

Finally, earlier and more systematic alignment of terminology and expectations across disciplines may help mitigate communication barriers inherent in interdisciplinary VR projects. These reflections highlight not only what proved effective but also which aspects require adaptation to improve future implementations.

Key Actions (Phase 2)

- Agree on 1 primary channel for asynchronous communication and 1 for formal decisions.
- Implement a shared task board (Kanban) and a source-of-truth repository for files.
- Set a fixed meeting cadence (weekly working meeting + monthly consortium steering).
- Define infrastructure responsibilities (room booking, device management, updates, and hygiene).

Phase 3: How to Build

Overview

This phase focuses on the practical development of VR applications for medical education. It outlines design considerations, interaction concepts, and technical decisions that guide the implementation of educationally meaningful VR simulations.

Educational Context

Defining Learning Objectives

To ensure sustainability, the educational context of a VR-based medical education project should be closely aligned with current competency-based learning objective guidelines, such as those outlined in the German National Competency-Based Learning Objectives Catalogue for Medicine (NKLM 2.0 [17]). This alignment helps ensure that the VR simulations directly address the core competencies and skills expected of medical graduates and educators. Furthermore, alignment with established national and international medical education frameworks helps keep the content relevant, which ultimately fosters not only institutional support but also adoption by other medical schools.

In the medical tr.AI.ning project, we initially focused on dermatological education and defined the corresponding learning objectives based on the NKLM 2.0 before the development of the actual application. A strong overall focus was placed on the competency "clinical reasoning," which refers to the process by which medical professionals gather and analyze patient information, generate differential diagnoses, and make informed decisions about patient management [21].

Discussions of Technical Feasibility

The requirements arising from the course design for the VR simulation must be discussed with the technical team members to assess their feasibility and, if necessary, revise the concept. This includes, for example, requirements for the environment, objects, and examination tools, as well as the behavior of virtual patients and the necessary interactions within the simulation.

Defining Course Structure

The course structure for VR-based courses in medical education should be designed to target specific, predefined learning objectives and requirements. A modular design approach divides the course into distinct sections, each focused on specific competencies, such as clinical procedures or clinical reasoning, allowing for independent updates to individual modules. A blended learning strategy combines immersive VR simulations with traditional lectures, workshops, or online materials, providing a comprehensive learning experience that supports both practical application and the necessary theoretical knowledge. As part of this strategy, theoretical input should precede the VR simulation to ensure that students have sufficient knowledge to complete the respective tasks in VR. Additionally, offering a VR tutorial session is advisable to help students become familiar with the navigation and controls.

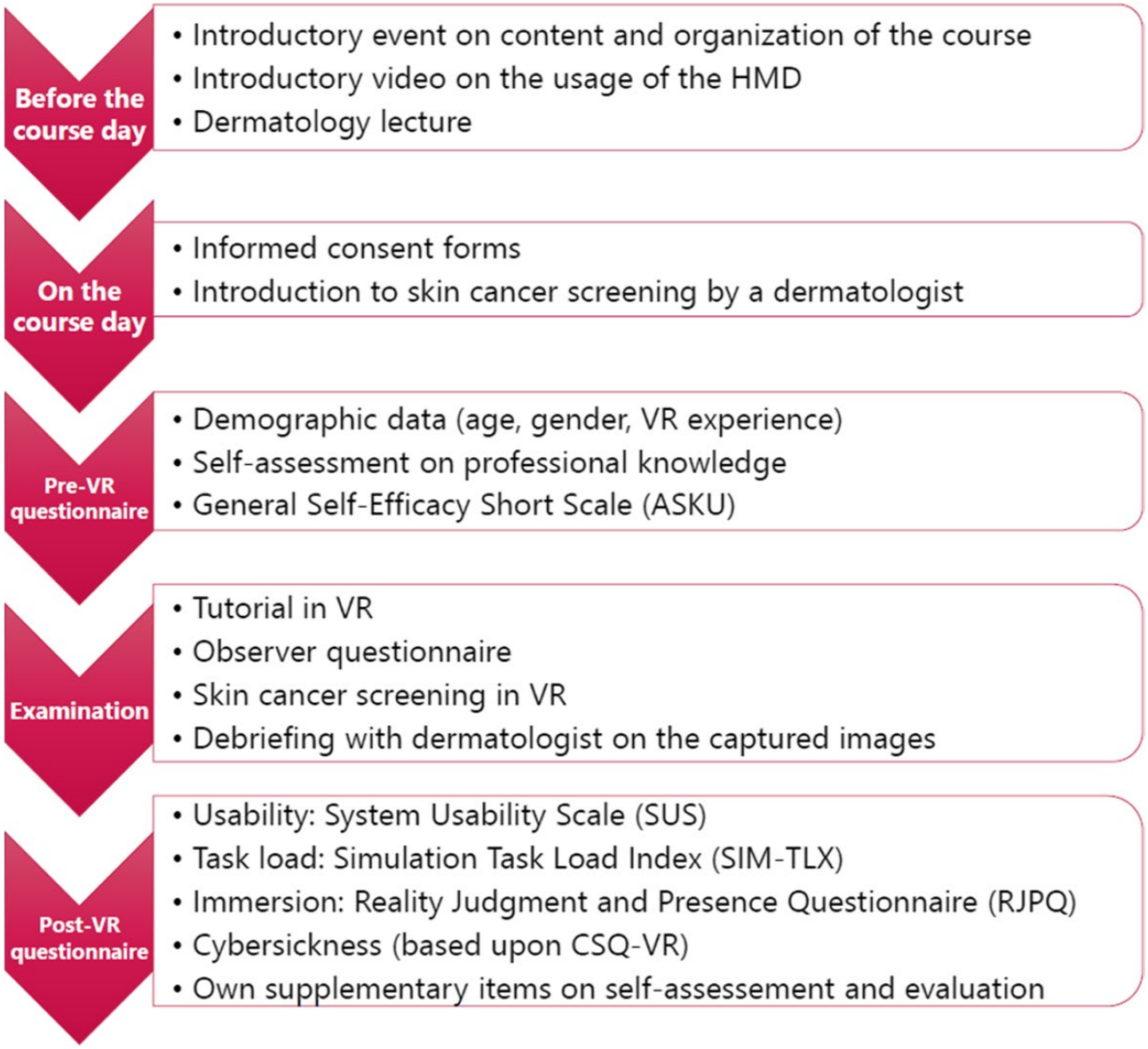
Courses can either be self-paced, allowing students to learn at their own pace and offering flexibility for their schedules, or instructor-led, ensuring structured guidance and personalized feedback from experts. Scenario-based learning within VR allows students to engage in real-world medical situations, promoting critical thinking and decision-making under realistic conditions.

Finally, integrating assessment and feedback mechanisms, such as formative and summative evaluations, ensures continuous tracking of student progress, alignment with competency-based educational goals, and the individual feedback students need to continuously improve their performance.

In the medical tr.AI.ning project, we started the course day by providing instructions and theoretical background knowledge on the strategy of a dermatological whole-body examination by a qualified dermatologist to ensure that students had sufficient theoretical knowledge to benefit adequately from practicing the procedure. After that, an interactive VR tutorial had to be completed, in which all necessary interactions could be practiced. Following the actual simulation, the results were discussed with the dermatologist, feedback was given, and questions could be asked. Figure 2 illustrates the overall structure and sequence of activities in the VR dermatology course (also see [14]).

XSL•FO
RenderX

Figure 2. Time schedule and content of the virtual reality (VR) course in dermatology established at Saarland University [14]. HMD: head-mounted display. CSQ-VR: Cybersickness in Virtual Reality Questionnaire.



Technical Foundation

Choosing VR Hardware

Several factors influence the decision regarding the most appropriate VR hardware. Performance requirements are crucial, as the hardware must support high-quality graphics and real-time interactions to create immersive and realistic medical scenarios.

Ease of use and comfort are also important, particularly for longer training sessions, so lightweight headsets with ergonomic designs are preferred.

Cost and scalability are major considerations, as the hardware must be affordable for institutions and scalable to accommodate multiple users simultaneously. Durability and ease of maintenance should be factored in, as the hardware will likely undergo frequent use, necessitating reliable performance and straightforward repair or replacement options. Mobility is another key consideration, that is, whether the VR setup will be room-based or portable, wire-free or PC-tethered, depending on the logistical possibilities and learning objectives. To meet hygiene standards, disinfection equipment should be used to clean the hardware between users, and single-use hygiene masks should be placed inside the HMD.

As graphic resolution plays a substantial role, not only in the context of cybersickness [22] but also in the accurate representation of medical conditions, the “Valve Index” by Valve was chosen for the medical tr.AI.ning project at both locations in 2022. The special Knuckles Controllers of the Valve Index were favored because of their finger-tracking capabilities, which allow the hands to be used as naturally as possible in medical scenarios. Although the VR hardware market is rapidly evolving, future compatibility and dissemination strategy should be considered when choosing the appropriate hardware. For example, in our case, relying on Steam’s Valve Index VR Kit made transfer to other faculties challenging because of Steam’s credit card–only acquisition policy.

XSL•FO
RenderX

Providing Appropriate Infrastructure

Ensuring appropriate infrastructure is crucial for the success of the project. The physical infrastructure should include dedicated rooms equipped with adequate space, power, and ventilation to accommodate VR hardware, such as HMDs, sensors, and powerful computers. These spaces need to be arranged to facilitate ease of movement and interaction (Figure 3).

Figure 3. An equipped course room for a student at the Münster campus, located within the premises of the LIMETTE (Lernzentrum für individualisiertes medizinisches Tätigkeitstraining - Learning Center for Individualized Medical Skills Training) simulation training center, featuring technical upgrades to support the use of virtual reality hardware.



On the technical side, IT support is essential to ensure that the VR hardware and software receive regular updates and troubleshooting when necessary. Network infrastructure is equally important, as VR simulations often require stable, high-bandwidth internet connections, particularly if cloud-based processing or collaborative simulations are involved. Additionally, the infrastructure should meet data management requirements, particularly if student performance data and feedback are being collected for evaluation or research purposes. This requires secure storage solutions that comply with data privacy standards, especially when handling sensitive information.

Finally, the infrastructure should be scalable to accommodate the growth of the project, whether through expansion to additional courses or by upgrading equipment as technology advances.

AI Frameworks, Methods, and Integration

AI for Interaction and Learner Support

Within the medical tr.AI.ning project, AI was explored as an enabling technology to support selected aspects of content creation and interaction design rather than as a central pedagogical driver of the VR course itself. At the time of project initiation in 2021, many AI-based approaches that are now widely available (eg, large language models [LLMs]) were still emerging or not yet sufficiently mature for reliable integration into medical education contexts. Consequently, several design decisions favored robust, manually implemented solutions over experimental AI-driven components to ensure clinical accuracy, usability, and curricular acceptance.

To facilitate realistic speech interaction with the virtual patients, the development of a specialized language interaction model was initially planned. To ensure that it would meet all requirements in terms of reliability and robustness, it was designed to build on established technologies and frameworks such as RASA (Really Awesome Software Automation) [23] to create a chatbot specifically adapted for use in medical

XSL•FO
RenderX

scenarios. Together with AI-supported speech-to-text and speech synthesis models from Microsoft Azure, the system would have evolved into a complete dialogue system while remaining flexible and predictable at its core through the use of decision trees and scenario-specific intention and action systems for the chatbot.

However, with the initial public preview release of ChatGPT at the end of 2022, OpenAI caused a significant disruption in the research community focused on the development of intelligent speech-enabled agent systems. Consequently, our efforts to develop a novel, cutting-edge interaction model for virtual medical scenarios were suspended. The subsequent rapid developments and possibilities that emerged in the aftermath of the LLM surge quickly became apparent, prompting an expeditious transition to a robust menu-based interaction system in the project's initial phase. Concurrently, the integration potential of novel LLM models was explored in small qualitative user studies.

Based on these findings, it became apparent that the models were not yet suitable for stable use in the medical curriculum. On the one hand, the additional overhead of the LLM caused the overall pipeline to be too slow to achieve a realistic conversational experience for the user. On the other hand, the model exhibited a tendency to frequently lose its role and hallucinate. Additionally, the capacity to effectively control tools was not yet sufficiently developed, which significantly hindered the synchronization of animations and actions. However, many of these issues have been mitigated or addressed in recent versions of LLMs. They currently offer great potential for the implementation of an intelligent speech-enabled agent system. A comprehensive evaluation within our system framework remains pending.

Considering the rapid development of AI technology, we recommend in-depth research on available tools and designing their integration in a flexible way to make short-term adjustments possible.

AI for Content Generation

AI also enabled us to explore new ways of generating content for learning applications that can easily be scaled up, allowing for more personalized learning experiences. We compared various image-generation techniques to create skin lesion images for our dermatological scenario. Ultimately, we opted to train our own diffusion model from scratch using publicly available data with real images [16]. We also added functionality to control the generation process, enabling us to create lesions directly on the virtual patient's skin texture. This allowed us to create multiple scenarios with different lesions placed in various locations on the virtual patient's body. For now, a medical expert has manually checked all of these lesions to ensure they are correct and do not contain unrealistic artifacts. However, it may be possible to fully automate the process in the future, providing an individual scenario for each student. Additionally, the generation process enables those preparing the course to customize the content to specific requirements. For example, we have also developed a tool that can generate lesions in a specific shape. This allows the focus to shift toward specific characteristics when identifying a malignant lesion, for example, more on texture than asymmetry.

Finally, the generated lesions can no longer be associated with a real patient, thereby helping to preserve patient privacy while still providing access to a nearly unlimited number of examples.

Interdisciplinary Communication

General Principles

Overall, interdisciplinary communication is a critical challenge that must be addressed effectively [24]. In addition to the typical difficulties faced by growing teams, there are further obstacles related to differing vocabularies and the potential for unrealistic expectations regarding the capabilities of other teams. Regular meetings are essential to quickly identify, uncover, and resolve misunderstandings. Additionally, visual representations of current development stages can help nontechnical team members stay aligned and informed.

Assigning Tasks

While this aspect may initially seem trivial, it is often overlooked in practice. Project tasks should be defined in great detail and assigned to specific individuals, making it clear who is responsible for each task. This reduces confusion and ensures that no critical aspect of the project is overlooked. It also facilitates better progress tracking, enabling project managers to identify potential delays or issues early on. Moreover, assigning tasks based on individuals' skills or expertise ensures that the right people handle the right responsibilities, thereby improving the overall quality of the project. This structured approach promotes better collaboration, as each team member understands their role and contribution to the project's success.

Defining Deadlines

Clear deadlines are essential for any project, as they provide structure, focus, and a sense of urgency, helping to keep the project on track. By setting deadlines, teams can prioritize tasks, allocate resources efficiently, and ensure steady progress toward milestones. In particular, for complex interdisciplinary projects, deadlines help synchronize the efforts of all disciplines, minimize bottlenecks, and ensure the timely delivery of the project.

In case of delays or when deadlines cannot be met, transparency with all cooperation partners is essential. Open communication helps to identify the underlying challenges and collaboratively develop solutions. One effective approach can be to reassign tasks to other team members or partners who are better positioned to complete them within the required time frame. This ensures that the project remains on schedule while maintaining quality.

Key Actions (Phase 3)

- Lock learning objectives and confirm technical feasibility with the development team.
- Select VR hardware based on visual fidelity, comfort, maintainability, and scalability.
- Establish an iterative build loop (prototype → medical expert review → user test → backlog).

XSL•FO
RenderX

- Decide whether AI is used for interaction support, content generation, or neither.
- Create a VR onboarding tutorial and define fallback options (cybersickness, accessibility).

Phase 4: How to Integrate

Overview

This phase focuses on the systematic integration of VR applications into medical curricula. It addresses iterative testing, pilot studies, curricular embedding, dissemination strategies, and long-term sustainability.

Iterative Testing During Development: Small Iteration Evaluations With Medical (Didactic) Experts

Small iterative evaluations with medical (didactic) experts are essential to ensure the educational validity and content quality of VR simulations. Conducting these evaluations regularly throughout the development process helps identify potential issues early, ensuring that the content aligns with clinical practices, educational goals, and frameworks such as the NKLM 2.0. These evaluations can be organized as short, focused sessions at the end of each development iteration. In practice, we recommend proper documentation to ensure that feedback is manageable and directly actionable.

Especially, iterative testing and early collaboration with health care professionals provide essential expertise in defining appropriate real-world medical scenarios, procedures, and patient interactions that the simulations must replicate. Their input is critical for ensuring that the VR environments reflect the complexity, decision-making processes, and challenges faced in actual clinical practice. Additionally, medical experts help ensure that the simulations align with current medical standards, protocols, and guidelines, which is vital for the validity of the VR training content.

Alongside clinical accuracy, realism and user experience (UX) should also be regularly evaluated. Feedback is then incorporated into subsequent iterations, allowing for continuous refinement. In our project, we regularly tested simulation prototypes with board-certified dermatologists to ensure the accuracy and medical validity of the simulation. Feedback was documented and directly influenced the next development steps.

Pilot Studies: Testing the Prototype in Small User Evaluations

Testing the prototype in small user evaluations ensures that the product is user-friendly and aligned with the intended learning objectives. Small-scale testing allows for early detection of usability issues, such as interface design flaws, technical glitches, or difficulties in navigating within the VR environment. It also provides valuable insights into the simulation's effectiveness in facilitating learning and whether students can successfully engage with the content to develop the targeted skills or knowledge. To conduct these evaluations, a small group of representative users, such as medical students or clinicians, should be selected to interact with the prototype.

Structured user testing sessions can be organized in which participants complete specific tasks within the VR simulation under observation. Based on our experience, this step should precede larger-scale deployment and ideally involve 5-10 representative students to capture diverse interactions with the VR environment. Feedback can be collected through postsession interviews, surveys, or direct observation of user interactions.

Quantitative metrics such as task completion time and error rates, along with qualitative feedback on ease of use and learning experience, should be gathered and analyzed. This iterative testing process helps identify areas for improvement before larger-scale deployment, ensuring that the final product is both effective and intuitive for its target audience.

Table 2 presents an overview of the evaluation activities carried out within the medical tr.AI.ning project that informed this tutorial, feasibility evaluation, and scientific contributions, as detailed below.

Table 2. Overview of evaluation activities informing this tutorial (medical tr.AI.ning, 2021-2024).

Evaluation type	Participants	Purpose/decision informed	Timing/setting
Feasibility study/curricular pilot	n=140 medical students	Technical feasibility + usability/presence/task load + perceived competence across teaching unit	Summer term 2023; University of Münster
Feasibility study/curricular pilot	n=58 medical students	Feasibility + usability/immersion/task load/cybersickness + perceived competence	Summer term 2023; Saarland University (Homburg)
Iterative SME[a] reviews (dermatology)	3-5 board-certified dermatologists	Clinical accuracy checks; scenario realism; and checklist completeness	Throughout development iterations (2022-2024)
Authoring tool workshops/usability tests (educators)	5-7 medical educators	Requirements + user interface comprehension + task success; and refinement of authoring workflow	Iterative workshops/interviews (2022-2024)

[a]SME: subject matter expert.

In our project, it was important that medical teachers could create their own VR training scenarios, even without prior VR knowledge, by using a dedicated authoring tool. Therefore, in an iterative process, the UX designers gathered information about the teachers' wishes and needs through workshops and interviews, leading to a user-centered design approach. Initial functional prototypes were repeatedly tested by the teachers to determine whether the functions were easily understood. The interface was developed through progressively refined iterations to ensure good usability and UX, and to help teachers identify

XSL•FO
RenderX

with the authoring tool and gain the confidence to create their own VR scenarios.

Throughout the project, we collected both qualitative data (comments, opinions, and suggestions for improvement) and quantitative data (time spent on tasks, error rates, and use of specific functions) to verify design decisions and tailor the application to specific user requirements in an iterative process.

Integration Into the Curriculum

Integration and Organization of the VR Course

Implementing VR-based courses in medical education requires meticulous planning to effectively address practical aspects. Supervision during the course, such as by student assistants, must be well coordinated, with clearly assigned roles for overseeing student interactions within the VR environment and addressing potential technical issues. Access to dedicated rooms and facilities equipped with VR hardware, including HMDs and computers, is essential, and scheduling these resources must be streamlined to avoid bottlenecks. Regular hardware and software maintenance, including timely updates and technical support, is critical to ensuring smooth operation during sessions.

Both staff and students should receive an introduction to the course design, including training on VR tools and how the simulations align with learning objectives. To enhance the scientific value of the project and promote continuous improvement, we recommend that the project be scientifically monitored through evaluation methods. This involves data collection, such as automatic logging of performance data, user feedback, and assessment of learning outcomes. Where appropriate and with prior consent, additional measures such as external assessments or video recordings can also be used. Finally, depending on the hardware used, it must be ensured that all wireless components are fully charged before sessions to allow uninterrupted use.

Depending on the course format and its learning objectives, it may be advisable to provide alternative learning options for students experiencing cybersickness, allowing everyone to achieve creditable learning outcomes regardless of constraints. This comprehensive organization ensures the smooth integration of VR simulations into the medical education curriculum while fostering effective operation and alignment with institutional settings.

VR Tutorial

A VR tutorial should guide users in learning to interact with the software. The structure of the tutorial should ensure that students learn basic interactions at the beginning, such as teleporting and basic object interactions, before introducing more specific functionalities, such as using tools like the digital dermatoscope in the medical tr.AI.ning application.

Interactions should be demonstrated step-by-step, allowing students to follow along and understand the controls. Later, students should be asked to apply their knowledge without specific instructions. For example, navigation should be among the first skills introduced: students should learn how to move around the virtual environment early in the tutorial. Subsequent tasks should require them to navigate without further guidance, reinforcing their understanding of movement controls through active use. This progression from guided learning to independent application ensures that, by the end of the tutorial, students have a solid grasp of the necessary VR controls and can confidently apply them during the actual simulation scenarios. In the medical tr.AI.ning tutorial, students start in a doctor's office and are familiarized with the environment. Gestures and functions to be learned are introduced within a medical context so that the subsequent transition to the actual medical examination simulation feels logical and intuitive.

Extensive Feasibility Evaluation

Establishing a clear assessment strategy is crucial to evaluating the effectiveness and acceptance of VR-based courses in medical education. This strategy should incorporate both formative and summative evaluations to measure acceptance among educators and students, as well as specific learning outcomes. In the initial stages, conducting extensive feasibility evaluations is particularly important. These evaluations should address key aspects such as usability, immersion, presence, cybersickness, and subjective learning outcomes. The findings from these studies provide valuable insights into the UX and educational impact, allowing for the refinement of subsequent prototypes. The respective evaluations of the medical tr.AI.ning project, based on the first courses conducted, have been published and are accessible for reference [14,25], providing detailed insights into the feasibility and effectiveness of the approach. In summary, both studies demonstrate technical feasibility and positive learner reception. Students reported high usability and immersion, alongside significantly increased self-assessed competence following VR training. The VR implementations were positively evaluated for their educational value while also drawing attention to resource and cost considerations relevant to broader curricular adoption.

By integrating these assessment practices, the project can ensure that the VR simulations align with educational goals, meet user expectations, and continuously improve based on empirical feedback.

Dissemination

Public Appearance

To ensure a consistent and unified visual identity for external correspondence, a logo should be developed to represent the project. The creation of a logo requires careful consideration and respect for the values of the individual working groups within the consortium. An iterative design process, with multiple review cycles, ensures that team members feel a sense of pride and unity in the final result. The logo should strike a balance between abstraction and inclusivity to fairly represent the diverse ideas and values contributed by all project partners. In addition, a project website should be developed to support the project's public presence and facilitate contact with potential cooperation partners. Furthermore, representation at relevant conferences ensures visibility, provides opportunities for hands-on experiences with the developed simulations, and opens possibilities for future scientific collaborations.

The logo for the medical tr.AI.ning project was developed by designers within our consortium. It integrates multiple concepts

XSL•FO
RenderX

into a cohesive design. The outer shape resembles a stylized brain, symbolizing the role of AI in the project. The sharp angles create the impression of a 3D ribbon, reflecting the VR application's immersive and realistic content. Additionally, the line's end points suggest the head and tail of a coiled snake, evoking the image of the Aesculapian staff, a symbol widely associated with medicine. This thoughtful integration of concepts underscores the project's interdisciplinary nature and its connection to medical education. Emphasizing the use of AI, the letters "AI" were highlighted in the project name. The green color of the logo was adopted as the official project color to create a unified visual identity. In addition to the logo and colors, it is important that the interaction design and interface of both the authoring tool and the VR environment be developed in a consistent style. This approach reinforces the perception of unity, allowing teachers and students to intuitively recognize all elements as part of a single, integrated system.

To further enhance visibility and outreach, a dedicated project website was developed [26]. The website features a video-based project overview, illustrative images, and continuously updated content, such as details on congress participation, publications, and reports from project meetings. The website reflects the essential stylistic elements of the authoring tool and VR environment interface to ensure a clear association with the project.

To convey professionalism and quality at conferences, the designers created branded materials, including business cards, leaflets, and a booth adorned with posters. These were complemented by a hands-on setup of the VR simulation, allowing attendees to directly experience the application. During public testing rounds, close attention was paid to immediate user feedback. Such settings often bring together experts in medical education and VR technology, providing valuable recommendations and opening opportunities for potential collaborations.

The project was showcased through lectures, exhibition stands, and workshops at various national and international congresses. Notable events included "Würtual Reality" in Würzburg, "Shift Medical" in Heidelberg, the German-speaking countries (Germany, Austria, and Switzerland; DACH region) "Annual Conference of the Association for Medical Education" (GMA) in Osnabrück, and the "International Association for Health Professions Education Conference" in Basel, Switzerland.

Networking Groups

To support sustainability and foster collaborations, consortium members should actively participate in specialized working groups. In our case, we prioritized networking through the VR Working Group of the GMA, taking on both active participation and leadership roles. This working group facilitates regular exchanges of knowledge and resources, providing updates on relevant congresses, publications, and other critical developments related to VR in medical education. The medical tr.AI.ning project was presented within this context, significantly enhancing its visibility and recognition among professional circles across the DACH region. This exposure sparked interest from several universities in adopting the project following the completion of software development. Furthermore, it paved the way for potential collaborations exploring the integration of VR technology in medical education.

Sustainability

To ensure the value and usability of the developed project outcomes beyond the project's life cycle, it is crucial to establish a distribution solution early in the process. A significant challenge in this endeavor is ensuring the transferability of the interface to various hardware systems. For example, in our case, the currently utilized finger-tracking functionality of Valve's Knuckles Controllers enables interaction with objects using natural grasping movements. However, this feature is not supported by most other current controller variants, necessitating the development of a fully button-based alternative to maintain compatibility across different systems.

Additionally, legal and formal considerations must be addressed to provide clear guarantees for all stakeholders, including both developers and end users. This involves reviewing each asset used in the project to identify potential licensing conflicts. Where such issues arise, suitable alternatives should be identified and implemented to ensure the legality and accessibility of the final product.

Providing a dedicated authoring tool, medical asset library, and open-source solutions can enable developers and medical educators to build upon the work of the VR project.

Scientific Contributions

To provide value to the scientific community and other developers, it is essential to disseminate collected data and research findings through online articles and open-access publications. In our project, several open-access articles have been published, contributing valuable insights into the integration of VR in medical education. These publications address various aspects of the project, including its overall objectives [8], students' perceptions and desires regarding the integration of VR into medical curricula [27], a scoping review of the current state of VR integration in medical education [1], and specific findings from implementing a VR simulation for skin cancer screening in the dermatology curriculum at both clinical sites involved in the project [14,25]. Collectively, these works offer a comprehensive perspective on the project's potential and the evolving role of VR in advancing medical training, serving as a resource for future research and development in this field.

Key Actions (Phase 4)

- Run small usability tests (5-10 learners) before scaling to curricular pilots.
- Embed VR with clear staffing, room logistics, device cleaning, and technical support.
- Use a predefined evaluation set (usability, presence, task load, cybersickness, and learning perception).
- Plan dissemination (website, conferences, and publications) and address licensing early.
- Define postproject ownership: maintenance, hardware strategy, and scenario authoring workflow.

XSL•FO
RenderX

Limitations

Several limitations should be considered when interpreting the scope of this work, particularly regarding its context-specific and transferable elements. First, the article follows a mixed format, combining aspects of a project report and a practice-oriented guideline. Its primary aim is to describe the development process of a multidisciplinary VR project rather than to evaluate educational outcomes. Accordingly, the work is predominantly descriptive, and conclusions regarding learning effectiveness or superiority over other teaching formats cannot be drawn from this paper.

The project was developed and implemented at 2 university sites (Münster and Saarland), and certain aspects are therefore context-specific. These include local infrastructural conditions, availability of VR hardware, existing technical expertise, and established interdisciplinary networks. Such factors influenced the pace and organization of the development process and may differ across institutions.

At the same time, the overarching project structure, development workflow, and interdisciplinary coordination strategies described in this paper are not tied to a specific institutional setting and may be transferable to other academic environments with comparable resources. The VR application itself is specific to dermatological skin cancer screening, a clinical domain that benefits from high visual fidelity and pattern recognition. Consequently, the developed assets and scenarios are not directly transferable to other medical disciplines. However, the underlying technical framework, including the modular software design and the AI-based generation of synthetic patient data to address data protection concerns, represents a transferable concept that may be adapted to other visually oriented medical fields, albeit with appropriate domain-specific modifications.

Resource requirements represent another important consideration. While the initial development phase was associated with substantial investments in time, coordination, and technical setup, the resulting course format can be implemented with comparatively moderate ongoing effort. Institutions without existing VR infrastructure or access to external funding may face higher entry barriers, particularly if limited hardware availability restricts scalability. Nevertheless, these constraints primarily affect project initiation rather than the applicability of the general development principles outlined in this article.

From a technical standpoint, development and testing were conducted using a single VR system (Valve Index). Although the software was designed with open-source compatibility and platform independence in mind, UX and performance may vary across different hardware environments. Finally, while the interdisciplinary collaboration constituted a key strength of the project, differences in disciplinary terminology and perspectives occasionally required additional coordination efforts. The final VR scenario focuses on a single-player setting and therefore only partially represents interpersonal and nonverbal aspects of clinical encounters, which should be considered when interpreting the scope of addressed competencies.

Conclusions

This tutorial translates the 3-year, multiinstitutional implementation experience of medical tr.AI.ning into a practical, phase-based workflow for educators and developers who want to establish VR in medical curricula. Across the 4 phases detailed in this tutorial, the central message is that successful VR adoption depends less on technology alone than on a clear educational rationale, interdisciplinary coordination, iterative feedback loops, and early planning for sustainability and dissemination.

The following conclusions summarize the key steps and recommendations of this tutorial and synthesize the central lessons learned from developing and integrating VR courses in medical education (Textbox 2).

XSL•FO
RenderX

Textbox 2. Key steps and recommendations.

1. Identify and justify educational needs

Before integrating a virtual reality (VR) course into the medical curriculum, conduct a comprehensive needs assessment to identify specific gaps in the curriculum. Ensure that VR adds unique value by addressing these gaps and improving the curriculum, rather than simply replacing existing teaching methods.

2. Assemble an interdisciplinary team

Form a diverse team with expertise in medical education, computer science, 3D design, user experience/user interface design, and VR development. Clearly define roles, including project managers, didactic coordinators, and technical coordinators, to streamline responsibilities and collaboration.

3. Define clear project goals and milestones

Set specific, measurable goals and milestones aligned with the desired educational outcomes early on, such as improving procedural skills or clinical decision-making. Establish clear deadlines to track progress effectively.

4. Use a structured project management framework

Implement a project management framework to organize development tasks, enable regular feedback loops, and ensure iterative progress throughout the project life cycle.

5. Align VR content with competency-based learning objectives

Design VR simulations to align with competency-based educational guidelines, ensuring relevance and standardization. This alignment ensures that the simulations directly support the desired learning outcomes.

6. Conduct iterative testing and evaluation

Perform regular, small-scale evaluations with medical, didactic, and user experience experts during development. This iterative approach ensures clinical accuracy, pedagogical effectiveness, and usability before broader deployment.

7. Choose appropriate VR hardware and infrastructure

Select VR hardware that meets the project's performance, comfort, and scalability needs. Ensure the infrastructure supports reliable use, including dedicated space, stable network connectivity, and regular maintenance protocols.

8. Organize efficient communication and data management

Establish clear communication channels integrated with project management tools. Utilize a central platform for updates, task tracking, and data sharing to avoid confusion and ensure a consistent information flow among team members.

9. Integrate VR into the curriculum thoughtfully

Plan for the seamless integration of VR into the existing curriculum early on, and train all educators and assistants to feel comfortable with the technology. Technical support should be readily available.

10. Plan for long-term sustainability and dissemination

Address sustainability early, including transferability to other hardware and legal aspects such as licensing. Promote the project through a professional website, conferences, and open-access publications to ensure visibility and encourage adoption by other institutions.

Acknowledgments

We thank all consortium members of the medical tr.AI.ning project for their cooperative work: University of Münster (Computer Vision & Machine Learning Systems [CVMLS], Institut für Ausbildung und Studienangelegenheiten [IfAS])—Benjamin Risse, Pascal Kockwelp, Leon Pielage, Bernhard Marschall, Anna Junga, Henriette Schulze, Ole Hätscher, and Niklas Tiefenbach; University of Applied Sciences Münster (VCLab, Department of Electrical Engineering and Computer Science [ETI], Münster School of Design [MSD], Institut Gesellschaft und Digitales [GUD])—Tina Glückselig, Kathrin Ungru, Philipp Bozdere, Mariel Kruithoff, and Sam Kastigen; Saarland University (Ubiquitous Media Technologies Lab [UMTL], Coordination Center Homburg for Education & Learning in Medicine [CHELM])—Antonio Krüger, Dimitar Valkov, André Zenner, Florian Daiber, Erum Manzoor, Norbert Graf, Marvin Mergen, Marcel Meyerheim, Thomas Jany, Joe Miebach, and Gaspar Eltze; University of Fine Arts Saar (xm:lab)—Michael Schmitz, Mert Akbal, Corbin Sassen, Frederieke Wagner, and Jennifer Schwarz.

Funding

This work was funded by the Bundesministerium für Bildung und Forschung (grant 16DHBKI080). The funder was not involved in the study design, data collection, analysis, interpretation, or writing of the manuscript.

Authors' Contributions

Conceptualization: M Mergen, M Meyerheim, HS, AJ, LP, PK, CS

XSL•FO
RenderX

Graphics and illustrations: AJ, M Mergen
Writing of original draft: M Mergen, M Meyerheim, HS, AJ, MS, TG, KU, PB, NG, BR
All authors have reviewed and approved the final version of the manuscript and consented to its submission to the journal.

Conflicts of Interest

None declared.

References

1. Mergen M, Graf N, Meyerheim M. Reviewing the current state of virtual reality integration in medical education - a scoping review. BMC Med Educ 2024 Jul 23;24(1):788 [FREE Full text] [doi: 10.1186/s12909-024-05777-5] [Medline: 39044186]
2. Helou S, Khalil N, Daou M, El Helou E. Virtual reality for healthcare: a scoping review of commercially available applications for head-mounted displays. Digit Health 2023;9:20552076231178619 [FREE Full text] [doi: 10.1177/20552076231178619] [Medline: 37312952]
3. Jiang H, Vimalesvaran S, Wang J, Lim K, Mogali S, Car L. Virtual reality in medical students' education: scoping review. JMIR Med Educ 2022 Feb 02;8(1):e34860 [FREE Full text] [doi: 10.2196/34860] [Medline: 35107421]
4. Hong M, Rozenblit J, Hamilton A. Simulation-based surgical training systems in laparoscopic surgery: a current review. Virtual Reality 2020 Sep 17;25(2):491-510 [FREE Full text] [doi: 10.1007/s10055-020-00469-z]
5. Chawla S, Devi S, Calvachi P, Gormley W, Rueda-Esteban R. Evaluation of simulation models in neurosurgical training according to face, content, and construct validity: a systematic review. Acta Neurochir (Wien) 2022 Apr;164(4):947-966 [FREE Full text] [doi: 10.1007/s00701-021-05003-x] [Medline: 35122126]
6. Barteit S, Lanfermann L, Bärnighausen T, Neuhann F, Beiersmann C. Augmented, mixed, and virtual reality-based head-mounted devices for medical education: systematic review. JMIR Serious Games 2021 Jul 08;9(3):e29080 [FREE Full text] [doi: 10.2196/29080] [Medline: 34255668]
7. Baniasadi T, Ayyoubzadeh S, Mohammadzadeh N. Challenges and practical considerations in applying virtual reality in medical education and treatment. Oman Med J 2020 May;35(3):e125 [FREE Full text] [doi: 10.5001/omj.2020.43] [Medline: 32489677]
8. Mergen M, Junga A, Risse B, Valkov D, Graf N, Marschall B. Immersive training of clinical decision making with AI driven virtual patients - a new VR platform called medical tr. GMS J Med Educ 2023;40(2):ZMA001600 [FREE Full text] [doi: 10.3205/ZMA001600]
9. Keicher F, Müller M, Ruf K, Härtel C, König S, Mühling T. Enhancing pediatric emergency training: the impact of virtual reality-simulation on medical student knowledge and learning experience—a pre-post intervention study. Virtual Reality 2025 Nov 20;29(4):175 [FREE Full text] [doi: 10.1007/s10055-025-01243-9]
10. Saab M, O'Donovan M, Koç İ, Kenny J, Hrdlička J, Wild J, et al. Evaluating virtual reality communication training in nursing and medical education: a multi-country cross-sectional study. Nurse Educ Today 2026 May;160:106985 [FREE Full text] [doi: 10.1016/j.nedt.2026.106985] [Medline: 41547319]
11. Einloft J, Meyer H, Bedenbender S, Morgenschweis M, Ganser A, Russ P, et al. Immersive medical training: a comprehensive longitudinal study of extended reality in emergency scenarios for large student groups. BMC Med Educ 2024 Sep 09;24(1):978 [FREE Full text] [doi: 10.1186/s12909-024-05957-3] [Medline: 39252064]
12. Lampropoulos G, del Bosque A, Fernández-Arias P, Vergara D. Virtual reality in medical education, healthcare education, and nursing education: an overview. MTI 2025 Jul 20;9(7):75 [FREE Full text] [doi: 10.3390/mti9070075]
13. Lie S, Helle N, Sletteland N, Vikman M, Bonsaksen T. Implementation of virtual reality in health professions education: scoping review. JMIR Med Educ 2023 Jan 24;9:e41589 [FREE Full text] [doi: 10.2196/41589] [Medline: 36692934]
14. Mergen M, Will L, Graf N, Meyerheim M. Feasibility study on virtual reality-based training for skin cancer screening: bridging the gap in dermatological education. Educ Inf Technol 2024 Sep 10;30(4):5251-5282 [FREE Full text] [doi: 10.1007/s10639-024-13019-w]
15. Córdova-Esparza D. AI-powered educational agents: opportunities, innovations, and ethical challenges. Information 2025 May 31;16(6):469 [FREE Full text] [doi: 10.3390/info16060469]
16. Pielage L, Schmidle P, Marschall B, Risse B. Interactive high-quality skin lesion generation using diffusion models for VR-based dermatological education. New York, NY: Association for Computing Machinery (ACM); 2025 Mar 24 Presented at: The 30th International Conference on Intelligent User Interfaces; March 24, 2025; Cagliari, Italy p. 878 URL: https://dl.acm.org/doi/10.1145/3708359.3712101 [doi: 10.1145/3708359.3712101]
17. NKLM 2.0. Medizinische Fakultätentag. URL: https://medizinische-fakultaeten.de/themen/studium/nklm-nklz/ [accessed 2024-06-13]
18. Boyde J. A Down-To-Earth Guide to SDLC Project Management: Getting Your System/Software Development Life Cycle Project Successfully Across the Line Using PMBOK Adaptively. North Charleston, SC: Createspace Independent Publishing Platform; Jul 2014.
19. Drumond C. What is scrum and how to get started. Atlassian. URL: https://www.atlassian.com/agile/scrum [accessed 2024-11-15]
20. Radigan D. Kanban. Atlassian. URL: https://www.atlassian.com/agile/kanban [accessed 2024-11-15]

XSL•FO
RenderX

21. Daniel M, Rencic J, Durning SJ, Holmboe E, Santen SA, Lang V. Clinical reasoning assessment methods: a scoping review and practical guidance. Academic Medicine 2019;94(6):902. [doi: 10.1097/acm.0000000000002618]
22. Biswas N, Mukherjee A, Bhattacharya S. “Are you feeling sick?” – A systematic literature review of cybersickness in virtual reality. ACM Comput Surv 2024 Jun 29;56(11):1-38 [FREE Full text] [doi: 10.1145/3670008]
23. Rasa Technologies GmbH. URL: https://rasa.com/docs/rasa/ [accessed 2025-07-01]
24. Henson V, Cobourn K, Weathers K, Carey C, Farrell K, Klug J, et al. A practical guide for managing interdisciplinary teams: lessons learned from coupled natural and human systems research. Social Sciences 2020 Jul 09;9(7):119 [FREE Full text] [doi: 10.3390/socsci9070119]
25. Junga A, Schmidle P, Pielage L, Schulze H, Hätscher O, Ständer S, medicaltr.AI.ning Consortium. New horizons in dermatological education: skin cancer screening with virtual reality. J Eur Acad Dermatol Venereol 2024 Dec;38(12):2259-2267 [FREE Full text] [doi: 10.1111/jdv.19960] [Medline: 38497674]
26. medical tr.AI.ning. URL: https://medical-training-project.de/ [accessed 2025-03-22]
27. Mergen M, Meyerheim M, Graf N. Towards integrating virtual reality into medical curricula: a single center student survey. Education Sciences 2023 May 08;13(5):477 [FREE Full text] [doi: 10.3390/educsci13050477]

Abbreviations

AI: artificial intelligence
CHELM: Coordination Center Homburg for Education & Learning in Medicine
CVMLS: Computer Vision & Machine Learning Systems
DACH: German-speaking countries (Germany, Austria, and Switzerland)
ETI: Department of Electrical Engineering and Computer Science
FHMS: University of Applied Sciences Münster
GMA: Association for Medical Education
GUD: Institut Gesellschaft und Digitales
HBKsaar: Hochschule der Bildenden Künste Saar
HMD: head-mounted display
IfAS: Institut für Ausbildung und Studienangelegenheiten
LLM: large language model
MSD: Münster School of Design
NKLM: German National Competency-Based Learning Objectives Catalogue for Medicine
RACI: Responsible, Accountable, Consulted, Informed
RASA: Really Awesome Software Automation
SMART: specific, measurable, achievable, relevant, and time-bound
UI: user interface
UMTL: Ubiquitous Media Technologies Lab
UX: user experience
VR: virtual reality

Edited by S Brini; submitted 20.Jul.2025; peer-reviewed by SS Fatima, C Jaiswal, R Ruparel; comments to author 09.Jan.2026; revised version received 05.Mar.2026; accepted 06.Mar.2026; published 01.May.2026.

Please cite as:
Mergen M, Junga A, Schulze H, Bozdere P, Kockwelp P, Pielage L, Sassen C, Glückselig T, Schmitz M, Ungru K, Graf N, Risse B, Meyerheim M
Developing and Integrating Virtual Reality Courses in Medical Education: Tutorial and Implementation Guideline Informed by Best Practices From the National Project “medical tr.AI.ning”
JMIR Med Educ 2026;12:e80976
URL: https://mededu.jmir.org/2026/1/e80976
doi:10.2196/80976
PMID:

©Marvin Mergen, Anna Junga, Henriette Schulze, Philipp Bozdere, Pascal Kockwelp, Leon Pielage, Corbin Sassen, Tina Glückselig, Michael Schmitz, Kathrin Ungru, Norbert Graf, Benjamin Risse, Marcel Meyerheim. Originally published in JMIR Medical Education (https://mededu.jmir.org), 01.May.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Education, is properly cited. The

XSL•FO
RenderX

complete bibliographic information, a link to the original publication on https://mededu.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Tutorial

Developing a Large Language Model–Based Feedback System for Case Report Writing in Rehabilitation Education: Tutorial

Yuuto Tonouchi[1], BScOT; Shunsuke Nakai[2,3], BScOT, DHSc; Kayo Murakami[1]; Yuki Kataoka[4,5,6,7,8], MD, MPH, DrPH

[1]Department of Rehabilitation, Kyoto Min-iren Asukai Hospital, Kyoto, Kyoto, Japan

[2]Department of Occupational Therapy, Faculty of Rehabilitation, Morinomiya University of Medical Sciences, Osaka, Osaka, Japan

[3]Graduate School of Rehabilitation Science, Osaka Metropolitan University, Osaka, Osaka, Japan

[4]Center for Postgraduate Clinical Training and Career Development, Nagoya University Hospital, Nagoya, Aichi, Japan

[5]Center for Medical Education, Graduate School of Medicine, Nagoya University, Nagoya, Aichi, Japan

[6]Department of Healthcare Epidemiology, Graduate School of Medicine and Public Health, Kyoto University, Kyoto, Kyoto, Japan

[7]Department of Systematic Reviewers, Scientific Research Works Peer Support Group (SRWS-PSG), Osaka, Osaka, Japan

[8]Department of Internal Medicine, Kyoto Min-iren Asukai Hospital, Kyoto, Kyoto, Japan

Corresponding Author:
Yuuto Tonouchi, BScOT
Department of Rehabilitation
Kyoto Min-iren Asukai Hospital
89, Tanaka-asukai
Sakyo-ku
Kyoto, Kyoto, 606-8226
Japan
Phone: 81 75 712 9091
Fax: 81 75 701 9183
Email: yuuto.0901el7@gmail.com

Abstract

Background: Novice health care staff often write case reports during early clinical training. However, many institutions lack structured feedback systems because of time constraints and instructor shortages. Large language models, a form of artificial intelligence (AI), offer new opportunities to enhance educational feedback, yet their application in clinical training requires careful design to ensure pedagogically appropriate and ethically sound outputs.

Objective: This tutorial provides a practical guide for educators without programming experience to develop an AI-based feedback system using 3 accessible tools: Dify (an AI workflow platform), Slack (a messaging app), and Google Apps Script. The system balances educational quality with operational efficiency while incorporating data privacy safeguards for clinical educational content.

Methods: We developed a feedback system comprising 4 AI chatbots with 2 distinct approaches: “loop-based” bots that promote clinical reasoning through iterative, comment-based feedback and “single-shot” bots for efficient proofreading and summarization tasks. The tutorial describes the system architecture; feedback design principles grounded in formative assessment theory; a step-by-step implementation guide; and data privacy safeguards, including a deidentification protocol and application programming interface–based data protection measures. To illustrate feasibility, we conducted a pilot implementation at a community care hospital from April to June 2024, involving 5 novice staff members and 5 instructors.

Results: A pilot implementation at a community care hospital demonstrated that the system was feasible to deploy and operate within routine clinical education workflows. Participant feedback indicated high usability and suggested that the iterative, comment-based feedback approach supported learner engagement while also identifying areas where feedback criteria required refinement to better match institutional educational goals.

Conclusions: This tutorial provides a reproducible framework for building a customized AI feedback system that combines comment-based iterative feedback with human-in-the-loop oversight and a multilayered data privacy framework. By following this guide, educators can implement an adaptive system tailored to their institutional context and clinical domain.

XSL•FO
RenderX

(JMIR Med Educ 2026;12:e76924) doi:10.2196/76924

KEYWORDS

large language model; education; artificial intelligence; AI; rehabilitation; case report; feedback

Introduction

Background

In medical education, writing is not merely a record of clinical events but a vital learning activity that supports reflection and growth. Reflective writing (RW), including the reflective elements embedded in case report writing, fosters core competencies such as clinical reasoning and professional identity formation [1,2]. However, writing becomes educationally meaningful only when feedback effectively guides learners toward explicit goals [3]. Previous studies confirm that supportive, structured, and longitudinal feedback maximizes the impact of RW [2,3].

Delivering such feedback at scale faces persistent challenges in clinical settings. Time constraints and limited instructional skills often prevent adequate feedback [1-3]. Consequently, this feedback loop frequently breaks down, hindering the professional development of novice staff.

Large language models (LLMs) have demonstrated utility for feedback, clinical reasoning, and research assistance, including health care contexts [4-6]. A recent review suggests that artificial intelligence (AI) reduces workload and improves efficiency, particularly in medical writing and administrative support [7]. These developments have increased interest in applying LLMs to educational feedback for RW.

However, realizing this potential requires purpose-built systems that go beyond general-purpose chatbot interactions, along with careful consideration of data privacy. This tutorial addresses that gap.

Aims and Target Audience

This tutorial provides a practical guide for educators without programming experience to develop an AI-based feedback system for case report writing in clinical education. Specifically, it offers (1) a conceptual framework distinguishing custom AI systems from standard chatbot use, (2) a step-by-step implementation guide using 3 accessible tools (Dify [LangGenius, Inc]; Slack [Slack Technologies, LLC]; and Google Apps Script [GAS; Google LLC]), (3) ethical and data privacy guidelines for deploying AI in medical education settings, and (4) empirical evaluation data from a pilot implementation at a community care hospital.

The primary audience is clinical educators who supervise novice staff writing case reports, experience time pressure when providing feedback, and lack programming expertise. Although the examples and prompts presented here focus on rehabilitation settings as a working illustration, educators in other medical fields can adapt this framework by modifying the prompts to suit their specific clinical domains and educational objectives.

Rationale for a Customized System

General-purpose chatbots (eg, a standard ChatGPT interface) pose substantial risks when used without customization in medical education. The medical domain requires specialized terminology and linguistic precision that general models may fail to process accurately [8]. Effective AI use in medical education depends on strategic prompt design; poorly constructed instructions can yield inaccurate or unsafe outputs [9]. Mastering such prompt design techniques requires technical expertise that most clinical educators do not possess [9], making accessibility a core design requirement. These considerations led us to develop a customized system rather than rely on off-the-shelf chatbot interactions.

Three requirements distinguish a custom educational system from standard chatbot use. First, standardization ensures fairness. In direct chatbot interaction, response quality depends entirely on the user's prompting skills. Novice staff may receive suboptimal feedback simply due to limited prompting skills. A custom system uses preprogrammed instructions, known as meta-prompts (Textbox 1), to evaluate every draft against the same rubric, ensuring consistent feedback for all learners. Second, integration enables visibility and progress management. When novice staff use personal chatbot accounts, educators cannot monitor their interactions. Integrating the system into a shared workspace enables educators to transparently track progress and intervene when necessary. Third, centralized logging supports educational auditing. To improve the curriculum, educators need to identify common errors among novice staff. A custom system automatically records all interactions, generating valuable data for analyzing trends in learner performance.

Textbox 1. Prompts vs meta-prompts (user input vs system instructions).

This system distinguishes between 2 types of text inputs:

1. Prompts (user input): the text written by the novice staff (eg, "Here is my draft...").
2. Meta-prompts (system instruction): hidden instructions preset by the educator. This text contains the persona (eg, "You are a senior physical therapist"), the rubric, and formatting rules. Separating these components ensures that the artificial intelligence consistently functions as a trained instructor, regardless of what the novice staff member inputs.

XSL•FO
RenderX

Methods

Highlights

This section is organized as follows. First, we describe the 3-component system architecture (Dify, Slack, and GAS) and the dataflow connecting them. Second, we present the feedback design principles and the specifications of 4 educational chatbots. Third, we provide a step-by-step implementation guide for reproducing the system. Fourth, we address ethical considerations and data privacy safeguards. Finally, we describe a pilot implementation conducted at a community care hospital to assess system usability and feasibility.

System Architecture and Dataflow

Figure 1 illustrates the operational workflow. From the novice staff's perspective, the system provides a straightforward interaction: the novice staff tags the relevant bot in Slack (eg, @Discussion_Bot) and submits a draft text. The system processes the request and provides specific feedback within the same thread. The novice staff then revises the text and replies to the bot again. This iterative conversation continues until the draft meets the required standards. This workflow off-loads the time-consuming revision cycle to the AI, reserving human feedback for clinical validity and professional judgment.

The system architecture consists of 3 components selected to enable implementation without advanced programming knowledge (Figure 2). Programming is a well-documented barrier for health care professionals seeking to develop AI applications [10]. We therefore adopted an approach that does not require specialized programming expertise and uses the following 3 components: Dify, Slack, and GAS.

Dify [11] is an open-source LLM application development platform that serves as the system's central processing engine. Hosted in the cloud, it receives user inputs, processes them through a predesigned workflow, and instructs the AI how to respond. Educators can build the AI's reasoning process visually, without writing code. We selected Dify because it offered the most accessible visual interface among available open-source LLM workflow platforms, combined with built-in application programming interface (API) publishing and conversation logging. The free tier is suitable for initial testing but has limited message credits; a paid plan may be required for sustained use with multiple learners. Data sovereignty considerations are discussed in the Ethical Considerations and Data Privacy section.

Slack [12] serves as the interactive chat interface for novice staff. Its private channel structure enables educators to monitor all novice staff-AI interactions in real time and to communicate with novice staff within the same workspace. The free tier limits message history retention, which we addressed through automatic logging to Google Sheets via GAS.

GAS functions as the technical bridge linking Slack and Dify, securely relaying messages between the 2 platforms. We chose GAS for its native integration with Google Workspace, enabling automatic logging to Google Sheets and administrator error notifications without additional infrastructure. Although Dify provides a built-in Slack plug-in for direct integration, it does not support the logging and oversight functions considered essential for educational management. When a novice staff sends a message on Slack, the data travels through GAS to reach Dify, and the response returns along the same path. Every interaction is automatically logged, visible to educators, and monitored for errors.

A key design feature is the iterative pass or revise loop implemented within Dify's Chatflow functionality (Figure 3). Unlike a simple chatbot that answers a question once, effective education requires iterative engagement [3,13]. A recent Association for Medical Education in Europe guide indicates that continuous, formative feedback aligned with explicit learning objectives develops competencies more effectively than one-time summative evaluation [14]. The system, therefore, implements a conditional workflow:

- Analyze: the system evaluates the novice staff's draft against the defined rubric.
- Decide the following:
  - If the draft meets the criteria, the system outputs "pass" and allows the novice staff to proceed.
  - If not, the system outputs "revise," provides targeted comments, and awaits the corrected draft.
- Repeat: this mechanism prevents novice staff from overlooking feedback and promotes iterative learning.

XSL•FO
RenderX

Figure 1. Operational flow comparison. AI: artificial intelligence.

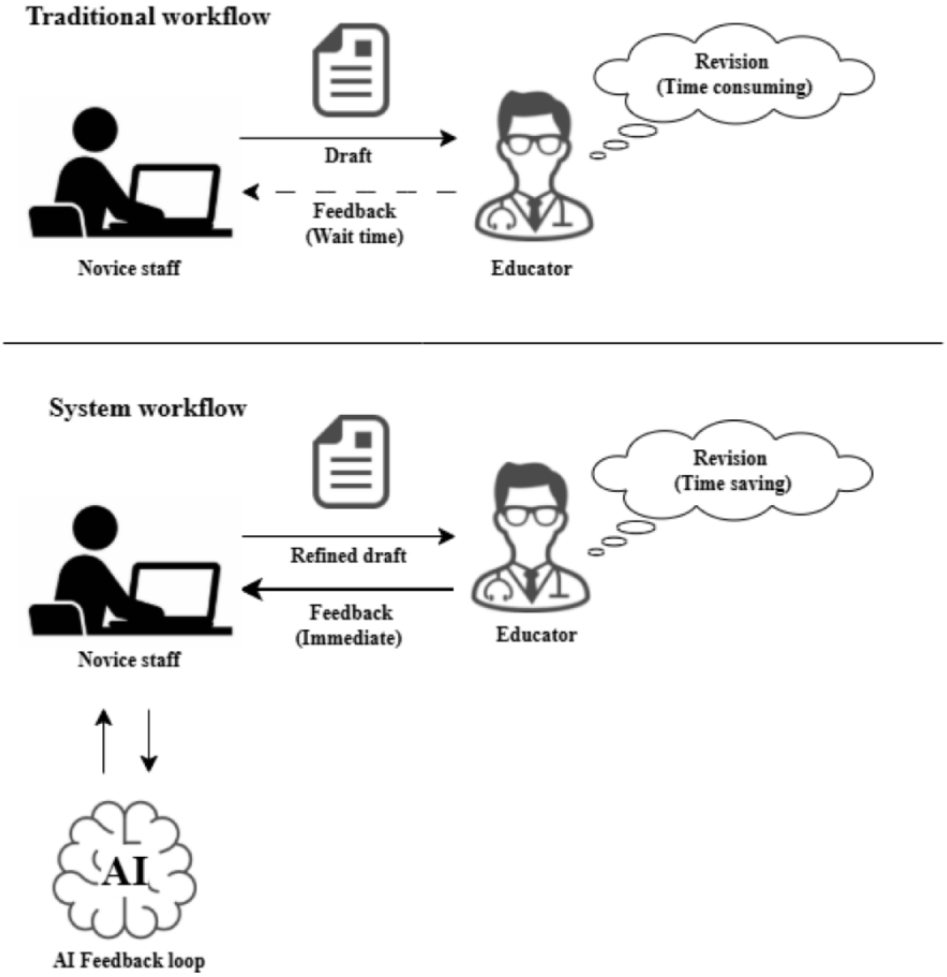
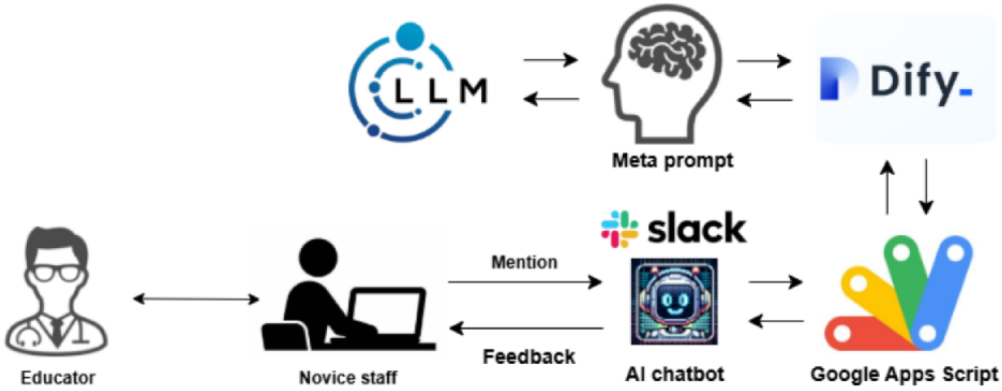
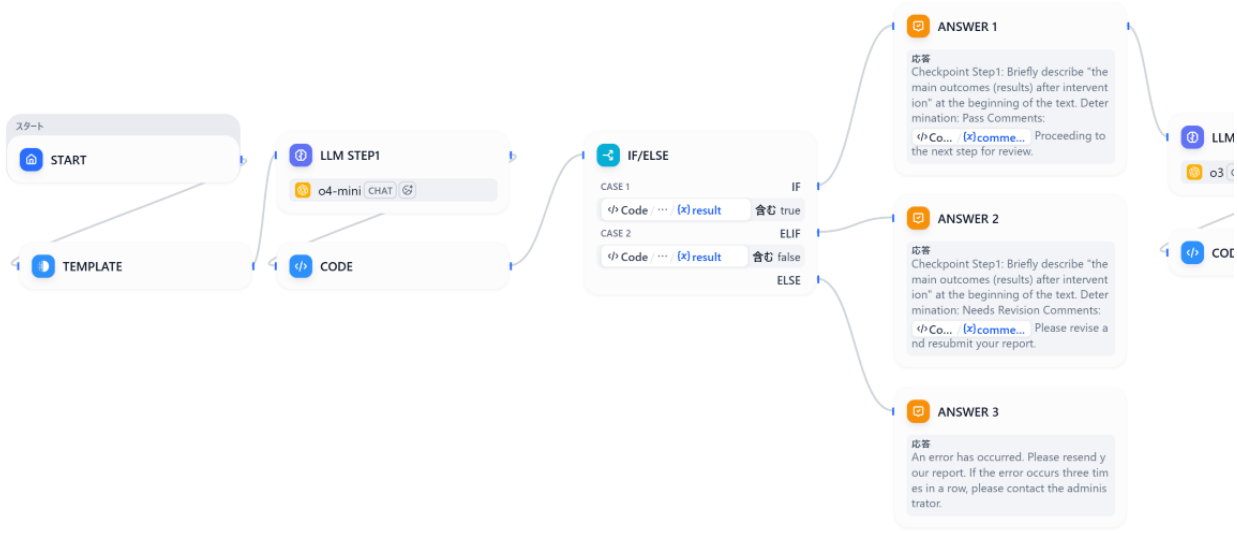


Figure 2. Three-component system architecture. AI: artificial intelligence. LLM: large language model.



XSL•FO
RenderX

Figure 3. Dify Chatflow for the conditional workflow. LLM: large language model.



Feedback Design and Bot Specifications

Overview

The feedback design rests on 2 complementary principles: comment-based feedback and the human-in-the-loop approach.

Comment-based feedback prioritizes written comments over direct correction. Research suggests that comments produce greater gains in writing proficiency than simple error location or direct correction [13] because comments compel learners to actively process information and revise their own work [14]. The system thus functions as an always-available virtual instructor, serving as a strict gatekeeper prior to human review. By off-loading the repetitive task of checking logical consistency and structure to the AI, educators can focus on the final clinical assessment.

The human-in-the-loop approach ensures that educators retain final decision-making authority; AI serves only as a support tool. The AI focuses on overall structure, logical consistency, and writing style, while avoiding specific clinical judgments to mitigate the risk of hallucinations (ie, plausible but incorrect medical advice). This system does not evaluate clinical validity or refine medical discussions. Qualified educators must provide professional feedback on these aspects, and novice staff should treat AI feedback as suggestions for improving writing structure and logic, not as clinical guidance.

On the basis of these principles, the system provides 4 distinct chatbots tailored to the report writing process (Table 1; refer to Multimedia Appendix 1 for detailed meta-prompts.) The system uses 2 distinct feedback approaches: a “loop-based” approach for iterative revision and a “single-shot” approach for efficient one-time processing.

XSL•FO
RenderX

Table 1. Specifications and logic of the 4 chatbots.

Chatbot name and type	Primary function	Evaluation criteria or prompt elements
Integration and interpretation bot (loop-based feedback)	Verifies clinical reasoning and checks consistency between assessment, problems, and goals.	• Passing criteria (checkpoints) • Description of the clinical overview • Clarification of focal points and rationale • Identification of specific clinical problems • Alignment of assessment data with identified problems • Formulation of treatment plan and prognosis
Discussion bot (loop-based feedback)	Evaluates argumentation and checks if the argument logically supports the conclusion based on evidence.	• Passing criteria (checkpoints) • Summary of key intervention outcomes • Integration of relevant literature and guidelines • Comparative analysis with previous studies and original conclusions
Proofreading bot (single-shot output)	Enhances text quality and acts as a domain-specific AI[a] editor to refine style and grammar for academic submission.	• Prompt elements • Detect and correct typographical and grammatical errors • Standardize tone to formal academic style • Refine expressions for clarity and conciseness • Ensure logical consistency and readability • Strengthen paragraph coherence
Abstract summarization bot (single-shot output)	Generates structured summaries and condenses the case report into a format suitable for conference or journal abstract submission.	• Output specifications • Generate a structured abstract text • Propose 2 candidate titles • Select 3 relevant keywords • Provide the final summarized text

[a]AI: artificial intelligence.

Loop-Based Chatbots

The 2 loop-based chatbots promote clinical reasoning through iterative, comment-based feedback:

- Integration and interpretation bot: verifies clinical reasoning by checking the logical consistency between assessment results, the problem list, and goal setting. The bot prompts revisions until all logical checkpoints are met.
- Discussion bot: evaluates whether the argument aligns with the interpretation and logically supports the conclusion. The bot prompts revisions until all checkpoints are satisfied.

Single-Shot Chatbots

The 2 single-shot chatbots perform efficient, single-pass tasks:

- Proofreading bot: corrects grammatical errors and improves clarity without altering the clinical meaning. It generates a polished draft in a single interaction.
- Abstract summarization bot: produces a concise summary of the completed case report in a format suitable for conference or journal submission.

Step-By-Step Implementation

Overview

This guide enables educators to build the system by following step-by-step instructions and copying the provided resources; no complex coding is required. The following subsections outline the role of each component and the key actions involved. Detailed instructions, including specific platform operations, configuration values, and troubleshooting guidance, are provided in Multimedia Appendix 2.

Prerequisites: Accounts and API Keys

Ensure the following are ready before starting:

- Dify account: (cloud version or self-hosted)
- Slack workspace: (educators need permission to create apps and add bots)
- Google account: (to access Google Drive and Sheets)
- API keys: an API key from an LLM provider (eg, OpenAI API key or Anthropic API key; Textbox 2).

Textbox 2. Understanding application programming interface (APIs) and API keys.

- What is an API? An API is a set of rules that allows different software systems to communicate. It acts as a standardized bridge, sending a digital request from your system to the artificial intelligence (AI) service and returning the response. This allows your application to use powerful external AI capabilities without hosting the underlying technology locally.
- What is an API key? An API key is a unique string of characters that acts as a digital ID card (authentication token). When Dify sends a request to a large language model provider (such as OpenAI), it uses this key to verify the user's identity and link use to a specific billing account. Warning: never share this key publicly. If compromised, unauthorized users may incur charges against your account.

XSL•FO
RenderX

Setting Up Dify

Dify serves as the AI processing engine. Implementation involves creating a new Chatflow-type application, importing the feedback workflow via the domain-specific language file (Multimedia Appendix 3), registering the LLM API key, and adjusting the meta-prompts to suit the target clinical domain. Once published, Dify generates an API key for use in subsequent integration steps.

Creating the Slack Application

After configuring Dify, the next step is to create the bot account in Slack. This involves registering a new app in the Slack API console, assigning the minimum necessary permissions (eg, reading messages, sending replies), and installing the bot in the target workspace. The resulting bot token is required for the integration step.

Configuring GAS

Dify handles the AI logic, and Slack provides the chat interface, but the 2 platforms cannot communicate directly. GAS serves as the bridge connecting them. Implementation involves creating a GAS project, deploying the relay script (Multimedia Appendix 4), and securely storing the API credentials from the previous steps as script properties. The script is then deployed as a web application so that Slack can send messages to it.

Final Configuration and Verification

The final step links all 3 components and verifies system function. Configuring Slack's event subscriptions feature to forward messages to the GAS end point completes the communication loop. A test message sent in Slack should trigger the full pipeline: Slack receives the message, GAS relays it to Dify, the LLM generates feedback, and the response returns to Slack while being logged in Google Sheets.

Data Privacy

Because this system transmits case report content through multiple external platforms, safeguarding protected health information (PHI) is essential. This section describes the data protection measures organized into 3 categories: administrative, technical, and supervisory safeguards.

Administrative Safeguards

Before submitting any draft to the system, novice staff must deidentify all patient information following the anonymization standards recommended by the International Committee of Medical Journal Editors [15]. The deidentification protocol involves 2 steps:

1. Remove all direct identifiers: patient names, medical record numbers, dates of birth, addresses, and contact information.
2. Generalize quasi-identifiers to prevent reidentification through combination: convert specific dates to relative timelines (eg, "Day 1 of admission" rather than a calendar date), express exact ages in decade ranges (eg, "a patient in their 80s"), and replace facility names with generic descriptors (eg, "a community rehabilitation hospital").

In the pilot implementation, we provided deidentification training to all participants prior to system use (refer to the Ethical Considerations section). Each novice staff member received individual instruction on the protocol, supplemented by a short instructional video demonstrating the 2-step process with concrete examples from rehabilitation case reports. We reinforced compliance through ongoing supervision during the pilot period.

Technical Safeguards

All communication with LLMs in this system occurs exclusively through the API, not through consumer-facing chat interfaces such as ChatGPT (OpenAI) or Claude.ai (Anthropic). This architectural choice provides a contractual safeguard: both OpenAI and Anthropic explicitly state in their API terms of service that data submitted through the API are not used to train or improve their models [16,17]. Both OpenAI and Anthropic retain API data for up to 30 days solely for abuse monitoring purposes, after which it is automatically deleted. Because data submitted to LLM providers cannot be recalled during this retention window, deidentification prior to submission constitutes the critical safeguard.

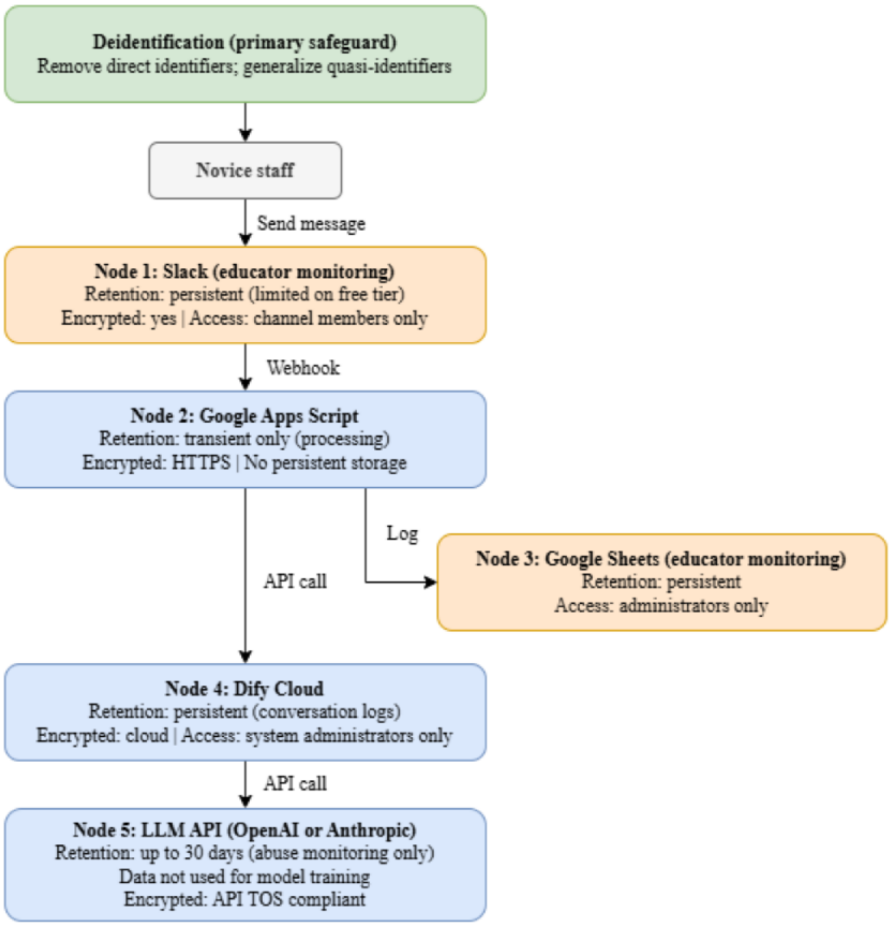
To provide a transparent risk assessment, we trace the complete data path (Figure 4). When a novice staff submits a message in Slack, the data traverses 5 external nodes before a response is returned:

1. Slack servers: Slack stores messages with encryption; access is restricted to invited channel members.
2. GAS: GAS processes data transiently during relay without persistent storage.
3. Google Sheets: conversation logs are persistently stored for educational analysis, with access restricted to designated administrators.
4. Dify Cloud: conversation logs are retained within the platform and are accessible only to the system administrator.
5. LLM API end point (OpenAI or Anthropic): the LLM provider processes data and retains it for up to 30 days under the nontraining policies described above.

For institutions requiring full data sovereignty, Dify offers a self-hosted deployment option that retains all data within the institution's own infrastructure, though this requires server infrastructure and technical expertise.

XSL•FO
RenderX

Figure 4. Dataflow and privacy safeguards across the 5 system nodes. API: application programming interface; LLM: large language model.



Real-Time Oversight and Incident Response

The system architecture supports real-time oversight. Because all interactions occur in shared Slack channels, educators can continuously monitor both novice staff submissions and AI responses, enabling immediate detection of any inadvertent PHI disclosure. The Google Sheets log provides a persistent, searchable record for periodic auditing. If PHI is inadvertently submitted, educators can take the following remediation steps:

- Delete the PHI-containing message from Slack.
- Remove the corresponding row from the Google Sheets log.
- Purge the conversation log from the Dify platform.
- The LLM provider automatically deletes data within the 30-day retention window.

The current system does not include automated PHI detection prior to submission; deidentification relies on the administrative safeguards described above. We identify the development of automated presubmission screening as a priority for future implementations (refer to the limitations and future directions section).

Transparency, Consent, and Scope of AI Feedback

Transparency is foundational to ethical AI adoption. Before introducing the system, educators should clearly explain to novice staff that they are communicating with an AI system and how their data will be processed. Educators should also define the system as an educational support tool for facilitating the drafting process, rather than a replacement for human supervision or a formal assessment of clinical capabilities.

To manage the risk of hallucinations, the system uses 2 strategies. First, the AI's feedback includes a disclaimer: "AI can make mistakes. Always verify medical facts with your supervisor." Second, the AI's role is limited to structural and logical feedback rather than clinical diagnosis, reducing the risk of harmful errors.

Ethical Considerations

The ethics committee of Kyoto Min-iren Asukai Hospital approved this pilot evaluation (2024-0502). All participating staff members received verbal and written explanations of the study's purpose and data handling procedures, and provided written informed consent. We explained deidentification requirements to all novice staff members individually and reinforced them through an instructional video prior to system use. For past case reports used in meta-prompt tuning, we implemented an opt-out procedure. Participants received no financial compensation.

Iterative Development Process

The current Dify-based system evolved through iterative refinement of an earlier prototype. The initial version used a

XSL•FO
RenderX

single-prompt chatbot approach. Early testing revealed 2 challenges: a ceiling effect in which most novice staff passed all checkpoints on their first submission, indicating that the meta-prompts were too lenient for meaningful learning, and declining instruction adherence as prompt length increased. These issues motivated the transition to the Dify-based Chatflow architecture, which evaluates checkpoints sequentially and allows educators to adjust passing criteria through its visual interface.

Some instructors also expressed concern that overreliance on AI feedback could diminish clinical reasoning skills. We therefore designed the system to limit AI feedback to structural and logical aspects and to prompt learners to reconsider their own reasoning rather than providing corrected example text.

Pilot Implementation

To inform the design and iterative refinement of this tutorial, we conducted a single-center pilot implementation at Kyoto Min-iren Asukai Hospital. This pilot used a prototype version with direct Slack-GAS-AI integration prior to the introduction of Dify. Because the user interface (Slack) and the fundamental feedback logic (iterative revisions) remained identical, the usability data are directly applicable to the Dify-based system.

Five novice rehabilitation staff members (3 physiotherapists and 2 occupational therapists) who joined in April 2024 and 5 educational instructors participated. The pilot ran from April to June 2024. Prior to this pilot, the facility had not used Slack for official communication; thus, this pilot represented the participants' first experience with the platform. During the pilot, participants tagged the specific bot (eg, "@Integration_Bot") in Slack to initiate iterative feedback on structure and logic. Once the content was finalized, they used the "@Proofreading_Bot" for grammatical corrections before submitting the final draft to their human supervisors.

We collected data through 2 methods. First, postimplementation surveys via Google Forms assessed the system's usability and acceptance. Novice staff completed the System Usability Scale [18] and rated the appropriateness of the AI feedback content. The System Usability Scale is a standardized 10-item questionnaire for measuring system usability. Scores above 68 indicate above-average usability, and scores above 85 indicate excellent usability. Instructors assessed changes in instructional efficiency and feedback appropriateness. Second, we analyzed use logs from the Slack-linked spreadsheet, including use frequency, revision counts per task, and the content of feedback.

Results

This section describes the pilot implementation experience, which informed the iterative refinement and final design of this tutorial.

The postimplementation survey response rate was 100% (5/5) for both instructors and novice staff. Detailed survey results are provided in Multimedia Appendix 5.

Novice staff found the system accessible, with a median System Usability Scale score of 90 (range 70-95), confirming that the approach lowered the barrier to adoption. All 5 staff members reported that the AI feedback was helpful for their learning (median 5, range 4-5). Instructor feedback identified areas where the tutorial framework requires further customization: while 80% (4/5) of staff members anticipated a reduction in future instructional burden (median 4, range 1-5), ratings for the appropriateness of AI feedback content were moderate (median 3, range 1-4), suggesting that the balance between structural feedback and clinical depth needs institution-specific calibration.

Log analysis confirmed that all 5 novice staff members actively used the system. Typically, they used the integration and interpretation bot 1 to 2 times and the proofreading bot once per report.

Qualitative feedback from novice staff highlighted the psychological benefits of the system, with comments such as "My mental burden is reduced because I can get feedback before showing it to a human". Instructors acknowledged efficiency gains, noting the ability to "focus on specialized feedback". However, some instructors expressed concern that reliance on the system could lead to a decline in writing skills and clinical reasoning abilities.

Discussion

Principal Findings

This tutorial presents a practical framework for building an LLM-based feedback system for case report writing in clinical education. The principal contribution is demonstrating that educators can construct a customized AI feedback system by combining 3 accessible tools (Dify, Slack, and GAS) while maintaining educational quality and data privacy.

The system embodies 2 key design principles. First, comment-based iterative feedback encourages active revision rather than passive correction, preserving the cognitive effort essential for developing clinical reasoning skills [3,13,14]. Second, the human-in-the-loop principle limits AI to structural and logical evaluation, reserving clinical validity assessment for qualified educators and mitigating the risk of unwarranted clinical authority [19].

The pilot implementation experience supported the design decisions underlying this tutorial. In particular, the transition from the prototype to the Dify-based architecture was directly motivated by 2 challenges observed during early testing: a ceiling effect indicating that initial meta-prompts were insufficiently stringent and declining instruction adherence as prompt complexity increased. These observations illustrate a core advantage of the framework presented here: educators can iteratively adjust passing criteria and feedback logic through Dify's visual interface, enabling continuous adaptation without programming expertise.

Comparison With Prior Work

To our knowledge, few studies have provided practical, reproducible guidance for clinical educators to build customized LLM-based feedback systems. Systematic reviews of LLMs in medical education have predominantly identified studies evaluating LLM performance on examinations and clinical knowledge assessments, with very few reporting empirical

XSL•FO
RenderX

implementation research [20,21]. Practical programming-free AI development guides in health care have focused on other domains, such as image classification [10], and exploratory studies in adjacent fields have examined similar accessible tools for providing formative feedback on student writing [22]. A step-by-step guide for building an LLM-based formative feedback system for clinical writing education has not been described. This tutorial addresses that gap.

The design principle of delegating structural and logical review to AI while reserving clinical judgment for educators reflects an emerging model in which AI supports, rather than replaces, expert judgment [19,20]. Prior studies on the use of AI in education have reported reductions in teachers' workload and stress, along with improvements in instructional efficiency [23,24]. The pilot findings were consistent: instructors reported the ability to "focus on specialized feedback" rather than routine text review. This division of labor may offer a scalable approach to feedback delivery in resource-constrained educational settings.

Accessibility was a key design consideration. Prior research indicates that AI-based educational tools positively influence learning efficiency and cognitive motivation [24,25]. The pilot results supported this, with high usability ratings achieved even among staff with no prior experience using the chat platform.

This tutorial demonstrates that rubric-based LLM evaluation can be customized for domain-specific educational purposes without programming expertise. Although the potential of AI as a writing support tool is well documented [26,27] and LLMs can reliably evaluate clinical writing against structured rubrics [28], these capabilities have typically required programming expertise, limiting accessibility for clinical educators.

Overreliance on AI feedback poses recognized risks, including cognitive deskilling and diminished independent reasoning [29,30]. A recent scoping review of ethical challenges in AI-assisted medical education underscores the need for hybrid learning models that balance AI assistance with traditional teaching [19]. This tutorial addresses these concerns through the human-in-the-loop principle: the system limits AI feedback to structural and logical aspects, and all AI-generated comments include disclaimers directing novice staff to verify content with their supervisors.

Limitations and Future Directions

This tutorial and its pilot evaluation have several limitations. First, the small sample size (5 novice staff and 5 instructors) limits causal inference. Second, as a single-center pilot, future research should involve multicenter collaborative studies. Third, the outcomes relied on subjective questionnaires. Future studies should analyze quantitative data such as the duration of instructor guidance and use independent reviewers to evaluate report quality objectively. Fourth, although we provided deidentification training through individual instruction and an instructional video, the pilot did not include an automated presubmission PHI detection mechanism. Recent studies have demonstrated that LLM-based systems can analyze conversation data while preserving privacy [31].

No incidents of PHI disclosure were identified during the pilot period, suggesting that the current approach, which aligns with standard security practices for cloud-based services, was sufficient within this controlled setting. However, future implementations should strengthen compliance through additional safeguards: supervisor-led presubmission review, prioritization of hypothetical cases during early adoption, and development of automated PHI detection systems. In contexts requiring stricter data governance, locally hosted LLMs may also be considered.

Two broader considerations apply to the adoption of this approach. The system relies on cloud-based services, which may not meet the data sovereignty requirements of all institutions. Additionally, because LLM capabilities and API policies evolve rapidly, educators should monitor LLM provider updates regarding data handling practices and model performance. Furthermore, while this tutorial provides detailed implementation instructions and reproducible resources, the step-by-step guide has not yet been independently tested by educators outside the development team.

Future directions include broader validation of the tutorial's reproducibility across diverse clinical settings, development of automated PHI detection as a presubmission safeguard, and investigation of the system's long-term impact on writing proficiency and clinical reasoning development.

Conclusions

This tutorial provides a reproducible guide for building a case report feedback system using Dify and Slack without programming expertise. The architecture enables educators to fine-tune AI logic through visual interfaces, separating AI reasoning from software code. The tutorial also addresses data privacy considerations for deploying AI systems that handle clinical educational content. By emphasizing human oversight at every stage and iterative refinement of feedback criteria, this framework supports educators in adapting the system to their institutional context and evolving educational needs.

Acknowledgments

The authors thank the rehabilitation staff at Kyoto Min-iren Asukai Hospital for their assistance in this project. During the preparation of this manuscript, the authors used Claude Opus 4.6 (Anthropic [17]) and Gemini 3.1 Pro (Google [32]) for the following purposes: drafting and editing assistance, literature search and information gathering, structural planning and outline development, and English language refinement. The authors reviewed, verified, and edited all artificial intelligence–generated content and take full responsibility for the accuracy and integrity of the final manuscript.

XSL•FO
RenderX

Funding

The application programming interface use fee for the pilot implementation was supported by a research fund from the Scientific Research Works Peer Support Group awarded to YK. The costs for the latest system development and article processing charges were covered by a 2025 research grant from the Shinohara Foundation. The funders had no role in the project design, data collection and analysis, decision to publish, or preparation of the manuscript.

Data Availability

The data supporting the findings of this work, including the developed meta-prompts and codes, are available in Multimedia Appendices 1 and 3-5. All other data are available upon request from the corresponding author.

Authors' Contributions

Conceptualization: YT, YK
Funding acquisition: YT, YK
Investigation: YT, SN, KM
Methodology: YT, YK
Project administration: YT
Resources: YT, YK
Supervision: YK
Validation: YT, SN, KM
Writing – original draft: YT
Writing – review & editing: YT, SN, KM, YK

Conflicts of Interest

None declared.

Multimedia Appendix 1
Meta-prompts used in the system.
[DOCX File , 44 KB - mededu_v12i1e76924_app1.docx]

Multimedia Appendix 2
Step-by-step implementation guide.
[DOCX File , 25 KB - mededu_v12i1e76924_app2.docx]

Multimedia Appendix 3
Dify domain-specific language file (importable file for reproduction).
[ZIP File (Zip Archive), 23 KB - mededu_v12i1e76924_app3.zip]

Multimedia Appendix 4
Google Apps Script code.
[TXT File , 7 KB - mededu_v12i1e76924_app4.txt]

Multimedia Appendix 5
Detailed questionnaire results.
[DOCX File , 29 KB - mededu_v12i1e76924_app5.docx]

References

1. Artioli G, Deiana L, De Vincenzo F, Raucci M, Amaducci G, Bassi MC, et al. Health professionals and students' experiences of reflective writing in learning: a qualitative meta-synthesis. BMC Med Educ 2021 Jul 22;21(1):394 [FREE Full text] [doi: 10.1186/s12909-021-02831-4] [Medline: 34294058]
2. Lim JY, Ong SY, Ng CY, Chan KL, Wu SY, So WZ, et al. A systematic scoping review of reflective writing in medical education. BMC Med Educ 2023 Jan 09;23(1):12 [FREE Full text] [doi: 10.1186/s12909-022-03924-4] [Medline: 36624494]
3. Burgess A, van Diggele C, Roberts C, Mellis C. Feedback in the clinical setting. BMC Med Educ 2020 Dec 03;20(Suppl 2):460 [FREE Full text] [doi: 10.1186/s12909-020-02280-5] [Medline: 33272265]
4. Dai W, Lin J, Jin H, Li T, Tsai YS, Gašević D. Can large language models provide feedback to students? A case study on ChatGPT. In: Proceedings of the 2023 IEEE International Conference on Advanced Learning Technologies. 2023 Presented at: ICALT 2023; July 10-13, 2023; Orem, UT. [doi: 10.1109/icalt58122.2023.00100]

XSL•FO
RenderX

5. Liang W, Zhang Y, Cao H, Wang B, Ding DY, Yang X, et al. Can large language models provide useful feedback on research papers? A large-scale empirical analysis. NEJM AI 2024 Jul 25;1(8). [doi: 10.1056/AIoa2400196]
6. Rossettini G, Cook C, Palese A, Pillastrini P, Turolla A. Pros and cons of using artificial intelligence chatbots for musculoskeletal rehabilitation management. J Orthop Sports Phys Ther 2023 Dec;53(12):728-734. [doi: 10.2519/jospt.2023.12000] [Medline: 37707390]
7. Meng X, Yan X, Zhang K, Liu D, Cui X, Yang Y, et al. The application of large language models in medicine: a scoping review. iScience 2024 Apr 23;27(5):109713 [FREE Full text] [doi: 10.1016/j.isci.2024.109713] [Medline: 38746668]
8. Heston TF, Khun C. Prompt engineering in medical education. Int Med Educ 2023 Aug 31;2(3):198-205. [doi: 10.3390/ime2030019]
9. Zaghir J, Naguib M, Bjelogrlic M, Névéol A, Tannier X, Lovis C. Prompt engineering paradigms for medical applications: scoping review. J Med Internet Res 2024 Sep 10;26:e60501 [FREE Full text] [doi: 10.2196/60501] [Medline: 39255030]
10. Hoseini SS, Dewar R. Empowering healthcare professionals with no-code artificial intelligence platforms for model development, a practical demonstration for pathology. Discoveries (Craiova) 2024 Mar 30;12(1):e182. [doi: 10.15190/d.2024.1] [Medline: 39734805]
11. Dify. URL: https://dify.ai/ [accessed 2026-06-01]
12. Slack. URL: https://slack.com/ [accessed 2026-06-01]
13. Biber D, Nekrasova T, Horn B. The effectiveness of feedback for L1-english and L2-writing development: a meta-analysis. ETS Res Rep Ser 2014 Aug 08;2011(1):i-99. [doi: 10.1002/j.2333-8504.2011.tb02241.x]
14. Lipnevich AA, Mattern K, Feddock C. Formative assessment and feedback in medical education: a practical guide: AMEE guide no. 189. Med Teach 2025 Oct 24;48(6):921-940. [doi: 10.1080/0142159x.2025.2569623]
15. International Committee of Medical Journal Editors. URL: http://www.ICMJE.org [accessed 2026-06-01]
16. Data controls in the OpenAI platform. OpenAI. URL: https://platform.openai.com/docs/guides/your-data [accessed 2026-05-31]
17. Is my data used for model training? Anthropic. URL: https://privacy.claude.com/en/articles/7996868-is-my-data-used-for-model-training [accessed 2026-05-31]
18. Brooke J. SUS: a 'quick and dirty' usability scale. In: Usability Evaluation in Industry. Boca Raton, FL: CRC Press; 1996.
19. Li X, Yan X, Lai H. The ethical challenges in the integration of artificial intelligence and large language models in medical education: a scoping review. PLoS One 2025 Oct 22;20(10):e0333411 [FREE Full text] [doi: 10.1371/journal.pone.0333411] [Medline: 41124146]
20. Lucas HC, Upperman JS, Robinson JR. A systematic review of large language models and their implications in medical education. Med Educ 2024 Nov;58(11):1276-1285. [doi: 10.1111/medu.15402] [Medline: 38639098]
21. Aster A, Laupichler MC, Rockwell-Kollmann T, Masala G, Bala E, Raupach T. ChatGPT and other large language models in medical education - scoping literature review. Med Sci Educ 2024 Nov 13;35(1):555-567. [doi: 10.1007/s40670-024-02206-6] [Medline: 40144083]
22. Venter J, Coetzee SA, Schmulian A. Exploring the use of artificial intelligence (AI) in the delivery of effective feedback. Assess Eval High Educ 2024 Oct 20;50(4):516-536. [doi: 10.1080/02602938.2024.2415649]
23. Younas A, Subramanian KP, Haziazi MA, Hussainy SS, Kindi AN. A review on implementation of artificial intelligence in education. Int J Res Innov Soc Sci 2023 Sep 14;VII(VIII):1092-1100. [doi: 10.47772/IJRISS.2023.7886]
24. Kasneci E, Sessler K, Küchemann S, Bannert M, Dementieva D, Fischer F, et al. ChatGPT for good? On opportunities and challenges of large language models for education. Learn Individ Differ 2023 Apr;103:102274. [doi: 10.1016/j.lindif.2023.102274]
25. Meyer J, Jansen T, Schiller R, Liebenow LW, Steinbach M, Horbach A, et al. Using LLMs to bring evidence-based feedback into the classroom: AI-generated feedback increases secondary students' text revision, motivation, and positive emotions. Comput Educ Artif Intell 2024 Jun;6:100199. [doi: 10.1016/j.caeai.2023.100199]
26. Ho WL, Koussayer B, Sujka J. ChatGPT: friend or foe in medical writing? An example of how ChatGPT can be utilized in writing case reports. Surg Pract Sci 2023 Jun 2;14:100185 [FREE Full text] [doi: 10.1016/j.sipas.2023.100185] [Medline: 39845855]
27. Abd-Alrazaq A, AlSaad R, Alhuwail D, Ahmed A, Healy PM, Latifi S, et al. Large language models in medical education: opportunities, challenges, and future directions. JMIR Med Educ 2023 Jun 01;9:e48291 [FREE Full text] [doi: 10.2196/48291] [Medline: 37261894]
28. Burke HB, Hoang A, Lopreiato JO, King H, Hemmer P, Montgomery M, et al. Assessing the ability of a large language model to score free-text medical student clinical notes: quantitative study. JMIR Med Educ 2024 Jul 25;10:e56342 [FREE Full text] [doi: 10.2196/56342] [Medline: 39118469]
29. Choudhury A, Chaudhry Z. Large language models and user trust: consequence of self-referential learning loop and the deskilling of health care professionals. J Med Internet Res 2024 Apr 25;26:e56764 [FREE Full text] [doi: 10.2196/56764] [Medline: 38662419]
30. Izquierdo-Condoy JS, Arias-Intriago M, Tello-De-la-Torre A, Busch F, Ortiz-Prado E. Generative artificial intelligence in medical education: enhancing critical thinking or undermining cognitive autonomy? J Med Internet Res 2025 Nov 03;27:e76340 [FREE Full text] [doi: 10.2196/76340] [Medline: 41183320]

XSL•FO
RenderX

31. Tamkin A, McCain M, Handa K, Durmus E, Lovitt L, Rathi A, et al. Clio: privacy-preserving insights into real-world AI use. arXiv Preprint posted online on December 18, 2024. [doi: 10.48550/arXiv.2412.13678]
32. Google Gemini. URL: https://gemini.google.com [accessed 2026-06-05]

Abbreviations

AI: artificial intelligence
API: application programming interface
GAS: Google Apps Script
LLM: large language model
PHI: protected health information
RW: reflective writing

Edited by T Leung; submitted 12.May.2025; peer-reviewed by M Koda; comments to author 09.Dec.2025; accepted 28.May.2026; published 15.Jun.2026.

Please cite as:
Tonouchi Y, Nakai S, Murakami K, Kataoka Y
Developing a Large Language Model–Based Feedback System for Case Report Writing in Rehabilitation Education: Tutorial
JMIR Med Educ 2026;12:e76924
URL: https://mededu.jmir.org/2026/1/e76924
doi:10.2196/76924
PMID:

©Yuuto Tonouchi, Shunsuke Nakai, Kayo Murakami, Yuki Kataoka. Originally published in JMIR Medical Education (https://mededu.jmir.org), 15.Jun.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Education, is properly cited. The complete bibliographic information, a link to the original publication on https://mededu.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Tutorial

How to Conduct Usability Testing of Medical Education Simulations: Methods Tutorial Using a Telemedicine Simulation for Maternal Care

Blake J Lesselroth[1,2], MD, MBI; Alexander Douglas[3], DO; Helen Monkman[1,2], PhD, MA; Romaric Marcilly[4,5], PhD; Karalane Bellavia[1], BA; Mataeo Anderson[6], BS; Jacob Van Buren[7], MD; Charles Parsons[8], BS; Sonakshi Mohanty[9], BS; Juell Homco[1], PhD, MPH; Morgan Richards[7], DO; Karen P Gold[7], MD, MSci

[1]Department of Medical Informatics, School of Community Medicine, University of Oklahoma - Tulsa, Tulsa, OK, United States

[2]School of Health Information Science, University of Victoria, Victoria, BC, Canada

[3]Department of Internal Medicine, Oklahoma State University, Tulsa, OK, United States

[4]Univ. Lille, CHU Lille, ULR 2694 - METRICS: Évaluation des technologies de santé et des pratiques médicales, F-59000, Lille, France

[5]Inserm, CIC-IT 1403, F-59000, Lille, France

[6]University of Oklahoma Polytechnic Institute, University of Oklahoma, Tulsa, OK, United States

[7]Department of Obstetrics and Gynaecology, School of Community Medicine, University of Oklahoma - Tulsa, Tulsa, OK, United States

[8]School of Community Medicine, University of Oklahoma - Tulsa, Tulsa, OK, United States

[9]Department of Computer Science and Design, University of California San Diego, San Diego, CA, United States

Corresponding Author:
Blake J Lesselroth, MD, MBI
Department of Medical Informatics
School of Community Medicine
University of Oklahoma - Tulsa
Schusterman Center
4502 E. 41st. Street
Tulsa, OK, 74135
United States
Phone: 1 9713033627
Fax: 1 9186603059
Email: Blake-Lesselroth@ou.edu

Abstract

Background: Clinical educators regularly use simulation-based education to help learners develop clinical reasoning, procedural skills, and communication strategies. However, differences between simulated and live clinical environments, such as missing case details or workflow interruptions, can reduce realism and divert learners' attention from the intended learning goals. While published design guidelines recommend pilot testing simulations, they do not explain how to systematically identify usability problems. Human factors evaluation methods can address this gap by identifying issues before piloting and providing design insights.

Objective: This tutorial describes our usability testing protocol for evaluating educational simulations before piloting with learners. Using a telemedicine maternal care simulation as an example, we show how mixed methods usability testing can identify technical, workflow, and assessment-related issues in simulation design.

Methods: We recruited obstetricians and family practitioners to test our simulation. To measure multiple usability dimensions concurrently, we used several published instruments, including an Agile Task Analysis (ATA), the Single Ease Question (SEQ), an adapted System Usability Scale (SUS), the Satisfaction With Simulation Experience Scale (SSES), and the NASA (National Aeronautics and Space Administration) Task Load Index (NASA-TLX). We also developed a competency assessment rubric to measure telemedicine competencies. We recorded observations on the ATA while testers used a think-aloud protocol. We then debriefed testers and administered the posttest questionnaires. We summarized quantitative data using descriptive statistics and analyzed qualitative data using theory-based deductive coding with published usability heuristics.

XSL•FO
RenderX

Results: Twelve testers identified 52 usability issues over 2 testing cycles; 15 issues required correction before piloting with learners. Revisions included changes to standardized patient scripts, synthetic patient data, and scoring rubric instructions. Testers assigned a median SEQ score of 6 or higher to 4 of the 6 workflow steps; most struggled with unfamiliar telemedicine tasks or completing a virtual physical examination. SUS scores ranged from 65 to 100. The median SUS score was 77.5, indicating above-average usability. The NASA-TLX subscale scores were highest for mental workload (median 60, IQR 56.3-63.8) compared with other subscales, such as physical or temporal workload (median 15.00, IQR 11.3-18.8 and median 20.00, IQR 11.3-43.8, respectively), suggesting the simulation imposed a high mental demand. The overall SSES score of 43.5 out of 50 suggested faculty testers thought the simulation offered valuable learning opportunities.

Conclusions: Following usability testing, we successfully piloted the simulation with 23 obstetrics and family medicine residents without encountering any implementation issues. Incorporating usability evaluation methods throughout the simulation development life cycle helped identify and mitigate design problems that might otherwise limit simulation fidelity and instructional effectiveness. Our tutorial showcases a modular mixed methods approach that can be adapted to study a range of simulation types and research questions related to workflow, technology, and learning objectives.

(JMIR Med Educ 2026;12:e102697) doi:10.2196/102697

KEYWORDS

human factors engineering; maternal health services telehealth simulation; medical education; simulation testing; simulation training; telehealth competencies; telemedicine; usability testing; user-centered design

Introduction

Background

Health professional programs frequently use simulation-based training as an instructional strategy to introduce new concepts, enrich learning, and provide trainees with feedback [1,2]. Educational simulations are faculty-designed instructional methods that provide learners with realistic, immersive clinical situations to develop decision-making skills [1]. The simulation may include human actors representing specific roles (eg, standardized patients [SPs] and health care professionals) or technologies that mimic people and the environment (eg, augmented reality [AR], virtual reality [VR], AI, and manikins). Simulations are an important bridge between the classroom and the bedside, providing a safe yet realistic environment for trainees to learn material, test strategies, and develop reasoning skills before real-world practice [3]. Educators typically design simulations around scenarios with a clinical challenge and a normative workflow. The scenarios, however, should be flexible and responsive, allowing learners to explore, adapt, and improvise [2,4]. It is also critical that simulations include context-sensitive feedback or guided reflection informed by educational theory and learning objectives [2].

To meet instructional objectives, experts and professional societies have published best-practice recommendations for simulation development [2,5-7]. These recommendations are largely conceptual and lack the granularity needed to guide design, testing, and implementation. Conceptual approaches to design alone do not ensure simulation quality [8,9]. It can be challenging to preserve realism and consistency given the many variables affecting case fidelity and the achievement of learning objectives [4]. Simulationists must ensure the accuracy of the scenario, environment, and digital or practical assets, such as medical equipment and synthetic patient data (ie, ecological validity) [9]. Educators must evaluate learners in ways that align with learning objectives (ie, content validity) [4,10]. Finally, learners need personalized feedback addressing the ability under assessment (ie, construct validity) [2,10]. Optimizing validity often requires educators to limit the number and types of discrepancies between the simulation and real-world clinical practice [11,12].

Discrepancies between the simulation and the real world, including design omissions (eg, incomplete scenarios, data, or objectives) or equipment issues, can limit the simulation's usability. These problems can consume cognitive resources that learners would otherwise apply to pedagogical tasks [11]. It therefore behooves developers to rigorously test the simulation from each person's perspective (eg, the learner, the instructor, and SPs) and correct usability flaws before piloting with intended learners [4]. We authored this tutorial to share our methods and demystify the testing process.

Problem Statement

Simulation development guidelines recommend rehearsing cases with representative learners and collecting feedback from content experts, colleagues, technical teams, and participants [2,4,8,9]. Yet simulation guidelines rarely include methods for testing design features or collecting performance data before piloting [2,4,8,13]. Recently, human factors researchers have published reports trying to address this methodological gap by applying usability measurement techniques [8,14,15]. The ISO defines usability as the "extent to which a system, product or service can be used by specified users to achieve specified goals with effectiveness, efficiency, and satisfaction in a specified context of use" [16]. Utility is what a system allows users to do, whereas usability is how easily and efficiently users can access and operate these functionalities. Together, these qualities contribute to the overall user experience (ie, "users' perceptions and responses resulting from the use and/or anticipated use of a system, product or service").

Usability testing is an evaluation method in which users interact with a product under realistic conditions to determine how effectively, efficiently, and satisfactorily a set of tasks can be completed [17]. Da Silva and Dubrowski [8], for example, used mixed methods to improve 2 simulations over 5 iterations, demonstrating how usability testing findings can inform

XSL•FO
RenderX

revisions. Still, their approach required large numbers of learners and relied, in part, on learner self-reports—an often unreliable proxy for educational effectiveness [18]. We outline here a pragmatic strategy and toolkit to address these limitations.

Our team at the University of Oklahoma designed a series of simulations to teach telemedicine competencies to residents providing online maternal health care. We adapted published usability methods to create a simulation evaluation protocol. Our protocol had 2 aims: to identify technical and content issues requiring immediate correction and to gauge the completeness and difficulty of the clinical tasks. We focused on the following three usability issue types: (1) technology failures that could disrupt workflow; (2) missing case details, patient-script language, or synthetic clinical data crucial for medical reasoning; and (3) faculty scoring guide instructions that might be ambiguous or difficult to apply.

Tutorial Aims and Goals

Our high-level aim with this tutorial is to provide medical educators, human factors experts, and simulationists with tools and instructions for usability testing educational simulations. Our learning objectives are to (1) explain how we adapted usability methods for this context, (2) share our data collection instruments with readers to facilitate replication (Multimedia Appendices 1-8), and (3) demonstrate our methods with a use case to illustrate how we applied our results (Textbox 1).

Textbox 1. Tutorial learning objectives.

After reading this tutorial, readers should be able to do the following:

- Describe how to apply usability testing frameworks when designing instructional medical simulations.
- Adapt published usability testing methods and instruments to test educational simulations.
- Explain how to interpret usability findings to improve simulation consistency and teaching effectiveness.

While the objectives for usability testing are straightforward, we recommend articulating a priori research questions to guide protocol design and measure selection. Readers replicating our methods will likely formulate research questions relevant to their simulation. For our study, we had the following 3 questions:

- What failure modes interfere with the function of the simulation?
- Can faculty evaluators use the scoring guide to measure learner performance?
- Is the simulation difficulty appropriate for the intended learners?

Methods

Theoretical Framework: Usability and User Experience

Usability and user experience are not intrinsic properties of a product or a simulation; instead, they are emergent properties of the work system, arising from the interaction of multiple elements: technology, learners, instructors, tasks, environment, and organizational context [12,19]. This holistic perspective emphasizes that evaluating usability requires examining the entire work system, not the interface or equipment in isolation. Consequently, usability evaluation of simulations should aim to replicate realistic conditions regarding participants, task complexity, environmental constraints, and instructor support. Regarding educational simulations, this means evaluating not only the tools and technology but also faculty roles, faculty-learner interactions, and scenario details [20].

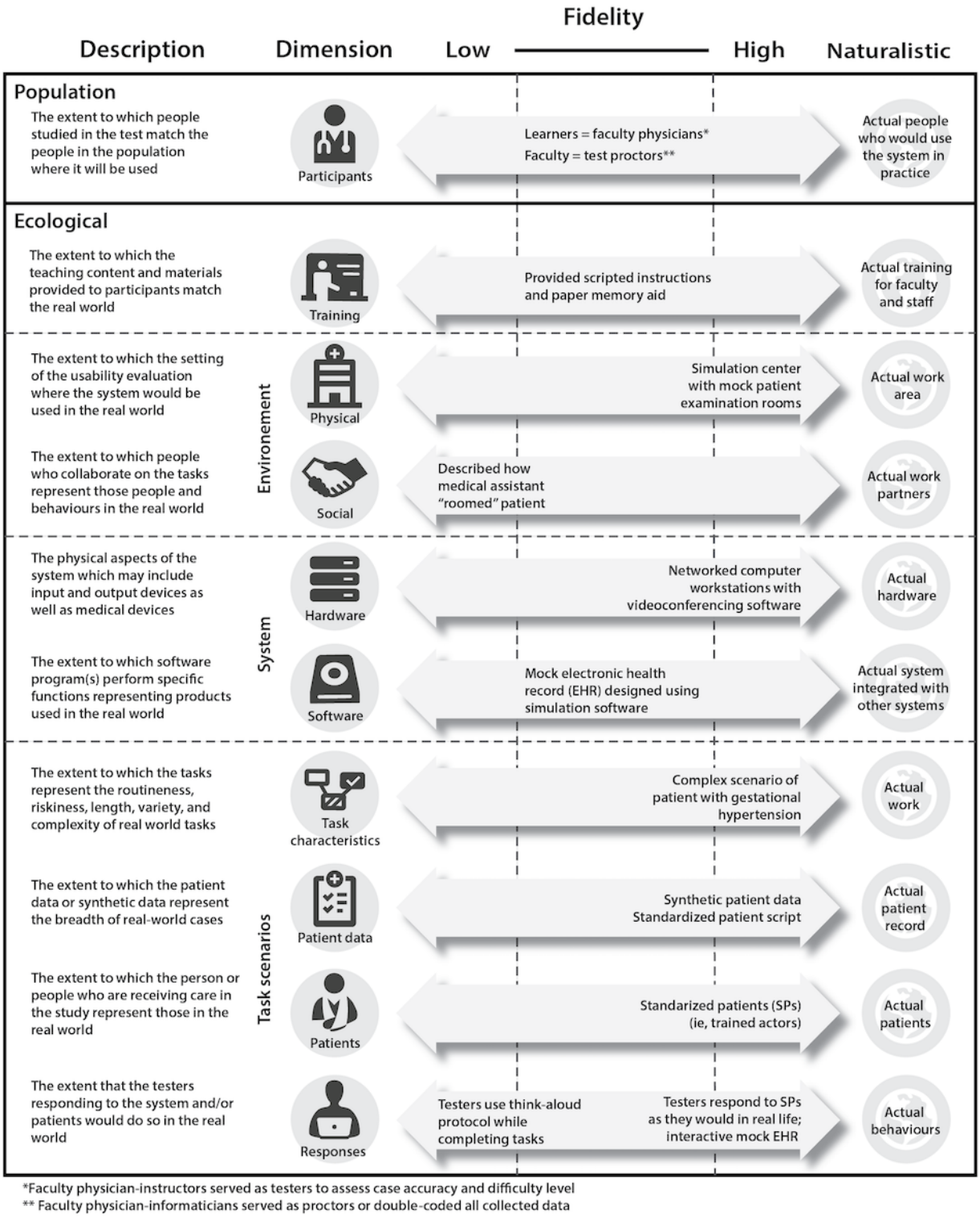
It is important to clarify how simulation usability can influence learners' performance. Human-computer interaction shows that poor user experience increases the likelihood of errors [21-23]. Usability problems add cognitive load, forcing users to dedicate attention, working memory, and problem-solving capacity to compensate for suboptimal design. As a result, learners confronted with unintentional simulation design problems use cognitive resources to manage these problems rather than engaging with the learning objectives. Usability issues can reduce participant engagement, hinder learning, and diminish the perceived credibility of the program [12]. Hence, while simulation guidelines recommend piloting with learners to uncover problems, we believe usability issues should be addressed before piloting or implementation [2,6].

We needed to solve several unique problems in educational simulations that are atypical for usability testing of medical devices or software. First, we needed to calibrate scenario difficulty to participants' experience level (eg, residents) and expertise (eg, obstetrics and gynecology [OB/GYN], family medicine). Second, faculty observers needed to evaluate learners for specific telehealth competencies to identify skill gaps. Third, we needed to include built-in challenges for the learner. Thus, we needed faculty participants to serve as observers to provide feedback on the scoring guide while gathering tester feedback on the simulation.

The theoretical framework informing our current approach is based on a model of external validity we created for usability evaluations of health information technology (Figure 1) [20]. External validity comprises (1) population validity (ie, the extent to which the sample represents the population) and (2) ecological validity (ie, the extent to which the setting matches the real world). While both types of validity are important, the fidelity level for each dimension depends on the evaluation objectives [12].

XSL•FO
RenderX

Figure 1. External validity framework for the application of usability testing in health. We adapted the figure to illustrate fidelity levels on a continuum for each dimension of our educational simulation.



*Faculty physician-instructors served as testers to assess case accuracy and difficulty level
** Faculty physician-informaticians served as proctors or double-coded all collected data

When selecting the degree of ecological validity, we needed to consider learner attributes, presimulation didactics, the clinical environment (ie, physical and social), the technology (ie, hardware and software), and the scenario (ie, task characteristics, patient data, patients, and responses). Figure 1 [20] shows how we mapped model constructs to our use case. Poor usability arising from interactions between dimensions could manifest as unclear instructions, confusing interface layouts, incomplete patient scripts, mismatched equipment, or poorly structured task sequences. We chose to test our simulation with faculty rather than with learners. This was an intentional choice to assess case accuracy and to determine whether faculty believed the tasks were appropriately calibrated for resident learners.

XSL•FO
RenderX

Use Case for Testing: A Telehealth Simulation for Residents Providing Maternal Care

During the COVID-19 pandemic, organizations embraced virtual care to enact social distancing precautions without limiting access [24-29]. This expansion highlighted educational gaps; practitioners were uncomfortable providing online care, and few training programs offered telemedicine instruction [30-35]. In response to this emerging need, the Association of American Medical Colleges (AAMC) published a white paper outlining a framework for telemedicine training in the United States [36]. The AAMC identified 20 competencies across six domains: (1) patient safety, (2) access and equity, (3) communication, (4) data collection, (5) technology, and (6) ethical and legal requirements. Since publication, medical schools and residencies have endeavored to teach this material using readings, didactics, and simulations [35,37-39].

At the University of Oklahoma School of Community Medicine (OUSCM), the Department of Medical Informatics partnered with the Department of Obstetrics and Gynecology to develop a telemedicine training program for maternal care [40]. Designed for OB/GYN and family and community medicine (FCM) residents, our simulations combined telemedicine instruction with vignettes managing maternal care conditions [41-43]. By training future physicians to provide online care to expectant mothers, we hoped to close a supply-demand mismatch contributing to local maternal care “deserts” [40,44-49].

Our educational program created the need—and the opportunity—to conduct usability tests of simulations before piloting with residents. The technology-forward nature of telemedicine inspired our team to adapt usability methods typically applied to health information technology. Although we have published preliminary reports describing individual tools, this is the first tutorial comprehensively describing our usability testing methods, findings, and downstream piloting experiences [41-43]. For illustrative purposes, we demonstrate our methods using a simulation of a fictional patient presenting for a telemedicine encounter to manage gestational hypertension.

Study Setting and Research Team

Our research team conducted all testing at our university’s simulation center in Tulsa, Oklahoma, United States. The team included 2 obstetricians (MR, KB) to provide subject-matter expertise, 4 informaticians (BJL, HM, RM, and JH) and 2 data analysts to design the test protocol (MA and SM), a proctor to supervise each test and code the data (MA or CP), and at least 1 additional observer (typically another member of the research team) to double-code observations. The proctor and observers played the role of faculty during simulation testing. The team also included an SP and a simulationist (KB). The simulationist is an expert in simulation design, SP education, and script writing. The simulation center was equipped with mock clinic rooms, remote monitoring stations, simulation software, and computers with videoconferencing software. Because this was a telemedicine simulation, our SPs connected to the simulation center remotely via videoconference.

To recreate our described approach, we recommend the following resources: a dedicated space with internet access, a trained team member to serve as the SP, videoconferencing software with recording capabilities, and a proctor for the session. We also recommend recruiting team members to observe and double-code findings. We found it helpful to have trained faculty on-site to debrief testers. While the examination rooms in our center increased the ecological validity of our testing, they were not required.

Data Collection Instruments

Overview

In this section, we describe each data collection instrument and our rationale for its inclusion (Table 1). Multimedia Appendices 1-8 provide instrument examples; Multimedia Appendices 9 and 10 are examples tools used for our specific simulation. We sought validated questionnaires to gather cross-sectional data on multiple usability dimensions. In our case, a “minimally viable product” for an educational simulation needed to satisfy three requirements: (1) the learner must be able to complete the scenario without interruption, (2) the learner must be able to review synthetic patient data needed for a clinical assessment, and (3) the faculty must be able to calculate competency scores for each learner.

XSL•FO
RenderX

Table 1. List of testing instruments used in our usability assessment protocol. We adapted usability dimensions and definitions from the ISO [16], Quesenbery [50], and Davis [51]. All instruments are available in Multimedia Appendices 1-9.

Instrument	Data collected	Measure type	Usability dimension	Usability dimension definition
Demographic screener (Multimedia Appendix 1)	Tester demographics, technology literacy, and attitudes and perceptions	Qualitative, binary (yes/no), and ordinal (Likert-type)	N/A[a]	N/A
ATA[b] (Multimedia Appendix 2)	Simulation step completion rates	Ordinal (trichotomous)	Effective (learner)	How completely and accurately learners can complete a task to reach their goals [50]
ATA (Multimedia Appendices 2, 7, and 8)	Usability issues	Qualitative	Error tolerance	How well a design prevents errors or helps users recover from mistakes [50]
SEQ[c] (Multimedia Appendix 3)	Task difficulty	Ordinal (Likert-type)	Easy to learn	The degree to which a person believes using a system would be free of effort [50]
Telemedicine competency rubric (Multimedia Appendix 9)	Clinical task completion rates, and competency scores	Task count and competency score	Effective (faculty)	How completely and accurately faculty could score the telemedicine rubric
Adapted SUS[d] (Multimedia Appendix 4)	User perceptions	Ordinal (Likert-type)	Ease of use	How intuitive the product is, ensuring learners can understand expectations with little to no additional instruction [51,52]
SSES[e] (Multimedia Appendix 4)	Learner perceptions	Ordinal (Likert-type)	Satisfaction	How relevant, valuable, and supportive the simulation was for acquiring skills and knowledge [53]
NASA-TLX[f] (Multimedia Appendix 5)	Cognitive load	Ordinal	Efficiency	The resources expended, including time, cognitive load, and effort in relation to the accuracy and completeness of goals [54]
Debrief interview guide (Multimedia Appendix 6)	Usability issues	Qualitative	Satisfaction	How pleasant, satisfying, and aesthetically pleasing the product is [16]
Debrief interview guide (Multimedia Appendix 6)	Usability issues	Qualitative	Fidelity	How completely and accurately the tasks align with real-world scenarios [11,55]

[a]N/A: not applicable.

[b]ATA: Agile Task Analysis.

[c]SEQ: Single Ease Question.

[d]SUS: System Usability Scale.

[e]SSES: Satisfaction With Simulation Experience Scale.

[f]NASA-TLX: NASA (National Aeronautics and Space Administration) Task Load Index.

Demographics Questionnaire

When considering population validity, it is helpful to create a user persona and recruit testers matching the persona's characteristics. Learners are the target users for educational simulations and can provide useful feedback. While it may seem counterintuitive, it has been our experience that faculty can also provide useful feedback. They have the domain expertise to critique simulation fidelity, task complexity, the need for additional scaffolding, and alignment with learning objectives [4]. Hence, we recommend recruiting faculty testers first, followed by learners as time and resources permit. For all recruited participants, we gathered descriptive demographic data to share with stakeholders (eg, educational leadership, executive sponsors, collaborative partners, funders) when presenting findings. We therefore used a demographics questionnaire with 3 sections: (1) tester demographics, (2) a technology literacy assessment, and (3) knowledge and attitude items about telemedicine (Multimedia Appendix 1). We confirmed that all testers provided maternal medical care. We asked whether they regularly used a computer, mobile technology, or videoconferencing software. Finally, we included a short questionnaire covering prior experiences with telemedicine. For the latter section, we based the content domains on research exploring the intersection between clinician attributes and successful telemedicine encounters [56].

Agile Task Analysis

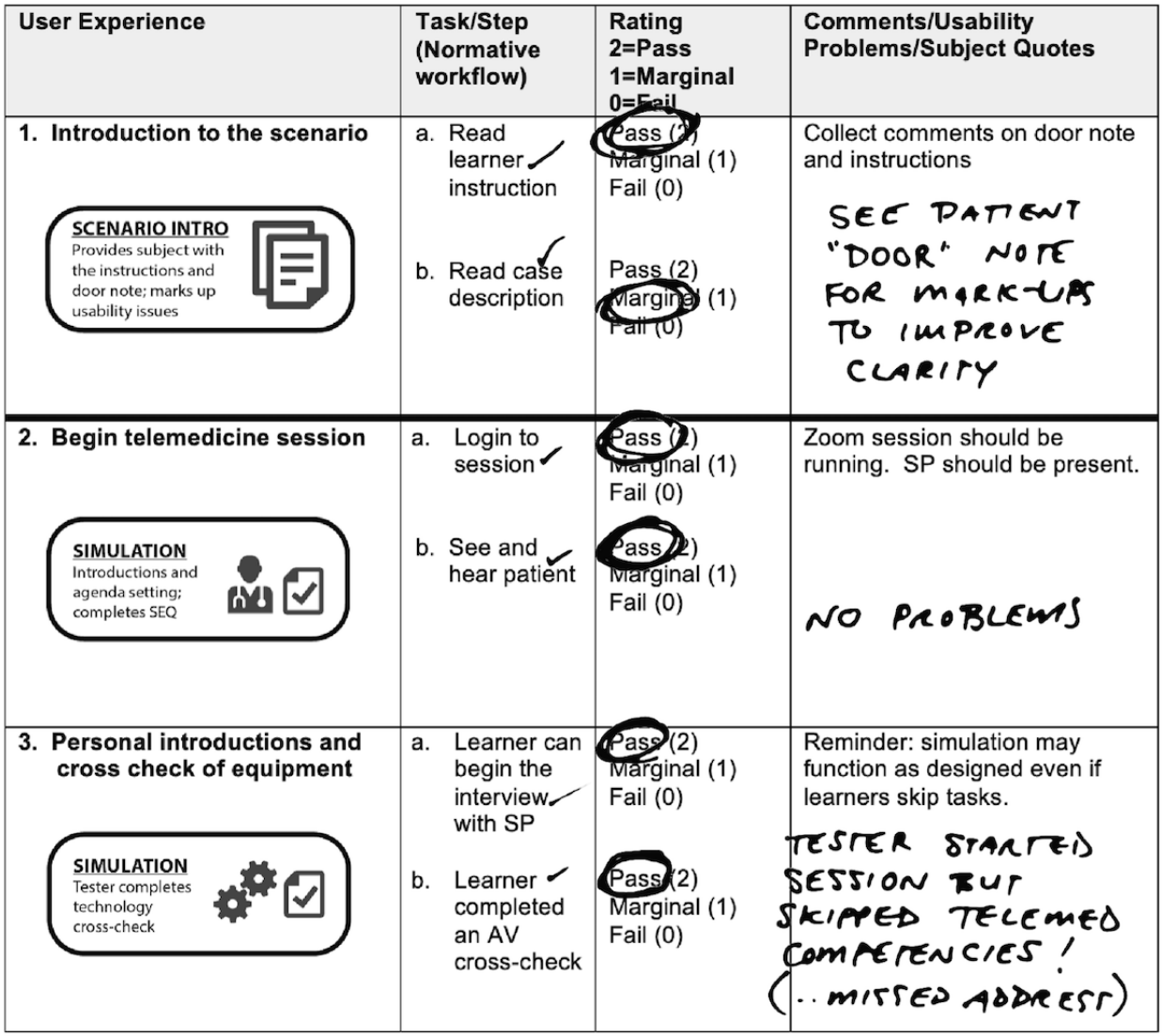
Agile Task Analysis (ATA) is a hybrid method that combines hierarchical task analysis (HTA) with qualitative observation during simulation testing (Multimedia Appendix 2) [57]. In a standard HTA, the researcher breaks complex activities into tasks and goals to understand how work is performed and what tools and information are required for each task [58].

XSL•FO
RenderX

Observations may include task lists, narrative details, workflows, and examples of tools, paper forms, or technology [59]. Borrowing from HTA, the ATA uses this framework to organize scenarios for simulation testing [57]. It also includes a data collection form that follows the scenario's normative workflow, allowing observers to record observations concurrently. Data can be quantitative or qualitative, including task completion rates, observed errors, usability issues, or user observations, such as quotes or design recommendations.

Researchers can modify the ATA based on the type of product or research questions. For example, the ATA can include screenshots of technology interfaces, informational handouts, or blueprints of the physical environment [57]. For this study, we listed six concept-level steps in our telemedicine workflow: (1) read learner instructions, (2) start a telemedicine session, (3) complete standard telemedicine tasks, (4) collect a patient history, (5) review examination and objective data, and (6) develop an assessment and plan (Figure 2 [57]). For each step, the research team tracked simulation performance using an ordinal trichotomous scale (ie, pass, marginal, and fail). We coded a step as "pass" if the tester completed it without assistance, "marginal" if the tester required proctor assistance, or "fail" if the tester encountered a critical fault in the simulation and needed to bypass the step entirely.

Figure 2. Excerpt from the Agile Task Analysis (ATA) form we used to evaluate our telehealth maternal care education simulation.



Telemedicine Competency Rubric

Educational simulations are often designed to challenge learners or to identify knowledge or skill gaps. A simulation step can work as designed even if the tester neglects a task or performs it incorrectly. We did not expect testers to complete every task. To differentiate simulation steps from telemedicine tasks, we included a telemedicine competency rubric (ie, faculty scoring guide) in the ATA to record testers' performance on telemedicine tasks during the simulation (Multimedia Appendix 2) [41].

Our telemedicine competency rubric was based on the AAMC white paper and operationalized measurement for 8 of the 20 listed telehealth competencies [36]. We selected 8 competencies for our simulation: 2 focusing on patient safety, 2 on

XSL•FO
RenderX

communication skills, 1 on data collection, 2 on telehealth technologies, and 1 on ethical requirements. The AAMC white paper includes definitions for each competency but not context-specific details for clinical care or explicit measurement strategies. We were forced to create a novel rubric that organized each competency into short sets of related, contextually embedded, and directly observable tasks (Multimedia Appendices 2 and 9) [60]. We summed correctly performed tasks to calculate composite competency scores. We operationalized competency into 3 levels: entrustable, approaching entrustment, and not entrustable [60,61]. A learner was "entrustable" if faculty believed the learner could be entrusted to complete all activities within a content domain without requiring direct supervision. Faculty marked a learner as "approaching entrustment" if they believed the learner needed some supervision to complete related clinical tasks. The tester needed to complete at least half of the tasks within a domain to be approaching entrustment. A learner was "not entrustable" if faculty believed the learner must be directly supervised throughout the activity. The proctor marked a tester "not entrustable" if the tester completed fewer than 2 tasks within a telemedicine domain. As a first step toward establishing interrater reliability, we required multiple research team members to score each tester independently during data collection and then resolved discrepancies through discussion and consensus (details are provided in the Data Recording and Adjudication section).

Single Ease Question

The Single Ease Question (SEQ) is a 7-point Likert-type question administered immediately after a task, asking the tester to rate the task's difficulty [23,62]. Tasks that score higher than 6 are considered "very easy" [63]. The SEQ offers insights that posttest questionnaires do not by dissociating individual steps from the overall user experience [64,65]. We included 1 SEQ after each of the 6 workflow tasks and 1 at the conclusion of the simulation (Multimedia Appendix 3).

Adapted System Usability Scale

The System Usability Scale (SUS) is a 10-item, 5-point Likert-type scale assessing users' self-reported ease of use with a product [23,52,65,66]. Researchers convert respondents' answers to a 100-point scale. Based on the distribution of scores across industries, an average score is 68 [52]. We adapted the SUS language to evaluate testers' perceptions of a simulation rather than a technology (Multimedia Appendix 4). Prior studies of the SUS show that substituting the word "system" with a contextually appropriate product name or referent does not degrade the psychometric properties of the instrument [67].

Satisfaction With Simulation Experience Scale

The Satisfaction With Simulation Experience Scale (SSES) is an 18-item instrument used to evaluate learner satisfaction with educational simulations [53]. It measures learners' engagement and perceptions of simulation effectiveness [53,68]. Although not a traditional usability measure, we used the SSES to capture faculty and resident perceptions of the simulation's educational value and treated it as a proxy for satisfaction. The SSES complements other usability measures (eg, SEQ, SUS, and think-aloud) but does not replace them. Items are grouped into 3 subscales: debriefing and reflection, clinical reasoning, and clinical learning. For this study, we adapted questions from the SSES clinical reasoning and clinical learning subscales, modifying the language for a telemedicine simulation (Multimedia Appendix 4).

NASA Task Load Index

To measure perceived simulation complexity, we administered the NASA (National Aeronautics and Space Administration) Task Load Index (NASA-TLX) after the simulation (Multimedia Appendix 5) [54]. The NASA-TLX was developed by NASA to measure perceived workload and cognitive overhead associated with a set of related tasks [69]. There are 6 subscales: mental demands, physical demands, temporal demands, perceived performance, effort, and frustration [54]. Each dimension is scored on a 100-point scale. Although there are no formal cut points, studies suggest scores ≤30 indicate a low workload, scores between 31 and 50 indicate a moderate workload, and scores >50 indicate a high workload [54,69-71]. The original NASA-TLX requires participants to assign a weight to each subscale. Researchers often omit the weighting step and use raw scores [69]. We used unweighted scores in our analysis.

Debrief Semistructured Interview

We used a semistructured debriefing guide to gather testers' final reflections (Multimedia Appendix 6). We designed our guide based on the Rapid Usability Evaluation method by Russ et al [72] and *Rocket Surgery Made Easy* [73] by Steve Krug. In both cases, the authors recommend asking probing questions during the observation session immediately following tasks and limiting the debrief to 1 or 2 open-ended follow-up questions. The researcher can use these questions to clarify usability issues or invite the participant to suggest changes for improving the design. We used questions and probes to explore the simulation's difficulty, the root causes of usability issues, and opportunities to improve the learning experience. We recorded responses in our usability issue log.

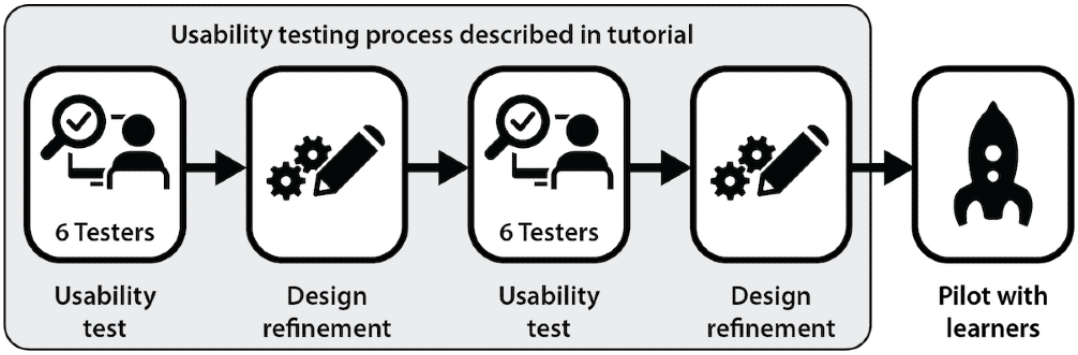
Testing Protocol

Overview

This section describes our testing protocol (Figure 3). While our general approach can be adapted to a range of educational simulations, the specifics of each simulation will dictate the number of testing cycles. We organized testing into sprints, gathering feedback from 6 testers before iterating on the design. We conducted 2 rounds of simulation testing and revisions (ie, sprints) to refine the simulation. We used new testers for each testing cycle to avoid data contamination. We conducted 2 refinement sessions, once at the midpoint of testing and once before piloting with target learners. We avoided making refinements to the simulation between refinement sessions.

XSL•FO
RenderX

Figure 3. Testing and refinement strategy. We used new testers for each round. The research team, which included content experts, served as faculty observers.



Recruitment

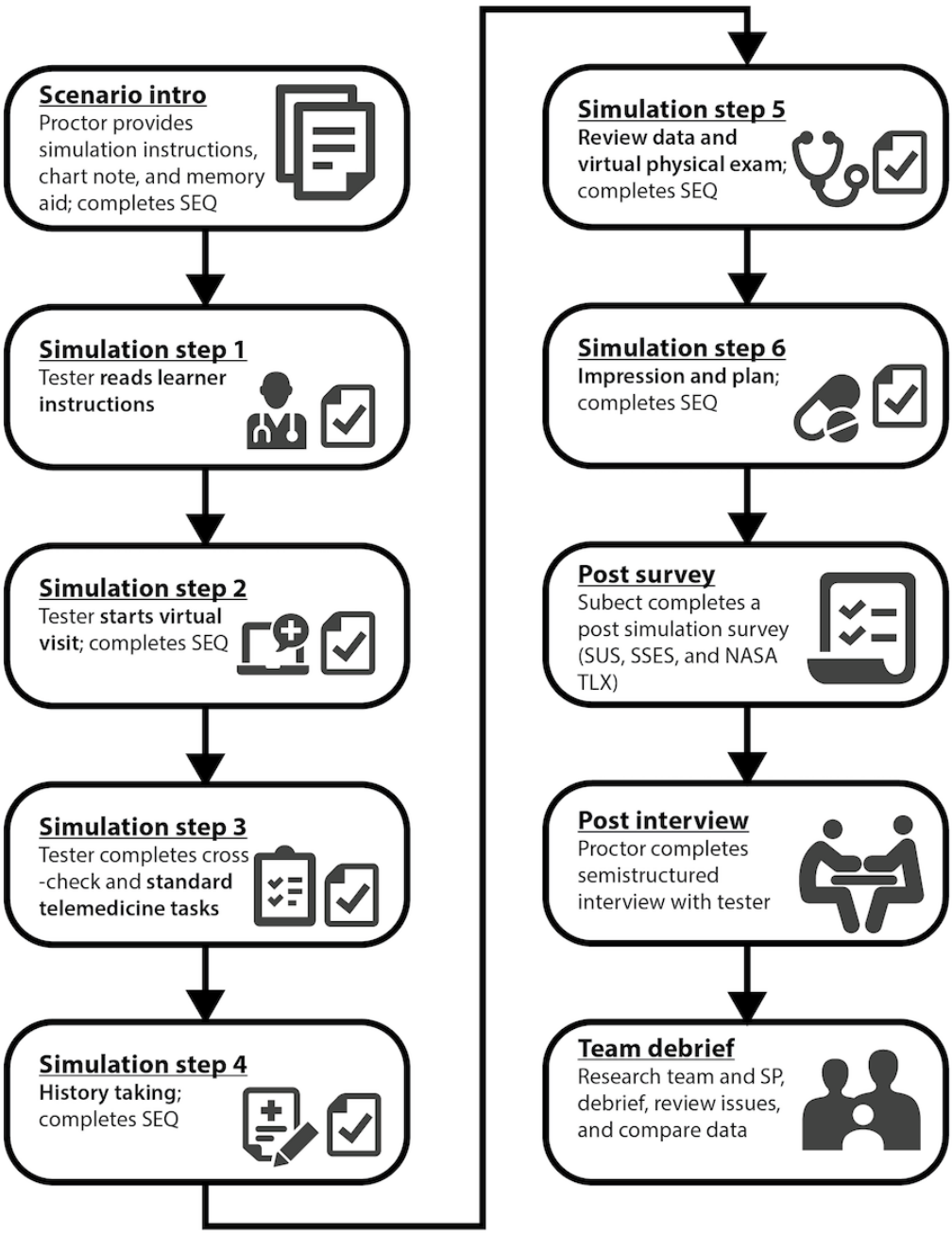
We used a purposive recruitment strategy, targeting faculty and residents to test the simulation and provide feedback on the scenario's accuracy and fidelity. Faculty and residents were eligible for inclusion if they (1) held an appointment at our institution, (2) had completed or were participating in a residency in obstetrics or family medicine, and (3) maintained an active obstetrics practice. Residents in obstetrics and family medicine are required to see obstetrics patients in a continuity clinic during training; they practice under the license of their supervising attending. We contacted every eligible practitioner at our institution and recruited the first 6 volunteers for each testing round, aiming to balance the number of obstetricians and family practitioners.

Tester Orientation

A member of the research team provided each tester with an overview of the study, including the curriculum's purpose and the high-level goals for the simulation (Figure 4). We secured verbal and written consent from each tester during the orientation to observe and record testing and tester insights. The demographic screening questionnaire included consent language and an explanation of how the data would be used. We entered data, excluding unique identifiers, into a spreadsheet for subsequent analysis. We did not include a telemedicine didactic before testing. We did not disclose scenario details, workflow expectations, or the competencies being evaluated. A proctor sat with the tester in a mock examination room with a computer workstation, microphones, and closed-circuit cameras (Figure 5). A proctor without clinical expertise can be advantageous as a neutral observer. Before starting the simulation, the proctor furnished the tester with the screening questionnaire. We adapted the Rapid Usability Evaluation method by Russ et al [72] and the methods outlined in *Rocket Surgery Made Easy* by Steve Krug [73]. The proctor read aloud a written script explaining the purpose of testing and the think-aloud technique (ie, testers verbalize their thoughts while interacting with the system and completing tasks) [74,75].

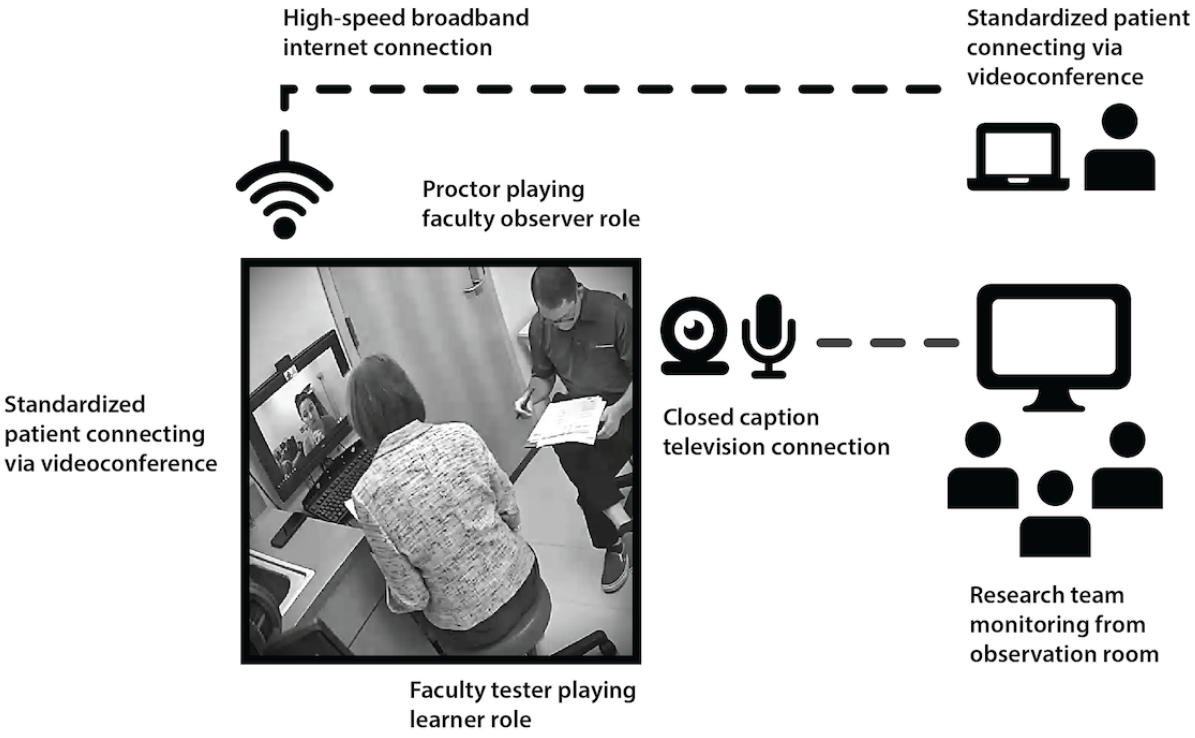
XSL•FO
RenderX

Figure 4. Schematic for the usability testing protocol. We conducted 2 rounds of formative usability testing before piloting with target learners. NASA-TLX: NASA Task Load Index; SEQ: Single Ease Question; SP: standardized patient; SSES: Satisfaction With Simulation Experience Scale; SUS: System Usability Scale.



XSL•FO
RenderX

Figure 5. Setting and configuration for usability testing.



Simulation Testing

During the simulation, the tester used the computer workstation with videoconferencing software (Zoom; Zoom Communications) to complete a mock encounter with our SP [76]. We furnished the tester with a summary of the case: a pregnant woman living in a rural community and presenting at 20 weeks' gestation for a routine follow-up appointment. The patient reported recording new high blood pressure values using a home digital blood pressure cuff.

In the first round, the proctor provided the tester with written case instructions, a mock clinical note with synthetic patient data, and a printed memory aid developed by the team to help with telemedicine tasks [77]. In the second round, rather than providing a mock clinical note, the tester reviewed synthetic patient data using a mock electronic health record (EHR) we designed with LearningSpace simulation software (Elevate Healthcare) [78]. We continued to give the tester written case instructions and our memory aid. The proctor did not provide technical instructions, additional medical information, or clinical feedback to the tester during the simulation. The other members of the research team watched the simulation on closed-circuit television and recorded observations simultaneously.

We expected the tester to establish rapport with the SP, troubleshoot technology problems, gather a history, review synthetic patient data, provide a clinical assessment, and communicate a treatment plan. The learner needed to gather enough clinical data to rule out preeclampsia, an obstetric emergency. During the simulation, the proctor recorded all observations on the ATA paper form. The proctor and observers recorded (1) whether each simulation step worked, (2) whether the tester had the opportunity to perform each competency-related task, (3) whether the tester performed the task successfully, and (4) whether the tester encountered any usability issues. Initially, the proctor asked the SEQ after each set of related tasks. However, we found this disrupted the simulation workflow and added cognitive overhead. Hence, the proctor asked all SEQs after the simulation.

Postsession Questionnaires and Semistructured Interviews

After the simulation, the proctor administered our postsession questionnaires (ie, SUS, SSES, and NASA-TLX) and conducted a semistructured interview (Multimedia Appendices 4-6). The interview provided an opportunity to explore areas of confusion, the root causes of usability issues, potential modifications to the simulation, and unmet educational needs.

Data Recording and Adjudication

For the rubric, the research team compared scores for each task and resolved discrepancies through discussion to reach consensus. When we identified vague or potentially confusing language, we revised the task description during the design and refinement phases. We did not calculate agreement statistics during testing. The Limitations section provides a more complete discussion of our rationale and next steps. We compiled these results into a "master" document. We then reviewed all qualitative data, including usability issues, observations, tester quotes, and interview responses. We summarized the qualitative findings in an issue log (Multimedia Appendices 7 and 8) for content analysis, coding, prioritization, solution ideation, and potential action.

Simulation Piloting

After we completed all usability testing and simulation refinements, we piloted the entire workshop (ie, a telemedicine didactic session, telemedicine simulation with the competency

XSL•FO
RenderX

rubric [Multimedia Appendix 9], and postsession questionnaire) with the residents. A detailed protocol for our workshop and simulation pilot falls outside the scope of this tutorial. We included a brief vignette of the pilot herewith (Textbox 2) to illustrate the downstream value of usability testing.

Textbox 2. Workshop and simulation vignette.

A complete description of workshop piloting with residents is beyond the scope of this report. We intend to describe the implementation methods and residents' scores in a companion manuscript. In this section, we include a short vignette of our piloting experience.

We conducted 2 separate pilots on different afternoons, 1 with 6 obstetrics and gynecology residents and 1 with 17 family and community medicine residents. All residents received a brief overview of the case, expectations for the simulated encounter, and a paper copy of the telemedicine memory aid. Residents used computer workstations in our testing center to interview a standardized patient and had online access to the mock electronic health record with synthetic patient data.

All participants completed the simulation without experiencing any unanticipated technical problems. Faculty were comfortable using the supplied evaluation rubric and had no difficulty providing individualized learner feedback after the simulation. Resident feedback on the simulation was generally favorable; most indicated the simulation provided a clear, safe, and constructive learning experience.

Data Analysis

Quantitative and Statistical Analysis

We calculated descriptive statistics for quantitative data, including demographic characteristics, simulation step completion rates, competency scores, SEQ scores, SUS scores, SSES scores, and raw NASA-TLX scores. Given the small sample size and the nonparametric nature of the data, we calculated raw counts, medians, and IQRs. Given the formative nature of the analysis, we did not conduct an a priori power analysis or calculate comparative statistics.

Qualitative Analysis

While a formal qualitative analysis is not always critical for pragmatic usability testing, we found it helpful for clarifying the root causes of issues, brainstorming potential solutions, and prioritizing fixes. Using coding methods described by Saldana, we conducted a content analysis of qualitative data gathered during the simulation and tester debriefing interviews [79]. We used simultaneous coding to assign a categorical usability code, a valence (ie, "positive" or "negative"), and a priority for action. For categorical codes, we used a deductive, theory-based coding strategy based on usability heuristics published by Zhang et al [80] (Multimedia Appendix 7) [21,81]. At least 2 members of the research team independently coded each finding. We then compared codes for concordance and resolved discrepancies through discussion. We assigned a provisional code with emergent supplementation to any findings that could not be classified according to the usability heuristics published by Zhang et al [80]. All codes and findings were recorded in a usability issue log (structural codes are provided in Multimedia Appendix 7; a sample log is provided in Multimedia Appendix 8).

Ethical Considerations

The testing described in this tutorial was reviewed by the University of Oklahoma Institutional Review Board and classified as a quality improvement initiative associated with normal educational activities. Nevertheless, we collected verbal and written consent from each tester to document the testing.

Results

Tester Demographics

We recruited 12 clinicians for usability testing (first round: 4 OB/GYNs and 2 FCMs; second round: 2 OB/GYNs and 4 FCMs). Table 2 provides demographic details.

XSL•FO
RenderX

Table 2. Tester demographic information and technology literacy.

Characteristic	Value	Mean (SD)	Median (IQR)
Demographic			
Age (years), range	25-60	38.9 (12.1)	37.0 (28.5-45.0)
Work experience (years), range	1-30	8.8 (11.1)	3.5 (1.0-10.5)
Specialty, n	OB/GYN[a]=6; FCM[b]=6	—[c]	—
Technology literacy, n			
Use computer once a week	Yes=12; no=0	—	—
Use portable technology	Yes=12; no=0	—	—
Use videoconferencing	Yes=12; no=0	—	—
Attitudes and perceptions, n			
Previously observed a telehealth encounter	Yes=11; no=1	—	—
Previously provided telehealth care	Yes=10; no=2	—	—
Previously received telehealth care	Yes=5; no=7	—	—
Received telemedicine instruction	Yes=4; no=8	—	—
Good understanding telemedicine practice	—	3.9 (0.70)	4.0 (4.0-4.0)
Familiar with types of technology	—	3.8 (0.8)	4.0 (3.8-4.0)
Think telemedicine is a good option	—	3.7 (1.3)	4.0 (4.0-4.0)
Will likely use telemedicine in practice	—	3.9 (0.8)	4.0 (4.0-4.0)

[a]OB/GYN: obstetrics and gynecology.

[b]FCM: family and community medicine.

[c]Not applicable.

Overall Results

Table 3 summarizes all usability measures, with key findings highlighted. We include detailed results for each measure in the following sections. The ATA identified 4 problematic steps in round 1 that we resolved by round 2. Across both rounds, we identified and addressed 15 critical usability issues. The virtual physical examination received the lowest SEQ ratings. During round 2, testers reported a median SUS score of 77.5 (IQR 73.1-80.0) the SUS and a median SSES score of 42 (IQR 41.0-46.75) out of a possible 50. The median NASA-TLX mental demand score in round 2 was 60.00 (IQR 56.3-63.8); a score >50 indicates a high workload. Proctors could not score all testers with the rubric during the first round. After modifying the rubric language during the first refinement phase, proctors successfully scored every tester.

XSL•FO
RenderX

Table 3. Summary table of quantitative results across instruments.

Usability dimension, instrument, and measured item	Frequency, n/N (%)	Median (IQR)	Maximum possible value
Effectiveness			
Agile Task Analysis			
Total number steps completed successfully	165/180 (92)	—[a]	—
Total number steps partially completed	11/180 (6)	—	—
Total number of failed steps	4/180 (2)	—	—
Error tolerance			
Agile Task Analysis			
Number of usability issues	52	—	—
Number of critical usability issues	15/52 (29)	—	—
Easy to learn			
Single Ease Question			
Reviewing materials	—	6.00 (6.00-6.25)	7.00
Starting encounter	—	6.00 (4.75-7.00)	7.00
Standardized telemedicine tasks	—	5.50 (5.00-6.25)	7.00
History collection	—	6.50 (5.00-7.00)	7.00
Virtual physical examination	—	4.50 (2.75-6.00)	7.00
Diagnosis and plan	—	6.00 (5.00-6.50)	7.00
Overall encounter	—	5.00 (5.00-6.00)	7.00
Ease of use			
Adapted System Usability Scale			
Total score round 1	—	73.80 (72.63-80.50)	100.00
Total score round 2	—	77.50 (73.13-80.00)	100.00
Total score overall	—	75.00 (71.88-81.25)	100.00
Satisfaction			
Satisfaction With Simulation Experience Scale			
Total score round 1	—	44.50 (39.5-45.0)	50.00
Total score round 2	—	42.00 (41.00-46.75)	50.00
Total score overall	—	43.50 (40.5-45.25)	50.00
Workload			
NASA[b] Task Load Index			
Physical demand round 1	—	7.50 (5.00-17.50)	100.00
Temporal demand round 1	—	12.50 (6.25-22.50)	100.00
Frustration level round 1	—	22.50 (10.00-42.50)	100.00
Perceived performance round 1	—	32.50 (26.25-35.00)	100.00
Mental demand round 1	—	37.50 (21.25-57.50)	100.00
Effort (mental and physical) round 1	—	42.50 (27.50-57.50)	100.00
Physical demand round 2	—	15.00 (11.30-18.80)	100.00
Temporal demand round 2	—	20.00 (11.30-43.80)	100.00
Frustration level round 2	—	25.00 (10.00-43.80)	100.00
Perceived performance round 2	—	32.50 (21.30-43.80)	100.00
Mental demand round 2	—	60.00 (56.30-63.80)	100.00

XSL•FO
RenderX

Usability dimension, instrument, and measured item	Frequency, n/N (%)	Median (IQR)	Maximum possible value
Effort (mental and physical) round 2	—	55.00 (46.30-63.80)	100.00
Mental demand overall	—	57.50 (23.75-61,25)	100.00
Effort (mental and physical) overall	—	50.00 (32.50-60.00)	100.00
Effectiveness			
Competency rubric[c]			
Incorporates telehealth	12/12 (100)	—	—
Attends to environment	11/12 (92)	—	—
Identifies and uses equipment	12/12 (100)	—	—
Develops rapport	12/12 (100)	—	—
Obtains and uses history	12/12 (100)	—	—
Troubleshoots technology	12/12 (100)	—	—
Prepares for escalating care	12/12 (100)	—	—
Adheres to professional requirements	10/12 (83)	—	—

[a]Not applicable.

[b]NASA: National Aeronautics and Space Administration.

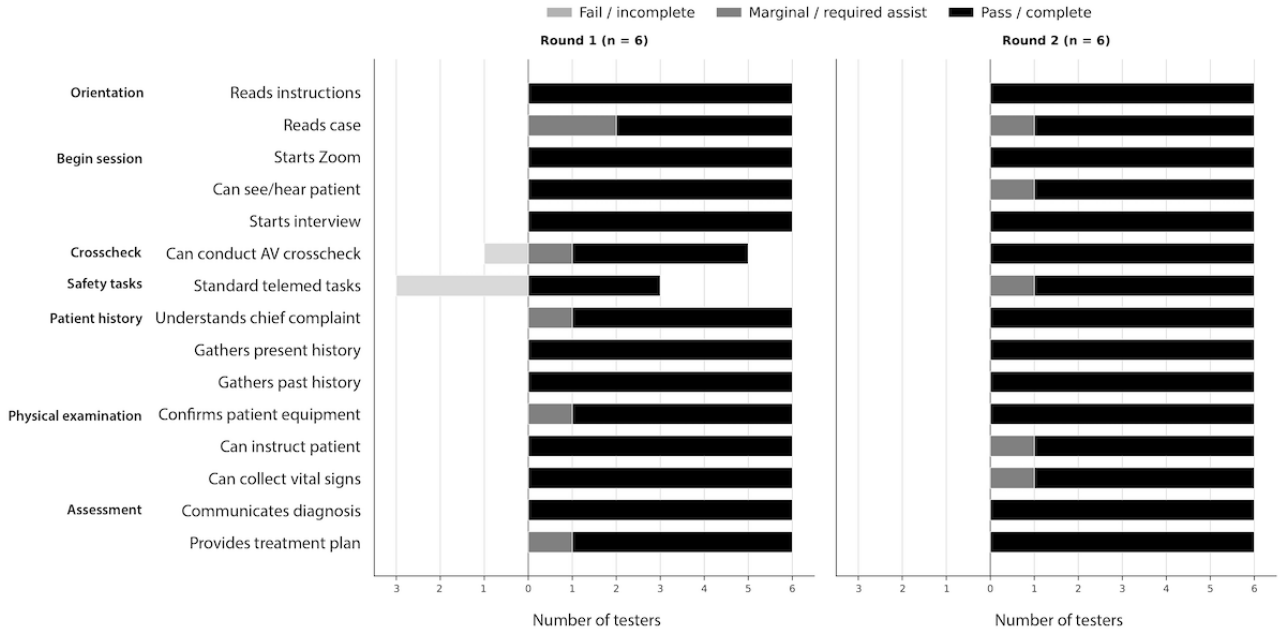
[c]Indicates the number of learners successfully assessed.

ATA Results

Simulation Step Completion Counts

Figure 6 shows simulation completion rates organized by workflow step. While most testers completed the steps without difficulty, not every tester completed every step. In the first round, testers skipped standard telemedicine patient safety tasks (eg, confirming the patient's physical location). This issue improved during the second round after we included a mock EHR with synthetic data and visual cues (eg, patient address and phone number fields).

Figure 6. Simulation step completion counts. We show scores for each testing round, separated by one refinement session. AV: audiovisual.



Usability Issues

We logged 52 usability findings across both testing rounds; 15 required corrections before piloting (Tables 3 and 4). We identified issues with the SP script, gaps in synthetic patient data, deviations in the simulation workflow, and problems with the competency rubric language. New issues emerged after the first refinement session due to updates to materials and assets (eg, the addition of a mock EHR).

XSL•FO
RenderX

Table 4. Excerpt from our usability issue log illustrating a sample of usability issues and remediation strategies organized according to the usability heuristics schema [80].

Test number	Finding	Code[a]	Valence	Priority	Fixed?	Intervention
7	The tester expected to see a flowsheet of blood pressure values in the mock electronic health record	Match	Negative	High	Yes	Added a flowsheet with additional blood pressure values to the LearningSpace software
8	Need to include photographs of physical findings, such as leg edema	Feedback	Negative	High	Yes	Gave the SP[b] printed photographs of physical examination findings sourced from licensed stock photography

[a]Code: usability heuristic code.

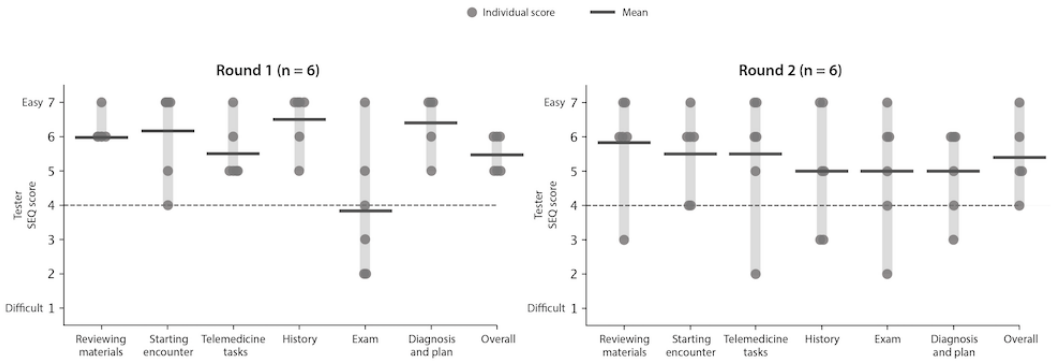
[b]SP: standardized patient.

Usability issues tended to fall into the following four heuristic categories: (1) user feedback, (2) match between the system and the real world, (3) error prevention, and (4) system visibility. For feedback, testers asked to see the patient's leg to check for edema. We provided SPs with photographs of ankle edema to show on camera. To align the system with the real world, testers requested a flowsheet of blood pressure values, which we added to our mock EHR. As an example of an error-prevention issue, faculty needed a clear deadline in the task definition to score learners scheduling follow-up laboratory tests. We added "within 1 week" to the rubric. As an example of a visibility issue, faculty could not tell when learners fabricated information to preserve the illusion of the simulation (eg, naming a fictional laboratory). We therefore added a question to the SP script that the learner would have to research (ie, the location of the laboratory nearest the patient's home).

SEQ Results

Most testers found the simulation instructions and steps easy to understand (Figure 7 [55]). Testers assigned a median SEQ score of 6 or greater to 4 of the 6 components. Testers said completing a virtual physical examination (median 4.50, IQR 2.75-6.00) was the most challenging step.

Figure 7. Single ease question results for both testing rounds. We show scores for each testing round, separated by one refinement session.

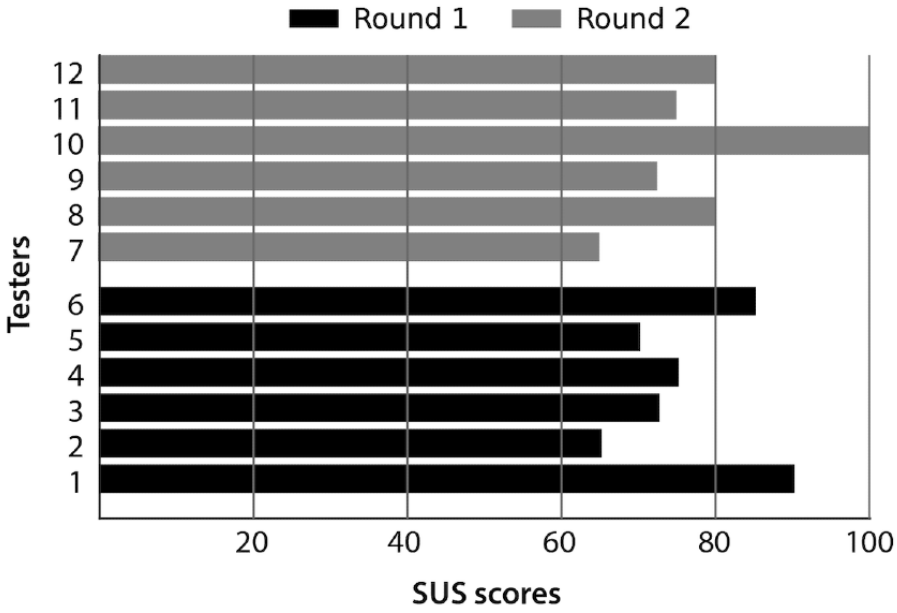


SUS Results

Testers reported some challenges with ease of use and learnability (Figure 8 [52]). SUS scores ranged from 65 to 100. For round 1, respondents' median score was 73.80 (IQR 70.63-82.50); for round 2, the median score was 77.50 (IQR 73.13-80.00). The inclusion of a mock EHR in round 2 may account for the increase in the median SUS score.

XSL•FO
RenderX

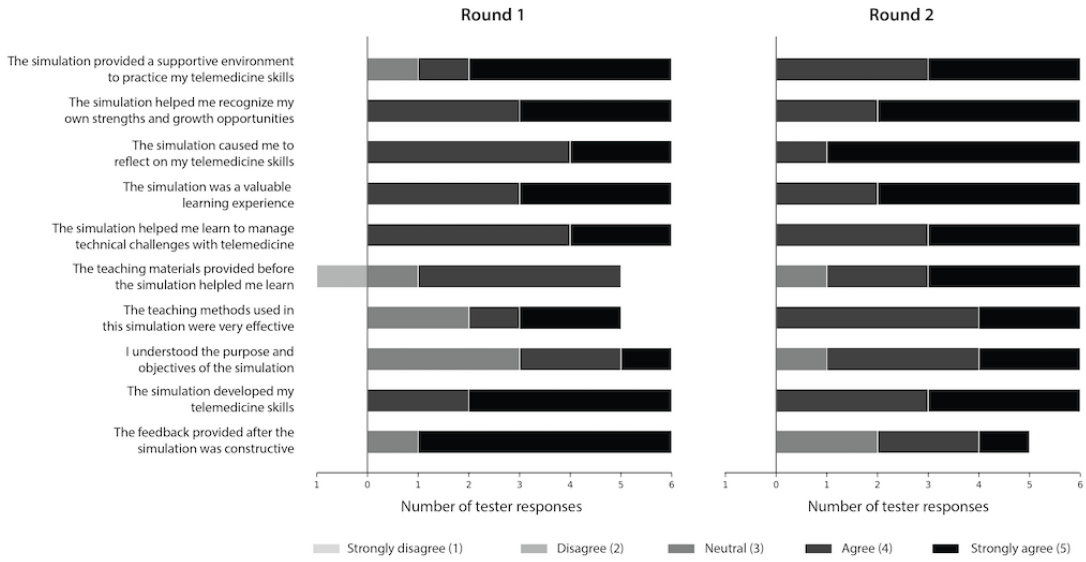
Figure 8. System Usability Scale (SUS) scores separated by testing rounds. We show scores for each testing round, separated by one refinement session.



SSES Results

The SSES scores indicate that most testers considered the simulation a valuable learning experience (Figure 9 [53]). One tester disagreed with the statement that the simulation tested their clinical abilities.

Figure 9. Satisfaction With Simulation Experience Scale scores. We show scores for each testing round, separated by one refinement session.

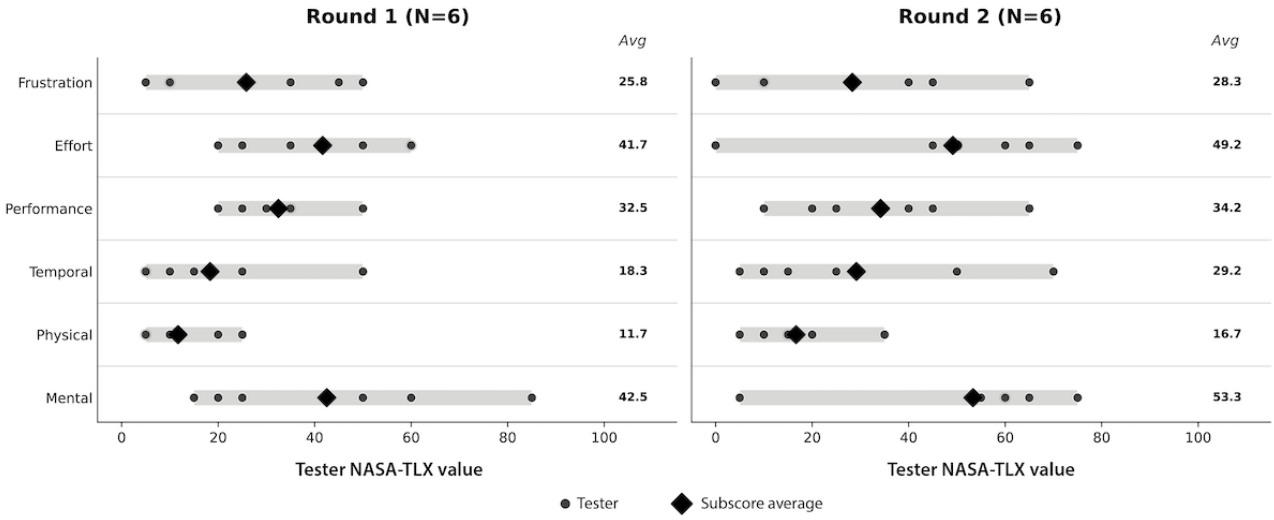


NASA-TLX Results

Testers reported considerable variation in perceived workload, depending on the subscale. Mental demand and effort were consistently the highest and increased from round 1 to round 2 (Figure 10 [69]). Mental workload was highest, with a median score of 37.50 (IQR 21.25-57.50) for round 1 and a median score of 60.00 (IQR 56.3-63.8) for round 2. The median scores for total effort (a composite of mental and physical effort) were 42.50 (IQR 27.50-57.50) and 55.00 (IQR 46.3-63.8) for rounds 1 and 2, respectively.

XSL•FO
RenderX

Figure 10. Unweighted NASA Task Load Index (NASA-TLX) scores organized by subscales. We show scores for each testing round, separated by one refinement session.

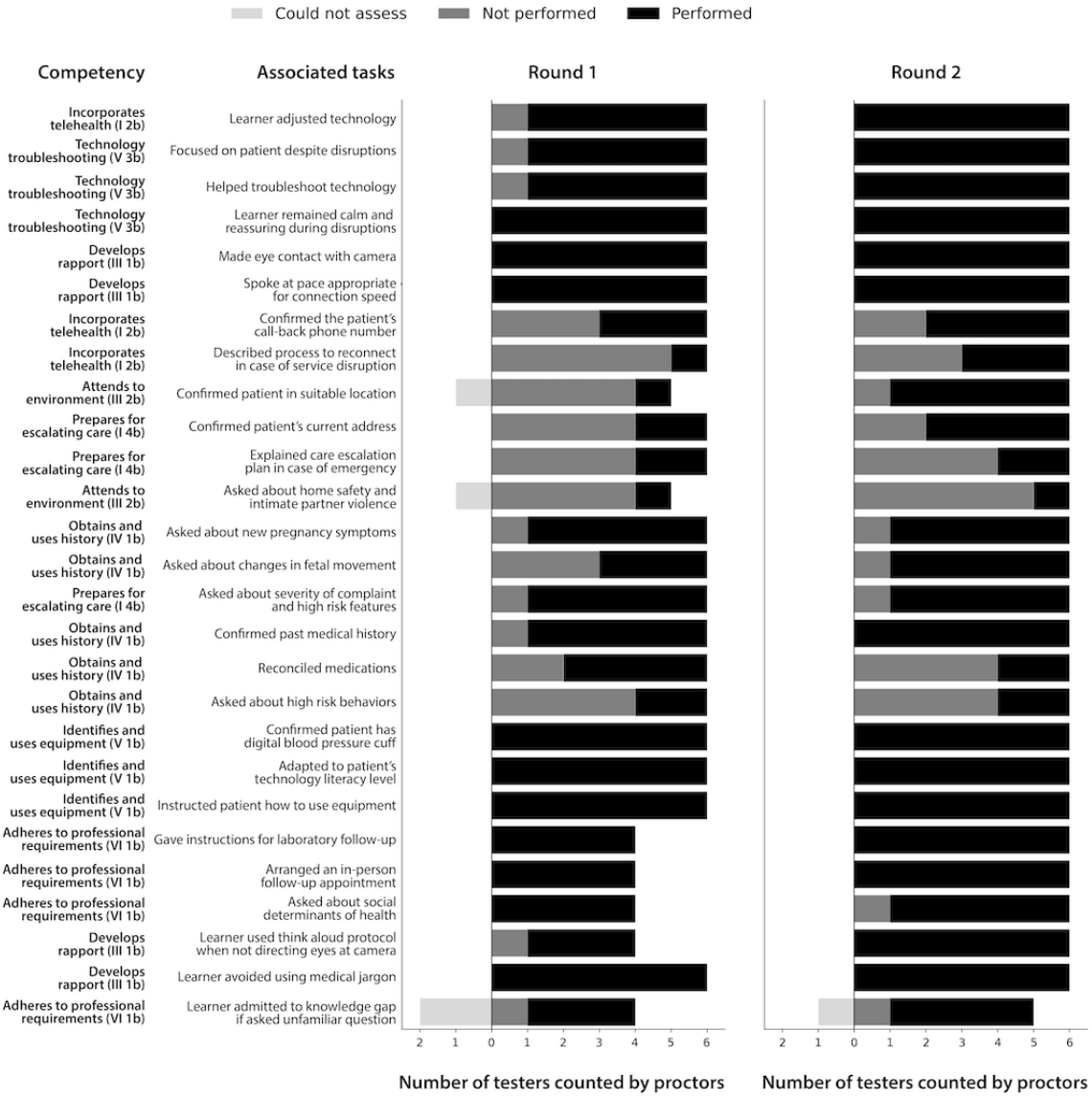


Competency Assessment

In the first round of testing, the proctor and observers could not score 3 tasks associated with 2 competencies (ie, "establishes therapeutic relationships and environments" and "supports solutions to ethical problems"). This problem was caused by ambiguities in the scoring guide's language (Figure 11 [36]). Revising the language improved usability. One persistent problem was identifying when a learner fabricated information (eg, improvising the name of a regional laboratory) to preserve the illusion of the simulation. We therefore modified the SP script before piloting to discourage learner improvisation.

XSL•FO
RenderX

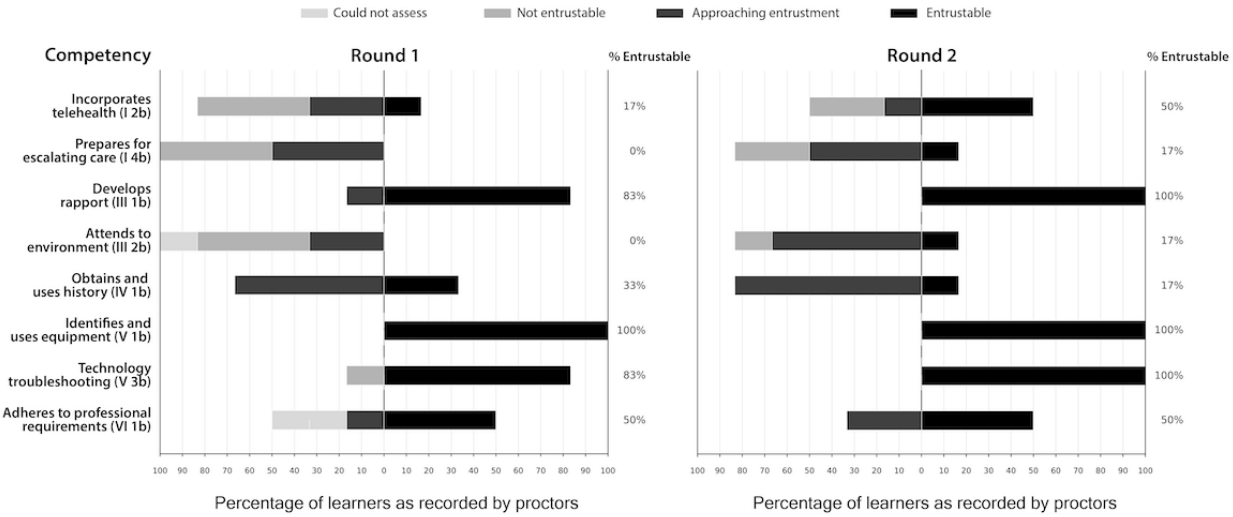
Figure 11. Individual task performance measured using a standardized rubric. Several tasks map to a single telehealth competency. We show task completion rates for each testing round, separated by one refinement session. Competencies were taken from the Association of American Medical Colleges (AAMC) white paper.



We summed proctor-scored task completion rates to calculate composite telemedicine competency scores (Figure 12 [36]). Testers tended to overlook telemedicine-specific tasks, such as gathering contact information, including the home address (AAMC competency I 2b); describing an emergency escalation plan (AAMC competency I 4b); or checking whether the patient was in a private setting and whether other household members were present (AAMC competency III 2b) [36].

XSL•FO
RenderX

Figure 12. Telemedicine competencies measured using our scoring rubric. Each competency is a composite score of several related tasks. We calculated competencies only for testers with complete data. We show scores for each testing round, separated by one refinement session. Competencies were taken from the Association of American Medical Colleges (AAMC) white paper.



Discussion

Overview

For this tutorial discussion, we first provide an abbreviated review of the most significant usability findings. We then offer a concept-level appraisal of our testing methods and insights for readers planning similar testing protocols. We conclude by comparing with the literature, highlighting gaps, unexplored research opportunities, and future directions.

Principal Findings

Our usability evaluation offered valuable insights into the effectiveness of our educational simulation and opportunities for quality improvement. The data gathered using the ATA enabled the development team to track and communicate improvements to the research team and simulation staff. We identified 52 usability issues requiring prioritization and made 15 design revisions before piloting with learners. By the second round of testing, we removed critical chokepoints related to the audiovisual cross-check and added a mock EHR with synthetic patient data and fill-in data fields. The mock EHR provided the learner with visual cues that improved the telemedicine workflow and task completion rates. Our qualitative data clarified the root causes of usability issues and provided an efficient method for prioritizing revisions. We traced our biggest challenges to the fidelity of blood pressure data and the reliability of the evaluation rubric. We subsequently focused revision efforts on clarifying details in patient scripts, adding synthetic patient data—including a blood pressure flowsheet to the mock EHR—and modifying scoring guide language and definitions in the telemedicine rubric.

It was crucial to analyze the simulation in accordance with the workflow steps, expected tasks, and learner competencies. By measuring simulation steps and task performance independently, we could see when testers completed all scripted steps but overlooked clinical tasks. This helped distinguish undesirable technical distractors from scripted challenges. By removing usability issues that diverted testers' attention from their tasks, testers could more effectively engage with the learning material. This also helped faculty more easily determine when simulation tasks were misaligned with learning objectives.

Testers successfully used the videoconferencing software, resolved technical issues, and established therapeutic rapport. By contrast, most testers neglected tasks necessary to optimize the telemedicine environment or prepare for a medical emergency. This is not unexpected; none of our testers received telemedicine training, whereas we gave a lecture to residents before the simulation.

For educators planning to include grading rubrics with their simulations, we recommend embedding scoring forms or debriefing guides into the testing instruments. In our example, we included the telemedicine competency rubric in the ATA. Our experience suggests a thorough appraisal of the evaluation materials by domain experts and participating faculty can improve construct validity by aligning simulation tasks and learning objectives with rubric terminology and instrument design. We also believe testing can improve instrument learnability and ease of use so that faculty can invest less time training proctors during piloting and implementation.

We used the SEQ, SUS, and NASA-TLX scores to compare our qualitative observations with usability benchmarks (ie, convergent validity). The range of SEQ scores (2 to 7) and SUS scores (65 to 100) helped identify the most challenging simulation components. The SEQ pinpointed issues in the synthetic data collected during the history and examination.

The NASA-TLX data helped estimate overall simulation difficulty. Responses indicated a relatively high score for mental workload compared with the other subscales (round 1: median score 37.50, IQR 21.25-57.50; round 2: median score 60.00, IQR 56.30-63.80). The scores may have increased due to the inclusion of a mock EHR. Considering these findings, we asked faculty whether the simulation should be modified. Obstetrics and family medicine faculty advised against changing the tasks

XSL•FO
RenderX

because they aligned with the educational goals and learning objectives. During piloting (illustrative vignette in Textbox 2; key tutorial insights in Textbox 3), residents were indeed able to complete the simulation but struggled with several tasks, including developing a preemptive care escalation plan and instructing SPs on the use of a digital blood pressure cuff. This suggests that our clinical scenario and materials provided sufficient scaffolding for learners to engage effectively with the material while gently challenging them with unfamiliar tasks.

Textbox 3. Key tutorial insights.

Key tutorial insights

- Agile Task Analysis combined with simulations using a "think aloud" protocol offers a practical strategy for identifying critical design issues.
- It is crucial to include instructional materials, scoring guides, and debriefing instruments in usability data collection instruments. This practice can help disambiguate simulation design issues from learner competency gaps.
- The System Usability Scale is a practical method for quickly benchmarking the overall perceived usability of an educational simulation.
- The addition of the Single Ease Question can help pinpoint the source of usability issues when global usability assessments raise concerns.
- The NASA (National Aeronautics and Space Administration) Task Load Index provides a "quick and dirty" method to estimate the difficulty of an educational simulation and guide future calibration.

Application of Findings to Simulation Development

This tutorial advances simulation science in several ways. First, we demonstrated how to leverage usability methods used in technology evaluations to systematically assess educational simulations before piloting with learners. Validated, theory-driven instruments can provide insights into simulation effectiveness, ease of use, and user satisfaction [22,82-84]. These methods are technology-agnostic and, with little modification, are suitable for evaluating simulation constructs such as engagement, cognitive load, skill acquisition, and competence.

While this study offers turnkey methods for formative usability testing, the choice of any single instrument should be driven by the learning objectives, the simulation's attributes, and the phase of the development life cycle. Not all methods are necessary for every simulation study. Researchers should add or remove components in a modular fashion. For example, user satisfaction with simulation data may be incidental to the educational program objectives; studies show self-assessment is an unreliable estimate of teaching effectiveness [85,86]. We included the SSES in our protocol because we believed fellow educators could provide useful appraisals of our simulation and its educational value.

Second, our approach addressed design challenges unique to educational simulations [22,23,74,87]. We concurrently evaluated simulation design features and learner competencies to align teaching modalities with educational objectives and to balance ecological validity with simulation complexity [4,11,12,20,88,89]. Our instruments were well suited to observing technology use and should be considered for technology-forward educational initiatives that include EHRs, mobile devices, artificial intelligence, and computerized decision support. Unlike traditional usability testing, our goal was not to remove opportunities for user error [17,21,73,80,90-92]. Educational simulations must enable learning through experimentation, discovery, and even failure [4,42,93,94]. We therefore used the ATA to record whether testers had the opportunity to demonstrate their skills during the simulation; the simulation could proceed as intended even if a tester failed a task.

Third, our methods can be used early in the product development life cycle with an array of testers [2,4,9]. While best-practice simulation guidelines recommend piloting simulations with learners, there are practical challenges [2]. Recruiting enough target learners can be impractical. Moreover, testing and piloting serve different but equally important roles. Because simulations can be challenging or stressful for learners, pilots may not be the ideal means of measuring usability dimensions such as learnability, error tolerance, and satisfaction [95-98]. Testing with experts can assess the fidelity of the case, the accuracy of the synthetic data, the appropriateness of the learning objectives, and the difficulty of tasks [41-43,99].

Fourth, we believe our methods are flexible, cost-effective, and scalable. We used low-cost tools that could be adapted to resource-constrained settings, such as smaller academic centers and schools in low- and middle-income countries [73,74,100]. Many of the instruments described herein, including the SUS, NASA-TLX, and SSES, have been validated in languages such as Chinese, Korean, Spanish, French, and Croatian [69,101-107]. These methods do not require large numbers of testers. We recommend multiple short testing rounds with as few as 5 testers; this provides sufficient data to guide refinements [108].

Comparison With Prior Work

With the expansion of advanced simulation modalities, such as AR, VR, and AI, software developers are increasingly recognizing the role of usability testing in the product development life cycle [109-114]. There are already protocols for a range of tools, including AR, VR, digital twins, and serious games [8,14,15]. While most reports focus on the technology, several have evaluated the broader simulation context. Anton and colleagues described an evaluation of 2 emergency medicine simulations that combined learner eye tracking and faculty ratings on the Non-Technical Skills for Surgeons Scale [14]. Silva et al [15] applied a more comprehensive approach by evaluating a simulation manikin for teaching cardiopulmonary resuscitation. They created a data collection instrument based on a basic life support guide and scored learners using a 16-item

XSL•FO
RenderX

Likert-type scale. The researchers recorded time on task and administered the SUS after each encounter. In this manner, they operationalized usability in terms of the dimensions of effectiveness, efficiency, and satisfaction. Together, these reports provide complementary perspectives on the evaluation of simulation-based instructional modules: Silva et al [15] focused on the physical usability of manikins, whereas Anton et al [14] focused on clinicians' cognitive skills.

Da Silva and Dubrowski [8] described one of the first attempts to incorporate usability methods early in the development life cycle. They conducted a series of design sprints to evaluate a conflict-management scenario between a nursing student and a senior nurse. The team tested 2 scenarios and recruited 20 nursing students. After each simulation, the team conducted exit interviews and administered the Simulation Design Scale (SDS)—a validated 20-item instrument measuring the quality and effectiveness of simulation-based learning experiences [115]. Unlike the SSES, which focuses on learners' satisfaction with the instructional approach, the SDS focuses on missing or underdeveloped simulation elements. Da Silva and Dubrowski [8] used the SDS as a quality metric and tracked improvement in scores across iterations. We believe this work is an important milestone toward developing a metrics-driven usability strategy.

Strengths, Limitations, and Next Steps to Improve Testing Protocols

We believe the application of usability methods to simulation-based education remains underexplored [41,42,99]. To our knowledge, our report and related preliminary work provide the most complete descriptions of mixed methods usability testing for educational simulations. Combining subjective measures (eg, user satisfaction and ease of use) with objective performance data (eg, task completion rates, error tolerance) can maximize insights into technical functionality and educational value [57,116,117].

These observations notwithstanding, several limitations may affect the generalizability of our work. We identified new usability issues after 2 refinement sessions, suggesting we had not reached saturation. Unfortunately, our team faced time and resource constraints; the number of iterations was, in part, dependent on curriculum deadlines and the number of available testers. To our knowledge, there is no universally accepted number of testers or testing rounds that guarantees researchers will find all usability issues. The usability literature, however, describes an issue-discovery curve with diminishing returns; as few as 5 testers may identify between 80% and 90% of critical issues [17]. Experts therefore recommend conducting several small testing cycles with 5 users at a time [17,118-121].

This report describes a single-site, single-SP, quality-improvement design. This raises several validity issues that usability specialists must consider. We conducted all usability testing with a single SP to control for simulation variability. However, our SP's familiarity with the case may have limited our ability to identify weak prompts, responses, or missing information. In the future, incorporating multiple SPs could reveal variations in script interpretation and behavior. We also observed variation in practice behaviors among subject-matter experts as a function of experience, specialty, and training. This suggests that institutions may inadvertently incorporate provincial practice patterns and local health system norms into their simulation designs. This could create standardization challenges, since usability studies tend to compare testers' performance against an established workflow or an approved set of product design requirements. In our example, we compared usability findings to local workflow and learning objectives. Researchers designing future studies may elect to use an external reference standard, such as a prerecorded best-practice encounter developed by a professional society.

We did not conduct interrater reliability tests for our rubric. In general, calculating agreement statistics falls outside the scope of usability testing during the product design phase. Instead, several members of the research team independently coded each task. We then compared scores and resolved discrepancies through discussion. Task descriptions with vague or ambiguous language were updated during the refinement phases. We believe the next step in our research program is to measure the interrater reliability of our rubric by asking faculty to independently rate learners who have completed the simulation and calculating an agreement statistic (eg, Cohen κ or Fleiss κ) [122].

We qualitatively coded all usability findings using the usability heuristics for health information technology by Zhang et al [80]. While useful for categorization, the codes did not indicate priority or provide insights for remediation. We believe there are opportunities to develop and validate usability heuristics specific to educational simulations. Ideally, codes informed by educational theory would provide design guidance to educators.

Finally, we see future opportunities to test additional usability measurement tools, including the SDS, heuristic inspections, and cognitive walkthroughs [17,22,115]. This could help enrich our logic models that connect usability tools to evaluation dimensions, such as realism, memorability, completeness, and predicted future learner performance. While videoconferencing may not be required for most simulations, other teams might use this capability to link distributed resources, such as testers, researchers, SPs, and software. To support these kinds of remote testing initiatives, researchers may distribute questionnaires digitally and trial wearable technologies, such as eye-tracking lenses and wearable activity trackers, including wristbands and smartwatches.

Conclusions

We believe this tutorial illustrates a practical, reproducible, and resource-sensitive method for usability testing educational simulations before piloting with target learners. While this report describes a single-site, single-SP, quality-improvement design, we strove to include published, validated instruments with claims of generalizability and relatively low-cost testing methods to permit replication or context-appropriate adaptation. We have included appendices with our instruments and encourage readers to adapt and apply these methods to a range of simulations. While we believe this monograph provides simulationists with a turnkey toolkit to conduct similar assessments, further studies using psychometrically validated instruments are needed to clarify the best educational use cases, evaluate additional constructs relevant to teaching and learning, and, when relevant,

XSL•FO
RenderX

incorporate additional tests to measure the interrater reliability of learner assessment instruments.

Acknowledgments

The views expressed in this publication are solely those of the authors and do not necessarily reflect the official policies of the US Department of Health and Human Services or the Health Resources and Services Administration, nor does mention of the department or agency names imply endorsement by the US Government.

We thank the Simulation Center of the University of Oklahoma-Tulsa (SCOUT) for providing simulation center space, standardized patients, and consultative support for this project. We thank Billi Coil and Guimy Castor for research coordination support. We thank Nancy Melton and Trisha Cook for administrative support.

The authors declare the use of generative AI (GenAI) at specific steps in the research and writing process. According to the GAIDeT taxonomy (2025), the following tasks were delegated to GenAI tools under full human supervision: code generation for Microsoft Excel formulas and spelling and grammar checking. The GenAI tools used included Grammarly, Microsoft Word 365 grammar and dictionary, and ChatGPT 5.5 (OpenAI). Responsibility for the final manuscript lies entirely with the authors. GenAI tools are not listed as authors and do not bear responsibility for the final manuscript. The authors did not use GenAI for text generation, ideation, protocol design, usability testing, data collection, final data analysis, or figure drafting.

Funding

Funding support for this project was provided in part by the University of Oklahoma School of Community Medicine (OUSCM) and the US Department of Health and Human Services, Health Resources and Services Administration (HRSA), as part of an approved federal and nonfederal award (grant number T34HP42150).

Data Availability

The datasets generated or analyzed during this study are available from the corresponding author on reasonable request.

Authors' Contributions

BJL, HM, JH, and SM developed the usability protocol. BJL, HM, and RM developed the theoretical framework and logic model. BJL, MR, KPG, and KB developed and revised the telemedicine simulation, including patient scripts, synthetic data, technology, and the evaluation rubric. BJL, KPG, MR, JVB, MA, and CP conducted all usability testing, data collection, and preliminary data analysis. BJL, HM, JH, AD, and RM conducted all final data analyses. BJL and MA generated all tables and figures. All authors contributed to the final manuscript and reviewed the submitted version.

Conflicts of Interest

BJL is the Editor-in-Chief of *JMIR Medical Education*. The authors declare no additional conflicts of interest.

Multimedia Appendix 1
Demographic questionnaire.
[DOCX File , 3806 KB - mededu_v12i1e102697_app1.docx]

Multimedia Appendix 2
Agile task analysis with embedded telemedicine competency rubric.
[DOCX File , 5427 KB - mededu_v12i1e102697_app2.docx]

Multimedia Appendix 3
List of Single Ease Questions (SEQ) for sets of related steps.
[DOCX File , 3805 KB - mededu_v12i1e102697_app3.docx]

Multimedia Appendix 4
Adapted post-simulation System Usability Scale (SUS) and Satisfaction with Simulation Experience Survey (SSES).
[DOCX File , 3806 KB - mededu_v12i1e102697_app4.docx]

Multimedia Appendix 5
Unweighted NASA Task Load Index (NASA-TLX).
[DOCX File , 3805 KB - mededu_v12i1e102697_app5.docx]

XSL•FO
RenderX

Multimedia Appendix 6
Semistructured simulation debrief guide.
[DOCX File , 3803 KB - mededu_v12i1e102697_app6.docx]

Multimedia Appendix 7
Structural codes data dictionary for usability issue coding.
[DOCX File , 4147 KB - mededu_v12i1e102697_app7.docx]

Multimedia Appendix 8
Usability issue log template with sample entries.
[DOCX File , 3804 KB - mededu_v12i1e102697_app8.docx]

Multimedia Appendix 9
Telehealth competency rubric used during resident piloting.
[DOCX File , 4496 KB - mededu_v12i1e102697_app9.docx]

Multimedia Appendix 10
Printed telemedicine memory aid.
[DOCX File , 5340 KB - mededu_v12i1e102697_app10.docx]

References

1. Elendu C, Amaechi DC, Okatta AU, Amaechi EC, Elendu TC, Ezeh CP, et al. The impact of simulation-based training in medical education: a review. Medicine (Baltimore) 2024;103(27):e38813 [FREE Full text] [doi: 10.1097/MD.0000000000038813] [Medline: 38968472]
2. Watts PI, McDermott DS, Alinier G, Charnetski M, Ludlow J, Horsley E, et al. Healthcare simulation standards of best practice™ simulation design. Clin Simul Nurs 2021;58:14-21. [doi: 10.1016/j.ecns.2021.08.009]
3. Aggarwal R, Mytton OT, Derbrew M, Hananel D, Heydenburg M, Issenberg B, et al. Training and simulation for patient safety. Qual Saf Health Care 2010;19(Suppl 2):i34-i43 [FREE Full text] [doi: 10.1136/qshc.2009.038562] [Medline: 20693215]
4. Alinier G. Developing high-fidelity health care simulation scenarios: a guide for educators and professionals. Simul Gaming 2010;42(1):9-26. [doi: 10.1177/1046878109355683]
5. Gliva-McConvey G, Nicholas CF, Nicholas L, editors. Comprehensive Healthcare Simulation: Implementing Best Practices in Standardized Patient Methodology. Cham, Switzerland: Springer; 2020.
6. Purva M, Nicklin J. ASPiH standards for simulation-based education: process of consultation, design and implementation. BMJ Simul Technol Enhanc Learn 2018;4(3):117-125 [FREE Full text] [doi: 10.1136/bmjstel-2017-000232] [Medline: 35520464]
7. Diaz-Navarro C, Laws-Chapman C, Moneypenny M, Purva M. The ASPiH Standards – 2023: guiding simulation-based practice in health and care. Int J Healthc Simul 2024. [doi: 10.54531/nyvm5886]
8. Da Silva C, Dubrowski A. Piloting simulations: a systematic refinement strategy. Cureus 2019;11(12):e6434 [FREE Full text] [doi: 10.7759/cureus.6434] [Medline: 31993271]
9. Lewis KL, Bohnert CA, Gammon WL, Hölzer H, Lyman L, Smith C, et al. The Association of Standardized Patient Educators (ASPE) Standards of Best Practice (SOBP). Adv Simul (Lond) 2017;2:10 [FREE Full text] [doi: 10.1186/s41077-017-0043-4] [Medline: 29450011]
10. Fanning RM, Gaba DM. The role of debriefing in simulation-based learning. Simul Healthc 2007;2(2):115-125. [doi: 10.1097/SIH.0b013e3180315539] [Medline: 19088616]
11. Marcilly R, Monkman H, Pelayo S, Lesselroth B. Usability evaluation ecological validity: is more always better? Healthcare (Basel) 2024;12(14):1417 [FREE Full text] [doi: 10.3390/healthcare12141417] [Medline: 39057560]
12. Dieckmann P, Gaba D, Rall M. Deepening the theoretical foundations of patient simulation as social practice. Simul Healthc 2007;2(3):183-193. [doi: 10.1097/SIH.0b013e3180f637f5] [Medline: 19088622]
13. Shin J, Lee H. Optimal usability test procedure generation for medical devices. Healthcare (Basel) 2023;11(3):296 [FREE Full text] [doi: 10.3390/healthcare11030296] [Medline: 36766871]
14. Anton NE, Cha JS, Hernandez E, Athanasiadis DI, Yang J, Zhou G, et al. Utilizing eye tracking to assess medical student non-technical performance during scenario-based simulation: results of a pilot study. Global Surg Educ 2023;2(1):49. [doi: 10.1007/s44186-023-00127-3] [Medline: 38414559]
15. Silva LDB, Gonçalves FT, Arantes HP, Garcia RA, Gasparin MRFR, dos Reis MM, et al. Usability evaluation of high-fidelity simulation manikin for cardiopulmonary resuscitation training for medical students. Res Biomed Eng) 2024;40(1):253-264. [doi: 10.1007/s42600-024-00340-z]

XSL•FO
RenderX

16. Ergonomics of Human-System Interaction—Part 210: Human-Centred Design for Interactive Systems. Geneva, Switzerland: International Organization for Standardization; 2019:33.
17. Barnum CM. Usability Testing Essentials: Ready, Set...Test!. Burlington, Massachusetts: Morgan Kaufmann; 2010.
18. Yates N, Gough S, Brazil V. Self-assessment: with all its limitations, why are we still measuring and teaching it? Lessons from a scoping review. Med Teach 2022;44(11):1296-1302. [doi: 10.1080/0142159X.2022.2093704] [Medline: 35786121]
19. Shackel B. The concept of usability. In: Bennett JL, Case D, Sandelin J, Smith M, editors. Visual Display Terminals: Usability Issues and Health Concerns. Englewood Cliffs, New Jersey: Prentice-Hall; 1984:4823.
20. Monkman H, Marcilly R, Lesselroth B. An integrated model of external validity usability evaluations in health care. Stud Health Technol Inform 2025;326:80-84 [FREE Full text] [doi: 10.3233/SHTI250242] [Medline: 40357606]
21. Nielsen J. Usability Engineering. 1st Edition. San Francisco, California: Morgan Kaufmann; 1994.
22. Harrington L, Harrington C. Usability Evaluation Handbook for Electronic Health Records. Chicago, Illinois: HIMSS; 2014.
23. Albert W, Tullis T. Measuring the User Experience: Collecting, Analyzing, and Presenting Usability Metrics. Waltham, Massachusetts: Morgan Kaufmann; 2013.
24. Bashshur R, Doarn CR, Frenk JM, Kvedar JC, Woolliscroft JO. Telemedicine and the COVID-19 pandemic, lessons for the future. Telemed J E Health 2020;26(5):571-573. [doi: 10.1089/tmj.2020.29040.rb] [Medline: 32275485]
25. Basu A, Kuziemsky C, de Araújo Novaes M, Kleber A, Sales F, Al-Shorbaji N, et al. Telehealth and the COVID-19 pandemic: international perspectives and a health systems framework for telehealth implementation to support critical response. Yearb Med Inform 2021;30(1):126-133 [FREE Full text] [doi: 10.1055/s-0041-1726484] [Medline: 33882598]
26. Clare CA. Telehealth and the digital divide as a social determinant of health during the COVID-19 pandemic. Netw Model Anal Health Inform Bioinform 2021;10(1):26 [FREE Full text] [doi: 10.1007/s13721-021-00300-y] [Medline: 33842187]
27. Eslami Jahromi M, Ayatollahi H. Utilization of telehealth to manage the COVID-19 pandemic in low- and middle-income countries: a scoping review. J Am Med Inform Assoc 2023;30(4):738-751 [FREE Full text] [doi: 10.1093/jamia/ocac250] [Medline: 36565464]
28. Thies KM, Gonzalez M, Porto A, Ashley KL, Korman S, Lamb M. Project ECHO COVID-19: vulnerable populations and telehealth early in the pandemic. J Prim Care Community Health 2021;12:21501327211019286 [FREE Full text] [doi: 10.1177/21501327211019286] [Medline: 34036832]
29. Wijesooriya NR, Mishra V, Brand PLP, Rubin BK. COVID-19 and telehealth, education, and research adaptations. Paediatr Respir Rev 2020;35:38-42 [FREE Full text] [doi: 10.1016/j.prrv.2020.06.009] [Medline: 32653468]
30. Alcocer Alkureishi M, Lenti G, Choo ZY, Castaneda J, Weyer G, Oyler J, et al. Teaching telemedicine: the next frontier for medical educators. JMIR Med Educ 2021;7(2):e29099 [FREE Full text] [doi: 10.2196/29099] [Medline: 33878011]
31. Belakovskiy A, Jones EK. Telehealth and medical education. Prim Care 2022;49(4):575-583. [doi: 10.1016/j.pop.2022.04.003] [Medline: 36357062]
32. Chike-Harris KE, Durham C, Logan A, Smith G, DuBose-Morris R. Integration of telehealth education into the health care provider curriculum: a review. Telemed J E Health 2021;27(2):137-149. [doi: 10.1089/tmj.2019.0261] [Medline: 32250196]
33. Khullar D, Mullangi S, Yu J, Weems K, Shipman SA, Caulfield M, et al. The state of telehealth education at U.S. medical schools. Healthc (Amst) 2021;9(2):100522. [doi: 10.1016/j.hjdsi.2021.100522] [Medline: 33548854]
34. Noronha C, Lo MC, Nikiforova T, Jones D, Nandiwada DR, Leung TI, Society of General Internal Medicine (SGIM) Education Committee. Telehealth competencies in medical education: new frontiers in faculty development and learner assessments. J Gen Intern Med 2022;37(12):3168-3173 [FREE Full text] [doi: 10.1007/s11606-022-07564-8] [Medline: 35474505]
35. Waseh S, Dicker AP. Telemedicine training in undergraduate medical education: mixed-methods review. JMIR Med Educ 2019;5(1):e12515 [FREE Full text] [doi: 10.2196/12515] [Medline: 30958269]
36. Telehealth competencies. Association of American Medical Colleges. URL: https://www.aamc.org/data-reports/report/telehealth-competencies [accessed 2026-08-14]
37. Cantone RE, Palmer R, Dodson LG, Biagioli FE. Insomnia telemedicine OSCE (TeleOSCE): a simulated standardized patient video-visit case for clerkship students. MedEdPORTAL 2019;15:10867 [FREE Full text] [doi: 10.15766/mep_2374-8265.10867] [Medline: 32051850]
38. Papanagnou D, Stone D, Chandra S, Watts P, Chang A, Hollander J. Integrating telehealth emergency department follow-up visits into residency training. Cureus 2018;10(4):e2433 [FREE Full text] [doi: 10.7759/cureus.2433] [Medline: 29876155]
39. Mulcare M, Naik N, Greenwald P, Schullstrom K, Gogia K, Clark S, et al. Advanced communication and examination skills in telemedicine: a structured simulation-based course for medical students. MedEdPORTAL 2020;16:11047 [FREE Full text] [doi: 10.15766/mep_2374-8265.11047] [Medline: 33365390]
40. Charron E, Castor G, Veach C, Chubb R, Lesselroth B, Sachs VESS, et al. Enhanced obstetric training to address maternity care workforce shortages in tribal, rural, and underserved communities: a case from Oklahoma. Healthc (Amst) 2025;13(2):100768. [doi: 10.1016/j.hjdsi.2025.100768] [Medline: 40926429]
41. Lesselroth B, Monkman H, Anderson M, Parsons C, Mohanty S, Lawson A, et al. Usability testing an educational telemedicine simulation: description of methods and preliminary findings. Stud Health Technol Inform 2025;327:273-277. [doi: 10.3233/SHTI250327] [Medline: 40380438]

XSL•FO
RenderX

42. Lesselroth B, Monkman H, Marcilly R, Mohanty S, Anderson M, Parsons C, et al. Hybrid usability analysis to improve an educational telemedicine simulation: task analysis and survey results. Stud Health Technol Inform 2025;326:59-63 [FREE Full text] [doi: 10.3233/SHTI250237] [Medline: 40357602]
43. Monkman H, Homco J, Bellavia K, Anderson M, Tanriverdi A, Lesselroth C. Iterative design of an educational telemedicine simulation: addressing usability issues to enhance the simulation. 2026 Presented at: Studies in Health Technology and Informatics; Opening the Personal Gate Between Technology and Health Care: Proceedings of MIE 2026; May 25-28, 2026; Genoa, Italy p. 2130-2134. [doi: 10.3233/SHTI260637]
44. Sonenberg A, Mason D. Maternity care deserts in the US. JAMA Health Forum 2023;4(1):e225541 [FREE Full text] [doi: 10.1001/jamahealthforum.2022.5541] [Medline: 36633853]
45. Fontenot J, Brigance C, Lucas R, Stoneburner A. Navigating geographical disparities: access to obstetric hospitals in maternity care deserts and across the United States. BMC Pregnancy Childbirth 2024;24(1):350. [doi: 10.1186/s12884-024-06535-7] [Medline: 38720255]
46. Atwani R, Robbins L, Saade G, Kawakita T. Association of maternity care deserts with maternal and pregnancy-related mortality. Obstet Gynecol 2025;146(2):181-188. [doi: 10.1097/AOG.0000000000005976] [Medline: 40570350]
47. Gleeson DE, Busch SH, Ickovics JR. State-level prevalence of maternity care deserts: association with healthcare access, utilization, and outcomes among medicaid recipients. AJPM Focus 2025;4(5):100362 [FREE Full text] [doi: 10.1016/j.focus.2025.100362] [Medline: 40735223]
48. Tanne JH. Nearly six million women in the US live in maternity care deserts. BMJ 2023;382:1878. [doi: 10.1136/bmj.p1878] [Medline: 37580083]
49. Adashi EY, O'Mahony DP, Cohen IG. Maternity care deserts: key drivers of the national maternal health crisis. J Am Board Fam Med 2025;38(1):165-167. [doi: 10.3122/jabfm.2024.240198R1] [Medline: 40268316]
50. Quesenbery W. The five dimensions of usability. In: Content and Complexity: Information Design in Technical Communication. New York: Routledge; 2014:93-114.
51. Davis FD. A Technology Acceptance Model for Empirically Testing New End-User Information Systems: Theory and Results. Cambridge, Massachusetts: Massachusetts Institute of Technology, Sloan School of Management; 1985.
52. Sauro J. A Practical Guide to the System Usability Scale: Background, Benchmarks & Best Practices. Denver, Colorado: Measuring Usability LLC; 2011.
53. Levett-Jones T, McCoy M, Lapkin S, Noble D, Hoffman K, Dempsey J, et al. The development and psychometric testing of the Satisfaction with Simulation Experience Scale. Nurse Educ Today 2011;31(7):705-710. [doi: 10.1016/j.nedt.2011.01.004] [Medline: 21288606]
54. Hart SG. NASA Task Load Index (TLX): Paper and Pencil Package, Version 1.0. Moffett Field, California: NASA Ames Research Center; 1986.
55. Sonderegger A, Sauer J. The influence of laboratory set-up in usability tests: effects on user performance, subjective ratings and physiological measures. Ergonomics 2009;52(11):1350-1361. [doi: 10.1080/00140130903067797] [Medline: 19851903]
56. Crotty BH, Hyun N, Polovneff A, Dong Y, Decker MC, Mortensen N, et al. Analysis of clinician and patient factors and completion of telemedicine appointments using video. JAMA Netw Open 2021;4(11):e2132917 [FREE Full text] [doi: 10.1001/jamanetworkopen.2021.32917] [Medline: 34735013]
57. Lesselroth BJ, Mastarone GL, Adams K, Tallett S, Kushniruk A, Borycki EM. Hybrid usability methods: practical techniques for evaluating health information technology in an operational setting. In: Patient-Centered Digital Healthcare Technology: Novel Applications for Next Generation Healthcare Systems. Stevenage, England: Institution of Engineering and Technology; 2020.
58. Kirwan B, Ainsworth L, Pendlebury G. Task analysis: the state of the art in process control. In: Contemporary Ergonomics. London, England: Taylor & Francis; 2020:479-484.
59. Stanton NA. Systematic human error reduction and prediction approach (SHERPA). In: Stanton NA, Hedge A, Brookhuis KA, Salas E, Hendrick HW, editors. Handbook of Human Factors and Ergonomics Methods. London, England: CRC Press; 2004:394-403.
60. ten Cate O. An updated primer on entrustable professional activities (EPAs). Rev Bras Educ Med 2019;43(1 suppl 1):712-720. [doi: 10.1590/1981-5271v43suplemento1-20190238.ING]
61. Cate OT. A primer on entrustable professional activities. Korean J Med Educ 2018;30(1):1-10 [FREE Full text] [doi: 10.3946/kjme.2018.76] [Medline: 29510603]
62. Sauro J, Dumas J. Comparison of three one-question, post-task usability questionnaires. 2009 Presented at: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI 2009); April 4-9, 2009; Boston, Massachusetts p. 1599-1608.
63. Sauro J, Lewis J. How the SEQ correlates with other task metrics. MeasuringU. 2026. URL: https://measuringu.com/how-the-seq-correlates-with-other-task-metrics/ [accessed 2026-02-05]
64. Sauro J. 10 things to know about the Single Ease Question (SEQ). MeasuringU. 2012. URL: https://measuringu.com/seq10/ [accessed 2026-02-04]
65. Russ-Jara AL, Saleem JJ, Herout J. A practical guide to usability questionnaires that evaluate clinicians' perceptions of health information technology. J Biomed Inform 2025;165:104822. [doi: 10.1016/j.jbi.2025.104822] [Medline: 40174758]

XSL•FO
RenderX

66. Brooke J. SUS: a quick and dirty usability scale. In: Jordan PW, Thomas B, Weerdmeester BA, McClelland IL, editors. Usability Evaluation in Industry. London, England: Taylor & Francis; 1996:4-7.
67. Lewis JR, Sauro J. The factor structure of the System Usability Scale. 2009 Presented at: First International Conference on Human Centered Design (HCD 2009); July 19-24, 2009; San Diego, California p. 94-103.
68. Williams B, Dousek S. The satisfaction with simulation experience scale (SSES): a validation study. J Nurs Educ Pract 2012;2(3):74. [doi: 10.5430/jnep.v2n3p74]
69. Hart SG. NASA-task load index (NASA-TLX): 20 years later. 2006 Presented at: Proceedings of the Human Factors and Ergonomics Society Annual Meeting; October 16-20, 2006; San Francisco, California. [doi: 10.1177/154193120605000909]
70. Young MS, Brookhuis KA, Wickens CD, Hancock PA. State of science: mental workload in ergonomics. Ergonomics 2015;58(1):1-17. [doi: 10.1080/00140139.2014.956151] [Medline: 25442818]
71. Grier RA. How high is high? A meta-analysis of NASA-TLX global workload scores. 2015 Presented at: Proceedings of the human factors and ergonomics society annual meeting; October 26-30, 2015; Los Angeles, California. [doi: 10.1177/1541931215591373]
72. Russ AL, Baker DA, Fahner WJ, Milligan BS, Cox L, Hagg HK, et al. A rapid usability evaluation (RUE) method for health information technology. AMIA Annu Symp Proc 2010;2010:702-706 [FREE Full text] [Medline: 21347069]
73. Krug S. Rocket Surgery Made Easy: The Do-It-Yourself Guide to Finding and Fixing Usability Problems. Berkeley, California: New Riders; 2009.
74. Kushniruk AW, Patel VL. Cognitive and usability engineering methods for the evaluation of clinical information systems. J Biomed Inform 2004;37(1):56-76 [FREE Full text] [doi: 10.1016/j.jbi.2004.01.003] [Medline: 15016386]
75. Kushniruk AW. Analysis of complex decision-making processes in health care: cognitive approaches to health informatics. J Biomed Inform 2001;34(5):365-376 [FREE Full text] [doi: 10.1006/jbin.2001.1021] [Medline: 12123153]
76. Zoom. Zoom Video Communications, Inc. 2013. URL: https://zoom.us/ [accessed 2026-07-29]
77. Parsons C, Monkman H, Jones B, Wolfinbarger A, Homco J, Anderson M. Co-creating a mnemonic with learners to support telehealth competency development during simulations. 2025 Presented at: Context Sensitive Health Informatics: AI for Social Good: Proceedings of CSHI 2025; Studies in Health Technology and Informatics, volume 326; May 23-24, 2025; Bradford, United Kingdom p. 137-141. [doi: 10.3233/shti250255]
78. LearningSpace. CAE Healthcare. URL: https://www.caehealthcare.com/learningspace/ [accessed 2026-08-14]
79. Saldana J. Fundamentals of Qualitative Research. New York: Oxford University Press; 2011.
80. Zhang J, Johnson TR, Patel VL, Paige DL, Kubose T. Using usability heuristics to evaluate patient safety of medical devices. J Biomed Inform 2003;36(1-2):23-30 [FREE Full text] [doi: 10.1016/s1532-0464(03)00060-1] [Medline: 14552844]
81. Shneiderman B. Designing the User Interface: Strategies for Effective Human-Computer Interaction. Bengaluru, India: Pearson Education; 2010.
82. Zhang J, Walji MF. TURF: toward a unified framework of EHR usability. J Biomed Inform 2011;44(6):1056-1067 [FREE Full text] [doi: 10.1016/j.jbi.2011.08.005] [Medline: 21867774]
83. Kushniruk A, Nohr C, Jensen S, Borycki EM. From usability testing to clinical simulations: bringing context into the design and evaluation of usable and safe health information technologies. Yearb Med Inform 2018;22(01):78-85. [doi: 10.1055/s-0038-1638836]
84. Lesselroth BJ, Monkman H, Adams K, Wood S, Corbett A, Homco J, et al. User experience theories, models, and frameworks: a focused review of the healthcare literature. Stud Health Technol Inform 2020;270:1076-1080. [doi: 10.3233/SHTI200327] [Medline: 32570547]
85. León SP, Panadero E, García-Martínez I. How accurate are our students? A meta-analytic systematic review on self-assessment scoring accuracy. Educ Psychol Rev 2023;35(4):106. [doi: 10.1007/s10648-023-09819-0]
86. Zheng B, He Q, Lei J. Informing factors and outcomes of self-assessment practices in medical education: a systematic review. Ann Med 2024;56(1):2421441 [FREE Full text] [doi: 10.1080/07853890.2024.2421441] [Medline: 39460558]
87. Sauro J, Lewis JR. Quantifying the User Experience: Practical Statistics for User Research. Cambridge, Massachusetts: Morgan Kaufmann; 2016.
88. Issenberg SB, McGaghie WC, Petrusa ER, Lee Gordon D, Scalese RJ. Features and uses of high-fidelity medical simulations that lead to effective learning: a BEME systematic review. Med Teach 2005;27(1):10-28. [doi: 10.1080/01421590500046924] [Medline: 16147767]
89. Issenberg SB, Scalese RJ. Best evidence on high-fidelity simulation: what clinical teachers need to know. Clin Teach 2007;4(2):73-77. [doi: 10.1111/j.1743-498X.2007.00161.x]
90. Carayon P. Handbook of Human Factors and Ergonomics in Health Care and Patient Safety. Mahwah, New Jersey: CRC Press; 2006.
91. Carayon P, Wetterneck TB, Rivera-Rodriguez AJ, Hundt AS, Hoonakker P, Holden R, et al. Human factors systems approach to healthcare quality and patient safety. Appl Ergon 2014;45(1):14-25 [FREE Full text] [doi: 10.1016/j.apergo.2013.04.023] [Medline: 23845724]
92. Kohn LT, Corrigan JM, Donladson MS. In: Kohn LT, Corrigan JM, Donaldson MS, editors. To Err is Human: Building a Safer Health System. Washington, District of Columbia: National Academies Press; 2000.

XSL•FO
RenderX

93. Monkman H, Kuziemsky C, Homco J, Liew A, Rodriguez K, Skaggs J, et al. Identifying failure modes in telemedicine: an instructional needs assessment. Stud Health Technol Inform 2023;304:39-43. [doi: 10.3233/SHTI230365] [Medline: 37347566]
94. Monkman H, Palmer R, Ijams S, Kollaja L, Rodriguez KA, Liew A, et al. Using simulations to train medical students for unanticipated technology failures in telemedicine. Stud Health Technol Inform 2022;294:775-779. [doi: 10.3233/SHTI220582] [Medline: 35612202]
95. Suh Y, Shim H, Lee HJ, Lee JH, Park MS, Choi S. Effects of subjective stress levels on learning effectiveness in high-fidelity simulation among healthcare workers. Acta Psychol (Amst) 2025;257:105075 [FREE Full text] [doi: 10.1016/j.actpsy.2025.105075] [Medline: 40472585]
96. Tremblay M, Lafleur A, Leppink J, Dolmans DHJM. The simulated clinical environment: cognitive and emotional impact among undergraduates. Med Teach 2017;39(2):181-187. [doi: 10.1080/0142159X.2016.1246710] [Medline: 27832706]
97. Tremblay ML, Rethans JJ, Dolmans D. Task complexity and cognitive load in simulation-based education: a randomised trial. Med Educ 2023;57(2):161-169. [doi: 10.1111/medu.14941] [Medline: 36151727]
98. Al-Ghareeb AZ, Cooper SJ, McKenna LG. Anxiety and clinical performance in simulated setting in undergraduate health professionals education: an integrative review. Clin Simul Nurs 2017;13(10):478-491. [doi: 10.1016/j.ecns.2017.05.015]
99. Lesselroth B, Monkman H, Bellavia K, Anderson M, van Buren J, Snider K, et al. Methods for usability testing an educational simulation: a case-study using a maternal telehealth scenario. Stud Health Technol Inform 2026;336:2125-2129. [doi: 10.3233/SHTI260636] [Medline: 42175304]
100. Nielsen J. Discount usability: 20 years. Nielsen Norman Group. 2009. URL: https://www.nngroup.com/articles/discount-usability-20-years/ [accessed 2009-02-06]
101. Sevilla-Gonzalez MDR, Moreno Loaeza L, Lazaro-Carrera LS, Bourguet Ramirez B, Vázquez Rodríguez A, Peralta-Pedrero ML, et al. Spanish version of the System Usability Scale for the assessment of electronic tools: development and validation. JMIR Hum Factors 2020;7(4):e21161 [FREE Full text] [doi: 10.2196/21161] [Medline: 33325828]
102. Wang Y, Lei T, Liu X. Chinese system usability scale: Translation, revision, psychological measurement. International Journal of Human–Computer Interaction 2019;36(10):953-963. [doi: 10.1080/10447318.2019.1700644]
103. Xiao YM, Wang ZM, Wang MZ, Lan YJ. The appraisal of reliability and validity of subjective workload assessment technique and NASA-task load index. Zhonghua Lao Dong Wei Sheng Zhi Ye Bing Za Zhi 2005;23(3):178-181. [Medline: 16124892]
104. Rubio S, Díaz E, Martín J, Puente JM. Evaluation of subjective mental workload: a comparison of SWAT, NASA-TLX, and workload profile methods. Appl Psychol 2004;53(1):61-86. [doi: 10.1111/j.1464-0597.2004.00161.x]
105. Kwon H, Yoou S. Validation of a Korean version of the satisfaction with simulation experience scale for paramedic students. Korean J Emerg Med Serv 2014;18(2):7-20 [FREE Full text] [doi: 10.14408/kjems.2014.18.2.007]
106. Smrekar M, Ledinski Fičko S, Kurtović B, Ilić B, Čukljek S, Tomac M, et al. Translation and validation of the Satisfaction with Simulation Experience scale: cross-sectional study. Cent Eur J Nurs Midw 2022;13(2):633-639. [doi: 10.15452/cejnm.2021.12.0038]
107. Clemmensen T. Templates for cross-cultural and culturally specific usability testing: results from field studies and ethnographic interviewing in three countries. Int J Hum Comput Interact 2011;27(7):634-669. [doi: 10.1080/10447318.2011.555303]
108. Lewis JR, Sauro J. Ten things to know about the RITE method. MeasuringU. URL: https://measuringu.com/rite-method/ [accessed 2026-02-06]
109. Brunzini A, Papetti A, Messi D, Germani M. A comprehensive method to design and assess mixed reality simulations. Virtual Real 2022;26(4):1257-1275. [doi: 10.1007/s10055-022-00632-8]
110. Birdsall S, McAlpin E, Grandhi U. Evaluating the usability and learning effectiveness of a virtual reality patient simulation for training nursing students: A pilot study. Teach Learn Nurs 2025;20(4):e1021-e1028. [doi: 10.1016/j.teln.2025.04.017]
111. Birrenbach T, Stuber R, Müller CE, Sutter P, Hautz WE, Exadaktylos AK, et al. Virtual reality simulation to enhance advanced trauma life support trainings: a randomized controlled trial. BMC Med Educ 2024;24(1):666 [FREE Full text] [doi: 10.1186/s12909-024-05645-2] [Medline: 38886688]
112. Mørk G, Bonsaksen T, Larsen OS, Kunnikoff HM, Lie SS. Virtual reality simulation in undergraduate health care education programs: usability study. JMIR Med Educ 2024;10:e56844 [FREE Full text] [doi: 10.2196/56844] [Medline: 39560982]
113. Rovati L, Gary PJ, Cubro E, Dong Y, Kilickaya O, Schulte PJ, et al. Development and usability testing of a patient digital twin for critical care education: a mixed methods study. Front Med (Lausanne) 2023;10:1336897 [FREE Full text] [doi: 10.3389/fmed.2023.1336897] [Medline: 38274456]
114. Lee Y, Kim SK, Eom M. Usability of mental illness simulation involving scenarios with patients with schizophrenia via immersive virtual reality: a mixed methods study. PLoS One 2020;15(9):e0238437 [FREE Full text] [doi: 10.1371/journal.pone.0238437] [Medline: 32936813]
115. Jeffries PR, Rizzolo MA. Designing and implementing models for the innovative use of simulation to teach nursing care of ill adults and children: a national, multi-site, multi-method study. NLN/Laerdal Project Summary Report. 2006. URL: https://www.nln.org/docs/default-source/uploadedfiles/professional-development-programs/read-the-nln-laerdal-project-summary-report-pdf.pdf [accessed 2026-08-14]

XSL•FO
RenderX

116. Keenan HL, Duke SL, Wharrad HJ, Doody GA, Patel RS. Usability: an introduction to and literature review of usability testing for educational resources in radiation oncology. Tech Innov Patient Support Radiat Oncol 2022;24:67-72 [FREE Full text] [doi: 10.1016/j.tipsro.2022.09.001] [Medline: 36262375]
117. Johnson SG, Potrebny T, Larun L, Ciliska D, Olsen NR. Usability methods and attributes reported in usability studies of mobile apps for health care education: scoping review. JMIR Med Educ 2022;8(2):e38259 [FREE Full text] [doi: 10.2196/38259] [Medline: 35767323]
118. Nielsen J. Why you only need to test with 5 users. Nielsen Norman Group. 2000. URL: https://www.nngroup.com/articles/why-you-only-need-to-test-with-5-users/ [accessed 2026-08-14]
119. Nielsen J, Landauer TK. A mathematical model of the finding of usability problems. 1993 Presented at: INTERACT '93 and CHI '93 Conference on Human Factors in Computing Systems; April 24-29, 1993; Amsterdam, Netherlands p. 206-213. [doi: 10.1145/169059.169166]
120. Bevan N, Barnum C, Cockton G, Nielsen J, Spool J, Wixon D. The “magic number 5”: is it enough for web testing? 2003 Presented at: CHI ’03 Extended Abstracts on Human Factors in Computing Systems; April 5-10, 2003; Fort Lauderdale, Florida p. 698-699. [doi: 10.1145/765891.765936]
121. Sauro J. A brief history of the magic number 5 in usability testing. MeasuringU. URL: https://measuringu.com/five-history/ [accessed 2026-08-14]
122. Gwet KL. Handbook of inter-rater reliability: the definitive guide to measuring the extent of agreement among raters. Gaithersburg, Maryland: Advanced Analytics, LLC; 2014.

Abbreviations

AAMC: Association of American Medical Colleges
AR: augmented reality
ATA: Agile Task Analysis
EHR: electronic health record
FCM: family and community medicine
HTA: hierarchical task analysis
NASA: National Aeronautics and Space Administration
NASA-TLX: NASA Task Load Index
OB/GYN: obstetrics and gynecology
OUSCM: University of Oklahoma School of Community Medicine
SDS: Simulation Design Scale
SEQ: Single Ease Question
SP: standardized patient
SSES: Satisfaction With Simulation Experience Scale
SUS: System Usability Scale
VR: virtual reality

Edited by A Stone; submitted 29.May.2026; peer-reviewed by LC Wetzlmair, K Grewal; comments to author 01.Jul.2026; revised version received 22.Jul.2026; accepted 27.Jul.2026; published 09.Sep.2026.

Please cite as:
Lesselroth BJ, Douglas A, Monkman H, Marcilly R, Bellavia K, Anderson M, Van Buren J, Parsons C, Mohanty S, Homco J, Richards M, Gold KP
How to Conduct Usability Testing of Medical Education Simulations: Methods Tutorial Using a Telemedicine Simulation for Maternal Care
JMIR Med Educ 2026;12:e102697
URL: https://mededu.jmir.org/2026/1/e102697
doi:10.2196/102697
PMID:

©Blake J Lesselroth, Alexander Douglas, Helen Monkman, Romaric Marcilly, Karalane Bellavia, Mataeo Anderson, Jacob Van Buren, Charles Parsons, Sonakshi Mohanty, Juell Homco, Morgan Richards, Karen P Gold. Originally published in JMIR Medical Education (https://mededu.jmir.org), 09.Sep.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Education, is properly cited. The

XSL•FO
RenderX

complete bibliographic information, a link to the original publication on https://mededu.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

A Competency Assessment Tool for Liver Contrast-Enhanced Ultrasound Among Sonography Trainees: Development and Multicenter Validation Study

Yanping Ma[1]; Bowen Zheng[1]; Jie Ren[1*]; Wen He[2*]; Lei Shang[3]; Hui Wang[4]; Gaiqin Xue[5]; Fang Nie[6]; Yanli Guo[7]; Yukun Luo[8]; Hongxia Zhang[2]; Xuan Wang[2]; Xiang Jing[9]; Ruifang Zhang[10]; Jianjun Yuan[11]; Rong Wu[12]; Xiaomao Luo[13]; Fenglin Dong[14]; Yan Luo[15]; Qiang Lu[15]; Jintang Liao[16]; Pintong Huang[17]; Qin Chen[18]; Jia'an Zhu[19]; Youfeng Xu[20]; Xiaoyan Xie[21]

[1]Department of Medical Ultrasonics, Third Affiliated Hospital of Sun Yat-sen University, 600 Tianhe Road, Guangzhou, China
[10]First Affiliated Hospital of Zhengzhou University, Zhengzhou, China
[11]Henan Provincial People's Hospital, Zhengzhou, China
[12]Shanghai First People's Hospital, Shanghai, China
[13]Yunnan Cancer Hospital, Kunming, China
[14]First Affiliated Hospital of Soochow University, Soochow, China
[15]West China Hospital of Sichuan University, Sihchuan, China
[16]Xiangya Hospital Central South University, Changsha, China
[17]Second Affiliated Hospital of Zhejiang University, Hangzhou, China
[18]Sichuan Academy of Medical Sciences & Sichuan Provincial People's Hospital, Chengdu, China
[19]Peking University People's Hospital, Beijing, China
[2]Beijing Tiantan Hospital, Capital Medical University, Beijing, China
[20]The Affiliated First Hospital of Ningbo University, Ningbo, China
[21]The First Affiliated Hospital, Sun Yat-sen University, Guangzhou, China
[3]Department of Health Statistics, School of Preventive Medicine, Air Force Medical University, Xi'an, China
[4]Department of Ultrasound, China-Japan Union Hospital of Jilin University, Jilin, China
[5]Shanxi Provincial Cancer Hospital, Taiyuan, China
[6]Lanzhou University Second Hospital, Lanzhou, China
[7]The First Hospital Affiliated to AMU, Chongqing, China
[8]Chinese PLA General Hospital, Beijing, China
[9]Tianjin Third Central Hospital, Tianjin, China
[*]these authors contributed equally

Corresponding Author:
Jie Ren
Department of Medical Ultrasonics, Third Affiliated Hospital of Sun Yat-sen University, 600 Tianhe Road, Guangzhou, China

Abstract

Background: Despite the increasing use of liver contrast-enhanced ultrasound (CEUS), no standardized, organ-specific tool exists for assessing trainees' cognitive competency. Existing instruments lack multidimensional evaluation and external validation, which limit training quality and patient safety.

Objective: This study aims to develop and validate the CEUS Competency Assessment for the Liver (CCAL), a tool to standardize the evaluation of theoretical and cognitive skills in liver CEUS.

Methods: This retrospective cohort study had four stages: (1) item generation from literature and guidelines, (2) 3 Delphi rounds (12/21, 57.1% female experts; mean age 51.0, SD 7.2 y) and cognitive testing (10 trainees) for content validity (>85% agreement) and clarity, (3) internal validation involving 100 sonography trainees (n=66, 66% assistant professors; n=27, 27% associate professors; n=7, 7% professors; n=70, 70% CEUS-naive) with 1 month or more of training at 4 hospitals (2018 - 2020), and (4) external validation on 579 sonography trainees (n=299, 51.6% assistant professors; n=183, 31.6% associate professors; n=97, 16.8% professors; n=439, 75.8% CEUS-naive) with 1 or more months of training at 25 multicenter sites (2018 - 2020). Primary outcomes were reliability (Cronbach α and Guttman split-half) and validity (exploratory factor analysis and item discrimination), while secondary outcomes included item difficulty distribution, pass rates, and between-group differences.

XSL•FO
RenderX

Results: The final CCAL comprised 74 items across 3 domains (demographics: 8, self-evaluation: 13, and objective questions: 53). Internal consistency was excellent (Cronbach α=0.916 and Guttman split-half=0.908). Factor analysis confirmed a 3-factor structure explaining 74.6% of the cumulative variance. Item analysis showed a balanced difficulty distribution (35/53, 66.0% easy; 17/53, 32.1% moderate; 1/53, 1.9% difficult items). Scores varied significantly by CEUS training duration (P<.001); trainees with 4 to 6 months or more than 6 months of training achieved higher self-evaluation ($F_{2,576}$=34.751, partial η^2=0.108, 95% CI 0.069-0.147) and objective question ($F_{2,576}$=34.374, partial η^2=0.107, 95% CI 0.068-0.146) scores than those with less than 3 months of training. Prior independent CEUS experience was also associated with higher scores on the self-evaluation (t_{577}=6.736, Cohen d=0.561, 95% CI 0.393-0.728; P<.001) and objective question sections (t_{577}=3.494, Cohen d=0.291, 95% CI 0.127-0.454; P=.001). Competency gaps were identified across specific competency domains: the pass rate ranged from 32.5% (188/579; image-guided intervention) to 81.3% (471/579; indications and contraindications).

Conclusions: The CCAL is the first liver-specific CEUS competency tool integrating objective knowledge testing with structured self-evaluation. Unlike generic or simulation-based tests, it is organ-specific and validated across a large multicenter cohort, providing a multidimensional framework that identifies specific competency gaps, most notably in advanced image-guided procedures, beyond global pass/fail outcomes. The tool supports competency-based medical education through rigorous data-driven assessment, targeted curricular improvements, cross-institutional benchmarking, and informed training standards. It evaluates cognitive competence exclusively and should not replace procedural assessment in high-stakes decisions. International validation and the establishment of criterion-related and predictive validity are essential next steps.

Trial Registration: Chinese Clinical Trial Registry ChiCTR2000035750; https://tinyurl.com/mpan5bc9

(JMIR Med Educ 2026;12:e86730) doi:10.2196/86730

KEYWORDS

contrast-enhanced ultrasound; CEUS; competency assessment; liver; validation; training; competency-based medical education; CBME

Introduction

Problem

Contrast-enhanced ultrasound (CEUS) has become an indispensable diagnostic tool for liver diseases, offering real-time, radiation-free characterization of focal liver lesions and dynamic depiction of tissue microcirculation [1]. Its clinical value has been endorsed by more than 20 international guidelines from organizations such as the World Federation for Ultrasound in Medicine and Biology (WFUMB) and the European Federation of Societies for Ultrasound in Medicine and Biology (EFSUMB) [2,3]. Nevertheless, the diagnostic accuracy of CEUS is highly operator-dependent [4]. Effective application of CEUS demands high technical proficiency from operators, as the procedure is time-sensitive and susceptible to visual interference [5]. Therefore, ensuring that trainees attain standardized and objectively verifiable competencies before transitioning to independent practice is essential not only for safeguarding training quality but also, more critically, for protecting patient safety and ensuring diagnostic accuracy.

In response to this need, competency-based medical education (CBME) has emerged as a paradigm in medical training, shifting the focus from time-based curricula to demonstrable outcomes that ensure trainees acquire the knowledge, skills, and attitudes necessary for independent practice [6,7]. This approach mandates the use of standardized, validated assessment tools to objectively evaluate clinical competencies and safeguard training quality. According to the Messick unified validity framework, validity is not a property of the test itself but of the interpretations and actions based on test scores. Therefore, assessment tools must gather evidence across 5 sources—content, response process, internal structure, relations to other variables, and consequences—to support high-stakes decisions such as certification or independent practice [8,9]. However, in the field of CEUS—particularly for liver applications, which represent a core clinical scenario—there is a lack of validated assessment tools to objectively evaluate the theoretical knowledge and operational competence of trainees.

Review of Relevant Scholarship

The importance of standardized competency assessment in medical education has been increasingly recognized. Recent systematic studies have shown that competency-based curricula require robust evaluation practices to ensure training quality, yet significant variation exists in how assessment tools are developed, implemented, and reported [10,11]. Alharbi [10] systematically reviewed evaluation practices in CBME curricula and found that, while questionnaires remain the primary assessment tool, numerous studies fail to clarify evaluation objectives, approaches, and standards, highlighting the need for more rigorous and transparent assessment methodologies. Similarly, a systematic review of undergraduate CBME assessment identified a range of constructivist techniques—including concept maps, portfolios, rubrics, and workplace-based assessments—as promising approaches, but noted that their implementation remains inconsistent and often lacks validity evidence [11].

Within the ultrasound education domain, recent years have witnessed significant efforts to establish competency frameworks. The ULTRA-Metrics study, a Delphi-derived international consensus initiative, developed a modular framework for assessing ultrasound competency across clinical applications and disciplines, identifying 3 core competency domains (experience, skills, and autonomy) and 4 skill domains (indication, acquisition, interpretation, and clinical integration)

XSL•FO
RenderX

[12]. This framework highlighted that existing approaches often lack generalizability, overemphasize exam counts, and fail to assess key skills such as interpretation, thereby limiting ultrasound's safe and effective integration into clinical practice. Similarly, Delphi consensus studies have been used to define ultrasound competencies in various specialties, including emergency medicine, anesthesiology, and radiology residency programs, reflecting a growing recognition of the need for competency-based ultrasound training [13,14]. However, these frameworks remain largely generic and have not been translated into organ-specific, validated assessment instruments.

Specifically, in the field of CEUS, several efforts have been made to develop competency assessment tools, yet substantial psychometric gaps persist. Jacobsen et al [15,16] developed a theoretical multiple-choice test for core CEUS competencies and gathered validity evidence according to the Messick framework, but the final 47-item test showed a considerable risk of false positives and false negatives when establishing a pass/fail score. The same research group also developed an immersive virtual reality simulation-based test with acceptable reliability (Cronbach's α=0.85), yet it carried a 23.6% risk of false negatives and focused exclusively on precontrast preparations rather than on comprehensive liver CEUS competency [17]. Furthermore, a scoping search of the literature on CEUS curricula and assessment revealed very limited published work in this area, confirming the scarcity of validated assessment tools.

Beyond CEUS, the broader ultrasound education literature provides instructive parallels. A systematic review of endoscopic ultrasound (EUS) competency assessment tools identified several instruments with strong validity evidence—including TEESAT (The EUS and Endoscopic Retrograde Cholangiopancreatography Skills Assessment Tool), GAPS-EUS (Global Assessment of Performance and Skills in EUS), and EUSAT (EUS Assessment Tool)—but noted that these tools remain specialty-specific and are not directly transferable to other ultrasound applications [18]. Similarly, organ-specific ultrasound assessment tools have been developed for various applications, including lung ultrasound (LUS)–Objective Structured Assessment of Ultrasound Skills (OSAUS) and cranial ultrasound, demonstrating the feasibility of organ-specific competency assessment [19,20]. Paradoxically, although the liver is a primary indication for CEUS, accounting for the majority of global CEUS examinations, no equivalent organ-specific tool has been developed for liver CEUS competency assessment. A 2024 study on CEUS Liver Imaging Reporting and Data System (LI-RADS) classification training for radiology residents demonstrated that dedicated training improved diagnostic performance, yet no standardized assessment tool accompanied that training, underscoring a critical gap between educational intervention and rigorous outcome evaluation [21].

Rationale and Research Questions

The convergence of the above evidence reveals a clear and urgent knowledge gap: there is currently no validated, organ-specific, and psychometrically sound competency assessment tool for liver CEUS. While generic ultrasound competency frameworks have been proposed, they lack the specificity needed for liver CEUS, which requires unique knowledge of hepatic vascular anatomy, LI-RADS classification, contrast kinetics, and image-guided interventional procedures. Furthermore, previously published CEUS tools have been limited by insufficient validity evidence (eg, undefined cutoffs) or overly narrow scope (eg, covering only precontrast preparations). This deficiency has direct clinical consequences; without robust evaluation criteria, high-stakes decisions regarding whether trainees can independently perform liver CEUS examinations become subjective, ultimately compromising patient safety and health care quality. Given the high operator dependency of CEUS and the growing clinical demand for liver examinations, developing a rigorous assessment tool is an urgent quality assurance priority.

Hypothesis, Aims, and Objectives

To address these gaps, the present study aimed to develop and validate the first standardized competency assessment tool specifically for liver CEUS (the CEUS Competency Assessment for the Liver [CCAL]). We hypothesized that the CCAL would demonstrate the following characteristics: (1) strong content validity, as established through a Delphi expert consensus process (content evidence); (2) high internal consistency reliability and satisfactory item discriminability (internal structure evidence); and (3) robust external validity across diverse training sites (relations to other variables evidence). Accordingly, through a Delphi process involving 21 experts and a multicenter validation study comprising over 600 trainees, this study sought to provide a reliable, psychometrically sound instrument for objective competency evaluation in liver imaging, thereby filling a critical void in CBME for CEUS training.

Methods

Ethical Considerations

This study involved human participants and adhered to the Declaration of Helsinki. Ethical approval was obtained from the Institutional Review Board (IRB) of The Third Affiliated Hospital of Sun Yat-sen University (IRB number: [2020] 02-137-01) and was registered with the Chinese Clinical Trial Registry (ChiCTR2000035750). Written informed consent was secured from all participants prior to enrollment, and they were informed of their right to withdraw at any time without academic or other repercussions. All collected data were anonymized and deidentified before analysis, with no personal identifiers retained in the research database. Participation was voluntary and part of routine assessment procedures; therefore, no compensation was provided. The manuscript and supplementary materials contain no images or content that could identify individual participants.

Study Procedures

Research Design

This study used a nonexperimental, observational, cross-sectional design with multiple-group comparisons. It was conducted from July 2020 to July 2021 and adopted a four-phase mixed methods design. It systematically followed established standards for the development of medical education assessment

XSL•FO
RenderX

tools as follows: (1) item generation, (2) content refinement, (3) psychometric validation, and (4) multicenter validation [22-26]. The methodology was carefully structured to address the recognized lack of standardized competency assessment instruments for liver CEUS.

Phase 1: Item Generation

Domain Identification

The assessment domains were established based on standardized curricula from Chinese CEUS training centers and are in accordance with the 2020 WFUMB CEUS Liver Guidelines [2]. The core domains include the principles of CEUS and device optimization, diagnostic characterization of liver lesions utilizing CEUS, and management of adverse reactions.

Initial Item Pool Generation

Six experts in CEUS independently developed items for the self-evaluation section (SES; a 5-point Likert scale) and the objective question section (OQS; multiple-choice or single-best-answer format). The content of the OQS included indications and contraindications, scanning protocols, and imaging characteristics of liver lesions.

Phase 2: Expert Consultation and Cognitive Testing

Expert Consensus Development

A 3-round modified Delphi process was conducted from September to December 2020 with 21 experts (n=12, 57.1% female experts; mean age 51.0, SD 7.2 y; mean ultrasound experience 25.5 , SD 6.8 y; mean CEUS experience 14.8 , SD 6.0 y; Table 1). In round 1, experts rated each item on a 5-point scale and suggested additions; items achieving >85% agreement were retained, while others were revised or removed. Rounds 2 and 3 involved aggregated feedback and rerating, and the process continued until consensus was reached (Multimedia Appendix 1). Items were blindly reviewed to mitigate groupthink and ensure independence. The method integrated clinical practice needs with existing guideline recommendations [16,27].

Table . Demographic and professional characteristics of the 21 experts who participated in the 3-round Delphi consensus process for content validation of the CEUS Competency Assessment for the Liver (CCAL) conducted from September 2020 to December 2020.

Category	Values, n
Sex	
Male	9
Female	12
Administrative position	
Department director	19
Other	2
Title	
Chief physician	19
Associate chief physician	1
Associate professor	1
Academic level	
Bachelor	5
Master	3
Doctor	13
Title in the academic committee	
Chairman	9
Vice chairman	5
Others	7
Geographic distribution	
South	7
North	7
Central	4
East	3
Specialty	
Abdomen	19
Intervention	2

XSL•FO
RenderX

Cognitive Testing

A pilot study was conducted with 10 trainees (5 novices with <1 mo of CEUS experience, 5 experienced with >6 mo of experience) who were interviewed regarding item clarity, comprehensibility, and wording ambiguity. Their feedback was used to inform subsequent revisions.

Phase 3: Internal Validation and Revision

Aim

A quantitative analysis was conducted utilizing internal validation trainees, and items of the trial version were screened to develop the final version in January 2021 [28,29].

Inclusion and Exclusion

Consecutive enrollment included trainees who met both of the following criteria: (1) completion of 1 or more months of CEUS training between January 2018 and December 2020 at 1 of 4 leading teaching hospitals—namely, the Chinese PLA General Hospital, Beijing Tiantan Hospital (Capital Medical University), the Second Hospital of Lanzhou University, and the Third Affiliated Hospital of Sun Yat-sen University; the training program comprised at least 30 hours per week of supervised hands-on scanning and structured didactic instruction; and (2) performance of more than 10 independent liver CEUS examinations after formal training completion. Exclusion criteria were as follows: (1) unwillingness to participate in the standardized competency assessment, and (2) prior completion of 2 or more CEUS training programs. All enrolled trainees were native Mandarin speakers and completed the assessment in Chinese. A total of 100 trainees met all eligibility criteria and were included in the study.

Participant Characteristics

Among 100 sonography trainees, 47% (n=47) held a bachelor's degree, 43% (n=43) a master's degree, and 10% (n=10) a doctorate. Professional titles included the following: 66% (n=66) assistant professors, 27% (n=27) associate professors, and 7% (n=7) full professors. Ultrasound experience was as follows: 48% (n=48) had 1 - 3 years, 20% (n=20) had 4 to 5 years, and 32% (n=32) had >5 years of experience. Before training, 70% (n=70) had no independent CEUS experience. CEUS training duration was as follows: 58% (n=58) trained for 1 to 3 months, 24% (n=24) for 4 to 6 months, and 18% (n=18) for more than 6 months. The mean number of CEUS examinations performed per trainee was 81.71 (SD 112.96; Multimedia Appendix 2).

Sampling Procedures

Participants were recruited consecutively from the 4 predetermined teaching hospitals. No random or stratified sampling was used; every eligible trainee was invited to participate. The participation rate was 100% (100/100) because the assessment was part of the routine training completion process. Self-selection at the individual level did not occur; institutional self-selection did not apply, as the 4 hospitals were preselected collaborators. No compensation was provided.

Sample Size, Power, and Precision

For internal validation, the sample size was determined pragmatically based on the number of eligible trainees available at the 4 sites during the study period (January 2018-December 2020). A minimum of 100 participants was considered adequate for initial item screening and exploratory factor analysis, following established recommendations for scale development (subject-to-item ratio ≥5:1). The achieved sample of 100 trainees exceeded this threshold for the OQS items (ratio ≈1.9:1, which is lower than the recommended 5:1; however, this was acceptable for preliminary validation and was compensated by the larger external validation sample).

Measures and Covariates

Primary measures were the SES (5-point Likert scale) and the OQS (multiple-choice or image-based). Covariates included age, sex, academic degree, professional title, years of ultrasound experience, prior independent CEUS experience (yes or no), CEUS training duration (mo), and the number of CEUS examinations performed during training. All covariates were collected using an 8-item demographic questionnaire during the assessment.

Data Collection and Quality of Measurements

The trial version was hosted online. Trainees accessed the assessment via a direct link and completed all items independently, with webcam monitoring and page-switching prohibited to prevent cheating. All participants were given a uniform time limit of 60 minutes. The website automatically recorded and stored the results and calculated the scores. The dataset contained no missing values; all enrolled participants fully completed the assessment, and no imputation methods were applied.

Phase 4: External Validation

Aim

To further expand the scope of validation, a multicenter external validation study was conducted in April 2021 [24].

Inclusion and Exclusion

Consecutive enrollment included trainees who met both of the following criteria: (1) completion of 1 month or more of CEUS training between January 2018 and December 2020 at 25 leading teaching centers distributed across multiple regions of China; the training program comprised at least 30 hours per week of supervised hands-on scanning and structured didactic instruction; and (2) performance of more than 10 independent liver CEUS examinations after formal training completion. The same exclusion criteria were applied. All enrolled trainees were native Mandarin speakers and completed the assessment in Chinese.

Participant Characteristics

Among the 579 sonography fellows, the majority held a bachelor's degree (n=336, 58.0%), were CEUS-naive before training (n=439, 75.8%), and received 1 to 3 months of CEUS training (n=337, 58.2%; Table 2). The mean number of CEUS examinations performed per trainee during training was 64.72 (SD 107.81).

XSL•FO
RenderX

Table . Demographic and clinical characteristics of trainees (N=579) in the external validation phase (phase 4) of the contrast-enhanced ultrasound (CEUS) competency assessment for the liver (CCAL) development study across 25 multicenter sites in China for liver CEUS competency assessment conducted in April 2021.

Characteristics	Values, n (%)
Sex	
Male	125 (21.6)
Female	454 (78.4)
Academic level	
Bachelor	336 (58.0)
Master	185 (31.9)
Doctor	58 (10.0)
Title	
Assistant professor	299 (51.6)
Associate professor	183 (31.6)
Professor	97 (16.8)
Years engaged in ultrasound	
1 - 3	204 (35.2)
4-5	89 (15.4)
>5	286 (49.4)
Had independent CEUS examination experience before training	
Yes	140 (24.2)
No	439 (75.8)
Duration of CEUS training (mo)	
1 - 3	337 (58.2)
4 - 6	116 (20.0)
>6	126 (21.8)

Sampling Procedures

Participants were recruited consecutively from 25 multicenter sites that agreed to participate (out of 45 invited; response rate n=25, 55.6%). At the trainee level, all eligible individuals were approached, and the participation rate was 100% (579/579). No individual self-selection occurred because the assessment was part of routine training completion. At the institutional level, self-selection was possible (centers chose to participate or decline). No compensation was provided.

Sample Size, Power, and Precision

Sample size was determined a priori using a 10:1 subject-to-item ratio for the final 53 OQS items, requiring at least 530 participants. The external validation cohort included 579 participants, exceeding the required sample size.

Measures and Covariates

The same measures (SES and OQS) and covariates (demographic characteristics, prior experience, training duration, and number of CEUS exams) were used as in phase 3. All covariates were collected via an 8-item demographic questionnaire during the assessment.

Data Collection and Quality of Measurements

Same as the internal validation part mentioned above.

Analytic Strategy

The Primary Outcomes

The primary outcomes included the following: (1) internal consistency reliability, with Cronbach α and the Guttman split-half coefficient ≥0.8 considered acceptable; (2) construct validity, assessed via exploratory factor analysis (principal axis factoring with varimax rotation). The Kaiser-Meyer-Olkin (KMO) measure of sampling adequacy was >0.8, and Bartlett test of sphericity was significant (P<.001), indicating that the data were suitable for factor analysis. The extracted factor structure was examined for alignment with the theoretically hypothesized structure; and (3) the index of discrimination (D), evaluated by the correlation between each item score and the total OQS score, where items were retained only if the discrimination index D was ≥0.4 [30,31].

The Secondary Outcomes

The secondary outcomes included the following:

1. The distribution of item difficulty (ie, the proportion of easy, moderate, and difficult items). The difficulty index

XSL•FO
RenderX

(H) was used to describe the difficulty of the OQS, where H of 0.3 or less indicated easy questions, 0.3<H<0.7 indicated moderately difficult questions, and H of 0.7 or more indicated difficult questions [32-34]. H was calculated as 1–P, where P is the proportion of correct responses.

2. Pass rates for each knowledge domain.
3. Between-group differences in overall OQS scores across trainee subgroups (eg, stratified by duration of CEUS training or prior clinical experience).

Item Analysis

Items exhibiting low difficulty (H≤0.3) or poor discrimination (D≤0.2) were removed after the first phase (n=13 items), which resulted in the final item set.

Statistical Analysis

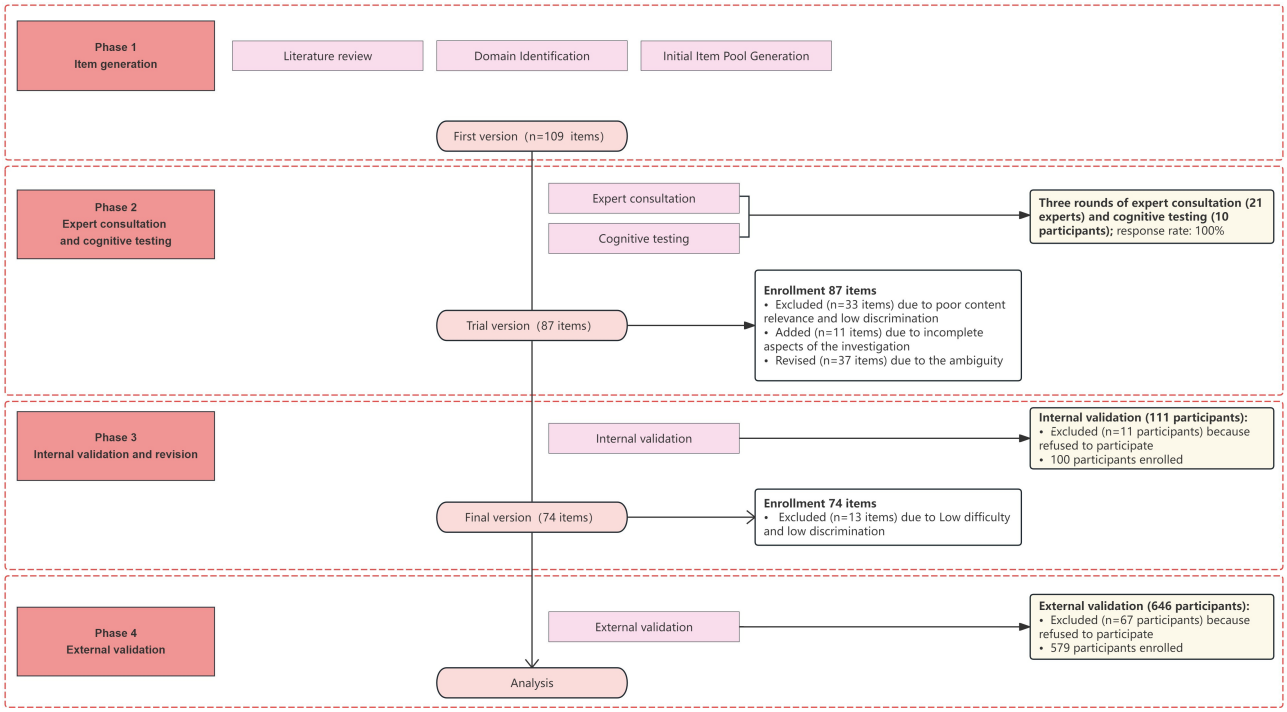
All statistical analyses were conducted using SPSS Statistics (version 26.0; IBM). Continuous variables are presented as the mean (SD), and categorical variables are presented as the frequency and percentage. The group comparisons used *t* tests, ANOVA, or nonparametric tests (Mann-Whitney *U* test and Kruskal-Wallis test) on the basis of data type and distribution. Associations were assessed with chi-square tests and Pearson correlations (P<.05, 2-tailed).

Results

Participant Flow

In phase 1, 6 experts generated items independently without formal enrollment or attrition tracking. In phase 2, all 21 invited experts and 10 recruited trainees completed the Delphi process and cognitive interviews, respectively, with no attrition. For phase 3 internal validation, 111 trainees met the eligibility criteria; 11 declined participation, and the remaining 100 were enrolled and completed the assessment in full, with no missing data or further withdrawals. In phase 4 external validation across 25 hospitals, 646 trainees were assessed for eligibility; 67 declined participation, and 579 were enrolled and completed the assessment with no missing data or withdrawals (Figure 1).

Figure 1. Retrospective study design and participant flow diagram of the 4-phase development and validation process of the contrast-enhanced ultrasound (CEUS) Competency Assessment for the Liver (CCAL) for trainees in liver CEUS, conducted across 25 multicenter sites in China from July 2020 to July 2021.



Recruitment

Recruitment was conducted in 4 phases. Phase 1 (item generation) took place in July 2020. Phase 2 (expert consultation and cognitive testing) ran from September to December 2020. Phase 3 (internal validation) was conducted in January 2021 at 4 preselected teaching hospitals. Phase 4 (external validation) occurred in April 2021. The study used a cross-sectional design with no repeated measures or follow-up.

Statistics and Data Analysis

Item Selection Process

To address the critical gap in the objective assessment of liver CEUS competency, this study developed the CCAL. An initial pool of 109 items (demographic: 8; SES: 24; OQS: 77) was reduced to a final version of 74 items (demographic: 8; SES: 13; OQS: 53) following 3 rounds of optimization (expert consultation, cognitive interviews, and internal validation). The final items span 3 clinically relevant domains: technical proficiency (eg, scanning protocols and device operation), diagnostic decision-making (eg, lesion characterization and

XSL•FO
RenderX

interpretation), and clinical judgment (eg, indication selection and complication management).

Psychometric Validation

The final scale was then systematically evaluated across multiple dimensions using an external validation cohort of 579 trainees.

Reliability

Reliability analysis demonstrated excellent internal consistency: Cronbach α=0.916 and Guttman split-half coefficient=0.908, indicating strong intercorrelations among items and consistent measurement of the construct.

Construct Validity

Construct validity was supported by a KMO of 0.921 and a significant Bartlett test (P<.001), confirming suitability for factor analysis and alignment with clinical competency requirements.

The scree test extracted 3 common factors, accounting for 74.6% of the cumulative variance, with all item factor loadings above 0.67. Factor 1 represents technical proficiency and encompasses device optimization and contraindication recognition, which reflect procedural execution capability. Factors 2 and 3 jointly capture diagnostic efficacy, specifically the improvement in diagnostic accuracy, which reflects interpretive competence and clinical decision-making skills (Table 3). This empirically derived factor structure strongly corresponds to the requirements of clinical competency.

Table . Factor loadings and dimension attribution of the self-evaluation section (SES) of the CEUS Competency Assessment for the Liver (CCAL) during the external validation phase (phase 4) among 579 trainees from 25 centers in China for the liver contrast-enhanced ultrasound (CEUS) competency assessment conducted in April 2021.

Items	Factor loading	Dimension attribution
Do you think it is necessary to perform CEUS examination in patients with suspected intrahepatic malignant lesions for the first time?	0.786	3
Do you think it is necessary to perform CEUS examination in patients with suspicious intrahepatic benign lesions for the first time?	0.855	3
During CEUS, I can find lesions that are not visible on 2D ultrasound.	0.672	2
I believe I have mastered the principles of CEUS technology.	0.900	1
I believe I have mastered the adjustment of the instruments during CEUS.	0.910	1
I believe I have mastered the indications and contraindications for CEUS.	0.863	1
I believe I can handle the adverse effects of CEUS with confidence.	0.854	1
I can use CEUS to qualitatively diagnose the vast majority of target lesions in general.	0.734	1
I believe CEUS has significantly improved the accuracy of my diagnosis.	0.857	2
I believe that the CEUS images and report descriptions of mine are normative.	0.832	1
I believe CEUS can improve the positive rate of tissue biopsy of mine.	0.837	2
How well do you think you have mastered the CEUS examination?	0.817	1
In order to accurately diagnose target lesions with comprehensive information, CEUS physicians need to be proficient in reading CT[a] and MRI[b] images.	0.754	2

[a]CT: computed tomography.

[b]MRI: magnetic resonance imaging.

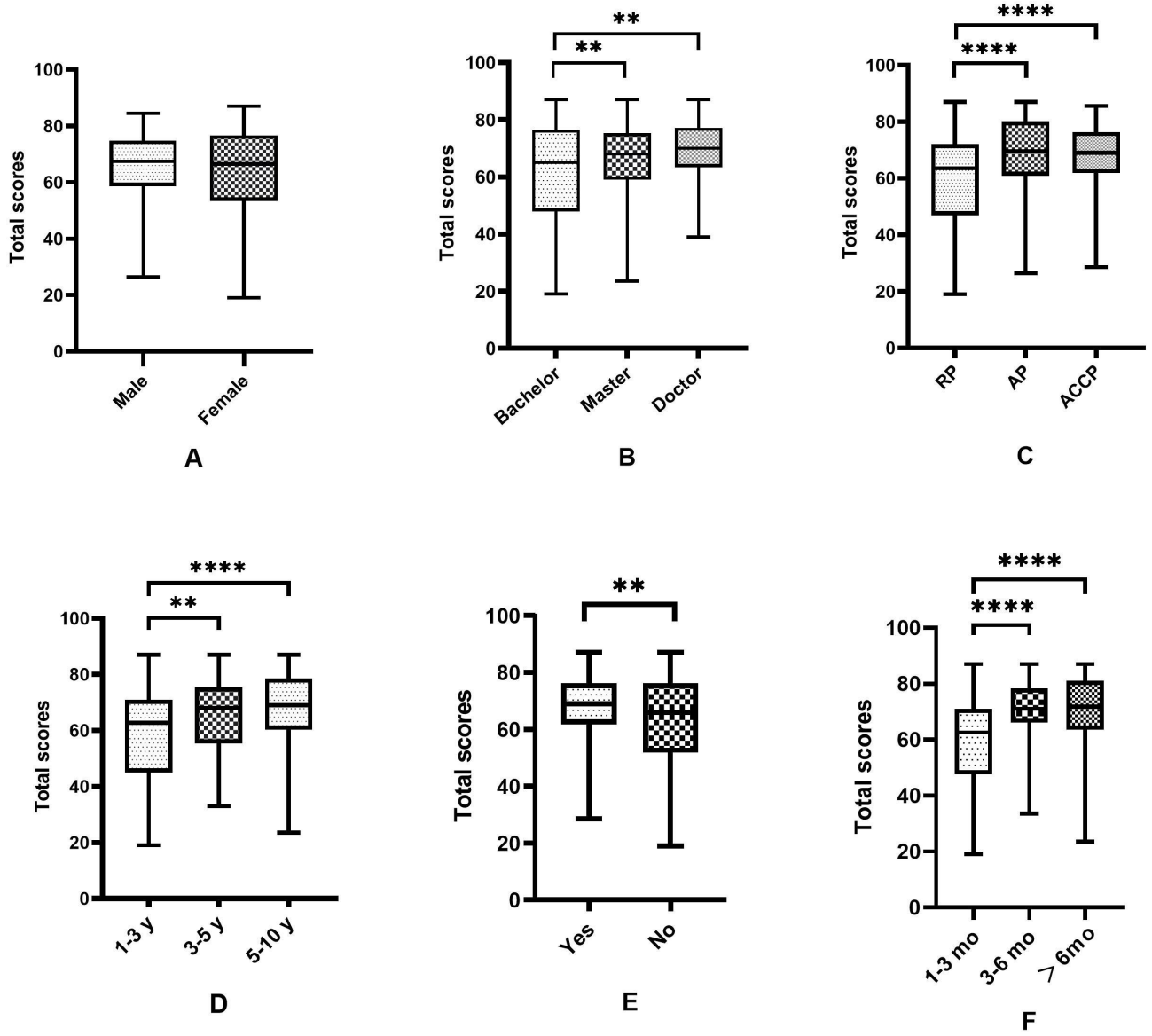
Discriminant Validity

Discriminant validity was examined to determine whether the scale distinguishes trainees by competence level. Scores differed significantly by experience and professional rank. Specifically, trainees with 4 or more months of CEUS training scored higher on both self-evaluation ($F_{2,576}$=34.751; partial η^2=0.108, 95%

XSL•FO
RenderX

CI 0.069-0.147) and objective questions ($F_{2,576}$=34.374; partial η^2=0.107, 95% CI 0.068-0.146) than those with less than 3 months (P<.001). Additionally, those with prior independent CEUS experience also scored higher on both the self-evaluation (t_{577}=6.736; P<.001; Cohen d=0.561, 95% CI 0.393-0.728) and objective question sections (t_{577}=3.494; P=.001; Cohen d=0.291, 95% CI 0.127-0.454). These findings confirm that the scale effectively differentiates trainees by ability level (Figure 2;Multimedia Appendices 3 and 4).

Figure 2. Box plots of total objective question section scores by trainee characteristics among 579 trainees in the external validation phase (phase 4) of the multicenter study of the Contrast-Enhanced Ultrasound Competency Assessment for the Liver (CCAL) for competency assessment in liver contrast-enhanced ultrasound, conducted in April 2021. (A) Total scores grouped by sex (male: fine black dot fill; female: coarse black-and-white checkerboard fill). (B) Total scores grouped by academic level (bachelor: fine black dot fill; master: coarse black-and-white checkerboard fill; doctor: fine grey-and-white checkerboard fill). (C) Total scores grouped by title (RP: resident physician, fine black dot fill; AP: assistant professor, coarse black-and-white checkerboard fill; ACCP: associate chief physician and chief physician, fine grey-and-white checkerboard fill). (D) Total scores grouped by years engaged in ultrasound (1-3 y: fine black dot fill; 3-5 y: coarse black-and-white checkerboard fill; 5-10 y: fine grey-and-white checkerboard fill). (E) Total scores grouped by prior independent CEUS experience (yes: fine black dot fill; no: coarse black-and-white checkerboard fill). (F) Total scores grouped by duration of CEUS training (1-3 mo: fine black dot fill; 3-6 mo: coarse black-and-white checkerboard fill; >6 mo: fine grey-and-white checkerboard fill). The ** denotes P<.01 and **** denotes P<.0001.



Spearman correlation analysis revealed a significant positive correlation between trainees' objective question scores and self-evaluation scores (rs=0.551; P<.001).

Difficulty

The OQS demonstrated excellent discrimination, with a discrimination index of 0.429, exceeding the established threshold for good discrimination (D≥0.4). Moreover, the OQS items were appropriately distributed across difficulty levels, covering a broad range of trainee ability: 66.0% (35/53, 95% CI 0.526 - 0.773) of items were categorized as simple questions (covering basic knowledge points), 32.1% (17/53, 95% CI 0.211 - 0.455) as medium-level questions (pertaining to clinical applications), and 1.9% (1/53, 95% CI 0.003 - 0.099) as challenging questions (involving advanced skills).

XSL•FO
RenderX

Identification of Weaknesses in Competency

Quantitative analysis of module-specific pass rates (Table 4) identified critical competency gaps: the pass rate for "indications and contraindications" was 81.3% (471/579, 95% CI 0.780 - 0.843), indicating solid command of fundamental knowledge, whereas the pass rate for "imaging features for procedure guidance" was only 32.5% (188/579, 95% CI 0.288 - 0.364), highlighting insufficient training in CEUS application for clinical procedures.

XSL•FO
RenderX

Table . Comparison of scores across different content domains (indications, scanning protocols, imaging features, and procedural guidance) of the objective question section (OQS) in the external validation phase (phase 4) among 579 trainees from 25 Chinese centers for liver contrast-enhanced ultrasound (CEUS) competency in April 2021.

Category	Indications and contraindications		Scanning procedures and general operating steps	Imaging features of space-occupying lesions in the liver		Imaging features used to guide puncture, tumor ablation, or other procedures.	
	Scores, mean (SD)	Pass rate, % (95% CI)	Scores, mean (SD)	Scores, mean (SD)	Pass rate, % (95% CI)	Scores, mean (SD)	Pass rate, % (95% CI)
Sex							
Male	12.73 (2.16)	81.6 (0.739 - 0.874)	13.58 (3.00)	35.38 (9.02)	79.2 (0.713 - 0.854)	4.56 (2.40)	38.4 (0.303 - 0.472)
Female	12.47 (2.57)	81.3 (0.774 - 0.846)	12.69 (3.28)[a]	34.58 (10.27)	78.6 (0.746 - 0.822)	4.94 (2.26)	30.8% (0.268 - 0.352)
Academic level							
Bachelor	12.31 (2.70)	85.7[b] (0.816 - 0.891)	12.29 (3.41)	33.18 (11.02)	78.9 (0.742 - 0.829)	4.77 (2.29)	32.4 (0.277 - 0.376)
Master	12.82 (2.18)	75.7 (0.690 - 0.813)	13.54 (2.81)[c]	36.42 (8.37)[c]	81.1 (0.748 - 0.861)	4.92 (2.37)	30.3 (0.241 - 0.372)
Doctor	12.85 (1.92)	74.1 (0.616 - 0.837)	14.18 (2.73)[c]	38.54 (6.17)[c]	70.7 (0.580 - 0.808)	5.19 (2.04)	39.7 (0.281 - 0.525)
Title							
Assistant professor	12.25 (2.61)	73.6[b] (0.683 - 0.783)	12.24 (3.40)	32.18 (10.68)	74.2[b] (0.690 - 0.789)	4.52 (2.28)	29.8 (0.249 - 0.352)
Associate professor	12.94 (2.40)[d]	92.9 (0.882 - 0.958)	13.64 (2.94)[d]	37.54 (8.53)[d]	85.8 (0.799 - 0.901)	5.30 (2.40)[d]	33.9 (0.274 - 0.410)
Professor	12.61 (2.17)	83.5 (0.749 - 0.896)	13.45 (2.87)[d]	37.46 (8.32)[d]	79.4 (0.703 - 0.862)	5.08 (1.94)	38.1 (0.291 - 0.484)
Years engaged in ultrasound							
1 - 3	11.88 (2.51)	72.5[b] (0.661 - 0.782)	11.84 (3.37)	31.32 (11.23)	74.5[b] (0.681 - 0.800)	4.40 (2.29)	32.4 (0.263 - 0.390)
4 - 5	12.74 (2.54)[e]	79.8 (0.703 - 0.868)	13.34 (3.20)[e]	35.12 (9.14)[e]	74.2 (0.642 - 0.821)	5.00 (2.32)	29.2 (0.208 - 0.394)
>5	12.93 (2.37)[e]	88.1 (0.838 - 0.914)	13.48 (2.97)[e]	37.09 (8.59)[e]	83.2 (0.785 - 0.871)	5.15 (2.24)[e]	33.6 (0.284 - 0.392)
Independent CEUS experience before training?							
Yes	12.79 (2.20)	83.6 (0.766 - 0.888)	13.58 (2.93)	37.14 (7.93)	76.4 (0.688 - 0.827)	5.02 (1.96)	32.9 (0.256 - 0.410)
No	12.44 (2.57)	80.6 (0.767 - 0.844)	12.66 (3.30)[f]	34.00 (10.49)[f]	79.5 (0.755 - 0.830)	4.80 (2.39)	32.3 (0.281 - 0.369)
Duration of CEUS training (mo)							
1 - 3	11.93 (2.49)	77.4[b] (0.727 - 0.816)	12.00 (3.20)	32.29 (10.63)	75.7[b] (0.708 - 0.799)	4.46 (2.24)	32.1 (0.273 - 0.372)
4 - 6	13.09 (1.78)[g]	93.1 (0.870 - 0.965)	14.13 (2.36)[g]	38.84 (6.35)[g]	86.2 (0.788 - 0.913)	5.59 (2.10)[g]	40.5 (0.320 - 0.496)
>6	13.60 (2.59)[g]	81.0 (0.732 - 0.869)	14.08 (3.30)[g]	37.60 (9.10)[g]	80.2 (0.724 - 0.862)	5.25 (2.38)[g]	26.2 (0.193 - 0.345)
Total	12.53 (2.50)	81.3 (0.780 - 0.8431)	12.88 (3.24)	34.76 (10.01)	78.7 (0.752 - 0.819)	4.86 (2.29)	32.5 (0.288 - 0.364)

[a]$P<.05$ compared with male participants.

[b]$P<.05$.

[c]$P<.05$ compared with participants with a bachelor's degree.

XSL•FO
RenderX

[d]$P<.05$ compared with participants who were assistant professors.

[e]$P<.05$ compared with participants who engaged in ultrasound for 1-3 years.

[f]$P<.05$ compared with participants without prior independent CEUS examination experience.

[g]$P<.05$ compared with participants who received CEUS training for 1-3 months.

Discussion

Principal Results

This study developed and validated the first organ-specific competency assessment tool for liver CEUS. Following a 3-round Delphi process and cognitive interviews, the CCAL was tested in internal and external multicenter cohorts. It demonstrated strong content validity, high internal consistency reliability, and robust construct and discriminant validity. Subdomain analysis revealed a pronounced competency gap in image-guided interventions, confirming the tool's ability to identify specific educational needs beyond global pass or fail outcomes.

Interpretations, Implications, and Comparison With Prior Work

The CCAL advances the field of CEUS competency assessment through several innovations that directly address the limitations of existing tools. Jacobsen et al [15,16] laid important groundwork by developing a core CEUS syllabus and operationalizing it into a 47-item multiple-choice test; however, their instrument was explicitly "not recommended for certification purposes" due to the risk of false positives and negatives, and its utility was limited to entry-level screening. The same research group later developed an immersive virtual reality simulation-based test focusing on precontrast preparations, yet it carried a 23.6% risk of false negatives and did not address comprehensive liver CEUS competency [17]. More broadly, other sonographic assessments, such as the OSAUS or generic radiology in-training examinations, assess general scanning skills or factual knowledge but lack the organ-specific, multidimensional construct needed to capture the unique demands of liver CEUS, including contrast-agent safety, dynamic enhancement patterns, and image-guided interventional decision-making [12,35,36]. These tools, while pioneering, were not organ-specific and lacked a self-assessment component, leaving an unmet need for a multidimensional, liver-dedicated assessment.

The CCAL addresses this gap through 5 defining strengths. With respect to content validity and theoretical foundation, item generation followed the 2020 WFUMB liver CEUS guidelines and was refined through a 3-round Delphi process involving 21 experts, which was followed by cognitive interviews, establishing the CCAL as the first liver-specific, expert-consensus-driven CEUS competency tool [2,37]. This expert-consensus approach stands in marked contrast to assessments developed in isolation by a single institution or those lacking systematic, evidence-informed item vetting. The high internal consistency and a 3-factor structure consistent with the theoretical blueprint confirm that the CCAL measures a coherent yet multidimensional construct—encompassing theoretical knowledge, image interpretation, and clinical judgment—rather than a single global ability. In comparison, prior CEUS tests either reported no reliability data or were confined to narrow procedural tasks [15-17]. The demonstration of known-group discriminant validity, with scores distinguishing trainees by experience and professional title, supports the tool's use for benchmarking progress along a developmental continuum, a critical requirement for competency-based education [38,39]. Methodologically, the CCAL used a 4-phase mixed methods design and is the first tool to integrate objective knowledge testing with structured self-evaluation within a single instrument, enabling multidimensional assessment of theoretical knowledge, self-awareness, and clinical decision-making rather than a unidimensional pass or fail judgment [40,41]. Finally, the online platform with remote proctoring and automated scoring ensures that the CCAL can be deployed at low cost across diverse clinical settings, lowering the barriers to standardized assessment in resource-limited environments [42].

Beyond global pass-or-fail decisions, the CCAL offers granular diagnostic capability through subdomain analysis. Pass rates were high in fundamental areas such as indications and contraindications but dropped substantially for advanced competencies, like image-guided navigation for biopsy and tumor ablation. This finding aligns with published training challenges in interventional ultrasound, where the cognitive load of integrating real-time imaging with needle manipulation is high and structured curricula are often lacking [43]. The ability to pinpoint such a specific competency gap enables program directors to target educational interventions, such as simulation-based training modules or dedicated rotations, rather than applying blanket remediation [44,45]. This function aligns with the principles of assessment for learning, whereby diagnostic information directly informs curricular improvement [46].

The incorporation of structured self-evaluation enhances the comprehensiveness and validity of the assessment. A statistically significant positive correlation between objective assessment and self-evaluation scores indicates that trainees demonstrate reasonable metacognitive awareness of their clinical competence. However, the presence of notable outliers—individuals who consistently overestimate or underestimate their performance by a clinically meaningful margin—raises important patient safety concerns and warrants targeted remediation [46,47]. Overestimation in image-guided procedures could lead to premature independent practice and an increased risk of complications [48]. Conversely, underestimation may unnecessarily erode confidence and limit procedural engagement [49]. By flagging such discrepancies, the CCAL enables supervisors to provide targeted feedback that fosters accurate self-assessment—a skill increasingly recognized as essential for lifelong learning and safe unsupervised practice [50,51]. This dual-modality reporting thus moves beyond traditional knowledge tests that offer no window into trainees' metacognitive awareness [52].

XSL•FO
RenderX

Collectively, these characteristics position the CCAL as a practical tool for implementing CBME in CEUS training. Standardized, multisource data on cognitive competence can inform entrustment decisions, document training progression for accreditation, and facilitate cross-institutional quality assurance, thereby addressing the persistent reliance on subjective, uncalibrated assessments that currently dominate the field. While the CCAL explicitly does not assess hands-on procedural skills and should not be used in isolation for high-stakes credentialing, within its intended scope, it provides a transparent, evidence-based foundation for evaluating and advancing the cognitive component of liver CEUS competence.

Limitations

This study had the following limitations. First, the validity evidence chain remains incomplete because of the absence of predictive validity verification—specifically, the correlation between test scores and real-time clinical decisions during procedures (eg, video-based assessments or supervisor ratings [53,54]). Second, easy items predominate (66% of the objective section), which may limit discrimination among advanced practitioners, although this was intentional given the target population (novices to competent operators) [55]. Third, standardization frameworks require further development as follows: norm-referenced benchmarks (eg, 60% as the passing threshold for first-year residents) remain unestablished, and institutional variations necessitate multilevel modeling to delineate confounding factors (eg, trainer expertise and equipment availability) [56]. Fourth, the CCAL was validated only in Chinese trainees across 25 Chinese hospitals; generalizability to non-Chinese populations (different curricula, certification, patient populations, and cultural self-assessment tendencies) remains unproven, requiring international validation [57,58]. Fifth, the relationship between self-perceived competency in the SES and objectively measured clinical performance needs further study, as we did not include direct observation (eg, live scanning or video-recorded exams)—the gold standard for criterion validity [41,59]. Future research should examine concurrent validity against supervisor ratings or objective structured assessment of technical skills (OSATS) during actual patient encounters [60]. Collectively, these advantages—content validity grounded in expert consensus, high reliability, robust construct validity, multicenter validation, a multidimensional construct, and fine-grained diagnostic capacity—constitute clear innovations over prior tools, while the acknowledged limitations delineate a focused agenda for future research.

Conclusions

The CCAL is the first organ-specific competency assessment tool for liver CEUS, innovatively combining objective knowledge testing with structured self-evaluation. Unlike generic or simulation-based CEUS tests, it is liver-specific and validated across a large multicenter cohort. It brings a standardized, refined framework that identifies specific competency gaps. Practically, the tool supports CBME by enabling data-driven assessment, enhancing rigor in CEUS examinations, guiding targeted curricular improvements, permitting cross-institutional benchmarking, and supplying evidence for specialty societies to update training standards, though further validation is required before high-stakes use. It must be emphasized that the CCAL evaluates theoretical competence exclusively and should not replace procedural assessment for certification decisions. Within its intended scope, the CCAL advances liver CEUS training toward greater standardization and evidence-based feedback. International validation and the establishment of criterion-related and predictive validity are essential next steps.

Acknowledgments

The authors thank Ping Chen, Chunlin Tang, Xia Zhang, Guangliang Huang, Dong Chen, Zhirui Zhao, Yingyu Cai, Junluo, Qing Zhou, Yanzhou, Minsheng Zhang, Yuemingming Jiang, Xiaofeng Wu, Yan Zhou, QinJiang, Ruili Wang, Quan Zheng, Yangyang Zhu, Yuxiang Wang, Dengke Teng, Tengfei Yu, Chao Zhang, Shengmin Zhang, and Wei Zhang for distributing links to the assessment during the internal and external validation processes. The authors also thank Lili Wu, Tao Wu, Yan Lv, Lei Tan, and Jieyang Jin for their valuable contributions to the literature review and item development.

The authors declare the use of generative AI (GenAI) during the research and writing process. According to the GenAI delegation taxonomy (GAIDeT; 2025), the following tasks were delegated to GenAI tools under full human supervision: literature search and systematization, proofreading and editing, text summarization, and reformatting. The GenAI tool used was DeepSeek (version 3.2; Hangzhou DeepSeek Artificial Intelligence Basic Technology Research Co, Ltd). The authors take full responsibility for the final manuscript. GenAI tools are not listed as authors and bear no responsibility for the final content. This declaration was submitted by YM.

Funding

This study has received funding by the National Key Research and Development Programme of the Ministry of Science and Technology (2023YFFO715305); the Rural and Social Development Science and Technology Major Programme of Guangzhou (2024B03J0154); the High-tech and Major Technology Projects of Guangzhou (2024P-GX25); the 5010 Clinical Research Project of Sun Yat-sen University (2022010); and the Municipal Schools (Institutes) Joint Funding Project Basic Research Programme encompasses the Municipal Key Laboratory Construction Project (2024A03J0284).

XSL•FO
RenderX

Data Availability

Data generated or analyzed during the study are available from the corresponding author upon reasonable request.

Authors' Contributions

YM and BZ are co-first authors
Conceptualization: WH (lead), JR (equal)
Data curation: HW, GX, FN, YG, Yukun Luo, HZ, XW, XJ, RZ, JY, RW, XL, FD, Yan Luo, QL, JL, PH, QC, JZ, YX, XX
Formal analysis: LS, YM
Funding acquisition: JR
Investigation: YM, JR
Methodology: YM, LS
Project administration: WH (lead), JR (equal)
Supervision: WH, JR
Writing – original draft: YM (lead), BZ (supporting)
Writing – review & editing: JR (lead), WH (supporting)

Conflicts of Interest

None declared.

Multimedia Appendix 1
Results of the 3-round Delphi expert consultation for content validation of the Contrast-Enhanced Ultrasound Competency Assessment for the Liver.
[DOCX File, 14 KB - mededu_v12i1e86730_app1.docx]

Multimedia Appendix 2
Demographic and clinical characteristics of trainees in the internal validation phase of the Contrast-Enhanced Ultrasound Competency Assessment for the Liver development study.
[DOCX File, 15 KB - mededu_v12i1e86730_app2.docx]

Multimedia Appendix 3
Univariate and multivariate analyses of factors associated with self-evaluation section (SES) scores in the external validation cohort.
[DOCX File, 18 KB - mededu_v12i1e86730_app3.docx]

Multimedia Appendix 4
Factors associated with objective question section (OQS) scores in the external validation cohort.
[DOCX File, 18 KB - mededu_v12i1e86730_app4.docx]

References

1. Lyshchik A, Kuon Yeng Escalante C, Siu Xiao T, et al. Contrast-enhanced US of high-risk indeterminate focal liver observations categorized as LR-4 or LR-M at CT/MRI. Radiology 2025 Jan;314(1):e240916. [doi: 10.1148/radiol.240916] [Medline: 39835975]
2. Dietrich CF, Nolsøe CP, Barr RG, et al. Guidelines and good clinical practice recommendations for contrast enhanced ultrasound (CEUS) in the liver - update 2020 - WFUMB in cooperation with EFSUMB, AFSUMB, AIUM, and FLAUS. Ultraschall Med 2020 Oct;41(5):562-585. [doi: 10.1055/a-1177-0530] [Medline: 32707595]
3. Sidhu PS, Cantisani V, Dietrich CF, et al. The EFSUMB guidelines and recommendations for the clinical practice of contrast-enhanced ultrasound (CEUS) in non-hepatic applications: update 2017 (long version). Ultraschall Med 2018 Apr;39(2):e2-e44. [doi: 10.1055/a-0586-1107] [Medline: 29510439]
4. Jenssen C, Gilja OH, Serra AL, et al. European Federation of Societies for Ultrasound in Medicine and Biology (EFSUMB) policy document development strategy: clinical practice guidelines, position statements, and technological reviews. Ultrasound Int Open 2019 Jan;5(1):E2-E10. [doi: 10.1055/a-0770-3965] [Medline: 30599040]
5. Ying X, Dong S, Zhao Y, Chen Z, Jiang J, Shi H. Research progress on contrast-enhanced ultrasound (CEUS) assisted diagnosis and treatment in liver-related diseases. Int J Med Sci 2025;22(5):1092-1108. [doi: 10.7150/ijms.101789] [Medline: 40027182]
6. Frank JR, Snell LS, Cate OT, et al. Competency-based medical education: theory to practice. Med Teach 2010;32(8):638-645. [doi: 10.3109/0142159X.2010.501190] [Medline: 20662574]

XSL•FO
RenderX

7. Holmboe ES, Sherbino J, Long DM, Swing SR, Frank JR. The role of assessment in competency-based medical education. Med Teach 2010 Aug;32(8):676-682. [doi: 10.3109/0142159X.2010.500704]
8. Cook DA, Kuper A, Hatala R, Ginsburg S. When assessment data are words: validity evidence for qualitative educational assessments. Acad Med 2016;91(10):1359-1369. [doi: 10.1097/ACM.0000000000001175]
9. Carrillo-Avalos BA, Leenen I, Trejo-Mejía JA, Sánchez-Mendiola M. Bridging validity frameworks in assessment: beyond traditional approaches in health professions education. Teach Learn Med 2025;37(2):229-238. [doi: 10.1080/10401334.2023.2293871] [Medline: 38108266]
10. Alharbi NS. Evaluating competency-based medical education: a systematized review of current practices. BMC Med Educ 2024 Jun 3;24(1):612. [doi: 10.1186/s12909-024-05609-6] [Medline: 38831271]
11. Gupta SK, Srivastava T. Assessment in undergraduate competency-based medical education: a systematic review. Cureus 2024 Apr;16(4):e58073. [doi: 10.7759/cureus.58073] [Medline: 38738047]
12. Reid S, Goffi A, Tsou E, et al. ULTRA-Metrics: Delphi-derived framework for assessing ultrasound competency. J Ultrasound Med 2026 Feb;45(2):383-400. [doi: 10.1002/jum.70074] [Medline: 41054916]
13. Haidar DA, Peterson WJ, Minges PG, et al. A consensus list of ultrasound competencies for graduating emergency medicine residents. AEM Educ Train 2022 Dec;6(6):e10817. [doi: 10.1002/aet2.10817] [Medline: 36425790]
14. Mizubuti GB, Maxwell S, Shatenko S, et al. Competencies for proficiency in basic point-of-care ultrasound in anesthesiology: national expert recommendations using Delphi methodology. Can J Anesth 2024 Jul;71(7):967-977. [doi: 10.1007/s12630-024-02746-w]
15. Jacobsen N, Nolsøe CP, Konge L, et al. Development of and gathering validity evidence for a theoretical test in contrast-enhanced ultrasound. Ultrasound Med Biol 2022 Feb;48(2):248-256. [doi: 10.1016/j.ultrasmedbio.2021.10.016] [Medline: 34815128]
16. Jacobsen N, Nolsøe CP, Konge L, et al. Contrast-enhanced ultrasound: development of syllabus for core theoretical and practical competencies. Ultrasound Med Biol 2020 Sep;46(9):2287-2292. [doi: 10.1016/j.ultrasmedbio.2020.05.011] [Medline: 32553690]
17. Jacobsen N, Larsen JD, Falster C, et al. Using immersive virtual reality simulation to ensure competence in contrast-enhanced ultrasound. Ultrasound Med Biol 2022 May;48(5):912-923. [doi: 10.1016/j.ultrasmedbio.2022.01.015]
18. Ceccacci A, Hothi H, Khan R, et al. Validity evidence for endoscopic ultrasound competency assessment tools: systematic review. Endosc Int Open 2024 Dec;12(12):E1465-E1475. [doi: 10.1055/a-2465-7283] [Medline: 39691740]
19. Di Pietro S, Mascolo M, Falaschi F, et al. Lung-ultrasound objective structured assessment of technical skills (LUS-OSAUS): utility in the assessment of lung-ultrasound trained medical undergraduates. J Ultrasound 2021 Mar;24(1):57-65. [doi: 10.1007/s40477-020-00454-x] [Medline: 32266687]
20. Tanaanantarak P, Kritsaneepaiboon S, Charoonratana V, et al. Cranial ultrasound performance assessment tool: development and validation. Med Ultrason 2024 Mar 27;26(1):83-90. [doi: 10.11152/mu-4307] [Medline: 38150694]
21. Dai T, Zhu H, Qiao M, et al. Training on contrast-enhanced ultrasound LI-RADS classification for resident radiologists: a retrospective comparison of performance after training. Insights Imaging 2024 Aug 14;15(1):205. [doi: 10.1186/s13244-024-01786-6] [Medline: 39143424]
22. Cheung JJH, Chen EW, Darani R, McCartney CJL, Dubrowski A, Awad IT. The creation of an objective assessment tool for ultrasound-guided regional anesthesia using the Delphi method. Reg Anesth Pain Med 2012;37(3):329-333. [doi: 10.1097/AAP.0b013e318246f63c] [Medline: 22354106]
23. Bick JS, Wanderer JP, Myler CS, Shaw AD, McEvoy MD. Standard setting for clinical performance of basic perioperative transesophageal echocardiography: moving beyond the written test. Anesthesiology 2017 Apr;126(4):718-728. [doi: 10.1097/ALN.0000000000001538] [Medline: 28169844]
24. Cho J, Choi EK, Kim IR, et al. Development and validation of Chemotherapy-induced Alopecia Distress Scale (CADS) for breast cancer patients. Ann Oncol 2014 Feb;25(2):346-351. [doi: 10.1093/annonc/mdt476] [Medline: 24379161]
25. Bedell A, Taft TH, Keefer L, Pandolfino J. Development of the Northwestern Esophageal Quality of Life Scale: a hybrid measure for use across esophageal conditions. Am J Gastroenterol 2016 Apr;111(4):493-499. [doi: 10.1038/ajg.2016.20] [Medline: 26881974]
26. Adegbola SO, Dibley L, Sahnan K, et al. Development and initial psychometric validation of a patient-reported outcome measure for Crohn's perianal fistula: the Crohn's Anal Fistula Quality of Life (CAF-QoL) scale. Gut 2021 Sep;70(9):1649-1656. [doi: 10.1136/gutjnl-2019-320553] [Medline: 33272978]
27. Nair R, Aggarwal R, Khanna D. Methods of formal consensus in classification/diagnostic criteria and guideline development. Semin Arthritis Rheum 2011 Oct;41(2):95-105. [doi: 10.1016/j.semarthrit.2010.12.001]
28. Phalswal U, Kalia R. A systematic approach to the development and validation process of a research tool: an overview. Int J Community Med Public Health 2026;13(2):1065-1073. [doi: 10.18203/2394-6040.ijcmph20260329]
29. Claessens F, Van der Auwera C, Seys D, et al. A multiphase, multicentre development and validation of two maturity tools assessing the implementation of the FlaQuM co-creation roadmap. Int J Qual Health Care 2024 Apr 30;36(2):mzae035. [doi: 10.1093/intqhc/mzae035]

XSL•FO
RenderX

30. Chauhan GR, Chauhan BR, Vaza JV, Chauhan PR. Relations of the number of functioning distractors with the item difficulty index and the item discrimination power in the multiple choice questions. Cureus 2023 Jul;15(7):e42492. [doi: 10.7759/cureus.42492] [Medline: 37644928]
31. Rezigalla AA, Eleragi AMESA, Elhussein AB, et al. Item analysis: the impact of distractor efficiency on the difficulty index and discrimination power of multiple-choice items. BMC Med Educ 2024;24(1):445. [doi: 10.1186/s12909-024-05433-y]
32. Shahat KA. Item analysis of multiple-choice question (MCQ)-based exam efficiency among postgraduate pediatric medical students: an observational, cross-sectional study from Saudi Arabia. Cureus 2024 Sep;16(9):e69151. [doi: 10.7759/cureus.69151] [Medline: 39262935]
33. Kim YH, Kim BH, Kim J, Jung B, Bae S. Item difficulty index, discrimination index, and reliability of the 26 health professions licensing examinations in 2024, Korea: a psychometric study. J Educ Eval Health Prof 2025;22:38. [doi: 10.3352/jeehp.2025.22.38]
34. Belay LM, Siebeck M, Fischer MR, Berndt M, Sendekie TY. Psychometric evaluation of the Ethiopian National Medical Licensing Examination: evidence from classical and item response theory analyses. BMC Med Educ 2026;26(1):150. [doi: 10.1186/s12909-026-08568-2]
35. Tolsgaard MG, Todsen T, Sorensen JL, et al. International multispecialty consensus on how to evaluate ultrasound competence: a Delphi consensus survey. PLoS ONE 2013;8(2):e57687. [doi: 10.1371/journal.pone.0057687] [Medline: 23469051]
36. Gomes SH, Trindade M, Petrisor C, et al. Objective structured assessment ultrasound skill scale for hyomental distance competence - psychometric study. BMC Med Educ 2023 Mar 22;23(1):177. [doi: 10.1186/s12909-023-04146-y] [Medline: 36949512]
37. Artino AR, La Rochelle JS, Dezee KJ, Gehlbach H. Developing questionnaires for educational research: AMEE Guide No. 87. Med Teach 2014 Jun;36(6):463-474. [doi: 10.3109/0142159X.2014.889814] [Medline: 24661014]
38. Zhao Y, Zhou P, Zhu W, et al. Validity evidence for simulator-based obstetric ultrasound competency assessment tool: a multi-center study. Ultraschall Med 2024 Apr;45(2):168-175. [doi: 10.1055/a-2122-6746] [Medline: 37402406]
39. Teslak KE, Post JH, Tolsgaard MG, Rasmussen S, Purup MM, Friis ML. Simulation-based assessment of upper abdominal ultrasound skills. BMC Med Educ 2024 Jan 3;24(1):15. [doi: 10.1186/s12909-023-05018-1] [Medline: 38172820]
40. Schifferdecker KE, Reed VA. Using mixed methods research in medical education: basic guidelines for researchers. Med Educ 2009 Jul;43(7):637-644. [doi: 10.1111/j.1365-2923.2009.03386.x] [Medline: 19573186]
41. Fraundorf SH, Caddick ZA, Nokes-Malach TJ, Rottman BM. Cognitive perspectives on maintaining physicians' medical expertise: III. Strengths and weaknesses of self-assessment. Cogn Res Princ Implic 2023 Aug 30;8(1):58. [doi: 10.1186/s41235-023-00511-z] [Medline: 37646932]
42. Ellaway R, Masters K. AMEE Guide 32: e-Learning in medical education Part 1: learning, teaching and assessment. Med Teach 2008 Jun;30(5):455-473. [doi: 10.1080/01421590802108331] [Medline: 18576185]
43. Liao SC, Shao SC, Gao SY, Lai ECC. Augmented reality visualization for ultrasound-guided interventions: a pilot randomized crossover trial to assess trainee performance and cognitive load. BMC Med Educ 2024 Sep 27;24(1):1058. [doi: 10.1186/s12909-024-05998-8] [Medline: 39334275]
44. Ramachandran R, Sharma K, Gupta N, Nagarajappa A, Kumar S, Roy A. A multimodal blended learning curriculum for ultrasound-guided diagnostics and needle navigation: development and early implementation in an Indian tertiary care setting. Front Med 2026;13:1697471. [doi: 10.3389/fmed.2026.1697471]
45. Nitsche JF, Gu C, Packard AT, Brost BC. Assessment of a task trainer and targeting tasks for ultrasound-guided invasive procedures in radiology residents. Acad Radiol 2025 Aug;32(8):4939-4946. [doi: 10.1016/j.acra.2025.04.056] [Medline: 40480936]
46. Neubauer R, Recker F, Weimer J, et al. Verifying learning success: assessment and certification of student ultrasound education. BMC Med Educ 2025 Jun 19;25(1):851. [doi: 10.1186/s12909-025-07469-0] [Medline: 40537759]
47. Deveze E, Traore A, Ribault N, et al. Self-assessment versus peer-assessment in microsurgery learning: a comparative retrospective study in a surgery residents cohort. J Surg Educ 2023 Oct;80(10):1472-1478. [doi: 10.1016/j.jsurg.2023.06.028] [Medline: 37524617]
48. Al-Maghrabi M, Mamede S, Schmidt HG, et al. Overconfidence, time-on-task, and medical errors: is there a relationship? Adv Med Educ Pract 2024;15:133-140. [doi: 10.2147/AMEP.S442689] [Medline: 38410282]
49. Tzamaras H, Sinz E, Yang M, Ng P, Moore J, Miller S. Competence over confidence: uncovering lower self-efficacy for women residents during central venous catheterization training. BMC Med Educ 2024 Aug 26;24(1):923. [doi: 10.1186/s12909-024-05747-x] [Medline: 39187825]
50. Zheng B, He Q, Lei J. Informing factors and outcomes of self-assessment practices in medical education: a systematic review. Ann Med 2024 Dec;56(1):2421441. [doi: 10.1080/07853890.2024.2421441] [Medline: 39460558]
51. Merkebu J, Veen M, Hosseini S, Varpio L. The case for metacognitive reflection: a theory integrative review with implications for medical education. Adv Health Sci Educ Theory Pract 2024 Sep;29(4):1481-1500. [doi: 10.1007/s10459-023-10310-2] [Medline: 38345690]

XSL•FO
RenderX

52. Yates N, Gough S, Brazil V. Self-assessment: with all its limitations, why are we still measuring and teaching it? Lessons from a scoping review. Med Teach 2022 Nov;44(11):1296-1302. [doi: 10.1080/0142159X.2022.2093704] [Medline: 35786121]
53. Downing SM. Validity: on meaningful interpretation of assessment data. Med Educ 2003 Sep;37(9):830-837. [doi: 10.1046/j.1365-2923.2003.01594.x] [Medline: 14506816]
54. Sinha A, Nimbalkar SM, Pujara RK, et al. SimCapture app video performance assessment versus real-time instructor-based performance evaluation of undergraduates in neonatal resuscitation—an agreement study. J Trop Pediatr 2024 Oct 4;70(6):fmae033. [doi: 10.1093/tropej/fmae033] [Medline: 39366743]
55. Sim SM, Rasiah RI. Relationship between item difficulty and discrimination indices in true/false-type multiple choice questions of a para-clinical multidisciplinary paper. Ann Acad Med Singap 2006 Feb;35(2):67-71. [Medline: 16565756]
56. Matsuyama Y, Nomura O, Oikawa S, Kikukawa M, Shimizu I, Gomi H. Competency-based medical education guidelines are context-based: lessons from national guidelines in five countries. Med Teach 2024 Sep;46(sup1):S38-S45. [doi: 10.1080/0142159X.2024.2351215] [Medline: 39545497]
57. Chen LY, Lue YJ, Wu PH, Kuo YL. Cross-cultural adaptation and validation of the Chinese version of the modified Fresno test for physical therapists. BMC Med Educ 2025 Jan 2;25(1):1. [doi: 10.1186/s12909-024-06615-4] [Medline: 39743510]
58. Andrés IR, Natalia BR, Pablo BH, Daniel AS, Milciades IP, John V. Transcultural adaptation of the virtual patient integration rating scale: a factor analysis. BMC Med Educ 2025 Mar 18;25(1):396. [doi: 10.1186/s12909-024-06571-z] [Medline: 40102821]
59. Gaeta TJ, Reisdorff E, Barton M, et al. The Dunning Kruger effect in resident predicted and actual performance on the American Board of Emergency Medicine in-training examination. J Am Coll Emerg Physicians Open 2024 Oct;5(5):e13305. [doi: 10.1002/emp2.13305] [Medline: 39463809]
60. Landreville JM, Wood TJ, Frank JR, Cheung WJ. Does direct observation influence the quality of workplace-based assessment documentation? AEM Educ Train 2022 Aug;6(4):e10781. [doi: 10.1002/aet2.10781] [Medline: 35903424]

Abbreviations

CBME: competency-based medical education
CCAL: Contrast-Enhanced Ultrasound Competency Assessment for the Liver
CEUS: contrast-enhanced ultrasound
EFSUMB: European Federation of Societies for Ultrasound in Medicine and Biology
EUS: endoscopic ultrasound
EUSAT: Endoscopic Ultrasound Assessment Tool
GAPS-EUS: Global Assessment of Performance and Skills in Endoscopic Ultrasound
IRB: Institutional Review Board
KMO: Kaiser-Meyer-Olkin
LI-RADS: Liver Imaging Reporting and Data System
LUS: lung ultrasound
OQS: objective question section
OSATS: objective structured assessment of technical skills
OSAUS: objective structured assessment of ultrasound skills
SES: self-evaluation section
TEESAT: The Endoscopic Ultrasound and Endoscopic Retrograde Cholangiopancreatography Skills Assessment Tool
WFUMB: World Federation for Ultrasound in Medicine and Biology

Edited by S Brini; submitted 29.Oct.2025; peer-reviewed by CR Cervantes-Sanchez, W He; revised version received 29.Jun.2026; accepted 01.Jul.2026; published 09.Nov.2026.

Please cite as:
Ma Y, Zheng B, Ren J, He W, Shang L, Wang H, Xue G, Nie F, Guo Y, Luo Y, Zhang H, Wang X, Jing X, Zhang R, Yuan J, Wu R, Luo X, Dong F, Luo Y, Lu Q, Liao J, Huang P, Chen Q, Zhu J, Xu Y, Xie X
A Competency Assessment Tool for Liver Contrast-Enhanced Ultrasound Among Sonography Trainees: Development and Multicenter Validation Study
JMIR Med Educ 2026;12:e86730
URL: https://mededu.jmir.org/2026/1/e86730
doi:10.2196/86730

XSL•FO
RenderX

© Yanping Ma, Bowen Zheng, Jie Ren, Wen He, Lei Shang, Hui Wang, Gaiqin Xue, Fang Nie, Yanli Guo, Yukun Luo, Hongxia Zhang, Xuan Wang, Xiang Jing, Ruifang Zhang, Jianjun Yuan, Rong Wu, Xiaomao Luo, Fenglin Dong, Yan Luo, Qiang Lu, Jintang Liao, Pintong Huang, Qin Chen, Jia'an Zhu, Youfeng Xu, Xiaoyan Xie. Originally published in JMIR Medical Education (https://mededu.jmir.org), 9.Nov.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Education, is properly cited. The complete bibliographic information, a link to the original publication on https://mededu.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

The Design and Evaluation of an Online Continuing Medical Education App for Medical Professionals in China: Quantitative Study

Xu Zhang[1*], MSc; Xianying He[1*], MM; Yuntian Chu[1], PhD; Dongqing Liu[1], PhD; Minzhao Lyu[2], PhD; Weiyi Wang[1], PhD; Haotian Chen[1], MSc; Meihao Ji[1], MBA; Fangfang Cui[1,3], MM; Jie Zhao[1,3], PhD

[1]National Engineering Laboratory for Internet Medical Systems and Applications, The First Affiliated Hospital of Zhengzhou University, 1st Jianshe Road, Zhengzhou, China

[2]School of Electrical Engineering and Telecommunications, University of New South Wales, Sydney, Australia

[3]Institute of Intelligent Medicine, Henan Academy of Innovations in Medical Science, Zhengzhou, China

[*]these authors contributed equally

Corresponding Author:
Fangfang Cui, MM
National Engineering Laboratory for Internet Medical Systems and Applications, The First Affiliated Hospital of Zhengzhou University, 1st Jianshe Road, Zhengzhou, China

Abstract

Background: As an emerging delivery mode of education, online continuing medical education (CME) increases the accessibility of high-quality medical training for professionals and students in China. Guoyuan (meaning “nationwide” in Chinese) is an online CME platform delivered via a mobile app and operated by the National Telemedicine Center of China since 2018, serving as an illustrative case of mobile online CME implementation.

Objective: We identified trends in the adoption and usage of the Guoyuan mobile online CME platform from 2018 to 2023 and provided evidence for the application and optimization of online CME.

Methods: We analyzed yearly usage data of the Guoyuan mobile app (The First Affiliated Hospital of Zhengzhou University) in 2018-2023 and collected surveys on the satisfaction and recognition of competency enhancement in online CME in each connected hospital in 2023. Using the IBM SPSS, the nonparametric Kruskal-Wallis *H* test was used to compare attendance across different disciplines, followed by post hoc pairwise comparisons for course types with significant differences and ordinal logistic regression analysis to examine factors influencing satisfaction with the online CME system and perceived competency enhancement among invited doctors.

Results: From 2018 to 2023, Guoyuan had 94,537 registered trainees, 1672 published course videos, and 1,878,437 attendances. Attendance was higher for courses in ophthalmology, otolaryngology, and pathology than in other disciplines (median attendance 610, IQR 105-2055 vs 283, IQR 106-690 participants). Based on a sample size of 245 participants, ordinal regression analysis showed that discipline category, professional title, and working years significantly influenced satisfaction. General practitioners showed lower overall satisfaction than internal medicine doctors (odds ratio [OR] 0.323, 95% CI 0.110-0.948; OR 0.251, 95% CI 0.087-0.729; and OR 0.196, 95% CI 0.066-0.585; *P*=.04; *P*=.01*;* *P*=.003). Junior titles reported higher audio-visual clarity (OR 3.151, 95% CI 1.178-8.427; *P*=.02) and process satisfaction (OR 4.939, 95% CI 1.674-14.576; *P*=.004). More experienced doctors had higher system usability (OR 1.102, 95% CI 1.012-1.200; *P*=.03) and process satisfaction (OR 1.141, 95% CI 1.044-1.247; *P*=.003). Recognition of online CME’s benefits was influenced by multiple factors. Greater clinical experience positively predicted recognition of clinical use (OR 1.106, 95% CI 1.004-1.218; *P*=.04), while an inverse association was observed with age (OR 0.894, 95% CI 0.802-0.996; *P*=.04). For research-related benefits, positive predictors included discipline category in obstetrics and gynecology compared to internal medicine (OR 6.217, 95% CI 1.236-31.258; *P*=.03) and junior professional title (OR 3.791, 95% CI 1.231-11.673; *P*=.02), whereas intensive care unit was a negative predictor compared to internal medicine (OR 0.111, 95% CI 0.014-0.893; *P*=.04).

Conclusions: Online mobile CME platforms have gained widespread adoption among medical professionals in China, particularly after the COVID-19 outbreak. However, substantial disciplinary disparities in course availability and user experience persist, indicating the need for further optimization of course design and software interaction.

(JMIR Med Educ 2026;12:e76299) doi:10.2196/76299

XSL•FO
RenderX

KEYWORDS

online continuing medical education; mobile medical education; trend; satisfaction; ability enhancement

Introduction

The unbalanced distribution of medical resources is a critical issue that significantly impacts the overall well-being of citizens, particularly in low-income countries [1]. This problem is even more acute in China, which has a population of more than 1.4 billion, with approximately 491 million citizens residing in rural areas that are currently experiencing shortages of high-quality medical resources [2,3]. Although the Chinese government is actively seeking effective measures to increase the coverage of medical resources [4], the issue of unbalanced distribution remains unsolved. In 2017, the number of physicians per 1000 population in eastern, middle, and southern China was 4.3, 3.8, and 3.5, respectively [5,6]. Prior research [7] has used the health resource agglomeration degree to describe the qualification of medical professionals. In 2018, the health resource agglomeration degree value for urban regions in China was 1.60 (sufficient), while the value for rural regions in China was 0.68 (severely insufficient). Furthermore, the annual growth rate for the number of medical professionals in rural regions is much lower than that in urban regions, making the problem increasingly severe over the years [8-10].

Continuing medical education (CME) is a critical component of medical education, aimed at enabling health care professionals to continuously acquire new knowledge and innovative technologies relevant to their specialties. It serves as a continuation, supplement, and enhancement of professional education, ultimately enhancing the technical competencies and service quality of medical practitioners. Without a program of active learning, no physician will be able to remain competent for more than a few years after graduation [11]. Common CME formats include lectures, conferences, printed materials, internet-based educational activities, and open discussions [12]. Online CME is one of the technologies aiming to provide accessible medical training and education for medical professionals at less-privileged hospitals in rural and remote regions without requiring them to physically attend classes held by leading institutes, which can be impractical due to financial and physical constraints [13]. As reported by prior research, online CME has gained popularity for medical education on an international scale [14-21]. During the COVID-19 pandemic, it has become a major form of medical education across many disciplines, such as anatomy and physiological classes in universities in Australia and New Zealand [17]. Interestingly, survey studies conducted in the United States and Australia [18,19] suggest that students and trainees in online anatomy and physiotherapy courses have reported better learning efficiency via online CME than in-person classes.

To the best of our knowledge, existing research primarily focuses on online learning among university students, while CME for health care professionals is mainly concentrated on discipline-specific training. In contrast, there is relatively little research on multidisciplinary and comprehensive CME for medical professionals. The increasing burden of chronic diseases in China has led to a growing demand for high-quality primary health care services. Online CME has a positive impact on physician performance, especially for primary care doctors. However, many primary health care professionals in China perceive a lack of online CME opportunities and have low participation rates, despite having high expectations [22]. A mobile online education model that enhances the accessibility of digital learning, facilitates large-scale implementation, and strengthens the capacity of primary health care systems urgently warrants broader promotion and application.

In this study, by introducing the Guoyuan platform (meaning “nationwide” in Chinese) as an illustrative example of a mobile online CME app serving hospitals in Henan, China’s most populous province, we propose a mobile online CME paradigm. Through this case, we demonstrate a model where educators disseminate knowledge across online channels, enabling learners to access educational content on demand by using fragmented time. By analyzing data from the Guoyuan platform spanning 2018 to 2023, together with questionnaire responses regarding online CME, this study seeks to investigate the platform’s curriculum characteristics and identify the key factors influencing doctors’ satisfaction and perceived competence improvement. Through these analyses, this study intends to provide a viable online CME model for developing regions and offer empirical evidence for the optimization of future online education programs, with the ultimate goal of reducing urban–rural disparities in health care resources and enhancing the professional competence of primary health care personnel.

Methods

The Guoyuan Mobile Online CME Platform

The Guoyuan mobile online CME platform is developed by the National Telemedicine Center of China and currently serves medical professionals in both premium tier-3 and less-capable tier-2 and tier‐1 hospitals primarily within Henan province (China is currently categorizing medical institutes into 3 tiers. In general, tier-3 hospitals, being the highest rank, are large-scale comprehensive or specialized medical centers at the top of the health care system; tier-2 hospitals function as regional medical centers focusing on common and frequently occurring diseases, while tier-1 hospitals are community health centers providing basic care, prevention, and rehabilitation services. The service coverage will be expanded to the entire nation in the next deployment phase of the platform. It delivers medical training and courses instructed by professors and highly qualified medical professionals in the tier-3 hospitals. Registered trainees in connected hospitals could access the online courses via an Android and iOS mobile app specifically designed for the platform. To motivate course instructors for high-quality content, enrollment fees are charged on a per-course basis, paid by the connected hospitals or individual trainees.

Within the client mobile app, after login and initialization, trainees could access a range of services, including their courses in progress, newly released courses under different disciplines, lecturer information, course materials, and publication database,

XSL•FO
RenderX

as visually shown in Figure 1. For each available course, in addition to scheduled live lectures that allow trainees to interact with course instructors in real time, lecture recordings and discussion forums are also available to trainees for after-class reviews and discussions. In addition to the core functions in online learning, a registered user could also have their friend circles socialize with colleagues in other medical institutes. Detailed descriptions of each available module on the mobile app are given in Multimedia Appendix 1.

Figure 1. Three example user pages of the Guoyuan mobile online continuing medical education app. CME: continuing medical education.



Data Collection

The dataset used in this study comprises the usage data of the regional online CME platform delivered via the Guoyuan mobile app over 6 years from 2018 to 2023. The scope of the dataset contains course metadata (eg, names, disciplines, instructors, sessions, enrollments, and attendance), connected hospitals, and registered medical professionals. In addition, to understand the satisfaction and recognition of competency enhancement in online CME, we created an online questionnaire. According to the sample size calculation formula as follows:

(1)n0=Zα/22 p(1−p)δ2 n=n01+n0−1N

Where n0 represents the initial sample size calculated for an infinite population, Zα/22 represents 95% CI, p represents an estimated population proportion of 0.5, and δ2 represents an acceptable margin of error of 0.1, N represents the total number of registered users on the Guoyuan app, and accounting for a 10% invalid response rate, the final calculated sample size was 108 participants. Based on the proportion of registered users by hospital level, the sample included 48 participants from tier-3 hospitals and 60 from tier 2 and tier 1 hospitals. The survey notice was distributed via the Guoyuan app to hospital administrators across the province, who were responsible for organizing local participation. The questionnaire was completed using WJX (an online survey software developed by a Chinese company). In total, we obtained 245 valid questionnaires, including 98 from tier 3 hospitals and 147 from tier 2 hospitals. The questionnaire encompassed 3 core sections: doctors' demographic information (eg, age, sex, professional title, education, and working experience), institutional profiles of their affiliated hospitals (eg, province, levels of hospital and which discipline the provider was from), and their satisfaction levels with online CME, alongside perceived competency enhancement through the online CME.

Statistical Analysis

The dataset we have collected consists of system logs that require data cleaning and preprocessing, which are performed using utilities in the Excel software (Microsoft Corp). Our analysis on the cleaned intermediary dataset leveraged SPSS software (version 25.0; IBM Corp). Absolute values, fractional

XSL•FO
RenderX

ratios, and statistical metrics such as median, mean values, and total values were used to describe the adoption of courses under different medical disciplines, their enrollment and attendance, and annual changes. In the line and bar charts, "yearly published courses," "average attendance," and "total attendance," indicate the number of newly released courses and their corresponding attendance for each year from 2018 to 2023. "Number of published courses," "median attendance per course," and "total attendance" shown in the table described the cumulative data from 2018 to 2023. The nonparametric Kruskal-Wallis *H* test was used to compare attendance across different disciplines, followed by post hoc pairwise comparisons for course types with significant differences. An ordinal logistic regression analysis was applied to examine the impact of hospital level, region, department classification, gender, and professional title [23] on satisfaction with the convenience of the online CME system, audio-video quality, and process. Furthermore, using data from inviting doctors, an ordinal logistic regression analysis was used to analyze the factors influencing the effectiveness of online CME in enhancing the skills of inviting doctors. The significance level was set at α=.05.

Ethical Considerations

This study was approved by the First Affiliated Hospital of Zhengzhou University (Ethical Code: 2023-KY-0488-002). The data from the Guoyuan app was limited to course-related statistics, such as the number of registrations and the number of course views, and the data collected from the questionnaire were limited to demographic information, such as gender and age. No sensitive or personally identifiable biological information was collected in any of the data. The doctors were aware that the consent information was uploaded to the National Telemedicine Center of China and agreed to the data analysis and utilization. Informed consent was obtained from all individual participants involved in the study. For secondary analyses of research data, we confirm that the original informed consent or institutional review board approval included provisions allowing for secondary analysis without additional consent from the participants. Participants in the study were not offered any form of compensation for their participation. No images of individual participants or users are included in the manuscript or supplementary material that would allow for identification.

Results

User Popularity

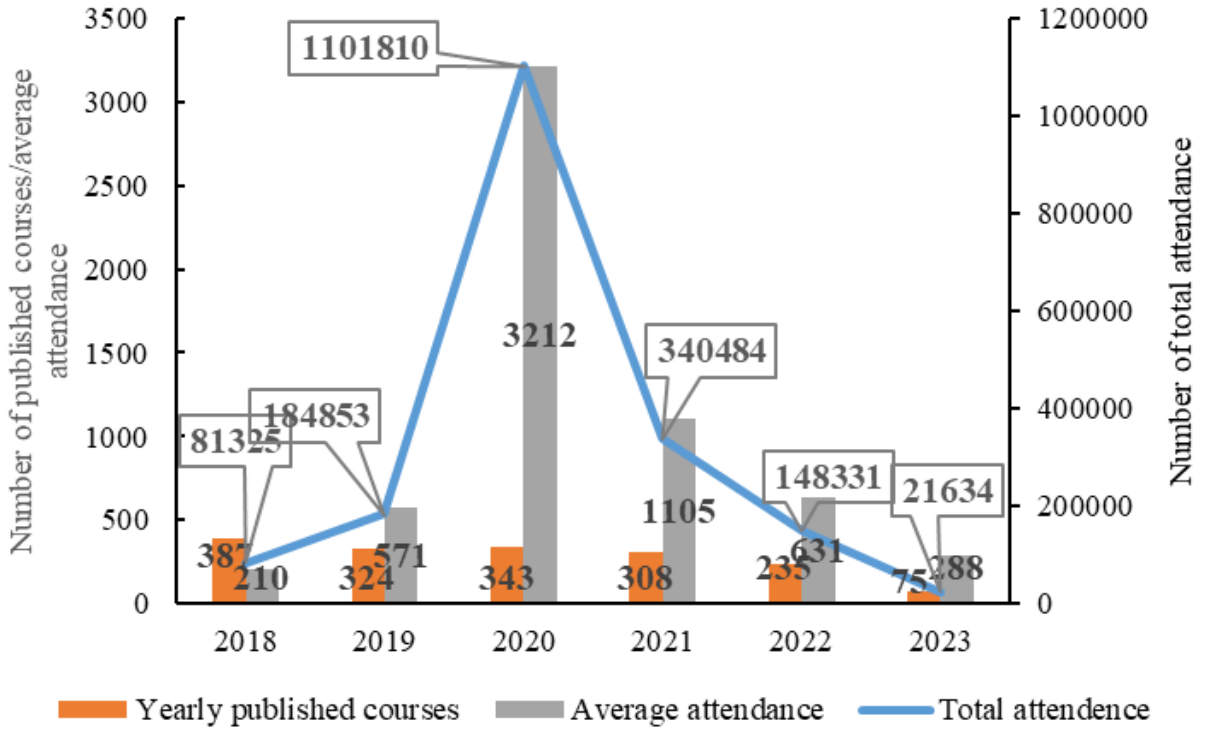
Guoyuan mobile online CME platform was launched in 2018, with cumulative registrations reaching 94,537 as of 2023. The annual registered users in 2018, 2019, 2020, and 2021 were 4902, 7879, 23,940, and 47,148 persons, respectively, demonstrating a marked growth trend with an average annual growth rate of 112.7%. The fastest growth occurred in 2020, with a year-on-year increase of 203.9% (23,940 in 2020 compared with 7879 in 2019), while the highest number of registrations was in 2021, reaching 47,148, with a year-on-year growth of 96.9% (47,148 in 2021 compared with 23,940 in 2020). From 2021 to 2023, the number of registrations showed a downward trend.

Available Courses and Attendance

During the period of 6 years from 2018 to 2023, a total of 1672 courses were published on the online CME platform. The total attendance for all courses during this period is 18,78,437 persons. The yearly numbers of published courses are depicted as the orange bar in Figure 2. In addition, the figure also shows the yearly course attendance represented by both aggregated count and averaged count per course, as annotated by the blue lines and gray bars, respectively. Notably, due to the COVID-19 outbreak, the total attendance in 2020 reached an abruptly high value of 1,101,810 persons, a 496.1% increase compared to 184,853 persons in 2019. From 2020 to 2023, the total attendance showed a downward trend, with the most significant decrease in 2023, which decreased by 85.4% (21,634 in 2023 compared with 148,331 in 2022). In general, the number of available courses remained roughly similar across years except for a significant decline in 2023. Both total attendance and average attendance peaked in 2020 before declining annually.

XSL•FO
RenderX

Figure 2. Number of available courses and their attendance.



We analyzed the availability of courses and their attendance under different medical disciplines (Table 1). Internal medicine was found to be the most dominant discipline in terms of total attendance, with cumulative viewership reaching 915,754. This is not surprising, as the discipline had 728 out of 1675 courses published, accounting for 43.5% of all courses on the online CME platform.

Table . Availability of courses and their attendance under different medical disciplines (H=93.182; P<.001).

Disciplines	Number of published courses	Attendance per course (Q_1 /Q_3), median (IQR)[a]	Total attendance
Pathology	105	283 (106-690)	98,247
Internal medicine	728	160 (36-914)[b,c]	915,754
Surgery	115	174 (48-712)[b,c]	102,911
Pediatrics	250	98 (26-804)[b,c,d]	121,186
Obstetrics and gynecology	121	94 (35-448)[b,c,d]	80,986
Ophthalmology and otolaryngology	103	610 (105-2055)	170,410
Traditional Chinese medicine	16	253 (180-940)	18,881
Pharmacy	71	70 (33-249)[b,c,d]	30,946
Nursing	88	104 (39-639)[b,c]	63,236
Others	78	407 (302-668)	282,964

[a]All comparisons were adjusted using the Bonferroni correction.

[b]Statistically significant difference compared to ophthalmology and otolaryngology.

[c]Statistically significant difference compared to others.

[d]Statistically significant difference compared to Pathology.

In terms of course availability, the second and the third popular disciplines were pediatrics and obstetrics and gynecology with their number of available courses as 250 out of 1675 (14.9%) and 121 out of 1675 (7.2%), respectively; whereas the second and the third popular disciplines ranked by their total attendance were ophthalmology and otolaryngology (170,410 views) and pediatrics (121,186 views). When comparing the median viewership across all disciplines, statistically significant differences were observed among disciplines (H=93.182, df=9; P<.001). Post hoc pairwise comparisons of course viewership

XSL•FO
RenderX

revealed that the median viewership for pathology was 283, which was higher than that of pediatrics, obstetrics and gynecology, and pharmacy (P=.009; P=.006; P=.003). The median viewership for ophthalmology and otolaryngology and other disciplines (eg, infection management, telemedicine, etc) was 610 and 407, respectively, which was higher than that of internal medicine, surgery, pediatrics, obstetrics and gynecology, pharmacy, and nursing (P<.001).

Online CME Satisfaction Evaluation

According to the evaluation from the survey, 233 out of 245 (95.1%) doctors, 224 out of 245 (91.4%) doctors, and 233 out of 245 (95.1%) doctors expressed satisfaction with system usability, audio and video clarity, service process, respectively. Using an ordinal regression model, we analyzed the impact of hospital level, region, professional title, age, discipline category, working years, telemedicine working years, and other variables on online CME satisfaction. The results (Table 2) showed that the discipline category was a factor influencing satisfaction across 3 dimensions. General practitioners had lower satisfaction in all 3 aspects of online CME compared to internal medicine doctors, with an odds ratio (OR) of 0.323 (95% CI 0.110-0.948), 0.251 (95% CI 0.087-0.729), and 0.196 (95% CI 0.066-0.585), respectively (P=.003).

XSL•FO
RenderX

Table . Ordinal logistic regression analysis results for satisfaction with online continuing medical education.

Characteristics	Satisfaction degree toward system usability		Satisfaction degree toward audio and video clarity		Satisfaction degree toward service process	
	OR (95% CI)[a]	*P* value	OR (95% CI)[a]	*P* value	OR (95% CI)[a]	*P* value
Hospital level (ref: tier-3)						
Tier-2 and tier - 1	1.897 (0.976-3.688)	.06	1.330 (0.691-2.560)	.39	1.356 (0.684-2.692)	.38
Region (ref: Western)						
Eastern	0.474 (0.188-1.195)	.11	1.041 (0.424-2.557)	.93	1.596 (0.611-4.164)	.34
Central	0.888 (0.434-1.816)	.75	1.107 (0.558-2.197)	.77	1.199 (0.584-2.463)	.62
Discipline (ref: Internal medicine)						
Others	0.511 (0.156-1.671)	.28	0.759 (0.234-2.462)	.65	1.033 (0.279-3.830)	.96
Surgery	0.833 (0.383-1.810)	.64	1.175 (0.554-2.492)	.67	1.341 (0.605-2.974)	.47
Pediatrics	0.903 (0.258-3.162)	.87	1.648 (0.453-6.001)	.50	1.652 (0.452-6.039)	.45
Obstetrics and gynecology	0.661 (0.218-2.005)	.47	0.632 (0.216-1.848)	.40	1.239 (0.372-4.126)	.73
Pathology	1.739 (0.433-6.979)	.44	2.245 (0.547-9.221)	.26	1.969 (0.489-7.936)	.34
General medicine	0.323 (0.110-0.948)	.04[b]	0.251 (0.087-0.729)	.01[b]	0.196 (0.066-0.585)	.003[b]
Traditional Chinese medicine	0.541 (0.096-3.056)	.49	0.506 (0.119-2.153)	.36	0.492 (0.095-2.545)	.40
Intensive care unit	0.151 (0.024-0.949)	.04[b]	1.645 (0.232-11.646)	.62	1.306 (0.176-9.663)	.80
Sex (ref: Female)						
Male	0.795 (0.440-1.435)	.48	0.671 (0.379-1.187)	.17	0.782 (0.429-1.425)	.42
Title (ref: Senior)						
Junior	2.619 (0.937-7.322)	.07	3.151 (1.178-8.427)	.02[b]	4.939 (1.674-14.576)	.004[b]
Intermediate	1.250 (0.574,2.724)	.57	1.940 (0.899-4.187)	.09	2.247 (0.994-5.080)	.05
Education (ref: PhD)						
Associate or below	0.586 (0.126-2.727)	.50	1.329 (0.308-5.730)	.70	1.583 (0.331-7.564)	.57
Bachelor	0.666 (0.188-2.359)	.53	1.098 (0.333-3.617)	.88	1.330 (0.371-4.770)	.66
Master	1.109 (0.311-3.960)	.87	1.489 (0.451-4.921)	.51	1.877 (0.518-6.800)	.34
Age	0.927 (0.848-1.014)	.10	0.931 (0.855-1.013)	.10	0.901 (0.823-0.987)	.03[b]
Working years	1.102 (1.012-1.200)	.03[b]	1.085 (0.999-1.177)	.05	1.141 (1.044-1.247)	.003[b]
Telemedicine working years	1.001 (0.937-1.068)	.10	1.013 (0.949-1.081)	.70	0.998 (0.931-1.070)	.97

[a]OR: odds ratio.

[b]*P* values <.05 are statistically significant.

In terms of system usability, working years and discipline category were factors affecting satisfaction. Doctors with more years of experience tended to have higher satisfaction with system usability, with OR 1.102 (95% CI 1.012-1.200; *P*=.03). Intensive care unit doctors had lower satisfaction with system operations compared to internal medicine doctors (OR 0.151, 95% CI 0.024-0.949; *P*=.04). Regarding audio-video clarity, professional title was a factor influencing satisfaction. Compared to senior titles, junior titles had higher satisfaction with the clarity of audio and video (OR 3.151, 95% CI 1.178-8.427; *P*=.02). For the service process, professional title, age, and working years were factors affecting satisfaction. Junior-level doctors reported higher satisfaction with the process than senior-level doctors, with OR 4.939 (95% CI 1.674-14.574; *P*=.004). Older doctors tended to be less satisfied with the process, with OR 0.901 (95% CI 0.823-0.987; *P*=.03). Those with more years of experience reported higher satisfaction, with OR 1.141 (95% CI 1.044-1.247; *P*<.003).

XSL•FO
RenderX

Recognition of the Ability Enhancement Through Online CME

We collected evaluations from invited doctors on the recognition for improved diagnostic and treatment capabilities and recognition for enhanced research capabilities. The results (Table 3) showed that 236 out of 245 doctors (96.3%) recognized that online CME can enhance their clinical diagnosis and treatment abilities, while 213 out of 245 doctors (86.9%) believed it can improve their research capabilities. Using an ordinal regression model, we analyzed the impact of hospital level, professional title, region, discipline category, age, education level, and other variables on recognition of the ability enhancement through online CME. Regarding the recognition of online CME in enhancing clinical diagnosis and treatment capabilities, age and years of experience were influencing factors. Clinical experience, measured by working years, was a significant positive factor in recognizing the clinical benefits of online CME (OR 1.106, 95% CI 1.004-1.218; *P*=.04). An independent association with age was also observed (OR 0.894, 95% CI 0.802-0.996; *P*=.04). For the recognition of online CME in improving research capabilities, discipline category and professional title were key factors. Compared to internal medicine, obstetrics and gynecology doctors were more likely to recognize the research benefits of online CME, while intensive care unit doctors were less likely to do so, with an OR of 6.217 (95% CI 1.236-31.258; *P*=.03) and 0.111 (95% CI 0.014-0.893; *P*=.04) respectively. Additionally, junior-level doctors were more likely than senior-level doctors to believe that online CME can enhance research capabilities (OR 3.791, 95% CI 1.231-11.673; *P*=.02).

XSL•FO
RenderX

Table . Ordinal logistic regression analysis results for recognition of enhancing capabilities.

Characteristics	Level of recognition for improving diagnostic and treatment capabilities		Level of recognition for enhanced research capabilities	
	OR (95% CI)[a]	*P* value	OR (95% CI)[a]	*P* value
Hospital level (ref: tier-3)				
Tier-2 and tier - 1	1.531 (0.721-3.250)	.27	1.539 (0.750-3.155)	.24
Region (ref: Western)				
Eastern	1.656 (0.589-4.650)	.34	0.953 (0.350-2.591)	.93
Central	1.202 (0.557-2.594)	.64	1.226 (0.565-2.663)	.61
Discipline (ref: Internal medicine)				
Others	0.759 (0.155-3.730)	.73	1.501 (0.321-7.007)	.61
Surgery	1.952 (0.828-4.601)	.13	1.493 (0.671-3.319)	.33
Pediatrics	1.663 (0.363-7.626)	.51	1.364 (0.337-5.519)	.66
Obstetrics and gynecology	1.680 (0.482-5.857)	.42	6.217 (1.236-31.258)	.03[b]
Pathology	2.848 (0.513-15.814)	.23	2.649 (0.601-11.671)	.20
General medicine	0.335 (0.101-1.109)	.07	0.496 (0.166-1.489)	.21
Traditional Chinese medicine	0.368 (0.040-3.344)	.38	0.322 (0.051-2.026)	.23
Intensive care unit	0.614 (0.086-4.362)	.63	0.111 (0.014-0.893)	.04[b]
Sex (ref: Female)				
Male	0.803 (0.413-1.561)	.52	0.792 (0.420-1.493)	.47
Title (ref: Senior)				
Junior	1.451 (0.458-4.597)	.53	3.791 (1.231-11.673)	.02[b]
Intermediate	0.916 (0.389-2.157)	.84	1.655 (0.719-3.806)	.24
Education (ref: PhD)				
Associate or below	3.545 (0.588-21.354)	.17	1.120 (0.210-5.959)	.90
Bachelor	1.668 (0.387-7.185)	.49	1.063 (0.277-4.084)	.93
Master	1.991 (0.451-8.781)	.36	0.822 (0.207-3.269)	.78
Age	0.894 (0.802-0.996)	.04[b]	0.940 (0.851-1.038)	.22
Working years	1.106 (1.004-1.218)	.04[b]	1.091 (0.997-1.194)	.06
Telemedicine working years	0.956 (0.891-1.026)	.21	0.966 (0.897-1.039)	.35

[a]OR: odds ratio.

[b]*P* values <.05 are statistically significant.

Discussion

Main Findings

As the emerging paradigm in medical education, online CME has become increasingly popular in recent years, as evidenced by prior studies on its adoption by students in medical schools [24,25]. Our study demonstrates that online CME is also being widely adopted by medical professionals in hospitals, as seen in the Guoyuan mobile online CME platform by the National Telemedicine Center of China. This platform offers a wide range of functionalities, including live streaming, offline lecture recording, literature archives, and discussion forums, as well as a comprehensive selection of courses taught by professors from medical schools and tier-3 hospitals across various disciplines, which has been widely accepted by hospital staff in Henan Province, China. Since 2018, the number of registered users and the average course attendance have shown a rapid increasing trend, particularly after the outbreak of the COVID-19 pandemic, followed by a gradual decline returning to normal levels, which is consistent with the conclusions drawn by previous studies [26-28]. The attendance for courses in ophthalmology, otolaryngology, and pathology was higher than that in other disciplines, with a median attendance of 610 and 283 participants, respectively. Discipline category, professional title, and working years were the main factors influencing

XSL•FO
RenderX

satisfaction and recognition of competency enhancement in online CME.

Service Coverage and Popularity Among Medical Disciplines

The regional online CME platform offers courses across a wide range of medical disciplines. Internal medicine was the most popular in terms of course availability, total attendance, and average participation, likely reflecting its numerous subdisciplines and growing clinical priorities such as respiratory medicine and oncology, with courses on lung cancer, chronic obstructive pulmonary disease, gastric cancer, and liver cancer [29-31]. Ophthalmology and otolaryngology courses recorded the highest median views. These fields involve key facial regions, including the eyes, ears, and maxillofacial areas, and are among the fastest-growing specialties in digital technology integration. Rapid adoption of innovations such as implant navigation, 3D-printed jawbone prostheses, and artificial intelligence–assisted orthodontic design has shortened innovation cycles compared with general surgery [32,33]. The continuous emergence of new technologies requires practitioners in these fields to update their cross-disciplinary knowledge frequently, resulting in high attendance rates. However, as reported in a previous study [34], tier-2 and tier-1 hospitals often lack competence in surgery, pathology, medical imaging, and pharmacy. Our findings further indicate that these disciplines are insufficiently represented in the current online CME offerings, highlighting the need to expand course coverage and strengthen targeted training to enhance professional capacity in less-resourced hospitals.

Influencing Factors and Optimization Strategies for Doctors' Satisfaction

Satisfaction with online CME varied across specialties and professional levels. General practitioners reported lower satisfaction than internal medicine doctors, likely due to limited hardware, slower networks, and outdated equipment in primary health care institutions compared to tertiary hospitals equipped with 5G systems [35]. Moreover, online CME content often emphasizes specialized knowledge, which mismatches the broader training needs of general practitioners, leading to cognitive overload and unmet expectations. Professional title also influenced satisfaction: junior doctors expressed higher satisfaction with audio-video quality and service process than senior doctors, as online CME aligns with their exam-oriented learning needs and digital-native familiarity with online systems [36]. Regarding system convenience and process satisfaction, doctors with more years of experience exhibited higher satisfaction. Senior doctors have witnessed the evolution of medical information systems over the years, gradually strengthening their technological adaptability and accumulating extensive experience in handling system errors, leading to higher tolerance for workflow disruptions. Overall, online CME satisfaction was relatively high, but general practitioners reported lower satisfaction levels. It is recommended to gather extensive feedback from general practitioners, increase relevant course offerings, and improve both the hardware infrastructure of primary health care institutions and software interaction experiences to further enhance the quality of online CME services.

Influencing Factors and Optimization Strategies for Doctors' Recognition of Competency Enhancement in Online CME

Analysis of doctors' recognition of online CME revealed that both age and years of experience significantly influenced perceptions of its value in improving clinical skills. Doctors with more years of experience tended to recognize its practical benefits, while age showed a modest negative effect. This may reflect a cognitive imbalance between confidence in personal experience and reliance on digital tools [37]. Their extensive clinical experience enhances their ability to judge the practicality of online CME content (a positive effect of experience), while their reluctance toward new technological platforms (a negative effect of age) creates a value perception conflict. In terms of research capability enhancement, junior doctors exhibited higher acceptance of online CME. Being in the early stages of research skill development, they often lack proficiency in literature retrieval and experimental design. Online CME effectively bridges these gaps, enhancing their confidence and competence. Overall, doctors demonstrated strong recognition of online CME's contribution to professional growth, though some age-related resistance remains. Future course design should adopt an "age-friendly" and "experience-adaptive" approach, integrating senior experts' practical knowledge with digital learning. Meanwhile, strengthening research-oriented modules aligned with career progression can further support junior doctors' academic development.

Trend Analysis and the Future of Online Mobile CME

During the 6-year period from 2018 to 2023, total attendance for online courses on the Guoyuan mobile app showed a rise followed by a decline, peaking in 2020 and gradually decreasing thereafter. With the outbreak of COVID-19 at the end of 2019, offline teaching was restricted, creating a strong demand for online learning across multiple disciplines such as epidemiology, statistics, sociology, and psychology [38]. To address the urgent need for pandemic-related training, a large number of COVID-19 courses were launched on the Guoyuan mobile app, while regular courses were temporarily reduced. Hospitals across Henan Province used the platform for centralized epidemic prevention and control training, resulting in a peak in attendance in 2020. After 2020, as course offerings and clinical activities gradually normalized, attendance levels also returned to the prepandemic baseline. Studies indicate that offline courses outperform online ones in interaction, discussion, and teaching quality [39], while integrating online CME with group learning can significantly enhance doctors' learning outcomes [40]. The Guoyuan mobile platform mainly offers prerecorded, self-directed courses, where engagement depends on doctors' initiative. Expanding course coverage and incorporating real-time interaction could improve participation and learning outcomes. Aligned with China's digital health strategy and the World Health Organization's framework for lifelong learning, this model illustrates how mobile CME platforms can enhance equitable access and standardize training quality. Although based on a single regional platform, the findings suggest broader

XSL•FO
RenderX

applicability. The Guoyuan model provides a scalable approach for strengthening professional competency and continuous education in other Chinese provinces and low- and middle-income countries with limited medical education resources.

Limitations

Although our findings provide a deep insight into the development of online CME in China, the study still has several limitations. First, the courses released by the Guoyuan education app mainly serve to improve the level of medical personnel in the first stage of promotion, while there are few courses focused on the popularization of medical knowledge among the general population. Future development of the app will enrich the courses' content and provide more courses on the popularization of basic medical knowledge. Second, considering the assessment requirements of continuing education for medical staff every year in China, we will introduce continuing education courses into the app to enhance the convenience of medical staff to participate in continuing education and expand the influence of the app. Additionally, this study is exploratory in nature, and its findings should be interpreted with certain limitations. The use of a voluntary response sample from a single platform may introduce self-selection bias and limit the generalizability of the conclusions. The analytical outcomes should be interpreted as a preliminary exploration of online CME dynamics. Future studies will use more rigorous sampling frameworks to further validate these exploratory findings and enhance the robustness of the results.

Conclusions

In this study, we examined the trends in the uptake of mobile online CME services by medical professionals in China through a representative study on a regional platform operated by the National Telemedicine Center of China. The platform started offering medical training via mobile apps in 2018 and is gaining popularity among staff in both tier-3 hospitals and tier-2 and tier‐1 hospitals over the years, particularly after the outbreak of COVID-19. In addition, the platform offered a wide range of courses covering multiple disciplines. Internal medicine was the most popular discipline in terms of the number of available courses, total attendance, and average attendance per course, and ophthalmology and otolaryngology courses had the highest median number of views. The professional title, discipline category, age, and working years were identified as the main factors influencing the satisfaction and recognition of competency enhancement in online CME. To further improve the quality of online CME services, it is necessary to increase the number of courses in disciplines with fewer offerings and conduct research on disciplines with lower satisfaction levels to optimize course structures and provide more targeted content. Additionally, efforts should be made to enhance the user interaction experience of the software platforms, ensuring accessibility and barrier-free usability for individuals across all age groups.

Funding

This research was supported by grants from the National Key R&D Program of China (grant number 2022ZD0160704), the Science and Technology Research Program in Henan Province (grant number 242102311021), the Key Provincial-Ministerial Co-Construction Project of the Henan Medical Science and Technology Research Program (grant number SBGJ202402097), the Joint Co-construction Project of the Henan Medical Science and Technology Research Program (grant number LHGJ20240288), the Basic Research and Applied Basic Research Program in Zhengzhou City (grant number 2024ZZJCYJ017), and Basic Research Operating Fund Project of Henan Academy of Innovations in Medical Science (JBKY250307). The funders of the study played no role in the design, data collection, or interpretation of the present work. The views expressed in the article are those of the authors and do not necessarily reflect the position of the funding bodies.

Data Availability

The datasets used and/or analyzed during this study are available from the corresponding author on reasonable request.

Authors' Contributions

FC, JZ, XZ, and XH made substantial contributions to the design of this work. FC, JZ, and XZ designed, led out, initiated the survey of this study, and drafted the initial article. XH conducted data analysis. DL and HC played an important role in the design and application of the app. YC and WW were involved in the discussion of article structure and results. XZ, ML, and MJ reviewed and revised the article.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Major functional modules in the Guoyuan mobile tele-education app.
[DOCX File, 16 KB - mededu_v12i1e76299_app1.docx]

References

XSL•FO
RenderX

1. Liu W, Liu Y, Twum P, Li S. National equity of health resource allocation in China: data from 2009 to 2013. Int J Equity Health 2016 Apr 19;15:68. [doi: 10.1186/s12939-016-0357-1] [Medline: 27093960]
2. Zhang X, Zhao L, Cui Z, Wang Y. Study on equity and efficiency of health resources and services based on key indicators in China. PLoS ONE 2015;10(12):e0144809. [doi: 10.1371/journal.pone.0144809] [Medline: 26679187]
3. Ding J, Hu X, Zhang X, Shang L, Yu M, Chen H. Equity and efficiency of medical service systems at the provincial level of China's mainland: a comparative study from 2009 to 2014. BMC Public Health 2018 Feb 5;18(1):214. [doi: 10.1186/s12889-018-5084-7] [Medline: 29402260]
4. Li D, Zhou Z, Si Y, et al. Unequal distribution of health human resource in mainland China: what are the determinants from a comprehensive perspective? Int J Equity Health 2018 Feb 27;17(1):29. [doi: 10.1186/s12939-018-0742-z] [Medline: 29486791]
5. Cao X, Bai G, Cao C, et al. Comparing regional distribution equity among doctors in China before and after the 2009 medical reform policy: a data analysis from 2002 to 2017. Int J Environ Res Public Health 2020 Feb 27;17(5):1520. [doi: 10.3390/ijerph17051520] [Medline: 32120925]
6. China health statistics yearbooks in 2019 [Web page in Chinese]. National Health Commission of the People's Republic of China. 2019. URL: https://www.nhc.gov.cn/mohwsbwstjxxzx/tjtjnj/202006/456391e2294244c593e871ac254f8aaa.shtml [accessed 2026-02-04]
7. Dai G, Li R, Ma S. Research on the equity of health resource allocation in TCM hospitals in China based on the Gini coefficient and agglomeration degree: 2009-2018. Int J Equity Health 2022 Oct 6;21(1):145. [doi: 10.1186/s12939-022-01749-7] [Medline: 36199086]
8. Anand S, Fan VY, Zhang J, et al. China's human resources for health: quantity, quality, and distribution. Lancet 2008 Nov 15;372(9651):1774-1781. [doi: 10.1016/S0140-6736(08)61363-X] [Medline: 18930528]
9. Dong E, Liu S, Chen M, et al. Differences in regional distribution and inequality in health-resource allocation at hospital and primary health centre levels: a longitudinal study in Shanghai, China. BMJ Open 2020 Jul 19;10(7):e035635. [doi: 10.1136/bmjopen-2019-035635] [Medline: 32690509]
10. Xu X, Zhou L, Antwi HA, Chen X. Evaluation of health resource utilization efficiency in community health centers of Jiangsu Province, China. Hum Resour Health 2018 Feb 20;16(1):13. [doi: 10.1186/s12960-018-0275-y] [Medline: 29463253]
11. Norman GR, Shannon SI, Marrin ML. The need for needs assessment in continuing medical education. BMJ 2004 Apr 24;328(7446):999-1001. [doi: 10.1136/bmj.328.7446.999] [Medline: 15105326]
12. Hassan Alshehri HM, Alshehri AMH, Alshehri AMH, et al. Primary health care professionals' opinions regarding continuing medical education: a cross sectional study. Medicine (Baltimore) 2024 Dec 13;103(50):e40865. [doi: 10.1097/MD.0000000000040865] [Medline: 39686415]
13. Walsh S, De Villiers MR, Golakai VK. Introducing an e-learning solution for medical education in Liberia. Ann Glob Health 2018 Apr 30;84(1):190-197. [doi: 10.29024/aogh.21] [Medline: 30873817]
14. Hanson AH, Krause LK, Simmons RN, et al. Dermatology education and the Internet: traditional and cutting-edge resources. J Am Acad Dermatol 2011 Oct;65(4):836-842. [doi: 10.1016/j.jaad.2010.05.049] [Medline: 21820206]
15. Danesh V, Rolin D, Hudson SV, White S. Telehealth in mental health nursing education: health care simulation with remote presence technology. J Psychosoc Nurs Ment Health Serv 2019 Jul 1;57(7):23-28. [doi: 10.3928/02793695-20190306-01] [Medline: 30888428]
16. Seymour-Walsh AE, Bell A, Weber A, Smith T. Adapting to a new reality: COVID-19 coronavirus and online education in the health professions. Rural Remote Health 2020 May;20(2):6000. [doi: 10.22605/RRH6000] [Medline: 32456441]
17. Pather N, Blyth P, Chapman JA, et al. Forced disruption of anatomy education in Australia and New Zealand: an acute response to the Covid-19 pandemic. Anat Sci Educ 2020 May;13(3):284-300. [doi: 10.1002/ase.1968] [Medline: 32306555]
18. Lellis-Santos C, Abdulkader F. Smartphone-assisted experimentation as a didactic strategy to maintain practical lessons in remote education: alternatives for physiology education during the COVID-19 pandemic. Adv Physiol Educ 2020 Dec 1;44(4):579-586. [doi: 10.1152/advan.00066.2020] [Medline: 32955344]
19. Wish-Baratz S, Crofton AR, Gutierrez J, Henninger E, Griswold MA. Assessment of mixed-reality technology use in remote online anatomy education. JAMA Netw Open 2020 Sep 1;3(9):e2016271. [doi: 10.1001/jamanetworkopen.2020.16271] [Medline: 32940677]
20. Nicklen P, Keating JL, Paynter S, Storr M, Maloney S. Remote-online case-based learning: a comparison of remote-online and face-to-face, case-based learning - a randomized controlled trial. Educ Health (Abingdon) 2016;29(3):195-202. [doi: 10.4103/1357-6283.204213] [Medline: 28406103]
21. Jiang Q, Pan J. The evolving hospital market in China after the 2009 healthcare reform. Inquiry 2020;57:46958020968783. [doi: 10.1177/0046958020968783] [Medline: 33124476]
22. Liu J, Mao Y. Continuing medical education and work commitment among rural healthcare workers: a cross-sectional study in 11 western provinces in China. BMJ Open 2020 Aug 4;10(8):e037985. [doi: 10.1136/bmjopen-2020-037985] [Medline: 32753451]
23. Shalabi KM, Almurdi MM. Satisfaction and attitudes towards online continuous medical education and its impact on clinical practice among physiotherapists. BMC Med Educ 2024 Jan 17;24(1):70. [doi: 10.1186/s12909-024-05049-2] [Medline: 38233905]

XSL•FO
RenderX

24. Sharma D, Bhaskar S. Addressing the Covid-19 burden on medical education and training: the role of telemedicine and tele-education during and beyond the pandemic. Front Public Health 2020;8:589669. [doi: 10.3389/fpubh.2020.589669] [Medline: 33330333]
25. Newman J, Liew A, Bowles J, Soady K, Inglis S. Podcasts for the delivery of medical education and remote learning. J Med Internet Res 2021 Aug 27;23(8):e29168. [doi: 10.2196/29168] [Medline: 34448719]
26. Wallace S, Schuler MS, Kaulback M, Hunt K, Baker M. Nursing student experiences of remote learning during the COVID-19 pandemic. Nurs Forum 2021 Jul;56(3):612-618. [doi: 10.1111/nuf.12568] [Medline: 33728660]
27. Becker SP, Breaux R, Cusick CN, et al. Remote learning during COVID-19: examining school practices, service continuation, and difficulties for adolescents with and without attention-deficit/hyperactivity disorder. J Adolesc Health 2020 Dec;67(6):769-777. [doi: 10.1016/j.jadohealth.2020.09.002] [Medline: 33069580]
28. Dost S, Hossain A, Shehab M, Abdelwahed A, Al-Nusair L. Perceptions of medical students towards online teaching during the COVID-19 pandemic: a national cross-sectional survey of 2721 UK medical students. BMJ Open 2020 Nov 5;10(11):e042378. [doi: 10.1136/bmjopen-2020-042378] [Medline: 33154063]
29. Global Burden of Disease 2019 Cancer Collaboration, Kocarnik JM, Compton K, et al. Cancer incidence, mortality, years of life lost, years lived with disability, and disability-adjusted life years for 29 cancer groups from 2010 to 2019: a systematic analysis for the global burden of disease study 2019. JAMA Oncol 2022 Mar 1;8(3):420-444. [doi: 10.1001/jamaoncol.2021.6987] [Medline: 34967848]
30. Vos T, Abajobir AA, Abate KH, et al. Global, regional, and national incidence, prevalence, and years lived with disability for 328 diseases and injuries for 195 countries, 1990–2016: a systematic analysis for the global burden of disease study 2016. The Lancet 2017 Sep;390(10100):1211-1259. [doi: 10.1016/S0140-6736(17)32154-2] [Medline: 28711585]
31. Zhou M, Wang H, Zeng X, et al. Mortality, morbidity, and risk factors in China and its provinces, 1990-2017: a systematic analysis for the global burden of disease study 2017. Lancet 2019 Sep 28;394(10204):1145-1158. [doi: 10.1016/S0140-6736(19)30427-1] [Medline: 31248666]
32. Jie B, Yao B, Li R, An J, Zhang Y, He Y. Post-traumatic maxillofacial reconstruction with vascularized flaps and digital techniques: 10-year experience. Int J Oral Maxillofac Surg 2020 Nov;49(11):1408-1415. [doi: 10.1016/j.ijom.2020.04.012] [Medline: 32404244]
33. Li JPO, Liu H, Ting DSJ, et al. Digital technology, tele-medicine and artificial intelligence in ophthalmology: a global perspective. Prog Retin Eye Res 2021 May;82:100900. [doi: 10.1016/j.preteyeres.2020.100900] [Medline: 32898686]
34. Jiang C, Gu J. History and current state of pathology in China. Virchows Arch 2013 Oct;463(4):599-608. [doi: 10.1007/s00428-013-1449-6] [Medline: 23881278]
35. Reid L, Button D, Brommeyer M. Challenging the myth of the digital native: a narrative review. Nurs Rep 2023 Apr 4;13(2):573-600. [doi: 10.3390/nursrep13020052] [Medline: 37092480]
36. Cui F, Ma Q, He X, et al. Implementation and application of telemedicine in China: cross-sectional study. JMIR Mhealth Uhealth 2020 Oct 23;8(10):e18426. [doi: 10.2196/18426] [Medline: 33095175]
37. Czaja SJ, Charness N, Fisk AD, et al. Factors predicting the use of technology: findings from the Center for Research and Education on Aging and Technology Enhancement (CREATE). Psychol Aging 2006 Jun;21(2):333-352. [doi: 10.1037/0882-7974.21.2.333] [Medline: 16768579]
38. Han L, Wu F. COVID-19 drives medical education reform to promote “healthy China 2030” action plan. Front Public Health 2024;12:1465781. [doi: 10.3389/fpubh.2024.1465781] [Medline: 39525459]
39. Rastogi A, Bansal A, Keshan P, Jindal A, Prakash A, Kumar V. Medical education in post-pandemic times: online or offline mode of learning? J Family Med Prim Care 2022 Sep;11(9):5375-5386. [doi: 10.4103/jfmpc.jfmpc_2305_21] [Medline: 36505568]
40. Dowling S, Minihan F, McNicholas C, et al. Benefits and limitations of the transfer online of Irish College of General Practitioners (ICGP) continuing medical education (CME) small group learning (SGL) during COVID-19: a national Delphi study. Rural Remote Health 2023 Jan;23(1):8157. [doi: 10.22605/RRH8157] [Medline: 36802741]

Abbreviations

CME: continuing medical education
OR: odds ratio

XSL•FO
RenderX

Edited by A Bahattab; submitted 21.Apr.2025; peer-reviewed by W Zhang, Z Ge; revised version received 24.Dec.2025; accepted 03.Jan.2026; published 23.Feb.2026.

Please cite as:
Zhang X, He X, Chu Y, Liu D, Lyu M, Wang W, Chen H, Ji M, Cui F, Zhao J
The Design and Evaluation of an Online Continuing Medical Education App for Medical Professionals in China: Quantitative Study
JMIR Med Educ 2026;12:e76299
URL: https://mededu.jmir.org/2026/1/e76299
doi:10.2196/76299

© Xu Zhang, Xianying He, Yuntian Chu, Dongqing Liu, Minzhao Lyu, Weiyi Wang, Haotian Chen, Meihao Ji, Fangfang Cui, Jie Zhao. Originally published in JMIR Medical Education (https://mededu.jmir.org), 23.Feb.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Education, is properly cited. The complete bibliographic information, a link to the original publication on https://mededu.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Global Learner Feedback on Continuing Medical Education–Accredited e-Learning Modules in Pediatric Endocrinology and Diabetes: Cross-Sectional Study

Yvonne G van der Zwan[1], MD, PhD; Conny van Wijngaard-de Vugt[2], MScBA; Abdulsalam Abu-Libdeh[3,4], MD, PhD; Evangelia Kalaitzoglou[5], MD, PhD; Zacharoula Karabouta[6], MD, PhD; Stenvert L S Drop[7], MD, PhD; Annemieke M Boot[8], MD, PhD; Sze May Ng[9,10], MD, PhD; Jan Idkowiak[11,12,13], MD, PhD

[1]Department of Paediatrics, Isala, Zwolle, The Netherlands

[10]Department of Pediatric Endocrinology, Mersey and West Lancashire Teaching Hospitals NHS Trust, Southport, United Kingdom

[11]Department of Metabolism and Systems Science, College of Medicine and Health, University of Birmingham, Edgbaston, IBR tower, level 2, Birmingham, United Kingdom

[12]Centre for Endocrinology, Diabetes and Metabolism, Birmingham Health Partners, University of Birmingham, Birmingham, United Kingdom

[13]Department of Endocrinology and Diabetes, Birmingham Children's Hospital, Birmingham Women's and Children's NHS Foundation Trust, Birmingham, United Kingdom

[2]WV Research, Advice and Management, Brunssum, The Netherlands

[3]Department of Pediatrics, Division of Pediatric Endocrinology, Makassed Hospital and Al-Quds Medical School, Jerusalem, Palestinian Territory, Occupied

[4]Division of Pediatric Endocrinology, Hadassah University Medical Center & Hadassah Medical School, Jerusalem, Israel

[5]Department of Pediatrics, University of Kentucky, Barnstable Brown Diabetes Center, Lexington, KY, United States

[6]2nd Paediatric Department, AHEPA University General Hospital, Thessaloniki, Greece

[7]Division of Endocrinology, Department of Pediatrics, Sophia Children's Hospital, Erasmus MC, Rotterdam, The Netherlands

[8]Department of Pediatric Endocrinology, University Medical Center Groningen, University of Groningen, Groningen, The Netherlands

[9]Edge Hill University, Ormskirk, England, United Kingdom

Corresponding Author:
Jan Idkowiak, MD, PhD
Department of Metabolism and Systems Science, College of Medicine and Health, University of Birmingham, Edgbaston, IBR tower, level 2, Birmingham, United Kingdom

Abstract

Background: The European Society for Paediatric Endocrinology (ESPE) e-Learning wesite is a free, globally accessible online resource to enhance learning in pediatric endocrinology and pediatric diabetes. The content is created by world-leading experts in pediatric endocrinology and pediatric diabetes and is closely aligned with published international consensus guidelines. In August 2022, 30 hours of e-learning courses received accreditation from the European Accreditation Council for Continuing Medical Education (CME). These CME courses cover three categories: (1) pediatric endocrinology, (2) pediatric diabetes, and (3) pediatric endocrinology in resource-limited settings.

Objective: This study aimed to assess learners' demographics and feedback from mandatory surveys after completion of CME e-learning courses and to identify areas for improvement.

Methods: The ESPE e-learning committee created a mandatory survey for each CME e-learning module. The survey includes baseline demographics and feedback on the quality of the learning content, assessed using a 5-point Likert scale. Data were extracted from the start of the CME modules in August 2022 until September 2025.

Results: A total of 567 surveys were completed: 286 (50.4%) in the category pediatric endocrinology, 225 (39.7%) in the category pediatric diabetes based on the International Society for Pediatric and Adolescent Diabetes guidelines, and 56 (9.9%) in the category pediatric endocrinology in resource-limited settings. There was global participation, with most learners practicing in Europe (n=333, 59%), followed by Asia (n=124, 22%), Africa (n=53, 9%), the Americas (North America, n=45, 8%; and South America, n=11, 2%), and Oceania (n=1, 0%). Most of the users indicated that they were medical experts (n=210, 37%), followed by fellows or residents (n=223, 39%), and medical students and nurses (n=29, 5% and n=32, 6%, respectively); overall, 10% (n=56) of learners practice in resource-limited countries. Overall, the learning content was well received for all modules regarding accessibility, organization, level of interest, improvement of learners' clinical practice, appropriateness of content, and provision of feedback (median Likert score 4, IQR 4-5). Learners' free-text feedback identified some areas for improvement, including

XSL•FO
RenderX

reducing text-heavy content and providing more graphical content and more interactive case reports. Most learners' free-text feedback consisted of encouraging and thankful comments.

Conclusions: The ESPE CME–accredited e-learning modules are well received, providing globally free CME education in pediatric endocrinology and pediatric diabetes. These findings support the continued development and promotion of open-access CME platforms, with the aim of improving global equity in specialist medical education and focusing on educational impact.

(JMIR Med Educ 2026;12:e67332) doi:10.2196/67332

KEYWORDS

e-learning; continuing medical education; endocrinology; European Society for Paediatric Endocrinology; ESPE; International Society for Pediatric and Adolescent Diabetes; ISPAD; pediatric endocrinology; pediatrics; diabetes; eHealth; digital health; medical education

Introduction

The European Society for Paediatric Endocrinology (ESPE) e-Learning website [1] is a free, globally accessible online tool to enhance learning in pediatric endocrinology (PE) and pediatric diabetes (PD) worldwide [2-4].

The ESPE e-Learning website was first published online in 2012. Since then, the content and use have expanded with courses available in PE and PE in resource-limited settings (RLSs). The e-learning committee of ESPE and the International Society for Pediatric and Adolescent Diabetes (ISPAD) are collaboratively responsible for establishing and maintaining the e-learning platform, as well as ensuring the development and quality of its content. This freely accessible online portal allows medical students, fellows, specialists, nurses, and tutors from around the world to share, contribute to, and expand their knowledge through interactive chapters and case-based learning, covering both core- and advanced-level modules [5,6]. During the COVID-19 pandemic, e-learning became an essential resource for continuous professional development, and online learning demonstrated that the effects of e-learning are equivalent to traditional learning [7-9]. e-Learning has now become a solution for overcoming training barriers posed by social distancing rules, offering education and learning opportunities to students, trainees, and physicians to help them maintain essential competencies and continue professional development. It is especially beneficial for individuals working residential shifts with work hour restrictions, which frequently conflict with in-person attendance at didactic lectures, and for those living in remote areas.

The ESPE e-learning website contains 80 chapters and more than 130 cases covering core and advanced learning courses on the most common endocrine disorders, including diabetes; of these, 16 chapters and 25 cases target health care professionals based in RLSs. The content includes normal physiology, pathophysiological mechanisms underlying endocrine diseases, diagnostic approaches or algorithms, and management based on international expert consensus and published guidelines. Direct feedback on questions from users is encouraged. Recently, two video masterclasses delivered by internationally recognized experts in the field have been added. Steering, conceptual, and content oversight is provided by the ESPE-ISPAD e-learning committee, which is responsible for creating and maintaining the e-learning website, ensuring up-to-date information and high-quality standards.

All courses are available in English, and all courses in the category *health care in resource-limited settings*, which is specifically intended for practitioners working in primary, secondary, and tertiary health care centers in resource-limited countries, are available in four additional languages (ie, French, Spanish, Swahili, and Chinese) [10].

Since August 2022, 30 credit hours of ESPE-ISPAD continuing medical education (CME) e-learning courses created by world-leading experts in PE and PD have received accreditation from the European Accreditation Council for CME (EACCME). These CME courses are typically based on consensus guidelines. They are organized in three categories with 10 courses each in the categories *pediatric endocrinology*, *pediatric diabetes*, and *health care in resource-limited settings*. The latter module is available in five different languages (ie, English, French, Spanish, Swahili, and Chinese). Each CME course takes approximately 1 to 1.5 hours of educational time, and completion of a course provides one European CME credit.

To promote global awareness and uptake of the CME-accredited modules, the ESPE e-learning platform was actively disseminated through multiple channels. These included presentations at international society meetings, such as ESPE and ISPAD, regular features in quarterly ESPE newsletters and bimonthly ISPAD newsletters, and targeted outreach via social media platforms, including Facebook and X. These efforts aimed to ensure broad visibility and accessibility of the platform across diverse geographic and professional audiences.

To date, little is known about how these accredited modules are used and perceived by health care professionals globally. Herein, we aimed to evaluate the geographical reach, user demographics, and learner feedback on the accredited CME modules, specifically examining how health care professionals from diverse regions and professional backgrounds engage with and evaluate the CME-accredited modules. Therefore, we conducted a descriptive cross-sectional analysis of the mandatory postcourse survey data collected between August 2022 and September 2025. We hypothesized that the modules would be positively received across user groups and that learners would identify specific areas for improvement. The findings may help inform educators, module developers, and professional societies in improving their approach to expand equitable access to high-quality, guideline-based CME in PE and PD.

XSL•FO
RenderX

Methods

Ethical Considerations

Participation in the feedback survey was entirely voluntary, and informed consent was implied by completion of the survey. Respondents were clearly informed that they could opt out at any time by not submitting the survey. No additional consent was required because this was a secondary analysis of anonymized data collected under the original consent process for CME module participation, which included agreement to provide feedback for service evaluation purposes.

Ethical approval was not deemed necessary as the analysis of the survey was conducted as an evaluation and quality improvement exercise to evaluate the educational offering. Data were collected anonymously from adults/professionals (no vulnerable population); no identifiable information was collected; there was no intervention, randomization, or direct implications on patient care; and the participation was entirely voluntary.

No financial or material compensation was provided to participants for completing the survey.

Survey Design and Setting

The reporting of this study adhered to the CHERRIES (Checklist for Reporting Results of Internet E-Surveys) to ensure methodological transparency and reproducibility [11]. The checklist was applied to all stages of the online evaluation process, including survey design, delivery, and data handling. The evaluation questionnaire was developed by the ESPE e-learning team and pretested with a small group of users to assess clarity, face validity, and technical functionality. The final version was implemented within the online education platform, which automatically invited participants to complete the anonymous feedback survey upon module completion (items 5 - 8). Participation was mandatory upon completion of the course; no personal data, as per General Data Protection Regulation requirements, were collected. No incentives were offered. Measures to prevent multiple submissions included limiting one response per user account and restricting survey access to registered participants only. Data completeness checks were built into the survey logic, and incomplete responses were excluded from the main analysis but retained for sensitivity assessment. All data were collected through a secure institutional server with encrypted transmission and stored in accordance with international data governance standards.

The survey was created by the ESPE e-learning committee to obtain information on learners' demographics and to evaluate the use and quality of the courses. The content, in accordance with Kirkpatrick's evaluation type level 1 (*reaction*) [12,13], included questions about learners' professional backgrounds and countries of residence. Feedback on the quality of the learning content, presentation, accessibility, and the anticipated impact on clinical practice was assessed using a 5-point Likert scale, ranging from strongly agree (Likert scale=1), agree (Likert scale=2), neutral (Likert scale=3), disagree (Likert scale=4), and strongly disagree (Likert scale=5). We used a different Likert scale for one specific question asking for the degree of difficulty of quiz questions provided in the modules (Likert scale 1=too easy, 2=easy, 3=appropriate, 4=difficult, and 5=too difficult). Likert scales were chosen to allow nuanced assessment of attitudes and perceptions, are well understood by respondents, and facilitate both descriptive and inferential statistical analyses. Additional feedback on content, user interface, or other issues was encouraged through an open free-text question. The detailed survey is provided in Multimedia Appendix 1. Data were analyzed from the launch of the CME module in August 2022 until September 2025. The professional background was categorized based on professional training and experience in PE and grouped as (1) medical expert or consultant, (2) fellow, (3) resident, (4) medical student, (5) nurse, and (6) other (free text). Health care officers from research-limited countries were asked if they were based in primary, secondary, or tertiary level settings to understand resources and facilities available (primary level: basic or rural with very limited laboratory and imaging facilities; secondary level: district or regional hospitals with limited laboratory and imaging facilities; and tertiary level: main or national referral hospitals with most but not all laboratory and imaging facilities [10]).

Visualization of Global Participation

Geographical distribution of participants completing the online CME modules was visualized using *Plotly* (version 5.24.0; Plotly Technologies Inc). Country names from the registration dataset were matched to International Organization for Standardization–recognized country centroids, and each country was represented by a proportional bubble plotted on a world map in a natural-earth projection.

Statistical Analysis

Descriptive statistics were applied. When analyzing the 5-point Likert scale, we used the median (IQR) as appropriate for an ordinal scale; however, we have also provided the mean and SD to describe the level of dispersion. As Likert responses are ordinal and nonnormally distributed, comparisons across the three modules were performed using the Kruskal-Wallis H test. Where relevant, P values <.05 were considered statistically significant. A stacked Likert plot was generated to visualize the proportional distribution of responses across all survey questions. Free-text feedback was independently coded into four categories by two authors. Discrepancies were discussed, and final classifications were reached by consensus. We have used Microsoft Excel and GraphPad Prism for descriptive statistics and graphical illustration of the data.

Results

From August 2022 to September 2025, a total of 567 surveys were completed. Of those, 286 (50.4%) were in the category pediatric endocrinology, 225 (39.7%) were in the category pediatric diabetes, and 56 (9.9%) were in the category pediatric endocrinology in RLS.

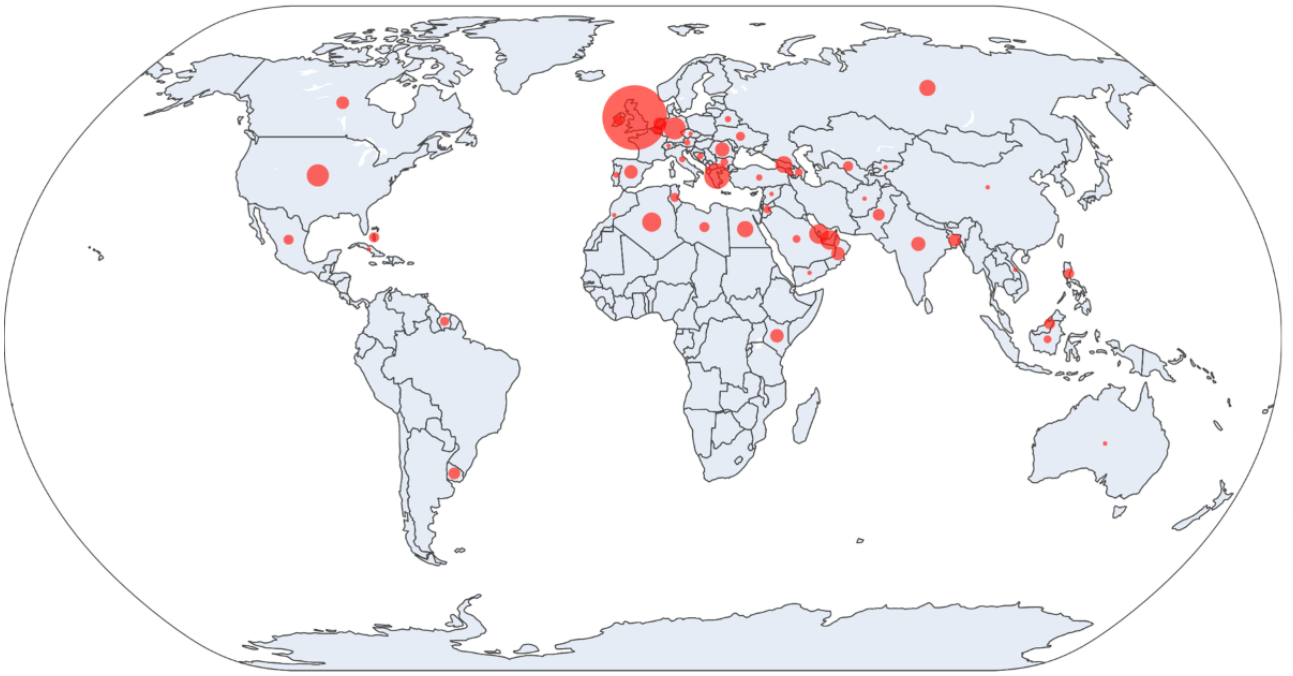
Learner's Background

There was global participation, but most learners were practicing in Europe (n=333, 59%), followed by Asia (n=124, 22%), Africa (n=53, 9%), the Americas (North America, n=45, 8%; and South America, n=11, 2%), and Oceania (n=1, 0%; Figure 1 and

XSL•FO
RenderX

Multimedia Appendix 2). Although the RLS courses are available in Mandarin, there was no participation from China.

Figure 1. Global distribution of participants completing online continuing medical education modules**.** Each bubble represents a country, with bubble size proportional to the number of completed modules. Red semitransparent bubbles are projected onto a natural-earth world map, with representative bubble sizes (n=5, 10, and 100) shown for scale.

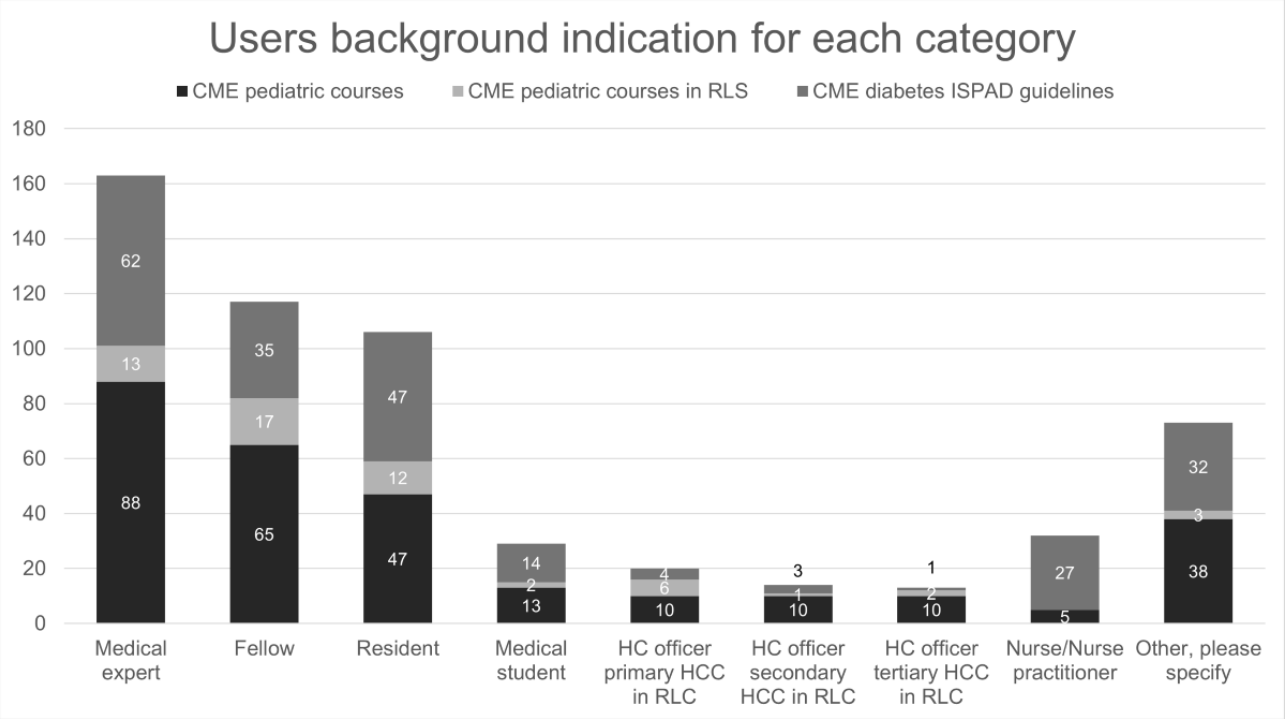


Most users indicated that they were medical experts (210/567, 37%), followed by fellows or residents (223/567, 39%). Medical students and nurses formed 5% and 6%, respectively, of the total group, and 13% were categorized under *other* (Figure 2). Relevant specifications given in the category *other* include lecturer, psychologist, patient advocate, diabetes educator, nutritionist, and education support worker.

Twelve percent (n=70/567) of learners practice in RLS countries, of which 20 work as health care officers in primary health care centers, 14 work in secondary health care centers, and 13 work in tertiary health care centers [14].

XSL•FO
RenderX

Figure 2. Representation of learners according to training background. CME: continuing medical education; HC: health care; HCC: health care center; ISPAD: International Society for Pediatric and Adolescent Diabetes; RLS: resource-limited setting.



Feedback on CME Modules

Across all survey items, overall learner satisfaction was high, with most responses falling within the *agree* or *strongly agree* categories (Table 1, Figure 3). The median scores for the majority of questions were 4.0, with means consistently above 4.0 for questions relating to content quality, clarity, motivation, competence, and perceived educational value. The feedback on the questions' difficulty level was also narrow, with the majority of responses falling into the *appropriate* category. The PE module generally received the highest mean ratings across items, whereas the RLS and PD modules showed similar, though slightly lower, response patterns. IQRs were narrow for most questions, indicating highly compressed distributions with limited variability. Kruskal-Wallis testing revealed statistically significant differences between modules for most survey items (P<.05), primarily due to a higher proportion of strongly positive responses in the PE module. Exceptions included the question on perceived improvement in performance, where module differences were not statistically significant. Items relating to preferences for additional open-ended questions (eg, bullet list and teacher feedback formats) received comparatively lower mean scores and showed greater variability. Overall, the findings indicate strong learner engagement and perceived educational value, with some variation in response patterns across modules (Table 1, Figure 3).

XSL•FO
RenderX

Table . Descriptive and comparative statistics for all Likert-scale survey questions across the three modules (pediatric endocrinology [PE], n=283; PE in resource-limited settings [RLS], n=56; and pediatric diabetes [PD], n=228).

Survey question[a] and modules	Likert score responses, n					Comparative statistics		
	1	2	3	4	5	Median (IQR)	Mean (SD)	*P* value
The online format was appropriate for the subject matter and I was able to access all components of the activity without difficulty.[b]								*<.001*
PE	0	5	9	137	132	4.0 (4.0-5.0)	4.40 (0.64)	
RLS	0	0	5	34	17	4.0 (4.0-5.0)	4.21 (0.59)	
PD	0	1	16	146	65	4.0 (4.0-5.0)	4.21 (0.58)	
The material was organized clearly for learning to occur.[b]								*<.001*
PE	0	4	9	138	132	4.0 (4.0-5.0)	4.41 (0.63)	
RLS	0	0	5	37	14	4.0 (4.0-4.2)	4.16 (0.57)	
PD	0	2	17	143	66	4.0 (4.0-5.0)	4.20 (0.60)	
The content of this chapter or case is interesting to me.[b]								*<.001*
PE	0	1	18	122	142	5.0 (4.0-5.0)	4.43 (0.63)	
RLS	0	1	5	35	15	4.0 (4.0-5.0)	4.14 (0.65)	
PD	0	1	14	139	74	4.0 (4.0-5.0)	4.25 (0.58)	
After studying this chapter or case, I feel motivated to learn more on the subject.[b]								*<.001*
PE	0	4	18	118	143	5.0 (4.0-5.0)	4.41 (0.68)	
RLS	0	0	5	34	17	4.0 (4.0-5.0)	4.21 (0.59)	
PD	0	1	18	141	68	4.0 (4.0-5.0)	4.21 (0.59)	
This activity will assist in the improvement of my competence.[b]								*.02*
PE	0	0	13	140	130	4.0 (4.0-5.0)	4.41 (0.58)	
RLS	0	0	2	38	16	4.0 (4.0-5.0)	4.25 (0.51)	
PD	0	1	11	134	82	4.0 (4.0-5.0)	4.30 (0.58)	
This activity will assist in the improvement of my performance.[b]								.12
PE	0	0	20	139	124	4.0 (4.0-5.0)	4.37 (0.61)	
RLS	0	0	2	38	16	4.0 (4.0-5.0)	4.25 (0.51)	
PD	0	1	12	135	80	4.0 (4.0-5.0)	4.29 (0.58)	
This activity will assist in the improvement of my patient outcomes.[b]								*<.001*
PE	0	0	26	128	129	4.0 (4.0-5.0)	4.36 (0.65)	
RLS	0	0	2	38	16	4.0 (4.0-5.0)	4.25 (0.51)	
PD	0	2	19	125	82	4.0 (4.0-5.0)	4.26 (0.64)	
The content and questions align with my knowledge level.[b]								*<.001*
PE	2	8	25	176	72	4.0 (4.0-5.0)	4.09 (0.72)	
RLS	0	0	5	45	6	4.0 (4.0-4.0)	4.02 (0.45)	
PD	0	4	19	156	49	4.0 (4.0-4.0)	4.10 (0.60)	
I like self-assessment with multiple choice questions.[b]								*.02*
PE	0	5	17	145	116	4.0 (4.0-5.0)	4.31 (0.67)	
RLS	0	3	4	38	11	4.0 (4.0-4.0)	4.02 (0.70)	
PD	0	1	27	142	58	4.0 (4.0-5.0)	4.13 (0.61)	
The feedback that is given after answering the questions is appropriate.[b]								*<.001*

XSL•FO
RenderX

Survey question[a] and modules	Likert score responses, n					Comparative statistics		
	1	2	3	4	5	Median (IQR)	Mean (SD)	*P* value
PE	1	3	19	158	102	4.0 (4.0-5.0)	4.26 (0.66)	
RLS	0	0	5	37	14	4.0 (4.0-4.2)	4.16 (0.57)	
PD	0	2	18	152	56	4.0 (4.0-4.0)	4.15 (0.58)	
I would like a few (more) open questions where answers are provided as a bullet list with relevant items.[b]								*<.001*
PE	8	27	62	123	63	4.0 (3.0-4.0)	3.73 (1.00)	
RLS	0	7	20	21	8	4.0 (3.0-4.0)	3.54 (0.89)	
PD	0	22	80	87	39	4.0 (3.0-4.0)	3.63 (0.88)	
I would like a few (more) open questions for classroom/teacher feedback.[b]								*.02*
PE	7	31	70	117	58	4.0 (3.0-4.0)	3.66 (1.00)	
RLS	0	6	25	21	4	3.0 (3.0-4.0)	3.41 (0.78)	
PD	0	34	81	76	37	3.0 (3.0-4.0)	3.51 (0.94)	
The content and questions are:[c]								*<.001*
PE	2	46	207	22	6	3.0 (3.0-3.0)	2.94 (0.59)	
RLS	0	3	46	7	0	3.0 (3.0-3.0)	3.07 (0.42)	
PD	0	9	190	22	7	3.0 (3.0-3.0)	3.12 (0.50)	

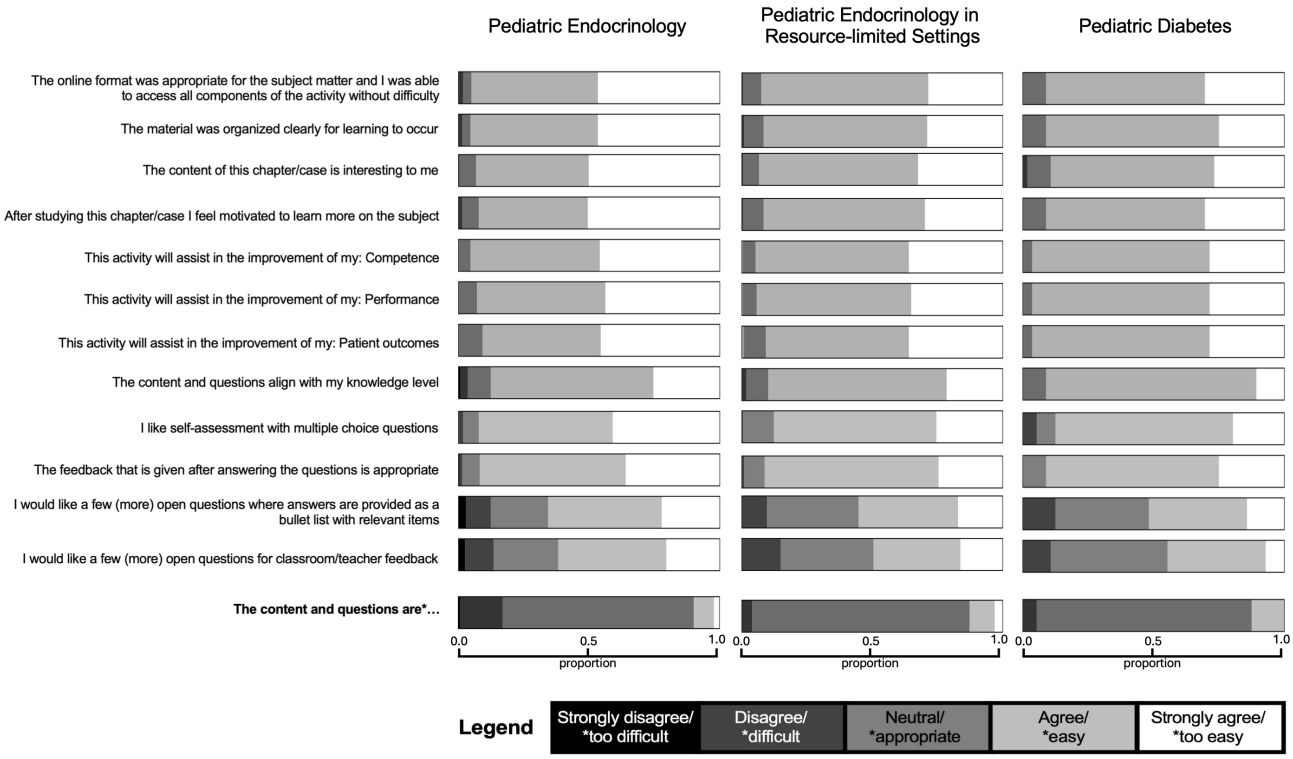
[a]For each question, response frequencies on a 5-point Likert scale were expanded into individual-level datasets to compute descriptive statistics, including total responses, median, mean, IQR, and SD. Differences in score distributions across modules were assessed using the Kruskal-Wallis H test, appropriate for ordinal, nonnormally distributed data. *P* values <.05 were considered statistically significant and are indicated by italics. Higher scores reflect more favorable evaluations.

[b]Likert scale responses: 1=strongly disagree; 2=disagree; 3=neutral; 4=agree; and 5=strongly agree.

[c]Likert scale responses: 1=too easy; 2=easy; 3=appropriate; 4=difficult; and 5=too difficult.

XSL•FO
RenderX

Figure 3. Stacked Likert distribution of responses across all survey questions. This figure displays the proportional distribution of Likert-scale responses, separate for each module (pediatric endocrinology, pediatric endocrinology in resource-limited settings, and pediatric diabetes). Each horizontal bar represents one survey question, subdivided into gray-shaded coded segments corresponding to the five Likert response categories. Segment widths reflect the proportion of total responses in each category. The plot illustrates the overall positive skew of responses across the curriculum, with most items demonstrating high levels of agreement (Likert 4‐5), while specific questions show greater dispersion, indicating more heterogeneous participant views.



Not all participants provided free-text feedback. Table 2 illustrates the distribution of free-text feedback received, categorized into four main categories, with the largest being various forms of appraisal. Improvement suggestions, such as requesting more interactive cases and algorithms, are encouraging. Layout issues are helpful to guide improvements, and similarly, critical comments about the difficulty of the content and excessive text are incentives to indicate more clearly the level of the course (core vs advanced), and replacing text-heavy chapters with bullet styles using artificial intelligence diligently.

XSL•FO
RenderX

Table . Free-text feedback from completed continuing medical education modules.

Domains	Overall (N=384), n (%)	PE[a] (n=197), n (%)	RLS[b] (N=26), n (%)	PD[c] (n=161), n (%)	Example comments
Praise or approval	274 (71.3)	145 (73.6)	21 (80.8)	108 (67.1)	Content is appropriate, good, excellent, perfect, helpful, informative, useful, adds knowledge and scientific thinking
Content suggestions	47 (12.2)	27 (13.7)	3 (11.5)	17 (10.6)	Needs more interactive cases, more algorithms, more detailed information, better explanations of multiple-choice answers, and more video content
Layout suggestions	28 (7.3)	13 (6.6)	2 (7.7)	13 (8.1)	Buttons need to be bigger, multiple replies in multiple-choice questions unclear, minimize menu left side of screen, and typos and missing sentences
Criticism or disapproval	35 (9.1)	12 (6.1)	0	23 (14.2)	Slightly *dry*, too difficult, too much text, too many references, and inappropriate for my work setting

[a]PE: pediatric endocrinology.

[b]RLS: PE in resource-limited setting.

[c]PD: pediatric diabetes.

Discussion

In an era where equitable access to CME is increasingly important, this study highlights the global uptake and positive reception of the ESPE CME–accredited e-learning modules made freely available in PE and PD, particularly among professionals in both high-resource and RLSs.

Principal Findings

Over a 3-year period, 567 learners from five continents completed postcourse surveys. The majority of participants were medical experts or trainees. Encouragingly, there were also learners practicing in RLSs. Feedback was overall very positive, with the majority of responses scoring 4 or 5 on a 5-point Likert scale. Free-text feedback suggested areas for improvement, including reducing text-heavy content and increasing interactivity and visual elements.

Implications of Findings

These findings support the value of free, guideline-based CME resources in PE and PD, particularly in expanding access to continuing education globally. The content is not only in line with the training requirements in PE and PD [15] but is also complementary to the requirements of the neonatology training curriculum [16]. Moreover, it is relevant for internists dealing with former pediatric patients with chronic congenital or acquired pediatric disorders [17].

A prerequisite of self-directed learning is that learners are internally motivated to take responsibility for their learning through a process in which they identify their own learning needs, use a variety of resources to meet these needs, and evaluate their learning to ensure that their learning needs have been met [18].

The high satisfaction ratings suggest that the ESPE platform meets these self-directed learning needs of a diverse professional audience. The engagement of health professionals, such as nurses and dietitians, also reflects the growing multidisciplinary nature of pediatric endocrine care. The feedback received provides actionable insights to improve the user experience and tailor content to different learning styles and clinical contexts.

Learners' free-text feedback has provided actionable insights for improving the ESPE e-learning modules. Common suggestions, such as adding more interactive cases, in particular when content is text-heavy; simplifying algorithms; and reducing text density, have informed content updates. The ESPE e-Learning website is moving to a new online platform, which allows for more visually engaging content and the use of artificial intelligence, which we anticipate will enhance learners' experience, engagement, and accessibility across varying levels of training. Categorizing feedback into themes also enables the committee to track trends and evaluate the impact of changes over time, ensuring the platform evolves in response to user needs.

XSL•FO
RenderX

Comparison With Prior Work

The EACCME implemented criteria for the accreditation of e-learning materials in 2009 and conducted an audit of CME–-Continuous Professional Development (events between 2017 and 2019), which included 385 e-learning materials [19]. However, no details on the assessment of these e-learning materials are provided, and studies focusing on accreditation of self-directed interactive asynchronous online learning designs are scarce [20,21]. Blomberg et al [22] highlighted a post-COVID-19 pandemic shift toward hybrid and digital CME formats, emphasizing accessibility and learner-centered design. Curran et al [20] concluded that an accredited, asynchronous e-learning module provides flexibility, accessibility, and scalability, particularly for geographically dispersed health professionals and described a structured approach to developing accredited e-learning modules, underscoring the importance of stakeholder engagement and iterative content development. Additionally, Tudor Car et al [23] demonstrated that digital problem–based learning is as effective as traditional methods for knowledge acquisition and may be superior for skill development.

Our results are consistent with earlier evaluations of the ESPE platform. Ng et al [4] reported sustained global engagement with the nonaccredited modules over a 10-year period. Kalaitzoglou et al [10] demonstrated the effectiveness of multilingual modules in improving access in RLSs, which aligns with our finding that 9% of users came from such contexts. Prior studies by Drop et al [5] and Kranenburg-van Koppen et al [6] emphasized the importance of interactive, case-based learning, a preference echoed in our learners' feedback. Our study adds to this literature by focusing specifically on the CME-accredited modules and providing structured postcourse feedback data.

This supports the notion that digital CME platforms, such as the ESPE e-learning portal, are well positioned to meet evolving global educational needs in PE and PD.

Strengths and Limitations

A key strength of this study is its global scope and inclusion of diverse professional roles, offering a comprehensive view of how the CME modules are used and perceived. The survey's mandatory nature ensured a high response rate. However, the study has limitations. It relies on self-reported data, which may be subject to response bias. The absence of pre- and post-module assessments limits our ability to measure learning outcomes objectively. Additionally, participation from certain regions (eg, China, Brazil, and Scandinavia) was lacking. Several factors may explain the lack of participation from certain regions. First, national CME accreditation policies vary widely; some countries may not recognize EACCME credits or may not require CME for specialist revalidation, reducing the incentive to engage with external platforms. Second, language barriers may persist despite the facilitating multilingual content, particularly if promotional materials or user interfaces are not localized. Third, awareness and visibility of the platform may be limited in regions where ESPE or ISPAD has fewer members. Fourth, cultural preferences for in-person or locally developed educational resources may influence uptake. Finally, technological access and digital literacy may still pose challenges in some areas, particularly in rural or underserved settings [24]. These insights underscore the need for targeted outreach strategies, local ambassadors, and region-specific promotional campaigns to improve awareness and adoption. Several authors emphasize the importance of partnerships with international societies, such as ESPE and ISPAD, and the support by their organizational infrastructure as of critical importance [4,22,25,26].

The accreditation credit points (European CME credit) are based on educational time spent and not on educational impact. In several papers, transitioning CME crediting from being time based to impact based is advocated with a plea for more rigorous quantitative or qualitative assessments. The assessment in the format of surveys should be appropriate to the intended goals or outcomes of the accredited education, measuring improvements in learner knowledge, skills, and competencies, professional performance, and ultimately in changes in patient health status [19,27,28]. The International Academy for Continuing Professional Development Accreditation created a shared set of international standards for accrediting CME [28]. De Leeuw et al [27] constructed a validated 7-step Medical E-Learning Evaluation Survey. The authors conclude that the Medical E-Learning Evaluation Survey is useful and understandable, adding value for e-learning creators. However, it is impossible to predict how motivated users will be to provide useful feedback. Therefore, a survey should always be accompanied by an in-depth focus group evaluation with the users. The adoption of international standards would establish a global framework, and substantive equivalency would enable international collaboration, reciprocity of credits, and broader access to high-quality education for health care professionals [28].

To enhance the platform's impact, future efforts should focus on developing pre- and postmodule assessments to evaluate knowledge gain, increasing interactivity through multimedia content and case simulations, strengthening outreach in underrepresented regions by partnering with local societies and ambassadors, and exploring the long-term impact of the modules on clinical practice and patient outcomes. The ESPE-ISPAD e-Learning platform has the potential to serve as a model for other specialties seeking to deliver equitable, high-quality CME globally.

Conclusions

The ESPE-ISPAD CME–accredited e-learning modules are widely accessed and well received by health care professionals globally, offering free, high-quality education in PE and PD. Moreover, the asynchronous self-paced designs afford greater convenience and flexibility for providers in accessing the CME courses at times that are best for them [26].

Learners reported high satisfaction with the content and delivery, and feedback highlighted opportunities to further enhance interactivity and accessibility.

These findings support the continued development and promotion of open-access CME platforms to improve global equity in specialist medical education. Ongoing efforts are needed to expand outreach, refine content based on user

XSL•FO
RenderX

feedback, and make accreditation more dynamic, evidence-based, and focused on educational impact.

Acknowledgments

The authors would like to thank contributing authors from many countries who have given many hours of their valuable time to provide this outstanding educational material. The authors gratefully acknowledge the support of the European Society for Paediatric Endocrinology (ESPE) and the International Society for Pediatric and Adolescent Diabetes (ISPAD) for facilitating the provision of the European Union of Medical Specialists (EUMS) CME certifications for the participants free of charge. They would also like to acknowledge QuestionPro for providing a nonprofit license. Apart from the free survey software, they also obtained access to QuestionPro's free survey templates.

Funding

The espe-elearning.org online learning website has received financial support from ESPE; ISPAD; Kyowa Kirin, Japan; Sandoz International, Germany; Alexion Pharmaceuticals, United States; and the European Union's Seventh Framework Program (FP7/2007-2013) under grant agreement 2014444. RLC Project has received financial support from ESPE, ISPAD, and Eli Lilly & Co, United States.

Data Availability

The datasets generated or analyzed during this study are available from the corresponding author on reasonable request.

Authors' Contributions

Conceptualization: YGVDZ, CVW-DV, JI
Data curation: CVW-DV
Formal analysis: CVW-DV, YGVDZ, JI
Funding acquisition: SMN, SLSD, JI
Methodology: YGVDZ, CVW-DV, JI
Project administration: CVW-DV
Software: CVW-DV, JI
Validation: YGVDZ, JI
Visualization: CVW-DV, JI, YGVDZ
Writing – original draft: YGVDZ, CVW-DV, SLSD, JI
Writing – review & editing: all authors

Conflicts of Interest

SMN is the current Chair of both ESPE and ISPAD E-learning Committees.
AMB is a past chair of the ESPE e-Learning Committee. SLSD is a past member, treasurer, and president of ESPE Council; a former member of the ESPE Caucasus and Central Asia School; and a former chief editor of espe-elearning.org

Multimedia Appendix 1
Content of the continuing medical education feedback survey.
[DOCX File, 27 KB - mededu_v12i1e67332_app1.docx]

Multimedia Appendix 2
Continuing medical education courses completed.
[XLSX File, 10 KB - mededu_v12i1e67332_app2.xlsx]

References

1. ESPE e-Learning. URL: www.espe-elearning.org [accessed 2025-09-30]
2. Grijpink-van den Biggelaar K, Drop SLS, Schuwirth L. Development of an e-learning portal for pediatric endocrinology: educational considerations. Horm Res Paediatr 2010;73(3):223-230. [doi: 10.1159/000284366] [Medline: 20197677]
3. Ng SM, de Beaufort C, Becker D, Wolfsdorf JI. A global e-learning initiative for pediatric diabetes and endocrinology: introduction and description. Pediatr Diabetes 2021 Jun;22(4):692-694. [doi: 10.1111/pedi.13214] [Medline: 33904237]
4. Ng SM, Kalaitzoglou E, Utari A, et al. Ten-year experience of a global and freely accessible e-learning website for pediatric endocrinology and diabetes. Horm Res Paediatr 2023;96(4):366-375. [doi: 10.1159/000527984] [Medline: 36349753]
5. Drop SLS, Kranenburg L, Reerds S, Grijpink K, Iotova V. The ESPE e-learning webportal: a global tool for instruction and formative assessment of pediatric endocrinology. Scr Sci Med 2014;46(3):11-18. [doi: 10.14748/ssm.v46i3.760]

XSL•FO
RenderX

6. Kranenburg-van Koppen LJC, Grijpink-van den Biggelaar K, Drop SLS. An interactive E-learning portal in pediatric endocrinology: practical experience. Pharmacy (Basel) 2013;1(2):160-171. [doi: 10.3390/pharmacy1020160]
7. Alzahrani M. Traditional learning compared to online learning during the COVID-19 pandemic: lessons learned from faculty's perspectives. Sage Open 2022 Apr;12(2):21582440221091720. [doi: 10.1177/21582440221091720]
8. Brika SKM, Chergui K, Algamdi A, Musa AA, Zouaghi R. E-learning research trends in higher education in light of COVID-19: a bibliometric analysis. Front Psychol 2022;12:762819. [doi: 10.3389/fpsyg.2021.762819] [Medline: 35308075]
9. Cook DA, Levinson AJ, Garside S, Dupras DM, Erwin PJ, Montori VM. Internet-based learning in the health professions: a meta-analysis. JAMA 2008 Sep 10;300(10):1181-1196. [doi: 10.1001/jama.300.10.1181] [Medline: 18780847]
10. Kalaitzoglou E, Majaliwa E, Zacharin M, et al. Multilingual global E-learning pediatric endocrinology and diabetes curriculum for front line health care providers in resource-limited countries: development study. JMIR Form Res 2020 Nov 5;4(11):e18555. [doi: 10.2196/18555] [Medline: 33151156]
11. Eysenbach G. Improving the quality of Web surveys: the Checklist for Reporting Results of Internet E-Surveys (CHERRIES). J Med Internet Res 2004 Sep 29;6(3):e34. [doi: 10.2196/jmir.6.3.e34] [Medline: 15471760]
12. Smidt A, Balandin S, Sigafoos J, Reed VA. The Kirkpatrick model: a useful tool for evaluating training outcomes. J Intellect Dev Disabil 2009 Sep;34(3):266-274. [doi: 10.1080/13668250903093125] [Medline: 19681007]
13. Heydari MR, Taghva F, Amini M, Delavari S. Using Kirkpatrick's model to measure the effect of a new teaching and learning methods workshop for health care staff. BMC Res Notes 2019 Jul 10;12(1):388. [doi: 10.1186/s13104-019-4421-y] [Medline: 31292006]
14. Khan T, Abimbola S, Kyobutungi C, Pai M. How we classify countries and people-and why it matters. BMJ Glob Health 2022 Jun;7(6):e009704. [doi: 10.1136/bmjgh-2022-009704] [Medline: 35672117]
15. Busiah K, Peet A, Tornese G, et al. The 2021 European training requirements in paediatric endocrinology and diabetes. Horm Res Paediatr 2021;94(11-12):441-447. [doi: 10.1159/000520073] [Medline: 34638127]
16. Santoro D, Zibulsky DA, Roehr CC, et al. Meeting the need for effective and standardized neonatology training: a pan-European Master's Curriculum. Pediatr Res 2024 Oct;96(5):1195-1200. [doi: 10.1038/s41390-024-03182-8] [Medline: 38702380]
17. Luger A, Durkan M, Christ-Crain M, Marques P. European society of Endocrinology curriculum and training recommendation in Endocrinology. Eur J Endocrinol 2024 Oct 29;191(5):R71-R77. [doi: 10.1093/ejendo/lvae141] [Medline: 39582372]
18. Zhu M, Bonk CJ, Berri S. Fostering self-directed learning in MOOCs: motivation, learning strategies, and instruction. Online Learn 2022;26(1):153-173. [doi: 10.24059/olj.v26i1.2629]
19. Sideris M, Rallis KS, Nicolaides M, et al. The UEMS experience in continuous medical education accreditation process: a "quo vadis" analysis of our global database. Ann Med Surg (Lond) 2024 Feb;86(2):689-696. [doi: 10.1097/MS9.0000000000001583] [Medline: 38333329]
20. Curran V, Glynn R, Whitton C, Hollett A. An approach to the design and development of an accredited continuing professional development e-learning module on Virtual Care. JMIR Med Educ 2024 Aug 8;10:e52906. [doi: 10.2196/52906] [Medline: 39119741]
21. Fiuzzi M. Outcomes and observations of on-line CME activities during the pandemic. J CME 2023;12(1):2167286. [doi: 10.1080/28338073.2023.2167286] [Medline: 36969485]
22. Blomberg D, Stephenson C, Atkinson T, et al. Continuing medical education in the post COVID-19 pandemic era. JMIR Med Educ 2023 Nov 15;9:e49825. [doi: 10.2196/49825] [Medline: 37966881]
23. Tudor Car L, Kyaw BM, Dunleavy G, et al. Digital problem-based learning in health professions: systematic review and meta-analysis by the digital health education collaboration. J Med Internet Res 2019 Feb 28;21(2):e12945. [doi: 10.2196/12945] [Medline: 30816846]
24. Fraillon J, Rožman M, editors. IEA International Computer and Information Literacy Study 2023: Assessment Framework: Springer; 2025. [doi: 10.1007/978-3-031-61194-0]
25. Craig A, Lawford H, Miller M, et al. Use of technology to support health care providers delivering care in low- and lower-middle-income countries: systematic umbrella review. J Med Internet Res 2025 Jun 18;27:e66288. [doi: 10.2196/66288] [Medline: 40533075]
26. Roth MJ, Maggio LA, Costello JA, Samuel A. E-learning interventions for quality improvement continuing medical education—a scoping review. J Contin Educ Health Prof 2025 Jan 1;45(1):35-43. [doi: 10.1097/CEH.0000000000000564] [Medline: 39028318]
27. de Leeuw RA, Westerman M, Walsh K, Scheele F. Development of an instructional design evaluation survey for postgraduate medical e-learning: content validation study. J Med Internet Res 2019 Aug 9;21(8):e13921. [doi: 10.2196/13921] [Medline: 31400102]
28. Regnier K, Campbell C, Griebenow R, et al. Standards for substantive equivalency between continuing professional development/continuing medical education (CPD/CME) accreditation systems. J Eur CME 2021 Jan 29;10(1):1874644. [doi: 10.1080/21614083.2021.1874644] [Medline: 33552679]

XSL•FO
RenderX

Abbreviations

CHERRIES: Checklist for Reporting Results of Internet E-Surveys
CME: continuing medical education
EACCME: European Accreditation Council for Continuing Medical Education
ESPE: European Society for Paediatric Endocrinology
ISPAD: International Society for Pediatric and Adolescent Diabetes
PD: pediatric diabetes
PE: pediatric endocrinology
RLS: resource-limited setting

Edited by B Lesselroth; submitted 12.Oct.2024; peer-reviewed by J Sarvestan, J Woldorf, Y Shen; revised version received 24.Nov.2025; accepted 03.Dec.2025; published 30.Jan.2026.

Please cite as:
van der Zwan YG, van Wijngaard-de Vugt C, Abu-Libdeh A, Kalaitzoglou E, Karabouta Z, Drop SLS, Boot AM, Ng SM, Idkowiak J Global Learner Feedback on Continuing Medical Education–Accredited e-Learning Modules in Pediatric Endocrinology and Diabetes: Cross-Sectional Study
JMIR Med Educ 2026;12:e67332
URL: https://mededu.jmir.org/2026/1/e67332
doi:10.2196/67332

© Yvonne G van der Zwan, Conny van Wijngaard-de Vugt, Abdulsalam Abu-Libdeh, Evangelia Kalaitzoglou, Zacharoula Karabouta, Stenvert L S Drop, Annemieke M Boot, Sze May Ng, Jan Idkowiak. Originally published in JMIR Medical Education (https://mededu.jmir.org), 30.Jan.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Education, is properly cited. The complete bibliographic information, a link to the original publication on https://mededu.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Upskilling in Healthy Longevity Medicine and Its Association With Physicians' Implementation Intent and Self-Reported Clinical Confidence: Cross-Sectional Observational Study

Evelyne Bischof[1,2], Dr med, MD, PhD; Dominika Wilczok[3,4]; James L Kirkland[5], MD, PhD, Dr Med; Bhirau Wilaksono[6,7], MD; Christine Yuan Huang[8], MD; Suwanna Suwannaphong[9], MD; Wanviput Sanphasitvong[9], MD; Dalila Čamdžić[10], MD; Carolina Hernandez[11], BSc; Yoko Madea[12], MD; Hidekazu Yamada[13], MD; Melissa Alexandre Fernandes[14], MD; Ricardo Gaminha Pacheco[15], MD, MBA; Fabiano M Serfaty[16], MD, PhD; Fernanda Calvo-Fortes[17], MSc; Amit Goldman[18], EMBA; Andrea B Maier[19,20,21], MD; Alexey Moskalev[22], PhD; Morten Scheibye-Knudsen[23], MD, PhD; Alex Zhavoronkov[24], PhD

[1]Sheba Longevity Center, Sheba Medical Center, Gray Faculty of Medical & Health Sciences, Tel Aviv University, Sheba Road 2, Ramat Gan, Israel
[10]Sarajevo School of Science and Technology, Sarajevo, Bosnia and Herzegovina
[11]Departent of Biomedicine, University of Barcelona, Barcelona, Spain
[12]Madea Clinic, Okayama, Japan
[13]Anti-Aging Center, Kindai University, Osaka, Japan
[14]Novartis Institutes for BioMedical Research, Basel, Switzerland
[15]Insilico Medicine Hong Kong Ltd, Hong Kong, China (Hong Kong)
[16]Departent of Biomedicine, Universidade do Estado do Rio de Janeiro, Rio de Janeiro, Brazil
[17]School of Nursing, Federal University of Minas Gerais, Belo Horizonte, Brazil
[18]Saïd Business School, University of Oxford, Oxford, United Kingdom
[19]Healthy Longevity Translational Research Program, Yong Loo Lin School of Medicine, National University of Singapore, Singapore
[2]Shanghai University of Medicine and Health Sciences, Shanghai, China
[20]NUS Academy for Healthy Longevity, Yong Loo Lin School of Medicine, National University of Singapore, Singapore
[21]Department of Human Movement Sciences, Faculty of Behavioral and Movement Sciences, Amsterdam Movement Sciences, Vrije Universiteit, Amsterdam, The Netherlands
[22]Petrovsky Russian Research Center for Surgery, Moscow, Russian Federation
[23]Center for Healthy Aging, Department of Cellular and Molecular Medicine, University of Copenhagen, Copenhagen, Denmark
[24]Insilico Medicine US Inc, Cambridge, MA, United States
[3]Duke University, Durham, NC, United States
[4]Duke Kunshan University, Kunshan, Jiangsu, China
[5]Center for Advanced Gerotherapeutics, Division of Endocrinology and Metabolism, Department of Medicine, Cedars Sinai Health Sciences University, Los Angeles, CA, United States
[6]Center for Healthy Longevity, Faculty of Medicine, Universitas Negeri Makassar, Indonesia
[7]Department of Biochemistry, Faculty of Medicine, Universitas Negeri Makassar, Indonesia
[8]Quantum Life Inc, Hong Kong, China (Hong Kong)
[9]VitalLife, Bumrungrad International Hospital, Bangkok, Thailand

Corresponding Author:
Evelyne Bischof, Dr med, MD, PhD
Sheba Longevity Center, Sheba Medical Center, Gray Faculty of Medical & Health Sciences, Tel Aviv University, Sheba Road 2, Ramat Gan, Israel

Related Article:

This is a corrected version. See correction statement: https://mededu.jmir.org/2026/1/e114811

Abstract

Background: Structured educational programs for physicians in healthy longevity medicine (HLM) remain scarce. No published data yet document the impact of longevity-focused medical education on physicians. This study assesses the ramifications of the

XSL•FO
RenderX

HLM curriculum, certified by the American Council for Continuing Medical Education, on physicians' confidence in their knowledge of HLM and clinical practice.

Objective: This study aimed to evaluate the impact of accredited HLM education on physicians' confidence in knowledge and practice patterns, examining self-reported integration of HLM principles, professional attitudes, and career trajectories to determine the translational value of structured curricula in the emerging medical discipline.

Methods: A cross-sectional online survey was conducted between March and April 2024 among physicians who had completed accredited HLM courses between January 2023 and February 2024. Invitations were sent globally to 590 eligible physicians; trainees and students were excluded. A total of 113 (19.2%) respondents completed the survey and were included in the analysis. The survey assessed self-reported changes in clinical implementation, confidence in HLM-related knowledge, and professional attitudes following course completion. Descriptive statistics and logistic regression analyses were performed (P<.05).

Results: Respondents represented 42 nationalities and were primarily trained in family medicine (n=31, 27.4%) and internal medicine (n=18, 15.9%). Overall, 96.5% (n=99) of the respondents reported increased confidence in HLM-related knowledge, with 47.8% (n=55) indicating substantial improvement. More than half of the respondents (n=63, 55.8%) reported integrating HLM principles into routine patient assessments, and 80.5% (n=91) of the respondents reported more frequent discussions related to health span–focused care. In addition, 23% (n=26) of the respondents initiated aging biomarker testing, 48.7% (n=55) increased the testing frequency, 52.2% (n=59) reported a shift in their perspective on aging, and 73.5% (n=83) anticipated full integration of HLM into mainstream medicine. Physicians practicing in specialized care demonstrated higher odds of reporting increased confidence in HLM knowledge compared with those in primary and preventive care (odds ratio 4.46, 95% CI 1.55 - 12.79; P=.005).

Conclusions: Accredited education in HLM is associated with enhanced confidence in HLM knowledge, increased clinical engagement with HLM practices, and a shift in aging-related care paradigms. These findings underscore the critical role of structured HLM curricula in bridging the translational gap between geroscience and everyday medical practice. Nevertheless, systemic health care barriers impede widespread implementation, warranting policy-level strategies to support health span–oriented education and care models.

(JMIR Med Educ 2026;12:e83779) doi:10.2196/83779

KEYWORDS

longevity; survey; continuing medical education; clinical training; professional development

Introduction

As populations age globally, there is an increasing need for rigorous, evidence-based education to prepare health care professionals to manage the complex, multimorbidity-driven needs of older adults. Healthy longevity medicine (HLM) represents an emerging clinical framework aimed at optimizing health span through proactive, evidence-based diagnostics and interventions grounded in geroscience—the discipline that studies the biological mechanisms of aging; their role in age-related chronic diseases; and the development of interventions to delay, prevent, or reverse these processes [1]. While geroscience has advanced rapidly over the past decade, translating these breakthroughs into clinical practice remains limited by a lack of data from prospective, well-designed clinical trials; formalized education; clinical practice protocols; and integration into existing health care systems [2]. Bridging this translational gap is essential to ensure that the benefits of geroscience reach patients in a timely, equitable, and scalable manner. Parallel to these scientific advances, new diagnostic and therapeutic approaches relevant to HLM are emerging, including dual-purpose interventions that target fundamental aging mechanisms while addressing disease-specific outcomes [3], as well as biological aging clocks, increasingly incorporating multimodal and artificial intelligence–based methods [4]. These developments expand the potential scope of care but also introduce complexity that exceeds conventional, disease-based training models. As a result, physicians require dedicated education not only to apply these tools responsibly in clinical settings but also to contribute meaningfully to clinical trial design, interpretation, and translational research efforts. At the same time, the rapid expansion of the "healthy longevity" field has been accompanied by a proliferation of nonstandardized, commercially driven, and sometimes misleading information, which can blur the distinction between evidence-based preventive medicine and unvalidated interventions. Therefore, structured educational initiatives grounded in established geriatric, preventive, and translational aging science may play a critical role in equipping clinicians with the skills needed to critically appraise emerging evidence, apply validated interventions appropriately, and safeguard patients against misinformation. Strengthening evidence-based training in this domain is essential not only for clinical quality and patient safety but also for maintaining professional and public trust as the field continues to evolve.

Despite this need, formal medical education in HLM remains scarce. Since 2020, the Longevity Education Hub has provided the first and only American Council for Continuing Medical Education–accredited educational framework designed specifically for practicing physicians and health care professionals. The courses represent a systematic attempt to equip physicians with the tools necessary to apply knowledge of the biology of aging in a clinical setting. Importantly, this curriculum has been adopted into the required coursework of medical schools within 2 major universities (State University of Massakar in Indonesia and King Mongkut's Institute of

XSL•FO
RenderX

Technology Ladkrabang in Thailand), marking a historic integration of HLM into formal medical education.

To expand physician education beyond the core continuing medical education (CME) courses, the first conference event dedicated to HLM, Longevity Medicine Day, was held in 2021 at the Aging Research and Drug Discovery Conference. This event became an annual feature and was expanded into the Healthy Longevity Medicine Track by 2024 [5,6]. In response to the growing interest in HLM among clinicians, the Healthy Longevity Medicine Society was established in 2022, providing a structured professional body to support clinicians and standardize practices within the field [7]. These milestones provided a theoretical and educational foundation for integrating HLM into clinical settings. Parallel to these scientific developments, institutional recognition of HLM has grown. In 2024, the Department of Health Abu Dhabi became the first regulatory body to officially recognize HLM protocols and initiate their integration into clinical settings [8].

Nevertheless, empirical data evaluating how such educational initiatives influence physicians' knowledge, attitudes, and clinical implementation remain limited. This gap is particularly consequential as HLM gains visibility in the consumer health sector, where the absence of validated clinical standards increases the risk of unproven interventions and commercial exploitation. Previous experience in other emerging or high-risk domains, such as pain management, demonstrates that mandatory and structured physician education can improve clinical confidence and promote safer practice patterns; for example, California's CME mandate in pain management was associated with self-reported positive practice changes in up to 90% of physicians at follow-up [9]. More broadly, CME-based interventions have been shown to enhance physician knowledge, confidence, and selected patient outcomes across medical disciplines [10,11].

Understanding this dynamic is crucial for guiding the next phase of curriculum development, policy adoption, and the broader institutionalization of HLM. In addition, any educational effort, particularly in emerging medical fields, requires systematic quality control to ensure its effectiveness and scalability. To meet this need, we conducted a cross-sectional study to evaluate physicians who have undergone such formal education in HLM. This study presents, to the best of our knowledge, the first global survey of physicians who have completed the HLM curriculum, evaluating its perceived impact, clinical integration, and barriers that remain. In doing so, we offer empirical insights into how frontline clinicians are shaping the transformation of geroscience into a defined, evidence-based clinical discipline.

Methods

Ethical Considerations

This study involved an anonymous, voluntary online survey of physicians who had completed accredited HLM courses.

In accordance with institutional and local policies, formal ethics board approval was not sought, as the study did not constitute human participant research requiring review. Institutional review board approval was not sought, as the study was designed and conducted as an anonymous, voluntary survey of physicians that collected no patient data, no sensitive personal information, and no identifiable data and posed no risk to participants. We followed the EC Horizon and the US Department of Health and Human Services (45 CFR 46.104(d)(2)) regulation [12].

No personal identifiers were collected. Informed consent was implied by participants' decision to complete the survey, and no financial or material compensation was provided. This study was conducted in accordance with the STROBE (Strengthening the Reporting of Observational Studies in Epidemiology; Checklist 1) guidelines for cross-sectional studies.

Study Design

This cross-sectional observational study was conducted between March and April 2024. Physicians who completed the Longevity Education Hub's online Longevity Medicine 101 and/or Longevity Medicine 201 courses self-administered and remotely completed the anonymous online survey assessing changes in clinical practice, confidence in HLM knowledge, and perspectives on HLM.

Participants

The study population was derived from all individuals who enrolled in the Longevity Medicine 101 and 102 courses offered through the Longevity Education Hub on Teachable between December 1, 2021, and March 1, 2024 (N=6464). Participants were excluded if they did not complete all course modules or did not pass the final quiz (n=3241, 50.1%), leaving 3223 (49.9%) course completers. Among them, those who did not report holding a Doctor of Medicine title upon course enrollment were excluded (n=1561, 24.1%), resulting in 1662 (25.7%) physicians. Junior physicians in training or medical students were further excluded (n=1072, 16.6%) because HLM is an emerging discipline without direct pathways to board-certified specialization; therefore, entry into HLM practice typically requires substantial clinical experience in one's primary discipline—an experience that trainees, by definition, lack. Therefore, trainees were excluded. The study survey was sent to all remaining 590 (9.1%) qualifying practicing physicians. Figure 1 illustrates the process of selecting the participants. The participants did not receive any monetary compensation for their responses.

Demographic data, including nationality and medical specialization, were collected. For analytical purposes, medical specialties indicated by the participants were grouped according to their primary role within the health care delivery continuum. Primary and preventive care and general practice (PPG) encompass specialties characterized by first-contact patient care, longitudinal responsibility, and a broad scope of practice, with a strong emphasis on prevention, risk stratification, and comprehensive health management. This group included general practice, family medicine, and related primary care disciplines, as displayed in Table 1. In contrast, specialized care and advanced procedures (SPEC) comprised specialties predominantly engaged through referral, with a more focused scope of practice centered on specific organ systems, disease categories, or procedural interventions.

XSL•FO
RenderX

Figure 1. Flowchart for participant recruitment. MD: Doctor of Medicine.

XSL•FO
RenderX

Table . Geographic distribution (derived from nationality), medical specialization, and professional status of survey respondents (N=6464 and n=113).

Item	Values, n (%)
Region (derived from nationality)	
Asia	24 (21.2)
Australia and Oceania	6 (5.3)
Africa	2 (1.8)
Europe	45 (39.8)
South America	11 (9.7)
North America	17 (15)
Middle East	8 (7)
Primary and preventive care and general practice	
Family medicine	31 (27.4)
Internal medicine	18 (15.9)
Public health	3 (2.6)
Integrative medicine	2 (1.8)
Community medicine	1 (0.9)
Functional medicine	1 (0.9)
Precision preventive medicine	7 (6.2)
Specialized care and advanced procedures	
Oncology	5 (4.4)
Orthopedics	3 (2.6)
Dermatology	3 (2.6)
Lifestyle medicine	3 (2.6)
Pain management	3 (2.6)
Endocrinology	3 (2.6)
Cardiology	3 (2.6)
Sports medicine	2 (1.8)
Clinical nutrition	1 (0.9)
Neuropsychiatry	1 (0.9)
Obstetrics and gynecology	1 (0.9)
Hematology	2 (1.8)
Regenerative medicine	1 (0.9)
Plastic surgery	1 (0.9)
Neurology	3 (2.6)
Urology	1 (0.9)
Psychiatry	2 (1.8)
Psycho-neuro-endocrino-immunology	1 (0.9)
MD[a]+MBA[b] in hospital administration	1 (0.9)
Medical genetics	3 (2.6)
Neurosurgery	1 (0.9)
Physical medicine and rehabilitation	1 (0.9)
General surgery	4 (3.5)
Transplantology	1 (0.9)
Practice status	

XSL•FO
RenderX

Item	Values, n (%)
Practicing MD (public)	34 (30.1)
Practicing MD (private)	61 (54)
Practicing MD (<40% clinical time)	7 (6.2)
Not practicing MD	11 (9.7)

[a]MD: Doctor of Medicine.

[b]MBA: Master of Business Administration.

Intervention

Longevity Education Hub's Longevity Medicine 101 and Longevity Medicine 201 courses provide structured, free-of-charge, accredited education in HLM for physicians. The courses are delivered in an online, asynchronous format with a total instructional duration of approximately 2 hours and 40 minutes and conclude with a final assessment requiring a minimum score of 80% for certification. The curriculum includes an introduction to HLM and core definitions; the epidemiology of aging; the scientific basis of aging from biogerontology to clinical translation; definitions and biological mechanisms of aging; therapeutic interventions and innovations in HLM, including geroprotectors and psychological aspects of aging; methods for measuring aging, including biological aging, biomarkers of aging, biochronology, and aging clocks; and the clinical application of HLM, encompassing diagnostic frameworks, structured clinical protocols, and integration of HLM into clinical practice. Emerging applications of artificial intelligence in longevity research and clinical workflows are also introduced.

Assessment

A 20-item survey instrument was developed using a structured 3-phase process. First, initial items were generated based on a comprehensive literature review and consultation with subject-matter experts in the field of HLM. This was followed by content validation conducted by a panel of 4 independent experts in medical education and HLM, ensuring the relevance and clarity of the survey items. Finally, the instrument was pilot-tested with a sample of 20 physicians, although responses from the pilot were not included in the final analysis. The finalized survey demonstrated high internal consistency, with a Cronbach α of 0.72. The mandatory questions took up to 10 minutes to complete. The primary outcomes of this study were self-reported confidence in HLM knowledge after the training, measured using a Likert scale, and the integration of HLM principles into clinical assessments, assessed through frequency-based response options. Secondary outcomes included the frequency of initiating longevity-related discussions with patients, the adoption and frequency of aging biomarker testing, the perceived efficacy of HLM education in addressing demographic challenges, shifts in perspectives on aging and longevity, and anticipated growth of HLM as a medical field. The survey also assessed participants' interest in further HLM education, their support for the accreditation of HLM as a specialty, and their willingness to join an HLM consortium. The last question was nonmandatory and open-ended, inviting participants to provide any additional comments.

The survey was distributed via email with a link to a Google Forms questionnaire. The distribution consisted of 2 send-outs spaced 2 weeks apart to maximize response rates. Two weeks after the initial invitation, all recipients received an automatic reminder, regardless of whether they had already completed the survey. To prevent multiple submissions, each participant was permitted to complete the survey only once.

Statistical Analysis

Statistical analyses were conducted in R (version 4.2.0; R Development Core Team Foundation for Statistical Computing). Data quality checks confirmed valid ranges and internal consistency of responses. Because the survey platform (Google Forms) required completion of all essential items before submission, there were no missing values for mandatory variables, and no imputation or casewise exclusion was performed. Descriptive statistics (counts and percentages) summarized demographics and survey responses overall and by group. Between-group differences (SPEC vs PPG) in course-related outcomes were assessed using separate logistic regression models (1 model per outcome). The group was included as the primary independent variable, with PPG serving as the reference category. Model results were reported as odds ratios (ORs), 95% CIs, and *P* values. Statistical significance was set at $P<.05$.

Results

The survey was sent to 590 eligible physicians, of whom 113 (19.2%) completed the survey (Figure 1). Table 1 summarizes the demographic and professional characteristics of the respondents, including nationality, medical specialty, and practice setting. The sample included physicians of 42 nationalities. Regional groupings follow standard continental divisions, with the Middle East presented separately as a commonly recognized geopolitical region in international health research. Regionally (based on nationality), the largest proportions of physicians were European (n=45, 39.8%), Asian (n=24, 21.2%), North American (n=17, 15%), and Middle Eastern (n=9, 8%). The most prevalent specialties were family medicine (n=31, 27.4%) and internal medicine (n=18, 25.6%). Regarding practice setting, 54% (n=61) of the respondents worked in private practice and 30.1% (n=34) in public settings.

Regarding the primary outcomes, the analysis of the responses indicated positive feedback on the impact of the course on participants' confidence in their knowledge of HLM. A total of 99 (96.5%) of the 113 respondents reported an increase in confidence in their knowledge, with 54 (47.8%) respondents noting a significant improvement, as shown in Figure 2. In terms

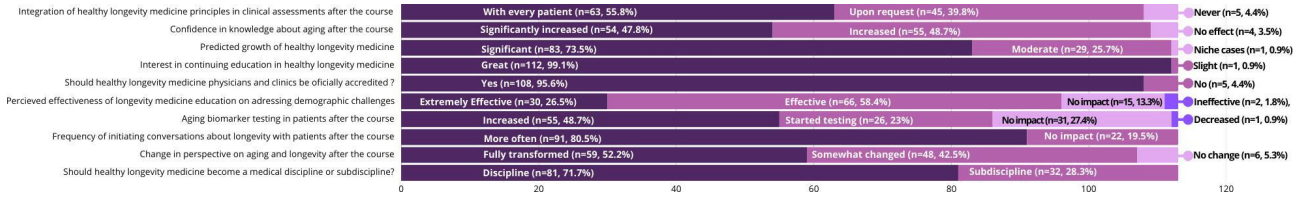
XSL•FO
RenderX

of clinical integration, 63 (55.8%) participants reported implementing HLM principles into routine assessments for all patients, with 45 (39.8%) participants applying these principles on special request (Figure 2).

Regarding secondary outcomes, 59 (52.2%) respondents indicated that the courses fully transformed their perspective on aging and healthy longevity (Figure 2). In addition, 26 (23%) participants who had not previously engaged in aging biomarker testing began incorporating it into their practice, while 55 (48.7%) respondents increased the frequency of such tests (Figure 2). Most of the participants (84.9%) perceived HLM education as an effective or extremely effective means of addressing demographic challenges (Figure 2). Furthermore, 91 (80.5%) respondents reported engaging in more frequent discussions about healthy longevity with their patients after completing the course (Figure 2). Looking ahead, 83 (73.5%) participants expected HLM to become fully integrated into mainstream medicine, while 29 (25.7%) participants believed it would remain in specialized clinics (Figure 2). The findings also show strong support for the formal accreditation of physicians and clinics specializing in HLM, with 108 (95.6%) respondents in favor (Figure 2). Most of the participants (n=112, 99.1%) participants expressed interest in further HLM education (Figure 2). Finally, 71.7% (n=81) of respondents believed that HLM will eventually be recognized as an official medical discipline (Figure 2).

Physicians reported various career impacts following HLM education. Most respondents (n=39, 34.5%) reported joining a group of HLM physicians. A notable proportion of respondents (n=36, 31.9%) indicated that the education did not result in significant changes to their medical career. In addition, 24 (21.2%) respondents reported opening a new HLM clinic, 4 (3.5%) joined a pre-existing HLM clinic, and 10 (8.9%) participants did both.

Figure 2. Impact of accredited healthy longevity medicine (HLM) education.



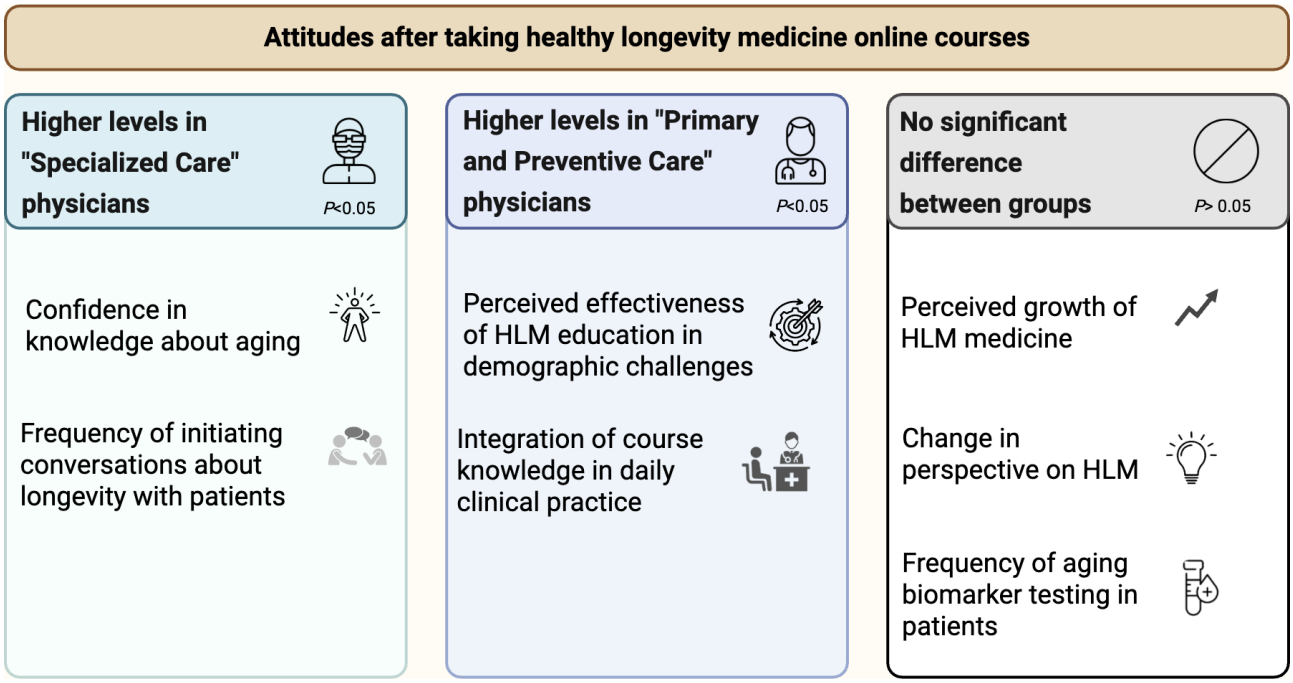
Logistic regression analysis revealed differences between groups' key educational takeaways, as illustrated in Figure 3. Confidence gain in HLM knowledge was significant in the SPEC group (OR 4.46, 95% CI 1.55 - 12.79; *P*=.005), with SPEC physicians having almost 4.5 times higher odds of reporting increased confidence after HLM education than PPG physicians. A higher frequency of HLM-related discussions was associated with a nearly threefold greater likelihood of occurring in the SPEC group (OR 2.98, 95% CI 1.30 - 6.83; *P*=.01). In contrast, the PPG physicians reported a greater perceived effectiveness of HLM education (OR 0.29, 95% CI 0.12 - 0.75; *P*=.01). The PPG group also had a greater integration of healthy longevity principles into clinical assessments (OR 0.70, 95% CI 0.53 - 0.91; *P*=.009).

Although none of the additional variables reached statistical significance, several trends emerged. Physicians in the SPEC group were more likely to report a more positive change in perspective on aging and healthy longevity after completing the courses (OR=0.389, 95% CI 0.13 - 1.14; *P*=.09), as well as to predict greater growth of HLM (OR 0.73, 95% CI 0.26 - 2.04; *P*=.55). However, PPG physicians reported more frequent testing of biomarkers relevant to aging processes (OR 1.49, 95% CI 0.71 - 3.14; *P*=.29).

The open-ended question regarding any additional comments the participants could have about HLM revealed multiple systemic barriers that continue to hinder its integration into clinical practice. The participants' answers revealed them to be financial disincentives embedded in fee-for-service models that disfavor preventive, time-intensive care; regional disparities in infrastructure and adoption of preventive frameworks; and the lack of standardized clinical guidelines translating scientific advances into routine medical workflows. The participants noted that the absence of formal regulatory recognition, limited accreditation pathways, and uneven diagnostic access further compound the challenge.

XSL•FO
RenderX

Figure 3. Differences in effects of healthy longevity medicine (HLM) education between primary and preventative care physicians and specialized care physicians.



Discussion

Principal Findings

This study found that accredited, online HLM courses are associated with significant improvements in physician-reported confidence in the knowledge of HLM and increased clinical integration of healthy longevity–focused practices.

More than half of the participants (n=63, 55.8%) reported routinely integrating HLM principles into assessments for all patients, while the rest applied these principles selectively, suggesting that educational gains translated into context-dependent, clinical behaviors. Notably, nearly all participants (n=109, 96.5%) reported increased confidence in HLM-related knowledge, with about half (n=54, 47.8%) reporting substantial improvement. Physicians in the SPEC group were 4 times more likely to report increased confidence than those in the PPG group. This difference may reflect how the course content aligned with routine clinical approaches across practice settings. Because several HLM concepts overlap with preventive and systems-based care commonly used in primary care, some PPG physicians may have perceived smaller relative gains in knowledge confidence despite completing the same course. The observed differences between SPEC and PPG physicians suggest that tailored educational approaches may be necessary to address the unique needs of diverse clinical sectors. The results support the hypothesis that structured online education can enhance physician engagement with HLM, suggesting a potential role for scalable digital learning in bridging knowledge gaps in HLM.

Beyond the primary outcomes, the findings indicate wider shifts following HLM education. More than half of the respondents (n=59, 55.2%) reported a complete transformation in their perspective on aging and healthy longevity, and the majority (n=96, 84.9%) perceived HLM education as an effective or extremely effective approach to addressing demographic challenges. Notably, 80.5% (n=91) of all physicians reported an increase in patient-facing healthy longevity conversations postcourse, indicating a meaningful shift in engagement strategies. Moreover, a general increase in biomarker use observed in both groups (n=55, 48.7% of all respondents increased their use) aligns with a broader movement within the HLM field toward earlier risk stratification and individualized preventive strategies. However, there is increasing recognition that biomarkers of aging are currently more robust as research and monitoring tools than as definitive clinical decision instruments [13-15]. Recent consensus efforts emphasize their value for standardizing outcomes in intervention studies and highlight the necessity of using a variety of biomarkers for accurate reflection of biological aging [16]. Despite the strong educational uptake, belief in the future growth of HLM and changes in perspectives on aging did not reach statistical significance in the regression model. Critically, nearly all respondents (n=108, 95.6%) support formal accreditation of physicians and clinics practicing HLM, indicating widespread professional demand for institutional recognition of this emerging field and underscoring the role of structured education as a catalyst for translating theory into practice.

In addition to educational outcomes, respondents pointed out several systemic barriers impeding the widespread adoption of HLM. One key obstacle was the fee-for-service payment structure, which rewards service volume rather than health outcomes and provides limited incentives for preventive care or time-intensive, individualized lifestyle counseling [17,18]. Another barrier to HLM implementation was the lack of official recognition of HLM by regulatory bodies, national-level regulatory constraints, few accreditation opportunities, and discrepancies in access to diagnostic tools and infrastructure depending on the country of practice. Literature shows that physician accreditation in medical subdomains is correlated

XSL•FO
RenderX

with better patient outcomes [19], while national board certification opportunities accompanied by CME education are associated with practice growth, career satisfaction, and better employment opportunities [20]. In this study, 95.6% (n=108) of physicians were in favor of formal accreditation of HLM, and 73.5% (n=83) of physicians expected HLM to significantly grow, eventually integrating into mainstream medicine. The responses represent an urgent need for coordinated efforts that integrate education, policy reform, and infrastructure development to facilitate the full-scale implementation of HLM across health care systems. Therefore, this study identified the need for structured education and certification programs, as well as formal board recognition, to enable scalable, evidence-based delivery of healthy longevity–focused care.

Limitations

There are several limitations to the study. As with all voluntary survey studies, the findings may be subject to self-selection and nonresponse bias, as physicians with a greater interest in HLM may have been more likely to participate. Outcomes were self-reported and may therefore be influenced by recall or social desirability bias. The cross-sectional design prevents causal inference, limiting conclusions about the long-term impact of HLM education on clinical practice. Selection bias is a concern, as participants were self-selected and may have had a pre-existing interest in HLM, potentially overestimating the adoption of HLM principles. The low response rate also introduces nonresponse bias, as respondents may not be representative of all physicians completing the course. In addition, self-reported data on confidence and practice changes are subject to recall and social desirability biases, lacking objective clinical validation. Regional differences in medical regulations and reimbursement models may further limit generalizability. Practice setting (public vs private) may plausibly influence physicians' perceptions of the applicability and validity of HLM, given differences in clinical autonomy, reimbursement structures, and preventive care implementation. Despite these limitations, this study provides novel insights into the role of structured online HLM education and underscores the need for further investigation into its clinical integration. Future research should assess the long-term impact and cost-effectiveness of HLM training while evaluating how delivery methods and systemic barriers influence clinical implementation and professional outcomes.

Conclusions

This study highlights the importance of accredited HLM education as a scalable tool for improving physicians' competency in HLM. The development and implementation of a designated curriculum constitute a critical step in bridging the translational gap between rapidly advancing theoretical discoveries in geroscience and their application in clinical settings. There is a growing need for structured, competency-based, and frequently updated educational programs that equip health care professionals with the necessary knowledge and skills to integrate gerodiagnostics (HLM diagnostics based on biomarkers of aging) and gerotherapeutics (drugs that affect pathways involved in aging to lessen the burden of age-related diseases as well as to prolong health span and potentially lifespan [21]) into patient care. Importantly, systematic studies of medical education serve as a form of quality control, ensuring the identification of educational gaps and clinical implementation challenges and enabling the iterative optimization of teaching methodologies and content relevance. This process is fundamental for further establishment of standardized training pathways and professional accreditation frameworks, which are essential prerequisites for recognizing HLM as a formal medical discipline. Furthermore, the scalability and long-term sustainability of such educational initiatives require strategic integration with health policy and regulatory efforts, fostering a harmonized approach that supports system-level adoption. Aligning educational standards with policy development not only enhances the credibility of HLM but also ensures that it becomes an integral part of national and international health systems, ultimately advancing the goal of extending health span and reducing the burden of age-related diseases.

Acknowledgments

JLK is supported by the Connor Fund, Robert J Ryan and Theresa W Ryan, and the Noaber Foundation. The authors thank Miao Zheng and Pheobe Zhang for their support in managing the Longevity Education Hub. During the preparation of this manuscript, the authors used ChatGPT-4o (OpenAI) for grammar, stylistic, and linguistic review. The authors subsequently reviewed and edited the content as necessary and assume full responsibility for the accuracy and integrity of the published work.

Funding

No external financial support or grants were received from any public, commercial, or not-for-profit entities for the research, authorship, or publication of this article.

Data Availability

The datasets generated or analyzed during this study are available from the corresponding author on reasonable request.

Authors' Contributions

Conceptualization: AZ, EB
Data curation: DW

XSL•FO
RenderX

Formal analysis: DW, EB, BW, CYH, RGP, MAF, SS, WS
Methodology: EB, DW, AZ, AM
Visualization: DW
Writing – original draft: DW, EB
Writing – review & editing: EB, DW, BW, CYH, SS, WS, DČ, CH, YM, HY, MAF, RGP, MS-K, AM, FS, FC-F, AG, ABM, JK, AZ

Conflicts of Interest

EB, DW, BW, CYH, SS, WS, DČ, CH, YM, HY, MAF, RGP, MS-K, AM, AG, FC-F, FMS, and AZ are volunteers of the Longevity Education Hub (LEH), a volunteer-run, noncommercial educational initiative offering free, online Healthy Longevity Medicine courses. None of the authors receives remuneration or holds an ownership interest in LEH. The following authors have additional nonfinancial roles in the initiative and its educational resources evaluated in this study: AZ and EB co-founded the initiative and conceptualized, developed, and authored course curricula and educational materials; MS-K and AM contributed scientific expertise and co-authored educational content; DW and DČ were responsible for communications, social media outreach, website coordination, and operational support, including the dissemination and accessibility of the courses. DČ additionally undertook course management and administrative coordination.
AZ is the founder and shareholder of Insilico Medicine, a generative AI and quantum computing–enabled longevity biotechnology company. ABM is chief medical officer of NU, a software company that provides analysis of health data and does not offer educational programs. ABM is also director of the NUS Academy for Healthy Longevity (disclosure, not a conflict of interest). All other authors declare no competing interests.

Checklist 1
STROBE checklist.
[DOCX File, 42 KB - mededu_v12i1e83779_app1.docx]

References

1. Kennedy BK, Berger SL, Brunet A, et al. Geroscience: linking aging to chronic disease. Cell 2014 Nov 6;159(4):709-713. [doi: 10.1016/j.cell.2014.10.039] [Medline: 25417146]
2. Newman JC, Milman S, Hashmi SK, et al. Strategies and challenges in clinical trials targeting human aging. J Gerontol A Biol Sci Med Sci 2016 Nov;71(11):1424-1434. [doi: 10.1093/gerona/glw149] [Medline: 27535968]
3. Thuault S. Drug discovery by AI trained on aging biology. Nat Aging 2024 Apr;4(4):437. [doi: 10.1038/s43587-024-00615-1] [Medline: 38580819]
4. Wilczok D. Deep learning and generative artificial intelligence in aging research and healthy longevity medicine. Aging (Albany NY) 2025 Jan 16;17(1):251-275. [doi: 10.18632/aging.206190] [Medline: 39836094]
5. Meron E, Thaysen M, Angeli S, et al. Meeting report: aging research and drug discovery. Aging (Milano) 2022 Jan 31;14(2):530-543 [FREE Full text] [doi: 10.18632/aging.203859]
6. Lyu YX, Fu Q, Wilczok D, et al. Longevity biotechnology: bridging AI, biomarkers, geroscience and clinical applications for healthy longevity. Aging (Albany NY) 2024 Oct 16;16(20):12955-12976. [doi: 10.18632/aging.206135] [Medline: 39418098]
7. Bonnes SLR, Strauss T, Palmer AK, et al. Establishing healthy longevity clinics in publicly funded hospitals. Geroscience 2024 Oct;46(5):4217-4223. [doi: 10.1007/s11357-024-01132-0] [Medline: 38512582]
8. Abu Dhabi sets standards for the world's first Healthy Longevity Medicine Centres. Department of Health - Abu Dhabi. URL: https://www.doh.gov.ae/en/news/abu-dhabi-sets-standards-for-the-world-first-healthy-longevity-medicine-centres [accessed 2026-02-27]
9. Leong L, Ninnis J, Slatkin N, et al. Evaluating the impact of pain management (PM) education on physician practice patterns--a continuing medical education (CME) outcomes study. J Cancer Educ 2010 Jun;25(2):224-228. [doi: 10.1007/s13187-010-0040-y] [Medline: 20204577]
10. Ames ML, Staffier KL, Kees A, et al. Online lifestyle medicine continuing medical education (CME) course completion predicts increases in clinician knowledge, confidence, and practice of lifestyle medicine. Am J Lifestyle Med 2024 Nov 5:15598276241279523. [doi: 10.1177/15598276241279523] [Medline: 39540184]
11. Cervero RM, Gaines JK. The impact of CME on physician performance and patient health outcomes: an updated synthesis of systematic reviews. J Contin Educ Health Prof 2015;35(2):131-138. [doi: 10.1002/chp.21290] [Medline: 26115113]
12. Federal Register: The Daily Journal of the United States Government. URL: https://unblock.federalregister.gov/ [accessed 2026-02-23]
13. Moqri M, Herzog C, Poganik JR, et al. Validation of biomarkers of aging. Nat Med 2024 Feb;30(2):360-372. [doi: 10.1038/s41591-023-02784-9] [Medline: 38355974]
14. Moqri M, Herzog C, Poganik JR, et al. Biomarkers of aging for the identification and evaluation of longevity interventions. Cell 2023 Aug 31;186(18):3758-3775. [doi: 10.1016/j.cell.2023.08.003] [Medline: 37657418]

XSL•FO
RenderX

15. Apsley AT, Etzel L, Ye Q, Shalev I. From population science to the clinic? Limits of epigenetic clocks as personal biomarkers. Epigenomics 2025 Dec;17(18):1447-1461. [doi: 10.1080/17501911.2025.2603880] [Medline: 41403206]
16. Perri G, French C, Agostinis-Sobrinho C, et al. An expert consensus statement on biomarkers of aging for use in intervention studies. J Gerontol A Biol Sci Med Sci 2025 Apr 7;80(5):glae297. [doi: 10.1093/gerona/glae297] [Medline: 39708300]
17. Morenz AM, Zhou L, Wong ES, Liao JM. Association between capitated payments and preventive care among U.S. adults. AJPM Focus 2023 Sep;2(3):100116. [doi: 10.1016/j.focus.2023.100116] [Medline: 37790668]
18. Kraft KB, Hoff EH, Nylenna M, Moe CF, Mykletun A, Østby K. Time is money: general practitioners' reflections on the fee-for-service system. BMC Health Serv Res 2024 Apr 15;24(1):472. [doi: 10.1186/s12913-024-10968-3] [Medline: 38622602]
19. Alkhenizan A, Shaw C. Impact of accreditation on the quality of healthcare services: a systematic review of the literature. Ann Saudi Med 2011;31(4):407-416. [doi: 10.4103/0256-4947.83204] [Medline: 21808119]
20. Lianov L, Rosenfeld R. The case for physician training and certification in lifestyle medicine. Am J Lifestyle Med 2023;17(2):326-330. [doi: 10.1177/15598276221138087] [Medline: 36896033]
21. Couteur DG, Barzilai N. New horizons in life extension, health span extension and exceptional longevity. Age Ageing 2022 Aug 2;51(8):afac156. [doi: 10.1093/ageing/afac156]

Abbreviations

CME: continuing medical education
HLM: healthy longevity medicine
OR: odds ratio
PPG: primary and preventive care and general practice
SPEC: specialized care and advanced procedures
STROBE: Strengthening the Reporting of Observational Studies in Epidemiology

Edited by A Stone; submitted 08.Sep.2025; peer-reviewed by A Aparicio, D Major, I Stambler, N Lidströmer, S Bonnes; accepted 30.Jan.2026; published 19.Mar.2026.

Please cite as:
Bischof E, Wilczok D, Kirkland JL, Wilaksono B, Huang CY, Suwannaphong S, Sanphasitvong W, Čamdžić D, Hernandez C, Madea Y, Yamada H, Fernandes MA, Gaminha Pacheco R, Serfaty FM, Calvo-Fortes F, Goldman A, Maier AB, Moskalev A, Scheibye-Knudsen M, Zhavoronkov A
Upskilling in Healthy Longevity Medicine and Its Association With Physicians' Implementation Intent and Self-Reported Clinical Confidence: Cross-Sectional Observational Study
JMIR Med Educ 2026;12:e83779
URL: https://mededu.jmir.org/2026/1/e83779
doi:10.2196/83779

© Evelyne Bischof, Dominika Wilczok, James L Kirkland, Bhirau Wilaksono, Christine Yuan Huang, Suwanna Suwannaphong, Wanviput Sanphasitvong, Dalila Čamdžić, Carolina Hernandez, Yoko Madea, Hidekazu Yamada, Melissa Alexandre Fernandes, Ricardo Gaminha Pacheco, Fabiano M Serfaty, Fernanda Calvo-Fortes, Amit Goldman, Andrea B Maier, Alexey Moskalev, Morten Scheibye-Knudsen, Alex Zhavoronkov. Originally published in JMIR Medical Education (https://mededu.jmir.org), 19.Mar.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Education, is properly cited. The complete bibliographic information, a link to the original publication on https://mededu.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Experiential Course Learning, Wellness, and Higher Education: Qualitative Descriptive Study

Adrianna Lorraine Watson[1], PhD; Neil E Peterson[1], PhD; Brandon Thatcher[1], MSN; Michael Thomas[1], DNP; Stacie Hunsaker[1], DNP; Cole Hooley[2], PhD; David Erekson[3], PhD; Adam Simpson[4], MS; Gregory Snow[4], PhD; Rachel Detrick[1]

[1]College of Nursing, Brigham Young University, 500 KMBLL, Provo, UT, United States

[2]School of Social Work, Brigham Young University, Provo, UT, United States

[3]Student Life, Brigham Young University, Provo, UT, United States

[4]Department of Statistics, Brigham Young University, Provo, UT, United States

Corresponding Author:
Adrianna Lorraine Watson, PhD
College of Nursing, Brigham Young University, 500 KMBLL, Provo, UT, United States

Abstract

Background: Undergraduate students, including those preparing for health professions, report high rates of psychological distress and underuse of traditional counseling services. Credit-bearing wellness courses that combine psychoeducation with experiential learning may offer a scalable, curriculum-based approach to supporting student well-being.

Objective: This qualitative study explored how undergraduate students described personal growth, coping, and lifestyle changes following participation in experiential wellness courses.

Methods: An anonymous postcourse online survey captured open-ended responses from students enrolled across 6 wellness course sections. The courses emphasized stress physiology, evidence-based coping strategies, and weekly experiential assignments. Narrative responses from 110 participants were analyzed inductively using the reflexive thematic analysis developed by Braun and Clarke within a constructivist-interpretivist paradigm.

Results: A total of six themes were identified: (1) healthy habits and practical lifestyle change; (2) stress management skills and mental health techniques; (3) self-reflection, awareness, and personal growth; (4) relevance and immediate applicability; (5) peer connection and discussion-based learning; and (6) course structure and opportunities for improvement. Students described adopting new coping strategies, developing greater self-awareness, and perceiving course content as relevant and applicable to their daily lives.

Conclusions: Students described experiential wellness courses as supportive of coping, self-awareness, and behavior change. These findings provide insight into how students engage with and interpret course-based wellness education. Curriculum-integrated approaches may represent a complementary strategy to support student well-being. Future research should examine these approaches across diverse populations and over time.

(JMIR Med Educ 2026;12:e88642) doi:10.2196/88642

KEYWORDS

mental health; nursing students; students; health education; health promotion; psychological stress; qualitative research; curriculum; psychological resilience

Introduction

Background

Mental health represents a global public health crisis. In 2019, approximately 970 million people were living with a mental disorder [1]. These conditions are shaped by structural determinants, such as food insecurity, violence, and childhood adversity, requiring responses that are context-sensitive and multifaceted [2]. Scalable, accessible approaches that extend beyond specialist care are therefore essential, particularly for young adults who will soon enter and sustain global health, education, and economic systems [3-6].

In the United States, an estimated 59.3 million adults experience mental illness annually, including 36.2% of individuals aged 18 to 25 years [7]. The burden is particularly pronounced among college students. Data from the Spring 2024 American College Health Association National College Health Assessment indicate that 19.5% of students screened positive for serious psychological distress, 25.9% for suicidality, and 48.5% reported

XSL•FO
RenderX

loneliness [8]. Despite this, only 25% received counseling in the prior year.

Longitudinal survey data further demonstrate a rising prevalence of depression (21%-38%) and anxiety (22%-34%) between 2014 and 2024, with 13% reporting serious suicidal ideation [9,10]. These trends reflect distress rates comparable to or exceeding those of the general population, alongside persistent underuse of services. Importantly, mental health needs and help-seeking behaviors vary across academic disciplines, underscoring the need for context-sensitive rather than uniform interventions [11].

A range of interventions, including digital platforms, cognitive-behavioral approaches, and mindfulness-based programs, have demonstrated moderate effectiveness in reducing depression and anxiety among university students [12]. However, systematic reviews highlight ongoing challenges, including variability in outcomes, methodological limitations, and issues with engagement and implementation [13]. Earlier work similarly identified gaps in prevention and early intervention strategies within higher education [14]. More recently, attention has shifted toward curriculum-integrated approaches, such as credit-bearing wellness courses that embed mental health promotion within academic environments. Emerging evidence suggests these interventions may improve stress, coping, and help-seeking behaviors [15], yet existing studies are often limited to single courses or specific populations and provide limited insight into student experience or mechanisms of impact [16-19]. This gap is particularly relevant given rising student distress and continued underuse of traditional services. To address this, this study explored undergraduate students' experiences of experiential wellness courses using a descriptive qualitative design with an inductive approach.

Aim

The purpose of this study was to explore students' experiences and the perceived effectiveness of experiential wellness courses. The research question guiding this inquiry was: How do undergraduate students describe their experiences, the perceived effectiveness, and the personal impact of experiential wellness courses?

Methods

Qualitative Approach and Research Paradigm

Reflexive Thematic Analysis

Narrative data were analyzed using the 6-phase reflexive thematic analysis (familiarization, coding, theme development, review, definition, and reporting) developed by Braun and Clarke [20]. This inductive approach supported repeated movement between the full dataset and individual meaning units, allowing patterns to be identified from participants' language while preserving interpretive depth. First-cycle coding was conducted manually, with analysts working line-by-line through the data using a flexible combination of in vivo and descriptive codes. In some excerpts, analysts also applied more than one code where multiple meanings were present. Codes were then discussed and iteratively refined through reflexive dialogue, memo writing, and repeated comparison across responses as themes developed. Table 1 summarizes the primary coding practices used in the early stages of analysis.

Table . Illustrative first-cycle coding practices[a].

Coding practice	Analytic function	Example
In vivo (participant-language) coding	Preserved participants' wording and stayed close to lived experience	"Feeling in control"
Descriptive coding	Summarized the main topic, action, or experience in a segment	"Improved sleep habits" and "time management strategies"
Simultaneous coding	Captured more than one meaning within a single excerpt	Same excerpt coded as coping skill use and identity shift

[a]These coding practices were used flexibly and interpretively rather than as fixed, mutually exclusive categories.

Paradigm

The study was situated within a constructivist-interpretivist paradigm, which assumes that meaning is coconstructed and context-dependent. Accordingly, the analysis focused on interpreting the range and depth of student experiences rather than quantifying response frequency. Within this approach, researcher subjectivity was understood as a resource for interpretation rather than a source of bias to be minimized.

Rationale

This approach aligns with the study's aim of understanding how students interpret and apply experiential wellness curricula. Reflexive thematic analysis provides a flexible yet systematic method for examining meaning-making processes related to coping, behavior change, and personal development.

Researcher Characteristics and Reflexivity

The analytic team brought complementary clinical, multidisciplinary, and lived perspectives to the study. The research team consisted of faculty researchers and student research assistants with backgrounds in nursing, social work, counseling psychology, and statistics. Recognizing that their professional roles could shape data interpretation, team members engaged in continuous reflexive practice. Before analysis, each analyst recorded positionality statements detailing personal wellness beliefs, prior exposure to college-health settings, and expectations about course impact. These statements were revisited during peer-debrief meetings to surface implicit assumptions. Differing interpretations were explored through reflexive discussion, with attention to how analytic perspectives

XSL•FO
RenderX

were shaped by researchers' disciplinary and experiential backgrounds [20,21].

Context and Sampling Strategy

The study was conducted across 6 undergraduate wellness courses offered over 4 semesters (2021 - 2022 and 2022 - 2023) at a private university in the Western United States. The courses were semester-long (14 - 16 weeks), credit-bearing electives, and varied in topical focus, including mindfulness-based wellness, stress management and resilience, positive psychology, cognitive and behavioral wellness, lifestyle behavior change, and integrative wellness or student development. Course content emphasized stress physiology, cognitive restructuring, mindfulness practices, behavior change strategies, and intentional self-care, with instruction delivered through weekly applied assignments, structured reflection (eg, journaling), and discussion-based learning. Table 2 summarizes the courses, including target learners, educational objectives, enrollment size, instructional format, and core content areas.

Table . Characteristics of experiential wellness courses

Course type or focus area	Target learners	Educational objectives	Enrollment size (approximately)	Instructional format	Core content areas
Mindfulness-based wellness	Undergraduate students (mixed majors; includes nursing/prenursing)	Develop awareness of stress responses and cultivate mindfulness-based coping strategies for emotional regulation	15 - 25	Small-group, discussion-based, and experiential	Mindfulness practices, breathwork, present-moment awareness, and stress reduction
Stress management and resilience	Undergraduate students (health and nonhealth majors)	Build practical stress management skills and enhance resilience for academic and personal challenges	20 - 30	Hybrid (lecture, experiential, and discussion)	Stress physiology, coping strategies, relaxation techniques, and resilience-building
Positive psychology and well-being	Undergraduate students (mixed majors)	Promote well-being through strengths-based approaches and meaning-making strategies	20 - 30	Discussion-based with reflective activities	Gratitude, strengths identification, purpose, and emotional well-being
Cognitive and behavioral wellness	Undergraduate students (including prehealth students)	Apply cognitive and behavioral strategies to improve thought patterns, emotional regulation, and behavior change	20 - 30	Hybrid (lecture and applied practice)	Cognitive restructuring, thought patterns, behavior modification, and self-regulation
Lifestyle and health behavior change	Undergraduate students (including exercise science and prehealth students)	Support development of sustainable health behaviors and habit formation	20 - 35	Experiential learning with weekly assignments	Sleep hygiene, time management, physical activity, and habit formation
Integrative wellness or student development	General undergraduate population (elective and nonhealth specific)	Encourage holistic self-awareness, personal development, and interpersonal well-being	15 - 25	Discussion-based	Self-reflection, goal setting, peer dialogue, and holistic wellness

Participants were a heterogeneous undergraduate sample that included nursing and prenursing students, other health-related majors (eg, exercise science), and nonhealth profession students. The courses functioned both as general education electives and as preparatory or complementary learning experiences for students pursuing health professions. A purposive sampling strategy was used to capture variation across pedagogical approaches, topical emphases, and disciplinary contexts within the university's wellness curriculum. All registered students in the 6 selected courses were invited to complete an anonymous end-of-semester Qualtrics survey with open-ended prompts. Inclusion criteria were age 18 years or older, English fluency, enrollment in 1 of the 6 target courses, and completion of at least 1 substantive open-ended response. Surveys with blank qualitative items were excluded. Because data were collected from a fixed set of courses, formal thematic saturation was not used as a stopping criterion. Instead, sampling prioritized conceptual breadth across course types to support transferability and the identification of cross-cutting patterns in student experience.

Ethical Considerations

The Brigham Young University Institutional Review Board ethically approved this study (IRB2021-192). Participants provided informed consent, and responses were deidentified on the online platform. Incentives for completion of the survey consisted of a US $5 gift card.

Data Collection Methods

Data were collected via an anonymous postcourse online survey administered through Qualtrics after course completion. All enrolled students were invited via email, and participation was

XSL•FO
RenderX

voluntary. Of the 125 consenting students, 110 (88%) provided usable responses. The survey system restricted submissions to one per participant, and no duplicate entries were identified. Survey design and reporting adhered to guidelines for web-based research [22].

Data Collection Instruments and Technologies

The research team developed open-ended survey prompts informed by literature on college student mental health, wellness education, and experiential learning. The prompts were designed to elicit perceptions of course effectiveness, engagement with experiential components, and perceived impact on well-being, including coping strategies, behavioral change, self-awareness, and personal growth. The instrument was developed collaboratively by a multidisciplinary team and was iteratively reviewed for clarity, relevance, and alignment with study aims. Consistent with the exploratory qualitative design, the instrument was not pilot tested or psychometrically validated.

Data Processing

Survey responses were exported to a secure, encrypted file. Identifying information was removed, and responses were assigned numeric codes. Data were organized in Microsoft Excel to facilitate coding and analysis.

Data Analysis

Analysis was conducted using reflexive thematic analysis [20]. A total of 2 analysts engaged in iterative, reflexive coding of the dataset, initially working through the data individually and then meeting regularly to discuss interpretations, explore alternative readings, and refine the evolving analytic framework. Coding was inductive and conducted line-by-line, using a flexible combination of in vivo (participant-language) and descriptive codes, with occasional simultaneous coding where excerpts reflected multiple meanings (Table 1).

Codes were not treated as fixed entities but were developed and refined through ongoing engagement with the data, supported by memo writing and an audit trail documenting analytical decisions. Through iterative comparison across responses, codes were grouped into candidate themes, which were subsequently reviewed and refined to ensure internal coherence and conceptual distinction. Themes were then defined and illustrated using representative participant quotations. Consistent with reflexive thematic analysis, the analytic process emphasized reflexivity, interpretive depth, and the coconstruction of meaning, rather than coding consensus or reliability. Reflexive journaling supported ongoing attention to how researchers' perspectives shaped interpretation throughout the analytic process.

Techniques to Enhance Trustworthiness

To enhance rigor, triangulation was operationalized through an iterative comparison of three sources: (1) student narrative responses, (2) relevant contemporary literature on wellness education and mental health interventions, and (3) input from external content experts. During analysis, the research team engaged in ongoing cycles of comparison between emerging codes and themes and existing literature. As preliminary patterns were identified, analysts conducted targeted literature checks to examine whether these patterns aligned with, extended, or diverged from prior findings. These comparisons were documented in analytic memos and discussed during regular team meetings, allowing for the refinement of theme definitions and ensuring that interpretations were both grounded in participant accounts and situated within the broader evidence base.

External expert review was incorporated after the initial theme development. A total of 2 subject-matter experts in counseling psychology and health promotion were provided with a summary of candidate themes, representative excerpts, and the analytic rationale. The experts were asked to assess conceptual clarity, coherence, and alignment with current knowledge in the field. Their feedback was discussed by the analytic team and used to refine theme boundaries, naming, and interpretation where appropriate. All changes informed by expert input were documented in the audit trail. This triangulated approach strengthened the credibility and confirmability of the findings by ensuring that interpretations were grounded in participant narratives while also being critically examined through both scholarly and expert perspectives.

Results

Participant Characteristics

Of the 125 students enrolled in wellness courses who consented to participate, 110 (88%) provided usable postcourse responses for analysis. Demographic information was available for the 125 consented participants. These participants were drawn from 6 courses offered across 4 semesters (2021‐2022 and 2022‐2023 academic years), representing a range of disciplines, including exercise science, interdisciplinary general education (honors), nursing, and student development; the latter reflects a general undergraduate or student affairs context rather than a health professions program (Table 3). Of the 125 students, the sample was predominantly female (n=115, 92%) and non-Hispanic White (n=113, 90.4%). Participants were relatively evenly distributed across academic years: first year (n=28, 22.4%), second year (n=32, 25.6%), third year (n=26, 20.8%), and fourth year (n=26, 20.8%), with a smaller proportion in their fifth year or beyond (n=13, 10.4%).

XSL•FO
RenderX

Table . Participant characteristics of consented students

Characteristic	Value, n (%)[a]
Total consented participants	125 (100)
Provided usable postcourse responses for qualitative analysis	110 (88.0)
Sex	
Female	115 (92.0)
Male	10 (8.0)
Race/ethnicity	
Non-Hispanic White	113 (90.4)
Other race/ethnicity	12 (9.6)
Academic year	
First year	28 (22.4)
Second year	32 (25.6)
Third year	26 (20.8)
Fourth year	26 (20.8)
Fifth year or beyond	13 (10.4)

[a]Percentages are based on the total consented sample and may not equal 100 due to rounding.

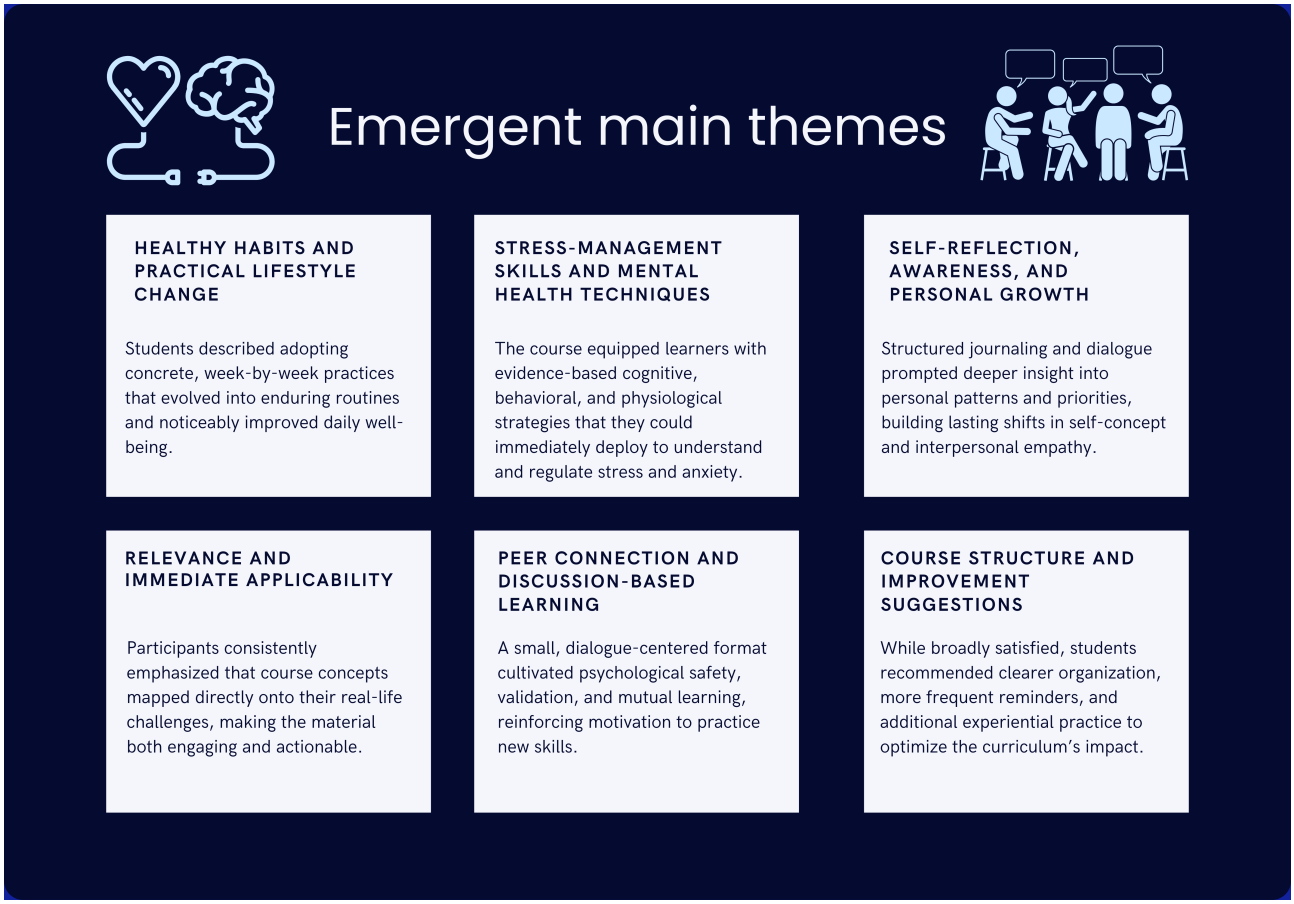
Emergent Themes

Overview

Analysis revealed six emergent main themes: (1) healthy habits and practical lifestyle changes; (2) stress management skills and mental health techniques; (3) self-reflection, awareness, and personal growth; (4) relevance and immediate applicability; (5) peer connection and discussion-based learning; and (6) course structure and opportunities for improvement. See Figure 1 for a brief overview of these themes. The 6 courses described above represent the instructional context from which participants were sampled; however, the themes presented below reflect cross-cutting patterns identified across all participants and course types. Themes did not map to individual courses but instead captured shared experiential outcomes that emerged across varied instructional formats, content areas, and student populations. The analysis, therefore, focuses on common patterns in how students described the impact of experiential wellness learning, rather than on course-specific differences.

XSL•FO
RenderX

Figure 1. Overview of main themes.



Theme 1: Healthy Habits and Practical Lifestyle Change

Participants across courses described their experiences in wellness courses as a catalyst for tangible, day-to-day transformation. Rather than learning abstract principles, they reported adopting new behaviors, routinizing them, and noticing measurable gains in well-being. One student captured the overarching impact succinctly: "It helped encourage healthy life habits that I saw having a large improvement in my life." The perception that change was both visible and personally meaningful was evident in every account.

A central driver of this shift was the rhythm of experiential homework. Weekly assignments were not viewed as traditional coursework but as opportunities to "learn how to be a happier person." Students repeatedly emphasized that these tasks translated concepts into action; by "consciously set[ting] apart time every week to help destress," they could evaluate strategies in real contexts and refine them over time.

Learners also spoke of accumulating skills in a deliberate, scaffolded sequence. As 1 participant explained:

> *We covered a new topic each week to focus on and apply into your life. It [built] on each other so that by the end of the semester, I had a lot of good new habits.*

This progressive layering fostered both competence and confidence, allowing practices to migrate from structured exercises to authentic habits embedded in daily routines.

Crucially, the motivation to change was experienced as self-endorsed. Students highlighted an "emphasis on growth for yourself and changing your lifestyle because you want to," describing the activities as "realistic, practical ways to be and feel better." Far from being externally imposed requirements, the assignments were perceived as invitations to experiment with well-being strategies and to keep what worked.

The tangible nature of the outcomes fortified engagement. As 1 learner reflected, "I liked putting into practice what we learned and seeing for myself the difference it made." Another credited the course grading structure for sustaining momentum, "I think the experiential assignments were most helpful too because we had motivation to reach our positive living goals since it was for a grade." Grades functioned less as extrinsic pressure than as a supportive scaffold, prompting sustained practice until new behaviors took root.

By the close of the term, many participants articulated a sense of ownership over their emerging lifestyles. One student concluded, "I loved the skills and habits I developed after taking the class," signaling that the practices had become woven into the fabric of everyday life. Narratives reveal a pattern of intentional, self-directed habit formation grounded in weekly experiential work, which yielded authentic, lived improvements in wellness.

XSL•FO
RenderX

Theme 2: Stress Management Skills and Mental Health Techniques

Students described the course as a practical workshop in evidence-based coping methods. Far from remaining abstract, concepts were embedded in hands-on activities that allowed immediate application. One participant captured this ethos by noting they "liked being able to try different stress-relieving activities," while another emphasized mastery of "simple skills to help me cope with stress… it has helped me so much." Across accounts, learners framed stress not as an unavoidable burden but as a challenge that could be actively managed with the right tools.

A recurring narrative centered on biological and cognitive insight. Several students highlighted that understanding "how the brain works" or "how the autonomic nervous system works and why" fundamentally shifted their appraisal of anxiety and pressure. This psychoeducation component laid the groundwork for skill adoption, "Learning some stress physiology and practical skills to manage stress, plus a shift in how I think about stress," signaled that knowledge and practice were mutually reinforcing.

Participants described a diverse toolkit that spanned cognitive, behavioral, and mindfulness-based techniques. Cognitive restructuring, time-management strategies, and relaxation exercises emerged as particularly salient. One student summarized the breadth of learning:

> *The most helpful things were learning about cognitive restructuring, learning about how the stress response works, and time-management skills. And relaxation skills.*

Another underscored the power of targeted practice, "We practiced simple techniques to acknowledge and cope with intrusive thoughts."

For some, the techniques proved strikingly consequential. The statement that "defusion has literally been a lifesaver and stopped me from cutting multiple times" points to the clinical potency of even brief interventions when they are delivered in an autonomy-supportive, experiential format. Collectively, student narratives suggest a progression: (1) students gained conceptual clarity about stress, (2) rehearsed concrete strategies in low-stakes settings, and (3) finally generalized these skills to real-world contexts where stress and anxiety naturally arise.

Learners framed these practices as foundational life capacities rather than course-bound tasks. They anticipated ongoing relevance because stress is "bound to be present" in adult life, positioning the acquired techniques as durable resources. The course, therefore, functioned as both an introduction to and a rehearsal of coping strategies, empowering students to meet future challenges with a richer, more specialized toolkit.

Theme 3: Self-Reflection, Awareness, and Personal Growth

A third strand of participant experience centered on deepened self-understanding. Learners consistently portrayed the course as a mirror that helped them see themselves (and others) with new clarity. One student remarked:

> *I learned so much about myself and how to better my own mental health. I've noticed a difference in how I approach problems and I will absolutely take what I've learned with me in the future.*

This forward-looking sentiment illustrates how insights gained in class were immediately translated into everyday decision-making.

Structured reflection activities invited students to identify priorities and patterns. The journal exercises, for instance, enabled 1 participant to "self-reflect on my life and see what my priorities are," while small-group dialogue allowed another to "open up," describing the process as an opportunity that "gave me new perspective on things." Such revelations frequently extended beyond the individual to a relational sphere; several students reported learning "how to identify personal signs of struggling with mental health as well as how I can recognize when others are struggling," underscoring heightened interpersonal awareness.

The personal nature of the coursework also fostered a sense of direction. One learner explained that the class "helped me figure out what I specifically want to work out in myself. It gave me direction," pointing to emergent goal-setting rooted in self-diagnosis. For others, customized assignments enhanced engagement, "My favorite part was how personalized it was...the way that the assignments catered to me individually to help me develop my habits was really cool and rewarding." Emphasis on tailoring appears to have strengthened both motivation and perceived relevance.

These narratives reveal that students began by examining internal states, progressed to recognizing behavioral patterns, and ultimately reoriented future actions in light of these insights. In the words of 1 participant, "The experience really improved how I view myself, others, and the world around me," signaling a broadened lens that blended self-care with empathic attunement to the well-being of others.

Theme 4: Relevance and Immediate Applicability

A concise but recurring narrative highlighted how readily students could transfer course material into everyday life. One participant summarized the sentiment in just 2 words: "relatable" and "applicable." Rather than perceiving the curriculum as theoretical or distant, learners repeatedly emphasized its direct fit with their current challenges and goals. As 1 student noted, "The concepts we learned about were super applicable and relevant to my life," underscoring ideas introduced in class that mapped neatly onto ongoing personal circumstances.

Several accounts pointed to a sense of novelty coupled with practicality. A participant celebrated the "fresh perspective it offered on happiness," adding that this outlook was "applicable...to my life." The description implies that relevance was not merely a function of content overlap but of insight; students encountered familiar struggles reframed through new, actionable lenses. Across comments, the prevailing tone suggested that applicability heightened motivation; when strategies resonated with real‐world contexts, learners became more willing to test them and, by extension, sustain engagement with the course.

XSL•FO
RenderX

Theme 5: Peer Connections and Discussion-Based Learning

A prominent feature that shaped students' experiences was the interpersonal fabric of the classroom. Participants did not merely learn about well-being; they learned with and through one another, crediting the small, dialogue-oriented format for amplifying insight and comfort. One learner captured this dynamic succinctly: "Getting to know my peers, talking as a group." Repeatedly, students described the course as "discussion based," emphasizing that conversation, rather than lecture, was its heartbeat.

Across courses, smaller class sizes appeared to support a more intimate scale, which participants perceived as allowing everyone to speak and be heard. For example, 1 participant called it "an open dialogue with my peers." Another echoed this sentiment, noting they "loved having a class that was small where the students were able to interact frequently with one another." Such proximity translated into psychological safety; learners reported that the setup "helped me participate more than I usually would," suggesting that the environment lowered barriers to sharing personal struggles and successes.

The relational tone of the course appeared to satisfy a need for validation and empathy. One student explained, "Discussions helped me feel understood and heard and helped me relate to other people," highlighting the reciprocity inherent in storytelling and listening. Another participant emphasized the value of mutual disclosure, stating that "Discussing with others and sharing my experiences when possible was very helpful and validating for me." This validation intensified the perceived relevance of the course content; hearing peers navigate similar challenges reinforced the applicability of the strategies being taught.

Peer connection was not an ancillary benefit but a central learning mechanism. Dialogue provided a richer context for theoretical concepts, encouraged reflection, and built a supportive community where new habits could take root. In this way, the social architecture of the course both mirrored and strengthened well-being objectives, demonstrating that personal growth is often cultivated most powerfully in the company of others.

Theme 6: Course Structure and Opportunities for Improvement

Alongside widespread enthusiasm, students offered constructive responses regarding the course's scaffolding. The most common critique concerned organizational clarity. One participant remarked, "There was a lack of direction, which led to a lack of organization. I would suggest a more detailed curriculum plan," signaling a desire for clearer road-mapping of weekly expectations and major assignments. Others focused on operational details that would streamline engagement; several learners asked instructors to "send out reminders during the week for our different 'challenges'" because those tasks were "easy to forget."

A second cluster of comments advocated for expanded experiential practice. Although students valued the hands-on components, some felt the dosage could be higher:

> *More experiential homework assignments. The practice is one of the most important outcomes off this class. The more practice we get, the more ways we can see implementation and start to create habits.*

This feedback suggests that the experiential spiral highlighted in earlier themes could be strengthened by allocating additional time or credit to real-world application.

Learners also imagined ways to broaden the course's reach. One participant urged, "I would invite more people to join. They really do not know what they are missing," while another proposed shifting focus to "how to make your smaller community flourish and not just a bigger institution," hinting at an interest in service-oriented or peer-mentoring extensions.

Notably, praise and critique co-existed. Even students who recommended tweaks often emphasized overarching satisfaction:

> *I honestly don't think I would change anything about the course. It was planned well and the instructor did a great job helping us learn material and giving us space to discuss amongst ourselves.*

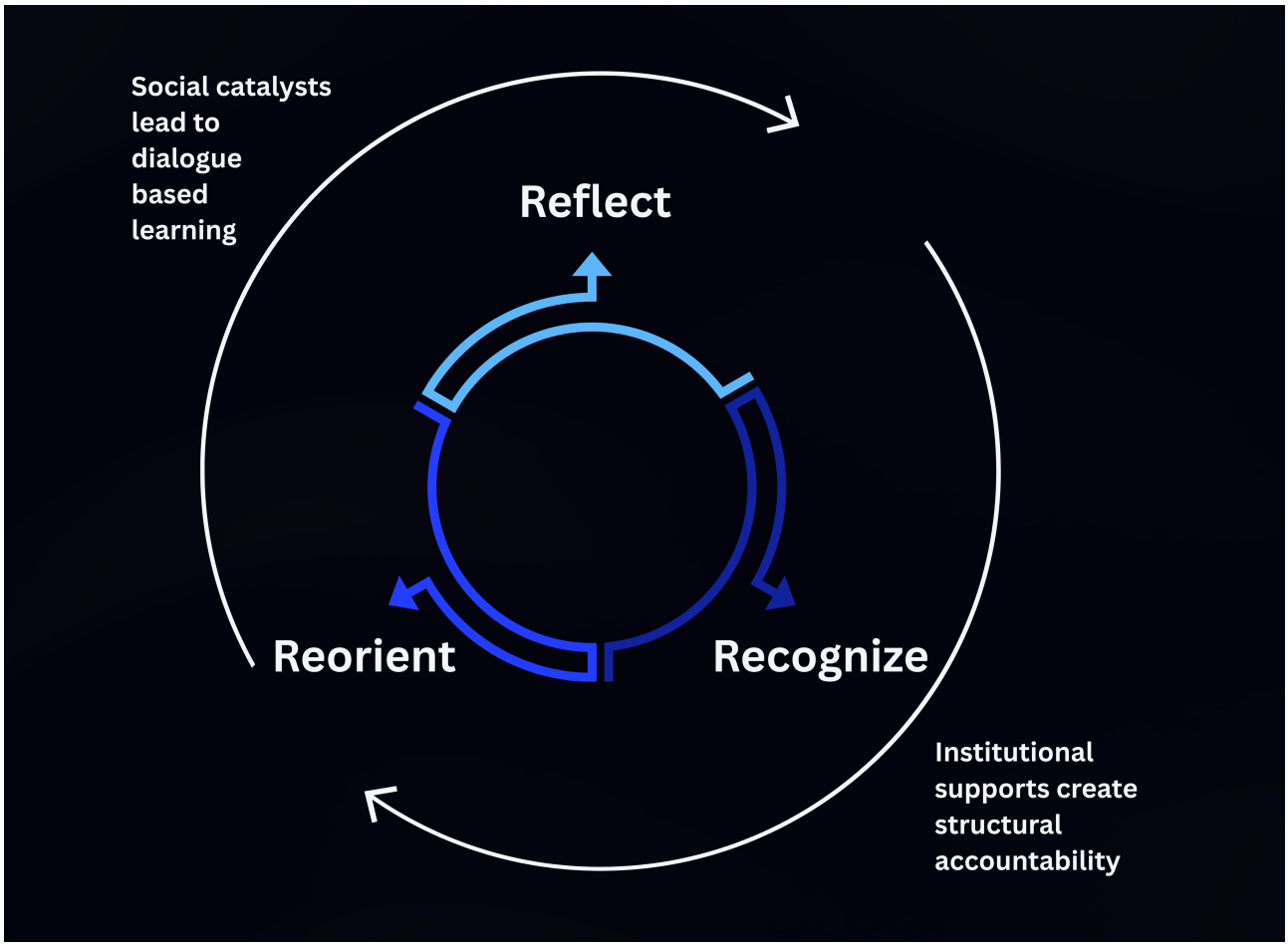
Participants sought fine-tuning (more transparent structure, timely reminders, and additional practice opportunities) rather than an overhaul, while affirming that the existing framework already delivered meaningful growth.

A Proposed Model of Undergraduate Wellness: "The Reflect-Recognize-Reorient Cycle"

Across themes, students described an iterative 3-part process that recurred throughout the semester (Figure 2).

XSL•FO
RenderX

Figure 2. The reflect-recognize-reorient (R^3) cycle.



Phase 1: Reflect

Students first turned their attention inward, identifying thoughts, emotions, and bodily cues they had previously overlooked. This mirrors the Kolb reflective observation and aligns with the Mezirow emphasis on critical self-reflection, yet our data foreground a distinctly affective, mindfulness-based stance that is not fully elaborated in either classical model [23-26].

Phase 2: Recognize

Having surfaced their internal states, students began to identify recurring behavioral patterns and situational triggers ("I spiral every time a deadline looms"). Conceptually, this bridges the Kolb abstract conceptualization and Mezirow meaning-making, but the reflect-recognize-reorient (R^3) cycle positions pattern recognition as a discrete, sense-making pivot rather than a by-product of reflection [23-26].

Phase 3: Reorient

Finally, students formulated concrete, forward-looking actions: scheduling micro-breaks, seeking peer support, or practicing paced breathing before exams. This corresponds to the Kolb active experimentation and the Mezirow reintegration; however, our participants described reorientation as value-driven, often linking new actions to a broader life purpose that surfaced in phases 1 - 2 [23-26].

Discussion

Principal Findings

Findings suggest that undergraduate students' experiences and perceived effectiveness of experiential wellness courses are characterized by 6 cross-cutting themes that illustrate how these courses may support student well-being. In this study, participants describe meaningful changes in daily behaviors, including the development of healthier habits and the application of practical coping strategies such as mindfulness, cognitive restructuring, and time management. In addition to skill acquisition, participants report increased self-awareness and personal growth, often accompanied by shifts in how they conceptualize stress, success, and self-care. Students consistently emphasize the immediate relevance and applicability of course content, which appears to enhance engagement and the sustained use of strategies. Peer interaction and discussion-based learning emerge as central mechanisms that foster validation, connection, and motivation. Although participants identify opportunities to improve course structure and organization, the findings collectively indicate that experiential, course-based wellness education supports both behavioral change and deeper meaning-making processes related to mental health and well-being.

XSL•FO
RenderX

Comparison With Literature

Alignment with and extension of existing research on mental health interventions suggest that participants may report improvements in well-being following exposure to structured interventions. For example, meta-analytic evidence demonstrates that cognitive-behavioral and mindfulness-based approaches produce moderate reductions in depression and anxiety among college students, and participants in this study similarly reported acquiring practical skills such as cognitive restructuring, relaxation techniques, and time management strategies [12]. These findings reinforce the effectiveness of skill-based approaches to mental health promotion within undergraduate populations.

The findings also extend current understanding by highlighting the importance of experiential and contextually embedded delivery. Previous research indicates that mental health interventions, particularly digital and self-guided formats, often encounter challenges related to engagement, uptake, and sustained use [13]. In contrast, participants in this study emphasize that embedding wellness strategies within structured, credit-bearing courses enhances both engagement and accountability. The integration of weekly experiential assignments and reflective activities appears to support sustained behavior change, suggesting that curriculum-based approaches may mitigate implementation challenges identified in prior research.

The findings further extend earlier work on prevention and early intervention in higher education. Foundational reviews identify a limited and inconsistent evidence base for preventive mental health interventions among university students, particularly for anxiety and depression [14]. Although recent studies demonstrate increased interest in course-based and curriculum-integrated approaches, much of the existing literature remains focused on discrete interventions or single-course implementations. Recent scoping reviews further highlight that while well-being interventions are increasingly implemented across higher education, many lack theoretical coherence and are not systematically integrated into academic curricula [27]. By examining student experiences across multiple wellness courses, this study identifies cross-cutting patterns of change, including the development of durable coping skills, increased self-awareness, and shifts in meaning-making.

The prominence of self-reflection, peer dialogue, and perceived relevance in these findings adds important nuance to the intervention literature. Prior studies primarily emphasize symptom reduction and measurable psychological outcomes, whereas participants in this study describe changes at both behavioral and identity levels. These findings suggest that wellness courses function not only as skill-building interventions but also as environments that support relational learning and meaning reconstruction. In doing so, the study bridges intervention-based research and transformative learning theory, offering a more integrated account of how mental health promotion occurs within educational settings.

Finally, the findings highlight the importance of structural and contextual features of the learning environment, including small class sizes and discussion-based formats. Participants consistently identified peer interaction and psychological safety as central to their learning experience, indicating that the instructional context plays a critical role in shaping outcomes. This perspective extends existing research by positioning classroom structure as an active component of mental health promotion rather than a neutral delivery mechanism. These findings align with broader literature demonstrating that academic stress is a pervasive driver of student distress and that effective interventions must address both individual coping strategies and the educational environments in which students learn [28].

The Kolb Experiential Learning in Practice

Students' accounts reveal an iterative learning spiral that closely mirrors the Kolb 4-mode cycle [23,24]. Experiential learning assignments provided concrete experience; guided journals and peer dialogue fostered reflective observation; minilectures on stress physiology or cognitive restructuring offered abstract conceptualization; and the following week's assignments invited active experimentation. Progressively, learners refined tactics for sleep hygiene, emotion regulation, and time management, demonstrating the Kolb premise that knowledge is continuously rebuilt as experience is tested, reflected upon, and reapplied [23,24].

The Mezirow Transformative Phases: From Skill to Meaning Change

While Kolb explains how learning moves, many participants described a deeper shift: questioning long-held beliefs, such as *success equals constant busyness*, and embracing new self-definitions ("I schedule rest with purpose") [25,26]. These narratives align with the Mezirow transformative learning theory: early assignments acted as disorienting dilemmas, prompting critical self-examination; dialogue enabled recognition of shared struggle; and students then explored options, planned actions, tried new roles, and reintegrated reshaped worldviews into daily life [25,26]. Thus, the experiential spiral sometimes culminated not merely in competence but in genuine perspective transformation.

The R^3 Cycle: A Pedagogical Bridge

The R^3 Cycle offers a concise synthesis of these findings, capturing the iterative process through which students move between self-awareness, pattern recognition, and intentional behavior change. Positioned alongside Kolb and Mezirow, the model highlights how experiential learning in this context extends beyond skill acquisition to include meaning-making and identity development. Table 3 provides a practical illustration of the model.

Contributions to the Wellness-Education Literature

This study advances the wellness-course literature in three key ways. First, it demonstrates that experiential learning curricula can support both discrete coping skill development and deeper shifts in meaning perspectives—outcomes that are often examined separately. Students not only reported acquiring practical strategies (eg, mindfulness, time management, and cognitive restructuring) but also described changes in how they conceptualized success, rest, and self-worth, bridging skill-based

XSL•FO
RenderX

and transformative learning perspectives. Second, the R^3 model offers a concise, practice-oriented framework for understanding how experiential learning translates into both behavioral and meaning-level change. Distilling complex transformational processes into 3 iterative phases allows the proposed model to suggest a structure that is adaptable to educational settings and amenable to operationalization in future research [25,26]. Third, the findings suggest that motivational and contextual factors (particularly experiential practice, reflection, and peer interaction) may help sustain the integration of new skills and perspectives over time. These insights extend existing learning theory by highlighting mechanisms that support persistence beyond the classroom [23-26]. Thus, these contributions can help shift the focus from whether wellness courses are effective to how and why they facilitate meaningful and sustained change in student well-being.

Limitations

These findings should be interpreted cautiously. First, they derive from a self-selected group at a single faith-based private university; respondents may differ from nonparticipants, and the campus culture, small class sizes, and elective credit structure may limit transferability. Second, the survey instrument was developed for the purposes of this study and was not formally pilot tested or psychometrically validated, which may affect the consistency and comparability of responses. Third, reliance on a single, postcourse open-ended survey limited opportunities for probing, introduced potential social desirability and recall bias, and provided limited evidence of sustained change over time. Fourth, all data were self-reported; objective indicators (eg, physiological stress measures or behavioral logs) were not collected to corroborate perceived gains. Fifth, the sample was predominantly female and non-Hispanic White, which may limit the transferability of findings to more diverse student populations and global nursing contexts, where experiences of stress, coping, and well-being may differ across sociocultural settings. Finally, although purposive sampling enabled the capture of diverse experiences across course types, the sample was drawn from a single institution and may not be fully transferable to other student populations or educational contexts. Together, these limitations underscore the need for mixed methods, multisite, and longitudinal research before broader generalization of the R^3 model or causal inferences can be made.

Implications for Health Profession Education

As universities face rising levels of student stress and increasing demand for mental health services, these findings suggest that wellness courses may serve as a scalable, curriculum-based approach to supporting student well-being. Participants reported meaningful gains in self-awareness, coping skills, and behavior change, indicating that such courses extend beyond knowledge acquisition to support personal growth. Given that only 45.5% of students rate their mental health as very good or excellent, and 41.7% identify stress as a barrier to academic success, the integration of wellness education into undergraduate curricula represents a timely and necessary response [8]. These courses may provide upstream support that could help alleviate pressure on campus counseling services. Such implications may also contribute to a broader institutional shift in viewing universities not only as academic environments but also as settings that shape lifelong well-being [27-29]. The findings of this study are consistent with calls for integrating wellness and trauma-informed approaches into nursing education, as existing curricula often lack structured preparation for managing stress and emotional burdens [30]. Wellness courses, therefore, represent both an educational innovation and a promising public health strategy.

The findings also offer several practical implications for educators in nursing and other health professions. First, experiential wellness strategies can be integrated into existing curricula without requiring the addition of standalone courses. Brief, structured activities such as reflective journaling, guided discussion, and applied skill practice (eg, stress regulation techniques) can be embedded within existing coursework to support student well-being. Second, instructional format plays a critical role in shaping student engagement. Small-group and discussion-based learning environments appear to enhance psychological safety, peer connection, and participation. Even within larger programs, incorporating periodic interactive or small-group elements may strengthen learning outcomes. Third, students consistently identify practical, immediately applicable skills as the most impactful. Incorporating content such as stress physiology, cognitive strategies, and time management into health professions education may support both academic success and professional resilience. These approaches align with the growing emphasis on prevention, self-regulation, and well-being in health care training [31]. Fourth, repetition and sustained practice emerge as essential for behavior change. Rather than relying on one-time interventions, wellness education may be more effective when delivered longitudinally through repeated, low-intensity experiential activities that support habit formation over time. Finally, these approaches are feasible within common institutional constraints. Many strategies identified (eg, reflection, peer dialogue, and applied exercises) require minimal financial investment and can be implemented without specialized infrastructure. This positions experiential wellness education as a scalable, curriculum-integrated approach to supporting student well-being in nursing and other health profession programs.

Conclusions

Experiential wellness courses offer a promising approach to integrating well-being into higher education. In this study, students described meaningful changes in coping, behavior, and self-awareness and perceived course content as relevant and applicable to their lives. These findings may provide insight into how students perceive, engage with, and interpret experiential wellness education. Such courses may highlight the value of combining psychoeducation, applied practice, reflection, and peer dialogue within academic settings. In the context of rising student distress, curriculum-integrated approaches may be considered a complementary strategy to support student well-being.

XSL•FO
RenderX

Acknowledgments

The authors thank the participating undergraduate students for their time and thoughtful responses. The authors also acknowledge the faculty and staff who supported recruitment and course coordination. The authors appreciate the guidance of the university's Office of Research Support and the contributions of the student research assistants who assisted with data management. The authors used ChatGPT (OpenAI) during manuscript development to critically appraise gaps in the logic and presentation of empirical findings. The tool was used to provide critical feedback for the authors' consideration. All manuscript content, including the analysis, conclusions, references, and citations, was reviewed and verified by the authors, who take full responsibility for the accuracy, originality, and integrity of the work.

Funding

This research received internal funding from Brigham Young University.

Data Availability

The datasets generated or analyzed during this study are available from the corresponding author on reasonable request.

Authors' Contributions

Conceptualization: NEP, BT, MT, SH, CH, DE
Data analysis: AS, GS
Formal analysis: NEP, BT, MT, SH, CH, DE, AS, GS
Funding acquisition: NEP, BT, MT, SH, CH, DE
Investigation: NEP, BT, MT, SH, CH, DE
Writing – review & editing: NEP, BT, MT, SH, CH, DE, ALW, RD
Writing – original draft: ALW, RD
All authors reviewed and approved the final version.

Conflicts of Interest

None declared.

References

1. Mental health. World Health Organization. 2025 Oct 8. URL: https://www.who.int/news-room/fact-sheets/detail/mental-health-strengthening-our-response [accessed 2026-04-23]
2. Kirkbride JB, Anglin DM, Colman I, et al. The social determinants of mental health and disorder: evidence, prevention and recommendations. World Psychiatry 2024 Feb;23(1):58-90. [doi: 10.1002/wps.21160] [Medline: 38214615]
3. Cara C, Hills M, Watson J. An Educator's Guide to Humanizing Nursing Education: Grounded in Caring Science: Springer Publishing Company; 2021.
4. Pei J, Amanvermez Y, Vigo D, et al. Sociodemographic correlates of mental health treatment seeking among college students: a systematic review and meta-analysis. Psychiatr Serv 2024 Jun 1;75(6):556-569. [doi: 10.1176/appi.ps.20230414] [Medline: 38291886]
5. Schroeder K, Sinko L, Ibrahim J, Sarwer DB. Supporting student learning and development through trauma-informed campuses. J Am Coll Health 2024 Dec;72(9):3681-3689. [doi: 10.1080/07448481.2023.2187647] [Medline: 36977341]
6. Zhang J, Peng C, Chen C. Mental health and academic performance of college students: knowledge in the field of mental health, self-control, and learning in college. Acta Psychol (Amst) 2024 Aug;248:104351. [doi: 10.1016/j.actpsy.2024.104351] [Medline: 38905949]
7. Key substance use and mental health indicators in the United States: results from the 2022 National Survey on Drug Use and Health. : Substance Abuse and Mental Health Services Administration; 2023 URL: https://www.issup.net/files/2025-07/SAMHSA%202022-nsduh.pdf [accessed 2026-04-23]
8. National College Health Assessment III: reference group executive summary spring 2024. : American College Health Association (ACHA); 2024 URL: https://www.acha.org/wp-content/uploads/NCHA-IIIb_SPRING_2024_REFERENCE_GROUP_EXECUTIVE_SUMMARY.pdf [accessed 2026-04-23]
9. The Healthy Minds Study: 2024-2025 data report. : The Healthy Minds Network; 2025 URL: https://healthymindsnetwork.org/wp-content/uploads/2025/09/2024-2025_HMS-National-Data-Report_Student.pdf [accessed 2026-04-23]
10. The Healthy Minds Study: 2014 data report. : The Healthy Minds Network; 2014 URL: https://healthymindsnetwork.org/wp-content/uploads/2019/04/HMS_national_DataReport_2014.pdf [accessed 2026-04-23]
11. Erekson D, Schmuck D, Lynn A, Toth L, Hoskin JM, Morrison S. The relationship between declared major, therapy utilization rates, and psychological functioning among college students. J College Stud Psychother 2023 Oct 2;37(4):318-334. [doi: 10.1080/87568225.2022.2039576]

XSL•FO
RenderX

12. Huang J, Nigatu YT, Smail-Crevier R, Zhang X, Wang J. Interventions for common mental health problems among university and college students: a systematic review and meta-analysis of randomized controlled trials. J Psychiatr Res 2018 Dec;107:1-10. [doi: 10.1016/j.jpsychires.2018.09.018] [Medline: 30300732]
13. Lattie EG, Adkins EC, Winquist N, Stiles-Shields C, Wafford QE, Graham AK. Digital mental health interventions for depression, anxiety, and enhancement of psychological well-being among college students: systematic review. J Med Internet Res 2019 Jul 22;21(7):e12869. [doi: 10.2196/12869] [Medline: 31333198]
14. Reavley N, Jorm AF. Prevention and early intervention to improve mental health in higher education students: a review. Early Interv Psychiatry 2010 May;4(2):132-142. [doi: 10.1111/j.1751-7893.2010.00167.x] [Medline: 20536969]
15. Walker J, Campbell I, Leary M. Undergraduate student perceptions of course-based mental wellness trainings. Acad Ment Health Well Being 2025;2(1). [doi: 10.20935/MHealthWellB7631]
16. Fernandez A, Howse E, Rubio-Valera M, et al. Setting-based interventions to promote mental health at the university: a systematic review. Int J Public Health 2016 Sep;61(7):797-807. [doi: 10.1007/s00038-016-0846-4] [Medline: 27364779]
17. Shapiro SL, Brown KW, Biegel GM. Teaching self-care to caregivers: effects of mindfulness-based stress reduction on the mental health of therapists in training. Train Educ Prof Psychol 2007;1(2):105-115. [doi: 10.1037/1931-3918.1.2.105]
18. Schiraldi GR, Brown SL. Preventive mental health education for functioning adults: stress, coping and mental health courses at the University of Maryland. Int J Emerg Ment Health 2002;4(1):57-63. [Medline: 12014294]
19. Alang S, McAlpine D. Treatment Modalities and Perceived Effectiveness of Treatment Among Adults With Depression. Health Serv Insights 2020;13:1178632920918288. [doi: 10.1177/1178632920918288] [Medline: 32425544]
20. Braun V, Clarke V. Thematic Analysis: A Practical Guide: Sage Publications; 2021.
21. Lincoln YS, Guba EG. Naturalistic Inquiry: Sage Publications; 1985.
22. Eysenbach G. Improving the quality of web surveys: the Checklist for Reporting Results of Internet E-Surveys (CHERRIES). J Med Internet Res 2004 Sep 29;6(3):e34. [doi: 10.2196/jmir.6.3.e34] [Medline: 15471760]
23. Kolb DA. Experiential Learning: Experience as the Source of Learning and Development: Prentice Hall; 1984. URL: https://www.researchgate.net/publication/235701029_Experiential_Learning_Experience_As_The_Source_Of_Learning_And_Development [accessed 2026-04-23]
24. Kolb DA. Experiential Learning: Experience as the Source of Learning and Development, 2nd edition: Pearson Education; 2015.
25. Mezirow J. Perspective transformation. Adult Educ 1978 Jan;28(2):100-110. [doi: 10.1177/074171367802800202]
26. Mezirow J. Transformative Dimensions of Adult Learning: Jossey-Bass; 1991.
27. Zhao GX, Zhao MY, Bean CR, Robbins AS, Mahrer NE. Well-being interventions in U.S. colleges: a scoping review from a positive higher education perspective. Front Psychol 2025;16:1601955. [doi: 10.3389/fpsyg.2025.1601955] [Medline: 40949350]
28. Pérez-Jorge D, Boutaba-Alehyan M, González-Contreras AI, Pérez-Pérez I. Examining the effects of academic stress on student well-being in higher education. Humanit Soc Sci Commun 2025;12(1):449. [doi: 10.1057/s41599-025-04698-y]
29. Chessman HM, Bourque AA. The student mental health epidemic on Our Nation's Campuses. Change Mag High Learn 2025 Mar 4;57(2):16-23. [doi: 10.1080/00091383.2025.2471260]
30. Watson A, Bond C, Madeux A, et al. When the wounded walk: exploring the lived experience of secondary traumatic stress in the emerging nursing workforce. J Adv Nurs 2025 Jul 18;81:1-20. [doi: 10.1111/jan.70086] [Medline: 40679288]
31. Watson A, Harper D, Peterson C, et al. A qualitative inquiry into stress and coping among baccalaureate students in preparation for entering the nursing workforce. J Prof Nurs 2025;58:122-131. [doi: 10.1016/j.profnurs.2025.03.010] [Medline: 40368488]

Abbreviations

R^3: reflect-recognize-reorient

Edited by B Lesselroth; submitted 28.Nov.2025; peer-reviewed by C Lei, F Moir, H Tong; revised version received 06.Apr.2026; accepted 07.Apr.2026; published 06.May.2026.

Please cite as:
Watson AL, Peterson NE, Thatcher B, Thomas M, Hunsaker S, Hooley C, Erekson D, Simpson A, Snow G, Detrick R
Experiential Course Learning, Wellness, and Higher Education: Qualitative Descriptive Study
JMIR Med Educ 2026;12:e88642
URL: https://mededu.jmir.org/2026/1/e88642
doi:10.2196/88642

XSL•FO
RenderX

© Adrianna Lorraine Watson, Neil E Peterson, Brandon Thatcher, Michael Thomas, Stacie Hunsaker, Cole Hooley, David Erekson, Adam Simpson, Gregory Snow, Rachel Detrick. Originally published in JMIR Medical Education (https://mededu.jmir.org), 6.May.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Education, is properly cited. The complete bibliographic information, a link to the original publication on https://mededu.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Efficacy of an Imaging Anatomy Virtual Simulation Experiment Course in Undergraduate Medical Education: Nonrandomized Controlled Trial

ACeng Li[1], MBBS; Lina Ma[1], MMed; Rui Peng[1], MBBS; Juan Shi[2], PhD; Mingming Zhang[2], PhD; Minwen Zheng[1], MD; Yayun Wang[3], PhD; Yayu Huang[4], MD; Jing Ren[1], MD

[1]Department of Radiology, Xijing Hospital, Air Force Medical University, No. 127, Changle West Road, Xi'an City, Shaanxi, China

[2]Teaching and Research Section of Human Anatomy and Histology and Embryology, Basic Medical College, Air Force Medical University, Xi'an City, Shaanxi, China

[3]National Basic Medical Experimental Teaching Demonstration Center, Basic Medical College, Air Force Medical University, Xi'an City, Shaanxi, China

[4]Internal Medicine Teaching and Research Office, Xijing Hospital, Air Force Medical University, Xi'an City, Shaanxi, China

Corresponding Author:
Jing Ren, MD
Department of Radiology, Xijing Hospital, Air Force Medical University, No. 127, Changle West Road, Xi'an City, Shaanxi, China

Abstract

Background: Human anatomy is the cornerstone of medicine. Currently, there is a lack of effective learning tools that can establish a correlation between 2D sectional anatomy and 3D stereoscopic anatomy and facilitate the conversion between the two.

Objective: This study aimed to evaluate the efficacy of the Imaging Anatomy Virtual Simulation Experiment (IAVSE) course in undergraduate education.

Methods: This nonrandomized controlled trial was conducted with 102 third-year medical students from the 2017 and 2018 cohorts: 50 medical students who elected to take the IAVSE course were assigned to the experimental group, and 52 students who took the traditional laboratory course served as the control group. Primary outcomes assessed by teaching evaluation included the third-year human anatomy theory scores and fourth-year medical imaging theory, practical test, and total scores. The secondary outcome was a 7-item Likert scale questionnaire based on Bloom's Taxonomy of Educational Objectives. Statistical analyses were performed using SPSS (IBM Corp), with the independent samples *t* test applied for normally distributed continuous data and the Mann-Whitney *U* test for nonnormally distributed data. A two-tailed $P<.05$ was considered statistically significant for all tests.

Results: The results of the teaching evaluation indicated a statistically significant difference between the experimental group and the control group. The experimental group achieved higher scores in human anatomy theory (experimental group: median 85, IQR 84.25 - 85.50, control group: median 81.75, IQR 78.13 - 83.00; $P<.001$), medical imaging theory (experimental group: median 61.03, IQR 60.41 - 61.53, control group: median 59.03, IQR 55.66 - 59.53; $P<.001$), practical testing (experimental group: median 22.50, IQR 21.50 - 23.00, control group: median 20.50, IQR 19.00 - 22.00; $P<.001$), and total scores (experimental group: mean 83.26, SD 2.58; control group: mean 78.46, SD 3.76; $P<.001$). Student feedback collected via a Likert questionnaire also revealed significantly higher ratings in the experimental group across multiple domains, including enjoyment, interactivity, participation, satisfaction, learning efficiency, usability, and acceptance (all *P* values <.001, except for serviceability, $P=.02$). Furthermore, the experimental group demonstrated a higher level of acceptance toward the virtual simulation course.

Conclusions: The IAVSE course effectively bridges the gap between human anatomy and medical imaging. It enhances students' spatial understanding and academic performance and stimulates their learning interest. It thus holds significant potential for broader application in undergraduate medical education.

(JMIR Med Educ 2026;12:e80393) doi:10.2196/80393

KEYWORDS

virtual simulation; human anatomy; undergraduate; undergraduate teaching; medical imaging

XSL•FO
RenderX

Introduction

The study of human anatomy is the cornerstone of medical education [1]. Through theoretical instruction, students develop a foundational understanding of the composition and morphological structure of human organs. Practical courses, particularly dissection courses, play a crucial role in anatomical teaching by facilitating the integration of theoretical knowledge with hands-on experiences [2,3].

However, traditional approaches to teaching human anatomy have primarily focused on general 3D anatomical structures, often overlooking the importance of 2D cross-sectional anatomy [4]. To understand the spatial relationships between different organs, students are required to develop a strong 3D spatial imagination [5,6]. This presented a significant challenge, especially when dealing with complex anatomical structures and their interrelationships [7]. As students progress to medical imaging in subsequent years, many find it difficult to translate 2D images into a coherent 3D understanding [8]. This difficulty is particularly pronounced among students with limited spatial visualization skills, adversely impacting their overall learning outcomes [9].

Recently, emerging virtual simulation technologies comprising augmented reality, mixed reality, and virtual reality (VR) have gained prominence in higher education driven by increased accessibility and affordability [10,11]. This technological advancement has prompted educational institutions to initiate research and development projects assessing their pedagogical efficacy when integrated into existing curricula [12]. A substantial body of evidence indicates that integrating these innovative approaches with conventional teaching methods significantly enhances the effectiveness of anatomy education for both instructors and learners [13]. As demonstrated in relevant studies, 3D visualization serves as a valuable educational tool, particularly in improving learning outcomes, by facilitating the mental representation of anatomical structures, a cognitive process essential for consolidating long-term knowledge [14,15]. Specifically, from a technical perspective, virtual simulation refers to the spatial superimposition of interactive virtual objects onto physical environments. Within medical education, this technology is being progressively adopted to provide multidimensional, intuitive learning resources that transform abstract concepts into tangible 3D models, thereby creating immersive learning experiences. These experiences are hypothesized to promote deeper understanding and knowledge retention [16]. Consequently, student enthusiasm for 3D visualization technologies in anatomical studies has increased notably, a trend mirrored by the booming market for anatomy-specific 3D learning tools [17].

This study was designed to address the critical gap between human anatomical knowledge and medical imaging interpretation. To this end, we developed the Imaging Anatomy Virtual Simulation Experiment (IAVSE) course, which specializes in imaging anatomy. This research is particularly significant as it systematically compares virtual simulation–based pedagogy with conventional didactic approaches within undergraduate medical education, using a dual evaluation framework that integrates objective competency assessments with comprehensive subjective feedback. Based on the core objectives of this study, the following research hypotheses are proposed for empirical investigation:

The VR-based IAVSE course functions as an effective transitional teaching tool connecting human anatomy and medical imaging in undergraduate medical education, resulting in improved performance in human anatomy and medical imaging assessments, heightened spatial reasoning ability, and greater knowledge retention.

Methods

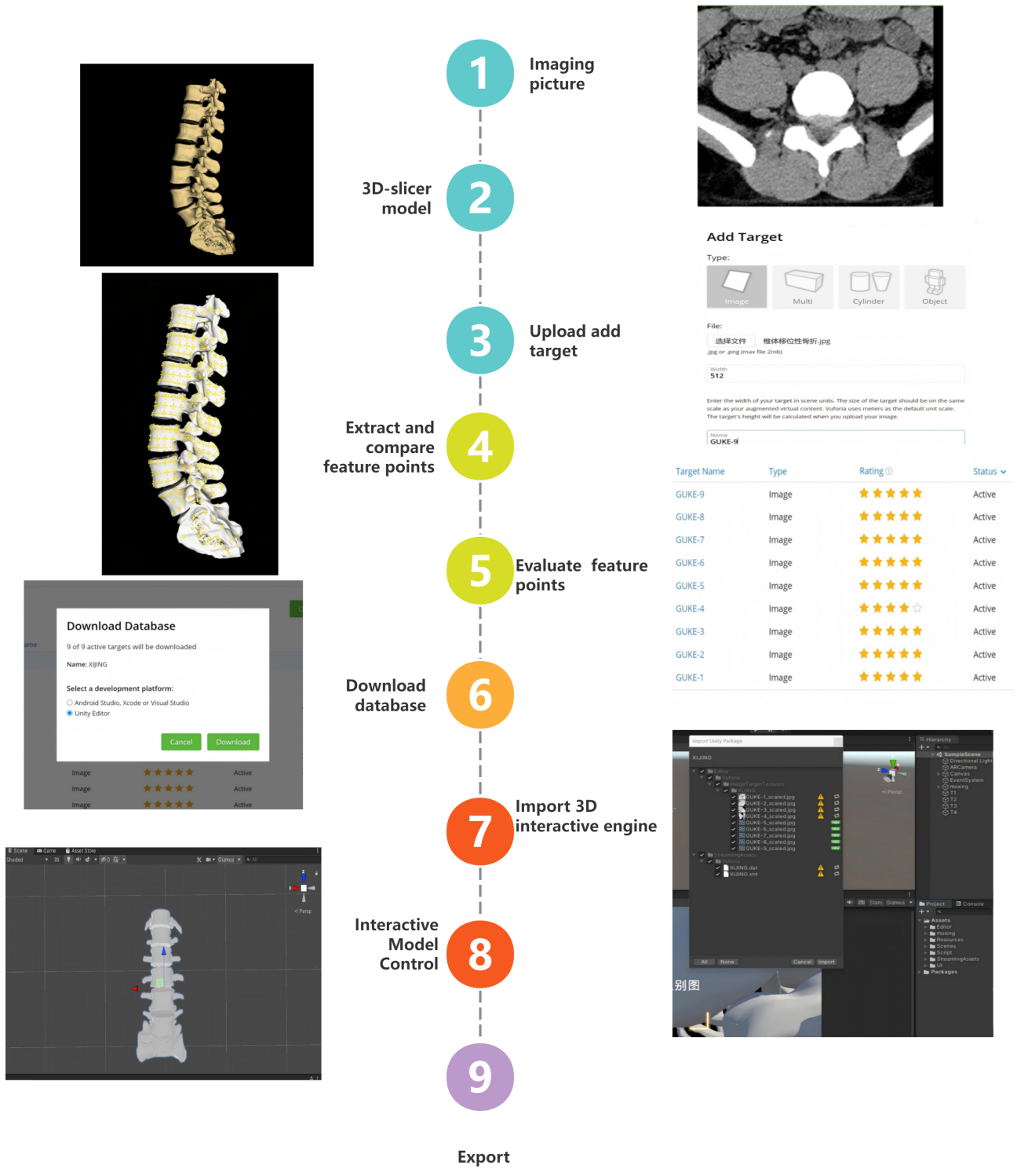
IAVSE Course

Development Process of IAVSE Course

The IAVSE course was collaboratively developed by the Human Anatomy Teaching and Research Section following a methodologically rigorous workflow (Figure 1). To meet the project requirements, Vuforia (Engine 9.x; PTC Inc), an industry-leading augmented reality engine, and Unity (2019.4 LTS), a robust application development platform, were selected as the foundational technologies. C# was used as the primary programming language for scripting interactive logic.

XSL•FO
RenderX

Figure 1. Development process of the Imaging Anatomy Virtual Simulation Experiment course.



Data Curation Phase

First, for case selection, clinically representative computed tomography or magnetic resonance imaging thin-section (<1 mm) imaging data were acquired from standardized diagnostic repositories.

Second, for preprocessing, all clinical computed tomography or magnetic resonance imaging datasets used in this study underwent a thorough deidentification process to strictly protect patient privacy. Specifically, before these datasets were used for course development, the professional software (RadiAnt DICOM Viewer; Medixant Sp. z o.o) had completely removed or anonymized all personal identifiable information of patients, including names, medical record numbers, dates of birth, and any other information that could be used to identify patients. Digital Imaging and Communications in Medicine datasets were exported with optimized windowing parameters.

3D Reconstruction Pipeline

First, for volumetric modeling, raw imaging data were processed using 3D Slicer v4.11 to generate surface-rendered anatomical models.

XSL•FO
RenderX

Second, for mesh optimization, models underwent topology-preserving decimation (polygon reduction >70%) and Laplacian smoothing for computational efficiency.

Model Target Preparation With Vuforia Engine 9.x

First, the optimized 3D model was uploaded to the Vuforia Developer Portal to create a model target, which served as a unique digital fingerprint for subsequent spatial registration in the VR teaching system.

Second, for algorithm implementation, a natural feature recognition algorithm (eg, scale-invariant feature transform variant) was used to extract scale-invariant key points (minimum 200 points/image). In tandem, the model target generator tool automatically extracted a dense set of scale-invariant geometric features (including vertex positions and normals) from the model's surface mesh, and these multidimensional features together laid a solid foundation for precise virtual model placement upon real-time recognition. Based on the extracted key points and geometric features, feature point libraries were constructed through nearest-neighbor ratio testing (threshold: 0.7); the random sample consensus algorithm was further applied for outlier rejection to eliminate false matching points caused by environmental interference, ensuring the accuracy of feature matching for spatial registration.

Third, for feature point stability and model target quality evaluation, the quality and robustness of the generated model target were rigorously evaluated to substantiate technical performance, with the model target generator tool ensuring uniform feature coverage across the model surface. Robustness was additionally tested against simulated variations in lighting and partial occlusion (typical real-world application interferences). For formal deployment, all model targets were required to meet strict quantitative metrics (core validation for engineering contributions), as follows:

- Vuforia star rating: ≥3 stars (a composite metric assessing feature quality, spatial distribution, and model complexity, ranging from 1 to 5 stars)
- Key point matching rate: >80% during runtime testing
- Tracking loss rate: <5 instances of continuous tracking loss per minute under standard experimental conditions.

Fourth, the finalized model target database, containing the model's feature descriptors, was downloaded for integration into the Unity development environment.

VR Application Development in Unity

The interactive VR learning module was developed in Unity 3D Engine (v2019.4). The downloaded Vuforia target database was imported, and the anatomical model was configured as a trackable target. Initial prototype testing confirmed a model tracking confidence level of ≥80%.

First, for XR platform configuration, the application was built for the Meta Quest 2. The Oculus XR Plugin (Unity Technologies), Vuforia (Engine 9.x), and Oculus Integration SDK (Meta Platforms) were installed via Unity's Package Manager. Player settings were configured for the Android platform with a minimum application programming interface level of 24 (Android 7.0 "Nougat") to match the Meta Quest 2 system requirements.

Second, for scene implementation, a VR camera rig was configured for stereoscopic rendering. The anatomical model was linked to the Vuforia Model Target behavior to ensure stable spatial registration. Single-pass stereo rendering was enabled to optimize graphical performance and frame rate stability.

Interaction Programming (C#)

First, for model manipulation, the users could grab, rotate, and translate the model using Oculus Touch controllers.

Second, for progressive disclosure, anatomical structures were organized in layers, allowing users to toggle their visibility incrementally via the controller trigger.

Third, for interactive labeling, mesh colliders were added to key structures. Selection was enabled via a controller ray cast, triggering the display of dynamic 3D text annotations at the hit point.

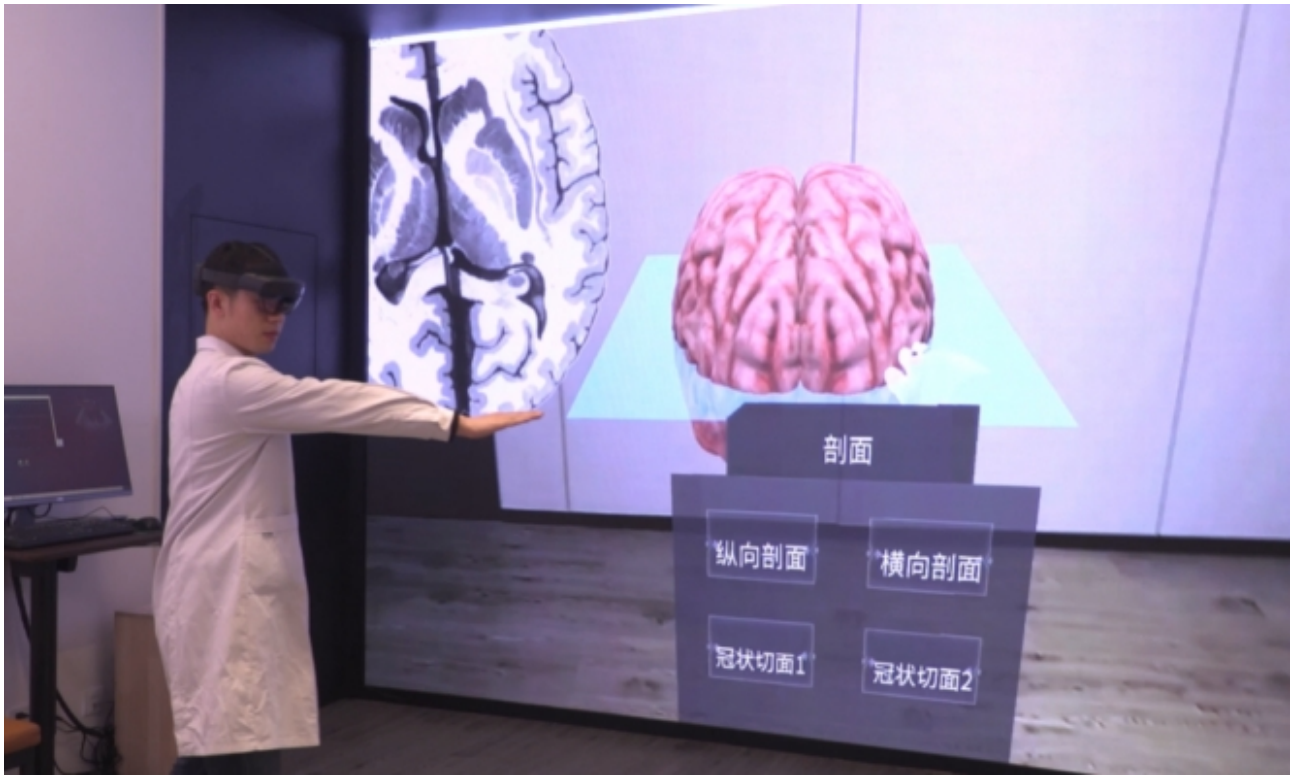
Fourth, for control scheme, the controller joystick was mapped to model rotation and translation, the trigger button cycled through anatomical layers, and the grip/side button reset the model to its initial position.

Infrastructure Configuration

The hardware platform consists of Meta Quest 2 stand-alone head-mounted displays with inside-out 6-degree-of-freedom tracking (Figure 2).

XSL•FO
RenderX

Figure 2. The classroom scene of the virtual simulation experiment in image anatomy focusing on the interaction between 3D anatomy and 2D tomographic images, incorporating game-based assessments for upgrades.



Interaction Paradigm

The interactive operation of VR devices can be reflected in three aspects: planar manipulation, volumetric exploration, and multimodal feedback. Specific details are provided below.

1. Planar manipulation: hand gestures (via optional bare-hand tracking or touch controllers) enable interaction with virtual menus and 2D interfaces.
2. Volumetric exploration: 6-degree-of-freedom manipulation enables real-time translational, rotational (360° axis-free), and translational adjustments.
3. Multimodal feedback: built-in 3D spatial audio and haptic feedback (via touch controllers) provide immersive sensory experiences, with haptic vibrations offering tactile confirmation for in-game interactions.

Study Design

Participant Recruitment and Group Allocation

A nonrandomized prospective cohort study was conducted, enrolling 102 third-year students from the 2017 and 2018 cohorts of the 8-year clinical medicine program at Air Force Medical University. During the human anatomy course, the IAVSE program was offered as an elective curriculum innovation.

Participants were allocated into two groups based on their course selection: (1) experimental group (n=50), with 20 students from the 2017 cohort and 30 from the 2018 cohort who elected to take the IAVSE course, and (2) control group (n=52), with 32 students from the 2017 cohort and 20 from the 2018 cohort who received the traditional laboratory course.

To enhance comparability between the groups, 2 key methodological safeguards were implemented:

1. Temporal alignment: both groups proceeded to complete an identical, standardized medical imaging curriculum in their subsequent (fourth) academic year.
2. Instructor consistency: the same teaching faculty supervised and delivered the imaging curriculum for both cohorts.

The sample size of this study was determined by the actual number of students who voluntarily selected the new-type elective course, which was uncontrollable in practice. Specifically, the experimental group consisted of 50 participants and the control group consisted of 52 participants, with a total sample size of 102. Due to the voluntary nature of the elective course, it was not feasible to predetermine or expand the sample size based on a priori power analysis, as student participation depended entirely on the actual registration situation of the new elective course. To verify the statistical power of the study, a post hoc power analysis was performed using G*Power 3.1 (Heinrich Heine University Düsseldorf) software. The results showed that with a total sample size of 102, a medium effect size of 0.5 (commonly used in educational intervention studies), and a significance level (α) of .05, the statistical power (1-β) of the study reached .81, which is higher than the commonly accepted threshold of .8. This indicates that the study had sufficient statistical power to detect statistically significant differences in learning outcomes between the 2 groups.

Third-Year Curriculum Framework

The third-year human anatomy curriculum for both cohorts was based on the textbook *Human Anatomy (Third Edition)* [18]. The curriculum was structured around two components:

1. Traditional theoretical instruction: a 60-hour theoretical course covering human anatomy.

XSL•FO
RenderX

2. Laboratory courses: totaling 40 hours. The laboratory component differed between groups. The control group completed a 40-hour traditional cadaver-based laboratory course. In contrast, the experimental group undertook a combined laboratory module, consisting of 24 hours of traditional cadaver-based sessions plus 16 hours of IAVSE course. All laboratory sessions were delivered synchronously with the theoretical lectures.

It is particularly noteworthy that due to the COVID-19 pandemic beginning in 2019, the school adjusted the teaching arrangements for key courses. The courses covered in this study for the 2017 and 2018 cohorts were delayed by 6 months. The 2017 cohort's human anatomy course commenced instruction in November 2020.

Control Group Pedagogy (Year 3)

The control group received the following standardized instructions are provided through: the following:

1. Traditional theoretical instruction includes syllabus-compliant PowerPoint presentations containing identical theoretical content and contact hours as the experimental group.

2. The laboratory course (traditional cadaver-based laboratory course, 40 hours) includes the following:

- Guided Ddemonstrations: Ffaculty-led reviews of key anatomical relationships.
- Active Llearning: Rrotation through prosected specimens and dissection stations and access to a video atlas system.
- Collaborative Llearning: Sstructured small-group discussions and instructor-facilitated question and answer Q&A sessions.
- Guided demonstrations: faculty-led reviews of key anatomical relationships
- Active learning: rotation through prosected specimens and dissection stations and access to a video atlas system
- Collaborative learning: structured small-group discussions and instructor-facilitated question and answer sessions

3. Independent Sstudy: includes Ppost-class review using textbook and digital slides, complemented by weekly self-assessment quizzes.

Experimental Group Intervention (Year 3)

The curriculum framework of the experimental group was as follows:

1. Traditional theoretical instructions are identical to the control group.
2. The laboratory course included 24 hours of traditional cadaver-based laboratory course plus 16 hours of IAVSE course. The IAVSE was designed as follows:
   - XR infrastructure conducted using Meta Quest 2 VR headsets, encompassing 6 one-credit modules covering major anatomical systems.

     Gamified learning program includes the following:
   - Core mechanics: real-time 3D model manipulation (rotation, zoom, and slicing) and anatomical "jigsaw puzzles" with haptic feedback
   - Formative assessment: competency checks using a composite scoring algorithm, including spatial reasoning tasks with automated feedback.

- Due to the limited availability of only 10 Meta Quest 2 units, it was not feasible for all participants in the experimental group to undergo VR training simultaneously. A total of 20 students from the 2017 cohort were divided into 2 groups, and 30 students from the 2018 cohort into 3 groups. To address this equipment constraint while preserving the statistical power of the sample size, a staggered group training protocol was adopted. Each VR training session was scheduled within 1 day of the corresponding theoretical course to maintain learning continuity. The entire VR-based IAVSE course was delivered over 7 to 8 weeks, with 2 to 3 training hours per week. All procedures, including preparatory guidance, hardware setup, and software navigation, were strictly standardized and administered by the same instructors to ensure instructional consistency and experimental reliability across all subgroups

1. Independent learning is also identical to the control group.

Fourth-Year Curriculum (Both Groups)

In the subsequent academic year, both cohorts transitioned to the same standardized medical imaging curriculum.

The core text used was the eighth edition of *Medical Imaging* [19].

Instructional standardization included (1) faculty with ≥5 years' specialized teaching experience and (2) system-based curriculum covering.

Teaching Evaluations Protocol

Human Anatomy Evaluation (Year 3)

Theoretical examination was administered postcurriculum completion, aligned with syllabus objectives. The following are the key parameters:

1. Total score: 100 points, difficulty index: 0.72 (calculated via pretest calibration)
2. Question types: multiple-choice (60 items, 1 point each) and short answer (4 structured questions, 10 points each)
3. Time allocation: 60 minutes
4. Content focus: core anatomical relationships and systemic integration

Medical Imaging Evaluation (Year 4)

A 2-stage competency assessment was implemented:

1. Theoretical examination (70 points):
   - Format: 35 multiple-choice questions (2 points each)
   - Duration: 25 minutes
   - Scope: radiographic principles and cross-sectional anatomy

1. Practice testing (30 points):
   - Task: diagnostic report writing for 3 clinical cases (10 points each)
   - Duration: 15 minutes per task

XSL•FO
RenderX

- Evaluation criteria: diagnostic accuracy (40%), anatomical correlation (30%), and clinical justification (30%)

Student Feedback

To complement the objective assessments, subjective student perceptions were coll

ected using a purpose-built questionnaire. All participants were asked to complete a 5-point Likert scale questionnaire (1=Strongly disagree, 5=Strongly agree) based on Bloom's Taxonomy of Educational Objectives [20,21]. The 7-item questionnaire includes (1) intrinsic motivation (enjoyment and satisfaction), (2) interactive engagement (participation, interactivity, and acceptance), and (3) pedagogical efficacy (learning efficiency and usability).

The questionnaire comprised 7 items designed to measure 3 key dimensions derived from the cognitive, affective, and psychomotor domains of Bloom's taxonomy:

- Cognitive domain: items assessed learning efficiency (related to comprehension, application, and analysis) and the usability of the tool (related to synthesis and evaluation).
- Affective domain: items assessed enjoyment (reflecting the receiving and responding levels) and overall satisfaction with the course (reflecting the valuing level).
- Psychomotor domain: items assessed participation (perception and set) and the quality of interactivity (guided and complex response). An item on course acceptance served as a composite measure of students' recognition of the interactive format and content.

The questionnaire was administered to students in both the experimental and control groups immediately following the intervention period.

Statistical Analysis

Statistical analysis was carried out by SPSS 26.0 (IBM Corp). The data distributions were assessed with the Kolmogorov-Smirnov test. The measurement data following normal distribution were expressed as mean (SDs), and the comparison between the groups was performed using 2 independent sample two-tailed *t* tests. Conversely, the measurement data that deviated from normal distribution were expressed as medians (IQRs), and Mann-Whitney *U* tests were used for group comparisons. Count data were presented as frequency and percentage (n/N, %), and chi-squared tests were applied for group comparisons. For the student feedback questionnaire, the internal consistency of the Likert scale items was assessed using Cronbach α, with a coefficient >.70 considered acceptable. The suitability of the data for exploratory factor analysis was evaluated using the Kaiser-Meyer-Olkin measure, with a value >0.60 deemed acceptable. A 2-tailed $P<.05$ was considered statistically significant for all tests.

Ethical Considerations

This study enrolled undergraduate medical students as participants. According to the Measures for the Review of Science and Technology Ethics (for trial implementation) (Guoke Fa Jian [2023] No. 167) issued by the Ministry of Science and Technology of the People's Republic of China [22], scientific and technological activities posing no more than minimal risk—defined as routine risks in daily life or risks equivalent to standard health checkups—qualify for simplified reviews or ethical exemption. The Measures for the Ethical Review of Life Science and Medical Research Involving Humans, issued by the Department of Science, Technology and Education of the National Health Commission [23], further clarifies that research using anonymized data that cause no bodily harm and exclude sensitive personal information is eligible for ethical review exemption. As a routine educational and teaching practice devoid of invasive interventions, this study fully meets the aforementioned exemption criteria; moreover, it aligns with the Charter of the Medical Ethics Committee of Air Force Medical University [24], which prioritizes the protection of participants' dignity, safety, and rights, and explicitly exempts routine medical teaching research, standard teaching practices, curricular interventions, and teaching efficacy evaluations from the full ethical review.

This research was conducted as an integral component of routine undergraduate medical education, with all data anonymized and deidentified prior to analysis and no personally identifiable information (eg, full names and student identification numbers) collected in strict adherence to the privacy and data protection provisions outlined in the aforementioned national regulations and institutional charter. Informed consent was waived given the study's integration into routine teaching activities, and complete data deidentification guaranteed the full protection of participants' legitimate rights and interests, consistent with the core ethical principles of respect for autonomy and risk minimization.

Results

Baseline Demographic Characteristics

The experimental (n=50) and control (n=52) groups demonstrated demographic equivalence, with no statistically significant intergroup differences in age or gender distribution. There was no statistically significant difference between the first- and second-year grade point averages of the 2 groups ($P>.05$). This demographic parity confirmed the groups' comparability for subsequent pedagogical comparisons (Table 1).

XSL•FO
RenderX

Table . Comparison of basic information of the 2 groups.

Project	Experimental group (n=50)	Control group (n=52)	*P* value
Sex ratio (female:male)	23:27	28:24	.42[a]
Age (y), median (IQR)	21 (20 - 22)	21 (20 - 23)	.45[b]
GPA1[d], mean (SD)	3.45 (0.23)	3.47 (0.21)	.07[c]
GPA2[e], mean (SD)	3.51 (0.20)	3.51 (0.17)	.44[c]

[a]Chi-square test.

[b]Wilcoxon Mann-Whitney *U* test.

[c]*t* test.

[d]GPA1: first-year grade point average.

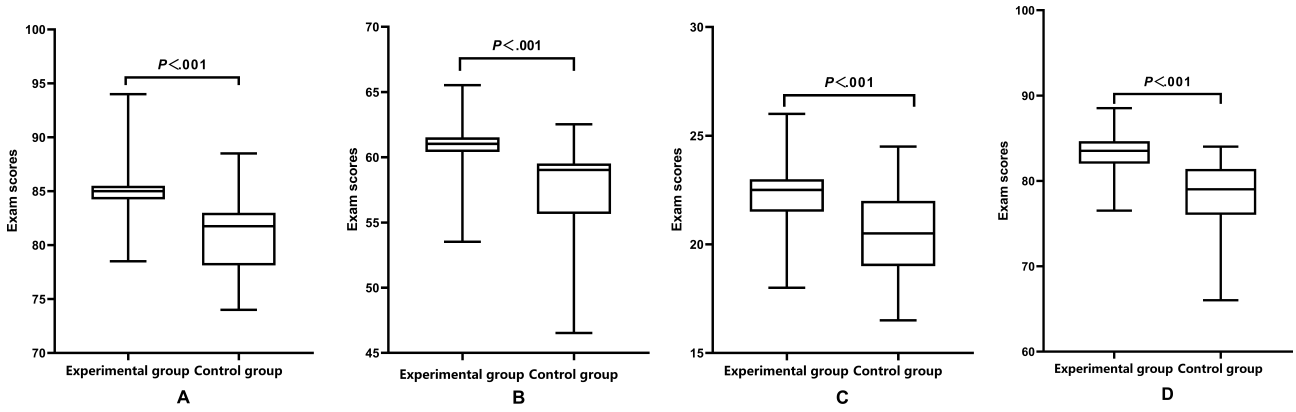
[e]GPA2: second-year grade point average.

Teaching Evaluation

Human Anatomy Theoretical Assessment

The theoretical assessment scores of the experimental group were higher than those of the control group (median 85.00, IQR 84.25 - 85.50 vs median 81.75, IQR 78.13 - 83.00), and the difference was statistically significant (*P*<.001, Figure 3).

Figure 3. Comparison of (A) human anatomy theory assessment, (B) medical imaging theory assessment, (C) practical testing, and (D) total scores of the 2 groups.



Medical Imaging Competency Evaluation

Following the completion of fourth-year medical imaging coursework, the experimental group achieved higher scores in theoretical assessment, practical testing, and total scores compared to the control group, with all differences reaching statistical significance (*P*<.001, Table 2 and Figure 3).

These significant statistical differences align with the study's core hypothesis, indicating that the IAVSE course effectively boosts core spatial reasoning ability for anatomical learning and offers direct quantitative evidence for its application in anatomical education.

Table . Comparison of medical imaging theory assessment, practical testing, and total scores of the 2 groups (points).

Project	Experimental group (n=50)	Control group (n=52)	*P* value
Theoretical examination, median (IQR)	61.03 (60.4 - 61.53)	59.03 (55.66 - 59.53)	<.001[a]
Practice testing, median (IQR)	22.50 (21.50 - 23)	20.50 (19 - 22)	<.001[a]
Total points, mean (SD)	83.26 (2.58)	78.46 (3.76)	<.001[b]

[a]Wilcoxon Mann-Whitney *U* test.

[b]*t* test.

Student Feedback

The survey achieved 100% response validity (102/102). As assessed by the Kolmogorov-Smirnov test, responses within each group deviated from a normal distribution (*P*<.05); therefore, response data are summarized using medians (IQRs).

Following the analytic procedures outlined in the *Methods* section, the psychometric properties of the questionnaire were

XSL•FO
RenderX

evaluated. The instrument demonstrated high reliability, with an overall Cronbach α=0.883. Validity assessment for factor analysis confirmed data suitability, yielding a Kaiser-Meyer-Olkin measure of 0.820 and a significant Bartlett test of sphericity (P<.001).

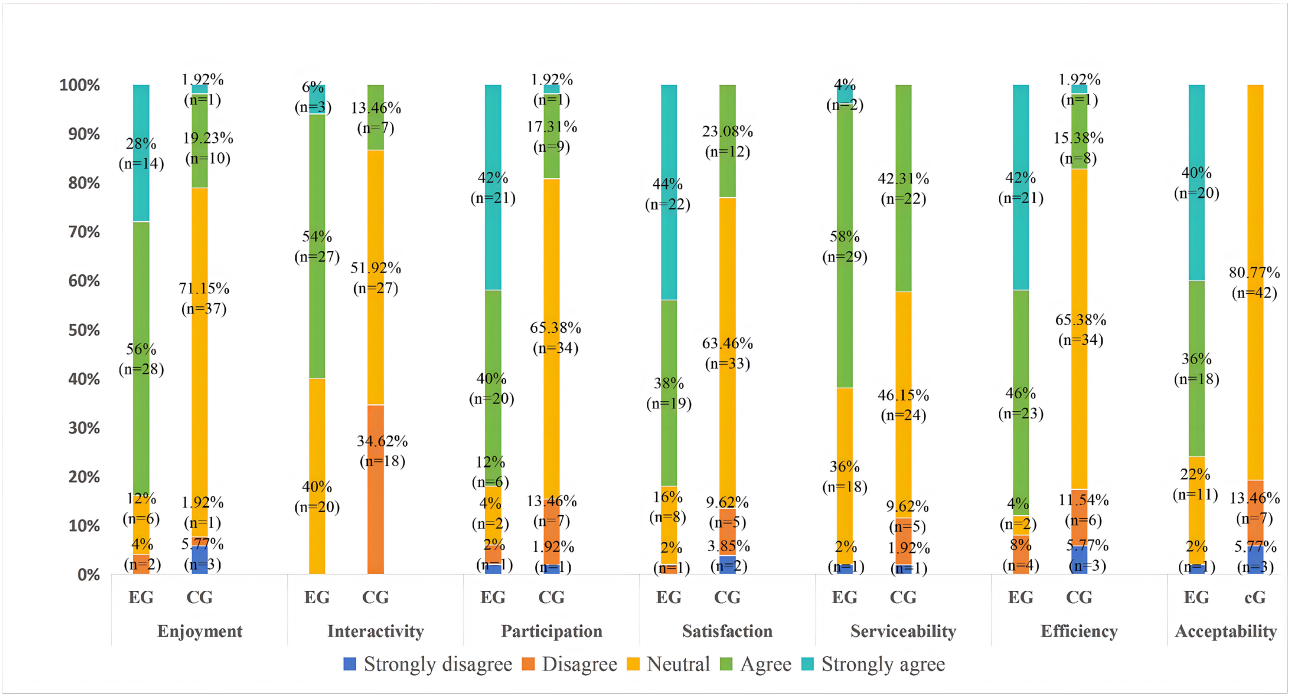
The comparative analysis of 5-point Likert scale responses revealed statistically significant advantages for the experimental group across 7 pedagogical dimensions (Table 3 and Figure 4). Such positive subjective feedback covering the cognitive, affective, and psychomotor domains indicates that the course optimizes learning experiences and facilitates competency development, thereby laying a solid foundation for better adapting to clinical practice requirements and enhancing clinical reasoning abilities.

Table . Evaluation of human anatomy (points).

Project	Experimental group (n=50), median (IQR)	Control group (n=52), median (IQR)	*Z* value	*P* value
Interesting	5 (4 - 5)	3 (3 - 3)	6.11	<.001[a]
Interactivity	4 (3 - 4)	3 (2 - 3)	5.74	<.001[a]
Degree of participation	4 (4 - 5)	3 (3 - 3)	6.13	<.001[a]
Degree of satisfaction	4 (4 - 5)	3 (3 - 3)	6.43	<.001[a]
Serviceability	4 (3 - 4)	3 (3 - 4)	2.37	.02[a]
Learning efficiency	4 (4 - 5)	3 (3 - 3)	6.56	<.001[a]
Acceptance level	4 (3.75 - 5)	3 (3 - 3)	7.41	<.001[a]

[a]Wilcoxon Mann-Whitney *U* test.

Figure 4. Evaluation of human anatomy. CG: control group; EG: experimental group.



Discussion

IAVSE Course as an Effective Bridge Program

This study pioneered the integration of a novel IAVSE course into the undergraduate medical curriculum. Using a mixed methods approach that combined objective instructor assessments with subjective student feedback, we evaluated the course's efficacy from complementary perspectives. Our findings indicate that implementing the IAVSE course following human anatomy instruction (1) was associated with significantly enhanced student performance in human anatomy and improved subsequent academic achievement in the fourth-year medical imaging course and (2) appeared to effectively foster students' comprehension of complex spatial relationships and their ability to mentally reconstruct 3D anatomy from 2D sectional images. Additionally, the course was linked to increased student engagement and self-reported learning efficacy. Taken together, these results suggest that the IAVSE course can serve as a valuable pedagogical bridge between human anatomy and medical imaging, highlighting its potential to enhance undergraduate medical education.

XSL•FO
RenderX

Synergistic Enhancement of Anatomical Understanding

The experimental group demonstrated statistically superior performance in human anatomy assessments (P<.01), validating the efficacy of the IAVSE course. This improvement can be attributed to 3 key mechanisms as follows:

1. Spatial cognition reinforcement: the IAVSE enables multiplanar manipulation of 3D anatomical models, facilitating mental rotation skills essential for understanding structural relationships.
2. Procedural memory formation: real-time virtual dissection exercises promote kinesthetic learning, aligning with evidence that haptic-enhanced simulations improve long-term knowledge retention by 23% to 28% compared to passive observation [25,26].
3. Collaborative learning optimization: the shared virtual environment fosters peer-to-peer knowledge exchange, addressing the specimen access limitations noted in traditional curricula [27,28].

Notably, our results extend previous findings by demonstrating a 15% to 20% performance gap in neurovascular anatomy modules—a critical predictor of surgical aptitude [29,30]. This advancement stems from the course's unique capacity to visualize deep anatomical strata through virtual layering techniques, overcoming the planar limitations of cadaveric dissection [31].

Foundational Preparation for Medical Imaging Competency

The results of this study show that the elective IAVSE course markedly improves the teaching efficacy of medical imaging in subsequent courses. This is demonstrated by the higher theory, practice, and total scores in medical imaging among fourth-grade students in the experimental group compared to those in the control group. Being proficient in understanding 2D tomographic anatomy, which entails grasping the organ and tissue structure, location, and adjacent morphology in sagittal and coronal sections, is essential for mastering medical imaging. However, traditional human anatomy courses mainly focus on 3D gross anatomical structures. As a result, when students first encounter medical imaging, they face difficulties in transforming these abstract 2D tomography images into a comprehensive understanding due to this disparity [32]. This challenge has persisted for many years and contributes to students' limited comprehension of medical imaging concepts. To tackle the teaching challenges, the team carried out a virtual simulation experiment in the image anatomy class. This approach was designed to enhance students' understanding of 3D anatomy and their interaction with 2D tomograms through virtual means [33]. Additionally, a game-based upgrade assessment was introduced to facilitate progress to subsequent questions and formative assessments. By synchronously mastering tomography anatomy during human anatomy learning, students were able to establish a solid foundation for future medical imaging courses, leading to significant enhancements in academic outcomes.

Demographic-Tailored Engagement Strategy

In comparison with the control group, the questionnaire results revealed a substantial improvement in students' interest, interactivity, participation, satisfaction, learning efficiency, availability, and acceptance in the experimental group. These findings suggest that the implementation of a virtual simulation experiment course on image anatomy effectively contributes to enhancing students' learning efficiency and interest [34]. Since 2018, the post-2000 generation has become the dominant demographic on university campuses. Having grown up in the information technology era, they possess active cognitive abilities and show a greater inclination to embrace novelty. The integration of simulation-based teaching methods has significantly enhanced students' learning interests by providing immersive experiences [35]. Incorporating game elements into courses, where progression to subsequent levels depends on correctly answering previous questions, increases student engagement and fosters a profound sense of accomplishment during knowledge acquisition. As a result, this approach makes the learning process more relaxed and enjoyable while stimulating overall positivity, increased satisfaction with learning outcomes, self-efficacy beliefs, and engagement with immersive technologies. It is noteworthy that the enrollment for the virtual simulation experiment course on image anatomy in 2018 saw a significant increase compared to that in 2017 (30 vs 20). This surge can be largely attributed to positive feedback and recommendations from previous students. Furthermore, it is worth mentioning that students from the 2019 and 2020 cohorts accounted for over 80% of the total enrollment, further validating their strong affinity for this course [36].

Limitations

While demonstrating pedagogical efficacy, this study presents key methodological constraints:

1. Generalizability limitations: the single-center cohort (N=102) restricts statistical power for subgroup analyses and generalizability. Institution-specific factors and cohort characteristics (eg, baseline digital literacy and learning preferences) may have constrained the applicability of results to other educational settings.
2. Design complexity: the conflation of VR simulation with traditional instruction prevented the isolation of the unique contribution of the VR component to the observed outcomes, leaving its specific mechanistic role unclear.
3. Temporal constraints: the 2-year evaluation window insufficiently captures longitudinal skill retention beyond undergraduate training.
4. Selection bias risk: as this was a newly developed elective course, and student choice was fully respected in enrollment, a nonrandomized study design was necessarily adopted, which may have introduced selection bias. Although baseline comparisons of gender, age, and prestudy GPA indicated acceptable group homogeneity, potential differences in learning motivation, digital literacy, and interest in anatomy were not evaluated using standardized questionnaires, and such pre-existing differences could have influenced the study outcomes.

XSL•FO
RenderX

5. Implementation constraints: the development of VR-based courses involves notable technical complexity and initial financial investment. Although VR may offer long-term cost and adaptability advantages [37], these initial barriers remained significant for widespread adoption during the study period.
6. A priori power analysis: it should be noted that no a priori power analysis was conducted in this study, which is mainly due to the uncontrollable sample size—this study focused on a new-type elective course, and the sample size was determined by the number of students who voluntarily registered, making it impossible to pre-estimate the sample size based on the power analysis. However, we supplemented a post hoc power analysis in the revised manuscript, and the results confirmed that the study had sufficient statistical power (1-β=.81) to detect significant differences in learning outcomes, which effectively mitigates this limitation.

Future Directions

The large-scale, multicenter randomized controlled trials (recruiting ≥500 participants across diverse institutional settings) should be conducted to strengthen the evidence base and enhance generalizability. Future studies with a randomized controlled design and comprehensive baseline psychological assessments are warranted to further validate the efficacy of the VR-based IAVSE program. Independent virtual teaching stations for comparison with traditional lecture stations should be established to isolate the impact of core teaching elements and identify the effective components of interventions. Longitudinal follow-up should be extended to students' entry into clinical residency training and their long-term professional development to assess the sustainability of educational outcomes. Subsequently, a comprehensive VR learning platform will be developed to integrate anatomical knowledge with clinical multidisciplinary modules and establish a complete learning framework, thereby more directly fostering students' core competencies, such as spatial reasoning ability and diagnostic acuity [38].

Conclusions

The IAVSE course represents a novel endeavor, serving as a vital conduit between human anatomy and medical imaging. It effectively addresses the interdisciplinary integration of these 2 domains and the inherent challenges therein. The primary objectives of this initiative are twofold: to enhance academic performance in human anatomy and to facilitate students' transition from visualizing 3D anatomical images to comprehending abstract 2D cross-sectional images by strengthening their spatial cognitive abilities. Consequently, this approach significantly improves the instructional efficacy of human anatomy and medical imaging courses while concurrently fostering students' learning interest. In turn, it creates a more engaging and enjoyable educational environment, conducive to in-depth knowledge acquisition. Ultimately, this innovative approach provides fresh insights and perspectives, offering valuable implications for advancing medical education through the integration of immersive simulation technologies.

Acknowledgments

The authors would like to give their heartfelt thanks to all the people who helped them with this paper. All authors are grateful for the support of all present and future participants as well as institutions. The authors declare the use of generative artificial intelligence (AI) during the research and writing process. According to the GAIDeT taxonomy (2025), the following tasks were delegated to generative AI tools under full human supervision: reformatting, proofreading, and editing. The generative AI tools used were Doubao and Language Tool. The responsibility for the final manuscript lies entirely with the authors. Generative AI tools are not listed as authors and do not bear responsibility for the final outcomes.

Funding

This study is funded by Shaanxi Undergraduate Education and Teaching Reform Research Key Project (23BZ089) and Key Project of Teaching and Research of Air Force Medical University (2023JXYJ003). The funding agencies do not play any role in the design, collection, analysis, or writing the manuscript.

Data Availability

The datasets generated or analyzed in this study will be publicly available.

Authors' Contributions

Conceptualization: JR
Data curation: LM (lead), RP (supporting)
Formal analysis: AL (lead), M Zheng (supporting)
Investigation: AL (lead), M Zheng (supporting)
Methodology: JR (lead), JS (supporting), M Zhang (supporting)
Resources: M Zheng (lead), RP (supporting), AL (supporting)
Software: RP (lead), LM (supporting)
Supervision: JR (lead), YW (supporting), YH (supporting)
Visualization: AL (lead), LM (supporting)

XSL•FO
RenderX

Writing—original draft: AL
Writing—review and editing: JR (lead), YW (supporting), YH (supporting)
The author(s) read and approved the final manuscript.

Conflicts of Interest

None declared.

References

1. Cawdell-Smith AJ, Pym RAE, Verrall RG, et al. Animal handling as an integrated component of animal and veterinary science programs at the University of Queensland. J Vet Med Educ 2007;34(5):542-549. [doi: 10.3138/jvme.34.5.542] [Medline: 18326761]
2. Arjomandi Rad A, Vardanyan R, Thavarajasingam SG, et al. Extended, virtual and augmented reality in thoracic surgery: a systematic review. Interact Cardiovasc Thorac Surg 2022 Jan 18;34(2):201-211. [doi: 10.1093/icvts/ivab241] [Medline: 34542639]
3. Clunie L, Morris NP, Joynes VCT, Pickering JD. How comprehensive are research studies investigating the efficacy of technology-enhanced learning resources in anatomy education? A systematic review. Anat Sci Educ 2018 May 6;11(3):303-319. [doi: 10.1002/ase.1762] [Medline: 29236354]
4. Dede C. Immersive interfaces for engagement and learning. Science 2009 Jan 2;323(5910):66-69. [doi: 10.1126/science.1167311] [Medline: 19119219]
5. Dhar P, Rocks T, Samarasinghe RM, Stephenson G, Smith C. Augmented reality in medical education: students' experiences and learning outcomes. Med Educ Online 2021 Dec;26(1):1953953. [doi: 10.1080/10872981.2021.1953953] [Medline: 34259122]
6. Duncan-Vaidya EA, Stevenson EL. The effectiveness of an augmented reality head-mounted display in learning skull anatomy at a community college. Anat Sci Educ 2021 Mar;14(2):221-231. [doi: 10.1002/ase.1998] [Medline: 32583577]
7. Efe IE, Çinkaya E, Kuhrt LD, Bruesseler MMT, Mührer-Osmanagic A. Neurosurgical education using cadaver-free brain models and augmented reality: first experiences from a hands-on simulation course for medical students. Medicina (Kaunas) 2023 Oct 9;59(10):1791. [doi: 10.3390/medicina59101791] [Medline: 37893509]
8. Finos K, Datta S, Sedrakyan A, Milsom JW, Pua BB. Mixed reality in interventional radiology: a focus on first clinical use of XR90 augmented reality-based visualization and navigation platform. Expert Rev Med Devices 2024 Aug;21(8):679-688. [doi: 10.1080/17434440.2024.2379925] [Medline: 39054630]
9. Ghosh SK. Cadaveric dissection as an educational tool for anatomical sciences in the 21st century. Anat Sci Educ 2017 Jun;10(3):286-299. [doi: 10.1002/ase.1649] [Medline: 27574911]
10. Hart LA, Wood MW, Weng HY. Mainstreaming alternatives in veterinary medical education: resource development and curricular reform. J Vet Med Educ 2005;32(4):473-480. [doi: 10.3138/jvme.32.4.473] [Medline: 16421831]
11. Hu X, Wang Y, Li J, et al. Three-dimensional multimodality image reconstruction as teaching tool for case-based learning among medical postgraduates: a focus on primary pelvic bone tumour education. BMC Med Educ 2023 Dec 12;23(1):944. [doi: 10.1186/s12909-023-04916-8] [Medline: 38087270]
12. Jeyakumar A, Dissanayake B, Dissabandara L. Dissection in the modern medical curriculum: an exploration into student perception and adaptions for the future. Anat Sci Educ 2020 May;13(3):366-380. [doi: 10.1002/ase.1905] [Medline: 31168930]
13. Jiang N, Jiang Z, Huang Y, et al. Application of augmented reality models of canine skull in veterinary anatomical education. Anat Sci Educ 2024;17(3):546-557. [doi: 10.1002/ase.2372] [Medline: 38238283]
14. Kuehn BM. Virtual and augmented reality put a twist on medical education. JAMA 2018 Feb 27;319(8):756-758. [doi: 10.1001/jama.2017.20800] [Medline: 29417140]
15. Kurul R, Ögün MN, Neriman Narin A, Avci Ş, Yazgan B. An alternative method for anatomy training: immersive virtual reality. Anat Sci Educ 2020 Sep;13(5):648-656. [doi: 10.1002/ase.1959] [Medline: 32163659]
16. Lewis TL, Burnett B, Tunstall RG, Abrahams PH. Complementing anatomy education using three-dimensional anatomy mobile software applications on tablet computers. Clin Anat 2014 Apr;27(3):313-320. [doi: 10.1002/ca.22256] [Medline: 23661327]
17. Liaw SY, Chua WL, et al. Desktop virtual reality versus face-to-face simulation for team-training on stress levels and performance in clinical deterioration: a Randomised Controlled Trial. J Gen Intern Med 2023 Jan;38(1):67-73. [doi: 10.1007/s11606-022-07557-7] [Medline: 35501626]
18. Li YQ. Human Anatomy, 3rd edition: Fourth Military Medical University Press; 2015.
19. Xu K, Gong QY, Han P. Medical Imaging, 8th edition: People's Medical Publishing House; 2018.
20. Montello MB, de Moura SAB, Abreu BJ. Bloom's taxonomy as a tool for educational objectives applied to human anatomy. Morphologie 2025 Mar;109(364):100943. [doi: 10.1016/j.morpho.2024.100943] [Medline: 39674085]
21. Tuma F, Nassar AK. Applying Bloom's taxonomy in clinical surgery: practical examples. Ann Med Surg (Lond) 2021 Sep;69(102656):102656. [doi: 10.1016/j.amsu.2021.102656] [Medline: 34429945]

XSL•FO
RenderX

22. Measures for ethical review of science and technology (trial) [Report in Chinese]. : Ministry of Science and Technology of the People's Republic of China; 2023 URL: https://figshare.com/articles/dataset/___/31742788?file=62812036 [accessed 2026-04-11]
23. Measures for the ethical review of life science and medical research involving humans [Report in Chinese]. : Department of Science, Technology and Education, National Health Commission of the People's Republic of China; 2023 URL: https://figshare.com/articles/dataset/Measures_for_the_Ethical_Review_of_Life_Science_and_Medical_Research_Involving_Humans/31742911?file=62812639 [accessed 2026-04-11]
24. Constitution of the medical ethics committee of air force medical university [Report in Chinese]. : Air Force Medical University (The Fourth Military Medical University); 2020 URL: https://figshare.com/articles/dataset/Constitution_of_the_Medical_Ethics_Committee_of_Air_Force_Medical_University/31742917?file=62812645 [accessed 2026-04-11]
25. Meyer ER, Cui D. Using stereoscopic virtual presentation for clinical anatomy instruction and procedural training in medical education. Adv Exp Med Biol 2023;1431(145-60):145-160. [doi: 10.1007/978-3-031-36727-4_7] [Medline: 37644291]
26. Moro C, Phelps C. Smartphone‐based augmented reality physiology and anatomy laboratories. Med Educ 2022 May;56(5):575-576. [doi: 10.1111/medu.14756]
27. Niedermair JF, Antipova V, Manhal S, et al. On the added benefit of virtual anatomy for dissection‐based skills. Anat Sci Educ 2023 May;16(3):439-451. [doi: 10.1002/ase.2234]
28. da Silva NP, Eloy S, Resende R. Robotic construction analysis: simulation with virtual reality. Heliyon 2022 Oct;8(10):e11039. [doi: 10.1016/j.heliyon.2022.e11039] [Medline: 36281420]
29. Ryan GV, Callaghan S, Rafferty A, Higgins MF, Mangina E, McAuliffe F. Learning outcomes of immersive technologies in health care student education: systematic review of the literature. J Med Internet Res 2022 Feb 1;24(2):e30082. [doi: 10.2196/30082] [Medline: 35103607]
30. Sinou N, Sinou N, Filippou D. Virtual reality and augmented reality in anatomy education during COVID-19 pandemic. Cureus 2023 Feb;15(2):e35170. [doi: 10.7759/cureus.35170] [Medline: 36949987]
31. Skandalakis GP, Chytas D, Paraskevas G, Noussios G, Salmas M, Fiska A. Virtual and augmented reality in anatomy education: need for comparison with other three-dimensional visualization methods. Morphologie 2022 Jun;106(353):141-142. [doi: 10.1016/j.morpho.2021.02.006] [Medline: 33762155]
32. Stanford W, Erkonen WE, Cassell MD, et al. Evaluation of a computer-based program for teaching cardiac anatomy. Invest Radiol 1994 Feb;29(2):248-252. [doi: 10.1097/00004424-199402000-00022] [Medline: 8169104]
33. Triepels CPR, Koppes DM, Van Kuijk SMJ, et al. Medical students' perspective on training in anatomy. Ann Anat 2018 May;217:60-65. [doi: 10.1016/j.aanat.2018.01.006] [Medline: 29501634]
34. Zammit C, Calleja-Agius J, Azzopardi E. Augmented reality for teaching anatomy. Clin Anat 2022 Sep;35(6):824-827. [doi: 10.1002/ca.23920] [Medline: 35657719]
35. Zeng W, Ilo C, Bowman D, Thompson J. Virtual loupes: a pilot study on the use of video passthrough augmented reality in plastic surgery. Surg Innov 2024 Oct;31(5):493-501. [doi: 10.1177/15533506241265544] [Medline: 39058601]
36. Zimmer L, Hatzl J, Uhl C, et al. Perspective or spectacle? Teaching thoracic aortic anatomy in a mixed reality assisted educational approach - a two-armed randomized pilot study. Langenbecks Arch Surg 2024 Sep 9;409(1):274. [doi: 10.1007/s00423-024-03463-0] [Medline: 39251463]
37. Garber AM, Meliagros P, Diener-Brazelle J, Dow A. Using virtual reality to teach medical students cross-coverage skills. Am J Med 2024 May;137(5):454-458. [doi: 10.1016/j.amjmed.2024.02.002] [Medline: 38364987]
38. Ogut E, Yildirim FB, Senol Y, Senol AU. Comprehensive evaluation of the educational impact and effectiveness of specialized study modules in cross-sectional anatomy: a study on student engagement and learning outcomes. BMC Med Educ 2025 Apr 10;25(1):514. [doi: 10.1186/s12909-025-07050-9] [Medline: 40211255]

Abbreviations

IAVSE: Imaging Anatomy Virtual Simulation Experiment
VR: virtual reality

XSL•FO
RenderX

Edited by S Guinez-Molinos, T Leung; submitted 09.Jul.2025; peer-reviewed by E Ogut, S Zhou, T Peng; revised version received 09.Mar.2026; accepted 12.Mar.2026; published 13.May.2026.

Please cite as:
Li A, Ma L, Peng R, Shi J, Zhang M, Zheng M, Wang Y, Huang Y, Ren J
Efficacy of an Imaging Anatomy Virtual Simulation Experiment Course in Undergraduate Medical Education: Nonrandomized Controlled Trial
JMIR Med Educ 2026;12:e80393
URL: https://mededu.jmir.org/2026/1/e80393
doi:10.2196/80393

© ACeng Li, Lina Ma, Rui Peng, Juan Shi, Mingming Zhang, Minwen Zheng, Yayun Wang, Yayu Huang, Jing Ren. Originally published in JMIR Medical Education (https://mededu.jmir.org), 13.May.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Education, is properly cited. The complete bibliographic information, a link to the original publication on https://mededu.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Experiences and Perceptions of Clinical and Graduate Medical Students Regarding AI in Syria: Cross-Sectional Study

Jouni Aljoudeh[1]; Abdulrahman Al Balkhi[2,3]; Yahia Ranjous[3,4]; Abdulrahman Shbani[5]; Danial Takieddin[3]; Mohammad Bashar Izzat[4], PhD

[1]Department of Otolaryngology, Faculty of Medicine, Damascus University, Damascus, Syrian Arab Republic

[2]Gastroenterology Department, Police Hospital, Harasta, Syrian Arab Republic

[3]Faculty of Medicine, Damascus University, Damascus, Syrian Arab Republic

[4]Department of Surgery, Faculty of Medicine, Damascus University, Damascus, Syrian Arab Republic

[5]Faculty of Medicine, Tartous University, Tartous city, Syrian Arab Republic

Corresponding Author:
Abdulrahman Shbani
Faculty of Medicine, Tartous University, Tartous city, Syrian Arab Republic

Abstract

Background: Artificial intelligence (AI) tools have revolutionized various aspects of education and health care in recent years. Their influence extends across multiple domains of medical education, from traditional learning to research and foreign language acquisition.

Objective: This study aims to evaluate the experiences and perceptions of AI tools usage in a low-resource setting and identify the factors influencing their adoption.

Methods: A cross-sectional study was conducted to evaluate the experiences with AI tools and perceptions regarding their future applications in education and health care among medical students in Syria. The sample was equally divided between clinical-year students and graduates. Chi-square tests analyzed differences based on demographics and experience, while Mann-Whitney *U* tests compared group perceptions of AI's future role. Factors studied included academic year, gender, German language learning, computer access, and research experience.

Results: Among 400 participants, AI tools were widely used for study preparation (228/400, 57% of participants), assignments (160/400, 40% of participants), and research. Clinical students used AI more than graduates for examination preparation (P<.001), creating cases (P=.03), and writing tasks (P<.001). Males used AI more for research (P=.004) or anatomy (P=.02); German learners relied on AI for language tasks. Despite 76% (304/400) of students believing AI would enhance residency training and 71.8% (287/400) of students supporting institutional policies, only 25.5% (102/400) of students expected career benefits. Ethical concerns were higher among females and researchers.

Conclusions: This study highlights the increasing reliance on AI tools among medical students and graduates for academic and clinical purposes. The highest usage was reported in study preparation, writing tasks, and clinical simulations. Significant differences in AI usage were observed based on academic level, gender, access to technology, and research experience. While perceptions were largely positive, concerns remained around ethical use, potential job displacement, and diminished human interaction in medicine. These findings underscore the importance of developing institutional policies to guide the ethical and effective integration of AI in medical education.

(JMIR Med Educ 2026;12:e84942) doi:10.2196/84942

KEYWORDS

artificial intelligence; AI; ChatGPT; medical education; cross-sectional; Syria

Introduction

Artificial intelligence (AI) refers to modern systems that use self-learning algorithms to perform tasks and generate responses that mimic human intelligence [1]. Despite its early beginnings, the surge in its popularity and success was in 2022 when OpenAI launched ChatGPT. Within just 5 days, the platform gained one million users, making it the second-fastest application in history to reach this milestone [2,3].

In the medical domain, AI has proved itself as a valuable tool for both education and health care. For students, AI assists in editing texts, generating medical case scenarios, summarizing complex topics, and simplifying difficult information, which saves time and enhances efficiency [4,5]. In health care, AI has proved itself as a valuable tool for medical practitioners. It has

XSL•FO
RenderX

passed the USMLEs (United States Medical Licensing Examinations) and proved its efficiency in disease diagnosis across multiple specialties. Its performance is particularly notable in specialties that rely on image analysis, such as radiology and pathology [6-10]. However, these advancements are not without doubts and concerns. The rapid growth of AI has sparked debates over ethical issues, job displacement in health care, and other challenges [11,12].

The integration of AI tools in medical education is crucial, as this new and fast-growing technology will shape the future of health care, despite its absence from traditional medical curricula. The World Medical Association has endorsed AI integration into medical education; however, its implementation is still not sufficient worldwide [13,14]. This gap presents a notable challenge in low-resource areas, where limited access to computers, reliable electricity, and internet connectivity further hinders adoption [15].

It is essential to understand how medical students engage with AI at different stages of their medical training. Clinical-year students (fourth and fifth years) and preresidency students (graduates and near-graduates) differ in both academic needs and clinical responsibilities. For instance, clinical students often rely on AI to prepare for examinations and explain complex information, whereas preresidency students tend to use AI for residency preparation and clinical decision-making. Comparing these groups provides valuable insights into whether AI experiences and perceptions evolve with advancing training and clinical practice.

Syria has endured a devastating war since 2011, resulting in millions of refugees both inside and outside the country [16]. The conflict has huge impacts on all aspects of life, including health care and medical education. Over half of medical facilities were destroyed or damaged during the war [17-19]. Moreover, the concurrent economic collapse and infrastructure devastation have created severe poverty and shortages in necessities, including food and electricity, among others [20]. One notable consequence is that most medical students now intend to immigrate after graduation, with Germany emerging as the primary destination for Syrian doctors [21-23]. These factors have significantly impacted medical education and access to knowledge.

Most existing literature originates from well-resourced settings, leaving limited knowledge about how AI is experienced and perceived in low recourse, conflict-affected contexts. In such settings, limited access to computers, electricity outages, poor internet connection, and lack of institutional support can hinder the effective use of AI. At the same time, these challenges may encourage greater reliance on AI tools as informal educational substitutes, offering access to information that is otherwise difficult to obtain. The implications of such use extend beyond individual learning, raising questions about digital divides, educational equity, and the risk of brain drain from already devastated health systems.

Understanding perceptions of AI among medical students requires consideration of individual and contextual factors that may shape technology acceptance. Prior literature suggests that demographic and training-related characteristics, such as gender and stage of clinical training, are associated with differing attitudes toward digital health technologies and AI adoption among medical trainees [24,25]. Additionally, prior exposure to AI tools has been consistently identified as a key determinant of perceived usefulness and acceptance within established technology acceptance frameworks [26].

Academic research experience may further influence perceptions of AI by enhancing familiarity with evidence-based practice and critical appraisal skills, potentially shaping trust and openness toward AI-supported tools in medical education and clinical contexts [26]. Accordingly, these factors were considered conceptually relevant for examining predictors of favorable AI perceptions among medical students.

This study aims to evaluate Syrian medical students' experiences with AI tools and their perceptions regarding AI's role in medical education and providing health care. We aim to examine whether war-related consequences—including limited computer access due to economic constraints and high emigration intentions—influence the experiences and perceptions of AI. Additionally, we aim to explore the differences across academic years to assess whether experiences and perceptions are evolving positively over time. These findings will provide insights into the readiness of Syria's future doctors to use AI tools in the future, despite having trained in an unsupportive environment for modern technologies.

Methods

Study Design and Setting

This study is a cross-sectional, descriptive survey conducted in April and May 2025 among medical students and recent graduates in Syria. The primary aim was to explore their experiences with AI tools and perceptions of AI's role in medical education and clinical practice.

Participants and Sampling

A total of 400 participants were enrolled, including clinical-year medical students (fourth and fifth years) and preresidency students (graduates and near-graduates). Participants were recruited through convenience sampling using academic student networks and widely used social media platforms, such as closed university-affiliated groups.

Exclusion criteria were preclinical students (first to third years) and postgraduate specialty trainees (residents). Inclusion criteria included current enrollment in clinical medical training or graduation within the past 2 years, and those who voluntarily agreed to complete the questionnaire. Participation was anonymous and uncompensated.

Note on Sampling Frame

Participants were drawn from multiple public medical faculties across various regions of Syria. The eligible population was estimated to include several thousand clinical-year students and recent graduates nationwide. Due to security and privacy considerations, specific university names are withheld.

XSL•FO
RenderX

Data Collection Instrument

Data were collected using a structured, self-administered online questionnaire delivered via a shared Google Forms (Google LLC) link, in accordance with CHERRIES (Checklist for Reporting Results of Internet E-Surveys) guidelines. The questionnaire was originally developed in English based on published literature addressing medical students' experiences and perceptions of AI in medical education and clinical practice. It was professionally translated into Arabic by experts in medical education and linguistics. The Arabic version was reviewed and refined by bilingual subject-matter experts to ensure linguistic accuracy, cultural appropriateness, and conceptual equivalence.

The perception items were intended to capture attitudinal views rather than to constitute a formal psychometric scale. Accordingly, internal consistency of the multi-item perception items was assessed using Cronbach α as an indicator of reliability. The final questionnaire was pilot-tested with a small group (≈30 students) to assess clarity and contextual relevance. Internal consistency of the multi-item perception scales was evaluated using Cronbach α, yielding acceptable reliability (α=.77 in pilot, α=.75 in full sample). As no modifications were made to the questionnaire following this assessment, these responses were retained and included in the final analytic dataset.

The survey consisted of 52 questions over 4 pages, organized into the following domains aligned with the Results section:

- Demographic Characteristics: age, gender, academic year, computer ownership (used as a proxy for socioeconomic status), language learning background (eg, German), and research experience.
- Prior Use of AI During Medical School: use of AI tools such as large language model–based chatbots (eg, ChatGPT) and AI-enabled visualization platforms across contexts, including study, examination preparation, clinical decision-making, and research writing.
- Intended Use of AI During Residency: anticipated or early-stage AI tool use for residency preparation and clinical reasoning among preresidency participants.
- Overall Perceptions of AI: measured via a 5-point Likert scale evaluating attitudes toward AI's role in education, ethics, patient care, and future impact.
- Factors Influencing AI Adoption: including clinical exposure, language learning goals, emigration intentions, and access to digital technology.

Survey Administration and Data Security

Access to the survey was restricted to researchers with password-protected accounts. The open survey link was disseminated via online academic student groups. Participation was voluntary, anonymous, and uncompensated, with no incentives offered. All questions were mandatory except for 2 optional items. A back button was available, but no summary review or adaptive questioning was used.

To protect participant confidentiality and minimize reidentification risk, no personally identifiable information (names, contact details, or institutional identifiers) or IP addresses were collected. Each Google Forms submission was treated as a unique response; duplicate or partial submissions were excluded after data cleaning.

Data Analysis

Quantitative data were analyzed using IBM SPSS Statistics (version 27; IBM Corp). Descriptive statistics (frequencies, percentages, means, or SDs) summarized participant characteristics and AI use. Inferential analyses included:

- Chi-square tests for associations between categorical variables (eg, gender and AI usage).
- Mann-Whitney *U* tests to compare AI perception differences between subgroups (eg, gender and AI usage). Due to the nonnormal distribution of Likert-scale data, perception items were treated as ordinal variables and summarized using medians and IQRs.
- Binary logistic regression to identify predictors of positive AI perceptions and behaviors, including gender, clinical status, prior AI experience, and research background. These variables were selected a priori based on theoretical relevance and evidence from prior literature on technology acceptance in medical education. Perception outcomes were dichotomized a priori into favorable ("agree" and "strongly agree") versus nonfavorable (all other responses). Prior AI usage experience was operationalized as a composite count variable reflecting the breadth of AI engagement. Each task was coded as present or absent and equally weighted. The total number of tasks represented cumulative prior AI exposure. Adjusted odds ratios (aORs) with 95% CIs were reported. A *P* value <.05 indicated statistical significance.

Given the exploratory nature of this study and the large number of comparisons performed, no formal adjustment for multiple comparisons was applied. Instead, exact *P* values were reported throughout to allow transparent interpretation of statistical significance.

Missing data were negligible, and all analyses were conducted using complete-case analysis without imputation.

Ethical Considerations

Ethical approval was obtained from the Ethical Approval for Biomedical Researchers committee (1612). All participants provided informed digital consent before survey commencement, after reading an introductory consent statement outlining this study's purpose and duration. Responses were collected anonymously and securely stored to protect privacy. Participants received no compensation for participation.

Results

Overview

Data collection took place between April and May 2025. By the end of the recruitment period in May 2025, a total of 400 clinical-year students and preresidency students had participated in this study. Data cleaning and verification were conducted, followed by statistical analyses completed in June 2025. The analysis focused on patterns of AI use, perceptions toward AI in medical education and patient care, and their associations with demographic characteristics. In the following analyses, the composite prior AI-use variable represents the number of

XSL•FO
RenderX

different educational and research tasks for which participants reported using AI; higher scores indicate AI use across a greater number of distinct tasks.

Demographic Characteristics

The sample was nearly gender-balanced, with 51.5% (206/400) of participants identifying as male and 48.5% (194/400) of participants as female. Most participants (336/400, 84%) were aged between 20 and 24 years. Half of the respondents were clinical-year students (200/400, 50%), and the other half were preresidency students. Notably, a large proportion (290/400, 72.5% of participants) expressed a desire to pursue postgraduate specialization abroad. In terms of technology and language exposure, 68.8% (275/400) of participants reported having access to a PC. German language study was reported by 39.8% (159/400) of participants, of whom 50.9% (81/159) of participants used AI tools to support their learning. Research experience was reported by 28.2% (113/400) of participants. Knowledge about AI was most commonly obtained through social media (294/400, 73.5% of participants), followed by reading research papers (60/400, 15% of participants) and discussions with IT-oriented peers (39/400, 9.8% participants; Table 1).

Table . Participant demographic data.

Characteristics	Participants
Gender, n (%)	
Man	206 (51.5)
Woman	194 (48.5)
Academic level, n (%)	
Clinical-years students	200 (50)
Preresidency students	200 (50)
Desire to travel abroad to complete specialization, n (%)	
Yes	290 (72.5)
No	110 (27.5)
German language study, n (%)	
Yes	159 (39.8)
No	241 (60.3)
Using AI[a] in German language learning, n (%)	
Yes	81 (51)
No	78 (49)
PC access, n (%)	
Yes	275 (68.8)
No	125 (31.3)
Research experience, n (%)	
Yes	113 (28.2)
No	287 (71.8)
Knowing about AI, n (%)	
Social media	294 (73.5)
IT-oriented peers	39 (9.75)
Read research about it	60 (15)
See a course about it	7 (1.75)

[a]AI: artificial intelligence.

Prior Use of AI During Medical School

Participants reported using AI for a variety of academic and clinical tasks. Use for studying or examination preparation was reported by 57% (228/400) of participants, while 40% (160/400) of participants used it to complete written assignments. AI was used to suggest research topics or questions by 35.5% (142/400) of participants and to assist in writing research papers by 33.8% (135/400) of participants. Additional uses included writing case reports (93/400, 23.3% participants), generating self-assessment questions (87/400, 21.8% participants), and drafting patient notes (47/400, 11.8% participants).

XSL•FO
RenderX

Use of AI-supported educational technologies during medical school was also common. Digital anatomy tools were used by 72.3% (289/400) of participants, computational pathology tools by 42.8% (171/400) of participants, and AI-generated clinical simulation cases by 22.5% (90/400) of participants (Table 2).

Table . Patterns of AI[a] use during medical school.

Survey questions	Participants
Tasks for which AI was used, n (%)	
Assist with studying or examination preparation	228 (57)
Generate questions to test yourself	87 (21.8)
Suggest research topics or questions	142 (35.5)
Help write research papers	135 (33.8)
Help write case reports	93 (23.3)
Help write patient notes	47 (11.8)
Help complete written assignments	160 (40)
Using advanced technology or AI, n (%)	
Digital anatomy	289 (72.3)
Computational pathology	171 (42.8)
AI-generated clinical cases	90 (22.5)

[a]AI: artificial intelligence.

Intended Use of AI During Residency

When asked about future use of AI during residency training, participants expressed high levels of interest. AI was considered useful for answering medical questions by 85.5% (342/400) of participants, exploring new medical topics or research by 82.8% (331/400) of participants, and supporting studying or examination preparation by 80.8% (323/400) of participants.

Further intended uses included writing research papers (300/400, 75% participants), writing case reports (261/400, 65.3% participants), writing patient notes (211/400, 52.8% participants), and assisting in clinical decision-making (208/400, 52% participants; Table 3).

Table . Intended applications of AI[a] during residency.

Tasks for which AI will be used	Participants
Assist with studying or examination preparation, n (%)	323 (80.8)
Help answer medical questions, n (%)	342 (85.5)
Explore new medical topics or research, n (%)	331 (82.8)
Help write research papers, n (%)	300 (75)
Help write case reports, n (%)	261 (65.3)
Help write patient notes, n (%)	211 (52.8)
Assist in clinical decision-making, n (%)	208 (52)

[a]AI: artificial intelligence.

Overall Perceptions of AI

Overview

Descriptive analysis of participant responses revealed generally favorable perceptions of AI in medical education and practice. Items were rated using a 5-point Likert scale, and positive perceptions were defined as the combined percentage of “agree” and “strongly agree” responses.

AI in Learning and Career Development

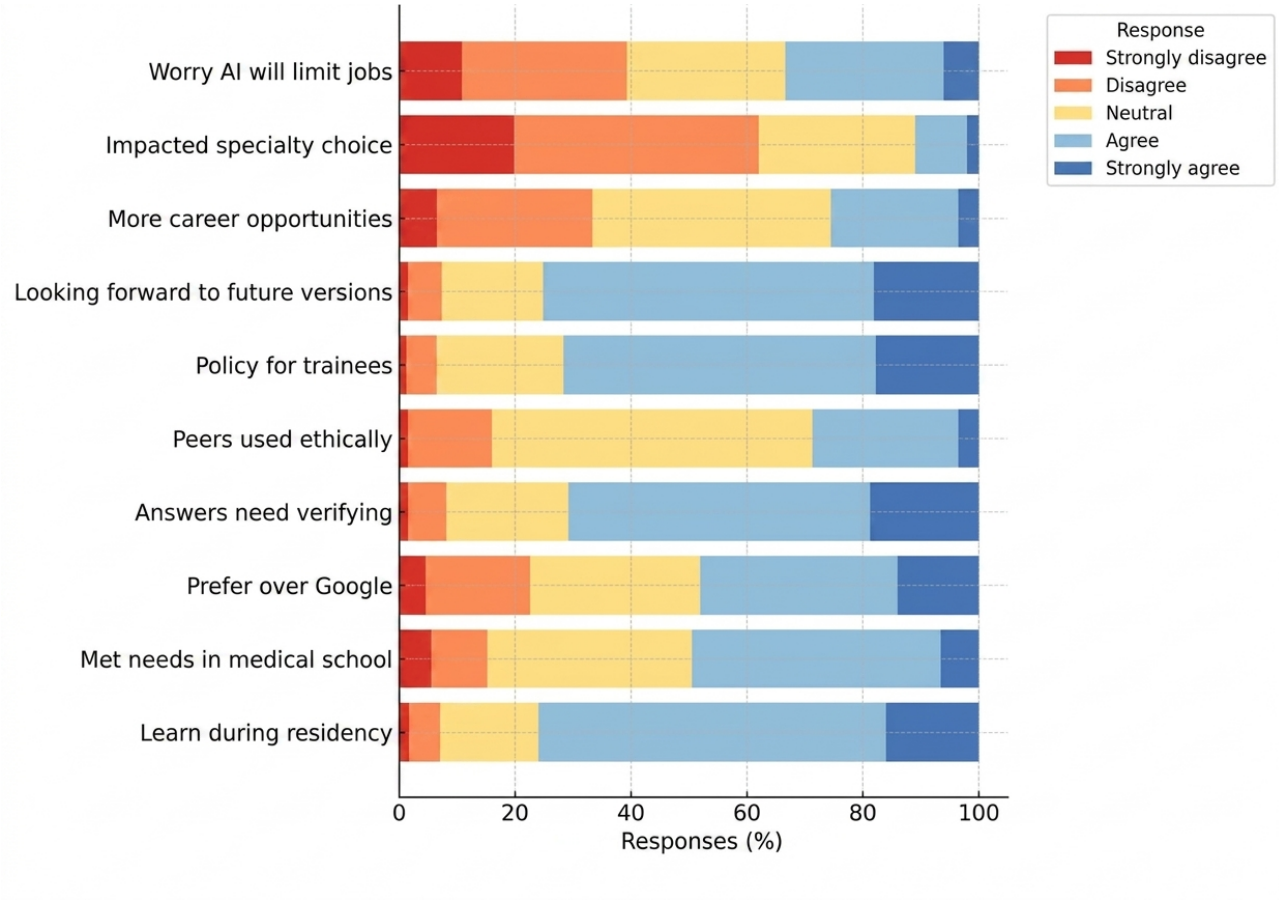
A total of 76% (304/400) of participants agreed that ChatGPT could improve their learning during residency, fewer participants agreed that AI tools effectively met their needs during medical school (189/400, 47.3% participants), and 48% (192/400) students preferred using ChatGPT over traditional resources such as Google or medical references.

The need to verify AI-generated responses was endorsed by 70.8% (283/400) of participants, 75.3% (301/400) of students expressed enthusiasm for using more advanced AI systems in their future careers, and just 28.8% (115/400) students reported that their peers have always used ChatGPT ethically. Institutional regulation of AI use by trainees was supported by 71.8% (287/400) of participants. In contrast, only 25.5% (102/400) of participants believed AI would create more career

XSL•FO
RenderX

opportunities, while 33.3% (133/400) of participants believed AI might limit job opportunities. AI was reported to have influenced specialty choice by 11% (44/400) of participants (Figure 1).

Figure 1. Perceptions of AI in education and career development. AI: artificial intelligence.

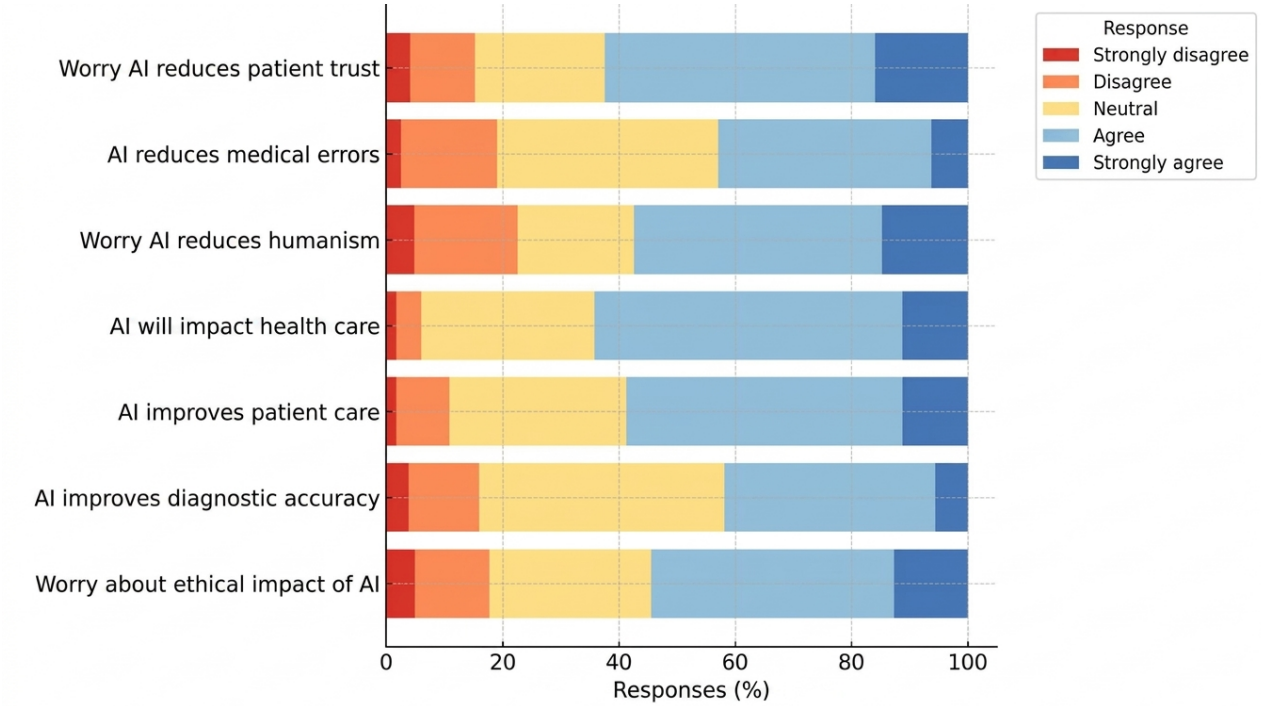


AI in Patient Care

Concerns regarding AI's ethical impact on health care were expressed by 54.6% (218/400) of participants, and 57.6% (230/400) of participants were concerned that AI might reduce the humanistic aspect of medicine. Despite these concerns, 58.8% (235/400) of participants believed AI would enhance patient care, 42.1% (168/400) of participants believed it would improve diagnostic accuracy, and 43.1% (172/400) of participants believed it would reduce medical errors. A majority (257/400, 64.3% participants) expected AI to have a major impact on the health care system, while 62.5% (250/400) of participants expressed concern that AI might reduce patient trust in physicians (Figure 2).

XSL•FO
RenderX

Figure 2. Perceptions of AI in patient care. AI: artificial intelligence.



Factors Associated With AI Use

A series of chi-square tests examined associations between demographic characteristics and patterns of AI use (academic level, gender, research experience, computer availability, and German language exposure). Significant associations were identified between subgroups and multiple AI use cases (Table 4).

XSL•FO
RenderX

Table . Prior AI[a] use by student characteristics.

AI use item and group characteristics	Values, n (%)	Chi-square (*df*)	*P* value
Assist with studying or examination preparation			
Gender		0.007 (1)	.93
Man	117 (56.8)		
Woman	111 (57.2)		
Academic level		13.219 (1)	<.001
Clinical-year student	132 (66)		
Preresidency student	96 (48)		
German language study		0.017 (1)	.90
Yes	90 (56.6)		
No	138 (57.3)		
PC access		1.855 (1)	.17
Yes	163 (59.3)		
No	65 (52)		
Research experience		0.292 (1)	.59
Yes	62 (54.9)		
No	166 (57.8)		
Generate questions to test yourself			
Gender		3.044 (1)	.08
Man	52 (25.2)		
Woman	35 (18)		
Academic level		5.303 (1)	.02
Clinical-year student	53 (26.5)		
Preresidency student	34 (17)		
German language study		0.358 (1)	.55
Yes	37 (23.3)		
No	50 (20.7)		
PC access		0.096 (1)	.76
Yes	61 (22.2)		
No	26 (20.8)		
Research experience		1.518 (1)	.22
Yes	20 (17.7)		
No	67 (23.3)		
Suggest research topics or questions			
Gender		4.258 (1)	.04
Man	83 (40.3)		
Woman	59 (30.4)		
Academic level		15.766 (1)	<.001
Clinical-year student	90 (45)		
Preresidency student	52 (26)		
German language study		0.110 (1)	.74
Yes	58 (36.5)		

XSL•FO
RenderX

AI use item and group characteristics	Values, n (%)	Chi-square (*df*)	*P* value
No	84 (34.9)		
PC access		0.579 (1)	.45
Yes	101 (36.7)		
No	41 (32.8)		
Research experience		4.252 (1)	.04
Yes	49 (43.4)		
No	93 (32.4)		
Help write research papers			
Gender		8.128 (1)	.004
Man	83 (40.3)		
Woman	52 (26.5)		
Academic level		5.915 (1)	.02
Clinical-year student	79 (39.5)		
Preresidency student	56 (28)		
German language study		8.305 (1)	.004
Yes	67 (42.1)		
No	68 (28.2)		
PC access		7.730 (1)	.005
Yes	105 (38.2)		
No	30 (24)		
Research experience		36.896 (1)	<.001
Yes	64 (56.6)		
No	71 (24.7)		
Help write case reports			
Gender		0.249 (1)	.62
Man	50 (24.3)		
Woman	42 (22.2)		
Academic level		2.368 (1)	.12
Clinical-year student	53 (26.5)		
Preresidency student	40 (20)		
German language study		0.055 (1)	.82
Yes	36 (22.6)		
No	57 (23.7)		
PC access		1.076 (1)	.30
Yes	68 (24.7)		
No	25 (20)		
Research experience		9.506 (1)	.002
Yes	38 (33.6)		
No	55 (19.2)		
Help write patient notes			
Gender		0.311 (1)	.58
Man	26 (12.6)		

XSL•FO
RenderX

AI use item and group characteristics	Values, n (%)	Chi-square (*df*)	*P* value
Woman	21 (10.8)		
Academic level		5.425 (1)	.02
Clinical-year student	31 (15.5)		
Preresidency student	16 (8)		
German language study		0.285 (1)	.59
Yes	17 (10.7)		
No	30 (12.4)		
PC access		1.231 (1)	.27
Yes	29 (10.5)		
No	18 (14.4)		
Research experience		1.278 (1)	.26
Yes	10 (8.8)		
No	37 (12.9)		
Help complete written assignments			
Gender		0.882 (1)	.35
Man	87 (42.2)		
Woman	73 (37.6)		
Academic level		12.042 (1)	<.001
Clinical-year student	97 (48.5)		
Preresidency student	63 (31.5)		
German language study		5.653 (1)	.02
Yes	75 (47.2)		
No	85 (35.3)		
PC access		0.436 (1)	.51
Yes	113 (41.1)		
No	47 (37.6)		
Research experience		0.403 (1)	.53
Yes	48 (42.5)		
No	112 (39)		
Digital anatomy			
Gender		5.158 (1)	.02
Man	159 (77)		
Woman	130 (67)		
Academic level		0.012 (1)	.91
Clinical-year student	145 (72.5)		
Preresidency student	144 (72)		
German language study		4.333 (1)	.04
Yes	124 (78)		
No	165 (68.5)		
PC access		10.286 (1)	<.001
Yes	212 (77.1)		
No	77 (61.6)		

XSL•FO
RenderX

AI use item and group characteristics	Values, n (%)	Chi-square (*df*)	*P* value
Research experience		10.976 (1)	<.001
Yes	95 (84.1)		
No	194 (67.6)		
Computational pathology			
Gender		1.049 (1)	.31
Man	83 (40.3)		
Woman	88 (45.4)		
Academic level		0.255 (1)	.613
Clinical-year student	88 (44)		
Preresidency student	83 (41.5)		
German language study		0.692 (1)	.41
Yes	72 (45.3)		
No	99 (41.1)		
PC access		1.406 (1)	.24
Yes	123 (44.7)		
No	48 (38.4)		
Research experience		1.110 (1)	.29
Yes	53 (46.9)		
No	118 (41.1)		
AI-generated cases for simulation			
Gender		1.086 (1)	.30
Man	42 (20.4)		
Woman	48 (24.7)		
Academic level		4.645 (1)	.03
Clinical-year student	54 (27)		
Preresidency student	36 (18)		
German language study		0.090 (1)	.76
Yes	37 (23.3)		
No	53 (22)		
PC access		1.135 (1)	.29
Yes	66 (24)		
No	24 (19.2)		
Research experience		0.144 (1)	.71
Yes	24 (21.2)		
No	66 (23)		

[a]AI: artificial intelligence.

Clinical-year students reported higher use of AI for studying or examination preparation compared with preresidency students (66% vs 48%; χ^2_1=13.21; *P*<.001). Similar differences were observed for generating self-assessment questions (26.5% vs 17%, χ^2_1=5.30; *P*=.02), suggesting research topics (45% vs 26%, χ^2_1=15.76; *P*<.001), writing research papers (39.5% vs 28%, χ^2_1=5.91; *P*=.02), completing written assignments (48.5% vs 31.5%, χ^2_1=12.04; *P*<.001), creating simulated clinical cases (27% vs 18%, χ^2_1=4.64; *P*=.03), and writing patient notes (15.5% vs 8%, χ^2_1=5.42; *P*=.02).

Gender differences were observed across several AI-use domains. Male participants reported greater use of AI for writing research papers (40.3% vs 26.8%; χ^2_1=8.12; *P*=.004), suggesting

XSL•FO
RenderX

research topics (40.3% vs 30.4%; χ^2_1=4.25; *P*=.04), and using digital anatomy tools (77.2% vs 67%; χ^2_1=5.15; *P*=.02).

Participants with research experience were significantly more likely to use AI for suggesting research topics (43.4% vs 32.4%; χ^2_1=4.25; *P*=.04), writing research papers (56.6% vs 24.7%; χ^2_1=36.89; *P*<.001), writing case reports (33.6% vs 19.2%; χ^2_1=9.50; *P*=.002), and using digital anatomy platforms (84.1% vs 67.6%; χ^2_1=10.97; *P*<.001).

Access to a PC was associated with greater use of AI for writing research papers (38.2% vs 24%; χ^2_1=7.73; *P*=.005) and for learning via digital anatomy platforms (77.1% vs 61.6%; χ^2_1=10.28; *P*<.001).

Students studying German reported slightly higher use of AI for language learning and for completing written assignments (47.2% vs 35.3%; χ^2_1=5.65; *P*=.02) and to help write research papers (42.1% vs 28.2%; χ^2_1=8.30; *P*=.004). This trend reflects a culturally motivated behavior, particularly among Syrian students, who often aspire to pursue postgraduate training in Germany due to the ongoing crisis. These students appeared to use AI to simulate German language acquisition, highlighting AI's potential role in educational mobility and international preparation.

Factors Associated With Perceptions of AI for Career and Education

Perceptions of AI for career and educational domains were compared across student subgroups (academic level, gender, research experience, access to a PC, and German language exposure) using Mann-Whitney *U* tests.

Preresidency students reported higher agreement that ChatGPT improves learning during residency compared with clinical-year students (median 4, IQR 4 - 4, vs 4, IQR 3 - 4; U=17,736; *P*=.03). They also showed stronger agreement that medical schools and residency programs should establish formal policies regulating AI use by trainees (median 4, IQR 4 - 4, vs 4, IQR 3 - 4; U=17,497; *P*=.02). In contrast, clinical-year students rated AI as more effective in meeting their needs during medical school than preresidency students (median 4, IQR 3 - 4, vs 3, IQR 3 - 4; U=16,501; *P*<.001).

Gender-based differences were observed across several items. Male students reported higher ratings for AI supporting learning during residency (median 4, IQR 4 - 4, vs 4, IQR 3 - 4; U=17,254.5; *P*=.007), preference for ChatGPT over other search engines (median 4, IQR 3 - 4, vs 3, IQR 2 - 4; U=16,667; *P*=.003), and enthusiasm toward future versions of AI (median 4, IQR 4 - 4, vs 4, IQR 3 - 4; U=17,264; *P*=.009).

Students without PCs reported greater influence of AI on residency specialty choice compared with those with computers (median 3, IQR 2 - 3, vs 3, IQR 2 - 3; U=15,059; *P*=.04). Conversely, students with PCs were less likely to agree that their peers consistently used AI ethically (median 3, IQR 3 - 3, vs 3, IQR 3 - 4; U=14,595; *P*=.007).

Students who had learned German rated AI as more helpful for learning during residency than those who had not (median 4, IQR 4 - 4, vs 4, IQR 3 - 4; U=16,350; *P*=.005), which may reflect AI's perceived value in language acquisition for migration-preparing students. Additionally, they showed enthusiasm toward future versions of AI (median 4, IQR 4 - 5, vs 4, IQR 3 - 4; U=16,136; *P*=.003).

Research experience was also associated with perception differences. Students with research experience were less confident in the accuracy of AI's outputs (median 4, IQR 4 - 5, vs 4, IQR 3 - 4; U=13,042; *P*<.001), more supportive of institutional AI policies (median 4, IQR 4 - 4.5, vs 4, IQR 3 - 4; U=13,535; *P*=.005), and more likely to agree that AI improves learning during residency (median 4, IQR 4 - 4, vs 4, IQR 3 - 4; U=14,272; *P*=.03). In contrast, students without research experience were more likely to perceive their peers as using AI ethically (median 3, IQR 3 - 4, vs 3, IQR 2 - 4; U=13,902; *P*=.01) and to report greater influence of AI on residency specialty choice (median 2, IQR 2 - 3, vs 2, IQR 1 - 3; U=14,175; *P*=.04). Detailed Mann-Whitney *U* test results for perceptions of AI in education and career are provided in Multimedia Appendix 1.

Factors Associated With Perceptions of AI for Patient Care

A series of Mann-Whitney *U* tests were conducted to compare perceptions of AI for patient care and professional practice among medical students' subgroups (academic level, gender, research experience, access to a PC, and German language exposure).

Gender again played an important role where male students reported significantly higher agreement that AI will improve diagnostic accuracy compared with female students (median 4, IQR 3 - 4, vs 3, IQR 3 - 4; U=15,489; *P*<.001), improve patient care (median 4, IQR 3 - 4, vs 4, IQR 3 - 4; U=15,281; *P*<.001), AI will have a major impact on health care (median 4, IQR 3 - 4, vs 4, IQR 3 - 4; U=16,171; *P*<.001), reduce medical errors (median 3, IQR 3 - 4, vs 3, IQR 3 - 4; U=16,931; *P*=.005), whereas female students worry more about the ethical impact of AI on health care (median 4, IQR 3 - 4, vs 3, IQR 3 - 4; U=17,410; *P*=.02).

Academic level was also associated with differences in perception. Preresidency students believed more that AI will have a major impact on health care during residency (median 4, IQR 3 - 4, vs 4, IQR 3 - 4; U=17,760; *P*=.03).

German language learners expressed higher agreement with the fact that AI will improve patient care (median 4, IQR 3 - 4, vs 4, IQR 3 - 4; U=16,308; *P*=.007), have a major impact on health care (median 4, IQR 3 - 4, vs 4, IQR 3 - 4; U=16,680; *P*=.02) and will enable them to make more accurate diagnoses (median 4, IQR 3 - 4, vs 3, IQR 3 - 4; U=15,716; *P*<.001).

Research experience was associated with increased concern that AI will reduce patient trust in physicians (median 4, IQR 3 - 4, vs 4, IQR 3 - 4; U=14,072; *P*=.03). Detailed Mann-Whitney *U* test results for perceptions of AI in patient care are available in Multimedia Appendix 2.

XSL•FO
RenderX

Predictors of Positive AI Perceptions

Multivariable binary logistic regression analyses were conducted to identify predictors of positive perceptions of AI across educational and clinical domains. Prior AI usage experience significantly increased the odds of a favorable perception in multiple domains: each additional AI‐use task raised the likelihood that students would agree that AI improves learning during residency (aOR=1.43, 95% CI 1.15‐1.79, *P*=.002), that AI was effective in meeting medical‐school needs (aOR=1.60, 95% CI 1.34‐1.92, *P*<.001), and that it would help reduce medical errors (aOR=1.14, 95% CI 1.01‐1.28, *P*=.04).

Clinical‐year status was consistently associated with more reserved views: clinical-year students were associated with lower odds of perceiving AI as beneficial for residency learning (aOR=0.26, 95% CI 0.11‐0.64, *P*=.003), less likely to anticipate using future versions (aOR=0.18, 95% CI 0.07‐0.49, *P*<.001), and less likely to expect AI to have a major impact on health care (aOR=0.24, 95% CI 0.09‐0.65, *P*=.005).

Gender differences appeared: female students had 3 times the odds of insisting on verifying AI's answers (aOR=3.12, 95% CI 1.36‐7.15, *P*=.007) and female students were twice as likely to worry about AI's ethical impact on health care (aOR=2.06, 95% CI 1.19‐3.54, *P*=.01).

Research experience did not significantly predict any positive perception after adjustment.

Discussion

Principal Findings

This multicenter cross-sectional survey of 400 medical students and recent graduates in the Syrian Arab Republic addresses a notable gap in the literature on AI in medical education in a low-resource, conflict-affected setting. We found that students increasingly rely on AI tools for educational and clinical purposes, with the highest reported uses being examination preparation (228/400, 57%), written assignments (160/400, 40%), and digital anatomy platforms (289/400, 72.3%). Usage patterns differed significantly by academic level (higher use among clinical-year students), gender, computer access, and research experience. Although participants were generally optimistic about AI tools' potential to enhance learning and residency training, many expressed ethical concerns and worries about reduced humanistic aspects of care; only 25.5% (102/400) of students expected direct career benefits. These findings both demonstrate the growing role of AI tools among Syrian medical trainees and underscore the need for institutional policies, targeted training, and efforts to address digital access gaps to ensure ethical and equitable integration. Thus, the summary of key findings is the high uptake of AI for study and anatomy tools; generally positive but ethically cautious perceptions; and notable subgroup differences by academic level, gender, access, and research experience.

Methodological Reminder

The cross-sectional design of this study limits causal inference; therefore, interpretations about drivers of use are based on observed associations rather than proven causal relationships.

Contextual Linkage

Overview

The national context—war-related damage, economic collapse, recurrent power cuts, and high emigration intentions—plausibly shapes patterns of access and attitudes reported here and helps explain some observed associations (eg, device ownership, language study, and intentions to train abroad). These contextual factors may influence who can access AI tools, who invests in language skills for migration, and which trainees consider leaving for postgraduate opportunities; we discuss these implications below.

Prior Use of AI During Medical School

A total of 57% (228/400) of participants reported using AI tools for examination preparation, 40% (160/400) of participants for written assignments, and 72.3% (289/400) of participants used digital anatomy tools. Further, 94% (376/400) of participants reported prior exposure to AI tools. These core results indicate high uptake of AI tools for study and academic tasks among our sample. When comparing with prior studies, reported rates vary by context and timing; we therefore present comparisons cautiously and note contextual differences (eg, study populations, timing, and local resources). For example, our experience figure is higher than a previous Syrian report (70%) [27], and our examination preparation rate (57%) exceeds a US study reporting that 48.9% of medical students use AI tools for study support [28]. Use for written assignments (40%) is higher than a UAE residency applicant sample (20.4%) [29], and use for research writing (33.8%) compares with the 24.8% in a Ugandan sample and 36.6% in a Malaysian study [30,31]. These contextual differences (population, timing, infrastructure, and educational policy) may explain variation between studies and caution against direct equivalence. Digital anatomy use (72.3%) is similar to a Canadian report of 68.1% [32].

Intended Use of AI During Residency

AI tools were reported by 85.5% (342/400) of participants as tools they would use to answer medical questions, and 82.8% (331/400) of participants indicated they would use AI tools to explore emerging medical topics or conduct research. These figures reflect intended or future use among students and recent graduates rather than measured practice among current residents. In contrast, another study among medical students revealed that 73% used AI to inquire about medical knowledge [33]. Similarly, in our study, 52% (208/400) of participants reported they would use AI tools to assist in clinical decision-making. Moreover, previous research has consistently demonstrated that some AI systems achieved high performance in certain diagnostic tasks, with reported accuracies in specific settings [34]; such performance metrics should be interpreted in context and do not imply direct transferability to routine clinical practice.

Factors Associated With AI Use

Analysis of the data reveals that the usage rate of AI tools was notably higher among males than among females across almost all fields, a pattern that aligns with findings from studies conducted in Sudan and the United Arab Emirates [35,36]. Moreover, this higher usage phenomenon may reflect differences

XSL•FO
RenderX

in attitudes, access, or self-reported behaviors rather than causal effects; we therefore describe these as associations rather than causes [37]. In addition, clinical students consistently demonstrated a greater likelihood of using AI tools across many tasks compared to both graduate and prospective students, reinforcing the notion that different academic cohorts adopt AI at varying intensities. Higher AI use among clinical students may reflect greater curricular demands or exposure; qualitative work is needed to confirm underlying reasons. However, our findings stand in contrast to a study conducted in China, which reported that graduate medical students exhibited greater awareness of medical AI usage than undergraduate students [38]. To examine socioeconomic influences, we used ownership of a PC as a proxy; access to a PC was positively associated with AI usage—an observation consistent with findings from Saudi Arabia [39].

Contextual Interpretation of Specific Variables

First, computer ownership likely reflects unequal access to digital infrastructure within the country; in the context of economic hardship and power instability, some students may lack reliable devices or connectivity, widen digital divides, and limit opportunities to use AI tools. Second, the association between German language study and higher AI use may reflect preparatory behavior among students aiming for postgraduate training abroad, who invest in language and digital tools to improve competitiveness. Third, intentions to specialize abroad raise concerns about potential brain drain: if those with greater access and digital literacy are also more likely to emigrate, local inequities in human resources for health could be exacerbated.

Although learning the German language was not statistically associated with most AI applications, students studying German reported slightly higher use of AI tools for language learning, particularly for completing written assignments (P=.02), and they used AI tools more frequently to assist with writing research papers than students not studying German. Previous studies have demonstrated that AI significantly impacts students' language learning [40]. This trend may reflect culturally motivated behavior among students preparing for postgraduate training abroad and highlights AI tools' potential role in facilitating language acquisition and international mobility.

Perceptions of AI Tools for Career, Education, and Patient Care

Overview

A total of 76% (304/400) of respondents indicated that AI tools can enhance learning outcomes, whereas an Indian study reported that 91.3% believed it could improve educational efficiency [41]. Educational policies, internet accessibility, and medical training approaches may all shape these opinions. Syrian students may face resource constraints that limit the practical benefits they can derive from AI tools. A total of 57.6% (230/400) of participants expressed concern that AI tools might diminish the human aspect of medicine. Despite technological advancements, the human touch remains a vital component of health care. AI tools should be regarded as tools to support compassionate care, rather than replace it [42]. Some students expressed concern about the potential impact of AI tools on the medical labor market. Further, 11% (44/400) of participants indicated that AI influenced their choice of specialty—a figure notably lower than the 25% reported in a study conducted in Ontario [43]. This may reflect the early stage of AI integration in Syrian clinical education or competing priorities, such as employment and migration. A total of 62.5% (250/400) of participants expressed concern that AI tools might reduce patients' trust in physicians, in contrast to a study conducted in Croatia and Slovakia, where 51.3% of students believed that patients would trust physicians less [44]. This difference may be attributed to the perception that the personal relationship between doctor and patient in Syria is deeply human and built on personal trust; therefore, the introduction of a technological element may be viewed as a direct threat to this conventional relationship. Despite some evident concerns—particularly regarding ethics and the human dimension of care—most students generally expressed optimism about integrating AI tools into both educational and clinical settings. Some skepticism existed regarding the ability of AI to enhance diagnostic accuracy, with only 42.1% (168/400) of participants agreeing to this notion. In comparison, a study conducted in London, United Kingdom, found that only 46% agreed that AI would improve clinical judgment [45]. This suggests that medical students do not view AI tools as a magical tool for enhancing diagnostics, but rather as a promising technology that should complement—rather than replace—human clinical reasoning. Approximately 33.3% (133/400) of respondents agreed that AI would limit job opportunities. In contrast, a study in Canada revealed that 98% of participants believed AI had the potential to replace some, most, or all medical activities in the field of radiation oncology [46]. This disparity may indicate that Syrian students view AI as an emerging technology designed to complement medical practice rather than to threaten it, or they might not have yet experienced practical demonstrations that reveal the autonomous capabilities of this technology.

Factors Associated With Perceptions of AI for Career and Education

Graduate and prospective graduate students expressed significantly more positive perceptions regarding AI tools' usefulness in learning during residency (P=.03). This finding suggests that AI is regarded as a valuable educational aid during clinical rotations. We attribute these perceptions to their increased exposure to clinical practice and heightened awareness of the educational challenges that AI tools can help address (eg, high caseloads, difficulties in accessing references, and time constraints). This finding is consistent with a study involving 1243 undergraduate and graduate students from 13 universities and 33 hospitals, which found that graduate medical students demonstrated a higher level of awareness regarding the use of medical AI than their undergraduate counterparts [38]. Moreover, these students were more inclined than clinical-year students to believe that medical schools and residency programs should develop policies regarding the use of AI tools by trainees (P=.02). This trend reflects their advanced awareness of the professional responsibilities and organizational challenges associated with these tools.

Male students rated AI significantly higher in supporting residency learning (P=.007), preference for ChatGPT over other

XSL•FO
RenderX

search engines (*P*=.003), and enthusiasm toward future versions of AI (*P*=.009). This may indicate that male students hold a stronger perception of the role of AI in developing clinical skills and preparing for postgraduate education.

The perceived potential of AI appeared to influence residency specialty choices more among students without PCs (*P*=.04). This may be attributed to their increased reliance on AI tools accessible via smartphones or public devices, which can serve as critical supports in making professional decisions when personal technological resources are limited. In contrast, students with PCs were more likely to believe that their peers had used ChatGPT less ethically than students without such resources (*P*=.007). This perception may stem from their deeper understanding of how these tools are used or from greater exposure to ethical discussions regarding the appropriate use of AI in research and academic tasks.

Those with experience were less confident in the accuracy of AI tools (*P*<.001), possibly due to their direct encounter with its shortcomings or biases, which led them to adopt a more cautious stance. They were also more supportive of medical schools that implement policies for the use of AI (*P*=.005), a perspective likely attributable to their awareness of the risks associated with unregulated use. Moreover, they held a more positive view regarding the utility of ChatGPT in supporting learning during residency (*P*=.03), reflecting their detailed understanding of its potential to enhance clinical education when applied within an appropriate framework. Additionally, among those without experience, residency choices were more influenced by the perceived capabilities of AI tools (*P*=.04), which may reflect a theoretical enthusiasm rather than one grounded in direct experience.

Factors Associated With Perceptions of AI for Patient Care

Gender also played a significant role. Male students demonstrated markedly higher agreement that AI tools would improve diagnostic accuracy (*P*<.001), enhance patient care (*P*<.001), exert a substantial impact on health care (*P*<.001), and reduce medical errors (*P*=.005). This finding is consistent with a study conducted in the United Arab Emirates, which reported that male students were more inclined than female students to believe that AI improves diagnostic accuracy (51.7% vs 39.7%), reduces medical errors (58.2% vs 40.8%), and enhances patient care (65.9% vs 54.6%) [29]. In contrast, female students expressed greater concern about the ethical implications of AI in health care (*P*=.02), reflecting their heightened sensitivity to the humanistic aspects of medical practice. This concern may be associated with fears of diminished human interaction or reduced trust in technology within clinical contexts.

German-language learners expressed significantly higher agreement with the statements that AI tools would improve patient care (*P*=.007), have a substantial impact on health care (*P*=.02), and enable more accurate diagnoses (*P*=.001). This may be related to the content of German-language medical programs or the medical culture associated with the German-speaking context, which may place greater emphasis on modern technologies and their clinical implications. Research experience was associated with greater concern that AI tools could diminish patients' trust in physicians (*P*=.03), possibly reflecting heightened awareness of ethical and trust issues among those engaged in research.

Limitations

While this study offers insights into AI tools' perceptions and use among Syrian medical trainees, limitations include convenience sampling, reliance on self-reported data subject to recall and social desirability bias, use of proxy variables (eg, computer ownership for socioeconomic status), and lack of prestudy sample size calculation; these factors limit generalizability and preclude causal inference. We explicitly note that PC ownership was used as a proxy for income; however, this is an imperfect and 1D indicator that does not capture the multidimensional nature of socioeconomic status. The cross-sectional design and self-reported measures limit causal inference; hypotheses about drivers of AI tool use require longitudinal or qualitative validation. Further, this study involved a large number of statistical comparisons across multiple domains. Although no formal adjustment for multiple comparisons was applied, exact *P* values were reported throughout to support transparent interpretation of the findings, and results with marginal statistical significance should be interpreted with caution. Additionally, items referring to "during residency" capture intended future use for most respondents and do not measure observed practice among current residents.

Further Studies

Future research should consider longitudinal or mixed method designs to better understand evolving perceptions of AI tools throughout medical training and early clinical practice. Qualitative methodologies, including interviews and focus groups, could deepen insights into the reasoning behind student attitudes—especially regarding trust, ethics, and specialty choices. Expanding the sample to include students from private institutions, rural areas, or international campuses would improve the representativeness of findings. Policymakers and educators should consider equity-focused interventions that address device and connectivity gaps and strategies to retain trained graduates to mitigate potential brain drain.

Conclusions

This study reveals a predominantly optimistic outlook among Syrian medical students and residents regarding the role of AI in education and clinical care. High levels of prior usage suggest both accessibility and growing technological literacy, though notable concerns—particularly ethical and relational—underscore the need for guided integration. Differences across gender, training stage, and socioeconomic proxies reflect the nuanced landscape of AI perception in medical education. While enthusiasm is widespread, cautious appraisal of AI's limitations and its potential to disrupt traditional human-centered care remains essential. As such, integrating AI into medical education must be performed strategically, emphasizing critical appraisal, ethical awareness, and reinforcement of compassionate clinical reasoning. These findings contribute meaningfully to the regional and global

XSL•FO
RenderX

discourse on responsible AI adoption in health professions education.

Acknowledgments

Artificial intelligence tools were only used in grammar correction and typesetting corrections.

Funding

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Authors' Contributions

JA wrote the results and analyzed the data, AAB wrote the Discussion, YR wrote the Introduction, AS wrote the Methods, DT collected the data, and MBI supervised the research. All authors participated in conceptualizing, reviewing the literature, preparing the survey, and revising the final draft. All authors read and approved this final paper.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Detailed Mann-Whitney U test results for perceptions of AI in education and career. AI: artificial intelligence.
[DOCX File, 3028 KB - mededu_v12i1e84942_app1.docx]

Multimedia Appendix 2
Detailed Mann-Whitney U test results for perceptions of AI in patient care. AI: artificial intelligence.
[DOCX File, 3021 KB - mededu_v12i1e84942_app2.docx]

Checklist 1
CHERRIES Checklist.
[DOCX File, 2579 KB - mededu_v12i1e84942_app3.docx]

References

1. Amisha , Pathania M, Rathaur VK. Overview of artificial intelligence in medicine. J Family Med Prim Care 2019 Jul;8(7):2328-2331. [doi: 10.4103/jfmpc.jfmpc_440_19] [Medline: 31463251]
2. Introducing ChatGPT. OpenAI. 2022 Nov 30. URL: https://openai.com/index/chatgpt/ [accessed 2026-04-24]
3. Duarte F. Number of ChatGPT users (January 2026). Exploding topics. 2026 Apr 2. URL: https://explodingtopics.com/blog/chatgpt-users [accessed 2026-04-24]
4. Sallam M. ChatGPT utility in healthcare education, research, and practice: systematic review on the promising perspectives and valid concerns. Healthcare (Basel) 2023 Mar 19;11(6):887. [doi: 10.3390/healthcare11060887] [Medline: 36981544]
5. Rasul T, Nair S, Kalendra D, Robin M, Ladeira WJ. The role of ChatGPT in higher education: benefits, challenges, and future research directions. JALT 2023 May 10;6(1):41-56. [doi: 10.37074/jalt.2023.6.1.29]
6. Kung TH, Cheatham M, Medenilla A, et al. Performance of ChatGPT on USMLE: potential for AI-assisted medical education using large language models. PLOS Digit Health 2023 Feb;2(2):e0000198. [doi: 10.1371/journal.pdig.0000198] [Medline: 36812645]
7. Gilson A, Safranek CW, Huang T, et al. How does ChatGPT perform on the United States Medical Licensing Examination (USMLE)? The implications of large language models for medical education and knowledge assessment. JMIR Med Educ 2023 Feb 8;9:e45312. [doi: 10.2196/45312] [Medline: 36753318]
8. McGenity C, Clarke EL, Jennings C, et al. Artificial intelligence in digital pathology: a systematic review and meta-analysis of diagnostic test accuracy. npj Digit Med 2024 May 4;7(1):114. [doi: 10.1038/s41746-024-01106-8] [Medline: 38704465]
9. Popa SL, Ismaiel A, Brata VD, et al. Artificial intelligence and medical specialties: support or substitution? Med Pharm Rep 2024 Oct;97(4):409-418. [doi: 10.15386/mpr-2696] [Medline: 39502760]
10. Daniyal M, Qureshi M, Marzo RR, Aljuaid M, Shahid D. Exploring clinical specialists' perspectives on the future role of AI: evaluating replacement perceptions, benefits, and drawbacks. BMC Health Serv Res 2024 May 9;24(1):587. [doi: 10.1186/s12913-024-10928-x] [Medline: 38725039]
11. Rony MKK, Parvin MR, Wahiduzzaman M, Debnath M, Bala SD, Kayesh I. "I wonder if my years of training and expertise will be devalued by machines": concerns about the replacement of medical professionals by artificial intelligence. SAGE Open Nurs 2024;10:23779608241245220. [doi: 10.1177/23779608241245220] [Medline: 38596508]

XSL•FO
RenderX

12. Astărăstoae V, Rogozea LM, Leaşu F, Ioan BG. Ethical dilemmas of using artificial intelligence in medicine. Am J Ther 2024;31(4):e388-e397. [doi: 10.1097/MJT.0000000000001693] [Medline: 38662923]
13. Chen CH, Mullen AJ. COVID-19 can catalyze the modernization of medical education. JMIR Med Educ 2020 Jun 12;6(1):e19725. [doi: 10.2196/19725] [Medline: 32501809]
14. WMA statement on augmented intelligence in medical care. World Medical Association. 2019 Oct 26. URL: https://www.wma.net/policies-post/wma-statement-on-augmented-intelligence-in-medical-care/ [accessed 2026-04-24]
15. Wahl B, Cossy-Gantner A, Germann S, Schwalbe NR. Artificial intelligence (AI) and global health: how can AI contribute to health in resource-poor settings? BMJ Glob Health 2018;3(4):e000798. [doi: 10.1136/bmjgh-2018-000798] [Medline: 30233828]
16. Syria emergency. UNHCR. URL: https://reporting.unhcr.org/operational/situations/syria-situation [accessed 2026-04-24]
17. Alnasser AA, Darwish M, Aldbis A, et al. Renewable resilience in conflict: lessons learned from Syria's solar-powered electric health vehicles. Front Public Health 2025;13:1560063. [doi: 10.3389/fpubh.2025.1560063] [Medline: 40241971]
18. Fouad FM, Sparrow A, Tarakji A, et al. Health workers and the weaponisation of health care in Syria: a preliminary inquiry for The Lancet-American University of Beirut Commission on Syria. Lancet 2017 Dec 2;390(10111):2516-2526. [doi: 10.1016/S0140-6736(17)30741-9] [Medline: 28314568]
19. Alhaffar M, Janos S. Public health consequences after ten years of the Syrian crisis: a literature review. Global Health 2021 Sep 19;17(1):111. [doi: 10.1186/s12992-021-00762-9] [Medline: 34538248]
20. Daher J. Dark times: syria struggles with increasingly longer power cuts. Al Majalla. 2024 May 27. URL: https://en.majalla.com/node/317941 [accessed 2026-04-24]
21. Sawaf B, Abbas F, Idris A, Al Saadi T, Ibrahim N. Specialty preference and intentions to study abroad of Syrian medical students during the crisis. BMC Med Educ 2018 Mar 16;18(1):39. [doi: 10.1186/s12909-018-1146-x] [Medline: 29548290]
22. Faiola A, Brady K. 6,000 Syrians work as doctors in Germany. Some weigh whether to stay or go. 2025 Mar 11. URL: https://www.washingtonpost.com/world/2025/03/11/germany-syrian-doctors-refugees/ [accessed 2026-04-24]
23. Soqia J, Yakoub-Agha L, Mohamad L, et al. Syrian crises effect on specialty choice and the decision to work in the country among residents of six major hospitals in Syria, Damascus. PLoS One 2024;19(2):e0295310. [doi: 10.1371/journal.pone.0295310] [Medline: 38329984]
24. Liu M, Cheng Y, Li S, et al. Exploring medical students' attitudes and perceptions toward artificial intelligence in medicine in Shandong Province, China. BMC Med Educ 2025 Dec 15;26(1):106. [doi: 10.1186/s12909-025-08465-0] [Medline: 41398585]
25. Haluza D, Wernhart A. Does gender matter? Exploring perceptions regarding health technologies among employees and students at a medical university. Int J Med Inform 2019 Oct;130:103948. [doi: 10.1016/j.ijmedinf.2019.08.008] [Medline: 31442846]
26. Heinrichs H, Kies A, Nagel SK, Kiessling F. Physicians' attitudes toward artificial intelligence in medicine: mixed methods survey and interview study. J Med Internet Res 2025 Aug 26;27:e74187. [doi: 10.2196/74187] [Medline: 40857713]
27. Swed S, Alibrahim H, Elkalagi NKH, et al. Knowledge, attitude, and practice of artificial intelligence among doctors and medical students in Syria: a cross-sectional online survey. Front Artif Intell 2022;5:1011524. [doi: 10.3389/frai.2022.1011524] [Medline: 36248622]
28. Zhang JS, Yoon C, Williams DKA, Pinkas A. Exploring the usage of ChatGPT among medical students in the United States. J Med Educ Curric Dev 2024;11:23821205241264695. [doi: 10.1177/23821205241264695] [Medline: 39092290]
29. Alkhaaldi SMI, Kassab CH, Dimassi Z, et al. Medical student experiences and perceptions of ChatGPT and artificial intelligence: cross-sectional study. JMIR Med Educ 2023 Dec 22;9:e51302. [doi: 10.2196/51302] [Medline: 38133911]
30. Ajalo E, Mukunya D, Nantale R, et al. Widespread use of ChatGPT and other artificial Intelligence tools among medical students in Uganda: a cross-sectional study. PLoS ONE 2025;20(1):e0313776. [doi: 10.1371/journal.pone.0313776] [Medline: 39787055]
31. Pallivathukal RG, Soe HHK, Donald PM, Samson RS, Ismail ARH. ChatGPT for academic purposes: survey among undergraduate healthcare students in Malaysia. Cureus 2024 Jan;16(1):e53032. [doi: 10.7759/cureus.53032] [Medline: 38410331]
32. Wood KN, McWatt S, Risha N, Mensink P. Analyzing the use of digital and physical anatomical teaching models in anatomy training. Adv Physiol Educ 2025 Jun 1;49(2):572-581. [doi: 10.1152/advan.00030.2025] [Medline: 40251704]
33. Weidener L, Fischer M. Artificial intelligence in medicine: cross-sectional study among medical students on application, education, and ethical aspects. JMIR Med Educ 2024 Jan 5;10:e51247. [doi: 10.2196/51247] [Medline: 38180787]
34. Rao A, Pang M, Kim J, et al. Assessing the utility of ChatGPT throughout the entire clinical workflow: development and usability study. J Med Internet Res 2023 Aug 22;25:e48659. [doi: 10.2196/48659] [Medline: 37606976]
35. Ahmed WMM, Abdalmotalib MM, Elbadawi MH, et al. Shaping the future of medical education: a cross-sectional study on ChatGPT attitude and usage among medical students in Sudan. PLoS ONE 2025;20(5):e0324698. [doi: 10.1371/journal.pone.0324698] [Medline: 40408361]
36. Sallam M, Elsayed W, Al-Shorbagy M, et al. ChatGPT usage and attitudes are driven by perceptions of usefulness, ease of use, risks, and psycho-social impact: a study among university students in the UAE. Front Educ 2024;9:1414758. [doi: 10.3389/feduc.2024.1414758]

XSL•FO
RenderX

37. Sunmboye K, Strafford H, Noorestani S, Wilison-Pirie M. Exploring the influence of artificial intelligence integration on personalized learning: a cross-sectional study of undergraduate medical students in the United Kingdom. BMC Med Educ 2025 Apr 17;25(1):570. [doi: 10.1186/s12909-025-07084-z] [Medline: 40247306]
38. Li Q, Qin Y. AI in medical education: medical student perception, curriculum recommendations and design suggestions. BMC Med Educ 2023 Nov 9;23(1):852. [doi: 10.1186/s12909-023-04700-8] [Medline: 37946176]
39. Al Shahrani A, Alhumaidan N, AlHindawi Z, et al. Readiness to embrace artificial intelligence among medical students in Saudi Arabia: a national survey. Healthcare (Basel) 2024 Dec 11;12(24):2504. [doi: 10.3390/healthcare12242504] [Medline: 39765931]
40. Chen A, Zhang Y, Jia J, Liang M, Cha Y, Lim CP. A systematic review and meta‐analysis of AI‐enabled assessment in language learning: design, implementation, and effectiveness. J Comput Assisted Learn 2025 Feb;41(1):e13064. [doi: 10.1111/jcal.13064]
41. Rani S, Kumari A, Ekka SC, Chakraborty R. Perception of medical students and faculty regarding the use of artificial intelligence (AI) in medical education: a cross-sectional study. Cureus 2025 Jan;17(1):e77514. [doi: 10.7759/cureus.77514] [Medline: 39958022]
42. Bhardwaj A. Promise and provisos of artificial intelligence and machine learning in healthcare. J Healthc Leadersh 2022;14:113-118. [doi: 10.2147/JHL.S369498] [Medline: 35898671]
43. Mehta N, Harish V, Bilimoria K, et al. Knowledge and attitudes on artificial intelligence in healthcare: a provincial survey study of medical students. MedEdPublish 2021;10(1):75. [doi: 10.15694/mep.2021.000075.1]
44. Malešević A, Kolesárová M, Čartolovni A. Encompassing trust in medical AI from the perspective of medical students: a quantitative comparative study. BMC Med Ethics 2024 Sep 2;25(1):94. [doi: 10.1186/s12910-024-01092-2] [Medline: 39223538]
45. Banerjee M, Chiew D, Patel KT, et al. The impact of artificial intelligence on clinical education: perceptions of postgraduate trainee doctors in London (UK) and recommendations for trainers. BMC Med Educ 2021 Aug 14;21(1):429. [doi: 10.1186/s12909-021-02870-x] [Medline: 34391424]
46. Kennedy T, Collie L, Nabhen JJ, Ynoe de Moraes F. Canadian oncology residents' knowledge of and attitudes towards artificial intelligence and machine learning. JCO 2022 Jun 1;40(16_suppl):e13587-e13587. [doi: 10.1200/JCO.2022.40.16_suppl.e13587]

Abbreviations

AI: artificial intelligence
aOR: adjusted odds ratio
CHERRIES: Checklist for Reporting Results of Internet E-Surveys
USMLE : United States Medical Licensing Examination

Edited by AH Sapci; submitted 27.Sep.2025; peer-reviewed by CY Su, E Nixon; revised version received 13.Mar.2026; accepted 05.Apr.2026; published 19.May.2026.

Please cite as:
Aljoudeh J, Al Balkhi A, Ranjous Y, Shbani A, Takieddin D, Bashar Izzat M
Experiences and Perceptions of Clinical and Graduate Medical Students Regarding AI in Syria: Cross-Sectional Study
JMIR Med Educ 2026;12:e84942
URL: https://mededu.jmir.org/2026/1/e84942
doi:10.2196/84942

© Jouni Aljoudeh, Abdulrahman Al Balkhi, Yahia Ranjous, Abdulrahman Shbani, Danial Takieddin, Mohammad Bashar Izzat. Originally published in JMIR Medical Education (https://mededu.jmir.org), 19.May.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Education, is properly cited. The complete bibliographic information, a link to the original publication on https://mededu.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Changes in Self-Reported Empathy After a Basic Medical Communication Course Among Premedical Students: Single-Group Longitudinal Study

Seong Ju Jeon[1]; Hyo Hyun Yoo[2], PhD

[1]College of Medicine, Jeonbuk National University, Jeonju-si, Jeonbuk-do, Republic of Korea

[2]Department of Medical Education, College of Medicine, Jeonbuk National University, 20, Geonji-ro, Deokjin-gu, Jeonju-si, Jeonbuk-do, Republic of Korea

Corresponding Author:
Hyo Hyun Yoo, PhD
Department of Medical Education, College of Medicine, Jeonbuk National University, 20, Geonji-ro, Deokjin-gu, Jeonju-si, Jeonbuk-do, Republic of Korea

Abstract

Background: Empathy is essential to patient-centered care, yet erosion of empathy has been reported during medical training. Although educational interventions may support empathy development, longitudinal evidence on whether these changes persist remains limited, particularly in non-Western settings.

Objective: This study examined changes in overall, emotional, and cognitive self-reported empathy scores and confidence in applying empathy at baseline, immediately after, and 6 months after a basic medical communication course among Korean premedical students. It also explored whether trajectories in self-reported empathy scores differed by gender.

Methods: In this single-group longitudinal study, 82.1% (119/145) of second-year premedical students completed surveys at all 3 time points. The required 15-week course included 12 instructional hours on medical communication theory, nonviolent communication, and peer clinical interview role-play. No formal communication- or empathy-focused instruction was provided during follow-up. Self-reported empathy scores were measured using the Korean student version of the Jefferson Scale of Empathy, and 2 additional items assessed confidence in applying empathy in daily life and in physician-patient relationships. Repeated-measures and mixed-design repeated-measures ANOVAs were used.

Results: Overall self-reported empathy scores changed significantly over time ($F_{2,236}$=13.41; P<.001), increasing from baseline (mean 5.30, SD 0.57) to immediately after the course (mean 5.55, SD 0.62) and remaining above baseline at 6 months (mean 5.48, SD 0.60). Emotional empathy scores increased from baseline to immediately after the course and remained above baseline at 6 months (mean 5.43, SD 0.58; mean 5.66, SD 0.63; and mean 5.59, SD 0.61, respectively; $F_{2,236}$=10.67; P<.001), as did cognitive empathy scores (mean 4.10, SD 1.18; mean 4.50, SD 1.30; and mean 4.51, SD 1.24, respectively; $F_{2,236}$=8.80; P<.001). Female students had higher overall empathy scores across assessments than male students ($F_{1,117}$=6.09; P=.02), but time-by-gender interactions were not significant for overall, emotional, or cognitive empathy ($F_{2,234}$=0.33 and P=.72, $F_{2,234}$=0.32 and P=.73, and $F_{2,234}$=0.12 and P=.89, respectively). Confidence in applying empathy in daily life increased from baseline to immediately after the course and remained higher at 6 months (mean 5.02, SD 1.04; mean 5.26, SD 0.97; and mean 5.26, SD 0.82, respectively; $F_{2,236}$=5.18; P=.006). Confidence in applying empathy in physician-patient relationships showed a significant overall time effect (mean 4.91, SD 1.09; mean 5.13, SD 1.04; and mean 5.13, SD 0.97, respectively; $F_{2,236}$=3.46; P=.03), but Bonferroni-adjusted pairwise comparisons were not significant (baseline vs immediately after the course, adjusted P=.07; baseline vs 6-month follow-up, adjusted P=.06; and immediately after the course vs 6-month follow-up, adjusted P>.99).

Conclusions: Participation in a basic medical communication course was associated with higher self-reported empathy scores and greater confidence in applying empathy in daily life, with gains maintained at 6 months; empathy trajectories did not differ significantly by gender. Controlled studies are needed to determine the independent and sustained effects of the course.

(JMIR Med Educ 2026;12:e92215) doi:10.2196/92215

KEYWORDS

empathy; medical education; premedical students; physician-patient relationship; communication education

XSL•FO
RenderX

Introduction

Empathy is the ability to understand and share another person's feelings from that person's perspective and comprises cognitive and emotional dimensions [1]. In medicine, empathy is essential for forming therapeutic relationships, enabling physicians to better understand patients' concerns and build trust [2,3]. Empathetic physician-patient relationships have been associated with better clinical outcomes, patient satisfaction, and treatment adherence [2,3]. Research on perceptions of good physicians also suggests that patients value physicians who demonstrate both clinical competence and empathy [4]. Together, these findings support the view that empathy is a core construct in medical education and assessment, underscoring the need to foster empathy-related competence early in medical training [4,5].

Despite its importance, empathy decline during medical training has been widely reported and is often described as an "erosion of empathy" [6-8]. Empathy levels have been reported to vary by gender and educational experience [9,10]. These findings suggest that empathy should not be treated as a fixed trait but as a construct that may change across medical education. Because empathy scores during medical school have also been linked to later empathic behavior in residency training [11], efforts to support empathy may be valuable from the early stages of training, before students enter more intensive clinical environments [6,12,13].

Although empathy may have personal or dispositional components, educational reinforcement is considered important for developing and sustaining empathy-related attitudes and behaviors [14]. While several international programs have aimed to enhance medical students' empathy [15,16], longitudinal evidence on the durability of such interventions remains limited [17], particularly in Korean medical education. Previous Korean studies have primarily examined differences by gender, academic year, and educational system or have focused on developing empathy measurement tools [18,19]. However, longitudinal evidence on changes in empathy after structured empathy- or communication-focused educational programs in Korean premedical settings remains limited. In addition to self-reported empathy, students' confidence in applying empathy may provide an exploratory indicator of their perceived readiness to use empathic communication.

Accordingly, this single-group longitudinal study aimed to examine changes in overall, emotional, and cognitive self-reported empathy scores and confidence in applying empathy among Korean premedical students at baseline, immediately after, and 6 months after a required basic medical communication course. It also explored whether empathy trajectories differed by gender.

Methods

Study Design and Setting

This study used a single-group longitudinal design to examine changes in self-reported empathy and confidence in applying empathy at 3 time points: before the course, immediately after the final examination, and 6 months after course completion. The study was conducted at Jeonbuk National University College of Medicine. The basic medical communication course was delivered during the second year of premedical education as a required face-to-face course.

Participants

Participants were second-year premedical students enrolled in the basic medical communication course. Of 145 eligible students, 119 (82.1%) completed surveys at all 3 assessment points and were included in the longitudinal analyses. The final sample included 63% (75/119) male students and 37% (44/119) female students.

Survey Administration

Surveys were administered online using Google Forms at 3 time points: before the first class, immediately after the final examination, and 6 months after course completion. At each time point, the survey link was distributed only to students enrolled in the basic medical communication course.

The first page of the online survey provided information about the study purpose and procedures, the voluntary nature of participation, and confidentiality. Students proceeded to the survey only after indicating informed consent by checking the consent item.

Students used an anonymous code to allow for longitudinal matching of responses across the 3 survey time points. Survey participation was voluntary and unrelated to attendance, examination scores, final course grades, or academic standing.

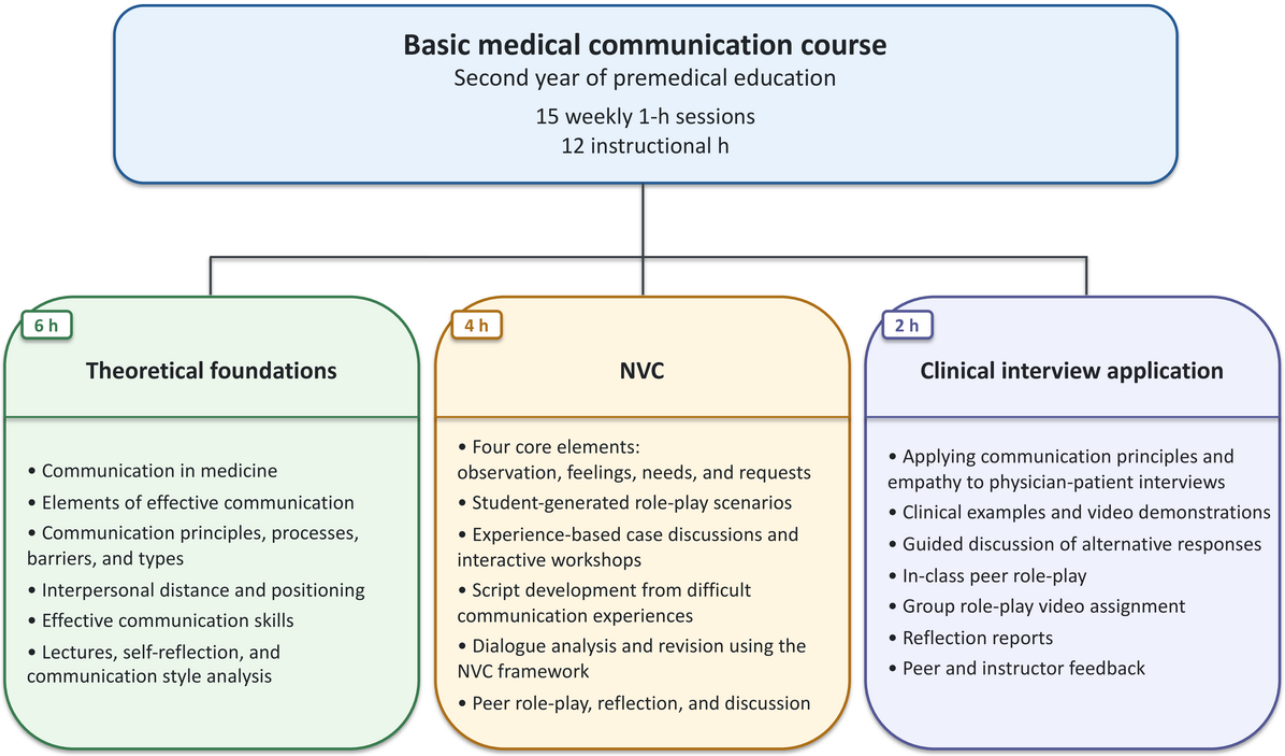
Overview of the Educational Program

The educational program evaluated in this study was the basic medical communication course, a required face-to-face course delivered during the second year of premedical education. Within the broader medical communication curriculum at our institution, this course precedes three subsequent required courses—Medical Communication 1, Medical Communication 2, and Medical Communication 3—as well as simulated patient interview training in another required course, Introduction to Clinical Medicine. No formal communication- or empathy-focused course was offered between the immediately postcourse assessment and the 6-month follow-up for this cohort.

The course comprised 15 weekly 1-hour sessions. Of these, 12 hours were instructional sessions, whereas 3 hours were allocated to orientation, the midterm examination, and the final examination. The 12 instructional hours were organized into 3 components: theoretical foundations of medical communication, nonviolent communication (NVC), and application of communication principles to clinical interview contexts. An overview of the course is shown in Figure 1.

XSL•FO
RenderX

Figure 1. Overview of the basic medical communication course. NVC: nonviolent communication.



The theoretical component comprised 6 hours and was delivered mainly through face-to-face lectures. This component covered the importance of communication in medicine, elements of effective communication, basic principles and processes of communication, barriers and types of communication, interpersonal distance and positioning, and effective communication skills. Learning activities included lectures, self-reflection, and communication style analysis.

The NVC component comprised 4 hours and included student-generated role-play scenarios, experience-based case discussions, and interactive workshops. Students were introduced to the 4 core elements of NVC: observation, feelings, needs, and requests [20]. In this component, students developed scripts based on recent experiences in which communication had been ineffective or difficult, analyzed the communication difficulty using the NVC framework, revised the dialogue, and practiced the revised conversation through peer role-play. These activities were followed by reflection and discussion.

The clinical interview application component comprised 2 hours and focused on connecting communication principles and empathy to physician-patient interview contexts. This component included lectures, clinical examples, video demonstrations of physician-patient interviews, guided discussion, and role-play scenarios. During class, students discussed possible responses to specific clinical communication situations and compared alternative responses in terms of contextual appropriateness. Students then practiced in-class peer role-play based on the clinical examples. As an additional assignment, students worked in groups of 7 to 10 to develop their own physician-patient communication scenario and produce a role-play video. Students submitted reflection reports with the video assignment. The submitted videos were reviewed in class, followed by peer and instructor feedback.

Measures

Self-reported empathy was assessed using the Korean student version of the Jefferson Scale of Empathy (S-JSE-K) adapted from the Jefferson Scale of Empathy [21] and translated and validated for Korean medical students [19,22]. The S-JSE-K consists of 18 items rated on a 7-point Likert scale and includes 2 domains: emotional empathy (16 items) and cognitive empathy (2 items). Scale and subscale scores were calculated as item means, with higher scores indicating greater self-reported empathy. Internal consistency in this sample was good for the total scale (Cronbach α=0.86) and emotional empathy (Cronbach α=0.88) and acceptable for cognitive empathy (Cronbach α=0.76).

In addition, 2 single items assessed confidence in applying empathy in daily life and in physician-patient relationships. These items were rated on the same 7-point Likert scale, with higher scores indicating greater perceived confidence.

Statistical Analysis

Means and SDs were calculated for overall empathy, emotional empathy, cognitive empathy, and the 2 confidence items at each assessment point. Repeated-measures ANOVA was used to examine within-subject changes over time, and Bonferroni-adjusted pairwise comparisons were conducted post hoc among baseline, immediately postcourse, and 6-month follow-up scores. Sphericity was assessed using the Mauchly test. To examine whether empathy trajectories differed by gender, mixed-design repeated-measures ANOVA was conducted with time as the within-subject factor and gender as the between-subject factor. Partial η^2 was reported as an effect size. All analyses were performed using PASW Statistics (version 18.0; SPSS Inc). Two-sided *P* values of less than .05 were considered statistically significant.

XSL•FO
RenderX

Ethical Considerations

The institutional review board of Jeonbuk National University approved this study (2020-02-002-002). All participants received information about the study purpose and procedures before survey completion and provided informed consent. Participation was voluntary and did not affect course grades or academic standing. No financial compensation, course credit, or other incentives were provided for survey participation. Survey data were analyzed in deidentified form. The study was conducted in accordance with the principles of the Declaration of Helsinki.

Results

Changes in Self-Reported Empathy

Repeated-measures ANOVA showed a significant effect of time on overall empathy ($F_{2,236}$=13.41; P<.001; partial η^2=0.10). Bonferroni-adjusted pairwise comparisons showed that overall empathy was higher immediately after the course than at baseline (adjusted P<.001) and remained higher at the 6-month follow-up than at baseline (adjusted P<.001). The immediately postcourse and 6-month follow-up scores did not differ significantly (adjusted P=.65).

Significant time effects were also observed for emotional empathy ($F_{2,236}$=10.67; P<.001; partial η^2=0.08) and cognitive empathy ($F_{2,236}$=8.80; P<.001; partial η^2=0.07). Emotional empathy was higher immediately after the course than at baseline (adjusted P<.001) and higher at follow-up than at baseline (adjusted P=.007), whereas the immediately postcourse and follow-up scores did not differ significantly (adjusted P=.51). Cognitive empathy was higher immediately after the course than at baseline (adjusted P=.002) and higher at follow-up than at baseline (adjusted P=.001), whereas the immediately postcourse and follow-up scores did not differ significantly (adjusted P>.99). Detailed Bonferroni-adjusted pairwise comparisons can be found in Table S1 in Multimedia Appendix 1. These results are summarized in Table 1 and illustrated in Figure 2.

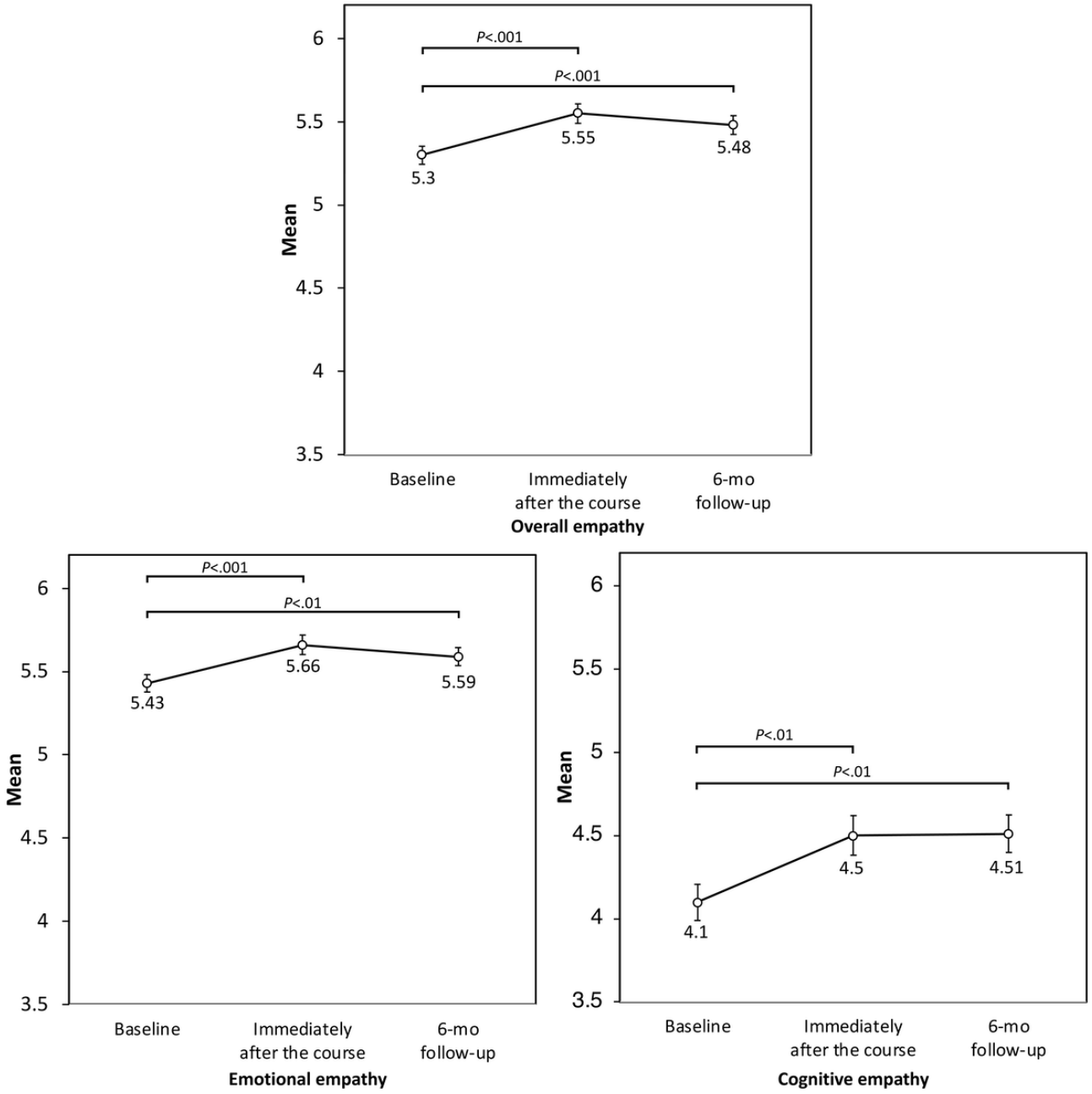
Table . Changes in self-reported empathy across time points[a].

	Baseline, mean (SD)	Immediately after the course, mean (SD)	6-mo follow-up, mean (SD)	*F* test (*df*)	*P* value	Partial η^2
Overall empathy	5.30 (0.57)	5.55 (0.62)	5.48 (0.60)	13.41 (2,236)	<.001	0.10
Emotional empathy	5.43 (0.58)	5.66 (0.63)	5.59 (0.61)	10.67 (2,236)	<.001	0.08
Cognitive empathy	4.10 (1.18)	4.50 (1.30)	4.51 (1.24)	8.80 (2,236)	<.001	0.07

[a]S-JSE-K scores were calculated as item means and range from 1 to 7, with higher scores indicating greater self-reported empathy.

XSL•FO
RenderX

Figure 2. Changes in overall, emotional, and cognitive empathy scores across time points. Error bars indicate SEs.



Gender Differences in Empathy Trajectories

At baseline, female students reported higher overall empathy scores than male students (mean 5.42, SD 0.58 vs mean 5.23, SD 0.56), and this pattern remained at the immediately postcourse assessment (mean 5.71, SD 0.56 vs mean 5.45, SD 0.63) and 6-month follow-up (mean 5.64, SD 0.52 vs mean 5.39, SD 0.64). Similar patterns were observed for emotional and cognitive empathy; gender-stratified means and SDs are provided in Table S2 in Multimedia Appendix 1. Mixed-design repeated-measures ANOVA showed no significant time-by-gender interaction for overall empathy ($F_{2,234}$=0.33; P=.72), emotional empathy ($F_{2,234}$=0.32; P=.73), or cognitive empathy ($F_{2,234}$=0.12; P=.89), suggesting no evidence that empathy trajectories differed by gender. Female students had higher mean scores across the 3 assessment points for overall empathy ($F_{1,117}$=6.09; P=.02; partial η^2=0.05), emotional empathy ($F_{1,117}$=4.91; P=.03; partial η^2=0.04), and cognitive empathy ($F_{1,117}$=4.78; P=.03; partial η^2=0.04). These results are summarized in Table 2 and illustrated in Figure 3.

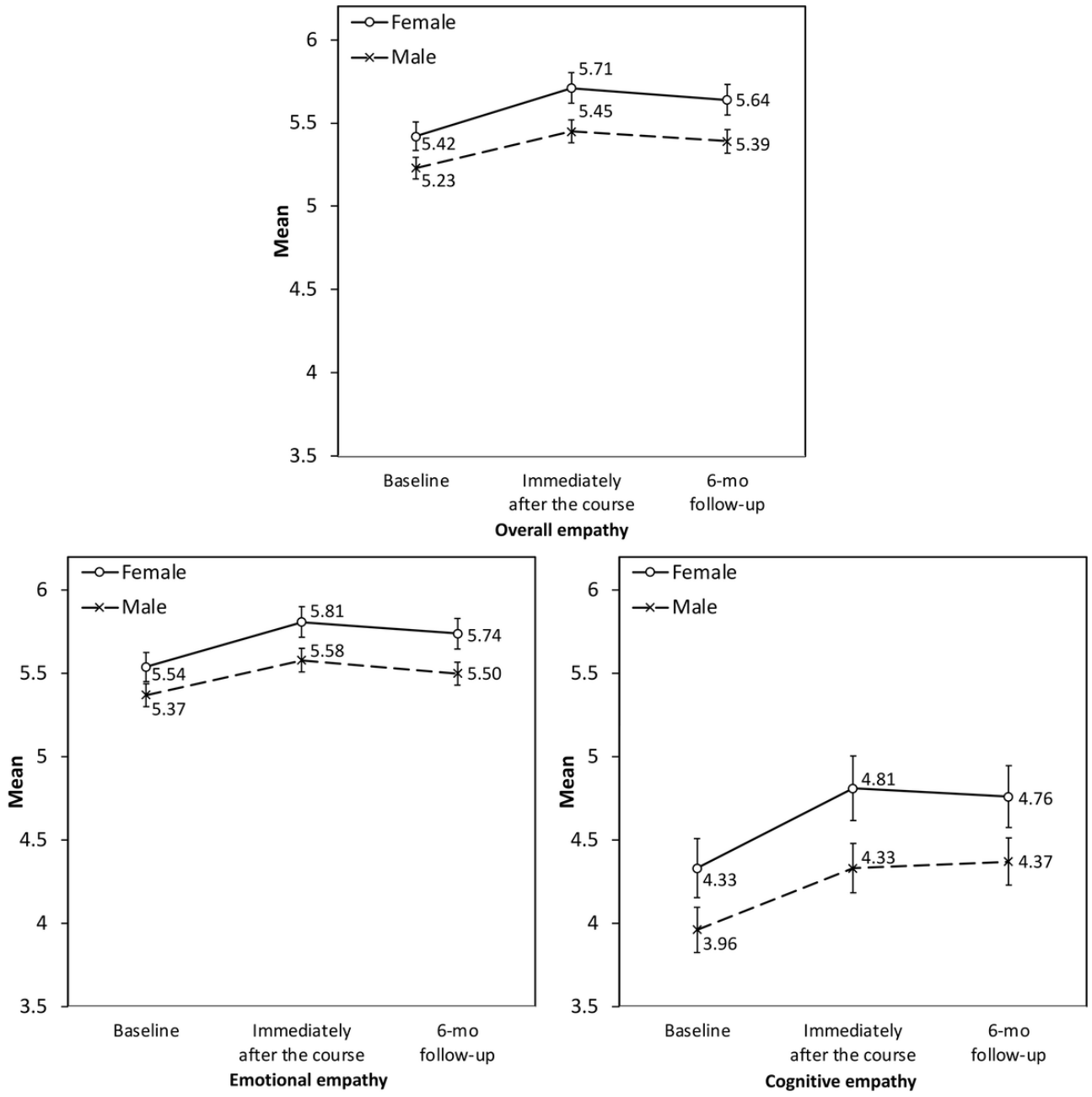
XSL•FO
RenderX

Table . Mixed-design repeated-measures ANOVA results for gender differences in empathy over time.

Effect	*F* test (*df*)	*P* value	Partial η^2
Overall empathy			
Time	13.48 (2,234)	<.001	0.10
Time × gender	0.33 (2,234)	.72	<0.01
Gender main effect	6.09 (1,117)	.02	0.05
Emotional empathy			
Time	10.77 (2,234)	<.001	0.08
Time × gender	0.32 (2,234)	.73	<0.01
Gender main effect	4.91 (1,117)	.03	0.04
Cognitive empathy			
Time	8.49 (2,234)	<.001	0.07
Time × gender	0.12 (2,234)	.89	<0.01
Gender main effect	4.78 (1,117)	.03	0.04

XSL•FO
RenderX

Figure 3. Changes in overall, emotional, and cognitive empathy scores across time points by gender. Error bars indicate SEs.



Changes in Confidence in Applying Empathy

Repeated-measures ANOVA showed a significant effect of time on confidence in applying empathy in daily life ($F_{2,236}$=5.18; P=.006; partial η^2=0.04). Bonferroni-adjusted pairwise comparisons showed that confidence in daily life was significantly higher immediately after the course than at baseline (adjusted P=.02) and remained significantly higher at the 6-month follow-up than at baseline (adjusted P=.02). The immediately postcourse and follow-up scores did not differ significantly (adjusted P>.99).

Confidence in applying empathy in physician-patient relationships also showed a significant overall effect of time ($F_{2,236}$=3.46; P=.03; partial η^2=0.03). However, Bonferroni-adjusted pairwise comparisons did not reach statistical significance for the comparison from baseline to immediately after the course (adjusted P=.07) or from baseline to follow-up (adjusted P=.06). The immediately postcourse and follow-up scores did not differ significantly (adjusted P>.99). These findings are summarized in Table 3 and illustrated in Figure 4. Detailed Bonferroni-adjusted pairwise comparisons are provided in Table S3 in Multimedia Appendix 1.

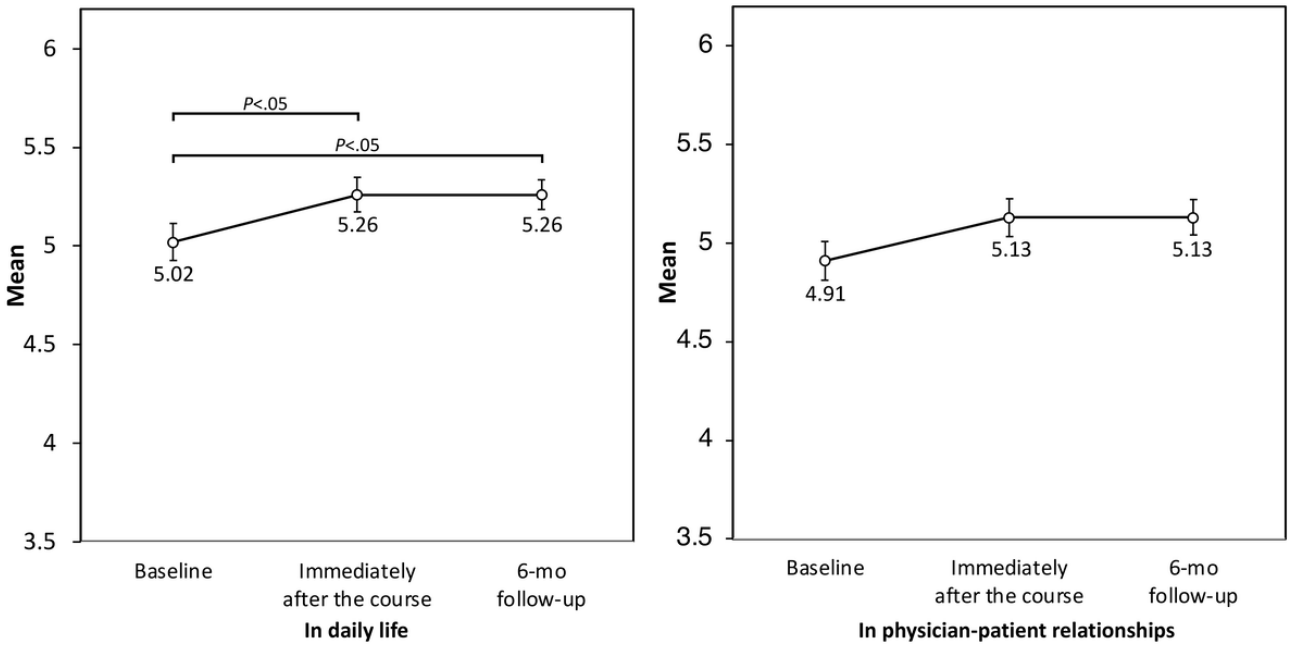
XSL•FO
RenderX

Table . Changes in confidence in applying empathy across time points[a].

	Baseline, mean (SD)	Immediately after the course, mean (SD)	6-mo follow-up, mean (SD)	*F* test (*df*)	*P* value	Partial η^2
Confidence in applying empathy in daily life	5.02 (1.04)	5.26 (0.97)	5.26 (0.82)	5.18 (2,236)	.006	0.04
Confidence in applying empathy in physician-patient relationships	4.91 (1.09)	5.13 (1.04)	5.13 (0.97)	3.46 (2,236)	.03	0.03

[a]Confidence scores range from 1 to 7, with higher scores indicating greater perceived confidence.

Figure 4. Changes in confidence in applying empathy in daily life and in physician-patient relationships across time points. Error bars indicate SEs.



Discussion

Principal Results

In this single-group longitudinal study, self-reported empathy scores increased during the course period and were maintained at 6 months, with similar patterns for emotional and cognitive empathy scores. Female students reported higher overall empathy scores than male students, but no significant time-by-gender interaction was observed, indicating similar longitudinal trajectories between male and female students. Confidence in applying empathy in daily life increased and remained above baseline at follow-up. Confidence in applying empathy in physician-patient relationships showed a significant overall time effect, although Bonferroni-adjusted pairwise comparisons were not statistically significant.

Comparison With Prior Work

The finding that self-reported empathy scores were higher after the course is consistent with prior work suggesting that empathy-related attitudes and communication skills can be supported through structured educational approaches [15,16,23-26]. A recent meta-analysis reported that empathy-focused interventions in medical students and physicians had moderate effects on both cognitive and affective empathy [27].

The findings of this study are also relevant to the literature on empathy trajectories during medical training. Several studies have reported decreases in empathy during medical school, especially during later clinical training [7,8,12,13]. However, evidence is not entirely consistent across settings. A longitudinal study in Singapore found no significant overall decline in empathy across medical education and suggested that curricular and sociocultural context may influence empathy trajectories [28]. In the Korean context, prior studies have focused mainly on empathy correlates, measurement, curricular attitudes, or clerkship-associated change rather than on longitudinal intervention studies in premedical students [18,19,29,30]. In this context, this study adds preliminary longitudinal evidence from a Korean premedical setting. That empathy scores were still elevated at 6 months is encouraging, although the modest attenuation after the immediately postcourse assessment suggests that a single short course may be insufficient to support durable long-term development without later reinforcement [17].

Because this study did not include a control group, the findings should be interpreted as longitudinal changes associated with

XSL•FO
RenderX

course participation rather than definitive evidence of course effectiveness. Although no formal communication- or empathy-focused instruction took place during follow-up, other unmeasured experiences may also have influenced the outcomes.

Emotional and Cognitive Empathy

Both emotional and cognitive empathy scores showed significant time effects and remained above baseline at the 6-month follow-up. Apparent differences between the 2 subscales should be interpreted cautiously. Although the raw mean increase was numerically larger for cognitive empathy, this should not be taken as evidence that the course had a stronger influence on cognitive perspective taking. The cognitive empathy subscale of the S-JSE-K contains only 2 items compared with 16 items in the emotional empathy subscale, and cognitive empathy scores also had a lower baseline mean and greater variability. Together, these features may have made the cognitive empathy score less precise and more sensitive to item-level variation.

The parallel increases in emotional and cognitive empathy may reflect the course's broader combination of conceptual learning and practice-based communication activities rather than any single instructional component [15,16,23,25-27]. Because the intervention was evaluated as a whole, this study cannot determine which specific components were most closely associated with changes in each empathy domain. Future studies should use more balanced multidimensional empathy measures, component-level process data, and observational outcomes to clarify how communication-focused education is associated with changes in emotional empathy, cognitive empathy, and empathic behavior.

Gender Differences

Female students reported higher empathy scores than male students across the 3 assessment points. This finding is consistent with those of previous studies reporting higher empathy scores among female medical students [31-35]. However, the absence of significant time-by-gender interactions indicates that the pattern of change over time was similar between male and female students. Thus, although female students had higher mean scores across assessments, the course-associated longitudinal trajectory of empathy scores did not appear to differ by gender.

Several explanations have been proposed for gender differences in self-reported empathy scores, including emotional socialization, responsiveness to interpersonal cues, communication expectations, and self-report tendencies [33,35,36]. These factors may be particularly relevant when empathy is measured using self-report instruments because responses may reflect not only empathic orientation but also gendered expectations about emotional expressiveness and interpersonal sensitivity [26]. In the Korean premedical context, these findings should be interpreted in light of broader sociocultural patterns in empathy and emotional expression [37,38]. Findings from Korean medical education studies have not been fully consistent, with at least one Korean study reporting an atypical gender pattern in empathy scores before clerkship [30]. Therefore, the higher mean scores among female students in this study should be interpreted as a descriptive group difference rather than evidence of fixed gender-based differences in empathic capacity. Future research should examine how gender, cultural context, self-report tendencies, and prior communication experiences interact to shape empathy score trajectories in Korean medical education.

Confidence in Applying Empathy

Confidence in applying empathy in daily life increased significantly from baseline to immediately after the course and remained significantly higher than baseline at follow-up. Confidence in applying empathy in physician-patient relationships also showed a significant overall time effect and remained numerically unchanged from immediately after the course to follow-up, although Bonferroni-adjusted pairwise comparisons from baseline were not statistically significant.

These findings suggest that students' perceived readiness to apply empathy may be more stable than their self-reported empathy scores over the 6-month period. Empathy scores reflect students' self-reported empathic orientation, whereas confidence items reflect perceived ability or readiness to use empathy in specific contexts. The slight decline in empathy scores at follow-up may reflect natural fluctuation or the challenge of sustaining empathy-related attitudes over time. In contrast, the relatively stable confidence scores may indicate that students retained a practical sense of how to apply empathic communication after the course.

However, confidence should not be equated with demonstrated empathic competence [39]. The confidence outcomes were measured using 2 items and were exploratory in nature. Future studies should examine whether confidence in applying empathy is associated with observed communication behaviors, standardized patient ratings, observer ratings, faculty assessments, or patient-reported empathy measures.

Implications for Medical Education

These findings suggest the potential educational value of introducing empathy and communication training during the premedical phase. Early exposure may help students acquire a shared language for patient-centered communication and begin linking empathy to observable communication behaviors. At the same time, the temporal pattern suggests that an introductory course should be viewed as a starting point rather than a complete solution. Most score changes occurred during the course itself, and the 6-month pattern was more consistent with short-term maintenance than with continued growth.

Although this pattern does not establish causality, it supports a longitudinal, spiral model in which early communication training is reinforced through later experiential and reflective opportunities [25,26,40,41]. Such integration may help students move beyond self-reported empathic orientation toward more stable empathic communication in authentic clinical settings.

Limitations

This study has several limitations. First, it was conducted at a single institution with Korean premedical students, which may limit generalizability to other institutions, countries, or stages of medical training. Second, the single-group design without a comparison group limits causal inference, and other unmeasured

XSL•FO
RenderX

experiences during follow-up may have influenced the outcomes. Third, all outcomes were based on self-report. Self-reported empathy and confidence may not directly correspond to empathic behavior in authentic clinical encounters [26,39]. Future studies should include behavioral assessments, standardized patient ratings, peer or faculty ratings, or patient-reported measures. Fourth, the cognitive empathy subscale of the S-JSE-K contains only 2 items, which may limit its sensitivity to detecting changes in cognitive empathy. Fifth, the 2 confidence items were single-item exploratory measures and were not validated as independent subscales. Finally, the follow-up period was limited to 6 months. Longer longitudinal studies are needed to examine whether course-associated changes persist into clinical training and whether they are reflected in observed physician-patient communication.

Conclusions

In this single-group longitudinal study, participation in a basic medical communication course was associated with higher self-reported empathy scores and greater confidence in applying empathy in daily life at 6 months among Korean premedical students. Controlled multisite studies using behavioral outcomes are needed to determine whether these early gains translate into sustained empathic communication during later clinical training.

Funding

The authors declared no financial support was received for this work.

Data Availability

The datasets generated and analyzed during this study are not publicly available due to participant confidentiality considerations but are available from the corresponding author on reasonable request.

Authors' Contributions

SJJ contributed to formal analysis and writing—original draft. HHY contributed to conceptualization, methodology, data curation, formal analysis, supervision, writing—original draft, and writing—review and editing. Both authors reviewed and approved the final manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Bonferroni-adjusted pairwise comparisons for self-reported empathy and confidence outcomes and gender-stratified descriptive statistics for self-reported empathy.
[PDF File, 145 KB - mededu_v12i1e92215_app1.pdf]

References

1. Coplan A, Goldie P. Empathy: Philosophical and Psychological Perspectives: Oxford University Press; 2011.
2. Hojat M, Louis DZ, Markham FW, Wender R, Rabinowitz C, Gonnella JS. Physicians' empathy and clinical outcomes for diabetic patients. Acad Med 2011 Mar;86(3):359-364. [doi: 10.1097/ACM.0b013e3182086fe1] [Medline: 21248604]
3. Kim SS, Kaplowitz S, Johnston MV. The effects of physician empathy on patient satisfaction and compliance. Eval Health Prof 2004 Sep;27(3):237-251. [doi: 10.1177/0163278704267037] [Medline: 15312283]
4. Leahy M, Cullen W, Bury G. "What makes a good doctor?" A cross sectional survey of public opinion. Ir Med J 2003 Feb;96(2):38-41. [Medline: 12674150]
5. Hemmerdinger JM, Stoddart SD, Lilford RJ. A systematic review of tests of empathy in medicine. BMC Med Educ 2007 Jul 25;7:24. [doi: 10.1186/1472-6920-7-24] [Medline: 17651477]
6. Chen D, Lew R, Hershman W, Orlander J. A cross-sectional measurement of medical student empathy. J Gen Intern Med 2007 Oct;22(10):1434-1438. [doi: 10.1007/s11606-007-0298-x] [Medline: 17653807]
7. Neumann M, Edelhäuser F, Tauschel D, et al. Empathy decline and its reasons: a systematic review of studies with medical students and residents. Acad Med 2011 Aug;86(8):996-1009. [doi: 10.1097/ACM.0b013e318221e615] [Medline: 21670661]
8. Hojat M, Vergare MJ, Maxwell K, et al. The devil is in the third year: a longitudinal study of erosion of empathy in medical school. Acad Med 2009 Sep;84(9):1182-1191. [doi: 10.1097/ACM.0b013e3181b17e55] [Medline: 19707055]
9. Hojat M, Gonnella JS, Nasca TJ, Mangione S, Vergare M, Magee M. Physician empathy: definition, components, measurement, and relationship to gender and specialty. Am J Psychiatry 2002 Sep;159(9):1563-1569. [doi: 10.1176/appi.ajp.159.9.1563] [Medline: 12202278]
10. Shankar PR, Piryani RM. Changes in empathy among first year medical students before and after a medical humanities module. Educ Med J 2013;5(1):e35-e42. [doi: 10.5959/eimj.v5i1.36]

XSL•FO
RenderX

11. Hojat M, Mangione S, Nasca TJ, Gonnella JS, Magee M. Empathy scores in medical school and ratings of empathic behavior in residency training 3 years later. J Soc Psychol 2005 Dec;145(6):663-672. [doi: 10.3200/SOCP.145.6.663-672] [Medline: 16334513]
12. Newton BW, Barber L, Clardy J, Cleveland E, O'Sullivan P. Is there hardening of the heart during medical school? Acad Med 2008 Mar;83(3):244-249. [doi: 10.1097/ACM.0b013e3181637837] [Medline: 18316868]
13. Hojat M, Mangione S, Nasca TJ, et al. An empirical study of decline in empathy in medical school. Med Educ 2004 Sep;38(9):934-941. [doi: 10.1111/j.1365-2929.2004.01911.x] [Medline: 15327674]
14. Batt-Rawden SA, Chisolm MS, Anton B, Flickinger TE. Teaching empathy to medical students: an updated, systematic review. Acad Med 2013 Aug;88(8):1171-1177. [doi: 10.1097/ACM.0b013e318299f3e3] [Medline: 23807099]
15. Kataoka H, Iwase T, Ogawa H, et al. Can communication skills training improve empathy? A six-year longitudinal study of medical students in Japan. Med Teach 2019 Feb;41(2):195-200. [doi: 10.1080/0142159X.2018.1460657] [Medline: 29683011]
16. Wündrich M, Schwartz C, Feige B, Lemper D, Nissen C, Voderholzer U. Empathy training in medical students - a randomized controlled trial. Med Teach 2017 Oct;39(10):1096-1098. [doi: 10.1080/0142159X.2017.1355451] [Medline: 28749198]
17. Baugh RF, Hoogland MA, Baugh AD. The long-term effectiveness of empathic interventions in medical education: a systematic review. Adv Med Educ Pract 2020;11:879-890. [doi: 10.2147/AMEP.S259718] [Medline: 33244286]
18. Lee BK, Bahn GH, Lee WH, Park JH, Yoon TY, Baek SB. The relationship between empathy and medical education system, grades, and personality in medical college students and medical school students. Korean J Med Educ 2009 Jun;21(2):117-124. [doi: 10.3946/kjme.2009.21.2.117] [Medline: 25813109]
19. Kim IS, Kim YH, Chang HI. An investigation on reliability and validity of the Korean Medical Students' Empathy Rating Scale. Korean J Psychopathol 2004;13(1-2):9-15 [FREE Full text]
20. Rosenberg MB, Chopra D. Nonviolent Communication: A Language of Life: Life-Changing Tools for Healthy Relationships, 3rd edition: PuddleDancer Press; 2015.
21. Hojat M, Mangione S, Nasca TJ, et al. The Jefferson Scale of Physician Empathy: development and preliminary psychometric data. Educ Psychol Meas 2001 Apr;61(2):349-365. [doi: 10.1177/00131640121971158]
22. Kim IS. An investigation on the reliability and validity of the Korean Medical Students' Empathy Rating Scale [master's thesis]. : Kyung Hee University; 2005 URL: https://www.riss.kr/link?id=T10091164 [accessed 2026-05-21]
23. Winefield HR, Chur-Hansen A. Evaluating the outcome of communication skill teaching for entry-level medical students: does knowledge of empathy increase? Med Educ 2000 Feb;34(2):90-94. [doi: 10.1046/j.1365-2923.2000.00463.x] [Medline: 10652060]
24. Spiro H. What is empathy and can it be taught? Ann Intern Med 1992 May 15;116(10):843-846. [doi: 10.7326/0003-4819-116-10-843] [Medline: 1482433]
25. Fragkos KC, Crampton PE. The effectiveness of teaching clinical empathy to medical students: a systematic review and meta-analysis of randomized controlled trials. Acad Med 2020 Jun;95(6):947-957. [doi: 10.1097/ACM.0000000000003058] [Medline: 31688037]
26. Kelm Z, Womer J, Walter JK, Feudtner C. Interventions to cultivate physician empathy: a systematic review. BMC Med Educ 2014 Oct 14;14:219. [doi: 10.1186/1472-6920-14-219] [Medline: 25315848]
27. Ngo HL, Sokolovic N, Jenkins JM. Teaching cognitive and affective empathy in medicine: a systematic review and meta-analysis of randomized controlled trials. Med Educ Online 2025 Dec;30(1):2501263. [doi: 10.1080/10872981.2025.2501263] [Medline: 40329527]
28. Samarasekera DD, Lee SS, Yeo SP, Ponnamperuma G. Development of student empathy during medical education: changes and the influence of context and training. Korean J Med Educ 2022 Mar;34(1):17-26. [doi: 10.3946/kjme.2022.216] [Medline: 35255613]
29. Lee I, Kim C, Lee H, Lee W, Jeong S. A study on attitude of medical students to medical communication and subject. J Pract Res Advert Public 2015 Aug;8(3):105-124. [doi: 10.21331/jprapr.2015.8.3.005]
30. Yeo S. Differences and changes in the empathy of Korean medical students according to gender and vocational aptitude, before and after clerkship. Korean J Med Educ 2019 Dec;31(4):343-355. [doi: 10.3946/kjme.2019.144] [Medline: 31813200]
31. Chatterjee A, Ravikumar R, Singh S, Chauhan PS, Goel M. Clinical empathy in medical students in India measured using the Jefferson Scale of Empathy-Student Version. J Educ Eval Health Prof 2017;14:33. [doi: 10.3352/jeehp.2017.14.33] [Medline: 29278905]
32. Duarte MI, Branco MC, Raposo ML, Rodrigues PJ. Empathy in medical students as related to gender, year of medical school and specialty interest. South-East Asian J Med Educ 2015;9(1):50-53. [doi: 10.4038/seajme.v9i1.97]
33. Elkin B, LaPlant EM, Olson AP, Violato C. Stability and differences in empathy between men and women medical students: a panel design study. Med Sci Educ 2021;31(6):1851-1858. [doi: 10.1007/s40670-021-01373-0] [Medline: 34956701]
34. Yu J, Lee S, Kim M, Lim K, Chang K, Lee M. Relationships between perspective-taking, empathic concern, and self-rating of empathy as a physician among medical students. Acad Psychiatry 2020 Jun;44(3):316-319. [doi: 10.1007/s40596-019-01114-x] [Medline: 31873921]
35. Hojat M, Gonnella JS, Mangione S, et al. Empathy in medical students as related to academic performance, clinical competence and gender. Med Educ 2002 Jun;36(6):522-527. [doi: 10.1046/j.1365-2923.2002.01234.x] [Medline: 12047665]

XSL•FO
RenderX

36. Roter DL, Hall JA, Aoki Y. Physician gender effects in medical communication: a meta-analytic review. JAMA 2002 Aug 14;288(6):756-764. [doi: 10.1001/jama.288.6.756] [Medline: 12169083]
37. Yang H, Kang SJ. Exploring the Korean adolescent empathy using the Interpersonal Reactivity Index (IRI). Asia Pacific Educ Rev 2020 Jun;21:339-349. [doi: 10.1007/s12564-019-09621-0]
38. Sohn S, Park MS. Understanding gender differences in Korean emotional expressions for increased cultural competency of social workers. Int Soc Work 2016;59(6):850-860. [doi: 10.1177/0020872815618735]
39. Hepner A, Cerutti B, Lüchinger R, Junod Perron N. Measuring empathy in OSCEs: a comparison of student, standardized patient, and observer rating tools. PEC Innov 2025;8:100452. [doi: 10.1016/j.pecinn.2025.100452] [Medline: 41567738]
40. Ward A, Gilligan C, Bennett-Weston A, Howick J. The development, delivery, and evaluation of novel longitudinal empathy-focused communication skills training at a UK medical school. PEC Innov 2025;7:100436. [doi: 10.1016/j.pecinn.2025.100436] [Medline: 41098203]
41. Howick J, Slavin D, Carr S, et al. Towards an empathic hidden curriculum in medical school: a roadmap. J Eval Clin Pract 2024 Jun;30(4):525-532. [doi: 10.1111/jep.13966] [Medline: 38332641]

Abbreviations

NVC: nonviolent communication
S-JSE-K: Korean student version of the Jefferson Scale of Empathy

Edited by T Gladman; submitted 26.Jan.2026; peer-reviewed by B Rubinstein, M Witry; revised version received 27.Apr.2026; accepted 03.May.2026; published 26.May.2026.

Please cite as:
Jeon SJ, Yoo HH
Changes in Self-Reported Empathy After a Basic Medical Communication Course Among Premedical Students: Single-Group Longitudinal Study
JMIR Med Educ 2026;12:e92215
URL: https://mededu.jmir.org/2026/1/e92215
doi:10.2196/92215

© Seong Ju Jeon, Hyo Hyun Yoo. Originally published in JMIR Medical Education (https://mededu.jmir.org), 26.May.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Education, is properly cited. The complete bibliographic information, a link to the original publication on https://mededu.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Impact of Community-Oriented Medical Education on Medical Students' Perceptions of Community Health Care: Qualitative Study

Kiyoshi Shikino[1], MD, MHPE, PhD; Kazuyo Yamauchi[1,2], MD, MHPE, PhD; Nobuyuki Araki[1], MD, PhD; Naoto Ozaki[1], MD; Yu Kamata[1], MD; Shinya Aoki[1,3], MD; Yota Katsuyama[1,4], MD, PhD; Daichi Sogai[1,5], MD; Mai Miyamoto[1,6], MD; Kensuke Yoshimura[7], MD, MPH, PhD; Takeshi Oki[8], MD, PhD; Shoichi Ito[1,9], MD, PhD

[1]Department of Community-Oriented Medical Education, Graduate School of Medicine, Chiba University, Chiba, Japan

[2]Department of Medical Education,, Dokkyo Medical University, Tochigi, Japan

[3]Department of General Medicine, Shioda Hospital, Katsuura, Japan

[4]Sanbunomori Clinic, Sanmu, Japan

[5]Department of General Medicine, Sanmu Medical Center, Sanmu, Japan

[6]Department of Internal Medicine, Isumi Medical Center, Isumi, Japan

[7]Center for Next Generation of Community Health, Chiba University Hospital, Chiba, Japan

[8]Trinity Neurology Clinic, Sakura, Japan

[9]Department of Medical Education, Graduate School of Medicine, Chiba University, Chiba, Japan

Corresponding Author:
Kiyoshi Shikino, MD, MHPE, PhD
Department of Community-Oriented Medical Education
Graduate School of Medicine
Chiba University
1-8-1, Inohana
Chu-ou-ku
Chiba, 2608677
Japan
Phone: 81 43 222 7171
Email: kshikino@gmail.com

Abstract

Background: Physician maldistribution remains a global challenge, with Japan's rural regions facing critical health care shortages. Regional quota programs aim to attract medical students to underserved areas; however, their effectiveness in fostering long-term commitment is uncertain. Community-oriented medical education (COME) programs aim to address this issue by developing students' understanding and dedication to rural health care.

Objective: This study investigated the impact of an enhanced COME program, featuring increased early clinical exposure and faculty development, on first-year regional quota medical students' perception of community health care at Chiba University.

Methods: We conducted a cross-sectional qualitative study comparing 2 cohorts, 20 students enrolled from the existing COME course (April-December 2021) and 20 from the revised course (April-December 2022). The revised course included an additional day of community-based clinical exposure supervised by COME-trained attending physicians. Students' written reflections were analyzed using qualitative content analysis and categorized according to the Fink Taxonomy of significant learning, comprising 6 domains, including foundational knowledge, application, integration, human dimension, caring, and learning how to learn. Reflections were synthesized into higher-order themes crosswalked to the Fink domains.

Results: Demographics were similar between the 2021 and 2022 cohorts. In 2021, 311 learning codes were identified across foundational knowledge (n=128), application (n=91), integration (n=40), human dimension (n=16), caring (n=30), and learning how to learn (n=6). In 2022, codes increased to 385, with notable growth in caring (n=58) and human dimension (n=57), alongside increases in learning how to learn (n=15) and integration (n=45). Theme-based synthesis identified four overarching themes: (1) community health care as an interconnected, resource-constrained system; (2) patient-centered relationships and trust through communication and teamwork; (3) emerging professional identity and responsibility toward community service; and (4) developing

XSL•FO
RenderX

a self-directed learning orientation for community practice. Qualitative analysis revealed that students gained a deeper understanding of patient-centered care, interprofessional collaboration, and social challenges in rural health care. The consistency in the foundational knowledge domain underscored a stable conceptual foundation, while the increase in affective and reflective domains reflected greater emphasis on interpersonal, value-oriented, and reflective learning in the revised cohort.

Conclusions: Enhancements of the COME program, including additional early clinical exposure and faculty development, were associated with improved students' perceptions of community health care. The increased focus on the caring and human dimension domains underscores the role of practical experiences in fostering collaboration, communication, and patient-centered care. The theme-based synthesis further suggests that the revised program prompted more frequent reflections on professional identity formation and self-directed learning while maintaining a stable foundation of community health care concepts. Mentorship by community hospital attendings, alongside structured clinical exposure, appears crucial in shaping medical students' understanding and commitment to rural medicine. Ongoing longitudinal evaluations are warranted to assess the sustained impact of COME programs on career trajectories in underserved areas.

(JMIR Med Educ 2026;12:e84406) doi:10.2196/84406

KEYWORDS

community health care; community-based medical education; community-oriented medical education; physician maldistribution; career perceptions; community hospital; community education; faculty development

Introduction

Physician maldistribution is a global concern affecting many nations [1-6]. In Japan, rural regions are facing a critical shortage of health care providers [7,8] due to urbanization [9-11]. To address this issue, numerous Japanese medical schools have introduced regional quota programs designed to motivate students to practice community or rural medicine [7,8,12]. These programs offer scholarships tied to service agreements in designated regions to address the scarcity of physicians. As of 2015, 70 out of 79 Japanese medical schools offered regional quota programs, enrolling 1541 students [7].

However, the effectiveness and ethical implications of these programs are debatable. Students assigned to regional quotas may become less inclined to work in their designated regions as they progress through their academic programs [8]. Therefore, it is essential to develop community-oriented medical education (COME) courses that not only impart knowledge but also cultivate students' commitment to serving in underserved communities [13].

The Japanese government has high expectations for community-based medical education (CBME) and has made it a compulsory component of medical training [13-15]. Through direct participation in CBME, medical students and resident physicians can acquire practical skills, knowledge, and attitudes pertinent to clinical practice [16]. Nonetheless, questions remain regarding the long-term sustainability and impact of these programs, including student retention in rural areas [14]. Thus, continuous evaluation and enhancement are critical to effectively mitigate health care provider shortages in rural areas in Japan [8].

Herein, we aimed to investigate the educational impact of an enhanced COME program, which incorporated increased early clinical exposure and faculty development for community hospital attendings, on first-year regional quota medical students at Chiba University. Specifically, we explored students' "career perceptions," defined as their developing understanding of professional identity, motivation toward community service, and long-term orientation toward rural medical practice. We hypothesized that participation in the enhanced COME program would strengthen students' sense of professional responsibility and commitment to working in underserved areas by deepening their reflection on real-world community health care experiences. In this study, the term "career perceptions" refers to students' evolving understanding of their future professional roles, motivations, and attitudes toward community health care. These perceptions encompass not only students' intentions to work in rural or underserved settings but also their sense of social responsibility, empathy, and professional identity as future physicians.

Methods

Study Design

A cross-sectional survey was conducted to assess the influence of a COME program, designed by the Chiba University School of Medicine, on medical students' career perceptions. In this study, "career perceptions" were defined as students' reflective understanding of their professional identity, motivation toward community service, and long-term orientation toward rural medical practice. The survey was designed to collect qualitative data from students' written reflections rather than quantitative responses, allowing for an in-depth exploration of their perceptions and learning experiences. Qualitative research typically involves data in the form of words instead of numbers and is particularly appropriate for understanding participants' experiences, thoughts, and motivations [17].

This study encompassed 2 cohorts, including 20 first-year regional quota medical students from April to December 2021, representing the existing course structure, and 20 students from April to December 2022, representing a revised course with additional early clinical exposure. Before participating in the COME program, the students had already completed the interprofessional education (IPE) step 1 course earlier in the academic year, which may have influenced their perceptions and learning outcomes in the COME program.

XSL•FO
RenderX

Participants and Settings

Participants were first-year medical students enrolled in Chiba University under a regional quota system, participating in a year-long COME program. All participants were Japanese medical students enrolled in the regional quota program at Chiba University. The 2021 (n=20) and 2022 cohorts (n=20) each consisted of first-year students aged 18-24 years and 18-21 years, respectively. Sex assigned at birth (male or female) was collected as a demographic variable. This program included didactic sessions and practical experiences designed to enhance students' understanding of and commitment to community health care.

Selection Process for Regional Quota Students

Students applying to Chiba University School of Medicine must choose between applying under the general admission track or the regional quota track. Admission decisions are based on the scores from the common test for university admissions, a written examination, and an interview. Students admitted under the regional quota track are obligated to work within Chiba Prefecture for 9 years after graduation, fulfilling a service agreement. In return, they receive financial support through educational subsidies as an incentive. This system is intended to attract and retain health care professionals in underserved areas of the prefecture.

COME at Chiba University

Chiba Prefecture is one of the regions in Japan with a significant disparity in the geographical distribution of physicians, creating an urgent need to ensure the retention of medical professionals in community health care settings. In response, the COME program was introduced to address this issue through targeted human resource development. The program aims to cultivate medical students' understanding of and commitment to practicing in underserved areas, thereby contributing to sustainable health care solutions in the region.

Facilities and Roles of Community Hospital Attending

In 2022, the COME program was implemented across 4 facilities, each staffed with attending physicians dedicated to community-based medical education. These community hospital attendings play a critical role in educating medical students and creating an optimal learning environment. Additionally, they participate in weekly collaborative faculty development sessions at Chiba University, integrating practical experience in community health care with theoretical frameworks, ensuring a comprehensive approach to medical education in underserved areas.

IPE Step 1 Course

Before participating in the COME program, all students had completed the IPE Step 1 course, which establishes the necessary professional attitudes for health care practitioners and develops their ability to communicate effectively with patients, service users, and students from other faculties. At the end of the course, students are expected to reflect on the requirements for their professional growth, contribute effectively to team efforts, and respect interprofessional collaboration. These competencies likely influenced students' perceptions and learning outcomes in the subsequent COME program.

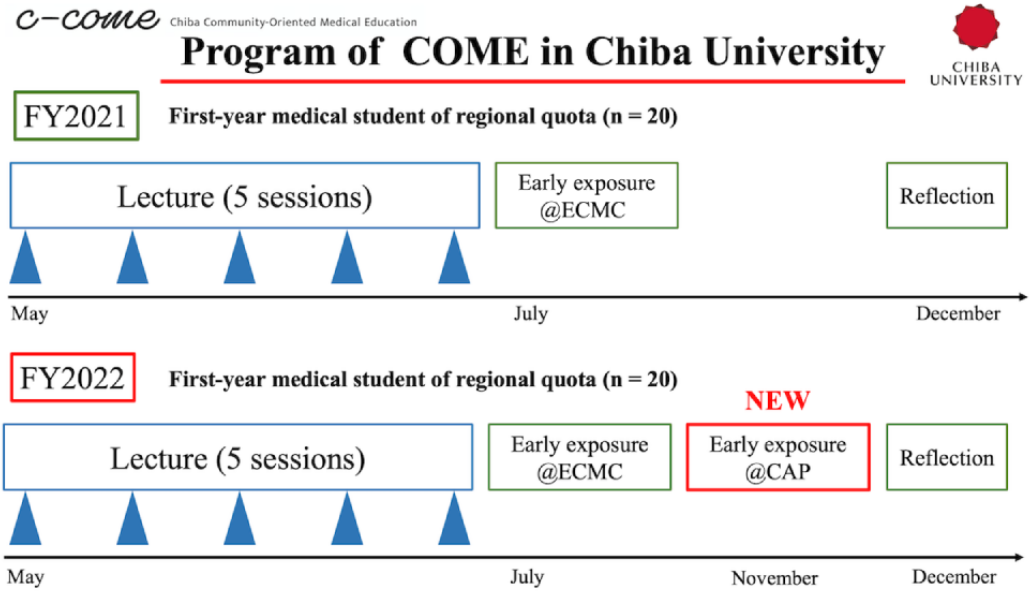
Enhancements to the COME Program

In 2021, the COME program consisted of five lecture sessions, one day of early clinical exposure, and subsequent reflection sessions (Figure 1). During this day, students mainly observed outpatient and inpatient activities at Eastern Chiba Medical Center, attended multidisciplinary team meetings, and learned about hospital functions and community health care delivery. Their participation was primarily observational. In 2022, the program was enhanced with an additional day of clinical exposure, orchestrated by COME-trained attending physicians serving as community preceptors. This additional day was implemented at community hospitals where students shadowed COME-trained attendings, interacted directly with patients, and discussed diagnostic reasoning and care planning. Students also engaged in reflective discussions with their mentors, receiving formative feedback on communication and professionalism. Early clinical exposure allowed students to learn directly from physicians who served as supervisors and educational leaders. Before clinical exposure, attending physicians engaged in targeted mentorship through personalized communication via email or online conferences to understand each student's individual needs and learning objectives. The faculty development program, consisting of 40 sessions annually, equipped these physicians with pedagogical skills to provide personalized guidance throughout the medical education process. These attending physicians, strategically positioned in regions with physician shortages, underwent comprehensive faculty development. This development entailed 40 sessions annually, each lasting 2 hours, and covered various pedagogical topics, including educational technology, simulation teaching, and assessment methodologies. These sessions were part of a broader faculty development program to enhance the physicians' overall educational capabilities and equip them to provide personalized guidance throughout the students' medical education. Overall, the 2022 revision represented a shift from passive observation to active engagement and mentorship-based learning, which deepened students' understanding of community health care and enhanced their sense of professional identity. These enhancements aimed to provide students with practical experiences and mentorship that deepen their understanding of community health care and address the challenges faced in underserved areas [18].

In Figure 1, it can be found that, in 2021, the program comprised 5 lecture sessions, one early clinical exposure at Eastern Chiba Medical Center (ECMC), and reflection activities. In 2022, the revised program added a new early clinical exposure component at hospitals staffed by COME-trained attending physicians (CAP), integrating faculty development initiatives to enhance community-based mentorship.

XSL•FO
RenderX

Figure 1. Structure and timeline of the community-oriented medical education (COME) program at Chiba University for first-year regional quota medical students in 2021 and 2022. CAP: community-oriented medical education–trained attending physicians; ECMC: Eastern Chiba Medical Center; FY: final year.



Data Collection

Students' Reports

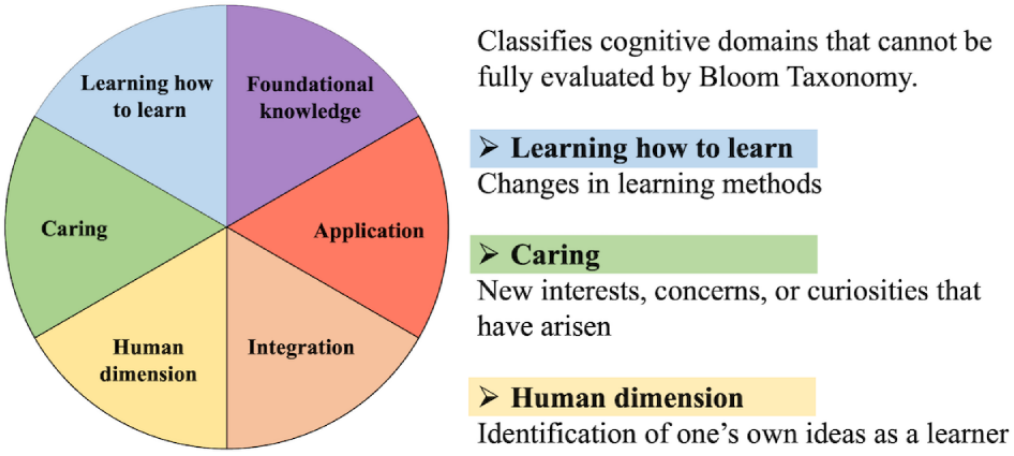
A qualitative inquiry was conducted. To measure the educational impact, we analyzed students' reports from 2021-2022 as qualitative data. These reports addressed 4 key questions regarding students' understanding of community medicine, health care challenges, their envisioned role in local health care, and the competencies required for these roles (Multimedia Appendix 1).

The Fink Taxonomy of Significant Learning

The Fink Taxonomy [19] is distinct in its categorization of cognitive domains, offering a broader scope than the Bloom Taxonomy [12,20-22]. It emphasizes the "Human dimension"–students' self-awareness as learners; "Caring" the development of new interests and concerns; and "Learning how to learn" the adaptation and improvement of students' learning approaches (Figure 2 and Multimedia Appendix 2) [19].

In Figure 2, the taxonomy classifies learning outcomes into 6 interrelated domains, foundational knowledge, application, integration, human dimension, caring, and learning how to learn. This framework guided the categorization of students' written reflections to evaluate changes in cognitive, affective, and interpersonal learning domains beyond what is captured by Bloom Taxonomy.

Figure 2. Framework of the Fink Taxonomy of significant learning used for qualitative content analysis.



XSL•FO
RenderX

Data Analysis

Content Analysis

Content analysis was conducted using the Fink Taxonomy of significant learning to evaluate students' perceptions of community health care. This taxonomy categorizes cognitive process dimensions into various categories and subcategories [23,24]. KS independently read and coded all transcripts. Subsequently, the descriptors were discussed, identified, and agreed upon by KS, K Yamauchi, and NA. Similar codes were then grouped into categories and subcategories as they emerged from the data. These categories and subcategories were regularly reviewed and discussed by authors experienced in qualitative research to ensure the credibility of the findings [25]. Data saturation was assessed through an iterative process of constant comparison. After each round of coding, the research team reviewed whether new concepts or categories emerged. When no additional themes or codes were identified across 3 consecutive transcripts, we considered both thematic and code saturation to have been achieved [26,27]. The authors discussed and reached a consensus on all further changes, achieving code saturation. We adopted an inductive and deductive content analysis approach to balance theory-driven and data-driven coding. Initially, inductive coding allowed themes to emerge naturally from the students' reflections, ensuring that novel insights were not constrained by pre-existing frameworks. Subsequently, deductive categorization was applied using the Fink Taxonomy to systematically organize the emergent themes into recognized dimensions of significant learning. This combined approach enhanced the analytical rigor and minimized the risk of forcing data into preset categories [22].

Concepts for each cognitive process dimension in the Fink Taxonomy [20] were analyzed, and the number of units of analysis for each concept was counted. The researchers then grouped similar codes into themes and checked to see which dimension of the cognitive process they corresponded to.

AI-Assisted Analytic Support (Generative AI)

Additionally, to enhance the rigor and transparency of our qualitative analysis, we complemented the human coding process with a generative artificial intelligence (AI) tool (ChatGPT 5; OpenAI) [25,26]. The AI tool was prompted to identify potential thematic clusters and semantic relationships within anonymized textual data, following best practices for AI-assisted qualitative research as outlined in recent literature [28-31]. Specifically, the AI was used to (1) propose candidate thematic clusters and relationships across codes/excerpts, (2) suggest alternative categorizations and labels, and (3) identify potential negative cases or boundary instances that did not fit preliminary themes. Only deidentified materials (eg, anonymized excerpts and/or code lists) were entered into the tool; no personal identifiers were provided. The AI outputs were critically reviewed, refined, and confirmed through team discussion to prevent model-induced bias. AI outputs were treated as hypotheses for discussion rather than definitive findings, and the research team documented whether each suggestion was adopted, refined, or rejected after checking against the original data. This triangulated approach improved coding consistency while ensuring that all final interpretations and coding decisions were made exclusively by human researchers [29,31,32]. Although formal interrater reliability statistics were not calculated, coding credibility was ensured through iterative peer review, team consensus discussions, and triangulation using AI-assisted content verification. Because we did not systematically archive full prompt-response transcripts during the earlier analytic iterations, we provide representative example prompts (and abbreviated outputs) in Multimedia Appendix 3 to enhance transparency and reproducibility.

Reflexivity

We conducted a qualitative study using content analysis to determine the effectiveness of the newly developed program in 2022 on students' perceptions of community health care. In qualitative research, researchers' backgrounds can significantly influence data interpretation and the construction of findings and conclusions, necessitating careful reflexivity. Reflexivity was used to account for researchers' potential biases.

In this study, all authors were clinical educators with experience in COME, providing the necessary expertise to understand and interpret the data accurately. The authors were trained in educational theory (KS, K Yamauchi, and SI) and primary care (SA, Y Kamata, DS, MM, and TO), enabling them to approach the analysis from a well-rounded perspective. This dual expertise ensured a comprehensive understanding of the educational and clinical aspects of community-oriented medical education programs. This reflective approach aimed to ensure the credibility and reliability of our findings and provide valuable insights into the factors contributing to successful training programs.

Ethical Considerations

This study was approved by the Ethics Review Committee of the Graduate School of Medicine, Chiba University (Approval No. 3425, approved on October 4, 2019). All work involving research on human participants complied with the Declaration of Helsinki. The procedures for obtaining informed consent were explained to the medical students, who were also informed that this study would not affect their grades. The researchers explained to the participants and obtained their informed and voluntary consent. No financial or material compensation was provided to participants for their involvement in this study.

Results

Content analysis of the reports from first-year regional quota medical students revealed distinct patterns across the academic years 2021 and 2022, categorized according to the Fink Taxonomy of significant learning (Tables 1 and 2). The demographic characteristics of participants were similar between the 2021 and 2022 cohorts. The median age was 19 years in both groups (2021: 18-24 years; 2022: 18-21 years), and the distribution of sex assigned at birth did not differ significantly (2021: 75% male, 25% female; 2022: 70% male, 30% female; P>.99).

XSL•FO
RenderX

Table 1. Results of qualitative content analysis of first-year regional quota medical students' reflection reports in the 2021 community-oriented medical education (COME) program at Chiba University (N=20), crosswalked to the Fink Taxonomy of significant learning.

Theme, key synthesis, and crosswalk to the Fink domains	Subcategories (code counts)
Theme 1: community health care as an interconnected, resource-constrained system	
Community health care is understood as a local system shaped by needs, constraints, and intersector collaboration.	
Primary domains: • Foundational knowledge	• Specific initiatives (n=48), challenges/current state (n=38), definition/role (n=26), aging (n=16)
Secondary domains: • Integration	• Challenges/responses (n=13), improvement of knowledge and skills (n=11), communication/cooperation (n=9), collaboration with community (n=7)
Theme 2: patient-centered relationships and trust through communication and teamwork	
Communication is positioned as the practical foundation of patient-centered care and trust, supported by collaboration.	
Primary domains: • Application	• Communication and trust building (n=15)
Secondary domains: • Human dimension • Caring	• Patient-centered health care and communication (n=6) • Patient-centered health care (n=7), communication and collaboration (n=5)
Theme 3: emerging professional identity and responsibility toward community service	
Students begin to articulate who they should become as physicians and how they should contribute to the community.	
Primary domains: • Human dimension • Caring	• Patient-centered health care and communication (n=6), understanding and importance of community health care (n=5), general practice and broad knowledge (n=3), realization of the community health care field (n=2) • Contribution and involvement in community health care (n=9), personal growth and learning (n=8), patient-centered health care (n=7), communication and collaboration (n=5), personal motivation and awareness (n=1)
Secondary domains: • Application	• Improve knowledge/skills (n=20), practice/problem-solving (n=18)
Theme 4: developing a self-directed learning orientation for community practice	
Students emphasize experiential learning and continuous improvement as preparation for community practice.	
Primary domains: • Learning how to learn	• Practical experience and activity participation (n=2), improvement of skills and knowledge (n=2), improvement of communication skills (n=1), continuous learning and understanding of community (n=1)
Secondary domains: • Application • Foundational knowledge	• Improvement of knowledge and skills (n=20), practice and problem-solving in community health care (n=18), communication and trust building (n=15), necessity and outlook of community health care (n=13), improvement of prevention and health awareness (n=9), use of information and communication technology (n=7), environmental preparation and collaboration in community health care (n=5), general practitioners and multidisciplinary knowledge (n=4) • Specific initiatives in community health care (n=48), challenges and current state of community health care (n=38), definition and role of community health care (n=26), aging population and community health care (n=16)

XSL•FO
RenderX

Table 2. Results of qualitative content analysis of first-year regional quota medical students' reflection reports in the revised 2022 community-oriented medical education (COME) program at Chiba University (n=20), crosswalked to the Fink Taxonomy of significant learning.

Theme, key synthesis, and crosswalk to the Fink domains	Subcategories (code counts)
Theme 1: community health care as an interconnected, resource-constrained system	
Community health care is understood as a local system shaped by needs, constraints, and intersector collaboration.	
Primary domains: • Foundational knowledge	• Quality and supply of health care (n=51), shortage of physicians and medical resources (n=41), importance of communication and consultation (n=23), aging population and the community health care (n=13)
Secondary domain: • Integration	• Medical resources and collaboration (n=15), sustainability and technology in health care (n=12), patient-centered care method (n=10)
Theme 2: patient-centered relationships and trust through communication and teamwork	
Communication is positioned as the practical foundation of patient-centered care and trust, supported by collaboration.	
Primary domains: • Human dimension • Application	• Patient-centered health care and communication (n=26) • Communication and trust building (n=15)
Secondary domain: • Caring	• Communication and collaboration (n=15), patient-centered health care (n=11)
Theme 3: emerging professional identity and responsibility toward community service	
Students begin to articulate who they should become as physicians and how they should contribute to the community.	
Primary domains: • Human dimension • Caring	• Challenges and improvements in community health care (n=21), importance and current state of community health care (n=10) • Contribution and involvement in community health care (n=12), personal motivation and awareness (n=10), personal growth and learning (n=10)
Secondary domains: • Integration	• Ethics and communication (n=8)
Theme 4: developing a self-directed learning orientation for community practice	
Students emphasize experiential learning and continuous improvement as preparation for community practice.	
Primary domains: • Learning how to learn	• Acquisition of broad knowledge (n=5), activities to broaden knowledge (n=4), interest and training (n=3), mandatory community health care training (n=2), improvement of communication skills (n=1)
Secondary domains: • Application	• Improvement of medical knowledge and skills (n=20), practice and problem-solving in community health care (n=20), community healthcare policy and support (n=8), executing ability and motivation (n=8), importance of community health care (n=6), use of information and communication technology (n=5)

Across both cohorts, we identified four higher-order themes that synthesize students' reflections and crosswalk to multiple domains of the Fink Taxonomy: (1) community health care as an interconnected, resource-constrained system; (2) patient-centered relationships and trust through communication and teamwork; (3) emerging professional identity and responsibility toward community service; and (4) developing a self-directed learning orientation for community practice (Tables 1 and 2; representative quotes are provided in Multimedia Appendices 4 and 5). While foundational understanding of community health care was evident in both years, the 2022 cohort more frequently expressed reflections aligned with affective and interpersonal learning (caring and human dimension) and reflective learning (learning how to learn), consistent with a more integrated pattern of "significant learning" in the Fink framework.

In Table 1, reports were analyzed using the Fink Taxonomy of significant learning, which categorizes learning into six domains. Subcategories and code counts are presented to support the theme-based synthesis and the crosswalk to the Fink domains.

XSL•FO
RenderX

Representative quotes supporting each theme and subcategory are provided in Multimedia Appendix 4.

In Table 2, reports were analyzed using the Fink Taxonomy of significant learning, which categorizes learning into 6 domains. Subcategories and code counts are presented to support the theme-based synthesis and the crosswalk to the Fink domains. Representative quotes supporting each theme and subcategory are provided in Multimedia Appendix 5.

For the academic year 2021, 311 codes were identified. In thematic terms, the 2021 reflections predominantly emphasized building a baseline understanding of community health care systems and challenges (theme 1; foundational knowledge and integration), accompanied by early attention to communication as a core clinical practice (theme 2; application), and emerging awareness of professional roles and values (theme 3; human dimension and caring). In the learning how to learn category, 6 codes were noted, including practical experience and participation in activities (n=2), improvement of skills and knowledge (n=2), improvement of communication skills (n=1), and continuous learning and understanding of the community (n=1). Students frequently reflected on the value of experiential learning, with one noting that "Practical experience enhances learning," emphasizing how direct participation reinforced their understanding of classroom content. The caring category included 30 codes, highlighting areas such as contribution to and involvement in community health care (n=9), personal growth and learning (n=8), patient-centered health care (n=7), communication and collaboration (n=5), and personal motivation and awareness (n=1). The human dimension category had 16 codes, emphasizing patient-centered health care and communication (n=6), understanding and importance of community health care (n=5), general practice and broad knowledge (n=3), and realization of the community health care field (n=2). The integration category comprised 40 codes, addressing challenges and responses in community health care (n=13), improvement of knowledge and skills (n=11), communication and cooperation (n=9), and collaboration with the community (n=7). The application category contained 91 codes, focusing on improvement of knowledge and skills (n=20), practice and problem-solving in community health care (n=18), communication and trust-building (n=15), necessity and outlook of community health care (n=13), improvement of prevention and health awareness (n=9), use of information and communication technology (n=7), environmental preparation and collaboration in community health care (n=5), and general practitioners and multidisciplinary knowledge (n=4). Finally, the foundational knowledge category included 128 codes, covering specific initiatives in community health care (n=48), challenges and the current state of community health care (n=38), the definition and role of community health care (n=26), and the aging population and community health care (n=16). One student reflected, "Community healthcare is necessary for supporting aging populations," illustrating growing awareness of the social challenges faced by Japan's communities.

In comparison, the academic year 2022 had 385 codes. The learning how to learn category expanded to 15 codes, covering the acquisition of broad knowledge (n=5), activities to broaden knowledge (n=4), interest and training (n=3), mandatory community health care training (n=2), and improvement of communication skills (n=1). The caring category rose significantly to 58 codes, with emphasis on communication and collaboration (n=15), contribution and involvement in community health care (n=12), patient-centered health care (n=11), personal growth and learning (n=10), and personal motivation and awareness (n=10). Students' reflections often highlighted a stronger sense of teamwork and responsibility, exemplified by one participant who stated, "Collaboration among healthcare providers enhances communication and patient care." Another remarked, "Personal motivation and awareness are key drivers for effective healthcare delivery," underscoring the affective learning fostered by the revised program. The human dimension category also increased to 57 codes, focusing on patient-centered health care and communication (n=26), challenges and improvements in community health care (n=21), and the importance and current state of community health care (n=10). The integration category contained 45 codes, addressing medical resources and collaboration (n=15), sustainability and technology in health care (n=12), patient-centered care methods (n=10), and ethics and communication (n=8). The application category contained 82 codes, emphasizing improvement of medical knowledge and skills (n=20), practice and problem-solving in community health care (n=20), communication and trust-building (n=15), community health care policy and support (n=8), execution ability and motivation (n=8), importance of community health care (n=6), and use of information and communication technology (n=5). Finally, the foundational knowledge category remained consistent with 128 codes, focusing on the quality and supply of health care (n=51), shortage of physicians and medical resources (n=41), importance of communication and consultation (n=23), and aging population and community health care (n=13).

Comparing the data from 2 years, a notable increase was observed in the total number of codes, particularly in the caring and human dimension categories in 2022. Rather than relying solely on descriptive counts, the theme-based synthesis suggests a shift in emphasis from predominantly system- and knowledge-oriented descriptions (theme 1) toward reflections that more frequently articulated professional identity, values, and relationships in care (themes 2-3), alongside clearer intentions for continued learning (theme 4). These patterns are consistent with a more integrated profile across the Fink domains in 2022, while maintaining a stable baseline of foundational knowledge.

Beyond the individual code counts, several higher-level patterns emerged across both cohorts. First, the number of foundational knowledge codes remained stable across the 2 years, indicating a consistent baseline of understanding of community health care concepts. Second, the revised 2022 program demonstrated a marked increase in codes within the caring and human dimension categories. Third, learning how to learn codes also increased in 2022. Taken together, the thematic synthesis (Tables 1 and 2) indicates that the 2022 cohort's reflections more often combined system understanding with interpersonal, value-oriented, and reflective learning elements, whereas the

XSL•FO
RenderX

2021 cohort's reflections were more concentrated in foundational and application-oriented descriptions.

Discussion

Summary of Key Findings

This study aimed to evaluate the educational impact of a COME program on first-year regional quota medical students at Chiba University. Through qualitative content analysis based on the Fink Taxonomy of significant learning, we identified substantial changes in students' perceptions and reflections following the program revision. The revised COME course, which incorporated additional early clinical exposure and faculty development for community hospital attendings, emphasized affective and interpersonal learning, as well as reflective learning—particularly within the caring, human dimension, and learning how to learn domains—while maintaining a stable baseline of foundational knowledge. These shifts were most apparent in themes 2-4, which foreground patient-centered relationships, emerging professional identity, and self-directed learning orientation. This aligns with the study's objective—to explore how early community-based experiences influence students' understanding of and commitment to community health care—and indicates that experiential learning opportunities in real clinical contexts contribute meaningfully to professional identity formation among first-year medical students.

Interpretation and Comparison With Previous Studies

Across both cohorts, several higher-level patterns emerged, offering broader insight into the educational impact of the revised COME program. While the number of foundational knowledge codes remained stable, indicating a consistent baseline understanding of community health care concepts, the revised 2022 program produced substantial increases in the caring, human dimension, and learning how to learn categories. Importantly, beyond these category-level changes, the theme-based synthesis (themes 1-4) indicates that students' reflections in 2022 more often integrated system understanding with interpersonal, value-oriented, and reflective elements of learning (themes 2-4). These findings suggest that the experiential and interpersonal elements of the program contributed to deeper affective, interpersonal, and reflective learning, extending beyond the acquisition of foundational knowledge.

Learning How to Learn and Lifelong Learning

The expansion of the learning how to learn category from 6 codes in 2021 to 15 in 2022 highlights the development of students' reflective and self-directed learning capacities.

This pattern aligns with theme 4 (developing a self-directed learning orientation for community practice), in which students increasingly recognized the importance of continuous learning and professional growth, as reflected in comments such as "Acquiring a broad range of knowledge is important for comprehensive healthcare" and "Ongoing interest and training are essential for professional development." This finding resonates with previous studies emphasizing the importance of reflective learning and self-regulation in developing lifelong learning competencies [32-34]. The structured faculty mentorship and feedback embedded in the revised program may have contributed to this growth by encouraging students to articulate their learning needs and apply theoretical knowledge to real-world contexts.

Caring and Human Dimension: Growth in Empathy and Collaboration

A major finding was the substantial increase in the caring (from 30 to 58 codes) and human dimension (from 16 to 57 codes) categories. These domains were central to themes 2 and 3, which foregrounded patient-centered relationships, teamwork, and emerging professional identity and responsibility toward community service. Students' reflections increasingly addressed interpersonal communication, collaboration, and patient-centered care. One student stated, "Collaboration among healthcare providers enhances communication and patient care," while another reflected, "Patient-centered healthcare ensures that the care provided meets the needs of the patient." These reflections demonstrate an affective shift toward empathy and teamwork, suggesting that early exposure to real patient encounters and mentorship in community hospitals fostered students' understanding of the relational aspects of medicine. This aligns with prior research demonstrating that community-based and interprofessional learning environments enhance empathy and communication among medical students [35-37]. Similarly, our findings echo studies emphasizing that authentic, longitudinal clinical experiences—particularly those in underserved settings—help shape professional identity and strengthen social accountability [38-40]. One student further noted, "Addressing the challenges in community healthcare requires continuous improvement efforts," capturing a sense of responsibility and ongoing self-improvement that mirrors the iterative nature of clinical reasoning and patient care.

Integration and Application: Connecting Theory to Practice

The integration category increased modestly (40 to 45 codes), while the application category slightly decreased (91 to 82 codes). In the theme-based synthesis, integration was the most visible in theme 1 (community health care as a system) and theme 3 (professional identity and responsibility), where students linked collaboration, sustainability/technology, patient-centered care methods, and ethics. For instance, a student commented, "Collaboration between medical institutions is essential for optimizing resources," illustrating recognition of systemic and organizational aspects of health care delivery. However, the slight decline in application codes suggests that opportunities for practical problem-solving and leadership in community health planning could be expanded. Despite this, students continued to value applied knowledge, as shown in reflections such as, "Healthcare providers must be adept at problem-solving to address diverse community health issues." This pattern implies that while the program successfully deepened conceptual and interpersonal understanding, future iterations could further reinforce practical application and leadership in community contexts. These insights are consistent with previous reports that emphasize the need for a balance

XSL•FO
RenderX

between affective learning and skill-based experiential education [41-43].

Foundational Knowledge: Consolidating Core Understanding

The foundational knowledge category remained stable with 128 codes in both years, reflecting consistent comprehension of fundamental community health care principles. This stable foundation corresponded to theme 1 in both cohorts, where students continued to demonstrate awareness of the structural challenges facing Japan's health care system, particularly physician shortages and the aging population. Statements such as "Ensuring high-quality healthcare is critical for meeting community needs" and "Many rural areas face significant shortages of doctors and medical resources" highlight this awareness. This stability suggests that students entered the program with a basic understanding of community health needs and retained this foundation while expanding their capacity for empathy, reflection, and teamwork through experiential learning [44,45].

Comparison With Existing Educational Models

The differences observed between the COME program and early clinical exposure within university hospitals further highlight the program's unique contribution. Unlike hospital-based experiences that often emphasize specialized and controlled environments, the COME program situates students in authentic community settings where resource limitations and multidisciplinary collaboration are inherent. Students reflected that learning from community physicians provided unique insights into the realities of rural health care. As one student expressed,

Learning from physicians embedded in the community provided valuable insights into the challenges and rewards of rural healthcare.

This finding aligns with prior literature suggesting that early, community-facing experiences cultivate adaptability, resource awareness, and social accountability [46,47]. Furthermore, recent evidence indicates that such community-facing clinical exposure not only enhances engagement and learning outcomes but also fosters the development of professional identity, social accountability, and collaborative competencies essential for clinical reasoning [46-48]. In addition, context-specific curriculum design is crucial for optimizing the educational impact of such programs. Understanding regional variations in students' study habits, learning preferences, and access to educational resources allows educators to adapt teaching methods—such as incorporating imaging-based learning modules, interprofessional collaboration, or problem-based approaches—to enhance engagement and relevance in local settings [49]. These insights highlight that while COME provides authentic, real-world exposure to health care challenges, tailoring curricular content and pedagogy to local educational contexts can further enhance learning outcomes and promote sustained motivation among medical students. By engaging with real patients and health care teams early in their education, students not only improved their communication skills but also developed a sense of mission toward addressing regional health disparities.

Limitations

This study has some limitations. First, the sample size was limited to first-year regional quota medical students at a single institution (n=40), which restricts the generalizability of the findings. The participants were from a single admission track; hence, demographic and baseline characteristics between the 2021 and 2022 cohorts were not formally analyzed, which may introduce selection bias. However, both cohorts had similar demographic profiles, showing comparable age ranges and distributions of sex assigned at birth. In addition, the results were derived from students' written reflections rather than objective performance-based assessments. Therefore, while these reflections provide valuable insights into students' perceptions and learning processes, they do not directly capture behavioral or performance-level outcomes. The study also relied on self-reported data, which may be subject to bias. The qualitative nature of the content analysis may also limit the reproducibility of the results, as it involves subjective data interpretation. Furthermore, the study used a cross-sectional design comparing 2 adjacent cohorts without a control group; thus, the observed improvements might reflect cohort effects, prior IPE exposure, or faculty-related factors rather than the revised COME program itself. The analysis also relied primarily on one coder, without interrater reliability testing, audit trails, or member checking, which could affect analytical rigor. In addition, applying the Fink Taxonomy a priori may have constrained data interpretation by fitting responses into predefined categories. However, previous qualitative studies have suggested that increases in code frequency and diversity—when analyzed within a validated theoretical framework—can indicate broader and deeper levels of learning and reflection [23,24]. Recent evidence also supports this interpretation, showing that analyses grounded in theoretically informed frameworks such as the Fink Taxonomy can meaningfully capture students' cognitive and affective development through code expansion and thematic diversification [50]. Therefore, in this study, the observed growth in the number and variety of codes (eg, "Caring," "Human dimension") was interpreted as a qualitative indicator of students' multidimensional learning development rather than a purely quantitative change. Nonetheless, no objective or longitudinal outcomes (eg, rural practice retention or sustained career engagement) were assessed. Another limitation is the short duration of the study, which spanned only 2 academic years; longer-term studies are needed to evaluate the sustained impact of the program on students' career choices and retention in rural health care settings. Finally, the study did not account for external factors that may have influenced students' perceptions, such as their personal experiences, socioeconomic backgrounds, or individual learning preferences. Tailoring educational strategies to align with students' learning styles and real-world experiential opportunities may foster more effective and personalized medical education, ultimately enhancing both theoretical understanding and practical competency [51]. Future research should incorporate such individualized approaches and

XSL•FO
RenderX

explore how these factors interact with community-based learning outcomes.

Conclusions

Enhancements to the COME program at Chiba University, including early clinical exposure and the involvement of faculty-development-trained community hospital attendings, significantly improved students' understanding of patient-centered and community-oriented health care. Our theme-based synthesis suggests that the revised program was associated with greater emphasis on patient-centered relationships, emerging professional identity, and self-directed learning (themes 2-4), while maintaining a stable baseline of foundational understanding (theme 1). These experiences cultivated empathy, reflective learning, and communication skills, which are essential for generalist physicians serving in underserved areas. Beyond these educational benefits, the findings have broader implications for medical education policy and workforce distribution. Structured community-based programs such as COME may contribute to addressing Japan's rural physician shortage by fostering students' identification with community health care roles early in their training. Continual evaluation and iterative refinement of COME programs are necessary to ensure that they remain responsive to regional health care needs and effectively cultivate physicians who are competent, compassionate, and socially accountable.

Acknowledgments

Generative AI tools were used only to support the qualitative data analysis in this study, specifically to assist in identifying potential patterns and alternative categorizations within the dataset. All final coding decisions and interpretations were made exclusively by the human researchers. No generative AI tools were used in the preparation, writing, or editing of the manuscript text.

Funding

No external financial support or grants were received from any public, commercial, or not-for-profit entities for the research, authorship, or publication of this article.

Data Availability

The dataset generated or analyzed during this study is available from the corresponding author on reasonable request.

Authors' Contributions

Conceptualization: KS, K Yamauchi, Y Kamata, SI
Methodology: KS, K Yamauchi, Y Kamata, SI
Investigation: KS, K Yamauchi, Y Kamata, SI
Data Curation: KS, K Yamauchi, NA, Y Kamata, IS
Formal Analysis: KS, K Yamauchi, NA, SI
Qualitative Analysis: KS, K Yamauchi, NA, SI
Writing – Original Draft: KS
Writing – Review & Editing: K Yamauchi, NA, Y Kamata, SA, DS, MM, Y Katsuyama, TO, K Yoshimura, SI
Supervision: SI
Project Administration: KS, K Yamauchi, SI
Resources: KS, K Yamauchi, Y Kamata, DS, MM, YY, TO, K Yoshimura, SI.
Visualization: KS

Conflicts of Interest

None declared.

Multimedia Appendix 1
Students' reports.
[DOCX File , 27 KB - mededu_v12i1e84406_app1.docx]

Multimedia Appendix 2
The Fink Taxonomy of significant learning.
[DOCX File , 27 KB - mededu_v12i1e84406_app2.docx]

Multimedia Appendix 3
Representative prompts and outputs for artificial intelligence–assisted qualitative content analysis.
[DOCX File , 32 KB - mededu_v12i1e84406_app3.docx]

XSL•FO
RenderX

Multimedia Appendix 4
Results of qualitative content analysis of first-year regional quota medical students' reflection reports in the 2021 community-oriented medical education program at Chiba University (n=20).
[DOCX File , 37 KB - mededu_v12i1e84406_app4.docx]

Multimedia Appendix 5
Results of qualitative content analysis of first-year regional quota medical students' reflection reports in the revised 2022 community-oriented medical education program at Chiba University (N=20).
[DOCX File , 33 KB - mededu_v12i1e84406_app5.docx]

References

1. Karan A, Negandhi H, Nair R, Sharma A, Tiwari R, Zodpey S. Size, composition and distribution of human resource for health in India: new estimates using national sample survey and registry data. BMJ Open 2019;9(4):e025979 [FREE Full text] [doi: 10.1136/bmjopen-2018-025979] [Medline: 31133622]
2. Morris S, Sutton M, Gravelle H. Inequity and inequality in the use of health care in England: an empirical investigation. Soc Sci Med 2005;60(6):1251-1266. [doi: 10.1016/j.socscimed.2004.07.016] [Medline: 15626522]
3. Guttmann A, Shipman S, Lam K, Goodman D, Stukel T. Primary care physician supply and children's health care use, access, and outcomes: findings from Canada. Pediatrics 2010;125(6):1119-1126. [doi: 10.1542/peds.2009-2821] [Medline: 20498170]
4. Mick SS, Lee SD, Wodchis WP. Variations in geographical distribution of foreign and domestically trained physicians in the United States: 'safety nets' or 'surplus exacerbation'? Soc Sci Med 2000;50(2):185-202. [doi: 10.1016/s0277-9536(99)00183-5] [Medline: 10619689]
5. Winkelmann J, Muench U, Maier CB. Time trends in the regional distribution of physicians, nurses and midwives in Europe. BMC Health Serv Res 2020;20(1):937 [FREE Full text] [doi: 10.1186/s12913-020-05760-y] [Medline: 33046077]
6. Wu J. Measuring inequalities in the demographical and geographical distribution of physicians in China: generalist versus specialist. Int J Health Plann Manage 2018;33(4):860-879. [doi: 10.1002/hpm.2539] [Medline: 29781216]
7. Ishikawa T, Fujiwara K, Ohba H, Suzuki T, Ogasawara K. Forecasting the regional distribution and sufficiency of physicians in Japan with a coupled system dynamics-geographic information system model. Hum Resour Health 2017;15(1):64 [FREE Full text] [doi: 10.1186/s12960-017-0238-8] [Medline: 28899406]
8. Koike S, Okazaki K, Tokinobu A, Matsumoto M, Kotani K, Kataoka H. Factors associated with regional retention of physicians: a cross-sectional online survey of medical students and graduates in Japan. Hum Resour Health 2023;21(1):85 [FREE Full text] [doi: 10.1186/s12960-023-00871-z] [Medline: 37885012]
9. Kobayashi Y, Takaki H. Geographic distribution of physicians in Japan. Lancet 1992;340(8832):1391-1393. [doi: 10.1016/0140-6736(92)92569-2] [Medline: 1360099]
10. Toyabe S. Trend in geographic distribution of physicians in Japan. Int J Equity Health 2009;8(1):5 [FREE Full text] [doi: 10.1186/1475-9276-8-5] [Medline: 19257879]
11. Tanihara S, Kobayashi Y, Une H, Kawachi I. Urbanization and physician maldistribution: a longitudinal study in Japan. BMC Health Serv Res 2011;11:260 [FREE Full text] [doi: 10.1186/1472-6963-11-260] [Medline: 21982582]
12. Yahata S, Takeshima T, Kenzaka T, Okayama M. Long-term impact of undergraduate community-based clinical training on community healthcare practice in Japan: a cross-sectional study. BMC Med Educ 2020;20(1):337 [FREE Full text] [doi: 10.1186/s12909-020-02258-3] [Medline: 33004049]
13. Fukui S, Shikino K, Nishizaki Y, Shimizu T, Yamamoto Y, Kobayashi H, et al. Association between regional quota program in medical schools and practical clinical competency based on general medicine in-training examination score: a nationwide cross-sectional study of resident physicians in Japan. Postgrad Med J 2023;99(1177):1197-1204. [doi: 10.1093/postmj/qgad059] [Medline: 37474744]
14. Shikino K, Tago M, Hirata R, Watari T, Sasaki Y, Takahashi H, et al. Educational strategies for general medicine education in accordance with the model core curriculum for medical education in Japan. J Gen Fam Med 2023;24(4):272-273 [FREE Full text] [doi: 10.1002/jgf2.619] [Medline: 37484122]
15. Medical Education Model Core Curriculum Expert Research Committee. The model core curriculum for medical education in Japan 2022. Ministry of Education, Culture, Sports, Science and Technology. 2023. URL: https://www.mext.go.jp/content/20250411-mxt_igaku-000028108_00003-2.pdf [accessed 2023-11-23]
16. Ohta R, Ryu Y, Sano C. The contribution of citizens to community-based medical education in Japan: a systematic review. Int J Environ Res Public Health 2021;18(4):1575 [FREE Full text] [doi: 10.3390/ijerph18041575] [Medline: 33562329]
17. Busetto L, Wick W, Gumbinger C. How to use and assess qualitative research methods. Neurol Res Pract 2020;2:14 [FREE Full text] [doi: 10.1186/s42466-020-00059-z] [Medline: 33324920]
18. Sandhu VK, Jose DM, Feldman CH. Underserved communities: enhancing care with graduate medical education. Rheum Dis Clin North Am 2020;46(1):167-178 [FREE Full text] [doi: 10.1016/j.rdc.2019.09.009] [Medline: 31757283]

XSL•FO
RenderX

19. Taxonomy of Significant Learning. The Peak Performance Center. URL: https://thepeakperformancecenter.com/educational-learning/thinking/blooms-taxonomy/taxonomy-of-significant-learning/ [accessed 2025-09-13]
20. Anderson LW, David R. A Taxonomy for Learning, Teaching and Assessing: a Revision of Bloom's Taxonomy. New York, NY: Longman; 2001.
21. Bloom BS. Taxonomy of educational objectives, the classification of educational goals. In: Handbook I: Cognitive Domain. New York, NY: Longmans, Green, and Co; 1956.
22. Morton DA, Colbert-Getz JM. Measuring the impact of the flipped anatomy classroom: the importance of categorizing an assessment by bloom's taxonomy. Anat Sci Educ 2017;10(2):170-175. [doi: 10.1002/ase.1635] [Medline: 27427860]
23. Elo S, Kyngäs H. The qualitative content analysis process. J Adv Nurs 2008;62(1):107-115. [doi: 10.1111/j.1365-2648.2007.04569.x] [Medline: 18352969]
24. Graneheim UH, Lindgren B, Lundman B. Methodological challenges in qualitative content analysis: a discussion paper. Nurse Educ Today 2017;56:29-34. [doi: 10.1016/j.nedt.2017.06.002] [Medline: 28651100]
25. Granero Lucchetti AL, Ezequiel ODS, Oliveira IND, Moreira-Almeida A, Lucchetti G. Using traditional or flipped classrooms to teach "Geriatrics and Gerontology"? Investigating the impact of active learning on medical students' competences. Med Teach 2018;40(12):1248-1256. [doi: 10.1080/0142159X.2018.1426837] [Medline: 29355063]
26. Guest G, Namey E, Chen M. A simple method to assess and report thematic saturation in qualitative research. PLoS One 2020;15(5):e0232076 [FREE Full text] [doi: 10.1371/journal.pone.0232076] [Medline: 32369511]
27. Kadam PD, Chuan HH. Erratum to: rectocutaneous fistula with transmigration of the suture: a rare delayed complication of vault fixation with the sacrospinous ligament. Int Urogynecol J 2016;27(3):505 [FREE Full text] [doi: 10.1007/s00192-016-2952-5] [Medline: 26811110]
28. Rossi M, Rehman S. Integrating artificial intelligence into telemedicine: evidence, challenges, and future directions. Cureus 2025;17(8):e90829. [doi: 10.7759/cureus.90829] [Medline: 40988810]
29. Bennis I, Mouwafaq S. Advancing AI-driven thematic analysis in qualitative research: a comparative study of nine generative models on Cutaneous Leishmaniasis data. BMC Med Inform Decis Mak 2025;25(1):124 [FREE Full text] [doi: 10.1186/s12911-025-02961-5] [Medline: 40065373]
30. Yue Y, Liu D, Lv Y, Hao J, Cui P. A practical guide and assessment on using ChatGPT to conduct grounded theory: tutorial. J Med Internet Res 2025;27:e70122 [FREE Full text] [doi: 10.2196/70122] [Medline: 40367510]
31. Bijker R, Merkouris SS, Dowling NA, Rodda SN. ChatGPT for automated qualitative research: content analysis. J Med Internet Res 2024;26:e59050 [FREE Full text] [doi: 10.2196/59050] [Medline: 39052327]
32. Marceau M, Vachon Lachiver ?, Lambert D, Daoust J, Dion V, Langlois M, et al. Assessment practices in continuing professional development activities in health professions: a scoping review. J Contin Educ Health Prof 2024;44(2):81-89. [doi: 10.1097/CEH.0000000000000507] [Medline: 37490015]
33. Main PAE, Anderson S. Evidence for continuing professional development standards for regulated health practitioners in Australia: a systematic review. Hum Resour Health 2023;21(1):23 [FREE Full text] [doi: 10.1186/s12960-023-00803-x] [Medline: 36941655]
34. Institute of Medicine (US) Committee on Planning a Continuing Health Professional Education Institute. Redesigning Continuing Education in the Healthcare Profession. Washington (DC): National Academies Press (US); 2010.
35. Steinmair D, Zervos K, Wong G, Löffler-Stastka H. Importance of communication in medical practice and medical education: an emphasis on empathy and attitudes and their possible influences. World J Psychiatry 2022;12(2):323-337 [FREE Full text] [doi: 10.5498/wjp.v12.i2.323] [Medline: 35317334]
36. Kerasidou A, Bærøe K, Berger Z, Caruso Brown AE. The need for empathetic healthcare systems. J Med Ethics 2020;47(12):e27 [FREE Full text] [doi: 10.1136/medethics-2019-105921] [Medline: 32709754]
37. Moudatsou M, Stavropoulou A, Philalithis A, Koukouli S. The role of empathy in health and social care professionals. Healthcare (Basel) 2020;8(1):26 [FREE Full text] [doi: 10.3390/healthcare8010026] [Medline: 32019104]
38. Kwame A, Petrucka PM. A literature-based study of patient-centered care and communication in nurse-patient interactions: barriers, facilitators, and the way forward. BMC Nurs 2021;20(1):158 [FREE Full text] [doi: 10.1186/s12912-021-00684-2] [Medline: 34479560]
39. Gilligan C, Powell M, Lynagh M, Ward BM, Lonsdale C, Harvey P, et al. Interventions for improving medical students' interpersonal communication in medical consultations. Cochrane Database Syst Rev 2021;2(2):CD012418 [FREE Full text] [doi: 10.1002/14651858.CD012418.pub2] [Medline: 33559127]
40. King A, Hoppe RB. "Best practice" for patient-centered communication: a narrative review. J Grad Med Educ 2013;5(3):385-393. [doi: 10.4300/JGME-D-13-00072.1] [Medline: 24404300]
41. Munthe C, Fumagalli D, Malmqvist E. Sustainability principle for the ethics of healthcare resource allocation. J Med Ethics 2021;47(2):90-97 [FREE Full text] [doi: 10.1136/medethics-2020-106644] [Medline: 33154090]
42. Lewis S, Damarell R, Tieman J, Trenerry C. Finding the integrated care evidence base in pubmed and beyond: a bibliometric study of the challenges. Int J Integr Care 2018;18(3):11 [FREE Full text] [doi: 10.5334/ijic.3975] [Medline: 30220894]
43. Institute of Medicine (US) Committee on Using Performance Monitoring to Improve Community Health. Improving Health in the Community: A Role for Performance Monitoring. Washington, DC: National Academies Press; 1997.

XSL•FO
RenderX

44. Mitchell P, Cribb A, Entwistle V. Vagueness and variety in person-centred care. Wellcome Open Res 2022;7:170 [FREE Full text] [doi: 10.12688/wellcomeopenres.17970.1] [Medline: 35865218]
45. Haruta J, Ando T, Fujishima S. How do medical students learn in an online community diagnostics program? BMC Med Educ 2023;23(1):15 [FREE Full text] [doi: 10.1186/s12909-023-04007-8] [Medline: 36627667]
46. Ewnte B, Yigzaw T. Early clinical exposure in medical education: the experience from debre tabor university. BMC Med Educ 2023;23(1):252 [FREE Full text] [doi: 10.1186/s12909-023-04221-4] [Medline: 37069522]
47. Ghanbari-Afra L, Adib-Hajbaghery M, Dianati M. Human caring: a concept analysis. J Caring Sci 2022;11(4):246-254 [FREE Full text] [doi: 10.34172/jcs.2022.21] [Medline: 36483689]
48. Ogut E, Yildirim FB, Senol Y, Senol AU. Comprehensive evaluation of the educational impact and effectiveness of specialized study modules in cross-sectional anatomy: a study on student engagement and learning outcomes. BMC Med Educ 2025;25(1):514 [FREE Full text] [doi: 10.1186/s12909-025-07050-9] [Medline: 40211255]
49. Barut C, Ogut E, Karaer E, Yavuz M. Anatomy study preferences of medical students in relation to gender, academic year and geographical distribution: considerations for anatomy education approaches. Bratisl Med J 2025 Mar 13;126(4):482-498. [doi: 10.1007/s44411-025-00078-9]
50. Ishizuka K, Shikino K, Takada N, Sakai Y, Ototake Y, Kobayashi T, et al. Enhancing clinical reasoning skills in medical students through team-based learning: a mixed-methods study. BMC Med Educ 2025;25(1):221 [FREE Full text] [doi: 10.1186/s12909-025-06784-w] [Medline: 39934738]
51. Ogut E, Senol Y, Yildirim F. Do learning styles affect study duration and academic success? Eur J Anat 2017;21(3):235-240.

Abbreviations

AI: artificial intelligence
CBME: community-based medical education
COME: community-oriented medical education
IPE: interprofessional education

Edited by A Stone, T Leung; submitted 19.Sep.2025; peer-reviewed by RS Gomaa Mahmoud, E Ogut, Aminuddin; comments to author 21.Oct.2025; revised version received 23.Dec.2025; accepted 30.Dec.2025; published 19.Jan.2026.

Please cite as:
Shikino K, Yamauchi K, Araki N, Ozaki N, Kamata Y, Aoki S, Katsuyama Y, Sogai D, Miyamoto M, Yoshimura K, Oki T, Ito S
Impact of Community-Oriented Medical Education on Medical Students' Perceptions of Community Health Care: Qualitative Study
JMIR Med Educ 2026;12:e84406
URL: https://mededu.jmir.org/2026/1/e84406
doi:10.2196/84406
PMID:

©Kiyoshi Shikino, Kazuyo Yamauchi, Nobuyuki Araki, Naoto Ozaki, Yu Kamata, Shinya Aoki, Yota Katsuyama, Daichi Sogai, Mai Miyamoto, Kensuke Yoshimura, Takeshi Oki, Shoichi Ito. Originally published in JMIR Medical Education (https://mededu.jmir.org), 19.Jan.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Education, is properly cited. The complete bibliographic information, a link to the original publication on https://mededu.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Comparative Efficacy of Simulation-Based and Traditional Training in Ultrasound-Assisted Regional Anesthesia for Medical Students: Randomized Controlled Trial

David Sánchez-Poveda[1*], MD; Juan Vicente-Mampel[2*], PT, PhD; Belén Curto[3*], PhD; Vidal Moreno[3*], PhD; Juan A García-Esteban[3*], PhD; Felipe Hernández-Zaballos[1*], MD, PhD; Pablo Alonso-Hernández[1*], PhD

[1]Pain Unit, Anaesthesiology Service, Complejo Asistencial Universitario de Salamanca (CAUSA), Salamanca, Spain

[2]Department of Physiotherapy, School of Medicine and Health Science, Valencia Catholic University Saint Vincent Martyr, Valencia, Spain

[3]Department of Computer Science and Automation, University of Salamanca, University of Salamanca, Salamanca, Spain

[*]all authors contributed equally

Corresponding Author:
Juan Vicente-Mampel, PT, PhD
Department of Physiotherapy
School of Medicine and Health Science
Valencia Catholic University Saint Vincent Martyr
Calle Quevedo, número 2
Valencia, 46001
Spain
Phone: 1 963637412
Email: juan.vicente@ucv.es

Abstract

Background: Ultrasound is very important in medicine and teaching, but there are not many formal training programs. We also do not know much about what students think. To be good at using ultrasound, one needs to learn technical, thinking, and seeing skills. This is especially true in regional anesthesia (RA), where mistakes in reading images can cause problems. Training with simulations is a safe and good way to learn these skills. Some models are helpful for teaching how to perform procedures using ultrasound.

Objective: This study aimed to evaluate the effectiveness, localization time, and success rate of traditional teaching versus a new simulation-based teaching method for RA designed by the investigators among undergraduate medical students.

Methods: A prospective, randomized controlled trial was conducted at the University of Salamanca from April 2022 to January 2023. A total of 34 medical students in their fourth to sixth academic years were randomly allocated to either a simulation-based training group using the Haptic Ultrasound Probe or a traditional teaching group. The simulation approach used a realistic probe replica and a software-based ultrasound environment, whereas the traditional method comprised a theoretical lecture and curated audiovisual materials. Two days after training, participants underwent a blinded assessment requiring the identification of peripheral nerve plexuses using an ultrasound device. The primary outcome measured was the successful identification of nerves, and the secondary outcome was the time taken to complete each procedure. Data were analyzed using an intention-to-treat approach.

Results: A total of 34 medical students (fourth to sixth years) were recruited to compare traditional teaching with simulation-based training in ultrasound-guided nerve localization. No statistically significant differences were found in the success rates between the groups. For the interscalene approach, the traditional teaching group achieved a 100% (17/17) success rate compared to 82% (14/17) in the simulation group (P=.07). The time to task completion was similar across most procedures. In the sciatic nerve division, the traditional teaching group was significantly faster, with a mean time of 42.4 (SD 39.5) seconds (P=.02). The regression models showed no significant interaction between the intervention type and academic year. Both teaching methods had positive educational impacts.

Conclusions: Simulation-based learning effectively supports competency acquisition in RA and offers a safe, scalable alternative to traditional methods. Its integration into medical curricula may standardize training, improve skill consistency, and enhance patient safety. Further multicenter studies with larger, diverse cohorts are needed to validate these benefits and guide implementation in medical education.

XSL•FO
RenderX

(JMIR Med Educ 2026;12:e77702) doi:10.2196/77702

KEYWORDS

education; ultrasound; anesthesia; nerve block; simulation

Introduction

Medical specialists are increasingly using ultrasound technology for diagnostic purposes and to guide therapeutic procedures owing to its numerous clinical advantages [1]. Some researchers have characterized portable ultrasound as the "visual stethoscope" of the 21st century [2], highlighting its growing relevance in modern medicine. Point-of-care ultrasound is a portable and versatile imaging modality that allows clinicians to perform rapid bedside assessments in various clinical scenarios. It is accessible and efficient [3,4]. Point-of-care ultrasound has become an essential tool in emergency medicine, critical care, and other specialties [5,6]. Consequently, ultrasound training has been incorporated into the undergraduate medical curriculum, encompassing both theoretical instruction and practical hands-on experience. However, the American Institute of Ultrasound in Medicine reports that only approximately one-third of medical schools in the United States have implemented specialized ultrasound training programs. In addition, there is a noticeable gap in the literature regarding medical students' views on optimal practices for ultrasound education [7]. Furthermore, there is limited understanding of whether its use enhances the acquisition of medical knowledge independent of technical skill development [8].

To effectively acquire ultrasound competencies, medical students must master a combination of technical, cognitive, and perceptual skills. Key areas include proper probe handling, such as the ability to maneuver the probe across anatomical surfaces, and the development of a 3D orientation to accurately visualize and interpret internal structures. Additionally, students must integrate real-time imaging with their anatomical knowledge to make precise adjustments during the examination, which is critical for both diagnostic accuracy and therapeutic efficacy [9]. Equally important is visual training, which enables students to recognize the anatomical structures within ultrasound images. This process requires repeated exposure and active problem-solving to enhance spatial awareness and the ability to discern subtle differences between tissues and organs [10]. These visual and interpretive skills must be grounded in a strong understanding of anatomy to ensure accurate interpretations.

Motor coordination also plays a pivotal role, particularly in interventional procedures that involve inserting needles into targeted soft tissues, where precision is essential to avoid complications. In the field of regional anesthesia (RA), the use of ultrasound has increased significantly over the past few decades because of its ability to enhance the safety and accuracy of procedures that were traditionally guided by anatomical landmarks and neurostimulation [11]. However, to fully realize these safety benefits, trainees must receive comprehensive education on the use of ultrasound technology as the equipment alone cannot guarantee a safe outcome. The number and complexity of RA techniques have also expanded, with some procedures requiring advanced sonoanatomical knowledge for successful performance [9]. In response to this evolution, Regional Anaesthesia UK (the UK division of the European Society of Regional Anaesthesia and Pain Therapy) proposed a classification system that groups RA techniques according to the level of knowledge and expertise required for their performance. Notably, most malpractice incidents associated with RA result from misinterpretation of ultrasound images or inadequate anatomical exploration [12].

Simulation-based learning provides valuable opportunities to develop these competencies in controlled settings. It allows both undergraduate and postgraduate students to practice without the ethical and logistical challenges of performing procedures on live patients or cadavers while still preserving the tactile feedback necessary for motor skill development [13]. Simulation training also enables repetition, feedback, and safe failure, which are key components of effective procedural training. As such, it has emerged as a highly promising approach for training in ultrasound-guided procedures. Several simulation-based teaching models have demonstrated their effectiveness in developing RA skills [14,15]. In this context, our research group developed a novel ultrasound simulator, the Haptic Ultrasound Probe (HUSP), which was officially registered on April 23, 2018, with the Spanish Trademark and Patent Office (*Oficina Española de Marcas y Patentes*) [16]. This study aimed to evaluate the effectiveness of a simulation-based teaching method using HUSP compared to traditional instruction for teaching RA to undergraduate medical students [1].

Methods

Study Design

This study was designed as a prospective, longitudinal, randomized controlled trial and conducted in accordance with the CONSORT (Consolidated Standards of Reporting Trials) guidelines (Multimedia Appendix 1) [17]. Participants were randomly and equally assigned to either a simulation-based education group or a traditional educational method group. Data were collected between April 2022 and January 2023.

Ethical Considerations

Participation was voluntary, and all students provided informed consent before their inclusion in this study. The study adhered to the ethical principles of the Declaration of Helsinki [18] and was approved by the institutional review board of the Department of Surgery, Faculty of Medicine and Health Sciences, University of Salamanca (protocol code 001-2018; July 13, 2018). Measures were implemented to protect participant privacy and confidentiality, including the anonymization of all collected data and secure storage of records. No compensation or incentives were provided to students.

XSL•FO
RenderX

Sample Size Calculation

This study used convenience sampling by recruiting participants who were easily accessible in a university. While this approach allows for efficient data gathering, it is crucial to recognize that it may limit the generalizability of the study due to potential selection biases. Therefore, an intention-to-treat analysis was performed to address this concern.

Participants

A total of 34 students from the Faculty of Medicine at the University of Salamanca (medical degree program) participated in this study. A sample was selected including those from the fourth, fifth, and sixth years of study. Both data collectors and analysts were unaware of the assigned interventions. The sample consisted of 32% (n=11) male and 68% (n=23) female individuals distributed in 2 independent groups. The inclusion criteria were as follows: (1) students enrolled in the medical degree program at the University of Salamanca; (2) students in the fourth, fifth, or sixth year; (3) students who voluntarily provided informed consent; (4) students willing to participate in all activities and assessments of the study; and (5) students with sufficient cognitive and physical capabilities to participate in the teaching interventions (simulation based).

Randomization and Blinding

All educational sessions were conducted by 2 instructors with extensive experience (>10 years) in ultrasound-guided RA following standardized teaching protocols. One of the researchers delivered both the traditional and simulation-based interventions, whereas the second researcher acted as an independent evaluator and remained blinded to group allocation. To minimize bias, group assignment was randomized using a table generated by an independent researcher in Microsoft Excel. A double-blind methodology was implemented; neither the participants nor the evaluator had knowledge of the group allocation or performance outcomes until the end of the study.

Procedures

This study investigated the simulation of interscalene, supraclavicular, and sciatic nerve blocks, including the external and internal popliteal divisions of the nerves, as they appear as Plan A and Plan B blocks in the Regional Anaesthesia UK guidelines. An Esaote MyLab Alpha ultrasound device, identical to that used to generate simulation ecosystem images, was used in the experiment. Two individuals with plexuses that were confirmed to have adequate ultrasound visibility before testing took part in the study. All participants were scheduled for an assessment 2 days after their simulation training with the HUSP simulator. The assessment required participants to identify plexuses in 3 live individuals. To ensure impartiality in student group allocations, a blind investigator who was not involved in providing the theoretical lesson or guiding the simulation ecosystem practice administered the tests. The time taken by each student to locate the plexus was recorded, and successful identification was noted (maximum identification time: 120 seconds). Any attempt exceeding 120 seconds was considered unsuccessful in locating the nerve plexus.

Educational Material

Traditional Educational Method

After a theoretical presentation by an expert, the students were exposed to instructor-curated audiovisual content to augment their comprehension. These materials included videos addressing the 3 principal topics covered in the theoretical lectures. The videos were carefully selected to complement the theoretical content by providing practical demonstrations and visual explanations. The instructor had the option of using YouTube videos, which are often freely accessible and exhibit a range of real-world examples, or consulting a textbook on RA or nerve blocks, which typically offers a more structured and academic perspective with detailed and evidence-based information. The choice between these resources depended on the teaching methodology and the desired depth of the material. While YouTube videos can present practical, real-life applications and demonstrations, textbooks generally provide a more comprehensive and authoritative source of knowledge supported by scholarly references.

Simulation Educational Method

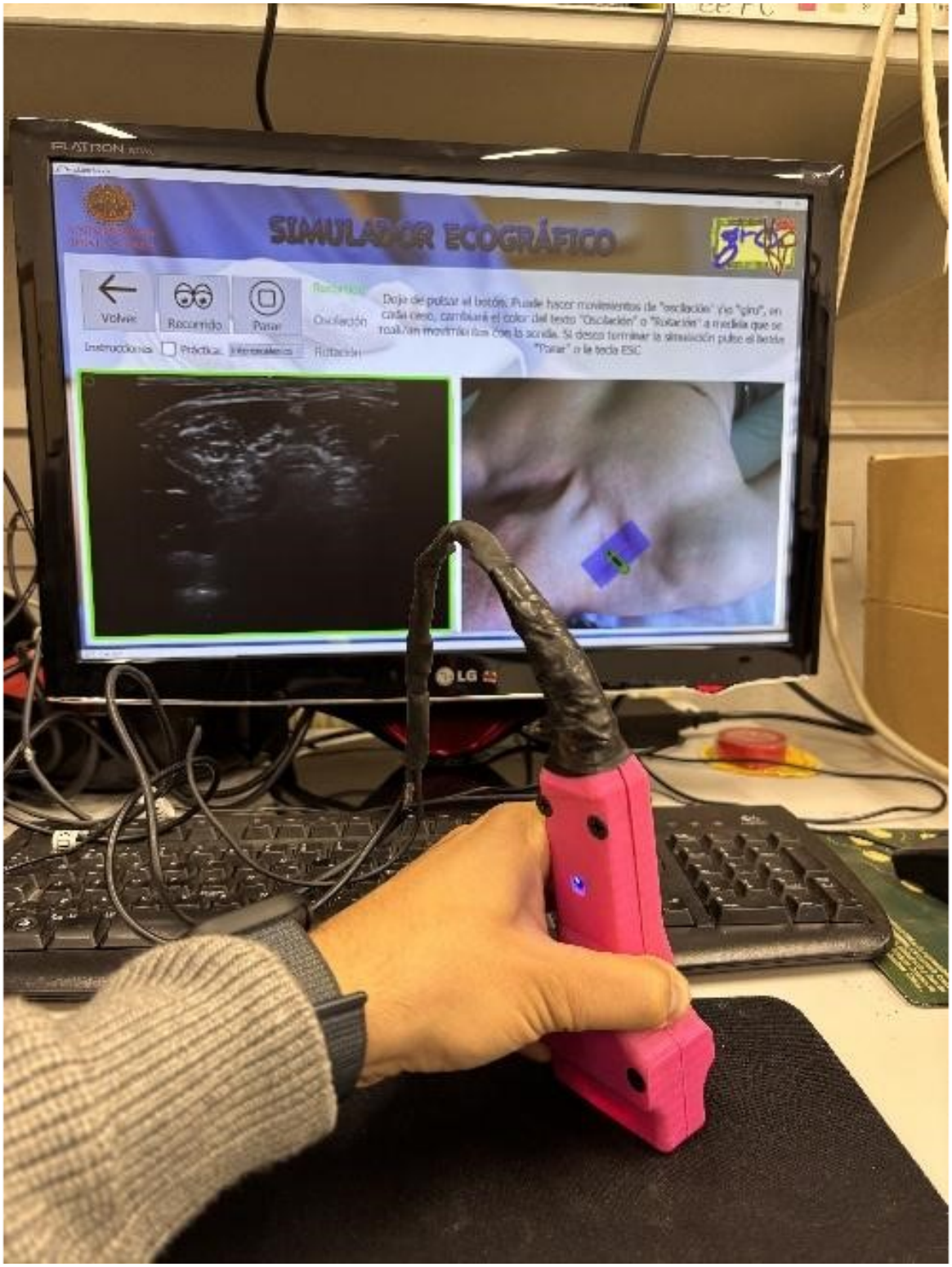
Following a concise theoretical introduction by the same expert, the simulation group engaged in a 2-hour practical session at the Faculty of Science, University of Salamanca. During this session, the participants practiced identifying the brachial plexus using the simulation system.

The ultrasound imaging system used comprised theoretical and practical components.

The Ultrasound Desktop Educational Simulator helped students in acquiring ultrasound images, comprehending the spatial relationship between anatomical structures in the ultrasound image and the overall 3D view of the patient's anatomy, and coordinating hand movements with structure visualization. A significant feature of the Ultrasound Desktop Educational Simulator was a realistic ultrasound probe replica that trainees could operate as if it were an actual ultrasound transducer (Figure 1).

XSL•FO
RenderX

Figure 1. Illustration of the probe's movements as recorded by the software, which are used to enhance the precision of the ultrasound-guided approach.



The simulation system screen displayed a 2D image of the corresponding anatomical pathways. The movement of the probe induced changes in both the skin surface and ultrasound images (Figure 2).

XSL•FO
RenderX

Figure 2. Simulated work environment designed for students, allowing them to perform tasks collaboratively in small groups.



A guide informed the students when they captured correct images (Figure 3). The simulation tool used in this study comprised 2 components: a probe replica that emulated a real ultrasound transducer and a multiprogram desktop application. The probe replica, designated as HUSP by its developers, contained sensors that monitored the trainee's hand movements. The software application generated ultrasound images that were prerecorded using real equipment based on the trainee's probe manipulations. It also enabled students to capture ultrasound images, anatomical surface images, and probe paths. Screen borders appeared when the student obtained correct ultrasound images. On April 23, 2018, the Spanish Trademark and Patent Office granted utility model number 201700521 to protect the HUSP device [16].

Figure 3. Depiction of the student highlighted in green, indicating correct probe positioning for the intervention.



XSL•FO
RenderX

Outcome Measures

The primary outcome variable was the success rate of ultrasound-guided peripheral nerve identification, recorded as a binary outcome (success or failure). The secondary outcome was the time (in seconds) required to complete each nerve localization procedure. The procedures were analyzed across 4 anatomical approaches: interscalene, supraclavicular, popliteal, and sciatic nerve division. The academic year of the students (fourth, fifth, or sixth year) was also included.

Statistical Analysis

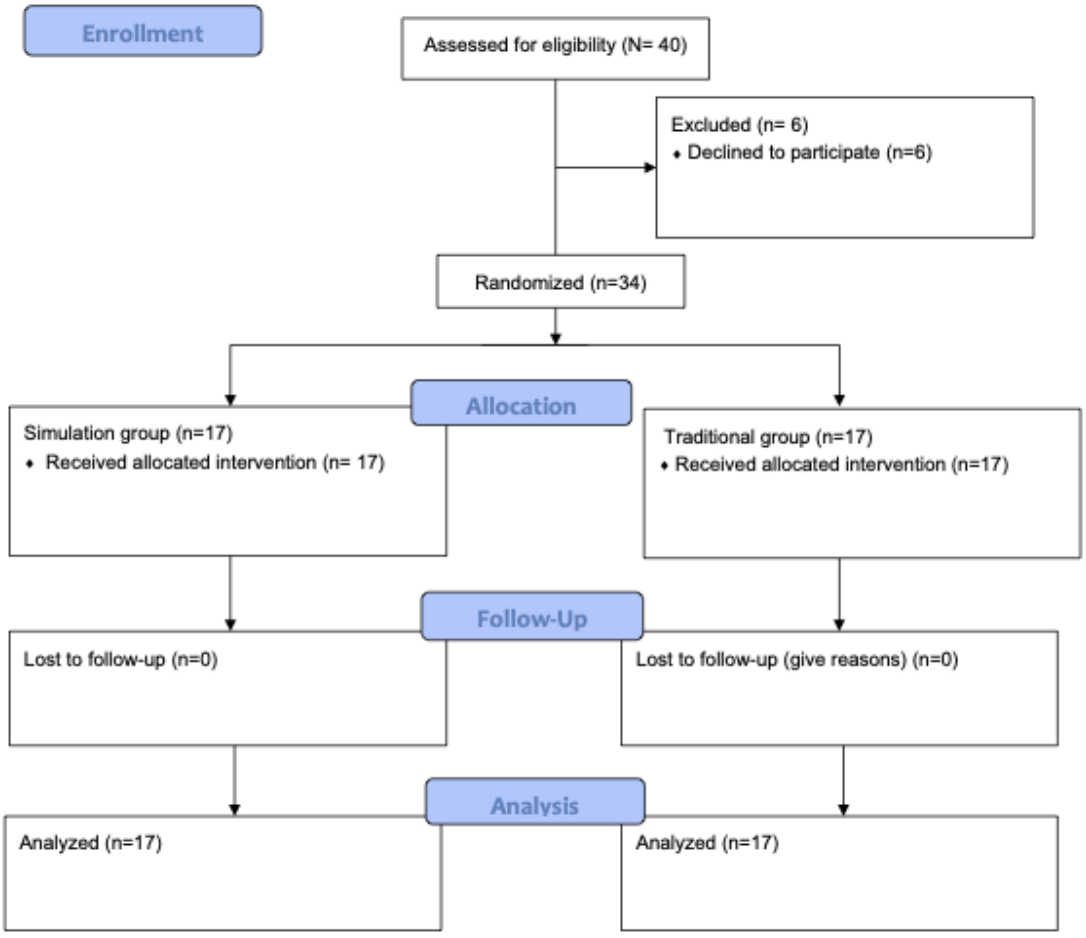
All data are expressed as the mean and SD. Descriptive statistics were used to summarize the participants' demographic characteristics (age, sex, and academic year) and procedure times (in seconds). Categorical variables were reported as absolute frequencies and percentages. Data normality was assessed by visually inspecting histograms. Additionally, the Shapiro-Wilk test was performed to formally evaluate the normality of the data. Given the pilot nature of this study and the limited number of participants per group, an intention-to-treat analysis was performed [19]. The primary outcome, success in nerve identification, was evaluated using the chi-square or Fisher exact tests to compare the proportions between the traditional and simulation-based teaching groups for each anatomical approach. For the secondary outcome (procedure time), mixed linear models were applied to assess the mean differences between groups, including the variable "academic year" as a covariate. This allowed for the evaluation of potential confounding and interaction effects on the results of the study. Interaction terms (group × time) were tested to explore whether the effectiveness of the intervention varied over time [19]. Unstandardized coefficients, 95% CIs, and *P* values were reported for all regression models. A significance level of $P<.05$ was considered statistically significant. All statistical analyses were conducted using the JASP software (version 0.16.3).

Results

Recruitment, Participant Flow, and Sample Characteristics

A cohort of 34 students from the Faculty of Medicine at the University of Salamanca (medical degree program) was recruited for this study. All students provided voluntary consent to participate (Figure 4).

Figure 4. Flowchart of the participants in this study.



The cohort comprised students in the fourth, fifth, and sixth years of the medical degree program. Comprehensive demographic data, including course year, sex, and age, are presented in Table 1. Furthermore, detailed information stratified by course year and duration of each ultrasound procedure is provided in Table 2.

XSL•FO
RenderX

Table 1. Demographic characteristics of undergraduate medical students participating in a randomized educational intervention study comparing a teaching simulator with traditional education for regional anesthesia training.

Variable	All participants (N=34)	Teaching simulator (n=17)	Traditional education (n=17)
Sex, n (%)			
Male	11 (32)	5 (29)	6 (35)
Female	23 (68)	12 (71)	11 (65)
Age (years), mean (SD)	21.8 (3.3)	22.41 (1.3)	22.4 (0.9)
Year of study, n (%)			
Fourth	11 (32)	6 (35.3)	5 (29)
Fifth	15 (44)	7 (41)	8 (47)
Sixth	8 (24)	4 (23.5)	4 (24)

Table 2. Comparison of procedure duration across different ultrasound-guided nerve block approaches between undergraduate medical students trained using a simulator-based method and traditional education.

Approach	Educational intervention (s), mean (SD)	
	Teaching simulator	Traditional education
Interscalene	38.6 (46.2)	31.41 (28.5)
Supraclavicular	51.2 (41.2)	55.7 (41.2)
Popliteal	37.8 (36.3)	42.4 (39.5)
Sciatic nerve division	10.8 (14.5)	24.1 (36.8)

Success Rate of the Intervention

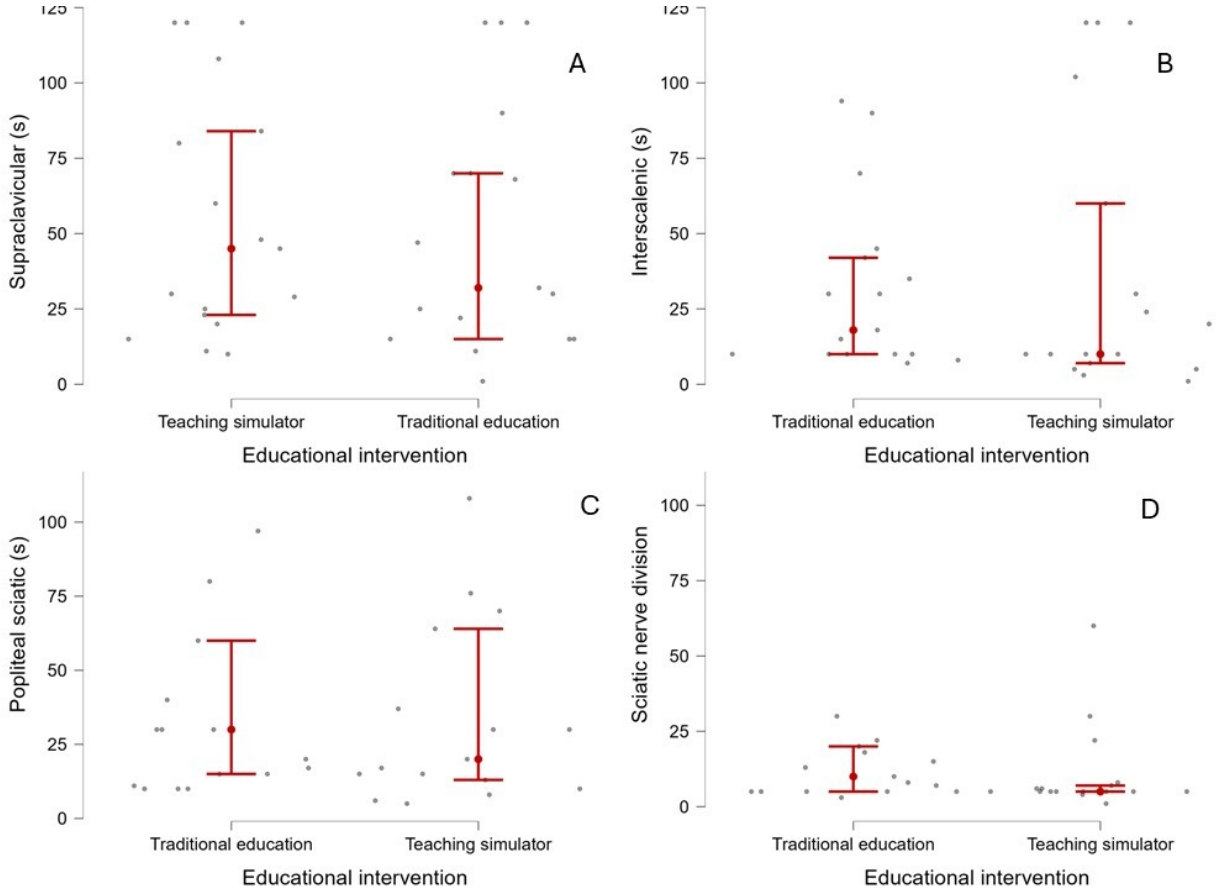
The results revealed no significant differences in successful nerve localization across the various approaches between the traditional teaching and simulation groups. Specifically, in the interscalene approach, 100% (17/17) of the students in the traditional teaching group successfully located the brachial plexus compared to 82% (14/17) in the simulation group, where 18% (3/17) of the students did not succeed. However, no statistically significant difference was observed (P=.07). Similarly, in the supraclavicular approach, both groups demonstrated comparable success, with 18% (3/17) of the students in each group unable to locate the relevant structure. In popliteal sciatic nerve localization, slightly fewer students in the simulation group (1/17, 6%) failed than in the traditional teaching group (2/17, 12%); however, this difference was not statistically significant (P=.07). Finally, for sciatic nerve division, all students in the simulation group successfully identified the division, whereas 12% (2/17) of the students from the traditional group, one from the fourth year and one from the fifth year, were unable to do so.

Supraclavicular Approach

Specifically, there were no statistically significant differences based on the educational interventions received. The reported values correspond to unstandardized coefficients (β), which represent the effect of each intervention in the original units of the outcome variable along with their P values and 95% CIs. Both educational interventions had positive impacts: the simulator group achieved β of 106.72 (95% CI 10.06-203.37; P=.03), whereas the traditional teaching group achieved β of 102.80 (95% CI 5.03-200.57; P=.40; Figure 5A). The academic year covariate in the regression model did not interact with the effect of the educational interventions (β=−10.43, 95% CI −29.79 to 8.91; P=.28).

XSL•FO
RenderX

Figure 5. Scatterplots showing the time (in seconds) required to perform 4 types of nerve blocks according to the educational intervention: teaching simulator vs traditional education (mean and CI).



Interscalenic Approach

The analysis revealed no statistically significant differences based on the educational methods used. Both intervention groups (traditional education and simulator-based education) exhibited positive effects. The simulator-based group achieved β of 79.38 (95% CI 24.78-164.73; *P*=.01), whereas the traditional teaching group achieved β of 74.36 (95% CI 9.03-145.23; *P*=.03; Figure 5B). In the regression model, the academic year covariate did not interact with the effect of the educational interventions (β=−8.69, 95% CI −22.39 to 4.52; *P*=.14).

Popliteal Sciatic Nerve Approach

The analysis demonstrated no statistically significant differences based on the educational methods. The simulator-based group attained β of 98.93 (95% CI 32.54-188.50; *P*=.07), whereas the traditional teaching group achieved β of 105.46 (95% CI 27.59-195.32; *P*=.01; Figure 5C). In the regression model, the academic year covariate did not interact with the effect of the educational interventions (β=−12.63, 95% CI −28.91 to 1.33; *P*=.08).

Sciatic Nerve Division Approach

The analysis demonstrated statistically significant differences based on the educational method used. The simulator-based group attained β of 60.72 (95% CI −3.63 to 125.07; *P*=.06), whereas the traditional teaching group achieved β of 74.67 (95% CI 9.57-139.77; *P*=.02; Figure 5D). Consequently, the traditional educational intervention demonstrated a statistically significant effect, whereas the simulator-based intervention did not reach statistical significance. In the regression model, the academic year covariate did not interact with the effect of the educational interventions (β=−10.22, 95% CI −23.11 to 2.66; *P*=.08).

Discussion

Principal Findings

This study evaluated the effectiveness of a simulation-based educational tool (HUSP) compared to traditional teaching for the acquisition of RA competencies in undergraduate medical students. Overall, no statistically significant differences were observed between the 2 groups regarding successful nerve localization across most approaches. In the interscalene approach, 100% (17/17) of the students in the traditional group successfully identified the brachial plexus compared with 82% (14/17) in the simulation group. Similar results were observed for the supraclavicular and popliteal approaches. In the sciatic nerve division approach, all students in the simulation group succeeded, whereas 12% (2/17) of the students in the traditional group failed, reflecting a statistically significant advantage of simulation-based training over traditional teaching in this specific case. No significant interaction was observed between the academic year and intervention type.

These findings align with those of previous studies supporting the use of simulation technologies for teaching ultrasound-guided procedures [20-22]. Simulation allows for safe and standardized practice of complex procedures and has

XSL•FO
RenderX

been shown to improve knowledge retention and practical skill acquisition among medical students [23]. In this study, the HUSP simulator enabled students to reach level 2 of the Kirkpatrick evaluation model, focusing on the acquisition of knowledge and skills following educational intervention. Specifically, students learned to identify the relevant anatomical landmarks for nerve block procedures using ultrasound. Achieving higher levels of clinical competence requires complementary training involving needle use, as found in cadaver laboratories or high-fidelity simulators. During this study, an additional software module integrating needle simulation was developed, enhancing the potential of the HUSP simulator to support advanced skill acquisition.

Chen et al [2] demonstrated the effectiveness of simulation-based learning and emphasized the need for clarification of optimal simulation modalities. Other studies have reported that using live models or fresh cadaver limbs in elective courses can improve anatomical understanding and provide a valuable clinical context [24]. The HUSP simulator offers an effective alternative, especially when access to live models or cadavers is limited in the clinical setting. Its customizable design allows educators to create tailored simulation exercises, which is a pedagogical advantage. The recent integration of needle simulation hardware has further improved clinical applicability. In terms of cost-effectiveness, the simulator is affordable; the hardware is low cost and compatible with any PC, with the software license being the main investment [25]. These characteristics make it a viable educational solution, particularly in resource-constrained settings [26].

A key strength of this study is its contribution to improving ultrasound education during the preclinical stage by addressing gaps identified in previous research [27]. Despite the growing interest in integrating ultrasound into undergraduate curricula, previous studies have faced limitations such as restricted access to resources, faculty availability, and methodological rigor [28]. The development of a simulator incorporating needle simulation represents a significant step toward comprehensive interventional training for the HUSP system. This study has several limitations. First, the relatively small sample size, dictated by the fixed academic cohort, may limit the generalizability of the findings, a common challenge in similar educational studies [15]. Second, the lack of blinding and the dual role of some authors as both instructors and researchers could have introduced bias. Furthermore, an intention-to-treat analysis was performed, ensuring that all participants were analyzed in their originally assigned groups regardless of protocol adherence, thereby minimizing bias and better reflecting real-world educational conditions.

Conclusions

Simulation-based learning is an effective tool for supporting the acquisition of competencies in RA and offers a practical, safe, and scalable alternative to traditional educational methods. Beyond individual skill acquisition, integrating simulators such as HUSP into medical curricula could standardize training, reduce variability in learner performance, and enhance patient safety by preparing students for clinical practice in a controlled environment. To strengthen the evidence and broaden its applicability, further multicenter studies with larger and more diverse cohorts are warranted, which may inform best practices and guide the implementation of simulation-based approaches in medical education programs.

Acknowledgments

The authors would like to thank all participants of this study. The authors verify and take full responsibility for the use of generative artificial intelligence in the preparation of this manuscript. The authors used Grammarly (Superhuman Platform Inc) to improve the grammar and clarity of the manuscript.

Data Availability

The datasets generated or analyzed during this study are available from the corresponding author on reasonable request.

Funding

No specific funding was received for this work.

Authors' Contributions

Conceptualization: DS-P, JV-M
Data curation: JV-M
Formal analysis: JV-M
Methodology: FH-Z, PA-H, VM
Software and hardware design: VM, BC, JAG-E
Validation: FH-Z, PA-H, VM, BC
Writing – original draft: DS-P
Writing – review & editing: JV-M, FH-Z, PA-H, DS-P
All authors have read and agreed to the published version of the manuscript.

XSL•FO
RenderX

Conflicts of Interest

None declared.

Multimedia Appendix 1
CONSORT checklist.
[PDF File (Adobe PDF File), 222 KB - mededu_v12i1e77702_app1.pdf]

References

1. Wang S, Hossack JA, Klibanov AL. From anatomy to functional and molecular biomarker imaging and therapy: ultrasound is safe, ultrafast, portable, and inexpensive. Invest Radiol 2020 Sep 17;55(9):559-572 [FREE Full text] [doi: 10.1097/RLI.0000000000000675] [Medline: 32776766]
2. Chen WT, Kang YN, Wang TC, Lin CW, Cheng CY, Suk FM, et al. Does ultrasound education improve anatomy learning? Effects of the Parallel Ultrasound Hands-on (PUSH) undergraduate medicine course. BMC Med Educ 2022 Mar 27;22(1):207 [FREE Full text] [doi: 10.1186/s12909-022-03255-4] [Medline: 35346161]
3. Spampinato MD, Luppi F, Cristofaro E, Benedetto M, Cianci A, Bachechi T, et al. Diagnostic accuracy of Point Of Care UltraSound (POCUS) in clinical practice: a retrospective, emergency department based study. J Clin Ultrasound 2024;52(3):255-264. [doi: 10.1002/jcu.23619] [Medline: 38059395]
4. Popat A, Harikrishnan S, Seby N, Sen U, Patel SK, Mittal L, et al. Utilization of point-of-care ultrasound as an imaging modality in the emergency department: a systematic review and meta-analysis. Cureus 2024 Jan;16(1):e52371 [FREE Full text] [doi: 10.7759/cureus.52371] [Medline: 38694948]
5. Szabó GV, Szigetváry C, Szabó L, Dembrovszky F, Rottler M, Ocskay K, et al. Point-of-care ultrasound improves clinical outcomes in patients with acute onset dyspnea: a systematic review and meta-analysis. Intern Emerg Med 2023 Mar;18(2):639-653 [FREE Full text] [doi: 10.1007/s11739-022-03126-2] [Medline: 36310302]
6. Cid-Serra X, Hoang W, El-Ansary D, Canty D, Royse A, Royse C. Clinical impact of point-of-care ultrasound in internal medicine inpatients: a systematic review. Ultrasound Med Biol 2022 Feb;48(2):170-179. [doi: 10.1016/j.ultrasmedbio.2021.09.013] [Medline: 34740496]
7. Allen AJ, White AB, Bacon DR, Beck Dallaghan GL, Jordan SG. Commentary on ultrasound instruction in undergraduate medical education: perspective from two students. Adv Med Educ Pract 2023 Jan;Volume 14:1-7. [doi: 10.2147/amep.s388044]
8. Shanks A, Darwish A, Cook M, Asencio I, Rouse C. Integration of ultrasound simulation to improve medical student knowledge and satisfaction on the obstetrics and gynecology clerkship. AJOG Glob Rep 2023 Aug;3(3):100228 [FREE Full text] [doi: 10.1016/j.xagr.2023.100228] [Medline: 37645647]
9. Woodworth GE, Carney PA, Cohen JM, Kopp SL, Vokach-Brodsky LE, Horn JE, et al. Development and validation of an assessment of regional anesthesia ultrasound interpretation skills. Reg Anesth Pain Med 2015;40(4):306-314. [doi: 10.1097/AAP.0000000000000236] [Medline: 26017720]
10. Shafqat A, Ferguson E, Thanawala V, Bedforth NM, Hardman JG, McCahon RA. Visuospatial ability as a predictor of novice performance in ultrasound-guided regional anesthesia. Anesthesiology 2015 Nov;123(5):1188-1197. [doi: 10.1097/ALN.0000000000000870] [Medline: 26397142]
11. Albrecht E, Chin KJ. Advances in regional anaesthesia and acute pain management: a narrative review. Anaesthesia 2020 Jan;75 Suppl 1:e101-e110 [FREE Full text] [doi: 10.1111/anae.14868] [Medline: 31903582]
12. Turbitt LR, Mariano ER, El-Boghdadly K. Future directions in regional anaesthesia: not just for the cognoscenti. Anaesthesia 2020 Mar 03;75(3):293-297 [FREE Full text] [doi: 10.1111/anae.14768] [Medline: 31268173]
13. Sekhavati P, Ramlogan R, Bailey JG, Busse JW, Boet S, Gu Y. Simulation-based ultrasound-guided regional anesthesia education: a national survey of Canadian anesthesiology residency training programs. Can J Anaesth 2025 Jan 06;72(1):173-180. [doi: 10.1007/s12630-024-02818-x] [Medline: 39107539]
14. Satapathy AR, Khalid IB, Fathil SM. Simulation in regional anaesthesia: a narrative review of its history, evolution and future prospects. J Clin Med 2024 Dec 26;14(1):67 [FREE Full text] [doi: 10.3390/jcm14010067] [Medline: 39797150]
15. Ashokka B, Law LS, Areti A, Burckett-St Laurent D, Zuercher RO, Chin KJ, et al. Educational outcomes of simulation-based training in regional anaesthesia: a scoping review. Br J Anaesth 2025 Feb;134(2):523-534. [doi: 10.1016/j.bja.2024.07.037] [Medline: 39358185]
16. Moreno V, Curto B, Garcia-Esteban JA, Hernández Zaballos F, Alonso Hernández P, Serrano FJ. HUSP: a smart haptic probe for reliable training in musculoskeletal evaluation using motion sensors. Sensors (Basel) 2018 Dec 29;19(1):101 [FREE Full text] [doi: 10.3390/s19010101] [Medline: 30597949]
17. Hopewell S, Chan AW, Collins GS, Hróbjartsson A, Moher D, Schulz KF, et al. CONSORT 2025 statement: updated guideline for reporting randomized trials. Nat Med 2025 Jun;31(6):1776-1783. [doi: 10.1038/s41591-025-03635-5] [Medline: 40229553]
18. WMA Declaration of Helsinki – ethical principles for medical research involving human participants. World Medical Association. URL: https://www.wma.net/policies-post/wma-declaration-of-helsinki/ [accessed 2026-01-29]

XSL•FO
RenderX

19. Twisk JWR. Applied longitudinal data analysis for epidemiology: a practical guide. 2nd ed. Cambridge, UK: Cambridge University Press; 2013.
20. Recker F, Neubauer R, Dong Y, Gschmack AM, Jenssen C, Möller K, et al. Exploring the dynamics of ultrasound training in medical education: current trends, debates, and approaches to didactics and hands-on learning. BMC Med Educ 2024 Nov 14;24(1):1311 [FREE Full text] [doi: 10.1186/s12909-024-06092-9] [Medline: 39543551]
21. Haji-Hassan M, Capra RD, Bolboacă SD. Efficacy of handheld ultrasound in medical education: a comprehensive systematic review and narrative analysis. Diagnostics (Basel) 2023 Dec 14;13(24):3665 [FREE Full text] [doi: 10.3390/diagnostics13243665] [Medline: 38132248]
22. Kaddoura R, Faraji H, Otaki F, Radhakrishnan R, Stanley A, Paulus A, et al. High-fidelity simulation versus case-based tutorial sessions for teaching pharmacology: convergent mixed methods research investigating undergraduate medical students' performance and perception. PLoS One 2024 Aug 16;19(8):e0302609 [FREE Full text] [doi: 10.1371/journal.pone.0302609] [Medline: 39150900]
23. Huang Q, Yan SY, Huang J, Guo Y, Zeng XT, Jin YH. Effectiveness of simulation-based clinical research curriculum for undergraduate medical students - a pre-post intervention study with external control. BMC Med Educ 2024 May 15;24(1):542 [FREE Full text] [doi: 10.1186/s12909-024-05455-6] [Medline: 38750452]
24. Huettner DP, Hahn LK, Noonan Haase MA, Jung FA, Orebaugh SL. Medical student instruction in peripheral nerve blockade utilizing fresh cadaver limbs in a simulation center. J Educ Perioper Med 2023;25(2):E703 [FREE Full text] [doi: 10.46374/volxxv_issue2_Noonan_Haase] [Medline: 37377509]
25. Ylönen M, Paajanen P, Kukkonen T, Torkki P, Paloneva J, Rosqvist E. From emergency room to operating room: multidisciplinary simulation training in emergency laparotomy for ruptured abdominal aortic aneurysm - learning outcomes and costs. Ann Vasc Surg 2025 Feb;111:341-350. [doi: 10.1016/j.avsg.2024.11.017] [Medline: 39617299]
26. Gasco J, Holbrook TJ, Patel A, Smith A, Paulson D, Muns A, et al. Neurosurgery simulation in residency training: feasibility, cost, and educational benefit. Neurosurgery 2013 Oct;73 Suppl 1:39-45. [doi: 10.1227/NEU.0000000000000102] [Medline: 24051881]
27. Draper R, Ellinger D, Trottier M, Armstrong K, Ma W. Ultrasound education in an elective course in the undergraduate medical curriculum improves students' subjective learning and clinical rotation and residency preparedness. Med Sci Educ 2023 Jun 11;33(3):737-746 [FREE Full text] [doi: 10.1007/s40670-023-01800-4] [Medline: 37501805]
28. Neubauer R, Bauer CJ, Dietrich CF, Strizek B, Schäfer VS, Recker F. Evidence-based ultrasound education? - A systematic literature review of undergraduate ultrasound training studies. Ultrasound Int Open 2024 Mar 26;10(CP):a22750702 [FREE Full text] [doi: 10.1055/a-2275-0702] [Medline: 39403546]

Abbreviations

CONSORT: Consolidated Standards of Reporting Trials
HUSP: Haptic Ultrasound Probe
RA: regional anesthesia

Edited by T Leung; submitted 18.May.2025; peer-reviewed by J Wilkinson, G McLeod; comments to author 20.Sep.2025; revised version received 09.Dec.2025; accepted 31.Dec.2025; published 03.Feb.2026.

Please cite as:
Sánchez-Poveda D, Vicente-Mampel J, Curto B, Moreno V, García-Esteban JA, Hernández-Zaballos F, Alonso-Hernández P Comparative Efficacy of Simulation-Based and Traditional Training in Ultrasound-Assisted Regional Anesthesia for Medical Students: Randomized Controlled Trial
JMIR Med Educ 2026;12:e77702
URL: https://mededu.jmir.org/2026/1/e77702
doi:10.2196/77702
PMID:

©David Sánchez-Poveda, Juan Vicente-Mampel, Belén Curto, Vidal Moreno, Juan A García-Esteban, Felipe Hernández-Zaballos, Pablo Alonso-Hernández. Originally published in JMIR Medical Education (https://mededu.jmir.org), 03.Feb.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Education, is properly cited. The complete bibliographic information, a link to the original publication on https://mededu.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Effectiveness of Artificial Intelligence–Assisted Peer Teaching in Orthopedic Clinical Education: Historical Cohort Study

Chengcheng Yu[1], MD; Fangcai Li[1*], MD; Ning Zhang[1], MD; Hejia Hu[1], MD; Hanxu Huang[1], MD; Jingkai Wang[1], MD; Yiqing Tao[1], MD; Yinan Wu[2*], MD

[1]Department of Orthopedics, Second Affiliated Hospital of Zhejiang University, Hangzhou, Zhejiang, China

[2]Zhejiang Cancer Hospital, Hangzhou, Zhejiang, China

[*]these authors contributed equally

Corresponding Author:
Yinan Wu, MD
Zhejiang Cancer Hospital
38 Guangji Road
Hangzhou, Zhejiang, 310022
China
Phone: 86 17816051881
Email: wuyn@zjcc.org.cn

Abstract

Background: Peer teaching is an established pedagogical approach in medical education; yet, traditional methods face challenges including inconsistent knowledge support, variable teaching quality, and limited scalability. Artificial intelligence (AI) large language models offer potential to augment peer teaching by providing on-demand access to medical knowledge and clinical reasoning support. However, AI integration within structured peer teaching has not been systematically evaluated in clinical education.

Objective: This study aims to evaluate the effectiveness of AI-assisted peer teaching compared to traditional peer teaching in orthopedic clinical education, with respect to knowledge acquisition, clinical skills development (particularly clinical reasoning), student engagement, and 3-month knowledge retention.

Methods: This historical cohort study compared 2 consecutive cohorts of medical students (aged 20-27 years, 108/190, 56.8% male) at the Second Affiliated Hospital, Zhejiang University School of Medicine, Hangzhou, China. All eligible students from each cohort were enrolled. The control group (2021 cohort, n=96, taught in 2024) received traditional peer teaching; the intervention group (2022 cohort, n=94, taught in 2025) received AI-assisted peer teaching with access to DeepSeek-V3. Primary outcomes were assessed using a validated 50-item multiple-choice examination (0-100 points) and a 4-station Objective Structured Clinical Examination (OSCE; 0-100 points) with standardized rubrics (intraclass correlation coefficient>0.85). Secondary outcomes included student engagement and satisfaction (5-point Likert scales) and AI usage metrics. Assessments were conducted at baseline, postintervention (8 weeks), and 3-month follow-up. Analysis of covariance adjusted for baseline knowledge, prior AI experience, and learning interest to address observed baseline imbalances.

Results: Using independent samples *t* tests (α=.05, 2-tailed), the AI-assisted group demonstrated significantly higher postintervention knowledge scores (mean 79.69, SD 8.41 vs mean 75.33, SD 9.26; mean difference=4.36, 95% CI 1.84-6.87; *P*<.001; Cohen *d*=0.49). OSCE total scores were significantly higher (mean 80.95, SD 7.57 vs mean 76.24, SD 9.23; mean difference=4.71, 95% CI 2.31-7.11; *P*<.001; *d*=0.56), with clinical reasoning showing the largest effect (mean difference=2.22, 95% CI 1.18-3.25; *P*<.001; *d*=0.61). Analysis of covariance adjusted results remained significant for all primary outcomes (adjusted knowledge difference=3.52, *P*=.002; adjusted OSCE difference=4.52, *P*<.001). At 3-month follow-up (174/190, 91.6%), the AI-assisted group maintained higher knowledge scores (mean 77.36, SD 8.60 vs mean 72.84, SD 10.42; mean difference=4.52, 95% CI 1.68-7.36; *P*=.002; *d*=0.47), with similar knowledge decay rates between groups.

Conclusions: This study provides the first systematic evidence that integrating AI tools within structured peer teaching enhances orthopedic clinical education across multiple domains, including knowledge acquisition, OSCE performance, and student engagement. Unlike prior studies examining AI as a stand-alone learning tool, this work demonstrates the synergistic potential of combining AI knowledge support with peer teaching's social learning benefits, with particularly strong effects on clinical

XSL•FO
RenderX

reasoning. These findings support scalable, cost-effective implementation of AI-augmented peer teaching, though randomized controlled trials are needed to confirm causality and determine optimal implementation strategies.

(JMIR Med Educ 2026;12:e87959) doi:10.2196/87959

KEYWORDS

artificial intelligence; peer teaching; medical education; orthopedic surgery; clinical reasoning; OSCE; Objective Structured Clinical Examination; educational technology; DeepSeek

Introduction

Background and Rationale

Peer teaching, defined as students learning from and with each other, has emerged as a valuable pedagogical approach in medical education [1,2]. The principles of cognitive and social congruence, which posit that peer teachers may better understand and address learners' difficulties due to shared recent learning experiences, underlie the effectiveness of peer teaching [3]. Meta-analyses have demonstrated that peer teaching is as effective as faculty teaching for knowledge acquisition while offering additional benefits including enhanced communication skills, professional identity development, and cost-effectiveness [4,5]. These advantages are particularly relevant in resource-constrained clinical education settings, where faculty shortages and increasing student enrollment create growing pressure on traditional teaching capacity [6].

Despite its benefits, traditional peer teaching faces several limitations. Peer teachers may lack comprehensive medical knowledge and clinical experience, potentially propagating misconceptions if not properly supervised [7]. The quality of peer teaching varies considerably depending on individual peer teachers' preparation and teaching abilities [8]. Peer teachers may struggle to provide evidence-based answers to complex clinical questions on demand, particularly in specialized fields like orthopedic surgery. The scalability of peer teaching is constrained by the availability of trained peer teachers and supervision requirements [9].

These challenges are amplified in orthopedic clinical education, where students must integrate knowledge across multiple domains including anatomy, biomechanics, radiology, and surgical decision-making [10]. Studies have consistently demonstrated that medical students exhibit inadequate musculoskeletal competency upon graduation, with the majority failing basic musculoskeletal knowledge assessments [11]. The cognitive complexity of orthopedic clinical reasoning, requiring simultaneous consideration of mechanism of injury, imaging findings, patient comorbidities, and treatment alternatives, makes it particularly challenging for peer teachers to address learners' questions comprehensively without access to a robust, on-demand knowledge support system. Orthopedic education therefore represents an ideal context for testing whether AI tools can meaningfully augment peer-mediated learning in a cognitively demanding clinical specialty.

Recent advances in artificial intelligence (AI), particularly large language models (LLMs), offer potential solutions to these challenges. AI tools can provide instant access to medical knowledge, assist with differential diagnosis, explain complex concepts, and support clinical reasoning development [12,13]. Studies have demonstrated that LLMs can perform at the level of medical students or junior physicians on standardized medical examinations [14], and preliminary evidence suggests that AI-assisted learning can improve examination performance and clinical decision-making skills in medical education [15,16]. AI tools have also shown promise in generating clinical reasoning prompts and modeling systematic diagnostic approaches [17,18]. However, concerns regarding AI-generated inaccuracies ("hallucinations") and the risk of cognitive offloading highlight the importance of structured implementation with appropriate quality safeguards [19]. The integration of AI tools specifically within peer teaching frameworks, where AI's knowledge capabilities could complement peer teaching's social learning advantages, remains underexplored. Given that orthopedic surgery requires integration of complex anatomical knowledge, biomechanical principles, and clinical decision-making, it represents an ideal test case for evaluating whether AI can meaningfully augment peer-mediated learning in a cognitively demanding clinical specialty [10].

From a theoretical perspective, the integration of AI within peer teaching can be understood through the lens of distributed cognition theory, which posits that cognitive processes are shared across individuals, tools, and the environment [20]. In an AI-assisted peer teaching model, the AI tool serves as a cognitive artifact that extends the collective knowledge base available during peer interactions, potentially transforming the quality of peer discourse. Complementarily, social constructivist learning theory suggests that knowledge is actively constructed through social interaction [21]; AI may enrich these interactions by providing a shared, accessible knowledge resource that both peer teachers and learners can draw upon to deepen their understanding. This theoretical framework predicts that AI integration would enhance not only individual knowledge acquisition but also the collaborative learning dynamics that distinguish peer teaching from self-directed study.

Knowledge Gaps

While both peer teaching and AI-assisted learning have been independently studied, their synergistic combination has not been systematically evaluated. Existing literature has not addressed whether AI tools can augment the strengths of peer teaching while mitigating its limitations. Furthermore, most studies on AI in medical education focus on individual learning, neglecting the collaborative and social learning aspects central to peer teaching [22]. There is also limited evidence on the long-term retention of knowledge acquired through AI-assisted educational interventions and on the differential effects of AI across clinical skill domains. This gap is particularly significant because the potential synergy between AI and peer teaching is

XSL•FO
RenderX

theoretically compelling: AI could address peer teaching's knowledge limitations while peer teaching could mitigate AI's lack of social and motivational dimensions; however, this hypothesis has not been empirically tested. Furthermore, no study has examined whether AI-assisted peer teaching differentially affects higher-order clinical competencies such as clinical reasoning compared to lower-order knowledge recall—a distinction with important implications for curriculum design and AI implementation strategies in medical education.

Objectives and Hypotheses

This study aimed to evaluate the effectiveness of AI-assisted peer teaching compared to traditional peer teaching in orthopedic clinical education. Specifically, we tested four hypotheses: (Hypothesis 1) AI-assisted peer teaching would enhance knowledge acquisition and performance on the Objective Structured Clinical Examination (OSCE); (Hypothesis 2) the intervention would particularly improve clinical reasoning skills; (Hypothesis 3) AI integration would increase student engagement and satisfaction; and (Hypothesis 4) learning advantages would persist at 3-month follow-up.

Methods

Study Design

This historical cohort study compared 2 consecutive cohorts of medical students receiving orthopedic clinical education at Zhejiang University School of Medicine Affiliated Second Hospital. The study followed the STROBE (Strengthening the Reporting of Observational Studies in Epidemiology) guidelines for reporting observational studies [23] and the JARS-Quant (Journal Article Reporting Standards for Quantitative Research) checklist [24].

Inclusion and Exclusion Criteria

Inclusion criteria were (1) enrollment in the required orthopedic clinical rotation at the Second Affiliated Hospital, Zhejiang University School of Medicine, Hangzhou, China; (2) completion of all required assessments (baseline, postintervention, and OSCE); and (3) provision of written informed consent. Exclusion criteria were (1) previous orthopedic clinical experience beyond the standard curriculum; (2) incomplete assessment data; or (3) withdrawal from the study.

Participant Characteristics

Participants were medical students from 2 consecutive academic cohorts. The control group comprised 2021 cohort students (n=96) who completed the orthopedic rotation in 2024, and the intervention group comprised 2022 cohort students (n=94) who completed the rotation in 2025. Baseline demographic and educational characteristics, including age, gender, prior AI experience, baseline knowledge scores, and self-reported learning interest, were collected and compared between groups (Table 1). Participants were medical students from 2 consecutive academic cohorts. The control group comprised 2021 cohort students (n=96) who completed the orthopedic rotation in 2024, and the intervention group comprised 2022 cohort students (n=94) who completed the rotation in 2025. Baseline demographic and educational characteristics, including age, gender, prior AI experience, baseline knowledge scores, and self-reported learning interest, were collected and compared between groups (Table 1).

Table 1. Baseline characteristics of medical students in traditional peer teaching (control, n=96, 2024 cohort) and artificial intelligence–assisted peer teaching (intervention, n=94, 2025 cohort) groups in orthopedic clinical education, Zhejiang University Affiliated Second Hospital, Hangzhou, China.

Characteristics	Control (n=96)	Intervention (n=94)	Test statistics	*P* value
Age (years), mean (SD)	23.43 (1.75)	23.13 (1.95)	t_{188}=1.12	.27
Male, n (%)	52 (54.2)	56 (59.6)	χ^2_1=0.37	.54
Prior AI[a] experience, n (%)	36 (37.5)	20 (21.3)	χ^2_1=5.26	.02
Baseline knowledge, mean (SD)	66.77 (9.71)	67.92 (9.75)	t_{188}=−0.82	.42
Learning interest, mean (SD)	3.98 (0.77)	3.71 (0.82)	t_{188}=2.31	.02

[a]AI: artificial intelligence.

Sampling Procedures

The study used universal sampling: all eligible students from each consecutive academic cohort enrolled in the required orthopedic clinical rotation were included. Data were collected at the Second Affiliated Hospital, Zhejiang University School of Medicine, Hangzhou, China, during the 2024 academic year (control group) and the 2025 academic year (intervention group). No recruitment incentives or monetary compensation were provided. The study used universal sampling: all eligible students from each consecutive academic cohort enrolled in the required orthopedic clinical rotation were included. Data were collected at the Second Affiliated Hospital, Zhejiang University School of Medicine, Hangzhou, China, during the 2024 academic year (control group) and the 2025 academic year (intervention group). No recruitment incentives or monetary compensation were provided.

Sample Size, Power, and Precision

Based on previous peer teaching studies showing effect sizes of d=0.4-0.6 [25], we calculated that 88 participants per group would provide 80% power to detect a medium effect size (d=0.5) at α=0.05 (2-tailed). Our achieved sample sizes (n=96 and n=94) exceeded this requirement, providing adequate statistical power for the primary analyses. Based on previous peer teaching studies showing effect sizes of d=0.4-0.6 [25], we calculated that 88 participants per group would provide 80% power to detect a medium effect size (d=0.5) at α=0.05 (2-tailed). Our

XSL•FO
RenderX

achieved sample sizes (n=96 and n=94) exceeded this requirement, providing adequate statistical power for the primary analyses.

Educational Interventions: Traditional Peer Teaching (Control Group)

The control group participated in a structured peer teaching program. Students were organized into small groups of 6-8 learners with 2 designated peer teachers (senior students who had completed the orthopedic rotation with distinction). Peer teaching sessions occurred twice weekly for 90 minutes each over 8 weeks (total: 16 sessions, 24 hours). Sessions covered core orthopedic topics including fracture management, joint disorders, spinal conditions, musculoskeletal infections, and orthopedic emergencies, with case discussions, physical examination demonstrations, and diagnostic reasoning exercises. Faculty members provided indirect supervision through presession briefings and postsession feedback, with direct observation in approximately 25% of sessions. Peer teachers had access to standard textbooks, clinical guidelines, and pre-prepared teaching materials developed by faculty.

AI-Assisted Peer Teaching (Intervention Group)

The intervention group received the same structural framework with integrated AI support. Students and peer teachers had access to DeepSeek-V3 (DeepSeek), a 671-billion-parameter mixture-of-experts LLM, through the official DeepSeek web application (chat.deepseek.com) and mobile app (iOS/Android). Students accessed the platform using individual accounts during the study period from February to June 2025. No custom fine-tuning, institutional configuration, or application programming interface integration was applied; the model was used in its publicly available, standard configuration with default parameters.

AI was integrated into peer teaching in four modes: (1) Presession preparation: peer teachers used AI to review topics, identify common misconceptions, and generate supplementary materials. (2) During sessions, students queried the AI for clarifications, evidence-based information, and alternative explanations. (3) Case-based learning: AI assisted in generating differential diagnoses, explaining pathophysiology, and modeling clinical reasoning. (4) Postsession: students used AI for self-directed study and follow-up questions.

Representative prompts provided during training included: “A 45-year-old male presents with acute onset low back pain radiating to the left leg after lifting. What are the key differential diagnoses and distinguishing features?” (clinical reasoning); “Explain the Garden classification of femoral neck fractures and its clinical significance for treatment decisions” (knowledge clarification); and “In a patient with suspected osteomyelitis, what laboratory and imaging findings support the diagnosis?” (case discussion). Students were encouraged to develop their own queries.

All participants received a 2-hour training session on effective AI use, including prompt formulation, critical evaluation of AI outputs, and recognition of AI limitations. Faculty review of AI interactions was conducted weekly, with approximately 10% of interaction logs randomly sampled and evaluated for factual accuracy by 2 faculty members with orthopedic expertise. Both groups received identical faculty-led lectures, clinical rotations, and patient encounters; peer teaching sessions were supplementary to the core curriculum. Both groups participated in peer teaching sessions of identical duration (16 sessions, 90 minutes each, over 8 weeks; total 24 hours) and frequency; the only difference between groups was the availability of AI support during and outside these sessions. The intervention group additionally received a one-time 2-hour AI training session prior to the start of peer teaching.

Measures and Covariates

Primary Outcomes

Primary outcomes included: (1) Postintervention Knowledge Score, a 50-question validated multiple-choice examination covering orthopedic basic science, clinical diagnosis, treatment principles, and emergency management (0-100 points); (2) Knowledge Gain, calculated as the difference between postintervention and baseline scores; and (3) OSCE Performance, a standardized 4-station OSCE assessing basic science knowledge application (0-25), clinical diagnosis skills (0-35), treatment principles understanding (0-25), and emergency management competencies (0-15), with a total score of 0-100. Additionally, four clinical skills dimensions were evaluated: physical examination, imaging interpretation, history taking, and clinical reasoning (each 0-25). All OSCE stations were evaluated by blinded faculty assessors using standardized rubrics with established interrater reliability (intraclass correlation coefficient>0.85) [26]. OSCE assessors underwent standardized training at the beginning of each assessment period using identical scoring manuals and calibration exercises to ensure consistency across cohorts. The same examination items and OSCE stations (identical case scenarios, scoring rubrics, and standardized patient scripts) were used for both cohorts.

Secondary Outcomes

Secondary outcomes included student engagement (discussion participation, question quality, and knowledge sharing) and satisfaction (peer teaching satisfaction, method effectiveness, and overall experience), each measured on 5-point Likert scales; AI usage metrics (weekly frequency, usability, utility, and accuracy perceptions) for the intervention group; and follow-up knowledge retention assessed at 3 months postintervention.

Covariates

Covariates included in adjusted analyses were baseline knowledge score (continuous), prior AI experience (dichotomous: yes/no), and self-reported learning interest (continuous, 5-point Likert scale). These were selected based on observed baseline imbalances between groups and their potential to confound the relationship between group assignment and outcomes.

Analytic Strategy

Based on previous peer teaching studies showing effect sizes of d=0.4-0.6 [25], we calculated that 88 participants per group would provide 80% power to detect a medium effect size (d=0.5) at α=0.05 (2-tailed). Our achieved sample size (n=96 and n=94) exceeded this requirement.

XSL•FO
RenderX

Normality of continuous outcome variables was assessed using the Shapiro-Wilk test and visual inspection of Q-Q plots. All primary outcomes met normality assumptions (Shapiro-Wilk $P>.05$), except clinical reasoning scores which showed mild nonnormality; Mann-Whitney *U* tests confirmed parametric results. Between-group comparisons used independent samples 2-tailed *t* tests for continuous variables and chi-square tests for categorical variables, with Mann-Whitney *U* tests as sensitivity analyses for Likert-scale data. Effect sizes were calculated using Cohen *d* (interpreted as small [0.2], medium [0.5], or large [0.8]). All tests were 2-tailed with $\alpha=.05$; 95% CIs are reported for all point estimates [27].

To address baseline imbalances, analysis of covariance (ANCOVA) was performed for all outcomes with baseline knowledge score, prior AI experience (yes/no), and self-reported learning interest as covariates. ANCOVA assumptions (linearity, homogeneity of regression slopes, and normality of residuals) were verified for each model. Missing data patterns were examined; follow-up missingness (8.4%) was assessed for randomness using Little's missing completely at random (MCAR) test, and sensitivity analyses using multiple imputation (m=20, Rubin's rules) were conducted. Secondary analyses included Pearson correlations between AI usage and outcomes, and preplanned subgroup analyses by prior AI experience, baseline knowledge tertiles, and gender. All analyses were conducted using Python 3.12 (scipy 1.x, statsmodels 0.14).

Ethical Considerations

This study was approved by the Institutional Review Board of the Second Affiliated Hospital, Zhejiang University School of Medicine (number Yan2026-0281). Written informed consent was obtained from all participants prior to the study. Students were informed that participation was voluntary, that declining would not affect their academic standing, and that they could withdraw at any time. All data were de-identified using randomly generated participant codes and stored on encrypted institutional servers accessible only to the research team. AI interaction logs were anonymized before analysis. No monetary compensation was provided. No images containing identifiable individuals are included in this manuscript. The educational intervention was implemented as part of an institutionally approved curricular improvement initiative; permission for curricular modifications was obtained from the Office of Medical Education at the Second Affiliated Hospital, Zhejiang University School of Medicine, prior to implementation.

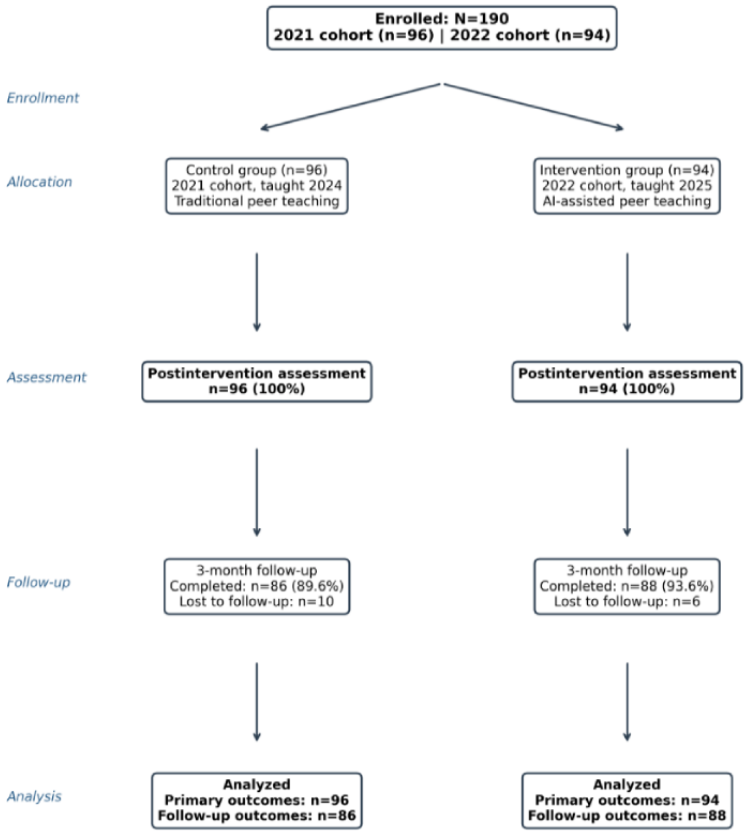
Results

Participant Characteristics

A total of 190 students were enrolled (96 in the control group and 94 in the intervention group). All participants completed baseline and postintervention assessments (100% completion rate). Follow-up data were obtained from 174 participants (91.6% retention rate: control 86/96, intervention 88/94), with no significant difference in follow-up completion between groups ($\chi^2_1=0.55$; $P=.46$). The participant flow is depicted in Figure 1.

XSL•FO
RenderX

Figure 1. Participant flow diagram for a historical cohort study comparing traditional peer teaching and artificial intelligence-assisted peer teaching among medical students in orthopedic clinical education, Second Affiliated Hospital, Zhejiang University School of Medicine, 2024-2025. All enrolled participants (N=190) completed baseline and post-intervention assessments (100% completion rate). At 3-month follow-up, 174 of 190 participants (91.6%) completed the assessment, with no significant difference in completion rates between groups ($\chi^2$1=0.55; *P*=.46). AI: artificial intelligence.



Baseline Characteristics

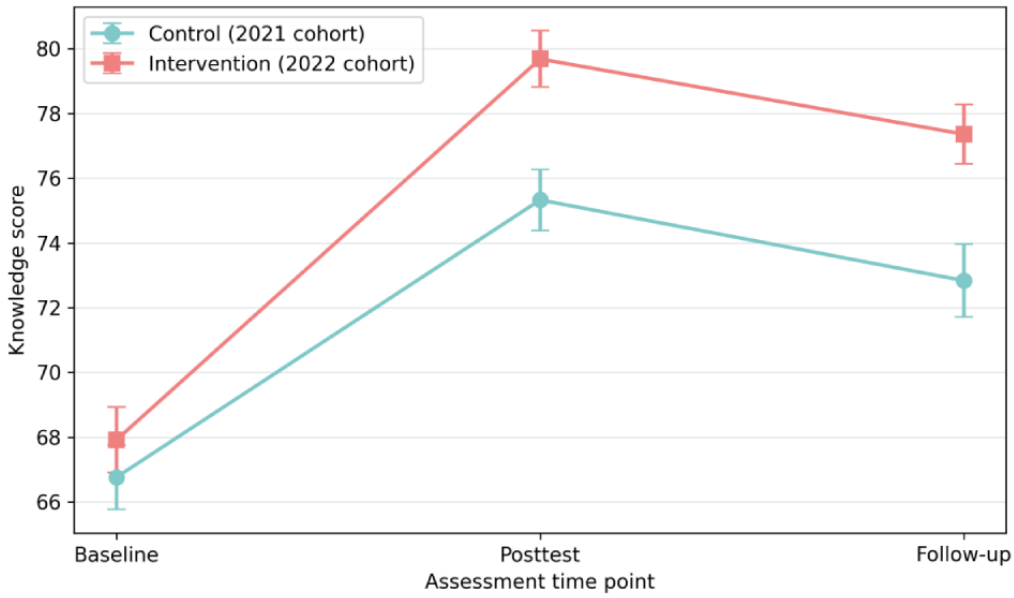
Table 1 presents baseline characteristics. The groups were comparable in age (mean 23.43, SD 1.75 vs mean 23.13, SD 1.95 years; *P*=.27), gender distribution (52/96, 54.2% vs 56/94, 59.6% male; *P*=.54), and baseline knowledge scores (mean 66.77, SD 9.71 vs mean 67.92, SD 9.75; *P*=.42). However, the control group had significantly higher prior AI experience (36/96, 37.5% vs 20/94, 21.3%; χ^2_1=5.26; *P*=.02) and baseline learning interest (mean 3.98, SD 0.77 vs mean 3.71, SD 0.82; *P*=.02). These imbalances were addressed through ANCOVA adjustment.

Primary Outcomes: Knowledge Acquisition

The AI-assisted group demonstrated significantly higher postintervention knowledge scores compared to the control group (mean 79.69, SD 8.41 vs mean 75.33, SD 9.26; mean difference=4.36, 95% CI 1.84-6.87; t_{188}=3.39; *P*<.001; Cohen *d*=0.49), representing a moderate effect (Figure 2). Knowledge gain from baseline to postintervention was also significantly greater in the AI-assisted group (mean 11.77, SD 8.80 vs mean 8.56, SD 9.19; mean difference=3.20, 95% CI 0.64-5.76; *P*=.02; *d*=0.36).

XSL•FO
RenderX

Figure 2. Knowledge development trajectory over time. Mean knowledge scores (SEM) at baseline, postintervention, and 3-month follow-up for control and artificial intelligence–assisted groups. The artificial intelligence–assisted group demonstrated significantly higher scores at both postintervention and follow-up timepoints.



OSCE Performance

The AI-assisted group achieved significantly higher OSCE total scores (mean 80.95, SD 7.57 vs mean 76.24, SD 9.23; mean difference=4.71, 95% CI 2.31-7.11; t_{188}=3.84, *P*<.001; *d*=0.56). Analysis of OSCE component scores revealed differential effects across domains (Table 2, Figure 3). Among knowledge domains, basic science (*P*=.006; *d*=0.41) and clinical diagnosis (*P*<.001; *d*=0.49) showed significant improvements, while treatment principles (*P*=.08) and emergency management (*P*=.89) did not differ significantly. Among clinical skills, clinical reasoning demonstrated the largest effect (t_{188}=4.18, *P*<.001; *d*=0.61), followed by physical examination (*P*<.001; *d*=0.54), while imaging interpretation and history taking showed no significant differences.

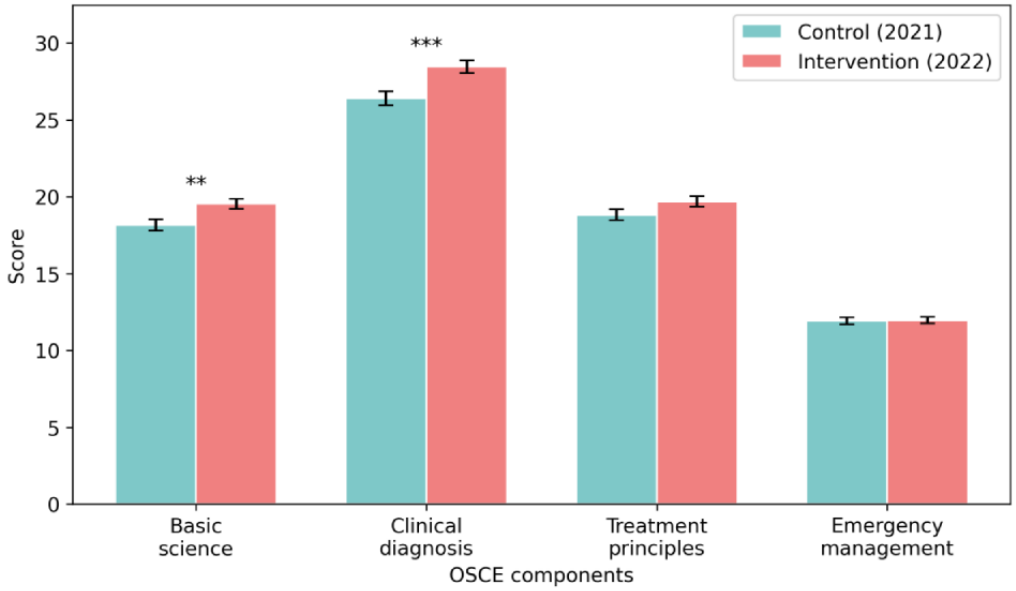
Table 2. Primary and secondary learning outcomes comparing traditional peer teaching (control, n=96) and artificial intelligence–assisted peer teaching (intervention, n=94) in orthopedic clinical education.

Outcome	Control mean (SD)	Intervention mean (SD)	Mean difference	95% CI	*t* test (*df*=188)	*P* value	Cohen *d*
Postintervention knowledge	75.33 (9.26)	79.69 (8.41)	4.36	1.84-6.87	3.39	<.001	0.49
Knowledge gain	8.56 (9.19)	11.77 (8.80)	3.20	0.64-5.76	2.45	.02	0.36
OSCE[a] total score	76.24 (9.23)	80.95 (7.57)	4.71	2.31-7.11	3.84	<.001	0.56
Basic science	18.17 (3.69)	19.54 (3.07)	1.38	0.41-2.34	2.79	.006	0.41
Clinical diagnosis	26.42 (4.48)	28.47 (3.93)	2.05	0.85-3.25	3.35	<.001	0.49
Treatment principles	18.83 (3.53)	19.71 (3.28)	0.88	−0.09 to 1.85	1.77	.08	0.26
Emergency management	11.93 (2.22)	11.97 (2.04)	0.04	−0.56 to 0.65	0.14	.89	0.02
Physical examination	18.66 (3.59)	20.50 (3.21)	1.84	0.87-2.80	3.72	<.001	0.54
Imaging interpretation	19.34 (2.89)	19.67 (2.98)	0.33	−0.51 to 1.16	0.77	.44	0.11
History taking	20.18 (3.76)	20.51 (3.33)	0.34	−0.67 to 1.35	0.66	.51	0.10
Clinical reasoning	18.06 (4.11)	20.28 (3.12)	2.22	1.18-3.25	4.18	<.001	0.61

[a]OSCE: Objective Structured Clinical Examination.

XSL•FO
RenderX

Figure 3. Objective Structured Clinical Examination (OSCE) component scores comparison. Bar chart comparing mean scores (SEM) across OSCE subscales between control and artificial intelligence–assisted groups. Asterisks indicate statistical significance: *P<.05, **P<.01, ***P<.001. Clinical reasoning skills showed the largest improvement in the artificial intelligence–assisted group.



Secondary Outcomes: Student Engagement and Satisfaction

The AI-assisted group demonstrated significantly higher discussion participation (mean 3.43, SD 0.74 vs mean 3.09, SD 0.85; P=.005; d=0.42) and knowledge sharing (mean 3.67, SD 0.75 vs mean 3.35, SD 0.92; P=.01; d=0.38). Question quality did not differ significantly (P=.31). Peer teaching satisfaction (mean 3.87, SD 0.69 vs mean 3.44, SD 0.86; P<.001; d=0.56) and perceived method effectiveness (mean 3.76, SD 0.68 vs mean 3.29, SD 0.92; P<.001; d=0.57) both showed moderate effect sizes. Overall experience showed a trend toward improvement (P=.08; d=0.26). Mann-Whitney U sensitivity analyses for all Likert-scale measures produced results consistent with parametric tests (Table 3 and Figure 4).

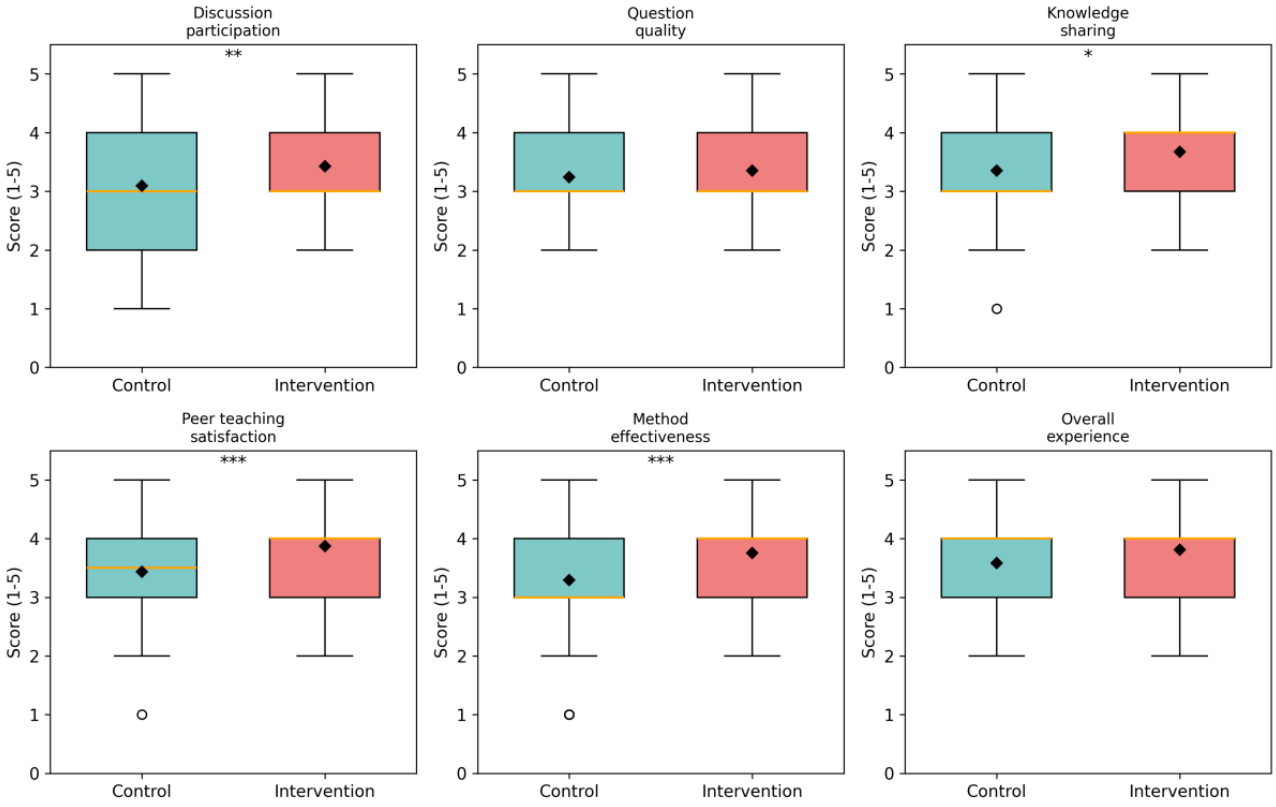
Table 3. Student engagement and satisfaction outcomes.

Measures[a]	Control mean (SD)	Intervention mean (SD)	95% CI	t test (df=188)	P value	Cohen d
Discussion participation	3.09 (0.85)	3.43 (0.74)	0.11-0.56	2.87	.005	0.42
Question quality	3.24 (0.78)	3.35 (0.74)	−0.10 to 0.33	1.01	.31	0.15
Knowledge sharing	3.35 (0.92)	3.67 (0.75)	0.08-0.55	2.59	.01	0.38
Peer teaching satisfaction	3.44 (0.86)	3.87 (0.69)	0.21-0.66	3.85	<.001	0.56
Method effectiveness	3.29 (0.92)	3.76 (0.68)	0.23-0.69	3.95	<.001	0.57
Overall experience	3.58 (0.90)	3.81 (0.83)	−0.02 to 0.47	1.79	.08	0.26

[a]All measures rated on 5-point Likert scale (1=strongly disagree, 5=strongly agree).

XSL•FO
RenderX

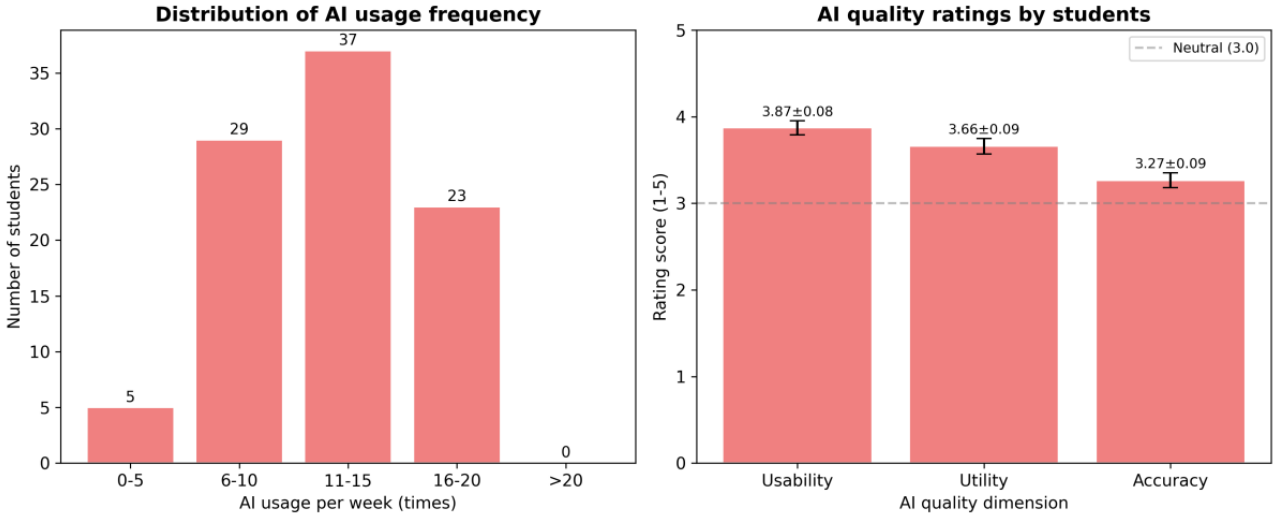
Figure 4. Student engagement and satisfaction outcomes. Box plots showing distribution of engagement and satisfaction measures for both groups. Boxes represent IQR, horizontal lines show medians, and whiskers extend to 1.5×IQR. Individual data points beyond whiskers represent outliers. *P<.05, **P<.01, ***P<.001.



AI Usage Patterns and Quality Perceptions

In the intervention group, students used the AI tool an average of 12.30 (SD 4.01) times per week (median 12, IQR 5-20). Most students (39.4%) used AI 11-15 times weekly; no students exceeded 20 uses, suggesting judicious rather than excessive use. Usability received the highest subjective quality rating (mean 3.87, SD 0.79), followed by utility (mean 3.66, SD 0.87) and perceived accuracy (mean 3.27, SD 0.83). It should be noted that these quality ratings reflect students' subjective perceptions rather than objective measures of AI output quality. AI utility perception was positively associated with postintervention knowledge (r=0.29; P=.004), while usage frequency showed no significant correlation (r=−0.08; P=.44), suggesting that engagement quality rather than quantity drove learning benefits (Figure 5).

Figure 5. Artificial intelligence usage patterns and quality ratings. The left panel shows the frequency distribution of weekly artificial intelligence usage among intervention group students. The right panel displays mean ratings (SEM) for three quality dimensions: usability, utility, and accuracy, all measured as subjective ratings on 5-point scales. AI: artificial intelligence.



XSL•FO
RenderX

Adjusted Analyses

After controlling for baseline knowledge, prior AI experience, and learning interest using ANCOVA, all primary findings remained statistically significant (Table S1 in Multimedia Appendix 1). Adjusted postintervention knowledge scores remained significantly higher in the AI-assisted group (adjusted mean difference=3.52, 95% CI 1.30-5.74; *P*=.002). Adjusted OSCE total scores showed a similar pattern (adjusted mean difference=4.52, 95% CI 2.07-6.97; *P*<.001). The adjusted effect for clinical reasoning remained the largest (adjusted mean difference=2.10, 95% CI 1.04-3.15; *P*<.001). Baseline knowledge was a significant covariate across all knowledge-related models (*P*<.001), while prior AI experience (*P*=.56 for knowledge; *P*=.73 for OSCE) and learning interest (*P*=.38 for knowledge; *P*=.86 for OSCE) were not significant covariates for any outcome, indicating that the observed baseline imbalances did not confound group differences.

Missing Data and Follow-up Knowledge Retention

Follow-up noncompleters (n=16, 8.4%) did not differ significantly from completers on baseline knowledge (*P*=.53), postintervention knowledge (*P*=.38), learning interest (*P*=.64), or group allocation (χ^2_1=0.55, *P*=.46), consistent with the MCAR pattern. Little's MCAR test formally confirmed this assumption (χ^2_7=7.67; *P*=.36). At 3-month follow-up, the AI-assisted group maintained significantly higher knowledge scores (mean 77.36, SD 8.60 vs mean 72.84, SD 10.42; mean difference=4.52, 95% CI 1.68-7.36; *P*=.002; *d*=0.47). Knowledge decay from postintervention to follow-up was similar between groups (intervention: 2.43 points; control: 2.64 points; *P*=.85). Multiple imputation sensitivity analyses (m=20, Rubin's rules) produced consistent results (adjusted difference=3.71, 95% CI 1.00-6.41; *P*=.008).

Subgroup Analyses

Subgroup analysis by prior AI experience revealed no significant interaction effect (*P* for interaction=.42). When stratified by baseline knowledge tertiles, effects were largest in the low-baseline subgroup (*d*=0.65), followed by middle (*d*=0.54) and high (*d*=0.37) tertiles. No significant gender differences were observed (*P* for interaction=.68).

Discussion

Principal Findings

This historical cohort study evaluated the effectiveness of AI-assisted peer teaching in orthopedic clinical education against four prespecified hypotheses. The findings supported all four hypotheses: students in the AI-assisted group demonstrated superior knowledge acquisition and OSCE performance (Hypothesis 1), with particularly pronounced improvement in clinical reasoning skills (Hypothesis 2); the AI-assisted group showed greater engagement and satisfaction (Hypothesis 3); and the learning advantages persisted at 3-month follow-up (Hypothesis 4). After adjusting for baseline imbalances using ANCOVA, all primary findings remained statistically significant, confirming the robustness of the observed effects [25].

These findings are noteworthy in the context of orthopedic education, where students frequently struggle with the integration of basic science and clinical knowledge required for musculoskeletal assessment and management [10,11]. The moderate to large effect sizes observed across multiple outcome domains suggest that AI augmentation addresses a genuine instructional gap in peer teaching rather than producing improvements attributable solely to novelty or increased engagement.

Interpretation of Findings: Knowledge Acquisition and Clinical Reasoning

The improvement in postintervention knowledge scores represents an educationally meaningful moderate effect, aligning with emerging evidence that AI tools can enhance medical knowledge acquisition in educational settings [14,15]. The superior knowledge gain in the AI-assisted group suggests that AI augmented the learning process by providing immediate access to evidence-based information when questions arose—a capability that peer teachers alone often lack [28]. Notably, the proportion of students achieving excellent scores was higher in the AI-assisted group for both knowledge and OSCE assessments, suggesting that AI shifted the entire performance distribution upward rather than merely benefiting a subset of learners. An important consideration is the direction of potential bias: the control group had higher prior AI experience and learning interest, which would favor the control group. Neither was a significant ANCOVA covariate, and the consistency between unadjusted and adjusted effect sizes further supports the conclusion that baseline differences did not confound the observed effects.

The improvement in clinical reasoning, which showed the largest effect size among all outcomes, warrants particular attention. Clinical reasoning, the cognitive process underlying diagnostic and therapeutic decision-making, is notoriously difficult to teach and assess [29,30]. AI tools may facilitate reasoning development through step-by-step differential diagnosis generation and systematic exploration of diagnostic pathways [17,18]. The differential OSCE improvement pattern is informative: strong effects were observed for clinical reasoning and physical examination, while imaging interpretation and history taking showed no significant differences. This pattern suggests that AI specifically enhances deliberative cognitive processes, tasks that benefit from structured analytical frameworks, rather than skills developed primarily through repeated patient contact or interpersonal communication [10].

These findings have important implications for the strategic deployment of AI in clinical education. The relative improvement in clinical reasoning is particularly notable given this competency's documented resistance to educational interventions [29]. The ANCOVA-adjusted OSCE effect was consistent with the unadjusted estimate, confirming that baseline imbalances favoring the control group produced conservative unadjusted estimates. AI integration should therefore be targeted toward cognitive tasks where structured knowledge support can scaffold learning, such as diagnostic reasoning, treatment planning, and evidence-based decision-making, rather than applied uniformly across all clinical competency domains.

XSL•FO
RenderX

Student Engagement and AI Usage Patterns

The AI-assisted group demonstrated higher discussion participation, knowledge sharing, satisfaction, and perceived effectiveness. Rather than displacing peer interaction, a concern frequently raised in the educational technology literature [19], AI appears to have enriched it by providing a richer substrate of information for productive peer exchanges. The nonsignificant difference in question quality suggests that AI enhanced discussion depth without fundamentally altering inquiry patterns. This finding is consistent with the collaborative learning literature suggesting that access to richer information sources can elevate the quality of peer discourse without replacing the social dynamics that make peer teaching effective [31].

An apparent paradox is that students' subjective perception of AI accuracy was only moderate, yet significant learning improvements were observed. This moderate accuracy rating may reflect well-calibrated critical evaluation skills fostered by the AI training session, which emphasized recognizing AI limitations [19]. The process of identifying and correcting AI errors may itself be pedagogically valuable, as error detection requires and reinforces domain knowledge [32]. This interpretation suggests that the combination of accessible but imperfect AI knowledge with human critical evaluation may be more educationally powerful than either element alone—a finding with implications for how AI tools should be framed in educational contexts.

The relationship between AI usage patterns and learning outcomes merits further consideration. AI utility perception, but not usage frequency, predicted learning outcomes, suggesting that students who engaged more thoughtfully with AI, critically evaluating its outputs, integrating AI-provided information with existing knowledge, and using AI as a springboard for deeper inquiry, derived greater educational benefit. This aligns with the broader educational technology literature indicating that the quality of technology-mediated learning interactions, rather than mere exposure duration, determines educational effectiveness [15,22]. Practical implications include the importance of training students not merely in how to use AI tools, but in how to engage with them critically and reflectively.

Subgroup Effects and Knowledge Retention

Stratified by baseline knowledge tertiles, the largest effect was observed in the low-baseline subgroup, followed by the middle and high subgroups, consistent with Vygotsky's zone of proximal development framework: AI may scaffold most effectively for students with the greatest growth potential, while high-performers face ceiling effects on standardized assessments [31]. The absence of significant interaction effects for prior AI experience or gender supports the feasibility of universal implementation, suggesting that AI-augmented peer teaching need not be restricted to technologically experienced students and can benefit learners across diverse demographic subgroups.

At 3-month follow-up, the AI-assisted group maintained higher knowledge scores with similar knowledge decay rates between groups, indicating that AI-assisted learning enhanced initial acquisition rather than fundamentally altering retention mechanisms [33]. The comparable decay rates suggest that AI-assisted learning does not create a dependency that leads to accelerated forgetting once AI access is removed. Rather, the enhanced initial learning appears to be genuinely encoded, consistent with the hypothesis that AI facilitated deeper processing during the learning phase rather than superficial information retrieval.

Whether the clinical reasoning advantages demonstrated under standardized assessment conditions translate into improved diagnostic accuracy and patient outcomes in authentic clinical settings remains an important question requiring longitudinal studies tracking learners into clinical practice [29]. Future studies should also investigate whether spaced retrieval practice combined with AI support could further enhance long-term retention and whether different AI implementation models produce differential learning effects across clinical training stages.

Limitations

This study has several limitations. First, the historical cohort design precludes definitive causal inference. Despite identical assessment instruments, curriculum content, and faculty assignments, temporal confounders, including evolving institutional emphasis on AI literacy, curriculum drift, and societal AI awareness, cannot be fully excluded. Causal conclusions cannot be definitively drawn; these findings should be interpreted as suggestive evidence requiring confirmation through randomized controlled trials.

Second, this single-center, single-specialty study limits generalizability. Third, subtle differences in assessment conditions across years cannot be entirely ruled out, and assessor blinding could have been compromised if assessors were aware of the AI initiative. Fourth, self-reported engagement and satisfaction measures are susceptible to social desirability and novelty bias. Fifth, total study time, including AI-assisted self-directed learning outside structured peer teaching sessions, was not quantified; therefore, we cannot rule out that the intervention group's additional unstructured AI usage contributed learning time beyond the equivalent scheduled peer teaching hours. Structured sessions may represent unaccounted additional learning time. Sixth, our quantitative usage log analysis did not capture the sophistication of individual student-AI interactions; qualitative research is needed. Seventh, potential risks of AI-assisted learning, including cognitive offloading and overreliance, were not directly assessed; future studies should include AI-free assessment conditions.

Conclusions

This study provides the first systematic evidence that integrating AI tools within structured peer teaching can meaningfully enhance orthopedic clinical education. Unlike prior studies examining AI as a stand-alone learning tool, unlike prior studies examining AI as a stand-alone individual learning tool, this work demonstrates the synergistic potential of combining AI knowledge support with peer teaching's social learning benefits, with particularly strong effects on clinical reasoning—a

XSL•FO
RenderX

competency that has historically proven resistant to educational interventions [29,30].

These findings carry practical implications. The intervention required minimal infrastructure (integration of a freely accessible AI tool) and modest training investment (a 2-hour session), suggesting feasibility for widespread adoption. However, implementation challenges include digital equity concerns, the risk of AI overreliance undermining peer interaction, the need for ongoing AI content validation, and faculty development requirements. Institutions should pilot implementations with robust evaluation frameworks before scaling [13,19].

Future research should prioritize randomized controlled trials, multicenter studies across specialties, longitudinal tracking into clinical practice, and qualitative studies of student-AI interaction patterns. As AI tools continue to evolve, the educational community must develop frameworks that maximize learning benefits while preserving the critical thinking and clinical judgment at the core of medical practice [22].

Acknowledgments

We thank all students who participated in this study and the faculty members who contributed to curriculum development and assessment. We acknowledge the support of the Department of Orthopedic Surgery at Zhejiang University School of Medicine Affiliated Second Hospital.

The authors declare the use of generative artificial intelligence (GenAI) in the research and writing process. According to the GAIDeT taxonomy (2025), the following tasks were delegated to GenAI tools under full human supervision:

- Proofreading and editing

- Reformatting

The GenAI tool used was: DeepSeek-V3 (DeepSeek, Hangzhou, China)

Responsibility for the final manuscript lies entirely with the authors.

GenAI tools are not listed as authors and do not bear responsibility for the final outcomes.

Declaration submitted by: All authors

YW and FL are co-corresponding authors.

Funding

This research was funded by the Joint Science and Technology Program Project of the Department of Science and Technology of the National Administration of Traditional Chinese Medicine and the Administration of Traditional Chinese Medicine of Zhejiang Province (GZY-ZJ-KJ-23080) and the National Natural Science Foundation of China (82503233). The funders had no role in study design, data collection, analysis, interpretation, or manuscript preparation.

Data Availability

The data that support the findings of this study are available on reasonable request from the corresponding author. The data are not publicly available due to participant privacy considerations.

Authors' Contributions

CY conceptualized the study, designed the educational intervention, coordinated data collection, performed statistical analyses, and drafted the manuscript. FL supervised the peer teaching program, contributed to study design, oversaw quality control of AI interactions, and critically revised the manuscript for important intellectual content. H Hu contributed to data collection and management. H Huang contributed to OSCE assessment coordination. JW contributed to the development of teaching materials. YT contributed to participant recruitment and follow-up data collection. YW conceived the study, provided overall supervision, secured funding, and critically revised the manuscript. All authors approved the final version and agree to be accountable for all aspects of the work.

Conflicts of Interest

The authors declare no conflicts of interest. DeepSeek had no role in study design, data collection, analysis, interpretation, or manuscript preparation.

Multimedia Appendix 1

XSL•FO
RenderX

Analysis of Covariance (ANCOVA) Results for All Primary and Secondary Outcomes, Adjusting for Baseline Knowledge, Prior AI Experience, and Learning Interest. Historical Cohort Study Comparing Traditional Peer Teaching (Control, n=96) and AI-Assisted Peer Teaching (Intervention, n=94), Second Affiliated Hospital, Zhejiang University School of Medicine, 2024–2025. [DOCX File , 11 KB - mededu_v12i1e87959_app1.docx]

References

1. Komócsi A, Csaba G, Csöngei EV, Fischer K, Filipánits K, Czopf L, et al. Academic performance and progression among near-peer tutors: a comparative analysis in undergraduate medical education. Med Teach 2026;48(3):394-404. [doi: 10.1080/0142159X.2025.2553658] [Medline: 40932010]
2. Ten Cate O, Durning S. Dimensions and psychology of peer teaching in medical education. Med Teach 2007;29(6):546-552. [doi: 10.1080/01421590701583816] [Medline: 17978967]
3. Zhang H, Liao AWX, Goh SH, Wu XV, Yoong SQ. Effectiveness of peer teaching in health professions education: a systematic review and meta-analysis. Nurse Educ Today 2022;118:105499. [doi: 10.1016/j.nedt.2022.105499] [Medline: 35961134]
4. Guraya SY, Abdalla ME. Determining the effectiveness of peer-assisted learning in medical education: a systematic review and meta-analysis. J Taibah Univ Med Sci 2020;15(3):177-184 [FREE Full text] [doi: 10.1016/j.jtumed.2020.05.002] [Medline: 32647511]
5. Rees EL, Quinn PJ, Davies B, Fotheringham V. How does peer teaching compare to faculty teaching? A systematic review and meta-analysis (.). Med Teach 2016;38(8):829-837. [doi: 10.3109/0142159X.2015.1112888] [Medline: 26613398]
6. Ahmed AI, Batt AM. Exploring faculty development initiatives in medical education in resource-limited settings: perspectives and challenges. BMC Med Educ 2025;25(1):1501 [FREE Full text] [doi: 10.1186/s12909-025-07848-7] [Medline: 41146191]
7. Zhang Y, Maconochie M. A meta-analysis of peer-assisted learning on examination performance in clinical knowledge and skills education. BMC Med Educ 2022;22(1):147 [FREE Full text] [doi: 10.1186/s12909-022-03183-3] [Medline: 35248051]
8. Yu TC, Wilson NC, Singh PP, Lemanu DP, Hawken SJ, Hill AG. Medical students-as-teachers: a systematic review of peer-assisted teaching during medical school. Adv Med Educ Pract 2011;2:157-172 [FREE Full text] [doi: 10.2147/AMEP.S14383] [Medline: 23745087]
9. Lerchenfeldt S, Mi M, Eng M. The utilization of peer feedback during collaborative learning in undergraduate medical education: a systematic review. BMC Med Educ 2019;19(1):321 [FREE Full text] [doi: 10.1186/s12909-019-1755-z] [Medline: 31443705]
10. Rosen J, Kartik P, Vella-Baldacchino M. From classroom to clinic: a scoping review of critical thinking and decision-making in orthopaedic education for medical students and residents. Curr Rev Musculoskelet Med 2025;19(1):9. [doi: 10.1007/s12178-025-10002-x] [Medline: 41430025]
11. Nishime N, Seibel M, Lindskog D, Wiznia D. Medical student musculoskeletal knowledge: examining the impact and value of an orthopaedic surgery clerkship using the freedman and bernstein examination. BMC Med Educ 2025;25(1):141 [FREE Full text] [doi: 10.1186/s12909-025-06673-2] [Medline: 39881308]
12. Temsah A, Alhasan K, Altamimi I, Jamal A, Al-Eyadhy A, Malki KH, et al. DeepSeek in healthcare: revealing opportunities and steering challenges of a new open-source artificial intelligence frontier. Cureus 2025;17(2):e79221. [doi: 10.7759/cureus.79221] [Medline: 39974299]
13. Boscardin CK, Gin B, Golde PB, Hauer KE. ChatGPT and generative artificial intelligence for medical education: potential impact and opportunity. Acad Med 2024;99(1):22-27 [FREE Full text] [doi: 10.1097/ACM.0000000000005439] [Medline: 37651677]
14. Kung TH, Cheatham M, Medenilla A, Sillos C, De Leon L, Elepaño C, et al. Performance of chatGPT on USMLE: potential for AI-assisted medical education using large language models. PLOS Digit Health 2023;2(2):e0000198 [FREE Full text] [doi: 10.1371/journal.pdig.0000198] [Medline: 36812645]
15. Wu Y, Zheng Y, Feng B, Yang Y, Kang K, Zhao A. Embracing chatGPT for medical education: exploring its impact on doctors and medical students. JMIR Med Educ 2024;10:e52483 [FREE Full text] [doi: 10.2196/52483] [Medline: 38598263]
16. Abd-Alrazaq A, AlSaad R, Alhuwail D, Ahmed A, Healy PM, Latifi S, et al. Large language models in medical education: opportunities, challenges, and future directions. JMIR Med Educ 2023;9:e48291 [FREE Full text] [doi: 10.2196/48291] [Medline: 37261894]
17. Savage T, Nayak A, Gallo R, Rangan E, Chen JH. Diagnostic reasoning prompts reveal the potential for large language model interpretability in medicine. NPJ Digit Med 2024;7(1):20 [FREE Full text] [doi: 10.1038/s41746-024-01010-1] [Medline: 38267608]
18. Hirosawa T, Kawamura R, Harada Y, Mizuta K, Tokumasu K, Kaji Y, et al. ChatGPT-generated differential diagnosis lists for complex case-derived clinical vignettes: diagnostic accuracy evaluation. JMIR Med Inform 2023;11:e48808 [FREE Full text] [doi: 10.2196/48808] [Medline: 37812468]
19. Eysenbach G. The role of chatGPT, generative language models, and artificial intelligence in medical education: a conversation with chatGPT and a call for papers. JMIR Med Educ 2023;9:e46885 [FREE Full text] [doi: 10.2196/46885] [Medline: 36863937]

XSL•FO
RenderX

20. Boyle JG, Walters MR, Jamieson S, Durning SJ. Distributed cognition: theoretical insights and practical applications to health professions education: AMEE guide No. 159. Med Teach 2023;45(12):1323-1333. [doi: 10.1080/0142159X.2023.2190479] [Medline: 37043405]
21. Durning SJ, Artino AR. Situativity theory: a perspective on how participants and the environment can interact: AMEE Guide no. 52. Med Teach 2011;33(3):188-199. [doi: 10.3109/0142159X.2011.550965] [Medline: 21345059]
22. Gilson A, Safranek CW, Huang T, Socrates V, Chi L, Taylor RA, et al. How does chatGPT perform on the United States medical licensing examination (USMLE)? The implications of large language models for medical education and knowledge assessment. JMIR Med Educ 2023;9:e45312 [FREE Full text] [doi: 10.2196/45312] [Medline: 36753318]
23. Wang X, Kattan MW. Cohort studies: design, analysis, and reporting. Chest 2020;158(1S):S72-S78. [doi: 10.1016/j.chest.2020.03.014] [Medline: 32658655]
24. Appelbaum M, Cooper H, Kline RB, Mayo-Wilson E, Nezu AM, Rao SM. Journal article reporting standards for quantitative research in psychology: the APA publications and communications board task force report. Am Psychol 2018;73(1):3-25. [doi: 10.1037/amp0000191] [Medline: 29345484]
25. McConnell MM, Monteiro S, Bryson GL. Sample size calculations for educational interventions: principles and methods. Can J Anaesth 2019;66(8):864-873. [doi: 10.1007/s12630-019-01405-9] [Medline: 31119552]
26. Yazbeck Karam V, Park YS, Tekian A, Youssef N. Evaluating the validity evidence of an OSCE: results from a new medical school. BMC Med Educ 2018;18(1):313 [FREE Full text] [doi: 10.1186/s12909-018-1421-x] [Medline: 30572876]
27. Windish DM, Diener-West M. A clinician-educator's roadmap to choosing and interpreting statistical tests. J Gen Intern Med 2006;21(6):656-660 [FREE Full text] [doi: 10.1111/j.1525-1497.2006.00390.x] [Medline: 16808753]
28. Schiff A, Salazar D, Vetter C, Andre J, Pinzur M. Results of a near-peer musculoskeletal medicine curriculum for senior medical students interested in orthopedic surgery. J Surg Educ 2014;71(5):734-737. [doi: 10.1016/j.jsurg.2014.01.007] [Medline: 24776870]
29. Daniel M, Rencic J, Durning SJ, Holmboe E, Santen SA, Lang V, et al. Clinical reasoning assessment methods: a scoping review and practical guidance. Acad Med 2019;94(6):902-912. [doi: 10.1097/ACM.0000000000002618] [Medline: 30720527]
30. Thampy H, Willert E, Ramani S. Assessing clinical reasoning: targeting the higher levels of the pyramid. J Gen Intern Med 2019;34(8):1631-1636 [FREE Full text] [doi: 10.1007/s11606-019-04953-4] [Medline: 31025307]
31. Cook DA. Creating virtual patients using large language models: scalable, global, and low cost. Med Teach 2025;47(1):40-42. [doi: 10.1080/0142159X.2024.2376879] [Medline: 38992981]
32. Klang E, Portugez S, Gross R, Lerner R K, Brenner A, Gilboa M, et al. Advantages and pitfalls in utilizing artificial intelligence for crafting medical examinations: a medical education pilot study with GPT-4. BMC Med Educ 2023;23(1):772. [doi: 10.1186/s12909-023-04752-w]
33. Timmer MCJ, Steendijk P, Arend SM, Versteeg M. Making a lecture stick: the effect of spaced instruction on knowledge retention in medical education. Med Sci Educ 2020;30(3):1211-1219 [FREE Full text] [doi: 10.1007/s40670-020-00995-0] [Medline: 34457784]

Abbreviations

AI: artificial intelligence
ANCOVA: analysis of covariance
JARS-Quant: Journal Article Reporting Standards for Quantitative Research
LLM: large language model
MCAR: missing completely at random
OSCE: Objective Structured Clinical Examination
STROBE: Strengthening the Reporting of Observational Studies in Epidemiology

Edited by S Brini; submitted 18.Nov.2025; peer-reviewed by M Metcalf, T Li, F Fukuzawa; comments to author 24.Feb.2026; revised version received 02.Jun.2026; accepted 03.Jun.2026; published 24.Jun.2026.

Please cite as:
Yu C, Li F, Zhang N, Hu H, Huang H, Wang J, Tao Y, Wu Y
Effectiveness of Artificial Intelligence–Assisted Peer Teaching in Orthopedic Clinical Education: Historical Cohort Study
JMIR Med Educ 2026;12:e87959
URL: https://mededu.jmir.org/2026/1/e87959
doi:10.2196/87959
PMID:

XSL•FO
RenderX

©Chengcheng Yu, Fangcai Li, Ning Zhang, Hejia Hu, Hanxu Huang, Jingkai Wang, Yiqing Tao, Yinan Wu. Originally published in JMIR Medical Education (https://mededu.jmir.org), 24.Jun.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Education, is properly cited. The complete bibliographic information, a link to the original publication on https://mededu.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Collaborative Simulation-Based Learning to Develop Competencies in Physical Therapy Students: Randomized Controlled Trial

Iván Cuyul-Vásquez[1,2*], PT, MSc; Hilda Aravena-Páez[3*], PT, MSc; Felipe Araya-Quintanilla[4], PT, PhD; Sergio Guinez-Molinos[5], PhD; Rodrigo Núñez-Cortés[6], PT, PhD; Héctor Gutiérrez Espinoza[7,8], PT, PhD

[1]Departamento de Procesos Terapéuticos, Facultad de Ciencias de la Salud, Universidad Católica de Temuco, Temuco, Chile

[2]Facultad de Ciencias de la Salud, Universidad Arturo Prat, Sede Victoria, Chile

[3]Facultad de Medicina, Departamento de Clínicas, Kinesiología, Universidad Católica del Norte, Coquimbo, Chile

[4]Escuela de Kinesiología, Facultad de Ciencias de la Rehabilitación y Calidad de Vida, Universidad San Sebastián, Santiago, Chile

[5]Laboratorio de Simulación e Informática Biomédica, Facultad de Medicina, Universidad de Talca, Talca, Chile

[6]Department of Physical Therapy, Faculty of Medicine, University of Chile, Santiago, Chile

[7]Faculty of Education, Universidad Autónoma de Chile, Av. Pedro de Valdivia 425, Santiago, Chile

[8]Universidad Espíritu Santo, Samborondón, Ecuador

[*]these authors contributed equally

Corresponding Author:
Héctor Gutiérrez Espinoza, PT, PhD
Faculty of Education, Universidad Autónoma de Chile, Av. Pedro de Valdivia 425, Santiago, Chile

Related Article:

This is a corrected version. See correction statement: https://mededu.jmir.org/2026/1/e112897

Abstract

Background: Although the implementation of simulation-based learning has progressed substantially in medicine, physical therapy education shows a lower level of development regarding standardized collaborative models.

Objective: This study aims to determine the effectiveness of collaborative versus individual simulation-based learning in developing clinical competence in undergraduate physical therapy students in the care of standardized patients with low back pain flare-ups.

Methods: A parallel-group randomized controlled trial was conducted. Thirty-one physical therapy undergraduate students were randomized (1:1) to a collaborative or individual simulation-based learning group. Students underwent 3 weeks of clinical simulation training. Standardized patient assessments were performed before and after the simulation-based learning intervention. Clinical competence was evaluated as the primary outcome, while effective communication and therapeutic alliance were evaluated as secondary outcomes.

Results: For clinical competence, a significant main effect of time was observed (η_p^2=0.739; P<.001), with equivalent pre-to-post improvements in both the collaborative (Δ=+6.17; P_{Bonf}<.001) and individual simulation-based learning groups (Δ=+5.39; P_{Bonf}<.001); the interaction effect was nonsignificant (P=.61). For communication, no significant time (P=.91) or group × time interaction (P=.83) effects were observed. Therapeutic alliance scores demonstrated a significant main effect of time (η_p^2=0.294; P=.008), indicating an overall decrease, though group-specific pre-to-post declines did not survive Bonferroni correction (both P_{Bonf}≥.11).

Conclusions: Collaborative simulation-based learning effectively develops physical therapy clinical competencies, and its effects are comparable to those of individual simulation-based learning with a standardized patient. However, there were no differences between the groups for effective communication or therapeutic alliance. Further research is warranted to assess the efficacy of these modalities in diverse clinical settings.

Trial Registration: OSF Registries 10.17605/OSF.IO/YFWHT; https://osf.io/yfwht/overview

(JMIR Med Educ 2026;12:e85442) doi:10.2196/85442

XSL•FO
RenderX

KEYWORDS

simulation training; patient simulation; low back pain; physical therapy; health education

Introduction

Assessment in health care education is critical to guarantee and certify clinical competence, an operational construct integrating cognition, clinical reasoning, communication, and psychomotor skills [1]. In physical therapy, acquiring these competencies depends heavily on clinical context and multidimensional domains, including patient-centric care and interpersonal skills [2]. Particularly, the management of musculoskeletal conditions such as chronic low back pain (CLBP) represents a cornerstone of professional practice [3]. Although low back pain is highly prevalent [4], its clinical course is often nonlinear, manifesting as acute, recurrent "flare-ups" characterized by sudden symptom worsening, severe emotional distress, catastrophizing, and fear-avoidance behaviors [5,6]. Consequently, developing clinical competencies within undergraduate curricula to effectively manage the complex intersection of tissue irritability, cognitive-behavioral barriers, and communication challenges during a CLBP flare-up is an urgent educational necessity.

Simulation-based learning in health care education has been proven to be an effective learning method by providing a similar condition of clinical practice in a controlled and safe environment [7,8]. Simulation-based experiences allow students to reflect and learn from their mistakes in a safe environment, ultimately enhancing patient safety [9,10]. Consequently, simulation-based education has experienced exponential growth [11] and has been used to assess professional competencies related to patient-centric care, practice psychomotor, and teamwork skills [12]. High-fidelity simulation, relying on expensive, technologically advanced manikin simulators or heavily staffed standardized patient programs, presents substantial financial barriers, making wide-scale implementation unsustainable for institutions with limited economic resources [13]. Despite these operational challenges, there is a scarcity of literature establishing how low-cost, high-fidelity simulation frameworks can be optimized to achieve clinical performance outcomes comparable to those of resource-intensive traditional modalities.

Simulation-based learning integrated within a peer-assisted learning framework represents a novel paradigm to address this pedagogical and economic gap [14]. Preliminary studies have demonstrated that students and educators are satisfied with what can be achieved through teamwork between peers, patients, and health care professionals [14-16]. Collaborative simulation-based learning is a structured learning model for the acquisition and assessment of clinical competencies in which small teams of students collaborate to design and perform clinical simulation scenarios [17,18]. In this approach, students occupy the center of the instructional design process, alternating roles between physical therapists, observers, and standardized patients, culminating in instructor-led debriefings [17]. By decentralizing the instructor's traditional role and eliminating the ongoing financial burden of external professional actors, this collaborative architecture theoretically maximizes contextual and social complexity while minimizing institutional operational costs [17,18]. Crucially, within the context of complex chronic conditions, the collaborative model may optimize student learning by distributing cognitive load, thereby preserving the emotional regulation and communication skills required to manage a challenging patient.

While the implementation of simulation-based learning has progressed substantially in medicine, physical therapy education shows a lower level of development regarding standardized collaborative models [19]. Furthermore, the efficacy of using collaborative simulation-based learning to address highly complex, psychosocially driven clinical scenarios, such as CLBP flare-ups, remains virtually unexamined. Although collaborative architectures are presumed to build clinical competencies cost-effectively, their comparative efficacy against traditional, resource-demanding individual simulations has not yet been established for outcomes that depend directly on interpersonal dynamics, such as patient perception and the therapeutic bond. Robust experimental data are required to determine if student-led collaborative designs compromise or enhance multidimensional outcomes, particularly those metrics highly sensitive to patient-provider interactions. Therefore, the aim of this study was to determine the effectiveness of collaborative versus individual simulation-based learning in developing clinical competence, effective communication, and therapeutic alliance among undergraduate physical therapy students during the care of standardized patients with CLBP flare-ups.

Methods

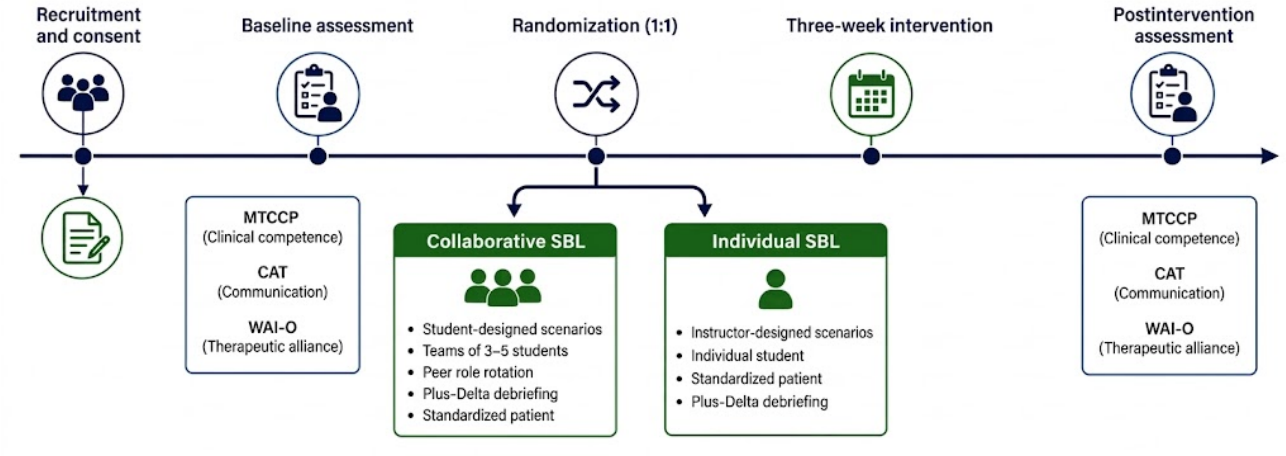
Design

The report of this study followed the recommendations of the Simulation-Based Research Extensions for the CONSORT (Consolidated Standards of Reporting Trials; Checklist 1) Statement [20] and the Modified Medical Education Research Study Quality Instrument [21].

This was a parallel-group randomized controlled trial with a pretest-posttest design comparing collaborative and individual simulation-based learning over 3 weeks in undergraduate physical therapy students (Figure 1). Participants were randomly assigned (1:1) to either collaborative or individual simulation-based learning and completed standardized patient assessments before and after the intervention. The learning assessment consisted of an individual standardized patient encounter, from which the primary outcome (clinical competence) and secondary outcomes (communication and therapeutic alliance) were evaluated. The study was conducted in the rehabilitation laboratories of the Faculty of Health Sciences of the Catholic University of Temuco, Chile.

XSL•FO
RenderX

Figure 1. Study design. CAT: Communication Assessment Tool; MTCCP: Measurement Tool for Clinical Competencies in Physiotherapy; SBL: simulation-based learning; WAI-O: Working Alliance Inventory–Observer Version.



Recruitment and Eligibility Criteria

Nonprobabilistic convenience sampling was used for this study. All eighth-semester physical therapy students (N=48) enrolled in the Musculoskeletal Rehabilitation course during 2022 were screened for eligibility. The inclusion criterion required the successful completion of the course. An independent researcher, who held no grading responsibilities or administrative ties to the students, conducted an informational session in December 2022. During this meeting, the study objectives and simulation-based experiences were detailed, emphasizing that participation was strictly voluntary and would result in no academic prejudice. Of the 48 eligible students, 31 (64.5%) provided written informed consent and were enrolled. Data collection and clinical simulation training took place between December 2022 and January 2023, coinciding with the start of the summer break.

Randomization and Blinding

Participants were randomly assigned to either an experimental group with collaborative simulation-based learning or a control group with individual simulation-based learning. A simple computer-based randomization (1:1) was performed via a website [22] by an assistant not associated with the research. The evaluators and the statistical analyst were blinded to group assignment. Allocation concealment was achieved through opaque sealed envelopes. Participants were blinded to the research hypothesis. Due to the nature of this research, the instructor who conducted the simulation-based experience could not be blinded to group assignment.

Intervention: Simulation-Based Learning

The learning objective was to apply a physical therapy intervention based on exercise and education to a standardized patient with CLBP. Collaborative simulation-based learning (experimental group) was carried out according to the structure proposed by Guinez-Molinos et al [18]: (1) educational design: the creation of the evidence-based guide provided the theoretical framework for students to design, simulate, and evaluate collaborative scenarios; (2) collaborative design: 3 teams of 3 to 5 students were formed. Each team designed a clinical case based on the theme covered in class during the week. Students were also encouraged to write a brief clinical history. Additionally, they assigned specific roles within their simulation scenario. Through this method, weekly active participation was promoted among all students. Each week, the roles rotated among the students in each group; (3) simulation: the student guide explained the clinical simulation procedure to 2 students from another team and controlled the time. Observing students identified opportunities for improvement; (4) debriefing: unlike the protocol proposed by Guinez-Molinos et al [18], a plus-delta debriefing was used [23]. Conversely, the control group underwent individual simulation-based learning, following a traditional structure that used standardized patients. Details of each simulation-based learning protocol are presented in Multimedia Appendix 1.

The procedures for simulation-based learning were standardized into five stages to ensure validity and comparability between the groups:

1. Pre-briefing: The prebriefing procedures were similar for both groups and led by the same instructor. Each week, a 1.5-hour face-to-face lecture on the theoretical content was held, followed by a 30-minute information session dedicated to the simulation-based experience (SBE) procedures and the general history of the standardized patient. In addition, an evidence-based guide for independent study was provided, summarizing the theoretical content related to the rehabilitation of patients with CLBP and acute flare-ups. This guide was included in the lectures [19].
2. Design: This stage lasted 20 minutes for both groups to balance exposure time. In the control group, the scenario design was carried out beforehand by the instructor, who also trained the standardized patients. During this time, the control group students reviewed their notes before the briefing began. In the experimental group, students designed the scenario in teams of 3 to 5 people. This process included creating a brief case history and assigning specific roles: patient, companion, observer, and student guide. The instructor provided the general characteristics of the case to ensure that the diagnosis and duration of pain were

XSL•FO
RenderX

comparable to those of the control group. The scenarios aimed to have students confront 3 types of patients with CLBP, presenting with nociceptive, neuropathic, or nociplastic pain. In the final 2 weeks, the scenarios included patients experiencing flare-ups. The characteristics of the scenarios are reported in Table S1 (Multimedia Appendix 2).

3. Briefing: This phase lasted a total of 7 minutes. The instructor presented the fiction contract and psychological safety guidelines, including establishing an atmosphere of trust, normalizing error, maintaining confidentiality, and creating a judgment-free environment. In the control group, the instructor explained the procedure to a randomly selected student, who then had 2 minutes to read the case history posted on the room door (Table S2 in Multimedia Appendix 2). In the experimental group, the student guide from each team was responsible for explaining the procedure to 2 students from another team, followed by a 2-minute reading of the case history, similar to the control group.
4. Scenario: The scenario lasted 10 minutes for both groups. Students had to gather information, reach a consensus, and implement a physical therapy intervention based on exercise and education for a standardized patient with an acute episode of CLBP. In both groups, the performance was observed by the instructor and peers via a private live YouTube stream to avoid interruptions to the scenario. In the control group, the student interacted directly with a standardized patient from a university theater company, who had been previously trained to ensure script consistency. In the experimental group, 2 students from an external team acted as physical therapists and performed the intervention within a scenario created by another group. The team that designed the case provided 2 students with the roles of patient and companion, alongside a student guide and observers.
5. Debriefing: Immediately after the scenario, a debriefing session of approximately 20 minutes was conducted by the same instructor for both groups. The Plus-Delta model was used [23], focusing on the analysis of the learning objective, clinical reasoning, and observed performance. During this phase, the instructor reinforced the continuity of the psychological safety guidelines initially established, encouraging reflection on opportunities for improvement and the successes achieved during the intervention.

Data Collection and Outcomes

Demographic data, such as age, sex assigned at birth (female or male), work, and number of courses, were collected in an online form. As potential confounder variables, the Academic Stressors Scale, the Academic Self-Efficacy Scale, and the R-14 Academic Resilience Scale were included in the online form.

All students were assessed individually before and after a 3-week simulation-based learning intervention (either collaborative or individual). The assessment consisted of a simulation-based scenario involving standardized patients. Clinical competency was defined as the primary outcome, while communication and therapeutic alliance served as secondary outcomes. Like other studies [24,25], students' patient-centered communication skills were evaluated by standardized patients during the scenario; these patients had been previously trained to use the specific evaluation instrument. To minimize recall bias, the standardized patients completed their evaluations immediately following each encounter.

Consistent with previously published protocols, clinical competence and therapeutic alliance were evaluated via video recordings by independent physical therapists who were unfamiliar with the students and remained blinded to group assignments [26]. All students' clinical performances were recorded using high-definition video and audio. To facilitate this, a camera equipped with an ambient microphone was positioned 3 m from the simulation scenario. This setup allowed for subsequent independent assessment by physical therapists with 5 years of experience in musculoskeletal rehabilitation. Prior to the assessment, these evaluators completed a 3-hour training session on the application of the instruments used to measure therapeutic alliance and clinical competence. This training involved evaluating 8 pilot video recordings, reviewing discrepancies, and receiving feedback from the principal investigator.

Learning Assessment

The pre- and post-intervention evaluations at the simulation-based scenario lasted 12 minutes, during which each student performed an intervention for a standardized patient experiencing a CLBP flare-up. Two clinical cases were developed to evaluate postintervention learning outcomes in both groups. A clinical case was used in the initial assessment, and another in the final assessment (Table S3 in Multimedia Appendix 2). The initial clinical characteristics were maintained to make the cases comparable, and only factors such as name, occupation, and family or occupational context were modified. To prepare the standardized patient and the physical space of the scenarios, the same training and review procedures were used as in the simulation-based learning group. In addition, in the training of the standardized patient, it was explained how to evaluate the communication of the students through a validated questionnaire.

Primary Outcome: Clinical Competence

This was evaluated with the Measurement Tool for Clinical Competencies in Physiotherapy (MTCCP), a valid and reliable (α=.98; intraclass correlation coefficient [ICC]=0.9) instrument to assess clinical competency in real and educational physiotherapy settings [27]. The MTCCP consists of 20 items, grouped into 2 subscales: professional behavior and clinical reasoning. The MTCCP uses a discrete measurement scale from 1 to 5. Those MTCCP items applicable to the simulated scenarios were selected for this research. Eight elements were considered to evaluate the video recordings (4 on the professional behavior subscale and 4 on the clinical reasoning subscale). Therefore, the maximum MTCCP score for this investigation was 40 points. A higher score indicates a better command of clinical competence, while lower scores indicate a worse command of clinical competence.

XSL•FO
RenderX

Secondary Outcomes

Communication

This was evaluated using the Communication Assessment Tool (CAT). The CAT is a valid and reliable instrument (α=.93) to assess real or standardized patients' perception of student patient-centered communication skills [24]. The CAT consists of 14 items grouped in 1 dimension and uses a discrete measurement scale from 1 to 5. The standardized patients answered the CAT with regard to each student at simulation-based assessment. A higher score indicates excellent communication, whereas lower scores indicate poor communication.

Therapeutic Alliance

This was evaluated with the observer version of the Working Alliance Inventory–Observer Version (WAI-O) [26]. WAI-O is a valid and reliable (internal consistency=0.98; ICC=0.92) instrument used to measure the therapeutic alliance within a session [28,29]. The WAI-O consists of 36 items grouped into 3 dimensions: bond, goals, and tasks. It uses a discrete measurement scale from 1 to 7. A higher score indicates a stronger therapeutic alliance, whereas lower scores indicate a weaker therapeutic alliance.

Statistical Methods

The sample size calculation was based on a power analysis "a priori compute required sample size" with an effect size (d=1.85) extracted from a previous and relevant study [30]. With an alpha of .05, power (1- β) of 95%, and Student t test, a minimum of 9 participants per group are needed. This sample estimate was increased by 20% after allowing for potential withdrawals, giving 11 participants in each group. The sample size calculation was performed using the G*Power 3.1.9.2 software (Department of Psychology at Heinrich Heine University Düsseldorf).

Descriptive statistics were used to describe the demographic and clinical characteristics of the participants. Continuous variables were the mean and SD, and categorical variables were the number and percentage.

Prior to inferential analyses, the tenability of the assumptions underlying the ANOVA model was systematically evaluated. The normality of each dependent variable was assessed within each group and time point using the Shapiro-Wilk test, supplemented by visual inspection of histograms and normal probability plots (Q-Q plots). The homogeneity of variances was assessed at each time point for all 3 outcomes using the Levene test. All contrasts were nonsignificant (P>.05). Given that all participants were assessed at both pre-intervention and post-intervention time points, data were analyzed using a 2 × 2 mixed-design ANOVA (also termed split-plot ANOVA), with group as the between-participants factor and time (pre vs post) as the within-participants (repeated measures) factor. Effect sizes were quantified using 2 complementary indices. Partial eta squared (η_p^2) was computed for each source of variance in the mixed ANOVA model and interpreted according to the benchmarks: small (η_p^2≈0.01), medium (η_p^2≈0.06), and large (η_p^2 ≥0.14) [31]. Cohen d was additionally calculated for all planned post hoc comparisons as the standardized mean difference (ΔM/SDΔ for within-group paired comparisons; ΔM/SDpooled for between-group independent comparisons) and interpreted as small (d≈0.20), medium (d≈0.50), and large (d≥0.80) [32]. Given the exploratory nature of the pairwise comparisons and the priori enumerable set of 4 contrasts per outcome, post hoc analyses were conducted using paired-samples Student t tests for within-group (pre-to-post) comparisons and independent-samples Student t tests for between-group comparisons at each time point. To control the familywise type I error rate, all P values were adjusted using the Bonferroni correction (α'=.05/4=.0125 per comparison). The Bonferroni procedure was selected on the grounds of its conceptual transparency, conservative control of familywise error, and suitability for a small, prespecified set of contrasts [33]. Statistical analyses were performed using IBM SPSS Statistics for Windows (version 26; IBM Corp.).

Ethical Considerations

This study was approved by the Research Ethics Committee of the Catholic University of Temuco (ID: 091302/22) and prospectively registered at OSF.io [34]. The study was conducted in accordance with the ethical principles of the Declaration of Helsinki. All participants were informed of the study procedures, the voluntary nature of participation, and their right to withdraw at any time without penalty. Written informed consent was obtained from all participants prior to enrollment. Privacy and confidentiality were maintained throughout the study. The participants in this study did not receive any form of compensation for their participation.

Results

Baseline Characteristics of Physical Therapy Students

The participants' timeline can be seen in the flow diagram (Figure 2). A total of 31 students agreed to participate in this research. After randomization, 16 students were assigned to the individual group and 15 to the collaborative group. Eight students dropped out of the study: 1 due to vacation and 7 due to seasonal work. Therefore, 11 students in the individual group and 12 in the collaborative group completed the study. The characteristics of the participants are shown in Table 1. None of the students had previous experience in simulation-based learning with standardized patients. No differences were observed in physical therapy student characteristics at the beginning of the study. Furthermore, no differences were found between groups at baseline for clinical competence (P=.42), effective communication (P=.16), and therapeutic alliance (P=.44).

XSL•FO
RenderX

Figure 2. CONSORT (Consolidated Standards of Reporting Trials) 2010 flowchart.

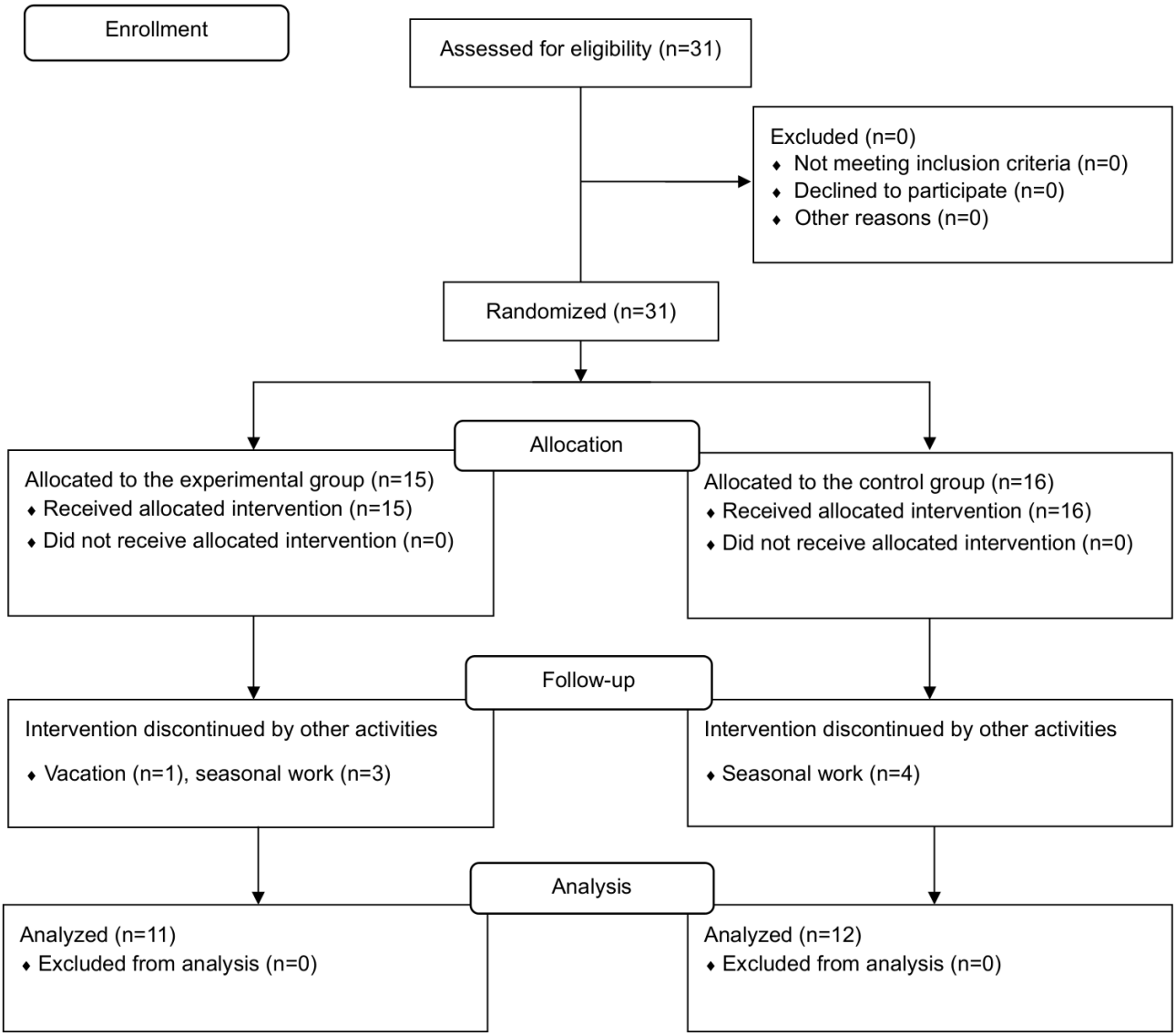


Table . Physical therapy students' characteristics[a].

Variable	Collaborative (n=12)	Individual (n=11)	Test statistic (*df*)	*P* value
Demographics				
Age (y), mean (SD)	22.6 (1.4)	23.4 (2.7)	−0.886[b] (21)	.39
Female sex, n (%)	9 (75.0)	8 (72.7)	0.02[c] (1)	.90
Outcome variables at baseline, mean (SD)				
MTCCP[d] total score	25.14 (4.33)	26.61 (4.15)	−0.827[b] (21)	.42
CAT[e] mean score	4.30 (0.39)	4.53 (0.36)	−1.470[b] (21)	.16
WAI-O[f] total score	117.69 (6.26)	120.09 (8.30)	−0.786[b] (21)	.44

[a]Between-group comparisons were performed using independent-samples Student *t* tests for continuous variables and Pearson's χ^2 test for categorical variables.

[b]*t* test.

[c]Chi-square test.

[d]MTCCP: Measurement Tool for Clinical Competencies in Physiotherapy.

[e]CAT: Communication Assessment Tool.

[f]WAI-O: Working Alliance Inventory–Observer Version.

XSL•FO
RenderX

The inferential results from the mixed ANOVA, the detailed descriptive statistics for all outcomes, and their corresponding interaction trajectories are summarized in Tables 2 and 3 and illustrated in Figure 3. Additionally, the complete results for the Bonferroni-corrected post hoc pairwise comparisons are provided in Table S4 (Multimedia Appendix 2).

Table . Results of the 2 × 2 mixed analysis of variance for Measurement Tool for Clinical Competencies in Physiotherapy (MTCCP), Communication Assessment Tool (CAT), and Working Alliance Inventory—Observer Version (WAI-O)[a].

Outcome and source of variance	*F* test (*df*)	*P* value	η_p^2
MTCCP			
Group	0.595 (1, 21)	.45	0.028
Time	59.47 (1, 21)	<.001	0.739
Group × time	0.264 (1, 21)	.61	0.012
CAT			
Group	3.906 (1, 21)	.06	0.157
Time	0.012 (1, 21)	.91	0.001
Group × time	0.046 (1, 21)	.83	0.002
WAI-O			
Group	0.266 (1, 21)	.61	0.013
Time	8.744 (1, 21)	.008	0.294
Group × time	0.546 (1, 21)	.47	0.025

[a]A mixed-design ANOVA was conducted with group (collaborative vs individual) as the between-participants factor and time (pre vs post) as the within-participants factor. The group × time interaction constitutes the primary inferential test of differential efficacy. η_p^2 = partial eta squared (Cohen, 1988 benchmarks: small≈0.01, medium≈0.06, large ≥0.14).

Table . Descriptive statistics for primary outcome variables by group and time point.

Outcome and group	Preintervention, mean (SD)	Postintervention, mean (SD)	Mean change (95% CI)	% Change
MTCCP[b]				
Collaborative	25.14 (4.33)	31.31 (3.73)	+6.17 (3.77 to 8.57)	+24.5
Individual	26.61 (4.15)	32.00 (2.78)	+5.39 (3.38 to 7.40)	+20.3
CAT[c]				
Collaborative	4.30 (0.39)	4.34 (0.39)	+0.04 (−0.35 to 0.43)	+0.8
Individual	4.53 (0.36)	4.52 (0.33)	−0.01 (−0.27 to 0.25)	−0.3
WAI-O[d]				
Collaborative	117.69 (6.26)	114.11 (6.49)	−3.58 (−9.40 to 2.24)	−3.0
Individual	120.09 (8.30)	114.15 (6.10)	−5.94 (−11.12 to −0.76)	−4.9

[a]Mean change = Post − Pre. 95% CI = 95% confidence interval of the mean change (paired). % Change = (Post − Pre)/Pre × 100.

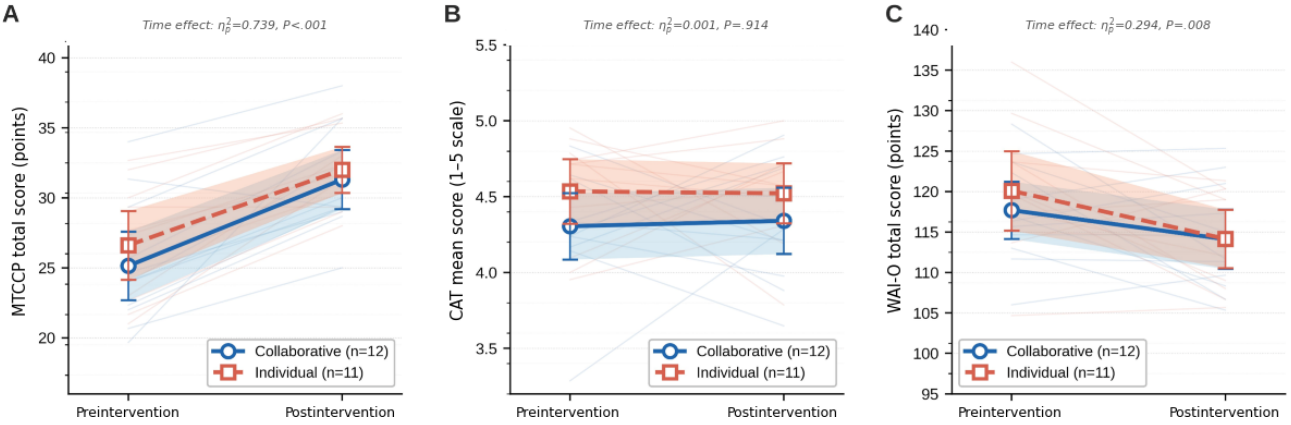
[b]MTCCP: Measurement Tool for Clinical Competencies in Physiotherapy.

[c]CAT: Communication Assessment Tool.

[d]WAI-O: Working Alliance Inventory–Observer Version.

XSL•FO
RenderX

Figure 3. Group × time interaction plots for Measurement Tool for Clinical Competencies in Physiotherapy (MTCCP), Communication Assessment Tool (CAT), and Working Alliance Inventory–Observer Version (WAI-O). Estimated marginal means + 95% CI; individual trajectories shown in translucent lines. (A) Results for clinical competence (MTCCP), (B) results for communication (CAT), and (C) results for therapeutic alliance (WAI-O).



Clinical Competence

The mixed ANOVA revealed a large main effect of time on MTCCP scores ($F_{1,21}$=59.47, η_p^2=0.739; P<.001), demonstrating substantial pre-to-post improvement in clinical competencies across both groups. The main effect of group was nonsignificant ($F_{1,21}$=0.595, η_p^2=0.028; P=.45), and the group × time interaction did not reach statistical significance ($F_{1,21}$=0.264, η_p^2=0.012; P=.61). However, the direction and magnitude of improvement were indicative of descriptive parity between conditions. Bonferroni-corrected post hoc analyses confirmed significant pre-to-post gains in both the collaborative group (t_{11}=−5.231, d=-1.510; P<.001 [corrected]) and the individual group (t_{10}=−5.979, d=1.803; P<.001 [corrected]). Between-group comparisons at preintervention (d=−0.346) and postintervention (d=−0.211) were small to moderate but nonsignificant after Bonferroni correction (both P_{Bonf}=1.00), reflecting a high degree of uncertainty due to the limited sample size.

Communication

For CAT scores, the mixed ANOVA yielded a negligible main effect of time ($F_{1,21}$=0.012, η_p^2=0.001; P=.91), and the group × time interaction was nonsignificant ($F_{1,21}$=0.046, η_p^2=0.002; P=.83). However, the main effect of group was characterized by a large effect size ($F_{1,21}$=3.906, η_p^2=0.157; P=.06), despite missing the conventional threshold for statistical significance. Descriptively, the individual group scored marginally higher. Given that the large effect size (η_p^2=0.157) alongside the nonsignificant P value, this trend strongly suggests a meaningful descriptive difference between the groups that cannot be conclusively confirmed or ruled out due to statistical underpowering.

Therapeutic Alliance

The mixed ANOVA indicated a moderate-to-large main effect of time for WAI-O scores ($F_{1,21}$=8.744, η_p^2=0.294; P=.008), reflecting a decrease in observer-rated working alliance from pre- to post-intervention across both conditions. The main effect of group ($F_{1,21}$=0.266, η_p^2=0.013; P=.61) and the group × time interaction ($F_{1,21}$=0.546, η_p^2=0.025; P=.47) were not statistically significant. Post hoc analyses revealed that the pre-to-post decline did not survive Bonferroni correction in either the collaborative group (t_{11}=1.637, d=−0.472; P_{Bonf}=.52) or the individual group (t_{10}=2.558, d=−0.771; P_{Bonf}=.11). Nevertheless, the moderate and large individual effect sizes reveal a descriptive reduction in alliance scores across both conditions, suggesting that true underlying changes may be obscured by statistical underpowering.

Discussion

Principal Findings

This study aimed to determine the effectiveness of collaborative versus individual simulation-based learning in developing clinical competence, effective communication, and therapeutic alliance among undergraduate physical therapy students during the care of standardized patients with CLBP flare-ups. Within the constraints of our sample size, our findings suggest that both simulation-based learning modalities led to within-group increases in clinical reasoning and professional behaviors, although no statistically significant differences were detected between the 2 modalities in clinical competence. However, observer-rated therapeutic alliance demonstrated a significant main effect of time, representing an overall decline across the intervention in both groups, while effective communication scores did not show statistical differences either between the groups or over time.

To the best of our knowledge, this study represents the first randomized experimental trial directly contrasting student-led collaborative scenario design against investigator-led individual simulation within physical therapy entry-to-practice curricula. The primary novelty of our work resides not merely in the logistical or financial advantages of collaborative simulation-based learning, but in providing preliminary empirical insight into a peer-assisted learning model for managing complex, psychosocially driven clinical conditions. Nevertheless, given the limited statistical power and the high uncertainty surrounding the resulting effect sizes, these findings should be considered exploratory rather than a definitive validation of the model's magnitude of effect.

XSL•FO
RenderX

Several studies have corroborated the effectiveness of simulation-based learning for the development of physical therapy competencies involving practitioner skills, empathy, motivation, and attitudes [35-37]. While collaborative simulation-based learning has been applied in cardiology emergency scenarios [16], where the perceived usefulness among students and teachers was very high, its comparative effectiveness has not yet been established. Our results suggest that collaborative simulation-based learning may serve as an effective and comparable alternative to individual simulation-based learning with standardized patients. Collaborative simulation-based learning introduces a paradigm shift rooted in social constructivism and cognitive load theory [17,38,39]. By placing students at the center of the instructional design process, requiring them to collaboratively analyze the literature, cocreate clinical scripts, and enact the patient role, the cognitive load associated with complex clinical decision-making is functionally distributed [17,18]. To successfully construct an authentic CLBP flare-up scenario, student dyads must actively negotiate and internalize what a patient experiencing centralized pain, catastrophizing, and fear-avoidance actually undergoes [5,6]. This metacognitive phase prior to execution could explain why the collaborative group achieved clinical competence scores similar to those of the individual group.

Regarding effective communication and the therapeutic alliance, both variables showed no significant differences between groups at the end of the study. While communication remained relatively stable over time, surprisingly, the therapeutic alliance decreased significantly between the pre- and post-intervention periods. Rather than reflecting a failure in teaching, this systematic reduction in the quality of the therapist-patient bond reflects a well-documented phenomenon in medical education, the cognitive rigidity of the novice physical therapist when faced with clinical complexity [40-42]. A flare-up of CLBP generates intense emotional distress and frustration in patients, which in turn complicates the relational context during the clinical encounter [5,6]. High-fidelity and complex standardized patient simulations can potentially induce significant anxiety and negative emotions in undergraduate students. These emotional responses may consume working-memory resources by increasing extraneous cognitive load, potentially hindering learning if total cognitive load exceeds working memory capacity [43,44]. Consequently, to preserve performance, students frequently retreat into a defensive, more technical, and highly prescriptive biomedical mode, focusing strictly on the physical execution of manual techniques or exercise parameters while neglecting affective validation, active listening, and the encouragement of positive expectations [45,46]. Given that an enhanced therapeutic alliance and effective communication are contextual factors capable of significantly reducing pain intensity [47], this procedural limitation represents a critical clinical error.

As not all students can participate in the clinical experience of meeting and interacting with this spectrum of patients, simulating a high-fidelity scenario provides an opportunity to practice and achieve clinical competencies in a safe learning environment. An example of a complex scenario is the management of a flare-up in patients with CLBP, as it impacts major elements of quality of life and involves multidimensional factors, such as a sense of disablement and shifts in mood or emotions, that the student must manage [5,6,48]. Active therapy with a patient-centered approach can modulate the patient's outcome through communication strategies such as active listening, empathy, building positive expectations, and providing reassuring information [45]. Therefore, promoting a therapeutic alliance and effective communication, while taking into account contextual factors, can be implemented as a key clinical education strategy within the university curriculum of health care professionals [45]. Through either collaborative or individual simulation-based learning, real-life experiences are represented in a guided and interactive manner; however, a collaborative approach may reduce overall costs of simulation, specifically with script development and the cost of actors [49].

Recommendations for Education

To integrate collaborative simulation-based learning into physical therapy or other health sciences curricula, several guidelines are recommended. First, it is necessary to identify the course, learning outcomes, and competencies to be developed to determine how many collaborative SBEs will be integrated. Second, the instructor must provide reading material that serves as the baseline input for students to create their clinical cases and scenarios. Third, students should be informed in advance of the simulation modality and their specific responsibilities. Fourth, during the collaborative simulation-based learning design phase, the instructor must facilitate the creative process by resolving doubts, defining roles, and reviewing the clinical coherence of the scenario and case against the intended learning outcome (eg, the general characteristics of the patient, presence of a companion, signs, symptoms, psychosocial factors, equipment, and spatial layout). Fifth, the execution of the scenarios is the sole responsibility of the designated student team, requiring members to fulfill the roles of patient, companion, timekeeper, and performance evaluator. Sixth, the debriefing phase can be adapted according to the methodology that best suits the instructor guiding the process.

From an educational curriculum standpoint, these preliminary findings suggest that collaborative simulation-based learning should not be restricted to an introductory method or a contingency plan for low-resource settings. Instead, it should be strategically integrated as a scaffolded instructional methodology. We recommend implementing collaborative peer-designed simulation during early entry-to-practice phases to build solid clinical reasoning and communicative baselines in psychologically safe, peer-supported environments. Once these distributed cognitive frameworks are consolidated, students can progressively transition toward individual encounters with professional standardized patients, culminating in immersion with real, complex clinical populations.

Strengths and Limitations

To the best of our knowledge, this is the first comparative experimental study directly evaluating the effectiveness of collaborative versus individual simulation-based learning. Following a previously defined and easily reproducible protocol,

XSL•FO
RenderX

we trained standardized patients for both the evaluation phases and the simulation-based learning interventions. The assessment sessions were video recorded to prevent students from experiencing heightened stress due to the physical presence of an in-room evaluator. Finally, the evaluation incorporated the dual perspectives of the standardized patient and a blinded external evaluator. However, our findings should be interpreted with caution due to several limitations. Although effective communication and the therapeutic alliance were evaluated, these relational skills were not explicitly included as formal content or core competencies to be developed during SBEs. Furthermore, while we performed a priori sample size calculation based on the existing literature, our final sample size limited the statistical power of the study. In underpowered educational research, these constraints can lead to overinflated effect sizes, a phenomenon that introduces a large amount of uncertainty regarding the true magnitude of the intervention's impact. Consequently, the unusually high effect sizes observed in our results must be interpreted with caution, as they may restrict the generalizability of the findings to broader physical therapy curricula. Future studies should incorporate larger and more diverse sample sizes, incorporate explicit post hoc power analyses, assessments of student stress and anxiety, and cost-effectiveness analyses, alongside qualitative evaluations of the collaborative simulation experience. Because this study included only physical therapy students, further research is warranted across other health professions, including interprofessional approaches.

Conclusions

In a complex clinical management context, both collaborative and individual simulation-based learning led to increases in physical therapy clinical competencies, with no statistically significant differences detected between the 2 instructional modalities. However, there were no differences between the groups for effective communication or therapeutic alliance. Given the high magnitude of improvement observed across both modalities, detecting a statistically significant differential effect between them would require larger sample sizes. Further research is warranted to assess the efficacy of these modalities in diverse clinical settings.

Acknowledgments

During the preparation of this manuscript, the authors used Gemini to improve the clarity, grammar, and formatting of the text. After using the tool, the authors reviewed and edited the content as needed and maintain full responsibility for the accuracy and integrity of the final publication.

Funding

This research was funded by the teaching innovation fund of the Catholic University of Temuco (PID411-5894).

Data Availability

In accordance with regulations and assuming permission by the Research Ethics Board, aggregated and anonymized data of students used in this study may be available upon request from the corresponding author.

Authors' Contributions

Conceptualization: IC-V, HA-P
Data curation: IC-V, HA-P, FA-Q
Formal analysis: IC-V, HA-P, FA-Q, HG-E
Investigation: IC-V, HA-P, SG-M, RN-C
Project administration: IC-V, HA-P
Visualization: IC-V, HA-P
Writing – original draft: IC-V, HA-P
Writing – review and editing: IC-V, HA-P, FA-Q, HG-E, SG-M, RN-C
All authors approved the final version to be published

Conflicts of Interest

SG-M serves on the JMIR Publications Editorial Board as an associate editor of JMIR Medical Education. However, he was not involved in the review process for this manuscript.

Multimedia Appendix 1
Individual and collaborative simulation-based learning protocols.
[DOCX File, 16 KB - mededu_v12i1e85442_app1.docx]

Multimedia Appendix 2
Additional tables included in the supplementary material.
[DOCX File, 27 KB - mededu_v12i1e85442_app2.docx]

XSL•FO
RenderX

Checklist 1
CONSORT checklist.
[PDF File, 132 KB - mededu_v12i1e85442_app3.pdf]

References

1. Epstein RM, Hundert EM. Defining and assessing professional competence. JAMA 2002 Jan 9;287(2):226-235. [doi: 10.1001/jama.287.2.226] [Medline: 11779266]
2. Vervaeke R, Lafrance S, Demont A. Core competencies for first contact physiotherapists in a direct access model of care for adults with musculoskeletal disorders: a scoping review. Musculoskeletal Care 2023 Dec;21(4):1353-1363. [doi: 10.1002/msc.1813] [Medline: 37641476]
3. Jenkins HJ, Brown BT, O'Keeffe M, Moloney N, Maher CG, Hancock M. Development of low back pain curriculum content standards for entry-level clinical training. BMC Med Educ 2024 Feb 12;24(1):136. [doi: 10.1186/s12909-024-05086-x] [Medline: 38347486]
4. Hoy D, March L, Brooks P, et al. The global burden of low back pain: estimates from the Global Burden of Disease 2010 study. Ann Rheum Dis 2014 Jun;73(6):968-974. [doi: 10.1136/annrheumdis-2013-204428] [Medline: 24665116]
5. Costa N, Ferreira ML, Setchell J, et al. A definition of "flare" in low back pain: a multiphase process involving perspectives of individuals with low back pain and expert consensus. J Pain 2019 Nov;20(11):1267-1275. [doi: 10.1016/j.jpain.2019.03.009] [Medline: 30904517]
6. Tan D, Hodges PW, Costa N, Ferreira M, Setchell J. Impact of flare-ups on the lives of individuals with low back pain: a qualitative investigation. Musculoskelet Sci Pract 2019 Oct;43:52-57. [doi: 10.1016/j.msksp.2019.06.003] [Medline: 31260931]
7. Cook DA, Hatala R, Brydges R, et al. Technology-enhanced simulation for health professions education: a systematic review and meta-analysis. JAMA 2011 Sep 7;306(9):978-988. [doi: 10.1001/jama.2011.1234] [Medline: 21900138]
8. Lamé G, Dixon-Woods M. Using clinical simulation to study how to improve quality and safety in healthcare. BMJ Simul Technol Enhanc Learn 2020 Mar;6(2):87-94. [doi: 10.1136/bmjstel-2018-000370] [Medline: 32133154]
9. Antila AK, Lindblom S, Louhiala P, Pyörälä E. Creating a safe space: medical students' perspectives on using actor simulations for learning communication skills. BMC Med Educ 2024 Oct 28;24(1):1225. [doi: 10.1186/s12909-024-06184-6] [Medline: 39468604]
10. Sarfati L, Ranchon F, Vantard N, et al. Human-simulation-based learning to prevent medication error: a systematic review. J Eval Clin Pract 2019 Feb;25(1):11-20. [doi: 10.1111/jep.12883] [Medline: 29383867]
11. Higham H. Simulation past, present and future—a decade of progress in simulation-based education in the UK. BMJ Simul Technol Enhanc Learn 2021;7(5):404-409. [doi: 10.1136/bmjstel-2020-000601] [Medline: 35515719]
12. Pritchard SA, Blackstock FC, Nestel D, Keating JL. Simulated patients in physical therapy education: systematic review and meta-analysis. Phys Ther 2016 Sep;96(9):1342-1353. [doi: 10.2522/ptj.20150500] [Medline: 26939603]
13. Maloney S, Haines T. Issues of cost-benefit and cost-effectiveness for simulation in health professions education. Adv Simul (Lond) 2016;1:13. [doi: 10.1186/s41077-016-0020-3] [Medline: 29449982]
14. Nunnink L, Thompson A, Alsaba N, Brazil V. Peer-assisted learning in simulation-based medical education: a mixed-methods exploratory study. BMJ Simul Technol Enhanc Learn 2021;7(5):366-371. [doi: 10.1136/bmjstel-2020-000645] [Medline: 35515740]
15. Sevenhuysen S, Farlie MK, Keating JL, Haines TP, Molloy E. Physiotherapy students and clinical educators perceive several ways in which incorporating peer-assisted learning could improve clinical placements: a qualitative study. J Physiother 2015 Apr;61(2):87-92. [doi: 10.1016/j.jphys.2015.02.015] [Medline: 25801365]
16. Guinez-Molinos S, Gomar-Sancho C. Collaborative clinical simulation in cardiologic emergency scenarios for medical students. An exploratory study on model applicability and assessment instruments. GMS J Med Educ 2021;38(4):Doc76. [doi: 10.3205/zma001472] [Medline: 34056065]
17. Guinez-Molinos S, Martínez-Molina A, Gomar-Sancho C, et al. A collaborative clinical simulation model for the development of competencies by medical students. Med Teach 2017 Feb;39(2):195-202. [doi: 10.1080/0142159X.2016.1248913] [Medline: 27841066]
18. Guinez-Molinos S, Gonzalez Díaz J, Gomar Sancho C, Espinoza P, Constenla G. A web platform (MOSAICO) to design, perform, and assess collaborative clinical scenarios for medical students: viewpoint. JMIR Med Educ 2021 Jan 26;7(1):e23370. [doi: 10.2196/23370] [Medline: 33496676]
19. Mori B, Carnahan H, Herold J. Use of simulation learning experiences in physical therapy entry-to-practice curricula: a systematic review. Physiother Can 2015;67(2):194-202. [doi: 10.3138/ptc.2014-40E] [Medline: 25931672]
20. Cheng A, Kessler D, Mackinnon R, et al. Reporting guidelines for health care simulation research: extensions to the CONSORT and STROBE statements. Simul Healthc 2016 Aug;11(4):238-248. [doi: 10.1097/SIH.0000000000000150] [Medline: 27465839]
21. Al Asmri M, Haque MS, Parle J. A modified Medical Education Research Study Quality Instrument (MMERSQI) developed by Delphi consensus. BMC Med Educ 2023 Jan 25;23(1):63. [doi: 10.1186/s12909-023-04033-6] [Medline: 36698117]
22. RANDOM.ORG. URL: https://www.random.org/ [accessed 2026-09-09]

XSL•FO
RenderX

23. Abulebda K, Auerbach M, Limaiem F. Debriefing techniques utilized in medical simulation. In: StatPearls [Internet]: StatPearls Publishing; 2024. URL: https://www.ncbi.nlm.nih.gov/books/NBK546660/ [accessed 2026-09-09]
24. Makoul G, Krupat E, Chang CH. Measuring patient views of physician communication skills: development and testing of the Communication Assessment Tool. Patient Educ Couns 2007 Aug;67(3):333-342. [doi: 10.1016/j.pec.2007.05.005] [Medline: 17574367]
25. Armijo-Rivera S, Behrens CC, Giaconi ME, et al. Validación de la versión en español de un instrumento de evaluación de la comunicación centrada en el paciente en OSCE. Educación Médica 2021 Jul;22(4):193-198. [doi: 10.1016/j.edumed.2020.12.007]
26. Waizmann V, Roussos A. Adaptación de Inventario de Alianza de Trabajo en su Versión Observador: WAI-O-A [Article in Spanish]. Anuario de investigaciones 2011;18:95-104 [FREE Full text]
27. Torres-Narváez MR, Vargas-Pinilla OC, Rodríguez-Grande EI. Validity and reproducibility of a tool for assessing clinical competencies in physical therapy students. BMC Med Educ 2018 Nov 23;18(1):280. [doi: 10.1186/s12909-018-1377-x] [Medline: 30470221]
28. Fenton LR, Cecero JJ, Nich C, Frankforter TL, Carroll KM. Perspective is everything: the predictive validity of six working alliance instruments. J Psychother Pract Res 2001;10(4):262-268. [Medline: 11696653]
29. Tichenor V, Hill CE. A comparison of six measures of working alliance. Psychother Theory Res Pract Train 1989;26(2):195-199. [doi: 10.1037/h0085419]
30. Forbes R, Mandrusiak A, Smith M, Russell T. Training physiotherapy students to educate patients: a randomised controlled trial. Patient Educ Couns 2018 Feb;101(2):295-303. [doi: 10.1016/j.pec.2017.08.009] [Medline: 28847438]
31. Lakens D. Calculating and reporting effect sizes to facilitate cumulative science: a practical primer for t-tests and ANOVAs. Front Psychol 2013 Nov 26;4:863. [doi: 10.3389/fpsyg.2013.00863] [Medline: 24324449]
32. Fritz CO, Morris PE, Richler JJ. Effect size estimates: current use, calculations, and interpretation. J Exp Psychol Gen 2012 Feb;141(1):2-18. [doi: 10.1037/a0024338] [Medline: 21823805]
33. Shi Q, Pavey ES, Carter RE. Bonferroni-based correction factor for multiple, correlated endpoints. Pharm Stat 2012;11(4):300-309. [doi: 10.1002/pst.1514] [Medline: 22588983]
34. Effectiveness of collaborative and traditional clinical simulation to develop clinical competencies in physical therapy students. Open Science Framework. URL: https://osf.io/yfwht/overview [accessed 2026-09-09]
35. Lin IH, Wang CY, Lin YN, Chen HC, Lin LF. Simulation-based holistic education in physiotherapy interns to increase empathy toward older adults and individuals with disabilities. BMC Geriatr 2022 Oct 12;22(1):795. [doi: 10.1186/s12877-022-03500-x] [Medline: 36224529]
36. Sandoval-Cuellar C, Alfonso-Mora ML, Castellanos-Garrido AL, et al. Simulation in physiotherapy students for clinical decisions during interaction with people with low back pain: randomised controlled trial. BMC Med Educ 2021 Jul 9;21(1):375. [doi: 10.1186/s12909-021-02812-7] [Medline: 34243767]
37. Wright A, Moss P, Dennis DM, et al. The influence of a full-time, immersive simulation-based clinical placement on physiotherapy student confidence during the transition to clinical practice. Adv Simul (Lond) 2018;3(1):3. [doi: 10.1186/s41077-018-0062-9] [Medline: 29484204]
38. Alnas S, Imran M. Effects of peer tutoring on academic achievement: a social constructivist perspective. Res J Societal Issues 2025 Sep;7(2):28-37. [doi: 10.56976/rjsi.v7i2.357]
39. Reedy GB. Using cognitive load theory to inform simulation design and practice. Clin Simul Nurs 2015 Aug;11(8):355-360. [doi: 10.1016/j.ecns.2015.05.004]
40. Williams L, Trebacz A, Atkinson M, Tiplady C. Exploring the barriers to clinical reasoning in high-fidelity simulation for undergraduate medical students: a phenomenological study. Simul Healthc 2026 Jun 1;21(3):143-150. [doi: 10.1097/SIH.0000000000000886] [Medline: 41020780]
41. Gabrys RL, Howell JW, Cebulski SF, Anisman H, Matheson K. Acute stressor effects on cognitive flexibility: mediating role of stressor appraisals and cortisol. Stress 2019 Mar;22(2):182-189. [doi: 10.1080/10253890.2018.1494152] [Medline: 30727804]
42. Kucukkelepce DS, Marasli İ, Mir SK. The relationship between cognitive flexibility and decision-making styles of midwifery and nursing students. BMC Med Educ 2026 Feb 18;26(1):410. [doi: 10.1186/s12909-026-08734-6] [Medline: 41709245]
43. Rogers BA, Franklin AE. Cognitive load experienced by nurses in simulation-based learning experiences: an integrative review. Nurse Educ Today 2021 Apr;99:104815. [doi: 10.1016/j.nedt.2021.104815] [Medline: 33640776]
44. Rogers T. Human information processing under stress. J Emerg Manag 2023;21(2):141-154. [doi: 10.5055/jem.0756] [Medline: 37270413]
45. Rossettini G, Carlino E, Testa M. Clinical relevance of contextual factors as triggers of placebo and nocebo effects in musculoskeletal pain. BMC Musculoskelet Disord 2018 Jan 22;19(1):27. [doi: 10.1186/s12891-018-1943-8] [Medline: 29357856]
46. Cuyul-Vásquez I, Barría JA, Perez NF, Fuentes J. The influence of verbal suggestions in the management of musculoskeletal pain: a narrative review. Phys Ther Rev 2019 Jul 4;24(3-4):175-181. [doi: 10.1080/10833196.2019.1639011]

XSL•FO
RenderX

47. Fuentes J, Armijo-Olivo S, Funabashi M, et al. Enhanced therapeutic alliance modulates pain intensity and muscle pain sensitivity in patients with chronic low back pain: an experimental controlled study. Phys Ther 2014 Apr;94(4):477-489. [doi: 10.2522/ptj.20130118] [Medline: 24309616]
48. Setchell J, Costa N, Ferreira M, Makovey J, Nielsen M, Hodges PW. Individuals' explanations for their persistent or recurrent low back pain: a cross-sectional survey. BMC Musculoskelet Disord 2017 Nov 17;18(1):466. [doi: 10.1186/s12891-017-1831-7] [Medline: 29149847]
49. Lateef F. Simulation-based learning: just like the real thing. J Emerg Trauma Shock 2010 Oct;3(4):348-352. [doi: 10.4103/0974-2700.70743] [Medline: 21063557]

Abbreviations

CAT: Communication Assessment Tool
CLBP: chronic low back pain
CONSORT: Consolidated Standards of Reporting Trials
ICC: intraclass correlation coefficient
MTCCP: Measurement Tool for Clinical Competencies in Physiotherapy
SBE: simulation-based experience
WAI-O: Working Alliance Inventory–Observer Version

Edited by A Stone; submitted 07.Oct.2025; peer-reviewed by B Wolden, K Schertzer; revised version received 14.Jul.2026; accepted 15.Jul.2026; published 21.Sep.2026.

Please cite as:
Cuyul-Vásquez I, Aravena-Páez H, Araya-Quintanilla F, Guinez-Molinos S, Núñez-Cortés R, Gutiérrez Espinoza H
Collaborative Simulation-Based Learning to Develop Competencies in Physical Therapy Students: Randomized Controlled Trial
JMIR Med Educ 2026;12:e85442
URL: https://mededu.jmir.org/2026/1/e85442
doi:10.2196/85442

© Iván Cuyul-Vásquez, Hilda Aravena-Páez, Felipe Araya-Quintanilla, Sergio Guinez-Molinos, Rodrigo Núñez-Cortés, Héctor Gutiérrez Espinoza. Originally published in JMIR Medical Education (https://mededu.jmir.org), 21.Sep.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Education, is properly cited. The complete bibliographic information, a link to the original publication on https://mededu.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Prescription Support Practice for Pharmacy Students: Pre-Post Educational Intervention Study

Fuka Aizawa[1,2*], PhD; Kenta Yagi[2,3], PhD; Tsukasa Higashionna[4], MS; Hirofumi Hamano[4], PhD; Shimon Takahashi[2,5], PhD; Yoshito Zamami[4], Prof Dr; Kazuaki Shinomiya[6], Prof Dr; Takahiro Niimura[1,2], PhD; Mitsuhiro Goda[2,7], Prof Dr; Kei Kawada[2,5], PhD; Keisuke Ishizawa[1,2,5*], Prof Dr

[1]Clinical Research Center for Developmental Therapeutics, Tokushima University Hospital, Tokushima, Japan

[2]Department of Clinical Pharmacology and Therapeutics, Graduate School of Biomedical Sciences, Tokushima University, 3-18-15 Kuramoto-cho, Tokushima, Japan

[3]Department of Pharmacy, Shimane University Hospital, Shimane, Japan

[4]Department of Pharmacy, Okayama University Hospital, Okayama, Japan

[5]Department of Pharmacy, Tokushima University Hospital, Tokushima, Japan

[6]Department of Pharmaceutical Care and Clinical Pharmacy, Tokushima Bunri University, Tokushima, Japan

[7]Department of Clinical Pharmacology and Therapeutics, Graduate School of Biomedical and Heath Sciences, Hiroshima University, Hiroshima, Japan

[*]these authors contributed equally

Corresponding Author:
Keisuke Ishizawa, Prof Dr
Clinical Research Center for Developmental Therapeutics, Tokushima University Hospital, Tokushima, Japan

Abstract

Background: In the field of team-based care, pharmacists are vital for optimizing medication therapy. However, many medical professionals lack the opportunity to learn how to propose prescription changes with precision.

Objective: This study aimed to address this knowledge gap by developing and assessing a new educational program for pharmacy students focused on prescription support and interprofessional collaboration.

Methods: We recruited 191 fifth-year pharmaceutical students during the 2022 - 2024 academic years. The program featured a 7-day intensive curriculum that included learning how to assist with prescriptions, analyzing clinical data, and engaging in role-playing exercises. A web-based questionnaire and a paper test were used to evaluate students' awareness and knowledge both before and after the program. Statistical analyses were performed to verify the significance of changes; we utilized the Wilcoxon signed-rank test for the ordinal data derived from the specific behavioral objectives and 2-tailed paired *t* tests for the interval data from the knowledge tests. The magnitude of change was quantified using *r* for Wilcoxon tests and Cohen dz for 2-tailed *t* tests, with 95% CI calculated to ensure the stability and reliability of the observed results.

Results: Analysis of the primary outcome specific behavioral objectives revealed statistically significant effects across all items (Wilcoxon signed-rank test; $P<.001$). Effect sizes ($r=0.505$ - 0.835) ranged from moderate to large, with particularly large effects observed in identifying contents issue ($r=0.835$, 95% CI 0.126-0.330; $P<.001$). Knowledge test scores showed significant improvement in the following 3 subjects: pharmacology ($r=-0.504$, 95% CI –0.215 to 0.127; $P<.001$), organic chemistry ($r=0.254$, 95% CI –0.148 to –0.193; $P=.004$), and communication ($r=0.221$, 95% CI –0.151 to –0.190; $P=.01$). No significant changes were observed in pathology or pharmacokinetics.

Conclusions: This program provides strong evidence that practical, hands-on learning with hospital pharmacists helps improve pharmacy students' professional skills and optimize pharmaceutical therapies in interprofessional care. By teaching pharmacists to effectively propose prescription changes, the program equips them to become integral members of interprofessional care, ultimately leading to optimized pharmaceutical care for patients.

(JMIR Med Educ 2026;12:e79545) doi:10.2196/79545

KEYWORDS

academic detailing; pharmaceutical clinical practice; prescription support; professional education; Interprofessional care

XSL•FO
RenderX

Introduction

Collaborative care, provided by a team of medical professionals, can enhance therapeutic effects [1,2]. Interprofessional care—in which each medical professional contributes their expertise—optimizes medical services that are becoming more specialized and advanced. However, multidisciplinary cooperation requires an understanding of the role of each medical professional while maintaining individual expertise and using it effectively to provide better medical care. In interprofessional care, pharmacists are specialists in medicine. To optimize physicians' prescriptions, pharmacists use their knowledge and skills—in dispensing medicines, formulating products, and managing supply—based on drug information [3-6]. Pharmacists can offer suggestions to ensure patient safety by promoting appropriate drug use and reducing adverse drug reactions [7]. Physicians view pharmacists' abilities positively in terms of patient outcomes and time spent in practice [8]. The decrease in side effects and improvement in pharmacotherapy owing to pharmacist-prescribed interventions are evident [9,10]. Additionally, improving the expertise of pharmacists can enable physicians to pay greater attention to patient care and appropriate pharmacotherapy. Furthermore, pharmacists can support other medical professionals, improving health care economics by optimizing prescriptions and decreasing medicine-related incidents.

Academic detailing is a strategy for collaboration and learning between physicians and pharmacists. It promotes evidence-based prescribing by physicians, with pharmacists serving as academic detailers who provide physicians with information on drug evidence. Initiatives in various countries cover a wide range of areas, including antibiotics and opioids [11-13]. These approaches by pharmacists have been shown to significantly influence physicians' prescribing behavior [14]. In Japan, pharmacists' duties have shifted in recent years from dispensing medication and other object-oriented tasks to patient-centered activities, such as medication counseling. Pharmacists are required to be part of specialized teams, such as antimicrobial stewardship teams and nutrition support teams, underscoring their input as members of the health care team. However, in Japan, while such activities as the pre-consultation clinical laboratory test and proxy prescription entry are practiced [15-17], implementation remains limited to a few facilities, with most practices involving postprescription verification tasks. Opportunities for physicians and pharmacists to collaborate on tasks within the clinical setting are scarce, indicating a significant barrier to routine.

During their college years, many medical staff undertake practical clinical training after obtaining basic knowledge and training in simulated patient care [18-20]. Practical training during college gives students the opportunity to apply their knowledge and practical experience in a clinical situation, enhance their expertise, learn a good bedside manner, and consider their life values. This experience enables students to provide appropriate care when they begin working in the field as medical professionals and further develop their skills within the context of the rules and regulations of the institution to which they are assigned. Many countries require 4 to 6 years to educate pharmaceutical students. In Japan, a 6-year pharmaceutical education system was implemented in 2006 to strengthen clinical pharmacy skills. The curriculum integrates foundational sciences with a mandatory 22-week clinical clerkship in hospitals and community pharmacies during the fifth year. More professional programs are structured in the United States and the United Kingdom, where pharmacist-led activities are more advanced. However, in many other countries, while professional learning is planned, the lack of advanced or practical programs and professional collaboration is perceived as a problem [21]. Pharmacists have comprehensive medical and pharmacological knowledge, which can be enhanced by specialized training. However, learning about interprofessional care and individually optimized treatment is not part of the standard training program. It is difficult for beginners to optimize their professional skills in busy clinical environments and use them to improve patient health care safety. In fact, concerns have been raised in some countries about the lack of awareness of the pharmaceutical profession and the continuing professional development for pharmacists. The systematic development of pharmacists' expertise and skills is necessary for them to become experienced and aware of processes when they begin clinical work. This clinical training focuses on reviewing the basic knowledge of pharmacists' procedures in community pharmacies and hospitals and practicing how to manage patients. Students work with tutors to formulate drug therapy plans by monitoring the treatment of patients in their charge, providing information on the occurrence of adverse drug reactions, and proposing an appropriate drug therapy plan for patients from a variety of drugs. Pharmaceutical risk-benefit, dosage, form, and cost-effectiveness must be considered when providing prescription support. It is also important to fully understand the physician's treatment plan and take steps to reflect the patient's wishes, contextual factors, and financial situation as well as the latest evidence. Students should be able to understand these complex steps in a limited period and consider what action they could take as pharmacists to proactively intervene in prescription design; however, there is no well-developed training program that focuses on these issues. In addition, pharmacists do not have learning-enabling programs, and the effective transfer of knowledge and skills is not taking place.

This study was designed to investigate the impact of an integrated learning module on the prescription recommendation process within the basic clinical practice curriculum in Japan. The specific aims were to determine whether this intervention would enhance students' awareness of interprofessional care and their knowledge and skills in pharmacotherapy planning. The outcome of the intervention was assessed through a pre- and postprogram evaluation using 2 questionnaires and a knowledge-based examination.

Methods

Study Design and Participants

This open-label, survey study was conducted at Tokushima University Hospital. In total, 191 pharmacy students (fifth-year students) who participated in clinical practice at Tokushima

XSL•FO
RenderX

University Hospital in the 2022 - 2023 academic year were enrolled in the program. The period of the study was a time of concern due to the coronavirus disease 2019 pandemic; however, none of the in-hospital regulations were in place, and all the procedures were conducted face-to-face. The students who desired clinical training at Tokushima University Hospital were recruited from 7 universities in Japan. They received 11 weeks of training in basic skills and patient care in the pharmacy before starting hospital training, in accordance with the training regulations. They were informed face-to-face about the study and given letters that obtained their consent to participate before the start of the practical training at the hospital. They were informed that all students would receive training as part of their clinical practice regardless of their participation in the study. Furthermore, refusal to provide consent would not impact their student performance in the clinical practice. Participants who provided informed consent were included in the analysis. We excluded those who did not provide informed consent, with missing data in the process of data collection, or whose lack of attendance made before-and-after comparisons impossible. All facilitators leading this program were pharmacists with doctoral degrees and at least 5 years of clinical experience. Furthermore, the lead facilitator held certification from the Japan Academic Detailing Research Association. This study defined a 5-point improvement in test scores as the effect size for the educational program. With an SD of 12 points, 80% power, and an α error of .05, the sample size was calculated. The study required 91 participants. Anticipating approximately 6 dropouts per year, the plan was to recruit 109 participants over 3 years (Checklist 1).

Procedures for Academic Detailing Program

All students joined an "academic detailing program" to learn the process of proposing prescriptions to physicians together with clinical practice in accordance with the guidelines for practical training of the Council on Pharmaceutical Education in Japan during the 11-week clinical practice period. No control group was established to ensure equal opportunities for learning within the same environment. The face-to-face program was designed to promote deeper thinking and increased cooperation with others. The students were randomly divided into small groups of 5 to 6, depending on the number of students accepted during the training period. Grouping involved researchers assigning numbers to participants and randomizing them to include both male and female participants. The program comprised three lectures over 5 days: (1) 2 days of face-to-face lectures and practice on how to review clinical papers and apply them to patients (1 pharmacist on staff); (2) half a day of face-to-face lectures on pharmacology, pathogenesis, chemistry, pharmacokinetics, and communication (30 minutes each, 4 pharmacists on staff); (3) 1 day of obtaining information and discussing the recommended drugs; and (4) preparation of materials for prescription proposals to physicians (or a staff pharmacist) and role-playing of prescription proposals among students. Two types of papers that had been accepted for submission were used to review the clinical papers [22,23]. During the paper review, the pharmacist gave lectures on reading papers, basic knowledge (diseases and drugs), guidelines, and other clinical studies related to the topic of the review. Each group read the papers and summarized the information on the worksheet. Then, the students discussed whether the evidence should be applied to the patient, including factors such as cost-effectiveness and patient willingness. The prescription proposal process focused on constipation. Students collected information on drugs used in Japan (excluding herbal medicines), including chemical structure, adaptation, dosage, administration, and contraindications. The students assessed the patients, discussed the drug therapy plans, and prepared handouts to assist physicians in prescribing the drugs. Based on the prescription proposal documents prepared by the program, students and pharmacists acting as physicians proposed prescriptions for 10 to 15 minutes and discussed therapeutic planning. After the session, the physicians decided whether or not to prescribe the prescriptions. Finally, the pharmacist who created the program provided feedback on improvements to the students' prescribing suggestions and shared examples of actual clinical experience.

Outcomes

The primary outcome measure was a change in specific behavioral objectives (SBOs) required in interprofessional care and prescribing support before and after clinical practice, with students requiring a score of at least 3 before undertaking the education program. The SBO change was collected using a web response, using a 5-point scale ranging from 1 (not possible) to 5 (possible). The questionnaire was developed by hospital pharmacists and was not related to teaching. It was based on common SBOs for pharmacy education, which were chosen from the Ministry of Education, Culture, Sports, Science and Technology, Japan [17]. We specifically selected SBOs that directly pertain to pharmacotherapy planning and prescription recommendation skills. This prioritization is based on the consensus that these competencies are the core clinical responsibilities of pharmacists in modern team–based medicine. By focusing on these high-priority items, we aimed to evaluate the practical and clinical readiness of students in scenarios that most closely mirror real-world pharmaceutical care. We selected 37 items related to our program from parts 1 to 10 (Table S1 in Multimedia Appendix 1). The original SBOs (Table S2 in Multimedia Appendix 1) comprised 11 items from part 11. The knowledge test included 30 questions, consisting of 6 questions each on chemistry, pharmacokinetics, pharmacology, pathology, and communication. To assess knowledge proficiency, 3 of the 6 questions were related to content not handled in the program but studied at university and in the clinical setting. Questionnaires and tests were administered at the beginning and end of the clinical practice. The names of the students were written on the questionnaires and tests at the time of collection for pre-post comparisons but were given serial numbers for anonymization in the analysis.

Ethical Considerations

This study was conducted in compliance with the Ethical Guidelines for Medical Research Involving Human Subjects and with the approval of the Ethics Committee of Tokushima University Hospital (approval number: 4193, certificate date: June 27, 2022). Informed consent was obtained from all individual participants included in the study. The authors affirm

XSL•FO
RenderX

that human research participants provided informed consent for publication of the images and all results. All data obtained and analyzed in the present study are included in this manuscript and not published elsewhere. All data were **de-identified** to ensure participant privacy. Personally identifiable information was replaced with **unique alphanumeric codes** (pseudonymization) to ensure that individuals cannot be identified from the dataset. Access to the key-code list was strictly restricted to the primary investigators. **No compensation**, monetary or otherwise, was provided to the participants for their involvement in this human subjects research. Participation was entirely voluntary.

Statistical Analysis

The primary outcome measure, the SBO questionnaire, is a 5-point ordinal scale. We performed a Wilcoxon signed-rank test for paired comparisons without assuming a specific distribution. The secondary outcome measure, the knowledge test, is an interval scale based on the number of correct answers. As the total score is out of 30 points, assuming normality, we used a 2-tailed paired *t* test. The subject-specific total scores are out of 6 points. Given the low number of scale levels and the high likelihood that normality assumptions would not hold, we analyzed these scores using the Wilcoxon signed-rank test. Similarly, the subject-specific total scores divided into related or unrelated categories are also out of 3 points; thus, we analyzed these scores using the Wilcoxon signed-rank test. The purpose of this study is to conduct an exploratory evaluation of the impact of the developed educational program. As this research design is not intended to verify intervention effects, adjustments for multiple comparisons were not necessary. For the pre-post difference in total knowledge test scores (out of 30 points), histograms and Q-Q plots were created to visually confirm normality. For all evaluation items, the standardized effect size and its 95% CI were calculated following the procedure set out below. For the SBO questionnaire using the Wilcoxon signed-rank test, the test's subject-specific total scores, and the related or unrelated subject-specific total scores, the effect size $r=Z/\mathrm{sqrt}(N)$ and its 95% CI were calculated. Here, Z is the standardized value of the Wilcoxon signed-rank test statistic (standardized test statistic), and N is the number of participants. The 95% CI was constructed using the normal approximation. For the total scores on the knowledge test, which used a 2-tailed paired *t* test, we calculated the effect size Cohen $dz=t/\mathrm{sqrt}(N)$ and its 95% CI. Here, t is the paired *t* test statistic, and N is the number of participants. The 95% CI was constructed using the normal approximation. The statistical software IBM SPSS Statistics (version 28.0.1.0; IBM Corp.) was used for the analysis.

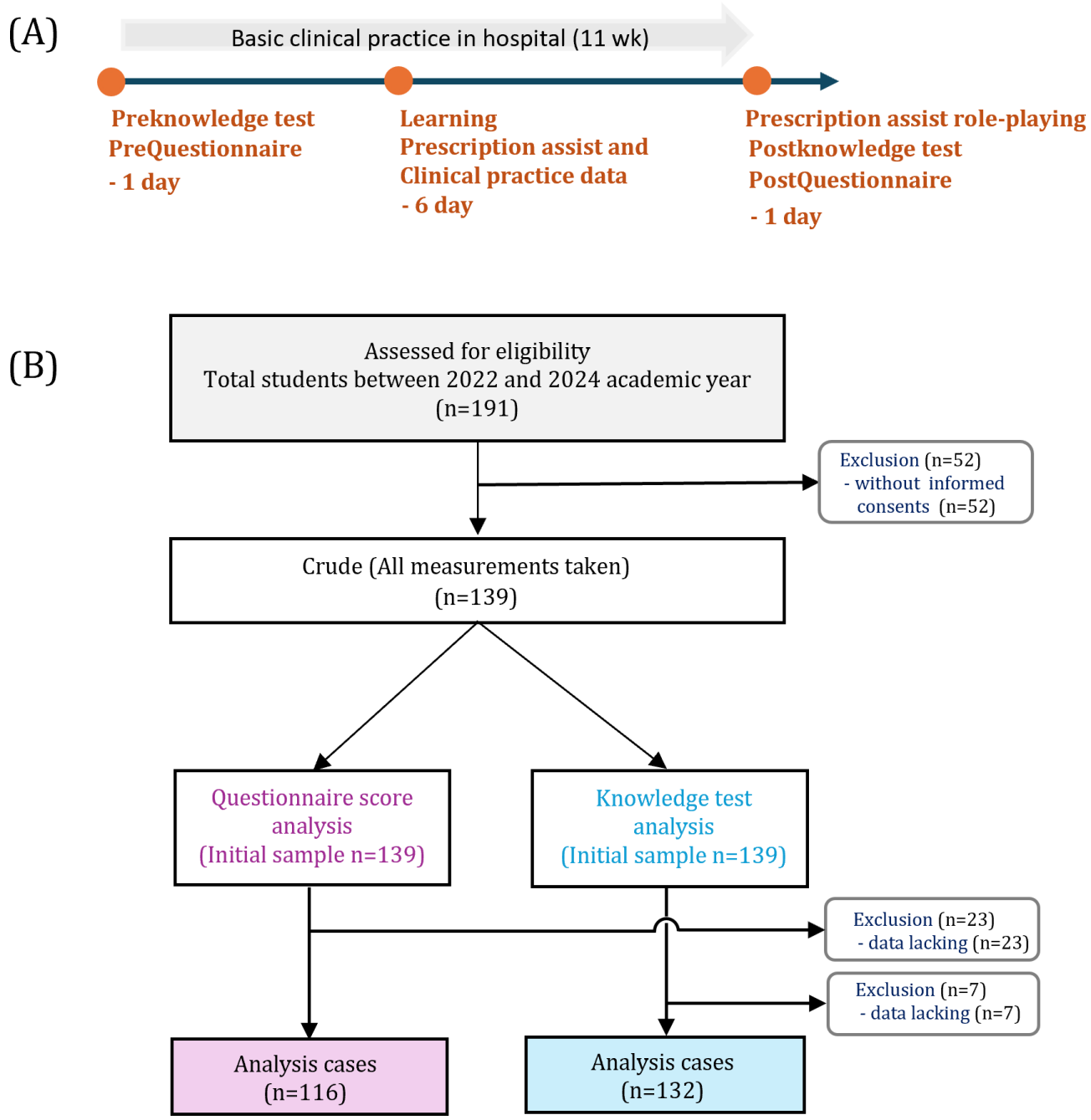
Results

Participants' Characteristics

The study was conducted from May 23, 2022, to February 11, 2024, and 191 students were included (Figure 1). As 52 participants did not provide consent to participate, they were excluded, resulting in 139 participants (Table S2 in Multimedia Appendix 1). The questionnaire and tests were administered to all students during clinical practice. Approximately 116 cases (58 male participants, 50%) were included in the questionnaire analysis, after excluding 23 cases (6 male participants, 26.1%) with invalid responses before or after the intervention. For the test analysis, 132 cases (59 male participants, 44.7%) were included, excluding 7 cases (4 male participants, 57.2%) with no data before or after the intervention. Among the data used for analysis, there were 87 (75%) affiliated students for the questionnaire and 100 (75.7%) for the knowledge test (Table S3 in Multimedia Appendix 1).

XSL•FO
RenderX

Figure 1. (A) Study design and (B) participant selection.



Raising Students' Awareness as Medical Professionals

For participants with baseline scores but no postintervention scores, the missing postintervention scores were imputed using the baseline scores (baseline observation carried forward), and the analysis was performed. The results showed no significant deviation from the original analysis. To evaluate the reliability of the measurement tools, Cronbach α coefficients were calculated. The SBO questionnaire demonstrated good-to-excellent internal consistency, with alpha values ranging from 0.700 to 0.912 across its factors.

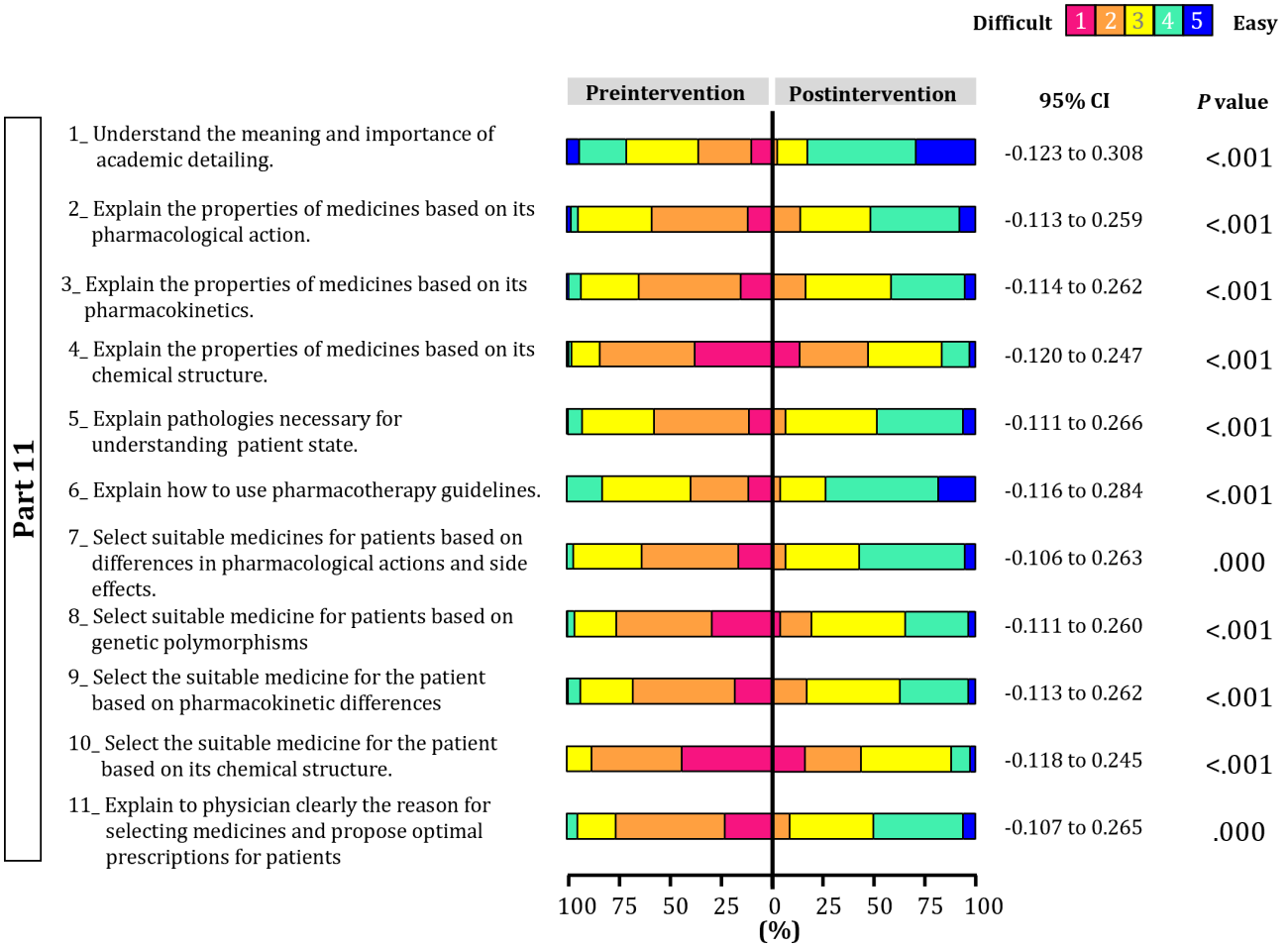
The analysis of the primary outcome revealed statistically significant effects across all items (Q1-1 to Q11-11; Figures 2 and 3, Multimedia Appendix 1; all $P<.001$). Effect sizes ranged from moderate to large (effect size: 0.505 - 0.835), with particularly large effects observed in Q3-2 (effect size: 0.835, 95% CI 0.126-0.330). The lower bound of the 95% CI exceeded zero for all items, suggesting that the observed effects were not due to chance but rather stable effects. For the secondary outcomes, unlike the primary outcome, no statistically significant effects were observed for most items (eg, Q2-1; $P=.59$, Table S1 in Multimedia Appendix 1). Effect sizes were widely distributed across both positive and negative values (effect size: –0.746 to 1.000), and the 95% CI for items showing no significant difference mostly crossed zero. Statistically significant effects were observed only for Q1-3 (effect size: 0.332, 95% CI 0.055-0.609; $P=.02$).

XSL•FO
RenderX

Figure 2. Changes in specific behavioral objectives in pharmacy students. Each color bar shows the changes in self-evaluation in (A) building trust with others, (B) active learning and development of future leaders, (C) handling medical information and planning drug therapy, and (D) interprofessional care. On a 5-point scale, the lower number indicates “difficult.” The x-axis shows the percentage of each respondent when the total is set at 100% (n=116).

XSL•FO
RenderX

Figure 3. Changes in original specific behavioral objectives in pharmacy students.

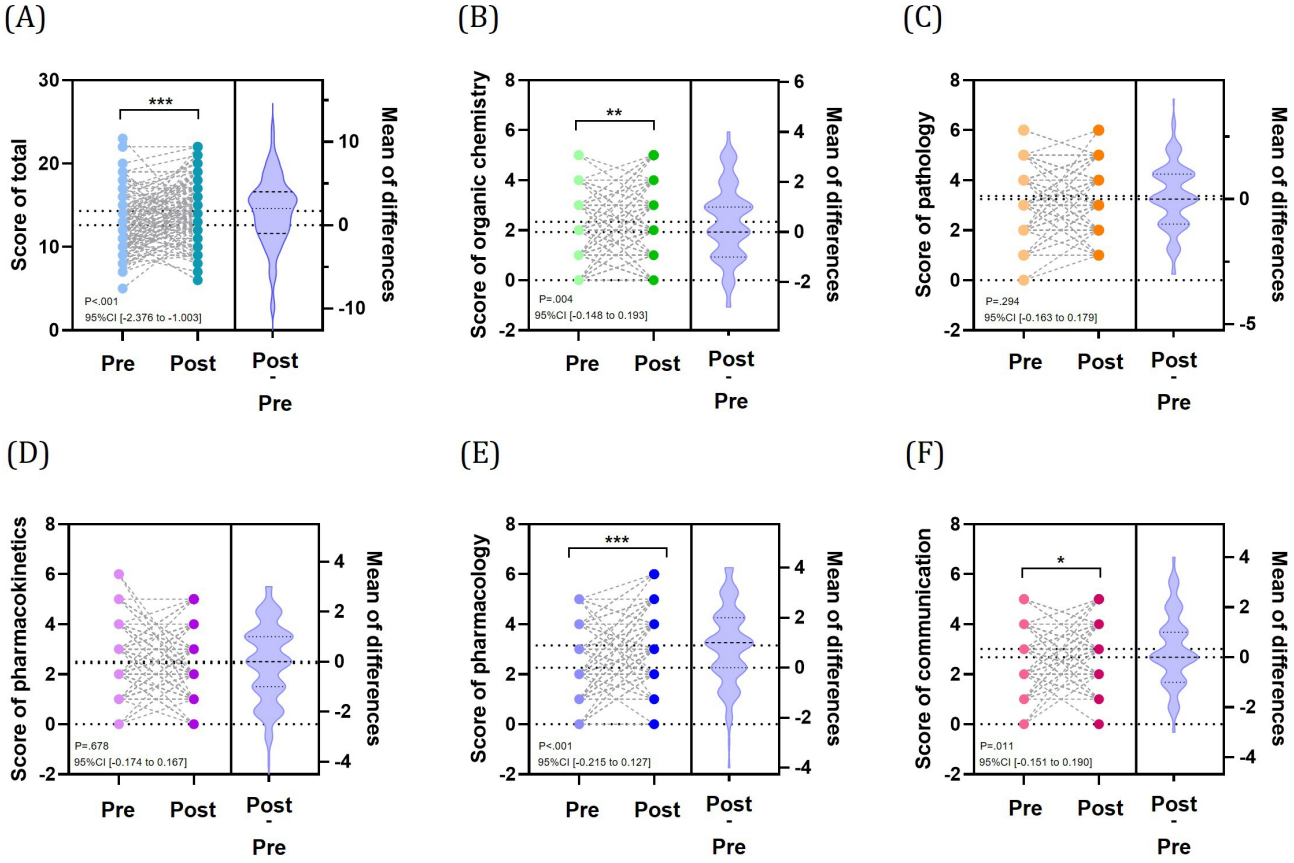


Knowledge Building for Prescribing Planning in Students

In the knowledge test of prescription support in the related program, no significant deviation from a normal distribution was observed for the pre-post difference in total knowledge test scores (out of 30 points). The 30-item knowledge test showed a Cronbach α of .427. Knowledge test scores (Figure 4) showed significant improvement in the following 3 subjects: pharmacology (effect size: −0.504, 95% CI −0.215 to −0.127; *P*<.001), organic chemistry (effect size: 0.254, 95% CI −0.148 to −0.193; *P*=.004), and communication (effect size: 0.221, 95% CI −0.151 to −0.190; *P*=.01). Meanwhile, no significant effects were observed for pathology (*P*=.29) and pharmacokinetics (*P*=.68). Program-specific analysis (Figure 5) revealed extremely significant effects: organic chemistry in the related program (effect size: 0.342; *P*<.001) and communication in the nonrelated program (effect size: 0.446; *P*<.001), with clear differences in effect size observed between programs. Particularly in pharmacology, while the overall results showed a large negative effect, the related program demonstrated a highly significant large positive effect (effect size: 0.624; *P*<.001), exhibiting the most pronounced difference between programs (Table S3 in Multimedia Appendix 1).

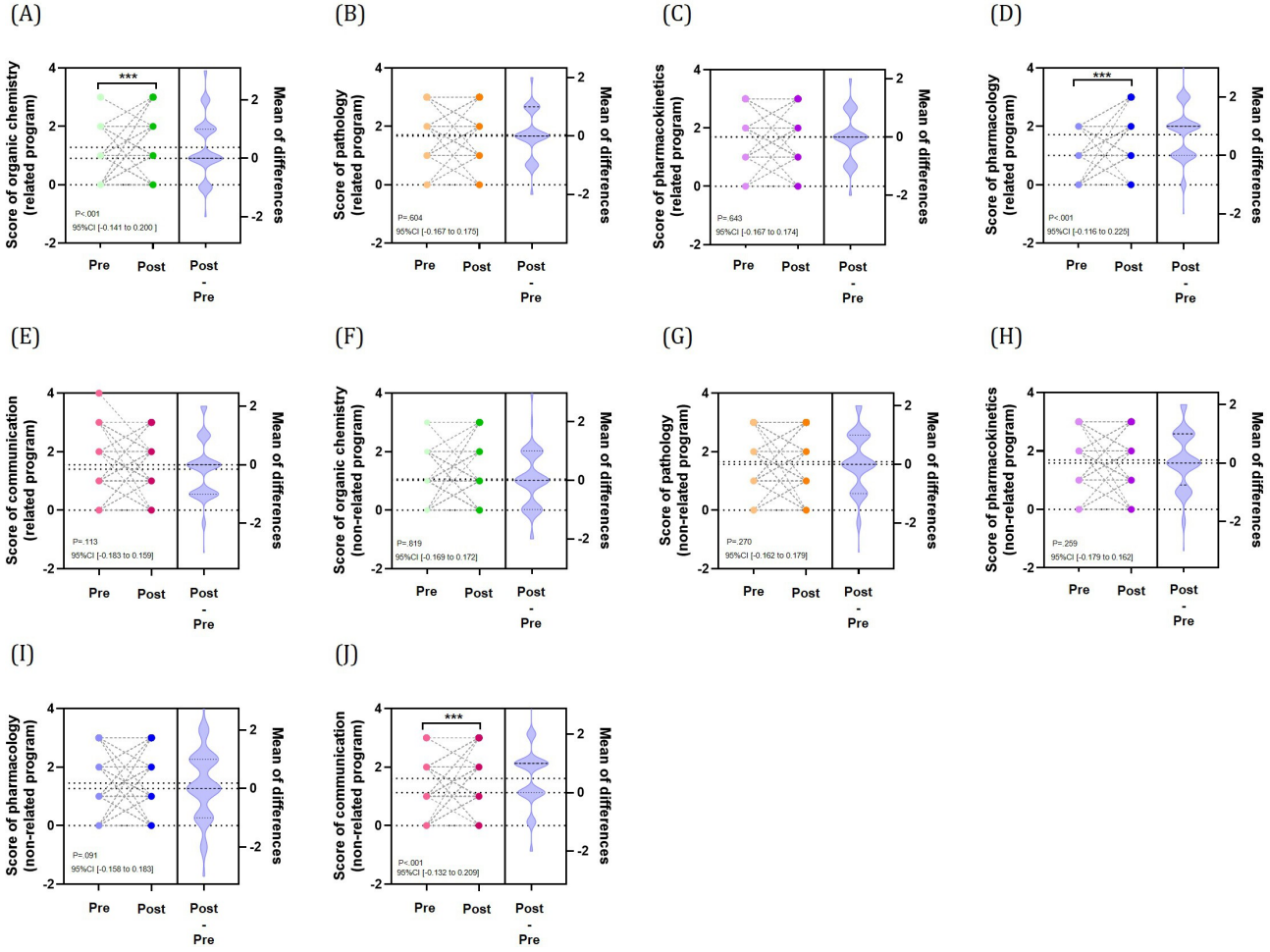
XSL•FO
RenderX

Figure 4. Alteration of knowledge proficiency in basic pharmacy. (A) Total score in basic pharmacy using a paired *t* test. Score of each domain using a Wilcoxon signed-rank test: (B) organic chemistry, (C) pathology, (D) pharmacokinetics, (E) pharmacology, and (F) communication. Violin plots for each figure indicate the mean of score difference before and after intervention. **P*<.05, ***P*<.01, ****P*<.001.



XSL•FO
RenderX

Figure 5. Knowledge proficiency analysis of subgroups in each domain (n=132). Changes in program-related knowledge proficiency questions: (A) organic chemistry, (B) pathology, (C) pharmacokinetics, (D) pharmacology, and (E) communication. Changes in program nonrelated knowledge proficiency questions: (F) organic chemistry, (G) pathology, (H) pharmacokinetics, (I) pharmacology, (J) and communication. Violin plots of each figure indicate the mean of score difference before and after intervention. A Wilcoxon signed-rank test was used for organic chemistry in changes in program-related knowledge proficiency questions and for pathology and communication in program nonrelated knowledge proficiency questions. ***P<.001.



Discussion

Principal Findings

This study aimed to evaluate a specialized educational program designed to enhance pharmacy students' ability to apply professional knowledge in interprofessional care settings during hospital practice. Our findings indicate that the program significantly improved students' self-reported confidence and objective knowledge levels. Notably, while over 70% of the students initially felt their communication skills were insufficient, this perception shifted positively following the intervention. The objective knowledge tests mirrored this improvement, particularly in pharmacology and the ability to apply pharmaceutical information to drug therapy planning.

They had opportunities to share information correctly and concisely with medical staff and worked in groups to summarize and present their research, analogous to a medical team environment. While students could learn the necessary elements for prescription support, the collaboration skills were generic. The program included students and pharmacists, using a simulated prescription proposal. These skills are obtained in clinical practice, and it is difficult to obtain them with static learning at universities. The importance of clinical practice has been mentioned in various countries, and optimal educational methods are currently under consideration [24-26]. Students are aware that health care professionals require continuous learning and education but lack understanding of how to gather information. Hospital practice requires learning about management in commercial pharmacies, and this requires evidence. We provided students with clinical research papers, secondary sources, and various guidelines. The students examined the validity of the evidence and determined whether it was appropriate. This helped them discriminate between sources to use them effectively. Almost all students answered with an understanding and ability to apply the medical and pharmaceutical information to plan drug therapy. By contrast, improvements in the ability to inform and discuss with medical staff were less significant. The students participated in conferences on antibiotic stewardship and nutritional management to learn the role of pharmacists in collaborative care. We designed the program within a limited duration to complete the basic program specified by the educational institutions of Japan. However, the limited time for implementing multidisciplinary collaboration was insufficient

XSL•FO
RenderX

and might have limited the status of improvement of practical items.

Clinical practice is expected to improve the knowledge of students with or without the present program. However, if pharmacists do not strongly intervene in student learning, then improvement in items not related to the program might not be expected. We assessed the improvement in students' knowledge levels by administering a knowledge test on the content covered and not covered in the program to supplement the self-assessments. This confirmed whether the self-reported knowledge acquisition was correct. The test scores increased significantly after the intervention; however, there were no significant differences between the 2 domains in terms of pharmacokinetics and pathophysiology. Students were expected to already know about medicines, and we chose constipation as the theme, because it is a common ailment in children and adults, and fewer types of medicines are available. Most drugs used to treat constipation act on the colon, and there are few drug-drug interactions. This might have made it difficult for students to understand the characteristics of each drug. However, the average score distribution of pathology regarding both program-related and nonprogram-related content was higher than that of the other regions in the preintervention period. There were no marked differences between the pre- and postintervention values. Considering that the relevance of physiology and pathophysiology is important for understanding the pharmacological effects of drugs, the pharmacology scores were not high before the intervention. The average pharmacology score improved at the end of the program, particularly for program-related questions. This may be due to the greater focus on assessing drug interactions to recommend a suitable drug to the simulated patient when preparing prescription recommendations for the physician. Drug interactions include drug formulation characteristics, such as dosage form, pH, chemical modifications, induction or antagonism of metabolic enzymes, and overlap of pharmacological action targets. The practice of prescribing proposals for a simulated patient—based on their concomitant medications and therapies—requires a problem-solving approach to learning, rather than a classroom lecture. Although there was an improvement in the knowledge of content not covered in the program, it was not as great as that of the program-covered content. This suggests that learning about pharmacology is more meaningful when it is combined with the pathophysiology of a disease or drug formulation. The distribution of scores on the organic chemistry questions was sparser than that in the other areas, although there was some improvement after the intervention. The results of the questionnaire on understanding structure-based features or selection improved after the intervention; however, more than half the students' answers in the questionnaire test were "below average" (mean score <3). The focus of organic chemistry at college is on reactions in the synthesis and properties of the synthesized compounds (eg, the presence of isomers, stability, and insolubility), not on understanding the structure, as it relates to pharmacology and kinetics. Drugs are synthesized to be stable and persistent by adding chemical modifications, such as insulin and statins, to endogenous molecules in the body. It is easy for students to speculate about the relationship between the structural properties and the disease or pharmacokinetics of the drug. However, the structures of natural molecules synthesized by molecular structural analysis are often complex, and it is difficult to link this knowledge to that in areas other than organic chemistry. Understanding the drug structure is necessary to estimate changes in formulation, drug stability, and pharmacokinetics. In this program, lectures on each specialty area were provided by hospital pharmacists, who discussed the structural characteristics of each laxative and the properties related to its structure in organic chemistry. The results of the preimplementation questionnaire revealed that the students lacked expertise in organic chemistry and pharmacokinetics. Lectures by hospital pharmacists may be useful for students to understand the meaning of organic chemistry in pharmacotherapy. However, this could be better remedied through combined lectures at universities—such as linking organic chemistry to pharmacology and pharmacokinetics—or collaborating with academic or clinical teachers in practice. Collaboration between universities and medical institutions is not well advanced. Sharing this program between both institutions might enable them to not only practice seamless clinical education but also to produce pharmacists with a research focus.

Limitations

This study has some limitations. First, it evaluated the educational effects of adding practical content as advanced learning to a basic program. To avoid differences in education, the results of a single group were compared before and after intervention. We were required to provide instruction on all items mandated within the prescribed period by Japanese educational institutions. Consequently, we needed to ensure equal educational opportunities for all participants, making it difficult to use a control group for comparison. While this study demonstrates the impact of the educational program we developed, it cannot fully exclude the influence of the natural course of the training period or other confounding factors. Nevertheless, the program-related content scored better in the question rating whether the content was program-related or not in the knowledge test. It seems there was a direct improvement in the program implementation. This study should be considered a pilot test, necessitating the implementation of more rigorous clinical research in the future. The second limitation was the duration of students' knowledge retention. The clinical practice was 11 weeks, and the evaluations were conducted at the beginning and end. Improvements in knowledge and understanding of the pharmacist's role in interprofessional care were observed during this period. However, we were unable to track whether the students sustained their learning when they became medical professionals. Therefore, a nationwide system for continuous data collection after graduation is required. The third limitation is the collaboration with other medical professionals. The program was only conducted with pharmacy students and hospital pharmacists. Although pharmacy students may spend time at medical team conferences, it is necessary to establish a long-term cross-sectional committee to assemble the contents of a common learning program. Much of the reporting on multidisciplinary collaboration is based on medical schools [27]. A previous study reported on interprofessional education

XSL•FO
RenderX

by medical and pharmacy students [28]. It is necessary to collaborate with institutional staff so as to establish this kind of learning system in the future. The fourth limitation is that this study concerns the internal consistency of the 30-item paper test. The calculated Cronbach α was .427, which is below the generally accepted threshold of .70. This low value may be attributed to the multidimensional nature of the test, which was designed to assess a broad range of knowledge rather than a single construct. Future studies should consider using validated standardized tests or refining the test items to improve reliability. Finally, this study evaluated the educational effects of the original program at a single center. The participating students belonged to different universities and a certain level of educational effect was expected for all students. However, it is unclear whether the same effect can be achieved at different hospitals because the medical institutions provided different educational resources. We plan to conduct this program at several medical institutions to determine its usefulness as an educational item for basic rather than advanced programs.

Conclusions

The key contribution of this study is that it measured and documented significant improvements in the ability of pharmacy students to provide direct collaboration with the members of interprofessional care, particularly with respect to drug information and prescription consultations. Our findings provide evidence that practical, hands-on learning facilitated by hospital pharmacists rather than static university lectures is essential for optimizing pharmaceutical therapies. By equipping students with the skills to propose evidence-based prescriptions, this program prepares them to become integral members of the health care team. Furthermore, as physicians increasingly recognize that pharmacist-led prescription support can optimize clinical processes, reduce medical costs, and minimize side effects [29]. By integrating the specialized expertise of physicians and pharmacists, this program represents an important step toward a seamless, coordinated, and advanced medical care system. Ultimately, fostering these collaborative competencies in clinical practice will lead to more individually optimized pharmacotherapy and improved patient outcomes in the evolving landscape of global interprofessional care.

Acknowledgments

The authors thank Masaya Kanda, Masateru Kondo, Takashi Bando, and Yumi Hasegawa at Tokushima University Hospital for teaching assistance. We also thank Editage for English language editing. We received statistical analysis instruction from Statcom Co Ltd (Tokyo, Japan).

Funding

No external financial support or grants were received from any public, commercial, or not-for-profit entities for the research, authorship, or publication of this study.

Data Availability

The datasets generated or analyzed during this study are not publicly available due to personal information protection and ethical guidelines of the authors' institution but are available from the corresponding author on reasonable request.

Authors' Contributions

FA contributed to writing the original draft preparation and conceptualization and performed intervention and analysis. KY and KS contributed to conceptualization and methodology. TH and ST contributed to analysis and writing–review and editing. HH and YZ contributed to methodology. TN, MG, and KK contributed to writing–review and editing. KI contributed to funding acquisition and supervision.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Supplementary tables on (1) additional information of specific behavioral objectives (2) additional information of specific behavioral objectives to prescription support practice, (2) basic characteristics of pharmacy students, and (4) all cases of analysis.
[DOCX File, 31 KB - mededu_v12i1e79545_app1.docx]

Checklist 1
CONSORT checklist.
[PDF File, 286 KB - mededu_v12i1e79545_app2.pdf]

References

XSL•FO
RenderX

1. Cadet T, Cusimano J, McKearney S, et al. Describing the evidence linking interprofessional education interventions to improving the delivery of safe and effective patient care: a scoping review. J Interprof Care 2024;38(3):476-485. [doi: 10.1080/13561820.2023.2283119] [Medline: 38124506]
2. Pradelli L, Risoli C, Summer E, et al. Healthcare professional perspective on barriers and facilitators of multidisciplinary team working in acute care setting: a systematic review and meta-synthesis. BMJ Open 2025 Mar 21;15(3):e087268. [doi: 10.1136/bmjopen-2024-087268] [Medline: 40118478]
3. Jairoun AA, Al-Hemyari SS, Shahwan M, Zyoud SH, El-Dahiyat F. Community pharmacist-led point-of-care eGFR screening: early detection of chronic kidney disease in high-risk patients. Sci Rep 2024 Mar 27;14(1):7284. [doi: 10.1038/s41598-024-56765-0] [Medline: 38538618]
4. Kim IJ, Ryu G, Rhie SJ, Kim HJ. Pharmacist interventions in Asian healthcare environments for older people: a systematic review and meta-analysis on hospitalization, mortality, and quality of life. BMC Geriatr 2024 Jun 12;24(1):513. [doi: 10.1186/s12877-024-05089-9] [Medline: 38867148]
5. Lenander C, Elfsson B, Danielsson B, Midlöv P, Hasselström J. Effects of a pharmacist-led structured medication review in primary care on drug-related problems and hospital admission rates: a randomized controlled trial. Scand J Prim Health Care 2014 Dec;32(4):180-186. [doi: 10.3109/02813432.2014.972062] [Medline: 25347723]
6. Briggs S, Pearce R, Dilworth S, Higgins I, Hullick C, Attia J. Clinical pharmacist review: a randomised controlled trial. Emerg Med Australas 2015 Oct;27(5):419-426. [doi: 10.1111/1742-6723.12451] [Medline: 26190132]
7. Novosadova M, Filip S, Molnarova V, Priester P, Svecova D. Clinical pharmacist in oncology palliative medicine: drug compliance and patient adherence. BMJ Support Palliat Care 2023 Dec;13(e3):e1308-e1317. [doi: 10.1136/spcare-2023-004212]
8. Hurley E, Walsh E, Foley T, Heinrich CH, Byrne S, Dalton K. General practitioners' perceptions of pharmacists working in general practice: a qualitative interview study. Fam Pract 2023 Mar 28;40(2):377-386. [doi: 10.1093/fampra/cmac115] [Medline: 36242543]
9. McGrory F, Elnaem MH. Pharmacist-led interventions for polypharmacy management in older adults: a systematic review of strategies and outcomes in the United Kingdom and the Republic of Ireland. Pharmacy (Basel) 2025 Aug 19;13(4):109. [doi: 10.3390/pharmacy13040109] [Medline: 40863706]
10. Kempen TGH, Koumi R, Sporrong SK. Pharmacists in general practice: what do they do? A qualitative case study. Int J Clin Pharm 2023 Dec;45(6):1472-1482. [doi: 10.1007/s11096-023-01619-4] [Medline: 37573267]
11. Bishop C, Yacoob Z, Knobloch MJ, Safdar N. Community pharmacy interventions to improve antibiotic stewardship and implications for pharmacy education: a narrative overview. Res Social Adm Pharm 2019 Jun;15(6):627-631. [doi: 10.1016/j.sapharm.2018.09.017] [Medline: 30279131]
12. Kuruc Poje D, Kuharić M, Posavec Andrić A, et al. Perspectives of primary care physicians on academic detailing for antimicrobial stewardship: feasibility and impact assessment. J Int Med Res 2024 Jan;52(1):3000605231222242. [doi: 10.1177/03000605231222242] [Medline: 38193298]
13. Kithulegoda N, Chu C, Tadrous M, et al. Academic detailing to improve appropriate opioid prescribing: a mixed-methods process evaluation. CMAJ Open 2023;11(5):E932-E941. [doi: 10.9778/cmajo.20210050] [Medline: 37848255]
14. Rome BN, Dancel E, Chaitoff A, et al. Academic detailing interventions and evidence-based prescribing: a systematic review. JAMA Netw Open 2025 Jan 2;8(1):e2453684. [doi: 10.1001/jamanetworkopen.2024.53684] [Medline: 39775805]
15. Katada Y, Nakagawa S, Minakata K, et al. Efficacy of protocol-based pharmacotherapy management on anticoagulation with warfarin for patients with cardiovascular surgery. J Clin Pharm Ther 2017 Oct;42(5):591-597. [doi: 10.1111/jcpt.12560] [Medline: 28503837]
16. Shibata Y, Koga Y, Sato Y, et al. Pharmacist intervention through protocol-based pharmacotherapy management is effective to ensure safety in invasive procedures for chronic liver disease. J Pharm Health Care Sci 2025 Aug 14;11(1):74. [doi: 10.1186/s40780-025-00456-z] [Medline: 40814072]
17. Ie K, Hirose M, Sakai T, et al. Medication optimization protocol efficacy for geriatric inpatients: a randomized clinical trial. JAMA Netw Open 2024 Jul 1;7(7):e2423544. [doi: 10.1001/jamanetworkopen.2024.23544] [Medline: 39078632]
18. Repas JM, Pruchnicki Coyle MC, Revelos FC, Winhoven MM, Li J, Legg JE. Improving self-perceived competencies of second-year pharmacy students through an introductory medication reconciliation rotation. Am J Pharm Educ 2024 Aug;88(8):100750. [doi: 10.1016/j.ajpe.2024.100750] [Medline: 38964505]
19. Bajis D, Chaar B, Basheti IA, Moles R. Pharmacy students' medication history taking competency: simulation and feedback learning intervention. Curr Pharm Teach Learn 2019 Oct;11(10):1002-1015. [doi: 10.1016/j.cptl.2019.06.007] [Medline: 31685169]
20. Wallman A, Vaudan C, Sporrong SK. Communications training in pharmacy education, 1995-2010. Am J Pharm Educ 2013 Mar 12;77(2):36. [doi: 10.5688/ajpe77236] [Medline: 23519011]
21. Etukakpan A, Uzman N, Ozer O, et al. Transforming pharmaceutical education: a needs-based global analysis for policy development. Explor Res Clin Soc Pharm 2023 Mar;9:100234. [doi: 10.1016/j.rcsop.2023.100234] [Medline: 36876147]
22. The Action to Control Cardiovascular Risk in Diabetes Study Group. Effects of intensive glucose lowering in type 2 diabetes. N Engl J Med 2008 Jun 12;358(24):2545-2559. [doi: 10.1056/NEJMoa0802743]

XSL•FO
RenderX

23. Hashimoto H, Abe M, Tokuyama O, et al. Olanzapine 5 mg plus standard antiemetic therapy for the prevention of chemotherapy-induced nausea and vomiting (J-FORCE): a multicentre, randomised, double-blind, placebo-controlled, phase 3 trial. Lancet Oncol 2020 Feb;21(2):242-249. [doi: 10.1016/S1470-2045(19)30678-3] [Medline: 31838011]
24. Marriott JL, Nation RL, Roller L, et al. Pharmacy education in the context of Australian practice. Am J Pharm Educ 2008 Dec 15;72(6):131. [doi: 10.5688/aj7206131] [Medline: 19325951]
25. Austin Z, Ensom MHH. Education of pharmacists in Canada. Am J Pharm Educ 2008 Dec 15;72(6):128. [doi: 10.5688/aj7206128] [Medline: 19325948]
26. Ryan M, Shao H, Yang L, et al. Clinical pharmacy education in China. Am J Pharm Educ 2008 Dec 15;72(6):129. [doi: 10.5688/aj7206129] [Medline: 19325949]
27. Ofstad W, Brunner LJ. Team-based learning in pharmacy education. Am J Pharm Educ 2013 May 13;77(4):70. [doi: 10.5688/ajpe77470] [Medline: 23716738]
28. Aizawa F, Hamano H, Okada N, et al. Assessing the effects of interprofessional education by hospital pharmacists on pharmaceutical students using a self-evaluation scale. J Pharm Health Care Sci 2024 Oct 1;10(1):61. [doi: 10.1186/s40780-024-00382-6] [Medline: 39354644]
29. Dyrkorn R, Langaas HC, Giverhaug T, Espnes KA, Rowett D, Spigset O. Academic detailing as a method of continuing medical education. Adv Med Educ Pract 2019;10:717-725. [doi: 10.2147/AMEP.S206073] [Medline: 31507335]

Abbreviations

SBO: specific behavioral objective

Edited by T Leung; submitted 23.Jun.2025; peer-reviewed by M Sturgill, M Ali, O Obadimu; revised version received 04.Jan.2026; accepted 31.Jan.2026; published 02.Mar.2026.

Please cite as:
Aizawa F, Yagi K, Higashionna T, Hamano H, Takahashi S, Zamami Y, Shinomiya K, Niimura T, Goda M, Kawada K, Ishizawa K
Prescription Support Practice for Pharmacy Students: Pre-Post Educational Intervention Study
JMIR Med Educ 2026;12:e79545
URL: https://mededu.jmir.org/2026/1/e79545
doi:10.2196/79545

© Fuka Aizawa, Kenta Yagi, Tsukasa Higashionna, Hirofumi Hamano, Shimon Takahashi, Yoshito Zamami, Kazuaki Shinomiya, Takahiro Niimura, Mitsuhiro Goda, Kei Kawada, Keisuke Ishizawa. Originally published in JMIR Medical Education (https://mededu.jmir.org), 2.Mar.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Education, is properly cited. The complete bibliographic information, a link to the original publication on https://mededu.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Youth-Elder Co-Learning Model in Psychiatric Long-Term Care Education: Mixed Methods Evaluation of Communication and Empathy Outcomes

Ke-Hsin Chueh[1,2,3], RN, PhD; Jeng-Wen Chen[4,5,6], MD, MSc; Li-Ang Lee[7,8,9], MD, MSc, PhD; Hsiang-I Hsu[10], RN, MS; Yen-Ju Lin[11], RN, MSN; Hsiao-Tzu Wang[2,12,13], RN, MSN

[1]Department of Nursing, College of Medicine, Fu Jen Catholic University, New Taipei City, Taiwan

[2]Bachelor Degree Program of Long-Term Care and Health Management, College of Medicine, Fu Jen Catholic University, New Taipei City, Taiwan

[3]Center for Medical Education, College of Medicine, Fu Jen Catholic University, New Taipei City, Taiwan

[4]Department of Otolaryngology-Head and Neck Surgery, Cardinal Tien Hospital and College of Medicine, Fu Jen Catholic University, New Taipei City, Taiwan

[5]Department of Otolaryngology–Head and Neck Surgery, National Taiwan University Hospital, Taipei, Taiwan

[6]Department of Hospital Management, Graduate Institute of Business Administration, Fu Jen Catholic University, New Taipei City, Taiwan

[7]School of Medicine, College of Life Science and Medicine, National Tsing Hua University, Hsinchu, Taiwan

[8]Department of Otorhinolaryngology - Head and Neck Surgery, Linkou Main Branch, Chang Gung Memorial Hospital, Taoyuan, Taiwan

[9]School of Medicine, College of Medicine, Chang Gung University, Taoyuan, Taiwan

[10]MS Program in Transdisciplinary Long Term Care, College of Medicine, Fu Jen Catholic University, New Taipei City, Taiwan

[11]Department of Nursing, Lo-Sheng Hospital, Ministry of Health and Welfare, New Taipei City, Taiwan

[12]Department of Nursing, Bali Psychiatric Center, Ministry of Health and Welfare, New Taipei City, Taiwan

[13]School of Nursing, College of Medicine, National Taiwan University, Taipei, Taiwan

Corresponding Author:
Hsiao-Tzu Wang, RN, MSN
Department of Nursing
Bali Psychiatric Center
Ministry of Health and Welfare
33, Huafushan, Bali Dist
New Taipei City, 249203
Taiwan
Phone: 886 2 26101660 ext 3630
Email: tzu.email@msa.hinet.net

Abstract

Background: Taiwan is projected to become a “super-aged” society by 2025, leading to an increasing demand for community psychiatric long-term care (LTC). This demographic shift necessitates frontline professionals equipped with specialized communication skills and deep empathy. However, traditional didactic teaching often fails to adequately prepare students for the complex emotional and practical challenges of real-world psychiatric caregiving.

Objective: This study aimed to evaluate the effectiveness of the innovative youth-elder co-learning instructional model, which uniquely integrates micro-movie discussions and an intergenerational empathy board game, on adult learners’ professional knowledge, communication competence, empathic development, and overall learning satisfaction.

Methods: A mixed methods, single-group, pre-post design was used. The educational intervention was implemented within an 18-week elective community psychiatric LTC course. Participants included 38 adult learners and continuing education students (aged 19-64 years). Notably, the majority of the cohort (n=29, 76.3%) had no prior practical experience in LTC. Quantitative data were collected using self-assessed and peer-evaluated scales for professional knowledge, communication competence, and empathy at pre-, mid-, and post-course time points, alongside an end-of-semester course student feedback survey. Qualitative data were systematically gathered through structured reflective journals and analyzed using a rigorous 6-phase thematic analysis framework.

XSL•FO
RenderX

Results: Students reported high course satisfaction rates, ranging from 92.4% to 95.3%. Quantitative analysis revealed a notable divergence: there were significant improvements in peer-evaluated outcomes (P<.001) and self-assessed communication competence (P=.004), but there was more conservative, statistically nonsignificant growth in self-assessed scores for professional knowledge (P=.14) and empathy (P=.09). This discrepancy likely reflects adult learners' heightened awareness of professional complexity and self-reflective humility. Furthermore, the qualitative thematic analysis uncovered the following three narrative shifts: (1) the dismantling of generational stereotypes through authentic, face-to-face interaction with real older adults; (2) an empathic awakening regarding the often-invisible burden of family caregivers, catalyzed by the micro-movies; and (3) the successful translation of theoretical nonviolent communication concepts into real-time clinical problem-solving during board game role-plays.

Conclusions: The youth-elder co-learning model shows promise as an innovative, experiential pedagogical approach. By bridging theoretical frameworks with authentic intergenerational contact, the intervention supported students in translating general empathic concepts into actionable communication competencies. However, given the exploratory nature of this study and the absence of a control group, the quantitative findings must be interpreted cautiously. Future research using randomized controlled trial designs across multiple institutions is warranted to establish definitive causal impacts.

(JMIR Med Educ 2026;12:e82812) doi:10.2196/82812

KEYWORDS

community psychiatric long-term care; LTC; youth-elder co-learning; professional knowledge; communication competence; empathy; long-term care

Introduction

The global aging population, with an increasing proportion of individuals aged 65 years and older, presents significant implications for health care education and societal structures [1]. Taiwan is on track to become a "super-aged society" by 2025, where over 20% of its population will be aged 65 years and above, intensifying the need for effective care and mental health services for older adults [2,3]. Mental disorders contribute substantially to the global burden of disease, with disability-adjusted life years and years lived with disability showing notable increases [4]. In Taiwan, approximately 187,000 individuals hold chronic mental illness cards, and this population experiences a shorter median life expectancy compared to the general population [5,6].

The chronic, recurrent, and functionally debilitating nature of mental illness, along with its associated economic burden, often places significant strain on caregivers [7]. Critical health care concerns include reducing relapse rates, facilitating patient community reintegration, and alleviating caregiver burden [8-10]. Despite the high demand for community psychiatric services, there is a widespread lack of public understanding regarding mental health issues, contributing to stigma and discrimination in various aspects of life, including health care, housing, and employment [11]. Stigma can worsen patient adaptation, self-esteem, and depression [12,13]. In health care, mental illness is often viewed as the primary issue, overshadowing other physical health concerns, leading to discriminatory attitudes and unequal treatment [14-16]. Caregivers' stigmatizing attitudes can further reduce empathy and support, while a lack of professional knowledge may compromise care quality [17,18]. Therefore, universities must equip students with accurate mental health care knowledge and empathetic attitudes to counter stigma [19]. Inclusive education principles should be adopted in promoting community psychiatric long-term care (LTC) education to eliminate discrimination and prejudice, ensuring equal access to learning and care opportunities [20].

The educational effectiveness of the youth-elder co-learning model encompasses 3 core mechanisms. First, intergenerational dialogue enables young and older participants to exchange experiences and perspectives on an equal footing, thereby fostering mutual understanding and empathy while challenging generational stereotypes [21]. Second, based on the theory of social constructivism, knowledge is co-constructed through authentic contexts and social interaction; by engaging in concrete tasks and collaborative activities, students can internalize abstract concepts into practical competencies [22]. Third, reciprocal learning emphasizes that both young and older individuals serve as learners and knowledge contributors. Through skill exchange and joint participation, participants enhance mutual understanding and respect, which can help mitigate generational stereotypes and encourage positive adjustments in professional attitudes [23].

In summary, the "youth-elder co-learning" teaching method facilitates young people's understanding of the life circumstances and challenges faced by older adults, thereby enhancing their empathy and sense of social responsibility [3,24]. Furthermore, it has a significant positive impact on young people's attitudes toward older adults, their comfort in interacting with them, and their interest in pursuing careers related to older adult care [23]. Within the youth-elder co-learning model, "micro-movies" and "empathy board games" serve as key interactive platforms to promote mutual understanding and empathy in community-based mental health care [24-28]. Micro-movies offer a narrative structure that enhances character and plot connections, accelerates story progression, and effectively uses classroom time [29]. They are effective tools for public communication and help learners understand the world, with storylines often based on real-life events and incorporating conflicts to leave a lasting impression [30,31]. Empathy board games combine education and communication in a playful format that is applicable across a wide age range [32-35]. They provide experiences that deepen understanding from different perspectives, identify alternative solutions, and encourage critical reflection and interpersonal

XSL•FO
RenderX

communication [36-39]. These games enhance understanding, learning, creativity [40,41], active participation [42], behavioral change [43], critical thinking [44], memory [39], and self-esteem [33]. The game dynamics create a positive atmosphere, changing learners' attitudes toward the content and encouraging empathy for characters, extending to real-life situations [45,46]. By successfully integrating local knowledge and culture, as demonstrated in Taiwan, board games can foster intergenerational communication and understanding [47]. In LTC environments, they can enhance caregiver-patient interaction and understanding, thereby improving physical and mental health [26,48]. Furthermore, cultivating and accurately evaluating empathy remains a cornerstone of professional health care education [49], which can be effectively measured using culturally validated instruments such as the Chinese version of the Jefferson Scale of Empathy [50].

Therefore, the development of an effective educational platform for community psychiatric LTC, incorporating youth-elder co-learning, micro-movies, and empathy board games, represents a valuable avenue. Traditional lecturing often falls short in preparing students for the subtle emotional demands of psychiatric care. This gap is particularly pronounced among nontraditional college students and adult learners in continuing education programs. Although these students often possess diverse life and professional experiences, they frequently lack specialized exposure to community psychiatric care environments. Selecting this population provides an opportunity to evaluate how interactive instructional models can support the connection between existing general life competencies and specialized clinical empathy. This approach is designed to support in-service students balancing work and study by facilitating interactive learning within restricted timeframes, managing knowledge depth, addressing mental illness stigma, and fostering self-efficacy. The youth-elder co-learning model guides students to interact and communicate with patients and caregivers, encouraging the development of professional knowledge, competence, and empathy.

Methods

Study Design

This study used a mixed methods, single-group pre-post design to evaluate the educational effectiveness of the youth-elder co-learning instructional model within a university setting. The reporting of this study adheres to the TREND (Transparent Reporting of Evaluations with Nonrandomized Designs) guidelines for the quantitative intervention evaluation and the SRQR (Standards for Reporting Qualitative Research) guidelines for the qualitative components.

Setting

The research was conducted within an 18-week elective community psychiatric LTC course at a private university. The course is a required component of the Continuing Education Degree Program in LTC and Health Management and is also available as an elective to students from other departments. The setting provided an authentic academic environment, deliberately structured to balance foundational theory with intensive practical application.

The study and data collection spanned the entire semester, with specific intervention periods occurring during the middle of the course (weeks 6-8) and late in the course (weeks 14-15). Data collection was strategically aligned with the curriculum. To provide a clear overview of the study's procedural sequence, the 18-week curriculum was systematically mapped. Table 1 illustrates the chronological integration of theoretical lectures and the specific timing of the youth-elder co-learning practical interventions, as well as the timeline for pre-, mid-, and post-course evaluations.

XSL•FO
RenderX

Table 1. Timeline of the youth-elder co-learning educational intervention process and data collection.

Time point	Educational content and intervention	Data collection and evaluation
Week 1	Course orientation, informed consent, and introduction to community psychiatric LTC[a]	Precourse evaluation: self-assessed knowledge, communication, and empathy scales
Weeks 2-5	Theoretical scaffolding: lectures on psychiatric symptom assessment, caregiver burden, and nonviolent communication	—[b]
Week 6	Youth-elder co-learning intervention phase 1: micro-movie A[c] discussion (with older adults)	—
Week 7	Holiday	—
Week 8	Youth-elder co-learning intervention phase 1: empathy board game practicum (with older adults)	—
Week 9	Midterm review and reflection	Midcourse evaluation: self-assessed scales and peer evaluation; qualitative: reflective journal
Weeks 10-13	Theoretical scaffolding: lectures on depression, suicide prevention, and substance abuse	—
Week 14	Youth-elder co-learning intervention phase 2: micro-movie B[d] discussion (with older adults)	—
Week 15	Youth-elder co-learning intervention phase 2: empathy board game practicum (with older adults)	—
Week 16	Course wrap-up and final reflection	Postcourse evaluation: self-assessed scales; qualitative: reflective journal
Week 17-18	Flexible self-directed learning weeks	Final evaluation: peer evaluation and Course Student Feedback Survey

[a]LTC: long-term care.

[b]Not applicable.

[c]Micro-movie A: the patient.

[d]Micro-movie B: unintentional mistake.

Participants and Study Size

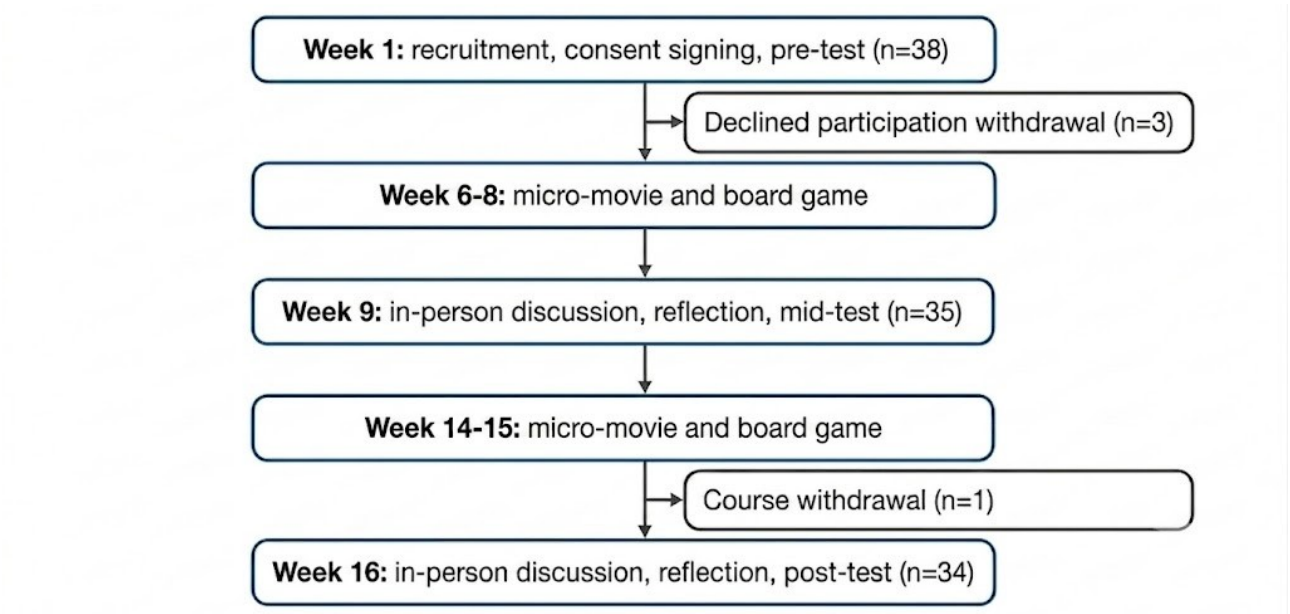
The target population for this study was intentionally selected from adult learners and in-service students enrolled in the community psychiatric LTC course. Participants were recruited using convenience sampling. The eligibility criteria for inclusion were (1) official enrollment in the course during the study semester and (2) provision of voluntary informed consent to participate in the study. Students who did not consent were excluded from the data collection but still participated in the regular course activities without penalty. Regarding the study size, it was intrinsically determined by the total enrollment capacity of the elective course for that semester. A formal a priori statistical power analysis was not conducted because the maximum possible sample size was fixed by institutional course limits.

Consequently, an initial cohort of 38 students who met the eligibility criteria formed our sample. Although the sample size was determined by institutional enrollment limits for the elective course, a post hoc power analysis conducted using G*Power software (version 3.1; Heinrich Heine University Düsseldorf) indicated that, with N=38 and α=.05, the study achieved a power (1–β) of 0.92 for detecting medium to large effect sizes. Given the exploratory nature of this intervention, this level of power suggests that the sample size was sufficient for a formative evaluation of the effectiveness of the youth-elder co-learning model.

Demographically, these participants were primarily second-year students spanning a wide age range (19-64 years). Crucially, the majority (29/38, 76.3%) reported having no prior practical experience in LTC or mental health care, making this cohort an ideal target population for evaluating an empathy- and communication-focused intervention. This student population represents a distinct cohort of adult learners and continuing education students who juggle full-time professional responsibilities with academic pursuits. While they bring mature cognitive frameworks to the classroom, their lack of clinical psychiatric familiarity often leads to implicit generational stereotypes or low caregiving confidence. Consequently, investigating this specific college student population allows for a robust assessment of how a targeted educational intervention can shape the professional competence of future frontline LTC providers. Regarding sampling sufficiency for the qualitative component, the inclusion of all 38 enrolled students ensured a comprehensive capture of diverse perspectives within this specific educational context. Data saturation was considered achieved as the analysis of the 38 reflective journals reached a point where no new themes or significant insights related to the youth-elder co-learning model emerged, and the identified patterns—such as the dismantling of generational stereotypes and empathic awakening—became consistently recurrent across the dataset. The research process flow is provided in Figure 1.

XSL•FO
RenderX

Figure 1. Research process flow.



The Youth-Elder Co-Learning Intervention

The youth-elder co-learning model was strictly guided by social constructivism. To put this theory into practice, the intervention was systematically organized into two primary modalities: micro-movie discussions and empathy board game practicums. Regarding the frequency and duration of the program, older adults were invited to the classroom during weeks 6-8 and weeks 14-15. In total, students and older adults met face-to-face for 4 dedicated intergenerational sessions, each lasting exactly 120 minutes. During the micro-movie discussions modality, students and older adults coviewed narrative micro-movies depicting authentic psychiatric LTC dilemmas, which were immediately followed by facilitated discussions to analyze caregivers' motives and identify potential support strategies.

In terms of the implementation procedures of the empathy board game, the practicums followed a strict 4-step sequential protocol to ensure high intervention fidelity and replicability. First, during the group setup phase, mixed groups comprising 4-5 students and 1-2 older adults were formed. Second, the scenario presentation phase required participants to draw cards that depicted authentic community psychiatric LTC challenges. Third, during the role-play and mediated interaction phase, participants used designated "emotion cards" to articulate hidden feelings, which actively encouraged students to practice nonviolent communication. Finally, the entire session concluded with a structured debriefing phase, which consisted of a 60-minute gameplay period followed by a 60-minute debriefing centered on emotional dynamics and practical clinical application.

Assessments and Data Sources

To mitigate measurement bias, data were sourced through multiple methods, incorporating 5 distinct quantitative measurement instruments. Specifically, the quantitative data sources comprised 4 self-assessed scales and one peer-assessed instrument. First, the self-assessed knowledge of community mental health LTC was measured using a 20-item scale developed by the research team, which demonstrated a content validity index (CVI) of 1.00 and a Kuder-Richardson 20 reliability coefficient of 0.64. Second, self-assessed communication competency was evaluated using a 0-10 scale focused on nonviolent and therapeutic communication skills, yielding a CVI of 1.00 and a Cronbach α of 0.96. Third, self-assessed empathy was captured via a modified 20-item Jefferson Scale of Empathy for Health Professions, which exhibited a CVI of 1.00 and a Cronbach α of 0.87. Fourth, to objectively assess the practical application of these competencies, a peer-assessed knowledge, communication, and empathy of community mental health scale was administered using a 5-point Likert rating system, demonstrating a CVI of 1.00 and a Cronbach α of 0.94. Finally, a self-assessed end-of-course student feedback survey used a 1-5 scale to measure perceived pedagogical benefits, students' preferences, and overall satisfaction with the teaching approach, showing a CVI of 1.00 and a Cronbach α of 0.92. To provide full transparency, the complete items for these designated measurement tools are detailed extensively in Multimedia Appendix 1.

Complementing these quantitative metrics, qualitative data were collected from reflective journals submitted voluntarily by participants at weeks 9 and 16. The original purpose of these journals was to serve as a structured formative assessment to capture students' internal cognitive shifts, emotional processing of caregiving burden, and real-time reflections during the intergenerational co-learning process. To ensure the depth of data, students were instructed to complete these journals independently in a private setting and were explicitly guided to move beyond superficial behavioral summaries, focusing instead on profound self-reflection regarding their interactive dynamics with the older adults. To systematically scaffold their responses and elicit deep reflection, students were provided with five specific semistructured guiding questions. First, students reflected on their initial impressions and subsequent changes in perspective regarding older adults with chronic mental illness after the co-learning sessions. Second, they evaluated the

XSL•FO
RenderX

specific communication challenges encountered during the board game practicums and the strategies used to resolve them. Third, the questions prompted students to analyze how the narrative micro-movies altered their subjective understanding of caregiving burden and family dynamics in psychiatric LTC. Fourth, students examined the ways in which the intergenerational dialogue stimulated their evolving professional empathy and clinical values. Finally, they articulated their comprehensive personal insights and future caregiving confidence regarding frontline community LTC.

Mixed Methods Data Integration

To ensure a robust and comprehensive evaluation of the youth-elder co-learning model, this study adopted a convergent parallel mixed methods design, wherein quantitative and qualitative data were collected concurrently during the 18-week intervention but were analyzed independently to maintain analytical integrity. The programmatic integration and triangulation of these 2 data streams were formally executed during the final interpretation and discussion phase using a joint-display triangulation protocol. Specifically, the quantitative findings from the self-assessed and peer-evaluated metrics served to establish the statistical magnitude, significance, and longitudinal trajectories of students' development in professional knowledge, communication competence, and empathy. Concurrently, the qualitative insights extracted from the thematic analysis of reflective journals were used to yield a deeper contextual understanding, effectively explaining the underlying mechanisms of the quantitative shifts by illustrating how and why students processed caregiving burdens and dismantled implicit generational stereotypes.

Through this convergence, the qualitative narratives directly complemented the statistical outcomes, providing concrete behavioral examples of empathic awakening that corresponded with the quantitative score increases. Furthermore, any convergent discrepancies—most notably the distinct variance observed between subjective self-assessments and objective peer evaluations—were systematically cross-examined and juxtaposed within the discussion. This integrative approach allowed the research team to mitigate potential measurement biases, maximize data validation, and construct a highly nuanced, multidimensional assessment of how the interactive instructional intervention shapes the professional competence of future frontline LTC providers.

Data Analysis and Quality Control

Regarding the quantitative analysis, all statistical data processing was performed using IBM SPSS Statistics for Windows (version 25.0). Descriptive statistics were used to summarize the baseline profiles of the participants. For inferential statistics, nonparametric tests were deliberately chosen due to the relatively small sample size and nonnormal data distribution. Specifically, the Wilcoxon signed-rank test was used to evaluate longitudinal changes across the 3 designated time points, with the threshold for statistical significance predefined at an α level of $P<.05$.

For the qualitative analysis, the qualitative component of this study was grounded in the interpretive paradigm, specifically using social constructivism as the underlying philosophical framework. This paradigm posits that knowledge is not a static reality to be discovered but is actively coconstructed through social interactions and the subjective meanings individuals assign to their experiences. By adopting this stance, the research aimed to explore how students meaningfully reconstructed their understanding of psychiatric care and empathy through the intergenerational dialogue with older adults. To operationalize this framework, the research used inductive thematic analysis, strictly adhering to the 6-phase framework outlined by Braun and Clarke [51]. This specific analytic approach was chosen for its theoretical flexibility and its suitability for an inductive, exploratory study. Given the novel nature of the youth-elder co-learning model, this method allowed patterns and themes to emerge directly from the students' lived experiences—unconstrained by preexisting theoretical frameworks—ensuring that the findings authentically reflected the nuances of intergenerational interaction.

The qualitative data processing followed a rigorous protocol. First, reflective journals were exported from the university's online learning platform and stored in secure, password-protected digital folders. During the data management phase, each journal was assigned a unique ID (eg, M1-M38) to ensure participant anonymity. Regarding the specific coding personnel and analytical software used, the systematic thematic coding process was executed manually and independently by 2 primary researchers from the research team who possess comprehensive expertise in qualitative nursing methodologies and health care education. Data entry, tracking, and open coding were managed systematically using a standardized digital matrix within Microsoft Excel. This manual spreadsheet approach facilitated close, iterative engagement with the raw texts, wherein the 2 coders mapped out meaning units—defined as specific sentences or paragraphs conveying a distinct learning experience—and directly paired them with open codes within the spreadsheet. Moving systematically from coding to generating, reviewing, and explicitly defining themes, any discrepancies in interpretation were resolved through iterative discussion between the researchers to ensure analytical rigor and trustworthiness.

In terms of quality control and researcher reflexivity, the research team—comprised of experts in nursing education and LTC management—acknowledged that their professional backgrounds and positive expectations regarding the youth-elder co-learning model could potentially influence the interpretation of the students' reflective journals. To ensure rigor, maintain high standards of trustworthiness, mitigate such biases, and alleviate power dynamics, the primary course instructor, who held a direct power relationship with the students, was strictly excluded from both the research implementation and analysis phases. Researcher triangulation was achieved through independent coding by 2 researchers, followed by consensus-building sessions to resolve any interpretive discrepancies. Throughout the thematic analysis, the researchers maintained a reflexive stance by documenting their internal reactions to the data and maintaining a detailed audit trail that documented all analytical decisions, coding revisions, and the evolution of theme definitions. Although member checking was

XSL•FO
RenderX

not feasible due to the anonymous nature of the submissions, the use of peer debriefing with a senior psychiatric nursing expert further validated that the identified themes were representative and grounded in the clinical educational context. Finally, to further enhance overall quality control and data integrity, all data collection was highly standardized using electronic capture methods (eg, Google Forms).

Ethical Considerations

To ensure the protection of human participants, this study was conducted in strict accordance with the Declaration of Helsinki and ethical guidelines. The research protocol underwent full review and was formally approved by the Fu Jen Catholic University Institutional Review Board (C112216). Before any data collection, written informed consent was obtained from all participants. During the recruitment phase in week 1, students were thoroughly briefed on the study's objectives, procedures, and their right to withdraw at any time without penalty. It was explicitly emphasized that participation was entirely voluntary and would not positively or negatively affect their course grades or academic standing. To safeguard privacy and confidentiality, all collected data were strictly deidentified. Quantitative assessments and qualitative reflective journals submitted via the online learning platform were stripped of personal identifiers and assigned unique alphanumeric codes (eg, "student 01"). Furthermore, the primary course instructor was intentionally blinded to the students' consent status and raw data throughout the semester to mitigate any potential power dynamics. Only independent research team members had access to the securely stored, anonymized dataset. Finally, regarding participant compensation, no financial incentives, gifts, or extra academic credits were provided to the students for their participation in this study.

Results

Quantitative Research Results

The study sample consisted predominantly of female participants (30/38, 78.9%). The average age was 34.47 (SD 15.83) years, ranging from 19 to 64 years, with the 21-30-year age group being the largest (21/38, 54.6%). Most students (29/38, 76.3%) reported no prior experience in LTC, indicating a lack of practical experience in the field (Table 2).

Table 2. Demographic characteristics of participants (N=38).

Characteristic	Values
Sex, n (%)	
Male	8 (21.1)
Female	30 (78.9)
Age (years)	
Mean (SD)	34.47 (15.83)
Range	19-64
Age groups (years), n (%)	
19-20	2 (5.3)
21-30	21 (54.6)
31-40	2 (5.3)
41-50	2 (5.3)
51-60	8 (20.8)
>60	3 (7.8)
LTC[a] experience (years)	
Mean (SD)	0.71 (1.73)
Range	0-6
LTC experience groups (years)	
0	29 (76.3)
<1	3 (7.8)
1 to <2	2 (5.3)
5	2 (5.3)
6	2 (5.3)

[a]LTC: long-term care.

The "youth-elder co-learning" teaching model showed significant improvement only in students' self-assessment of communication skills (z=–2.88; P<.004 from mid to postcourset and z=–1.98; P=.048 from pre- to postcourse). However, there

XSL•FO
RenderX

was no statistically significant improvement in self-assessments of professional knowledge or empathy. In contrast, peer evaluations demonstrated effective progress in all areas: professional knowledge (z=–4.93; P<.001), communication skills (z=–4.94; P<.001), and empathy (z=–4.55; P<.001) (Table 3).

Table 3. Differences in knowledge, competency, and empathy pre- and postintervention (N=38).

Outcome measure	Scores, mean (SD)	Statistical comparison	z	P value[a]
Knowledge				
Precourse (T0; n=38)	11.16 (3.15)	Pre-mid	–0.07	.95
Midcourse (T1; n=35)	11.29 (3.66)	Pre-post	–1.89	.06
Postcourse (T2; n=34)	12.09 (4.05)	Mid-post	–1.49	.14
Knowledge (peer)				
Midcourse (T1; n=35)	2.22 (0.68)	Mid-post	–4.93	<.001
Postcourse (T2; n=35)	3.22 (0.66)	—[b]	—	—
Communication				
Precourse (T0; n=38)	6.45 (2.24)	Pre-mid	–0.01	.99
Midcourse (T1; n=35)	6.54 (2.23)	Pre-post	–1.98	.048
Postcourse (T2; n=34)	7.53 (1.67)	Mid-post	–2.88	.004
Communication (peer)				
Midcourse (T1; n=35)	2.39 (0.65)	Mid-post	–4.94	<.001
Postcourse (T2; n=35)	3.31 (0.63)	—	—	—
Empathy				
Precourse (T0; n=38)	84.97 (18.88)	Pre-mid	–0.95	.34
Midcourse (T1; n=35)	83.06 (17.73)	Pre-post	–0.52	.61
Postcourse (T2; n=34)	89.03 (15.50)	Mid-post	–1.67	.09
Empathy (peer)				
Midcourse (T1; n=35)	2.45 (0.80)	Mid-post	–4.55	<.001
Post course (T2; n=35)	3.33 (0.72)	—	—	—

[a]P values were calculated using the Wilcoxon signed-rank test.

[b]Not applicable.

Overall, student satisfaction with the teaching activities ranged from 92.4% to 95.3%, indicating high acceptance. The micro-movie course and the "youth-elder co-learning" program had the highest satisfaction (95.3%), while the board game course achieved 92.9% satisfaction. Students also reported increased confidence in their ability to work in community mental health LTC after course completion (92.4% satisfaction; Table 4).

Table 4. End-of-course student feedback (N=34).

Item	Scores, mean (SD)	Standardized scores[a] (%)	Ranking
Overall satisfaction with the course arrangement this semester?	4.68 (0.59)	93.5	3
Satisfaction with the "micro-movie" course this semester?	4.76 (0.50)	95.3	1
Satisfaction with the "board game" course this semester?	4.65 (0.65)	92.9	4
Did the "youth-elder co-learning" program improve your satisfaction with community mental health LTC[b]?	4.76 (0.50)	95.3	1
After completing the course, do you feel more confident about working in community mental health LTC[b]?	4.62 (0.65)	92.4	5

[a]Standardized score = (original score / maximum score) × 100%.

[b]LTC: long-term care.

XSL•FO
RenderX

Qualitative Findings

The thematic analysis of students' voluntary reflective journals revealed a notable shift in their cognitive and emotional understanding of psychiatric LTC. Rather than viewing older adults solely through a lens of clinical detachment, students described a dynamic process of developing professional awareness. The qualitative data converged into three interconnected narratives: dismantling generational stereotypes, experiencing a heightened empathic awareness regarding caregiver burden, and translating theoretical communication models into practical clinical skills.

Dismantling Generational Stereotypes Through Authentic Interaction

Initially, some adult learners admitted to holding subtle preconceived notions about older adults in psychiatric settings, often anticipating them to be rigid, defensive, or difficult to engage. However, the direct, face-to-face interactions facilitated by the youth-elder co-learning empathy board game actively disrupted these assumptions. Students noted that the structured yet collaborative environment allowed them to perceive the older adults as multifaceted individuals with rich life histories, rather than merely as care recipients. As one student vividly reflected on this intergenerational bridge, "It was quite special. Their thoughts are actually just as clear as ours. Even though they are much older, they are still learning... interacting with them made me realize they often treat us like their own grandchildren, filled with warmth and love" (M6). This authentic contact forced learners to confront their internal biases, replacing initial apprehension with renewed openness and interpersonal warmth.

Empathic Awakening and the Recognition of Caregiver Burden

The micro-movie discussions served as a powerful emotional catalyst within the curriculum. By analyzing highly realistic portrayals of psychiatric caregiving dilemmas, students moved beyond superficial "perspective-taking" to experience deep emotional resonance. The visual narratives effectively translated abstract textbook concepts into tangible human struggles, shedding light on the immense, often invisible, psychological pressure borne by family caregivers. A recurring sentiment among the participants was the sudden realization of the systemic and emotional complexities inherent in LTC. One participant powerfully noted, "Caregivers are the 'second patients'—I truly understand that now. I used to think 'empathy' was just imagination, but especially after watching the micro-movie, I seemingly truly 'saw' the hardships and struggles of caregivers, and can better appreciate what it means to 'understand another's powerlessness and choices'" (M36). This affective shift supported the expansion of their theoretical empathy into an actionable intent to provide holistic support.

Translating Theory Into Practice via Nonviolent Communication

Perhaps a notable pedagogical outcome extracted from the journals was the students' ability to translate abstract communication theories into actionable, real-time skills. During the situational role-play segments of the board game, students frequently encountered simulated, high-tension communication breakdowns. By using the specific game mechanics (such as cards representing emotions and moods), they practiced articulating underlying needs and actively listening, directly applying nonviolent communication techniques. Students reported that these moments of simulated friction provided valuable learning opportunities, encouraging them to adapt. "After actually simulating the situations of different roles in the game, I realized that everyone carries different pressures and emotions in their hearts... this experience taught me to listen more and be more inclusive in communication, reminding myself to maintain respect and care no matter who I face," expressed another student (M3). Consequently, this experiential learning process supported their confidence in managing complex clinical interactions without resorting to authoritative or dismissive behaviors.

Overall, the qualitative data indicate that the course supported the integration of theoretical knowledge while enhancing empathy, cognitive flexibility, tolerance, care, and respect within complex social and caregiving situations. Throughout the youth-elder co-learning intervention, students demonstrated a more nuanced understanding of vulnerable groups. They reported improvements in communication competencies—specifically in active listening, emotional expression, and conflict resolution—alongside increased empathy, emotional awareness, and self-management. These learning outcomes suggest the potential value of the immersive, interactive teaching methods used in the curriculum. By creating genuine emotional connections through micro-movies and intergenerational board games, the pedagogical approach guided students from cognitive knowledge acquisition to a heightened level of affective understanding and empathy, positively impacting both their future professional practice and their daily interpersonal relationships.

Convergence and Contextualization of Mixed Methods Findings

To present a holistic evaluation of the youth-elder co-learning model, the quantitative statistical trajectories and qualitative narrative insights were actively contextualized together, demonstrating clear cross-methodological convergence and mutuality. Specifically, the significant longitudinal increases observed in the quantitative communication competency and empathy scores were directly illuminated by the students' concurrent journal reflections. For instance, the statistical improvement in therapeutic communication from pre- to postintervention directly corresponded with qualitative themes describing how the "emotion cards" in the board game prompted students to move past academic theories and actively practice nonviolent communication in real-time intergenerational dialogues. This qualitative context complements the data by offering behavioral evidence corresponding with the observed numeric increase.

Furthermore, the qualitative data deeply informed and clarified complex patterns within the quantitative results, particularly regarding the statistical variance between subjective self-assessments and objective peer evaluations. While the quantitative data showed a temporary plateau in students'

XSL•FO
RenderX

self-assessed empathy scores at midcourse, the qualitative narratives contextualized this anomaly as a notable "empathic awakening." The journals revealed that as students engaged in deep, face-to-face dialogue with older adults, they developed an intense awareness of the heavy caregiving burden and their own clinical naivety. This newfound professional humility led them to score themselves more critically on self-assessments, even though their objective peer evaluations simultaneously reflected a high level of clinical competence and interpersonal sensitivity. Consequently, by contextualizing these 2 data streams together, the results provide a much more transparent, multidimensional, and authentic representation of how the interactive intervention shapes both the measurable competencies and the internal psychological growth of future frontline LTC providers.

Discussion

Principal Findings

The study's results affirm that the "youth-elder co-learning" model effectively enhanced students' peer-evaluated professional knowledge, communication skills, and empathy, as well as self-assessed communication competence within the Community Mental Health LTC course. The model successfully fostered interaction and understanding between younger students and older adults (including care recipients and caregivers), thereby deepening learning experiences. These findings strongly support the social constructivist approach, demonstrating that when students and older adults actively coconstruct knowledge through game-based interactions, the translation from theoretical empathy to practical communication competence is significantly enhanced.

Reflection on Practical Challenges Encountered in Youth-Elder Co-Learning Teaching

A notable discrepancy emerged: students' self-assessments of professional knowledge and empathy did not show statistically significant improvement, unlike peer evaluations. This suggests that students may hold a more conservative self-perception of their competencies. This phenomenon could be explained by factors such as students' humble dispositions, self-doubt, or an increased awareness of the complexity of professional roles as the course progressed. As students gained deeper insight into the expectations and challenges of LTC settings, they might have become less confident in self-rating highly. This cautious self-assessment could reflect both modesty and a genuine recognition of the depth and difficulty of the knowledge and empathy in LTC.

Furthermore, the unique demographic characteristics of our sample—primarily adult learners and in-service professionals in a continuing education program—may provide additional context for this cautious self-assessment. Unlike traditional young adult college students, these nontraditional students (spanning ages 19-64 years) bring extensive life experience to the classroom. This maturity may lead to a deeper, more realistic appreciation of the complexities involved in psychiatric LTC, thereby making them more self-reflective, humble, and stringent when evaluating their own professional knowledge and empathic growth.

These findings align with Xu et al [52], who noted that students often underestimate their mastery of complex skills due to uncertainty about performance standards or growing recognition of real-world professional challenges. As formative assessments expose students to authentic tasks, some become more self-critical, revealing that personal values and increased awareness can suppress self-assessment scores, even amid actual progress. Likewise, Collier-Sewell et al [53] found that nursing students' growing awareness of professional complexity was associated with more critical self-evaluation, a phenomenon that may extend to learners in other applied disciplines. Their findings reinforce the idea that as students deepen their understanding of practice realities, they may become more reserved in assessing their own competence, not due to lack of progress, but due to heightened awareness and reflective caution.

These findings highlight the importance of addressing not only students' skill and knowledge growth but also shifts in their internal perceptions, cognitive frameworks, and evaluative processes. Teaching assessments should consider how students evolve in their reflective awareness, aligning with previous research that self-assessment can be influenced by personal values, humility, and recognition of real-world complexities [52-54].

Integration of Micro-Movies and Situational Role-Play in the Community Mental Health LTC Course

The teaching team's blended approach, fusing micro-movie–based teaching and situational role-playing within the youth-elder co-learning model, effectively integrated emotional engagement with practical application, promoting deeper learning in professional knowledge, communication competence, and empathic development. Micro-movies provided vivid, emotionally rich representations of caregiving scenarios, making abstract issues tangible and allowing students to understand the psychological burdens and moral dilemmas of caregivers [55,56]. Role-playing further enhanced this by enabling active dialogue and empathy-building with peers and older adults, a strategy shown to cultivate professional competencies and deepen affective learning outcomes [57]. This combination encouraged students to reflect on how individual backgrounds, resources, context, and both physical and psychological states influence choices and behaviors. Students reported approaching others with greater openness, reduced judgment, and increased emotional intelligence and interpersonal understanding. These interactive and experiential pedagogical strategies supported both cognitive understanding and the emotional resonance necessary for developing core competencies in LTC education [58,59].

Comparison With Previous Studies on Empathy and Intergenerational Education

Our findings regarding the enhancement of communication skills and empathy through the youth-elder co-learning model strongly resonate with recent literature on intergenerational education. Consistent with studies by Trujillo-Torres et al [21] and Pillemer et al [23], our qualitative and peer-assessed results

XSL•FO
RenderX

demonstrate that structured intergenerational dialogue effectively dismantles generational stereotypes and fosters mutual respect. However, our study extends these foundational concepts by applying them specifically to the high-stress, stigmatized context of community psychiatric LTC, demonstrating that intergenerational co-learning is not only beneficial for general socialization but is also a vital pedagogical tool for specialized clinical communication.

Furthermore, regarding empathy development, our outcomes align closely with recent nursing education research. Cho and Kim [58] and Huang et al [59] emphasized that immersive, simulation-based interventions are highly effective in cultivating empathy among health care students. While many previous studies rely on standardized patients or peer-to-peer role-play, our youth-elder co-learning model integrates real older adults into the simulation via the "empathy board game." This authentic interaction operationalizes empathy in real-time. It compels students to move beyond passive "perspective-taking" and actively translate theoretical compassion into practical, on-the-spot communication strategies. The detailed reflective feedback and interpersonal interactions reported by our students suggest that combining game-based learning with active intergenerational contact offers valuable pedagogical opportunities to support empathic development beyond standard classroom lectures.

Limitations

Despite demonstrating meaningful outcomes, this study has several limitations. First, the relatively small sample size and convenience sampling may limit the generalizability of the findings. Second, the self-assessment tools, while carefully developed, were subject to personal bias, influenced by cultural or psychological factors such as modesty or self-criticism. Third, peer assessments, though valuable, may also be affected by interpersonal dynamics within the class. Fourth, the reliance on a single-group pre-post design without a control group at a single university means that we cannot definitively rule out the influence of external confounding factors—such as the natural passage of time or concurrent coursework—on students' development. Consequently, the findings may strongly reflect the specific institutional culture or the unique demographic makeup of this particular continuing education program. Future research should consider incorporating a control group design and replicating the study across different universities, student populations, and educational settings. Longitudinal studies would also be beneficial to explore the sustained impact of the youth-elder co-learning model on professional competence and empathic communication over time.

Conclusion

This study suggests that the youth-elder co-learning model—using micro-movies and intergenerational empathy board games—shows promise as an innovative pedagogical approach for students preparing for community psychiatric LTC. By providing authentic, guided interactions with real older adults, the intervention successfully helped adult learners bridge theoretical knowledge with practical communication and empathic skills. The study contributes to the growing evidence supporting experiential and emotionally engaging teaching methods in LTC education, especially within community mental health.

Through this formative evaluation, several key takeaways emerged to guide future educational research. First, the observed discrepancy between students' self-assessments and objective peer evaluations highlights a critical methodological issue that warrants deeper future exploration. Second, the strategic integration of emotional resonance and cognitive learning can facilitate meaningful improvements in students' developing communication competence and empathic awareness. Finally, future LTC course designs should increasingly incorporate interactive, reflective, and immersive teaching strategies that simulate real-life caregiving scenarios to cultivate students' comprehensive clinical abilities.

However, given the exploratory nature of this study and the absence of a control group, these preliminary results must be interpreted cautiously. While peer evaluations and detailed qualitative feedback indicate meaningful learning patterns, the lack of a comparison cohort prevents us from definitively attributing all observed longitudinal improvements solely to the youth-elder co-learning intervention. These findings highlight the potential value of incorporating active intergenerational contact into LTC curricula, but further rigorous research—specifically using randomized controlled trial designs across multiple institutions—is required to establish definitive causal impacts. Ultimately, by combining emotional engagement with cognitive understanding, the youth-elder co-learning approach fosters not only professional knowledge acquisition but also reflexive capacity and interpersonal sensitivity, which are critical for preparing future nursing and health care professionals to meet the complex demands of community psychiatric care.

Acknowledgments

During the preparation of this paper, the authors used Gemini (Google) for language polishing. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the final content of the publication.

Data Availability

The datasets generated and analyzed during this study are not publicly available due to the inclusion of sensitive qualitative reflections from a small participant cohort but are available from the corresponding author on reasonable request.

XSL•FO
RenderX

Funding

This study was funded by the Ministry of Education Teaching Practice Research Program.

Authors' Contributions

KHC contributed to conceptualization, methodology, supervision, formal analysis, and writing—original draft preparation. JWC contributed to conceptualization, supervision, resources, and writing—review and editing. LAL contributed to conceptualization, supervision, resources, and writing—review and editing. HIH contributed to investigation, data curation, and formal analysis. YJL contributed to data curation. HTW contributed to methodology, intervention design, instruction, and writing—original draft preparation.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Study questionnaires and assessment instruments.
[DOCX File , 47 KB - mededu_v12i1e82812_app1.docx]

References

1. Chang YC, Chen YM. The current situation, challenges and opportunities to implement integrated care in "The Alley Long-Term Care Service C Station"- the view of World Health Organization's ICOPE MESO level approach. J Long Term Care 2020;24(2):93-109 [FREE Full text]
2. Chen HC, Yang CY. Applying scaffold teaching approach to cultivate the cultural competence of elderly service workers. J Taiwan Health Care Assoc 2020(21):36-71 [FREE Full text]
3. Wang LY. Promoting intergenerational co-learning in an aging society: the concepts and practice of intergenerational learning. J Educ Res 2020(315):104 [FREE Full text]
4. GBD 2019 Mental Disorders Collaborators. Global, regional, and national burden of 12 mental disorders in 204 countries and territories, 1990-2019: a systematic analysis for the global burden of disease study 2019. Lancet Psychiatry 2022;9(2):137-150 [FREE Full text] [doi: 10.1016/S2215-0366(21)00395-3] [Medline: 35026139]
5. Statistics on valid certificates for major illnesses and injuries under the National Health Insurance Program (May 2026). Taiwan Ministry of Health and Welfare, Health Insurance Department. URL: https://www.nhi.gov.tw/ch/cp-5034-fd764-3028-1.html [accessed 2026-06-27]
6. World Mental Health Day Health and Welfare Statistics Bulletin (Statistics Department). Taiwan Ministry of Health and Welfare, Department of Statistics. URL: https://dep.mohw.gov.tw/DOS/cp-5112-63761-113.html [accessed 2026-06-27]
7. Hsieh WT, Yang HY, Lai FT. Key demands on the caregivers of mental patients. J Health Manag 2021;22(1):23-45 [FREE Full text]
8. Li YM. Burden, attachment, filial piety and psychological well-being among caregivers of patient with neurocognitive disorder. Chung Shan Medical University, Department of Psychology. 2021. URL: https://doi.org/10.6834/csmu202100052 [accessed 2026-06-27]
9. Chang WB. The distance between mental illness and long-term care: discuss long-term care system for mental patients after Long Fa Tang incident in 2018. Soc Sci Law Pol Rev 2021(9):1-41 [FREE Full text]
10. Lu MJ, Wang CS, Hsu PC, Huang CL. Needs, dilemmas, and policy suggestions related to patients with mental health long-term care. Nurs J 2021;68(2):25-31 [FREE Full text]
11. Goda AMI, Elhamid SAA, Wassif GO. Mental health literacy: a comparative study on stigmatizing attitude and help-seeking behavior towards mental disorders between adolescents and adults. J Egypt Public Health Assoc 2025;100(1):7 [FREE Full text] [doi: 10.1186/s42506-025-00184-0] [Medline: 40293553]
12. Silver S. Stigma and severe mental illness: research summary. Treatment Advocacy Center 2024:1-9 [FREE Full text]
13. Corrigan PW, Al-Khouja MA. Three agendas for changing the public stigma of mental illness. Psychiatr Rehabil J 2018;41(1):1-7 [FREE Full text] [doi: 10.1037/prj0000277]
14. Cunningham R, Imlach F, Haitana T, Every-Palmer S, Lacey C, Lockett H, et al. It's not in my head: a qualitative analysis of experiences of discrimination in people with mental health and substance use conditions seeking physical healthcare. Front Psychiatry 2023;14:1285431 [FREE Full text] [doi: 10.3389/fpsyt.2023.1285431] [Medline: 37908598]
15. Cunningham R, Imlach F, Lockett H, Lacey C, Haitana T, Every-Palmer S, et al. Do patients with mental health and substance use conditions experience discrimination and diagnostic overshadowing in primary care in Aotearoa New Zealand? Results from a national online survey. J Prim Health Care 2023;15(2):112-121 [FREE Full text] [doi: 10.1071/hc23015]
16. Hallyburton A. Diagnostic overshadowing: an evolutionary concept analysis on the misattribution of physical symptoms to pre-existing psychological illnesses. Int J Ment Health Nurs 2022;31(6):1360-1372 [FREE Full text] [doi: 10.1111/inm.13034] [Medline: 35718951]

XSL•FO
RenderX

17. Gunasekaran S, Tan GTH, Shahwan S, Goh CMJ, Ong WJ, Subramaniam M. The perspectives of healthcare professionals in mental health settings on stigma and recovery - A qualitative inquiry. BMC Health Serv Res 2022;22(1):888 [FREE Full text] [doi: 10.1186/s12913-022-08248-z] [Medline: 35804378]
18. Yang YY, Wang C, Cai J, Mu YF, Zhou L, Wang YZ, et al. Mental illness-related stigma and its associated factors among primary healthcare professionals in rural China. Front Psychiatry 2025;16:1519527 [FREE Full text]
19. Arbanas G, Bosnjak D, Sabo T. Impact of a nursing in psychiatry course on students' attitudes toward mental health disorders. J Psychosoc Nurs Ment Health Serv 2018;56(3):45-51. [doi: 10.3928/02793695-20171024-01] [Medline: 29117427]
20. Florian L. On the necessary co-existence of special and inclusive education. Int J Incl Educ 2019;23(7-8):691-704 [FREE Full text] [doi: 10.1080/13603116.2019.1622801]
21. Trujillo-Torres JM, Aznar-Díaz I, Cáceres-Reche MP, Mentado-Labao T, Barrera-Corominas A. Intergenerational learning and its impact on the improvement of educational processes. Educ Sci 2023;13(10):1019 [FREE Full text] [doi: 10.3390/educsci13101019]
22. Çelik Ç, Polat S. The relationship between intergenerational knowledge sharing and intergenerational learning levels among teachers. J Intergener Relatsh 2022;21(4):427-441 [FREE Full text] [doi: 10.1080/15350770.2022.2084202]
23. Pillemer K, Nolte J, Schultz L, Yau H, Henderson CR, Cope MT, et al. The benefits of intergenerational wisdom-sharing: a randomized controlled study. Int J Environ Res Public Health 2022;19(7):4010 [FREE Full text] [doi: 10.3390/ijerph19074010] [Medline: 35409693]
24. Chi CL. Intergenerational co-learning: an attempt and practice of introducing intergenerational learning models in service learning courses. J Serv Learn Soc Link 2023(6):1-24 [FREE Full text]
25. Yen HC. Apply the action research to explore the implementation process and learning effect of cross-generational learning - a case study of Ji-fong Presbyterian Church in Wufeng District, Taichung City. National Chi Nan University, Department of Counseling Psychology and Human Resource Development. 2022. URL: https://doi.org/10.6837/ncnu202200252 [accessed 2026-06-27]
26. Chen YF, Janicki S. A cognitive-based board game with augmented reality for older adults: development and usability study. JMIR Serious Games 2020;8(4):e22007 [FREE Full text] [doi: 10.2196/22007] [Medline: 33315015]
27. Kleijberg M, Ahlberg BM, Hilton R, Tishelman C. Death, loss and community-perspectives from children, their parents and older adults on intergenerational community-based arts initiatives in Sweden. Health Soc Care Community 2020;28(6):2025-2036 [FREE Full text] [doi: 10.1111/hsc.13014] [Medline: 32412151]
28. Leong KS, Klainin-Yobas P, Fong SD, Wu XV. Older adults' perspectives on intergenerational programmes at a Singaporean senior day-care centre: a qualitative study. Health Soc Care Community 2022;30(1):e222--e233 [FREE Full text]
29. Ng SFE, Man TW. A study on integrating the teaching of microfilm appreciation into filial piety education. School Admin 2020(128):155-179 [FREE Full text]
30. Chueh KH. Application of interdisciplinary micro-movie production in medical education. Taiwan Med Educ Soc 2022 [FREE Full text]
31. Mathew A, Mathew J. A study to assess the effectiveness of movie based educational package in teaching psychiatry for undergraduate nursing students in selected nursing college at Thiruvalla. Asian J Nurs Educ Res 2018;8(2):275-277 [FREE Full text] [doi: 10.5958/2349-2996.2018.00053.8]
32. Ferreira PC, Veiga Simão AM, Paiva A, Martinho C, Prada R, Ferreira A, et al. Exploring empathy in cyberbullying with serious games. Comput Educ 2021;166:104155 [FREE Full text] [doi: 10.1016/j.compedu.2021.104155]
33. Olympio PCDAP, Alvim NAT. Board games: gerotechnology in nursing care practice. Rev Bras Enferm 2018;71 Suppl 2:818-826 [FREE Full text] [doi: 10.1590/0034-7167-2017-0365] [Medline: 29791637]
34. Sincharoenkul K, Tongtep N, Boonlamp L. Supervised classification of board games for active learning: enhancing business knowledge and skills. 2020 Presented at: 2020 IEEE ECTI-CON Conference; June 24-27, 2020; Phuket, Thailand URL: https://doi.org/10.1109/ECTI-CON49241.2020.9158301 [doi: 10.1109/ecti-con49241.2020.9158301]
35. Al Fatta H, Maksom Z, Zakaria MH. Game-based learning and gamificationarching for definitions. Int J Simul Syst Sci Technol 2018;19(6):41.1-41.5 [FREE Full text] [doi: 10.5013/ijssst.a.19.06.41]
36. d'Astous A, Gagnon K. An inquiry into the factors that impact on consumer appreciation of a board game. J Consum Mark 2007;24(2):80-89 [FREE Full text]
37. Hunsucker AJ. Board games as a platform for collaborative learning. 2016 Presented at: Meaningful Play 2016; October 20-22, 2016; Lansing, MI URL: https://www.researchgate.net/profile/Andrew-Hunsucker/publication/309385174_Board_Games_as_a_Platform_For_Collaborative_Learning/links/580cda1b08ae2cb3a5e3a394/Board-Games-as-a-Platform-For-Collaborative-Learning.pdf
38. Dahlin JE, Fenner R, Cruickshank H. Evaluating simulations and games as tools for expanding student perspectives on sustainability. 2015 Presented at: The 7th International Conference on Engineering Education for Sustainable Development; June 9-12, 2015; Vancouver, BC p. 1-8 URL: https://open.library.ubc.ca/soa/cIRcle/collections/52657/items/1.0064673
39. Boghian I, Cojocariu VM, Popescu CV, Mâ ă L. Game-based learning: using board games in adult education. J Educ Sci Psychol 2019;9(1) [FREE Full text]

XSL•FO
RenderX

40. Azizan MT, Mellon N, Ramli RM, Yusup S. Improving teamwork skills and enhancing deep learning via development of board game using cooperative learning method in reaction engineering course. Educ Chem Eng 2018;22:1-13 [FREE Full text] [doi: 10.1016/j.ece.2017.10.002]
41. Silveira MS. Creativity and learning through building board games in HCI courses. 2020 Presented at: ACM ITiCSE (Innovation and Technology in Computer Science Education) conference; June 15-19, 2020; Trondheim, Norway URL: https://doi.org/10.1145/3341525.3387374
42. Noda S, Shirotsuki K, Nakao M. The effectiveness of intervention with board games: a systematic review. BioPsychoSoc Med 2019;13:22 [FREE Full text] [doi: 10.1186/s13030-019-0164-1] [Medline: 31641371]
43. Douglas BD, Brauer M. Gamification to prevent climate change: a review of games and apps for sustainability. Curr Opin Psychol 2021;42:89-94. [doi: 10.1016/j.copsyc.2021.04.008] [Medline: 34052619]
44. Duncan KJ. Examining the effects of immersive game-based learning on student engagement and the development of collaboration, communication, creativity and critical thinking. TechTrends 2020;64(3):514-524 [FREE Full text] [doi: 10.1007/s11528-020-00500-9]
45. Ilić Rajković A, Senić Ružić M, Ljujić B. Board games as educational media: creating and playing board games for acquiring knowledge of history. 2019 Presented at: IARTEM 2017 Conference Proceedings; September27-29, 2019; Lisbon, Portugal URL: https://doi.org/10.21344/iartem.v11i2.582
46. Krause A, Headley S, Bryant D, Watkin A, Woodard CM, Sinicki S. Game-based teaching methodology and empathy in ethics education. Int Christian Community Teacher Educ J 2018;13(2):3 [FREE Full text]
47. Chen MC, Kao HS. Research of community knowledge incorporated into a board game: development and application of the board game happiness gueilai. Humanities Soc Sci Res 2022;16(2):1-31 [FREE Full text]
48. Pogge EK, Davis LE. Fostering mindfulness in continuing pharmacy education using a board game: initial experiences and perceptions. J Contin Educ Health Prof 2021;41(3):230-233 [FREE Full text] [doi: 10.1097/CEH.0000000000000351] [Medline: 33929360]
49. Hojat M, Gonnella JS, Nasca TJ, Mangione S, Vergare M, Magee M. Physician empathy: definition, components, measurement, and relationship to gender and specialty. Am J Psychiatry 2002;159(9):1563-1569 [FREE Full text] [doi: 10.1176/appi.ajp.159.9.1563] [Medline: 12202278]
50. Cheng JF, Lai YM, Livneh H, Tsai TY. Establishing reliability and validity of the Chinese version of the jefferson scale of empathy (healthcare providers version). Nurs J 2011;58(2):41-48 [FREE Full text] [Medline: 21455891]
51. Braun V, Clarke V. Using thematic analysis in psychology. Qual Res Psychol 2008;3(2):77-101 [FREE Full text] [doi: 10.1191/1478088706qp063oa]
52. Xu X, Shen W, Atiquil Islam AYM, Zhou Y. A whole learning process-oriented formative assessment framework to cultivate complex skills. Humanit Soc Sci Commun 2023;10(1):1-15 [FREE Full text] [doi: 10.1057/s41599-023-02200-0]
53. Collier-Sewell F, Atherton I, Mahoney C, Kyle RG, Hughes E, Lasater K. Competencies and standards in nurse education: the irresolvable tensions. Nurse Educ Today 2023;125:105782 [FREE Full text] [doi: 10.1016/j.nedt.2023.105782] [Medline: 36921541]
54. Araújo AA, Kalinowski M, Paixao M, Graziotin D. Towards emotionally intelligent software engineers: understanding students' self-perceptions after cooperative learning experience. 2025 Presented at: Proc IEEE/ACM CHASE 2025; June 24-26, 2025; Manhattan, New York City URL: https://doi.org/10.1109/CHASE66643.2025.00035 [doi: 10.1109/chase66643.2025.00035]
55. Park H, Cho H. Effects of nursing education using films on perception of nursing, satisfaction with major, and professional nursing values. J Nurs Res 2021;29(3):e150 [FREE Full text] [doi: 10.1097/JNR.0000000000000433] [Medline: 33867508]
56. Choenarom C, Samputtanon J. The effectiveness of a question-embedded movie clips learning program in nursing students: a quasi-experimental pretest-posttest study. JMIR Nurs 2025;8:e71111 [FREE Full text] [doi: 10.2196/71111] [Medline: 40637259]
57. Benko E, Peršolja M. Nursing students' views of the impact of geriatric role-play workshops on professional competencies: survey. BMC Nurs 2023;22(1):203 [FREE Full text] [doi: 10.1186/s12912-023-01373-y] [Medline: 37316872]
58. Cho MK, Kim MY. Effectiveness of simulation-based interventions on empathy enhancement among nursing students: a systematic literature review and meta-analysis. BMC Nurs 2024;23(1):319 [FREE Full text] [doi: 10.1186/s12912-024-01944-7] [Medline: 38734606]
59. Huang HP, Tien Y, Lin YC, Yu IC, Chien NH. Effects of empathy mapping and mini-simulation on second-year nursing students' empathy and communication self-confidence: a quasi-experimental study. BMC Med Educ 2025;25(1):109 [FREE Full text] [doi: 10.1186/s12909-025-06686-x] [Medline: 39844127]

Abbreviations

CVI: content validity index
LTC: long-term care
SRQR: Standards for Reporting Qualitative Research
TREND: Transparent Reporting of Evaluations with Nonrandomized Designs

XSL•FO
RenderX

Edited by A Stone; submitted 24.Sep.2025; peer-reviewed by Y Luo, M Chatzimina; comments to author 03.Apr.2026; revised version received 24.May.2026; accepted 26.May.2026; published 08.Jul.2026.

Please cite as:
Chueh KH, Chen JW, Lee LA, Hsu HI, Lin YJ, Wang HT
Youth-Elder Co-Learning Model in Psychiatric Long-Term Care Education: Mixed Methods Evaluation of Communication and Empathy Outcomes
JMIR Med Educ 2026;12:e82812
URL: https://mededu.jmir.org/2026/1/e82812
doi:10.2196/82812
PMID:42418241

©Ke-Hsin Chueh, Jeng-Wen Chen, Li-Ang Lee, Hsiang-I Hsu, Yen-Ju Lin, Hsiao-Tzu Wang. Originally published in JMIR Medical Education (https://mededu.jmir.org), 08.Jul.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Education, is properly cited. The complete bibliographic information, a link to the original publication on https://mededu.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Andragogic Model Curriculum for One-Year ACGME-Accredited Fellowship Programs: Single-Center Educational Improvement Project

Romy Yun[1], MD; Allyson Farran[2], MD; Echo Rowe[1], MD; Dina Sheira[3], BA; Christian Jackson[1], MS; Thomas J Caruso[1], MD, PhD

[1]Department of Anesthesiology, Perioperative, and Pain Medicine, Stanford University School of Medicine, Center for Academic Medicine, 453 Quarry Road, MC 5663, Palo Alto, CA, United States

[2]Department of Anesthesiology, Perioperative, and Pain Medicine, University of Alberta School of Medicine, Edmonton, AB, Canada

[3]Stanford University School of Medicine, Stanford, CA, United States

Corresponding Author:
Romy Yun, MD
Department of Anesthesiology, Perioperative, and Pain Medicine, Stanford University School of Medicine, Center for Academic Medicine, 453 Quarry Road, MC 5663, Palo Alto, CA, United States

Abstract

Background: The number of 1-year Accreditation Council for Graduate Medical Education (ACGME) fellowships continues to grow. The ACGME recommends a holistic curriculum with nonclinical areas, inclusive of educational sessions. Given the competing demands between clinical skill development, educational pursuits, and work-hour restrictions, we propose an andragogic curriculum using pediatric anesthesiology as the model fellowship.

Objective: The primary objective was to improve fellows' perceptions of their educational experience during their fellowship year after implementing an andragogic holistic curriculum. Secondary objectives assessed improvements in diversity, equity, and inclusion (DEI) training and resources.

Methods: This was a single-center educational improvement project completed at Lucile Packard Children's Hospital Stanford. Data were collected between 2014 and 2024. The new curriculum was introduced in 2021 - 2022 and involved 12 different teaching modalities rooted in andragogic principles. A statistical process control p-chart was used to analyze the primary outcome based on the ACGME annual program evaluation. Outcomes were analyzed using censored regression modeling or a *t* test, depending on the presence of ceiling effects.

Results: From 2014 to 2024, 58 of 60 pediatric anesthesiology fellows completed the ACGME survey. A break in the statistical process control p-chart for educational content scores occurred during 2021 - 2022, when the new curriculum was introduced. The mean difference was 0.89 (P<.001). Scores in DEI improved (mean difference 0.52; P=.03), and no difference was noted in resources (mean difference −0.13; P=.98).

Conclusions: Introduction of an andragogic curriculum into a pediatric anesthesiology fellowship program was associated with more favorable perceptions of educational content and DEI training.

(JMIR Med Educ 2026;12:e81570) doi:10.2196/81570

KEYWORDS

education; medical; curriculum; learning; fellowship

Introduction

Fellowship programs have expanded to meet the demand for subspecialized physicians [1]. One-year Accreditation Council for Graduate Medical Education (ACGME) programs continue to increase, with 75 currently registered [2]. Historically, fellowship training has been taught through clinical apprenticeships and didactics, which are resource and time intensive [3,4]. While subspecialty programs yield clinical competency, expertise in nonclinical areas has been less emphasized.

The ACGME defines core competencies and specialty-specific milestones for fellowship programs. Expanded ACGME milestones include quality improvement; diversity, equity, and inclusion (DEI) training; well-being; and dedicated resources for scholarly activity. Despite short durations of fellowships compared to residencies, ACGME programmatic requirements are relatively similar, creating challenges for program directors seeking to meaningfully incorporate these requirements [5,6].

XSL•FO
RenderX

Given the increasing number of ACGME milestones, an andragogic approach may provide an effective framework and incorporate individual learning styles across trainees [3]. This approach leverages the previous experiences of a learner and puts them in a more active role within the education process. Using established curricular frameworks from Knowles' Andragogic Model and Kolb's Experiential Learning Cycle, we developed an enhanced fellowship curriculum that leverages andragogic principles to integrate ACGME clinical and nonclinical curricular recommendations into a single year. This approach tailors the experience to mature learners with an active role in their educational experience [3,7].

Although this study describes the integration of all ACGME competencies into a 1-year curriculum using the Pediatric Anesthesiology Fellowship as the model, the components are relatable to other ACGME programs. The primary aim of this curricular redesign was to improve fellows' perceptions of their educational experience.

Methods

Setting

This was an educational improvement project completed between 2014 and 2024 at Stanford University School of Medicine within the ACGME-approved Pediatric Anesthesiology Fellowship. The fellows practice at Lucile Packard Children's Hospital Stanford, a level-one trauma center with 20 operating rooms. The training occurs over a single year, with approximately 8 fellows in postgraduate year five or greater.

Intervention

During the 2021 - 2022 year, an andragogic curriculum was integrated into the fellowship program. Before this new curriculum, fellow education consisted of after-hours didactics, one simulation session, and an apprenticeship model of clinical care. The new curricular framework relied on Knowles' Andragogic Model combined with Kolb's Experiential Learning Cycle [8,9]. These models were preferred for their established application in health professions education and direct alignment with the characteristics of the fellowship learner population. Fellows enter the program as fully qualified physicians with professional identities, diverse prior clinical experiences, and intrinsic motivation to achieve subspeciality competence. Knowles' Andragogic Model was chosen for its emphasis on the role of the independent self-directed learner, prioritized problem-centered approaches, and application of learning into daily life [8]. Kolb's Experiential Learning Cycle was selected as a complementary framework since it provides a mechanistic account of how learning occurs within the experiential environment [9,10]. While Knowles' principles guided the overall curriculum design and ensured that it was self-directed, problem-centered, and responsive to fellows' prior experience, Kolb's cycle was used to structure the learning process itself. Specifically, the curriculum was designed for concrete clinical experiences to be systematically followed by structured reflection, didactic conceptualization, and opportunities for active experimentation, rather than solely relying on experience.

Curricular components with various delivery platforms were implemented that prioritized the unique characteristics of fellowship-level learners and the structured process by which direct clinical experiences are converted into meaningful learning. A 12-component syllabus that addressed clinical and nonclinical ACGME requirements was developed, which included experiential, self-directed, problem-based, and asynchronous remote learning platforms (Table 1) [11]. Within these 12 learning modalities and trainings, each curricular component was curated around chapters in a foundational pediatric anesthesiology textbook and core knowledge requirements dictated by ACGME pediatric anesthesiology milestones [12]. Learning modalities within the new curriculum included directed readings, didactic lectures, quality improvement training, problem-based learning discussions, simulation, workshops, professional and behavioral health series, doctoring with care series, oral board preparation series, DEI training, journal clubs, and self-assessment in addition to daily clinical care responsibilities (Multimedia Appendix 1). Andragogic principles such as self-concept, need to know, problem-centered orientation, prior experience, readiness to learn, and internal motivation were applied to the learning modalities within the new curriculum (Table 1). The principle of problem-centered orientation was embedded through case-based conferences, simulations, and workshops. Internal motivation was applied to coaching conversations, and reflections focused on growth and ownership of a quality improvement training module and a quality improvement project of their choosing. Didactic sessions and directed readings highlighted the need to know and readiness to learn principles with explicit framing of clinical context. Additionally, the stages within Kolb's Experiential Learning Cycle (concrete experience, reflective observation, abstract conceptualization, and active experimentation) were used within the curricular components to ensure that clinical experiences were followed by thoughtful reflection, feedback, and conceptualization (Table 1).

XSL•FO
RenderX

Table . Integration of andragogic curricular components.

Knowles' principles	Kolb's stages	Curricular component	Description	Level of integration
Need to know and readiness to learn	Abstract conceptualization	Directed readings	• Weekly topic readings correlated to organ-based learning blocks from core textbook	Asynchronous
Need to know and readiness to learn	Abstract conceptualization	Didactic lectures[a]	• Approximately 25 midday lectures with topics aligned with readings/questions in organ-based blocks	In-person or remote
Internal motivation	Abstract conceptualization	QI[b] training	• QI web-based modules • Morbidity and mortality monthly conferences	Asynchronous and in-person or remote
Problem-centered orientation and prior experience	Abstract conceptualization	Problem-based learning discussions[a]	• Case-based discussions correlated to organ-based learning blocks	In-person or remote
Problem-centered orientation and prior experience	Abstract conceptualization and active experimentation	Simulation[a]	• Case-based scenarios practicing technical and nontechnical skills • In situ, virtual, and extended reality simulations offered	In-person and experiential
Problem-centered orientation and readiness to learn	Abstract conceptualization	Workshops	• Hands-on educational workshops • Examples: pacemaker, point of care ultrasound, regional anesthesia cadaver lab, vascular access	In-person
Internal motivation	Reflective observation	PBHS[c] and DWCS[a,d]	• DWCS: Opportunities for reflection, self-compassion, community-building • PBHS: Roundtable discussions on physician well-being, finances, nutrition, sleep deprivation, burnout	In-person
Self-concept and internal motivation	Abstract conceptualization and reflective observation	Oral board preparation series	• Interactive sessions geared toward oral board preparation • Led by current faculty oral board examiners	In-person and interactive learning
Internal motivation and need to know	Reflective observation	DEI[e] training[a]	• DEI course sequence occurring every other month • Topic examples: health disparities, use of perioperative interpreters, gender diverse patients, racial disparities in medicine	In-person and introspective learning
Prior experience and need to know	Abstract conceptualization and reflective observation	Journal clubs[a]	• Group-based discussions led by trainees on selected journal articles	In-person or remote and flipped classroom

XSL•FO
RenderX

Knowles' principles	Kolb's stages	Curricular component	Description	Level of integration
Self-concept	Reflective observation	Self-assessment	• Assigned practice questions aligned with readings in organ-based learning blocks • Mid and end of year self-assessment exams similar to the pediatric anesthesiology board exam questions • Board specific in-training exam	Asynchronous and in-person or remote

[a]During workday hours.

[b]QI: quality improvement.

[c]PBHS: professional and behavioral health series.

[d]DWCS: doctoring with care series.

[e]DEI: diversity, equity, and inclusion.

Outcomes and Data Collection

The primary outcome assessed fellows' perception of educational content. This was selected because it demonstrated a broad representation of nonclinical components highlighted by ACGME guidelines. Secondary outcomes explored DEI training and programmatic resources. The DEI category was selected since it assessed reflective practice and inclusivity, and the resources category assessed the program's educational infrastructure.

Outcomes were measured with the ACGME trainee survey, which assesses education quality and compliance with ACGME requirements [13]. Sections within the ACGME survey are standardized into preset categories. The educational content subsections include a broad range of questions on well-being, burnout, and scholarly activities. The resources subsections assess protected time for educational and professional activities and overall institutional support. The DEI subsections include questions regarding inclusivity, diversity of patient populations, and engagement of a program's DEI efforts. Survey responses are based on a 5-point Likert scale and are annually completed by the fellows. ACGME fellow survey results from 2014 to 2024 were collected from the educational content and resources categories on the survey. Data were collected from 2019 to 2024 regarding DEI training, a new category introduced in 2019. Each category contained several subsection scores, which were averaged by the ACGME to provide the overall category mean.

Statistical Analysis

To analyze the primary outcome, a statistical process control p-chart monitored a change in fellows' responses to the educational content subsection of the ACGME survey. The Western Electric rules assessed special cause variation [14]. Rule one in Western Electric states that special cause variation occurs when a data point falls outside the 3 SD upper control limit from the center line. This occurred in the 2021 - 2022 academic year, leading to the establishment of a new center line. Mean outcomes were assessed using censored regression modeling if there was a ceiling effect or a *t* test if no ceiling effect. A *P* value less than .05 was considered significant. All analyses were performed using Excel (Microsoft Corporation) and R statistical software (version 4.5.1; R Foundation for Statistical Computing; *censReg* package) [15].

Ethical Considerations

This project received a waiver of consent by the Stanford University Institutional Review Board (eProtocol No. 87506). Data were deidentified by the ACGME before reporting the results to the program. There was no compensation given to participants.

Results

Demographics

Of the 60 pediatric anesthesiology fellows, 58 completed the ACGME survey from 2014 to 2024, with an overall response rate of 97%. During the data collection period, 56% (n=34) of the fellows identified as female and 44% (n=26) as male.

Primary Outcome

Special cause variation was noted in the statistical process control p-chart for mean scores for educational content during the 2021 - 2022 academic year, which coincides with the implementation of the new curriculum. Given a ceiling effect, censored regression was used to estimate the mean difference. A mean difference of 0.89 (SE 0.27) from 2014 to 2021 was observed compared to 2022 - 2024 ($P<.001$).

Secondary Outcomes

The mean difference between DEI training scores did not exhibit a ceiling effect; therefore, a *t* test was used to analyze the results. For DEI training (n=36), we demonstrated a difference of 0.52 (SE 0.2; $P=.03$). Fellows' perceptions of program resources (n=58) had a ceiling effect, and no estimated difference was observed (mean difference −0.13, SE 0.24; $P=.98$).

Discussion

Holistic integration of all ACGME curricular sections is challenging to implement in shorter-duration training programs. This project assessed learners' perception of an enhanced pediatric anesthesiology fellowship curriculum. We relied on

XSL•FO
RenderX

multiple andragogic principles on established curricular frameworks to guide the integration of recommended ACGME components. The curriculum improved fellows' perception of their educational experience and their DEI training, with no difference in perception of programmatic resources.

The key components of the enhanced curriculum expanded on didactics to include several curricular threads grounded in andragogic learning theories. Self-directed, asynchronous learning improves trainee engagement and provides a sense of agency [8,16,17]. We used a set of professionally created online modules to teach quality improvement methodology, which fellows completed at their own pace. Additionally, self-directed readings with affiliated questions were aligned with organ-based learning blocks throughout the year. Experiential learning provided opportunities to build knowledge and skill sets in realistic clinical scenarios with structured feedback via case-based problem-based learning discussions, procedural skills workshops, and simulations [18]. For our simulation sessions, the fellows experienced in situ simulations that occurred during typical working hours in their practice environment. Because integration of curriculum components into the workday is associated with improved trainee perception of work-life balance, several curricular components were held during daytime hours [19,20]. Additionally, a professionalism and behavioral health monthly thread was developed to improve resiliency and self-care [21].

Implementation of DEI curricula remains challenging with limited evidence on best practices to guide effective strategies to fulfill ACGME common program requirements [22]. In this report, the improvement in DEI training was due to two new core curricular components. These components were influenced by special study modules, which are programs that intentionally provide educational opportunities to explore topics beyond the core curriculum [23]. Understanding that adult learners value critical thinking, introspection, and clinical application, the DEI sessions consisted of monthly midday lectures that challenged fellows' implicit biases and explored social determinants of health within diverse patient populations. By also including dedicated time for self-reflection, called doctoring with care, fellows were provided the opportunity to process challenging patient and personal interactions [21]. Implementation of well-designed modules alongside interactive learning environments increased fellow satisfaction within these domains. However, the enhanced curriculum did not demonstrate improvements in programmatic resources. Although surprising, variation within subsections in this category may account for the lack of difference.

This project had several limitations. First, inherent to most improvement projects, there was no control group. However, the demographic constituency of the fellows did not change during the project period. Second, while standardized across the country, there is annual variation in the ACGME subsection questions. Despite this variance, the ACGME routinely uses its survey to highlight programmatic trends, similar to our methodology. Third, this enhanced curriculum was applied to a single fellowship. Although we believe this specific andragogic approach would be beneficial for other shorter-duration programs, this hypothesis is untested. Fourth, there was no method to account for natural shifts that occur over time, such as ACGME expectations, institutional culture, or national attention to DEI, that could explain these improvements. Fifth, there is a wide range of other unmeasured items that could impact trainee experience not captured by the ACGME survey. Lastly, our data relied on self-reported perception scores from the ACGME survey. Although we intentionally chose a standardized measure to improve the generalizability of our findings to other national training programs, we did not include objective educational outcomes such as examination scores. Future studies should investigate if these changes translate into measurable improvements in fellow performance and whether these gains are practically meaningful.

Introduction of an andragogic curriculum to the Pediatric Anesthesiology fellowship program was associated with more favorable perceptions of the educational experience. Adoption of aspects of this curricular framework may improve the educational content of other shorter-duration training programs. Future studies will explore additional curricular enhancements and their effects on 1-year training programs beyond pediatric anesthesiology to better control for potential confounders and natural variations within ACGME expectations.

Acknowledgments

The authors would like to recognize the following for their dedication to fellowship education: Travis Reece-Nguyen, MD; Andrea Murray, MD; Ellen Wang, MD; Michael Lennig, MD; Mohammad Esfahanian, MD; Asheen Rama, MD; Louise Furukawa, MD; Rita Agarwal, MD; Bernadett Mahaney; and Jessica Martinez.

Funding

The authors report no external funding source for this study.

Authors' Contributions

Conceptualization: RY, ER, TJC
Data curation: RY, AF, CJ, TJC
Investigation: RY, AF, TJC
Methodology: RY, TJC
Writing – original draft: RY, AF

XSL•FO
RenderX

Writing – review & editing: RY, ER, DS, TJC
Visualization: AF, DS
Formal analysis: CJ
Project administration: TJC
Supervision: TJC

Conflicts of Interest

None declared.

Multimedia Appendix 1
Learning modalities within the new curriculum.
[PNG File, 168 KB - mededu_v12i1e81570_app1.png]

References

1. 7,700 resident physicians placed into fellowship training positions through the NRMP's medicine and pediatric specialties match. National Resident Matching Program. 2023 Nov 29. URL: https://www.nrmp.org/about/news/2023/11/7700-resident-physicians-placed-into-fellowship-training-positions-through-the-nrmps-medicine-and-pediatric-specialties-match [accessed 2024-10-27]
2. Accreditation data system (ADS). Accreditation Council for Graduate Medical Education. URL: https://apps.acgme.org/ads/public [accessed 2024-11-02]
3. Caruso TJ, Qian J, Lawrence K, Armstrong-Carter E, Domingue BW. From Socrates to virtual reality: a historical review of learning theories and their influence on the training of anesthesiologists. J Educ Perioper Med 2020 Apr 1;22(2):E638. [doi: 10.46374/volxxii-issue2-Caruso] [Medline: 32939366]
4. Ahmad M, Tariq R. History and evolution of anesthesia education in United States. J Anesth Clin Res 2017;8(6). [doi: 10.4172/2155-6148.1000734]
5. Ambardekar AP, Eriksen W, Ferschl MB, et al. A consensus-driven approach to redesigning graduate medical education: the Pediatric Anesthesiology Delphi Study. Anesth Analg 2023 Mar 1;136(3):437-445. [doi: 10.1213/ANE.0000000000006128] [Medline: 35777829]
6. Cladis FP, Waisel D. Creating the pediatric anesthesiology consultant: you can't have your cake and eat it too (unless you change the recipe). Anesth Analg 2023 Mar 1;136(3):434-436. [doi: 10.1213/ANE.0000000000006218] [Medline: 36806231]
7. Gaiser RR. The adult learner: is it necessary to understand for teaching in anesthesiology. Int Anesthesiol Clin 2010;48(3):1-12. [doi: 10.1097/AIA.0b013e3181e6e7f3] [Medline: 20616634]
8. Knowles M. S. Andragogy in Action: Applying Modern Principles of Adult Learning: Jossey-Bass; 1984.
9. Kolb DA, Boyatzis RE, Mainemelis C. Experiential learning theory: previous research and new directions. In: Sternberg RJ, Zhang L, editors. Perspectives on Thinking, Learning, and Cognitive Styles: Routledge; 2014:227-247. [doi: 10.4324/9781410605986-9]
10. Yardley S, Teunissen PW, Dornan T. Experiential learning: AMEE Guide No. 63. Med Teach 2012;34(2):e102-e115. [doi: 10.3109/0142159X.2012.650741] [Medline: 22289008]
11. Program requirements, FAQs, and applications. Accreditation Council for Graduate Medical Education. URL: https://www.acgme.org/specialties/anesthesiology/program-requirements-and-faqs-and-applications [accessed 2024-11-02]
12. Cote CJ, Lerman J, Anderson B, editors. A Practice of Anesthesia for Infants and Children, 7th edition: Elsevier; 2024.
13. Survey information. ACGME International. URL: https://www.acgme-i.org/accreditation-process/survey-information/ [accessed 2026-06-03]
14. Kurth CD. Introducing quality improvement. Paediatr Anaesth 2013 Jul;23(7):569-570. [doi: 10.1111/pan.12167] [Medline: 23738595]
15. Henningsen A. censReg: censored regression (Tobit) models. The Comprehensive R Archive Network. 2024 May. URL: https://cran.r-project.org/web/packages/censReg/index.html [accessed 2024-11-02]
16. Burford B, Whittle V, Vance GHS. The relationship between medical student learning opportunities and preparedness for practice: a questionnaire study. BMC Med Educ 2014 Oct 21;14(1):223. [doi: 10.1186/1472-6920-14-223] [Medline: 25331443]
17. Trivedi SP, Rodman A, Eliasz KL, Soffler MI, Sullivan AM. Finding the right combination for self-directed learning: a focus group study of residents' choice and use of digital resources to support their learning. Clin Teach 2024 Dec;21(6):e13722. [doi: 10.1111/tct.13722] [Medline: 38233893]
18. Mukhalalati BA, Taylor A. Adult learning theories in context: a quick guide for healthcare professional educators. J Med Educ Curric Dev 2019 Apr 10;6:2382120519840332. [doi: 10.1177/2382120519840332] [Medline: 31008257]
19. Patel RS, Sekhri S, Bhimanadham NN, Imran S, Hossain S. A review on strategies to manage physician burnout. Cureus 2019 Jun 3;11(6):e4805. [doi: 10.7759/cureus.4805] [Medline: 31404361]

XSL•FO
RenderX

20. Moen P, Lam J, Ammons S, Kelly EL. Time work by overworked professionals: strategies in response to the stress of higher status. Work Occup 2013 May 1;40(2):79-114. [doi: 10.1177/0730888413481482] [Medline: 24039337]
21. Szuster RR, Onoye J, Matsu C. Presence, Resilience, and Compassion Training in Clinical Education (PRACTICE): a follow-up evaluation of a resident-focused wellness program. J Grad Med Educ 2023 Apr;15(2):237-243. [doi: 10.4300/JGME-D-22-00422.1] [Medline: 37139203]
22. Huang G, Wang C, Rosenzweig T, Moquin R, Markowitz S, Ablordeppey E. Enhancing anesthesiology: a survey of diversity, equity, and inclusion in residency curricula. J Educ Perioper Med 2025 Apr 8;27(1):E736. [doi: 10.46374/VolXXVII_Issue1_Ablordeppey] [Medline: 40207081]
23. Ogut E, Yildirim FB, Senol Y, Senol AU. Comprehensive evaluation of the educational impact and effectiveness of specialized study modules in cross-sectional anatomy: a study on student engagement and learning outcomes. BMC Med Educ 2025 Apr 10;25(1):514. [doi: 10.1186/s12909-025-07050-9] [Medline: 40211255]

Abbreviations

ACGME: Accreditation Council for Graduate Medical Education
DEI: diversity, equity, and inclusion

Edited by W Jerjes; submitted 02.Aug.2025; peer-reviewed by E Ogut, J Wilkinson; revised version received 17.May.2026; accepted 19.May.2026; published 23.Jun.2026.

Please cite as:
Yun R, Farran A, Rowe E, Sheira D, Jackson C, Caruso TJ
Andragogic Model Curriculum for One-Year ACGME-Accredited Fellowship Programs: Single-Center Educational Improvement Project
JMIR Med Educ 2026;12:e81570
URL: https://mededu.jmir.org/2026/1/e81570
doi:10.2196/81570

© Romy Yun, Allyson Farran, Echo Rowe, Dina Sheira, Christian Jackson, Thomas J Caruso. Originally published in JMIR Medical Education (https://mededu.jmir.org), 23.Jun.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Education, is properly cited. The complete bibliographic information, a link to the original publication on https://mededu.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Evaluating Large Language Model–Based Automated Scoring in a Voice-Based Virtual Standardized Patient Platform for Medical Students: Cross-Sectional Agreement Study

Xiaoxing Gao[1], MD; Xiaoming Huang[2], MD; Rongrong Hu[3], MD; Li Zhang[4], MD; Huiting Liu[5], MD; Bingqing Zhang[2], MD; Chong Wei[6], MD; Wei Qiu[7], MD; Mengyu Zhang[8], MD; Xuefeng Sun[1,9], MD

[1]Department of Pulmonary and Critical Care Medicine, Peking Union Medical College Hospital, 1 Shuaifuyuan, Dongcheng District, Beijing, China
[2]Department of General Internal Medicine, Peking Union Medical College Hospital, Beijing, China
[3]Department of Nephrology, Peking Union Medical College Hospital, Beijing, China
[4]Department of Rheumatology, Peking Union Medical College Hospital, Beijing, China
[5]Department of Infectious Disease, Peking Union Medical College Hospital, Beijing, China
[6]Department of Haematology, Peking Union Medical College Hospital, Beijing, China
[7]Department of Medical Oncology, Peking Union Medical College Hospital, Beijing, China
[8]Department of Neurology, Peking Union Medical College Hospital, Beijing, China
[9]Department of Internal Medicine, Peking Union Medical College Hospital, Beijing, China

Corresponding Author:
Xuefeng Sun, MD
Department of Pulmonary and Critical Care Medicine, Peking Union Medical College Hospital, 1 Shuaifuyuan, Dongcheng District, Beijing, China

Abstract

Background: Large language model (LLM)–powered virtual standardized patients (VSPs) enable scalable clinical skills practice, but the validity of AI-generated scores relative to faculty ratings remains unclear.

Objective: This study aimed to assess agreement between LLM-generated and faculty ratings of history-taking and communication performance and to examine the influence of rater and case heterogeneity.

Methods: In this cross-sectional study, 92 fourth-year medical students completed one of three 15-minute voice-based VSP cases (fever, diarrhea, and cough). Ten blinded faculty raters scored performance (0‐100 points total; 0‐50 points per domain). AI scores were generated by DeepSeek-V3 using a calibrated prompt. Agreement was evaluated using mixed-effects models, intraclass correlation coefficients (ICC [2,1]), Spearman correlations, mean absolute error (MAE), Bland-Altman analysis, and variance partition coefficients (VPC).

Results: Median total scores were similar for AI and faculty (median 93.0, IQR 89.0-95.0 vs median 94.0, IQR 91.0-95.0). Rater variability accounted for 37% of residual variance in faculty total scores (VPC=0.37). AI total scores were positively associated with faculty total scores (β=0.37, 95% CI 0.26‐0.48; *P*<.001; Spearman ρ=0.50, 95% CI 0.34‐0.65). Absolute agreement was moderate (ICC[2,1]=0.51, 95% CI 0.34‐0.65), with MAE of 3.11 points. Mixed-effects Bland-Altman analysis showed a small, not statistically significant mean bias (1.26 points, 95% CI −0.48 to 3.01; *P*=.16) and 95% limits of agreement from −4.95 to 7.48 (width=12.43 points), with proportional bias (β_proportional bias=−0.55; *P*<.001). Agreement was stronger for information gathering (β=0.46; ρ=0.49; ICC=0.54; VPC=0.23) than for communication (β=0.27; ρ=0.28; ICC=0.29; VPC=0.52). A sensitivity analysis in the lowest quartile showed attenuated but consistent agreement (ICC=0.38).

Conclusions: LLM-based scoring in a VSP showed moderate agreement with faculty ratings, performing better for information gathering than for communication. Due to rater and case heterogeneity, ceiling effects, and proportional bias, this method is suitable for formative use and enhanced sampling in programmatic assessment but not for independent, high-stakes summative decisions.

(JMIR Med Educ 2026;12:e95578) doi:10.2196/95578

KEYWORDS

large language model; virtual standardized patient; information gathering; communication skills; rater variability; mixed-effects model

XSL•FO
RenderX

Introduction

Background

History-taking and physician-patient communication are core competencies in undergraduate medical education. Traditional teaching relies on faculty-facilitated role-play and standardized patients (SPs), which are resource-intensive and constrained by faculty time and SP availability. As a result, students often have limited opportunities for repeated practice and timely feedback.

Virtual standardized patients (VSPs) are computer-based simulations designed to portray patients with defined clinical presentations and respond dynamically to learner inquiries. They offer a scalable supplement or alternative. Virtual patients have long been used to support history-taking and clinical reasoning, and reviews show positive learner perceptions and modest gains in knowledge and skills compared with traditional methods [1-3]. Earlier systems largely used branching logic and rule-based scoring. Recent large language models (LLMs) can support free-text, contextually appropriate dialogue, enabling more natural and flexible patient simulations [4].

Beyond dialogue, LLMs can generate automated scores and feedback, potentially increasing feedback frequency while reducing faculty workload [5]. However, the usefulness of automated assessment depends on validity and reliability. For summative uses (eg, progression decisions), strong agreement with expert judgment and robust validity evidence are essential; for formative use, moderate agreement may be acceptable if feedback is timely, specific, and actionable [6].

Notably, even trained human raters in objective structured clinical examinations (OSCEs) and workplace-based assessments often show only moderate interrater reliability and substantial variability in severity and leniency [7,8]. Therefore, AI-human agreement should be interpreted in the context of human-human variability, rather than assuming a flawless human "gold standard." Few empirical studies have systematically compared LLM-driven scoring of clinical communication with human examiners in authentic teaching settings—especially within integrated VSP platforms [5,9,10].

We developed an LLM-based VSP system for voice-based history-taking and communication practice. Students conducted spoken interviews with an LLM-driven virtual patient; the system then generated scores and narrative feedback. We evaluated the system by comparing LLM-generated scores with faculty ratings across multiple instructional groups and clinical cases.

This study aimed to evaluate the agreement between LLM-generated and faculty ratings of history-taking and communication performance in an LLM-powered VSP platform and to examine how rater and case heterogeneity influenced agreement to determine appropriate educational use boundaries for LLM-based scoring.

Research Questions

This study addressed three questions: (1) How closely do LLM total scores agree with faculty ratings? (2) How do rater and case factors affect AI-human agreement? (3) Is the psychometric agreement between LLM-based scoring and faculty ratings sufficient to support formative and/or summative assessment uses?

Methods

Study Design and Setting

This cross-sectional observational study was conducted at Peking Union Medical College Hospital. Each student completed a single VSP encounter via a voice-based dialogue. In each encounter, the VSP acted as the patient, while students role-played the physician to conduct a history-taking interview. Encounters were limited to 15 minutes. Immediately after the encounter, the supervising faculty member provided ratings while blinded to AI-generated scores.

Participants

Ninety-two fourth-year medical students enrolled in a diagnostics course participated. Prior to the experiment, all participants had completed 7 weeks of history-taking training, including 1 human SP practice session per week, and they were offered voluntary access to an LLM-powered VSP platform aligned with the SP curriculum. Students could initiate unlimited, untimed multiturn interviews with any case, at any time, outside scheduled SP sessions.

As preparation for the end-of-term history-taking examination, we conducted this VSP-based exercise. Students were informed that this was an AI-simulated practice session; they were not told that AI scores would be compared with faculty ratings for research. Ten faculty raters scored students using a standardized rubric, with brief orientation but no formal calibration.

LLM-Powered VSP System

The VSP system used DeepSeek-V3 (DeepSeek AI, accessed via API, temperature=0.0) for scoring, Doubao-1.5 (ByteDance) for patient dialogue, paraformer-realtime-v2 (Alibaba) for automatic speech recognition (ASR), and CosyVoice-v2 (Alibaba) for text-to-speech synthesis. Prior to deployment, we iteratively calibrated the LLM scoring prompt using 45 historical records from the 2024 academic year (independent cohort, no overlap in students, cases, or raters). The five-stage calibration involved (1) rubric-aligned prompt drafting, (2) initial testing on 15 records, (3) item-level discrepancy analysis, (4) iterative refinement targeting systematic errors, and (5) validation on the held-out 30 records. The prompt was frozen before deployment. The complete prompt and detailed calibration procedure are provided in Multimedia Appendix 1.

Three clinical cases (fever, diarrhea, and cough) were developed by the same clinical education team to represent common outpatient presentations of comparable complexity and expected duration (15 min). Participants were randomly assigned to one of 3 cases using a simple randomization procedure (drawing identical paper lots). Group sizes varied (case 1: n=32; case 2: n=35; case 3: n=25) because of random allocation.

Scoring Dimensions and Rubrics

Immediately after each encounter, 2 independent 100-point scores were generated using the same 50:50 weighting for

XSL•FO
RenderX

information gathering and communication. Information gathering (0‐50 points) assessed completeness and relevance of the clinical history through approximately 40 checklist items. Communication (0‐50 points) assessed process quality through 10 criteria, including clinical reasoning, question appropriateness, summarization, transitional language, and humanistic care. Communication criteria were weighted from 3 to 8 points based on clinical importance; the complete rubric is provided in Multimedia Appendix 1.

Statistical Analysis

Analyses were conducted in R (version 4.3.2; R Foundation for Statistical Computing). The unit of analysis was the student (n=92); each student contributed 1 paired set of AI-generated and faculty-assessed scores from a single VSP encounter. Because each encounter was scored by 1 faculty rater, clustering by rater was handled using mixed-effects models with rater-level random intercepts. Given bounded (0‐100) scores with ceiling tendencies, results were summarized using median (25th-75th percentile). All tests were 2-sided (α=.05).

The primary analysis used a linear mixed-effects model: human_total ~ AI_total + case + (1|rater). A random-slope specification (AI_total|rater) was explored; singular fits were interpreted as insufficient information to support between-rater slope variability. Between-rater clustering was summarized using the variance partition coefficient (VPC). Absolute agreement was summarized with intraclass correlation coefficient (ICC[2,1]; 2-way random-effects, single-measure, absolute agreement). A mixed-effects Bland-Altman model estimated mean bias and 95% limits of agreement, with tests for proportional bias. Full model specifications are in Multimedia Appendix 2.

All analyses were repeated for information gathering and communication subscales. Prespecified sensitivity analyses refit the primary model using robust mixed-effects estimation and refit without case adjustment. Additional summaries included Spearman ρ and mean absolute error (MAE).

Student Feedback Survey

After completing the VSP encounter, students were invited to complete an anonymous online feedback questionnaire. The survey assessed 4 dimensions: overall satisfaction (0‐10 scale), perceived convenience of the VSP platform (1‐5 Likert scale), perceived realism compared with human SP interactions (1‐5 Likert scale), and perceived feedback value relative to human SPs (1‐5 Likert scale). Responses were collected via a web-based form (Wenjuanxing) immediately after each encounter. Descriptive statistics (mean and response distribution) were computed for each item. The questionnaire was developed de novo for this study to capture platform-specific user experience and was not adapted from a previously validated instrument.

Ethical Considerations

This study used deidentified student data from routine teaching activities. The Peking Union Medical College Hospital Ethics Committee approved this secondary analysis (I-26PJ0511) and granted a waiver of written informed consent. Students had been informed that their encounters might be used for educational research. The study adhered to the principles of the Declaration of Helsinki.

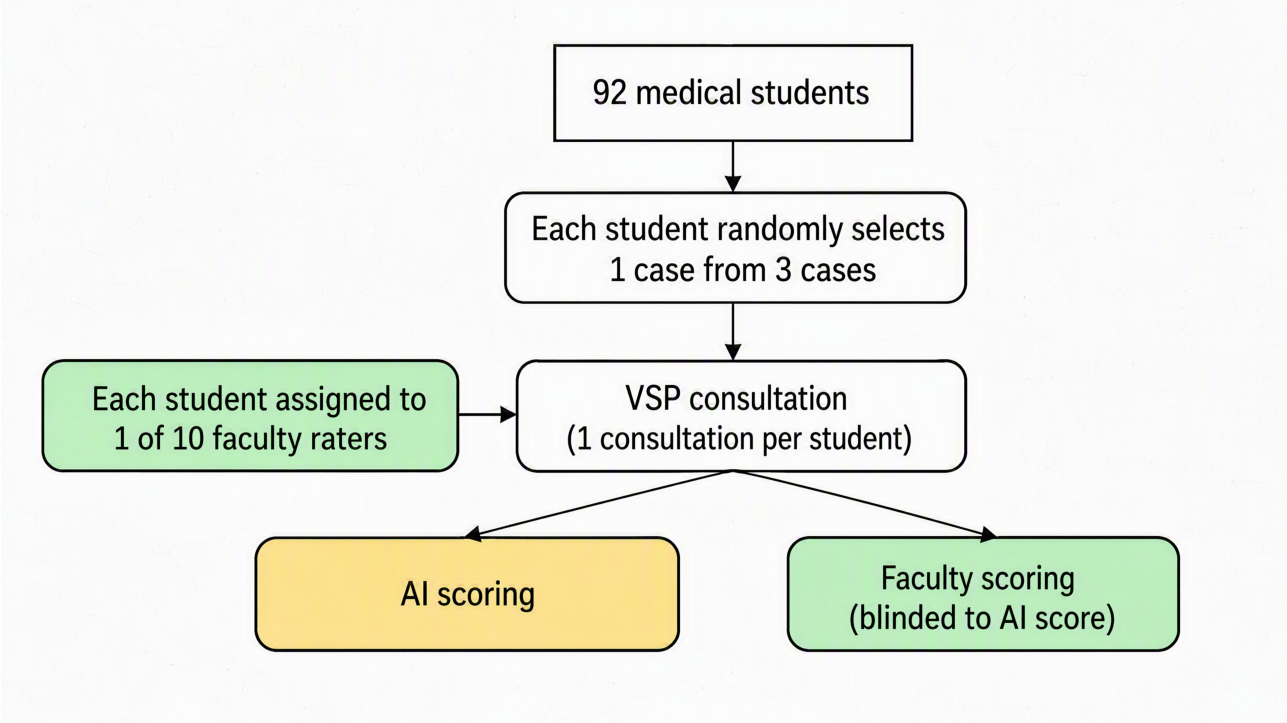
Results

Descriptive Statistics

The study design flowchart is shown in Figure 1. All 92 students successfully completed their VSP encounters without system crashes or encounter restarts. No encounters were excluded due to technical failures. Occasional ASR errors were informally noted by supervising faculty, primarily affecting medical terminology and rapid speech; students had been trained to correct such errors by repeating or rephrasing their questions.

XSL•FO
RenderX

Figure 1. Flowchart of the study design. VSP: virtual standardized patient.



Ninety-two medical students each contributed 1 paired observation and were rated by 1 of 10 independent raters (8 - 10 students each, median 9, IQR 9-10). Students were randomized to 3 cases (case 1: n=32; case 2: n=35; case 3: n=25). Overall scores were high, with evidence of ceiling effects: median AI total score 93.0 (IQR 89.0-95.0), median human total score 94.0 (IQR 91.0-95.0); 17 (18.5%) students scored above 95 on AI totals and 17 (18.5%) on human totals, with 6 (6.5%) students scoring above 95 on both. AI total scores ranged from 77 to 100 and human total scores from 84 to 99. Score distributions were compressed (total-score SDs 3.4-4.6: AI 4.6, human 3.4; subscale SDs 1.5-3.0: AI information 3.0, human information 2.9, AI communication 3.0, human communication 1.5). Median total scores were similar across cases, whereas rater medians varied substantially (AI 89.0 - 94.0; human 87.2 - 95.0), indicating meaningful between-rater severity differences (Table 1). Histograms are provided in Multimedia Appendix 3.

XSL•FO
RenderX

Table . Descriptive statistics of AI and human ratings across competency dimensions.

Group	AI total score	Human total score	AI information score	Human information score	AI communication score	Human communication score
Overall (n=92), median (IQR)	93.0 (89.0-95.0)	94.0 (91.0-95.0)	46.0 (44.0 - 48.0)	46.0 (44.5 - 47.0)	47.0 (45.0 - 49.0)	48.0 (47.0 - 49.0)
Case, median (IQR)						
Case 1 (n=32)	92.0 (89.8-95.2)	94.0 (91.5 - 95.2)	47.0 (45.0 - 48.0)	46.5 (44.9 - 47.2)	46.5 (44.8 - 48.2)	48.0 (47.0 - 49.0)
Case 2 (n=35)	93.0 (89.5 - 95.0)	94.0 (92.0 - 95.0)	46.0 (44.0 - 47.0)	46.0 (45.0 - 46.2)	47.0 (46.0-49.0)	48.0 (46.0 - 49.0)
Case 3 (n=25)	92.0 (89.0 - 94.0)	93.0 (88.5 - 95.0)	45.0 (44.0 - 47.0)	45.0 (40.5 - 47.0)	47.0 (44.0 - 48.0)	48.0 (47.0 - 49.0)
Rater, median (IQR)						
Rater 1 (n=9)	94.0 (91.0 - 95.0)	95.0 (94.0 - 95.0)	47.0 (45.0 - 47.0)	46.0 (46.0 - 47.0)	49.0 (46.0 - 49.0)	48.0 (48.0 - 49.0)
Rater 2 (n=9)	94.0 (92.0 - 96.0)	94.0 (93.0 - 94.0)	48.0 (47.0 - 49.0)	46.0 (44.5 - 47.0)	46.0 (45.0 - 47.0)	48.0 (47.0 - 48.0)
Rater 3 (n=9)	93.0 (91.0 - 95.0)	95.0 (95.0-97.0)	48.0 (46.0 - 49.0)	47.0 (46.0 - 47.0)	47.0 (44.0 - 47.0)	49.0 (49.0 - 49.0)
Rater 4 (n=9)	93.0 (90.0 - 95.0)	91.0 (90.0 - 93.5)	45.0 (44.0 - 47.0)	45.0 (44.5 - 45.0)	48.0 (44.0 - 50.0)	48.0 (46.0 - 49.0)
Rater 5 (n=8)	93.0 (89.0 - 94.0)	94.5 (92.6 - 96.5)	47.0 (45.0 - 48.0)	45.8 (44.0 - 47.8)	46.0 (43.5 - 46.1)	48.8 (48.0 - 49.0)
Rater 6 (n=10)	93.0 (90.5 - 95.0)	94.0 (93.0 - 94.8)	47.0 (45.2 - 47.0)	47.0 (47.0 - 48.0)	47.0 (46.2 - 48.0)	47.0 (46.0 - 47.0)
Rater 7 (n=10)	91.0 (89.2 - 93.5)	87.2 (86.2 - 91.2)	44.0 (41.2 - 45.0)	40.5 (39.0 - 44.5)	47.5 (45.5 - 49.8)	47.0 (46.0 - 47.9)
Rater 8 (n=10)	92.5 (88.2 - 94.8)	91.2 (90.2 - 92.0)	45.0 (43.2 - 45.8)	44.8 (44.0 - 45.8)	48.5 (47.2 - 49.0)	46.5 (46.0 - 47.8)
Rater 9 (n=8)	93.0 (92.0 - 95.2)	95.0 (93.5 - 96.2)	47.0 (46.0 - 47.0)	47.0 (46.8 - 48.0)	47.0 (45.8 - 48.2)	48.0 (46.0 - 49.0)
Rater 10 (n=10)	89.0 (86.5 - 91.2)	92.0 (89.2 - 95.8)	44.0 (42.2 - 45.0)	44.0 (40.2 - 45.8)	44.5 (43.2 - 47.8)	49.0 (49.0 - 49.8)

Rater Clustering

A linear mixed-effects model with fixed effects for the AI total score and case and a random intercept for rater showed notable rater dependence (Var[rater]=3.05; Var[residual]=5.29), yielding a VPC of 0.37. This indicates that 37% of the unexplained variance in human total scores was attributable to between-rater differences in average scoring severity or leniency. Accounting for rater clustering improved fit versus ordinary least squares (ΔAkaike information criterion=19.10), supporting the presence of nontrivial rater heterogeneity [11].

Agreement Between AI and Human Total Scores

In the primary mixed model, the AI total score was positively associated with the human total score (β=0.37, 95% CI 0.26 - 0.48; *P*<.001; Table 2 and Figure 2). Case indicators were not statistically significant (Case 2: *P*=.86; Case 3: *P*=.23). Agreement was moderate: Spearman ρ=0.50, ICC(2,1)=0.51, and MAE=3.11 points. Mixed-effects Bland-Altman analysis showed a mean bias that did not reach statistical significance (1.26 points, 95% CI −0.48 to 3.01; *P*=.16) but wide 95% limits of agreement (−4.95 to 7.48; width=12.43 points), indicating educationally meaningful individual-level discrepancies. Proportional bias was present (β_proportional bias=−0.55; *P*<.001), with AI assigning relatively higher scores at the lower end of the performance distribution (Figure 3).

Table . Model-based association and AI-human agreement.

Outcome	LMM[a], β (95% CI)	*P* value	Spearman ρ	ICC[b](2,1)	Mean bias (human–AI)	Conditional 95% LoA[c]	Proportional bias, β_proportional bias	Proportional bias, *P* value
Total score	0.37 (0.26 - 0.48)	<.001	0.50	0.51	1.26	−4.95 to 7.48	−0.55	<.001
Information	0.46 (0.30 - 0.63)	<.001	0.49	0.54	−0.38	−5.52 to 4.77	−0.10	.40
Communication	0.27 (0.19 - 0.34)	<.001	0.28	0.29	1.29	−1.69 to 4.28	−0.94	<.001

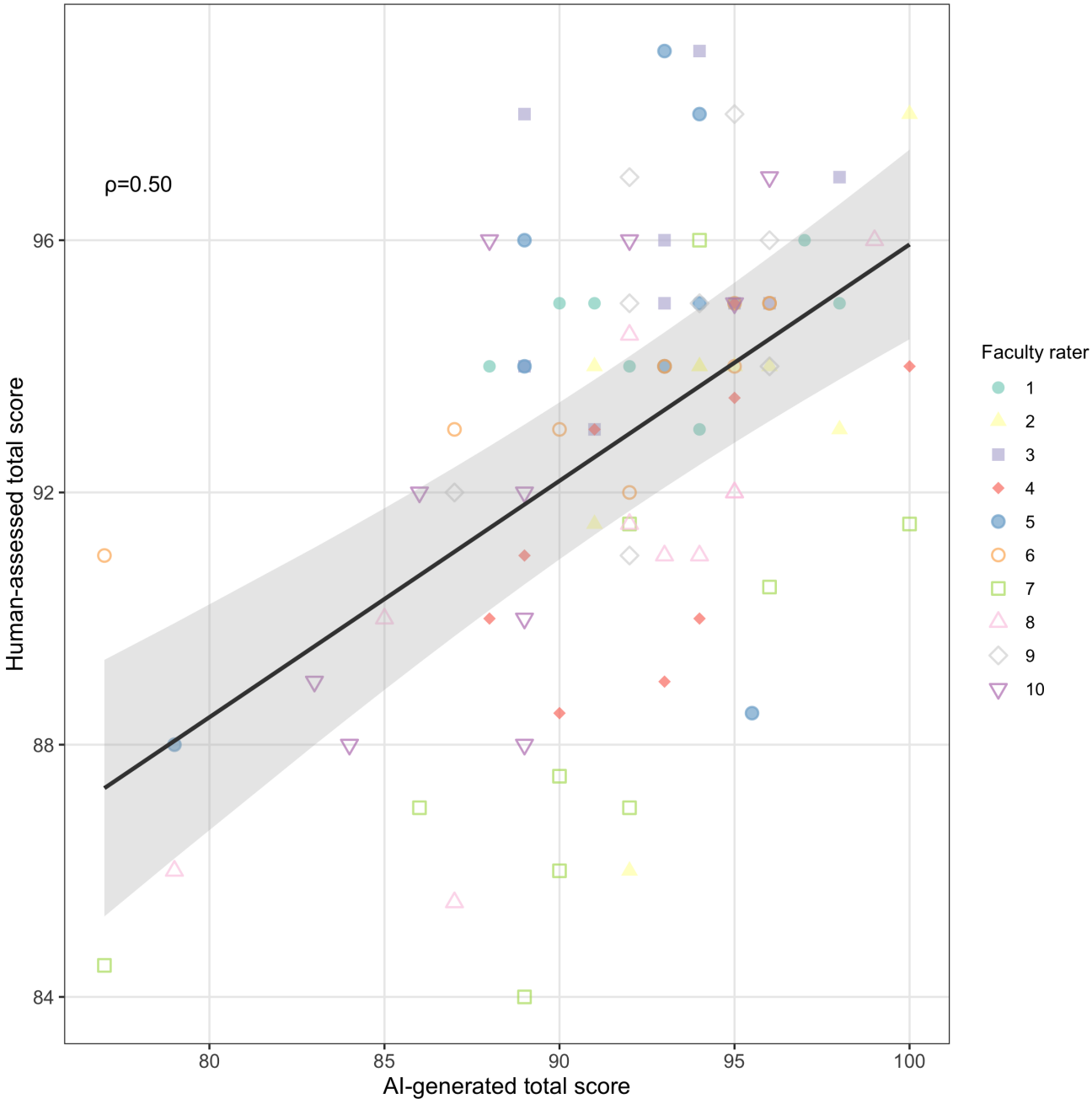
[a]LMM: linear mixed-effects model.

[b]ICC: intraclass correlation coefficient.

[c]LoA: limits of agreement.

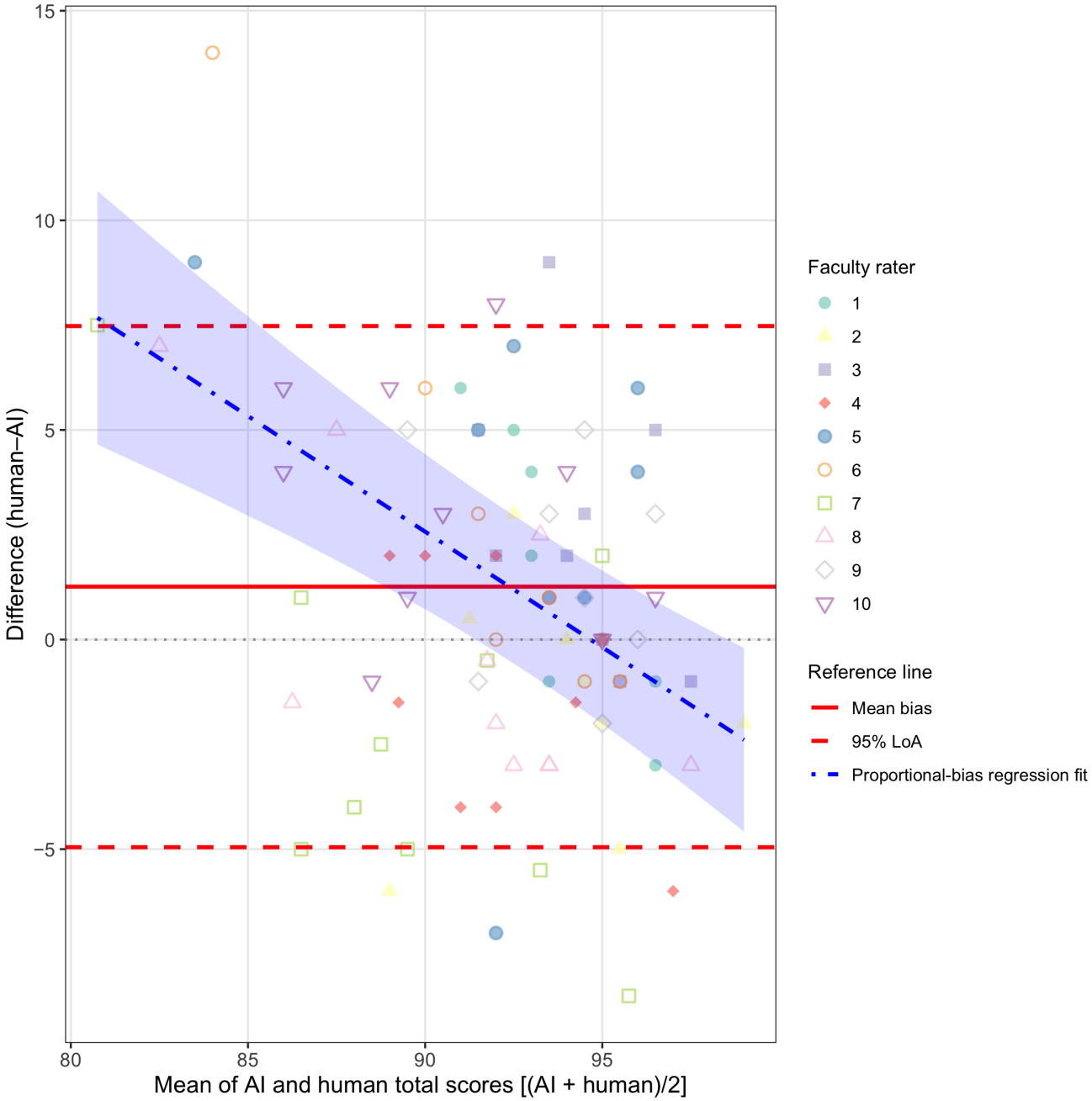
XSL•FO
RenderX

Figure 2. Scatter plot showing the association between AI-generated total scores and human-assessed total scores. Each point represents an individual medical student's score pair (n=92). The solid black line indicates the linear regression fit from the mixed-effects model (β=0.37, 95% CI 0.26 - 0.48; *P*<.001), and the gray shaded area represents the 95% CI for the regression line.



XSL•FO
RenderX

Figure 3. Mixed-effects Bland-Altman plot for AI and human total scores. The solid red line indicates the mean bias (1.26 points, 95% CI −0.48 to 3.01), which did not reach statistical significance at the group level (*P*=.16). The dashed red lines represent the 95% limits of agreement (LoA; −4.95 to 7.48; width=12.43 points). The blue line shows proportional bias (β_proportional bias=−0.55, 95% CI −0.75 to −0.35; *P*<.001), indicating that AI assigned relatively higher scores at the lower end of the performance distribution.



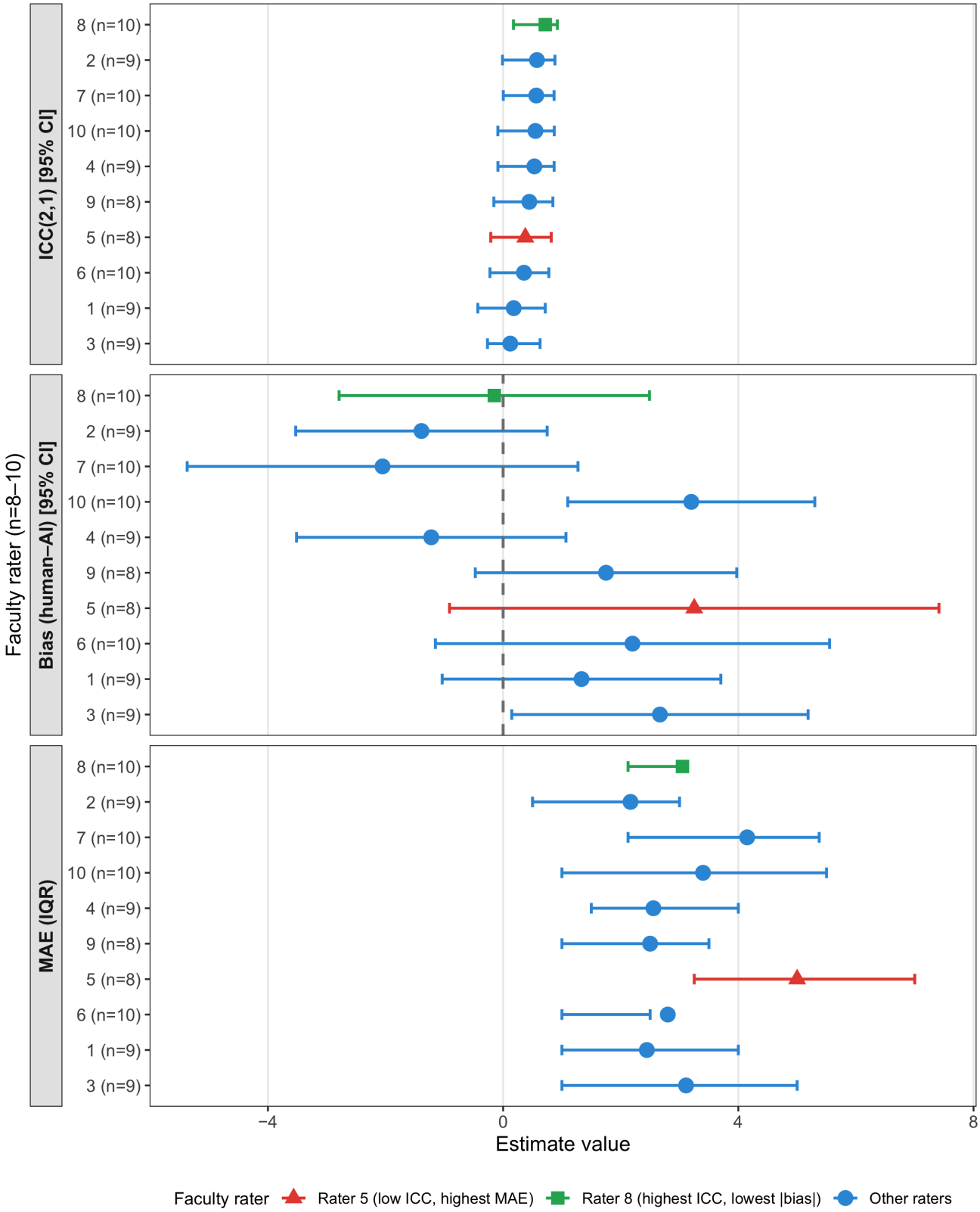
Subgroup Patterns

Rater-stratified agreement varied widely (ICC[2,1] estimates 0.57 - 0.93; Spearman ρ 0.21 - 0.93; MAE 2.17 - 5.00) with imprecise estimates (n=8 - 10 per rater). Mean AI-human bias also varied in direction across raters (–2.05 to 3.25), consistent with rater-specific scoring stringency or leniency or local calibration effects (Figure 4).

Case-stratified results suggested highest agreement in case 1 (ICC=0.63; ρ=0.62; MAE=2.83) and lowest in case 3 (ICC=0.36; ρ=0.32; MAE=4.24). These case-level estimates should be interpreted descriptively given limited per-subgroup sample sizes.

XSL•FO
RenderX

Figure 4. Forest (dot-whisker) plot of rater-level heterogeneity in AI-human agreement for total scores. Each row represents an independent rater (n=8 - 10), ordered by ICC(2,1) point estimate. Points show estimates and whiskers show 95% CIs for ICC(2,1), bias (mean Human–AI; vertical line at 0), and mean absolute error (MAE). Given small per-rater sample sizes, results are descriptive. ICC: intraclass correlation coefficient.



Sensitivity Analyses

Prespecified sensitivity analyses examined model robustness. Refitting the primary mixed-effects model with robust Huber estimation yielded a nearly identical AI coefficient (β=0.35, 95% CI 0.24 - 0.46; *P*<.001). Omitting case adjustment also produced an essentially unchanged estimate (β=0.36, 95% CI 0.25 - 0.47; *P*<.001). Full details are provided in Multimedia Appendix 4.

XSL•FO
RenderX

Subscale Analyses

AI-human agreement differed by subscale, with stronger alignment for information gathering than for communication (Table 2). For information gathering, β=0.46 (*P*<.001); ρ=0.49; ICC=0.54 (95% CI 0.38 - 0.67); VPC=0.23. For communication, β=0.27 (*P*<.001); ρ=0.28; ICC=0.29 (95% CI 0.08 - 0.47); VPC=0.52, indicating greater rater dependence and lower reproducibility.

Student Feedback

Of the 92 participating students, 63 (68.5%) completed the postencounter feedback questionnaire. Overall satisfaction was moderate (mean 6.4, SD 1.9 on a 10-point scale). Students highly valued the convenience of the VSP platform (mean 4.1, SD 0.8 on a 5-point scale). However, perceived realism was low (mean 1.8, SD 0.9 on a 5-point scale), and perceived feedback value relative to human SPs was below the scale midpoint (mean 2.4, SD 1.1 on a 5-point scale). Full survey results, including item-level response distributions, are provided in Multimedia Appendix 3.

Discussion

Principal Findings

In this study of a voice-based, LLM-powered VSP used by fourth-year medical students, 3 main findings emerged. First, AI-generated and faculty-assigned total scores showed moderate agreement, with stronger alignment for information gathering than for communication. Second, rater differences accounted for 37% of residual variance in total scores and 52% in communication scores, indicating that apparent AI-human disagreement partly reflects human-human variability. Third, proportional bias indicated that AI tended to assign relatively higher scores at the lower end of the performance distribution.

Interpretation and Psychometric Implications

From a psychometric perspective, the magnitude of AI-human agreement observed here does not support unsupervised, high-stakes use for individual learners. Although mean bias was small and did not reach statistical significance (1.26 points, 95% CI −0.48 to 3.01; *P*=.16), and the 95% limits of agreement spanned 12.43 points on a 100-point scale, wide enough that an individual learner could plausibly be classified differently depending on whether the AI or a faculty rater provided the score. Ceiling effects further complicate interpretation: with 18% of students scoring above 95 and total score SDs of only 3.4 to 4.6 points, the observed ICC of 0.51 should be treated as a lower-bound estimate. A range-restriction sensitivity analysis (Thorndike Case 2 correction) yielded a corrected ICC of approximately 0.72 (Multimedia Appendix 2). Proportional bias (AI assigning relatively higher scores to lower-performing students) further cautions against high-stakes use, particularly for communication skills.

At the same time, the overall direction and robustness of the AI-human association, which were stronger than those often observed between individual human raters in clinical performance assessments, suggested practical utility for low-stakes formative contexts [5,6,12,13]. In these contexts, the primary goal is not precise interchangeability with a single rater but scalable feedback and increased sampling of performance.

Validity Argument

This study provided criterion-related validity evidence by quantifying associations with faculty judgments. However, it addressed psychometric feasibility (score agreement) rather than educational feasibility (impact on learning outcomes), which requires dedicated longitudinal studies. Validity also depends on content alignment, response process coherence, and consequences [6]; these domains require further investigation before broader adoption—especially for higher-stakes purposes—can be justified.

Subscale-Specific Performance and Equity

Information gathering showed higher AI-human agreement and lower rater variance than communication. This aligned with the relative objectivity of checklist-based content compared with the judgment-laden nature of communication behaviors [14]. In practice, this makes AI scoring potentially useful for identifying omitted history elements, quantifying completeness, and tracking progress over time.

Communication scoring, however, remains challenging. Subjective judgments about rapport, empathy, and professionalism depend on nuanced, context-sensitive interpretation that current LLMs may not reliably replicate [15,16]. This is also an equity-relevant domain: communication judgments—whether by humans or AI—may be sensitive to accent, language proficiency, culturally patterned interaction styles, or disability-related differences. These potential differential effects should be tested explicitly rather than assumed away [17].

Rater Heterogeneity and Faculty Development

Rater-stratified results showed wide variability in AI-human agreement across faculty, reinforcing that human raters themselves are not interchangeable. Because students were nested within raters in this design, a single human score cannot be treated as an error-free reference; rather, it reflects one draw from a distribution of possible faculty judgments. This variability also limited generalizability: AI-human agreement observed here depended on the particular mix of raters and their scoring tendencies [18,19].

Rather than treating faculty ratings as a fixed “gold standard,” a more defensible target is whether AI scores fall within the range of acceptable human variation. If AI-human disagreement is comparable to human-human disagreement, AI scores can be considered psychometrically equivalent to an additional human rater—a standard that has been proposed for automated scoring in other domains [20-22].

Relation to Prior Work

Prior work on LLMs in medical education has emphasized written examination performance [23-25]. This study instead evaluated an LLM as an assessor of interactive clinical performance, foregrounding questions of rater variability and validity for performance scores. Our results aligned with emerging evidence that AI-human agreement in clinical

XSL•FO
RenderX

communication assessment is moderate rather than excellent [9,10,26], and they extended that literature by quantifying how much apparent AI-human disagreement reflected human-human variability [6]. Importantly, moderate AI-human agreement does not necessarily imply invalid AI scoring; rather, it highlights that both AI and human raters bring distinct sources of error and bias [20-22].

Limitations

This study has several limitations. Data were drawn from a single institution with predominantly high-performing students, and ceiling effects may limit generalizability to other learner populations. Each student completed only 1 encounter evaluated by a single faculty rater, precluding assessment of within-student stability and human interrater reliability. Faculty received brief rubric orientation without structured calibration, which may have introduced variability in scoring standards. Although we examined quantitative score agreement, we did not evaluate the quality or educational impact of AI-generated narrative feedback. Students' awareness of interacting with an AI system may have influenced communication behaviors. ASR error rates were not systematically recorded, limiting assessment of potential transcription bias. The scoring prompt was aligned with local faculty norms, and the analyses relied on specific LLM models, prompts, and cases; these factors may limit generalizability across institutions and over time, underscoring the need for multisite validation. Finally, the student feedback questionnaire was newly developed for this study and has not undergone formal psychometric validation (eg, content validity or reliability testing); the survey results should therefore be interpreted as exploratory and descriptive.

Student Experience and Perceived Utility

Student feedback on the VSP platform highlighted a tension between convenience and authenticity. While learners appreciated the on-demand accessibility of AI-based practice, they rated the realism of AI-simulated patient interactions and the perceived value of AI-generated feedback lower than comparable ratings for human SP encounters. These perceptions likely reflect current limitations in conversational naturalness, which may attenuate engagement and learning transfer. Improving dialogue fidelity and closing the perceived feedback quality gap represent key priorities for VSP refinement. The juxtaposition of moderate satisfaction with low perceived realism suggests that learner experience is shaped by factors beyond scoring accuracy alone, underscoring the need for multidimensional evaluation of VSP systems.

Practical Implications

Given the current evidence, we suggest that LLM-based scoring in VSPs can be used to prioritize practice opportunities and deliver rapid feedback—especially for information gathering—where scalability and frequent practice are central and where perfect interchangeability with a faculty score is not required. High-stakes use should remain human-supervised and, if used at all, should be embedded in hybrid workflows where AI provides an additional perspective rather than a final decision.

Within programmatic assessment, AI scores may be most defensible for increasing sampling density, flagging learners for coaching, and triangulating with other evidence (OSCEs, workplace-based assessments, and supervisor narratives), rather than serving as a standalone progression metric.

Future Directions

Future work should develop behavioral benchmarks that disentangle AI error from legitimate variation in faculty judgment, characterize differential item functioning across student subgroups, and test rater calibration interventions that leverage AI scores to improve human rating consistency.

Conclusions

In this undergraduate medical course, an LLM-based VSP generated history-taking and communication scores that showed moderate agreement with human raters, with greater convergence for information gathering than for communication skills. Substantial variability among human raters indicated that human scoring was itself an imperfect reference standard. Overall, these findings supported LLM-based scoring as a feasible, scalable tool for formative assessment and practice, while cautioning against unsupervised high-stakes application. Ongoing validation should prioritize rater calibration, communication scoring refinement, and equity auditing as these technologies mature in medical education.

Acknowledgments

The authors thank the faculty raters and students who participated in this study.

Funding

This work was supported by the Undergraduate Education and Teaching Reform Project of Peking Union Medical College (grant 2025bkjg011).

Data Availability

The deidentified dataset supporting the findings of this study is openly available on Zenodo at [27]. The dataset includes student-level AI and faculty scores, case assignments, rater identifiers, and item-level scoring details.

Authors' Contributions

Conceptualization: XG, XH, XS

XSL•FO
RenderX

Data curation: XG
Formal analysis: XG
Funding acquisition: XS
Investigation: XG, RH, LZ, HL, BZ, CW, WQ, MZ
Methodology: XS
Project administration: XG
Resources: XS
Software: XS
Supervision: XH, XS
Validation: XS
Visualization: XS
Writing—original draft: XG
Writing—review and editing: XG, XH, XS
All authors approved the final manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Complete large language model scoring prompt and iterative calibration procedure.
[DOCX File, 30 KB - mededu_v12i1e95578_app1.docx]

Multimedia Appendix 2
Detailed statistical model specifications.
[DOCX File, 39 KB - mededu_v12i1e95578_app2.docx]

Multimedia Appendix 3
Score distribution histograms and student feedback survey results.
[DOCX File, 389 KB - mededu_v12i1e95578_app3.docx]

Multimedia Appendix 4
Rater-stratified agreement metrics, sensitivity analyses, and lower-quartile analysis.
[DOCX File, 32 KB - mededu_v12i1e95578_app4.docx]

References

1. Kononowicz AA, Zary N, Edelbring S, Corral J, Hege I. Virtual patients--what are we talking about? A framework to classify the meanings of the term in healthcare education. BMC Med Educ 2015 Feb 1;15:11. [doi: 10.1186/s12909-015-0296-3] [Medline: 25638167]
2. Kononowicz AA, Woodham LA, Edelbring S, et al. Virtual patient simulations in health professions education: systematic review and meta-analysis by the digital health education collaboration. J Med Internet Res 2019 Jul 2;21(7):e14676. [doi: 10.2196/14676] [Medline: 31267981]
3. Cook DA, Triola MM. Virtual patients: a critical literature review and proposed next steps. Med Educ 2009 Apr;43(4):303-311. [doi: 10.1111/j.1365-2923.2008.03286.x] [Medline: 19335571]
4. Abd-Alrazaq A, AlSaad R, Alhuwail D, et al. Large language models in medical education: opportunities, challenges, and future directions. JMIR Med Educ 2023 Jun 1;9:e48291. [doi: 10.2196/48291] [Medline: 37261894]
5. Jukiewicz M, Wyrwa M. Can ChatGPT replace the teacher in assessment? A review of research on the use of large language models in grading and providing feedback. Appl Sci 2026;16(2):680. [doi: 10.3390/app16020680]
6. Cook DA, Zendejas B, Hamstra SJ, Hatala R, Brydges R. What counts as validity evidence? Examples and prevalence in a systematic review of simulation-based assessment. Adv Health Sci Educ Theory Pract 2014 May;19(2):233-250. [doi: 10.1007/s10459-013-9458-4] [Medline: 23636643]
7. Homer M. Pass/fail decisions and standards: the impact of differential examiner stringency on OSCE outcomes. Adv Health Sci Educ Theory Pract 2022 May;27(2):457-473. [doi: 10.1007/s10459-022-10096-9] [Medline: 35230590]
8. Brannick MT, Erol-Korkmaz HT, Prewett M. A systematic review of the reliability of objective structured clinical examination scores. Med Educ 2011 Dec;45(12):1181-1189. [doi: 10.1111/j.1365-2923.2011.04075.x] [Medline: 21988659]
9. Tekin M, Yurdal MO, Toraman Ç, Korkmaz G, Uysal İ. Is AI the future of evaluation in medical education?? AI vs. human evaluation in objective structured clinical examination. BMC Med Educ 2025 May 1;25(1):641. [doi: 10.1186/s12909-025-07241-4] [Medline: 40312328]

XSL•FO
RenderX

10. Schiött J, Ivegren W, Borg A, Parodis I, Skantze G. Using LLMs to grade clinical reasoning for medical students in virtual patient dialogues. In: Proceedings of the 26th Annual Meeting of the Special Interest Group on Discourse and Dialogue: Association for Computational Linguistics; 2025. URL: https://aclanthology.org/2025.sigdial-1.56/ [accessed 2026-08-12]
11. Burnham KP, Anderson DR. Multimodel inference: understanding AIC and BIC in model selection. Sociol Methods Res 2004;33(2):261-304. [doi: 10.1177/0049124104268644]
12. Yamamoto A, Koda M, Ogawa H, et al. Enhancing medical interview skills through AI-simulated patient interactions: nonrandomized controlled trial. JMIR Med Educ 2024 Sep 23;10:e58753. [doi: 10.2196/58753] [Medline: 39312284]
13. Liu Y, Shi C, Wu L, et al. Development and validation of a large language model-based system for medical history-taking training: prospective multicase study on evaluation stability, human-AI consistency, and transparency. JMIR Med Educ 2025 Aug 29;11:e73419. [doi: 10.2196/73419] [Medline: 40882613]
14. Zhou K, Giorgi JM, Mani P, Xu P, Liang D, Tan C. From feedback to checklists: grounded evaluation of AI-generated clinical notes. In: Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: Association for Computational Linguistics; 2025. [doi: 10.18653/v1/2025.emnlp-industry.104]
15. Dorrestein L, Ritter C, De Mol Z, et al. Validity evidence for communication skills assessment in health professions education: a scoping review. BMJ Open 2025 Sep 5;15(9):e096799. [doi: 10.1136/bmjopen-2024-096799] [Medline: 40912699]
16. Hodges B. Assessment in the post-psychometric era: learning to love the subjective and collective. Med Teach 2013 Jul;35(7):564-568. [doi: 10.3109/0142159X.2013.789134] [Medline: 23631408]
17. Cleland JA, Knight LV, Rees CE, Tracey S, Bond CM. Is it me or is it them? Factors that influence the passing of underperforming students. Med Educ 2008 Aug;42(8):800-809. [doi: 10.1111/j.1365-2923.2008.03113.x] [Medline: 18715477]
18. Anthony CJ, Styck KM, Volpe RJ, Robert CR. Using many-facet Rasch measurement and generalizability theory to explore rater effects for direct behavior rating-multi-item scales. Sch Psychol 2023 Mar;38(2):119-128. [doi: 10.1037/spq0000518] [Medline: 36174169]
19. Uto M. A Bayesian many-facet Rasch model with Markov modeling for rater severity drift. Behav Res Methods 2023 Oct;55(7):3910-3928. [doi: 10.3758/s13428-022-01997-z] [Medline: 36284065]
20. Roberts RH, Ali SR, Hutchings HA, Dobbs TD, Whitaker IS. Comparative study of ChatGPT and human evaluators on the assessment of medical literature according to recognised reporting standards. BMJ Health Care Inform 2023 Oct;30(1):e100830. [doi: 10.1136/bmjhci-2023-100830] [Medline: 37827724]
21. Prentice S, Benson J, Kirkpatrick E, Schuwirth L. Workplace-based assessments in postgraduate medical education: a hermeneutic review. Med Educ 2020 Nov;54(11):981-992. [doi: 10.1111/medu.14221] [Medline: 32403200]
22. Hyde S, Fessey C, Boursicot K, MacKenzie R, McGrath D. OSCE rater cognition - an international multi-centre qualitative study. BMC Med Educ 2022 Jan 3;22(1):6. [doi: 10.1186/s12909-021-03077-w] [Medline: 34980099]
23. Kung TH, Cheatham M, Medenilla A, et al. Performance of ChatGPT on USMLE: potential for AI-assisted medical education using large language models. PLOS Digit Health 2023 Feb;2(2):e0000198. [doi: 10.1371/journal.pdig.0000198] [Medline: 36812645]
24. Gilson A, Safranek CW, Huang T, et al. How does ChatGPT perform on the United States Medical Licensing Examination (USMLE)? The implications of large language models for medical education and knowledge assessment. JMIR Med Educ 2023 Feb 8;9:e45312. [doi: 10.2196/45312] [Medline: 36753318]
25. Burke HB, Hoang A, Lopreiato JO, et al. Assessing the ability of a large language model to score free-text medical student clinical notes: quantitative study. JMIR Med Educ 2024 Jul 25;10:e56342. [doi: 10.2196/56342] [Medline: 39118469]
26. Dhillon IK, Lee GK, Hu S. Evaluating the pediatric behavior guidance of students based on actual clinical transcripts scored by faculty and large language models: pilot comparative study. JMIR Med Educ 2026 Jun 12;12:e83376. [doi: 10.2196/83376] [Medline: 42285034]
27. Sun X. AI score and human score of medical students' VSP interview. Zenodo. 2026. URL: https://zenodo.org/records/20774435 [accessed 2026-08-26]

Abbreviations

ASR: automatic speech recognition
ICC: intraclass correlation coefficient
LLM: large language model
MAE: mean absolute error
OSCE: objective structured clinical examination
SP: standardized patient
VPC: variance partition coefficient
VSP: virtual standardized patient

XSL•FO
RenderX

Edited by JWY Lee; submitted 19.Mar.2026; peer-reviewed by MA Vicente, S Cui; revised version received 28.Jul.2026; accepted 30.Jul.2026; published 24.Sep.2026.

Please cite as:
Gao X, Huang X, Hu R, Zhang L, Liu H, Zhang B, Wei C, Qiu W, Zhang M, Sun X
Evaluating Large Language Model–Based Automated Scoring in a Voice-Based Virtual Standardized Patient Platform for Medical Students: Cross-Sectional Agreement Study
JMIR Med Educ 2026;12:e95578
URL: https://mededu.jmir.org/2026/1/e95578
doi:10.2196/95578

© Xiaoxing Gao, Xiaoming Huang, Rongrong Hu, Li Zhang, Huiting Liu, Bingqing Zhang, Chong Wei, Wei Qiu, Mengyu Zhang, Xuefeng Sun. Originally published in JMIR Medical Education (https://mededu.jmir.org), 24.Sep.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Education, is properly cited. The complete bibliographic information, a link to the original publication on https://mededu.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Teaching Acute Coronary Syndrome High-Risk ECG Interpretation and Clinical Decision-Making Through FOAMed Videos and Podcast Versus Print-Based Materials Among Emergency Care Providers: Randomized Controlled Mixed Methods Trial

Wiebke Turner[1], MD; Klaus Fessele[2], Dr med; Martin Fandler[3], Dr med univ; Jan P Ehlers[1], Dr med vet; Julia Nitsche[1], Dr rer med

[1]Department of Human Medicine, Chair of Didactics and Educational Research in Healthcare, Witten/Herdecke University, Alfred-Herrhausen-Straße 50, Witten, North Rhine-Westphalia, Germany

[2]Emergency Department, Nuremberg Hospital, Nuremberg, Bavaria, Germany

[3]ADAC HEMS Academy, Weßling, Bavaria, Germany

Corresponding Author:
Wiebke Turner, MD
Department of Human Medicine, Chair of Didactics and Educational Research in Healthcare, Witten/Herdecke University, Alfred-Herrhausen-Straße 50, Witten, North Rhine-Westphalia, Germany

Abstract

Background: Accurate interpretation of high-risk acute coronary syndrome (ACS) electrocardiograms (ECGs) is essential for early diagnosis and timely reperfusion, yet substantial deficits persist across health care professions. Digital self-learning formats such as FOAMed (Free Open Access Medical Education) are widely used, but their effectiveness has rarely been evaluated for complex, high-risk ACS ECG patterns. Existing ECG education studies often focus on students or single professional groups and established ST-segment elevation myocardial infarction (STEMI) criteria, leaving newer guideline-recognized STEMI equivalents, selected emerging occlusion myocardial infarction (OMI)–related patterns, and interprofessional emergency care underrepresented.

Objective: This study aimed to compare the effectiveness of FOAMed podcast and videos versus traditional print-based materials for teaching high-risk ACS ECG patterns and related clinical decision-making in emergency providers.

Methods: We conducted a prospective, interprofessional, controlled mixed methods trial across 5 training sites in Germany. Paramedics, prehospital emergency physicians, and emergency department clinicians received either a FOAMed multimedia module or print-based materials through concealed allocation; deviations from the intended 1:1 ratio resulted from participant no-shows. The intervention consisted of a 30-minute supervised self-learning session. In total, 103 participants were allocated to FOAMed (n=45) or print-based materials (n=58). Two coprimary outcomes were assessed: ECG interpretation accuracy and text-based ACS clinical decision-making. Secondary outcomes included subjective confidence, learning experience, and exploratory qualitative free-text responses. Outcome assessment was automated and blinded; mixed ANOVA was the primary analysis. The study was not prospectively registered because it assessed educational outcomes in health care professionals rather than patient health outcomes.

Results: All 103 participants completed the study. Both groups improved, with greater gains in the FOAMed group: ECG interpretation increased from 55% to 65.5% and text-based ACS clinical decision-making from 45% to 68%, versus 57% to 60% and from 47% to 63%, respectively, in the print-based group. Effect sizes were η^2=0.055 for ECG interpretation and η^2=0.044 for clinical decision-making. Exploratory subgroup analyses provided no evidence of differential effects across age, gender, or professional background and were likely underpowered. Qualitative responses (46 and 37 entries) provided contextual insights into perceived clarity, engagement, and practical relevance supporting the quantitative findings.

Conclusions: This study is innovative in directly comparing a curated FOAMed multimedia module with selected print-based materials in an interprofessional emergency care population. It differs from existing research by focusing on subtle, emerging ischemic patterns and evaluating realistic, time-limited self-learning formats. The findings provide evidence that curated FOAMed resources can produce greater short-term improvements in ECG interpretation and text-based ACS clinical decision-making than traditional print-based materials in this setting. Although implications for clinical performance remain hypothetical, concise, high-quality digital modules may represent a practical supplement to structured continuing education in emergency care.

(JMIR Med Educ 2026;12:e87587) doi:10.2196/87587

XSL•FO
RenderX

KEYWORDS

acute coronary syndrome; electrocardiogram interpretation; high-risk ECG signs; STEMI equivalents; clinical decision-making; FOAMed; digital learning; podcast-based education; video-based learning; self-directed learning; electrocardiogram; Free Open Access Medical Education

Introduction

The rapid and correct interpretation of electrocardiograms (ECGs) is crucial for the initial treatment of acute coronary syndrome (ACS) [1]. In addition to classic electrocardiographic ST-segment elevation myocardial infarctions (STEMI), various guideline-recognized STEMI equivalents and selected high-risk ACS ECG patterns have been described in recent years. These include left bundle branch block and right bundle branch block in the context of ACS, hyperacute T waves, de Winter and Wellens sign, shark fin sign (as a special subtype of STEMI), Aslanger sign, semi-STEMI, and left main occlusion ECG [2-8]. Recent work further emphasizes the occlusion myocardial infarction (OMI) paradigm and the diagnostic relevance of subtle high-risk ACS ECG patterns [9,10]. Despite their clinical relevance, many of these patterns were only recently incorporated into ACS guidelines or remain insufficiently represented [1]. At the same time, several ECG phenomena closely resemble these high-risk ACS ECG patterns and are therefore referred to as "STEMI mimics" [11]. Misinterpretation of high-risk ACS ECG patterns—particularly STEMI equivalents—can lead to delayed recognition and reperfusion, whereas STEMI mimics primarily contribute to false-positive activations rather than treatment delays.

International meta-analyses and studies show that the average accuracy of ECG interpretation among medical students and physicians is only around 54% [12]. More recent analyses confirm persistent deficits across health care professions [13], with diagnostic accuracy varying by age, gender, and expertise [14]. Recent prospective data further show that even senior emergency physicians frequently miss major ECG abnormalities, with only moderate agreement compared to cardiologists [15]. Comparable deficits have been reported among emergency medical services personnel [16], and in some cases, uncertainty is so pronounced that it has been described as "ECGphobia" [17], with observational work further showing that paramedics frequently misinterpret ACS-related ECG patterns [18]. Additional systematic review evidence similarly indicates that paramedics' ability to identify STEMI varies substantially and improves with targeted ECG training [19]. These deficits affect not only rare patterns, but also central, potentially life-threatening ECG patterns such as classic STEMI findings, which are not recognized in up to 44% of cases [13,20].

Primary contributors to these deficits include insufficient training, limited exposure, and a lack of structured ECG education [21]. Emergency medical technicians likewise report a strong need for additional ECG training [16]. Standardized training that includes both established guideline-recognized STEMI equivalents and selected emerging or previously described high-risk ACS ECG patterns is therefore essential to strengthen diagnostic competencies in emergency care.

However, access to structured continuing education is often limited by time constraints, staffing shortages, and financial pressures [22]. At the same time, digitalization offers new learning opportunities for flexible learning. Podcasts, videos, and other freely available formats in the spirit of the Free Open Access Medical Education (FOAMed) movement are increasingly being used [23], although they remain insufficiently integrated into formal curricula [23]. Digital learning formats offer time- and location-independent education, address practice-relevant content, and accommodate different learning preferences. According to cognitive load theory and the cognitive theory of multimedia learning, audiovisual formats can reduce extraneous load and support more efficient processing of complex information [24,25]. However, evidence regarding the effectiveness and implementation of digital learning in health sciences education remains heterogeneous [26]. Although numerous studies have evaluated ECG teaching interventions [27,28], no randomized studies have compared FOAMed-based self-learning formats with traditional print materials for teaching high-risk ACS ECG patterns. This represents a significant research gap: although the need for structured, practical, and efficient continuing education in emergency medicine is undisputed, robust data on which formats are most effective for which content are lacking. Previous studies have focused primarily on medical students or physicians [27,28], while evidence from emergency medical personnel remains scarce [16,18,19]. The aim of the study is to evaluate the effectiveness of digital FOAMed self-learning formats (in this case, podcast and videos) compared with traditional print materials (guidelines and textbooks) in teaching selected high-risk ACS ECG patterns. The study assesses objective knowledge gain, clinical decision-making, subjective confidence, and perceived learning experience. Additionally, differences between established and newer ECG patterns and between professional groups in emergency medicine will be analyzed. Specifically, this study addresses the following research questions:

1. What influence does 30 minutes of using various self-learning training media have on learning success?
   a. Regarding interpreting ECG
   b. Regarding knowledge in dealing with potential myocardial infarction patients
2. What impact does 30 minutes of using various self-training media have on the sense of safety?
3. How do participants rate the learning experience of the training media they used?
4. Which groups of people show the greatest increase in knowledge with which training medium?
5. What is the current level of knowledge of the newer high-risk ACS ECG patterns among decision-makers in emergency medicine?

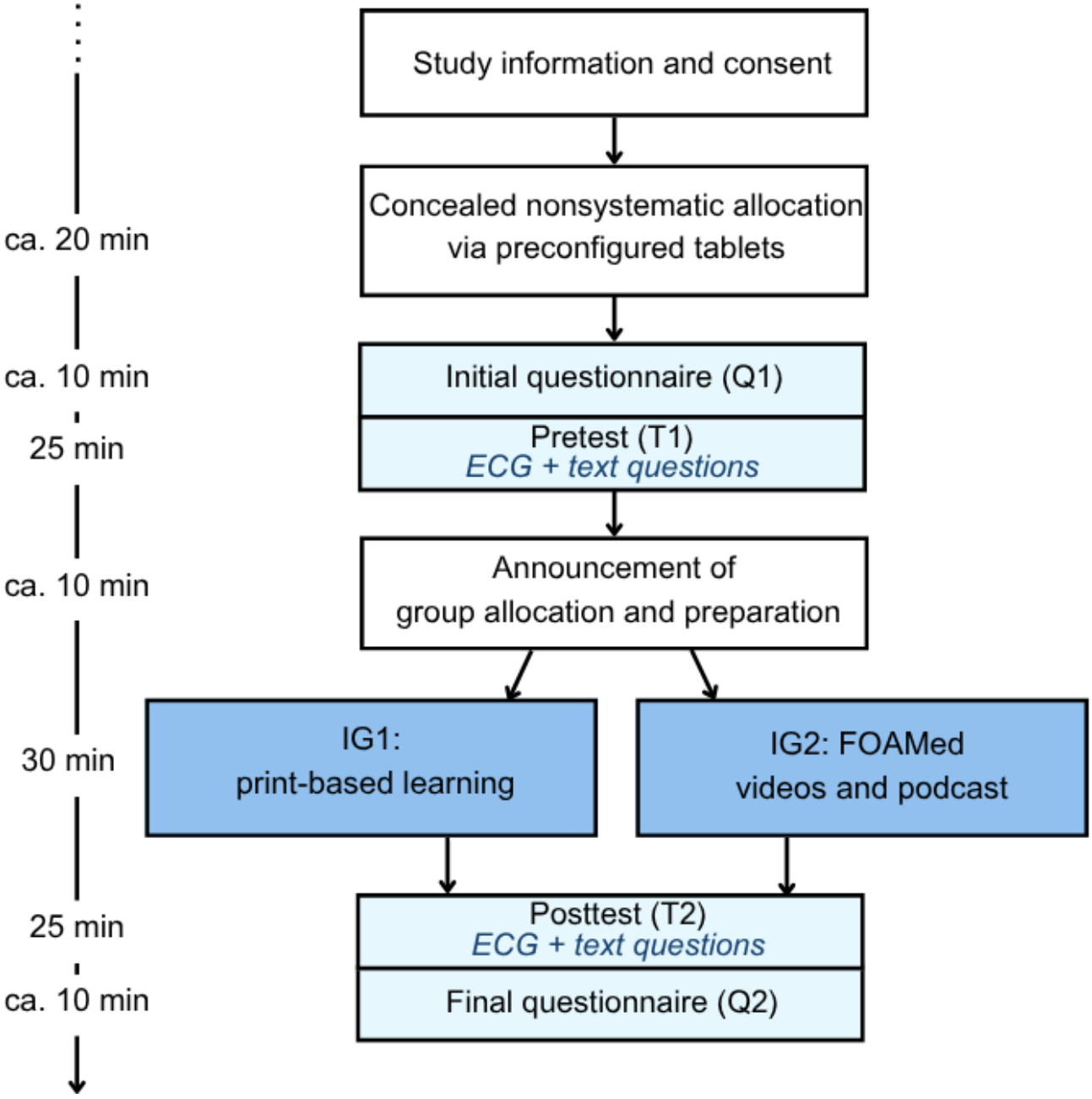
XSL•FO
RenderX

Methods

Trial Design

This study was designed as a prospective, parallel-group controlled educational trial with 2 intervention arms and a pre-post assessment structure. The trial followed a superiority framework, comparing 2 self-directed learning formats. No changes to the trial protocol or prespecified outcomes occurred after trial commencement. No publicly accessible trial protocol or separate statistical analysis plan exists for this minimal-risk educational trial. All methodological details, including prespecified outcomes and statistical procedures, are fully reported in this paper. Figure 1 provides a visual summary of the study design, including concealed, nonsystematic allocation, timing of assessments, and the sequence of both intervention arms. The study was not prospectively registered in a primary trial registry. According to the International Committee of Medical Journal Editors (ICMJE), registration is not required for studies in which health care providers are assigned to intervention and control groups when the purpose is to assess effects on the providers themselves, such as knowledge or attitudes, rather than on the health outcomes of their patients [29]. As this study evaluated the effect of educational interventions on participants' ECG interpretation and knowledge, without assessing patient health outcomes, prospective trial registration was therefore considered not applicable. This paper was prepared in accordance with the CONSORT (Consolidated Standards of Reporting Trials) 2025 guidelines for reporting randomized controlled trials [30].

Figure 1. Study design of a controlled educational intervention comparing FOAMed (Free Open Access Medical Education) videos and podcast with print-based learning materials among emergency care providers in Germany in 2024. ECG: electrocardiogram; IG1: intervention group 1; IG2: intervention group 2.



XSL•FO
RenderX

Participants

Participants were paramedics and physicians regularly involved in ACS-related decision-making in emergency care. Inclusion criteria were age 18 - 65 years, ≥30 working hours per week, and active clinical responsibility in emergency medicine. Exclusion criteria were medical students without decision-making authority, trainees, and members of the study team.

Recruitment occurred through invitations to hospital departments and emergency services, with remaining slots publicly advertised via social media. Five in-person study events were conducted in North Rhine-Westphalia, Germany, in January 2024. Patients or members of the public were not involved in the design, conduct, reporting, or dissemination plans of this study.

Because the trial involved only self-directed educational materials without any patient contact or patient-facing procedures, patient or public involvement was not applicable.

Trial Setting

The trial took place in classroom-based training sessions at 5 emergency-medicine teaching locations in North Rhine-Westphalia, Germany. All assessments and interventions were administered on-site under supervised conditions.

Study Procedure

The study consisted of 2 assessment points: a pretest (Questionnaire 1/Test 1; Q1/T1) immediately before the intervention and a posttest (Questionnaire 2/Test 2; Q2/T2) directly after the 30-minute learning phase. Both assessments were completed individually on the tablets provided using the same interface and response format. After completing the pretest, the intervention instructions were presented via the projector, and all participants began the 30-minute learning phase simultaneously. The posttest followed immediately under identical supervised conditions.

Interventions

Print-Based Learning

Participants in the print-based learning group (intervention Group 1; IG1) received full, unmarked versions of established textbooks and current ACS guidelines. Materials included the following:

- *EKG-Kurs für Isabel* (Trappe and Schuster, 8th ed, 2020; ISBN 978 - 3132200302)
- *EKG in der Notfallmedizin* (Schnelle, 1st ed, 2017; ISBN 978 - 3943174694)
- Deutsche Gesellschaft für Kardiologie (DGK) non–ST-segment elevation acute coronary syndrome (NSTE-ACS) guideline (2020)
- DGK STEMI guideline (2017)
- European Society of Cardiology (ESC) ACS guideline (2023)
- American College of Cardiology (ACC) Chest Pain Consensus (2022)

The DGK guidelines were available in German, while the ESC guideline and the ACC Chest Pain Consensus were provided in English, reflecting their original publication languages. The textbooks were selected based on a prestudy survey of 1012 emergency care providers identifying the most used and recommended ECG resources, ensuring ecological validity and relevance for the target population. All materials were provided in full length without marked chapters. Except for the consensus paper, all print-based materials included a table of contents that participants could use to navigate relevant ACS-related sections. The core components of the print-based intervention were unrestricted access to full textbook and guideline materials and self-directed navigation of ACS-relevant content.

Digital Learning

Participants in the digital learning group (intervention group 2; IG2) completed a curated multimedia module consisting of the following:

- Nine short YouTube (Google LLC) videos from "Nerdfallmedizin"
- A podcast episode from "PinUpDocs" on high-risk ACS ECG patterns

All digital materials were frozen for the duration of the study, delivered via preconfigured iPads, and could be paused or replayed. The complete list of all FOAMed materials (titles, URLs, and durations) and interface screenshots is provided in Multimedia Appendix 1. The detailed learning objectives, exact sequence of multimedia elements, and the Template for Intervention Description and Replication (TIDieR)–compliant description of the intervention are fully reported in the "Methods" section.

Delivery and Fidelity

Both interventions were delivered in a supervised, face-to-face classroom setting. Each participant had an individual desk facing the projector screen. A 30-minute timer was displayed on the projector, which was also used to introduce the intervention and explain the procedure. All participants started simultaneously, and the study lead together with an uninvolved assistant ensured a quiet environment without conversation. Both were available throughout to address technical issues but did not provide instructional guidance. No co-interventions or additional educational activities were provided during the study period.

Participants were asked in advance to bring their own Bluetooth-compatible headphones, which were paired with the tablets before the intervention started. Each participant had an individual tablet for completing the assessments; IG2 additionally used the tablet to access the multimedia module. In IG1, each participant received one copy of each listed textbook and one copy of each printed, ring-bound guideline, ensuring that all materials were individually available and not shared. Participants could take handwritten notes on a provided A5 notepad, although notes were not permitted during the posttests.

Both intervention arms addressed the same overarching topic—recognition and management of high-risk ACS ECG patterns—but the materials reflected the real-world differences between print-based and digital educational resources. The print-based group had unrestricted access to current ACS guidelines, which fully cover high-risk ECG concepts, while

XSL•FO
RenderX

the textbooks reflected the publication cycle of print media. The FOAMed materials presented the same concepts in a curated multimedia format. These differences represent authentic educational ecosystems rather than methodological imbalance.

Outcomes

Overview

The primary outcome was improvement in ACS-related diagnostic performance from T1 to T2, assessed through two coprimary components: (1) ECG interpretation accuracy (high-risk ACS ECG: yes, no, or unsure) and (2) text-based clinical decision-making items, consisting of true, false, or unsure statements on ACS management. Both components were administered at T1 and T2 and analyzed as coprimary outcomes. No formal trial-level success criterion was prespecified, because the two coprimary outcomes represent distinct diagnostic domains in which different learning formats could theoretically perform better.

Secondary outcomes included the following:

- Subjective confidence in ECG interpretation and ACS management
- Learning experience ratings
- Qualitative explanations for changes in perceived confidence

Clinical decision-making was assessed using text-based emergency scenarios, representing hypothetical decision-making rather than observed clinical behavior. Confidence and sense of safety were measured as self-reported perceptions and do not reflect objective clinical competence. All outcomes were prespecified. No harms or unintended effects occurred or were reported. As the intervention consisted solely of self-directed educational materials without any clinical procedures or patient contact, no adverse events were expected.

Test Development and Validity

All ECG and text-based items were developed under the supervision of a senior cardiologist and emergency physician (KF), who reviewed all items for clinical relevance, clarity, and unambiguous ECG morphology. ECG cases were primarily drawn from an internal case archive of a cardiology emergency department; additional cases were sourced from Dr Smith's ECG blog, an educational resource founded by Stephen W Smith, MD, Professor of Emergency Medicine at the University of Minnesota [31], and subsequently verified for diagnostic accuracy by the supervising emergency cardiologist. Face validity was further assessed through pretesting with 8 target-group representatives (4 physicians and 4 paramedics), who evaluated comprehensibility and assigned difficulty ratings on a 1 - 3 scale. Based on mean difficulty ratings, 2 parallel test forms of comparable difficulty were created for T1 and T2, and 5 items (5 ECGs and 5 text questions) were repeated across both tests to anchor difficulty. Internal consistency metrics (eg, Cronbach α) and item discrimination indices were not calculated because the instrument was intentionally designed as a heterogeneous, content-validated assessment combining ECG pattern recognition and guideline-based text questions, which measure distinct constructs and do not meet the psychometric assumptions required for such analyses. Sample ECG items and text questions are provided in Multimedia Appendix 2 to support transparency and face and content validity.

Methodological Quality Assessment (Medical Education Research Study Quality Instrument)

The methodological quality of the quantitative component of this study was evaluated using the Medical Education Research Study Quality Instrument (MERSQI) [32]. The scoring was performed by 2 independent reviewers from the Chair of Didactics and Educational Research in Healthcare who were not involved in the study. Both reviewers first scored all 6 MERSQI domains independently and subsequently discussed 2 minor discrepancies before reaching full consensus. Domain-level scores and justifications are provided in the MERSQI checklist (Checklist 1). The MERSQI score applies exclusively to the quantitative part of the study, as the instrument is not designed to evaluate qualitative components.

Sample Size

An a priori power analysis (G*Power 3.1; Faul, Erdfelder, Lang, and Buchner; mixed ANOVA; Cohen f=0.25; α=.05; power=0.80) indicated a required sample size of 126 participants. No interim analyses or stopping guidelines were planned.

Allocation Procedure: Sequence Generation

Allocation was achieved through practically random distribution of preconfigured tablets. Half of the tablets contained the FOAMed multimedia module, while the other half contained only the assessment interface. Tablets were prepared manually before participant arrival; no computer-generated random sequence was used.

Allocation Concealment

Tablets were placed nonsystematically on tables, and all devices displayed identical locked screens, preventing visibility of group assignment. Neither participants nor study personnel could infer allocation prior to unlocking the devices.

Implementation

An uninformed assistant distributed the tablets, and participants freely chose seats and generated an individual study code. Because the tablets were indistinguishable and group assignment became visible only upon unlocking, allocation was fully concealed. Participants assigned to the print-based group received the printed textbooks and guidelines only after unlocking their tablet.

The unequal final group sizes (58 vs 45) resulted from participant no-shows: preconfigured tablets that were not unlocked did not contribute to the realized allocation. Seat choice did not influence allocation because tablet content could not be inferred from appearance or placement, and free seating therefore did not introduce selection bias.

Blinding

Blinding of participants and study personnel was not feasible due to the nature of the educational interventions. Outcome

XSL•FO
RenderX

assessment was automated and identical across groups, minimizing assessor bias.

Statistical Methods (Quantitative Analysis)

Overview

Descriptive statistics were calculated in Microsoft Excel. Primary and secondary outcomes were analyzed using mixed ANOVA (between-subject factor: group; within-subject factor: time). Mixed ANOVA served as the primary analysis, and no additional mixed-effects modeling was required due to the complete paired dataset. All point estimates are reported with 95% CIs. No statistical adjustments for multiple comparisons were applied. Secondary outcomes and subgroup analyses are therefore interpreted as exploratory and likely underpowered. Trichotomous ECG and text-based responses (correct, incorrect, or unsure) were dichotomized into correct versus incorrect, with "unsure" coded as incorrect.

Assumption checks for the mixed ANOVA were conducted. Normality of residuals was assessed using Shapiro-Wilk tests, homogeneity of variances using Levene test, and equality of covariance matrices using the Box M test; all assumptions were sufficiently met. Because the within-subject factor included only 2 time points, sphericity was automatically satisfied. Percentage scores were treated as continuous variables, which is considered acceptable for ANOVA when distributions are not severely skewed and sample sizes are adequate. Baseline comparability between groups was confirmed by similar T1 means (Table 1). As a sensitivity analysis, nonparametric comparisons (Mann-Whitney U for between-group differences at T1 and T2; Wilcoxon signed-rank tests for within-group changes) yielded the same directional effects, providing additional reassurance regarding the consistency of the findings. Mean differences and 95% CIs are reported for all outcomes.

Table . Pre- and postintervention test performance in ECG[a] interpretation and text-based ACS[b] decision-making among participants in a controlled educational intervention comparing FOAMed[c] videos and podcast with print-based learning materials in Germany in 2024[d].

	Average of correct answers	
Test component	Print-based media	FOAMed videos and podcast
ECG T1	57	55
ECG T2	60	65.5
Knowledge development	+3%P[e]	+10.5%P
Text questions T1	47	45
Text questions T2	63	68
Knowledge development	+16%P	+23%P

[a]ECG: electrocardiogram.

[b]ACS: acute coronary syndrome.

[c]FOAMed: Free Open Access Medical Education.

[d]The table presents average correct answers for electrocardiogram and text-based questions at baseline (T1) and postintervention (T2), as well as calculated knowledge development for both intervention groups. Full inferential statistics are provided in Tables S1 and S2 in Multimedia Appendix 3.

[e]%P denotes percentage points, indicating the absolute difference between two percentages.

Analyses were conducted on all allocated participants with available data (complete-case analysis). Because no participants were lost between T1 and T2, all allocated participants contributed complete paired data. Therefore, the intention-to-treat (ITT) and per-protocol (PP) populations were identical, and the primary analysis effectively reflects both ITT and PP principles.

Missing Data

Missing data were assessed at both the participant and item level. No participant-level missingness occurred, as all participants completed the pretest, intervention, and posttest consecutively within a single supervised session (0%). For all ECG interpretation items and clinical decision-making text questions, the response option "I am too unsure" was available and counted as a valid response. In addition, the survey platform (LimeSurvey; LimeSurvey GmbH) required a response before participants could proceed, ensuring complete data for these test items (0% item-level missingness).

For the media evaluation items, overall item-level missingness was 1.2% (24/2043). For the perceived safety ratings, overall missingness was similarly low at 1.2% (6/514). Free-text questions were optional; therefore, non-responses were expected and not considered missing data in the statistical sense.

Because missingness across all quantitative variables was minimal and no inferential analyses required imputation or assumptions about the missing-data mechanism, Missing Completely At Random (MCAR) testing was not applicable.

Qualitative Analysis (Mayring)

The qualitative component was exploratory and aimed at contextualizing the quantitative findings by capturing participants' reasons for feeling equally safe or unsafe or less safe after the intervention. The two open-ended survey items were optional and presented only to participants who had indicated unchanged or decreased perceived safety. The exact original German wording was as follows:

XSL•FO
RenderX

- "Beim Erkennen herzkatheter-pflichtiger EKG-Bilder fühle ich mich jetzt gleich (un-)sicher / unsicherer, weil…"
- "Im Umgang mit potenziellen Myokardinfarkt-Patient:innen fühle ich mich jetzt gleich (un-)sicher / unsicherer, weil…"

A total of 46 free-text responses were provided for the question on high-risk ACS ECG interpretation (29 from the print-based group and 17 from the FOAMed group), out of 54 eligible participants. For the question on managing potential myocardial infarction patients, 37 responses were provided (21 print-based and 16 FOAMed), out of 51 eligible participants. As the open-ended items were optional and placed at the end of the survey, the qualitative dataset represents a subset of the quantitative sample.

All responses were transferred verbatim into an Excel matrix and inductively coded following the principles of summary content analysis according to Mayring [33]. The first author conducted the initial coding and assigned content categories, followed by a second step of thematic grouping. To enhance reliability, a didactics-affine internist with experience in qualitative coding reviewed all categories within a coding circle; one minor category adjustment was required, and all discrepancies were resolved by consensus. Irrelevant statements were excluded, and unclear formulations were retained and marked as such. Given the brevity of the free-text responses (typically one short sentence), the qualitative analysis was descriptive in nature and limited in interpretive depth. The responses nevertheless provided sufficient descriptive variation to identify recurring patterns relevant to the study aim. Although the qualitative component consisted of 2 brief free-text questions within a single data-collection wave, thematic saturation was effectively reached, as no new categories emerged during the inductive summary content analysis and all responses clustered within a small set of recurring themes. The qualitative analysis was carried out by 2 emergency-medicine researchers with prior experience in ECG education and qualitative coding. Their professional background and prior familiarity with ACS-related decision-making were explicitly acknowledged during the analytic process, and reflexive discussions were used to limit potential influence on category formation.

The free-text questions were intentionally offered only to participants who reported unchanged or decreased perceived safety, as the aim was to explore reasons for limited confidence gains. This selective sampling strategy aligns with the exploratory purpose of contextualizing heterogeneous changes in perceived safety and was not intended to achieve thematic saturation in the comprehensive qualitative sense.

Methodological integrity was supported through coder consensus, iterative refinement of categories, and grounding all interpretations in verbatim participant statements. Given the exploratory nature and the brevity of responses, the qualitative findings are presented as interpretive insights rather than fully developed thematic conclusions.

Ethical Considerations

Ethical Approval

The study was reviewed and approved by the Ethics Committee of Witten Herdecke University (reference number 139/2023). As the study involved health care providers as participants and assessed the effects of educational interventions on their knowledge and ECG interpretation, without patient involvement or assessment of patient health outcomes, prospective trial registration was not considered applicable according to ICMJE guidance [29].

Informed Consent

All participants received written study information and provided informed consent prior to participation. Participation was voluntary, and withdrawal was possible at any time without consequences.

Privacy and Confidentiality

All data were collected in a pseudonymized manner using participant-generated individual study codes. No personal identifiers were recorded, and no reidentification key existed within the study team. Although theoretical reidentification would only be possible through external personnel files, the study team had no access to such records. All datasets were stored in accordance with General Data Protection Regulation (GDPR) requirements on secure institutional servers.

Compensation

Participants did not receive financial compensation. They were offered the opportunity to enter a raffle for the textbooks used during the study and to attend an optional webinar.

Image Use and Identifiability

No identifiable images of participants were collected or included in the paper or supplementary materials. Therefore, no image-related consent was required.

A completed CONSORT-eHEALTH (V1.6) checklist is provided as a checklist file in Checklist 2.

Results

Overview

The primary outcome was improvement in ACS-related diagnostic performance from T1 to T2, assessed through 2 coprimary components: ECG interpretation accuracy (high-risk ACS ECG: yes, no, or unsure) and text-based clinical decision-making items (true, false, or unsure). These 2 components were analyzed jointly to capture short-term diagnostic knowledge gain across both intervention groups. Secondary outcomes included subjective confidence in ECG interpretation and ACS management, as well as learning experience ratings. All decision-making results reflect performance in simulated, text-based scenarios rather than real-world clinical actions and therefore represent hypothetical rather than observed clinical behavior.

Participants and Study Implementation

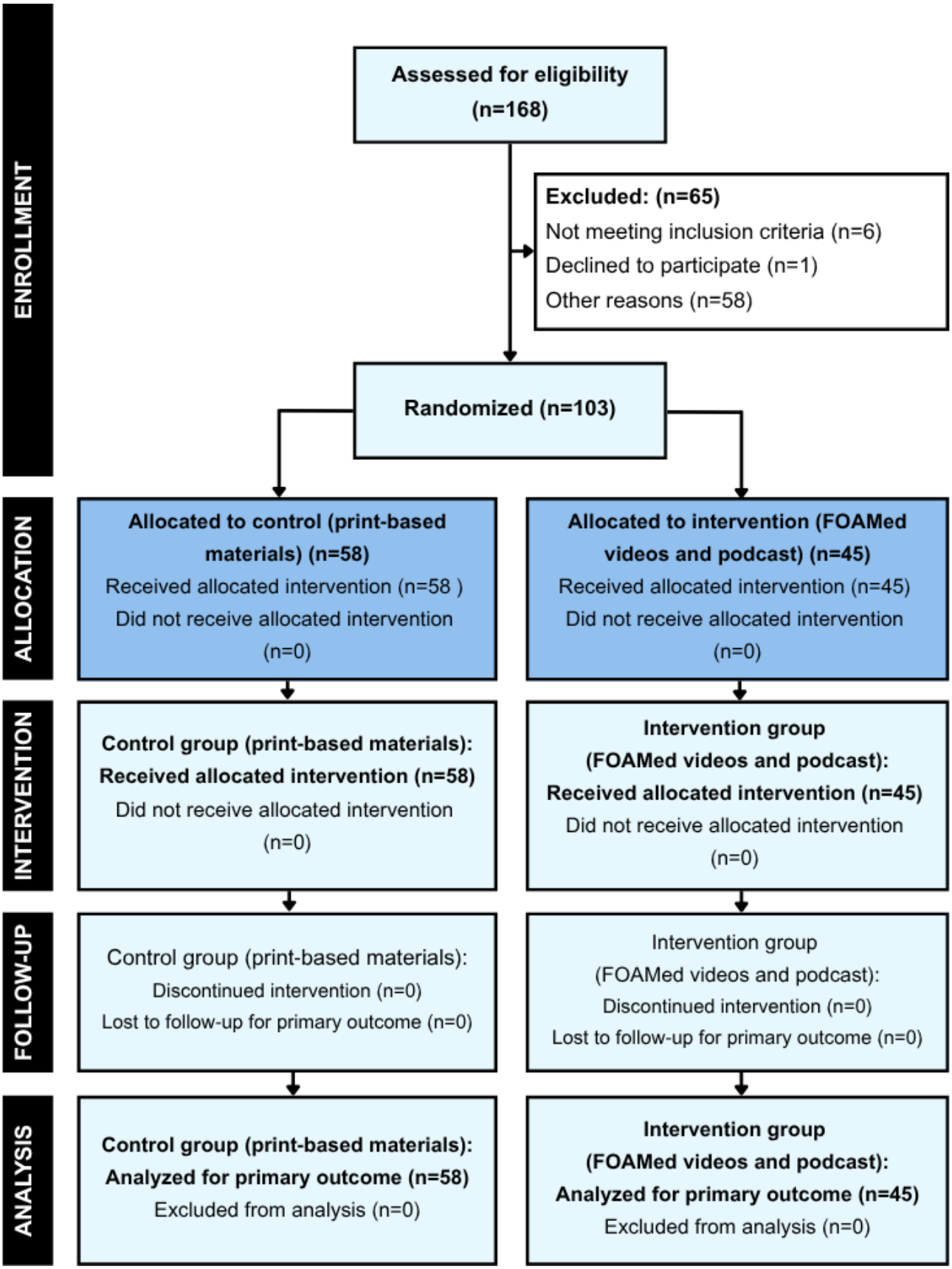
Of the 168 people who originally registered, 103 participants ultimately took part in the study and were allocated. Six of the 65 people who withdrew did not meet the inclusion criteria (including trainees and medical staff not working in emergency medicine), and one doctor declined due to organizational constraints (no free parking). The most significant dropouts

XSL•FO
RenderX

were the 58 last-minute cancellations due to staff shortages in the emergency services and clinics, illness (their own or their children's), and travel problems (train cancellations and car breakdowns). Nine of these 58 people did not show up for their appointments without giving notice. The following CONSORT diagram (Figure 2) visualizes all dropouts. All allocated participants completed both T1 and T2 assessments; therefore, ITT and PP analyses were identical.

XSL•FO
RenderX

Figure 2. CONSORT (Consolidated Standards of Reporting Trials) 2025 flow diagram of participant progression through the educational intervention. CONSORT 2025 flow diagram showing participant enrollment, allocation, follow-up, and analysis in a controlled educational intervention among emergency care providers in Germany in 2024. FOAMed: Free Open Access Medical Education.



The final sample consisted of 71 paramedics and 32 physicians. Three paramedics were also studying medicine. Among the physicians, 21 worked regularly as prehospital emergency physicians, 16 regularly in an emergency department, and 1 in

XSL•FO
RenderX

cardiology (multiple responses possible). Further characteristics of the participants are shown in Table 2.

XSL•FO
RenderX

Table . Sociodemographic and professional characteristics of the participants in a controlled educational intervention comparing FOAMed[a] videos and podcast with print-based learning materials in Germany in 2024[b].

Characteristic	Total, n (%)	Intervention group	
		Print-based media, n (%)	FOAMed videos and podcasts, n (%)
Participants	103 (100)	58 (56.3)	45 (43.7)
Gender			
Women	40 (38.8)	25 (62.5)	15 (37.5)
Men	63 (61.2)	33 (52.3)	30 (47.6)
Nonbinary	0 (0)	0 (0)	0 (0)
Age (years)			
18 - 25	20 (19.4)	11 (55.0)	9 (45.0)
26 - 35	38 (36.9)	21 (55.3)	17 (44.7)
36 - 45	29 (28.1)	17 (58.6)	12 (41.4)
46 - 55	10 (9.7)	7 (70.0)	3 (30.0)
56 - 65	6 (5.8)	2 (33.3)	4 (66.7)
Median age (years)	35	35.1	35
Profession			
Paramedics	71 (68.9)	39 (54.9)	32 (55.1)
Simultaneously medical student	3 (4.2)	3 (100)	0 (0.0)
Medical doctors	32 (31.1)	19 (59.4)	13 (40.6)
Regularly working as a prehospital emergency physician[c]	21 (65.6)	13 (61.9)	8 (38.1)
Regularly working in emergency department[c]	16 (50.0)	10 (62.5)	6 (37.5)
Working in cardiology department[c]	1 (3.1)	0 (0.0)	1 (100.0)
Work experience (years)			
<1	12 (11.65)	6 (50.0)	6 (50.0)
1 - 3	24 (23.3)	14 (58.3)	10 (41.7)
4 - 7	24	16 (66.7)	8 (33.3)
8 - 10	9 (8.7)	6 (66.7)	3 (33.3)
>10	34 (33)	16 (47.0)	18 (53.0)
ECG[d] interpretations per week (n=101)			
1 - 10	52 (50.5)	27 (51.9)	25 (48.1)
11 - 25	40 (38.8)	24 (60.0)	16 (40.0)
>25	9 (8.7)	6 (66.7)	3 (33.3)
Sense of safety regarding ECG interpretation			
Safe	9 (8.7)	5 (55.5)	4 (44.5)
Rather safe	47 (45.6)	28 (59.6)	19 (40.4)
Rather unsafe	42 (40.8)	21 (50.0)	21 (50.0)
Unsafe	4 (3.9)	3 (75.0)	1 (25.0)

[a]FOAMed: Free Open Access Medical Education.

[b]The table summarizes demographic variables (gender and age), professional background differentiated into paramedics (including 3 simultaneously enrolled medical students) and medical doctors (with multiple possible roles: regularly working as prehospital emergency physicians, in emergency

XSL•FO
RenderX

departments, or in cardiology), work experience, weekly electrocardiogram interpretation frequency, and self-reported sense of safety regarding electrocardiogram interpretation, presented for the total sample and stratified by intervention group. The italicized entries indicate subcategories within the same response option. Multiple selections were possible, which is why the totals do not sum to 100%.

[c]Multiple responses possible.

[d]ECG: electrocardiogram.

The concealed, nonsystematic allocation resulted in a largely balanced distribution across both intervention groups. Individual peculiarities (eg, all 3 medical students in IG1, the only physician working in cardiology in IG2) had no relevant influence on the group balance.

The allocation of materials was controlled, and there is evidence that participants used the learning resources assigned to them. Protocol adherence was thus ensured. The study design, sample, type of data collected, validity of the evaluation tools, data analysis chosen, and the measured outcomes resulted in a MERSQI score [32] of 16.5 out of a maximum of 18 points. Furthermore, the requirements of the 7-step model of self-directed learning [34] and the methodological recommendations for ECG competence measurement [35] were considered.

Quantitative Findings

Both intervention groups showed a significant increase in knowledge after the 30-minute learning unit, with this increase being more pronounced in the FOAMed group (IG2; podcast + video) than in the print-based group (IG1; guideline + textbook).

For ECG interpretation, IG2 improved from 55% (T1) to 65.5% (T2), whereas IG1 increased from 57% to 60%. The mixed ANOVA showed a statistically significant interaction between test time (T1 vs T2) and intervention group, assuming sphericity ($F_{1,101}$=5.85, P=.02, partial η^2=0.055, 95% CI 0.001 to 0.158).

For the text-based questions, IG2 improved from 45% (T1) to 68% (T2), while IG1 increased from 47% to 63%. Here, too, there was a statistically significant interaction between test time and intervention group ($F_{1,101}$=4.65, P=.03, partial η^2=0.044, 95% CI 0.000 to 0.142).

The results of this development in knowledge are visualized in Table 1. Detailed inferential statistics (means, SDs, within-group change scores, between-group differences, 95% CIs, P values, and effect sizes) are provided in Tables S1 and S2 in Multimedia Appendix 3. CIs are reported to allow readers to assess the precision of the estimates.

IG2 (FOAMed podcast and videos) thus achieved significantly higher learning gains in both areas.

About recognizing ECG patterns requiring emergency cardiac catheterization, 63.6% (28/44) of IG2 participants (FOAMed videos and podcast) reported an increase in their sense of confidence, whereas only 34.5% (20/58) of IG1 participants made this statement. Regarding treatment decisions, 61.4% (27/44) of IG2 participants (FOAMed videos and podcast) reported an increase in their sense of confidence, whereas only 39.3% (22/56) of IG1 participants made this statement.

The FOAMed group (IG2) rated their training unit significantly more positively than the print-based group (IG1). The FOAMed video unit scored highest in the overall evaluation with a mean value of 1.57 (Likert scale: 1=completely agree, 5=completely disagree), followed by the textbooks (mean 2.2), the podcast (mean 2.4), and guidelines (mean 2.6). IG2 also showed the highest willingness to recommend (mean 1.7), while IG1 was more skeptical (mean 2.7). FOAMed videos were also rated as particularly accessible and easy to get used to (mean 4.1 on a scale of 1 = “very difficult to get used to” to 5 = “not at all”), whereas traditional print materials were rated much more critically. In the free-text responses, these were repeatedly described as dry, confusing, or too extensive (eg, “Guidelines are really confusing and there are no illustrative examples. Even in both textbooks, despite many illustrations, there are not enough example ECG” or “For the short time available, textbooks and guidelines are rather unwieldy. I had to read several chapters before I found the answer to a specific question.”) The training content was also very frequently described as very difficult and complicated (eg, “[Topic] is very complicated” or “Too many STEMI equivalents and exceptions”), which was not reflected in the evaluation by the FOAMed group. Table 3 below shows the specific media ratings by the participants.

XSL•FO
RenderX

Table . Evaluation of learning materials (books, guidelines, videos, and podcast) by participants of a controlled educational intervention comparing FOAMed[a] videos and podcasts with print-based learning materials in Germany in 2024[b].

	Books and guidelines (n=58)[c]		FOAMed videos and podcast (n=45)[d]	
	Books (n=58[e])	Guidelines (n=52[e])	Videos (n=44[e])	Podcast (n=30[e])
Entertaining	2.9	3.4	2.0 (n=43[f])	2.4 (n=28[f])
User-friendly	2.8	3.4	1.3	2.4
Comprehensible	2.3	2.9	1.5	2.4
Informative	1.7	1.9	1.4	1.6 (n=29[f])
Trustworthy	1.55	1.3 (n=51[f])	1.65	1.9
Mean value of the individual medium	2.25	2.58	1.5	2.41

[a]FOAMed: Free Open Access Medical Education.

[b]The table presents Likert-scale ratings (1=completely agree; 5=completely disagree) for perceived qualities of each medium, including entertainment value, user-friendliness, comprehensibility, informativeness, and trustworthiness. Mean values for each individual medium and average values for the 2 intervention units (books and guidelines vs FOAMed videos and podcast) are shown, alongside overall satisfaction, intention to recommend, and perceived need to get used to the medium. Participants only rated the materials they used during the self-learning session, and item-specific sample sizes vary because individual questions could be skipped.

[c]Mean value of the individual media: books=2.25; guidelines=2.58. Mean value of the intervention unit: books and guidelines=2.4. Overall satisfaction: books and guidelines: mean 2.7, n=57. Intention to recommend: books and guidelines: mean 2.7, n=57. Needing to get used to the medium (mean values): books=3.4; guidelines=2.2.

[d]Mean value of the individual media: videos=1.5; podcasts=2.41. Mean value of the intervention unit: videos and podcasts=1.9. Overall satisfaction: videos and podcasts: mean 1.8, n=45. Intention to recommend: videos and podcasts: mean 1.7, n=45. Needing to get used to the medium (mean values): videos=4.1; podcasts=2.8.

[e]Participants only rated the materials they actually used during the self-learning session; unused media were not evaluated.

[f]Individual item counts differ because participants could skip questions; missing responses were not imputed.

An exploratory subgroup analysis provided no evidence of differential effects, including gender, and was likely underpowered.

Qualitative Findings

The inductive summary content analysis yielded recurring descriptive categories that explained why some participants did not experience increased perceived safety after the intervention. Across both questions, participants most frequently cited *Zeitmangel*, *Orientierungsschwierigkeiten*, *Komplexität des Themas*, *Ungeeignetes Medium*, *Fehlendes Feedback*, *Merk-Schwierigkeiten*, *Fehlender Praxisbezug*, *Bereits stabiles Vorwissen*, *and Bereits bestehende Wissensdefizite*. These categories represent structural learning barriers and contextual factors that influenced perceived safety and were descriptive in nature rather than in-depth qualitative themes.

In IG1 (print-based materials), participants most often reported difficulties related to the lack of ECG examples, insufficient feedback, time pressure, unclear structure of the materials, and excessive demands due to the scope or technical language. Overall, IG1 participants perceived the topic as more difficult and cognitively demanding, frequently describing the material as complex, hard to navigate, or overwhelming within the limited time available. Several participants stated that the knowledge they had gained had no effect on their sense of safety, as local guidelines and not individual knowledge determined their clinical decision-making (“There are local guidelines and treatment pathways, so the newly acquired knowledge does not change anything” or “I feel equally (un)certain, as the treatment of ACS patients is determined by an SOP.”).

In IG2 (FOAMed videos + podcasts), didactic shortcomings were mentioned less frequently. Instead, participants more often referred to individual learning problems such as lack of prior knowledge, difficulties in remembering, or insufficient consolidation of newly acquired content. Some participants reported that they would need repeated exposure to the material to feel safer, or that the short duration of the learning unit limited their ability to internalize the content.

Overall, the qualitative findings illustrate why measurable knowledge gains did not uniformly translate into increased perceived safety. They highlight structural barriers (eg, time constraints, medium-related challenges, and lack of feedback) and individual factors (eg, prior knowledge and cognitive load) that help explain heterogeneous changes in perceived safety across groups.

In general, the knowledge gaps in the pre- and posttests affected both newer and established ECG patterns equally, with particularly significant improvements in the identification of de Winter sign and the left main occlusion ECG. STEMI mimics continued to be misinterpreted with striking frequency, for example, confusing pericarditis or hyperkalemia with STEMI equivalents, and frequent incorrect answers to everyday guideline content, such as required additional leads, recommended time intervals until the first 12-lead ECG diagnosis, or acetylsalicylic acid (ASA) dosage.

XSL•FO
RenderX

Discussion

Principal Findings

This controlled trial showed that a single 30-minute self-learning session led to clear improvements in ECG interpretation and text-based ACS-related clinical decision-making across all participants. In line with our primary objective, the curated FOAMed multimedia module produced greater short-term learning gains than the selected print-based reference materials under the 30-minute self-learning condition. These findings reflect differences between 2 real-world educational ecosystems rather than a categorical comparison of digital versus print media. Consistent with our secondary aims, FOAMed participants also reported larger increases in perceived confidence when recognizing high-risk ACS ECG patterns and making treatment decisions. Across all analyses, learning benefits were comparable for established and newer high-risk ACS ECG patterns, and no evidence of differential effects was observed; these exploratory analyses were likely underpowered.

The qualitative findings provide interpretive depth to the quantitative results by illustrating why measurable knowledge gains did not uniformly translate into increased perceived safety. Participants frequently described structural barriers such as limited time, difficulty navigating print-based materials, lack of feedback, and the inherent complexity of high-risk ECG patterns. Notably, IG1 participants perceived the topic itself as more difficult and cognitively demanding, suggesting that the medium shaped not only usability but also the perceived complexity of the content. This pattern may align with cognitive load theory [24,25]: some respondents described the print-based materials as difficult to navigate, which could suggest increased extraneous load and reduced cognitive resources for confidence-building. However, this exploratory analysis cannot establish a causal mechanism. In addition, several IG1 participants expressed a more defensive stance toward the relevance of the learning content (eg, reliance on standard operating procedure [SOP] rather than individual diagnostic judgment), which may be consistent with reduced perceived applicability or increased cognitive demands [24,25]. These qualitative insights may offer a possible explanation why confidence improved less consistently in IG1 despite measurable knowledge increases. Together, these responses suggest a possible explanation for the divergent quantitative pattern: some respondents perceived the FOAMed formats as less complex and cognitively demanding, whereas print-based materials were more often described as challenging to navigate. This analysis cannot establish a causal mechanism. Thus, the qualitative component adds explanatory value beyond the quantitative outcomes and helps interpret heterogeneous changes in perceived safety.

Importantly, these improvements in clinical decision-making refer to performance in hypothetical, text-based ACS scenarios rather than real-world clinical behavior, and confidence and sense of safety represent self-reported perceptions rather than objective measures of clinical competence.

Emergency medicine education has increasingly incorporated digital and multimedia-based learning, and our findings are broadly consistent with evidence supporting digital approaches for short-term educational outcomes in healthcare education [26,36,37].

Systematic review evidence suggests that simulation-based training can improve clinical skills acquisition and retention, although the underlying evidence is heterogeneous in quality [36]. Similarly, a systematic review of digital learning of clinical skills found evidence of beneficial effects on learners' academic performance, while also highlighting heterogeneity across digital modalities and outcomes [37].

In ECG education specifically, asynchronous electronic modules (e-modules) incorporating narrated videos, interactive questions, and feedback have shown potential as an alternative to traditional didactic teaching [38]. Evidence for podcasts is more limited: a recent randomized study reported educational benefits of podcast-based objective structured clinical examination (OSCE) training [39], while systematic review evidence suggests improvements in immediate knowledge retention but remains insufficient to establish broader effects on learning or clinical outcomes [40,41]. These benefits are consistent with cognitive load theory and the cognitive theory of multimedia learning, which posit that audiovisual formats can reduce extraneous load and facilitate the integration of complex visual information such as ECG patterns [24,25]. Our results extend this literature by demonstrating that FOAMed-based self-learning is not only well-received but also more effective than print-based materials for teaching high-risk ACS ECG patterns, a domain known to impose substantial cognitive demands and to be associated with persistent interpretation deficits among both physicians and paramedics [13,18,19]. The observation that print-based materials were perceived as more difficult to navigate is consistent with principles of cognitive load theory, according to which unnecessary demands associated with information presentation may increase extraneous cognitive load and reduce the cognitive resources available for learning [24,25]. Taken together, these findings suggest that curated digital FOAMed resources may better support efficient, practice-relevant learning in emergency care settings than traditional guideline- or textbook-based approaches.

A second key finding of this trial is that a brief, highly structured FOAMed module—combining podcasts and short videos—appeared particularly well suited to support self-directed learning of complex, high-risk ACS ECG patterns in a time-constrained emergency care context. This is consistent with evidence showing that simulation-based and digital learning formats can support clinical skills acquisition and educational outcomes when appropriately designed and aligned with learning objectives [36,37]. Recent work on podcasts in medical education similarly suggests that audio-based resources can promote flexible, self-paced learning and are generally well accepted by learners, although robust evidence on objective learning outcomes remains limited and heterogeneous [39-41]. Our findings extend this literature by demonstrating that FOAMed podcasts and videos can outperform traditional guideline and textbook materials for short, focused upskilling in ACS ECG interpretation and decision-making among practicing emergency care providers. Moreover, the strong learner acceptance and perceived confidence gains observed in

XSL•FO
RenderX

the FOAMed are consistent with prior work on self-directed learning and the use of freely accessible educational resources among emergency care professionals [23,42,43]. Such resources may help reduce barriers to continuing education by providing flexible, practice-relevant learning that can be integrated into time-constrained clinical work.

The interprofessional relevance of these findings is underscored by the persistent ECG interpretation deficits across emergency care professions. Prior studies have shown that both physicians and paramedics frequently struggle with ECG interpretation, including recognition of high-risk ACS patterns and STEMI equivalents, with accuracy remaining suboptimal even among experienced clinicians [13,14,18,19]. Recent work further emphasizes the need to recognize a broader spectrum of ischemic ECG patterns beyond classic STEMI criteria, including visually subtle patterns that may be overlooked [10,11]. Against this backdrop, the exploratory subgroup analyses provided no evidence of differential effects, although these analyses were likely underpowered. Within this limitation, the pattern of findings suggests that FOAMed-based learning may offer a scalable and equitable approach to strengthening diagnostic competencies across diverse emergency care roles [23,42,43]. This is particularly relevant given the well-described challenges of maintaining structured continuing education in shift-based systems, where staffing shortages, time constraints, and limited access to formal training opportunities can hinder ongoing skill development [22,44]. By demonstrating that even a brief, self-learning digital intervention can enhance both recognition of high-risk ACS ECG patterns and confidence in clinical decision-making, our findings highlight the potential of FOAMed resources to contribute to interprofessional preparedness, although such implications remain hypothetical. This study also contributes to the broader literature on ECG education by addressing several gaps that have been highlighted in recent research. Prior studies have demonstrated that digital learning formats, including asynchronous e-modules and other structured digital approaches, can improve ECG interpretation skills among medical students and clinicians [38,45]. However, most existing interventions focus on ECG interpretation more broadly, while evidence on teaching newer high-risk ACS and OMI-related ECG patterns remains limited [9,10,27,28]. Recent literature highlights that these patterns are visually subtle, may be underrepresented in conventional ECG teaching, and can be misinterpreted in clinical practice [9-11,27,28]. By incorporating selected high-risk ACS ECG patterns—including guideline-recognized STEMI equivalents and previously described emerging OMI-related patterns—into a controlled design and comparing FOAMed resources directly with traditional print materials, our study extends prior work by demonstrating that digital, case-based multimedia formats can support the teaching of these clinically relevant diagnostic concepts. Furthermore, while earlier studies have often been restricted to single professions or student populations, our interprofessional sample reflects the real-world diversity of emergency care teams and addresses the need for practice-oriented ECG training across professional groups [16,18,19,27]. In this way, the study provides new evidence that curated FOAMed resources can support the acquisition of both established and emerging ischemic ECG patterns across multiple emergency care roles.

Taken together, these findings highlight several contributions to this work. First, the study is innovative in directly comparing a curated FOAMed multimedia module with full-length, deliberately selected print-based reference materials—including the current ACS guideline and a structured textbook chapter—for teaching complex high-risk ACS ECG patterns in an interprofessional emergency care population. Second, it extends existing research by focusing on subtle, emerging ischemic patterns and by evaluating realistic, time-limited self-learning formats rather than isolated digital tools or student samples [27,28,38]. Third, the results contribute new empirical evidence that curated FOAMed resources can improve short-term ECG interpretation and ACS decision-making compared with traditional print-based materials under constrained learning conditions. Finally, while real-world implications remain hypothetical, the findings suggest that concise, high-quality digital modules may represent a feasible and accessible supplement to structured continuing education in shift-based emergency care environments [23,42,43].

Limitations

This study has several limitations that should be considered when interpreting the findings. First, voluntary participation introduces the possibility of self-selection bias, as individuals with a particular interest in ECG education or digital learning may have been more likely to enroll. Although baseline confidence levels were broadly comparable across groups, recruitment through social media for the final available study slots may have influenced participants' attitudes toward the media formats. Minor deviations from the intended 1:1 allocation resulted solely from participant no-shows, leaving some preallocated seats unoccupied; this represents a small design limitation. Because the allocation was practically random but not based on a formal computer-generated sequence, the study should be interpreted as using concealed, nonsystematic allocation rather than strict algorithmic randomization. Second, the study focused on a highly specific set of selected high-risk ACS ECG patterns, including guideline-recognized STEMI equivalents and previously described emerging patterns, which limits the generalizability of the results to other clinical topics or educational contexts. Interactive digital formats with feedback functions—which may provide additional educational benefits [46]—were deliberately excluded to maintain feasibility and comparability with print-based materials. Similarly, the decision to contrast freely accessible FOAMed resources with comprehensive but often costly guideline and textbook materials reflected a pragmatic design choice but restricts conclusions about other types of digital or print-based instruction. A further methodological limitation concerns potential allegiance bias, as 2 authors had prior involvement with one FOAMed source; although the selection of materials followed predefined criteria and external verification, prior familiarity cannot be fully excluded as an influencing factor.

Additionally, the fixed 30-minute learning period was perceived by several participants as insufficient for engaging deeply with the material. It is plausible that greater learning gains might

XSL•FO
RenderX

have been achieved under more flexible or realistic time conditions. Long-term retention of knowledge could not be assessed within the study timeframe, and follow-up data on sustainability are therefore not yet available. Additionally, although the sample size exceeded 30 participants per group, the target number derived from a priori power analysis was not fully reached. Nevertheless, the achieved sample size provided sufficient data to detect the observed between-group effects, although the study may have been underpowered for smaller effects and subgroup analyses [47-49]. Because no statistical adjustments for multiple comparisons were applied, the risk of type I error is increased, and secondary findings should be interpreted as exploratory.

The qualitative component is subject to additional limitations. The free-text responses were brief (typically one short sentence), optional, and provided only by a subset of participants (54 and 51 eligible respondents), which limits contextual depth and representativeness. As a result, the inductively derived categories reflected descriptive patterns rather than in-depth qualitative themes. Nevertheless, the qualitative findings offer meaningful contextual insights into heterogeneous changes in perceived safety and help explain why knowledge gains did not uniformly translate into increased confidence. They should therefore be interpreted as exploratory, complementary evidence that supports and enriches the quantitative results rather than as stand-alone qualitative themes.

Conclusion

This controlled trial demonstrates that, under a 30-minute self-learning condition, a curated FOAMed multimedia module can produce greater short-term improvements in ECG interpretation accuracy and text-based ACS decision-making than the full-length print-based reference materials. By directly comparing the tested curated FOAMed module with the tested full-length print-based materials, the study provides novel evidence that the curated FOAMed module was experienced as more accessible within the specific 30-minute condition, rather than supporting general conclusions about multimedia formats. Although the ECG topic is inherently complex, participants experienced the tested FOAMed module as more approachable and less cognitively demanding when delivered through FOAMed formats rather than the print-based texts. The qualitative findings suggest that some print-based materials are more complex and less directly applicable, which may offer a possible explanation for the less consistent confidence gains in the print group despite measurable knowledge improvement. These findings extend existing ECG education research by showing that even emerging ischemic patterns—often underrepresented in guidelines and difficult to teach in conventional formats—can be effectively conveyed through short, focused digital modules. All findings reflect immediate postintervention knowledge and self-reported perceptions only; the study did not assess real-world clinical decision-making, diagnostic performance, patient outcomes, or long-term retention.

The results also highlight the hypothetical potential of FOAMed to complement formal training structures, particularly in shift-based emergency care environments where time constraints and limited access to structured education remain persistent barriers. Integrating high-quality, peer-reviewed digital resources into continuing education may have potential implications, but such implications remain hypothetical and require empirical testing, as this study assessed short-term knowledge gains rather than observed clinical performance. Future research should examine long-term retention, clinical impact, and the optimal combination of digital and interactive learning strategies to further enhance ECG education and emergency cardiovascular care.

Acknowledgments

We thank Thieme and S+K for providing textbooks used in the print-based group, and Apple Inc for supporting data collection by loaning iPads and keyboards free of charge as part of an Apple Education Program. We also acknowledge the Notfallguru project for contributing books for the prize draw.

The authors declare the use of generative AI (GenAI) in the research and writing process. According to the GAIDeT (Generative AI Delegation Taxonomy; 2025), the following tasks were delegated to GenAI tools under full human supervision: proofreading and editing, summarizing text, translation, and reformatting. The GenAI tools used were Microsoft Copilot (version not specified by provider) and DeepL Translator 26.5.1. Responsibility for the final manuscript lies entirely with the authors. GenAI tools are not listed as authors and do not bear responsibility for the final outcomes. This declaration was submitted by Wiebke Turner (corresponding author) on behalf of all authors. "Summarizing selected passages" refers to generating concise summaries of author-written sections from a previous dissertation to support restructuring and drafting.

Funding

The authors declared no financial support was received for this work.

Data Availability

The datasets generated and analyzed during this study (including deidentified Excel files, the questionnaires used, SPSS syntax for inferential analyses, and the coding framework for the qualitative analysis) are not publicly available due to participant privacy considerations but are available from the corresponding author upon reasonable request.

XSL•FO
RenderX

Authors' Contributions

WT conducted the controlled educational trial, developed the entry and exit questionnaires, created the text-based questions, selected the ECG cases, performed the descriptive analyses, and drafted the manuscript. KF provided cardiological supervision, contributed to the selection of ECGs and formulation of text questions, ensured cardiological accuracy, and critically revised the manuscript. MF was responsible for participant recruitment, participated in the coding circle for the qualitative Mayring analysis, and critically revised the manuscript. JPE, as head of the department and doctoral supervisor, advised on study design and procedures, and critically reviewed the manuscript. JN performed the statistical analyses, provided continuous advice on scientific methodology, and contributed to the didactic background and manuscript editing. All authors approved the final version of the manuscript.

Conflicts of Interest

Two authors (MF and KF) had previously contributed to the FOAMed project Nerdfallmedizin, and some of the videos used in this study were originally produced by the Nerdfallmedizin team. These materials were created long before the dissertation project was conceived and before the study tasks were defined. WT is part of the broader Nerdfallmedizin project but was not involved in creating or editing any of the specific materials used in this trial. The selection of digital materials was based on objective criteria (national reach, FOAMed reputation, prior awards) and was independently verified by a senior emergency cardiologist. The authors' prior involvement in Nerdfallmedizin did not influence the selection of materials; the videos were chosen because they represented the most widely used, award-winning, and highest-quality FOAMed resources available at the time of study conception. Participants were informed in the study invitation that Nerdfallmedizin, PinUpDocs, Thieme, S+K Verlag, and Apple were involved, but no information was provided about which materials would be used in which study arm. All digital content was frozen for the duration of the study, and the standardized study design, allocation procedure, and identical 30-minute exposure were implemented to minimize allegiance bias and ensure equal learning opportunities across both groups. No financial benefits resulted from this engagement. The project was selected for the intervention because, at the time of study conception, Nerdfallmedizin was the most widely subscribed FOAMed blog and YouTube channel in the German-speaking emergency medicine community. The authors declare no other conflicts of interest.

Multimedia Appendix 1
Intervention materials (digital and print resources used in the study).
[PDF File, 353 KB - mededu_v12i1e87587_app1.pdf]

Multimedia Appendix 2
Example test items.
[DOCX File, 2169 KB - mededu_v12i1e87587_app2.docx]

Multimedia Appendix 3
Detailed inferential statistics for electrocardiogram interpretation accuracy and text-based acute coronary syndrome clinical decision-making.
[PDF File, 48 KB - mededu_v12i1e87587_app3.pdf]

Checklist 1
Application of the MERSQI checklist to this study.
[PDF File, 76 KB - mededu_v12i1e87587_app4.pdf]

Checklist 2
CONSORT-eHEALTH checklist (V 1.6.x).
[PDF File, 691 KB - mededu_v12i1e87587_app5.pdf]

References

1. Byrne RA, Rossello X, Coughlan JJ, et al. 2023 ESC Guidelines for the management of acute coronary syndromes. Eur Heart J 2023 Oct 12;44(38):3720-3826. [doi: 10.1093/eurheartj/ehad191] [Medline: 37622654]
2. Ibanez B, James S, Agewall S, et al. 2017 ESC Guidelines for the management of acute myocardial infarction in patients presenting with ST-segment elevation: the task Force for the management of acute myocardial infarction in patients presenting with ST-segment elevation of the European Society of Cardiology (ESC). Eur Heart J 2018 Jan 7;39(2):119-177. [doi: 10.1093/eurheartj/ehx393] [Medline: 28886621]
3. Sovari AA, Assadi R, Lakshminarayanan B, Kocheril AG. Hyperacute T wave, the early sign of myocardial infarction. Am J Emerg Med 2007 Sep;25(7):859. [doi: 10.1016/j.ajem.2007.02.005] [Medline: 17870505]

XSL•FO
RenderX

4. de Winter RW, Adams R, Verouden NJW, de Winter RJ. Precordial junctional ST-segment depression with tall symmetric T-waves signifying proximal LAD occlusion, case reports of STEMI equivalence. J Electrocardiol 2016;49(1):76-80. [doi: 10.1016/j.jelectrocard.2015.10.005] [Medline: 26560436]
5. de Zwaan C, Bär FW, Janssen JH, et al. Angiographic and clinical characteristics of patients with unstable angina showing an ECG pattern indicating critical narrowing of the proximal LAD coronary artery. Am Heart J 1989 Mar;117(3):657-665. [doi: 10.1016/0002-8703(89)90742-4] [Medline: 2784024]
6. Jaiswal AK, Shah S. Shark fin electrocardiogram: a deadly electrocardiogram pattern in ST-Elevation Myocardial Infarction (STEMI). Cureus 2021 Jun;13(6):e15989. [doi: 10.7759/cureus.15989] [Medline: 34336480]
7. Aslanger E, Yıldırımtürk Ö, Şimşek B, et al. A new electrocardiographic pattern indicating inferior myocardial infarction. J Electrocardiol 2020;61:41-46. [doi: 10.1016/j.jelectrocard.2020.04.008] [Medline: 32526537]
8. Rokos IC, French WJ, Mattu A, et al. Appropriate cardiac cath lab activation: optimizing electrocardiogram interpretation and clinical decision-making for acute ST-elevation myocardial infarction. Am Heart J 2010 Dec;160(6):995-1003. [doi: 10.1016/j.ahj.2010.08.011] [Medline: 21146650]
9. Ricci F, Martini C, Scordo DM, et al. ECG patterns of occlusion myocardial infarction: a narrative review. Ann Emerg Med 2025 Apr;85(4):330-340. [doi: 10.1016/j.annemergmed.2024.11.019] [Medline: 39818676]
10. Aslanger E, Aggul B, Yildirimturk O, Karabay CY, Degertekin M. The new occlusive/non-occlusive MI paradigm to remedy the failure of high risk non-ST elevation MI criteria in the real world - insights from the DIFOCCULT-3 trial. Eur Heart J Suppl 2026 Mar 30;28(Supplement_3). [doi: 10.1093/eurheartjsupp/suag056.033]
11. Grautoff S, Fessele K, Fandler M, Knappen N, Gotthardt P. STEMI mimics: ST elevations on ECG: alternative diagnoses to acute coronary occlusion. Med Klin Intensivmed Notfmed 2023 Feb;118(1):35-44. [doi: 10.1007/s00063-021-00882-5] [Medline: 34709428]
12. Cook DA, Oh SY, Pusic MV. Accuracy of physicians' electrocardiogram interpretations: a systematic review and meta-analysis. JAMA Intern Med 2020 Nov 1;180(11):1461-1471. [doi: 10.1001/jamainternmed.2020.3989] [Medline: 32986084]
13. Kashou AH, Noseworthy PA, Beckman TJ, et al. ECG interpretation proficiency of healthcare professionals. Curr Probl Cardiol 2023 Oct;48(10):101924. [doi: 10.1016/j.cpcardiol.2023.101924] [Medline: 37394202]
14. Attar ET. ECG interpretation abilities in clinical practice: examining the role of expertise, age, and gender. Medicine (Baltimore) 2025 Aug 8;104(32):e42401. [doi: 10.1097/MD.0000000000042401] [Medline: 40797395]
15. Perrichot A, Vaittinada Ayar P, Taboulet P, et al. Assessment of real-time electrocardiogram effects on interpretation quality by emergency physicians. BMC Med Educ 2023 Sep 18;23(1):677. [doi: 10.1186/s12909-023-04670-x] [Medline: 37723508]
16. Yokoyama M, Iyama K, Honda T, et al. Attitude of emergency medical technicians toward electrocardiogram - needs of electrocardiogram training courses and other learning opportunities. Circ J 2024 Jul 25;88(8):1315-1321. [doi: 10.1253/circj.CJ-23-0469] [Medline: 37704439]
17. Sia CH, Chew NWS, Cheong CWS, et al. Fear of electrocardiogram interpretation (ECGphobia) among medical students and junior doctors. Singapore Med J 2022 Dec;63(12):763-768. [doi: 10.11622/smedj.2021078] [Medline: 34157806]
18. Wyld L, Clegg L, Griffin K. Paramedic electrocardiogram interpretation in acute coronary syndrome. J Paramed Pract 2023 Dec 2;15(12):496-504. [doi: 10.12968/jpar.2023.15.12.496]
19. Funder JL, Ross L, Ryan S. How effective are paramedics at interpreting ECGs in order to recognize STEMI? A systematic review. Australas J Paramed 2020 Jan;17:1-9. [doi: 10.33151/ajp.17.772]
20. Barthelemy FX, Segard J, Fradin P, et al. ECG interpretation in Emergency Department residents: an update and e-learning as a resource to improve skills. Eur J Emerg Med 2017 Apr;24(2):149-156. [doi: 10.1097/MEJ.0000000000000312] [Medline: 26287803]
21. Breen CJ, Kelly GP, Kernohan WG. ECG interpretation skill acquisition: a review of learning, teaching and assessment. J Electrocardiol 2022;73:125-128. [doi: 10.1016/j.jelectrocard.2019.03.010] [Medline: 31005264]
22. Hock SM, Cassara M, Aghera A, Saloum D, Bentley SK. Attending physicians as simulation learners: summary of current practices and barriers in emergency medicine. Clin Exp Emerg Med 2024 Jun;11(2):224-228. [doi: 10.15441/ceem.23.137] [Medline: 38286502]
23. Morgan J. Free Open Access Medical Education (FOAMed) use in medical students: a literature review. BMC Med Educ 2025 Jan 4;25(1):20. [doi: 10.1186/s12909-024-06392-0] [Medline: 39755631]
24. Mayer RE. Applying the science of learning: evidence-based principles for the design of multimedia instruction. Am Psychol 2008 Nov;63(8):760-769. [doi: 10.1037/0003-066X.63.8.760] [Medline: 19014238]
25. Sweller J. Cognitive load during problem solving: effects on learning. Cogn Sci 1988 Apr;12(2):257-285. [doi: 10.1207/s15516709cog1202_4]
26. Regmi K, Jones L. A systematic review of the factors - enablers and barriers - affecting e-learning in health sciences education. BMC Med Educ 2020 Mar 30;20(1):91. [doi: 10.1186/s12909-020-02007-6] [Medline: 32228560]
27. Oh SY, Cook DA, Van Gerven PWM, et al. Physician training for electrocardiogram interpretation: a systematic review and meta-analysis. Acad Med 2022 Apr 1;97(4):593-602. [doi: 10.1097/ACM.0000000000004607] [Medline: 35086115]
28. Ardekani A, Hider AM, Rastegar Kazerooni AA, et al. Surfing the clinical trials of ECG teaching to medical students: a systematic review. J Educ Health Promot 2023;12:107. [doi: 10.4103/jehp.jehp_780_22] [Medline: 37288415]

XSL•FO
RenderX

29. Clinical trials registration: frequently asked questions. International Committee of Medical Journal Editors. URL: https://www.icmje.org/about-icmje/faqs/clinical-trials-registration/ [accessed 2026-08-09]
30. Hopewell S, Chan AW, Collins GS, et al. CONSORT 2025 statement: updated guideline for reporting randomised trials. Lancet 2025 Apr 14:S0140-6736(25)00672-5. [doi: 10.1016/S0140-6736(25)00672-5] [Medline: 40245901]
31. Dr Smith's ECG Blog. URL: https://drsmithsecgblog.com/ [accessed 2026-09-09]
32. Reed DA, Cook DA, Beckman TJ, Levine RB, Kern DE, Wright SM. Association between funding and quality of published medical education research. JAMA 2007 Sep 5;298(9):1002-1009. [doi: 10.1001/jama.298.9.1002] [Medline: 17785645]
33. Mayring P. Qualitative Content Analysis: Foundations and Techniques, 12th edition: Beltz; 2015. URL: https://www.fachportal-paedagogik.de/literatur/vollanzeige.html?FId=3232124 [accessed 2026-08-30]
34. Taylor TAH, Kemp K, Mi M, Lerchenfeldt S. Self-directed learning assessment practices in undergraduate health professions education: a systematic review. Med Educ Online 2023 Dec;28(1):2189553. [doi: 10.1080/10872981.2023.2189553] [Medline: 36919556]
35. Cook DA, Oh SY, Pusic MV. Assessments of physicians' electrocardiogram interpretation skill: a systematic review. Acad Med 2022 Apr 1;97(4):603-615. [doi: 10.1097/ACM.0000000000004140] [Medline: 33913438]
36. Ajemba MN, Ikwe C, Iroanya JC. Effectiveness of simulation-based training in medical education: assessing the impact of simulation-based training on clinical skills acquisition and retention: a systematic review. World J Adv Res Rev 2024;21(1):1833-1843 [FREE Full text] [doi: 10.30574/wjarr.2024.21.1.0163]
37. McGee RG, Wark S, Mwangi F, et al. Digital learning of clinical skills and its impact on medical students' academic performance: a systematic review. BMC Med Educ 2024 Dec 18;24(1):1477. [doi: 10.1186/s12909-024-06471-2] [Medline: 39696150]
38. Olvet DM, Sadigh K. Comparing the effectiveness of asynchronous e-modules and didactic lectures to teach electrocardiogram interpretation to first year US medical students. BMC Med Educ 2023 May 22;23(1):360. [doi: 10.1186/s12909-023-04338-6] [Medline: 37217893]
39. Dupont V, Guerrot D, Bouazzi L, et al. Evaluating podcasts as a tool for OSCE training: a randomized trial using generative AI-powered simulation. BMC Med Educ 2025 Mar 24;25(1):429. [doi: 10.1186/s12909-025-06675-0] [Medline: 40128772]
40. Wang S, Meckling G, Hanafi S, et al. The effects of podcasts on medical education: a systematic review. Research Square. Preprint posted online on Oct 3, 2023. [doi: 10.21203/rs.3.rs-3297034/v1]
41. Caldwell KE, Zarate-Rodriguez JG, Fox JC, Yaeger L, Wise PE. Listen up: a systematic review of the utilization and efficacy of podcasts for medical education. Global Surg Educ 2024;3(1):107. [doi: 10.1007/s44186-024-00301-1]
42. Bryant J, Zucca A, Turon H, Sanson-Fisher R, Morrison A. Attitudes towards and engagement in self-directed learning among paramedics in New South Wales, Australia: a cross sectional study. BMC Med Educ 2023 Oct 12;23(1):759. [doi: 10.1186/s12909-023-04740-0] [Medline: 37828464]
43. Khadpe J, Belcher CN, Damore K, et al. Emergency medicine procedures: a review of approved instructional resources from the world of Free Open Access Medical Education. Cureus 2023 Sep;15(9):e45986. [doi: 10.7759/cureus.45986] [Medline: 37900377]
44. Ghattas AHS, Abdou HA. Challenges and best practices for moving forward in interprofessional collaboration in critical care units: nurses' perspectives. BMC Nurs 2025 Mar 25;24(1):317. [doi: 10.1186/s12912-025-02860-0] [Medline: 40133871]
45. Chaumont C, Morgat C, Ollitrault P, et al. How to improve medical students' ECG interpretation skills? Multicenter survey and results of a comparative study evaluating a new educational approach. BMC Med Educ 2024 Sep 9;24(1):979. [doi: 10.1186/s12909-024-05929-7] [Medline: 39252033]
46. Bognár L, Fauszt T, Váraljai M. The impact of online quizzes on student success. Int J Emerg Technol Learn 2021;16(11):225. [doi: 10.3991/ijet.v16i11.21679]
47. Bortz J, Schuster C. Statistics for Humanities and Social Scientists, 7th edition: Springer-Verlag; 2010. URL: https://www.lehrbuch-psychologie.springernature.com/statistik-f%C3%BCr-human-und-sozialwissenschaftler-7-aufl-2010 [accessed 2026-08-30] [doi: 10.1007/978-3-642-12770-0]
48. Kähler WM. Statistical Data Analysis: Understanding Methods and Using Them Skillfully with SPSS, 7th edition: Vieweg + Teubner; 2011. URL: https://pollux-fid.de/r/kxp-666231788 [accessed 2026-08-30]
49. Tavakoli H. A Dictionary of Research Methodology and Statistics in Applied Linguistics: Rahnamā; 2013. URL: https://search.worldcat.org/de/title/954825122 [accessed 2026-08-30]

Abbreviations

ACC: American College of Cardiology
ACS: acute coronary syndrome
ASA: acetylsalicylic acid
CONSORT: Consolidated Standards of Reporting Trials
DGK: Deutsche Gesellschaft für Kardiologie
e-module: electronic module
ECG: electrocardiogram

XSL•FO
RenderX

ESC: European Society of Cardiology
FOAMed: Free Open Access Medical Education
GDPR: General Data Protection Regulation
ICMJE: International Committee of Medical Journal Editors
IG1: intervention group 1
IG2: intervention group 2
ITT: intention-to-treat
MCAR: Missing Completely At Random
MERSQI: Medical Education Research Study Quality Instrument
NSTE-ACS: non–ST-segment elevation acute coronary syndrome
OCSE: Objective Structured Clinical Examination
OMI: Occlusion Myocardial Infarction
PP: per-protocol
Q1: baseline questionnaire
Q2: final evaluation questionnaire
SOP: standard operating procedure
STEMI: ST-segment elevation myocardial infarction
T1: pretest
T2: posttest
TIDieR: Template for Intervention Description and Replication

Edited by S Brini; submitted 11.Nov.2025; peer-reviewed by LTO Cornejo, N Spitznagel; revised version received 11.Aug.2026; accepted 14.Aug.2026; published 15.Sep.2026.

Please cite as:
Turner W, Fessele K, Fandler M, Ehlers JP, Nitsche J
Teaching Acute Coronary Syndrome High-Risk ECG Interpretation and Clinical Decision-Making Through FOAMed Videos and Podcast Versus Print-Based Materials Among Emergency Care Providers: Randomized Controlled Mixed Methods Trial
JMIR Med Educ 2026;12:e87587
URL: https://mededu.jmir.org/2026/1/e87587
doi:10.2196/87587

© Wiebke Turner, Klaus Fessele, Martin Fandler, Jan P Ehlers, Julia Nitsche. Originally published in JMIR Medical Education (https://mededu.jmir.org), 15.Sep.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Education, is properly cited. The complete bibliographic information, a link to the original publication on https://mededu.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Influence of Admission Pathways on Learning Strategies, Assessment Engagement, and Academic Performance Among First-Year Medical Students: Mixed Methods Retrospective Observational and Cross-Sectional Survey Study

Issarawan Keadkraichaiwat, MSc; Chantacha Sitticharoon, MD, PhD; Punyapat Maprapho, MD; Nisa Jangboon, BEd, MSc; Nadda Wannarat, BA, MEd

Faculty of Medicine Siriraj Hospital, Mahidol University, 2 Wanglang Rd, Siriraj, Bangkoknoi, Thailand

Corresponding Author:
Chantacha Sitticharoon, MD, PhD
Faculty of Medicine Siriraj Hospital, Mahidol University, 2 Wanglang Rd, Siriraj, Bangkoknoi, Thailand

Abstract

Background: Medical school admission pathways are designed to select suitable applicants, with different approaches potentially impacting students' learning behaviors and performance.

Objective: This study aimed to compare students' self-regulated learning (SRL) strategies, assessment engagement statistics (AES), nongrading evaluation (Outstanding ["O"]/Satisfactory ["S"]/Unsatisfactory ["U"]) preferences, and academic performance across admission pathways, and analyze correlations and linear regression models among summative scores, AES, and course learning outcome (CLO) scores.

Methods: This mixed methods retrospective observational and cross-sectional survey study used census sampling with selection criteria of all enrolled first-year medical students in 2021 (N=319) across 4 admission pathways: academic (n=23), quota (n=6), test (n=261), and rural (n=29). Demographics included age (19 - 24 years) and sex (167/319, 52.4% male). AES, CLO scores, and summative scores were obtained from institutional databases. Two system-embedded institutional questionnaires assessed SRL strategies (316/319, 99.1% response rate) and "O"/"S"/"U" preferences (299/319, 93.7% response rate). Outcome measures included SRL strategies, AES, "O"/"S"/"U" preferences, CLO scores, and summative scores. Statistical significance was set at $P<.05$.

Results: When compared among pathways, using one-way ANOVA with Fisher least significant difference post hoc tests, the academic group reported significantly higher mean (with 95% CI) goal setting (4.35, 4.07 - 4.63), enthusiasm (4.43, 4.18 - 4.69), and lower stress during study (2.64, 2.15 - 3.12), while the rural group showed higher pre-examination stress (4.38, 4.10 - 4.66) (all $P<.05$). Most academic (14/22, 63.6%), quota (5/6, 83.3%), and test students (132/243, 54.3%) preferred "O"/ "S"/"U," while the rural students preferred "S"/"U" (13/28, 46.4%). The academic group showed significantly higher CLO and summative scores but fewer total and intentional attempts and instances of first-pass and highest scoring attempts (all $P<.05$), whereas the rural group showed significantly lower CLO and summative scores and higher instances of first-pass and highest scoring attempts (all $P<.05$). For correlation analyses, using Pearson correlation coefficient, summative scores were positively correlated with CLO scores and number of passings and negatively with first-pass attempts. For multiple linear regression analyses, summative scores were positively influenced by number of passings for each CLO and CLO scores and negatively influenced by instances of first-pass attempts and highest scoring attempts. Overall, the academic group demonstrated higher academic performance and fewer attempts and instances of first-pass and highest scoring attempts, while the rural group showed lower academic performance, requiring more attempts for first-passing CLOs.

Conclusions: Admission pathways significantly influence students' SRL strategies, AES, evaluation preferences, and academic performance. This study is innovative in analyzing these interconnected components within a single cohort, unlike prior research that examined them separately. By integrating assessment-engagement analytics with SRL data, it offers equity-oriented evidence on how admission systems shape learning behaviors and academic trajectories. These findings provide actionable insights for inclusive curriculum design and early identification of at-risk students. Real-world implications include targeted mentoring, SRL-focused interventions, and assessment reforms balancing academic rigor with psychological safety.

(JMIR Med Educ 2026;12:e68636) doi:10.2196/68636

XSL•FO
RenderX

KEYWORDS

admission systems; self-regulated learning strategies; assessment engagement statistics; academic outcome; course learning outcome; nongrading evaluation

Introduction

To address the diversity of medical curricula, institutions use various admission pathways, which may impact students' academic outcomes, self-regulated learning (SRL) strategies, assessment engagement statistics, and evaluation preferences. In comparisons among admission pathways, there were differences in academic outcomes and the likelihood of on-time graduation [1]. In terms of academic performance, students admitted through the Open National Unified Admission track, characterized by the highest competitiveness and a reserved quota for underprivileged students, predominantly comprising children of military personnel and government employees within the Ministry of Education in Jordan, had significantly higher graduate grade point average (GPA) than those in other tracks, whereas students from the other pathway track and an international track exhibited the lowest graduating GPA [1]. Furthermore, students entering through the GENERAL admission pathway in New Zealand demonstrated significantly higher GPAs in years 2 - 3 than students admitted through the Māori and Pacific Admission Scheme, which was designed to address shortages in the Māori and Pacific health workforce, reduce significant disparities in health outcomes, and uphold the rights of Māori within the New Zealand pathway [2]. In addition, multiple regression analysis demonstrated that the undergraduate entry pathway positively contributed to GPA in years 2 - 3 in New Zealand [2].

For students from various academic backgrounds, health-related undergraduate degree students consistently surpassed those from BioMed, Science, Humanities, and Business undergraduate degree programs in performance on medical science, clinical practice, objective structured clinical examinations, and overall performance when compared to other pathways in Australia [3].

For graduation rates, students admitted through the competitive Open National Unified Admission track and the reserved quota for underprivileged student track achieved the highest percentage of on-time graduation, followed by the children of university staff track in Jordan [1]. Similarly, students who gained admission through the GENERAL pathway in New Zealand exhibited a greater likelihood of completing their intended program than those from the Māori and Pacific Admission Scheme pathway [2].

Admission to the Faculty of Medicine Siriraj Hospital in the 2021 academic year followed the Thai University Central Admission System (TCAS) framework and aligned with the broader educational policy goals of Thailand's national medical school admission framework, which emphasizes not only academic merit but also equity, inclusion, and workforce distribution [4].

Importantly, at the time of the 2021 admission cycle, TCAS is structured as 4 sequential rounds and operates as a binding system, that is, once students are accepted and confirm their place in an earlier round, they cannot participate in subsequent rounds. This structure ensures efficient allocation, reduces redundancy, and facilitates early selection of high-priority candidates. Round 1 emphasizes portfolio submission and interviews, making it suitable for applicants with strong academic records and extracurricular achievements, while round 2 follows a quota system designed for students from specific regions or with special talents. Round 3 relies on standardized examinations such as the Thai General Aptitude Test, the Thai Professional Aptitude Test (TPAT), and core subjects, favoring applicants with strong test performance, whereas round 4 enables institutions with unfilled seats to admit students through their own criteria, offering opportunities for applicants not accepted in earlier rounds [5].

At our institution, via 3 TCAS rounds, 4 pathways (groups) were established: round 1—the academic group targeting high-potential students with demonstrated cognitive strengths and scientific aptitude, particularly those participating in national science Olympiads; round 2—the quota group, promoting diversity by admitting students with exceptional talents in music or sports and individuals holding a bachelor's degree; and round 3—the central admission test pathway (test group) and the pathway for students required to serve as rural doctors (rural group), prioritizing academic achievement through competitive entrance examinations [5,6]. Notably, students in the rural group were officially enrolled under the Faculty of Medicine, Praboromarajchanok Institute, but undertook their preclinical study together with students of the Faculty of Medicine Siriraj Hospital.

Differences in admission pathways may be associated with variations in students' SRL behaviors. According to Zimmerman's cyclical model, SRL encompasses 3 interrelated phases: the forethought phase, which involves goal setting, task analysis, and motivational beliefs; the performance phase, which includes self-control, monitoring, and strategy implementation; and the self-reflection phase, which focuses on self-evaluation and adaptive reactions after learning tasks [7]. SRL may influence students across various admission systems, since it influences how they approach learning tasks, manage challenges, and engage with evaluation, and plays a crucial role in their academic success, particularly among students with different backgrounds [8].

In addition, nongrading or mastery-based systems, such as pass/fail or pass/satisfactory/outstanding formats, have gained attention for their potential to reduce stress [9], promote intrinsic motivation [10-12], and foster deeper, self-directed learning [10]. These approaches are especially relevant in high-stakes disciplines such as medicine, where long-term competence and well-being are critical [13], probably reducing competitiveness and anxiety while promoting equity and fostering collaborative learning environments [10,14,15]. Despite the growing adoption of nongrading evaluation systems in medical education as part of curricular reform, limited evidence exists on how students

XSL•FO
RenderX

from different admission tracks perceive these systems or how such reforms influence their SRL strategies and academic performance.

Student feedback is widely recognized as a crucial mechanism for curriculum evaluation and improvement, serving purposes such as quality assurance, course and program revision, and the evaluation of teaching quality and assessment tasks, while also being critical to student learning by influencing motivation, engagement, self-reflective learning, and overall performance, as well as supporting iterative refinement and long-term acceptance [16-19]. Understanding the perceptions of students from different admission pathways is particularly important, as their diverse educational, social, and cultural backgrounds may shape how they interpret and adapt to new evaluation systems, potentially influencing not only their motivation and learning behaviors but also the effectiveness of these systems in supporting diverse student cohorts [20].

Comparisons of academic outcomes among various admission pathways have been partly reported; however, comparisons of preclinical students' SRL strategies, assessment engagement statistics, preferences regarding nongrading evaluation, and course learning outcomes (CLOs) and summative scores have not been studied. This study aimed to (1) compare students' SRL strategies, assessment engagement statistics, preferences regarding nongrading evaluation, and academic performance among the academic, quota, test, and rural groups; (2) determine correlations among summative scores, students' assessment engagement statistics, and CLO scores in each group; and (3) identify factors contributing to summative scores in each group through multiple linear regression analyses. By addressing these gaps, our study provides a comprehensive understanding of how admission pathways may shape not only academic outcomes but also learning behaviors, essential for designing inclusive medical curricula that support all learners [7].

Methods

Study Design

This study used a mixed methods design consisting of a retrospective observational component and a cross-sectional survey component. The retrospective observational component included assessment engagement statistics and CLO scores obtained from the Siriraj E-Learning and Education Community, as well as summative examination scores retrieved from the Siriraj Campus Management System (SiCMs). The cross-sectional component consisted of 2 system-embedded institutional questionnaires: questionnaire 1 assessing students' SRL strategies and questionnaire 2 assessing preferences regarding nongrading evaluation, both administered within the SiCMs platform. The integration of quantitative and qualitatively informed data enabled a comprehensive examination of learning behaviors and academic performance across admission pathways.

Although this study adopted a mixed methods framework, the qualitative component was limited to a qualitatively informed instrument development process based on literature review and expert input rather than primary qualitative data collection. Therefore, formal qualitative data saturation was not applicable. Instead, content validity was ensured through iterative expert review and consensus to confirm comprehensive coverage of relevant educational constructs.

Study Protocol

Participants included all first-year medical students enrolled in the academic year 2021 at the Faculty of Medicine Siriraj Hospital, Mahidol University, using a census sampling approach. Selection criteria included all officially enrolled students, with no exclusions applied.

Ethical Considerations

This study used routinely collected educational data and system-embedded questionnaires. The research protocol was reviewed and approved by the Siriraj Institutional Review Board. The initial protocol was granted exemption from full review under protocol number 370/2565 (exempt). A subsequent amendment, which included additional demographic data (age, school region, and hometown region), was approved under certificate of approval number 731/2025. For the retrospective observational component, the institutional review board granted a waiver of written informed consent because the data were obtained from existing institutional databases, posed minimal risk, and were analyzed in deidentified, aggregate form. For the cross-sectional component, participation involved completion of system-embedded questionnaires as part of routine institutional processes; therefore, no additional consent was required for secondary analysis of anonymized data, as approved by the institutional review board. All datasets were anonymized prior to export, with direct identifiers removed, and were stored on password-protected computers with restricted access. No financial or academic incentives were provided, and participation had no impact on academic standing. No identifiable images or sensitive personal information was included in the analysis.

Admission to the Faculty of Medicine Siriraj Hospital

For the academic group, applicants had to be high school students who had achieved gold or silver medals at the National Academic Olympiad in the following fields: Physics, Chemistry, Biology, Mathematics, Computer Science (Informatics), Geography, Astronomy, and Astrophysics; or have completed Camp 2 in Earth and Space Science. In addition, applicants could be high school students in the year of the training camp or those selected to attend the Academic Olympiad Training Camp Round 1 in Biology, Chemistry, Physics, Computer Science, or Mathematics. Furthermore, they were required to have a GPA of at least 3.00 in Science, Mathematics, English, Thai, and Social Studies. Selection was primarily based on portfolio assessment, including activities, academic achievements, and other special skills or talents, as well as a statement of purpose. Subsequently, they underwent multiple mini-interviews (MMIs), which were constructed based on the conceptual framework of the objective structured clinical examination [21], and final selection was determined according to MMIs criteria.

For the quota group, applicants must (1) have talent in sports or music at national or international levels; (2) have eligibility

XSL•FO
RenderX

for the Mahidol Medical Scholars Program, with a GPA of at least 3.00 in Science, Mathematics, English, Thai, and Social Studies; or (3) be holding a bachelor's degree with a GPA of at least 3.25 in any field in Thailand, with age not exceeding 35 years. Applicants were primarily selected based on academic examinations organized by the TCAS, accounting for 70%, including 20% Mathematics, 40% Sciences (Physics, Chemistry, and Biology), 20% English, 10% Thai, and 10% Social Studies, as well as 30% Ordinary National Educational Test for music talent, or General Aptitude Test/Professional Aptitude Test for sport talents and Mahidol Medical Scholar Program, or the TPAT organized by the Consortium of Thai Medical Schools for bachelor's graduates. Finally, they underwent MMIs for definitive selection.

For the test group, applicants were high school students selected through academic examinations weighted at 70% TCAS and 30% TPAT, as described for the quota group, along with a general interview, including both a standard personal interview and a psychiatric assessment, as well as a health examination, to screen out individuals unsuitable for medical studies. The eligibility criteria for passing these assessments were based on the national medical school admission standards, which ensure that applicants are free from any physical or mental health conditions that could hinder their medical education, clinical training, or future medical practice. These criteria are defined in the official 2016 regulation on student eligibility and include, for example, severe psychiatric disorders, active communicable diseases, significant physical disabilities, and uncorrectable hearing or visual impairments.

For the rural group, high school students were admitted under Thailand's national "Doctors for Rural Areas" policy, implemented through the Collaborative Project to Increase Production of Rural Doctors, which was established to address physician shortages in underserved areas and to reflect Thailand's national "Doctors for Rural Areas" policy, aiming to promote equity in health care distribution [22]. Students from designated provinces must have their names in the house records of the specified province continuously for not less than 5 years, attend a school in the designated area, and receive government financial support throughout medical training. Upon graduation, graduates are obligated to return to their home province or designated health region to work in public hospitals as part of a binding service requirement [22]. They are required to serve 3 years in Ministry of Public Health hospitals, with a financial penalty of approximately US $11,300 imposed for noncompliance [22]. They were selected based on a GPA of at least 3.00, a 70% academic examination score through TCAS as the test group, and 30% interview performance. Although they were officially enrolled under the Faculty of Medicine, Praboromarajchanok Institute, they studied their preclinical years alongside the Faculty of Medicine Siriraj Hospital's students, receiving the same curriculum delivery, teaching and assessment processes, access to facilities and equipment, academic and mentoring support, and extracurricular opportunities. After finishing the preclinical curriculum, they continued their clinical training at the Clinical Education Center, Ratchaburi Hospital.

Curriculum

The Doctor of Medicine program at the Faculty of Medicine Siriraj Hospital, Mahidol University, Thailand, spans 6 years, divided into the preclinical years (years 1 - 3) and the clinical years (years 4 - 6). The latest curriculum, implemented in 2021, introduced several changes, including defining specific CLOs, transferring 4 basic biomedical science courses from year 2 of the 2017 curriculum to year 1 of the 2021 curriculum, changing from compensatory to noncompensatory evaluation, and shifting from grading to nongrading evaluation as Outstanding ("O"), Satisfactory ("S"), or Unsatisfactory ("U").

Assessment and Evaluation

For each CLO performance, students were permitted unlimited attempts, except for course 4, which allowed only 1 attempt with an 80% passing threshold. For summative assessment, examinations for 4 courses were conducted simultaneously on the same day with a minimum pass mark of 60% for courses 1 - 3 and 50% for course 4. For evaluation, students who failed the CLO assessment or the initial attempt at a summative assessment received an "X" grade, indicating that they had not yet passed. If these students subsequently passed the CLO within the specified time frame and achieved at least a 60% score on their second attempt at the summative assessment, they were awarded an "S" grade, but they forfeited eligibility for an "O" grade. To qualify for an "O" grade, students had to attain summative scores of at least 85% in courses 1, 2, and 4, and 80% in course 3, without receiving any "X" grades in any CLO, attitudinal assessments, or summative assessments. A "U" grade was given to students who failed the remediation of the CLO examination, the second attempt at the summative examination, or the assessment of their attitude.

Assessment Engagement Statistics

Students' assessment engagement statistics comprised 319 students: 23 in the academic group, 6 in the quota group, 261 in the test group, and 29 in the rural group. These statistics encompassed various aspects, including the number of total, intentional, and unintentional attempts; instances of first-pass attempt and highest scoring attempt; the number of passings for each CLO; and (an) additional attempts after passing each CLO. Across the 4 courses, course 1 included 5 CLOs, course 2 included 4 CLOs, course 3 included 5 CLOs, and course 4 included 5 CLOs.

For each CLO, the number of total attempts is the total number of times the students accessed the assessment platform. The number of intentional attempts is defined as attempts in which ≥50% of the items were completed. The number of unintentional attempts is defined as attempts in which <50% of the items were completed. The instances of the first-pass attempt are the first instances when the students successfully meet the passing level. The instances of the highest-scoring attempt are the instances when the students achieve the maximum score. The number of passings for each CLO is the count of instances in which the students successfully meet the passing level. The additional attempts after passing each CLO are the number of subsequent instances after students have already passed.

XSL•FO
RenderX

Questionnaires

The questionnaires were developed collaboratively by the authors under the supervision of the deputy dean for undergraduate education at the Faculty of Medicine Siriraj Hospital. The questionnaires regarding students' SRL strategies and preferences regarding nongrading evaluation among the academic, quota, test, and rural groups were administered to first-year medical students during the academic year 2021, following the summative examination.

Questionnaire 1, rated on a Likert scale from 1 (strongly disagree) to 5 (strongly agree), assessed students' perspectives and experiences across four domains: (1) teaching and learning, (2) CLOs and program learning outcomes, (3) nongrading evaluation, and (4) academic obstacles, included to identify barriers perceived by students in adapting to the new structure, learning activities, and nongrading evaluation system.

Questionnaire 1 was initially based on a university-standard template and was subsequently refined using items adapted from 2 well-established instruments: first, the University Students Learning Strategies Questionnaire, which contributed items addressing goal-setting and emotional barriers to learning [23], for example, "I try to set goals before studying," "I feel discouraged from studying," and "I have concerns that interfere with my learning." Second, the Academic Motivation Scale, college version, which informed items related to intrinsic motivation (eg, "I enjoy and feel satisfied when learning new things" and "I feel happy when I achieve each course learning outcome") and amotivation (eg, "I feel like I waste time participating in class activities") [24].

Questionnaire 1 represents 3 phases of SRL strategies. The Forethought Phase (P1), representing motivation and beliefs, included items such as goal setting and enthusiasm toward learning and teaching activities. The Performance Phase (P2), reflecting monitoring and strategy use, encompassed concerns or obstacles that hindered studying such as the perceived impact of unlimited CLO assessment attempts. The Self-Reflection Phase (P3), representing assessment and self-reaction, included items related to satisfaction with nongrading evaluation and overall learning experiences gained from the current curriculum. It also explored students' perceptions of the nongrading "O"/"S"/"U" evaluation, the applicability of knowledge to future clinical and national assessments, and overall satisfaction with the curriculum. Both positively and negatively worded statements were included to support internal consistency testing.

The overall questionnaire 1 response rate was 99.1% (316/319), including 100% (23/23) in the academic group, 100% (6/6) in the quota group, 98.9% (258/261) in the test group, and 100% (29/29) in the rural group. The proportions of respondents were 7.3% (23/316) in the academic group, 1.9% (6/316) in the quota group, 81.6% (258/316) in the test group, and 9.2% (29/316) in the rural group.

Questionnaire 2, which allowed multiple responses, was developed to explore students' preferences regarding the newly implemented "O"/"S"/"U" grading system, with a particular focus on their perspectives toward the inclusion of the "O" grade, in relation to aspects of SRL strategies (Forethought, Performance, and Self-Reflection phases). While the "O" grade may serve as a form of academic recognition that rewards effort, motivation, and high performance, several student concerns also emerged, particularly regarding its association with extrinsic motivation, heightened anxiety, and increased competitiveness. Although no literature has directly addressed the specific effects of the "O" grade itself, studies on A-F grading suggest that top grades may similarly function as motivators and thus can be reasonably used to approximate the motivational impact of the "O" grade in this context [25]. Based on these recurring themes, the research team identified measurable items that addressed key constructs frequently cited in the literature on grading systems, including stress [9], competition [10,26], study focus [26], effort, and life-goal setting.

Although this questionnaire did not undergo a full validation process and did not adopt an existing standardized instrument, its content was informed by student feedback gathered during monthly online visits by the deputy dean and assistant dean (the corresponding author, CS) of undergraduate education with the first-year cohort, internal faculty discussions, and relevant literature on the psychological and behavioral impacts of grading practices. A qualitatively informed survey design approach was used to generate a focused list of 5 questions designed to capture concrete student perspectives, thereby enabling clearer interpretation of their positions on key issues as part of the validity argument. A simplified yes/no response format was deliberately selected to enhance clarity, improve response process validity, and encourage full participation by prompting students to take a definitive stance on each item.

The overall response rate was 93.7% (299/319), with participation rates of 95.7% (22/23) in the academic group, 100% (6/6) in the quota group, 93.1% (243/261) in the test group, and 96.6% (28/29) in the rural group. The proportions of respondents were 7.4% (22/299) in the academic group, 2.0% (6/299) in the quota group, 81.3% (243/299) in the test group, and 9.4% (28/299) in the rural group.

To ensure content validity, both questionnaires were reviewed by 3 medical education experts. Their feedback was incorporated to improve item clarity, coverage, and logical flow. The finalized questionnaire was administered at the end of the academic year but before grade announcement to minimize potential bias.

Construct validity for questionnaire 1 was examined using exploratory factor analysis (principal component extraction with Varimax rotation). Sampling adequacy was acceptable (Kaiser-Meyer-Olkin Measure of Sampling Adequacy=0.812), and Bartlett's test of sphericity was significant (χ^2_{136}=2253.45; P<.001). Four factors were extracted, explaining 62.9% of the total variance after rotation. Factor loadings supported a theoretically consistent structure aligned with SRL domains: (1) teaching and learning, (2) CLOs and program learning outcomes, (3) nongrading evaluation, and (4) academic obstacles. Most communalities were ≥0.50. Internal consistency was acceptable (Cronbach α=0.729). Content validity was ensured through expert review and adaptation from validated instruments.

XSL•FO
RenderX

For questionnaire 2, the first item assessed grading system preference and was analyzed descriptively because it represents a categorical choice rather than a latent construct. The remaining 5 dichotomous items assessed perceived consequences of the "O" grade (stress, competition, focus, effort, and goals). When treated as a single composite, internal consistency was low (Cronbach α=0.349), indicating that these items did not form a unidimensional scale. Interitem correlations suggested 2 distinct clusters—perceived costs and perceived benefits—with negative correlations between clusters. Given this multidimensional structure and low reliability, items were analyzed individually. Construct validity was not established for questionnaire 2; however, content validity was supported through expert review and literature-informed item development.

Academic Performance

Academic performance was represented by CLO scores and summative scores, both comprising a total of 319 students: 23 in the academic group, 6 in the quota group, 261 in the test group, and 29 in the rural group.

Missing Data Handling

Missing data were assessed using the SPSS (version 30; IBM Corp) Missing Value Analysis module. The dataset for assessment engagement statistics and CLO scores included 319 students, and summative scores were available for all 319 students. No missing data were observed for questionnaire 1 (N=316) and questionnaire 2 (N=299), as well as for summative scores. For assessment engagement statistics and CLO scores, a small proportion of missing data occurred exclusively in the test group due to some students not attempting specific CLO assessments or data unavailability. Specifically, missing data were observed for CLO 1 (3 students), CLO 2 (5 students), CLO 3 (3 students), CLO 4 (6 students), and CLO 5 (4 students) in course 1; for CLO 1 (3 students), CLO 2 (3 students), CLO 3 (6 students), and CLO 4 (7 students) in course 2; and for CLOs 1 - 5 (3 students) in course 3. Across all CLO variables, the proportion of missing data ranged from 0.94% to 2.19%, which is below the commonly accepted threshold of 5%.

Little's Missing Completely at Random (MCAR) test was performed on the combined dataset of assessment engagement statistics and CLO scores using the SPSS Missing Value Analysis module. The test was nonsignificant (χ^2_{464}=0.000; P=1.000), indicating that the observed missing data pattern was consistent with MCAR. Because the proportion of missing data was very small and the MCAR assumption was satisfied, complete case analysis was applied. Multiple imputation was not performed.

Reporting Guidelines

To enhance transparency and clarity of reporting, this study adhered to EQUATOR-aligned guidelines appropriate to its methodological structure. The STROBE (Strengthening the Reporting of Observational Studies in Epidemiology) guideline was applied to the retrospective observational components [27], and the GRAMMS (Good Reporting of A Mixed Methods Study) guideline was applied to the mixed methods framework [28]. Completed STROBE and GRAMMS checklists have been uploaded as supplementary files.

Statistics

For statistical analysis, SPSS (version 18) was used. Participant demographics and students' preferences regarding nongrading evaluation are presented as frequencies (N) and percentages. Students' assessment engagement statistics are reported as ranges of means and medians. Comparisons of SRL, assessment engagement statistics, and CLO and summative scores among students in the academic, quota, test, and rural groups were analyzed using one-way ANOVA and are expressed as means with 95% CIs. Post hoc analyses using Fisher least significant difference test were performed, as appropriate. The correlations of summative scores with students' assessment engagement statistics and CLO scores within the academic, quota, test, and rural groups were analyzed using Pearson correlation coefficient (R) with 95% CIs. Multiple linear regression analyses were conducted to identify factors significantly associated with summative scores in these groups, with regression coefficients reported alongside their 95% CIs. Statistical significance was set at P<.05.

Results

Participant Demographics by Admission Pathway

The demographic characteristics of participants across the 4 TCAS admission groups—academic, quota, test, and rural—demonstrated distinct patterns in age, sex, school region, and hometown region, as shown in Table 1. Most students were aged between 18 and 20 years, except in the quota group, where one-third were aged 21 - 25 years and another one-third 26 - 30 years. Accordingly, the mean age was highest in the quota group (23.7 years), compared with 19.4 - 19.5 years in the other groups. Male students predominated in the academic (15/23, 65.2%), test (134/261, 51.3%), and rural (17/29, 58.6%) groups, whereas the quota group was primarily female (5/6, 83.3%). By school region, Bangkok was most common in the test group (147/261, 56.3%), the central region predominated in the academic group (15/23, 65.2%), and the western region was most frequent in the rural group (22/29, 75.9%). The quota group showed a more even distribution, with 50.0% (3/6) from the central region and 16.7% (1/6) each from the west, south, and Bangkok. By hometown region, Bangkok accounted for the largest share in the test group (122/261, 46.7%); the academic group was distributed across Bangkok (9/23, 39.1%), central (5/23, 21.7%), and northeast (5/23, 21.7%); and the rural group primarily originated from the western region (23/29, 79.3%). The quota group consisted of one-third each from Bangkok and the central region, with smaller proportions from the west and south (1/6, 16.7% each).

XSL•FO
RenderX

Table . Participant demographics by admission pathway.

Factors	Academic group (N=23), n (%)	Quota group (N=6), n (%)	Test group (N=261), n (%)	Rural group (N=29), n (%)
Age (years)				
18 - 20	23 (100.0)	2 (33.4)	246 (94.2)	29 (100.0)
21 - 25	N/A[a]	2 (33.3)	15 (5.8)	N/A
26 - 30	N/A	2 (33.4)	N/A	N/A
Mean	19.4	23.7	19.5	19.5
Median	19.4	24.0	19.4	19.5
Range	18.7 - 20.0	19.4 - 29.7	18.0 - 22.7	18.8 - 20.0
Sex				
Male	15 (65.2)	1 (16.7)	134 (51.3)	17 (58.6)
Female	8 (34.8)	5 (83.3)	127 (48.7)	12 (41.4)
School region				
Bangkok	3 (13.0)	1 (16.7)	147 (56.3)	N/A
Central	15 (65.2)	3 (50.0)	90 (34.5)	7 (24.1)
North	1 (4.4)	N/A	2 (0.8)	N/A
Northeast	3 (13.0)	N/A	9 (3.5)	N/A
East	1 (4.4)	N/A	3 (1.1)	N/A
West	N/A	1 (16.7)	N/A	22 (75.9)
South	N/A	1 (16.7)	10 (3.8)	N/A
Hometown region				
Bangkok	9 (39.1)	2 (33.3)	122 (46.7)	N/A
Central	5 (21.7)	2 (33.3)	59 (22.6)	6 (20.7)
North	1 (4.4)	N/A	3 (1.2)	N/A
Northeast	5 (21.7)	N/A	24 (9.2)	N/A
East	2 (8.7)	N/A	11 (4.2)	N/A
West	N/A	1 (16.7)	8 (3.1)	23 (79.3)
South	1 (4.4)	1 (16.7)	34 (13.0)	N/A

[a]N/A: not applicable.

Comparisons of SRL Strategies Among Academic, Quota, Test, and Rural Groups

Comparisons of SRL strategies among academic, quota, test, and rural groups are shown in Table 2. In the aspect of teaching and learning, students in the academic group reported significantly higher levels of actively setting goals before studying than those in the quota group (P=.044; Table 2). The 95% CI for the academic group was narrow (4.07 - 4.63), indicating high precision. In contrast, the quota group showed a much wider 95% CI (2.58 - 4.75), reflecting lower precision due to both the small sample size and the high variability in students' responses within this group (Table 2). Students in the academic group also reported significantly higher levels of feeling enthusiastic about learning and teaching activities than students in the rural group (P=.042; Table 2). The 95% CI for the academic group (4.18 - 4.69) was narrow, suggesting stable estimates, whereas the rural group's moderately wide 95% CI (3.65 - 4.28) indicates greater variability in enthusiasm levels among students in this pathway (Table 2).

XSL•FO
RenderX

Table . Comparisons of self-regulated learning strategies among academic, quota, test, and rural groups.

Questions		Academic group (N=23) Mean (95% CI)	Quota group (N=6) Mean (95% CI)	Test group (N=258) Mean (95% CI)	Rural group (N=29) Mean (95% CI)
Teaching and learning					
1	I actively set goals for myself before studying. (P1[a])	4.35 (4.07 - 4.63)	3.67 (2.58-4.75)[b]	4.22 (4.13 - 4.31)	4.10 (3.79 - 4.41)
2	I am enthusiastic about learning and teaching activities. (P1[a])	4.43 (4.18 - 4.69)	3.83 (2.80 - 4.87)	4.12 (4.02 - 4.22)	3.97 (3.65–4.28)[b]
3	I am happy and satisfied when I learn something new. (P[3c])	4.39 (4.08 - 4.70)	4.50 (3.93 - 5.07)	4.41 (4.32 - 4.50)	4.34 (4.11 - 4.58)
4	The teaching and learning management of this curriculum helps me experience learning in an excellent manner. (P3[c])	4.39 (4.05 - 4.73)	4.67 (4.12 - 5.21)	4.38 (4.29 - 4.46)	4.31 (4.10 - 4.52)
5	Overall, I am satisfied with the learning experience gained from the current curriculum. (P3[c])	4.35 (4.14 - 4.56)	4.67 (4.12 - 5.21)	4.35 (4.27 - 4.43)	4.38 (4.12 - 4.64)
CLOs[d] and PLOs[e]					
6	I feel happy when I achieve the CLOs each time. (P3[c])	4.83 (4.66 - 4.99)	5.00 (5.00 - 5.00)	4.62 (4.55 - 4.70)	4.55 (4.36 - 4.74)
7	I like the opportunity to have unlimited attempts at achieving the CLOs. (P2[f])	4.74 (4.54 - 4.93)	4.83 (4.40 - 5.26)	4.85 (4.80 - 4.90)	4.76 (4.56 - 4.95)
8	I think that the knowledge and experience gained from studying in the first year can be applied in future medical studies. (P3[c])	4.61 (4.36 - 4.86)	5.00 (5.00 - 5.00)	4.67 (4.61 - 4.74)	4.72 (4.55 - 4.90)
9	I think that the knowledge and experience gained from studying will be applicable to the preclinical comprehensive examination and the National License Examination step 1. (P3[c])	4.74 (4.54 - 4.93)	4.83 (4.40 - 5.26)	4.64 (4.57 - 4.71)	4.62 (4.41 - 4.83)
10	The current examination method (once/semester) allows me to study happily. (P3[c])	3.17 (2.64 - 3.71)	3.33 (2.48 - 4.19)	3.40 (3.24 - 3.56)	3.24 (2.81 - 3.67)
Nongrading evaluation					

XSL•FO
RenderX

Questions		Academic group (N=23) Mean (95% CI)	Quota group (N=6) Mean (95% CI)	Test group (N=258) Mean (95% CI)	Rural group (N=29) Mean (95% CI)
11	I am happy to learn through the nongrading evaluation (“O”/”S”/“U”). (P3[c])	4.35 (4.07 - 4.63)	5.00 (5.00–5.00)[b]	4.77 (4.70 - 4.84)[g]	4.86 (4.73 - 5.00)[g]
12	I agree with changing from the grading to the nongrading evaluation (“O”/”S”/“U”) for all preclinical courses. (P3[c])	4.35 (4.07 - 4.63)	5.00 (5.00–5.00)[b]	4.79 (4.72 - 4.86)[h]	4.90 (4.78 - 5.01)[h]
Academic obstacles					
13	I feel like I’m wasting my time participating in the teaching and learning activities. (P2[f])	1.83 (1.47 - 2.19)	1.67 (0.81 - 2.52)	2.07 (1.91 - 2.22)	2.00 (1.55 - 2.45)
14	I have concerns or problems that prevent me from studying. (P2[f])	2.13 (1.60 - 2.66)	1.83 (1.04 - 2.62)	2.60 (2.45 - 2.76)	2.76 (2.38 - 3.13)
15	I don’t feel like studying. (P1[a])	1.87 (1.43 - 2.31)	2.00 (1.06 - 2.94)	2.37 (2.22 - 2.52)	2.38 (1.98 - 2.78)
16	I have stress when reading books or following lessons during studies. (P2[f])	2.64 (2.15 - 3.12)	3.00 (1.67 - 4.33)	3.27 (3.14 - 3.40)[g]	3.52 (3.20 - 3.83)[g]
17	I have stress when reading books before exams. (P2[f])	3.70 (3.11 - 4.29)[i]	3.50 (1.78 - 5.22)	3.97 (3.85 - 4.10)[i]	4.38 (4.10 - 4.66)

[a]P1 represents Motivation and Beliefs (motivation and beliefs (Forethought Phase).

[b]*P*<.05 compared with the academic group.

[c]P3 represents assessment and self-reaction (Self-Reflection Phase).

[d]CLOs: course learning outcomes.

[e]PLOs: program learning outcomes.

[f]P2 represents Monitoring and Strategy Use (Performance Phase).

[g]*P*<.01 compared with the academic group.

[h]*P*<.001 compared with the academic group.

[i]*P*<.05 compared with the rural group.

In the aspect of nongrading evaluation, students in the academic group reported significantly lower levels of agreement with learning via nongrading evaluation (“O”/“S”/“U”) than the quota group (*P*=.012), test group (*P*=.001), and rural group (*P*=.001) (Table 2). The 95% CI for the academic group (4.07 - 4.63) was relatively narrow, indicating high precision and consistent responses, whereas the quota group showed a ceiling effect (5.00 - 5.00) suggesting no variability, and the 95% CIs for the test (4.70 - 4.84) and rural groups (4.73 - 5.00) were very narrow, reflecting precise estimates (Table 2). Similarly, students in the academic group reported significantly lower agreement with changing from grading to nongrading evaluation (“O”/“S”/“U”) for all preclinical courses than the quota group (*P*=.011), test group (*P*<.001), and rural group (*P*<.001; Table 2). The 95% CI for the academic group (4.07 - 4.63) was narrow, suggesting stable estimates, whereas the quota group again showed a ceiling effect (5.00 - 5.00), and the test (4.72 - 4.86) and rural groups (4.78 - 5.01) had narrow 95% CIs, indicating high precision (Table 2).

In the aspect of academic obstacles, students in the academic group rated significantly lower levels of stress when reading books or following lessons during studies than students in the test (*P*=.007) and rural groups (*P*=.003; Table 2). The 95% CI for the academic group (2.15 - 3.12) was moderately wide, indicating greater variability in perceived stress, whereas the test (3.14 - 3.40) and rural groups (3.20 - 3.83) showed

XSL•FO
RenderX

narrower 95% CIs reflecting more consistent responses, and the quota group demonstrated a wide 95% CI (1.67 - 4.33), indicating low precision due to both high variability and a small sample size (Table 2).

Furthermore, students in the rural group rated significantly higher levels of stress when reading books before examinations than students in the academic (P=.02) and test groups (P=.046; Table 2). The 95% CI for the rural group (4.10 - 4.66) was relatively narrow, indicating high precision, while the 95% CI for the academic group (3.11 - 4.29) was moderately wide, suggesting moderate precision, and the 95% CI for the test group (3.85 - 4.10) was narrow, reflecting precise estimates; the quota group had an extremely wide 95% CI (1.78 - 5.22), indicating high variability and small sample size (Table 2).

Students' Preferences Regarding Nongrading Evaluation

Students' preferences regarding nongrading evaluation and the "O" grade in academic, quota, test, and rural groups are shown in Table 3. Students in the academic, quota, and test groups predominately selected "O"/"S"/"U" for 63.6% (14/22), 83.3% (5/6), and 54.3% (132/243), respectively, while students in the rural group mainly selected "S"/"U" at 46.4% (13/28; Table 3). Furthermore, students in the academic and quota groups reported that "O" mainly made them more focused on studying and led to increased effort with equal percentage (15/22, 68.2% and 3/6, 50%, respectively; Table 3). In addition, students in the test and rural groups reported that "O" mostly led to increased effort (161/243, 66.3% and 18/28, 64.3%, respectively; Table 3).

Table . Students' preferences regarding nongrading evaluation.

	Academic group (N[a]=22), n (%)	Quota group (N=6), n (%)	Test group (N=243), n (%)	Rural group (N=28), n (%)
Choosing the evaluation				
"S"[a]/"U"[b]	7 (31.8)	1 (16.7)	90 (37.0)	13 (46.4)
"O"[c]/"S"/"U"	14 (63.6)	5 (83.3)	132 (54.3)	12 (42.9)
Both "S"/"U" and "O"/"S"/"U"	1 (4.6)	N/A[d]	21 (8.7)	3 (10.7)
Topics				
"O" makes students feel stressed. (P3[e])	10 (45.5)	N/A	83 (34.2)	11 (39.3)
"O" leads to increased competition. (P2[f])	13 (59.1)	1 (16.7)	98 (40.3)	11 (39.3)
"O" makes students more focused on studying. (P2)	15 (68.2)	3 (50.0)	116 (47.7)	12 (42.9)
"O" leads to increased effort. (P2)	15 (68.2)	3 (50.0)	161 (66.3)	18 (64.3)
"O" leads to having more goals in life. (P1[g])	13 (59.5)	1 (16.7)	104 (42.8)	13 (46.4)

[a]"S": satisfactory.

[b]"U": unsatisfactory.

[c]"O": outstanding.

[d]N/A: not applicable.

[e]P3 represents assessment and self-reaction (Self-Reflection Phase).

[f]P2 represents monitoring and strategy use (Performance Phase).

[g]P1 represents motivation and beliefs (Forethought Phase).

XSL•FO
RenderX

Students' Assessment Engagement Statistics in Performing Each CLO

Students' assessment engagement statistics in performing each CLO of students in academic, quota, test, and rural groups are shown as ranges of means and medians in Table 4. These assessment engagement statistics include multiple aspects of assessment engagement, namely, total attempts, intentional and unintentional attempts, first-pass attempts, highest scoring attempts, passing instances, and additional attempts after passing for each CLO.

Table . Students' assessment engagement statistics in performing each course learning outcome.

Factors	Academic group (N=23)		Quota group (N=6)		Test group (N=258)		Rural group (N=29)	
	Mean (range)	Median (range)	Mean (range)	Median (range)	Mean (range)	Median (range)	Mean (range)	Median (range)
Number of total attempts of each CLO[a].	2.0 - 9.1	1 - 7	3.5 - 29.0	2.5 - 19	2.6 - 15.3	2 - 8	2.4 - 12.3	2 - 10
Number of intentional attempts of each CLO.	1.9 - 6.3	1 - 5	2.0 - 6.3	2 - 5.5	2.0 - 7.5	2 - 6	2.3 - 9.8	2 - 9
Number of unintentional attempts of each CLO.	0 - 5.1	0	0 - 22.8	0 - 8.5	0 - 10.9	0	0 - 6.5	0
Instances of first-pass attempt of each CLO.	1.4 - 3.6	1 - 3	1.3 - 4.0	1 - 3	1.4 - 4.5	1 - 3	1.9 - 6.5	1 - 5
Instances of highest scoring attempt of each CLO.	1.7 - 5.7	1 - 4	1.7 - 6.0	1.5 - 5	1.9 - 6.7	2 - 5	2.2 - 9.5	2 - 9
Number of passings for each CLO.	1.5 - 3.0	1 - 2	1.5 - 4.0	1 - 3.5	1.4 - 3.3	1 - 2	1.4 - 2.9	1 - 2
Additional attempts after passing each CLO.	0.5 - 2.7	0 - 2	0.5 - 4.0	0 - 3	0.6 - 3.0	0 - 1	0.4 - 3.4	0 - 1

[a]CLO: course learning outcome.

Comparisons of Assessment Engagement Statistics Among Students in Academic, Quota, Test, and Rural Groups

Comparisons of assessment engagement statistics among students in academic, quota, test, and rural groups are shown in Figure 1. For each CLO, the number of total attempts was significantly higher in the test group compared with the academic group in course 1 CLO 1 (P=.044) and course 2 CLO 2 (P=.049) and compared with the rural group in course 1 CLO 3 (P=.03), as well as significantly higher in the quota group compared with the academic (P=.02) and rural groups (P=.041) in course 3 CLO 2 (Figure 1A). The number of intentional attempts was significantly lower in the academic group in course 3 CLO 1 compared with the rural group (P=.027; Figure 1B). The number of unintentional attempts was significantly higher in the quota group compared with the rural group in course 1 CLO 3 (P=.045), course 3 CLO 2 (P=.043), and course 4 (P=.048), and compared with the academic group in course 3 CLO 2 (P=.03), as well as significantly higher in the test group compared with the rural group in course 1 CLO 3 (P=.042; Figure 1C). The instances of first-pass attempt were significantly lower in the academic group compared with the rural group in course 1 CLO 1 (P=.04), course 2 CLO 1 (P=.008) and 2 (P=.004), course 3 CLO 1 (P=.03), CLO 4 (P=.044), and CLO 5 (P=.03), and compared with the quota group in course 3 CLO 2 (P=.02), as well as significantly lower in the test group compared with the rural group in course 2 CLO 1 (P=.02) and CLO 2 (P=.03) and course 3 CLO 1 (P=.04) and 4 (P=.005) (Figure 1D). The instances of highest scoring attempts were significantly higher in the rural group compared with the academic group in course 2 CLO 2 (P=.02) and compared with the academic group (P=.01), quota group (P=.03), and test group (P=.02) in course 3 CLO 1 (Figure 1E). The number of passings in each CLO was significantly lower in the rural group compared with the test group in course 1 CLO 1 (P=.03), as well as significantly higher in the quota group compared with the academic group (P=.004), test group (P<.001), and rural group

XSL•FO
RenderX

(P<.001) in course 2 CLO 4 (Figure 1F). The additional attempts after passing each CLO were significantly higher in the quota group compared with the academic group (P=.01), test group (P=.004), and rural group (P=.001) in course 2 CLO 4 (Figure 1G).

Across total attempts (Figure 1A) and unintentional attempts (Figure 1C), the test group consistently demonstrated narrow 95% CIs indicating high precision, the academic and rural groups showed 95% CIs of narrow to moderate width indicating high to moderate precision, and the quota group exhibited extremely wide 95% CIs reflecting low precision due to substantial variability and a small sample size.

Across intentional attempts (Figure 1B), first-pass attempts (Figure 1D), highest scoring attempts (Figure 1E), number of passings (Figure 1F), and additional attempts after passing (Figure 1G), a similar precision pattern was observed, in which the test group consistently showed narrow 95% CIs indicating high precision, the academic and rural groups demonstrated moderate width 95% CIs reflecting moderate precision, and the quota group exhibited the widest 95% CIs indicating low precision due to high variability and its small sample size.

XSL•FO
RenderX

Figure 1. Comparisons of assessment engagement statistics among students in academic, quota, test, and rural groups. Data are presented as mean with 95% CIs. *P<.05 compared with students in the academic group; $^{\#}P$<.05, $^{\#\#}P$<.001 compared with students in the quota group; and $^{\$}P$<.05 compared with students in the test group. (A) Number of total attempts of each CLO. (B) Number of intentional attempts of each CLO. (C) Number of unintentional attempts of each CLO. (D) The instance(s) of first-pass attempt of each CLO. (E) The instance(s) of highest scoring attempt of each CLO. (F) Number of passing(s) each CLO. (G) Additional attempt(s) after passing each CLO. CLO: course learning outcome.

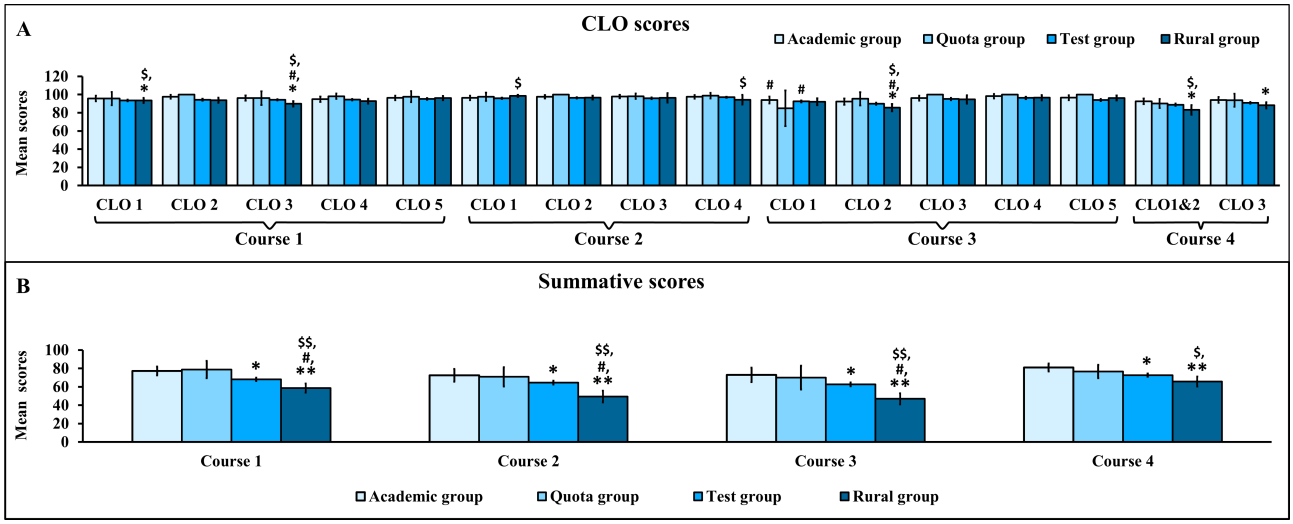
XSL•FO
RenderX

Comparisons of CLO and Summative Scores Among Students in Academic, Quota, Test, and Rural Groups

Comparisons of CLO and summative scores among students in academic, quota, test, and rural groups are shown in Figure 2.

Some CLO scores (Figure 2A) and all summative scores (Figure 2B) in courses 1 - 4 were significantly lower in the rural group when compared with the academic, quota, and test groups (P<.05 all). Across CLO scores (Figure 2A) and summative scores (Figure 2B), a similar precision pattern was observed, in which the test group consistently showed narrow 95% CIs indicating high precision, the academic and rural groups demonstrated moderate-width 95% CIs reflecting moderate precision, and the quota group exhibited the widest 95% CIs indicating low precision due to substantial variability and a small sample size.

Figure 2. Comparisons of CLO and summative scores among students in academic, quota, test, and rural groups. Data are presented as mean with 95% CIs. *P<.05, **P<.001 compared with students in the academic group; $^{\#}P$<.05 compared with students in the quota group; and $^{\$}P$<.05, $^{\$\$}P$<.001 compared with students in the test group. (A) CLO scores. (B) Summative scores. CLO: course learning outcome.



Correlations of Summative Scores With Students' Assessment Engagement Statistics and CLO Scores Within the Course

Correlations of summative scores with students' assessment engagement statistics and CLO scores within the course of students in academic, quota, test, and rural groups are shown in Multimedia Appendix 1.

In the academic group, summative scores of each course were positively correlated with number of passings each CLO of courses 1 and 3 (R=0.415 - 0.580), additional attempts after passing each CLO of course 1 (R=0.415 - 0.555), and CLO scores of all courses (R=0.432 - 0.707) but were negatively correlated with number of total attempts (R=−0.450) and number of unintentional attempts (R=−0.511) of course 1 and instances of first-pass attempt of courses 1 and 2 (R=(−0.437) – (−0.491)) (P<.05 all; Multimedia Appendix 1).

In the quota group, summative scores of each course exhibited positive correlation with number of total attempts (R=0.814) and number of intentional attempts (R=0.926) of course 2 and CLO scores of course 3 (R=0.844) but had negative correlations with number of total attempts (R=−0.964) and number of unintentional attempts (R=−0.845) of course 3 (P<.05 all; Multimedia Appendix 1).

In the test group, summative scores of each course were significantly positively correlated with number of passings each CLO (R=0.164 - 0.233) and additional attempts after passing each CLO (R=0.125 - 0.209) of courses 1, 2, and 3; CLO scores of all courses (R=0.129 - 0.253); and number of intentional attempts of course 3 (R=0.177) but were significantly negatively correlated with number of total attempts (R=(−0.128) – (−0.233)) and number of unintentional attempts (R=(−0.148) –(−0.208)) of courses 1 and 3; number of intentional attempts of course 1 (R=−0.134); instances of first-pass attempt of courses 1, 2, and 3 (R=(−0.123) – (−0.344)); and instances of highest scoring attempt of courses 1 and 2 (R=(−0.153) – (−0.160)) (P<.05 all; Multimedia Appendix 1).

In the rural group, summative scores of each course showed positive correlations with CLO scores of courses 1, 3, and 4 (R=0.420 - 0.725) and number of unintentional attempts (R=0.393) and number of passings each CLO (R=0.390) of course 3 but exhibited negative correlations with number of intentional attempts of courses 2 and 3 (R=(−0.399)–(−0.577)); instances of first-pass attempt of courses 2 and 3 (R=(−0.388) – (−0.655); and instances of highest scoring attempt of courses 2 and 3 (R=(−0.441) – (−0.614)) (P<.05 all; Multimedia Appendix 1).

Notably, the scores for course 1 CLO 2, course 2 CLO 2, and course 3 CLO 3, 4, and 5 in the quota group did not correlate with the summative scores because all students in this group received a score of 100 for these CLOs. Across correlations between summative scores and assessment engagement statistics and CLO scores, a similar precision pattern was observed, in which the test group consistently showed narrow 95% CIs indicating high precision, the academic and rural groups demonstrated moderate-width 95% CIs reflecting moderate precision, and the quota group exhibited the widest 95% CIs indicating low precision due to substantial variability and a small sample size (Multimedia Appendix 1).

XSL•FO
RenderX

Multiple Linear Regression Analyses of Summative Scores

Multiple linear regression analyses of summative scores in academic, quota, test, and rural groups are shown in Table 5.

XSL•FO
RenderX

Table . Multiple linear regression analyses of summative scores.

Model	*R*	R^2	*P* value		Coefficient	SE	*t* test (*df*)	*P* value	95% CI
Academic group									
Summative scores of course 1									
1	0.630	0.397	.001[a]						
				Constant	−40.301	31.689	−1.272 (21)	.217	−106.203 to 25.600
				CLO5[b] scores	1.221	0.329	3.717 (21)	.001[a]	0.538 to 1.905
2	0.722	0.521	<.001[c]						
				Constant	−17.197	30.682	−0.560 (20)	.581	−81.198 to 46.804
				CLO5 scores	0.912	0.330	2.769 (20)	.012[d]	0.225 to 1.600
				Number of passings (CLO1)	2.349	1.033	2.273 (20)	.034[d]	0.193 to 4.504
3	0.784	0.614	<.001[c]						
				Constant	−1.143	29.210	−0.039 (19)	.969	−62.280 to 59.994
				CLO5 scores	0.764	0.311	2.456 (19)	.024[d]	0.113 to 1.415
				Number of passings (CLO1)	2.112	0.957	2.206 (19)	.040[d]	0.108 to 4.116
				Number of unintentional attempts (CLO4)	−0.466	0.217	−2.148 (19)	.045[d]	−0.920 to −0.012
Summative scores of course 2									
1	0.491	0.241	.017[d]						
				Constant	88.704	6.934	12.793 (21)	<.001[c]	74.284 to 103.124
				Instances of first-pass attempt (CLO3)	−10.648	4.125	−2.582 (21)	.017[d]	−19.226 to −2.070
Summative scores of course 3									
1	0.707	0.500	<.001[c]						
				Constant	−119.198	42.038	−2.835 (21)	.010[d]	−206.621 to (−31.776)
				CLO5 scores	1.991	0.435	4.580 (21)	<.001[c]	1.087 to 2.894
2	0.778	0.606	<.001[c]						
				Constant	−160.196	42.113	−3.804 (20)	.001[a]	−248.043to −72.350
				CLO5 scores	2.285	0.415	5.504 (20)	<.001[c]	1.419 to 3.150
				Number of passings (CLO4)	8.538	3.678	2.322 (20)	.031[d]	0.866 to 16.209

XSL•FO
RenderX

Model	*R*	R^2	*P* value		Coefficient	SE	*t* test (*df*)	*P* value	95% CI
Summative scores of course 4									
1	0.520	0.271	.011[d]						
				Constant	16.399	23.234	0.706 (21)	.488	−31.920 to 64.717
				CLO1 and 2 scores	0.699	0.250	2.793 (21)	.011[d]	0.178 to 1.219
Quota group									
Summative scores of course 2									
1	0.926	0.857	.008[a]						
				Constant	45.455	5.454	8.335 (4)	.001[a]	30.313 to 60.596
				Number of intentional attempts (CLO1)	7.614	1.553	4.903 (4)	.008[a]	3.302 to 11.925
Summative scores of course 3									
1	0.964	0.929	.002[a]						
				Constant	86.429	2.719	31.782 (4)	<.001[c]	78.878 to 93.979
				Number of total attempts (CLO4)	−4.107	0.567	-7.243 (4)	.002[d]	−5.681 to −2.533
Test group									
Summative scores of course 1									
1	0.344	0.118	<.001[c]						
				Constant	73.796	1.251	58.999 (256)	<.001[c]	71.333 to 76.259
				Instances of first-pass attempt (CLO3)	−2.653	0.453	−5.856 (256)	<.001[c]	−3.545 to −1.761
2	0.399	0.159	<.001[c]						
				Constant	70.442	1.546	45.554 (254)	<.001[c]	67.396 to 73.487
				Instances of first-pass attempt (CLO3)	−2.649	0.444	−5.963 (254)	<.001[c]	−3.523 to −1.774
				Number of passing (CLO5)	1.765	0.498	3.545 (254)	<.001[c]	0.784 to 2.745
3	0.433	0.187	<.001[c]						
				Constant	71.500	1.565	45.689 (253)	<.001[c]	68.418 to 74.582
				Instances of first-pass attempt (CLO3)	−2.558	0.439	−5.833 (253)	<.001[c]	−3.422 to −1.695

XSL•FO
RenderX

Model	*R*	R^2	*P* value		Coefficient	SE	*t* test (*df*)	*P* value	95% CI
				Number of passing (CLO5)	1.638	0.492	3.327 (253)	.001[a]	0.668 to 2.608
				Number of unintentional attempts (CLO5)	−0.344	0.116	−2.953 (253)	.003[a]	−0.573 to −0.115
Summative scores of course 2									
1	0.260	0.068	<.001[c]						
				Constant	71.040	1.797	39.530 (253)	<.001[c]	67.501 to 74.580
				Instances of first-pass attempt (CLO3)	−3.164	0.739	−4.282 (253)	<.001[c]	−4.620 to −1.709
2	0.370	0.137	<.001[c]						
				Constant	16.729	12.185	1.373 (252)	.171	−7.269 to 40.727
				Instances of first-pass attempt (CLO3)	−3.289	0.713	-4.614 (252)	<.001[c]	−4.693 to −1.885
				CLO3 scores	0.568	0.126	4.503 (252)	<.001[c]	0.320 to 0.817
3	0.402	0.162	<.001[c]						
				Constant	21.178	12.145	1.744 (251)	.082	−2.742 to 45.098
				Instances of first-pass attempt (CLO3)	−3.222	0.704	−4.574 (251)	<.001[c]	−4.610 to −1.835
				CLO3 scores	0.472	0.130	3.643 (251)	<.001[c]	0.217 to 0.727
				Number of passing (CLO2)	2.260	0.833	2.714 (251)	.007[a]	0.620 to 3.900
4	0.420	0.176	<.001[c]						
				Constant	24.308	12.157	2.000 (250)	.047[d]	0.366 to 48.251
				Instances of first-pass attempt (CLO3)	−2.881	0.719	−4.009 (250)	<.001[c]	−4.296 to −1.465
				CLO3 scores	0.455	0.129	3.525 (250)	.001[a]	0.201 to 0.709
				Number of passing (CLO2)	3.663	1.065	3.441 (250)	.001[a]	1.566 to 5.760
				Instances of highest scoring attempts (CLO2)	−1.625	0.777	−2.093 (250)	.037[d]	−3.155 to −0.096
Summative scores of course 3									

XSL•FO
RenderX

Model	*R*	R^2	*P* value		Coefficient	SE	*t* test (*df*)	*P* value	95% CI
1	0.211	0.045	<.001[c]						
				Constant	59.682	1.385	43.099 (256)	<.001[c]	56.955 to 62.409
				Number of passing (CLO1)	1.032	0.299	3.456 (256)	.001[a]	0.444 to 1.621
2	0.310	0.096	<.001[c]						
				Constant	63.643	1.701	37.406 (255)	<.001[c]	60.292 to 66.994
				Number of passing (CLO1)	1.014	0.291	3.483 (255)	.001[a]	0.441 to 1.587
				Instances of first-pass attempt (CLO3)	−2.142	0.560	−3.822 (255)	<.001[c]	−3.246 to −1.038
3	0.354	0.125	<.001[c]						
				Constant	60.276	2.043	29.508 (254)	<.001[c]	56.253 to 64.299
				Number of passing (CLO1)	0.735	0.303	2.429 (254)	.016[d]	0.139 to 1.332
				Instances of first-pass attempt (CLO3)	−2.570	0.572	−4.493 (254)	<.001[c]	−3.697 to −1.444
				Number of passing (CLO4)	3.520	1.219	2.888 (254)	.004[a]	1.120 to 5.921
4	0.405	0.164	<.001[c]						
				Constant	61.009	2.012	30.317 (253)	<.001[c]	57.046 to 64.972
				Number of passing (CLO1)	0.679	0.297	2.288 (253)	.023[d]	0.095 to 1.264
				Instances of first-pass attempt (CLO3)	−2.201	0.571	−3.857 (253)	<.001[c]	−3.325 to −1.077
				Number of passing (CLO4)	5.381	1.312	4.102 (253)	<.001[c]	2.797 to 7.964
				Number of total attempts (CLO4)	−1.469	0.429	−3.424 (253)	.001[a]	−2.313 to −0.624
Summative scores of course 4									
1	0.357	0.127	<.001[c]						
				Constant	32.361	6.595	4.907 (259)	<.001[c]	19.374 to 45.347
				CLO1 and 2 scores	0.454	0.074	6.150 (259)	<.001[c]	0.308 to 0.599

XSL•FO
RenderX

Model	*R*	R^2	*P* value		Coefficient	SE	*t* test (*df*)	*P* value	95% CI
2	0.395	0.156	<.001[c]						
				Constant	13.440	9.095	1.478 (258)	.141	−4.470 to 31.349
				CLO1 and 2 scores	0.354	0.080	4.433 (258)	<.001[c]	0.197 to 0.512
				CLO3 scores	0.305	0.103	2.973 (258)	.003[a]	0.103 to 0.507
Rural group									
Summative scores of course 1									
1	0.482	0.232	.008[a]						
				Constant	−20.619	27.819	−0.741 (27)	.465	−77.700 to 36.462
				CLO2 scores	0.845	0.296	2.857 (27)	.008[a]	0.238 to 1.453
Summative scores of course 2									
1	0.655	0.429	<.001[c]						
				Constant	75.026	6.184	12.132 (27)	<.001[c]	62.337 to 87.715
				Instances of first-pass attempt (CLO1)	−10.182	2.261	−4.504 (27)	<.001[c]	−14.820 to −5.543
Summative scores of course 3									
1	0.441	0.194	.017[d]						
				Constant	59.697	5.668	10.532 (27)	<.001[c]	48.067 to 71.328
				Instances of highest scoring attempt (CLO4)	−5.853	2.294	-2.551 (27)	.017[d]	−10.560 to −1.145
2	0.600	0.360	.003[a]						
				Constant	57.041	5.250	10.865 (26)	<.001[c]	46.249 to 67.832
				Instances of highest scoring attempt (CLO4)	−6.060	2.086	−2.905 (26)	.007[a]	−10.347 to −1.773
				Number of passing (CLO2)	1.325	0.511	2.592 (26)	.015[d]	0.274 to 2.375
3	0.748	0.559	<.001[c]						
				Constant	54.567	4.502	12.119 (26)	<.001[c]	45.294 to 63.840
				Instances of highest scoring attempt (CLO4)	−6.331	1.767	−3.584 (25)	.001[a]	−9.969 to −2.692
				Number of passing (CLO2)	1.444	0.434	3.327 (25)	.003[a]	0.550 to 2.337

XSL•FO
RenderX

Model	*R*	R^2	*P* value		Coefficient	SE	*t* test (*df*)	*P* value	95% CI
				Number of unintentional attempts (CLO1)	1.718	0.511	3.364 (25)	.002[a]	0.666 to 2.769
Summative scores of course 4									
1	0.725	0.526	<.001[c]						
				Constant	3.731	11.465	0.325 (27)	.747	−19.793 to 27.256
				CLO1 and 2 scores	0.745	0.136	5.476 (27)	<.001[c]	0.466 to 1.025

[a]*P*<.01.

[b]CLO: course learning outcome.

[c]*P*<.001.

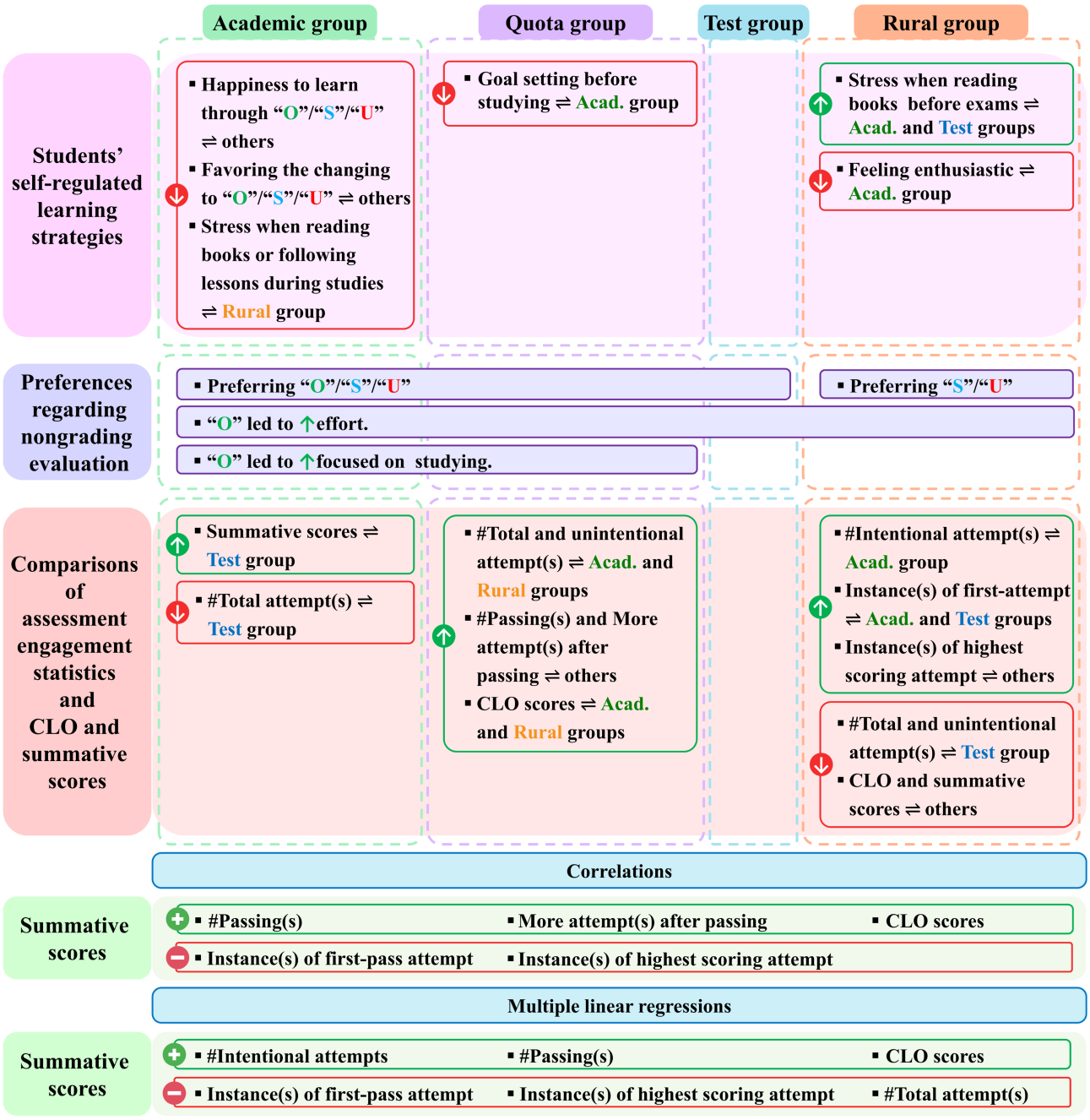
[d]*P*<.05.

In the academic group, with summative scores as the dependent variable, positive influences included CLO scores and number of passings for each CLO, while a negative influence comprised instances of first-pass attempts (Table 5). In the quota group, a positive contribution was the number of intentional attempts, while a negative contribution was the number of total attempts (Table 5). In the test group, positive factors comprised number of passings for each CLO and CLO scores, while negative factors included instances of first-pass attempts, instances of highest scoring attempts of CLO, and number of total attempts (Table 5). In the rural group, by setting summative scores as the dependent variable, positive influences included CLO scores, number of passings for each CLO, and number of unintentional attempts, while negative contributions comprised instances of first-pass attempts and instances of highest scoring attempts of CLO (Table 5).

Across all predictors and constant terms in the multiple linear regression models predicting summative scores, the test group generally showed narrow 95% CIs for predictors and moderately wide 95% CIs for constants, indicating overall moderate to high precision; the quota group exhibited moderate 95% CIs, reflecting moderate precision; and the academic and rural groups showed moderate to wide 95% CIs, especially wide for constants, indicating moderate to low precision (Table 5). A summary of results is shown in Figure 3.

XSL•FO
RenderX

Figure 3. The summary of results. Acad: academic group, CLO: course learning outcome; "O": Outstanding, others: other groups, "S": Satisfactory, "U": Unsatisfactory, : compared with; : increased; : decreased; : positive correlations; : negative correlations.



Discussion

Principal Findings

This study aimed to compare SRL strategies, assessment engagement behaviors, preferences for nongrading evaluation, and academic performance across 4 admission pathways (academic, quota, test, and rural); examine correlations among summative scores, CLO scores, and engagement metrics; and identify predictors of summative scores using multiple linear regression. In line with these objectives, the main findings indicate that students in the academic group demonstrated stronger SRL strategies, fewer but more effective assessment engagement behaviors, and higher CLO and summative scores; and the rural group exhibited lower SRL, greater academic obstacles, higher numbers of assessment attempts, and the lowest academic performance, while the quota and test groups showed intermediate patterns. Most academic, quota, and test students preferred "'O"/"S"/"U," whereas rural students favored "S"/"U." Summative scores exhibited positive correlations with CLO scores and number of passings, whereas they showed negative correlations with first-pass attempts. In the multiple linear regression analyses, summative scores were contributed positively by number of passings and CLO scores, whereas they were predicted negatively by instances of first-pass attempts and highest scoring attempts.

When comparing SRL strategies across admission pathways, students in the academic group showed stronger goal setting and greater enthusiasm for learning activities than other groups.

XSL•FO
RenderX

They also reported less stress during routine study, whereas rural students experienced more stress before examinations than their peers.

Basically, the academic group, composed of science Olympiads students who participate in science competitions, generally demonstrates a high interest in science and learning, comes from competitive backgrounds, and possesses higher levels of general cognitive abilities [29]. Furthermore, they typically have a self-concept of ability [30], self-efficacy [29], short-term and long-term goals [29], perseverance [31], and self-discipline [31], all indicative of SRL [29,32]. In addition, they usually have support from peers, family, and teachers [29,33]. Therefore, it is not surprising that they exhibited higher levels of active goal setting and enthusiasm, showed less stress, and were more likely to receive higher scores. This finding aligns with recent research showing that medical graduates from more educationally privileged backgrounds (ie, those who attended fee-paying schools) were significantly more likely to succeed in postgraduate medical examinations [34]. In contrast, the rural group, typically from less competitive backgrounds, rated lower levels of enthusiasm but higher levels of stress when learning alongside students from high academic backgrounds [35]. These students may face unique challenges, such as limited access to educational resources and support systems [35-37].

Interestingly, students in the academic group were less satisfied with learning through nongrading evaluation and less supportive of adopting it for all preclinical courses than other groups. Most students in the academic, quota, and test pathways favored the inclusion of the "O" grade because they felt that it encouraged focus and effort, whereas rural students showed a slight preference for the simpler "S"/"U" system. These results reflect that the academic group might have familiarity with competitive academic environments [29,32] and a preference for evaluation systems that differentiate performance and offer visible recognition of achievement. The "O" grade appears to serve as an external motivating factor by reinforcing effort and focus through a tangible reward for students in most groups [10]. Students in the rural group showed a slight preference for the simpler "S"/"U" system, which was possibly due to a desire to reduce pressure or stress associated with competitive assessments, or differing motivational orientations shaped by prior educational experiences [35-37].

For assessment engagement and academic performance, students in the academic group achieved higher CLO and summative scores and completed fewer attempts overall, including early passes and highest scoring attempts, compared with other groups. These results suggest that the academic group, with a science Olympiads background, had strong SRL abilities that contributed to efficient learning and better outcomes.

In contrast, rural students showed lower CLO and summative scores and needed more attempts to pass each CLO, indicating greater challenges in regulating their learning, which may stem from factors that hinder their ability to effectively manage learning [36]. These results are consistent with a previous study in Jordan, which demonstrated that students from remote areas performed worse than those admitted through the competitive pathway or the academically excellent track [1].

In the correlation analysis of summative scores, the number of passings for each CLO, additional attempts after passing, and CLO scores consistently showed positive associations, whereas instances of first-pass attempts and highest scoring attempts showed negative associations. Similarly, in the multiple linear regression analyses, the number of intentional attempts, number of passings, and CLO scores contributed positively to summative scores, whereas total attempts, first-pass attempts, and highest scoring attempts contributed negatively. Taken together, these findings indicate that formative mastery—rather than the frequency of attempts—is the strongest predictor of summative success. This interpretation aligns with prior research showing that repeated unsuccessful attempts often reflect ineffective regulation or shallow processing strategies, which in turn predict lower achievement [38,39].

These empirical patterns align well with theoretical perspectives from Zimmerman's cyclical SRL model, in which the Forethought Phase—comprising goal setting and strategic planning—shapes subsequent performance behaviors [7]. Students requiring multiple first-pass attempts or repeated highest scoring attempts may reflect weaker planning or monitoring processes—consistent with evidence that inadequate forethought predicts inefficient learning behaviors and lower achievement [7,39]. Conversely, the positive contributions of CLO scores and purposeful engagement align with findings that according to Zimmerman's SRL model, self-monitoring enables learners to detect progress and develop the efficacy needed to perform at a high level of skill, while adaptive reactions increase the effectiveness of students' learning strategies, and self-reflection feeds forward into subsequent forethought processes [7]. Similar relationships between SRL components and academic performance have been reported in medical education, where proactive goal setting and metacognitive regulation strongly predict examination outcomes [40,41].

These interpretations are based on associations and should not be construed as evidence that modifying a single engagement metric would necessarily lead to improved summative scores without concurrent changes in broader learning strategies and support. Nevertheless, addressing these challenges requires equitable access to resources, supportive environments, and targeted instruction [42]. Our findings emphasize how differences in learning behavior and assessment engagement statistics are strongly associated with academic performance. The academic group's pattern of fewer but more effective attempts, coupled with higher CLO and summative scores, likely reflects their stronger SRL skills. While performance-based distinctions such as the "O" grade may enhance motivation and engagement for some students, they may also exacerbate stress or feelings of disadvantage for others. To promote equity and acceptance of nongrading systems, educational strategies should include clear communication of purpose, tailored feedback, and early orientation—particularly for students less familiar with performance-based academic cultures. In contrast, the rural group required more attempts and demonstrated lower performance, suggesting a need for earlier support and closer monitoring. These insights underscore the importance of using assessment engagement metrics as early indicators for identifying at-risk students. Interventions such as targeted

XSL•FO
RenderX

mentoring, structured remediation, and proactive counseling, along with curriculum designs that promote SRL, may help foster equitable academic success across diverse student populations.

Limitations

This study has several limitations. First, there were disparities in sample sizes across admission groups, particularly in the quota group, which included only 6 students. This small number may have limited the statistical power of comparisons and correlation analyses within and between groups. In addition, a few students in the quota group exhibited an unusually high number of CLO attempts, contributing to outliers that may have affected the robustness and generalizability of the findings, as reflected by the very wide 95% CIs observed in this group, indicating low precision due to high variability and the small sample size. Second, the number of allowed attempts for CLO assessments varied across courses; notably, only 1 attempt was permitted for course 4, which limited the ability to comprehensively analyze assessment engagement behaviors across all courses. Third, although descriptive demographic variables were added, including age, sex, school region, and hometown region, more detailed socioeconomic and family background information (eg, parental education, first-generation college or medical school status, and household income) was not comprehensively collected in the SiCMs. The absence of these variables limits interpretation of whether contextual factors may have contributed to the observed group differences, and future studies should incorporate such measures.

Conclusions

Admission pathways play a crucial role in shaping students' learning strategies, assessment engagement behaviors, and academic performance. These findings highlight the need for medical schools to adopt equity-oriented curricular approaches that not only diversify selection processes but also actively support students with varying learning profiles throughout their training. Integrating assessment-engagement metrics with SRL data provides a scalable, data-driven framework for early identification of at-risk students and enables timely, targeted interventions—such as personalized mentoring, structured remediation, and SRL-focused instructional support. Moreover, understanding group-specific preferences toward nongrading evaluation underscores the importance of designing assessment reforms that balance academic rigor with psychological safety, ensuring fairness and acceptance across diverse student cohorts, while also recognizing that SRL strategies and engagement behaviors interact with these preferences in shaping learning outcomes. Prior research often examined these factors separately [1-3], so integrating them conceptually strengthens the evidence for equity-focused curriculum design and early support strategies. In this context, our study differs from existing research by concurrently examining SRL strategies, assessment engagement behaviors, preferences regarding nongrading evaluation, and academic performance across 4 distinct admission pathways within a single preclinical cohort. This integrated analytic approach—linking engagement metrics with SRL data and performance outcomes—provides new insights into how student characteristics and learning behaviors interact within diverse admission systems. These findings generate novel, actionable knowledge for designing equity-driven educational policies and leveraging learning analytics to inform curriculum development in medical education.

Acknowledgments

For offering invaluable guidance and assistance that greatly aided in the successful completion of this research, the authors would like to express their thanks to Associate Professor Rungnirand Praditsuwan, Former Deputy Dean for Undergraduate Education at the Faculty of Medicine Siriraj Hospital, Mahidol University. For preparing the tables, the authors would like to thank Nutradee Makphon. For manuscript formatting and data checking, the authors thank Inpreeya Inpaen for her assistance.

Funding

This study was financially supported by the Medical Council of Thailand (Police General Dr Jongjate Aojanepong Foundation) and the Faculty of Medicine Siriraj Hospital Medical Education Research Fund (IO: R016661001).

Data Availability

The datasets generated during and/or analyzed during this study are available from the corresponding author on reasonable request.

Authors' Contributions

CS and IK made substantial contributions to the conception and design. CS, IK, NJ, and NW were responsible for data acquisition. IK, CS, and PM performed the analysis and interpretation of data. IK, CS, and PM drafted the paper or revised it critically for important intellectual content. Finally, all authors approved the final version for publication.

Conflicts of Interest

None declared.

Multimedia Appendix 1

XSL•FO
RenderX

Correlations of summative scores with students' assessment engagement statistics and course learning outcome scores within the course. Blue shades represent significant positive correlations, red shades represent significant negative correlations, and gray shades represent correlation not computable due to zero variance.
[DOCX File, 51 KB - mededu_v12i1e68636_app1.docx]

Checklist 1
STROBE checklist.
[PDF File, 92 KB - mededu_v12i1e68636_app2.pdf]

Checklist 2
GRAMMS checklist.
[PDF File, 10 KB - mededu_v12i1e68636_app3.pdf]

References

1. Tamimi A, Hassuneh M, Tamimi I, et al. Admission criteria and academic performance in medical school. BMC Med Educ 2023 Apr 21;23(1):273. [doi: 10.1186/s12909-023-04251-y] [Medline: 37085824]
2. Curtis E, Wikaire E, Jiang Y, et al. Examining the predictors of academic outcomes for indigenous Māori, Pacific and rural students admitted into medicine via two equity pathways: a retrospective observational study at the University of Auckland, Aotearoa New Zealand. BMJ Open 2017 Aug;7(8):e017276. [doi: 10.1136/bmjopen-2017-017276]
3. Aston-Mourney K, McLeod J, Rivera LR, McNeill BA, Baldi DL. Prior degree and academic performance in medical school: evidence for prioritising health students and moving away from a bio-medical science-focused entry stream. BMC Med Educ 2022 Oct 4;22(1):700. [doi: 10.1186/s12909-022-03768-y] [Medline: 36195862]
4. TCAS64 round 3 application guide. version 1. Council of University Presidents of Thailand. 2021 May 6. URL: https://www.admissionpremium.com/uploads/contents/2021051090735.pdf [accessed 2025-08-25]
5. Chanmaneewong S, Yasri P. Challenges faced by high school students in Thailand when preparing for medical school admissions: a comparative study of regular Thai programmes and international schools. Worldte 2023;2(1):1-9. [doi: 10.33422/worldte.v2i1.51]
6. "Study anywhere to become a doctor, but not all are the same": Siriraj revises medical education and transforms the medical curriculum to meet learners' needs for the 2021 academic year, emphasizing the development of competent and compassionate physicians. Faculty of Medicine Siriraj Hospital, Mahidol University. 2020 Oct 22. URL: https://www.si.mahidol.ac.th/th/hotnewsdetail.asp?hn_id=2584 [accessed 2025-08-25]
7. Zimmerman BJ. Becoming a self-regulated learner: an overview. Theory Pract 2002 May;41(2):64-70. [doi: 10.1207/s15430421tip4102_2]
8. Villatoro Moral S, de-Benito Crosseti B. Self-regulation of learning and the co-design of personalized learning pathways in higher education: a theoretical model approach. J Interactive Media Educ 2022 Sep 12;2022(1). [doi: 10.5334/jime.749]
9. Ali M, Asim H, Edhi AI, et al. Does academic assessment system type affect levels of academic stress in medical students? A cross-sectional study from Pakistan. Med Educ Online 2015;20:27706. [doi: 10.3402/meo.v20.27706] [Medline: 26112353]
10. White CB, Fantone JC. Pass-fail grading: laying the foundation for self-regulated learning. Adv Health Sci Educ Theory Pract 2010 Oct;15(4):469-477. [doi: 10.1007/s10459-009-9211-1] [Medline: 20012686]
11. Wilkinson TJ, Wells JE, Bushnell JA. What is the educational impact of standards-based assessment in a medical degree? Med Educ 2007 Jun;41(6):565-572. [doi: 10.1111/j.1365-2923.2007.02766.x] [Medline: 17518836]
12. Wilkinson T. Pass/fail grading: not everything that counts can be counted. Med Educ 2011 Sep;45(9):860-862. [doi: 10.1111/j.1365-2923.2011.04018.x] [Medline: 21848711]
13. Hill MR, Goicochea S, Merlo LJ. In their own words: stressors facing medical students in the millennial generation. Med Educ Online 2018 Dec;23(1):1530558. [doi: 10.1080/10872981.2018.1530558] [Medline: 30286698]
14. Moir F, Yielder J, Sanson J, Chen Y. Depression in medical students: current insights. Adv Med Educ Pract 2018;9:323-333. [doi: 10.2147/AMEP.S137384] [Medline: 29765261]
15. Iyer AA, Hayes C, Chang BS, et al. Should medical school grading be tiered or pass/fail? A scoping review of conceptual arguments and empirical data. Acad Med 2025 Aug 1;100(8):975-985. [doi: 10.1097/ACM.0000000000006085] [Medline: 40359252]
16. Sosibo L. The views of academics on the use of student feedback for curriculum improvement. J Educ 2010 Dec;49:113-132 [FREE Full text]
17. Williams A. Delivering effective student feedback in higher education: an evaluation of the challenges and best practice. Int J Res Educ Sci 2024 May;10(2):473-501. [doi: 10.46328/ijres.3404]
18. Bendermacher GWG, De Grave WS, Wolfhagen IHAP, Dolmans DHJM, Oude Egbrink MGA. Shaping a Culture for Continuous Quality Improvement in Undergraduate Medical Education. Acad Med 2020 Dec;95(12):1913-1920. [doi: 10.1097/ACM.0000000000003406] [Medline: 32287081]

XSL•FO
RenderX

19. Broyles I, Savidge M, Schwalenberg-Leip E, Thompson K, Lee R, Sprafka S. Stages of concern during curriculum change. J Int Assoc Med Sci Educ 2007;17(1):14-26 [FREE Full text]
20. Li IW, Jackson D. Influence of entry pathway and equity group status on retention and the student experience in higher education. High Educ 2024 May;87(5):1411-1431. [doi: 10.1007/s10734-023-01070-4]
21. Eva KW, Rosenfeld J, Reiter HI, Norman GR. An admissions OSCE: the multiple mini-interview. Med Educ 2004 Mar;38(3):314-326. [doi: 10.1046/j.1365-2923.2004.01776.x] [Medline: 14996341]
22. Nithiapinyasakul A, Arora R, Chamnan P. Impact of a 20-year collaborative approach to increasing the production of rural doctors in Thailand. Int J Med Educ 2016 Dec 22;7(414–6):414-416. [doi: 10.5116/ijme.582f.4d3b] [Medline: 28008137]
23. Martín Cabrera E, García LA, Torbay Betancor Á. Factorial structure and reliability of a questionnaire on learning strategies in university students: CEA-U. An Psicol 2007 Jun;23(1):1-6 [FREE Full text]
24. Vallerand RJ, Pelletier LG, Blais MR, Briere NM, Senecal C, Vallieres EF. The Academic Motivation Scale: a measure of intrinsic, extrinsic, and amotivation in education. Educ Psychol Meas 1992 Dec;52(4):1003-1017. [doi: 10.1177/0013164492052004025]
25. Schinske J, Tanner K. Teaching more by grading less (or differently). CBE Life Sci Educ 2014;13(2):159-166. [doi: 10.1187/cbe.cbe-14-03-0054] [Medline: 26086649]
26. Spring L, Robillard D, Gehlbach L, Simas TAM. Impact of pass/fail grading on medical students' well-being and academic outcomes. Med Educ 2011 Sep;45(9):867-877. [doi: 10.1111/j.1365-2923.2011.03989.x] [Medline: 21848714]
27. Vandenbroucke JP, von Elm E, Altman DG, et al. Strengthening the Reporting of Observational Studies in Epidemiology (STROBE): explanation and elaboration. Ann Intern Med 2007 Oct 16;147(8):W163-W194. [doi: 10.7326/0003-4819-147-8-200710160-00010-w1] [Medline: 17938389]
28. O'Cathain A, Murphy E, Nicholl J. The quality of mixed methods studies in health services research. J Health Serv Res Policy 2008 Apr;13(2):92-98. [doi: 10.1258/jhsrp.2007.007074] [Medline: 18416914]
29. Tschisgale PL, Steegh A, Petersen S, Kubsch M, Wulff P, Neumann K. Are science competitions meeting their intentions? a case study on affective and cognitive predictors of success in the Physics Olympiad. Discip Interdscip Sci Educ Res 2024 Mar;6(1):10. [doi: 10.1186/s43031-024-00102-y]
30. Campbell JR. Early identification of mathematics talent has long-term positive consequences for career contributions. Int J Educ Res 1996;25(6):497-522. [doi: 10.1016/S0883-0355(97)86728-6]
31. Verna MA, Feng AX. American Chemistry Olympians achieve the highest level of equity. J Educ Res 2002;12(1):101-105 [FREE Full text]
32. Cho KK, Marjadi B, Langendyk V, Hu W. The self-regulated learning of medical students in the clinical environment—a scoping review. BMC Med Educ 2017 Jul 10;17(1):112. [doi: 10.1186/s12909-017-0956-6] [Medline: 28693468]
33. Campbell JR, Feng AX. Comparing adult productivity of American Mathematics, Chemistry, and Physics Olympians With Terman's Longitudinal Study. Roeper Rev 2010 Dec 30;33(1):18-25. [doi: 10.1080/02783193.2011.530203]
34. Ellis R, Knapton A, Cannon J, Lee AJ, Cleland J. A multivariate analysis examining the relationship between sociodemographic differences and UK graduates' performance on postgraduate medical exams. Med Teach 2025 Dec;47(12):2006-2020. [doi: 10.1080/0142159X.2025.2513426] [Medline: 40512226]
35. Wilkinson TJ, McKenzie JM, Ali AN, Rudland J, Carter FA, Bell CJ. Identifying medical students at risk of underperformance from significant stressors. BMC Med Educ 2016 Feb 2;16(1):43. [doi: 10.1186/s12909-016-0565-9] [Medline: 26837428]
36. Rodriguez F, Rivas MJ, Matsumura LH, Warschauer M, Sato BK. How do students study in STEM courses? Findings from a light-touch intervention and its relevance for underrepresented students. PLoS One 2018;13(7):e0200767. [doi: 10.1371/journal.pone.0200767] [Medline: 30063744]
37. Boateng BA, Thomas B. How can we ease the social isolation of underrepresented minority students? Acad Med 2011 Oct;86(10):1190. [doi: 10.1097/ACM.0b013e31822be60a] [Medline: 21955702]
38. Alvi E, Iqbal Z, Masood F, et al. A qualitative account of the nature and use of self-regulated learning (SRL) strategies employed by university students. Aust J Teach Educ 2016;41(8):40-59. [doi: 10.14221/ajte.2016v41n8.3]
39. Rovers SFE, Stalmeijer RE, van Merriënboer JJG, Savelberg HHCM, de Bruin ABH. How and why do students use learning strategies? A mixed methods study on learning strategies and desirable difficulties with effective strategy users. Front Psychol 2018;9:2501. [doi: 10.3389/fpsyg.2018.02501] [Medline: 30618932]
40. Bursalı Boz N, Oz H. The role of goal setting in metacognitive awareness as a self-regulatory behavior in foreign language learning. Int Online J Educ Teach 2018 Jul;5:662-671 [FREE Full text]
41. Morisano D, Locke EA. Goal setting and academic achievement. In: Hattie JAC, Anderman EM, editors. International Guide to Student Achievement: NY: Routledge; 2013:45-48 URL: https://psycnet.apa.org/record/2013-13643-014 [accessed 2025-08-25]
42. O'Neal L, Perkins A. Rural exclusion from science and academia. Trends Microbiol 2021 Nov;29(11):953-956. [doi: 10.1016/j.tim.2021.06.012] [Medline: 34274196]

Abbreviations

AES: assessment engagement statistics

XSL•FO
RenderX

CLO: course learning outcome
GPA: grade point average
GRAMMS: Good Reporting of A Mixed Methods Study
MCAR: Missing Completely at Random
MMI: multiple mini-interview
O: outstanding
S: Satisfactory
SiCMs: Siriraj Campus Management System
SRL: self-regulated learning
STROBE: Strengthening the Reporting of Observational Studies in Epidemiology
TCAS: Thai University Central Admission System
TPAT: Thai Professional Aptitude Test
U: Unsatisfactory

Edited by S Brini; submitted 12.Nov.2024; peer-reviewed by C Wallwey, L Harrison; revised version received 15.Dec.2025; accepted 21.Dec.2025; published 02.Feb.2026.

Please cite as:
Keadkraichaiwat I, Sitticharoon C, Maprapho P, Jangboon N, Wannarat N
Influence of Admission Pathways on Learning Strategies, Assessment Engagement, and Academic Performance Among First-Year Medical Students: Mixed Methods Retrospective Observational and Cross-Sectional Survey Study
JMIR Med Educ 2026;12:e68636
URL: https://mededu.jmir.org/2026/1/e68636
doi:10.2196/68636

© Issarawan Keadkraichaiwat, Chantacha Sitticharoon, Punyapat Maprapho, Nisa Jangboon, Nadda Wannarat. Originally published in JMIR Medical Education (https://mededu.jmir.org), 2.Feb.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Education, is properly cited. The complete bibliographic information, a link to the original publication on https://mededu.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Developing and Validating a Coding Scheme for Clinical Reasoning in History Taking Using Generative AI–Based Virtual Patients: Systematic Text Condensation Approach

Naping Chen[1], MM; Luzhen Tang[2], BMgt; Yang Liu[1], MM; Changmin Lin[3], MD; Zijian Li[2], BSc, BEng; Chujun Shi[1], BM; Mengyu Xia[2], BSc; Dragan Gasevic[4], DCS, Prof Dr; Danijela Gasevic[5], MD, PhD; Jinbin Zheng[6], MM; Yizhou Fan[2*], PhD; Xinyu Li[4*], DCS

[1]Department of Clinical Skills Training Center, Shantou University Medical College, Shantou, Guangdong, China

[2]Graduate School of Education, Peking University, 5th Yiheyuan Road, Beijing, China

[3]Office of Teaching Affairs, Shantou University Medical College, Shantou, Guangdong, China

[4]Faculty of Information Technology, Monash University, 25 Exhibition Walk, Clayton, Melbourne, Australia

[5]School of Public Health and Preventive Medicine, Monash University, Melbourne, Australia

[6]Department of Cardiology, The First Affiliated Hospital of Shantou University Medical College, Shantou, Guangdong, China

[*]these authors contributed equally

Corresponding Author:
Yizhou Fan, PhD
Graduate School of Education, Peking University, 5th Yiheyuan Road, Beijing, China

Abstract

Background: Effective history taking helps clinicians identify key symptoms and form accurate hypotheses. Generative artificial intelligence (GenAI)–based virtual patients (VPs) are increasingly used to simulate and practice history taking. However, there is currently no straightforward approach to effectively identify students' clinical reasoning activities during these interactions, which limits the ability to provide instructional feedback.

Objective: This study aims to develop and validate a coding scheme to identify medical students' history-taking behaviors during interactions with GenAI-based VPs.

Methods: Second-year medical students (N=210) participated in 5 history-taking cases with GenAI-based VPs, yielding 1030 dialogues. Researchers applied the systematic text condensation method to these dialogue data from cases 1 to 4 to inductively develop a coding scheme and validate coding consistency. Subsequently, the dialogue data from case 5 were used to assess the correlation between the students' history-taking behaviors and their academic performance, including diagnostic accuracy, history-taking checklist scores, clinical knowledge test scores, and postencounter form scores.

Results: A coding scheme comprising 12 behaviors across 3 dimensions—clinical reasoning behaviors, information gathering behaviors, and social interaction behaviors—was developed with high interrater reliability ($\kappa \geq 0.85$). The correlation analysis revealed that key clinical reasoning behaviors, such as *summarizing and integrating* and *logical organization*, showed significant positive correlations with multiple performance metrics, underscoring their importance in fostering clinical competence. In contrast, information gathering behaviors such as *specifying symptoms* and *routine question* were associated with clinical knowledge and thoroughness of history taking but were less predictive of diagnostic accuracy.

Conclusions: This study developed a reliable, theory-informed coding scheme that can identify students' questioning behaviors during history taking with GenAI-based VPs. The scheme effectively captures higher-order cognitive strategies and provides valuable insights into the development of clinical reasoning in medical students. This approach offers a scalable and efficient way to integrate real-time feedback into future medical education, fostering personalized learning and advancing competency-based assessments in clinical training.

(JMIR Med Educ 2026;12:e84347) doi:10.2196/84347

KEYWORDS

clinical reasoning; coding scheme; history taking; virtual patients; medical education; generative artificial intelligence

XSL•FO
RenderX

Introduction

Clinical reasoning is a core cognitive process in which clinicians gather, interpret, and analyze patient information to generate and test diagnostic hypotheses and ultimately make appropriate clinical decisions [1-3]. Clinical reasoning begins early during the initial patient encounter, as initial cues are used to generate a small set of hypotheses that guide subsequent information gathering and are iteratively tested during history taking, physical examination, and diagnostic testing [3-5]. As the first stage of clinical reasoning in patient encounters, history taking is a key step for information gathering and exerts a central influence on subsequent reasoning and decision-making [6-9]. Through patient-centered interviewing, targeted questioning, and attentive listening, effective history taking helps clinicians identify salient information, detect subtle clues, explore symptoms comprehensively, and rule out critical red-flag conditions [10,11]. Theoretically, clinicians' questioning strategies during history taking can be explained by hypothetical-deductive reasoning [5], illness script theory [12], and dual-process models [1,13-15]. Early hypothesis generation is closely linked to diagnostic accuracy [16]. In parallel, novices frequently experience challenges in eliciting and discriminating relevant cues during history taking, underscoring the need for assessment approaches that capture clinical reasoning in this phase and enable targeted educational interventions [17,18].

Current assessment methods evaluating students' clinical reasoning abilities in history taking have several limitations. Traditional scoring tools, such as checklist-based systems during history taking, predominantly assess students' recall abilities rather than their active clinical reasoning processes [18,19]. Students achieve high scores mainly by remembering and mechanically checking off predetermined items, potentially inflating their performance while inadequately reflecting actual clinical reasoning skills [20]. Additionally, more advanced assessment formats, such as key feature questions, script concordance tests, and postencounter written reflections [18,21,22], attempt to indirectly measure clinical reasoning through evaluating the quantity and accuracy of collected patient information and diagnoses [23] but are limited in capturing the dynamic, ongoing cognitive process occurring during real-time patient interactions [24] and providing personalized and timely feedback to learners.

To better prepare students for real-world clinical practice, educational initiatives and assessments must move beyond rote questioning to foster and evaluate the dynamic integration of history-taking skills and clinical reasoning abilities as an interactive process [7,25]. Innovative and practical methods are required to directly capture how students' history-taking strategies interact with their clinical reasoning skills [26,27]. Hasnain et al [28] classified history-taking strategies by analyzing how frequently students inquired about key symptoms and how thoroughly they explored presenting complaints. However, such surface-level metrics, although useful for identifying hypothesis-driven inquiry [29], fail to capture the deeper cognitive strategies central to clinical reasoning [8]. Haring et al [30] identified observable indicators of clinical reasoning across 4 domains—student actions, patient responses, conversational flow, and data gathering efficiency—through expert evaluation of students' video-recorded performances. Both Hasnain et al [28] and Haring et al [30] relied primarily on expert subjective judgment, reflecting a lack of clearly operationalized behavioral definitions. Although this approach effectively captures the interactive nature of clinical reasoning, it depends heavily on expert input and labor-intensive video analysis, limiting its scalability and consistency [31]. Operationalizing the qualitative framework proposed by Haring et al [31], Fürstenberg et al [32] developed the Clinical Reasoning Indicators-History Taking-Scale (CRI-HT-S), a structured assessment scale encompassing domains such as "focusing questions," "creating context," and "securing information." Despite these refinements, the CRI-HT-S still relies on subjective evaluation and intensive rater training; demonstrates only moderate internal consistency [33]; and is limited in capturing the depth and nuance of students' reasoning, as shown in large-scale implementations [34]. Taken together, these studies reveal a critical gap: the limited ability to conduct standardized analyses using extractable dialogue data.

The growing adoption of virtual patient (VP) simulations, including conversational agents and generative artificial intelligence (GenAI)–based VPs, has created interactive, low-risk environments in which students can repeatedly practice and refine clinical reasoning skills in authentic clinical scenarios [35-38]. In addition to supporting experiential learning, these platforms generate structured conversational data, offering unique opportunities for real-time analysis of reasoning behaviors. In response to these opportunities and the limitations of existing assessment methods, this study aimed to develop and validate a novel coding scheme tailored to capture and assess medical students' clinical reasoning processes and history-taking competence during consultation dialogues with a GenAI-based VP. Such a scheme could offer the potential to enable real-time, large-scale tracking of learners' evolving clinical reasoning skills, translating spontaneous conversational behaviors into direct, interpretable, and reliable indicators of proficiency. By operationalizing clinical reasoning in this way, the proposed scheme could support both individualized feedback and rigorous, scalable assessment in technology-enhanced medical education. To evaluate the effectiveness of this approach, this study addresses the following research question: to what extent can a coding scheme reliably and validly capture medical students' clinical reasoning competence during consultation dialogues with a GenAI-based VP?

Methods

Participants

Participants were second-year medical students at a university medical college in Shantou, China, enrolled in their second semester (N=210). At the time of data collection (April to May 2024), students had completed learning about systemic anatomy and were studying preclinical medicine, including cardiovascular and respiratory sciences. During this period, students began to construct their foundational illness scripts and develop basic skills in history taking and symptom recognition. Because participants had not yet received formal training in complex

XSL•FO
RenderX

diagnostic reasoning, we selected common chest pain presentations aligned with their curriculum to reduce construct-irrelevant variance due to advanced specialty knowledge.

Ethical Considerations

This study was approved by the Ethical Committee of Medical College, Shantou University, China (SUMC-2024 - 064). All participants received both oral and written information regarding the study's objectives and the voluntary nature of their participation. Written informed consent was obtained from all participants prior to data collection, with the explicit right to withdraw at any time without penalty, whereupon their data would be deleted and excluded from analysis. To ensure privacy and confidentiality, all collected data were anonymized and coded for analysis. Identifiable information was stored on secure, password-protected systems accessible only to the research team. Participants received a financial compensation of ¥20 (US $2.9) for each clinical case completed.

Data Collection Design

Students engaged in history-taking and clinical reasoning exercises with a GenAI-based VP operationalized as a chatbot [39]. The chatbot used a tailored prompt designed for GPT-3.5, provided in Multimedia Appendix 1. The chatbot was integrated in FLoRA [40,41] to facilitate these interactions. FLoRA is a Moodle-integrated digital learning platform specifically designed to record the entirety of each student-VP dialogue for comprehensive subsequent analysis. The study incorporated 5 representative "chest pain" cases—spontaneous pneumothorax, stable angina, aortic dissection, acute pulmonary embolism, and acute pericarditis—selected for their clinical relevance and diversity in diagnostic challenges. Across all 5 cases, a total of 1030 consultation dialogues were collected, with each case involving between 205 and 207 of the 210 participants (Table 1). For this study, we defined a consultation dialogue as an exchange between a medical student and a VP. A dialogue turn referred to 1 question-answer pair between the student and the VP. Case 1 yielded a higher average number of dialogue turns compared to subsequent cases. This increase in dialogue turns can be attributed to students' initial tendency to ask redundant questions and engage in exploratory interactions while acclimating to the FLoRA system.

Table . Dialogue turn statistics collected from 5 cases in the study.

Case	Participants, n	Total dialogue turns, n	Dialogue turns per participant, mean (SD)
1	206	18,792	91.2 (31.2)
2	207	14,753	71.3 (13.7)
3	205	15,379	75.0 (11.8)
4	206	15,723	76.3 (13.3)
5	206	15,911	77.2 (16.8)

Coding Scheme Development

The development of the coding scheme followed a rigorous, iterative process that integrated deductive theoretical grounding with inductive semantic analysis.

Deductive Foundation

The deductive coding framework was informed by prior empirical studies that examined observable behaviors during medical history taking and their relationship to clinical reasoning. In particular, we drew on the behavioral classifications described by Hasnain et al [28] and Haring et al [30], as both studies explicitly linked observable student actions during history taking to indicators of diagnostic reasoning. Hasnain et al [28] identified specific positive and negative history-taking behaviors associated with diagnostic competence, providing a structured set of behavioral indicators. Haring et al [30] used expert observation and qualitative analysis of video-recorded encounters to identify moments where clinical reasoning was demonstrated or absent in student performance. These frameworks were used as conceptual starting points for developing the initial coding scheme. The behavioral categories were subsequently adapted and operationalized to fit the context of student interactions with GenAI-based VPs. Rather than adopting them verbatim, we translated each construct into explicit, dialogue-level linguistic and interactional indicators and developed decision rules to ensure consistent coding in the interactions with GenAI-based VPs. Where necessary, we refined and split constructs to increase analytic precision (eg, separating summarizing-as-integration from summarizing-as-restatement; Multimedia Appendix 2).

Inductive Refinement

To ensure that the scheme captured the nuances of student-VP interactions, we further refined these categories inductively using Malterud's [42] systematic text condensation (STC). In the study, "meaning units" referred to students' individual utterances during interactions with GenAI-based VPs. Each utterance was treated as a discrete unit for analysis, with 1 or more codes assigned as needed on the basis of the specific intent of the inquiry. In defining each code, we drew on the principles of discourse and content analysis [43], framing the codes around 3 aspects: linguistic form, pragmatic intent, and contextual function. When inductive findings extended beyond or diverged from the initial deductive categories, we expanded the conceptual scope of existing constructs, differentiated them into analytically distinct subcategories, or introduced new codes to ensure comprehensive capture of observed behaviors. This process resulted in a theory-informed framework refined and supplemented through systematic semantic analysis. The

XSL•FO
RenderX

evolution from these theoretical indicators to the finalized codes, detailed through the 4-step STC, is presented in Table 2.

Table . Coding scheme development process based on systematic text condensation.

Step	Focus (Malterud's terminology)	Description
1	Total impression—from chaos to themes	NC reviewed all consultation dialogues (cases 1 - 4) to gain a holistic understanding. Through a combination of deductive mapping from established theory [28,30] and initial inductive semantic analysis, we identified core themes such as symptom inquiry, pathophysiological thinking, and other emergent interaction behaviors, forming a preliminary thematic scheme.
2	Identification and sorting of meaning units—from themes to codes	Dialogues from case 1 were analyzed to identify "meaning units." These units were sorted and compared against the initial themes. Through this inductive process, redundant codes were removed and similar codes merged; for example, *summarizing* was refined into 2 distinct codes based on semantic synthesis levels—*summarizing and integrating* and *summarizing and restating*.
3	Condensation—from code to meaning	The refined scheme was applied to dialogue data from cases 2 - 4 to ensure robustness across diverse clinical contexts. Through cross-case comparison, we condensed meaning units into validated "condensates," further clarifying the boundaries and definitions of each code to ensure consistency in the analysis of student behaviors.
4	Synthesis—from condensation to descriptions and concepts	The final condensates were synthesized into 12 formalized codes across 3 dimensions. This step involved a final conceptual refinement to ensure that the scheme accurately reflected clinical reasoning constructs. The resulting scheme was finalized as a standardized tool applicable to varied student-VP[a] interaction scenarios.

[a]VP: virtual patient.

Coding Scheme Refinement

To further ensure reliability, 2 authors (NC and YL), with backgrounds in clinical practice and medical education, independently conducted an initial thematic analysis of the case 1 dialogue data and collaboratively refined the coding scheme through iterative discussions. Cases 1 to 4 served as the iterative development and refinement dataset during the construction of the coding scheme. Then, NC and YL randomly selected 20 students and coded their dialogue data from cases 1 to 4. Discrepancies were resolved through iterative discussions to refine code definitions and boundaries. This evolution process, including the merging of redundant codes and the refinement of meaning units, is detailed in Multimedia Appendix 2.

For interrater reliability calibration, subsets of dialogue turns were randomly sampled and independently coded. Cohen κ was calculated on prediscussion ratings, and the calibration was conducted in multiple iterative rounds with progressively refined code definitions and decision rules until all codes achieved the predefined reliability threshold (κ≥0.85). A detailed account of the multiround reliability assessment procedure is provided in the *Results* section. Once the coding scheme was finalized, all case dialogues were fully coded, with 2 authors (NC and YL) each independently coding half of the dataset.

Applying and Validating the Coding Scheme

The coding scheme was applied to the entire corpus of 80,558 dialogue turns collected from 210 students over the 5-week study period. All dialogue data were manually coded by researchers, ensuring that the coding scheme captured the full volume of dialogues. For a detailed breakdown of all code frequencies and their distribution from case 1 to case 5, please see Multimedia Appendix 2.

To evaluate the coding scheme in a phase distinct from its iterative development, we conducted correlation analyses using case 5 (acute pericarditis). Cases 1 to 4 were used for coding scheme development and refinement; thus, using case 5 for this analysis reduced the potential bias between scheme construction and validation. Moreover, case 5 was the most diagnostically complex case in the sequence, providing a context in which clinical reasoning behaviors could be examined under higher complexity and stability. By coding these dialogues, we were able to systematically compare student behaviors as captured by the coding scheme against widely accepted criteria for assessing clinical reasoning and history-taking skills. Specifically, the coded data were correlated with multiple external benchmarks validated in prior research, allowing us to examine whether code frequencies were associated with meaningful aspects of clinical reasoning performance. These

XSL•FO
RenderX

included students' clinical knowledge test scores; diagnostic accuracy, defined dichotomously as correct or incorrect; history-taking checklist scores, adapted from the Kalamazoo Essential Elements Communication Checklist and Medical Licensing Examination criteria; and scores from a modified postencounter form [20,44-47]. These measures provided a comprehensive context for validating the relevance and sensitivity of the coding scheme to key aspects of student performance.

To quantitatively assess the relationships between code frequencies and established assessment criteria (referred to as concurrent validity) [48], we conducted Pearson correlation analysis for continuous codes and point-biserial correlation analysis or chi-square tests for binary codes. These analyses were based on the counts of codes per participant, the frequency of individual code occurrence, and students' scores on all other assessments described above. All statistical analyses were performed in Python (Python Software Foundation), with a significance threshold of $P<.05$. To control for multiple comparisons, we applied the Benjamini-Hochberg method for P value correction. By triangulating findings from various statistical tests across multiple performance metrics, this validation process provided empirical support for the theoretical coherence and practical applicability of the final coding scheme within the context of VP-based clinical reasoning education.

Results

Final Coding Scheme

The finalized coding scheme comprised 12 behavioral codes aimed at capturing students' clinical reasoning, information gathering, and social interaction behaviors during history taking (see Textbox 1 for an overview of these dimensions and their related codes). All student utterances were assigned at least 1 code across the 3 behavioral dimensions, ensuring comprehensive coverage of the dialogue corpus. A complete example of student-patient dialogue coding is provided in Multimedia Appendix 3.

Textbox 1. Dimensions and codes for identifying students' history-taking behaviors.

Dimension 1: clinical reasoning behaviors

- Pathophysiological question
- Relevant response
- Summarizing and integrating
- Logical organization

Dimension 2: information gathering behaviors

- Specifying symptoms
- Routine question
- Summarizing and restating
- Checking
- Repeating question
- Fuzzy question

Dimension 3: social interaction behaviors

- Facilitative communication
- Off-topic statement

Detailed Explanation of Codes, Definitions, and Examples

Dimension 1: Clinical Reasoning Behaviors

This dimension referred to the cognitive synthesis of patient data to generate targeted, hypothesis-driven inquiries. It focused on how students analyzed the obtained information to formulate purposeful questions that test or refine clinical diagnostic possibilities.

Pathophysiological Question

When students posed specific, hypothesis-driven questions grounded in pathophysiological thinking—such as asking about radiating pain at anatomically relevant sites, identifiable triggers or alleviating factors, or relevant past or family history—these behaviors were coded as *pathophysiological question*. The code also applied when students provided diagnostic explanations or suggestions in response to patient concerns. The defining feature of this behavior is its clear diagnostic intent. For example, in a case involving stable angina, the following responses of students were coded as *pathophysiological question*:

> *Okay, have you had any heart problems before?*
>
> *The preliminary diagnosis is angina, but I'll confirm it after you have an ECG.*
>
> *Did the pain always occur when you were lifting heavy objects or climbing slopes?*

XSL•FO
RenderX

Relevant Response

When patients mentioned diagnostically significant information such as characteristic symptoms, key clinical signs, or high-risk factors, the responses of students who recognized and responded to these features were coded as *relevant response*. The defining feature of this behavior is the student's ability to pick up on clinical cues provided by the patient and follow up with targeted inquiries. For example, the following student responses were coded as *relevant response*:

> *For how long have you been taking them? [in case of pulmonary embolism, responding to oral contraceptive use]*
>
> *How long after taking the medication does the pain relief occur? [in case of stable angina, responding to nitroglycerin use]*

Summarizing and Integrating

Summarizing and integrating the patient's information in a coherent and structured manner was coded as *summarizing and integrating*, a binary code. This behavior was not limited to repetition but involved logical reorganization of collected data, as illustrated in the following example:

> *You experienced sudden, persistent pain in your right chest this morning after forcefully blowing up a balloon, which lasted about 2 minutes before gradually easing, but it hasn't been completely relieved even after resting. You've had bilateral chest tightness that persists until now, and it worsens noticeably when walking quickly or climbing stairs....*

Logical Organization

This code was assigned on the basis of the diagnostic logic underlying a question rather than its sequential order. It identified the exploration of pertinent positive or negative associated symptoms clinically relevant to the chief complaint. Specifically, while *pathophysiological question* evaluated the mechanistic depth of an inquiry, *logical organization* captured the student's strategic ability to link the primary complaint with other clinical signs to rule in or rule out potential diagnoses, as illustrated in the following examples:

> *Do you have hematuria/reduced urine output? [in case of aortic dissection]*
>
> *Do you have shortness of breath/cyanosis? [in case of pulmonary embolism]*

Dimension 2: Information Gathering Behaviors

This dimension encompassed the student's adherence to standardized history-taking protocols and the application of foundational interviewing techniques. It focused on the procedural completeness of data collection as prescribed by the medical curriculum, ensuring a systematic acquisition of patient information.

Specifying Symptoms

This code included questions aimed at systematically exploring the characteristics of the symptom. Typical aspects included onset, duration, location, quality, severity, and aggravating or relieving factors. Unlike *logical organization*, which evaluated the diagnostic logic used to link multiple symptoms, *specifying symptoms* was strictly confined to the descriptive characterization of a single symptom. Furthermore, while *pathophysiological question* explored underlying pathophysiological mechanisms, *specifying symptoms* focused on the external attributes of the clinical presentation. Some examples are presented below:

> *What do you think triggered your chest pain today?*
>
> *Are there any factors that worsen or relieve the pain?*

Routine Question

Routine question referred to items commonly asked in all clinical interviews, regardless of the specific presenting problem. These included standard elements such as basic patient information, past medical history, personal and family history, and review of systems—questions that resembled those found on clinical interview checklists. This category also included commonly used open-ended questions in routine history taking.

> *What discomfort brought you to the clinic?*
>
> *Do you have any other discomfort that occurred at the same time as this chest pain?*
>
> *How old is your father?*

Summarizing and Restating

This code was applied when students summarized collected information by restating it without reorganization or synthesis. Unlike *summarizing and integrating*, which required thematic integration of data, *summarizing and restating* was characterized by a "linear playback" of facts, typically in chronological order, where the summary lacked a diagnostic or structured approach to reformulation. *Summarizing and restating* and *summarizing and integrating* are binary and mutually exclusive; any summarizing action must be categorized as one or the other but not both. An example of *summarizing and restating* is given below:

> *Two minutes ago, you came to the hospital due to chest pain and shortness of breath. You haven't sought medical attention before, nor have you taken any medication....*

Checking

This code was applied when students sought to confirm or clarify patient-reported information, typically in response to unclear or seemingly inconsistent statements. *Checking* was often used in combination with other codes, such as *routine question*, *specifying symptoms*, or *logical organization*, when confirmation occurred alongside other history-taking approaches. An example of *checking* behavior is shown below:

> ***Patient:*** *It lasted for about 2 minutes.*
>
> ***Student:*** *So, after 2 minutes, the pain became the same as it is now?*

Repeating Question

Repeating question was characterized by students asking about the same clinical information point multiple times, sometimes using different expressions or synonymous terms. However, whether such synonymous expressions would be coded as

XSL•FO
RenderX

repeating question depended on the student's level of clinical knowledge. Among lower-year medical students, this behavior is often associated with limited understanding of clinical terminology. For instance, they may treat synonymous terms—such as "shortness of breath" and "labored breathing"—as distinct symptoms. Consequently, such instances may result in alternative coding. The following example illustrates a *repeating question* behavior where the semantic intent remained redundant despite the change in phrasing:

> ***Student:*** *Do you experience shortness of breath or chest tightness? [logical organization]*
>
> ***Patient:*** *I feel my heart racing, sweating, and lightheaded, but without shortness of breath or chest tightness.*
>
> *[....17 rounds later]*
>
> ***Student:*** *Is your breathing normal when you have chest pain? [repeating question]*

Fuzzy Question

Fuzzy question involved the repeated use (2 or more times) of vague, open-ended questions within a single section of history taking, typically intended to prompt patients to provide additional information voluntarily. This definition excluded appropriate opening inquiries, which were coded as *routine question*. An example illustrating the distinction between *fuzzy question* and *routine question* is provided below:

> ***Student:*** *How was your health before? [routine question]*
>
> ....
>
> ***Student:*** *Do you have any other medical conditions besides high blood pressure? [routine question]*
>
> ....
>
> ***Student:*** *Any other diseases? [fuzzy question]*

Dimension 3: Social Interaction Behaviors

This dimension involved facilitative verbal interactions that serve to maintain rapport and ensure a smooth conversational flow. It focused on empathetic expressions, active listening cues, and other supportive utterances that foster the patient-interviewer relationship rather than direct clinical data collection.

Facilitative Communication

Facilitative communication included greetings, small talk, brief responses, transitional phrases, terminology explanations, expressions of reassurance, and nondiagnostic patient education. When brief conversational elements (eg, "Okay") appeared at the beginning of a question, they were not independently coded as *facilitative communication*. Instead, the entire utterance was coded according to the category assigned to the question itself (eg, *fuzzy question*, *routine question*, *logical organization*, and *specifying symptoms*), in order to avoid overcoding. However, if they were semantically distinct from other coded behaviors, dual coding was applied, assigning both *facilitative communication* and the relevant category. The following examples illustrate independent *facilitative communication* behaviors:

> *Hello, I'm Dr. Guo. [greeting]*
>
> *Take it easy. [reassurance]*
>
> *Hemoptysis means you coughed up and noticed blood. [terminology explanation]*

Off-Topic Statement

Utterances that were incomplete, illogical, or entirely unrelated to the clinical case and therefore could not be reasonably classified under any other code were categorized as *off-topic statement* to ensure coding integrity and comprehensive inclusion of all dialogue segments:

> *What kind of bird do you keep, pigeons?*
>
> *Was the balloon sucked in?*

Coding Combinations

To capture the complexity of student-patient interactions, multiple codes were assigned to a single utterance whenever it encompassed more than one functional meaning unit. This approach was not restricted to predefined combinations but was driven by the specific intent of the inquiry. This granular coding allowed for a more nuanced assessment of how students navigated between routine inquiry, symptom exploration, and diagnostic reasoning. A comprehensive taxonomy of these combinations and illustrative examples are provided in Multimedia Appendix 3.

Interrater Reliability of the Coding Scheme

Interrater reliability was established through a 4-round calibration of the 12-code scheme. In each round, we randomly sampled student utterances (≤100 utterances per student) and ensured that κ estimation for each code under evaluation was based on at least 80 independently labeled utterances [49]. Round 1 included 1034 utterances from 13 students; because several codes occurred too infrequently to support separate κ estimation, we assessed reliability at a higher level by pooling *pathophysiological question*, *relevant response*, and *summarizing and integrating* into 1 category for κ computation (κ=1.00), pooling *summarizing and restating*, *checking*, *repeating question*, and *fuzzy question* into another category for κ computation (κ=0.96) and pooling *facilitative communication* and *off-topic statement* into a group for κ computation (κ=1). For the higher-frequency individual codes, strong agreement was likewise observed: *logical organization* (κ=1.00, 95% CI 1.00-1.00), *routine question* (κ=0.99; 95% CI 0.96-1.00), and *specifying symptoms* (κ=0.96; 95% CI 0.89-1.00). After each round, 2 raters independently coded the sampled utterances, and Cohen κ was computed on the prediscussion ratings. Codes (or pooled code groups) meeting the criterion (κ≥0.85) were removed from subsequent rounds, and remaining disagreements were reviewed to refine code definitions and decision rules. Rounds 2, 3, and 4 sampled 600 utterances from 8 students, 200 utterances from 2 students, and 100 utterances from 2 students, respectively, again maintaining ≥80 labeled utterances per remaining code (or pooled group). This process continued until all codes (or pooled code groups) achieved κ ≥0.85. We report κ with 95% confidence intervals estimated via bootstrap resampling (10,000 resamples).

XSL•FO
RenderX

Associations Between Codes and Performance—Validity

The coding scheme demonstrated strong construct validity, as evidenced by significant correlations between specific behaviors and performance metrics (Table 3). Table S1 in Multimedia Appendix 2 shows the basic statistical distribution (minimum, maximum, mean, and SD) of student behavior codes and performance metrics. This provides an overview of the data's range and variability. Advanced clinical reasoning behaviors—specifically *summarizing and integrating* and *logical organization*—emerged as the primary differentiators of clinical proficiency. *Summarizing and integrating* was the most robust predictor, showing consistent, significant correlations across all measures, including diagnostic accuracy (*P*=.048). Similarly, *logical organization* served as a key indicator of success. It highlights that maintaining internal coherence by linking chief complaints with logically related symptoms is vital for an accurate diagnosis.

Table . Correlation between behavior codes and performance metrics in case 5 (acute pericarditis).[a]

Code	Diagnostic accuracy		History-taking checklist score		Clinical knowledge test score		Postencounter form score	
	r/(χ^2)	*P* value	*r*	*P* value	*r*	*P* value	*r*	*P* value
Dimension 1: clinical reasoning behaviors								
Pathophysiological question	0.082	.619	0.187	.019	0.154	.084	0.119	.273
Relevant response	−0.062	.653	0.082	.418	−0.014	.914	−0.025	.871
Summarizing and integrating	8.389	.048	0.252	<.001	0.236	.006	0.239	.012
Logical organization	0.186	.048	0.570	<.001	0.178	.044	0.144	.164
Dimension 2: information gathering behaviors								
Specifying symptoms	0.030	.730	0.344	<.001	0.262	<.001	0.067	.643
Routine question	0.141	.184	0.380	<.001	0.032	.873	0.188	.042
Summarizing and restating	0.190	.730	−0.071	.447	−0.012	.914	0.050	.643
Checking	−0.044	.715	0.019	.790	−0.129	.134	0.010	.889
Repeating question	0.049	.715	0.068	.447	−0.108	.214	0.051	.643
Fuzzy question	−0.070	.650	−0.057	.456	0.032	.873	−0.062	.643
Dimension 3: social interaction behaviors								
Facilitative communication	−0.006	.929	0.106	.268	0.141	.108	0.017	.878
Off-topic statement	0.080	.619	0.060	.456	0.008	.914	0.053	.643

[a]Pearson correlation analysis was used for continuous codes, while point-biserial correlation or χ^2 test was used for binary codes (*summarizing and integrating* and *summarizing and restating*) and binary performance (diagnostic accuracy). The *P* value was corrected with the Benjamini-Hochberg method.

In contrast, while foundational behaviors such as *specifying symptoms* and *routine question* contributed to the thoroughness of history taking, they were less predictive of diagnostic accuracy. Notably, *specifying symptoms* also showed a strong correlation with clinical knowledge scores (*P*<.001), suggesting that detailed symptom exploration is closely linked to a student's underlying knowledge base. Beyond these primary predictors, *pathophysiological question* was associated with higher thoroughness of history taking (*P*=.019), while *routine question* showed a narrow correlation with the postencounter form score (*P*=.042).

These findings suggest a clear hierarchy: while basic inquiry and data collection form the bulk of the encounter, clinical excellence is defined by high-level synthesis and the internal logical coherence of the inquiry process.

XSL•FO
RenderX

Discussion

Principal Findings

This study developed an operationalized, dialogue-based coding scheme for capturing medical students' behaviors during history taking with a GenAI-based VP. Building upon deductive blueprints, we used STC-informed iterative refinement to translate theoretical constructs into 12 observable behaviors. These behaviors, spanning clinical reasoning, information gathering, and social interaction, demonstrated high interrater agreement after structured calibration. We further provided validity evidence based on correlations with other variables: behaviors reflecting higher-order synthesis and diagnostic organization (eg, *summarizing and integrating* and *logical organization*) were more consistently associated with performance outcomes than routine or socially facilitative utterances.

Differentiating Deep Reasoning From Surface-Level Behaviors

The positive associations of *summarizing and integrating* and *logical organization* with multiple performance indicators are consistent with prior work linking diagnostic expertise to structured synthesis and clinically meaningful organization of information [22,50,51]. *Summarizing and integrating* reflects learners' ability to integrate and reorganize clinical data into a coherent representation. *Logical organization* shifts the focus from the sequence of questions to their clinical relevance, thereby operationalizing diagnostic organization more directly than order-based approaches [30].

In contrast, *pathophysiological question* was infrequent and only weakly related to the thoroughness of history taking, which likely reflects participants' early stage of biomedical knowledge and limited capacity to translate mechanistic concepts into effective hypothesis testing in real time [52,53]. *Relevant response* also showed no clear associations with performance, suggesting that novice learners may respond to salient cues reactively rather than strategically, consistent with underdeveloped illness scripts and hypothesis formulation skills [15,54,55].

Although *routine question* and *specifying symptoms* were common, their weaker links to performance suggest a "wide-net" approach emphasizing checklist completeness over interpretive reasoning [18,19]. Similarly, *checking*, *fuzzy question*, and *repeating question* may reflect attempts to manage uncertainty or cognitive load rather than deliberate diagnostic strategy [56]. *Facilitative communication* and *off-topic statement* primarily serve social or conversational functions, with no relevance to clinical problem-solving [57].

Overall, the pattern of associations supports the scheme's sensitivity to the semantic quality of inquiry: it distinguishes higher-order, logically connected reasoning moves from procedural data collection and nonclinical talk and thus provides a plausible foundation for future automated assessment and feedback.

Comparative Advantages and Generalizability of Process-Oriented Assessment

Building on prior work [28,30], we extend existing descriptions of history taking by introducing additional, operationalized behavioral indicators organized into 3 dimensions. The scheme is designed for AI-mediated learning environments: its explicit definitions support reliable human coding and provide a structured target for future automated classification and feedback. Compared with established clinical reasoning assessments, this dialogue-based approach offers complementary advantages for capturing reasoning in action. Instruments such as script concordance tests [21], clinical reasoning problems [58], and postencounter forms [18] benchmark learners' judgments against expert responses but provide limited visibility into how questions are selected and adapted during an encounter [18,58]. Oral examinations can elicit reasoning through conversation but are susceptible to examiner effects and variable reliability [59]. Objective structured clinical examinations (OSCEs) and key feature questions, while widely used, often prioritize checklist completeness or discrete decision points and may not represent the continuity of reasoning across an interview [20,60]. In contrast, our scheme codes the pragmatic intent of student utterances, enabling process-oriented assessment from naturalistic learner-patient dialogues.

Recent GenAI-enabled training systems highlight the feasibility of scalable simulation and feedback (eg, Brügge et al [34]) but commonly rely on existing inventories rather than modeling fine-grained dialogue behaviors. Our scheme addresses this gap by providing an interpretable representation of reasoning-related dialogue moves that can complement GenAI-based simulation and feedback.

Finally, the stability of code distributions across cases (Table S2 in Multimedia Appendix 2) suggests that the scheme may capture generalizable reasoning patterns beyond a single case. Ongoing work is extending the coding scheme to additional presenting complaints (eg, fever, dyspnea, and abdominal pain) to further test generalizability and reliability.

Educational Implications

The findings of this study have significant implications for the evolution of digital medical education in the age of AI.

First, the scheme provides an interpretable "intermediate representation" of learners' dialogue moves that is amenable to automation. Because the codes are defined in terms of pragmatic intent (eg, hypothesis testing vs descriptive symptom elicitation) rather than case-specific keywords, they can serve as targets for supervised natural language processing classifiers and GenAI-based annotators [61,62]. Importantly, any move toward automated coding should be framed as an empirical program rather than an assumption: future work should report classification performance (eg, macro-F1 for imbalanced codes) [63], robustness across cases and model versions, and calibration for high-stakes use before deployment in assessment contexts.

Second, a dialogue-level behavioral profile can enable more actionable formative feedback than checklist completeness alone. Code patterns such as frequent *routine question* responses without evidence of synthesis, limited *logical organization*, or

XSL•FO
RenderX

failure to respond to diagnostically salient cues can be translated into specific, teachable feedback messages and targeted prompts during subsequent practice sessions [64]. For example, learners who rarely demonstrate *logical organization* could receive scaffolds that explicitly cue key discriminating positives or negatives for competing diagnoses, whereas learners who do not produce *summarizing and integrating* statements could be prompted to generate a 1- to 2-sentence problem representation before advancing to new topic areas. Such feedback is potentially scalable, but it should be evaluated for educational impact (eg, improvement in subsequent dialogue behaviors and diagnostic accuracy) and for unintended consequences (eg, gaming the codes or reduced attention to rapport) [65].

Finally, the scheme supports longitudinal, process-oriented assessment of developing reasoning. By tracking observable reasoning behaviors across repeated encounters, educators can complement outcome measures (eg, diagnostic accuracy) with indicators of the reasoning process and progression, enabling growth-oriented competence monitoring over time [66]. Because the codes capture generalizable discourse functions rather than disease-specific scripts, the approach may support transfer across clinical domains and stages of training; nevertheless, claims about lifelong applicability require validation across learner levels, institutions, and authentic clinical contexts [67]. Overall, the scheme offers a structured bridge between history-taking instruction and data-driven, feedback-oriented educational systems while highlighting the need for rigorous evaluation of automation and downstream educational consequences.

Limitations and Future Directions

Several limitations warrant consideration. First, the study was conducted at a single institution and included only second-year students, which may limit generalizability. Second, although interrater reliability was high, manual coding limited its scalability in real-time settings. Third, while this study used 5 chest pain cases to ensure a controlled baseline appropriate for second-year students' cognitive stage, the current scope remains a limitation regarding generalizability. However, its focus on underlying cognitive mechanisms, rather than on disease-specific keywords, suggests potential utility in other domains. Our ongoing follow-up study involving 8 additional cases (eg, abdominal pain and fever) will further validate this cross-domain robustness and build upon the coding stability observed in the preliminary analysis.

Future work should explore the automation of the coding process using GenAI to enable real-time assessment and feedback. Longitudinal studies tracking students' reasoning development over time and across clinical encounters would help validate the scheme's utility. Further integration with formative assessments, OSCEs, or AI-driven learning platforms could support its transition from a research tool to an embedded component of clinical education at scale.

Conclusions

We developed an operationalized 12-code scheme to identify students' clinical reasoning–relevant behaviors during history taking with GenAI-based VPs. The scheme demonstrated high interrater agreement and provided validity evidence based on correlations with other variables, with synthesis and diagnostic organization behaviors showing stronger alignment with performance outcomes than procedural behaviors. This work establishes a reproducible foundation for process-oriented assessment of history taking and supports future research on automated coding and feedback across broader clinical domains and learner populations.

Acknowledgments

The authors used ChatGPT and Gemini (generative artificial intelligence [AI] tools) for manuscript editing, language refinement, and formatting. Specifically, these tools provided suggestions for sentence structure and clarity. The authors reviewed, verified, and modified all AI-generated suggestions and take full responsibility for the final content and results of this study.

Funding

This research was funded by the National Natural Science Foundation of China (No. 62407001), Guangdong Provincial Undergraduate Teaching Quality and Teaching Reform Project for 2025 (Document No. 4 of the Department of Higher Education of Guangdong Province, 2026, 678), and Shantou University Medical College Teaching Reform and Research Project (2025, No. 36).

Data Availability

The datasets generated or analyzed during this study are available from the corresponding author on reasonable request.

Authors' Contributions

NC contributed to study design, development of the coding scheme, data coding, and manuscript drafting. LT contributed to data collection and analysis. YL assisted in the development of the coding scheme and data coding. CL, ZL, MX, CS, and JZ participated in data collection. Dragan Gasevic and Danijela Gasevic contributed to study design and critical manuscript revision. XL and YF conceptualized the study, led data interpretation, and oversaw manuscript revision. All authors have reviewed and approved the final manuscript and agree to be accountable for all aspects of the work. XL (xinyu.li1@monash.edu) and YF (fyz@pku.edu.cn) are co-corresponding authors for this paper.

XSL•FO
RenderX

Conflicts of Interest

None declared.

Multimedia Appendix 1
Prompt template for generative artificial intelligence–based virtual patients.
[DOCX File, 18 KB - mededu_v12i1e84347_app1.docx]

Multimedia Appendix 2
Detailed development of the coding scheme, correlation with performance metrics, and cross-case distribution.
[DOCX File, 41 KB - mededu_v12i1e84347_app2.docx]

Multimedia Appendix 3
Sample of student–virtual patient dialogue with codes.
[DOCX File, 21 KB - mededu_v12i1e84347_app3.docx]

References

1. Eva KW. What every teacher needs to know about clinical reasoning. Med Educ 2005 Jan;39(1):98-106. [doi: 10.1111/j.1365-2929.2004.01972.x] [Medline: 15612906]
2. Hawks MK, Maciuba JM, Merkebu J, et al. Clinical reasoning curricula in preclinical undergraduate medical education: a scoping review. Acad Med 2023 Aug 1;98(8):958-965. [doi: 10.1097/ACM.0000000000005197] [Medline: 36862627]
3. National Academies of Sciences, Engineering, and Medicine. The diagnostic process. In: Improving Diagnosis in Health Care: National Academies Press; 2015. [doi: 10.17226/21794] [Medline: 26803862]
4. Young M, Thomas A, Lubarsky S, et al. Drawing boundaries: the difficulty in defining clinical reasoning. Acad Med 2018 Jul;93(7):990-995. [doi: 10.1097/ACM.0000000000002142] [Medline: 29369086]
5. Elstein AS, Shulman LS, Sprafka SA. Medical Problem Solving: An Analysis of Clinical Reasoning: Harvard University Press; 1978.
6. Bowen JL. Educational strategies to promote clinical diagnostic reasoning. N Engl J Med 2006 Nov 23;355(21):2217-2225. [doi: 10.1056/NEJMra054782] [Medline: 17124019]
7. Elstein AS, Schwartz A. Clinical problem solving and diagnostic decision making: selective review of the cognitive literature. BMJ 2002 Mar 23;324(7339):729-732. [doi: 10.1136/bmj.324.7339.729] [Medline: 11909793]
8. Cooper N, Bartlett M, Gay S, et al. Consensus statement on the content of clinical reasoning curricula in undergraduate medical education. Med Teach 2021 Feb;43(2):152-159. [doi: 10.1080/0142159X.2020.1842343] [Medline: 33205693]
9. Setrakian J, Gauthier G, Bergeron L, Chamberland M, St-Onge C. Comparison of assessment by a virtual patient and by clinician-educators of medical students' history-taking skills: exploratory descriptive study. JMIR Med Educ 2020 Mar 12;6(1):e14428. [doi: 10.2196/14428] [Medline: 32163036]
10. Bickley LS, Szilagyi PG, Hoffman RM, Soriano RP. Bates' Guide to Physical Examination and History Taking, 13th edition: Wolters Kluwer; 2020.
11. Nichol JR, Sundjaja JH, Nelson G. Medical history. In: StatPearls [Internet]: StatPearls Publishing; 2024. [Medline: 30484996]
12. Schmidt HG, Norman GR, Boshuizen HP. A cognitive perspective on medical expertise: theory and implication. Acad Med 1990 Oct;65(10):611-621. [doi: 10.1097/00001888-199010000-00001] [Medline: 2261032]
13. Schwartz A, Elstein AS. Clinical reasoning in medicine. In: Higgs J, Jones M, Loftus S, Christensen N, editors. Clinical Reasoning in the Health Professions, 3rd edition: Elsevier; 2008.
14. Croskerry P. A universal model of diagnostic reasoning. Acad Med 2009 Aug;84(8):1022-1028. [doi: 10.1097/ACM.0b013e3181ace703] [Medline: 19638766]
15. Yazdani S, Hosseinzadeh M, Hosseini F. Models of clinical reasoning with a focus on general practice: a critical review. J Adv Med Educ Prof 2017 Oct;5(4):177-184. [Medline: 28979912]
16. Barrows HS, Norman GR, Neufeld VR, Feightner JW. The clinical reasoning of randomly selected physicians in general medical practice. Clin Invest Med 1982;5(1):49-55. [Medline: 7116714]
17. van Merriënboer JJG, Sweller J. Cognitive load theory in health professional education: design principles and strategies. Med Educ 2010 Jan;44(1):85-93. [doi: 10.1111/j.1365-2923.2009.03498.x] [Medline: 20078759]
18. Thampy H, Willert E, Ramani S. Assessing clinical reasoning: targeting the higher levels of the pyramid. J Gen Intern Med 2019 Aug;34(8):1631-1636. [doi: 10.1007/s11606-019-04953-4] [Medline: 31025307]
19. Stivers T, Heritage J. Breaking the sequential mold: answering 'more than the question' during comprehensive history taking. Text Talk 2001;21(1-2):151-185. [doi: 10.1515/text.1.21.1-2.151]
20. Park WB, Kang SH, Lee YS, Myung SJ. Does objective structured clinical examinations score reflect the clinical reasoning ability of medical students? Am J Med Sci 2015 Jul;350(1):64-67. [doi: 10.1097/MAJ.0000000000000420] [Medline: 25647834]

XSL•FO
RenderX

21. Charlin B, van der Vleuten C. Standardized assessment of reasoning in contexts of uncertainty: the script concordance approach. Eval Health Prof 2004 Sep;27(3):304-319. [doi: 10.1177/0163278704267043] [Medline: 15312287]
22. Lai JH, Cheng KH, Wu YJ, Lin CC. Assessing clinical reasoning ability in fourth-year medical students via an integrative group history-taking with an individual reasoning activity. BMC Med Educ 2022 Jul 26;22(1):573. [doi: 10.1186/s12909-022-03649-4] [Medline: 35883069]
23. Cheng KH, Lee CY, Wu YJ, Lin CC. Using group history-taking and individual reasoning to identify shortcomings in clinical reasoning for medical students. J Med Educ Curric Dev 2024;11:23821205241280946. [doi: 10.1177/23821205241280946] [Medline: 39290776]
24. Im S, Kim DK, Kong HH, Roh HR, Oh YR, Seo JH. Assessing clinical reasoning abilities of medical students using clinical performance examination. Korean J Med Educ 2016 Mar;28(1):35-47. [doi: 10.3946/kjme.2016.8] [Medline: 26838567]
25. Windish DM, Price EG, Clever SL, Magaziner JL, Thomas PA. Teaching medical students the important connection between communication and clinical reasoning. J Gen Intern Med 2005 Dec;20(12):1108-1113. [doi: 10.1111/j.1525-1497.2005.0244.x] [Medline: 16423099]
26. Ilgen JS, Humbert AJ, Kuhn G, et al. Assessing diagnostic reasoning: a consensus statement summarizing theory, practice, and future needs. Acad Emerg Med 2012 Dec;19(12):1454-1461. [doi: 10.1111/acem.12034] [Medline: 23279251]
27. Gruppen LD. Clinical reasoning: defining it, teaching it, assessing it, studying it. West J Emerg Med 2017 Jan;18(1):4-7. [doi: 10.5811/westjem.2016.11.33191] [Medline: 28115999]
28. Hasnain M, Bordage G, Connell KJ, Sinacore JM. History-taking behaviors associated with diagnostic competence of clerks: an exploratory study. Acad Med 2001 Oct;76(10 Suppl):S14-S17. [doi: 10.1097/00001888-200110001-00006] [Medline: 11597860]
29. Hauer KE, Ten Cate O, Boscardin C, Irby DM, Iobst W, O'Sullivan PS. Understanding trust as an essential element of trainee supervision and learning in the workplace. Adv Health Sci Educ Theory Pract 2014 Aug;19(3):435-456. [doi: 10.1007/s10459-013-9474-4] [Medline: 23892689]
30. Haring CM, Cools BM, van Gurp PJM, van der Meer JWM, Postma CT. Observable phenomena that reveal medical students' clinical reasoning ability during expert assessment of their history taking: a qualitative study. BMC Med Educ 2017 Aug 29;17(1):147. [doi: 10.1186/s12909-017-0983-3] [Medline: 28851340]
31. Haring CM, Klaarwater CCR, Bouwmans GA, et al. Validity, reliability and feasibility of a new observation rating tool and a post encounter rating tool for the assessment of clinical reasoning skills of medical students during their internal medicine clerkship: a pilot study. BMC Med Educ 2020 Jun 19;20(1):198. [doi: 10.1186/s12909-020-02110-8] [Medline: 32560648]
32. Fürstenberg S, Helm T, Prediger S, Kadmon M, Berberat PO, Harendza S. Assessing clinical reasoning in undergraduate medical students during history taking with an empirically derived scale for clinical reasoning indicators. BMC Med Educ 2020 Oct 19;20(1):368. [doi: 10.1186/s12909-020-02260-9] [Medline: 33076879]
33. Bußenius L, Kadmon M, Berberat PO, Harendza S. Evaluating the Global Rating scale's psychometric properties to assess communication skills of undergraduate medical students in video-recorded simulated patient encounters. Patient Educ Couns 2022 Mar;105(3):750-755. [doi: 10.1016/j.pec.2021.06.001] [Medline: 34112546]
34. Brügge E, Ricchizzi S, Arenbeck M, et al. Large language models improve clinical decision making of medical students through patient simulation and structured feedback: a randomized controlled trial. BMC Med Educ 2024 Nov 28;24(1):1391. [doi: 10.1186/s12909-024-06399-7] [Medline: 39609823]
35. Talbot TB, Sagae K, John B, Rizzo AA, Playa C. Designing useful virtual standardized patient encounters. Presented at: Interservice/Industry Training, Simulation and Education Conference (I/ITSEC) 2012; Dec 26-29, 2012.
36. Maicher KR, Stiff A, Scholl M, et al. Artificial intelligence in virtual standardized patients: combining natural language understanding and rule based dialogue management to improve conversational fidelity. Med Teach 2022 Nov 8:1-7. [doi: 10.1080/0142159X.2022.2130216] [Medline: 36346810]
37. Hamilton A, Molzahn A, McLemore K. The evolution from standardized to virtual patients in medical education. Cureus 2024 Oct;16(10):e71224. [doi: 10.7759/cureus.71224] [Medline: 39525234]
38. Holderried F, Stegemann-Philipps C, Herschbach L, et al. A generative pretrained transformer (GPT)-powered chatbot as a simulated patient to practice history taking: prospective, mixed methods study. JMIR Med Educ 2024 Jan 16;10:e53961. [doi: 10.2196/53961] [Medline: 38227363]
39. Chen N, Tang L, Huang X, et al. Research on virtual standardized patient based on ChatGPT-4 [Article in Chinese]. Chin J Med Educ 2025;45(1):44-49. [doi: 10.3760/cma.j.cn115259-20240307-00225]
40. Li X, Fan Y, Li T, et al. The FLoRA engine: using analytics to measure and facilitate learners' own regulation activities. J Learn Anal 2025 Jan 23;12(1):391-413. [doi: 10.18608/jla.2025.8349]
41. Li X, Li T, Yan L, et al. FLoRA: an advanced AI-powered engine to facilitate hybrid human-AI regulated learning. Comput Educ 2026 Apr;243:105527. [doi: 10.1016/j.compedu.2025.105527]
42. Malterud K. Systematic text condensation: a strategy for qualitative analysis. Scand J Public Health 2012 Dec;40(8):795-805. [doi: 10.1177/1403494812465030] [Medline: 23221918]
43. Kennedy TJT, Regehr G, Baker GR, Lingard L. Point-of-care assessment of medical trainee competence for independent clinical work. Acad Med 2008 Oct;83(10 Suppl):S89-S92. [doi: 10.1097/ACM.0b013e318183c8b7] [Medline: 18820510]

XSL•FO
RenderX

44. Makoul G. Essential elements of communication in medical encounters: the Kalamazoo consensus statement. Acad Med 2001 Apr;76(4):390-393. [doi: 10.1097/00001888-200104000-00021] [Medline: 11299158]
45. Milota MM, van Thiel G, van Delden JJM. Narrative medicine as a medical education tool: a systematic review. Med Teach 2019 Jul;41(7):802-810. [doi: 10.1080/0142159X.2019.1584274] [Medline: 30983460]
46. Fürstenberg S, Oubaid V, Berberat PO, Kadmon M, Harendza S. Medical knowledge and teamwork predict the quality of case summary statements as an indicator of clinical reasoning in undergraduate medical students. GMS J Med Educ 2019;36(6):Doc83. [doi: 10.3205/zma001291] [Medline: 31844655]
47. Berger AJ, Gillespie CC, Tewksbury LR, et al. Assessment of medical student clinical reasoning by "lay" vs physician raters: inter-rater reliability using a scoring guide in a multidisciplinary objective structured clinical examination. Am J Surg 2012 Jan;203(1):81-86. [doi: 10.1016/j.amjsurg.2011.08.003] [Medline: 22172486]
48. Messick S. Validity. In: Linn RL, editor. Educational Measurement: Macmillan Publishing Co, Inc; 1989:13-103.
49. Shaffer DW, Ruis AR. How we code. In: International Conference on Quantitative Ethnography: Springer International Publishing; 2021:62-77. [doi: 10.1007/978-3-030-67788-6_5]
50. Monajemi A. The role of biomedical knowledge in clinical reasoning: bridging the gap between two theories. Int J Body Mind Culture 2014;1(2):102-106. [doi: 10.22122/ijbmc.v1i2.16]
51. Harendza S, Berberat PO, Kadmon M. Assessing competences in medical students with a newly designed 360-degree examination of a simulated first day of residency: a feasibility study. J Community Med Health Educ 2017;07(4):550. [doi: 10.4172/2161-0711.1000550]
52. Edelbring S, Parodis I, Lundberg IE. Increasing reasoning awareness: video analysis of students' two-party virtual patient interactions. JMIR Med Educ 2018 Feb 27;4(1):e4. [doi: 10.2196/mededu.9137] [Medline: 29487043]
53. Woods NN. Science is fundamental: the role of biomedical knowledge in clinical reasoning. Med Educ 2007 Dec;41(12):1173-1177. [doi: 10.1111/j.1365-2923.2007.02911.x] [Medline: 18045369]
54. Schmidt HG, Mamede S. How to improve the teaching of clinical reasoning: a narrative review and a proposal. Med Educ 2015 Oct;49(10):961-973. [doi: 10.1111/medu.12775] [Medline: 26383068]
55. Marcum JA. An integrated model of clinical reasoning: dual-process theory of cognition and metacognition. J Eval Clin Pract 2012 Oct;18(5):954-961. [doi: 10.1111/j.1365-2753.2012.01900.x] [Medline: 22994991]
56. Durning SJ, Jung E, Kim DH, Lee YM. Teaching clinical reasoning: principles from the literature to help improve instruction from the classroom to the bedside. Korean J Med Educ 2024 Jun;36(2):145-155. [doi: 10.3946/kjme.2024.292] [Medline: 38835308]
57. Kurtz S, Draper J, Silverman J. Teaching and Learning Communication Skills in Medicine, 2nd edition: CRC Press; 2017.
58. Higgs J, Jones MA, Loftus S, Christensen N. Clinical Reasoning in the Health Professions, 4th edition: Elsevier Health Sciences; 2018.
59. Wass V, Wakeford R, Neighbour R, Van der Vleuten C, Royal College of General Practitioners. Achieving acceptable reliability in oral examinations: an analysis of the Royal College of General Practitioners membership examination's oral component. Med Educ 2003 Feb;37(2):126-131. [doi: 10.1046/j.1365-2923.2003.01417.x] [Medline: 12558883]
60. Hrynchak P, Takahashi SG, Nayer M. Key-feature questions for assessment of clinical reasoning: a literature review. Med Educ 2014 Sep;48(9):870-883. [doi: 10.1111/medu.12509] [Medline: 25113114]
61. Devlin J, Chang MW, Lee K, Toutanova K. Pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Association for Computational Linguistics; 2019:4171-4186. [doi: 10.18653/v1/N19-1423]
62. Kung TH, Cheatham M, Medenilla A, et al. Performance of ChatGPT on USMLE: potential for AI-assisted medical education using large language models. PLOS Digit Health 2023 Feb;2(2):e0000198. [doi: 10.1371/journal.pdig.0000198] [Medline: 36812645]
63. Labrak Y, Bazoge A, Morin E, Gourraud PA, Rouvier M, Dufour R. BioMistral: a collection of open-source pretrained large language models for medical domains. In: Findings of the Association for Computational Linguistics: ACL 2024: Association for Computational Linguistics; 2024:5848-5864. [doi: 10.18653/v1/2024.findings-acl.348]
64. Sinclair AC, Gesel SA, LeJeune LM, Lemons CJ. A review of the evidence for real-time performance feedback to improve instructional practice. J Spec Educ 2020 Aug;54(2):90-100. [doi: 10.1177/0022466919878470]
65. Mat Sanusi KA, Iren D, Fanchamps N, Geisen M, Klemke R. Virtual virtuoso: a systematic literature review of immersive learning environments for psychomotor skill development. Education Tech Research Dev 2025 Apr;73(2):909-949. [doi: 10.1007/s11423-025-10449-2]
66. Holmboe ES, Yamazaki K, Hamstra SJ. The evolution of assessment: thinking longitudinally and developmentally. Acad Med 2020 Nov;95(Supplement_2):S7-S9. [doi: 10.1097/ACM.0000000000003649] [Medline: 32769451]
67. Riopel MA, Benham S, Landis J, Falcone S, Harvey S. The clinical reasoning assessment tool for learning from standardized patient experiences: a pilot study. Internet J Allied Health Sci Pract 2022;20(4):9. [doi: 10.46743/1540-580X/2022.2204]

Abbreviations

AI: artificial intelligence

XSL•FO
RenderX

CRI-HT-S: Clinical Reasoning Indicators-History Taking-Scale
GenAI: generative artificial intelligence
OSCE: objective structured clinical examination
STC: systematic text condensation
VP: virtual patient

Edited by S Tsuei, T Leung; submitted 18.Sep.2025; peer-reviewed by A Njoku, J Odeniya, K Naushad; revised version received 09.Mar.2026; accepted 09.Mar.2026; published 13.Apr.2026.

Please cite as:
Chen N, Tang L, Liu Y, Lin C, Li Z, Shi C, Xia M, Gasevic D, Gasevic D, Zheng J, Fan Y, Li X
Developing and Validating a Coding Scheme for Clinical Reasoning in History Taking Using Generative AI–Based Virtual Patients: Systematic Text Condensation Approach
JMIR Med Educ 2026;12:e84347
URL: https://mededu.jmir.org/2026/1/e84347
doi:10.2196/84347

© Naping Chen, Luzhen Tang, Yang Liu, Changmin Lin, Zijian Li, Chujun Shi, Mengyu Xia, Dragan Gasevic, Danijela Gasevic, Jinbin Zheng, Yizhou Fan, Xinyu Li. Originally published in JMIR Medical Education (https://mededu.jmir.org), 13.Apr.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Education, is properly cited. The complete bibliographic information, a link to the original publication on https://mededu.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

“Optimizing Learning in Integrated Curriculum”—Comparative Effectiveness of Online and Face-to-Face Formative Assessments: Mixed Methods Study

Amira Salem Ismail[1,2*], MBBS, MSc, PhD; Sarah Mohamed Hussein[3,4*], MBBS, MSc, MD; Mohamed El-Shafey[5,6*], MBBS, MSc, PhD; Ahmed Farid Al-Neklawy[3,7*], MBBS, MSc, MD; Ahd A Mansour[8*], MLS, MSc, PhD; Shatha Ghazi Felemban[8*], MLS, MSc, PhD; Shimaa Elaraby[1,2*], MBBS, MSc, PhD

[1]Medical Education Department, Fakeeh College for Medical Sciences, Fakeeh Care Group, Alhamara, Jeddah, Saudi Arabia

[2]Medical Education Department, Faculty of Medicine, Suez Canal University, Round Road, Ismailia, Egypt

[3]Clinical Sciences Department, MBBS Program, Fakeeh College for Medical Sciences, Fakeeh Care Group, Jeddah, Saudi Arabia

[4]Department of Public Health, Community Medicine, Environmental Medicine and Occupational Medicine, Faculty of Medicine, Suez Canal University, Ismailia, Egypt

[5]Physiological Sciences Department, MBBS Program, Fakeeh College for Medical Sciences, Fakeeh Care Group, Jeddah, Saudi Arabia

[6]Department of Anatomy and Embryology, Faculty of Medicine, Mansoura University, Mansoura, Dakahlia, Egypt

[7]Department of Anatomy and Embryology, Faculty of Medicine, Ain Shams University, Cairo, Egypt

[8]Medical Laboratory Sciences Department, MLS Program, Fakeeh College for Medical Sciences, Fakeeh Care Group, Jeddah, Saudi Arabia

[*]all authors contributed equally

Corresponding Author:
Amira Salem Ismail, MBBS, MSc, PhD
Medical Education Department, Fakeeh College for Medical Sciences, Fakeeh Care Group, Alhamara, Jeddah, Saudi Arabia

Abstract

Background: Assessment is a critical component of teaching and learning and serves as the foundation for how learners demonstrate success in achieving learning objectives. Formative assessments (FAs) and timely feedback play a crucial role in integrated curricula, whereas basic and clinical sciences are taught in a coordinated manner. Feedback-based FA supports student learning, and teachers can determine learning gaps to monitor progress in learning. Based on existing evidence, limited literature compared the effect of online versus onsite FA on summative performance in a fully integrated curriculum.

Objective: This study aimed to examine the effectiveness of online versus on-site FAs and feedback on summative assessment in the integrated medical curriculum.

Methods: This study used an exploratory mixed methods approach to delving into students’ experiences with face-to-face versus online FA and feedback, and its effect on their summative performance in the integrated Bachelor of Medicine, Bachelor of Surgery program. This study was conducted at Fakeeh College for Medical Sciences in Jeddah, Saudi Arabia. A total of 143 consenting students were recruited into the study. The students in the study were distributed voluntarily into 2 groups regardless of age, sex, or academic performance. Group 1 (n=92) was assigned to receive online FAs and immediate online feedback throughout the module using the Speedwell system. However, Group 2 (n=51) was assigned to receive onsite FAs and face-to-face feedback throughout the module in the examination hall in the college. The quantitative part of the study involved analyzing student scores of summative assessments in 2 groups exposed to online and onsite FA and feedback. The qualitative part aimed to explore students’ perceptions of FA and feedback.

Results: The passing rate in summative examinations (quiz, midmodule, and final) was higher in the onsite group (61.2%, 51%, and 62.7%, respectively) compared with the online group (53.3%, 48.3%, and 45.7%, respectively). However, the difference was statistically significant only in the quiz examination. Four key themes were identified from the qualitative analyses regarding participants’ different experiences of FA and feedback: the accessibility of the examination format facilitates flexibility in learning; FA is a means of recognizing learning opportunities; FAs help shift student attitudes toward learning; and the last theme is opportunities for discussion and personalized feedback.

Conclusions: This research sheds light on the intricate interplay between assessment modalities and student learning outcomes by demonstrating that onsite FA followed by onsite feedback is more effective than online FA and feedback in fostering student engagement and promoting deep understanding and improving students’ performance in summative examinations. Thereafter, this study contributes to the ongoing discourse surrounding effective assessment practices in contemporary educational settings.

XSL•FO
RenderX

(JMIR Med Educ 2026;12:e84935) doi:10.2196/84935

KEYWORDS

formative assessment; feedback; online assessment; assessment; student engagement

Introduction

The integrated curriculum is now a cornerstone of modern medical education, bridging the gap between theoretical knowledge and clinical practice through integration and embracing relevance [1]. However, the effectiveness of the integrated curricula depends heavily on the assessment strategies used that support deep learning and enhance the knowledge application and skill development [2]. These assessment strategies are a cornerstone in medical education, serving not only as a tool to evaluate learners' achievement of curricular outcomes but also as a driver of learning itself [3]. In addition, assessment can regulate learning with the delivery of feedback as a formative assessment (FA) or document achievement as a summative assessment [4].

The FA plays a pivotal role in evaluating learning progress and enhances it through continuous feedback mechanisms [5]. In medical curricula, FAs take the form of short tests with different types of questions that represent the content across different disciplines [6]. The aim of this type of assessment is the application of knowledge in clinically relevant contexts [7]. The FA is introduced to students to support learning through the knowledge gaps and guide the improvement at the level of individual students and the curriculum too [8]. FA is usually accompanied by formative feedback. Although these 2 terms are distinct, FA and formative feedback are complementary strategies [9]. The FA refers to the activities and different types of questions that provide evidence of learning, and the feedback refers to the information provided in response to that evidence [10]. FA is widely used as a powerful tool that provides learners with prompt feedback that identifies strengths and learning gaps across different disciplines, which helps students to bring relevance to their study and makes it more authentic [11]. The ongoing feedback loop provided through FA promotes critical thinking, helps students to have self-regulation strategies, and ensures that students are better prepared for high-stakes summative assessments [12]. Importantly, FA in an integrated curriculum does not merely test discrete knowledge areas; rather, it encourages students to synthesize information and apply it in clinically relevant contexts. Moreover, the FA can help in identifying areas requiring further integration and application [13]. Research studies provide evidence that FA and feedback serve as valuable strategies that facilitate student understanding, promote enhanced cognitive engagement, and foster self-regulated learning [14-16]. However, these effects vary depending on the implementation strategies of FA and feedback [17].

As digital technologies have become more common in medical education, FAs and feedback are now given in both traditional face-to-face settings and online settings [18]. Online assessments offer flexibility, immediate feedback, and opportunities for adaptive learning, while face-to-face assessments foster interactive dialogue, personalized clarification, and the development of professional communication skills [19]. Despite these differences, evidence remains limited and inconsistent regarding the comparative effectiveness of these modalities in influencing students' summative performance within an integrated curriculum [20].

Available evidence indicates that a positive and significant correlation exists between students' scores in online formative tests and their performance in summative assessments. Notably, this relationship appears more pronounced among students with higher entry qualifications [21]. According to a previous study, individualized or personalized feedback received online was found to be more effective for several students compared with face-to-face feedback or peer feedback [22]. Although online feedback was found accessible and acceptable by the majority of the students, it is not enough to ensure student engagement and depends heavily on students' self-regulation and motivation [23]. Nevertheless, limited research has been conducted comparing online FAs with automated feedback to onsite FAs with face-to-face feedback in the context of medical student performance in summative assessments [24]. Therefore, this study aimed at examining the effectiveness of online versus on-site FAs and feedback on summative assessment performance in a fully integrated curriculum. Moreover, the study explored the students' perception and experience of both assessment and feedback modalities for a better understanding of the variation in the summative performance. The study hypothesis was that onsite FA followed by face-to-face feedback is associated with higher summative assessment performance than online FA and online feedback. The study hypothesis is supported by the socio-constructivist perspective that highlights the idea that knowledge develops through interaction, dialogue, and the shared construction of meaning. These processes tend to emerge more naturally in face-to-face settings, where learners and educators can engage in real-time discussion and collaborative sense-making. Additionally, onsite feedback enhances the affective domain of learning, including confidence, motivation, and instructor scaffolding [25]. Additionally, the explanatory sequential mixed methods design was implemented to test the hypothesis through a quantitative approach, followed by deep exploration through qualitative inquiries on how students perceived different modalities of FA and feedback.

Method

Study Approach

This study used an explanatory sequential mixed methods quasi-experimental design to examine the association between the format of FA and feedback (online vs onsite) and subsequent summative students' performance. The study further sought to explore the students' experience of onsite versus online FA and feedback on their summative performance. In accordance with the explanatory sequential mixed approach, the study was conducted into 2 phases. The quantitative phase was implemented to compare the summative assessment performance

XSL•FO
RenderX

between students exposed to online versus onsite FA and feedback. The findings from the first phase informed the development of the qualitative component in 2 key approaches: first, the refinement of the interview guide to explore the differences in students' experience and perceived feedback effectiveness of the 2 different modalities of FA and feedback. Additionally, the quantitative performance outcomes supported the purposive sampling to ensure the representation of students with different achievement levels. This design helps in gaining a more complete picture than a standalone quantitative or qualitative study, as it integrates the benefits of both methods. It facilitates adding qualitative questions to the quantitative part to explore the students' different perspectives toward the face-to-face versus online FA [26]. The quantitative arm of the study involved analyzing student marks for summative assessment between 2 groups exposed to online and onsite FA and feedback. The qualitative arm aimed to explore student perceptions on this matter and was guided by an interpretive phenomenological methodology [26]. This approach assumes that individuals construct their behavior and attitudes in conjunction with their interpretation of the world around them, considering their social and cultural context. Therefore, this approach is considered valuable in understanding the student experiences associated with diverse modes of FA and feedback as investigated in this study.

Study Settings

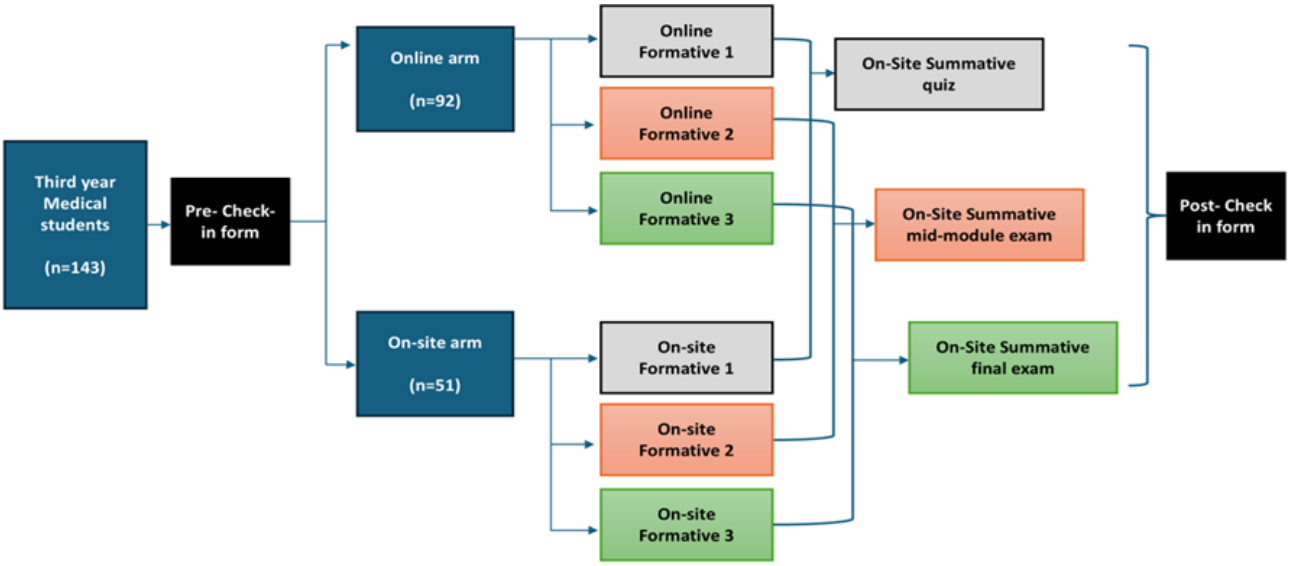
This study was conducted at Fakeeh College for Medical Sciences (FCMS) in Jeddah, Saudi Arabia. The research took place during the academic year 2019 - 2020, over 6 weeks of the integrated Endocrine and Reproductive module delivered to medical students.

Study Design and Participants Sampling

An online recruitment form was distributed to all third-year medical students, and 143 consenting students were recruited into the study. Thereafter, the students in the study were distributed voluntarily into 2 groups regardless of age, sex, or academic performance.

Group 1 (n=92) was assigned to receive online FAs and immediate online feedback throughout the module using the Speedwell system, the online examination software used in the local educational setting (Speedwell Software Ltd., 2013 - 2024). Speedwell System (Version 4.3.10088). Group 2 (n=51) was assigned to receive onsite FAs and face-to-face immediate feedback throughout the module in the examination hall in the college. All students completed the same summative assessments, as illustrated in Figure 1.

Figure 1. Distribution of the study groups of the undergraduate medical students at Fakeeh College for Medical Sciences, Jeddah, Saudi Arabia (2022), mixed method design (143 students).



The FCMS uses a structured test blueprint for formative and summative examinations for the integrated Endocrine and Reproductive module. There was a total of 3 formative examinations taken throughout the 6 weeks of the module, each taken 3 days before the respective summative examination. The FA had the same number and types of questions as summative examinations. The FAs were conducted in weeks 2, 3, and 5, preceding the summative assessments, which included a quiz, midmodule examination, and final examination, respectively. The FAs consisted of a range of question types, including multiple-choice questions, modified essay questions, and long essay questions. The questions on the formative examinations were designed to assess learning and help provide feedback to students. The questions were based on what was learned during the module leading up to that point. The scoring of these questions varied based on assigned weights. The following table (Table 1) outlines the distribution of questions and scoring used in the FAs.

After the summative examination, 2 focus groups were conducted to explore participants' experiences in relation to face-to-face FA and feedback as opposed to online modalities.

XSL•FO
RenderX

Table . The distribution of the number of questions and scoring used in the formative assessments (mixed method design).

Question types	Number	Score per question	Total score
MCQs[a]	10	1.5	15
MEQs[b]	1	2	2
LEQs[c]	1	3	3
Total	12		20

[a]MCQ: Multiple-Choice Question.

[b]MEQ: Modified Essay Question.

[c]LEQ: Long Essay Question.

Data Collection

Quantitative Data

A baseline assessment of student perceptions on FAs and feedback was sought using a check-in form on the first day of the semester for comparison with a similar evaluation at the end of the module.

Following each FA, students took 3 summative assessments throughout the module (quiz, midmodule examination, and final examination at the end of the module). Scores on the summative assessment were then recorded at the end of each summative examination.

As students began to complete the formative or summative examinations, the researchers moved on to grading, recording scores, and providing feedback to students. The researcher kept a journal throughout the module to record observations and field notes. The observations were anything that happened while students were taking formative and summative examinations. Some common observations made were common mistakes students made on the quizzes, the amount and type of feedback delivered to students, and any adjustments or changes in feedback to meet the needs of students.

Qualitative Data

Study participants were purposively invited for a focus group discussion to explore further their face-to-face and online modalities of FA and onsite and online feedback. The consenting individuals (n=20) were randomly allocated to 2 focus groups, each group consisting of 10 students, a mix of individuals who had experienced both modalities of FA in the study design. The qualitative sample size was guided by the information power principle, which states that the more relevant information the study participants hold in relation to the study aim and objectives, the fewer participants are required [26]. This principle is influenced by the specificity of the research aims and the depth of the dialogue obtained during the focus groups. In this study, participants were purposively selected from students at the same level who had experienced both online and face-to-face FAs within an integrated course. Because these participants had direct and relevant experience with the assessment modalities under investigation, the interviews generated rich and focused data aligned with the study objectives. Data collection continued until sufficient informational depth was achieved to address the research questions.

The focus groups were conducted using structured interview guides. The guide was pilot tested with 5 students who were not included in the final study sample to ensure relevance and clarity of the questions. Minor modifications are applied for the guide, including wording and sequencing of questions based on feedback obtained from the pilot testing. The interviews were conducted in English.

The focus group discussions were conducted smoothly with a nonhierarchical atmosphere. The data collectors played the role of facilitators. They asked open-ended questions and established a 2-way dialogue, avoiding dominating the discussion. The students had an equal opportunity of participation without fear of judgment or punishment. The focus group was audio-recorded and transcribed verbatim.

Data Analysis

Quantitative Data Analysis

The summative grades of all students who were enrolled in both arms of the study, namely those who underwent either online FAs or attended onsite FAs before the summative assessment, were gathered and recorded within an Excel spreadsheet. The statistical package of the social science statistical program software version 22 (SPSS Inc) was used for analyzing quantitative data. Descriptive statistics were used to assess the overall and group-specific performance in the summative assessment, in the form of percentages, median, and range. Additionally, inferential statistics were used to evaluate whether there were statistically significant differences in the pass rate of the summative examination between the 2 groups.

The summative examination marks did not adhere to a normal distribution, as determined by the Shapiro-Wilk test. Therefore, the Mann-Whitney *U* test was used to evaluate any differences in student performance between the online and onsite FA groups. The summative examinations (quiz, midmodule theory, and final theory) have different maximum scores. Therefore, to ensure standardization of results and to display the pattern of the scores for the online and the onsite groups, we transformed these row scores into Z-scores. The pass mark is 50% in all summative examinations.

Qualitative Data Analysis

The focus group discussions were transcribed verbatim and analyzed manually following Braun and Clarke's [26] approach to thematic analysis [27]. An inductive coding approach was used to generate themes from data without a predefined

XSL•FO
RenderX

theoretical framework. Two researchers engaged in independent coding of the transcripts to achieve data familiarization. To ensure consistency and credibility of the analytic approach, any discrepancies between their codes were addressed through discussion and consensus, resulting in a final codebook used for the analysis. The qualitative data analysis followed 5 stages: immersing in the data through repeated reading of transcripts, coding, combining preliminary codes into categories and subthemes, and formulating overarching themes that conveyed the meaning of the dataset, thereby facilitating the write-up of the qualitative data [27]. Throughout the study, researchers maintained reflexivity by engaging in note-taking [14,15,28]. This served to recognize individual perceptions related to the study and to ensure that conclusions were grounded in the data derived from the study.

Ethical Considerations

The research design applied in our study was approved by the Institutional Review Board at DSFH (approval number: 282/IRB/2022). Informed consent to participate was obtained from all the participants. In addition, the participants were volunteers, and no compensation was provided. All participants' information was handled in accordance with strict privacy and confidentiality standards; identifying details have been omitted, and any potentially identifiable information or images have been included only with explicit written informed consent.

Results

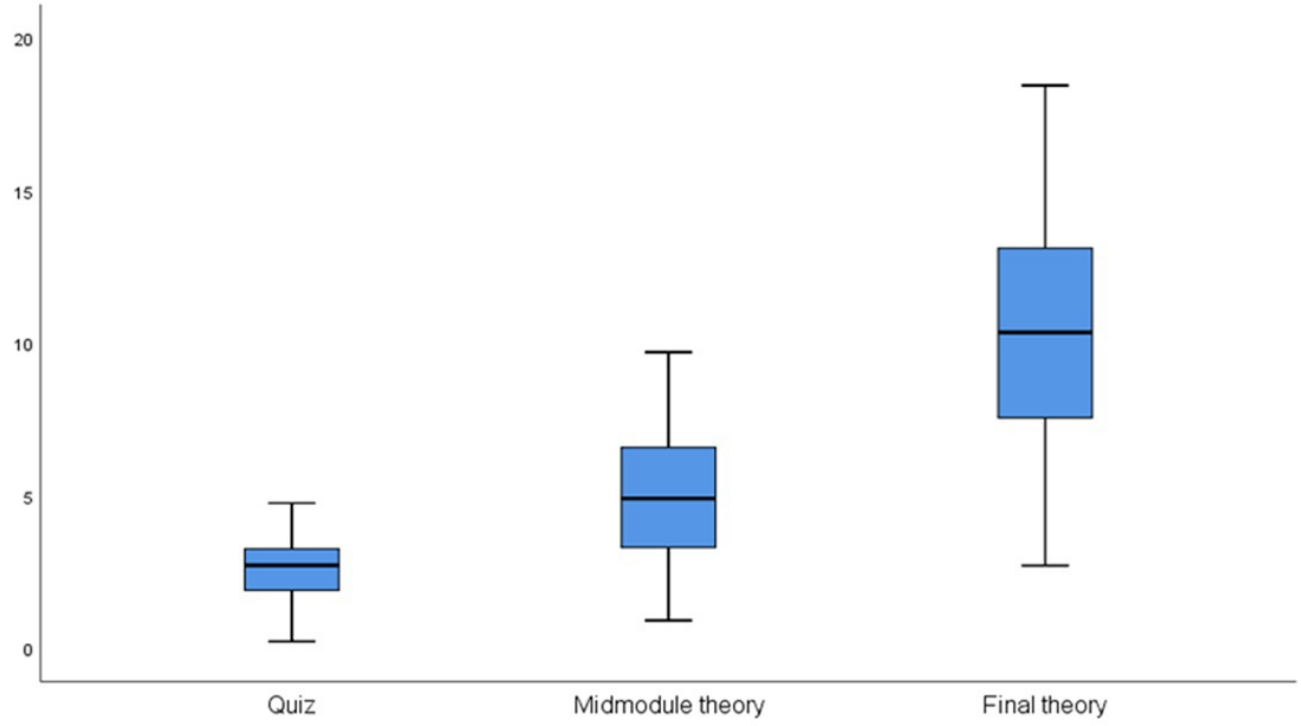
Quantitative Results

This study was carried out on a cohort of 143 third-year students enrolled in the endocrine module at FCMS. Fifty-one students were enrolled in the onsite group who received onsite FA and feedback, while 92 students were enrolled in the group who received online FA and feedback. Four students were absent from the summative quiz examination, and 5 students were absent from the summative midmodule examination.

There were 3 summative examinations: a quiz, a midmodule examination, and a final theory examination. Each examination was evaluated using a different grading scale, with scores out of 5, 10, and 20, respectively.

The median and the range of marks obtained for each of the 3 summative examinations were calculated. The median of the quiz was 2.71, with a range from 0.21 to 4.75 marks. The median of the midmodule examination was 4.9, with a range from 0.90 to 9.70 marks. The final theory examination had a median of 10.35, and a range from 2.70 to 18.45 marks (Figure 2).

Figure 2. Distribution of student marks in 3 summative examinations assessed using different modes of delivery among undergraduate medical students at Fakeeh College for Medical Sciences, Jeddah, Saudi Arabia (2022), mixed method design (143 students).



The median marks and range for the 3 summative examinations were calculated separately for the 2 student groups: the online FA and feedback, and the onsite FA and feedback groups. The onsite group had a higher median score of 2.80 marks (range 0.71 to 4.75) for the summative quiz compared with the online group, which received a median score of 2.64 marks (range 0.21 to 4.5). This difference was found to be statistically significant (P=.03).

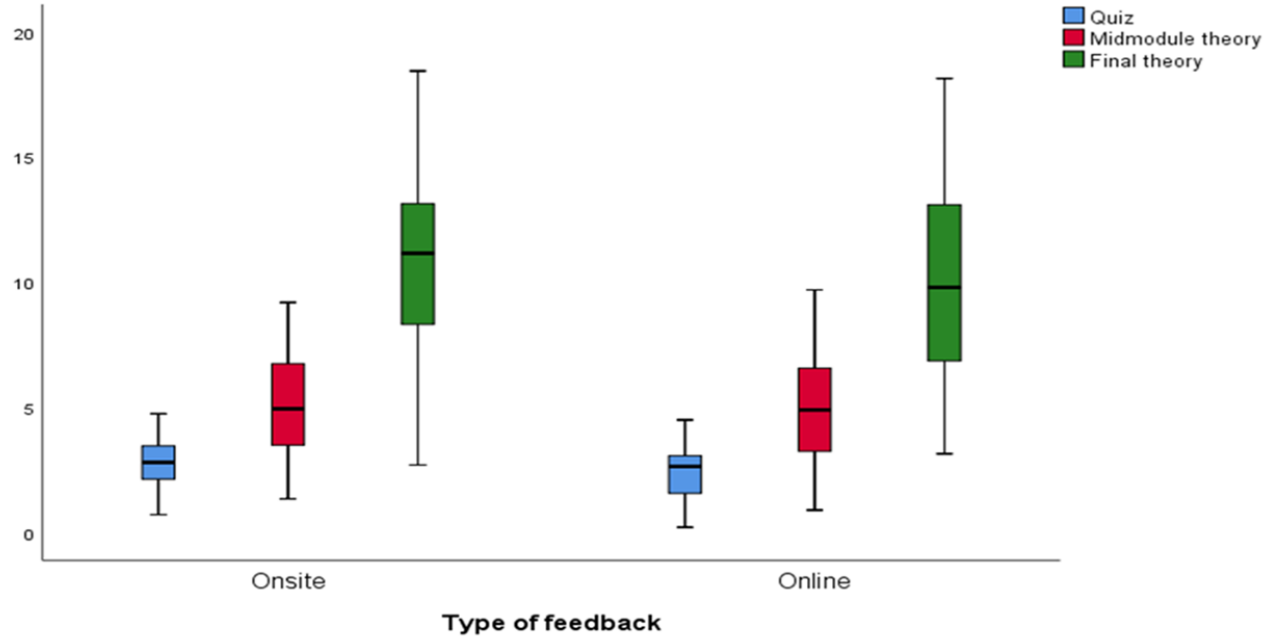
There was no statistically significant difference in the median scores for the midmodule theory examination between the onsite group (4.95 marks, range 1.35 to 9.2) and the online group (4.90 marks, range 0.9 to 9.7).

The onsite group also had a higher median score for the final theory examination (11.16 marks, range 2.7 to 18.54) compared with the online group (median: 9.8 marks, range 3.15 to 18.50).

XSL•FO
RenderX

However, this difference was not statistically significant (Figure 3).

Figure 3. Distribution of student marks in the 3 summative examinations according to mode of formative feedback delivery among undergraduate medical students at a private medical college, Jeddah, Saudi Arabia (2022) (n=143 students).



The pass rates for the 3 summative assessments were calculated and compared between 2 cohorts of students: those who received online feedback and those who received onsite feedback. Table 2 illustrates the pass rates for each examination in both groups. The results indicate that the pass rate for the summative quiz was higher among students in the onsite group (61.2%) compared with those in the online group (53.3%). Similarly, the pass rate for the midmodule theory examination was higher in the onsite group (51%) compared with the online group (48.3%). However, these differences did not reach statistical significance. In contrast, the pass rate for the final examination theory was notably higher among students in the onsite group (62.7%) than among their online counterparts (45.7%). This discrepancy approached statistical significance, with a *P* value of .05 (Table 2). The pass mark is 50% in all summative examinations.

The summative examinations (quiz, midmodule theory, and final theory) have different denominators. Therefore, these scores were transformed into Z-scores to display the pattern of the scores for the online and the onsite groups and to conduct a standardized comparison.

Table . Pass rates of summative examinations according to mode of delivery among undergraduate medical students: a mixed method study at Fakeeh College for Medical Sciences, Jeddah, Saudi Arabia (2022) (n=143).

Summative examinations	Mode of delivery		*P* value[a]
	Onsite, n (%)	Online, n (%)	
Quiz examination (n=139): pass	30 (61.2)	48 (53.3)	.37
Midmodule theory examination (n=138): pass	25 (51.0)	43(44.3)	.76
Final examination theory examination (n=143): pass	32 (62.7)	42 (45.7)	.05

[a]Statistically significance was set at *P*<.05.

Qualitative Results

The analysis of focus group discussions revealed 4 overarching themes that encapsulated participants' diverse experiences of FA and feedback. These themes were derived from a total of 17 subthemes, offering further insights into the participants' perspectives. These findings are presented below according to these 4 main themes in Table 3.

XSL•FO
RenderX

Table . Thematic analysis of the students' focus groups mixed method study at Fakeeh College for Medical Sciences, Jeddah, Saudi Arabia (2022) (n=143).

Theme	Description	Students' quotes
Theme 1: Accessibility of the examination format facilitates flexibility in learning.	This theme describes how the flexibility of examination timing and place contributed to its accessibility. Many participants expressed a preference for online formative assessments, citing their comfort and flexibility. Online assessments provide students with the convenience of completing their assessments at their own pace, allowing them to schedule their assessments around their academic workload and other commitments. This flexibility was not available with face-to-face assessments. Participants also noted that online assessments were less stressful since they eliminated the need for photocopying questions because they could revisit the examinations after submitting them. They also reported that online format was saving their time and provided more opportunities to reselect their answers. Therefore, there was a general consensus among the majority of participants that online formative assessments offered improved flexibility, accessibility, and reduced examination-related stress compared with traditional face-to-face assessments.	• "I am totally with online formative assessment, online formative assessment more flexible and accessible for example, I will take the exam after I learned all the required material and the stress related to this is very low." [FG1] • "I prefer online formative assessment, (it is) more comfortable to me and the questions were more accessible than the paper-based formative." [FG1] • "The online formative assessment provides one with the opportunities to revisit the exam questions." [FG1]
Theme 2: Formative assessment is a means of recognizing learning opportunities.	This theme describes the learning opportunities provided by the formative assessment and its effect on students' preparation. Students acknowledged that formative assessment provided a valuable opportunity to assess their knowledge, identify any areas of weakness, and recognize the need for adjustments to their learning strategies. They found this process helpful in guiding them toward meeting the necessary standards and effectively preparing for future summative examinations.	• "We test ourselves by taking the formative assessment and identifying our gaps in knowledge, so we can revisit some concept again before the summative." [FG1] • "The formative assessments guide me about the main concepts and topics that should be studied they also help me to improve my performance." [FG1]
Theme 3: Formative assessments help shift student attitudes toward learning	The theme describes how formative assessment change students' attitudes toward their study. The majority of students perceived online assessments as less stressful due to the flexibility it provided in scheduling examinations and the continued access to the assessment questions. This allowed for later reference and identification of individual learning needs, which was not possible in traditional face-to-face formative assessments. However, the reduced stress associated with online assessments was also found to lead to decreased commitment and motivation toward learning and studying for subsequent assessments. In contrast, onsite formative assessments simulated a similar stressful examination environment experienced during summative assessments. Although the associated stress in this context was seen as a drawback, it was considered beneficial in preparing students for the upcoming summative assessments. Overall, it was agreed upon by the majority that both online and onsite formative assessments influenced learning in positive and negative ways. However, it was believed that some level of examination-related stress was helpful in motivating students toward learning, making the onsite format more advantageous in this regard.	• "I will take the exam after I learned all the required material, (therefore) the stress related to this is very low so, recalling of information was easy." [FG1] • "I feel online assessment gives me the opportunity to modify my learning style as availability of questions online makes me more comfortable in identifying my learning needs. However, it (online formatives) did not give me the real experience of the summative exam and most of the time I was not committed to study well before the formative assessment." [FG1] • "Being onsite and facing onsite formative assessment gives you the same experience of being in summative exam. With the stress related to exam… of course the stress related to the summative is much higher. Being in the same situation gives you the opportunity to test yourself in similar situations." [FG2]

XSL•FO
RenderX

Theme	Description	Students' quotes
Theme 4: Opportunities for discussion and personalized feedback	This theme describes student perceptions toward feedback received through online and onsite modes. Students found the immediate provision of standardized feedback, including correct answers and the rationale was an advantage of online formative assessments compared with face-to-face feedback. However, this did not evoke the same level of enthusiasm among the majority of students. Face-to-face feedback allowed students to engage in discussions with their tutors, seek further clarification beyond the answers provided, and receive more personalized support. Hence, online feedback was considered less effective compared with face-to-face feedback due to the lack of communication and interactive discussion opportunities. Therefore, despite associated merits related to online feedback, the majority of students considered face-to-face feedback to be more beneficial.	• "Feedback with the rationale of the right answer gives you opportunities about the good application exercises with discussion and rationale sometimes." [FG2] • "Online feedback was good however it lacks communication and discussions also elaboration. I think being committed and responsible for formative assessment affects your performance in summative assessment." [FG1]

Integration of Quantitative and Qualitative Results

A joint display was developed to demonstrate how qualitative themes provide contextual explanations for the quantitative results, as shown in Table 4.

The integration table provides a more comprehensive understanding of the role of FAs and feedback within the integrated curriculum. It indicates that although online FAs increase participation through flexibility and accessibility, onsite FAs may be more effective in supporting academic performance and examination preparedness due to opportunities for interaction, personalized feedback, and exposure to authentic assessment conditions.

XSL•FO
RenderX

Table . Joint display of integrated quantitative and qualitative results.

Quantitative results	Qualitative themes	Representative student perspectives	Integrated interpretation
The results demonstrated that 36% of the students were enrolled in the onsite group who received onsite formative assessment and feedback, while 64% of the students were enrolled in the group who received online formative assessment and feedback.	Accessibility of the examination format facilitates flexibility in learning	Students described that online formative assessment provide flexibility and convenience regarding timing and remote access	The students' preference of online formative assessment is aligned with the student perspective that accessibility and convenience of online formative assessment increases flexibility in learning.
The results indicate that the pass rates for the 3 summative examinations were consistently higher among students in the onsite group. Moreover, in the quiz, the onsite group had higher median score than online group which was statistically significant.	Opportunities for discussion and personalized feedback	Students reported that onsite formative assessment and feedback allowed students to engage in discussions with their instructors, seek further clarification beyond the answers provided, and receive more personalized support. Hence, face-to-face feedback was considered more effective compared with online feedback due to the availability of communication and interactive discussion opportunities	Better performance and pass rates of face-to-face formative assessment aligns with the students' perception that face-to-face formative assessment and feedback facilitate immediate clarification and interaction with instructors.
No statistically significant differences were observed in overall summative examinations scores between students who completed online versus onsite formative assessments except for the quiz.	Formative assessment is a means of recognizing learning opportunities	Students acknowledged that formative assessment provided a valuable opportunity to assess their knowledge, identify any areas of weakness, and recognize the need for adjustments to their learning strategies	Formative assessments support students in identifying learning gaps and preparing for summative assessments This leads to similar overall scores of summative examinations despite the mode of delivery.
Similar overall academic outcomes across the 2 modalities of formative assessment and feedback despite differences in pass rates.	Formative assessments help shift student attitudes toward learning	Students perceived online assessments as less stressful, which supported flexibility but sometimes reduced motivation. Conversely, onsite assessments created examination-like pressure that helped prepare them for summative examinations	While online assessments promote accessibility and reduce stress, onsite assessments may enhance motivation and examination preparedness by simulating the pressure of summative assessment environments. The qualitative data revealed, the formative assessment helps shift student attitude toward learning supporting the results that revealed similar overall academic outcomes across the 2 modalities.

Discussion

Principal Findings

This study used an explanatory sequential mixed methods quasi-experimental approach to investigate the differential impact of onsite versus online FAs and feedback on summative assessment performance. In this design, the qualitative results informed the qualitative inquiry that was used to explore and conceptualize the students' experience and perception of the different formats of the FA and feedback. The effectiveness was assessed by comparing summative examination grades between students who received online and onsite formative feedback. Students in the onsite group achieved a higher passing rate across all summative assessments (quizzes, midterm, and final examinations) compared with their online counterparts. However, the differences were statistically significant only in the quiz examination.

Qualitative data provides important context for quantitative data interpretation, as the analysis of the data identified 4 recurring themes. Most of the students believed that online FA followed by online feedback was more flexible and accessible for them. The examination-related stress was reduced, especially with the opportunity for photocopying ideal questions and answers for revision. On the other hand, they reported that face-to-face FA followed by onsite feedback caused more stress because it was less accessible and not flexible. The participants also emphasized the importance of FA in identifying gaps that would help to improve their summative performance. Although both modalities facilitate learning, notable differences were observed in how students perceive the 2 experiences. Online assessments appear to promote self-directed learning and reduce anxiety. However, it was observed that onsite feedback was preferred by the study participants due to its ability to foster in-depth discussion and elaboration during solving application cases. This form of feedback motivated their learning to a greater extent compared with online feedback. These results also supported the quantitative data where superior summative quiz performance was observed in onsite FA and feedback, and these results reflect a better ability for students in this group to identify and address their learning gaps through more structured, engaged supervised experience, as the quiz was the first summative exposure for

XSL•FO
RenderX

students in both groups. Additionally, combining the qualitative results with students' perceptions and experience from the qualitative part conceptualizes the inconsistent outcome of online FA and feedback, although it was perceived as convenient and less stressful.

A multitude of research studies underline the significance of FA in enhancing subsequent student performance in summative assessments [14,15,28]. In addition to these findings, there is evidence of a positive association between students' engagement in course activities and their final summative grades [24]. This evidence aligns well with the outcomes of this study, where participants consistently highlighted the value of FA and feedback in supporting their performance in summative assessments. Moreover, it has been confirmed that FA has a strong effect on the academic outcomes of poor learners [15]. However, the effectiveness of formative feedback is not clear-cut, as the effectiveness of feedback is likely to reflect variations with varying approaches and student characteristics [12]. This study sought to investigate this impact further by examining the impact of online and onsite FAs on students' subsequent summative performance.

The literature on whether online feedback is superior to face-to-face feedback remains unclear, with existing studies frequently presenting contradictory findings. Previous studies indicate no significant differences in summative outcomes between learners who received online as opposed to onsite FAs [2,25,29]. However, other studies indicate a positive relationship between online formative test scores and summative assessment scores in students exposed to online formative test groups [30].

The students in this study expressed reduced stress while taking online FAs. This finding is supported by a study that examined the impact of online assessments on students' behavior and beliefs in examinations. Additionally, according to Ice et al [31], students who received audio feedback had a higher percentage of satisfaction, relative to those who received their text feedback from online channels. The clarity of information conveyed through audio feedback, as well as instructors providing clearer direction, enthusiasm, and interactivity, were cited as facilitators of its perceived effectiveness [29].

However, in this study, the number of students who attended online FA and feedback was much higher than that of students who attended the onsite FA and feedback. Previous studies have indicated that the choice of adopting an online mode of assessment could facilitate students in self-regulation of learning, in contrast to the conventional onsite mode of delivery. This could be due to the access to instant feedback in the process of online FAs, which might ease the process of learning for learners and thus facilitate better academic performance by students on the following tasks [24,32].

In line with these findings, study participants acknowledged the benefits of online FAs and the subsequent feedback they received. However, the findings also revealed that the optional nature of online assessments was associated with a decrease in motivation toward learning. Consequently, student performance during the summative assessments varied. These findings suggest that online assessment and feedback may be more effective for students who possess higher levels of capability, motivation, and self-regulation in their learning process [25,27]. An earlier study has supported this claim and showed that the engagement of students in the learning process can only be made possible through students' willingness and desire to be involved in their own learning [33].

One of these challenges, as indicated by the participants in the study, is the limited opportunities for communication, discussion, and meaningful engagement with tutors in seeking clarification on the feedback they received and on related topics. These findings align with a previous study that examined learners' perceptions of the deficiencies in online learning. The study emphasized that although learners may perceive online learning as more comfortable, the effectiveness of communication is diminished due to the absence of nonverbal cues in online feedback [34]. The notion of incorporating nonverbal cues in the delivery of online feedback is an interesting idea worth exploring further, especially in the context of the benefits associated with this. Furthermore, it is worth noting that traditional modes of providing feedback, which involve personalized and comprehensive face-to-face interaction, as well as the supportive social environment they create, have been acknowledged to motivate students to perform at a higher level [34]. There is also evidence that when teacher feedback is coupled with student engagement, this will mediate the effect between student achievement and feedback [35]. It seems that online feedback is a new challenge in replicating these benefits. However, this perspective is being challenged as research suggests that the effectiveness of the learning process is not solely determined by the mode of information delivery, but rather by the quality of interaction among peers and teachers, as well as the way the instructional content is presented [35]. Accordingly, student satisfaction is prompted mostly by active discussions and engagement among module participants. This suggests that online modes of instruction have the potential to offer opportunities for such engagement. Further studies should aim to investigate how online assessments can be proposed to provide experiences similar to those offered in traditional face-to-face settings, to enhance students' overall learning.

Various studies have explored the challenges that tutors face when interacting with students online. These challenges have many aspects, including but not limited to the psychological aspects of the use of technology, the quality of faculty teaching relative to technical literacy, and the readiness of technical infrastructure. This is all established as an important requirement [36,37]. In addition, another challenge in online delivery is the requirement for effective pedagogical strategies to provide a variety of engagement opportunities, thus promoting motivation and creating a meaningful learning experience for students [38]. This suggests that online modality in teaching and learning can be challenging due to the feelings of the students as disconnected from their classmates and instructors if interactions are not provided. These findings emphasize the importance of continued faculty development and the necessity to explore the perceptions of teachers, the challenges they face, and potential strategies to bridge any gaps associated with the delivery of online teaching and assessments. However, these aspects were not specifically addressed in this study; they warrant consideration for future research [29].

XSL•FO
RenderX

Limitations

Several factors may have influenced the study outcomes and should be considered when interpreting the findings. Student-related variables, such as motivation and self-regulation skills, were not formally measured and could act as potential confounders. Additionally, allowing students to optionally select the assessment and feedback format may have introduced variability in engagement. The quasi-experimental design also limits causal inference due to the absence of randomization and control over confounding variables. Moreover, the qualitative data did not specifically explore how FA and feedback affected preparation for summative assessments. The use of nonprobability, voluntary sampling further limits the generalizability of the results. Finally, although qualitative coding was conducted collaboratively with disagreements resolved through discussion and consensus, formal interrater reliability statistics were not calculated. While consensus-based coding is widely accepted in qualitative research, the lack of a formal reliability measure may reduce the reproducibility of the analytic process.

Conclusions

The onsite formative examination improves the performance of students in summative examinations more than online FA. The qualitative results revealed that onsite FA and feedback are better for deeper understanding and long-term knowledge retention, especially in an integrated module, while the online format offers more flexibility, accessibility, and reduces anxiety. These results suggest revisiting the approach of using FA in teaching and learning. However, given the presence of unmeasured confounders, further research is needed to clarify the mechanism of improving the summative performance through a specific FA and feedback format.

Acknowledgments

The research group for this study acknowledges the financial support provided by the Fakeeh College for Medical Sciences through the Vice Deanship for Postgraduate Studies and Scientific Research. Moreover, the research team acknowledges the participation of Dr Ben Kumwenda and Dr Dilmini Karunaratne in reviewing the whole manuscript.

Funding

Fakeeh College for Medical Sciences, Jeddah, Saudi Arabia, funded this research. No external financial support or grants were received from any public, commercial, or not-for-profit entities for the research, authorship, or publication of this article.

Data Availability

The datasets used or analyzed during this study are available from the corresponding author on reasonable request.

Authors' Contributions

All authors contribute and participate in conception and design or acquisition of data. ASI analyzed and interpreted the qualitative data of this study and participated in the methodology section. SMH analyzed and interpreted the quantitative data of this study and participated in the methodology section. AAM, MES, AFAN, and SGF participated in writing the background, introduction, and discussion sections. SE wrote the discussion sections and reviewed the whole research manuscript. All authors reviewed and approved the manuscript.

Conflicts of Interest

None declared.

References

1. Bader M, Burner T, Hoem Iversen S, Varga Z. Student perspectives on formative feedback as part of writing portfolios. Assess Eval High Educ 2019 Oct 3;44(7):1017-1028. [doi: 10.1080/02602938.2018.1564811]
2. Mitra NK, Barua A. Effect of online formative assessment on summative performance in integrated musculoskeletal system module. BMC Med Educ 2015 Mar 3;15(1):29. [doi: 10.1186/s12909-015-0318-1] [Medline: 25884641]
3. Black P, Wiliam D. Developing the theory of formative assessment. Educ Asse Eval Acc 2009 Feb;21(1):5-31. [doi: 10.1007/s11092-008-9068-5]
4. Black P, Wiliam D. Assessment and classroom learning. Assess Educ Princ Policy Pract 1998 Mar;5(1):7-74. [doi: 10.1080/0969595980050102]
5. Sadler DR. Formative assessment and the design of instructional systems. Instr Sci 1989 Jun;18(2):119-144. [doi: 10.1007/BF00117714]
6. Goodwin RL, Nathaniel TI. Effective feedback strategy for formative assessment in an integrated medical neuroscience course. Med Sci Educ 2023 Jun;33(3):747-753. [doi: 10.1007/s40670-023-01801-3] [Medline: 37501810]
7. Brenner JM, Bird JB, Willey JM. Formative assessment in an integrated curriculum: identifying at-risk students for poor performance on USMLE step 1 using NBME custom exam questions. Acad Med 2017 Nov;92:S21-S25. [doi: 10.1097/ACM.0000000000001914] [Medline: 29065019]

XSL•FO
RenderX

8. Wander B, Fernandes ATF, Daudt CVG, Gomes MQ, Pinto MEB. Curriculum integration in the formative assessment of distance continuing medical education: the use of integrative activities. Rev Bras Educ Med 2022;46(1). [doi: 10.1590/1981-5271v46.1-20210176.ing]
9. Gikandi JW, Morrow D, Davis NE. Online formative assessment in higher education: a review of the literature. Comput Educ 2011 Dec;57(4):2333-2351. [doi: 10.1016/j.compedu.2011.06.004]
10. Xiao Y, Yang M. Formative assessment and self-regulated learning: how formative assessment supports students' self-regulation in English language learning. System 2019 Apr;81:39-49. [doi: 10.1016/j.system.2019.01.004]
11. El-Yassin HD. Integrated assessment in medical education. J Contemp Med Sci 2015 Sep 1(4):36-38 [FREE Full text]
12. Lipnevich AA, Mattern K, Feddock C. Formative assessment and feedback in medical education: a practical guide: AMEE guide No. 189. Med Teach 2025 Oct 24:1-20. [doi: 10.1080/0142159X.2025.2569623] [Medline: 41134874]
13. Hattie J, Timperley H. The power of feedback. Rev Educ Res 2007 Mar;77(1):81-112. [doi: 10.3102/003465430298487]
14. Shute VJ. Focus on formative feedback. Rev Educ Res 2008 Mar;78(1):153-189. [doi: 10.3102/0034654307313795]
15. Cohen A, Singh D. Effective student feedback as a marker for student success. SAJHE 2020;35(4). [doi: 10.20853/34-5-4259]
16. Bennett RE. Formative assessment: a critical review. Assess Educ Princ Policy Pract 2011 Feb;18(1):5-25. [doi: 10.1080/0969594X.2010.513678]
17. Brookhart SM. Expanding views about formative classroom assessment: a review of the literature. In: Formative Classroom Assessment: Theory into Practice: Teachers College Press; 2007.
18. Birenbaum M, DeLuca C, Earl L, et al. International trends in the implementation of assessment for learning: Implications for policy and practice. Policy Futures Educ 2015 Jan;13(1):117-140. [doi: 10.1177/1478210314566733]
19. Li H. How is formative assessment related to students' reading achievement? findings from PISA 2009. Assess Educ Princ Policy Pract 2016 Oct;23(4):473-494. [doi: 10.1080/0969594X.2016.1139543]
20. Moreno J, Pineda AF. A framework for automated formative assessment in mathematics courses. IEEE Access 2020 Feb;8:30152-30159. [doi: 10.1109/ACCESS.2020.2973026]
21. Yan Z, Brown GTL. Assessment for learning in the Hong Kong assessment reform: a case of policy borrowing. Stud Educ Eval 2021 Mar;68:100985. [doi: 10.1016/j.stueduc.2021.100985]
22. Robertson SN, Humphrey SM, Steele JP. Using technology tools for formative assessments. JEO 2019 Jul;16(2):2. [doi: 10.9743/JEO.2019.16.2.11]
23. Bhagat KK, Spector JM. Formative assessment in complex problem-solving domains: the emerging role of assessment technologies. J Educ Techno Soc 2017;20(4):312-317 [FREE Full text]
24. Ozan C, Kincal RY. The effects of formative assessment on academic achievement, attitudes toward the lesson, and self-regulation skills. Educ Sci Theory Pract 2018 Feb;18(1):85-118 [FREE Full text] [doi: 10.12738/estp.2018.1.0216]
25. Creswell JW. Research designs. In: Qualitative, Quantitative, and Mixed Methods Approaches: SAGE Publications; 2009.
26. Braun V, Clarke V. Successful Qualitative Research: A Practical Guide for Beginners: SAGE Publications Ltd; 2013.
27. Stodel EJ, Thompson TL, MacDonald CJ. Learners' perspectives on what is missing from online learning: interpretations through the community of inquiry framework. IRRODL 2006 Dec;7(3):1-24. [doi: 10.19173/irrodl.v7i3.325]
28. Azam F, Irshad K, Shaheen A, Moin H, Javed N, Ahmer H. Online versus conventional paper based formative assessment: do they predict summative scores? Pak J Phsyiol 2021 Mar 31;17(1):46-50. [doi: 10.69656/pjp.v17i1.1312]
29. Cassady JC, Gridley BE. The effects of online formative and summative assessment on test anxiety and performance. J Technol Learn Assess 2005 Oct 1;4(1) [FREE Full text]
30. Henly DC, Reid AE. Use of the web to provide learning support for a large metabolism and nutrition class. Biochem Molecular Bio Educ 2001 Nov;29(6):229-233. [doi: 10.1111/j.1539-3429.2001.tb00129.x]
31. Ice P, Curtis R, Phillips P, Wells J. Using asynchronous audio feedback to enhance teaching presence and students' sense of community. J Asynchronous Learn Netw 2007;11(2):3-25. [doi: 10.24059/olj.v11i2.1724] [Medline: 40151300]
32. Miller RL, Amsel E, Kowalewski BM, Beins BC, Keith KD, Peden BF. Promoting student engagement. In: Volume 1: Programs, Techniques and Opportunities: Society for the Teaching of Psychology; 2011.
33. Yang L, Chiu MM, Yan Z. The power of teacher feedback in affecting student learning and achievement: insights from students' perspective. Educ Psychol (Lond) 2021 Aug 9;41(7):821-824. [doi: 10.1080/01443410.2021.1964855]
34. Al Kuhayli H, Pilotti M, El Alaoui K, Cavazos SE, Hassan SA, Al Ghazo R. An exploratory non-experimental design of self-assessment practice. Int J Assess Eval 2019;26(1):49-65. [doi: 10.18848/2327-7920/CGP/v26i01/49-65]
35. Mulvenon S. A critical review of research on formative assessment: the limited scientific evidence of the impact of formative assessment in education. Pract Assess Res Eval 2021 Sep 15. [doi: 10.7275/jg4h-rb87]
36. Zhao CG, Liao L. Metacognitive strategy use in L2 writing assessment. System 2021 Jun;98:102472. [doi: 10.1016/j.system.2021.102472]
37. Yan X, Zhang C, Fan JJ. "Assessment knowledge is important, but …": how contextual and experiential factors mediate assessment practice and training needs of language teachers. System 2018 Jun;74:158-168. [doi: 10.1016/j.system.2018.03.003]
38. van der Vleuten CPM, Schuwirth LWT, Driessen EW, et al. A model for programmatic assessment fit for purpose. Med Teach 2012;34(3):205-214. [doi: 10.3109/0142159X.2012.652239] [Medline: 22364452]

XSL•FO
RenderX

Abbreviation

FA: formative assessment
FCMS: Fakeeh College for Medical Sciences

Edited by A Stone, T Leung; submitted 27.Sep.2025; peer-reviewed by M Groenier, T Mesbah; revised version received 14.Mar.2026; accepted 16.Mar.2026; published 05.May.2026.

Please cite as:
Ismail AS, Hussein SM, El-Shafey M, Al-Neklawy AF, Mansour AA, Felemban SG, Elaraby S
"Optimizing Learning in Integrated Curriculum"—Comparative Effectiveness of Online and Face-to-Face Formative Assessments: Mixed Methods Study
JMIR Med Educ 2026;12:e84935
URL: https://mededu.jmir.org/2026/1/e84935
doi:10.2196/84935

© Amira Salem Ismail, Sarah Mohamed Hussein, Mohamed El-Shafey, Ahmed Farid Al-Neklawy, Ahd A Mansour, Shatha Ghazi Felemban, Shimaa Elaraby. Originally published in JMIR Medical Education (https://mededu.jmir.org), 5.May.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Education, is properly cited. The complete bibliographic information, a link to the original publication on https://mededu.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Virtual Reality and Gamification for Assessing Technical Aptitude, Cognitive Abilities, and Personality Characteristics in Surgical Residency Selection: Validation Study

Noa Gazit[1,2], PhD; Gilad Ben-Gal[2*], PhD; Ron Eliashar[1*], MD

[1]Department of Otolaryngology/HNS, Faculty of Medicine, Hebrew University of Jerusalem, Hadassah Medical Center, Kalman Ya'akov Man 1, Jerusalem, Israel

[2]Department of Prosthodontics, Faculty of Dental Medicine, Hebrew University of Jerusalem, Hadassah Medical Center, Jerusalem, Israel

[*]these authors contributed equally

Corresponding Author:
Noa Gazit, PhD
Department of Otolaryngology/HNS, Faculty of Medicine, Hebrew University of Jerusalem, Hadassah Medical Center, Kalman Ya'akov Man 1, Jerusalem, Israel

Abstract

Background: Assessment of technical aptitude, cognitive abilities, and personality characteristics is important in selecting candidates for surgical training. Currently, the selection of surgical training candidates does not systematically include objective assessment of these variables. Instead, it relies heavily on traditional selection methods, such as academic achievement, letters of recommendation, and interviews, whose presumed relationships with later performance are based on limited and inconsistent evidence.

Objective: This study examined evidence for validity based on relationships with other variables—a key source of validity evidence—to support the use of 2 novel tools for objectively assessing surgical training candidates: a nonimmersive (screen-based) virtual reality laparoscopic technical aptitude test, and a game-based assessment of cognitive abilities and personality characteristics.

Methods: This study had 3 phases, all focused on establishing validity based on relationships with other variables. In Phase 1, we evaluated convergent and discriminant evidence of validity by assessing the correlation between interns' performance in the 2 novel selection tests and in 4 established psychometric instruments for assessing dexterity, visuospatial ability, intelligence, and personality. In Phase 2, we evaluated evidence for the test-criterion relationship by assessing the correlation between residents' performance in the 2 novel tests and their concurrent residency performance evaluations. In this phase, we also assessed evidence for the fairness of the tests between genders. In Phase 3, we focused on the technical aptitude test and evaluated evidence for its relationship with training level by administering the test to a sample of expert surgeons and comparing their performance with that of the residents and interns from the previous phases.

Results: Interns' scores on the 2 novel selection tests were correlated with scores on the relevant established psychometric instruments, providing convergent and discriminant evidence (Phase 1). Residents' scores on the 2 novel tests were significantly correlated with relevant performance criteria (Phase 2). In addition, no evidence for gender bias in the tests was found. Finally, based on data collected in all 3 phases, we found evidence for expert-novice differences in the technical aptitude test scores, such that scores were correlated with surgical experience.

Conclusions: The findings provide validity evidence supporting the use of the novel virtual reality–based technical aptitude test and a game-based assessment of cognitive abilities and personality characteristics in selecting candidates for surgical training. However, because the test-criterion evidence was obtained using a concurrent design, further prospective longitudinal studies are required to determine whether these assessments predict subsequent residency performance.

(JMIR Med Educ 2026;12:e82515) doi:10.2196/82515

KEYWORDS

resident selection; assessment; surgical training; technical aptitude; cognitive abilities; personality characteristics; surgical simulators; gamification; game-based assessment

XSL•FO
RenderX

Introduction

Background

The selection of candidates for surgical training is a crucial step in ensuring the quality and safety of patient care. The primary objective of the selection process is to identify the most suitable candidates, those who will perform well both during their residency and ultimately as independent surgeons. For an effective selection process for surgical training, it is widely accepted that prospective candidates should be assessed objectively on their technical aptitude (eg, dexterity, coordination, and visuospatial ability), cognitive abilities (eg, deductive and inductive reasoning, learning ability, and concentration), and personality characteristics (eg, decision-making, stress tolerance, and communication skills) [1-7].

In practice, many surgical training programs have formalized and structured selection procedures. However, these procedures rely predominantly on traditional selection methods such as academic achievement, letters of recommendation, and interviews, which do not systematically incorporate standardized, performance-based assessments [8,9]. Evidence regarding the relationships between these traditional methods and subsequent clinical and operative performance during residency is mixed and context-dependent, with several studies demonstrating only weak to modest associations [10-14]. Efforts to identify more suitable tools for objectively assessing candidates' capacities and characteristics have included surrogate tests for assessing technical aptitude (indirect indicators of nonspecific technical abilities, usually in paper-and-pencil or computerized formats); medical exams (eg, the USMLE [United States Medical Licensing Examination]) for assessing cognitive abilities; and self-report questionnaires for assessing personality characteristics (eg, the "Big Five" personality traits, emotional intelligence, and grit). However, there is as yet no consistent evidence that these methods improve the selection of surgical residents [4,10,15-17]. The limited and inconsistent validity evidence supporting these approaches may partly explain why their use has remained largely confined to experimental or pilot contexts, with limited translation into routine, large-scale selection practice.

This research examines an alternative approach to the selection of surgical residents based on innovative simulation assessment methods [18,19], similar to those used in the selection process for aircrew [20]. Through simulation, examinees are exposed to controlled situations designed to elicit behaviors relevant to the assessment of particular competencies. These situations can be designed so that they mimic the challenges and tasks associated with real-life surgical work. Therefore, these methods are expected to be more effective than traditional methods of assessment or other surrogates tried thus far.

Simulative testing can be conducted either in the real world by evaluators or actors, or on a computer or a simulator, using innovative methods such as virtual reality (VR) and gamification. VR is broadly defined as a computer-generated simulation of a realistic 3D environment in which users interact with digitally constructed stimuli in real time [21]. VR systems are characterized by varying degrees of presence (the subjective sense of "being there"), immersion (the extent to which the system technologically replaces or augments sensory input), and interactivity (the user's ability to influence the virtual environment through active responses) [22]. According to established classifications, VR technologies range from nonimmersive desktop-based systems displayed on standard monitors to fully immersive head-mounted display environments. Nonimmersive systems, while providing lower levels of sensory input, nonetheless allow real-time interaction with a computer-generated environment and performance-based feedback.

Gamification refers to the incorporation of game design elements, such as goals, rules, feedback systems, scoring mechanisms, levels, and reward structures, into nongame contexts in order to enhance engagement, motivation, and behavioral activation [23]. This approach has led to the development of game-based assessments (GBAs), which embed psychometrically informed measurement principles within interactive digital environments [24-27]. GBAs emerged within educational technology as part of a broader shift toward digital and data-driven assessment, alongside efforts to enhance learning through adaptive personalization [28]. Rather than relying solely on final test scores, such assessment approaches analyze in-game behavioral data, such as response times, movement trajectories, action sequences, error patterns, decision latency, and changes in strategy following feedback, to infer underlying cognitive and noncognitive competencies [29]. Advances in computational psychometrics then allow modeling these behavioral traces using principled statistical frameworks in order to link observable digital actions to latent constructs, such as problem-solving, working memory, persistence, visuospatial reasoning, flexibility, or decision-making under uncertainty [30]. In parallel, evidence-centered design frameworks provide structured approaches for aligning task design, observed behaviors, and intended score interpretations within immersive digital environments [31]. In the domain of personnel selection, GBAs improve on traditional self-report questionnaires or decontextualized cognitive tests because they allow the inference of job-relevant skills, abilities, and personality characteristics from observable gameplay behaviors, including response times, decision patterns, error rates, learning curves, and adjustments in strategy following feedback or changing task demands. These behavioral traces are captured continuously and transformed into quantifiable performance indicators, as described above.

Importantly, the VR and gamification paradigms are not mutually exclusive. In particular, a simulation-based assessment may use VR technology while simultaneously incorporating gamified design elements such as structured goals, scoring systems, feedback mechanisms, and progressive task levels. This study focuses on the validation of 2 separate tools: a game-based assessment administered on a standard computer for measuring cognitive abilities and personality characteristics; and a screen-based laparoscopic virtual reality simulator for assessing technical aptitude. However, the VR-based test also incorporates several gamified elements intended to structure performance and enhance engagement (see the description of

XSL•FO
RenderX

the tests in the "Methods" section for more details). Both tools are designed to support standardized measurement and defensible score interpretation in a high-stakes surgical selection context.

VR and gamification are promising directions in assessment that have numerous advantages over conventional methods [24-26]. First, they promote a more positive assessment experience that reduces examinees' stress levels and increases their engagement and motivation. Second, the use of computerized systems allows collecting rich, high-resolution spatiotemporal behavioral data, which provides a great deal of information about each participant's performance [32,33]. Since performance is assessed continuously throughout the tasks based on multiple parameters, this results in more valid and reliable assessments. Finally, these assessments are based on an automated scoring system, which eliminates the bias associated with human assessments [34,35] or self-report questionnaires [36,37]. These features of VR-based simulation and game-based assessment may contribute to more valid and reliable evaluation of candidates' skills and abilities.

The implementation of VR and gamification in assessment is a relatively new direction in personnel selection in general, and in the selection of candidates for surgical training in particular. Although a few studies have provided initial evidence regarding the potential of using VR simulators to assess candidates' technical aptitude [38-42], there is not yet sufficient evidence supporting the validity of these simulators for selecting candidates for surgical training [43] (evidence for test content, response process, internal structure, relationships with other variables, and consequences, as described by Messick [44,45]). In addition, studies exploring the implementation of VR or gamification for assessing the cognitive abilities and personality characteristics of candidates for medical residency programs (surgical or nonsurgical) are scarce.

Study Objectives

This paper is part of a larger research project addressing the systematic development and validation of 2 new simulation-based assessments for resident selection which integrate novel VR and gamification techniques [46,47]: a technical aptitude test performed on the nonimmersive (screen-based) Lap-X-VR laparoscopic simulator [48]; and a computerized GBA of cognitive abilities and personality characteristics, which comprises 3 video games designed specifically for assessment.

Consistent with contemporary validity theory, validity is not a property of a test itself but of the interpretation and proposed use of its scores [49]. According to Messick's unified framework, validity is supported by evidence from multiple sources, including content, response processes, internal structure, relationships with other variables, and consequences [44,45]. In 2 previous studies [46,47], we presented initial evidence primarily addressing content, response process, and internal structure to support the proposed interpretation and use of these tests in surgical residency selection. The aim of this study was to extend this validation program by examining evidence based on relationships with other variables—that is, the degree to which the relationships of the test scores with other variables are consistent with the construct underlying the proposed interpretation of the scores [44,45].

Specifically, this study addressed the following research questions:

1. Do scores on the VR-based technical aptitude test and the GBA demonstrate convergent and discriminant associations with established psychometric measures of related and unrelated constructs?
2. Are test scores associated with residency performance evaluations, thereby providing evidence of test-criterion relationships?
3. Do VR technical aptitude scores differentiate between interns, residents, and expert surgeons (expert-novice differences)?
4. Given previously observed gender differences in test scores, is there evidence of differential prediction with respect to gender?

To address these questions, we conducted a multiphase validation study involving interns, residents, and expert surgeons, using correlational and group comparison analyses within a validity framework. We hypothesized that (1) both assessments would exhibit construct-consistent patterns of convergent and discriminant associations, with stronger relationships between each test and theoretically aligned constructs than with unrelated measures, (2) scores on both assessments would be positively associated with relevant dimensions of residency performance evaluations, (3) scores on the VR technical aptitude test would increase with surgical experience, and (4) although mean gender differences might be observed, the tests would not demonstrate evidence of differential prediction.

The findings are intended to inform stakeholders responsible for the selection of candidates for surgical training, including program directors, medical educators, and researchers seeking evidence-based tools for high-stakes decision-making.

Methods

Overview of the Validation Study Design

This study had 3 main phases, each designed primarily to collect evidence for one set or type of relationships with other variables: relationships with other tests measuring similar and different constructs (convergent and discriminant evidence), test-criterion relationships, and expert-novice differences [44,45,50]. Additional analyses, including incremental evidence of validity and fairness (differential prediction), were conducted within this overarching framework of relationships with other variables. Table 1 provides an overview of the validity evidence collected and the analytic approaches used to evaluate each source.

XSL•FO
RenderX

Table . Summary of evidence collected to evaluate validity (relationships with other variables) and fairness.

Type of evidence	Definition	Relevant study phases[a]	Evidence collected in the study
Convergent and discriminant evidence	Evidence based on assessing the relationship between the selection tests and other instruments. Convergent evidence for validity refers to the relationships between test scores and other measures intended to assess the same or similar constructs, while discriminant evidence for validity refers to the relationships between test scores and other measures intended to assess different constructs	Phase 1	Associations between the novel selection tests and 4 established measures of dexterity and coordination, visuospatial ability, intelligence, and personality[b]. The relationship between the VR[c]-based technical aptitude test and both dexterity and coordination and visuospatial ability (the PPT[d] and MRT[e]), and between the GBA[f] and both intelligence and personality (the RAPM[g] and the mini-IPIP[h]), was considered convergent evidence; the relationship between the VR-based technical aptitude test and intelligence and personality, and between the GBA and dexterity and coordination and visuospatial ability, was considered discriminant evidence.
Test-criterion relationships	Evidence based on assessing the relationship between the selection tests and relevant external performance criteria, reflecting the extent to which test scores meaningfully predict measures of real-world performance	Phase 2	Associations between the novel selection tests and 16 residency performance criteria (1 technical skills criterion and 15 nontechnical skills criteria) were examined in a concurrent design. The criteria were assessed using structured evaluations completed independently by 3 to 4 supervising surgeons per resident. The technical skills criterion was used to evaluate the VR-based technical aptitude test, while the 15 nontechnical criteria were used to evaluate the GBA.
Incremental evidence	Evidence based on examining whether a test explains unique variance in a relevant outcome beyond that accounted for by other measures, thereby supporting its incremental contribution within the assessment framework	Phase 2	Associations were examined to determine whether each selection test predicted its theoretically relevant performance criterion above and beyond the other test. The incremental contribution of the VR-based technical aptitude test was evaluated for the technical skills criterion beyond the GBA, and the incremental contribution of the GBA was evaluated for the aggregated nontechnical performance criteria beyond the VR-based technical aptitude test.
Relationship with training level (expert-novice differences)	Evidence based on examining the relationship between selection test scores and theoretically relevant variables (eg, training level for the technical aptitude test), reflecting the extent to which these associations are consistent with the construct underlying the proposed interpretation of the scores	All phases	Performance on the VR-based technical aptitude test was compared across interns, residents, and expert surgeons to evaluate whether scores increased with level of surgical expertise. Expert-novice differences were not evaluated for the GBA.

XSL•FO
RenderX

Type of evidence	Definition	Relevant study phases[a]	Evidence collected in the study
Fairness (differential prediction)	Evidence based on examining whether test scores predict relevant performance criteria equivalently across demographic groups (eg, gender)	Phase 2	For each selection test, regression models were estimated to determine whether gender significantly moderated the relationship between test scores and the relevant performance criterion (technical skills for the VR-based technical aptitude test and the aggregated nontechnical performance criteria for the GBA). Sensitivity analyses were conducted, including prior simulator experience and prior video game experience as covariates, to assess the robustness of the findings.

[a]The study included 3 phases: Phase 1 (administration to interns), Phase 2 (administration to residents), and Phase 3 (administration to expert surgeons).

[b]For details on these tests and their sources, see under "Phase 1: Administration to Interns," below, and Table S1 in the Multimedia Appendix 1.

[c]VR: virtual reality.

[d]PPT: Purdue Pegboard Test.

[e]MRT: Mental Rotation Test.

[f]GBA: game-based assessment.

[g]RAPM: Raven Advanced Progressive Matrices.

[h]mini-IPIP: short version of the International Personality Item Pool.

In Phase 1, the 2 novel selection tests (ie, both the VR technical aptitude test and the GBA of cognitive abilities and personality characteristics) were administered to a sample of interns, along with 4 established psychometric instruments commonly used for assessing dexterity and coordination, visuospatial ability, intelligence, and personality. Based on the data collected in this phase, we evaluated evidence for the relationships between scores on the 2 novel selection tests and other tests measuring similar and different constructs (convergent and discriminant evidence). In Phase 2, the 2 novel tests were administered to a sample of residents, and their performance in residency was assessed using an evaluation form filled in by their supervisors in a concurrent design (obtaining test scores and criterion information simultaneously). This phase allowed us to assess evidence for test-criterion relationships. In addition, the data from this phase were used to assess the incremental contribution of each test and evaluate evidence for the possibility of gender bias in the tests. Finally, in Phase 3, the VR technical aptitude test was administered to a sample of senior surgeons. Their scores were then compared with those of the interns and residents from the previous phases in order to test for expert-novice differences (ie, the relationship between the technical aptitude scores and test-takers' training level). Expert-novice differences were assessed only for the VR technical aptitude test and not for the GBA of cognitive abilities and personality characteristics because evidence regarding expert-novice differences is relevant only when the construct being measured is hypothesized to be related to training status and, therefore, should differ between the groups [49,51,52]. Unlike the technical aptitude test, the tasks used in the GBA do not resemble real-life surgical tasks or challenges. Therefore, expert-novice differences are not relevant to the GBA.

Formal a priori power calculations to determine sample sizes for the 3 phases were not performed due to the feasibility-based recruitment framework and the fixed pool of eligible participants. Based on practical and logistical considerations, we aimed to recruit a minimum of 60 participants per training level while including all available individuals. Given the final sample (n=216), the study was adequately powered to detect small-medium effect sizes for between-group comparisons, within-group correlations, and regression analyses.

Ethical Considerations

This study was approved by the ethics committee of the Hebrew University of Jerusalem (approval no. 13032023) and was conducted in accordance with the principles of the Declaration of Helsinki. All participants provided informed consent prior to participation. Participant data were stored using unique anonymized identifiers; the key linking these identifiers to real identities was kept in a password-protected file stored offline, ensuring that no identifying information was accessible online. Interns received US $70 and residents US $50 as monetary compensation. Both groups were also provided with feedback on their performance in the test, presented as percentile rankings relative to their respective samples.

The Selection Tests

This section briefly describes the 2 novel tests examined in this research.

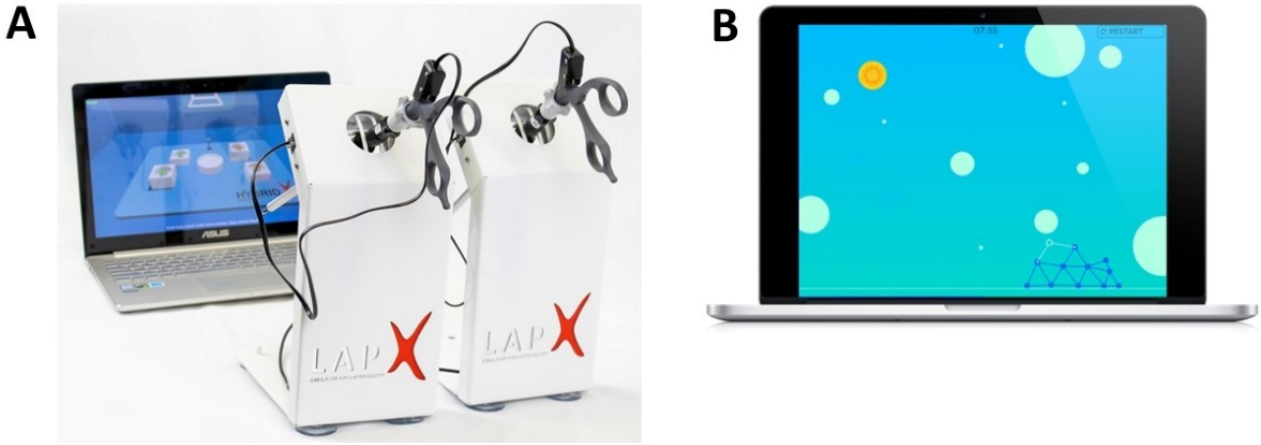
The VR Technical Aptitude Test

The VR technical aptitude test was developed and designed to assess technical aptitude among candidates for surgical training with no former surgical experience or knowledge [46]. The test consists of 10 tasks designed to assess the psychomotor and perceptual abilities (coordination, ambidexterity, movement precision, visuospatial ability, and depth perception) needed to perform key tasks relevant to minimally invasive surgery (MIS), including grasping and transferring objects, cutting with scissors, scope handling, and suturing with a needle [40,53]. The test is performed on the Lap-X-VR laparoscopic simulator (Medical-X;

XSL•FO
RenderX

Figure 1A) [48] and takes about 50 minutes to complete. The Lap-X-VR is a nonimmersive, screen-based laparoscopic VR simulator. Examinees manipulate real laparoscopic instruments connected to a computer system displaying a 3D virtual operative field on a standard monitor. Unlike fully immersive VR systems, it does not use head-mounted displays, stereoscopic visualization, or spatial head tracking. The simulated operative field is displayed on a conventional monitor, and immersion is therefore limited to a screen-based interaction.

Figure 1. Illustrations of the 2 selection tests evaluated in the study. (A) The technical aptitude test was performed on the LAP-X-VR laparoscopic simulator. (B) The game-based assessment of cognitive abilities and personality characteristics. For illustration, only one task of each test is presented.



Although primarily designed as a VR-based surgical simulator, the test also incorporates several gamified elements intended to structure performance and enhance engagement, including clearly defined task goals, real-time performance feedback, scoring metrics, and progressive task completion requirements. As such, it reflects the integration of VR technology and gamification principles within a single assessment platform.

Test scores are calculated based on the aggregation of performance data recorded by the simulator in each task for the following parameters: success rate (%), time (s), number of mistakes, path length (cm), and, where relevant, percent of time within scope (%). Final test scores are presented on a scale with a mean of 100 and an SD of 20, with higher scores indicating better technical aptitude. This test has been shown to have high reliability (Cronbach α=0.83) and discrimination between examinees (mean task discrimination 0.5, SD 0.1) [46]. For a full description of the test development and the specific tasks included, refer to Gazit et al [46].

GBA of Cognitive Abilities and Personality Characteristics

The GBA test [47] was developed and designed to assess 14 cognitive abilities and personality characteristics relevant to surgical training [7]: planning, problem-solving, ingenuity, goal orientation, self-reflection, endurance, analytical thinking, learning ability, flexibility, concentration, conformity, multitasking, working memory, and precision. The test was developed in cooperation with Benchmark.games Ltd, a company that produces GBAs for organizational hiring and recruitment. The test is administered on a standard computer and takes about 60 minutes to complete (Figure 1B). The test consists of 3 video games, designed by a team of psychometricians and psychologists. Each video game was designed to assess a set of specific competencies relevant to surgical residents. In each game, all actions of examinees (eg, mouse movements and key presses) are recorded and logged. These raw data are then transformed into higher-level variables that describe a set of meaningful measurements (eg, time to first response, time between actions, accuracy, number of steps, and learning curve). Then, competency scores are calculated by aggregating the relevant variables, with higher weight given to variables characterized by larger variance between candidates. Competency scores are presented on a scale of 1 - 10. Final test scores are calculated by averaging the 14 individual competency scores. As with the VR technical aptitude test described above, total scores are then presented on a scale with a mean of 100 and an SD of 20. This test has been shown to have high reliability (Cronbach α=0.76) and discrimination between examinees (mean game discrimination 0.4, SD 0.2) [47]. For a comprehensive description of the test development process, the structure and objectives of the 3 games, the specific cognitive and personality competencies assessed, and the theoretical and empirical rationale underlying their design, see Gazit et al [47].

Note that, consistent with established standards for high-stakes assessment [45] and common practice in commercially developed game-based selection tools [24,27], both Gazit et al [47] and this paper report only the tool's general scoring framework and psychometric properties. Detailed operational definitions of specific behavioral indicators and exact weighting coefficients are proprietary to Benchmark.games Ltd and protected under contractual confidentiality agreements, precluding disclosure of precise scoring parameters. Maintaining confidentiality of specific scoring rules also serves test security purposes in high-stakes contexts. In simulation-based assessments, validity depends partly on candidates responding naturally to task demands rather than optimizing behavior toward explicitly known scoring indicators. Disclosure of detailed indicator-competency mappings could facilitate superficial behavioral adjustments without corresponding changes in the underlying competencies, thereby introducing construct-irrelevant variance.

XSL•FO
RenderX

Phase 1: Administration to Interns

Participants

Seventy-six medical interns from 2 hospitals in Israel participated in Phase 1. To recruit participants, invitations were posted in relevant Facebook (Meta Platforms, Inc) and WhatsApp (WhatsApp LLC) groups along with the contact information of the research coordinator. To improve the likelihood that the sample would represent the population of candidates for surgical training, only interns who were interested in pursuing a surgical career were eligible to participate in the study. Recruitment continued until we had at least 60 participants.

Procedure

Each intern was invited to 2 sessions. During the first session, participants completed the 2 novel selection tests. The order in which the tests were administered was randomized across participants. In the second session, participants completed 4 established psychometric instruments measuring dexterity and coordination, visual-spatial ability, intelligence, and personality (see below). Again, the order in which the tests were administered was randomized across participants. The first session lasted approximately 2 hours, and the second about 75 minutes.

The 4 established psychometric instruments were as follows:

1. The Purdue Pegboard Test (PPT) [54,55], a general measure of manual dexterity and bimanual coordination. In the original test, scores are calculated separately for each of the 4 tasks included in the test. In this study, we also calculated total test scores by transforming the 4 raw task scores into z-scores (ie, distributions with a mean of 0 and an SD of 1) and then averaging the z-score values. This was done in order to simplify the interpretation of the relationship between scores on this test and on the selection tests evaluated in this study.
2. The Mental Rotation Test (MRT) [56,57], a general measure of depth perception and visuospatial ability. The original version of this test was developed by Vandenberg and Kuse [56]. In this study, we used a redrawn version of the MRT (the MRT-A) created by Peters et al [57].
3. The Raven's Advanced Progressive Matrices (RAPM) [58], a nonverbal test used to measure general intelligence and abstract reasoning.
4. The short version of the International Personality Item Pool (mini-IPIP) [59], which measures the personality traits captured in the well-established five-factor model (extraversion, agreeableness, conscientiousness, neuroticism, and openness).

These are all commonly used, well-studied tests that have been validated for assessment of competencies similar to the competencies measured in the 2 novel tests, making them appropriate for evaluating convergent and discriminant evidence of validity. Full descriptions of the 4 tests can be found in Table S1 in the Multimedia Appendix 1.

At the end of the second session, participants provided general demographic information (age, gender, and dominant hand). Participants also reported their previous experience using laparoscopic simulators and playing video games, both on 5-point Likert scales (1=no experience, 5=very extensive experience).

Phase 2: Administration to Residents

Participants

Seventy-five residents from 2 hospitals in Israel participated in Phase 2. To recruit participants, emails and WhatsApp messages were sent to eligible residents asking for their participation. Email addresses and phone numbers of potential participants were obtained from hospital websites or from the database of the Israeli medical association. We contacted only residents in 5 surgical fields characterized by extensive use of MIS techniques: general surgery, gynecology, orthopedics, otorhinolaryngology and head and neck surgery, and urology. Only residents in their second year of residency or above were eligible to participate in the study. We set this exclusion criterion to ensure that participating residents would have sufficient experience in residency that (1) their skills and characteristics would be distinct from those of the interns in Phase 1, and (2) their performance could be reliably evaluated by their supervisors. Recruitment continued until we had at least 60 participants, with at least 5 from each of the 5 surgical fields mentioned above. Participants in this phase had completed on average 3.5 (SD 1.8) years of specialist surgical training.

Procedure

Each resident was invited to one 2-hour session. Participants first completed the 2 novel tests. As in Phase 1, the order in which the tests were administered was randomized across participants. They then provided general demographic information (age, gender, dominant hand, and surgical specialty), and reported their previous experience using laparoscopic simulators and playing video games as in Phase 1.

We assessed the residents' performance in their training program through structured performance evaluations completed by participants' supervisors (eg, their department directors, training program directors, or other senior expert surgeons who worked closely and directly with the residents). The evaluations assessed residents on 16 dimensions, including the competencies assessed by the 2 novel selection tests (see below). Each resident was evaluated by 3 to 4 supervisors, depending on the specialty. Supervisors filled in all evaluation forms independently, without knowing the ratings of other evaluators, and were blinded to participants' scores in the 2 selection tests. Participants were informed that we would contact their departments in order to collect their performance evaluations, and we asked them to sign a consent form for this purpose. Four residents declined to sign the form, and so evaluations were used for only 71 of the 75 residents in the sample.

The evaluation form was designed using online survey software (Qualtrics; Qualtrics International Inc) and was sent to supervisors by email. Each supervisor was asked to read an introduction and presentation of the study's aims before completing the evaluation forms, one for each resident in their department. The evaluations assessed residents' performance in 16 dimensions: (1) medical knowledge (relative to the

XSL•FO
RenderX

resident's stage in the specific training program), (2) technical skills (again, relative to the resident's stage in the program), (3) communication with patients and their families, (4) communication with medical staff and teamwork, (5) integrity, (6) diligence, (7) learning ability, (8) decision-making and problem-solving, (9) self-criticism and ability to learn from mistakes, (10) thoroughness, (11) organization and planning, (12) physical and mental endurance, (13) stress tolerance, (14) creativity and cognitive flexibility, (15) motivation, and (16) general assessment of performance in the residency program. Evaluations were provided on a 5-point Likert scale, where 1=very low and 5=very high. Evaluators were required to provide ratings for each of the 16 dimensions.

The evaluation form was developed based on existing forms used in the participating departments to assess residents' performance and progress. As such, the form included not only competencies assessed by the 2 selection tests examined in this study (eg, technical skills, decision-making and problem-solving, and learning ability), but also competencies not assessed by the 2 selection tests examined here (eg, communication with patients and their families, communication with medical staff and teamwork, and integrity). This decision was made for 2 reasons. First, it allows us to examine whether correlations between the selection tests and competencies that are expected to be related to them are greater than correlations between the selection tests and competencies that are not expected to be related. Second, it enables us to consider residents' scores on the selection tests in light of a full and comprehensive assessment of their performance in the residency program.

Phase 3: Administration to Expert Surgeons

Participants

Sixty-five senior surgeons from 2 hospitals in Israel participated in Phase 3. To recruit participants, emails and WhatsApp messages were sent to eligible surgeons asking for their participation. Email addresses and phone numbers of potential participants were obtained from hospital websites or from the Israeli Medical Association database. Eligible surgeons were specialists in 1 of the 5 surgical fields from which we drew the resident sample in Phase 2, selected for their extensive use of MIS techniques (general surgery, gynecology, orthopedics, otorhinolaryngology and head and neck surgery, and urology), and had at least 10 years of experience with MIS. Recruitment continued until we had at least 60 participants, with at least 5 from each of the 5 surgical fields mentioned above. Participating expert surgeons had, on average, 16.7 (SD 9.6) years of experience with MIS. The expert surgeons were not compensated for their participation.

Procedure

Each expert was invited to 1 session lasting about 50 minutes. At the session, participants first completed the VR technical aptitude test. They then provided demographic information (age, gender, dominant hand, and surgical specialty), and reported their previous experience using laparoscopic simulators and playing video games as in the previous phases.

Validation and Analyses

Following the contemporary framework of validity [44,45], we collected evidence on relationships between the 2 novel tests and other variables. In Phase 1, we collected evidence on the relationship between the selection tests and 4 established psychometric instruments measuring similar and different constructs (convergent and discriminant evidence). Toward this end, we used Pearson correlations to assess the associations between the novel selection test scores and scores on the 4 established psychometric instruments (the PPT, MRT, RAPM, and mini-IPIP). Looking first at the VR technical aptitude test, to evaluate convergent evidence for validity, we computed the correlations between interns' scores on this test and their scores on 2 established psychometric instruments assessing similar competencies—the PPT and the MRT. Those correlations are expected to be relatively strong. We then evaluated discriminant evidence for the VR test by computing the correlations between scores on this test and on 2 established psychometric instruments assessing different competencies—the RAPM and the mini-IPIP. Those correlations are expected to be relatively weak. Turning to the GBA, this time, correlations with the RAPM and the mini-IPIP were used to assess convergent evidence, and correlations with the PPT and the MRT were used to assess discriminant evidence. Finally, correlations between scores on the 2 novel tests were calculated as another source of discriminant evidence.

In Phase 2, we collected evidence for test-criterion relationships based on correlations between the novel selection test scores and measures of performance on each relevant criterion extracted from residency performance evaluations. To do so, we first averaged, for each participant, the scores for each dimension in the evaluation form provided by all supervisors who evaluated that resident (at least 3 supervisors for each participant, as described above). This process resulted in a set of 16 criterion scores for each participant. As another preliminary step, we calculated the average of all dimension scores, excluding the technical skills and general performance dimensions, to produce a mean performance evaluation without technical skills. We then turned to our main analysis—namely, examining the correlations between the 2 selection test scores and the 17 criterion assessments (the 16 dimension scores and the mean performance evaluation without technical skills). We used Pearson correlations to assess the relationships between participants' scores on the selection tests and their criterion scores. Due to the hierarchical structure of the data (in that residents are nested in different surgical fields and departments), the correlations were calculated separately for each surgical field, and then their weighted mean was calculated across the surgical fields. The correlations were conducted to evaluate prespecified, theory-driven hypotheses regarding the relationships between each selection test and specific performance criteria. Because each analysis addressed a distinct, theory-driven hypothesis and the outcomes were conceptually related rather than independent, formal familywise error correction was not applied [60,61]. Results were interpreted in light of the overall pattern of associations rather than on the basis of isolated statistically significant findings. We expected relatively high correlations between each selection test and its

XSL•FO
RenderX

relevant criteria (ie, between technical aptitude test scores and the technical skills criterion, and between GBA scores and all other criterion scores). In addition, we expected that the GBA would be correlated more strongly with those criteria capturing competencies similar to those assessed by the GBA (eg, learning ability and decision-making) than those capturing competencies not assessed by the GBA (eg, communication with patients and their families and communication with medical staff and teamwork).

Based on the data collected in Phase 2, we also examined incremental evidence of validity. To this end, we estimated nested multilevel (random-intercept) models to assess whether each selection test explained unique variance in its relevant residency performance criterion beyond the other test. Residents were modeled as nested within surgical fields. For each criterion, we compared models including a single predictor with models including both predictors using likelihood ratio tests. Changes in model fit and marginal R^2 (ie, the variance explained by the fixed effects only) were examined to evaluate the incremental contribution of each selection test beyond the other test. We expected that each test would demonstrate a unique and statistically significant contribution to its theoretically aligned performance criterion beyond the alternative selection test.

Following Phase 3, we used the VR-based technical aptitude test scores obtained in all 3 phases to examine evidence based on the relationship between test performance and training level (expert-novice differences). For this purpose, we analyzed the VR technical aptitude test scores of expert surgeons obtained in Phase 3 alongside the comparable data obtained from interns and residents in Phases 1 and 2, using a 1-way between-subjects ANOVA. We hypothesized that there would be a positive correlation between training level and technical aptitude test scores, such that groups with more advanced training would have higher scores (ie, experts would score higher than residents and residents higher than interns).

Finally, based on the data of residents in Phase 2, we conducted a fairness analysis to determine whether the selection test discriminates on the basis of gender. Group bias is considered present if the predicted values of the criterion based on test scores differ between the examined groups (ie, if the same selection test score predicts different criterion scores for individuals from the different groups of interest). Toward this end, we conducted a multiple regression analysis for each selection test which included the relevant criterion score (technical skills for the technical aptitude tests and the mean performance evaluation without technical skills for the GBA) as a dependent variable, and test scores and gender as predictors [62,63]. Because evaluators differed between surgical specialties, criterion scores were standardized within each specialty to minimize potential rater effects prior to conducting the regression analyses. Differential prediction was evaluated by testing for slope and intercept differences between the gender groups, and by testing for systematic deviations from the common regression line (evidence for bias increases as the sum of deviations for a specific gender group grows larger). We hypothesized that, although mean gender differences in test scores might be observed, the relationship between test scores and performance criteria would not differ by gender, indicating no evidence of differential prediction.

To examine whether differential prior exposure influenced fairness-related inferences, we also conducted sensitivity analyses in which the fairness regression models were reestimated, including prior laparoscopic simulator experience and prior video game experience as covariates. Based on prior findings indicating modest associations between simulator and/or video game experience and performance on the VR technical aptitude test [46], and between video game experience and benchmark performance [47], the primary specification included simulator experience and video game experience as covariates in the VR technical aptitude model, and video game experience as a covariate in the GBA model.

All tests were 2-sided, and the level of statistical significance was set to .05. Statistical analyses were performed using R (version 4.3.2; R Core Team; R Foundation for Statistical Computing).

Results

Participant Characteristics

The demographic characteristics of the 76 interns, 75 residents, and 65 expert surgeons who participated in the study are presented in Table 2.

XSL•FO
RenderX

Table . Demographic characteristics of the participants.

Characteristic	Interns (n=76)	Residents (n=75)	Experts (n=65)
Age (years), mean (SD)	26.8 (4.0)	34.2 (4.8)	54.5 (9.8)
Sex (female), n (%)	31 (41)	23 (31)	16 (25)
Left dominant hand, n (%)	6 (8)	5 (7)	8 (12)
Surgical specialty, n (%)			
General surgery	—[a]	19 (25)	18 (28)
Gynecology	—	17 (23)	12 (18)
Orthopedics	—	16 (21)	16 (25)
Otorhinolaryngology and head and neck surgery	—	13 (17)	12 (18)
Urology	—	10 (13)	7 (11)
Experience with MIS[b] simulators, n (%)			
No experience	59 (78)	40 (53)	7 (11)
Little experience	16 (21)	30 (40)	19 (29)
Moderate experience	1 (1)	3 (4)	27 (42)
Considerable experience	0 (0)	2 (3)	11 (17)
Very extensive experience	0 (0)	0 (0)	1 (2)
Experience with video games, n (%)			
No experience	10 (13)	18 (24)	22 (34)
Little experience	19 (25)	30 (40)	31 (48)
Moderate experience	26 (34)	18 (24)	8 (12)
Considerable experience	14 (18)	6 (8)	3 (5)
Very extensive experience	7 (9)	3 (4)	1 (2)

[a]Not applicable.

[b]MIS: minimally invasive surgery.

In what follows, we present the results according to their relevance for the different types of validity evidence and for fairness.

Convergent and Discriminant Evidence for Validity: Phase 1

Means and SDs of scores on the 4 established instruments are shown in Table S2 in the Multimedia Appendix 1. Correlations between scores on the 2 novel selection tests and the 4 established instruments are presented in Table 3.

XSL•FO
RenderX

Table . Correlations between scores on the 2 novel selection tests and the 4 established psychometric instruments (Phase 1).

Established instrument	VR[a] technical aptitude test			GBA[b] of cognitive abilities and personality		
	r	95% CI	*P* value	*r*	95% CI	*P* value
PPT[c]	0.33	0.11 to 0.52	.003	0.12	−0.11 to 0.33	.30
MRT[d]	0.59	0.42 to 0.72	<.001	0.38	0.17 to 0.56	.001
RAPM[e]	0.32	0.10 to 0.51	.005	0.54	0.36 to 0.68	<.001
mini-IPIP[f] scale						
Extraversion	−0.08	−0.30 to 0.15	.50	0.02	−0.21 to 0.25	.86
Agreeableness	−0.13	−0.34 to 0.10	.26	−0.12	−0.33 to 0.11	.30
Conscientiousness	0.20	−0.03 to 0.41	.08	0.29	0.07 to 0.48	.01
Neuroticism	−0.21	−0.42 to 0.02	.07	−0.25	−0.45 to −0.03	.03
Openness	0.01	−0.22 to 0.24	.93	0.14	−0.09 to 0.35	.23

[a]VR: virtual reality.

[b]GBA: game-based assessment.

[c]PPT: Purdue Pegboard Test.

[d]MRT: Mental Rotation Test.

[e]RAPM: Raven Advanced Progressive Matrices.

[f]mini-IPIP: short version of the International Personality Item Pool.

As expected, scores on the PPT were significantly correlated with scores on the technical aptitude test, but not with scores on the GBA. Scores on the MRT were significantly correlated with scores on both the technical aptitude test and the GBA, although the correlation with the former was considerably stronger (0.59 vs 0.38). For the RAPM and the mini-IPIP, we obtained an inverse pattern, again largely in keeping with our expectations. Specifically, scores on the RAPM were significantly correlated with scores on both the technical aptitude test and the GBA, but this time the correlation with the GBA was substantially stronger (0.54 vs 0.32). For the mini-IPIP, the picture was somewhat more complicated. While none of the 5 scales of the mini-IPIP were significantly correlated with scores on the technical aptitude test, the correlations with 2 of these scales, for conscientiousness and neuroticism, were marginally significant (conscientiousness: r_{74}=0.20, 95% CI −0.03to 0.41, *P*=.08); neuroticism: r_{74}=−0.21, 95% CI −0.42to 0.02, *P*=.07). However, scores for both of those scales correlated significantly with scores on the GBA. In addition, we found a moderately significant correlation between scores on the technical aptitude test and the GBA (r_{74}=0.37, 95% CI 0.16-0.55, *P*=.001).

Overall, the correlations between scores on the novel tests and established psychometric instruments measuring similar competencies were stronger than the correlations with established psychometric instruments measuring different competencies. Likewise, correlations between each of the established instruments and the relevant novel selection test measuring similar competencies were higher than those with the selection test measuring different competencies. Therefore, the convergent and discriminant evidence presented supports the validity of the 2 novel selection tests.

Evidence for Test-Criterion Relationship and Incremental Contribution: Phase 2

To evaluate the evidence for the test-criterion relationship, we first calculated for each resident 16 criterion scores, as described above. Means and SDs of the criterion scores are shown in Table S3 in the Multimedia Appendix 1. The intercorrelations among the 16 performance criteria ranged from 0.21 to 0.89 (mean *r* 0.55), indicating moderate shared variance across domains. As noted above, the number of raters evaluating each resident ranged from 3 to 4, depending on the surgical field. Therefore, to ensure the quality of the criterion scores, we first assessed interrater reliability using the intraclass correlation coefficient (ICC) reliability index. The ICC was calculated separately for each dimension and each surgical field, based on a mean rating (3≤k≤4) consistency 2-way mixed-effects model [64]. The ICC estimates and their 95% CIs are shown in Table S4 in the Multimedia Appendix 1. Overall, the ICC estimates were high, ranging from 0.46 to 0.96 with a median value of 0.78, indicating good interrater reliability of the criterion evaluations.

Then, we examined the correlations between the 2 selection test scores and the 17 criterion assessments (the 16 dimension scores and the mean performance evaluation without technical skills). As noted above, the correlations were first calculated separately for each surgical field (results for each specialty are shown in Table S5 in the Multimedia Appendix 1). We then calculated their weighted mean across the surgical fields. The mean correlations are presented in Table 4.

XSL•FO
RenderX

Table . Correlations between scores on the 2 selection tests and the criteria scores (Phase 2).

Criterion	VR[a] technical aptitude test			GBA[b] of cognitive abilities and personality		
	r	95% CI	*P* value	*r*	95% CI	*P* value
Medical knowledge	0.20	−0.03 to 0.41	.10	0.40	0.18 to 0.58	<.001
Technical skills	0.61	0.44 to 0.74	<.001	0.32	0.09 to 0.51	.007
Communication with patients and their families	0.05	−0.18 to 0.28	.68	0.23	0.00 to 0.44	.05
Communication with medical staff and teamwork	0.13	−0.11 to 0.35	.28	0.24	0.01 to 0.45	.04
Integrity	0.13	−0.11 to 0.35	.28	0.36	0.14 to 0.55	.002
Diligence	0.13	−0.11 to 0.35	.28	0.32	0.09 to 0.51	.007
Learning ability	0.15	−0.09 to 0.37	.21	0.41	0.19 to 0.59	<.001
Decision-making and problem-solving	0.28	0.05 to 0.48	.02	0.40	0.18 to 0.58	<.001
Self-criticism and ability to learn from mistakes	0.16	−0.08 to 0.38	.18	0.33	0.10 to 0.52	.005
Thoroughness	0.06	−0.17 to 0.29	.62	0.15	−0.09 to 0.37	.21
Organization and planning	−0.07	−0.30 to 0.17	.56	0.24	0.01 to 0.45	.04
Physical and mental endurance	0.24	0.01 to 0.45	.04	0.30	0.07 to 0.49	.01
Stress tolerance	0.30	0.07 to 0.49	.01	0.32	0.09 to 0.51	.007
Creativity and cognitive flexibility	0.27	0.04 to 0.47	.02	0.34	0.11 to 0.53	.004
Motivation	0.15	−0.09 to 0.37	.21	0.33	0.10 to 0.52	.005
General assessment	0.19	−0.04 to 0.40	.11	0.46	0.25 to 0.63	<.001
Mean performance evaluation without technical skills	0.17	−0.07 to 0.39	.16	0.43	0.21 to 0.60	<.001

[a]VR: virtual reality.

[b]GBA: game-based assessment.

As in Phase 1, as a secondary analysis, we calculated the correlation between residents' scores on the technical aptitude test and the GBA. This correlation was again found to be moderate and significant (r_{69}=0.40, 95% CI 0.18-0.58, *P*<.001).

As expected, scores on the technical aptitude test were strongly and significantly correlated with the technical skills criterion. In addition, significant (though weak) correlations were found between technical aptitude test scores and 3 criteria: endurance, stress tolerance, and creativity and cognitive flexibility. In contrast, GBA scores were significantly correlated with all criteria (including the technical skills criterion) except thoroughness, with the strongest correlations emerging for medical knowledge, learning ability, decision-making and problem-solving, and general assessment of performance.

Given the moderate intercorrelation between the 2 selection tests and their overlapping associations with several performance criteria, incremental validity analyses were conducted to examine whether each test accounted for unique variance in its theoretically relevant performance criterion above and beyond the other assessment. Specifically, we examined whether the VR-based technical aptitude test uniquely predicted technical skills evaluations, and whether the GBA uniquely predicted overall performance excluding technical skills. Nested multilevel (random-intercept) models were compared using likelihood ratio tests. For technical skills, adding the VR-based technical aptitude test significantly improved model fit (χ^2_1=21.18, *P*<.001). In the full model, the VR-based test remained a significant predictor while controlling for GBA scores (β=0.03, SE=0.01, 95% CI 0.01-0.05), with marginal R^2 increasing from 0.119 to 0.363. Conversely, for performance excluding technical skills, adding GBA scores significantly improved model fit (χ^2_1=8.61, *P*=.003). In the full model, the GBA remained a significant predictor while controlling for VR-based technical aptitude test scores (β=0.02, SE=0.01, 95% CI 0.01-0.04), with marginal R^2 increasing from 0.091 to 0.183. These findings

XSL•FO
RenderX

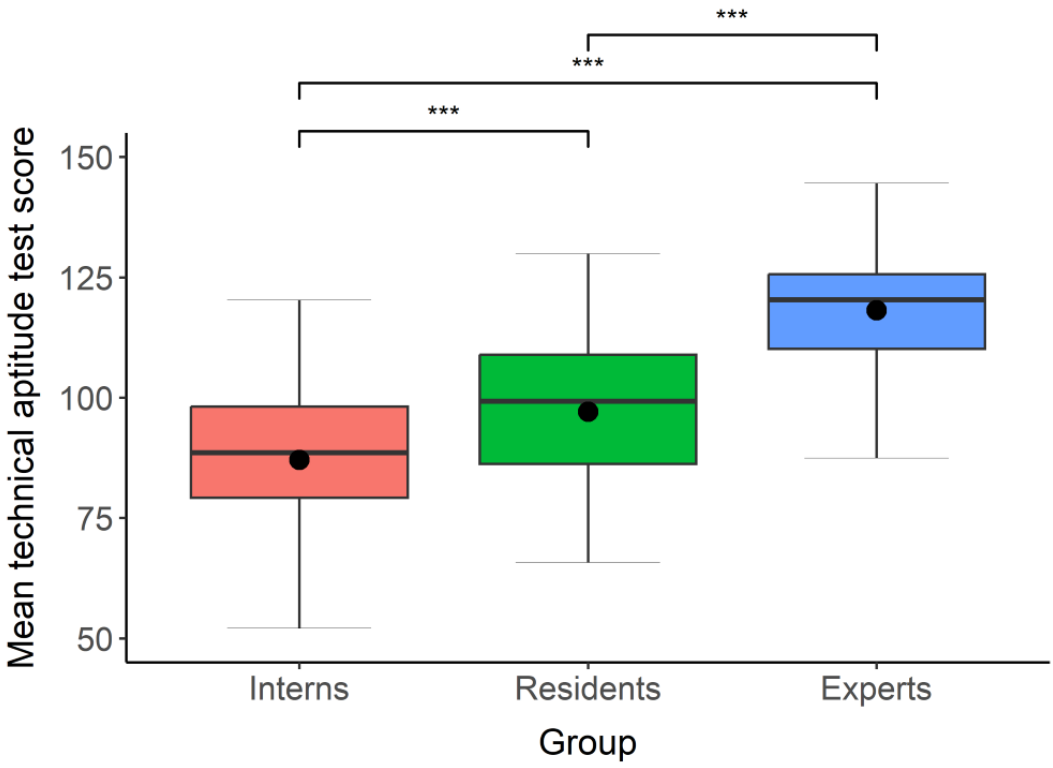
indicate that each assessment contributes unique variance to its domain-relevant performance outcome.

Overall, the correlation between scores on the technical aptitude test and its theoretically relevant criterion (ie, technical skills) was substantially stronger than that between GBA scores and the same criterion. Conversely, correlations between GBA scores and performance criteria excluding technical skills were generally stronger than those observed for the technical aptitude test, particularly for the general assessment of performance and the mean performance evaluation without technical skills. At the same time, both tests exhibited significant associations with some nontarget criteria, indicating that the constructs assessed by the 2 tests are partially overlapping rather than fully distinct. The incremental validity analyses further clarified this pattern: the VR-based technical aptitude test uniquely predicted technical skills above and beyond GBA scores, while the GBA uniquely predicted broader nontechnical performance beyond the VR-based test. Together, these findings suggest that the 2 assessments are related but not redundant, and that each contributes distinct and complementary variance to domain-relevant performance outcomes. Accordingly, the pattern of test-criterion relationships contributes to the accumulating evidence regarding the construct validity of both selection tests within Messick's framework.

Evidence for Relationship With Training Level: Phases 1, 2, and 3

We compared the technical aptitude scores of participating interns, residents, and experts via a 1-way between-subjects ANOVA, followed by post hoc comparisons using the Tukey honestly significant difference test. The mean scores of the 3 groups are presented in Figure 2. As expected, we found statistically significant differences between the groups in technical aptitude test scores ($F_{2,211}$=72.1, P<.001, η2=0.41). In post hoc comparisons using the Tukey honestly significant difference test, the mean score of the expert surgeons (mean 118.2, SD 13.1) was significantly higher than the mean scores both of residents (mean 97.1, SD 16.1, P<.001, difference 95% CI 14.8-27.2) and of interns (mean 87.1, SD 30.9, P<.001, difference 95% CI 24.9-37.3). In addition, the residents' mean score was significantly higher than that of the interns (P<.001, difference 95% CI 4.1-16.0). We also separately compared the scores of the 3 groups for each of the performance parameters assessed by the simulator (success rate, time, number of mistakes, path length, and percent of time within scope; refer to Table S6 and Figure S1 in the Multimedia Appendix 1). The pattern of results obtained when examining each parameter separately was similar to the pattern obtained for the final technical aptitude test scores, suggesting that group differences exist for all performance parameters.

Figure 2. Scores of interns, residents, and expert surgeons on the virtual reality technical aptitude test. Higher scores indicate greater technical aptitude. Within each box, the horizontal bar indicates the median, the circle indicates the mean, and the lower and upper boundaries indicate the first and third quartiles. The vertical lines outside the boxes (whiskers) indicate values within 1.5x the IQR from the upper to lower quartile (or the minimum and the maximum if within 1.5x the IQR of the quartiles). ***, P<.001.



XSL•FO
RenderX

Finally, we also calculated correlations between technical aptitude test scores and more specific measures of surgical experience: number of years in surgical training among residents, and number of years of experience with MIS among the expert surgeons. Both correlations were significant (residents: r_{73}=0.26, 95% CI 0.03-0.47, P=.02; experts: r_{63}=0.50, 95% CI 0.29-0.67, P<.001), providing further support for the relationship with training level.

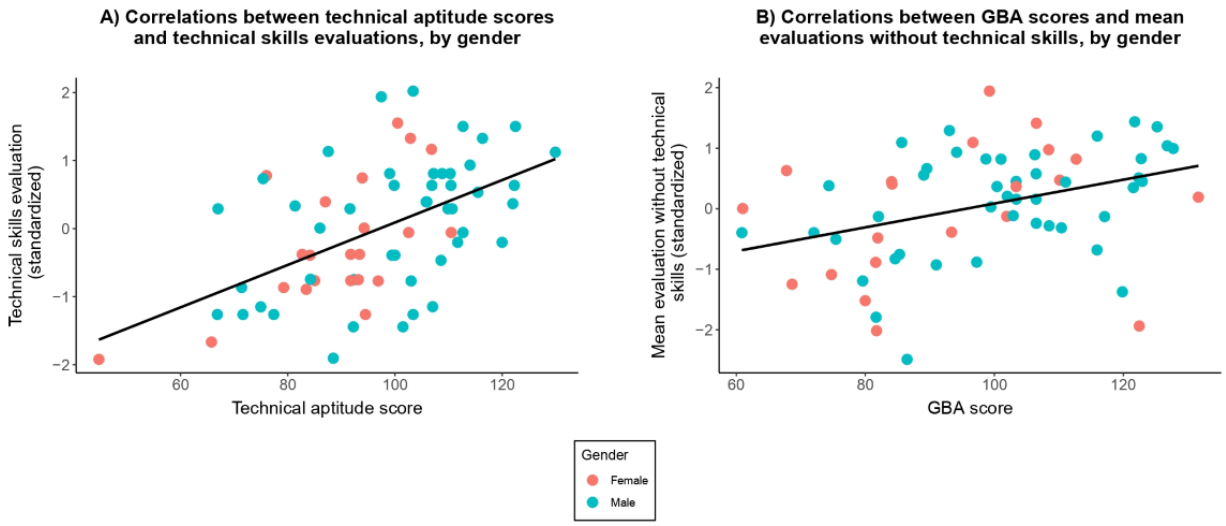
Gender Bias Analyses: Phase 2

Based on the data obtained from residents in Phase 2, we conducted a fairness analysis to assess the possibility of differential prediction with respect to gender. First, we compared the scores obtained separately by males and females in the selection tests and in the relevant evaluation rating criteria (technical skills evaluation ratings and mean evaluation ratings without technical skills). We found that males scored significantly higher than females in both selection tests (technical aptitude test scores: mean difference 10.7, SD 15.3, 95% CI 2.75 - 18.65, t_{69}=2.67, P=.009, Cohen d=0.7; GBA scores: mean difference 9.0, SD 18.0, 95% CI 0.16 - 17.84, t_{69}=2.03, P=0.046, Cohen d=0.5). In the relevant evaluation ratings (the criteria), we found a significant difference between males and females in the technical skills evaluation ratings (mean difference 0.5, SD 0.6, 95% CI 0.11 - 0.90, t_{69}= 2.27, P=.03, Cohen d=0.8), but only a marginally significant difference in mean evaluation ratings without technical skills (mean difference 0.2, SD 0.4, 95% CI −0.02to 0.50, t_{69}=1.87, P=.07, Cohen d=0.5).

Next, we conducted 2 multiple regression analyses. The first of these included technical aptitude test scores (a continuous variable), gender (a dummy variable), and their interaction as predictors, and the technical skills criterion as the dependent variable. The second included GBA scores (a continuous variable), gender (a dummy variable), and their interaction as predictors, and the mean evaluation ratings without the technical skills criterion as the dependent variable (refer to Table S7 in the Multimedia Appendix 1 for the regression statistics). In both regression analyses, the correlation between the selection test and the criterion was significant (in the first regression, t_{64}=3.6, P<.001; in the second regression, t_{64}=2.70, P=.009). However, the effects of gender (in the first regression: t_{64}=0.25, P=.80; in the second regression: t_{64}=−0.006, P=.99) and of the test-gender interaction (in the first regression: t_{64}=0.76, P=.45; in the second regression: t_{64}=−0.44, P=.66) were not significant. Prior simulator and video game experience were modestly associated with relevant selection test scores (r=0.11-0.16; Table S8 in the Multimedia Appendix 1). Sensitivity analyses incorporating prior simulator experience and prior video game experience (technical aptitude test model) and prior video game experience (GBA model) as covariates yielded substantively similar conclusions, meaning that the inclusion of differential prior exposure did not materially alter the pattern of regression coefficients or the absence of significant gender-by-test interactions (results are provided in Table S9 in the Multimedia Appendix 1). In other words, the predictive relationship between test scores and performance criteria was comparable for men and women even after accounting for prior simulator and gaming experience. Thus, the regression analyses found no evidence for gender bias.

As a final step, we examined the deviations of female and male residents from the common regression lines (Figure 3). The sum of the deviations for both genders in both regression models was close to zero (since for both genders, negative and positive deviations largely canceled each other out). Thus, this analysis also suggests that the prediction is not biased toward one of the genders. Overall, these findings do not support the existence of bias in either of the 2 selection tests.

Figure 3. Scatter plots of the correlations between scores on the selection tests and the evaluation criteria by gender. The black lines indicate the common regression lines for both genders. Observations above the regression lines represent positive deviations from the regression lines, and observations below the regression lines represent negative deviations. GBA: game-based assessment.



XSL•FO
RenderX

Discussion

Principal Findings

This study examined evidence based on relationships with other variables for validation of 2 simulation-based digital assessments—a VR-based technical aptitude test and a GBA of cognitive abilities and personality characteristics—proposed for use in surgical residency selection.

Consistent with our hypotheses, the VR-based technical aptitude test demonstrated construct-consistent patterns of association. Scores were significantly correlated with established psychometric measures of dexterity and visuospatial ability, and with concurrent supervisor evaluations of technical skills among residents. In addition, scores increased systematically with level of surgical experience, differentiating between interns, residents, and expert surgeons. These findings support the interpretation that the VR test captures psychomotor and perceptual abilities relevant to surgical performance, and they reinforce the underlying assumption that higher scores reflect greater technical aptitude and, consequently, stronger technical performance during surgical training.

Similarly, the GBA demonstrated theoretically aligned associations. GBA scores were significantly related to established measures of intelligence and personality, and to supervisor-rated nontechnical performance dimensions in residency. As expected, the findings supported convergent and discriminant patterns consistent with the underlying construct: stronger associations were observed with competencies conceptually aligned with the GBA (eg, in Phase 1: neuroticism, which is related to stress tolerance, and conscientiousness, which is related to thoroughness; in Phase 2: decision-making and problem-solving, learning ability, creativity and cognitive flexibility, and stress tolerance) than with less related interpersonal dimensions (eg, in Phase 1: extraversion and agreeableness, which are related to behavior toward other people; in Phase 2: communication with patients and their families, and communication with medical staff and teamwork). These findings support the interpretation that the GBA measures nontechnical competencies relevant for resident selection, and reinforce the assumption that higher scores reflect higher levels of the assessed competencies.

With respect to fairness, although mean gender differences were observed on both selection tests, there was no evidence of differential prediction. The relationship between test scores and relevant performance criteria did not differ by gender, and findings remained robust after accounting for prior simulator and video game experience. These results suggest that individuals with comparable levels of relevant competencies—regardless of gender—are expected to achieve similar predicted performance outcomes based on test scores, and therefore, there is no evidence that the tests are systematically biased on the basis of gender.

We must note aspects of the findings of this study which suggest that the 2 novel tests do not fully differentiate between the assessment of technical aptitude, on the one hand, and of cognitive abilities and personality on the other. In particular, scores on each test are also correlated with measures not directly related to the construct the test is intended to measure ("construct-irrelevant variance"), suggesting some similarities between the constructs being assessed in the 2 tests. Specifically, scores on the technical aptitude test were significantly correlated with a measure of intelligence (RAPM scores in Phase 1), and with ratings of residents' decision-making and problem-solving, physical and mental endurance, stress tolerance, and creativity and cognitive flexibility in Phase 2. In addition, scores on the technical aptitude test were correlated with the personality dimensions of conscientiousness and neuroticism at a marginally significant level. With respect to the GBA, scores on this test were significantly correlated with a measure of visuospatial ability (MRT scores in Phase 1) and with technical skills in Phase 2.

Despite this construct-irrelevant variance, however, the stronger associations observed between each test and its theoretically relevant measures, together with the incremental validity findings, indicate that each assessment primarily reflects its intended construct despite partial overlap. Specifically, the incremental analyses demonstrated that the VR-based technical aptitude test uniquely predicted technical skills above and beyond GBA scores, while the GBA uniquely predicted broader nontechnical performance beyond the VR-based test. As such, the findings suggest that the 2 assessments provide complementary, nonredundant information for selection decisions. In this regard, it is also important to note that simulation-based assessments involve complex behavioral tasks that require the simultaneous engagement of multiple competencies [65]. Consequently, some degree of cross-domain association is theoretically expected, as performance in such environments may also draw upon decision-making, stress tolerance, and the ability to learn from feedback. These findings are in line with many previous findings assessing the relationship between performance on surgical simulators and cognitive abilities (eg, intelligence, memory, perceptual speed, and reasoning) [66-68] and personality characteristics (eg, stress tolerance, motivation, conscientiousness, and neuroticism) [69-72], and with studies showing correlations between nontechnical skills (such as decision-making and judgment) and surgical performance [72-74].

Comparison to the Literature

This study extends 2 previous studies [46,47], which described the development of the 2 novel tests examined here and presented initial evidence for their validity. Although the findings of those studies supported the potential of the tests in the selection of candidates for surgical training, the validity evidence they presented was partial, dealing primarily with content, internal structure, and response process evidence. In addition, although those studies identified gender differences in test scores, they were unable to evaluate potential gender bias because they did not include external criterion measures necessary to assess differential prediction. This study continues the validation process, focusing on evidence derived from relationships between scores on the novel tests and other variables: specifically, relationships with other tests measuring similar and different constructs, test-criterion relationships, and

XSL•FO
RenderX

relationships with training level. This study also addresses the issue of gender bias in the tests.

More broadly, our findings contribute to the existing literature examining the use of surgical simulators and game-based assessments for selection in general, as well as specifically for the selection of candidates for surgical training. Although digital simulation-based assessments are considered promising alternatives to traditional selection methods due to their unique advantages—such as the ability to simulate realistic "job-sample" tasks and to generate assessments based on large volumes of objective performance data [18,19,24-26,75]—empirical evidence supporting their use in surgical residency selection remains nascent.

In the broader literature on the use of VR surgical simulators, most validation studies have focused on the use of simulated tasks for training purposes or for assessing the proficiency of residents and practicing surgeons (eg, for feedback, credentialing, or examination) rather than for selection [76-78]. Prior research has demonstrated that performance on VR surgical simulators correlates with operating room performance [19,79,80], that substantial variability exists in learning curves among trainees and surgeons [81-83], and that simulator tasks can effectively discriminate between individuals at different levels of expertise [84-86]. However, as emphasized in the contemporary unified framework of validity described by Messick [44,45], evidence supporting the use of simulated tasks for training or proficiency assessment cannot be automatically generalized to their use for high-stakes selection decisions. It is therefore essential to examine validity evidence specifically in relation to the intended use of simulators as part of a selection process before incorporating them into such decisions.

Evidence specifically supporting the use of simulated tasks for resident selection remains limited [38-42]. Cope and Fenton-Lee [38] found no performance differences between interns who expressed interest in surgical careers and those who did not, suggesting a lack of self-selection. Jardin et al [39] and Salgado et al [42] reported no significant correlations between simulator task performance and traditional academic metrics such as USMLE scores, grades, or interview ratings, leading to the suggestion that technical aptitude assessment may capture competencies not reflected in conventional selection tools. Gallagher et al [40,41] incorporated simulator-based technical tasks into a broader multimethod selection system and reported some validity evidence; however, the evidence pertained to the overall selection system rather than to the simulator-based technical assessment specifically. None of the aforementioned studies used a systematic process to develop a comprehensive test for assessing candidates' technical aptitude based on accepted psychometric procedures (eg, developing a test blueprint, systematic selection of tasks, and developing a scoring system), or provided significant validity evidence for the use of surgical simulators in the selection process (test content, response process, internal structure, relationships to other variables, and consequences) [43]. In addition, most of these studies used MIS simulators to assess candidates for higher surgical education or surgical fellowships, so their assessment of technical skills is not applicable to candidates without previous surgical experience. In this context, our previous [46] and current studies address this gap by systematically developing and validating a simulation-based technical aptitude assessment for candidates without prior surgical experience.

As GBAs are still relatively new, only a limited number of studies have examined their potential for assessing cognitive abilities and personality characteristics in hiring and recruitment contexts [87-92]. Simons et al [87] developed a GBA to assess aspects of general intelligence and showed associations with established intelligence measures. Wiernik et al [88] designed and validated a GBA to assess cognitive and noncognitive competencies relevant to cyber occupations in the US Air Force, presenting evidence related to content (subject-matter expert input), internal structure (reliability and factor analysis), and relationships with other variables (convergent and discriminant evidence). Similarly, Landers et al [89] developed a theory-driven GBA of general cognitive ability for personnel selection, and revealed evidence for internal consistency, test-criterion relationships with job performance outcomes, and fairness across demographic groups. Haizel et al [90], Quwaider et al [91], and van Lankveld et al [92] explored the use of video game behaviors to infer five-factor personality traits, and provided only preliminary evidence regarding associations with self-report personality inventories. Collectively, this body of work provides initial support for the feasibility of designing game-based environments to assess cognitive abilities, job-relevant competencies, and personality traits for recruitment and occupational screening purposes. However, those studies were conducted in corporate, military, or experimental settings and did not apply game-based assessment to medical education or surgical residency selection. Accordingly, the combined findings of our earlier [47] and present validation studies represent the first systematic examination of the use of a game-based assessment to evaluate cognitive abilities and personality characteristics among candidates for surgical training, and to provide validity evidence supporting its use in surgical residency selection.

In addition, the findings of this study can be situated within the broader literature on the assessment of cognitive abilities and personality traits in the context of selection for surgical training. Traditionally, resident selection processes have relied on proxies such as curricula vitae, letters of recommendation, and unstructured interviews, as well as standardized cognitive measures (eg, academic examinations) [8,93]. Other approaches have included self-report questionnaires assessing personality traits, emotional intelligence, and grit. However, there is currently no consistent evidence that these methods substantially improve the selection of surgical residents [4,17]. In contrast, the GBA examined in this study is specifically designed to operationalize relevant cognitive and nontechnical competencies through interactive tasks that capture dynamic performance indicators (eg, response patterns and learning curves), capturing competencies that have been identified as important for surgical training (eg, decision-making, stress tolerance, and cognitive flexibility) [1-7]. As such, the GBA examined here reflects contemporary assessment approaches that prioritize behavior-based evidence and more realistic task contexts [29].

XSL•FO
RenderX

Gender Differences

Although we found no evidence of psychometric bias with respect to gender, significant mean differences in selection test scores were observed, with males scoring higher on both tests. It is possible that the gender differences in test scores found in our study and in our 2 previous validation studies [46,47] stem from the specific set of competencies assessed by the tests, or from the format in which the tests were administered. Various studies on sex and gender differences suggest that males and females tend to perform better on different kinds of challenges or tasks. For example, on average, men have been found to score significantly higher than women in tasks involving visuospatial perception and technical skills [75,81,84,94-102], and there is some evidence for a male advantage in intelligence tests [95]. For their part, women are thought to excel (again, on average) in tasks that require verbal abilities [95] and interpersonal skills [103,104]. The tests examined in this study were procedural and did not require verbal ability. In addition, the GBA examined in this study does not assess interpersonal skills, teamwork, leadership, integrity, and other cognitive abilities and personality characteristics relevant for selecting surgical residents, some of which might favor women. Therefore, it is possible that a more comprehensive selection process, including a broader assessment of nontechnical and verbal skills, would decrease or eliminate the gender difference found here. Future studies should examine whether other types of GBAs, such as gamified situational judgment tests [25], or other assessment methods, could lead to improvements in gender parity.

Probing deeper into the gender differences found in this study, these may be influenced, at least in part, by situational, environmental, or experiential factors. Prior exposure to video gaming, technical hobbies, and early engagement with digital tools has been proposed as an important contributor to gender differences in performance on technology-mediated and spatially demanding tasks [105-108]. Such exposure may influence familiarity with the testing format, but it may also contribute to the development of relevant cognitive and procedural skills over time [109,110]. Thus, observed differences could reflect experiential pathways (including broader patterns of socialization and access to opportunities) rather than inherent ability differences, and should therefore be considered when interpreting mean score disparities in VR-based or game-based assessments. Consistent with this perspective, findings from our 2 prior studies indicated that a portion of the gender difference in technical aptitude scores was explained by prior self-reported simulator and video game experience [46], and that the gender difference in GBA performance was largely accounted for by differences in self-reported video game experience [47]. In this study, adjusting for these exposure variables did not affect our findings regarding differential prediction or psychometric fairness. Nonetheless, the mean score differences we found may partly reflect differences in prior experience and opportunity structures. Future research should examine these mechanisms more directly, for example, by incorporating objective measures of prior exposure, experimentally manipulating practice opportunities, or evaluating whether structured familiarization procedures before testing attenuate group-level disparities.

Another possibility, not mutually exclusive with the first, is that the gender differences we found reflect, at least to some extent, a “stereotype threat” effect. According to this explanation, negative stereotypes associated with sex- or gender-typical performance could lead to performance-related anxiety in members of that gender, which ultimately diminishes their performance [95]. Given that surgery is still a male-dominated field, it is possible that this phenomenon affected, to some degree, the performance of female participants in our study. In addition, to the extent that stereotype threat operated among the residents in our sample, this could have affected their performance both in the selection tests and more broadly during their residency, creating a false correlation between their test scores and their performance evaluations. This hypothesis is supported by evidence that stereotype threat indeed may have an effect on surgical trainees [111,112].

Finally, although we found no evidence for gender bias in the tests in the psychometric sense, we recognize that the absence of statistical bias does not resolve all fairness concerns in high-stakes selection contexts. Our analyses indicate that test scores predict relevant performance criteria similarly for men and women; however, differences in average scores may still result in different selection rates when score-based cutoffs are applied. It is therefore important to distinguish between statistical fairness, which concerns the measurement properties of a test, and the practical consequences of its use in selection, which may affect representation even when psychometric bias is absent.

From a measurement perspective, selection tools that meet accepted standards of validity and fairness can improve the accuracy of selection decisions. At the same time, questions related to equity, diversity, and workforce composition involve institutional and policy considerations that extend beyond what psychometric validation alone can address and relate to how selection tools are implemented in practice.

In this broader context, expanding selection processes to include a more comprehensive assessment of nontechnical competencies, such as interpersonal skills, may enhance the overall validity of selection decisions and help mitigate differential outcomes associated with specific ability profiles [103,104]. In addition, institutions may choose to pursue representation goals through policy-level decisions that are external to the assessment tools themselves, such as the use of different selection cutoffs, separate norms, or targeted allocation strategies. Taken together, these considerations underscore the need to interpret psychometric fairness findings cautiously and within a wider applied and societal context.

Implications

Currently, most surgical programs select residents using traditional methods, whose relationships with later clinical and operative performance have limited and inconsistent empirical support [10-14]. In this study, we present evidence supporting the potential of a VR technical aptitude test and a GBA for objectively assessing cognitive abilities and personality characteristics for the selection of candidates for surgical training.

XSL•FO
RenderX

The combined use of a VR-based technical aptitude test and a gamified behavioral assessment may offer a structured and standardized approach to evaluating multiple competency domains considered relevant for surgical training [1-7]. The technological features of VR and GBAs allow automated administration and scoring, standardized testing conditions, greater engagement of examinees, and high-resolution behavioral data capture. These characteristics may reduce certain sources of human subjectivity in scoring and enhance feasibility in large-scale contexts. In addition, the integration of assessment results across technical, cognitive, and personality-related domains may provide selection committees with a broader representation of candidate performance, offering an in-depth understanding of each candidate's strengths and weaknesses. Such information could potentially support more informed and structured decision-making, for example, by helping selection committees consider the fit between a candidate's profile and the characteristics of a given training program. Assessment results may also help educators identify areas where incoming residents might benefit from targeted supervision or skills development. At the program level, aggregated data could contribute to ongoing evaluation of training priorities and curricular design.

The importance of developing and validating structured assessment tools is particularly salient in the current context of surgical education, which faces challenges including work-hour restrictions, increasing technological complexity, and growing economic pressures to enhance efficiency in the operating room [8]. Within this landscape, strengthening the evidentiary foundation of selection processes represents a critical priority. The findings of this study contribute to the cumulative process of validity evidence generation for technology-enhanced assessments in surgical education. The conceptual framework and validation approach may also inform the development of technology-enhanced assessments in other medical specialties and professional domains.

As a final note, we acknowledge that the extent to which the advantages of the proposed tools translate into improved decision-making remains to be determined empirically, particularly in light of the current absence of longitudinal evidence linking pretraining assessment scores to subsequent performance during residency. Future research should extend this work through longitudinal designs, cross-program collaborations, and the development of more standardized performance criteria across training institutions.

Research Limitations and Future Directions

This study examines a comprehensive assessment of technical aptitude, cognitive abilities, and personality characteristics. Other key strengths include the breadth of the validity evidence provided and the large sample of interns, residents, and expert surgeons.

This study also has limitations. First, while our sample was large and included doctors with varying levels of experience, all participants came from one country, which limits the generalizability of the results. However, the competencies we assessed, such as technical skills in laparoscopic procedures, are fundamental aspects of surgical training and are typically standardized across medical education programs worldwide. This standardization, along with common international guidelines and best practices in surgical training, suggests that these competencies are unlikely to be distributed differently among doctors from different countries. Therefore, we believe our findings still provide valuable insights relevant to a broader international context.

Second, residency performance outcomes were based on supervisor ratings, which are inherently subjective and may be influenced by rater-related biases, including halo effects and lack of full independence among raters working within the same clinical environment. Although each resident was evaluated by multiple supervisors (3 or 4), and interrater reliability was high, complete independence of ratings cannot be assumed within a shared clinical environment where supervisors interact regularly. Indeed, the intercorrelations among the 16 performance criteria ranged from 0.21 to 0.89 (mean r 0.55), indicating moderate shared variance across domains. This pattern suggests that the evaluations may partly reflect an overall impression of resident performance. That said, the variability in both correlation magnitudes and mean ratings across criteria indicates that supervisors did not rate all domains uniformly high or low, which would be expected under a strong halo effect. Future research should incorporate additional objective, behavioral, or independently assessed performance indicators to further strengthen criterion-related evidence and reduce the risk of shared method variance.

Third, although this study provided evidence for test-criterion relationships, these data were collected using a concurrent design. Accordingly, based on the current evidence, we cannot conclude definitively that administering these tests to candidates prior to residency would predict their subsequent performance during training. Prospective longitudinal research, in which selection assessments are administered before residency and performance is evaluated at later stages, is required to establish predictive utility in high-stakes selection contexts.

A fourth limitation concerns the proprietary nature of the GBA scoring system. Although the general scoring framework and psychometric properties are described, detailed operational definitions of specific behavioral indicators and exact weighting parameters cannot be publicly disclosed due to contractual and trade secret protections. Consequently, full computational replication of the scoring procedure by independent researchers is not possible. Nevertheless, consistent with established standards for educational and psychological testing [45], this study evaluates score interpretation through relationships with other variables, which remain fully open to empirical scrutiny.

Fifth, the digital format of both selection tests raises the possibility that prior familiarity with gaming or computerized environments may influence performance. Although this study, as well as our 2 previous validation studies [46,47], found only weak associations between self-reported gaming experience and test scores, this potential source of construct-irrelevant variance cannot be entirely excluded. Indeed, some prior research has reported performance advantages for individuals with video game experience in both GBAs [107,108] and surgical simulator tasks [113,114]. Future research should more

XSL•FO
RenderX

systematically examine whether prior exposure to digital gaming environments confers an unintended advantage and, if so, explore design-based strategies to mitigate such effects—for example, through standardized familiarization periods, structured practice sessions prior to scored tasks, interface simplification, or task designs that reduce construct-irrelevant variance arising from digital interface familiarity.

Finally, the gamified test presented in our study does not assess all cognitive abilities and personality characteristics relevant to the selection of surgical residents. In particular, it does not include important competencies such as interpersonal skills, teamwork, leadership, and integrity. Future studies should examine whether additional assessment methods could complement the proposed tools in capturing these competencies. Established approaches in selection research include traditional situational judgment tests, structured self-report questionnaires, and simulation-based assessments with trained observers rating interpersonal behaviors [115]. In parallel, recent technological developments offer promising extensions, such as gamified situational judgment tests [25], as well as emerging approaches using machine learning to model interpersonal behaviors from rich behavioral data [116]. Integrating such methods may enable a more comprehensive and ecologically valid assessment of nontechnical competencies relevant to surgical training selection.

Conclusions

This study presents validity evidence regarding the use of a VR-based technical aptitude test and a game-based assessment of cognitive abilities and personality characteristics in the selection of candidates for surgical training. The findings extend 2 previous validation studies by contributing additional evidence within the domain of relations to other variables, as well as evidence regarding gender fairness. Although this study does not empirically establish downstream effects on residency outcomes, attrition, or patient care, the rigorous validation of structured assessment approaches may represent an important step toward more evidence-informed selection practices. To achieve a more comprehensive evaluation of competencies relevant to surgical training, these assessments may be considered as components of a broader, multimodal selection framework that also includes measures of interpersonal and communication skills.

Acknowledgments

Generative artificial intelligence (AI) tools were used solely for language editing and to improve clarity. The authors are fully responsible for all content, analyses, and interpretations presented in the manuscript.

Funding

This work was supported by the Israel Science Foundation (grant 1830/20).

Data Availability

The datasets used and analyzed during this study are available from the corresponding author on reasonable request.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Supporting materials for study analyses.
[DOCX File, 208 KB - mededu_v12i1e82515_app1.docx]

References

1. Cuschieri A, Francis N, Crosby J, Hanna GB. What do master surgeons think of surgical competence and revalidation? Am J Surg 2001 Aug;182(2):110-116. [doi: 10.1016/s0002-9610(01)00667-5] [Medline: 11574079]
2. Baldwin PJ, Paisley AM, Brown SP. Consultant surgeons' opinion of the skills required of basic surgical trainees. Br J Surg 1999 Aug;86(8):1078-1082. [doi: 10.1046/j.1365-2168.1999.01169.x] [Medline: 10460649]
3. Dean B, Jones L, Garfjeld Roberts P, Rees J. What is known about the attributes of a successful surgical trainer? A systematic review. J Surg Educ 2017;74(5):843-850. [doi: 10.1016/j.jsurg.2017.01.010] [Medline: 28392267]
4. Bann S, Darzi A. Selection of individuals for training in surgery. Am J Surg 2005 Jul;190(1):98-102. [doi: 10.1016/j.amjsurg.2005.04.002] [Medline: 15972179]
5. Grantcharov TP, Reznick RK. Training tomorrow's surgeons: what are we looking for and how can we achieve it? ANZ J Surg 2009 Mar;79(3):104-107. [doi: 10.1111/j.1445-2197.2008.04823.x] [Medline: 19317771]
6. Gardner AK, Cavanaugh KJ, Willis RE, et al. Great expectations? Future competency requirements among candidates entering surgery training. J Surg Educ 2020;77(2):267-272. [doi: 10.1016/j.jsurg.2019.09.001] [Medline: 31606376]
7. Gazit N, Ben-Gal G, Eliashar R. Using job analysis for identifying the desired competencies of 21st-century surgeons for improving trainees selection. J Surg Educ 2023 Jan;80(1):81-92. [doi: 10.1016/j.jsurg.2022.08.015] [Medline: 36175291]

XSL•FO
RenderX

8. Schaverien MV. Selection for surgical training: an evidence-based review. J Surg Educ 2016;73(4):721-729. [doi: 10.1016/j.jsurg.2016.02.007] [Medline: 27133583]
9. Collins JP, Doherty EM, Traynor O. Selection into surgical education and training. In: Nestel D, Dalrymple K, AR PJ, editors. Advancing Surgical Education: Springer; 2019:157-170. [doi: 10.1007/978-981-13-3128-2_15]
10. Bowe SN, Laury AM, Gray ST. Associations between otolaryngology applicant characteristics and future performance in residency or practice: a systematic review. Otolaryngol Head Neck Surg 2017 Jun;156(6):1011-1017. [doi: 10.1177/0194599817698430] [Medline: 28349776]
11. Harfmann KL, Zirwas MJ. Can performance in medical school predict performance in residency? A compilation and review of correlative studies. J Am Acad Dermatol 2011 Nov;65(5):1010-1022. [doi: 10.1016/j.jaad.2010.07.034] [Medline: 21612841]
12. Kenny S, McInnes M, Singh V. Associations between residency selection strategies and doctor performance: a meta-analysis. Med Educ 2013 Aug;47(8):790-800. [doi: 10.1111/medu.12234] [Medline: 23837425]
13. Oldfield Z, Beasley SW, Smith J, Anthony A, Watt A. Correlation of selection scores with subsequent assessment scores during surgical training. ANZ J Surg 2013 Jun;83(6):412-416. [doi: 10.1111/ans.12176] [Medline: 23647783]
14. Stephenson-Famy A, Houmard BS, Oberoi S, Manyak A, Chiang S, Kim S. Use of the interview in resident candidate selection: a review of the literature. J Grad Med Educ 2015 Dec;7(4):539-548. [doi: 10.4300/JGME-D-14-00236.1] [Medline: 26692964]
15. Louridas M, Szasz P, de Montbrun S, Harris KA, Grantcharov TP. Can we predict technical aptitude? Ann Surg 2016;263(4):673-691. [doi: 10.1097/SLA.0000000000001283]
16. Maan ZN, Maan IN, Darzi AW, Aggarwal R. Systematic review of predictors of surgical performance. Br J Surg 2012 Dec;99(12):1610-1621. [doi: 10.1002/bjs.8893] [Medline: 23034658]
17. Gardner AK, Dunkin BJ. Evaluation of validity evidence for personality, emotional intelligence, and situational judgment tests to identify successful residents. JAMA Surg 2018 May 1;153(5):409-416. [doi: 10.1001/jamasurg.2017.5013] [Medline: 29282462]
18. Gardner AK, Ritter EM, Paige JT, Ahmed RA, Fernandez G, Dunkin BJ. Simulation-based selection of surgical trainees: considerations, challenges, and opportunities. J Am Coll Surg 2016 Sep;223(3):530-536. [doi: 10.1016/j.jamcollsurg.2016.05.021] [Medline: 27321389]
19. Kramp KH, van Det MJ, Hoff C, Veeger N, ten Cate Hoedemaker HO, Pierie J. The predictive value of aptitude assessment in laparoscopic surgery: a meta-analysis. Med Educ 2016 Apr;50(4):409-427. [doi: 10.1111/medu.12945] [Medline: 26995481]
20. Bor R, Eriksen C, Hubbard T, King RE. Pilot Selection: Psychological Principles and Practice: CRC Press; 2019. [doi: 10.4324/9780429492105]
21. Mandal S. Brief introduction of virtual reality and its challenges. Int J Sci Eng Res 2013;4:304-309 [FREE Full text]
22. Mäkinen H, Haavisto E, Havola S, Koivisto JM. User experiences of virtual reality technologies for healthcare in learning: an integrative review. Behav Inf Technol 2022 Jan 2;41(1):1-17. [doi: 10.1080/0144929X.2020.1788162]
23. Seaborn K, Fels DI. Gamification in theory and action: a survey. Int J Hum Comput Stud 2015 Feb;74:14-31. [doi: 10.1016/j.ijhcs.2014.09.006]
24. Landers RN, Sanchez DR. Game‐based, gamified, and gamefully designed assessments for employee selection: definitions, distinctions, design, and validation. Int J Selection Assessment 2022 Mar;30(1):1-13. [doi: 10.1111/ijsa.12376]
25. Georgiou K, Gouras A, Nikolaou I. Gamification in employee selection: the development of a gamified assessment. Int J Selection Assessment 2019 Jun;27(2):91-103 [FREE Full text] [doi: 10.1111/ijsa.12240]
26. Gomez MJ, Ruipérez-Valiente JA, Clemente FJG. A systematic literature review of game-based assessment studies: trends and challenges. IEEE Trans Learning Technol 2023;16(4):500-515. [doi: 10.1109/TLT.2022.3226661]
27. Ramos-Villagrasa PJ, Fernández-Del-Río E, Castro Á. Game-related assessments for personnel selection: a systematic review. Front Psychol 2022;13:952002. [doi: 10.3389/fpsyg.2022.952002] [Medline: 36248590]
28. Zourmpakis AI, Kalogiannakis M, Papadakis S. The effects of adaptive gamification in science learning: a comparison between traditional inquiry-based learning and gender differences. Computers 2024;13(12):324. [doi: 10.3390/computers13120324]
29. Su F, Zou D. A systematic review of game-based assessment in education in the past decade. Knowl Manag E-Learn 2024 Sep 30;16:451-476. [doi: 10.34105/j.kmel.2024.16.021]
30. von Davier AA, Deonovic B, Yudelson M, Polyak ST, Woo A. Computational psychometrics approach to holistic learning and assessment systems. Front Educ 2019;4:69. [doi: 10.3389/feduc.2019.00069]
31. Udeozor C, Chan P, Russo Abegão F, Glassey J. Game-based assessment framework for virtual reality, augmented reality and digital game-based learning. Int J Educ Technol High Educ 2023;20(1):36. [doi: 10.1186/s41239-023-00405-6]
32. Tarr MJ, Warren WH. Virtual reality in behavioral neuroscience and beyond. Nat Neurosci 2002 Nov;5 Suppl:1089-1092. [doi: 10.1038/nn948] [Medline: 12403993]
33. Parsons TD. Virtual reality for enhanced ecological validity and experimental control in the clinical, affective and social neurosciences. Front Hum Neurosci 2015;9:660. [doi: 10.3389/fnhum.2015.00660] [Medline: 26696869]

XSL•FO
RenderX

34. Bendick M, Nunes AP. Developing the research basis for controlling bias in hiring. J Soc Issues 2012 Jun;68(2):238-262. [doi: 10.1111/j.1540-4560.2012.01747.x]
35. Tews MJ, Stafford K, Zhu J. Beauty revisited: the impact of attractiveness, ability, and personality in the assessment of employment suitability. Int J Select Assess 2009 Mar;17(1):92-100. [doi: 10.1111/j.1468-2389.2009.00454.x]
36. Donaldson SI, Grant-Vallone EJ. Understanding self-report bias in organizational behavior research. J Bus Psychol 2002 Dec;17(2):245-260. [doi: 10.1023/A:1019637632584]
37. Fisher RJ, Katz JE. Social-desirability bias and the validity of self-reported values. Psychol Mark 2000 Feb;17(2):105-120. [doi: 10.1002/(SICI)1520-6793(200002)17:2<105::AID-MAR3>3.0.CO;2-9]
38. Cope DH, Fenton-Lee D. Assessment of laparoscopic psychomotor skills in interns using the MIST Virtual Reality Simulator: a prerequisite for those considering surgical training? ANZ J Surg 2008 Apr;78(4):291-296. [doi: 10.1111/j.1445-2197.2007.04440.x] [Medline: 18366403]
39. Jardine D, Hoagland B, Perez A, Gessler E. Evaluation of surgical dexterity during the interview day: another factor for consideration. J Grad Med Educ 2015 Jun;7(2):234-237. [doi: 10.4300/JGME-D-14-00546.1] [Medline: 26221441]
40. Gallagher AG, O'Sullivan GC, Neary PC, et al. An objective evaluation of a multi-component, competitive, selection process for admitting surgeons into higher surgical training in a national setting. World J Surg 2014 Feb;38(2):296-304. [doi: 10.1007/s00268-013-2302-4] [Medline: 24146198]
41. Gallagher AG, Neary P, Gillen P, et al. Novel method for assessment and selection of trainees for higher surgical training in general surgery. ANZ J Surg 2008 Apr;78(4):282-290. [doi: 10.1111/j.1445-2197.2008.04439.x] [Medline: 18366402]
42. Salgado J, Grantcharov TP, Papasavas PK, Gagne DJ, Caushaj PF. Technical skills assessment as part of the selection process for a fellowship in minimally invasive surgery. Surg Endosc 2009 Mar;23(3):641-644. [doi: 10.1007/s00464-008-0033-7] [Medline: 18813975]
43. Korndorffer JR Jr, Kasten SJ, Downing SM. A call for the utilization of consensus standards in the surgical education literature. Am J Surg 2010 Jan;199(1):99-104. [doi: 10.1016/j.amjsurg.2009.08.018] [Medline: 20103073]
44. Messick S. Standards of validity and the validity of standards in performance asessment. Educ Meas 1995 Dec;14(4):5-8. [doi: 10.1111/j.1745-3992.1995.tb00881.x]
45. American Educational Research Association, American Psychological Association, National Council on Measurement in Education. Standards for Educational and Psychological Testing: American Educational Research Association; 2014.
46. Gazit N, Ben-Gal G, Eliashar R. Development and validation of an objective virtual reality tool for assessing technical aptitude among potential candidates for surgical training. BMC Med Educ 2024 Mar 14;24(1):286. [doi: 10.1186/s12909-024-05228-1] [Medline: 38486166]
47. Gazit N, Ben-Gal G, Eliashar R. Game-based assessment of cognitive abilities and personality characteristics for surgical resident selection: a preliminary validation study. JMIR Med Educ 2025 Aug 15;11:e72264. [doi: 10.2196/72264] [Medline: 40815821]
48. Kawaguchi K, Egi H, Hattori M, Sawada H, Suzuki T, Ohdan H. Validation of a novel basic virtual reality simulator, the LAP-X, for training basic laparoscopic skills. Minim Invasive Ther Allied Technol 2014 Oct;23(5):287-293. [doi: 10.3109/13645706.2014.903853] [Medline: 24773373]
49. Cook DA, Hatala R. Validation of educational assessments: a primer for simulation and beyond. Adv Simul 2016 Jan;1(1):1-12. [doi: 10.1186/s41077-016-0033-y]
50. Downing SM. Validity: on the meaningful interpretation of assessment data. Med Educ (Chicago Ill) 2003 Sep;37(9):830-837. [doi: 10.1046/j.1365-2923.2003.01594.x]
51. Cook DA. Much ado about differences: why expert-novice comparisons add little to the validity argument. Adv in Health Sci Educ 2015 Aug;20(3):829-834. [doi: 10.1007/s10459-014-9551-3]
52. Borgersen NJ, Naur TMH, Sørensen SMD, et al. Gathering validity evidence for surgical simulation: a systematic review. Ann Surg 2018 Jun;267(6):1063-1068. [doi: 10.1097/SLA.0000000000002652] [Medline: 29303808]
53. Silvennoinen M, Mecklin JP, Saariluoma P, Antikainen T. Expertise and skill in minimally invasive surgery. Scand J Surg 2009;98(4):209-213. [doi: 10.1177/145749690909800403] [Medline: 20218416]
54. TIFFIN J, ASHER EJ. The Purdue pegboard; norms and studies of reliability and validity. J Appl Psychol 1948 Jun;32(3):234-247. [doi: 10.1037/h0061266] [Medline: 18867059]
55. Lawson I. Purdue Pegboard test. Occup Med (Chic Ill) 2019 Aug 22;69(5):376-377. [doi: 10.1093/occmed/kqz044]
56. Vandenberg SG, Kuse AR. Mental rotations, a group test of three-dimensional spatial visualization. Percept Mot Skills 1978 Oct;47(2):599-604. [doi: 10.2466/pms.1978.47.2.599] [Medline: 724398]
57. Peters M, Laeng B, Latham K, Jackson M, Zaiyouna R, Richardson C. A redrawn Vandenberg and Kuse mental rotations test: different versions and factors that affect performance. Brain Cogn 1995 Jun;28(1):39-58. [doi: 10.1006/brcg.1995.1032] [Medline: 7546667]
58. Raven J, Raven JC, Court JH. The Advanced Progressive Matrices In: Manual for Raven's Progressive Matrices and Vocabulary Scales: Harcourt Assessment; 1998.
59. Donnellan MB, Oswald FL, Baird BM, Lucas RE. The mini-IPIP scales: tiny-yet-effective measures of the Big Five factors of personality. Psychol Assess 2006 Jun;18(2):192-203. [doi: 10.1037/1040-3590.18.2.192] [Medline: 16768595]

XSL•FO
RenderX

60. Rubin M. When to adjust alpha during multiple testing: a consideration of disjunction, conjunction, and individual testing. Synthese 2021 Dec;199(3-4):10969-11000. [doi: 10.1007/s11229-021-03276-4]
61. Streiner DL, Norman GR. Correction for multiple testing: is there a resolution? Chest 2011 Jul;140(1):16-18. [doi: 10.1378/chest.11-0523] [Medline: 21729890]
62. Arvey RD, Renz GL. Fairness in the selection of employees. J Bus Ethics 1992 May;11(5-6):331-340. [doi: 10.1007/BF00870545]
63. Bartlett CJ, Bobko P, Mosier SB, Hannan R. Testing for fairness with a moderated multiple regression strategy: an alternative to differential analysis. Pers Psychol 1978 Jun;31(2):233-241. [doi: 10.1111/j.1744-6570.1978.tb00442.x]
64. Koo TK, Li MY. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. J Chiropr Med 2016 Jun;15(2):155-163. [doi: 10.1016/j.jcm.2016.02.012] [Medline: 27330520]
65. Levy R. Psychometric and evidentiary advances, opportunities, and challenges for simulation-based assessment. Educ Assess 2013 Jul;18(3):182-207. [doi: 10.1080/10627197.2013.814517]
66. Mathias AP, Vogel P, Knauff M. Different cognitive styles can affect performance in laparoscopic surgery skill training. Surg Endosc 2020 Nov;34(11):4866-4873. [doi: 10.1007/s00464-019-07267-y] [Medline: 31823045]
67. Groenier M, Schraagen JMC, Miedema HAT, Broeders IAMJ. The role of cognitive abilities in laparoscopic simulator training. Adv Health Sci Educ 2014 May;19(2):203-217. [doi: 10.1007/s10459-013-9455-7]
68. Jungmann F, Gockel I, Hecht H, et al. Impact of perceptual ability and mental imagery training on simulated laparoscopic knot-tying in surgical novices using a Nissen fundoplication model. Scand J Surg 2011;100(2):78-85. [doi: 10.1177/145749691110000203] [Medline: 21737382]
69. Maschuw K, Schlosser K, Kupietz E, Slater EP, Weyers P, Hassan I. Do soft skills predict surgical performance?: a single-center randomized controlled trial evaluating predictors of skill acquisition in virtual reality laparoscopy. World J Surg 2011 Mar;35(3):480-486. [doi: 10.1007/s00268-010-0933-2] [Medline: 21190109]
70. Wetzel CM, Black SA, Hanna GB, et al. The effects of stress and coping on surgical performance during simulations. Ann Surg 2010 Jan;251(1):171-176. [doi: 10.1097/SLA.0b013e3181b3b2be] [Medline: 20032721]
71. Hattori M, Egi H, Hasunuma N. Conscientiousness counts: how personality traits impact laparoscopic surgical skill improvement in medical students. J Surg Educ 2023 Oct;80(10):1412-1417. [doi: 10.1016/j.jsurg.2023.07.015] [Medline: 37596108]
72. Rosendal AA, Sloth SB, Rölfing JD, Bie M, Jensen RD. Technical, non-technical, or both? A scoping review of skills in simulation-based surgical training. J Surg Educ 2023 May;80(5):731-749. [doi: 10.1016/j.jsurg.2023.02.011] [Medline: 36906398]
73. Riem N, Boet S, Bould MD, Tavares W, Naik VN. Do technical skills correlate with non-technical skills in crisis resource management: a simulation study. Br J Anaesth 2012 Nov;109(5):723-728. [doi: 10.1093/bja/aes256] [Medline: 22850221]
74. Gillespie BM, Harbeck E, Kang E, Steel C, Fairweather N, Chaboyer W. Correlates of non-technical skills in surgery: a prospective study. BMJ Open 2017 Jan 30;7(1):e014480. [doi: 10.1136/bmjopen-2016-014480] [Medline: 28137931]
75. Ali A, Subhi Y, Ringsted C, Konge L. Gender differences in the acquisition of surgical skills: a systematic review. Surg Endosc 2015 Nov;29(11):3065-3073. [doi: 10.1007/s00464-015-4092-2] [Medline: 25631116]
76. Agha RA, Fowler AJ. The role and validity of surgical simulation. Int Surg 2015 Feb;100(2):350-357. [doi: 10.9738/INTSURG-D-14-00004.1] [Medline: 25692441]
77. McCluney AL, Vassiliou MC, Kaneva PA, et al. FLS simulator performance predicts intraoperative laparoscopic skill. Surg Endosc 2007 Nov;21(11):1991-1995. [doi: 10.1007/s00464-007-9451-1] [Medline: 17593434]
78. Paisley MAM, Baldwin PJ, Paterson-Brown S. Validity of surgical simulation for the assessment of operative skill. Br J Surg 2001 Nov;88(11):1525-1532. [doi: 10.1046/j.0007-1323.2001.01880.x] [Medline: 11683753]
79. Kundhal PS, Grantcharov TP. Psychomotor performance measured in a virtual environment correlates with technical skills in the operating room. Surg Endosc 2009 Mar;23(3):645-649. [doi: 10.1007/s00464-008-0043-5] [Medline: 18622548]
80. Matsuda T, McDougall EM, Ono Y, et al. Positive correlation between motion analysis data on the LapMentor virtual reality laparoscopic surgical simulator and the results from videotape assessment of real laparoscopic surgeries. J Endourol 2012 Nov;26(11):1506-1511. [doi: 10.1089/end.2012.0183] [Medline: 22642549]
81. Moglia A, Morelli L, Ferrari V, Ferrari M, Mosca F, Cuschieri A. Distribution of innate psychomotor skills recognized as important for surgical specialization in unconditioned medical undergraduates. Surg Endosc 2018 Oct;32(10):4087-4095. [doi: 10.1007/s00464-018-6146-8] [Medline: 29541863]
82. Moglia A, Ferrari V, Morelli L, et al. Distribution of innate ability for surgery amongst medical students assessed by an advanced virtual reality surgical simulator. Surg Endosc 2014 Jun;28(6):1830-1837. [doi: 10.1007/s00464-013-3393-6] [Medline: 24442679]
83. Louridas M, Szasz P, Fecso AB, et al. Practice does not always make perfect: need for selection curricula in modern surgical training. Surg Endosc 2017 Sep;31(9):3718-3727. [doi: 10.1007/s00464-017-5572-3] [Medline: 28451813]
84. McDougall EM, Corica FA, Boker JR, et al. Construct validity testing of a laparoscopic surgical simulator. J Am Coll Surg 2006 May;202(5):779-787. [doi: 10.1016/j.jamcollsurg.2006.01.004] [Medline: 16648018]

XSL•FO
RenderX

85. van Dongen KW, Tournoij E, van der Zee DC, Schijven MP, Broeders I. Construct validity of the LapSim: can the LapSim virtual reality simulator distinguish between novices and experts? Surg Endosc 2007 Aug;21(8):1413-1417. [doi: 10.1007/s00464-006-9188-2] [Medline: 17294307]
86. Zhang A, Hünerbein M, Dai Y, Schlag PM, Beller S. Construct validity testing of a laparoscopic surgery simulator (Lap Mentor®). Surg Endosc 2008 Jun;22(6):1440-1444. [doi: 10.1007/s00464-007-9625-x]
87. Simons A, Wohlgenannt I, Zelt S, Weinmann M, Schneider J, vom Brocke J. Intelligence at play: game-based assessment using a virtual-reality application. Virtual Real 2023 Sep;27(3):1827-1843. [doi: 10.1007/s10055-023-00752-9]
88. Wiernik BM, Raghavan M, Caretta TR, Coovert MD. Developing and validating a serious game‐based assessment for cyber occupations in the US Air Force. Int J Sel Assess 2022 Mar;30(1):27-47. [doi: 10.1111/ijsa.12378]
89. Landers RN, Armstrong MB, Collmus AB, Mujcic S, Blaik J. Theory-driven game-based assessment of general cognitive ability: design theory, measurement, prediction of performance, and test fairness. J Appl Psychol 2022 Oct;107(10):1655-1677. [doi: 10.1037/apl0000954] [Medline: 34672652]
90. Haizel P, Vernanda G, Wawolangi KA, Hanafiah N. Personality assessment video game based on The Five-Factor Model. Procedia Comput Sci 2021;179:566-573. [doi: 10.1016/j.procs.2021.01.041]
91. Quwaider M, Alabed A, Duwairi R. Shooter video games for personality prediction using five factor model traits and machine learning. Simul Model Pract Theory 2023 Jan;122:102665. [doi: 10.1016/j.simpat.2022.102665]
92. van Lankveld G, Spronck P, Den Herik J, Arntz A. Games as personality profiling tools. Presented at: 2011 IEEE Conference on Computational Intelligence and Games (CIG 2011); Aug 31 to Sep 3, 2011; Seoul, Korea (South p. 197-202. [doi: 10.1109/CIG.2011.6032007]
93. Lipman JM, Colbert CY, Ashton R, et al. A systematic review of metrics utilized in the selection and prediction of future performance of residents in the United States. J Grad Med Educ 2023 Dec;15(6):652-668. [doi: 10.4300/JGME-D-22-00955.1] [Medline: 38045930]
94. Thorson CM, Kelly JP, Forse RA, Turaga KK. Can we continue to ignore gender differences in performance on simulation trainers? J Laparoendosc Adv Surg Tech 2011 May;21(4):329-333. [doi: 10.1089/lap.2010.0368]
95. Kheloui S, Jacmin-Park S, Larocque O, et al. Sex/gender differences in cognitive abilities. Neurosci Biobehav Rev 2023 Sep;152:105333. [doi: 10.1016/j.neubiorev.2023.105333] [Medline: 37517542]
96. Linn MC, Petersen AC. Emergence and characterization of sex differences in spatial ability: a meta-analysis. Child Dev 1985 Dec;56(6):1479-1498. [doi: 10.2307/1130467] [Medline: 4075870]
97. Masters MS. The gender difference on the Mental Rotations test is not due to performance factors. Mem Cogn 1998 May;26(3):444-448. [doi: 10.3758/BF03201154]
98. Maeda Y, Yoon SY. A meta-analysis on gender differences in mental rotation ability measured by the Purdue Spatial Visualization Tests: Visualization of Rotations (PSVT:R). Educ Psychol Rev 2013 Mar;25(1):69-94. [doi: 10.1007/s10648-012-9215-x]
99. Grantcharov TP, Bardram L, Funch-Jensen P, Rosenberg J. Impact of hand dominance, gender, and experience with computer games on performance in virtual reality laparoscopy. Surg Endosc 2003 Jul;17(7):1082-1085. [doi: 10.1007/s00464-002-9176-0] [Medline: 12728373]
100. Elneel FHF, Carter F, Tang B, Cuschieri A. Extent of innate dexterity and ambidexterity across handedness and gender: implications for training in laparoscopic surgery. Surg Endosc 2008 Jan;22(1):31-37. [doi: 10.1007/s00464-007-9533-0] [Medline: 17965919]
101. White MT, Welch K. Does gender predict performance of novices undergoing Fundamentals of Laparoscopic Surgery (FLS) training? Am J Surg 2012 Mar;203(3):397-400. [doi: 10.1016/j.amjsurg.2011.09.020] [Medline: 22364906]
102. Lin D, Pena G, Field J, et al. What are the demographic predictors in laparoscopic simulator performance? ANZ J Surg 2016 Dec;86(12):983-989. [doi: 10.1111/ans.12992] [Medline: 25645288]
103. Sugawara A, Ishikawa K, Motoya R, Kobayashi G, Moroi Y, Fukushima T. Characteristics and gender differences in the medical interview skills of Japanese medical students. Intern Med 2017;56(12):1507-1513. [doi: 10.2169/internalmedicine.56.8135] [Medline: 28626175]
104. Graf J, Smolka R, Simoes E, et al. Communication skills of medical students during the OSCE: gender-specific differences in a longitudinal trend study. BMC Med Educ 2017 May 2;17(1):75. [doi: 10.1186/s12909-017-0913-4] [Medline: 28464857]
105. Feng J, Spence I, Pratt J. Playing an action video game reduces gender differences in spatial cognition. Psychol Sci 2007 Oct;18(10):850-855. [doi: 10.1111/j.1467-9280.2007.01990.x] [Medline: 17894600]
106. Cherney ID. Mom, let me play more computer games: they improve my mental rotation skills. Sex Roles 2008 Dec;59(11-12):776-786. [doi: 10.1007/s11199-008-9498-z]
107. Kim YJ, Shute VJ. The interplay of game elements with psychometric qualities, learning, and enjoyment in game-based assessment. Comput Educ 2015 Sep;87:340-356. [doi: 10.1016/j.compedu.2015.07.009]
108. Ventura M, Shute V. The validity of a game-based assessment of persistence. Comput Human Behav 2013 Nov;29(6):2568-2572. [doi: 10.1016/j.chb.2013.06.033]
109. Granic I, Lobel A, Engels R. The benefits of playing video games. Am Psychol 2014 Jan;69(1):66-78. [doi: 10.1037/a0034857] [Medline: 24295515]

XSL•FO
RenderX

110. Reynaldo C, Christian R, Hosea H, Gunawan AAS. Using video games to improve capabilities in decision making and cognitive skill: a literature review. Procedia Comput Sci 2021;179:211-221. [doi: 10.1016/j.procs.2020.12.027]
111. Myers SP, Dasari M, Brown JB, et al. Effects of gender bias and stereotypes in surgical training: a randomized clinical trial. JAMA Surg 2020 Jul 1;155(7):552-560. [doi: 10.1001/jamasurg.2020.1127] [Medline: 32432669]
112. Milam LA, Cohen GL, Mueller C, Salles A. Stereotype threat and working memory among surgical residents. Am J Surg 2018 Oct;216(4):824-829. [doi: 10.1016/j.amjsurg.2018.07.064] [Medline: 30249337]
113. Lynch J, Aughwane P, Hammond TM. Video games and surgical ability: a literature review. J Surg Educ 2010;67(3):184-189. [doi: 10.1016/j.jsurg.2010.02.010] [Medline: 20630431]
114. Chalhoub E, Tanos V, Campo R, et al. The role of video games in facilitating the psychomotor skills training in laparoscopic surgery. Gynecol Surg 2016 Nov;13(4):419-424. [doi: 10.1007/s10397-016-0986-9]
115. Patterson F, Knight A, Dowell J, Nicholson S, Cousans F, Cleland J. How effective are selection methods in medical education? A systematic review. Med Educ (Chicago Ill) 2016 Jan;50(1):36-60. [doi: 10.1111/medu.12817]
116. Goldberg SB, Tanana M, Imel ZE, Atkins DC, Hill CE, Anderson T. Can a computer detect interpersonal skills? Using machine learning to scale up the Facilitative Interpersonal Skills task. Psychother Res 2021 Mar;31(3):281-288. [doi: 10.1080/10503307.2020.1741047] [Medline: 32172682]

Abbreviations

GBA: game-based assessment
ICC: intraclass correlation coefficient
mini-IPIP: short version of the International Personality Item Pool
MIS: minimally invasive surgery
MRT: Mental Rotation Test
PPT: Purdue Pegboard Test
RAPM: Raven's Advanced Progressive Matrices
USMLE: United States Medical Licensing Examination
VR: virtual reality

Edited by B Lesselroth; submitted 16.Aug.2025; peer-reviewed by M Iso-Mustajarvi, S Papadakis, TE Carver, T Mühling; revised version received 08.Mar.2026; accepted 07.Apr.2026; published 25.May.2026.

Please cite as:
Gazit N, Ben-Gal G, Eliashar R
Virtual Reality and Gamification for Assessing Technical Aptitude, Cognitive Abilities, and Personality Characteristics in Surgical Residency Selection: Validation Study
JMIR Med Educ 2026;12:e82515
URL: https://mededu.jmir.org/2026/1/e82515
doi:10.2196/82515

© Noa Gazit, Gilad Ben-Gal, Ron Eliashar. Originally published in JMIR Medical Education (https://mededu.jmir.org), 25.May.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Education, is properly cited. The complete bibliographic information, a link to the original publication on https://mededu.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Ambiguity Detection in Medical Exams via Large Language Models: Retrospective Cross-Sectional Pilot Study

Romain Lombardi[1,2], MSc, MD; Alexandre Destere[3,4], PharmD, PhD; Jean Dellamonica[1,2,5], MD, PhD; Alexandre O Gérard[3], MD, PhD; Mathieu Jozwiak[2,5], MD, PhD

[1]Critical Care Unit, Pasteur 2 University Hospital, 30 Voie Romaine, Nice, France

[2]UR2CA, Unité de Recherche Clinique Côte d'Azur, Université Côte d'Azur, Nice, France

[3]Department of Clinical Pharmacology and Pharmacovigilance Center, Medical Centre, Université Côte d'Azur, Nice, France

[4]Maasai Team, Laboratoire J.A. Dieudonné, Université Côte d'Azur, Inria, CNRS, Nice, France

[5]Critical Care Unit, Archet 1 University Hospital, Nice, France

Corresponding Author:
Romain Lombardi, MSc, MD
Critical Care Unit, Pasteur 2 University Hospital, 30 Voie Romaine, Nice, France

Abstract

Background: Large language models (LLMs) have emerged as promising tools in medical education due to their ability to understand, generate, and reason with natural language. Their ability to simulate expert reasoning suggests a potential for supporting quality control in assessment design. In this study, the use of LLMs in identifying ambiguous or poorly constructed exam items in critical care academic assessments was evaluated.

Objective: The study aimed to develop automated ambiguity and quality scores to objectively assess individual questions and entire exam components.

Methods: We analyzed 264 questions from academic exams conducted over 3 academic years (2023-2025) at the Medical School of Université Côte d'Azur. Questions were drawn from 4 docimological formats: progressive clinical cases (PCC), mini-PCC, key feature problems, and isolated question sequences (IQS). Each element was submitted to 4 LLMs (ChatGPT, Gemini Pro, Le Chat, and DeepSeek) without prompt engineering. Performance was evaluated using the official correction key. We applied 4 binary diagnostic tags based on model agreement and self-reported ambiguity: ambiguity, low performance, incoherence, and subjective ambiguity. These tags generated a composite ambiguity score and contributed to a weighted quality score for each exam component.

Results: LLMs achieved mean scores in the same range as students, with no significant differences across academic years and significantly higher performance on the mini-PCC and IQS formats (P=.049 and P=.04, respectively). IQS items had the highest ambiguity scores (54 items received a score of 2 in both 2023 and 2024, and 53 items retained the same score). Tag patterns revealed frequent issues with ambiguity and inconsistency. Quality scores varied across academic years. IQS predominantly showed moderate ambiguity (score 2), with occasional instances of strong signals. There was no significant difference in quality based on author specialty or seniority (P=.08 and P=.44, respectively).

Conclusions: In this pilot study, LLMs may offer a preliminary framework to proactively detect ambiguous exam questions and estimate the overall quality of an exam. Integrating these tools into the assessment design process could potentially reduce the need for postexam corrections and may help improve fairness and clarity in medical evaluations.

(JMIR Med Educ 2026;12:e82702) doi:10.2196/82702

KEYWORDS

large language model; LLM; medical education; exams; ambiguity detection; automated scoring; quality assessment; docimology; artificial intelligence; critical care; emergency

Introduction

Large language models (LLMs) are a class of artificial intelligence (AI) systems based on neural network architectures designed for natural language understanding and generation [1]. To function efficiently and deliver robust performance, LLMs are trained on massive amounts of textual data. They are designed to identify and learn recurrent patterns embedded in natural language [2]. LLMs have become an integral part of daily applications. Chatbot platforms such as ChatGPT, Google Gemini, DeepSeek, and Mistral are leading examples of the widespread use of natural language interactions [3-7]. LLMs are also extensively applied in machine translation, with systems like DeepL and Google Translate providing accurate and

XSL•FO
RenderX

context-aware translations [8]. Furthermore, these models play a growing role in education by generating automated exercises (eg, Socratic) and in software development by providing intelligent code assistance (eg, GitHub Copilot) [9].

LLMs have already been studied in a wide range of medical applications across various disciplines, and they appear to have a promising future [1,2]. Numerous studies in the fields of intensive care and perioperative medicine have investigated the various applications of LLMs. For example, Chung et al [10] demonstrated how these models can be used for perioperative risk stratification. These models can also be used to generate databases from the daily unstructured medical data in the medical sheets of critically ill patients [11]. LLMs can also serve as valuable tools for triaging patients and prioritizing those requiring emergency department care based on severity [12]. In more unconventional applications, their use has been explored for generating recommendations on the most relevant clinical topics. However, the results remain mixed and require expert oversight [13]. Their increasing use in clinical reasoning necessitates both evaluation and rigorous testing to ensure their validity.

LLMs are increasingly being explored as important tools in the fields of pedagogy and medical education [14,15]. These models can be used to train anesthesia residents by generating relevant information and synthesizing content from multiple sources [16]. They can assist with clinical diagnosis, which is particularly valuable for students in the early stages of their medical education [17]. One of the major limitations of clinical examination simulation is the limited availability of standardized patients. ChatGPT may offer a potential solution by serving as a substitute for human actors [18]. AI could enhance not only medical students' learning but also the evaluation and quality assurance of medical assessments. Maitland et al [19] attempted to have LLMs sit the Membership of the Royal Colleges of Physicians written examinations in the United Kingdom. The results showed that LLMs outperformed human candidates. The fine-tuned models also enabled the identification of 8 types of errors, such as factual errors, context errors, and omission errors. This offers a potential way to improve the quality of exam questions. In a preliminary study, Gérard et al [20] evaluated local and European pharmacology exams, highlighting how LLMs can help in identifying ambiguous questions.

The primary objective of this study was to develop a methodology for detecting ambiguous or low-quality questions in academic exams using various LLMs.

Methods

Objectives

We used LLMs as expert reasoning agents to simulate medical students' behavior and identify exam questions that might be poorly constructed or ambiguous. The secondary objectives were to develop an ambiguity scoring system and a quality score at the case level (ie, aggregating multiple questions), and to analyze quality variations according to academic year, assessment format, and the author's academic title.

Data Sources and Study Population

We analyzed academic exams taken by medical students at the Université Côte d'Azur. Only questions from the "Emergency and Intensive Care" course unit were included in this study. The questions were written either by an intensivist or an emergency physician, and the author was either a clinical fellow, an associate professor, or a full professor. A total of 3 academic exam cohorts were analyzed: 2023, 2024, and 2025. An exam was composed of several docimological components, and each component consisted of multiple questions, such as multiple-choice questions (MCQs), short-answer questions (SAQs), extended matching questions, or single-best-answer questions. The docimological components consist of four types: (1) progressive clinical cases (PCC; extended clinical scenarios with sequenced MCQs), (2) key feature problem (KFP; short critical decision-making vignettes), (3) mini-PCC (mPCC; condensed clinical cases), and (4) isolated question sequence (IQS; standalone items [MCQs or SAQs]) [21].

The year 2025 corresponded to the evaluation of master's level 2 (M2) students, while the years 2024 and 2023 corresponded to evaluations of both master's level 1 (M1) and M2 students. Additionally, the 2023 cohort also included evaluations of master's level 3 (M3) students. Each docimological component varied across the academic years.

Each exam consisted of between 3 and 8 different docimological components. A total of 264 questions were analyzed. All the exam scores were normalized to a 20-point scale. Partial credit was awarded based on the number of discrepancies between the participant's response and the official answer key, specifically false positives or incorrect options omitted. Participants received a score of 0.5 for 1 discrepancy, 0.2 for 2 discrepancies, and no points if more than 2 discrepancies were identified. A minimum score of 10 out of 20 was required to pass the exam.

The official correction and scoring rubric, the global performance scores of students per question and per docimological component, the responses and performance scores from the LLMs, and manual annotations from a human investigator were available for all questions.

To ensure consistency and minimize bias, all questions were carefully standardized in their format and presentation. No alterations or prompt engineering were applied to either the questions or answer options, allowing for an impartial assessment of the LLMs' inherent capabilities. The exam questions, extracted from PDF files, were submitted to each LLM, which was asked to choose the appropriate answer relying solely on its prior knowledge and training.

LLMs Evaluation Procedure and Ambiguity Detection

We analyzed the results of the 4 most prominent LLMs: ChatGPT-4, Gemini 2.5 Pro, Mistral Le Chat, and DeepSeek R1. Each exam question was submitted to the LLMs in two formats: (1) full-context input, simulating access to the complete clinical case; and (2) sequential-input, where questions were revealed one by one, mimicking an exam scenario. We deliberately adopted a zero-shot prompting strategy without task-specific parameter tuning or extensive prompt engineering. The rationale was to simulate a real-world usage scenario where

XSL•FO
RenderX

educators or examiners interact with LLMs as "out-of-the-box" tools without requiring specialized expertise in AI engineering, and to assess how the intrinsic phrasing of each test item naturally triggers consistency or divergence across different model architectures, minimizing the risk of over-tuning the models.

LLM responses were graded manually by the investigator according to the official correction, using the same scale applied to student answers. We also asked each LLM to identify any ambiguity in the question based on its own defined criteria. From these scores, three metrics were computed per question: the mean score of the LLMs, the SD of the LLMs' scores, and the number of LLMs that failed to answer correctly. Of note, no significant differences were observed between the 2 input formats.

We implemented a tagging system to assess the quality and clarity of each question. This system assigned binary diagnostic tags based on LLM performance patterns. These tags were designed to identify specific aspects of flawed question design and were applied automatically to each question to minimize the impact of subjective human bias. The tags used are as follows:

- Ambiguity tag: assigned when multiple LLMs provide incorrect answers with a diversity of distinct wrong responses, indicating disagreement in correctness, reasoning, or interpretation of the question. The tag was set to 1 if at least 2 LLMs answered incorrectly with at least 2 different incorrect answers among them.
- Low performance tag: this tag indicates that the question is challenging or potentially flawed due to universally poor performance by the LLMs. The tag is assigned if the mean normalized score of all LLMs is below 0.5, combined with low variance among their scores (SD<0.3). The low performance tag aims to identify items that are potentially "unanswerable" or structurally flawed. While difficulty and ambiguity are distinct, a universal failure of multiple expert-level LLMs with low variance often points toward a lack of clarity or a misleading structure rather than intended pedagogical selectivity.
- Incoherence tag: captured inconsistency in LLMs' understanding of the question. The tag was assigned when the measure of dispersion in LLM scores exceeded 0.3 (SD>0.3).
- Subjective ambiguity detection is assigned when the LLMs themselves identify ambiguity in the question based on evaluation criteria that they have autonomously established. The tag was set to 1 if at least 2 LLMs reported the presence of ambiguity according to their self-defined standards.

Table 1 summarizes the different tags and their assignment criteria.

XSL•FO
RenderX

Table . Summary of definitions and criteria for tag attribution.

Tag names	Definition	Operational criteria	Rationale
Ambiguity tag	Multiple LLMs[a] provide incorrect answers with a diversity of wrong responses.	Tag=1 if at least 2 LLMs answer incorrectly and if at least 2 different answers	• Reducing false positives arising from multiple LLMs converging on the same wrong answer reflects a shared misconception rather than true ambiguity. • The diversity of wrong answers suggests multiple plausible interpretations or poorly defined answer keys.
Low performance tag	The question is challenging or potentially flawed due to universally poor performance by the LLMs	Tag=1 if the mean normalized score of all LLMs is below 0.5 and SD<0.3	• Combined criterion distinguished truly difficult but valid questions from ambiguous or flawed questions (displaying greater variability or discrepancy between LLMs and student results).
Incoherence tag	Captures inconsistency in LLMs' understanding of the question	Tag=1 if SD>0.3	• Prevents outliers from skewing variability measures and ensures that incoherence reflects genuine divergence in interpretation rather than isolated anomalies.
Subjective ambiguity detection tag	Assigned when LLMs identify ambiguity based on their own evaluation criteria	Tag=1 if at least 2 LLMs report the presence of ambiguity	• This approach leverages the LLMs' capacity for self-assessment. Requiring agreement from multiple models helps ensure that the detected ambiguity is not idiosyncratic to a single LLM.

[a]LLM: large language model.

Each question was automatically tagged with the 4 binary flags described earlier. The sum of these tags formed a composite ambiguity score ranging from 0 to 3 (ambiguity score = ambiguity tag + low performance tag + incoherence tag + subjective ambiguity detection tag).

A score of 0 indicated "No issues detected," a score of 1 indicated "Minor concern," a score of 2 indicated "Moderate ambiguity," and a score of 4 indicated "Strong signal of item flaw or misleading structure." The low performance and incoherence tags are mutually exclusive, as they depend on opposing SD thresholds.

The thresholds for flagging potentially problematic questions were systematically determined using a combination of visual inspection and data-driven analysis, specifically elbow-point detection, to ensure that tag assignments were objectively grounded in the performance distribution. The elbow-point detection technique identifies the inflection point in the sorted distribution of each metric, representing the transition from typical to anomalous values.

Quality Assessment for Docimological Component

To calculate the quality score for each docimological component, we first determined the average score obtained by the LLMs for each question. We then applied a weighting based on the ambiguity score of each question according to the following formula:

$$(1) Qi=20\times Si$$

where Q_i is the quality score for question i (on a 20-point scale), and S_i, ranging from 0 to 1, is the average score of LLMs for question i.

$$(2) Scorecomponent=\sum i=1NQi\times wi\sum i=1Nwi$$

where N is the total number of questions, Q_i is the quality score for question i (on a 20-point scale), and w_i is the weight assigned to question i with:

$$(3) wi=11+Ai$$

where w_i is the weight assigned to question i, and A_i is the ambiguity score for the question i.

Using this system, we were able to assign a qualitative description to each docimological component based on the score obtained, as follows:

- If score ∈ [18-20]: excellent, clear, precise, and unambiguous writing.
- If score ∈ [15-18]: very good, well-structured writing with minor ambiguities.

XSL•FO
RenderX

- If score ∈ [12-15]: moderate, understandable writing, but with recurring ambiguities.
- If score ∈ [10-12]: poor, confused writing with repeated ambiguities.
- If score ∈ [0-10]: insufficient, vague, and ambiguous writing impacts reliability.

Components with a quality score below 15 were re-evaluated by the authors to clarify the identified ambiguities.

Statistical Analysis

The results obtained by the students were expressed in terms of mean, median, and maximum scores, whereas the LLMs' results were presented as mean only. We used the Mann-Whitney *U* test to compare the performance of LLMs and students, with the results expressed in terms of *P* values. We compared the quality score results according to the writer's medical specialty, professional career level (clinical fellow, associate professor, or full professor), and the type of docimological component. A Student 2-tailed *t* test was used when the distribution was normal, as assessed by the Shapiro-Wilk test; otherwise, a Mann-Whitney *U* test was applied. We considered a *P* value less than .05 to be statistically significant.

Ethical Considerations

All materials analyzed were anonymized. No individual student responses were accessed. LLMs were used only to simulate reasoning without personalization. According to French law (Articles L1121-1 et seq. of the French Public Health Code) [22], this study did not qualify as biomedical research involving human participants. Since it involved neither interventions on individuals nor the processing of identifiable personal data, it was exempt from ethics committee review. Consequently, under French regulations [23], this type of research did not require approval from an institutional ethics board.

Development Environment

All the analyses and development were performed in Python (Python Software Foundation) version 3.13.3, along with the following libraries and their respective versions: *pandas* 2.2.3, *NumPy* 2.2.6, *seaborn* 0.13.2, *matplotlib* 3.10.3, *scikit-learn* 1.6.1, *statsmodels* 0.14.4, *SciPy* 1.15.3, and *UpSetPlot* 0.9.0.

Results

Performance Comparison

A total of 264 exam questions were analyzed and submitted to the 4 LLM models (ChatGPT, Gemini Pro, Le Chat, and DeepSeek). Students achieved relatively consistent results over the years, with average scores ranging from 11.59 (SD 2.18) to 14.0 (SD 1.80) (Table 2). Performance was higher on KFPs for students. The results obtained by LLMs across academic years were heterogeneous. Superior performance was achieved by ChatGPT and Gemini Pro. No significant difference was observed between the LLMs' results and the students' results when comparing across academic years. However, when analyzing performance by docimological component, the LLMs outperformed students in the mPCC and IQS formats (*P*=.049 and *P*=.04; respectively).

XSL•FO
RenderX

Table . Performance comparison between large language models (LLMs) and students' scores. For the LLMs, only the mean score is reported.

	Students			LLM model				*P* value
	Median (IQR)	Mean (SD)	Min-max	ChatGPT	Gemini Pro	Le Chat	DeepSeek	
Academic level								
2023								
M1[a]	11.85 (10.09-13.15)	11.59 (2.18)	5.35-16.55	15.28	18.44	12.28	11.8	.05
M2[b]	12.80 (11.68-13.71)	12.64 (1.48)	7.39-18.42	14.11	16.77	12.32	12.27	.39
M3[c]	12.53 (11.62-14.03)	11.77 (4.19)	20 (0-20)	—[d]	—[d]	—[d]	—[d]	.40
2024								
M1	14.14 (12.87-15.25)	14.0 (1.80)	7.54-15.82	15.91	14.36	15.75	14.30	.17
M2	13.07 (11.96-14.15)	13.08 (1.63)	8.40-17.36	12.88	16.30	13.21	12.52	.61
2025								
M2	11.68 (10.61-12.74)	11.67 (1.56)	7.35-16.46	11.32	12.38	11.79	11.96	.74
Docimological component								
PCC[e]	12.67 (10.18-14.86)	12.07 (4.49)	0.0-20	13.32	16.40	13.10	11.81	.16
mPCC[f]	11.85 (8.35-15.0)	11.35 (5.26)	0.0-20	13.16	15.14	13.00	12.78	.049
KFP[g]	13.33 (8.0-15.33)	11.91 (5.71)	0.0-20	14.95	14.19	14.95	14.19	.16
IQS[h]	10.14 (9.3-13.40)	10.63 (4.15)	0.0-18.4	14.14	16.22	12.71	12.49	.04

[a]M1: master's level 1.

[b]M2: master's level 2.

[c]M3: master's level 3.

[d]The exam cases provided to M2 and M3 students during faculty examinations are strictly identical.

[e]PCC: progressive clinical cases.

[f]mPCC: mini–progressive clinical cases.

[g]KFP: key feature problem.

[h]IQS: isolated question sequence.

Ambiguity Score and Tag Repartition

Figure 1 shows the distribution of ambiguity scores by item according to academic year and docimological component. An ambiguity score of 2 was the most frequently observed, regardless of academic year or type of docimological component. Specifically, 54 items received a score of 2 in both 2023 and 2024, while 53 items had the same score in the IQS category, and 37 and 34 items in the PCC and mPCC categories, respectively. The 2024 academic year had the highest number of ambiguity scores at 0, and there was no score greater than 2 in 2023 and 2025. IQSs exhibited the highest number of scores above 2, with a total of 5 items scoring 3, but also the highest number of scores at 0, with 36 questions in total.

For the academic year 2023, the low-performance and ambiguity tags were over-represented compared to the tags inconsistency and subjective ambiguity detection, regardless of the docimological component studied (Figure 2). The total ambiguity score was higher for IQS and PCC, with scores of 35 and 32, respectively. An increase in the number of subjective ambiguity detection tags was observed in 2024, while the distribution of the other tags remained unchanged. During that year, the total ambiguity score for mPCC increased to 27, compared to 10 in 2023. No PCC case was included that year. KFPs had the lowest total ambiguity score during the same period. In 2025, the various tags appeared less frequently than in previous years. The subjective ambiguity detection tag was not used, and the maximum total ambiguity score for the PCC cases was 18.

Figure S1 in Multimedia Appendix 1 shows the most common tag combinations. The most frequent tag combinations were ambiguity and incoherence (78, 29.5% items), low performance

XSL•FO
RenderX

and ambiguity (52, 19.7% items), and incoherence, subjective ambiguity detection, and ambiguity (5, 1.9% items). The ambiguity tag alone appeared in 36 items, corresponding to 13.6% of cases, and the low performance tag was never found alone.

Figure 1. Distribution of ambiguity scores by item according to (A) academic year and (B) docimological component. No scores greater than 3 are shown, as the tagging criteria make this situation impossible. IQS: isolated question sequence; KFP: key feature problem; mPCC: mini–progressive clinical cases; PCC: progressive clinical cases.

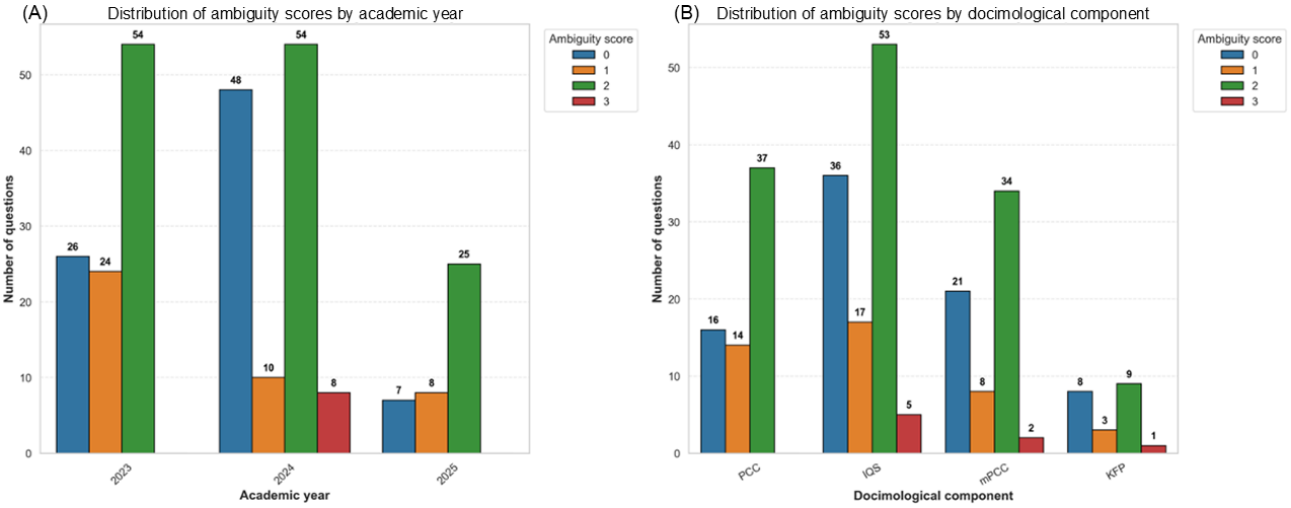
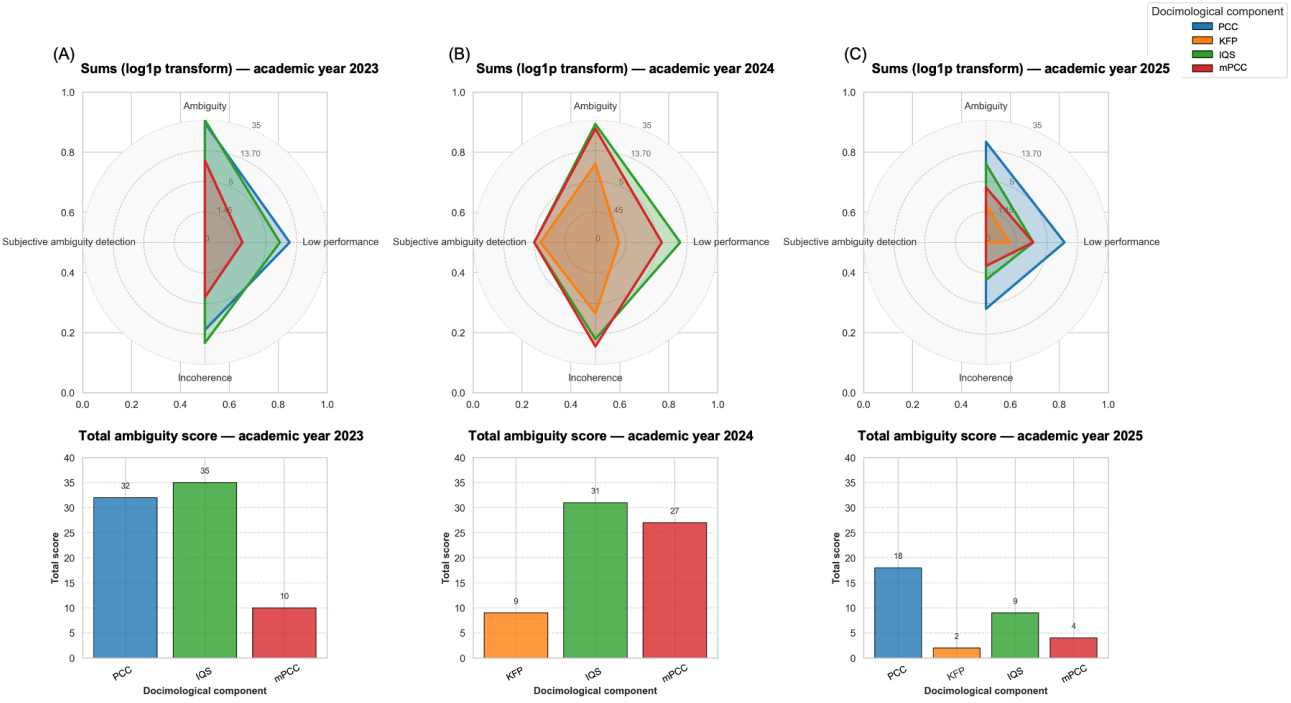


Figure 2. Tag repartition by docimological component according to (A) academic year 2023, (B) academic year 2024, and (C) academic year 2025. To increase the visibility of the differences between the distributions of the different tags, the axes of the radar plots were transformed using log1p. The total ambiguity score results from the sum of all the ambiguity scores. IQS: isolated question sequence; KFP: key feature problem; mPCC: mini–progressive clinical cases; PCC: progressive clinical cases.

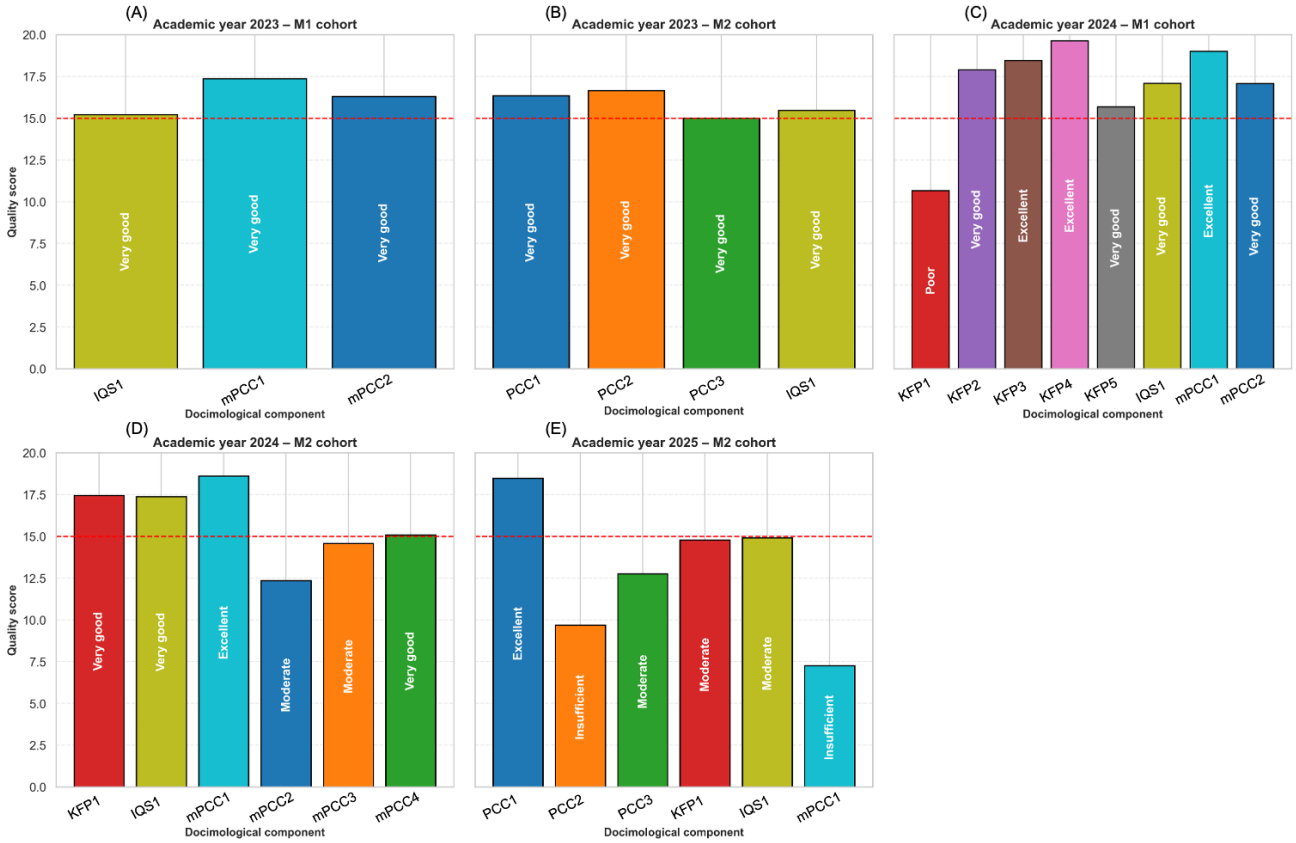


Quality Score

All the docimological elements for the faculty exams in 2023 exceeded the validity threshold, with a qualitative score of 15 out of 20, and all of them received a "Very good" rating (Figure 3). For the 2024 M1 cohort, only 1 component, KFP1, fell below the validity threshold with a qualitative rating of "Poor," whereas for the M2 cohort, 2 components were below this threshold with ratings of "Moderate." Four of 6 items did not reach the 15 out of 20 threshold for the 2025 M2 cohort, with the lowest scores observed for PCC2 and mPCC1, both receiving a qualitative rating of "Insufficient." Figure 3 also shows an expansion in the diversity of docimological element types over the years, with all formats (PCC, mPCC, KFP, and IQS) represented in 2025. In 2023, and for the M1 academic exam, only the IQS and mPCC formats were included. IQSs were represented in each academic exam. Elements with a score of 15 were reviewed by the authors, who confirmed the presence of ambiguities in the wording, structure, or context.

XSL•FO
RenderX

Figure 3. Qualitative score distribution by academic year and cohort. (A) 2023—M1 cohort, (B) 2023—M2 cohort, (C) 2024—M1 cohort, (D) 2024—M2 cohort, and (E) 2025—M2 cohort. IQS: isolated question sequence; KFP: key feature problem; mPCC: mini–progressive clinical cases; PCC: progressive clinical cases.



No significant differences were found between the groups when comparing results based on quality scores according to the writer's specialty and career level, nor when comparing results by career level and the docimological component (Table S1 in Multimedia Appendix 2). No significant difference in career level was found between clinical fellows and associate professors (P=.44). The difference between emergency writers and intensive care unit or critical care unit writers appeared more pronounced, approaching statistical significance (P=.08). After adjustment for career level, the quality gap between mPCC and KFPs remained more marked than for the remaining docimological types (P=.12 and P=.16, respectively). The number of contributors involved in writing the exams increased during 2025, particularly with more clinical fellows participating.

Discussion

Principal Findings

In our comparison of LLMs with average student scores, no significant differences were found across academic years. However, mean scores varied between LLMs, and additional differences appeared when analyzing performance by type of docimological component. One reason for this variation may be the specific training approach and intended purpose of each model. ChatGPT and Gemini Pro are the oldest and most widely known models. A study focusing on the responses of various LLMs to medical questions related to tuberculosis highlighted the absence of overall performance differences between the models but revealed disparities depending on the type of content submitted [24]. It is important to emphasize that this study was not designed to establish the superiority of 1 specific LLM over others. Rather, building upon the methodological framework previously described by our team in Gérard et al [20], our approach focuses on the synergy of an ensemble of models. Using multiple LLMs allows for a convergence of evidence, which increases the reliability of ambiguity detection by filtering out idiosyncratic errors or "hallucinations" from individual models.

There was an unequal distribution of tag types and ambiguity scores depending on the academic year and the type of assessment elements. The highest ambiguity score was observed for IQSs. One potential explanation for the disparity in distribution across academic years, though not statistically confirmed, might be differences in authorship. In 2024 and 2025, a greater number of contributors, particularly clinical fellows, were involved in writing the exams. While these younger authors are often tasked with drafting clinical case scenarios, our analysis did not reach statistical significance to definitively confirm a difference in quality based on academic rank. This lack of significance is likely due to a lack of statistical power when subdividing the 264 items by year, format, and seniority. This idea is supported by several studies. For example, Oyibo et al [25] examined a cohort of 65 junior doctors and found that young professionals produced case scenarios that were rated as effective and clear in 98.5% of cases. Additionally, they were enthusiastic about the writing process. Furthermore, case writing is an effective way to develop advanced writing

XSL•FO
RenderX

skills, which are particularly useful for creating clinical case scenarios [26]. However, in our study, we did not find a statistically significant difference in the quality of clinical cases according to professional grade or specialty. Another explanation can be found in the type of docimological component. The construction of decontextualized single-question items is challenging and leads to a lack of validity and low rigor. Several authors describe IQSs as poor evaluators that encourage memorization rather than critical thinking, making them less discriminative [27,28].

Our results suggest that quality scores could provide a valuable framework for evaluating the ambiguity and overall quality of academic exam cases *a priori*. Unlike current methods that assess items *a posteriori* based on student performance, our approach aims to identify "structural fragility" before the exam is administered. Most evaluations of these components are currently conducted retrospectively, after the exam has been administered [29]. These items may be removed or may affect the exam's discriminatory power, leading to student frustration. Their removal is costly in terms of time and budget and undermines educators' credibility and trust among students. Defining a score before submission could provide valuable feedback and may facilitate proactive efforts to improve clarity and revise ambiguous items [30,31].

This work highlights the growing importance of LLMs in the evaluation and validation of academic exams. This study extends the work initiated by Gérard et al [20] in the domain of pharmacological exams by incorporating a rigorous methodology for question evaluation, enabling real-time improvement. In contrast to their findings, our results revealed fewer statistically significant differences between the performance of LLMs and medical students. Our study encompassed a wider array of question formats, including radiological interpretation tasks and docimological components, which were not addressed in the original study. Recent studies have shown that LLMs can do more than just detect ambiguity; they can also analyze complex medical questions, identify nuances in item clarity, and provide consistent evaluative feedback [32,33]. Furthermore, LLMs have been explored as valuable tools for designing and revising educational assessments. They improve item quality and reduce ambiguity through iterative review processes [19]. Our study presented a practical application of LLMs in an understudied field of medical education. LLMs were not used to replace students in taking exams but to offer a promising approach to addressing ambiguities and docimological errors. Our methodology adopts a "human-in-the-loop" approach, where LLMs serve as automated screening tools rather than final judges of construct validity. By flagging questions with low performance or high ambiguity, the system is intended to assist pedagogical experts in concentrating their review on items most likely to contain docimological errors. This synergy between automated detection and human expertise ensures that the distinction between a "challenging but fair" question and an "ambiguous" one is ultimately determined by the educator.

Given the pilot nature of this study and the specific context of critical care assessments, these results should be interpreted with caution. While the trends observed are encouraging, they represent preliminary insights that require validation through larger, multicentric studies across diverse medical disciplines.

Strengths

This was an original research study in terms of both its thematic focus and its methodological design. We introduced a fully objective tagging methodology aimed at reducing the risk of bias associated with human interpretation. This approach has the potential to be applied beyond the fields of critical care and emergency medicine. The academic exams at the Medical School of Université Côte d'Azur incorporated a variety of docimological formats, including KFP, PCC, mPCC, IQS, MCQ, single-best-answer questions, and SAQ. This study enabled an evaluation of this broad spectrum and an analysis of the various types of errors that may occur during the design of such items. Including this diversity enhances external validity and supports its applicability in real-world educational settings. Comparative analyses with students are often lacking in AI-related studies. Similar to the approach taken by Gérard et al [20], conducting such an analysis provides additional insight into the difficulty level of the assessed questions. This approach allows for a type of *a priori* evaluation, in which the exams are assessed before students take them. This is different from *a posteriori* methods, which rely on student performance data. Thus, this *a priori* evaluation could significantly reduce the logistical and human resources required for postexamination modifications and reviews while supporting fairness among students. To ensure maximum objectivity, we selected the four most widely used LLMs: ChatGPT, Gemini, Le Chat, and DeepSeek. This choice reduced the risk of systematic errors inherent to a single model and was intended to enhance the robustness of the results by converging ambiguity detection across multiple models. Furthermore, no prompt engineering was used to maintain the most realistic and standardized conditions possible.

Limitations

One of the main limitations of this study was the absence of a systematic human gold standard for the entire dataset. Consequently, we must acknowledge that some ambiguities identified by the LLMs may reflect intrinsic model limitations or reasoning issues rather than actual flaws in the questions. To mitigate this risk of "model noise," we used an ensemble strategy where ambiguity is only flagged upon the disagreement of multiple independent architectures. Rather than definitive proofs of error, these scores should be viewed as objective indicators of an item's structural fragility, designed to guide educators toward the questions most in need of manual refinement. Nevertheless, we used four different LLMs to minimize this potential bias. We only used general-purpose models for the analysis of exam questions, although it is known that domain-specific LLMs, such as Med-PaLM 2, exist [34]. These specialized models may demonstrate superior reasoning abilities in complex clinical scenarios and a greater capacity to detect ambiguities in medical questions. However, this would reduce the generalizability of the methodology to domains outside of medicine. Furthermore, the absence of task-specific tuning and the reliance on the generative intuition of LLMs may introduce certain biases. As this study did not use few-shot examples or chain-of-thought prompting to standardize the

XSL•FO
RenderX

reasoning process, the "ambiguity scores" should be interpreted as indicators of potential fragility rather than definitive proofs of item flaws. Future research should investigate whether advanced prompting techniques or fine-tuning on human-validated "gold standards" can further isolate model noise from item-specific ambiguity. The monocentric nature of the study population, which is limited to the field of intensive care and emergency medicine, may limit the generalizability of the results. However, the methodology is transferable and can be applied to other specialties. To minimize this bias and avoid focusing on a single student cohort, we included multiple academic years, thereby introducing variation among authors. The use of the "subjective ambiguity" tag may have led to confusion and introduced interpretation bias, as its definition relies on LLMs' ability to detect ambiguity. This undermines transparency, as the criteria were defined by the models themselves and were not externally verifiable, thereby increasing the risk of a "black box" effect. While our thresholds were derived from data-driven inflection points, we acknowledge that sensitivity analyses across different medical specialties would further validate the stability and generalizability of these assignments. Furthermore, our analysis of quality based on author seniority and specialty may have been limited by a lack of statistical power. Although the total sample of 264 questions is substantial, the fragmentation into multiple categories (academic years, docimological formats, and authors) resulted in small effect sizes that were insufficient to reach statistical significance.

Future Work

To ensure the external validity of this methodology, our next step involves a large-scale, multicenter validation study spanning multiple medical specialties. Future research will focus on the external validation of these scores by correlating LLM-generated fragility signals with student psychometric data, such as bimodal response distributions or item discrimination indices. This will allow us to establish the predictive validity of our *a priori* tool against the *a posteriori* gold standard of student performance. Both qualitative and quantitative assessments of examinations will be implemented before and after the validation tool is deployed. Once validated, the system will be adopted for the *a priori* evaluation of faculty examination quality at the local level using LLMs. An online platform will provide easy access to this solution, facilitating its integration into academic routines.

Conclusions

Our findings suggest a potential role that LLMs can play in medical education. We analyzed 264 exam questions across 3 academic years using four different LLMs. A standardized evaluation process was applied, combining automated scoring with 4 binary diagnostic tags to generate composite ambiguity and weighted quality scores. This methodology enabled a more standardized assessment of both individual questions and entire exam components. Overall, the models achieved performance levels comparable to students, with significantly higher scores on mPCC and IQS formats. IQS items exhibited the highest ambiguity scores, frequently associated with incoherence and low-performance tags. Ambiguity scores varied by year, with a predominance of moderate ambiguity (score 2) and occasional strong signals in IQS. Across all cohorts, IQS items were consistently represented, while the 2025 cohort included the full range of docimological formats. No significant differences in exam quality were observed based on the authors' specialty or academic rank. Given the exploratory and monocentric nature of this pilot study, these results represent hypothesis-generating observations that necessitate further prospective validation.

LLMs act as intelligent assistants in a human-in-the-loop framework, flagging potential issues for expert review to improve the overall clarity and fairness of medical assessments. As these technologies evolve, more research is needed to explore their applications across disciplines and institutions.

Acknowledgments

The authors would like to thank the medical students of Université Côte d'Azur, whose examination results were analyzed in this study. The authors are also grateful to the professors and instructors who generously provided their examinations and official answer keys, which were essential to our analyses.

Funding

No external financial support or grants were received from any public, commercial, or not-for-profit entities for the research, authorship, or publication of this article.

Data Availability

The datasets generated or analyzed during this study are not publicly available due to legal restrictions prohibiting the dissemination of faculty examination materials; however, they are available from the corresponding author upon reasonable request.

Authors' Contributions

Conceptualization: RL, MJ, AOG, AD, JD
Data curation: RL, MJ, AOG, AD, MJ, JD
Formal analysis: RL
Investigation: RL, MJ
Methodology: RL, MJ, AOG

XSL•FO
RenderX

Supervision: MJ, AOG, AD, JD
Validation: MJ, AOG, AD, JD
Writing – original draft: RL
Writing – review & editing: MJ, AOG, AD, JD

Conflicts of Interest

None declared.

Multimedia Appendix 1
Upset plot of tag combinations.
[PNG File, 279 KB - mededu_v12i1e82702_app1.png]

Multimedia Appendix 2
Comparison of quality scores between career levels, specialties, and docimological components.
[XLSX File, 10 KB - mededu_v12i1e82702_app2.xlsx]

References

1. Omiye JA, Gui H, Rezaei SJ, Zou J, Daneshjou R. Large language models in medicine: the potentials and pitfalls: a narrative review. Ann Intern Med 2024 Feb;177(2):210-220. [doi: 10.7326/M23-2772] [Medline: 38285984]
2. Lu Y, Aleta A, Du C, Shi L, Moreno Y. LLMs and generative agent-based models for complex systems research. Phys Life Rev 2024 Dec;51:283-293. [doi: 10.1016/j.plrev.2024.10.013] [Medline: 39486377]
3. Ouyang L, Wu J, Jiang X, et al. Training language models to follow instructions with human feedback. Presented at: 36th Conference on Neural Information Processing Systems (NeurIPS 2022); Nov 28 to Dec 9, 2022. [doi: 10.52202/068431-2011]
4. Gemini Team, Anil R, Borgeaud S, et al. Gemini: a family of highly capable multimodal models. arXiv. Preprint posted online on Dec 19, 2023. [doi: 10.48550/ARXIV.2312.11805]
5. DeepSeek-AI, Liu A, Feng B, et al. DeepSeek-V2: a strong, economical, and efficient mixture-of-experts language model. arXiv. Preprint posted online on May 7, 2024. [doi: 10.48550/ARXIV.2405.04434]
6. Jiang AQ, Sablayrolles A, Mensch A, et al. Mistral 7B. arXiv. Preprint posted online on Oct 10, 2023. [doi: 10.48550/ARXIV.2310.06825]
7. Briganti G. How ChatGPT works: a mini review. Eur Arch Otorhinolaryngol 2024 Mar;281(3):1565-1569. [doi: 10.1007/s00405-023-08337-7] [Medline: 37991499]
8. Wu Y, Schuster M, Chen Z, et al. Google's neural machine translation system: bridging the gap between human and machine translation. arXiv. Preprint posted online on Sep 26, 2016. [doi: 10.48550/ARXIV.1609.08144]
9. Chen M, Tworek J, Jun H, et al. Evaluating large language models trained on code. arXiv. Preprint posted online on Jul 7, 2021. [doi: 10.48550/ARXIV.2107.03374]
10. Chung P, Fong CT, Walters AM, Aghaeepour N, Yetisgen M, O'Reilly-Shah VN. Large language model capabilities in perioperative risk prediction and prognostication. JAMA Surg 2024 Aug 1;159(8):928-937. [doi: 10.1001/jamasurg.2024.1621] [Medline: 38837145]
11. Madden MG, McNicholas BA, Laffey JG. Assessing the usefulness of a large language model to query and summarize unstructured medical notes in intensive care. Intensive Care Med 2023 Aug;49(8):1018-1020. [doi: 10.1007/s00134-023-07128-2] [Medline: 37338549]
12. Shekhar AC, Kimbrell J, Saharan A, Stebel J, Ashley E, Abbott EE. Use of a large language model (LLM) for ambulance dispatch and triage. Am J Emerg Med 2025 Mar;89:27-29. [doi: 10.1016/j.ajem.2024.12.032] [Medline: 39675178]
13. Balta KY, Javidan AP, Walser E, Arntfield R, Prager R. Evaluating the appropriateness, consistency, and readability of ChatGPT in critical care recommendations. J Intensive Care Med 2025 Feb;40(2):184-190. [doi: 10.1177/08850666241267871] [Medline: 39118320]
14. Benítez TM, Xu Y, Boudreau JD, et al. Harnessing the potential of large language models in medical education: promise and pitfalls. J Am Med Inform Assoc 2024 Feb 16;31(3):776-783. [doi: 10.1093/jamia/ocad252] [Medline: 38269644]
15. Shool S, Adimi S, Saboori Amleshi R, Bitaraf E, Golpira R, Tara M. A systematic review of large language model (LLM) evaluations in clinical medicine. BMC Med Inform Decis Mak 2025 Mar 7;25(1):117. [doi: 10.1186/s12911-025-02954-4] [Medline: 40055694]
16. Guthrie E, Levy D, Del Carmen G. The operating and anesthetic reference assistant (OARA): a fine-tuned large language model for resident teaching. Am J Surg 2024 Aug;234:28-34. [doi: 10.1016/j.amjsurg.2024.02.016] [Medline: 38365551]
17. Hadi A, Tran E, Nagarajan B, Kirpalani A. Evaluation of ChatGPT as a diagnostic tool for medical learners and clinicians. PLoS One 2024;19(7):e0307383. [doi: 10.1371/journal.pone.0307383] [Medline: 39083523]
18. Wang C, Li S, Lin N, et al. Application of large language models in medical training evaluation-using ChatGPT as a standardized patient: multimetric assessment. J Med Internet Res 2025 Jan 1;27:e59435. [doi: 10.2196/59435] [Medline: 39742453]

XSL•FO
RenderX

19. Maitland A, Fowkes R, Maitland S. Can ChatGPT pass the MRCP (UK) written examinations? Analysis of performance and errors using a clinical decision-reasoning framework. BMJ Open 2024 Mar 15;14(3):e080558. [doi: 10.1136/bmjopen-2023-080558] [Medline: 38490655]
20. Gérard AO, Merino D, Labriffe M, et al. Evaluating and leveraging large language models in clinical pharmacology and therapeutics assessment: from exam takers to exam shapers. Br J Clin Pharmacol 2025 Jun 10;10. [doi: 10.1002/bcp.70137] [Medline: 40495266]
21. Cohen Aubart F, Lhote R, Hertig A, et al. Progressive clinical case-based multiple-choice questions: an innovative way to evaluate and rank undergraduate medical students. Rev Med Interne 2021 May;42(5):302-309. [doi: 10.1016/j.revmed.2020.11.006] [Medline: 33518414]
22. Code de la santé publique [Article in French]. Légifrance (RÉPUBLIQUE FRANÇAISE Liberté, Égalité, Fraternité). URL: https://www.legifrance.gouv.fr/loda/id/LEGISCTA000006190263 [accessed 2026-05-15]
23. Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing directive 95/46/EC (General Data Protection Regulation) (text with EEA relevance). : European Union; 2016 URL: https://eur-lex.europa.eu/eli/reg/2016/679/oj/eng [accessed 2026-05-15]
24. Dastani M, Mardaneh J, Rostamian M. Large language models' capabilities in responding to tuberculosis medical questions: testing ChatGPT, Gemini, and Copilot. Sci Rep 2025 May 23;15(1):18004. [doi: 10.1038/s41598-025-03074-9] [Medline: 40410343]
25. Oyibo SO. Developing a beginner's guide to writing a clinical case report: a pilot evaluation by junior doctors. Cureus 2019 Dec 13;11(12):e6370. [doi: 10.7759/cureus.6370] [Medline: 31886095]
26. Oyibo SO, Brij SO. A medical article publishing club for junior doctors: a quantitative and qualitative analysis. Cureus 2018 Dec 7;10(12):e3701. [doi: 10.7759/cureus.3701] [Medline: 30788190]
27. Schuwirth LWT, van der Vleuten CPM. Different written assessment methods: What can be said about their strengths and weaknesses? Med Educ 2004 Sep;38(9):974-979. [doi: 10.1111/j.1365-2929.2004.01916.x] [Medline: 15327679]
28. Walsh JL, Harris BHL, Smith PE. Single best answer question–writing tips for clinicians. Postgrad Med J 2017 Feb;93(1096):76-81. [doi: 10.1136/postgradmedj-2015-133893] [Medline: 27371033]
29. Rezigalla AA. Item analysis: concept and application. In: Firstenberg MS, Stawicki SP, editors. Medical Education for the 21st Century: IntechOpen; 2022:1-17. [doi: 10.5772/intechopen.100138]
30. Iñarrairaegui M, Fernández-Ros N, Lucena F, et al. Evaluation of the quality of multiple-choice questions according to the students' academic level. BMC Med Educ 2022 Nov 11;22(1):779. [doi: 10.1186/s12909-022-03844-3] [Medline: 36369070]
31. Boateng GO, Neilands TB, Frongillo EA, Melgar-Quiñonez HR, Young SL. Best practices for developing and validating scales for health, social, and behavioral research: a primer. Front Public Health 2018;6:149. [doi: 10.3389/fpubh.2018.00149] [Medline: 29942800]
32. Nori H, King N, McKinney SM, Carignan D, Horvitz E. Capabilities of GPT-4 on medical challenge problems. arXiv. Preprint posted online on Mar 20, 2023. [doi: 10.48550/ARXIV.2303.13375]
33. Singhal K, Tu T, Gottweis J, et al. Toward expert-level medical question answering with large language models. Nat Med 2025 Mar;31(3):943-950. [doi: 10.1038/s41591-024-03423-7] [Medline: 39779926]
34. Singhal K, Azizi S, Tu T, et al. Large language models encode clinical knowledge. Nature New Biol 2023 Aug;620(7972):172-180. [doi: 10.1038/s41586-023-06291-2] [Medline: 37438534]

Abbreviations

AI: artificial intelligence
IQS: isolated question sequence
KFP: key feature problem
LLM: large language model
MCQ: multiple-choice question
mPCC: mini–progressive clinical cases
PCC: progressive clinical cases
SAQ: short-answer question

XSL•FO
RenderX

Edited by R Pellegrino; submitted 20.Aug.2025; peer-reviewed by E Bai, JL Cracowski, L Haudebourg; revised version received 12.Jan.2026; accepted 20.Apr.2026; published 26.May.2026.

Please cite as:
Lombardi R, Destere A, Dellamonica J, Gérard AO, Jozwiak M
Ambiguity Detection in Medical Exams via Large Language Models: Retrospective Cross-Sectional Pilot Study
JMIR Med Educ 2026;12:e82702
URL: https://mededu.jmir.org/2026/1/e82702
doi:10.2196/82702

© Romain Lombardi, Alexandre Destere, Jean Dellamonica, Alexandre O Gérard, Mathieu Jozwiak. Originally published in JMIR Medical Education (https://mededu.jmir.org), 26.May.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Education, is properly cited. The complete bibliographic information, a link to the original publication on https://mededu.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Exploring the Feasibility of an Examiner-Worn Neck-Mounted Camera for Objective Structured Clinical Examination Assessment: Pilot Feasibility Study

Miwa Sekine[1], PhD; Yuji Nishizaki[1], MD, MPH, PhD; Amane Endo[1], MD, PhD; Takasuke Ogawa[1], MD, PhD; Yoshihide Takeshita[1], MD, PhD; Motomi Nasu[1], MD, PhD; Hiroo Wada[1], MD, PhD; Chizuko Miyamoto[2], PhD; Michiko Oguro[3], PhD; Chie Otake[2], MS; Yasuhiko Konishi[1], MD, PhD; Yuichi Tomiki[1], MD, PhD

[1]Division of Medical Education, Faculty of Medicine, Juntendo University, Bunkyo, Tokyo, Japan

[2]Chiba Faculty of Nursing, Tokyo Healthcare University, Funabashi City, Chiba, Japan

[3]Graduate School of Nursing Science, St. Luke's International University, Chuo City, Tokyo, Japan

Corresponding Author:
Yuji Nishizaki, MD, MPH, PhD
Division of Medical Education
Faculty of Medicine
Juntendo University
2-1-1 Hongo
Bunkyo, Tokyo, 113-8421
Japan
Phone: 81 3 3813 3111 ext 3548
Email: ynishiza@juntendo.ac.jp

Abstract

Background: The Objective Structured Clinical Examination (OSCE) is a prevalent method for evaluating clinical competence in medical education. As OSCEs become increasingly standardized and resource intensive, alternative evaluation methods are being explored, particularly because of the limited availability of certified examiners. However, few studies have investigated whether wearable technologies can support OSCE assessment. Wearable devices may provide a means of recording clinical skills from the examiner's perspective.

Objective: This pilot study, conducted in 2024, aimed to investigate the feasibility of using an examiner-worn neck-mounted camera for recording OSCE scenarios and to evaluate the evaluability of clinical performance from the recorded footage.

Methods: In total, 9 experienced medical educators participated in a simulated OSCE scenario involving electrocardiogram lead placement. All participants completed the initial live assessment and the postuse questionnaire, while 8 of 9 (89%) participants completed the subsequent video-based reassessment. Video recordings from both a fixed camera and a neck-mounted camera (THINKLET) were used to assess the evaluability of each OSCE item. Following a washout period, evaluators reassessed the neck-mounted camera recordings by using the original checklist, while fixed-camera recordings were used to judge the evaluability of each item. Agreement between live and video-based scoring was analyzed using percent agreement and the Cohen κ coefficient. A postevaluation questionnaire captured evaluators' experiences with the wearable device.

Results: Cohen κ ranged from 0.258 to 0.913 (mean 0.67, SD 0.20). Across checklist observations, more items were judged to be evaluable in the neck-mounted camera recordings than in the fixed-camera recordings, particularly for tasks requiring observation of fine motor skills. Evaluators reported generally positive experiences with the device, although some noted issues related to audio quality, comfort, posture restriction, and limited visibility at low angles.

Conclusions: Although further investigation is needed, this pilot study suggests that an examiner-worn neck-mounted camera may be a valuable supplementary assessment tool for selected OSCE tasks. Further work is needed to refine the device, standardize recording protocols, and clarify how it can best support review and verification alongside live evaluation.

(JMIR Med Educ 2026;12:e87483) doi:10.2196/87483

KEYWORDS

objective structured clinical examination; OSCE; wearable camera; medical education; assessment; feasibility study

XSL•FO
RenderX

Introduction

The Objective Structured Clinical Examination (OSCE) is a globally recognized method for evaluating clinical competence in medical education [1]. According to the pyramid of clinical competence by Miller [2], OSCEs correspond to the "shows how" level of assessment, which assesses practical performance in simulated environments. In Japan, the OSCE was officially introduced in 2005 following several trial phases [3], and its role has expanded significantly in response to the transition from observation-based to hands-on clinical clerkships under the student physician system, which allows medical students to participate directly in patient care [4]. Medical students are now required to demonstrate essential clinical skills and professional attitudes before commencing clinical clerkships, with the OSCE serving as a critical evaluation tool [5].

In Western countries, medical schools in Australia and Germany have implemented various OSCE formats, often in alignment with national standards [6,7]. In Asia, Korea adopted OSCEs in 1994 despite institutional challenges [8], while Taiwan experienced significant changes following national policy shifts [9]. However, the structure and implementation of OSCEs vary considerably among institutions, including the number and duration of stations, target clinical competencies, and examiner training systems. Educational frameworks, regulatory bodies, and available human and technical resources all influence these variations.

Although the traditional OSCE provides benefits such as standardization, objectivity, and reliability [10,11], it also poses operational challenges, including significant costs, staffing demands, and logistical complexities [6]. These practical limitations have led to critical discussions concerning the feasibility and scalability of OSCEs in real-world educational contexts [12-14]. In response to these challenges, virtual OSCEs have gained prominence, particularly during the COVID-19 pandemic, necessitating the evaluation of new competencies such as "webside manner" and remote communication [15].

In 2023, Japan implemented a national certification system mandating that OSCE examiners undergo formal training and accreditation, signaling a move toward standardizing the quality of clinical skills assessment [16]. This policy reflects Japan's effort to standardize OSCE assessment quality at the national level and illustrates the increasingly stringent requirements placed on OSCE examiners. However, the training and certification process, as required by the Common Achievement Tests Organization, is resource intensive and often impractical for faculty members who are already overburdened. Consequently, in October 2024, national guidelines were amended to allow OSCE sessions to be conducted with only 1 certified examiner per testing room, provided that the evaluation can subsequently be verified using video recordings [16].

Currently, most OSCE stations are monitored using fixed cameras. However, these systems frequently encounter limitations such as restricted angles and obstructed views, particularly when assessors or examinees block the field of view or when fine motor procedures are involved [17]. Such limitations impede accurate post hoc evaluations and may compromise the reliability of skill assessments.

Given the logistical challenges associated with securing certified examiners and the technical limitations of fixed-camera recordings, such as blind spots and obstructed views, there is an increasing need for alternative methods that can ensure reliable assessment with fewer personnel. To the best of our knowledge, no published study has evaluated the use of a neck-mounted wearable camera for OSCE assessment, and no prior research has specifically examined the THINKLET system. This study addresses this gap by investigating the feasibility of using an examiner-worn neck-mounted camera during OSCE assessments and its potential role as a supplementary tool for reviewing selected procedural tasks in settings with limited examiner availability.

Methods

Study Design

We undertook this pilot study to evaluate the feasibility and utility of a neck-mounted wearable camera (THINKLET, Fairy Devices Inc) for retrospective analysis in a simulated OSCE environment. Additionally, the study incorporated a questionnaire to assess the device's usability and acceptability. The OSCE scenario focused on the placement of 12-lead electrocardiogram (ECG) electrodes, a skill frequently evaluated in OSCEs. We selected the ECG electrode placement because it requires close observation of hand positioning and electrode placement relative to anatomical landmarks, which can be difficult to capture adequately using fixed-position cameras.

Participants

A fifth-year medical student volunteered to serve as the examinee. Nine faculty members, all with experience in clinical education, participated as evaluators. Demographic data were available for 8 evaluators; 1 evaluator did not complete the demographic questionnaire. All 9 evaluators took part in the initial live assessment; however, 1 of 9 (11%) evaluators did not participate in the subsequent video-based reassessment. We summarized continuous variables using means and SDs and presented categorical variables as frequencies and percentages.

Scenario and Checklist Development

The research team developed the OSCE scenario and checklist in accordance with the publicly accessible guideline, "Learning and Evaluation Items for Skills and Attitudes Required for Students Participating in Clinical Clerkships," published by the Common Achievement Tests Organization. The scenario encompassed a standard patient interaction and a procedural task related to the placement of ECG electrodes. The checklist, refined through expert consultation, comprised 26 binary items, with responses recorded as "1=performed" or "0=not performed."

Procedures

Evaluation 1 (Initial Assessment)

During the initial assessment phase, each evaluator wore a neck-mounted camera (THINKLET) while participating in the

XSL•FO
RenderX

simulated OSCE. Before each session, we conducted a brief test recording to confirm the camera angle and audio functionality. Concurrent recordings from the wearable camera and a stationary camera were obtained. Assessments were conducted on October 26, 2024 (6/9 evaluators, 67%) and December 16, 2024 (3/9 evaluators, 33%), with real-time scoring using a 26-item checklist. The complete checklist items are provided in Table S1 of Multimedia Appendix 1.

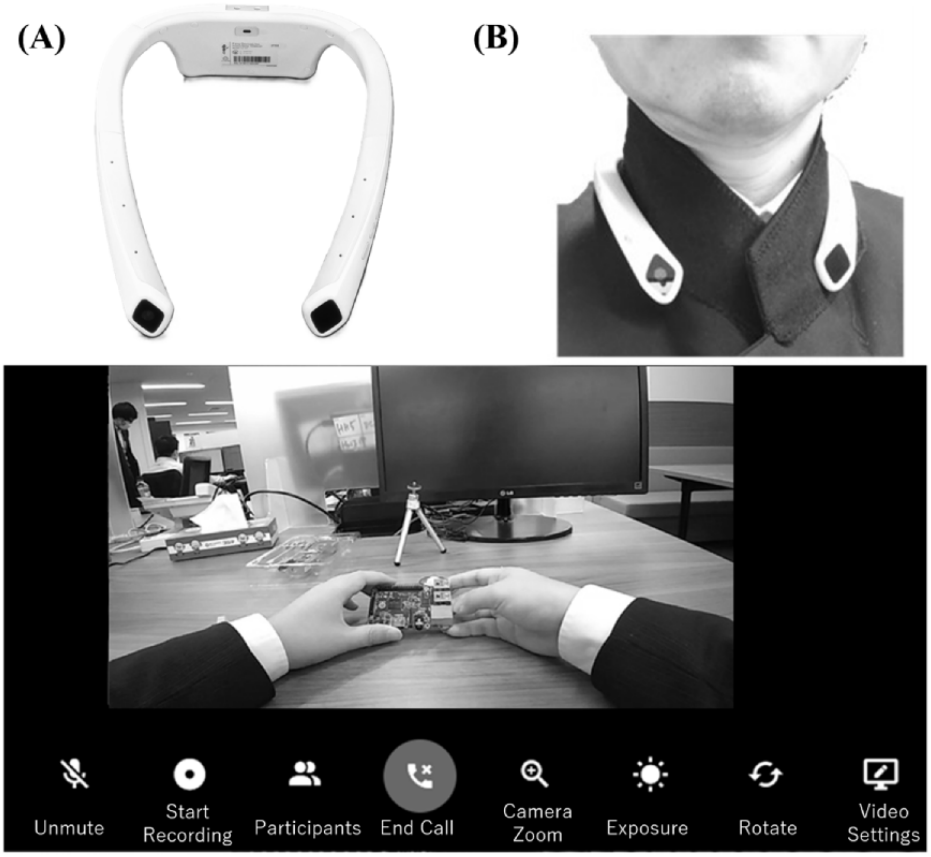
Evaluation 2 (Video-Based Reassessment)

Following a washout period of approximately 1 month, evaluators reassessed the same OSCE performance using neck-mounted camera recordings and the original 26-item checklist. We disseminated requests for reassessment on December 2, 2024, to 6 of 9 (67%) evaluators and on January 17, 2025, to 3 of 9 (33%) evaluators. Of the 9 evaluators, 8 (89%) completed the video-based reassessment; 1 (11%) evaluator did not respond and was consequently excluded from the agreement analysis. For the fixed-camera recordings, evaluators did not rescore the performance; rather, they judged whether each checklist item was evaluable or not evaluable based on the visibility of the relevant action in the footage.

Devices

The neck-mounted wearable camera used in this study was the THINKLET system, operated through the on-premises version of the Mirai Connect platform (Mirait One Systems Corporation). This device comprises an integrated camera and microphone, facilitating real-time communication via long term evolution or Wi-Fi. For this study, we used video and audio recordings. We recorded offline on a general-purpose PC and subsequently stored the recordings on a dedicated server. We captured audio in pulse code modulation 24-bit 48 kHz format and provided it to evaluators as complete video files. We obtained video recordings from both the fixed camera and the neck-mounted camera (refer to Multimedia Appendix 2 for an example of wearable camera footage). Figure 1 illustrates the configuration of the neck-mounted device. The device was rented from Mirait One Systems Corporation as a commercial service. The company provided technical support for device operation and troubleshooting but had no role in the study design, data collection, data analysis, interpretation of the results, or manuscript preparation.

Figure 1. Overview of the neck-mounted wearable camera and its setup in the Objective Structured Clinical Examination environment. (A) Front view of the neck-mounted wearable camera device, (B) example of the device worn around the neck, and (C) example of a live video feed captured from the neck-mounted camera during a manual task, as displayed in an online conferencing interface.



Questionnaire

Following the evaluations, each participant completed a questionnaire designed to assess 2 primary aspects: the comfort and usability of the wearable device during use, and the perceived ease and effectiveness of evaluating student performance using footage from wearable devices rather than fixed-camera footage. The questionnaire comprised 8 closed-ended items rated on a 5-point Likert scale, along with 2 open-ended questions. The questionnaire items on device usability and video-based evaluation are provided in Tables S2 and S3 of Multimedia Appendix 1.

XSL•FO
RenderX

Data Analysis

Descriptive Statistics

We undertook a descriptive analysis of the demographic data and the evaluators' backgrounds. We reported means and SDs for continuous variables and presented categorical variables as frequencies and percentages.

Agreement Between Live and Video-Based Scoring

We analyzed agreement between evaluation 1 and the video-based reassessment using percentage agreement and Cohen κ coefficient [18] and excluded items designated as "not evaluable" or missing data. We interpreted κ values in accordance with guidelines established by Landis and Koch [19].

Evaluability by Camera Type

For each checklist item, we descriptively compared evaluability ("evaluable" vs "not evaluable") between recordings from a fixed camera and those from a neck-mounted camera. Results are presented as counts and percentages of evaluable and nonevaluable observations for each camera type.

Questionnaire Analysis

We summarized responses to the Likert-scale items by calculating frequencies and percentages. We used thematic coding to analyze free-text responses, facilitating the identification of recurring content categories. Furthermore, we manually applied a basic sentiment classification, categorized as positive, negative, or neutral, to each response based on its overall tone. We selected this manual approach to ensure contextual accuracy and minimize the risk of misclassification due to nuanced expressions or cultural subtleties that automated sentiment analysis tools frequently overlook.

Software

We performed data analyses using Microsoft Excel and R (version 4.3.3; R Foundation for Statistical Computing).

Ethical Considerations

The Juntendo University Ethical Review Board (approval number E24-0093) approved this study. The study followed the Strengthening the Reporting of Observational Studies in Epidemiology reporting guidelines. All participants, including the student examinee and faculty evaluators, provided written informed consent before participation, including consent for video recording and publication of selected materials. The student examinee received compensation for participation, whereas the faculty evaluators did not receive financial compensation. Because the study involved video and audio recordings that could contain identifiable information, the data were stored on a secure institutional server with access restricted to the research team. Identifiable information, such as faces and name badges, was masked in publication materials when appropriate. A selected video excerpt was submitted for publication.

Results

Participant Characteristics

A total of 9 evaluators participated. Demographic data were available for 8 of 9 (89%) evaluators, as 1 of 9 (11%) evaluators did not complete the demographic questionnaire. Among these 8 evaluators, 7 (88%) were men and 1 (13%) was a woman, with a mean age of 42.1 (SD 8.9) years and a mean of 14.9 years (SD 9.1) since graduation. All respondents had graduated from a medical school. Overall, 3 (38%) evaluators had prior experience as officially certified OSCE examiners, having completed the formal national accreditation process, while the remaining 5 (63%) had no such certification. In total, 5 (56%) evaluators had never previously participated in an OSCE evaluation, whereas 3 (33%) had done so ≥5 times.

Agreement Between Live and Video-Based Scoring

Item-Based Agreement

Among the 26 checklist items, most demonstrated high concordance (75% to 100%) between the initial live assessments and video-based reassessments. Three items—removal of clothing (item 4), placement of chest lead C2 (item 12), and postprocedure tidying up (item 25)—achieved 100% agreement. In contrast, item 20, communication during the examination: instructing the patient to relax, had the lowest agreement at 50%. Table S4 in Multimedia Appendix 1 provides detailed item-level results.

Evaluator-Based Agreement

The level of agreement between each evaluator's initial and reassessment scores varied, with percentage agreement ranging from 70% to 96% and Cohen κ coefficient ranging from 0.258 to 0.913. The mean κ across all evaluators was 0.67 (SD 0.20). One evaluator showed only fair agreement (Cohen κ=0.26). Table S5 in Multimedia Appendix 1 provides detailed evaluator-specific results.

Evaluability by Camera Type

Figure 2 presents visual examples of the perspectives captured by the neck-mounted and fixed cameras. In the fixed-camera recordings, some tasks requiring fine motor skills, such as electrode placements at C5 and C6, were more frequently judged as not evaluable than in the neck-mounted camera recordings (Table 1). Across all checklist observations, 206 were judged evaluable in the neck-mounted camera recordings, compared with 185 in the fixed-camera recordings (Table 2). Several electrode placement tasks, including C5 and C6 placement, were more frequently judged evaluable in the neck-mounted camera recordings than in the fixed-camera recordings (Table 1). Multimedia Appendix 2 provides an example of this type of footage.

XSL•FO
RenderX

Figure 2. Comparison of camera perspectives during an Objective Structured Clinical Examination station on 12-lead electrocardiogram electrode placement. (A) View captured from the neck-mounted camera worn by the examiner, (B) view captured from a foot-side fixed camera, and (C) view captured from a head-side fixed camera.



XSL•FO
RenderX

Table 1. Comparison of evaluability: fixed vs neck-mounted camera.

Objective Structured Clinical Examination checklist item	Evaluation feasibility, n (%)	
	Fixed camera (n=8)	Neck-mounted camera (n=8)
Patient confirmation and preparation		
Hand hygiene: disinfect hands	8 (100)	8 (100)
Patient identification: confirm the patient's identity using full name and date of birth	8 (100)	8 (100)
Explanation of procedure: briefly explain the purpose and process of the ECG[a] to the patient	8 (100)	8 (100)
Clothing removal: ask the patient to remove upper-body garments, socks, and stockings	8 (100)	8 (100)
Privacy protection: provide appropriate coverage using a towel	8 (100)	8 (100)
Preparation of ECG equipment and environment		
Environmental check: ensure no other electrical devices are within 10 feet of the ECG machine	8 (100)	8 (100)
Device check: turn on the ECG device and confirm that it is functioning properly	8 (100)	8 (100)
Electrode placement		
Order of attachment: attach limb electrodes first, followed by precordial electrodes	7 (100)[b]	8 (100)
Limb lead placement: attach electrodes to the palmar side of the wrists and medial side of the ankles	7 (87.5)	8 (100)
Precordial lead preparation: palpate the sternal angle and identify the second intercostal space	7 (87.5)	8 (100)
Precordial lead placement: place C1 at the fourth intercostal space, right sternal border	7 (87.5)	8 (100)
Precordial lead placement: place C2 at the fourth intercostal space, left sternal border	7 (87.5)	8 (100)
Precordial lead placement: place C4 at the intersection of the fifth intercostal space and midclavicular line	6 (75)	8 (100)
Precordial lead placement: place C3 midway between C2 and C4	6 (75)	8 (100)
Precordial lead placement: place C5 at the same level as C4 on the left anterior axillary line	4 (50)	7 (87.5)
Precordial lead placement: place C6 at the same level as C4 on the left midaxillary line	4 (50)	7 (87.5)
Precordial lead placement: confirm the correct order from C1 to C6	6 (75)	8 (100)
Precordial lead placement: ensure that adjacent electrodes do not touch each other	7 (87.5)	8 (100)
Procedure and communication		
Communication during attachment: inform the patient before attaching the electrodes	8 (100)	8 (100)
Communication during measurement: encourage the patient to relax and remain still	8 (100)	8 (100)
Postprocedure handling		
Electrode removal: carefully remove electrodes to avoid skin irritation	7 (100)[b]	8 (100)
End-of-test communication: notify the patient that the measurement is complete before removing the electrodes	8 (100)	8 (100)
Electrode removal order: first remove the precordial electrodes, then the limb electrodes	8 (100)	8 (100)
Electrode removal order: remove the right leg (ground) electrode last	6 (75)	8 (100)
Cleanup: prepare electrodes and related items appropriately for next use	8 (100)	8 (100)
Closure		
Completion communication: inform the patient that the test is complete and instruct them to get dressed	8 (100)	8 (100)

[a]ECG: electrocardiogram.

[b]n=7 evaluators.

XSL•FO
RenderX

Table 2. Descriptive comparison of item evaluability between fixed camera and examiner-worn neck-mounted camera recordings[a].

Category	Fixed camera		Neck-mounted camera	
	Evaluable (n=185), n (%)	Not evaluable (n=21), n (%)	Evaluable (n=206), n (%)	Not evaluable (n=2), n (%)
Patient confirmation and preparation	40 (21.6)	0 (0)	40 (19.4)	0 (0)
Preparation of electrocardiogram equipment and environment	16 (8.6)	0 (0)	16 (7.8)	0 (0)
Electrode placement	68 (36.8)	19 (90.5)	86 (41.7)	2 (100)
Procedure and communication	16 (8.6)	0 (0)	16 (7.8)	0 (0)
Postprocedure handling	37 (20)	2 (9.5)	40 (19.4)	0 (0)
Closure	8 (4.3)	0 (0)	8 (3.9)	0 (0)

[a]The table presents a descriptive comparison of evaluability between the fixed camera and the neck-mounted camera, based on the number of evaluable and not evaluable observations in each category.

Questionnaire Results

In the context of video-based evaluation using wearable camera footage, 7 (88%) of 8 evaluators rewound and rewatched the videos during assessment. In total, 3 (38%) of 8 evaluators reported difficulty hearing audio, and 3 (38%) reported difficulty observing the desired area. Half (n=4, 50%) of the evaluators agreed, and 1 (13%) strongly agreed, that video quality was equivalent to that of on-site evaluation. Most (n=7, 88%) evaluators strongly agreed or agreed that video-based evaluation should be integrated into OSCE assessment. Tables S6 and S7 in Multimedia Appendix 1 provide detailed results, and Table 3 shows a summary of free-text feedback.

XSL•FO
RenderX

Table 3. Summary of free-text feedback.

Category and feedback	Coded comments, n	Representative comments
Usability and comfort		
Positive	11	• "I was initially worried the device might fall off, but I gradually got used to it." • "I didn't feel any burden from the weight, which was good." • "I was able to use it without any discomfort."
Negative	4	• "A bit hot." • "The back of my head felt warm." • "It might feel slightly hot in summer if worn for a long time." • "I felt heat around my neck."
Technical issues and attachment		
Positive	1	• "Honestly, I was surprised by the high quality of the wearable device."
Negative	13	• "It would be good if the position could be fixed. If it slips, I worry the camera will show what's below the body." • "I felt I needed to be aware that the camera's view might differ from what I was actually seeing." • "How to hold the evaluation sheet needs consideration." • "It was hard to decide where to put the binder." • "I was concerned the evaluation sheet might appear on screen." • "Though the device worked during the check, during the evaluation, it only showed my neck and glasses." • "Depending on how its worn, wearable devices may not capture low-positioned items." • "Familiarity may help, but placing the evaluation form was difficult." • "It was difficult to write on the form, which delayed evaluation." • "Audio cuts out." • "Audio issues." • "I was somewhat conscious about what the camera was capturing, but it didn't affect my evaluation." • "I was worried the evaluation notebook might be visible."
Benefits of video-based evaluation		
Positive	8	• "I could evaluate just like in a real-time situation." • "Using this device could reduce the burden on OSCE evaluators." • "Fixed cameras can't capture hands, but wearables can." • "Replaying improves evaluation accuracy."
Negative	0	• None

The evaluators' positive comments commonly noted good usability and lack of discomfort, while their negative comments frequently mentioned heat sensation, device slippage, and difficulty handling the evaluation sheet. Counts represent the number of coded comments derived from the free-text responses.

Discussion

Principal Findings

This study investigated the feasibility of using an examiner-worn neck-mounted camera for OSCE assessment. Agreement between live and video-based scoring varied across evaluators, with Cohen κ values ranging from 0.258 to 0.913. In the descriptive comparison of camera perspectives, more checklist items were judged evaluable in neck-mounted camera recordings than in fixed-camera recordings, particularly for selected electrode placement tasks requiring fine motor observation. Evaluator feedback indicated both perceived usefulness and practical limitations. Although many evaluators supported incorporating video-based assessment into OSCE practice, others reported technical challenges, including audio issues, visibility of the target area, heat sensation, device slippage, and difficulty handling the evaluation sheet (Table 3). Together, these findings suggest that an examiner-worn neck-mounted camera may be useful as a supplementary tool for review and verification in selected OSCE tasks, rather than as a replacement for live assessment. Further refinement of device setup and recording conditions is needed before broader implementation.

This study directly compared live and video-based OSCE assessments using neck-mounted and fixed-camera recordings in a simulated educational scenario. Our methodology provided descriptive information regarding the evaluability of individual checklist items. We also gathered and reported evaluator background information, thereby enhancing the transparency of the study context. Furthermore, we incorporated both quantitative and qualitative feedback on device usability to support the interpretation of the feasibility findings.

The COVID-19 pandemic, which began in 2019, prompted a significant shift toward remote and online methods across

XSL•FO
RenderX

various sectors, including medical education [20]. As telemedicine, online learning, and virtual assessments proliferated [21,22], the adoption of wearable technologies also gained traction [22].

Wearable devices have garnered increasing attention as instruments for clinical education and performance evaluation [23]. The expansion of remote and hybrid teaching during the COVID-19 pandemic further accelerated interest in their application for clinical skills training. These devices provide a first-person perspective, offering wearable camera recordings that facilitate detailed observation of fine motor skills and examinee behaviors [24,25]. In addition to passive recording, certain devices also support real-time, bidirectional communication, enabling remote instructors or supervisors to deliver immediate feedback during clinical tasks. This interactive capability further enhances their utility in remote education and simulation-based training environments. In online education, these recordings have been associated with increased student satisfaction and perceived educational effectiveness [23,26]. These characteristics may make wearable cameras a potentially relevant option for examiner-perspective recording in OSCE assessment.

Although wearable devices offer distinct advantages, they may not be suitable for all assessment domains; for example, fixed third-person cameras may be more appropriate for evaluating nonverbal communication [25]. Wearable and related observational technologies have been explored in health research, OSCE-related contexts, and other applied settings [27-31]. In this study, we focused specifically on the potential role of an examiner-worn neck-mounted camera in OSCE assessment.

Although various wearable devices, including smart glasses and action cameras, have been used globally to evaluate communication with standardized patients and develop instructional content [32,33], they are not without limitations. Reported challenges encompass device weight, image instability, limited field of view, and privacy concerns [34,35]. Smart glasses have been associated with visual distraction and diminished attentional focus because of their design, including a limited field of view and camera instability. These limitations may impede clear observation of procedural tasks and compromise the reliability of recorded footage, particularly in high-stakes assessment settings such as OSCEs [36-38]. These challenges indicate the importance of ergonomic and design considerations in wearable camera systems. These considerations are particularly relevant to the use of examiner-worn cameras in OSCE assessment, where stable first-person recording of procedural tasks is required. Research has identified smart glasses as less user-friendly regarding operability and comfort [36], which influenced our decision to implement a neck-mounted camera design. In this pilot study, evaluator feedback suggested that the neck-mounted camera was generally acceptable to some users and may offer practical advantages for reviewing selected procedural tasks.

Nonetheless, we observed some technical limitations. For example, visibility was occasionally obstructed when evaluators held the checklist close at hand, which sometimes blocked the camera's view. In addition, there were challenges related to audio capture because the recording software did not allow remote control of the microphone's on-off function due to privacy regulations. Consequently, evaluators could not easily confirm whether audio was being recorded, which occasionally made it difficult to ensure adequate audio quality for precise assessment.

Furthermore, we noted variability in interrater agreement (Cohen κ=0.258-0.913), potentially influenced by inconsistencies in video framing. In certain instances, key procedural steps were not clearly visible, and evaluators reported misaligned camera angles. These issues likely resulted from variations in how the device was worn, as factors such as shirt collars, ties, or fabric thickness affected the camera's orientation and stability.

To address these challenges, future efforts should focus on standardizing attachment protocols, introducing supplementary supports to minimize clothing-related interference, enhancing microphone sensitivity, and providing user interfaces that indicate audio recording status. Conducting test recordings before actual assessments may also help ensure optimal video framing and sound quality. Moreover, broader implementation will require the establishment of standardized OSCE video-recording protocols, structured evaluator training programs, and secure data governance frameworks that comply with relevant ethical and legal standards.

Limitations

This pilot study had some limitations. First, the small sample size and limited number of evaluators constrained the generalizability of the findings. Additionally, the examinee was a fifth-year medical student participating in a mock OSCE, which may not fully replicate actual examination conditions. Second, the evaluators were predominantly faculty members from a single institution, with only 1 evaluator from another university, thereby limiting the diversity of perspectives. Third, the study focused exclusively on a single OSCE scenario, ECG electrode placement, rendering it challenging to extrapolate the findings to other clinical skills or specialties. Additionally, this study did not evaluate scoring accuracy against an external reference standard; therefore, the findings should not be interpreted as evidence that neck-mounted recordings improve assessment accuracy. Furthermore, agreement analyses were based on reassessment using the neck-mounted recordings, whereas fixed-camera recordings were used only to judge whether checklist items were evaluable. Technical issues, including audio quality, visibility, and device positioning, also limited the consistency of video-based reassessment. Future research involving larger sample sizes, multi-institutional collaboration, and multiple OSCE stations is necessary to validate the broader applicability of neck-mounted cameras.

Conclusions

This pilot study investigated the feasibility of using an examiner-worn neck-mounted camera in preclinical OSCE assessments. A greater number of checklist items were judged evaluable in neck-mounted camera recordings than in fixed-camera recordings, particularly for selected tasks requiring fine motor observation. These findings suggest that an

XSL•FO
RenderX

examiner-worn neck-mounted camera may be useful as a supplementary tool for review and verification in selected OSCE tasks rather than as a replacement for live assessment. However, technical and operational challenges remain, indicating the need for further refinement of device design and evaluation protocols in medical education before broader implementation.

Acknowledgments

The authors thank Mr Yoshiki Tsumura for volunteering as the simulated Objective Structured Clinical Examination examinee, and Dr Naoto Nishizaki, Dr Daishi Hirano, and Mirait One Systems Corporation for arranging the rental of the wearable camera device and providing technical support. Additionally, the authors thank all participants for their cooperation in this study and Editage for English-language editing.

Data Availability

The datasets generated and analyzed during this study, including video recordings featuring participants, are not publicly available due to ethical restrictions and participant confidentiality. However, they are available from the corresponding author upon reasonable request and with approval from the institutional ethics committee.

Funding

The Japan Medical Education Foundation Grant-in-Aid for Research (2024) funded this study.

Authors' Contributions

All listed authors met the 4 criteria for authorship as defined by the International Committee of Medical Journal Editors and required by *Academic Medicine*: (1) substantial contributions to the conception and design of the study, acquisition of data, or analysis and interpretation of data; (2) drafting the manuscript or revising it critically for important intellectual content; (3) final approval of the version to be published; and (4) agreement to be accountable for all aspects of the work.
Task-specific contributions are summarized as follows:
Conceptualization: MS, YN, CM, MO, CO, YK, YT
Data curation: MS, YN
Formal analysis: MS
Funding acquisition: YN
Investigation: AE, TO, YT, MN, HW, CM, MO, CO
Methodology: YN, YK
Project administration: YN
Software: MS
Supervision: YN
Validation: MS, AE, TO, YT, MN, HW
Visualization: MS, YN
Writing—original draft: MS, YN, YK, YT
Writing—review and editing: YN, AE, TO, YT, MN, HW, CM, MO, CO, YK, YT
All authors reviewed and approved the final version of the manuscript and agree to be accountable for all aspects of the work.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Detailed Objective Structured Clinical Examination checklist, interrater agreement, and questionnaire items or responses used in the study.
[DOCX File , 31 KB - mededu_v12i1e87483_app1.docx]

Multimedia Appendix 2
A 60-second excerpt from a neck-mounted camera (THINKLET, Fairy Devices Inc) worn by the examiner during a simulated Objective Structured Clinical Examination station on 12-lead electrocardiogram electrode placement.
[MP4 File (MP4 Video), 49707 KB - mededu_v12i1e87483_app2.mp4]

References

1. Harden RM, Stevenson M, Downie WW, Wilson GM. Assessment of clinical competence using objective structured examination. Br Med J 1975 Feb 22;1(5955):447-451 [FREE Full text] [doi: 10.1136/bmj.1.5955.447] [Medline: 1115966]

XSL•FO
RenderX

2. Miller GE. The assessment of clinical skills/competence/performance. Acad Med 1990 Sep;65(9 Suppl):S63-S67. [doi: 10.1097/00001888-199009000-00045] [Medline: 2400509]
3. Wass V, Van der Vleuten C, Shatzer J, Jones R. Assessment of clinical competence. Lancet 2001 Mar 24;357(9260):945-949. [doi: 10.1016/s0140-6736(00)04221-5]
4. Common Achievement Tests Organization. URL: https://www.mhlw.go.jp/content/10803000/000932162.pdf [accessed 2025-04-09]
5. Zayyan M. Objective structured clinical examination: the assessment of choice. Oman Med J 2011 Jul;26(4):219-222 [FREE Full text] [doi: 10.5001/omj.2011.55] [Medline: 22043423]
6. Heal C, D'Souza K, Banks J, Malau-Aduli BS, Turner R, Smith J, et al. A snapshot of current Objective Structured Clinical Examination (OSCE) practice at Australian medical schools. Med Teach 2019 Apr;41(4):441-447. [doi: 10.1080/0142159X.2018.1487547] [Medline: 30261798]
7. Müller S, Dahmen U, Settmacher U. Application of the Objective Structured Clinical Examination (OSCE) in German medical schools: an inventory [Article in German]. Gesundheitswesen 2018 Dec;80(12):1099-1103. [doi: 10.1055/s-0042-116435] [Medline: 27813044]
8. Lee YM, Ahn DS. The OSCE: a new challenge to the evaluation system in Korea. Med Teach 2006 Jun;28(4):377-379. [doi: 10.1080/01421590600627086] [Medline: 16807181]
9. Lin CW, Tsai TC, Sun CK, Chen DF, Liu KM. Power of the policy: how the announcement of high-stakes clinical examination altered OSCE implementation at institutional level. BMC Med Educ 2013 Jan 24;13:8 [FREE Full text] [doi: 10.1186/1472-6920-13-8] [Medline: 23347392]
10. Witheridge A, Ferns G, Scott-Smith W. Revisiting Miller's pyramid in medical education: the gap between traditional assessment and diagnostic reasoning. Int J Med Educ 2019 Oct 25;10:191-192 [FREE Full text] [doi: 10.5116/ijme.5d9b.0c37] [Medline: 31655795]
11. Moin M. Objective Structured Clinical Examination (OSCE); strengths and hurdles in implementation. Pak J Ophthalmol 2020 Sep 14;36(4):312-314. [doi: 10.36351/pjo.v36i4.1129]
12. Elshama SS. How to design and apply an Objective Structured Clinical Examination (OSCE) in medical education? Iberoam J Med 2020 Nov 06;3(1):51-55. [doi: 10.53986/ibjm.2021.0009]
13. Gormley G. Summative OSCEs in undergraduate medical education. Ulster Med J 2011 Sep;80(3):127-132 [FREE Full text] [Medline: 23526843]
14. Patrício MF, Julião M, Fareleira F, Carneiro AV. Is the OSCE a feasible tool to assess competencies in undergraduate medical education? Med Teach 2013 Jun;35(6):503-514. [doi: 10.3109/0142159X.2013.774330] [Medline: 23521582]
15. Blamoun J, Hakemi A, Armstead T. A guide for medical students and residents preparing for formative, summative, and virtual Objective Structured Clinical Examination (OSCE): twenty tips and pointers. Adv Med Educ Pract 2021 Aug 29;12:973-978. [doi: 10.2147/AMEP.S326488]
16. Opinion on the Common Achievement Test from 2020 onwards. Ministry of Health, Labour and Welfare. 2024. URL: https://www.mhlw.go.jp/content/001319749.pdf [accessed 2025-04-05]
17. Wilby KJ, Diab M. Key challenges for implementing a Canadian-based Objective Structured Clinical Examination (OSCE) in a Middle Eastern context. Can Med Educ J 2016 Dec 05;7(2):e4-e9. [doi: 10.36834/cmej.36720]
18. Cohen J. A coefficient of agreement for nominal scales. Educ Psychol Meas 1960 Apr;20(1):37-46. [doi: 10.1177/001316446002000104]
19. Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics 1977 Mar;33(1):159-174. [doi: 10.2307/2529310]
20. Goudarzi H, Onozawa M, Takahashi M. Impact of the Covid-19 pandemic and ensuing online teaching on pre-clinical medical education. BMC Med Educ 2024 Jan 17;24(1):66 [FREE Full text] [doi: 10.1186/s12909-023-04967-x] [Medline: 38233801]
21. Wosik J, Fudim M, Cameron B, Gellad ZF, Cho A, Phinney D, et al. Telehealth transformation: COVID-19 and the rise of virtual care. J Am Med Inform Assoc 2020 Jun 01;27(6):957-962 [FREE Full text] [doi: 10.1093/jamia/ocaa067] [Medline: 32311034]
22. Rose S. Medical student education in the time of COVID-19. JAMA 2020 Jun 02;323(21):2131-2132. [doi: 10.1001/jama.2020.5227] [Medline: 32232420]
23. Kwon OY. Online clinical skills education using a wearable action camera for medical students. J Exerc Rehabil 2022 Dec 27;18(6):356-360 [FREE Full text] [doi: 10.12965/jer.2244460.230] [Medline: 36684536]
24. Devine T, Gormley C, Doyle P. Lights, camera, action: using wearable camera and interactive video technologies for the teaching and assessment of lab experiments. Int J Innov Sci Math Educ 2015 Jan;23(2):22-33 [FREE Full text]
25. Zahl DA, Schrader SM, Edwards PC. Student perspectives on using egocentric video recorded by smart glasses to assess communicative and clinical skills with standardised patients. Eur J Dent Educ 2018 May;22(2):73-79 [FREE Full text] [doi: 10.1111/eje.12217] [Medline: 27380732]
26. Sheng A, Chubbs-Payne A, Horner E, Issa M, Leung FH, Malek M. Using POV wearable technology as a tool in virtual teaching sessions to supplement clinical skills training: a medical student perspective. Univ Ottawa J Med 2024 Jul 19;14(1):69-78. [doi: 10.18192/uojm.v14i1.6483]

XSL•FO
RenderX

27. Doherty AR, Hodges SE, King AC, Smeaton AF, Berry E, Moulin CJ, et al. Wearable cameras in health: the state of the art and future possibilities. Am J Prev Med 2013 Mar;44(3):320-323. [doi: 10.1016/j.amepre.2012.11.008] [Medline: 23415132]
28. Grima-Murcia MD, Sanchez-Ferrer F, Ramos-Rincón JM, Fernández E. Use of eye-tracking technology by medical students taking the objective structured clinical examination: descriptive study. J Med Internet Res 2020 Aug 21;22(8):e17719 [FREE Full text] [doi: 10.2196/17719] [Medline: 32821060]
29. Mardonova M, Choi Y. Review of wearable device technology and its applications to the mining industry. Energies 2018 Mar 04;11(3):547. [doi: 10.3390/en11030547]
30. Sandy M, Goldburt T, Carapezza E. Mobile, portable lightweight wireless video recording solutions for homeland security, defense,law enforcement applications. Proc SPIE 2015:94560A [FREE Full text] [doi: 10.1117/12.2193820]
31. Nahavandi D, Alizadehsani R, Khosravi A, Acharya UR. Application of artificial intelligence in wearable devices: opportunities and challenges. Comput Methods Programs Biomed 2022 Jan;213:106541. [doi: 10.1016/j.cmpb.2021.106541] [Medline: 34837860]
32. Fukui S, Shimizu T, Nishizaki Y, Shikino K, Yamamoto Y, Kobayashi H, et al. The utility of wearable cameras in developing examination questions and answers on physical examinations: preliminary study. JMIR Med Educ 2024 Jul 19;10:e53193 [FREE Full text] [doi: 10.2196/53193] [Medline: 39037348]
33. Tully J, Dameff C, Kaib S, Moffitt M. Recording medical students' encounters with standardized patients using Google Glass: providing end-of-life clinical education. Acad Med 2015 Mar;90(3):314-316. [doi: 10.1097/ACM.0000000000000620] [Medline: 25551855]
34. McDevitt S, Hernandez H, Hicks J, Lowell R, Bentahaikt H, Burch R, et al. Wearables for biomechanical performance optimization and risk assessment in industrial and sports applications. Bioengineering (Basel) 2022 Jan 13;9(1):33 [FREE Full text] [doi: 10.3390/bioengineering9010033] [Medline: 35049742]
35. Zarepour E, Hosseini M, Kanhere SS, Sowmya A, Rabiee HR. Applications and challenges of wearable visual lifeloggers. Computer 2017 Mar;50(3):60-69. [doi: 10.1109/mc.2017.82]
36. Kim SK, Lee Y, Yoon H, Choi J. Adaptation of extended reality smart glasses for core nursing skill training among undergraduate nursing students: usability and feasibility study. J Med Internet Res 2021 Mar 02;23(3):e24313 [FREE Full text] [doi: 10.2196/24313] [Medline: 33650975]
37. Lee Y, Kim SK, Yoon H, Choi J, Kim H, Go Y. Integration of extended reality and a high-fidelity simulator in team-based simulations for emergency scenarios. Electronics 2021 Sep 06;10(17):2170. [doi: 10.3390/electronics10172170]
38. Muroi K, Kyogoku S, Sakano Y, Sakamoto H, Nakazeko K, Koyama K, et al. An analysis of the effectiveness of reflective learning through watching videos recorded with smart glasses-with multiple views (student, patient, and overall) in radiography education. PLoS One 2024 Jan 2;19(1):e0296417 [FREE Full text] [doi: 10.1371/journal.pone.0296417] [Medline: 38165849]

Abbreviations

ECG: electrocardiogram
OSCE: Objective Structured Clinical Examination

Edited by J Gentges; submitted 10.Nov.2025; peer-reviewed by J Kelly, D Hein; comments to author 30.Jan.2026; accepted 13.Apr.2026; published 27.May.2026.

Please cite as:
Sekine M, Nishizaki Y, Endo A, Ogawa T, Takeshita Y, Nasu M, Wada H, Miyamoto C, Oguro M, Otake C, Konishi Y, Tomiki Y
Exploring the Feasibility of an Examiner-Worn Neck-Mounted Camera for Objective Structured Clinical Examination Assessment: Pilot Feasibility Study
JMIR Med Educ 2026;12:e87483
URL: https://mededu.jmir.org/2026/1/e87483
doi:10.2196/87483
PMID:

©Miwa Sekine, Yuji Nishizaki, Amane Endo, Takasuke Ogawa, Yoshihide Takeshita, Motomi Nasu, Hiroo Wada, Chizuko Miyamoto, Michiko Oguro, Chie Otake, Yasuhiko Konishi, Yuichi Tomiki. Originally published in JMIR Medical Education (https://mededu.jmir.org), 27.May.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Education, is properly cited. The complete bibliographic

XSL•FO
RenderX

information, a link to the original publication on https://mededu.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Learning Analytics of a National Entrustable Professional Activities Platform: Cross-Sectional Study of System-Level Constraints on Advanced Entrustment in Competency-Based Medical Education

Jeng-Wen Chen[1,2,3*]; Rui-Bin Yu[4,5*]; Kai-Nan Lin[1,2*]; Yukiko Ono[3,6]; Xuan-Bing Yu[3]; Wei-Chung Hsu[7,8,9]; Pa-Chun Wang[10,11,12]; Li-Jen Liao[13]; Chun-Hou Liao[11,14]; Mingchih Chen[6,15]

[1]Department of Otolaryngology–Head and Neck Surgery, Cardinal Tien Hospital and College of Medicine, Fu Jen Catholic University, New Taipei City, Taiwan

[2]Department of Otolaryngology–Head and Neck Surgery, National Taiwan University Hospital, Taipei, Taiwan

[3]Department of Hospital Management, Graduate Institute of Business Administration, Fu Jen Catholic University, New Taipei City, Taiwan

[4]Department of Otolaryngology–Head and Neck Surgery, Cardinal Tien Hospital and Fu Jen Catholic University, New Taipei City, Taiwan

[5]Institute of Educational Administration and Evaluation, University of Taipei, Taipei, Taiwan

[6]Data Science Center, College of Medicine, Fu Jen Catholic University, New Taipei City, Taiwan

[7]Department of Otolaryngology, Head and Neck Surgery, National Taiwan University Hospital and Children's Hospital, Taipei, Taiwan

[8]Sleep Center, National Taiwan University Hospital, Taipei, Taiwan

[9]Department of Otolaryngology, College of Medicine, National Taiwan University, Taipei, Taiwan

[10]Department of Otolaryngology, Cathay General Hospital, Taipei, Taiwan

[11]School of Medicine, Fu-Jen Catholic University, New Taipei City, Taiwan

[12]Department of Medical Research, China Medical University Hospital, China Medical University, Taichung, Taiwan

[13]Far Eastern Memorial Hospital, New Taipei City, Taiwan

[14]Division of Urology, Department of Surgery, Cardinal Tien Hospital, New Taipei City, Taiwan

[15]Artificial Intelligence Development Center, Fu Jen Catholic University, New Taipei City, Taiwan

[*]these authors contributed equally

Corresponding Author:
Jeng-Wen Chen
Department of Otolaryngology–Head and Neck Surgery
Cardinal Tien Hospital and College of Medicine
Fu Jen Catholic University
362 ZhongZheng Rd.
Xindian Dist.
New Taipei City, 23148
Taiwan
Phone: 886 2 22193391 ext 67451
Email: 086365@mail.fju.edu.tw

Abstract

Background: Competency-based medical education (CBME) relies on entrustable professional activities (EPAs) and Clinical Competency Committee (CCC) deliberation to support defensible decisions about trainee progression. As digital assessment platforms increasingly aggregate workplace-based assessment data across training programs, large-scale learning analytics can provide new insights into how entrustment decisions are generated and interpreted within CBME systems. However, little is known about how national assessment infrastructures influence patterns of entrustment attainment.

Objective: This study examined national CCC summative entrustment decisions to identify system-level factors associated with attainment of expected supervision levels across residency training.

Methods: We conducted a cross-sectional analysis of nationwide CCC entrustment data derived from the Emyway digital assessment platform used by all accredited otolaryngology–head and neck surgery residency programs in Taiwan. The dataset included 3504 summative entrustment decisions across 12 EPAs for 292 residents. Observed supervision levels were compared

XSL•FO
RenderX

with prespecified targets for each training year. Logistic regression models were used to identify factors associated with attainment of expected supervision levels, including training stage, EPA sequencing, CCC review cadence, and program characteristics.

Results: Median supervision levels increased across training years but did not consistently reach expected targets in later stages. Expected entrustment targets were met in 2558 of 3504 decisions (73%). Underattainment was most pronounced in senior training years, particularly R4 and R5. In multivariable analyses, delayed EPA sequencing (adjusted odds ratio 0.75, 95% CI 0.63-0.91) and lower CCC review cadence (adjusted odds ratio 0.41, 95% CI 0.28-0.61) were independently associated with lower attainment. Program-level variability in entrustment outcomes was also observed.

Conclusions: Nationwide learning analytics of CCC entrustment decisions revealed a reproducible late-stage gap between expected and observed autonomy in CBME training. This pattern was associated with features of assessment system design that may be modifiable rather than trainee characteristics alone. National benchmarking of digital assessment data can therefore provide actionable insights to optimize EPA sequencing, CCC governance, and evidence generation within CBME systems.

(JMIR Med Educ 2026;12:e95066) doi:10.2196/95066

KEYWORDS

competency-based medical education; learning analytics; entrustable professional activities; clinical competency committees; workplace-based assessment; digital assessment systems

Introduction

Competency-based medical education (CBME) aims to ensure that progression through training reflects demonstrated clinical competence rather than time alone [1-5]. Entrustable professional activities (EPAs) [6,7] operationalize this goal by linking clinical work to supervision levels that reflect increasing autonomy [8-10]. Within programmatic assessment systems, Clinical Competency Committees (CCCs) synthesize multiple workplace-based assessments (WBAs) [11,12] to determine whether trainees can be entrusted with progressively independent clinical responsibilities [10,13-16].

As CBME implementation expands, digital assessment platforms have become central infrastructures for collecting and aggregating assessment data across training programs [5]. These platforms enable large-scale analysis of assessment patterns, creating opportunities to apply learning analytics to understand how entrustment decisions are generated within complex educational systems [17]. While prior research has examined individual EPA assessments and local CCC processes [18,19], fewer studies have analyzed national entrustment data to identify structural factors that influence progression across training programs.

Understanding these system-level dynamics is particularly important for high-stakes entrustment decisions [20]. If patterns of entrustment attainment systematically deviate from expected trajectories, this may reflect constraints [21] in the design or implementation of assessment systems rather than limitations in trainee development alone [22,23]. National digital assessment platforms, therefore, provide a unique opportunity to examine whether features of assessment architecture—such as EPA sequencing, committee review cadence, and program context—shape observed entrustment outcomes [16,24,25].

Recent scholarship has highlighted the potential of learning analytics to strengthen CCC processes and increase the value of data-informed decision-making in CBME [16,26]. Yet much of the existing literature remains focused on single institutions, local dashboards, or implementation experiences rather than national-scale analyses of summative entrustment patterns across programs. As a result, relatively little is known about how digital assessment infrastructures can be used to detect system-level bottlenecks in entrustment progression or to identify modifiable features of assessment design associated with under-attainment of expected supervision targets.

National EPA infrastructures offer a particularly useful setting for this type of analysis because they combine standardized frameworks with cross-program data collection [27]. In Taiwan, the Taiwan Society of Otorhinolaryngology-Head and Neck Surgery (ORL–HNS) has implemented a national EPA-based digital assessment system through the Emyway platform, which has accumulated more than 35,000 formative EPA records since 2022 [21,28,29]. Building on this infrastructure, a nationwide pilot in 2025 required all accredited ORL–HNS residency programs to upload CCC summative entrustment decisions for each resident and each EPA through the same electronic platform. This created a national dataset spanning programs, residents, EPAs, and training stages, providing an opportunity to examine summative entrustment decisions through a learning analytics lens [16,30].

Using nationwide CCC summative data derived from a digital EPA assessment platform, this study examined patterns of entrustment attainment across residency training. We sought to identify system-level factors associated with meeting expected supervision targets and to explore how national learning analytics can inform the design and governance of CBME assessment systems.

Methods

Study Design and Data Source

We conducted a national cross-sectional analysis of summative entrustment decisions recorded through a shared digital assessment platform (Emyway) supporting CBME implementation across all accredited ORL–HNS residency programs in Taiwan [28]. The platform is a cloud-based educational information system that records WBAs and CCC summative entrustment decisions across institutions. The CCC summative entrustment decisions are entered into Emyway by program administrators after CCC meetings, using standardized

XSL•FO
RenderX

EPA definitions and supervision-level criteria. The platform stores these decisions in a centralized database, enabling aggregation across programs for national benchmarking and research.

In 2025, the Taiwan Society of ORL–HNS implemented a nationwide pilot requiring all programs to submit CCC summative entrustment-supervision levels for each resident and each EPA through the platform. Reporting occurred twice during the year (January and July). During each cycle, programs submitted one CCC summative entrustment level per resident-EPA combination, reflecting committee-level judgments synthesized from longitudinal WBA data. The resulting dataset comprised residents from 35 accredited training programs nationwide and reflects formal summative entrustment decisions used to determine trainee progression within the national EPA framework.

Digital Assessment Platform

Emyway supports EPA-based programmatic assessment by enabling faculty to submit WBA records, aggregating performance data across EPAs, and providing dashboards for CCC review [28,29]. Summative entrustment decisions were recorded within the Emyway platform through structured CCC processes involving multiple faculty reviewers, with standardized EPA definitions and supervision level criteria (Level 1: observation only; Level 2: direct supervision; Level 3: indirect supervision; Level 4: independent practice; Level 5: supervise others), supporting the consistency and reliability of the recorded data. Following CCC meetings, program administrators enter final summative entrustment decisions into the platform, where the data are stored in a centralized database that supports national benchmarking and research.

National EPA Framework

The national ORL–HNS EPA framework comprises 12 specialty-specific EPAs covering the major domains of clinical practice (Multimedia Appendix 1). Each EPA is evaluated using a 5-level entrustment-supervision scale adapted from established CBME frameworks. The framework defines expected supervision levels for each stage of training (R1-R5) (Table 1). The expected entrustment targets were defined a priori based on the national ORL–HNS EPA framework developed by the Taiwan Society of Otolaryngology–Head and Neck Surgery through expert consensus during implementation. These targets represent the intended progression of supervision levels across training years and are used operationally for programmatic assessment and benchmarking, rather than being derived from empirical validation. Summative entrustment decisions represent CCC judgments synthesized from formative EPA assessments collected within the platform.

Table 1. Expected entrustment-supervision levels used for summative assessment according to resident seniority.

EPA[a] number	EPA title	Summative entrustment-supervision levels for respective resident seniority				
		R1	R2	R3	R4	R5
EPA01	(Airway) Assessing and managing patients with airway presentations	2	3	3	4	5
EPA02	(Foreign Body) Assessing and managing patients with suspicious foreign body presentations	2	3	3	4	5
EPA03	(Bleeding) Assessing and managing patients with upper aerodigestive tract bleeding presentations	2	3	3	4	5
EPA04	(Vertigo) Assessing and managing patients with vertigo	2	3	3	4	5
EPA05	(Infection) Assessing and managing patients with head and neck infections	2	3	3	4	5
EPA06[b]	(Head and Neck) Assessing and managing patients with head and neck masses (including oral cavity tumors)	2	2	3	4	5
EPA07	(Ear) Assessing and managing patients with ear and hearing diseases	2	3	3	4	5
EPA08	(Sinonasal) Assessing and managing patients with sinonasal diseases	2	3	4	4	5
EPA09	(Larynx) Assessing and managing patients with laryngopharyngeal diseases (voice/speech/language/dysphagia)	2	3	3	4	5
EPA10[b]	(Sleep Disordered Breathing) Assessing and managing patients with sleep-disordered breathing	2	2	3	4	5
EPA11[b]	(Plasty) Assessing and managing patients with facial plastic and reconstructive surgery	2	2	3	3	4
EPA12	(Oral Presentation) Case presentation	2	3	4	4	5

[a]EPA: entrustable professional activity.

[b]Delayed EPAs: EPA06, EPA10, and EPA11.

Participants

The study population included all residents enrolled in accredited ORL–HNS residency programs in Taiwan during 2025; no additional exclusion criteria were applied. Residency training spans 5 specialty years (R1-R5), corresponding to postgraduate years PGY3-PGY7 following completion of a 2-year postgraduate general training program. Faculty

XSL•FO
RenderX

participants were attending physicians who contributed workplace-based EPA assessments that informed CCC deliberations.

Variables

CCC ratings were recorded using a five-level entrustment-supervision scale [10,31,32]. The primary outcome was dichotomized as meeting versus not meeting the expected entrustment-supervision level because the national CBME framework uses prespecified stage-based targets to monitor learner progression and determine whether residents have achieved the expected level of entrustment for each EPA (Table 1).

Independent variables represented features of the assessment system and program context. EPA sequencing was defined based on the typical order of clinical exposure within the national ORL–HNS curriculum. EPAs related to common outpatient care and emergency presentations are introduced earlier, whereas those involving subspecialty procedures, complex consultations, or advanced decision-making occur later as experience accumulates. Accordingly, early versus delayed EPAs reflect curricular design and expected learning progression rather than assessment logistics. Delayed EPAs were defined as EPAs that are typically expected to be achieved at later stages of training due to their higher complexity, requirement for advanced clinical judgment, and dependence on accumulated experience. In this study, EPA6, EPA10, and EPA11 were categorized as delayed EPAs based on expert consensus within the national specialty framework and prior curriculum design. CCC review cadence was classified as either 2 reviews per year or one review per year and was operationalized using platform-recorded submission frequency in Emyway during 2025. Programs were categorized based on whether they submitted CCC summative entrustment decisions in both scheduled cycles (January and July) or in only one cycle; this variable was derived from platform records rather than self-report. Resident-level covariates included training year (R1-R5) and sex. Program-level variables included hospital level (medical center vs nonmedical center), geographic region, and faculty size (number of attending physicians involved in resident training and assessment). Variables were selected a priori based on prior literature and theoretical considerations related to programmatic assessment systems. For factor analyses, data from the July 2025 reporting cycle were used to ensure contemporaneous assessment across EPAs.

Statistical Analysis

Descriptive statistics were used to summarize resident, program, and assessment characteristics. Differences between assessments that met or did not meet expected entrustment targets were examined using chi-square tests. We considered potential sources of bias, including measurement variability in CCC entrustment decisions and residual confounding by program and training level; selection bias was minimized through nationwide inclusion, and remaining biases were addressed through model adjustment and sensitivity analyses. The unit of analysis was the CCC summative entrustment decision for each resident-EPA combination. Although some residents contributed decisions from more than one review cycle, each decision was made through a formal CCC process involving different faculty members and EPA-specific evidence inputs; therefore, these decisions were analyzed as distinct observations. Associations between explanatory variables and attainment of expected entrustment levels were evaluated using logistic regression models. Univariable analyses were first performed, followed by multivariable logistic regression including hospital level, faculty size, geographic region, CCC review cadence, EPA sequencing, resident training year, and sex. Results are reported as odds ratios with 95% CIs. Stratified analyses were conducted by resident training year and EPA sequencing to assess the consistency of associations across training stages. Developmental patterns in entrustment progression were examined by comparing observed CCC supervision levels with predefined training targets, summarized using medians with IQRs. All tests were 2-sided with statistical significance set at $P<.05$. Analyses were performed using SAS (version 9.4; SAS Institute, Cary, NC, USA).

Ethical Considerations

The study was approved by the Institutional Review Board of Cardinal Tien Hospital (CTH-112-2-1-002). Participants provided electronic informed consent prior to accessing the Emyway platform by reviewing the “Training-Related Data Collection and Privacy Information” and selecting the agreement option. Participation was uncompensated. All data were deidentified prior to analysis, and access to the dataset was restricted to the research team through secure, password-protected servers.

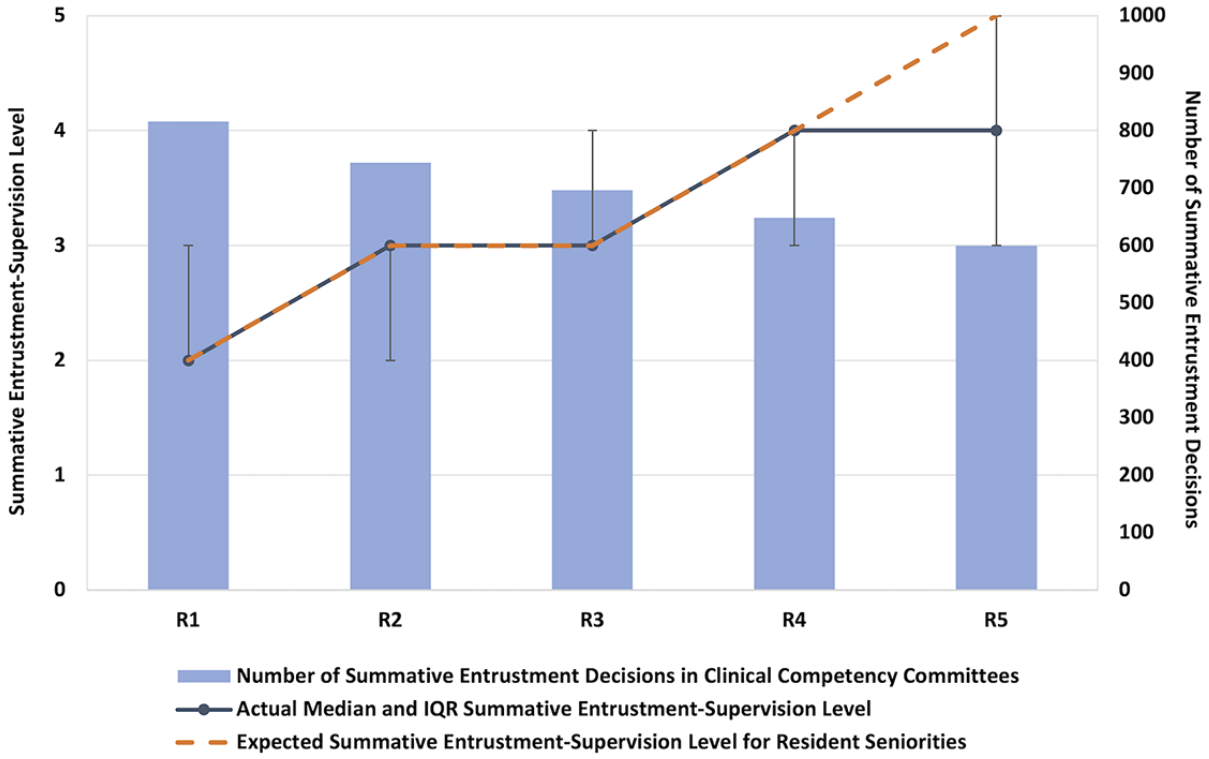
Results

Progression of Entrustment-Supervision Levels Across Training Years

Observed CCC summative entrustment-supervision levels were compared with prespecified targets for each training year (Figure 1). Median (IQR) observed levels increased across training stages—from 2 (2-3) in R1, to 3 (2-3) in R2, 3 (3-4) in R3, and 4 (3-4) in both R4 and R5. Despite this upward progression, observed supervision levels did not consistently reach the expected targets in later training years.

XSL•FO
RenderX

Figure 1. Progression of Clinical Competency Committee (CCC) summative entrustment-supervision levels across residency years in Taiwan otorhinolaryngology-head and neck surgery training programs. The solid line with error bars represents the observed median (IQR) CCC summative entrustment-supervision level by residency year, overlaid with the expected entrustment-supervision targets for each training stage (dashed line). Shaded bars (right y-axis) indicate the number of CCC summative entrustment decisions contributing to each residency year. Expected targets were prespecified according to the national framework (R1: Level 2; R2-R3: Level 3; R4: Level 4; R5: Level 5).



Median observed supervision levels achieved expectations in R1 through R4 but fell below expected targets in R5. The discrepancy between expected and observed supervision levels increased with advancing training year. The largest gap was observed in R5, where the expected supervision level was Level 5, whereas the median observed level remained Level 4 (IQR 3-4).

Characteristics by Achievement Status

The dataset included 3504 CCC summative entrustment decisions across 12 EPAs for 292 residents nationwide. Sixteen residents contributed data from only one CCC reporting cycle (July 2025), accounting for 192 observations. Overall, 2558 assessments (73%) met the expected entrustment-supervision targets, whereas 946 assessments (27%) did not. Comparisons of assessments that met versus did not meet expected targets are shown in Table 2. Nonattainment was significantly associated with larger resident quotas, northern geographic region, advanced training year, lower CCC review cadence (1 vs 2 reviews per year), and delayed EPA sequencing (all $P<.001$). Nonattainment was most frequent in senior training years, particularly R4 (20.2%) and R5 (42.2%), in programs conducting a single CCC review per year (15.8%), and in delayed EPAs (28.2%). In contrast, hospital level, municipality type, faculty size, and resident sex were not significantly associated with achievement status.

XSL•FO
RenderX

Table 2. Characteristics of institutions, entrustable professional activity (EPA) assessments, and participants by achievement status.

Variable	Meet expectations (n=2558), n (%)	Did not meet expectations (n=946), n (%)	*P* value
Program characteristics			
Hospital level			.34
Medical center	1890 (73.9)	714 (75.5)	
Nonmedical center	668 (26.1)	232 (24.5)	
Municipality type			.35
Special	2216 (86.6)	808 (85.4)	
Nonspecial	342 (13.4)	138 (14.6)	
Resident quota			<.001[a]
≥3	913 (35.7)	419 (44.3)	
2	874 (34.2)	278 (29.4)	
1	771 (30.1)	249 (26.3)	
Number of faculty			.93
>10	1543 (60.3)	569 (60.2)	
≤10	1015 (39.7)	377 (39.9)	
Geographic region			<.001[a]
Northern	1274 (49.8)	538 (56.9)	
Central	638 (24.9)	238 (25.2)	
Southern	556 (21.7)	140 (14.8)	
Eastern	90 (3.5)	30 (3.2)	
Resident characteristics			
Seniority			<.001[a]
R1	701 (27.4)	115 (12.2)	
R2	596 (23.3)	148 (15.6)	
R3	603 (23.6)	93 (9.8)	
R4	457 (17.9)	191 (20.2)	
R5	201 (7.9)	399 (42.2)	
Sex			.41
Male	1786 (69.8)	674 (71.3)	
Female	772 (30.2)	272 (28.8)	
EPA sequencing			
Planned training stage			.007[a]
Delayed	609 (23.8)	267 (28.2)	
Early	1949 (76.2)	679 (71.78)	
Assessment cadence			
Two assessments			<.001[a]
No	43 (1.7)	149 (15.8)	
Yes	2515 (98.3)	797 (84.2)	

[a]Statistically significant results, defined as $P<.05$.

XSL•FO
RenderX

Factors Associated With Achieving Expected Entrustment Levels

Results of the univariable and multivariable logistic regression analyses are presented in Table 3.

In univariable analyses, training in the southern region was associated with higher odds of meeting expected entrustment targets. In contrast, larger resident quotas, delayed EPA sequencing, lower CCC review cadence, and advancing training year were associated with lower odds of attainment. Compared with R1, the likelihood of meeting expected supervision targets declined in R2, R4, and R5. Hospital level, municipality type, faculty size, and resident sex were not significantly associated with attainment.

In the multivariable model, 2 system-level variables remained independently associated with lower attainment: lower CCC review cadence (adjusted odds ratio [aOR] 0.41, 95% CI 0.28-0.61) and delayed EPA sequencing (EPA6, EPA10, and EPA11), defined as higher-complexity activities typically achieved later in training, were associated with lower entrustment attainment compared with early EPAs (aOR 0.75, 95% CI 0.63-0.91). The resident quota was excluded from the multivariable model due to strong dependency on faculty number, which led to redundancy and model instability. The faculty number was retained as it more directly reflects supervision capacity. Training in the southern region remained positively associated with attainment (aOR 1.42, 95% CI 1.12-1.79). Resident seniority also remained a strong predictor, with significantly lower attainment observed in R2, R4, and R5 relative to R1. Hospital level, faculty size, resident sex, and other geographic regions were not independently associated with attainment after adjustment.

XSL•FO
RenderX

Table 3. Univariable and multivariable logistic regression analyses of factors associated with achieving the expected entrustment-supervision level.

Variable	Univariable model		Multivariable model	
	OR[a] (95% CI)	*P* value	Adjusted OR (95% CI)	*P* value
Program characteristics				
Hospital level				
Medical center vs Nonmedical center	0.92 (0.77-1.09)	.34	0.85 (0.66-1.09)	.21
Municipality type				
Special vs Nonspecial	1.11 (0.89-1.37)	.35	—[b]	—
Resident quota				
2 vs 1	1.02 (0.83-1.24)	.88	—	—
≥3 vs 1	0.70 (0.59-0.85)	<.001	—	—
Number of faculty				
>10 vs ≤10	1.01 (0.87-1.17)	.93	1.11 (0.88-1.39)	.37
Geographic region				
Central vs Northern	1.13 (0.95-1.36)	.18	0.99 (0.81-1.21)	.92
Southern vs Northern	1.68 (1.36-2.07)	<.001	1.42 (1.12-1.79)	.003[c]
Eastern vs Northern	1.27 (0.83-1.94)	.28	1.31 (0.81-2.12)	.27
Resident characteristics				
Seniority				
R2 vs R1	0.66 (0.51-0.86)	.002	0.69 (0.53-0.91)	.008[c]
R3 vs R1	1.06 (0.79-1.43)	.68	1.10 (0.82-1.48)	.51
R4 vs R1	0.39 (0.30-0.51)	<.001	0.40 (0.31-0.52)	<.001[c]
R5 vs R1	0.08 (0.06-0.11)	<.001	0.11 (0.08-0.14)	<.001[c]
Sex				
Male vs Female	0.93 (0.79-1.10)	.41	0.93 (0.77-1.13)	.48
EPA[d] sequencing				
Delayed vs Early	0.79 (0.67-0.94)	.007	0.75 (0.63-0.91)	.003[c]
Assessment cadence				
2 CCC[e] evaluations				
No vs Yes	0.09 (0.06-0.13)	<.001	0.41 (0.28-0.61)	<.001[c]

[a]OR: odds ratio.

[b]Not applicable.

[c]Statistically significant results, defined as $P<.05$.

[d]EPA: entrustable professional activity.

[e]CCC: Clinical Competency Committee.

Stratified Analyses

Stratified analyses suggested heterogeneity of associations across training stages and EPA sequencing (Table 4 and Table S1 in Multimedia Appendix 2).

When stratified by training year, hospital level and faculty size showed differing associations across stages. Training in a medical center was positively associated with attainment in R2 but negatively associated in R5, whereas larger faculty size showed the opposite pattern. Regional differences were most evident in R5, where attainment was higher outside the northern region. Sex was associated with attainment only in R2. Lower CCC review cadence showed the strongest negative association in R2 but was not significant in R5.

XSL•FO
RenderX

Table 4. Multivariable logistic regression analyses of factors associated with achieving the expected entrustment-supervision level, stratified by resident seniority.

Variable, seniority	OR[a] (95% CI)	*P* value
Program characteristics		
Hospital level		
Medical center vs Nonmedical center		
Overall	0.85 (0.66-1.09)	.21
R1	1.68 (0.91-3.11)	.09
R2	1.92 (1.07-3.46)	.029[b]
R3	0.69 (0.35-1.36)	.28
R4	0.64 (0.38-1.07)	.09
R5	0.32 (0.18-0.58)	<.001[b]
Number of faculty		
>10 vs ≤ 10		
Overall	1.11 (0.88-1.39)	.37
R1	1.09 (0.61-1.94)	.78
R2	0.53 (0.30-0.93)	.027[b]
R3	1.34 (0.74-2.41)	.34
R4	1.14 (0.71-1.81)	.59
R5	2.69 (1.52-4.77)	.001[b]
Geographic region		
Central vs Northern		
Overall	0.99 (0.81-1.21)	.92
R1	0.79 (0.49-1.28)	.33
R2	1.10 (0.69-1.76)	.69
R3	0.86 (0.52-1.43)	.56
R4	0.93 (0.59-1.49)	.77
R5	2.26 (1.42-3.59)	.001[b]
Southern vs Northern		
Overall	1.42 (1.12-1.79)	.003[b]
R1	1.44 (0.84-2.48)	.19
R2	1.12 (0.66-1.90)	.67
R3	1.50 (0.73-3.08)	.27
R4	1.09 (0.67-1.76)	.73
R5	2.73 (1.60-4.65)	<.001[b]
Eastern vs Northern		
Overall	1.31 (0.81-2.12)	.27
R1	5.35 (0.68-41.98)	.11
R2	0.81 (0.32-2.06)	.66
R3	2.43 (0.29-20.59)	.42
R4	0.80 (0.30-2.09)	.65
R5	2.77 (1.07-7.15)	.036[b]
Resident characteristics		

XSL•FO
RenderX

Variable, seniority	OR[a] (95% CI)	*P* value
Sex		
Male vs Female		
Overall	0.93 (0.77-1.13)	.48
R1	1.34 (0.87-2.08)	.19
R2	1.67 (1.02-2.74)	.043[b]
R3	0.83 (0.49-1.41)	.49
R4	0.74 (0.48-1.16)	.19
R5	0.78 (0.52-1.18)	.24
EPA[c] sequencing		
Delayed vs Early		
Overall	0.75 (0.63-0.91)	.003[b]
R1	0.36 (0.24-0.54)	<.001[b]
R2	2.67 (1.57-4.55)	<.001[b]
R3	0.44 (0.28-0.70)	.001[b]
R4	0.69 (0.47-1.00)	.052
R5	0.91 (0.60-1.37)	.64
Assessment cadence		
2 CCC[d] evaluations		
No vs Yes		
Overall	0.41 (0.28-0.61)	<.001[b]
R1	N/A[e]	N/A
R2	0.02 (0.00-0.14)	<.001[b]
R3	N/A	N/A
R4	N/A	N/A
R5	0.83 (0.51-1.33)	.43

[a]OR: odds ratio.

[b]Statistically significant results, defined as $P<.05$.

[c]EPA: entrustable professional activity.

[d]CCC: Clinical Competency Committee.

[e]N/A: Estimates for R1, R3, and R4 were not reported because of insufficient sample size.

Associations between delayed EPA sequencing and attainment also varied by training year. Lower attainment was observed in R1 and R3, higher attainment in R2, and no clear association in R5. When stratified by EPA sequencing, training in the southern region remained positively associated with attainment for both early and delayed EPAs (Table S1 in Multimedia Appendix 2). Larger faculty size was associated with higher attainment only for delayed EPAs. Seniority effects differed by sequencing: for early EPAs, attainment declined progressively with advancing training year, whereas for delayed EPAs, attainment was higher in R2 but lower in R4 and R5. Lower CCC review cadence remained negatively associated with attainment across both sequencing groups.

Program-level analyses demonstrated substantial variability between training programs. Relative to the reference program, approximately half of the nonreference programs showed significantly different odds of meeting expected entrustment targets, with both higher and lower attainment observed (Tables S2 and S3 in Multimedia Appendix 2).

Discussion

Principal Findings

This national analysis of CCC entrustment decisions demonstrates that aggregation of WBA data alone is insufficient to ensure alignment between expected and documented trainee autonomy. Despite the synthesis of longitudinal evidence through CCC deliberation, attainment of prespecified

XSL•FO
RenderX

entrustment targets declined in the most advanced stages of training, revealing a reproducible late-stage gap between expected and observed supervision levels. Importantly, this pattern was not random in distribution; 2 system-level features—delayed sequencing of complex EPAs and lower CCC review cadence—were independently associated with reduced attainment. These findings suggest that advanced entrustment decisions are influenced by both trainee capability and the architecture of the assessment systems used to generate and interpret evidence. Viewed through the lens of learning analytics, national digital assessment platforms may function as diagnostic infrastructures for CBME, revealing structural constraints on trainee progression and identifying system-level design features that are associated with how competence is documented, interpreted, and ultimately trusted.

Seniority Gap and Late-Stage Entrustment Plateau

A central finding of this study is the persistence of a late-stage discrepancy between expected and observed entrustment levels. The expected entrustment targets were defined by expert consensus within the national EPA framework and represent intended progression rather than empirically validated thresholds; thus, the observed nonattainment rate should be interpreted relative to these consensus-based benchmarks. Although supervision levels increased progressively across training years, the median CCC entrustment level plateaued at Level 4 in the most advanced stage of training, despite an expected Level 5 target. This pattern indicates that senior residents frequently approached—but did not consistently reach—the threshold for unsupervised practice as judged by CCCs. Rather than reflecting isolated measurement variability, the consistency of this pattern across programs suggests a structural feature of the assessment system influencing how advanced competence is documented.

Several mechanisms may plausibly contribute to this phenomenon. First, the evidentiary burden required to justify Level 5 entrustment is inherently higher than that required for intermediate supervision levels [33]. CCCs typically require repeated demonstrations of safe and independent clinical performance across varied contexts before endorsing unsupervised practice [34]. When opportunities to demonstrate such autonomy are limited, evidence supporting the highest entrustment level may be less likely to accumulate even among highly competent trainees [35]. In programmatic assessment systems, this phenomenon has been described as an “evidence threshold effect,” in which higher-stakes decisions demand stronger or more consistent documentation before committees are willing to endorse them [36,37].

Second, the clinical contexts encountered by senior residents may paradoxically make documentation of autonomy more difficult. Advanced trainees often manage more complex cases or work within higher-risk environments where supervisors remain appropriately involved in decision-making [28,38]. In such situations, supervisors may exercise greater caution when documenting unsupervised practice, even when the trainee demonstrates high competence. Consequently, observed supervision levels may reflect contextual supervisory practices rather than a lack of capability. Similar patterns have been described in digital assessment systems where entrustment documentation reflects both learner performance and clinical workflow constraints [39].

Third, the timing of complex EPAs may compress the period during which evidence of independent practice can accumulate. When advanced EPAs are introduced late in training, residents may have fewer opportunities to demonstrate repeated independent performance before CCC review. This shortened observation window can limit the evidentiary base available to committees and may contribute to conservative decision thresholds. Within digital assessment environments, the delayed introduction of complex competencies has been associated with reduced opportunity for iterative feedback and progressive autonomy documentation [40].

Fourth, limited opportunities to formally supervise novice learners may partly explain the lower attainment of Level 5 entrustment. In Taiwan’s otolaryngology residency training, senior residents may guide junior residents or medical students during ward care, outpatient preparation, emergency consultations, and selected procedural settings. However, these supervisory activities are often informal, context-dependent, and not consistently documented as EPA evidence. As a result, CCCs may have insufficient direct evidence to support Level 5 entrustment, even when senior residents demonstrate high clinical competence.

Together, these mechanisms suggest that the observed plateau in senior entrustment levels reflects an interaction between trainee progression and assessment system architecture. In digital CBME infrastructures, learning analytics derived from aggregated assessment data can help reveal these structural dynamics by identifying patterns of progression that are not readily visible within individual programs [5].

Influence of CCC Review Cadence on Entrustment Outcomes

Another important finding of this study is the association between CCC review cadence and attainment of expected entrustment levels. Programs conducting only a single CCC review within an academic year showed significantly lower odds of meeting expected supervision targets compared with programs conducting 2 reviews annually. Because residents within these programs accumulate comparable numbers of EPAs, the difference is unlikely to reflect assessment volume alone. Instead, CCC review cadence likely influences how evidence is synthesized and interpreted over time. More frequent committee reviews allow CCC members to examine developmental trajectories across multiple time points, confirm whether performance patterns are stable, and provide feedback that can guide targeted clinical exposure or documentation practices. Earlier feedback cycles also enable trainees and supervisors to address evidentiary gaps before final summative decisions are made. In contrast, when only a single review occurs annually, the evidentiary window is compressed, and recent or atypical events may disproportionately influence committee judgments.

Digital assessment platforms have increasingly enabled longitudinal visualization of learner performance through

XSL•FO
RenderX

dashboards and aggregated data displays, allowing committees to review developmental trajectories rather than isolated observations [41]. These tools have been proposed as mechanisms to strengthen CCC deliberation and support data-informed decision-making in CBME [5]. From a systems perspective, CCC review cadence therefore represents a modifiable governance feature within national CBME infrastructures. When combined with digital assessment platforms capable of aggregating longitudinal evidence, justified review cycles may improve the interpretability of assessment data and reduce uncertainty surrounding advanced entrustment decisions.

Reversal Pattern in R2 Entrustment Attainment

An additional notable finding was the reversal pattern observed in R2, where attainment appeared higher for delayed EPAs but lower for early EPAs compared with R1. One possible explanation is that delayed EPAs in R2 are encountered in more structured and supervised contexts, often involving selected cases with closer faculty guidance. In contrast, early EPAs are performed more frequently in routine clinical settings, where expectations for independent performance may be higher, and variability in workload and assessment opportunities may be greater. This difference in context may contribute to the observed pattern. However, this interpretation remains speculative and warrants further investigation using longitudinal and WBA-level data.

Program-Level Variation in Entrustment Decisions

Substantial variability in attainment rates between programs was also observed in this study. Such variability indicates that CCC entrustment outcomes reflect not only resident performance but also local implementation practices and institutional contexts [42]. Differences in case mix, supervisory structures, and clinical exposure may legitimately influence how autonomy is granted within different training environments [43].

However, wide variation in summative entrustment decisions may also arise from differences in CCC calibration, documentation practices, or local thresholds for endorsing unsupervised practice [44]. In digital CBME systems, these variations can become visible when assessment data from multiple institutions are aggregated and analyzed collectively. National learning analytics infrastructures, therefore, provide an opportunity to distinguish contextually appropriate variation from potential inconsistency in assessment practices [40,45,46].

National benchmarking of entrustment outcomes may help training programs evaluate whether local practices align with broader system patterns. When combined with faculty development and CCC calibration initiatives, such benchmarking can strengthen the fairness and comparability of high-stakes progression decisions while preserving necessary contextual flexibility.

Regional Variation in Entrustment Attainment

Programs in the southern region consistently demonstrated higher entrustment attainment across analyses. One possible explanation is regional variation in clinical exposure, supervisory practices, or CCC calibration, which may influence how entrustment decisions are documented. Differences in case mix, faculty-resident interaction patterns, or assessment culture may also contribute. However, as these factors were not directly measured, this finding should be interpreted cautiously. Further qualitative and mixed methods studies are needed to clarify the mechanisms underlying regional variation.

Implications for Digital CBME Systems

The findings of this study highlight the growing role of digital assessment infrastructures in shaping CBME implementation. Historically, assessment systems primarily served to document trainee performance within individual programs. Increasingly, however, national digital platforms enable aggregation and analysis of assessment data across institutions, creating opportunities for large-scale learning analytics [46-48].

Such systems allow educational leaders to examine patterns of trainee progression, identify structural constraints within training pathways, and evaluate how design features of assessment systems influence entrustment decisions. When interpreted alongside programmatic assessment theory, these analytics can guide targeted adjustments to EPA sequencing, CCC review cadence, and evidence sampling strategies.

In this sense, national digital assessment platforms function not only as repositories of assessment data but also as diagnostic infrastructures for CBME governance. By enabling cross-program benchmarking and longitudinal data analysis, these systems support more transparent and defensible progression decisions while providing actionable insights for improving training system design.

Limitations

Several limitations should be considered when interpreting these findings. First, the cross-sectional design precludes causal inference and does not capture individual longitudinal trajectories of entrustment progression. Although the nationwide dataset provides a comprehensive system-level snapshot of CCC decisions, future studies using longitudinal data across multiple training cycles could more precisely examine developmental patterns in entrustment. Second, several contextual variables that may influence CCC judgments were not available in the analytic dataset, including variation in underlying WBA volume, quality, distribution, observed patterns, rotation-specific clinical exposure, case complexity, rater composition, and institutional autonomy policies. These factors may contribute to residual variability in entrustment outcomes across programs. Third, CCC decisions are inherently context dependent, reflecting both trainee performance and local supervisory practices. While such contextual variation may limit strict comparability across institutions, it is an expected feature of programmatic assessment systems and can provide meaningful information for national benchmarking and quality improvement. Fourth, although each CCC summative entrustment decision was treated as a distinct observation, some residents contributed multiple decisions across EPAs or review cycles. Residual within-resident correlation may therefore remain, and future analyses using multilevel or clustered models could further clarify resident- and program-level sources of variation. Moreover, while this binary outcome aligns with CCC progression decisions, it

XSL•FO
RenderX

reduces the informational richness of the ordinal entrustment scale and may obscure smaller differences between adjacent supervision levels. Future studies using ordinal or longitudinal analytic approaches may provide a more granular understanding of entrustment progression. Additionally, the generalizability of these findings to other CBME contexts, even those with structured EPA frameworks and centralized assessment systems, remains uncertain and requires further study to confirm. Finally, this study examined summative entrustment decisions without linking them to downstream clinical outcomes. Future research integrating digital assessment data with measures of independent clinical performance and patient safety outcomes would further strengthen the evidence base supporting entrustment decisions within CBME.

Conclusions

Using nationwide CCC entrustment data derived from a digital assessment platform, this study demonstrates that under-attainment of expected supervision targets in CBME is not random but systematically associated with structural features of assessment system design. Under-attainment clustered in advanced training stages, in late-sequenced EPAs, and in programs with lower CCC review cadence, suggesting the presence of a seniority gap within current assessment architectures. These findings indicate that documented entrustment outcomes reflect not only trainee capability but also how assessment evidence is generated, aggregated, and interpreted within programmatic assessment systems. National benchmarking of CCC decisions, enabled by shared digital assessment infrastructures, can therefore provide actionable learning analytics to identify structural constraints on trainee progression and inform system-level improvements in EPA sequencing, CCC governance, and evidence sampling. In this way, national digital assessment infrastructures can evolve from data repositories into strategic learning analytics systems that support continuous improvement of competency-based training.

Acknowledgments

We thank the Taiwan Society of Otorhinolaryngology-Head and Neck Surgery and all its faculties and resident physicians for using the Joint Commission of Taiwan's Emyway platform. We also thank the information technology team of Dalin Tzu Chi Hospital for their support with the platform. Additionally, we are grateful for the administrative assistance provided by Chiu-Ping Wang, Shu-Hwei Fan, Uan-Shr Jan, and Wan-Ning Luo in this project. They received no additional compensation for their contributions. The authors declare the use of generative AI in the research and writing process. According to the GAIDeT taxonomy (2025), the following tasks were delegated to GAI tools under full human supervision: Proofreading and editing. The GAI tool used was ChatGPT-Thinking 5.5. Responsibility for the final manuscript lies entirely with the authors. GAI tools are not listed as authors and do not bear responsibility for the final outcomes.

Funding

This study was supported by the National Science and Technology Council of the Republic of China (Taiwan) under grants NSTC 110-2511-H-567-001-MY2 and NSTC 112-2410-H-567-001-MY3, and was partially funded by Cardinal Tien Hospital under grants CTH113A-2212 and CTH114A-2210. This study was also funded by the Ministry of Health and Welfare of the Republic of China (Taiwan) under the Healthy Taiwan Sprout grant C-0002. The funders had no role in the design or conduct of the study; the collection, management, analysis, or interpretation of the data; the preparation, review, or approval of the manuscript; or the decision to submit the manuscript for publication.

Data Availability

Data were obtained from the Joint Commission of Taiwan and the Taiwan Society of Otorhinolaryngology-Head and Neck Surgery Emyway database with permission for use. The datasets used and analyzed during this study are available from the corresponding author on reasonable request.

Authors' Contributions

JWC, RBY, KNL, WCH, PCW, and LJL contributed to the conceptualization and investigation of the study. JWC, YO, and XBY were responsible for data curation, while YO and XBY conducted the formal analysis. The methodology was developed by JWC, YO, and XBY. Project administration and provision of resources were managed by JWC and WCH. Supervision was provided by JWC, WCH, CHL, and MCC. JWC, RBY, YO, and XBY prepared the original draft of the manuscript, and JWC, RBY, KNL, YO, XBY, WCH, PCW, LJL, CHL, and MCC contributed to the review and editing of the manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Taiwan Society of Otorhinolaryngology–Head and Neck Surgery Entrustable Professional Activities Assessment Framework for Resident Physician Training.

XSL•FO
RenderX

[PDF File (Adobe PDF File), 791 KB - mededu_v12i1e95066_app1.pdf]

Multimedia Appendix 2
Multivariable logistic regression analyses of factors associated with achieving the expected entrustment–supervision level, stratified by entrustable professional activity (EPA) sequencing and program-level fixed-effect estimates from logistic regression (maximum likelihood estimates) comparing each residency training program (HospID) with the reference program (H01) for attainment of the expected Clinical Competency Committee (CCC) summative entrustment–supervision level. Data on the program-level odds ratios (ORs) and Wald 95% CIs for attainment of the expected CCC summative entrustment–supervision level, comparing each training program (HospID) with the reference program (H01).
[DOCX File , 79 KB - mededu_v12i1e95066_app2.docx]

References

1. Cooper D, Holmboe ES. Competency-based medical education at the front lines of patient care. N Engl J Med 2025;393(4):376-388. [doi: 10.1056/NEJMra2411880] [Medline: 40700689]
2. Lucey CR, Thibault GE, Ten Cate O. Competency-based, time-variable education in the health professions: crossroads. Acad Med 2018;93(3S Competency-Based, Time-Variable Education in the Health Professions):S1-S5. [doi: 10.1097/ACM.0000000000002080] [Medline: 29485479]
3. Carraccio CL, Englander R. From flexner to competencies: reflections on a decade and the journey ahead. Acad Med 2013;88(8):1067-1073. [doi: 10.1097/ACM.0b013e318299396f] [Medline: 23807096]
4. Hoffman M, Dedow L, Boscamp J. Competency-based and less time-bound: a new approach to the macro-structure of a medical school curriculum. Med Educ Online 2024;29(1):2343205 [FREE Full text] [doi: 10.1080/10872981.2024.2343205] [Medline: 38626425]
5. Hsiao CT, Chou FC, Hsieh CC, Chang LC, Hsu CM. Developing a competency-based learning and assessment system for residency training: analysis study of user requirements and acceptance. J Med Internet Res 2020;22(4):e15655 [FREE Full text] [doi: 10.2196/15655] [Medline: 32286233]
6. Carraccio C, Englander R, Gilhooly J, Mink R, Hofkosh D, Barone MA, et al. Building a framework of entrustable professional activities, supported by competencies and milestones, to bridge the educational continuum. Acad Med 2017;92(3):324-330. [doi: 10.1097/ACM.0000000000001141] [Medline: 26959225]
7. Encandela JA, Shaull L, Jayas A, Amiel JM, Brown DR, Obeso VT, et al. Entrustable professional activities as a training and assessment framework in undergraduate medical education: a case study of a multi-institutional pilot. Med Educ Online 2023;28(1):2175405 [FREE Full text] [doi: 10.1080/10872981.2023.2175405] [Medline: 36794397]
8. Shorey S, Lau TC, Lau ST, Ang E. Entrustable professional activities in health care education: a scoping review. Med Educ 2019;53(8):766-777. [doi: 10.1111/medu.13879] [Medline: 30945329]
9. Ten Cate O, Schwartz A, Chen HC. Assessing trainees and making entrustment decisions: on the nature and use of entrustment-supervision scales. Acad Med 2020;95(11):1662-1669. [doi: 10.1097/ACM.0000000000003427] [Medline: 32324633]
10. ten Cate O, Scheele F. Competency-based postgraduate training: can we bridge the gap between theory and clinical practice? Acad Med 2007;82(6):542-547. [doi: 10.1097/ACM.0b013e31805559c7] [Medline: 17525536]
11. Alharbi NS. Evaluating competency-based medical education: a systematized review of current practices. BMC Med Educ 2024;24(1):612 [FREE Full text] [doi: 10.1186/s12909-024-05609-6] [Medline: 38831271]
12. Bok HGJ, Teunissen PW, Favier RP, Rietbroek NJ, Theyse LFH, Brommer H, et al. Programmatic assessment of competency-based workplace learning: when theory meets practice. BMC Med Educ 2013;13:123 [FREE Full text] [doi: 10.1186/1472-6920-13-123] [Medline: 24020944]
13. Ten Cate O, Hart D, Ankel F, Busari J, Englander R, Glasgow N, et al. Entrustment decision making in clinical training. Acad Med 2016;91(2):191-198. [doi: 10.1097/ACM.0000000000001044] [Medline: 26630606]
14. Ten Cate O, Carraccio C, Damodaran A, Gofton W, Hamstra SJ, Hart DE, et al. Entrustment decision making: extending Miller's pyramid. Acad Med 2021;96(2):199-204. [doi: 10.1097/ACM.0000000000003800] [Medline: 33060399]
15. ten Cate O. Entrustability of professional activities and competency-based training. Med Educ 2005;39(12):1176-1177. [doi: 10.1111/j.1365-2929.2005.02341.x] [Medline: 16313574]
16. Carney PA, Sebok-Syer SS, Pusic MV, Gillespie CC, Westervelt M, Goldhamer MEJ. Using learning analytics in clinical competency committees: increasing the impact of competency-based medical education. Med Educ Online 2023;28(1):2178913 [FREE Full text] [doi: 10.1080/10872981.2023.2178913] [Medline: 36821373]
17. Peng CR, Schertzer KA, Caretta-Weyer HA, Sebok-Syer SS, Lu W, Tansomboon C, et al. Assessment of entrustable professional activities using a web-based simulation platform during transition to emergency medicine residency: mixed methods pilot study. JMIR Med Educ 2021;7(4):e32356 [FREE Full text] [doi: 10.2196/32356] [Medline: 34787582]
18. Karpinski J, Fantaye AW, Hendry P, Wiesenfeld L, Whiting S, Lochnan H, et al. Implementation, operationalization, and optimization of clinical competency committees in postgraduate training programs: a scoping review. Acad Med 2025;100(12):1543-1554. [doi: 10.1097/ACM.0000000000006180] [Medline: 40758982]

XSL•FO
RenderX

19. Montgomery KB, Mellinger JD, McLeod MC, Jones A, Zmijewski P, Sarosi GA, et al. Decision-making confidence of clinical competency committees for entrustable professional activities. JAMA Surg 2024;159(7):801-808. [doi: 10.1001/jamasurg.2024.0809] [Medline: 38717759]
20. Chiang YH, Yu HC, Chung H, Chen JW. Implementing an entrustable professional activities programmatic assessments for nurse practitioner training in emergency care: a pilot study. Nurse Educ Today 2022;115:105409. [doi: 10.1016/j.nedt.2022.105409] [Medline: 35636245]
21. Lin CP, Hsu WC, Wang GL, Su HC, Chung HC, Hung SH, et al. Factors associated with attainment of ad hoc entrustability among Taiwan otolaryngology resident physicians: a nationwide cross-sectional study. BMC Med Educ 2025;25(1):1135 [FREE Full text] [doi: 10.1186/s12909-025-07679-6] [Medline: 40751193]
22. Younas A, Khan RA, Yasmin R. Entrustment in physician-patient communication: a modified Delphi study using the EPA approach. BMC Med Educ 2021;21(1):497 [FREE Full text] [doi: 10.1186/s12909-021-02931-1] [Medline: 34544394]
23. Kwan J, Crampton R, Mogensen LL, Weaver R, van der Vleuten CPM, Hu WCY. Bridging the gap: a five stage approach for developing specialty-specific entrustable professional activities. BMC Med Educ 2016;16:117 [FREE Full text] [doi: 10.1186/s12909-016-0637-x] [Medline: 27097981]
24. Ekstrom V. Digital choice architecture in medical education: applying behavioral economics to online learning environments. JMIR Med Educ 2026;12:e86497 [FREE Full text] [doi: 10.2196/86497] [Medline: 41650469]
25. Holmboe ES, Yamazaki K, Nasca TJ, Hamstra SJ. Using longitudinal milestones data and learning analytics to facilitate the professional development of residents: early lessons from three specialties. Acad Med 2020;95(1):97-103 [FREE Full text] [doi: 10.1097/ACM.0000000000002899] [Medline: 31348058]
26. Chan T, Sebok-Syer S, Thoma B, Wise A, Sherbino J, Pusic M. Learning analytics in medical education assessment: the past, the present, and the future. AEM Educ Train 2018;2(2):178-187 [FREE Full text] [doi: 10.1002/aet2.10087] [Medline: 30051086]
27. Cheung WJ, Bhanji F, Gofton W, Hall AK, Karpinski J, Richardson D, et al. Design and implementation of a national program of assessment model - integrating entrustable professional activity assessments in Canadian specialist postgraduate medical education. Perspect Med Educ 2024;13(1):44-55 [FREE Full text] [doi: 10.5334/pme.956] [Medline: 38343554]
28. Guo FC, Chen YT, Hsu WC, Wang PC, Chen M, Chen JW. EMYWAY workplace-based entrustable professional activities assessments in otolaryngology residency training: a nationwide experience. Otolaryngol Head Neck Surg 2025;172(4):1242-1253. [doi: 10.1002/ohn.1104] [Medline: 39739526]
29. Chen JW, Tu HL, Chang CH, Hsu WC, Wang PC, Liao CH, et al. Automated evaluation of reflection and feedback quality in workplace-based assessments by using natural language processing: cross-sectional competency-based medical education study. JMIR Med Educ 2025;11:e81718 [FREE Full text] [doi: 10.2196/81718] [Medline: 41124694]
30. Krecko LK, Jung S, Martin S, Krebsbach C, Rosser AA, Stahl C, et al. Enhancing the value of surgical entrustable professional activities through integrative learning analytics. J Surg Educ 2023;80(10):1370-1377. [doi: 10.1016/j.jsurg.2023.07.018] [Medline: 37596105]
31. Ten Cate O, Chen HC, Hoff RG, Peters H, Bok H, van der Schaaf M. Curriculum development for the workplace using entrustable professional activities (EPAs): AMEE guide no. 99. Med Teach 2015;37(11):983-1002. [doi: 10.3109/0142159X.2015.1060308] [Medline: 26172347]
32. Ten Cate O. Nuts and bolts of entrustable professional activities. J Grad Med Educ 2013;5(1):157-158 [FREE Full text] [doi: 10.4300/JGME-D-12-00380.1] [Medline: 24404246]
33. Santen SA, Wolff MS, Saxon K, Juneja N, Bassin B. Factors affecting entrustment and autonomy in emergency medicine: "How much rope do I give them?". West J Emerg Med 2019;20(1):58-63 [FREE Full text] [doi: 10.5811/westjem.2018.10.39843] [Medline: 30643602]
34. Bohnen JD, George BC, Williams RG, Schuller MC, DaRosa DA, Torbeck L, et al. The feasibility of real-time intraoperative performance assessment with SIMPL (System for Improving and Measuring Procedural Learning): early experience from a multi-institutional trial. J Surg Educ 2016;73(6):e118-e130. [doi: 10.1016/j.jsurg.2016.08.010] [Medline: 27886971]
35. Baugh TP, Franzese CB. Extremes in otolaryngology resident surgical case numbers: an update. Otolaryngol Head Neck Surg 2017;156(6):1060-1066. [doi: 10.1177/0194599817696305] [Medline: 28319672]
36. Ryan MS, Gielissen KA, Shin D, Perera RA, Gusic M, Ferenchick G, et al. How well do workplace-based assessments support summative entrustment decisions? a multi-institutional generalisability study. Med Educ 2024;58(7):825-837. [doi: 10.1111/medu.15291] [Medline: 38167833]
37. Yu P, Fang C, Liu X, Fu W, Ling J, Yan Z, et al. Performance of ChatGPT on the Chinese postgraduate examination for clinical medicine: survey study. JMIR Med Educ 2024;10:e48514 [FREE Full text] [doi: 10.2196/48514] [Medline: 38335017]
38. Tiyyagura G, Balmer D, Chaudoin L, Kessler D, Khanna K, Srivastava G, et al. The greater good: how supervising physicians make entrustment decisions in the pediatric emergency department. Acad Pediatr 2014;14(6):597-602. [doi: 10.1016/j.acap.2014.06.001] [Medline: 25439158]
39. Preiksaitis C, Ashenburg N, Bunney G, Chu A, Kabeer R, Riley F, et al. The role of large language models in transforming emergency medicine: scoping review. JMIR Med Inform 2024;12:e53787 [FREE Full text] [doi: 10.2196/53787] [Medline: 38728687]

XSL•FO
RenderX

40. Preiksaitis C, Hughes J, Kabeer R, Dixon W, Rose C. Quantifying emergency medicine residency learning curves using natural language processing: retrospective cohort study. JMIR Med Educ 2025;11:e82326 [FREE Full text] [doi: 10.2196/82326] [Medline: 41364786]
41. Friedman KA, Raimo J, Spielmann K, Chaudhry S. Resident dashboards: helping your clinical competency committee visualize trainees' key performance indicators. Med Educ Online 2016;21:29838 [FREE Full text] [doi: 10.3402/meo.v21.29838] [Medline: 27037226]
42. Teunissen PW, Kogan JR, Ten Cate O, Gruppen LD, Lingard LA. Learning in practice: a valuation of context in time-variable medical training. Acad Med 2018;93(3S Competency-Based, Time-Variable Education in the Health Professions):S22-S26. [doi: 10.1097/ACM.0000000000002070] [Medline: 29485483]
43. Clarke MJ, Edwards RA, White BAA. Surgical assessments: current state, ideal state, and considerations for implementation. a qualitative interview study of faculty and trainees. J Surg Educ 2025;82(7):103534. [doi: 10.1016/j.jsurg.2025.103534] [Medline: 40393345]
44. Karpinski J, Stewart J, Oswald A, Dalseg TR, Atkinson A, Frank JR. Competency-based medical education at scale: a road map for transforming national systems of postgraduate medical education. Perspect Med Educ 2024;13(1):24-32 [FREE Full text] [doi: 10.5334/pme.957] [Medline: 38371306]
45. Powell L, Nour R, Zidoun Y, Kaladhara S, Al Suwaidi H, Zary N. A web-based public health intervention for addressing vaccine misinformation: protocol for analyzing learner engagement and impacts on the hesitancy to vaccinate. JMIR Res Protoc 2022;11(5):e38034 [FREE Full text] [doi: 10.2196/38034] [Medline: 35451967]
46. Ahmed AAE, Biswas A, Bempong-Ahun N, Perić I, O'Flynn EP. Barriers and enablers to the production of open access medical education platforms: scoping review. JMIR Med Educ 2025;11:e65306 [FREE Full text] [doi: 10.2196/65306] [Medline: 41202285]
47. Elamin A, Obeidat M, Davis G. The ePortfolio in UK cardiology training: time for a new digital platform? Br J Cardiol 2021;28(3):31 [FREE Full text] [doi: 10.5837/bjc.2021.031] [Medline: 35747697]
48. Grosjean J, Dufour F, Benis A, Januel JM, Staccini P, Darmoni SJ. Digital health education for the future: the SaNuRN (santé numérique rouen-nice) consortium's journey. JMIR Med Educ 2024;10:e53997 [FREE Full text] [doi: 10.2196/53997] [Medline: 38693686]

Abbreviations

aOR: adjusted odds ratio
CBME: competency-based medical education
CCC: Clinical Competency Committee
EPA: entrustable professional activity
ORL–HNS: Otorhinolaryngology-Head and Neck Surgery
WBA: workplace-based assessment

Edited by A Stone; submitted 13.Mar.2026; peer-reviewed by S Jung, AS Moura, M Ncube; comments to author 03.Apr.2026; revised version received 25.Apr.2026; accepted 28.Apr.2026; published 27.May.2026.

Please cite as:
Chen JW, Yu RB, Lin KN, Ono Y, Yu XB, Hsu WC, Wang PC, Liao LJ, Liao CH, Chen M
Learning Analytics of a National Entrustable Professional Activities Platform: Cross-Sectional Study of System-Level Constraints on Advanced Entrustment in Competency-Based Medical Education
JMIR Med Educ 2026;12:e95066
URL: https://mededu.jmir.org/2026/1/e95066
doi:10.2196/95066
PMID:

©Jeng-Wen Chen, Rui-Bin Yu, Kai-Nan Lin, Yukiko Ono, Xuan-Bing Yu, Wei-Chung Hsu, Pa-Chun Wang, Li-Jen Liao, Chun-Hou Liao, Mingchih Chen. Originally published in JMIR Medical Education (https://mededu.jmir.org), 27.May.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Education, is properly cited. The complete bibliographic information, a link to the original publication on https://mededu.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Artificial Intelligence for Assessment and Feedback in Medical Education: Bibliometric Mapping Study and Thematic Evidence Map

Zihang Zhao[1,2*], MD; Zihan Liu[1,2*], MSc; Liang Guo[1,2*], MSc; Teng Pan[3], MSc; Yousheng Zhang[4], MSc; Chenxiang Miao[2], BM; Yiting Ge[2], BM; Yipeng Wang[2], BM; Xin Hu[1,2], MD; Xin Wang[1,2], MD; Ruipeng Zhang[1,2], MD; Zhiyong Hou[1,2], MD

[1]Department of Orthopedic Surgery, The Third Hospital of Hebei Medical University, Shijiazhuang, Hebei, China

[2]Hebei Medical University, Shijiazhuang, Hebei, China

[3]Department of Oncology, Shijiazhuang People's Hospital, Shijiazhuang, Hebei, China

[4]Faculty of Medicine, Dentistry and Health Sciences, The University of Melbourne, Melbourne, Victoria, Australia

[*]these authors contributed equally

Corresponding Author:
Zhiyong Hou, MD
Department of Orthopedic Surgery
The Third Hospital of Hebei Medical University
No. 139 Ziqiang Road
Shijiazhuang, Hebei, 050051
China
Phone: 86 18533112800
Email: drzyhou@hebmu.edu.cn

Abstract

Background: Artificial intelligence (AI), particularly generative AI and large language models, is increasingly used for assessment-related tasks in medical education. Existing overviews often address AI in medical education broadly, limiting assessment-specific interpretation of functions, settings, learner stages, and responsible-AI reporting domains.

Objective: This study aims to map the literature on AI for assessment and feedback in medical education, including publication trends, bibliometric structure, assessment functions, AI types, settings, learner stages, and reporting of validity, reliability, fairness, integrity, transparency, human oversight, implementation, and governance.

Methods: We conducted a bibliometric mapping study incorporating structured thematic evidence-map coding. Web of Science Core Collection, Scopus, and PubMed were searched from January 1, 2015, to April 8, 2026. Document selection and main evidence-map coding were based primarily on titles, abstracts, and bibliographic metadata, with targeted ambiguity resolution. Because reporting domains may appear mainly in full-text sections, all 435 included records underwent full-text sensitivity analysis for the 8 reporting domains. Coding reliability was assessed by two coders using percent agreement and Cohen κ before adjudication. Exploratory subgroup analyses and an excluding-2026 partial-year sensitivity analysis were conducted.

Results: Searches identified 14,968 records; 435 were included after deduplication and selection. Overall, 399 (91.7%) records were indexed in the post-ChatGPT period. Generative AI was coded in 310 (71.3%) records, and large language models in 301 (69.2%) records. In the assessment-function umbrella analysis, learner performance evaluation accounted for 270 (62.1%) records, feedback for 93 (21.4%) records, assessment content generation for 65 (14.9%) records, and other or unclear functions for 7 (1.6%) records. The most common settings were board-style examinations (n=151, 34.7%) and written examinations (n=88, 20.2%); undergraduate medical education was the most represented learner stage (n=172, 39.5%). Full-text–confirmed reporting was most frequent for reliability (n=288, 66.2%) and implementation (n=231, 53.1%); intermediate for validity (n=158, 36.3%), fairness (n=132, 30.3%), and transparency (n=130, 29.9%); and less frequent for governance (n=57, 13.1%), human oversight (n=46, 10.6%), and integrity (n=26, 6%). Stage 1 κ values ranged from 0.785 to 0.895, and stage 2 κ values ranged from 0.809 to 0.880. Excluding 65 partial-year 2026 records did not change the overall interpretation.

Conclusions: The indexed English-language literature on AI for assessment and feedback in medical education expanded rapidly in the post-ChatGPT period and was concentrated in generative AI, large language models, examination-oriented assessment,

XSL•FO
RenderX

and undergraduate medical education. Future studies should complement examination benchmarking with authentic assessment contexts, distinguish assessment content generation from learner-facing evaluation and feedback, and report responsible assessment domains more consistently.

(JMIR Med Educ 2026;12:e98949) doi:10.2196/98949

KEYWORDS

artificial intelligence; generative artificial intelligence; large language model; medical education; learner assessment; feedback; bibliometric analysis; evidence map; summative assessment; validity; fairness; governance

Introduction

Artificial intelligence (AI) is increasingly shaping medical education through learning analytics, simulation, adaptive learning, clinical reasoning support, assessment, feedback, and faculty development [1-4]. Earlier applications often involved machine learning, deep learning, natural language processing, computer vision, or rule-based systems applied to defined educational or clinical-learning tasks [1,2,4,5]. More recently, generative artificial intelligence (GenAI) and large language models have changed the field because they can generate explanations, assessment items, simulated dialogue, feedback, and other educational content with relatively low technical barriers for educators and learners [6-10].

Assessment and feedback deserve focused attention within this broader literature because assessment in medical education is not only a measurement activity [11]. It supports learning, informs feedback, contributes to judgments of competence, and may influence progression, remediation, certification, and professional trust [12-15]. When AI is used to generate assessment materials, score responses, interpret performance, or provide feedback, the relevant concerns extend beyond technical accuracy to include validity, reliability, fairness, academic integrity, transparency, human oversight, implementation, accountability, and governance [16-20].

Existing reviews and bibliometric studies have provided valuable overviews of AI in medical education and health professions education [3,6,21,22]. These broad syntheses may combine studies of curriculum, learning support, simulation, attitudes toward AI, assessment, institutional readiness, and faculty development. As a result, they may not isolate which assessment functions are being studied, which learner stages and settings are represented, and how often assessment-quality or governance domains are explicitly reported. This distinction has become more important in the GenAI era, when examination benchmarking, question generation, automated feedback, and performance interpretation may appear within the same broad literature but carry different educational implications.

A further distinction is needed between assessment content generation and learner-facing evaluation or feedback. A study that uses AI to generate multiple-choice questions, a study that evaluates a model's performance on licensing-style examinations, and a study that uses AI to support feedback on learner performance all relate to assessment, but they do not answer the same educational question. Content generation concerns the production or refinement of assessment materials, whereas learner-facing assessment involves scoring, performance evaluation, interpretation, feedback, or decisions about learner progress. This distinction follows from assessment theory, which links evidence requirements to intended interpretations, uses, and consequences [11-13,15].

This study aimed to map the literature on AI for assessment and feedback in medical education through a bibliometric mapping study and thematic evidence map. Specifically, we examined temporal publication trends, bibliometric structure, assessment functions, AI types, assessment settings, learner stages, and reporting of validity, reliability, fairness, integrity, transparency, human oversight, implementation, and governance. We also conducted a full-text sensitivity analysis for these reporting domains, assessed 2-stage coding reliability, disaggregated assessment content generation from learner-facing assessment functions, and evaluated exploratory subgroup and partial-year sensitivity patterns.

Methods

Study Design and Reporting Framework

We conducted a bibliometric mapping study incorporating structured thematic evidence-map coding to characterize the literature on AI for assessment and feedback in medical education [23-25]. The design emphasizes field characterization, including publication trends, bibliometric structure, assessment functions, settings, learner stages, AI categories, and reporting domains. Reporting was guided by the BIBLIO (bibliometric reviews of biomedical literature) checklist, where applicable [26], with additional evidence-map and review-methodology references used to support the structured thematic evidence-map component [24,25]. The completed BIBLIO checklist is provided in Multimedia Appendix 1. Because this study was designed as a bibliometric mapping study, no protocol was registered.

The analytic unit was the screened bibliographic record. Document selection and the main evidence-map coding were conducted primarily from titles, abstracts, and bibliographic metadata, with targeted ambiguity resolution when needed. Because reporting domains such as fairness, integrity, transparency, human oversight, implementation, and governance may appear primarily in full-text sections, we conducted a full-text sensitivity analysis for the 8 reporting domains across all 435 included records.

Methodological transparency was supported through a 2-layer reporting structure in which the main manuscript describes the study design and analytic workflow, and the supplementary materials provide database-specific search strategies, coding

XSL•FO
RenderX

definitions, sensitivity analyses, agreement tables, and bibliometric-parameter settings.

Data Sources and Search Strategy

We searched Web of Science Core Collection, Scopus, and PubMed from January 1, 2015, to April 8, 2026. The 3-database strategy was selected to balance multidisciplinary citation indexing, broad scholarly coverage, and biomedical indexing relevant to medical education, assessment, and AI.

The search strategy combined terms related to AI, including machine learning, deep learning, GenAI, large language models, ChatGPT, GPT, foundation models, and language models; assessment and feedback, including scoring, grading, examinations, testing, objective structured clinical examinations, item generation, question generation, performance evaluation, automated scoring, formative assessment, and summative assessment; and medical education, including medical students, undergraduate medical education, graduate medical education, residency, clerkship, internship, continuing medical education, and continuing professional development. Database-specific search strings, limits, and final hit counts are provided in Multimedia Appendix 2.

The final searches identified 14,968 records: 4381 from Web of Science Core Collection, 7256 from Scopus, and 3331 from PubMed.

Eligibility Criteria

Records were eligible if they addressed AI-supported assessment or feedback in a medical education context. Assessment was defined as any AI-supported activity used to generate, score, interpret, or provide feedback on learner performance or assessment content in a medical education context.

Eligible records addressed learner assessment, performance evaluation, automated scoring, feedback, examination activity, item generation, question generation, formative or summative assessment, interpretation of learner performance, or related assessment processes. Eligible educational populations included medical students, residents, graduate medical trainees, physicians, clinicians, continuing medical education participants, or mixed medical education learner groups. Eligible publication types included original research, validation studies, reviews, scoping reviews, systematic reviews, bibliometric or mapping studies, and mixed methods studies.

We excluded records focused only on clinical diagnosis, prognosis, image interpretation, clinical prediction, or decision support without a medical education assessment or feedback context. We also excluded records focused only on teaching support, curriculum delivery, admissions, recruitment, patient education, nonmedical learner populations, or general AI commentary without an assessment or feedback component. Nonanalytic publication types, including conference abstracts, editorials, news items, and letters, were excluded. The search was limited to English-language records indexed in the selected databases.

Deduplication and Document Selection

Database exports were merged and deduplicated through multistage matching based on DOI, PubMed identifier, normalized title, publication year, and available bibliographic metadata. After multistage deduplication, 8378 records remained for document selection.

Document selection was conducted primarily at the title and abstract level because the analytic unit was the screened bibliographic record, and the study aimed to map the structure and characteristics of a large assessment-focused literature. Bibliographic metadata were used to support record identification and classification. For records with insufficient title or abstract information, targeted ambiguity resolution was conducted through supplementary web-assisted review and, when necessary, PDF-assisted review. This ambiguity-resolution process was used to clarify eligibility and main evidence-map categories while preserving the bibliographic-record-based mapping design.

Following document selection, 7943 records were excluded, and 435 records comprised the final screened analysis cohort.

Data Extraction and Evidence-Map Coding

For each included record, we extracted bibliometric metadata including publication year, authors, journal, DOI, country, and institution information derived from author affiliations, keywords, cited references, and citation-related fields used for bibliometric visualization.

We then applied a structured evidence-map coding framework to classify publication type, study design, AI type, assessment function, assessment-function umbrella, assessment setting, learner stage, GenAI relevance, post-ChatGPT period, and reporting domains visible in titles or abstracts. AI type categories included large language model, GenAI, machine learning, deep learning, natural language processing, computer vision, rule-based or expert systems, wearable or sensor-based AI, and other AI approaches.

Assessment function was coded using nonmutually exclusive categories, including formative assessment, summative assessment, feedback, automated scoring, performance evaluation, interpretation of learner performance, item generation, question generation, and other functions. To address conceptual overlap, we also coded an assessment-function umbrella variable distinguishing learner performance evaluation, feedback, assessment content generation, and other or unclear functions. Question generation and item generation were treated as assessment content generation and were not automatically coded as summative assessment unless the record explicitly linked them to examination use, learner performance evaluation, or assessment decision-making.

Assessment setting and learner stage were coded separately. Learner-stage categories included undergraduate medical education, residency, graduate medical education, physician or clinician continuing education, continuing medical education, mixed or multiple learner stages, and unclear. Because individual records could involve more than 1 AI type, assessment function, or assessment setting, evidence-map

XSL•FO
RenderX

categories were treated as nonmutually exclusive unless otherwise specified. Additional coding definitions and operational details are provided in Multimedia Appendix 3.

Full-Text Sensitivity Analysis for Reporting Domains

The main evidence-map categories remained based on the final manually reviewed coding based on titles and abstracts, with targeted ambiguity resolution. The initial reporting-domain variables captured explicit reporting visible in titles or abstracts of validity, reliability, fairness, integrity, transparency, human oversight, implementation, and governance. These estimates, based on titles and abstracts, were retained as the baseline.

Because these domains may be discussed mainly in full-text Methods, Results, Discussion, Limitations, tables, figures, supplementary materials, or governance-related sections, we conducted a full-text sensitivity analysis for all 8 reporting domains across all 435 included records. A domain was coded as present only when the full text explicitly reported, discussed, evaluated, or proposed considerations relevant to that domain. Coding was conservative; domains were not inferred from the topic, method, or study design alone. If the correct full text could not be verified or was unavailable, domain status was coded as unclear and not inferred.

Integrity was defined as explicit attention to academic integrity, assessment security, or safeguards against misuse of AI in assessment contexts. Human oversight was defined as explicit attention to human review, supervision, escalation, or educator involvement in AI-supported assessment or feedback processes. Transparency was defined as explicit reporting or discussion of model explainability, disclosure, interpretability, model reporting, or transparency of AI-supported assessment processes. Governance was defined as explicit attention to policy, accountability, regulation, institutional governance, privacy or data governance, or formal guidance for responsible AI use.

Full-text–confirmed estimates were used as the primary basis for interpreting reporting-domain findings. Estimates based on titles and abstracts were reported as the baseline and sensitivity comparison.

Coding Reliability and Adjudication

We assessed coding reliability through a 2-stage independent coding process across the full cohort of 435 records. Two coders used the same codebook. Before independent coding, the coders reviewed the codebook and coding examples to calibrate the interpretation of category definitions. In stage 1, the coders independently coded the evidence-map categories coded from titles and abstracts using available bibliographic information. Stage 1 variables included publication type, study design, AI type exact set, assessment function exact set, assessment-function umbrella, assessment setting, learner stage, GenAI relevance, and post-ChatGPT period.

In Stage 2, the 2 coders independently coded the 8 full-text reporting domains: validity, reliability, fairness, integrity, transparency, human oversight, implementation, and governance. Agreement was calculated before adjudication using percent agreement and Cohen κ. For multilabel variables, exact-set agreement was used. Disagreements were resolved through consensus review against the codebook and source record, and final adjudicated codes were assigned after discussion. For low-prevalence variables, including integrity and human oversight, Cohen κ was interpreted alongside percent agreement because prevalence can affect κ estimates. Detailed agreement tables are provided in Multimedia Appendix 3.

Bibliometric and Evidence-Map Analysis

We summarized annual publication counts, publication types, study designs, AI types, assessment functions, assessment settings, learner stages, GenAI relevance, and post-ChatGPT status using counts and percentages. The post-ChatGPT period was defined operationally as records indexed from November 2022 onward, corresponding to the public release of ChatGPT.

Bibliometric analyses characterized journal output, citation patterns, country-level collaboration, institutional contribution, author contribution, keyword structure, thematic evolution, and recent co-citation structure. Country and institution fields were derived from author affiliation metadata after harmonization. Country-level findings were interpreted as patterns within the indexed English-language records captured by the search, not complete global research output. Weighted author metrics were used to reduce inflation from large multiauthor records. Bibliometric software, thresholds, normalization, clustering, layout, counting, and weighting parameters are reported in Multimedia Appendix 3.

Evidence maps were constructed for the assessment function by AI type and assessment setting by learner stage. Because evidence-map categories were nonmutually exclusive, matrix counts represent coded study characteristics, not mutually exclusive study strata.

Subgroup and Partial-Year Sensitivity Analyses

We conducted exploratory subgroup analyses for full-text–confirmed reporting domains by GenAI relevance, post-ChatGPT period, and assessment content generation status. Fisher exact tests were used for subgroup comparisons, with Benjamini-Hochberg correction for multiplicity. All 8 reporting domains were retained in the supplementary subgroup tables. In the main text, subgroup findings are reported selectively and interpreted cautiously as descriptive patterns.

Because 2026 was incomplete at the time of the search, annual trend figures identify 2026 as a partial year through April 8, 2026. We also conducted an excluding-2026 sensitivity analysis to evaluate whether the main reporting-domain patterns were robust to the removal of partial-year records. Detailed subgroup and partial-year sensitivity results are provided in Multimedia Appendix 3.

Ethical Considerations

This study analyzed bibliographic records and published reports and did not involve human participants, patients, private identifiable information, or intervention delivery. Ethics review and informed consent were not required.

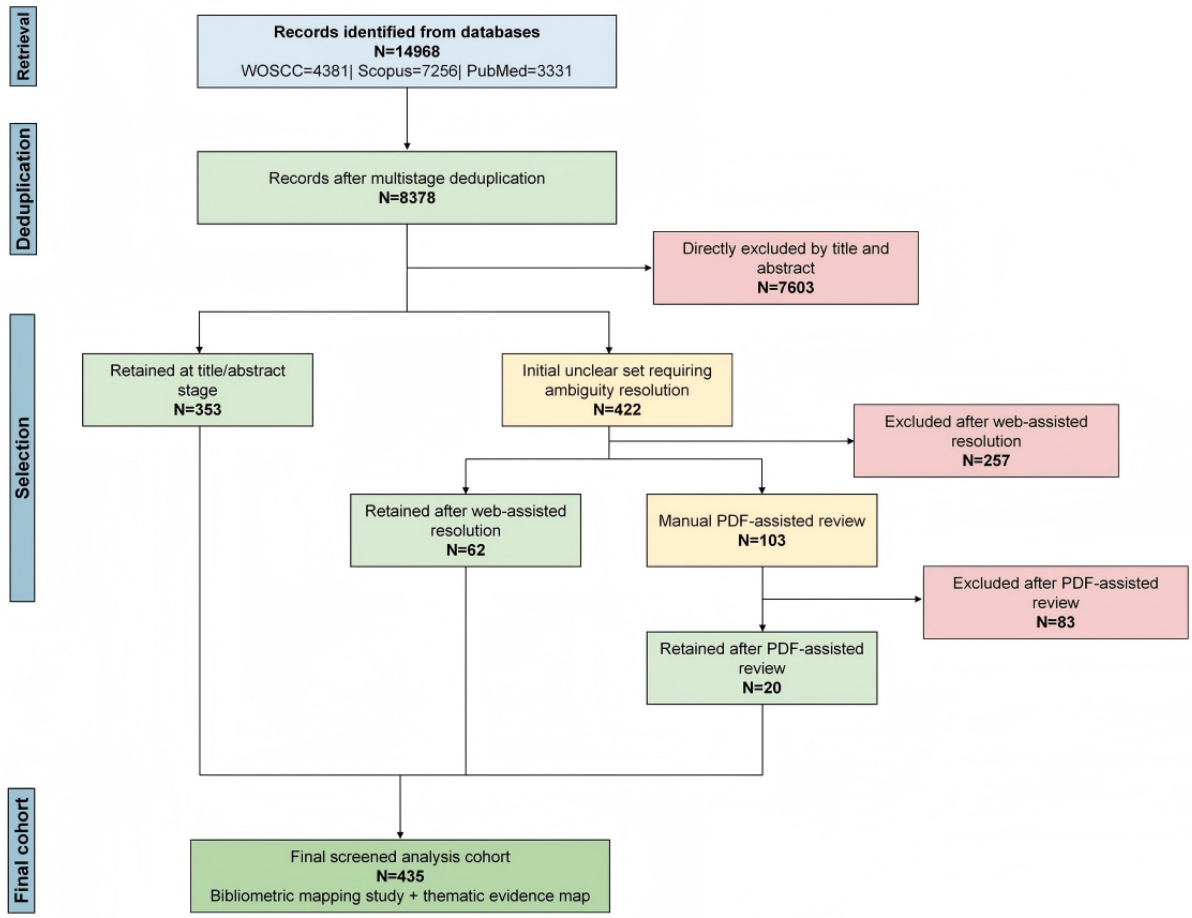
XSL•FO
RenderX

Results

Study Selection and Temporal Growth

The database searches identified 14,968 records, including 4381 from Web of Science Core Collection, 7256 from Scopus, and 3331 from PubMed. After multistage deduplication, 8378 records remained for document selection. During document selection, 7943 records were excluded, leaving 435 records in the final screened analysis cohort (Figure 1).

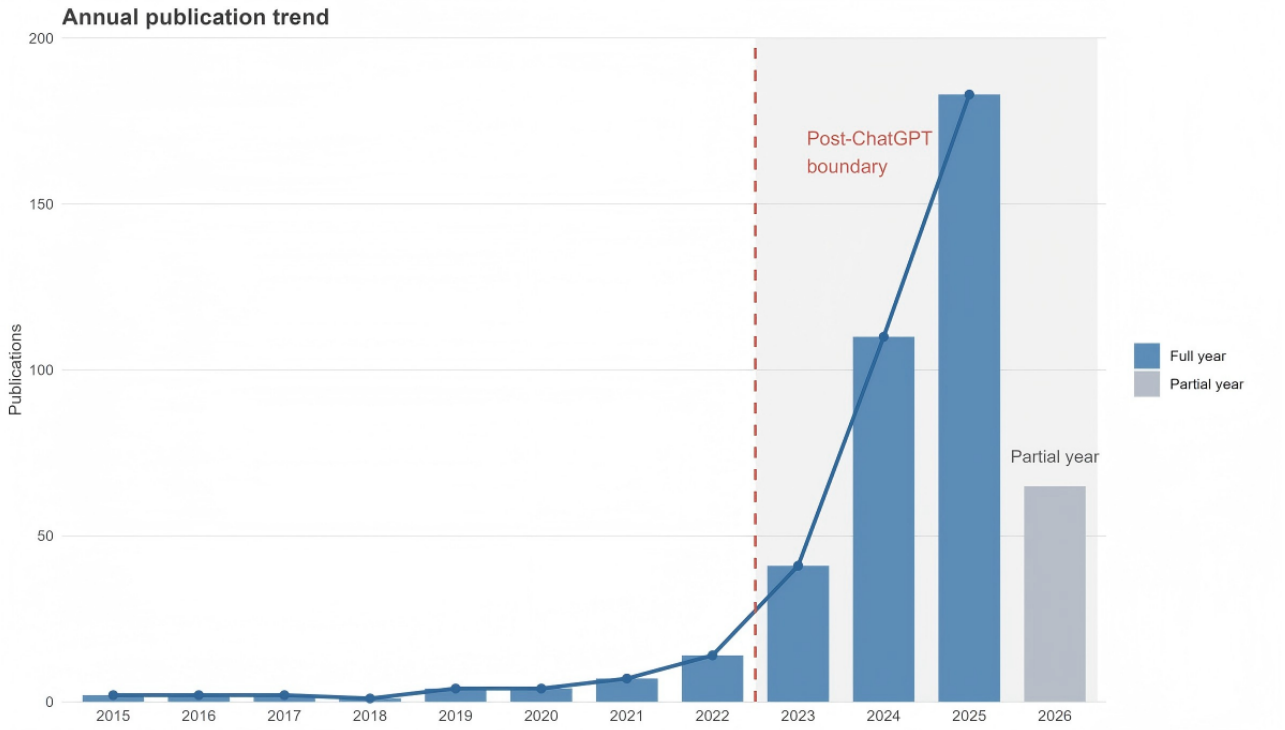
Figure 1. Study selection flow diagram. Records were identified from Web of Science Core Collection, Scopus, and PubMed; deduplicated through a multistage matching workflow; and screened during document selection. The final cohort included records retained at the title and abstract stage, after web-assisted ambiguity resolution, and after manual PDF-assisted review. The diagram accounts for all 8378 records entering document selection, comprising 7943 exclusions and 435 included records.



Original research accounted for 334 (76.8%) records, followed by validation studies (n=46, 10.6%), reviews (n=25, 5.7%), and mixed methods studies (n=19, 4.4%). Annual output remained low through 2021, with 2 records in 2015, 2 in 2016, 2 in 2017, 1 in 2018, 4 in 2019, 4 in 2020, and 7 in 2021. Output then increased to 14 records in 2022, 41 in 2023, 110 in 2024, and 183 in 2025. By the search date of April 8, 2026, 65 records had been indexed in 2026; this year was treated as a partial year. Overall, 399 (91.7%) records were indexed in the post-ChatGPT period, defined as November 2022 onward (Figure 2).

XSL•FO
RenderX

Figure 2. Annual publication trend in artificial intelligence for assessment and feedback in medical education, 2015-2026. Bars show the number of records by publication year. The shaded region denotes the post-ChatGPT period, operationally defined as records indexed from November 2022 onward. The line connects complete calendar years through 2025. The year 2026 is shown as a partial year through the final search date of April 8, 2026.

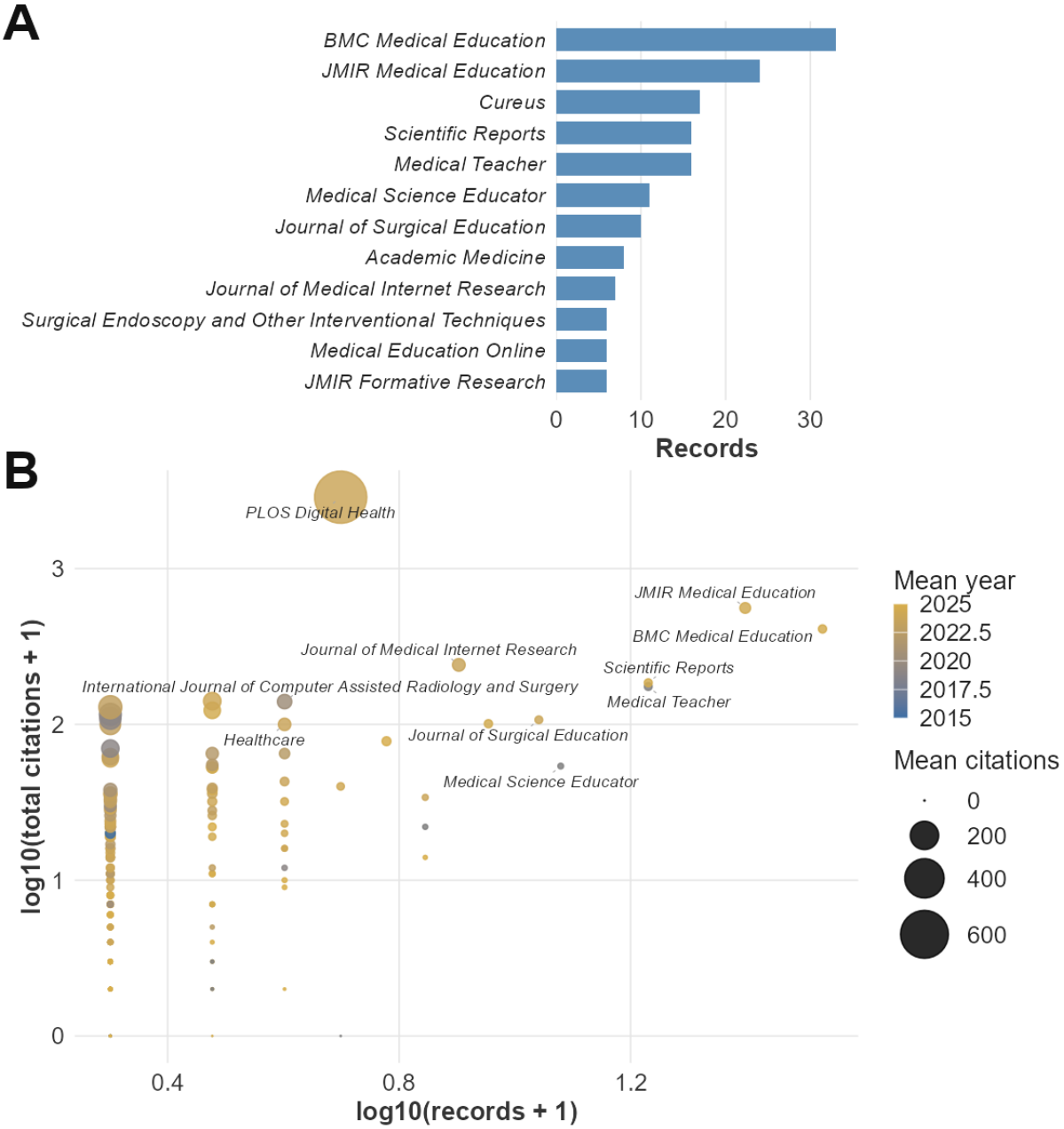


Bibliometric Structure and Publication Characteristics

The final screened analysis cohort was distributed across a broad journal base, with output concentrated in medical education, digital health, and specialty education journals (Figure 3). Based on the final manually reviewed coding worksheet, the leading journals by record count were *BMC Medical Education* (n=32, 7.4%), *JMIR Medical Education* (n=24, 5.5%), *Cureus* (n=17, 3.9%), *Scientific Reports* (n=16, 3.7%), and *Medical Teacher* (n=16, 3.7%). In the journal landscape, *JMIR Medical Education* and *BMC Medical Education* combined visible publication volume with citation activity. *Journal of Medical Internet Research* and *PLOS Digital Health* had smaller record counts but high citation indicators, reflecting the visibility of highly cited digital health and large language model assessment papers. These citation indicators should be interpreted cautiously because publication timing and early benchmark visibility can strongly influence citation accumulation.

XSL•FO
RenderX

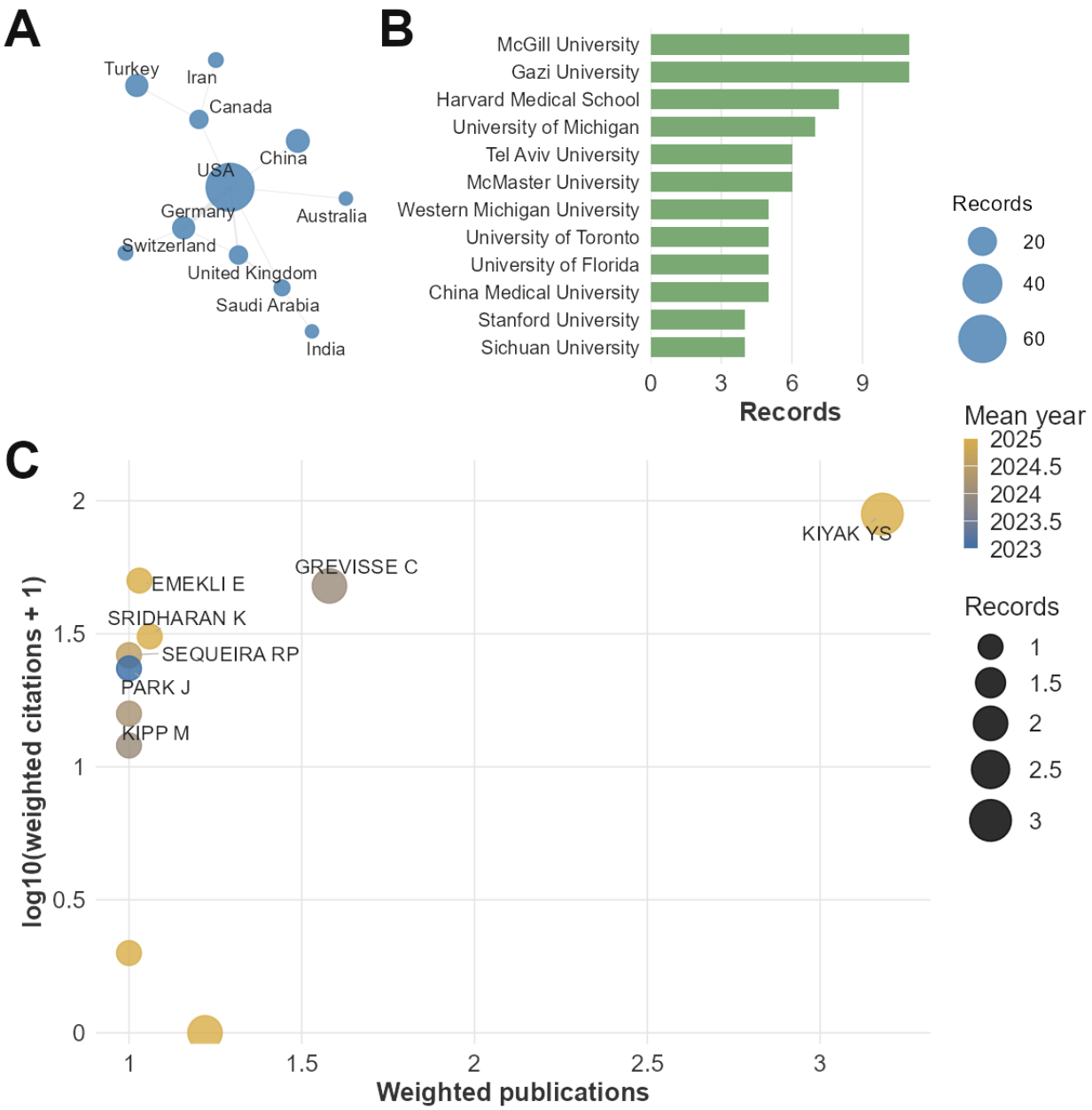
Figure 3. Journal landscape of the final screened analysis cohort. Panel A shows the leading journals by publication count. Panel B plots journal output against citation indicators, with bubble size representing average citations per record and color indicating mean publication year. Counts and citation indicators reflect indexed English-language records captured by the search and should be interpreted descriptively.



Country-level collaboration was interpreted as a pattern within the indexed English-language records captured by the search, not as complete global research output (Figure 4). Within this indexed corpus, the United States was the largest contributing country and the main collaboration hub in the coauthorship network, with 62 (14.3%) records. China (n=13, 3%), Germany (n=12, 2.8%), and Turkey (n=12, 2.8%) were the next most represented countries. The collaboration network indicated international linkage among several leading contributors, but the English-language and database-indexing restrictions limit conclusions about global research capacity or regional participation.

XSL•FO
RenderX

Figure 4. Collaboration and contributor structure in the indexed English-language corpus. Panel A shows the country-level collaboration network, with node size proportional to record count and line thickness proportional to collaboration strength. Panel B shows leading institutions by record count. Panel C shows leading authors using weighted publication output and citation profile, with color indicating mean publication year. Country, institution, and author patterns reflect records captured by the search and do not represent complete global output.



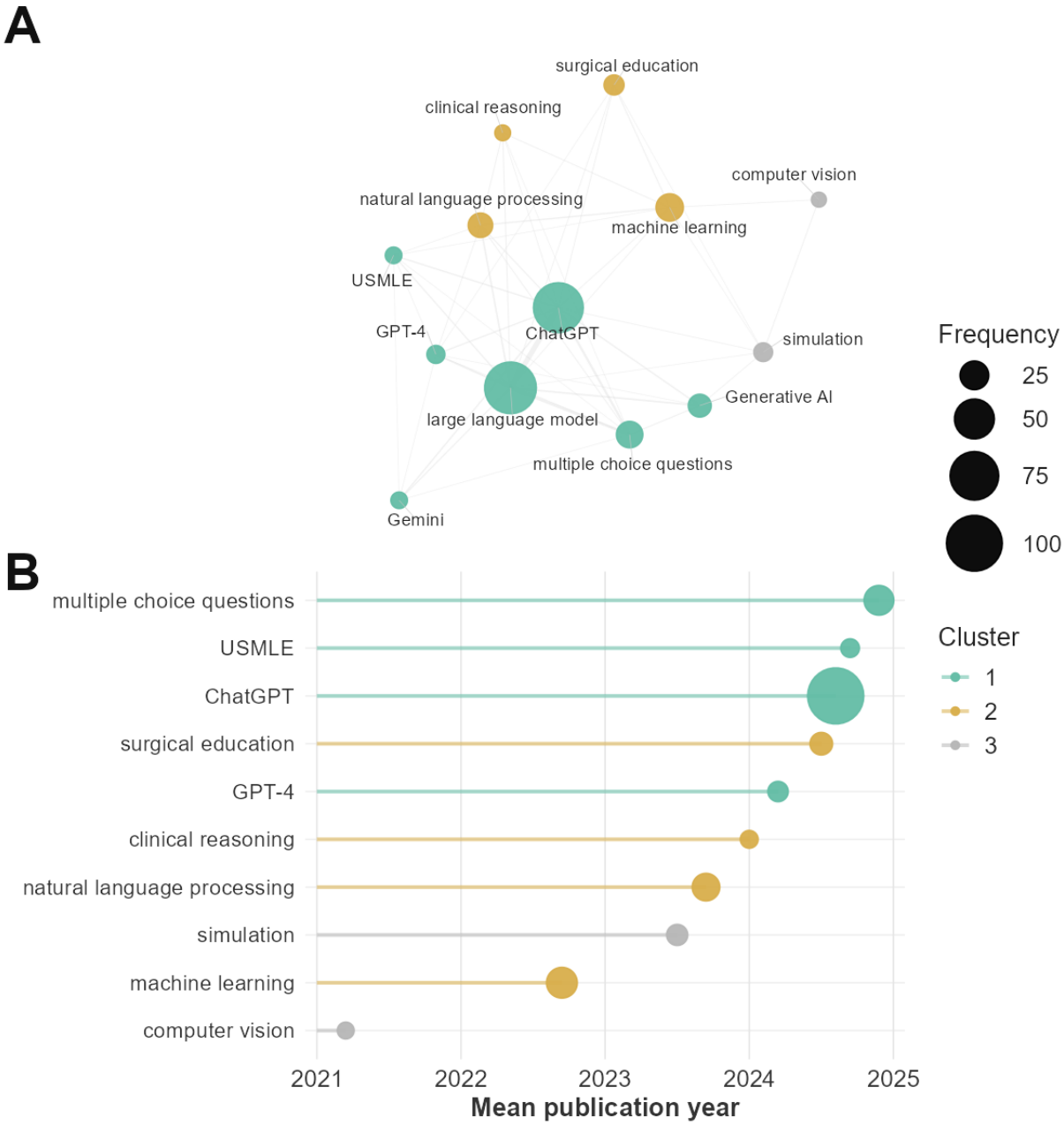
Institution and author summaries were retained as descriptive bibliometric indicators. In the supplementary institution table, McGill University, Gazi University, and the University of Michigan appeared among the leading institution entries. In the weighted author landscape, Kiyak YS showed the strongest weighted publication profile, with Grevisse C and Emekli E also appearing prominently in the combined output and citation space. These author- and institution-level findings should be interpreted cautiously because publication timing, team size, affiliation-string harmonization, and citation accumulation can influence bibliometric visibility.

The keyword structure reflected the field's recent shift toward GenAI and examination-oriented assessment (Figure 5). Frequent normalized keywords included large language model (n=109, 25.1%), ChatGPT (n=101, 23.2%), machine learning (n=29, 6.7%), multiple-choice questions (n=27, 6.2%), and natural language processing (n=23, 5.3%). The co-occurrence structure showed a recent cluster centered on large language models, ChatGPT, GPT-4, multiple-choice questions, and the United States Medical Licensing Examination (USMLE), alongside clusters linking natural language processing or machine learning with clinical reasoning, surgical education, simulation, and computer vision. The recent post-ChatGPT co-citation heatmap showed a compact intellectual structure anchored by early examination-benchmarking papers and synthesis work, including studies by Kung et al [8], Gilson et al [9], and Sallam [7] (Multimedia Appendix 4). This pattern suggests that the recent assessment-focused literature has

XSL•FO
RenderX

developed rapidly around a relatively concentrated set of GenAI and examination-benchmarking references.

Figure 5. Keyword structure and thematic evolution of the literature. Panel A shows the co-occurrence network of normalized higher-frequency keywords. Panel B shows temporal profiles of leading keywords across the study period, highlighting movement from earlier machine-learning and simulation themes toward recent large language model and generative artificial intelligence assessment themes.



Evidence Map of Assessment Functions, Settings, and Learner Stages

Across the final manually reviewed evidence-map coding, GenAI was coded in 310 records (71.3%) and large language models in 301 records (69.2%). The most common assessment functions were summative assessment (n=168, 38.6%), feedback (n=93, 21.4%), question generation (n=59, 13.6%), and automated scoring (n=42, 9.7%).

To separate assessment content generation from learner-facing assessment activities, we added an assessment-function umbrella analysis. Learner performance evaluation accounted for 270 records (62.1%), feedback for 93 records (21.4%), assessment content generation for 65 records (14.9%), and other or unclear functions for 7 records (1.6%). Question generation and item generation were treated as assessment content generation and were not automatically interpreted as summative assessment unless explicitly linked to learner evaluation, examination use, or assessment decision-making.

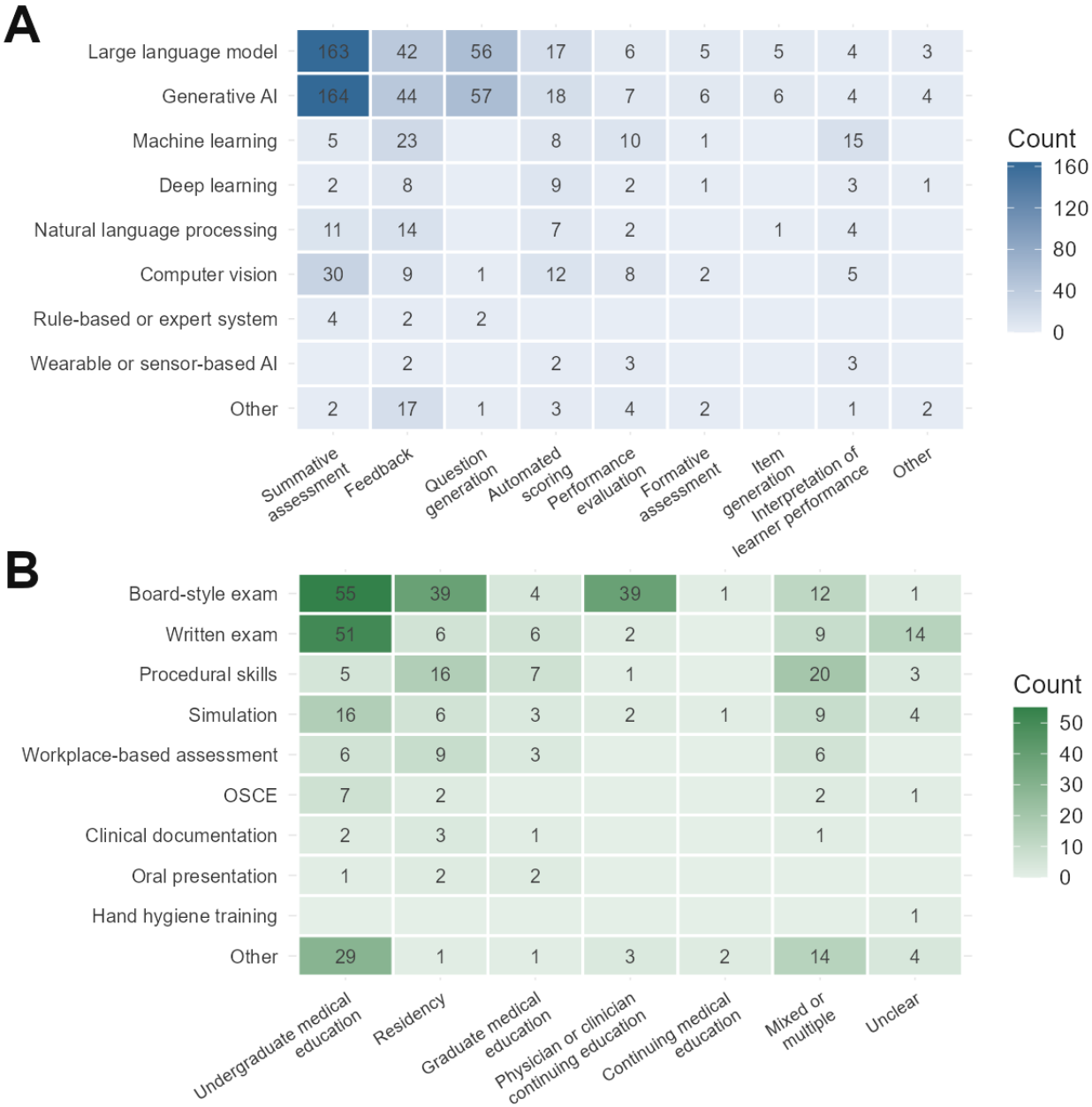
The most common assessment settings were board-style examinations (n=151, 34.7%), written examinations (n=88, 20.2%), procedural skills (n=52, 12%), and simulation (n=41, 9.4%). Undergraduate medical education was the most represented learner stage (n=172, 39.5%), followed by residency (n=84, 19.3%) and mixed or multiple learner stages (n=73, 16.8%).

XSL•FO
RenderX

The assessment function by AI type matrix showed concentration of GenAI and large language model records in summative assessment, question generation, and feedback (Figure 6). The largest intersections were GenAI with summative assessment (164 records), large language models with summative assessment (163 records), GenAI with question generation (57 records), large language models with question generation (56 records), and GenAI with feedback (44 records).

The assessment setting by learner stage matrix showed that board-style and written examinations were concentrated most clearly in undergraduate medical education, with 55 records in the board-style examination by the undergraduate medical education cell and 51 records in the written examination by the undergraduate medical education cell. Board-style examination records also appeared in residency and physician or clinician continuing education, with 39 records in each of those cells.

Figure 6. Evidence maps of assessment use. Panel A shows the assessment function by artificial intelligence type. Panel B shows the assessment setting by learner stage. Counts are based on nonmutually exclusive evidence-map coding categories and therefore represent coded study characteristics, not mutually exclusive study strata. Question generation and item generation were treated as assessment content generation and were not automatically interpreted as summative assessment unless explicitly linked to learner evaluation, examination use, or assessment decision-making.



Full-Text Reporting-Domain Sensitivity Analysis

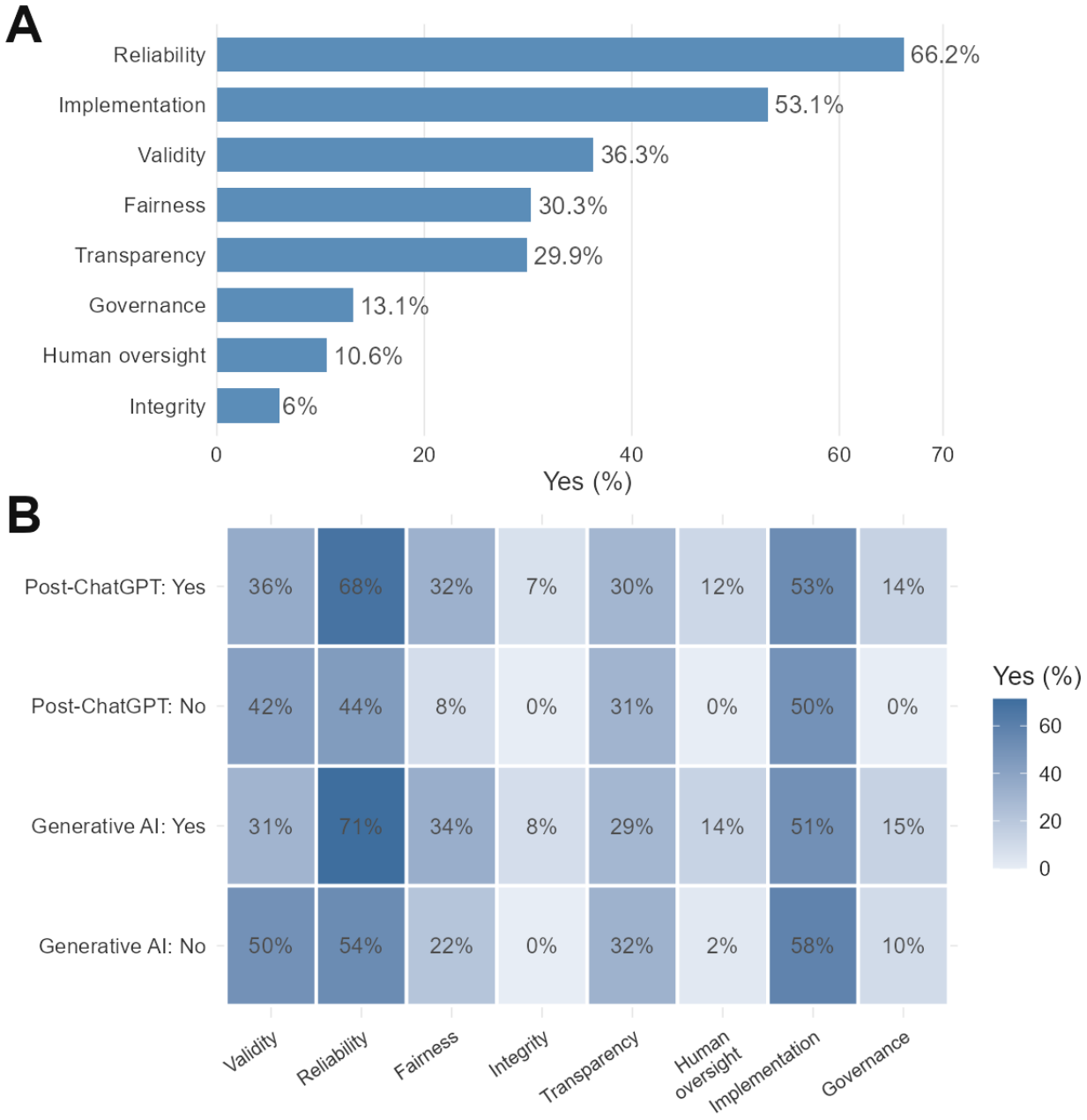
The reporting-domain variables visible in titles or abstracts were retained as the baseline. Because several reporting domains may be discussed mainly in full-text sections, we conducted a full-text sensitivity analysis across all 435 included records and used the full-text–confirmed estimates as the primary basis for interpreting reporting-domain findings.

In the full-text–confirmed analysis, reliability was identified in 288 (66.2%) records and implementation in 231 (53.1%) records (Figure 7). Validity, fairness, and transparency were identified

XSL•FO
RenderX

at intermediate levels: validity in 158 (36.3%) records, fairness in 132 (30.3%) records, and transparency in 130 (29.9%) records. Governance, human oversight, and integrity were identified less frequently: governance in 57 (13.1%) records, human oversight in 46 (10.6%) records, and integrity in 26 (6%) records.

Figure 7. Full-text–confirmed reporting of assessment quality, implementation, and governance domains. Panel A shows the percentage of records with full-text–confirmed reporting for each domain. Panel B shows descriptive subgroup percentages by post-ChatGPT period and generative artificial intelligence relevance. Reporting-domain estimates are based on the full-text sensitivity analysis across all 435 included records. Subgroup percentages are descriptive; Fisher exact tests with multiplicity correction are reported in Multimedia Appendix 3.



Compared with the baseline based on titles and abstracts, full-text–confirmed estimates increased for validity from 84 (19.3%) to 158 (36.3%), a net increase of 74 records; reliability from 107 (24.6%) to 288 (66.2%), a net increase of 181 records; fairness from 45 (10.3%) to 132 (30.3%), a net increase of 87 records; integrity from 11 (2.5%) to 26 (6%), a net increase of 15 records; transparency from 19 (4.4%) to 130 (29.9%), a net increase of 111 records; and implementation from 212 (48.7%) to 231 (53.1%), a net increase of 19 records. Human oversight remained unchanged at the aggregate level, with 46 records (10.6%) in both the baseline based on titles and abstracts and full-text–confirmed analysis. Governance changed from 62 records (14.3%) in the baseline based on titles and abstracts to 57 records (13.1%) in the full-text–confirmed analysis, a net decrease of 5 records after full-text adjudication.

These findings indicate that coding based on titles and abstracts underestimated several reporting domains, particularly reliability, transparency, fairness, and validity. The full-text sensitivity analysis refined the interpretation of reporting domains while leaving the main evidence-map categories based

XSL•FO
RenderX

on the final manually reviewed coding based on titles and abstracts, with targeted ambiguity resolution.

Coding Reliability

Stage 1 agreement for evidence-map categories coded from titles and abstracts was calculated across the full cohort of 435 records. Agreement was 393 (90.3%) for publication type, with Cohen κ of 0.785; 385 (88.5%) for study design, with Cohen κ of 0.843; 391 (89.9%) for AI type exact set, with Cohen κ of 0.859; 391 (89.9%) for assessment function exact set, with Cohen κ of 0.868; 392 (90.1%) for assessment-function umbrella, with Cohen κ of 0.826; 391 (89.9%) for assessment setting, with Cohen κ of 0.870; 400 (92%) for learner stage, with Cohen κ of 0.895; 407 (93.6%) for GenAI relevance, with Cohen κ of 0.851; and 423 (97.2%) for post-ChatGPT period, with Cohen κ of 0.834.

Stage 2 agreement for full-text reporting-domain coding was also calculated across the full cohort of 435 records. Agreement was 411 (94.5%) for validity, with Cohen κ of 0.877; 400 (92%) for reliability, with Cohen κ of 0.830; 414 (95.2%) for fairness, with Cohen κ of 0.880; 427 (98.2%) for integrity, with Cohen κ of 0.809; 413 (94.9%) for transparency, with Cohen κ of 0.873; 422 (97%) for human oversight, with Cohen κ of 0.819; 404 (92.9%) for implementation, with Cohen κ of 0.858; and 420 (96.6%) for governance, with Cohen κ of 0.830. For low-prevalence domains such as integrity and human oversight, percent agreement was interpreted alongside Cohen κ because prevalence can influence κ estimates.

Subgroup and Partial-Year Sensitivity Analyses

Exploratory subgroup analyses for all 8 full-text–confirmed reporting domains are reported in Multimedia Appendix 3. Fisher exact tests with Benjamini-Hochberg correction were used to support these comparisons. In the main text, subgroup findings are summarized selectively and should be interpreted cautiously as descriptive and noncausal.

In the GenAI subgroup comparison, GenAI-related records had higher full-text–confirmed reporting of reliability than records not coded as involving GenAI (221/310, 71.3% vs 67/125, 53.6%), fairness (105/310, 33.9% vs 27/125, 21.6%), integrity (26/310, 8.4% vs 0/125, 0%), and human oversight (43/310, 13.9% vs 3/125, 2.4%). Validity was identified less often in GenAI-related records than in records not coded as involving GenAI (95/310, 30.6% vs 63/125, 50.4%), and transparency was similar between groups (90/310, 29% vs 40/125, 32%). These subgroup differences may reflect differences in publication period, assessment context, and application type.

In the post-ChatGPT comparison, the pre-ChatGPT subset was small (36 records). Reliability, fairness, human oversight, and governance were identified more often in the post-ChatGPT subset, whereas transparency was similar across periods. Detailed counts, unadjusted *P* values, and multiplicity-adjusted *P* values are provided in Multimedia Appendix 3.

Because 2026 was a partial year through April 8, 2026, we conducted an excluding-2026 sensitivity analysis. Excluding the 65 records from 2026, there were 370 records. Full-text–confirmed reporting-domain estimates changed by 0 to 3.9 percentage points. The largest absolute change was for transparency, which changed from 130 of 435 (29.9%) in the full cohort to 96 of 370 (25.9%) after excluding 2026. The overall interpretation was unchanged: reliability and implementation were the most frequently identified domains, validity, fairness, and transparency were identified at intermediate levels, and integrity, human oversight, and governance remained less frequently identified.

Discussion

Principal Findings

This bibliometric mapping study and thematic evidence map showed rapid growth in the literature on AI for assessment and feedback in medical education, particularly after the public release of ChatGPT. The mapped literature was concentrated in GenAI and large language model applications, with prominent representation of summative assessment, board-style examinations, written examinations, and undergraduate medical education.

The full-text sensitivity analysis showed uneven reporting of assessment-quality, implementation, and governance domains: reliability and implementation were frequently identified; validity, fairness, and transparency were identified at intermediate levels; and integrity, human oversight, and governance remained less frequently identified. The full-text sensitivity analysis also showed that coding based on titles and abstracts underestimated several reporting domains, supporting the decision to interpret reporting-domain findings primarily using full-text–confirmed estimates while retaining estimates based on titles and abstracts as the baseline comparison.

Assessment Focus, Learner Stages, and Educational Implications

The prominence of board-style examinations and written examinations is understandable in the context of GenAI. Standardized questions provide accessible test beds, defined answer structures, and clear scoring benchmarks. They also allow rapid comparison across models and versions. For this reason, examination-benchmarking and evaluation studies have been useful for demonstrating what large language models can and cannot do on constrained knowledge-assessment tasks [8-10,27,28]. These studies are valuable for defining technical questions, including model performance, response consistency, prompt sensitivity, and knowledge retrieval.

However, benchmark performance alone is not sufficient to guide AI use in medical education assessment. Authentic assessment often requires observing learners in context, interpreting performance over time, integrating multiple sources of evidence, and making judgments that are educationally and professionally consequential [11-15]. Workplace-based assessment, procedural skills assessment, simulation, entrustment-related judgment, clinical reasoning assessment, and longitudinal feedback depend on context, supervision, relationships, and human interpretation in ways that differ from multiple-choice examination performance [29-33]. In the present map, board-style and written examinations were more common than procedural skills and simulation. This distribution suggests

XSL•FO
RenderX

that the current assessment-focused AI literature is more developed for structured examination contexts than for assessment environments that require contextual judgment and integration of human observation.

The learner-stage distribution further reinforces this interpretation. Undergraduate medical education was the most represented learner stage, followed by residency and mixed or multiple learner stages. Undergraduate contexts are important sites for AI literacy, early clinical reasoning, and examination preparation [6,34-37]. At the same time, AI-supported assessment in residency, graduate medical education, continuing medical education, and clinician continuing professional development may raise different questions about workplace performance, entrustment, certification, supervision, and accountability. Future work should therefore examine AI-supported assessment across the continuum of medical training without assuming that evidence from undergraduate examination contexts generalizes to all learner stages.

The distinction between assessment content generation and learner-facing assessment is also central. AI-generated questions, cases, rubrics, or feedback templates may support educators and assessment systems, but producing assessment content is not the same as evaluating learner performance. Learner-facing assessment involves judging what a learner knows or can do, determining what feedback is appropriate, and sometimes informing progression or remediation. Therefore, the assessment-function disaggregation helps clarify the structure of the field. Assessment content generation may require evidence of item quality, alignment, difficulty, bias, and review processes [11-16,38]. Learner-facing scoring or feedback requires evidence about validity, reliability, fairness, transparency, human oversight, and educational consequences [16-20].

Feedback is a particularly important area for future development. AI-supported feedback could provide timely explanations, identify patterns in learner performance, draft individualized comments, or help educators manage feedback workload. Yet feedback is not only information delivery; it depends on learner interpretation, credibility, relationship, timing, and opportunities for improvement [39-41]. For learner-facing AI assessment or feedback studies that function as educational interventions, educational-intervention reporting guidance such as GREET (Guideline for Reporting Evidence-Based Practice Educational Interventions and Teaching) may help clarify the educational context, intervention components, implementation conditions, learner outcomes, and feedback processes [42]. The most educationally defensible uses of AI feedback may therefore be hybrid workflows in which AI supports generation, organization, or pattern recognition while educators retain responsibility for contextualization, escalation, and high-stakes judgment. This framing positions AI as a tool that may augment feedback processes when appropriately validated and governed [43].

Full-Text Reporting Domains, Governance, and Responsible Assessment

The full-text sensitivity analysis refined the reporting-domain findings. In the baseline based on titles and abstracts, several domains appeared uncommon. After full-text review across all 435 records, reliability and implementation were frequently identified; validity, fairness, and transparency were identified at intermediate levels; and integrity, human oversight, and governance remained less frequent. This pattern supports a more nuanced interpretation: reporting was uneven, not uniformly sparse.

This distinction matters because assessment places particular demands on responsible AI use. Reliability and implementation are necessary, but they are not sufficient. Validity is needed to support the intended interpretation and use of assessment results [11-16,38]. Fairness, transparency, and human oversight are also needed so educators and learners can understand, question, and appropriately rely on AI-supported outputs [17-20]. The intermediate full-text reporting of validity, fairness, and transparency suggests that these issues are present in the literature, but they are not yet consistently integrated across assessment-focused studies.

The lower reporting of integrity, human oversight, and governance is particularly relevant for medical education assessment. Integrity concerns whether AI changes what is being assessed, creates opportunities for misuse, or undermines confidence in the interpretation of learner performance. Human oversight concerns who reviews AI outputs, when educators intervene, and how disagreements or uncertain outputs are handled. Governance concerns institutional policy, accountability, privacy, data governance, acceptable use, escalation, and monitoring [17-20,43]. These domains are especially important when AI-supported systems contribute to feedback, scoring, performance interpretation, or progression-related judgments.

The subgroup analyses were exploratory and should not be interpreted causally. GenAI-related records differed from records not coded as involving GenAI in the publication period, assessment context, and application type, and the pre-ChatGPT subset was small. Still, the descriptive patterns reinforce the need for more explicit reporting as the field matures. In assessment contexts, responsible AI reporting is not an administrative add-on. It shapes whether a tool can be trusted, how it should be supervised, what decisions it can support, and how learners and educators can challenge or interpret its outputs.

Comparison With Prior Work

This study builds on broad reviews and bibliometric analyses of AI in medical education, which have described rapid growth in AI-related educational scholarship and highlighted applications in curriculum design, simulation, clinical reasoning, feedback, learner support, and institutional readiness [3,6,21,22]. Those studies are essential for understanding the wider educational landscape. The present study contributes a narrower but complementary view by isolating assessment and feedback.

This assessment-specific focus changes what becomes visible. In broad AI-in-medical-education syntheses, assessment may appear as one application among many. In the present map, assessment-related work shows a distinct profile: strong post-ChatGPT growth, concentration in GenAI and large language models, prominence of examination-oriented settings, and comparatively less representation of authentic assessment contexts such as workplace-based assessment, procedural

XSL•FO
RenderX

assessment, and longitudinal performance interpretation. The evidence-map approach also highlights the importance of separating assessment content generation from learner-facing evaluation and feedback.

The full-text reporting-domain analysis also extends prior work. Responsible AI and ethical guidance in health professions education have emphasized fairness, transparency, accountability, human involvement, privacy, and governance [17-20,43]. Our results show how often related domains are explicitly identifiable in the assessment-focused literature. The finding that reliability and implementation were frequently identified, validity, fairness, and transparency were intermediate, and integrity, human oversight, and governance were less frequent, provides a more granular view than broad claims about underreporting across all domains.

Finally, the bibliometric component complements prior field-level studies by showing how the assessment-focused subfield is structured around recent GenAI and examination-benchmarking work. This is useful for medical educators and researchers because it identifies both the current center of activity and areas where the evidence base remains less developed. Future assessment-focused studies can build on benchmark work while expanding into authentic performance contexts, learner development, faculty judgment, and governance of AI-supported assessment systems.

Limitations

This study has several limitations. First, the search was limited to Web of Science Core Collection, Scopus, and PubMed, and to English-language records. Relevant publications indexed only in other databases or published in other languages may have been missed. Country and institution findings should therefore be interpreted as patterns within the indexed English-language literature captured by the search, not as complete global research output.

Second, this was a bibliometric mapping study with structured evidence-map coding. The study maps the literature and its reported characteristics but does not evaluate intervention effectiveness, compare educational outcomes, or assess risk of bias for individual empirical studies.

Third, document selection and the main evidence-map categories were based primarily on titles, abstracts, and bibliographic metadata, with targeted ambiguity resolution when needed. This approach was aligned with the bibliographic-record-based mapping purpose of the study, but some records may have been incompletely characterized at the level of titles and abstracts. To address this limitation for domains especially likely to appear outside abstracts, we added a full-text sensitivity analysis for validity, reliability, fairness, integrity, transparency, human oversight, implementation, and governance across all 435 included records.

Fourth, evidence-map categories were nonmutually exclusive. This was appropriate for representing multidimensional records that involved more than 1 AI type, assessment function, setting, or learner stage, but counts should not be interpreted as mutually exclusive study strata. Fifth, subgroup analyses were exploratory. Although Fisher exact tests with Benjamini-Hochberg correction were conducted, subgroup patterns may reflect differences in publication period, study type, assessment context, or AI application and should not be interpreted as independent associations.

Sixth, citation-based indicators are influenced by publication timing, journal indexing, database coverage, and the short time window since the release of ChatGPT. Recent records may not yet have accumulated citations, whereas early benchmark papers may appear especially prominent because they became early anchors for subsequent work. Finally, 2026 was a partial year through April 8, 2026. We marked 2026 as partial in the annual trend figure and conducted an excluding-2026 sensitivity analysis; the main interpretation of reporting-domain patterns was unchanged.

Conclusions

The literature on AI for assessment and feedback in medical education has expanded rapidly, particularly in the post-ChatGPT period. Within the mapped indexed English-language literature, research was concentrated in GenAI, large language models, summative assessment, board-style and written examinations, and undergraduate medical education. The full-text sensitivity analysis showed that reporting of assessment-quality, implementation, and governance domains was uneven: reliability and implementation were frequently identified; validity, fairness, and transparency were identified at intermediate levels; and integrity, human oversight, and governance remained less frequent.

Future research should complement examination benchmarking with more authentic assessment contexts, including workplace-based assessment, simulation, procedural skills assessment, longitudinal feedback, and performance interpretation across the continuum of medical training. Studies should distinguish assessment content generation from learner-facing evaluation and feedback, and should report more consistently on validity, fairness, transparency, integrity, human oversight, implementation, and governance. For medical education assessment, the central question is not only whether AI can perform isolated assessment tasks, but also how AI-supported workflows can improve learning and assessment while preserving accountability, fairness, and human educational judgment.

Acknowledgments

The authors thank Mr Doudou Kim for his assistance in the preparation of this manuscript.

ChatGPT was used only to assist with English-language polishing during manuscript revision and was not used for literature screening, eligibility assessment, evidence-map coding, data analysis, or interpretation. The authors reviewed and approved all final manuscript content, analyses, interpretations, and conclusions.

XSL•FO
RenderX

Data Availability

Summary data supporting the findings are provided in the article and the Multimedia Appendices. Because the source records were obtained from third-party bibliographic databases subject to their respective licensing terms, raw database exports are not publicly shared. Additional derived summary tables are available from the corresponding author on reasonable request, subject to database licensing restrictions. The R scripts used to generate the derived summary tables, sensitivity analyses, agreement analyses, subgroup analyses, partial-year sensitivity analyses, and figures will be made publicly available in a GitHub repository after acceptance.

Funding

This study was supported by the Natural Science Foundation of Hebei Province (Grant number H2025206300), the 2025 Hebei Province Government-funded Training Program for Clinical Medical Professionals (Grant number ZF2025138), the Clinical Innovation Research Team of Hebei Medical University (Grant number 2022LCTD-B35), the Third Hospital of Hebei Medical University University-Hospital Collaboration Fund Program (Grant number ydsyqjjh202513), and the Hebei Province Introduction of Foreign Intelligence Projects (2026). The funders had no role in the design and conduct of the study; collection, management, analysis, and interpretation of the data; preparation, review, or approval of the manuscript; or the decision to submit the manuscript for publication.

Authors' Contributions

Conceptualization: ZZ (lead), ZL (equal), LG (equal), ZH (lead)
Data curation: ZZ (lead), ZL (equal), LG (equal)
Formal analysis: ZZ (lead), ZL (equal), LG (equal)
Investigation: ZZ (lead), ZL (equal), LG (equal), TP (supporting), YZ (supporting), CM (supporting), YG (supporting), YW (supporting)
Methodology: ZZ (lead), ZL (equal), LG (equal), RZ (supporting), ZH (lead)
Supervision and manuscript revision: RZ (equal), ZH (equal)
Visualization: ZZ (lead), ZL (equal), LG (equal)
Writing – original draft: ZZ (lead), ZL (equal), LG (equal)
Writing – review and editing: RZ (lead), ZH (lead), TP (supporting), YZ (supporting), CM (supporting), YG (supporting), YW (supporting), XH (supporting), XW (supporting)

Conflicts of Interest

None declared.

Multimedia Appendix 1
Completed BIBLIO reporting checklist. Completed checklist showing where BIBLIO reporting items are addressed in the revised manuscript and supplementary materials.
[DOCX File , 27 KB - mededu_v12i1e98949_app1.docx]

Multimedia Appendix 2
Detailed search strategies for Web of Science Core Collection, Scopus, and PubMed. Database-specific search strings; search limits; final search date of April 8, 2026; search period from January 1, 2015, to April 8, 2026; and final hit counts.
[DOCX File , 17 KB - mededu_v12i1e98949_app2.docx]

Multimedia Appendix 3
Supplementary methods and results. Supplementary methodological reporting, coding definitions, bibliometric software and parameter settings, full-text sensitivity coding rules, study selection and annual trend details, descriptive evidence-map tables, assessment-function disaggregation, full-text reporting-domain summaries, reporting shifts from estimates visible in titles or abstracts to full-text-confirmed estimates, two-stage agreement analyses, subgroup Fisher exact tests with Benjamini-Hochberg correction, excluding-2026 partial-year sensitivity analysis, and bibliometric summaries.
[XLSX File (Microsoft Excel File), 38 KB - mededu_v12i1e98949_app3.xlsx]

Multimedia Appendix 4
Supplementary co-citation heatmap of the recent post-ChatGPT assessment literature. Co-citation heatmap showing the recent intellectual structure of the post-ChatGPT literature on artificial intelligence for assessment and feedback in medical education.
[PNG File , 110 KB - mededu_v12i1e98949_app4.png]

References

XSL•FO
RenderX

1. Chan KS, Zary N. Applications and challenges of implementing artificial intelligence in medical education: integrative review. JMIR Med Educ 2019;5(1):e13930 [FREE Full text] [doi: 10.2196/13930] [Medline: 31199295]
2. Masters K. Artificial intelligence in medical education. Med Teach 2019;41(9):976-980. [doi: 10.1080/0142159X.2019.1595557] [Medline: 31007106]
3. Gordon M, Daniel M, Ajiboye A, Uraiby H, Xu NY, Bartlett R, et al. A scoping review of artificial intelligence in medical education: BEME guide no. 84. Med Teach 2024;46(4):446-470 [FREE Full text] [doi: 10.1080/0142159X.2024.2314198] [Medline: 38423127]
4. Han ER, Yeo S, Kim MJ, Lee YH, Park KH, Roh H. Medical education trends for future physicians in the era of advanced technology and artificial intelligence: an integrative review. BMC Med Educ 2019;19(1):460 [FREE Full text] [doi: 10.1186/s12909-019-1891-5] [Medline: 31829208]
5. Kolachalama VB, Garg PS. Machine learning and medical education. NPJ Digit Med 2018;1:54 [FREE Full text] [doi: 10.1038/s41746-018-0061-1] [Medline: 31304333]
6. Michalczak M, Zgoda W, Michalczak J, Żądło A, Nasser A, Tokarek T. Artificial intelligence in medical education: a narrative review. AI 2025;6(12):322. [doi: 10.3390/ai6120322]
7. Sallam M. ChatGPT utility in healthcare education, research, and practice: systematic review on the promising perspectives and valid concerns. Healthcare (Basel) 2023;11(6):887 [FREE Full text] [doi: 10.3390/healthcare11060887] [Medline: 36981544]
8. Kung TH, Cheatham M, Medenilla A, Sillos C, De Leon L, Elepaño C, et al. Performance of ChatGPT on USMLE: potential for AI-assisted medical education using large language models. PLOS Digit Health 2023;2(2):e0000198 [FREE Full text] [doi: 10.1371/journal.pdig.0000198] [Medline: 36812645]
9. Gilson A, Safranek CW, Huang T, Socrates V, Chi L, Taylor RA, et al. How does ChatGPT perform on the United States medical licensing examination (USMLE)? The implications of large language models for medical education and knowledge assessment. JMIR Med Educ 2023;9:e45312 [FREE Full text] [doi: 10.2196/45312] [Medline: 36753318]
10. Huh S. Are ChatGPT's knowledge and interpretation ability comparable to those of medical students in Korea for taking a parasitology examination?: A descriptive study. J Educ Eval Health Prof 2023;20:1 [FREE Full text] [doi: 10.3352/jeehp.2023.20.1] [Medline: 36627845]
11. Epstein RM. Assessment in medical education. N Engl J Med 2007;356(4):387-396. [doi: 10.1056/NEJMra054784] [Medline: 17251535]
12. van der Vleuten CPM, Schuwirth LWT. Assessing professional competence: from methods to programmes. Med Educ 2005;39(3):309-317. [doi: 10.1111/j.1365-2929.2005.02094.x] [Medline: 15733167]
13. van der Vleuten CPM, Schuwirth LWT, Driessen EW, Dijkstra J, Tigelaar D, Baartman LKJ, et al. A model for programmatic assessment fit for purpose. Med Teach 2012;34(3):205-214. [doi: 10.3109/0142159X.2012.652239] [Medline: 22364452]
14. Schuwirth LWT, Van der Vleuten CPM. Programmatic assessment: from assessment of learning to assessment for learning. Med Teach 2011;33(6):478-485. [doi: 10.3109/0142159X.2011.565828] [Medline: 21609177]
15. Holmboe ES, Sherbino J, Long DM, Swing SR, Frank JR. The role of assessment in competency-based medical education. Med Teach 2010;32(8):676-682. [doi: 10.3109/0142159X.2010.500704] [Medline: 20662580]
16. Cook DA, Hatala R. Validation of educational assessments: a primer for simulation and beyond. Adv Simul (Lond) 2016;1:31 [FREE Full text] [doi: 10.1186/s41077-016-0033-y] [Medline: 29450000]
17. Masters K. Ethical use of artificial intelligence in health professions education: AMEE guide No. 158. Med Teach 2023;45(6):574-584 [FREE Full text] [doi: 10.1080/0142159X.2023.2186203] [Medline: 36912253]
18. Ethics and governance of artificial intelligence for health. World Health Organization. Geneva: World Health Organization; 2021 Jun 01. URL: https://www.who.int/publications/i/item/9789240029200 [accessed 2026-06-01]
19. Liu X, Cruz Rivera S, Moher D, Calvert MJ, Denniston AK, SPIRIT-AI and CONSORT-AI Working Group. Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI extension. Nat Med 2020;26(9):1364-1374 [FREE Full text] [doi: 10.1038/s41591-020-1034-x] [Medline: 32908283]
20. Vasey B, Nagendran M, Campbell B, Clifton DA, Collins GS, Denaxas S, et al. Reporting guideline for the early-stage clinical evaluation of decision support systems driven by artificial intelligence: DECIDE-AI. Nat Med 2022;28(5):924-933. [doi: 10.1038/s41591-022-01772-9] [Medline: 35585198]
21. Wang S, Yang L, Li M, Zhang X, Tai X. Medical education and artificial intelligence: web of science-based bibliometric analysis (2013-2022). JMIR Med Educ 2024;10:e51411 [FREE Full text] [doi: 10.2196/51411] [Medline: 39388721]
22. Wang Y, Chang C, Shi W, Liu H, Huang X, Jiao Y. How AI is transforming medical education: bibliometric analysis. JMIR Med Educ 2025;11:e75911 [FREE Full text] [doi: 10.2196/75911] [Medline: 41252190]
23. Maggio LA, Ninkov A, Frank JR, Costello JA, Artino AR. Delineating the field of medical education: bibliometric research approach(es). Med Educ 2022;56(4):387-394 [FREE Full text] [doi: 10.1111/medu.14677] [Medline: 34652832]
24. Miake-Lye IM, Hempel S, Shanman R, Shekelle PG. What is an evidence map? A systematic review of published evidence maps and their definitions, methods, and products. Syst Rev 2016;5:28 [FREE Full text] [doi: 10.1186/s13643-016-0204-x] [Medline: 26864942]
25. Grant MJ, Booth A. A typology of reviews: an analysis of 14 review types and associated methodologies. Health Info Libr J 2009;26(2):91-108 [FREE Full text] [doi: 10.1111/j.1471-1842.2009.00848.x] [Medline: 19490148]

XSL•FO
RenderX

26. Montazeri A, Mohammadi S, M Hesari P, Ghaemi M, Riazi H, Sheikhi-Mobarakeh Z. Preliminary guideline for reporting bibliometric reviews of the biomedical literature (BIBLIO): a minimum requirements. Syst Rev 2023;12(1):239 [FREE Full text] [doi: 10.1186/s13643-023-02410-2] [Medline: 38102710]
27. Wang C, Li S, Lin N, Zhang X, Han Y, Wang X, et al. Application of large language models in medical training evaluation-using ChatGPT as a standardized patient: multimetric assessment. J Med Internet Res 2025;27:e59435 [FREE Full text] [doi: 10.2196/59435] [Medline: 39742453]
28. Sakelaris PG, Novotny KV, Borvick MS, Lagasca GG, Simanton EG. Evaluating the use of artificial intelligence as a study tool for preclinical medical school exams. J Med Educ Curric Dev 2025;12:23821205251320150 [FREE Full text] [doi: 10.1177/23821205251320150] [Medline: 40008118]
29. Norcini J, Burch V. Workplace-based assessment as an educational tool: AMEE Guide No. 31. Med Teach 2007;29(9):855-871. [doi: 10.1080/01421590701775453] [Medline: 18158655]
30. Ten Cate O. Nuts and bolts of entrustable professional activities. J Grad Med Educ 2013;5(1):157-158 [FREE Full text] [doi: 10.4300/JGME-D-12-00380.1] [Medline: 24404246]
31. ten Cate O, Scheele F. Competency-based postgraduate training: can we bridge the gap between theory and clinical practice? Acad Med 2007;82(6):542-547. [doi: 10.1097/ACM.0b013e31805559c7] [Medline: 17525536]
32. Kogan JR, Holmboe ES, Hauer KE. Tools for direct observation and assessment of clinical skills of medical trainees: a systematic review. JAMA 2009;302(12):1316-1326. [doi: 10.1001/jama.2009.1365] [Medline: 19773567]
33. Kirubarajan A, Young D, Khan S, Crasto N, Sobel M, Sussman D. Artificial intelligence and surgical education: a systematic scoping review of interventions. J Surg Educ 2022;79(2):500-515. [doi: 10.1016/j.jsurg.2021.09.012] [Medline: 34756807]
34. Grunhut J, Marques O, Wyatt ATM. Needs, challenges, and applications of artificial intelligence in medical education curriculum. JMIR Med Educ 2022;8(2):e35587 [FREE Full text] [doi: 10.2196/35587] [Medline: 35671077]
35. Paranjape K, Schinkel M, Nannan Panday R, Car J, Nanayakkara P. Introducing artificial intelligence training in medical education. JMIR Med Educ 2019;5(2):e16048 [FREE Full text] [doi: 10.2196/16048] [Medline: 31793895]
36. McCoy LG, Nagaraj S, Morgado F, Harish V, Das S, Celi LA. What do medical students actually need to know about artificial intelligence? NPJ Digit Med 2020;3:86. [doi: 10.1038/s41746-020-0294-7] [Medline: 32577533]
37. Park SH, Do K, Kim S, Park JH, Lim Y. What should medical students know about artificial intelligence in medicine? J Educ Eval Health Prof 2019;16:18 [FREE Full text] [doi: 10.3352/jeehp.2019.16.18] [Medline: 31319450]
38. Cook DA, Beckman TJ. Current concepts in validity and reliability for psychometric instruments: theory and application. Am J Med 2006;119(2):166.e7-166.e16. [doi: 10.1016/j.amjmed.2005.10.036] [Medline: 16443422]
39. van de Ridder JMM, Stokking KM, McGaghie WC, ten Cate OTJ. What is feedback in clinical education? Med Educ 2008;42(2):189-197. [doi: 10.1111/j.1365-2923.2007.02973.x] [Medline: 18230092]
40. Ramani S, Krackov SK. Twelve tips for giving feedback effectively in the clinical environment. Med Teach 2012;34(10):787-791. [doi: 10.3109/0142159X.2012.684916] [Medline: 22730899]
41. Watling CJ, Ginsburg S. Assessment, feedback and the alchemy of learning. Med Educ 2019;53(1):76-85. [doi: 10.1111/medu.13645] [Medline: 30073692]
42. Phillips AC, Lewis LK, McEvoy MP, Galipeau J, Glasziou P, Moher D, et al. Development and validation of the guideline for reporting evidence-based practice educational interventions and teaching (GREET). BMC Med Educ 2016;16(1):237 [FREE Full text] [doi: 10.1186/s12909-016-0759-1] [Medline: 27599967]
43. Cohen IG, Babic B, Gerke S, Xia Q, Evgeniou T, Wertenbroch K. How AI can learn from the law: putting humans in the loop only on appeal. NPJ Digit Med 2023;6(1):160 [FREE Full text] [doi: 10.1038/s41746-023-00906-8] [Medline: 37626155]

Abbreviations

AI: artificial intelligence
BIBLIO: bibliometric reviews of biomedical literature
GenAI: generative artificial intelligence
GREET: Guideline for Reporting Evidence-Based Practice Educational Interventions and Teaching
USMLE: United States Medical Licensing Examination

XSL•FO
RenderX

Edited by N Khamis; submitted 21.Apr.2026; peer-reviewed by L Mosca, T Tokarek, C Rose, SS Fatima; comments to author 20.May.2026; accepted 15.Jun.2026; published 02.Jul.2026.

Please cite as:
Zhao Z, Liu Z, Guo L, Pan T, Zhang Y, Miao C, Ge Y, Wang Y, Hu X, Wang X, Zhang R, Hou Z
Artificial Intelligence for Assessment and Feedback in Medical Education: Bibliometric Mapping Study and Thematic Evidence Map
JMIR Med Educ 2026;12:e98949
URL: https://mededu.jmir.org/2026/1/e98949
doi:10.2196/98949
PMID:

©Zihang Zhao, Zihan Liu, Liang Guo, Teng Pan, Yousheng Zhang, Chenxiang Miao, Yiting Ge, Yipeng Wang, Xin Hu, Xin Wang, Ruipeng Zhang, Zhiyong Hou. Originally published in JMIR Medical Education (https://mededu.jmir.org), 02.Jul.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Education, is properly cited. The complete bibliographic information, a link to the original publication on https://mededu.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Evaluating the Pediatric Behavior Guidance of Students Based on Actual Clinical Transcripts Scored by Faculty and Large Language Models: Pilot Comparative Study

Ishreen Kaur Dhillon, BDS, MDS; Gabriel Keng Yan Lee, BDS, MPH; Shijia Hu, BDS, PhD

Faculty of Dentistry, National University of Singapore, 9 Lower Kent Ridge Road, Singapore, Singapore

Corresponding Author:
Shijia Hu, BDS, PhD
Faculty of Dentistry, National University of Singapore, 9 Lower Kent Ridge Road, Singapore, Singapore

Abstract

Background: Personalized feedback improves the clinical pediatric behavior guidance performance of students but is prohibitively time-consuming to provide. Large language models (LLMs) can automate the process of evaluating clinical sessions but are limited to text-only input and consistency issues.

Objective: This study compared the use of text-only transcripts against the use of video recordings for evaluating the clinical behavior guidance performance of dental students. Additionally, the consistency and accuracy of LLMs in evaluating the transcripts were compared against a human assessor.

Methods: This study was conducted by using 40 video-recorded clinical encounters involving final-year dental students who were managing patients aged between 4 and 12 years at the Faculty of Dentistry, National University of Singapore. The videos were scored by using a previously validated pediatric behavior guidance scale. Clinical encounters were transcribed verbatim and scored by a study member using a modified version of the scale (nonverbal components removed). The time taken to rate the transcripts was recorded. Video scores were compared with transcript scores. Both the free-to-use version and the paid version of the ChatGPT LLM were also used to score the transcripts; consistency was evaluated and compared against the human assessor.

Results: The average time taken to rate the transcripts (mean 12, range 3-25 min) was significantly (P<.001) lower than the average video length (mean 73, range 37-120 min). Comparing transcript scores with video scores resulted in a consistency intraclass correlation coefficient of 0.830 (95% CI 0.679 - 0.910; P<.001), demonstrating good reliability. Comparing transcript scores with the free-to-use LLM's and paid LLM's scores yielded an absolute agreement intraclass correlation coefficient of 0.729 (95% CI 0.475 - 0.859; P<.001) and 0.670 (95% CI 0.377 - 0.825; P<.001), respectively, demonstrating moderate agreement. The LLMs were inconsistent, producing variable scores with the same prompt. The free-to-use and paid versions produced the same score for all 3 runs in only 7 (18%) and 4 (10%) of the 40 clinical encounters, respectively.

Conclusions: Using transcripts to evaluate students' clinical behavior guidance was time-saving for faculty, demonstrated good agreement with video-based evaluation, and could improve clinical teaching. Although LLMs can automate the task, improvements are needed to improve their consistency and accuracy.

(JMIR Med Educ 2026;12:e83376) doi:10.2196/83376

KEYWORDS

clinical mentoring; dental education; large language models; pediatric dentistry; dental students

Introduction

Managing pediatric dental patients is stressful for dental students and inexperienced practitioners, particularly when communicating with fearful and uncooperative children. Dental students experience 3 times the stress levels when compared to seasoned specialists [1]. Although lectures and seminars deliver theoretical knowledge, they are limited in terms of actual application [2]. Experiential learning during clinical sessions is essential for applying knowledge received from didactic teaching. Central to this process is guidance from faculty (ie, clinical supervisor), whose feedback helps to transform knowledge into clinical proficiency [3]. Real-time faculty feedback offers the greatest potential due to the immediacy and individualized feedback provided [4]. However, it is impossible to provide continuous instructional and coaching feedback for every student, as faculty often supervise multiple students during clinical sessions.

An alternative to real-time faculty feedback is the utilization of video-recorded clinical sessions, which have been shown to significantly improve dental students' pediatric guidance scores in clinical situations [5]. Although video feedback enables faculty to evaluate the entire session for each student and provide feedback targeted at processes, it is prohibitively

XSL•FO
RenderX

time-consuming. Each session can last up to 90 minutes, requiring full viewing before meaningful feedback can be developed and delivered.

The advancement of artificial intelligence (AI)–powered large language models (LLMs) allows for the evaluation of clinical interactions and generation of feedback on the improvement of pediatric behavior guidance, which can help to improve the process of teaching such guidance. Recently, an LLM was used to simulate patient-physician interactions for medical students; the clinical decision-making performance of the students who received feedback from the LLM significantly improved in subsequent patient-physician interactions when compared to that of students who did not receive any feedback [6]. In another study, an LLM (ie, a chatbot) provided feedback on medical students' history-taking performance that was comparable to feedback from human assessors [7]. However, these studies were based on simulated clinical scenarios, with none examining the use of LLMs in actual clinical situations. The use of LLMs to provide feedback on behavior guidance in pediatric dentistry can address the issue of the extensive time requirement for providing personalized feedback based on video-recorded sessions—an area that has yet to be explored.

LLMs have inherent shortcomings that may limit their ability to evaluate clinical interactions. Currently, input is limited to only text, which does not consider other important facets of communication, including tone, physical gestures, and nonverbal cues. As such, an LLM's ability to evaluate clinical interactions may not be wholly representative [8]. Furthermore, a recent study found that using a commercially available LLM to evaluate patient risk based on computer-simulated clinical data resulted in moderately correlated (r=0.605) risk scores and only 56% agreement on a diagnosis category, calling into question the consistency of LLMs in scoring clinical interactions [9].

The primary aim of this study was to determine if using text-only input was sufficient for evaluating pediatric behavior guidance performance in a clinical setting when compared to using video-recorded clinical interactions. The secondary aim was to examine the consistency and accuracy of commercially available LLMs in evaluating pediatric behavior guidance when compared against a human assessor.

Methods

Ethical Considerations

This study was approved by the National University of Singapore Institutional Review Board (NUS-IRB-2024 - 934). This study was reported according to the Chatbot Assessment Reporting Tool (CHART) guidelines [10] (Checklist 1). Informed consent was obtained from participants. The participants were not provided with any monetary compensation. All data were deidentified before data analysis was done.

Study Design

This study was conducted by using a set of 50 video-recorded clinical encounters involving final-year dental students who were managing pediatric patients aged between 4 and 12 years at the Faculty of Dentistry, National University of Singapore. The video-recorded clinical encounters were previously evaluated by a pediatric dentistry faculty member using a validated pediatric behavior guidance scale from an earlier study [5].

The video-recorded encounters were screened, and those not conducted primarily in English were excluded from this study (8 encounters). An accredited commercial company manually transcribed the videos and annotated the speakers to produce transcripts of the encounters for evaluation. Videos with audio that was too poor in quality for accurate transcription were excluded (2 encounters). A total of 40 encounters were successfully transcribed for this study. As this was a pilot study on using clinical transcripts to score behavior guidance, no sample size calculation was conducted.

A single study member (IKD) rated the transcripts by using a modified version of the previously validated scale [5], and the time taken by the human assessor to rate each encounter was recorded. The same study member randomly rerated 20% (8/40) of the transcripts, and the intrarater reliability bias–adjusted κ score was calculated to be 0.94. The pediatric behavior guidance scale was modified to remove nonverbal components, which are not captured with a text-only approach, giving the transcript (text-only) scale a maximum score of 15 (Figure 1), whereas the original scale had a maximum score of 20 (with higher scores denoting better clinical performance of behavior guidance techniques) in the original study (Figure 2) [5]. For comparison, the video scores from the original study were similarly modified to remove nonverbal components, resulting in a maximum possible score of 15. The transcript scores were compared to the video scores and the modified video scores without nonverbal components. This helped to determine if text-only information could be used to score clinical pediatric behavior guidance performance as reliably as clinical videos.

XSL•FO
RenderX

Figure 1. Modified rating scale for the rating of pediatric behavior guidance performance via the transcript (text-only) approach [11,12].

Video:	Rater:		Time taken:
Category	**Positive Technique**		**Undesirable Technique**
Guidance (direction)	Directs immediate behaviour by command • Giving clear direction and specific instructions		Examples • Coercing (Bribing with toys, threats) • *Negative reinforcement (verbally reprimanding/ punishment for undesired behaviour) • Verbally puts down (e.g. You are too old to be crying) • Statements of criticism (e.g. you are a being a bad patient) Qualitative grading
	Sets rules and limits • Good level of control and sense of assertiveness		
	*Enhancing control • Providing the patient a signal (e.g., raising a hand) to use if he becomes uncomfortable		
Guidance (feedback)	Provides specific feedback • Positive reinforcement and descriptive praise		
	*Explains in age appropriate language • Tell-show-do (2) • Explaining procedures using simple language (2)		
Empathy	Questions for feelings • Ask-tell-ask (inquiring about patient's feelings towards procedures) (2)		
	Reassuring • An empathetic approach with verbal reassurance		
Verbalization	*Communicative guidance • Rapport building (questions and active/reflective listening) (2)		
	Distraction • Using imagination and stories (2)		
Total score /15	Each bullet point is worth 1 or 2 points for a total of 15 points, undesirable techniques portion is graded qualitatively Categories and techniques adapted from Zhuo 2011: "Systematic review of the effect of dental staff behaviour on child dental patient anxiety and behaviour" Patient Education and counseling. 2011;85(1):4-13 *Additional Basic behavior guidance techniques from AAPD Best practice guideline: Behaviour guidance for the pediatric dental patient		

Figure 2. Original rating scale with nonverbal components [11,12].

Video:		Rater:	
Category	**Positive technique**		**Undesirable technique**
Guidance (direction)	Directs immediate behaviour by command • Giving clear direction and specific instructions		Examples • Coercing (Bribing with toys, threats) • *Negative reinforcement (verbally reprimanding/ punishment for undesired behaviour) • Verbally puts down (e.g. You are too old to be crying) • Inappropriate use of holding and restrain • Statements of criticism (e.g. you are a being a bad patient) • *Inappropriate tone of voice (showing anger, irritation) Qualitative grading
	Sets rules and limits • Good level of control and sense of assertiveness		
	*Enhancing control • Providing the patient a signal (e.g., raising a hand) to use if he becomes uncomfortable		
Guidance (feedback)	Provides specific feedback • Positive reinforcement and descriptive praise		
	*Explains in age appropriate language • Tell-show-do (2) • Explaining procedures using simple language (2)		
Empathy	Questions for feelings • Ask-tell-ask (inquiring about patient's feelings towards procedures) (2)		
	Reassuring • An empathetic approach • *Non-verbal communication with posture, facial expression and body language		
Physical contact	Appropriate level of working contact • Reassuring touch accompanied by verbal reassurance (2)		
Verbalization	Loud (clear) and firm voice • Appropriate used of voice control		
	*Communicative guidance • Rapport building (questions and active/reflective listening) (2)		
	*Distraction • Using imagination (stories, singing) (2) • Use of break during stressful procedure		
Total score /20	Each bullet point is worth 1 or 2 points for a total of 20 points, undesirable techniques portion is graded qualitatively Categories and techniques adapted from Zhuo 2011: "Systematic review of the effect of dental staff behaviour on child dental patient anxiety and behaviour" *Patient Education and counseling. 2011;85(1):4-13* *Additional Basic behavior guidance techniques from AAPD Best practice guideline: Behaviour guidance for the pediatric dental patient		

The free-to-use (GPT-4o) and paid (o4-mini-high) versions of ChatGPT (OpenAI) were guided with prompts (Multimedia Appendix 1) to rate behavior guidance performance by using the study rating scale and background information on pediatric behavior guidance (ie, from *The Reference Manual of Pediatric Dentistry* by the American Academy of Pediatric Dentistry) [5,11]. A clinical encounter transcript was uploaded to a new chat with the same prompt, and the score was recorded. This was repeated 3 times for each clinical encounter. The scores were marked as consistent if they matched in all 3 runs; if the scores were not consistent, the score with the higher frequency (matched in 2 out of 3 runs) or the middle score was used for the final analysis. Variability was reported as the average difference and the range of the difference between the highest and lowest scores for each encounter. The free-to-use LLM's scores and the paid LLM's scores were compared to the transcript scores to determine if LLMs could perform as well as humans when scoring behavior guidance performance.

XSL•FO
RenderX

Statistical analyses were conducted by using SPSS 29.0 statistical software (IBM Corporation). Descriptive statistics were used to report the consistency of LLM scoring, variability of LLM scores, length of videos, and time taken to rate transcripts. Comparisons were conducted by using the chi-square test and Wilcoxon signed rank test. Intraclass correlation (consistency) was conducted to compare video scores against modified video scores without nonverbal components and transcript scores to determine the reliability between the scales. Intraclass correlation (absolute agreement) and interclass correlation (Pearson coefficient correlation with bias adjustment) were conducted between transcript scores, modified video scores without nonverbal components, and LLM (free-to-use and paid versions) scores to determine the agreement of the scales.

Results

A total of 40 clinical encounters were included in the final analysis. The average video length was 73 (range 37-120) minutes, and the average time taken to rate the transcripts was 12 (range 3-25) minutes. The time taken to rate the transcripts was significantly (P<.001) lower than the time taken to rate the videos.

The video scores were compared to the video scores without nonverbal components, resulting in a consistency intraclass correlation coefficient (ICC) of 0.978 (95% CI 0.959 - 0.988; P<.001), which demonstrates excellent reliability.

The average video score (maximum=20) was 11.64 (SD 3.57), the average video score without nonverbal components (maximum=15) was 8.93 (SD 2.90), and the average transcript score (maximum=15) was 8.93 (SD 3.24). Comparing transcript scores to video scores resulted in a consistency ICC of 0.830 (95% CI 0.679 - 0.910; P<.001), demonstrating good reliability, and comparing transcript scores to video scores without nonverbal components resulted in an absolute agreement ICC of 0.834 (95% CI 0.684 - 0.912; P<.001), demonstrating good agreement (Table 1). The interclass correlation analysis showed similar results (Multimedia Appendix 2).

Table . Agreement of video scores, video scores without nonverbal components, and large language model (LLM) scores when compared against transcript scores.

	Average score (SD)	Compared with transcript scores, ICC[a] (95% CI; *P* value)	Level of reliability/agreement[b]
Transcript scores (maximum=15)	8.93 (3.24)	—[c]	—
Video scores (maximum=20)	11.64 (3.57)	Consistency: 0.830 (0.679 - 0.910; <.001)	Good
Video scores without nonverbal components (maximum=15)	8.93 (2.90)	Absolute agreement: 0.834 (0.684 - 0.912; <.001)	Good
LLM (free-to-use) scores (maximum=15)	10.13 (3.32)	Absolute agreement: 0.729 (0.475 - 0.859; <.001)	Moderate
LLM (paid) scores (maximum=15)	9.33 (3.14)	Absolute agreement: 0.670 (0.377 - 0.825; <.001)	Moderate

[a]ICC: intraclass correlation coefficient.

[b]ICC ranges: above 0.90=excellent; between 0.75 and 0.90=good; between 0.50 and 0.75=moderate; below 0.50=poor [13].

[c]Not applicable.

The free-to-use LLM's average score (maximum=15) was 10.13 (SD 3.32), and the paid LLM's average score (maximum=15) was 9.33 (SD 3.14). Comparing transcript scores to the free-to-use LLM's scores resulted in an absolute agreement ICC of 0.729 (95% CI 0.475 - 0.859; P<.001), demonstrating moderate agreement, and comparing transcript scores to the paid LLM's scores resulted in an absolute agreement ICC of 0.670 (95% CI 0.377 - 0.825; P<.001), also demonstrating moderate agreement.

Both the free-to-use and paid versions of ChatGPT had low consistency, producing the same score in all 3 runs for only 7 out of 40 (18%) and 4 out of 40 (10%) clinical encounters, respectively. Additionally, the free-to-use version of ChatGPT had a significantly (P=.004) smaller average range of scores (mean 1.73, SD 1.18; range 0-4) when compared to that of the paid version of ChatGPT (mean 2.85, SD 1.70; range 0-7) (Table 2).

XSL•FO
RenderX

Table . Consistency and variability of the unpaid and paid versions of ChatGPT.

	Free-to-use ChatGPT (GPT-4o)	Paid ChatGPT (o4-mini-high)	*P* value
Consistency (among 3 runs with the same prompt for each clinical encounter [N=40])[a], n (%)	7 (18)	4 (10)	.33
Variability (average difference for each encounter)[b], mean (SD; range)	1.73 (1.18; 0-4)	2.85 (1.70; 0-7)	.004[c]

[a]Chi-square test.

[b]Wilcoxon signed rank test.

[c]*P*<.005.

Discussion

This study found that using only transcripts to evaluate the clinical pediatric behavior guidance performance of dental students demonstrated good agreement when compared to using video evaluation. Moreover, transcripts required significantly less time to evaluate when compared to faculty evaluating full videos. Although the commercially available LLMs demonstrated moderate agreement with the faculty rater, they showed a lack of consistency, generating different scores under the same input parameters.

These findings have implications for resource optimization, as LLMs allow faculty to provide personalized feedback to more students, and the time saved allows for more clinical encounters per student. Adopting a theoretical framework of providing formative feedback continuously during training will have the greatest impact on student learning outcomes, which, in our case, was the improvement of clinical pediatric behavior guidance [14]. The use of LLMs can automate this task, thereby reducing the time burden on faculty. However, similar to a previous study, the commercially available LLMs in our study showed a lack of consistency, generating different scores under the same input parameters [9]. As the range of the LLM scores was relatively narrow and showed moderate agreement with the human assessor, future studies should evaluate if this lack of consistency would substantially alter the feedback provided or if the feedback could still be educationally useful. ChatGPT has been evaluated for other automated scoring applications, such as essay marking, with studies showing moderate to excellent correlation between ChatGPT and human assessors when using real-life datasets [15,16]. However, these studies also found wide task-dependent [15] and assessor-dependent [16] variation, and they cautioned the need for good rubric and prompt design and further LLM development before large-scale adoption.

A strength of this study was the use of real-life clinical encounters, as opposed to previous studies' use of simulated clinical encounters [6,7,9,17]. Notably, the greater language variation present in the dataset could explain the LLMs' large variation in scores and poor consistency. Further, while secondary refinement of a prompt after the first run improves the consistency of subsequent runs [18], this effect was lost when starting a new chat. This means that prompt engineering for every encounter or the use of an extremely complex and detailed prompt was necessary, making LLMs impractical for large-scale translation. Additionally, this study examined verbatim clinical encounter transcripts that contained informal language, such as slang and colloquialisms. Previous research showed that current LLMs have issues with processing informal language [19] and require substantial fine-tuning to achieve better performance [20]. This limitation in the LLMs used in this study possibly contributed to the large variation and poor consistency in scoring.

A limitation of this study was the use of only closed-source LLMs for the evaluation of pediatric behavior guidance. Both the free-to-use version and the paid version of ChatGPT resulted in similarly low consistency, with the free-to-use version demonstrating lower variability. It was difficult to speculate why the free-to-use version outperformed the paid version, as both were closed-source LLMs, but this warrants further investigation in future studies. Although closed-source LLMs provide several advantages, such as higher performance, easy implementation, and general accessibility due to significantly higher financial and technical backing, open-source LLMs are adaptable and cost-effective, thus allowing them to be customized and fine-tuned to specific domains, such as patient communication [21]. The transparency and inclusivity of open-source frameworks, such as Large Language Model Meta AI (LLaMA), could better address the specific requirements of dealing with the large variation and informal language present in real-life clinical communication datasets [22]. Future studies should compare the consistency and variability of closed-source and open-source LLMs.

In conclusion, using transcripts to evaluate students' clinical behavior guidance was time-saving for the faculty and demonstrated good agreement with using videos for evaluation. Although LLMs can further automate the task, more work is needed to improve their consistency and accuracy. These findings can be used to improve the provision of targeted feedback to students for clinical teaching.

Acknowledgments

The authors would like to thank Associate Professor Catherine Hong for her work on the video scoring.

XSL•FO
RenderX

Funding

This study was supported by the Faculty of Dentistry R Fund from the National University of Singapore, and article processing fee support was provided by the Faculty of Dentistry, National University of Singapore.

Data Availability

The datasets generated or analyzed during this study are available from the corresponding author (SH) upon reasonable request.

Authors' Contributions

IKD is considered the first author, and SH is the senior author. IKD, GKYL, and SH conceived the idea and were responsible for funding acquisition. IKD and SH participated in data collection. SH conducted the data analysis. IKD, GKYL, and SH led the interpretation and writing. All authors revised the manuscript for important intellectual content.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Prompts used to guide ChatGPT for the rating of pediatric behavior guidance performance.
[PNG File, 534 KB - mededu_v12i1e83376_app1.png]

Multimedia Appendix 2
Interclass (Pearson coefficient) correlation of transcript scores against video scores and video scores without nonverbal components.
[PDF File, 35 KB - mededu_v12i1e83376_app2.pdf]

Checklist 1
CHART checklist.
[PDF File, 251 KB - mededu_v12i1e83376_app3.pdf]

References

1. Davidovich E, Pessov Y, Baniel A, Ram D. Levels of stress among general practitioners, students and specialists in pediatric dentistry during dental treatment. J Clin Pediatr Dent 2015;39(5):419-422. [doi: 10.17796/1053-4628-39.5.419] [Medline: 26551363]
2. York KM, Mlinac ME, Deibler MW, Creed TA, Ganem I. Pediatric behavior management techniques: a survey of predoctoral dental students. J Dent Educ 2007 Apr;71(4):532-539. [Medline: 17468315]
3. Bimstein E, Azari AF, Sotto JJ, Riley JL3. Students' perceptions about pediatric dental behavior guidance in an undergraduate four-year dental curriculum. J Dent Educ 2009 Dec;73(12):1366-1371. [Medline: 20007491]
4. Wisniewski B, Zierer K, Hattie J. The power of feedback revisited: a meta-analysis of educational feedback research. Front Psychol 2020 Jan;10:3087. [doi: 10.3389/fpsyg.2019.03087] [Medline: 32038429]
5. Hu S, Mok BYY, Tok WW, Wong ML, Hong CHL. Teaching pediatric behavior management in student dentists with constructive video feedback from faculty. J Dent Educ 2021 Aug 12. [doi: 10.1002/jdd.12756] [Medline: 34383296]
6. Brügge E, Ricchizzi S, Arenbeck M, et al. Large language models improve clinical decision making of medical students through patient simulation and structured feedback: a randomized controlled trial. BMC Med Educ 2024 Nov 28;24(1):1391. [doi: 10.1186/s12909-024-06399-7] [Medline: 39609823]
7. Holderried F, Stegemann-Philipps C, Herrmann-Werner A, et al. A language model-powered simulated patient with automated feedback for history taking: prospective study. JMIR Med Educ 2024 Aug 16;10:e59213. [doi: 10.2196/59213] [Medline: 39150749]
8. Abd-Alrazaq A, AlSaad R, Alhuwail D, et al. Large language models in medical education: opportunities, challenges, and future directions. JMIR Med Educ 2023 Jun 1;9:e48291. [doi: 10.2196/48291] [Medline: 37261894]
9. Heston TF, Lewis LM. ChatGPT provides inconsistent risk-stratification of patients with atraumatic chest pain. PLoS One 2024 Apr 16;19(4):e0301854. [doi: 10.1371/journal.pone.0301854] [Medline: 38626142]
10. CHART Collaborative, Huo B, Collins GS, et al. Reporting guideline for chatbot health advice studies: the CHART statement. JAMA Netw Open 2025 Aug 1;8(8):e2530220. [doi: 10.1001/jamanetworkopen.2025.30220] [Medline: 40747871]
11. American Academy of Pediatric Dentistry. Behavior guidance for the pediatric dental patient. In: The Reference Manual of Pediatric Dentistry: American Academy of Pediatric Dentistry; 2025:379-399 URL: https://www.aapd.org/globalassets/media/policies_guidelines/bp_behavguide.pdf [accessed 2026-06-08]
12. Zhou Y, Cameron E, Forbes G, Humphris G. Systematic review of the effect of dental staff behaviour on child dental patient anxiety and behaviour. Patient Educ Couns 2011 Oct;85(1):4-13. [doi: 10.1016/j.pec.2010.08.002] [Medline: 20807676]

XSL•FO
RenderX

13. Koo TK, Li MY. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. J Chiropr Med 2016 Jun;15(2):155-163. [doi: 10.1016/j.jcm.2016.02.012] [Medline: 27330520]
14. Hattie J, Timperley H. The power of feedback. Rev Educ Res 2007 Mar;77(1):81-112. [doi: 10.3102/003465430298487]
15. Quah B, Zheng L, Sng TJH, Yong CW, Islam I. Reliability of ChatGPT in automated essay scoring for dental undergraduate examinations. BMC Med Educ 2024 Sep 3;24(1):962. [doi: 10.1186/s12909-024-05881-6] [Medline: 39227811]
16. Manning J, Baldwin J, Powell N. Human versus machine: the effectiveness of ChatGPT in automated essay scoring. Innovations in Education and Teaching International 2025;62(5):1500-1513. [doi: 10.1080/14703297.2025.2469089]
17. Webb JJ. Proof of concept: using ChatGPT to teach emergency physicians how to break bad news. Cureus 2023 May 9;15(5):e38755. [doi: 10.7759/cureus.38755] [Medline: 37303324]
18. Wang L, Chen X, Deng X, et al. Prompt engineering in consistency and reliability with the evidence-based guideline for LLMs. NPJ Digit Med 2024 Feb 20;7(1):41. [doi: 10.1038/s41746-024-01029-4] [Medline: 38378899]
19. Wuraola I, Dethlefs N, Marciniak D. Understanding slang with LLMs: modelling cross-cultural nuances through paraphrasing. Presented at: 2024 Conference on Empirical Methods in Natural Language Processing; Nov 12-16, 2024; Miami, Florida p. 15525-15531. [doi: 10.18653/v1/2024.emnlp-main.869]
20. Sun Z, Hu Q, Gupta R, Zemel R, Xu Y. Toward informal language processing: knowledge of slang in large language models. arXiv. Preprint posted online on Apr 2, 2024. [doi: 10.48550/arXiv.2404.02323]
21. Xu J, Ding Y, Bu Y. Position: open and closed large language models in healthcare. arXiv. Preprint posted online on Jan 17, 2025. [doi: 10.48550/arXiv.2501.09906]
22. Manchanda J, Boettcher L, Westphalen M, Jasser J. The open source advantage in large language models (LLMs). arXiv. Preprint posted online on Dec 16, 2024. [doi: 10.48550/arXiv.2412.12004]

ABBREVIATIONS

AI: artificial intelligence
CHART: Chatbot Assessment Reporting Tool
ICC: intraclass correlation coefficient
LLaMA: Large Language Model Meta AI
LLM: large language model

Edited by D Chartash; submitted 01.Sep.2025; peer-reviewed by AM Meldrum, M Pang; revised version received 06.May.2026; accepted 08.May.2026; published 12.Jun.2026.

Please cite as:
Dhillon IK, Lee GKY, Hu S
Evaluating the Pediatric Behavior Guidance of Students Based on Actual Clinical Transcripts Scored by Faculty and Large Language Models: Pilot Comparative Study
JMIR Med Educ 2026;12:e83376
URL: https://mededu.jmir.org/2026/1/e83376
doi:10.2196/83376

© Ishreen Kaur Dhillon, Gabriel Keng Yan Lee, Shijia Hu. Originally published in JMIR Medical Education (https://mededu.jmir.org), 12.Jun.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Education, is properly cited. The complete bibliographic information, a link to the original publication on https://mededu.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Evaluation of the Bridge-in, Objective, Preassessment, Participatory Learning, Postassessment, and Summary–Based Instructional Model in Disaster Nursing: Quasi-Experimental Study

Li Pei*, MSc; Taoyu Jia*, BSc; Yaxin Liu, BSc; Xiao-li Pang, PhD; Muzi Li, BSc; Haoying Dou, MSc

Tianjin University of Traditional Chinese Medicine, No.10 Poyang Lake Road, Tuanbo New City West, Jinghai District, Tianjin, China

*these authors contributed equally

Corresponding Author:
Haoying Dou, MSc
Tianjin University of Traditional Chinese Medicine, No.10 Poyang Lake Road, Tuanbo New City West, Jinghai District, Tianjin, China

Abstract

Background: The undergraduate stage is critical for developing nursing students' disaster nursing competencies; however, traditional lecture-based teaching often fails to promote active learning, and innovative instructional models remain underused in this field. The BOPPPS (Bridge-in, Objective, Preassessment, Participatory Learning, Postassessment, and Summary) model is a student-centered, closed-loop instructional framework that emphasizes active participatory learning and real-time feedback.

Objective: This study aimed to evaluate the effects of a BOPPPS-based blended teaching model on disaster nursing competencies and academic outcomes among undergraduate nursing students.

Methods: This study used a quasi-experimental design with 77 undergraduate nursing students allocated to 3 groups: an experimental group receiving BOPPPS-based blended teaching (n=27), a control group with traditional lecture-based teaching (n=25), and a comparison group without systematic instruction (n=25). Questionnaires were used to assess students' self-perceived disaster nursing knowledge, professional operational skills, and overall comprehensive competencies. The final examination assessed students' mastery of theoretical knowledge and applied reasoning.

Results: All 3 groups showed significant within-group improvements in disaster competence (all P<.001, n=77). Intergroup comparisons revealed significant differences (H=8.843, P=.01, η^2=0.093); post hoc analysis showed that the experimental group (mean rank=48.79) scored significantly higher than the comparison group (mean rank=29.76, P_{adj}=.009), but not significantly higher than the control group (mean rank=38.32, P_{adj}=.34). The experimental group had a significantly higher overall course score than the control group (86.93, SD 5.08 vs 81.60, SD 6.90; t_{50}=−3.186, P=.002).

Conclusions: The implementation of a disaster nursing curriculum positively impacts students' disaster nursing competency. Since the BOPPPS-based blended teaching model promotes active student engagement and improves teaching quality, we recommend its integration into disaster nursing education.

(JMIR Med Educ 2026;12:e77728) doi:10.2196/77728

KEYWORDS

disaster nursing education; BOPPPS; competency; curriculum construction; undergraduate; Bridge-in, Objective, Preassessment, Participatory Learning, Postassessment, and Summary

Introduction

Background

Disaster events present a significant threat to social development and human survival. Over recent years, there has been an increasing trend in the frequency of disasters worldwide [1]. Disasters can impose enormous stress on communities, leading to widespread disruption of daily life and challenging the resilience of social structures [2]. Disaster nursing is a crucial aspect of emergency relief, involving caregivers who use their specialized knowledge and disaster-related skills to provide urgent assistance to those affected by disasters. The primary goal is to minimize the adverse effects on public health and safety, reducing the damage caused by the disaster event [3]. Nurses are an essential component of the rescue team and play a crucial role in the overall functioning of the rescue system. They ensure that rescue operations run effectively while providing compassionate care to those in need. During and after disasters, nurses can critically assist patients, enhancing the quality of care and supporting recovery efforts [4]. The disaster nursing competencies of nurses directly impact the efficiency of disaster relief efforts [5]. Enhancing nurses' disaster nursing skills enables them to respond swiftly and effectively during

XSL•FO
RenderX

emergencies, thereby minimizing the societal losses caused by disasters. Research has demonstrated that nurses frequently experience heightened levels of physical and mental stress during periods of uncertainty and shock following disasters [6]. Although there is generally a high level of willingness among nursing students to participate in disaster relief, many are still reluctant to engage in rescue operations involving infectious diseases or man-made disasters [7]. The root cause of this reluctance among nursing students stems from complex psychological and structural factors, rather than a mere apprehension toward disaster events. Empirical evidence from our previous mixed methods inquiry regarding the baseline status indicates that students' deployment willingness and response behaviors are profoundly shaped by their foundational disaster literacy, self-efficacy, and career adaptability [8]. Perceived deficits in disaster nursing competency can trigger situational anxiety when anticipating high-risk environments, thereby undermining students' psychological readiness. Consequently, implementing targeted disaster nursing education that bridges the gap between theoretical knowledge and psychological adjustment is crucial to dismantle these cognitive and psychological barriers, thereby comprehensively enhancing students' actual preparedness.

In China, disaster nursing education is still in the early stages of development and remains in its infancy, lacking a systematic and comprehensive academic framework as well as standardized teaching methodologies [9]. Consequently, only a limited number of colleges and universities currently offer specialized courses in this domain, with existing programs relying heavily on traditional lecture-based teaching. This conventional paradigm introduces severe pedagogical limitations, such as placing an overemphasis on book-based theoretical knowledge, unbalancing the relationship between teaching and learning, and marginalizing students' central role in the learning process while suppressing their clinical initiative and creativity [10]. To overcome these pedagogical limitations, modern disaster education requires alternative active learning frameworks that reposition students as central participants. BOPPPS, a structured instructional design model, was introduced by education scholar Douglas Kerr in 1978 [11]. This model provides a framework for creating effective and engaging lessons by emphasizing 6 key components: Bridge-in (B), Objective (O), Preassessment (P), Participatory Learning (P), Postassessment (P), and Summary (S). Unlike traditional lecture-based teaching, this instructional approach provides teachers with a clearer framework for structuring lessons. By using preassessment and postassessment scores as timely feedback, educators can flexibly adjust their teaching strategies and master students' learning progress. This teaching model emphasizes the creation of a student-centered learning environment, supporting students in managing their own learning content and providing flexible learning spaces that encourage purposeful engagement and demonstration of knowledge. This teaching model, which focuses on cultivating students' curiosity, has proven effective in boosting learning satisfaction and significantly improving instructional efficacy [12]. Concurrently, research shows that early preparedness training during undergraduate study enhances the development of disaster nursing skills [13].

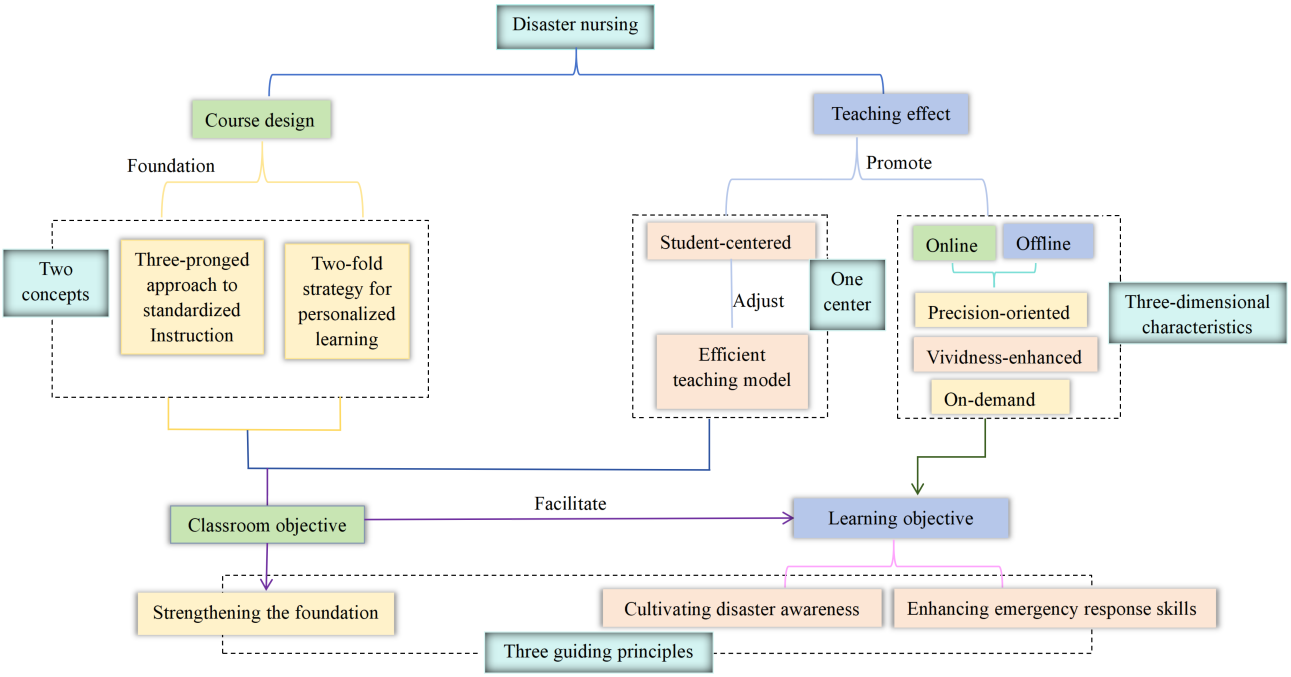
Over the past decade, various medical disciplines in China have gradually integrated the BOPPPS teaching model. Despite this widespread multidisciplinary adoption, empirical applications within the highly specialized field of disaster nursing remain largely absent from domestic research. The course is designed to enable students to (1) acquire core knowledge of disaster nursing, (2) understand the principles of emergency response and basic first aid, and (3) develop nursing competencies for common disaster scenarios. Accordingly, we conducted a quasi-experimental study to address the following research question: Is the BOPPPS-based blended teaching model more effective than traditional lecture-based teaching in improving undergraduate nursing students' disaster nursing competency and academic performance? We hypothesized that, compared with the control group, the experimental group would achieve significantly greater gains in disaster nursing competency and significantly higher scores on the objective final examination. The findings of this study are intended to inform nursing educators, disaster nursing curriculum designers, and clinical emergency training managers seeking to strengthen course design and instructional strategies in disaster nursing education.

Conceptual Framework

Drawing on the BOPPPS model, we constructed a blended teaching framework for disaster nursing education that comprises 4 interrelated components (Figure 1). Component 1 (standardization with personalization) establishes an integrated online-offline teaching structure in which content, resources, and assessment standards are unified and classroom-based real-time feedback enables differentiated guidance. Component 2 (student-centered pedagogy) activates students' intrinsic motivation through clear learning objectives and participatory learning contexts, thereby shifting them from passive reception to active inquiry. Component 3 (technological support) uses the Chaoxing Learning Platform to integrate online resources, in-person lectures, and after-class interaction, enhancing learning efficiency. Component 4 (nursing talent development) follows three guiding principles: consolidating the knowledge foundation, fostering systematic and critical thinking, and enhancing knowledge transfer ability. This conceptual framework guided the subsequent instructional design and outcome evaluation of the course.

XSL•FO
RenderX

Figure 1. The conceptual framework of curriculum construction.



Methods

Study Design

We conducted this study during the second semester of the 2023 to 2024 academic year. To implement the study effectively, we systematically developed syllabi, instructional designs, evaluation methods, and quality control measures. We developed the curriculum design and instructional modules of the disaster nursing course based on the International Council of Nurses (ICN) Core Competencies in Disaster Nursing Version 2.0, jointly issued by the ICN and the World Health Organization (WHO) [14]. To bridge the gap between these macro standards and undergraduate requirements, our research group used a 2-round Delphi expert consultation to systematically select and adapt the specific course topics, learning objectives, and competencies. We convened a panel of 15 multidisciplinary experts, all of whom had more than 10 years of professional experience and intermediate or higher titles. Through this consensus-building process, we refined the initial 5 primary and 14 secondary domains based on expert feedback from the first round, which optimized specific item phrasings (eg, modifying “first-aid skills” to “on-site rescue skills”) and added critical dimensions, such as rehabilitation for postdisaster psychological issues, while explicitly considering cross-disciplinary integration with other nursing tracks. Consequently, this rigorous adaptation resulted in a curriculum with a total of 27 instructional hours comprising 9 hours of online instruction and 18 hours of face-to-face instruction, organized into 5 teaching units and 13 content items, as shown in Table 1.

XSL•FO
RenderX

Table . Course framework and clock hours.

Teaching unit and content	Face-to-face instruction	Online instruction	Total hours
Chapter 1: Introduction			
Concept and classification of disasters	2	—[a]	2
Disaster emergency response system	0.5	1	1.5
Management of disaster stages	0.5	—	0.5
Chapter 2: Common disaster rescue and care			
Natural disaster nursing	4	2	6
Nursing in technological disasters	3	3	6
Nursing care for common medical conditions in disaster rescue	3	—	3
Basic on-site disaster rescue skills	—	3	3
Chapter 3: Psychological intervention in disaster rescue			
Overview, psychological characteristics, and interventions for affected populations	0.5	—	0.5
Mental health maintenance and intervention for rescue workers	0.5	—	0.5
Chapter 4: Postdisaster rehabilitation and nursing			
Rehabilitation of common neurological disorders after disasters	1	—	1
Rehabilitation of common musculoskeletal disorders after disasters	1	—	1
Chapter 5: Hygiene and epidemic prevention in disaster rescue			
Postdisaster environmental health protection	1	—	1
Population health protection	1	—	1
Total	18	9	27

[a]Not applicable.

This study included 3 groups. We determined group assignments based on the students' voluntary enrollment in the elective course sessions. Participants registered for the Tuesday class comprised the experimental group and received BOPPPS-based blended teaching. Those enrolled in the Monday session formed the control group and received traditional lecture-based teaching. Additionally, we used convenience sampling to recruit a comparison group from students in the same academic cohort who did not enroll in the elective. Participants in this group did not receive any disaster nursing education, and their data served as a baseline for determining whether the specialized curriculum was the sole driver of improvements in students' disaster response capabilities. Both the control and experimental groups attended the course in parallel and completed all assigned learning tasks; Table 2 details the specific curricular differences. To ensure the comparability of baseline levels across groups, all 3 cohorts participated in an identical core nursing curriculum, with the "Disaster Nursing" elective course serving as the sole independent variable. During the intervention, we administered precourse and postcourse questionnaires to all groups, assessing self-perceived disaster nursing knowledge, professional operational skills, and overall comprehensive competencies to verify the intervention's efficacy. At the end of the semester, students in the experimental and control groups completed a unified, comprehensive final examination assessing theoretical knowledge and applied reasoning, thereby comparing the BOPPPS-based blended teaching with traditional lecture-based teaching.

XSL•FO
RenderX

Table . Comparison of instructional activities and design.

Phase	BOPPPS[a]-based blended teaching (experimental group, n=27)	Traditional lecture–based teaching (control group, n=25)	Overlap (shared elements)
Preclass	Bridge-in, Objective, Preassessment: Students previewed courseware, reviewed learning objectives, and completed diagnostic quizzes via the Chaoxing platform.	General autonomous preview without structured guidance or digital tracking.	Same textbook and preparatory topics, completed the same preassessment.
In-class (1)	Participatory Learning: Students engaged in active learning through classroom presentations and group discussions.	Teacher-led lecturing with Q&A sessions.	Same curriculum standards and core knowledge points.
In-class (2)	Postassessment: Real-time knowledge testing using the interactive "Discussion" function of the Chaoxing platform to monitor mastery.	Standardized posttest conducted at the end of the session to assess immediate mastery.	Both groups completed the same posttest content for outcome evaluation.
In-class (3)	Summary: Structured conclusion by teachers based on real-time feedback and knowledge mapping.	Teacher-led summary of key points from the lecture.	Same concluding knowledge framework.
Postclass	Consolidation: Knowledge reinforcement through Chaoxing resources integrated into the consolidation stage of BOPPPS, with structured review.	Knowledge review via personal notes using the same Chaoxing resources.	Same reference materials and study goals.

[a]BOPPPS: Bridge-In, Objective, Preassessment, Participatory Learning, Postassessment, and Summary.

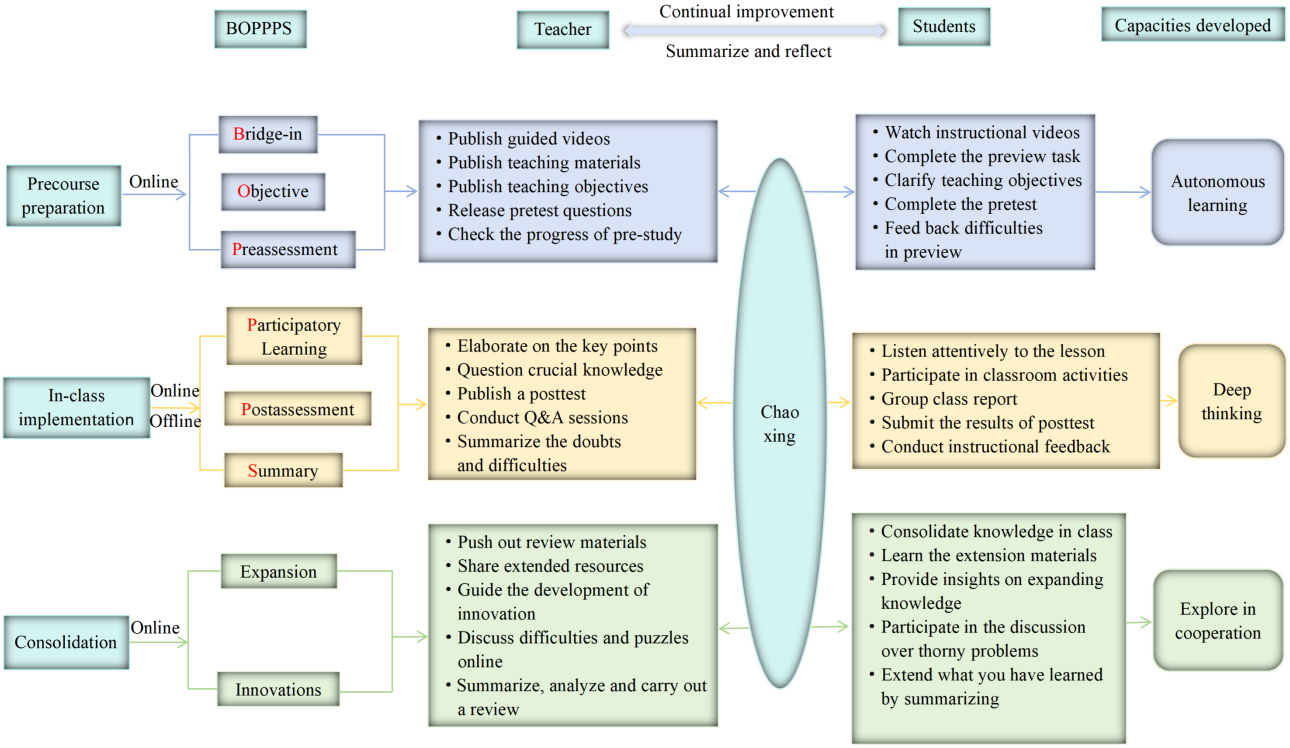
The control group primarily used the traditional lecture-based teaching, with the following structured steps: in the precourse phase, teachers clarified the teaching objectives by outlining the essential textbook knowledge and skills that students needed to master. They meticulously planned the teaching content to ensure alignment with these objectives, selected and collected relevant case studies and supplementary materials to support learning, and designed appropriate teaching methods to effectively deliver the curriculum. Second, during in-class instruction, the teacher explained the theoretical knowledge from the textbook, supplemented by PowerPoint slides that included videos related to disaster nursing and case studies to enhance understanding. Students were encouraged to outline key points from the textbook content and take detailed notes during the lectures. The teacher then guided students' thinking through the analysis of provided cases, fostering active participation in group discussions and responding to teacher-led questions. Finally, the teacher summarized the key points covered in the class and addressed any questions raised by the students, ensuring that all concepts were clearly understood.

The experimental group delivered the course based on the BOPPPS teaching model, incorporating course resources from the Chaoxing Learning Platform. This teaching model comprised 3 main phases: preclass preparation, implementation during the lesson, and consolidation and feedback at the end of the lesson. The 6 core components of the BOPPPS model, including Bridge-in, Objective, Preassessment, Participatory Learning, Postassessment, and Summary, span both the preclass and in-class phases. This structure enabled students to focus and systematize their knowledge effectively. The preclass preparation stage: (1) Bridge-in: Teachers distributed introductory videos and instructional materials related to the disaster nursing course to students via the Chaoxing Learning Platform, designed to present knowledge content more intuitively and engagingly, thereby deepening students' cognitive impressions and enhancing the quality of preclass preparation. (2) Objective: The teacher clearly defined the teaching objectives and delivered the guided-learning content to students via the online platform. Students then completed the preclass guided learning tasks on the Chaoxing Learning Platform, establishing clear learning expectations before class, thereby making in-class activities more targeted and effective. (3) Preassessment: Students completed the corresponding preassessment questions based on the teaching materials released by the teacher, thereby clarifying their current level of knowledge mastery and identifying weak points to focus on in their subsequent learning. The in-class implementation phase: (4) Participatory Learning: Teachers systematically reviewed and analyzed the key content of newly introduced courses, facilitating teacher-student interactions through the Chaoxing Learning Platform. This included posing questions on critical knowledge points and encouraging students to engage in in-depth discussions and rapid responses to relevant queries. To elevate student engagement, we designed a series of in-class participatory activities, including group presentations and case analyses, fostering an interactive learning environment. (5) Postassessment: During the postassessment phase, teachers posted follow-up test questions and facilitated Q&A sessions to clarify any remaining doubts. Additionally, they guided students in summarizing their reflections via the Chaoxing Learning Platform. Students submitted their quiz answers on time during class, provided constructive feedback on their learning experiences, and posed questions following submission. (6) Summary: Upon completion of the teaching of key content

XSL•FO
RenderX

and related classroom activities, the teacher provided a comprehensive summary of the material covered, highlighting key points and clarifying any challenging areas. Postcourse consolidation and feedback phase: After class, students used the course materials available on the Chaoxing Learning Platform to consolidate and reinforce what had been covered during the lesson. The specific teaching process of the BOPPPS-based blended curriculum construction is shown in Figure 2.

Figure 2. Flowchart of Chaoxing Learning Platform-Bridge-In, Objective, Preassessment, Participatory Learning, Postassessment, and Summary (BOPPPS) blended instructional design.



Course Implementation Context

At the beginning of each semester, the university's Academic Affairs Office publishes an official administrative document regarding the optimization of the undergraduate elective course system, thereby formally activating the application and development process for elective courses across the entire university. To align with the practical requirements of course development, our research team first guided students in implementing a multimedia project aimed at promoting disaster nursing knowledge, while conducting a preliminary needs assessment to evaluate nursing students' baseline disaster competencies and learning willingness. Researchers regularly compiled student feedback and, in conjunction with the assessment results, dynamically optimized the elective course content within the university's overarching curriculum framework, thereby providing empirical evidence to support the design and approval of the course. In accordance with the standards outlined in the regulatory documents, our team submitted a comprehensive course proposal to the academic administration, detailing the course rationale, learning objectives, curriculum content, and implementation plan. The department responsible for the course established an independent evaluation panel comprising at least 3 senior academic experts holding associate professor titles or higher, alongside invited external peer experts to ensure optimal review objectivity. Using uniform quality assessment standards, this review panel conducted a comprehensive evaluation of the course across multiple critical dimensions, including the feasibility of the learning objectives, the appropriateness of the class schedule, and the structural fidelity of the syllabus. Curricula securing a comprehensive score of 80 or higher were eligible for development approval upon successful validation. Having met the required standard, the disaster nursing course was formally included in the undergraduate talent cultivation scheme and officially commenced during the second semester of the 2023 to 2024 academic year.

Participants

This study recruited 77 third-year undergraduate nursing students from Tianjin University of Traditional Chinese Medicine, employing a quasi-experimental design. We recruited undergraduate nursing students prior to the commencement of the academic semester, using an open-enrollment framework to offer the disaster nursing course as an elective to all third-year nursing students. The inclusion criteria were as follows: (1) full-time undergraduate nursing students; (2) no prior experience with disaster nursing courses that integrated the traditional curriculum; and (3) voluntary participation in this study and signing of the informed consent form. To implement the recruitment process, the research team promoted information about the elective course through the university's channels, and students enrolled voluntarily based on their learning interests.

XSL•FO
RenderX

Questionnaire Survey

General Information Questionnaire

The research team developed a self-administered general information questionnaire to collect participants' general information, including home location (county/town vs city), whether they voluntarily chose the nursing program, whether they had been taught disaster prevention knowledge and skills by their parents, and whether they had attended relevant practical training.

Disaster Nursing Competency Scale

To evaluate the course's effectiveness, we adopted a dual approach combining questionnaires and a final examination. For the questionnaire component, all 77 participants completed precourse and postcourse tests via the Questionnaire Star platform, yielding a 100% effective response rate. This study employed the Disaster Nursing Competency Scale (DNCS) developed by Hu [15] to assess participants' disaster nursing competencies. This instrument is theoretically grounded in the core-competency framework for disaster nursing jointly established by the ICN and the WHO, and was originally developed using a mixed methods approach involving a literature review, 2 rounds of Delphi expert consultations, and a pilot survey. The scale comprises three first-level indicators, 11 second-level indicators, and 50 third-level indicators. The first-level indicators include knowledge system, professional competence, and comprehensive qualities. The second-level indicators include disaster nursing foundational knowledge, disaster nursing specialized knowledge, assessment and observation skills, nonemergency response capabilities, on-site emergency care skills, disaster nursing management abilities, psychological crisis intervention skills, communication and collaboration abilities, physical fitness, psychological resilience, and moral and ethical standards. In the original development study, the scale demonstrated sound validity and reliability: the expert-rated Scale-level Content Validity Index was 0.95, correlation coefficients between first- and second-order dimensions ranged from 0.810 to 0.982 and from 0.394 to 0.976, respectively, and the Cronbach α and split-half reliability of the full scale reached 0.984 and 0.920, respectively. We used a 5-point Likert scale for all responses, with 1 indicating "very poor" and 5 indicating "very good"; total scores range from 50 to 250, with higher scores indicating greater disaster nursing competency. We used the original version of the scale directly, without any adjustments or omissions of any items, to preserve its structural integrity; in the current sample, the scale also demonstrated high internal consistency, with a Cronbach α of 0.927.

Academic Achievement Assessment

To comprehensively evaluate students' academic achievement, this study established a multidimensional course assessment framework. This framework derived the final course grade from a process assessment (40%) and a final examination (60%). The process assessment, accounting for 40% of the total course grade, was derived from objective data automatically collected through the university's online learning platform. The process assessment comprised 5 weighted modules, with the combined weights totaling 100%: (1) chapter task points (25%). This module assessed the completion of video- and audio-based learning tasks. Students who completed all task points received full credit; otherwise, the score was calculated in proportion to the completed task points. (2) Chapter quizzes (20%). We administered chapter quizzes during each online class session to assess mastery of the session content. The module score was calculated as the average percentage across all chapter quizzes. (3) Group problem-based learning tasks (40%). Each group assignment was combined as a weighted average of instructor evaluation (60%), intergroup peer evaluation (10%), intragroup peer evaluation (20%), and self-evaluation by group members (10%). The instructor's evaluation criteria included the appropriateness of the presentation materials, content accuracy, clarity of organization and logical rigor, quality of PowerPoint design, and the presenter's delivery and demeanor. Each assignment was scored on a 100-point scale, and the module score was the average of all group assignment scores. (4) Attendance (10%). This module captured the attendance rate, defined as the number of sessions attended divided by the total number of required sessions. (5) Class discussions (5%). This module tracked points earned by students: 20 points for initiating a new topic, 5 points for replying to a topic, and 5 points for receiving a "like," with a maximum total of 100 points. We scored all 5 modules on a 0 to 100 scale and aggregated them according to the weights above to produce the process assessment score. We developed the final examination for the disaster nursing course based on the course syllabus, which was structured according to the ICN/WHO disaster nursing core competency framework. To clarify the content basis of this objective academic assessment, the research team analyzed the assessment domains covered by the final examination. The intended learning objectives of the course were for students to be able to: (1) describe the concepts, classification, characteristics, and management phases of disasters and the structure of disaster emergency rescue systems; (2) apply nursing principles in common disaster rescue scenarios; (3) demonstrate knowledge of essential first-aid and emergency response skills in disaster settings; (4) identify major postdisaster nursing and rehabilitation needs; and (5) recognize common psychological responses after disasters and identify basic principles of psychological crisis intervention, and explain key principles of postdisaster public health protection and epidemic prevention. Because several learning objectives overlapped conceptually, the final examination was not designed with a strict one-to-one correspondence between each item and a single learning objective. Instead, it was structured and analyzed according to 5 assessment domains, each corresponding to one or more learning objectives. These domains included foundations of disaster nursing, common disaster rescue nursing and first aid, psychological intervention in disaster rescue, postdisaster rehabilitation and nursing, and hygiene and epidemic prevention in disaster rescue.

At the conclusion of the curriculum, we administered the summative evaluation through a standardized, invigilated offline final examination. The instructional emphasis of the course syllabus dictated the score distribution. We designed the 100-point assessment instrument to evaluate multiple cognitive domains. The detailed arrangement of question types and score

XSL•FO
RenderX

distribution is as follows: (1) 25 single-choice questions (50 points), which primarily assessed students' mastery of foundational theoretical knowledge across the major course modules, including disaster concepts and classification, disaster rescue principles, common disaster rescue nursing, psychological intervention, and postdisaster public health and epidemic prevention; (2) 15 true or false items (30 points), which mainly evaluated rapid judgment and recognition of correct actions in disaster-related situations, including on-site first aid, emergency response, safety principles, and postdisaster infection prevention measures; and (3) a case-based essay question (20 points), which tested higher-order reasoning and decision-making by requiring students to analyze a complex disaster scenario, identify key problems, and propose appropriate rescue and nursing interventions. To ensure methodological rigor and objective scoring, the final examination was administered under strict proctoring in a controlled examination environment. After the final examination, one senior instructor graded all responses according to a predefined, standardized scoring rubric to maintain scoring consistency. Upon completion of grading, the scores were entered into the course-assessment database for automatic calculation of the overall course score. The resulting final overall course score was then released to students via the university's encrypted digital learning management system.

Statistical Analysis

This study performed data entry and analysis using SPSS statistical software (version 27.0; IBM Corp). Categorical variables were summarized as frequency counts (n) and percentages (%), with the chi-square test evaluating between-group differences. The Shapiro-Wilk test assessed the normality of continuous variables. For normally distributed data, we reported variables as means (SDs), and used the independent-samples *t* test to compare the 2 groups. For nonnormally distributed data, we expressed variables as median (IQR) and applied the Mann-Whitney *U* test for 2 groups. Within-group pretest versus posttest paired comparisons were performed using the Wilcoxon signed-rank test, with effect size *r* calculated as $r=|Z|/n$. According to Cohen [16] criteria, *r* values of 0.1, 0.3, and 0.5 indicate small, medium, and large effects, respectively. For comparisons among 3 groups, the Kruskal-Wallis *H* test compared the distributions, which were delineated using mean ranks. When the *H* test revealed a significant overall difference, Dunn post hoc test with Bonferroni correction was used for the pairwise comparisons. Furthermore, ordinal eta-squared (η^2) captured the effect size via the formula $\eta^2=(H-k+1)/(N-k)$. Based on Cohen conventional benchmarks, η^2 values of 0.01, 0.06, and 0.14 denote small, medium, and large effects, respectively [17]. Two-tailed $P<.05$ or $P_{adj}<.05$ for post hoc analyses defined statistical significance.

Ethical Considerations

This study obtained ethical clearance from the Medical Ethics Committee of Tianjin University of Traditional Chinese Medicine. The study involved the collection and analysis of data from the teaching intervention, student questionnaires, and course assessments. All participants provided informed consent before completing the questionnaire. The researchers also clarified to all participants that their course assessment scores would be used solely for data analysis within the scope of their informed consent. To ensure academic fairness, participants were explicitly informed that participation in or withdrawal from the survey would have no impact on their academic evaluations or final grades.

Results

Comparison of General Information Among the 3 Groups

Table 3 presents the baseline characteristics of the 77 participants by cohort. The 3 cohorts exhibited no significant differences in terms of home location, whether they chose the nursing program of their own volition, or whether they had attended practical training in disaster prevention skills. Among baseline characteristics, only parental teaching of disaster prevention knowledge and skills differed significantly across groups ($P=.04$). This variable was treated as a descriptive baseline characteristic and was not included as a covariate in subsequent analyses.

XSL•FO
RenderX

Table . Comparison of general information for undergraduate nursing students (N=77).

Variable	Experimental group (n=27), n (%)	Control group (n=25), n (%)	Comparison group (n=25), n (%)	Chi-square (*df*)	*P* value
Home location				1.299 (2)	.52
Counties and towns	15 (56)	16 (64)	12 (48)		
City	12 (44)	9 (36)	13 (52)		
Do you choose your nursing program by yourself?				4.069 (2)	.13
Yes	14 (52)	13 (52)	19 (76)		
No	13 (48)	12 (48)	6 (24)		
Have you been taught disaster prevention knowledge and skills by your parents?				6.432 (2)	.04
Yes	10 (37)	14 (56)	18 (72)		
No	17 (63)	11 (44)	7 (28)		
Have you attended any practical training on disaster prevention skills?				4.069 (2)	.13
Yes	14 (52)	19 (76)	13 (52)		
No	13 (48)	6 (24)	12 (48)		

Comparison of Precourse and Postcourse Disaster Nursing Competency Scores and Score Gains Across the 3 Groups

As shown in Table 4, the results showed that undergraduate nursing students' level of disaster nursing competency improved after the course implementation. On the postassessment, the experimental group scored higher than the control group (experimental: 210.00, IQR 194.00 - 242.00 vs control: 200.00, IQR 186.50 - 210.50). Wilcoxon signed-rank tests revealed significant within-group differences across all groups: control and comparison groups (z=−4.37, r=0.874, P<.001) and experimental group (z=−4.54, r=0.874, P<.001).

Table . Comparison of the disaster nursing competency scores before and after the educational intervention (N=77).

Group	Students, n	Pretest, median (IQR)	Posttest, median (IQR)	*z*	*r*	*P* value
Control group (didactic instruction)	25	148.00 (134.00 - 165.00)	200.00 (186.50 - 210.50)	−4.37	0.874	<.001
Comparison group (no instruction)	25	155.00 (143.00 - 176.00)	200.00 (181.00 - 214.00)	−4.37	0.874	<.001
Experimental group (BOPPPS[a] model)	27	138.00 (127.00 - 152.00)	210.00 (194.00 - 242.00)	−4.54	0.874	<.001

[a]BOPPPS: Bridge-In, Objective, Preassessment, Participatory Learning, Postassessment, and Summary.

Table 5 presents the medians (IQRs) of disaster nursing competency gain scores (posttest − pretest) for the 3 groups, along with the between-group comparisons of these values. The Kruskal-Wallis H test showed a statistically significant difference among the 3 groups (H=8.843, P=.01, η^2=0.093). Dunn post hoc test with Bonferroni correction showed that the experimental group (mean rank=48.79) scored significantly higher than the comparison group (mean rank=29.76, z=2.968, P_{adj}=.009), but no statistically significant difference was observed between the experimental and control groups (mean rank=38.32, z=−1.589, P_{adj}=.34), nor between the control group and the comparison group (z=1.353, P_{adj}=.53).

XSL•FO
RenderX

Table . Comparison of the disaster nursing competency gain scores among the 3 groups (N=77).

	Students, n	Median (IQR)	Mean rank	*P* value	*H* test	η^2
Group				.01	8.843	0.093
Control group (didactic instruction)	25	53.00 (36.50 - 69.50)	38.32			
Comparison group (no instruction)	25	40.00 (25.00 - 62.00)	29.76			
Experimental group (BOPPPS[a] model)	27	61.00 (49.00 - 86.00)	48.79			

[a]BOPPPS: Bridge-In, Objective, Preassessment, Participatory Learning, Postassessment, and Summary.

Comparison of Academic Achievement Outcomes Between the Experimental and Control Groups

As shown in Table 6, the BOPPPS-based blended model produced significantly higher overall course scores in the experimental group (mean 86.93, SD 5.08) than in the control group (mean 81.60, SD 6.90; t_{50}=−3.186, *P*=.002). Grade distributions shifted toward higher bands in the experimental group, with more students scoring in the 90 to 99 and 80 to 89 ranges and fewer in the 70 to 79 range, while the number of nonpassing scores remained comparable across groups (Figure 3). For the process assessment section, the score was higher in the experimental group (91.60, IQR 86.60 - 94.00 vs 88.20, IQR 82.85 - 90.65; *z*=1.969, *P*=.049). On the final examination, the experimental group also scored significantly higher than the control group (88.00, IQR 82.00 - 92.00 vs 80.00, IQR 72.00 - 87.00; *z*=3.107, *P*=.002). Further analysis of the exam submodule data revealed that the score advantage of the experimental group was mainly concentrated in the single-choice questions section, in which the experimental group scored significantly higher than the control group (42.00, IQR 40.00 - 46.00 vs 40.00, IQR 34.00 - 42.00; *z*=2.101, *P*=.04). By contrast, the true or false and case-based essay sections showed no significant differences (true/false: *z*=1.235, *P*=.22; case-based essay: *z*=0.739, *P*=.46).

XSL•FO
RenderX

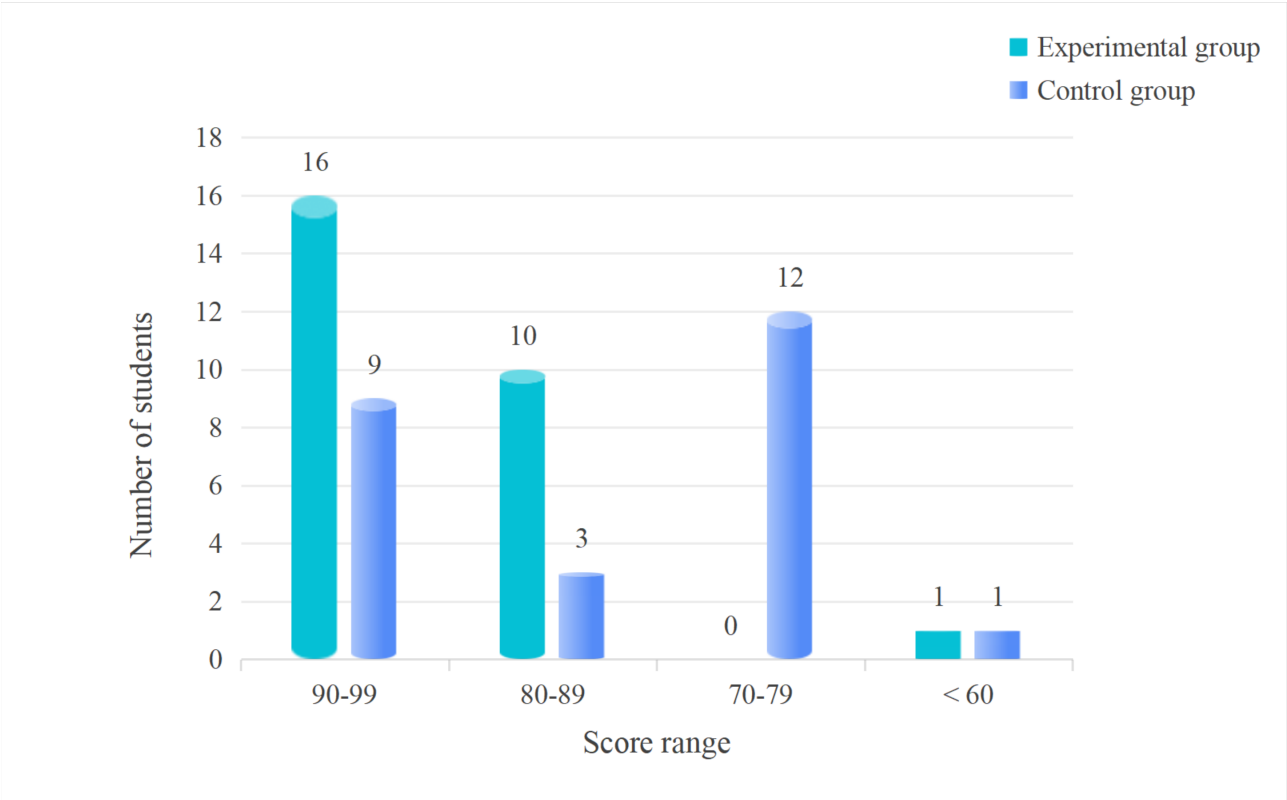
Table . Comparison of course assessment components between the experimental and control groups (N=52).

Assessment component and group	Students, n	Median (IQR)	Mean (SD)	*t* test (*df*)/*z*	*P* value
Process assessment				1.969	.049
Control group	25	88.20 (82.85 - 90.65)	—[a]		
Experimental group	27	91.60 (86.60 - 94.00)	—		
Final examination				3.107	.002
Control group	25	80.00 (72.00 - 87.00)	—		
Experimental group	27	88.00 (82.00 - 92.00)	—		
Single-choice questions				2.101	.04
Control group	25	40.00 (34.00 - 42.00)	—		
Experimental group	27	42.00 (40.00 - 46.00)	—		
True/false questions				1.235	.22
Control group	25	26.00 (26.00 - 29.00)	—		
Experimental group	27	28.00 (26.00 - 28.00)	—		
Case-based essay question				0.739	.46
Control group	25	18.00 (5.00 - 19.00)	—		
Experimental group	27	18.00 (18.00 - 18.00)	—		
Overall course score				−3.186 (50)	.002
Control group	25	—	81.60 (6.90)		
Experimental group	27	—	86.93 (5.08)		

[a]Not applicable.

XSL•FO
RenderX

Figure 3. Overall course scores by group.



We analyzed the distribution of final examination items across assessment domains based on the course syllabus and the primary content assessed by each item. The final examination covered 5 assessment domains, each corresponding to one or more course learning objectives. Common disaster rescue nursing and first aid accounted for the largest proportion of the total score, followed by foundations of disaster nursing, hygiene and epidemic prevention in disaster rescue, psychological intervention in disaster rescue, and postdisaster rehabilitation and nursing. Postdisaster rehabilitation and nursing was assessed mainly through the second subquestion of the case-based essay item and represented a relatively small proportion of the total score. Table 7 presents the distribution of question formats and scores across assessment domains.

Table . Distribution of final examination items across assessment domains and related learning objectives.

Assessment domain	Related learning objectives	Single-choice questions	True/false questions	Case-based essay question	Score
Foundations of disaster nursing	Objective 1	4, 13, 14, 15, 16, 25	28, 29	N/A[a]	16
Common disaster rescue nursing and first aid	Objectives 2 and 3	1, 2, 5, 6, 7, 8, 9, 11, 12, 17, 24	30, 31, 32, 33, 34, 35, 36, 37, 38, 39	41(1)	52
Psychological intervention in disaster rescue	Objective 5	18, 19, 20, 22	26	N/A	10
Postdisaster rehabilitation and nursing	Objective 4	N/A	N/A	41(2)	10
Hygiene and epidemic prevention in disaster rescue	Objective 5	3, 10, 21, 23	27, 40	N/A	12
Total	Objectives 1 - 5	25 items	15 items	1 item	100

[a]N/A: not applicable.

Discussion

Principal Findings

In this study, we employed a quasi-experimental design to analyze quantitative data to evaluate the effectiveness of the BOPPPS-based blended teaching model. The study yielded 3 principal findings. First, regarding the disaster nursing competency, the experimental group demonstrated greater improvement in competency scores than the control and comparison groups. Second, in the final examination, the

XSL•FO
RenderX

experimental group achieved significantly higher total scores than the control group, primarily in the single-choice section. Third, the experimental group scored significantly higher in overall course score than the control group. These results partially confirm the hypothesis: the BOPPPS-based blended teaching is more effective than traditional lecture-based teaching or no systematic instruction in improving disaster nursing competency and objective academic outcomes.

Implications of Findings

These findings offer several implications. First, the BOPPPS-based blended teaching model positively affected both students' self-assessed competence and objective course grades, suggesting that it enhances not only perceived ability but also quantifiable learning outcomes, which supports its broader adoption in disaster nursing courses. Second, the Bridge-in and Preassessment components helped students enter a learning mindset early. This integration of online self-study with in-person instruction addresses a key limitation of traditional lecture-based teaching in promoting active learning. Third, because the experimental and control groups did not differ significantly in true/false questions or case-based essay questions, future courses should incorporate more simulation-based training, situational judgment tasks, and case analysis to support the development of higher-order applied competencies.

Comparison to the Literature

Compared with previous studies, the present study addresses several limitations identified in earlier work through its structured design. Tussing et al [18] synthesized the literature on disaster preparedness education for nurses and nursing students and recommended digitalized teaching approaches; however, as a narrative review, their conclusions relied largely on inferential synthesis of previously published literature and lacked primary empirical data. The present study addresses this gap by providing quantitative evidence of the effectiveness of the BOPPPS-based blended teaching model in disaster nursing. Prior studies have demonstrated that high-immersion simulation approaches, such as virtual reality, can enhance disaster preparedness among nursing students [19], but these approaches typically require substantial hardware investment, and instructors cannot readily observe students' operational details during simulation, which hampers timely feedback. In contrast, the BOPPPS-based blended teaching model implemented in this study relies on routine online learning resources and incorporates a structured real-time classroom feedback loop, offering a cost-effective alternative for disaster nursing education. Phan et al [20] conducted a low-cost disaster simulation exercise but collected data only after the intervention, which limited the ability to attribute score improvements to the teaching method. The present study addressed this by including a control group and employing a pretest-posttest design, allowing both within-group and between-group comparisons and providing stronger quantitative evidence that the observed score gains can be attributed to the teaching method. Glauberman et al [21] reported the DAIS tabletop exercise primarily through qualitative description and experience summary, without quantitative comparative evidence. It should be noted that DAIS focuses on interprofessional team collaboration, whereas the present study focuses on intraprofessional competency development through the BOPPPS-based blended teaching model. The findings of the present study can therefore serve as a foundation for single-discipline disaster nursing curricula and provide a basis for future designs that incorporate interprofessional collaboration. Accordingly, in a context where systematic disaster nursing education frameworks remain underdeveloped in China [9], the model proposed in this study represents a highly scalable and effective implementation strategy.

Effects on Disaster Nursing Competency

Analysis of the DNCS scores indicated significant within-group improvements among all participants, suggesting that different forms of learning may help enhance students' disaster nursing competency. Subsequent intergroup comparisons revealed significant differences in disaster nursing competency between the 3 groups. In terms of competency score gains, the experimental group showed the highest mean rank (48.79), followed by the control group (38.32) and the comparison group (29.76). This pattern suggests that the experimental group had the highest overall ranking in changes in disaster nursing competency. Dunn post hoc test showed a significant difference between the experimental group and the comparison group (P_{adj}=.009), whereas the differences between the experimental and control groups (P_{adj}=.34) and between the control and comparison groups did not reach statistical significance. These findings suggest that systematic teaching intervention may play an important role in improving disaster nursing competency, and that the BOPPPS-based blended teaching model showed a clearer beneficial effect than no systematic training. Although the experimental group had a higher mean rank than the control group, the difference between these 2 groups did not reach statistical significance. Therefore, this study cannot confirm a clear statistical advantage of the BOPPPS-based blended teaching over traditional lecture-based teaching. Future studies with larger sample sizes and greater statistical power should further verify this trend. Notably, although the control group had a higher mean rank than the comparison group, this difference also did not reach statistical significance. This finding suggests that only relying on traditional lecture-based teaching or relying on students' existing professional experience may be insufficient to fully promote disaster nursing competency. Cho and Kwon [22] similarly reported that traditional disaster nursing education may not adequately develop students' practical abilities and has limitations in maintaining long-term learning effects and improving students' responses to nonroutine disaster situations. Therefore, future disaster nursing education should further strengthen structured, interactive, and scenario-based teaching designs to support the integrated development of students' disaster nursing knowledge, skills, and clinical reasoning.

Furthermore, a significant difference was observed between the experimental and comparison groups but not between the experimental and control groups. This finding may be partly attributable to methodological limitations. First, the study may have lacked sufficient statistical power to detect subtle

XSL•FO
RenderX

differences in pedagogical outcomes between the two intervention models. The experimental and control groups included only 27 and 25 participants, respectively. Although the overall effect size was moderate (η^2=0.093), the achieved statistical power remained limited for detecting a medium-sized between-group effect. Larger multicenter studies are therefore needed to determine whether the trend toward higher scores in the BOPPPS teaching group represents a modest but genuine pedagogical advantage. Second, the DNCS is a subjective self-assessment instrument, and its sensitivity may be insufficient to capture changes in behavioral performance or subtle differences in skill mastery between two advanced teaching models. To address this measurement constraint, our protocol incorporated objective final examinations in addition to subjective self-assessment scales. Although the self-assessment scale did not reveal significant between-group differences, the objective final examination showed significantly higher scores in the BOPPPS teaching group than in the control group. This finding suggests that the BOPPPS model may be more effective in fostering classroom participation and deepening students' understanding of the course content. This finding is consistent with the results reported by Dai et al [23], confirming that the BOPPPS-based blended teaching model yields more significant educational outcomes in nursing courses compared to traditional teaching methods and effectively improves students' academic performance. Therefore, implementing specialized disaster nursing courses constitutes a pivotal professional empowerment initiative, equipping future frontline nurses with the essential technical capabilities required to respond to large-scale public health emergencies. This course not only fulfills nursing students' interest in learning about this specialized field but also significantly improves their integrated mastery of disaster theory and practice through systematic training. Research indicates that robust disaster competence and knowledge foster a heightened willingness to engage in relevant activities and assume corresponding responsibilities [4,24]. Consequently, the competency gains achieved through the BOPPPS-based blended teaching model not only fortify nursing students' professional qualifications but also instill the confidence necessary for active engagement and effective performance in real-world disaster response scenarios.

A notable finding requiring explanation is the improvement observed in the comparison group, which received no formal disaster nursing education; we offer several possible interpretations. First, an inherent selection bias and baseline imbalance heavily influenced student trajectories. Specifically, the comparison group demonstrated the highest baseline scores on the pretest (155 points) compared to the experimental (138 points) and control groups (148 points). This difference indicates that students who did not take the course already possessed a stronger foundation of existing knowledge, and their higher initial learning ability may have made it easier for them to achieve further self-directed improvement over time. Second, a test effect or learning effect may have emerged: participants completed the DNCS during the pretest, and this prior contact may have influenced the comparison group students' awareness of disaster nursing concepts, thereby prompting them to engage in informal, self-directed learning. Third, intergroup interference inevitably blurred the boundaries of the intervention. Contamination may have occurred through informal peer discussions, as all 3 groups were drawn from the same academic cohort and shared common spaces (eg, dormitories, clinical placement sites). Comparison group students may have discussed disaster nursing topics with peers in the instructed groups, partially attenuating the observed differences between groups. Fourth, after a full semester of study, the natural accumulation and maturation of students' overall clinical knowledge would also lead to a slight increase in their self-assessed professional competence. The study design did not include procedures to monitor and avoid such intergroup interference, which is a limitation of this study. Future research could address this issue by conducting cluster randomization at the institutional level or by staggering the study periods for different intervention groups. These considerations further highlight the necessity of using objective assessment metrics, such as final exams, in conjunction with self-assessment scales when evaluating teaching effectiveness.

Effects on Academic Achievement Outcomes

The results of this study indicate that the experimental group, which adopted the BOPPPS-based blended teaching model, significantly outperformed the control group in both process assessment and final exam scores. Figure 3 further demonstrates that the experimental group had a higher proportion of students scoring in the excellent (90-99) and good (80-89) ranges for their overall course score. In contrast, fewer students in the experimental group fell into the intermediate score range (70-79), while the number of nonpassing scores was comparable between both groups. The results of this study align with previous research, indicating that blended teaching models have more positive effects compared to single traditional teaching methods [11]. This is because the BOPPPS-based blended teaching model excels by dividing the teaching and learning process into six distinct components: Bridge-in, Objective, Preassessment, Participatory Learning, Postassessment, and Summary. In this study, these components were organized into 3 phases: before, during, and after the lesson. This structured approach provides a clear learning path, significantly enhances students' initiative, and helps them maintain better focus [25]. Studies have demonstrated that simulated mixed reality exercises enhance nursing students' knowledge and understanding of real disaster scenarios. Notably, participatory learning, which is an indispensable component of this process, plays a crucial role in developing students' practical problem-solving abilities [26,27]. Within the BOPPPS model, participatory learning effectively integrates theoretical knowledge and skills, promoting a student-centered approach to education [28].

Regarding process assessment, there was only a marginally significant difference between the 2 groups. This result demonstrates the rigorous process management employed in this study, with both groups maintaining consistency in behavioral compliance, including classroom attendance, completion of daily assignments, and participation in online learning. Therefore, the final examination score can serve as an objective core metric reflecting differences in students' mastery of knowledge under different instructional interventions. In terms of the achievement of instructional content and objectives, this study found that the BOPPPS intervention was effective in

XSL•FO
RenderX

helping students master basic disaster concepts, classifications, and rescue systems (Objective 1); nursing principles applied in common disaster scenarios such as fires and mudslides (Objective 2); the mastery of basic first aid and self-rescue skills, such as handling spinal fractures and treating burns (Objective 3); and postdisaster psychological disorders and the implementation of public health and epidemic prevention principles (Objective 5). However, the current assessment tools included relatively few items addressing Learning Objective 4, which focused on long-term postdisaster care and rehabilitation needs. As a result, the observed differences across question formats and total final examination scores may not be sufficient to determine whether the two groups differed in their deeper understanding of rehabilitation nursing. This limitation suggests that future course assessments should refine the distribution and weighting of items across learning objectives and include more diverse approaches, such as case-based essay questions, to ensure more comprehensive coverage of all intended learning outcomes. The results of this study show that the experimental group's scores were significantly higher than those of the control group. Within specific sections, the experimental group's advantage was primarily concentrated in the single-choice questions section, while no significant differences were observed in the true or false questions and the case-based essay questions. This pattern of an advantage that manifests selectively in single-choice items is consistent with the broader measurement evidence reported by Mee et al [29], who systematically analyzed question difficulty, discrimination, and response latency between objective choice questions and short-answer formats in medical student examinations. Their evaluation substantiated the conclusion that choice items can robustly segregate students based on their actual proficiency levels. The single-choice questions (totaling 50 points) in this exam are designed to comprehensively cover the course's learning objectives, with content extending to specific core topics such as the international triage standards, first aid principles for sudden drowning and food poisoning, and public health measures following earthquakes and floods. In the BOPPPS-based blended teaching, teachers facilitated diverse activities such as classroom presentations and group discussions. These interactive processes enabled students to exchange ideas and share opinions, fostering critical thinking and teamwork skills. By leveraging students' individual strengths, these activities enhanced the interactivity of the teaching process and established effective communication and feedback mechanisms between teachers and students, as well as among students. The integration of an immediate feedback mechanism into instruction effectively facilitates students' mastery of foundational knowledge across diverse domains and systematically promotes the construction of an integrated theoretical framework [30]. Consequently, students in the experimental group scored significantly higher on single-choice questions (P=.04), and this improvement contributes to the achievement of core instructional objectives. The true or false questions primarily assessed students' rapid recognition of appropriate emergency response principles and procedures in disaster scenarios, whereas the case-based essay question emphasized the integration of clinical presentations, triage priorities, emergency interventions, and postdisaster nursing needs. These findings suggest that the effect of short-term BOPPPS-based blended teaching on rapid situational judgment and higher-order case-based reasoning remains to be further examined. Compared with basic knowledge recognition, these abilities depend more heavily on situational experience, sequential decision-making practice, exposure to complex scenarios, and feedback on clinical decision-making. This interpretation is consistent with previous systematic review evidence indicating that simulation-based instruction can support the development of clinical reasoning in undergraduate nursing students [31]. Therefore, classroom-based structural optimization alone may not be sufficient to produce clear between-group differences within a short course period. Future course designs should consider extending instructional hours and incorporating more in-depth case discussions, scenario-based simulation training, and situational judgment tasks to better foster students' higher-order competencies. In summary, the BOPPPS-based blended teaching markedly promotes students' capacity to construct and retain core theoretical knowledge across multiple modules, thereby advancing the achievement of Objectives 1, 2, 3, and 5 through enhanced structured engagement. In medical education, the BOPPPS model has been effectively utilized across multiple disciplines, proving to be a highly effective teaching framework. It promotes sustainable development in students' future careers and aligns with the current trend in medical education, which emphasizes ability cultivation and practical application.

Limitations

This study has several limitations. First, the sample comprised only 77 students from a single institution, which may limit the generalizability of the findings. Second, a few baseline characteristics differed significantly between groups. Although the primary outcomes were analyzed using nonparametric tests without covariate adjustment, residual confounding cannot be fully excluded. Third, the study's one-semester timeframe did not permit the evaluation of long-term knowledge retention or the application of skills in professional settings. Fourth, our recruitment strategy introduced a self-selection bias because the disaster nursing curriculum was administered as an elective course. Fifth, the final examination was newly developed for this study and has not been validated; its content and construct validity remain to be established. Finally, despite efforts to maintain group separation, the quasi-experimental design within a single institution cannot fully eliminate the possibility of inter-group contamination. Future research should therefore prioritize multi-institutional, longitudinal investigations with larger sample sizes that account for baseline balancing and covariate adjustment, validate the examination through expert review and repeated administration, and integrate simulated disaster scenarios to evaluate the BOPPPS model's long-term impact on students' clinical competencies and professional growth.

Conclusions

This study compared the effectiveness of a blended teaching model based on the BOPPPS framework with traditional teaching methods in disaster nursing education for undergraduate students. It investigated the impact of the course curriculum on students' disaster nursing competency. The results indicated

XSL•FO
RenderX

that students taught using the BOPPPS-based blended teaching model demonstrated a significantly higher overall course score than those who received traditional lecture-based teaching. The disaster nursing curriculum was found to be effective in improving students' disaster response skills. As an innovative teaching model, BOPPPS effectively engages students' initiative, enhances their interest in learning, and develops essential professional skills and emergency response capabilities for disaster scenarios. Future studies could further validate the effectiveness and applicability of BOPPPS by integrating various teaching modes and utilizing advanced educational technologies. This would not only provide valuable insights for enhancing disaster nursing education at other Chinese universities but also establish a strong foundation for cultivating nursing professionals with advanced disaster management competencies.

Acknowledgments

Language-editing assistance was used during the preparation of this manuscript. Generative AI tools (Claude Opus 4.7, Anthropic; Kimi k1.5, Moonshot AI) were consulted solely to improve grammar, clarity, and sentence structure, as English is not the authors' first language. The tools were used exclusively for linguistic refinement and reference formatting support. All scientific concepts, interpretations, arguments, and conclusions presented in this manuscript were independently developed by the authors. The authors take full responsibility for the intellectual content, accuracy, and integrity of the work.

Funding

This research was supported by the Tianjin University of Traditional Chinese Medicine 2023 Education and Teaching Reform Research Project (2023YJY016) and the Tianjin Higher Education Bachelor's Teaching Quality and Teaching Reform Research Program (A251006304).

Authors' Contributions

LP contributed to conceptualization, investigation, resources, methodology, data curation, project administration, and manuscript review and editing. TJ analyzed the data, interpreted the statistical results, wrote the initial manuscript draft, and revised the manuscript. YL contributed to statistical analysis, interpretation of statistical results, and visualization, and wrote the initial manuscript. XLP contributed to writing-review and editing, validation, funding acquisition, and resources. ML contributed to data curation. HD provided supervision, project administration, and funding acquisition. All authors reviewed and approved the final manuscript.

Conflicts of Interest

None declared.

References

1. Stimpson JP, Rashed AL, Pandya J, Baudot EC, Whitfill J, Ortega AN. Health equity in the wake of disasters and extreme weather: evidence from an umbrella review. Health Aff Sch 2025 Nov;3(11):qxaf207. [doi: 10.1093/haschl/qxaf207] [Medline: 41229451]
2. Han W, Liang C, Jiang B, Ma W, Zhang Y. Major natural disasters in China, 1985-2014: occurrence and damages. Int J Environ Res Public Health 2016 Nov 10;13(11):1118. [doi: 10.3390/ijerph13111118] [Medline: 27834899]
3. Zhang D, Zhang L, Gong A. Development of disaster nursing in China: from the spirit of Nightingale to COVID-19. Disaster Med Public Health Prep 2021 Apr;15(2):e32-e35. [doi: 10.1017/dmp.2020.468] [Medline: 33357254]
4. Taskiran G, Baykal U. Nurses' disaster preparedness and core competencies in Turkey: a descriptive correlational design. Int Nurs Rev 2019 Jun;66(2):165-175. [doi: 10.1111/inr.12501] [Medline: 30734270]
5. van Esch A, Daehnert E, Bramer WM, Gommers D, van Mol MMC. Disaster preparedness and needed competencies among critical care nurses: a scoping review. BMC Nurs 2026 Jan 9;25(1):127. [doi: 10.1186/s12912-025-04275-3] [Medline: 41507847]
6. Nekooei Moghaddam M, Saeed S, Khanjani N, Arab M. Nurses' requirements for relief and casualty support in disasters: a qualitative study. Nurs Midwifery Stud 2014 Apr;3(1):e9939. [doi: 10.17795/nmsjournal9939] [Medline: 25414897]
7. Tayebi Z, Norouzinia R, Moatadelro Z, Pour AF, Nourian B. Nursing students' willingness to respond in disasters: a cross sectional study of facilitators and barriers. BMC Nurs 2024 Jun 20;23(1):416. [doi: 10.1186/s12912-024-02088-4] [Medline: 38902686]
8. Zhou M, Pei L. Research on status and influencing factors of disaster literacy among medical students based on mixed methods [Article in Chinese]. Health Vocational Education 2025;43(23):132-137. [doi: 10.20037/j.issn.1671-1246.2025.23.32]
9. Zhang YY, Zhu LL, Sheng Y, Li XH, Xu XH, Wang QY. Disaster nursing development in China and other countries: a bibliometric study. J Nurs Scholarsh 2018 Sep;50(5):567-576. [doi: 10.1111/jnu.12401] [Medline: 29998630]

XSL•FO
RenderX

10. Ma X, Zeng D, Wang J, Xu K, Li L. Effectiveness of bridge-in, objective, pre-assessment, participatory learning, post-assessment, and summary teaching strategy in Chinese medical education: a systematic review and meta-analysis. Front Med (Lausanne) 2022;9:975229. [doi: 10.3389/fmed.2022.975229] [Medline: 36186766]
11. Hu K, Ma RJ, Ma C, Zheng QK, Sun ZG. Comparison of the BOPPPS model and traditional instructional approaches in thoracic surgery education. BMC Med Educ 2022 Jun 9;22(1):447. [doi: 10.1186/s12909-022-03526-0] [Medline: 35681190]
12. Kılıç N, Şimşek N. The effects of psychological first aid training on disaster preparedness perception and self-efficacy. Nurse Educ Today 2019 Dec;83:104203. [doi: 10.1016/j.nedt.2019.104203] [Medline: 31683114]
13. Li Z, Cai X, Zhou K, et al. Effects of BOPPPS combined with TBL in surgical nursing for nursing undergraduates: a mixed-method study. BMC Nurs 2023 Apr 23;22(1):133. [doi: 10.1186/s12912-023-01281-1] [Medline: 37088853]
14. Core competencies in disaster nursing: version 2.0. : International Council of Nurses; 2019 URL: https://www.icn.ch/resources/publications-and-reports/core-competencies-disaster-nursing-version-20 [accessed 2026-04-07]
15. Hu X. Research on competency-based disaster nursing training curriculum setting [Article in Chinese] [Master's thesis]. : Huazhong University of Science and Technology; 2020. [doi: 10.27157/d.cnki.ghzku.2019.001828]
16. Cohen J. Statistical Power Analysis for the Behavioral Sciences, 2nd edition: Routledge; 1988. [doi: 10.4324/9780203771587]
17. Fiel Peres F. Effect sizes for nonparametric tests. Biochem Med (Zagreb) 2026 Feb 15;36(1):010101. [doi: 10.11613/BM.2026.010101] [Medline: 41399660]
18. Tussing TE, Chesnick H, Jackson A. Disaster preparedness: keeping nursing staff and students at the ready. Nurs Clin North Am 2022 Dec;57(4):599-611. [doi: 10.1016/j.cnur.2022.06.008] [Medline: 36280298]
19. Shujuan L, Mawpin T, Meichan C, Weijun X, Jing W, Biru L. The use of virtual reality to improve disaster preparedness among nursing students: a randomized study. J Nurs Educ 2022 Feb;61(2):93-96. [doi: 10.3928/01484834-20211213-05] [Medline: 35112954]
20. Phan Q, Geller DE, Broughton AS, Swan BA, Wells JS. Evaluating a low-cost disaster preparedness simulation for prelicensure nursing students. Disaster Med Public Health Prep 2023 Mar 1;17:e343. [doi: 10.1017/dmp.2022.280] [Medline: 36855255]
21. Glauberman GHR, Wong LCK, Bray ML, Katz AR. Disaster aftermath interprofessional simulation: promoting nursing students' preparedness for interprofessional teamwork. J Nurs Educ 2020 Jun 1;59(6):353-356. [doi: 10.3928/01484834-20200520-11] [Medline: 32497240]
22. Cho M, Kwon M. Sustainable disaster nursing education through functional exercises and simulation: effects on knowledge, problem-solving, and learning outcomes. Sustainability 2025;17(20):9165. [doi: 10.3390/su17209165]
23. Dai Y, He Q, Lei L. Effect of the BOPPPS-based blended teaching model in undergraduate nursing education: a quasi-experimental study. Sci Rep 2026 Feb 12;16(1):8580. [doi: 10.1038/s41598-026-37120-x] [Medline: 41680311]
24. Deci EL, Ryan RM. The "what" and "why" of goal pursuits: human needs and the self-determination of behavior. Psychol Inq 2000;11(4):227-268. [doi: 10.1207/S15327965PLI1104_01]
25. Liu XY, Lu C, Zhu H, et al. Assessment of the effectiveness of BOPPPS-based hybrid teaching model in physiology education. BMC Med Educ 2022 Mar 30;22(1):217. [doi: 10.1186/s12909-022-03269-y] [Medline: 35354465]
26. Alberti S, Motta P, Ferri P, Bonetti L. The effectiveness of team-based learning in nursing education: a systematic review. Nurse Educ Today 2021 Feb;97:104721. [doi: 10.1016/j.nedt.2020.104721] [Medline: 33385942]
27. Labrague LJ, Hammad K, Gloe DS, et al. Disaster preparedness among nurses: a systematic review of literature. Int Nurs Rev 2018 Mar;65(1):41-53. [doi: 10.1111/inr.12369] [Medline: 28295314]
28. Chen L, Tang XJ, Chen XK, Ke N, Liu Q. Effect of the BOPPPS model combined with case-based learning versus lecture-based learning on ophthalmology education for five-year paediatric undergraduates in Southwest China. BMC Med Educ 2022 Jun 7;22(1):437. [doi: 10.1186/s12909-022-03514-4] [Medline: 35668389]
29. Mee J, Pandian R, Wolczynski J, et al. An experimental comparison of multiple-choice and short-answer questions on a high-stakes test for medical students. Adv Health Sci Educ Theory Pract 2024 Jul;29(3):783-801. [doi: 10.1007/s10459-023-10266-3] [Medline: 37665413]
30. Thijssen DHJ, Hopman MTE, van Wijngaarden MT, Hoenderop JGJ, Bindels RJM, Eijsvogels TMH. The impact of feedback during formative testing on study behaviour and performance of (bio)medical students: a randomised controlled study. BMC Med Educ 2019 Apr 3;19(1):97. [doi: 10.1186/s12909-019-1534-x] [Medline: 30943962]
31. Theobald KA, Tutticci N, Ramsbotham J, Johnston S. Effectiveness of using simulation in the development of clinical reasoning in undergraduate nursing students: a systematic review. Nurse Educ Pract 2021 Nov;57:103220. [doi: 10.1016/j.nepr.2021.103220] [Medline: 34781195]

Abbreviations

BOPPPS: Bridge-in, Objective, Preassessment, Participatory Learning, Postassessment, and Summary
DNCS: Disaster Nursing Competency Scale
ICN: International Council of Nurses
WHO: World Health Organization

XSL•FO
RenderX

Edited by A Bahattab, B Lesselroth; submitted 19.May.2025; peer-reviewed by B Dovdon, K Kruger, N Simsek; revised version received 22.Aug.2026; accepted 02.Sep.2026; published 05.Oct.2026.

Please cite as:
Pei L, Jia T, Liu Y, Pang XL, Li M, Dou H
Evaluation of the Bridge-in, Objective, Preassessment, Participatory Learning, Postassessment, and Summary–Based Instructional Model in Disaster Nursing: Quasi-Experimental Study
JMIR Med Educ 2026;12:e77728
URL: https://mededu.jmir.org/2026/1/e77728
doi:10.2196/77728

© Li Pei, Taoyu Jia, Yaxin Liu, Xiao-li Pang, Muzi Li, Haoying Dou. Originally published in JMIR Medical Education (https://mededu.jmir.org), 5.Oct.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Education, is properly cited. The complete bibliographic information, a link to the original publication on https://mededu.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

The Effect of a Traditional Chinese Medicine Course on Western Medicine Students' Attitudes Toward Traditional Chinese Medicine: Self-Controlled Pre-Post Questionnaire Study

Haoyu He[1,2,3], MD; Hanjin Cui[4], MD; Li Guo[5]; Yanhui Liao[6], MD

[1]Department of Psychology, College of Education, Hunan First Normal University, Changsha, Hunan, China

[2]Research Center for Innovative Development of Teacher Education in the New Era, Changsha, China

[3]Hunan Key Laboratory for Children's Psychological Development and Brain & Cognition Science, Changsha, China

[4]Institute of Integrative Medicine, Xiangya Hospital, Central South University, Changsha, Hunan, China

[5]Department of Educational Administration, The Second People's Hospital of Hunan Province (Brain Hospital of Hunan Province), Changsha, China

[6]Department of Psychiatry, Sir Run Run Shaw Hospital, Zhejiang University School of Medicine, 3 East Qingchun Road, Hangzhou, Zhejiang, China

Corresponding Author:
Yanhui Liao, MD
Department of Psychiatry, Sir Run Run Shaw Hospital, Zhejiang University School of Medicine, 3 East Qingchun Road, Hangzhou, Zhejiang, China

Abstract

Background: Traditional Chinese medicine (TCM) hasbeen widely used to treat various diseases in China for thousands of years and has shown satisfactory effectiveness. However, many surveys found that TCM receives little recognition from Western medicine (WM) physicians and students. At present, TCM is offered as a compulsory course for WM students in WM schools.

Objective: This study aimed to investigate whether TCM courses can affect the WM students' attitude toward TCM.

Methods: WM students from Xiangya Medical School were invited to completeaweb-based questionnaire before and immediately after a TCM course. Their attitude toward TCM and treatment preferences for different kinds of diseases were tested. The Attitude Scale of TCM (ASTCM) was used. The main part of the ASTCM was designed to measure the attitude of medical students towardTCM. It consisted of 18 items, divided into cognitive dimension (5 terms), emotional dimension (8 terms), and behavioral tendencyfactor (5 terms).

Results: Finally, the results of 118 five-year program (FYP) and 36 eight-year program (EYP) students were included. For FYP students, there was a significant increase in the total mean score (66.42, SD 7.66 vs 71.43, SD 7.38;*P*<.001) of ASTCM after the TCM course. Significant increases in mean scores of the 3 factors of attitude (cognition: 21.64, SD 2.08 vs 22.90, SD 1.94; affection: 25.21, SD 4.39 vs 27.96, SD 4.4; and behavioral tendency: 19.577, SD 3.02 vs 20.58, SD 2.76; *P*<.001)were also observed. Except for the score of behavioral tendency (17.50, SD 3.54 vs 18.78, SD 3.22; *P*=.02), a significant increase was not detected in total score, cognition, and affection in EPY students (total score: mean 60.36, SD 10.53 vs mean 62.92, SD 10.05; cognition: mean 20.50, SD 2.73 vs mean 20.69, SD 2.73; and affection: mean 22.36, SD 6.32 vs mean 23.44, SD 5.84; all *P*>.05). The treatment preference of FYP students in acute (*P*=.02), chronic (*P*=.003), and physical diseases (*P*=.02) showed remarkable change. A major change was also detected in internal diseases (*P*=.02), surgical diseases (perioperative period; *P*=.01), and mental illnesses (*P*=.02) in EYP students. This change mainly appeared as a decline in WM preference and an increase in TCM and WM preference.

Conclusions: The study showed that earlier exposure to the TCM course increased the positive attitude toward TCM in students majoring in WM. The results provide some suggestions for arraging TCM courses in WM schools.

(JMIR Med Educ 2026;12:e55972) doi:10.2196/55972

KEYWORDS

traditional Chinese medicine; TCM; Chinese medicine; natural medicine; naturopathy; naturopathic; Western medicine students; Western medicine; modern medicine; medical students; attitude

Introduction

Traditional Chinese medicine (TCM) has been widely used to treat various diseases in China for more than 2500 years, and it has become an integral part of Chinese culture [1]. On the basis of its holistic theory and rich practical experiences, TCM offers a diversity of therapeutic strategies for complex diseases to balance the Yin and Yang disorders of the human body [2,3].

XSL•FO
RenderX

However, Western medicine (WM) has become the mainstream medicine in modern China [4]. The number of TCM hospitals is only approximately one-fifth of the number of WM hospitals, with the number of TCM practitioners less than a quarter of that of WM physicians [5]. Medical colleges mainly focus on WM teaching, except for several colleges of TCM [6]. The curricula that WM students learn include basic and clinical medical courses, which are similar to European and American programs. Basic medical courses include human anatomy, embryology, biochemistry, physiology, pathology, pharmacology, and so on. These courses provide a theoretical foundation for clinical medicine. Clinical medicine courses focus more on practice, including internal medicine, surgery, obstetrics and gynecology, pediatrics, neurology, infectious diseases, and so on. These courses help studentsmaster the diagnostic and treatment skills needed for managing diseases. Thus, TCM receives less recognition from WM physicians and students [4]. A study among Hong Kong WM physicians showed that only 11.9% and 9.4% of them had ever considered or referred their patients to TCM, respectively [7]. Another survey found that medical students appeared to become more negative toward TCM after learning WM [8]. In addition, the positive attitude held by WM students toward TCM was less than half reported by some other surveys [8-10]. The present situation due to the negative attitude of WM students toward TCM greatly hinders the inheritance, development, and innovation of TCM in China [11]. Attitude refers to a constant psychological inclination that individuals hold toward other individuals, ideas, emotions, or events. It encompasses a subjective evaluation and influences the behavior of individuals [12]. Cognition, affection, and behavioral tendency are the 3 critical components of attitude. Among them, the cognitive component is the foundation of the other components of attitude. The emotional component is the core and key of attitude, which affects both cognitive and behavioral components. Behavioral tendency components can affect people's future reactions to attitude objects. The 3 components are interrelated and mutually restrictive [13]. Consequently, cultivating a suitable attitude toward TCM will serve to foster a precise perspective on TCM for WM students.

TCM courses in WM schools are a direct way to rectify the attitude. A survey conducted at the Capital Medical University suggested that after finishing the TCM course, more than 65% and 71.4% of WM students approved the safety and efficacy of TCM, respectively; 94.3% of students believed that TCM was worth learning; and 82.9% of students found that TCM was helpful for clinical practice [14]. It is believed that the course lets WM students understand the basic ideas, theories, and skills of TCM. Meanwhile, the course helps them understand the differences between TCM and WM, broadens their mind, and enriches the strategies of diagnosis and treatment. Students should possess a profound understanding of both TCM and WM and offer better medical service for patients in their future clinical practice [14,15].

Xiangya School of Medicine was founded in 1914 by the Hunan Society of Education and the Yale Association. It was the first WM school in China. At present, at Xiangya School of Medicine of Central South University, TCM is offered as a compulsory course for grade 3 WM students in the 5-year program (FYP) and grade 5 in the 8-year program (EYP). It was the first time for both of them to access the TCM course. TCM is, however, less recognized by WM physicians and students, posing challenges for its integration into modern medical education as a compulsory course. This study used a self-controlled design with online questionnaires administered before and after the TCM course to WM students. The aim of the study was to explore whether TCM courses can exert an impact on the attitude of WM students toward TCM. The a priori hypothesis is that the TCM course will significantly improve WM students' attitudes toward TCM, particularly among FYP students, because of their relatively less established WM knowledge base compared to EYP students. The results are most likely to be used by medical educators, medical education researchers, WM students, future clinicians, and health care policy makers.

Methods

Description of TCM Course

According to the teaching arrangement of the dean's office at Xiangya Hospital, the TCM course for FYP and EYP students was taught by the same teaching group in the second semester of 2022—from April 15, 2022, to June 10, 2022, once a week for students in FYP, and from June 21, 2022, to August 15, 2022, once a week for students in EYP. Each theoretical part (20 class h in total) was followed by an internship (20 class h in total). The course includes the basic theory of TCM, diagnostics of TCM, Chinese pharmacy, Chinese medical formulas, acupuncture, and moxibustion. All the students were required to study *Approaching TCM*, which is a synchronous online course recorded by teachers from the Department of TCM in Xiangya Hospital. In addition, they were required to finish the self-examination after each chapter. The TCM course was delivered by teachers from the department of TCM in Xiangya Hospital. All the teachers are clinical physicians with medium or senior titles, holding qualification certificates for physicians and teacher qualifications from colleges and universities. *Traditional Chinese Medicine, Ninth Edition*, published by the People's Health Publishing House, was used as the textbook. According to the syllabus, the basic theory of TCM, diagnostics of TCM, traditional Chinese pharmacology, science of prescriptions, and acupuncture and moxibustion were taught.

Participants and Procedure

The participants comprised 145 grade 3 students in an FYP and 37 grade 5 students in an EYP. The students were asked to complete the same questionnaire before and after the TCM course online.

Ethical Considerations

This study was approved by the Ethics Committee of Hunan First Normal University (202202; Multimedia Appendix 1). Informed consent was obtained from all participants by choosing "yes" at the beginning of the online questionnaire. The information provided by participants will not be disclosed to any third party and will be used solely for the purposes of this research. All participant identifiers have been anonymized to ensure confidentiality. As the intervention involved regular

XSL•FO
RenderX

academic coursework during a natural semester and posed no risk of harm, no compensation was provided to participants on ethical grounds.

Measures

The questionnaire comprised four parts. The first part was the informed consent, including the purpose and risk of the study, the benefits of participation, and confidentiality. Part two was the basic demographic information(eg, name, sex, age, grade, and educational program). Part three was the Attitude Scale of TCM (ASTCM) developed by Wu [16]. This scale consists of 18 items divided into cognitive dimension (5 terms), emotional dimension (8 terms), and behavioral tendency factor (5 terms). The reliability test showed that internal consistency coefficients (ranging from 0.78 to 0.85), split-half reliability (ranging from 0.64 to 0.84), and retest reliability (ranging from 0.62 to 0.85) were significant. The validity test showed that the explanatory rate of the 3 factors from exploratory factor analysis was 53.58%. Confirmatory factor analysis showed that the root mean square error of approximation, comparative fit index, and non-normed fit index were 0.067, 0.903, and 0.995, respectively. The ASTCM was designed to measure the attitude of medical students toward TCM, and the reliability and validity indicators meet the measurement requirements. Physical illness is the opposite of “mental illness.” There are pathological changes in the structure of body tissues (organs) or disruptions in physiological functions. Part 4 was set to assess treatmentpreferences for different kinds of diseases [17], for example, whether participants would choose TCM, WM, or a combination of the two to treat functional disorders.

Statistical Analysis

Analyses were carried out using SPSS (version 23; IBM Corp). For data with homogeneous variance, the intergroup differences in attitude scores (grouped by demographic information) were analyzed using an independent sample 2-tailed *t* test, and intragroup differences (self-controlled) were analyzed using a paired 2-tailed *t* test. For data with heterogeneous variance, the intergroup differences in attitude scores (grouped by demographic information) were analyzed using the Mann-Whitney *U* test, and the intragroup differences (self-controlled) were analyzed using the Wilcoxon *W* test.

Paired sample 2-tailed *t* tests were used to analyze changes in the total, cognition, affection, and behavioral tendency scores between the pretest and posttest. The McNemar test was used for analyzing changes in the selection preference of different therapeutic strategies. Cohen *d* was used to measure the effect sizes of changes from pre- to postcourse using the following formula: d=tmatchedn211-ρ,where *n* is the paired sample size and ρ refers to the correlation coefficient. Cohen suggested that *d*=0.2 be considered a “small” effect size, 0.5 represents a “medium” effect size, and 0.8 represents a “large” effect size [18].

Results

The data were collected from April 15, 2022, to August 15, 2022. In total, 118 FYP and 36 EYP students completed the before and after tests, and the data analysis was completed.

Basic Demographic Information of Participants

There were 118 and 36 questionnaires returned from 145 students in the FYP and 37 students in the EYP, respectively. Among the returned questionnaires, 118 (FYP) and 36 (EYP) were valid with before and after tests completed. Finally, 118 (FYP) and 36 (EYP) questionnaires were included for statistical analysis.

The basic demographic information of participants is presented in Table 1. The sample was 39.8% (n=47) male and 60.2% (n=71) female participants in FYP, and 38.9% (n=14) male and 61.1% (n=22) female participants in EYP(question 6 of the questionnaire). The average age was 20.64 (SD 0.70) years and 22.56 (SD 0.56) years for FYP and EYP, respectively. Of 118 FYP students, 18 (15.3%) reported that they had relatives engaged in TCM clinical or scientific research (question 9), whereas this percentage was 16.7% (6/36) among EYP students. Moreover, 74.6% (88/118) of FYP students and 66.7% (24/136) of EYP students reported that they had received TCM treatment (question 10). In addition, 64.4% (76/118) of FYP students and 52.8% (19/36) of EYP students reported that there were celebrities in the field of TCM whom they admired (question 11; Table 1).

XSL•FO
RenderX

Table . Basic demographic information of participants.

Characteristics	FYP[a] students (n=118)	EYP[b] students (n=36)
Sex, n (%)		
Male	47 (39.8)	14 (38.9)
Female	71 (60.2)	22 (61.1)
Age (y), mean (SD)	20.64 (0.70)	22.56 (0.56)
Students whose relatives engaged in TCM[c] clinical or scientific research, n (%)	18 (15.3)	6 (16.7)
Students who had received TCM treatment, n (%)	88 (74.6)	24 (66.7)
Students who admired celebrities in the field of TCM, n (%)	76 (64.4)	19 (52.8)

[a]FYP: 5-year program.

[b]EYP: 8-year program.

[c]TCM: traditional Chinese medicine.

There was no significant difference in the attitude pretest scores of FYP students on questions 5, 9, and 10, but there was a significant difference on question 11. On question 11, the total attitude score and 3-factor scores of students who answered yes were significantly higher than those who answered no (Table 2). We did not find significant differences in the pretest scores and 3-factor scores of EYP students in terms of demographic factors.

Table . The difference in the attitude pretest scores of 5-year program students (n=118) on question 11.

Score	Do you have any celebrities in the field of TCM[a] whom you admire?		*t* test[b] (*df*)	*P* value	Cohen *d*
	Yes (n=76), mean (SD)	No (n=42), mean (SD)			
Total score	68.26 (7.15)	63.07 (7.50)	3.711 (116)	<.001	0.71
Cognition score	21.97 (2.03)	21.02 (2.07)	2.421 (116)	.02	0.46
Affection score	25.95 (4.19)	23.88 (4.46)	2.506 (116)	.01	0.48
Behavior tendency score	20.34 (3.04)	18.17 (2.76)	3.848 (116)	<.001	0.74

[a]TCM: traditional Chinese medicine.

[b]Independent sample 2-tailed *t* test.

The Effect of TCM Course on Attitude Toward TCM

For FYP students, there was a significant increase in the total mean score of ASTCM after the TCM course (from 66.42, SD 7.66 to 71.43, SD 7.38; *P*<.001). A significant increase was also observed in the mean scores of the 3 factors: cognition (from 21.64, SD 2.08 to 22.90, SD 1.94; *P*<.001), affection (from 25.21, SD 4.39to 27.96, SD 4.4; *P*<.001), and behavioral tendency (from 19.57, SD 3.02 to 20.58, SD 2.76; *P*<.001). Interestingly, except for the behavioral tendency mean score (from 17.50, SD 3.54 to 18.78, SD 3.22; *P*=.02), a significant increase was not detected in the total mean score (from 60.36, SD 10.53 to 62.92, SD 10.05; *P*=.19), cognition (from 20.50, SD 2.73 to 20.69, SD 2.73; *P*=.69), and affection of ASTCM (from 22.36, SD 6.32 to 23.44, SD 5.84; *P*=.34) in EPY students (Table 3). The effect sizes (Cohen *d*) are presented in Table 3.

XSL•FO
RenderX

Table . Changes in attitude scores of Western medicine students.

Score	FYP[a] (n=118)					EYP[b] (n=36)				
	Precourse, mean (SD)	Postcourse, mean (SD)	*t* test[c] (*df*)	*P* value	Cohen *d*	Precourse, mean (SD)	Postcourse, mean (SD)	*t* test (*df*)	*P* value	Cohen *d*
Total score	66.42 (7.66)	71.43 (7.38)	−7.206 (117)	<.001	−0.67	60.36 (10.53)	62.92 (10.05)	−1.353 (35)	.19	−0.25
Cognition score	21.64 (2.08)	22.90 (1.94)	−6.276 (117)	<.001	−0.63	20.50 (2.73)	20.69 (2.73)	−0.397 (35)	.69	−0.07
Affection score	25.21 (4.39)	27.96 (4.45)	−6.912 (117)	<.001	−0.6	22.36 (6.32)	23.44 (5.84)	−0.976 (35)	.34	−0.18
Behavioral tendency score	19.57 (3.02)	20.58 (2.76)	−3.892 (117)	<.001	−0.34	17.50 (3.54)	18.78 (3.22)	−2.354 (35)	.02	−0.38

[a]FYP: 5-year program.

[b]EYP: 8-year program.

[c]Paired sample 2-tailed *t* test.

In addition, we found that the attitude scores of FYP students significantly increased after the TCM course, regardless of whether they admired a Chinese medicine celebrity. However, no significant difference was observed between students who admired a TCM celebrity and those who did not in the posttest attitude scores(Table 4).

Table . The difference in the attitude scores between pretest and posttest of 5-year program students on question 11 (n=118).

Score	Do you have any celebrities in the field of TCM[a] whom you admire?									
	Yes (n=76)					No (n=42)				
	Precourse, mean (SD)	Postcourse, mean (SD)[b]	*t* test[c] (*df*)	*P* value	Cohen *d*	Precourse, mean (SD)	Postcourse, mean (SD)[b]	*t* test (*df*)	*P* value	Cohen *d*
Total score	68.26 (7.15)	72.54 (5.13)	−5.803 (75)	<.001	−0.60	63.07 (7.50)	69.63 (10.04)	−4.472 (41)	<.001	−0.70
Cognition score	21.97 (2.03)	23.14 (1.46)	−5.221 (75)	<.001	−0.58	21.02 (2.07)	22.45 (2.55)	−3.606 (41)	.001	−0.61
Affection score	25.95 (4.19)	28.42 (3.70)	−5.608 (75)	<.001	−0.55	23.88 (4.46)	27.12 (5.52)	−4.140 (41)	<.001	−0.64
Behavioral tendency score	20.34 (3.04)	20.97 (2.13)	−2.120 (75)	.04	−0.22	18.17 (2.76)	19.86 (3.55)	−3.548 (41)	.001	-0.53

[a]TCM: traditional Chinese medicine.

[b]Due to the heterogeneity of variance, we compared the posttest attitude scores between groups using the Mann-Whitney *U* test, and the result showed no significant difference.

[c]Paired sample 2-tailed *t* test.

The Effect of TCM Course on Selection Preference of Different Therapeutic Strategies

For FYP students, there was a significant difference in the selection preferences of different therapeutic strategies for acute (*P*=.025), chronic (*P*=.02), and physical diseases (*P*=.004). After the TCM course, the percentage of students who preferred WM for acute diseases decreased from 94.1% (n=111) to 83.9% (n=99). The preference for the combination of TCM and WM (TCM and WM) increased from 5.1% (n=6) to 14.4% (n=17), and the preference for TCM increased from 0.8% (n=1) to 1.7% (n=2). For chronic diseases, fewer students preferred WM (from n=10, 8.5% to n=1, 0.8%), whereas more students preferred TCM (from n=53, 44.9% to n=72, 61%). As for physical diseases, the percentage of students preferring WM declined (from n=56, 47.5% to n=39, 33.1%), whereas the preference for TCM and WM increased (from n=52, 44.1% to n=78, 66.1%; Table 5).

For EYP students, the significant difference in the selection preferences for different therapeutic strategies mainly focused on internal diseases (*P*=.02), surgical diseases in the perioperative period (*P*=.01), and mental illnesses (*P*=.02). More students realized that TCM and WM was a better choice (from n=12, 33.3%-52.8%) for internal diseases. The percentage of students who preferred TCM and WM for perioperative surgical diseases nearly tripled from 8.3% (n=3) to 25% (n=9). There was also an approximately 20% point increase in WM and TCM

XSL•FO
RenderX

preference for mental illnesses (from n=51, 41.7% to n=22, 61.1%) with a marked decline in WM (n=19, 52.8% to n=13, 36.1%; Table 3).

Table . Changein treatment preference of Western medicine (WM) students (n=118).

Program, disease, treatment preference, and data type	Before course, n (%)	After course, n (%)	*P* value[a]
FYP[b]			
Acute diseases			.02
TCM[c]	1 (0.8)	2 (1.7)	
WM	111 (94.1)	99 (83.9)	
TCM and WM	6 (5.1)	17 (14.4)	
Chronic diseases			.003
TCM	53 (44.9)	72 (61)	
WM	10 (8.5)	1 (0.8)	
TCM and WM	55 (46.6)	45 (38.1)	
Somatic disease			.02
TCM	10 (8.5)	7 (5.9)	
WM	56 (47.5)	39 (33.1)	
TCM and WM	52 (44.1)	78 (61)	
EYP[d]			
Internal diseases			.02
TCM	4 (11.1)	2 (5.6)	
WM	20 (55.6)	15 (41.7)	
TCM and WM	12 (33.3)	19 (52.8)	
Surgical diseases (perioperative period)			.01
TCM	1 (2.8)	1 (2.8)	
WM	32 (88.9)	26 (72.2)	
TCM and WM	3 (8.3)	9 (25)	
Mental illnesses			.02
TCM	2 (5.6)	1 (2.8)	
WM	19 (52.8)	13 (36.1)	
TCM and WM	15 (41.7)	22 (61.1)	

[a]McNemar test.

[b]FYP: 5-year program.

[c]TCM: traditional Chinese medicine.

[d]EYP: 8-year program.

Discussion

Principal Findings

This study was conducted in natural classes, based on the real teaching arrangements of the dean's office at Xiangya Hospital. The study examined the effect of TCM courses on the attitude toward TCM in students majoring in WM using the ASTCM. The study found that the course significantly increased the total score and the scores of the 3 factors of ASTCM in FYP students but did not change these scores in EYP students, except for the score of the behavioral factor. The study also found a significant change in therapeutic strategy preferences before and after TCM learning, with the percentage reduction in WM and increase in TCM or TCM and WM. The alteration in therapeutic approach inclination primarily manifested in acute, chronic, and somatic ailments among FYP scholars and in internal, perioperative surgical, and psychological disorders in EYP learners. The results suggested that TCM courses increased a positive attitude toward TCM for WM students.

Basic Demographic Information of Participants

The demographic results showed that there were more female students in both FYP and EYP groups. The result was similar

XSL•FO
RenderX

to the gender ratio officially reported by Central South University. The age of EYP students was 2 years older than that of FYP students. This was in line with the reality. An interesting result was that FYP students who admired Chinese medicine celebrities had a significantly better attitude toward TCM than those who did not admire Chinese medicine celebrities. This, to some extent, indicates that celebrity worship had a positive impact on attitudes toward TCM [19].

The Effect of TCM Course on Cognitive Component of Attitude in FYP

This study reported a more positive attitude toward TCM in FYP students after completing the TCM course. The results were supported by another study carried out at the Capital Medical University, which found that WM students became more positive toward TCM after learning about it [20]. Attitude is a disposition toward or against a specified phenomenon, person, or thing. An attitude comprises cognitive, affective, and behavioral tendency components [21]. It is believed that attitude canbe changed by interventions or messages delivered at a particular time from the perspective of cognition, affection, and behavioral tendency [22].

TCM is generally considered to be mysterious, profound, and lacking a scientific basis [23]. By learning the basic theory of TCM, students gained a preliminary understanding of the basic theoretical framework of TCM. They learned that TCM is not mystical but has a solid foundation to support. They also learned how TCM physicians diagnose diseases with their unique syndrome differentiation system by learning the diagnostics of TCM. They witnessed how to use so-called “flowers and plants” (herbs) to treat diseases by learning Chinese pharmacy and medical formulas. In the acupuncture and moxibustion courses, the mysterious meridians were demystified and presented to students. All the abovementioned courses are believed to improve the students’ attitude toward TCM from the perspective of cognition.

The Effect of TCM Course on Affective Component of Attitude in FYP

Improvement in the affection component of attitude was also observed after the TCM course. It is a form of identification and resonance. First, TCM was acknowledged by students rationally with correct cognition via theoretical lectures. Second, identification and resonance of students were achieved through internships, which are considered the most crucial experiences for undergraduate medical students [24]. The students witnessed the remarkable therapeutic effect of TCM. The cases showed that 2 doses of medicine cured a 3-month persistent cough, and immediate acupuncture for acute knee injuries were often shown. Some students even experienced TCM for dysmenorrhea: the entire diagnosis and treatment process, from pulse feeling to prescription. Finally, during the study, many famous TCM physicians were known to students and became their admired seniors. It greatly increased the affection component of attitude, as shown in the result of question 11.

The Effect of TCM Course on Behavioral Tendency Component of Attitude in FYP

The behavioral tendency component of attitude also improved after the TCM course among FYP students. The clinical internship after the theoretical lecture was also conducted to change the behavioral component of attitude. The internship aimed at integrating TCM theory with practice. In this way, abstract TCM theories and methods would be specified [25]. For example, students visited patients with different syndromes, and they were taught the yin and yang attributes of syndromes according to the Yin-Yang theory. The internship also showed the students a complete diagnosis and treatment process of TCM, which included 4 diagnostic methods (ie, observation, auscultation and olfaction, inquiry, and pulse feeling and palpation), syndrome differentiation, and treatment as well [26]. During the internship, students experienced some unique diagnostic methods of TCM, such as tongue diagnosis, pulse feeling, and palpation. Students even tried the needling sensation (Deqi) themselves in the internship. Hence, the clinical internship afforded the students a hands-on and introspective encounter with TCM, rather than mere hearsay and boastful claims. It was of great importance for students to learn about the unique advantages of TCM, to eliminate prejudices about TCM, and, finally, to make a more objective and comprehensive judgment about TCM.

The Effect of Chinese Medicine Celebrity and TCM Course on Attitude in FYP

There was also an interesting result that before the course, students who admired Chinese medicine celebrity worship had a more positive attitude toward TCM compared with those who did not. However, after the course, there was no intergroup difference in attitudes between these 2 groups of students, although there were intragroup differences. In other words, the course made the attitudes of the 2 groups of students, which were originally different, become consistent and significantly more positive. The hierarchical effect may explain this interesting result. Students who did not admire Chinese medicine celebrities may follow a standard learning hierarchy. In this learning hierarchy, cognition was the foundation, which in turn affects emotions and ultimately behavioral tendencies. Students who admired Chinese medicine celebrities followed an experiential hierarchy, in which emotions were the foundation that influenced behavior and ultimately cognition.

Why the TCM Course Did Not Affect Cognitive and Affective Components of Attitude in EYP

Compared with FYP students, it was very interesting that TCM courses did not significantly change the total score or the scores in the cognition and affection domains in EYP students. Some facts attracted our attention that might be reasons for quite different results. First, 5th-grade EYP students had completed WM diagnostics and accessed almost all the clinical professional courses, including internal medicine, surgery, gynecology, and pediatrics, alike. Moreover, they had determined their future professional major and selected their doctoral supervisors. Therefore, long-term WM clinical training had already established an almost complete WM knowledge system and the mode of thinking for students in EYP. It was difficult to be

XSL•FO
RenderX

changed by TCM learning at a relatively later stage. It was a formidable challenge for EYP students to acknowledge and assimilate TCM, given their preconceived learning of WM, because of the fact that these 2 domains are founded on divergent theories and modalities [27-29]. However, for FYP students, 3-year learning in basic and clinical bridging courses has not yet fully westernized their thinking. They were relatively simple to change through TCM learning. This might be one of the critical reasons why total scores and scores in cognition and affection did not increase in EYP. The findings of this study were in accordance with the report of Hon et al [8]that preclinical students were more positive toward TCM compared to students in clinical years, the latter showing a more negative attitude toward TCM after studying WM. In addition, the small sample size included (36 EYP students) might be another important reason for the absence of a difference between before and after tests. It is difficult to avoid the fact that the college entrance examination enrollment of the EYP is very limited. Future studies should adopt more rigorous experimental designs to address this limitation, for example, a combination of effects of several consecutive years should be feasible.

The Effect of TCM Course on Behavioral Tendency Component of Attitude in EYP

The behavioral tendency component of attitude among EYP students was improved after the course. Behavioral intention is different from actual behavior, but it is an individual's behavioral readiness toward a certain object, representing a possible behavioral tendency. The result suggests that after learning TCM, EYP students gained more in-depth exposure to another medical strategy. It is greatly different from WM yet effective, well founded, and uniquely advantageous, and can solve some problems that WM cannot solve. Therefore, they may be willing to use TCM for health maintenance, to use the knowledge of TCM to safeguard their parents' health, to learn more about TCM, and to consider using TCM to solve patients' illnesses in future clinical work.

The Effect of TCM Course on Selection Preference of Different Therapeutic Strategies in FYP

The purpose of this part of the questionnaire was to test the behavior tendency in specific clinical conditions. The test was designed to simulate a physician's daily work, such as making an appropriate treatment plan for different types of diseases or, at least, giving proper referrals for different patients. After completing the TCM course, more FYP students realized that TCM and WM were very effective in treating acute diseases, which broke the stereotype that TCM was only suitable for chronic diseases. Many real cases they witnessed during their TCM internship may have influenced their selection preference. For example, a TCM physician cured a child with unexplained persistent high fever with a combination of xiaochaihu decoction and gypsum soup. As for chronic diseases, more FYP students tended to choose TCM treatment after the course. These diseases often have a long course and complex condition, requiring long-term management and treatment. The TCM course taught students that TCM relies on its unique theoretical system and rich practical experience and has demonstrated significant advantages and potential [30]. It emphasizes a holistic concept and treatment with syndrome differentiation for chronic diseases, which helps physicians treat patients using the overall situation, comprehensively considering the patients' physical condition, lifestyle habits, emotional factors, and other aspects. As for organic diseases, more FYP students realized that the strategy of using both TCM and WM is better. This result is reasonable and explainable. It is widely recognized that WM is the mainstream and effective treatment for organic diseases. However, during the internship, students found that participation of TCM in the treatment of organic diseases accelerated recovery, minimized side effects, and increased patient comfort. Thus, they tended to use TCM and WM for organic diseases.

The Effect of TCM Course on Selection Preference of Different Therapeutic Strategies in EYP

After the TCM course, more EYP students chose TCM and WM to treat internal diseases. EYP students had completed many clinical courses, so they were more likely to divide diseases into internal and surgical diseases according to their scope and treatment methods. Generally speaking, internal medicine diseases can be treated with medication. By learning the basic theory of TCM, along with Chinese pharmacy and Chinese medical formulas, they understood how TCM cures diseases. It is greatly different from that of WM but is a good supplement and promotion to WM. However, for surgical diseases, surgical treatment remains the first choice undoubtedly. Nevertheless, EYP students found that the TCM therapies, such as acupuncture and massage, help restore gastrointestinal, cognitive, and limb functions in postoperative patients. Therefore, they preferred to use a combination of TCM and WM to address perioperative disorders. Mental illness is another major category of disease. The therapeutic effect of antipsychotic drugs is limited, and there are many side effects. Mental illness belongs to the category of emotional disorders in TCM. By learning TCM etiology and pathogenesis, students understood that emotional disorders are related to various factors, such as organ dysfunction, yin-yang imbalance, and so on. Common treatment methods include regulating qi and relieving depression, calming the heart and eliminating phlegm, purging the liver and clearing fire, and so on. EYP students also found that TCM and WM treatment showed a satisfactory effect in patients with mental illness. Thus, more EYP students preferred a TCM and WM strategy for mental illness. The results additionally aligned with a previous study, which found that TCM was not practiced in isolation but in conjunction with WM [17]. The increased participation of TCM in treatment was expected to enhance its effectiveness.

Implications of Findings

The findings of the present study highlight the importance of integrating TCM into WM education and practice. It will lead to more holistic and patient-centered care, enhance clinical outcomes, and contribute to the development of integrated health care policies. Future research should continue to explore the long-term effects of TCM education on WM students' clinical practice and the potential benefits of integrating TCM and WM in various health care settings.

XSL•FO
RenderX

Comparison With Prior Literature

The findings of this study align with and extend several key themes observed in the existing literature on the integration of TCM and WM, particularly in the context of medical education and attitude formation. This study builds on existing literature by confirming the positive impact of TCM education on WM students' attitudes and by extending the understanding of how this impact varies by educational stage and specific disease contexts. The detailed analysis of attitude components and the introduction of novel factors, such as celebrity worship, provide new insights that can inform future research and educational practices in the integration of TCM and WM.

Strengths and Limitations

This study has several distinct advantages that contribute to its overall quality and relevance. The study used a self-controlled design, which is a robust method for assessing changes in attitudes over time. By comparing the same group of students before and after the TCM course, the study effectively controls for individual differences and external variables, providing a more accurate assessment of the impact of the TCM course. The study used the ASTCM, a validated and reliable tool designed specifically to measure medical students' attitudes toward TCM through a multidimensional analysis. This ensured that the measurements are accurate, reproducible, and reflective of the true changes in students' attitudes. Some inevitable limitations in this study should be pointed out. First, a self-administered questionnaire was used instead of an interview. The reliability of the answers may be discounted. The qualitative interview should be included to make the study more persuasive in our future work. Second, the questionnaire is a kind of survey based on text reading and filling in answers. To avoid the feeling of boredom in participants, the questionnaire was designed to be relatively brief. Thus, it was impossible to have an in-depth examination. Moreover, it is an online questionnaire. Researchers possess a limited understanding of the actual identity of the participant, the mood of the participant at that time, and whether the process of finishing the questionnaire was influenced by others. Finally, the sample size of EYP students was relatively small. In the 2-tailed *t* test, the smaller sample size makes it harder to gain differential results. Thus, the results may be influenced. A sufficient sample size or combination of effects of EYP for several consecutive years is expected in the future.

Conclusions

The study showed that offering a TCM course increased a positive attitude toward TCM in students majoring in WM. The results will provide recommendations with regard to the TCM curriculum arrangement in institutions of WM learning.

Acknowledgments

The authors would like to thank Dan Han, School of Medicine, Zhejiang University, for data collection and analysis.

Funding

This study was supported by grants from the Scientific Research Project of Hunan Provincial Department of Education (22B0876).

Conflicts of Interest

None declared.

Multimedia Appendix 1
Ethical approval document.
[PDF File, 173 KB - mededu_v12i1e55972_app1.pdf]

References

1. Zhang YL, Wang YL, Yan K, et al. Traditional Chinese medicine formulae QY305 reducing cutaneous adverse reaction and diarrhea by its nanostructure. Adv Sci (Weinh) 2024 Feb;11(5):e2306140. [doi: 10.1002/advs.202306140] [Medline: 38044276]
2. Liu S, Zhu JJ, Li JC. The interpretation of human body in traditional Chinese medicine and its influence on the characteristics of TCM theory. Anat Rec (Hoboken) 2021 Nov;304(11):2559-2565. [doi: 10.1002/ar.24643] [Medline: 34117702]
3. Weng H, Chen J, Ou A, Lao Y. Leveraging representation learning for the construction and application of a knowledge graph for traditional Chinese medicine: framework development study. JMIR Med Inform 2022 Sep 2;10(9):e38414. [doi: 10.2196/38414] [Medline: 36053574]
4. Jin L, Lin L, Liu G, et al. Inconsistency between physicians' attitudes and behaviors toward TCM: a questionnaire survey in Chengdu. Medicine (Baltimore) 2024 Mar 22;103(12):e37596. [doi: 10.1097/MD.0000000000037596] [Medline: 38518017]
5. National Health Commission. China Health Statistics Yearbook of 2022: Peking Union Medical College Press; 2022. URL: https://www.nhc.gov.cn/mohwsbwstjxxzx/tjtjnj/202305/49beded3bd984669bfe9089c6f231cf5.shtml [accessed 2026-01-02]
6. Liu X, Feng J, Liu C, et al. Medical education systems in China: development, status, and evaluation. Acad Med 2023 Jan 1;98(1):43-49. [doi: 10.1097/ACM.0000000000004919] [Medline: 35947483]

XSL•FO
RenderX

7. Chung VCH, Hillier S, Lau CH, Wong SYS, Yeoh EK, Griffiths SM. Referral to and attitude towards traditional Chinese medicine amongst western medical doctors in postcolonial Hong Kong. Soc Sci Med 2011 Jan;72(2):247-255. [doi: 10.1016/j.socscimed.2010.10.021] [Medline: 21145150]
8. Hon KLE, Leung TF, Tse HM, et al. A survey of attitudes to traditional Chinese medicine among Chinese medical students. Am J Chin Med 2005;33(2):269-279. [doi: 10.1142/S0192415X05002904] [Medline: 15974486]
9. Hon EKL, Lee K, Tse HM, et al. A survey of attitudes to traditional Chinese medicine in Hong Kong pharmacy students. Complement Ther Med 2004 Mar;12(1):51-56. [doi: 10.1016/j.ctim.2003.12.002] [Medline: 15130573]
10. Hon KLE, Twinn SF, Leung TF, Thompson DR, Wong Y, Fok TF. Chinese nursing students' attitudes toward traditional Chinese medicine. J Nurs Educ 2006 May;45(5):182-185. [doi: 10.3928/01484834-20060501-08] [Medline: 16722501]
11. Yu F, Gao J. Challenges that the law of traditional Chinese medicine has to face. Cap Food Med 2016;23(9):21-22 [FREE Full text]
12. Ng TKC, Lo MF, Fong BYF, Yee HHL. Predictors of the intention to use traditional Chinese medicine (TCM) using extended theory of planned behavior: a cross-sectional study among TCM users in Hong Kong. BMC Complement Med Ther 2022 Apr 22;22(1):113. [doi: 10.1186/s12906-022-03598-x] [Medline: 35459198]
13. Corchado Rodri´guez JM, Uden L, Wang LSL, Yang H, editors. The 8th International Conference on Knowledge Management in Organizations:Social and Big Data Computing for Knowledge Management, 1st edition: Springer; 2014. [doi: 10.1007/978-94-007-7287-8]
14. Gan X, Yang J, Yang B, Yuan J. Problems and strategies of traditional Chinese medicine teaching in western medical colleges. J Pract Intern Med Tradit Chin Med 2020;34(7):11-14. [doi: 10.13729/j.issn.1671-7813.Z20200253]
15. Xu K. Thoughts on traditional Chinese medicine teaching in western medical colleges. Chin Med Med Mod Distance Educ China 2021;4(9):19-21. [doi: 10.3969/j.issn.1672-2779.2021.04.007]
16. Wu Z. Development of Chinese Medicine Attitude Scale [Master's thesis]. : Beijing University of Traditional Chinese Medicine; 2011.
17. Harmsworth K, Lewith GT. Attitudes to traditional Chinese medicine amongst western trained doctors in the People's Republic of China. Soc Sci Med 2001 Jan;52(1):149-153. [doi: 10.1016/s0277-9536(00)00124-6] [Medline: 11144912]
18. Nakagawa S, Cuthill IC. Effect size, confidence interval and statistical significance: a practical guide for biologists. Biol Rev Camb Philos Soc 2007 Nov;82(4):591-605. [doi: 10.1111/j.1469-185X.2007.00027.x] [Medline: 17944619]
19. Lu Y. Study on the phenomenon of celebrity worship among college students in the new era [Master's thesis]. : Anhui University of Engineering; 2023 URL: https://tinyurl.com/3wy9kh2m [accessed 2026-01-02]
20. Xuan Y, Haihong L, Tao S, et al. Investigation and analysis of traditional Chinese medicine curriculum in clinical specialty of western medicine. Int J Tradit Chin Med 2018;40(6):554-557. [doi: 10.3760/cma.j.issn.1673-4246.2018.06.018]
21. Altmann TK. Attitude: a concept analysis. Nurs Forum (Auckl) 2008 Jul;43(3):144-150. [doi: 10.1111/j.1744-6198.2008.00106.x]
22. Albarracin D, Shavitt S. Attitudes and attitude change. Annu Rev Psychol 2018 Jan 4;69(69):299-327. [doi: 10.1146/annurev-psych-122216-011911] [Medline: 28841390]
23. Huang Z, Miao J, Chen J, et al. A traditional Chinese medicine syndrome classification model based on cross-feature generation by convolution neural network: model development and validation. JMIR Med Inform 2022 Apr 6;10(4):e29290. [doi: 10.2196/29290] [Medline: 35384854]
24. Yang H, Xiao X, Wu X, et al. Virtual standardized patients versus traditional academic training for improving clinical competence among traditional Chinese medicine students: prospective randomized controlled trial. J Med Internet Res 2023 Sep 20;25:e43763. [doi: 10.2196/43763] [Medline: 37728989]
25. Zhou Y, Shen H. Research and exploration on traditional Chinese medicine as an internship in western medical colleges. Curr Educ Res 2020(29):122-124.
26. Lu L, Lu T, Tian C, Zhang X. AI: Bridging ancient wisdom and modern innovation in traditional Chinese medicine. JMIR Med Inform 2024 Jun 28;12:e58491. [doi: 10.2196/58491] [Medline: 38941141]
27. Li A, Yajie L, Guoqiang C. Research on teaching methods of traditional Chinese medicine in colleges and universities of western medicine. Med Inf 2018;31(12):21-22. [doi: 10.3969/j.issn.1006-1959.2018.12.007]
28. Hu S. Traditional Chinese medicine teaching material tailored for western hospital school: "Xiehe edition" traditional Chinese medicine. Med J Peking Union Med Coll Hosp 2012;3(2):242-243 [FREE Full text]
29. Liu XH, Jin F, Hsu J, Li DN, Chen W. Comparing smartphone apps for traditional Chinese medicine and modern medicine in China: systematic search and content analysis. JMIR Mhealth Uhealth 2021 Mar 24;9(3):e27406. [doi: 10.2196/27406] [Medline: 33759786]
30. Wang J, Chen L, Wang Z, Fang X, Zhen Y. Perceptions, attitudes, beliefs, and practices toward traditional Chinese medicine therapies for herpes zoster: a cross-sectional study of healthcare professionals. Risk Manag Healthc Policy 2024;17:1239-1251. [doi: 10.2147/RMHP.S461149] [Medline: 38765781]

Abbreviations

ASTCM: Attitude Scale of Traditional Chinese Medicine

XSL•FO
RenderX

EYP: 8-year program
FYP: 5-year program
TCM: traditional Chinese medicine
WM: Western medicine

Edited by B Lesselroth; submitted 01.Jan.2024; peer-reviewed by L Yu, LM Chen; revised version received 19.Apr.2025; accepted 13.May.2025; published 16.Jan.2026.

Please cite as:
He H, Cui H, Guo L, Liao Y
The Effect of a Traditional Chinese Medicine Course on Western Medicine Students' Attitudes Toward Traditional Chinese Medicine: Self-Controlled Pre-Post Questionnaire Study
JMIR Med Educ 2026;12:e55972
URL: https://mededu.jmir.org/2026/1/e55972
doi:10.2196/55972

© Haoyu He, Hanjin Cui, Li Guo, Yanhui Liao. Originally published in JMIR Medical Education (https://mededu.jmir.org), 16.Jan.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Education, is properly cited. The complete bibliographic information, a link to the original publication on https://mededu.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Interactive, Image-Based Modules as a Complement to Prosection-Based Anatomy Laboratories: Multicohort Evaluation

Caroline Sumner[1*], BS; Sami L Case[2*], PhD; Samuel Franklin[3], PhD; Kristen Platt[3], PhD

[1]College of Medicine, University of Kentucky, Lexington, KY, United States

[2]Department of Biomedical Sciences, College of Veterinary Medicine and Biomedical Sciences, Colorado State University, Fort Collins, CO, United States

[3]Department of Neuroscience, College of Medicine, University of Kentucky, Lexington, KY, United States

[*]these authors contributed equally

Corresponding Author:
Kristen Platt, PhD
Department of Neuroscience
College of Medicine
University of Kentucky
780 Rose Street
MS 207 Medical Sciences Building
Lexington, KY, 40536
United States
Phone: 1 8592186648
Fax: 1 8593235946
Email: platt.kristen@uky.edu

Abstract

Background: As medical and allied health curricula adapt to increasing time constraints, ethical considerations, and resource limitations, digital innovations are becoming vital supplements to donor-based anatomy instruction. While prior studies have examined the effectiveness of prosection versus dissection and the role of digital tools in anatomy learning, few resources align interactive digital modules directly with hands-on prosection experiences.

Objective: This project addresses that gap by introducing an integrated, curriculum-aligned platform for self-guided cadaveric learning.

Methods: We created Anatomy Interactives, a web-based laboratory manual structured to complement prosection laboratories for MD, DPT, and PA students. Modules were developed using iSpring Suite (iSpring Solutions Incorporated) and included interactive labeled images, donor photographs, and quiz-style self-assessments. Learners engaged with modules before, during, or after laboratory sessions. PA/DPT and MD students completed postcourse surveys evaluating module use and perceived impact. MD student examination scores from a 2023 cohort (no module access) were compared to a 2024 cohort (with access) to evaluate effectiveness.

Results: A total of 147 students completed the survey (31 PA/DPT and 116 MD). The majority reported using modules for 1-2 hours per week and found them helpful for both written and laboratory examinations. MD students in the 2024 cohort performed better on all 3 examinations compared to the 2023 cohort, with 2 examination median differences reaching statistical significance (Mann-Whitney U, $P<.001$). Qualitative feedback highlighted accessibility, content reinforcement, and user engagement as key benefits.

Conclusions: Interactive modules integrated with prosection laboratories enhanced learner engagement and performance. This hybrid digital-donor model shows promise for scalable, learner-centered gross anatomy education.

(JMIR Med Educ 2026;12:e85028) doi:10.2196/85028

KEYWORDS

digital learning; education; medical; gross anatomy; prosection; hybrid anatomy education

XSL•FO
RenderX

Introduction

Gross anatomy remains a foundational component of health professions education, providing essential knowledge for clinical reasoning, procedural competency, and disease management [1]. However, students in prosection-based courses often struggle to transition from routine passive study to active, tactile engagement during laboratory sessions. Most existing digital resources are intended for use before or after the laboratory rather than functioning as interactive tools to engage the learner at the dissection table. As a result, students may feel disconnected from the donor during prosection, limiting the depth and integration of their learning experience.

Historically, cadaveric dissection has been regarded as the pedagogical gold standard for anatomy education [2,3]. However, contemporary anatomical curricula face mounting logistical, ethical, and financial constraints. These constraints include high costs for donor procurement and preservation, limited laboratory space and faculty, exposure to formaldehyde and other chemicals, and ethical concerns. These forces are driving the adoption of alternative teaching approaches such as prosection, peer-assisted learning, and digital platforms [4].

A growing body of literature suggests that prosection-based learning can be as effective as, and in some cases more efficient than, dissection in facilitating anatomical knowledge acquisition and clinical relevance [5-7]. A comparative assessment demonstrated that nonmedical undergraduate students reported similar academic outcomes and levels of engagement regardless of whether they were taught through dissection or prosection [3]. Likewise, in a study of veterinary students, prosections prepared by qualified staff outperformed traditional dissection in promoting both efficiency and student satisfaction [8]. Coker et al [9] report that students who learned by prosection performed better on questions specifically involving complex or deep structures, but simultaneously demonstrated higher satisfaction with and preference for a dissection-based model.

Concurrently, the integration of digital technologies into anatomy curricula is transforming how learners engage with spatial content. Radiographic imaging is a time-tested methodology that has been incorporated into undergraduate medical education with increasing frequency [10-12]. Ultrasound-based anatomy instruction has grown in popularity in recent years [13,14]. In addition, meta-analytic findings underscore the effectiveness of virtual- and augmented-reality platforms in reinforcing anatomical understanding and enhancing learner autonomy [15,16]. This shift was accelerated by the COVID-19 pandemic, during which hybrid and online formats became critical for continuity in medical education [17].

Other commercially available digital tools serve to augment the didactic anatomy learning experience. For example, the Anatomage and Sectra digital dissection technologies rely on the Visible Human Project work to generate a donor experience in physical format reminiscent of a donor table [18,19]. Use of virtual dissection as a supplement to donor dissection has been viewed favorably by students [20]. The Anatomage also offers unique advantages, such as blood flow modeling that cannot be replicated in a donor [21]. However, these devices are often proposed as an alternative to the study of the donor body and do not address the gap wherein digital media can complement the study of the prosected donor in real time. Models such as hybrid prosection-based anatomy laboratories demonstrated high student satisfaction, with learners appreciating the flexibility and clarity offered by digital supplements [7].

Several institutions have developed innovative digital platforms to complement or enhance traditional gross anatomy instruction, exemplifying how technology can be leveraged to support donor-based learning. For instance, the University of Michigan's BlueLink program offers students an extensive library of presentation slide-like materials rich with cadaveric images, facilitating self-directed study and review [22]. These resources, which also include laboratory manuals and interactive files, are designed to provide a flexible and widely accessible digital ecosystem for anatomy education. Similarly, Texas Tech University Health Sciences Center has developed a gross anatomy program featuring a varied suite of online tools [23]. Their offerings include prelaboratory cadaveric dissection videos, practice quizzes for reinforcing knowledge of cadaveric material, and various other interactive learning modules, providing multifaceted support for students beyond traditional laboratory settings.

From a theoretical perspective, multimodal learning and constructivist principles provide a useful lens for understanding why integrated digital prosection tools may enhance anatomy learning. Multimodal learning theory posits that students build stronger and more durable knowledge when information is presented through complementary channels, such as visual, spatial, textual, and tactile, which is an especially relevant consideration in gross anatomy [24]. Constructivist frameworks further suggest that learners deepen understanding by actively engaging with material, generating connections, and iteratively refining mental models during hands-on exploration, and this approach has been shown to facilitate more meaningful learning than traditional methods in the context of human anatomy [25]. Past constructivist approaches in anatomical sciences education include techniques such as casting and creating models [26,27]. Digital resources that integrate with prosection-based laboratory work have the potential to support meaning-making in real time, helping students actively construct anatomical understanding while drawing on multiple modes of representation.

Despite the growing prevalence of prosection-based anatomy coursework and the surge of fully virtual learning models, few resources are designed specifically to bridge digital learning with hands-on prosected donor experiences in a structured, curriculum-aligned format. The Anatomy Interactives platform was developed to fill this gap, providing guided modules that integrate interactive digital content with cadaveric prosections [28]. Unlike static resources or purely virtual simulations, this tool is designed to reinforce laboratory-based learning through active engagement, self-assessment, and iterative feedback in a digital format while students are present in the gross anatomy laboratory.

This study evaluates the effectiveness of the Anatomy Interactives platform in 2 different student populations and implementation formats. This includes analyses of student

XSL•FO
RenderX

engagement, self-reported preparedness, and academic performance. This work will interest medical and health professions students and educators whose classroom is the cadaveric laboratory. These learners often face challenges integrating didactic anatomical knowledge with hands-on experiences, particularly when transitioning between digital prelaboratory preparation and in-laboratory prosection-based learning. By comparing student experiences and performance across 2 cohorts, one with access to the digital modules and one without, we aim to provide evidence for the impact of integrated, digital-prosection hybrid models on anatomy education for future health professionals.

Methods

Ethical Considerations

All work with living subjects (ie, surveys, website traffic, and grades) was reviewed by the University of Kentucky Institutional Review Board (IRB) and determined to be exempt (IRB Protocol #94449) from human subjects research oversight. All procedures comply with the World Medical Association Declaration of Helsinki. The donor registration form of the University of Kentucky Willed Body Program includes language of consent for "education, general research, and/or to further innovative technologies." This project used anonymous, optional surveys. No identifiable or sensitive personal data were collected, and all responses were stored without linkage to participants. Informed consent was collected via the cover page of the survey with approved study details, contact information for the University of Kentucky Office of Research Integrity, and a final question asking, "Do you consent to your anonymized survey data being used in future research publications or other research purposes?" Participants received no compensation for taking part in the survey.

Curricular Context

We implemented this educational resource as a required component of the first-year gross anatomy courses at the University of Kentucky, serving learners in the MD, DPT, and PA programs. During the course laboratory days, the medical students rotate between physical examination with point-of-care ultrasound, prosection in the gross anatomy laboratory, and radiology-focused integrated clinical case topic sessions. This manuscript only considers the gross anatomy component of the laboratory experiences. All medical students participated in donor-based laboratory sessions using prosected donors, while PA and DPT students used an alternating dissection model (half dissected one day; the other half dissected the opposite day, with peer teaching at the end of each session—a long-used dissection model) [29]. We specifically designed the Anatomy Interactives modules to supplement these in-person experiences. Students had no prerequisite training in anatomy beyond the concurrent lectures offered within the course. Facilitators (anatomy instructors) had no specific training on the Anatomy Interactives.

Development of Anatomy Interactives

Overview

We created the Anatomy Interactives website to function as a digital laboratory manual, aiming to enhance anatomical structure identification and spatial reasoning through the use of prosected donor images.

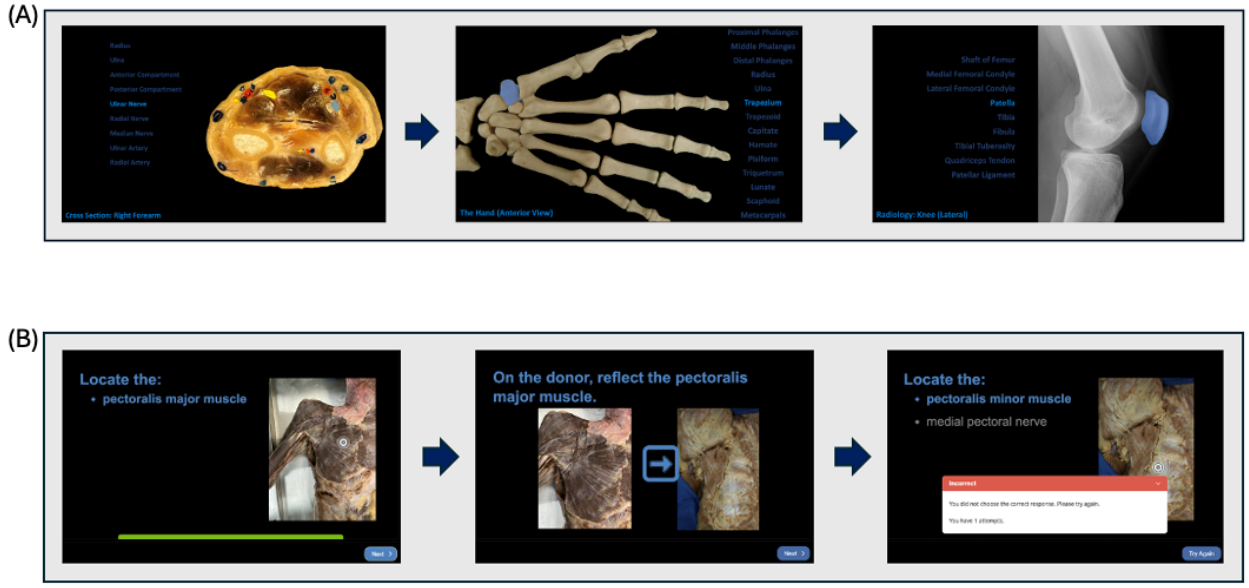
Our primary goal was to provide a guided, active learning experience that reinforced anatomical concepts outside of the traditional laboratory setting. We executed the development process in 2 distinct phases, representing the type of materials that we created.

Phase 1: Foundational Content Module Development

In this phase, we generated labeled anatomical images from diverse modalities, including arteriograms, cross-sections, computed tomography scans, radiographs, and osteological photographs (Figure 1A). We embedded these images with hover-over identification features, where selecting an anatomical term would highlight the corresponding structure on the image, and vice versa. We initially built this content in Microsoft PowerPoint and subsequently exported it to HTML format using iSpring Suite 11 to ensure web accessibility.

XSL•FO
RenderX

Figure 1. Sample view of Anatomy Interactives Website. (A) Examples of fixed labeled images from the hover-over identification resource. Screenshots from cross-sectional, osteology, and radiology image sets showing labeled anatomical structures. Structures and corresponding names are highlighted for visual reinforcement of anatomical terminology. (B) Example Anatomy Interactives module interface: pectoral muscle identification task. Screenshot from the pectoral region module showing a pin-on-image question asking students to identify a muscular structure. Immediate feedback is provided upon incorrect identification, with instructional guidance on adjusting the donor to reveal the underlying structure.



Phase 2: Prosection Laboratory Module Development

In the second phase, we focused on developing interactive, donor-based modules (Figure 1B). We dedicated each module to a specific anatomical subregion (eg, foot and thigh) and incorporated a series of image-based questions, such as pin-on-image tasks, multiple-choice questions, and check-all-that-apply questions. We augmented these with pop-up content designed to provide relevant clinical and conceptual context. Each module culminated in a score based on accuracy, offering students formative feedback on their understanding. We then grouped multiple modules to comprise a regional "laboratory" based on curricular structure. We developed six laboratories: (1) back and upper limb, (2) lower limb, (3) thorax, (4) abdomen, (5) pelvis, and (6) head and neck.

Implementation

Overview

At the beginning of the semester, we oriented students to the Anatomy Interactives platform via an in-class demonstration. We hosted the modules on the institution's learning management system, and they were available on demand throughout the course. We suggested students use the modules before, during, and after their scheduled laboratory sessions to reinforce learning. During the laboratory time for the MD students, the main exercise was for them to complete the module with their group (approximately 8 students per group, with 2 donors). However, we did not mandate completion of the modules for course credit for any student (MD, PA, and DPT). Students used these modules using university-sponsored iPads and any personal devices that could access the internet. The modules were all made available to students from the start of their gross anatomy course, allowing them to integrate both types of resources into prelaboratory preparation and ongoing review throughout the semester.

MD Students

Students in the MD program had 80-minute gross anatomy prosection laboratory sessions 6 weeks out of the 9-week course. MD students collaborated with their table group to use the digital Prosection Laboratory Modules as they worked with their donor. Students answered multiple-choice questions or select-all-that-apply questions as a team, and compared the location of the item on the hotspot questions to the anatomy of their group's specific donor. We made the Foundational Content Modules available for students to self-study on their own time to prepare for assessments. The fall 2024 cohort of MD students had access to the Anatomy Interactives modules as part of their standard curriculum. The fall 2023 cohort of MD students, serving as a comparison group, completed the same gross anatomy course without access to these modules. Both cohorts completed the course with the same professors with nearly identical assessments. Both groups had full-time access to the anatomy laboratory and participated in the same number of small-group laboratory sessions, the only difference being the use of the modules.

PA/DPT students

In spring 2024, we provided PA/DPT students access during their shared dissection-based gross anatomy laboratory course. The format of this course consisted of 2 days of 2-hour lectures followed by 2-hour laboratories per week for 16 weeks. We encouraged students to use the Foundational Content Modules and Prosection Laboratory Modules as self-study to prepare for laboratories and course assessments.

XSL•FO
RenderX

Assessment and Evaluation Strategy

Overview

We used a mixed-methods approach to evaluate the usage and effectiveness of the Anatomy Interactives platform. This strategy included student surveys, website usage data, and measures of academic performance.

Surveys

Surveys were distributed at the conclusion of the respective course to all PA/DPT students and MD students who had access to the modules. The surveys aimed to assess perceived utility, engagement, and the learning impact of the platform. Surveys included questions including which modules were used, time spent using the modules, time spent in the laboratory, opinions regarding score correlation with module usage, opinions regarding preparation level and module usage, study purposes when using the modules, and use of modules with the donor or in a stand-alone study. We collected open-ended responses pertaining to likes, dislikes, and recommendations for the modules and analyzed them thematically.

Website Analytics

We collected web traffic data from Google Analytics. These included metrics such as daily active users and usage patterns, particularly around examination periods and laboratory sessions, to understand engagement trends.

Academic Performance

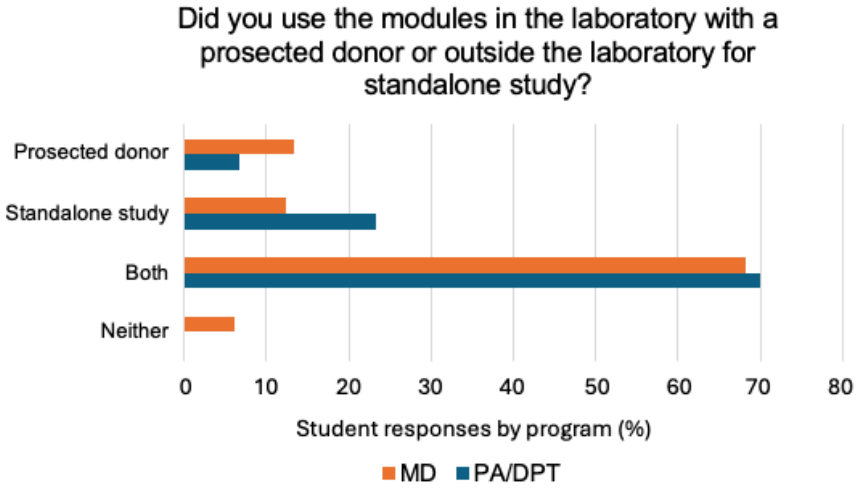
We obtained written and practical anatomy examination scores from the Office of Medical Education (determined as not human research per the University of Kentucky IRB) for 2 MD student cohorts: the 2023 cohort (N=206) without access to the modules, and the 2024 cohort (N=205) with access. To ensure baseline academic comparability between student cohorts, we compared Medical College Admission Test (MCAT) scores, overall undergraduate grade point averages (uGPA), and science undergraduate grade point averages (sGPA) for both MD cohorts to test for differences. We then conducted a comparison of examination performance between these 2 MD cohorts to assess the educational impact of the Anatomy Interactives modules. We first tested for normality of the data and determined that our data was not normally distributed. Thus, we analyzed significance using an unpaired Mann-Whitney *U* test with statistical significance set at $P<.05$ and assessed effect size by calculating the rank-biserial correlation coefficient (r_{rb}). Some examination scores were not available in the dataset, and therefore, the number of participants changes per examination; this was accounted for in the data analysis. Statistical analyses were conducted in GraphPad Prism (version 10.6.1; GraphPad Software, LLC). r_{rb} was manually calculated using the simplified formula as previously described [30].

Results

Analysis of the student surveys (N=31 for PA/DPT and N=116 for MD) and website trafficking data provided significant insight into the usage of the modules by students.

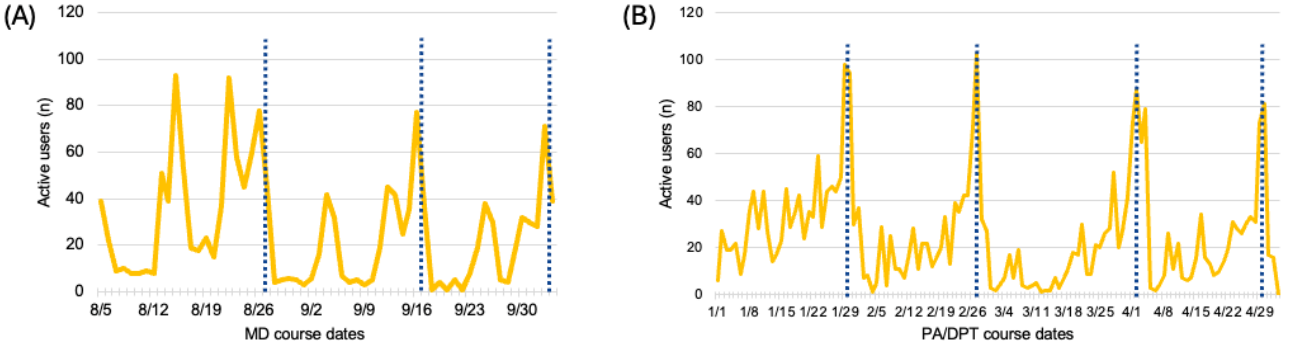
For the PA/DPT cohort, 23 out of 31 (74%) students reported using the modules for 1-2 hours per week in addition to the 5-6 hours spent with the donor in allotted laboratory time. Of the medical students surveyed, 85 out of 116 (73%) students report using the laboratory modules 1-2 hours per week. Regarding the nature of their usage, 22 out of 31 (70%) PA/DPT students and 79 out of 116 (68%) medical students used the modules both with the donor and as a stand-alone study tool (Figure 2). The website data indicate the modules are heavily trafficked leading up to the examination, the number of users increasing drastically in the 1-3 days leading up to an examination for PA/DPT and medical students (Figure 2). Module page traffic also increases on laboratory days for medical students, as well as the day before and the day after a laboratory session (Figure 3). The laboratory modules had much higher traffic than the hover-over labeled images.

Figure 2. Student self-report of module usage type. Percentage of surveyed MD students (n=116) reporting use of the Anatomy Interactives modules in each context: with the donor during laboratory sessions, as a stand-alone study tool, or both. Percentage of surveyed PA/DPT students (n=31) reporting use of the modules in each context.



XSL•FO
RenderX

Figure 3. Daily users of the Anatomy Interactives platform for MD and PA/DPT courses. (A) Number of MD student users per day during the MD gross anatomy course (August 5, 2024, to October 24, 2024). (B) Number of PA/DPT student users per day during the PA/DPT gross anatomy course (January 2, 2024, to April 30, 2024). Spikes correspond to laboratory sessions and examination periods. Vertical dashed lines indicate examination dates.



When asked about the relationship between their module usage and examination score, 23 out of 31 (74%) PA/DPT students reported that the modules helped them on their written examination, and 26 out of 31 (84%) believed the modules helped them on their laboratory examination. The PA/DPT cohort felt more prepared for the examination when they had a laboratory module to support the curriculum. The medical student survey revealed 80 out of 116 (69%) of that cohort believed their assessment scores reflected the time and effort they spent on modules.

To examine the modules' impact, examination scores from 2 different cohorts were compared (Table 1). First-year medical students in 2023 did not use the modules as part of their laboratory donor experience, while the medical students in 2024 did have access to the new resource. For all 3 measures we used to test comparability (MCAT scores, uGPA, and sGPA), the data were not normally distributed (Shapiro-Wilk test, $P<.001$) and not significantly different (Mann-Whitney U, $P<.001$) between the 2 MD cohorts. All MD cohort examination score data failed normality testing (Shapiro-Wilk test, $P<.001$), so groups were compared using nonparametric testing. Briefly, Mann-Whitney U tests were used to test for cohort examination score differences, and r_{rb} values were calculated and used to assess effect size. Table 1 summarizes medians, P values, Mann-Whitney U, and the r_{rb} values from these tests. Students in the 2024 cohort scored higher on both the written and practical portions across examination 1 and examination 3, with moderate effect sizes for the written portions and total examination scores.

Table 1. Comparison of anatomy examination scores between MD cohorts with and without module access. Median written, laboratory, and total scores for 3 examinations in the 2023 cohort (no module use) and the 2024 cohort (module use), with P values, Mann-Whitney U, and rank-biserial correlations.

	2023 MD cohort, median (IQR), %	2024 MD cohort, median (IQR), %	P value	Mann-Whitney U	Rank-biserial correlation
Examination 1 medians					
Written portion	82 (74-88)	86 (79-94)	<.001	15,130	.28
Laboratory portion	90 (85-95.5)	92.5 (85-97.5)	.01	18,153	.14
Total	86.1 (78.9-90)	88.9 (82.2-94.4)	<.001	15,778	.25
Examination 2 medians					
Written portion	83 (76-90)	84 (78-90)	.13	19,201	.09
Laboratory portion	92.5 (87.5-95)	92.5 (87.5-95)	.39	19,984	.05
Total	87.8 (81.1-91.1)	87.8 (82.2-92.2)	.16	19,327	.08
Examination 3 medians					
Written portion	85 (79.5-92)	92 (84-96)	<.001	12,420	.41
Laboratory portion	92.5 (87.5-95)	92.5 (87.5-97.5)	.49	20,092	.04
Total	87.8 (83.3-92.2)	92.2 (87.5-97.5)	<.001	14,645	.30

When asked about the purpose of their module usage, 27 PA/DPT students used the modules for general studying, 8 before laboratory for preparation, 24 on laboratory day for preparation, 23 for understanding current content, and 25 for targeted examination review. A review of medical students' qualitative feedback on the modules revealed students appreciated the accessibility of the modules, reinforcement of course content, engagement provided by the active-learning style, and the complementary nature of the modules and the prosected donor experience. Accessibility and flexibility were mentioned as advantages by 43 students, with 27 stating an appreciation for the convenience and ease of using the modules on their own time, and 16 mentioning they were helpful during independent examination preparation. The complementary

XSL•FO
RenderX

nature of the modules to the prosected donor experience was highlighted by 19 students who detailed the usefulness of having the resource available during the laboratory session. Furthermore, 41 students mentioned liking how the modules support course content and reinforce lecture material. Technological characteristics were mentioned by many students, with 29 students giving positive feedback on the technology and user interface of the modules, while 48 students thought it was an area for improvement.

Discussion

Principal Findings

The Anatomy Interactives digital laboratory manual achieved its intended goal of supplementing the prosection-based gross anatomy curriculum with an interactive, learner-centered resource that promoted structure recognition, spatial understanding, and self-directed learning. Moving toward a student-centered learning approach in anatomy may have positive effects on student performance [31]. While learners report that student-centered approaches promote their independent learning and peer communication, they perceive that faculty-led learning increases their anatomical knowledge [32]. Before the implementation of the modules at our institution, the prosection laboratory had become a completely passive faculty-led environment with minimal student participation. By offering a digitally guided experience aligned with the prosected donor material, this innovation supported student engagement and potentially improved academic performance across multiple learner groups.

Both MD and PA/DPT learners found the modules helpful in preparing for examinations and reinforcing course content. Website analytics corroborated student usage of the modules as a study material outside of class, as website usage spiked before laboratory sessions and assessments. This suggests that the modules were integrated into learners' study routines in a timely and meaningful way. It is still uncertain whether engagement with various digital media in anatomical education correlates to improved student outcomes and satisfaction with the learning experience [33]. However, the goal of implementing the digital modules during laboratory time was to promote engagement as a means to foster a constructivist learning environment [34]. That is to say, during the laboratory, the modules were a student-centered method to have learners interact with the donors and the course content in a meaningful way, such that they could develop or "construct" their own knowledge base as it related to the gross anatomical structures. The outcomes observed suggest that we successfully addressed the initial challenge of providing anatomy learners with a resource that bridges traditional prosection-based instruction and flexible, learner-driven study to enhance both student satisfaction and academic performance.

Most notably, medical students who used the modules in 2024 outperformed the previous cohort (2023) on both written and practical assessments, even after controlling for prior academic performance (ie, MCAT scores, uGPA, and sGPA), indicating a potential causal relationship between module usage and examination performance. While the use of MCAT scores is an imperfect means of normalizing between cohorts, it is a method that has been used by similar studies [9]. More work is warranted to assess whether the use of the digital laboratory modules directly improves course performance compared to baseline (faculty-led prosection session with no digital modules).

Recent work has drawn light to the importance of educational design that creates flourishing medical and health sciences learners and practitioners [35]. Independent of developing student knowledge recall, the anatomy laboratory offers a unique environment in which to promote the holistic development of future clinicians [36,37]. In fact, with careful course design, the anatomy laboratory can be harnessed as a means to promote the development of flourishing physicians and health care providers [38]. The Anatomy Interactives laboratory modules could be enhanced in future iterations to encourage learners to embrace a sense of meaning and belonging in line with these goals.

Limitations

This educational resource was implemented and evaluated within a single institution, limiting the generalizability of the results. The anatomy curriculum and access to prosected donors may differ at other institutions, affecting feasibility and relevance. While the resource is available for use in any program, the module progression is modeled after the University of Kentucky's gross anatomy curriculum and may not directly follow along the curriculum of other institutions. To access the resource, students simply need a device that connects to the internet. To implement the use of the modules in a digital-donor laboratory experience, institutions must have access to prosected donors and a space to conduct a laboratory session, likely with course director oversight.

Additionally, the evaluation relied on student self-reports and examination scores as primary metrics, which, while informative, do not fully capture long-term knowledge retention, clinical application, or behavioral change. The module usage and examination data were examined on a cohort level, rather than correlated on an individual level, which limits the interpretation of their association. Furthermore, although MCAT scores, uGPA, and sGPA were used to control for incoming academic aptitude, unmeasured variables such as instructor differences, schedule differences, or peer learning dynamics could have influenced performance outcomes.

Future Directions

During implementation, we found that introducing the resource early in the course and orienting students to its features increased usage. Embedding the modules via a link from the learning management system allowed easy access and made integration into student study workflows more seamless. However, we also learned that variability in student technology preferences and learning styles may affect engagement; while some praised the interactivity of the modules, others noted opportunities for improvement in interface design and navigation.

Future iterations of the Anatomy Interactives platform will aim to improve user interface design, enhance accessibility across devices, and incorporate more clinically relevant content, such as radiographic correlations and pathology examples. Broader implementation at multiple institutions would help test

XSL•FO
RenderX

generalizability and effectiveness across diverse learner populations and curricula. Additionally, incorporating longitudinal tracking and correlating module usage with clinical anatomy performance in later coursework or clerkships could strengthen claims about its educational impact.

On a broader scale, this work supports ongoing shifts in anatomy education toward multimodal, learner-driven resources. The findings may inform curricular policy regarding the balance between traditional laboratory experiences and flexible digital tools, particularly as programs navigate logistical constraints, student wellness, and educational equity.

Acknowledgments

The authors would like to acknowledge the donors who gifted their bodies to the Willed Body Program at the University of Kentucky. Generative AI was not used in any part of producing this manuscript.

Funding

The article processing fee (APF) was paid by the University of Kentucky Department of Neuroscience Research Enhancement Award (KP). The funder had no involvement in this study's design, data collection, analysis, interpretation, or the writing of this paper.

Data Availability

The datasets generated or analyzed during this study are not publicly available due to concerns of student educational data protection, but are available in an anonymous format from the corresponding author on reasonable request.

Authors' Contributions

Conceptualization: CS (equal), SF (equal), SLC (equal), KP (equal)
Data curation: CS (equal), SLC (equal), KP (equal)
Formal analysis: CS (lead), SLC (lead), KP (supporting)
Methodology: CS (equal), SF (equal), SLC (equal), KP (equal)
Writing – original draft: CS (equal), SLC (equal), KP (equal)
Writing – review & editing: CS (equal), SLC (equal), KP (equal)

Conflicts of Interest

None declared.

References

1. Wilson AB, Miller CH, Klein BA, Taylor MA, Goodwin M, Boyle EK, et al. A meta-analysis of anatomy laboratory pedagogies. Clin Anat 2018;31(1):122-133. [doi: 10.1002/ca.22934] [Medline: 28612403]
2. Yeager VL. Learning gross anatomy: dissection and prosection. Clin Anat 1996;9(1):57-59. [doi: 10.1002/(SICI)1098-2353(1996)9:1<57::AID-CA12>3.0.CO;2-9] [Medline: 8838283]
3. McWatt SC, Newton GS, Umphrey GJ, Jadeski LC. Dissection versus prosection: a comparative assessment of the course experiences, approaches to learning, and academic performance of non-medical undergraduate students in human anatomy. Anat Sci Educ 2021;14(2):184-200. [doi: 10.1002/ase.1993] [Medline: 32539226]
4. Quek FF. Revolutionising anatomy education: the current role of digital technologies in enhancing anatomy learning for undergraduate medical students. Cureus 2024;16(12):e75919. [doi: 10.7759/cureus.75919] [Medline: 39711926]
5. McMenamin PG, McLachlan J, Wilson A, McBride JM, Pickering J, Evans DJR, et al. Do we really need cadavers anymore to learn anatomy in undergraduate medicine? Med Teach 2018;40(10):1020-1029. [doi: 10.1080/0142159X.2018.1485884] [Medline: 30265177]
6. Lackey-Cornelison WL, Bauler LD, Smith J. A comparison of the effectiveness of dissection and prosection on short-term anatomic knowledge retention in a reciprocal peer-teaching program. Adv Physiol Educ 2020;44(2):239-246. [doi: 10.1152/advan.00031.2020] [Medline: 32412383]
7. Schurr AF, Burg BJ, Dickinson E, Granatosky MC. No cuts, no buts: satisfaction of first-year medical students with a hybrid prosection-based model for learning gross anatomy during the COVID-19 pandemic. Anat Sci Educ 2022;15(5):827-838 [FREE Full text] [doi: 10.1002/ase.2205] [Medline: 35726438]
8. Clement SM, Ubben TA, Yates DT. Cadaveric prosections prepared by qualified instructional staff were more efficient and effective teaching modalities for veterinary gross anatomy than in-class dissections by students. J Vet Med Educ 2024;51(5):593-609. [doi: 10.3138/jvme-2024-0031] [Medline: 39499889]

XSL•FO
RenderX

9. Coker C, Rogers RS, Freed B, Steele R, Kinde MN, Danilova G, et al. Battle of the sections: student outcomes and course feedback support combined prosection and dissection laboratory formats to maximize student success. Anat Sci Educ 2024;17(5):1012-1025. [doi: 10.1002/ase.2420] [Medline: 38570916]
10. Hussey D, Shaw AV, Brian PL, Lazarus MD. Learning head and neck anatomy through a radiological imaging platform. MedEdPORTAL 2022;18:11230 [FREE Full text] [doi: 10.15766/mep_2374-8265.11230] [Medline: 35342790]
11. Marsland MJ, Tomic D, Brian PL, Lazarus MD. Pelvis anatomy tutorial using radiological images. MedEdPORTAL 2018;14:10778 [FREE Full text] [doi: 10.15766/mep_2374-8265.10778] [Medline: 30800978]
12. Lack CM, Pena J, Grubb EXL, Shen E, Hiatt KD, Benayoun MD, et al. Cranial nerve anatomy using a modular and multimodal radiologic approach. MedEdPORTAL 2022;18:11261 [FREE Full text] [doi: 10.15766/mep_2374-8265.11261] [Medline: 35720637]
13. So S, Patel RM, Orebaugh SL. Ultrasound imaging in medical student education: impact on learning anatomy and physical diagnosis. Anat Sci Educ 2017;10(2):176-189. [doi: 10.1002/ase.1630] [Medline: 27286419]
14. Bell FE, Neuffer FH, Haddad TR, Epps JC, Kozik ME, Warren BC. Active learning of the floor of mouth anatomy with ultrasound. Anat Sci Educ 2019;12(3):310-316. [doi: 10.1002/ase.1839] [Medline: 30414266]
15. McBain KA, Habib R, Laggis G, Quaiattini A, M Ventura N, Noel GPJC. Scoping review: the use of augmented reality in clinical anatomical education and its assessment tools. Anat Sci Educ 2022;15(4):765-796. [doi: 10.1002/ase.2155] [Medline: 34800073]
16. Odogwu TS, Mohamed EAH, Mishu L, Umahi I. Effect of virtual reality simulation on anatomy learning outcomes: a systematic review. Cureus 2025;17(4):e81893. [doi: 10.7759/cureus.81893] [Medline: 40342485]
17. Attardi SM, Harmon DJ, Barremkala M, Bentley D, Brown KM, Dennis JF, et al. An analysis of anatomy education before and during COVID-19: August-December 2020. Anat Sci Educ 2022;15(1):5-26 [FREE Full text] [doi: 10.1002/ase.2152] [Medline: 34758206]
18. Al-Redouan A, Dudin A, Urbanek AJ, Olsson E, Kachlik D. Visible human project based applications can prompt integrating cross-sectional anatomy into the medical school curriculum when combined with radiological modalities: a three-year cross-sectional observational study. Ann Anat 2025;257:152357. [doi: 10.1016/j.aanat.2024.152357] [Medline: 39577816]
19. Ackerman MJ. The visible human project. Stud Health Technol Inform 2022;288:134-140. [doi: 10.3233/SHTI210988] [Medline: 35102835]
20. Darras KE, Spouge R, Hatala R, Nicolaou S, Hu J, Worthington A, et al. Integrated virtual and cadaveric dissection laboratories enhance first year medical students' anatomy experience: a pilot study. BMC Med Educ 2019;19(1):366 [FREE Full text] [doi: 10.1186/s12909-019-1806-5] [Medline: 31590672]
21. Kavvadia E, Katsoula I, Angelis S, Filippou D. The anatomage table: a promising alternative in anatomy education. Cureus 2023;15(8):e43047 [FREE Full text] [doi: 10.7759/cureus.43047] [Medline: 37692592]
22. Alsup K, Fox G. University of Michigan BlueLink Photographic Atlas. 2014. URL: https://sites.google.com/a/umich.edu/bluelink/resources/bluelink [accessed 2025-12-13]
23. Gest TR. Texas Tech University Health Sciences Center Gross Anatomy Home. 2000. URL: https://anatomy.elpaso.ttuhsc.edu/index.html [accessed 2012-12-13]
24. Mayer RE. Using multimedia for e‐learning. J Comput Assisted Learn 2017;33(5):403-423 [FREE Full text] [doi: 10.1111/jcal.12197]
25. Mota MF, Mata FRD, Aversi-Ferreira TA. Constructivist pedagogic method used in the teaching of human anatomy. Int J Morphol 2010;28(2):369-374. [doi: 10.4067/s0717-95022010000200005]
26. Hermiz DJ, O'Sullivan DJ, Lujan HL, DiCarlo SE. Constructivist learning of anatomy: gaining knowledge by creating anatomical casts. Anat Sci Educ 2011;4(2):98-104. [doi: 10.1002/ase.206] [Medline: 21337719]
27. Rodenbaugh DW, Lujan HL, DiCarlo SE. Learning by doing: construction and manipulation of a skeletal muscle model during lecture. Adv Physiol Educ 2012;36(4):302-306 [FREE Full text] [doi: 10.1152/advan.00093.2012] [Medline: 23209011]
28. Platt K, Franklin S, Case S. University of Kentucky College of Medicine Anatomy Interactives. 2021. URL: https://medapp.mc.uky.edu/anatomyinteractives [accessed 2025-12-11]
29. Brueckner JK, MacPherson BR. Benefits from peer teaching in the dental gross anatomy laboratory. Eur J Dent Educ 2004;8(2):72-77. [doi: 10.1111/j.1600-0579.2003.00333.x] [Medline: 15059083]
30. Wendt H. Dealing with a common problem in social science: a simplified rank‐biserial coefficient of correlation based on the U statistic. Euro J Social Psych 1972;2(4):463-465 [FREE Full text] [doi: 10.1002/ejsp.2420020412]
31. Wilson AB, Brown KM, Misch J, Miller CH, Klein BA, Taylor MA, et al. Breaking with tradition: a scoping meta-analysis analyzing the effects of student-centered learning and computer-aided instruction on student performance in anatomy. Anat Sci Educ 2019;12(1):61-73. [doi: 10.1002/ase.1789] [Medline: 29659155]
32. Abdullah E, Lone M, Balta JY. Student-centered learning in the anatomy laboratory: medical students' perspective. Med Sci Educ 2020;30(4):1459-1464 [FREE Full text] [doi: 10.1007/s40670-020-01094-w] [Medline: 34457813]
33. Pickering JD, Swinnerton BJ. Exploring the dimensions of medical student engagement with technology-enhanced learning resources and assessing the impact on assessment outcomes. Anat Sci Educ 2019;12(2):117-128 [FREE Full text] [doi: 10.1002/ase.1810] [Medline: 30168911]

XSL•FO
RenderX

34. Mukhalalati BA, Taylor A. Adult learning theories in context: a quick guide for healthcare professional educators. J Med Educ Curric Dev 2019;6:2382120519840332 [FREE Full text] [doi: 10.1177/2382120519840332] [Medline: 31008257]
35. Maurana CA, Fritz JD, Witten AA, Williams SE, Ellefson KA. Advancing flourishing as the north star of medical education: a call for personal and professional development as key to becoming physicians. Med Teach 2024;46(12):1539-1543. [doi: 10.1080/0142159X.2024.2412791] [Medline: 39481004]
36. Ghosh SK. Cadaveric dissection as an educational tool for anatomical sciences in the 21st century. Anat Sci Educ 2017;10(3):286-299. [doi: 10.1002/ase.1649] [Medline: 27574911]
37. Čavka M, Petanjek Z, Zanze Beader M, Mathew T, Brassett C, Nagy AD, et al. Balancing tradition and innovation: rethinking the dichotomy in anatomy teaching. Croat Med J 2025;66(1):62-70 [FREE Full text] [doi: 10.3325/cmj.2025.66.62] [Medline: 40047163]
38. Kranzlein J, Sharpe B, Robinson M, Burbank E, Pearson WG. Repurposing the anatomy lab: a developmental model for physician formation and human flourishing. Anat Sci Educ 2025;18(9):897-904. [doi: 10.1002/ase.70103] [Medline: 40717029]

Abbreviations

IRB: Institutional Review Board
MCAT: Medical College Admission Test
sGPA: science undergraduate grade point average
uGPA: undergraduate grade point average

Edited by J Gentges; submitted 29.Sep.2025; peer-reviewed by M Pascoe, A Al-Redouan; comments to author 03.Nov.2025; accepted 08.Dec.2025; published 13.Jan.2026.

Please cite as:
Sumner C, Case SL, Franklin S, Platt K
Interactive, Image-Based Modules as a Complement to Prosection-Based Anatomy Laboratories: Multicohort Evaluation
JMIR Med Educ 2026;12:e85028
URL: https://mededu.jmir.org/2026/1/e85028
doi:10.2196/85028
PMID:

©Caroline Sumner, Sami L Case, Samuel Franklin, Kristen Platt. Originally published in JMIR Medical Education (https://mededu.jmir.org), 13.Jan.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Education, is properly cited. The complete bibliographic information, a link to the original publication on https://mededu.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Disaster Medicine Training for Medical Students in Lebanon: Quasi-Experimental Comparison of e-Learning and Face-to-Face Modalities

Ali Msheik[1,2], MD, MSc; Ruben Peralta[3,4], MD, MAMSE; Zeinab Al Mokdad[5], MSc; Christine Bartulec[6], MD; Linda Abou-Abbas[6,7], PhD; Dima Anani[1], MD; Hussein Al Mawla[1], MD; Hassan Salame[8], MD; Mohamad Moussa[1], MD; Awsan Bahattab[9], MD, PhD

[1]Faculty of Medical Sciences, Lebanese University, Beirut, Beyrouth, Lebanon

[2]Neurological Surgery, Neuroscience Institute, Hamad Medical Corporation, Doha, Qatar

[3]Trauma Surgery, Hamad Medical Corporation, Doha, Qatar

[4]Department of Surgery, Universidad Nacional Pedro Henríquez Ureña, Santo Domingo, Dominican Republic

[5]Medical Ethics, Faculty of Medical Sciences, Lebanese University, Beirut, Lebanon

[6]International Committee of the Red Cross, Beirut, Lebanon

[7]Institut National de Santé Publique, Epidémiologie Clinique et Toxicologie-Liban, Beirut, Lebanon

[8]Head of the Program Evaluation Committee, Faculty of Medical Sciences, Lebanese University, Beirut, Lebanon

[9]Center for Research and Training in Disaster Medicine, Università degli Studi del Piemonte Orientale “Amedeo Avogadro”, Piemonte, Italy

Corresponding Author:
Ali Msheik, MD, MSc
Neurological Surgery
Neuroscience Institute
Hamad Medical Corporation
Doha
3050
Qatar
Phone: 974 72157731
Email: dr.alimsheik@gmail.com

Abstract

Background: Despite global advocacy for its integration into medical curricula, disaster medicine (DM) education remains underdeveloped, especially in fragile settings where such training is urgently needed. In Lebanon, a country facing political and economic crises, students face significant barriers to in-person education.

Objective: This study aims to evaluate the effectiveness of e-learning versus face-to-face (F2F) approaches in improving knowledge retention and provides insight into the practical considerations of implementing DM courses in such settings.

Methods: This quasi-experimental study used a Solomon 4-group design to evaluate e-learning and F2F DM courses for second- to fifth-year medical students at the Lebanese University. A total of 205 participants, stratified by academic year, were divided between the 2 modalities. Knowledge was assessed before the course, after the course, and at 1-month follow-up. Confidence, competency, and satisfaction were evaluated after the course using validated tools.

Results: Of 205 participants, 56.6% (n=116) favored e-learning. Both modalities improved knowledge and knowledge retention, with no significant difference between the 2 groups. Fifth-year students achieved the highest gains in knowledge, particularly in the e-learning group. Similarly, no significant difference in satisfaction was observed across modalities, although F2F was preferred overall, except among fifth-year students, who preferred e-learning. Confidence levels were also similar across both modalities, but F2F scored higher for skills like triaging. Feedback emphasized the relevance of the course and advocated for integration of DM into the medical curriculum, and adding practical sessions.

Conclusions: Integrating DM education into the fifth-year curriculum, prior to externship, can enhance preparedness and promote knowledge retention and application in real-world disaster settings. The study shows that e-learning is an effective modality for improving knowledge acquisition and retention in DM. Although feasibility and efficiency were not measured directly, the successful delivery of the course to geographically dispersed students suggests potential practical advantages. Combining F2F

XSL•FO
RenderX

practical sessions for specific targeted topics in a blended curriculum is recommended to further enhance medical students' confidence for future disaster response. These findings support broader policy efforts to institutionalize DM in medical curricula, particularly in fragile and resource-limited settings.

(JMIR Med Educ 2026;12:e80409) doi:10.2196/80409

KEYWORDS

disaster medicine; education, medical, undergraduate; educational technology; teaching methods; program evaluation; knowledge; surveys and questionnaires; self-efficacy; Lebanon

Introduction

The need for disaster medicine (DM) training in medical education has been recognized since the 1970s, but its integration into the curricula has remained limited until the early 2000s, when DM was incorporated into the curricula of US medical schools. Since then, international frameworks and associations such as the Sendai Framework for Disaster Risk Reduction [1], and the Association of American Medical Colleges and the World Association for Disaster and Emergency Medicine have called for the incorporation of DM and mass casualty incident education in medical schools using scientific evidence-based and flexible approaches [2,3]. Despite these recommendations, limited efforts have been made by universities to implement such training [2].

The lack of DM education has left medical students unprepared to participate in disaster response [4,5]. In contrast, those who have participated in elective DM courses have reported positive results, with high satisfaction, and improved understanding, knowledge, attitudes, and skills toward disaster response [2,6-9]. Still, evaluating these courses remains complicated, due to a lack of standardization across curricula and assessment methods [10,11]. Furthermore, the availability of such training is concentrated in high-income countries and Global North institutes, with limited accessibility for global participation, particularly in fragile and conflict-affected countries, which disproportionately experience higher frequency and impact of disasters [12,13].

Lebanon, home to 5.4 million people and hosting over a million refugees, presents an urgent need for DM education. The country is facing repeated manmade and public health disasters compounded by political instability and an economic crisis, which highlighted the country's fragile health care system [12,14-18]. Yet, undergraduate medical curricula still lack DM education, which remains limited to postgraduate in-service training for practitioners and emergency physicians only. Incorporating DM as a component of the medical curricula constitutes a critical strategy in medical education in Lebanon.

The Lebanese University Faculty of Medical Sciences (LUFMS) is the only public medical school and serves many low-income students all over the country, with around 130 graduates annually. Due to ongoing hostilities and the economic crisis, many low-income students from remote areas are unable to afford the money required to travel to and from campus. The recent COVID-19 pandemic has led to the development of e-learning alternatives that have made it even more accessible to students from these areas. e-learning can provide a cost-effective, equitable, and even practical and interactive solution for DM education [18-24]. However, implementation of such training in crisis-affected settings poses operational challenges [25,26].

This study aimed to evaluate the effectiveness of e-learning versus face-to-face (F2F) approaches in DM education for improving knowledge acquisition and retention among medical students. DM education was delivered through both e-learning and F2F modalities because DM is not included in the undergraduate medical curriculum at LUFMS, and no standardized format for its delivery exists. Evaluating both modalities was therefore essential to determine the most feasible, equitable, and effective approach for nationwide implementation in a context where financial constraints, transportation barriers, and recurrent disruptions frequently hinder in-person attendance. This study also assessed students' self-perceived confidence, competence, satisfaction, and engagement to inform the acceptability, effectiveness, and feasibility of implementing such courses in fragile or disaster-prone settings.

Methods

DM Course Development

The online course used the International Committee of the Red Cross (ICRC) DM e-learning tool. This e-learning tool was an existing, predeveloped ICRC course and was not created by the authors for this study. The development of the ICRC tool was based on the standard principles that are used by multiple courses and learning resources, including the major incident medical management and support course, World Health Organization, and Médecins Sans Frontières, with each module referencing these standards as relevant [27-36]. Development of the F2F course was based on the course contents of the e-learning tool. Both e-learning and F2F DM courses used in this study were based on 10 core topics (Table 1). The training is designed to promote effective learning through interactive, experience-based methods that encourage active engagement and reflection. This approach is informed by constructivist theory, which emphasizes that learners build knowledge through meaningful experiences. To support motivation and a sense of autonomy, the training incorporates principles from self-determination theory, offering flexible learning options that allow individuals to take ownership of their learning process. Additionally, the structure of the content is guided by cognitive load theory, ensuring that information is presented in a clear and manageable way to enhance understanding and retention. The design of the course schedule incorporated a degree of flexibility to accommodate institutional curriculum requirements and the blended e-learning modality used. Of note, principles

XSL•FO
RenderX

of responder and scene safety, including self-safety, hazard identification, and ensuring a safe environment before initiating triage or patient care, were incorporated across multiple modules, particularly the introduction to DM, prehospital disaster management, and triage modules, consistent with ICRC, World Health Organization, Médecins Sans Frontières, and major incident medical management and support course standards.

Table 1. Topics and subtopics included in the courses in this study.

Topic	Subtopic
Introduction to disaster medicine and taxonomy	• Disaster medicine definitions "disaster" and "MCI[a]" • Difference between DM[b] and EM[c] • Public health principles in disasters • Disaster management phases: 3 phases • The concept of "hazard vulnerability analysis"
Triage in disasters	• MCI triage definitions • Different triaging systems • EM versus DM triaging
Hospital disaster preparation and response	• Laws and DM • Internal vs external hospital incidents • What is an MCI plan? • Integration of medical staff
Health consequences of different disasters	• Types of disasters and the impact of each • Natural vs man-made disasters: the health impact
Prehospital disaster management	• MCI: disposition, treatment, and transport • Disaster plans and the control chain • Functional response rate
Psychosocial care	• Psychic reactions and disaster exposure • Treatment of incident stress reactions
Presentation of past disasters and public health emergencies	• Beirut post explosion • COVID-19 pandemic
Pediatric injuries in MCI	• Trauma assessment and triage • Pediatric-specific injuries • Psychological and emotional support • Pediatric resuscitation
MCI and burn injuries	• Burn severity index • Burn resuscitation • Wound care and dressing • Pain management • Psychological support
Management of the dead bodies after disasters	• Disaster victim identification • Forensic pathology and autopsy • Temporary mortuary facilities • Body handling • Documentation • Cultural and religious considerations • Public health and hygiene

[a]MCI: mass casualty incident.

[b]DM: disaster medicine.

[c]EM: emergency medicine.

Study Design

This is a quasi-experimental controlled study using a Solomon 4-group design [37]. Quasi-experimental designs are appropriate when random assignments are impractical, allowing for the evaluation of interventions in real-world settings while maintaining a degree of control over confounding variables [38]. In the Lebanese context, constraints such as transportation difficulties, financial limitations, and geographical barriers rendered random allocation to e-learning and F2F groups unfeasible. Therefore, participants were assigned based on their ability to attend on-campus sessions, considering factors like

XSL•FO
RenderX

proximity and access to reliable transportation rather than their personal preference or academic performance, and selection bias related to motivation was minimal. The design balances trade-offs between internal and external validity, feasibility, inclusiveness, and applicability in educational real-life settings in resource-constrained environments [39]. Furthermore, the Solomon 4-group design enhances the robustness of the study by controlling confounding variables and potential pretest sensitization effects [37].

Target Population

The target population consisted of medical students between the second and fifth academic years. All students from the second to the fifth academic years received identical course content, as DM is not part of the formal curriculum at any level. Standardizing the content ensured comparability across groups and isolated the effect of the delivery modality rather than differences in academic year. No interns were included in this study.

Sampling

The total number of the targeted population was 512. The sample size was calculated using G*Power software to detect the effect size (Cohen *d*=0.4) for the primary outcome (posttraining effectiveness difference between e-learning and F2F, measured by the significant difference in knowledge scores between the groups), with a significance level of .05, a desired power of 0.80, and a 2-tailed test. The estimated minimum sample size was 174 participants (87 in each group). The population count per group was determined with the assistance of delegates from each year for the medical students via convenience sampling, such as open registration on a first-come, first-served basis, regardless of their academic year. Group assignments were determined based on students' logistical capabilities. Those with the ability to attend on-campus sessions, considering factors such as proximity to campus, access to reliable transportation, and affordability, were assigned to the F2F cohort. Conversely, students facing logistical barriers that precluded on-campus attendance were placed in the e-learning cohort.

DM Course Implementation

The study involved delivering a DM course using identical educational materials through 2 modalities: e-learning and F2F sessions. The course was delivered over 3 days per week for 2 weeks. The F2F sessions were implemented by the principal investigator at LUFMS. All synchronous e-learning sessions were also delivered by the principal investigator to ensure standardization across modalities. The digital version for the e-learning groups was a set of modules built onto a learning management system, which was delivered to the respective subgroups in a synchronized fashion by the principal investigator, with the lectures on the same dates to the assigned subgroups. Both groups received a digital version of the respective teaching material at the end of each session. while the F2F group received the slides for the respective session.

Evaluation Framework and Outcomes

The first 2 levels of Kirkpatrick's model were used to assess knowledge, knowledge retention, learning, and satisfaction. Additionally, self-efficacy, as a predictor of skill application, was measured through perceived behavioral outcomes, specifically confidence (self-perceived competence), and was guided by social cognitive theory [40].

1. Assessment of the students' knowledge: Students' knowledge assessment consisted of 20 multiple-choice questions (Table 2). The 20-item multiple-choice questions assessment included 2 questions per core topic, as summarized in Table 2; the full questionnaire is provided in Multimedia Appendix 1. Each question had 5 possible answers, and only 1 answer was correct.

Based on Solomon 4-group design model, students within each year and learning modality were divided randomly into 2 groups [37]. Each academic year was subdivided into 4 subgroups (A, B, C, and D; Table 3).

XSL•FO
RenderX

Table 2. Multiple-choice question topics per core subject.

Topic	Multiple choice questions
Introduction to disaster medicine and taxonomy	• Question 1: Which statement is correct about DM[a]?
Triage in disasters	• Question 2: What is the purpose of triaging in a medical emergency? • Question 3: In the triage system, which color is typically associated with the highest priority level?
Hospital disaster preparedness and response	• Question 4: What is the role of MOH[b] in DM? • Question 5: What is an MCI[c] plan? • Question 6: Role of hospitals and prehospital units in preparedness • Question 7: Which means of communication is most important in DM?
Health consequences of different disasters	• Question 8: Which type of disaster affects the respiratory system? • Question 9: Which statement is correct?
Prehospital and hospital disaster management	• Question 10: Who controls during a disaster? • Question 11: What is the role of an incident commander? • Question 12: In an MCI, who is typically responsible for declaring a patient deceased? • Question 13: What is the role of hospitals in disaster preparedness?
Psychosocial care	• Question 14: What does PTSD[d] stand for? • Question 15: What are treatment modalities for stress during disasters?
Presentation of past disasters and public health emergencies	• Question 16: Recall the hospitals damaged during the incident • Question 17: What is COVID-19 pandemic most significant impact worldwide?
Pediatric injuries in MCI	• Question 18: In an MCI involving pediatric injuries, which of the following is essential to identify first?
MCI and burn injuries	• Question 19: In the context of mass casualty incidents involving burn injuries, which aspect is essential for initial assessment?
Management of the dead bodies after disasters	• Question 20: When managing dead bodies after disasters, which of the following should be prioritized first?

[a]DM: disaster management.

[b]MOH: Ministry of Health.

[c]MCI: mass casualty incident.

[d]PTSD: posttraumatic stress disorder.

Table 3. Subgroup classification of the patients within each academic year and the timing of the assessments based on Solomon 4-group design model.

Time points	Medical students: second-fifth year (subgroups A-D)
First week	Phase I: precourse assessment: Subgroups A and B had face-to-face sessions, and subgroups C and D had e-learning in the first and second weeks
Second week “end of course”	Phase II: postcourse assessment online survey for all subgroups
1 month post course completion	Phase III: assessment for all subgroups

The assessment was done over 3 phases:

- Precourse assessment: conducted 15 minutes before the course for subgroups A and C only.
- Postcourse assessment: conducted after the end of the final session for all subgroups. The pre- and post-assessments used the same questionnaire details.
- Follow-up assessments: conducted 1 month after the finalization of the DM educational course to evaluate knowledge retention.

1. Assessment of the satisfaction and recommendations of students: the assessment of satisfaction was done using a tool developed by Han et al [38]. The tool consisted of questions with a rating scale (Multimedia Appendix 2) [35,37]:
   - “Reaction” of the participants: how they felt and thought about the course
   - “Learning” of the participants: how they described the increase in knowledge after the course.

In addition, 3 open questions addressed the experience of the participants. The first question addressed the most unique aspect of the course. The second question addressed the suggestions

XSL•FO
RenderX

of the participants for improvement of the course. The last question requested additional comments from the participants.

1. Evaluation of self-efficacy: a postcourse evaluation of confidence was conducted for all subgroups in each academic year. The evaluation of the confidence level was measured using a validated Disaster Preparedness Evaluation Tool [38,39]. The original version was adapted to the educational material implemented in the course (Multimedia Appendix 3) [35,36]. Students will be asked their level of confidence on a Likert scale as "strongly disagree," "disagree," "neutral," "agree," and "strongly agree."

Data Collection Tools and Methods

The student's ID, their academic year, and evaluation data were confidentially collected using the Google Forms survey tool, which was converted to a Microsoft Excel sheet and saved anonymously (Multimedia Appendix 4). The handling and access of the data were limited to the principal investigator to ensure data security. All data were stored in accordance with institutional data protection policies.

Reliability Assessment

A post hoc reliability analysis check of the scale used to assess the knowledge, satisfaction, and confidence level showed excellent internal consistency (Cronbach α was 0.82, 0.92, and 0.963, respectively).

Data Analysis

Quantitative Analysis

Descriptive statistics, including percentages, means, and SD values, were used to describe the sample and summarize survey scores. Knowledge scores were reported as the number of correct responses. Repeated measures ANOVA, also referred to as a within- and between-subjects ANOVA for correlated samples, was used to detect any overall differences between related means at different points in time. Chi-square and Fisher exact tests were used to evaluate differences in the items of the satisfaction scale and confidence scale across the different options. An independent-sample *t* test was used to compare satisfaction and confidence scores in the overall sample and stratified by year of education between the 2 learning modalities. Data were analyzed using the Statistical Package for the Social Sciences (version 27.0; IBM Corp). A *P* value less than .05 was considered significant.

Qualitative Analysis

Open question responses were inductively analyzed and coded, and recurrent themes were presented in percentages.

Ethical Considerations

This project was conducted according to the applied ethical guidelines and received the necessary approvals from the relevant ethical review boards. The study was approved by the Institutional Review Board of Al Zahraa Hospital University Center under reference number 12/2023. Additionally, the study received an exemption from the Ethics Review Board of the ICRC in Geneva. The ICRC Ethics Review Board confirmed that the research proposal presented minimal risks to participants, which were adequately addressed and minimized through the proposed procedures. The exemption reference number for this study is 0623. Informed consent was obtained from all participants, and strict confidentiality was maintained in the handling of all data to ensure participants' privacy and well-being throughout the study.

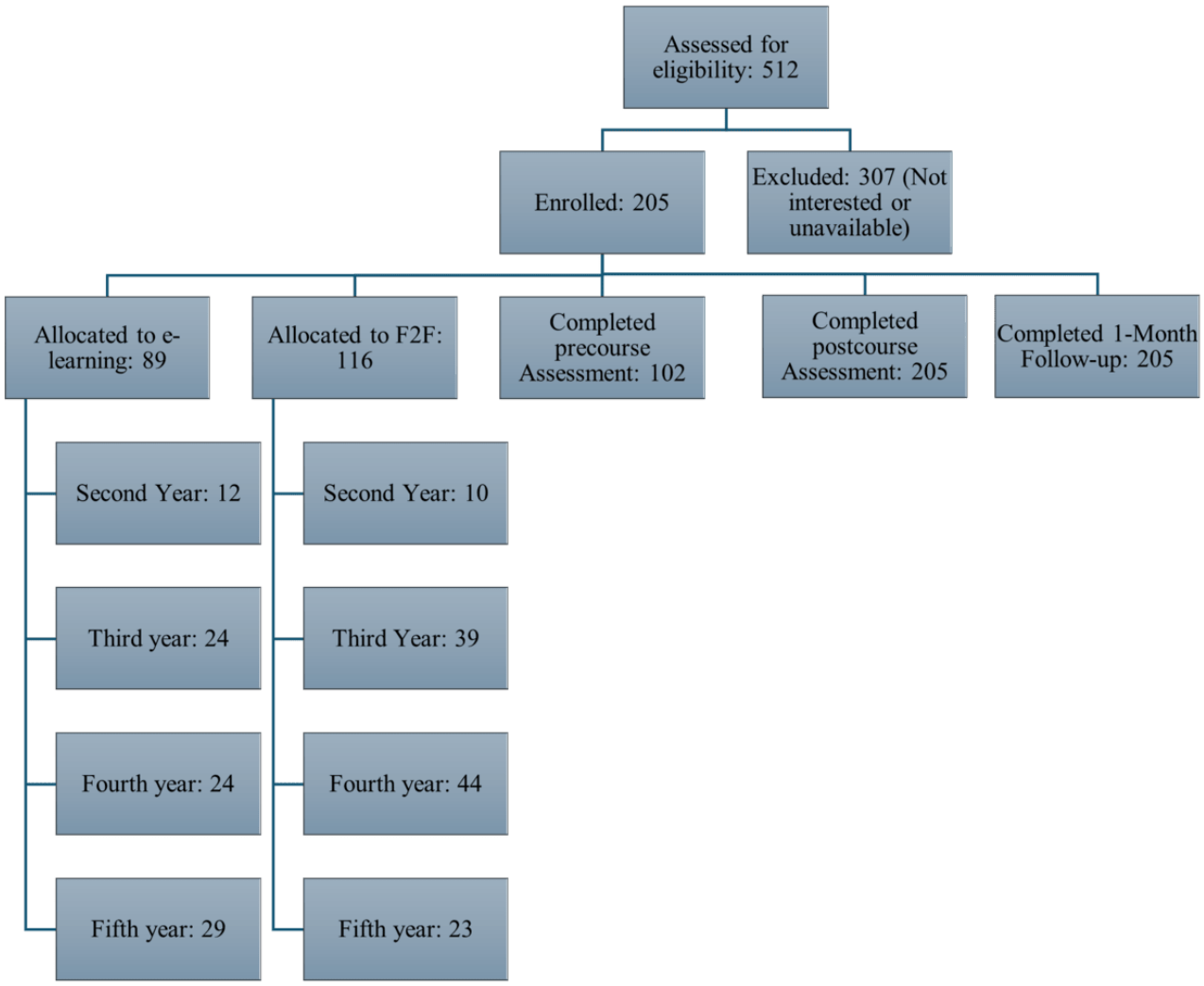
Results

Descriptive Statistics

Participants were recruited, and the course was conducted between March 2024 and June 2024. The participant flowchart is presented in Figure 1. Out of 512 students at LUFMS from the second to the fifth academic years, 205 (40%) voluntarily participated in the study, of which 116 (56.6%) joined the e-learning cohort, and 89 (43.4%) joined the F2F cohort between March and June 2024. The third and fourth academic year students represented 30.7% (63/205) and 33.2% (68/205) of the participants, respectively. Fifth academic year students were 25.4% (52/205), and the second academic year students were the least (22/205, 10.7%). The distribution of students according to the crosstabulation of the academic year and learning modality is depicted in Figure 1. Participation rates from the second and fifth academic years were comparable across the F2F and e-learning groups, with 12 (13.9%) and 10 (8.6%) students from the second year, and 29 (32.5%) and 23 (19.8%) students from the fifth year, respectively. Nearly 61.9% (39/63) of the third academic year students and 64.7% (44/66) of the fourth academic year students joined the F2F cohort.

XSL•FO
RenderX

Figure 1. Participant flow diagram showing distribution by academic year and learning modality (F2F vs e-learning). F2F: face-to-face.



Assessment of Precourse Sensitization

Based on Solomon 4-group design model [37], a summarized comparison for the overall sample is presented in Table 4. The analysis shows that pretested and non-pretested groups demonstrated comparable scores at baseline, at the immediate postcourse assessment, and at 1-month follow-up (*P*=.60, .40, and .64, respectively). These results confirm that the precourse assessment did not introduce sensitization effects or bias the learning outcomes, thereby supporting the internal validity of the study design. Hence, precourse assessment is not a confounding factor and did not affect the outcome across the subgroups.

Table 4. Summary of Solomon 4-group analysis for the overall sample.

Overall	Pretest	Posttest	Post 1 month
F2F[a]			
Experimental, mean (SD)	6.3 (2.3)	12.3 (1.8)	14.3 (2.4)
Control, mean (SD)	—[b]	12.2 (1.7)	15.7 (3.9)
e-Learning			
Experimental, mean (SD)	6.6 (2.3)	12 (2.1)	14.6 (3.5)
Control, mean (SD)	—	13.5 (3.7)	16.4 (4.4)
P value between groups	.60	.40	.64

[a]F2F: face-to-face.

[b]Not applicable.

XSL•FO
RenderX

Assessment of the Knowledge Level in the Overall Sample and Across the Academic Years

For the overall sample, both F2F and e-learning groups demonstrated significant improvements from precourse to immediate postcourse and post 1 month, with no significant statistical difference between both modalities overall (*P*=.40 and .64) and across each year (Multimedia Appendix 5). The F2F cohort showed a mean increase of 5.9 (95% CI 5.35-6.45; *P*<.001) and 7.9 (95% CI 7.34-8.46; *P*<.001) from precourse to postcourse and post 1 month, respectively (Table 5). Similarly, the e-learning group exhibited a mean increase of 5.4 (95% CI 4.69-6.11; *P*<.001) and 8 (95% CI 7.11-8.89; *P*<.001) over the same periods. In the second year, the F2F group showed a significant increase of 5.6 (95% CI 3-8.2; *P*=.04) and 7.6 (95% CI 5.68-9.52; *P*=.002) from precourse to postcourse and post 1 month, respectively. However, the e-learning group did not show significant improvement, with a mean increase of 4 (95% CI 0.83-7.17; *P*=.74) and 7 (95% CI 4.57-9.43; *P*=.14), indicating less effectiveness in the second year. Of note, all mean change values were calculated at the individual level and then averaged within each subgroup; no pooled scoring was used.

Table 5. Mean change in knowledge scores from precourse to postcourse and 1-month follow-up.

Cohort	Sample size	Postcourse, mean (SD)		1-month follow-up, mean (SD)		*P* value
		Mean (SD)	95% CI	Mean (SD)	95% CI	
F2F total[a]	116	5.9 (3)	5.35-6.45	7.9 (3.1)	7.34-8.46	<.001
e-learning total	89	5.4 (3.4)	4.69-6.11	8 (4.3)	7.11-8.89	<.001
Second-year F2F	10	5.6 (4.2)	3-8.20	7.6 (3.1)	5.68-9.52	.002
Second-year e-learning	12	4 (5.6)	0.83-7.17	7 (4.3)	4.57-9.43	.14
Third-year F2F	39	5.3 (2.1)	4.64-5.96	7.3 (2.3)	6.58-8.02	<.001
Third-year e-learning	24	5.8 (3.2)	4.52-7.08	8.3 (4.3)	6.58-10.02	<.001
Fourth-year F2F	44	7.1 (3.6)	6.04-8.16	9 (3.9)	7.85-10.15	.003
Fourth-year e-learning	24	4.9 (3.3)	3.58-6.22	7.2 (4.8)	5.28-9.12	<.001
Fifth-year F2F	23	6.1 (2.7)	5-7.20	8.2 (3.4)	6.81-9.59	<.001
Fifth-year e-learning	29	6.1 (3)	5.01-7.19	10.1 (2.6)	9.15-11.05	<.001

[a]F2F: face-to-face.

The fifth-year students exhibited the highest mean score increases within their respective learning modalities. The F2F group showed a mean increase of 6.1 (95% CI 5-7.2; *P*=.005) and 8.2 (95% CI 6.81-9.59; *P*<.001) from precourse to postcourse and post 1 month, respectively. The e-learning group showed a comparable increase of 6.1 (95% CI 5.01-7.19; *P*=.005) and 10.1 (95% CI 9.15-11.05; *P*<.001) over the same periods.

Assessment of the Satisfaction Level in the Overall Sample and Across the Academic Years

The results show that the total satisfaction score (SD) was slightly higher for the F2F cohort (mean 87.6, SD 11.3; 95% CI 85.54-89.66) compared to e-learning (mean 86, SD 14.2; 95% CI 83.05-88.95). Only the e-learning cohort of the fifth academic year students showed a higher score (mean 90.2, SD 11.7; 95% CI 85.94-94.46) compared to F2F (mean 87.5, SD 9.6; 95% CI 83.58-91.42; Table 6). No statistically significant differences in satisfaction scores were observed between F2F and e-learning modalities in the overall sample (*P*=.38) or within any academic year subgroup.

XSL•FO
RenderX

Table 6. Mean change in the satisfaction score compared to the precourse assessment: at the postcourse assessment, directly postcourse.

Cohort	Postcourse assessment, mean (SD)	Sample size, n	95% CI	*P* value[a]
F2F[b] total	87.6 (11.3)	116	85.54-89.66	.38
e-learning total	86 (14.2)	89	83.05-88.95	—[c]
Second-year F2F	87.6 (15)	10	78.3-96.9	.81
Second-year e-learning	86.2 (11.32)	12	79.8-92.6	—
Third-year F2F	87.5 (11.5)	39	83.89-91.11	.51
Third-year e-learning	84.9 (16.65)	24	78.24-91.56	—
Fourth-year F2F	87.6 (11.5)	44	84.2-91	.35
Fourth-year e-learning	84.5 (13.6)	24	79.06-89.94	—
Fifth-year F2F	87.5 (9.6)	23	83.58-91.42	.37
Fifth-year e-learning	90.2 (11.7)	29	85.94-94.46	—

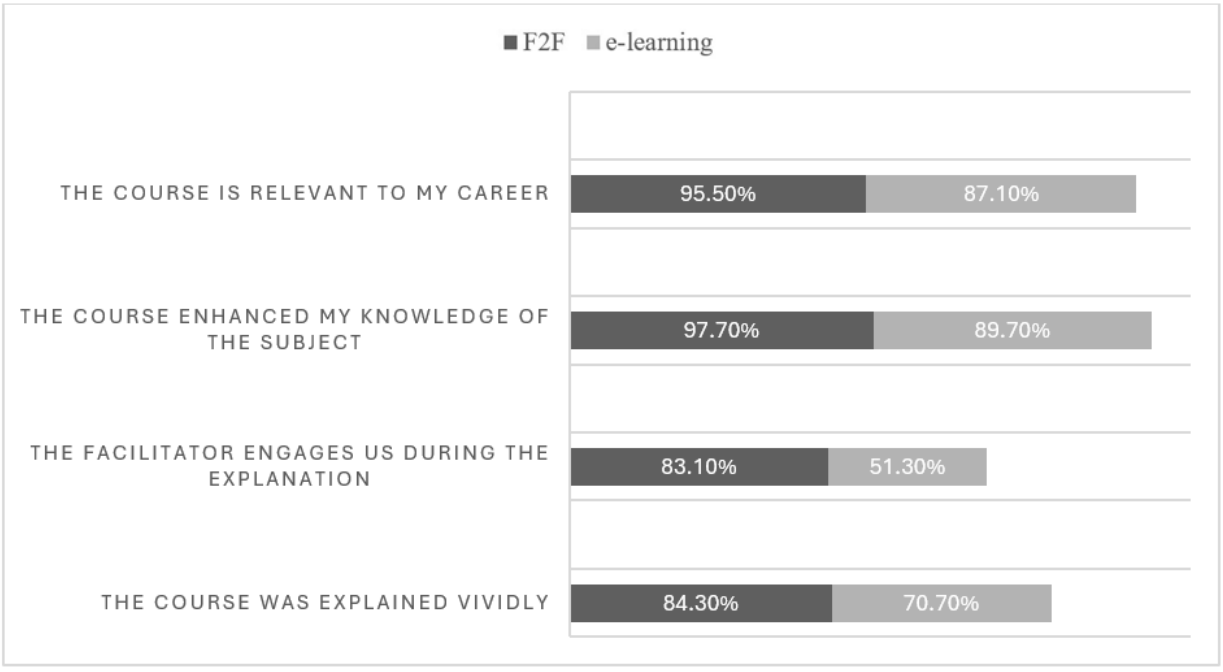
[a]*P* values calculated using independent-sample Welch *t* tests.

[b]F2F: face-to-face.

[c]Not available.

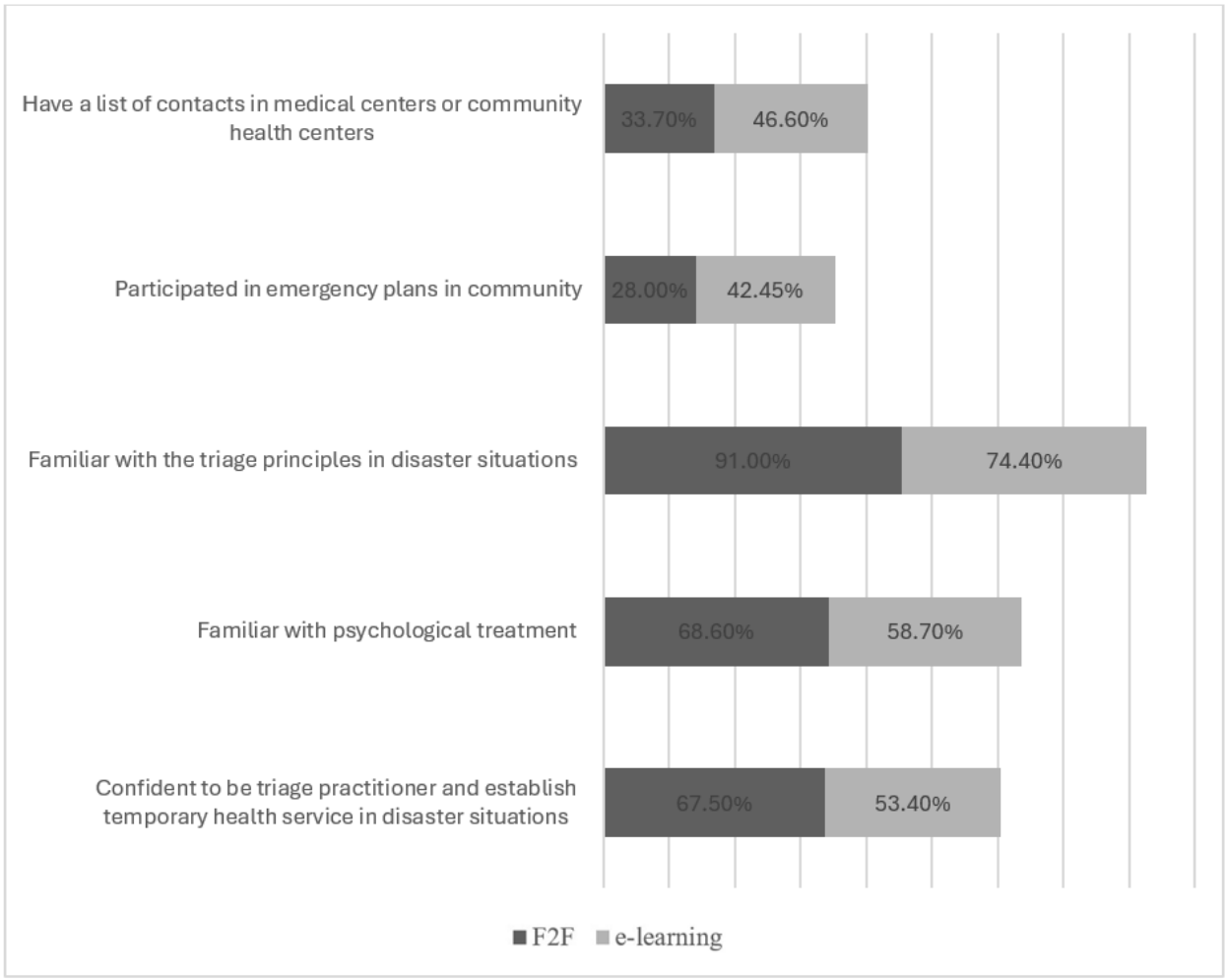
Although the comparison of the total satisfaction between F2F and e-learning showed statistically nonsignificant *P* values, comparison of the individual satisfaction scores of each question between the 2 modalities was statistically significant for only 4 questions in favor of F2F (Figure 2). Students were mostly satisfied with the vividness of the course and its relevance to their medical career, enhancement of their knowledge of disaster management, and their engagement during the course (Figure 3).

Figure 2. Horizontal bar graph showing the percentage of satisfaction of the participants with each learning modality in the statistically significant questions. F2F: face-to-face.



XSL•FO
RenderX

Figure 3. Horizontal bar graph showing the percentage of confidence of the participants with each learning modality in the statistically significant questions. F2F: face-to-face.



Assessment of the Confidence Level in the Overall Sample and Across the Academic Years

The results show that the total confidence score (SD) was similar between F2F (mean 125.5, SD 20.6; 95% CI 121.75-129.25) and e-learning (mean 126.5, SD 23.8; 95% CI 121.56-131.44). Only the F2F cohort of the second academic year students showed scores that were like the overall score, yet better than e-learning, with a mean F2F score of 125.4 (SD 22.5; 95% CI 111.45-139.35) compared to a mean of 116 (SD 13.9; 95% CI 108.14-123.86) for e-learning (Table 7). The e-learning cohort of the fifth academic year students showed higher scores compared to F2F, with a F2F score of mean 124.9 (SD 20.1; 95% CI 116.69-133.11) compared to mean 135.3 (SD 21.71; 95% CI 127.4-143.2) for e-learning. No statistically significant differences in confidence scores were observed between learning modalities in the overall sample (P=.75). A marginal trend favoring the e-learning group was noted in the fifth-year subgroup (P=.08), although it did not reach statistical significance.

XSL•FO
RenderX

Table 7. Mean change in the confidence score compared to the precourse assessment: at the postcourse assessment, directly postcourse.

Cohort	Postcourse assessment, mean (SD)	Sample size, n	95% CI	*P* value[b]
F2F[a] total	125.5 (20.6)	116	121.75-129.25	.75
e-learning total	126.5 (23.8)	89	121.56-131.44	—[c]
Second-year F2F	125.4 (22.5)	10	111.45-139.35	.27
Second-year e-learning	116 (13.9)	12	108.14-123.86	—
Third-year F2F	124.7 (23.1)	39	117.45-131.95	.99
Third-year e-learning	124.6 (28.4)	24	113.24-135.96	—
Fourth-year F2F	127 (18.7)	44	121.47-132.53	.70
Fourth-year e-learning	125 (21.1)	24	116.56-133.44	—
Fifth-year F2F	124.9 (20.1)	23	116.69-133.11	.08
Fifth-year e-learning	135.3 (21.71)	29	127.4-143.2	—

[a]F2F: face-to-face.

[b]*P* values calculated using independent-sample Welch *t* tests.

[c]Not available.

Although the comparison of the total confidence between F2F and e-learning showed no statistically significant *P* values (Table 7, the comparison of the individual confidence score of each question between the 2 modalities was statistically significant for only 5 questions. Students were mostly confident about their knowledge of triaging principles in a disaster through F2F learning (105/116, 91%) versus e-learning (66/89, 74.4%). Similarly, students expressed a higher familiarity with psychological treatment and the ability to triage patients in disaster situations through F2F learning (80/116, 68.6% and 78/116, 67.5%, respectively) relative to e-learning (52/89, 58.7% and 48/89, 53.4%, respectively). The confidence level was below 50% for either modality when students were asked about participation in community emergency plans and having contacts in community health centers.

Unique Aspect of the Course

The participants highlighted the distinctiveness of the course. Around 85% (174/205) of the participants appreciated the relevance of the course to real-life situations and disaster management. Nearly 75% (153/205) of the participants found engagement with a variety of topics, including mass casualty incidents, triaging, and the management of specific disasters like white phosphorus. The innovative teaching methods and the adaptation of the course to the Lebanese situation were praised by 70% (143/205) of the participants, while 65% (133/205) valued the case studies that helped to effectively manage the disasters.

Areas of Improvement

Around 60% (123/205) of the participants suggested the inclusion of interactive and practical sessions to reinforce through hands-on practice. Nearly 55% (112/205) of the participants suggested improvement of the scheduling of the sessions, while some proposed that it should become a mandatory part of the curriculum. Half of the participants noted that the artificial intelligence voice used in the e-learning course was mildly distracting. Around 45% (92/205) suggested that live sessions would enhance the course's effectiveness, especially for complex medical terms and scenarios.

Additional Comments

About 80% (164/205) of the participants expressed appreciation for the course and gratitude for the opportunity to learn about disaster management. Nearly 70% (143/205) recognized the importance of the course, considering the recent events in Lebanon. A desire for the incorporation of similar courses into the curriculum was noted by 65% (133/205) of the participants. There were also suggestions from 55% (112/205) of participants to expand the course to include more practical exercises, community-based projects, and timely distribution of course materials to ensure better knowledge retention.

Discussion

Principal Findings

This study evaluated whether e-learning is comparable to F2F instruction in DM training for LUFMS medical students and contributes to a growing body of evidence supporting the effectiveness of e-learning [40-45]. The results demonstrated significant improvements in knowledge and retention for both modalities, with no statistical difference between them (*P*=.40). However, F2F training resulted in higher confidence in practical skills like triage (91% vs 74.4%), while e-learning resulted in superior retention among fifth-year students. Satisfaction was high in both groups, although qualitative feedback emphasized the need for more hands-on sessions (60%). The results have direct implications for the delivery of DM education and the expectation of outcomes, especially in countries of protracted conflict like Lebanon [46,47]. A blended DM training approach with the skill-building benefits of targeted F2F sessions offers the most pragmatic solution. Still, the equivalence in knowledge gains supports e-learning as a viable alternative in resource-limited settings.

While both delivery methods were effective across academic years, the higher mean gains in knowledge and retention

XSL•FO
RenderX

observed among fifth-year e-learning students can be attributed to their advanced medical knowledge compared to students in the preceding academic years. The LUFMS curriculum typically incorporates diagnostic courses within the fifth year, and students commonly undertake a 6-week externship after the fifth academic year. This confluence of advanced knowledge and practical experience may have motivated the fifth-year students to excel in the assessments, recognizing the opportunity to apply their learning in real-life and clinical settings [6]. Additionally, the fifth academic year students were the only group who reported satisfaction and confidence levels with e-learning higher compared to F2F due to their 2-year antecedent exposure to online learning. Given their advanced medical knowledge and diagnostic expertise, the students' ability to integrate prior knowledge with new concepts acquired through e-learning and independent research likely played a significant role in their favorable outcomes.

Both modalities showed low confidence in community emergency plan participation and health center contacts. This can be attributed to the fact that the participants have not engaged in hospital situations and do not interact with hospital staff except their academic mentors. Therefore, they have limited knowledge about the emergency plans in hospitals. This suggests DM implementation should be timed with clinical rotations to maximize its relevance and effectiveness.

Limitations

This study has several limitations, which stem from the difficulty of conducting similar studies in a volatile context such as Lebanon, a country with a history of disasters and emergencies, which also impacts the educational settings. Nevertheless, the findings are particularly relevant and have direct application and impact in practice. Voluntary recruitment and the lack of random assignments of the participants may have compromised internal validity, although the real-world applicability, along with sufficient sample size, enhances the study's external validity, especially for comparable settings. Furthermore, the application of the Solomon 4-group design model classification of the participants allowed for control confounding and pretest sensitization [37], enhancing the internal validity of the study. The short follow-up period limits the conclusion about long-term retention. Still, the significant increase in knowledge over time suggests the consolidation of learning over time due to its real application in their setting. Finally, the reliance on self-reported learning and confidence can introduce bias, but the results were also triangulated by qualitative data and objective assessment with a validated tool [48-51], which also enhances the applicability and internal validity of the study.

Although "lack of interest" was the most frequently selected reason for nonparticipation, this explanation likely underrepresents the true contextual barriers faced by students. Informal feedback indicated that transportation costs, fuel shortages, intermittent electricity and internet access, work obligations, and exam schedules limited their ability to participate. Therefore, the high nonparticipation rate should be interpreted within Lebanon's unstable socioeconomic context, where such logistical constraints, not genuine disinterest in DM, are more plausible explanations for nonenrollment.

Conclusions

This study demonstrates that e-learning and F2F DM education are comparably effective in improving knowledge retention, confidence, and perceived competence among Lebanese medical students. The findings support integrating a DM course into the fifth-year curriculum before externships to enhance preparedness and application of skills. A blended approach combining theoretical e-learning with practical F2F sessions may further strengthen practical competencies. These results provide evidence to guide curriculum development and DM training implementation at the LUFMS.

Acknowledgments

The authors would like to thank all students, faculty, and department who contributed to the fulfillment of this project. All authors declared that they had insufficient funding to support open access publication of this manuscript, including from affiliated organizations or institutions, funding agencies, or other organizations. JMIR Publications provided article processing fee (APF) support for the publication of this article.

Funding

This study was funded by the Agence Française de Développement as part of its partnership project with the International Committee of the Red Cross and Rafik Hariri University Hospital. AFD had no role in the conceptualization, design, data collection, analysis, decision to publish, or preparation of the article.

Data Availability

Data are available with the corresponding author upon request.

Authors' Contributions

Conceptualization: AM, CB, HS, and RP
Methodology: AM, CB, DA, and AB
Ethical approval: ZM and HAM
Project administration: AM, HAM, MM, and AB

XSL•FO
RenderX

Supervision: MM, AB, and RP
Teaching: AM
Resources: AM and AB
Investigation: AM, HAM, and RP
Data curation: AM, LA, and AB
Formal analysis: AM, LA, and MM
Validation: LA and MM
Original draft: AM
Review and editing: AM, CB, ZM, HS, MM, and AB
Visualization: AM

Conflicts of Interest

None declared.

Multimedia Appendix 1
Questions for knowledge for phases I, II, and III.
[PDF File (Adobe PDF File), 336 KB - mededu_v12i1e80409_app1.pdf]

Multimedia Appendix 2
Questions for the evaluation of the satisfaction of participants in the DM course utilizing the Likert scale.
[DOCX File , 18 KB - mededu_v12i1e80409_app2.docx]

Multimedia Appendix 3
Questions for evaluations of the confidence of participants in the disaster management educational course utilizing the Likert scale.
[DOCX File , 19 KB - mededu_v12i1e80409_app3.docx]

Multimedia Appendix 4
All data collection schemes.
[DOCX File , 15 KB - mededu_v12i1e80409_app4.docx]

Multimedia Appendix 5
Mean knowledge scores for the four groups over the three assessment phases in the overall sample and stratified by year of education and according to the Solomon 4-group design.
[DOCX File , 19 KB - mededu_v12i1e80409_app5.docx]

References

1. Kim PY, Sawe HR, Wallis LA. The Sendai Framework and emergency care. Afr J Emerg Med 2016 Mar;6(1):3-4 [FREE Full text] [doi: 10.1016/j.afjem.2015.09.001] [Medline: 30456056]
2. Ingrassia PL, Ragazzoni L, Tengattini M, Carenzo L, Della Corte F. Nationwide program of education for undergraduates in the field of disaster medicine: development of a core curriculum centered on blended learning and simulation tools. Prehosp Disaster Med 2014 Oct;29(5):508-515 [FREE Full text] [doi: 10.1017/S1049023X14000831] [Medline: 25155942]
3. Chong K, Abdul NF, Musa S, Shahron S, Anuar N. Challenges and opportunities of disaster education program among UUM students. 2018. URL: https://ui.adsabs.harvard.edu/abs/2018AIPC.2016b0038L/abstract [accessed 2026-01-03]
4. Ragazzoni L, Ingrassia PL, Gugliotta G, Tengattini M, Franc JM, Corte FD. Italian medical students and disaster medicine: awareness and formative needs. Am J Disaster Med 2013;8(2):127-136. [doi: 10.5055/ajdm.2013.0119] [Medline: 24352928]
5. International FOMSA. IFMSA Policy Statement on Disasters and Emergencies 2012. Accessed February 6. 2014. URL: http://imccam2012.files.wordpress.com/2012/07/psam12-disasters-and-emergencies.pdf [accessed 2026-01-03]
6. Kaji AH, Coates W, Fung C. A disaster medicine curriculum for medical students. Teach Learn Med 2010;22(2):116-122. [doi: 10.1080/10401331003656561] [Medline: 20614377]
7. Mortelmans LJ, De Cauwer HG, Van Dyck E, Monballyu P, Van Giel R, Van Turnhout E. Are Belgian senior medical students ready to deliver basic medical care in case of a H5N1 pandemic? Prehosp Disaster Med 2009;24(5):438-442. [doi: 10.1017/s1049023x00007287] [Medline: 20066648]
8. Hermann S, Gerstner J, Weiss F, Aichele S, Stricker E, Gorgati E, et al. Presentation and evaluation of a modern course in disaster medicine and humanitarian assistance for medical students. BMC Med Educ 2021;21(1):610 [FREE Full text] [doi: 10.1186/s12909-021-03043-6] [Medline: 34893083]

XSL•FO
RenderX

9. Bhattacharya S, Singh A, Semwal J, Marzo RR, Sharma N, Goyal M, et val. Impact of a training program on disaster preparedness among paramedic students of a tertiary care hospital of North India: A single-group, before-after intervention study. J Educ Health Promot 2020;9:5 [FREE Full text] [doi: 10.4103/jehp.jehp_423_19] [Medline: 32154300]
10. Ashcroft J, Byrne M, Brennan PA, Davies RJ. Preparing medical students for a pandemic: a systematic review of student disaster training programmes. Postgrad Med J 2021;97(1148):368-379 [FREE Full text] [doi: 10.1136/postgradmedj-2020-137906] [Medline: 32518075]
11. Voicescu GT, Lamine H, Lo on i AE, Lupan-Mure an EM, Luka S, Ulerio JG, et al. Monitoring and evaluation in disaster management courses: a scoping review. BMC Med Educ 2025;25(1):188 [FREE Full text] [doi: 10.1186/s12909-025-06659-0] [Medline: 39910476]
12. Bahattab AA, Linty M, Trentin M, Truppa C, Hubloue I, Della Corte F, et al. Availability and characteristics of humanitarian health education and training programs: a web-based review. Prehosp Disaster Med 2022;37(1):132-138. [doi: 10.1017/S1049023X21001333] [Medline: 34928199]
13. Saeed Bahattab AA, Amat Camacho N, von Reding H, Rogers M, Lajolo CP, Salio F, et al. Climate, conflict, and epidemic: global and regional disaster profiles and implications for emergency medical teams. Prehosp Disaster Med 2025 May 21;40(S1):s3-s3. [doi: 10.1017/s1049023x25000305]
14. Lebanon. World Bank Group. URL: https://data.worldbank.org/country/lebanon [accessed 2026-01-03]
15. Lebanon profile-timeline. BBC. 2021 Jan 08. URL: https://www.bbc.com/news/world-middle-east-14649284 [accessed 2024-01-03]
16. Beirut explosion: what we know so far.: BBC; 2020 Aug 5. URL: https://www.bbc.com/news/world-middle-east-53668493 [accessed 2024-01-03]
17. Crisis briefing. International Committee of the Red Cross. URL: https://www.rescue.org/country/lebanon [accessed 2024-01-03]
18. Lebanon. World Health Organization. 2023 Apr 03. URL: https://www.who.int/countries/lbn/ [accessed 2026-01-03]
19. Bahattab A, Caviglia M, Martini D, Hubloue I, Della Corte F, Ragazzoni L. Scenario-based e-simulation design for global health education: theoretical foundation and practical recommendations. J Med Internet Res 2023;25:e46639 [FREE Full text] [doi: 10.2196/46639] [Medline: 37902810]
20. Zarifsanaiey N, Farrokhi M, Karimian Z, Hoseini S, Chahartangi F, Shahraki HR. Lesson learned from assessing teachers' and students' perspectives regarding the quality of e-learning in medical education during the pandemic: a mixed-methods study. BMC Med Educ 2024;24(1):171 [FREE Full text] [doi: 10.1186/s12909-024-05160-4] [Medline: 38388898]
21. Imran R, Fatima A, Elbayoumi Salem I, Allil K. Teaching and learning delivery modes in higher education: looking back to move forward post-COVID-19 era. The International Journal of Management Education 2023;21(2):100805. [doi: 10.1016/j.ijme.2023.100805]
22. Mahdavi Ardestani SF, Adibi S, Golshan A, Sadeghian P. Factors influencing the effectiveness of e-learning in healthcare: a fuzzy ANP study. Healthcare (Basel) 2023;11(14) [FREE Full text] [doi: 10.3390/healthcare11142035] [Medline: 37510477]
23. De Felice S, Hamilton A, Ponari M, Vigliocco G. Learning others is good, others is better: the role of social interaction in human acquisition of new knowledge. Philos Trans R Soc Lond B Biol Sci 2023;378(1870):20210357 [FREE Full text] [doi: 10.1098/rstb.2021.0357] [Medline: 36571126]
24. Luke K. Twelve tips for using synchronous virtual classroom technologies in medical education. MedEdPublish (2016) 2021;10:66 [FREE Full text] [doi: 10.15694/mep.2021.000066.1] [Medline: 38486516]
25. Alarifi BN, Song S. Online vs in-person learning in higher education: effects on student achievement and recommendations for leadership. Humanit Soc Sci Commun 2024;11(1). [doi: 10.1057/s41599-023-02590-1]
26. Bahattab A, Zain O, Linty M, Amat Camacho N, Von Schreeb J, Hubloue I, et al. Development and evaluation of scenario-based e-simulation for humanitarian health training: a mixed-methods action research study. BMJ Open 2024;14(8):e079681 [FREE Full text] [doi: 10.1136/bmjopen-2023-079681] [Medline: 39107011]
27. Major Incident Medical Management and Support (MIMMS) course. URL: https://www.coemed.org/courses/course_desc/mimms [accessed 2024-01-03]
28. Health Emergency and Disaster Risk Management Framework.: WHO; 2019. URL: https://www.who.int/publications/i/item/health-emergency-and-disaster-risk-management-framework [accessed 2024-01-03]
29. Naidoo M, Lee J, Trelles M, Wallis L, Chu KM. Preventing avoidable hospital admissions after emergency care in humanitarian settings: a cross-sectional review of Médecins Sans Frontières emergency departments. BMJ Open 2021;11(7):e049785 [FREE Full text] [doi: 10.1136/bmjopen-2021-049785] [Medline: 34257097]
30. Desmond M, Schwengel D, Chilson K, Rusy D, Ingram K, Ambardekar A, et al. Paediatric patients in mass casualty incidents: a comprehensive review and call to action. Br J Anaesth 2022;128(2):e109-e119 [FREE Full text] [doi: 10.1016/j.bja.2021.10.026] [Medline: 34862001]
31. Schaefer T, Nunez L. Burn resuscitation and management. Orlando, FL: StatPearls Publishing; Oct 03, 2024.
32. Hughes A, Almeland SK, Leclerc T, Ogura T, Hayashi M, Mills J, et al. Recommendations for burns care in mass casualty incidents: WHO Emergency Medical Teams Technical Working Group on Burns (WHO TWGB) 2017-2020. Burns 2021;47(2):349-370 [FREE Full text] [doi: 10.1016/j.burns.2020.07.001] [Medline: 33041154]

XSL•FO
RenderX

33. Suwalowska H, Amara F, Roberts N, Kingori P. Ethical and sociocultural challenges in managing dead bodies during epidemics and natural disasters. BMJ Glob Health 2021 Nov;6(11) [FREE Full text] [doi: 10.1136/bmjgh-2021-006345] [Medline: 34740913]
34. Craig P, Dieppe P, Macintyre S, Michie S, Nazareth I, Petticrew M, et al. Developing and evaluating complex interventions: the new Medical Research Council guidance. BMJ 2008;337:a1655 [FREE Full text] [doi: 10.1136/bmj.a1655] [Medline: 18824488]
35. Glasgow RE, Emmons KM. How can we increase translation of research into practice? Types of evidence needed. Annu Rev Public Health 2007;28:413-433. [doi: 10.1146/annurev.publhealth.28.021406.144145] [Medline: 17150029]
36. Corpuz JCG. A dignified death: management of dead bodies during COVID-19. J Public Health (Oxf) 2022 Jun 27;44(2):e281-e282 [FREE Full text] [doi: 10.1093/pubmed/fdab182] [Medline: 34096608]
37. French D, Miles L, Elbourne D, Farmer A, Gulliford M, Locock L, et al. Reducing bias in trials from reactions to measurement: the MERIT study including developmental work and expert workshop. Health Technol Assess 2021;25(55):1-72 [FREE Full text] [doi: 10.3310/hta25550] [Medline: 34553685]
38. Han SJ, Chun J. Validation of the disaster preparedness evaluation tool for nurses-the korean version. Int J Environ Res Public Health 2021;18(3) [FREE Full text] [doi: 10.3390/ijerph18031348] [Medline: 33540899]
39. Wang J, Lu S, Sun X, Wang F, Wan M, Chen H, et al. Psychometric evaluation of the disaster preparedness evaluation tool (dpet) on emergency nurses in mainland China: two cross-sectional studies. Disaster Med Public Health Prep 2022;16(3):1083-1090. [doi: 10.1017/dmp.2021.39] [Medline: 33949305]
40. Sobaih AEE, Palla IA, Baquee A. Social media use in e-learning amid COVID 19 pandemic: Indian students' perspective. Int J Environ Res Public Health 2022;19(9) [FREE Full text] [doi: 10.3390/ijerph19095380] [Medline: 35564771]
41. Bahattab A, Trentin M, Hubloue I, Della Corte F, Ragazzoni L. Humanitarian health education and training state-of-the-art: a scoping review. Front Public Health 2024;12:1343867 [FREE Full text] [doi: 10.3389/fpubh.2024.1343867] [Medline: 39135925]
42. Roman M, Plopeanu A. The effectiveness of the emergency eLearning during COVID-19 pandemic. The case of higher education in economics in Romania. International Review of Economics Education 2021;37:100218. [doi: 10.1016/j.iree.2021.100218]
43. Maatuk AM, Elberkawi EK, Aljawarneh S, Rashaideh H, Alharbi H. The COVID-19 pandemic and E-learning: challenges and opportunities from the perspective of students and instructors. J Comput High Educ 2022;34(1):21-38 [FREE Full text] [doi: 10.1007/s12528-021-09274-2] [Medline: 33967563]
44. Fauzi MA. E-learning in higher education institutions during COVID-19 pandemic: current and future trends through bibliometric analysis. Heliyon 2022;8(5):e09433 [FREE Full text] [doi: 10.1016/j.heliyon.2022.e09433] [Medline: 35600443]
45. Schilly K, Huhn M, Visker JD, Cox C. Evaluation of a disaster preparedness curriculum and medical students' views on preparedness education requirements for health professionals. Disaster Med Public Health Prep 2024;18:e8. [doi: 10.1017/dmp.2023.230] [Medline: 38282524]
46. Xia Y, Hu Y, Wu C, Yang L, Lei M. Challenges of online learning amid the COVID-19: college students' perspective. Front Psychol 2022;13:1037311 [FREE Full text] [doi: 10.3389/fpsyg.2022.1037311] [Medline: 36619114]
47. Doumi R, Alassaf W, Aljahany M, Aldobeaban S, Mirghani HA, BinSaleh Y, et al. The effectiveness and benefits of disaster simulation training for undergraduate medical students in Saudi Arabia. Adv Med Educ Pract 2024;15:707-714 [FREE Full text] [doi: 10.2147/AMEP.S453879] [Medline: 39050119]
48. Mubarak AF, Amiruddin R, Gaus S. The effectiveness of disaster prevention and mitigation training for the students in disaster prone areas. IOP Conf. Ser.: Earth Environ. Sci 2019;235:012055. [doi: 10.1088/1755-1315/235/1/012055]
49. Barrimah I, Adam I, Al-Mohaimeed A. Disaster medicine education for medical students: is it a real need? Med Teach 2016;38 Suppl 1:S60-S65. [doi: 10.3109/0142159X.2016.1142515] [Medline: 26984036]
50. Tipton E, Hallberg K, Hedges LV, Chan W. Implications of small samples for generalization: adjustments and rules of thumb. Eval Rev 2017 Oct;41(5):472-505. [doi: 10.1177/0193841X16655665] [Medline: 27402612]
51. Stuart EA, Bradshaw CP, Leaf PJ. Assessing the generalizability of randomized trial results to target populations. Prev Sci 2015;16(3):475-485 [FREE Full text] [doi: 10.1007/s11121-014-0513-z] [Medline: 25307417]

Abbreviations

DM: disaster medicine
LUFMS: Lebanese University Faculty of Medical Sciences
F2F: face-to-face
ICRC: International Committee of the Red Cross

XSL•FO
RenderX

Edited by T Gladman; submitted 11.Jul.2025; peer-reviewed by H Hanada, J Wilkinson; comments to author 03.Nov.2025; revised version received 05.Dec.2025; accepted 14.Dec.2025; published 28.Jan.2026.

Please cite as:
Msheik A, Peralta R, Al Mokdad Z, Bartulec C, Abou-Abbas L, Anani D, Al Mawla H, Salame H, Moussa M, Bahattab A
Disaster Medicine Training for Medical Students in Lebanon: Quasi-Experimental Comparison of e-Learning and Face-to-Face Modalities
JMIR Med Educ 2026;12:e80409
URL: https://mededu.jmir.org/2026/1/e80409
doi:10.2196/80409
PMID:

©Ali Msheik, Ruben Peralta, Zeinab Al Mokdad, Christine Bartulec, Linda Abou-Abbas, Dima Anani, Hussein Al Mawla, Hassan Salame, Mohamad Moussa, Awsan Bahattab. Originally published in JMIR Medical Education (https://mededu.jmir.org), 28.Jan.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Education, is properly cited. The complete bibliographic information, a link to the original publication on https://mededu.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Dependability of Entrustable Professional Activity Portfolios in Otorhinolaryngology Residency: Nationwide Generalizability and Decision Study

Jeng-Wen Chen[1,2,3*], MD, MSc; Ching-Lin Hsieh[4,5,6*], PhD; Chun-Hsiang Chang[1*], MD, MSc; Ya-Hui Wang[7], PhD; Wei-Chung Hsu[3,8,9], MD, PhD; Pa-Chun Wang[10,11,12], MD, PhD; Li-Jen Liao[13], MD; Chun-Hou Liao[14], MD, PhD; Mingchih Chen[2,15], PhD; Okki Dhona Laksmita[4], PhD, RN

[1]Department of Otorhinolaryngology–Head and Neck Surgery, Cardinal Tien Hospital and Fu Jen Catholic University, 362, ZhongZheng Rd. Xindian Dist., New Taipei City, Taiwan

[10]Department of Otolaryngology, Cathay General Hospital, Taipei, Taiwan

[11]Department of Otolaryngology, School of Medicine, Fu-Jen Catholic University, New Taipei City, Taiwan

[12]Department of Medical Research, China Medical University Hospital, Taichung, Taiwan

[13]Department of Otolaryngology, Far Eastern Memorial Hospital, New Taipei City, Taiwan

[14]Division of Urology, Department of Surgery, Cardinal Tien Hospital and Fu Jen Catholic University, New Taipei City, Taiwan

[15]Artificial Intelligence Development Center, Fu Jen Catholic University, New Taipei City, Taiwan

[2]Department of Hospital Management, Graduate Institute of Business Administration, Fu Jen Catholic University, New Taipei City, Taiwan

[3]Department of Otorhinolaryngology–Head and Neck Surgery, National Taiwan University Hospital, Taipei, Taiwan

[4]School of Occupational Therapy, College of Medicine, National Taiwan University, Taipei, Taiwan

[5]Department of Physical Medicine and Rehabilitation, National Taiwan University Hospital, Taipei, Taiwan

[6]Department of Occupational Therapy, College of Medical and Health Sciences, Asia University, Taichung, Taiwan

[7]Medical Research Center, Cardinal Tien Hospital and Fu Jen Catholic University, New Taipei City, Taiwan

[8]Sleep Center, National Taiwan University Hospital, Taipei, Taiwan

[9]Department of Otolaryngology, College of Medicine, National Taiwan University, Taipei, Taiwan

[*]these authors contributed equally

Corresponding Author:
Jeng-Wen Chen, MD, MSc
Department of Otorhinolaryngology–Head and Neck Surgery, Cardinal Tien Hospital and Fu Jen Catholic University, 362, ZhongZheng Rd. Xindian Dist., New Taipei City, Taiwan

Abstract

Background: Competency-based medical education increasingly uses digital workplace assessment platforms to collect longitudinal evidence of trainees' readiness for progressive responsibility. Entrustable professional activities (EPAs) translate competencies into observable clinical work, but individual workplace observations are context-dependent and may not support dependable progression decisions. Empirical guidance remains limited on how many observations, faculty raters, EPA titles, and workplace settings are needed for a dependable EPA portfolio.

Objective: This study aimed to estimate EPA portfolio dependability for competency-based progression review using nationwide data from a digital workplace assessment platform in otorhinolaryngology-head and neck surgery residency training.

Methods: We conducted a retrospective, nationwide cross-sectional generalizability and decision study using routinely collected EPA-based workplace assessment data from Taiwan's E-MyWay platform. The study included completed EPA observations from accredited otorhinolaryngology-head and neck surgery residency programs between August 2022 and May 2026. Faculty-assigned entrustment-supervision ratings were scored on a 5-level national EPA scale. The resident was the object of measurement, whereas faculty rater, EPA title, and workplace setting were measurement facets. The generalizability study partitioned rating variance into resident, facet, interaction, and residual components, which were subsequently used in decision-study projections of the generalizability coefficient, Phi coefficient, and absolute standard error of measurement under alternative portfolio configurations.

Results: The analytic sample included 45,526 EPA observations involving 466 residents and 448 faculty raters across 35 training programs, 11 EPA titles, and 5 workplace settings. Residents had a median of 87 (IQR 47-124) observations and were assessed by a median of 9 (IQR 6-12) faculty raters across 11 (IQR 10-11) EPA titles and 5 (IQR 4-5) workplace settings. Single EPA observations were insufficient for resident-level summative interpretation. In the primary generalizability study,

XSL•FO
RenderX

residual/encounter-level variance was the largest component, accounting for 40.1% of total variance; resident variance accounted for 26.9%, faculty rater variance accounted for 10.8%, and resident × faculty rater interaction for 10.4%. A study-defined well-distributed 10-observation portfolio achieved a generalizability coefficient of 0.77 but an absolute dependability coefficient (Phi) of 0.65. A 20-observation portfolio across 5 faculty raters, 5 EPA titles, and 3 workplace settings yielded a Phi of 0.76, and a 30-observation portfolio across 6 faculty raters, 6 EPA titles, and 4 workplace settings yielded a Phi of 0.80.

Conclusions: Nationwide digital workplace assessment data showed that EPA portfolio dependability reflected both observation counts and sampling breadth across faculty raters, EPA titles, and workplace settings. These study-defined portfolio configurations may help clinical competency committees, residency programs, and specialty societies monitor portfolio completeness before progression deliberation. They should complement, rather than replace, narrative feedback, case complexity, performance trajectory, and committee judgment.

(JMIR Med Educ 2026;12:e104628) doi:10.2196/104628

KEYWORDS

entrustable professional activities; competency-based medical education; generalizability theory; clinical competency committee; postgraduate medical education

Introduction

Digital workplace assessment platforms are increasingly used to collect longitudinal evidence for progression decisions [1-3]. In postgraduate competency-based medical education (CBME) [4,5], these systems can capture entrustable professional activity (EPA) [6,7] observations across faculty raters, clinical tasks, and workplace settings [8]. However, the availability of large volumes of digital assessment data does not by itself ensure that resident portfolios contain sufficient, well-distributed evidence for defensible progression review [9,10].

A persistent challenge for digital EPA systems is determining when accumulated workplace observations become an adequate evidence portfolio [11,12]. Individual EPA observations are useful for formative feedback, but summative progression decisions require aggregation across contextually variable encounters [13]. Without empirical benchmarks, programs may under-sample performance, risking unstable decisions, or over-sample, increasing assessment burden without proportional gains in decision quality [14].

Digital platforms create an opportunity to examine these sampling problems at scale by preserving structured information about raters, EPA titles, workplace settings, and longitudinal assessment exposure. Otorhinolaryngology-head and neck surgery is an informative specialty context, with residents assessed across emergency, inpatient, outpatient, procedural, and operating-room environments [15,16]. This breadth makes the specialty well suited for studying how EPA portfolios should be sampled across clinical tasks and settings. Taiwan offers a unique national context because, since 2022, the Taiwan Society of Otorhinolaryngology-Head and Neck Surgery (TSO-HNS) has implemented a specialty EPA framework through the E-MyWay platform across accredited residency programs, enabling standardized national capture of EPA ratings, faculty identifiers, resident identifiers, EPA titles, and workplace settings [17-20]. This infrastructure allows portfolio dependability to be examined as a learning analytics question rather than as a single-program quality assurance exercise [21,22].

This study used nationwide E-MyWay data [17-20] to examine EPA portfolio dependability as a multidimensional sampling problem in CBME [4-7,14]. Rather than focusing only on the number of EPA observations, we evaluated how resident-level portfolio dependability varied jointly with observation count and breadth across faculty raters, EPA titles, and workplace settings [23,24]. By combining routine digital workplace-based assessment data with generalizability study (G-study) and decision study (D-study) modeling [23-25], this study aimed to provide portfolio-design evidence for programmatic assessment in competency-based otorhinolaryngology residency training [2,9,26].

Methods

Study Design and Participants

We conducted a retrospective, nationwide cross-sectional G-study and D-study using routinely collected EPA-based workplace assessment data from Taiwan's E-MyWay platform [20]. The study evaluated whether faculty-assigned entrustment-supervision ratings, aggregated across resident assessment portfolios, provided dependable evidence for resident-level progression review in otorhinolaryngology-head and neck surgery residency training. The intended interpretation was resident entrustment readiness across the specialty EPA framework; the intended use was programmatic review and progression decision-making.

Otorhinolaryngology-head and neck surgery residency training in Taiwan lasts 4.5 years, spanning postgraduate year 3 through postgraduate year 7. During the study period, all 35 accredited programs using E-MyWay were included, representing the national implementation network for this specialty. Participating programs were distributed across all 6 National Health Insurance Administration regional divisions: Taipei (n=16), Northern (n=2), Central (n=7), Southern (n=4), Kao-Ping (n=5), and Eastern (n=1), reflecting broad national coverage of accredited training sites.

Digital Workplace Assessment Platform

E-MyWay is a web app–based national CBME platform, developed by the Joint Commission of Taiwan, to capture

XSL•FO
RenderX

structured EPA assessment records [20]. In otorhinolaryngology-head and neck surgery, it enables standardized documentation of resident identifier, faculty rater identifier, EPA title, workplace setting, entrustment-supervision rating, narrative resident reflection, and faculty feedback [18]. The TSO-HNS developed the framework through an iterative national expert-consensus process beginning in August 2020. A CBME workgroup comprising residency review committee members, program directors, specialty experts, and physician educators used Glaser's modified state-of-the-art approach, followed by internal and external quantitative and qualitative review. The framework was finalized in June 2021, and stage-specific expected entrustment-supervision levels were agreed upon during pilot implementation in 2021 to 2022. This process was an expert-consensus and pilot-refinement process rather than a formal Delphi study. The platform supports program-level monitoring and permits aggregation of EPA observations into resident portfolios; the expectations are summarized in Table 1 [20]. The expected levels were provided for curricular context only and were not used to transform observed ratings or calculate the G-study or D-study estimates.

XSL•FO
RenderX

Table . Distribution of entrustable professional activity (EPA)–based workplace assessment records across 11 EPA titles and expected entrustment-supervision levels by resident seniority.

EPA[a] title	Assessment records, n (%)[b]	A priori expected entrustment-supervision levels by resident seniority[c]				
		PGY3[d]	PGY4	PGY5	PGY6	PGY7
EPA01 (Airway) Assessing and managing patients with airway presentations	3693 (8.1)	2	3	3	4	5
EPA02 (Foreign Body) Assessing and managing patients with suspicious foreign body presentations	3263 (7.2)	2	3	3	4	5
EPA03 (Bleeding) Assessing and managing patients with upper aerodigestive tract bleeding presentations	3139 (6.9)	2	3	3	4	5
EPA04 (Vertigo) Assessing and managing patients with vertigo	2346 (5.2)	2	3	3	4	5
EPA05 (Infection) Assessing and managing patients with head and neck infections	4033 (8.9)	2	3	3	4	5
EPA06[e] (Head and Neck) Assessing and managing patients with head and neck masses	7935 (17.4)	2	2	3	4	5
EPA07 (Ear) Assessing and managing patients with ear and hearing diseases	5521 (12.1)	2	3	3	4	5
EPA08 (Sinonasal) Assessing and managing patients with sinonasal diseases	6071 (13.3)	2	3	4	4	5
EPA09 (Larynx) Assessing and managing patients with laryngopharyngeal diseases (voice/speech/language/dysphagia)	3729 (8.2)	2	3	3	4	5
EPA10[e] (Sleep Disordered Breathing) Assessing and managing patients with sleep-disordered breathing	3443 (7.6)	2	2	3	4	5

XSL•FO
RenderX

EPA[a] title	Assessment records, n (%)[b]	A priori expected entrustment-supervision levels by resident seniority[c]				
		PGY3[d]	PGY4	PGY5	PGY6	PGY7
EPA11[e] (Plasty) Assessing and managing patients with facial plastic and reconstructive surgery	2353 (5.2)	2	2	3	3	4

[a]EPA: entrustable professional activity.

[b]The analytic dataset included EPA01-EPA11 only. EPA12 was present in the original data source but was excluded owing to its later introduction and inconsistent availability across the analytic window.

[c]Expected entrustment-supervision levels were established a priori through the Taiwan Society of Otorhinolaryngology-Head and Neck Surgery iterative expert-consensus and pilot-refinement process. They represent curricular reference expectations for each integrated EPA by training stage, rather than phase-specific or empirically derived cut points.

[d]PGY: postgraduate year.

[e]Delayed EPAs: EPA06, EPA10, and EPA11.

Portfolio Definitions

An observation was defined as a completed EPA-based workplace assessment record. A resident assessment portfolio was defined as the accumulated set of EPA observations for an individual resident. Portfolio size referred to the number of observations, whereas portfolio breadth referred to sampling across faculty raters, EPA titles, and workplace settings. The dataset had an unbalanced, partially crossed, multi-institutional structure, with residents and faculty raters largely clustered within training hospitals but partially crossed through workplace assessments.

Data Source

We analyzed EPA-based workplace assessment records from the August 2022 to May 2026 academic years. The data extraction date was May 1, 2026. Each record represented one completed workplace observation in which a faculty rater assessed the component of an EPA demonstrated during a specific clinical encounter and workplace setting and assigned an entrustment-supervision level for that observed performance.

The primary analysis included EPA01 through EPA11, which constituted the consistently implemented specialty EPA framework during the study period [17]. EPA12 was excluded owing to its introduction in 2024 and inconsistent availability across the analytic window. Conference-room assessments were also excluded because they represented case presentations rather than true workplace-based assessments. The eligible workplace settings were operating room, ward, including intensive care unit, emergency department, outpatient clinic, and consultation. Records were included if they contained complete data for entrustment–supervision level, resident identifier, faculty rater identifier, EPA title, and workplace setting.

Outcome and Measurement Facets

The primary outcome was the faculty-assigned entrustment-supervision rating, scored on a 5-level scale aligned with the national specialty EPA framework [6,20]. The 5-level scale used common supervision anchors: observation without active participation, direct and proactive supervision, indirect and reactive supervision, independent practice, and supervision of junior learners. Higher scores indicated greater readiness for less supervised practice. The resident was specified as the object of measurement to align with the intended inference of resident-level EPA portfolio dependability for programmatic review and progression deliberation. Faculty rater, EPA title, and workplace setting were treated as measurement facets given their roles as major sources of sampling variation in routine workplace assessment. EPA title was included as a facet to examine aggregate portfolio dependability across the specialty EPA framework rather than certification-level dependability for each individual EPA.

Selected 2-way interactions—resident × faculty rater, resident × EPA title, and faculty rater × EPA title—were modeled to capture major sources of score variation. Residual/encounter-level variance was interpreted as remaining assessment-level or encounter-level variation not explained by the modeled resident, faculty rater, EPA title, workplace setting, or interaction components. This component may include occasion-specific resident performance, patient or case factors, unmodeled contextual variation, higher-order interactions, rater inconsistency, and measurement error. The same clinical encounter was not independently rated by multiple faculty raters; therefore, residual/encounter-level variance was not interpreted as conventional interrater reliability.

Although entrustment-supervision ratings are ordinal, treating them as approximately numeric is therefore a modeling assumption. We used this approach to support variance decomposition and D-study projection, and we interpreted the results as estimates of portfolio dependability rather than direct measures of clinical competence.

G-Study

We conducted a G-study to decompose variance in entrustment-supervision ratings [25]. Consistent with the measurement-facet framework described above, the primary model was an intercept-only random-effects linear mixed-effects model that included random components for resident, faculty rater, EPA title, workplace setting, resident × faculty rater,

XSL•FO
RenderX

resident × EPA title, faculty rater × EPA title, and residual error. This model was selected to estimate aggregate resident-level portfolio dependability across the specialty EPA framework rather than EPA-specific dependability for each individual EPA. The full model specification is provided in Multimedia Appendix 1. Given the unbalanced, partially crossed, multi-institutional data structure, variance components were interpreted in relation to the intended resident-level inference rather than as properties of individual observation records. Variance components were expressed as proportions of total variance to quantify resident signal, rater effects, task effects, workplace effects, interaction effects, and encounter-level variation.

D-Study

Using variance components from the primary G-study, we conducted D-studies to estimate the projected dependability of alternative EPA portfolio configurations [23,24,27,28]. The Phi coefficient was emphasized as the primary dependability index for absolute decisions, consistent with the criterion-referenced nature of resident-level programmatic review and progression deliberation. The generalizability coefficient was reported secondarily as an index of relative dependability, and the absolute standard error of measurement (SEM) was calculated to quantify expected measurement error in the original rating metric.

D-study scenarios varied the number of observation records, faculty raters, EPA titles, and workplace settings to model both portfolio size and breadth. The D-study portfolio configurations were specified as practical, study-defined scenarios in which portfolio size and breadth increased together. The 10-, 20-, 30-, 40-, 60-, median observed, and 100-observation scenarios were selected to represent a range of plausible portfolio structures, from smaller portfolios for routine programmatic review to broader portfolios that may be considered before more consequential progression deliberation. These configurations were not empirically optimized cutoffs or validated progression standards; rather, they were model-based scenarios used to examine how projected dependability changed when observations were distributed across increasing numbers of faculty raters, EPA titles, and workplace settings. Phi ≥0.70 and Phi ≥0.80 were used as prespecified psychometric reference points for interpreting projected dependability, not as validated standards for competence, certification, or progression.

Validity Framework

We interpreted dependability as one source of validity evidence for using EPA scores in progression review [26]. Within the Messick framework [29], this study focused on internal structure evidence by examining how much variation in entrustment ratings was attributable to residents relative to measurement facets and residual error. We did not evaluate other sources of validity evidence, including content, response processes, relations to other variables, or consequences [30]. Accordingly, the findings should be interpreted as one component of a broader validity argument for EPA-based progression decisions [26,31-34].

Portfolio Concentration

To characterize portfolio imbalance arising from opportunistic workplace-based assessment, we conducted a descriptive portfolio concentration analysis. For each resident, we calculated the proportion of observations contributed by the most frequent faculty rater, the most frequent EPA title, and the most frequent workplace setting. These resident-level proportions were summarized using medians and interquartile ranges. This analysis was intended to describe, but not statistically correct, observed sampling imbalance.

Sensitivity Analyses

We performed 3 sensitivity analyses to examine model robustness. First, an EPA title × workplace setting sensitivity model assessed whether task-context combinations materially changed the variance-component structure. Second, to address potential hospital or program-level effects, we fitted a hospital-aware sensitivity model that added random intercepts for resident hospital and faculty hospital to the primary G-study model. Hospital codes were modeled separately for residents and faculty raters to reflect the separate recording of resident and faculty hospital affiliations in the dataset. This model was used to evaluate whether between-hospital differences, including potential differences in curriculum, clinical exposure, assessment culture, or rater stringency, materially changed the variance-component structure. Hospital was treated as a contextual sensitivity specification rather than a primary measurement facet, consistent with the intended inference of aggregate resident-level portfolio dependability across the specialty EPA framework rather than hospital comparison.

Third, to examine whether case complexity accounted for residual/encounter-level variation, we fitted a complexity-adjusted sensitivity model that added case complexity as a fixed effect to the primary G-study model. Case complexity was available for 45,513 of 45,526 observations; 31,148 (68.4%) observations were classified as basic/routine and 14,365 (31.6%) as advanced/nonroutine, with 13 observations missing case-complexity data. The 13 observations with missing case-complexity data were excluded from this sensitivity analysis, yielding an analytic sample of 45,513 observations. This model was used to assess whether adjustment for case complexity materially changed the variance-component structure.

Statistical Analysis

All analyses were conducted using R (version 4.5.3; R Foundation for Statistical Computing). Data management and descriptive analyses used the tidyverse package, and linear mixed-effects models were estimated with lme4. D-study projections were derived from the estimated variance components. Observed resident portfolios were summarized by comparing each resident's assessment portfolio with the study-defined portfolio configurations for Phi ≥0.70 and Phi ≥0.80.

Ethical Considerations

This study was approved by the Institutional Review Board of Cardinal Tien Hospital (CTH-114-3-5-019). All data were analyzed in deidentified form. Electronic consent for research

XSL•FO
RenderX

use was obtained at initial E-MyWay platform enrollment and activation. Data were obtained from the TSO-HNS E-MyWay database with permission. The study is reported in accordance with applicable STROBE (Strengthening the Reporting of Observational Studies in Epidemiology) guidance for observational studies, with adaptations for G-study and D-study reporting (Checklist 1).

Results

Resident Assessment Exposure

The E-MyWay platform yielded 45,526 complete EPA-based workplace assessment records from 466 residents and 448 faculty raters across 35 accredited training programs, covering 11 EPA titles and 5 workplace settings. Residents had a mean of 97.7 (SD 64.9) observations and a median of 87 (IQR 47 - 124; range, 1 - 501) observations. Each resident was assessed by a median of 9 (IQR 6 - 12) faculty raters, across a median of 11 (IQR 10 - 11) EPA titles, and in a median of 5 (IQR 4 - 5) workplace settings.

EPA-specific assessment volume is summarized in Table 1. The most frequently assessed EPAs were EPA06, head and neck/oral mass assessment and management (n=7935 observations, 17.4%); EPA08, sinonasal disease assessment and management (n=6071, 13.3%); and EPA07, ear and hearing disease assessment and management (n=5521, 12.1%). The least frequently assessed were EPA04, vertigo assessment and management (n=2346, 5.2%), and EPA11, facial plastic and reconstructive surgery assessment and management (n=2353, 5.2%).

Variance Components From the G-Study

Variance components from the primary G-study are presented in Table 2. Residual/encounter-level variance was the largest component, accounting for 40.1% of total variance. Resident variance accounted for 26.9%, faculty rater variance accounted for 10.8%, and the resident × faculty rater interaction accounted for 10.4%. EPA title, workplace setting, resident × EPA title interaction, and faculty rater × EPA title interaction accounted for 4.2%, 2.5%, 2.0%, and 3.1%, respectively.

Table . Variance components from the generalizability study model.

Variance component[a]	Variance	Total variance (%)
Resident	0.32	26.9
Faculty rater	0.13	10.8
EPA[b] title	0.05	4.2
Workplace setting	0.03	2.5
Resident × faculty rater	0.12	10.4
Resident × EPA title	0.02	2.0
Faculty rater × EPA title	0.04	3.1
Residual/encounter-level variance	0.47	40.1

[a]Resident variance represents universe-score variance. Resident × faculty rater, resident × EPA title, and residual variance contribute to relative error variance. Absolute error variance additionally includes faculty rater, EPA title, workplace setting, and faculty rater × EPA title variance after adjustment for the corresponding numbers of raters, EPA titles, workplace settings, and observations in each projected portfolio.

[b]EPA: entrustable professional activity.

D-Study Portfolio Projections

D-study projections are summarized in Table 3. A well-distributed 10-observation portfolio across 3 faculty raters, 3 EPA titles, and 2 workplace settings yielded a G coefficient of 0.77 but a Phi coefficient of 0.65 and an absolute SEM of 0.42.

A study-defined well-distributed 20-observation portfolio across 5 faculty raters, 5 EPA titles, and 3 workplace settings yielded a G coefficient of 0.86, a Phi coefficient of 0.76, and an absolute SEM of 0.32, reaching the prespecified Phi ≥0.70 psychometric reference point. A study-defined well-distributed 30-observation portfolio across 6 faculty raters, 6 EPA titles, and 4 workplace settings yielded a G coefficient of 0.89, a Phi coefficient of 0.80, and an absolute SEM of 0.28, reaching the stricter Phi ≥0.80 psychometric reference point. Thus, increasing the number of observations without adequate distribution across raters, EPAs, and settings was insufficient to ensure higher absolute dependability. Together, these model-based projections showed higher absolute dependability in portfolio configurations with larger observation counts and broader sampling across faculty raters, EPA titles, and workplace settings.

Additional sampling produced smaller gains. Phi increased to 0.84 with 40 observations and 0.87 with 60 observations. The median observed portfolio scenario—approximately 87 observations across 9 faculty raters, 11 EPA titles, and 5 workplace settings—yielded a Phi of 0.87.

XSL•FO
RenderX

Table . Decision study projections for alternative assessment designs[a].

Portfolio scenario	Observations	Faculty raters	EPA[b] titles	Workplace settings	G coefficient	Phi coefficient	Absolute SEM[c]
10-observation well-distributed portfolio	10	3	3	2	0.77	0.65	0.42
20-observation well-distributed portfolio	20	5	5	3	0.86	0.76	0.32
30-observation well-distributed portfolio	30	6	6	4	0.89	0.80	0.28
40-observation well-distributed portfolio	40	8	8	4	0.91	0.84	0.25
60-observation well-distributed portfolio	60	9	10	5	0.93	0.87	0.22
Median-based observed well-distributed portfolio	87	9	11	5	0.94	0.87	0.21
100-observation well-distributed portfolio	100	10	11	5	0.94	0.88	0.21

[a]Each scenario represents a study-defined, well-distributed portfolio in which observations, faculty raters, EPA titles, and workplace settings vary together.

[b]EPA: entrustable professional activity.

[c]SEM: standard error of measurement.

Faculty-Rater Breadth and Observation Density

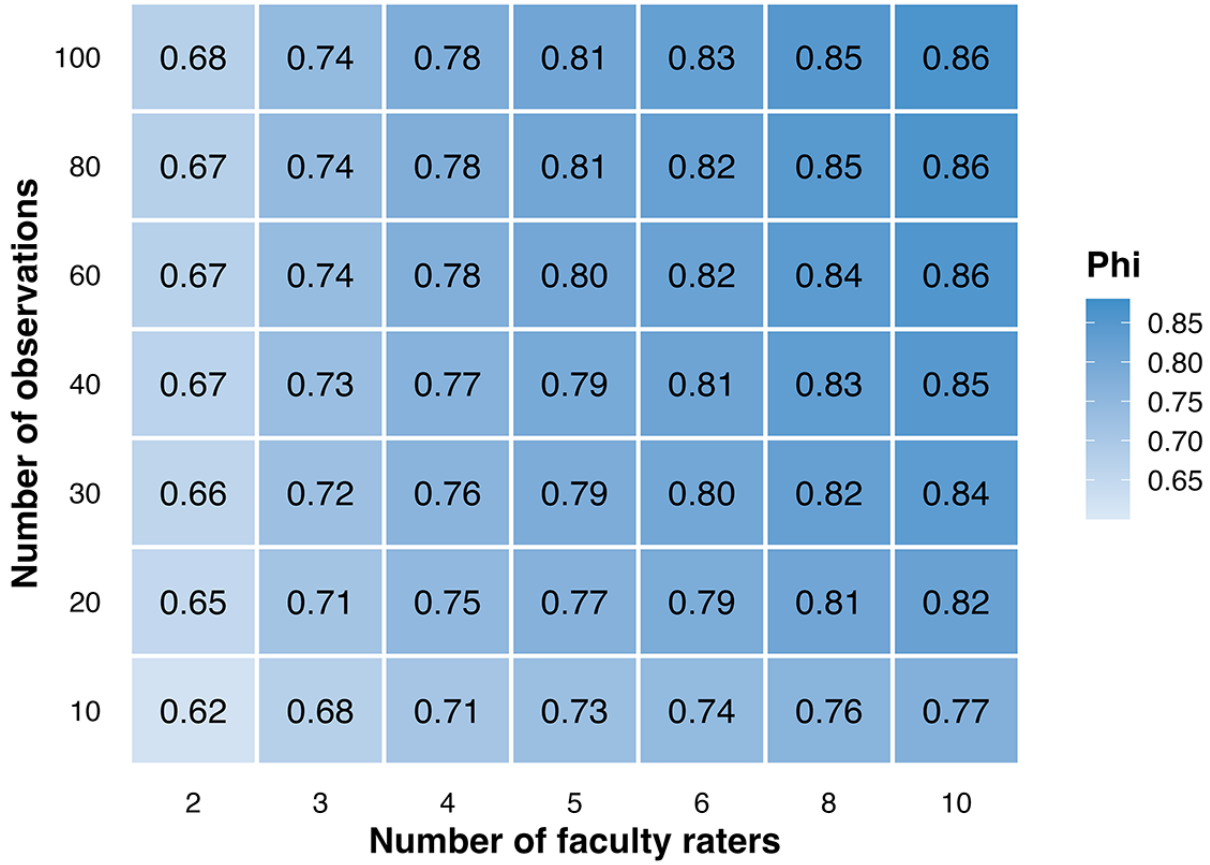
The D-study heatmap examined the joint contribution of observation count and faculty-rater breadth while holding EPA title breadth at 6 and workplace-setting breadth at 4 (Figure 1).

Across the displayed grid, projected Phi coefficients ranged from 0.62 to 0.86. Phi coefficients increased with both larger observation counts and broader faculty-rater representation.

Under this fixed EPA and workplace-setting structure, the lowest displayed observation count reaching Phi ≥0.80 was 20 observations with 8 or 10 faculty raters. With faculty-rater breadth fixed at 6 raters, the first displayed scenario reaching Phi ≥0.80 was 30 observations. The complementary D-study curve (Figure S1 in Multimedia Appendix 1) showed that the Phi coefficient increased from approximately 0.65 at 10 observations to 0.80 at 30 observations, while the G coefficient increased from approximately 0.77 to 0.89. From 40 to 100 observations, the Phi coefficient increased from approximately 0.84 to 0.88, and the G coefficient increased from approximately 0.91 to 0.94.

XSL•FO
RenderX

Figure 1. Decision-study projections of Phi coefficient by number of observations and faculty raters. Entrustable professional activity-title breadth was fixed at 6 and workplace-setting breadth was fixed at 4.



Observed Portfolio Configurations and Concentration

Observed resident portfolios were compared with projected D-study portfolio configurations (Table 4). Most residents met the observation-count reference values, with 447 (95.9%) residents having at least 20 observations and 412 (88.4%) residents having at least 30 observations. When faculty-rater, EPA-title, and workplace-setting breadth were also considered, 398 (85.4%) residents met the 20-observation well-distributed portfolio criteria, defined as at least 20 observations, 5 faculty raters, 5 EPA titles, and 3 workplace settings. A total of 330 (70.8%) residents met the stricter 30-observation well-distributed portfolio criteria, defined as at least 30 observations, 6 faculty raters, 6 EPA titles, and 4 workplace settings. Fewer residents met the study-defined well-distributed portfolio configurations than met the corresponding raw observation-count reference values.

Resident portfolios showed some concentration. The median proportion of observations contributed by the most frequent faculty rater within each resident's portfolio was 27.0% (IQR 21.6% - 36.7%), by the most frequent EPA title was 20.2% (IQR 15.7% - 25.0%), and by the most frequent workplace setting was 51.2% (IQR 41.4% - 64.2%).

XSL•FO
RenderX

Table . Observed resident portfolio relative to projected decision-study portfolio configurations (N=466).

Portfolio configuration	Residents, n (%)
At least 20 observations	447 (95.9)
At least 30 observations	412 (88.4)
At least 5 faculty raters	416 (89.3)
At least 6 faculty raters	377 (80.9)
At least 5 EPA[a] titles	464 (99.6)
At least 6 EPA titles	462 (99.1)
At least 3 workplace settings	448 (96.1)
At least 4 workplace settings	402 (86.3)
20-observation well-distributed portfolio[b]	398 (85.4)
30-observation well-distributed portfolio[c]	330 (70.8)

[a]EPA: entrustable professional activity.

[b]The 20-observation well-distributed portfolio was defined as ≥20 observations, ≥5 faculty raters, ≥5 EPA titles, and ≥3 workplace settings.

[c]The 30-observation well-distributed portfolio was defined as ≥30 observations, ≥6 faculty raters, ≥6 EPA titles, and ≥4 workplace settings.

Sensitivity Analyses

Sensitivity analysis results are presented in Table S1 (Multimedia Appendix 1). In the EPA title × workplace setting sensitivity model, the EPA title × workplace setting interaction accounted for 0.3% of total variance, and residual/encounter-level variance accounted for 40.2%. In the hospital-aware model, resident hospital and faculty hospital accounted for 1.1% and 0.8% of total variance, respectively. The main variance-component pattern remained stable, with resident variance accounting for 25.8%, and residual/encounter-level variance accounting for 39.8%.

In the complexity-adjusted sensitivity model, which included 45,513 observations after excluding 13 records with missing case-complexity data, advanced/nonroutine cases were associated with entrustment-supervision ratings 0.34 points lower than basic/routine cases. Residual/encounter-level variance decreased modestly from 40.1% in the primary model to 39.1% after adjusting for case complexity, while the main variance-component pattern remained stable.

Discussion

Principal Findings and Implications

This nationwide digital platform study found that EPA-based progression review depends on both observation density and portfolio breadth. Single EPA observations were insufficient for dependable resident-level interpretation, whereas well-distributed portfolios showed higher projected dependability. A 20-observation portfolio across 5 faculty raters, 5 EPA titles, and 3 workplace settings reached the prespecified Phi ≥0.70 psychometric reference point, while a 30-observation portfolio across 6 faculty raters, 6 EPA titles, and 4 workplace settings reached the stricter Phi ≥0.80 reference point. These findings provide internal structure validity evidence for EPA-based workplace assessment and indicate that defensible progression judgments depend not only on assessment volume but also on purposeful sampling across assessors, tasks, and contexts.

The main contribution of this study is to reframe EPA portfolio dependability as a multidimensional sampling issue in programmatic assessment [9,25]. In many CBME implementations, portfolio completeness is often discussed primarily in terms of the number of completed assessments [2]. Our findings suggest that assessment count alone is insufficient for dependable resident-level interpretation. Instead, dependability depends on how assessment evidence is distributed across faculty raters, EPA titles, and workplace settings [23,24]. This perspective is relevant to medical education, providing a practical way for programs, competency committees, and digital assessment platforms to monitor not only whether residents have enough assessment records but also whether those records are sufficiently broad for the intended progression decision [20,35,36]. Therefore, the findings support portfolio-design thinking rather than simple numeric assessment targets.

Our findings extend prior work on entrustment-derived workplace assessment. Kelleher et al [23] identified rater variance as a major source of score variation in internal medicine assessments and showed that reliability improved with aggregation across raters and encounters. Ryan et al [14] multi-institutional Core EPA study found limited learner-attributable variance in many workplace-based assessment datasets and cautioned against using such assessments alone for high-stakes summative decisions. Tanaka et al [37] similarly showed that anesthesiology EPAs differed in their capacity to discriminate resident performance. Against this background, our findings suggest that EPA-based assessment can support dependable resident-level judgments when portfolios are deliberately sampled across raters, EPAs, and workplaces. The relatively higher dependability estimates observed in this study may reflect the postgraduate specialty context, alignment between EPAs and routine clinical work, greater resident-level performance variation, or features of national implementation [12].

XSL•FO
RenderX

The generalizability of these findings should be understood in 2 ways. The analytic framework may be transferable: other EPA-based or workplace-based assessment systems can use G-study and D-study methods to examine how portfolio dependability changes with assessment count, rater breadth, EPA breadth, and workplace-setting breadth [23-27]. However, the numerical estimates are context dependent. The projected portfolio configurations in this study reflect the Taiwan otorhinolaryngology EPA framework, the E-MyWay platform structure, local rating practices, workplace settings, and the variance components observed in this dataset. Therefore, these configurations should not be interpreted as universal requirements for demonstrating competence. Specialties with more or fewer EPAs, different case-mix structures, different assessment cultures, or predominantly nonprocedural or longitudinal clinical work would require context-specific modeling.

A central implication is that observation count is necessary but insufficient. Programmatic assessment often emphasizes accumulating multiple low-stakes observations, but a portfolio concentrated within a narrow set of raters, tasks, or settings may remain vulnerable to rater bias, task specificity, and contextual restriction [38]. Our results shift the question from "How many assessments are enough?" to "What configuration of evidence is dependable enough for the intended interpretation?" For CBME systems, portfolio monitoring should therefore be multidimensional, incorporating observation density, assessor diversity, EPA breadth, and workplace-setting breadth [24,39]. Digital platforms can operationalize this principle by tracking not only portfolio size but also rater diversity, EPA coverage, and workplace-setting breadth. With the resident as the object of measurement, these projected configurations refer to aggregate resident-level portfolio dependability across EPA-based assessment records, not certification-level dependability for each individual EPA. Therefore, they should not be interpreted as evidence that assessment of a subset of EPA titles is sufficient for entrustment across all core activities.

The variance-component findings reinforce this point. Resident variance accounted for a meaningful proportion of score variation, supporting the interpretation that EPA ratings captured differences in residents' entrustment readiness. However, resident variance should not be interpreted as a pure measure of competence, as it may reflect differences in training stage, exposure opportunity, program context, and individual performance. At the same time, residual/encounter-level variance was the largest component, and faculty rater variance and resident × faculty rater interaction variance were also substantial. This pattern should not be viewed simply as measurement failure. Workplace performance is situated, relational, and contingent on patient acuity, task complexity, clinical setting, and supervisor judgment. In surgical and procedurally intensive specialties, this contextual variability may be particularly pronounced, given that competence integrates diagnostic reasoning, procedural skill, perioperative decision-making, team coordination, and risk management. EPA ratings are therefore best understood as contextually sampled indicators of readiness rather than fixed traits revealed through isolated encounters [13].

The sensitivity analyses further supported the robustness of the primary variance structure. In the hospital-aware sensitivity analysis, resident-hospital and faculty-hospital variance components were small, and the main variance pattern remained stable after adding these contextual random effects. This finding supports the interpretation that the primary results reflected aggregate resident-level portfolio dependability rather than being driven mainly by hospital-level clustering. Nevertheless, hospital codes may not fully capture all programmatic differences in curriculum, clinical exposure, assessment culture, faculty development, or rater calibration [12,38].

The complexity-adjusted sensitivity model also supported the primary interpretation. Although advanced/nonroutine cases were associated with lower entrustment-supervision ratings, adjustment for measured case complexity only modestly reduced residual/encounter-level variance and did not materially alter the primary variance-component pattern. These findings suggest that measured case complexity contributed to rating variation but did not account for most encounter-level variability, reinforcing the need to aggregate evidence across multiple observations, faculty raters, EPA titles, and workplace settings. Future assessment implementation may reduce unexplained encounter-level variation by strengthening behavioral anchors, improving rater orientation, documenting case complexity more systematically, and ensuring that resident portfolios are sampled across multiple raters, EPA titles, and workplace settings.

This study contributes to a limited evidence base on EPA-based summative assessment in postgraduate surgical specialty training. Much of the existing reliability literature has emerged from undergraduate medical education or nonsurgical postgraduate contexts [27,40,41]. Otorhinolaryngology-head and neck surgery provides an informative test case, spanning medical and surgical practice, including acute airway and bleeding emergencies, head and neck infections, otologic and sinonasal disease, sleep-disordered breathing, head and neck tumors, and facial plastic and reconstructive conditions across emergency, inpatient, outpatient, procedural, and operating-room environments. The finding that dependable decisions are achievable only through broadly sampled portfolios suggests that EPA-based summative assessment can be psychometrically defensible in surgical specialty training when systems are designed to capture the breadth of clinical work rather than rely on convenience-based observations [42].

These findings have practical implications for clinical competency committees (CCCs). CCCs are often asked to make progression decisions from heterogeneous evidence but may lack explicit guidance for judging whether the quantitative observation evidence base is sufficiently broad. The projected portfolio configurations in this study provide model-based psychometric reference points: approximately 20 well-distributed observations may support lower-stakes resident-level portfolio review, whereas more consequential progression deliberation may require broader evidence, such as 30 observations across multiple faculty raters, EPA titles, and workplace settings. These configurations should not be applied mechanically as competence standards. CCCs must still incorporate narrative feedback, case complexity, performance trajectory, professionalism concerns, and contextual knowledge.

XSL•FO
RenderX

Rather, the configurations may help committees and programs judge whether quantitative entrustment data are sufficiently distributed to support defensible deliberation [35,36].

Importantly, the projected portfolio configurations should be interpreted as model-based psychometric evidence for portfolio design, not as validated standards for competence, certification, or promotion. This cross-sectional G-study/D-study estimated internal structure evidence and projected dependability under selected sampling configurations, but it did not evaluate whether these configurations improve the accuracy, fairness, or consequences of actual progression decisions [26,34,43]. Future studies should link EPA portfolio dependability to independent educational outcomes, including CCC decisions, remediation, delayed progression, graduation outcomes, and subsequent clinical performance [35,36].

The results also speak to assessment burden [2,11]. Excessive assessment expectations may overwhelm faculty and residents, encourage superficial completion, and weaken feedback quality, whereas insufficient sampling may lead to unstable high-stakes decisions. In this dataset, the median observed resident portfolio exceeded the stricter projected portfolio configuration, suggesting feasibility within a mature national digital assessment system. Yet approximately 71% (330/466) of residents met all criteria for the stricter 30-observation scenario, despite many having sufficient raw observation counts. This gap suggests that portfolio breadth, rather than assessment volume alone, may have constrained portfolio completeness for a substantial subgroup. Thus, implementation should focus not only on increasing the number of assessments but also on monitoring portfolio distribution. A platform dashboard could flag residents who meet raw observation counts but lack sufficient rater breadth, EPA-title breadth, or workplace-setting breadth before CCC review.

Finally, this study should be interpreted as one component of a broader validity argument [34,43]. Generalizability theory provided internal structure evidence by estimating the extent to which entrustment ratings reflected resident differences relative to rater, task, context, interaction, and residual/encounter-level variation. D-studies linked this evidence to intended use by projecting resident-level dependability under alternative portfolio designs. This approach aligns with contemporary validity thinking: validity resides not in the assessment tool alone but in the interpretation and use of assessment data for a specific decision. Formative EPA observations are context-dependent assessments intended primarily to support feedback and learning within specific clinical encounters. They should not be used individually, or in small numbers, to classify a resident as performing above or below stage-specific expectations. Such classification requires longitudinal synthesis of evidence through a summative process, such as CCC deliberation. Reliability is necessary but not sufficient; content alignment, response processes, relations to other variables, consequences, and CCC deliberation remain essential [23].

Limitations

This study has limitations. First, the numerical D-study projections were derived from one mature national digital workplace assessment platform in Taiwan otorhinolaryngology-head and neck surgery residency training. Therefore, they should not be generalized as universal portfolio requirements. The estimates are conditional on the Taiwan specialty EPA framework, E-MyWay platform structure, local rating practices, workplace settings, hospital/program context, and observed variance components. Although the hospital-aware sensitivity model showed small resident-hospital and faculty-hospital variance components, hospital codes may not fully capture program-level differences in curriculum, clinical exposure, assessment culture, faculty development, or rater calibration. Other specialties or assessment systems would require context-specific modeling.

Second, workplace-based assessment data were opportunistic and unbalanced, reflecting clinical workflow, case availability, faculty availability, resident rotation assignments, and local assessment culture rather than balanced or randomized sampling of resident performance. The portfolio concentration analysis characterized this imbalance but did not statistically correct sampling bias. Therefore, the D-study projections should be interpreted as model-based estimates from observed real-world assessment data, not as evidence from a controlled sampling design.

Third, faculty-assigned entrustment-supervision ratings were ordinal judgments but were treated as approximately numeric to permit variance decomposition and D-study projection within the classical G-theory framework. This analytic approximation simplifies the ordered categorical nature and the social complexity of entrustment judgment. Ordinal models, such as cumulative-link mixed models or many-facet ordinal models, may provide complementary evidence in future work.

Fourth, quantitative ratings do not capture the full richness of narrative feedback, resident reflection, performance trajectory, phase-of-care information, or CCC deliberation. Although case complexity was included in a sensitivity analysis, it was captured only as basic/routine versus advanced/nonroutine, and phase of care was not separately modeled. EPA12 was also excluded, owing to its inconsistent implementation across the analytic period. In addition, the D-study projections assumed well-distributed portfolios and should not be interpreted to mean that any 20 or 30 observations would yield equivalent dependability, nor that the projected configurations establish sufficient evidence for entrustment in each individual EPA.

Finally, the present study evaluated portfolio dependability and internal-structure evidence rather than whether shorter formative EPA portfolios could identify residents performing above or below stage-specific expectations. The analyzed EPA records were ad hoc, low-stakes, formative, and context dependent; each observation was intended to support feedback within a specific clinical encounter rather than summative classification of overall resident performance. Stage-relative classification generally requires longitudinal synthesis through CCC summative entrustment decisions. Although CCC decisions became available in the E-MyWay database beginning in January 2025, they were not available across the full analytic period and were not consistently temporally aligned with the formative EPA portfolios analyzed in this study. In addition, as CCC deliberations may incorporate longitudinal EPA evidence,

XSL•FO
RenderX

using contemporaneous CCC decisions to define performance subgroups would require careful temporal separation and complementary outcomes to minimize circular interpretation.

Future Directions

Future studies should prospectively align formative EPA portfolios accumulated before each CCC review with subsequent CCC summative entrustment decisions. Additional outcomes, including remediation, delayed progression, promotion, and subsequent clinical performance, should be incorporated to provide evidence beyond the EPA ratings themselves. Such longitudinal analyses could determine whether residents performing clearly above or below stage-specific expectations can be identified with fewer observations while accounting for the context-dependent nature of formative workplace assessment. These analyses should also examine whether portfolio configurations associated with prespecified dependability targets vary by postgraduate year and by individual EPA. Additional work should examine how narrative feedback, resident reflection, and case complexity can be integrated into summative validity arguments. Common scale anchors alone do not demonstrate successful rater calibration. Prospective studies should therefore use standardized cases, video-recorded encounters, or independently duplicated ratings to evaluate interrater agreement directly and determine whether structured faculty calibration improves rating consistency. Cross-specialty comparisons are needed to determine whether medical, surgical, and hybrid specialties require different sampling architectures. Finally, future research should test whether real-time learning analytics dashboards improve portfolio completeness, reduce assessment burden, and strengthen CCC preparedness.

Conclusions

This nationwide cross-sectional G- and D-study showed that digital EPA assessment portfolios require purposeful sampling across faculty raters, EPA titles, and workplace settings to support dependable resident-level progression review. The findings support study-defined projected portfolio configurations, including a 30-observation well-distributed portfolio that reached the stricter Phi ≥0.80 psychometric reference point. These model-based dependability estimates should complement, rather than replace, narrative evidence, case complexity, performance trajectory, and CCC deliberation.

Acknowledgments

We thank the Taiwan Society of Otorhinolaryngology-Head and Neck Surgery and all participating faculty members and resident physicians for using the Joint Commission of Taiwan's E-MyWay platform. We also thank the information technology team of Dalin Tzu Chi Hospital for their support with the platform. Additionally, we are grateful for the administrative assistance provided by Chiu-Ping Wang, Shu-Hwei Fan, Uan-Shr Jan, and Wan-Ning Luo in this project. They received no additional compensation for their contributions. The content is solely the responsibility of the authors and does not necessarily represent the official views of the Taiwan Society of Otorhinolaryngology-Head and Neck Surgery. The authors declare the use of generative AI (GAI) in the research and writing process. According to the GAIDeT taxonomy (2025), the following tasks were delegated to GAI tools under full human supervision: proofreading and editing. The GAI tool used was ChatGPT (GPT-Thinking 5.5; OpenAI). Responsibility for the final manuscript lies entirely with the authors. GAI tools are not listed as authors and do not bear responsibility for the final outcomes. Declaration submitted by JWC.

Funding

This study was supported by the National Science and Technology Council of the Republic of China (Taiwan) under grants NSTC 110-2511-H-567-001-MY2 and NSTC 112-2410-H-567-001-MY3, and was partially funded by Cardinal Tien Hospital under grants CTH113AK-2220 and CTH114A-NDMC-2229. This study was also funded by the Ministry of Health and Welfare of the Republic of China (Taiwan) under the Healthy Taiwan Sprout Project grant C-0002. The funders had no role in the design or conduct of the study; the collection, management, analysis, or interpretation of the data; the preparation, review, or approval of the manuscript; or the decision to submit the manuscript for publication.

Data Availability

The deidentified data analyzed in this study are not publicly available because they were obtained from the Taiwan Society of Otorhinolaryngology–Head and Neck Surgery E-MyWay database with permission and include trainee assessment records. Reasonable requests for data access may be considered subject to society approval, institutional review requirements, and applicable data-use agreements.

Authors' Contributions

Conceptualization: JWC, CLH, CHC, WCH, PCW, LJL
Data curation: ODL, CLH, JWC
Formal analysis: ODL, CLH, YHW
Funding acquisition: WCH, PCW, JWC
Methodology: JWC, ODL, CLH, YHW
Project administration: JWC, WCH, CHL, MC

XSL•FO
RenderX

Supervision: JWC, CLH, WCH, CHL, MC
Writing – original draft: JWC, CLH, CHC
Writing – review and editing: JWC, CLH, CHC, YHW, WCH, PCW, LJL, CHL, MC, ODL.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Supplementary generalizability and decision study model specifications, dependability projections, and sensitivity variance-component analyses.
[DOCX File, 179 KB - mededu_v12i1e104628_app1.docx]

Checklist 1
STROBE checklist.
[DOCX File, 30 KB - mededu_v12i1e104628_app2.docx]

References

1. Hsiao CT, Chou FC, Hsieh CC, Chang LC, Hsu CM. Developing a competency-based learning and assessment system for residency training: analysis study of user requirements and acceptance. J Med Internet Res 2020 Apr 14;22(4):e15655. [doi: 10.2196/15655] [Medline: 32286233]
2. Szulewski A, Braund H, Dagnone DJ, et al. The assessment burden in competency-based medical education: how programs are adapting. Acad Med 2023 Nov 1;98(11):1261-1267. [doi: 10.1097/ACM.0000000000005305] [Medline: 37343164]
3. Domann M, Richters C, Witti MJ, Stadler M. Medical students' acceptance of digital entrustable professional activities: results of a cohort study. JMIR Med Educ 2026 May 4;12:e87605. [doi: 10.2196/87605] [Medline: 42081711]
4. Long DM. Competency-based residency training: the next advance in graduate medical education. Acad Med 2000 Dec;75(12):1178-1183. [doi: 10.1097/00001888-200012000-00009] [Medline: 11112714]
5. Alharbi NS. Evaluating competency-based medical education: a systematized review of current practices. BMC Med Educ 2024 Jun 3;24(1):612. [doi: 10.1186/s12909-024-05609-6] [Medline: 38831271]
6. ten Cate O, Scheele F. Competency-based postgraduate training: can we bridge the gap between theory and clinical practice? Acad Med 2007 Jun;82(6):542-547. [doi: 10.1097/ACM.0b013e31805559c7] [Medline: 17525536]
7. Ten Cate O, Taylor DR. The recommended description of an entrustable professional activity: AMEE Guide No. 140. Med Teach 2021 Oct;43(10):1106-1114. [doi: 10.1080/0142159X.2020.1838465] [Medline: 33167763]
8. Mun M, Chanchlani S, Lyons K, Gray K. Transforming the future of digital health education: redesign of a graduate program using competency mapping. JMIR Med Educ 2024 Oct 31;10:e54112. [doi: 10.2196/54112] [Medline: 39481124]
9. Pack R, Lingard L, Watling C, Cristancho S. Beyond summative decision making: illuminating the broader roles of competence committees. Med Educ 2020 Jun;54(6):517-527. [doi: 10.1111/medu.14072] [Medline: 31991000]
10. Preiksaitis C, Hughes J, Kabeer R, Dixon W, Rose C. Quantifying emergency medicine residency learning curves using natural language processing: retrospective cohort study. JMIR Med Educ 2025 Dec 9;11:e82326. [doi: 10.2196/82326] [Medline: 41364786]
11. Nel D, Marty AP, Frick S, Hennus MP, Linsenmeyer M. Addressing practical and conceptual challenges in workplace-based assessment. In: ten Cate O, Burch VC, Chen HC, Chou FC, Hennus MP, editors. Entrustable Professional Activities and Entrustment Decision-Making in Health Professions Education, 1st edition: Ubiquity Press; 2024:237-247. [doi: 10.5334/bdc.t]
12. Driessen E, Scheele F. What is wrong with assessment in postgraduate training? Lessons from clinical practice and educational research. Med Teach 2013 Jul;35(7):569-574. [doi: 10.3109/0142159X.2013.798403] [Medline: 23701250]
13. Brown A, La J, Keri MI, et al. In EPAs we trust, is quality and safety a must? A cross-specialty analysis of entrustable professional activity guides. Med Teach 2025 Jan;47(1):134-142. [doi: 10.1080/0142159X.2024.2323177] [Medline: 38527417]
14. Ryan MS, Gielissen KA, Shin D, et al. How well do workplace-based assessments support summative entrustment decisions? A multi-institutional generalisability study. Med Educ 2024 Jul;58(7):825-837. [doi: 10.1111/medu.15291] [Medline: 38167833]
15. Cabrera-Muffly C, Cusumano C, Freeman M, et al. Milestones 2.0: otolaryngology resident competency in the postpandemic era. Otolaryngol Head Neck Surg 2022 Apr;166(4):605-607. [doi: 10.1177/01945998211033235] [Medline: 34313511]
16. Chen JX, Yu SE, Miller LE, Gray ST. A needs assessment for the future of otolaryngology education. Otolaryngol Head Neck Surg 2023 Jul;169(1):192-193. [doi: 10.1177/01945998221128292] [Medline: 36125895]
17. Lin CP, Hsu WC, Wang GL, et al. Factors associated with attainment of ad hoc entrustability among Taiwan otolaryngology resident physicians: a nationwide cross-sectional study. BMC Med Educ 2025 Aug 1;25(1):1135. [doi: 10.1186/s12909-025-07679-6] [Medline: 40751193]

XSL•FO
RenderX

18. Chen JW, Tu HL, Chang CH, et al. Automated evaluation of reflection and feedback quality in workplace-based assessments by using natural language processing: cross-sectional competency-based medical education study. JMIR Med Educ 2025 Oct 22;11:e81718. [doi: 10.2196/81718] [Medline: 41124694]
19. Chen JW, Yu RB, Lin KN, et al. Learning analytics of a national entrustable professional activities platform: cross-sectional study of system-level constraints on advanced entrustment in competency-based medical education. JMIR Med Educ 2026 May 27;12:e95066. [doi: 10.2196/95066] [Medline: 42202270]
20. Guo FC, Chen YT, Hsu WC, Wang PC, Chen M, Chen JW. EMYWAY workplace-based entrustable professional activities assessments in otolaryngology residency training: a nationwide experience. Otolaryngol Head Neck Surg 2025 Apr;172(4):1242-1253. [doi: 10.1002/ohn.1104] [Medline: 39739526]
21. Bojic I, Mammadova M, Ang CS, et al. Empowering health care education through learning analytics: in-depth scoping review. J Med Internet Res 2023 May 17;25:e41671. [doi: 10.2196/41671] [Medline: 37195746]
22. Chiang YH, Yu HC, Chung HC, Chen JW. Implementing an entrustable professional activities programmatic assessments for nurse practitioner training in emergency care: a pilot study. Nurse Educ Today 2022 Aug;115:105409. [doi: 10.1016/j.nedt.2022.105409] [Medline: 35636245]
23. Kelleher M, Kinnear B, Sall D, et al. A reliability analysis of entrustment-derived workplace-based assessments. Acad Med 2020 Apr;95(4):616-622. [doi: 10.1097/ACM.0000000000002997] [Medline: 31567170]
24. Dunne D, Gielissen K, Slade M, Park YS, Green M. WBAs in UME-how many are needed? A reliability analysis of 5 AAMC core EPAs implemented in the internal medicine clerkship. J Gen Intern Med 2022 Aug;37(11):2684-2690. [doi: 10.1007/s11606-021-07151-3] [Medline: 34561828]
25. Briesch AM, Swaminathan H, Welsh M, Chafouleas SM. Generalizability theory: a practical guide to study design, implementation, and interpretation. J Sch Psychol 2014 Feb;52(1):13-35. [doi: 10.1016/j.jsp.2013.11.008] [Medline: 24495492]
26. Kinnear B, Kelleher M, May B, et al. Constructing a validity map for a workplace-based assessment system: cross-walking Messick and Kane. Acad Med 2021 Jul 1;96(7S):S64-S69. [doi: 10.1097/ACM.0000000000004112] [Medline: 34183604]
27. Suneja M, Hanrahan KD, Kreiter C, Rowat J. Psychometric properties of entrustable professional activity-based objective structured clinical examinations during transition from undergraduate to graduate medical education: a generalizability study. Acad Med 2025 Feb 1;100(2):179-183. [doi: 10.1097/ACM.0000000000005719] [Medline: 38534117]
28. Choo EK, Woods R, Walker ME, O'Brien JM, Chan TM. The quality of assessment for learning score for evaluating written feedback in anesthesiology postgraduate medical education: a generalizability and decision study. Can Med Educ J 2023 Dec;14(6):78-85. [doi: 10.36834/cmej.75876] [Medline: 38226296]
29. Calhoun AW, Scerbo MW. Preparing and presenting validation studies: a guide for the perplexed. Simul Healthc 2022 Dec 1;17(6):357-365. [doi: 10.1097/SIH.0000000000000667] [Medline: 35470343]
30. Ryan MS, Richards A, Perera R, et al. Generalizability of the Ottawa Surgical Competency Operating Room Evaluation (O-SCORE) scale to assess medical student performance on core EPAs in the workplace: findings from one institution. Acad Med 2021 Aug 1;96(8):1197-1204. [doi: 10.1097/ACM.0000000000003921] [Medline: 33464735]
31. Johnson J, Schwartz A, Lineberry M, Rehman F, Park YS. Development, administration, and validity evidence of a subspecialty preparatory test toward licensure: a pilot study. BMC Med Educ 2018 Aug 1;18(1):176. [doi: 10.1186/s12909-018-1294-z] [Medline: 30068394]
32. O'Keeffe DA, Traynor O, Tekian A, Park YS. Evaluating the validity of national multiassessment system in postgraduate surgical training: a retrospective cohort study. J Surg Educ 2024 Nov;81(11):1709-1719. [doi: 10.1016/j.jsurg.2024.08.008] [Medline: 39299054]
33. Labbé M, Young M, Nguyen LHP. Validity evidence as a key marker of quality of technical skill assessment in OTL-HNS. Laryngoscope 2018 Oct;128(10):2296-2300. [doi: 10.1002/lary.27085] [Medline: 29330861]
34. Cook DA, Brydges R, Ginsburg S, Hatala R. A contemporary approach to validity arguments: a practical guide to Kane's framework. Med Educ 2015 Jun;49(6):560-575. [doi: 10.1111/medu.12678] [Medline: 25989405]
35. Schumacher DJ, Martini A, Bartlett KW, et al. Key factors in clinical competency committee members' decisions regarding residents' readiness to serve as supervisors: a national study. Acad Med 2019 Feb;94(2):251-258. [doi: 10.1097/ACM.0000000000002469] [Medline: 30256253]
36. Duitsman ME, Fluit CRMG, van Alfen-van der Velden JAEM, et al. Design and evaluation of a clinical competency committee. Perspect Med Educ 2019 Feb;8(1):1-8. [doi: 10.1007/s40037-018-0490-1] [Medline: 30656533]
37. Tanaka P, Park YS, Chen J, Macario A. Reliability and utility of anesthesiology entrustable professional activities assessed with a mobile web application. J Clin Anesth 2025 Sep;106(111922):111922. [doi: 10.1016/j.jclinane.2025.111922] [Medline: 40609217]
38. Branfield Day L, Miles A, Ginsburg S, Melvin L. Resident perceptions of assessment and feedback in competency-based medical education: a focus group study of one internal medicine residency program. Acad Med 2020 Nov;95(11):1712-1717. [doi: 10.1097/ACM.0000000000003315] [Medline: 32195692]
39. Fahim C, Wagner N, Nousiainen MT, Sonnadara R. Assessment of technical skills competence in the operating room: a systematic and scoping review. Acad Med 2018 May;93(5):794-808. [doi: 10.1097/ACM.0000000000001902] [Medline: 28953567]

XSL•FO
RenderX

40. Violato C, Cullen MJ, Englander R, et al. Validity evidence for assessing entrustable professional activities during undergraduate medical education. Acad Med 2021 Jul 1;96(7S):S70-S75. [doi: 10.1097/ACM.0000000000004090] [Medline: 34183605]
41. Violato C, Englander R, Dale E, Gauer JL. Implementing core entrustable professional activities in undergraduate medical education: a psychometric study. Acad Med 2025 May 1;100(5):585-591. [doi: 10.1097/ACM.0000000000005907] [Medline: 39485120]
42. Sigurdsson V, Ten Cate O. Do summative entrustment decisions actually lead to entrustment? Clin Teach 2024 Feb;21(1):e13668. [doi: 10.1111/tct.13668] [Medline: 37817015]
43. Tavares W, Rowland P, Dagnone D, McEwen LA, Billett S, Sibbald M. Translating outcome frameworks to assessment programmes: implications for validity. Med Educ 2020 Oct;54(10):932-942. [doi: 10.1111/medu.14287] [Medline: 32614480]

Abbreviations

CBME: competency-based medical education
CCC: clinical competency committee
D-study: decision study
EPA: entrustable professional activity
G-study: generalizability study
SEM: standard error of measurement
STROBE: Strengthening the Reporting of Observational Studies in Epidemiology
TSO-HNS: Taiwan Society of Otorhinolaryngology-Head and Neck Surgery

Edited by A Stone; submitted 18.Jun.2026; peer-reviewed by AS Moura, F Rizzuti, M Libby Weaver; revised version received 04.Aug.2026; accepted 26.Aug.2026; published 05.Oct.2026.

Please cite as:
Chen JW, Hsieh CL, Chang CH, Wang YH, Hsu WC, Wang PC, Liao LJ, Liao CH, Chen M, Laksmita OD
Dependability of Entrustable Professional Activity Portfolios in Otorhinolaryngology Residency: Nationwide Generalizability and Decision Study
JMIR Med Educ 2026;12:e104628
URL: https://mededu.jmir.org/2026/1/e104628
doi:10.2196/104628

© Jeng-Wen Chen, Ching-Lin Hsieh, Chun-Hsiang Chang, Ya-Hui Wang, Wei-Chung Hsu, Pa-Chun Wang, Li-Jen Liao, Chun-Hou Liao, Mingchih Chen, Okki Dhona Laksmita. Originally published in JMIR Medical Education (https://mededu.jmir.org), 5.Oct.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Education, is properly cited. The complete bibliographic information, a link to the original publication on https://mededu.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Attitudes Toward and Preconditions for Digital Game–Based Learning Among Public Health and Health Sciences Students: Cross-Sectional Online Survey Study

Kamil J Wrona[1], Dr PH; Joanna Albrecht[2], Dr PH; Anne Krümmel[3], MSc; Leona Aschentrup[3], MSc; Christoph Dockweiler[2], Dr PH; Timothy Mc Call[4], Dr PH

[1]School of Public Health, Bielefeld University, Universitätsstraße 25, Bielefeld, Germany

[2]Department of Social Sciences, Faculty of Arts and Humanities, University of Siegen, Siegen, Germany

[3]School of Public Health, Bielefeld University, Universitätsstraße 25, Bielefeld, Germany

[4]School of Public Health, Medical School OWL, Bielefeld University, Bielefeld, Germany

Corresponding Author:
Kamil J Wrona, Dr PH
School of Public Health, Bielefeld University, Universitätsstraße 25, Bielefeld, Germany

Abstract

Background: Digital game–based learning (DGBL) is gaining traction in medical and nursing education, but its integration into public health and health sciences curricula remains limited. This study addresses this research gap by exploring student perspectives, use patterns, and expectations regarding DGBL within German public health study programs.

Objective: The study investigated how public health and health sciences students perceive DGBL and identified conditions that promote or hinder its integration into academic training for public health professionals.

Methods: A cross-sectional online survey was conducted from January 2021 to March 2024. A total of 444 students participated, with 326 completing the questionnaire. The instrument combined closed- and open-ended questions on sociodemographic characteristics, digital media use, attitudes toward DGBL, and structural and pedagogical requirements. Quantitative data were analyzed descriptively using SPSS; qualitative data were evaluated using MAXQDA and Kuckartz structured content analysis.

Results: The sample consisted predominantly of full-time students enrolled in bachelor's or master's programs in public health and health communication. Digital media are widely used for communication (n=320, 98.2%), information search (n=291, 89.3%), and entertainment (n=268, 82.2%) in daily student routines. In contrast, video gaming and online shopping play a minor role. Students expressed a generally positive attitude toward DGBL, particularly in practical learning contexts such as exam preparation, visualizing complex content, and knowledge checks. High agreement was found for using DGBL in modules such as Prevention and Health Promotion (n=277, 84.9%), Epidemiology and Statistics (n=271, 83.2%), and Population Medicine (n=251, 76.7%). Preferred devices include desktop PCs, tablets, and smartphones, while immersive technologies such as virtual reality or augmented reality glasses or consoles are less favored. Students highlighted specific prerequisites for DGBL integration: a functioning technical infrastructure (n=321, 98.5%), media-literate teaching staff (n=311, 95.4%), and compliance with data protection standards (n=303, 93%). Furthermore, students valued user-friendly design (n=255, 78.2%), technical reliability (n=299, 91.7%), clear feedback (n=236, 72.4%), and flexibility regarding the time and place of use (n=307, 94.2%). Game elements such as animations, multimedia, and reward systems were appreciated, whereas features such as 3D graphics, avatars, and multiplayer modes were considered less important.

Conclusions: The findings show a high willingness among students to engage with DGBL, especially when its design emphasizes clarity, usability, and relevance to course content. Emotional safety, flexible access, and pedagogical integration are more decisive than immersive game features. To effectively implement DGBL in public health education, developers and educators should cocreate low-threshold, didactically sound applications that align with students' preferences and study realities. Future research should assess learning outcomes and emotional effects longitudinally and explore scalable implementation strategies.

(JMIR Med Educ 2026;12:e83347) doi:10.2196/83347

KEYWORDS

serious games; digital game–based learning; education; survey; public health; health sciences

XSL•FO
RenderX

Introduction

Background

The teaching and training of public health professionals face challenges resulting from globalization, demographic and epidemiological changes, and increasing socially induced health inequalities. To develop solutions that serve the interests of populations, it is crucial for research and practice that students acquire the necessary qualifications to develop interdisciplinary approaches to policy issues [1]. It is challenging to train students at the Bachelor's, Master's, and doctoral levels to acquire comparable, internationally standardized skills that meet the diverse requirements of the labor market [2]. Digital teaching and learning methods can support the acquisition of skills in this regard and have also been introduced into university didactics at German universities [3-5]. Digital teaching and learning methods may include the use of PowerPoint, online learning platforms with lecture recordings, or online-based exercises and examinations [6]. Digital formats such as serious games or multimedia tools, in which students use online videos to create and learn their own explanatory models, are also already being used in teaching practice [7]. However, few concepts have yet been used within Public Health or Health Sciences programs, although the advantages of e-learning have been clearly identified [8].

Academic Training for Public Health Professionals

Today, public health is understood by the World Health Organization (WHO) as the science and professional practice needed to prevent disease, prolong life, and promote psychosocial and physical health [9]. It is seen not only as a theory but also as a scientific discipline that ideally operates in all policy areas ("health in all policies") [10] and forms the basis for the practice of population-based health policy [1]. German universities, as well as universities of applied sciences and arts, impart public health knowledge to academically qualified professionals in bachelor's, master's, and doctoral programs. Public health education aims to train experts to contribute to planning and decision-making processes in health care and health-related policy fields; qualify them to develop, implement, and evaluate health prevention and promotion measures; and prepare them for research and teaching in Public Health [11].

To help public health professionals develop the knowledge, skills, and abilities to act appropriately to achieve set goals, the Association of Schools of Public Health in the European Region (ASPHER) lists competencies related to working with individuals and groups as well as organizations and systems. They range from methods and population health to policy, economics, and management. Prospective public health professionals are expected to generate new knowledge in the field of public health using a range of scientific methods, to develop the capacity for critical reflection, and to review and evaluate existing knowledge from different disciplines. They need to be able to distinguish between scientific and practical knowledge as a basis for application- and needs-oriented public health research and practice. In addition, they should develop broad, interdisciplinary thinking and working skills in the context of lifelong learning, particularly in view of increasingly frequent job changes. Therefore, interdisciplinarity is not only a core element of public health for solving public health problems but also a central challenge for professional development in Germany [12].

On the basis of the development of public health with regard to constantly changing environments and target groups, Blättner and Dierks [2] formulate 4 central challenges for public health teaching: medical vs population competence in public health; views vs perspectives of public health students; individual vs population medicine and public health ethics; and national vs international public health, including the role of the state in public health.

Digital Game Applications in the Field of Health

Play and learning are evolutionarily anchored in human development and are closely linked [13]. In the context of international higher education, the training of medical professionals (eg, physicians and nurses) has already begun to use sophisticated e-learning applications under the designation of *digital serious games* and *digital game–based learning* (DGBL) [3].

DGBL has been defined by Prensky [14] as "[...] any learning game on a computer or online" and thus refers exclusively to digital games. The definition emphasizes the learning process in the context of the game. The user should perceive themselves as learning through a high level of engagement, predominantly as if they were playing [14]. Serious games in research and practice currently refer to digital applications such as computer or video games, although there is no clear consensus on the definition [15]. However, it is clear that the focus is on education rather than entertainment [15,16]. Tolks and Lampert [17] have developed the following criteria for serious games: the aim of the game is to convey content that is not primarily for entertainment purposes, the game must contain pedagogical aspects, and these pedagogical structures must be subordinate to the entertaining factors.

The design and development of serious games for the education of health professionals are also very diverse, and only a few publications report on the (software) development process used for the games [18]. In the context of higher education, serious games and DGBL have already been successfully used in various study programs [4,19-21]. For example, Peng [22] found a positive effect of an educational serious game in the area of healthy eating in the public health subfield of health communication and prevention. In this context, an increase in nutrition and weight-related knowledge, an intention to eat healthier, and greater self-efficacy were achieved. In the field of nursing, Chang et al [23] reported that nursing students who participated in an integrated DGBL program showed significantly higher learning performance, self-efficacy, willingness to learn, and learning satisfaction than those in a control group. As some suitable software already exists in the training of medical personnel, the use of such gaming software for public health professionals is also conceivable. There is also considerable overlap between the 2 professions regarding the acquisition of competencies, although there are only a few serious games available for these target groups.

XSL•FO
RenderX

Several reviews document the effectiveness of various serious gaming elements in (academic) health care education [24-27]. An initial review emphasizes that games can increase learning enjoyment and improve knowledge retention in the long term, although the evidence is limited in some cases [27]. More recent studies show that serious games are at least as effective as traditional learning methods and can improve knowledge, skills, and educational satisfaction [26]. Tori et al [24] also show that gamification elements such as scoring systems and level advancement achieve positive learning effects, particularly in physiotherapy, psychology, and physical education. Another review examined adaptive e-learning elements and found that their effectiveness is inconsistent. However, the adaptation process appears to be more beneficial for the acquisition of skills than for the acquisition of pure factual knowledge [25].

Although numerous studies have demonstrated the effectiveness of serious games in higher education within the health sector, research on the specific attitudes of public health students toward the use of DGBL in public health study programs or health sciences remains limited. There is a lack of empirical evidence on which applications are perceived as beneficial and which factors influence their acceptance and implementation. The aim of the present study is to provide a deeper understanding of health sciences students' attitudes toward DGBL in public health teaching as well as its possible applications. For this purpose, the application possibilities, performance, and effort expectations as well as the facilitating and inhibiting conditions for the implementation of digital games in public health education at universities will be explicated. Ultimately, the results of the study will allow conclusions to be drawn about possible uses and the design of serious games to promote their implementation and acceptance in public health study programs or health sciences. The following research questions will be explored:

- What are the attitudes of students toward the use of DGBL in public health study programs or health sciences?
- What are the advantages of using DGBL in public health study programs or health sciences?
- What factors promote and inhibit the use of DGBL in public health study programs or health sciences from the perspective of students?

Methods

Overview

To obtain an assessment of the possible future use of DGBL in public health study programs or health sciences and to structure the associated attitudes and needs of students, a structured web-based survey was conducted. The cross-sectional online survey was developed during the winter semester of 2020-2021 as part of a learning module of the Master of Public Health program at Bielefeld University and administered via the Unipark (Tivian XI GmbH) survey platform.

Ethical Considerations

Ethics approval to conduct the study was obtained from Bielefeld University. The ethics committee of Bielefeld University reviewed the research project in accordance with the ethical guidelines of the German Psychological Society and the Professional Association of German Psychologists and found the study to be ethically unobjectionable. The application has been filed under the number 2023-243.

Study Design and Setting

The questionnaire comprised 22 main questions (or statements) with 130 subquestions and required approximately 20 to 30 minutes to complete. It consisted of closed- and open-ended questions divided into 7 content units, including information on sociodemographic data, course of study, current use of digital media, and students' affinity for technology. Students' affinity for technology was assessed using the validated *Technikaffinität–Elektronische Geräte* (TA-EG; Affinity for Technology–Electronic Gadgets) questionnaire [28], which includes the subscales of enthusiasm for technology, perceived competence in handling technology, trust in technological systems, and technology-related fears. Internal consistency of the subscales ranged from acceptable to good (Cronbach α). Enthusiasm for technology showed good internal consistency (α=0.8), while perceived competence in handling technology (α=0.74), trust in technological systems (α=0.65), and technology-related fears (α=0.69) showed acceptable internal consistency. In addition, the questionnaire included self-developed items on attitudes toward DGBL, students' personal assessment of digital media use during their studies, motivation, and social influences on opinion formation and acceptance. The DGBL items were based on a standardized point-and-click adventure game scenario developed for the Department of Epidemiology.

Although the TA-EG scales [28] are a validated instrument, all additional items were developed specifically for this study to capture attitudes and behaviors related to DGBL in the target population. These items were reviewed by experts in public health education to ensure content validity and clarity.

The DGBL scenario and survey items were subsequently pilot-tested with a small group of 5 students from different study programs. Participants completed the survey and provided qualitative feedback in a structured debriefing session focusing on clarity, wording, and technical implementation. On the basis of this feedback, minor revisions were made to improve comprehensibility and usability. Following the pilot phase, a pretest was conducted with 24 students at Bielefeld University during a course session. The pretest focused on feasibility aspects, including completion time, clarity of items, and technical functionality. As the study primarily relied on previously validated instruments, no formal statistical analyses of reliability or validity were conducted at this stage. Instead, the pretest served to evaluate the overall feasibility of the study procedures.

Participants completed the full survey under real study conditions. Subsequently, feasibility aspects and questionnaire structure were discussed, and minor wording adjustments were implemented based on participant feedback. In addition, technical and content-related aspects of the scenario were refined based on pilot testing to ensure clarity and usability across different devices and study contexts.

XSL•FO
RenderX

The DGBL scenario was implemented as a standardized point-and-click adventure. All participants were provided with the same online version, including identical instructions, navigation options, and interactive elements. The scenario was designed to reflect realistic decision-making situations relevant to public health practice.

Recruitment and Data Collection

To include different student cohorts, the survey was extended across several semesters. The survey was conducted from January 2021 to March 2024 to include students from multiple semesters and study years across the participating health-related programs. This extended period allowed the study to capture attitudes and experiences from both earlier- and later-stage students, ensuring representation of different cohorts within the programs. The survey was administered in German via an online platform targeting students in health-related programs, including public health, Prevention and Health Promotion, and Medicine, regardless of study level (bachelor's or master's). Participants were recruited using an open, convenience sampling approach: invitations were sent via university email distribution lists. For Fulda University of Applied Sciences, one of the authors has a teaching assignment and coordinated access to the student distribution lists in collaboration with the university administration. In addition, the survey was promoted in relevant social media groups on Facebook (Meta Platforms Inc), LinkedIn (Microsoft Corp), Xing (New Work SE), and WhatsApp (Meta Platforms Inc). Three reminders were sent to the institutions enlisted in Textbox 1 to reach out to previously contacted cohorts as well as newly enrolled cohorts in public health or health sciences programs.

Eligibility criteria were minimal. Given the exploratory nature of the study, no formal sample size calculation was conducted.

Textbox 1. Overview of universities for the recruitment of the sample.

Bielefeld University

- Bachelor of Science
- Master of Science
- Doctoral program
- Bachelor of Arts
- Master of Arts
- Postgraduate studies
- Guest students

Siegen University

- Bachelor of Science
- Master of Science

Fulda University of Applied Sciences

- Bachelor of Science
- Master of Science

Bielefeld University of Applied Sciences and Arts

- Bachelor of Science
- Master of Science
- Bachelor of Arts
- Master of Arts

Statistical Analysis

The data were imported from the Unipark survey platform into the respective analysis software and processed. To structure the analysis and evaluation of the data, variables were checked for their suitability for the corresponding questions, assigned, and coded accordingly.

For the answers to closed questions, the responses of all participants who fully completed the survey were considered. SPSS (version 30; IBM Corp) software was used for the descriptive statistical analysis (frequency and location parameter analyses) of the dataset. For open-ended questions, all participants who provided an answer were included, and the responses were translated into English. The responses were then content analyzed according to Rädiker and Kuckartz [29] using MAXQDA (VERBI Software) analysis software.

Results

Sample Description and Distribution of Characteristics

A total of 326 students from health-related programs completed the questionnaire. The majority were female (n=277), and the largest age group was 18 to 25 years (n=123). Most participants

XSL•FO
RenderX

were enrolled in public health (n=114) or health communication (n=98) programs, studied full time (n=298), and were pursuing either a bachelor's degree (n=198) or a master's degree (n=121). Furthermore, the median study duration was 4 semesters (Table 1).

XSL•FO
RenderX

Table . Demographic data of the sample (N=326).

Demographic characteristics	Participants, n (%)
Gender[a]	
Female	277 (85)
Male	42 (12.9)
Diverse	3 (0.9)
Not specified	2 (0.6)
Age (years)[b]	
18-25	123 (37.7)
26-30	33 (10.1)
31-35	14 (4.3)
36-40	4 (1.2)
>40	7 (2.1)
Degree program	
Public Health	114 (35)
Health Communication	98 (30.1)
Health Sciences	35 (10.7)
Health Promotion	32 (9.8)
Health	11 (3.4)
Health Education	10 (3.1)
Health Promotion and Management	6 (1.8)
Applied Health Sciences	5 (1.5)
International Health	5 (1.5)
Public Health Nutrition	4 (1.2)
Health Technology	2 (0.6)
Digital Biomedical and Health Science	1 (0.3)
Health and Nursing Sciences	1 (0.3)
Physiotherapy	1 (0.3)
Dietetics	1 (0.3)
Type of program	
Bachelor	198 (60.7)
Master	121 (37.1)
Doctoral program or free doctorate	3 (0.9)
Postgraduate studies or second degree program	2 (0.6)
Guest auditor	1 (0.3)
Not specified	1 (0.3)
Organizational form of the study program	
Full-time program	298 (91.4)
Part-time program	26 (8)
Doctoral program	1 (0.3)
Visiting program	1 (0.3)
Semester	
1	22 (6.7)

XSL•FO
RenderX

Demographic characteristics	Participants, n (%)
2	71 (21.8)
3	27 (8.3)
4	70 (21.5)
5	35 (10.7)
6	46 (14.1)
>6	55 (16.9)

[a]Data on "gender" were available for 324 participants; percentages for individual categories are based on the total sample (N=326).

[b]Data on "age" were available for 181 participants; percentages for individual categories are based on the total sample (N=326).

Overall affinity for technology (TA-EG score) was moderate (median 2.95, SD 0.38). Trust in technology showed the highest mean score (median 3.6, SD 0.56), followed by fear-related attitudes (mean 3, SD 0.44) and competence in using technology (median 2.75, SD 0.44). Surprisingly, enthusiasm had the lowest values among the subscales (median 2.2, SD 0.73; Table 2).

Table . *Technikaffinität–Elektronische Geräte* (Affinity for Technology–Electronic Gadgets) score of the sample (N=326).

Scales	Score, median (SD)
Enthusiasm for technology	2.2 (0.72863)
Competence in using technology	2.75 (0.43531)
Fear of technology	3 (0.43722)
Trust in technology	3.6 (0.562)
Overall score	2.9474 (0.3812)

Use of Digital Media in General and in Everyday University Life

Digital media were most frequently used for communication (n=320, 98.2%), information search (n=291, 89.3%), and entertainment (n=268, 82.2%), while work-related activities occurred daily in approximately half of the sample (n=180, 55.2%). Video gaming and online shopping were less common (Multimedia Appendix 1).

Most students preferred desktop PCs, tablets, or smartphones as end devices for DGBL and expressed a positive intention to use DGBL in the future (agree: n=140, 42.9%; tend to agree: n=94, 28.8%). DGBL was perceived as particularly useful in modules such as Prevention and Health Promotion, Epidemiology and Statistics, and Population Medicine, with practical applications (including exam preparation, knowledge checks, and learning new content) rated most positively. Open-ended responses emphasized that DGBL should supplement, not replace, face-to-face teaching and support self-directed learning.

Gaming motivations were primarily entertainment, relaxation, and boredom, while competitive or stress-relief motives were less relevant. Digital media in everyday university life were rated as most important for internet access, PCs or tablets, and streaming services, while print media, computer games, and educational games were less relevant. Regarding the specific use of educational games, most participants stated that they did not use them (n=306, 93.9%). Only a small minority of students (n=20, 6.1%) reported using educational games, most frequently Duolingo (Duolingo Inc) and Babbel (Babbel AG), while other apps (eg, Atomas, BioApp, Cross Logic, Icare, Comma Rules, King of Mathematics, educational games for children, 10 Finger Training, Speexx [Speexx GmbH], Simulation Game, Studysmarter [StudySmarter GmbH], and other logic games) were used individually (<5% each; Multimedia Appendix 2).

Attitudes of Students Toward the Use of DGBL in Public Health Study Programs and Health Sciences

In response to a DGBL-oriented learning game scenario (Multimedia Appendix 3), students predominantly preferred desktop PCs (n=119, 36.5%), tablets (n=100, 30.7%), or smartphones (n=91, 27.9%) as end devices, while virtual reality or augmented reality glasses and game consoles were rarely chosen (n = 6, <2%). Most participants reported a willingness to use DGBL in the future (agree: n=140, 42.9%; tend to agree: n=94, 28.8%), whereas 9.5% (n=31) disagreed and 18.7% (n=61) did not provide information.

Perceived Potentials of DGBL in Public Health Study Programs and Health Sciences

Regarding the perceived usefulness of DGBL across modules (Multimedia Appendix 4), the highest levels of agreement were found for Prevention and Health Promotion (n=277, 85%), Epidemiology and Statistics (n=271, 83%), Population Medicine (n=251, 77%), and Environment and Health (n=271, 83%). Moderate agreement was observed for demography, health system, policy, or sociology, and health economics or management, whereas nursing sciences (n=104, 32%) and health services research (n=124, 38%) received lower levels of agreement. Open-ended responses highlighted additional applications in eHealth, communication strategy, scientific writing, and other specialized modules, with 1 participant emphasizing potential usefulness "in all modules of all subjects"

XSL•FO
RenderX

(R12). Refer to Multimedia Appendix 5 for participants' personal assessments of the use of DGBL (N=326).

Participants rated practical DGBL applications highly, especially exam preparation (n=279, 85.6%), knowledge checks (n=282, 86.5%), and acquiring new knowledge (n=273, 83.8%). Moderate support existed for methodological training (n=252, 77.3%), motivation to study (n=242, 74.2%), visualizing practical applications (n=242, 74.2%), error analysis, personalized learning profiles, self-determined action, and co-design (which received majority agreement but lower percentages; Multimedia Appendix 5).

Open-ended comments emphasized that DGBL should be used to supplement, rather than replace, face-to-face teaching. In addition, it should support self-study and be piloted in suitable modules to allow the accumulation of practical experience. Participants also emphasized that DGBL should be applied where it is didactically meaningful, supplementing rather than replacing face-to-face teaching and supporting low-threshold self-study, motivation, and structured learning. Some students suggested broader integration of serious games, while others noted that more experience is needed to assess DGBL's potential, recommending piloting in suitable modules (R119, R146, R214, R250, and R292). The open-ended field also revealed that participants considered the use of DGBL to be useful in other modules, including e–Public Health or eHealth (n=3, 0.9%); communication strategy (n=2, 0.6%); and scientific writing (small paragraphs), strategic research, social research, nursing education, psychology, statistics, biomedicine, and counseling and interviewing (each n=1, 0.3%). One person also emphasized that DGBL would be useful "in all modules of all subjects" (R12).

In an open-ended question, the respondents were able to emphasize their key comments on potential uses of DGBL in public health teaching. In terms of possible applications, DGBL should be used specifically where it makes didactic sense:

> *DGBL should only be used where it makes sense.* [R119]

However, according to some participants, DGBL applications should not replace face-to-face teaching, but merely supplement it, especially for the voluntary consolidation of content or learning control (R264). In addition, DGBL applications can motivate and structure self-study phases in a low-threshold manner (R146, R214, and R317). Some students emphasized the desire for a fundamental use of DGBL in their studies:

> *I would like to learn more through serious games.* [R129 and R250]

Others highlight that the idea of DGBL is still too abstract and that more experience with DGBL needs to be gained to enable a well-founded assessment of possible applications (R184 and R292). This could be remedied by piloting different DGBL applications in suitable modules (R292).

Promoting and Inhibiting Conditions for the Use of DGBL in Public Health Study Programs and Health Sciences

Students highlighted procedural and structural facilitators (Multimedia Appendix 6). Most agreed on the importance of data protection (n=303, 93%), technical infrastructure (n=321, 98.5%), and teaching staff skills (n=311, 95.4%). Adequate time allocation (n=310, 95.1%), central IT support (n=263, 80.7%), accessibility (n=303, 92.9%), end devices (n=247, 75.8%), and tutors (n=235, 72.1%) were also relevant. A majority agreed that prior skills may be necessary (n=183, 56.1%).

Additional conditions focused on usability and learning experience (Multimedia Appendix 7). Key requirements were ease of use (n=318, 97.5%), clarity (n=314, 96.3%), feedback (n=315, 96.6%), the ability to play anytime and anywhere (n=307, 94.2%), and technical reliability (n=299, 91.7%). A playful experience without time pressure (n=270, 82.8%), multimedia elements (n=235, 72.1%), reward systems (n=230, 70.5%), excitement (n=223, 68.4%), and realistic storylines (n=190, 58.3%) were also valued. In contrast, 3D graphics, immersion, avatar use, and multiplayer modes were less relevant. Nonuse was mainly reported because of concerns about additional sedentary PC activities (Multimedia Appendix 8).

Social influence (Multimedia Appendix 9) came primarily from fellow students (n=248, 76.2%) and lecturers (n=249, 76.6%), with lesser influence from university information sources, friends, partners, family, social media, and the press.

During gameplay, motivating factors included appropriate challenges (n=314, 96.3%), reflection on learning (n=297, 91.1%), feedback (n=300, 92%), individual performance measurement (n=242, 90.8%), and personal addressing of content (n=259, 79.3%). Social interaction (n=186, 57.1%), immersion in the history (n=197, 60.5%), realism of the story (n=180, 55.3%), team play, and competition were less influential (Multimedia Appendix 10).

In a final open-ended question, the respondents were able to emphasize their key comments on promoting and inhibiting conditions for the use of DGBL in public health teaching. The content and visual design of DGBL should be visually appealing, authentic, and serious without appearing childish (R87 and R184). In addition, clear feedback loops should be integrated so that players immediately recognize whether their answers are right or wrong and have a sense of participation and success (R164). From 1 participant's own experience, the consideration of social aspects in video games was considered relevant but not decisive for learning success:

> *Many of the video games that helped me on a personal level had no social aspect.* [R162]

Regarding access to DGBL, "soft digitalization" is preferred, which refers to gradual infrastructural adjustments rather than abrupt implementation:

> *Instead of hard digitalization = soft digitalization!* [R111]

XSL•FO
RenderX

Discussion

Principal Results

Most students reported a positive attitude toward DGBL in public health and health sciences, with strong support for future implementation. DGBL was perceived as particularly useful for Prevention and Health Promotion, Epidemiology and Statistics, and for exam preparation and knowledge monitoring. Desktop PCs, tablets, and smartphones were all considered suitable devices.

Successful implementation was seen as contingent on reliable technical infrastructure, media-competent teaching staff, data protection, and high usability. Students emphasized flexibility, clarity, and a pressure-free learning environment. Although multimedia elements were generally welcomed, immersive features such as 3D graphics or multiplayer formats were considered less relevant.

The technology affinity profile revealed moderate competence and trust but low enthusiasm and high technology-related fear, indicating cautious engagement with DGBL. Accordingly, students favored structured, predictable, and functionally transparent designs. Peer and instructor support were identified as key factors for acceptance.

Overall, students expressed openness to DGBL when embedded in clear, supportive, and goal-oriented learning contexts that emphasize personal progress, feedback, and emotional safety over competitive or immersive mechanisms.

Comparison With Prior Work

Previous literature has broadly demonstrated the educational potential of serious games in health care training. Chang et al [23] have highlighted that DGBL can significantly improve learning performance, self-efficacy, and satisfaction among nursing students. Similarly, Peng [22] showed that educational games in health communication settings could effectively promote healthier dietary behaviors. However, these examples predominantly focus on applied medical settings or individual behavior change and often target medical or nursing students rather than the broader interdisciplinary scope of public health education. Systematic reviews [24-27] support the claim that serious games can enhance learning enjoyment, knowledge retention, and skill acquisition across multiple health disciplines. Nevertheless, the scope of these reviews primarily includes physiotherapy, psychology, or general medical education. The present study addresses this oversight by concentrating on public health students, a population largely neglected in the literature. Moreover, although prior studies [19-21] have examined the use of serious games in various academic contexts, few have focused on students' attitudinal and contextual prerequisites for adopting DGBL. Our study contributes by empirically documenting both the facilitators and barriers to DGBL integration, including infrastructural needs, pedagogical considerations, and motivational dynamics, such as feedback and challenge. In contrast to previous works, we offer a user-centered view grounded in the actual preferences, reservations, and expectations of students enrolled in public health–related programs. Rather than evaluating a specific DGBL application, as seen in most intervention-based designs [18], our research used a scenario-based survey to explore students' anticipated acceptance and perceived utility of DGBL across a variety of modules and use cases. This broader approach allows a more generalizable understanding of DGBL's potential role in public health curricula. Finally, although the *Hochschulforum Digitalisierung* [3] and other German initiatives have promoted digital innovation in higher education, their recommendations often remain theoretical or focus on conventional digital tools such as e-learning platforms and lecture recordings. In contrast, our study assesses students' receptiveness to more immersive, interactive forms of digital education such as serious games and identifies practical implementation factors such as device preferences, flexibility in timing, and game design expectations. In sum, we complement and extend prior research by focusing on a previously underexplored academic population—public health students—and by detailing both their general attitudes toward DGBL and the specific structural, pedagogical, and social factors shaping its potential uptake in university-level health education.

Limitations

One limitation of this study is the occasional technical issue affecting the display of the DGBL scenario within the survey. At least 2 participants reported that the example scenario was not rendered correctly when using Safari on a MacBook. This may have affected their understanding of the scenario and increased the abstraction level of the subsequent questions, potentially influencing their responses. Future studies should ensure compatibility across different browsers and devices to improve data quality and user experience. In addition, usability testing should be conducted prior to data collection to identify and minimize such issues.

Another limitation concerns the sample composition and recruitment. Although students from multiple universities and health-related programs were included, recruitment via social media and voluntary participation may have led to self-selection bias, as students with greater interest in digital learning were likely more inclined to participate. Due to the anonymous online format, it was theoretically possible for students to complete the questionnaire multiple times. However, as invitations were primarily sent to new students each semester, the likelihood of multiple participation was minimized. Because the total number of students reached is unknown, a traditional response rate could not be calculated. Consequently, the sample represents only a subset of the student population in Germany and cannot be generalized to all public health and health sciences students.

The extended data collection period from 2021 to 2024 also presents a limitation, as it spans different phases of the COVID-19 pandemic and the postpandemic period. Students' experiences and familiarity with online learning may have changed during this time, potentially influencing attitudes toward DGBL. Temporal changes were not analyzed in this exploratory study. Future research could examine cohort or time effects to account for such variation.

Finally, the analysis was limited to descriptive statistics and qualitative content analysis. Although this approach allowed an exploratory understanding of student attitudes and conditions

XSL•FO
RenderX

for DGBL use, causal conclusions or inferential generalizations are not possible. Future studies could apply mixed methods approaches or longitudinal study designs to assess effectiveness, behavioral outcomes, and changes in attitudes over time.

Conclusions

Our study addresses an underexplored area by examining DGBL in public health and health sciences education in Germany. The findings indicate an overall positive attitude among students toward DGBL, particularly for practice-oriented applications such as exam preparation, knowledge testing, and the visualization of complex content. Students identified Prevention and Health Promotion, Epidemiology and Statistics, and Population Medicine as especially suitable domains. Effective implementation of DGBL requires adequate technical infrastructure, media-competent teaching staff, and a strong focus on usability, data protection, and pedagogical integration. Given students' moderate technological competence but high technology-related fear and low enthusiasm, DGBL designs should prioritize clarity, reliability, and psychological accessibility over immersive or complex features. Academic peers and instructors play a key role in shaping acceptance. Future research should assess longitudinal learning effects and implementation feasibility across institutional contexts to support the sustainable integration of DGBL into public health and health sciences curricula.

Acknowledgments

The authors express their gratitude to the seminar participants who contributed to the development of the study and to all study participants for their valuable contributions.

Funding

This study was funded by Bielefeld University. The authors bear sole responsibility for the content of this publication.

Authors' Contributions

KJW, JA, and AK drafted the manuscript. KJW and CD were responsible for the research project and study planning. TM, JA, and KJW developed the methodology. JA conducted the analysis, and JA and KJW jointly interpreted the results. JA and KJW finalized the Discussion and Conclusions sections. All authors discussed and critically revised the manuscript for important intellectual content and approved the final version.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Use of digital media (N=326).
[PNG File, 99 KB - mededu_v12i1e83347_app1.png]

Multimedia Appendix 2
Reasons for playing computer or video games (N=326).
[PNG File, 98 KB - mededu_v12i1e83347_app2.png]

Multimedia Appendix 3
Example scenario.
[PNG File, 868 KB - mededu_v12i1e83347_app3.png]

Multimedia Appendix 4
Assessment of the usefulness of digital game–based learning in public health–related modules (N=326).
[PNG File, 132 KB - mededu_v12i1e83347_app4.png]

Multimedia Appendix 5
Personal assessment of the use of digital game–based learning (N=326).
[PNG File, 115 KB - mededu_v12i1e83347_app5.png]

Multimedia Appendix 6
Assessment of the framework conditions for the use of digital game–based learning at the university (N=326).
[PNG File, 118 KB - mededu_v12i1e83347_app6.png]

XSL•FO
RenderX

Multimedia Appendix 7
Assessment of the framework conditions for the use of digital game–based learning in health-related programs (N=326).
[PNG File, 204 KB - mededu_v12i1e83347_app7.png]

Multimedia Appendix 8
Personal assessment of the nonuse of digital game–based learning (N=326).
[PNG File, 133 KB - mededu_v12i1e83347_app8.png]

Multimedia Appendix 9
Assessment of social influence on the use of digital game–based learning at the university (N=326).
[PNG File, 119 KB - mededu_v12i1e83347_app9.png]

Multimedia Appendix 10
Assessment of motivation during the game (N=326).
[PNG File, 173 KB - mededu_v12i1e83347_app10.png]

References

1. Gerlinger T, Schmucker R. 20 jahre public health – 20 jahre politik für eine gesunde gesellschaft? In: Schott T, Hornberg C, editors. Die Gesellschaft und ihre Gesundheit [Book in German]: VS Verlag für Sozialwissenschaften; 2011:69-83. [doi: 10.1007/978-3-531-92790-9_3]
2. Blättner B, Dierks ML. Lehre in public health: entwicklung und herausforderungen in Deutschland. In: Schmidt-Semisch H, Schorb F, editors. Public Health [Book in German]: Springer VS; 2021:179-193. [doi: 10.1007/978-3-658-30377-8_11]
3. Wannemacher K, Jungermann I, Scholz J, Tercanli H, von Villiez A. Digitale lernszenarien im hochschulbereich [Report in German]. : Hochschulforum Digitalisierung; 2016 URL: https://hochschulforumdigitalisierung.de/wp-content/uploads/2023/09/HFD-AP-Nr-15_Digitale-Lernszenarien.pdf [accessed 2026-08-20]
4. Kauffeld S, Othmer J. Handbuch Innovative Lehre [Book in German]: Springer Fachmedien Wiesbaden GmbH; 2019. [doi: 10.1007/978-3-658-22797-5]
5. Strahringer S, Leyh C, editors. Gamification und Serious Games: Grundlagen, Vorgehen und Anwendungen [Book in German]: Springer Vieweg Wiesbaden; 2017. [doi: 10.1007/978-3-658-16742-4]
6. Arnold P, Kilian L, Thillosen A, Zimmer G. Handbuch E-Learning: Lehren und Lernen mit Digitalen Medien [Book in German]: W. Bertelsmann Verlag; 2013.
7. Schubert T, Gerth M, editors. Multimediales Lehren und Lernen an der Martin-Luther-Universität Halle-Wittenberg: Befunde und Ansätze aus dem Forschungsförderprogramm des Zentrums für Multimediales Lehren und Lernen (LLZ) [Book in German], 1st edition: Martin-Luther-Universität Halle-Wittenberg; 2021. [doi: 10.25673/38465]
8. Gerhardus A, Munko T, Kolip P, Schilling I, Schlingmann K, editors. Lehren und Lernen in den Gesundheitswissenschaften [Book in German]: Hogrefe Verlag; 2019. [doi: 10.1024/85930-000]
9. Hurrelmann K, Razum O, editors. Handbuch Gesundheitswissenschaften [Book in German]: Beltz Juventa; 2016.
10. The 8th Global Conference on Health Promotion, Helsinki, 2013. World Health Organization. 2013. URL: https://www.who.int/teams/health-promotion/enhanced-wellbeing/eighth-global-conference [accessed 2026-09-03]
11. Gerlinger T, Babitsch B, Blättner B, et al. Situation and perspectives of public health in Germany--research and teaching. Position paper of the German Public Health Association [Article in German]. Gesundheitswesen 2012;74(11):762-766. [doi: 10.1055/s-0032-1330011] [Medline: 23175125]
12. Razum O, Dockweiler C. Generalists or specialists? Interdisciplinarity in, and scientific development of, public health in Germany [Article in German]. Gesundheitswesen 2015;77(11):831-833. [doi: 10.1055/s-0034-1395629] [Medline: 25643291]
13. Eibl-Eibesfeldt I. Human Ethology: Routledge; 2017.
14. Prensky M. Digital Game-Based Learning: Paragon House; 2007.
15. Wiemeyer J. Serious Games für die Gesundheit [Book in German]: Springer Wiesbaden; 2016. [doi: 10.1007/978-3-658-15472-1]
16. Klopfer E, Osterweil S, Salen K. Moving learning games forward. The Education Arcade Massachusetts Institute of Technology. 2009. URL: https://education.mit.edu/wp-content/uploads/2018/10/MovingLearningGamesForward_EdArcade.pdf [accessed 2026-08-20]
17. Tolks D, Lampert C. Abgrenzung von serious games zu anderen lehr- und lernkonzepten. In: Dadaczynski K, Schiemann S, Paulus P, editors. Gesundheit Spielend Fördern: Potenziale und Herausforderungen von Digitalen Spieleanwendungen für die Gesundheitsförderung und Prävention [Book in German]: Beltz Juventa; 2016:191-217.
18. Olszewski AE, Wolbrink TA. Serious gaming in medical education: a proposed structured framework for game development. Simul Healthc 2017 Aug;12(4):240-253. [doi: 10.1097/SIH.0000000000000212] [Medline: 28027076]

XSL•FO
RenderX

19. Eckardt L, Körber S, Becht EJ, Plath A, Al Falah S, Robra-Bissantz S. Führen serious games zu lernerfolg? – Ein vergleich zum frontalunterricht. In: Strahringer S, Leyh C, editors. Gamification und Serious Games [Book in German]: Springer Vieweg; 2017:139-150. [doi: 10.1007/978-3-658-16742-4_11]
20. Fischer H, Heinz M, Schlenker L, Münster S, Follert F, Köhler T. Die gamifizierung der hochschullehre – potenziale und herausforderungen. In: Strahringer S, Leyh C, editors. Gamification und Serious Games [Book in German]: Springer Vieweg; 2017:113-125. [doi: 10.1007/978-3-658-16742-4_9]
21. Robra-Bissantz S, Bott OJ, Kleinefeld N, Neu K. In: Zickwolf K, editor. Teaching Trends 2018. Die Präsenzhochschule und die Digitale Transformation [Book in German]: Waxmann; 2019. [doi: 10.25656/01:17455]
22. Peng W. Design and evaluation of a computer game to promote a healthy diet for young adults. Health Commun 2009 Mar;24(2):115-127. [doi: 10.1080/10410230802676490] [Medline: 19280455]
23. Chang CY, Chung MH, Yang JC. Facilitating nursing students' skill training in distance education via online game-based learning with the watch-summarize-question approach during the COVID-19 pandemic: a quasi-experimental study. Nurse Educ Today 2022 Feb;109:105256. [doi: 10.1016/j.nedt.2021.105256] [Medline: 34968932]
24. Tori AA, Tori R, Nunes FD. Serious game design in health education: a systematic review. IEEE Trans Learn Technol 2022;15(6):827-846. [doi: 10.1109/TLT.2022.3200583]
25. Fontaine G, Cossette S, Maheu-Cadotte MA, et al. Efficacy of adaptive e-learning for health professionals and students: a systematic review and meta-analysis. BMJ Open 2019 Aug 28;9(8):e025252. [doi: 10.1136/bmjopen-2018-025252] [Medline: 31467045]
26. Gentry SV, Gauthier A, L'Estrade Ehrstrom B, et al. Serious gaming and gamification education in health professions: systematic review. J Med Internet Res 2019 Mar 28;21(3):e12994. [doi: 10.2196/12994] [Medline: 30920375]
27. Blakely G, Skirton H, Cooper S, Allum P, Nelmes P. Educational gaming in the health sciences: systematic review. J Adv Nurs 2009 Feb;65(2):259-269. [doi: 10.1111/j.1365-2648.2008.04843.x] [Medline: 19032512]
28. Karrer K, Glaser C, Clemens C, Bruder C. Technikaffinität erfassen – der fragebogen TA-EG. In: Lichtenstein A, Clemens C, editors. Der Mensch Im Mittelpunkt Technischer Systeme [Book in German]: VDI Verlag; 2009:196.
29. Rädiker S, Kuckartz U. Analyse Qualitativer Daten mit MAXQDA [Book in German]: Springer VS Wiesbaden; 2018. [doi: 10.1007/978-3-658-22095-2]

Abbreviations

ASPHER: Association of Schools of Public Health in the European Region
DGBL: digital game–based learning
TA-EG: *Technikaffinität–Elektronische Geräte*
WHO: World Health Organization

Edited by T Gladman; submitted 01.Sep.2025; peer-reviewed by D Dinh, S Guinez-Molinos; revised version received 21.Jul.2026; accepted 23.Jul.2026; published 10.Sep.2026.

Please cite as:
Wrona KJ, Albrecht J, Krümmel A, Aschentrup L, Dockweiler C, Mc Call T
Attitudes Toward and Preconditions for Digital Game–Based Learning Among Public Health and Health Sciences Students: Cross-Sectional Online Survey Study
JMIR Med Educ 2026;12:e83347
URL: https://mededu.jmir.org/2026/1/e83347
doi:10.2196/83347

© Kamil J Wrona, Joanna Albrecht, Anne Krümmel, Leona Aschentrup, Christoph Dockweiler, Timothy Mc Call. Originally published in JMIR Medical Education (https://mededu.jmir.org), 10.Sep.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Education, is properly cited. The complete bibliographic information, a link to the original publication on https://mededu.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Impact of a Structured Training Program on Medical Student Confidence and Behavior During Their First Radial Arterial Puncture: Comparative Study

Camille Rolland-Debord[1,2], MD, PhD; Lucien Juret[1], MD; Mathilde Simon[1], MD; Abdallah El Mouhajer[1], MD; Cécile Londner[3], MD; Capucine Morélot-Panzini[2,4], MD, PhD; Cécile Chenivesse[5], MD, PhD; Thomas Similowski[4,6], MD, PhD

[1]Service de Pneumologie, CHU Clermont-Ferrand, Université Clermont Auvergne, 58 Rue Montalembert, Clermont-Ferrand, France

[2]Service de Pneumologie, Département R3S, F-75013, Groupe Hospitalier Universitaire, Assistance Publique–Hôpitaux de Paris, Sorbonne Université, Hôpital Pitié-Salpêtrière, Paris, France

[3]Service des Pathologies du Sommeil, Hôpital Pitié-Salpêtrière, Département R3S, F-75013, Groupe Hospitalier Universitaire, Assistance Publique–Hôpitaux, Sorbonne Université, Paris, France

[4]Neurophysiologie Respiratoire Expérimentale et Clinique, F-75005, INSERM UMRS1158, Sorbonne Université, Paris, France

[5]Service de Pneumologie et Immuno-Allergologie, Centre Hospitalier Universitaire de Lille, Lille, France

[6]Département R3S, F-75013,Hôpital Pitié-Salpêtrière, Groupe Hospitalier Universitaire, Assistance Publique–Hôpitaux, Sorbonne Université, Paris, France

Corresponding Author:
Camille Rolland-Debord, MD, PhD
Service de Pneumologie, CHU Clermont-Ferrand, Université Clermont Auvergne, 58 Rue Montalembert, Clermont-Ferrand, France

Abstract

Background: Radial artery puncture is a common clinical procedure essential for assessing gas exchange but is frequently perceived as stressful by inexperienced operators, who fear causing pain to their patients. Despite its practical relevance, formal training in this procedure is inconsistently integrated into medical curricula. This study evaluated whether a structured training program—combining theoretical instruction, simulation-based practice, and debriefing—could influence students' procedural confidence and decision-making and patient experience during their first clinical arterial puncture.

Objective: This study aimed to determine whether structured simulation-based training influences medical students' anxiety, confidence, and technical performance and patient experience during their first arterial puncture.

Methods: Third-year medical students who had never performed an arterial puncture were assigned to 1 of 2 groups: a structured training group (group 1) or a control group receiving informal or no specific training (group 2). After performing their first arterial puncture under supervision, students completed a questionnaire assessing apprehension, satisfaction, and confidence. The decision to use local anesthesia, puncture success, and patient-rated pain and apprehension were also recorded. A total of 67 students participated (group 1: n=24, 35.8%; group 2: n=43, 64.2%), with 61 patients included. Statistical comparisons were performed using the Fisher exact and nonparametric Mann-Whitney *U* tests (α=.05).

Results: Self-reported apprehension and confidence were similar between groups. However, group 1 students were significantly less likely to use local anesthesia compared to group 2 students (7/20, 35% vs 28/36, 77.8%, respectively; *P*=.003), suggesting greater procedural confidence. First-attempt success rates were comparable (group 1: 3/13, 23.1%; group 2: 14/29, 48.3%; *P*=.18). Median patient-reported pain scores were numerically but not statistically significantly lower when anesthesia was used (2.1, IQR 1.2 - 4.0 vs 4.8, IQR 2.1 - 6.4; *P*=.08).

Conclusions: Structured training influenced students' behavior during their first arterial puncture, reducing reliance on anesthesia despite similar levels of self-reported apprehension. Although confidence ratings did not differ, behavioral indicators suggested improved self-efficacy and readiness for clinical performance. These findings support the behavioral impact of structured procedural education and call for future research using validated assessment tools and long-term follow-up.

(JMIR Med Educ 2026;12:e78086) doi:10.2196/78086

KEYWORDS

arterial puncture; medical education; simulation-based training; procedural confidence; medical students; local anesthesia; clinical skills; competency-based learning

XSL•FO
RenderX

Introduction

Performing a radial arterial puncture to assess gas exchange can be challenging for medical and nursing students due to the risk of causing significant pain at a small puncture site densely packed with nociceptive C-fiber–containing structures, including the artery itself and nearby tissues such as the periosteum [1,2]. Several studies have explored strategies to improve patient tolerance. Topical anesthetics have shown limited effectiveness [3,4], and cryoanalgesia, though promising, is seldom used [5]. Local infiltration anesthesia offers the greatest pain reduction but is considered impractical and is infrequently applied [6-8]. Moreover, reducing needle gauge does not alleviate pain and may prolong the procedure [9]. Patout et al [9] demonstrated that preprocedural anxiety in patients significantly correlates with increased pain perception during radial arterial puncture. Importantly, patient anxiety is likely to be heightened by the anxiety of the operator, whose fear of causing discomfort can lead to anticipatory stress. This, in turn, may further complicate the procedure, creating a vicious circle that amplifies the risk of pain.

Structured training programs have been shown to reduce anxiety and build confidence in nursing [10,11] and medical students [12-14] performing clinical procedures. In the particular case of radial arterial puncture, Collins et al [13] implemented a 5-step curriculum for respiratory care students combining theory, simulation, and peer assessment, which resulted in high interrater performance scores. Similarly, simulation-based training has been shown to enhance confidence and technical skills in core procedures among medical students, including those involving invasive techniques [15,16].

In this context, we conducted a national survey of arterial puncture teaching practices in French pneumology departments and found that only a minority used structured programs to train medical students (Multimedia Appendix 1). At our own institution, a systematic observation of unsupervised arterial puncture attempts by third-year medical students on a simulator revealed a high frequency of procedural errors (Multimedia Appendix 1).

Building on these observations and the available literature, we designed a structured training program combining theoretical instruction, simulation-based practice, and supervised hands-on experience. The aim of this study was to evaluate whether adding a structured theoretical and simulation-based training session influenced medical students' procedural confidence, preprocedural anxiety, and first-attempt performance during radial arterial puncture compared with usual heterogeneous bedside teaching. We hypothesized that structured preparation would be associated with increased confidence and reduced anxiety, as well as a possible improvement in performance.

Methods

Ethical Considerations

This study was approved by the institutional review board of the National Institute of Health and Medical Research (reference IRB00003888). All data were collected anonymously, and no personal identifiers were recorded at any stage. Informed consent was obtained from all participants prior to study participation. Participation was voluntary, and students could decline or withdraw at any time without any academic or clinical consequences. No financial or other compensation was provided for participation.

Recruitment Procedure

This study was conducted in hospitals affiliated with the Faculty of Health, Sorbonne University, one of the medical schools in the Paris area.

All third-year medical students (General Medical Sciences Diploma) completing a 3-month clinical rotation in pulmonology were eligible for inclusion if they performed their first arterial puncture on a patient and had not previously received simulation-based training for this procedure. Each eligible student was invited to complete the online questionnaire within 24 hours of performing the procedure (Multimedia Appendix 1). At the time of rotation assignment, students selecting a pulmonology placement were verbally informed that they would be invited to complete a short online questionnaire after performing their first arterial puncture. This information was reiterated at the beginning and end of the rotation and again during the end-of-semester pulmonology examination. At the start of each rotation, senior physicians in the participating departments were informed of the study via email and provided with a link to the questionnaire for distribution to their students. Data were collected using an online questionnaire specifically developed for this study and hosted on the SurveyMonkey platform (SurveyMonkey Inc; Multimedia Appendix 1). The initial draft was developed by the study investigators and reviewed for face and content validity by a panel of 5 academic pulmonologists and medical educators from Sorbonne University. The instrument was pilot-tested among medical students before implementation.

Student Groups

Overview

Students were assigned to groups based solely on their clinical placement independently of the study investigators. Both groups followed the usual teaching of their respective departments. These practices were heterogeneous and could involve bedside instruction by a nurse, a more senior medical student, a resident, or an attending physician. In some cases, students were verbally instructed before entering the patient's room or performed the procedure after having observed it on another patient. No intervention was made to modify these existing teaching practices. Regarding the clinical procedure itself, the use of topical anesthesia was left up to the students in accordance with local departmental policies. The students were divided as follows:

- Group 1—students rotating in the respiratory medicine department of Pitié-Salpêtrière Hospital, Sorbonne University, who received the usual teaching plus an additional structured training program before their first arterial puncture

XSL•FO
RenderX

- Group 2—students rotating in other respiratory medicine departments of Sorbonne University who received usual teaching only

For sensitivity analysis, group 2 was then subdivided into students who had received bedside demonstration or guidance before their first arterial puncture (group 2+) and those who had not (group 2−). This subdivision allowed us to capture the full spectrum of existing informal learning situations within the "usual practice" group.

Structured Training Program (Group 1)

Overview

Students in group 1 participated in a structured theoretical and practical training session that had been routinely implemented in the department in previous years. The session was delivered by a single instructor and organized into 2 groups of approximately 7 students each. The total duration varied slightly between sessions but was at least 1 hour. The session began with a structured discussion of the indications and techniques for arterial puncture, followed by a briefing outlining the simulation objectives, procedural steps to be reproduced, and key safety principles.

Theoretical Component

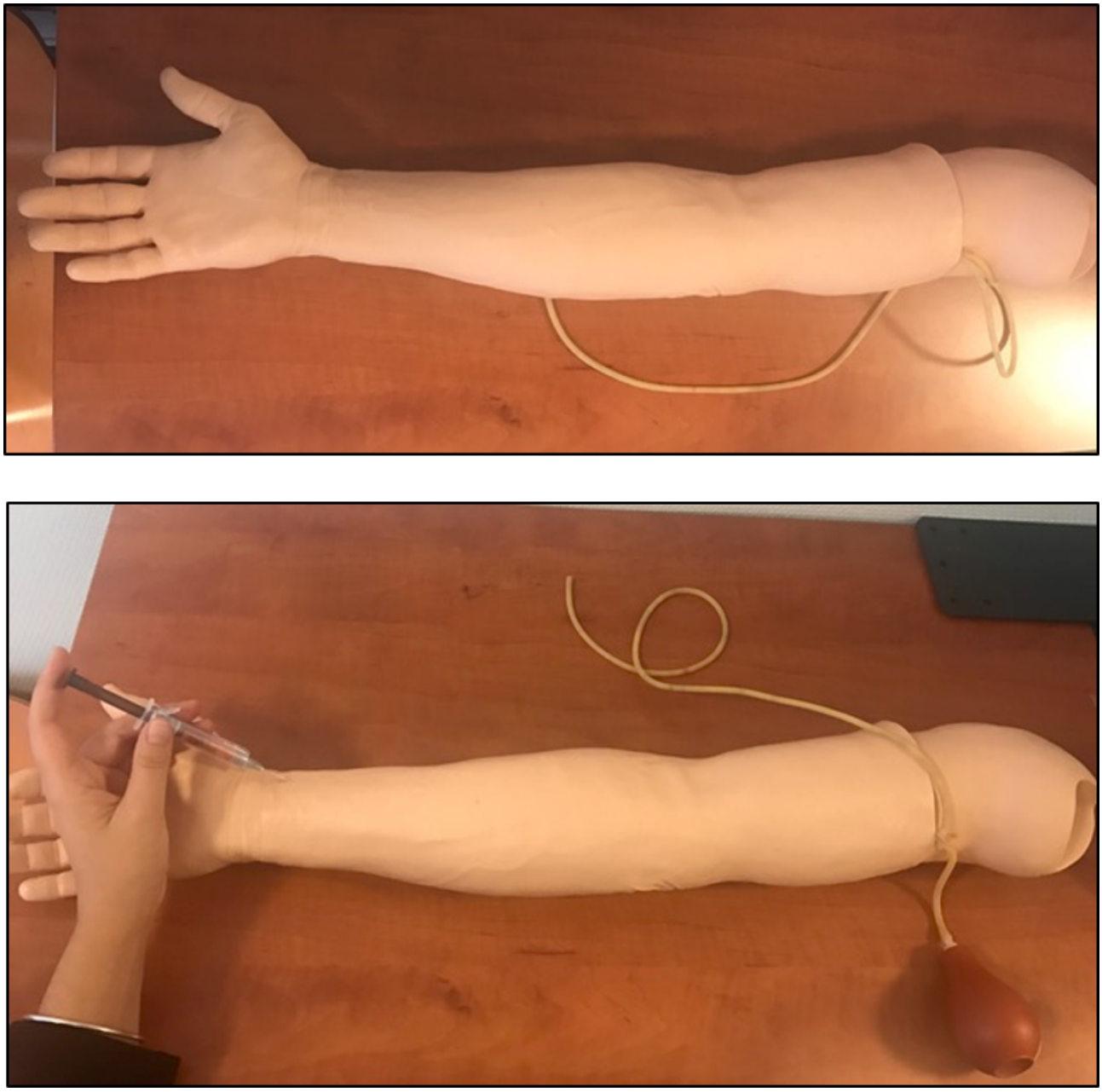
A dedicated slide deck (Multimedia Appendix 1) was developed based on the *New England Journal of Medicine* resource on radial artery puncture [17] incorporating annotated screenshots from the procedural video. The content was reviewed for accuracy and pedagogical coherence by 5 academic pulmonologists. Each student received a printed copy of the material at the end of the session and was granted permanent online access.

Practical Component

The practical portion of the session consisted of individual hands-on training using an artificial arm simulator (Laerdal Medical; Figure 1), which allowed for palpation of a simulated pulse and visualization of arterial blood return. Each student repeated the procedure until they achieved self-reported confidence.

XSL•FO
RenderX

Figure 1. Procedural arm simulator with a pseudopulsatile artery and circulating red liquid (Laerdal Medical).



Procedural steps were practiced systematically, including patient communication, site preparation, needle positioning and insertion, troubleshooting in case of failure, and needle disposal. Students were encouraged to observe and discuss each other's attempts to reinforce learning through peer observation.

Peer-to-peer exercises were also incorporated, including palpation of the radial artery groove, performance of the Allen test, patient communication, and patient positioning and preparation.

A standardized communication script (Multimedia Appendix 1) written and reviewed by the same 5 academic pulmonologists was introduced and read aloud by the instructor and then practiced in pairs by students. Each participant received a printed copy for review before performing the procedure in a clinical setting. The script included a standard introduction, an explanation of the procedure, information regarding expected discomfort, and clarification about the absence of unambiguously validated analgesic strategies.

Structured Debriefing

Each session concluded with a structured debriefing conducted according to the Haute Autorité de Santé simulation guidelines. Students were first invited to express their emotions and initial impressions (reaction); then analyze their actions and underlying reasoning (analysis); and, finally, summarize key learning points and objectives for improvement (synthesis). Incorrect decisions or technical errors were explored to identify their causes and were addressed through guided feedback, theoretical reinforcement, and targeted procedural adjustments. Throughout the session, the instructor fostered psychological safety; encouraged self-assessment; used open-ended questioning; and provided constructive, nonjudgmental feedback.

XSL•FO
RenderX

At the conclusion of the session, students were reminded that they would later be asked to complete an online questionnaire following their first arterial puncture on a patient.

Usual Teaching Only (Group 2)

Students in group 2 received the usual teaching routinely provided in the departments to which they had been assigned. Teaching practices varied and could include bedside instruction and/or verbal explanations given before entering the patient's room. Instruction could be provided by a nurse, a senior medical student, a resident, or an attending physician. In some cases, students performed the procedure after first observing it on another patient. No intervention was introduced to modify these usual teaching practices.

Evaluation Criteria

Evaluation criteria included several dimensions of student and patient experience. For students, the following variables were assessed: global anticipatory apprehension, apprehension about causing pain, apprehension about procedural failure, satisfaction with the procedure, apprehension about repeating the procedure, and confidence in performing it again. The decision to use topical anesthesia was also recorded, as well as the success rate of the arterial puncture. For patients, perceptions of the procedure were evaluated through 2 measures: anticipatory apprehension and pain experienced during the puncture. Both students and patients rated their perceptions using a 10-cm visual analogue scale (VAS) ranging from 0 ("none at all") to 10 ("maximum imaginable").

Statistical Analysis

Statistical analyses were performed using Prism (GraphPad Software). Categorical variables were reported as counts and percentages, whereas continuous variables were presented as means and SDs or medians and IQRs depending on their distribution. Given the limited sample size and the ordinal nature of several variables, nonparametric statistics were used. The primary analysis compared group 1 and group 2. Group comparisons were conducted as follows: (1) categorical variables were compared using the Fisher exact test, and (2) continuous variables were compared using the Mann-Whitney *U* test. A sensitivity analysis was performed to compare group 1 with the group 2 subgroups: group 2+ (students who received prior training) and group 2− (students who did not receive prior training). All tests were 2 sided, with the type I error rate set at α=.05.

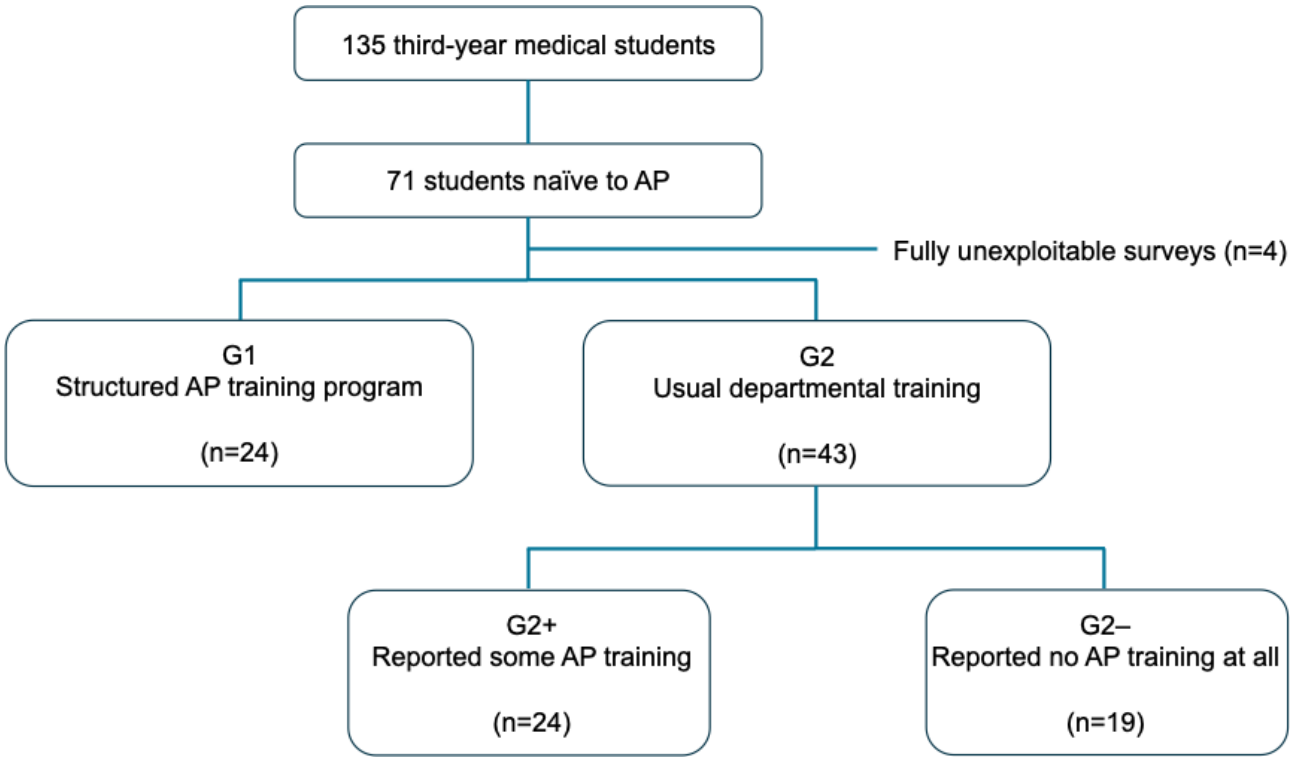
Results

Participants

During the study period, 135 third-year medical students were assigned to the participating departments. Of these 135 students, 71 (52.6%) were entirely naive to arterial puncture and performed their first procedure during their rotation (Figure 2). The cohort included 78.9% (56/71) women and 21.1% (15/71) men, with a median age of 21 (IQR 20 - 21) years; 90.1% (64/71) were right-handed. A total of 83.1% (59/71) had previously observed an arterial puncture performed by a nurse or physician. All participants began completing the survey immediately after the procedure, but 21.1% (15/71) submitted incomplete responses. These included 16.7% (4/24) of the students from group 1, a total of 16.3% (7/43) from group 2, and 5.6% (4/71) whose group affiliation could not be determined. These cases were excluded from analysis, yielding a final sample of 67 students. Group 1 included 35.8% (24/67) of the students. Among the 43 students in group 2, a total of 19 (44.2%) reported having received no prior training (group 2−), and 24 (55.8%) reported some form of training (group 2+), which typically consisted of theoretical explanations and observation of procedures performed by nurses, senior students, or physicians.

XSL•FO
RenderX

Figure 2. Flowchart of the study. AP: arterial puncture; G1: students assigned to the combined theoretical and practical structured AP training program; G2: students assigned to usual teaching; G2+: subgroup of G2 consisting of students who reported receiving some prior AP training; G2−: subgroup of G2 consisting of students who reported receiving no prior AP training at all.



Attendants

In total, 92.9% (52/56) of the students were accompanied during their first arterial puncture (19/19, 100% in the group 2− subgroup). Attendants were nurses in 52.2% (35/67) of cases, medical trainees in 31.3% (21/67) of cases (including fourth- to sixth-year medical students and residents more advanced in training than the study participants, who were all in their third year), and senior physicians in the remaining 16.4% (11/67) of cases.

Medical Students' Perceptions

No statistically significant differences were observed between group 1 and group 2 in terms of global anticipatory apprehension, apprehension about causing pain, apprehension about procedural failure, satisfaction with the procedure, apprehension about repeating the procedure, or confidence in repeating the procedure (Table 1). However, anticipatory apprehension about causing pain was significantly lower in group 1 compared to group 2− (median 4.0, IQR 2.0 - 7.0 vs median 6.4, IQR 4.2 - 8.5; *P*=.04; Table 2 and Figure 3). Conversely, postprocedure satisfaction was significantly higher in group 2− than in group 1 (median 8.2, IQR 2.3 - 10.0 vs median 3.6, IQR 1.8 - 7.2; *P*=.03; Table 2 and Figure 3).

Table . Students' perception of the procedure: group 1 vs group 2.[a]

	Group 1, median (IQR)	Group 2, median (IQR)	*P* value
Global anticipatory apprehension	4.1 (2.2-5.0)	4.0 (2.0-5.6)	.67
Apprehension about causing pain	5.0 (2.3-7.2)	5.0 (3.1-6.9)	.60
Apprehension about procedure failure	4.3 (2.3-6.2)	4.4 (2.0-6.9)	.84
Satisfaction with the procedure	3.6 (1.8-7.3)	5.6 (2.4-9.4)	.09
Apprehension about repeating the procedure	1.9 (0-4.6)	2.3 (0.3-3.9)	.87
Confidence in repeating the procedure	7.0 (5.1-7.8)	6.0 (4.4-8.0)	.80

[a]Group 1: students assigned to the combined theoretical and practical structured training program (data from 20/24, 83.3% of the students; incomplete survey forms excluded). Group 2: other students (data from 43/46, 93.5% of the students; incomplete survey forms excluded). Comparisons between the groups were performed using the Mann-Whitney *U* test.

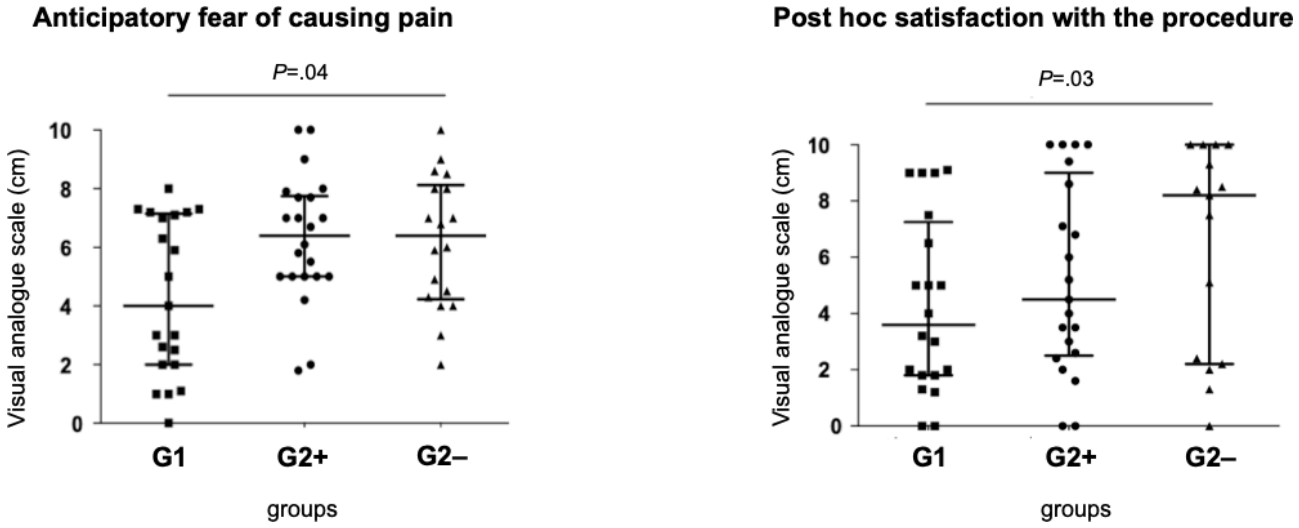
XSL•FO
RenderX

Table . Students' perception of the procedure: group 1 vs group 2–.[a]

	Group 1, median (IQR)	Group 2–, median (IQR)	*P* value
Global anticipatory apprehension	4.2 (1.0-6.1)	4.7 (1.7-6.2)	.44
Apprehension about causing pain	4.0 (2.0-7.0)	6.4 (4.2-8.5)	.04
Apprehension about procedure failure	5.0 (3.0-7.0)	5.2 (3.0-7.2)	.56
Satisfaction with the procedure	3.6 (1.8-7.2)	8.2 (2.3-10.0)	.03
Apprehension about repeating the procedure	1.8 (0-4.6)	2.5 (1.0-3.8)	.58
Confidence in repeating the procedure	7.0 (5.0-7.7)	6.0 (5.0-9.2)	.87

[a]Group 1: students assigned to the combined theoretical and practical structured training program (n=24). Group 2–: subgroup of group 2 consisting of students who reported receiving no prior training (n=19). Comparisons between the groups were performed using the Mann-Whitney *U* test.

Figure 3. Sensitivity analysis. G1: students assigned to the combined theoretical and practical structured arterial puncture training program; G2+: subgroup of group 2 consisting of students who reported receiving some prior arterial puncture training; G2−: subgroup of group 2 consisting of students who reported receiving no prior arterial puncture training at all.



Use of Topical Anesthesia

Overall, 62.5% (35/56) of the students elected to use anesthesia before performing the arterial puncture. Among them, 45.7% (16/35) used an EMLA 5% patch, 51.4% (18/35) used EMLA 5% cream, and 2.9% (1/35) used subcutaneous lidocaine. Anesthesia use was significantly less frequent in group 1 compared to group 2 (7/20, 35% vs 28/36, 77.8%; *P*=.003). The use of anesthesia did not lead to statistically significant differences in students' perceptions or patients' reported experiences, although procedural pain and self-reported confidence showed nonsignificant trends (Table 3).

XSL•FO
RenderX

Table . Patients' and students' experience based on the use of preprocedure anesthesia. Comparisons between groups were performed using the Mann-Whitney *U* test.

	Without prior anesthesia, median (IQR)	With prior anesthesia, median (IQR)	*P* value
Patients			
Anticipatory apprehension	4 (2-6.2)	3 (1.5-5.7)	.50
Procedural pain	4.8 (2.1-6.4)	2.1 (1.2 - 4)	.08
Students			
Global anticipatory apprehension	3.8 (1.8-5.8)	4.1 (2-5.1)	.89
Apprehension about causing pain	5 (2-7.3)	5 (3.4-6.9)	.57
Apprehension about procedure failure	4.3 (2-7.2)	4.4 (2-6.2)	.82
Satisfaction with the procedure	4 (2-9)	5.1 (2 - 8.8)	.61
Apprehension about repeating the procedure	1.9 (0-5)	2.4 (0-4.1)	.93
Confidence in repeating the procedure	7.1 (5.3-8.6)	5.4 (4.3-7.9)	.07

Arterial Puncture Success

The first-attempt success rate was 23.1% (3/13) in group 1 and 48.3% (14/29) in group 2 (*P*=.18). Within group 2, students with no prior training (group 2−) achieved a notably higher first-attempt success rate than those with some prior training (group 2+): 71.4% (10/14) vs 28.6% (4/14), respectively (*P*=.06). The overall success rate within the first 3 attempts was 65% (13/20) in group 1 and 80.6% (29/36) in group 2 (*P*=.20). The median number of attempts required for success was significantly higher in group 1 than in group 2 (median 1, IQR 0.5 - 2 vs median 1, IQR 0 - 1; *P*=.02). The number of procedures requiring supervisor intervention after 3 failed attempts was identical in both groups (7/20, 35% and 7/36, 19.4%).

Patient Description and Experiences

Arterial punctures were performed by students on 61 patients (median age 65, IQR 57 - 76 years; n=38, 62.3% women and n=23, 37.7% men). The number of patients (n=61) was lower than the number of participating students because some students performed their first arterial puncture on the same patient for logistical reasons. A total of 26.2% (n=16) of the patients were receiving oral analgesics, primarily paracetamol. Most patients (n=41, 67.2%) had previously undergone multiple radial arterial punctures.

Overall, anticipatory apprehension was rated at a median of 3.5 (IQR 2.0 - 6.0) on the VAS, with no statistically significant difference between patients who did not receive anesthesia and those who did (median 4.0, IQR 2.0 - 6.2 vs median 3.0, IQR 1.5 - 5.7; *P*=.50). Overall procedural pain was rated at a median of 2.7 (IQR 1.2 - 5.4). Pain scores were higher in the group without anesthesia than in the group with anesthesia (median 4.8, IQR 2.1 - 6.4 vs median 2.1, IQR 1.2 - 4.0), although the difference did not reach statistical significance (*P*=.08).

Discussion

Principal Findings

In a context where the training provided to French medical students for arterial puncture appears highly variable and generally brief—typically lasting less than 30 minutes—and where unsupervised students without specific training demonstrate frequent procedural errors (Multimedia Appendix 1), our results show that a structured training program can meaningfully influence behavior. Specifically, students who completed the structured program (group 1) were significantly less likely to choose local anesthesia before the procedure compared to those who received only informal or no training (group 2; 7/20, 35% vs 28/36, 77.8%; *P*=.003) despite having similar levels of self-reported anticipatory apprehension. This occurred even though the rate of successful punctures did not differ significantly between groups. From the patients' perspective, no statistically significant differences were found in anticipatory apprehension or procedural pain based on anesthesia use. Nonetheless, the difference in pain ratings exceeded the minimal clinically important difference on a VAS—typically estimated at 13 mm (1.3 cm) [18]—suggesting a clinically meaningful effect even in the absence of statistical significance. These findings are consistent with those of prior studies that report limited impact of topical anesthetics and highlight the role of operator anxiety in shaping patient experience [4,5,9].

Recent reviews in respiratory care education highlight that simulation—whether of low or high fidelity—enhances clinical reasoning, adherence to safety standards, and learner engagement [14]. To our knowledge, this study is among the first to examine the impact of structured training on medical students performing their first arterial puncture. While no direct parallels exist in this population, similar findings have been reported in other health care student groups. For example, Collins et al [13] implemented a 5-step curriculum for respiratory care students combining theory, simulation, and peer

XSL•FO
RenderX

assessment, which resulted in high interrater performance scores. In nursing education, structured programs using artificial arm simulators have improved procedural success rates and reduced errors [11], and a competency-based model allowed nurses to reach proficiency levels typically associated with physicians [19]. Given these precedents—and the findings from our preliminary observations (Multimedia Appendix 1)—our study focused not on procedural skill acquisition but on the impact of structured training on students' apprehension and confidence. In other domains, simulation-based training and structured debriefing have been shown to reduce anxiety and enhance self-confidence. For instance, Schmidt et al [20] demonstrated reduced anxiety in students presenting clinical cases after receiving video-based feedback, and Toy et al [21] reported improved confidence and competence following simulation-based training in core invasive procedures. Similar outcomes have been reported with academic detailing, an approach that typically refers to individualized educational outreach for practicing clinicians [22] that can influence clinical decision-making and has also been shown to improve confidence, competence, and critical thinking [23]. In our study, students who completed the structured training program (group 1) reported levels of apprehension and confidence similar to those of students with informal or no training (group 2) both before and after the procedure. While these results might initially appear less favorable, group 1 students were significantly less likely to use local anesthesia (7/20, 35% vs 28/36, 77.8%; *P*=.003), revealing a notable divergence between self-reported perception and actual behavior. This suggests that the training influenced confidence in ways not fully captured by subjective ratings. Choosing to perform a first arterial puncture without anesthesia may reflect a stronger sense of preparedness despite lingering concerns about causing pain. Supporting this interpretation, group 1 students showed significantly less apprehension about causing pain than students with no prior training (group 2−). Group 1 students also reported lower postprocedure satisfaction than group 2− students, which may reflect more realistic expectations shaped by the training and a narrower gap between anticipation and experience. The lower use of local anesthesia by group 1 students may also be due to the educational content of the program, which did mention the limited effectiveness of topical anesthetics for arterial puncture according to the literature. However, group 1 students were not advised against their use and were taught how to administer them. This leads us to believe that increased self-confidence did play a role in the observed difference, possibly in addition to or in synergy with adherence to the course content.

This study has several limitations. First, the structured training was not delivered by certified simulation instructors as recommended by the French National Authority for Health. This may have affected its quality—particularly the debriefing, widely regarded as the most critical element of simulation-based learning. Second, the educational interventions received by students in both groups were inherently heterogeneous. Teaching practices varied across hospital sites, ranging from bedside instruction by a nurse, senior medical student, resident, or attending physician to situations in which students performed the procedure with minimal prior observation. This design intentionally reflected the authentic diversity of clinical training conditions rather than an artificially standardized environment to assess the added value of a structured program under real-world conditions. Such heterogeneity may have introduced uncontrolled differences in baseline exposure. Future studies could address this by comparing structured training with a standardized "usual practice" curriculum to more precisely isolate the effects of the structured training. Third, we relied on nonvalidated VASs to assess students' apprehension and confidence. While these were practical and easy to administer, their psychometric properties are uncertain, limiting the reliability and generalizability of the findings. Future studies should consider using validated instruments for anxiety and procedural self-efficacy, as well as integrating frameworks such as entrustable professional activities to better capture learners' readiness for unsupervised clinical practice. Fourth, contextual and patient-related factors could not be controlled across hospital sites. Differences in patient characteristics or case complexity may have influenced the technical difficulty of arterial puncture and contributed to variations in first-attempt success rates. Additional limitations include the potential demotivating effect of a lengthy questionnaire, which may have contributed to incomplete responses and the absence of data on the time interval between training and the procedure. Although students had continuous access to theoretical materials, we cannot confirm that they reviewed them prior to performing the puncture, which may have attenuated the effects of the training.

Conclusions

Structured training for arterial puncture influences students' behavior in meaningful ways even when self-reported confidence and apprehension appear unchanged. The reduced reliance on local anesthesia among trained students may reflect greater readiness to take responsibility for a potentially painful procedure. These findings support the value of formalized training and highlight the need for validated tools and long-term evaluation to fully capture its educational and clinical impact. Although exploratory and limited by methodological constraints, this study provides initial evidence that such pedagogical frameworks could be meaningfully integrated into undergraduate medical curricula in pulmonology and related disciplines.

Funding

This study was conducted by the authors within the realm of their academic duties and received no dedicated funding.

Data Availability

The datasets generated or analyzed during this study are available from the corresponding author on reasonable request.

XSL•FO
RenderX

Authors' Contributions

CR-D, CL, CM-P, CC, and TS conceptualized the study. CR-D conducted the data analysis. CR-D, LJ, and TS analyzed and interpreted the data. CR-D, LJ, and TS coordinated manuscript writing. MS and AEM contributed to manuscript writing and revision. All authors read and approved the final manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Survey, teaching slide, and script.
[DOCX File, 310 KB - mededu_v12i1e78086_app1.docx]

References

1. Dubey R, Singh V, Khetarpal M, Singha SK, Sharma K, Neema P. A comparative study of the ease of radial artery cannulation using the palpation method in patients undergoing surgery under general anaesthesia: pre-induction versus post-induction. Cureus 2025 Feb 16;17(2):e79115. [doi: 10.7759/cureus.79115] [Medline: 40109821]
2. Gibbons RC, Zanaboni A, Saravitz SM, Costantino TG. Ultrasound guidance versus landmark-guided palpation for radial arterial line placement by novice emergency medicine interns: a randomized controlled trial. J Emerg Med 2020 Dec;59(6):911-917. [doi: 10.1016/j.jemermed.2020.07.029] [Medline: 32917440]
3. Aaron SD, Vandemheen KL, Naftel SA, Lewis MJ, Rodger MA. Topical tetracaine prior to arterial puncture: a randomized, placebo-controlled clinical trial. Respir Med 2003 Nov;97(11):1195-1199. [doi: 10.1016/s0954-6111(03)00226-9] [Medline: 14635973]
4. Tran NQ, Pretto JJ, Worsnop CJ. A randomized controlled trial of the effectiveness of topical amethocaine in reducing pain during arterial puncture. Chest 2002 Oct;122(4):1357-1360. [doi: 10.1378/chest.122.4.1357] [Medline: 12377864]
5. Haynes JM. Randomized controlled trial of cryoanalgesia (ice bag) to reduce pain associated with arterial puncture. Respir Care 2015 Jan;60(1):1-5. [doi: 10.4187/respcare.03312] [Medline: 25516991]
6. Beaumont M, Goret M, Orione C, et al. Effect of local anesthesia on pain during arterial puncture: the GAEL randomized placebo-controlled trial. Respir Care 2021 Jun;66(6):976-982. [doi: 10.4187/respcare.08328] [Medline: 33758058]
7. Lightowler JV, Elliott MW. Local anaesthetic infiltration prior to arterial puncture for blood gas analysis: a survey of current practice and a randomised double blind placebo controlled trial. J R Coll Physicians Lond 1997;31(6):645-646. [Medline: 9409498]
8. Valero Marco AV, Martínez Castillo C, Maciá Soler L. Local anesthesia in arterial puncture: nurses' knowledge and attitudes. Arch Bronconeumol 2008 Jul;44(7):360-363. [doi: 10.1016/s1579-2129(08)60062-2] [Medline: 18727888]
9. Patout M, Lamia B, Lhuillier E, et al. A randomized controlled trial on the effect of needle gauge on the pain and anxiety experienced during radial arterial puncture. PLoS ONE 2015 Sep 25;10(9):e0139432. [doi: 10.1371/journal.pone.0139432] [Medline: 26407017]
10. Hernández-Padilla JM, Granero-Molina J, Márquez-Hernández VV, Suthers F, Fernández-Sola C. Development and psychometric evaluation of the arterial puncture self-efficacy scale. Nurse Educ Today 2016 May;40:45-51. [doi: 10.1016/j.nedt.2016.02.008] [Medline: 27125149]
11. Vahedian-Azimi A, Rahimi-Bashar F, Pourhoseingholi MA, et al. Effect of the specific training course for competency in doing arterial blood gas sampling in the intensive care unit: developing a standardized learning curve according to the procedure's time and socioprofessional predictors. Biomed Res Int 2021 Feb 13;2021:2989213. [doi: 10.1155/2021/2989213] [Medline: 33628776]
12. McMillan HJ, Writer H, Moreau KA, et al. Lumbar puncture simulation in pediatric residency training: improving procedural competence and decreasing anxiety. BMC Med Educ 2016 Aug 8;16:198. [doi: 10.1186/s12909-016-0722-1] [Medline: 27502925]
13. Collins KP, Russian CJ, Gonzales JF. Teaching health care students the radial arterial puncture procedure. J Allied Health 2016;45(4):283-288. [Medline: 27915362]
14. Sreedharan JK, Gopalakrishnan GK, Jose AM, et al. Simulation-based teaching and learning in respiratory care education: a narrative review. Adv Med Educ Pract 2024 May 29;15:473-486. [doi: 10.2147/AMEP.S464629] [Medline: 38826695]
15. Cook DA, Hatala R, Brydges R, et al. Technology-enhanced simulation for health professions education: a systematic review and meta-analysis. JAMA 2011 Sep 7;306(9):978-988. [doi: 10.1001/jama.2011.1234] [Medline: 21900138]
16. Gent D, Kainth R. Simulation-based procedure training (SBPT) in rarely performed procedures: a blueprint for theory-informed design considerations. Adv Simul (Lond) 2022 May 8;7(1):13. [doi: 10.1186/s41077-022-00205-4] [Medline: 35527267]
17. Dev SP, Hillmer MD, Ferri M. Videos in clinical medicine. Arterial puncture for blood gas analysis. N Engl J Med 2011 Feb 3;364(5):e7. [doi: 10.1056/NEJMvcm0803851] [Medline: 21288091]

XSL•FO
RenderX

18. Gallagher EJ, Liebman M, Bijur PE. Prospective validation of clinically important changes in pain severity measured on a visual analog scale. Ann Emerg Med 2001 Dec;38(6):633-638. [doi: 10.1067/mem.2001.118863] [Medline: 11719741]
19. Dodds S, Williamson G. Nurse-led arterial blood gas sampling for patients. Nurs Times 2007 Feb;103(8):44-45. [Medline: 17333877]
20. Schmidt M, Freund Y, Alves M, et al. Video-based feedback of oral clinical presentations reduces the anxiety of ICU medical students: a multicentre, prospective, randomized study. BMC Med Educ 2014 May 22;14:103. [doi: 10.1186/1472-6920-14-103] [Medline: 24885005]
21. Toy S, McKay RS, Walker JL, Johnson S, Arnett JL. Using learner-centered, simulation-based training to improve medical students' procedural skills. J Med Educ Curric Dev 2017 Mar 15;4:2382120516684829. [doi: 10.1177/2382120516684829] [Medline: 29349329]
22. Soumerai SB, Avorn J. Principles of educational outreach ('academic detailing') to improve clinical decision making. JAMA 1990 Jan 26;263(4):549-556. [Medline: 2104640]
23. Luetsch K, Wong G, Rowett D. A realist synthesis of educational outreach visiting and integrated academic detailing to influence prescribing in ambulatory care: why relationships and dialogue matter. BMJ Qual Saf 2023 Dec 14;33(1):43-54. [doi: 10.1136/bmjqs-2022-015498] [Medline: 37142414]

Abbreviations

VAS: visual analogue scale

Edited by T Leung; submitted 26.May.2025; peer-reviewed by D Cunha, T Schreckenbach; revised version received 27.Nov.2025; accepted 31.Dec.2025; published 18.Feb.2026.

Please cite as:
Rolland-Debord C, Juret L, Simon M, El Mouhajer A, Londner C, Morélot-Panzini C, Chenivesse C, Similowski T
Impact of a Structured Training Program on Medical Student Confidence and Behavior During Their First Radial Arterial Puncture: Comparative Study
JMIR Med Educ 2026;12:e78086
URL: https://mededu.jmir.org/2026/1/e78086
doi:10.2196/78086

© Camille Rolland-Debord, Lucien Juret, Mathilde Simon, Abdallah El Mouhajer, Cécile Londner, Capucine Morélot-Panzini, Cécile Chenivesse, Thomas Similowski. Originally published in JMIR Medical Education (https://mededu.jmir.org), 18.Feb.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Education, is properly cited. The complete bibliographic information, a link to the original publication on https://mededu.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Perceptions and Attitudes of Medical Students Toward the Integration of Large Language Models in Medical Education: Cross-Sectional Survey in China

Cheng Zhao[1], MD; Weiqian Yan[2], MD; Long Wang[3], MD; Jing Wu[4], PhD; Herve Pasteur Ndikuriyo[1], MD; Renhe Yu[5], PhD

[1]Department of Urology, Xiangya Hospital, Central South University, No.87 Xiangya Road, Kaifu District, Changsha, China

[2]Department of Neurology, The Second Xiangya Hospital, Central South University, Changsha, China

[3]Department of Urology, The Third Xiangya Hospital, Central South University, Changsha, China

[4]Department of Education, Xiangya Hospital, Central South University, Changsha, China

[5]Department of Epidemiology and Health Statistics, Xiangya School of Public Health, Central South University, Changsha, China

Corresponding Author:
Cheng Zhao, MD
Department of Urology, Xiangya Hospital, Central South University, No.87 Xiangya Road, Kaifu District, Changsha, China

Abstract

Background: Although artificial intelligence (AI) is being rapidly integrated into medical education, insights from medical students, particularly in the Chinese context, remain limited.

Objective: This study was designed to explore Chinese medical students' perceptions of and attitudes toward the integration of AI into medical education, as well as the factors that may influence their perspectives. The findings of this research offer valuable insights to assist medical educators in the effective implementation of these innovative educational approaches.

Methods: On the basis of the estimated total number of clinical medical students at the target institutions, the sample size was calculated to be 379. A web-based questionnaire survey was designed to investigate the acceptance level of medical students toward the application of AI. The questionnaire consisted of 14 questions across 4 dimensions, which included demographic characteristics, perceptions of AI application, willingness, and concerns. Each dimension contained 3 to 4 questions. Descriptive statistics were used to tabulate the frequency of each variable. Chi-square tests and multiple regression analyses were conducted to measure the influencing factors.

Results: A total of 566 cross-sectional online surveys were distributed from December 2023 to January 2024 through a snowball sampling method. Finally, 490 medical students from various local tertiary medical centers were involved. Overall, a majority of the participants showed a positive attitude toward future learning and the usage of AI, manifested as totally willing to acquire relevant knowledge (222/490, 45.3%), totally willing to use AI tools (230/490, 46.9%), and totally desiring that schools would offer AI-related courses (230/490, 46.9%). However, there is still a large proportion (392/490, 80.0%) of participants who expressed concerns regarding ethical issues. The findings also indicated that gender and educational level were significantly correlated with the AI application. Specifically, regression analysis indicated that male participants were more inclined to acquire AI information through social media (odds ratio 0.458, 95% CI 0.33 - 0.67; *P*<.001) and that male or graduate-level participants were more likely to use AI for academic writing purposes (odds ratio 0.476, 95% CI 0.38 - 0.82; *P*=.001 for male; odds ratio 1.552, 95% CI 1.32 - 1.77; *P*=.009 for graduate students, respectively).

Conclusions: Our findings indicate that a general awareness of AI's role in medical education is evident among students. However, subgroup-specific differences must be taken into account, particularly when designing and optimizing educational strategies integrated with AI. This consideration is critical to ensuring that such tools align with the diverse learning needs of distinct student groups.

(JMIR Med Educ 2026;12:e66381) doi:10.2196/66381

KEYWORDS

artificial intelligence; medical education; large language models; Chinese medical students; perceptions; attitudes; cross-sectional survey

XSL•FO
RenderX

Introduction

Artificial intelligence (AI) is increasingly used across various sectors, such as finance, health care, and education [1-4]. In the field of medical education, AI holds the potential to enhance teaching and learning by improving students' critical thinking and communication skills [5]. However, while numerous studies have examined the views of medical educators, the perspectives of medical students themselves have received less attention. As modern medical education emphasizes collaboration between educators and learners, a thorough understanding of students' attitudes is crucial for designing and implementing effective teaching strategies [6,7].

As China undergoes a broad transformation driven by AI, the distribution of educational resources in medicine remains uneven [8]. Currently, only a limited number of medical schools have established specialized AI courses, primarily located in economically developed regions [9]. This regional disparity in educational resources has led to widespread misconceptions about AI among medical students, which in turn may amplify ethical and privacy-related risks [10]. While Chinese policymakers acknowledge the need for regulation in this rapidly evolving field, most existing policies remain high-level guidelines, providing limited practical direction for specific sectors such as medical education [11].

To address the aforementioned issues, a cross-sectional study was conducted. The primary goal of this study was to (1) investigate the perceptions, attitudes, and willingness of medical students in China toward the integration of large language models (LLMs) into medical education; (2) identify key influencing factors associated with these perceptions and attitudes; and (3) provide empirical evidence to inform the design of targeted AI-related educational strategies for Chinese medical institutions and discuss the regional disparities while using this novel technology. No a priori hypotheses were formulated, given the exploratory nature of investigating region-specific perspectives on LLM integration in medical education.

Methods

Study Population

A web-based questionnaire survey was conducted in the central south region of China, and all participants were recruited from 2 regional tertiary hospitals in Changsha and 3 tertiary hospitals affiliated with Central South University. The inclusion criteria were as follows: (1) full-time students majoring in clinical medicine, (2) each participant could respond only once per account, (3) the total response time for completing the questionnaire was longer than 60 seconds, and (4) participants were recruited between December 2023 and January 2024. The exclusion criteria included (1) inability to complete the questionnaire in Chinese, (2) a history of cognitive impairment or mental illness, and (3) no prior clinical practice experience. Specifically, the assessment of cognitive impairment or mental illness was conducted via a self-reported question at the start of the questionnaire: "Have you ever been diagnosed with a cognitive disorder or mental illness by a professional physician?" Participants who answered "yes" were automatically excluded from the survey to ensure data quality and protect their well-being. This screening method was conducted in compliance with the Declaration of Helsinki [12]. Finally, we used the CHERRIES (Checklist for Reporting Results of Internet E-Surveys) to guide our report [13].

Sample Size Calculation

The sample size for this cross-sectional survey was calculated to estimate the proportion of clinical medical students holding positive attitudes toward AI applications in medical education. The following formula for finite populations was used:

Formula: $n=[N\times z^2\times p(1-p)]/[(N-1)\times e^2+z^2\times p(1-p)]$

Parameters: N=total population size (estimated at 2000 clinical students across target hospitals), z=1.96 (for a 95% confidence level), *P*=.50 (conservative proportion assuming maximum variability), and e=0.05 (desired margin of error)

Calculation:

$n=[N\times z^2\times p(1-p)]/[(N-1)\times e^2+z^2\times p(1-p)]$

$=[2000\times(1.96)^2\times0.5\times.5]/[(1999\times0.0025)+(3.8416\times0.25)]$

$=[2000\times3.8416\times.25]/[4.9975+0.9604]$

$=1920.8/5.9579\approx322$

Nonresponse adjustment (15% attrition):

$n_adjusted=322/0.85\approx379$

A minimum of 379 participants was required.

Questionnaire Design

The questionnaire was developed based on a comprehensive review of the existing literature on related topics and was modified according to local perceptions in China [14,15]. As most of the participants included in our study were preclinical medical students, the questionnaire items were designed to avoid esoteric or ambiguous research-related or clinical questions as much as possible. Before the formal investigation, a pilot test of the questionnaires was conducted in this study. Three experts (LW, JW, and WY) were invited to conduct a systematic content validity evaluation. Subsequently, 20 students were recruited from different clinical disciplines, maintaining a male-to-female ratio of 1:1. The participants included both undergraduate and postgraduate students. These medical students completed the questionnaire to identify any encountered issues. On the basis of their feedback, the questionnaire was further refined. It should be noted that the responses from the pilot test were excluded from the final data analysis. The final questionnaire, consisting of 14 questions, was published on the WJX platform [16]. It addressed four dimensions: demographic characteristics, perceptions, willingness, and concerns. Each dimension encompassed 3 to 4 questions, and depending on the research target, question formats were either single or multiple choices.

Distribution of Questionnaire

We distributed the online questionnaire through "WeChat," (Tencent Holdings Limited) a widely used real-time messaging application in China. The team members contacted educators

XSL•FO
RenderX

from various hospitals through WeChat, informed them of the purpose of the survey, and then forwarded the questionnaire to them in WeChat groups of students from different grades. Medical students who agreed to participate scanned the QR code to anonymously answer the questionnaire online. To further enhance the reach and minimize potential bias, we used a snowball sampling method by requesting initial respondents to forward the survey link to other medical students in their networks [17].

Ethical Considerations

Informed consent and instructions were displayed at specific locations in the questionnaire. Participants were required to complete an informed consent form, which stated, "You acknowledge and agree that the data collected may be used for research purpose," as a prerequisite for proceeding with the survey. Participants could freely modify their selected options before final submission and retract their questionnaire at any time after delivery.

Given the nature of the study (anonymous, noninterventional, and without collection of personal or sensitive data), it was exempt from formal institutional review board review and approval, in accordance with the ethical guidelines of the Medical Ethics Committee of Xiangya Hospital, Central South University. The study was conducted in full compliance with relevant ethical principles, ensuring the protection of participants' privacy and rights throughout the research process. As the survey was voluntary and designed to take no more than 10 minutes to complete, no financial or nonfinancial compensation was provided to participants.

Data Analysis

After questionnaire collection, data cleaning was conducted. The questionnaires were screened based on the following predetermined criteria: (1) responses with excessive missing data, accounting for two-thirds of the total questions; (2) responses exhibiting uniform or highly regular answer selections; and (3) questionnaires completed in less than 60 seconds. These time thresholds were established as reasonable response times based on the pilot test conducted previously.

We used SPSS Statistics (version 26; IBM Corporation) for data analysis. First, descriptive analysis was performed to summarize the frequency distribution characteristics of the variables. Subsequently, subgroup analysis using chi-square tests was conducted to explore the associations between different demographic characteristics. Finally, regression analysis was applied to control for confounding factors and further investigate the associations between factors such as gender, age, and education level with the study objectives. If $P<.05$, differences were considered statistically significant. The reliability of the items from the willingness and concerns dimensions was tested using the Cronbach α coefficient analysis. The results of the Cronbach α were considered to have an acceptable reliability, as the generally accepted rule is that α values of 0.6 to 0.7 indicate an acceptable level of reliability, and 0.8 or greater is a very good level [18].

Results

Demographic Characteristics

A total of 566 questionnaires were distributed online, and 490 valid responses were collected (a response rate of 86.6%). The demographic characteristics of the participants are presented in Table 1. Overall, the gender distribution of participants was balanced (248/490, 50.7%, men vs 242/490, 49.3%, women), with the majority concentrated in the age range of 20 to 25 years (304/490, 62.0%), and only 32 individuals aged 35 years and older (32/490, 6.5%). The majority of participants were undergraduate students (258/490, 52.7%), followed by master's candidates (134/490, 27.3%), PhD candidates (74/490, 15.1%), and postdoctoral fellows (16/490, 3.3%). To clarify the Chinese medical education model, the following definitions apply to participant degree programs: (1) undergraduate students, that is, students pursuing a 5-year Bachelor of Medicine degree with a primary focus on foundational medical courses and who are not concurrently enrolled in other degree programs; (2) graduate students, including master's (3-y program) and PhD (3 - to 4-y program) students, who specialize in clinical subfields or medical research and are required to complete clinical rotations and publish academic papers; and (3) postdoctoral fellows, who are researchers who have completed a PhD and engage in advanced clinical or research work at hospitals; while not "students" in the traditional sense, they were included because they continue to participate in medical education training programs and represent the upper end of the medical learner spectrum. This classification aligns with national medical education standards in China [19].

XSL•FO
RenderX

Table . Demographic characteristics of study (N=490).

Characteristics	Values (%)
Sex	
Male	248 (50.7)
Female	242 (49.3)
Age (y)	
<20	8 (1.6)
21 - 25	304 (62.0)
26 - 30	100 (20.4)
31 - 35	46 (9.4)
>35	32 (6.5)
Educational level	
Junior college	8 (1.6)
Undergraduate degree	258 (52.7)
Master's degree	134 (27.3)
PhD	74 (15.1)
Postdoctorate	16 (3.3)

The specific frequency distribution is as follows: The age range of participants is from 18 to 38 years, and the median age is 23.1 (18.7-31.6) years. The extreme values were defined as ages less than 20 (8/490, 1.6%) or greater than 35 (32/490, 6.5%) years. Specifically, the proportion of participants aged less than 20 years was (8/490, 1.6%), all of whom were undergraduate students in their first or second year. Most participants (304/490, 62.0%) were aged 21 to 25 years, corresponding to undergraduate fourth- or fifth-year students or first-year postgraduates, followed by those aged 26 to 30 years (100/490, 20.4%), who were second- or third-year postgraduates, and those aged older than 30 years (78/490, 15.9%), who were PhD or postdoctoral fellows. This distribution is consistent with the age structure of medical students in China, where most enter postgraduate studies after 23 years of age [19].

Perceptions of AI Application

This set of questions primarily focused on participants' perceptions of the application of LLMs. The description of participants' perceptions of LLMs is presented in Tables 2-4. Subsequently, we conducted chi-square tests to investigate the correlation between other factors and perceptions. First, male participants were more likely to have been exposed to information related to LLMs before the conduction of the survey (158/248, 63.7% vs 132/242, 54.5% for women; χ^2_1=5.4; *P*=.02), as well as educational level also contributed to this issue (undergraduates: 162/258, 62.7%; master's students: 80/134, 59.7%; PhD candidates: 30/74, 40.5%; χ^2_1=19.854; *P*=.001). Second, gender (180/248, 72.5% vs 130/242, 53.7% for women; χ^2_1=18.7; *P*<.001), age (21 - 25 y: 142/304, 46.7%; 26 - 30 y: 60/100, 60%; and 31 - 35 y: 38/46, 82%; χ^2_1=26.4; *P*<.001), and educational level (undergraduates: 170/258, 65.8%; master's students: 68/134, 50.7%; PhD candidates: 56/74, 75.6%; χ^2_1=38.7; *P*<.001) were all correlated with the way that participants primarily acquired LLM-related information. A clear trend was observed that as education level increased, participants were more likely to access LLM information via academic journals or friends chatting and to recognize the value of LLM for research. Third, gender (206/248, 83.0% vs 168/242, 69.4% for women; χ^2_1=12.6; *P*<.001), age (21 - 25 y: 222/304, 73.0%; 26 - 30 y: 84/100, 84.0%; 31 - 35 y: 40/46, 87.0%; χ^2_1=25.5; *P*<.001), and educational level (undergraduates: 186/258, 72.1%; master's students: 110/134, 82.1%; PhD candidates: 60/74, 81.1%; χ^2_1=22.5; *P*<.001) also influenced participants' views on the application methods of LLMs in the medical field, especially in assisting in academic writing and literature translation. Finally, age (21 - 25 y: 232/304, 76.3%; 26 - 30 y: 84/100, 84%; 31 - 35 y: 42/46, 91.3%; χ^2_1=25.7; *P*<.001) and educational level (undergraduates: 198/258, 76.7%; master's students: 110/134, 82.1%; PhD candidates: 70/74, 94.6%; χ^2_1=25.0: *P*<.001) were also correlated with the challenges of LLMs in medical applications, specifically concerns about privacy protection. These trends indicated that older participants expressed greater concern about privacy risks compared to younger participants, which may reflect 2 key aspects: first, a higher demand for AI applications in the academic domain among more experienced learners, and second, a stronger awareness of data protection among this group.

XSL•FO
RenderX

Table . Chi-squared test between gender and categorical variables.

Variable	Gender, n		Pearson χ^2 (*df*)	*P* value[a]
	Male	Female		
Have you ever been exposed to information related to large language models?			5.4 (1)	.02
Yes	158	132		
No	90	110		
In what ways have you primarily acquired information about large language models?			18.7 (1)	<.001
Social media				
Yes	180	130		
No	68	112		
Friend chatting				
Yes	118	138		
No	130	104		
What do you think large language models can be used for in the field of medicine?			12.6 (1)	<.001
Assistance in writing academic papers				
Yes	206	168		
No	42	74		
Translation of literature				
Yes	186	160		
No	62	82		

[a]*P*<.05 was considered statistically significant.

XSL•FO
RenderX

Table . Chi-squared test between age and categorical variables.

Variable	Age (y), n					Pearson χ^2 (*df*)	*P* value[a]
	<20	21 - 25	26 - 30	31 - 35	>35		
In what ways have you primarily acquired information about large language models?						26.4 (1)	<.001
Friend chatting							
Yes	2	142	60	38	14		
No	6	162	40	8	18		
What do you think large language models can be used for in the field of medicine?						25.5 (1)	<.001
Assistance in writing academic papers							
Yes	2	222	84	40	26		
No	6	82	16	6	6		
Translation of literature							
Yes	0	208	80	36	22		
No	8	96	20	10	10		
What challenges do you believe large language models encounter in their applications within the field of medicine?						25.7 (1)	<.001
Privacy protection							
Yes	2	232	84	42	30		
No	6	72	16	4	2		

[a]*P*<.05 was considered statistically significant.

XSL•FO
RenderX

Table . Chi-squared test between education level and categorical variables.

Variable	Education level, n					Pearson χ^2 (*df*)	*P* value[a]
	Junior college	Undergraduate degree	Master's degree	PhD	Postdoctorate		
Have you ever been exposed to information related to large language models?						19.9 (1)	.001
Yes	6	162	80	30	4		
No	2	96	54	44	12		
In what ways have you primarily acquired information about large language models?						38.7 (1)	<.001
Academic journal							
Yes	2	102	64	36	12		
No	6	156	70	38	4		
Social media							
Yes	4	170	68	56	12		
No	4	88	66	18	4		
Lectures							
Yes	0	86	18	14	6		
No	8	172	116	60	10		
Friend chatting							
Yes	2	108	94	38	14		
No	6	150	40	36	2		
What do you think large language models can be used for in the field of medicine?						22.6 (1)	<.001
Summarize medical research papers							
Yes	6	216	124	72	14		
No	2	42	10	2	2		
Assistance in writing academic papers							
Yes	2	186	110	60	16		
No	6	72	24	14	0		
Translation of literature							
Yes	2	172	102	56	14		
No	6	86	32	18	2		
What challenges do you believe large language models encounter in their applications within the field of medicine?						25.0 (1)	<.001
Privacy protection							
Yes	4	198	110	70	8		
No	4	60	24	4	8		

[a]*P*<.05 was considered statistically significant.

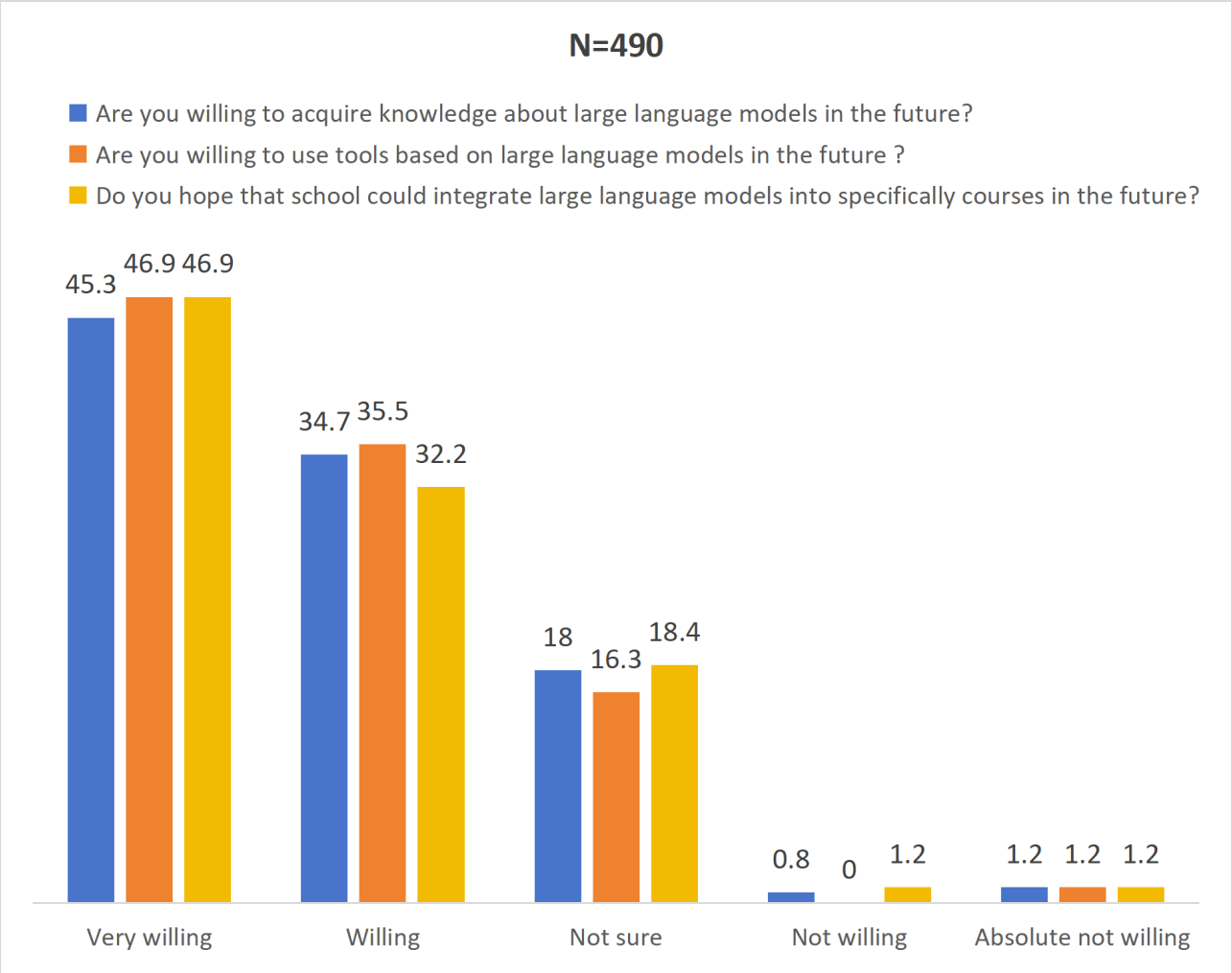
Willingness

The 3 questions in this dimension examined the participants' attitudes regarding their future studies and the application of or training courses related to LLMs. The responses were recorded using a 5-point Likert scale, ranging from 1 (totally agree) to 5 (totally disagree). Overall, all the participants demonstrated a positive willingness to accept and engage with AI-related

XSL•FO
RenderX

learning and training. Specifically, the proportion of participants who chose “completely agree” accounted for nearly half of the total number in all 3 questions. Additionally, the median scores were all relatively low (with lower scores indicating a higher degree of acceptance by the participants). Specific data are present in Figure 1.

Figure 1. Descriptive statistical analysis of the willingness domain. Numbers represent the percentage of participants in each category.

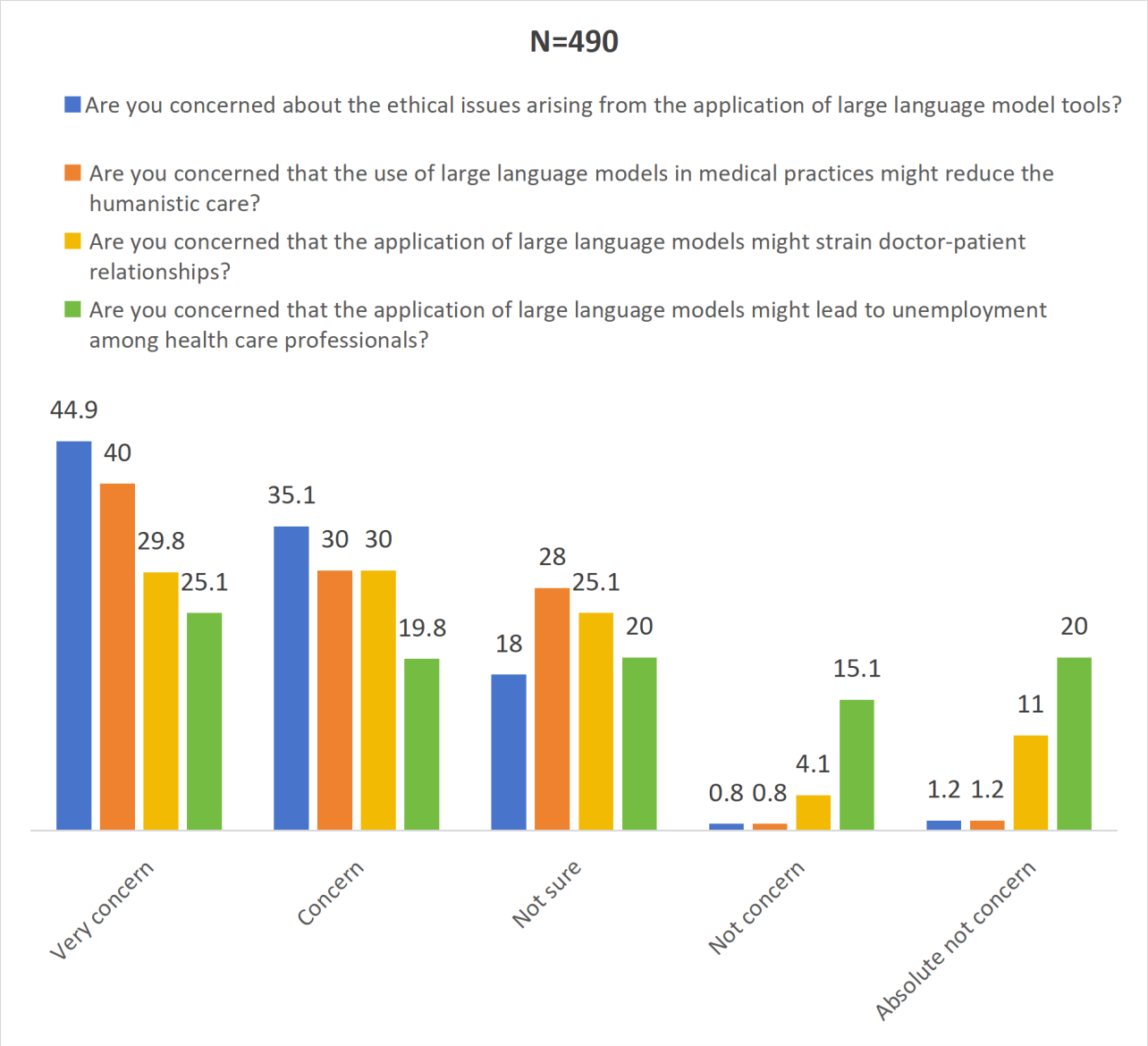


Concerns

We also assessed potential concerns about the integration of LLMs into medical education. First, the majority of participants expressed concerns about ethical issues, with more than half of them choosing to totally agree. However, in the subsequent questions related to medical issues, the number of participants choosing to totally agree largely decreased compared to the ethical questions. Moreover, regarding the issue of whether LLMs would lead to the unemployment of physicians, the vast majority of participants did not express any significant concern, with the median scores being relatively high. Specific data are shown in Figure 2.

XSL•FO
RenderX

Figure 2. Descriptive statistical analysis of the concerns domain. Numbers represent the percentage of participants in each category.



Regression Analysis

After adjusting for gender, age, and education, we observed that both gender and educational level had a significant influence on participants' awareness and attitudes. Specifically, male participants were more inclined to acquire relevant information through social media (odds ratio [OR] 0.458, 95% CI 0.33 - 0.67; *P*<.001). Furthermore, after adjusting for other variables, male participants (OR 0.476, 95% CI 0.38 - 0.82; *P*=.001) and graduate students (OR 1.552, 95% CI 1.32 - 1.77; *P*=.009) both showed a strong inclination toward using AI for academic writing. Data are presented in Table 5.

XSL•FO
RenderX

Table . Binary logistic regression between the categorical variables and different options.

Characteristics	Social media		Assistance in writing academic papers	
	P value	OR[a] (95% CI)	*P* value	OR (95% CI)
Sex	<.001[b]		.001[b]	
Male		Ref.[c]		Ref.
Female		0.458 (0.33 - 0.67)		0.476 (0.38 - 0.82)
Education level	.81		.009[b]	
Undergraduate degree		Ref.		Ref.
Graduate degree		1.033 (0.95 - 1.21)		1.552 (1.32 - 1.77)

[a]OR: odds ratio.

[b]*P*<.05 was considered statistically significant.

[c]Ref.: reference.

Discussion

Principal Findings

To summarize the principal findings of this study: Among 490 valid responses from Chinese medical students in Hunan, (1) participants generally held a positive attitude toward LLM integration into medical education, with 45.3% expressing full agreement with learning AI-related knowledge and 46.9% willing to use AI tools; (2) gender and educational level significantly influenced LLM-related perceptions—men were more likely to access LLM information via social media and use LLMs for academic writing, while postgraduates compared with undergraduates showed stronger recognition of the LLM's academic value; (3) ethical concerns were the most prominent worry, followed by privacy risks, but few participants feared LLM-induced physician unemployment; and (4) most students acquired LLM information through informal channels (social media or friends chatting), with only a few relying on formal medical school courses (academic journals or lectures).

Implications of Findings

Importantly, this study does not claim that students possess "substantial foundational knowledge of AI"—instead, it reports students' self-reported awareness and use of AI. No conclusions can be drawn about students' actual mastery of AI foundational knowledge because of the self-reported data limitations. However, for AI course design, the study's findings highlight specific gaps rather than definitive knowledge levels: (1) informal information channels, such as social media, may lead to an incomplete understanding of AI principles—thus, courses should include foundational modules on AI basics; (2) high ethical concerns suggest a need for modules on AI ethics (eg, case studies of AI-related academic misconduct and guidelines for patient data protection); and (3) the preference for research applications indicates that courses should integrate AI tools for thesis writing, literature review, and data analysis (aligned with Chinese medical education requirements). Future research should use objective assessments to identify specific knowledge gaps and inform the targeted curriculum design.

Comparison to the Literature

This study's findings align with some conventional knowledge in medical education while revealing region-specific differences.

Findings aligning with existing literature include the following. First, informal AI learning channels: consistent with one study [20], our results show that most Chinese medical students (63.2% and 52.2%) access AI information via nonformal channels (social media or friends chatting) rather than school courses, which aligns with global trends [20,21]. Specifically, 1 study has demonstrated that, on the premise of enhancing transparency and interpretability, leveraging social media as a channel helps expand the audience of medical research and provides opportunities for the early intervention of diseases [22]. However, another study also indicates that social media are associated with risks of data misuse, and its fundamental flaw lies in the lack of procedural standards and extremely low transparency. As an emerging field, AI will further amplify the risks when students acquire relevant information through social media platforms [23]. Second, gender differences in AI attitudes: our findings indicated that male participants were more likely to use AI (68.2% vs 51.2% of female participants), whereas female participants expressed greater concerns about risks (72.3% vs 54.8% of male participants). This result is consistent with surveys from other studies [24-26], which suggest cross-cultural consistency in gender-related AI perceptions. Third, low formal AI education coverage: our data (25.3% of students with formal AI courses) align with a Chinese survey reporting that regional disparities of AI-related instruction exist among medical schools in China, confirming the existence of a national education gap [9].

Novel or region-specific findings are as follows. First, preference for research over clinical AI use: unlike Western studies [27-29], where medical students prioritize AI for clinical tasks, our participants favored research applications (76.3% vs 32.7% for clinical use). This difference could be explained by China's medical education requirements, which mandate research experience and paper publication for postgraduates [19]. Second, low fear of AI-induced unemployment: only 12.7% of our participants worried about physician unemployment due to AI. This may reflect China's large health

XSL•FO
RenderX

care workforce gap [30], where AI is viewed as a complement rather than a replacement for physicians. Notably, all participants in our study were recruited from large tertiary hospitals located in provincial capitals; therefore, this conclusion may be subject to selection bias. Third, high ethical concerns focused on academic misconduct: while Western studies highlight clinical ethics [31,32], our participants prioritized academic ethics more likely due to strict academic integrity policies in Chinese medical schools [33].

At the policy level, the value of policy guidance in the field of AI applications has gained growing recognition. These policy frameworks not only establish effective norms for AI implementation in practical scenarios but also play a pivotal role in mitigating risks associated with AI deployment [34]. China's regulatory environment is characterized by centralized management and alignment with national strategic priorities, emphasizing that technological development should serve the overall interests of society, safeguard national security, and uphold social stability and moral norms [11]. This top-down approach facilitates coordinated deployment of AI in health care and education, ensuring consistency with long-term national goals. For example, the "Implementation Opinions on Promoting and Regulating the Application and Development of "AI+Healthcare" (issued by the National Health Commission and 4 other ministries in November 2025) sets targets for 2027 and 2030, explicitly supporting AI application in clinical scenarios [35]. Moreover, the "Reference Guidelines for AI Application Scenarios in the Health Industry" (issued by the National Health Commission in November 2024) defines 84 AI application directions in health care, including medical image analysis and intelligent drug research—providing clear guidance for designing AI education modules aligned with clinical needs [36]. These policies' emphasis on "application-oriented AI development" aligns with our finding that medical students prioritize AI use for academic and clinical tasks, highlighting the need for education to bridge the gap between policy goals and student practice. In contrast, the regulatory environment in Western countries adopts a more decentralized and issue-specific focus. For example, Europe prioritizes robust protection of personal privacy and data sovereignty, as exemplified by the General Data Protection Regulation and the EU AI Act [37], which classifies medical AI as "high risk" and mandates strict transparency and safety requirements. The United States emphasizes fostering technological innovation and market competition, often relying on industry self-regulation supplemented by targeted federal guidelines [38]. These measures exemplify China's proactive efforts toward digital transformation in the health care and education sectors, with policy acting as a catalyst for cross-sector collaboration.

Notably, disparities in resource allocation between different regions may give rise to structural differences in the AI literacy levels of medical students [39]. This gap is unlikely to be alleviated in the short term. In Hunan Province, as a relatively underdeveloped province located in the central-southern area of China, there are relatively fewer high-quality AI medical resources compared to economically developed coastal provinces. Although the "Artificial Intelligence+Teacher Team Construction Project" initiated by Hunan Province may promote the balanced development of education within the province [40], this policy is unlikely to be promoted to other provinces. Moreover, the concept that this policy views AI as a tool rather than a decision-making entity may inhibit students' critical thinking about the ethical risks of AI, thereby increasing the potential risks of its usage. Therefore, as the primary cohort of future users of AI-related applications, their insights are critical to facilitating the effective and contextually relevant adoption of this transformative technology, particularly in aligning its use with the unique demands of medical practice and education. These findings are not only regionally relevant but also have a significant impact on the digital transformation of medical education nationwide. However, when applying these findings to other regions, adjustments must be made based on the differences in the local policy and medical ecosystems.

Limitations

Notably, this study has several limitations. First, cross-sectional studies, although useful for identifying correlations between variables such as the use of LLMs and learning outcomes, are limited in their ability to establish causality. Second, while the sample is representative of the targeted region, its regional specificity may preclude full generalization to the broader population of Chinese medical students. Third, the relatively small sample size of the questionnaire survey may render certain conclusions vulnerable to random variation; expanding the sample size could potentially alter the robustness of these results. Furthermore, selection bias may exist, as medical students with limited interest in the medical applications of LLMs and those who are currently experiencing emotional distress were likely less motivated to participate in the survey. This nonresponse bias may have led to an overrepresentation of positive attitudes toward the use of LLMs in medical contexts. Finally, the study lacks data from two key stakeholders in the medical education ecosystem: practicing physicians and patients. Their perspectives on AI applications in medical practice and education would serve as a valuable complement to the current findings.

Conclusions

In summary, while the momentum toward AI integration in medicine is undeniable, its healthy development relies on coordinated efforts to advance technology responsibly, implement proactive regulation, and prioritize the diverse perspectives of those most impacted by its implementation. This approach will not only address the limitations identified in this study but also lay the groundwork for an AI-enabled medical ecosystem that is equitable, safe, and aligned with the core goals of health care and medical education.

XSL•FO
RenderX

Funding

This work was supported by the Education and Teaching Reform Project of Central South University (2023jy102 to WY; 2023jy077 to CZ); the Degree & Postgraduate Education Reform Project of Central South University (2023JGB010 to WY; 2024JGB141 to CZ); the Degree & Postgraduate Education Reform Project of Hunan Province (2023JGYB019 to WY); and the Natural Science Foundation of Hunan Province (2024JJ5572 to CZ).

Data Availability

The minimal dataset necessary to replicate the study findings is available from the corresponding author upon reasonable request.

Authors' Contributions

WY collected the data and drafted the manuscript. LW and JW performed the literature review and questionnaire design. RY conducted statistical analysis. HPN performed language polishing. CZ conceptualized the study and revised the manuscript.

Conflicts of Interest

None declared.

References

1. Cao Y. Using AI and big data analytics to support entrepreneurial decisions in the digital economy. Sci Rep 2025 Oct 22;15(1):36933. [doi: 10.1038/s41598-025-20871-4] [Medline: 41125685]
2. Roy P, Ghose B, Singh PK, Tyagi PK, Vasudevan A. Artificial intelligence and finance: a bibliometric review on the trends, influences, and research directions. F1000Res 2025;14:122. [doi: 10.12688/f1000research.160959.1] [Medline: 39911727]
3. Reddy S. Generative AI in healthcare: an implementation science informed translational path on application, integration and governance. Implement Sci 2024 Mar 15;19(1):27. [doi: 10.1186/s13012-024-01357-9] [Medline: 38491544]
4. Boscardin CK, Gin B, Golde PB, Hauer KE. ChatGPT and generative artificial intelligence for medical education: potential impact and opportunity. Acad Med 2024 Jan 1;99(1):22-27. [doi: 10.1097/ACM.0000000000005439] [Medline: 37651677]
5. Dave M, Patel N. Artificial intelligence in healthcare and education. Br Dent J 2023 May;234(10):761-764. [doi: 10.1038/s41415-023-5845-2] [Medline: 37237212]
6. Westerdahl F, Carlson E, Wennick A, Borglin G. Bachelor nursing students´ and their educators´ experiences of teaching strategies targeting critical thinking: A scoping review. Nurse Educ Pract 2022 Aug;63:103409. [doi: 10.1016/j.nepr.2022.103409] [Medline: 35868062]
7. Alamrani MH, Alammar KA, Alqahtani SS, Salem OA. Comparing the effects of simulation-based and traditional teaching methods on the critical thinking abilities and self-confidence of nursing students. J Nurs Res 2018 Jun;26(3):152-157. [doi: 10.1097/jnr.0000000000000231] [Medline: 29016466]
8. Tong W, Zhang X, Zeng H, Pan J, Gong C, Zhang H. Reforming China's secondary vocational medical education: adapting to the challenges and opportunities of the AI era. JMIR Med Educ 2024 Aug 15;10:e48594. [doi: 10.2196/48594] [Medline: 39149865]
9. Hu N, Jiang XQ, Wang YD, et al. Status and perceptions of ChatGPT utilization among medical students: a survey-based study. BMC Med Educ 2025 Jun 4;25(1):831. [doi: 10.1186/s12909-025-07438-7] [Medline: 40468340]
10. Lee H. Assessment of digital capabilities by 9 countries in the alliance for healthy cities using AI: cross-sectional analysis. JMIR Form Res 2025 Feb 7;9:e62935. [doi: 10.2196/62935] [Medline: 39924308]
11. He J, Baxter SL, Xu J, Xu J, Zhou X, Zhang K. The practical implementation of artificial intelligence technologies in medicine. Nat Med 2019 Jan;25(1):30-36. [doi: 10.1038/s41591-018-0307-0] [Medline: 30617336]
12. World Medical Association. World Medical Association Declaration of Helsinki: ethical principles for medical research involving human subjects. JAMA 2013 Nov 27;310(20):2191-2194. [doi: 10.1001/jama.2013.281053] [Medline: 24141714]
13. Eysenbach G. Improving the quality of Web surveys: the Checklist for Reporting Results of Internet E-Surveys (CHERRIES). J Med Internet Res 2004 Sep 29;6(3):e34. [doi: 10.2196/jmir.6.3.e34] [Medline: 15471760]
14. Tangadulrat P, Sono S, Tangtrakulwanich B. Using ChatGPT for clinical practice and medical education: cross-sectional survey of medical students' and physicians' perceptions. JMIR Med Educ 2023 Dec 22;9:e50658. [doi: 10.2196/50658] [Medline: 38133908]
15. Alkhaaldi SMI, Kassab CH, Dimassi Z, et al. Medical student experiences and perceptions of ChatGPT and artificial intelligence: cross-sectional study. JMIR Med Educ 2023 Dec 22;9:e51302. [doi: 10.2196/51302] [Medline: 38133911]
16. Wenjuanxing. 2025. URL: https://www.wjx.cn/ [accessed 2025-12-31]
17. Sarangi PK, Panda BB, P S, Pattanayak D, Panda S, Mondal H. Exploring radiology postgraduate students' engagement with large language models for educational purposes: a study of knowledge, attitudes, and practices. Indian J Radiol Imaging 2025 Jan;35(1):35-42. [doi: 10.1055/s-0044-1788605] [Medline: 39697505]
18. Tao W, Yang J, Qu X. Utilization of, perceptions on, and intention to use AI chatbots among medical students in China: national cross-sectional study. JMIR Med Educ 2024 Oct 28;10:e57132. [doi: 10.2196/57132] [Medline: 39466038]

XSL•FO
RenderX

19. Tie H, Luo Y, Chen D. Thinkings on the reform of medical education system in China. Med Educ Online 2024 Dec 31;29(1):2302677. [doi: 10.1080/10872981.2024.2302677] [Medline: 39611706]
20. Ravšelj D, Keržič D, Tomaževič N, et al. Higher education students' perceptions of ChatGPT: a global study of early reactions. PLoS ONE 2025;20(2):e0315011. [doi: 10.1371/journal.pone.0315011] [Medline: 39908277]
21. Gupta S, Kattapuram TM, Patel TY. Social media's role in the perception of radiologists and artificial intelligence. Clin Imaging 2020 Dec;68:158-160. [doi: 10.1016/j.clinimag.2020.06.003] [Medline: 32623195]
22. Hameed S, Nauman M, Akhtar N, Fayyaz MAB, Nawaz R. Explainable AI-driven depression detection from social media using natural language processing and black box machine learning models. Front Artif Intell 2025;8:1627078. [doi: 10.3389/frai.2025.1627078] [Medline: 41018736]
23. Kostygina G, Kim Y, Seeskin Z, LeClere F, Emery S. Disclosure standards for social media and generative artificial intelligence research: toward transparency and replicability. Soc Media Soc 2023;9(4):20563051231216947. [doi: 10.1177/20563051231216947] [Medline: 38239338]
24. McDermott H, Choudhury N, Lewin-Runacres M, Aemn I, Moss E. Gender differences in understanding and acceptance of robot-assisted surgery. J Robot Surg 2020 Feb;14(1):227-232. [doi: 10.1007/s11701-019-00960-z] [Medline: 31049775]
25. Shalaby R, Vuong W, Hrabok M, et al. Gender differences in satisfaction with a text messaging program (Text4Hope) and anticipated receptivity to technology-based health support during the COVID-19 pandemic: cross-sectional survey study. JMIR Mhealth Uhealth 2021 Apr 15;9(4):e24184. [doi: 10.2196/24184] [Medline: 33750738]
26. Dutta B, Peng MH, Sun SL. Modeling the adoption of personal health record (PHR) among individual: the effect of health-care technology self-efficacy and gender concern. Libyan J Med 2018 Dec;13(1):1500349. [doi: 10.1080/19932820.2018.1500349] [Medline: 30037314]
27. Karimov Z, Allahverdiyev I, Agayarov OY, Demir D, Almuradova E. ChatGPT vs UpToDate: comparative study of usefulness and reliability of Chatbot in common clinical presentations of otorhinolaryngology-head and neck surgery. Eur Arch Otorhinolaryngol 2024 Apr;281(4):2145-2151. [doi: 10.1007/s00405-023-08423-w] [Medline: 38217726]
28. Uche-Anya E, Anyane-Yeboa A, Berzin TM, Ghassemi M, May FP. Artificial intelligence in gastroenterology and hepatology: how to advance clinical practice while ensuring health equity. Gut 2022 Sep;71(9):1909-1915. [doi: 10.1136/gutjnl-2021-326271] [Medline: 35688612]
29. Shickel B, Ozrazgat-Baslanti T, Bihorac A. Artificial intelligence at the bedside for the prevention, detection, and management of acute kidney injury. Curr Opin Nephrol Hypertens 2025 Nov 1;34(6):483-490. [doi: 10.1097/MNH.0000000000001112] [Medline: 40916929]
30. Dong E, Sun X, Xu T, et al. Measuring the inequalities in healthcare resource in facility and workforce: a longitudinal study in China. Front Public Health 2023;11:1074417. [doi: 10.3389/fpubh.2023.1074417] [Medline: 37006575]
31. Badawy W, Zinhom H, Shaban M. Navigating ethical considerations in the use of artificial intelligence for patient care: a systematic review. Int Nurs Rev 2025 Sep;72(3):e13059. [doi: 10.1111/inr.13059] [Medline: 39545614]
32. Kleebayoon A, Wiwanitkit V. Ethics for artificial intelligence use in clinical pharmacology. Indian J Pharmacol 2024 May 1;56(3):224-225. [doi: 10.4103/ijp.ijp_729_23] [Medline: 39078188]
33. Jin Y, Wei J, Zhang J, et al. A study on the knowledge, attitude, and practice of research integrity among medical professionals in Ningxia, China. BMC Med Educ 2024 Nov 23;24(1):1355. [doi: 10.1186/s12909-024-06380-4] [Medline: 39580401]
34. Preiksaitis C, Rose C. Opportunities, challenges, and future directions of generative artificial intelligence in medical education: scoping review. JMIR Med Educ 2023 Oct 20;9:e48785. [doi: 10.2196/48785] [Medline: 37862079]
35. Implementation opinions on promoting and regulating the application and development of "AI + healthcare". National Health Commission of the People's Republic of China. 2025. URL: https://www.gov.cn/zhengce/202511/content_7047015.htm [accessed 2025-12-31]
36. National Health Commission of the People's Republic of China reference guidelines for AI application scenarios in the health industry. CNR. 2024. URL: http://www.cnr.cn/newscenter/native/gd/20251110/t20251110_527425853.shtml [accessed 2025-12-31]
37. Data protection under GDPR. European Union. URL: https://europa.eu/youreurope/business/dealing-with-customers/data-protection/data-protection-gdpr/ [accessed 2025-12-31]
38. Artificial intelligence program: research on AI/ML-based medical devices. US Food and Drug Administration. URL: https://www.fda.gov/medical-devices/medical-device-regulatory-science-research-programs-conducted-osel/artificial-intelligence-program-research-aiml-based-medical-devices [accessed 2025-12-31]
39. Gumilar KE, Indraprasta BR, Hsu YC, et al. Disparities in medical recommendations from AI-based chatbots across different countries/regions. Sci Rep 2024 Jul 24;14(1):17052. [doi: 10.1038/s41598-024-67689-0] [Medline: 39048640]
40. Changsha city, hunan province, is taking multiple measures to promote the development of digital teachers. Ministry of Education of the People's Republic of China. 2025. URL: http://www.moe.gov.cn/jyb_xwfb/s6192/s222/moe_1750/202509/t20250916_1413597.html [accessed 2025-12-31]

Abbreviations

AI: artificial intelligence

XSL•FO
RenderX

CHERRIES: Checklist for Reporting Results of Internet E-Surveys
LLM: large language model
OR: odds ratio

Edited by B Lesselroth; submitted 11.Sep.2024; peer-reviewed by OA Aboshady, O Allela, PK Sarangi; accepted 03.Dec.2025; published 09.Mar.2026.

Please cite as:
Zhao C, Yan W, Wang L, Wu J, Ndikuriyo HP, Yu R
Perceptions and Attitudes of Medical Students Toward the Integration of Large Language Models in Medical Education: Cross-Sectional Survey in China
JMIR Med Educ 2026;12:e66381
URL: https://mededu.jmir.org/2026/1/e66381
doi:10.2196/66381

© Cheng Zhao, Weiqian Yan, Long Wang, Jing Wu, Herve Pasteur Ndikuriyo, Renhe Yu. Originally published in JMIR Medical Education (https://mededu.jmir.org), 9.Mar.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Education, is properly cited. The complete bibliographic information, a link to the original publication on https://mededu.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Questionnaires on Perceptions of Artificial Intelligence in Health Care Among Health Care Students: Cross-Cultural Translation Into French and Linguistic Validation

Sylvain Kotzki[1], PharmD, PhD; Calvin Massonnet Turner[1], MEng; Nicolas Vuillerme[1,2], PhD

[1]Grenoble INP, LIG Sangria, CNRS, UMR 5217, Université Grenoble Alpes, Bureau 315, Bâtiment Jean Roget, Campus Santé, Domaine de La Merci, Grenoble, France

[2]Institut Universitaire de France, Paris, France

Corresponding Author:
Sylvain Kotzki, PharmD, PhD
Grenoble INP, LIG Sangria, CNRS, UMR 5217, Université Grenoble Alpes, Bureau 315, Bâtiment Jean Roget, Campus Santé, Domaine de La Merci, Grenoble, France

Abstract

Background: Artificial intelligence (AI) is rapidly transforming health care by enhancing diagnostic accuracy, optimizing clinical workflows, and supporting decision-making across all health disciplines. As AI-driven tools are progressively introduced into health systems, educating future professionals about AI has become a critical priority to ensure safe, ethical, and effective use. Although several validated English-language questionnaires exist to assess medical students' perceptions and readiness on AI in medicine, no French-language equivalents are currently available, which limits their use in francophone settings and hampers international comparisons. To bridge this gap and enable comparable, evidence-based assessment of AI perceptions among French health care students, rigorous cross-cultural adaptation of validated instruments is essential.

Objective: This study aimed to translate, culturally adapt, and linguistically validate 5 established English-language questionnaires assessing medical students' perceptions of AI in medicine to produce French versions suitable for subsequent psychometric validation and use across health care training programs.

Methods: We followed international guidelines for the cross-cultural adaptation of self-report measures, combining independent forward translations, reconciliation, backward translation, expert committee review, and cognitive debriefing. Two bilingual translators first produced independent French versions of each questionnaire, which were reconciled into a single draft. A third bilingual translator, blinded to the original instruments, then performed backward translation into English. An expert panel reviewed all versions to ensure conceptual equivalence and to adapt items for applicability across health professions. Finally, cognitive testing was conducted with 38 French health care students (in medicine, pharmacy, adapted physical activity and health, nursing, and midwifery) to assess clarity, comprehensibility, and acceptability with iterative revisions made until consensus was reached.

Results: During forward translation, wording discrepancies were observed for 73.6% (148/201) of expressions, but only 1.0% (2/201) of items required resolution due to meaning differences. In the backward translation step, 97.0% (195/201) of expressions were judged to be conceptually equivalent to the originals; the remaining 3.0% (6/201) of expressions were revised after discussion. Cognitive debriefing with students led to minor wording modifications in 26.4% (53/201) of expressions to improve clarity and readability without altering the underlying concepts.

Conclusions: We produced French-language versions of 5 widely used questionnaires assessing health care students' perceptions of AI in medicine, following a rigorous cross-cultural translation, adaptation, and linguistic validation process. These instruments preserve conceptual equivalence with their English originals and provide standardized tools to document AI-related knowledge, attitudes, and intentions among French-speaking health care students. This work lays the groundwork for subsequent psychometric studies of these French-language versions of questionnaires used in diverse health care training programs.

(JMIR Med Educ 2026;12:e76572) doi:10.2196/76572

KEYWORDS

health care students; artificial intelligence perception; AI perception; cross-cultural adaptation; curriculum; linguistic validation; artificial intelligence; AI

XSL•FO
RenderX

Introduction

Artificial intelligence (AI) is revolutionizing health care by providing transformative tools for early detection, diagnosis, decision-making, therapeutic management, and patient follow-up [1-5]. Over the past 2 decades, the digitization of data and processes worldwide has propelled a massive transformation in patient care models and strategies [6]. However, integrating AI into health care implies more than just new technical tools; it demands a profound reorganization of health systems and clinical practices. While predictive medicine, which anticipates patient needs, and personalized medicine, which optimizes care from the moment of diagnosis, offer unprecedented prospects [7], this new landscape raises at least 2 crucial issues: trust in inherently opaque algorithms [8] and the responsibility of health professionals using them in health care [9].

Despite encouraging results from certain AI algorithms [10], numerous health care professionals remain cautious [11]. Indeed, over 60% report reluctance to adopt AI because of limited transparency [12], concerns regarding validation [13], and ethical [14] as well as security [15] risks. Asan et al [16] have shown that user training is key to fostering trust and clarifying responsibilities. Similarly, Yakar et al [17] demonstrate that practitioners with limited AI knowledge fear being replaced by automated systems, stressing the need for education that addresses both technical and ethical dimensions. Moreover, multiple studies highlight how specialized training programs can strengthen AI literacy, covering theoretical knowledge, practical skills, and ethical reflection, although the precise competencies to teach remain debated [18-20]. In France, such training programs are only beginning to be integrated into medical and paramedical curricula [21], revealing an initial lack of readiness that might hinder appropriate use and acceptance of AI.

Therefore, assessing health care students' knowledge, perceptions, and motivations regarding AI is critical for designing targeted educational strategies. In other countries, surveys have already illuminated students' expectations and concerns about AI in health care [22-24]. To achieve this goal, surveys based on validated questionnaires are necessary for gathering data that is both comparable and actionable. Thus far, several questionnaires have been specifically designed to assess medical students' perceptions of AI in medicine. Among them, we selected five instruments as follows: (1) the Medical Artificial Intelligence Readiness Scale for Medical Students (MAIRS-MS) [25], (2) the questionnaire by Sit et al [26], (3) the questionnaire by Boillat et al [27], (4) the questionnaire by Li et al [28], and (5) the questionnaire by Park et al [29]. In their original form, these instruments provide robust conceptual frameworks for exploring AI readiness, perceived impact on specialties, behavioral intentions to learn about AI, and perceived barriers and risks, but they are all available only in English and have been developed in specific educational and cultural contexts.

First, the MAIRS-MS by Karaca et al [25] assesses readiness to integrate AI into clinical practice, focusing on theoretical knowledge, practical skills, and ethical considerations. Second, the questionnaire by Sit et al [26] explores UK medical students' perceptions of AI, especially its impact on radiology, by probing both their optimism and concerns regarding technological substitution. Third, Boillat et al [27] offer a questionnaire designed to identify both obstacles and facilitators to AI adoption among health care professionals and medical students, addressing challenges such as algorithmic complexity and the risk of dehumanizing care. Fourth, Li et al [28] investigate factors influencing medical students' intentions to pursue AI training, taking into account social pressure, the perceived usefulness of AI in medicine, and self-confidence in mastering the technology. Finally, Park et al [29] delve into how AI shapes medical students' perceptions of medical specialties and their career choices by assessing their confidence levels and primary sources of information about these emerging technologies.

Collectively, these 5 questionnaires [25-29], which are specifically addressed to medical students, provide a comprehensive and complementary perspective on the challenges and opportunities related to training future health care professionals in AI. However, since they are currently available only in English, their application to French-speaking settings remains limited. Moreover, none of these instruments has yet undergone a systematic cross-cultural linguistic adaptation for French-speaking health care students, which prevents their direct use for curriculum development and comparative research in francophone contexts.

In this context, our study aimed to translate, culturally adapt, and linguistically validate 5 key instruments, producing French versions that are conceptually equivalent to the originals and suitable for use across different health training programs. In addition to these conceptual and educational challenges, the rigorous adaptation of perception questionnaires requires a structured cross-cultural procedure (translation, cultural adaptation, and linguistic validation) distinct from subsequent psychometric validation, as recommended in international guidelines for self-report measures [30]. Although the original instruments were designed for medical students, we adapted the wording so that they can be administered to students from different health professions in the French context; however, their psychometric properties and applicability across diverse health training programs will need to be confirmed in subsequent studies.

Methods

The transcultural adaptation of the questionnaires into French was carried out following international guidelines for the cross-cultural adaptation of self-report measures, combining independent forward translations, backward translation, expert committee review, and cognitive debriefing [30-32]. Operationally, the procedure was implemented using a methodology adapted from Epstein et al [33], previously applied to French-language instruments.

Ethical Considerations

This study has received ethics approval from the Ethics Committee for the Integrity and Ethics of Research in Health Professions Education of the Société Internationale Francophone

XSL•FO
RenderX

d'Education Médicale (1210-2023, issued on November 26, 2023).

This study corresponded to a qualitative cognitive debriefing or pilot linguistic validation phase (ie, the validation of version D of each questionnaire, the version sent to students), aimed at assessing the clarity, comprehensibility, and acceptability of translated items. In accordance with applicable French regulations for minimal-risk educational research, written informed consent was not required. Participants were required to read the study information sheet before participation, and oral informed consent was obtained prior to collecting their feedback. Participants could decline or withdraw at any time without consequences for their training or assessment.

Feedback was collected and analyzed for the sole purpose of improving linguistic clarity and conceptual equivalence of the French versions. Data were handled confidentially, with access restricted to the research team. The study followed institutional procedures and applicable French data protection requirements (law No. 78-17 of 6 January 1978, as amended), including provisions related to the Commission nationale de l'informatique et des libertés, where applicable. Results are reported in the aggregate, and no identifying information is disclosed in this paper.

No compensation, payment, or academic credit was provided.

Questionnaires

Five questionnaires were targeted for cross-cultural validation in this study. The first questionnaire, MAIRS-MS, proposed by Karaca et al [25], evaluates the readiness of Turkish medical students to integrate AI into their future clinical practice. It includes 22 items distributed among 4 factors: cognitive (8 items), aptitude (8 items), vision (3 items), and ethical (3 items). Responses are recorded on a 5-point Likert scale ranging from 1 ("strongly disagree") to 5 ("strongly agree"), yielding a total score between 22 and 110 points, with higher scores indicating greater readiness for AI integration in medical practice.

The second questionnaire, developed by Sit et al [26], explores UK medical students' perceptions of AI, with particular emphasis on its impact on radiology. It includes 15 questions in total, 11 of which are formatted as a 5-point Likert scale ranging from 1 ("strongly disagree") to 5 ("strongly agree") and 1 with a 5-point Likert scale ranging from 1 ("extremely useful") to 5 ("not at all useful"). These items assess medical students' knowledge of AI, their perceptions of its role in the evolution of the medical profession, and their potential apprehensions about technological substitution.

The third questionnaire, proposed by Boillat et al [27], consists of 17 questions designed to measure the familiarity of medical students and physicians with AI and to identify perceived barriers to its adoption in clinical practice. This questionnaire is structured around four themes: (1) familiarity with AI; (2) AI training; (3) factors facilitating AI implementation, including obstacles such as the complexity of algorithms, data availability, and trust in AI tools; and (4) risks associated with AI development, such as the dehumanization of care or the gradual loss of clinical skills. Three of the themes are formatted as a 5-point Likert scale ranging from 1 ("strongly disagree") to 5 ("strongly agree"), and 1 theme is formatted as a 5-point Likert scale ranging from 1 ("never heard of it") to 5 ("can confidently explain it").

The fourth questionnaire, proposed by Li et al [28], comprises 23 items that examine the factors influencing the motivation of medical students to learn about AI and integrate these technologies into their training. This questionnaire includes 5 subscales: the perceived relevance of AI in medicine, the influence of peers and teachers, confidence in one's ability to understand and use AI, basic knowledge of AI, and behavioral intention to learn AI. Responses are recorded on a 6-point Likert scale ranging from 1 ("not true of me at all") to 6 ("very much true of me).

Finally, the fifth questionnaire, proposed by Park et al [29], focuses on medical students' perceptions of the impact of AI on medical practice and its influence on specialty choice. It seeks to assess the extent to which AI affects students' career prospects and includes 6 multiple-choice questions, 2 Likert-scale items, and open-ended questions to capture qualitative insights.

Forward Translation

The forward translation of each questionnaire from the original English into French was performed by 2 independent bilingual translators. One translator (translator 1) was a native English speaker, while the other (translator 2) was a native French speaker. Translator 1 was familiar with the concepts evaluated by the 5 questionnaires, while translator 2 was not informed of their intended purpose. The 2 translations were later compiled, and in cases of disagreement, the 2 translators were consulted to reach an agreement. Each questionnaire was then thoroughly reviewed to correct errors in spelling, grammar, and punctuation and to ensure that terminology and style were faithful to the original English version.

Backward Translation

Following this, backward translation was conducted. Version A of each questionnaire was translated from French back into English (version B) by a bilingual translator whose native language was English (back translator) and who had no expertise in health care or AI. The back translator was instructed to translate each questionnaire literally. The original English version and the back-translated version were then compared and reconciled to produce a synthesized French version (version C), ensuring translation accuracy and highlighting any potentially confusing wording.

Adaptation of Questionnaires to All Health Training Courses

Version C was then submitted to a health professional to propose adjustments that would extend the applicability of the questionnaires to all health training courses. For each questionnaire, this health professional was asked to suggest modifications when questions were considered exclusively proper for physician training or medical practice, thereby producing an extended version (version D).

XSL•FO
RenderX

Final Validation

The final validation phase involved conducting a preliminary pilot test with native French-speaking participants representative of the target population to assess the clarity, intelligibility, and acceptability of the translations. Between October 10, 2024, and December 1, 2025, version D was sent to 38 students enrolled in various initial health training programs in France (women: n=27, 71.1%; men: n=11, 28.9%), including pharmacy (n=10, 26.3%), medicine (n=8, 21.1%), adapted physical activity and health (n=8, 21.1%), nursing (n=6, 15.8%), and midwifery (n=6, 15.8%) programs. They were asked to review the intelligibility of the questions and the proposed answers in version D, and their feedback was synthesized to produce version E.

Data Analysis

All qualitative feedback obtained during the final verification phase was analyzed using a thematic approach. Comments were independently coded by 2 researchers to identify issues related to clarity, comprehension, terminology, and formatting. Any discrepancies in coding were resolved through discussion until consensus was reached. On the basis of these findings, problematic items were iteratively revised to enhance semantic equivalence with the original questionnaires and improve readability in French. No quantitative statistical analyses were conducted during this phase, as the primary objective was to ensure content validity and linguistic clarity of the translated questionnaires.

Results

Characteristics of the Translators and the Participants

The age, gender, and specific cultural characteristics of the translators and the student participants are detailed in Table 1.

Table . Age, gender, and cultural characteristics of the translators and the participants.

Step	Number of translators or participants	Age (y), mean (SD; range)	Gender	Cultural characteristics
Forward translation	2 translators	41.5 (13.5; 28-55)	1 man and 1 woman	1 native English speaker and 1 native French speaker; 1 familiar with the concept of AI[a] and 1 unfamiliar with the concept of AI
Backward translation	1 translator	37 (—[b])	1 woman	1 native English speaker
Adaptation of health studies	1 expert	40 (—)	1 man	1 health care professional with more than 10 years of experience
Final verification	38 students	23.6 (5.57; 18-50)	27 women and 11 men	10 pharmacy students, 8 medical students, 8 adapted physical activity and health care students, 6 nursing students, and 6 midwifery students

[a]AI: artificial intelligence.

[b]Not applicable.

Forward Translation

Across the 5 instruments, the 2 forward translators proposed identical wording for 26.4% (53/201) of the expressions, whereas 73.6% (148/201) of the expressions differed in vocabulary, register, or syntax. In almost all cases, these discrepancies did not affect the underlying meaning, and a consensual French wording was obtained through discussion and recontextualization of the items. Only 2 items required arbitration because the alternative translations conveyed slightly different concepts. For example, for the MAIRS-MS item “I find valuable to use AI for education,” translator 1 suggested “J’estime qu’il est utile d’utiliser l’IA à des fins d’éducation,” whereas translator 2 proposed “Je trouve valorisant l’utilisation de l’IA à des fins de pédagogie.” After discussion, the first option was retained because it emphasized the perceived usefulness of AI in education rather than the personal gratification implied by “valorisant.” Similarly, for 1 item in the questionnaire by Sit et al [26] concerning students’ anticipated understanding of methods to evaluate AI algorithm performance, the agreed wording (“je maîtriserai davantage les méthodes utilisées pour évaluer les performances des algorithmes d’IA dans la santé”) was preferred over alternatives focusing on “connaissance” alone to reflect the level of competence implied in the original English version. In the remaining items, minor stylistic variations (eg, “Je peux...” vs “Je suis capable de...”) were resolved in favor of the simplest and most natural phrasing for health care students, without altering the constructs being measured.

Backward Translation

When the back-translated English versions (version B) were compared with the original questionnaires, no loss or change in meaning was identified for 3 instruments (Karaca et al [25], Sit et al [26], and Park et al [29]). Overall, 97.0% (195/201) of the expressions were judged to be conceptually equivalent to the originals, and only 3.0% (6/201) of the expressions required revision after discussion with the back translator. In the

XSL•FO
RenderX

questionnaire by Boillat et al [27], only 1 expression exhibited a variation in meaning between translator 3's version and the original version. The proposals "Do you think it is necessary to receive more in-depth training in order to" from version B and "Would you benefit from more training to" from the original version were discussed with the back translator, which resulted in replacing the expression "Pourriez-vous bénéficier d'une formation plus poussée pour" ("Could you benefit from further training for") in version A with "Pensez-vous nécessaire de bénéficier d'une formation plus approfondie pour" ("Do you think it is necessary to benefit from more in-depth training to") in version C. In the questionnaire by Li et al [28], 5 phrases showed meaning discrepancies and led to minor wording adjustments in the French version to better reflect the original notions of "health care solutions," future concerns about the development of medical AI, and actual interaction with medical AI applications (eg, replacing "les solutions de soins" ["care solutions"] with "les solutions de santé" ["health solutions"] and "J'ai étudié des applications médicales" ["I have studied medical apps"] with "J'ai interagi avec des applications médicales" ["I have used medical apps"]). All remaining differences between the back-translated and original versions were considered stylistic and did not require changes to the French wording.

Adaptation of Questionnaires to All Health Training Courses

The health professional who reviewed version C proposed several modifications to extend the scope of the 5 questionnaires to all students in health training. For all questionnaires, and when no change in meaning or interpretation was involved, the health professional systematically proposed the following substitutions: "médecine" ("medicine") was replaced with "santé" ("health"), "médecins" ("physicians") was replaced with "professionnels de santé" ("health professionals"), "radiologie" ("radiology") was replaced with "imagerie médicale" ("medical imaging"), "diplôme de médecine" ("medical degree") was replaced with "diplôme de professionnel de santé" ("health professional degree"), "pratiques cliniques" ("clinical practices") was replaced with "pratiques professionnelles" ("professional practices"), and "applications médicales" ("medical applications") was replaced with "applications en santé" ("applications in health"). In the questionnaire by Boillat et al [27], the list of medical specializations was replaced by a list of "professions de santé" ("health professions") to align with the objective of creating an extended version of the questionnaire. All these modifications were later incorporated into version D of each questionnaire.

Final Validation

The extended versions of the 5 questionnaires (version D) were submitted to 38 native French-speaking health care students from 5 different health training programs (pharmacy, medicine, adapted physical activity and health, nursing, and midwifery). Each student was invited to propose modifications to ensure the intelligibility of both the questions and the response options. A total of 53 expressions were subject to revision proposals for improved clarity. For example, the phrase "Je peux établir des plans de travail compatibles avec l'IA" ("I can build AI-enabled worktops") was replaced with "Je suis capable d'organiser des flux de travail compatibles avec l'intelligence artificielle" ("I'm able to organize AI-enabled workflows"), a version deemed more intelligible by the student panel. The complete French versions of the 5 questionnaires are available in Multimedia Appendix 1.

Discussion

Principal Findings

In this study, we translated, culturally adapted, and linguistically validated 5 established English-language questionnaires [25-29] to assess French health care students' perceptions of AI, using a rigorous forward-backward translation process, expert arbitration, and cognitive testing with 38 French students from 5 different health training programs (pharmacy, medicine, adapted physical activity and health, nursing, and midwifery). We produced French versions that maintain strong conceptual equivalence with the original versions while ensuring clarity and accessibility for an academic audience.

The health care sector is undergoing a profound digital transformation, making AI literacy an essential competency for future health professionals [18,19]. Capturing the in-depth views, knowledge, and expectations of French health care students on AI is a crucial first step, and perception surveys provide an effective means to gather these insights. For instance, a 2020 European survey of 451 medical students found that although 40% were preparing for an increasingly digitized health system, more than half felt inadequately trained in digital technologies and nearly 85% called for enhanced digital health education in the medical curriculum [22]. Similarly, in the United States, a study published in 2022 revealed that 91.5% of the 390 surveyed medical students believed that AI training would be beneficial for their future practice, yet 91.2% reported that their institutions provided little or no formal instruction on the subject [24]. A 2022 Canadian survey involving 2167 students across 10 health disciplines indicated that 79% of respondents anticipated that AI would impact their careers within the next decade [23]. Even initially skeptical students recognize the need to acquire fundamental AI skills in their training, despite some concerns, for example, an Austrian study highlighted skepticism among medical students who viewed digitalization as a potential threat to the physician-patient relationship [34]. Collectively, these studies highlight both a strong demand for AI education and persistent concerns regarding its impact on professional roles.

Insights from these surveys can directly inform curriculum design by prioritizing fundamental AI concepts, ethical and legal issues, and practical hands-on training. European medical students expressed a desire for dedicated courses addressing health data management, ethical and legal issues, and other digital health topics, alongside more practical, hands-on training through workshops or exercises that illustrate the use of AI and digital health tools [22]. Similarly, American medical students favored learning formats such as short lectures (70%), optional dedicated courses (48%), and question and answer sessions with experts (44%), and they emphasized the need to focus on

XSL•FO
RenderX

fundamental AI concepts and criteria for its application in medicine rather than solely on technical aspects [24].

In this context, our French adaptations of 5 complementary questionnaires [25-29] fill an important gap in French-speaking contexts by providing, for the first time, a coherent set of standardized tools that, once fully validated, will support the systematic assessment of AI readiness, perceived impact on specialties, behavioral intentions to engage with AI, and perceived barriers and risks among health care students. Beyond enabling cross-sectional surveys, these instruments can also be used longitudinally to monitor changes in perceptions over time, compare cohorts or curricula, and evaluate the impact of new educational interventions related to AI and digital health.

Limitations

This study has several limitations that must be acknowledged. First, the cognitive testing was conducted with a convenience sample of 38 students from 5 different health training programs within a single institution. Although this multiprofessional sample improves the face validity and perceived applicability of the French versions across health curricula, it remains a single-center, relatively small pretest, typical of exploratory cognitive debriefing in cross-cultural adaptation studies and comparable to previous work using similar forward-backward methodologies [35,36]. Second, we did not perform any psychometric analyses (eg, internal consistency, factor structure, and construct validity) in this study; these will be the focus of subsequent validation work.

Future Research

Future work should build on this linguistic validation by conducting full psychometric evaluations of the French versions of these instruments in larger and more diverse samples of health care students. Key steps will include assessing internal consistency and test-retest reliability; examining the factor structure of each questionnaire through exploratory and confirmatory factor analyses; and investigating construct validity by relating scores to relevant constructs such as digital health literacy, previous exposure to AI, or attitudes toward technology. Future studies should follow established methodological standards for evaluating the measurement properties of self-report instruments, such as those proposed in the Consensus-Based Standards for the Selection of Health Measurement Instruments and in classical psychometric frameworks for scale development [37]. Multiprofessional and multicenter studies will also be essential to examine the generalizability of these instruments across different health professions and educational settings, thereby extending the preliminary multiprofessional cognitive testing conducted in this study.

Beyond cross-sectional validation, future studies could also explore measurement invariance across disciplines and institutions and assess the responsiveness of these instruments to educational interventions introducing AI and digital health content into curricula. In the longer term, empirical data from these validation studies may support the development of a shortened, transdisciplinary questionnaire that captures core AI-related perceptions and readiness across health professions, suitable for routine use in curriculum design and longitudinal monitoring of AI literacy in health education.

Conclusions

In this study, we translated, culturally adapted, and linguistically validated into French 5 established English-language questionnaires originally designed to assess medical students' perceptions of AI in health care, and we adapted their wording for use among health care students in the French context. These adaptations fill an important gap in francophone contexts by providing standardized, conceptually equivalent instruments to document AI-related knowledge, attitudes, and intentions across health training programs. Once fully validated through psychometric studies in larger and more diverse samples, these questionnaires could be used in a robust manner for curriculum design, evaluation of educational interventions, and international comparisons of AI literacy in health education.

Funding

The study was supported by the French National Research Agency under the France 2030 program (grant 23-CMAS-0035) and by the Multidisciplinary Institute in Artificial intelligence Cluster (grant ANR-23-IACL-0006). This work also forms part of a broader translational and interdisciplinary research program, Santé Numérique Grenoble Intelligence Artificielle.

Data Availability

The data that support the findings of this study are available from the corresponding author SK upon reasonable request.

Authors' Contributions

SK contributed to conceptualization, methodology, software, validation, formal analysis, investigation, data curation, original draft writing, manuscript review and editing, supervision, funding acquisition, and project administration. CMT contributed to software, validation, formal analysis, investigation, data curation, original draft writing, and manuscript review and editing. NV contributed to conceptualization, methodology, original draft writing, manuscript review and editing, supervision, funding acquisition, and project administration.

Conflicts of Interest

None declared.

XSL•FO
RenderX

Multimedia Appendix 1
Translation and linguistic validation in French of the 5 questionnaires.
[DOCX File, 39 KB - mededu_v12i1e76572_app1.docx]

References

1. Sorrentino FS, Gardini L, Fontana L, et al. Novel approaches for early detection of retinal diseases using artificial intelligence. J Pers Med 2024 Jun 26;14(7):690. [doi: 10.3390/jpm14070690] [Medline: 39063944]
2. Chatzipanagiotou OP, Loukas C, Vailas M, et al. Artificial intelligence in hepatocellular carcinoma diagnosis: a comprehensive review of current literature. J Gastroenterol Hepatol 2024 Oct;39(10):1994-2005. [doi: 10.1111/jgh.16663] [Medline: 38923550]
3. Ouanes K, Farhah N. Effectiveness of artificial intelligence (AI) in clinical decision support systems and care delivery. J Med Syst 2024 Aug 12;48(1):74. [doi: 10.1007/s10916-024-02098-4] [Medline: 39133332]
4. Bilgin GB, Bilgin C, Burkett BJ, et al. Theranostics and artificial intelligence: new frontiers in personalized medicine. Theranostics 2024;14(6):2367-2378. [doi: 10.7150/thno.94788]
5. Geoghegan L, Scarborough A, Wormald JCR, et al. Automated conversational agents for post-intervention follow-up: a systematic review. BJS Open 2021 Jul 6;5(4):zrab070. [doi: 10.1093/bjsopen/zrab070] [Medline: 34323916]
6. Iyamu I, Xu AXT, Gómez-Ramírez O, et al. Defining digital public health and the role of digitization, digitalization, and digital transformation: scoping review. JMIR Public Health Surveill 2021 Nov 26;7(11):e30399. [doi: 10.2196/30399] [Medline: 34842555]
7. Khera R, Asnani AH, Krive J, et al. Artificial intelligence to enhance precision medicine in cardio-oncology: a scientific statement from the American Heart Association. Circ Genom Precis Med 2025 Apr;18(2):e000097. [doi: 10.1161/HCG.0000000000000097] [Medline: 39989357]
8. Steerling E, Siira E, Nilsen P, Svedberg P, Nygren J. Implementing AI in healthcare-the relevance of trust: a scoping review. Front Health Serv 2023;3:1211150. [doi: 10.3389/frhs.2023.1211150] [Medline: 37693234]
9. Cestonaro C, Delicati A, Marcante B, Caenazzo L, Tozzo P. Defining medical liability when artificial intelligence is applied on diagnostic algorithms: a systematic review. Front Med 2023 Nov 27;10:1305756. [doi: 10.3389/fmed.2023.1305756]
10. Retamero JA, Gulturk E, Bozkurt A, et al. Artificial intelligence helps pathologists increase diagnostic accuracy and efficiency in the detection of breast cancer lymph node metastases. Am J Surg Pathol 2024 Jul 1;48(7):846-854. [doi: 10.1097/PAS.0000000000002248] [Medline: 38809272]
11. Mohsin Khan M, Shah N, Shaikh N, Thabet A, Alrabayah T, Belkhair S. Towards secure and trusted AI in healthcare: a systematic review of emerging innovations and ethical challenges. Int J Med Inform 2025 Mar;195:105780. [doi: 10.1016/j.ijmedinf.2024.105780] [Medline: 39753062]
12. Naik N, Hameed BMZ, Shetty DK, et al. Legal and ethical consideration in artificial intelligence in healthcare: who takes responsibility? Front Surg 2022;9:862322. [doi: 10.3389/fsurg.2022.862322] [Medline: 35360424]
13. Ellahham S, Ellahham N, Simsekler MCE. Application of artificial intelligence in the health care safety context: opportunities and challenges. Am J Med Qual 2020;35(4):341-348. [doi: 10.1177/1062860619878515] [Medline: 31581790]
14. Díaz-Rodríguez N, Del Ser J, Coeckelbergh M, López de Prado M, Herrera-Viedma E, Herrera F. Connecting the dots in trustworthy artificial intelligence: from AI principles, ethics, and key requirements to responsible AI systems and regulation. Inf Fusion 2023 Nov 1;99:101896. [doi: 10.1016/j.inffus.2023.101896]
15. Habli I, Lawton T, Porter Z. Artificial intelligence in health care: accountability and safety. Bull World Health Organ 2020 Apr 1;98(4):251-256. [doi: 10.2471/BLT.19.237487] [Medline: 32284648]
16. Asan O, Bayrak AE, Choudhury A. Artificial intelligence and human trust in healthcare: focus on clinicians. J Med Internet Res 2020 Jun 19;22(6):e15154. [doi: 10.2196/15154] [Medline: 32558657]
17. Yakar D, Ongena YP, Kwee TC, Haan M. Do people favor artificial intelligence over physicians? A survey among the general population and their view on artificial intelligence in medicine. Value Health 2022 Mar;25(3):374-381. [doi: 10.1016/j.jval.2021.09.004] [Medline: 35227448]
18. Ng FYC, Thirunavukarasu AJ, Cheng H, et al. Artificial intelligence education: an evidence-based medicine approach for consumers, translators, and developers. Cell Rep Med 2023 Oct 17;4(10):101230. [doi: 10.1016/j.xcrm.2023.101230] [Medline: 37852174]
19. Schubert T, Oosterlinck T, Stevens RD, Maxwell PH, van der Schaar M. AI education for clinicians. eClinicalMedicine 2025 Jan;79:102968. [doi: 10.1016/j.eclinm.2024.102968] [Medline: 39720600]
20. Gazquez-Garcia J, Sánchez-Bocanegra CL, Sevillano JL. AI in the health sector: systematic review of key skills for future health professionals. JMIR Med Educ 2025 Feb 5;11:e58161. [doi: 10.2196/58161] [Medline: 39912237]
21. Order of November 10, 2022 relating to basic digital health training for health students. Légifrance. 2022. URL: https://www.legifrance.gouv.fr/eli/arrete/2022/11/10/SPRH2220195A/jo/texte [accessed 2025-03-13]
22. Machleid F, Kaczmarczyk R, Johann D, et al. Perceptions of digital health education among European medical students: mixed methods survey. J Med Internet Res 2020 Aug 14;22(8):e19827. [doi: 10.2196/19827] [Medline: 32667899]
23. Teng M, Singla R, Yau O, et al. Health care students' perspectives on artificial intelligence: countrywide survey in Canada. JMIR Med Educ 2022 Jan 31;8(1):e33390. [doi: 10.2196/33390] [Medline: 35099397]

XSL•FO
RenderX

24. Liu DS, Sawyer J, Luna A, et al. Perceptions of US medical students on artificial intelligence in medicine: mixed methods survey study. JMIR Med Educ 2022 Oct 21;8(4):e38325. [doi: 10.2196/38325] [Medline: 36269641]
25. Karaca O, Çalışkan SA, Demir K. Medical artificial intelligence readiness scale for medical students (MAIRS-MS) - development, validity and reliability study. BMC Med Educ 2021 Feb 18;21(1):112. [doi: 10.1186/s12909-021-02546-6] [Medline: 33602196]
26. Sit C, Srinivasan R, Amlani A, et al. Attitudes and perceptions of UK medical students towards artificial intelligence and radiology: a multicentre survey. Insights Imaging 2020 Feb 5;11(1):14. [doi: 10.1186/s13244-019-0830-7] [Medline: 32025951]
27. Boillat T, Nawaz FA, Rivas H. Readiness to embrace artificial intelligence among medical doctors and students: questionnaire-based study. JMIR Med Educ 2022 Apr 12;8(2):e34973. [doi: 10.2196/34973] [Medline: 35412463]
28. Li X, Jiang MYC, Jong MSY, Zhang X, Chai CS. Understanding medical students' perceptions of and behavioral intentions toward learning artificial intelligence: a survey study. Int J Environ Res Public Health 2022 Jul 18;19(14):8733. [doi: 10.3390/ijerph19148733] [Medline: 35886587]
29. Park CJ, Yi PH, Siegel EL. Medical student perspectives on the impact of artificial intelligence on the practice of medicine. Curr Probl Diagn Radiol 2021;50(5):614-619. [doi: 10.1067/j.cpradiol.2020.06.011] [Medline: 32680632]
30. Beaton DE, Bombardier C, Guillemin F, Ferraz MB. Guidelines for the process of cross-cultural adaptation of self-report measures. Spine (Phila Pa 1986) 2000 Dec;25(24):3186-3191. [doi: 10.1097/00007632-200012150-00014]
31. Wild D, Grove A, Martin M, et al. Principles of good practice for the translation and cultural adaptation process for patient-reported outcomes (PRO) measures: report of the ISPOR task force for translation and cultural adaptation. Value Health 2005;8(2):94-104. [doi: 10.1111/j.1524-4733.2005.04054.x] [Medline: 15804318]
32. Sousa VD, Rojjanasrirat W. Translation, adaptation and validation of instruments or scales for use in cross-cultural health care research: a clear and user-friendly guideline. J Eval Clin Pract 2011 Apr;17(2):268-274. [doi: 10.1111/j.1365-2753.2010.01434.x] [Medline: 20874835]
33. Epstein J, Santo RM, Guillemin F. A review of guidelines for cross-cultural adaptation of questionnaires could not bring out a consensus. J Clin Epidemiol 2015 Apr;68(4):435-441. [doi: 10.1016/j.jclinepi.2014.11.021] [Medline: 25698408]
34. Baumgartner M, Sauer C, Blagec K, Dorffner G. Digital health understanding and preparedness of medical students: a cross-sectional study. Med Educ Online 2022 Dec;27(1):2114851. [doi: 10.1080/10872981.2022.2114851] [Medline: 36036219]
35. Soulard J, Vaillant J, Baillet A, Vuillerme N. Translation and French linguistic validation of the Bath Ankylosing Spondylitis Functional Index and the Bath Ankylosing Spondylitis Global Score in patients with axial spondyloarthritis. Curr Rheumatol Rev 2023;19(4):449-454. [doi: 10.2174/1573397118666220829124234] [Medline: 36043728]
36. El Boudi I, Riant M, Bellier A, Vuillerme N. French versions of 4 English questionnaires on problematic smartphone use: cross-cultural linguistic translation and adaptation study. Interact J Med Res 2025 Feb 26;14:e53958. [doi: 10.2196/53958] [Medline: 40009839]
37. Mokkink LB, Terwee CB, Patrick DL, et al. The COSMIN checklist for assessing the methodological quality of studies on measurement properties of health status measurement instruments: an international Delphi study. Qual Life Res 2010 May;19(4):539-549. [doi: 10.1007/s11136-010-9606-8] [Medline: 20169472]

Abbreviations

AI: artificial intelligence
MAIRS-MS: Medical Artificial Intelligence Readiness Scale for Medical Students

Edited by J Gentges, T Leung; submitted 26.Apr.2025; peer-reviewed by DS Courvoisier, HF Many; revised version received 03.Dec.2025; accepted 04.Dec.2025; published 19.Mar.2026.

Please cite as:
Kotzki S, Massonnet Turner C, Vuillerme N
Questionnaires on Perceptions of Artificial Intelligence in Health Care Among Health Care Students: Cross-Cultural Translation Into French and Linguistic Validation
JMIR Med Educ 2026;12:e76572
URL: https://mededu.jmir.org/2026/1/e76572
doi:10.2196/76572

© Sylvain Kotzki, Calvin Massonnet Turner, Nicolas Vuillerme. Originally published in JMIR Medical Education (https://mededu.jmir.org), 19.Mar.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction

XSL•FO
RenderX

in any medium, provided the original work, first published in JMIR Medical Education, is properly cited. The complete bibliographic information, a link to the original publication on https://mededu.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Exploring Ambient Artificial Intelligence to Enhance Learning and Feedback During Operating Room-to-Intensive Care Unit Handoffs: Co-Design and Simulation Study

Laleh Jalilian[1], MD, MS; Federico Lorenzo Barra[2], MD; Tristan Grogan[3], MS; Jason Lee[1], MD; Achuta Kadambi[4], PhD

[1]Department of Anesthesiology and Perioperative Medicine, University of California, Los Angeles, 757 Westwood Plaza, Suite 3325, Los Angeles, CA, United States

[2]Department of Education - General Directorate, Azienda Ospedaliero-Universitaria "Maggiore della Carità", Novara, Italy

[3]Department of Medicine Statistics Core, David Geffen School of Medicine, University of California, Los Angeles, 1100 Glendon Avenue, 8th Floor,, Los Angeles, CA, United States

[4]Electrical and Computer Engineering, University of California, Los Angeles, 420 Westwood Plaza,, Los Angeles, CA, United States

Corresponding Author:
Laleh Jalilian, MD, MS
Department of Anesthesiology and Perioperative Medicine, University of California, Los Angeles, 757 Westwood Plaza, Suite 3325, Los Angeles, CA, United States

Abstract

Background: Operating room (OR)-to-intensive care unit (ICU) handoffs are among the most complex and high-risk communication events in perioperative care. Despite the implementation of structured checklists, trainees often receive limited feedback on their communication skills, and simulation-based education rarely provides objective data on communication performance and checklist adherence. This study explores how an ambient artificial intelligence (AI) handoff assistant used during simulation-based training of OR-to-ICU handoff discussions can enhance clinical communication training and AI literacy by mapping spoken handoff discussions to handoff checklist items, providing immediate feedback on checklist item omissions, and generating a structured handoff note that functions as a feedback-rich learning artifact.

Objective: This study aims to co-design and evaluate an ambient AI handoff assistant that transcribes spoken OR-to-ICU handoff communication, maps the discussion to handoff checklist items, generates a structured handoff note for educational review, and provides immediate feedback on handoff completeness during simulated OR-to-ICU handoff discussions in a low-fidelity educational setting.

Methods: A 2-phase mixed-methods study was conducted within the University of California, Los Angeles, Department of Anesthesiology and Perioperative Care (July-October 2025). Phase 1 comprised co-design interviews with 4 clinician educators to identify limitations of current handoff training and inform AI feature development. Phase 2 involved an error analysis, as well as evaluations of usability, workload, and educational impact, conducted through ten 60-minute simulation sessions with pairs of medical students and first-year residents. Quantitative measures included the Physician Task Load Index, System Usability Scale, and a postsimulation survey; qualitative data from co-design sessions and simulation debrief interviews were thematically analyzed.

Results: Educators highlighted inconsistent checklist use and the absence of objective feedback on learners' communication skills as key areas that could benefit from structured documentation of handoff discussions using AI. Error analysis of the ambient AI handoff assistant revealed a mean of 3.6 (SD 1.2) errors per note, with incorrect output being the most frequent error type. There was no statistically significant difference between the ambient AI handoff assistant and the paper checklist with respect to the Physician Task Load Index and System Usability Scale measures. Trainees valued real-time transcripts and structured handoff notes for reflection of communication practices, and exposure to AI documentation errors enhanced critical thinking and awareness of AI technology limitations.

Conclusions: The ambient AI handoff assistant mapped simulated handoff discussions to checklist items and generated a structured handoff note, facilitating reflection on team-based communication skills in handoff education. Imperfections in the AI's output encouraged critical appraisal of its capabilities and prompted discussion about automation complacency, suggesting that AI-assisted simulations can foster both communication and digital literacy skills essential for future AI-enabled clinical practice.

(JMIR Med Educ 2026;12:e85666) doi:10.2196/85666

XSL•FO
RenderX

KEYWORDS

simulation training; handoff communication; human-computer interaction; ambient artificial intelligence; prompt engineering

Introduction

Effective handoff and transitions of care discussions are essential for ensuring safe and high-quality patient care [1-4]. Operating room (OR)-to-intensive care unit (ICU) handoffs are particularly complex and high risk, requiring coordinated communication among surgical, anesthesia, and ICU teams under conditions of cognitive load, time pressure, and frequent interruptions [5-7]. Incomplete communication transfer during OR-to-ICU handoffs remains a preventable contributor to medical errors and adverse patient outcomes [6,8,9], and structured handoff checklists have been developed to standardize communication and improve information transfer [10-15]. Despite the introduction of checklists, adherence to checklists remains variable [16], and trainees often receive limited instruction or feedback on communication skills [17]. Handoff skills are frequently learned informally through observation, with variable role modeling, inconsistent feedback, and few opportunities for deliberate practice [18]. As a result, trainees may be unaware of omissions or deficiencies in their communication and may not learn important relational communication skills. Without objective assessment, gaps between knowledge and performance may persist, contributing to preventable harm. Simulation-based education offers a safe environment to practice handoffs, but most simulations rely on subjective observation. Learners may overestimate the completeness of their handoffs, and without objective records or data-driven feedback, meaningful self-assessment and improvement are limited [19].

Recent advances in ambient artificial intelligence (AI) present new opportunities to address this educational gap. Ambient AI systems use natural language processing and large language models (LLMs) to passively capture spoken communication and generate summarized documentation of the transcript. Applied to checklist-based handoffs, such systems are being explored for their potential to identify which checklist items are verbalized, map spoken communication to structured checklist elements [20], and provide feedback on the completeness of information exchange. This capability may enhance handoff education from subjective assessment to measurable performance analytics, enabling iterative practice and self-reflection. Despite growing interest in AI applications for documentation and workflow support, no prior studies have examined the use of ambient AI for enhancing handoff communication skills to medical trainees [21,22]. This gap is particularly striking given that standardized handoff checklists already provide structured frameworks that align well with LLM-based mapping through prompt engineering, and simulation-based handoff training is widely implemented yet lacks robust, objective assessment tools [23] of verbal discussions [24].

To address this need, we developed and evaluated an ambient AI-assisted handoff simulation platform designed to support learning and feedback during OR-to-ICU handoff educational sessions. The system listens to trainee handoff discussions, maps verbalized content to checklist items, and generates structured documentation identifying which items were omitted. The objectives of this study were to (1) identify and co-design features of an ambient AI handoff assistant for OR-to-ICU handoff education through needs assessment and interviews with clinician educators, focusing on limitations of current handoff training and design requirements for AI-mediated feedback, (2) assess the performance and accuracy of the LLM-based checklist validation in identifying discussed and omitted handoff items, and (3) evaluate the usability, workload, and learning impact of the AI-assisted handoff simulation platform during low-fidelity sessions, while exploring learner perceptions of the AI-generated feedback, specifically its influence on communication, reflection, and perceived educational and clinical value.

Methods

Overview

A 2-phase, mixed-methods study was conducted within the University of California, Los Angeles (UCLA), Department of Anesthesiology and Perioperative Care (July-October 2025). The COREQ (Consolidated Criteria for Reporting Qualitative Research) [25] (Checklist 1) framework was used to guide research reporting.

Study Design

We used a human-centered, co-design methodology to develop an ambient AI handoff assistant. Co-design was selected for its suitability in simulation-based health care research, as it enables clinicians to surface workflow challenges and limitations, shape system features, and envision technology integration in high-stakes team communication [26,27]. The study was conducted in two phases: (1) needs assessment and co-design interviews with OR-to-ICU quality improvement leaders, aiming to (i) identify limitations of the current handoff educational modules, (ii) elicit expectations and perceptions of automated handoff discussion capture using an ambient AI handoff assistant, and (iii) prioritize design features for the ambient AI handoff assistant for handoff education; and (2) low-fidelity simulation sessions to evaluate the educational use of the ambient AI handoff assistant, with mixed-methods assessment of experience, acceptance, usability, and workload.

Participants and Setting

For the Needs Assessment and Co-Design workshop, 4 clinician educators participated in semistructured 45-minute interviews (remote or in person). Their years of professional experience ranged from 9 to 24, with a mean of 16 (SD 4.9) years. The decision to interview educators individually was dictated by their availability to meet with the study team. In the simulation phase, 10 simulation sessions were conducted in a low-fidelity setting. Each session involved 2 medical students or first-year interns who had no prior exposure to the OR-to-ICU handoff process at UCLA, who played the roles of surgery and anesthesia, and played the role of the ICU attending. New participants were recruited for each simulation session to

XSL•FO
RenderX

minimize bias from repeated exposure. Demographic breakdown of simulation participants is listed in Table 1.

Table . Overview of simulation study participants (n=20).

Participant type	Number of participants
What is your current training level?	
Medical student	7
Resident	13
Sex	
Male	7
Female	10
Nonbinary	2
Prefer not to answer	1
Do you have experience using an ambient AI[a] scribe? (An AI scribe refers to a tool that automatically generates documentation from ambient audio or speech input during clinical encounters, such as DAX Copilot or Nabla.)	
Yes	2
No	18
Do you have experience using generative AI technologies? (Examples include ChatGPT, Claude, Med-PaLM, Microsoft Copilot, or other tools that use artificial intelligence to generate content, summarize information, or support documentation.)	
No, I have never used any AI-generated technologies.	1
Yes, I've tried one briefly out of curiosity.	3
Yes, I've used them occasionally for general tasks (eg, answering questions, drafting emails).	9
Yes, I've used them occasionally for clinical or work-related tasks (eg, summarizing notes, generating documentation).	1
Yes, I use them frequently and consider them an integral part of my workflow.	6

[a]AI: artificial intelligence.

Needs Assessment and Co-Design Interviews

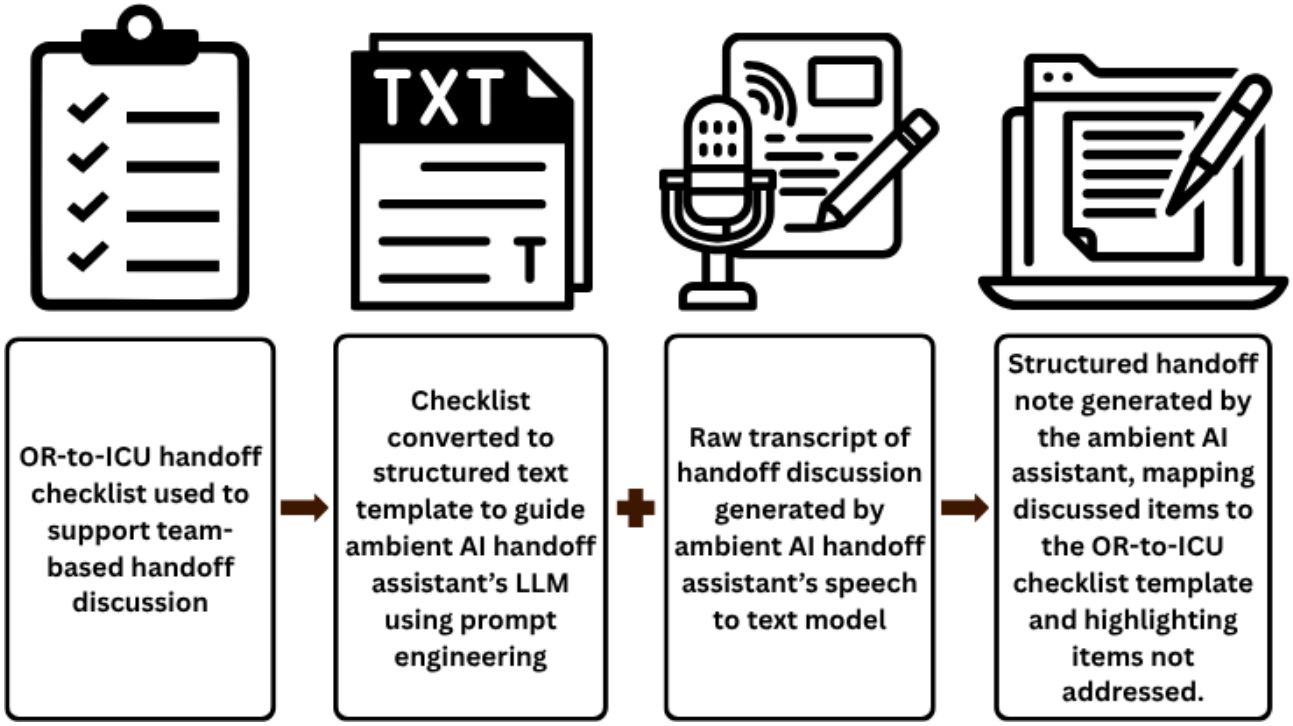
The primary study author interviewed 4 clinician educators in a 45-minute semistructured interview fashion using the Clinician Educator Interview guide in Multimedia Appendix 1 after obtaining verbal consent. The interviewer was a clinician trained in the use of qualitative research techniques. Given the interviewer's role as a faculty peer, reflexivity was emphasized to reduce potential bias. Interviews began with reflections on recent OR-to-ICU handoffs and handoff education to ground discussions in real clinical practice. Clinician educators identified limitations of current handoff educational methods and potential roles for ambient AI in encouraging educational discussion on checklist completeness and communication quality. The interviewer then introduced a design probe, an ambient AI handoff assistant to be used during handoff education simulation sessions with junior trainees. A functional prototype was used to ground discussions in concrete interactions. Using whiteboards/Miro and the ambient AI handoff assistant prototype, participants sketched desired AI functions (eg, transcription verification, structured note review, and interaction styles) and collaboratively prototyped ideal AI-supported simulation workflows. Transcripts were member-checked and coded in Atlas.ti (ATLAS.ti Scientific Software Development GmbH) using a grounded theory approach until thematic saturation was reached.

Ambient AI Handoff Assistant Technology Development

Based on insights from the needs assessment and co-design interviews, we developed an ambient AI handoff assistant to support both clinician educators and learners during OR-to-ICU handoff simulation sessions. The assistant digitally displays the UCLA OR-to-ICU Handoff Checklist, provides real-time visual transcription of the clinical discussion, and automatically generates a structured handoff note where the discussion is mapped to individual checklist items for learner review. It also highlights checklist items that were not discussed, helping educators and learners identify omissions and communication gaps. A commercial ambient AI scribe was not used because existing systems lack the capability to map spoken handoff dialogue directly to a structured handoff checklist for the purpose of assessing checklist adherence. Meeting this requirement necessitated custom prompt engineering of an LLM aligned specifically to UCLA's OR-to-ICU handoff checklist. Figure 1 illustrates the technical framework used to map handoff discussion content to corresponding checklist items.

XSL•FO
RenderX

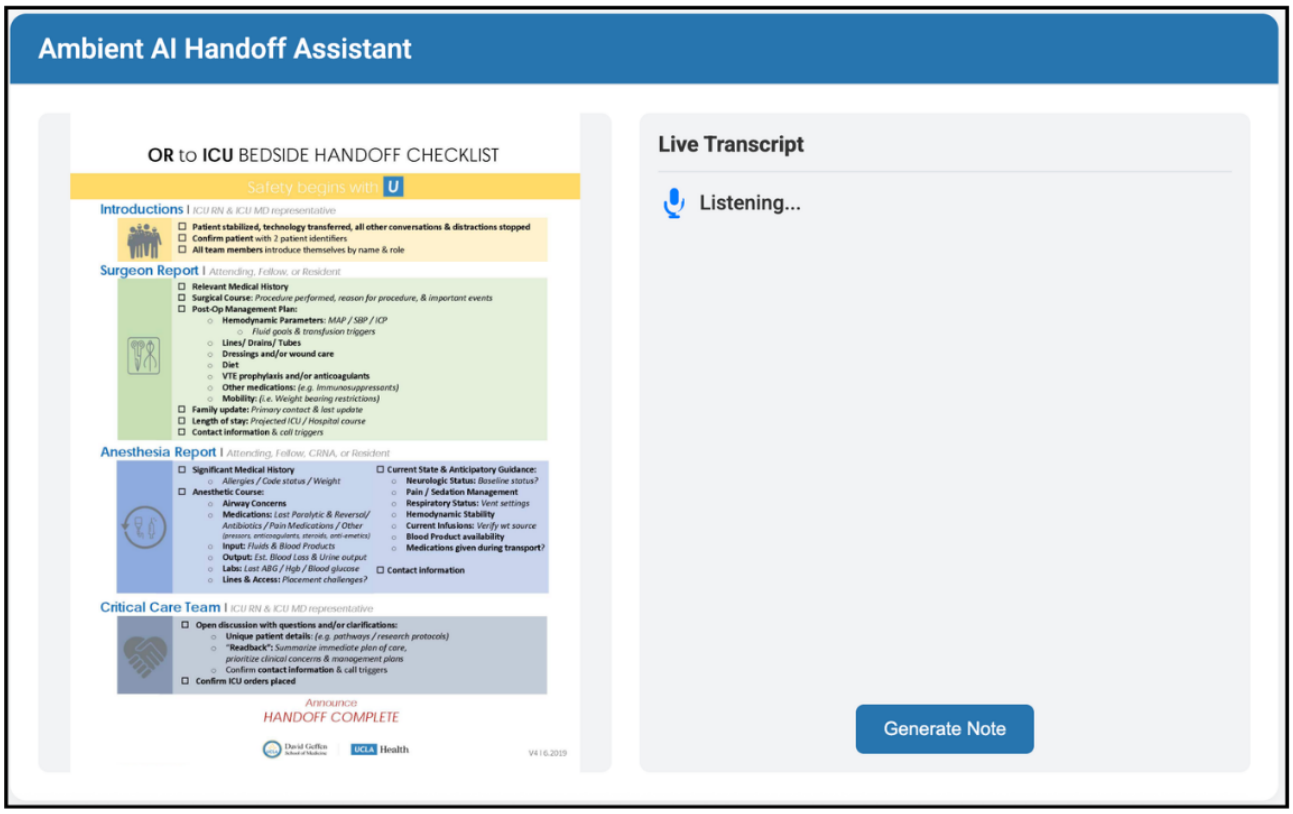
Figure 1. Standardized handoff checklists, such as the University of California Los Angeles operating room-to-intensive care unit Handoff Checklist, can be converted into structured text templates to guide the ambient artificial intelligence handoff assistant in generating documentation aligned with the desired format. Figure adapted from Jalilian et al [28]. AI: artificial intelligence; ICU: intensive care unit; LLM: large language model; OR: operating room.



Google AI Studio was used to develop the ambient AI handoff assistant, a single-page application built with HTML5, CSS3, and TypeScript, running entirely within the client's web browser to ensure low latency and data privacy. Real-time audio capture and transcription are handled by the Gemini 2.5 Flash Live application programming interface (API), configured for continuous listening to capture the full clinical dialogue. The captured transcript is processed by the Gemini 2.5 Flash LLM via the @google/genai SDK. No fine-tuning of the model was performed; instead, a zero-shot prompt engineering approach was used. The prompt requirements and handoff template were specified by clinical domain experts to align with UCLA's OR-to-ICU handoff checklist and are listed in Multimedia Appendix 2. The prompt uses explicit "Chain of Thought" instructions, requiring the model to extract factual data, resolve conflicting information by prioritizing the most recent statement, and strictly mark missing fields as "N/A." To ensure structured and reliable data extraction, the system leverages the model's JSON mode with a strict responseSchema, forcing the output into a predictable JSON object containing a formatted OR-to-ICU handoff note. The organization of transcription content into the final checklist structure and the identification of items labeled "not discussed" are performed via deterministic post-processing logic in TypeScript, rather than a separate predictive model. This logic parses the AI-generated JSON, compares it against a master checklist of 38 clinical items, and flags any item as 'missed' if the value is empty or remains "N/A." The application's core logic is managed by a finite state machine that guides the user through distinct phases: listening, processing, verifying, and summarizing. Auditory feedback and status prompts are delivered via the Gemini 2.5 Flash Live API's SpeechSynthesis interface, with a preselected "en-GB" female voice to ensure a consistent user experience. Screenshots of the different interface views of the ambient AI Handoff Assistant are visible in Figure 2 and Figure 3.

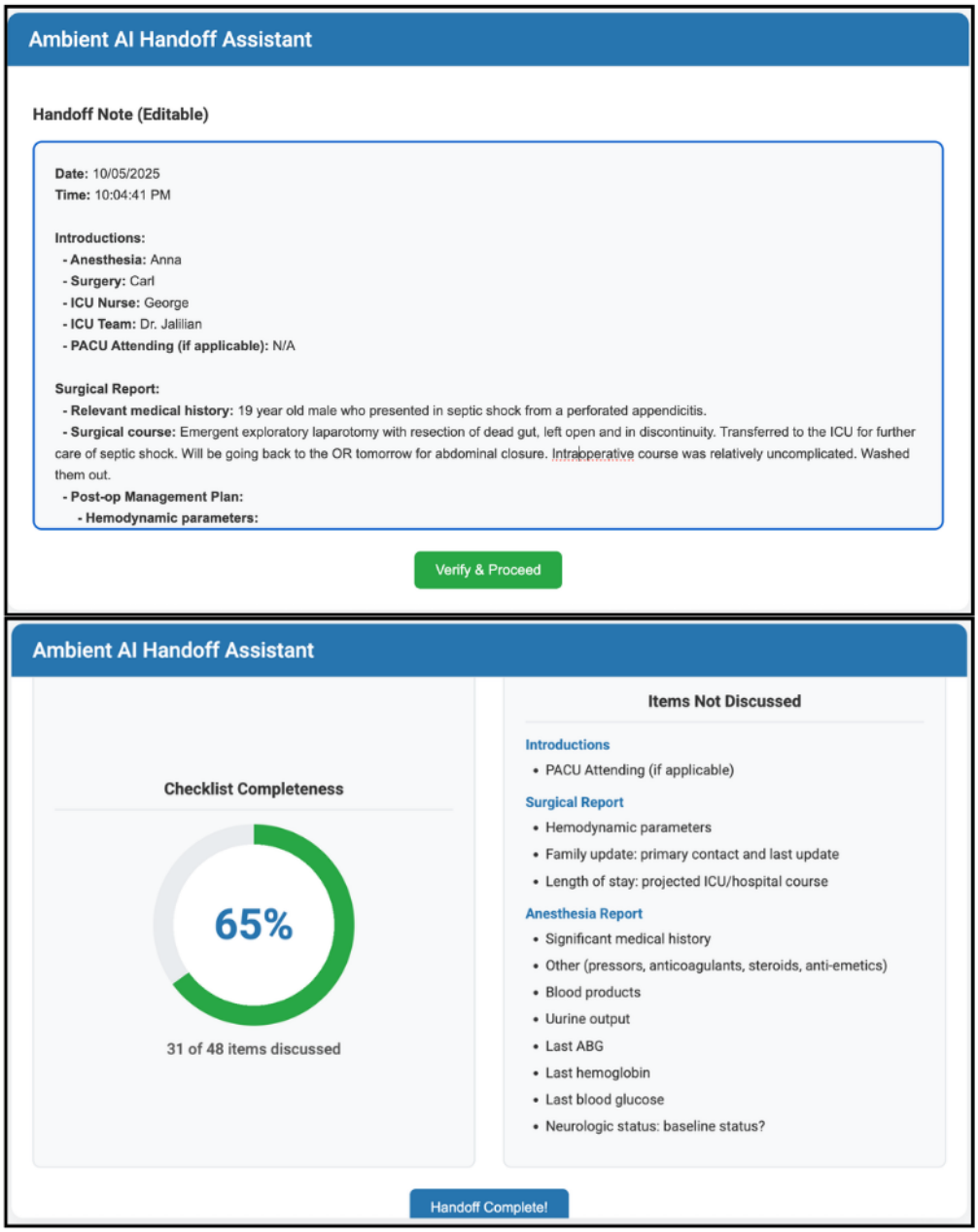
XSL•FO
RenderX

Figure 2. Screenshot of the ambient artificial intelligence handoff assistant pairing digital checklist display with visualization of real-time transcription. The University of California Los Angeles operating room-to-intensive care unit handoff checklist is digitally displayed as a visual cognitive aid for learners, who can simultaneously view a real-time transcription of the discussion in the right-hand panel. AI: artificial intelligence; ICU: intensive care unit; OR: operating room.



XSL•FO
RenderX

Figure 3. Screenshot of the ambient artificial intelligence handoff assistant as a supportive tool for checklist omission detection. In the top panel, handoff participants view an editable note organized in a checklist-based format, in which discussed content is mapped to its corresponding checklist items, facilitating post-hoc identification of omissions compared with a traditional narrative format. In the bottom panel, items flagged by the artificial intelligence Handoff Assistant as "not discussed" are displayed at the end of the session, providing explicit automated feedback to further support gap identification. Both the checklist-based note in the top panel and the AI-flagged omissions in the bottom panel operate after the discussion has concluded, serving complementary roles by enabling structured human review while also providing explicit, automated feedback to reinforce omission identification. AI: artificial intelligence; ICU: intensive care unit; OR: operating room.



Error Analysis of the AI Handoff Assistant

Five simulated OR-to-ICU handoff scenarios were developed using a previously established agentic framework [29] to evaluate the ambient AI handoff assistant in generating handoff notes and displaying undiscussed items. The scenarios represented a range of surgical cases commonly encountered in the ICU. Two clinician educators (LJ and JL) simulated the surgeon and anesthesiologist roles by reading the scenarios aloud while the ambient AI handoff assistant operated in the background. After each session, the clinician educators reviewed the AI-generated handoff note for errors. Each identified error was categorized into one of the following error types: omission, irrelevant or misplaced text, wrong output, or addition. This categorization was coded independently by 2 clinician educators (LJ and JL), and disagreements were discussed until consensus was reached.

Simulation Testing of the Ambient AI Handoff Assistant

Ten 60-minute simulation sessions in a nonclinical, low-fidelity setting were conducted with pairs of medical students or first-year resident physicians who played the roles of anesthesiology and surgery team members during an OR-to-ICU

XSL•FO
RenderX

handoff. Participants were recruited through voluntary email invitations and had no academic or evaluative relationship with the study faculty. Demographic characteristics are presented in Table 1. Each simulation was facilitated by faculty experts in handoff communication who held no supervisory role over participants. Reflexivity was emphasized to mitigate potential bias. Two distinct simulation scenarios (A and B) were generated from a previously developed agentic workflow [29]. The study used a within-team crossover design in which each dyad completed two simulated handoffs: (1) Control: a standard verbal handoff using the institutional paper checklist, and (2) Intervention: a handoff supported by the ambient AI handoff assistant. Each team served as its own control to reduce variability in communication skills, teamwork, and experience. Condition order (AI-first vs. control-first) and case scenario (A vs B) were randomly assigned and counterbalanced to control for order and case effects. Following both simulations, participants completed a postsimulation survey and a semistructured debriefing interview (Multimedia Appendix 3). Debriefing interviews were transcribed and returned to participants for member checking. All participants provided verbal informed consent.

Statistical Analysis

For the needs assessment and co-design phase, inductive content analysis was used to synthesize design insights. Two researchers (LJ and JL) independently coded transcripts, sketches, and field notes to identify (1) limitations in current handoff education and (2) desired ambient AI handoff assistant features. Coding discrepancies were resolved by consensus, and themes were consolidated into preliminary design requirements for the handoff AI assistant. The Educator Co-design codebook is listed as Codebook A in Multimedia Appendix 4.

Quantitative data from the 4-item Physician Task Load Index (PTL) [30] and System Usability Scale (SUS) were analyzed descriptively. Each participant completed both the control (paper checklist) and intervention (ambient AI assistant) conditions in a within-subjects crossover design. For each measure, paired-samples analyses were conducted to compare participant scores between conditions. The normality of paired differences was evaluated using the Shapiro–Wilk test. When assumptions of normality were satisfied, paired *t*-tests were used to compare mean scores between the control and AI conditions. Results are reported as means (standard deviations) for normally distributed data. All estimates include 95% CIs for the mean differences. Effect sizes for paired 2-tailed *t* tests were expressed as Cohen *d* (standardized mean difference for paired data). Although 2 participants worked together in each of 10 simulation sessions, the paired 2-tailed *t* test was selected as the primary analytic approach because each participant served as their own control, and between-team clustering was minimal. All analyses were conducted using Python (version 3.11; Python Software Foundation), with statistical significance set at $P<.05$ (2-tailed).

For the simulation debriefings, qualitative data from debriefings were coded independently by two researchers (LJ and JL) using Atlas.ti, with themes refined through iterative consensus. Structured observations were analyzed to capture communication dynamics and were triangulated with interview findings to assess human-technology interaction. The Learner Simulation Codebook is listed as Codebook B in Multimedia Appendix 4.

Ethical Considerations

This study was approved by the UCLA Institutional Review Board #25 - 1998. Verbal informed consent was obtained from all participants before data collection. Participants were informed of the study purpose, procedures, potential risks and benefits, and their right to decline participation or withdraw at any time. Measures were taken to protect participant privacy and maintain confidentiality, including deidentification of data and secure storage of study materials.

Results

Needs Assessment and Co-Design Thematic Analysis

Themes from the co-design interviews with educators are listed in Table 2. Educators described ongoing challenges in teaching closed-loop communication and ensuring complete information transfer during handoffs. The OR-to-ICU checklist was viewed as a valuable cognitive aid for structuring handoff discussions across surgical, anesthetic, and ICU domains, yet there was recognition that its real-world use was inconsistent and that trainees could inadvertently skip over checklist items. Educators emphasized the need for greater situational awareness and systematic progression through each checklist item during real-time communication.

XSL•FO
RenderX

Table . Thematic analysis of needs assessment and co-design sessions on ambient AI handoff assistants for operating room-to-intensive care unit handoff education. Representative quotes are drawn from clinician educator interviews conducted during the co-design phase (n=4).

Theme	Key insight	Representative quote
Promoting professional communication	Awareness of AI[a] documentation may encourage more mindful, deliberate, and professional communication among trainees.	*"I'd be more careful of what I'm saying."*
Enhancing reflective learning	AI-generated transcripts and note summaries may support self-assessment and reflection, helping learners review and improve their communication.	*"Having a transcript to go back and refer to adds another level of feedback."*
Exposure and readiness for AI	Exposure in simulation may allow trainees to experience AI-assisted documentation and assess its accuracy and use in realistic settings.	*"Ambient AI scribes might be one of the first AI technologies that perioperative doctors might be exposed to. Would be interesting to see how accurate the handoff assistant is in documenting the discussion."*
Calibration of trust in AI output for learners	EAmbient AI Handoff Assistant may be more inaccurate than commercial ambient AI scribes, but exposure to AI transcription inaccuracies may enable trust calibration in AI.	*"[Learners] need to see the AI make mistakes. It doesn't matter how accurate [the ambient AI handoff assistant] is, because exposure to the errors is instructive in and of itself."*
Reducing documentation burden in simulation environments	Automating note-taking and checklist reconciliation may allow educators to focus on observation and higher-level feedback rather than manual documentation during teaching.	*"It'd be nice to focus on the students instead of taking notes."*
Ethical and privacy considerations in training	Educators emphasize the ethical obligation to protect learners' privacy, psychological safety, and autonomy when using AI in training.	*"If they know they're being recorded, they might not ask questions they think will sound dumb."*
Trust and validation of AI output for educators	Educators are uncertain of the accuracy of AI-generated content and need experience to understand its educational use.	*"If it can't tell which handoff checklist items the trainee did or did not mention, I wouldn't rely on it. It might not be the point of a tool like this though."*
Design requirements for ambient AI handoff assistant	Digital checklist display, real-time display of transcription, mapping transcript content to checklist categories in the format of a structured checklist-aligned handoff note, and explicitly labeling checklist items as "discussed" or "not discussed" based on transcript evidence would support clinician educators.	*"I don't want to bring a paper checklist to one of these sessions, so we should digitally display it as part of the module."*

[a]AI: artificial intelligence.

> *We've taught the residents to use the checklist for years, but in the heat of the moment, they skip parts of the checklist.* [Clinician Educator]

Educators felt that many trainees struggled with the interpersonal dimensions of team communication, such as speaking hesitantly, not introducing themselves, or not making eye contact with other members of the handoff team. Educators highlighted that in high-stakes transitions of care, clarity, confidence, and engagement are essential safety and communication behaviors that promote shared understanding. Developing these expressive and relational communication skills was seen as critical yet under-taught.

> *We don't teach the residents how to speak in a handoff. We focus on what checklist items need to be discussed, not so much on delivery, tone, or engagement with the team.* [Clinician Educator]

Traditional feedback was described as subjective and infrequent, offering limited improvement, especially if an attending physician does not accompany a trainee to the ICU for a handoff.

> *Without a record of what was actually said, feedback is just opinion. Everyone remembers the handoff a little bit differently.* [Clinician Educator]

Educators perceived AI-generated documentation as a promising way of creating an objective record of what was actually communicated during handoffs. In developing the ambient AI handoff assistant, they noted that AI-assisted documentation technologies could more concretely identify communication omissions and translate them into specific, actionable learning points, while maintaining a safe environment for learners to critically evaluate the limitations of AI.

> *My understanding is that AI output isn't perfect, but I guess that would be part of the learning. Trainees would see how the AI interprets them. Maybe it might make them question it or think critically about accuracy.* [Clinician Educator]

Educators felt that the design requirements of the ambient AI handoff assistant should include digital display of the checklist, real-time visualization of transcription, and an AI-generated handoff note where each checklist item is populated with what

XSL•FO
RenderX

the learners discussed to enable educator validation of checklist completeness. Beyond the educational module, they suggested that commercially available ambient AI scribe technologies could potentially enhance handoff processes in the future through the generation of a handoff note for the receiving ICU team to use as a reference of the discussion. Introducing ambient AI first in a simulated setting would allow educators and Handoff QI leaders to evaluate their use, feasibility, and user acceptance in simulated handoff workflows before considering any potential real-world use.

Error Analysis of the Ambient AI Handoff Assistant

There were 18 errors noted in 5 of 5 draft notes (average of 3.6 errors per draft note). Error type by product is detailed in Table 2, with "wrong output" being the most frequent error across products.

Physician Task Load and System Usability Score

The paired-samples 2-tailed *t* test showed no significant difference in perceived task load between the paper checklist (mean 12.9, SD 16) and the AI handoff assistant (mean 15.8, SD 8.3), a difference of 2.9 (95% CI −5.9 to 11.7; *P*=.50). This suggests that the AI tool did not meaningfully change participants' subjective workload during simulated handoffs. Participants rated the AI assistant and paper checklist as similarly usable, with mean SUS scores of 81.4 (SD 11.3) for the paper checklist and 74.3 (SD 13) for the AI assistant, a difference of 2.9 (95% CI −0.6 to 14.8; *P*=.06). Although the AI assistant trended slightly lower in usability ratings, this difference was not quite statistically significant (Tables 3 and 4).

Table . Physician task load and System Usability Scale (n=20).

Metric	Paper checklist, mean (SD)	Ambient AI[c] handoff assistant, mean (SD)	*t* test (df)	*P* value	Cohen *d*
PTL[a]	12.9 (16)	15.8 (8.3)	0.68 (19)	.50	0.15
SUS[b]	81.4 (11.3)	74.3 (13.0)	−1.93 (19)	.06	−0.43

[a]PTL: Physician Task Load Index.

[b]SUS: System Usability Scale.

[c]AI: artificial intelligence.

Table . Postsimulation survey results[a] (n=20).

Postsimulation survey results	Mean (SD)	95% CI
The AI[b]-generated handoff note and missed-items list helped me recognize strengths and weaknesses in my handoff communication.	3.9 (0.9)	3.5-4.3
The AI assistant's feedback complemented faculty debriefing and peer discussion.	4.3 (0.6)	4-4.6
I was able to distinguish between accurate and inaccurate information in the AI-generated documentation.	3.8 (1)	3.2-4.3
I exercised appropriate caution in interpreting and relying on AI-generated documentation.	3.9 (0.9)	3.5-4.3
Reviewing the AI-generated note and missed-items list increased my confidence in performing OR[c]-to-ICU[d] handoffs.	4 (1)	3.5-4.4
The AI feedback helped me feel more prepared to apply structured handoff practices in clinical settings.	4.2 (0.9)	3.8-4.6

[a]Response options of strongly disagree (1), disagree (2), neither agree nor disagree (3), agree (4), and strongly agree (5)

[b]AI: artificial intelligence.

[c]OR: operating room.

[d]ICU: intensive care unit.

A post-hoc power analysis was performed to estimate the minimum detectable effect size under our study's sample size of 20 paired observations. Assuming a 2-sided α=.05, this sample provides 80% power to detect a standardized mean difference of 0.66. In contrast, the observed effect sizes in our data were 0.43 for the SUS outcome and –0.15 for the task load outcome, which correspond to smaller magnitudes than this detectable threshold. To achieve 80% power to detect an effect size of 0.43, a sample size of approximately 45 would be required, whereas detecting an effect as small as 0.15 would require roughly 351 participants. These estimates may guide

XSL•FO
RenderX

the design of future studies with adequate power to detect more modest effects.

Simulation Debrief Thematic Analysis

Themes from debrief interviews with trainees are captured in Table 5. Trainees emphasized the educational value of an ambient AI handoff assistant in simulated handoff training. AI-generated transcripts and structured notes were felt to provide objective feedback, allowing learners to identify omissions, reflect on their communication, and discuss the limits of AI. These outputs transformed team discussions into concrete learning records that supported debriefing and self-assessment, addressing a gap in handoff education often reliant on subjective recall (Table 6).

Table . Frequency counts, percentages, definitions, and examples of error types in the artificial intelligence-generated operating room-to-intensive care unit handoff note (n=5).

Error type	Ambient AI[a] handoff assistant, n (%)	Definition	Example
Omission	5 (27.7)	AI assistant leaves out key information that was discussed	*"No mention of estimated blood loss when it was discussed"*
Misplaced text	3 (16.7)	AI assistant output is technically correct but not in appropriate checklist item	*"Estimated blood loss was discussed but placed in the wrong checklist item"*
Wrong output	7 (38.9)	AI assistant provides an incorrect response	*"Phenylephrine was the spoken vasopressor medication but ephedrine was documented instead"*
Addition	3 (16.7)	AI assistant adds inappropriate or irrelevant information	*"Estimated blood loss was described only as 'minimal,' but the AI output documents it as 'minimal (~50 mL).'*
Total	18 (100)		

[a]AI: artificial intelligence.

XSL•FO
RenderX

Table . Thematic analysis of simulation debrief. Representative quotes are drawn from learner participants conducted during the simulation phase (n=20).

Theme	Key insight	Representative quote
Exposure and readiness for AI[a]	Learners viewed the simulation as a valuable first exposure to AI tools in clinical communication, where they could evaluate AI capabilities and accuracy.	*"This was my first time using something like this. I was surprised at how well it actually captured what I said."*
Reflective and professional learning	Reviewing AI-generated transcripts encouraged self-assessment and helped learners recognize patterns in their communication and teamwork.	*"Seeing the transcript was good. I didn't realize I skipped over things when I talked through a case."*
Feedback integration and educational use	Learners discuss how they or faculty could use AI transcripts to guide debriefing, identify missed information, or support feedback loops.	*"I liked the feature that highlighted what we didn't discuss."*
Communication awareness	Awareness of being recorded made learners more deliberate in how they presented information during handoffs.	*"Knowing it was recording made me think about what I was saying."*
Peer dynamics and team communication	Learners sometimes adjust how they communicate with teammates when AI is present.	*"We were more formal because we knew the AI was picking up everything."*
Trust, accuracy, and error sensitivity	Learners noted AI's limitations and inaccuracies in speech-to-text transcription and note generation, emphasizing the need for human verification.	*"It made mistakes! It was interesting to see how some of the medical language was mistakenly transcribed, like "mcgs" being transcribed as "Mike's." I think you will need to check it."*
Translation to real-world clinical practice	Learners connected the simulation to future practice, envisioning how AI scribes could improve communication and documentation in real OR[b]-to-ICU[c] handoff workflows.	*"If this worked in the real ICU, the generated note would keep everyone on the same page and be really helpful to refer back to."*
Emotional and psychological responses to AI	Learners sometimes felt self-conscious knowing AI is "listening."	*"At first it felt weird knowing it was recording everything. I got used to it though."*
Ethical and privacy awareness in training	Learners reflected about consent, data storage, and fairness in educational AI use.	*"I'd want to know who can access the data from these sessions."*

[a]AI: artificial intelligence.

[b]OR: operating room.

[c]ICU: intensive care unit.

> *Seeing the AI transcript right after the handoff was great. It was interesting to see what I actually said versus what I thought I said.* [Learner]

Exposure to the ambient AI handoff assistant's errors also promoted critical thinking and awareness of the technology's limitations.

> *Seeing the mistakes the AI made in its transcription was helpful. We can't just go on auto-pilot with it.* [Learner]

Trainees further envisioned how AI scribe technology might impact real-world OR-to-ICU workflows to document complex handoffs, enhance continuity, and strengthen accountability.

> *If it could document the handoff automatically, it would help with continuity, especially for the ICU team coming on shift.* [Learner]

They did express reservations regarding data governance of ambient AI technologies.

> *If AI gets implemented in the real-world, it should be clear what data is actually being stored.* [Learner]

Discussion

Principal Findings

This study designed and evaluated an ambient AI handoff assistant for documenting OR-to-ICU handoff discussions within a medical education context. Because OR-to-ICU transitions are among the most complex and high-risk discussions in perioperative care, they provided an ideal setting for exploring how ambient AI can support medical trainees in promoting awareness of their communication and checklist completion during high-stakes discussions. Introducing the ambient AI handoff assistant to transcribe participants' discussions appeared to enhance perceptions of learning and educational value, as reflected in both survey feedback and debrief interviews. Further, PTL scores did not differ significantly between the AI scribe and the traditional checklist. This may suggest that verbally delivering a complex OR-to-ICU handoff is inherently challenging for inexperienced trainees, regardless of whether they are supported by a paper checklist or the AI-based handoff assistant. SUS scores did not differ significantly between the paper checklist and the AI-based handoff assistant, suggesting that digitally displayed checklists may be acceptable to users as a cognitive aid. Participants reported that the AI-generated

XSL•FO
RenderX

note functioned as a valuable record of the discussion, helping them verify checklist completion, promoting reflective conversations about communication, and enabling them to better understand the limitations of AI-assisted tools. These observations align with prior work suggesting that documentation of verbal discussions can enhance reflective learning and promote deliberate practice in team communication [31,32]. These findings also align with prior research showing that new technologies can be perceived as equally usable as established tools once users recognize their educational and workflow benefits [33].

Because learners were medical students and first-year interns with limited semantic understanding or familiarity with complex anesthesia and surgical terminology, they adhered closely to the scripted simulation scenarios. As a result, their handoffs were generally comprehensive, creating a ceiling effect on traditional performance measures such as completeness and limiting our ability to assess whether the ambient AI handoff assistant improved completeness. However, qualitative feedback indicated that learners valued the opportunity to practice structured team-based communication in a simulated setting and to engage with speech-to-text technology. This experience prompted broader discussions about communication strategies in high-stakes scenarios, focusing on the importance of closed-loop communication, clear and confident speech, and shared situational awareness to ensure effective information transfer. Further, interacting with imperfect AI output proved instructive, as learners observed transcription and checklist-mapping errors firsthand. In one instance, a learner with a foreign accent observed that her articulation of medical terminology was frequently inaccurately transcribed by the ambient AI handoff assistant, prompting a discussion about biases in training data and the limitations of speech-to-text systems. Trainees discussed the importance of vigilance of stereotyped assumptions and biased outputs with AI-generated content. Further, the limitations of the ambient AI handoff assistant in recognizing complex perioperative medical terminology also led to transcription and checklist mapping errors. Learners felt these inaccuracies became a productive feature of the simulation, enabling them to calibrate their trust in ambient AI outputs and also to understand how these technologies are assistive but still require human oversight. This led to discussion on automation complacency and its implications for patient safety with the use of AI technologies, prompting learner participants to reflect on the importance of verifying AI outputs [34,35] in future AI-assisted clinical environments. This fostered critical appraisal skills and transformed the AI assistant from a passive documentation tool into an active learning aid that promoted vigilance, AI literacy, and reflective communication. Future iterations of AI-assisted handoff training could vary the fidelity of AI output to examine how learners calibrate trust and maintain situational awareness when interacting with interactive systems.

Although this project primarily examined how ambient AI could support handoff education through automated documentation and verbal checklist workflow capture, it also demonstrated that high-stakes, multi-speaker handoff discussions can be structured into clinically and pedagogically meaningful documentation. The ambient AI handoff assistant was designed to map spoken communication to predefined checklist elements, thereby supporting the Educator during handoff simulation sessions and allowing greater focus on the Learners rather than on checklist reconciliation. During the error analysis phase, we assessed the ambient AI handoff assistant's ability to map what was discussed to its corresponding checklist item. Most errors were attributed to difficulties in recognizing medical terminology and variability in speaker articulation. This finding was expected, given the inherent limitations of the Gemini 2.5 Flash Live API on which the system relies. Further, there were more transcription errors with the learners in simulation compared with the prior error analysis, likely reflecting the greater acoustic variability of real learning environments. Together, these factors contributed to downstream checklist mapping errors that resulted in imperfect checklist completeness assessment functionality of the ambient AI handoff assistant. Importantly, the goal of developing this system was to create an educational platform that would enable some transcription feedback of a high-stakes discussion, exposure to ambient AI, and assessment of whether AI could map spoken team communication to handoff checklist items as a feedback tool. Both educators and learners reflected on whether ambient AI technologies could play a role in real-world handoffs and used the simulation experience as a basis for discussing whether commercial ambient AI scribes might provide value in real-world OR-to-ICU handoff clinical workflows.

Simulation as Essential Infrastructure for AI Evaluation in Health Care

Our work highlights the importance of simulation-based testing as essential infrastructure for preparing the medical workforce to engage safely with AI technologies [36]. Simulation provided a safe environment for experimentation, where medical trainees experienced an AI-assisted handoff workflow firsthand. It also allowed them to gain exposure to the limitations of AI technologies without exposing patients to risk. The findings also informed educators, prompting consideration of how AI assistance could be integrated into other educational modules. Simulation made it possible to observe behavioral and human-factor dynamics, such as how trainees modified their language when they knew a recording technology was in use. Further, it allowed Handoff QI leaders to consider whether to trial commercial ambient AI scribes in real-world handoff workflows and whether such technologies may play supportive roles in these workflows. These findings align with emerging frameworks for AI safety evaluation that emphasize controlled, risk-free environments where failure modes and sociotechnical evaluation can be systematically explored [37].

Clinician-Led Innovation as a Model for Health Care AI Development

This study demonstrates that clinician-led development of AI technologies can produce viable AI proof-of-concept platforms for medical education purposes. This approach contrasts with traditional technology-driven innovation models where external developers create solutions that may not align with educational needs. The participatory design methodology [38,39] used in our co-design interviews exemplifies how domain expertise can

XSL•FO
RenderX

be integrated into AI system development for real-world educational needs. In this project, clinicians designed and developed the ambient AI handoff assistant, which enabled them to gain a more realistic understanding of the AI's capabilities and limitations. Furthermore, our experience suggests that clinicians can effectively translate clinical workflows into technical requirements. As an example, the development of the OR-to-ICU handoff note template using prompt engineering in Figure 3 [28], where items in the discussion are mapped to the checklist, demonstrates how clinical expertise can guide AI system design to produce outputs that are clinically meaningful.

Limitations and Future Directions

This study was limited by its exploratory design, small sample size, and simulated environment. Novelty effects and learning curves were evident, as participants included trainees with varying degrees of medical experience and almost no exposure to the OR-to-ICU handoff process at UCLA. The single-site design and use of UCLA's handoff checklist may also limit generalizability to other institutions with differing handoff checklist structures. However, the methodological framework developed in this study was designed to be adaptable across institutional handoff formats. Further, the error analysis was limited in scope and designed to assess feasibility rather than comprehensively evaluate model performance. Transcription errors that are specific to this ambient AI handoff assistant may not generalize to commercial ambient AI scribe platforms, which differ in accuracy, editing workflows, and user interaction. These differences may influence learners' perceptions of ambient AI, particularly prior to direct exposure to commercial ambient AI scribe systems, and give them an inaccurate impression of how real-world commercial ambient AI scribe technologies work. Additionally, this analysis focused on checklist completeness and did not assess handoff sequence, which may affect comprehension and retention. Completeness alone may not reflect handoff quality; future studies should evaluate whether quantitatively more complete handoffs improve knowledge acquisition and retention among receiving clinicians. Additionally, further evaluation in larger samples and among more experienced trainee populations is needed to assess generalizability, usability, and effectiveness of the handoff assistant in supporting handoff education. Despite these limitations, simulation demonstrated the feasibility of using ambient AI technology to train learners in adhering to institutional checklists, reflecting on their communication practices, and gaining exposure to ambient AI technologies.

Conclusions

This study highlights the potential of an ambient AI handoff assistant to enhance handoff education by converting real-time team dialogue into structured documentation for reflection and feedback. Through participatory co-design and simulation-based evaluation, we identified design requirements for integrating AI into handoff training that preserve authentic team-based communication and enrich reflective learning. The ambient AI handoff assistant demonstrated that complex team-based discussions can be mapped to real-world checklists. Although scripted simulations limited the ability to assess the handoff assistant's impact on promoting improved checklist completeness, the AI-generated note emerged as a powerful learning tool that prompted critical discussion on communication clarity, calibration of trust in AI, and the boundaries of AI automation in human-AI collaboration. Imperfect AI output from the ambient AI handoff assistant encouraged learners to verify and question AI-generated information, fostering critical skills required for the safe use of AI in clinical settings. Taken together, these findings underscore how ambient AI technologies can support reflective, team-based learning while laying the groundwork for interactive AI systems that enhance safety, communication, and education in future clinical workflows.

Acknowledgments

ChatGPT (OpenAI) was used during manuscript preparation to assist with grammar refinement and to improve overall readability. The authors reviewed and approved all content and take full responsibility for the accuracy and integrity of the final text. The authors would like to acknowledge Selene Mak, PhD, and Emily Patterson, PhD, for their suggestions and feedback.

Funding

No external funding was provided.

Data Availability

The datasets generated or analyzed during this study are available from the corresponding author on reasonable request.

Authors' Contributions

All authors meet the ICMJE criteria for authorship. Each made substantial contributions to the conception or design of the work. All authors participated in drafting the manuscript or revising it critically for important intellectual content, and approved the final version to be published.

Conflicts of Interest

None declared.

XSL•FO
RenderX

Multimedia Appendix 1
Needs assessment and co-design interview script with clinician educators.
[DOCX File, 2490 KB - mededu_v12i1e85666_app1.docx]

Multimedia Appendix 2
Representative prompt examples for the ambient AI assistant.
[DOCX File, 2547 KB - mededu_v12i1e85666_app2.docx]

Multimedia Appendix 3
Debrief interview guide for simulation learner participants who used the ambient AI handoff assistant.
[DOCX File, 2489 KB - mededu_v12i1e85666_app3.docx]

Multimedia Appendix 4
Final codebook for educator co-design and learner simulation sessions on the use of the AI in handoff education.
[DOCX File, 2491 KB - mededu_v12i1e85666_app4.docx]

Checklist 1
COREQ checklist.
[PDF File, 112 KB - mededu_v12i1e85666_app5.pdf]

References

1. Singh H, Thomas EJ, Petersen LA, Studdert DM. Medical errors involving trainees: a study of closed malpractice claims from 5 insurers. Arch Intern Med 2007 Oct 22;167(19):2030-2036. [doi: 10.1001/archinte.167.19.2030] [Medline: 17954795]
2. Borowitz SM, Waggoner-Fountain LA, Bass EJ, Sledd RM. Adequacy of information transferred at resident sign-out (in-hospital handover of care): a prospective survey. Qual Saf Health Care 2008 Feb;17(1):6-10. [doi: 10.1136/qshc.2006.019273] [Medline: 18245212]
3. Kitch BT, Cooper JB, Zapol WM, et al. Handoffs causing patient harm: a survey of medical and surgical house staff. Jt Comm J Qual Patient Saf 2008 Oct;34(10):563-570. [doi: 10.1016/s1553-7250(08)34071-9] [Medline: 18947116]
4. Petersen LA, Brennan TA, O'Neil AC, Cook EF, Lee TH. Does housestaff discontinuity of care increase the risk for preventable adverse events? Ann Intern Med 1994 Dec 1;121(11):866-872. [doi: 10.7326/0003-4819-121-11-199412010-00008] [Medline: 7978700]
5. Odom-Forren J. Accurate patient handoffs: imperative for patient safety. J Perianesth Nurs 2007 Aug;22(4):233-234. [doi: 10.1016/j.jopan.2007.06.003] [Medline: 17666293]
6. Nagpal K, Arora S, Vats A, et al. Failures in communication and information transfer across the surgical care pathway: interview study. BMJ Qual Saf 2012 Oct;21(10):843-849. [doi: 10.1136/bmjqs-2012-000886] [Medline: 22773891]
7. Krishnamoorthy V, Raghunathan K. Anesthesia care handovers and risk of adverse outcomes. JAMA 2018 Jun 5;319(21):2235-2236. [doi: 10.1001/jama.2018.4442] [Medline: 29872855]
8. Dutra M, Monteiro MV, Ribeiro KB, Schettino GP, Kajdacsy-Balla Amaral AC. Handovers among staff intensivists: a study of information loss and clinical accuracy to anticipate events. Crit Care Med 2018 Nov;46(11):1717-1721. [doi: 10.1097/CCM.0000000000003320] [Medline: 30024429]
9. Abraham J, Meng A, Sona C, Wildes T, Avidan M, Kannampallil T. An observational study of postoperative handoff standardization failures. Int J Med Inform 2021 Jul;151:104458. [doi: 10.1016/j.ijmedinf.2021.104458] [Medline: 33932762]
10. Lane-Fall MB, Pascual JL, Peifer HG, et al. A partially structured postoperative handoff protocol improves communication in 2 mixed surgical intensive care units: findings from the handoffs and transitions in critical care (HATRICC) prospective cohort study. Ann Surg 2020 Mar;271(3):484-493. [doi: 10.1097/SLA.0000000000003137] [Medline: 30499797]
11. Catchpole KR, de Leval MR, McEwan A, et al. Patient handover from surgery to intensive care: using Formula 1 pit-stop and aviation models to improve safety and quality. Paediatr Anaesth 2007 May;17(5):470-478. [doi: 10.1111/j.1460-9592.2006.02239.x] [Medline: 17474955]
12. Hebballi NB, Gupta VS, Sheppard K, et al. Standardization of pediatric noncardiac operating room to intensive care unit handoffs improves communication and patient care. J Patient Saf 2022 Sep 1;18(6):e1021-e1026. [doi: 10.1097/PTS.0000000000000986] [Medline: 35985048]
13. Abraham J, King CR, Meng A. Ascertaining design requirements for postoperative care transition interventions. Appl Clin Inform 2021 Jan;12(1):107-115. [doi: 10.1055/s-0040-1721780] [Medline: 33626584]
14. Schmidt RF, Vibbert MD, Vernick CA, et al. Standardizing postoperative handoffs using the evidence-based IPASS framework through a multidisciplinary initiative improves handoff communication for neurosurgical patients in the neuro-intensive care unit. J Clin Neurosci 2021 Oct;92:67-74. [doi: 10.1016/j.jocn.2021.07.039] [Medline: 34509265]
15. Parent B, LaGrone LN, Albirair MT, et al. Effect of standardized handoff curriculum on improved clinician preparedness in the intensive care unit. JAMA Surg 2018 May 1;153(5):464. [doi: 10.1001/jamasurg.2017.5440] [Medline: 29299602]

XSL•FO
RenderX

16. Conn Busch J, Wu J, Anglade E, Peifer HG, Lane-Fall MB. So many ways to be wrong: completeness and accuracy in a prospective study of OR-to-ICU handoff standardization. Jt Comm J Qual Patient Saf 2023 Aug;49(8):365-372. [doi: 10.1016/j.jcjq.2023.05.001] [Medline: 37316396]
17. Massa S, Wu J, Wang C, Peifer H, Lane-Fall MB. Interprofessional training and communication practices among clinicians in the postoperative ICU handoff. Jt Comm J Qual Patient Saf 2021 Apr;47(4):242-249. [doi: 10.1016/j.jcjq.2020.12.004] [Medline: 33451897]
18. Lazzara EH, Keebler JR, Simonson RJ, Agarwala A, Lane-Fall MB. Navigating the challenges of performing anesthesia handoffs and conducting anesthesia handoff research. Int Anesthesiol Clin 2020;58(1):32-37. [doi: 10.1097/AIA.0000000000000260] [Medline: 31800413]
19. Patterson ES, Wears RL. Patient handoffs: standardized and reliable measurement tools remain elusive. Jt Comm J Qual Patient Saf 2010 Feb;36(2):52-61. [doi: 10.1016/s1553-7250(10)36011-9] [Medline: 20180437]
20. Gershov S, Braunold D, Spektor R, Ioscovich A, Raz A, Laufer S. Automating medical simulations. J Biomed Inform 2023 Aug;144:104446. [doi: 10.1016/j.jbi.2023.104446] [Medline: 37467836]
21. Davis J, Roach C, Elliott C, Mardis M, Justice EM, Riesenberg LA. Feedback and assessment tools for handoffs: a systematic review. J Grad Med Educ 2017 Feb;9(1):18-32. [doi: 10.4300/JGME-D-16-00168.1] [Medline: 28261391]
22. Blackmore A, Kasfiki EV, Purva M. Simulation-based education to improve communication skills: a systematic review and identification of current best practice. BMJ Simul Technol Enhanc Learn 2018;4(4):159-164. [doi: 10.1136/bmjstel-2017-000220] [Medline: 35519010]
23. Winkler-Schwartz A, Bissonnette V, Mirchi N, et al. Artificial intelligence in medical education: best practices using machine learning to assess surgical expertise in virtual reality simulation. J Surg Educ 2019;76(6):1681-1690. [doi: 10.1016/j.jsurg.2019.05.015] [Medline: 31202633]
24. Wartman SA, Combs CD. Medical education must move from the information age to the age of artificial intelligence. Acad Med 2018 Aug;93(8):1107-1109. [doi: 10.1097/ACM.0000000000002044] [Medline: 29095704]
25. Tong A, Sainsbury P, Craig J. Consolidated criteria for reporting qualitative research (COREQ): a 32-item checklist for interviews and focus groups. Int J Qual Health Care 2007 Dec;19(6):349-357. [doi: 10.1093/intqhc/mzm042] [Medline: 17872937]
26. Bate P, Robert G. Experience-based design: from redesigning the system around the patient to co-designing services with the patient. Qual Saf Health Care 2006 Oct;15(5):307-310. [doi: 10.1136/qshc.2005.016527] [Medline: 17074863]
27. Sanders EBN, Stappers PJ. Co-creation and the new landscapes of design. CoDesign 2008 Mar;4(1):5-18. [doi: 10.1080/15710880701875068]
28. Jalilian L, Lukac P, Lane-Fall M. Enhancing structured team communication in acute care settings with ambient AI scribes. J Am Med Inform Assoc 2025 Nov 1;32(11):1785-1789. [doi: 10.1093/jamia/ocaf166] [Medline: 41003645]
29. Barra FL, Rodella G, Costa A, et al. From prompt to platform: an agentic AI workflow for healthcare simulation scenario design. Adv Simul 2025 May 16;10(1):29. [doi: 10.1186/s41077-025-00357-z]
30. Trockel M, Bohman B, Lesure E, et al. A brief instrument to assess both burnout and professional fulfillment in physicians: reliability and validity, including correlation with self-reported medical errors, in a sample of resident and practicing physicians. Acad Psychiatry 2018 Feb;42(1):11-24. [doi: 10.1007/s40596-017-0849-3] [Medline: 29196982]
31. Duff JP, Morse KJ, Seelandt J, et al. Debriefing methods for simulation in healthcare: a systematic review. Simul Healthc 2024 Jan 1;19(1S):S112-S121. [doi: 10.1097/SIH.0000000000000765] [Medline: 38240623]
32. Eppich W, Cheng A. Promoting Excellence and Reflective Learning in Simulation (PEARLS): development and rationale for a blended approach to health care simulation debriefing. Simul Healthc 2015 Apr;10(2):106-115. [doi: 10.1097/SIH.0000000000000072] [Medline: 25710312]
33. Salwei ME, Carayon P, Hoonakker PLT, et al. Workflow integration analysis of a human factors-based clinical decision support in the emergency department. Appl Ergon 2021 Nov;97:103498. [doi: 10.1016/j.apergo.2021.103498] [Medline: 34182430]
34. Dzindolet MT, Peterson SA, Pomranky RA, Pierce LG, Beck HP. The role of trust in automation reliance. Int J Hum Comput Stud 2003 Jun;58(6):697-718. [doi: 10.1016/S1071-5819(03)00038-7]
35. Parasuraman R, Molloy R, Singh IL. Performance consequences of automation-induced “Complacency”. Int J Aviat Psychol 1993 Jan;3(1):1-23. [doi: 10.1207/s15327108ijap0301_1] [Medline: 34182430]
36. Bhutiani M, Hester DL, Lonsdale HJ. Artificial intelligence and the simulationists: more iterations needed. Simul Healthc 2024 Aug 1;19(4):268-269. [doi: 10.1097/SIH.0000000000000803] [Medline: 38922444]
37. Biro JM, Handley JL, Mickler J, et al. The value of simulation testing for the evaluation of ambient digital scribes: a case report. J Am Med Inform Assoc 2025 May 1;32(5):928-931. [doi: 10.1093/jamia/ocaf052] [Medline: 40116912]
38. Spinuzzi C. The methodology of participatory design. Tech Commun 2005 May 1;52(2):163-174 [FREE Full text]
39. Clemensen J, Larsen SB, Kyng M, Kirkevold M. Participatory design in health sciences: using cooperative experimental methods in developing health services and computer technology. Qual Health Res 2007 Jan;17(1):122-130. [doi: 10.1177/1049732306293664] [Medline: 17170250]

XSL•FO
RenderX

Abbreviations

AI: artificial intelligence
API: application programming interface
COREQ: Consolidated Criteria for Reporting Qualitative Research
ICU: intensive care unit
LLM: large language model
OR: operating room
PTL: Physician Task Load Index
SUS: System Usability Scale
UCLA: University of California, Los Angeles

Edited by A Hasan Sapci; submitted 24.Oct.2025; peer-reviewed by A Vannucci, MS Mashuk; revised version received 23.Feb.2026; accepted 05.Apr.2026; published 02.Jul.2026.

Please cite as:
Jalilian L, Barra FL, Grogan T, Lee J, Kadambi A
Exploring Ambient Artificial Intelligence to Enhance Learning and Feedback During Operating Room-to-Intensive Care Unit Handoffs: Co-Design and Simulation Study
JMIR Med Educ 2026;12:e85666
URL: https://mededu.jmir.org/2026/1/e85666
doi:10.2196/85666

© Laleh Jalilian, Federico Lorenzo Barra, Tristan Grogan, Jason Lee, Achuta Kadambi. Originally published in JMIR Medical Education (https://mededu.jmir.org), 2.Jul.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Education, is properly cited. The complete bibliographic information, a link to the original publication on https://mededu.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Medical Students' Experiences With an Integrated Surgical Suturing Training Module Using Simulation Models and Asynchronous Videos: Interpretive Qualitative Study

Álvaro Herrera Alcaíno[1], MD; Consuelo Vade Martínez[2]; Benjamín Painemal Rivera[2]; Daniela Serey Torres[2]; Sebastián Puentes Bravo[2]; Ignacio Leal Lizama[2]; Josefina Darrigol Parra[2]; Camila Riquelme Bahamondes[3], MD

[1]Faculty of Medicine, Universidad San Sebastián, Lota 2465, Santiago, Metropolitan Region, Chile

[2]Universidad San Sebastián, Santiago, Metropolitan Region, Chile

[3]Faculty of Medicine, Universidad San Sebastián, Concepción, Biobio Region, Chile

Corresponding Author:
Álvaro Herrera Alcaíno, MD
Faculty of Medicine, Universidad San Sebastián, Lota 2465, Santiago, Metropolitan Region, Chile

Abstract

Background: Digital educational resources, including asynchronous video-based materials and simulation models, are increasingly used in undergraduate medical education to support procedural skills training. Although both approaches have demonstrated educational value, there is limited qualitative evidence on how the pedagogical integration of these approaches is experienced in real-world clinical training environments, particularly during time-constrained surgical internships.

Objective: This study aimed to explore medical students' and an instructor's experiences with an integrated suturing skills module combining asynchronous instructional videos and simulation-based practice during a surgical internship.

Methods: An interpretive qualitative study was conducted with sixth-year medical students during a surgical internship at a medical school in Chile. The intervention consisted of a suturing simulation module supported by asynchronous instructional videos. Data were collected through 2 focus groups with 20 students and a semistructured interview with 1 instructor after completion of the module. All sessions were audio-recorded, transcribed verbatim, and pseudonymized. Data were analyzed using reflexive thematic analysis following an inductive approach. The Technological Pedagogical Content Knowledge framework was used as a sensitizing lens during the interpretive phase. Ethics approval was obtained prior to data collection, and all participants provided informed consent.

Results: Participants described asynchronous videos as facilitating prior preparation by allowing flexible, self-paced review of procedural steps, which supported students' perceived readiness for hands-on practice. The integration of video-based resources with simulation-based training was perceived as enabling more focused engagement during supervised sessions, shifting attention from procedural recall to technical execution. Participants also reported perceived increases in confidence when approaching early clinical tasks. However, logistical constraints, including limited time for individualized feedback and challenges related to the physical learning environment, were identified. Some participants also expressed ongoing uncertainty regarding specific techniques despite prior video exposure.

Conclusions: The integration of asynchronous instructional videos with simulation-based practice was perceived as a coherent educational strategy that supported students' preparation for procedural training within a demanding clinical context. The findings highlight the potential of asynchronous resources to function as preparatory scaffolding that may enable more efficient use of supervised training time. This study contributes to the qualitative understanding of how integrated digital and simulation-based approaches are experienced in surgical internships, supporting their consideration in curriculum design while emphasizing the need for careful implementation and contextual adaptation.

(JMIR Med Educ 2026;12:e90563) doi:10.2196/90563

KEYWORDS

medical education; surgical education; simulation-based learning; suturing; qualitative research

Introduction

Digital educational resources, including asynchronous video-based learning materials and simulation models, have been increasingly incorporated into undergraduate medical education to support the acquisition of procedural skills [1-3]. In surgical education, these approaches have been used to facilitate deliberate practice, promote learner autonomy, and

XSL•FO
RenderX

enhance preparedness for hands-on clinical training, particularly in settings where opportunities for supervised practice may be limited [2-6]. Recent studies have further highlighted the accelerated adoption of digital and blended learning strategies in health professions education, particularly following shifts toward flexible and technology-enhanced learning environments [7,8].

Instructional videos have been widely adopted as preparatory tools for procedural training, allowing learners to review techniques repeatedly, control the pace of learning, and access content flexibly according to their individual needs [6,9]. Prior research has shown that video-based resources can support self-regulated learning and procedural familiarization, which may be especially relevant during clinically intensive training periods such as surgical clerkships or internships [10,11]. However, most studies have focused on the effectiveness or usability of videos as isolated educational tools, with less attention being paid to how they are integrated into broader pedagogical designs and authentic clinical training environments [3,5,11]. More recent literature suggests that the educational impact of video-based learning depends not only on access to content but also on how these resources are pedagogically integrated into structured learning activities [12].

Simulation-based training has also been recognized as a key strategy for teaching surgical skills in a safe and controlled environment, enabling repeated practice without risk to patients [12,13]. Simulation models have been shown to support the development of technical skills and complement clinical exposure, particularly for basic surgical procedures such as suturing [13-15]. In Latin American medical education contexts, simulation-based approaches for teaching surgical skills have likewise been described as relevant and feasible, highlighting their educational value in settings characterized by high clinical demands and limited protected training time [1,14]. Recent work has also emphasized the role of simulation in supporting early skill acquisition and confidence in resource-constrained and clinically intensive training environments [15].

Despite the growing use of both asynchronous videos and simulation in surgical education, there is limited qualitative evidence exploring how students and instructors experience the combined and pedagogically integrated use of these resources within authentic clinical training contexts. In particular, little is known about how such integrated approaches are perceived and used by students during surgical internships, where time constraints, workload, and clinical responsibilities may shape engagement with educational interventions.

Understanding learners' and instructors' experiences with integrated digital and simulation-based strategies may provide valuable insights into how these approaches function in practice and how they can be meaningfully aligned with pedagogical goals and clinical training demands.

The aim of this study was to explore medical students' and an instructor's experiences with an integrated suturing skills module combining asynchronous instructional videos and simulation-based practice during a surgical internship. Although previous studies have examined video-based learning and simulation independently, less attention has been paid to how these resources are pedagogically integrated within time-constrained clinical environments and how such integration shapes students' engagement with procedural learning.

Methods

Study Design

This study used an interpretive qualitative design to explore how medical students and an instructor experienced the integration of asynchronous instructional videos and simulation-based practice within a surgical internship. A qualitative approach was appropriate because the aim was not to measure technical performance but to examine participants' perceptions and experiences regarding these educational strategies in a real-world clinical training context. This design enabled an in-depth understanding of how learners engaged with preparatory video resources and how these were integrated with hands-on simulation during a clinically demanding stage of training.

Data were analyzed using reflexive thematic analysis, following the approach described by Braun and Clarke [16,17]. This approach supported the identification of patterns of meaning across participants' accounts. Consistent with its epistemological assumptions, the analytic process emphasized reflexive interpretation rather than coding reliability or consensus.

Description of the Intervention

The intervention consisted of a suturing simulation module supported by asynchronous instructional videos implemented at the beginning of the surgical internship.

The asynchronous component included 5 short instructional videos (approximately 5 minutes each) demonstrating simple interrupted, continuous, vertical mattress, horizontal mattress, and intradermal sutures. Viewing was recommended but not mandatory, and engagement was not formally monitored.

The in-person component consisted of a supervised 2-hour workshop using individual simulation models and suturing kits. The instructor provided real-time feedback, and the videos remained accessible during practice.

Participants and Recruitment

Participants were sixth-year medical students enrolled in a surgical internship. All students who completed the module were eligible.

Approximately 1 month after the intervention, students were invited in person by a member of the research team. Participation was voluntary and unrelated to academic evaluation.

A total of 20 students participated. They were typically aged 23 to 25 years and had prior exposure to suturing with variable practice opportunities.

One faculty instructor also took part in a semistructured interview. The instructor had experience teaching procedural skills and did not participate in student evaluation.

Sample adequacy was considered appropriate based on the concept of information power, which suggests that studies addressing focused research aims with specific and

XSL•FO
RenderX

information-rich samples may require fewer participants to generate meaningful qualitative insights. In this study, the aim was narrowly defined, and participants shared a highly specific educational context.

In addition, the use of focus groups and an instructor interview enabled the generation of rich, experience-based data. The application of reflexive thematic analysis supported an in-depth interpretation of participants' accounts, contributing to sufficient information power to address the study aim.

Data Collection

Data were collected through focus groups with students and a semistructured interview with the instructor.

Two focus groups were conducted, each including approximately 8 to 12 students, for a total of 20 participants. Each session lasted approximately 75 minutes and was conducted in person in a meeting room within the medical school. Focus groups took place approximately 1 month after completion of the suturing simulation module, allowing participants to reflect on their experiences after initial exposure to clinical practice during the surgical internship.

In addition, a semistructured individual interview lasting approximately 45 minutes was conducted with the instructor who facilitated the workshop. The interview explored the instructor's perspectives on the design, implementation, and perceived value of the intervention.

A semistructured discussion guide was developed by the research team based on the study objectives and informed by existing literature on simulation-based and video-assisted learning in medical education. The guide was designed to explore key experiential domains aligned with the research aim, including participants' perceptions of the asynchronous videos, experiences during the simulation workshop, perceived facilitators and barriers to learning, and perceived relevance to clinical practice.

These domains were defined to reflect the study's interpretive focus on how participants experienced and made sense of the integration of digital and simulation-based learning strategies within a clinical training context.

The guide was developed collaboratively by the research team and refined through iterative discussion to ensure clarity, relevance, and alignment with the study objectives prior to data collection.

All focus groups and the interview were audio-recorded with participants' consent, transcribed verbatim, and pseudonymized prior to analysis.

Focus groups were moderated by a member of the research team who was not involved in student assessment or direct clinical supervision to reduce potential hierarchical influences on participants' responses.

Data Analysis

Data were analyzed using reflexive thematic analysis, following the approach described by Braun and Clarke [14]. The analysis involved an iterative and recursive process, including familiarization with the data, generation of initial codes, development of candidate themes, review and refinement of themes, and production of the final analytic narrative.

All transcripts were read multiple times to achieve immersion in the data. Initial coding was conducted in a data-driven manner, allowing codes to be generated inductively from participants' accounts. Four members of the research team participated in the analytic process. Coding and theme development were conducted through ongoing, collaborative engagement with the data, supported by regular analytic meetings in which interpretations were discussed, codes refined, and themes developed.

Consistent with the principles of reflexive thematic analysis, these discussions were not intended to establish interrater reliability or achieve coding consensus but rather to support reflexive engagement with the data and deepen interpretive analysis.

During the later stages of analysis, the Technological Pedagogical Content Knowledge (TPACK) framework was used as a sensitizing lens to interpret how technological resources (asynchronous videos), pedagogical strategies (simulation-based instruction), and disciplinary content (suturing techniques) interacted within participants' experiences [18]. This interpretive lens supported the organization and contextualization of themes without constraining the inductive coding process.

In addition, analytic attention was given to cross-cutting experiential dimensions, including motivation, perceived learning, facilitators, barriers, and perceived transfer to clinical practice, which were used as organizing categories to support thematic interpretation. This study was reported in accordance with the Consolidated Criteria for Reporting Qualitative Research (COREQ) guidelines. A complete checklist is provided as Checklist 1.

The research team included faculty members and medical students with experience in medical education. Throughout the analytic process, the research team acknowledged their dual role as educators and researchers and engaged in reflexive discussions to consider how this positioning might influence data interpretation.

No qualitative data analysis software was used. Data organization and coding were conducted manually by the research team through iterative and collaborative engagement with the transcripts.

Ethical Considerations

The study was approved by the scientific ethics committee of *Universidad San Sebastián* (40-25). Ethics approval was obtained prior to the commencement of data collection. All participants received information about the study and provided informed consent prior to participation. Participation was voluntary, and students were explicitly informed that their decision to participate or decline would not affect their academic evaluation or standing. All data were anonymized and managed confidentially throughout the research process. Participants did

XSL•FO
RenderX

not receive financial or other compensation for their participation.

Results

Participants Characteristics

The study included 20 sixth-year medical students who were beginning their surgical internship and had completed a simulated suturing module supported by asynchronous instructional videos, as well as 1 faculty member who facilitated the workshop and supervised students during their clinical training. Students were typically aged between 23 and 25 years and had prior introductory exposure to suturing techniques, although opportunities for repeated supervised practice varied.

All participating students had completed the full educational module prior to taking part in the focus groups. All participants completed the data collection activities, and no withdrawals were recorded.

Primary Results

Analysis of the focus groups and the faculty interview revealed convergent perceptions regarding the use of asynchronous videos in combination with hands-on practice using simulation models during the surgical internship. Participants described how this combined approach was experienced and used within a context characterized by high clinical demands and limited time for structured teaching.

The findings are organized into four interrelated themes: (1) preparation for hands-on practice, (2) integration of asynchronous videos and simulation-based learning, (3) perceived confidence and readiness for clinical tasks, and (4) logistical and contextual constraints.

Preparation for Hands-on Practice

Students described asynchronous videos as a preparatory resource that supported their engagement with hands-on practice, allowing them to approach the workshop with prior familiarity with the procedures and reducing the need for real-time procedural recall. They emphasized the ability to access the content at their own pace, review specific segments, and engage with the material without interfering with clinical responsibilities:

> *The videos were short, well summarized, and concise, so they didn't take much time to review.* [Student, focus group 1]

Several participants highlighted that this flexibility was particularly relevant within the demanding schedule of the surgical internship:

> *Not everyone has the opportunity to practice at home, so having this instance is really helpful, especially at the beginning of the internship.* [Student, focus group 2]

Integration of Asynchronous Videos and Simulation-Based Practice

Participants described the combination of prior video exposure and supervised hands-on practice as enabling more focused engagement during the workshop, shifting attention from procedural recall to technical execution. Students reported that reviewing the procedures beforehand allowed them to focus during the workshop on performing the techniques rather than recalling procedural steps:

> *The tutor played the videos while we were doing the activity, which really helped guide us during practice.* [Student, focus group 1]

> *Being able to see the procedures beforehand allowed me to focus more on doing the sutures properly instead of thinking about what the next step was.* [Student, focus group 1]

Students also described the availability of individual simulation kits as a facilitator of continuous practice during the session:

> *It was great that each student had their own kit, because we didn't have to wait for others to practice.* [Student, focus group 2]

Perceived Confidence and Readiness for Clinical Tasks

Students described a sense of increased confidence and perceived readiness when approaching early clinical tasks, which they associated with having had the opportunity to engage in structured and supported practice prior to clinical exposure:

> *It helped me start clinical practice feeling more confident, because suturing is something that's requested from the very first shifts.* [Student, focus group 2]

Some participants also noted that the workshop provided an opportunity to revisit and practice skills that were often assumed to be already acquired but not consistently reinforced under supervision:

> *It's something that people often assume you already know how to do well, but in reality you only learn by practicing.* [Student, focus group 1]

Logistical and Contextual Constraints

Although participants generally described the module positively, they also identified challenges related to time constraints and limited opportunities for individualized feedback during the workshop:

> *There were some sutures where I wasn't sure if they were right or wrong, because there wasn't enough time to review everyone's work.* [Student, focus group 2]

These constraints were described in relation to the number of students per instructor and the limited duration of the session.

From the faculty perspective, the use of asynchronous videos was described as supporting a more focused use of in-person time, allowing attention to be directed toward technical aspects and student questions:

> *They arrived with a prior foundation, which made it possible to focus more on improving technique rather than explaining everything from scratch.* [Faculty member]

XSL•FO
RenderX

The faculty member also identified logistical challenges related to the physical space and audio conditions, while noting that prior exposure to the videos influenced how the workshop was experienced:

> *The difficulties were more related to the physical space than to the videos themselves; the important thing was that they had already reviewed the material beforehand.* [Faculty member]

Perceived Applicability to Other Areas of Training

Participants described the potential applicability of this educational approach to other areas of the surgical internship, particularly for skills that are expected in clinical practice but not always explicitly taught:

> *It would be useful to have this kind of preparation in other areas as well, like trauma or wound care, because these are skills that are expected but not always taught beforehand.* [Student, focus group 2]

From the faculty perspective, the experience was described as a potential basis for broader integration of audiovisual resources and simulation within internship training:

> *It's a tool that could be maintained and used with other groups of students.* [Faculty member]

Discussion

Principal Findings

This study explored how medical students and an instructor experienced an asynchronous video-supported suturing simulation module during a surgical internship. Overall, participants described the integrated use of asynchronous videos and supervised simulation-based practice as a useful and contextually relevant educational strategy in a setting characterized by high clinical demands and limited protected teaching time.

Across themes, participants highlighted the role of asynchronous videos as a preparatory resource, the integration of video-based and hands-on learning during the workshop, and a perceived sense of confidence and readiness when approaching early clinical tasks. At the same time, logistical constraints related to time, feedback opportunities, and physical space were also identified.

Interpretation in Relation to Existing Literature

The findings of this study align with prior research suggesting that asynchronous video-based learning may support flexible access to instructional content and allow learners to engage with procedural knowledge at their own pace. In this study, participants described how this flexibility was particularly relevant within the demanding context of a surgical internship, where time for structured learning is often limited [4,5]. Recent studies have further emphasized the role of asynchronous video-based learning in supporting self-regulated learning and flexible engagement in clinically intensive training environments, reinforcing its relevance in contexts with competing clinical demands [19,20].

In addition, the perceived value of combining asynchronous videos with simulation-based practice is consistent with the literature emphasizing the role of multimodal educational strategies in procedural skills training. Rather than functioning as a replacement for supervised instruction, participants described asynchronous videos as a preparatory scaffold that enabled more focused engagement during hands-on practice [7,10]. More recent literature has highlighted that integrating digital resources with simulation-based instruction may enhance the efficiency of in-person training by allowing learners to arrive with foundational procedural understanding, thereby optimizing opportunities for feedback and technical refinement [21,22].

Importantly, this study extends the existing literature by providing qualitative insights into how this combination of resources is experienced within a real-world clinical training environment [23]. Although previous studies have examined video-based learning or simulation independently, the findings of this study highlight how their integration may be perceived as supporting the organization and use of limited instructional time during surgical internships [24]. In this sense, this study contributes to a growing body of research that moves beyond evaluating isolated educational tools and instead examines how integrated pedagogical designs are experienced and enacted within authentic clinical training settings.

A key contribution of this study lies in conceptualizing asynchronous videos not merely as supplementary resources but as a form of preparatory pedagogical scaffolding within clinically demanding training environments. This scaffolding appears to enable a redistribution of cognitive and instructional effort, allowing supervised sessions to focus more explicitly on technical refinement and feedback. In this sense, the integration of digital and simulation-based resources may be understood as a strategy for optimizing the pedagogical use of limited clinical teaching time.

The use of the TPACK framework as an interpretive lens further supports understanding how technological resources (videos), pedagogical approaches (simulation-based instruction), and disciplinary content (suturing techniques) interact within situated learning experiences. From a TPACK perspective, participants' experiences suggest that the perceived value of the intervention emerged from the interaction among technological resources (videos), pedagogical design (simulation-based instruction), and disciplinary content (suturing skills), rather than from any single component in isolation [8].

From a curricular perspective, these findings suggest that asynchronous instructional videos may function as preparatory scaffolding that allows simulation-based workshops to focus more effectively on hands-on practice and feedback. In contexts where teaching time is constrained, such as surgical internships, this approach may enable a more pedagogically efficient use of in-person instructional time [7,11].

Participants' accounts also highlight the importance of ensuring adequate opportunities for individualized feedback and appropriate instructor-to-student ratios during simulation sessions. Additionally, logistical factors such as physical space and audio conditions were described as influencing the learning

XSL•FO
RenderX

experience, suggesting that implementation requires attention not only to educational design but also to practical conditions.

The perceived applicability of this approach to other areas of clinical training suggests potential for broader integration of asynchronous and simulation-based strategies within undergraduate medical education. However, such adaptations should consider contextual factors and resource availability.

Limitations

Several limitations should be considered when interpreting these findings. First, the study was conducted within a single medical school and involved students from one internship cohort, which may limit the transferability of the results to other institutional contexts. Second, the faculty perspective was represented by a single instructor interview, which restricts the diversity of teaching perspectives included in the analysis.

Third, the study focused on participants' perceptions and did not include objective measures of procedural skill performance. In addition, data collection was conducted approximately one month after completion of the intervention. This timing may have influenced participants' recollections, as their experiences could have been shaped by subsequent clinical exposure during the surgical internship. This temporal distance may have introduced recall bias and influenced how participants interpreted and evaluated the perceived usefulness of the module. Social desirability bias during focus group discussions may also have shaped participants' responses.

Engagement with the asynchronous video component was not formally measured, and individual exposure to the videos may have varied across participants. Educational practices and clinical training structures in Chile may also differ from those in other regions, which should be considered when interpreting the broader applicability of the findings.

The findings are best understood as contextually situated insights that may inform transferability to similar educational settings rather than broad generalization.

Finally, the involvement of the research team in the educational context may have influenced participants' responses, which was addressed through reflexive discussions during the analytic process.

Conclusions

This study provides qualitative insights into how medical students and an instructor experienced the integration of asynchronous video resources and simulation-based practice within a surgical internship. Participants perceived this combined approach as supporting preparation for hands-on learning and shaping how supervised training time was used. These findings suggest that asynchronous instructional videos may function as a form of preparatory pedagogical scaffolding that enables a more focused and efficient use of simulation-based training within clinically demanding environments. Further research is needed to explore how such approaches can be adapted across different contexts and to examine their relationship with objective measures of procedural learning.

Acknowledgments

The authors thank the participating medical students and the instructor for their time and willingness to share their experiences. The authors also acknowledge the institutional support of *Universidad San Sebastián*, Puerto Montt campus, for the implementation of the simulated module. The authors used an artificial intelligence–based language model (ChatGPT version; OpenAI) to support the drafting and linguistic editing of the manuscript. All content was critically reviewed and approved by the authors, who take full responsibility for the accuracy and integrity of the work.

Funding

This project was fully funded by *Universidad San Sebastián*. This study was supported by *Vicerrectoría Académica* and *Centro de Investigación para la Educación Superior* (grant USS-FIN-2026_APC_VRA-004).

Data Availability

Due to the qualitative nature of the study and the potential risk of participant identification, the data generated and analyzed are not publicly available. Anonymized excerpts may be made available from the corresponding author upon reasonable request.

Authors' Contributions

CVM, BPR, DST, SPB, ILL, and JDP contributed to study conceptualization, methodology, data collection, collaborative data analysis, and preparation of the original draft. AHA and CRB contributed to collaborative data analysis, manuscript review and editing, and supervision of the research team.

Conflicts of Interest

None declared.

Checklist 1
COREQ checklist.
[PDF File, 79 KB - mededu_v12i1e90563_app1.pdf]

XSL•FO
RenderX

References

1. Kyaw BM, Saxena N, Posadzki P, et al. Virtual reality for health professions education: systematic review and meta-analysis by the digital health education collaboration. J Med Internet Res 2019 Jan 22;21(1):e12959. [doi: 10.2196/12959] [Medline: 30668519]
2. Prober CG, Khan S. Medical education reimagined: a call to action. Acad Med 2013 Oct;88(10):1407-1410. [doi: 10.1097/ACM.0b013e3182a368bd] [Medline: 23969367]
3. Valencia-Ortiz R, Garay U, Cabero-Almenara J. Students' and teachers' perceptions of students' use of the internet and its relationship with the study modality [Article in Spanish]. Revista De Educación a Distancia (RED) 2020;20(62). [doi: 10.6018/red.411781]
4. Kononowicz AA, Woodham LA, Edelbring S, et al. Virtual patient simulations in health professions education: systematic review and meta-analysis by the digital health education collaboration. J Med Internet Res 2019 Jul 2;21(7):e14676. [doi: 10.2196/14676] [Medline: 31267981]
5. Rapp AK, Healy MG, Charlton ME, Keith JN, Rosenbaum ME, Kapadia MR. YouTube is the most frequently used educational video source for surgical preparation. J Surg Educ 2016;73(6):1072-1076. [doi: 10.1016/j.jsurg.2016.04.024] [Medline: 27316383]
6. Chen F, Lui AM, Martinelli SM. A systematic review of the effectiveness of flipped classrooms in medical education. Med Educ 2017 Jun;51(6):585-597. [doi: 10.1111/medu.13272] [Medline: 28488303]
7. Almohammadi NH, Elnugomi NM, Atta Elmannan AA, Zalat MM. Perspectives of medical students in using blended learning: a thematic analysis. Adv Med Educ Pract 2025;16:837-847. [doi: 10.2147/AMEP.S515941] [Medline: 40416900]
8. Gordon M, Patricio M, Horne L, et al. Developments in medical education in response to the COVID-19 pandemic: a rapid BEME systematic review: BEME guide no. 63. Med Teach 2020 Nov;42(11):1202-1215. [doi: 10.1080/0142159X.2020.1807484] [Medline: 32847456]
9. McGaghie WC, Issenberg SB, Petrusa ER, Scalese RJ. A critical review of simulation-based medical education research: 2003-2009. Med Educ 2010 Jan;44(1):50-63. [doi: 10.1111/j.1365-2923.2009.03547.x] [Medline: 20078756]
10. Moncada A, García R, Rodríguez M, Fernández A, Varela CL, Terán AJ. Simulation as a tool for surgical teaching and learning [Article in Spanish]. Venez J Surg 2023;76(2):120-125. [doi: 10.48104/RVC.2023.76.2.14]
11. Riddle EW, Kewalramani D, Narayan M, Jones DB. Surgical simulation: virtual reality to artificial intelligence. Curr Probl Surg 2024 Nov;61(11):101625. [doi: 10.1016/j.cpsurg.2024.101625] [Medline: 39477664]
12. Bland T, Guo M, Dousay TA. Multimedia design for learner interest and achievement: a visual guide to pharmacology. BMC Med Educ 2024 Feb 5;24(1):113. [doi: 10.1186/s12909-024-05077-y] [Medline: 38317141]
13. Aggarwal R, Mytton OT, Derbrew M, et al. Training and simulation for patient safety. Qual Saf Health Care 2010 Aug;19 Suppl 2:i34-i43. [doi: 10.1136/qshc.2009.038562] [Medline: 20693215]
14. Houshyari AB, Bahadorani M, Tootoonchi M, Gardiner JJ, Peña RA, Adibi P. Medical education and information and communication technology. J Educ Health Promot 2012;1:3. [doi: 10.4103/2277-9531.94411] [Medline: 23555106]
15. Armijo-Rivera S, Machuca-Contreras F, Raul N, et al. Characterization of simulation centers and programs in Latin America according to the ASPIRE and SSH quality criteria. Adv Simul (Lond) 2021 Nov 12;6(1):41. [doi: 10.1186/s41077-021-00188-8] [Medline: 34772461]
16. Braun V, Clarke V. Reflecting on reflexive thematic analysis. Qual Res Sport Exerc Health 2019;11(4):589-597. [doi: 10.1080/2159676X.2019.1628806]
17. Braun V, Clarke V. Using thematic analysis in psychology. Qual Res Psychol 2006 Jan;3(2):77-101. [doi: 10.1191/1478088706qp063oa]
18. Mishra P, Koehler MJ. Technological pedagogical content knowledge: a framework for teacher knowledge. Teach Coll Rec 2006;108(6):1017-1054. [doi: 10.1111/j.1467-9620.2006.00684.x]
19. Ajjawi R, Kent F. Understanding realist reviews for medical education. J Grad Med Educ 2022 Jun;14(3):274-278. [doi: 10.4300/JGME-D-22-00334.1] [Medline: 35754632]
20. Bain R, Lee J, Soars S, et al. Barriers and enablers to medical students' engagement in research and clinical academic careers: a mixed-methods study. BMJ Open 2025 Oct 15;15(10):e102988. [doi: 10.1136/bmjopen-2025-102988] [Medline: 41093348]
21. Li M, Hong Y, Wu A, et al. The effectiveness of blended learning in nursing and medical education: an umbrella review. Nurse Educ Pract 2025 Jul;86:104421. [doi: 10.1016/j.nepr.2025.104421] [Medline: 40483931]
22. Gong J, Ruan M, Yang W, et al. Application of blended learning approach in clinical skills to stimulate active learning attitudes and improve clinical practice among medical students. PeerJ 2021;9:e11690. [doi: 10.7717/peerj.11690] [Medline: 34221742]
23. Lembo D, Abate Daga F, Calì C, et al. Early introduction of simulation in the medical curriculum: the MedInTo perspective. Front Med (Lausanne) 2024;10:1280592. [doi: 10.3389/fmed.2023.1280592] [Medline: 38239607]
24. Nazirova X. Integrating simulation-based and technology-enhanced learning in contemporary medical education: insights from surgery, internal medicine, and urology. J Clin Biomed Res 2026;2(4):61-72 [FREE Full text]

XSL•FO
RenderX

Abbreviations

COREQ: Consolidated Criteria for Reporting Qualitative Research
TPACK: Technological Pedagogical Content Knowledge

Edited by A Stone; submitted 02.Jan.2026; peer-reviewed by J Liu, M Kalem; revised version received 04.May.2026; accepted 06.May.2026; published 09.Jul.2026.

Please cite as:
Herrera Alcaíno Á, Vade Martínez C, Painemal Rivera B, Serey Torres D, Puentes Bravo S, Leal Lizama I, Darrigol Parra J, Riquelme Bahamondes C
Medical Students' Experiences With an Integrated Surgical Suturing Training Module Using Simulation Models and Asynchronous Videos: Interpretive Qualitative Study
JMIR Med Educ 2026;12:e90563
URL: https://mededu.jmir.org/2026/1/e90563
doi:10.2196/90563

© Álvaro Herrera Alcaíno, Consuelo Vade Martínez, Benjamín Painemal Rivera, Daniela Serey Torres, Sebastián Puentes Bravo, Ignacio Leal Lizama, Josefina Darrigol Parra, Camila Riquelme Bahamondes. Originally published in JMIR Medical Education (https://mededu.jmir.org), 9.Jul.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Education, is properly cited. The complete bibliographic information, a link to the original publication on https://mededu.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Enhancing Engagement With Endocrine Guidelines and Fostering Medical Student Interest Through Concise Medical Information Cines: Qualitative Co-Design Study

Rahul Sagu[1], MBChB, FHEA; Sangamithra Ravi[1], BSc FHEA; Aspasia Manta[2], MSc, MD; Punith Kempegowda[2,3], MSc, MBBS, MD, PhD, SFHEA, FRCP

[1]Birmingham Medical School, College of Medicine and Health, University of Birmingham, Birmingham, United Kingdom

[2]Department of Applied Health Sciences, School of Health Sciences, College of Medicine and Health, University of Birmingham, Birmingham, United Kingdom

[3]Department of Endocrinology and Diabetes, Queen Elizabeth Hospital Birmingham, University Hospitals Birmingham NHS Foundation Trust, Mindelsohn Way, Birmingham, United Kingdom

Corresponding Author:
Punith Kempegowda, MSc, MBBS, MD, PhD, SFHEA, FRCP
Department of Applied Health Sciences, School of Health Sciences, College of Medicine and Health, University of Birmingham, Birmingham, United Kingdom

Abstract

Background: There is a need to modernize the dissemination of clinical guidelines, making them more accessible and engaging for health care professionals. Concise Medical Information Cines (CoMICs) are peer-reviewed videos created by medical students that distill complex guidelines into learner-friendly visuals.

Objective: This study aimed to describe the process of co-designing an audiovisual version of a clinical guideline and explore the experiences of co-designing audiovisual guideline summaries using the CoMICs model.

Methods: A 4-part CoMICs series on glucocorticoid-induced adrenal insufficiency was codeveloped by clinicians and medical students through 10 iterative steps. A patient version of these CoMICs was then created in multiple languages. Semistructured interviews with authors, reviewers, and student collaborators assessed the clarity, usability, trustworthiness, and educational value of these CoMICs. Reflexive thematic analysis then identified key themes.

Results: CoMICs improved guideline accessibility, comprehension, and global adaptability, while the collaborative process promoted interdisciplinary learning and underscored the efficacy of audiovisual tools for complex content. Student collaborators reported greater confidence in interpreting and communicating clinical guidance, renewed interest in endocrinology, and a deeper appreciation of its academic dimensions.

Conclusions: Cocreating audiovisual resources, such as CoMICs, enhances guideline dissemination. Student involvement can foster curiosity, encourage academic career pathways, and reshape engagement with evidence-based medicine.

(JMIR Med Educ 2026;12:e83711) doi:10.2196/83711

KEYWORDS

guidelines; education; endocrinology; dissemination; implementation

Introduction

Clinical practice guidelines are a cornerstone of evidence-based medicine, offering systematically developed recommendations to optimize patient care [1]. Despite their importance, ensuring that guidelines are effectively disseminated and implemented in clinical settings remains a persistent challenge. Health care professionals often face barriers such as time constraints, cognitive overload, lack of awareness of updates, and perceived limited contextual relevance—factors that reduce the usability of traditional long-form, text-based documents in fast-paced environments [2].

Dissemination refers to the targeted distribution of evidence-based knowledge and materials to specific audiences to increase awareness and understanding [3]. Conventional dissemination approaches, such as printed summaries, educational events, audit and feedback loops, and reminder systems, have shown variable success [4]. While multifaceted interventions tend to outperform single strategies, their implementation is often constrained by organizational resources, and they fail to achieve consistent improvements in clinical behaviors or patient outcomes. Implementation, in this context, involves systematically integrating that knowledge into real-world practice to improve health care outcomes. More importantly, these strategies frequently overlook the importance

XSL•FO
RenderX

of engagement, accessibility, and real-time utility—features that are increasingly essential in today's diverse and dynamic health care systems [5]. These limitations underscore the need for supplementary dissemination strategies that can deliver key recommendations in more accessible, engaging, and time-efficient formats [1].

Emerging audiovisual tools offer an opportunity to bridge this gap. By reducing cognitive load, enhancing recall, and improving user engagement, visual formats can complement traditional learning formats [6]. One such approach is Concise Medical Information Cines (CoMICs)—short, peer-reviewed, visually engaging videos designed to translate complex clinical guidance into digestible content [7]. Using infographics, narration, and visual storytelling, CoMICs aim to make guidelines more approachable and usable for both early-career and experienced professionals.

In parallel with growing calls to diversify how students experience specialty fields and engage with research, coproducing educational resources such as CoMICs allows students to gain exposure to academic medicine in a hands-on, creative manner. The Endocrine Society's Medical School Engagement Program similarly seeks to broaden medical students' exposure to endocrinology [8]. By aligning with the goals of the Medical School Engagement Program, CoMICs offer a complementary, hands-on platform that both disseminates core guideline content and actively engages students in the specialty. While the earlier CoMICs study investigated the dissemination and engagement patterns of specialty-focused educational videos on social media platforms, the translation of a clinical guideline into a structured audiovisual format has not been examined in previous CoMICs research. Therefore, in this study, we describe the process of codeveloping a 4-part CoMICs series based on the glucocorticoid-induced adrenal insufficiency (GIAI) guideline, showcasing an innovative approach to translating clinical guidance into accessible resources. We also explored medical students' experiences of co-designing the CoMICs and how this influenced their learning.

Methods

Overview

This study was conducted between October 2024 and May 2025 at the University of Birmingham. The process of CoMICs creation, aligned with the knowledge-to-action framework, has been described in detail in our previous publication [7]. To inform the adoption of CoMICs for guideline dissemination, we conducted a structured exploratory literature review using PubMed. The search strategy focused on 3 intersecting themes: dissemination methods, barriers to the uptake of clinical guidelines, and health care professional behaviors. The review was limited to English-language articles involving health care professionals but included studies from all publication years to ensure comprehensive coverage. Key findings highlighted the limitations of traditional dissemination techniques, such as printed materials, educational sessions, and audit-feedback cycles, and underscored the need for more engaging and accessible formats, particularly in time-pressured clinical environments [4]. These insights shaped the rationale for the CoMICs approach. A 4-part audiovisual series was created to convey the core content of the Joint European Society of Endocrinology and Endocrine Society guideline on GIAI. The primary focus of this work is qualitative, centered on the co-design and evaluation of CoMICs. The literature review served as a supporting framework for this study. Each video distilled key diagnostic, therapeutic, and follow-up recommendations into concise, visually engaging episodes. Crucially, medical students played an active role throughout the ten-step CoMICs development process, which included the following:

1. The CoMICs team contacted the corresponding guideline author to establish collaboration, define the scope of the project, and confirm participation in an iterative review.
2. Guideline authors shared supplementary educational resources (eg, slide decks, figures, explanatory notes, or key references) to support accurate interpretation of the guideline.
3. The CoMICs team developed an initial script and visual storyboard using PowerPoint (Microsoft Corp), translating key guideline recommendations into simplified, learner-friendly visual narratives.
4. The storyboard underwent clinical review by guideline authors, who provided feedback on accuracy, emphasis, and clarity, after which suggested revisions were incorporated.
5. A preliminary video draft without narration or background audio was produced to enable focused review of structure, visuals, and content flow.
6. Feedback from guideline authors, clinical experts, and peer reviewers was collated by the CoMICs team and used to refine content, visuals, and sequencing.
7. A prefinal version incorporating narration and refined visuals was produced and circulated to the authors for final clinical validation.
8. Final comments and minor revisions were addressed to ensure consistency with the guideline and clarity for the intended audience.
9. The completed CoMIC was approved by the corresponding author and prepared for professional use.
10. Approved videos were disseminated through clinical networks and social media platforms to maximize accessibility, reach, and engagement [9].

This process not only supported the technical production of CoMICs but also served as a structured educational opportunity, exposing students to academic publishing, clinical content curation, and interdisciplinary teamwork in endocrinology. The CoMICs videos were disseminated through multiple online platforms to maximize reach and engagement. All videos were uploaded to YouTube (Google LLC) as the primary hosting platform and shared via social media channels, including X (X Corp, formerly Twitter), Instagram (Meta Platforms, Inc), and WhatsApp (Meta Platforms, Inc) groups within the CoMICs network. The posts were accompanied by short educational captions and reposted by CoMICs members on the aforementioned platforms to enhance discoverability.

XSL•FO
RenderX

Multilingual Patient-Focused CoMICs Development

In parallel, a patient-facing version of the CoMICs series was developed to improve health literacy and access among lay audiences. These 4 short videos were co-designed with individuals living with adrenal insufficiency and deliberately avoided technical language. Patient advocates provided feedback throughout script development, ensuring clarity, relevance, and the inclusion of relatable, everyday examples. Practical guidance on symptom recognition, treatment adherence, and self-management was integrated.

To further enhance accessibility, multilingual versions were created. Medical students fluent in English and another language of interest to this project produced initial transcripts, which were then reviewed by native speakers with clinical or health communication experience. This ensured that idiomatic clarity, cultural appropriateness, and medical accuracy were preserved across language versions. Languages for the multilingual video adaptations were selected pragmatically based on the connections within the CoMICs network and the availability of fluent volunteer translators. This approach ensured linguistic accuracy while allowing culturally relevant adaptations for each audience. This process enhanced the reach and relevance of the CoMICs series to non–English-speaking audiences, promoting global health equity [10].

Qualitative Interviews to Explore Experience and Impact

Purposive sampling was used to recruit participants with direct involvement in the CoMICs project or relevant expertise. These included the following:

- Medical students and health care professionals involved in the CoMICs cocreation
- Clinical experts who authored or peer-reviewed the GIAI guideline

Among the CoMICs team members interviewed, there were 10 medical students and 2 physicians. The medical students were in their penultimate or final year of study, while the 2 physicians included 1 foundation physician and 1 honorary research clinical fellow. All participants demonstrated an interest in endocrinology, having been active members of the CoMICs initiative for at least 1 year and engaged in ongoing development of the CoMICs guideline. Their participation reflected both their exploratory career interests and commitment to evidence-based knowledge dissemination.

Semistructured interviews explored participants' experiences with clinical guidelines, perceptions of the CoMICs format, and the personal or professional impact of their involvement. Students were also asked to reflect on how the experience shaped their views on endocrinology, academic collaboration, and the role of creative dissemination in medicine. The interview guide was aligned with the 2 objectives of the study and focused on clarity, usability, coproduction dynamics, and suggestions for future improvement. All interviews were audio-recorded and transcribed verbatim. Saturation was considered reached when participants could no longer provide additional insights or perspectives relevant to the research questions.

Ethical Considerations

The study was reviewed and received ethics approval from the Science, Technology, Engineering, and Mathematics Committee at the University of Birmingham (ERN_2965-Aug2024). All data were anonymized to safeguard the participants' information. Written informed consent was obtained from all participants before the interview, in accordance with institutional guidance and the approved ethics protocol. Participation was voluntary, and participants were not compensated for their time.

Data Analysis

Interview data were analyzed using reflexive thematic analysis following the six-phase framework by Braun and Clarke [10]:

1. Familiarization: transcripts were reviewed in depth, with initial impressions documented.
2. Generating initial codes: RS undertook repeated familiarization with the transcripts before manually coding meaningful segments of text. Initial codes were developed inductively and generated directly from the data at a semantic level. Coding was data-driven and iterative, with reflexive notes maintained throughout to document analytic decisions and support ongoing refinement. Preliminary codes were shared with the research team, and team discussions informed the refinement and consolidation of the initial coding set.
3. Constructing initial themes: related codes were clustered to form preliminary candidate themes.
4. Reviewing themes: themes were refined for coherence and distinctiveness, with weak or overlapping themes revised or removed.
5. Defining and naming themes: final themes were clearly defined, named, and supported with representative quotations through group discussion. Generative artificial intelligence assistance was used only to refine subthemes, and all analytic decisions were confirmed through team consensus.
6. Producing the report: themes were organized in relation to the study objectives and integrated with relevant literature. This approach enabled a nuanced understanding of the practical and pedagogical value of audiovisual dissemination and provided insight into how cocreation experiences can foster medical students' interest in specialty areas and academic pathways. The themes were contextualized using findings from the literature review, highlighting the potential for CoMICs to address longstanding dissemination challenges while also serving as a platform for early academic engagement.

Results

Overview

The 4-part guideline CoMICs series for health care professionals has accumulated 1218 views on YouTube and 1331 views on X (as of November 22, 2025) since its creation. The patient and public versions were created in English, Bengali, Serbian, Tamil, Greek, Georgian, and Brazilian Portuguese and have been viewed 1884 times on YouTube and 2420 times on X (as of November 22, 2025) since their creation. The final CoMICs

XSL•FO
RenderX

videos are publicly accessible via YouTube (Multimedia Appendix 1).

Thematic Analysis of Interviews

Overview

A total of 15 participants—12 (80%) medical students and 3 (20%) senior health care professionals—involved in the development or review of the GIAI CoMICs series were interviewed. Thematic analysis revealed 5 overarching themes with associated subthemes and codes (Table 1), with supporting illustrative quotes.

Table . Thematic analysis results.

Theme and subtheme	Code
Accessibility and usability	• Use NICE[a] for guidelines • Written guidelines are too long • Prefer short videos • Guidelines difficult to navigate in clinical settings • Multilingual versions broaden access • Supports diverse learner backgrounds
Visual and cognitive engagement	
Quick refreshers	• Use videos as refreshers before clinics • Videos used during breaks or on commutes
Clinician use cases	• Visuals improve learning and retention • Animations clarify physiology and treatments
Credibility and trust	• Expert review adds legitimacy • Professional tone builds trust • Referencing and consistency enhance credibility
Empowerment through cocreation	
Workflow friction	• Students gained confidence and communication skills • Challenges in feedback and team coordination • Improved time and project management skills
Inclusivity and cultural reach	• Multilingual versions increase reach • Non-English speakers benefit from translations • Can be shared with patients from diverse backgrounds

[a]NICE: National Institute for Health and Care Excellence.

Accessibility and Usability

Participants reported that traditional guidelines were often inaccessible in time-constrained settings, particularly during clinical work or exam preparation. Students and clinicians alike praised CoMICs as practical alternatives that distilled core messages into short, visually engaging formats. Furthermore, participants highlighted that multilingual CoMICs videos broadened accessibility and usability for parents and learners from diverse linguistic backgrounds, enhancing inclusivity and reach, especially where written English guidelines pose barriers:

> *Even just a quick 2-minute summary video is more useful than scrolling through 30 pages.* [Medical student]

Visual and Cognitive Engagement

CoMICs' design—combining narration, visual metaphors, and sequential storytelling—was credited with enhancing comprehension and retention. Students appreciated the ability to review content quickly and repeatedly, while health care professionals used the videos as just-in-time learning tools before clinical duties:

> *The animations helped me understand the physiology without flipping through the textbook.* [CoMICs student team member]

Credibility and Trust

The rigorous review process, which involved guideline authors and clinical experts, was critical to establishing credibility.

XSL•FO
RenderX

Students valued knowing their contributions would be clinically validated, while reviewers endorsed the professional standard of the final product:

> *If you know the guideline authors reviewed it, it feels legitimate—not just a student project.* [Medical student]

Empowerment Through Cocreation

Student participants expressed a strong sense of empowerment, seeing their work translated into tangible educational resources. They highlighted growth in communication, critical thinking, and teamwork skills. The experience also offered early exposure to academic endocrinology and knowledge translation, reinforcing interest in future academic careers:

> *I never thought a video I helped make would be used by actual doctors.* [Medical student]

Despite some communication delays and revision bottlenecks, students viewed the challenges as valuable professional learning opportunities.

Inclusivity and Cultural Reach

Multilingual adaptations were particularly valued for expanding access. Students noted that having videos in multiple languages allowed them to share resources with patients and peers from diverse backgrounds, enhancing equity in education and patient engagement:

> *The Bangla version is great—I could share it with patients too.* [Medical student]

This inclusive element positioned CoMICs as a globally relevant tool that could strengthen both practitioner education and patient-facing communication.

Discussion

Principal Findings

This study represents a novel contribution to the field of clinical guideline dissemination and medical education. To the best of our knowledge, it is among the first to explore the cocreation of audiovisual, narrative-based clinical resources—CoMICs—that are not only grounded in evidence-based guidelines but also designed for multilingual and multicultural relevance. Unlike previous CoMICs initiatives that focused on broad health topics and social media engagement, this study uniquely evaluates the transformation of a single clinical guideline into structured audiovisual outputs and examines the educational impact of the cocreation process on medical students. The visual structure of CoMICs supports the cognitive load theory by minimizing extraneous load and enabling learners to process guideline content more efficiently [11]. Previous reviews have emphasized the limited effectiveness of traditional dissemination methods (eg, printed guidelines and lectures) unless paired with contextual relevance, interactivity, and credible authorship [5,7,12,13]. CoMICs address these limitations through visual storytelling, coproduction with guideline authors, and multilingual delivery, making them a more adaptable and engaging format for real-world use. This aligns with emerging innovative dissemination practices, which emphasize concise, visually driven formats to improve the reach and uptake of health evidence [14].

Furthermore, this study supports growing calls in the health care education literature for multifaceted, user-driven dissemination models [15]. Notably, CoMICs serve not only as tools to improve practitioner guideline uptake but also as platforms for early-career professional development. Medical students involved in the production process gained firsthand experience in academic endocrinology, communication, interprofessional collaboration, and health education. Such active involvement can foster future interest in both clinical specialties and research careers—an impact rarely explored in the dissemination literature.

Our findings can be contextualized within dissemination science frameworks. The Framework for Knowledge Transfer by Lavis et al [16] proposes that effective dissemination depends on 5 core questions: what should be transferred, to whom, by whom, how, and with what effect. CoMICs directly addresses each of these domains: it condenses evidence-based guideline messages (what), targets medical students and patients (to whom), involves credible experts and peers (by whom), delivers content through short multilingual videos and social media (how), and measures engagement and understanding (with what effect). By mapping CoMICs to these principles, our approach provides a structured, theoretically grounded model for disseminating evidence through digital media.

The strength of this study lies in its dual-impact design, which simultaneously addresses the challenges of guideline dissemination and supports the development of future academic clinicians. The cocreative model provided students with meaningful, mentored engagement in real-world implementation efforts, resulting in a finished product with practical utility for health care professionals and patients alike. The inclusion of multilingual and culturally relevant content further extends CoMICs' global reach, especially in resource-limited or linguistically diverse health care settings. Additionally, by capturing reflections from both clinical experts and student collaborators, the study offers a rich, balanced qualitative perspective on how innovation, collaboration, and mentorship can enhance both educational quality and clinical relevance.

Limitations

We did not measure actual changes in clinical behavior or patient outcomes. While the findings indicate strong potential, objective evidence of impact remains limited. Moreover, because the interviewees were directly involved in developing the resources, positive bias may have influenced responses, potentially limiting the generalizability of the findings. It is also important to note that quantitative metrics, such as video views and engagement figures, were included solely to contextualize the qualitative findings, as the primary aim of this study was to explore the co-design experience rather than to evaluate dissemination outcomes. Furthermore, evaluating the broader impact of CoMICs on students and patients who viewed the content was beyond the scope of this work. Future research [1-7] should focus on guideline adherence, decision-making confidence, and quality of patient care. Several user-generated

XSL•FO
RenderX

insights, such as using the videos as “quick refreshers before clinics” or “during breaks or commutes,” suggest practical pathways for future implementation. These patterns indicate that CoMICs could be integrated into just-in-time learning strategies, on-shift clinical preparation, and student revision routines. Future work should examine how such informal use cases can be systematically incorporated into guideline uptake workflows and clinical education pathways.

Conclusions

CoMICs offer a practical and engaging solution to barriers in guideline dissemination, improving accessibility and relevance through visual storytelling and multilingual adaptation. Importantly, involving medical students in their creation fosters early exposure to academic endocrinology and professional skill development, supporting interest in research and specialty careers. Future work should focus on evaluating clinical outcomes, cost-effectiveness, and integration into wider implementation systems, ensuring scalability across specialties and settings.

Acknowledgments

The authors gratefully acknowledge the DEVI (dissemination, education & value-added implementation) collaboration—Josh Banerjee, Pranav Viswanath Iyer, Vardhan Venkatesh, and Mukunth Kowsik—for their contributions to the production of the guideline Concise Medical Information Cines (CoMICs). They also thank Cristina Ronchi and Leonie van Hulsteijn for their valuable review of the guideline CoMICs during development. They are also grateful to the Department of Applied Health Sciences, School of Health Sciences, College of Medicine and Health, University of Birmingham, for supporting this study. We used the generative artificial intelligence tool (GPT-5.1) [17] to refine the language and grammar during manuscript preparation, which were further reviewed and revised by the study group.

Funding

This research received no particular grant from any funding agency in the public, commercial, or not-for-profit sectors. PK received support from the National Institute for Health and Care Research (NIHR) through his Advanced Clinician Scientist Fellowship (grant NIHR303671). Additionally, he was supported by the Midlands Patient Safety Research Collaboration and the NIHR Race, Equity, and Diversity in Careers Incubator. The views expressed in this study are those of the author and do not necessarily reflect the official positions of the NIHR or the Department of Health and Social Care.

Data Availability

The datasets generated or analyzed during this study are available from the corresponding author on reasonable request.

Authors' Contributions

Conceptualization: PK (lead), RS (equal), AM (supporting).
Data curation: RS (lead), PK (supporting).
Formal analysis: RS (lead), PK (supporting).
Funding acquisition: PK.
Investigation: RS (lead), PK (equal), SR (supporting).
Methodology: RS (lead), PK (equal).
Project administration: RS (lead), PK (equal).
Resources: PK (lead), RS (equal), SR (supporting).
Supervision: PK (lead), AM (supporting).
Validation: RS (lead), PK (equal), SR (supporting), AM (supporting).
Visualization: RS (lead), PK (supporting), SR (supporting).
Writing – original draft: RS (lead).
Writing – review & editing: RS (lead), PK (equal), SR (supporting), AM (supporting).

Conflicts of Interest

None declared.

Multimedia Appendix 1
YouTube hyperlinks to the 4-part glucocorticoid-induced adrenal insufficiency guideline Concise Medical Information Cines series and the 7 multilingual patient versions.
[DOCX File, 15 KB - mededu_v12i1e83711_app1.docx]

References

XSL•FO
RenderX

1. Panteli D, Legido-Quigley H, Reichebner C, Ollenschläger G, Schäfer C, Busse R. Clinical practice guidelines as a quality strategy. In: Busse R, Klazinga N, Panteli D, editors. Improving Healthcare Quality in Europe: Characteristics, Effectiveness and Implementation of Different Strategies: European Observatory on Health Systems and Policies; 2019.
2. McArthur C, Bai Y, Hewston P, Giangregorio L, Straus S, Papaioannou A. Barriers and facilitators to implementing evidence-based guidelines in long-term care: a qualitative evidence synthesis. Implement Sci 2021 Jul 9;16(1):70. [doi: 10.1186/s13012-021-01140-0] [Medline: 34243789]
3. Owoeye OB, Rauvola RS, Brownson RC. Dissemination and implementation research in sports and exercise medicine and sports physical therapy: translating evidence to practice and policy. BMJ Open Sport Exerc Med 2020 Nov 13;6(1):e000974. [doi: 10.1136/bmjsem-2020-000974] [Medline: 33304607]
4. Légaré F, Adekpedjou R, Stacey D, et al. Interventions for increasing the use of shared decision making by healthcare professionals. Cochrane Database Syst Rev 2018 Jul 19;7(7):CD006732. [doi: 10.1002/14651858.CD006732.pub4] [Medline: 30025154]
5. Squires JE, Sullivan K, Eccles MP, Worswick J, Grimshaw JM. Are multifaceted interventions more effective than single-component interventions in changing health-care professionals' behaviours? An overview of systematic reviews. Implement Sci 2014 Oct 6;9:152. [doi: 10.1186/s13012-014-0152-6] [Medline: 25287951]
6. Allen Moore W, Russell Smith A. Effects of video podcasting on psychomotor and cognitive performance, attitudes and study behaviour of student physical therapists. Innov Educ Teach Int 2012;49(4):401-414. [doi: 10.1080/14703297.2012.728876]
7. Elhariry M, Malhotra K, Goyal K, Bardus M, Simba And CoMICs, Kempegowda P. A SIMBA CoMICs initiative to cocreating and disseminating evidence-based, peer-reviewed short videos on social media: mixed methods prospective study. JMIR Med Educ 2024 Oct 30;10:e52924. [doi: 10.2196/52924] [Medline: 39475500]
8. Medical school engagement program. Endocrine Society. 2025. URL: https://www.endocrine.org/our-community/expanding-field/medical-school-engagement-program [accessed 2025-08-18]
9. Venkatesh V, Mathiopoulou M, Kumar K, et al. Multilingual menopause comics: combating misinformation and enhancing global health literacy through peer-reviewed educational videos. Endocr Abstr 2025;110:1076. [doi: 10.1530/endoabs.110.P1076]
10. Braun V, Clarke V. Using thematic analysis in psychology. Qual Res Psychol 2006 Jan;3(2):77-101. [doi: 10.1191/1478088706qp063oa]
11. Young JQ, Van Merrienboer J, Durning S, Ten Cate O. Cognitive Load Theory: implications for medical education: AMEE Guide No. 86. Med Teach 2014 May;36(5):371-384. [doi: 10.3109/0142159X.2014.889290] [Medline: 24593808]
12. van der Wees PJ, Jamtvedt G, Rebbeck T, de Bie RA, Dekker J, Hendriks EJ. Multifaceted strategies may increase implementation of physiotherapy clinical guidelines: a systematic review. Aust J Physiother 2008;54(4):233-241. [doi: 10.1016/s0004-9514(08)70002-3] [Medline: 19025503]
13. Prior M, Guerin M, Grimmer-Somers K. The effectiveness of clinical guideline implementation strategies--a synthesis of systematic review findings. J Eval Clin Pract 2008 Oct;14(5):888-897. [doi: 10.1111/j.1365-2753.2008.01014.x] [Medline: 19018923]
14. Ross-Hellauer T, Tennant JP, Banelytė V, et al. Ten simple rules for innovative dissemination of research. PLoS Comput Biol 2020 Apr;16(4):e1007704. [doi: 10.1371/journal.pcbi.1007704] [Medline: 32298255]
15. Chan WV, Pearson TA, Bennett GC, et al. ACC/AHA special report: clinical practice guideline implementation strategies: a summary of systematic reviews by the NHLBI implementation science work group: a report of the American College of Cardiology/American Heart Association Task Force on Clinical Practice Guidelines. J Am Coll Cardiol 2017 Feb 28;69(8):1076-1092. [doi: 10.1016/j.jacc.2016.11.004] [Medline: 28132746]
16. Lavis JN, Robertson D, Woodside JM, McLeod CB, Abelson J, Knowledge Transfer Study Group. How can research organizations more effectively transfer research knowledge to decision makers? Milbank Q 2003;81(2):221-248. [doi: 10.1111/1468-0009.t01-1-00052] [Medline: 12841049]
17. ChatGPT 5.1. OpenAI. URL: https://openai.com/index/gpt-5-1/ [accessed 2025-05-25]

Abbreviations

CoMICs: Concise Medical Information Cines
GIAI: glucocorticoid-induced adrenal insufficiency

XSL•FO
RenderX

Edited by A Stone; submitted 10.Sep.2025; peer-reviewed by C Provvidenza, J Kaur, MS Shafi; revised version received 22.Dec.2025; accepted 24.Dec.2025; published 27.Jan.2026.

Please cite as:
Sagu R, Ravi S, Manta A, Kempegowda P
Enhancing Engagement With Endocrine Guidelines and Fostering Medical Student Interest Through Concise Medical Information Cines: Qualitative Co-Design Study
JMIR Med Educ 2026;12:e83711
URL: https://mededu.jmir.org/2026/1/e83711
doi:10.2196/83711

© Rahul Sagu, Sangamithra Ravi, Aspasia Manta, Punith Kempegowda. Originally published in JMIR Medical Education (https://mededu.jmir.org), 27.Jan.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Education, is properly cited. The complete bibliographic information, a link to the original publication on https://mededu.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Web-Based Virtual Environment Versus Face-To-Face Delivery for Team-Based Learning of Anesthesia Techniques Among Undergraduate Medical Students: Randomized Controlled Trial

Darunee Sripadungkul[1], MD; Suhattaya Boonmak[1], MD; Monsicha Somjit[1], MD; Narin Plailaharn[1], MD; Wimonrat Sriraj[1], MD; Polpun Boonmak[1], MD

Department of Anesthesiology, Faculty of Medicine, Khon Kaen University, Muang Khon Kaen, Khon Kaen, Thailand

Corresponding Author:
Suhattaya Boonmak, MD
Department of Anesthesiology, Faculty of Medicine
Khon Kaen University
123 Ni Muang
Muang Khon Kaen, Khon Kaen,
Thailand
Phone: 66 813698630
Email: suhabo@kku.ac.th

Abstract

Background: Foundational knowledge of anesthesia techniques is essential for medical students. Team-based learning (TBL) improves engagement. Web-based virtual environments (WBVEs) allow many learners to join the same session in real time while being guided by an instructor.

Objective: This study aimed to compare a WBVE with face-to-face (F2F) delivery of the same TBL curriculum in terms of postclass knowledge and learner satisfaction.

Methods: We conducted a randomized, controlled, assessor-blinded trial at a Thai medical school from August 2024 to January 2025. Eligible participants were fifth-year medical students from the Faculty of Medicine, Khon Kaen University, who attended the anesthesiology course at the department of anesthesiology. Students who had previously completed the anesthesiology course or were unable to comply with the study protocol were excluded. They were allocated to one of the groups using a computer-generated sequence, with concealment of allocation to WBVE (on the Spatial platform) or F2F sessions. Both groups received identical 10-section content in a standardized TBL sequence lasting 130 minutes. Only the delivery mode differed (Spatial WBVE vs classroom F2F). The primary outcome was the postclass multiple-choice questionnaire score. The secondary outcome was learner satisfaction. Individual knowledge was assessed before and after the session using a 15-item questionnaire containing multiple-choice questions via Google Forms. Satisfaction was measured immediately after class on a 5-point Likert scale. Outcome scoring and data analysis were blinded to group assignment. Participants and instructors were not blinded.

Results: In total, 79 students were randomized in this study (F2F: n=38, 48%; WBVE: n=41, 52%). We excluded 2% (1/41) of the students in the WBVE group due to incomplete data. There were complete data for the analysis for 78 participants (F2F: n=38, 49%; WBVE: n=40, 51%). Preclass scores were similar between groups (F2F: mean 6.03, SD 2.05; WBVE: mean 6.20, SD 2.04). Postclass knowledge did not differ significantly (F2F: mean 11.24, SD 1.93; WBVE: mean 10.40, SD 2.62; mean difference 0.88, 95% CI –0.18 to 1.94; *P*=.12). Learner satisfaction favored F2F learning across multiple domains, including overall course satisfaction. Overall satisfaction favored F2F learning (mean difference 0.42, 95% CI 0.07-0.77; *P*=.01). Both groups ran as planned. No adverse events were reported. No technical failures occurred in the WBVE group.

Conclusions: In this trial, WBVE-delivered TBL produced similar short-term knowledge gains to F2F delivery, but learner satisfaction was lower in the WBVE group. Unlike many previous studies, this trial compared WBVE and F2F delivery while keeping the TBL curriculum and prespecified outcomes identical across groups. These findings support WBVEs as a scalable option when physical space, learner volume, or constraints are present. However, lower satisfaction in the WBVE highlights the real-world need for improved facilitation, user experience design, and technical readiness before broader implementation.

Trial Registration: Thai Clinical Trials Registry TCTR20240708012; https://www.thaiclinicaltrials.org/show/TCTR20240708012

(JMIR Med Educ 2026;12:e80097) doi:10.2196/80097

XSL•FO
RenderX

KEYWORDS

anesthesia; computer-assisted instruction; distance; education; internet; learning; medical; problem-based learning; students; teaching; virtual reality

Introduction

Foundational knowledge of anesthesia techniques is important for medical students. Medical students are required to have knowledge of anesthesia administration, pharmacology, procedures, complication management, and interpretation of anesthesia records. Traditional didactic lectures offer a clear, efficient way to organize and deliver this content. It also allows broad coverage of key concepts in a limited time. However, limited interactivity can reduce engagement and encourage passive learning [1]. Interactive lectures address these limitations with small-group discussions, problem-solving exercises, and simulations. These activities create a more active classroom and support clinical reasoning. They also improve knowledge retention and provide immediate feedback to guide learning [2,3].

Team-based learning (TBL) is an effective approach to promote collaborative and active learning. TBL links to constructivist learning theory, the interactive-constructive-active-passive framework, and cognitive load theory. A simple sequence includes individual and team readiness assurance, application exercises, and immediate feedback. Students can move beyond just completing tasks to engaging with the content and peers. Additionally, it encourages a deeper understanding of concepts through discussion. Students can focus on what matters, think more deeply, and retain knowledge longer [4-6]. However, TBL also has its limitations. Challenges include managing group dynamics, resolving conflicts arising from differing opinions, managing time effectively, and distributing the workload. It also requires appropriate activity design, resources, and assessments [3,7,8].

Interactive lectures and TBL are used to facilitate the topic of anesthesia techniques for fifth-year medical students in our department. This learning technique demonstrates motivation, problem-based solving, and communication. Advances in medical education technology, such as web-based virtual environments (WBVEs), can support synchronous interaction, shared workspaces, and simulation-like experiences. WBVEs are virtual worlds that enable real-time interactions between users and digital objects through technologies such as virtual and augmented reality. They provide immersive experiences, facilitate collaboration, and support interactive and realistic simulations [9-13]. Compared with conventional TBL, WBVEs may make classes easier to join, easier to expand to more students, and more supportive of group work [7,12,14].

However, comparative evidence remains limited. Existing studies are constrained by heterogeneous reporting, small sample sizes, and nonrandomized designs, particularly in anesthesia education [15-17]. Moreover, learners' diverse preferences and learning styles should be considered when implementing WBVEs [18]. These limitations justify a randomized trial that isolates delivery mode by comparing WBVE with face-to-face (F2F) instruction under an equivalent TBL design. We conducted a randomized controlled trial comparing WBVE (on the Spatial platform [19]) versus F2F delivery of the same TBL curriculum in anesthesia techniques among fifth-year medical students. Both groups used identical objectives, materials, facilitation, and assessments. Only the delivery mode differed. The primary outcome was postclass knowledge, and the secondary outcome was learner satisfaction. We hypothesized that there would be between-group differences in postclass knowledge and learner satisfaction.

Methods

Study Design

This study was a randomized, controlled, single-blinded trial. Double blinding was not feasible as participants were aware of their assigned groups. This study adhered to the CONSORT-EHEALTH (Consolidated Standards of Reporting Trials of Electronic and Mobile Health Applications and Online Telehealth) guidelines [20] to ensure comprehensive and transparent reporting of randomized controlled trial data.

Participants

We recruited fifth-year medical students from the Faculty of Medicine, Khon Kaen University, who attended the anesthesiology course at the department of anesthesiology. Students who had previously completed the anesthesiology course or were unable to comply with the study protocol were excluded. All participants provided written informed consent. The study was conducted from August 26, 2024, to January 7, 2025. Withdrawal criteria included technical issues, such as internet disruptions or software glitches on the Spatial platform [19].

Randomization and Recruitment

We used cluster randomization at the class level to minimize contamination between teaching groups. Students were grouped into clusters based on their scheduled learning sessions. Each cluster comprised approximately 20 students. Clusters were then randomly allocated in a 1:1 ratio to either the WBVE or F2F group using computer-generated random numbers. We used a block size of 2 to ensure that each group had a similar number of clusters. The randomization sequence was generated and implemented by investigators who were not involved in teaching or assessment to reduce the risk of allocation bias. For clusters randomized to the WBVE group, students who declined research participation received F2F teaching delivered by another instructor, scheduled at the same time and following an identical teaching plan. For clusters randomized to the F2F group, students who declined research participation attended the same F2F session with the same instructor but were not asked to complete the posttest or satisfaction questionnaire.

XSL•FO
RenderX

Interventions

Preintervention Process

After recruitment into the study, all participants in both groups were given access to the learning material via online learning in Google Classroom, including a slide presentation and supplementary video material, 2 days before the teaching session. On the teaching day, students first received a 5-minute briefing about study procedures and learning objectives. They then completed a questionnaire containing 15 multiple-choice questions (MCQs) before the test in 15 minutes to assess baseline knowledge. Participants also completed a brief questionnaire capturing student characteristics. The test specification table is provided in Multimedia Appendix 1.

Session Allocation and Team Setup

After baseline assessments, students attended their assigned session (either a WBVE or an F2F session). In each group, students were organized into 4 TBL teams of 5 to 6 students, supervised by 1 instructor. Each session lasted 130 minutes.

Content and Structure Are Common to Both Groups

Both groups participated in a single, standardized 130-minute TBL session. The content was identical, organized into 10 sections on anesthesia techniques (listed in Multimedia Appendix 1). Each section began with core knowledge, followed by a problem-solving scenario aligned with the same learning objectives. Students progressed sequentially through the modules to complete the required tasks.

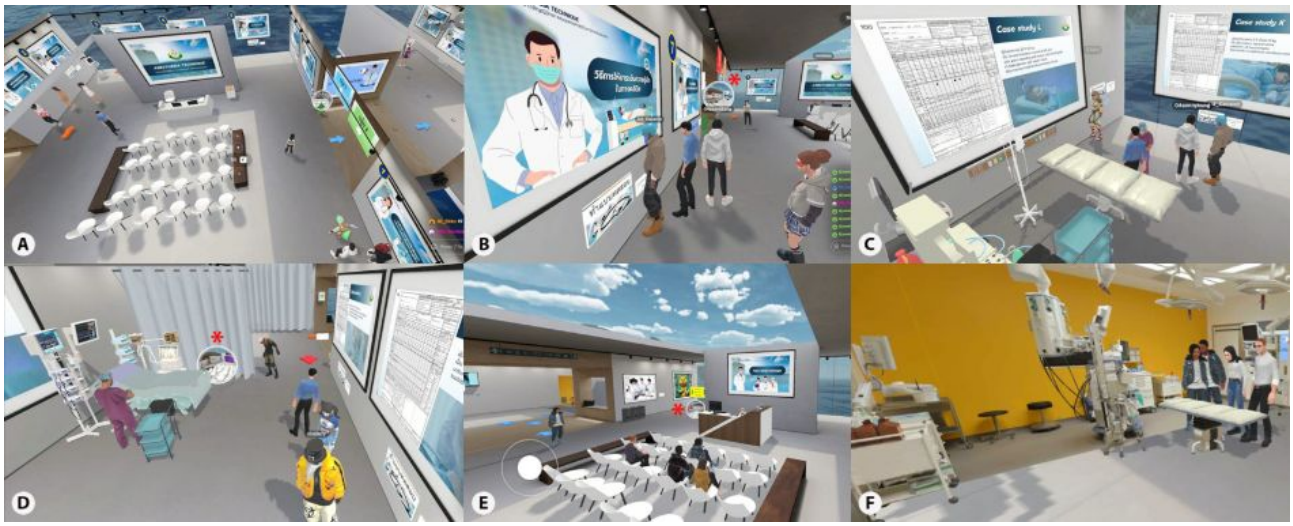
F2F Group

The 10 modules described earlier were delivered F2F using a standard TBL sequence. At the start of the session, the instructor reviewed the learning objectives and ground rules and confirmed team assignments. For each module (Multimedia Appendix 1), students first received a brief minilecture to consolidate core concepts. Subsequently, teams completed a scenario-based application task at their respective tables. They recorded a single team answer. They received immediate feedback, followed by a short whole-class debrief that compared rationales across teams. The room layout comprised 4 separated team tables with a central screen, and teams were asked to discuss quietly to minimize cross talk. Materials were identical to those in the other group (handouts mirroring the slides and identical question stems). Any clarifying questions were addressed at the table before the debrief.

WBVE Group

Sessions ran on the Spatial platform (Spatial Systems, Inc; Thai localization) [19] using desktop computers on the university's secure local network. On first log-in, students created an account and entered a virtual space with 3 areas (a classroom, an operating room, and a common room). Each was equipped with detailed 3D anesthesia equipment to support the team's application tasks. The same 10 interactive modules described earlier were delivered in sequence. Each module began with brief core content, followed by a scenario-based application task aligned with the learning objectives. Students moved through the environment with keyboard-controlled avatars and collaborated in real time via built-in voice and text chat. Embedded multimedia (short videos, slide decks, interactive cases, and simple educational games) supported engagement. The instructor monitored team rooms, offered real-time guidance, provided immediate feedback, and answered questions as they arose. Any clarifying questions were addressed in the common room before the debrief. A detailed description of the WBVE is provided in Figure 1.

Figure 1. The web-based virtual environment ran on the Spatial platform: (A) A classroom. Students learned modules 1 to 7 in this classroom, organized into 4 teams of 5 to 6 students each, supervised by 1 instructor. (B) A classroom. Students navigated the environment using keyboard-controlled avatars and collaborated in real time via built-in voice and text chat. Students moved to the operating room by clicking the button (*). (C) An operating room. Students learned modules 8 and 9 in this room, including how to read anesthetic record data. Each group also discussed 4 case studies for team-based learning. (D) An operating room. After finishing the questions, students moved to the common room by clicking the button (*). (E) A common room. Students in each group asked clarifying questions in this room, after which the instructor conducted a debriefing for module 10. If students wanted to review the material again or go to an operating room or a classroom, they could return by clicking the button (*). (F) A common room. A simulated operating room had been created to allow students to learn in a realistic virtual environment before entering the actual operating room. If students had extra time while waiting for their classmates, they could explore this room. Various games were also available to help train their minds.



XSL•FO
RenderX

Postintervention Process

Immediately after the teaching session, 15 minutes were allocated for outcome measurement. All students completed a questionnaire containing 15 MCQs after the test via Google Forms, using the exact test specification as before the test, under invigilated conditions. They then completed a satisfaction questionnaire in Google Forms (5-point Likert scale; instrument and scoring details are provided in Multimedia Appendix 1). The forms were identical for both groups and were accessed through locked links. Answer keys were withheld until all submissions had been received. In addition, we enabled automatic scoring and time stamp logging, restricted responses to a single submission per account, and exported raw data directly from Google Forms into the analysis dataset.

Data Collection

Pre- and postintervention knowledge evaluations were assessed using a 15-item MCQ test mapped to a test specification table (Multimedia Appendix 1). The authors designed and developed a questionnaire on students' satisfaction with the learning process, drawing on previous studies [21,22], and adjusted its content for use in a Thai context (Multimedia Appendix 1). Student satisfaction was evaluated immediately after the intervention using a 21-item survey, with responses on a 5-point Likert scale (1=strongly disagree, 2=disagree, 3=neither agree nor disagree, 4=agree, and 5=strongly agree). The questionnaire's content validity was confirmed by 3 experts. Internal consistency was excellent, with a Cronbach α of 0.95. We also recorded gender, age, grade point average, and experience with WBVEs (including frequency, proficiency, and comfort).

Statistical Analysis

Data were analyzed using Stata/SE (version 18.0; StataCorp) for Windows. Descriptive statistics summarized participant characteristics. Categorical variables were presented as counts and percentages, and continuous variables were presented as means and SDs. Proportions were calculated using nonmissing denominators. Analyses followed the intention-to-treat principle. Because only 1 (1%) of the 79 participants had missing primary outcome data, analyses were conducted using complete-case analysis. Missing completely at random testing and multiple imputation were not performed due to negligible missingness. Between-group comparisons were conducted using linear mixed modeling. We reported the mean differences and 95% CIs. For the primary outcome (knowledge), we used posttest score as the dependent variable, study group as the factor, and pretest score as a covariate. For multiple tests that were analyzed, the false discovery rate was controlled at 5% (Benjamini-Hochberg method). Item-level Likert responses were summarized descriptively as means (SDs).

The required sample size was calculated using the postlearning knowledge score from 63 fifth-year medical students who studied the anesthesia technique in the previous academic year (2023), with a mean of 57.94% (SD 14.07). To detect a 10% difference in knowledge score (type I error of 0.05) with 80% power, we determined that a sample size of 32 participants in each group would be required. The dropout rate was accounted for in the sample size, which included at least 40 participants per group, representing approximately 10% of the total dropout rate.

Ethical Considerations

This study received full board approval from the human research ethics committee of Khon Kaen University (HE671294). It was also registered with the Thai Clinical Trials Registry before participant enrollment (TCTR20240708012). This study was conducted in accordance with the World Medical Association Declaration of Helsinki and institutional policies. Participants received clear written information about the study objectives, procedures, potential risks, and data management before participating. Participation was voluntary, and they could withdraw at any time without affecting their academic standing. Those who agreed to take part then provided written informed consent. Study-related data were not collected from them. After participating in this study, students created unique codes instead of using their names, which enabled us to link pre- and posttest data without knowing individual identities to protect privacy and confidentiality. Data were anonymized for analysis and stored on password-protected systems with access restricted to the investigator's team. Students did not receive any financial or in-kind compensation.

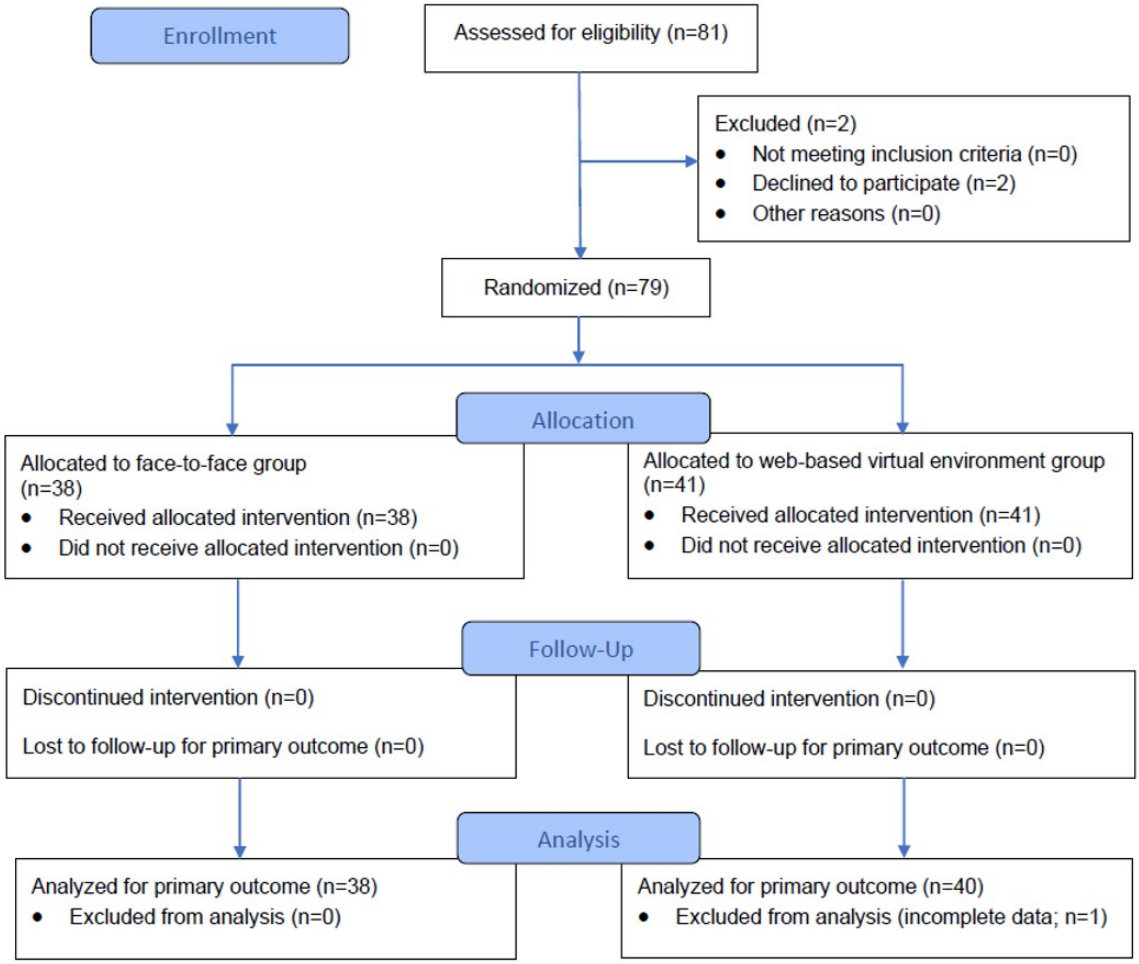
Results

Overview

A total of 81 participants were assessed for eligibility; 2 (2%) participants declined to participate. Both were in clusters randomized to the F2F group; therefore, they received F2F teaching as part of the regular course. However, their data were not included in the analysis. Hence, 79 participants were enrolled in the study and allocated into 2 groups: the F2F group (n=38, 48%) and the WBVE group (n=41, 52%); 2% (1/41) of the participants in the WBVE group had incomplete outcome data and were excluded. A total of 78 (99%) participants had complete data in the analysis (Figure 2).

XSL•FO
RenderX

Figure 2. Flow diagram of the participants in the study.



Participant Characteristics

As shown in Table 1, 78 participants were included, with 38 (49%) assigned to the F2F group and 40 (51%) to the WBVE group. The groups were comparable in gender, age, and grade point average. The proportion of male participants was slightly higher in the F2F group (25/38, 66%) compared to the WBVE group (21/40, 52%). Most participants in both groups reported limited previous exposure to WBVEs, with more than 60% (51/78) of the participants classified as novices and a few reporting intermediate or higher proficiency. Comfort levels with the technology were predominantly neutral.

XSL•FO
RenderX

Table 1. Baseline characteristics by group in a cluster-randomized trial of a face-to-face (F2F) group versus a web-based virtual environment (WBVE) group for team-based learning among fifth-year medical students (N=78).

	F2F group (n=38)	WBVE group (n=40)
Male, n (%)	25 (66)	21 (52)
Age (y), mean (SD)	22.5 (0.6)	22.5 (0.7)
Grade point average (0-4), mean (SD)	3.50 (0.34)	3.44 (0.33)
Experience with WBVEs, n (%)		
Frequency		
No	19 (50)	26 (65)
Sometimes	18 (47)	14 (35)
Frequently	1 (3)	0 (0)
Proficiency		
Novice	24 (63)	27 (67)
Beginner	10 (26)	7 (17)
Intermediate	4 (11)	5 (13)
Advanced	0 (0)	1 (3)
Expert	0 (0)	0 (0)
Comfort level		
Discomfort	0 (0)	2 (5)
Neutral	27 (71)	27 (67)
Comfort	11 (29)	11 (28)

Knowledge Outcomes

As shown in Table 2, both groups demonstrated significant improvement in knowledge after the intervention. The WBVE group had slightly higher mean pretest scores than the F2F group (mean 6.20, SD 2.04 vs mean 6.03, SD 2.05 for WBVE vs F2F), but lower posttest scores (mean 10.40, SD 2.62 vs mean 11.24, SD 1.93 for WBVE vs F2F). However, after adjustment using linear mixed modeling and the Benjamini-Hochberg procedure, no statistically significant differences between groups were observed in either pretest or posttest scores (P=.70 and P=.12, respectively).

XSL•FO
RenderX

Table 2. Knowledge (15-point scale) and learner satisfaction (Likert 1-5) by group in a cluster-randomized trial comparing face-to-face (F2F) versus web-based virtual environment (WBVE) team-based learning among fifth-year medical students (N=78).

	F2F group (n=38), mean (SD)	WBVE group (n=40), mean (SD)	Mean difference (95% CI)	*P* value
Knowledge (15-point scale)[a]				
Pretest knowledge score	6.03 (2.05)	6.20 (2.04)	–0.17 (–1.07 to 0.72)	.70
Posttest knowledge score	11.24 (1.93)	10.40 (2.62)	0.88 (–0.18 to 1.94)	.12
Learner satisfaction (Likert scale 1-5)[b]				
Learning topic				
Intellectually challenging and stimulating	4.7 (0.5)	4.4 (0.6)	0.34 (0.05 to 0.62)	.02[c]
Valuable learning gained	4.9 (0.3)	4.6 (0.6)	0.29 (0.04 to 0.54)	.02[c]
Increased interest in the subject	4.6 (0.7)	4.5 (0.7)	0.10 (–0.21 to 0.41)	.53
Learning process				
Instructor				
Instructor enthusiasm	4.9 (0.3)	4.7 (0.6)	0.22 (–0.01 to 0.45)	.07
Presentation style holds interest	4.8 (0.6)	4.4 (0.9)	0.36 (–0.01 to 0.74)	.06
Clear explanations	4.8 (0.5)	4.3 (0.8)	0.46 (0.08 to 0.85)	.01[c]
Facilitates note-taking	4.6 (0.7)	4.4 (1.0)	0.28 (–0.11 to 0.67)	.17
Friendly to students	4.9 (0.2)	4.9 (0.4)	0.10 (–0.06 to 0.25)	.25
Welcomes questions or help	4.9 (0.3)	4.8 (0.5)	0.12 (–0.07 to 0.31)	.25
Genuine interest in students	4.9 (0.3)	4.8 (0.5)	0.07 (–0.10 to 0.24)	.46
Presents alternative viewpoints	4.8 (0.4)	4.6 (0.6)	0.22 (–0.04 to 0.48)	.11
Technique				
Well-prepared materials	4.8 (0.4)	4.3 (1.0)	0.49 (0.08 to 0.90)	.01[c]
Encourages class discussion	4.9 (0.4)	4.1 (1.0)	0.77 (0.27 to 1.27)	<.001[c]
Invites the sharing of ideas	4.8 (0.4)	3.9 (1.1)	0.92 (0.34 to 1.50)	<.001[c]
Encourages and answers questions	4.9 (0.4)	4.6 (0.7)	0.32 (0.04 to 0.60)	.02[c]
Supports student expression	4.9 (0.3)	4.5 (0.8)	0.42 (0.07 to 0.76)	.009[c]
Technical satisfaction (audio, video, and media)	4.7 (0.5)	4.4 (0.8)	0.28 (–0.06 to 0.62)	.11
Satisfaction with teaching methods	4.9 (0.3)	4.3 (0.9)	0.54 (0.12 to 0.97)	.004[c]
Learning outcome				
Content covers objectives	4.9 (0.3)	4.7 (0.5)	0.27 (0.05 to 0.50)	.01[c]
Consistent with current knowledge	4.9 (0.4)	4.7 (0.5)	0.22 (–0.02 to 0.45)	.07
Overall course satisfaction	4.9 (0.4)	4.5 (0.8)	0.42 (0.07 to 0.77)	.01[c]

[a]Knowledge scores were on a 15-point scale (0-15); higher scores indicated greater knowledge.

[b]Likert score (1=very dissatisfied, 2=dissatisfied, 3=neither satisfied nor dissatisfied, 4=satisfied, and 5=very satisfied).

[c]Statistically significant difference at $P<.05$.

Learner Satisfaction

As shown in Table 2, learner satisfaction across multiple domains was generally high in both groups. However, several aspects favored the F2F group. For the learning topic, the F2F group reported higher levels of intellectual challenge and stimulation (mean difference 0.34; 95% CI 0.05-0.62; $P=.02$) as well as greater perceived value in learning gained (mean difference 0.29; 95% CI 0.04-0.54; $P=.02$). For the learning

XSL•FO
RenderX

process, the evaluation was divided into instructor- and technique-related components. The F2F group rated significantly higher on clear explanations (mean difference 0.46; 95% CI 0.08-0.85; *P*=.01) on instructor-related items. It also rated significantly higher on well-prepared materials (mean difference 0.49; 95% CI 0.08-0.90; *P*=.01), encouragement of class discussion (mean difference 0.77; 95% CI 0.27-1.27; *P*<.001), an invitation to share ideas (mean difference 0.92; 95% CI 0.34-1.50; *P*<.001), encouragement and answering questions (mean difference 0.32; 95% CI 0.04-0.60; *P*=.02), support for student expression (mean difference 0.42; 95% CI 0.07-0.76; *P*=.009), and satisfaction with teaching methods (mean difference 0.54; 95% CI 0.12-0.97; *P*=.004). In terms of learning outcomes, the F2F group rated significantly higher on content coverage (mean difference 0.27; 95% CI 0.05-0.50; *P*=.01). While overall course satisfaction was significantly higher in the F2F group mean 4.9, SD 0.4 vs mean 4.5, SD 0.8; mean difference 0.42; 95% CI 0.07-0.77; *P*=.01.

Harms

Neither group experienced adverse events, privacy breaches, or unintended effects. The WBVE session ran as planned, and no technical issues (audio or latency) were observed. There were no withdrawals due to technical problems.

Discussion

Principal Findings

This study compared knowledge and learner satisfaction between F2F and WBVE learning for anesthesia techniques. Both groups were comparable in terms of characteristics, including their experience with WBVEs. All participants showed improvement in knowledge outcomes. However, posttest scores did not show statistically significant differences between groups. Learner satisfaction was generally high in both groups, while it was consistently higher in the F2F group. The F2F group reported significantly higher satisfaction with the learning topic, the learning process (related to learning techniques), and the learning outcome. Overall course satisfaction was higher in the F2F group.

Participant Characteristics

The baseline characteristics of participants in both groups were comparable. There were no significant differences in gender, age, or academic performance. Most participants in both groups reported limited previous experience with WBVEs, with more than 60% (51/78) of the participants classified as novices. Most (54/78, 69%) participants reported neutral comfort ratings. This suggests that students were not very confident with WBVEs but were not uncomfortable with them. This limited exposure may have influenced both their comfort level and engagement with the WBVEs [23,24]. Previous studies have reported that WBVEs are not an obstacle to learning for medical students with limited experience with them [23-25]. The findings indicate that additional training and familiarization sessions with WBVEs may be necessary to enhance learner confidence and optimize engagement [23-25].

Knowledge Outcome

Pre- and posttest results showed improvement in both groups. Posttest scores also did not show statistically significant differences between groups. This suggests that both the WBVE and F2F delivery are effective for teaching. This is consistent with previous studies demonstrating the potential of WBVEs to facilitate cognitive learning and performance in medical students [25-29]. Despite technical and practical challenges, the use of WBVEs can improve knowledge and performance [25-29].

Learner Satisfaction

Our data reported that overall satisfaction was high in both groups. However, across multiple domains of learner satisfaction, the F2F instructional modality was favored. The F2F group reported significantly higher ratings for intellectual stimulation and for the value of what they learned. They also felt that instructors provided clearer explanations and more effectively encouraged idea sharing, questioning, answering, and student expression. Our data showed the benefit of the F2F format in promoting effective engagement, interactive learning dynamics, and real-time interpersonal communication. These findings suggest that conducting TBL in WBVEs is faced with several limitations. Technical constraints and students' limited familiarity with WBVE may affect learner satisfaction [3,23,24]. Additionally, the instructor's facilitation techniques during WBVE should be a concern [23,24].

Cultural context may help explain the differences in satisfaction [29-32]. However, it should be viewed as a possible moderator rather than a fixed or deterministic cause. Thai medical students may defer to instructors and be comfortable with structured guidance and supervision. This may favor F2F settings, where nonverbal cues and immediate instructor feedback are more noticeable. Consistent with previous work in Asia, students have reported greater satisfaction with traditional methods [30]. This appears to be influenced by personality, self-efficacy, and their expectations about instructional design and delivery [30]. In contrast, studies and systematic reviews from Europe and the United States have often found higher satisfaction with WBVEs [29,31]. However, cultural context does not uniformly predict outcomes. Our data and previous studies observed variability across participants that appears to depend on facilitation quality, previous exposure, technology self-efficacy, language demands, and device access [29-32]. During the COVID-19 pandemic, Thai medical students adapted quickly [32]. They reported greater satisfaction with new anesthesiology learning technologies than their teachers. These suggest that growing familiarity can narrow culture-based gaps over time [32]. We view cultural context as one of several interacting factors that affect learner satisfaction.

The variability in learner satisfaction likely reflects implementation issues rather than only the delivery method [29-32]. In WBVEs, instructors cannot rely on being physically in the room, so they must think more carefully about how they facilitate the learning process [10,12,29]. It is essential to use strategies that build social presence and keep students engaged [10,12,29]. Additionally, technical issues, such as latency, audio problems, and complex navigation, can create more burdens

XSL•FO
RenderX

[10,12]. When the sound is delayed, the flow of discussion is broken. It becomes hard to know when to speak or respond [10,12]. Complex navigation may be engaging at first, but students must spend time figuring out how to move between virtual rooms or use the interface. They end up focusing on the platform rather than the TBL tasks and the content itself [2,10,12]. Students' satisfaction is likely to depend on how well learning is facilitated [10,12]. Reliable devices and technical support should be included [10,12]. Therefore, future work should concentrate on strengthening facilitation and improving the user experience.

Implications

Our findings indicate that WBVEs can yield knowledge gains comparable to those of F2F instruction. WBVEs may be especially valuable in settings with limited classroom space, tight schedules, or large cohorts [10,12,16]. However, running TBL in WBVEs requires careful design and facilitation to keep students engaged and improve their satisfaction [2,10,12,29]. Specifically, this includes clearer turn taking, more visible feedback, simpler navigation within the virtual environment, and fewer technical obstacles [2,10,12,29]. This study's main contribution is to separate delivery mode from pedagogy. We directly compared a WBVE with F2F delivery under an identical TBL design using assessor-blinded, prespecified outcomes. These design features extend previous work, which often combined different instructional methods or did not report use and outcome measures as clearly [15-17]. This approach can be implemented with medical students, especially in anesthesiology education.

Limitations

This study has several limitations. A single-center study may limit the generalizability of the findings. The relatively small sample size may limit their applicability to broader and more diverse student populations. Next, a double-masked study was not feasible due to the nature of the intervention. This may have introduced performance or response bias. Additionally, participants had limited previous exposure to WBVEs. This may have dampened their engagement and satisfaction. Cultural factors specific to Thai medical students should be a concern. A preference for structured, instructor-led learning may have contributed to the lower satisfaction reported with WBVE.

Future Directions

Future research should include larger sample sizes and multicenter trials. The generalizability of the findings across diverse learner populations should be of concern. In addition, future studies should examine how students' familiarity with WBVEs affects their satisfaction and learning outcomes. They should test whether repeated exposure helps students feel more comfortable and engaged. Moreover, researchers should focus on facilitation techniques to identify the most effective approach for WBVE-based TBL. Simulation-based features within WBVEs can make the environment more interactive and better aligned with students' learning needs. Further work should also study instructional designs tailored to Asian medical education. These designs should incorporate strategies that support collaborative learning and respect cultural norms.

Prior Presentation

Part of this study was previously presented as an oral presentation at the SANCON-ASPA 2025 conference—Scaling New Heights in Pediatric Anesthesia and Beyond (24th Annual Conference of the Society of Anesthesiologists of Nepal [SANCON] and 21st Meeting of the Asian Society of Pediatric Anesthesiologists [ASPA]), held in Kathmandu, Nepal, on April 5, 2025. The abstract was presented under the title "Metaverse versus in-person: impact on anesthesia education and student satisfaction."

Conclusions

This trial is innovative in isolating the delivery mode from pedagogy by directly comparing WBVE to F2F delivery within an identical TBL structure with prespecified end points. This design differs from much of the existing literature, in which shifts in instructional strategy often accompany technological changes. It is challenging to attribute effects solely to the delivery mode. Our findings contribute evidence that a WBVE can deliver comparable short-term knowledge gains in a single standardized session. It has practical implications for scaling instruction when classroom space, faculty time, or scheduling is limited. Nevertheless, lower satisfaction in WBVE underscores the need for real-world implementation. We require stronger facilitation strategies, improved social presence, simplified navigation, and robust technical support to achieve an experience comparable to F2F.

Acknowledgments

The authors would like to thank Kaewjai Thepsuthammarat of the clinical epidemiology unit for her expert biostatistical assistance and Suwitcha Sripadungkul for his support and encouragement. Artificial intelligence–based software (Grammarly) was used solely for grammar and language editing. Professional language editors at the Language Institute of Khon Kaen University also reviewed the final manuscript. No artificial intelligence tools were used for data analysis, content generation, or interpretation. The authors remain fully responsible for the accuracy and scientific content.

Funding

This study was funded by the Faculty of Medicine, Khon Kaen University, Thailand (grant IN67086). The funder had no role in the study design, data collection, analysis, interpretation, or manuscript writing.

XSL•FO
RenderX

Data Availability

The datasets generated or analyzed during this study are available from the corresponding author on reasonable request.

Authors' Contributions

Conceptualization: DS, SB, PB
Data curation: DS, SB, MS, NP, PB
Formal analysis: DS, SB, PB
Investigation: DS, SB, MS, NP, PB
Methodology: DS, SB, MS, NP, WS, PB
Project administration: DS, SB, MS, NP, WS, PB
Supervision: DS, SB, MS, NP, WS, PB
Visualization: DS, SB, MS, NP, WS, PB
Writing—original draft: DS, SB, MS, NP, WS, PB
Writing—review and editing: DS, SB, MS, NP, WS, PB

Conflicts of Interest

None declared.

Multimedia Appendix 1
Instructional package: lesson plan, user guide, and assessment instruments.
[DOCX File , 7270 KB - mededu_v12i1e80097_app1.docx]

Multimedia Appendix 2
CONSORT-eHEALTH (V 1.6.1).
[PDF File (Adobe PDF File), 409 KB - mededu_v12i1e80097_app2.pdf]

References

1. Solymos O, O'Kelly P, Walshe CM. Pilot study comparing simulation-based and didactic lecture-based critical care teaching for final-year medical students. BMC Anesthesiol 2015 Oct 21;15(1):153 [FREE Full text] [doi: 10.1186/s12871-015-0109-6] [Medline: 26490826]
2. Anas S, Kyrou I, Rand-Weaver M, Karteris E. The effect of online and in-person team-based learning (TBL) on undergraduate endocrinology teaching during COVID-19 pandemic. BMC Med Educ 2022 Mar 22;22(1):120 [FREE Full text] [doi: 10.1186/s12909-022-03173-5] [Medline: 35193577]
3. Liu H, Mi XF, Huang ZZ, Heng BC, Shen WL. Challenges and strategies in developing team-based learning in Chinese medical education. Med Teach 2020 Aug 08;42(11):1243-1249. [doi: 10.1080/0142159x.2020.1801995]
4. Wekerle C, Daumiller M, Janke S, Dickhäuser O, Dresel M, Kollar I. Putting ICAP to the test: how technology-enhanced learning activities are related to cognitive and affective-motivational learning outcomes in higher education. Sci Rep 2024 Jul 15;14(1):16295 [FREE Full text] [doi: 10.1038/s41598-024-66069-y] [Medline: 39009608]
5. Szulewski A, Howes D, van Merriënboer JJ, Sweller J. From theory to practice: the application of cognitive load theory to the practice of medicine. Acad Med 2021 Jan 01;96(1):24-30. [doi: 10.1097/ACM.0000000000003524] [Medline: 32496287]
6. Roossien L, van Klaveren LM, de Vos R, Dolmans DH, Boerboom TB. Team-based learning (TBL) in health professions education: a systematic review on its conceptual elements and outcomes. Med Teach 2025 May 28;47(12):1933-1947. [doi: 10.1080/0142159x.2025.2506723]
7. Mulugeta H, Zemedkun A. Implementation of team-based learning for a clinical module of the Ethiopian undergraduate anesthesia curriculum and students' perspectives: a pilot cross-sectional study. Adv Med Educ Pract 2023;14:1413-1424 [FREE Full text] [doi: 10.2147/AMEP.S437710] [Medline: 38116477]
8. Ho CL. Team-based learning: a medical student's perspective. Med Teach 2019 Oct 01;41(9):1087. [doi: 10.1080/0142159X.2018.1555370] [Medline: 30707851]
9. Wang Y, Li C, Qu L, Cai H, Ge Y. Application and challenges of a metaverse in medicine. Front Robot AI 2023 Dec 11;10:1291199 [FREE Full text] [doi: 10.3389/frobt.2023.1291199] [Medline: 38152305]
10. Lewis KO, Popov V, Fatima SS. From static web to metaverse: reinventing medical education in the post-pandemic era. Ann Med 2024 Dec;56(1):2305694 [FREE Full text] [doi: 10.1080/07853890.2024.2305694] [Medline: 38261592]
11. Seddon I, Rosenberg E, Houston 3rd SK. Future of virtual education and telementoring. Curr Opin Ophthalmol 2023 May 01;34(3):255-260. [doi: 10.1097/ICU.0000000000000945] [Medline: 36995108]
12. Ahuja AS, Polascik BW, Doddapaneni D, Byrnes ES, Sridhar J. The digital metaverse: applications in artificial intelligence, medical education, and integrative health. Integr Med Res 2023 Mar;12(1):100917 [FREE Full text] [doi: 10.1016/j.imr.2022.100917] [Medline: 36691642]

XSL•FO
RenderX

13. Bhugaonkar K, Bhugaonkar R, Masne N. The trend of metaverse and augmented and virtual reality extending to the healthcare system. Cureus 2022 Oct;14(9):e29071 [FREE Full text] [doi: 10.7759/cureus.29071] [Medline: 36258985]
14. MacNeill H, Masters K, Nemethy K, Correia R. Online learning in health professions education. Part 1: teaching and learning in online environments: AMEE guide no. 161. Med Teach 2024 Jan 24;46(1):4-17 [FREE Full text] [doi: 10.1080/0142159X.2023.2197135] [Medline: 37094079]
15. Ichikura K, Watanabe K, Moriya R, Chiba H, Inoue A, Arai Y, et al. Online vs. face-to-face interactive communication education using video materials among healthcare college students: a pilot non-randomized controlled study. BMC Med Educ 2024 Jul 10;24(1):746. [doi: 10.1186/S12909-024-05742-2]
16. Tokalić R, Poklepović Peričić T, Marušić A. Similar outcomes of web-based and face-to-face training of the GRADE approach for the certainty of evidence: randomized controlled trial. J Med Internet Res 2023 Jul 06;25:e43928 [FREE Full text] [doi: 10.2196/43928] [Medline: 37279050]
17. Ashokka B, Law LS, Areti A, Burckett-St Laurent D, Zuercher RO, Chin K, et al. Educational outcomes of simulation-based training in regional anaesthesia: a scoping review. Br J Anaesth 2025 Mar;134(2):523-534. [doi: 10.1016/j.bja.2024.07.037] [Medline: 39358185]
18. Chinnapun D, Narkkul U. Enhancing learning in medical biochemistry by teaching based on VARK learning style for medical students. Adv Med Educ Pract 2024 Sep;15:895-902 [FREE Full text] [doi: 10.2147/AMEP.S472532] [Medline: 39359490]
19. Spatial Systems. URL: https://www.spatial.io [accessed 2024-04-10]
20. Eysenbach G, CONSORT-EHEALTH Group. CONSORT-EHEALTH: improving and standardizing evaluation reports of Web-based and mobile health interventions. J Med Internet Res 2011 Dec 31;13(4):e126 [FREE Full text] [doi: 10.2196/jmir.1923] [Medline: 22209829]
21. Corbalan M, Plaza I, Hervas E, Aldabas E, Arcega F. Reduction of the students' evaluation of education quality questionnaire. In: Proceedings of the 2013 Federated Conference on Computer Science and Information Systems. 2013 Presented at: FCCSIS '13; September 8-11, 2013; Krakow, Poland p. 659-659 URL: https://ieeexplore.ieee.org/document/6644083
22. Fatani TH. Student satisfaction with videoconferencing teaching quality during the COVID-19 pandemic. BMC Med Educ 2020 Oct 31;20(1):396 [FREE Full text] [doi: 10.1186/s12909-020-02310-2] [Medline: 33129295]
23. Elston P, Canale GP, Ail G, Fisher N, Mahendran M. Twelve tips for teaching in virtual reality. Med Teach 2024 May 25;46(4):495-499. [doi: 10.1080/0142159X.2023.2285396] [Medline: 38006603]
24. Hussain Z, Ng DM, Alnafisee N, Sheikh Z, Ng N, Khan A, et al. Effectiveness of virtual and augmented reality for improving knowledge and skills in medical students: protocol for a systematic review. BMJ Open 2021 Aug 16;11(8):e047004 [FREE Full text] [doi: 10.1136/bmjopen-2020-047004] [Medline: 34400451]
25. Rodriguez-Florido M, Reyes-Cabrera JJ, Melián A, Hernández-Flores CN, Ruiz-Azola J, Maynar M. Feasibility of teaching and assessing medical students in the metaverse: design and features for its learning efficiency. J New Approaches Educ Res 2024 May 06;13(1):9. [doi: 10.1007/S44322-024-00009-6]
26. Plotzky C, Loessl B, Kuhnert B, Friedrich N, Kugler C, König P, et al. My hands are running away - learning a complex nursing skill via virtual reality simulation: a randomised mixed methods study. BMC Nurs 2023 Jul 27;22(1):222 [FREE Full text] [doi: 10.1186/s12912-023-01384-9] [Medline: 37370124]
27. Schoeb DS, Schwarz J, Hein S, Schlager D, Pohlmann PF, Frankenschmidt A, et al. Mixed reality for teaching catheter placement to medical students: a randomized single-blinded, prospective trial. BMC Med Educ 2020 Dec 16;20(1):510 [FREE Full text] [doi: 10.1186/s12909-020-02450-5] [Medline: 33327963]
28. Liu S, Xie M, Zhang Z, Wu X, Gao F, Lu L, et al. A 3D hologram with mixed reality techniques to improve understanding of pulmonary lesions caused by COVID-19: randomized controlled trial. J Med Internet Res 2021 Oct 10;23(9):e24081 [FREE Full text] [doi: 10.2196/24081] [Medline: 34061760]
29. Li Q, Duan H, Zhou X, Sun X, Tao L, Lu X. The use of metaverse in medical education: a systematic review. Clin Med (Lond) 2025 May;25(3):100315 [FREE Full text] [doi: 10.1016/j.clinme.2025.100315] [Medline: 40246026]
30. Elshami W, Taha MH, Abuzaid M, Saravanan C, Al Kawas S, Abdalla ME. Satisfaction with online learning in the new normal: perspective of students and faculty at medical and health sciences colleges. Med Educ Online 2021 Dec 11;26(1):1920090 [FREE Full text] [doi: 10.1080/10872981.2021.1920090] [Medline: 33974523]
31. Birrenbach T, Zbinden J, Papagiannakis G, Exadaktylos AK, Müller M, Hautz WE, et al. Effectiveness and utility of virtual reality simulation as an educational tool for safe performance of COVID-19 diagnostics: prospective, randomized pilot trial. JMIR Serious Games 2021 Oct 08;9(4):e29586 [FREE Full text] [doi: 10.2196/29586] [Medline: 34623315]
32. Sripadungkul D, Sripadungkul S, Boonmak S, Boonmak P. Differences in perception of online anesthesiology between Thai medical students and teachers during the COVID-19 pandemic. Korean J Med Educ 2023 Mar;35(1):45-53 [FREE Full text] [doi: 10.3946/kjme.2023.248] [Medline: 36858376]

Abbreviations

CONSORT-EHEALTH: Consolidated Standards of Reporting Trials of Electronic and Mobile Health Applications and Online Telehealth

XSL•FO
RenderX

F2F: face-to-face
MCQ: multiple-choice question
TBL: team-based learning
WBVE: web-based virtual environment

Edited by S Brini, T Leung; submitted 04.Jul.2025; peer-reviewed by T Gladman; comments to author 03.Nov.2025; revised version received 23.Dec.2025; accepted 24.Dec.2025; published 15.Jan.2026.

Please cite as:
Sripadungkul D, Boonmak S, Somjit M, Plailaharn N, Sriraj W, Boonmak P
Web-Based Virtual Environment Versus Face-To-Face Delivery for Team-Based Learning of Anesthesia Techniques Among Undergraduate Medical Students: Randomized Controlled Trial
JMIR Med Educ 2026;12:e80097
URL: https://mededu.jmir.org/2026/1/e80097
doi:10.2196/80097
PMID:

©Darunee Sripadungkul, Suhattaya Boonmak, Monsicha Somjit, Narin Plailaharn, Wimonrat Sriraj, Polpun Boonmak. Originally published in JMIR Medical Education (https://mededu.jmir.org), 15.Jan.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Education, is properly cited. The complete bibliographic information, a link to the original publication on https://mededu.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Virtual Reality in Training and Assessment Among Clinical Students and Lecturers at a Nigerian University: A Phenomenological Study

Abiola Olubusola Komolafe[1], PhD; Omotade Adebimpe Ijarotimi[2], MBChB, MPH; Oluseye Ademola Okunola[3], PhD; Olufemi Mayowa Adetutu[4], PhD; Ayodeji Oludola Oluwatope[5], PhD; Olatunde Abiona[6], PhD; Ojo Melvin Agunbiade[3], PhD; Adeboye Titus Ayinde[7], PhD; Stephen Babatunde Aregbesola[8], BChD; Olawale Babatunde Akinwale[9], PhD; Babatope Ayodeji Kolawole[10], MBChB; Lanre Idowu[11], PhD; Alaba Adeyemi Adediwura[12], PhD; Olayinka Donald Otuyemi[13], BDS, MSc, MPH

[1]Department of Nursing Science, Faculty of Basic Medical Sciences, Obafemi Awolowo University, Ile-Ife, Osun, Nigeria
[2]Department of Obstetrics, Gynecology and Perinatology, Faculty of Clinical Sciences, Obafemi Awolowo University, Ile-Ife, Osun, Nigeria
[3]Department of Sociology and Anthropology, Faculty of Social Sciences, Obafemi Awolowo University, Ile-Ife, Osun, Nigeria
[4]Department of Demography and Social Statistics, Faculty of Social Sciences, Obafemi Awolowo University, Ile-Ife, Osun, Nigeria
[5]Department of Computer Science and Engineering, Faculty of Technology, Obafemi Awolowo University, Ile-Ife, Osun, Nigeria
[6]Department of Computer Information Services, College of Arts and Sciences, Indiana University Northwest, Gary, IN, United States
[7]Department of Psychology, Faculty of Social Sciences, Obafemi Awolowo University, Ile-Ife, Osun, Nigeria
[8]Department of Oral and Maxillofacial Surgery, Faculty of Dentistry, Obafemi Awolowo University, Ile-Ife, Osun, Nigeria
[9]Department of Electronic and Electrical Engineering, Faculty of Technology, Obafemi Awolowo University, Ile-Ife, Osun, Nigeria
[10]Department of Medicine, Faculty of Clinical Sciences, Obafemi Awolowo University, Ile-Ife, Osun, Nigeria
[11]Department of Educational Technology and Library Studies, Faculty of Education, Obafemi Awolowo University, Ile-Ife, Osun, Nigeria
[12]Department of Educational Foundation and Counselling, Faculty of Education, Obafemi Awolowo University, Ile-Ife, Osun, Nigeria
[13]Department of Child Dental Health, Faculty of Dentistry, Obafemi Awolowo University, College of Health Sciences, Ile-Ife, Osun, Nigeria

Corresponding Author:
Olayinka Donald Otuyemi, BDS, MSc, MPH
Department of Child Dental Health
Faculty of Dentistry
Obafemi Awolowo University, College of Health Sciences
Road 2, OAU Campus
Ile-Ife, Osun, 220001
Nigeria
Phone: 234 08037191734
Email: ootuyemi@yahoo.com

Abstract

Background: Virtual reality (VR) technology is increasingly used in health care professionals' education as a novel tool for teaching, learning, and assessment.

Objective: This study explored the experiences of clinical students and lecturers with VR for training and assessment at a Nigerian institution. It also explored students' perceptions of the usefulness of VR in improving their clinical abilities, knowledge retention, engagement, and overall learning experience.

Methods: A qualitative research study was conducted among 24 clinical students and 8 clinical lecturers. A developed Virtual reality model to TRain and Assess Clinical Students (VTRACS) was used to train and assess clinical students using clinical scenarios. Data were collected through 4 focus group discussions conducted among the clinical students and 8 in-depth interviews conducted among the clinical lecturers. Trustworthiness was maintained, and ethical approval for the study was obtained. The focus group discussions and in-depth interviews were audio-recorded, transcribed verbatim, and analyzed using NVivo (version 11; QSR International).

XSL•FO
RenderX

Results: Many of the participants had no previous experience with VR in teaching and learning activities, but judging from their engagement with VTRACS, they defined VR as an alternative learning method (alternative to the traditional physical method). Major themes emerging from the study were expression of excitement, simple and useful innovation, proficiency enhancement, challenges with innovation, and uniformity. The clinical students adjudged VTRACS as an educational supplement with a feeling of unlimited learning access, enhancing clinical abilities while positively impacting their confidence and reducing clinical errors. The participants also described the objectivity and standardization of clinical scenarios as drivers of uniformity in training and assessment of clinical students. The participants were, however, concerned about the loss of empathy with the use of VTRACS, which may negatively impact the affective domain of learning.

Conclusions: The use of VR in the teaching and assessment of clinical students at a Nigerian university is perceived as a complementary method of learning that increases skill acquisition, provides unlimited access to training, and enhances proficiency. While VR is considered to be engaging and beneficial to health care professionals' education, there is a need for its effective incorporation into clinical courses and mitigation of challenges such as cost and technology to ensure the realization of the full potential of VR in health care professionals' education.

(JMIR Med Educ 2026;12:e75021) doi:10.2196/75021

KEYWORDS

virtual reality; virtual patient; clinical students; health care professionals' education; digital technology; Nigeria; VR in education; educational technology

Introduction

The challenges posed by the restriction of physical contact during and in the aftermath of the COVID-19 outbreak seem to have suddenly highlighted the need to develop alternative methods of handling teaching and learning activities that require physical contact with a view to attaining desired results [1-3]. In Nigeria, medical training, whose pedagogical design revolves around physical clinical assessment, is no exception.

The increasing adoption of technology, including virtual reality (VR), as an emerging tool in health care professionals' educational training processes, backed by empirical research, reinforces the credibility of this perspective [4]. VR simulations are engaging, interactive, and realistic, and can improve the learning experience and enable new approaches to the training and assessment of clinical trainees. This technology enables students to participate in complicated clinical scenarios in a controlled and safe setting, allowing for the development of important skills without the hazards associated with real-world patient encounters.

VR has enormous potential in medical education, particularly in resource-limited countries, such as Nigeria. Traditional clinical training approaches frequently encounter obstacles, such as restricted access to clinical facilities, insufficient patient interactions, and inconsistent teaching resources [5,6]. Digital technologies have had a tremendous impact on education, resulting in the implementation of integrated information and communication technology (ICT) initiatives. However, concerns about teaching and learning quality persist, particularly during the COVID-19 pandemic and in other situations that limit physical contact. Many educational institutions have limited experience and digital capacity, which results in learning gaps and disparities in the quality of education and training received by clinical students. This has prompted educational institutions to improve their digital capacity and plan for a successful digital transformation. A nonsystematic literature analysis demonstrated that ICT integration has an impact not only on student achievement but also on school-related factors and stakeholders. Factors associated with digital transformation play an important role in effective and efficient changes in educational institutions [7]. VR presents an opportunity to address these issues by providing standardized, high-fidelity simulations that are accessible to all students regardless of their geographical location.

The adoption of VR in Nigerian health care professionals' education is still in its nascent stage, and there is a need for empirical research to understand its impact and effectiveness. The use of a developed VR model in clinical students' instruction and assessment has the potential to alter health care professionals' education in Nigeria [8]. This study adds to the growing body of literature on the use of VR in education by conducting a contextualized analysis of its impact at a Nigerian university. The findings of this study can help guide future VR implementations in health care professionals' education and ensure that this technology meets the unique challenges faced by Nigerian clinical students. Understanding the experiences may provide actionable information to the educators, policymakers, and technology developers on future VR implementation in health care professionals' education to ensure the technology addresses the uniqueness of health care professionals' education, including the potential benefits and limitations of VR in enhancing health care professionals' education in Nigeria. This study aimed to explore the experiences of clinical students following their engagement with a developed VR technology software, the Virtual reality model to TRain and Assess Clinical Students (VTRACS), as part of their training and assessment at a Nigerian university. The study hypothesized that the use of the VR model (VTRACS) in training and assessment of clinical students would promote objectivity in clinical assessment, improve students' clinical abilities, and overall learning experience.

XSL•FO
RenderX

Methods

Setting

This study adopted a phenomenological design to explore the experiences of students and faculty members using the VTRACS software in the College of Health Sciences. Phenomenology is especially appropriate for this study because it allowed the researchers to uncover the essence of how participants perceived, interpreted, and made meaning of VR in their educational environment. Phenomenology was applied in the purposeful sampling of participants who had direct experience with VR and in exploring their personal experiences and interpretations of the technology in the results and discussion, rather than focusing on how VR was used. The college comprises the Faculty of Basic Medical Sciences with 2 programs, namely Nursing Science and Medical Rehabilitation, the Faculty of Dentistry with 1 program, and the Faculty of Clinical Science with 1 program, although it has 2 departments: Medicine and Surgery.

The study was conducted using a haptic VR simulation platform, VTRACS, with limited haptic feedback, delivering immersive clinical scenarios through standalone VR headsets (Meta Quest 2). The simulations were designed to replicate common patient care scenarios, such as performing vital signs checks, abdominal examinations (for pregnancy and appendicitis), and conducting tooth extraction, focusing on skill acquisition in nursing, medical, and dental education. The software used a learner-centered approach for the training and assessment of clinical students across the 3 faculties within the College of Health Sciences of the university [8]. Participants interacted with the virtual environment using hand-held controllers, allowing for navigation, action selection, and patient engagement. While the system provided a high level of visual and auditory immersion, haptic feedback was limited to vibrations; no tactile sensations or physical resistance were simulated. Each VR session lasted approximately 20-25 minutes and was conducted in a controlled laboratory environment with technical support available throughout. The content was aligned with curricular objectives and developed to reflect realistic clinical workflows and patient interactions.

Recruitment Strategy

A total of 32 participants (8 clinical lecturers and 24 clinical students) were recruited for the study. A purposive sampling technique was used to recruit 6 clinical students and 2 clinical lecturers each from Nursing Science, Dentistry, Medicine, and Surgery, who had previously been involved in training and assessment with the VTRACS model. The choice of 6 students from each department was based on the rationale that, unlike quantitative studies, qualitative research focuses on rich, detailed insights from a smaller, carefully selected sample. Research shows that saturation often occurs within 6-12 interviews per homogeneous group, making 6 students per department a reasonable and efficient choice. It also makes data analysis more manageable, especially when using thematic analysis, as was done in this study. Equal representation ensures that comparisons across departments are valid and not skewed by unequal sample sizes. This is crucial when exploring discipline-specific experiences with VR tools in clinical training. Equal numbers help maintain a balanced view of the phenomenon across different clinical disciplines.

The inclusion criteria were being a clinical student and having been trained and assessed in clinical examination with VTRACS, while clinical students who had not been trained or assessed with VTRACS were excluded from the study. The provost of the college and head of departments identified and nominated clinical lecturers with a broad range of disciplinary backgrounds and wide experience who would give diverse insights for the in-depth interviews (IDIs). The students and lecturers had 2 supervised exposures to the VTRACS model before the interviews were conducted.

Data Collection

VTRACS was used to teach and assess 24 clinical students on clinical scenarios in their field of study. Thereafter, data were collected from the students and the clinical lecturers using an interview guide with 4 open-ended questions (Table 1). Four focus group discussions (FGDs), with an average duration of 45 minutes, involving 6 students from each of the 4 departments, were conducted to explore their experiences with the VTRACS model. The basis for selecting 6 students per department was to ensure equality in the representation of the students. In addition, 8 IDIs lasting an average of 30 minutes were conducted with clinical lecturers from the 4 departments to explore their experiences with the VTRACS model. Qualitative rigor of the study was maintained. To ensure credibility, the research team built rapport with the participants and engaged them in the use of VR before the interviews were conducted. Credibility was also enhanced through member checking, where interview transcripts were given to a few participants who were willing to review their transcripts to confirm accuracy and resonance. Participants' feedback from the process of member-checking confirmed accuracy and true representation of the interviews. Various strategies contributed to enhancing confirmability in this study. First, data triangulation was ensured by collecting information from multiple sources, such as clinical students in all the faculties in the college and the clinical lecturers. Second, multiple methods, such as FGDs and IDIs, were used to ensure consistency across different datasets. Third, probing followed the flow of the interviews rather than leading, and clarifications were sought where and when necessary. For dependability, documentation of the research process was done transparently through verbatim transcription of all the recorded interviews. All FGDs and IDIs were audio-recorded and transcribed verbatim. An audit trail was maintained, and the researchers kept audio recordings, analysis notes, and coding details. The description of the setting, participants, and context of exploring participants' experiences with the use of VTRACS allows for the transferability of the study. Qualitative data analysis software, NVivo (version 11; QSR International), was used in organizing data into nodes and themes.

XSL•FO
RenderX

Table 1. Sociodemographic characteristics of participants (N=32).

Variable	Faculties (n=8), n (%)	Students (n=24), n (%)
Age (years)		
21-25	—[a]	20 (83.3)
26-30	—	4 (16.7)
31-35	—	—
36-40	—	—
41-45	3 (37.5)	—
36-40	2 (25)	—
≥46	3 (37.5)	—
Sex		
Male	4 (50)	12 (50)
Female	4 (50)	12 (50)
Faculty or department		
Dentistry	2 (25)	6 (25)
Nursing	2	6
Medicine	2	6
Surgery	2	6
Level of study		
Lecturing	8 (100)	—
400	—	12 (50)
500	—	10 (41.7)
600	—	2 (8.3)

[a]Not applicable.

Ethical Considerations

Ethical approval was obtained from the Health Research Ethics Committee of the Obafemi Awolowo University Teaching Hospitals Complex, Ile-Ife, Nigeria, with protocol number ERC/2022/11/21. The VTRACS project team contacted the potential participants and explained to them the objectives of the study and the purpose of the interviews. Written informed consent was obtained from all the participants, and they were assured of their confidentiality and their right to withdraw from the study at any time without penalty. None of the participants who gave consent to participate in the FGD and IDI withdrew their consent.

Results

Characteristics of Participants

A total of 32 clinical students and faculty members participated across the 3 faculties in the College of Health Sciences, as shown in Table 1. Twenty-four clinical students, 6 from each of the 4 departments or faculties, took part in 4 homogeneous FGDs, and 8 clinical lecturers, 2 from each department or faculty, participated in IDIs.

Themes and Nodes From Participants' Experience

The results of the study yielded 5 major themes and 13 nodes. The experiences of the students and the lecturers were largely similar, except that lecturers reported more concerns compared to the students, possibly because of generational differences (Gen X and Gen Z) in exposure to and interaction with technology. Textbox 1 provides the themes and their corresponding nodes.

XSL•FO
RenderX

Textbox 1. Themes and nodes from participants' experience with the Virtual reality model to TRain and Assess Clinical Students (VTRACS).

Expression of excitement

- Novice
- Complementary learning method
- Imagination rather than reality

Simple and useful innovation

- Ease of use
- Compatible with the task

Proficiency enhancement

- Unlimited learning access
- Confidence
- Fewer medical errors

Challenges with innovation

- Limited patient complexity and feedback
- Loss of empathy
- Financial burden of innovation

Uniformity

- Standardized clinical scenarios
- Objectivity in clinical assessment

Theme 1: Expression of Excitement

The introduction of VTRACS for learning by clinical students prepared them for real-world situations, allowing learning and continuous assessment in furtherance of academic excellence and technical expertise. This theme, "expression of excitement," emerged from the clinical students' description of their experience with VTRACS as being novice, as an alternative learning method, and as imagination rather than reality. Many of the students and clinical lecturers did not have previous experience with VR for teaching and learning activities, but based their understanding on their interaction with VTRACS. The participants said:

> *I haven't used it (VR) before…* [VTRACS DENTISTRY FGD 1]
>
> *Prior to this, I've not had…I've only heard about it, I've not taken part in it.* [VTRACS IDI MEDICINE 2]
>
> *I have not heard about it before…I have not had any experience about it before* [VTRACS IDI NURSING 2]
>
> *I've had contact…using VR, but not really in terms of education…education, because there are so many games that require the use of VR and all that…* [VTRACS IDI DENTISTRY 2]

The definition of VR varies due to the differences in the experiences of the participants with VR. However, a common consensus in the definition of VR is its recognition as an alternative learning method. The participants said:

> *Virtual reality is more or less an alternative to physical practical where you still have the real-life experience scenario at least to a large extent, and you are able to carry out procedures to an extent like the way you would physically.* [VTRACS NURSING FGD 3]
>
> *To the best of my knowledge, by virtual reality, it's just like a form of alternative to the usual traditional way of learning…well, I do not really have the details of how it works. I know it has to do with technological modifications* [VTRACS IDI MEDICINE 2]

The responses from participants in the FGDs and IDIs describe VTRACS as a virtual tool that enhances learning for users who may not have the opportunity for physical training. While some participants viewed it as an alternative, the overall perspective more accurately reflects VTRACS as a complementary method to the traditional physical methods adopted in teaching, particularly in contexts where hands-on experience with real-life patients is essential and available.

The participants demonstrated unfamiliarity with the usage of VTRACS in their statements, but explained their perception of the versatility of VTRACS in teaching and learning from their first experience with it. The participants described the performance of procedures with VTRACS as imagination rather than reality. The imaginative nature of VTRACS immerses users into a different reality where they can explore different and novel learning components. For instance, the participants said:

XSL•FO
RenderX

> *Virtual reality is like imagination…someone's imagination like…defined from what you think the reality will be as the name implies – virtual reality, like it's an imagination.* [VTRACS NURSING FGD 4]

> *It's a simulation…of the clinical procedure which can be used to teach students or…especially online without having an actual patient there. So, I think it's a very...it's a very good initiative.* [VTRACS DENTISTRY FGD 2]

The participant acknowledged VTRACS as a reliable learning platform in the online environment where face-to-face engagements are not possible. This component of VTRACS coincides with the dire need for a substitute method of learning as a result of the limitations imposed by the pandemic. The statement, "It is a very good initiative," describes the positive perception of VR as an alternative teaching method among clinical students. The new pedagogical method seen in VTRACS and the use of imagination in the performance of clinical procedures, which stems from hands-on experience with VTRACS, gave participants a feeling of excitement. Although participants explained that it did not feel the same as physical practice, this is a limitation in replicating reality, as the use and adoption of VTRACS may not replace traditional physical training in clinical practice, but serves as an alternative.

> *It was really exciting, comfortable, and actually interesting.* [VTRACS NURSING, FGD 1]

> *I also found the experience interesting and…comfortable* [VTRACS SURGERY, FGD 6]

> *….it was fun, although it wasn't like the physical one.* [VTRACS MEDICINE, FGD 4]

> *I had a…I had a nice experience, and it was interesting, very interesting…* [VTRACS IDI MEDICINE 1]

> *I mean, the experience is something you want to look toward* [VTRACS IDI DENTISTRY 1]

Although the participants were excited, there was diversity in excitement. While the majority of both students and lecturers found the experience with VTRACS pleasant, comfortable, and interesting, 2 lecturers expressed unpleasantness in the form of headache and dizziness. One of them said:

> *It was not that pleasant to me when I placed the… accessories, …the one that was put on the head…(headset) okay, so, when I…when I wore it, I discovered that I was having a kind of headache during the process* [VTRACS IDI NURSING 2]

This experience of discomfort reinforces the existence of individual differences and health disparities in the use of technologies.

Theme 2: Simple and Useful Innovation

As revealed by the participants' testimonies, they experienced VR in various ways. Participants who described their experiences as simple also expressed various uses of VR for games, education, watching films, and video demonstrations. This acquaintance with VR made the learning of VTRACS easier because they had previous knowledge of the functioning of VR technological innovation.

> *It (VTRACS) picked my interest,….and when there is interest, learning becomes much easier. It was quite easy and very logical and systematic in a way. …it's very clear, and it's well arranged and very logical.* [VTRACS NURSING, FGD 5]

> *It was easy for me to navigate…. the operations and then the…the keys to use, and the instructions were very easy to apply and very easy to follow…..it was so easy to use for me.* [VTRACS SURGERY, FGD 3]

> *It gives room for multiple practice…it's easier…students are able to practice more.* [VTRACS DENTISTRY, FGD 6]

However, some of the participants who had no previous experience with VR found VTRACS interesting but complex at first interaction. For example, one participant said:

> *it was fun, …..although it was difficult to learn at first, but with patience to adapt to a new system. It's more difficult than physical one. Yeah, maybe because I've not tried it so many times, ….if it's physical, I know where to place…I know how to reach…* [VTRACS MEDICINE, FGD 4]

Some lecturers expressed difficulty with hand-eye coordination compared to the students. They, however, opined that learning new technology at their age is good for their cognition. In the words of one of them:

> *They (procedural steps in VTRACS) were logical. The only thing was coordinating between the brain and the eye movement…that took some time, but like I said, practice makes perfect… one will be stimulating his or her brain in doing such things. If there is something you've not been doing before, and you start doing …..… it will help each participant whether old or young, particularly the older ones like me.* [VTRACS IDI SURGERY 1]

Theme 3: Proficiency Enhancement

The theme "proficiency enhancement" was generated from 3 nodes, namely: unlimited learning access, confidence, and fewer medical errors. The nodes evolved from the participants' responses to the benefits of VTRACS in clinical students' training. The participants believed it was an avenue for students to learn better and gain more knowledge, even beyond the boundaries of the classroom.

> *Possible benefits could be that in terms of time flexibility…the students don't always have to be in the hospital settings…they can learn outside hospital settings…it's flexible* [VTRACS IDI MEDICINE 2]

This flexibility, in the opinion of the lecturers, would also give them more time to do other things because they do not have to be physically present to supervise the students at all practice sessions.

XSL•FO
RenderX

The participants' responses reflected an opportunity for seamless learning, transcending physical barriers and limitations of availability. For instance, one of the participants said:

> *It (VTRACS) can be used to teach a large class, …. when there is a need for physical distancing, when we have to conduct exams, and we don't have the patients on ground, we can use the virtual experience to conduct the clinical exams…* [VTRACS IDI SURGERY 2]

Also, the responses from the participants emphasized the benefits of this technology to society. VTRACS, as a virtual platform for learning, equips learners better and helps them replicate these procedures in real-life situations. This further boosts the confidence of clinical professionals through guidance in performing real-life procedures and reduces cases of medical errors. The urge to be in the field of medicine and clinical sciences may be further encouraged through the adoption and use of this technology, thereby producing more medical professionals to cater to the health needs of an ever-evolving society.

> *…if students have known so many procedures using VR, and they have understood very well, they've learnt it, …it will reduce some errors that are committed in the hospitals* [VTRACS IDI NURSING 2]

> *I think if we implement this virtual reality now, it will help both the students and the staff at large…the society at large, …, to be able to …train more students…more doctors will be available* [VTRACS SURGERY, FGD 3]

Theme 4: Challenges With Innovation

This theme captures the experiential and financial concerns expressed by participants regarding the implementation and use of VTRACS. As exciting as most of the participants found VTRACS, some participants described VTRACS as unable to replicate the variability and realism of real patient interactions, limiting opportunities for comprehensive clinical feedback. They also expressed concern that the lack of real human engagement reduced their ability to connect emotionally, potentially diminishing the development of empathy. One of the participants said:

> *…a shortcoming of this virtual reality program is that no two patients are the same, so there is a way you will treat a live patient and that you get a particular type of feedback that will help you become a better doctor, that you won't get from virtual reality.* [VTRACS DENTISTRY, FGD 3]

> *I think the disadvantage is that you will probably be unable to show empathy* [VTRACS DENTISTRY FGD 2]

> *The challenges I think will be that…there could be a possibility of losing touch with the empathy that comes with…. …I mean, the affective aspect* [VTRACS IDI MEDICINE 2]

The participants noted that VTRACS may limit the development of empathy due to the absence of real human interaction. In addition, the lack of comprehensive, responsive feedback in VTRACS may be a potential barrier to professional growth. However, some participants indicated that VTRACS could reduce the anxiety often associated with assessments in traditional physical settings, thereby enhancing performance and proficiency. As one participant explained:

> *Most times, students make mistakes in the physical method of assessment, not because they don't know it, but because of anxiety. With this VR, the fact that you are wearing it (headset), and you can't see the person assessing you, you are comfortable to put what you already know into practice.* [VTRACS NURSING, FGD 3]

Some of the participants, students and lecturers alike, expressed concerns about the financial implications of acquiring the new technology, which may negatively affect sustainability. One student said:

> *…that means every student will have it, and obviously, the financial burden will be there.* [VTRACS NURSING, FGD 6]

One of the lecturers also said:

> *It might be expensive. I don't know the cost of this one, but I know with the way things are, it will be.* [VTRACS IDI SURGERY 1]

Another lecturer said:

> *I don't know how much one costs (headset)…in terms of cost containment, I don't know if it's that cost-effective in the sense that how many of them will be needed for the number of students that we have.* [VTRACS IDI NURSING 2]

These concerns were, however, not sufficient to deter them from embracing VTRACS because they perceived the concerns as challenges that could be overcome.

Theme 5: Uniformity

The theme "uniformity" emerged as a result of participants' acknowledgment that VTRACS ensures standardized clinical scenarios for all clinical students in teaching and examination, thereby promoting objectivity in clinical assessment. The participants said:

> *….it (VTRACS) is also standardized…. what the patient is presenting with will be standardized for everybody, not that one patient will have an extremely difficult case, and one person will have an easy case, and they will still be score based on that.* [VTRACS DENTISTRY, FGD 3]

Discussion

Principal Findings

The use of VR in the training and assessment of clinical students presents transformative potential for health care professionals' education, particularly within the context of a Nigerian university. This qualitative study explored the experiences of clinical students engaged in VR technology (VTRACS) as part of their educational curriculum, highlighting their perceptions

XSL•FO
RenderX

of the innovation, usefulness, and the benefits and challenges associated with this approach. Overall, participants expressed excitement about using VTRACS and recognized it as a valuable educational supplement, particularly in contexts where access to traditional hands-on training is constrained. The objectivity in assessment and standardized clinical scenarios was described as a driver of uniformity in clinical training and assessment by the participants. The flexibility and accessibility of VR-based simulation were seen as key strengths, allowing learners to engage with clinical scenarios in a safe, controlled environment, thus enhancing clinical abilities. However, participants also expressed concern about the limited patient complexity and feedback, and the potential loss of empathy with the use of VTRACS.

Comparison With Previous Work

The findings of this study reveal that VR technology significantly enhances the learning experience of clinical students. The immersive nature of VR allows students to engage in near-realistic clinical scenarios, providing a safe and controlled environment to practice and hone their skills. Richards [9] posited that mixed-reality technologies in health care professionals' education can boost student engagement by combining virtual and augmented reality imagery. Bridge et al [10] also demonstrated that a VR environment improved mindfulness meditation and allowed pseudo-anonymous interactions with peers and tutors. The students in this study reported an increased level of engagement and motivation, noting that VR simulations are more interactive and enjoyable than traditional learning methods. This aligns with the findings of Lau et al [11], who conducted a study involving nurses as participants and reported that the majority similarly found the immersive VR experience engaging and experienced a strong sense of presence. Both studies underscore the potential of VR to enhance learner motivation through immersive, interactive environments, thereby supporting improved learning outcomes by providing a dynamic and stimulating educational experience.

One of the primary benefits identified by the participants in this study was improvement in clinical skill acquisition and confidence. VR simulations offer repetitive practice opportunities in a safe, controlled environment, without the risk of harm to patients. This repetitive practice is crucial for mastering clinical procedures and building confidence. Supporting this, Kennedy et al [12], who compared traditional and interactive VR instructions for clinical students, found that those who received VR training made 40% fewer errors in a simulated practical setting compared to those who received traditional training, highlighting VR's effectiveness in reducing human error and enhancing procedural knowledge. Similarly, Lau et al [11] reported that VR can improve nurses' knowledge and skills, though they emphasized the need for further refinement of VR prototypes to enhance the user experience. In this study, students also noted that the immediate feedback provided by the VR system enabled them to promptly identify and correct mistakes, contributing to more effective and self-directed learning. This finding aligns with a systematic review of 31 studies, which demonstrated that VR-based education significantly improves knowledge and skill outcomes among health care professionals when compared to traditional learning methods, reinforcing VR's positive impact on clinical competence and learner confidence [13].

The integration of VR into health care professionals' education has the potential to address accessibility and equity issues. Lakshminarayanan et al [14] affirmed that affordability and integrated experience creators, when combined with augmented reality (AR) or VR, can facilitate low-cost data collection in remote and rural locations. Evidence has shown that the use of VR in health care, particularly in surgical education, is gaining popularity owing to possible advantages such as enhanced preoperative resident exposure and improved patient safety [15]; however, VR may not completely imitate organic tissues or generate realistic scenarios, and persistent challenges, such as inconsistent internet connections and implementation costs, may impede equitable adoption among academic centers. Ibrahim et al [16] investigated the incorporation of equity, diversity, and inclusion (EDI) into clinical simulation instruction for health care trainees and clinicians. They found that the integration of EDI in clinical simulation education improves self-awareness, communication, insight, knowledge, self-efficacy, and competence. The recommendations include a systematic approach to incorporating EDI, creating a digital repository of EDI-focused scenarios, cocreating simulations with individuals from varied backgrounds, ensuring a secure learning environment, and conducting more rigorous research to enhance the science of clinical simulations.

Participants in this study generally perceived VTRACS as a valuable tool for enhancing learning, particularly in contexts where access to traditional hands-on or physical training is limited. In resource-limited settings such as Nigeria, access to high-quality clinical training facilities and real patient interactions can be limited [17]. VR provides a means of bridging this access gap by offering standardized and high-fidelity simulations accessible to all students, regardless of their geographical location or institutional resources. This democratization of training resources can contribute to leveling the playing field and ensuring that all students receive comprehensive and equitable training. Andigema et al [18] emphasized the importance of action and collaboration in implementing and scaling up artificial intelligence advances to improve health outcomes and ensure that African populations receive the health care they deserve. As immersive technologies such as VR continue to evolve, their strategic integration into health care professionals' education has the potential not only to transform learning experiences but also to drive equitable health care delivery on a global scale.

The ability to engage in safe, repeatable, and immersive clinical simulations was highlighted as a key benefit of VTRACS. While some participants initially described VTRACS as an alternative to conventional training, their overall reflections highlighted important challenges with the innovation, such as limited realism and feedback and a reduced capacity to foster empathy, which are typically addressed more effectively in hands-on, physical settings. These challenges underscore the importance of positioning VTRACS not as an alternative but as a complementary training method. This perspective aligns with evidence from previous studies, which support the use of VR in simulation-based learning primarily as a supplement to

XSL•FO
RenderX

traditional physical methods, unless realistic haptic feedback is integrated to replicate tactile experiences [19-21]. For example, a systematic review by Rangarajan et al [19] found that haptic-enabled VR simulations significantly enhanced surgical performance compared to nonhaptic VR, highlighting the critical role of tactile realism in effectively substituting physical training. Although the VR platform used in this study includes limited haptic feedback through controller vibrations, it lacks tactile sensations or physical resistance, which constrains its ability to fully replicate real-world interactions or convey empathy. Despite its strong visual and auditory immersion, this limitation reduces the system's overall realism.

In addition to pedagogical concerns, this study also identified logistical and technical barriers that could impact the effective implementation of VR tools such as VTRACS. Participants reported challenges, including unfamiliarity with the VR learning format, difficulty operating hand controllers, and occasional physical discomfort. These findings are consistent with those reported by Lau et al [11], who observed similar usability issues in their study. These challenges underscore the need to address both ergonomic design and user onboarding when integrating VR-based training into educational settings.

While virtual and augmented reality technology is gaining traction in higher education due to its immersive learning potential, the high cost of acquiring and maintaining VR hardware and software remains a major constraint, particularly for institutions operating under limited budgets. For instance, Marks and Thomas [22] described a purpose-built VR laboratory that housed 26 Oculus Rift headset units. Over 5 teaching sessions, 4833 students were taught in the laboratory, with notable increases in student engagement, especially in the Faculty of Engineering, and a 250% increase (n=1016) in student enrollment. Although 71.5% (n=211) of the students reported improved learning outcomes, the initiative was supported by subsidized funding, highlighting potential cost barriers to wider adoption. Conversely, the shift toward remote education during the COVID-19 pandemic prompted the exploration of VR as a more scalable and cost-effective alternative. A retrospective case study comparing high-fidelity manikin-based simulation and VR-based learning found that VR was 22% less time-consuming and 40% more cost-efficient, supporting its feasibility as a pedagogical solution in resource-constrained environments [23].

Furthermore, technical issues, such as software glitches and hardware malfunctions, were noted by some students, which can disrupt the learning process and cause frustration. A learning curve was also noted, with some users requiring additional training to engage confidently with the platform. These findings echo previous studies that stress the importance of robust institutional support, reliable technology infrastructure, and user training when integrating VR into educational settings. Collectively, these insights reinforce the idea that while VR can meaningfully augment health care professionals' education, its role should be carefully framed to complement rather than replace traditional hands-on methods.

Cultural and contextual factors also play crucial roles in the implementation of VR in health care professionals' education. In the Nigerian context, factors such as technological infrastructure, institutional readiness, and cultural attitudes toward technology adoption can influence the success of VR integration. The COVID-19 pandemic has highlighted the crucial role of educational technology in the global education system. However, probably because of a lack of resources, Nigeria has faced challenges in sustaining educational services, resulting in a shift in the school calendar, longer graduation dates, and lower research output. To maintain sustainability, the Nigerian education system must adopt a pragmatic approach to innovation by using the diffusion of innovation theory. However, many practitioners lack the requisite skills and attitudes, require in-service training, and face challenges in having a methodical approach to purchasing and installing ICT infrastructure [24].

Following the use of the headset, some participants raised the issue of headache and dizziness. VR sickness, also known as visually induced motion sickness, can be described as a physiological, unpleasant experience resulting from a mismatch between users' vestibular and visual stimuli, which can negatively impact the VR user experience [25]. However, certain factors, such as fear, previous experience of motion sickness, and previous exposure to VR, can also influence participants' experience of VR sickness [26]. Evidence shows that the number of sessions and the intervals between sessions are important factors in increasing adaptation effects and minimizing motion sickness [27-29]. It is usually advisable to start with short sessions, adjust headset settings for comfort, take breaks, avoid intense content, and consult a doctor if the VR sickness persists.

It is essential to tailor the VR content to reflect specific clinical scenarios and cultural nuances relevant to the Nigerian health care system to ensure its effectiveness and acceptance among students and educators. The positive experiences reported by the students in this study suggest a promising future for the integration of VR into clinical education. To maximize the potential of VR, institutions should consider investing in robust technical infrastructure, providing ongoing training and support for both students and faculty, and continuously evaluating and updating VR content to ensure its relevance and accuracy. This is supported by a study that assessed the educational applications of AR and VR across a variety of educational disciplines, concentrating on their impact on student motivation, learning results, engagement, and overall learning experiences [30]. Through simulations, this study investigated how AR and VR can improve information retention and skill acquisition while also encouraging active learning, collaboration, and critical thinking. The study also examined the potential of AR and VR in remote education and made practical recommendations for instructors to effectively incorporate these technologies into their teaching practices [30].

Strengths and Limitations

The qualitative research design used in this study was well-suited, as it enabled an in-depth exploration of students' experiences, perceptions, and contextual interpretations related to the use of VTRACS in their learning processes. One of the primary advantages of qualitative research in educational and media didactic contexts is its ability to capture the complexity and nuance of learner engagement with emerging technologies.

XSL•FO
RenderX

Through rich, narrative data, qualitative methods used in this study provide insight into how students interpret the effectiveness, usability, and emotional impact of digital learning tools, which are often not easily measurable through quantitative approaches. This is especially relevant when evaluating immersive technologies such as VR, where user experience, perceived realism, and pedagogical value can vary significantly across individuals and contexts.

However, qualitative approaches also have inherent limitations. First, the findings of this study are based on a relatively smaller, nonrepresentative sample, which limit generalizability. Second, the subjective nature of data collection and analysis introduces potential biases from both participants and researchers, despite efforts to ensure reflexivity and transparency. Third, in the context of media didactics, where technological tools are often applied across diverse educational settings, the findings should be interpreted with caution and viewed as context-dependent rather than universally applicable. Nevertheless, the insights gained from qualitative studies can serve as a foundation for further research and development.

Future Directions

Future work could build on these findings through mixed methods or quantitative study designs to validate trends observed in the qualitative data. Moreover, the findings can inform the iterative design of educational technologies, training programs for lecturers, and institutional implementation strategies, ensuring that pedagogical innovation remains grounded in the lived experiences of learners.

Conclusion

The use of VR in the teaching and assessment of clinical students is an emerging development in medical education in Nigeria. The use of VTRACS in the training and assessment of clinical students in a Nigerian university was perceived as a complementary learning method that provides continuous access to training and enhances proficiency. While there are hurdles such as cost and technological issues, students' overall favorable feedback demonstrates VR's promise as a valuable tool in health care professionals' education to promote objectivity and improve clinical abilities. Tackling these problems may unlock the full potential of VR in the health care professionals' education and the development of competent and confident health care professionals in Nigeria.

Acknowledgments

We thank all the students and faculty members who participated in the FGDs and in-depth interviews. The authors also acknowledge the financial support of the Tertiary Education Trust Fund under the National Research Fund Grant scheme (reference no TETF/DR&D/CE/NRF2020/CC/24/VOL.1). The authors declare that the funder did not influence the study design; data collection, analysis, or interpretation; writing of the paper; and decision to submit for publication.

Data Availability

The data supporting the findings of this study are available from the corresponding author upon reasonable request.

Funding

This study was funded by the Tertiary Education Trust Fund through the National Research Fund Grant scheme (reference no TETF/DR&D/CE/NRF2020/CC/24/VOL.1). The funder had no role in study design, data collection, analysis, interpretation, or manuscript writing.

Authors' Contributions

AOK contributed to the conceptualization, funding acquisition, methodology, resources, investigation, formal analysis, validation, visualization, writing of the original draft, and review and editing of the manuscript. OAI contributed to the conceptualization, funding acquisition, methodology, resources, investigation, validation, and writing, review, and editing. OAO contributed to the conceptualization, funding acquisition, methodology, and resources. OM Adetutu contributed to the conceptualization, funding acquisition, and methodology. AOO contributed to the conceptualization, funding acquisition, software, and resources. OA contributed to the conceptualization, funding acquisition, software, and resources. OM Agunbiade contributed to the conceptualization, funding acquisition, methodology, formal analysis, and validation. ATA contributed to the conceptualization, funding acquisition, and methodology. SBA contributed to the conceptualization, funding acquisition, methodology, resources, and investigation. OBA contributed to the conceptualization, funding acquisition, and software. BAK contributed to the conceptualization, funding acquisition, methodology, and resources. LI contributed to the conceptualization, funding acquisition, and methodology. AAA contributed to the conceptualization, funding acquisition, and methodology. ODO contributed to the conceptualization, funding acquisition, methodology, resources, investigation, validation, project administration, and supervision.

Conflicts of Interest

None declared.

References

XSL•FO
RenderX

1. Ossai EN. Impact of COVID-19 on medical education and the challenges: how prepared is Nigeria? Pan Afr Med J 2020;37(Suppl 1):45 [FREE Full text] [doi: 10.11604/pamj.supp.2020.37.45.24915] [Medline: 33552373]
2. Komolafe AO, Ijarotimi OA, Adetutu OM, Okunola OA, Ojo TO, Oyinlola FF, et al. Acceptability, appropriateness and feasibility of webinar in strengthening research capacity in COVID-19 era in Nigeria. Niger Postgrad Med J 2022;29(4):288-295 [FREE Full text] [doi: 10.4103/npmj.npmj_167_22] [Medline: 36308257]
3. Otuyemi OD, Jebose OJ. Experience and perceptions of Nigerian undergraduate dental students to virtual learning during the COVID-19 pandemic. Niger J Clin Pract 2023 May;26(5):566-573 [FREE Full text] [doi: 10.4103/njcp.njcp_265_22] [Medline: 37357471]
4. Barteit S, Lanfermann L, Bärnighausen T, Neuhann F, Beiersmann C. Augmented, Mixed, and Virtual Reality-Based Head-Mounted Devices for Medical Education: Systematic Review. JMIR Serious Games 2021 Jul 08;9(3):e29080 [FREE Full text] [doi: 10.2196/29080] [Medline: 34255668]
5. Muhumuza A, Najjuma JN, MacIntosh H, Sharma N, Singhal N, Hollaar GL, et al. Understanding the barriers and enablers for postgraduate medical trainees becoming simulation educators: a qualitative study. BMC Med Educ 2023 Jan 14;23(1):28 [FREE Full text] [doi: 10.1186/s12909-022-03995-3] [Medline: 36641526]
6. Elendu C, Amaechi DC, Okatta AU, Amaechi EC, Elendu TC, Ezeh CP, et al. The impact of simulation-based training in medical education: A review. Medicine (Baltimore) 2024 Jul 05;103(27):e38813 [FREE Full text] [doi: 10.1097/MD.0000000000038813] [Medline: 38968472]
7. Timotheou S, Miliou O, Dimitriadis Y, Sobrino SV, Giannoutsou N, Cachia R, et al. Impacts of digital technologies on education and factors influencing schools' digital capacity and transformation: A literature review. Educ Inf Technol (Dordr) 2023;28(6):6695-6726 [FREE Full text] [doi: 10.1007/s10639-022-11431-8] [Medline: 36465416]
8. Akinwale O, Abiona O, Oluwatope A, Otuyemi O, Ijarotimi O, Olubusola Komolafe A, et al. Designing a virtual reality system for clinical education and examination. Computers & Education: X Reality 2024 Dec 01;5(17):100083-100041 [FREE Full text] [doi: 10.1016/j.cexr.2024.100083]
9. Richards S. Student Engagement Using HoloLens Mixed-Reality Technology in Human Anatomy Laboratories for Osteopathic Medical Students: an Instructional Model. Med Sci Educ 2023 Feb;33(1):223-231 [FREE Full text] [doi: 10.1007/s40670-023-01728-9] [Medline: 36691419]
10. Bridge P, Mehta J, Keane P, El-Sayed O, Mackay S, Ketterer S, et al. A virtual reality environment for supporting mental wellbeing of students on remote clinical placement: A multi-methods evaluation. Nurse Educ Today 2024 Jul;138:106184 [FREE Full text] [doi: 10.1016/j.nedt.2024.106184] [Medline: 38537594]
11. Lau ST, Siah RCJ, Dzakirin Bin Rusli K, Loh WL, Yap JYG, Ang E, et al. Design and Evaluation of Using Head-Mounted Virtual Reality for Learning Clinical Procedures: Mixed Methods Study. JMIR Serious Games 2023 Aug 30;11:e46398 [FREE Full text] [doi: 10.2196/46398] [Medline: 37647108]
12. Kennedy G, Pedram S, Sanzone S. Improving safety outcomes through medical error reduction via virtual reality-based clinical skills training. Safety Science 2023 Sep 01;165(17):106200-106241 [FREE Full text] [doi: 10.1016/j.ssci.2023.106200]
13. Kyaw BM, Saxena N, Posadzki P, Vseteckova J, Nikolaou CK, George PP, et al. Virtual Reality for Health Professions Education: Systematic Review and Meta-Analysis by the Digital Health Education Collaboration. J Med Internet Res 2019 Jan 22;21(1):e12959 [FREE Full text] [doi: 10.2196/12959] [Medline: 30668519]
14. Lakshminarayanan V, Ravikumar A, Sriraman H, Alla S, Chattu VK. Health Care Equity Through Intelligent Edge Computing and Augmented Reality/Virtual Reality: A Systematic Review. J Multidiscip Healthc 2023;16:2839-2859 [FREE Full text] [doi: 10.2147/JMDH.S419923] [Medline: 37753339]
15. Laspro M, Groysman L, Verzella A, Kimberly L, Flores R. The Use of Virtual Reality in Surgical Training: Implications for Education, Patient Safety, and Global Health Equity. Surgeries 2023 Nov 29;4(4):635-646 [FREE Full text] [doi: 10.3390/surgeries4040061]
16. Mutch J, Golden S, Purdy E, Chang CHX, Oliver N, Tallentire VR. Equity, diversity and inclusion in simulation-based education: constructing a developmental framework for medical educators. Adv Simul (Lond) 2024 May 16;9(1):20 [FREE Full text] [doi: 10.1186/s41077-024-00292-5] [Medline: 38750552]
17. Afolabi OF, Ugwu CN, Eze CO, Serges-Anthanase DF, Onwuchekwa CO, Ude MA. Quality of Health Care in Resource Limited Settings-Patients' Satisfaction Survey at a Mission Hospital in South East, Nigeria. Open J Intern Med 2021;11(04):201-209 [FREE Full text] [doi: 10.4236/ojim.2021.114016]
18. Andigema AS, Ngnokam N, Cyrielle T, Kamga N, Danaëlle L, Ekwelle E. Transforming African Healthcare with AI: Paving the Way for Improved Health Outcomes. J Transl Med Epidemiol 2024 Sep 01;7(1):1046. [doi: 10.47739/2333-7125/1046]
19. Rangarajan K, Davis H, Pucher PH. Systematic Review of Virtual Haptics in Surgical Simulation: A Valid Educational Tool? J Surg Educ 2020;77(2):337-347. [doi: 10.1016/j.jsurg.2019.09.006] [Medline: 31564519]
20. Tan J, Karim M, Tamanna R, Kim S, Patel B. Comparing Learning Outcomes of Virtual Reality (VR) Simulators Using Haptic Feedback Versus Box Trainer (BT) in Laparoscopic Training: A Systematic Review and Meta-Analysis. Cureus 2025 Feb;17(2):e78910. [doi: 10.7759/cureus.78910] [Medline: 40091919]

XSL•FO
RenderX

21. Chahal B, Aydin A, Ahmed K. Virtual reality vs. physical models in surgical skills training. An update of the evidence. Curr Opin Urol 2024 Jan 01;34(1):32-36 [FREE Full text] [doi: 10.1097/MOU.0000000000001145] [Medline: 37962178]
22. Marks B, Thomas J. Adoption of virtual reality technology in higher education: An evaluation of five teaching semesters in a purpose-designed laboratory. Educ Inf Technol (Dordr) 2022;27(1):1287-1305 [FREE Full text] [doi: 10.1007/s10639-021-10653-6] [Medline: 34257511]
23. Bumbach MD, Culross BA, Datta SK. Assessing the Financial Sustainability of High-Fidelity and Virtual Reality Simulation for Nursing Education: A Retrospective Case Analysis. Comput Inform Nurs 2022 Sep 01;40(9):615-623. [doi: 10.1097/CIN.0000000000000916] [Medline: 35524349]
24. Ogwu EN, Emelogu NU, Azor RO, Okwo FA. Educational Technology Adoption in Instructional Delivery in the New Global reality. Educ Inf Technol (Dordr) 2023;28(1):1065-1080 [FREE Full text] [doi: 10.1007/s10639-022-11203-4] [Medline: 35875827]
25. Salimi Z, Ferguson-Pell MW. Motion sickness and sense of presence in a virtual reality environment developed for manual wheelchair users, with three different approaches. PLoS One 2021;16(8):e0255898 [FREE Full text] [doi: 10.1371/journal.pone.0255898] [Medline: 34411151]
26. Chattha U, Janjua U, Anwar F, Madni T, Cheema M, Janjua S. Motion Sickness in Virtual Reality: An Empirical Evaluation. IEEE Access 2020 Dec;8(6):130486-130499. [doi: 10.1109/ACCESS.2020.3007076]
27. Domeyer JE, Cassavaugh ND, Backs RW. The use of adaptation to reduce simulator sickness in driving assessment and research. Accid Anal Prev 2013 Apr;53:127-132. [doi: 10.1016/j.aap.2012.12.039] [Medline: 23416680]
28. Mackrous I, Lavallière M, Teasdale N. Adaptation to simulator sickness in older drivers following multiple sessions in a driving simulator. Gerontechnology 2014 Sep 01;12(2):101-111. [doi: 10.4017/gt.2013.12.2.004.00]
29. Kasegawa C, Itaguchi Y, Yamawaki Y, Miki M, Hayashi M, Miyazaki M. Effects of within-day intervals on adaptation to visually induced motion sickness in a virtual-reality motorcycling simulator. Sci Rep 2024 Sep 22;14(1):21302 [FREE Full text] [doi: 10.1038/s41598-024-71526-9] [Medline: 39307847]
30. AlGerafi M, Zhou Y, Oubibi M, Wijaya T. Unlocking the Potential: A Comprehensive Evaluation of Augmented Reality and Virtual Reality in Education. Electronics 2023 Sep 20;12(18):3953-3941. [doi: 10.3390/electronics12183953]

Abbreviations

AR: augmented reality
EDI: equity, diversity, and inclusion
FGD: focus group discussion
ICT: information and communication technology
IDI: in-depth interview
VR: virtual reality
VTRACS: Virtual reality model to TRain and Assess Clinical Students

Edited by T Leung; submitted 27.Mar.2025; peer-reviewed by D Lerner, R Omari; comments to author 02.Jul.2025; revised version received 14.Aug.2025; accepted 30.Nov.2025; published 26.Jan.2026.

Please cite as:
Komolafe AO, Ijarotimi OA, Okunola OA, Adetutu OM, Oluwatope AO, Abiona O, Agunbiade OM, Ayinde AT, Aregbesola SB, Akinwale OB, Kolawole BA, Idowu L, Adediwura AA, Otuyemi OD
Virtual Reality in Training and Assessment Among Clinical Students and Lecturers at a Nigerian University: A Phenomenological Study
JMIR Med Educ 2026;12:e75021
URL: https://mededu.jmir.org/2026/1/e75021
doi:10.2196/75021
PMID:

©Abiola Olubusola Komolafe, Omotade Adebimpe Ijarotimi, Oluseye Ademola Okunola, Olufemi Mayowa Adetutu, Ayodeji Oludola Oluwatope, Olatunde Abiona, Ojo Melvin Agunbiade, Adeboye Titus Ayinde, Stephen Babatunde Aregbesola, Olawale Babatunde Akinwale, Babatope Ayodeji Kolawole, Lanre Idowu, Alaba Adeyemi Adediwura, Olayinka Donald Otuyemi. Originally published in JMIR Medical Education (https://mededu.jmir.org), 26.Jan.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Education, is properly cited. The complete bibliographic information, a link to the original publication on https://mededu.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Medical Students' Experiences With Virtual Reality Simulation Training: Qualitative Study

Vanshika Sharma[1*], BA, MSc; Sohee Park[1*], BSc, MSc; Alexandra Voinescu[1*], MSc, PhD; Chris Jacobs[1*], BChir, MRes, MD(Res)

Department of Psychology, University of Bath, Bath, United Kingdom

[*]all authors contributed equally

Corresponding Author:
Chris Jacobs, BChir, MRes, MD(Res)
Department of Psychology
University of Bath
Claverton Down
Bath, BA2 7AY
United Kingdom
Phone: 44 0 1225 388388
Email: cj511@bath.ac.uk

Abstract

Background: Beyond its applications in other settings, virtual reality (VR) technology has gained attention in medical education, offering immersive learning experiences. Previous research has demonstrated its potential as an educational tool in medical settings, highlighting enhanced educational outcomes, skill acquisition and retention, standardized training experiences, and the promotion of active learning. However, there is still a dearth of research exploring various aspects of VR user experiences, with most studies focusing on its effect on skill acquisition. Limited qualitative research further hinders an in-depth understanding of user experiences, restricting a comprehensive overview of VR's potential in medical education.

Objective: This study explored subjective experiences with VR simulation training and its perceived benefits and challenges among medical students in the United Kingdom, using the 5 domains of the Immersive Technology Evaluation Measure (ITEM).

Methods: In July 2024, 15- to 20-minute in-person interviews were conducted with 11 medical students who had completed the immersive VR training consisting of the assessment and treatment of a virtual patient. Guided by the 5 domains of the ITEM as preconceived themes, a deductive thematic analysis was used to explore individual experiences with the training, embedded within narrative responses.

Results: Findings aligned with the 5 a priori ITEM domains of system usability, immersion, motivation, cognitive load, and debriefing. Within these predefined domains, new subthemes emerged that enhanced the understanding of user experience. Participants reported usability barriers involving accessibility, technical issues, and limited variability in scenarios. Immersion was generally strong due to realistic environments, although reduced interactivity constrained authenticity. Motivation was reflected in active engagement and a greater sense of preparedness for clinical practice. Cognitive load was associated with divided attention, physical effects, and a need for clearer guidance and familiarization. Ultimately, participants valued debriefing sessions as valuable opportunities for reflection and reinforcing knowledge.

Conclusions: VR training fosters immersion and motivation, but its effectiveness depends on balancing technical usability with cognitive demands. Future integration should prioritize design variability and structured debriefing to optimize learning outcomes. Refinement of immersive VR training in clinical education is also warranted, alongside further research in broader contexts and longitudinal use.

(JMIR Med Educ 2026;12:e74301) doi:10.2196/74301

KEYWORDS

health professions education; immersive technology; medical education; simulation; simulation-based education; usability; user experience; virtual reality

XSL•FO
RenderX

Introduction

Virtual reality (VR) technology has emerged as a promising tool in various industries, including marketing [1], advertising [2,3], and education [4-6]. Using VR in health care, specifically in medical education, has been of growing interest in generating lifelike medical scenarios to provide students with more realistic and interactive learning experiences [7].

VR consists of computer-generated virtual environments experienced through various technical devices, including motion tracking, audio, and interactive hand controllers combined with head-mounted displays (HMDs) that generate 360° first-person perspectives of the scenario [8-10]. Such devices create a highly immersive virtual environment for users by offering real-time engagement in a realistic virtual environment, which extends beyond traditional simulations.

VR, with its immersive environment, can be highly beneficial in medical education. For instance, HMD-based simulations enhance ecological validity by mirroring the complexity and unpredictability of real-world clinical practice [9,11]. Learners in VR environments can be immersed in virtual hospitals, where they can explore interactive anatomical models, interact with patients, and practice surgical procedures more realistically [12,13]. Reflecting these benefits, systematic reviews have consistently highlighted that immersive VR-based medical simulations may be more effective than traditional methods across multiple outcomes, including clinical and soft skills, performance, self-efficacy, and stress and anxiety reduction [12-15].

Another benefit of this advanced technology is that it offers an immersive learning environment, allowing students to engage in highly realistic simulations of clinical settings without risking patient safety [16]. As such, VR simulations can increase confidence by reducing anxiety and stress and increasing self-efficacy [15,17]. Additionally, it allows students to make errors and derive lessons from them without encountering real-world repercussions, thereby cultivating confidence and competence [18]. In line with this, VR simulations have been found to improve surgeon safety behaviors through medical training in a risk-free setting, enhancing both understanding and adherence to safety protocols [19].

Numerous studies and meta-analyses have shown that VR training enhances both technical and nontechnical skills while fostering a positive learning environment. A systematic review by Zhao et al [20] reported that VR training significantly improved medical students' clinical skills and performance compared to traditional techniques. Similarly, Harrington et al [21] found that VR training can match or even surpass conventional methods across a range of medical procedures and skills. Supporting this, Kyaw et al [22] conducted a systematic review and meta-analysis demonstrating that VR interventions improved posttraining knowledge and skills relative to traditional education. A further review comparing HMD simulations with traditional teaching formats (eg, slides, videos, 3D-printed or silicon models, and non-HMD simulators such as LapSim) showed that immersive VR accelerated the learning curve and increased motivation [12]. More recent umbrella reviews also reinforce these findings, highlighting that VR simulations improve learners' skills, performance, and engagement [14,15].

Another potential advantage of VR in medical education is its ability to standardize training experiences, feedback, and assessment. Allowing for bias-free evaluations, providing real-time feedback, and enabling the recording of training data, it ensures consistent, reproducible training scenarios across learners, addressing the variability often encountered in traditional clinical training [15]. This standardization may contribute to more equitable learning opportunities and assessment practices while reducing the workload of instructors, as VR may require less supervision time [15]. For example, VR-based standardized patient encounters led to more consistent and objective assessments of medical students' communication skills than traditional methods [23].

Despite initial cost concerns, advances in VR technology and affordable commercial HMDs may make training more accessible [24]. For example, a cost-effectiveness analysis by Farra et al [25] suggested that despite high initial costs, VR-based medical training programs may lead to significant cost savings over 3 years by reducing the need for physical resources and instructor time. Recent studies also indicated that simulation training may reduce the time required to achieve clinical competency and optimize training strategies [26].

User experience studies have recorded generally positive feedback on VR training; however, issues such as discomfort from the hardware or simulator sickness, a condition similar to motion sickness that occurs during or shortly after exposure to VR, characterized by symptoms such as nausea, dizziness, headache, disorientation, and eye strain, have been noted in several studies [11,12,15]. Nevertheless, recent research advancements have demonstrated promising results in mitigating simulator sickness and enhancing user comfort in VR systems intended for application across fields [27,28].

While existing research has substantially advanced the use of VR in medical training, much of the literature has primarily examined its role in skill acquisition and learning [15,29]. Comparatively less attention has been given to user experience–related factors, although a small number of studies have explored usability and technology adoption [30,31]. Moreover, the predominance of quantitative methodologies in this area [29] may further limit insight into the subjective and experiential aspects of VR engagement, which are not always fully captured through numerical measures alone. In line with this observation, a recent review highlights the relative scarcity of qualitative studies investigating the impact of VR on medical education [32].

Considering the limitations, this study used the domains of the Immersive Technology Evaluation Measure (ITEM) [33,34], a validated instrument, to comprehensively explore the user experience, precisely that of medical students, with VR technology in medical education. The ITEM was developed to capture the user experience of immersive technology in medical education [35]. The quantitative and qualitative content analyses of survey responses led to an online consensus meeting with key stakeholders, including doctors, allied health educators,

XSL•FO
RenderX

technicians, and simulation administrators [33], on key aspects of measuring user experience, which guided the development of the 5 ITEM domains: immersion, motivation, cognitive load, system usability, and debriefing [33,34].

Grounded in the Model of Immersive Technology in Healthcare Education (MITHE), which explains multisensory experiences using interactions between technology interfaces, immersion, cognitive load, and motivational states [34], these domains serve as valid constructs for evaluating the effect of immersive technology on medical education. Furthermore, the domains were developed with expert consensus, capturing essential considerations for assessing user experience and ensuring relevance to real-world medical education settings. Integrating these domains, this research will address the limitations of prior studies by exploring underresearched user experience factors, which will provide insights into the potential of implementing VR technology in medical education.

Therefore, this qualitative study explores medical students' experiences of VR simulation training, examining perceived benefits and challenges through the ITEM framework [33-35]. By capturing nuanced user perspectives, the study aims to inform how VR can be effectively integrated into medical education to support meaningful learning and more efficient training environments. Applying the ITEM framework to narrative data also enables further evaluation of its use in immersive medical education research.

Methods

Design

The study used a qualitative research design to explore the experiences of VR training among medical students with a deductive thematic lens, enabling the exploration of the use of the ITEM framework in understanding the experiences. Considering that a preexisting theoretical model from health care education guided our study, the postpositivist paradigm was considered most appropriate. Postpositivism recognizes the existence of an external reality while acknowledging that complete objectivity is unattainable, as observations are inevitably shaped by subjectivity [36,37]. The paradigm aligns with the deductive and theory-informed nature of our approach, while still recognizing the subjective dimensions inherent in experiential accounts.

Procedure

Before the simulation, participants received a demonstration of using the VR devices in groups (4 students on day 1 and 7 students on day 2). Two teaching medical professionals were available throughout the simulation. They were registered medical practitioners with more than 2 years of teaching experience, working full-time as clinical educators for medical undergraduate education in the Department of Teaching at Great Western Hospital (GWH), Swindon, United Kingdom. Participants were instructed to undertake a medical consultation with a patient in VR using verbal prompts and to manage the randomly assigned clinical scenario. They were not informed of the diagnosis and had to assess the virtual patient to determine the diagnosis as if it were a real case. They completed the scenario using instructional materials embedded within the simulation. Teaching doctors and a virtual nurse provided clinical and technical assistance as needed. Each participant completed the simulation one after the other, while other members of the group were present. The simulation lasted between 20 and 30 minutes. The sessions took place over 2 consecutive days across 2 time slots.

After completing all scenarios, the 2 teaching medical professionals conducted a debrief with the entire group. This session lasted 10-15 minutes on average and discussed each student's interactions with the simulation, along with the relevance of their clinical decisions to patient care and diagnosis. After the debrief, the study author (VS) conducted individual interviews with participants to explore their experiences with the VR training, which were recorded via Microsoft Teams [38]. Each interview lasted approximately 15-20 minutes.

Participants

The study was conducted with medical students (n=11; 7 females and 4 males; age range 22-23 years) on clinical placement at the GWH. Participants were recruited from educational sessions led by the teaching medical professionals, and participation was voluntary. Participants opted to gain additional simulation experience via VR in identifying and managing sepsis, asthma, and anaphylaxis.

A purposive sampling strategy was used, wherein only students who had completed the VR simulation were invited to take part in the interview. The inclusion criteria specified that participants must be undergraduate medical students engaged in simulation-based teaching at GWH. For safety reasons, we excluded students living with epilepsy or other medical conditions that could be adversely affected by VR. Recruitment continued until no new themes were emerging from the data, indicating that theoretical saturation had been reached, resulting in a final sample size of 11 participants.

Students were informed that participation was entirely voluntary and could withdraw from participation at any time during the interview without providing a reason. They could also request the withdrawal of their data following participation until the data was fully anonymized.

Materials

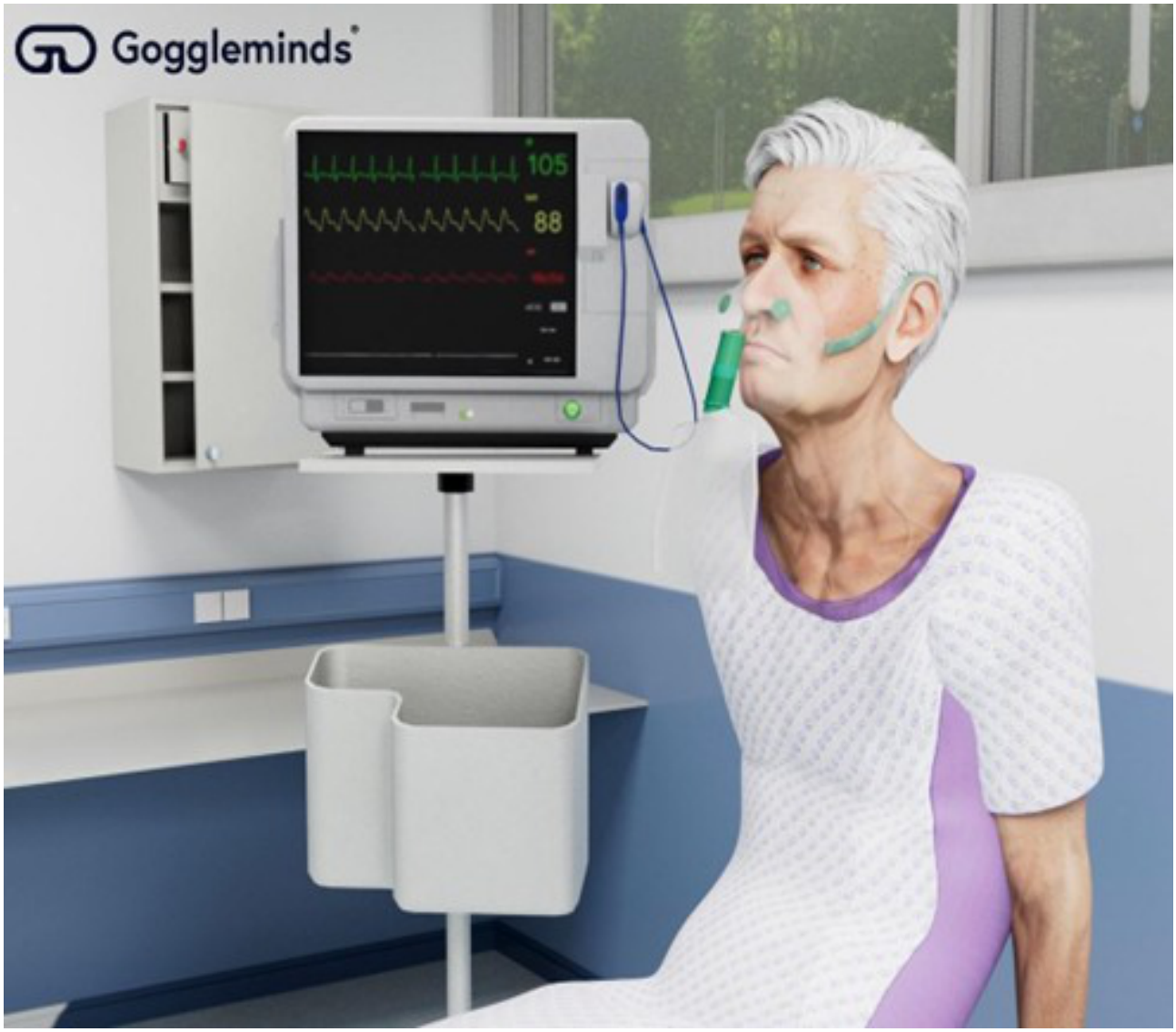
VR Simulations

The VR clinical simulation used in the study was developed based on the prior simulation outlined in the case study by Jacobs et al [39] and instructional design principles. The simulation was created by Goggleminds and operated on a Meta Quest 2 HMD with 2 controllers. The virtual environments were designed to resemble acute medical settings in the United Kingdom, and the virtual patients exhibited typical characteristics of the illnesses presented (Figure 1). The room allowed freedom of movement in the simulated resuscitation room, which was approximately 4 m × 4 m in size and mapped to a UK hospital. Participants were ambulatory and could walk and touch objects within the environment using the controllers. The virtual room was equipped with necessary clinical

XSL•FO
RenderX

instruments, such as a thermometer for measuring temperature and a mask for delivering oxygen.

Figure 1. Virtual reality image of a patient presented to participants.



The simulation incorporated 3 clinical scenarios (sepsis, asthma, and anaphylaxis). Each participant was randomly assigned to one of the scenarios and had to assess and treat the illness of the virtual patient. The clinical scenarios required participants to take a branched-decision patient history, perform clinical assessments, and medically manage the presenting illnesses in an appropriate manner. The clinical content was developed per national guidelines [40]. The VR sepsis training followed Merrill's [41] 5 instructional principles aligned with Bloom's taxonomy [42], providing an experiential, problem-centered simulation.

Simulation content and participant procedural steps were tested in a mannequin simulation environment prior to transfer of content to VR. Participants had access to a clinical history of the virtual patient at any point during the simulation by interacting with the virtual patient. Instructional materials were embedded in the simulation to guide participants through scenarios with a nonbranching logic sequence without logic operators or conditions for further action. A virtual nurse was present to execute participant commands, such as ordering tests or selecting treatments from a dropdown menu. Instructional cues and visual prompts guided decision-making, while embedded feedback and a facilitated debrief upon completion of the simulation supported reflection and consolidation of learning.

Semistructured Interview

A semistructured interview was conducted to explore medical students' experiences of the VR training. The interview began with 3 background questions (job title, years of experience, and training duration), followed by 15 open-ended questions and prompts inviting participants to elaborate on their experiences. Selected questions were adapted from Swinnen et al [43], who used Bowen's [44] framework for interview design, and refined to align with the ITEM framework and the specific context of VR in health care education. Additional questions were developed by the research team to capture feasibility considerations relevant to implementation (Multimedia Appendix 1).

XSL•FO
RenderX

Bowen's [44] framework served as a complementary tool framing pragmatic aspects of feasibility, since it is particularly suited to assessing health care interventions that have already demonstrated effectiveness, shifting the feasibility to the phase of implementation. While Bowen's [44] framework was reviewed to shape question development, the ITEM framework was adopted as the primary informing framework, serving as the conceptual foundation of the interview questions.

Grounded in the MITHE, ITEM provided a theory-driven structure for examining experiential, cognitive, and motivational aspects of immersive learning [34]. The final interview schedule was systematically reviewed to ensure alignment with the 5 ITEM domains (immersion, motivation, cognitive load, system usability, and debriefing) while incorporating relevant feasibility considerations derived from Bowen's framework (Multimedia Appendix 1).

Data Analysis

Interview transcripts were analyzed using deductive thematic analysis [45], guided by the 5 domains of the ITEM framework: immersion, motivation, cognitive load, system usability, and debriefing [33,34]. Predefined themes derived from ITEM structured the initial coding, while subthemes were allowed to emerge within each domain to capture nuances in participants' experiences. This approach ensured that the analysis remained theoretically grounded while accommodating new insights from the data. The process was undertaken within a postpositivist paradigm, recognizing both the value of a structured, theory-informed framework and the interpretive role of the researchers in meaning-making.

Following the interviews, all transcripts were prepared verbatim and pseudonymized, and any identifying data were edited/redacted by VS before analysis, resulting in 11 transcripts. The thematic analysis followed the stages described by Braun and Clarke [46], including familiarization, code generation, code combination, theme review, determination of theme significance, and reporting.

Initial code generation was conducted using NVivo (version 12; Lumivero) [47] by 2 study authors (VS and SP) who independently reviewed the transcripts and reached consensus on the codes. As the study used preconceived themes, codes were categorized under the relevant ITEM domains, with subthemes generated as necessary. Subsequent procedures of the analysis were conducted by VS, SP, and the other 2 authors (AV and CJ) to generate the final 5 themes, 10 subthemes, and 23 codes through researcher consensus.

Ethical Considerations

Ethical approval for the study was obtained from the University of Bath (Data & Digital REC; reference 5416-5731). Following the University of Bath's Research Data policy, fully anonymised data was stored on university-managed servers and temporarily on a password-protected, encrypted personal computer. Only the researchers had access to the data during the study. Informed consent forms and fully anonymised data will be stored for at least 10 years in the University of Bath Research Data Archive. Participants were provided with a brief study overview in the advertising email. They received a detailed information sheet and provided written consent confirming their understanding of participation before proceeding with the study. All participants received £10 (US $13.69) as compensation for their time and participation in the study.

Results

Overview

The analysis of medical students' experiences with VR training was structured around the 5 domains of the ITEM framework: immersion, motivation, cognitive load, system usability, and debriefing. These domains highlighted both the strengths and limitations of VR as a training tool (Figure 2). Within each domain, subthemes were developed from participant narratives, resulting in 5 overarching themes, 10 subthemes, and 23 codes agreed upon through researcher consensus. The findings illustrated the interrelated nature of the ITEM domains, showing how immersive, motivational, cognitive, and usability factors interacted to shape learning experiences. This dynamic interaction reflected the broader mechanisms described in MITHE [34], where multisensory engagement and iterative practice underpin deeper understanding and skill refinement (Figure 3).

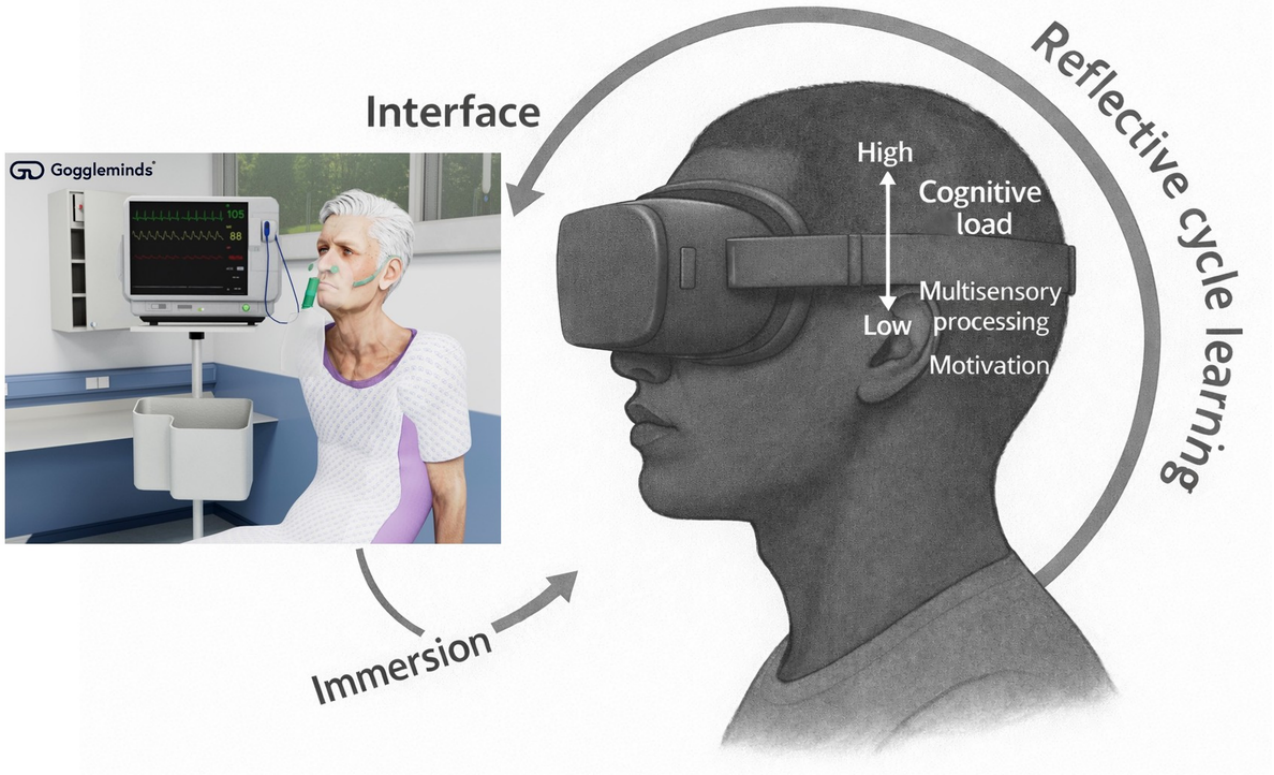
XSL•FO
RenderX

Figure 2. Thematic map with themes, subthemes, codes, and exemplary quotes. VR: virtual reality.



XSL•FO
RenderX

Figure 3. Adapted version of the Model of Immersive Technology in Healthcare Education.



System Usability

System usability covers the issues around the VR training that could limit its efficient usage. The theme covers 3 different issues, denoted as subthemes: accessibility, technology, and variability. Accessibility covers how the physical discomfort of the headset and the cost of training can limit its access. Technology includes the technical limitations of this design and the need for more time for familiarization. Finally, variability reflects the need for having diverse scenarios and levels of difficulty for practice.

Participants raised some concerns regarding the usability of the VR system. One key issue was the wearability of the headsets, which reduced the accessibility of the training. One participant explained, “I wear glasses, and they usually cause a little indent on my nose because of the pressure” (P6). Another participant noted, “It’s a relief to take it off after a session” (P3), indicating that the physical discomfort limited the engagement with the system. Additionally, the perceived cost of implementing VR training was identified as a barrier to its ubiquitous use in medical education (P5).

Participants also encountered minor technical limitations like limited test options and occasional incorrect labels that disrupted their learning. One participant described the frustration caused by software issues, stating, “The bugs and glitches meant that certain options weren’t available to us” (P2). "Another raised the faults within the simulation, remarking, “So for example, like I think this was like the kind of glitch of like getting a medication and the nurse being like this isn’t needed right now, and it kind of is the correct thing to do ... so often things just don’t quite marry up" (P10).

Some suggestions for improving system usability also centered on familiarization. Participants voiced the difficulties faced in using the unfamiliar VR system and recommended allocating time for familiarization. One participant noted, “... but it was quite a lot of information to take on ... it would be helpful to have some kind of extra time outside of the scenario to familiarise ourselves ...” (P11).

The need for more significant variability in the VR training was highlighted across participants. One key issue was the lack of diversity in these training resources, which some participants considered a space for improvement. One participant noted, “It will be nice if there was like a subsection like anaphylaxis ... it will be nice to pick a sort of difficulty level ...” (P9), underscoring the importance of variability in options and levels of difficulty in each scenario. Additionally, participants expressed a desire for scenarios beyond the standard scenarios they typically train for in medical training (P3).

Immersion

Immersion emerged as another central theme, comprising simulation, replication, and realism fidelity. Simulation highlights limitations in practicing hands-on skills and patient interaction within the simulation due to language model constraints. In contrast, replication includes a positive experience of the rehearsal of some manual skills and clinical routines, while realism fidelity reflects the consensus for providing authentic patient scenarios, realistic clinical environments, and a safe, low-stakes context for realistic training experience.

Participants frequently highlighted the realistic clinical environments they encountered through the sensory inputs in the VR training. Participants remarked, “... you can listen with a stethoscope and there are actually heart sounds ... it does do

XSL•FO
RenderX

a good job of simulating real life practise ...” (P5), “I quite liked that everything was a bit intuitive, so the room was like small, but it was representative of the normal hospital room there” (P3), illustrating how the VR training effectively incorporated visual and auditory aspects in actual clinical settings into the VR environment to enhance realism. Additionally, participants noted that the virtual patients closely resembled real-life patients with the illnesses presented in each scenario (P5). Furthermore, participants valued the opportunity to engage in realistic practices without safety risks, which encouraged them to explore a broader range of patient assessments and management techniques (P2).

Many participants praised VR for its high-level replication of clinical practice overall. One participant commented, “... it was a bit more realistic in the sense of having to turn the oxygen on ... rather than everything being done with one click” (P11). This suggests that VR training realistically mimics step-by-step clinical procedures, enhancing the immersive experience. Additionally, participants frequently acknowledged the benefit of the VR simulation for requiring manual skills needed in medical practice. For instance, one participant noted, “You can actually have to grab things and apply them ... say the port of simulator and put it on the patient” (P6), reflecting how the training facilitated the exertion of manual skills that medical professionals ought to be equipped with in real life.

However, some noted limitations in realism within simulation. Some participants reported limited tactile realism in simulation, claiming that the training did not allow them to practice practical skills involving direct and more sophisticated manual interaction (P5). Additionally, participants identified a lack of realism in the training due to limited patient interaction. One suggestion for improvement was to enhance communication options in VR, as noted by a participant: “I’d give the patients more talking options. I think in clinical practice, one thing that is highlighted to us is the communication with our patient ...” (P7). Addressing this in VR training could better prepare medical students for effective patient interactions.

Motivation

Motivation reflects the actuating role of the training with education as a subtheme. Education as a subtheme highlights how VR training fostered medical students to actively engage in learning independently through simulated clinical scenarios and to apply theory to practice during the training. The sense of preparedness denotes how the training helped them prepare for practicing their skills in real-life scenarios, providing a sense of self-assurance.

VR was praised for encouraging active learning, requiring them to actively engage in decision-making and problem-solving on their own during clinical scenarios. One participant explained, “But as long as the person who’s training you stays quiet ... It forces that decision-making process ... as opposed to studying from a textbook or notes where learning is more passive” (P5). This self-directed and active engagement was a significant advantage, helping translate theoretical knowledge into practice. One participant reflected, “VR training helps me think step-by-step about things like how to prescribe fluids or who to ask for help with bloods” (P3). Hence, self-efficacy and agency are promoted as motivations to drive learning.

Another key outcome of the training was an increased sense of preparedness as future medical professionals. Several participants noted that practicing clinical protocols in the virtual environment enhanced their preparedness for real-life scenarios. One remarked, “It made me more confident about running through a protocol in the moment” (P4). This immersive experience was perceived as a valuable tool for learning, fostering confidence, and improving clinical skill retention.

Cognitive Load

Cognitive load reflects both the inherent cognitive demands of completing the clinical task (ie, intrinsic load) and the additional effort created by navigating the VR system (ie, extraneous load), which for some students contributed to discomfort and divided attention. Two subthemes emerged: diffusion of focus and prior guidance. The theme also includes some physical effects experienced by the participants, which may have surfaced due to the cognitive discomfort, such as motion sickness. Diffusion of focus as a subtheme includes the reported displacement of focus from the training to using the VR system itself. Finally, prior guidance includes the need for training and orientation to be better prepared to use the technology and the inclusion of assistive components in the system to better navigate the training.

Participants experienced varying cognitive load during the training, as reported by physical impacts. A notable issue was motion sickness, with a few participants reporting discomfort. One participant shared, “I had a lot of times where I was zooming in and out, which made me really dizzy” (P1).

In addition to physical discomfort, many found that using the system and navigating the VR environment (P5) created a cognitive burden, distracting them from clinical tasks. One participant noted, “Half my focus was actually on figuring out how the system works” (P1). This challenge, coupled with the need for prior knowledge to use the VR training effectively, left some participants feeling unprepared for the sessions. One respondent mentioned, “If you don’t have the knowledge going in ... it feels like a waste of time” (P2), implying that improved preparatory materials or instructions could alleviate the cognitive burden and enhance their learning experience.

Additionally, the need for training assistance in the system was suggested, as noted by one participant: “... maybe something like voice recognition ... like maybe some sort of assistant thing could pop up” (P9), reiterating that additional assistive support before and even during the simulation could benefit medical students in VR training.

Debriefing

Debrief emerged as a central theme, with reflection identified as a key subtheme. Reflection encompassed the positive aspects of both the critical review of the training session and the consolidation of learning, underscoring its importance in reinforcing clinical understanding and supporting the translation of experience into improved practice.

XSL•FO
RenderX

Across the board, participants regarded the debriefing sessions following VR training as one of the most beneficial aspects of the program. These sessions allowed participants to reflect on their experiences, discuss mistakes, and consolidate their learning. One participant stated, “What helps consolidate the memory is the debrief after the session ... I learn the most when I’m able to practice and make mistakes” (P7). Another echoed this, noting, “Where I got the most learning was in the debrief” (P5). This feedback indicates that debriefing is crucial in reinforcing the learning outcomes of the VR training by reflecting on the experiences and consolidating the knowledge gained.

Summary

The results indicate that VR training offers significant benefits, including immersive learning experiences, enhanced motivation, and reflection through debriefing. Acknowledging the benefits, participants also identified areas for improvement, such as expanding communication options and enhancing tactile realism to facilitate more immersive learning. Additionally, they emphasized the need to reduce physical and cognitive burdens associated with the training. The findings also highlight the importance of improving system usability and training variability while addressing technical and headset wearability concerns.

Discussion

Overview

Despite extensive research on the implementation of VR in medical education, few studies have explored a broad range of user experience factors. Moreover, prior research has predominantly relied on quantitative methods, limiting a comprehensive understanding of participants’ experiences with VR. This study begins to narrow these gaps by qualitatively exploring medical students’ VR training experiences, using semistructured interviews and deductive thematic analysis [45] to provide in-depth insights beyond numerical data. Guided by the 5 domains of the ITEM [33,34], the study aligned with the 5 themes, 10 subthemes, and 23 codes to explore broader key constructs for assessing VR in medical education. Adopting a postpositivist, theory-informed perspective, the discussion interprets these findings in relation to established models of immersive learning, highlighting how the ITEM domains interact and reflect the experiences of potential users (ie, medical students).

System Usability: Technical and Psychological Dimensions

The participants frequently highlighted usability challenges, especially those pertaining to wearability. Students wearing glasses reported discomfort due to headset pressure, limiting engagement, reflecting on findings by Kardong-Edgren et al [48] on similar usability issues in VR medical training. Additionally, although participants found VR training immersive and educationally valuable, they reported concerns regarding its costs. The participants reported that the cost of implementation would hinder the ubiquitous use of VR training, which aligns with a review that emphasized the need to address cost issues of VR in medical education [29].

Technical limitations disrupted the VR training in some instances, which participants often reported to impact their experiences with the training. For example, there were instances where the VR environment froze for a few seconds, the sound stopped, or the headset was out of order, which may have negatively affected user experience. Although such limitations were resolved during the training, such as by restarting the VR environment and the headset, these emphasize the importance of usability and comfort in VR design, advocating hardware adjustments to improve user experience [49].

These technical limitations not only affected usability but also undermined participants’ sense of autonomy and competence, which are critical elements in motivation for learning. Past research has identified presence and immersion as key factors that positively influence experience in VR, and are associated with perceived autonomy and competence [50]. For instance, presence and immersion may enhance the sense of autonomy by increasing participants’ perceived control over their virtual behaviors and decision-making [51]. Similarly, users’ perceived competence may increase with greater presence and immersion in the simulation, particularly through seamless interactions with the virtual environment [51]. As a result, the intermittent pause of the training sessions to resolve technical issues may potentially have disrupted the immersion and sense of presence in VR of participants, reducing the sense of autonomy and competence. Our findings indicate that system usability is crucial not only for ease of use but also for enhancing learners’ immersion and presence, which in turn support their sense of autonomy and competence during training.

Participants frequently expressed the need for variability within the VR training. Specifically, they emphasized the need for a broader range of scenarios in terms of diseases, as well as options for treatment embedded within each scenario. Although participants appreciated the realistic clinical scenarios relevant to their future practice, they suggested expanding the range to include a wider variety of diseases. Additionally, participants suggested including drug subsections when selecting treatment options in the simulation, along with adjustable difficulty levels to enhance comprehensiveness and applicability. Recent evidence from educational VR studies [52-54] supports that flexible scenario difficulty and customization enhance motivation for learning by reinforcing competence and autonomy. This highlights the importance of variability in VR medical training to enhance motivation for learning, ultimately enhancing learning outcomes.

The need for sufficient time to familiarize oneself with VR technology was a recurring theme. Many participants initially felt overwhelmed by the VR setup process, including using the controllers and navigating the scenarios within the environment. They were concerned that learning the VR setup might disrupt their focus on the clinical tasks and suggested a dedicated preparatory period to allow sufficient time for familiarization. Such remarks are aligned with the findings by Ghanbarzadeh and Ghapanchi [55], which highlight that familiarity with VR platforms reduces the learning curve, allowing students to engage more effectively with clinical tasks. Findings by Jung and Lee [56] support this approach, recommending dedicated VR practice sessions to help participants familiarize themselves

XSL•FO
RenderX

with the environment and controllers, thereby enhancing their sense of presence and better preparing them for clinical practice.

Immersion and Cognitive Processing

The interviews revealed that participants appreciated the high fidelity of the VR training, emphasizing its close resemblance to real clinical environments. They appreciated the realistic visual representation of the clinical apparatus, patients, and auditory elements, including heart sounds, aligning with prior research demonstrating that greater realism fidelity enhances immersion [57]. Participants also acknowledged that the training successfully replicated clinical practice, incorporating patient assessment and equipment selection within the scenarios. The inclusion of manual tasks, such as requiring them to handle simulators and adjusting oxygen levels using controllers, further anchored users in the clinical settings, showing high resemblance to real-life clinical practice. These narratives are consistent with a study that demonstrated the importance of active participation, including administering diagnostic assessments and therapeutic measures using procedural skills, to strengthen the sense of presence and immersion [58].

Participants also valued the opportunity for low-stakes clinical practice through the training, enabling repetitive training of clinical skills and clinical procedures without safety risks. This aligns with the concept of deliberate practice introduced by Ericsson et al [59], a method shown to improve performance in medical education by providing ample opportunities for gradual refinement of learner performance [60]. For learners to achieve deliberate practice, engagement in focused and repetitive practice of skills is essential [61,62], which is often limited in real clinical settings [60,63] due to safety risks. In this context, immersive VR medical training serves as a platform for clinicians to repeatedly practice skills in a risk-free environment and further analyze their performances for skill improvement. Thus, the VR training is an effective tool for medical students to achieve deliberate practice [16], and is anticipated to enhance their performances in future clinical settings, driving the benefits demonstrated in the previous studies [15-19].

Despite these strengths, some participants highlighted limitations that diminished realism. While the training included manual tasks, participants reported limitations in performing practical and sophisticated skills using their hands because the simulation required using controllers to complete manual tasks, restricting haptic feedback and natural hand movements. In the medical context, studies have shown that haptic feedback enhances realism by providing tactile sensations, thereby making VR training more immersive [64]. Additionally, research indicates that allowing users to perform object-specific natural hand gestures improves immersion in VR training [65]. Despite the importance, realistic haptic feedback is still viewed as a major missing piece [66], and naturalistic hand-object interaction is still limited [67], which may considerably diminish the degree of immersion. The narratives, therefore, illustrate the limitations of current VR technologies, calling for the incorporation of these features to include precise manual tasks, which would enhance realism.

Another limitation noted was the lack of patient interaction, including limited verbal exchanges and opportunities for history-taking. Patient interaction is critical to effective health care delivery and improves therapeutic outcomes and well-being [68,69]. Additionally, simulations with interactive virtual patients enhance realism and foster essential clinical reasoning and history-taking skills [70]. These underscore that sufficient training is needed for medical students for effective patient interaction in actual clinical settings. Therefore, incorporating interactive features into VR training could enhance immersion and prepare medical students for real-world patient interactions, enhancing training outcomes.

Motivation and Self-Determination Theory

Participants acknowledged the VR training for its ability to foster self-directed and active learning. They described the training as a more engaging alternative to traditional text-based education, requiring independent decision-making and systematic thinking. This is consistent with studies showing the efficacy of VR-based training in promoting self-directed and active learning in simulated environments [18,71]. Participants also experienced self-directed and active learning through applying theoretical knowledge to practice, remarking that the training encouraged them to consider treatment steps systematically based on their theoretical knowledge. This aligns with the previous findings, highlighting that VR training fosters active and independent learning accomplished by applying theoretical knowledge in practice, which is crucial for medical education [72].

Self-determination theory (SDT) [73], a human motivational theory that focuses on understanding autonomy functioning, provides a valuable framework for understanding these findings, specifically in the context of VR-based learning. According to SDT, intrinsic motivation leads to high academic performance, which has also been demonstrated in medical education [73-75]. According to the theory, maintaining and enhancing motivation requires satisfying 3 basic needs: autonomy, competence, and relatedness [73]. Autonomy refers to an individual's perceived control over their actions [76], competence refers to the ability to effectively interact with the environment and the confidence to achieve desired outcomes [77], and relatedness refers to the need to form meaningful connections and interact with others during an experience [78].

From an SDT perspective, the VR training supported autonomy by allowing participants to engage in self-directed learning and independently apply theoretical knowledge to practical scenarios, actively helping them to have control over their actions. Additionally, it is possible that participants achieved competence through interacting with realistic patients and a clinical environment to some extent, taking actions to achieve desired clinical outcomes. Therefore, the narratives demonstrate the potential of the VR training as a tool for fostering motivation in medical students if technical optimization is entailed.

The participants' sense of autonomy in SDT is also closely linked to a sense of agency in VR, defined as the user's perception of control and ability to influence their virtual environment and its consequences [79]. The VR training allowed medical students to interact with virtual medical equipment and make step-by-step independent decisions in simulated scenarios. These components are likely to increase a sense of agency,

XSL•FO
RenderX

which is crucial for promoting learners' autonomy and self-directed learning in VR environments [80,81]. Thus, VR training that provides a high sense of agency likely supports students' autonomy, enhances motivation, and improves learning outcomes [73,82].

However, given that the technological limitations potentially diminished the sense of presence and immersion, this motivational triad may have been vulnerable to disruption. Technical limitations, headset discomfort, or confusing interface design that were remarked by participants might have directly undermined the sense of immersion and presence, leading to diminished feelings of competence and autonomy [50,51], thereby diminishing motivation. Participants who struggled to use the system effectively expressed frustration and dependency, an example of how usability issues can thwart psychological needs and reduce motivation even in otherwise engaging environments. Similar findings in immersive learning research show that poor usability elevates extraneous cognitive load and negates the benefits of motivation [83]. Hence, motivation in VR environments is contingent on technological reliability as much as on instructional quality, calling for technological optimization in VR training to foster motivation for learning.

Although they may have experienced limited motivation, participants highlighted that the training provided a sense of preparedness, giving them a sense of readiness to apply these skills in real-life contexts. Among medical students, active learning sessions and practical applications have been shown to help students feel a sense of preparedness, boosting confidence levels [84]. This may indicate that the structure of the training that requires active learning and provides opportunities to apply their knowledge into practice successfully helped students to prepare for future clinical practice, highlighting the importance of incorporating such components in medical education using VR.

Cognitive Load and Managing Complexity

Participants frequently described VR training as cognitively demanding, particularly when navigating virtual environments. Dedicated attention to the interface and VR environment interrupted the full concentration on the clinical tasks and created additional cognitive strain, echoing prior findings that immersive learning environments may inadvertently elevate extraneous cognitive load [85]. Physical effects, including motion sickness and visual fatigue, further compounded this burden, reflecting how physical discomfort can interfere with concentration [86]. While some cognitive challenge is pedagogically beneficial, excessive demand risks overshadowing core learning objectives. As seen in the narratives of participants, providing assistive components embedded within the simulation and preparatory materials could reduce unnecessary load and enable learners to focus on clinical reasoning. This is consistent with the recommendations of Howard and Lee [87], who emphasize attentional guidance and pretraining interventions to balance complexity and cognitive efficiency.

Crucially, from the narratives, system usability emerged as a core determinant of cognitive load. Participants who struggled with headset wearability or technological issues reported cognitive strain and distraction, suggesting that design flaws amplify mental workload. Thus, usability and cognitive load should be considered reciprocal: improved interface design directly alleviates cognitive demands, while excessive mental effort can in turn exacerbate perceived usability issues. This relationship reinforces that technological design and cognitive design are inseparable aspects of effective simulation pedagogy.

Reported cognitive load can also be examined through the lens of realism and fidelity. According to cognitive load theory [88], all novel information is primarily processed by working memory, which is limited in capacity and duration. Therefore, in virtual environments that provide high immersion, users may encounter a flood of new information, increasing their cognitive load [89]. For instance, immersive VR was found to increase cognitive load and decrease cognitive engagement [90], which may impact learning. Additionally, immersive VR was perceived as more distracting, potentially due to the increased cognitive load [91]. From these findings, it can be inferred that high immersion may enhance the sense of agency and competence in learners, leading to increased motivation. However, it can simultaneously increase cognitive load when learners allocate too many attentional resources to managing the interface rather than to clinical reasoning. Given that a familiarization period reduces cognitive load in VR users [89], the potential increase in cognitive load from high-fidelity VR can be alleviated by providing the preparatory materials suggested by participants. Additionally, this highlights the need to balance fidelity with the cognitive resources required to optimize training and promote positive learning outcomes.

Debriefing to Integrate Reflection and Relatedness

Debriefing emerged as a critical process through which learners transformed immersive experiences into lasting knowledge. Participants consistently described postsimulation discussions as where "the most learning happened," providing practice experience. Debriefing allowed errors to be reframed as opportunities for growth, consolidating procedural and conceptual understanding. These findings align with Fanning and Gaba [92] and McGaghie et al [93], who identified structured debriefing as essential to simulation-based learning. From an SDT perspective, debriefing fulfilled the need for relatedness by creating a supportive environment where learners receive feedback, validation, and social connection, as factors strongly associated with sustained motivation [94-96]. Additionally, debriefing serves as a crucial component for improving performance through deliberate practice. [59]. According to Ericsson et al [59], debriefing enabled learners to gradually refine and improve their performance through immediate expert feedback after training, providing opportunities for reflection and correction of choices, and ultimately facilitating deliberate practice [62].

In this sense, debriefing served a restorative function within the ITEM system. Where cognitive load or usability barriers temporarily undermined autonomy, competence, and confidence, debriefing reestablished equilibrium, reinforcing learning, and reengaging motivation. This balancing role highlights why debriefing should not be considered an adjunct but a core pedagogical domain within immersive education. Additionally, the feedback from medical practitioners during debriefing further

XSL•FO
RenderX

enabled learners to reflect on their performance, which is essential for the improvement of performance and skills. Overall, narratives highlight the importance of debriefing in medical education to facilitate motivation and performance improvement essential for future clinical practice.

Synthesis and Educational Implications

Overall, the findings indicate that immersive VR training involves a complex interaction between immersion, usability, cognitive load, motivation, and debriefing. High immersion may enhance realism and engagement, but without sufficient usability and scaffolding, it can contribute to cognitive overload. Similarly, motivation, which is crucial for learning, depends on the fulfillment of autonomy and competence, which may be disrupted by technical difficulties or disorientation. Debriefing appears to function as a recalibrating mechanism, supporting these psychological needs through reflection and feedback. Additionally, its role as an opportunity to receive constructive feedback and reflect on performance further highlights the importance of debriefing for performance improvement and enhanced motivation. The educational value of VR, therefore, arises not from any individual domain but from the integration and balance among them.

Moreover, such a comprehensive interpretation of narratives emphasizes the importance of balancing technological and psychological fidelity, which would potentially guide future VR simulations. According to the narratives, supporting autonomy (through user control), competence (through graduated difficulty and responsive feedback), and relatedness (through debriefing and collaboration) can sustain engagement and deepen learning, while cognitive load should be managed for optimal learning. To achieve this, future VR simulations can incorporate prebriefing, intuitive interfaces, adaptive difficulty levels, variability in scenarios, and ergonomic hardware to reduce extraneous load. Equally, structured debriefing should remain a central component to reinforce reflection, motivation, and further skill improvement. Ultimately, the goal is not maximal immersion but optimal integration, where cognitive, affective, and technological dimensions align to create meaningful, sustainable learning experiences.

Limitations

While this research offers valuable insights into VR adoption in medical education, some limitations are noted. Participants only underwent the training once before being interviewed without further engagement in the training. This may not have accurately reflected the experience of regular VR use in training over time, which potentially limits our understanding of implementing the technology in medical education. The retrospective nature of the responses also could have introduced information bias due to potential memory inaccuracies. Additionally, the ITEM framework used for data analysis might have restricted the focus to responses within its scope, suggesting the use of an inductive approach for future exploration.

The study's scope was limited to medical students from one UK hospital. This qualitative research aimed to gain in-depth insights into the experiences and perceptions of a specific group of VR training users; thus, the results cannot be generalized to other professions. Future research may encompass a broader population, including various job roles, expertise levels, academic institutions, and geographical locations, to provide a more comprehensive understanding of perceptions.

Conclusions

This qualitative study examined medical students' subjective experiences with VR training. Participants valued the immersive learning experience, highlighting its role in fostering motivation and providing effective debriefing. However, areas for improvement emerged, including the need to reduce cognitive load, enhance realism, and improve system usability. The findings suggest that the effectiveness of VR-based learning depends on balancing these domains rather than maximizing any single aspect. Incorporating these findings could guide advancements in VR as a reliable tool for medical education. Future research should use longitudinal designs and include diverse populations to broaden understanding of VR's effectiveness in medical education, which can help determine whether initial perceptions of excitement or apprehension, potentially influenced by the novelty effect of VR, are sustained, diminished, or evolve with repeated use. Such research would also provide insight into the long-term educational impact of VR training, including knowledge retention, transfer of skills to clinical practice, and the durability of motivational and immersive benefits. By aligning technological fidelity with psychological and pedagogical principles, immersive VR can evolve into a sustainable and effective component of medical training.

Acknowledgments

We would like to express our deepest gratitude to Goggleminds for providing logistical support with data collection. We also extend our sincere appreciation to Jennifer Cherrett for her contribution to the data analysis process.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Interview schedule and its alignment with the frameworks (table and map format).
[DOCX File , 1046 KB - mededu_v12i1e74301_app1.docx]

XSL•FO
RenderX

References

1. Cowan K, Ketron S. A dual model of product involvement for effective virtual reality: The roles of imagination, co-creation, telepresence, and interactivity. J Bus Res 2019;100:483-492. [doi: 10.1016/j.jbusres.2018.10.063]
2. Kitsopoulou M, Lappas G. AR/VR technologies in advertising: a scoping review on empirical studies on the effectiveness of AR/VR in advertising. AIP Conf Proc 2023;2909(1):030012. [doi: 10.1063/5.0185352]
3. Wedel M, Bigné E, Zhang J. Virtual and augmented reality: advancing research in consumer marketing. Int J Res Mark 2020;37(3):443-465. [doi: 10.1016/j.ijresmar.2020.04.004]
4. Campos E, Hidrogo I, Zavala G. Impact of virtual reality use on the teaching and learning of vectors. Front Educ 2022;7. [doi: 10.3389/feduc.2022.965640]
5. Liu R, Wang L, Lei J, Wang Q, Ren Y. Effects of an immersive virtual reality‐based classroom on students' learning performance in science lessons. Br J Educ Technol 2020;51(6):2034-2049. [doi: 10.1111/bjet.13028]
6. Miller MD, Castillo G, Medoff N, Hardy A. Immersive VR for organic chemistry: impacts on performance and grades for first-generation and continuing-generation university students. Innov High Educ 2021;46(5):565-589. [doi: 10.1007/s10755-021-09551-z]
7. Mistry D, Geevarughese S, Brock CA, Nelson J, Drew D, Goddard CM, et al. Utilizing a virtual reality matrix in medical education. Cureus 2024;16(8):e66446. [doi: 10.7759/cureus.66446] [Medline: 39132090]
8. Parsons T, Gaggioli A, Riva G. Virtual reality for research in social neuroscience. Brain Sci 2017;7(4):1-22. [doi: 10.3390/brainsci7040042]
9. Rizzo AS, Koenig ST. Is clinical virtual reality ready for primetime? Neuropsychology 2017;31(8):877-899. [doi: 10.1037/neu0000405] [Medline: 29376669]
10. Bell IH, Pot-Kolder R, Rizzo A, Rus-Calafell M, Cardi V, Cella M, et al. Advances in the use of virtual reality to treat mental health conditions. Nat Rev Psychol 2024;3(8):552-567. [doi: 10.1038/s44159-024-00334-9]
11. Barteit S, Lanfermann L, Bärnighausen T, Neuhann F, Beiersmann C. Augmented, mixed, and virtual reality-based head-mounted devices for medical education: systematic review. JMIR Serious Games 2021;9(3):e29080 [FREE Full text] [doi: 10.2196/29080] [Medline: 34255668]
12. Xu X, Mangina E, Campbell AG. HMD-Based virtual and augmented reality in medical education: a systematic review. Front Virtual Real 2021;2:692103. [doi: 10.3389/frvir.2021.692103]
13. Elendu C, Amaechi DC, Okatta AU, Amaechi EC, Elendu TC, Ezeh CP, et al. The impact of simulation-based training in medical education: a review. Medicine (Baltimore) 2024;103(27):e38813 [FREE Full text] [doi: 10.1097/MD.0000000000038813] [Medline: 38968472]
14. Tene T, López DFV, Aguirre PEV, Puente LMO, Vacacela CV. Virtual reality and augmented reality in medical education: an umbrella review. Front Digit Health 2024;6:1365345 [FREE Full text] [doi: 10.3389/fdgth.2024.1365345] [Medline: 38550715]
15. Mergen M, Graf N, Meyerheim M. Reviewing the current state of virtual reality integration in medical education: a scoping review. BMC Med Educ 2024;24(1):788 [FREE Full text] [doi: 10.1186/s12909-024-05777-5] [Medline: 39044186]
16. Pottle J. Virtual reality and the transformation of medical education. Future Healthc J 2019;6(3):181-185 [FREE Full text] [doi: 10.7861/fhj.2019-0036] [Medline: 31660522]
17. Junga A, Kockwelp P, Valkov D, Schulze H, Bozdere P, Hätscher O, et al. Teach the unteachable with a virtual reality (VR) brain death scenario - 800 students and 3 years of experience. Perspect Med Educ 2025;14(1):44-54 [FREE Full text] [doi: 10.5334/pme.1427] [Medline: 39897616]
18. Ruthenbeck GS, Reynolds KJ. Virtual reality for medical training: the state-of-the-art. J Simul 2017;9(1):16-26. [doi: 10.1057/jos.2014.14]
19. Mazur L, Butler L, Mitchell C, Lashani S, Buchanan S, Fenison C, et al. Effect of immersive virtual reality teamwork training on safety behaviors during surgical cases: nonrandomized intervention versus controlled pilot study. JMIR Med Educ 2025;11:e66186 [FREE Full text] [doi: 10.2196/66186] [Medline: 40311122]
20. Zhao G, Fan M, Yuan Y, Zhao F, Huang H. The comparison of teaching efficiency between virtual reality and traditional education in medical education: a systematic review and meta-analysis. Ann Transl Med 2021;9(3):252 [FREE Full text] [doi: 10.21037/atm-20-2785] [Medline: 33708879]
21. Harrington CM, Kavanagh DO, Quinlan JF, Ryan D, Dicker P, O'Keeffe D, et al. Development and evaluation of a trauma decision-making simulator in Oculus virtual reality. Am J Surg 2018;215(1):42-47. [doi: 10.1016/j.amjsurg.2017.02.011] [Medline: 28262203]
22. Kyaw BM, Saxena N, Posadzki P, Vseteckova J, Nikolaou CK, George PP, et al. Virtual reality for health professions education: systematic review and meta-analysis by the digital health education collaboration. J Med Internet Res 2019;21(1):e12959 [FREE Full text] [doi: 10.2196/12959] [Medline: 30668519]
23. Maicher KR, Zimmerman L, Wilcox B, Liston B, Cronau H, Macerollo A, et al. Using virtual standardized patients to accurately assess information gathering skills in medical students. Med Teach 2019;41(9):1053-1059. [doi: 10.1080/0142159X.2019.1616683] [Medline: 31230496]

XSL•FO
RenderX

24. Landau M, Comeaux M, Mortell T, Boyle R, Imbrescia K, Chaffin AE. Characterizing the untapped potential of virtual reality in plastic and reconstructive surgical training: a systematic review on skill transferability. JPRAS Open 2024;41:295-310 [FREE Full text] [doi: 10.1016/j.jpra.2024.06.015] [Medline: 39188661]
25. Farra S, Gneuhs M, Hodgson E, Kawosa B, Miller ET, Simon A, et al. Comparative cost of virtual reality rraining and live exercises for training hospital workers for evacuation. Comput Inform Nurs 2019;37(9):446-454 [FREE Full text] [doi: 10.1097/CIN.0000000000000540] [Medline: 31166203]
26. Ng D, Yip BHK, Young AL, Yip WWK, Lam NM, Li KK, et al. Cost-effectiveness of virtual reality and wet laboratory cataract surgery simulation. Medicine (Baltimore) 2023;102(40):e35067 [FREE Full text] [doi: 10.1097/MD.0000000000035067] [Medline: 37800761]
27. Dopsaj M, Tan W, Perovic V, Stajic Z, Milosavljevic N, Paessler S, et al. Novel neurodigital interface reduces motion sickness in virtual reality. Neuroscience Letters 2024;825:137692. [doi: 10.1016/j.neulet.2024.137692]
28. Ang S, Quarles J. Reduction of cybersickness in head mounted displays use: a systematic review and taxonomy of current strategies. Front Virtual Real 2023;4. [doi: 10.3389/frvir.2023.1027552]
29. Kim HY, Kim EY. Effects of medical education program using virtual reality: a systematic review and meta-analysis. Int J Environ Res Public Health 2023;20(5):3895. [doi: 10.3390/ijerph20053895]
30. Mäkinen H, Haavisto E, Havola S, Koivisto J. User experiences of virtual reality technologies for healthcare in learning: an integrative review. Behav Inf Technol 2020;41(1):1-17. [doi: 10.1080/0144929x.2020.1788162]
31. Zhang L, Wang CX. User experience of virtual reality in healthcare clinical training. In: Ahram T, Karwowski W, editors. Human Factors in Design, Engineering, and Computing. New York: AHFE International; 2024:1149-1156.
32. Dhar E, Upadhyay U, Huang Y, Uddin M, Manias G, Kyriazis D, et al. A scoping review to assess the effects of virtual reality in medical education and clinical care. Digit Health 2023;9:20552076231158022 [FREE Full text] [doi: 10.1177/20552076231158022] [Medline: 36865772]
33. Jacobs C, Foote G, Williams M. Evaluating user experience with immersive technology in simulation-based education: a modified Delphi study with qualitative analysis. PLoS One 2023;18(8):e0275766 [FREE Full text] [doi: 10.1371/journal.pone.0275766] [Medline: 37531361]
34. Jacobs C, Wheeler J, Williams M, Joiner R. Cognitive interviewing as a method to inform questionnaire design and validity: Immersive Technology Evaluation Measure (ITEM) for healthcare education. Comput Educ X Real 2023;2:100027. [doi: 10.1016/j.cexr.2023.100027]
35. Jacobs C, Foote G, Joiner R, Williams M. A narrative review of immersive technology enhanced learning in healthcare education. Int Med Educ 2022;1(2):43-72. [doi: 10.3390/ime1020008]
36. Popper, Karl. The Logic of Scientific Discovery. New York: Basic Books; 1959.
37. Spencer R, Pryce JM, Walsh J. Philosophical approaches to qualitative research. In: Pryce JM, editor. The Oxford Handbook of Qualitative Research. London: Oxford University Press; 2014:81-98.
38. Version 25306.804.4102.7193. Microsoft Teams. 2025. URL: https://learn.microsoft.com/en-us/officeupdates/teams-app-versioning [accessed 2026-01-17]
39. Jacobs C, Vaidya K, Medwell L, Old T, Joiner R. Case study of virtual reality sepsis management - instructional design and ITEM outcomes. J Vis Commun Med 2023;46(3):168-177 [FREE Full text] [doi: 10.1080/17453054.2023.2280611] [Medline: 37990908]
40. Suspected SR, diagnosisearly M. Suspected sepsis: recognition, diagnosis and early management. National Institute for Health and Care Excellence. URL: https://www.nice.org.uk/guidance/ng51 [accessed 2023-12-20]
41. Merrill MD. First principles of instruction. Educ Technol Res Dev 2002;50(3):43-59. [doi: 10.1007/BF02505024]
42. Adams NE. Bloom's taxonomy of cognitive learning objectives. J Med Libr Assoc 2015;103(3):152-153 [FREE Full text] [doi: 10.3163/1536-5050.103.3.010] [Medline: 26213509]
43. Swinnen N, de Bruin ED, Guimarães V, Dumoulin C, De Jong J, Akkerman R, et al. The feasibility of a stepping exergame prototype for older adults with major neurocognitive disorder residing in a long-term care facility: a mixed methods pilot study. Disabil Rehabil 2023:1-15. [doi: 10.1080/09638288.2023.2182916] [Medline: 36824039]
44. Bowen DJ, Kreuter M, Spring B, Cofta-Woerpel L, Linnan L, Weiner D, et al. How we design feasibility studies. Am J Prev Med 2009;36(5):452-457 [FREE Full text] [doi: 10.1016/j.amepre.2009.02.002] [Medline: 19362699]
45. Braun V, Clarke V. Using thematic analysis in psychology. Qual Res Psychol 2006;3(2):77-101. [doi: 10.1191/1478088706qp063oa]
46. Clarke V, Braun V, editors. Successful Qualitative Research: A Practical Guide for Beginners. London: SAGE Publishing; 2013.
47. Version 12.6.1. NVivo 12. 2018. URL: https://docs.cyberark.com/credential-providers/latest/en/content/landingpages/cp-wn-rn-12.6.1.htm?TocPath=Get%20Started%7CRelease%20notes%7C____17 [accessed 2026-01-17]
48. Kardong-Edgren S, Breitkreuz K, Werb M, Foreman S, Ellertson A. Evaluating the usability of a second-generation virtual reality game for refreshing sterile urinary catheterization skills. Nurse Educ 2018;44(3):137-141. [doi: 10.1097/NNE.0000000000000570] [Medline: 31009442]
49. Blum J, Rockstroh C, Göritz AS. Development and pilot test of a virtual reality respiratory biofeedback approach. Appl Psychophysiol Biofeedback 2020;45(3):153-163 [FREE Full text] [doi: 10.1007/s10484-020-09468-x] [Medline: 32361963]

XSL•FO
RenderX

50. Ijaz K, Ahmadpour N, Wang Y, Calvo RA. Player experience of needs satisfaction (PENS) in an immersive virtual reality exercise platform describes motivation and enjoyment. Int J Hum-Comput Interact 2020;36(13):1195-1204. [doi: 10.1080/10447318.2020.1726107]
51. Xiong SR, Ho SS, Tan W, Li BJ, Lisak G. Virtual environment, real impacts: a self-determination perspective on the use of virtual reality for pro-environmental behavior interventions. Environ Commun 2024;18(5):628-647. [doi: 10.1080/17524032.2024.2361270]
52. Afyouni I, Rehman FU, Qamar AM, Ghani S, Hussain SO, Sadiq B, et al. A therapy-driven gamification framework for hand rehabilitation. User Model User-Adapt Interact 2017;27(2):215-265. [doi: 10.1007/s11257-017-9191-4]
53. Alves S, Callado A, Jucá P. Evaluation of graphical user interfaces guidelines for virtual reality games. : IEEE; 2020 Presented at: 2020 19th Brazilian Symposium on Computer Games and Digital Entertainment (SBGames); 2020 Nov 7-10; Recife, Brazil. [doi: 10.1109/sbgames51465.2020.00020]
54. Avola D, Cinque L, Foresti GL, Marini MR. An interactive and low-cost full body rehabilitation framework based on 3D immersive serious games. J Biomed Inform 2019;89:81-100. [doi: 10.1016/j.jbi.2018.11.012] [Medline: 30521854]
55. Ghanbarzadeh R, Ghapanchi A. A literature review on the use of three-dimensional virtual worlds in higher education. In: Choi DH, Dailey-Hebert A, Estes JS, editors. Current and Prospective Applications of Virtual Reality in Higher Education. New York: IGI Global Scientific Publishing; 2020:21-47.
56. Jung M, Lee M. Educational needs for virtual reality-based simulation training in the nursing care of adult respiratory patients: an explanatory sequential mixed-methods study. BMC Med Educ 2024;24(1):1068 [FREE Full text] [doi: 10.1186/s12909-024-06100-y] [Medline: 39350184]
57. Schott D, Heinrich F, Stallmeister L, Moritz J, Hensen B, Hansen C. Is this the vReal life? Manipulating visual fidelity of immersive environments for medical task simulation. IEEE 2023:1171-1180. [doi: 10.1109/ismar59233.2023.00134]
58. Mühling T, Späth I, Backhaus J, Milke N, Oberdörfer S, Meining A, et al. Virtual reality in medical emergencies training: benefits, perceived stress, and learning success. Multimed Syst 2023;29(4):2239-2252. [doi: 10.1007/s00530-023-01102-0]
59. Ericsson KA, Krampe RT, Tesch-Römer C. The role of deliberate practice in the acquisition of expert performance. Psychol Rev 1993;100(3):363-406. [doi: 10.1037//0033-295x.100.3.363]
60. Mitchell SA, Boyer TJ. Deliberate practice in medical simulation. In: StatPearls. Treasure Island, FL: StatPearls Publishing; 2023.
61. Ericsson KA. Deliberate practice and the acquisition and maintenance of expert performance in medicine and related domains. Acad Med 2004;79(Suppl 10):S70-S81. [doi: 10.1097/00001888-200410001-00022]
62. Ericsson KA. Deliberate practice and acquisition of expert performance: a general overview. Acad Emerg Med 2008;15(11):988-994 [FREE Full text] [doi: 10.1111/j.1553-2712.2008.00227.x] [Medline: 18778378]
63. Issenberg SB, McGaghie WC, Petrusa ER, Lee Gordon D, Scalese RJ. Features and uses of high-fidelity medical simulations that lead to effective learning: a BEME systematic review. Med Teach 2005;27(1):10-28. [doi: 10.1080/01421590500046924] [Medline: 16147767]
64. Boulanger P, Wang T, Hanzaki M. Improving medical simulation using virtual reality augmented by haptic proxy. In: Rath M, Samal TK, editors. Modern Development and Challenges in Virtual Reality. London: IntechOpen; 2022.
65. Schäfer A, Reis G, Stricker D. Controlling continuous locomotion in virtual reality with bare hands using hand gestures. In: Zachmann G, Alcañiz Raya M, Bourdot P, Marchal M, tefanucci J, Yang X, editors. Lecture Notes in Computer Science. Cham: Springer; 2022:191-205.
66. Harris DJ, Buckingham G, Wilson MR, Vine SJ. Virtually the same? How impaired sensory information in virtual reality may disrupt vision for action. Exp Brain Res 2019;237(11):2761-2766 [FREE Full text] [doi: 10.1007/s00221-019-05642-8] [Medline: 31485708]
67. Mangalam M, Oruganti S, Buckingham G, Borst CW. Enhancing hand-object interactions in virtual reality for precision manual tasks. Virtual Real 2024;28(4):166. [doi: 10.1007/s10055-024-01055-3]
68. Cramm JM, Nieboer AP. The importance of productive patient-professional interaction for the well-being of chronically ill patients. Qual Life Res 2015;24(4):897-903 [FREE Full text] [doi: 10.1007/s11136-014-0813-6] [Medline: 25267102]
69. Kulińska J, Rypicz ?, Zatońska K. The impact of effective communication on perceptions of patient safety - a prospective study in selected polish hospitals. Int J Environ Res Public Health 2022;19(15):9174 [FREE Full text] [doi: 10.3390/ijerph19159174] [Medline: 35954529]
70. Sezer B, Sezer TA, Teker GT, Elcin M. Developing a virtual patient: design, usability, and learning effect in communication skills training. BMC Med Educ 2023;23(1):891 [FREE Full text] [doi: 10.1186/s12909-023-04860-7] [Medline: 37993846]
71. Kothgassner OD, Goreis A, Kafka JX, Van Eickels RL, Plener PL, Felnhofer A. Virtual reality exposure therapy for posttraumatic stress disorder (PTSD): a meta-analysis. Eur J Psychotraumatol 2019;10(1):1654782 [FREE Full text] [doi: 10.1080/20008198.2019.1654782] [Medline: 31489138]
72. Walling A, Istas K, Bonaminio GA, Paolo AM, Fontes JD, Davis N, et al. Medical student perspectives of active learning: a focus group study. Teach Learn Med 2017;29(2):173-180. [doi: 10.1080/10401334.2016.1247708] [Medline: 28033488]
73. Ryan RM, Deci EL. Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being. Am Psychol 2000;55(1):68-78. [doi: 10.1037/0003-066X.55.1.68]

XSL•FO
RenderX

74. Ganotice FA, Chan KMK, Chan SL, Chan SSC, Fan KKH, Lam MPS, et al. Applying motivational framework in medical education: a self-determination theory perspectives. Med Educ Online 2023;28(1):2178873 [FREE Full text] [doi: 10.1080/10872981.2023.2178873] [Medline: 36812020]
75. Ten Cate TJ, Kusurkar RA, Williams GC. How self-determination theory can assist our understanding of the teaching and learning processes in medical education. AMEE Guide No. 59. Med Teach 2011;33(12):961-973. [doi: 10.3109/0142159X.2011.595435]
76. Decharms R, Carpenter V. Measuring motivation in culturally disadvantaged school children. J Exp Educ 1968;37(1):31-41. [doi: 10.1080/00220973.1968.11011086]
77. Harter S. Effectance motivation reconsidered toward a developmental model. Hum Dev 1978;21(1):34-64. [doi: 10.1159/000271574]
78. Baumeister RF, Leary MR. The need to belong: desire for interpersonal attachments as a fundamental human motivation. Psychol Bull 1995;117(3):497-529. [Medline: 7777651]
79. Braun N, Debener S, Spychala N, Bongartz E, Sörös P, Müller HHO, et al. The senses of agency and ownership: a review. Front Psychol 2018;9:535 [FREE Full text] [doi: 10.3389/fpsyg.2018.00535] [Medline: 29713301]
80. Cesari V, D'Aversa S, Piarulli A, Melfi F, Gemignani A, Menicucci D. Sense of agency and skills learning in virtual-mediated environment: a systematic review. Brain Sci 2024;14(4):350 [FREE Full text] [doi: 10.3390/brainsci14040350] [Medline: 38672002]
81. Code J. Agency for learning: intention, motivation, self-efficacy and self-regulation. Front Educ 2020;5. [doi: 10.3389/feduc.2020.00019]
82. Kusurkar RA, Croiset G, Mann KV, Custers E, Ten Cate O. Have motivation theories guided the development and reform of medical education curricula? A review of the literature. Acad Med 2012;87(6):735-743. [doi: 10.1097/ACM.0b013e318253cc0e] [Medline: 22534597]
83. Mergen M, Will L, Graf N, Meyerheim M. Feasibility study on virtual reality-based training for skin cancer screening: bridging the gap in dermatological education. Educ Inf Technol 2024;30(4):5251-5282. [doi: 10.1007/s10639-024-13019-w]
84. Erlich DR, Shaughnessy AF. Student–teacher education programme (STEP) by step: transforming medical students into competent, confident teachers. Med Teac 2014;36(4):322-332. [doi: 10.3109/0142159x.2014.887835]
85. Makransky G, Borre‐Gude S, Mayer RE. Motivational and cognitive benefits of training in immersive virtual reality based on multiple assessments. J Comput Assist Learn 2019;35(6):691-707. [doi: 10.1111/jcal.12375]
86. Mittelstaedt JM, Wacker J, Stelling D. VR aftereffect and the relation of cybersickness and cognitive performance. Virtual Real 2018;23(2):143-154. [doi: 10.1007/s10055-018-0370-3]
87. Howard MC, Lee J. Pre‐training interventions to counteract seductive details in virtual reality training programs. Hum Resour Dev Q 2019;31(1):13-29. [doi: 10.1002/hrdq.21378]
88. Sweller J, van Merrienboer JJG, Paas FGWC. Cognitive architecture and instructional design. Educ Psychol Rev 1998;10(3):251-296 [FREE Full text] [doi: 10.1023/a:1022193728205]
89. Mayer RE. Multimedia Learning. 2nd ed. Cambridge: University Press; 2009.
90. Birbara N, Pather N. Instructional design of virtual learning resources for anatomy education. Adv Exp Med Biol 2021;1317:75-110. [doi: 10.1007/978-3-030-61125-5_5] [Medline: 33945133]
91. Parong J, Mayer RE. Cognitive and affective processes for learning science in immersive virtual reality. J Comput Assist Learn 2020;37(1):226-241. [doi: 10.1111/jcal.12482]
92. Fanning RM, Gaba DM. The role of debriefing in simulation-based learning. Simul Healthc 2007;2(2):115-125. [doi: 10.1097/SIH.0b013e3180315539] [Medline: 19088616]
93. McGaghie WC, Issenberg SB, Petrusa ER, Scalese RJ. A critical review of simulation-based medical education research: 2003-2009. Med Educ 2010;44(1):50-63. [doi: 10.1111/j.1365-2923.2009.03547.x] [Medline: 20078756]
94. Deci EL, Ryan RM. The "what" and "why" of goal pursuits: human needs and the self-determination of behavior. Psychol Inq 2000;11(4):227-268. [doi: 10.1207/s15327965pli1104_01]
95. Markwell L, Velten J, Partridge J, Porter J. Virtual motivation: the psychological and transfer of learning effects of immersive virtual reality practice. J Mot Learn Dev 2023;11(2):359-373. [doi: 10.1123/jmld.2022-0084]
96. Orsini C, Evans P, Binnie V, Ledezma P, Fuentes F. Encouraging intrinsic motivation in the clinical setting: teachers' perspectives from the self-determination theory. Eur J Dent Educ 2016;20(2):102-111. [doi: 10.1111/eje.12147] [Medline: 25864943]

Abbreviations

GWH: Great Western Hospital
HMD: head-mounted display
ITEM: Immersive Technology Evaluation Measure
MITHE: Model of Immersive Technology in Healthcare Education
SDT: self-determination theory
VR: virtual reality

XSL•FO
RenderX

Edited by D Chartash; submitted 02.Apr.2025; peer-reviewed by G Jones, M Mergen, M Sanverdi; comments to author 26.Aug.2025; revised version received 09.Jan.2026; accepted 12.Jan.2026; published 11.Feb.2026.

Please cite as:
Sharma V, Park S, Voinescu A, Jacobs C
Medical Students' Experiences With Virtual Reality Simulation Training: Qualitative Study
JMIR Med Educ 2026;12:e74301
URL: https://mededu.jmir.org/2026/1/e74301
doi:10.2196/74301
PMID:

©Vanshika Sharma, Sohee Park, Alexandra Voinescu, Chris Jacobs. Originally published in JMIR Medical Education (https://mededu.jmir.org), 11.Feb.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Education, is properly cited. The complete bibliographic information, a link to the original publication on https://mededu.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

e-Learning, Distance Education, and Virtual and Augmented Reality in Orthopedic Training: European Cross-Sectional Survey of Trainee Acceptance Guided by the Technology Acceptance Model and Unified Theory of Acceptance and Use of Technology

Adam Tibor Schlegl[1,2], MD, PhD; Marko Ostojić[3,4], MD, PhD; Ines Unterfrauner[5], MD; Gianluca Ciolli[6,7], MD; Panayiotis D Megaloikonomos[8], MD, PhD; Vasilios G Igoumenou[9], MD, PhD; Michele Mercurio[10], MD; Thomas Stark[11], MD; David Kordic[12], MD; Martijn Dietvorst[13], MD, PhD; Filipe Lima Santos[14], MD; Viktória Nyakas[1], MD; Luca Tóth[15,16], MD, PhD; András Komócsi[17], MD, PhD, DSC; Péter Maróti[2,18], MD, PhD; András Matuz[19,20], PhD

[1]Department of Orthopaedics, Medical School, University of Pecs, Pecs, Hungary
[2]Medical Skills Education and Innovation Centre, Medical School, University of Pecs, Pecs, Hungary
[3]Traumatology Department, University Hospital Sisters of Mercy, Zagreb, Croatia
[4]Osteon Orthopedics and Sports Medicine Clinic, Mostar, Bosnia and Herzegovina
[5]Department of Orthopedics, Balgrist University Hospital, University of Zurich, Zurich, Switzerland
[6]Department of Orthopaedics and Traumatology, Agostino Gemelli University Polyclinic, Rome, Italy
[7]Università Cattolica del Sacro Cuore, Rome, Italy
[8]First Department of Orthopaedic Surgery, Attikon University General Hospital, National and Kapodistrian University of Athens, Athens, Greece
[9]Department of Spine Surgery, Medius Klinik Nürtingen, Nürtingen, Germany
[10]Department of Orthopedic and Trauma Surgery, "Renato Dulbecco" University Hospital, Magna Graecia University, Catanzaro, Italy
[11]Orthopedics and Trauma Surgery Clinic, University Hospital of Basel, Basel, Switzerland
[12]Department of Orthopaedics, University Clinical Hospital Mostar, Mostar, Bosnia and Herzegovina
[13]Department of Orthopaedic Surgery and Trauma, Máxima Medisch Centrum, Eindhoven, The Netherlands
[14]Serviço de Ortopedia e Traumatologia, Centro Hospitalar de Vila Nova de Gaia, Vila Nova de Gaia, Portugal
[15]Department of Neurosurgery, Medical School, University of Pecs, Pecs, Hungary
[16]Institute for Translational Medicine, Medical School, University of Pecs, Pecs, Hungary
[17]Heart Centre, Medical School, University of Pecs, Pecs, Hungary
[18]3D Printing and Visualization Centre, Medical School, University of Pecs, Pecs, Hungary
[19]Department of Behavioural Sciences, Medical School, University of Pecs, Pecs, Hungary
[20]Szentágothai Research Centre, University of Pecs, Pecs, Hungary

Corresponding Author:
András Komócsi, MD, PhD, DSC
Heart Centre
Medical School
University of Pecs
Ifjúság str. 13
Pecs, 7624
Hungary
Phone: 36 72 536001
Email: komocsi.andras@pte.hu

Abstract

Background: Digital technologies increasingly shape postgraduate medical education, yet orthopedic and trauma training face unique challenges because of the tactile, procedurally focused skills involved. Digital tools partially address these needs, but gaps remain, particularly across diverse European contexts.

Objective: Our primary aim was to quantitatively assess predictors of digital learning technology acceptance (e-learning, distance education, and virtual reality [VR] and augmented reality [AR]) among European orthopedic and trauma trainees, drawing on

XSL•FO
RenderX

the technology acceptance model (TAM) and the unified theory of acceptance and use of technology (UTAUT) as conceptual guides. Specifically, we examined how perceived usefulness and perceived ease of use (TAM), alongside performance expectancy, effort expectancy, social influence, and facilitating conditions (UTAUT), related to trainees' acceptance of digital technologies. These constructs guided variable selection and grouping, attitudinal scale design, and interpretation of how individual and contextual factors shape acceptance of digital learning tools in orthopedic training. Secondary aims were to describe adoption and attitude patterns, identify attitudinal trainee profiles, and examine contextual associations (eg, workplace type and national gross domestic product [GDP]).

Methods: We distributed a multinational survey via European trainee federations and used validated scales to assess digital competence and attitudes and gathered demographic data (n=217 across 29 European countries). We administered the questionnaire in English; however, respondents who self-reported English proficiency below the intermediate level were excluded from the analyses to minimize potential comprehension-related bias. The survey assessed digital experience, self-reported digital competence, and attitudes toward e-learning, distance education, and VR/AR, and collected detailed demographic and workplace data. Expert review, cognitive pretesting, and pilot testing ensured validity and clarity. Analytical methods included Wilcoxon tests, ANOVA, clustering, multinomial logistic regression, and factor analysis to ensure the reliability and validity of attitudinal measures.

Results: e-Learning technologies were the most widely adopted, whereas VR/AR tools were less frequently used despite high average attitude ratings (mean 4.07, SD 0.88). Cluster analysis identified 3 distinctive groups—enthusiastic, supportive, and hesitant—that differed significantly in digital competence and acceptance profiles. Digital competence and national GDP emerged as significant predictors of group membership, consistent with TAM/UTAUT expectations that perceived capability and contextual facilitating conditions shape acceptance. Variation in attitudes was further associated with workplace type and regional resource disparities, underscoring the influence of contextual factors on technology adoption.

Conclusions: European orthopedic trainees show broad support for digital innovations, preferring VR/AR despite low use. Preliminary evidence supports digital competence as a key mediator of acceptance, with GDP and workplace disparities predicting profiles (hesitant vs enthusiastic). Competence-first strategies and targeted resource equity (eg, low-GDP subsidies), together with policy adjustments, may address regional disparities. Future longitudinal and multimethod studies are needed to test causal pathways implied by TAM and UTAUT and to evaluate the generalizability of these findings across specialties and educational contexts.

(JMIR Med Educ 2026;12:e79418) doi:10.2196/79418

KEYWORDS

augmented reality; digital learning; distance education; e-learning; medical education; orthopedic training; virtual reality

Introduction

In postgraduate medical education, digitalization has progressed in ways that are particularly consequential for procedural specialties [1-4]. Early work in the 1990s established the concept of computer-based surgical simulation and virtual reality (VR) as a training analogue to aviation simulators, creating a foundation for skills rehearsal outside the operating room [1,3-5]. Particularly in surgical fields where technical expertise depends on repeated practice, evidence from the 2000s supported 2 parallel developments: on the one hand, internet-based instruction became an effective modality for knowledge acquisition compared with no intervention and, in many settings, comparable to noninternet formats; and on the other hand, simulation-based education matured into a structured approach for deliberate practice and feedback [6]. For orthopedics, these shifts were not simply "generic" educational upgrades: online platforms enabled scalable delivery of radiograph-based case interpretation, fracture classification modules, and arthroscopy video instruction, while simulation provided a response to reduced hands-on exposure and increasing expectations for patient safety [7,8]. During the 2010s, orthopedic simulation moved beyond low-resolution prototypes toward higher-fidelity VR systems with performance metrics and growing evidence of skill transfer. A multicenter randomized blinded trial demonstrated transfer validity from a VR arthroscopic knee simulator to improved operating-room performance, highlighting that VR can train clinically meaningful technical behaviors [9]. In parallel, haptics-integrated platforms emerged for fixation procedures; for example, a randomized controlled trial on a VR haptics-enabled dynamic hip screw simulator showed measurable training effects without compromising patient safety priorities [10]. Since 2020, the COVID-19 pandemic has accelerated the digital delivery of surgical teaching, while immersive simulation evidence has expanded further, including orthopedic trials showing that adding haptic feedback to immersive VR improves performance and safety metrics in basic bone-drilling tasks [11]. More recently, VR-based simulation has increasingly incorporated automated performance assessment, including artificial intelligence–based approaches, to support scalable feedback and competency tracking [12]. Recent orthopedic-specific syntheses also support VR as an effective adjunct for learning arthroplasty techniques, reinforcing the relevance of acceptance and implementation questions in contemporary training environments [13].

The technological development in the field of medical education has significantly impacted orthopedic and trauma training, where the hands-on nature of the specialty poses unique challenges for digital education [14-16]. In orthopedics and trauma, technical performance is strongly guided by tactile and proprioceptive information from bone–implant interaction and manual reduction maneuvers. For example, fracture reduction

XSL•FO
RenderX

requires continuous modulation of traction, rotation, and alignment under variable resistance, while drilling and fixation rely on the operator's ability to regulate force and torque and recognize changes in bone "feel" to avoid plunge or malposition. In contrast to screen-mediated minimally invasive and robotic procedures, where natural haptic feedback is attenuated, and operators must rely predominantly on visual cues to judge applied force, orthopedic trauma procedures frequently involve direct bone handling and high-load maneuvers in which tactile–proprioceptive information remains central to safe reduction and fixation [17-19]. Accordingly, many digital training solutions (including VR and augmented reality [AR] platforms) still lack robust tactile feedback, which is a critical component in the development of manual skills within surgical disciplines. While some studies report limited measurable benefits of artificial tactile feedback, many users indicate that it enhances the perceived realism and fidelity of simulation environments. Moreover, haptic feedback has been shown to potentially improve specific technical abilities, particularly tasks requiring fine motor control [20]. A recent study showed that absent haptic cues in laparoscopic training caused excessive force and increased tissue damage risk [21]. Digital display variability (eg, monitors, tablets, and head-mounted displays) further impairs visualization and depth perception [20,22]. A 2022 study found haptic-augmented VR superior to nonhaptic VR for orthopedic drilling accuracy, safety, and skills [11]. Most studies concur that haptic feedback is essential for optimal outcomes; advanced systems such as integrated VR/AR simulators or haptic gloves offer promise [23]. Yet user perceptions remain underexplored, with European studies noting heterogeneous validation and uneven hospital integration [24,25]. Cadaveric and supervised exposure still excels for complex orthopedics, necessitating hybrid curricula blending virtual and real-tissue simulation to meet European standards [26].

Despite widespread use, research on trainees' attitudes toward VR/AR—especially in orthopedics—remains limited, focusing mainly on technical and clinical rather than attitudinal outcomes. While VR/AR shows promise in preoperative planning and simulation, its adoption in postgraduate education, particularly outside the United States, lacks critical evaluation [14,27-29]. The technology acceptance model (TAM) and unified theory of acceptance and use of technology (UTAUT) frameworks address this: TAM emphasizes perceived usefulness (enhanced performance) and perceived ease of use (low effort). UTAUT adds performance expectancy, effort expectancy, social influence, and facilitating conditions (organizational and technical support) [30,31]. In this study, we used the definition of distance education as a structured educational approach characterized by instruction occurring between geographically or temporally separated learners and instructors, using various instructional materials [32,33]. e-Learning), by contrast, refers to teaching and learning activities facilitated through electronic media, encompassing a wide array of devices, software, and digital content [33]. As an umbrella term, e-learning includes diverse methods, with distance education as one application requiring telecommunication. As an umbrella term, e-learning encompasses VR/AR; however, debates persist in simulation-heavy fields such as orthopedics on whether immersive AR/VR qualifies as e-learning or distinct procedural simulation because of its haptic and skill focus—here, we situate VR/AR within e-learning in the context of training following the COVID-19 pandemic [34]. By situating VR/AR within e-learning, this study illustrates how these theoretical models guide understanding of user engagement with cutting-edge educational tools.

The hybrid TAM–UTAUT model suits orthopedic training: TAM addresses perceived usefulness and ease of use for e-learning, VR, and AR; UTAUT incorporates social influence and facilitating conditions for hierarchical residencies with institutional disparities. Here, digital competence reflects effort expectancy; workplace and gross domestic product (GDP) proxy facilitating conditions [35,36]. Accordingly, the aims of the study were as follows:

- Quantitatively assess acceptance predictors for e-learning, distance education, and VR/AR among European orthopedic trainees, drawing on TAM and UTAUT constructs as conceptual guides for scale selection and interpretation. This addresses limited attitudinal research in procedural specialties beyond technical outcomes.
- Characterize adoption rates and attitude patterns across modalities. Building on low VR/AR use despite high enthusiasm, this maps baseline engagement.
- Identify distinct attitudinal profiles via cluster analysis. Profiles (eg, enthusiastic, supportive, and hesitant) reveal digital competence gradients, aligning with UTAUT effort expectancy.
- Examine associations with contextual factors (workplace type and national GDP). Socioeconomic disparities in digital access—evident in European Union medical training—represent UTAUT facilitating conditions, explaining heterogeneous adoption [30,31].

These aims inform equitable policies for digital integration in diverse European residencies and help ensure that medical device and software development companies effectively integrate digital tools to meet the evolving needs of trainees and institutions in a postpandemic educational landscape.

Methods

Ethical Considerations

This study was approved by the Hungarian United Ethical Review Committee for Research in Psychology (reference number 2023-56). Participants provided informed consent at the beginning of the survey. Data were collected anonymously, and no incentives were offered.

Survey Design and Validation

A self-administered, anonymous online survey was developed, covering demographics, digital competencies, and attitudes toward e-learning, distance education, and VR/AR among orthopedic and trauma trainees, using validated scales (Attitude Toward E-Learning Scale [37], Distance Education Perception Scale [38], and VR/AR attitude items [39]) on a 5-point Likert scale (1=strongly disagree and 5=strongly agree). We did not alter the individual items of the validated scales. Instead, we selected relevant question groups from the original lengthy

XSL•FO
RenderX

questionnaire without modifying their wording or content. This approach preserved the integrity and construct validity of the instrument. Expert review, cognitive pretesting, and pilot testing with 10 trainees (qualitative feedback confirmed clarity and comprehensibility) ensured validity and clarity [40]. All items of the survey were presented in English to ensure consistency across countries and to avoid potential discrepancies introduced by translation.

The e-learning and VR/AR-based learning questionnaires demonstrated reliability in our cohort with Cronbach α values of 0.837 and 0.931, respectively. The Distance Education Perception Scale questionnaire consists of 4 subscales: student's perception, equipment facility, facility and support of the institution, and time management. The first 3 subscales demonstrated high degrees of reliability, ranging from 0.813 to 0.891; however, the time management subscale showed a low level of reliability (α=0.344) and therefore was excluded from the calculations.

Demographics included age, gender, training location (country and city), workplace type (university, regional, city, and rural hospitals), and previous experience with distance education, e-learning, medical simulation, and VR/AR simulation. Digital competence questions covered basic and advanced tools (eg, expertise ranging from basic competencies such as installing a program to advanced skills such as programming or network setup) and VR/AR familiarity (Multimedia Appendix 1).

The structure of the attitudinal scales and analytical framework was informed by the TAM and UTAUT frameworks. In this context, perceived usefulness and ease of use (TAM constructs) were operationalized through items assessing engagement and perceived effectiveness of e-learning, distance education, and VR/AR tools. UTAUT dimensions were represented by social influence and facilitating conditions (eg, equipment availability and workplace resources). These theoretical constructs guided variable grouping, factor structure validation, and subsequent analyses exploring how contextual and personal factors (including GDP, workplace type, and digital competence) influence acceptance and attitudes toward educational technologies.

Participant Recruitment and Data Collection

Due to the lack of comprehensive registries enumerating orthopedic residents by country, a precise estimation of the eligible population and a predefined sample size were not feasible. Therefore, this study was conducted as a time-bounded, voluntary cross-sectional survey. Participants were recruited via Federation of Orthopaedic and Trauma Trainees in Europe newsletters and mailing lists (estimated reach: approximately 1000-1200 trainees across 27 countries), plus targeted emails and social media posts through national orthopedic resident societies. No incentives offered, reducing response bias from economic and regional disparities. A 2-week data collection period with a 1-week reminder yielded 235 responses (217 analyzed after exclusions). In the absence of centralized registries, an exact response rate was unavailable; however, diverse demographics (29 countries and balanced GDP and workplace types) support generalizability.

The survey was hosted on Google Forms, ensuring ease of access and data security. GDP per capita data were obtained from the International Monetary Fund (2022) and categorized into low (<US $29,000), medium (US $29,000-US $40,000), and high (>US $40,000) GDP groups (tertiles of 29 respondent countries), aligning with World Bank thresholds (the European Union average is US $41,806).

Statistical Analysis

To evaluate trainees' attitudes toward e-learning and distance education, mean scores for survey items were compared with the neutral midpoint (3.0) using Wilcoxon signed-rank tests due to nonnormal distributions indicated by Shapiro-Wilk tests (all $P<.01$). Effect sizes were reported as rank-biserial correlations (r_{bs}), with values of 0.10, 0.30, and 0.50 interpreted as small, medium, and large effects, respectively. To examine whether attitudes differed depending on the type of workplace and the economic context of the country, separate 2-way ANOVAs were run with workplace type (university vs nonuniversity clinics) and GDP per capita (low, medium, or high) as between-subject factors, and attitude variables (ie, scores on subscales of the attitude questionnaires) as outcomes. Significant interactions were followed up by simple effect analyses. All post hoc analyses were adjusted for multiple testing using the Bonferroni correction.

A priori power analyses using G*Power (version 3.1.9.7 [41]) indicated that the minimum sample size required for ANOVA was 155 (α=5%; power=80%), and thus the final sample of 217 (see below) had adequate statistical power to detect main effects as well as the interactions examined. Differences in experience levels and digital competencies across domains were tested using the Friedman test. Significant results were followed up by Wilcoxon tests with Bonferroni correction. Spearman partial correlations were used to assess relationships among the measures while controlling for digital competence and experience with digital learning. When an attitude variable was correlated with either competence or experience, the other served as the control variable. No control variables were applied for the correlation between competence and experience. Spearman correlations were chosen because Shapiro-Wilk tests indicated nonnormal distributions for all variables (all $P<.05$), except for the student's perception subscale.

We used k-means cluster analysis to identify distinct subgroups of participants with similar attitudinal profiles toward digital learning. The optimal cluster count was determined using the elbow and silhouette methods (Multimedia Appendix 2). Cluster-wise stability was assessed using the mean Jaccard similarity value obtained via bootstrapping (n=1000). Mean Jaccard similarity values >0.75 and >0.85 were interpreted as indicators of stable or very stable clusters, respectively [42].

After clustering, the dataset was aggregated to calculate mean values for each cluster, providing insight into the dominant attitudes and experiences regarding digital learning across groups. Multinomial logistic regression was conducted to examine how background variables (eg, digital competence, experience, and demographics) predicted cluster membership and to better understand differences between subgroups.

XSL•FO
RenderX

To identify the underlying factor structure and consolidate conceptually overlapping constructs into a clearer and more parsimonious model, exploratory factor analysis (EFA) was conducted on the complete set of items from all 3 attitude questionnaires. Because items were ordinal, a polychoric correlation matrix was computed, and robust maximum-likelihood extraction with oblique (Promax) rotation was used. Factor retention was guided by (1) parallel analysis and scree plot inspection and (2) interpretability; primary loadings ≥0.40, cross-loadings <0.30, and communalities ≥0.30 were used as criteria for item retention. The full EFA diagnostics, including the parallel-analysis scree plot, communalities, and item loadings, are provided in Multimedia Appendix 3.

To confirm the factor structure identified by EFA, confirmatory factor analysis was conducted using robust maximum likelihood estimation (estimator="MLR") to account for nonnormality. Model adequacy was evaluated using standard fit indices and recommended thresholds: comparative fit index (CFI) and Tucker-Lewis index (TLI) ≥0.90 (acceptable) or ≥0.95 (excellent), root mean square error of approximation (RMSEA) ≤0.08 (acceptable) or ≤0.06 (good), and standardized root mean square residual (SRMR) ≤0.08. Standardized factor loadings ≥0.50 were interpreted as adequate. Model refinement involved reviewing modification indices (MI>10) and applying theoretically justified correlations between residuals where appropriate.

Statistical significance was set at $P<.05$ (2-tailed) for all analyses unless Bonferroni-corrected otherwise. All analyses were conducted using SPSS (version 28; IBM Corp) and R (version 4.4.0; R Foundation for Statistical Computing). R packages used are listed in Multimedia Appendix 4.

Results

Overview

In total, 235 responses from 29 countries (geographical distribution shown in Multimedia Appendix 5) were received, of which 217 were included in the analysis after excluding responses with beginner-level English proficiency (n=17) or incomplete data (n=1) to minimize potential bias arising from language-related comprehension issues. Excluded respondents did not significantly differ from included respondents in terms of basic demographics (age: U=1482.5; P=.17; r_{bs}=0.2 and gender: χ^2_1=0.02; P=.88); however, trainees working at university clinics were underrepresented among the excluded participants (χ^2_1=5.3, P=.02). A detailed description is provided in Table 1. The dataset is available elsewhere [43].

Table 1. Participant demographics and background characteristics.

Characteristics	Values
Sample size, n	217
Age (years), mean (SD); range	31.5 (4.38); 24-45
Training duration (years), mean (SD); range	4.31 (2.20); 1-7
Gender, n (%)	
Male	156 (71.9)
Female	61 (28.1)
English proficiency, n (%)	
Intermediate	95 (43.7)
Professional	111 (51.2)
Mother tongue	11 (5.1)
Workplace type, n (%)	
University clinical center or national referral hospital	116 (53.5)
Regional/provincial or county referral hospital	46 (21.2)
City hospital	37 (17.1)
Town or rural hospital	18 (8.3)
GDP[a] per capita group (US $), n (%)	
Low (< 29,000)	88 (40.6)
Medium (29,000-40,000)	69 (31.8)
High (>40,000)	60 (27.6)

[a]GDP: gross domestic product.

XSL•FO
RenderX

Experience and Competence Levels

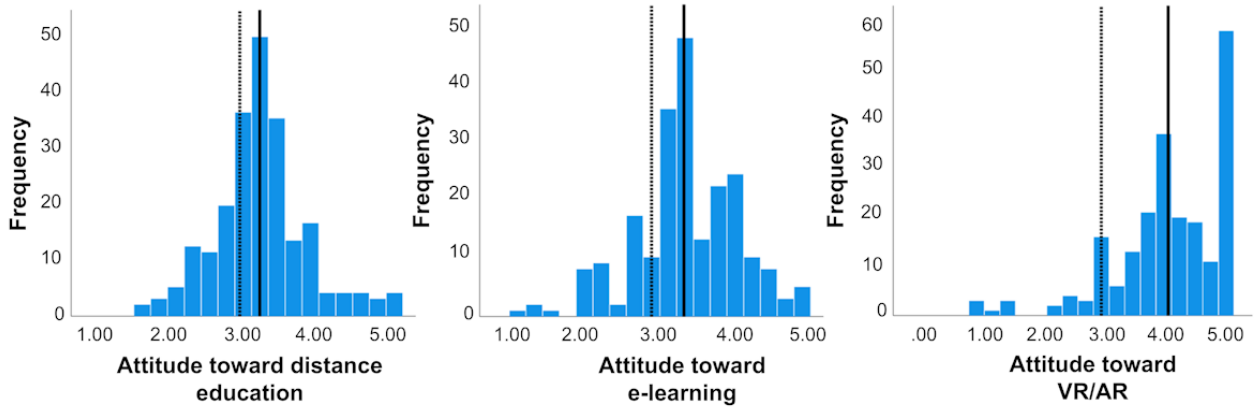
Participants reported their experience across 4 domains: distance education, e-learning, medical simulation, and VR/AR. The Friedman test indicated significant differences in self-reported experience across domains (χ^2_3=285.0; *P*<.001). Post hoc Wilcoxon signed-rank tests indicated that participants' experience with e-learning was significantly higher than their experience with distance education (Wilcoxon Z=3.68; *P*=.001; r_{bs}=0.44), medical simulation (Z=9.30; *P*<.001; r_{bs}=0.88), and VR/AR (Z=11.12; *P*<.001; r_{bs}=0.97). Experience with distance education was also significantly higher than experience with medical simulation (Z=6.62; *P*<.001; r_{bs}=0.61) and VR/AR (Z=9.88; *P*<.001; r_{bs}=0.86). Finally, experience with medical simulation was significantly higher than that with VR/AR (Z=8.34; *P*<.001; r_{bs}=0.88).

The Friedman test showed significant differences in self-reported digital competence (χ^2_4=542.9; *P*<.001). Post hoc analyses revealed that competence in basic tools and online platforms did not differ significantly (Z=2.76; *P*=.06; r_{bs}=0.31), but both were rated significantly higher than each of the other domains of competence (all *z*>10.51, all *P*<.001, and all effect sizes large per Cohen r_{bs} > 0.97). Nearly 80% of participants reported "advanced" or "expert" competence in both domains. Competence in digital problem-solving was rated significantly higher than both competence in advanced technologies (Z=5.55; *P*<.001; r_{bs}=0.55) and VR/AR (Z=9.60; *P*<.001; r_{bs}=0.88). Finally, self-reported competence in advanced technologies was significantly higher than that of VR/AR (Z=7.20; *P*<.001; r_{bs}=0.70). For digital problem-solving, advanced technologies, and VR/AR, 37.79% (82/217), 24.88% (54/217), and 11.06% (24/217) reported "advanced" or "expert" level of competence, respectively. Multimedia Appendix 6 provides a more detailed description of within-domain distributions.

Attitudes Toward e-Learning and Distance Education

Attitudes toward e-Learning and distance education varied among participants. Distance education recorded a mean score of 3.27 (SD 0.63) and a median of 3.18 (IQR 2.95-3.55), indicating moderately positive attitudes, with consistent responses (SD 0.63; skewness 0.37). Statistical analysis revealed significant positivity in the attitudes toward distance education (Wilcoxon Z=6.10; *P*<.001; r_{bs}=0.49; mean 3.27 (SD 0.63); 95% CI 3.19-3.36). e-Learning followed with a mean of 3.43 (SD 0.69) and a median of 3.44 (IQR 3.11-3.89), exhibiting slightly more favorable attitudes with moderate variability (SD 0.69). The Wilcoxon test confirmed the positivity of these attitudes (Z=8.03; *P*<.001; r_{bs}=0.64; mean 3.43 (SD 0.69); 95% CI 3.33-3.52). VR/AR received the highest mean score of 4.07 (SD 0.88) and a mode of 5, reflecting the most positive attitudes, albeit with greater variability (SD 0.88, skewness –1.17), and the analysis significantly supported these positive perceptions (Z=10.94; *P*<.001; r_{bs}=0.89; mean 4.07 (SD 0.88); 95% CI 3.96-4.19). Negative skewness in VR/AR attitudes suggests a ceiling effect, potentially reflecting aspirational responding. Overall, VR/AR elicited the strongest positive responses, although its variability suggested mixed familiarity or application, whereas distance education and e-learning garnered more tempered but consistent favorability (Figure 1).

Figure 1. Histograms of the mean scores on the attitude questionnaires. The dashed line indicates the neutral midpoint (3.00), whereas the solid line indicates the sample mean. AR: augmented reality; e-learning: electronic learning; VR: virtual reality.



Cluster Analysis

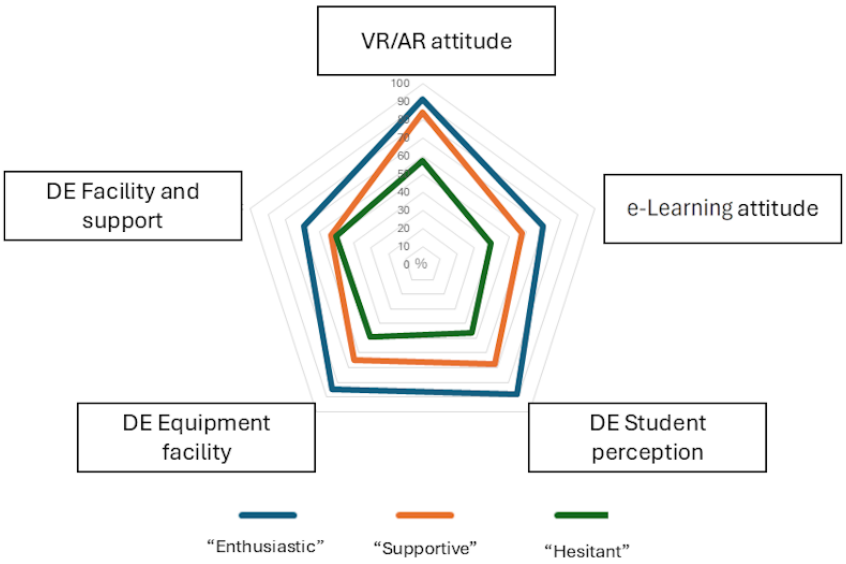
K-means clustering based on responses to the attitude scales, optimized for internal coherence and separation, suggested 3 clusters that provided an optimal balance of granularity and distinctiveness. The average silhouette score was 0.27 for the 3-cluster solution, whereas it was 0.26 and 0.21 for the 2- and 4-cluster solutions, respectively (Multimedia Appendix 2 provides a detailed description of cluster number optimization). The clusters were tentatively named "enthusiastic" (47/217, 21.66%), "supportive" (129/217, 59.45%), and "hesitant" (41/217, 18.89%) (Figure 2). All 3 clusters showed high levels of stability according to mean Jaccard similarities (0.85, 0.86, and 0.89), confirming the robustness of the 3-cluster solution (mean Jaccard similarities for the 2-cluster solution: 0.87 and 0.88; for the 4-cluster solution: 0.57, 0.30, 0.57, and 0.74). The attitudes were measured using scales adapted from validated sources, including the Attitude Toward E-Learning Scale, the Distance Education Perception Scale, and items related to perceptions of VR/AR applications in learning. Enthusiastic respondents showed the highest endorsements across all attitude scales. Supportive individuals expressed favorable attitudes but with moderate enthusiasm compared with the enthusiastic cluster. Hesitant respondents exhibited the lowest agreement

XSL•FO
RenderX

with positive statements regarding all educational technology aspects (Figure 2). Significant differences among clusters were confirmed by 1-way ANOVAs across all attitude dimensions, each showing large effect sizes (η_p^2>0.14): distance education student's perception, $F_{2,214}$=179.85; P<.001; η_p^2=0.63; distance education equipment facility, $F_{2,214}$=132.94; P<.001; η_p^2=0.55; distance education facility and support, $F_{2,214}$=52.27; P<.001; η_p^2=0.33; e-learning engagement, $F_{2,214}$=105.04; P<.001; η_p^2=0.50; and VR/AR engagement, $F_{2,214}$=50.77; P<.001; η_p^2=0.32.

The results of multinomial logistic regression analysis are summarized in Table 2. Global model fit indices indicated that the multinomial regression model provided a significantly better fit than the null model (χ^2_{14}=45.27; P<.001) with a McFadden pseudo-R^2 of 0.11, suggesting a modest yet meaningful amount of variance in cluster membership was accounted for by the predictors. The findings highlighted digital competence as a significant predictor in differentiating between the enthusiastic and hesitant clusters (estimate=0.22; SE=0.07; Z=3.22; P=.001; and odds ratio [OR] 1.25, 95% CI 1.09-1.43) and between the supportive and hesitant clusters (estimate=0.18; SE=0.06; Z=3.07; P=.002; OR 1.19, 95% CI 1.07-1.33). This indicates that higher levels of digital competence are associated with increased odds of belonging to the enthusiastic or supportive clusters rather than the hesitant cluster, suggesting that perceived capability with digital technologies plays a key role in shaping trainees' attitudinal profiles. Economic conditions, indicated by GDP, also played a significant role in distinguishing between the enthusiastic and hesitant groups (estimate=–0.03; SE=0.01; Z=–2.44; P=.02; OR 0.97, 95% CI 0.95-0.99) Similarly, GDP was a significant predictor of cluster membership when the enthusiastic and supportive groups were compared (estimate=–0.03; SE=0.01; Z=–3.02; P=.003; OR 0.97, 95% CI 0.95-0.99), suggesting that trainees from countries with lower GDP were more likely to belong to the enthusiastic cluster. Gender, age, training year, workplace type (university vs nonuniversity clinic), and prior digital experience were not significant predictors (all P>.05). Given that demographic and training characteristics, as well as overall experience with various forms of digital learning, did not strongly differentiate clusters, attitudinal patterns may be largely independent of these background factors.

Figure 2. Radar chart illustrating the distribution of standardized (0%-100%) mean scores across 3 clusters: "enthusiastic," "supportive," and "hesitant." Scores were normalized based on each dimension's theoretical maximum to ensure comparability across variables. The 5 dimensions include attitude toward VR/AR, attitude toward e-learning, distance education (DE) student's perception, DE equipment facility, and DE facility and support.



XSL•FO
RenderX

Table 2. Multinomial regression analysis comparing attitudinal clusters in distance education.

Clusters and predictors	Estimate	SE	Wilcoxon Z	*P* value	OR[a] (95% CI)
Enthusiastic (47/ 217 , 21.66%) –hesitant (41/ 217 , 18.89%)					
Intercept	-5.07	2.35	-2.16	.03	0.01 (0.00-0.63)
Digital_experience	0.1	0.07	1.38	.17	1.11 (0.96-1.28)
Digital_competence[b]	0.22	0.07	3.22	.001	1.25 (1.09-1.43)
Training_year	0.09	0.14	0.65	.52	1.1 (0.83-1.46)
Sex					
Male–female	0.5	0.58	0.86	.39	1.65 (0.52-5.19)
Workplace					
University clinic– nonuniversity clinic	-0.6	0.48	-1.24	.21	0.55 (0.21-1.41)
GDP[c] (US $1000)[b]	-0.03	0.01	-2.44	.02	0.97 (0.95-0.99)
Age (years)	0.03	0.07	0.38	.70	1.03 (0.89-1.19)
Supportive (129/ 217 , 59.45%) –hesitant (41/ 217 , 18.89%)					
Intercept	-0.45	1.9	-0.24	.81	0.64 (0.02-26.52)
Digital_experience	0	0.06	0.02	.99	1 (0.89-1.12)
Digital_competence[b]	0.18	0.06	3.07	.002	1.19 (1.07-1.33)
Training_year	0.11	0.12	0.93	.35	1.11 (0.89-1.39)
Sex					
Male–female	0	0.42	0.01	.99	1 (0.44-2.29)
Workplace					
University clinic–nonuniversity clinic	-0.39	0.39	-1.01	.31	0.67 (0.31-1.45)
GDP_(US $1000)	0	0.01	0.04	.97	1 (0.98-1.02)
Age (years)	-0.04	0.06	-0.66	.51	0.96 (0.85-1.08)
Enthusiastic (47/ 217 , 21.66%) –supportive (129/ 217 , 59.45%)					
Intercept	-4.62	1.78	-2.60	.009	0.01 (0.00-0.32)
Digital_experience	0.10	0.06	1.75	.08	1.11 (0.99-1.24)
Digital_competence	0.05	0.05	0.94	.35	1.05 (0.95-1.16)
Training_year	-0.01	0.11	-0.12	.91	0.99 (0.79-1.23)
Sex					
Male–female	0.50	0.49	1.02	.31	1.64 (0.63-4.28)
Workplace					
University clinic–nonuniversity clinic	-0.21	0.37	-0.56	.58	0.81 (0.39-1.69)
GDP (US $1000)[b]	-0.03	0.01	-3.02	.003	0.97 (0.95-0.99)
Age (years)	0.07	0.06	1.23	.22	1.07 (0.96-1.19)

[a]OR: odds ratio.

[b]Denotes statistically significant values.

[c]GDP: gross domestic product.

Attitude as a Function of Workplace and GDP of the Country Hosting the Training Program

The results of the ANOVAs investigating the effects of GDP and workplace on attitudes are depicted in Figure 3 and Multimedia Appendix 7. The main effect of workplace reached significance for the student's perception subscale ($F_{1,211}$=6.95; P=.009; η_p^2=0.03). Trainees from nonuniversity clinics had significantly more positive attitudes toward distance learning than trainees from university clinics.

The analysis of the equipment facility subscale yielded a significant GDP main effect ($F_{2,211}$=5.02; P=.007; η_p^2=0.05).

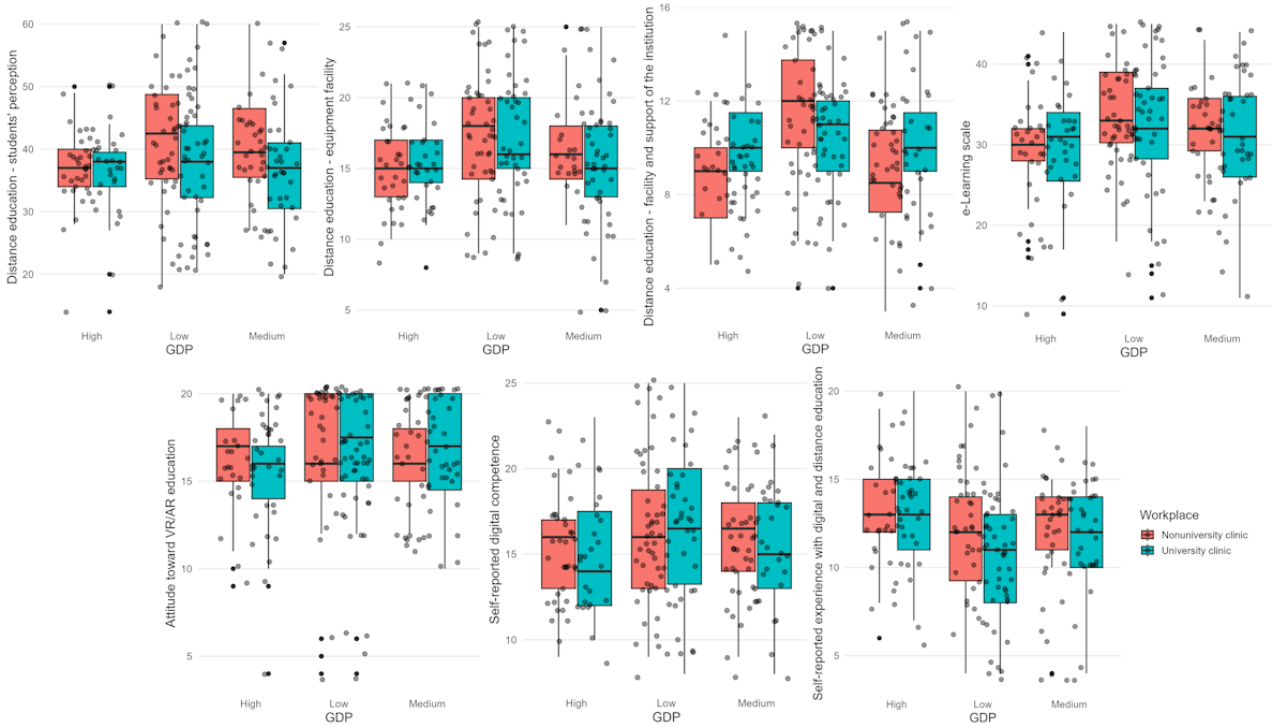
XSL•FO
RenderX

Corrected post hoc tests showed that the scores of trainees from low-GDP countries were significantly higher than those of trainees from high-GDP countries (mean difference –2.07, 95% CI –3.64 to –0.49; t_{211}=–3.10; 2-tailed *P*=.007); Cohen *d*=–0.52). Similarly, there was a significant GDP main effect for facility and support of the institution ($F_{2,211}$=7.84; *P*<.001; η_p^2=0.07), as well as a significant Workplace × GDP interaction ($F_{2,211}$=5.16; *P*=.007; η_p^2=0.05). Simple effects analyses revealed that the GDP main effect was only significant for trainees from nonuniversity clinics ($F_{2,211}$=13.24; *P*<.001; η_p^2=0.11). Within this group, scores of trainees from low-GDP countries were significantly higher than those of trainees from medium-GDP (mean difference 2.17, 95% CI 0.33 to 4.01; t_{211}=3.39; 2-tailed *P*=.01); Cohen *d*=–1.11) and high-GDP countries (mean difference –2.73, 95% CI –4.35 to –1.11; t_{211}=–4.85; 2-tailed *P*<.001; Cohen *d*=–1.11). The analyses also showed a significant workplace effect within the group of trainees from high-GDP countries ($F_{1,211}$=6.03; *P*=.02; η_p^2=0.03): scores were higher for trainees from university clinics than for other trainees.

The analysis of attitudes toward e-learning also yielded a significant GDP main effect ($F_{2,211}$=4.59; *P*=.01; η_p^2=0.04): trainees from low-GDP countries showed significantly more positive attitude toward e-learning than trainees from high-GDP countries (mean difference –3.62, 95% CI –6.45 to –0.80; t_{211}=–3.03; 2-tailed *P*=.008; Cohen *d*=–0.51).

Finally, a significant GDP main effect was found for experience with technologies ($F_{2,211}$=3.65; *P*=.03; η_p^2=0.03): trainees from high-GDP countries had significantly more experience than trainees from low-GDP countries (mean difference 1.48, 95% CI 0.14 to 2.82; t_{211}=2.60; 2-tailed *P*=.03; Cohen *d*=0.44). No further significant main effects or interactions were found.

Figure 3. Results of the questionnaire scores as a function of workplace and gross domestic product (GDP) of the country hosting the training program.

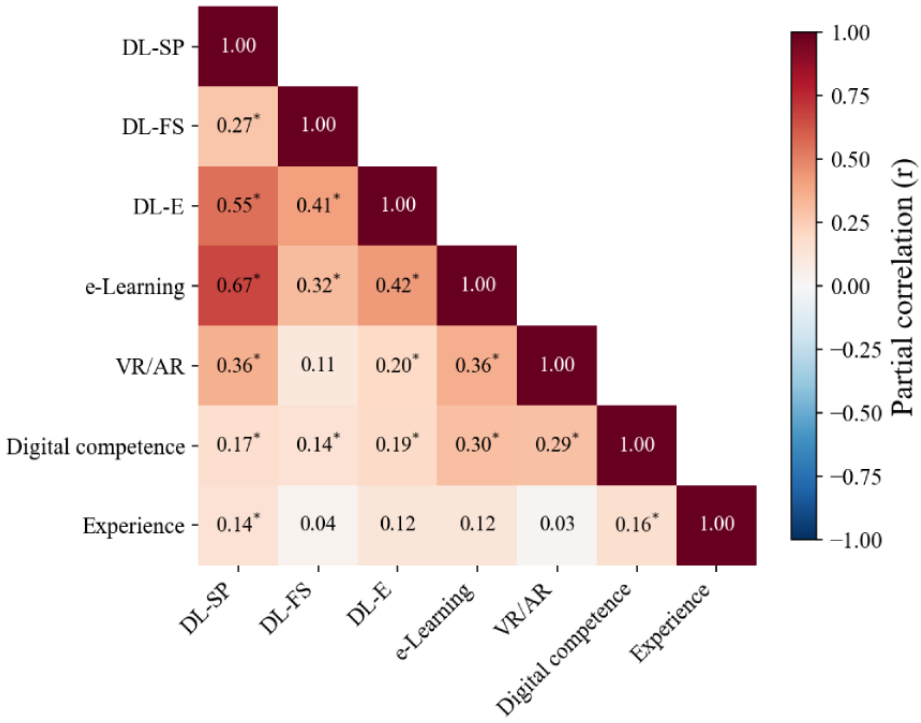


Associations Among Digital Competence, Experience, and Attitude

Correlation analyses (Figure 4) showed digital competence was strongly associated with favorable attitudes across all learning modalities, especially in low- and medium-GDP countries (Multimedia Appendix 8). Prior experience also correlated with more positive attitudes, highlighting the role of familiarity. Positive views of one modality often extended to others, reflecting broad openness to digital education, consistent with theoretical frameworks emphasizing usability and prior exposure.

XSL•FO
RenderX

Figure 4. Correlations and partial correlations between attitude scales, digital competence, and experience with digital technology. Control variables were digital competence and experience with digital learning. When competence or experience was correlated with an attitude variable, the other variable served as the control variable. No control variables were applied when examining the correlation between competence and experience. *Indicates statistical significance (*P*<.05). DL-E: distance learning–equipment facility; DL-FS: distance learning–facility and support of the institution; DL-SP: distance learning–students' perceptions; e-Learning: electronic learning; VR/AR: virtual reality/augmented reality.



Factor Analysis of Attitude Items

The Mardia test showed significant multivariate skewness (b_{1p}=307.11; *P*<.001) and kurtosis (b_{2p}=1379.52; *P*<.001), indicating violation of multivariate normality. Consequently, a polychoric correlation matrix was used in all subsequent analyses. Factorability was supported by the Bartlett test of sphericity (χ^2_{595}=7216.854; *P*<.001), the Kaiser-Meyer-Olkin index =0.78, and the determinant of the correlation matrix (3.97×10^{-16}). Interitem correlations were generally moderate (median r =0.368, range 0-0.925). Parallel analysis indicated 6 factors; however, conceptual overlap among related dimensions motivated a theory-driven consolidation (eg, distance education engagement and conditions → distance education perception), yielding a more parsimonious 4-factor solution. In this reduced model, after rotation, the 4 factors explained 20.8%, 14.2%, 11.8%, and 10.7% of the variance (total=57.5%). No Heywood cases were identified. More information is available in Multimedia Appendix 3.

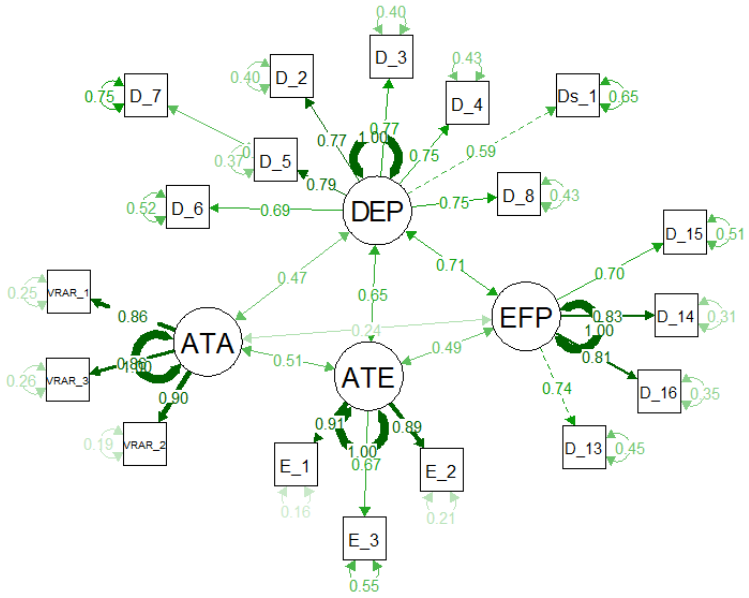
For model refinement, items with strong primary loadings and adequate communalities were retained, while theoretical interpretability was also considered. Further, 8 items (e-learning items 6 and 10; distance education items 7, 10, 12, and 20-22) were removed because of low communalities. Distance education item 8 was excluded because of a low loading (0.38) and cross-loading. e-Learning items 7-9 and 11 were excluded because they loaded on a theoretically implausible factor (distance education perception rather than e-learning).

The resulting model comprised 4 latent constructs: distance education perception, equipment facility perception, attitude toward VR/AR, and attitude toward e-learning. This structure was evaluated using confirmatory factor analysis. The initial model showed acceptable fit (χ^2_{203}=370.57; CFI=0.938; TLI=0.93; RMSEA=0.062; SRMR=0.052), although modification indices indicated potential improvement. Accordingly, a residual covariance was specified between distance education items 13 and 14 (MI=33.49), resulting in improved fit (ΔCFI=0.01; ΔTLI=0.011; ΔRMSEA=–0.005; ΔSRMR=0.002). Subsequent indices suggested an additional residual covariance between e-learning items 1 and 2 (MI=24.52). In both cases, the items were semantically similar and reflected closely related constructs, justifying correlated residuals (Multimedia Appendix 3).

The final model (Figure 5) demonstrated good fit (χ^2_{201}=323.203; CFI=0.955; TLI=0.948; RMSEA=0.053; SRMR=0.053). Further modifications were not pursued despite several MIs>10 to avoid overfitting. The model showed moderate to high correlations among the constructs, reflecting their conceptual interdependencies within the domain of e-learning environments. In more detail, interfactor correlations demonstrated meaningful relationships, with distance education perception positively associated with equipment facility perception, attitude toward VR/AR, and attitude toward e-learning, and equipment facility perception significantly correlated with both attitude toward VR/AR and attitude toward e-learning, reflecting theoretical interdependencies across educational dimensions. Findings from this more parsimonious model were consistent with the partial correlation patterns reported earlier, supporting the stability of the factor structure.

XSL•FO
RenderX

Figure 5. Interfactor correlations among constructs in the confirmatory factor analysis model. ATA: attitude toward VR/AR; ATE: attitude toward e-learning; DEP: distance education perception; EFP: equipment facility perception.



Discussion

Overview

Digital tools have advanced medical education, particularly in procedural fields such as orthopedics, where tactile skills challenge remote learning. VR/AR simulations enabling safe rehearsal (eg, arthroscopy knee simulators show operating room transferability, and hip arthroscopy VR matches benchtop training) [44-46]. This study highlights the importance of digital competence and familiarity in shaping attitudes, aligning with findings from other disciplines such as nursing and parallel surgical evidence from veterinary medicine [47-50].

Our finding that VR/AR elicited the most positive attitudes (mean 4.07) despite limited exposure aligns with recent umbrella reviews showing VR's growing dominance in surgical training and procedural skill acquisition [34]. The findings of Sommer et al [51] regarding VR's reduction of errors and improvement of completion times in surgical simulations (60 students; 2-arm randomized trial) are consistent with the trainee optimism reported in this study. This is corroborated by the findings from Pettinelli et al [13], whose systematic review with a total sample size of 148 residents or students affirmed that VR is effective for learning specific arthroplasty techniques (no difference between VR-trained and control groups; P=.56; I^2=85%), further supporting the notion that residents are open to integrating this technology into their training repertoire. However, similar to the narrative review by Ivanov et al [52], we note that statistical significance in performance outcomes remains inconsistent, underscoring the need for larger longitudinal studies.

Structural equation modeling (SEM) revealed factors that influence clinical cognition, health behavior, or the acceptance of digital learning modalities [53-55]. Our analysis demonstrates that engagement with distance education shows a strong association with acceptance, suggesting a bidirectional influence between perceptions and intent in shaping residents' comfort with e-learning platforms and VR/AR tools. This finding underscores the importance of fostering engagement through distance education to build familiarity with advanced digital tools. SEM associations (eg, distance education acceptance) are correlational, precluding causality; bidirectionality is likely (eg, comfort shapes perceptions and vice versa). Longitudinal studies are needed to test directions and moderators such as competence.

Our SEM findings, which link digital confidence to VR/AR acceptance, support the "blended digital comfort" strategies described by Cabero-Almenara et al [56] (online questionnaire involving 17,301 Latin American students). For example, pairing VR modules with hands-on workshops could reduce effort expectancy (a UTAUT construct), as observed in proof-of-concept mixed-reality ward rounds by Bala et al [57] involving 11 students, in which 82% of trainees reported improved interaction. Tool proficiency is key for acceptance and integration.

The interplay between e-learning dimensions, such as blended digital comfort's influence on e-learning attitudes, emphasizes the need to address broader digital literacy and confidence. Consistent with prior research (multinational survey involving 223 students), positive attitudes were more pronounced in settings with lower access to conventional resources, such as noncentral hospitals and low-GDP countries, where distance and e-learning provide critical educational support [58]. This is particularly important for designing interventions that integrate multiple technologies effectively.

Beyond economic resources, sociocultural factors also shape regional differences in technology adoption; national cultural dimensions (uncertainty avoidance, individualism/collectivism, and power distance) moderate links among ease of use, perceived usefulness, and intention to adopt [59]. Trainees from lower-GDP countries showed more positive attitudes, possibly reflecting adaptive innovation cultures rather than mere scarcity, whereas high-GDP regions may have more rigid educational norms and accreditation structures that slow adoption

XSL•FO
RenderX

(Multimedia Appendix 9 presents the Hofstede dimensions of countries included in this study) [60]. Effective VR-based orthopedic education therefore requires strategies tailored to both economic capacity and cultural context: in lower-income settings, low-cost standalone VR, local VR "champions," and team-based modules can support uptake, whereas in higher-income settings, integration into accredited curricula, assessment frameworks, and faculty development programs helps overcome inertia. Across contexts, culturally responsive leadership, clear standards, and evaluation using frameworks such as the Consolidated Framework for Implementation Research and reach, effectiveness, adoption, implementation, and maintenance (RE-AIM) are needed to sustain engagement, equity, and long-term integration [61,62].

TAM and UTAUT provide a useful interpretive lens for our attitudinal patterns, linking them to concepts such as perceived usefulness, ease of use, and contextual enablers. In our model, attitude toward e-learning and distance education perception map onto perceived usefulness and performance expectancy, with digital competence and experience reflecting effort expectancy, and workplace type and GDP representing facilitating conditions and social influence. Digital competence showed positive associations with attitudes toward e-learning, distance education, and VR/AR—patterns consistent with TAM and UTAUT concepts, while institutional effects align with UTAUT's emphasis on environmental support and broader trends in which novel tools elicit optimism despite limited familiarity, although we did not estimate a full structural model (eg, with R^2 for acceptance) [63,64]. Perceived ease of use, digital confidence, and facilitating conditions—access to hardware, connectivity, and institutional backing—are therefore critical and can be strengthened through user-friendly tools, targeted training, institutional support, subsidized technology, and partnerships with providers to promote equitable access [65].

Implications for Educational Design

The findings suggest a tiered approach to educational design. At the core (immediate) level, digital competence training should be embedded within curricula through targeted workshops and organized testing campaigns to enhance familiarity and effort expectancy. At the supportive (medium-term) level, access to digital tools should be subsidized to address disparities associated with lower GDP and strengthen facilitating conditions. Collaborative programs and partnerships between institutions in lower- and higher-GDP countries should also be promoted to support resident training, faculty development, and the sharing of educational resources and programs. At the advanced (long-term) level, VR/AR modules and e-learning platforms can be integrated into hybrid educational models, with content and implementation strategies tailored to local cultural and contextual needs.

Limitations

Limitations include the unknown size of the trainee population, uneven geographic distribution, inability to characterize nonrespondents, and reliance on self-reported data. Due to the absence of standardized, publicly available databases enumerating orthopedic resident characteristics across European countries, it is not possible to ascertain whether the proportions observed in our sample reflect the true population. Uneven geographic representation limits the ability to extrapolate findings uniformly across all European contexts. Exclusion bias might have affected the results because of the higher exclusion rate among trainees working in nonuniversity clinics (10 of 18 exclusions were due to beginner-level English proficiency). The reliance on self-reported attitude data may have introduced social desirability bias, potentially inflating positive responses toward digital tools, as evidenced by the VR ceiling. No social desirability scale was used. Nonetheless, the findings are likely generalizable because diverse demographics (29 countries and balanced GDP and workplace types) support generalizability, as do global similarities in medical training and digital education challenges. Future multimethod validation (eg, objective logs) is needed; longitudinal studies are needed to test causality.

Future Perspectives

Further studies should assess the effects of sustained digital platform exposure on attitudes and outcomes over time, using mixed methods designs. Qualitative components should examine VR/AR as innovative tools for orthopedic and trauma trainee education. Phenomenological studies can explore trainees' lived experiences with VR/AR tools, while institutional case studies can reveal local implementation mechanisms. Quantitative designs (exploratory, experimental, and longitudinal) should assess skills acquisition via the objective structured assessment of technical skills and the objective structured clinical examination across residency. Translational research should examine the impact of digital skills on patient outcomes and their correlations with attitudes, motivation, and retention. All studies should be grounded in self-directed learning, cognitive load, and deliberate practice theories for design and interpretation.

Based on the findings, future research should focus on 3 main directions:

- Longitudinal randomized controlled trials (priority): track TAM and UTAUT pathways over time; test interventions (eg, competence training) on actual method adoption in multisite trials (n>500).
- Qualitative studies: phenomenology or interviews elucidate mechanisms (eg, barriers among hesitant trainees and drivers of enthusiasm in low-GDP countries).
- Mixed methods evaluations: assess policy impacts (eg, subsidies and curricula) via surveys and usage logs in hybrid programs across Europe's diverse socioeconomic landscape.

To operationalize these research directions and secure funding, we propose a scalable 4-step approach (Textbox 1) aligned with European Union schemes (Erasmus+ and Horizon Europe) and frameworks (Consolidated Framework for Implementation Research for determinants; RE-AIM, and dynamic sustainability framework for sustainability). Each includes milestones, risks, and mitigations (more detailed suggestions are provided in Multimedia Appendix 10).

XSL•FO
RenderX

Textbox 1. Scalable 4-step implementation and funding roadmap for virtual reality/augmented reality (VR/AR) surgical training in the European Union.

Step 1 (year 1)

Generate feasibility evidence via pilot studies in 3 to 5 centers (approximately 30-50 trainees). End points: recruitment rates, usability, and faculty burden. Fund curriculum and cross-border activities via Erasmus+ Small-scale Partnerships; mitigate heterogeneous infrastructure with device specifications and train-the-trainer programs.

Step 2 (years 2-3)

Conduct preregistered multicenter studies (n≥150 trainees) with standardized outcomes (objective structured assessment of technical skills and the Surgical Council on Resident Education), efficiency analyses, and cost-benefit analyses. Compare virtual reality (VR) and augmented reality (AR) with traditional methods; mitigate bias via cluster designs and harmonized training.

Step 3 (years 3-4)

Establish 2 to 4 regional VR and AR hubs for training, content sharing, and faculty development, including low-resource sites. Leverage the Society in Europe for Simulation Applied to Medicine and European Digital Innovation Hubs networks; mitigate fragmentation via interoperability, shared procurement, and equitable access.

Step 4 (years 4-5)

Conduct multicountry scale-out with RE-AIM (reach, effectiveness, adoption, implementation, and maintenance) evaluation; test sustainable models (eg, subscription-based approaches). This approach mirrors European Union initiatives; mitigate delays via shared contracts and General Data Protection Regulation (GDPR) templates for scalable, equitable VR and AR education.

Conclusion

Analysis of this international multicountry survey highlights that foundational engagement and confidence are pivotal for promoting advanced adoption of VR and AR technologies and optimizing the use of digital tools. Programs should prioritize an interactive, well-supported, and blended approach to accommodate varying levels of digital familiarity and preferences.

The application of TAM and UTAUT in this study supports their utility in understanding technology adoption in medical education. By addressing perceived usefulness, ease of use, social influence, and facilitating conditions, institutions can design strategies to enhance digital tool integration. In this context, tailored implementation strategies should be developed to optimize digital education adoption. These theoretical advancements inform future research and guide pragmatic policy design aimed at integrating digital tools into postgraduate orthopedic training.

Our findings indicate that digital competence strongly influences acceptance of technology-enhanced education among orthopedic trainees. Therefore, implementation strategies should prioritize competence-building initiatives, such as structured digital literacy programs, as foundational steps. Following competency development, scalable interventions such as VR module integration can be implemented more effectively.

This study advances the understanding of factors influencing digital learning adoption in education, providing a foundation for designing effective, equitable, and innovative learning interventions aligned with trainees' needs and preferences.

Orthopedic training must advance beyond pandemic-driven digital adoption. While e-learning remains a foundational element, the untapped potential of VR and AR technologies necessitates targeted investments in infrastructure, competency-based curricula, and cross-institutional collaborations. Consistent with the TAM and UTAUT, addressing socioeconomic inequities through policies (such as establishing European Union–funded VR hubs in low-GDP regions) can translate trainee enthusiasm into measurable clinical proficiency.

Beyond educational equity, improving access to high-quality digital and simulation-based training is expected to have downstream clinical relevance, as randomized-trial evidence suggests that simulation training can improve patient-related outcomes in surgery. While our findings are European, the proposed competence-first and equity-oriented implementation pathway could be adapted through partnerships with institutions in the Global South, aligned with World Health Organization priorities for equitable digital health capacity building and the broader global agenda for strengthening surgical systems and workforce development [66,67].

Acknowledgments

The authors gratefully acknowledge the expert assistance of Péter Hillebrand, whose skillful image editing and creative development of the graphical abstract greatly contributed to the quality of this work.

The authors declare the use of generative artificial intelligence (GAI) in the writing process. According to the GAIDeT taxonomy (2025), the following tasks were delegated to GAI tools under full human supervision: proofreading and editing, translation, and reformatting. The GAI tool used was Perplexity (Perplexity AI). Responsibility for the final manuscript lies entirely with the authors. GAI tools are not listed as authors and do not bear responsibility for the final outcomes.

XSL•FO
RenderX

Funding

The study was supported by the National Laboratory of Cooperative Technologies project (project number 2022-2.1.1-NL-2022-00012), provided by the Ministry of Culture and Innovation from the National Research, Development, and Innovation Fund, and financed by the National Laboratories Programme.

Data Availability

The datasets generated or analyzed during this study are available in the Mendeley repository [43].

Authors' Contributions

Conceptualization: ATS (lead), AM (equal)
Data curation: MO (equal), IU (equal), GC (equal), TS (equal), FLS (equal), VN (equal)
Formal analysis: AK (lead), AM (equal), ATS (supporting), PDM (supporting), VGI (supporting), MD (supporting)
Funding acquisition: PM
Investigation: MO (equal), IU (equal), GC (equal), VGI (equal), DK (equal), MD (equal)
Methodology: ATS (lead), AM (equal), PM (supporting), AK (supporting)
Software: AM (lead), ATS (supporting), MD (supporting), FLS (supporting)
Supervision: PM (lead), AK (equal), LT (supporting)
Validation: AK (lead), PDM (supporting), VGI (supporting), DK (supporting), VN (supporting), LT (supporting), AM (supporting)
Visualization: AM (lead), TS (supporting), FLS (supporting)
Writing – original draft: ATS (lead), AM (equal), MO (supporting), MM (supporting), DK (supporting), LT (supporting), AK (supporting), PM (supporting)

Conflicts of Interest

None declared.

Multimedia Appendix 1
Survey questions.
[DOCX File , 24 KB - mededu_v12i1e79418_app1.docx]

Multimedia Appendix 2
Elbow and Silhouette methods.
[DOCX File , 47 KB - mededu_v12i1e79418_app2.docx]

Multimedia Appendix 3
Results of the factor analyses.
[DOCX File , 100 KB - mededu_v12i1e79418_app3.docx]

Multimedia Appendix 4
R packages.
[DOCX File , 32 KB - mededu_v12i1e79418_app4.docx]

Multimedia Appendix 5
Geographical distribution of respondents.
[DOCX File , 418 KB - mededu_v12i1e79418_app5.docx]

Multimedia Appendix 6
Analyses of experience and competence levels.
[DOCX File , 633 KB - mededu_v12i1e79418_app6.docx]

Multimedia Appendix 7
Results of the questionnaire scores as a function of workplace and gross domestic product (GDP) of the country hosting the training program.
[DOCX File , 390 KB - mededu_v12i1e79418_app7.docx]

Multimedia Appendix 8

XSL•FO
RenderX

Associations of attitude variables with experience and digital competence as a function of gross domestic product (GDP) category.
[DOCX File , 598 KB - mededu_v12i1e79418_app8.docx]

Multimedia Appendix 9
Characteristics of low-, medium-, and high–gross domestic product (GDP) countries based on Hofstede's dimensions of culture.
[DOCX File , 16 KB - mededu_v12i1e79418_app9.docx]

Multimedia Appendix 10
Detailed description of the operationalization of the suggested research directions.
[DOCX File , 17 KB - mededu_v12i1e79418_app10.docx]

References

1. Ogundiya O, Rahman TJ, Valnarov-Boulter I, Young TM. Looking back on digital medical education over the last 25 years and looking to the future: narrative review. J Med Internet Res 2024;26:e60312 [FREE Full text] [doi: 10.2196/60312] [Medline: 39700490]
2. Chowdhury PN, Vaish A, Puri B, Vaishya R. Medical education technology: past, present and future. Apollo Med 2024;21(4):374-380. [doi: 10.1177/09760016241256202]
3. Muttappallymyalil J, Mendis S, John LJ, Shanthakumari N, Sreedharan J, Shaikh RB. Evolution of technology in teaching: blackboard and beyond in medical education. Nepal J Epidemiol 2016;6(3):588-592 [FREE Full text] [doi: 10.3126/nje.v6i3.15870] [Medline: 27822404]
4. Han H, Resch DS, Kovach RA. Educational technology in medical education. Teach Learn Med 2013;25 Suppl 1:S39-S43. [doi: 10.1080/10401334.2013.842914] [Medline: 24246105]
5. Satava RM. Virtual reality surgical simulator: the first steps. Surg Endosc 1993;7(3):203-205. [doi: 10.1007/BF00594110] [Medline: 8503081]
6. Cook DA, Levinson AJ, Garside S, Dupras DM, Erwin PJ, Montori VM. Internet-based learning in the health professions: a meta-analysis. JAMA 2008;300(10):1181-1196. [doi: 10.1001/jama.300.10.1181] [Medline: 18780847]
7. Stirling ERB, Lewis TL, Ferran NA. Surgical skills simulation in trauma and orthopaedic training. J Orthop Surg Res 2014;9:126 [FREE Full text] [doi: 10.1186/s13018-014-0126-z] [Medline: 25523023]
8. Bogár PZ, Tóth L, Rendeki S, Mátyus L, Németh N, Boros M, et al. [The present and the future of medical simulation education in Hungary]. Orv Hetil 2020;161(26):1078-1087. [doi: 10.1556/650.2020.31761] [Medline: 32541086]
9. Cannon WD, Garrett WE, Hunter RE, Sweeney HJ, Eckhoff DG, Nicandri GT, et al. Improving residency training in arthroscopic knee surgery with use of a virtual-reality simulator: a randomized blinded study. J Bone Joint Surg Am 2014;96(21):1798-1806. [doi: 10.2106/JBJS.N.00058] [Medline: 25378507]
10. Sugand K, Akhtar K, Khatri C, Cobb J, Gupte C. Training effect of a virtual reality haptics-enabled dynamic hip screw simulator. Acta Orthop 2015;86(6):695-701 [FREE Full text] [doi: 10.3109/17453674.2015.1071111] [Medline: 26168925]
11. Gani A, Pickering O, Ellis C, Sabri O, Pucher P. Impact of haptic feedback on surgical training outcomes: a randomised controlled trial of haptic versus non-haptic immersive virtual reality training. Ann Med Surg (Lond) 2022;83:104734 [FREE Full text] [doi: 10.1016/j.amsu.2022.104734] [Medline: 36389184]
12. Bogar PZ, Virag M, Bene M, Hardi P, Matuz A, Schlegl AT, et al. Validation of a novel, low-fidelity virtual reality simulator and an artificial intelligence assessment approach for peg transfer laparoscopic training. Sci Rep 2024;14(1):16702 [FREE Full text] [doi: 10.1038/s41598-024-67435-6] [Medline: 39030307]
13. Pettinelli NJ, Lee AY, Lee MS, Mahatme RJ, Gillinov SM, Jimenez AE. Virtual reality is an effective tool for learning techniques in arthroplasty: a systematic review and meta-analysis. JAAOS Glob Res Rev 2023;7(6). [doi: 10.5435/jaaosglobal-d-23-00078]
14. Cate G, Barnes J, Cherney S, Stambough J, Bumpass D, Barnes CL, et al. Current status of virtual reality simulation education for orthopedic residents: the need for a change in focus. Global Surg Educ 2023;2(1):46 [FREE Full text] [doi: 10.1007/s44186-023-00120-w] [Medline: 38013875]
15. Dedeilia A, Sotiropoulos MG, Hanrahan JG, Janga D, Dedeilias P, Sideris M. Medical and surgical education challenges and innovations in the COVID-19 era: a systematic review. In Vivo 2020;34(3 Suppl):1603-1611 [FREE Full text] [doi: 10.21873/invivo.11950] [Medline: 32503818]
16. Schlégl Á, Pintér Z, Kovács A, Kopjár E, Varga P, Kardos D, et al. Teaching basic surgical skills using homemade tools in response to COVID-19. Acad Med 2020;95(11):e7 [FREE Full text] [doi: 10.1097/ACM.0000000000003586] [Medline: 32657784]
17. Zhu Q, Liang B, Wang X, Sun X, Wang L. Force–torque intraoperative measurements for femoral shaft fracture reduction. Comput Assist Surg 2016;21(sup1):37-44. [doi: 10.1080/24699322.2016.1240311]
18. Benjamin M, Sabri O. Using haptic feedback in a virtual reality bone drilling simulation to reduce plunge distance. Cureus 2021;13(9):e18315 [FREE Full text] [doi: 10.7759/cureus.18315] [Medline: 34722082]

XSL•FO
RenderX

19. Trute RJ, Alijani A, Erden MS. Visual cues of soft-tissue behaviour in minimal-invasive and robotic surgery. J Robot Surg 2024;18(1):401. [doi: 10.1007/s11701-024-02150-y] [Medline: 39508918]
20. Schostek S, Schurr MO, Buess GF. Review on aspects of artificial tactile feedback in laparoscopic surgery. Med Eng Phys 2009;31(8):887-898. [doi: 10.1016/j.medengphy.2009.06.003] [Medline: 19595620]
21. von Bechtolsheim F, Bielert F, Schmidt S, Buck N, Bodenstedt S, Speidel S, et al. Can you feel the force just right? Tactile force feedback for training of minimally invasive surgery-evaluation of vibration feedback for adequate force application. Surg Endosc 2024;38(7):3917-3928. [doi: 10.1007/s00464-024-10919-3] [Medline: 38834723]
22. Rangarajan K, Davis H, Pucher PH. Systematic review of virtual haptics in surgical simulation: a valid educational tool? J Surg Educ 2020;77(2):337-347. [doi: 10.1016/j.jsurg.2019.09.006] [Medline: 31564519]
23. Shilaskar S, Shinde M, Tadavi S, Shevale P. VR based medical procedure simulator with haptic feedback gloves. 2024 Presented at: International Conference on Emerging Systems and Intelligent Computing (ESIC); February 9-10, 2024; Bhubaneswar p. 533-537. [doi: 10.1109/esic60604.2024.10481573]
24. Woodward CJ, Khan O, Aydın A, Dasgupta P, Sinha J. Simulation-based training in orthopedic surgery: a systematic review. Curr Probl Surg 2025;63:101676 [FREE Full text] [doi: 10.1016/j.cpsurg.2024.101676] [Medline: 39922638]
25. Harris M, Burke J, Yassin N. Technology-enhanced assessment of surgical training. Bull R Coll Surg Engl 2024;106(S1):28-33. [doi: 10.1308/rcsbull.tb2024.5]
26. Salaja B, Feeley A, Feeley I, Sheehan E, Merghani K. Virtual reality simulation in orthopaedic surgical training during periods of restricted clinical hours: a systematic review. J Surg Simul 2021:18-28. [doi: 10.1102/2051-7726.2021.0003]
27. McKinney B, Dbeis A, Lamb A, Frousiakis P, Sweet S. Virtual reality training in unicompartmental knee arthroplasty: a randomized, blinded trial. J Surg Educ 2022;79(6):1526-1535. [doi: 10.1016/j.jsurg.2022.06.008] [Medline: 35863958]
28. Wolf MA, Pizanis A, Fischer G, Langer F, Scherber P, Stutz J, et al. COVID-19: a catalyst for the digitization of surgical teaching at a German University Hospital. BMC Med Educ 2022;22(1):308 [FREE Full text] [doi: 10.1186/s12909-022-03362-2] [Medline: 35459175]
29. Zhao Y, Watterston J. The changes we need: Education post COVID-19. J Educ Chang 2021;22(1):3-12 [FREE Full text] [doi: 10.1007/s10833-021-09417-3] [Medline: 38624465]
30. Davis FD. Perceived usefulness, perceived ease of use, and user acceptance of information technology. MIS Quarterly 1989;13(3):319. [doi: 10.2307/249008]
31. Venkatesh, Morris, Davis, Davis. User acceptance of information technology: toward a unified view. MIS Quarterly 2003;27(3):425. [doi: 10.2307/30036540]
32. Lewis KO, Cidon MJ, Seto TL, Chen H, Mahan JD. Leveraging e-learning in medical education. Curr Probl Pediatr Adolesc Health Care 2014;44(6):150-163. [doi: 10.1016/j.cppeds.2014.01.004] [Medline: 24981664]
33. Moore JL, Dickson-Deane C, Galyen K. e-Learning, online learning, and distance learning environments: are they the same? Internet High Educ 2011;14(2):129-135. [doi: 10.1016/j.iheduc.2010.10.001]
34. Tene T, Vique López DF, Valverde Aguirre PE, Orna Puente LM, Vacacela Gomez C. Virtual reality and augmented reality in medical education: an umbrella review. Front Digit Health 2024;6:1365345 [FREE Full text] [doi: 10.3389/fdgth.2024.1365345] [Medline: 38550715]
35. Lee JWY, Tan JY, Bello F. Technology acceptance model in medical education: systematic review. JMIR Med Educ 2025;11:e67873 [FREE Full text] [doi: 10.2196/67873] [Medline: 40668663]
36. Lee AT, Ramasamy RK, Subbarao A. Understanding psychosocial barriers to healthcare technology adoption: a review of TAM technology acceptance model and unified theory of acceptance and use of technology and UTAUT frameworks. Healthcare (Basel) 2025;13(3):250 [FREE Full text] [doi: 10.3390/healthcare13030250] [Medline: 39942440]
37. Soriano GP. Psychometric properties of ‘attitude towards e-learning scale’ among nursing students. Int J Educ Sci 2020;30(1-3):1-5. [doi: 10.31901/24566322.2020/30.1-3.1135]
38. Özkaya G, Aydin MO, Alper Z. Distance education perception scale for medical students: a validity and reliability study. BMC Med Educ 2021;21(1):400 [FREE Full text] [doi: 10.1186/s12909-021-02839-w] [Medline: 34311725]
39. Shen S, Xu K, Sotiriadis M, Wang Y. Exploring the factors influencing the adoption and usage of augmented reality and virtual reality applications in tourism education within the context of COVID-19 pandemic. J Hosp Leis Sport Tour Educ 2022;30:100373. [doi: 10.1016/j.jhlste.2022.100373] [Medline: 35069015]
40. Artino AR, Phillips AW, Utrankar A, Ta AQ, Durning SJ. "The Questions Shape the Answers": assessing the quality of published survey instruments in health professions education research. Acad Med 2018;93(3):456-463. [doi: 10.1097/ACM.0000000000002002] [Medline: 29095172]
41. Faul F, Erdfelder E, Lang A, Buchner A. G*Power 3: a flexible statistical power analysis program for the social, behavioral, and biomedical sciences. Behav Res Methods 2007;39(2):175-191. [doi: 10.3758/bf03193146] [Medline: 17695343]
42. Hennig C. Cluster-wise assessment of cluster stability. Comput Stat Data Anal 2007;52(1):258-271. [doi: 10.1016/j.csda.2006.11.025]
43. Orthopedic and trauma trainees’ attitude towards e-learning and distance learning: an European-wide survey. Mendeley Data.: Elsevier (Mendeley Data) URL: https://data.mendeley.com/datasets/xct57d995v/draft?a=ba28b364-da6c-4463-912f-d5f5c7a5a66b [accessed 2026-06-01]

XSL•FO
RenderX

44. Soffar A, Elkohail A, Elbanna M, Kassem HA, Murhekar S, Millat MS, et al. Virtual reality simulation for orthopedic surgical training: a narrative review of current evidence and educational impact. Cureus 2025;17(12):e99367. [doi: 10.7759/cureus.99367] [Medline: 41416020]
45. Clarke E. Virtual reality simulation-the future of orthopaedic training? A systematic review and narrative analysis. Adv Simul (Lond) 2021;6(1):2. [doi: 10.1186/s41077-020-00153-x] [Medline: 33441190]
46. Gamieldien H, Kruger N, Dey R. MB ChB fifth-year student response to e-learning in orthopaedic surgery during COVID-19. Afr J Health Prof Educ 2023:15-20. [doi: 10.7196/ajhpe.2023.v15i1.1629]
47. Bond M, Buntins K, Bedenlier S, Zawacki-Richter O, Kerres M. Mapping research in student engagement and educational technology in higher education: a systematic evidence map. Int J Educ Technol High Educ 2020;17(1):2. [doi: 10.1186/s41239-019-0176-8]
48. Elnoor OS, Hundal JS, Chahal US, Kansal SK. Assessment of postgraduate veterinary students' information literacy competencies and attitude towards e-learning. Indian J Econ Dev 2017;13(2):339. [doi: 10.5958/2322-0430.2017.00185.8]
49. Langegård U, Kiani K, Nielsen SJ, Svensson P. Nursing students' experiences of a pedagogical transition from campus learning to distance learning using digital tools. BMC Nurs 2021;20(1):23 [FREE Full text] [doi: 10.1186/s12912-021-00542-1] [Medline: 33468132]
50. Lera M, Taxtsoglou K, Iliadis C, Frantzana A, Kourkouta L. Nurses' attitudes toward lifelong learning via new technologies. Asian Pac Isl Nurs J 2020;5(2):89-102 [FREE Full text] [doi: 10.31372/20200502.1088] [Medline: 33043137]
51. Sommer GM, Broschewitz J, Huppert S, Sommer CG, Jahn N, Jansen-Winkeln B, et al. The role of virtual reality simulation in surgical training in the light of COVID-19 pandemic: Visual spatial ability as a predictor for improved surgical performance: a randomized trial. Medicine (Baltimore) 2021;100(50):e27844 [FREE Full text] [doi: 10.1097/MD.0000000000027844] [Medline: 34918632]
52. Ivanov V, Klygach A, Shterenberg S, Strelkov S, Levy J. Advances in agumented reality (AR) for medical simulation and training. 3C Tecnol Glosas Innov Apl Pyme 2020(Special Issue 5):303-312 [FREE Full text] [doi: 10.17993/3ctecno.2020.specialissue5.303-312]
53. Cianciolo AT, Williams RG, Klamen DL, Roberts NK. Biomedical knowledge, clinical cognition and diagnostic justification: a structural equation model. Med Educ 2013;47(3):309-316. [doi: 10.1111/medu.12096] [Medline: 23398017]
54. Teleki S, Zsidó AN, Lénárd L, Komócsi A, Kiss EC, Tiringer I. Role of received social support in the physical activity of coronary heart patients: the health action process approach. Appl Psychol Health Well Being 2022;14(1):44-63. [doi: 10.1111/aphw.12290] [Medline: 34166561]
55. Tick A. An extended TAM Model, for evaluating eLearning acceptance, digital learning and smart tool usage. Acta Polytech Hung 2019;16(9):213-233. [doi: 10.12700/APH.16.9.2019.9.12]
56. Cabero-Almenara J, Gutiérrez-Castillo JJ, Guillén-Gámez FD, Gaete-Bravo AF. Digital competence of higher education students as a predictor of academic success. Tech Know Learn 2022;28(2):683-702. [doi: 10.1007/s10758-022-09624-8]
57. Bala L, Kinross J, Martin G, Koizia LJ, Kooner AS, Shimshon GJ, et al. A remote access mixed reality teaching ward round. Clin Teach 2021;18(4):386-390. [doi: 10.1111/tct.13338] [Medline: 33786988]
58. Fidalgo P, Thormann J, Kulyk O, Lencastre JA. Students' perceptions on distance education: a multinational study. Int J Educ Technol High Educ 2020;17(1):18. [doi: 10.1186/s41239-020-00194-2]
59. Jan J, Alshare KA, Lane PL. Hofstede's cultural dimensions in technology acceptance models: a meta-analysis. Univ Access Inf Soc 2022;23(2):717-741. [doi: 10.1007/s10209-022-00930-7]
60. Jippes M, Majoor GD. Influence of national culture on the adoption of integrated medical curricula. Adv Health Sci Educ Theory Pract 2011;16(1):5-16 [FREE Full text] [doi: 10.1007/s10459-010-9236-5] [Medline: 20658354]
61. Damschroder LJ, Reardon CM, Widerquist MAO, Lowery J. The updated consolidated framework for implementation research based on user feedback. Implement Sci 2022;17(1):75 [FREE Full text] [doi: 10.1186/s13012-022-01245-0] [Medline: 36309746]
62. Glasgow RE, Harden SM, Gaglio B, Rabin B, Smith ML, Porter GC, et al. RE-AIM planning and evaluation framework: adapting to new science and practice with a 20-year review. Front Public Health 2019;7:64 [FREE Full text] [doi: 10.3389/fpubh.2019.00064] [Medline: 30984733]
63. Cabero-Almenara J, Llorente-Cejudo C, Palacios-Rodríguez A, Gallego-Pérez Ó. Degree of acceptance of virtual reality by health sciences students. Int J Environ Res Public Health 2023;20(8):5571 [FREE Full text] [doi: 10.3390/ijerph20085571] [Medline: 37107853]
64. Safi S, Thiessen T, Schmailzl KJ. Acceptance and resistance of new digital technologies in medicine: qualitative study. JMIR Res Protoc 2018;7(12):e11072 [FREE Full text] [doi: 10.2196/11072] [Medline: 30514693]
65. Goh J, Clapham M. Attitude to e-learning among newly qualified doctors. Clin Teach 2014;11(1):20-23. [doi: 10.1111/tct.12117] [Medline: 24405914]
66. Meling TR, Meling TR. The impact of surgical simulation on patient outcomes: a systematic review and meta-analysis. Neurosurg Rev 2021;44(2):843-854 [FREE Full text] [doi: 10.1007/s10143-020-01314-2] [Medline: 32399730]
67. Global Strategy on Digital Health 2020-2025. Geneva: World Health Organization; 2021.

XSL•FO
RenderX

Abbreviations

AR: augmented reality
CFI: comparative fit index
EFA: exploratory factor analysis
GDP: gross domestic product
GDPR: General Data Protection Regulation
MI: modification indices
OR: odds ratio
rbs: rank-biserial correlations
RE-AIM: research and reach, effectiveness, adoption, implementation, and maintenance
RMSEA: root mean square error of approximation
SEM: structural equation modeling
SRMR: standardized root mean square residual
TAM: technology acceptance model
TLI: Tucker-Lewis Index
UTAUT: unified theory of acceptance and use of technology
VR: virtual reality

Edited by A Stone; submitted 23.Jun.2025; peer-reviewed by Y Asada, R Nooripour; comments to author 04.Oct.2025; revised version received 08.Apr.2026; accepted 08.Apr.2026; published 10.Jul.2026.

Please cite as:
Schlegl AT, Ostojić M, Unterfrauner I, Ciolli G, Megaloikonomos PD, Igoumenou VG, Mercurio M, Stark T, Kordic D, Dietvorst M, Santos FL, Nyakas V, Tóth L, Komócsi A, Maróti P, Matuz A
e-Learning, Distance Education, and Virtual and Augmented Reality in Orthopedic Training: European Cross-Sectional Survey of Trainee Acceptance Guided by the Technology Acceptance Model and Unified Theory of Acceptance and Use of Technology
JMIR Med Educ 2026;12:e79418
URL: https://mededu.jmir.org/2026/1/e79418
doi:10.2196/79418
PMID:

©Adam Tibor Schlegl, Marko Ostojić, Ines Unterfrauner, Gianluca Ciolli, Panayiotis D Megaloikonomos, Vasilios G Igoumenou, Michele Mercurio, Thomas Stark, David Kordic, Martijn Dietvorst, Filipe Lima Santos, Viktória Nyakas, Luca Tóth, András Komócsi, Péter Maróti, András Matuz. Originally published in JMIR Medical Education (https://mededu.jmir.org), 10.Jul.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Education, is properly cited. The complete bibliographic information, a link to the original publication on https://mededu.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Live Podcasting as an Educational Intervention in Dentomaxillofacial Radiology: Controlled Cohort Study

Anna-Lena Hillebrecht[1], Dr Med Dent; Daniel Fritzsche[2], Med Dent; Thamar Voss[3], Dipl Psych, Prof Dr phil; Anne Kruse[2], Dr Med Dent, Priv Doz; Andreas Keßler[1,4], Dr Med Dent, Priv Doz, MSc; Kirstin Vach[5], Dipl Math, PhD; Markus Jörg Altenburger[6], Prof Dr Med Dent; Rainer Schmelzeisen[7], Prof Dr Med, Dr Med Dent; Wiebke Semper-Hogg[7], Dr Med Dent, Priv Doz

[1]Department of Prosthetic Dentistry, Center for Dental Medicine, Faculty of Medicine, Medical Center, University of Freiburg, Hugstetter Strasse 55, Freiburg, Germany

[2]Department of Operative Dentistry and Periodontology, Center for Dental Medicine, Faculty of Medicine, Medical Center, University of Freiburg, Freiburg, Germany

[3]Department of Educational Science, University of Freiburg, Freiburg, Germany

[4]Department of Conservative Dentistry and Periodontology, University Hospital, Ludwig-Maximilians University of Munich, Munich, Germany

[5]Institute of Medical Biometry and Statistics, Medical Center, University of Freiburg, Freiburg, Germany

[6]Department of Orthodontics, Center for Dental Medicine, Medical Center, Faculty of Medicine, University of Freiburg, Freiburg, Germany

[7]Department of Oral and Maxillofacial Surgery, Center for Dental Medicine, Faculty of Medicine, Medical Center, University of Freiburg, Freiburg, Germany

Corresponding Author:
Anna-Lena Hillebrecht, Dr Med Dent
Department of Prosthetic Dentistry, Center for Dental Medicine, Faculty of Medicine, Medical Center, University of Freiburg, Hugstetter Strasse 55, Freiburg, Germany

Abstract

Background: Podcasts are increasingly used in health professions education; however, most formats are asynchronous and noninteractive. Didactically grounded, synchronous implementations in dental curricula are scarce.

Objective: This study aims to design, implement, and evaluate a synchronous, case-based live podcast (LP) as a didactic teaching format in dentomaxillofacial radiology.

Methods: In a controlled cohort study with 2 third-year cohorts (N=41), the intervention group (IG; n=21, 51%) received weekly case-based LP sessions in addition to standard teaching, while the control group (CG; n=20, 49%) received standard teaching only. Acceptability was evaluated 6 months postcourse using the 27-item student evaluation questionnaire and open-text responses. Knowledge was assessed immediately after the course with a 21-item radiology knowledge test, and after 6 months, with a 15-item interdisciplinary clinical application test.

Results: The primary outcome was student-reported acceptability of the LP format. It was rated highly by students in the Student Evaluation Questionnaire (mean out of 10: structure 9.76, interactivity 9.62, interdisciplinary relevance 9.55). Qualitative feedback was assessed highlighting motivation, authenticity, and discussion quality. In the radiology knowledge test, no group differences were observed (IG: n=21, 51% vs CG: n=20, 49%; *P*=.37). In the interdisciplinary clinical application test, the IG outperformed the CG in restorative dentistry (median 5, IQR 4-5 vs median 4, IQR 3-5; *P*=.02; *r*=0.38) and in item-level analysis (15/21, 71% vs 40%; *P*=.04; ϕ=0.64).

Conclusions: The LP format represents a feasible, scalable, and low-threshold approach to fostering clinical reasoning in dental curricula, particularly at the transition to clinical training. While radiology-specific theoretical competencies did not differ between the groups, students consistently rated the LP as more engaging and motivating compared to standard lectures.

(JMIR Med Educ 2026;12:e77980) doi:10.2196/77980

KEYWORDS

case-based learning; clinical reasoning; dental education; digital learning; health professions education; interdisciplinary teaching; podcast; student engagement; synchronous learning

XSL•FO
RenderX

Introduction

Overview

In the landscape of health professions education, digital audio formats, particularly podcasts, have gained substantial momentum [1-5]. Their flexibility, accessibility, and conversational tone have made them an increasingly attractive medium for knowledge dissemination and learner engagement [4,6]. While podcasts are now widely accepted as a complementary teaching method in undergraduate medical education, their use remains predominantly asynchronous and prerecorded, with limited interactivity and integration into curricular strategies [1]. Despite their potential, podcasts have rarely been explored as intentionally designed didactic tools grounded in contemporary educational theory. While the demand for flexible and digital teaching formats in dental education continues to rise, there is a striking lack of pedagogically grounded concepts specifically addressing the use of podcasts within the dental curriculum [5,7]. To better understand students' needs in modern education, a preliminary needs assessment was conducted among undergraduate dental students at the University of Freiburg. The results indicated that while students appreciated classical media such as lectures and textbooks for foundational learning, they strongly favored digital formats for more advanced clinical content. Podcasts, although not yet widely used for learning, were perceived as a promising format for flexible, self-directed studies. These findings highlighted a clear interest in diverse, interactive, and clinically relevant digital learning formats [8].

Rationale and Objectives

Recent initiatives such as the student-led technology committee highlight the growing recognition of students as active stakeholders in shaping digital education, demonstrating how learner involvement can drive innovation and relevance in curricular technologies [9]. At the same time, in dentistry, clinical reasoning (CR) is gaining increasing importance, particularly as dental treatment needs are rising among vulnerable populations, including older adults and patients with complex medical conditions [10,11]. CR is particularly evident in dentomaxillofacial radiology (DMFR), where diagnostic decisions based on radiographic interpretation significantly influence treatment planning. CR in this context includes not only the technical analysis of imaging data but also interdisciplinary judgment, risk assessment, and patient-centered decision-making. CR refers to the cognitive process by which health care professionals collect, interpret, and synthesize patient information to make informed diagnostic and therapeutic decisions. CR remains a central yet pedagogically challenging competence in medical and dental education. Developing CR requires authentic, situated learning experiences that bridge disciplines and promote critical discussion, an area in which traditional lecture-based formats often fall short.

In recent years, traditional lecture-based formats have increasingly lost their appeal among students, raising the question of whether innovative and more interactive approaches can reengage learners and potentially revitalize attendance in academic teaching. This study introduces and evaluates the live podcast (LP) as an innovative digital teaching format that combines the accessibility of podcasting with the didactic strengths of interactive, case-based learning. To our knowledge, this is the first systematic implementation of an LP format within the dental curriculum. The LP was not conceived as a mere content delivery tool but as a didactically grounded educational intervention designed to foster clinical reasoning, interdisciplinary thinking, and student engagement through authentic expert dialogue. The objective was to introduce a didactically designed LP format within undergraduate DMFR and evaluate its feasibility and student-reported acceptability and to explore potential effects on theoretical and interdisciplinary knowledge compared with standard teaching.

The LP format was hypothesized to be feasible, well accepted by students, and to improve knowledge scores in domains requiring case-based reasoning compared with standard instruction.

Methods

This study primarily presents and evaluates the LP as a novel teaching format in DMFR education, focusing on its instructional design, implementation, and initial effectiveness.

Curricular Context

Undergraduate dental education in Germany is structured as a 5-year program, consisting of a 2-year preclinical phase (basic sciences such as anatomy, physiology, biochemistry, and histology, along with introductory dental subjects), followed by a 3-year clinical phase. The curriculum is nationally regulated and guided by the National Competence-Based Learning Objectives Catalogue in Dentistry. DMFR is introduced at the beginning of the clinical training (fifth semester) in close connection with basic dental instruction and is subsequently applied throughout the clinical phase. The course includes radiographic techniques, diagnostic interpretation, radiation protection, and ethical considerations, integrating theoretical instruction with practical training (intraoral radiographs, including periapical radiographs, bitewing radiographs, and occlusal radiographs; panoramic radiographs such as orthopantomogram; and cone-beam computed tomography). The LP sessions were designed based on National Competence-Based Learning Objectives Catalogue in Dentistry objectives, particularly Z 21 (clinical diagnostics, radiographic imaging, and radiation protection) and Z 5 (dentist as medical expert), thereby ensuring consistency with national competency standards.

Didactic and Conceptual Design

The LP was developed as a synchronous, case-based learning format and implemented as a mandatory component of the DMFR course for third-year undergraduate dental students at the University of Freiburg. Each weekly 45-minute session followed a structured design: anonymized real-life cases were discussed in moderated dialogue between a dental radiologist and a specialist from a complementary dental or medical discipline (eg, pediatric dentistry, prosthodontics, oral surgery, and radiology). Discussions focused on radiographic diagnosis, differential diagnoses, and interdisciplinary treatment planning.

XSL•FO
RenderX

All episodes were designed with explicit learning goals that emphasized diagnostic competence and clinical reasoning, in line with the interdisciplinary nature of DMFR (Tables 1 and 2).

Table . Curriculum framework for the live podcast: episodes, topics, disciplines, imaging modalities (periapical radiographs [PA], bitewing radiographs [BW], occlusal radiographs [OC], panoramic radiographs such as orthopantomogram [OPG], and cone-beam computed tomography [CBCT]), and radiographic diagnoses.

Episode	Topic	Discipline	Imaging modalities in this episode	Radiographic diagnoses
1	Pediatric dentistry	Pediatric dentistry	PA; BW; OPG	Approximal caries and apical periodontitis
2	Endodontology	Endodontics	PA; CBCT for complex cases	Irreversible pulpitis and apical periodontitis
3	Caries diagnostics	Restorative dentistry	BW; PA	Bitewing caries; secondary caries versus artifacts
4	General prosthodontic planning	Prosthodontics	OPG; PA	Tooth loss, apical periodontitis, bone loss, and TMJ[a] findings
5	Special care dentistry	Special care dentistry	OPG; PA	Tooth loss, caries, bone loss, and radiation protection issues
6	General radiology	Oral and maxillofacial radiology	PA; BW; OPG; CBCT; MRI[b]	Modality characteristics (intraoral, BW, CBCT, and MRI)
7	History of dentistry	History of medicine or dentistry	Historical imaging methods	Early radiographs and OPG
8	Artificial intelligence in dentistry	Informatics; oral radiology	PA; OPG; CBCT	AI[c]-assisted detection of caries and periodontal bone loss
9	Orthodontics	Orthodontics	OPG; lateral cephalogram; PA; OC	Tooth development; retained teeth
10	Periodontology	Periodontology	PA; BW; OPG	Horizontal and vertical bone loss
11	Surgical removal of wisdom teeth	Oral and maxillofacial surgery	OPG; CBCT for complex cases	Impactions, cysts, and nerve injury risk

[a]TMJ: temporomandibular joint.

[b]MRI: magnetic resonance imaging.

[c]AI: artificial intelligence.

XSL•FO
RenderX

Table . Curriculum framework for the live podcast: episodes, topics, covered domains, interdisciplinary aspects, and learning goals.

Episode	Topic	Covered domains	Interdisciplinary aspects	Learning goals
1	Pediatric dentistry	Imaging and diagnostics in primary and mixed dentition, trauma management, prevention, and developmental diagnostics	Pediatrics, and oral and maxillofacial surgery (fracture management)	Apply pediatric-specific diagnostic strategies, and select imaging in pediatric trauma and explain informed consent in minors
2	Endodontology	Pain diagnostics, pulp vitality assessment, and decision-making in endodontic emergencies	Oral and maxillofacial surgery and restorative dentistry	Strengthen case-based reasoning and interpret findings in irreversible pulpitis and differentiate periapical pathologies
3	Caries diagnostics	Primary and secondary caries, diagnostics, and prevention	Prevention, pediatrics, prosthodontics, and endodontics	Enhance diagnostic accuracy and apply imaging in caries diagnostics and distinguish true lesions from imaging artifacts
4	General prosthodontic planning	Prosthetic concepts, treatment planning, and radiographic considerations	Geriatrics, orthodontics, oral and maxillofacial surgery, and restorative dentistry	Integrate diagnostics into treatment planning, and use OPG[a] in planning and develop interdisciplinary prosthodontic strategies
5	Special care dentistry	Prosthetic care for people with disabilities, barriers to care, and ethics	Medicine, nursing, and ethics (focus on vulnerable and medically complex patients)	Link ethics with radiology, and apply radiation protection in vulnerable groups and critically reflect on imaging indications
6	General radiology	Physics; ALARA[b] principle; imaging modalities	Cross-disciplinary relevance of imaging decisions	Build safety and modality competence, and apply ALARA and differentiate indications for imaging modalities
7	History of dentistry	Evolution of diagnostics and imaging	History of medicine	Promote scientific reflection, and describe historical developments in dental diagnostics and radiology
8	Artificial intelligence in dentistry	Digital diagnostics and decision support	Ethics and data science	Foster critical, future-oriented thinking; evaluate benefits and limitations of AI[c]-assisted diagnostics
9	Orthodontics	Malocclusions and growth diagnostics	Pediatrics and prosthodontics	Relate radiology to growth analysis, and explain OPG or CBCT[d] and model analysis in orthodontics and assess impacted teeth
10	Periodontology	Periodontal diagnostics, disease progression, and treatment planning	Prosthodontics and implantology	Strengthen diagnostic integration, and identify periodontal bone loss radiographically and link findings to implant and prosthodontic planning
11	Surgical removal of wisdom teeth	Indications, surgical techniques, and preoperative planning	Radiology, anesthesiology, and prosthodontics	Link diagnostics with surgery, and select OPG versus CBCT appropriately and assess surgical risks radiographically

[a]OPG: orthopantomogram.

[b]ALARA: As Low As Reasonably Achievable.

[c]AI: artificial intelligence.

[d]CBCT: cone-beam computed tomography.

XSL•FO
RenderX

Technical Setup

Sessions were held in a lecture hall with professional audiovisual infrastructure. At least 2 microphones transmitted the voices of the podcasters via the hall sound system, and an additional microphone was available for student interaction. A projector displayed radiographs and clinical photographs. Each session was recorded live with sound levels adjusted by a technician. Recordings were postprocessed in Adobe Audition (version 23.11; Adobe Inc) and uploaded via Panopto (version 14.0; Panopto Inc) to the official university learning platform, ensuring high audio quality and asynchronous access for enrolled students (Multimedia Appendix 1).

Participants and Study Design

All third-year dental students enrolled in the mandatory DMFR course at the University of Freiburg were eligible. In total, 41 students participated in 2 consecutive cohorts: a control group (CG; n=20, 49%; with n=12, 60% women and n=8, 40% men) that attended traditional lectures covering DMFR principles, radiographic techniques, and clinical relevance, and an intervention group (IG; n=21, 51%; with n=11, 52% women and n=10, 48% men) that attended the LP sessions. No exclusion criteria were applied because all students were required to complete the course as part of the curriculum. The cohorts received comparable curricular content from the same teaching staff, ensuring baseline comparability of exposure. No pretest was conducted because both groups had identical prior curricular exposure. All students completed the subsequent knowledge assessments. Item-level correctness and point-based domain scores were analyzed across restorative dentistry, oral surgery, and orthodontics. The 3 assessment instruments used in this study were the interdisciplinary clinical application test (ICAT; Multimedia Appendix 2), the radiology knowledge test (RKT; Multimedia Appendix 3), and the Student Evaluation Questionnaire (SEQ; Multimedia Appendix 4).

Assessment of Student-Reported Acceptability

To determine the primary outcome, student-reported acceptability was evaluated 6 months after the LP using an anonymous questionnaire with 27 rating items (SEQ), on a 10-point unipolar scale (1=I strongly disagree, 10=I strongly agree). The questionnaire covered organizational aspects, atmosphere, interactivity, memorability, long-term benefit, integration of theory and practice, reflective engagement, social learning, and overall evaluation. Selected domains were stratified by gender. An optional open-text field captured qualitative impressions that were analyzed descriptively.

Assessment of Learning Outcomes

Knowledge served as the secondary outcome and was evaluated at 2 time points to capture both immediate and longer-term learning effects.

Immediately after the course: IG and CG completed a DMFR radiology-specific competence test (RKT), with 21 single-choice items (F1-F21) covering radiation protection, diagnostic imaging, and interpretation.

Six months after the course, IG and CG participated in an interdisciplinary multiple-choice test (ICAT) with 15 case-based items (5 items each for restorative dentistry, oral surgery, and orthodontics), addressing radiographic interpretation, interdisciplinary treatment planning, and clinical application of radiological findings.

Statistical Analysis

No formal sample size calculation was performed. As a feasibility study embedded in routine teaching, the available cohort size (N=41) defined the sample. This aligns with the exploratory aim to evaluate acceptability and didactic value rather than to demonstrate definitive knowledge effects. Descriptive statistics (mean, SD, median, and IQR) were calculated for all outcomes. Chi-square tests were applied for item-specific correctness, and Wilcoxon rank-sum tests were used for group comparisons of point-based scores and gender differences. The Fisher exact test was applied in the case of single-choice items. Effect sizes (r) were reported where appropriate. Given the exploratory design, no adjustment for multiple testing was applied. Statistical analyses were performed in Stata (version 19; StataCorp) with a significance level of α=.05.

Ethical Considerations

The study involved human participants and was conducted in accordance with institutional and national research regulations and the Declaration of Helsinki. The study was reviewed informally by the ethics committee of the University of Freiburg (reference number: 25-1479). As it served internal educational quality assurance without collection of personal or health-related data, formal approval was not required. Informed consent was obtained from all participants during their voluntary enrollment in the course after they had received detailed information about the study purpose, procedures, data use, and their right to withdraw at any time without negative consequences. Data were collected and analyzed anonymously and treated confidentially in accordance with data protection regulations. Participation was voluntary and no financial or other compensation was provided.

Results

Overview

A total of 41 third-year dental students participated in the study (IG: n=21, 51%; CG: n=20, 49%). All students completed the RKT immediately after the course, and all participated in the ICAT interdisciplinary follow-up test 6 months later. Student-reported acceptability of the LP format was assessed 6 months postcourse; 21 students (100% of the IG) completed the SEQ.

Primary Outcome: Student-Reported Acceptability

To determine student-perceived acceptability of the LP format, 21 students completed the SEQ 6 months after the course. Overall, the LP received consistently high ratings across all evaluated dimensions, including atmosphere, organization, perceived usefulness, and interdisciplinary relevance. The mean scores were above 9/10 in nearly all categories, indicating very high acceptance and perceived educational value. No meaningful gender−related differences were observed (Table 3).

XSL•FO
RenderX

Table . Student Evaluation Questionnaire (SEQ): evaluation of the live podcast in different categories stratified by gender.

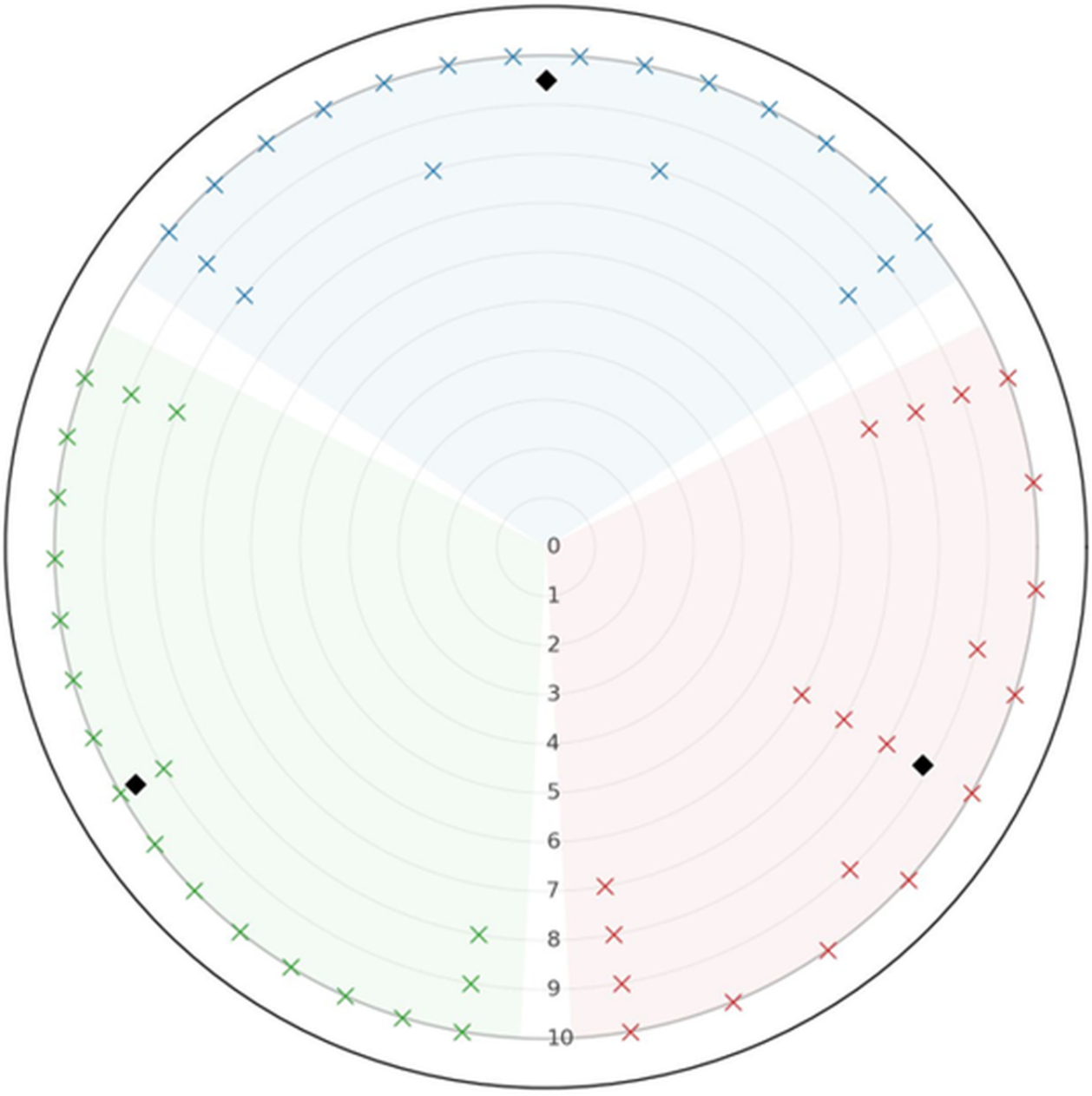
Category	Women (n=11; score: 1-10), mean (SD)	Men (n=10; score: 1-10), mean (SD)
Atmosphere	9.91 (0.30)	9.40 (0.97)
Organization	9.82 (0.40)	9.70 (0.48)
Useful teaching method	9.73 (0.65)	9.56 (0.73)
Scope of knowledge	9.18 (1.32)	9.56 (1.01)
Interdisciplinary connections	9.45 (0.82)	9.67 (0.50)
Personal perception	6.64 (1.21)	7.67 (1.94)
Addition to the lecture	8.64 (1.75)	8.78 (1.39)
Interest in topics	9.91 (0.30)	9.30 (0.95)

No meaningful gender-related differences were identified for the total evaluation score (P=0.18) or for any subscale (P=0.32–0.79).

Student ratings for the 3 selected evaluation items as illustrative examples are shown in Figure 1. Overall, the majority of participants provided high ratings, indicating strong agreement across all 3 items. Specifically, for “learning from the experiences of others,” 20% (n=4) of participants rated 8, 10% (n=2) rated 9, and 70% (n=14) rated 10. For “encouraged reflection,” 5% (n=1) rated 6, 15% (n=3) rated 7, 15% (n=3) rated 8, 20% (n=4) rated 9, and 45% (n=9) rated 10. For “useful teaching method,” 10% (n=2) rated 8, 15% (n=3) rated 9, and 75% (n=15) rated 10. All other rating categories received no responses.

XSL•FO
RenderX

Figure 1. Distribution of participant ratings for the Student Evaluation Questionnaire (SEQ) items. The radial plot visualizes the distribution of all ratings (1-10) for the items "learning from the experiences of others" (blue), "encouraged reflection" (red), and "useful teaching method" (green; each n=20). Each cross represents 1 individual response, while black diamonds (◆) indicate mean values. No ratings were given for scores 1-5. For "learning from the experiences of others," 20% (n=4) of participants rated 8, 10% (n=2) rated 9, and 70% (n=14) rated 10. For "encouraged reflection," 5% (n=1) rated 6, 15% (n=3) rated 7, 15% (n=3) rated 8, 20% (n=4) rated 9, and 45% (n=9) rated 10. For "useful teaching method," 10% (n=2) rated 8, 15% (n=3) rated 9, and 75% (n=15) rated 10. All other rating categories received no responses. A unipolar rating scale from 1 (I strongly disagree) to 10 (I strongly agree) was used, with higher values indicating a more favorable evaluation.



Qualitative Findings

In addition to the quantitative results, students provided written feedback at the end of the course. Selected anonymized comments (translated from German) illustrate how the LP was perceived:

> *The podcasts were scheduled at a good time before lunch, which encouraged further discussion of the topics in the cafeteria afterwards. The content was very stimulating.*

> *In some sessions, time was even too short, as the discussions were highly engaging. A second microphone for students would sometimes have been helpful.*

> *I sincerely appreciate the effort that went into these sessions. It was a great experience to learn the material in such an interesting way through contributions from many different lecturers.*

> *Through the podcast I realized how much dentistry has to offer. The podcast, with its many insights into different specialties, gave me renewed motivation and interest in my studies. After four semesters of rote*

XSL•FO
RenderX

> *learning, there could not have been a better introduction to the clinical phase.*
>
> *I really enjoyed attending the podcast lectures. They were varied and never felt like dry teaching, yet I learned a great deal. The relaxed atmosphere, such as sitting on the sofa, created an environment that did not feel like studying, although one gained a lot of knowledge.*

These student voices indicate that the LP sessions were experienced as innovative, motivating, and discussion-stimulating, creating a positive learning atmosphere at the transition from preclinical to clinical studies.

Secondary Outcomes: Knowledge Acquisition

In the ICAT, item-level analysis revealed a significantly higher proportion of correct responses in the IG compared with the CG in restorative dentistry (P=.04; ϕ=0.64). No statistically significant differences were observed in oral surgery (P=.85; ϕ=0.005) or orthodontics (P=.14; ϕ=0.34).

Across all ICAT domains combined, no statistically significant difference in overall performance was observed between groups (P=.37; ϕ=0.49). A distribution-based analysis showed that in the IG, 7 students answered 3 items correctly, 13 students answered 2 items correctly, and 1 student answered 1 item correctly, with no student scoring 0 correct answers. In contrast, in the CG, 4 students answered 3 items correctly, 12 students answered 2 items correctly, 2 students answered 1 item correctly, and 2 students answered no items correctly. This distribution indicates a trend toward higher overall performance in the IG, although the difference did not reach statistical significance (Table 4).

Table . Interdisciplinary clinical application test (ICAT): point-based domain scores by group (Wilcoxon rank-sum test; α=.05)[a].

Domain	IG[b] (points), median (IQR)	IG (points), mean (SD)	CG[c] (points), median (IQR)	CG (points), mean (SD)	Effect size *r*
Restorative dentistry	5 (4-5)	4.62 (0.67)	4 (3-5)	3.75 (1.25)	0.38
Oral surgery	5 (3-5)	3.95 (1.46)	5 (3.5-5)	4.10 (1.37)	0.05
Orthodontics	5 (5-5)	5.00 (0.00)	5 (5-5)	4.60 (1.23)	0.23

[a]Median, IQR, mean, and SD values are presented for the intervention and control groups. The effect size (r) is shown for each domain. Only the comparison in restorative dentistry reached statistical significance (P=.02). No significant differences were found for oral surgery or orthodontics.

[b]IG: intervention group.

[c]CG: control group.

No statistically significant differences were observed in the RKT between IG and CG for the overall score or any subdomain (all P>.05). The total RKT score was numerically higher in the IG (median 14, IQR 13-15) than in the CG (median 13, IQR 11-14), but this difference did not reach statistical significance (P=.17; r=0.22), indicating comparable short-term acquisition of radiology-specific theoretical competences in both groups.

In the ICAT, analysis of item-level correctness revealed a significantly higher performance in the IG (75/105, 71%) than in the CG (40/100, 40%) in restorative dentistry (P=.04; ϕ=0.64). No significant differences were observed in oral surgery (P=.85; ϕ=0.005) or orthodontics (P=.14; ϕ=0.34). Overall correctness across all domains was slightly higher in the IG (161/315, 51%) compared with the CG (147/300, 49%), though this difference was not statistically significant (P=.37; ϕ=0.49; Table 4).

Discussion

Principal Findings

This study demonstrates the development and feasibility of a novel teaching format in undergraduate dental education: the live podcast. To our knowledge, this is the first systematically designed and evaluated synchronous podcast format embedded within a dental curriculum. Unlike traditional podcasts, the LP was conceived as a didactically grounded intervention that combines digital accessibility with case-based, interactive learning. Our results provide initial evidence that LPs are feasible and highly accepted. The consistently high SEQ ratings in structure, clarity, and interactivity suggest that LPs foster an engaging learning environment, particularly at the transition from preclinical to clinical training.

Students in the intervention group performed significantly better in the ICAT in the restorative dentistry domain, which indicates that authentic case discussions supported by expert dialogue may enhance diagnostic reasoning. Although no significant differences were observed across all domains in ICAT, descriptive trends point to positive effects on interdisciplinary understanding. Beyond knowledge outcomes, very high acceptance ratings and open-ended comments show that students perceived the LP as motivating, discussion-enhancing, and well aligned with their learning preferences. In line with contemporary student-voice literature, the LP addressed the cognitive dimension of clinical reasoning and also fostered engagement, reflexivity, and ownership through its interactive and dialogic structure [12]. Goh and Lim [13] argue that dental learning is increasingly conceptualized as situated and relational. The LP exemplifies this by embedding reasoning in authentic, interdisciplinary practice.

The RKT did not reveal significant differences between the 2 groups. For theoretical radiology competencies, both formats appear equally effective. Taken together with the very high acceptance ratings in the SEQ, this suggests that the added value of the LP lies less in measurable short-term knowledge gains and more in providing a motivating, engaging, and student-centered learning environment with an interdisciplinary focus [1].

XSL•FO
RenderX

Evidence from this study suggests that synchronous, audio-based teaching formats such as LPs can act as effective didactic tools in complex, interdisciplinary areas of dental education, including DMFR. By fostering active engagement, real-time dialogue, and case-based reflection, LPs extend the potential of both traditional lectures and asynchronous e-learning. The positive reception aligns with prior literature on the educational benefits of podcasts. Kaplan et al [14] emphasize brevity, narrative structure, and authentic expert dialogue as advantages aligned with adult learning principles and conducive to learner motivation and retention. Consistent with these findings, students in our study reported in the SEQ high levels of didactic clarity, structure, and perceived longer-term benefit, dimensions repeatedly highlighted in podcast-based education [15]. The conversational and interdisciplinary character of the LP mirrors the informal, yet focused tone described by Kaplan et al [14] as central to engagement and professional connectedness. Narrative and storytelling, which support sense-making, empathy, and professional identity formation [16], were intentionally embedded through authentic case discussions. Qualitative feedback indicated that these elements were experienced as engaging and professionally meaningful, resonating with sociocultural learning theory [14,17]. In line with work that identifies podcasts as vehicles for democratizing discourse and fostering reflection [18], live podcasting may strengthen interdisciplinary awareness, reflective reasoning, and inclusive dialogue in dental education.

Core disciplinary knowledge was tested in the RKT immediately after the course to capture short-term acquisition of radiology-specific competences in line with constructive alignment principles. Interdisciplinary case−based knowledge in the ICAT, by contrast, was assessed after 6 months because durable learning and knowledge transfer typically manifest after a consolidation phase. By separating immediate mastery from delayed application, the design aimed to minimize test contamination and to capture distinct learning outcomes targeted by the LP. The combination of synchronous interaction and asynchronous access corresponds to evolving learner expectations for hybrid, customizable formats [19]. The ability to revisit discussions, catch up on missed sessions, and selectively review content increases long-term educational use and accommodates diverse learning needs. These findings are consistent with prior institutional needs assessments showing strong preferences for flexible, multimodal tools [8].

Didactic Implications for Teaching Practice

LPs are a cost-effective, low-threshold format that integrates readily into blended learning environments. The synchronous, case-based design suits the development of clinical reasoning and interdisciplinary thinking. Opportunities for learner questions and moderated dialogue promote interaction, reflection, and active engagement. High ratings in structure, relevance, and perceived benefit are known drivers of acceptance in higher education [20]. Embedding LPs as curriculum-aligned supplements, supported by preparatory and follow-up materials, can deepen learning and support knowledge retention. LPs are intended to complement, not replace, lectures so that students benefit from interactive digital formats while still using the lecture hall for consolidation and further integration of knowledge. Because LPs rely on existing audiovisual infrastructures and can be easily streamed, recorded, and integrated into learning management systems, this teaching format is scalable and transferable to other health professions programs.

Limitations

The sample was small and restricted to a single center, which limits generalizability. Group allocation used consecutive cohorts rather than randomization, which may introduce selection bias. No pretest of baseline knowledge was conducted, and individual prior knowledge could not be controlled. The ICAT comprised 15 multiple-choice items and assessed midterm retention only, so effects on long-term learning and clinical performance remain unclear. The evaluation of student-reported acceptability in the SEQ relied on self-report, which may be affected by social desirability.

Future Directions

Future research should use randomized, adequately powered, multicenter designs with baseline testing and correction for multiple comparisons. A longer follow-up is needed to examine durability. Performance-based outcomes such as objective structured clinical examinations and workplace-based assessments should be included to evaluate transfer to clinical decision-making and patient care. Studies could also compare LP design variants, for example, the degree of interactivity, scaffolding materials, and image integration.

Conclusions

The LP was feasible and well accepted in the undergraduate DMFR course. It was associated with a significant knowledge gain in restorative dentistry and yielded results comparable to lectures in radiology-specific competences, while students rated it as more engaging and motivating. LPs, therefore, represent a scalable complement to standard instruction with potential to strengthen clinical reasoning and student-centered learning.

Acknowledgments

The authors thank Jonas Wagner, Sebastian Ollinger, Finn Weisser, Cumhur Yörük, and Robin Gommel for advice and technical support, the interdisciplinary experts who contributed as podcast guests, and the students for their active participation in the live podcast and the assessments.

XSL•FO
RenderX

Funding

This work was funded entirely by internal resources of the University of Freiburg with no external financial support, and we acknowledge the support by the Open Access Publication Fund of the University of Freiburg.

Data Availability

All relevant data are included in this article.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Live podcast episode 2—endodontology (case-based teaching material) PDF of the case-based teaching slides. The material includes learning objectives, a trauma-related case (teeth 21/22), sequential clinical and radiographic images (periapical radiographs [PA] and cone-beam computed tomography [CBCT]), key reasoning steps for pain diagnostics and pulp vitality assessment, take-home messages, and an optional self-test multiple-choice question (MCQ). All radiographs and clinical photographs are shown in anonymized form with separate patient consent for publication.
[PDF File, 624 KB - mededu_v12i1e77980_app1.pdf]

Multimedia Appendix 2
Interdisciplinary clinical application test (ICAT).
[PDF File, 82 KB - mededu_v12i1e77980_app2.pdf]

Multimedia Appendix 3
Radiology knowledge test (RKT).
[PDF File, 441 KB - mededu_v12i1e77980_app3.pdf]

Multimedia Appendix 4
Student Evaluation Questionnaire (SEQ).
[PDF File, 139 KB - mededu_v12i1e77980_app4.pdf]

References

1. Kelly JM, Perseghin A, Dow AW, Trivedi SP, Rodman A, Berk J. Learning through listening: a scoping review of podcast use in medical education. Acad Med 2022 Jul 1;97(7):1079-1085. [doi: 10.1097/ACM.0000000000004565] [Medline: 34935729]
2. Wang LKP, Paidisetty PS, Sathyanarayanan S. Using podcasts to study efficiently and increase exposure to medical specialties. Acad Med 2023 Feb 1;98(2):157. [doi: 10.1097/ACM.0000000000005075] [Medline: 36696295]
3. Cho D, Cosimini M, Espinoza J. Podcasting in medical education: a review of the literature. Korean J Med Educ 2017 Dec;29(4):229-239. [doi: 10.3946/kjme.2017.69] [Medline: 29207454]
4. Cook DA, Triola MM. What is the role of e-learning? Looking past the hype. Med Educ 2014 Sep;48(9):930-937. [doi: 10.1111/medu.12484] [Medline: 25113119]
5. Condo M, McKay J, Dassinger J, Nolan B. Podcasting as an innovative tool for National Physical Therapy Examination preparation. Med Sci Educ 2023 Jun;33(3):631-633. [doi: 10.1007/s40670-023-01792-1] [Medline: 37501798]
6. Surapaneni KM. Implications of using podcasts as handy tools for continuing professional development in health professions. Med Sci Educ 2023 Aug;33(4):1021. [doi: 10.1007/s40670-023-01830-y] [Medline: 37546187]
7. Apel Z, Fagundes NCF, Sharmin N, et al. Social media in oral health education: a scoping review. Eur J Dent Educ 2025 Feb;29(1):50-63. [doi: 10.1111/eje.13053] [Medline: 39462438]
8. Kruse AB, Isailov-Schöchlin M, Giesler M, Ratka-Krüger P. Which digital learning strategies do undergraduate dentistry students favor? A questionnaire survey at a German university. GMS J Med Educ 2023;40(4):Doc49. [doi: 10.3205/zma001631] [Medline: 37560041]
9. Shenson JA, Adams RC, Ahmed ST, Spickard A. Formation of a new entity to support effective use of technology in medical education: the student technology committee. JMIR Med Educ 2015 Sep 17;1(2):e9. [doi: 10.2196/mededu.4676] [Medline: 27731843]
10. Annamma LM, Varma SR, Abuttayem H, et al. Current challenges in dental education—a scoping review. BMC Med Educ 2024 Dec 23;24(1):1523. [doi: 10.1186/s12909-024-06545-1] [Medline: 39716191]
11. Prosser GM, Louca C, Radford DR. Potential educational and workforce strategies to meet the oral health challenges of an increasingly older population: a qualitative study. BDJ Open 2022 Mar 10;8(1):6. [doi: 10.1038/s41405-022-00098-5] [Medline: 35273157]

XSL•FO
RenderX

12. Subramanian J, Anderson VR, Morgaine KC, Thomson WM. The importance of “student voice” in dental education. Eur J Dent Educ 2013 Feb;17(1):e136-e141. [doi: 10.1111/j.1600-0579.2012.00773.x] [Medline: 23279401]
13. Goh AYS, Lim AD. Reconceptualising dentist learning—a commentary. Eur J Dent Educ 2022 Aug;26(3):639-642. [doi: 10.1111/eje.12741] [Medline: 34954843]
14. Kaplan H, Verma D, Sargsyan Z. What traditional lectures can learn from podcasts. J Grad Med Educ 2020 Jun;12(3):250-253. [doi: 10.4300/JGME-D-19-00619.1] [Medline: 32595838]
15. Back DA, von Malotky J, Sostmann K, Hube R, Peters H, Hoff E. Superior gain in knowledge by podcasts versus text-based learning in teaching orthopedics: a randomized controlled trial. J Surg Educ 2017;74(1):154-160. [doi: 10.1016/j.jsurg.2016.07.008] [Medline: 27651055]
16. Dahlstrom MF. Using narratives and storytelling to communicate science with nonexpert audiences. Proc Natl Acad Sci U S A 2014 Sep 16;111(Suppl 4):13614-13620. [doi: 10.1073/pnas.1320645111] [Medline: 25225368]
17. Kahlke R, Bates J, Nimmon L. When I say … sociocultural learning theory. Med Educ 2019 Feb;53(2):117-118. [doi: 10.1111/medu.13626] [Medline: 29943464]
18. Kamal S, Trivedi SP, Essien UR, Nematollahi S. Podcasting: a medium for amplifying racial justice discourse, reflection, and representation within graduate medical education. J Grad Med Educ 2021 Feb;13(1):29-32. [doi: 10.4300/JGME-D-20-00990.1] [Medline: 33680297]
19. Riddell J, Robins L, Brown A, Sherbino J, Lin M, Ilgen JS. Independent and interwoven: a qualitative exploration of residents' experiences with educational podcasts. Acad Med 2020 Jan;95(1):89-96. [doi: 10.1097/ACM.0000000000002984] [Medline: 31517682]
20. Venkatesh V, Morris MG, Davis GB, Davis FD. User acceptance of information technology: toward a unified view. MIS Q 2003 Sep 1;27(3):425-478. [doi: 10.2307/30036540]

Abbreviations

CG: control group
CR: clinical reasoning
DMFR: dentomaxillofacial radiology
ICAT: interdisciplinary clinical application test
IG: intervention group
LP: live podcast
RKT: radiology knowledge test
SEQ: Student Evaluation Questionnaire

Edited by P Kanzow; submitted 29.May.2025; peer-reviewed by AV Mahuli, DP Valenzuela; accepted 19.Nov.2025; published 05.Jan.2026.

Please cite as:
Hillebrecht AL, Fritzsche D, Voss T, Kruse A, Keßler A, Vach K, Altenburger MJ, Schmelzeisen R, Semper-Hogg W
Live Podcasting as an Educational Intervention in Dentomaxillofacial Radiology: Controlled Cohort Study
JMIR Med Educ 2026;12:e77980
URL: https://mededu.jmir.org/2026/1/e77980
doi:10.2196/77980

© Anna-Lena Hillebrecht, Daniel Fritzsche, Thamar Voss, Anne Kruse, Andreas Keßler, Kirstin Vach, Markus Jörg Altenburger, Rainer Schmelzeisen, Wiebke Semper-Hogg. Originally published in JMIR Medical Education (https://mededu.jmir.org), 5.Jan.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Education, is properly cited. The complete bibliographic information, a link to the original publication on https://mededu.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Toward Human-Centered Pedagogy in the Age of Generative AI Among Higher Education Educators: Qualitative Phenomenological Study

Zuheir Khlaif, PhD; Bilal Hamamra, PhD

Learning and Education Leadership, United Arab Emirates University, Al Ain, Abu Dhabi, United Arab Emirates

Corresponding Author:
Zuheir Khlaif, PhD
Learning and Education Leadership, United Arab Emirates University, Al Ain, Abu Dhabi, United Arab Emirates

Abstract

Background: Generative AI (GenAI) expands possibilities for content generation, feedback, personalization, and instructional support in higher education. Its growing use also challenges established assumptions about teaching, assessment, professional expertise, academic integrity, and educational responsibility. Although research has examined technology adoption and AI competencies, less is known about how educators experience and interpret the pedagogical transformation associated with GenAI in everyday teaching.

Objective: This study explored how higher education educators experience GenAI-mediated pedagogical transformation, focusing on pedagogical practices, professional roles and expertise, educator agency, human-AI relationships, and associated opportunities and tensions.

Methods: A qualitative phenomenological inquiry involved 23 educators from 3 Palestinian universities. Purposive maximum-variation sampling sought variation in academic rank, discipline, teaching experience, and experience with AI tools. Individual semistructured, in-depth interviews were conducted from March to April 2026, audio-recorded, transcribed verbatim, and returned to participants for review. Data were analyzed using reflexive thematic analysis informed by phenomenology. Initial coding was inductive and grounded in participants' accounts; human-centered AI and sociomateriality subsequently informed interpretation as sensitizing concepts. Themes were developed through iterative comparison of transcripts, analytic memos, and alternative interpretations. Informational sufficiency was observed after approximately 20 interviews, with 3 additional interviews used to examine the developing thematic structure.

Results: The analysis generated 8 interconnected themes. Educators described a shift from content transmission toward mediation, inquiry, reflective learning, creative exploration, and process-oriented assessment. Professional expertise increasingly centered on facilitation, pedagogical design, contextual interpretation, ethical guidance, and evaluation of AI-generated content. Human-AI collaboration involved bounded and integrated configurations, with humans retaining responsibility for judgment and educational consequences. Educator agency appeared through decisions about AI participation, independent student reasoning, and the boundaries of human expertise. Credibility, bias, authorship, authenticity, and academic integrity became embedded as pedagogical concerns. Participants valued opportunities for creativity, personalization, flexibility, and instructional support while expressing concerns about dependence, reduced intellectual effort, and weakened human connection.

Conclusions: GenAI-mediated pedagogical transformation depends on educators' alignment of AI use with learning purposes, the allocation of responsibilities, and the preservation of ethical accountability and meaningful relationships. The resulting human-centered pedagogy framework positions pedagogical intentionality, contextual judgment, negotiated agency, and relational presence as central to responsible integration. Universities should complement technical training and institutional policies with support for pedagogical design, critical evaluation, process-oriented assessment, and ethical reasoning.

(JMIR Med Educ 2026;12:e105172) doi:10.2196/105172

KEYWORDS

generative AI; higher education; relational pedagogy; qualitative research; phenomenology

XSL•FO
RenderX

Introduction

Background

AI is rapidly transforming higher education, challenging traditional pedagogical paradigms and prompting educators to reconsider how teaching and learning are designed and enacted [1-3]. Recent advances in generative AI (GenAI), adaptive learning technologies, and intelligent tutoring systems have created new opportunities to personalize instruction, enhance feedback, and support data-informed educational decision-making [4,5]. However, integrating these technologies into educational practice also presents pedagogical, ethical, and professional challenges, requiring educators to continuously adapt their teaching approaches and decision-making processes [6,7].

GenAI now participates in knowledge creation, feedback, assessment, and curriculum work, placing educators at the point where technical capability becomes pedagogical practice [8,9]. Their decisions determine how AI-generated content is evaluated, contextualized, and integrated into student learning [6,10]. Although research has documented AI adoption, effectiveness, academic integrity, and implementation, less attention has been given to how educators interpret changing roles, authority, and responsibility in everyday teaching [4,11-13].

AI in Higher Education: From Technological Innovation to Pedagogical Change

AI scholarship in higher education has often emphasized efficiency, personalization, and assessment, framing technology primarily as an instrument for improving established educational processes. GenAI complicates this view because it participates directly in explanation, drafting, feedback, and academic discourse, intervening in how knowledge is produced and authorized in settings defined by inquiry, authorship, and disciplinary judgment [5,11]. Unlike earlier digital tools that mainly organized or delivered learning, generative systems produce content during the pedagogical encounter. Their significance therefore arises from how generated material changes academic engagement, evaluation, and the distribution of intellectual work [7,14].

These capabilities create opportunities for adaptation and personalized support while simultaneously intensifying concerns about academic integrity, overreliance, mediation, and responsibility. Educators must, therefore, determine which activities may benefit from AI, which forms of judgment should remain human, and how educational value can be sustained when AI-generated content becomes part of knowledge production [8,15]. GenAI thus represents not merely another technological resource but a development that may require educators to reconsider the purposes, processes, and relationships underlying teaching and learning.

Emerging Pedagogy in the AI Era

Emerging pedagogy concerns how teaching, knowledge mediation, and learning relationships are organized under conditions of AI participation. The central question is therefore not simply whether educators use AI, but how educational purposes and practices change when AI-generated content enters instruction, assessment, and feedback. The findings in the study [16] and [17] conceptualize this transformation through teacher positionality, rationality, and functionality, shifting attention from technical proficiency toward the roles and interactions through which authority, creativity, feedback, and meaning-making are organized.

The increasing availability of explanations, examples, and AI-generated content weakens information possession as the primary basis for teacher authority. Educators increasingly evaluate, contextualize, and direct AI outputs, placing discernment, relational guidance, and ethical interpretation at the center of pedagogical work [15]. Emerging AI-mediated pedagogy nevertheless remains contested. Creativity and flexibility coexist with concerns about authenticity, agency, judgment, and dependence, requiring educators to decide when to trust, challenge, limit, or redirect AI [9,11,18]. Consequently, pedagogical transformation in the AI era involves not only adopting new tools but also renegotiating how educational purposes, responsibilities, and relationships are enacted.

Educators' Lived Experiences in the AI Era

Competency models specify what educators should know and be able to do with AI, but they do not fully explain how teaching with AI is experienced within institutional expectations, disciplinary values, and changing professional identities. A phenomenological perspective therefore provides an important lens for examining uncertainty, adaptation, and the meanings educators construct through human-machine relations [3,17]. This focus is particularly important in higher education, where authorship, critique, originality, disciplinary judgment, and professional responsibility shape academic practice.

Examining educators' lived experiences can reveal how they interpret and negotiate tensions involving authority, control, agency, and ethics as AI becomes embedded in routine teaching [13,16]. Such an approach moves beyond the questions of whether educators adopt AI or possess sufficient technical competencies and instead examines how AI-mediated practices are understood, experienced, and incorporated into educators' professional identities and pedagogical decision-making.

Theoretical Orientation: Human-Centered AI and Sociomateriality

Two complementary theoretical perspectives provide the conceptual orientation for interpreting educators' experiences. First, human-centered AI emphasizes high levels of human control alongside capable automation and foregrounds self-efficacy, responsibility, safety, and trustworthiness [19]. Applied to higher education, this perspective focuses on educators' judgment about when AI can support learning, where boundaries are needed, and how accountability remains human.

Second, sociomateriality treats practice as constituted through relations among people, technologies, institutional norms, and local contexts [20]. This perspective supports interpreting AI-mediated pedagogy as enacted through interconnected educator-student-AI relations rather than as an outcome of technology alone. Together, these perspectives provide complementary ways of understanding pedagogical

XSL•FO
RenderX

transformation: human-centered AI foregrounds human judgment, responsibility, and control, whereas sociomateriality highlights the relational and contextual conditions through which AI-mediated practices are enacted. In the present study, both perspectives operated as sensitizing concepts during later interpretation rather than as a fixed deductive codebook; first-cycle coding remained grounded in participants' accounts.

Research Gap

Despite the growing body of scholarship on AI in higher education, an important gap remains concerning how educators experience and interpret pedagogical change after GenAI becomes part of routine teaching practice. Existing conceptual scholarship identifies changes in teacher positionality, classroom relations, authorship, agency, and ethical judgment [14-16,18], while empirical research has examined adoption, competencies, effectiveness, and implementation. However, qualitative evidence on how educators themselves make sense of these changes and enact them in everyday pedagogical practice remains limited.

In particular, insufficient attention has been given to how educators negotiate the relationship between AI affordances and human pedagogical judgment, how their professional roles and agency are reconstructed, and how they manage tensions between efficiency and educational purpose, technological assistance and human responsibility, and AI-supported activity and authentic learning. Addressing this gap requires attention to educators' lived experiences and meaning-making rather than focusing exclusively on technological adoption or competency. Accordingly, this study adopts a phenomenologically informed approach to examine how higher education educators experience and interpret emerging pedagogy in the AI era.

Study Purpose and Contribution

This study aimed to investigate higher education educators' experiences of emerging pedagogy in the context of GenAI, with particular attention to changes in professional roles and agency, the ways educators exercise pedagogical judgment, and the opportunities and tensions associated with AI-mediated teaching. The inquiry therefore focused on relationships, judgment, and educational purpose, as well as technological tools and competencies [7,10]. An educator-centered interpretation of pedagogical change can clarify support needs relevant to faculty development, curriculum design, and institutional policy [21,22]. Understanding how educators negotiate AI-mediated teaching can also inform professional development that connects technical capability with critical evaluation, ethical judgment, and context-sensitive pedagogical design [17,23,24]. By focusing on educators' lived experiences, this study seeks to contribute an empirically grounded understanding of how pedagogical purposes, professional roles, human-AI relationships, and educational responsibilities are being reconstructed as GenAI becomes increasingly embedded in higher education.

Research Questions

The study addressed the following research questions:

1. How do higher education educators experience emerging pedagogy in the AI era?
2. How do higher education educators make sense of the influence of AI on their pedagogical roles, judgment, and agency?
3. How do educators manage the pedagogical opportunities and tensions that arise when teaching in AI-mediated higher education environments?

Accordingly, this study aimed to explore how higher education educators experience pedagogical transformation in the context of GenAI, with particular attention to changes in pedagogical roles, professional judgment, agency, human-AI relationships, and the opportunities and tensions associated with AI-mediated teaching.

Methods

Research Design

We used a qualitative phenomenological design to explore how higher education educators experienced and interpreted pedagogical transformation in the age of GenAI. Phenomenology guided the inquiry toward the meanings attributed to experience, while reflexive thematic analysis supported the interpretation of patterned meanings across participants' accounts. This methodological combination enabled attention to individual meaning-making and shared experiential patterns without treating themes as objective entities discovered independently of the researchers' interpretive work [25-27].

Research Context and Participants

The study was conducted across 3 large comprehensive universities in Palestine that offer undergraduate and postgraduate programs in medicine and health sciences, engineering, education, business, the humanities, and other fields. Each institution had begun encouraging GenAI use through professional development or teaching initiatives, creating varied settings for examining educators' experiences.

The study included 23 higher education educators with direct experience using GenAI in assessment design, feedback, course preparation, instructional support, or learning activities. Participants were recruited through purposive maximum-variation sampling [28,29]. Institutional contacts and professional academic networks circulated invitations to educators who met 2 inclusion criteria: employment in higher education and practical experience integrating GenAI into teaching. No invited participants declined participation or withdrew from the study. A demographic questionnaire asked participants to classify their experience as beginner, intermediate, or advanced according to the duration, frequency, and breadth of teaching-related AI use. These self-reported categories supported variation in the sample and did not represent a formal competency assessment.

Maximum-variation sampling captured shared and context-specific experiences across disciplines, ranks, teaching tenure, and self-reported AI experience. Table 1 summarizes the sample, and Appendix D in Multimedia Appendix 1 links P1 to P23 with disciplinary background, rank, teaching experience, and AI experience.

XSL•FO
RenderX

Table . Demographic characteristics of participants (N=23).

Variable and category	Frequency, n (%)
Gender	
Man	12 (52.2)
Woman	11 (47.8)
Academic rank	
Lecturer	6 (26.1)
Assistant professor	9 (39.1)
Associate professor	5 (21.7)
Professor	3 (13.0)
Years of teaching experience	
1 - 5 years	5 (21.7)
6 - 10 years	7 (30.4)
11 - 15 years	6 (26.1)
More than 15 years	5 (21.7)
Discipline	
Humanities and social sciences	3 (13.0)
Education	5 (21.7)
Engineering and technology	4 (17.4)
Business and economics	4 (17.4)
Medical and health sciences	7 (30.4)
Experience with AI tools	
Beginner	5 (21.7)
Intermediate	11 (47.8)
Advanced	7 (30.4)

Data Collection

Data collection was conducted from March to April 2026 through a single individual semistructured, in-depth interview with each participant. Both researchers were men faculty members with PhD degrees and had prior experience conducting qualitative research and teaching qualitative research courses in higher education. Both researchers conducted the interviews. Before the interviews, participants were informed about the researchers' academic and research backgrounds, as well as the purpose of the study.

The interview guide used broad questions and flexible probes to elicit experiences, interpretations, and reflections. Literature on teacher positionality, pedagogical transformation, human-AI collaboration, professional agency, ethics, and human-centered AI informed the guide's domains, while the wording remained open enough for participants to introduce unanticipated meanings. Before data collection, the interview guide was pilot-tested through a cognitive interview to assess the clarity, relevance, and comprehensibility of the questions. Appendix C in Multimedia Appendix 1 provides the interview protocol and representative questions.

Each interview lasted approximately 45 to 60 minutes. Face-to-face interviews were conducted at locations within the participants' universities chosen by the participants, whereas a small number of participants preferred online interviews conducted in the evening. No individuals other than the participant and the researcher were present during the interviews. The study included no classroom observations and therefore used no observation protocol.

Before each interview, participants received information about the purpose of the study, voluntary participation, confidentiality, withdrawal rights, and audio recording, after which written consent was obtained. Interviews were audio-recorded with permission and transcribed verbatim. Field notes documented interview context and researchers' preliminary reflections; they were not treated as observational data. Interview transcripts were returned to participants for review and to allow them to correct, add, or clarify their responses. No participants requested changes to their transcripts.

Data Analysis

Interview data were analyzed using reflexive thematic analysis informed by the phenomenological orientation [25,26,30]. Repeated transcript reading supported familiarization, followed by first-cycle inductive coding of meaningful segments close to participants' accounts. Both researchers independently coded the interview transcripts before discussing and refining the

XSL•FO
RenderX

developing coding framework. Differences in coding and interpretation were addressed through reflexive discussion. The researchers did not use voting, calculate intercoder agreement, or designate one researcher as the final decision maker. When coding differences occurred, they returned to the relevant transcript and discussed the assumptions underlying each interpretation. They also considered other participants' accounts, reflexive memos, and the study aim. This process led to clearer code definitions and the combination of overlapping codes. They retained distinct codes when they represented meaningfully different experiences. The researchers also revised the boundaries and descriptions of categories and themes when necessary. Differences in theme generation were handled through the same process. Each proposed theme was examined for internal coherence, distinctiveness, and support across the dataset. Decisions were documented in the audit trail. Competing interpretations were examined further until the researchers developed a shared interpretive account consistent with reflexive thematic analysis. Human-centered AI and sociomateriality supplied sensitizing concepts during later interpretation, but they did not determine the initial code set or impose themes in advance. Codes were iteratively compared and grouped into conceptual categories, which were refined into subthemes and overarching themes through recursive movement among transcripts, codes, memos, and developing interpretations. The team examined how participants connected pedagogical purpose, professional role, human-AI practice, agency, and ethics. Reflexive memos and an audit trail documented coding choices, alternative interpretations, category development, and theme revision. NVivo 15 (Lumivero, LLC) supported data organization, retrieval, and comparison. To enhance methodological transparency, Appendix A in Multimedia Appendix 1 illustrates the progression from interview excerpts to codes, categories, subthemes, and themes. Appendix B (Multimedia Appendix 1) presents the complete coding framework, and Appendix C (Multimedia Appendix 1) provides the semistructured interview protocol.

Data Saturation

Data collection and analysis proceeded concurrently, allowing the team to assess informational sufficiency as the interpretation developed. After approximately 20 interviews, additional accounts no longer introduced substantively new meanings or conceptual relations and instead elaborated or qualified the existing analysis. This pattern approximated meaning saturation, as distinguished from early code saturation [31]. Sample adequacy was also considered through information power, including the focused aim, participants' relevant experience, and the richness of the interviews [30,31]. Three additional interviews tested the stability and completeness of the developing thematic structure.

Trustworthiness

We addressed credibility, transferability, dependability, and confirmability [32]. Credibility was supported through in-depth interviews, repeated transcript engagement, verbatim quotations linked to interpretive claims, and participant feedback. Preliminary themes were shared with participants who agreed to provide feedback, allowing them to comment on whether the interpretations reflected their experiences. An external qualitative researcher also independently coded approximately 10% of the transcripts using the developing framework. Differences in interpretation prompted analytic dialogue, clarification of code boundaries, and revision of the framework. This external review supported critical dialogue and was not used for intercoder reliability testing, consistent with the reflexive character of the analysis.

Transferability was supported through a detailed description of Palestine, the institutional setting, recruitment, inclusion criteria, participant characteristics, and maximum variation, allowing readers to assess relevance to other contexts. Dependability rested on an audit trail documenting recruitment, data collection, coding choices, category development, alternative interpretations, theme revision, and methodological decisions [28]. Appendix A in Multimedia Appendix 1 illustrates analytical progression, and Appendix B in Multimedia Appendix 1 presents the complete coding framework. Confirmability was supported through analytic memos, reflexive journals, transcript checks, and records of alternative interpretations. These materials connected the final account to participants' words while acknowledging the researchers' interpretive role.

Researcher Reflexivity

Given the phenomenological orientation, the researchers maintained reflexivity throughout data collection, analysis, and interpretation [27,33]. Both researchers were men, faculty members with PhD degrees, and had prior experience conducting qualitative research and teaching qualitative research courses in higher education. Both researchers conducted the interviews. Before the interviews, participants were informed about the researchers' academic and research backgrounds and the study's purpose. The team comprised scholars with experience in educational technology, qualitative research, and GenAI in higher education. Collectively, they developed the protocol, conducted interviews, reviewed transcripts, coded data, developed themes, and examined the final interpretation. This background provided contextual familiarity and also created assumptions that required continuing scrutiny. Bracketing was practiced as an ongoing reflexive discipline, not as a claim that prior knowledge could be completely suspended. Interviewers recorded reflexive notes after data collection, and the team maintained journals, analytic memos, and an audit trail during coding and theme development. Regular discussions examined emerging interpretations against the transcripts and considered plausible alternatives. No prior supervisory, instructional, or evaluative relationships existed between researchers and participants, which reduced potential power asymmetries.

Ethical Considerations

The study received ethical approval from the Research Department, Ministry of Higher Education (approval number March 2026/21). Before participating, all educators received information about the study's purpose and procedures, the voluntary nature of participation, their right to decline to answer any question, and their right to withdraw from the study without penalty. We obtained written informed consent from all participants before conducting the interviews.

XSL•FO
RenderX

We protected participant privacy and confidentiality throughout the research process. Each participant was assigned a pseudonymous alphanumeric identifier (P1-P23), which was used in transcripts, data analysis, and reporting. Identifying information was removed from the research data and manuscript, and demographic and professional information was limited to details necessary for interpreting the findings to minimize the risk of deductive identification. Interview recordings and transcripts were stored securely and were accessible only to the research team. Participants received no financial or material compensation for their participation in the study.

Reporting Guideline

This qualitative study was reported in accordance with the COREQ (Consolidated Criteria for Reporting Qualitative Research [34]), a 32-item checklist for qualitative studies involving interviews and focus groups. The completed COREQ checklist is provided as Checklist 1.

Results

Participant Characteristics

The study included 23 higher education educators from 3 Palestinian universities. The participants included 12 (52.2%) men and 11 (47.8%) women. Gender was self-reported by participants. No participants identified with a different gender category. Additional participant characteristics, including academic rank, teaching experience, discipline, and experience with AI tools, are presented in Table 1.

Following the presentation of participant characteristics, the analysis generated 8 interconnected themes describing educators' experiences of pedagogical transformation in the context of GenAI.

Overview of Thematic Findings

Educators experienced GenAI as a reorganization in pedagogy, assessment, professional roles, and responsibilities. They identified possibilities for reflection, creativity, personalization, and innovation alongside uncertainty involving agency, authenticity, dependence, ethics, and human connection. The 8 themes represent interconnected dimensions of this experience.

Theme 1: Pedagogical Transformation

Participants located pedagogical change in knowledge mediation, assessment, and learning design. AI prompted the reconsideration of classroom interaction and educational purpose, not the routine addition of another tool.

From Transmission to Mediation

Educators reported moving away from transmission-based teaching toward a mediating role focused on helping students critically interpret and evaluate AI-generated material. As P1 explained:

> *I am no longer just explaining content. I spend much more time helping students evaluate and use what AI gives them. Instead of providing immediate answers, I encourage them to question the AI's responses, compare different sources, and think about why one answer may be more appropriate than another. I feel that my role has shifted from being the main source of knowledge to becoming someone who guides students in making informed judgments about the information they receive.* [P1]

P1's account locates pedagogical authority in guided evaluation. Asking students to question outputs, compare sources, and justify the suitability of an answer makes the educator responsible for cultivating judgment while students assume greater responsibility for evaluating information. Authority therefore shifts from possessing answers to orchestrating reasoned inquiry.

Process-Oriented Pedagogy

P4 described a growing emphasis on learning processes alongside final products:

> *Now I care much more about how students used AI in developing their work than simply looking at the final answer. I want to understand the prompts they created, how they refined the responses, and why they accepted or rejected particular suggestions. Watching this process tells me much more about their learning than reading the finished assignment alone.* [P4]

P4's account recasts assessment as evidence of decision-making. Prompts, revisions, acceptances, rejections, and explanations made students' reasoning visible and allowed the educator to evaluate how learning developed through interaction with AI.

Metacognitive and Reflective Pedagogy

P5 incorporated AI into activities that promoted students' awareness of their own thinking:

> *Whenever students submit AI-assisted work, I ask them to explain how they used the tool and what they decided to change. I want them to reflect on which ideas came from AI, which ideas were their own, and why they made certain decisions. These conversations help students realize that learning is not about accepting AI's answers but about understanding their own thinking and improving it.* [P5]

P5's practice required students to distinguish generated suggestions from their own decisions and explain why they revised particular ideas. The AI-supported activity, therefore, became a prompt for metacognition when reflection and accountability were built into the task.

Fluid and Evolving Pedagogy

Participants also emphasized that pedagogy in the AI era remains dynamic and continually evolving. As P3 stated:

> *This is not a fixed teaching model yet. Every semester I find myself changing my approach because both the technology and my students are changing. What worked a few months ago may no longer be effective, so I keep experimenting, reflecting, and adjusting my teaching. I think we are all still trying to understand what good teaching with AI really looks like.* [P3]

P3's account defines "good teaching with AI" as context-sensitive experimentation, reflection, and revision. The uncertainty does not indicate an absence of pedagogical purpose;

XSL•FO
RenderX

it required continuing professional learning as technological capabilities and student practices change.

Theme 2: Creative and Inquiry-Based Pedagogy

The second theme concerns inquiry, creativity, and authentic learning. Educators described AI as useful when it stimulated curiosity, supported exploration, and drew students into sustained engagement with complex questions.

Pedagogy of Wonder

Participants emphasized that the educational value of AI lay not in accelerating answers but in encouraging students to ask richer and more meaningful questions. As P2 explained:

> *The value is not getting an answer faster, but using AI to push students toward better questions. I encourage my students to keep asking the AI different questions, challenge the responses, and explore alternative explanations instead of accepting the first answer they receive. When students become curious enough to continue questioning, I know the learning process has become much deeper than simply searching for information.* [P2]

P2 evaluated AI through the quality of questions it enabled students to formulate. Educational value emerged when an initial response generated challenge, comparison, and further inquiry, so curiosity and critical exploration displaced answer speed as the measure of learning.

Iterative Inquiry

Participants also described AI-supported learning as an iterative process in which students continuously refined their understanding through repeated interaction with the technology. As P6 reflected:

> *The process became more iterative. Students ask, revise, question, and go back again with AI. They rarely stop after the first response because every interaction leads to another question or another revision. I noticed that students became much more comfortable experimenting with different ideas and improving their work through repeated dialogue with the AI.* [P6]

P6's account presents learning as recursive: each response invites questioning, revision, and another attempt. Understanding develops through repeated evaluation and refinement instead of a fixed sequence from instruction to assessment.

AI-Enhanced Creativity

Educators treated generated suggestions as starting points whose value depended on student interpretation and transformation. As P11 explained:

> *AI helps generate possibilities, but the real creativity comes from how students transform those ideas. I tell my students that AI can give them many starting points, but it cannot decide which ideas are meaningful or how they should be combined. Creativity happens when students evaluate, modify, and build on those suggestions to create something that reflects their own thinking.* [P11]

Creativity, therefore, resided in selecting, combining, and reshaping possibilities into work that reflected the student's judgment; generation alone did not constitute creative authorship.

Challenge- or Project-Based Pedagogy

Educators also designed authentic tasks that required students to apply AI within real-world problems. As P8 explained:

> *I started designing tasks where students use AI to solve real problems instead of simply completing routine assignments. Rather than asking them to reproduce information, I ask them to investigate authentic situations, evaluate different solutions, and justify the decisions they make. AI becomes one resource within a much larger process of analysing and solving complex problems.* [P8]

Project- and challenge-based tasks placed AI within analysis, justification, and problem solving. Technology supported inquiry while students retained ownership of decisions and outcomes.

Theme 3: Repositioning the Educator's Role

Overview

GenAI also altered how educators understood professional value. Facilitation, curation, ethical mentorship, design, and contextual judgment became central to their accounts of expertise.

Teacher positionality shift. Participants described moving away from seeing themselves as the primary source of knowledge toward becoming facilitators who help students assess increasingly complex information environments. As P7 explained:

> *Students can get information anywhere now, so my role is more about guiding and framing their learning. Instead of spending most of my time delivering content, I focus on helping students decide which information is credible, how different ideas connect, and why certain sources are more trustworthy than others. I feel that my value as a teacher comes from helping students make sense of information rather than simply providing it.* [P7]

P7 located authority in framing, connection, and credibility assessment. Expertise rested on helping students make sense of abundant information, not on controlling access to it.

Educator as a Curator

Participants also emphasized that AI-generated content required continuous human evaluation, adaptation, and contextualization before it could be integrated into teaching. As P10 reflected:

> *AI can produce content quickly, but I still have to choose what is accurate and relevant for my course. Sometimes the information is incomplete, lacks the appropriate clinical context, or does not match my students' level of understanding. I spend a great deal of time reviewing, adapting, and reorganizing the material so that it supports my learning objectives rather than simply accepting what AI generates.* [P10]

XSL•FO
RenderX

AI shifted professional effort toward selecting, refining, and contextualizing content. Curation, therefore, required disciplinary knowledge, awareness of student level, and alignment with learning objectives.

Educator as an Ethical Mediator

Beyond instructional responsibilities, participants described assuming a more explicit ethical role in helping students use AI responsibly. As P9 explained:

> *I now spend more time discussing authorship, bias, and responsible AI use as part of teaching. Students often assume that if AI produces an answer, it must be correct, so we have conversations about transparency, academic integrity, and the importance of verifying information. These ethical discussions have become just as important as teaching the course content itself.* [P9]

P9's account places ethical guidance within routine higher education pedagogy. Transparency, verification, authorship, and bias become part of disciplinary learning, so technical competence cannot be separated from professional and societal responsibility.

Shifting Professional Identity

Participants further reflected on how AI had prompted them to reconsider their own professional identity and purpose as educators. As P15 stated:

> *AI has forced me to rethink what makes me valuable as a teacher beyond delivering information. I realised that students need someone who can challenge their thinking, encourage reflection, and help them apply knowledge responsibly in real situations. The technology has changed the questions I ask myself about teaching, and in many ways it has strengthened my understanding of what my role should be.* [P15]

P15 interpreted AI as a prompt to reconstruct their professional identity around challenge, reflection, responsible application, and meaningful learning relationships. The perceived source of value moved from content delivery to contextual and ethical guidance.

Theme 4: Human-AI Relational Pedagogy

A fourth theme concerns the relational patterns through which educators organized AI-supported work. Participants used interpersonal metaphors to describe how AI influenced task development and instructional decision-making while retaining human accountability for interpretation and judgment.

Rationality With AI

P13 described AI as shaping interactions and task development: "It feels less like using software and more like working with a participant that keeps shaping the task." This metaphor described the iterative influence of generated responses on pedagogical thinking; it did not attribute consciousness or moral agency to the system. Human educators remained responsible for evaluating and directing each contribution.

P16's account locates educational value in the interpretive work through which educators and students evaluate, adapt, and refine generated material. Cocreation describes a process of joint activity without assigning moral agency to AI.

Bounded collaboration ("centaur" configuration). P14 described a divided arrangement in which AI-generated options were provided while humans retained final judgment: "I let AI generate options, but I keep the final judgment and evaluation for myself and the students." The term identifies a workflow with explicit boundaries, where efficiency is accepted and interpretive responsibility remains human.

Integrated collaboration ("cyborg" configuration). P13 described continuous AI-assisted iteration across the teaching process: "Sometimes I work with AI back and forth through the whole teaching process, almost like an ongoing collaboration." The term identifies an integrated workflow, not a fusion of human and machine agency. Continuous involvement increased the need for contextual evaluation and professional accountability.

Theme 5: Negotiating Agency in AI-Mediated Teaching

Participants described agency as continual decisions about when, how, and to what extent AI should influence teaching and assessment. Technological assistance remained subordinate to educator-defined purposes and professional autonomy.

Retaining Pedagogical Control

P17 emphasized deliberate control over when and how AI entered teaching: "I use AI strategically, but I do not let it define the lesson or the assessment." AI functioned as a resource for educator-defined purposes, preserving responsibility for learning objectives and assessment design.

P18's account shows why contextual reading remained irreducibly human: classroom dynamics, student needs, and disciplinary expectations required situated interpretation. AI suggestions acquired pedagogical value only through that judgment.

P19 set boundaries according to the educational purpose of the task. Selective use of support protected independent reasoning and intellectual effort while allowing support where it strengthened learning.

Agency Under Pressure

P20 reported institutional and societal pressure to adopt AI because it was associated with innovation: "There is pressure to use AI because it is seen as progress, but I still need to decide what serves learning." This account separates technological uptake from pedagogical improvement and locates responsible adoption in professional judgment.

Theme 6: Critical and Ethical Pedagogy

Critical and ethical reasoning formed part of everyday AI-mediated pedagogy. Educators linked technical use to the capacity to assess reliability, bias, authorship, and consequence.

Bias and Accuracy Assessment

P21 treated critical evaluation as a learning outcome, requiring students to identify errors, biases, and contextual limits before using the generated material.

XSL•FO
RenderX

Algorithmic Literacy

P22 extended AI literacy beyond operation to include an understanding of how systems produce responses, where errors arise, and why evaluation remains necessary.

Academic Integrity and Authenticity

P4's concern moved assessment toward evidence of reasoning, including prompts, revisions, and explanations, so authenticity could be evaluated through the process as well as the product.

Ethics as Pedagogy

P7's account shows ethics entering classroom dialogue, assessment, and feedback as a guiding principle for responsible engagement with AI and its social consequences.

Theme 7: Pedagogical Opportunities and Tensions

Educators evaluated AI through the interplay between pedagogical gains and educational costs. Efficiency, personalization, and access were weighed against dependence, reduced intellectual effort, and weakened human connection.

Efficiency in Planning and Feedback

P6 valued time savings when they released effort for discussion, individual support, and deeper teaching; efficiency mattered because of its pedagogical use.

Personalized and Differentiated Learning

P9 viewed AI as expanding the range of support available to students, while the educator remained responsible for adapting that support to individual needs.

Asynchronous and Ubiquitous Learning

P12 linked AI to learning beyond fixed classroom time and space: "Students can now continue engaging with course content and feedback outside the classroom in new ways." This flexibility expands access to practice and feedback while increasing the need for clear expectations and educator guidance.

Overdependence on AI

P14's concern identifies intellectual struggle as part of learning. Immediate generation could weaken independent thought when it displaces questioning and sustained engagement.

Loss of human connection. P18 treated dialogue, empathy, and trust as conditions of learning that technological support could not reproduce. Excessive reliance therefore risked distancing students from relational presence.

Theme 8: Human-Centered Pedagogy

The final theme brings the preceding tensions into a human-centered orientation. Participants judged AI through empathy, responsibility, professional judgment, and the preservation of meaningful interaction.

Preserving Care and Presence

P2's account defines care as responsiveness to students' emotions, needs, and contexts. Technology could assist learning, but it could not assume this relational responsibility.

Meaningful Learning Beyond Efficiency

P11 distinguished productivity from educational quality. Speed mattered only when it supported reflection, critical thinking, and deep understanding.

Human Responsibility in Cocreated Learning

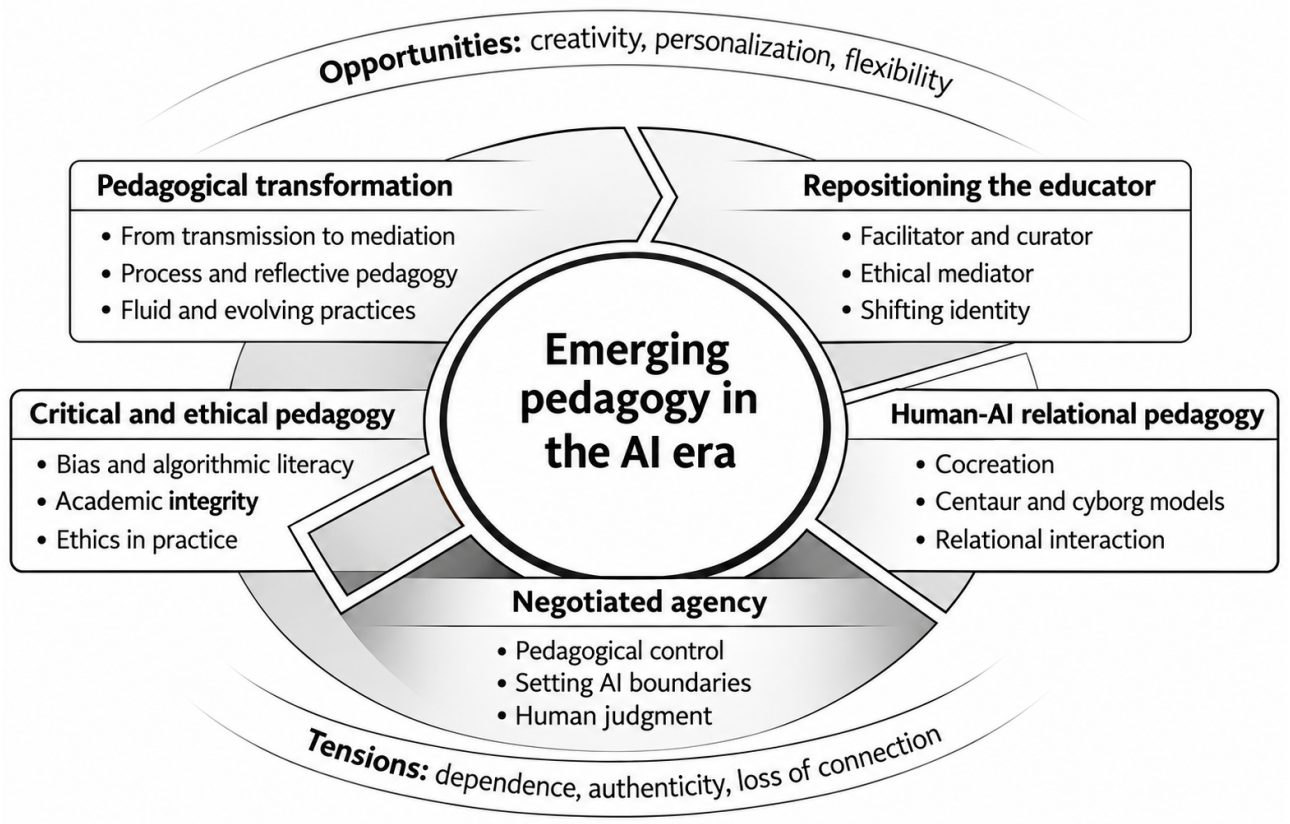
P16 retained human accountability for instructional design, ethical decisions, and the meaning of learning even when AI contributed to knowledge production.

A Human-Centered Conceptual Framework for Emerging Pedagogy in the AI Era

We developed an empirically grounded interpretive framework comprising pedagogical transformation, educator repositioning, human-AI relational practice, negotiated agency, and critical-ethical pedagogy (Figure 1). Opportunities and tensions form the surrounding conditions, while human-centered pedagogy provides the organizing orientation. The synthesis arose from recurrent links across accounts and neither depends on code frequency nor claims causal validation.

XSL•FO
RenderX

Figure 1. A human-centered conceptual framework for emerging pedagogy in the AI era.



The dimensions are recursive: shifts in educational purpose reshape professional roles; role reconstruction influences patterns of AI-supported work; these patterns prompt decisions about control and responsibility; and ethical judgment informs design, assessment, engagement, and boundaries.

Unlike adoption, literacy, digital competence, or instructional-function models, the framework interprets how pedagogy is reconstructed after AI enters everyday practice [3,15,25]. Human-centered AI clarifies control and accountability, while sociomateriality explains practice through relations among educators, students, systems, disciplines, and institutions [19,20].

Discussion

Principal Findings

This study explored how higher education educators experienced pedagogical transformation in the context of GenAI, focusing on changes in pedagogical practice, professional expertise, agency, human-AI relationships, and the opportunities and tensions associated with AI-mediated teaching. The findings show that educators experienced GenAI as more than an additional instructional technology. Its integration prompted them to reconsider how teaching, assessment, professional responsibility, and educational relationships should be organized. Participants described a movement from content transmission toward mediation, inquiry, reflection, creative exploration, and process-oriented assessment. They repositioned professional expertise around pedagogical design, contextual judgment, ethical guidance, and the allocation of responsibilities between humans and AI systems. Although they valued creativity, personalization, flexibility, and instructional support, they remained concerned about dependence, authenticity, academic integrity, weakened student effort, and reduced human interaction. Overall, the educational value of GenAI depended on how educators aligned its use with pedagogical purposes, ethical responsibilities, and the relational conditions of learning.

Interpretation, Comparison With Previous Research, and Implications

The movement from content transmission toward inquiry, reflection, and evaluation indicates that pedagogical transformation cannot be understood solely as technology adoption. Previous research has highlighted the potential of GenAI to support personalization, feedback, creativity, and instructional efficiency [15,34]. The present findings qualify these claims by showing that technological affordances do not automatically produce meaningful learning. Participants designed activities that required students to interrogate AI-generated outputs, evaluate their credibility, revise them, and explain their decisions. They also made prompting, reasoning, and revision visible within the learning process. AI-generated content, therefore, became an object of critical examination rather than a finished academic product. Technological capability created possibilities for change, but educators' intentional design and metacognitive guidance determined whether these possibilities contributed to learning [8,9,14].

This transformation also reconstructed professional expertise. As GenAI increasingly supported information retrieval, summarization, explanation, and content generation, participants located their professional value less in transmitting information

XSL•FO
RenderX

and more in facilitation, curation, contextual interpretation, pedagogical design, and ethical mediation. This represents a more substantial change than adding technical competencies to an established teaching role [3]. AI-generated information still required interpretation within disciplinary, cultural, institutional, and classroom contexts. Educators' authority therefore rested on recognizing students' needs, guiding through uncertainty, evaluating AI-generated content, and connecting knowledge with disciplinary standards and educational purposes. Technical competence acquired educational value only when combined with pedagogical, relational, and interpretive expertise.

The findings further clarify the relational character of human-AI collaboration. Participants described bounded collaboration, in which AI generated ideas or alternatives while humans retained explicit evaluative authority, and integrated collaboration, in which AI-assisted iteration continued throughout a task. Both configurations could support questioning, idea development, and revision [7,14]. Neither configuration, however, was inherently pedagogically effective. Their value depended on how educators established boundaries, distributed responsibilities, interpreted outputs, and maintained accountability. From a sociomaterial perspective, learning emerged through relationships among educators, students, AI systems, disciplinary expectations, assessment practices, and institutional conditions rather than through the technology alone [14]. Consequently, the same AI function could support inquiry in one context but encourage dependence or superficial engagement in another.

Educator agency was similarly expressed through professional discretion rather than a binary choice between adoption and resistance. Participants decided when AI should contribute, which activities required independent student reasoning, how AI-generated outputs should be evaluated, and where human expertise needed to remain central. This interpretation extends beyond adoption models, which can explain why educators use technology but offer less insight into how they allocate cognitive, pedagogical, and ethical responsibilities after adoption. Consistent with human-centered AI, the findings position professional judgment as the mechanism through which technological capabilities are combined with meaningful human control and accountability [19].

Ethical reasoning was embedded within pedagogical practice rather than treated solely as an external regulatory requirement. Previous scholarship has commonly addressed AI ethics through governance, privacy, academic integrity, transparency, and technological risk [18]. Participants incorporated these concerns into teaching and assessment by asking students to examine credibility, identify bias, consider authorship, disclose AI use, and evaluate the consequences of relying on AI-generated content. AI literacy therefore extended beyond operational competence to include critical evaluation, intellectual responsibility, and informed judgment [35]. This suggests that responsible AI use cannot be developed through institutional rules alone. It must also be cultivated through instructional activities, assessment, dialogue, and feedback.

Participants' simultaneous recognition of opportunities and concerns should not be interpreted as inconsistency. They valued AI's capacity to support creativity, personalization, flexibility, and instructional assistance while expressing concern about dependence, weakened effort, reduced authenticity, and diminished human interaction [4,35]. These positions reflected evaluations of different pedagogical consequences. AI use was viewed positively when it supported inquiry, critical thinking, student agency, and meaningful engagement. It was viewed more cautiously when it replaced reasoning, obscured authorship, or weakened educator-student interaction. Educators' attitudes were therefore conditional on the purposes, boundaries, and relationships associated with particular uses of AI.

The human-centered pedagogy framework integrates these findings by connecting pedagogical transformation, educator repositioning, human-AI relational practice, negotiated agency, and critical-ethical pedagogy. It extends models centered primarily on technology adoption, competence, or instructional function [18,34,36]. The framework identifies pedagogical intentionality as the connection between technological affordances and educational value, contextual judgment as a foundation of professional expertise, responsibility allocation as an expression of agency, and ethical reasoning as part of pedagogy. These dimensions are recursive: changes in assessment reshape educators' roles, reconstructed roles influence how tasks are distributed between humans and AI, and these arrangements generate further decisions about transparency, accountability, and oversight.

The framework also demonstrates the complementary contributions of human-centered AI and sociomateriality. Human-centered AI explains the continuing importance of professional judgment, responsibility, oversight, and appropriate boundaries. Sociomateriality explains how these decisions are enacted through relationships among educators, students, AI systems, disciplinary practices, and institutional conditions [20,22]. Pedagogical agency is therefore neither located entirely within educators nor transferred to technology. It is exercised through how educators organize human and technological contributions while retaining responsibility for their educational consequences. Because the framework was developed interpretively from educators' accounts, it remains open to refinement and empirical examination in other disciplines, institutions, and cultural settings.

These findings have practical implications for faculty development, curriculum, assessment, and institutional policy. Professional development should connect technical knowledge with learning design, process-oriented assessment, critical evaluation, ethical reasoning, and responsible human-AI collaboration. Educators need opportunities to design activities that make students' prompting, reasoning, verification, revision, and justification visible. They also need support in determining when AI enhances learning and when it replaces the intellectual effort or interpersonal engagement an activity is intended to develop. Professional judgment should therefore be treated as a central capacity for AI integration rather than as an outcome that follows automatically from technical competence.

Universities should also translate general principles concerning bias, credibility, authorship, authenticity, transparency, and academic integrity into classroom and assessment practices.

XSL•FO
RenderX

Ethical guidance becomes educationally meaningful when students apply it while evaluating AI-generated content and explaining their decisions. Institutional policies should establish shared expectations for responsibility and transparency while preserving educators' capacity to make context-sensitive decisions. Uniformly restrictive approaches may prevent valuable experimentation, whereas overly permissive policies may weaken academic standards and leave educators without sufficient guidance. Finally, universities should ensure that efficiency and personalization do not displace interaction, mentorship, care, and educator presence. AI integration should be evaluated not only by what it automates but also by whether it strengthens pedagogical quality, student agency, ethical responsibility, and meaningful human relationships.

Limitations and Future Research

This study should be interpreted in light of several limitations. The findings were derived from interviews with 23 educators across 3 Palestinian universities. The qualitative design supported an in-depth examination of lived experiences and professional judgment, but the findings are contextually situated and are not intended to be statistically generalized to all higher education institutions. Institutional policies, technological infrastructure, disciplinary traditions, and sociocultural expectations may shape AI-mediated pedagogy differently elsewhere. The contextual detail and variation within the sample may nevertheless help readers assess the transferability of the findings to comparable settings.

The study focused on educators' accounts and did not include students, institutional leaders, or policymakers. It therefore explains how educators understood pedagogical transformation but cannot establish whether other stakeholders experienced or interpreted AI-supported practices similarly. Future research should incorporate multiple perspectives and examine where understandings of responsible AI use converge or differ. Research involving students could investigate metacognition, creativity, critical AI literacy, academic integrity, dependence, and perceptions of educator presence.

The findings were also based on reported experiences rather than direct observations of classroom practices or analyses of AI-supported learning activities. Participants' accounts provided access to the interpretations and professional reasoning central to the study's aim, but they may not fully represent how pedagogical decisions are enacted. Future studies could combine interviews with classroom observations, assessment tasks, student work, AI interaction records, or institutional documents to examine relationships between stated pedagogical intentions and enacted practices.

GenAI technologies, institutional policies, and educational practices are developing rapidly. The findings, therefore, represent a particular period within an ongoing transition. Longitudinal research is needed to examine how professional roles, pedagogical judgment, and the allocation of human-AI responsibilities evolve. Finally, the proposed framework is interpretive rather than predictive or causal. Comparative, mixed methods, and quantitative studies could refine and examine relationships among its dimensions across disciplines, institutions, and cultural contexts.

Conclusions

The growing presence of GenAI in higher education requires more than the adoption of new tools or the development of technical competencies. It requires universities to reconsider how educational purpose, professional authority, student agency, and responsibility are defined when teaching and knowledge production increasingly involve AI systems. The central challenge is not simply whether GenAI should be used but how its use can be organized so that technological capabilities strengthen rather than displace critical thinking, meaningful assessment, ethical responsibility, and human relationships.

The human-centered pedagogy framework provides a foundation for addressing this challenge by placing pedagogical intentionality, contextual judgment, negotiated responsibility, and relational engagement at the center of AI-supported education. Responsible integration cannot be achieved through technical training, regulation, or access to technology alone. Universities must enable educators and students to question AI-generated outputs, justify decisions, establish appropriate boundaries, and remain accountable for the educational consequences of AI use. As GenAI continues to evolve, the quality of higher education will depend not only on what these systems can produce but also on whether institutions preserve the professional judgment, student agency, ethical reasoning, and human connection that give education its purpose.

Generative AI Use

The generative AI tool ChatGPT-5 was used during manuscript preparation to assist with language editing, clarity, and organization. The authors have critically reviewed and revised all AI-assisted content and take full responsibility for the accuracy, integrity, and final content of the manuscript.

Funding

The authors declared no financial support was received for this work.

Data Availability

The interview data generated and analyzed during this study are not publicly available because of participant confidentiality and ethical restrictions. Deidentified data may be available from the corresponding author upon reasonable request, subject to ethical and institutional requirements.

XSL•FO
RenderX

Authors' Contributions

Conceptualization: ZK, BH
Formal analysis: ZK, BH
Investigation: ZK, BH
Methodology: ZK, BH
Resources: BH
Visualization: BH
Writing – original draft: ZK, BH
Writing – review & editing: ZK, BH

Conflicts of Interest

ZK was a previous associate editor for *JMIR Medical Education*.

Multimedia Appendix 1
Coding book for analyzing higher education educators' lived experiences of emerging pedagogy in the AI era.
[DOCX File, 26 KB - mededu_v12i1e105172_app1.docx]

Checklist 1
COREQ checklist.
[PDF File, 7 KB - mededu_v12i1e105172_app2.pdf]

References

1. Çifçi İ, Çetin G, Şahin MA, Karatay C. Educators' AI interactions in higher education. Ankara Uluslararası Sosyal Bilimler Dergisi 2024:52-64 [FREE Full text]
2. Fitzgerald R, Amantha Kumar J, Roe J, Roehrer E, Yang J. Editorial: framing the future with a research agenda for artificial intelligence in higher education. J Univ Teach Learn Pract 2025;22(4). [doi: 10.53761/jwt7ra63]
3. Zou D, Xie H, Kohnke L. Navigating the future: establishing a framework for educators' pedagogic artificial intelligence competence. Eur J Educ 2025 Jun;60(2):e70117. [doi: 10.1111/ejed.70117]
4. Mishra HR, Sarvaiya MD. Exploring the impact of artificial intelligence tools in engineering pedagogy: a qualitative survey of academic experiences. Int J Res Appl Sci Eng Technol 2024;12(1):60-66. [doi: 10.22214/ijraset.2024.57864]
5. Ashwin S, Yeter IH. Exploring faculty members' artificial intelligence literacy through the lens of the TPACK framework: a qualitative study. In: 2025 ASEE Annu Conf Expo: American Society for Engineering Education (ASEE); 2025:11. [doi: 10.18260/1-2--56507]
6. Cambra-Fierro JJ, Blasco MF, López-Pérez MEE, Trifu A. ChatGPT adoption and its influence on faculty well-being: an empirical research in higher education. Educ Inf Technol 2025 Feb;30(2):1517-1538. [doi: 10.1007/s10639-024-12871-0]
7. Chiu TKF. Future research recommendations for transforming higher education with generative AI. Comput Educ Artif Intell 2024 Jun;6:100197. [doi: 10.1016/j.caeai.2023.100197]
8. Medina-Gual L, Parejo JL. University students' engagement with artificial intelligence: a cluster analysis of learner profiles in AI literacy. Tech Know Learn 2026 Mar;31:291-309. [doi: 10.1007/s10758-025-09926-7]
9. Siddiqui MT, Mansoori MV, Siddiqui MA, Yadav A. AI-enabled pedagogy: advancing education through innovative teaching tools and the AI-TEACH model. J Inform Educ Res 2025;5(1). [doi: 10.52783/jier.v5i1.2261]
10. Alshehri B. Pedagogical paradigms in the AI era: insights from Saudi educators on the long-term implications of AI integration in classroom teaching. Int J Educ Sci Arts 2023;2(8):159-180. [doi: 10.59992/IJESA.2023.v2n8p7]
11. Arvin N, Hoseinabady M, Bayat B, Zahmatkesh E. Teacher experiences with AI-based educational tools. AI Tech Behav Soc Sci 2023;1(2):26-32. [doi: 10.61838/kman.aitech.1.2.5]
12. Karkoulian S, Sayegh N, Sayegh N. ChatGPT unveiled: understanding perceptions of academic integrity in higher education - a qualitative approach. J Acad Ethics 2025 Sep;23(3):1171-1188. [doi: 10.1007/s10805-024-09543-6]
13. Taufiq Hail GAM, Yusof SAM, Rashid A, El-Shekeil I, Lutfi A. Exploring factors influencing Gen Z's acceptance and adoption of AI and cloud-based applications and tools in academic attainment. Emerg Sci J 2024;8(3):815-836. [doi: 10.28991/ESJ-2024-08-03-02]
14. Creely E, Carabott K. Teaching and learning with AI: an integrated AI-oriented pedagogical model. Aust Educ Res 2025;52(6):4633-4654. [doi: 10.1007/s13384-025-00913-6]
15. Zawacki-Richter O, Bai JYH, Lee K, Slagter van Tryon PJ, Prinsloo P. New advances in artificial intelligence applications in higher education? Int J Educ Technol High Educ 2024;21(1):32. [doi: 10.1186/s41239-024-00464-3]
16. Qureshi I. The impact of AI on teacher roles: towards a collaborative human-AI pedagogy. AI EDIFY J 2025;2(1):1-11 [FREE Full text]
17. Ayyoub AM, Khlaif ZN, Shamali M, et al. Advancing higher education with GenAI: factors influencing educator AI literacy. Front Educ 2025;10:1530721. [doi: 10.3389/feduc.2025.1530721]

XSL•FO
RenderX

18. Zhou X, Zhang JJ, Chan C. Unveiling students' experiences and perceptions of artificial intelligence usage in higher education. J Univ Teach Learn Pract 2024;21(6). [doi: 10.53761/xzjprb23]
19. Shneiderman B. Human-centered artificial intelligence: reliable, safe and trustworthy. Int J Hum Comput Interact 2020;36(6):495-504. [doi: 10.1080/10447318.2020.1741118]
20. Orlikowski WJ. Sociomaterial practices: exploring technology at work. Organ Stud 2007 Sep;28(9):1435-1448. [doi: 10.1177/0170840607081138]
21. Bayaga A. Leveraging AI-enhanced and emerging technologies for pedagogical innovations in higher education. Educ Inf Technol 2025;30(1):1045-1072. [doi: 10.1007/s10639-024-13122-y]
22. Strzelecki A. Students' acceptance of ChatGPT in higher education: an extended unified theory of acceptance and use of technology. Innov High Educ 2024;49(2):223-245. [doi: 10.1007/s10755-023-09686-1]
23. Burneo-Arteaga P, Lira Y, Murzi H, Balula A, Costa AP. Capability-based training framework for generative AI in higher education. Front Educ 2025;10:1594199. [doi: 10.3389/feduc.2025.1594199]
24. Nevárez Montes J, Elizondo-Garcia J. Faculty acceptance and use of generative artificial intelligence in their practice. Front Educ 2025;10:1427450. [doi: 10.3389/feduc.2025.1427450]
25. Braun V, Clarke V. Reflecting on reflexive thematic analysis. Qual Res Sport Exerc Health 2019;11(4):589-597. [doi: 10.1080/2159676X.2019.1628806]
26. Braun V, Clarke V. Thematic Analysis: A Practical Guide: SAGE Publications; 2021.
27. Khlaif ZN, Alshakhshir R, Hamamra B, Joma AI. Assistive generative AI for visually impaired learners: personalization and inclusion in higher education. Int J Artif Intell Educ 2026 Mar;36(1-2):100008. [doi: 10.1016/j.ijaied.2026.100008]
28. Creswell JW, Poth CN. Qualitative Inquiry and Research Design: Choosing Among Five Approaches, 4th edition: SAGE Publications; 2018.
29. Patton MQ. Qualitative Research & Evaluation Methods: Integrating Theory and Practice, 4th edition: SAGE Publications; 2015.
30. Braun V, Clarke V. Using thematic analysis in psychology. Qual Res Psychol 2006;3(2):77-101. [doi: 10.1191/1478088706qp063oa]
31. Hennink MM, Kaiser BN, Marconi VC. Code saturation versus meaning saturation: how many interviews are enough? Qual Health Res 2017 Mar;27(4):591-608. [doi: 10.1177/1049732316665344] [Medline: 27670770]
32. Lincoln YS, Guba EG. Naturalistic Inquiry: SAGE Publications; 1985.
33. Finlay L. Negotiating the swamp: the opportunity and challenge of reflexivity in research practice. Qual Res 2002 Aug;2(2):209-230. [doi: 10.1177/146879410200200205]
34. Tong A, Sainsbury P, Craig J. COnsolidated Criteria for Reporting Qualitative Research (COREQ): a 32-item checklist for interviews and focus groups. Int J Qual Health Care 2007 Dec;19(6):349-357. [doi: 10.1093/intqhc/mzm042] [Medline: 17872937]
35. Mishra P, Warr M, Islam R. TPACK in the age of ChatGPT and Generative AI. J Digit Learn Teach Educ 2023 Oct 2;39(4):235-251. [doi: 10.1080/21532974.2023.2247480]
36. Long D, Magerko B. What is AI literacy? Competencies and design considerations. Proc CHI Conf Hum Factors Comput Syst 2020:1-16. [doi: 10.1145/3313831.3376727]

Abbreviations

COREQ: Consolidated Criteria for Reporting Qualitative Research
GenAI: generative AI

Edited by A Stone; submitted 21.Jun.2026; peer-reviewed by A Derakhshan, C Gabarre, D Itmazi; revised version received 22.Aug.2026; accepted 24.Aug.2026; published 06.Nov.2026.

Please cite as:
Khlaif Z, Hamamra B
Toward Human-Centered Pedagogy in the Age of Generative AI Among Higher Education Educators: Qualitative Phenomenological Study
JMIR Med Educ 2026;12:e105172
URL: https://mededu.jmir.org/2026/1/e105172
doi:10.2196/105172

© Zuheir Khlaif, Bilal Hamamra. Originally published in JMIR Medical Education (https://mededu.jmir.org), 6.Nov.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium,

XSL•FO
RenderX

provided the original work, first published in JMIR Medical Education, is properly cited. The complete bibliographic information, a link to the original publication on https://mededu.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Medical Students' Acceptance of Digital Entrustable Professional Activities: Results of a Cohort Study

Maximilian Domann; Constanze Richters, PhD; Matthias J Witti, PhD; Matthias Stadler, Prof Dr

Institute of Medical Education, LMU University Hospital, LMU Munich, Pettenkoferstr. 8A, Munich, Bavaria, Germany

Corresponding Author:
Matthias Stadler, Prof Dr
Institute of Medical Education, LMU University Hospital, LMU Munich, Pettenkoferstr. 8A, Munich, Bavaria, Germany

Abstract

Background: Digital entrustable professional activities (EPAs) in simulated environments may accelerate competency acquisition, but adoption depends on learner acceptance. The Technology Acceptance Model (TAM) posits that perceived usefulness (PU) and perceived ease of use (PEU) shape attitudes (AT) and, in turn, behavioral intention (BI).

Objective: This study aimed to examine medical students' acceptance of digital EPAs and to test the hypothesized TAM relationships among PU, PEU, AT, and BI.

Methods: Clinical-phase medical students at Ludwig Maximilian University of Munich completed a TAM-based survey (7-point Likert scales) after reading a canonical analog EPA and its digital counterpart. Confirmatory analyses comprised bivariate correlations and hierarchical regressions testing TAM paths. Exploratory analyses comprised paired-sample two-tailed *t* tests comparing analog versus digital ratings and path modeling to evaluate global TAM fit.

Results: Between 70 and 72 medical students provided complete, usable responses, depending on the construct. Mean ratings were favorable (≈5/7). Internal consistency was acceptable (ω=.67–.80). Within the digital EPAs, PU strongly predicted AT (β=.59; *P*<.001), and AT predicted BI (β=.58; *P*<.001). For the analog EPAs, PU (β=.54; *P*<.001) and PEU (β=.28; *P*=.005) predicted AT; both AT (β=.42; *P*<.001) and PU (β=.36; *P*=.002) predicted BI. Attitudes were modestly higher for analog versus digital (M=5.18 vs 4.87; t_{71}=−2.50, *d*=−0.30; *P*=.02), but PU, PEU, and BI did not differ significantly. The path models indicated excellent fit for both formats (comparative fit index=1; root mean square error of approximation=0; standardized root mean square residual ≤.01).

Conclusions: Students reported high acceptance of digital EPAs. Acceptance was driven primarily by PU (via AT), whereas PEU contributed to AT only for analog EPAs. Implementation should emphasize demonstrable educational value and cultivate positive attitudes; subsequent work should link acceptance to actual use and learning outcomes.

International Registered Report Identifier (IRRID): RR2-10.2196/59326

(JMIR Med Educ 2026;12:e87605) doi:10.2196/87605

KEYWORDS

medical education; entrustable professional activities; simulation; technology acceptance model; digital learning; undergraduate medical education; EPA; TAM

Introduction

Entrustable professional activities (EPAs) have reshaped medical education by shifting the focus toward competency-based learning. EPAs break down complex clinical skills into manageable units, enabling learners to acquire proficiency progressively while allowing teaching staff to structure supervision and maximize learning outcomes [1,2]. In competency-oriented curricula, responsibilities and levels of autonomy are made explicit and can be increased as competence grows, culminating in independent performance with appropriate oversight [3,3-7]. Such transparency benefits students, educators, and institutions by aligning expectations and clarifying which tasks can be performed at which level of supervision [7,8].

In parallel, simulation-based training has become integral to contemporary medical education. Simulations create realistic, interactive environments in which students can safely practice essential cognitive and motor skills without risk to patients. Computer-based simulations in particular provide dynamic representations of clinical scenarios and have demonstrated benefits for decision-making and procedural learning in higher education [7,9-12]. In domains such as surgery, where theoretical knowledge must be combined with hands-on technique, simulation-based training has shown promise for accelerating skill acquisition and transfer to clinical performance [12]. Recent reviews further highlight that advances in digital technologies, including immersive simulations and data-driven feedback systems, are increasingly shaping how

XSL•FO
RenderX

simulation-based training is designed and implemented in medical education [13].

Despite the established value of both EPAs and simulation-based training, their systematic integration into digital EPAs remains underexplored. We use the term "digital EPAs" to denote entrustable tasks instantiated in simulated or virtual environments so that learners can rehearse the same competencies before encountering them in real clinical care. Digital EPAs could enhance learning efficiency, enable standardized exposure to complex scenarios, deliver immediate and scalable feedback, and capture rich performance data to inform programmatic assessment. Recent evidence from virtual simulation environments further suggests that such digital formats may also enhance learners' motivation and perceived value of training, thereby supporting acceptance alongside learning outcomes [14]. Beyond medical education, recent work across clinical professions highlights that digital health competencies are increasingly expected in clinical practice, yet remain inconsistently embedded in formal training and competency frameworks, underscoring the need for structured educational approaches [15]. At the same time, digital EPAs may introduce new barriers, including usability burdens, cognitive overload, or mismatches with learners' expectations and existing curricular workflows. These competing forces make it insufficient to assume that established benefits of simulation or EPAs will automatically translate to digital EPAs. Rather, acceptance by end users must be established empirically before large-scale implementation is warranted [9,12,16].

To examine technology adoption in this context, the Technology Acceptance Model (TAM) provides a robust explanatory framework. TAM posits two central beliefs—perceived usefulness (PU) and perceived ease of use (PEU)—that shape attitudes (AT) toward a technology and, in turn, behavioral intention (BI) to use it (and ultimately actual usage) [17,18]. PU captures the belief that a system will enhance performance; PEU reflects the effort required to operate it. Together, these constructs explain why learners adopt educational technologies and provide a practical framework to identify leverage points for implementation in medical education.

In medical education, where novel digital tools continue to proliferate, applying TAM can yield actionable insights into which features matter most for learner uptake and sustained use. A recent systematic review confirms the continued relevance of the TAM in medical education, demonstrating its applicability across a broad range of digital learning technologies and educational contexts [19]. Prior work applying TAM in educational contexts has demonstrated that PU and PEU can meaningfully predict AT and BI toward a wide range of learning technologies [20,21]. However, the specific TAM predictors of acceptance for digital EPAs have not yet been characterized. While prior methodological work has outlined approaches for developing digital health–related EPAs and defining their structure and content, empirical evidence on learners' acceptance of such digital EPA formats remains scarce [16].

By operationalizing TAM constructs for this setting and testing their associations with behavioral intention among clinical-phase medical students, this study seeks to provide evidence to guide the design of learner-centered digital EPAs and inform decisions about their integration into medical curricula. By focusing on learners in the clinical phase, where competency acquisition and supervised autonomy are most salient, this study addresses a critical evidence gap at the intersection of EPAs, simulation, and technology acceptance. The findings aim to inform the design and implementation of digital EPA curricula and to support educators and developers in aligning functionality and user experience with the factors that drive adoption.

Based on the predictions of the TAM, we hypothesize that:

- Hypothesis 1: the more useful medical students find digital EPAs in medical simulations, the higher their behavioral intention to use this technology as part of their studies will be
- Hypothesis 2: the easier medical students find the use of digital EPAs in medical simulations to be, the greater their behavioral intention to use this technology will be
- Hypothesis 3: the more positive the attitude of medical students toward digital EPAs in medical simulations is, the higher their behavioral intention to use this technology will be

In addition, we will also test these hypotheses for analog EPAs to exploratively compare the fit of the TAM for analog EPAs to that of digital EPAs.

Methods

Overview

The methodology closely aligns with the Research Protocol, which describes it in more detail [17]. Further, the digital EPA condition was assessed using a standardized narrative description of a hypothetical digital EPA rather than interaction with a functional software system, as no such system was available at the time of the study.

Participants and Recruitment

Overview

This cohort study targeted medical students in the clinical phase of their studies at Ludwig Maximilian University (LMU) Munich. This population was optimally suited for the application of digital EPAs, as the clinical phase centers on acquiring specific clinical skills. The survey link was distributed through major student email lists (eg, the "Module 6" distribution list), ensuring that only students in the clinical phase received the invitation. Participation was voluntary, and students could complete the questionnaire at their convenience; overall, no incentives were offered.

Inclusion Criteria

Participants had to be at least in their third year of medical school (having completed the preclinical foundational courses in anatomy, physiology, biochemistry, etc) and must have passed the first medical licensing examination. This ensured that all respondents had a sufficient clinical and educational background to appraise EPAs in context.

XSL•FO
RenderX

Exclusion Criteria

Students needed a sufficient command of the German language to understand the study materials and survey, since the questionnaire and EPA descriptions were provided in German.

Survey Instrument and Hypotheses

We developed a web-based questionnaire using SoSciSurvey based on the TAM, adapting an existing one from Abdul Ghani et al [18] to the context of digital EPAs [19].

The survey included items (rated on 7-point Likert-type scales) to measure four TAM constructs: PU, PEU, AT, and BI. Each construct was assessed with a multiitem scale, and reliability (internal consistency) was measured with McDonald ω. Prior to fielding the survey, the questionnaire was reviewed for clarity and technical functionality and was internally pilot-tested to ensure usability and correct item presentation. Data collection took place between May 2025 and June 2025.

Study Procedure and Materials

Because EPAs (especially in digital form) are a relatively unfamiliar concept for students, we provided participants with a concrete example of an EPA and its digital counterpart before they answered the survey. Specifically, participants were shown a description of a sample analog EPA (a clinical task with defined supervision levels) and a description of the same task implemented as a hypothetical digital EPA. These materials were developed based on the guidelines by Cate and Taylor [2] for formulating EPAs. The digital EPA was presented conceptually (in narrative text form) since no fully functional digital platform was available; students were asked to imagine using a system that would allow them to perform the EPA in a simulated digital environment. All materials, as well as the full descriptions of the example EPA (analog version and conceptual digital version), are available in the project's Open Science Framework repository [22] as well as Multimedia Appendices 1-4. Both German and English versions of these descriptions were provided (the survey itself was administered in German).

After reviewing the EPA descriptions, participants proceeded to the TAM questionnaire. They were instructed to respond to the TAM items twice: once reflecting on the analog EPA scenario and once for the digital EPA scenario. In other words, for each TAM construct (PU, PEU, AT, and BI), students provided separate ratings regarding (1) the traditional analog EPA and (2) the digital EPA. This within-subjects design allowed direct comparison of acceptance between the analog and digital formats.

Statistical Analysis

All analyses were conducted using the jamovi software [20]. We first computed descriptive statistics for each TAM construct (mean and SD) and evaluated internal consistency (McDonald ω) for each multi-item scale. We examined bivariate correlations among all TAM constructs (PU, PEU, AT, and BI) separately for the digital EPA ratings and the analog EPA ratings.

To test our hypotheses about the determinants of attitude and intention, we performed hierarchical multiple regressions. For attitude (AT) as the outcome, we entered PEU in the first model (to assess its effect alone) and then added PU in the second model, for both digital and analog conditions. This allowed us to see whether PU provided additional explanatory power for attitude beyond PEU. Similarly, for behavioral intention (BI) as the outcome, we first entered attitude alone (Model 1), then added PU in a second model, analyzing digital and analog conditions separately. In each case, we report standardized regression coefficients (β), *P* values, the proportion of variance explained (R^2), and the change in R^2 for the stepwise addition.

We supplemented the regression analyses with path modeling to simultaneously evaluate the TAM relationships (PEU → AT, PU → AT, AT → BI, and direct PU → BI) and to compare model fit for the digital and analog data. Model fit was assessed with the comparative fit index (CFI), Tucker-Lewis Index (TLI), root mean square error of approximation (RMSEA), and standardized root mean square residual.

Finally, although not specified in the original protocol, we conducted exploratory paired-sample 2-tailed *t* tests to directly compare students' mean ratings for the analog EPA versus the digital EPA on each TAM construct. This was done to gage any overall preference for one format over the other (eg, whether attitudes toward the analog EPA were significantly different from attitudes toward the digital EPA, etc). Effect sizes for mean differences were calculated using Cohen *d*.

All hypothesis tests used a significance level of α=.05 (2-tailed). The target sample size (n=68) was determined a priori via power analysis (G*Power 3.1) for detecting moderate effect sizes with power 0.80 at α=.05 in regression models [23].

CHERRIES Checklist

The online survey was conducted and reported in accordance with the CHERRIES (Checklist for Reporting Results of Internet E-Surveys) checklist. The questionnaire was distributed over multiple pages with a limited number of items per page. Participants were able to review and modify their responses before final submission. No randomization or adaptive questioning was applied. To prevent multiple entries from the same individual, the survey platform used standard technical measures (eg, cookies and IP-based restrictions). No personally identifiable information was collected. View rates and participation rates could not be calculated, as access logs of unique visitors were not available.

Ethical Considerations

This study was approved by the ethics committee of the LMU University Hospital, Munich (project No. 23 - 0577). Before accessing the questionnaire, participants were presented with an electronic informed consent page describing the study purpose, voluntary participation, approximate survey duration, and data handling procedures. No personally identifiable information was collected, and all data were stored on secure servers in accordance with institutional data protection regulations. Participants received no compensation for their participation.

XSL•FO
RenderX

Results

Overview

We first examined students' overall evaluations of analog and digital EPAs to establish baseline acceptance and scale reliability. Depending on the construct, between 70 and 72 students provided complete, usable responses (the exact *n* varies slightly because some participants left a few items blank for certain scales). All analyses below reflect the data from those who provided complete responses for the given measures.

Descriptive Statistics and Reliability

Overall, students' evaluations of both the digital EPA and the analog EPA were positive. Mean scores on the 7-point scales were around 5 for all TAM constructs (Table 1). For the digital EPA, the mean scores (with SDs) were 4.94 (SD 1.10) for perceived usefulness, 4.94 (SD 1.21) for perceived ease of use, 4.87 (SD 1.13) for attitude, and 4.85 (SD 1.05) for behavioral intention. Analog EPAs were rated slightly higher on average, with means ranging from 4.98 to 5.18. All constructs demonstrated acceptable internal consistency, with McDonald ω coefficients between .67 and .80 across the scales. The lower coefficients (particularly ω≈.67) are plausible in this early-stage application of an adapted TAM instrument to a novel and partly hypothetical learning concept, where item interpretations may still vary across respondents. Moreover, the relatively small number of items per construct and the homogeneous sample can attenuate internal consistency estimates.

Table . Descriptive statistics and internal consistency of Technology Acceptance Model constructs (n=70 - 72).

Construct	Digital EPAs[a], mean (SD)	n	Digital ω	Analog EPAs, mean (SD)	n	Analog ω
Perceived usefulness (PU)	4.94 (1.10)	72	.75	5.17 (1.09)	72	.80
Perceived ease of use (PEU)	4.94 (1.21)	70	.76	5.02 (0.99)	70	.67
Attitude (AT)	4.87 (1.13)	72	.79	5.18 (0.97)	72	.73
Behavioral intention (BI)	4.85 (1.05)	72	.74	4.98 (1.02)	72	.72

[a]EPA: entrustable professional activity.

Overall, these descriptive results indicate generally favorable evaluations of both analog and digital EPAs, with mean ratings close to the upper end of the scale and acceptable internal consistency across constructs.

Exploratory Comparison of Analog vs Digital

To explore whether students evaluated analog and digital EPAs differently overall, we conducted paired-sample 2-tailed *t* tests comparing ratings across formats. These exploratory analyses (which were not specified in the study protocol) revealed that attitudes toward the analog EPA were slightly but significantly more positive than attitudes toward the digital EPA (t_{71}=–2.50, *d*=–0.30; *P*=.02). In contrast, there were no statistically significant differences between analog and digital formats in perceived usefulness (t_{71}=–1.46; *P*=.15), perceived ease of use (t_{69}=–0.39; *P*=.70), or behavioral intention (t_{71}=–0.73; *P*=.47). Thus, apart from a small difference in attitude favoring the analog EPA, students' evaluations of usefulness, ease of use, and behavioral intention did not differ substantially between formats.

Correlations Among TAM Constructs

Before testing the hypothesized TAM relationships in regression analyses, we examined bivariate correlations among all TAM constructs for both digital and analog EPAs. Bivariate correlations between all pairs of constructs are presented in Table 2 for the digital and analog versions, respectively. As expected, the TAM variables were moderately to strongly intercorrelated. For digital EPAs (Table 2), perceived usefulness had a strong positive correlation with attitude (*r*=.64; *P*<.001), and attitude in turn was strongly correlated with behavioral intention (*r*=.70; *P*<.001). Perceived ease of use showed moderate positive correlations with both perceived usefulness (*r*=.45; *P*<.001) and attitude (*r*=.55; *P*<.001). For analog EPAs (Table 2), a very similar pattern emerged: perceived usefulness correlated strongly with attitude (*r*=.66; *P*<.001), and attitude correlated strongly with behavioral intention (*r*=.65; *P*<.001). The correlation between perceived ease of use and perceived usefulness was of moderate size (*r*=.41; *P*<.001), as was the correlation between perceived ease of use and attitude (*r*=.49; *P*<.001). Overall, the correlation patterns were consistent with the TAM and justified proceeding to regression and path analyses.

XSL•FO
RenderX

Table . Correlations among Technology Acceptance Model constructs for digital entrustable professional activities (n=70-72)[a].

Variable	Perceived ease of use	Perceived usefulness	Attitude	Behavioral intention
Perceived ease of use	—	.41	.49	.46
Perceived usefulness	.45	—	.66	.61
Attitude	.55	.64	—	.65
Behavioral intention	.50	.48	.70	—

[a]Values below the diagonal are for analog entrustable professional activities, values above the diagonal are for digital entrustable professional activities. All correlations *P*<.001.

Predictors of AT

We next examined which TAM variables predicted students' attitudes toward EPAs using hierarchical regression analyses. For digital EPAs, perceived ease of use alone (Model 1) explained 31% of the variance in attitude, but in this model PEU's effect was not statistically significant (β=.14; *P*=.29). When perceived usefulness was added (Model 2), the model's explanatory power increased significantly (ΔR^2=.17, total R^2=.48; *P*<.001). In the final model for digital EPAs, PU emerged as a strong positive predictor of attitude (β=.59; *P*<.001), whereas PEU remained nonsignificant (β=.14; *P*=.29).

For analog EPAs, perceived usefulness alone accounted for 44% of the variance in attitude (Model 1, β=.68; *P*<.001). Adding perceived ease of use (Model 2) produced a small but significant increase in explained variance (ΔR^2=.06, total R^2=.51; *P*=.005). In the final analog model, both predictors had significant effects: PU (β=.54; *P*<.001) and PEU (β=.28; *P*=.01) each independently contributed to a positive attitude. In summary, perceived usefulness emerged as the dominant predictor of attitude in both formats, whereas perceived ease of use contributed additionally only in the analog condition.

Predictors of BI

Subsequently, we analyzed predictors of behavioral intention to use EPAs. For digital EPAs, students' attitude toward the digital EPA was by far the dominant predictor of their behavioral intention. Attitude alone explained 48% of the variance in BI (Model 1, β=.58; *P*<.001). Adding perceived usefulness did not significantly improve the model (ΔR^2=.01; *P*=.40), and in the full model attitude remained the only significant predictor of BI (β=.58; *P*<.001, while PU's direct effect was β=.09; *P*=.17, not significant). This indicates that for digital EPAs, behavioral intention was primarily driven by students' attitude, and any influence of perceived usefulness operated through attitude rather than as a separate direct effect.

For analog EPAs, attitude was also a strong predictor of BI, accounting for 43% of variance on its own (Model 1, β=.42; *P*<.001). However, in contrast to the digital case, adding perceived usefulness yielded a significant improvement (ΔR^2=.07; *P*=.002), bringing the total R^2 to .51 for Model 2. In the final analog model, both attitude (β=.42; *P*<.001) and PU (β=.36; *P*=.002) had significant independent effects on behavioral intention. Thus, for analog EPAs, students' intention to use was jointly determined by their attitude and their perceived usefulness of the EPA.

Path Modeling

Finally, to evaluate the overall fit of the hypothesized TAM structure, we conducted path modeling for both digital and analog EPAs. We specified a model in which PEU and PU both load onto AT (with direct paths PEU→AT and PU→AT), AT loads onto BI (AT→BI), and PU also has a direct path to BI. This corresponds to the classic TAM with attitude as a mediator and a possible direct effect of usefulness on intention. The model was estimated separately for the digital EPA ratings and the analog EPA ratings. Model fit was excellent for both the digital model and the analog model: CFI=1, TLI=1.05 for both; RMSEA=0; standardized root mean square residual=0.004 (digital) and 0.002 (analog).

These fit indices generally indicate that the hypothesized TAM relationships almost perfectly reproduced the observed covariance patterns in our data. We interpret these findings cautiously since global indices such as CFI or RMSEA may appear near perfect even in modest samples when models are small and comparatively constrained. Accordingly, we focus primarily on the pattern and magnitude of the path estimates, which converged with the hierarchical regression findings.

The path estimates were consistent with the regression findings. In the digital EPA model, perceived usefulness had a significant positive effect on attitude, and attitude had a strong positive effect on behavioral intention. The direct path from PU to BI was not significant (in line with our earlier result that attitude mediated the effect of PU). In the analog EPA model, both perceived usefulness and perceived ease of use significantly influenced attitude, and both attitude and PU significantly influenced behavioral intention. In both models, the vast majority of the explainable variance in BI was accounted for (which is reflected in the high R^2 values of ~0.5 in the regressions and the excellent fit indices here). Overall, the path models corroborated the regression findings and supported the applicability of the TAM structure to both EPA formats.

Discussion

Principal Findings

This study examined medical students' acceptance of digital EPAs within simulated learning environments, guided by the TAM. Several important findings emerged. First, perceived usefulness was the strongest predictor of students' attitudes toward using digital EPAs, and attitude in turn was the primary driver of behavioral intention. For digital EPAs, once attitude was taken into account, perceived usefulness no longer had a

XSL•FO
RenderX

direct effect on behavioral intention, suggesting that the influence of usefulness operated mainly by shaping favorable attitudes.

Second, while perceived ease of use predicted attitudes toward analog EPAs, its contribution was not statistically significant for digital EPAs. This pattern may reflect that usability-related considerations were less salient in the hypothetical digital condition and should therefore be interpreted cautiously.

Third, students' attitudes toward the analog EPA were slightly more favorable than their attitudes toward the digital EPA, although the mean scores for both formats were high overall (both approaching 5 on a 7-point scale). Finally, path modeling confirmed an excellent fit for the hypothesized TAM pathways in both conditions, underscoring the robustness of the TAM framework in this new educational context.

Together, these findings highlight that medical students are generally open to the concept of digital EPAs, but their acceptance seems to be primarily associated with perceived educational value in this study context, while usability considerations may become more salient once learners interact with a functional system. In practical terms, this pattern suggests that effective implementation of digital EPAs will require making the educational benefits transparent and directly aligned with students' learning goals, rather than focusing exclusively on technical usability or novelty.

Comparison With Prior Work

The dominance of perceived usefulness as a predictor of attitudes and intentions is consistent with a large body of TAM research in educational technology. Similar patterns have been reported in studies of simulation-based learning, virtual reality training, and digital assessment tools in medical education, where perceived educational value consistently outweighs usability considerations during early adoption phases [9,12,13]. Prior studies on e-learning platforms and immersive digital assessment tools likewise show that perceived usefulness is often a stronger driver of intention than perceived ease of use [16, 21]. Our findings extend this evidence to the novel domain of digital EPAs. While ease of use is usually considered a critical factor when learners first encounter new technologies, its weaker role in our data may reflect the nature of our intervention: the digital EPA was presented conceptually rather than as a hands-on system. This aligns with recent findings from virtual simulation research suggesting that learners' initial acceptance is primarily driven by perceived relevance and learning value when interaction with a fully functional system is not yet possible [14].

We also observed that attitudes toward the analog EPA were slightly more positive than attitudes toward the digital EPA. This echoes previous observations that learners often exhibit a preference for traditional or familiar tools when a novel digital innovation is first introduced. This kind of "familiarity advantage" may explain why analog EPAs scored marginally higher in our sample. Importantly, however, the differences we found were small, and overall attitudes toward the digital EPA were still quite favorable. This suggests that resistance to this form of digital innovation is not deep-seated and may diminish with increased exposure. Such familiarity effects have also been described in competency-based medical education, where learners initially favor established assessment formats before developing trust in novel digital implementations [15].

Interestingly, in the analog condition, perceived ease of use significantly predicted attitude in addition to perceived usefulness (both contributed to higher attitude in the regression model). This aligns with TAM's original formulation, in which both usability and usefulness shape attitudes. In the analog context, "ease of use" rather refers to the clarity, structure, and manageability of the EPA tasks (since no technology interface is involved). Students who perceive the analog EPA process as straightforward and not burdensome are more likely to develop a positive attitude toward it. For the digital EPA, by contrast, usefulness alone was sufficient to explain attitude, which may indicate that students either took ease of use for granted in a hypothetical digital system or did not consider it salient without actual experience of the interface.

Our findings also reinforce theoretical assumptions about the mediating role of attitude in technology acceptance. In both the analog and digital models, behavioral intention was strongly predicted by attitude, whereas any direct effects of perceived usefulness on intention became negligible once attitude was accounted for. Interventions to foster acceptance should therefore not only demonstrate the usefulness of a digital EPA (eg, by showcasing how it improves learning outcomes or efficiency), but also aim to cultivate positive attitudes, for example, by providing early positive experiences, highlighting success stories or endorsements from peers, and linking the tool to students' professional identity and values.

Strengths and Limitations

This study offers several strengths. It represents one of the first empirical investigations of medical students' acceptance of digital EPAs, applying the well-established TAM in this context. We obtained satisfactory reliability for all TAM scales, and by using both regression analyses and path modeling, we provided converging evidence for the robustness of the findings. Additionally, by having students consider both a traditional analog EPA and a hypothetical digital EPA, the study provides novel insights into the similarities and differences in what drives acceptance across these formats. This within-subject comparison strengthens our conclusions about the unique contribution of ease of use in the analog format and the primary importance of usefulness in the digital format.

Nevertheless, certain limitations must be acknowledged. First, the sample was drawn from a single medical school, which may limit the generalizability of the results. Cultural, institutional, and curricular factors can influence technology acceptance. For example, the degree to which EPAs and simulation-based learning are already embedded in a curriculum, and more generally, the availability of digital infrastructure as well as institutional attitudes toward educational innovation may shape students' baseline expectations and acceptance of digital EPAs. Acceptance patterns observed at LMU Munich may therefore differ from those at institutions with less exposure to competency-based curricula or fewer simulation resources. On the other hand, focusing on one institution provided a relatively

XSL•FO
RenderX

homogeneous context (all students shared a similar curriculum and culture), which reduced extraneous variance and allowed a clearer interpretation of the TAM relationships. Future research should replicate these findings in larger and more diverse multicenter cohorts to test their robustness across different educational settings.

Second, the study relied on self-report questionnaires for both predictors and outcomes. While surveys are indispensable for capturing perceptions and intentions (and are the standard approach in early-stage TAM research), they do not directly measure actual usage behavior. We cannot be certain that what students *say* they intend to do would translate into real adoption of a digital EPA if one were available. That said, self-report measures of TAM constructs have been widely validated and are valuable for establishing a baseline of acceptance before a technology is fully implemented. Future studies should complement self-reported acceptance data with objective usage analytics and measures of learning outcomes once digital EPA platforms are introduced. This would help determine whether high perceived acceptance indeed leads to sustained engagement and educational benefits.

Third, the "digital EPA" in this study was presented only conceptually rather than as a fully operational system. Students read a description and were asked to imagine using the digital EPA, instead of interacting with a real software application. This design choice was a practical necessity (since no such software existed yet) and allowed us to gage baseline acceptance before investing in development. However, it likely reduced the salience of some usability factors and may have emphasized perceived usefulness. Accordingly, our findings pertain primarily to acceptance of the concept of a digital EPA. In addition, the conceptual and relatively constrained nature of the model, together with the modest sample size, may have contributed to the very high global fit indices observed in the path models. Such near-perfect fit values are not uncommon in simple, theory-driven models and should therefore be interpreted with caution, with greater emphasis placed on the consistency and plausibility of the estimated path coefficients. The relatively small sample size further increases the risk that effect sizes and correlations may be overestimated or appear more stable than they would in larger samples. It is therefore possible that some statistically significant associations observed in this study would be attenuated or no longer reach significance in larger, more heterogeneous cohorts. Replication with substantially larger samples is necessary to assess the robustness of the observed relationships. Once a functional digital EPA platform is developed, future research should assess actual user experience, including any usability hurdles, and examine how these real-world factors might alter the acceptance model. It is possible that ease of use will play a larger role when students can directly experience the interface, features, and potential frustrations of a digital EPA tool.

Finally, the cross-sectional design of this study limits our ability to draw causal conclusions or to observe changes over time. We captured students' perceptions at a single point after a brief exposure to the concept. It remains unknown how acceptance might evolve with more prolonged or repeated exposure to digital EPAs, or after students have the opportunity to use such tools in practice. Longitudinal studies would be valuable to see if initial attitudes and intentions lead to actual usage and if they change through increased familiarity. Additionally, experimental designs (for instance, providing some students with access to a prototype digital EPA and comparing outcomes with a control group) could strengthen causal inference about the impact of digital EPAs on learning and acceptance. Despite these limitations, our cross-sectional survey provides an efficient initial evidence base that can inform future research directions and offers actionable guidance for the implementation of digital EPAs in educational practice. For educators, digital EPAs should be explicitly aligned with curricular learning objectives and assessment frameworks, so that students can clearly recognize how engagement contributes to competency progression and entrustment decisions. For developers, perceived usefulness may be strengthened by providing targeted, actionable feedback, visualizing performance data in relation to EPA levels, and ensuring that digital EPAs address authentic clinical tasks that students perceive as directly relevant to their future practice.

Conclusions

This study provides initial empirical evidence that medical students are open to the use of digital EPAs, with acceptance primarily driven by perceived usefulness and mediated by attitude. Ease of use, while important for acceptance of analog EPAs, did not emerge as a decisive factor for the digital EPA in this conceptual, hypothetical evaluation, likely reflecting the absence of interaction with a functional system. The slightly more favorable attitudes toward the analog EPA likely reflect familiarity with traditional learning tools rather than any fundamental resistance to digitalization.

By showing that usefulness and positive attitudes are central levers for shaping students' intent to use digital EPAs, our findings have practical implications for both developers and educators. For developers of digital EPA platforms, the message is clear: focus on features that provide tangible educational benefits and ensure the tool is reliable and adds clear value to student learning. For educators and curriculum designers, the task is to build enthusiasm and buy-in for integrating digital EPAs in a way that aligns with valued learning activities and demonstrating success stories to cultivate positive attitudes. With careful attention to demonstrating educational value and maintaining a student-centered user experience, digital EPAs have the potential to enrich competency-based medical education and better prepare students for their clinical responsibilities.

Acknowledgments

In preparing this manuscript, generative artificial intelligence (ChatGPT-4, OpenAI, February 2024 version) was used only for language refinement and stylistic editing. No content creation, data analysis, or scientific writing was performed by the AI. The

XSL•FO
RenderX

authors confirm that all intellectual content is the product of their own work, and they take full responsibility for the accuracy and integrity of the content.

Funding

MD was supported by a Promotionsstipendium (doctoral scholarship) from the Medical Faculty of LMU Munich during the conduct of this research.

Data Availability

All data generated by this study will be made available in anonymized form on the Open Science Framework (OSF) repository after completion of the study.

Authors' Contributions

MD was responsible for creating the study materials and visuals and drafted the initial manuscript (including translating materials and results into English). CR, MJW, and MS provided substantial contributions through critical revision and editing of the manuscript. MS also served as the project supervisor and oversaw the research administration. All authors read and approved the final manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Analog and digital EPA descriptions (English).
[PDF File, 66 KB - mededu_v12i1e87605_app1.pdf]

Multimedia Appendix 2
Analoge und digitale EPA-Beschreibung (Deutsch).
[PDF File, 87 KB - mededu_v12i1e87605_app2.pdf]

Multimedia Appendix 3
Hypothesen und Fragenkatalog der TAM-Umfrage (Deutsch).
[PDF File, 112 KB - mededu_v12i1e87605_app3.pdf]

Multimedia Appendix 4
Hypotheses and questionnaire items used in the TAM survey (English).
[PDF File, 109 KB - mededu_v12i1e87605_app4.pdf]

References

1. Hamed O, Verstegen D, Alsheikh G, et al. International Delphi study on developing entrustable professional activities (EPAs) for master's learners in health professions education. BMC Med Educ 2024 Nov 26;24(1):1362. [doi: 10.1186/s12909-024-06377-z] [Medline: 39587515]
2. Ten Cate O, Taylor DR. The recommended description of an entrustable professional activity: AMEE Guide No. 140. Med Teach 2021 Oct;43(10):1106-1114. [doi: 10.1080/0142159X.2020.1838465] [Medline: 33167763]
3. Cangiarella J, Gillespie C, Shea JA, Morrison G, Abramson SB. Accelerating medical education: a survey of deans and program directors. Med Educ Online 2016;21:31794. [doi: 10.3402/meo.v21.31794] [Medline: 27301381]
4. Holzhausen Y, Maaz A, Renz A, Bosch J, Peters H. Development of entrustable professional activities for entry into residency at the Charité Berlin. GMS J Med Educ 2019;36(1):Doc5. [doi: 10.3205/zma001213] [Medline: 30828605]
5. Smetana LK, Bell RL. Computer simulations to support science instruction and learning: a critical review of the literature. Int J Sci Educ 2012 Jun;34(9):1337-1370. [doi: 10.1080/09500693.2011.605182]
6. Rafiq A, Sethi A. Entrustable professional activities for bedside clinical teachers. BMC Med Educ 2024 Aug 18;24(1):887. [doi: 10.1186/s12909-024-05876-3] [Medline: 39155387]
7. Chernikova O, Heitzmann N, Stadler M, Holzberger D, Seidel T, Fischer F. Simulation-based learning in higher education: a meta-analysis. Rev Educ Res 2020 Aug;90(4):499-541. [doi: 10.3102/0034654320933544]
8. Mao RQ, Lan L, Kay J, et al. Immersive virtual reality for surgical training: a systematic review. J Surg Res 2021 Dec;268:40-58. [doi: 10.1016/j.jss.2021.06.045] [Medline: 34284320]
9. Roche E. Advances in undergraduate and postgraduate education- bedside, simulation and E-learning. Global Pediatrics 2024 Sep;9:100188. [doi: 10.1016/j.gpeds.2024.100188]

XSL•FO
RenderX

10. Li K, Li Q, Wang L, Wang X, Ding X, Liu B. Virtual simulation experiments in medical education: technology acceptance, learning outcomes, and motivational impacts. BMC Med Educ 2025 Oct 21;25(1):1462. [doi: 10.1186/s12909-025-08025-6] [Medline: 41121033]
11. Merolli M, Ahmed O, McCreesh K, Remedios L, Butler-Henderson K. Are physiotherapists expected to be competent in digital health practice? Meta-synthesis of international physiotherapy practice competency standards. Physiother Theory Pract 2024 Dec;40(12):2988-2999. [doi: 10.1080/09593985.2023.2299202] [Medline: 38214509]
12. Marangunić N, Granić A. Technology acceptance model: a literature review from 1986 to 2013. Univ Access Inf Soc 2015 Mar;14(1):81-95. [doi: 10.1007/s10209-014-0348-1]
13. Lee JWY, Tan JY, Bello F. Technology acceptance model in medical education: systematic review. JMIR Med Educ 2025 Jul 16;11:e67873. [doi: 10.2196/67873] [Medline: 40668663]
14. Holden RJ, Karsh BT. The technology acceptance model: its past and its future in health care. J Biomed Inform 2010 Feb;43(1):159-172. [doi: 10.1016/j.jbi.2009.07.002] [Medline: 19615467]
15. Or C. Thirty-five years of the technology acceptance model: insights from meta-analytic structural equation modelling. OTESSAJournal 2024;4(3):1-26. [doi: 10.18357/otessaj.2024.4.3.66]
16. Merolli M, Gray K. Demonstrating digital health clinical competence in practice: a method for developing entrustable professional activities. Stud Health Technol Inform 2024 Jan 25;310:1221-1225. [doi: 10.3233/SHTI231159] [Medline: 38270009]
17. Domann M, Richters C, Stadler M. Student acceptance of digital entrustable professional activities: a study protocol. Open Science Framework. URL: https://osf.io/yzsb9/?view_only=88ce4b2f45bc4056a8ed7e9d6c5dddf2 [accessed 2026-04-15]
18. Abdul Ghani MT, Hamzah M, Ramli S, et al. A questionnaire-based approach on technology acceptance model for mobile digital game-based learning. J Glob Bus Soc Entrepreneursh 2019;5:11-21 [FREE Full text]
19. SoSci Survey – the solution for professional online questionnaires. SoSci Survey (Version 3501). 2024. URL: https://www.soscisurvey.de [accessed 2026-04-15]
20. The jamovi project. Jamovi. 2025. URL: https://www.jamovi.org [accessed 2026-04-15]
21. Davis FD. Perceived usefulness, perceived ease of use, and user acceptance of information technology. MIS Q 1989 Sep 1;13(3):319-340. [doi: 10.2307/249008]
22. Student acceptance of digital entrustable professional activities: a study protocol. Open Science Framework. URL: https://osf.io/yzsb9/?view_only=88ce4b2f45bc4056a8ed7e9d6c5dddf2 [accessed 2026-04-15]
23. Faul F, Erdfelder E, Lang AG, Buchner A. G*Power 3: a flexible statistical power analysis program for the social, behavioral, and biomedical sciences. Behav Res Methods 2007 May;39(2):175-191. [doi: 10.3758/bf03193146] [Medline: 17695343]

Abbreviations

AT: attitude (toward using the technology)
BI: behavioral intention
CFI: comparative fit index
CHERRIES: Checklist for Reporting Results of Internet E-Surveys
EPA: entrustable professional activity
LMU: Ludwig Maximilian University
OSF: Open Science Framework
PEU: perceived ease of use
PU: perceived usefulness
RMSEA: root mean square error of approximation
TAM: Technology Acceptance Model

Edited by P Kanzow; submitted 11.Nov.2025; peer-reviewed by K Butler-Henderson, YF Ahmed; revised version received 09.Mar.2026; accepted 12.Mar.2026; published 04.May.2026.

Please cite as:
Domann M, Richters C, Witti MJ, Stadler M
Medical Students' Acceptance of Digital Entrustable Professional Activities: Results of a Cohort Study
JMIR Med Educ 2026;12:e87605
URL: https://mededu.jmir.org/2026/1/e87605
doi:10.2196/87605

© Maximilian Domann, Constanze Richters, Matthias J Witti, Matthias Stadler. Originally published in JMIR Medical Education (https://mededu.jmir.org), 4.May.2026. This is an open-access article distributed under the terms of the Creative Commons

XSL•FO
RenderX

Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Education, is properly cited. The complete bibliographic information, a link to the original publication on https://mededu.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Olfactory Cues and Patient Safety Behavior in Virtual Reality Medical Simulation: Controlled Trial

Anna Junga[1], MD, Dr med; Pascal Kockwelp[2], BSc, MSc; Jennifer Dabel[1], BSc, MSc, PhD; Sönke Helmut Scherzer[1], BSc, MSc, PhD; Benjamin Risse[2], Dipl. Inf, PhD; Markus Holling[3], MD, Dr med, MHBA; Bernhard Marschall[1], MD, Dr med; Hendrik Friederichs[4], MA, MME, MD, Dr med

[1]Institute of Education and Student Affairs, University of Münster, Albert-Schweitzer-Campus 1A, Building A6, Münster, North Rhine-Westphalia, Germany

[2]Institute for Geoinformatics & Faculty of Mathematics and Computer Science, University of Münster, Münster, North Rhine-Westphalia, Germany

[3]Department of Neurosurgery, University Hospital Münster, Münster, North Rhine-Westphalia, Germany

[4]Research Group Medical Education, Medical School OWL, Bielefeld University, Bielefeld, North Rhine-Westphalia, Germany

Corresponding Author:
Anna Junga, MD, Dr med
Institute of Education and Student Affairs, University of Münster, Albert-Schweitzer-Campus 1A, Building A6, Münster, North Rhine-Westphalia, Germany

Abstract

Background: Simulation-based medical education is essential for improving patient safety. In virtual reality (VR)–based simulation, immersion is primarily generated through visual and auditory cues, while other sensory modalities are typically absent. This sensory limitation may reduce the emergence of authentic safety-relevant behaviors. Olfaction plays an important role in clinical reasoning, risk perception, and self-protective behavior and is closely linked to memory and emotion. Although olfactory cues have been shown to influence hand hygiene behavior in real or simulated environments, their targeted integration into fully immersive VR-based medical simulation has not been systematically examined.

Objective: This study aimed to investigate whether adding a real olfactory cue (disinfectant scent) to a fully virtual clinical simulation increases patient safety–relevant behavior, specifically hand hygiene compliance (hand disinfection and glove usage).

Methods: In a controlled study at the University of Münster (during the winter term 2025-2026), 95 medical students participated in a VR-based clinical simulation. Study rooms were preassigned to either an olfactory intervention or a control condition, and participants selected their room without knowledge of the assigned condition (quasi-random allocation by room). Hand hygiene and glove use were automatically tracked as outcomes. Odds ratios (ORs) were calculated to assess the effect of the intervention on these behaviors.

Results: The olfactory intervention nearly tripled the odds of hand disinfection (8/43, 18.6% vs 20/52, 38.5%; an absolute risk increase of 19.9 percentage points; OR=2.73, 95% CI 1.09 - 7.40; *P*=.04; number needed to treat=5), while no significant difference was observed for glove use (OR=1.44, 95% CI 0.62 - 3.45; *P*=.40).

Conclusions: The integration of a real olfactory cue into a fully immersive VR medical simulation significantly increased hand disinfection behavior, particularly after patient contact, but did not affect glove use. These findings suggest that olfactory augmentation can selectively reinforce safety-relevant behaviors in digital training environments. Incorporating real-world sensory cues into VR may represent a simple yet effective design strategy to enhance behavioral authenticity and patient safety outcomes in simulation-based medical education.

Trial Registration: German Clinical Trials Register DRKS00039472; https://drks.de/search/en/trial/DRKS00039472

(JMIR Med Educ 2026;12:e93148) doi:10.2196/93148

KEYWORDS

virtual reality; undergraduate medical education; hand hygiene; olfactory; patient safety; simulation; immersion; brain death

Introduction

Simulation in medical education has become an essential strategy for improving health care quality and, most importantly, patient safety. Modern simulation modalities—ranging from manikins and standardized patients to fully immersive virtual reality (VR) environments—enable learners to practice technical skills, develop situational awareness, and recognize safety-relevant cues in a controlled, risk-free setting. Effective simulation is thus not only about skill acquisition but also about fostering behaviors that prevent harm to patients.

XSL•FO
RenderX

Across all simulation modalities, immersion—the subjective perception of a simulated environment as "real"—is critical for learner engagement and the transfer of knowledge to clinical practice. It is useful to distinguish immersion, the objective sensory and technical fidelity of the system, from presence, the subjective sense of being there, and from embodiment, the experience of the virtual body as one's own. These constructs are conceptually distinct but jointly shape how learners act in virtual environments. Immersion arises from the integration of visual, auditory, tactile, and contextual cues, with the surrounding environment often contributing substantially in traditional simulation centers. Purpose-built clinical simulation environments aim to replicate real-world contexts as faithfully as possible, integrating multiple sensory inputs to strengthen both realism and learner engagement [1,2]. Hybrid solutions, such as SimuScape [3], have further demonstrated that combining real interaction partners with projected environments can enhance learning while remaining economically feasible.

VR-based simulation, in contrast, creates immersion almost entirely through digitally generated environments, enabling learners to actively explore and interact with complex clinical situations rather than passively observe them. This interactivity is enabled by head-mounted displays (HMDs) and hand-tracking technologies. However, the sensory experience in VR is largely restricted to virtual visual and auditory cues, while tactile feedback remains limited and other sensory modalities are typically absent [4-6]. As a result, clinically relevant safety behaviors may not reliably emerge, even when scenarios are cognitively well-designed. Previous work by Junga et al [7,8] indicates that, in VR-based simulations, hand hygiene–related behavioral targets are not consistently achieved, presumably because the level of situational realism is insufficient to trigger safety-relevant actions.

This raises a central design challenge for VR-based medical education: How can digital simulations be augmented to elicit authentic, real-world patient safety behavior? In this context, olfactory cues are of particular interest. Unlike visual or auditory stimuli in VR, there is no widely used technology to render smell, so it must originate from the physical world. Olfaction, therefore, represents the only manipulatable sensory input in VR-based simulations that is entirely real. Introducing real smells, such as disinfectants, into a VR scenario may help learners cross the virtual threshold, experiencing the situation in a way that more closely resembles clinical practice and thereby increasing the likelihood of authentic patient safety behavior.

Evidence from cognitive psychology suggests that learning and recall are enhanced when the sensory context of learning closely resembles that of application, a phenomenon known as context-dependent memory [9]. In clinical practice, olfactory cues play a multifaceted role in patient safety and clinical decision-making. They may provide diagnostic information (eg, ketone odor in diabetic ketoacidosis or alcohol odor in unconscious patients), signal environmental or personal risk (eg, gas leaks or intoxication-associated aggression), and support rapid situational assessment. Olfaction is also closely linked to memory formation and emotional processing, making it a powerful but often implicit driver of behavior.

Despite this, olfactory input has rarely been integrated into VR-based medical education. For non-VR simulation, Birnbach et al [10] conducted a landmark randomized study in 2013 investigating the effect of a fresh citrus scent on hand hygiene compliance. In the control group (n=86), the compliance rate was 51%, whereas the scent group (n=79) achieved a significantly higher rate of 80% (P<.001). A subsequent study by King et al [11] confirmed these findings in a surgical intensive care unit. Exposure to a citrus scent resulted in improved hand hygiene compliance (46.9% vs 15.0%; P<.001), corresponding to an odds ratio (OR) of approximately 5.0. In contrast, Schmidtke et al [12] reported differing results. In a large crossover study conducted at hospital ward entrances, none of the priming interventions led to a consistent increase in hand hygiene behavior.

Notably, all previous studies investigating olfactory effects on hand hygiene were conducted in real-world or simulated environments. To date, the targeted integration of real olfactory cues into fully immersive VR-based medical simulation has not been systematically examined. Building on our previous findings showing that immersion levels and the visual representation of virtual hands can measurably influence hygiene behavior [7,8], the present study investigates whether adding real olfactory cues to an otherwise fully virtual environment can enhance hand hygiene practices, including prepatient and postpatient contact disinfection.

By conceptualizing olfactory augmentation as a design element in digital simulation rather than a contextual add-on, this study addresses a critical gap in VR-based medical education. Based on the heterogeneous evidence with positive findings from controlled simulation environments and negative findings from real-world settings, we hypothesize that exposure to real olfactory cues (disinfectant scent) during a VR simulation with defined patient contact increases hand hygiene behavior compared with a VR simulation without olfactory augmentation. As a second primary outcome, we examine whether the olfactory intervention also influences glove use.

Methods

Setting and Participants

The study was carried out as part of the "Brain Death/Death by Neurologic Criteria" (BD/DNC) examination course at the University of Münster during the winter term 2025-2026. The course is scheduled for the seventh semester of the medical curriculum. A VR-based digital simulation depicting a clinical scenario for BD/DNC examination served as the research environment [13-15]. The VR environment replicates a realistic intensive care unit in which participants can interact with a virtual patient and perform clinical examinations. The course is part of a block week focusing on organ donation and has been delivered in a VR-based format since 2021. Further information about the course structure and procedures can be found in Junga et al [13,16]. Every student received a technical and medical preparation video podcast to overcome knowledge and computer literacy inequalities.

XSL•FO
RenderX

Although participation in the course was mandatory, participation in the study was voluntary. Students who declined to participate did not experience any disadvantages. Students enrolled in the course were recruited and assigned to 1 of 2 conditions through quasi-random allocation by room. Rooms were preassigned to either a control condition ("NoSmell"), in which participants performed the standard brain death examination, or an intervention condition ("Smell"), in which the same simulation was augmented with an olfactory stimulus. Participants selected their room without knowledge of the assigned condition or about the assignment at all. Inclusion criteria included enrollment in medical school, course attendance, and willingness to participate in the study.

Ethical Considerations

The study was conducted in accordance with the Declaration of Helsinki and was approved by the responsible ethics committee (Ethik-Kommission Westfalen-Lippe; application 2022 - 736-b-S; approved on November 9, 2022). Participation in the study was voluntary and independent of course participation, completion, and assessment, and declining to participate entailed no disadvantage. Written informed consent was obtained from all participants prior to allocation through a digital survey form. All study data were collected and analyzed in pseudonymized form using 6-character study codes, and no directly identifying information was retained at the analysis stage. No financial or nonfinancial compensation was provided for study participation.

Sample Size Calculation

The required sample size was determined a priori using G*Power (version 3.1; Heinrich Heine University Düsseldorf) [17]. For the primary dichotomous outcomes (hand disinfection and glove use, each coded as true or false), a medium effect size (w=0.30) [18] was assumed, and a χ^2_1 test for a 2×2 contingency table was specified. With an α level of .05 and a statistical power of 80%, the required total sample size was 88. To account for potential dropouts, a minimum of 90 participants was targeted for recruitment.

Outcome Objectives and Necessary Adjustments

In accordance with the recommendations of the Commission for Hospital Hygiene and Infection Prevention (KRINKO) at the Robert Koch Institute, hand disinfection is mandatory before patient contact or contact with the patient's environment (eg, bed, side table, and catheter bag), before aseptic procedures, and after contact with potentially contaminated material or the patient and their surroundings. The use of medical gloves is recommended for "protection of the wearer against contamination with blood, secretions, and excretions, including pathogens, and indirectly to interrupt chains of infection" [19].

For the purposes of this study, the functionalities for using gloves and a hand disinfectant dispenser were specifically implemented and enabled within the VR environment (Figure 1A, B). To use the dispenser, at least one hand had to be placed underneath it, and a hissing sound was generated to confirm the interaction. To use the gloves, the participant's hands had to touch the glove box, after which the gloves were attached automatically.

The software was further modified so that, in addition to interactions associated with a potential risk of exposure to bodily fluids (eg, gag reflex, tracheal reflex, and caloric-induced nystagmus), the examination procedure deliberately involved contact with a simulated extensive skin abrasion in the facial area (Figure 1C, D). This design rendered glove use mandatory for hygiene, in addition to hand disinfection, which is strictly required in intensive care settings.

As demonstrated in Junga et al [8], the representation of the hands significantly affects glove usage. Hence, the application was modified to provide the avatar with a bare hand model, as opposed to the previous one, which was equipped with gloves by default.

The objectives stated in Table 1 were identified in accordance with the recommendations for hand hygiene issued by the Robert Koch Institute [19]. Primary and explorative outcomes were analyzed separately. All actions were automatically tracked and documented by the software. Before and after the course, participants completed a survey on demographics and prior VR experience. No item assessed scent perception or other manipulation-check questions.

All end points were operationalized as binary variables. Hand disinfection was considered to have been performed if the participant activated the virtual disinfectant dispenser with at least one hand. Glove use was recorded if the participant interacted with the glove box using at least one hand, thereby triggering the glove application within the virtual environment.

XSL•FO
RenderX

Figure 1. In-game screenshots of the implemented hygiene stations and clinical application. (A) Activation of the disinfectant dispenser. (B) Interaction with the glove box to equip protective gear. (C, D) Subsequent patient examination (head stabilization and corneal reflex test) performed with gloved hands.

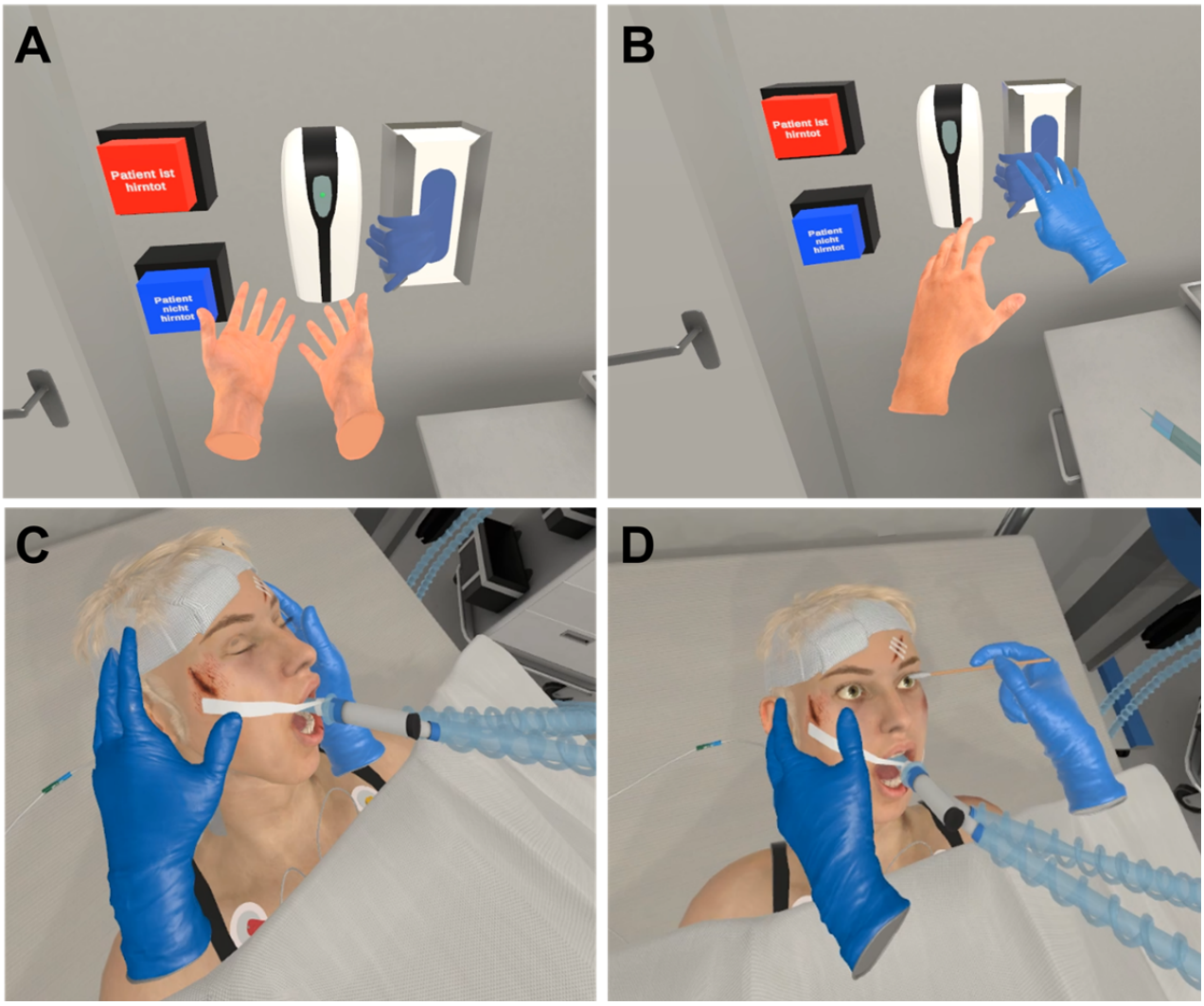


Table . Primary and explorative objectives used in this study.

Task-implicit objectives (hand hygiene)	Measure
Primary outcome	
Hand disinfection was applied overall (at least one hand)	True or false
Gloves were put on overall (at least one hand)	True or false
Explorative outcome	
Gloves were put on before first patient contact (at least one hand)	True or false
Hand disinfection was applied before first direct patient contact (at least one hand)	True or false
Hand disinfection was applied after last patient contact (at least one hand)	True or false

Olfactory Intervention Procedure

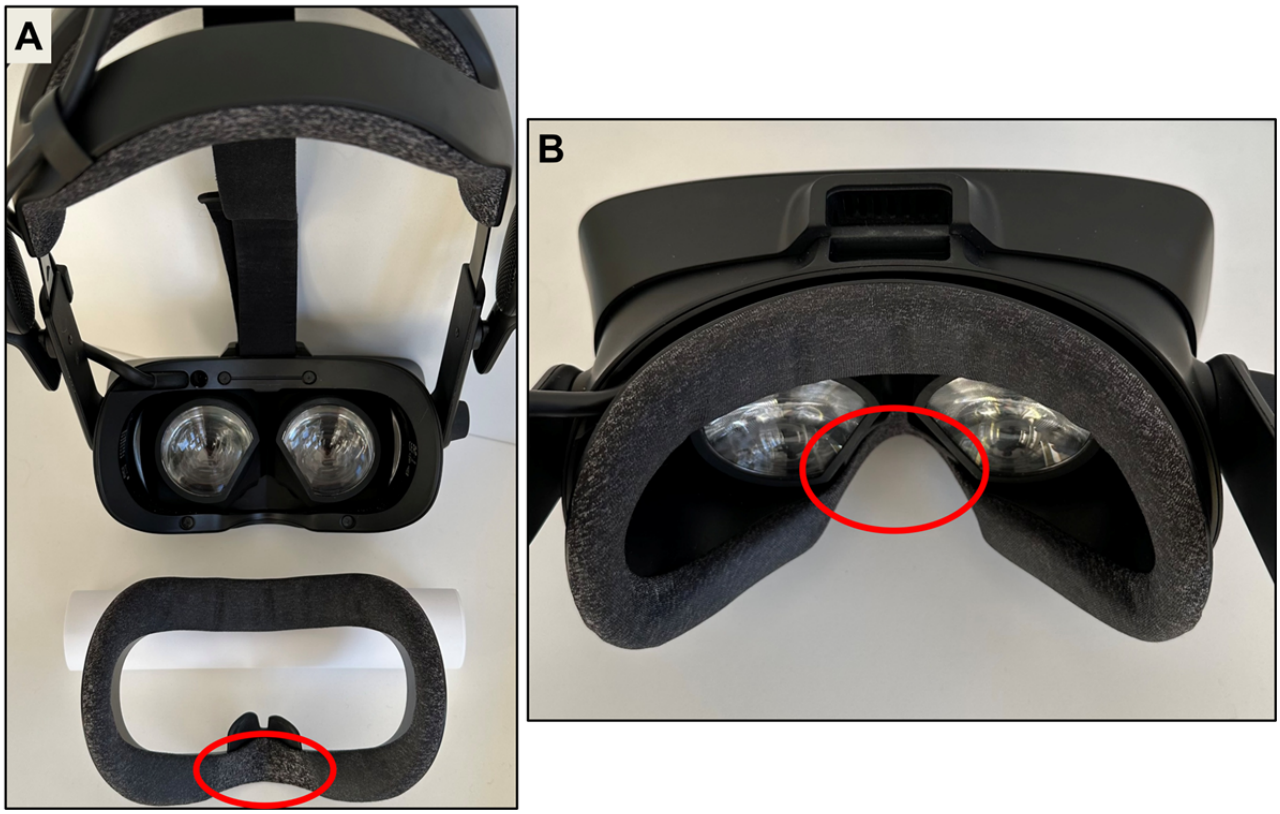
In 3 of the 6 rooms, an olfactory stimulus was applied by placing 5 drops (approximately 0.25 mL) of an alcohol-based hand disinfectant (desmanol care, Schülke & Mayr GmbH) onto the facial interface of the VR headset, specifically at the upper edge of the nasal cutout (Figure 2). This application was repeated immediately before each new rotation (<3 min) to avoid differences in smell intensity.

Due to the use of additional hygienic face masks, there was no direct skin contact with the students' faces, but the fluid was placed less than 5 mm from the bridge of the nose without any visible manipulation of the HMD. Participants were blinded to both the presence and the purpose of the olfactory intervention.

XSL•FO
RenderX

The use of hand disinfectants and gloves in the virtual environment was automatically tracked by the system, and the data were exported after course completion.

Figure 2. Inside views of the Valve Index virtual reality headset (view from above) showing the area of applied disinfectant on the facial interface (red). (A) Headset with facial interface removed for better visibility. (B) Headset with facial interface attached.



Materials

The VR scenario was delivered using the *BD/DNC Examination* software (version December 2025; University of Münster) [13-15] on a Valve Index HMD with Knuckles controllers (Valve Corporation). Data were collected using LimeSurvey and analyzed with R (version 4.5.2; R Foundation for Statistical Computing) using the tidyverse [20] and broom [21] packages.

Statistical Methods

Primary Analysis

The effect of the olfactory intervention on the 2 primary end points (hand disinfection and glove use) was evaluated using logistic regression on the intention-to-treat sample of all 95 allocated participants, with intervention condition entered as the sole predictor. Parameter estimation was performed using the maximum likelihood method. Results are reported as ORs with 95% CIs. In addition, standardized regression coefficients are reported, obtained by fitting the model to a standardized version of the dataset. CIs were obtained by the profile-likelihood method. Model fit was quantified using Tjur R^2, a pseudo-coefficient of determination appropriate for logistic regression.

Sensitivity Analyses

Results were verified using Fisher exact and Pearson chi-square tests. Effect size was calculated as follows: Cohen $w = \sqrt{(\chi^2/N)}$.

Exploratory End Points

Given low expected cell counts (<5) for some end points, Fisher exact test was used throughout to calculate *P* values, whereas the corresponding ORs and 95% CIs were obtained from logistic regression (with a Haldane-Anscombe correction for an empty cell in initial glove use). Exploratory end points were reported with unadjusted *P* values and should therefore be interpreted with appropriate caution.

The study was reported in accordance with the CONSORT-EHEALTH (Consolidated Standards of Reporting Trials of Electronic and Mobile Health Applications and Online Telehealth) checklist, version 1.6.1. The completed checklist is provided in Checklist 1.

Results

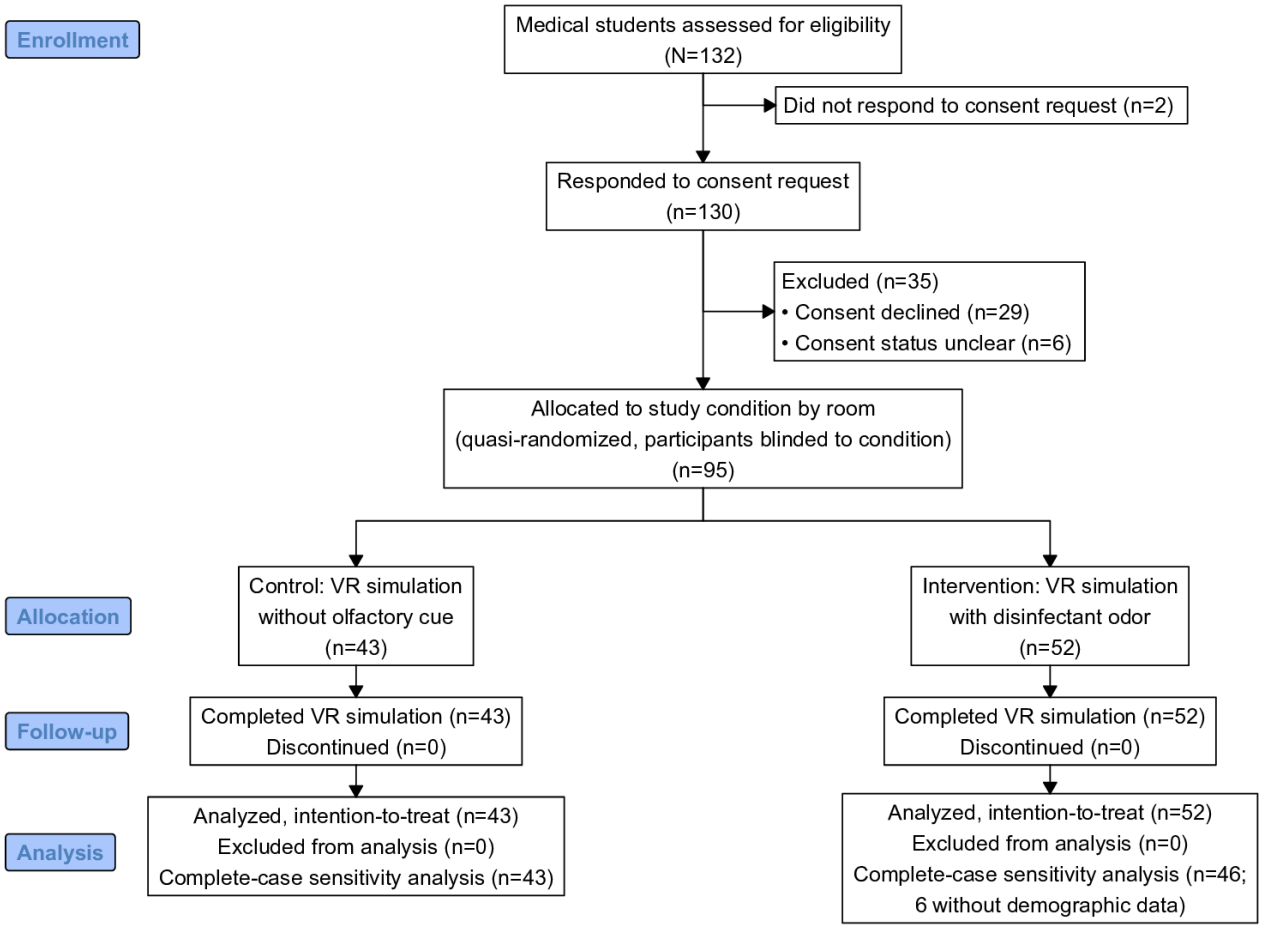
Inclusion Criteria

In the winter term 2025-2026, 132 individuals were assessed for eligibility, of which 130 responded to the consent request. Of these, 35 were excluded: 29 refused consent, and the consent status of 6 remained unclear. The remaining 95 participants were allocated to the control group (n=43) or the intervention group (n=52). All 95 allocated participants completed the VR simulation and were analyzed according to their assigned condition (intention-to-treat analysis). Six participants in the intervention group did not provide demographic data and were therefore excluded from the complete-case sensitivity analyses only (complete-case: n=89, comprising 43 control and 46

XSL•FO
RenderX

intervention participants). The minor imbalance in group sizes reflects the quasi-random allocation by room, as final numbers were determined only after all sessions had been completed. The flow of participants through the study is shown in Figure 3.

Figure 3. Participant flow through the study. Allocation to the intervention (olfactory cue) or control condition was performed at room level (quasi-random allocation); participants were blinded to their assigned condition. All allocated participants completed the simulation and were included in the intention-to-treat analysis. VR: virtual reality.



Participant Characteristics

As shown in Table 2, participants had a mean age of 23.6 (SD 2.6; range 21‐35) years. Age did not differ significantly between conditions (24.0 vs 23.2 y; $t_{67.8}$=1.51; *P*=.14). Female students (59/89, 66.3%) were overrepresented in the sample, consistent with the sex composition of medical students in Germany [22]. However, the sex distribution did not differ significantly between groups (32/43, 74.4% vs 27/46, 58.7% female participants; χ^2_1=2.46; *P*=.12).

In the subsequent survey about prior HMD experience, only 3 participants reported owning a personal VR headset, all of whom were male. However, the majority of students (52/89, 58.4%; regularly+once or rarely) had previously used a VR headset at least once (for further information, see Table 3).

XSL•FO
RenderX

Table . Baseline characteristics (N=89).

Characteristic	Overall (N=89)	Control (n=43)	Intervention (n=46)	*P* value
Age (y)				.14[a]
Mean (SD)	23.6 (2.6)	24.0 (3.1)	23.2 (1.9)	
Median (IQR)	23 (22-24)	23 (22–25)	23 (22–23)	
Range	21 - 35	21 - 35	21 - 29	
Sex, n (%)				.12[b]
Female	59 (66.3)	32 (74.4)	27 (58.7)	
Male	30 (33.7)	11 (25.6)	19 (41.3)	

[a]Independent-samples *t* test (2-tailed, Welch correction for unequal variances).

[b]Pearson chi-square test (2-tailed, *df*=1).

Table . Characteristics of participants, divided into total (N=89), intervention ("Smell", n=46), and control group ("NoSmell", n=43).

Characteristic	Total (N=89), n (%)	Smell (n=46), n (%)	NoSmell (n=43), n (%)
Owns HMD[a]	3 (3.37)	1 (2.17)	2 (4.65)
HMD experience			
Never used	37 (41.57)	19 (41.30)	18 (41.86)
Once or rarely	50 (56.18)	27 (58.70)	23 (53.49)
Regularly	2 (2.24)	0 (0)	2 (4.65)

[a]HMD: head-mounted display.

Primary Outcome

Hand Disinfection

Hand disinfection was performed by 8 of 43 participants (18.6%) in the control group compared with 20 of 52 participants (38.5%) in the intervention group (Table 4; Figure 4). To predict hand disinfection, a logistic regression model using maximum likelihood estimation was fitted (formula: hand disinfection ~ intervention). The explanatory power of the model was weak (Tjur R^2=0.05), which is not uncommon given a single predictor and dichotomous behavioral data. A single environmental cue is not expected to account for the full variance of hygiene behavior, which is shaped by knowledge, attitudes, habits, time pressure, and contextual factors. The relevant question is whether the cue exerts an incremental, practically meaningful effect. The model intercept, corresponding to the control condition (intervention is 0), was −1.48 (*P*<.001). The effect of the intervention was statistically significant (β=1.01, SE 0.49; *P*=.04), corresponding to an OR of 2.73 (95% CI 1.09-7.40). This indicates that the intervention approximately tripled the odds of hand disinfection (Table 5).

Table . Results of the smell intervention on students' hygiene behavior from the primary and explorative analyses (N=95).

Implicit objective	NoSmell (n=43), n (%)	Smell (n=52), n (%)	OR (95% CI)	*P* value
Primary outcomes				
Hand disinfection overall	8 (18.6)	20 (38.5)	2.73 (1.09 - 7.40)[a]	.04
Use of gloves overall	13 (30.2)	20 (38.5)	1.44 (0.62 - 3.45)[a]	.40
Exploratory outcomes				
Initial disinfection	3 (7.0)	2 (3.8)	0.53 (0.07 - 3.37)[b]	.66
Initial use of gloves	2 (4.7)	0 (0.0)	0.16 (0.01 - 3.38)[b]	.20
Final disinfection	1 (2.3)	10 (19.2)	10.00 (1.80 - 187.60)[b]	.01

[a]Logistic regression.

[b]For exploratory outcomes, the OR and 95% CI were obtained from logistic regression (with a Haldane-Anscombe correction for an empty cell in initial glove use outcome), and the *P* value was obtained from Fisher exact test.

XSL•FO
RenderX

Figure 4. Forest plot (effect of olfactory intervention) for primary (blue) and exploratory (gray) end points.

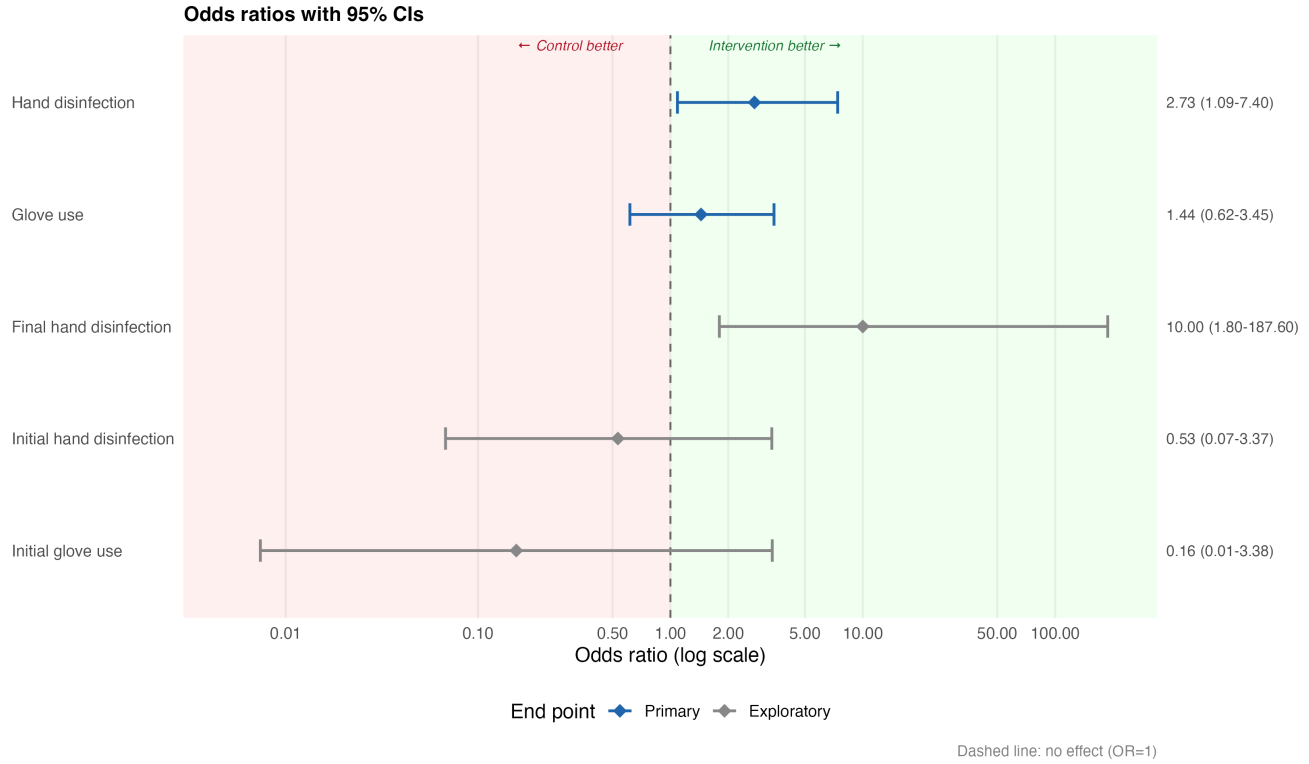


Table . Logistic regression models for primary outcomes[a].

Outcome and parameter	β (SE)	OR (95% CI)	*P* value
Hand disinfection			
Intercept	−1.48 (0.39)	—[b]	<.001
Intervention	1.01 (0.49)	2.73 (1.09-7.40)	.04
Glove use			
Intercept	−0.84 (0.33)	—	.01
Intervention	0.37 (0.44)	1.44 (0.62-3.45)	.40

[a]Model fit (hand disinfection): Tjur R^2=0.05; AIC=114.6. Model fit (glove use): Tjur R^2=0.01; AIC=126.0. AIC refers to Akaike information criterion.
[b]Not applicable.

Convergent tests supported this finding: Fisher exact test (*P*=.04) and Pearson chi-square test (χ^2_1=4.47; *P*=.04). The observed effect size was small to medium (Cohen *w*=0.22).

In absolute terms, the intervention increased the probability of hand disinfection by 19.9 percentage points, corresponding to a number needed to treat of 5. In practical terms, this means that for every 5 students exposed to the olfactory intervention, one additional individual would perform hand disinfection who otherwise would not have done so under standard conditions.

Use of Gloves

Gloves were used by 13 of 43 participants (30.2%) in the control group, compared with 20 of 52 participants (38.5%) in the intervention group. The logistic regression model predicting glove use (formula: glove use ~ intervention) showed weak explanatory power (Tjur R^2=0.01). The effect of the intervention was not statistically significant (β=0.37, SE 0.44; *P*=.40), corresponding to an OR of 1.44 (95% CI 0.62-3.45). Fisher exact test confirmed the absence of a statistically significant effect (*P*=.52; Table 4; Figure 4). The absolute risk difference was 8.2 percentage points in favor of the intervention group—an effect that could not be distinguished from chance, given the sample size.

Sensitivity Analyses

The intervention effect on hand disinfection was robust across all prespecified sensitivity analyses. In the complete-case sample (n=89), the unadjusted OR was 2.81 (95% CI 1.09 - 7.75; *P*=.04). Adjustment for sex did not materially change the estimate (adjusted OR=2.58, 95% CI 0.99 - 7.20; *P*=.06). The marginally attenuated significance reflects reduced power in the smaller subsample rather than substantive confounding, given the nonsignificant baseline sex imbalance and the stable point estimate. Fisher exact test (*P*=.04) and the chi-square test (χ^2_1=4.47; *P*=.04) further supported the finding. For glove use, no sensitivity analysis indicated a statistically significant effect. These convergent results indicate that the primary conclusion

XSL•FO
RenderX

is not driven by the choice of analytic sample or by adjustment for baseline covariates.

Exploratory End Points

For final hand disinfection, defined as disinfection immediately when leaving the patient environment, a pronounced difference between groups was observed. Only 1 of 43 control participants (2.3%) exhibited this behavior, compared with 10 of 52 participants (19.2%) in the intervention group. Fisher exact test yielded an OR of 9.82 (95% CI 1.29-443.13; *P*=.01). The wide CI reflects the low event rate in the control group and warrants cautious interpretation of this exploratory finding (Figure 3). No statistically significant differences between groups were observed for initial hand disinfection (3/43, 7.0% vs 2/52, 3.8%; OR=0.53, 95% CI 0.07 - 3.37; *P*=.66) or for initial glove use (2/43, 4.7% vs 0/52, 0.0%; OR=0.16, 95% CI 0.01 - 3.38; *P*=.20).

Summary of Findings

Table 4 and Figure 4 present the complete results for all end points, and further information on the analysis is provided in Table 5. The olfactory intervention significantly increased hand disinfection (the first primary end point), whereas glove use (the second primary end point) did not show a statistically significant difference between groups. In the exploratory analyses, the effect on final hand disinfection was particularly pronounced. However, this finding should be interpreted with caution given its exploratory nature and the low event rates observed.

Discussion

Impact of Smell on Hand Hygiene and Glove Use

The primary hypothesis, that olfactory cues would increase hand hygiene behavior, was partially supported. Hand disinfection increased significantly under olfactory exposure, whereas glove use showed no significant effect. Furthermore, exposure to the scent of disinfectants increased postcontact hand disinfection, whereas no effects were observed for initial hand disinfection or initial glove use. Because these behavioral changes were observed within a simulated environment, they correspond to Kirkpatrick level 3 (behavior) within simulation and serve as a proxy for, rather than direct evidence of, transfer to real-world clinical practice.

To contextualize these findings, it is important to note that not all hand hygiene behaviors are triggered by the same cues. While initial hand disinfection, which directly targets patient safety, represents a highly trained and routinized behavior, which is done straight as the first action while entering the room, the decision to use gloves is typically reassessed in each situation. This assessment depends on perceived exposure risks, such as contact with bodily fluids—either overt (eg, wound exudate) or subtle (eg, sweat)—as well as the need to protect one's own skin during wet procedures. These evaluations are guided by sensory impressions.

Our findings suggest that positively valenced cues, such as the clean scent of disinfectants, may reinforce reactive behaviors (eg, final hand disinfection) rather than initiating preventive routines, whereas other behaviors, such as glove use or initial hand hygiene, may require more salient or negatively valenced cues to be affected. Beyond this descriptive pattern, the dissociation between hand disinfection and glove use is itself theoretically informative. The scent of disinfectant carries a strong, repeated real-world association with the act of hand disinfection, whereas no comparable associative link exists for glove use. This selective effect suggests that the olfactory cue operates through specific stimulus-response priming rather than a generalized activation of patient safety behavior, implying that effective sensory augmentation may need to match cues to target behaviors based on learned associations. As glove removal was not possible within the simulation, no conclusions regarding postglove hand hygiene can be drawn from the present data. In line with the existing literature, our findings confirm that positively valenced olfactory cues can also exert a beneficial effect on hand hygiene behavior in VR-based simulations, similar to the effects reported by Birnbach et al and King et al [10,11]. Notably, in contrast to these prior studies, which used a citrus scent, the present study used the smell of a disinfectant as the olfactory stimulus.

In contrast, Schmidtke et al [12] were unable to demonstrate a consistent effect of olfactory priming on hand hygiene behavior in a real-world hospital entrance setting, suggesting that the effectiveness of olfactory cues may depend on contextual specificity and clearly defined patient contact situations. To our knowledge, no prior studies have investigated the influence of olfactory cues on medical glove use, indicating a relevant gap in the current literature.

Implications for Patient Safety and Simulation Design

Beyond the behavior-specific effects, these results raise broader questions regarding the role of sensory modalities in VR-based patient safety training. Visual and auditory cues are inherently embedded in contemporary VR systems and can be rendered digitally with high fidelity. In contrast, olfactory stimuli cannot be generated virtually and must originate from the physical world. Thus, smell represents the only deliberately introduced real sensory cue in this study, effectively crossing the boundary between virtual and physical experience—a boundary we refer to as the *virtual threshold*.

This concept relates to several established constructs in the simulation and cognitive psychology literature. Presence and embodiment, defined above, both depend on sensory congruence between the simulated scene and learners' real-world expectations. Ecological validity, in turn, refers to the correspondence between the simulation and the clinical situation it is intended to prepare for. From a memory perspective, the principle of context-dependent retrieval predicts that behaviors learned and habitually performed under specific sensory conditions are more readily activated when those conditions are reinstated. Within this framework, the disinfectant scent can be understood as a sensory retrieval cue that reinstantiates clinically learned associations within an otherwise digitally generated environment, thereby narrowing the gap between virtual and clinical contexts.

Despite its central role in clinical practice, olfaction has largely been neglected in VR-based medical education. As demonstrated

XSL•FO
RenderX

in this study, olfactory cues can influence safety-relevant behavior even when visual and auditory cues alone are insufficient. Evidence on olfactory augmentation in immersive medical simulation remains sparse. The most recent systematic review was published in 2016 by Kent et al [23] and primarily investigated olfactory cues in non-VR simulation contexts, leaving the role of real olfactory input in immersive VR-based medical education largely unexplored.

Insights from psychology and entertainment-oriented VR research suggest that olfactory cues can enhance emotional salience, contextual memory, and behavioral realism [24-26]. However, these findings have rarely been translated into medical education, where multisensory perception is fundamental to safe clinical action.

Theoretical Considerations, Limitations, and Implications for VR-Based Medical Education

This study should be interpreted in light of both theoretical mechanisms and methodological constraints relevant to immersive medical education. Immersive VR aims not only to convey procedural knowledge but also to elicit self-relevant clinical behavior. In this context, it is important to distinguish between immersion (objective system properties), presence (the subjective feeling of "being there"), and embodiment, which refers to the experience of the virtual body as one's own. Embodiment is commonly conceptualized along the dimensions of body ownership, agency, and self-location and has been identified as a key determinant of behavior in virtual environments, particularly when actions involve self-protection or bodily risk [27-29].

Although olfactory cues do not directly modify visual body representation, they may indirectly reinforce embodiment by strengthening the coupling between bodily actions and real-world physiological expectations. In clinical settings, smells associated with contamination, cleanliness, or disinfection are closely linked to bodily vulnerability and habitual protective routines. Introducing such cues from the physical world into an otherwise fully virtual environment may therefore increase the salience of the virtual body as a potential site of risk. This mechanism could help shift learner behavior from abstract rule compliance toward self-protective, behaviorally realistic responses, such as postcontact hand disinfection. This proposed mechanism was not directly tested in the present study and remains a hypothesis for future investigation.

The relatively low baseline rate of hand disinfection in the control group warrants explicit consideration. Several factors plausibly contribute. First, the novelty of the VR environment may divert attention from routine behaviors. Second, the environmental cues that normally prompt hand hygiene in real settings, such as ambient scent, tactile feedback from the dispenser, and the physical proximity of cleansed surfaces, are absent or only weakly present in a purely audiovisual VR scene. Third, learners may operate with a reduced perceived consequence in an obviously simulated context. Under this reading, the olfactory intervention does not enhance hygiene behavior beyond real-world baselines but rather restores cue-triggered behavior under impoverished sensory conditions. This interpretation does not diminish the practical value of the finding because, if VR environments inherently suppress hygiene-relevant behavior, sensory augmentation is precisely what is required to elicit ecologically valid training responses.

However, several limitations constrain the interpretation and generalizability of these findings. First, the olfactory intervention was limited to a single, positively connoted scent associated with cleanliness. While appropriate for feasibility and safety, this does not reflect the full range of clinically relevant olfactory cues. Future studies should include aversive or negatively valenced odors, which may exert stronger or qualitatively different effects on risk perception and behavior.

Second, participant allocation was constrained by faculty-assigned groups, resulting in limited group sizes and minor imbalances. Although the sample size was sufficient to detect the main effects, these constraints may have reduced statistical precision and may have limited subgroup analyses. Six participants were excluded from the analyses due to missing data, all of whom were in the intervention group, reducing the sample from 95 to 89. Sex distribution showed a nonsignificant imbalance (P=.12). Because allocation was quasi-random by room rather than by individual-level randomization, the primary intention-to-treat analysis was complemented by covariate-adjusted sensitivity analyses, and the intervention effect on hand disinfection remained robust across these models. Because participants were allocated within 6 rooms, clustering at the room level cannot be fully excluded, and with only 6 clusters, formal multilevel modeling was not feasible. Third, the behavioral effects observed were behavior-specific: olfactory augmentation influenced overall and postcontact hand disinfection but not initial hand hygiene or glove use. This suggests that olfactory cues may preferentially reinforce reactive, situationally triggered behaviors rather than preventive or purely rule-based routines.

Fourth, the study was conducted within a single VR clinical scenario, which limits extrapolation to other procedures, specialties, or safety behaviors. Broader evaluation across multiple scenarios and end points is required to determine the generalizability of the effects of olfactory augmentation. Finally, technical constraints of the VR tracking system limited the granularity of behavioral measurement. Certain interaction sequences, such as returning to patient contact after hand disinfection outside the tracked examinations, could not be reliably detected, potentially leading to an overestimation of compliance. Future implementations should incorporate more fine-grained temporal and interaction tracking to better capture hygiene-critical behavior.

Fifth, no formal manipulation check was conducted to verify whether participants consciously perceived the olfactory stimulus. We therefore cannot determine whether the effect operates through conscious recognition, such as identifying the disinfectant scent and responding accordingly, or through subconscious priming. Future studies should include a brief postsimulation item assessing scent perception, which would enable stratified analyses of participants who did and did not consciously notice the scent and help adjudicate between conscious and subconscious mechanisms.

XSL•FO
RenderX

A further question concerns the transferability of these findings to experienced clinicians. Our sample comprised medical students with limited clinical experience, whose hygiene behaviors are not yet fully automated. Experienced practitioners might rely on highly routinized motor patterns that are less susceptible to environmental cues, a ceiling interpretation, or they might instead have more strongly encoded cue-behavior associations and therefore respond more readily to congruent sensory primes. Both predictions are plausible and yield different design implications for VR-based continuing education, and which one holds is an empirical question for future studies.

Taken together, these findings and limitations highlight both the promise and the current constraints of olfactory augmentation in VR-based medical education. Future research should systematically vary olfactory valence, improve behavioral tracking fidelity, optimize randomization procedures, and expand scenario diversity. From an educational perspective, the results suggest that olfactory cues may serve as a targeted design element to promote behaviorally realistic, safety-relevant actions in immersive simulations, particularly when the educational goal is not knowledge acquisition alone but the elicitation of authentic clinical behavior.

Conclusions

This study provides empirical evidence that olfactory cues—an integral yet previously neglected component of clinical environments—can meaningfully enhance behavioral outcomes in immersive VR-based medical education. The introduction of a positively connoted disinfectant scent led to a 19.9 percentage point increase in hand hygiene behavior during simulation training, suggesting that real-world sensory augmentation can strengthen safety-relevant actions beyond purely visual and auditory immersion.

In contrast, glove use was not significantly influenced by the olfactory intervention, indicating that different hygiene behaviors may rely on distinct cognitive mechanisms. While hand disinfection appears susceptible to context-driven sensory priming, glove use may depend more strongly on rule-based or procedural decision processes.

Taken together, these findings highlight the potential of integrating physical sensory cues into digital simulation environments to enhance behavioral realism and patient safety training. Future research should investigate the effects of diverse olfactory stimuli across scenarios and learner populations to further elucidate the mechanisms and generalizability of sensory augmentation in VR-based medical education.

Acknowledgments

We thank Philipp Bozdere for adjusting the patient’s facial texture. A large language model (ChatGPT, OpenAI) was used exclusively for language editing and stylistic refinement of the manuscript, with no involvement in data generation, analysis, or interpretation.

Funding

We acknowledge support from the Open Access Publication Fund of the University of Münster.

Data Availability

The dataset used during this study is available from the corresponding author upon reasonable request.

Authors' Contributions

Conceptualization: AJ
Methodology: AJ, JD, BM, HF
Formal analysis: AJ, HF
Investigation: AJ, PK, JD
Resources: AJ, JD, BR, MH, BM
Data curation: AJ, PK
Writing – original draft: AJ, HF
Writing – review & editing: AJ, JD, SHS, BR, MH, BM, HF
Supervision: AJ, BM, HF
Project administration: AJ, JD
Software development: PK, BR, MH
Visualization: JD, HF
Validation: SHS, BM, HF
Funding acquisition: BR, MH, BM

Conflicts of Interest

None declared.

Checklist 1

XSL•FO
RenderX

CONSORT-EHEALTH (V 1.6.1) checklist.
[PDF File, 609 KB - mededu_v12i1e93148_app1.pdf]

References

1. Hibbeler B. Studienhospital Münster: Alles Simulanten! [Article in German]. Deutsches Ärzteblatt. URL: https://www.aerzteblatt.de/archiv/studienhospital-muenster-alles-simulanten-e5c89e97-098f-40a1-9790-9efbaa1819e3 [accessed 2026-07-21]
2. Doering S, Schneider G, Burgmer M, et al. Evaluation des Praktikums »Psychosomatik und Psychotherapie« mit standardisierten Patienten [Article in German]. Z Psychosom Med Psychother 2010 Dec 1;56(4):385-398. [doi: 10.13109/zptm.2010.56.4.385] [Medline: 21243608]
3. Thies AL, Weissenstein A, Haulsen I, Marschall B, Friederichs H. Experiences with a simulated learning environment - the SimuScape©: virtual environments in medical education. J Technol Sci Educ 2014;4(1):48-57. [doi: 10.3926/jotse.88]
4. Wohlgenannt I, Simons A, Stieglitz S. Virtual reality. Bus Inf Syst Eng 2020 Oct;62(5):455-461. [doi: 10.1007/s12599-020-00658-9]
5. Abdlkarim D, Di Luca M, Aves P, et al. A methodological framework to assess the accuracy of virtual reality hand-tracking systems: a case study with the Meta Quest 2. Behav Res Methods 2024 Feb;56(2):1052-1063. [doi: 10.3758/s13428-022-02051-8] [Medline: 36781700]
6. Li Z, Bujić M, Kerouš B, Hamari J. Natural manipulation of virtual reality: the effects of kinesthetic gloves with force feedback on user performance and experience for accurate 3D object manipulation. Int J Hum Comput Stud 2025 May;199:103507. [doi: 10.1016/j.ijhcs.2025.103507]
7. Junga A, Schulze H, Scherzer S, et al. Immersive learning in medical education: analyzing behavioral insights to shape the future of VR-based courses. BMC Med Educ 2024 Dec 3;24(1):1413. [doi: 10.1186/s12909-024-06337-7] [Medline: 39627755]
8. Junga A, Bozdere P, Hätscher O, et al. Rethinking interaction design - special implications for interaction concepts in medical education using virtual reality. Virtual Real 2025;29(3):98. [doi: 10.1007/s10055-025-01180-7]
9. Godden DR, Baddeley AD. Context ‐ dependent memory in two natural environments: on land and underwater. British J Psychol 1975 Aug;66(3):325-331. [doi: 10.1111/j.2044-8295.1975.tb01468.x]
10. Birnbach DJ, King D, Vlaev I, Rosen LF, Harvey PD. Impact of environmental olfactory cues on hand hygiene behaviour in a simulated hospital environment: a randomized study. J Hosp Infect 2013 Sep;85(1):79-81. [doi: 10.1016/j.jhin.2013.06.008] [Medline: 23910402]
11. King D, Vlaev I, Everett-Thomas R, Fitzpatrick M, Darzi A, Birnbach DJ. “Priming” hand hygiene compliance in clinical environments. Health Psychol 2016 Jan;35(1):96-101. [doi: 10.1037/hea0000239] [Medline: 26214075]
12. Schmidtke KA, Aujla N, Marshall T, et al. A crossover randomized controlled trial of priming interventions to increase hand hygiene at ward entrances. Front Public Health 2022;9:781359. [doi: 10.3389/fpubh.2021.781359] [Medline: 35111716]
13. Junga A, Kockwelp P, Valkov D, et al. Teach the unteachable with a virtual reality (VR) brain death scenario - 800 students and 3 years of experience. Perspect Med Educ 2025;14(1):44-54. [doi: 10.5334/pme.1427] [Medline: 39897616]
14. Kockwelp P, Junga A, Valkov D, Marschall B, Holling M, Risse B. Towards VR simulation-based training in brain death determination. Presented at: 2022 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW); Mar 12-16, 2022. [doi: 10.1109/VRW55335.2022.00065]
15. Kockwelp P, Meyerheim M, Valkov D, et al. VR-based competence training at scale: teaching clinical skills in the context of virtual brain death examination. Proc ACM Hum-Comput Interact 2024 Jun 17;8:1-31. [doi: 10.1145/3664635]
16. Schmidt H, Becker J, Friederichs H, et al. Interdisziplinäres Lehr-Modul zur Transplantation für das klinische Medizinstudium – Konzeption, Durchführung und Evaluationsergebnisse [Interdisciplinary teaching module for transplantation: concept, management and evaluation]. Transplantationsmedizin 2010;22:248-254.
17. Faul F, Erdfelder E, Lang AG, Buchner A. G*Power 3: a flexible statistical power analysis program for the social, behavioral, and biomedical sciences. Behav Res Methods 2007 May;39(2):175-191. [doi: 10.3758/bf03193146] [Medline: 17695343]
18. Cohen J. Statistical Power Analysis for the Behavioral Sciences: Routledge; 2013. [doi: 10.4324/9780203771587]
19. Empfehlung der Kommission für Krankenhaushygiene und Infektionsprävention (KRINKO) beim Robert Koch-Institut (RKI). Händehygiene in Einrichtungen des Gesundheitswesens [Article in German]. Bundesgesundheitsblatt Gesundheitsforschung Gesundheitsschutz 2016 Sep;59(9):1189-1220. [doi: 10.1007/s00103-016-2416-6] [Medline: 27558147]
20. Wickham H. tidyverse: easily install and load the “tidyverse”. CRAN: The Comprehensive R Archive Network. 2016. URL: https://cran.r-project.org/web/packages/tidyverse/index.html [accessed 2026-08-03]
21. Robinson D, Hayes A, Couch S, Hvitfeldt E. broom: Convert Statistical Objects into Tidy Tibbles. CRAN: The Comprehensive R Archive Network. 2014. URL: https://cran.r-project.org/web/packages/broom/index.html [accessed 2026-08-03]
22. Studierende: Deutschland, Semester, Nationalität, Geschlecht, Studienfach [Article in German]. Statistisches Bundesamt (Destatis). 2026. URL: https://www-genesis.destatis.de/datenbank/online/url/3d2dcf50 [accessed 2026-02-09]

XSL•FO
RenderX

23. Kent SJW, Kent FH, Brown CW, Morrison IG, Morse JC. Should we add smells in simulation training? A systematic review of smells in healthcare-related simulation training. BMJ Simul Technol Enhanc Learn 2016;2(1):19-22. [doi: 10.1136/bmjstel-2015-000064] [Medline: 35516447]
24. Pützer A, Wolf OT. Odours as context cues of emotional memories—the role of semantic relatedness. Acta Psychol (Amst) 2021 Sep;219:103377. [doi: 10.1016/j.actpsy.2021.103377] [Medline: 34293594]
25. Herz RS. A naturalistic analysis of autobiographical memories triggered by olfactory visual and auditory stimuli. Chem Senses 2004 Mar;29(3):217-224. [doi: 10.1093/chemse/bjh025] [Medline: 15047596]
26. Dinh HQ, Walker N, Hodges LF, Song C, Kobayashi A. Evaluating the importance of multi-sensory input on memory and the sense of presence in virtual environments. Presented at: Proceedings IEEE Virtual Reality (Cat No 99CB36316); Mar 13-17, 1999. [doi: 10.1109/VR.1999.756955]
27. Banakou D, Groten R, Slater M. Illusory ownership of a virtual child body causes overestimation of object sizes and implicit attitude changes. Proc Natl Acad Sci U S A 2013 Jul 30;110(31):12846-12851. [doi: 10.1073/pnas.1306779110] [Medline: 23858436]
28. Kilteni K, Groten R, Slater M. The sense of embodiment in virtual reality. Presence 2012 Nov;21(4):373-387. [doi: 10.1162/PRES_a_00124]
29. Slater M. Place illusion and plausibility can lead to realistic behaviour in immersive virtual environments. Philos Trans R Soc Lond B Biol Sci 2009 Dec 12;364(1535):3549-3557. [doi: 10.1098/rstb.2009.0138] [Medline: 19884149]

Abbreviations

BD/DNC: Brain Death/Death by Neurologic Criteria
CONSORT-EHEALTH: Consolidated Standards of Reporting Trials of Electronic and Mobile Health Applications and Online Telehealth
HMD: head-mounted display
OR: odds ratio
VR: virtual reality

Edited by A Stone; submitted 23.Feb.2026; peer-reviewed by D Winterton, P Joseph; revised version received 15.Jul.2026; accepted 15.Jul.2026; published 19.Aug.2026.

Please cite as:
Junga A, Kockwelp P, Dabel J, Scherzer SH, Risse B, Holling M, Marschall B, Friederichs H
Olfactory Cues and Patient Safety Behavior in Virtual Reality Medical Simulation: Controlled Trial
JMIR Med Educ 2026;12:e93148
URL: https://mededu.jmir.org/2026/1/e93148
doi:10.2196/93148

© Anna Junga, Pascal Kockwelp, Jennifer Dabel, Sönke Helmut Scherzer, Benjamin Risse, Markus Holling, Bernhard Marschall, Hendrik Friederichs. Originally published in JMIR Medical Education (https://mededu.jmir.org), 19.Aug.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Education, is properly cited. The complete bibliographic information, a link to the original publication on https://mededu.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Ambient AI Scribes to Create Educational Feedback Notes for Medical Students: Randomized Trial

Jaideep S Talwalkar[1], MD; David Chartash[2], PhD; Lisa Zhang[2], MD; Michael Makutonin[2], MD; Conrad W Safranek[3], BSc; Anne Elizabeth Sidamon-Eristoff[3], AB; Lee H Schwamm[4], MD; Donald S Wright[2,5], MD, MHS

[1]Departments of Medicine and Pediatrics, Yale School of Medicine, 367 Cedar Street, Bldg D, New Haven, CT, United States

[2]Department of Emergency Medicine, Yale School of Medicine, New Haven, CT, United States

[3]Yale School of Medicine, New Haven, CT, United States

[4]Departments of Neurology and Biomedical Informatics and Data Sciences, Yale School of Medicine, New Haven, CT, United States

[5]Department of Veterans Affairs, Connecticut VA Healthcare System, West Haven, CT, United States

Corresponding Author:
Jaideep S Talwalkar, MD
Departments of Medicine and Pediatrics, Yale School of Medicine, 367 Cedar Street, Bldg D, New Haven, CT, United States

Related Article:

Comment in: https://mededu.jmir.org/2026/1/e105029

Abstract

Background: High-quality observation and feedback contribute to the development of clinical competence and professional growth in medical education. Faculty often struggle to translate verbal observations into written feedback because of documentation burden and competing demands. Ambient artificial intelligence (AI) scribes, already adopted in clinical practice, may address this challenge by capturing verbal exchanges and generating structured notes.

Objective: The purpose of this study was to examine the use of ambient AI scribes to generate educational feedback notes during a formative medical interviewing workshop for first-year medical students in March and April 2025.

Methods: Thirteen instructors were randomized to control (human-only) or intervention (AI scribe–assisted) workflows to complete narrative feedback forms. The intervention group used an AI scribe to generate transcripts of student-instructor encounters, which were then summarized into feedback notes using a large language model and edited by the instructors before submission. All narratives were scored using the Evaluation of Feedback Captured Tool (EFeCT). Factual accuracy of a subsample of unedited AI feedback summaries was reviewed against the source transcripts. Task load and usability were measured using NASA Task Load Index and System Usability Scale, respectively.

Results: Instructors submitted feedback on 92.2% (94/102) of the students. EFeCT scores on the scale from 0 to 5 were higher for human-edited AI narratives (median 3.00, IQR 2.00-4.00) and unedited AI summaries (median 3.00, IQR 3.00-4.00) than for human-only narratives (median 2.00, IQR 1.75-3.00; P<.001). Human-only narratives were shorter than AI-assisted outputs (P<.001). Review of 117 AI-generated feedback elements showed a 6.8% (n=8) mischaracterization and 1.7% (n=2) hallucination rate, with most errors corrected during editing. Task load was high, and usability was marginal in both the control and intervention groups, with no significant differences (P=.31 and P=.40, respectively).

Conclusions: An ambient AI scribe–assisted workflow improved the quality of written narrative feedback with no observed increase in instructor effort compared to human-only documentation. Although occasional inaccuracies required review, this innovation has the potential to transform feedback documentation.

(JMIR Med Educ 2026;12:e89996) doi:10.2196/89996

KEYWORDS

ambient scribe; artificial intelligence; AI; feedback; medical student education; formative assessment; competency-based education

XSL•FO
RenderX

Introduction

Problem: Documentation Burden and Quality of Feedback in the Modern Assessment Paradigm

High-quality observation and feedback contribute to the development of clinical competence and professional growth in medical education. Effective feedback provides learners with specific and actionable information on their performance, which can promote deliberate practice, motivation, and engagement with ongoing feedback [1-4]. While verbal feedback is critical for collaborative discussion [5] and occurs frequently in clinical [1] and simulation settings [6], it may be limited as a tool for longitudinal reflection as learners recall few feedback points when given verbal feedback alone [5,7]. Written feedback supports self-regulated learning by allowing trainees and their supervisors to review past narratives, set goals, and refine strategies for improvement [1,8]. By providing detailed, context-rich guidance, narrative feedback can help learners focus efforts on achieving professional outcomes expected of them during training [1,9].

In recognition of the value of both verbal and written feedback, accreditation bodies such as the Liaison Committee on Medical Education emphasize the importance of defined competencies within a robust assessment system that includes narrative feedback [10]. Unfortunately, medical educators faced with competing responsibilities and increased educational documentation burden [11] struggle to convert real-time verbal observations into meaningful written feedback [1]. As a result, rather than serving as a tool to promote learning, formative assessment is often reduced to a checklist activity disconnected from best practices for feedback. Brief, nonspecific narratives are common and diminish the educational value of written feedback [11,12].

Solution: Ambient Artificial Intelligence Scribes in Educational Practice

Ambient artificial intelligence (AI) refers to AI embedded into environments, working continuously in the background to support human tasks [13]. Ambient AI scribes have seen rapid adoption to address the growing documentation burden in clinical practice. Ambient AI scribes passively capture and document physician-patient conversations into structured clinical notes [14]. There is early evidence suggesting that these tools reduce administrative burden while maintaining or improving documentation quality [15,16]. The potential for this technology to extend beyond patient care into medical education is compelling. By capturing and structuring verbal formative feedback provided by faculty in real time, ambient AI scribes could generate written records of learner performance. These "educational feedback notes" could transform narrative workflows similar to the way in which AI-generated clinical notes are transforming clinical practice [17].

Gap: No Use in Medical Education Settings

Despite the promise of ambient AI scribes in clinical settings, their application in medical education remains unexplored. No studies have described the use of these tools in directly observed encounters or small-group teaching settings, examined whether they can reliably capture narrative feedback in teaching contexts, compared how their outputs align with best practices for written feedback, or determined whether faculty perceive them as usable while actively teaching. No extensions to summarize data into educational feedback notes currently exist. The potential for ambient AI scribes to transform feedback documentation in education remains untested.

This study is the first to our knowledge to examine the use of ambient AI scribes to generate educational feedback notes during directly observed standardized patient (SP) encounters. Specifically, we aimed to (1) evaluate the quality and accuracy of narrative feedback captured by an ambient AI scribe workflow compared to that of feedback provided by humans without AI assistance during a formative medical interviewing workshop for first-year medical students and (2) assess the usability of this technology by medical educators.

Methods

Educational Setting

This study took place in spring 2025 as part of a formative medical interviewing workshop for all first-year medical students at Yale School of Medicine. During the workshop, each medical student spent 20 minutes conducting a complete history with an SP while being observed by an instructor and 3 peers. Students and instructors were permitted to call "time-out" for interspersed guidance and feedback [6]. Upon completion of the interview, 10 minutes were allotted for self-, peer, and instructor feedback. SPs provided no feedback. The exercise was repeated until all 4 students in the room had interviewed separate SPs portraying different scenarios and received individual verbal feedback. After the workshop, instructors completed a postsession feedback narrative (Multimedia Appendix 1) for each student within a learning management system (Medtrics Lab LLC). The workshop was cycled over 4 afternoons, with each student participating once. Instructors devoted 3 and a half hours per workshop, including presession faculty development. This was the seventh iteration of the workshop in the first-year clinical skills curriculum, with each workshop structured similarly, offering students opportunities for repeat practice as their foundational knowledge grows during medical school.

Participants and Platforms

Study participants were 13 clinician educators who had signed up to teach medical interviewing workshops based on their availability before the study was announced. All had completed standard training on teaching medical interviewing, facilitating small groups, working with SPs, and giving feedback. All had facilitated other medical interviewing workshops earlier in the academic year and completed feedback forms associated with those sessions. Instructors were informed of the study via email and in-person announcements.

All instructors received the same study-specific training on form completion and use of the verbatim transcription feature within the health system's AI scribe (Abridge Inc), a generative AI tool based on a large language model (LLM) built using medical notes. The AI scribe transcripts were extracted separately from

XSL•FO
RenderX

the final medical note that Abridge produces and then fed manually into a private instance of GPT-4o (OpenAI) within the university's secure infrastructure (Clarity Platform; Multimedia Appendix 2) to produce a feedback summary that ignored the direct medical content and instead focused on student feedback. Other tools included a data collection instrument (Qualtrics International Inc) and the learning management system. Those wishing to opt out were instructed to inform a course administrator not involved in the study.

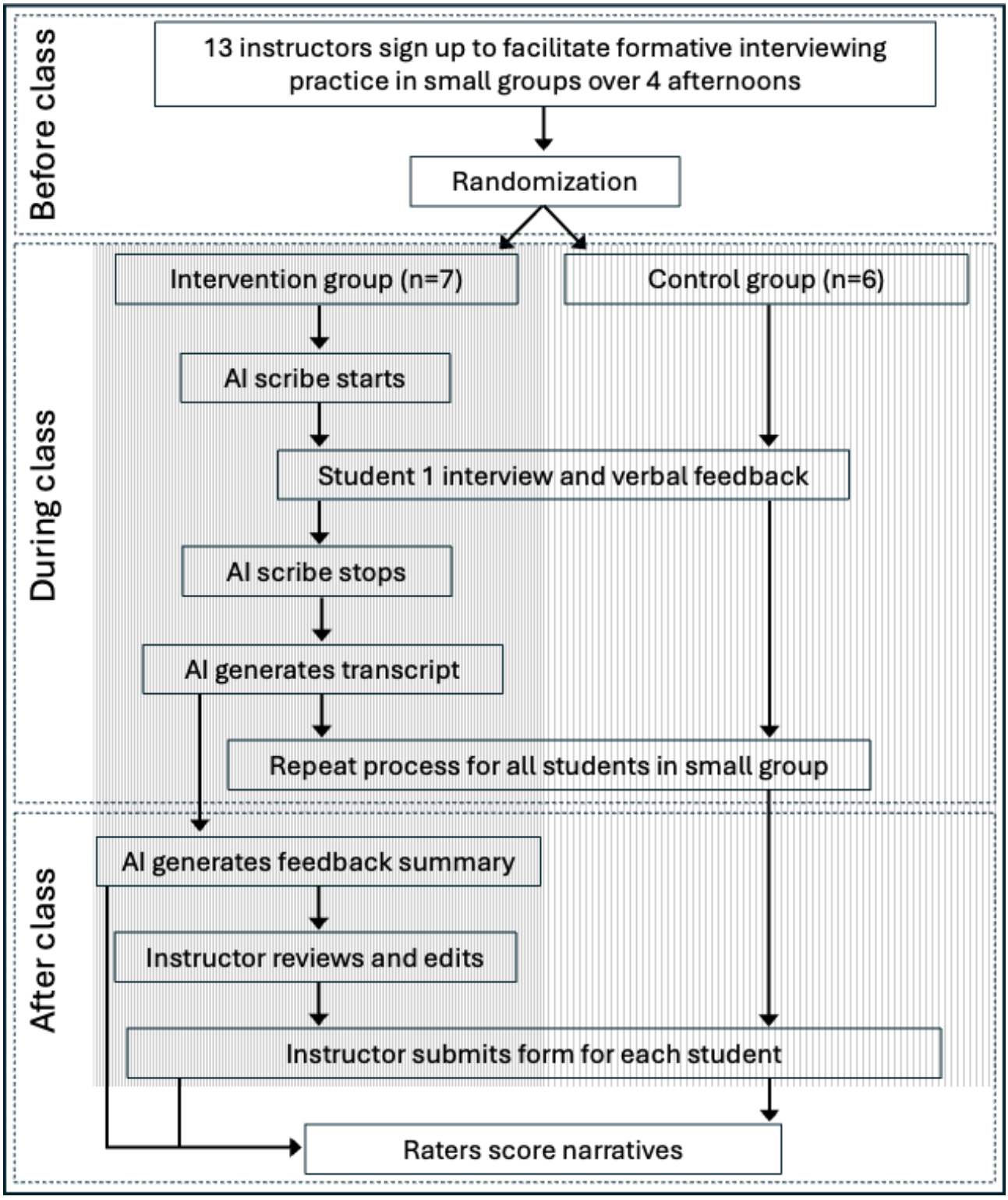
While the entire class of 104 medical students involved in the workshop was given the option to opt out, no students or instructors did.

Feedback Workflows and Data Collection

Instructors were randomized to the control or intervention group through random number generation to reduce bias of fixed effects from prior experience with the educational activity and technology, thereby isolating effects on the quality of the narratives (Figure 1). Instructors were informed of their group assignments immediately prior to the workshop. Students were unaware of their instructor's group assignment, and all instructors placed their smartphones on a counter centrally located in the room at the start of the session. Control group instructors did not activate the AI scribe, and after completing the workshop, composed the "human narratives" on the postsession feedback forms without AI assistance. Intervention group instructors activated the AI scribe, which ran unobtrusively on a smartphone in the background, during each interview and feedback session to create a session transcript. After the workshop, they used a zero-shot prompt (ie, a prompting strategy in which the LLM is asked to perform a task without being given examples, training, or fine-tuning) [18,19] within the Clarity Platform (Multimedia Appendix 3) to extract the educational feedback comments and use them to create a feedback summary from each transcript. The prompt was written to summarize feedback verbalized during the educational session, not to create new feedback. Instructors reviewed and edited these outputs as they deemed necessary before entering them as final narratives, thus turning unedited AI feedback summaries into human-edited AI narratives.

XSL•FO
RenderX

Figure 1. Depiction of the study workflow. Instructors were randomized to the control or intervention group; the latter used an artificial intelligence (AI) scribe to create a feedback summary. Three sources of narratives were analyzed: unedited AI, human-edited AI, and human only (control). Raters were blinded to the source of the narratives.



All text created through this process was submitted via a data collection instrument designed to preserve student anonymity. The instrument also prompted instructors to complete the NASA Task Load Index [20] and System Usability Scale [21] to measure instructor cognitive load and general usability, respectively.

Review of Narratives

Feedback narratives were assessed using the Evaluation of Feedback Captured Tool (EFeCT), a feedback quality scoring tool consisting of 5 elements, each of which represents an aspect of good written feedback. One point is given for each element present, for a total score of 0 to 5 [22]. Two blinded, independent raters (DC and LZ) assigned scores asynchronously for each narrative. Reconciliation of mental models for each element

XSL•FO
RenderX

was performed after 10 and 25 narratives. After all narratives were scored, the 2 raters synchronously reconciled their scores into a final score for each element for each narrative using the constant comparative method. This final reconciled score was passed to a third rater (JST) to be confirmed. In addition to the use of EFeCT, the raters flagged whether they thought the narrative was generated through AI assistance (binary variable).

Evaluation of Factuality of AI Feedback Summaries

Eight unedited AI feedback summaries were randomly sampled from the intervention cohort to benchmark any potential factual inaccuracies as an exploratory secondary analysis. These summaries were manually delimited into concepts by identifying each independent feedback statement that relied on an observed behavior in the simulation. Each individual feedback concept in the unedited AI feedback summary was cross-referenced with the full transcript of the session by an author (DSW) and classified for factuality (ie, whether the information in the written text corresponded to a real-world fact) [23]. Feedback was classified as correct if it matched the content of the original transcript, mischaracterized if it misrepresented the original context or meaning of the interaction in the transcript, or hallucinated if it referenced events not present in the transcript. Each feedback element classified as mischaracterized or hallucinated was then compared to the human-edited AI narrative to see whether the error was resolved.

Statistical Analysis

Analyses were conducted in R (version 4.2.2; R Foundation for Statistical Computing). For the EFeCT dataset, total EFeCT feedback scores and word counts were compared across sources (human only, unedited AI, and human-edited AI) using Kruskal-Wallis tests, with pairwise Wilcoxon rank-sum tests with Holm correction for multiple comparisons. Individual EFeCT subscores (elements B-F) were treated as binary outcomes and compared across groups using the Fisher exact test and then pairwise when the omnibus test was significant with Holm correction. Factuality analysis was conducted using descriptive statistics. The Fisher exact test was used to compare completion rates. We tested assumptions of normality (Shapiro-Wilk) and homogeneity of variance (Levene) prior to the cognitive load and usability analysis, in which we used independent-sample 2-tailed *t* tests. All tests were 2 tailed with significance set at an α value of .05.

Ethical Considerations

This study received an exemption from review by the Yale University Institutional Review Board due to its educational nature on January 23, 2025 (protocol ID 2000039478).

Results

Each of the 13 instructors taught 1 to 4 afternoons of the 4-workshop cycle. Of the 104 students in the class, 2 (1.9%) were absent due to illness. Instructors submitted feedback forms via the data collection instrument on 92.2% (94/102) of the students. The 46.2% (6/13) of instructors randomized to the control group taught a mean of 1.7 (SD 0.5) sessions and submitted human-only narratives for 84.2% (32/38) of the students in their groups, all of which were submitted on the day of the session. The 53.8% (7/13) of instructors in the intervention group taught a mean of 2.6 (SD 1.1) sessions and submitted unedited AI feedback summaries for 95.3% (61/64) of the students and human-edited AI narratives for 96.9% (62/64) of the students, a completion rate significantly higher than that of the control group (P=.049). All were submitted within 1 week of the session, with 95.2% (59/62) submitted on the day of the session.

Median EFeCT scores were 2.00 (IQR 1.75-3.00) for human-only narratives (control group), 3.00 (IQR 2.00-4.00) for human-edited AI narratives, and 3.00 (IQR 3.00-4.00) for unedited AI feedback summaries (Table 1). When narratives were grouped by feedback generation method, a significant difference in total EFeCT scores was observed (P<.001; Figure 2). Subsequent pairwise comparisons demonstrated higher EFeCT scores in both the human-edited and unedited AI feedback methods relative to the human-only narratives (P=.005 and P<.001, respectively), whereas no difference was observed between the 2 AI methods (P=.11). Similar significant overall and pairwise effects were observed on 4 of the 5 EFeCT component elements (Table 2 and Figure 3).

XSL•FO
RenderX

Table . Review of narratives, including factuality analysis.

Outcome	Narrative feedback workflow			Groupwise comparison	
	Human-only narratives (n=32)	Human-edited AI narratives[a] (n=62)	Unedited AI narratives (n=61)	*P* value	Effect size estimate (ε^2)
EFeCT[b] score, median (IQR)	2.00 (1.75-3.00)	3.00 (2.00-4.00)	3.00 (3.00-4.00)	<.001[c]	0.11
Word count, median (IQR)	45.50 (33.00-62.00)	312.50 (207.25-355.00)	344.00 (310.00-396.00)	<.001[d]	0.40
Mischaracterizations among AI feedback elements, n (%)	—[e]	1/117 (0.9)	8/117 (6.8)	—	—
Hallucinations among AI feedback elements, n (%)	—	2/117 (1.7)	2/117 (1.7)	—	—

[a]AI: artificial intelligence.

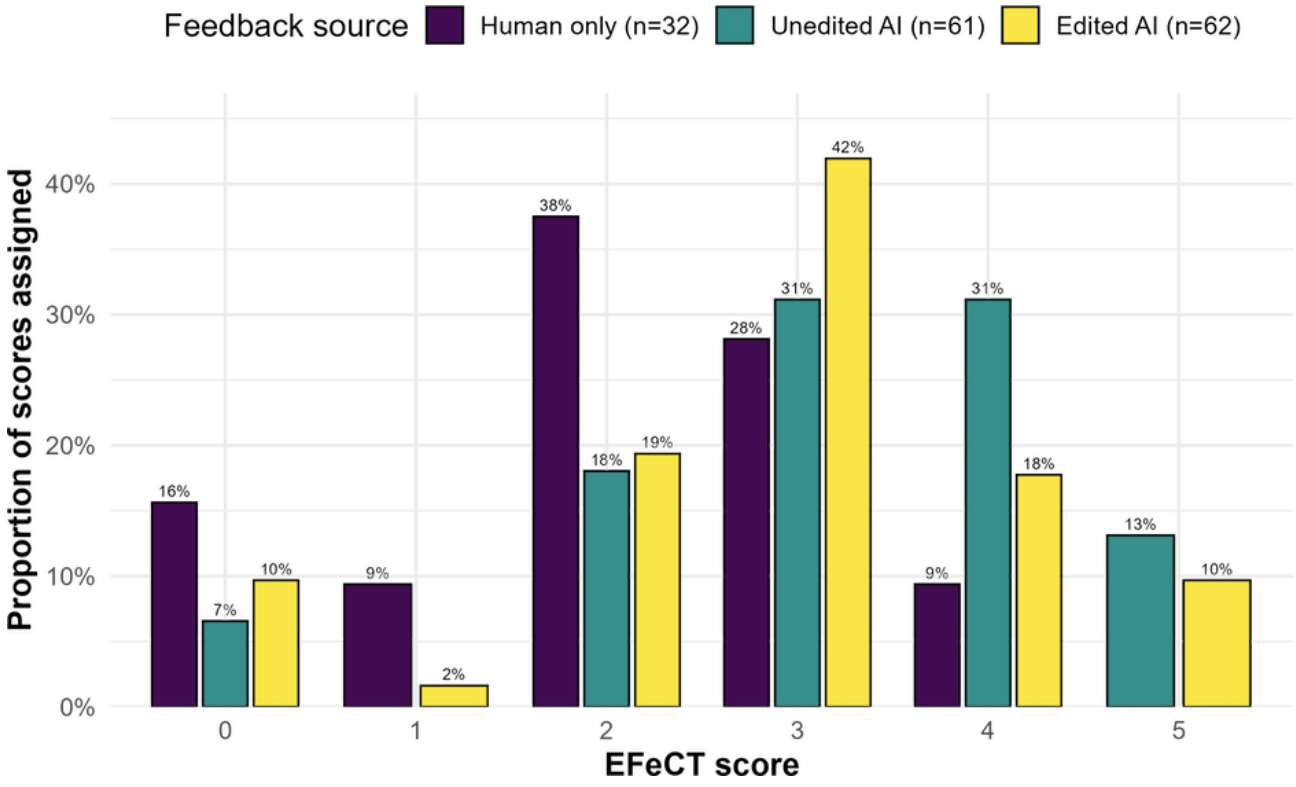
[b]EFeCT: Evaluation of Feedback Captured Tool.

[c]Via the Kruskal-Wallis test. Additional pairwise comparisons demonstrated higher scores for edited (*P*=.005) and unedited (*P*<.001) AI narratives relative to human-only narratives.

[d]Additional pairwise comparisons demonstrated that all differences were statistically significant.

[e]Not applicable.

Figure 2. Evaluation of Feedback Captured Tool (EFeCT) scores by narrative source: human only, unedited artificial intelligence (AI), and edited AI. The EFeCT consists of 5 elements, each of which represents an aspect of good written feedback. One point is given for each element present for a total score of 0 to 5.



XSL•FO
RenderX

Table . Evaluation of Feedback Captured Tool (EFeCT) element comparisons.

Element and pairwise comparison	Pairwise *P* value[a]	Risk difference	Groupwise comparison[b]	
			P value	Effect size, Cramér *V*
What did the learner do? Interpretation: explicit statements that the participant performed a medical interview or completed a medical history.			.01	0.21
Unedited AI[c] vs human-edited AI	.56	0.06		
Human only vs human-edited AI	.06	–0.25		
Human only vs unedited AI	.01	–0.31		
Context: when, who, where? Interpretation: a positive score was assigned if the raters were able to recognize which patient was being interviewed based on the context provided.			.001	0.27
Unedited AI vs human-edited AI	.06	0.18		
Human only vs human-edited AI	.08	–0.15		
Human only vs unedited AI	.002	–0.31		
How did the learner do? Interpretation: there was an explicit statement verbalized that the participant performed a medical interview AND a qualifier was provided stating how well they performed.			.70	0
Unedited AI vs human-edited AI	__[d]	—		
Human only vs human-edited AI	—	—		
Human only vs unedited AI	—	—		
What was done well or needs improvement? Interpretation: any explicit statement of feedback on a specific task.			.01	0.23
Unedited AI vs human-edited AI	.76	0.03		
Human only vs human-edited AI	.05	–0.20		
Human only vs unedited AI	.02	–0.23		
How was it done well or how can it be improved? Interpretation: any explicit statement that highlighted a specific element of a task that was done well or needed improvement.			<.001	0.31
Unedited AI vs human-edited AI	.11	0.11		
Human only vs human-edited AI	.03	–0.24		
Human only vs unedited AI	<.001	–0.36		

[a]Pairwise comparisons via the Fisher exact test with Holm correction for multiple comparisons; effect size reported as the risk difference (proportion meeting the criterion in group 1 minus group 2). A positive risk difference indicates higher EFeCT element attainment in the first-listed group.

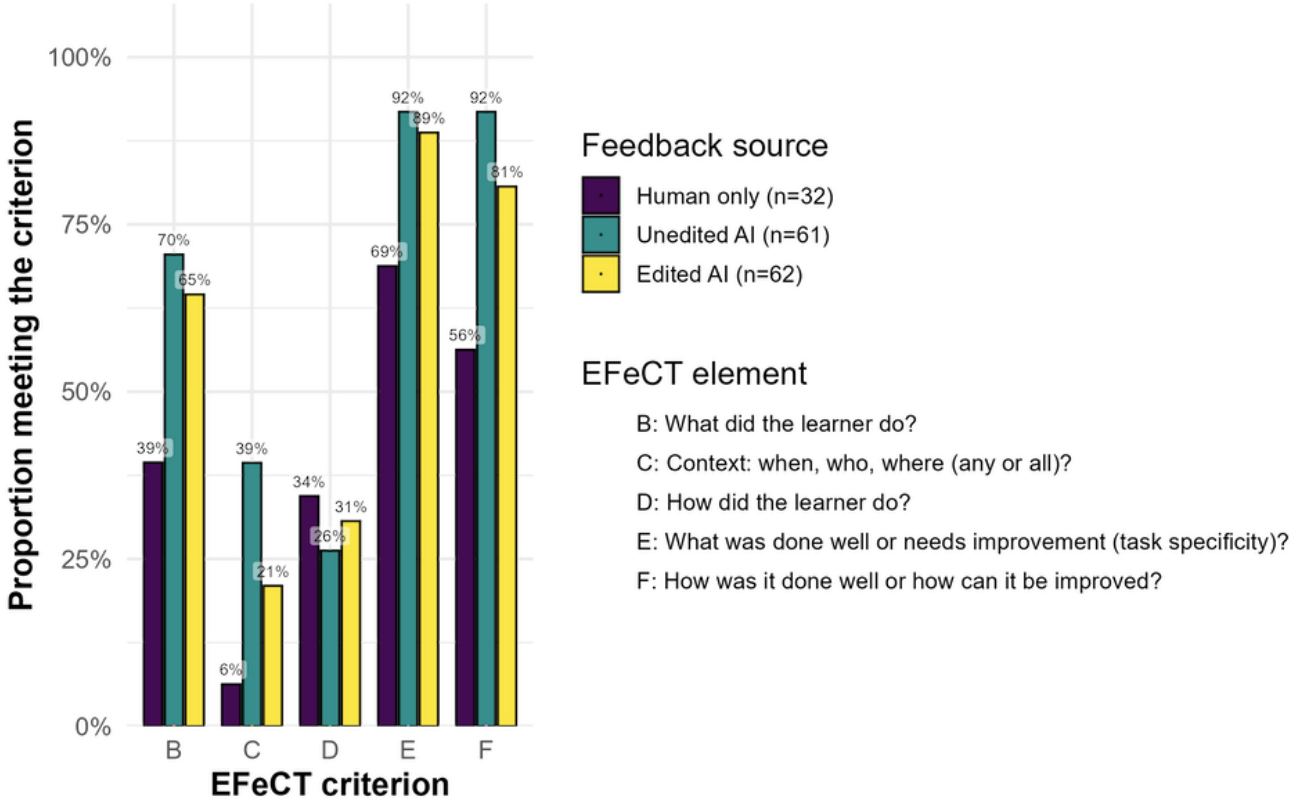
[b]Via the Fisher exact test on 3 × 2 tables.

[c]AI: artificial intelligence.

[d]Not applicable.

XSL•FO
RenderX

Figure 3. Evaluation of Feedback Captured Tool (EFeCT) scores by element. The EFeCT consists of 5 elements. One point is given for each element present. Element lettering begins with B because the "A" designation is used when the narrative section is blank, which did not apply to this study. AI: artificial intelligence.



Human-only narratives were significantly shorter (median 45.50, IQR 33.00-62.00 words) than those generated in human-edited (median 312.50, IQR 207.25-355.00 words) and unedited (median 344.00, IQR 310.00-396.00 words) AI feedback workflows (P<.001 in both cases). Human-edited AI narratives were shorter than unedited summaries (P=.001; Multimedia Appendix 4). A total of 117 AI-generated feedback elements were manually reviewed for factuality, demonstrating 8 (6.8%) mischaracterizations and 2 (1.7%) hallucinations in comparison to the source transcript. Most errors were corrected by instructors prior to submission of human-edited AI narratives (Table 1). Raters correctly identified 93.5% (58/62) of AI-assisted narratives and 100% (32/32) of human-only narratives.

Task load related to completion of the narratives as measured using the NASA Task Load Index did not differ between the control (mean 47.08, SD 12.39) and intervention (mean 54.32, SD 9.45) groups (P=.31; mean difference 95% CI –8.28 to 22.75), with scores indicating high cognitive load for both workflows [24]. Similarly, general usability as measured using the System Usability Scale did not differ between the control (mean 58.75, SD 19.55) and intervention (mean 49.70, SD 13.65) groups (P=.40; mean difference 95% CI –33.25 to 15.15), with scores indicating marginal acceptability [21].

Discussion

Quality of Narratives

The integration of an ambient AI scribe into educational workflows improved the quality of written feedback in a medical student workshop with no observed increase in instructor effort. Feedback narratives generated with AI assistance received significantly higher scores on the EFeCT instrument than those written by clinician educators without AI assistance. By passively capturing verbal feedback and restructuring it into written narratives, AI scribes may address a long-standing challenge in competency-based medical education: translating verbal observations into high-quality written narratives. This is the first study to our knowledge to explore the use of ambient AI scribes to create educational feedback notes, with improvement in documentation quality [15,16] and timeliness [25] similar to that observed with ambient scribes in clinical care.

The improvement observed with AI-assisted narratives is educationally meaningful: a 1-point increase on the EFeCT instrument exceeded the gains reported in the original validation study by Ross et al [22], where multiyear faculty development and feedback were required to achieve a smaller degree of improvement. Several factors may explain why AI-assisted narratives were scored higher on nearly all elements of the EFeCT instrument. Because the EFeCT instrument rewards inclusion of specific feedback elements, the longer AI-assisted narratives may have scored higher because they provided the detail needed to explicitly include those elements. The zero-shot LLM, guided only by a simple prompt, drew on a broad knowledge base of feedback principles, generating more structured, context-rich narratives. In contrast, despite prior extensive training in delivering verbal and written feedback, instructor-written narratives omitted best practice suggestions, as has been described previously. In an integrative review on

XSL•FO
RenderX

feedback practices, Bing-You et al [26] identified inadequate feedback as a frequent theme in the literature even among experienced educators and those who had received feedback training. Gingerich et al [12] highlight a tendency for faculty to either omit written feedback or provide only vague, nonspecific comments, particularly when they feel that essential feedback has already been delivered verbally or is considered "unwritable" due to social or relational dynamics. Additionally, instructors may have relied on the assumption that students will sufficiently retain verbal feedback given during the encounter despite evidence to the contrary [7]. Finally, despite having volunteered, instructors may have faced competing professional demands that accumulated during the 3.5-hour workshop, potentially limiting time spent on feedback forms [1,11].

The impact of an AI-assisted workflow in narrative feedback may be even greater outside of research settings. In this study, all instructors were clinician educators who were aware that their notes would be reviewed as part of the research protocol. This may have resulted in more detailed narratives from instructors in the control group than they would have submitted otherwise. Additionally, nearly all narratives were submitted immediately. In real-world educational environments, where delays in form completion are common and specificity is often lost due to poor recall [27], the potential benefit of ambient AI scribes may be even greater. Alternatively, observed differences in completion rates may be attributable to novelty of the intervention workflow, warranting iterative evaluation of sustained impact.

Accuracy and Reliability of AI Outputs

As a secondary, exploratory outcome, we also attempted to measure the rate of hallucinations and mischaracterizations. Within the sample of unedited feedback summaries that were examined, factual inconsistencies with the underlying transcript were rare; however, the certainty of this estimate is limited by the size of the sample reviewed. Instructors corrected most of the errors in the unedited AI feedback summaries, resulting in shorter human-edited AI narratives that retained higher EFeCT scores. This finding underscores the importance of human oversight over AI outputs. Accurate documentation in medical education remains paramount given its importance for learner development, institutional accountability, and public trust in training outcomes [9].

Task Load and Usability

We did not observe reductions in task load or improvements in usability when using AI scribes compared to the traditional manual human workflow. Both approaches were rated as cognitively demanding and only marginally usable. Notably, the manual workflow itself was associated with high task load, echoing prior work showing that educational documentation can add extraneous demands that detract from core teaching activities [1,11,28]. Cognitive load theory dictates that, when administrative processes create high extraneous load, educators have fewer cognitive resources available for essential instructional and clinical tasks [29]. This dynamic likely contributes to the tendency of faculty to abbreviate written feedback despite their awareness of best practices.

Our findings contrast with reports from clinical medicine, where ambient AI scribes have consistently reduced cognitive burden and improved usability for clinicians [14,16,30,31]. The discrepancy likely reflects the research workflow we used, which required manual copying and pasting among 3 separate platforms (Abridge, Clarity, and Medtrics), all of which were used at no added cost as they were already available within the university and health system. By design, the copy-paste workflow introduced inefficiencies to capture interim outputs for study purposes that would be unnecessary in practice. In a production-grade system, this would be replaced by automated data flow, which may lead to better task load and usability relative to our intervention, so our findings related to cognitive load and usability should be interpreted with caution. Additionally, with well-designed systems, cognitive load [29] and usability [21] typically improve with experience, suggesting that adoption over time may mitigate some of the present challenges as workflows become smoother and more familiar.

Limitations

This study has several limitations. It was conducted at a single institution with a small number of highly trained educators, limiting generalizability. Participants were diligent in reviewing AI outputs, which may not reflect the behaviors of a broader faculty population in usual teaching environments. However, in real-world practice, we can reasonably expect human-only narratives to be less diligently submitted, potentially further increasing the observed differences in quality between AI-assisted and human-only narratives.

Additionally, the AI workflow relied on robust verbal feedback having occurred as the scribe was not enabled to generate its own feedback. We used this safeguard to ensure that feedback was consistent with session objectives.

Second, some contamination may have occurred if instructors in the intervention arm occasionally defaulted to human-only narratives, potentially due to usability challenges. Raters were excellent at identifying AI-assisted narratives, so the rare occasions when raters misattributed an AI-generated narrative to the human-only group raise the possibility that technical difficulties or workflow barriers led some intervention instructors to manually compose narratives. If this occurred, the observed differences in quality between AI-assisted and human-only narratives may, in fact, underestimate the true effect of AI assistance.

Finally, while the EFeCT instrument provides a framework for evaluating written feedback, this was its first application to our knowledge to AI-generated narratives. The EFeCT instrument focuses on structure and content but does not capture tone or learner-perceived utility.

Future Directions

Next steps should expand on these findings in several directions. It is essential to engage learners and clinical coaches to determine whether AI-assisted narratives are useful or actionable for reflection and growth. Through prompt engineering, AI outputs can be iteratively refined to more closely align with the needs of learners and coaches and adjusted to improve factuality [32]. Additionally, integrating AI scribe content into streamlined

XSL•FO
RenderX

workflows and data pipelines could reduce task load and facilitate adoption. Finally, applications in other educational settings—such as clinical precepting and bedside teaching—represent promising areas where ambient AI could help close the gap between frequent verbal feedback and inadequate written documentation.

Conclusions

Ambient AI scribes represent a promising innovation in medical education, where the challenge of creating high-quality narrative feedback persists. In this study, AI-assisted feedback narratives—produced through the combination of automated transcription and zero-shot prompting of an LLM—were of higher quality than human-only narratives with no observed increased human effort. While usability gains were not realized in this research workflow, integration into streamlined educational systems could replicate the successes of ambient AI scribes observed in clinical practice. With continued refinement and oversight, ambient AI scribes have the potential to strengthen feedback culture; support longitudinal assessment; and, ultimately, enhance the learning experience for medical students.

Acknowledgments

The views expressed in this article are those of the authors and do not necessarily reflect the position or policy of the US Department of Veterans Affairs or the US government. The authors used large language models in the preparation of the manuscript to assist with literature search (OpenEvidence), formatting references (GPT-4o; OpenAI), and mock peer review (GPT-4o). GPT-4o was accessed using a generative artificial intelligence tool housed within Yale University's secure infrastructure in which inputs are not added to the external training dataset. There is no plagiarism of text or images related to our use of artificial intelligence.

Funding

No external funding was received.

Data Availability

The datasets generated or analyzed during this study are not publicly available due to the need to protect student privacy, as the data were collected within a single class of students at a single medical school. The data are available from the corresponding author on reasonable request.

Authors' Contributions

JST contributed to conceptualization, formal analysis, investigation, methodology, project administration, resources, software, supervision, visualization, and writing—original draft. DC contributed to conceptualization, data curation, investigation, methodology, validation, and writing—original draft. LZ contributed to data curation, investigation, validation, and writing—review and editing. MM contributed to data curation, formal analysis, validation, and writing—review and editing. CWS contributed to conceptualization, methodology, and writing—review and editing. AES-E contributed to conceptualization, methodology, and writing—review and editing. LHS contributed to conceptualization, methodology, resources, software, supervision, and writing—review and editing. DSW contributed to data curation, formal analysis, investigation, methodology, software, validation, visualization, and writing—review and editing.

Conflicts of Interest

LHS reports consulting with Medtronic, LifeImage, Genentech, and Penumbra unrelated to this work; serving as a content advisor on digital health to the *Stroke* editorial board; holding an artificial intelligence–related patent pending (US2024031761); and representing Yale New Haven Health System on a hospital advisory committee for Abridge, an ambient artificial intelligence clinical documentation company. All other authors declare no other conflicts of interest.

Multimedia Appendix 1
Postsession form completed by the instructor for each student. This standard formative assessment form is used across curricular sessions during which medical students are directly observed performing the patient care skills of history taking or physical examination. Instructors compose a feedback narrative in the final open-response box.
[PNG File, 223 KB - mededu_v12i1e89996_app1.png]

Multimedia Appendix 2
Clarity Platform technical specifications.
[DOCX File, 14 KB - mededu_v12i1e89996_app2.docx]

Multimedia Appendix 3
Zero-shot prompt in GPT-4o used by the intervention group.

XSL•FO
RenderX

[DOCX File, 14 KB - mededu_v12i1e89996_app3.docx]

Multimedia Appendix 4
Sample feedback notes submitted by instructors to first-year medical students following the formative medical interviewing workshop. Three representative samples from each workflow are presented with the removal of student identifiers. Control group instructors created the narratives manually. Intervention group instructors created the narratives with the use of an ambient artificial intelligence scribe.
[DOCX File, 19 KB - mededu_v12i1e89996_app4.docx]

References

1. Cooper D, Holmboe ES. Competency-based medical education at the front lines of patient care. N Engl J Med 2025 Jul 24;393(4):376-388. [doi: 10.1056/NEJMra2411880] [Medline: 40700689]
2. Watling CJ, Ginsburg S. Assessment, feedback and the alchemy of learning. Med Educ 2019 Jan;53(1):76-85. [doi: 10.1111/medu.13645] [Medline: 30073692]
3. Veloski J, Boex JR, Grasberger MJ, Evans A, Wolfson DB. Systematic review of the literature on assessment, feedback and physicians' clinical performance: BEME Guide No. 7. Med Teach 2006 Mar;28(2):117-128. [doi: 10.1080/01421590600622665] [Medline: 16707292]
4. Lai MM, Roberts N, Mohebbi M, Martin J. A randomised controlled trial of feedback to improve patient satisfaction and consultation skills in medical students. BMC Med Educ 2020 Aug 20;20(1):277. [doi: 10.1186/s12909-020-02171-9] [Medline: 32819352]
5. Natesan S, Jordan J, Sheng A, et al. Feedback in medical education: an evidence-based guide to best practices from the Council of Residency Directors in Emergency Medicine. West J Emerg Med 2023 May 5;24(3):479-494. [doi: 10.5811/westjem.56544] [Medline: 37278777]
6. Talwalkar JS, Cyrus KD, Fortin AH. Twelve tips for running an effective session with standardized patients. Med Teach 2020 Jun;42(6):622-627. [doi: 10.1080/0142159X.2019.1607969] [Medline: 31033363]
7. Humphrey-Murto S, Mihok M, Pugh D, Touchie C, Halman S, Wood TJ. Feedback in the OSCE: what do residents remember? Teach Learn Med 2016;28(1):52-60. [doi: 10.1080/10401334.2015.1107487] [Medline: 26787085]
8. Sargeant JM, Mann KV, van der Vleuten CP, Metsemakers JF. Reflection: a link between receiving and using assessment feedback. Adv Health Sci Educ Theory Pract 2009 Aug;14(3):399-410. [doi: 10.1007/s10459-008-9124-4] [Medline: 18528777]
9. Holmboe ES, Sherbino J, Long DM, Swing SR, Frank JR. The role of assessment in competency-based medical education. Med Teach 2010;32(8):676-682. [doi: 10.3109/0142159X.2010.500704] [Medline: 20662580]
10. Publications. Liaison Committee on Medical Education. URL: https://lcme.org/publications/ [accessed 2026-05-10]
11. Szulewski A, Braund H, Dagnone DJ, et al. The assessment burden in competency-based medical education: how programs are adapting. Acad Med 2023 Nov 1;98(11):1261-1267. [doi: 10.1097/ACM.0000000000005305] [Medline: 37343164]
12. Gingerich AN, Lingard L, Sebok-Syer SS, Watling CJ, Ginsburg S. "Praise in public; criticize in private": unwritable assessment comments and the performance information that resists being written. Acad Med 2024 Nov 1;99(11):1240-1246. [doi: 10.1097/ACM.0000000000005839] [Medline: 39137257]
13. Tierney AA, Gayre G, Hoberman B, et al. Ambient artificial intelligence scribes: learnings after 1 year and over 2.5 million uses. NEJM Catalyst 2025 Apr 16;6(5). [doi: 10.1056/CAT.25.0040]
14. Gams M, Gu IYH, Härmä A, Muñoz A, Tam V. Artificial intelligence and ambient intelligence. J Ambient Intell Smart Environ 2019;11(1):71-86. [doi: 10.3233/AIS-180508]
15. Cain CH, Davis AC, Broder B, et al. Quality assurance during the rapid implementation of an AI-assisted clinical documentation support tool. NEJM AI 2025;2(4). [doi: 10.1056/AIcs2400977]
16. Tierney AA, Gayre G, Hoberman B, et al. Ambient artificial intelligence scribes to alleviate the burden of clinical documentation. NEJM Catalyst 2024 Feb 21;5(3). [doi: 10.1056/CAT.23.0404]
17. Fu P. Transforming clinical workflows with artificial intelligence (AI)-based technologies. In: Zheng K, Westbrook J, Patel VL, editors. Reengineering Clinical Workflow in the Digital and AI Era: Springer; 2025:35-55. [doi: 10.1007/978-3-031-82971-0_3]
18. Liu J, Liu F, Wang C, Liu S. Prompt engineering in clinical practice: tutorial for clinicians. J Med Internet Res 2025;27:e72644-e72644. [doi: 10.2196/72644] [Medline: 40955776]
19. CHART Collaborative, Huo B, Collins GS, et al. Reporting guideline for chatbot health advice studies: the CHART statement. JAMA Netw Open 2025 Aug 1;8(8):e2530220. [doi: 10.1001/jamanetworkopen.2025.30220] [Medline: 40747871]
20. Hart SG. NASA-Task Load Index (NASA-TLX); 20 years later. Proc Hum Factors Ergon Soc Annu Meet 2006 Oct;50(9):904-908. [doi: 10.1177/154193120605000909]
21. Bangor A, Kortum PT, Miller JT. An empirical evaluation of the System Usability Scale. Int J Hum Comput Interact 2008 Jul;24(6):574-594. [doi: 10.1080/10447310802205776]

XSL•FO
RenderX

22. Ross S, Hamza D, Zulla R, Stasiuk S, Nichols D. Development of and preliminary validity evidence for the EFeCT feedback scoring tool. J Grad Med Educ 2022 Feb;14(1):71-79. [doi: 10.4300/JGME-D-21-00602.1] [Medline: 35222824]
23. Saurí R, Pustejovsky J. Are you sure that this happened? Assessing the factuality degree of events in text. Comput Linguist 2012 Jun;38(2):261-299. [doi: 10.1162/COLI_a_00096]
24. Prabaswari AD, Basumerda C, Utomo BW. The mental workload analysis of staff in study program of private educational organization. IOP Conf Ser Mater Sci Eng 2019;528:012018. [doi: 10.1088/1757-899X/528/1/012018]
25. Moura LM, Mishuris RG, Metlay JP, et al. Hybrid ambient clinical documentation and physician performance: work outside of work, documentation delay, and financial productivity. J Gen Intern Med 2026 Apr;41(5):1294-1303. [doi: 10.1007/s11606-025-09979-5] [Medline: 41249645]
26. Bing-You R, Varaklis K, Hayes V, Trowbridge R, Kemp H, McKelvy D. The feedback tango: an integrative review and analysis of the content of the teacher-learner feedback exchange. Acad Med 2018 Apr;93(4):657-663. [doi: 10.1097/ACM.0000000000001927] [Medline: 28991848]
27. Lee GB, Chiu AM. Assessment and feedback methods in competency-based medical education. Ann Allergy Asthma Immunol 2022 Mar;128(3):256-262. [doi: 10.1016/j.anai.2021.12.010] [Medline: 34929390]
28. Li SX, Li CMF, Jenkins ME, Venance SL, Florendo-Cumbermack A. Insights from the transition to competency-based medical education in neurology programs. Can J Neurol Sci 2023 Nov 23:1-5. [doi: 10.1017/cjn.2023.318] [Medline: 37994542]
29. Young JQ, Van Merrienboer J, Durning S, Ten Cate O. Cognitive load theory: implications for medical education: AMEE guide no. 86. Med Teach 2014 May;36(5):371-384. [doi: 10.3109/0142159X.2014.889290] [Medline: 24593808]
30. Wright DS, Kanaparthy NS, Melnick ER, et al. The effect of ambient artificial intelligence scribes on trainee documentation burden. Appl Clin Inform 2025 Aug;16(4):872-878. [doi: 10.1055/a-2647-1142] [Medline: 40602775]
31. Olson KD, Meeker D, Troup M, et al. Use of ambient AI scribes to reduce administrative burden and professional burnout. JAMA Netw Open 2025 Oct 1;8(10):e2534976. [doi: 10.1001/jamanetworkopen.2025.34976] [Medline: 41037268]
32. Debnath T, Siddiky MN, Rahman ME, et al. A comprehensive survey of prompt engineering techniques in large language models. TechRxiv. Preprint posted online on Oct 28, 2025. [doi: 10.36227/techrxiv.174140719.96375390/v1]

Abbreviations

AI: artificial intelligence
EFeCT: Evaluation of Feedback Captured Tool
LLM : large language model
SP: standardized patient

Edited by M Montagna; submitted 19.Dec.2025; peer-reviewed by F Fukuzawa, O Ng, Y Asada; revised version received 09.Apr.2026; accepted 29.Apr.2026; published 28.May.2026.

Please cite as:
Talwalkar JS, Chartash D, Zhang L, Makutonin M, Safranek CW, Sidamon-Eristoff AE, Schwamm LH, Wright DS
Ambient AI Scribes to Create Educational Feedback Notes for Medical Students: Randomized Trial
JMIR Med Educ 2026;12:e89996
URL: https://mededu.jmir.org/2026/1/e89996
doi:10.2196/89996

© Jaideep S Talwalkar, David Chartash, Lisa Zhang, Michael Makutonin, Conrad W Safranek, Anne Elizabeth Sidamon-Eristoff, Lee H Schwamm, Donald S Wright. Originally published in JMIR Medical Education (https://mededu.jmir.org), 28.May.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Education, is properly cited. The complete bibliographic information, a link to the original publication on https://mededu.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Development of Therapeutic Competencies in Health Care Students: Qualitative Focus Group Study Using 360-Degree Video and Virtual Reality Technology

Camilla Lauritzen[1], PhD; Charlotte Reedtz[1], PhD; Kjersti Bergum Kristensen[1], PhD; Vår Mathisen[2], PhD; Eva Therese Næss[3], ClinPsy; Rigmor Furu[4], MD; Hege Nermo[5], PhD; Rita Jentoft[2], MD

[1]Regional Center for Child and Youth Mental Health, Faculty of Health Sciences, UiT The Arctic University of Norway, Tromso, Norway

[2]Department of Health and Care Sciences, Faculty of Health Sciences, UiT The Arctic University of Norway, Tromso, Norway

[3]Department of Psychology, Faculty of Health Sciences, UiT The Arctic University of Norway, Tromso, Norway

[4]Centre for Health Education Development, Faculty of Health Sciences, UiT The Arctic University of Norway, Tromso, Norway

[5]Department of Dentistry, Faculty of Health Sciences, UiT The Arctic University of Norway, Tromso, Norway

Corresponding Author:
Camilla Lauritzen, PhD
Regional Center for Child and Youth Mental Health
Faculty of Health Sciences
UiT The Arctic University of Norway
Postboks 6050 Stakkevollan
Tromso, 9037
Norway
Phone: 47 77645871
Email: camilla.lauritzen@uit.no

Abstract

Background: Therapeutic competence is a critical skill for health care professionals, encompassing communication, interaction, and guidance in vulnerable situations. Virtual reality (VR) and 360-degree video technologies have emerged as innovative tools in health care education, offering immersive and interactive learning experiences. However, there is limited research on their effectiveness in developing therapeutic competencies among health care students.

Objective: This pilot study aimed to explore the feasibility, usability, and perceived educational value of a virtual learning resource using VR and 360-degree video to enhance therapeutic competence in health care students.

Methods: A virtual learning resource was developed, consisting of three modules: (1) a virtual home visit, (2) observation of therapeutic conversations using a 360-degree video, and (3) practice of therapeutic conversations in a simulated environment using VR. The resource was tested with students (n=12) from occupational therapy, psychology, and dentistry programs. Data were collected through focus group interviews after the students completed the modules. Thematic analysis was conducted to identify key themes related to the educational value and learning outcomes of the resource.

Results: The analysis revealed four key themes: (1) active exploration, where students engaged deeply with the material and contextualized theoretical knowledge; (2) observation, which provided practical insights into therapeutic conversations; (3) practice and reflection, which allowed students to refine their skills and build confidence; and (4) translation of theoretical knowledge into practical skills. Students reported that the resource was engaging, immersive, and effective in promoting learning compared to traditional teaching methods. Some students found the VR experience intense but valuable for skill development.

Conclusions: This pilot study demonstrates the feasibility and potential educational value of integrating VR and 360-degree video into health care education. The findings provide preliminary insights into the resource's ability to enhance therapeutic competence and student engagement. Future research should focus on larger, multi-institutional studies to validate these findings and assess the resource's impact on measurable learning outcomes.

(JMIR Med Educ 2026;12:e75776) doi:10.2196/75776

KEYWORDS

health care education; virtual learning resources; virtual reality; therapeutic competence; immersive learning

XSL•FO
RenderX

Introduction

Background

Students pursuing health professional education are preparing for professions where they will interact with people in vulnerable life situations. Therefore, it is crucial that these students develop knowledge and skills in interaction, communication, and guidance during their education. In the context of health professions, this can be described as developing therapeutic competencies. Therapeutic competence can be defined as specific skills in conversations and interactions with patients and their families. These skills can be acquired through interactions between students and teachers [1].

Therapeutic competence encompasses a range of skills, including empathy, active listening, and the ability to adapt communication to the patient's needs and context. It also involves understanding the broader social and familial network of the patient, as well as the impact of the patient's condition on their family. In this paper, we explore how the development of therapeutic competence can be enhanced through an interactive virtual learning resource, using 360-degree video and virtual reality (VR) technology in teaching. VR and 360-degree videos have been increasingly recognized as effective tools in health care education. These technologies provide immersive and interactive learning experiences, allowing students to practice skills in a safe and controlled environment [2,3].

The learning resource is centered around supporting a family where one of the parents is experiencing mental illness. However, the methodology and therapeutic approach are quite generic and can be applied to several other severe conditions affecting a family, such as parental cancer, injuries, or substance use disorders.

Therapeutic competence also refers to the ability to assess and understand the patient's overall situation as part of a larger social and familial network. This holistic assessment includes the patient's roles and responsibilities, which may involve caring for children. An important aspect of this competence is understanding how the patient's illness affects their family, particularly their children, and adapting health care to the patient's life situation. Children whose parents experience significant illnesses, such as mental illness or substance use disorders, may require support, information, and follow-up. This perspective on therapeutic responsibilities expands the unit of care provision from a narrow focus on the patient to the wider family and caregiving system and is commonly referred to as a "family focus" in health services [4].

In Norway, health care personnel have a legal obligation to identify whether patients have children, following the revision of the Health Personnel Act in 2010 [5]. This legal provision requires health care personnel providing care to adult patients not only to clarify whether the patient has minor children but also to assess what information or support the family might need. The legislation applies to all illnesses that might impact everyday functioning and parenting, not just mental illnesses [5]. However, extensive national evaluations show that efforts to change practices in hospitals so that all patients with children are assessed and offered support are progressing very slowly [6-9]. It has also been revealed that children are rarely included in conversations with health care personnel when parents are severely ill [10]. The challenges of successfully including children are related to obstacles both within families themselves and within the health care system [10,11]. Internationally, several studies document similar issues. Health care personnel across professions struggle to establish systematic routines that encompass the family situation of the patient [12-14].

Lack of time, competence, and competing tasks are among the reasons why many children do not receive the support and information they need when their parents are severely ill [11]. Part of the explanation may be that health care personnel perceive this work as additional rather than an integrated part of health care [9,11]. Additionally, health care personnel find it challenging to talk to children about difficult topics, such as substance abuse and mental illness in their parents [15].

Acquiring skills in communication, interaction, and guidance for families is thus central to the education of health care personnel. Health care professionals frequently encounter situations that demand therapeutic competence [16]. This requires an understanding of the situation and the ability to interact with the patient in an appropriate manner [17]. Therefore, it is essential that education focuses on developing therapeutic competencies, particularly skills in communication and interaction [16].

Approaches to Learning Therapeutic Competencies

Emphasizing the active role of students in knowledge development and applying educational approaches that encompass interactivity is an important starting point for developing therapeutic competence [1]. Chi and Wylie [18] describe 4 learning modes, namely passive, active, constructive, and interactive, which were tested in a 5-year project. The results showed that constructive and interactive activities, involving interaction, increased learning outcomes. The students who participated learned significantly more through interactive learning forms compared to traditional 1-way communication from teacher to student [2,3].

Observation can also be an important learning arena for students training to become health care professionals. Studies have shown that health care personnel lack competence in conducting professional conversations with children [4,5], and many are unsure of how to proceed in conversations with families [6,7]. However, by observing family and child conversations, students can gain knowledge about how such conversations can be conducted, although this is challenging to achieve in practice with families in highly vulnerable life situations. Learning through observation is often referred to as the apprenticeship model, which primarily involves the master first explaining or demonstrating how to perform various tasks, followed by beginners observing and trying the tasks themselves [8,9].

Interactive learning approaches and observation should be followed by a component of reflection to optimize the learning of new skills. Reflection is considered an essential component of the learning process, especially when it comes to acquiring

XSL•FO
RenderX

new skills in practice [10]. Reflection involves a conscious and systematic assessment of one's own experiences, actions, and results.

Another aspect to consider is that in the field of children of parents with severe illness, encounters with families where one or both parents face challenges due to illness may reveal children who are not receiving adequate care at home. In such situations, health care professionals may have a legal obligation to report these cases to agencies, such as social services. However, studies have shown that health care personnel largely underreport neglect [11]. To fulfill their reporting responsibilities, it is important that students training to become health care professionals are also trained to recognize risk factors and signs of neglect when interacting with families where serious illness is involved. Reflection promotes a deeper understanding, as it develops the ability to think analytically and solve problems creatively [12]. Additionally, reflection can help build confidence in the role of a health care professional, as one becomes more aware of one's own competence and potential for development.

Health care personnel report that it can be difficult to ask about sensitive topics, and they are often unsure of what to do with the information they receive [13]. A final aspect that should be emphasized when students are acquiring therapeutic competence and competence in family-focused treatment is the opportunity to practice. To develop skills in asking difficult questions and responding appropriately, it is necessary to obtain practice.

It is ethically challenging to allow students to practice on real patients as interns, especially when it involves complicated life challenges and illness profiles in patients. Practicing through role-play is common, and role-play can be a good way to train professionals for situations that will arise in real life. However, it can be challenging to get all students to participate in role-play in a way that provides learning outcomes because many may find it demanding to be in focus [14]. Furthermore, topics addressed through role-play can trigger negative emotions that may lead the learning process in an undesirable direction [15]. Practicing in simulated and virtual situations can thus be a good alternative. By using virtual learning resources, health science students can practice, among other things, establishing contact and building alliances in an arena where the students' skills have no consequences for the patient, as they might in an internship situation. Additionally, the virtual situation provides greater opportunities for repeated practice.

Study Objectives and Research Questions

This pilot study explores the feasibility, usability, and perceived educational value of a virtual learning resource designed to enhance therapeutic competence. The study aims to provide preliminary insights that can inform the development of larger-scale evaluations and more robust research designs in the future.

The research question in the study is "How can a virtual learning resource promote knowledge and skills in therapeutic competencies when interacting with patients who have minor children?"

Methods

Overview

The current project aimed to strengthen therapeutic competence in students training to become health professionals by developing virtual and interactive learning resources. The project received funding from the Norwegian Directorate of Higher Education and Skills to develop and test the learning resource. Alongside the development and testing of the resource, an approach inspired by action research was adapted, in the sense that researchers were closely involved in the development and testing process [19]. An important prerequisite was to facilitate systematic development and improvement based on students' experiences with using the learning resource.

While the study was inspired by action research principles, it does not fully align with the iterative cycles typically associated with action research. Instead, the study focused on piloting and refining the learning resource based on student feedback.

The Learning Resource: Design, Structure, and Learning Objectives

Learning Objectives

To strengthen therapeutic competence development in assessing and offering support to patients who have children, we developed a virtual learning resource for students training to become health care professionals. The objective of the learning resource is to give students and health care personnel insight into family-focused practice and to provide knowledge about how children are affected by their parents' illness. Furthermore, the learning objective is to teach students what kind of information they need to understand the parents' illness and its consequences. Additionally, the resource provides an example of how to conduct conversations with parents and children of different ages. The learning resource facilitates an arena for students to practice and reflect on their own performance and encompasses student-active learning forms, observational learning, reflection, and practice. The learning resource consists of 3 modules, which are described below.

Modules

Module 1

In this module, students go on a virtual home visit to an example family, the Hansen family's house. This is done as preparatory work through a flipped-classroom approach. Students follow a learning path where they gradually gain more insight into the family's risk and protective factors. The learning resource provides room for practice and reflection, with interactive tasks and exploration of the case (the family) that encourages participants to assess and develop their own practice. The virtual home visit is easily done using a computer and does not require special equipment. Students walk through the family's home and acquire knowledge about both the topic area of children as relatives and the example family, the Hansens. This is done by exploring various clickable interactive elements placed in different rooms in the home.

XSL•FO
RenderX

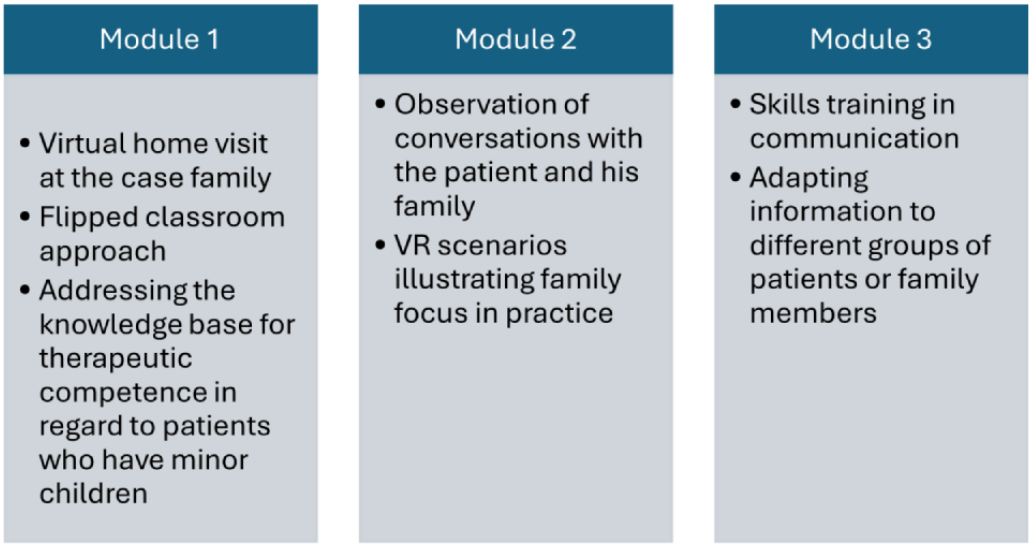
Module 2

In this module, students observe conversations with families using 360-degree video and VR headsets. VR refers to a technology that allows the user to interact in a computer-generated environment that mimics reality. VR for skill training has proven to be as effective in teaching as traditional exercises [3]. One of the unique strengths of using VR is that it can provide experiences that would otherwise not be available, such as observation in virtual rooms and practicing specific skills. In this module, the use of VR offers auditory and visual experiences that create a sense of presence in situations that would otherwise not be accessible and can thus enrich the learning process through realistic and engaging simulations.

Module 3

The final module opens opportunities for practicing conversations with the family and the children. With VR headsets on, students sit in the therapy room and interact with the family themselves. The 360-degree videos provide the experience of the family coming into the room to the student as a therapist. There is no dialogue in this module, but the family enters the room and looks directly at the student. It is then up to the student to engage in conversations with the family and the children of different ages. Specific tasks have been created for the module to help the students practice how to express themselves. The exercises for module 3 can be done alone or in groups of students. Figure 1 provides an overview of the entire learning resource.

Figure 1. Overview of the modules in the learning resource. VR: virtual reality.



Participants and Procedure in Testing the Learning Resource

The testing of the learning resource was conducted with 2 groups of students recruited from a faculty of health sciences. A total of 12 students participated, including 2 students each from the occupational therapy, psychology, and dentistry programs in each group. The students had limited knowledge of family-focused treatment before the testing. However, they had received some instruction in therapeutic competence and communication, mainly theoretical with elements of role-play. Module 1 was completed at home as preparation, while modules 2 and 3 were completed in a learning lab at the faculty. Based on the experiences from the first group, adjustments were made to the setup for test group 2.

The students used VR headsets for modules 2 and 3. The researchers were present during the entire testing period and took notes consecutively. After finishing the third module, 2 researchers interviewed the students in focus groups. The interviews were transcribed by a transcription service and quality assured by the researchers who conducted the interviews. Analyses of the initial interview revealed minor errors and weaknesses in the learning resource, which were then adjusted before test group 2 tried the resource. The results from test group 2 led to further fine-tuning of the learning resource, and a teacher guide was developed for pedagogical support. The learning resource and guide were then made available in Canvas Commons. Canvas Commons is an integrated content repository for Canvas that lets educators discover, share, and import course materials—like modules, assignments, and quizzes—across courses and institutions. The learning resource is published with open access in line with the principles of open education. We are currently in the process of translating the resource to English, making it freely available for use in international education.

Analyses

The text data in this study were analyzed thematically. Thematic analysis is based on selected themes by comparing information from all informants in a study [20]. Text data were systematically reviewed by the first author, and the content was sorted into categories by theme. The thematic analysis followed Braun and Clarke's 6-step framework, which includes familiarization with the data, generating initial codes, searching for themes, reviewing themes, defining and naming themes, and producing the report [21]. Codes were developed inductively based on recurring patterns in the focus group transcripts and

XSL•FO
RenderX

were iteratively refined by 2 researchers to ensure consistency and reliability.

The overarching thematic basis for the analysis has been the educational value and learning outcomes of the learning resource. Through the analysis of the interviews, it was identified how the learning resource contributed to learning by (1) exploration; (2) observation; (3) practicing, reflecting, and discussing; and (4) translation of theoretical knowledge into practical skills.

Ethical Considerations

The project was exempt from ethics review approval by the regional ethical committee, as it does not involve the collection of health information. Privacy considerations have been assessed and approved by SIKT (Norwegian Agency for Shared Services in Education and Research). All informants in the study have been presented with written information material and have signed informed consent forms. The participating students received a small compensation fee for taking part in the testing of the resource.

Results

Overview

In the analysis of the focus group interviews, several themes were identified that illustrate how a virtual learning resource can promote knowledge and skills in therapeutic competencies when interacting with patients who have minor children. Overall, the results showed that the learning resource engaged the students and contributed to an active exploration of the learning material, leading to a deeper understanding and better retention of information. Another overarching finding was that the learning resource was perceived by students as a more interactive and engaging learning experience than traditional teaching. Below, we delve into the 4 categories that emerged in the analysis. These were (1) learning through exploration, (2) learning through observation, (3) learning through practicing, reflection, and discussion, and (4) translating what has been learned into practical skills.

Learning Through Exploration

Students found that the virtual resources facilitated an active exploration process where they could engage directly with the material in an interactive way. Students described how the preparatory work in module 1 allowed them to explore various aspects of the family's situation, not just by passively reading, but by investigating and assessing the family situation and viewing the situation from different perspectives. The interactive exploration led to a deeper understanding of the topic and made it easier to remember what they had learned compared to traditional methods, for example, receiving a lecture with Microsoft PowerPoint on the subject matter. The dialogue below illustrates this:

> *Did it make it a bit easier to grasp the health legislation this way, compared to having something presented on a PowerPoint? You're nodding, yes.* [Interviewer]
>
> *Yes, that's a very good example, with the legislation... I hate when it's introduced on a slide. It's something I can't manage to learn when it comes on PowerPoint, whereas here I could almost remember the paragraph word by word afterward.* [Student]
>
> *I found it very helpful when the legislation was very connected to the situations and the examples. They were tailored to the situations you had either gone through or were going to go through, so it was a bit like: "okay, that applies here, and that applies there." It made much more sense. Because when you're reading it in a textbook, it's like: "Yes, it probably applies there." But when you could see why, it was way better.* [Student 2]
>
> *So, it was easier to remember it?* [Interviewer]
>
> *Yes, really.* [Student 2]
>
> *Speaking of what he said about retaining the information, because in our educational program, we constantly ask for clinical examples. Put it into context, give us something to attach it to, and don't just list an 80-page PowerPoint with legislation. We can't grasp that. But receiving something concrete, I think, can be used in many other contexts too, and it is so much easier to learn when you have something to attach it to.* [Student 1]

In both focus groups, there was broad agreement that being able to explore the topic themselves, in this case through a virtual home visit, was perceived as promoting learning. It made abstract theory about the consequences of being a child of parents with various life challenges, as well as training in health legislation, more concrete and comprehensible.

Some students found it somewhat intense to work with the topic in this way. The quote below illustrates this:

> *Before I came here, I had an expectation that it would be interesting and educational. I'm generally interested in therapeutic competence, communication forms, and such things. But I didn't think it would be as all-encompassing as it actually was. When I did module 1 on the PC at home yesterday, I even found it more educational than a PowerPoint. So, now in VR mode, I almost felt my brain boiling, because there was so much info. I started thinking like: "Who should I look at? Should I look at her or him? Who am I most curious about in this family?" And that's a good thing, but yes, very intensive learning. So, it's a great resource, but I noticed I was a bit caught off guard. I didn't know it would be so many stimuli and impressions.* [Student]

This finding shows that it can also be overwhelming to work with such topics in VR, and that teachers therefore must consider that various reactions may arise in the student group. Discomfort can also be a source of development and lead to students experiencing a high learning outcome. Being able to explore the case on their own and become familiar with the challenges the example family had was perceived as relevant and educational.

XSL•FO
RenderX

Another student expressed this as follows:

> *I think that often when we do case-based learning, the information that is important in the case is pointed out to you. But in this exploration, you get a lot of information to assess yourself. With being in such an interactive resource, you get the opportunity to choose for yourself what you want to focus your attention on. You practice much more on noticing the important things, compared to when you just get presented with something. When you have to look for it yourself, you also practice focusing your attention. So, I thought that was very cool.* [Student]

In the conversation about exploring the topic area on their own using the interactive, virtual home visit in module 1, the students reported that it had a high educational value. The students said that it resembled reality. When students are in internships or start practicing as healthcare personnel after their education is finished, they don't necessarily know what the patients' or families' issues are. They have to use their developing therapeutic competence and knowledge to assess risks and problem areas to work on moving forward. The overarching finding in this category was that the learning resource was perceived as a very practice-oriented and effective way to learn these skills.

Learning Through Observation

The students reported that observing therapeutic conversations and communication between the therapist, patient, and the patient's family was very educational. The observations allowed them to see the acquired theory applied in practice. Students discussed how they observed different communication styles in conversations with children versus adults and how the therapist adapted their language and methods according to the client's age and level of understanding. This provided students with a practical understanding of how to tailor therapeutic techniques to different client groups.

> *Observing these conversations can be used as a starting point for skills training and give us the opportunity to practice therapeutic competence aimed at children, adolescents, adults, and older adults. It allows us to see what we need to do as therapists to bring out the best possible information about their life situation, and in what way we can use ourselves as tools to acquire information.* [Student]

Another student expressed it as follows:

> *It was very educational because you get to see a real situation of how both the therapist and those coming to the therapist express themselves, and how their body language is. In my case, I learn a lot by viewing examples, like in an internship, by watching how my supervisor does it, and this was similar.* [Student]

The overarching finding from this category showed that students experienced high learning outcomes from observing a conversation between a therapist and a family.

Learning Through Practicing, Reflection, and Discussion

VR offers a unique opportunity not only to observe but also to actively engage through simulated therapeutic sessions. This leads to reflection and discussion, which are crucial for learning and developing therapeutic competence. Participants appreciated the opportunity to discuss and reflect on their actions and decisions in the VR scenarios, as shown in the dialogue below:

> *How did you find reflecting along the way? Useful or not useful?* [Interviewer]
>
> *I generally think just seeing that you have different ways of perceiving things, and that someone can make you aware of something you absolutely weren't aware of. We discovered that we disagreed on how we perceived the situation, and I think it's great to see different viewpoints.* [Student 1]
>
> *Yes, and it triggers a much longer thought process too.* [Student 2]
>
> *I thought it was very educational, even though I felt tired at the end.* [Student 3]
>
> *I thought the questions were good. The questions allowed us to reflect on several possibilities and perspectives instead of giving us the answer immediately.* [Student 4]

This excerpt from one of the focus group interviews highlights the value of questions that stimulate reflection. This helps them understand and internalize the learning material better.

Another important pedagogical value that emerged in the focus groups was the opportunities the VR resource provided for practice. The dialogue below illustrates this:

> *I noticed that for me to practice my therapeutic competence, I need a certain calm. Because I noticed that I didn't quite know where to start at first, because I was stressed by them [the family] staring at me. It takes practice, I think, because you have to be able to be in a certain situation several times to be able to do it well. So, I felt that I got into it the more I practiced.* [Student 1]
>
> *To get into a kind of flow. I started to think about my body language eventually. After a while, I started to think: "How am I sitting now, actually? Is it okay to sit like this?" But at first, I couldn't even think about what to say.* [Student 2]

The overarching result in this category is that students experienced good learning outcomes from trying out, practicing, reflecting, and discussing with each other using this learning resource.

Translating What Has Been Learned Into Practical Skills

Finally, the students discussed how they can apply what they have learned through VR in their professional practice. This includes applying knowledge and skills in real therapeutic situations, such as handling sensitive topics with clients or adapting communication styles based on the client's needs and

XSL•FO
RenderX

situation. Students reflected on how they can use insights from the learning resource to improve their practice, especially in how they interact with clients and handle complex and sensitive issues. The dialogue below illustrates this:

> *So, what do you think about conducting such a conversation now? What would it be like?* [Interviewer]
>
> *One of the things I'm a bit nervous about is exactly that. But it feels a bit easier now. I feel like I know more.* [Student]

Students also reported that they have become more aware of what the term therapeutic competence entails. They also mentioned that they had become more attentive to the whole family and the importance of involving relatives. The dialogue below illustrates this:

> *I'm also a bit curious about what you now consider to be therapeutic competence. What do you think now? Is it clear to you what it is?* [Interviewer]
>
> *It's a whole bag of things, with listening and hearing, tips, tricks, what works, what doesn't, reflecting, getting them to become aware of things, and not just like: "Pull yourself together, get well!"* [Student 1]
>
> *I became much more aware that maybe we should try to involve families more often.* [Student 2]

The overarching result in this category is that, in terms of translating theoretical knowledge into actual therapeutic competence, the students had positive learning outcomes by using the learning resource.

Discussion

Principal Findings

This pilot study explored the feasibility, usability, and perceived educational value of a virtual learning resource designed to enhance therapeutic competence in healthcare students. The findings suggest that the resource has the potential to promote active exploration, observation, and practice in a safe and controlled environment. Students reported that the resource was engaging and effective in promoting learning compared to traditional teaching methods. The study's findings align with existing literature on the benefits of interactive and immersive learning approaches in healthcare education [2,18].

The results also showed that the learning resource provided students with the opportunity to observe and practice therapeutic conversations in a safe environment. This aligns with the need to develop skills in adapting communication and interaction to different patient groups. These skills are crucial for responsible professional practice. VR technology offers a unique opportunity to learn through observing simulated, yet realistic, therapeutic conversations. VR technology can simulate complex interactions with patients and their families, providing students with some experience, yet avoiding the ethical concerns that may arise from practicing directly on patients. This is particularly relevant in education, where direct practice with real patients can be challenging or difficult to achieve, such as in the early stages of educating health professionals. Compared to role-play, where students themselves play the different roles, using films of relevant scenarios where students practice being a therapist who observes, listens, and interacts through VR technology appears more direct and thus professionally relevant than inventing thoughts, feelings, and behaviors in people they play themselves. Furthermore, the students developed new concepts and understanding through reflection and discussion of what they had observed in the scenarios. They could relate legislation and therapeutic techniques to practical situations, showing that they moved from concrete experience to abstract conceptualization [22]. Translating what they had observed into concrete skills seemed, therefore, easier using this learning resource.

The learning resource is designed for students to reflect on their observations, often referred to as reflective observation [22]. Using reflection as a key component in the learning process, the opportunity to reflect on their observations was provided, and the students appreciated this option. The students found that the questions in the learning resource stimulated deeper thinking, reflection, and discussion, which is crucial for developing a deeper understanding and ability to adapt to new challenges. By discussing how the VR experience allowed them to see theory applied in practice, they felt they gained a deeper understanding of the material. Reflection on what they observed, such as communication styles and the therapist's adaptation to the client's age, thus helped strengthen their learning.

The learning resource gave the students an opportunity to practice and refine their interaction skills in a controlled environment before applying them in real situations. Practice is essential for developing skills. Students reported that they experienced increased confidence and flow in communication after repeated exercises in the VR environment. This underscores the importance of practice for achieving mastery and confidence in therapeutic practice. Trying out new skills in a safe environment can be essential for active experimentation. The students reported that practice in VR helped them achieve flow and confidence in therapeutic conversations, which is a critical part of the learning process. Practicing in VR was perceived as activating and engaging, and the students found that the resource provided a deeper understanding as well as promoted the development of skills.

Translating theoretical knowledge into practical skills can be difficult to achieve with traditional learning methods. Students reported that the VR experience provided a more vivid and realistic understanding of the subject matter, contributing to more effective learning. Students described the VR experience as particularly valuable because it allowed them to "step into" scenarios relevant to their field of study. This feeling of being part of the scenario contributed to a more authentic and engaging learning experience. The students felt that visualization contributed to a better understanding of complex subject matter. The learning resource created a unique opportunity to explore, observe, and interact in complex situations in a way that text or images cannot. In light of Kolb's experiential learning cycle [22], one can say that the learning resource provides a concrete experience that is essential to start the learning cycle. Students could see body language, hear tone of voice, and observe interactions in a controlled but realistic environment. This helped them understand complex subject matter and integrate theoretical

XSL•FO
RenderX

and practical aspects of the therapeutic competence they will need in practice. Through exploration, observation, reflection, and discussion, the learning resource thus contributes to the acquisition of experience, laying the foundation for the development of therapeutic competence.

VR-based practice offers significant advantages for developing foundational therapeutic communication skills, such as active listening, empathy, and adapting communication to different client groups. The immersive nature of VR allows students to observe and practice interactions in realistic scenarios, which can enhance their confidence and readiness for real-world practice [23]. However, certain aspects of authentic therapeutic interaction, such as emotional attunement, relational pacing, and the influence of cultural and familial contexts, may be more challenging to replicate in a virtual environment. These elements often require nuanced, real-time adjustments that are best developed through direct interaction with patients and guided supervision. To address these limitations, the learning resource was complemented with structured reflection activities that allowed students to explore the subtleties of therapeutic relationships in greater depth.

Therapeutic communication is a complex skill that develops progressively throughout a health care professional's education and career [1,16]. For students in the early stages of their training, foundational skills, such as active listening, empathy, and basic conversational techniques, are critical [18,22]. The virtual learning resource aligns with these developmental needs by providing a safe and controlled environment for students to practice these foundational skills. Unlike residents or more experienced practitioners, students may have limited exposure to real-life therapeutic interactions, making simulated environments particularly valuable for bridging the gap between theoretical knowledge and practical application. Future studies could explore how the resource might be adapted for learners at different stages of their professional development, including residents and practicing clinicians, to address more advanced aspects of therapeutic communication.

Equity and fairness are critical considerations in the implementation of VR-based learning environments. Learners come from diverse cultural, linguistic, and socioeconomic backgrounds, which can influence their communication styles and access to technology. Similarly, patients and families bring unique cultural and contextual factors that shape their communication needs and expectations. To support equitable outcomes, the learning resource could be adapted to include scenarios that reflect a wide range of cultural and familial contexts. Additionally, structured reflection activities and supervisory guidance could help learners develop cultural competence and inclusive communication practices. Future iterations of the resource should also consider accessibility features, such as subtitles or alternative formats, to ensure that all learners can fully engage with the material.

Practical Implications

Although VR can have significant educational benefits, several challenges must be addressed to maximize educational value and learning outcomes. It is important to consider and assess the ethical aspects of using VR in education, because simulations of sensitive topics can have unforeseen consequences. Teachers using such learning resources should be prepared for emotional reactions and have a plan for how to handle such reactions. As with all teaching on sensitive topics that are perceived as challenging for students, it requires educators to help students navigate these challenges by providing support and guidance throughout the learning process and through supervision.

User-friendliness and technical support are important prerequisites when applying virtual resources to teaching situations [23]. Participants in the focus groups reported the need for technical support and guidance when using VR, emphasizing the importance of accessible and effective technical assistance. Another challenge is the cost and availability of VR technology. Although the prices of VR equipment have fallen in recent years, costs can still be a barrier for many educational institutions. Furthermore, effective use of VR requires appropriate software and updates, which can also incur significant costs. It also requires some logistics to ensure that equipment is charged and updated with new software so that it can be easily used.

It is also important to consider how VR can be integrated and systematically implemented in educational programs to maximize learning outcomes. To ensure that VR technology becomes an integrated part of educational programs, educational institutions must consider both infrastructural and pedagogical adjustments. This requires anchoring with management, investment in necessary technology, organization of training, use, and guidance at the institution. By using VR technology, however, one can achieve more targeted competence development. In many health professions, VR scenarios can be tailored to meet the need for developing specific clinical and therapeutic skills.

Limitations

This pilot study involved a small sample size (n=12) and a single-case context, which limits the generalizability of the findings. The study was designed to provide preliminary insights into the feasibility and perceived educational value of the virtual learning resource, rather than to draw definitive conclusions about its effectiveness. Future research should include larger, more diverse student groups and multi-institutional studies to validate the findings and assess the resource's impact on measurable learning outcomes. Additionally, comparative studies with control groups using traditional teaching methods would provide a more robust evaluation of the resource's relative advantages and limitations.

While the study was inspired by action research principles, it did not fully adhere to the iterative cycles typically associated with action research. Instead, the study focused on piloting and refining the learning resource based on student feedback. Future research could adopt a more comprehensive action research approach, incorporating multiple iterations and cycles of feedback to further enhance the resource and its implementation.

Conclusion

This pilot study demonstrates the feasibility and potential educational value of integrating VR and 360-degree video into health care education. The participating students experienced

XSL•FO
RenderX

that the virtual learning resource promoted the learning of therapeutic competence. Interactive learning, practice, and reflection were important elements that contributed to the students' development of skills necessary to meet the demands of health professions. The findings provide preliminary insights into the resource's ability to enhance therapeutic competence and student engagement. Future research should focus on larger, multi-institutional studies to validate these findings and assess the resource's impact on measurable learning outcomes.

Acknowledgments

The authors would like to acknowledge The Resource Center for Teaching, Learning, and Technology (RESULT) at UiT - The Arctic University of Norway. Furthermore, we would like to acknowledge the students who participated in testing and developing the learning resource. We used a locally adapted large language model for translation and language editing purposes. The model used is named chat.uit.no and is based on OpenAI ChatGPT-4.

Funding

The project received funding from the Norwegian Directorate for Higher Education and Skills: 2019/10223.

Data Availability

The datasets generated or analyzed during this study are not publicly available because they consist of qualitative interview data that cannot be adequately anonymized, and written consent to store the data in an open repository was not obtained prior to the study.

Authors' Contributions

Conceptualization: CL Data curation: CL, ETN, RF, RJ Formal analysis: CL Funding acquisition: CL, RF, RJ Project administration: CL, RF, RJ Writing – original draft: CL Writing – review & editing: CL, CR, KBK, VM, ETN, RF, HN, RJ

Conflicts of Interest

None declared.

References

1. Jentoft R, Mathisen V. , & Mathisen, V. In: Arntzen C, Gramstad A, Jentoft R, editors. Ergoterapi - mangfold og muligheter (Occupational Therapy: Diversity and Opportunities). Oslo: Orcana Academic Publisher; 2015.
2. Blair C, Walsh C, Best P. Immersive 360° videos in health and social care education: a scoping review. BMC Med Educ 2021;21(1):590 [FREE Full text] [doi: 10.1186/s12909-021-03013-y] [Medline: 34819063]
3. Kaplan AD, Cruit J, Endsley M, Beers SM, Sawyer BD, Hancock PA. The effects of virtual reality, augmented reality, and mixed reality as training enhancement methods: a meta-analysis. Hum Factors 2021;63(4):706-726. [doi: 10.1177/0018720820904229] [Medline: 32091937]
4. Foster K, O'Brien L, Korhonen T. Developing resilient children and families when parents have mental illness: a family-focused approach. Int J Ment Health Nurs 2012;21(1):3-11. [doi: 10.1111/j.1447-0349.2011.00754.x] [Medline: 21692961]
5. Parr J. Om Lovdata. Lov og Rett 1989;28(7):330-331 [FREE Full text] [doi: 10.18261/issn1504-3061-1989-07-06]
6. Lauritzen C, Reedtz C, Rognmo K, Nilsen MA, Walstad A. Identification of and support for children of mentally ill parents: a 5 year follow-up study of adult mental health services. Front Psychiatry 2018;9:507 [FREE Full text] [doi: 10.3389/fpsyt.2018.00507] [Medline: 30386268]
7. Lauritzen C, Kolmannskog AB, Iversen AC. Family assessment conversations as a tool to support families affected by parental mental illness: a retrospective review of electronic patient journals. Int J Ment Health Syst 2018;12:18 [FREE Full text] [doi: 10.1186/s13033-018-0199-x] [Medline: 29760769]
8. Reedtz C, Jensaas E, Storjord T, Kristensen KB, Lauritzen C. Identification of children of mentally ill patients and provision of support according to the norwegian health legislation: a 11-year review. Front Psychiatry 2021;12:815526 [FREE Full text] [doi: 10.3389/fpsyt.2021.815526] [Medline: 35095621]
9. Ruud T, Birkeland B, Faugli A, et al. Barn som pårørende: resultater fra en multisenterstudie [Children of ill parents: results from a multi-center study]. Fafo. 2015. URL: https://www.fafo.no/publikasjoner/eksterne-rapporter/barn-som-parorende [accessed 2025-12-29]
10. Kristensen KB, Lauritzen C, Reedtz C. Support for children of parents with mental illness: an analysis of patients' health records. Front Psychiatry 2022;13:778236 [FREE Full text] [doi: 10.3389/fpsyt.2022.778236] [Medline: 35356710]

XSL•FO
RenderX

11. Lauritzen C, Reedtz C, Van Doesum K, Martinussen M. Factors that may facilitate or hinder a family-focus in the treatment of parents with a mental illness. J Child Fam Stud 2015;24(4):864-871 [FREE Full text] [doi: 10.1007/s10826-013-9895-y] [Medline: 25814823]
12. Laletas S, Reupert A, Goodyear M, Morgan B. Pathways of care: targeting the early childhood sector for early intervention. Adv Ment Health 2015;13(2):139-152. [doi: 10.1080/18387357.2015.1066293]
13. Laletas S, Goodyear M, Reupert A. Parental mental illness: cross-sectional analysis of family focused practice within the early childhood sector. J Child Fam Stud 2017;27(5):1650-1660. [doi: 10.1007/s10826-017-0992-1]
14. Grant A, Reupert A, Maybery D, Goodyear M. Predictors and enablers of mental health nurses' family-focused practice. Int J Ment Health Nurs 2019;28(1):140-151. [doi: 10.1111/inm.12503] [Medline: 29949231]
15. Lauritzen C, Reedtz C. Support for children of service users in Norway. Mental Health Practice 2013;16(10):12-18. [doi: 10.7748/mhp2013.07.16.10.12.e875]
16. Mota P, Selby K, Gouveia A, Tzartzas K, Staeger P, Marion-Veyron R, et al. Difficult patient-doctor encounters in a Swiss university outpatient clinic: cross-sectional study. BMJ Open 2019;9(1):e025569 [FREE Full text] [doi: 10.1136/bmjopen-2018-025569] [Medline: 30782753]
17. Weingarten M, Guttman N, Abramovitch H, Margalit RS, Roter D, Ziv A, et al. An anatomy of conflicts in primary care encounters: a multi-method study. Fam Pract 2010;27(1):93-100. [doi: 10.1093/fampra/cmp082] [Medline: 19948564]
18. Chi M, Wylie R. The ICAP framework: linking cognitive engagement to active learning outcomes. Educ Psychol 2014;49(4):219-243 [FREE Full text] [doi: 10.1080/00461520.2014.965823]
19. Ulvik M, Riese H, Roness D. Å Forske På Egen Praksis. Aksjonsforskning Og Andre Tilnærminger Til Profesjonell Utvikling I Utdanningsfeltet [Researching One's Own Practice: Action Research and Other Approaches to Professional Development in the Field of Education]. Bergen: Fagbokforlaget; 2016.
20. Thagaard T. Systematikk og innlevelse. In: En Innføring I Kvalitative Metoder [Systematics and Empathy: an Introduction to Qualitative Methods]. Bergen: Fagbokforlaget; 2018.
21. Braun V, Clarke V. Using thematic analysis in psychology. Qual Res Psychol 2008;3(2):77-101 [FREE Full text] [doi: 10.1191/1478088706qp063oa]
22. Kolb DA. Experiential Learning: Experience as the Source of Learning and Development. New Jersey: Pearson Education; 2015.
23. Lie SS, Helle N, Sletteland NV, Vikman MD, Bonsaksen T. Implementation of virtual reality in health professions education: scoping review. JMIR Med Educ 2023;9:e41589 [FREE Full text] [doi: 10.2196/41589] [Medline: 36692934]

Abbreviations

SIKT: Norwegian Agency for Shared Services in Education and Research
VR: virtual reality

Edited by AH Sapci, MD; submitted 10.Apr.2025; peer-reviewed by A Ugbor, R Ayub; comments to author 15.Jun.2025; revised version received 20.Nov.2025; accepted 14.Dec.2025; published 02.Feb.2026.

Please cite as:
Lauritzen C, Reedtz C, Kristensen KB, Mathisen V, Næss ET, Furu R, Nermo H, Jentoft R
Development of Therapeutic Competencies in Health Care Students: Qualitative Focus Group Study Using 360-Degree Video and Virtual Reality Technology
JMIR Med Educ 2026;12:e75776
URL: https://mededu.jmir.org/2026/1/e75776
doi:10.2196/75776
PMID:

©Camilla Lauritzen, Charlotte Reedtz, Kjersti Bergum Kristensen, Vår Mathisen, Eva Therese Næss, Rigmor Furu, Hege Nermo, Rita Jentoft. Originally published in JMIR Medical Education (https://mededu.jmir.org), 02.Feb.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Education, is properly cited. The complete bibliographic information, a link to the original publication on https://mededu.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Comparing the Weighted Gain Score and a Rasch-Based Approach for Estimating Learning Outcomes in Medical Education: Quantitative Study

Rauf Aliyev[1], MD; Joy Backhaus[1], MSc; Silke Hammer[2], MD; Sarah König[1], MME, MD

[1]Institute of Medical Teaching and Medical Education Research, University Hospital Würzburg, Josef-Schneider-Str. 2/D6, Würzburg, Germany
[2]Institute of Diagnostic and Interventional Radiology, University Hospital Würzburg, Würzburg, Germany

Corresponding Author:
Sarah König, MME, MD
Institute of Medical Teaching and Medical Education Research, University Hospital Würzburg, Josef-Schneider-Str. 2/D6, Würzburg, Germany

Abstract

Background: Pretest-posttest designs are widely used to estimate learning gain in studies evaluating educational interventions in medical education. The Weighted Gain Score (WGS) was proposed to reduce bias associated with differences in baseline performance.

Objective: This study evaluated the statistical and inferential properties of the WGS by comparing it to Rasch Learning Gain (RLG) across 3 datasets.

Methods: The WGS implements a weighting coefficient that includes the parameter μ, which linearly rescales the difference between pretest and posttest percentage scores. We examined the effect of varying μ (30, 50, and 70) on learning gain calculations and compared the results with those obtained using RLG. The following three datasets were analyzed: (1) a small illustrative dataset demonstrating the mathematical behavior of the WGS, (2) an empirical dataset from a previous educational evaluation study, and (3) a randomly generated binomial dataset designed to examine the metric under larger sample conditions.

Results: Changing the parameter μ in the WGS affected the magnitude of the calculated learning gains: lower μ-values produced larger gain estimates, whereas higher μ-values produced smaller estimates. Despite these differences in scale, the WGS and RLG correlated strongly in both the empirical dataset (r=0.93; P<.001) and the simulated dataset (r=0.92; P<.001); variation in μ did not alter the inferential results. Both methods identified the same interaction effect in the empirical dataset.

Conclusions: The WGS produced results highly consistent with those of RLG while requiring substantially lower computational complexity. The metric can be applied to both small and large datasets and allows μ to function as an adjustment coefficient for calibrating learning gain estimates across cohorts without altering inferential conclusions.

(JMIR Med Educ 2026;12:e75516) doi:10.2196/75516

KEYWORDS

medical education; teaching quality; curriculum evaluation; learning gain; pretest-posttest design; Rasch model; Weighted Gain Score

Introduction

Teaching quality in medical education is a complex construct encompassing curriculum design, instructional methods, teaching expertise, learner engagement, and assessment practices [1-4]. High-quality teaching in this context contributes to the development of competent physicians and thereby influences the quality of patient care [5,6].

Among the various aspects of teaching quality in medical education, student learning outcomes represent one measurable indicator frequently used in program evaluation and educational research [7-11]. However, interpreting learning outcomes as indicators of teaching effectiveness requires caution, as they are influenced by multiple factors beyond instructional quality. These include student motivation, prior knowledge, learning strategies, teacher enthusiasm, and learning activities occurring outside the formal curriculum [12,13]. To account for these influences, educational research often focuses not only on absolute performance but also on changes in performance over time. The concept of learning gain represents a widely used approach to capturing students' learning progress. In educational research, learning gain is commonly operationalized by assessing students before (pretest) and after (posttest) an educational intervention. The difference between pretest and posttest scores is then interpreted as an indicator of learning gain attributable, at least in part, to the educational intervention [14-16]. However, calculating learning gain is not trivial, as

XSL•FO
RenderX

simple difference scores may lead to biased estimates depending on students' baseline knowledge. One simple approach is raw gain, which is calculated as the arithmetic difference between posttest and pretest scores. However, raw gain scores exhibit a negative correlation with baseline performance (ie, pretest scores) and are also affected by ceiling effects, meaning that students with lower pretest scores may appear to exhibit larger gains simply because they have more room for improvement [17-19].

To address these limitations, several modified gain metrics have been proposed. One widely used approach is the normalized gain introduced by Hake [20], which expresses the observed pretest-posttest gain relative to the maximum possible gain. Although this metric has been applied extensively in educational research [21,22], it also has important methodological limitations. It remains dependent on baseline performance, may inflate gains for students with high pretest scores, and behaves inconsistently when posttest scores fall below pretest scores or when pretest scores approach the maximum value [16,23].

Taken together, existing gain metrics may distort estimates of learning gain, particularly in cohorts with heterogeneous baseline knowledge. Many of these metrics either remain strongly dependent on baseline performance or require complex psychometric modeling. This highlights the need for approaches that provide statistically robust yet practically applicable estimates of learning gain in educational evaluation.

A recently proposed metric developed by our workgroup, the "Weighted Gain Score" (WGS), aims to address these limitations by applying a weighting coefficient that adjusts gain calculations according to students' baseline performance [16]. However, the statistical and inferential properties of this metric have not yet been systematically investigated. To address this gap, we evaluated the WGS by comparing it with Rasch Learning Gain (RLG), a Rasch model–based approach for estimating learning gain that served as the benchmark in our study [24]. Specifically, we addressed the following research questions:

- Does the WGS produce inferential results comparable to those produced by RLG?
- Can the parameter μ in the WGS be adjusted for different cohorts to calibrate learning gain calculations without altering inferential conclusions?

Through this analysis, we aimed to clarify the statistical behavior of the WGS and explore its potential applicability for the evaluation of educational interventions.

Methods

Metric WGS

The mathematical foundation of the WGS lies in the use of the weighting coefficient "pre/μ," which linearly transforms the difference between pretest and posttest percentage scores (denoted as "pre" and "post" in equation 1), thereby adjusting for pretest variability [16]. Formally, the WGS is defined as:

$$(1) WGS=(post-pre)\times(pre/\mu)$$

To illustrate the computation, consider a hypothetical student with a pretest score of 40% and a posttest score of 70%.

For μ=50, the WGS is calculated as: WGS = (70 – 40) × (40/50) = 30 × 0.8 = 24.

If μ is increased to 70, the same performance yields: WGS = (70 – 40) × (40/70) = 30 × 0.57 = 17.14.

This example illustrates that increasing μ reduces the magnitude of the calculated gain while preserving the relative ordering of observations. When posttest scores fall below pretest scores, the WGS assumes negative values, indicating a decrease in performance.

Originally, the parameter μ used in the weighting coefficient was defined as the average pretest score of a cohort. It was constrained to integer values between 1 and 100, consistent with the percentage format of "pre" and "post." In the original formulation, its value was set at 50 as a default reference value [16]. In this study, μ is interpreted as an adjustment coefficient that functions as a scaling parameter for learning gain calculations. Changing its value proportionally rescales the calculated gain scores: higher values of μ lead to smaller gain estimates, whereas lower values produce larger gain estimates. Importantly, this modification represents a linear transformation of the calculated values and therefore does not alter the underlying statistical relationships among observations.

To examine the influence of this parameter on the stability of the WGS, we tested 3 calibration levels in our datasets: μ=30, μ=50, and μ=70. These values represent 3 nonextreme points within the possible range of 1 to 100, allowing us to evaluate the behavior of the WGS across low, moderate, and high scaling conditions.

Rasch Model and RLG

The Rasch model is a fundamental concept in modern psychometric measurement. The probability that a student answers a specific item correctly depends on 2 key factors: the student's ability and the difficulty of the item. In the Rasch framework, a student's latent ability is denoted by θ, whereas item difficulty is represented by β. When a student's ability exceeds the difficulty of an item, the probability of answering correctly increases, and vice versa [25]. Because the Rasch model allows the estimation of individual students' abilities independently of the specific test items used, it is widely applied in educational measurement and medical education research [26]. With this in mind, we selected RLG as a reference method for evaluating the WGS.

We applied the dichotomous 1-parameter logistic Rasch model. Item parameters were estimated using conditional maximum likelihood estimation. On the basis of the fitted model, person abilities were subsequently calculated using maximum likelihood estimation separately for the pretest (θ_{pre}) and posttest (θ_{post}) data [27].

As indicated in equation 2, RLG was defined as the difference between the estimated posttest and pretest abilities. This difference represents the change in latent ability on the Rasch measurement scale and serves as an estimate of individual learning gain across the instructional intervention [24,28].

$$(2) RLG=(\theta post-\theta pre)$$

XSL•FO
RenderX

To ensure the validity of Rasch-based ability estimates, we examined global model fit indicators. Item infit and outfit statistics ranged between 0.7 and 1.3, which is generally considered acceptable for the Rasch model. In addition, person reliability exceeded 0.8, and separation indices were >2, indicating satisfactory measurement precision.

Datasets

Three datasets were used to examine the behavior of the WGS under different analytical conditions:

1. The illustrative dataset (n=10): a small artificial dataset designed to illustrate the mathematical behavior of the parameter μ within the WGS metric
2. The empirical dataset (n=170): a dataset consisting of real-world data derived from a previously published educational evaluation study [29], used to examine the behavior of the WGS under authentic educational conditions and to perform inferential statistical analyses
3. The simulated dataset (n=1000): a randomly generated binomial dataset designed to mirror the structure of the empirical dataset while providing a larger sample size, allowing the behavior of the parameter μ to be examined independently of the empirical data

The Illustrative Dataset

Following the design of the simulated dataset in our previous study [16], we created an artificial dataset by combining different pretest scores with varying levels of raw gain in test performance, defined as the absolute difference between posttest and pretest scores. Pretest scores ranged from 1 to 10 points, and the gain in performance was simulated by increasing test scores by 1 to 4 points. To avoid potential ceiling effects, the analysis included only combinations in which the sum of pretest scores and the simulated gains did not exceed the maximum of 10 points. The sample size of the illustrative dataset was set at 10. RLG was not applicable here, as Rasch model–based estimation requires larger sample sizes to obtain stable parameter estimates [30].

The Empirical Dataset

The empirical dataset originated from a prospective educational study conducted at the University Medical Center Göttingen in Göttingen, Germany [29]. The study compared the learning gain of students attending a traditional lecture on goiter with that of students using a corresponding video podcast (vodcast) within the teaching module “Operative Medicine.” The study was conducted over 2 consecutive semesters using a pretest-posttest design based on 9 multiple-choice test items. A total of 170 students participated. Students were additionally surveyed regarding their learning dispositions, which resulted in the classification of participants into 2 groups: “traditional learners” and “digital natives.” A total of 35 students (20.59%) could not be clearly assigned to either group and were therefore excluded from group-based analyses. Consequently, 135 (79.41%) students were included in the 2-way ANOVA examining the interaction between teaching format and learning disposition (Table 1).

Table . Distribution of students according to teaching format and learning disposition in the empirical and simulated datasets.

Datasets and teaching formats	Traditional learners, n (%)	Digital natives, n (%)
Empirical dataset (N=135)		
Lecture	38 (28.15)	34 (25.19)
Vodcast	28 (20.74)	35 (25.93)
Simulated dataset (N=1000)		
Lecture	259 (25.9)	210 (21)
Vodcast	250 (25)	281 (28.1)

The Simulated Dataset

The simulated dataset was generated using a random binomial distribution, assuming a 50% probability of correctly answering a hypothetical examination item. This probability was applied to 9 multiple-choice items in both the pretest and the posttest scores, reflecting the structure of the “empirical dataset.” Apart from the larger sample size, the primary difference between the empirical and simulated datasets was the random allocation of group variables. Two variables were simulated: teaching format and learning disposition. Both variables were coded dichotomously. For consistency in labeling, the simulated variables were named analogously to those in the empirical dataset, although they represent random group assignments rather than actual instructional formats or learning characteristics. Each simulated student had a 50% probability of being assigned to each category (Table 1). The sample size for the simulated dataset was 1000.

Statistical Analysis

All simulations and statistical analyses were conducted using the R software suite (version 4.1.2; R Foundation for Statistical Computing) [31]. Rasch modeling was performed using the *eRm* package [32].

To examine the relationship between the 2 learning gain metrics, Pearson correlation coefficients were calculated between the WGS and RLG scores.

To investigate potential interaction effects between teaching format and learning disposition, we conducted a 2-way ANOVA. Post hoc comparisons were performed using Bonferroni-adjusted contrasts. Effect sizes were reported as partial η^2, and 95% CIs were calculated where appropriate.

Normality of the dependent variables was assessed using the Shapiro-Wilk test and visual inspection of Q-Q plots. Minor deviations from normality were observed, which are common

XSL•FO
RenderX

in bounded percentage scores (0% to 100%) frequently used in educational assessments. Given the present sample sizes and the absence of influential outliers, ANOVA was considered sufficiently robust to moderate violations of the normality assumption.

Homogeneity of variances across groups was evaluated using the Levene test and the Brown-Forsythe test, both of which indicated no statistically significant differences in variance between groups. All statistical tests were 2-sided, and a significance level of $P<.05$ was applied.

Ethical Considerations

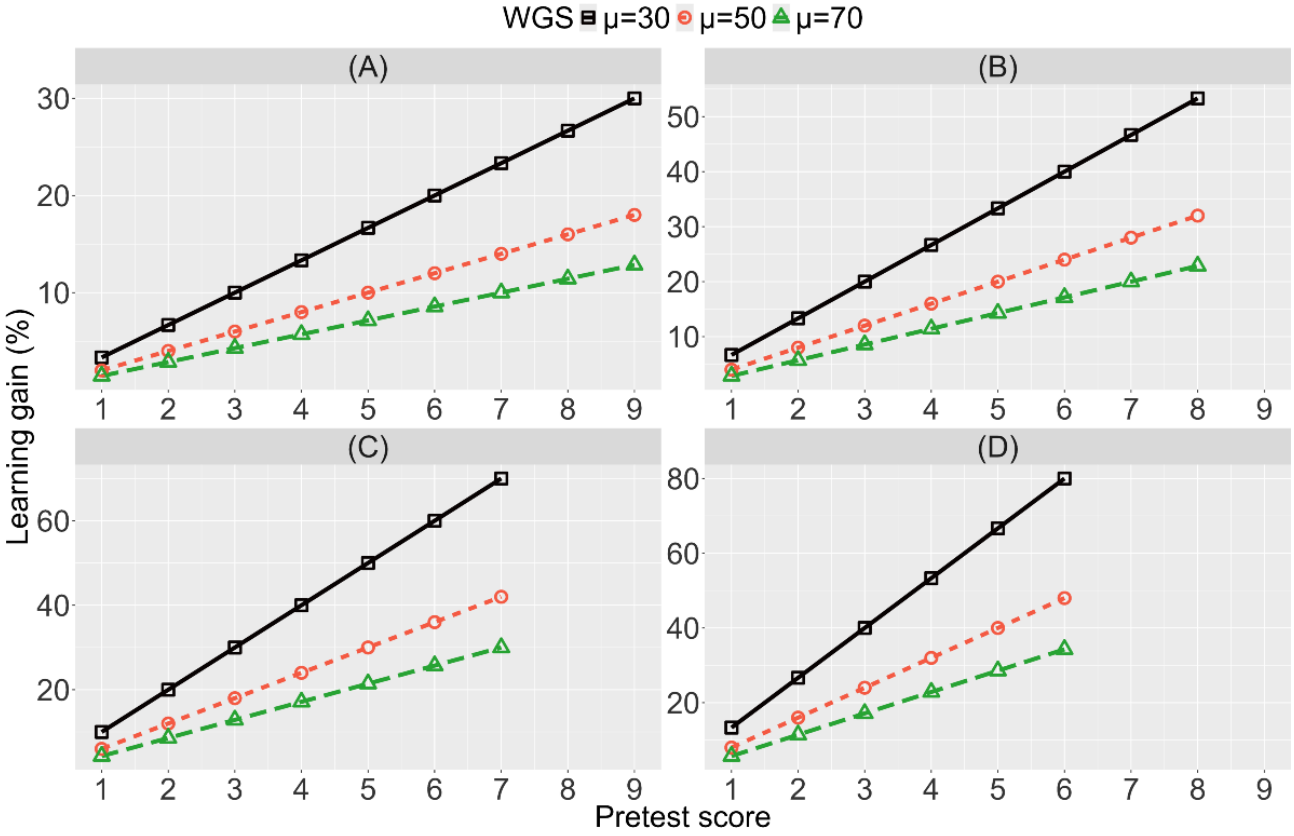
The empirical data analyzed within this work were reviewed and judged by the local institutional review and ethics board (Medical Ethics Committee, University Medical Center Göttingen) as not representing medical or epidemiological research on human participants and, therefore, were assessed using a simplified assessment protocol. The project was approved without any reservation under proposal number 1-11-14.

Results

Effect of the Parameter μ on WGS Learning Gain Estimates

The illustrative dataset demonstrates the mathematical effect of varying μ (30, 50, and 70) on the WGS. Changes in μ systematically altered the slope of the WGS learning gain plots (Figure 1). Each subplot represents a different raw gain scenario, ranging from 1 to 4 points. As the μ-value increased, the slope of the learning gain curve decreased, resulting in smaller WGS values for the same pretest score. For example, with a gain of 1 point and a pretest score of 6, the WGS was approximately 20% for μ=30 and decreased to <10% for μ=70. This pattern remained consistent across all 4 gain scenarios, illustrating that increasing μ reduces the magnitude of the calculated learning gain while preserving the relative ordering of observations.

Figure 1. Effect of varying μ (30, 50, and 70) on Weighted Gain Score (WGS) learning gain estimates in the illustrative dataset. (A) Gain of 1 point, (B) gain of 2 points, (C) gain of 3 points, and (D) gain of 4 points.

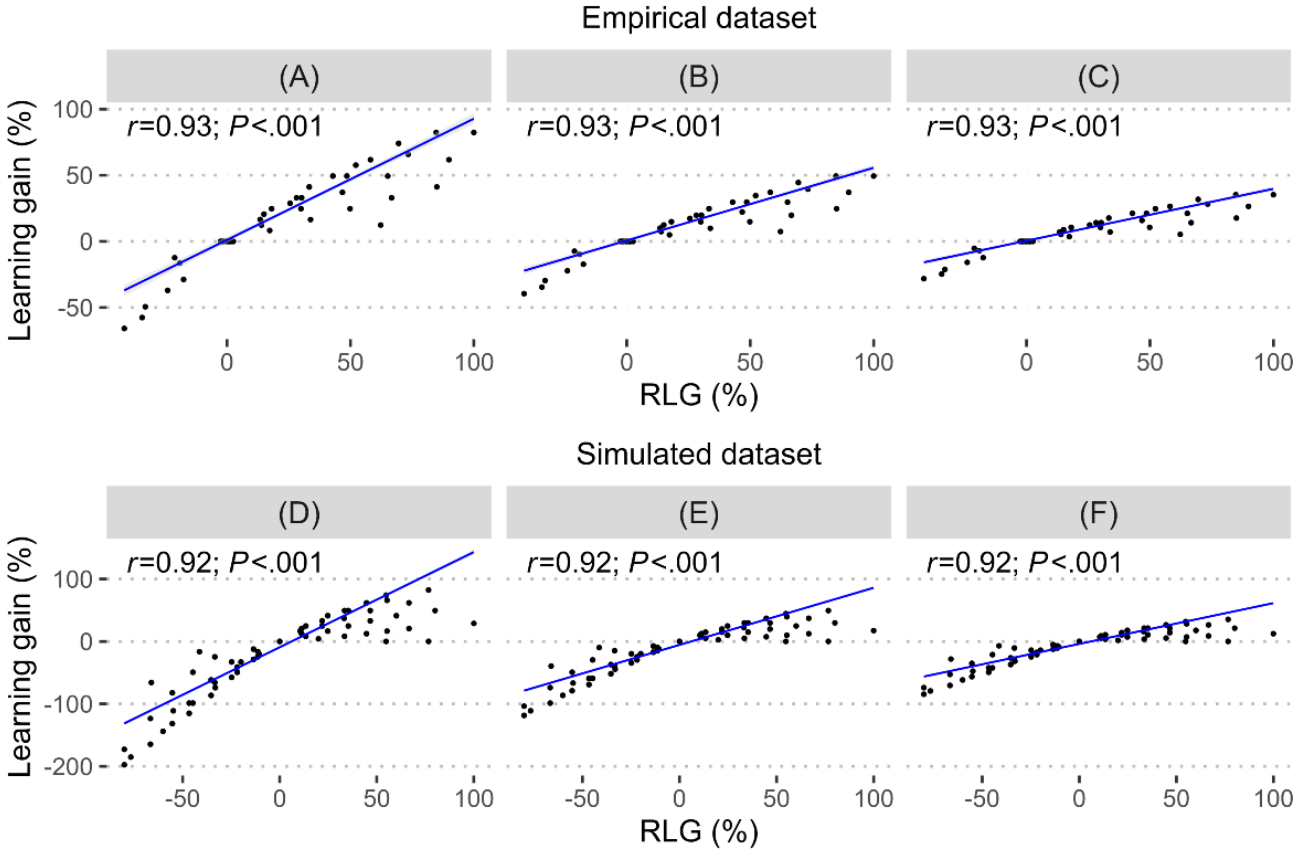


Correlation Analysis Between WGS and RLG

The WGS demonstrated a strong positive correlation with RLG across all tested μ-values (Figure 2). In the empirical dataset, the Pearson correlation coefficient was consistently high ($r=0.93$; $P<.001$). A similarly strong relationship was observed in the simulated dataset ($r=0.92$; $P<.001$). The correlation coefficients remained identical across the tested μ-values (30, 50, and 70) in both datasets.

XSL•FO
RenderX

Figure 2. Correlation between Weighted Gain Score (WGS) and Rasch Learning Gain (RLG) in the empirical and simulated datasets. (A) Empirical dataset with μ=30, (B) empirical dataset with μ=50, (C) empirical dataset with μ=70, (D) simulated dataset with μ=30, (E) simulated dataset with μ=50, and (F) simulated dataset with μ=70.

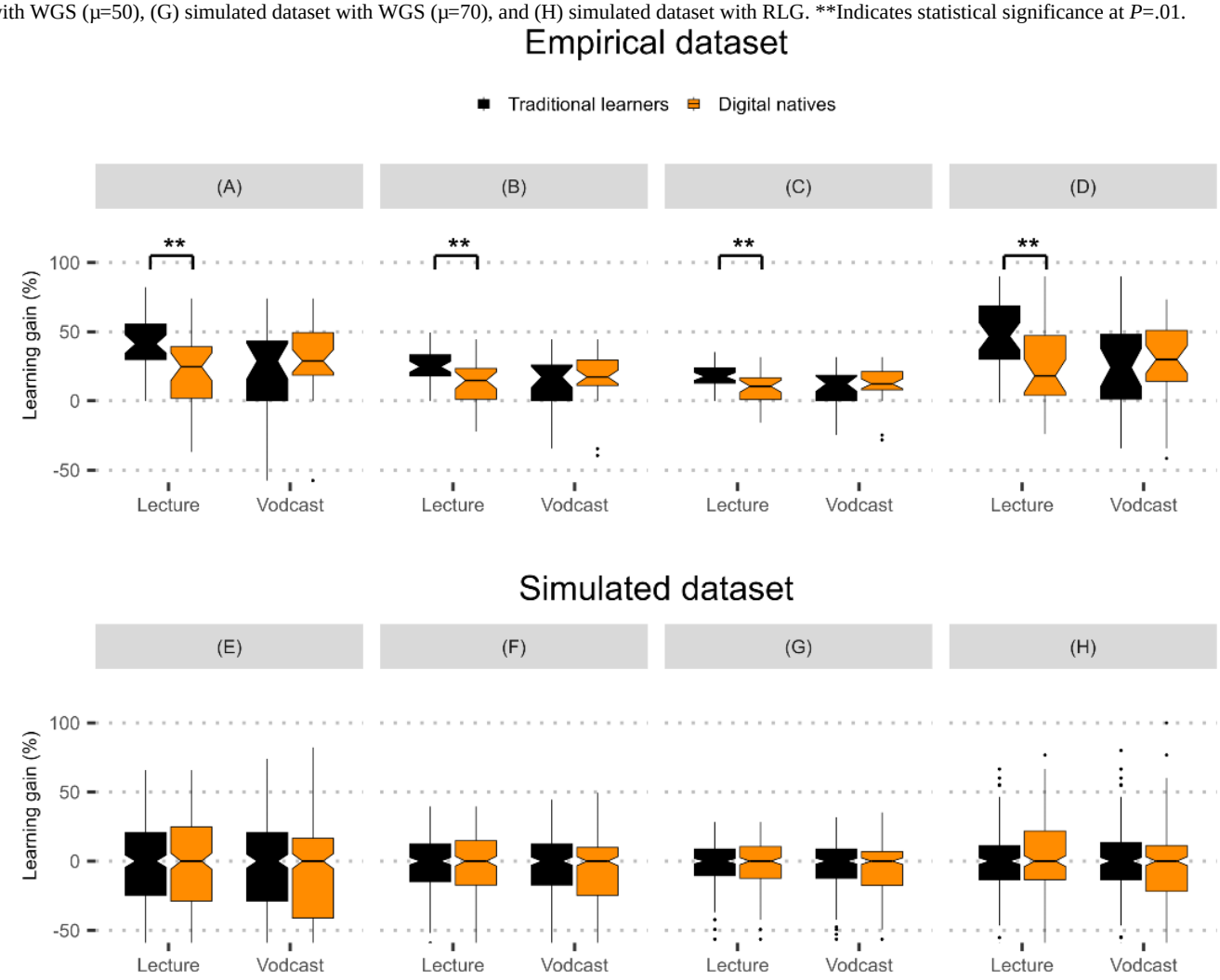


Analysis of Interaction Effects Using WGS and RLG

All 3 calibrations of the WGS (μ=30, μ=50, and μ=70) detected a significant interaction effect between teaching format and learning disposition in the empirical dataset (Figure 3). Traditional learners displayed higher learning gains in the lecture format than digital natives ($F_{1,131}$=6.51; *P*=.01; partial η^2=0.05). For μ=50, the mean difference was −11.64 (95% Bonferroni-adjusted CI −21.46 to −1.83; *P*=.01). Corresponding estimates were −19.41 for μ=30 (95% Bonferroni-adjusted CI −35.80 to −3.04; *P*=.01) and −8.32 for μ=70 (95% Bonferroni-adjusted CI −15.33 to −1.31; *P*=.01).

XSL•FO
RenderX

Figure 3. Learning gain estimates calculated using Weighted Gain Score (WGS) and Rasch Learning Gain (RLG), depicting the interaction between teaching format and learning disposition in the empirical and simulated datasets. (A) Empirical dataset with WGS (μ=30), (B) empirical dataset with WGS (μ=50), (C) empirical dataset with WGS (μ=70), (D) empirical dataset with RLG, (E) simulated dataset with WGS (μ=30), (F) simulated dataset with WGS (μ=50), (G) simulated dataset with WGS (μ=70), and (H) simulated dataset with RLG. **Indicates statistical significance at *P*=.01.



RLG also detected this interaction effect ($F_{1,131}$=6.75; *P*=.01; partial η^2=0.05) with a mean difference of −19.91 (95% Bonferroni-adjusted CI −36.80 to −3.05; *P*=.01), confirming the interaction pattern observed in the original study from which our empirical dataset was derived [16,29].

In the simulated dataset, no significant interaction between teaching format and learning disposition was observed when learning gains were calculated using the WGS, regardless of the μ-value applied ($F_{1,996}$=0.39; *P*=.53; partial η^2<0.001; Figure 3). Similarly, RLG did not reveal any significant difference in performance between the groups ($F_{1,996}$=1.10; *P*=.29; partial η^2=0.001). Because teaching format and learning disposition were randomly assigned in the simulated dataset, we did not necessarily expect any interaction effect.

Discussion

Inferential Behavior of WGS Compared With RLG

A robust method for calculating learning gain is essential for capturing students' learning progress following an educational intervention and for providing interpretable indicators of educational effectiveness. Such a method should be statistically sound, transparent, and practically applicable within evaluation processes.

This study evaluated the statistical behavior of the WGS, a method designed to estimate learning gain in a way that is both methodologically robust and straightforward to implement. The first research question examined whether the WGS yields inferential results comparable to those obtained with RLG. Our findings demonstrated a strong inferential correspondence between the 2 methods. The WGS produced learning gain estimates that correlated highly with those derived from RLG, while also identifying the same interaction effect in the empirical dataset as the Rasch model–based approach. Importantly, these inferential conclusions remained stable across all tested μ-values (30, 50, and 70). The identical correlation coefficients between the WGS and RLG and the unchanged ANOVA results indicate that modifying the parameter μ linearly rescales learning gain estimates. Consequently, varying μ changes the magnitude of WGS values but does not affect statistical inference.

Robustness of WGS Under Nonnormally Distributed Data

Neither the empirical nor the simulated dataset fully satisfied the assumption of normality, although no substantial skewness

XSL•FO
RenderX

was observed. In medical education research, deviations from normality are common, particularly in pretest-posttest designs [17,33]. A ceiling effect occurs when pretest scores approach the maximum possible value, limiting the measurable improvement, whereas a floor effect arises when pretest performance is concentrated near the minimum score in a difficult test. Very easy items tend to produce ceiling effects, whereas very difficult items may lead to floor effects. Despite deviations from normality, the WGS demonstrated stable inferential behavior across the empirical and simulated datasets, suggesting a degree of robustness. This finding is consistent with previous research indicating that parametric methods such as ANOVA and correlation analyses are generally robust to moderate violations of normality, particularly in samples of the size examined in this study [34-36]. Nevertheless, future research is needed to examine the behavior of the WGS across a broader range of distributional scenarios to better establish its reliability.

Applicability of WGS in Small Samples

The illustrative dataset demonstrates that the WGS yields interpretable results even with very small sample sizes. In contrast, Rasch model–based approaches typically require substantially larger samples to ensure stable estimation of item parameters and person abilities [24,26]. This distinction is particularly relevant in educational settings with small cohorts, such as specialized teaching modules, pilot courses, or resource-intensive instructional interventions. In such contexts, the WGS may represent a practical alternative method for estimating learning gain because it does not rely on complex parameter estimation.

More broadly, transparent feedback on learning outcomes supports the continuous development of teaching practices, as evidence suggests that feedback on educational performance encourages educators to engage in reflective improvement of their teaching [37-39].

The Role of μ as an Adjustment Coefficient

The second research question examined whether the parameter μ in the WGS can be adjusted across different cohorts to calibrate learning gain calculations without altering inferential outcomes. In the original study introducing the WGS, μ was defined as the average pretest score of a cohort. Our findings suggest that the role of μ can be understood more broadly. Rather than representing solely the cohort mean, μ functions as a scaling parameter that allows calibration of the learning gain metric. To reflect this role more accurately, we interpret μ in this study as an adjustment coefficient that can be modified depending on the analytical purpose of the evaluation. On the basis of the results of this study, 3 conceptual adjustment strategies can be distinguished: absolute adjustment, relative adjustment, and routine evaluation. A decision framework for selecting μ is provided in Multimedia Appendix 1.

Absolute Adjustment: Monitoring of Cohort Learning

Absolute adjustment refers to the use of a fixed μ-value to estimate learning gain within a stable scaling framework. When μ remains constant, differences in learning gain across courses, time points within the curriculum, or different cohorts can be interpreted without recalibration of the metric, thereby ensuring cross-cohort comparability. This approach supports standardized monitoring of educational outcomes, for example, when evaluating curricular developments over time or comparing modules within a program. Observed differences in learning gain may arise from multiple factors, including instructional design, assessment characteristics, or cohort composition. Maintaining a fixed μ ensures that such differences remain visible and can be attributed to substantive factors.

Relative Adjustment: Evaluation of Teaching Interventions

Relative adjustment enables comparison of teaching interventions across cohorts with heterogeneous characteristics. In educational practice, cohorts often differ in characteristics such as demographics, motivation, workload, or external contextual influences [13,40]. When learning gain is used to compare instructional formats, such heterogeneity may affect the interpretation of outcomes. Under a relative adjustment strategy, a μ-value may be calibrated separately for each cohort, allowing the scaling of learning gain calculations to reflect cohort-specific baseline conditions. Although this approach does not eliminate potential confounding factors, it may reduce systematic bias associated with heterogeneous starting conditions. This strategy is particularly useful when learning gain is evaluated without strict requirements for cross-cohort comparability, but with a focus on fair comparison of teaching interventions within specific cohorts or instructional contexts.

Routine Evaluation: Selecting μ in Practice

In routine applications, when learning gain is estimated without strict requirements for cross-cohort comparability or cohort-specific calibration, μ may be selected pragmatically based on the cohort's mean pretest performance. For example, cohorts with mean pretest scores approximately 50% of the maximum achievable score may be assigned μ=50, whereas cohorts with substantially higher or lower baseline knowledge may be assigned correspondingly higher (eg, ≥70) or lower (eg, ≤30) μ-values. This pragmatic approach enables straightforward estimation of learning gain while preserving a transparent and easily interpretable scaling of the WGS metric.

Limitations

One limitation of the WGS arises when a student obtains a pretest score of zero, which results in a calculated learning gain of zero regardless of posttest performance. In practice, such cases are unlikely in multiple-choice assessments because guessing and prior knowledge increase the probability of obtaining at least 1 correct response [41]. In the empirical dataset, no student recorded a pretest score of zero, and in the simulated dataset, a negligible number (3 out of 1000 students) achieved zero points on the pretest score. One possible strategy is to exclude such observations from the analysis. However, this may reduce statistical power and introduce bias if students with low baseline scores are systematically underrepresented. Alternatively, a small positive offset (pseudocount) could be added to avoid undefined computations, analogous to continuity corrections used in categorical data analysis [42,43]. The implications of such adjustments should be examined in future

XSL•FO
RenderX

methodological studies, for example, through sensitivity analyses comparing different handling strategies for zero-baseline observations [44].

A further limitation concerns the sample size of the empirical dataset (n=170). Although cohort sizes of this magnitude are common in single-semester cohorts at German medical faculties, they are slightly below commonly cited recommendations for stable Rasch parameter estimation, which often suggest sample sizes of approximately 150 to 200 participants or more [45]. Nevertheless, global model fit indicators in the empirical dataset (infit and outfit statistics, person reliability, and separation indices) were within acceptable ranges, supporting the interpretability of the RLG-based estimates despite the moderate sample size.

Another limitation relates to the test length used in the simulated dataset, which consisted of 9 multiple-choice items to mirror the empirical dataset. Because measurement reliability generally increases with test length [46-50], the limited number of items may reduce measurement precision and restrict the generalizability of the findings. Therefore, future research should examine the performance of the WGS in assessments with larger item sets that more closely reflect the scope of medical examinations.

Finally, both datasets exhibited deviations from normality**,** although homogeneity of variances across groups was supported by the Levene and Brown-Forsythe tests, and no influential outliers were observed. Previous methodological research indicates that ANOVA and Pearson correlation are generally robust to moderate violations of normality, particularly in samples of the present size [34-36]. Therefore, we consider the impact of nonnormality on the inferential conclusions to be limited.

Conclusions and Future Research

This study evaluated the WGS as a method for estimating learning gain in pretest-posttest educational designs. Our findings indicate that the WGS provides robust and easily interpretable estimates while remaining computationally simple. Rather than replacing established psychometric models, the WGS may complement existing approaches, particularly in routine educational evaluations.

Future research should further develop the WGS as a broadly applicable evaluation instrument. In particular, establishing a methodologically sound calibration framework for μ will be essential, including empirically grounded decision models that guide μ-selection according to the evaluation purpose, such as cohort monitoring or comparative evaluation of teaching interventions. In addition, integrating the WGS into structured program evaluations, including longitudinal monitoring across courses, will be important for assessing its generalizability across educational contexts.

Future work may also explore the integration of the WGS within Bayesian test-theoretical frameworks [51]. By incorporating prior information and updating gain estimates as new data become available, Bayesian approaches could further improve the precision and contextual sensitivity of WGS-based learning gain estimates. Further studies should also examine the behavior of the WGS under different distributional conditions to better establish its robustness.

Acknowledgments

The authors sincerely thank Simone Kann and Michael Schuler for their valuable insights and thoughtful suggestions, as well as Andrew Entwistle for his contribution to the revision of this manuscript.

Funding

This research did not receive funding from any specific grant provided by public, commercial, or not-for-profit agencies.

Data Availability

The data supporting the findings of this study are provided as a multimedia appendix to facilitate full reproducibility.

Authors' Contributions

All authors were involved in the conception and/or design of the study and contributed critically to the final preparation of this study, including approving the final version of the manuscript. In particular, SK conceived and designed the study, wrote the final study protocol, and drafted the manuscript. RA conducted the study, collected the results, and analyzed the data. SH and JB analyzed the data and performed and verified the statistical analyses.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Decision framework for selecting the calibration parameter μ in Weighted Gain Score calculations according to the evaluation objective (absolute adjustment, relative adjustment, or routine evaluation).
[DOCX File, 194 KB - mededu_v12i1e75516_app1.docx]

XSL•FO
RenderX

References

1. Charalambous CY, Praetorius AK, Sammons P, Walkowiak T, Jentsch A, Kyriakides L. Working more collaboratively to better understand teaching and its quality: challenges faced and possible solutions. Stud Educ Eval 2021 Dec;71(1):101092. [doi: 10.1016/j.stueduc.2021.101092]
2. Gibson KA, Boyle P, Black DA, Cunningham M, Grimm MC, McNeil HP. Enhancing evaluation in an undergraduate medical education program. Acad Med 2008 Aug;83(8):787-793. [doi: 10.1097/ACM.0b013e31817eb8ab] [Medline: 18667897]
3. Litzelman DK, Stratos GA, Marriott DJ, Skeff KM. Factorial validation of a widely disseminated educational framework for evaluating clinical teachers. Acad Med 1998 Jun;73(6):688-695. [doi: 10.1097/00001888-199806000-00016] [Medline: 9653408]
4. Noor C, Hozan CT, Vîlceanu N, Bon ea MG. A review of the effectiveness of the role of various components in medical education. Arch Pharm Pract 2023;14(4):155-159. [doi: 10.51847/LrElkFGJAO]
5. McGaghie WC, Issenberg SB, Cohen ER, Barsuk JH, Wayne DB. Medical education featuring mastery learning with deliberate practice can lead to better health for individuals and populations. Acad Med 2011 Nov;86(11):e8-e9. [doi: 10.1097/ACM.0b013e3182308d37] [Medline: 22030671]
6. Gould BE, Grey MR, Huntington CG, et al. Improving patient care outcomes by teaching quality improvement to medical students in community-based practices. Acad Med 2002 Oct;77(10):1011-1018. [doi: 10.1097/00001888-200210000-00014] [Medline: 12377677]
7. Schiekirka-Schwake S, Anders S, von Steinbüchel N, Becker JC, Raupach T. Facilitators of high-quality teaching in medical school: findings from a nation-wide survey among clinical teachers. BMC Med Educ 2017 Sep 29;17(1):178. [doi: 10.1186/s12909-017-1000-6] [Medline: 28962568]
8. Schiekirka S, Reinhardt D, Beißbarth T, Anders S, Pukrop T, Raupach T. Estimating learning outcomes from pre- and posttest student self-assessments: a longitudinal study. Acad Med 2013 Mar;88(3):369-375. [doi: 10.1097/ACM.0b013e318280a6f6] [Medline: 23348083]
9. Gruppen LD. Outcome-based medical education: implications, opportunities, and challenges. Korean J Med Educ 2012 Dec;24(4):281-285. [doi: 10.3946/kjme.2012.24.4.281] [Medline: 25813324]
10. Harden RM. AMEE guide no. 14: outcome-based education: part 1-an introduction to outcome-based education. Med Teach 1999 Jan;21(1):7-14. [doi: 10.1080/01421599979969]
11. Haverkamp N, Barth J, Schmidt D, Dahmen U, Keis O, Raupach T. Position statement of the GMA committee “teaching evaluation”. GMS J Med Educ 2024;41(2):Doc19. [doi: 10.3205/zma001674] [Medline: 38779701]
12. Fraenkel JR, Wallen NE, Hyun HH. How to Design and Evaluate Research in Education, 8th edition: McGraw-Hill; 2012.
13. Cook DA, Beckman TJ. Reflections on experimental research in medical education. Adv Health Sci Educ Theory Pract 2010 Aug;15(3):455-464. [doi: 10.1007/s10459-008-9117-3] [Medline: 18427941]
14. Colt HG, Davoudi M, Murgu S, Zamanian Rohani N. Measuring learning gain during a one-day introductory bronchoscopy course. Surg Endosc 2011 Jan;25(1):207-216. [doi: 10.1007/s00464-010-1161-4] [Medline: 20585964]
15. McGrath C, Guerin B, Harte E, Frearson M, Manville C. Learning gain in higher education. RAND Corporation. 2015. URL: https://www.rand.org/pubs/research_reports/RR996.html [accessed 2026-05-25]
16. Westphale S, Backhaus J, Koenig S. Quantifying teaching quality in medical education: the impact of learning gain calculation. Med Educ 2022 Mar;56(3):312-320. [doi: 10.1111/medu.14694] [Medline: 34767274]
17. Šimkovic M, Träuble B. Robustness of statistical methods when measure is affected by ceiling and/or floor effect. PLoS One 2019;14(8):e0220889. [doi: 10.1371/journal.pone.0220889] [Medline: 31425561]
18. Bereiter C. Some persisting dilemmas in the measurement of change. In: Harris CW, editor. Problems in Measuring Change: University of Wisconsin Press; 1963:3-20.
19. Prieler J, Raven J. Problems in the measurement of change (with particular reference to individual change [gain] scores) and their potential solution using IRT. In: Uses and Abuses of Intelligence: Studies Advancing Spearman and Raven’s Quest for Non-Arbitrary Metrics: Royal Fireworks Press; 2008:173-210.
20. Hake RR. Interactive-engagement versus traditional methods: a six-thousand-student survey of mechanics test data for introductory physics courses. Am J Phys 1998 Jan;66(1):64-74. [doi: 10.1119/1.18809]
21. Coletta VP, Phillips JA. Interpreting FCI scores: normalized gain, preinstruction scores, and scientific reasoning ability. Am J Phys 2005 Dec;73(12):1172-1182. [doi: 10.1119/1.2117109]
22. Nissen JM, Talbot RM, Thompson AN, Van Dusen B. Comparison of normalized gain and Cohen’s d for analyzing gains on concept inventories. Phys Rev Phys Educ Res 2018;14:010115. [doi: 10.1103/PhysRevPhysEducRes.14.010115]
23. Marx JD, Cummings K. Normalized change. Am J Phys 2007;75(1):87-91. [doi: 10.1119/1.2372468]
24. Pentecost TC, Barbera J. Measuring learning gains in chemical education: a comparison of two methods. J Chem Educ 2013 Jul;90(7):839-845. [doi: 10.1021/ed400018v]
25. Bond TG, Fox CM. Applying the Rasch Model: Fundamental Measurement in the Human Sciences, 3rd edition: Routledge; 2015.
26. Downing SM. Item response theory: applications of modern test theory in medical education. Med Educ 2003 Aug;37(8):739-745. [doi: 10.1046/j.1365-2923.2003.01587.x] [Medline: 12945568]

XSL•FO
RenderX

27. Embretson SE, Reise SP. Item Response Theory: Psychology Press; 2013.
28. Wallace CS, Bailey JM. Do concept inventories actually measure anything? Astron Educ Rev 2010;9(1). [doi: 10.3847/AER2010024]
29. Backhaus J, Huth K, Entwistle A, Homayounfar K, Koenig S. Digital affinity in medical students influences learning outcome: a cluster analytical design comparing vodcast with traditional lecture. J Surg Educ 2019;76(3):711-719. [doi: 10.1016/j.jsurg.2018.12.001] [Medline: 30833205]
30. Chen WH, Lenderking W, Jin Y, Wyrwich KW, Gelhorn H, Revicki DA. Is Rasch model analysis applicable in small sample size pilot studies for assessing item characteristics? An example using PROMIS pain behavior item bank data. Qual Life Res 2014 Mar;23(2):485-493. [doi: 10.1007/s11136-013-0487-5] [Medline: 23912855]
31. R Core Team. The R Project for Statistical Computing. R Foundation for Statistical Computing. 2013. URL: http://www.R-project.org [accessed 2026-05-25]
32. Mair P, Hatzinger R. Extended Rasch modeling: the eRm package for the application of IRT models in R. J Stat Soft 2007;20(9):1-20. [doi: 10.18637/jss.v020.i09]
33. Micceri T. The unicorn, the normal curve, and other improbable creatures. Psychol Bull 1989;105(1):156-166. [doi: 10.1037/0033-2909.105.1.156]
34. Bishara AJ, Hittner JB. Testing the significance of a correlation with nonnormal data: comparison of Pearson, Spearman, transformation, and resampling approaches. Psychol Methods 2012 Sep;17(3):399-417. [doi: 10.1037/a0028087] [Medline: 22563845]
35. Knief U, Forstmeier W. Violating the normality assumption may be the lesser of two evils. Behav Res Methods 2021 Dec;53(6):2576-2590. [doi: 10.3758/s13428-021-01587-5] [Medline: 33963496]
36. Havlicek LL, Peterson NL. Robustness of the Pearson correlation against violations of assumptions. Percept Mot Skills 1976 Dec;43(3_suppl):1319-1334. [doi: 10.2466/pms.1976.43.3f.1319]
37. Boerboom TBB, Stalmeijer RE, Dolmans DHJM, Jaarsma DADC. How feedback can foster professional growth of teachers in the clinical workplace: a review of the literature. Studies in Educational Evaluation 2015 Sep;46:47-52. [doi: 10.1016/j.stueduc.2015.02.001]
38. Scheeler MC, Ruhl KL, McAfee JK. Providing performance feedback to teachers: a review. Teach Educ Spec Educ 2004;27(4):396-407. [doi: 10.1177/088840640402700407]
39. Evans C, Howson CK, Forsythe A. Making sense of learning gain in higher education. Higher Education Pedagogies 2018;3(1):1-45. [doi: 10.1080/23752696.2018.1508360]
40. Ewert A, Sibthorp J. Creating outcomes through experiential education: the challenge of confounding variables. Journal of Experiential Education 2009 Jan 1;31(3):376-389. [doi: 10.5193/JEE.31.3.376]
41. Kubinger KD, Gottschall CH. Item difficulty of multiple choice tests dependant on different item response formats—an experiment in fundamental research on psychological assessment. Psychol Sci 2007;49(4):361-374.
42. Weber F, Knapp G, Ickstadt K, Kundt G, Glass Ä. Zero-cell corrections in random-effects meta-analyses. Res Synth Methods 2020 Nov;11(6):913-919. [doi: 10.1002/jrsm.1460] [Medline: 32991790]
43. Sweeting MJ, Sutton AJ, Lambert PC. What to add to nothing? Use and avoidance of continuity corrections in meta-analysis of sparse data. Stat Med 2004 May 15;23(9):1351-1375. [doi: 10.1002/sim.1761] [Medline: 15116347]
44. Aung NM, Jurak I, Mehmood S, Axon E. Sensitivity analysis in meta-analysis: a tutorial. Cochrane Evid Synth Methods 2026 Jan;4(1):e70067. [doi: 10.1002/cesm.70067] [Medline: 41497796]
45. O'Neill TR, Gregg JL, Peabody MR. Effect of sample size on common item equating using the dichotomous Rasch model. Appl Meas Educ 2020 Jan;33(1):10-23. [doi: 10.1080/08957347.2019.1674309]
46. Downing SM. Reliability: on the reproducibility of assessment data. Med Educ 2004 Sep;38(9):1006-1012. [doi: 10.1111/j.1365-2929.2004.01932.x] [Medline: 15327684]
47. Tavakol M, Dennick R. Making sense of Cronbach's alpha. Int J Med Educ 2011 Jun 27;2:53-55. [doi: 10.5116/ijme.4dfb.8dfd] [Medline: 28029643]
48. de Vet HC, Mokkink LB, Mosmuller DG, Terwee CB. Spearman-Brown prophecy formula and Cronbach's alpha: different faces of reliability and opportunities for new applications. J Clin Epidemiol 2017 May;85:45-49. [doi: 10.1016/j.jclinepi.2017.01.013] [Medline: 28342902]
49. Brown W. Some experimental results in the correlation of mental abilities. Br J Psychol 1904-1920 1910 Oct;3(3):296-322. [doi: 10.1111/j.2044-8295.1910.tb00207.x]
50. Spearman C. Correlation calculated from faulty data. Br J Psychol 1904-1920 1910 Oct;3(3):271-295. [doi: 10.1111/j.2044-8295.1910.tb00206.x]
51. Rindskopf D. Overview of Bayesian statistics. Eval Rev 2020 Aug;44(4):225-237. [doi: 10.1177/0193841X19895623] [Medline: 31894697]

Abbreviations

WGS: Weighted Gain Score
RLG: Rasch Learning Gain

XSL•FO
RenderX

Edited by A Bahattab; submitted 18.May.2025; peer-reviewed by MS Shafi, S Alzaabi, TA Valencia-Perez; revised version received 19.Mar.2026; accepted 22.Apr.2026; published 16.Jun.2026.

Please cite as:
Aliyev R, Backhaus J, Hammer S, König S
Comparing the Weighted Gain Score and a Rasch-Based Approach for Estimating Learning Outcomes in Medical Education: Quantitative Study
JMIR Med Educ 2026;12:e75516
URL: https://mededu.jmir.org/2026/1/e75516
doi:10.2196/75516

© Rauf Aliyev, Joy Backhaus, Silke Hammer, Sarah König. Originally published in JMIR Medical Education (https://mededu.jmir.org), 16.Jun.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Education, is properly cited. The complete bibliographic information, a link to the original publication on https://mededu.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Evaluating the Influence of MCAT Scores on Medical Student Selection and Performance Using a Machine Learning Approach: Retrospective Cohort Study

Michele Cherfane[1], PhD; Marc Ghanem[1], MD; Georges Choueiry[1], MS; Selim M Nasser[1,2], MD; Tamina Elias-Rizk[1,2], MD; Sola Aoun Bahous[1,2], MD, PhD

[1]Gilbert and Rose-Marie Chagoury School of Medicine, Lebanese American University, Blat, PO box 36 Byblos, Byblos, Byblos, Lebanon
[2]Lebanese American University Medical Center-Rizk Hospital, Beirut, Lebanon

Corresponding Author:
Sola Aoun Bahous, MD, PhD
Gilbert and Rose-Marie Chagoury School of Medicine, Lebanese American University, Blat, PO box 36 Byblos, Byblos, Byblos, Lebanon

Abstract

Background: The Medical College Admission Test (MCAT) has been central to medical school admissions in North America, though its necessity in holistic selection processes remains debated. The COVID-19 pandemic's suspension of MCAT testing sessions allowed institutions to explore alternative admission criteria. Additionally, global challenges in test administration underscore the vulnerability of systems dependent on a single standardized test.

Objective: This study aimed to (1) assess whether including MCAT scores in admissions decisions improves the prediction of medical school performance compared with grade point average (GPA) and interview-based selection and (2) evaluate whether machine learning (ML) can generate viable MCAT score predictions when testing is unavailable.

Methods: We conducted a retrospective cohort study of 1898 applicants to the Lebanese American University School of Medicine (2009 - 2023). Among 349 admitted students with MCAT scores, we compared admission composites including vs excluding the MCAT in relation to academic outcomes (Med 1 - 4 final grades) and clinical performance (Med 1 - 2 objective structured clinical examination [OSCE] scores). Students were stratified into tertiles to examine tier-specific effects. To model an MCAT alternative, we trained an ensemble ML model (least absolute shrinkage and selection operator, kernel ridge, gradient boosting, elastic net, and light gradient boosting machine [LightGBM]) using data from 1583 applicants (2009 - 2020) to predict MCAT scores from cumulative GPA, core GPA, interview scores, merit points, and honors participation. The model was validated on 315 applicants admitted during the MCAT suspension (2021 - 2023), and its impact on admission rankings and performance correlations was evaluated.

Results: Including the MCAT in the admissions composite did not meaningfully improve prediction of overall academic performance (r=0.63, 95% CI 0.56-0.69 with MCAT vs r=0.62, 95% CI 0.55-0.68 without MCAT; P=.63). However, it significantly weakened prediction of clinical skills (OSCE: r=0.37, 95% CI 0.27-0.46 with MCAT vs r=0.48, 95% CI 0.40-0.56 without MCAT; P<.001). In tertile analyses, the MCAT modestly improved prediction among top-performing students but eliminated predictive validity in the bottom tertile (r=0.12, 95% CI –0.07 to 0.30; P=.21 vs r=0.28, 95% CI 0.10-0.44 without MCAT; P=.003). The ML model explained 36% of MCAT variance (R^2=0.36 with 95% CI 0.32-0.41; root mean squared error ≈10% of score range). Incorporating predicted MCAT scores reduced correlations with subsequent performance across all outcomes. Although rankings based on predicted MCAT scores strongly correlated with original rankings (r=0.80, 95% CI 0.77-0.83), 6.3% (4/64) to 9.4% (6/64) of admission decisions would have changed.

Conclusions: Within this single-institution study, MCAT scores provided limited incremental validity for predicting academic performance and reduced prediction of clinical skills. ML-based predictions of MCAT scores introduced sufficient error to affect admissions decisions, supporting a robust holistic admissions process during temporary MCAT disruptions while highlighting the need for validation across diverse institutional settings before broader generalization.

(JMIR Med Educ 2026;12:e94197) doi:10.2196/94197

KEYWORDS

Medical College Admission Test; MCAT; machine learning; medical school admission; medical school performance; student selection

XSL•FO
RenderX

Introduction

Background

Selecting future medical students is a high-stakes endeavor that shapes the future physician workforce. Admissions committees strive for a holistic review that balances academic excellence with personal attributes. In North American models, this means combining metrics like undergraduate grade point average (uGPA) and Medical College Admission Test (MCAT) scores with evaluations of nonacademic qualities such as interview performance and situational judgment tests [1-4]. This multifactor approach aims to identify well-rounded candidates who possess the knowledge, skills, and character to succeed as physicians.

The MCAT has long held a central role as a standardized measure of applicants' academic preparedness. Since its introduction, the examination has evolved from a predominantly knowledge-based assessment into a competency-oriented tool designed to evaluate scientific reasoning, critical analysis, and understanding of behavioral and social determinants of health [5]. This evolution reflects broader shifts in medical education toward competency-based training and more holistic admissions practices. Several studies evaluating the newer versions of the MCAT have demonstrated its predictive validity, particularly for student performance in the first year of medical school, and, to a lesser extent, throughout subsequent stages of training [4,6-9]. Despite this, debate has persisted within the medical education community regarding whether the exam carries disproportionate weight in admissions decisions, with some advocating for its reduced emphasis or elimination [10].

These debates extend beyond medical education and reflect broader discussions regarding standardized testing in higher education. Supporters of standardized examinations argue that they provide a common benchmark across heterogeneous educational systems and help ensure comparability among applicants from diverse academic backgrounds [11,12]. Critics, however, have raised concerns that excessive reliance on test-based selection may unintentionally reinforce socioeconomic and educational inequities, particularly given unequal access to preparatory resources, variations in educational opportunity, and differential application patterns across socioeconomic groups [13]. Recent national data from MCAT examinees further demonstrate persistent socioeconomic disparities in examination performance and the likelihood of applying to medical school, reinforcing concerns regarding equity in admissions [14]. These debates have also been informed by broader psychometric phenomena, including the Flynn effect, which describes the long-term increase in standardized cognitive test scores across generations, estimated at approximately 2 to 3 IQ points per decade across many populations, and highlights that population-level changes may influence the interpretation and ongoing evaluation of standardized assessments [15,16]. Consequently, many medical schools have increasingly adopted holistic review approaches that seek to balance cognitive metrics with assessment of interpersonal, ethical, and professional attributes [17].

Importantly, the North American medical admissions model differs substantially from that used in many other regions of the world. While many countries admit students directly into medical education following secondary school, North American institutions generally require completion of a prior undergraduate degree and evaluate applicants through multiple layers of academic and nonacademic assessment. Within this framework, the MCAT serves not only as a measure of academic preparedness but also as a standardized comparator across diverse undergraduate institutions and educational pathways [12]. At the same time, the growing use of situational judgment tests and structured interviews reflects increasing recognition that no single examination fully captures the competencies required for future physicians [17,18].

Admissions During the COVID-19 Pandemic

The widespread suspension of the MCAT exam during the COVID-19 pandemic created a unique opportunity that allowed institutions to explore alternative admissions criteria, relying more heavily on grade point average (GPA) and interviews [19]. While GPA is a consistent academic predictor, interview scores have been shown to be less reliable and predictive of success [20-25], may unintentionally favor some demographic groups [26], and can be influenced by applicants' prior academic performance [27]. Beyond the pandemic, global challenges in MCAT administration, ranging from limited test center capacity to policy restrictions, such as the elimination of the examination in Quebec following new language requirements, underscore the vulnerability of admissions systems that depend heavily on a single standardized test. These experiences encouraged investigation into innovative approaches, such as predictive modeling, that could support admissions decision-making when standardized testing is disrupted.

In Lebanon, the MCAT was suspended for 2 consecutive years due to the pandemic, requiring the Lebanese American University's (LAUs) medical school admissions to rely solely on interview scores and uGPA. This experience raised 2 fundamental questions: How essential is the MCAT to making sound admissions decisions, and can technology offer a stand-in for the MCAT when it is unavailable?

On the one hand, if admissions without the MCAT yield similar outcomes, it would suggest that the test, while useful, may be somewhat redundant with other indicators (like GPA). On the other hand, the MCAT might add a unique standardization that, in its absence, could be approximated using predictive modeling. Recent advances in machine learning (ML) make it conceivable to predict an applicant's likely MCAT score from their academic record and other attributes. If such an ML model is accurate, it could serve as a temporary tool for admissions committees during periods when traditional testing is disrupted (eg, pandemics, natural disasters, or sociopolitical instability). Importantly, framing ML in this context is not to replace the MCAT permanently but to support decision-making when the test cannot be administered.

Study Objectives

Building on this context, the present study had two prespecified objectives: (1) to assess whether including MCAT scores in

XSL•FO
RenderX

admissions scoring improves prediction of medical school performance compared with decisions made without the exam and (2) to evaluate the feasibility of an ML-based model that predicts MCAT scores from other admissions data as a potential alternative when the exam is unavailable.

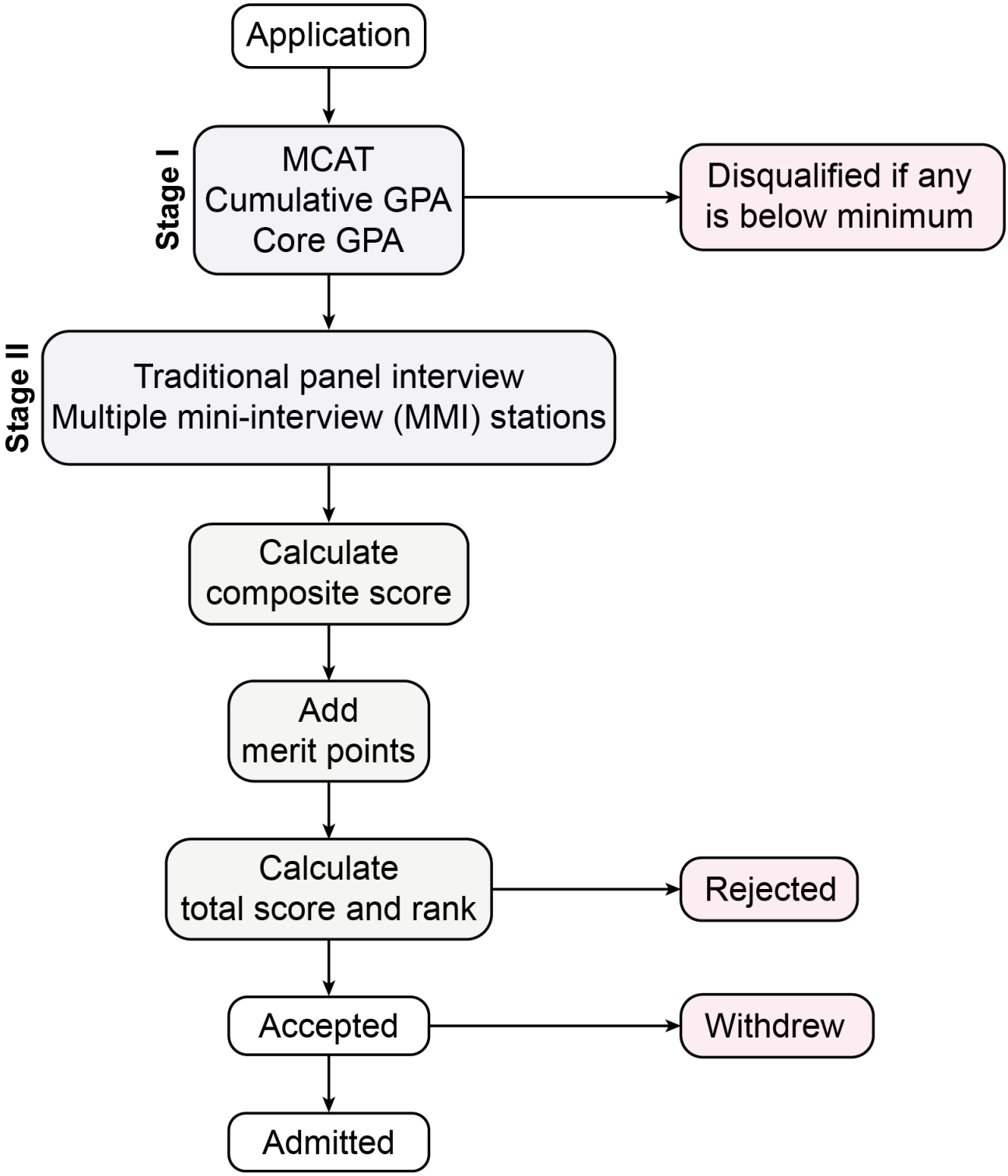
Methods

Overview of the Admission Process

The LAU School of Medicine (SOM) follows the American model of medical education and offers a 4-year Doctor of Medicine program. Established in 2009, the SOM welcomed its first cohort the same year, initially enrolling 23 students; class size then increased progressively, in line with university strategy, to 64 students. Admission is conducted through a 2-stage process (Figure 1). In the first stage, applications are reviewed based on academic criteria including MCAT and uGPA scores. The uGPA incorporates both the cumulative GPA and a core GPA calculated from grades in prerequisite courses such as biology, biochemistry, and other disciplines considered essential for medical studies by the school. Applicants must meet specific minimum thresholds for each component (MCAT, cumulative GPA, and core GPA), below which they are automatically disqualified from consideration.

XSL•FO
RenderX

Figure 1. Admission process at the Lebanese American University (LAU) School of Medicine. GPA: grade point average; MCAT: Medical College Admission Test.



In the second stage, selected candidates are invited for interviews designed to assess nonacademic qualities. These interviews include traditional panel interviews, as well as multiple mini-interview (MMI) stations, which were introduced at the school in 2018. The outcomes from both stages are combined into a final composite score, with weightings determined by the admissions committee. Finally, applicants enrolled in the honors program and those with exceptional talents, research experience, or unique achievements are awarded up to 2 merit points (over 100), which are added to their composite score. Over time, the relative weighting of composite score components has been adjusted through continuous programmatic evaluation of the admissions process.

The admissions committee itself reflects a broad representation from across the university. Membership includes SOM faculty from basic science and clinical departments, faculty from undergraduate programs, as well as medical students, graduates of the SOM, and representatives from various university offices. Admissions decisions follow a structured deliberation process:

XSL•FO
RenderX

they begin with a framework developed by admissions leadership that aligns with the school's mission and institutional priorities, incorporate data from internal validation studies, and undergo iterative discussion and consensus-building among committee members before final approval.

During the 2 academic years when the MCAT was suspended due to the COVID-19 pandemic, the committee maintained the same deliberative process while adjusting the admissions model. In place of the MCAT, the committee adopted an alternative weighting scheme that emphasized uGPA and interview performance. This weighting was based on the collective wisdom and consensus of the committee and subsequently approved through the school's established procedures for modifying admissions processes. As a result, admissions decisions during this period continued to follow the same rigorous, mission-driven, and data-informed framework used in all other years (Table 1). To ensure consistent implementation, interviewers received standardized guidance and scoring rubrics (adapted for remote delivery where needed), panel composition and scoring protocols were preserved, minimum thresholds for uGPA and core GPA remained unchanged, and merit or honors point allocation followed existing rules; all pandemic-era interview ratings were recorded in the institutional database, and interviewers were asked to flag any connectivity or format issues to support individualized review of borderline applicants.

Table . Weight distribution of admission composites before and during the Medical College Admission Test (MCAT) suspension.

Component	Cohorts of 2009 - 10 to 2020 - 21	Cohorts of 2021 - 22 and 2022 - 23
Cumulative GPA[a], %	25 - 30	40
Core GPA, %	20	35
MCAT, %	30 - 35	—[b]
Interview[c], %	15 - 25	25

[a]GPA: grade point average.

[b]Not available.

[c]Includes traditional interviews with multiple mini-interview stations since 2018.

Study Design and Data Collection

This study is a retrospective cohort analysis of 1898 candidates who submitted complete applications to the Doctor of Medicine program at LAU from 2008 (for the 2009 - 2010 academic year, when the program was launched) to 2021 (for the 2022 - 2023 academic year). For each applicant, we collected data on cumulative GPA, core GPA, MCAT score (when available), and interview grade. For admitted candidates, additional performance measures were collected, consisting of end-of-year grades for Med 1 through Med 4 and average objective structured clinical examination (OSCE) scores from Med 1 and Med 2. Throughout the study period, the SOM maintained standardized grading procedures, assessment frameworks, and OSCE evaluation methods across cohorts. Although minor curricular refinements occurred over time, the core competency framework and evaluation principles remained consistent.

Analytical Approach (Aligned with Objectives)

Evaluating the Role of the MCAT in Predicting Medical School Performance (Objective 1)

To address objective 1, we recalculated admissions composites with and without the MCAT and compared their associations with subsequent medical school outcomes. In other words, we examined whether the actual admissions model that includes MCAT scores shows a stronger association with subsequent student performance compared with an MCAT-excluded admissions model. For the 349 admitted students in years when the MCAT was administered, we recalculated a hypothetical admissions composite that excluded the MCAT, using the weighting scheme that was adopted by the admissions committee during the COVID-19 suspension period (the right column of Table 1).

This approach generated 2 parallel admission scores for each student: 1 incorporating the MCAT and 1 excluding it. We then tested, using Pearson correlation analysis, the association between each score and subsequent medical school performance, as measured by final-year grades from Med 1 to Med 4 and average OSCE scores from Med 1 and Med 2. We also examined the nonlinear relationship between admission scores (with and without MCAT) and academic outcomes (average OSCE scores and average final-year grades) using generalized additive models with cubic splines. For each outcome, we had 2 models: 1 using the admission score with MCAT and 1 without it. To adjust for overfitting, given our small sample size of 349 admitted students, model performance was evaluated using R^2 obtained from 10-fold cross-validation. The additional value provided by MCAT was determined by calculating the difference between R^2 of the model with MCAT and R^2 of the model without MCAT. The 95% CIs for this difference were calculated using 500 bootstrap resamples.

To examine whether the value of MCAT differs across students' admission ranks, we tested the interaction between MCAT score and admission rank in a linear regression model with average final-year grade as the outcome, controlling for cumulative GPA, core GPA, interview score, honors program status, and merit score. Because a single interaction term assumes a uniform linear change across ranks, we complemented this analysis by stratifying students into tertiles based on admission composite scores: top third (n=119), middle third (n=117), bottom third (n=113), and recalculated the correlation with the average final-year grade for each tier. This allowed us to determine the predictive value of MCAT specifically for the most critical

XSL•FO
RenderX

students in the bottom tier, who are at the greatest academic risk.

Developing an ML Model to Predict MCAT (Objective 2)

To address objective 2, we trained and validated an ensemble ML model to predict MCAT scores from preadmission data and assessed the impact of predicted scores on rankings and outcome correlations. For this, we used data from 1583 medical school applicants with available MCAT scores from academic years 2009 - 2010 to 2020 - 2021 to train an ensemble of 5 models to predict MCAT scores based on the following preadmission information: core GPA, cumulative GPA, interview score, merit points, and participation in honors programs. The predictions from all 5 models were then averaged with equal weights to produce final MCAT score estimates. We then used this model to predict MCAT scores for applicants who did not take the exam due to the COVID-19 pandemic during the 2021 to 2022 (n=167) and 2022 to 2023 (n=148) academic years and who qualified for the 2 stages of admission. These predicted scores were incorporated into the overall admission scoring alongside undergraduate GPA and interview scores (using the prepandemic weighting shown in Table 1) to produce new rankings. We then (1) assessed the agreement between rankings produced with and without the predicted MCAT scores and (2) compared the associations between admission scores and medical school performance metrics, including final-year grades from Med 1 to Med 3 (Med 4 data were not yet available for these cohorts at the time of the study) and average OSCE grades from Med 1 and Med 2.

Statistical Analysis

To address objective 1 and analyze the impact of MCAT, we calculated the correlation coefficients between each admission score (with and without MCAT) and medical school grades. The difference between these correlations was tested using the Steiger Z test.

To address objective 2 and develop an ML model that predicts the MCAT score, the dataset was split into training (60%), validation (10%), and testing (30%) sets using stratified random sampling to maintain score distribution across sets. The validation set was used to assess model behavior during development and to confirm the equal-weight averaging strategy. Final model performance was evaluated on the independent test set. Missing values were imputed using MissForest, a method capable of handling nonlinear relationships and complex interactions [28]. We used an ensemble of 5 models that combined least absolute shrinkage and selection operator, kernel ridge, gradient boosting, elastic net, and light gradient boosting machine (LightGBM) models. All admission scores were scaled before modeling, as some of the models used are sensitive to the scale of predictors and variables with larger ranges would otherwise dominate the regularization penalty and therefore bias the results. Performance was evaluated using R^2 (proportion of variance explained) and root mean squared error (RMSE, average prediction error on the original MCAT scale after denormalization). Feature importance was assessed through permutation importance analysis, which measures the decrease in model performance when each feature is randomly shuffled. We then used the model to predict MCAT scores for admissions during COVID-19, and therefore produce new rankings based on it, as an alternative to the existing method. Agreement between these 2 admission scoring methods was assessed using the Kendall rank correlation coefficient (τ). All statistical tests were 2-sided, and $P<.05$ was considered the threshold for statistical significance.

Data were collected and aggregated using Python (version 3.13.2), and subsequent analyses were performed in R (version 4.5.1; R Foundation for Statistical Computing).

Ethical Considerations

This study was deemed exempt from full review by the Lebanese American University Institutional Review Board. All admission and medical school performance information was deidentified prior to analysis. Individual informed consent was not obtained. This was a retrospective study using existing student data routinely collected and maintained by the institution for administrative and academic purposes. No direct contact with students occurred for the purposes of this study. The review board waived the requirement for individual informed consent.

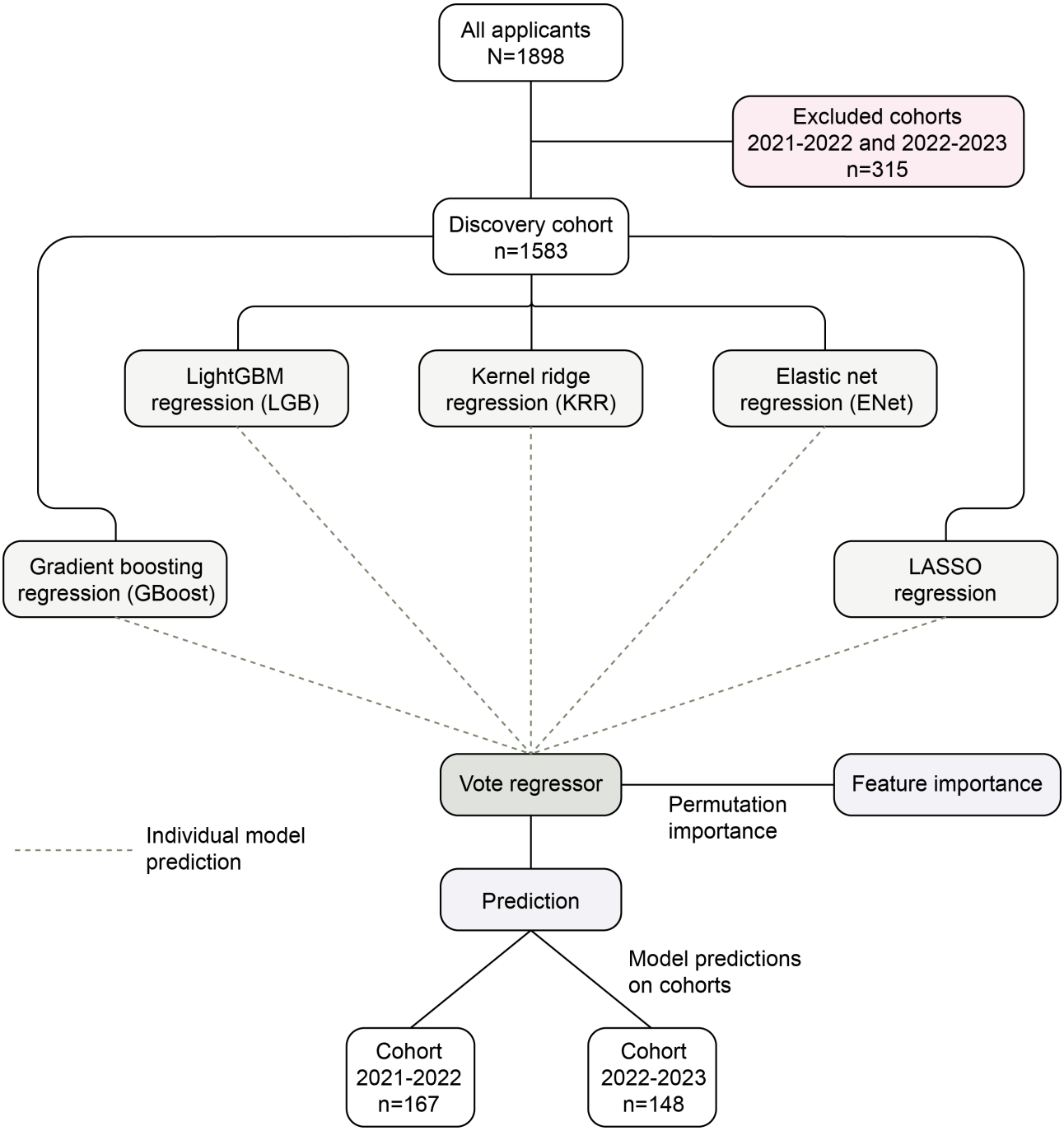
Results

Sample Description

Our sample consists of 1898 candidates with complete applications from 2009 to 2023. The subset used for modeling excludes the last 2 cohorts of 2021 to 2022 and 2022 to 2023, who did not take the MCAT exam due to the COVID-19 pandemic (Figure 2). The description of this subset is presented in Table 2.

XSL•FO
RenderX

Figure 2. Steps in modeling the Medical College Admission Test (MCAT) score. GBM: gradient-boosting machine; LASSO: least absolute shrinkage and selection operator.



XSL•FO
RenderX

Table . Description of the subset used in modeling Medical College Admission Test (MCAT) scores (N=1583)[a].

Feature	Values	Missing values
Honors program participation, n (%)		
No	1480 (93.49)	—[b]
Yes	103 (6.51)	—
Merit points, n (%)		
0	1539 (97.22)	—
1	36 (2.27)	—
2	8 (0.51)	—
Scaled interview score (over 100), mean (SD)	81 (11)	678
Cumulative GPA[c] score (over 100), mean (SD)	51.1 (34)	12
Core GPA score (over 100), mean (SD)	37.3 (36)	26
MCAT score (over 100), mean (SD)	67 (12)	—

[a]GPA and interview scores were transformed to a common 0 - 100 normalized scale for analysis. Therefore, the reported values represent normalized scores rather than conventional GPA values.

[b]Not applicable.

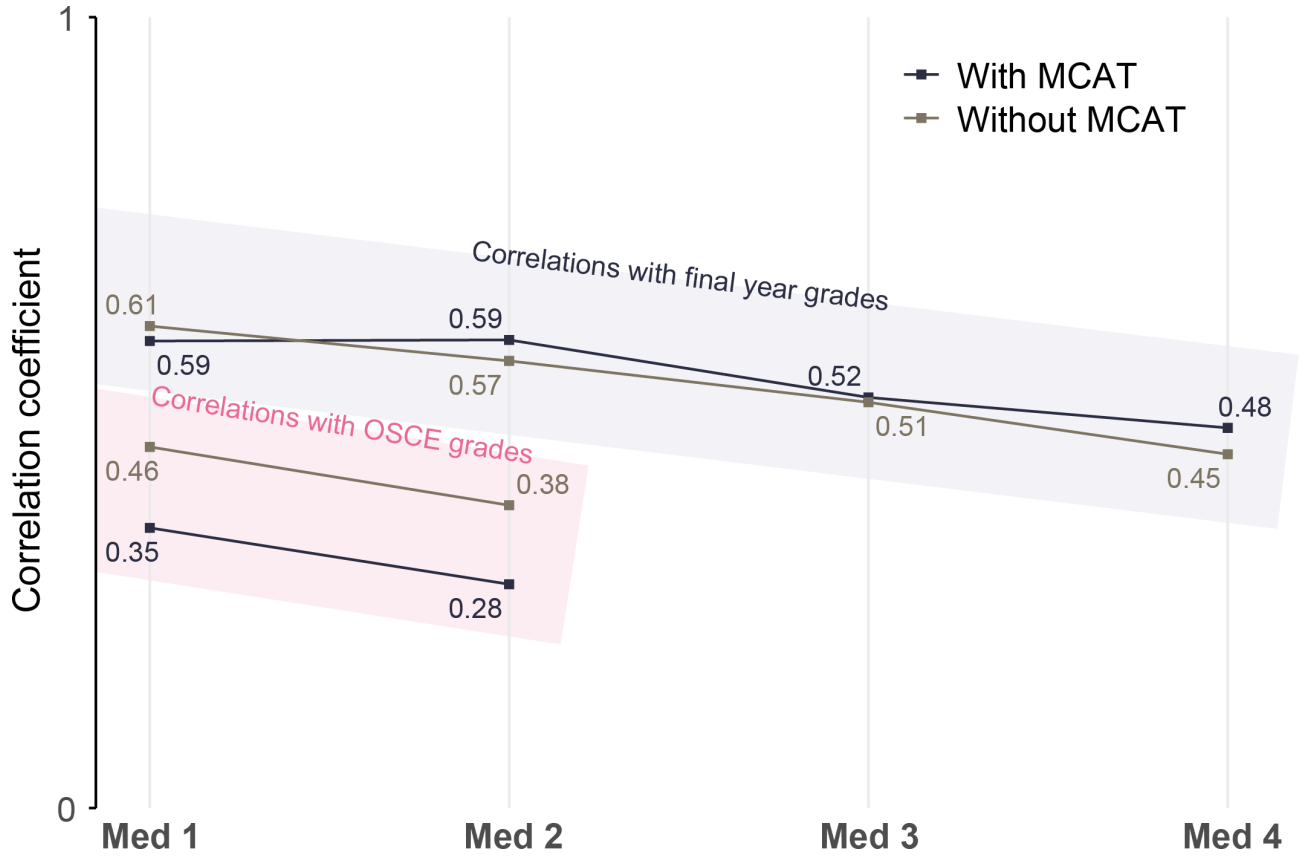
[c]GPA: grade point average.

Impact of Including the MCAT Score in Admissions Decisions on Future Performance (Objective 1)

Including the MCAT score in the admissions composite did not improve prediction of overall academic performance (n=349 admitted students with MCAT scores and complete outcome data; academic years 2012 - 2013 till 2018 - 2019). The correlation between the admissions composite score and the average final grade (an aggregate of Med 1 - 4 performance) was 0.63 (95% CI 0.56-0.69) when the MCAT was included, compared to 0.62 (95% CI 0.55-0.68) without it, and this difference was not statistically significant (Steiger Z=0.48, P=.63). By contrast, for clinical performance as measured by the average OSCE scores from Med 1 and Med 2, the correlation was significantly lower (Steiger Z=−5.35, P<.001) when the MCAT was included (r=0.37, 95% CI 0.27-0.46) than when it was excluded (r=0.48, 95% CI 0.40-0.56). Year-by-year correlations are presented in Figure 3, which shows that this pattern was consistent over time. Nonlinear analysis using generalized additive model was consistent with linear findings. For final-year grades, the difference in R^2 (with MCAT-without MCAT) was ΔR^2=−0.01 (95% CI −0.04 to 0.03), indicating no meaningful contribution of the MCAT to predictive accuracy. For OSCE grades, the model without MCAT outperformed the model with MCAT, with ΔR^2=−0.11 (95% CI –0.16 to –0.07), suggesting that including the MCAT in the admission score reduced the prediction of OSCE performance.

XSL•FO
RenderX

Figure 3. Comparison of the 2 admission models based on the correlation between their scores and grades from Med 1 to Med 4. MCAT: Medical College Admission Test; OSCE: objective structured clinical examination.



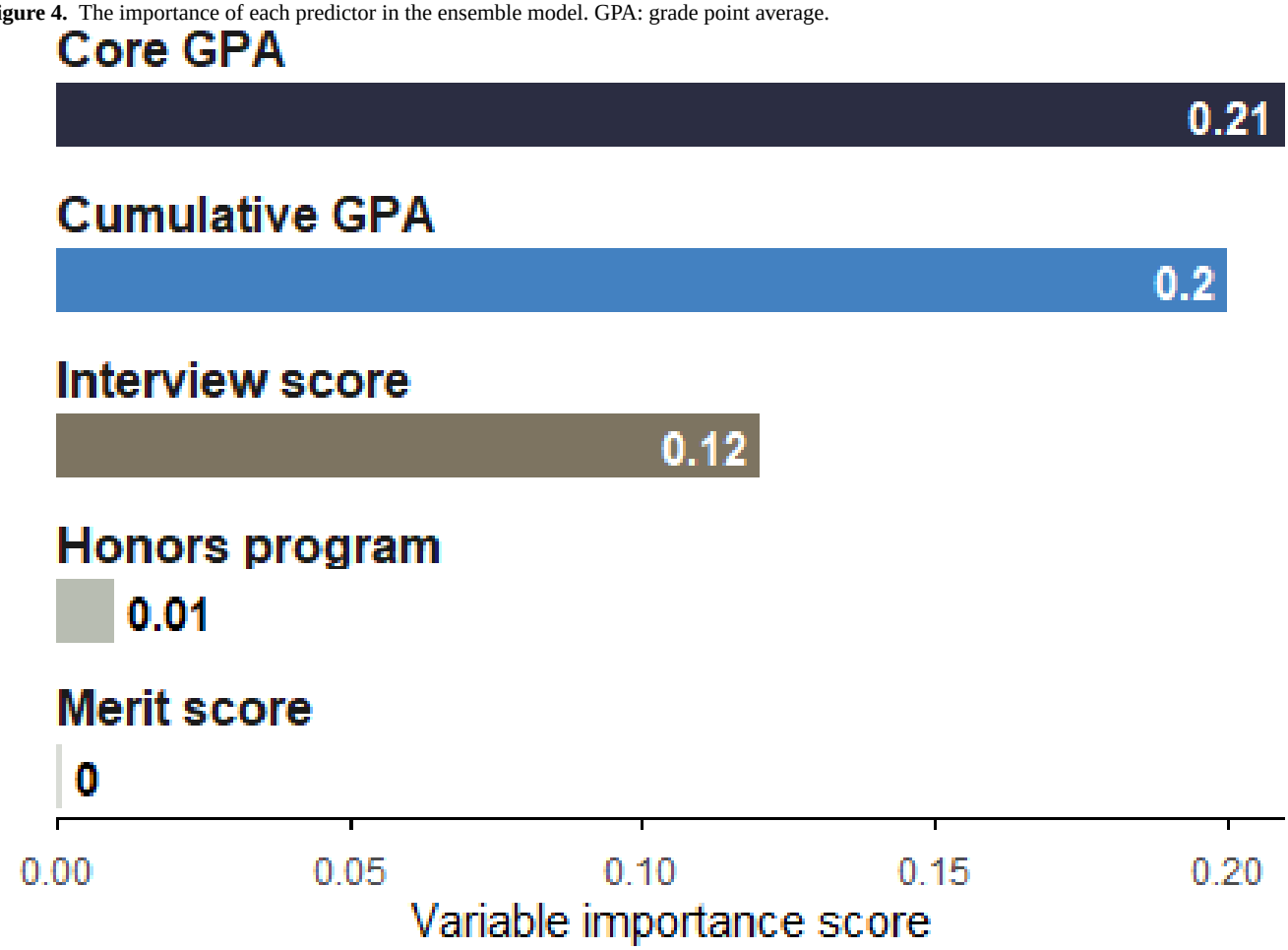
After dividing the admitted cohort (n=349) into 3 tertiles based on admission composite scores including MCAT, we recalculated the correlation with the average final-year grade for each tier. In the top third (n=119), the score including MCAT showed stronger correlation than the score without MCAT, with r=0.59 (95% CI 0.46-0.69) vs 0.52 (95% CI 0.37-0.64) and both P<.001. In the middle tertile (n=117), the correlation was nearly identical whether MCAT was included (r=0.33, 95% CI 0.15-0.48; P<.001) or excluded (r=0.34, 95% CI 0.17-0.49; P<.001). Most notably, in the bottom tertile (n=113), the score including MCAT showed no significant correlation with final-year grades (r=0.12, 95% CI −0.07 to 0.30; P=.21), whereas the score without MCAT maintained a significant small-to-moderate correlation (r=0.28, 95% CI 0.10-0.44; P=.003). However, interaction between MCAT score and admission rank was not statistically significant in a linear regression model predicting average final-year grade (β coefficient=−0.002, 95% CI −0.004 to <0.001; P=.07), although the model may have been underpowered to detect a small interaction effect.

Modeling an MCAT Alternative and Its Decision Impact (Objective 2)

We trained an ensemble model to predict the MCAT score using core GPA, cumulative GPA, interview score, merit score, and participation in honors programs. The model achieved an out-of-sample R^2 of 0.36 (95% CI 0.32-0.41) and an RMSE of 0.10 on the normalized scale, meaning it explained 36% of the variability in MCAT scores and predicted them with an average error of 0.1. Since the MCAT score is normalized (ranging from 0-1 in our analysis), this average error of 0.1 corresponds to 10% of the score range (eg, ~5 - 6 MCAT points if mapping to 472 - 528), which is large enough to reorder applicants near admissions cutoffs. The feature importance plot (Figure 4) shows that core and cumulative GPAs were the most influential factors in predicting the MCAT score, while honors and merit points have little to no effect.

XSL•FO
RenderX

Figure 4. The importance of each predictor in the ensemble model. GPA: grade point average.



We used the model to predict MCAT scores for the cohorts during which the exam was suspended (2021 - 2022 and 2022 - 2023) and to calculate new rankings and admission scores using the predicted MCAT. The Kendall rank correlation between the original admission rankings (based solely on GPA and interviews during the MCAT suspension period) and the model-based ranking was 0.80 (95% CI 0.77-0.83; *P*<.001), indicating a strong overall agreement. We found that, had predicted MCAT scores been used, 6 (9.4%) of the 64 admitted students in 2021 - 2022 and 4 (6.3%) of 64 in 2022 - 2023 would not have been admitted, with an equal number of applicants from the rejected pool selected instead. With respect to merit scholarships, 1 award in 2021 - 2022 and 2 awards in 2022 to 2023 would have been reassigned to different students.

However, the new admission scores (with predicted MCAT) showed weaker correlations across all measures of future performance in medical school compared with the actual admission scores from those years (Table 3). This was expected, since the model's 10% average prediction error introduced significant noise into the new scores.

Table . Correlation between admission scores, including and excluding predicted modeling of Medical College Admission Test (MCAT) scores and later performance in medical school.

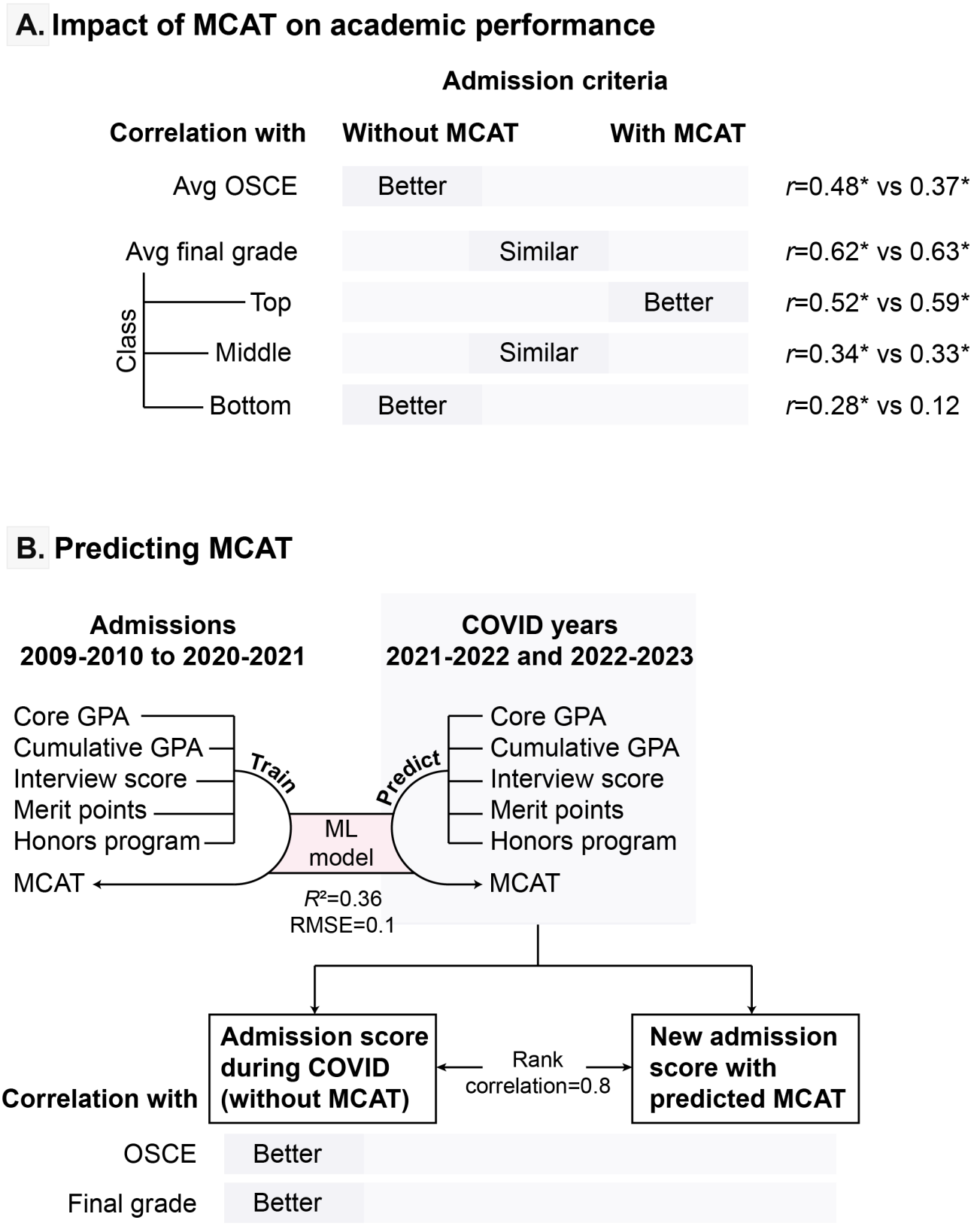
Grade type and medical year	Sample size, n	Original admission score		Model-based admission score	
		Correlation coefficient (95% CI)	*P* value	Correlation coefficient (95% CI)	*P* value
OSCE[a]					
1	129	0.28 (0.11-0.43)	.002	0.24 (0.07-0.40)	.006
2	126	0.33 (0.17-0.48)	<.001	0.27 (0.10-0.43)	.002
Final-year grade					
1	128	0.48 (0.33-0.60)	<.001	0.42 (0.27-0.56)	<.001
2	126	0.46 (0.31-0.59)	<.001	0.39 (0.24-0.53)	<.001
3	63	0.33 (0.09-0.54)	.008	0.23 (-0.01-0.46)	.07

[a]OSCE: objective structured clinical examination.

XSL•FO
RenderX

Figure 5 summarizes the key findings across all analyses. Including the MCAT in the admissions score did not improve the prediction of final-year grades and was associated with worse prediction of OSCE grades. Furthermore, when we predicted MCAT scores from other admission criteria and incorporated it into the admission score, predictive performance was also reduced for both outcomes.

Figure 5. High-level summary of the results. (A) Comparison of correlations with academic performance of admission scores calculated with and without Medical College Admission Test (MCAT). (B) A predicted MCAT score was used to calculate an admission score during COVID years, and its impact on future academic performance was analyzed. *A statistically significant correlation (P<.05; Table 3). GPA: grade point average; ML: machine learning; OSCE: objective structured clinical examination; RMSE: root mean squared error.



XSL•FO
RenderX

Discussion

Principal Findings

This study examined 2 related questions: the incremental validity of observed MCAT scores for predicting medical school performance (objective 1) and the feasibility and decision-impact of ML-predicted MCAT scores when testing is unavailable (objective 2). By addressing both the role of MCAT within holistic admissions and the potential for algorithmic substitution, the findings contribute to ongoing discussions regarding standardized testing in medical education and the use of data-driven tools in admission selection.

MCAT and Academic Performance (Objective 1)

Our analysis indicates that within our institution and study population, the inclusion of MCAT scores did not substantially enhance the prediction of short-term academic outcomes beyond an already rigorous admissions process incorporating undergraduate GPA and structured interviews. Admissions composites generated during the temporary MCAT suspension showed associations with subsequent academic performance that were largely comparable to those observed under the traditional MCAT-inclusive model. Practically, this implies that uGPA plus structured interviews (and other holistic factors) can provide a robust basis for selection in the short term.

Our findings, however, should not be interpreted as evidence against the value of standardized testing altogether. Historically, the MCAT has served as a standardized comparator across heterogeneous educational backgrounds within the North American graduate-entry medical admissions model. Rather, our results suggest that the incremental contribution of MCAT scores may become smaller when admissions systems already incorporate multiple academic and nonacademic measures.

These findings should be interpreted alongside the broader literature. Large multisite studies and recent meta-analyses have consistently demonstrated that MCAT scores are moderately associated with medical school and licensing examination performance [7,8,29-31]. However, single-institution studies have also shown comparable early outcomes with MCAT-blinded admissions after controlling for uGPA [32]. Thus, while our results suggest that robust holistic processes can mitigate the short-term impact of MCAT suspension at our school, they should not be interpreted as demonstrating limited incremental value of the MCAT across medical schools. Institutional differences in applicant pools, admissions weightings, curricular structure, assessment methods, and resource availability may produce different outcomes elsewhere.

There are several reasons why our results failed to show an incremental value of the MCAT in predicting future performance. First, our holistic admissions process may already capture much of what the MCAT measures. Evidence suggests that when noncognitive assessments are robust, the incremental validity of standardized tests diminishes [18,33]. Second, our relatively small sample size (n=349) and single-institution design limit statistical power to detect small effect sizes and restrict generalizability. Finally, and most importantly, our correlations are not corrected for restriction of range, a limitation we address in detail below.

The finding that including the MCAT in admissions scoring weakened the prediction of clinical skills performance, as measured by the average OSCE grades in Med 1 and Med 2, is entirely consistent with existing literature demonstrating that the MCAT primarily predicts knowledge-based examination performance rather than clinical assessments. For example, Saguil et al [34] found no significant association between MCAT scores and OSCE performance, step 2 clinical skills subscores, or postgraduate year 1 program director evaluations, findings that have been supported by other reports [6,7,35,36].

This pattern is conceptually expected because the MCAT primarily assesses scientific knowledge and critical reasoning, whereas OSCEs evaluate communication, clinical reasoning, professionalism, and interpersonal skills [37,38], consistent with studies demonstrating stronger associations between MMIs and clinical performance [39,40], which may explain why the non-MCAT composite (with greater relative weight on interviews) showed stronger correlations with OSCE outcomes.

Another interesting, and unexpected, finding was that among students in the lower tertile of admission scores (where we would have expected the MCAT to strengthen prediction), the MCAT-inclusive composite was not significantly correlated with final grades, as opposed to the MCAT-excluded composite. This suggests that for students at the margin of admission decisions, where selection decisions are most difficult and consequential, MCAT scores may fail to add predictive value. Students selected with lower MCAT scores (bottom tertile) may have other compensatory (noncognitive) abilities that contribute to their future success.

ML Model Performance (Objective 2)

The ML model showed limited success in predicting MCAT scores (R^2=0.36, RMSE=10%), explaining only 36% of the variance despite incorporating cumulative and core GPAs, interview scores, merit points, and honors program participation.

Although an R^2 of 0.36 may be considered substantial in some applied settings [41], it is modest for predicting one standardized academic measure from other academic variables, suggesting that the MCAT captures information not fully reflected by GPA and interview performance. Nevertheless, even perfect prediction of MCAT scores would have limited impact in our setting, as our primary finding was that the MCAT itself added negligible incremental validity to admission decisions.

Incorporating predicted MCAT scores into the composite for the 2 cohorts admitted during the MCAT suspension produced rankings that were strongly aligned with the original admissions outcomes, yet it would have altered the admission of 10 students across the 2 years, with those students replaced by others from the previously rejected pool. Importantly, the model's RMSE corresponds to an error of several MCAT points, large enough to shift rank order for applicants near the admissions cutoff. Accordingly, despite strong overall rank correlation, predicted MCAT values should not be treated as interchangeable substitutes for observed MCAT scores in high-stakes decision-making.

XSL•FO
RenderX

These findings do not reflect a flaw in the modeling strategy. We used a robust pipeline and an ensemble of linear and nonlinear learners, yet prediction error remained large enough to influence rank order near the admissions threshold. This likely reflects an inherent limitation of attempting to reconstruct a standardized test score from a limited set of admissions variables that were not designed to reproduce the MCAT at the individual level. In practice, while predicted MCAT scores may approximate population-level trends, they are not sufficiently precise for high-stakes individual ranking decisions.

These findings highlight an important limitation of current ML approaches. While ML may support tasks such as initial applicant screening [42], detecting biases in human decisions [43], or identifying patterns in large datasets [44], our results suggest that it is not sufficiently accurate to replace standardized test scores in high-stakes admissions decisions.

A key takeaway from this modeling strategy is the importance of considering multiple metrics for model performance. While R^2 is a commonly used measure, it does not provide a complete picture on its own. The RMSE, which reflects the model's average error, should also be considered. Although there is no universally accepted threshold for RMSE, it should be interpreted based on the specific context.

Limitations

This study has several important limitations that warrant consideration when interpreting our findings.

First, our findings come from one medical school with a specific admissions process, curriculum, and student population, which limits generalizability to other institutions and contexts. Consequently, external validity is limited. Other schools that rely more heavily on standardized testing, that use different interview formats (eg, interview calibration and rubric design), or that have distinct curricular or assessment systems may observe different incremental validity for the MCAT. Multi-institutional studies with variation in admissions models and outcome measures are required before recommending broader policy change. Second, our analyses were limited to admitted students, introducing restriction-of-range effects that likely reduced the magnitude of observed correlations relative to the full applicant pool. Because outcome data were unavailable for rejected applicants, formal correction for range restriction could not be performed. Therefore, our findings likely underestimate true population-level associations and should be interpreted cautiously when compared with range-corrected studies. This limitation likely affects both the MCAT and non-MCAT admissions models, although the extent of its impact cannot be determined from our data. Therefore, the comparative analyses should be interpreted cautiously. Third, we relied only on medical school grades and early OSCE scores (Med 1 - 2 only) as performance indicators. We did not examine longer-term outcomes such as licensing examination performance (United States Medical Licensing Examination [USMLE] Steps), residency matching success, clinical performance in residency, or ultimate career trajectories. Fourth, we used Pearson correlations to assess predictive validity, which capture only linear relationships. Important nonlinear patterns or interaction effects may exist that our analytical approach did not detect. Finally, our admissions process evolved over the study period, particularly with MMI introduction in 2018 and weight adjustments. While we accounted for these changes analytically, the evolving nature of our process introduces some heterogeneity. Because of the relatively small sample size, we were unable to perform meaningful sensitivity analyses by time period (eg, pre-MMI vs post-MMI implementation), and temporal changes may therefore have influenced the observed associations. Additionally, the COVID-19 pandemic brought further potential confounders, such as remote or virtual interview formats and short-term curricular modifications, which may have affected interview scores, early OSCE experiences, and the relationships between admissions metrics and outcomes. These pandemic-related changes could therefore have contributed to greater variability in the pandemic cohorts and should be taken into account when generalizing our findings.

Conclusions

In this single-institution study, we addressed 2 related questions: whether including MCAT scores adds incremental validity to admissions decisions and whether an ML model can substitute for the MCAT when testing is unavailable. Our findings suggest that when the MCAT is temporarily unavailable, a well-designed holistic admissions process emphasizing careful GPA review and interviews maintained short-term student quality. Accordingly, the value of this analysis lies less in proposing a new admissions tool than in demonstrating the current limitations of ML as a substitute for standardized testing in this context. Rather than attempting to recreate MCAT scores, efforts should focus on strengthening validated noncognitive assessments, particularly interviews such as the MMI, while maintaining individualized review of applicants near the admissions threshold. Institutions considering reduced reliance on the MCAT should validate such changes locally (eg, via retrospective simulations or pilot cohorts) and monitor medium-term and long-term outcomes. More broadly, within the context of this single-institution study, our findings support viewing standardized testing as one component of a comprehensive, evidence-informed admissions process, rather than as a standalone admission criterion. Further, multi-institutional studies are needed before broader conclusions regarding the incremental value of the MCAT can be drawn and to evaluate alternative approaches for MCAT estimation.

Acknowledgments

All authors declared that they had insufficient funding to support open access publication of this manuscript, including from affiliated organizations or institutions, funding agencies, or other organizations. JMIR Publications provided article processing fee (APF) support for the publication of this article.

XSL•FO
RenderX

ChatGPT, specifically the GPT-5 model from OpenAI, was used to improve the clarity and readability of some sentences in the manuscript. All AI-generated outputs were reviewed and revised by the authors, who take full responsibility for the content of the manuscript.

Funding

The authors declared no financial support was received for this work.

Data Availability

The datasets generated or analyzed during this study are available from the corresponding author upon reasonable request.

Authors' Contributions

Conceptualization: SAB (lead), MC (equal)
Data curation: MG
Formal analysis: MC (lead), MG (equal), GC (supporting)
Methodology: MC (lead), MG (equal), SAB (supporting)
Supervision: SAB (lead), TER (supporting), SMN (supporting)
Validation: SAB (lead), GC (supporting)
Visualization: MG (lead), GC (equal)
Writing – original draft: MC (lead), SAB (equal), MG (supporting)
Writing – review & editing: SAB (lead), MC (supporting), TER (supporting), SMN (supporting), GC (supporting)

Conflicts of Interest

None declared.

References

1. Kirch DG. Transforming admissions: the gateway to medicine. JAMA 2012 Dec 5;308(21):2250-2251. [doi: 10.1001/jama.2012.74126] [Medline: 23212504]
2. Lovell B. What do we know about coaching in medical education? A literature review. Med Educ 2018 Apr;52(4):376-390. [doi: 10.1111/medu.13482] [Medline: 29226349]
3. Schreurs S, Cleutjens KB, Muijtjens AMM, Cleland J, Oude Egbrink MGA. Selection into medicine: the predictive validity of an outcome-based procedure. BMC Med Educ 2018 Sep 17;18(1):214. [doi: 10.1186/s12909-018-1316-x] [Medline: 30223816]
4. Dunleavy DM, Kroopnick MH, Dowd KW, Searcy CA, Zhao X. The predictive validity of the MCAT exam in relation to academic performance through medical school: a national cohort study of 2001-2004 matriculants. Acad Med 2013 May;88(5):666-671. [doi: 10.1097/ACM.0b013e3182864299] [Medline: 23478635]
5. Kirch DG, Mitchell K, Ast C. The new 2015 MCAT: testing competencies. JAMA 2013 Dec 4;310(21):2243-2244. [doi: 10.1001/jama.2013.282093] [Medline: 24302080]
6. Donnon T, Paolucci EO, Violato C. The predictive validity of the MCAT for medical school performance and medical board licensing examinations: a meta-analysis of the published research. Acad Med 2007 Jan;82(1):100-106. [doi: 10.1097/01.ACM.0000249878.25186.b7] [Medline: 17198300]
7. Busche K, Elks ML, Hanson JT, et al. The validity of scores from the new MCAT exam in predicting student performance: results from a multisite study. Acad Med 2020 Mar;95(3):387-395. [doi: 10.1097/ACM.0000000000002942] [Medline: 31425189]
8. Violato C, Gauer JL, Violato EM, Patel D. A study of the validity of the new MCAT exam. Acad Med 2020 Mar;95(3):396-400. [doi: 10.1097/ACM.0000000000003064] [Medline: 31702690]
9. Casey PM, Palmer BA, Thompson GB, et al. Predictors of medical school clerkship performance: a multispecialty longitudinal analysis of standardized examination scores and clinical assessments. BMC Med Educ 2016 Apr 27;16(1):128. [doi: 10.1186/s12909-016-0652-y] [Medline: 27121276]
10. Eskander A, Shandling M, Hanson MD. Should the MCAT exam be used for medical school admissions in Canada? Acad Med 2013 May;88(5):572-580. [doi: 10.1097/ACM.0b013e31828b85af] [Medline: 23524935]
11. Kuncel NR, Hezlett SA. Assessment. Standardized tests predict graduate students' success. Science 2007 Feb 23;315(5815):1080-1081. [doi: 10.1126/science.1136618] [Medline: 17322046]
12. Julian ER. Validity of the Medical College Admission Test for predicting medical school performance. Acad Med 2005 Oct;80(10):910-917. [doi: 10.1097/00001888-200510000-00010] [Medline: 16186610]
13. Grbic D, Jones DJ, Case ST. The role of socioeconomic status in medical school admissions: validation of a socioeconomic indicator for use in medical school admissions. Acad Med 2015 Jul;90(7):953-960. [doi: 10.1097/ACM.0000000000000653] [Medline: 25629949]

XSL•FO
RenderX

14. Shahriar AA, Englander R, Che Y, Carter K. Socioeconomic background, exam performance, and probability of medical school application among all MCAT examinees, 2017 to 2019. BMC Med Educ 2026 Mar 29;26(1):727. [doi: 10.1186/s12909-026-08940-2] [Medline: 41904533]
15. Pietschnig J, Voracek M. One century of global IQ gains: a formal meta-analysis of the Flynn effect (1909-2013). Perspect Psychol Sci 2015 May;10(3):282-306. [doi: 10.1177/1745691615577701] [Medline: 25987509]
16. Trahan L, Stuebing KK, Fletcher JM, Hiscock MK. The Flynn effect: a meta-analysis. Psychol Bull 2014 Sep;140(5):1332-1360. [doi: 10.1037/a0037173] [Medline: 24979188]
17. Witzburg RA, Sondheimer HM. Holistic review--shaping the medical profession one applicant at a time. N Engl J Med 2013 Apr 25;368(17):1565-1567. [doi: 10.1056/NEJMp1300411] [Medline: 23574032]
18. Patterson F, Knight A, Dowell J, Nicholson S, Cousans F, Cleland J. How effective are selection methods in medical education? A systematic review. Med Educ 2016 Jan;50(1):36-60. [doi: 10.1111/medu.12817] [Medline: 26695465]
19. Corridon PR. Admissions to medical school during the COVID-19 era without the MCAT. J Med Educ Curric Dev 2021;8:23821205211014898. [doi: 10.1177/23821205211014898] [Medline: 33997289]
20. Burkhardt JC, Stansfield RB, Vohra T, Losman E, Turner-Lawrence D, Hopson LR. Prognostic value of the multiple mini-interview for emergency medicine residency performance. J Emerg Med 2015 Aug;49(2):196-202. [doi: 10.1016/j.jemermed.2015.02.008] [Medline: 25937476]
21. DeVaul RA, Jervey F, Chappell JA, Caver P, Short B, O'Keefe S. Medical school performance of initially rejected students. JAMA 1987 Jan 2;257(1):47-51. [doi: 10.1001/jama.1987.03390010051027] [Medline: 3783902]
22. Eva KW, Reiter HI. Where judgement fails: pitfalls in the selection process for medical personnel. Adv Health Sci Educ Theory Pract 2004;9(2):161-174. [doi: 10.1023/B:AHSE.0000027479.14637.6c] [Medline: 15141133]
23. Kelly ME, Dowell J, Husbands A, et al. The fairness, predictive validity and acceptability of multiple mini interview in an internationally diverse student population--a mixed methods study. BMC Med Educ 2014 Dec 21;14(1):267. [doi: 10.1186/s12909-014-0267-0] [Medline: 25528046]
24. Kreiter CD, Yin P, Solow C, Brennan RL. Investigating the reliability of the medical school admissions interview. Adv Health Sci Educ Theory Pract 2004;9(2):147-159. [doi: 10.1023/B:AHSE.0000027464.22411.0f] [Medline: 15141132]
25. Salvatori P. Reliability and validity of admissions tools used to select students for the health professions. Adv Health Sci Educ Theory Pract 2001;6(2):159-175. [doi: 10.1023/a:1011489618208] [Medline: 11435766]
26. Shaw DL, Martz DM, Lancaster CJ, Sade RM. Influence of medical school applicants' demographic and cognitive characteristics on interviewers' ratings of noncognitive traits. Acad Med 1995 Jun;70(6):532-536. [doi: 10.1097/00001888-199506000-00015] [Medline: 7786374]
27. Gay SE, Santen SA, Mangrulkar RS, Sisson TH, Ross PT, Zaidi NLB. The influence of MCAT and GPA preadmission academic metrics on interview scores. Adv Health Sci Educ Theory Pract 2018 Mar;23(1):151-158. [doi: 10.1007/s10459-017-9779-9] [Medline: 28501933]
28. Stekhoven DJ, Bühlmann P. MissForest--non-parametric missing value imputation for mixed-type data. Bioinformatics 2012 Jan 1;28(1):112-118. [doi: 10.1093/bioinformatics/btr597] [Medline: 22039212]
29. Hanson JT, Busche K, Elks ML, et al. The validity of MCAT scores in predicting students' performance and progress in medical school: results from a multisite study. Acad Med 2022 Sep 1;97(9):1374-1384. [doi: 10.1097/ACM.0000000000004754] [Medline: 35612915]
30. MCAT validity data report. : Association of American Medical Colleges (AAMC); 2024 URL: https://www.aamc.org/media/78266/download [accessed 2026-08-13]
31. Harvey B, Ahsan MD, Chandler IR, et al. Forecasting medical school success: a systematic review and meta-analysis of the predictive value of MCAT scores. J Natl Med Assoc 2025 Dec;117(6):479-489. [doi: 10.1016/j.jnma.2025.08.108] [Medline: 40973544]
32. Teng Y, Dong T, Durning SJ, Jung E, Saguil A, Liotta R. Blinding of admissions committee to MCAT scores: is there a difference in subsequent medical school performance between cohorts? Mil Med 2023 May 18;188(Suppl 2):7-10. [doi: 10.1093/milmed/usad010] [Medline: 37201485]
33. Dore KL, Kreuger S, Ladhani M, et al. The reliability and acceptability of the multiple mini-interview as a selection instrument for postgraduate admissions. Acad Med 2010 Oct;85(10 Suppl):S60-S63. [doi: 10.1097/ACM.0b013e3181ed442b] [Medline: 20881706]
34. Saguil A, Dong T, Gingerich RJ, et al. Does the MCAT predict medical school and PGY-1 performance? Mil Med 2015 Apr;180(4 Suppl):4-11. [doi: 10.7205/MILMED-D-14-00550] [Medline: 25850120]
35. Gauer JL, Wolff JM, Jackson JB. Do MCAT scores predict USMLE scores? An analysis on 5 years of medical student data. Med Educ Online 2016;21:31795. [doi: 10.3402/meo.v21.31795] [Medline: 27702431]
36. Callahan CA, Hojat M, Veloski J, Erdmann JB, Gonnella JS. The predictive validity of three versions of the MCAT in relation to performance in medical school, residency, and licensing examinations: a longitudinal study of 36 classes of Jefferson Medical College. Acad Med 2010 Jun;85(6):980-987. [doi: 10.1097/ACM.0b013e3181cece3d] [Medline: 20068426]
37. Khan KZ, Ramachandran S, Gaunt K, Pushkar P. The objective structured clinical examination (OSCE): AMEE guide no. 81. Part I: an historical and theoretical perspective. Med Teach 2013 Sep;35(9):e1437-e1446. [doi: 10.3109/0142159X.2013.818634] [Medline: 23968323]

XSL•FO
RenderX

38. Chang O, Holbrook AM, Lohit S, et al. Comparability of objective structured clinical examinations (OSCEs) and written tests for assessing medical school students' competencies: a scoping review. Eval Health Prof 2023 Sep;46(3):213-224. [doi: 10.1177/01632787231165797] [Medline: 36959750]
39. Eva KW, Rosenfeld J, Reiter HI, Norman GR. An admissions OSCE: the multiple mini-interview. Med Educ 2004 Mar;38(3):314-326. [doi: 10.1046/j.1365-2923.2004.01776.x] [Medline: 14996341]
40. Pau A, Jeevaratnam K, Chen YS, Fall AA, Khoo C, Nadarajah VD. The multiple mini-interview (MMI) for student selection in health professions training - a systematic review. Med Teach 2013 Dec;35(12):1027-1041. [doi: 10.3109/0142159X.2013.829912] [Medline: 24050709]
41. Ferguson CJ. An effect size primer: a guide for clinicians and researchers. Prof Psychol Res Pract 2009;40(5):532-538. [doi: 10.1037/a0015808]
42. Triola MM, Reinstein I, Marin M, et al. Artificial intelligence screening of medical school applications: development and validation of a machine-learning algorithm. Acad Med 2023 Sep 1;98(9):1036-1043. [doi: 10.1097/ACM.0000000000005202] [Medline: 36888969]
43. Keir G, Hu W, Filippi CG, Ellenbogen L, Woldenberg R. Using artificial intelligence in medical school admissions screening to decrease inter- and intra-observer variability. JAMIA Open 2023;6(1):ooad011. [doi: 10.1093/jamiaopen/ooad011] [Medline: 36819893]
44. Sarker IH. Machine learning: algorithms, real-world applications and research directions. SN Comput Sci 2021;2(3):160. [doi: 10.1007/s42979-021-00592-x] [Medline: 33778771]

Abbreviations

GPA: grade point average
LAU: Lebanese American University
LightGBM: light gradient boosting machine
MCAT: Medical College Admission Test
ML: machine learning
MMI: multiple mini-interview
OSCE: objective structured clinical examination
RMSE: root mean squared error
SOM: School of Medicine
uGPA: undergraduate grade point average
USMLE: United States Medical Licensing Examination

Edited by D Chartash; submitted 26.Feb.2026; peer-reviewed by M Raghavan, S Sharma, W Peijnenburg; revised version received 10.Jul.2026; accepted 11.Jul.2026; published 01.Sep.2026.

Please cite as:
Cherfane M, Ghanem M, Choueiry G, Nasser SM, Elias-Rizk T, Aoun Bahous S
Evaluating the Influence of MCAT Scores on Medical Student Selection and Performance Using a Machine Learning Approach: Retrospective Cohort Study
JMIR Med Educ 2026;12:e94197
URL: https://mededu.jmir.org/2026/1/e94197
doi:10.2196/94197

© Michele Cherfane, Marc Ghanem, Georges Choueiry, Selim M Nasser, Tamina Elias-Rizk, Sola Aoun Bahous. Originally published in JMIR Medical Education (https://mededu.jmir.org), 1.Sep.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Education, is properly cited. The complete bibliographic information, a link to the original publication on https://mededu.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

What Platform Scores Miss: Multidimensional Evaluation of AI Teaching Agents in Medical Education

Hui Zhang[1], MM; Lihui Qu[2], PHD; Jianmin Zheng[1], MMgmt; Yi Xiong[1], MSc; Hongbo Bai[3], MSc; Ruiying Ji[3], MSc; Guohui Liu[3], MSc; Wanling Chen[4], MD; Zirui Cheng[5], MD; Youbang Chen[3], MSc; Chun-tao Yang[3], PhD

[1]Office of Academic Affairs, Guangzhou Medical University, Guangzhou, China

[2]Department of Physiology, School of Basic Medical Sciences, The Fourth Affiliated Hospital of Guangzhou Medical University, Guangzhou, China

[3]Guangzhou Municipal and Guangdong Provincial Key Laboratory of Protein Modification and Disease, Department of Physiology, School of Basic Medical Sciences, Guangzhou Medical University, 1 Xinzao Road, Guangzhou, Guangdong, China

[4]Department of Endocrinology, The First Affiliated Hospital of Guangzhou Medical University, Guangzhou, China

[5]Zhongshan School of Medicine, Sun Yat-Sen University, Guangzhou, China

Corresponding Author:
Chun-tao Yang, PhD
Guangzhou Municipal and Guangdong Provincial Key Laboratory of Protein Modification and Disease, Department of Physiology, School of Basic Medical Sciences, Guangzhou Medical University, 1 Xinzao Road, Guangzhou, Guangdong, China

Abstract

Background: Large language model (LLM)–based AI teaching agents are increasingly used in medical education, yet their pedagogical quality is typically judged by platform-generated scores whose scoring criteria are undisclosed and may not reflect the teaching quality of the agent.

Objective: This study aimed to develop and validate a multidimensional rubric for evaluating AI teaching agents and to examine the correspondence between platform scores and rubric-based teaching quality.

Methods: Eight AI teaching agents covering an endocrinology curriculum were deployed across 4 role-play paradigms (patient, student, expert, and family). Twenty-two fourth-year medical students generated 167 dialogues, which were scored both by the platform and by an independently applied 8-dimension rubric (100 points, covering knowledge accuracy, pedagogical guidance, knowledge coverage, role-play quality, difficulty calibration, medical safety, student engagement, and feedback quality). Each dialogue was scored 4 times by a primary evaluator (Claude Opus 4.8; mean within-model SD 0.36), with 2 additional LLMs as robustness checks; 40 dialogues spanning all agents were rescored by a medical-education expert for validation.

Results: Platform and rubric rankings diverged for most agents: the agent ranked third by the platform ranked last on rubric-based quality, and the platform's fourth-ranked agent ranked first. Agents differed most on knowledge-related dimensions (knowledge coverage coefficient of variation=27.3%) and least on role-play quality (coefficient of variation=5.7%), while difficulty calibration was a shared weakness. In a case-level observation, one agent revised specifically to strengthen empathy attained high role-play quality yet the lowest knowledge coverage of all agents. AI scores agreed with expert ratings at the total-score level (intraclass correlation coefficient=0.51) and on cognitive-process dimensions, but agreement was low for the more subjective dimensions. Student gender showed no detectable effect, though this analysis was underpowered.

Conclusions: In this exploratory study, platform-generated scores reflected a construct different from agent teaching quality and should be used as a complement rather than as the sole quality indicator. The 8-dimension rubric provides a transparent, standardized alternative that reveals differences missed by platform scores, including a lack of association between empathy and knowledge coverage that warrants attention in future agent design.

(JMIR Med Educ 2026;12:e96819) doi:10.2196/96819

KEYWORDS

AI teaching agents; educational data science; educational evaluation; endocrinology; large language models; LLMs; multidimensional rubric; precision education; role-play

Introduction

Bridging the gap between theoretical knowledge and clinical competence remains a central challenge in medical education, particularly in disciplines requiring multisystem integration such as endocrinology [1,2]. Simulated patient encounters are among the most effective approaches for developing clinical reasoning and communication skills, but their scalability is constrained by the cost and limited availability of standardized human patients [3-7]. Large language model (LLM)–based AI

XSL•FO
RenderX

agents have emerged as a promising alternative, achieving clinical fidelity comparable to standardized human patients at substantially lower cost [8-13], and they are now being adopted in medical curricula at a pace that outstrips our ability to evaluate them. Yet nearly all existing implementations adopt a single role-play paradigm in which the AI portrays a patient and the student acts as a physician, without examining whether alternative role configurations produce different teaching outcomes, or how the quality of any such agent should be judged in the first place.

These developments expose a more fundamental gap: we lack a validated way to judge whether an AI teaching agent teaches well. Existing evaluation frameworks address clinical intervention outcomes or clinical task accuracy [14-16] but not the teaching-specific dimensions that determine educational value, such as adaptive difficulty calibration, knowledge coverage, and formative feedback quality. In practice, quality is often inferred from the aggregate scores generated by commercial teaching platforms, yet these scores are produced by undisclosed criteria that cannot be independently verified, are not decomposable by dimension, and are not comparable across agents, leaving it unclear what they actually measure and whether they reflect the teaching quality of the agent or merely the surface characteristics of a student's responses. This evaluation gap has a direct consequence: without a trustworthy measure of teaching quality, the design variables that might improve these agents, including role-play configuration, content domain, and learner characteristics, such as gender, cannot be studied systematically and remain largely unexplored within a single controlled context [6,17].

Endocrinology is well suited to investigating these questions, spanning molecular mechanisms to chronic disease management within a single curricular module, yet remaining heavily reliant on traditional didactics, with simulation-based approaches underrepresented [18-20]. Against this background, we deployed 8 AI teaching agents covering a complete endocrinology curriculum, each employing a distinct role-play paradigm, and analyzed 167 student-agent dialogues. Because the platform's built-in scores proved uninterpretable as measures of teaching quality, we developed a purpose-built 8-dimension rubric, validated against multiple independent LLM evaluators and a human expert, as a transparent external benchmark. To our knowledge, this is the first study to (1) propose and validate a multidimensional framework for evaluating AI teaching agents in medical education, (2) examine how role-play design, content domain, and learner gender relate to teaching quality within a unified curriculum, and (3) compare platform-generated scores against a transparent external standard to clarify what each actually captures. Beyond the specific agents studied, this work offers a reusable approach to a problem the field will increasingly face: how to see what an AI teaching agent is actually teaching.

Methods

Study Design and Participants

This cross-sectional observational study was conducted during the autumn semester of 2025 at Guangzhou Medical University with 22 fourth-year undergraduate students majoring in basic medical sciences, who participated as part of their endocrinology curriculum. The cohort comprised 15 female and 7 male students. Each student interacted with all 8 AI teaching agents sequentially over the course of the semester. Owing to individual absences, the number of completed dialogues per agent ranged from 19 to 22, yielding 167 valid dialogues in total.

Ethical Considerations

All participants were informed of the study purpose, and the study was approved by the institutional review board of Guangzhou Medical University (202607001).

AI Teaching Agent Design and Deployment

Eight AI teaching agents were constructed and deployed on the Chaoxing e-Learning platform (Beijing Century Chaoxing Information Technology Development Co, Ltd), a widely used commercial platform in Chinese higher education. Each agent covered 1 chapter of the endocrinology curriculum. Two chapters addressed foundational content (a general introduction and organ morphology), and 6 addressed clinical or disease-oriented content (hypothalamic-pituitary diseases, thyroid diseases, adrenal diseases, glucose metabolism disorders, lipid metabolism disorders, and calcium-phosphorus metabolism disorders).

Each agent employed a distinct role-play scenario assigned to 1 of 4 role paradigms: patient (A1, A4, A5), student (A2, A3), expert (A6, A7), and family member (A8); full configurations are provided in Table 1. The 4 paradigms were chosen to represent 4 prototypical interaction scenarios in clinical medicine (doctor and patient, teacher and student, expert consultation, and family communication), thereby sampling a range of communicative and pedagogical demands rather than instantiating a single learning theory a priori. Where relevant, the resulting behaviors could be related post hoc to established educational concepts; for example, the student paradigm elicits scaffolding within the learner's zone of proximal development [20,21]. We emphasize that because each agent combined 1 role paradigm with 1 content chapter, role and content were confounded by design.

Key implementation differences among the agents lay in the assigned AI persona and the reciprocal role assigned to the student, which together determined the communicative register and the pedagogical stance of each dialogue (Table 1). For example, within the expert paradigm, agent A6 positioned the student as a learner in a mentorship model, whereas agent A7 positioned the student as a patient, providing both a physician and a patient perspective. Agent A8 was deliberately revised after the initial deployment of the first 7 agents, when their empathic engagement was judged to be limited; only its prompt was modified, while all other settings were held constant, to strengthen an empathic family-member persona. For each agent, the research team specified the instructional content, learning objectives, knowledge sources, role-play scenario, and safety and behavioral constraints; the complete per-agent configuration, including prompts, is provided in Multimedia Appendix 1.

XSL•FO
RenderX

Table . Overview of AI teaching agent configurations.

Agent	Content topic	Role paradigm	AI role	Student role
A1	General introduction to endocrinology	Patient	Patient (endocrinology inpatient)	Physician (endocrinology intern)
A2	Organ morphology	Student	Learner (first-year medical student)	Instructor (anatomy and histology teacher)
A3	Hypothalamic-pituitary diseases	Student	Learner (average-performing student)	Instructor (top-performing class representative)
A4	Thyroid diseases	Patient	Patient (inquisitive, detail-seeking patient)	Physician (patient's physician)
A5	Adrenal diseases	Patient	Patient (Cushing syndrome patient)	Physician (endocrinologist)
A6	Glucose metabolism disorders	Expert	Expert (senior diabetes clinician-researcher)	Learner (basic medical science student)
A7	Lipid metabolism disorders	Expert	Expert (attending physician)	Patient (patient with obesity)
A8	Calcium-phosphorus metabolism disorders	Family	Family member (older patient with osteoporosis)	Caregiver (grandchild)

Data Collection

Two types of data were collected from each student-agent dialogue. First, the platform's built-in scoring system generated a single aggregate score per dialogue as an automated assessment of student performance. The platform does not disclose how these scores are generated, and its scoring mechanism could not be independently obtained or verified; only aggregate scores were available to instructors, with no dimensional breakdown. Second, complete dialogue transcripts were exported from the platform for independent external evaluation.

Eight-Dimension Pedagogical Evaluation Rubric

To enable standardized cross-agent comparison of teaching quality, we developed an 8-dimension evaluation rubric (Table 2) informed by existing frameworks for AI conversational agents in health care [14], AI agent evaluation in clinical medicine [16], and educational agent effectiveness research [17], supplemented by expert discussion within the research team. Unlike platform-generated scoring, the rubric was designed to assess agent teaching quality rather than student response quality. The complete scoring prompt is provided in Multimedia Appendix 2.

Dimension weights were assigned by the research team on conceptual grounds: medical knowledge accuracy and pedagogical guidance, as the core teaching functions, received the highest weight (20 points each), whereas student engagement elicitation and feedback quality, as supportive dimensions, received the lowest (5 points each). We note as a limitation that these weights were set without formal student or educator input, and that empirically derived weights are a direction for future work.

Table . Eight-dimension pedagogical evaluation rubric for AI teaching agents.

Dimension	Points (total=100)	Scoring focus
Medical knowledge accuracy	20	Factual correctness and absence of medical errors
Pedagogical guidance ability	20	Use of questioning, scaffolding, and structured guidance
Knowledge-point coverage efficiency	15	Breadth and completeness of curriculum content addressed
Role-play quality	15	Consistency, naturalness, and appropriateness of role portrayal
Adaptive difficulty calibration	10	Adjustment of complexity in response to student level
Medical safety boundary	10	Avoidance of harmful advice and appropriate referral behavior
Student engagement elicitation	5	Prompting of active student participation and critical thinking
Assessment and feedback quality	5	Specificity and constructiveness of formative feedback

XSL•FO
RenderX

Rubric-Based Scoring and Validation

All 167 dialogue transcripts were scored across the 8 dimensions using LLMs operating under standardized instructions [15]. To assess the robustness and reliability of the scoring, 3 LLMs from different developers were used as evaluators: Claude Opus 4.8 (Anthropic), representing a leading international model, and DeepSeek (DeepSeek V4 Artificial Intelligence Basic Technology Research Co, Ltd) and Qwen 2.5 (Alibaba Cloud), two widely used Chinese LLMs selected for their strong Chinese-language capability given that all dialogues used Chinese. DeepSeek and Qwen were added at the recommendation of a reviewer and on the basis of our recent study of mainstream domestic and international LLMs for medical assessment tasks [22].

All scoring was performed programmatically through each model's application programming interface under a fixed request configuration, with an intercall interval of 1.0 second and up to 3 retries per call. To promote deterministic, reproducible outputs, DeepSeek and Qwen were queried with temperatures of 0 and top-p of 0.01; Claude Opus 4.8 was queried at its default setting. Each model independently applied the identical 8-dimension rubric and scored every dialogue 4 times, and the mean of the 4 runs was used as that model's final score. This repeated-scoring design followed directly from our recent finding that the reliability of LLM-based scoring is better captured by repeated runs and cross-run variance than by single-run performance [22]. Within-model reproducibility across the 4 runs was quantified by the SD of the repeated scores; reproducibility was high, indicating that scores were stable rather than single, unrepeatable judgments [23].

To calibrate the AI-based scoring against human judgment, a medical-education expert independently rescored a stratified subset of 40 dialogues using the identical 8-dimension rubric. The subset was stratified by Claude total score to span high, medium, and low quality and to cover all 8 agents, and the expert was blinded to the AI scores. Agreement between AI and expert ratings was quantified by the Pearson correlation and the intraclass correlation coefficient (ICC), computed at the total-score level and per dimension [24]. Because rescoring was performed by a single expert, expert-to-expert reliability could not be computed; this is addressed in the Limitations section.

Statistical Analysis

Given nonnormal score distributions and heterogeneous variances across agents (Shapiro-Wilk and Levene tests), nonparametric methods were used throughout. Platform-generated scores were compared across agents using the Kruskal-Wallis H test, with post hoc Dunn tests and Bonferroni correction for pairwise comparisons. Effect sizes were reported as η^2 for omnibus tests and Cohen d for selected contrasts. Between-agent variation in each rubric dimension was summarized using the coefficient of variation (CV).

Comparisons across role paradigms were treated as descriptive, because the independent unit was the agent and each role group comprised only 1 to 3 agents; formal inferential testing was therefore not performed at the agent level. The divergence between platform-based and rubric-based agent rankings was characterized through concrete rank comparisons rather than a single agent-level correlation coefficient, which would be unstable at this sample size. Learner gender differences were examined across dimensions using Mann-Whitney U tests, with results interpreted in light of the small and imbalanced sample. The correspondence between rubric-based agent quality and students' end-of-course chapter examination performance was examined descriptively. All analyses were conducted in Python (version 3.11), with a 2-sided significance threshold of $P<.05$.

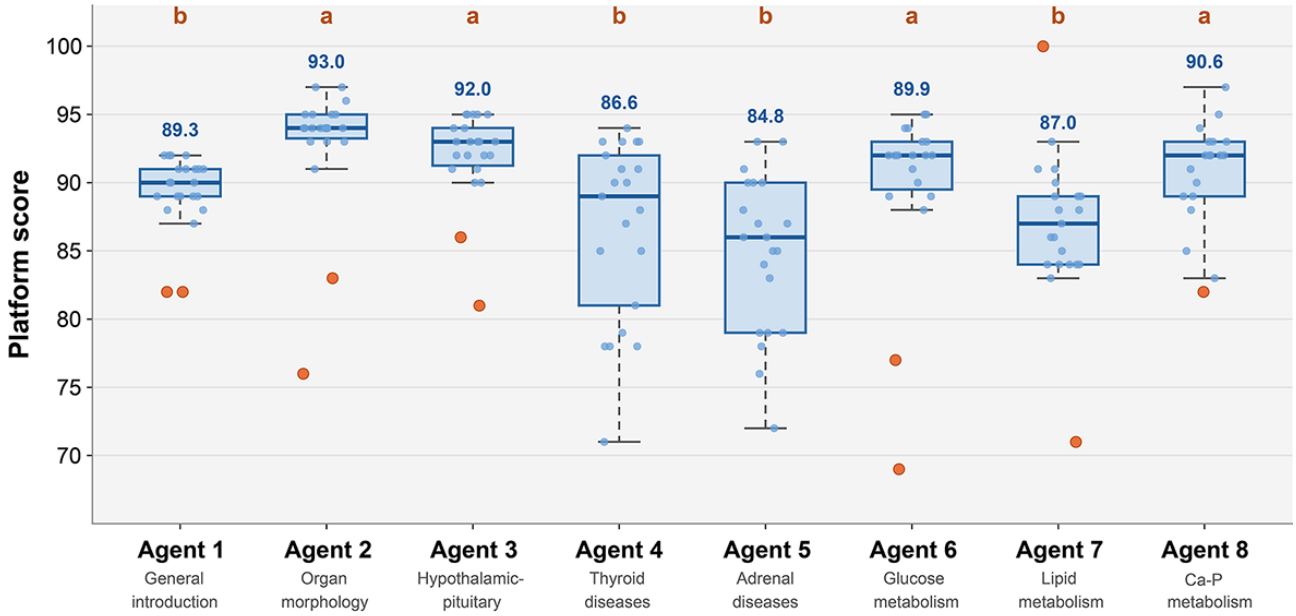
Results

Platform-Generated Scores Varied Across Agents

Chaoxing platform–generated scores from 167 student-agent dialogues differed significantly across the 8 agents (Kruskal-Wallis $H_7=57.981$; $P<.001$; $\eta^2=0.321$; Figure 1), with 9 significant pairwise differences identified by post hoc Dunn tests (Bonferroni-corrected $P<.05$). Basic science agents (A2, A3) received the highest scores, while clinical disease agents (A4, A5, A7) received the lowest. While these differences indicate that agent design influenced student performance under platform scoring, the platform does not disclose how its scores are generated, and its scoring mechanism could not be independently verified, precluding any interpretation of what drove them or whether they reflect genuine agent teaching quality. Therefore, the 8-dimension rubric was applied as an independent external audit, reported in the following sections.

XSL•FO
RenderX

Figure 1. Distribution of Chaoxing platform–generated scores across 8 agents (22 students; 167 valid dialogues). Box plots show median, IQR, and 1.5×IQR whiskers; individual scores are jittered dots; outliers (>1.5×IQR) are in orange. Blue numerals above boxes denote means. Agents not sharing a common letter (a vs b) differ significantly (Bonferroni-corrected P<.05).



LLM Evaluators Differed in Scoring Discrimination

To evaluate the robustness of the rubric-based scoring, the 8-dimension rubric was independently applied by 3 LLM evaluators from different developers (Claude, DeepSeek, and Qwen), each scoring all 167 dialogues 4 times, with high within-model reproducibility (Table 3; mean within-model SD≤0.36).

Table . Comparison of 3 LLM evaluators on the 8-dimension rubric.

Metric	Claude (Opus 4.8)	DeepSeek V4	Qwen 2.5
Dialogues scored (×4 runs), n	167	167	167
Intrarater SD (score stability)[a]	0.36	0.06	0.17
Agent-total score range	61 - 92	82 - 95	75 - 100
Agent-total SD (discrimination)[b]	11.6	4.0	7.3
Overall ceiling rate[c], %	17	30	51
Mean achievement rate, %	82	90	92
Rank correlation with Claude (ρ)[d]	—[e]	0.64	0.71

[a]Intrarater SD: mean within-model SD across the 4 runs (lower=more stable).

[b]Agent-total SD: SD of the 8 agent-level totals (higher=better discrimination).

[c]Overall ceiling rate: proportion of dialogue×dimension cells scored at the maximum (higher=stronger ceiling effect).

[d]Rank correlation: Spearman ρ of agent rankings with Claude (both P<.05).

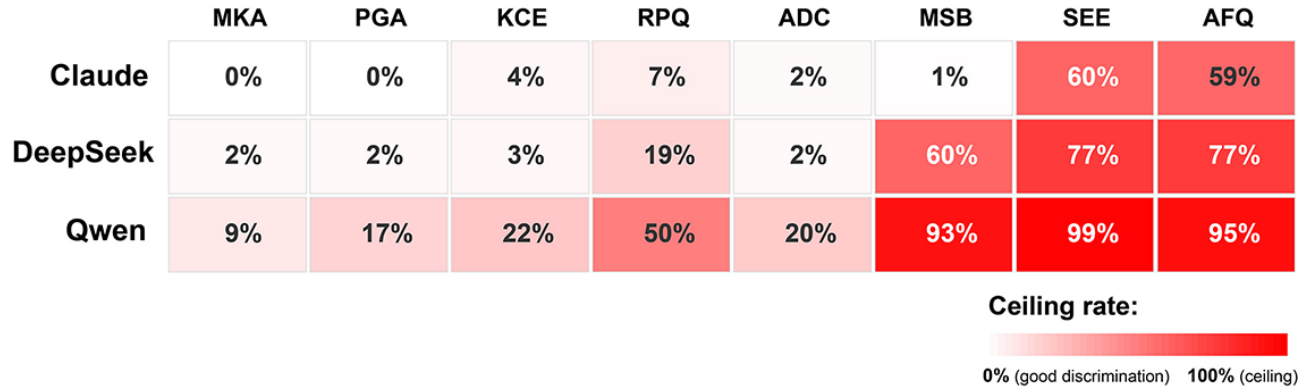
[e]Not applicable.

The 3 models agreed broadly on the relative ranking of agents (Spearman ρ=0.64 and 0.71 for Claude vs DeepSeek and Qwen) but differed substantially in discrimination. DeepSeek and Qwen showed pronounced leniency, scoring large proportions of dialogues at the ceiling (Figure 2; Qwen: above 90% on medical safety boundary, student engagement elicitation, and assessment and feedback quality; DeepSeek: 60-77% on the same dimensions). This compressed between-agent differences, yielding narrow score ranges (82 - 95 and 75 - 100) and low discrimination (SD of agent total scores was 4.0 and 7.3). Claude showed a minimal ceiling effects except on the two 5-point dimensions (student engagement elicitation, assessment and feedback quality), with the widest range (61-92) and greatest discrimination (SD 11.6).

Because the analysis aimed to resolve differences in teaching quality among agents, an evaluator able to discriminate between them was required; Claude was therefore used as the primary evaluator, with its agreement against human expert judgment examined separately. The leniency of the other evaluators is itself a relevant finding, indicating that ceiling effects should be anticipated when LLMs serve as evaluators of educational quality.

XSL•FO
RenderX

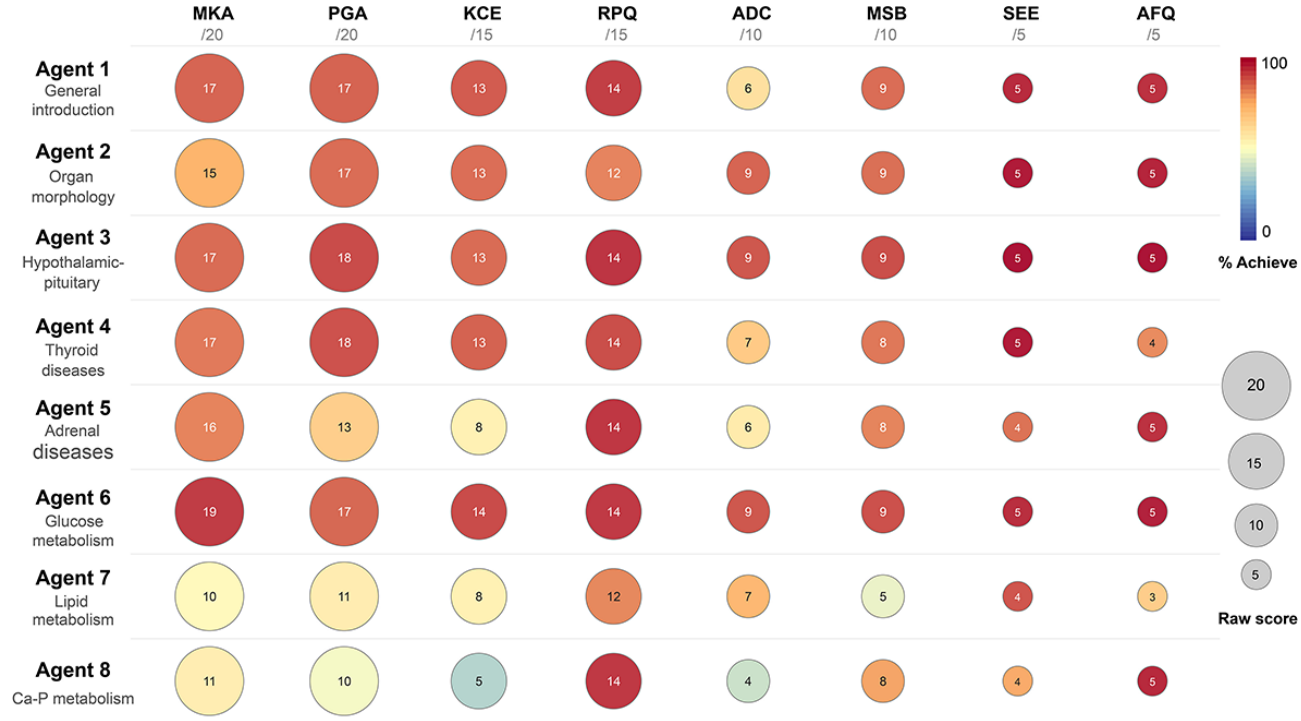
Figure 2. Ceiling effects across the 3 large language model evaluators on the 8 scoring dimensions. Each cell shows the proportion of the 167 dialogues scored at the maximum for that dimension; darker red indicates a stronger ceiling effect (less discrimination). ADC: adaptive difficulty calibration; AFQ: assessment and feedback quality; KCE: knowledge-point coverage efficiency; MKA: medical knowledge accuracy; MSB: medical safety boundary; PGA: pedagogical guidance ability; RPQ: role-play quality; SEE: student engagement elicitation.



Agents Differed Mainly in Knowledge, Not in Role Performance

Applying the 8-dimension rubric, agent total scores ranged widely (60.6 - 91.6; Figure 3), with A6 and A3 scoring highest and A7 and A8 lowest. This variation was not uniform across dimensions. It was largest on the 2 knowledge-related dimensions, namely knowledge-point coverage (CV=27.3%) and medical knowledge accuracy (CV=18.5%), indicating that agents differed most in what and how accurately they taught. In contrast, role-play quality was the most uniform dimension (CV=5.7%), with all agents sustaining their assigned personas comparably well. Adaptive difficulty control was low across nearly all agents, emerging as a shared weakness rather than a source of between-agent variation. Thus, the agents were broadly comparable in role enactment but diverged mainly in the knowledge content they delivered.

Figure 3. Eight-dimension rubric profiles of the 8 AI teaching agents. Scores are the mean of 4 independent runs by Claude Opus 4.8. Bubble size indicates the dimension's maximum score, color indicates the achievement rate; and number indicates the raw score. ADC: adaptive difficulty calibration; AFQ: assessment and feedback quality; KCE: knowledge-point coverage efficiency; MKA: medical knowledge accuracy; MSB: medical safety boundary; PGA: pedagogical guidance ability; RPQ: role-play quality; SEE: student engagement elicitation.



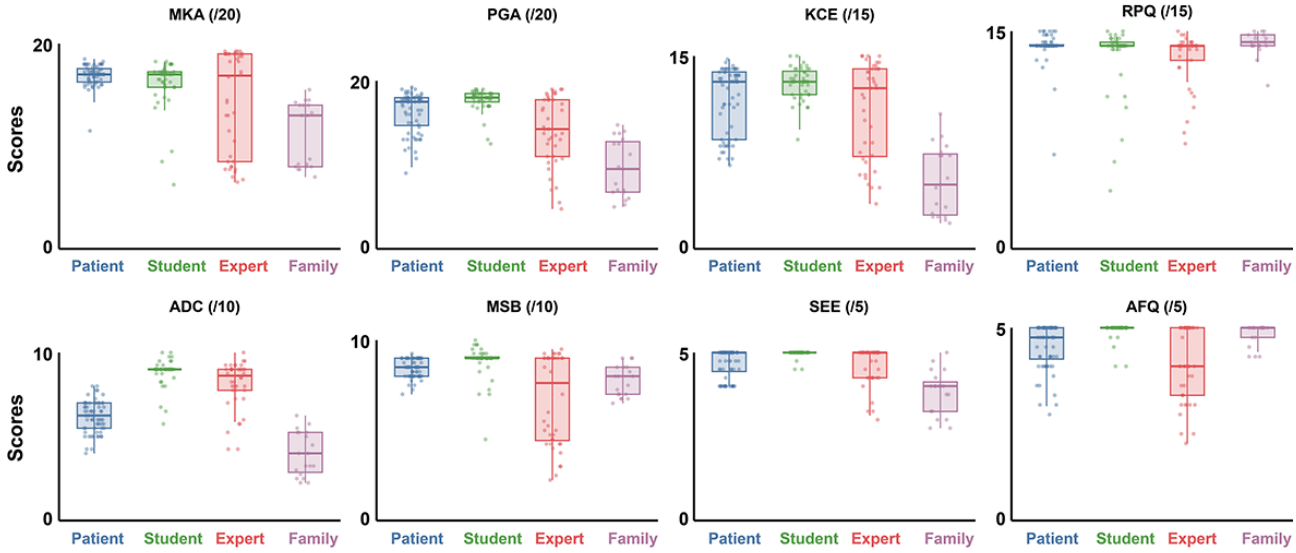
Role Paradigms Shaped Dimension Profiles, but Empathy Did Not Ensure Knowledge Coverage

Grouping agents by role paradigm revealed distinct dimension profiles (Figure 4). At the group level, student-type agents showed the strongest pedagogical guidance (group-mean pedagogical guidance ability was 17.7), whereas expert-type agents showed the lowest medical safety (group-mean medical safety boundary was 6.7). The family-member agent illustrated a notable dissociation between empathy and knowledge: it attained one of the highest role-play scores (role-play quality=14.1) yet the lowest knowledge-point coverage of all agents (knowledge-point coverage efficiency=5.3) and the lowest total score (60.6). Notably, this agent had been

XSL•FO
RenderX

deliberately revised after the first 7 agents showed generally limited empathic engagement in student use; only its prompt was adjusted—while all other settings were held constant—to strengthen the empathic, family-member persona. Despite this targeted enhancement of empathy, it remained weakest in knowledge coverage, indicating that strengthening empathy did not translate into broader knowledge delivery. How to balance affective engagement and knowledge coverage within a single agent thus remains an open question. Because each role group comprised only 2 to 3 agents, these between-group comparisons are descriptive; formal inferential testing was not performed at the agent level.

Figure 4. Eight-dimension score distributions across the 4 role paradigms: patient (A1, A4, A5; n=64 dialogues), student (A2, A3; n=44), expert (A6, A7; n=40), and family member (A8; n=19). Each panel shows 1 dimension (scored out of its maximum, indicated in parentheses). Boxes show median and IQR; points are individual dialogue scores. Between-group differences are descriptive; formal inferential testing was not performed at the agent level. ADC: adaptive difficulty calibration; AFQ: assessment and feedback quality; KCE: knowledge-point coverage efficiency; MKA: medical knowledge accuracy; MSB: medical safety boundary; PGA: pedagogical guidance ability; RPQ: role-play quality; SEE: student engagement elicitation.



Student Gender Showed No Detectable Effect on Teaching Quality Scores

Male (n=7) and female (n=15) students were compared across all 8 dimensions using Mann-Whitney *U* tests. No significant difference was detected in any dimension or in total score (79.1 vs 80.1; *P*=.82; all dimension-level *P*>.05; Cohen *d*=0.07 for total score). Given the small, imbalanced sample and the very small observed effect size, this analysis was substantially underpowered; the absence of a detectable difference should therefore not be interpreted as evidence of gender-equitable delivery, as a type II error cannot be excluded.

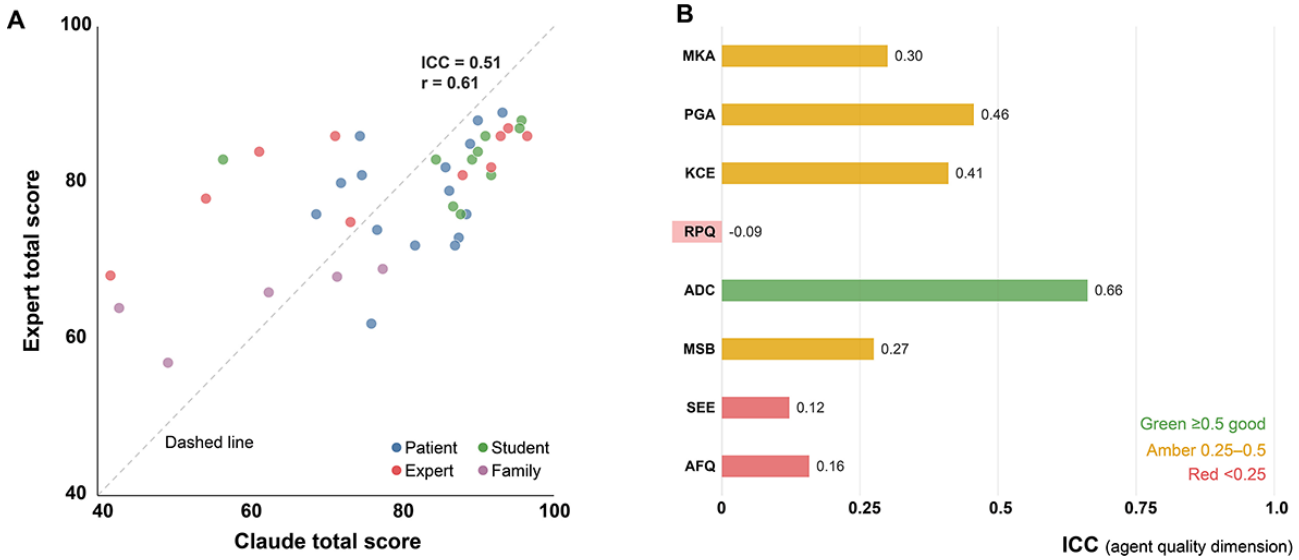
AI Scores Agreed With Expert Ratings on Overall and Cognitive-Process Dimensions

To validate the rubric-based AI scoring, 40 dialogues spanning all 8 agents and the full quality range were independently rescored by a medical-education expert using an identical rubric, blinded to the AI scores (Figure 5A). AI and expert ratings were closely aligned in magnitude (mean totals 79.1 vs 78.5) and were positively correlated at the total-score level (*r*=0.61; ICC=0.51). Because the two derive from a highly consistent automated rater and a single subjective human expert, close but not exact agreement was expected; the aim was to confirm convergence on the same construct, not identity.

Agreement was strongest on the cognitive-process dimensions, ranging from moderate for knowledge accuracy to highest for adaptive difficulty control. For role-play quality and the two 5-point dimensions, the ICC was near zero or negative, but this reflects a ceiling effect rather than rater disagreement: both AI and expert rated these dimensions uniformly high (for role-play quality, 13.4 vs 13.3 out of 15), leaving minimal between-agent variance (Figure 5B). Because the ICC indexes relative discrimination, it is driven toward zero when there is little variance to rank, and its low value here does not indicate that the raters disagreed; on the contrary, they consistently converged on high scores [24]. The moderate agreement on knowledge accuracy likewise does not undermine the study's key knowledge-related finding, which rests on a rank-level difference too large to depend on rating precision: A8 scored markedly lowest on knowledge coverage under both AI and expert scoring.

XSL•FO
RenderX

Figure 5. Agreement between Claude AI and expert ratings on 40 dialogues, independently scored by a medical-education expert using the same 8-dimension rubric. (A) Total score agreement: expert versus Claude total scores (the dashed line indicates perfect agreement). (B) Per-dimension agreement represented by intraclass correlation coefficient (ICC) per dimension (green≥0.5; amber 0.25-0.5; red<0.25). ADC: adaptive difficulty calibration; AFQ: assessment and feedback quality; KCE: knowledge-point coverage efficiency; MKA: medical knowledge accuracy; MSB: medical safety boundary; PGA: pedagogical guidance ability; RPQ: role-play quality; SEE: student engagement elicitation.

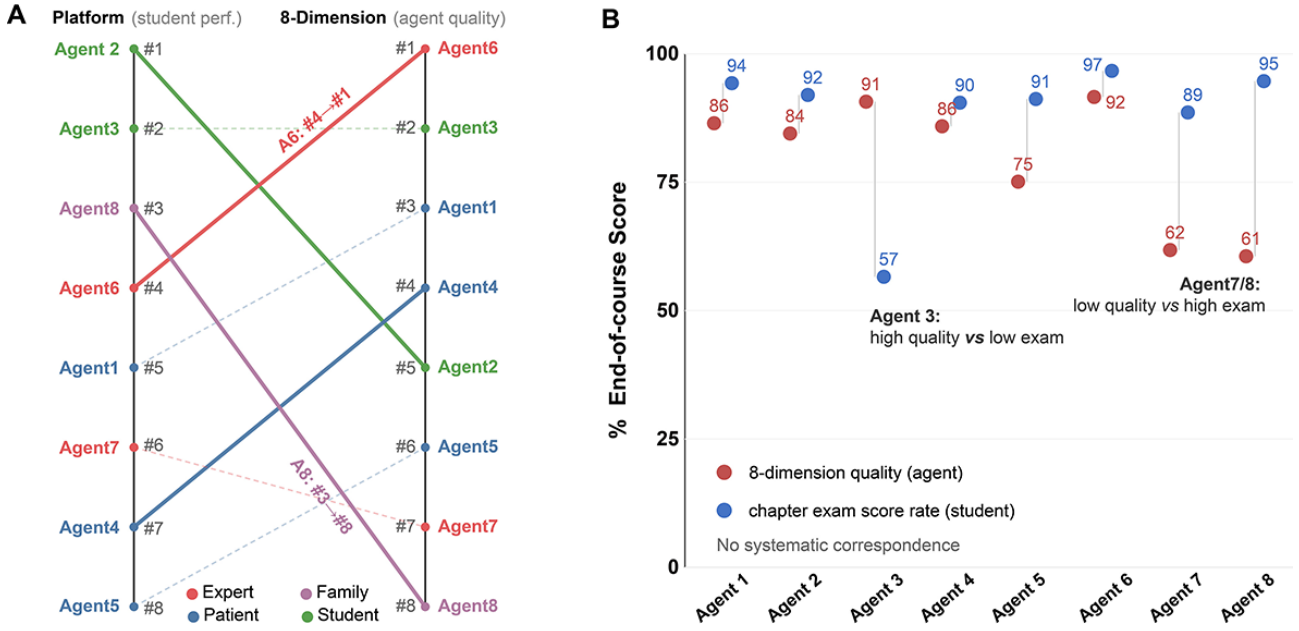


Platform and Examination Performance Captured Constructs Distinct From Rubric-Based Agent Quality

Agent rankings from platform scores and from the 8-dimension rubric diverged substantially (Figure 6A). Two agents illustrate this most clearly: A8 ranked third by platform score but last by rubric quality, and A6 rose from fourth to first. Rather than indicating that either measure is invalid, this divergence reflects that the two capture different constructs—platform scores index student performance during the interaction, whereas the rubric indexes the teaching quality of the agent itself. The two are therefore complementary rather than interchangeable.

There was a similar lack of association between rubric quality and end-of-course chapter examination performance (Figure 6B): higher agent quality did not correspond to higher chapter examination scores (eg, A3 showed high quality but the lowest examination score rate, whereas A7 and A8 showed low overall quality but high examination scores). This absence of a systematic association should be interpreted cautiously, given the small number of agents, the near-ceiling examination scores, and the lack of individual baseline data; it indicates only that agent teaching quality and this particular examination did not track one another, not that better teaching is ineffective.

Figure 6. Platform and examination performance versus rubric-based agent quality. (A) Rank divergence: agent rankings by platform score (student performance) vs by 8-dimension quality (agent quality); 1=best. (B) Eight-dimension agent quality versus end-of-course chapter examination score rate.



XSL•FO
RenderX

Discussion

This study set out to test whether platform-generated scores, the default metric for AI teaching agents in many educational platforms, adequately capture teaching quality, and to ask what a transparent, multidimensional alternative would reveal. The findings speak less to any single agent than to how the field should evaluate, and design, conversational AI tutors.

Evaluation is not a solved problem, and the default metric may mislead. The most consequential implication is that the metric an educator happens to have at hand can point in the wrong direction. When a platform score and a criterion-based rubric rank the same agents in nearly opposite orders, the choice of metric is not a technical detail but a determinant of which agents get adopted, refined, or discarded. Crucially, this does not make either metric wrong. A platform score that reflects student performance during an interaction and a rubric that reflects the agent's teaching behavior are answering different questions; the problem arises only when one is silently substituted for the other. As AI tutors proliferate faster than the frameworks to evaluate them [11,14], this argues for treating evaluation itself as a first-class research object, reporting what a metric measures, and pairing convenient platform analytics with transparent, education-specific criteria rather than trusting either alone [17].

Automated evaluation is promising but must be bounded. Using LLMs to score pedagogical quality is attractive because it scales, but our results caution against naive trust [25,26]. Independently developed models converged on the ordering of agents yet differed markedly in leniency, showing that agreement on ranking can coexist with disagreement on absolute standards, and that some models are too lenient to discriminate at all [13,23]. Human calibration therefore remains indispensable [27], and it is not uniform across constructs: automated and expert judgments aligned well on cognitive-process dimensions but poorly on more subjective, affective ones [27]. This suggests a pragmatic division of labor, in which automated scoring is trusted most for the dimensions where it is demonstrably calibrated, while human judgment is reserved for those where meaning is contested [12].

Designing for the whole of teaching, not its most visible parts, and comparing role paradigms, rather than the field's default patient-doctor script [8,9], exposed a design tension that is easy to miss when engagement is the headline goal. An agent optimized to be more empathic became more engaging without becoming more knowledgeable, a reminder that affective and cognitive teaching functions do not automatically move together [17], and that improving the salient, likable qualities of a tutor can leave its instructional substance untouched. Rather than seeking a single best role, this points toward designing for complementarity, orchestrating different pedagogical strengths across an interaction [16], and toward evaluation that can detect when a more engaging agent is not in fact teaching more. The most uniform finding across every agent, namely persistent weakness in adapting difficulty and providing formative feedback, reinforces the same lesson: the harder and less visible pedagogical skills are precisely those that current prompt-based designs deliver least well [9,28].

Future directions should consider 3 points. First, disentangling role from content requires a factorial design that crosses role paradigms with content domains, so that the effect of how an agent teaches can be separated from what it teaches [5]. Second, establishing educational value, rather than teaching quality alone, requires learning-outcome designs with pre- and postlearning knowledge assessment and individual-level linkage [10,19], ideally using more discriminating assessments than a near-ceiling course examination. Third, the persistent deficits in adaptive difficulty and feedback motivate designs that move beyond static prompting toward dynamic learner modeling and memory-augmented tutoring [8,16], evaluated against the dimensions on which current agents underperform. More broadly, the 8-dimension evaluation framework introduced here offers a transferable reference for the development of future teaching agents, and of medical teaching agents in particular, by making explicit the pedagogical dimensions against which such agents should be designed and assessed.

Undeniably, this is a small, single-institution study in one discipline, and its design confounds role paradigm with content domain, each agent being one role and one chapter, so role-related observations are descriptive rather than causal and no independent content-domain effect can be claimed. Because role groups contain few agents, group comparisons are descriptive and were not tested inferentially at the agent level. Expert calibration rested on a single expert, precluding expert-to-expert reliability, and was weaker for subjective dimensions. The gender comparison was underpowered and cannot support claims of equitable delivery, the rubric weights were team-assigned without formal stakeholder input, and teaching quality rather than learning gain was the outcome assessed. These constraints frame the study as an exploratory, framework-building effort whose specific estimates await confirmation in larger, factorial, outcome-linked designs.

Platform scores answer a different question than teaching quality and should complement rather than replace criterion-based evaluation. By making the dimensions of teaching explicit, and by validating those judgments against expert and cross-model agreement, a multidimensional rubric reveals distinctions that convenient metrics miss, including where more engaging agents are not more instructive and where all current agents falter. Beyond any single result, the study suggests that how we evaluate AI teaching agents is inseparable from how we improve them.

XSL•FO
RenderX

Acknowledgments

The authors declare that generative AI was used in the preparation of this manuscript. Claude Opus 4.8 was employed to enhance the language and readability of the draft. The final version was thoroughly reviewed, revised, and approved by the authors, who take full responsibility for the content.

Funding

This work was supported by the following grants: the 2026 Educational Science Planning and Teaching Reform Projects of Guangzhou Medical University (2026JXGG02) and the 2025 Guangdong Provincial Graduate Education Innovation Program (2025JGXM_139).

Data Availability

The datasets generated and analyzed during the current study are available from the corresponding author on reasonable request.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Complete configuration of the 8 AI teaching agents.
[PDF File, 198 KB - mededu_v12i1e96819_app1.pdf]

Multimedia Appendix 2
The complete 8-dimension scoring prompt used by all evaluators.
[PDF File, 123 KB - mededu_v12i1e96819_app2.pdf]

References

1. Long C, Tsai K, Napolitano JD, Khawaja RA, Leung AM. The current status of preclinical endocrine education in U.S. medical schools. Endocr Pract 2022 Aug;28(8):744-748. [doi: 10.1016/j.eprac.2022.04.008] [Medline: 35452814]
2. Hu MH, Han YT, Lyu J, Li LM. Advantages and challenges of applying artificial intelligence in medical education. Zhonghua Liu Xing Bing Xue Za Zhi 2026 Feb 10;47(2):200-206. [doi: 10.3760/cma.j.cn112338-20250512-00309] [Medline: 41765658]
3. Jiang Y, Fu X, Wang J, et al. Enhancing medical education with chatbots: a randomized controlled trial on standardized patients for colorectal cancer. BMC Med Educ 2024 Dec 20;24(1):1511. [doi: 10.1186/s12909-024-06530-8] [Medline: 39707245]
4. Öncü S, Torun F, Ülkü HH. AI-powered standardised patients: evaluating ChatGPT-4o's impact on clinical case management in intern physicians. BMC Med Educ 2025 Feb 20;25(1):278. [doi: 10.1186/s12909-025-06877-6] [Medline: 39979969]
5. Xiao J, Fu X. Is the use of standardized patients more effective than role-playing in medical education? A meta-analysis. Front Med (Lausanne) 2025 Jun 18;12:1601116. [doi: 10.3389/fmed.2025.1601116] [Medline: 40606465]
6. Huang LJ, Huang HC, Chuang CL, et al. Role-play of real patients improves the clinical performance of medical students. J Chin Med Assoc 2021 Feb 1;84(2):183-190. [doi: 10.1097/JCMA.0000000000000431] [Medline: 32925298]
7. Flanagan OL, Cummings KM. Standardized patients in medical education: a review of the literature. Cureus 2023 Jul 17;15(7):e42027. [doi: 10.7759/cureus.42027] [Medline: 37593270]
8. Yu H, Zhou J, Li L, et al. Simulated patient systems powered by large language model-based AI agents offer potential for transforming medical education. Commun Med (Lond) 2025 Dec 19;6(1):27. [doi: 10.1038/s43856-025-01283-x] [Medline: 41420084]
9. Brügge E, Ricchizzi S, Arenbeck M, et al. Large language models improve clinical decision making of medical students through patient simulation and structured feedback: a randomized controlled trial. BMC Med Educ 2024 Nov 28;24(1):1391. [doi: 10.1186/s12909-024-06399-7] [Medline: 39609823]
10. Peng J, Zhang H, Tu X, et al. Effectiveness of AI-assisted medical education for Chinese undergraduate medical students: a meta-analysis. BMC Med Educ 2025 Aug 27;25(1):1207. [doi: 10.1186/s12909-025-07770-y] [Medline: 40866973]
11. Aster A, Laupichler MC, Rockwell-Kollmann T, Masala G, Bala E, Raupach T. ChatGPT and other large language models in medical education - scoping literature review. Med Sci Educ 2024 Nov 13;35(1):555-567. [doi: 10.1007/s40670-024-02206-6] [Medline: 40144083]
12. Vrdoljak J, Boban Z, Vilović M, Kumrić M, Božić J. A review of large language models in medical education, clinical decision support, and healthcare administration. Healthcare (Basel) 2025 Mar 10;13(6):603. [doi: 10.3390/healthcare13060603] [Medline: 40150453]

XSL•FO
RenderX

13. Elhilali A, Ngo ASH, Reichenpfader D, Denecke K. Large language model-based patient simulation to foster communication skills in health care professionals: user-centered development and usability study. JMIR Med Educ 2025 Dec 12;11:e81271. [doi: 10.2196/81271] [Medline: 41385781]
14. Ding H, Simmich J, Vaezipour A, Andrews N, Russell T. Evaluation framework for conversational agents with artificial intelligence in health interventions: a systematic scoping review. J Am Med Inform Assoc 2024 Feb 16;31(3):746-761. [doi: 10.1093/jamia/ocad222] [Medline: 38070173]
15. Gu J, Jiang X, Shi Z, et al. A survey on LLM-as-a-judge. Innovation (Camb) 2026 Jun 1;7(6):101253. [doi: 10.1016/j.xinn.2025.101253] [Medline: 42254963]
16. Gorenshtein A, Omar M, Glicksberg BS, Nadkarni GN, Klang E. AI agents in clinical medicine: a systematic review. medRxiv. 2025 Aug 26 p. 2025.08.22.25334232. [doi: 10.1101/2025.08.22.25334232] [Medline: 40909853]
17. Xu X, Cao X, Wu Q. Impact of educational agents on student's learning outcomes: a meta-analysis. Front Psychol 2026 Feb 24;17:1707196. [doi: 10.3389/fpsyg.2026.1707196] [Medline: 41815245]
18. Alshammri F, Abdulshakour M, Chen L, et al. Pediatric endocrinology education among trainees: a scoping review. Clin Teach 2025 Feb;22(1):e70011. [doi: 10.1111/tct.70011] [Medline: 39743233]
19. Myers SE, Bender NR, Seidel MA, Weinstock RS. Diabetes SPECIAL (Students Providing Education on Chronic Illness and Lifestyle): a novel preclinical medical student elective. Perspect Med Educ 2021 Oct;10(5):312-315. [doi: 10.1007/s40037-020-00641-w] [Medline: 33349906]
20. Daemicke AK, Galt NJ, Samonds KE, Bergan-Roller HE. Challenging endocrinology students with a critical-thinking workbook. Adv Physiol Educ 2020 Mar 1;44(1):72-79. [doi: 10.1152/advan.00101.2019] [Medline: 32057267]
21. Hodge TJ, Potter LJ, Helsby CJ. Scaffolding simulation activities for medical students learning cardiopulmonary assessment: a retrospective study. Cureus 2025 Apr 10;17(4):e82013. [doi: 10.7759/cureus.82013] [Medline: 40351920]
22. Zhang H, Qu L, Bai H, et al. Beyond accuracy: evaluating the reliability of large language models for medical assessment. Front Artif Intell 2026 Jul 8;9:1832829. [doi: 10.3389/frai.2026.1832829] [Medline: 42488377]
23. Thomas LDW, Romasanta AKG, Pujol Priego L. Jagged competencies: measuring the reliability of generative AI in academic research. J Bus Res 2026 Jan;203:115804. [doi: 10.1016/j.jbusres.2025.115804]
24. Mehta S, Bastero-Caballero RF, Sun Y, et al. Performance of intraclass correlation coefficient (ICC) as a reliability index under various distributions in scale reliability studies. Stat Med 2018 Aug 15;37(18):2734-2752. [doi: 10.1002/sim.7679] [Medline: 29707825]
25. Croxford E, Gao Y, First E, et al. Evaluating clinical AI summaries with large language models as judges. NPJ Digit Med 2025 Nov 5;8(1):640. [doi: 10.1038/s41746-025-02005-2] [Medline: 41193667]
26. Zhai G, Bar M, Cowan AJ, et al. AI for evidence-based treatment recommendation in oncology: a blinded evaluation of large language models and agentic workflows. Front Artif Intell 2025 Dec 9;8:1683322. [doi: 10.3389/frai.2025.1683322] [Medline: 41446897]
27. Teckwani SH, Wong AHP, Luke NV, Low ICC. Accuracy and reliability of large language models in assessing learning outcomes achievement across cognitive domains. Adv Physiol Educ 2024 Dec 1;48(4):904-914. [doi: 10.1152/advan.00137.2024] [Medline: 39514712]
28. Bogenschutz K, Demeter J, Knoderer CA. Artificial intelligence in medical education: promoting active learning with a customized chatbot tool. J Physician Assist Educ 2026 Jun 1;37(2):216-220. [doi: 10.1097/JPA.0000000000000742] [Medline: 41603621]

Abbreviations

CV: coefficient of variation
ICC: intraclass correlation coefficient
LLM: large language model

Edited by W Jerjes; submitted 01.Apr.2026; peer-reviewed by S Liu, Y Yang; revised version received 10.Jul.2026; accepted 16.Jul.2026; published 03.Sep.2026.

Please cite as:
Zhang H, Qu L, Zheng J, Xiong Y, Bai H, Ji R, Liu G, Chen W, Cheng Z, Chen Y, Yang CT
What Platform Scores Miss: Multidimensional Evaluation of AI Teaching Agents in Medical Education
JMIR Med Educ 2026;12:e96819
URL: https://mededu.jmir.org/2026/1/e96819
doi:10.2196/96819

XSL•FO
RenderX

© Hui Zhang, Lihui Qu, Jianmin Zheng, Yi Xiong, Hongbo Bai, Ruiying Ji, Guohui Liu, Wanling Chen, Zirui Cheng, Youbang Chen, Chun-tao Yang. Originally published in JMIR Medical Education (https://mededu.jmir.org), 3.Sep.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Education, is properly cited. The complete bibliographic information, a link to the original publication on https://mededu.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Implementation and Evaluation of a Cancer Immunotherapy Elective for Medical Students: Mixed Methods Descriptive Study

Mark Raynor[1], MD; Rivers Hock[1], MD; Brandon Godinich[1], MBA, MHA; Satish Maharaj[2,3], MD; Houriya Ayoubieh[1], MD; Cynthia Perry[1], PhD; Jessica Chacon[1], PhD

[1]Department of Medical Education, Foster School of Medicine, Texas Tech University Health Sciences Center El Paso, 130 Rick Francis, El Paso, TX, United States

[2]Department of Hematology/Oncology, University of South Florida, Tampa, FL, United States

[3]Moffitt Cancer Center, Tampa, FL, United States

Corresponding Author:
Jessica Chacon, PhD
Department of Medical Education, Foster School of Medicine, Texas Tech University Health Sciences Center El Paso, 130 Rick Francis, El Paso, TX, United States

Abstract

Background: Cancer immunotherapy represents a transformative advancement in oncology, offering new avenues for treating malignancies by harnessing the immune system. Despite its growing clinical relevance, immunotherapy remains underrepresented in undergraduate medical education, particularly in curricula integrating foundational immunology with clinical application. To address this gap, we developed and implemented a fully online elective for fourth-year medical students focused on core immunology concepts, immunotherapy mechanisms, Food and Drug Administration–approved treatments, immune-related adverse events, and patient-centered therapeutic decision-making.

Objective: This study aimed to evaluate the effectiveness of an asynchronous-synchronous online cancer immunotherapy elective in improving medical student knowledge, engagement, and critical-thinking skills. We hypothesized that participation in the elective would be associated with perceived improvements in knowledge and clinical preparedness and inform future strategies for integrating cancer immunotherapy into medical curricula.

Methods: We conducted a mixed methods study with fourth-year medical students enrolled in a 2-week elective at a US medical school. The curriculum included a self-paced foundational module, online discussion board, and a capstone oral presentation requiring students to propose a novel immunotherapy approach. Participants completed pre- and postcourse quizzes assessing immunotherapy knowledge and an anonymous postcourse Likert-scale survey. Quantitative data were summarized descriptively, and Likert responses were reported using medians and IQRs. Due to the small sample size, 2-tailed unpaired *t* tests comparing pre- and postcourse quiz averages were underpowered to detect statistically significant differences. Qualitative data were analyzed using inductive thematic analysis with investigator triangulation.

Results: A total of 35 students completed the elective, and 20 submitted the postcourse survey (response rate: 57%). Across all Likert-scale items, students reported a median response of 5 (Strongly Agree), with IQR values ranging from 0 to 1, indicating uniformly positive perceptions and minimal variability in their evaluation of the course. Descriptively, average postcourse quiz scores were higher than precourse scores, suggesting improved conceptual understanding. Qualitative thematic analysis revealed three major themes: (1) increased confidence engaging with complex immunotherapy mechanisms, (2) appreciation for the flexibility and interactivity afforded by the hybrid asynchronous-synchronous model, and (3) enhanced understanding of the real-world clinical application of immunotherapy across interdisciplinary settings.

Conclusions: Descriptive quantitative and qualitative findings suggest that a targeted online cancer immunotherapy elective may enhance learners' perceived knowledge and critical-thinking capacity related to emerging cancer therapies. The course's hybrid structure offered flexibility, accessibility, and potential scalability. As immunotherapy continues to expand in clinical practice, this model provides a promising framework for integration into medical curricula. Future work should include larger cohorts and longitudinal follow-up into residency to more rigorously assess educational impact.

(JMIR Med Educ 2026;12:e71628) doi:10.2196/71628

KEYWORDS

medical education; immunotherapy; cancer; undergraduate medical education; cancer immunotherapy

XSL•FO
RenderX

Introduction

Immunotherapy, heralded as a beacon of hope in the oncological landscape, stands at the forefront of medical innovation, offering novel avenues for combating cancer [1-3]. As immune checkpoint blockade, cellular therapies, and cancer vaccinology assume increasing prominence, proficiency in cancer immunotherapy is becoming an essential competency for future health care practitioners [1,2,4,5]. Yet, despite its transformative clinical impact, formal education on cancer immunotherapy remains fragmented or absent in most undergraduate medical curricula. Prior studies have highlighted that health care trainees often feel underprepared to discuss or apply immunotherapy principles in clinical settings, reflecting a broader disconnect between rapid scientific advances and medical education [6,7].

Existing educational work in this area has largely centered on postgraduate oncology programs, continuing medical education, or highly specialized workshops. There remains a critical gap in evidence-based, integrative, and educational models that can effectively prepare medical students, before residency, for the clinical realities of modern cancer care. This lack of structured exposure creates downstream challenges: learners enter rotations or residency without foundational immunotherapy literacy, clinicians report variability in trainee preparedness, and the rapidly evolving therapeutic landscape continues to outpace traditional curricula.

To address this gap, we developed the *Basics and Implementation of Cancer Immunotherapy* elective at the Texas Tech University Health Sciences Center (TTUHSC) El Paso Paul L. Foster School of Medicine. This fully online elective was intentionally designed to provide accessible foundational knowledge while fostering higher-order thinking, clinical reasoning, and innovation in the context of rapidly evolving immunotherapy treatments.

The goal of the course was to deliver content and create a sustainable, adaptable, and evaluable model for broader implementation. Therefore, we grounded the curriculum in 2 complementary frameworks: RE-AIM (Reach, Effectiveness, Adoption, Implementation, and Maintenance) and Kirkpatrick's model. RE-AIM is an implementation science framework typically used in public health and program dissemination research [8,9]. It guided the structural design of the elective by prompting us to consider who the elective could reach, how well it might function across learner groups, what resources and faculty expertise were required for adoption, how consistently it could be delivered in an online environment, and how it could be maintained or scaled over time.

In contrast, Kirkpatrick's model, which focuses on Reaction, Learning, Behavior, and Results, provided a pedagogical framework for evaluating educational impact at the learner level. This dual framework approach allowed us to simultaneously examine (1) learner satisfaction, knowledge, and perceived skill development, and (2) broader programmatic considerations such as feasibility, reproducibility, and sustainability within an already crowded medical curriculum [10-14]. By integrating these 2 models, we aimed to ensure both a high-quality educational experience and a design that could be reliably implemented or adapted by other institutions.

This study aimed to evaluate the effectiveness of the elective in enhancing student knowledge and critical thinking using a mixed methods approach. Specifically, we sought to answer the research question: Does participation in this elective improve students' knowledge, perceptions of course relevance, and critical thinking in cancer immunotherapy? We hypothesized that students completing the elective would demonstrate measurable knowledge gains, report a positive educational experience, and show enhanced critical thinking related to immunotherapy. Findings from this study can guide future curricular efforts aimed at integrating contemporary immunotherapy concepts into undergraduate medical education in a structured and measurable way.

Methods

Rationale for Elective Development

Medical students at Foster School of Medicine receive limited exposure to cancer immunotherapy during the preclerkship phase because the integrated curriculum and density of required content constrain opportunities for in-depth coverage [15-17]. To address this gap, faculty designed a fourth-year elective grounded in a spiral learning model that builds on students' foundational immunology knowledge from the first 2 preclerkship years and reinforces concepts introduced during their clinical rotations.

Curricular Design

We selected a virtual elective to match student preferences for greater fourth-year flexibility and to leverage the institution's expanded capacity for high-quality online learning primarily developed during the COVID-19 era [18-21]. The elective directly addressed institutional needs identified through annual curriculum review and supported the School of Medicine's priority to prepare graduates for rapidly evolving biomedical innovations, such as immunotherapy. Within the Department of Medical Education, the elective advanced key curricular priorities: (1) integrating foundational basic sciences, such as immunology, with clinical reasoning; (2) broadening elective options that promote scholarly inquiry; and (3) expanding technology-enhanced and competency-based instruction.

Faculty with expertise in cancer immunotherapy, biochemistry and cell biology, clinical genetics, oncology, educational design, and online pedagogy led curriculum development. The team followed a staged design process: we mapped core immunotherapy competencies, identified gaps across the preclinical and clerkship curriculum, and iterated content through consultation with our clinical oncology faculty. Medical students actively contributed to early course development. A small group of students co-developed the immunology primer by drawing on their research experience and peer and faculty feedback to ensure clarity, relevance, and appropriate depth. These students also conducted an informal needs assessment through conversations with classmates, confirming that many trainees felt underprepared to discuss contemporary cancer immunotherapies during clinical rotations.

XSL•FO
RenderX

The faculty built the final course using interactive synchronous and asynchronous components, including the student-developed primer, structured discussion boards, and a culminating student presentation. These elements aimed to deepen engagement with the scientific literature and support inquiry–driven learning [18-20]. Constructivist learning theory, interdisciplinary integration [22], and reflective practice guided the instructional approach. Expert consensus, national immunotherapy competency frameworks, and student-identified needs informed topic selection. Basic science and clinical faculty peer-reviewed all content before launch to ensure scientific rigor, clinical relevance, and instructional quality.

Pedagogical Framework, Principles, and Competencies and Standards

Developing a fourth-year elective for medical students on cancer immunotherapy required deliberate alignment with pedagogical frameworks, instructional principles, established competencies, and educational standards to meet the learning objectives. Basic science faculty (immunology, biochemistry, and cell biology) and clinical faculty (internal medicine, oncology, and clinical genetics) jointly oversaw the course.

Constructivist and Experiential Foundations

The design of the elective drew on constructivist and experiential learning theories to promote knowledge construction through active engagement, inquiry, and reflective practice. We incorporated Kirkpatrick's Four-Level Model and the RE-AIM framework to guide course design and evaluation. These frameworks ensured that learning activities targeted measurable outcomes, real-world application, and iterative improvement based on faculty and student feedback. The course combined asynchronous modules, faculty-led synchronous discussions, and a final oral presentation to support both individual learning and collaborative inquiry.

Kirkpatrick's Four-Level Model

We applied the Kirkpatrick's Model [10] to evaluate student learning and course impact.

- Level 1. Reaction: We collected postcourse survey feedback to assess engagement and satisfaction with course structure and content.
- Level 2. Learning: We measured knowledge acquisition using asynchronous pre- and postquizzes focused on immunology basics and the interplay between the immune system and the tumor microenvironment.
- Level 3. Behavior: We evaluated critical thinking by assigning discussion prompts requiring students to propose novel therapies and apply theoretical concepts in innovative ways.
- Level 4. Results: We examined how effectively students identified and proposed solutions to gaps in current immunotherapy practices during their oral presentations.

RE-AIM Framework

In parallel, we used the RE-AIM framework to evaluate implementation and scalability.

- *Reach*: We monitored enrollment numbers and demographic diversity.
- *Effectiveness*: We analyzed improvements in quiz scores and qualitative feedback on learning outcomes.
- *Adoption*: We noted the course's integration into the fourth-year elective catalog and student uptake across multiple intended specialties.
- *Implementation*: We assessed delivery fidelity, faculty coordination, and alignment of synchronous and asynchronous components.
- *Maintenance*: We planned annual updates to incorporate emerging immunotherapy advances and maintain curricular relevance.

This dual framework approach enabled a comprehensive evaluation of learner experience and overall educational impact.

Pedagogical Framework

We provided a pedagogical framework for evaluating educational impact at the learner level:

- *Constructivist approach* [23]: Active learning was emphasized where students constructed their understanding of cancer immunotherapy through exploration, inquiry, and problem-solving.
- *Interdisciplinary integration* [24]: Principles from basic sciences, such as immunology, cancer biology, pharmacology, and molecular biology, were integrated to provide a holistic understanding of cancer immunotherapy.
- *Reflective practice* [25,26]: Students were encouraged to reflect on their learning, clinical experiences, and possible immune adverse events. Students were additionally challenged to consider their patient population, particularly those underrepresented in medicine when considering immunotherapy for cancer treatment.

Principles

Principles were integrated as follows:

- *Evidence-based* [27]: All content and teaching methods were grounded in current scientific evidence and clinical practice guidelines. A consultation with experts in the cancer immunotherapy field at the basic science and clinical settings was conducted.
- *Patient-centered* [28,29]: The importance of patient perspectives and personalized treatment approaches in cancer care was emphasized.
- *Ethical awareness* [30]: Ethical dilemmas, such as access to treatment, clinical trial participation, and the implications and importance of genetic testing in cancer immunotherapy, were discussed.

Competencies and Standards [31-34]

We developed the competencies and standards to meet the learning objectives:

- *Knowledge base*: Ensure that students understand the fundamental principles of cancer immunology, including the role of immune checkpoints, innate and adaptive immunity, cytokines, and monoclonal antibodies in cancer treatment.
- *Clinical skills*: Ensure that students develop a foundational understanding of how clinicians evaluate candidates for

XSL•FO
RenderX

immunotherapy, monitor treatment responses, and identify common immune-related adverse events.
- *Critical thinking*: Oversee student ability to analyze current immunotherapy options and formulate evidence-based innovative treatment.
- *Communication skills*: Provide feedback on student ability to effectively communicate with faculty and peers regarding basic science concepts, cancer immunotherapy treatment, and potential side effects.
- *Professionalism*: Provide student feedback on demonstrating empathy, cultural competence, and ethical behavior in all aspects of immunotherapy care and professional interactions.

Elective Structure

We developed the elective structure for enhancing student knowledge and critical thinking:

- *Pre-elective quiz*: We developed the pre-elective quiz to measure students' baseline understanding of basic immunology and foundational cancer immunotherapy concepts. Faculty selected quiz topics by mapping them to the elective's core learning objectives and to national immunotherapy competency frameworks [35,36]. The prequiz intentionally emphasized basic foundational knowledge, including innate and adaptive immunity, immune cell function, antigen recognition, tumor and immune system interactions, and introductory immunotherapy mechanisms, to ensure that all learners entered the elective with a minimum level of readiness. To determine the length of the quiz, faculty reviewed comparable formative and summative assessments used in preclerkship immunology courses at TTUHSC El Paso and agreed on a concise 15-item structure that would prevent test fatigue while allowing coverage of core concepts. Before administering the quiz, basic science faculty and clinical faculty members conducted a peer review to evaluate question clarity, clinical accuracy, and alignment with learning outcomes. The team performed an informal bias and sensitivity review to ensure that items did not disadvantage students based on specialty interest or prior research experience. Faculty also performed a preliminary item difficulty appraisal using standard immunology benchmarks to ensure that the quiz appropriately captured foundational, not advanced, knowledge. Students completed the pre-elective quiz electronically on the first day of the course. Faculty used group-level response patterns to identify shared conceptual gaps and to tailor emphasis during synchronous discussions.

- *Immunology primer review*: Students reviewed a self-paced immunology primer that reinforced foundational concepts needed to analyze immunotherapy mechanisms and clinical applications. The primer served as the common baseline for all subsequent discussions.

- *Introduction to cancer immunotherapy assigned readings*: Assigned readings introduced the historical evolution of immunotherapy, early discoveries in tumor immunology, and the scientific rationale behind modern immune-based cancer treatments. Students were expected to extract key mechanistic principles and discuss them during faculty-facilitated sessions.

- *Interactive discussion board (collaborative inquiry)*: Over 2 weeks, students participated in a moderated online discussion board (Canvas LMS). Faculty posted weekly prompts connecting mechanistic immunology to clinical controversies (eg, "Should combination immunotherapy be used upfront in all patients with high tumor burden?"). Students responded to peers, synthesized evidence from the literature, and reflected on the ethical and clinical implications of their arguments. Faculty monitored discussions, asked probing questions, and clarified misconceptions in real time.

Faculty facilitators led structured online discussions using a consistent set of prompts, such as the following:

- What immunologic mechanism explains this patient's response or nonresponse?
- What biomarkers would guide treatment selection?
- How would you monitor for immune-related adverse events, and which patient population is at most risk?
- What therapeutic alternatives exist, and what is the evidence supporting them?

This structure ensured clinically rigorous, evidence-driven dialogue while encouraging students to integrate mechanistic immunology with real-world oncology decision-making.

- *Ethical considerations, patient perspectives, and social determinants of health*: To ensure robust and meaningful discussion of ethics and equity, the faculty developed structured ethical prompts and probes related to access, affordability, and geographic disparities in immunotherapy delivery. Examples included the following: (1) How do insurance status and socioeconomic factors influence a patient's ability to access checkpoint inhibitors? (2) What ethical considerations arise when offering chimeric antigen receptor (CAR) T-cell therapy to patients in rural or underserved regions? (3) How should clinicians navigate treatment decisions when cultural beliefs or health literacy affects understanding of immunotherapy risks?

- *Future directions and research (student-led inquiry)*: Students explored emerging topics such as next-generation CAR T-cell constructs, tumor microbiome effects on immunotherapy, predictive biomarkers, and novel combination strategies. Faculty provided a curated list of high-impact reviews and seminal papers. Each student selected a topic, analyzed current evidence, evaluated limitations of existing studies, and proposed future research directions and a novel therapy. Students presented their findings during a final oral presentation and received faculty feedback grounded in scientific rigor and clinical relevance.
- *Postelective quiz*: We designed the postelective quiz to parallel the prequiz in structure, length, and difficulty so that performance changes reflected student learning rather than differences in test format. Like the prequiz, the postquiz included 15 multiple-choice questions aligned with the same foundational and applied immunotherapy concepts. To ensure content validity, basic science and

XSL•FO
RenderX

clinical faculty mapped each postquiz item to the learning objectives addressed during the elective, including immune mechanisms of action, indications for Food and Drug Administration (FDA)–approved immunotherapies, and immune-related adverse events. Faculty conducted the same peer review, clarity check, and bias review procedures used for the prequiz. Student anonymity prevented individual-level tracking; therefore, we analyzed quiz results at the aggregate group level, comparing pre- and postcourse means to identify trends in knowledge acquisition. Students completed the postelective quiz immediately after the final oral presentations. Faculty used the results to evaluate learning outcomes, assess overall instructional effectiveness, and identify content areas requiring refinement in future course iterations.

Primer Development and Foundational Preparation

To strengthen students' readiness for advanced immunotherapy concepts, we provided a self-paced immunology primer at the start of the course. A basic scientist with immunology expertise and a clinical oncologist collaboratively selected the primer topics to ensure alignment with learning objectives and relevance to both foundational and clinical content. Medical students then drafted and refined the primer content, drawing on their experience in the preclerkship curriculum to identify which concepts were most challenging and most critical for clinical application during the clerkship curriculum. This student-driven development process (1) incorporated learners' perspectives to better target foundational gaps and (2) supported student contributors' professional growth through scholarly authorship and early medical education experience. Faculty reviewed and approved all content to ensure accuracy, clarity, and educational value. Integrating student voices added a peer-informed dimension to the primer, helping bridge the gap between didactic instruction and practical understanding.

Knowledge Assessment

We created pre- and postelective quizzes (Multimedia Appendix 1) to gauge students' baseline knowledge, measure learning gains, and evaluate the effectiveness of course content and teaching methods. On the first day of the elective, students completed the pre-elective quiz. Afterward, they received the immunology primer (Multimedia Appendix 2) as a foundational resource. The primer introduced key immunology topics, including innate and adaptive immunity, immune-cell functions, and mechanisms of immune activation. It emphasized how these principles govern the detection and elimination of cancer cells and highlighted the interplay between the immune system and the tumor microenvironment. The primer also incorporated recent advances such as checkpoint inhibitors and CAR T-cell therapy, enabling students to connect core immunological principles to cutting-edge cancer treatments. By reinforcing these concepts, the primer equipped students with the background needed to understand the complexities of cancer immunotherapy.

Table 1 outlines the key topics and learning objectives that guided students through essential features of the tumor microenvironment and the application of immunotherapy in cancer treatment. Students engaged with these objectives through an online discussion board (Canvas LMS) and delivered an oral presentation on the final day of the elective.

XSL•FO
RenderX

Table . List of topics and specific objectives for the online discussion and final presentation.

Topic	Objectives
Discuss tumor microenvironment and the basis for using immunotherapy.	• Outline the mechanism of how cancer cells lead to T-cell exhaustion. • Know the basics of other tumor microenvironment effects such as angiogenesis and extracellular matrix changes.
Describe the mechanism of action of immunotherapy techniques.	• Identify the role of PD-1[a], PD-L1[b], CTLA-4[c] inhibitors, and CAR[d] T-cells in cancer therapy.
What are the most common combinations of immunotherapy drugs used at the present time?	• Discuss current effective treatments available such as for melanoma and lung cancer. • Identify promising clinical trials and new developments in immunotherapy.
What are the adverse events commonly associated with immunotherapy?	• Identify common immunotherapy-related adverse events. • Know the current terminology used to classify immunotherapy adverse events.
What patient population is more at risk for adverse events while taking immunotherapy?	• Identify risk factors based on patient demographics, underlying medical conditions, and cancer-type specific reactions. • Determine whether a patient is a good candidate for immunotherapy given risk factors.
Novel therapies: Research current treatment modalities as well as developments in the field. Propose a new therapy protocol, given your new knowledge of immunotherapy.	• Propose a new therapy based on the understanding of immunotherapy mechanisms based on literature review. • Differentiate your innovative treatment from what is currently being used or investigated.

[a]PD-1: programmed death-1.

[b]PD-L1: programmed death ligand-1.

[c]CTLA-4: cytotoxic T-lymphocyte antigen 4.

[d]CAR: chimeric antigen receptor.

Participants

Participants in this study consisted of fourth-year medical students enrolled at TTUHSC El Paso. We used purposive sampling to target students in their final year who were eligible and available to participate in a 2-week online elective. Students qualified for inclusion in the research study if they voluntarily enrolled in the elective and provided informed consent to participate in the research study by completing postelective surveys. Participation required students to agree to the study procedures and complete all associated assessments. A total of 35 students enrolled in the elective, and 20 of those students completed the postcourse survey.

We focused on fourth-year students because they typically select electives that align with their intended specialty or postgraduate training plans (Table 2). We chose an online format for this elective based on prior institutional success with virtual coursework and to provide maximal flexibility for students who may travel for residency interviews during this period [18].

Table 2 shows the residency programs that students who completed the cancer immunotherapy elective applied to. The distribution illustrates their specialty preferences, with a notably higher number pursuing internal medicine compared with other fields.

XSL•FO
RenderX

Table . Residency programs medical students completing the cancer immunotherapy elective applied to.

Residency program	Students, n
Internal medicine	5
Pediatrics	3
Anesthesiology	3
Emergency medicine	2
Family medicine	2
Med/Peds	2
Otolaryngology	1
Pathology	1
Psychiatry	1
Total number of students	20

Assessment Instruments

We assessed students' knowledge using pre- and postcourse quizzes that immunology and oncology faculty collaboratively designed. Faculty with expertise in immunotherapy mechanisms, curriculum design, and clinical applications created quiz items and aligned them with the elective's core learning objectives. They also drew on national competency frameworks and relevant literature to support construct and content validity. We evaluated student discussion responses and oral presentations for critical thinking, integration of knowledge, and innovation, providing a complementary behavioral measure aligned with Kirkpatrick level 3 outcomes.

Survey Instruments

Oral Presentation Rubric Design and Validation

We evaluated student presentations using a structured rubric intentionally designed to measure four core constructs aligned with the elective's learning objectives:

1. *Conceptual understanding*: Accuracy and depth of immunology and immunotherapy knowledge, including mechanisms of action and tumor-immune interactions.
2. *Critical thinking and innovation*: Ability to synthesize literature, evaluate evidence, identify therapeutic gaps, and propose novel or evidence-based strategies.
3. *Clinical application*: Integration of clinical relevance, patient considerations, and implications for treatment decision-making.
4. *Communication and professionalism*: Clarity of oral communication, logical organization, visual design, and professional engagement with peers and faculty.

Rubric Development

A team of basic science and clinical faculty collaboratively drafted the rubric. To ensure content validity, faculty mapped each rubric domain directly to the course's stated competencies and to national guidelines for cancer immunotherapy education. The team reviewed representative presentation topics from prior years to ensure that the rubric captured the range of cognitive skills expected of fourth-year medical students.

Validation and Quality Assurance

We strengthened the rubric through a multistep validation process:

- *Peer review*: Three faculty with expertise in immunology, internal medicine, and oncology independently reviewed the rubric for accuracy, alignment with competencies, and clarity of scoring criteria.
- *Bias review*: Educational specialists examined the rubric for ambiguous language, construct-irrelevant variance, and potential bias related to topic selection or communication style.
- *Pilot testing*: Faculty applied the rubric to sample presentation summaries from a previous elective offering. This process allowed the team to refine wording, add performance anchors, and calibrate expectations.
- *Calibration session*: Before evaluation began, faculty reviewers participated in a group-norming session. They scored 2 anonymized sample presentations, compared ratings, and discussed scoring discrepancies to promote interrater consistency.

Scoring and Reliability

Two faculty evaluators independently scored each oral presentation using the finalized rubric. We measured interrater reliability by comparing paired scores across all rubric domains. Reviewers resolved discrepancies through discussion until reaching consensus. This dual-review system strengthened score reliability and helped ensure that the rubric consistently captured higher-order reasoning, conceptual accuracy, and clinical integration.

Postcourse Survey Instrument

We developed a postcourse survey to evaluate students' perceptions of the elective and capture qualitative feedback to inform iterative course improvements. We did not use a previously published instrument because no validated tool exists that specifically measures medical students' perceptions of a cancer immunotherapy elective or assesses learning experiences aligned with RE-AIM and Kirkpatrick evaluation frameworks for a cancer immunotherapy course. Instead, we designed a context-specific instrument grounded in established principles of educational assessment and program evaluation.

XSL•FO
RenderX

The postcourse survey consisted of Likert-scale items, open-ended questions, and demographic prompts and was administered anonymously through Qualtrics. To ensure content validity, the course faculty collaboratively generated the survey items. We mapped each item to the course objectives, the targeted competencies, and the relevant Kirkpatrick levels (Reaction and Learning) to confirm alignment. We refined the survey through faculty peer review, focusing on clarity, item relevance, cognitive load, and potential bias in item phrasing.

To enhance face validity, the draft survey was reviewed by 2 educational researchers unaffiliated with the course who provided feedback on structure, scale selection, and representativeness of domains (usefulness, relevance, comprehension, engagement, and areas for improvement). We revised the instrument accordingly. Due to the small sample size and anonymity requirements, we did not calculate reliability coefficients; however, reviewers assessed the consistency of underlying constructs and the appropriateness of response options.

The final instrument assessed the following domains:

- perception of elective usefulness and relevance;
- comprehension of immunotherapy concepts and techniques;
- engagement with course materials and faculty-led discussions;
- identification of areas requiring improvement; and
- perceived closing of knowledge gaps.

This approach allowed us to capture meaningful student feedback while ensuring that the instrument was appropriate for the elective's goals and context.

Data Collection

We administered pre- and postcourse quizzes electronically to assess group-level knowledge acquisition related to basic immunology and cancer immunotherapy. To maintain full anonymity in accordance with institutional review board (IRB) requirements, we collected all quiz responses without any identifiers or linkage markers. No individual-level identifiers were used; therefore, we analyzed all pre- and postcourse quiz data strictly in aggregate and did not attempt to match responses across time points. This approach allowed us to compare group performance before and after the elective while fully protecting student privacy.

Following course completion, we distributed an optional, anonymous postcourse survey through the Qualtrics platform. The survey included Likert-scale items and open-ended prompts assessing students' perceptions of course effectiveness, relevance of the content, confidence in applying immunotherapy concepts, and suggestions for improvement. Course faculty, including educational researchers, developed the survey to ensure alignment with course objectives and with Kirkpatrick level 1 (Reaction) and level 2 (Learning) outcomes.

Ethical Considerations

This study was reviewed by the TTUHSC El Paso IRB and determined to meet the criteria for exemption from full board review as a minimal-risk study (protocol no. E23116). In accordance with the approved IRB protocol, the requirement for written informed consent was waived. Participants were provided with an information statement describing the study purpose, planned use of data, and the voluntary nature of participation prior to study involvement. No identifiable information was collected, no compensation was provided, and participant privacy and confidentiality were maintained throughout data collection, analysis, and reporting.

Quantitative Data Analysis

We used a mixed methods approach to evaluate course outcomes. For the quantitative analysis, we compared group-level pre- and postcourse quiz scores using a 2-tailed unpaired *t* test. We did not link responses at the individual level, so we analyzed all scores in aggregate. The unpaired *t* test allowed us to assess differences in group means while preserving student anonymity.

When designing the knowledge assessment, we did not generate formal power calculations because this elective enrolled a small cohort (n=35), and the study design prioritized anonymity over paired testing, which limited our ability to estimate individual-level effect sizes. Instead, faculty experts in medical education, immunology, internal medicine, and oncology developed the quiz with the expectation, based on content coverage and alignment with the learning objectives, that students would show modest but measurable improvement in foundational immunotherapy knowledge. Given the cohort size and the use of unpaired aggregate scores, the study was underpowered to detect small differences between pre- and postcourse means.

Qualitative Data Analysis

For qualitative analysis, we examined open-ended survey responses using inductive thematic analysis, following Braun and Clarke's 6-step methodology [37]. We used Braun and Clarke's 6-step approach [37] to thematic analysis, a widely applied framework for identifying and interpreting patterns in qualitative data. This process involves becoming familiar with the data, generating initial codes, organizing codes into preliminary themes, refining and reviewing those themes, defining and naming them, and producing the final analytic narrative. We did not apply any a priori theoretical framework; instead, we allowed themes to emerge directly from the data. Two investigators (JC and HA) independently reviewed all responses and generated initial, descriptive codes that reflected recurring ideas in student comments (eg, "clarity of content," "workload concerns," "value of clinical discussions," "ethical considerations," "confidence applying immunotherapy concepts," and "suggestions for improvement"). The investigators met to compare and refine these codes, resolve discrepancies through consensus, and organize related codes into preliminary categories (eg, "perceived learning gains," "course structure," "engagement and interaction," and "areas needing clarification"). We iteratively refined a codebook as additional patterns emerged. Thematic saturation occurred when no new codes or concepts appeared in subsequent rounds of review, ensuring that the final themes accurately captured the breadth of student feedback and strengthened the trustworthiness of the analysis.

XSL•FO
RenderX

Implementing Student Feedback

As part of the postcourse survey, we asked students to identify areas for improvement and propose refinements to the elective. We systematically reviewed all feedback and used it to guide iterative course modifications. Student suggestions directly informed several enhancements implemented in subsequent offerings. For example, many students requested clearer guidance before the oral presentation. In response, we added a detailed grading rubric, sample presentation topics, and a faculty-led question and answer session to clarify expectations. Students also recommended streamlining the weekly workload. To address this, we adjusted reading assignments and aligned asynchronous modules more intentionally with live discussion topics to reduce redundancy and cognitive load.

These changes demonstrated our commitment to continuous quality improvement and student-centered course design. The iterative refinement process, informed by both qualitative and quantitative feedback, reflects the practical application of the RE-AIM and Kirkpatrick frameworks to enhance course delivery and educational effectiveness.

Study Duration

The study timeline encompassed data collection, analysis, and dissemination of findings over an estimated 1-2 months following the conclusion of the elective. Timely dissemination of findings ensures prompt feedback and informed iterative elective design and delivery improvements, thereby fostering continuous quality enhancement in medical education at TTUHSC El Paso.

Results

Knowledge Assessment (Pre- and Postcourse Quiz)

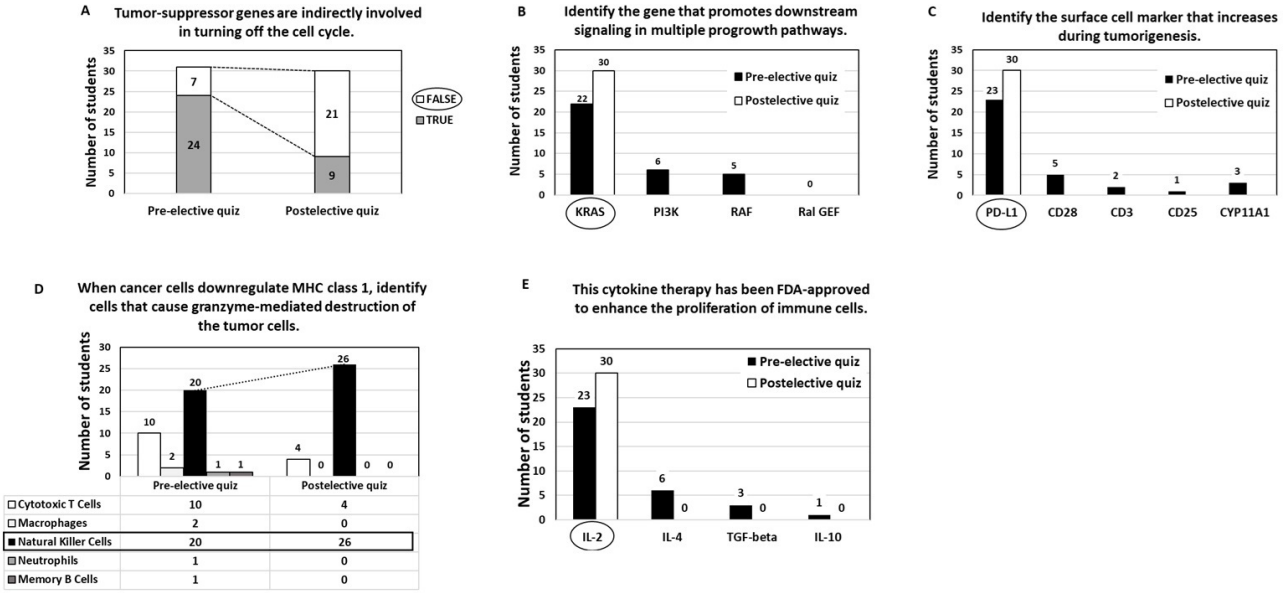
A total of 35 students completed the elective, and all quiz data were analyzed in aggregate. Once the students completed the course, they took a postelective quiz. Compared with their performance on the pre-elective quiz, fourth-year medical students demonstrated an improvement in their understanding of basic immunology and cancer immunotherapy topics after completing the elective. For example, students were able to accurately identify the role of tumor-suppressor genes (Figure 1A), the function of *KRAS* (Figure 1B), and the roles of PD-L1 (Figure 1C) and Natural Killer cells (Figure 1D) within the tumor microenvironment. In addition, students could correctly pinpoint which cytokine therapy was FDA-approved for cancer treatment (Figure 1E).

Figure 1 illustrates the enhanced understanding and knowledge gained by students in these critical areas of cancer biology and immunotherapy after completing the elective. Although the 2-tailed unpaired *t* test did not show statistical significance, postelective quiz scores trended higher than pre-elective scores, suggesting a potential improvement in knowledge. This trend supports the educational value of the elective and highlights the need for further study with paired data.

The study design did not link responses at the individual level, and no power calculation was conducted to estimate detectable effect size; therefore, the analysis remained exploratory rather than inferential. The aggregate data nevertheless provided useful information about baseline and end-of-course performance and will guide future refinement of assessment methods, including the use of paired data and sample-size estimation.

XSL•FO
RenderX

Figure 1. Student proficiency in key concepts of tumor biology and immunotherapy. Compared with pre-elective performance, postelective assessments demonstrated improved student ability to identify and apply foundational concepts in cancer biology and immunotherapy. Specifically, students more accurately recognized the role of tumor-suppressor genes in regulating cell cycle control and preventing malignant transformation (A), as well as the oncogenic function of *KRAS* in driving aberrant intracellular signaling pathways involved in tumor growth and survival (B). In addition, students showed increased understanding of immune regulatory mechanisms within the tumor microenvironment, including the role of PD-L1 in mediating immune checkpoint inhibition (C) and the contribution of natural killer cells to innate antitumor immune responses (D) Students also correctly identified the Food and Drug Administration–approved cytokine therapy used clinically to enhance immune cell proliferation and activity, reflecting improved recognition of immunomodulatory treatment strategies. (E) Together, these results indicate enhanced conceptual understanding across molecular, cellular, and immunologic domains central to contemporary cancer immunotherapy following completion of the elective. FDA: Food and Drug Administration; IL: interleukin-2; MHC: major histocompatibility complex; TGF: transforming growth factor.



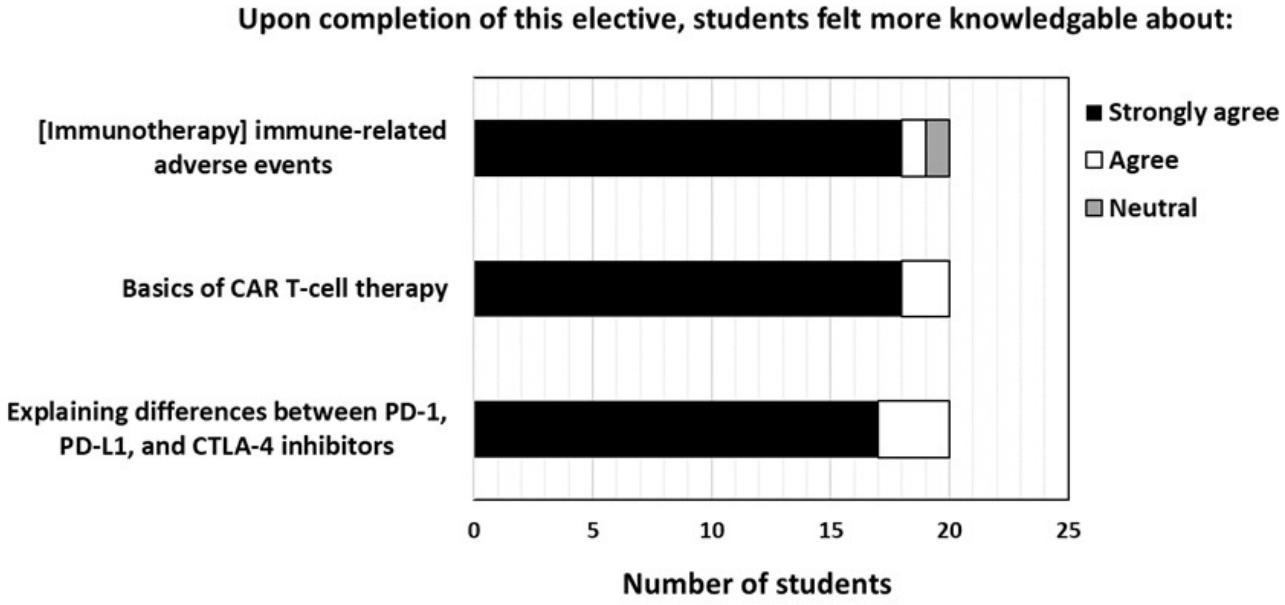
Postcourse Survey

A total of 35 students completed the elective, and 20 students completed the postcourse survey. Figure 2 and Multimedia Appendix 3 summarize the distribution of responses. Students consistently rated the elective favorably across all surveyed domains.

Evaluation of student perceptions demonstrated meaningful gains in their understanding of immune-related adverse events and core immunotherapy mechanisms. Students expressed increased confidence in identifying and describing the types of immune-related adverse events observed in patients undergoing immunotherapy (Figure 2, top). They also reported greater mastery of fundamental CAR T-cell principles (Figure 2, middle). Importantly, every student agreed that the course improved their ability to distinguish between programmed death-1 (PD-1), programmed death ligand-1 (PD-L1), and cytotoxic T-lymphocyte antigen 4 (CTLA-4) (Figure 2 bottom).

XSL•FO
RenderX

Figure 2. Exploring medical students' understanding of immune-related adverse events and cancer immunotherapy treatments. Students felt knowledgeable about various types of immune-related adverse events observed in some patients receiving immunotherapy (top). Students felt more knowledgeable about the basics of CAR T-cell therapy (middle). All students agreed that the cancer immunotherapy course enabled them to explain the differences between PD-1, PD-L1, and CTLA-4 (bottom). CAR: chimeric antigen receptor; CTLA-4, cytotoxic T-lymphocyte antigen 4; PD-1: programmed death-1; PD-L1, programmed death ligand-1.



Students agreed or strongly agreed that the course was instrumental in learning current immunotherapy techniques (Multimedia Appendix 3A). Students highlighted that the course provided a solid foundation and deeper understanding of cutting-edge immunotherapy practices. Overall, students felt well prepared to apply these techniques in real-world settings.

In addition, the course significantly enhanced students' ability to think critically about developing novel cancer immunotherapies. Nineteen out of twenty students agreed or strongly agreed that the course's comprehensive curriculum and emphasis on innovative thinking fostered their analytical skills and creativity (Multimedia Appendix 3B). Through understanding basic science concepts and reviewing the current literature, students were encouraged to explore and devise new strategies for cancer treatment. This critical thinking approach not only deepened their understanding of current immunotherapy techniques but also inspired them to contribute to future advancements in the field.

Finally, in a unanimous consensus, medical students have expressed overwhelming satisfaction with the comprehensive coverage of cancer immunotherapy in their recent course. Each student acknowledged that the course effectively addressed and filled significant gaps in their understanding of immunotherapy (Multimedia Appendix 3C). From foundational concepts to advanced applications, the material provided clarity on complex mechanisms and cutting-edge research in the field. This collective sentiment underscores the course's success in equipping future health care professionals with essential knowledge and skills crucial for addressing the challenges of modern cancer treatment.

Twenty students completed the postelective survey. Table 3 shows median scores with IQRs for all Likert-scale items. Students reported overwhelmingly positive perceptions of the cancer immunotherapy elective, with consistently high ratings across all survey items. For every item, the median response was 5 (Strongly agree), indicating strong endorsement of the course's effectiveness. Students agreed that the elective was useful for their learning regarding current immunotherapy techniques (median5, IQR 0) and that the introductory immunology overview supported their understanding of later content (median 5, IQR 0.25).

Students also felt that the course enhanced their ability to think critically about developing novel immunotherapy approaches (median 5, IQR 0) and improved their ability to explain differences among PD-1, PD-L1, and CTLA-4 inhibitors (median 5, IQR 0). Similarly, learners reported increased knowledge of CAR T-cell therapy (median 5, IQR 0) and felt that the amount of course material was appropriate to address gaps in their prior knowledge (median 5, IQR 0).

Students further indicated that their understanding of immune-related adverse events improved as a result of the elective (median 5, IQR 0). The only item with notable response variability was students' confidence in identifying steps needed to manage patients receiving immunotherapy during residency (median 5, IQR 1.0), suggesting greater heterogeneity in perceived readiness for clinical application.

Overall, these results demonstrate consistently strong satisfaction with the elective and meaningful perceived gains in foundational and applied cancer immunotherapy knowledge. Collectively, these findings suggest high perceived educational value and alignment between course objectives and learner-reported outcomes. A summary of the highest-rated domains is demonstrated in Table 3.

XSL•FO
RenderX

Table . Twenty students completed the postelective survey[a].

Item	Median	IQR
Q1. Course usefulness for learning immunotherapy techniques.	5	0
Q2. Immunology overview helped understanding.	5	0.25
Q3. Course enabled critical thinking about novel immunotherapy approaches.	5	0
Q4. Ability to explain PD-1[b], PD-L1[c], and CTLA-4[d] differences.	5	0
Q5. Increased knowledge of CAR[e] T-cell therapy.	5	0
Q6. Amount of information was appropriate.	5	0
Q7. Increased knowledge of immune-related adverse events.	5	0
Q8. Ability to identify steps to manage patients receiving immunotherapy.	5	1.0

[a]Median scores and IQRs indicate overwhelmingly positive perceptions of the elective, with all items receiving a median of 5 (Strongly agree). These findings reflect strong alignment between the course objectives and students' reported gains in cancer immunotherapy knowledge.

[b]PD-1: programmed death-1.

[c]PD-L1: programmed death ligand-1.

[d]CTLA-4: cytotoxic T-lymphocyte antigen 4.

[e]CAR: chimeric antigen receptor.

Qualitative Themes From Open-Ended Responses

Inductive thematic analysis of open-ended survey responses revealed 4 dominant themes that collectively illustrate how students experienced the elective:

- Theme 1: Increased Conceptual Clarity reflected students' reports of gaining a clearer understanding of immune checkpoint pathways, mechanisms of T-cell activation, and distinctions among PD-1, PD-L1, and CTLA-4.
- Theme 2: Strengthened Ability to Apply Concepts Clinically captured students' increased confidence in connecting mechanistic immunology to therapeutic decision-making, recognition of immune-related adverse events, and understanding FDA-approved therapies such as CAR T-cell therapy.
- Theme 3: Value of Structured, Faculty-Guided Discussion highlighted the importance of guided prompts and faculty facilitation in promoting deeper inquiry and enabling sustained dialogue around ethics, access to care, cost barriers, disparities in clinical trials, and social determinants of health shaping immunotherapy outcomes.
- Theme 4: Recommendations for Course Enhancement included requests for clearer expectations for oral presentations, streamlined weekly readings, and expanded case-based discussions, particularly those incorporating literature focused on marginalized populations.

These qualitative insights not only informed targeted course refinements but also aligned with the evaluative dimensions of the RE-AIM and Kirkpatrick models described in the "Methods" section. Students also expressed strong appreciation for the structured discussion prompts, faculty engagement, and the primer's effectiveness in scaffolding learning.

Discussion

Principal Findings

This mixed methods evaluation demonstrated that a fourth-year online cancer immunotherapy elective improved students' foundational understanding of immunotherapy principles, strengthened their ability to apply mechanistic knowledge to clinical scenarios, and enhanced their confidence discussing emerging therapies. Students showed measurable gains in core immunology and immunotherapy concepts on the postelective knowledge assessment, and their final oral presentations provided evidence of higher-order reasoning, integration of basic and clinical science, and accurate articulation of complex mechanisms such as PD-1/PD-L1 and CTLA-4 pathways, CAR T-cell engineering, and immune-related toxicities. Qualitative analysis of postelective survey responses further revealed that students valued the elective's relevance, clarity, and emphasis on clinical translation, noting that it addressed significant gaps in their prior training and supported their ability to think critically about current and emerging cancer immunotherapies. Collectively, these findings indicate that the elective achieved its intended goal of enhancing immunotherapy competency among senior medical students.

Implications of Findings

The findings from this evaluation demonstrate that a structured, 2-week online cancer immunotherapy elective can enhance medical students' self-reported understanding of fundamental immunotherapy concepts, mechanisms of action, and FDA-approved treatments. Students also indicated that the elective strengthened their ability to critique scientific literature and articulate clinical applications, skills reinforced through

XSL•FO
RenderX

structured discussion prompts and a scored oral presentation. These results suggest that even a short, focused elective can support competency development in emerging therapeutic domains.

The mixed methods data further indicate that pairing asynchronous foundational content (eg, the immunology primer) with guided synchronous discussions and a final presentation provides an effective instructional sequence for fourth-year learners. Students consistently rated the course as useful, relevant, and appropriately rigorous, reflecting the value of integrating preexisting knowledge, analytic tasks, and application-based assessments. This supports the feasibility of using similar instructional designs to introduce other rapidly evolving biomedical topics.

In addition, the elective's structure, which included evidence-based discussion prompts and opportunities for students to explain mechanisms to peers, offered a model that other institutions may adapt when integrating novel scientific content into the curriculum. Implementing a peer-informed approach to resource development (eg, the student-written primer) may also enhance perceived relevance and learning efficiency for medical students.

These implications extend primarily to curriculum design, instructional sequencing, and feasibility of implementation in undergraduate medical education. While the course addressed selected ethical and access-related issues within immunotherapy during discussions, the study did not evaluate growth in students' understanding of social determinants of health or culturally competent care. Future research using targeted assessments could explore these domains more explicitly.

Comparison With Literature

The integration of immunotherapy into undergraduate medical education remains limited, yet our findings mirror a growing body of evidence supporting earlier and more structured exposure to advanced cancer therapeutics. Several studies have reported that medical trainees feel underprepared to discuss immunotherapy, interpret mechanism-of-action data, or counsel patients about associated risks and benefits, a gap that our elective similarly sought to address [38]. Recent educational interventions have been developed during graduate medical education, including short courses, case-based modules, and interdisciplinary workshops.

Other research teams have emphasized the value of combining foundational basic science reviews with clinical case discussions to enhance long-term retention and clinical reasoning, an approach consistent with our spiral design and primer-supported structure [39,40]. These parallels support the relevance and effectiveness of our elective's hybrid instructional format.

There is also a small but emerging literature on the importance of tailoring immunotherapy education to local contexts, including programs serving diverse or medically underserved communities. These studies underscore the need for cultural and demographic relevance when teaching advanced therapeutics, a theme that aligns with, but does not overextend beyond, the reflections shared by our students regarding practice readiness and communication skills [41-45].

Together, these converging findings suggest that structured, evidence-informed immunotherapy curricula can enhance medical students' conceptual understanding and clinical preparedness. They also highlight the broader national momentum toward incorporating cutting-edge oncologic therapeutics earlier in medical training. Future multi-institutional studies will be essential to compare outcomes across diverse curricular models and to determine best practices for sustainable integration.

Strengths and Limitations

A major strength of this elective was its alignment with up-to-date clinical and scientific standards in cancer immunotherapy. Due to the field evolving rapidly, faculty routinely reviewed and updated course content using evidence-based guidelines from the American Society of Clinical Oncology and the Society for Immunotherapy of Cancer. This process ensured that students engaged with current FDA-approved therapies, emerging research directions, and clinically relevant case examples, thereby reinforcing the scientific rigor and clinical applicability of the curriculum.

Another strength was the elective's intentional design for a modular, hybrid structure, blending asynchronous learning with synchronous faculty-led discussion, allowing the course to adapt to varying institutional schedules and curricular models. In addition, its use of structured discussion prompts and a capstone innovation project created flexible instructional components that other medical schools could readily adopt or tailor to their own contexts.

The course also benefited from strong interdisciplinary collaboration. Faculty from immunology, oncology, biochemistry, clinical genetics, and internal medicine jointly developed and implemented the elective, ensuring cohesion across basic science and clinical content. Student cocreation, particularly in the development of the immunology primer, added a peer-informed dimension that enhanced clarity, relevance, and learner engagement.

Despite these strengths, several limitations warrant attention. First, participation in both the elective and the postcourse survey was voluntary, introducing potential self-selection and response bias that may limit the generalizability of the findings. Second, the retrospective nature of some perception-based survey items may have contributed to recall bias. Third, although the pre- and postcourse quizzes assessed knowledge gains, all data were analyzed at the aggregate level to maintain anonymity. The absence of individually paired responses limited our ability to conduct more detailed analyses of individual learning trajectories or matched-pair statistical testing.

The elective also faced implementation challenges, including variable student schedules, faculty workload constraints, and occasional technological barriers inherent to online instruction. Students on clinical rotations sometimes experienced difficulties synchronizing schedules with synchronous components. While some students pursued optional shadowing opportunities with oncologists, the elective did not include a formal clinical rotation, representing a potential area for future enhancement.

XSL•FO
RenderX

To address these challenges, we incorporated contingency measures such as recorded sessions, flexible deadlines, and alternative participation formats for discussion activities. Regular faculty meetings and scheduled check-ins with students allowed the instructional team to respond to emerging needs in real time. Attrition was minimal, as all enrolled students completed the elective; however, not all students completed the postcourse survey.

Lessons Learned

Implementing this online cancer immunotherapy elective provided valuable insights into curriculum development, learner engagement, and program scalability. Early and meaningful collaboration between basic scientists, clinicians, and students was essential to ensure content relevance and clarity. Notably, the cocreation of the immunotherapy primer by medical students offered a peer-informed perspective that enhanced both accessibility and professional development for the authors.

Flexibility in course delivery, through asynchronous modules supplemented by live sessions, proved crucial for accommodating students' clinical schedules and different learning styles. However, we learned that clear alignment between asynchronous and synchronous content was essential to avoid redundancy and promote continuity. Later course iterations were adjusted to better integrate these components.

Embedding structured opportunities for innovation, such as the final oral presentation task, fostered critical thinking and interdisciplinary application. Students were encouraged to propose novel therapeutic strategies, and faculty feedback emphasized this capstone experience as one of the most valuable elements of the elective.

Continuous improvement based on student feedback was also key. The anonymous survey asked students to identify strengths and areas for improvement, which led to enhancements in discussion prompts, session timing, and technical delivery. These efforts reinforced a culture of responsiveness and reflective teaching and demonstrated the importance of iterative design in building a sustainable, scalable intervention.

Future Directions

Future research should explore how this elective contributes to long-term competence in cancer immunotherapy, especially during postgraduate training. While short-term outcomes demonstrated knowledge gain and student satisfaction, it remains critical to assess whether these improvements translate to clinical proficiency, leadership in immunotherapy initiatives, and patient-centered decision-making.

A key next step is longitudinal tracking of alumni who completed the elective, evaluating how early exposure to cancer immunotherapy influences clinical performance, advocacy, and continued professional development. This may include follow-up surveys, interviews, or program director assessments to gauge real-world application of course content.

Alternatively, conducting a cross-sectional cohort study comparing students who completed the elective with those who did not would offer insight into differences in long-term knowledge retention, confidence in discussing immunotherapy with patients, and likelihood of pursuing related research or clinical specialties.

To improve generalizability, future iterations of the elective should be piloted across additional medical schools with varying curricular structures and student demographics. Multi-institutional collaborations could help evaluate the adaptability, feasibility, and sustainability of this instructional model and identify institution-specific facilitators and barriers. Incorporating paired pre-post knowledge assessments or longitudinal tracking of clinical application skills may also strengthen future evaluation designs.

Expanding the elective to include interprofessional learners, such as pharmacy, nursing, and physician assistant students, could enhance team-based education and cancer care delivery. Finally, multisite implementation at diverse institutions will be critical for testing the generalizability and adaptability of the curriculum, and future iterations must remain agile to incorporate emerging therapies and updated guidelines in this rapidly evolving field.

Acknowledgments

We thank the students for participating in the course and the research study.

Funding

No external financial support or grants were received from any public, commercial, or not-for-profit entities for the research, authorship, or publication of this article.

Data Availability

Data can be accessed upon request to the corresponding author.

Authors' Contributions

Conceptualization: MR, RH, BG, SM, HA, CP, JC
Data curation: RH, HA, JC
Formal analysis: RH, HA, JC
Investigation: MR, RH, SM, HA, CP, JC

XSL•FO
RenderX

Methodology: MR, RH, SM, HA, CP, JC
Project administration: SM, HA, CP, JC
Resources: MR, RH, BG, SM, HA, CP, JC
Software: MR, RH, HA, JC
Supervision: SM, HA, CP, JC
Validation: SM, HA, CP, JC
Visualization: MR, RH, HA, JC
Writing – original draft: MR, RH, BG, SM, HA, CP, JC
Writing – review & editing: MR, RH, BG, SM, HA, CP, JC

Conflicts of Interest

None declared.

Multimedia Appendix 1
Pre- and postelective quiz instrument. This quiz assessed medical students' baseline and postcourse knowledge of cancer immunotherapy, including mechanisms of action, clinical applications, and immune-related adverse events. The same set of questions was administered before and after the elective to evaluate knowledge gains and support course effectiveness analysis.
[PDF File, 143 KB - mededu_v12i1e71628_app1.pdf]

Multimedia Appendix 2
Immunology primer overview. This primer provided foundational content on the immune system, including innate and adaptive immunity, key immune cells, cytokine signaling, and tumor immunology concepts. It was designed to ensure that students entered the elective with a baseline understanding necessary to engage in advanced cancer immunotherapy discussions.
[PDF File, 1298 KB - mededu_v12i1e71628_app2.pdf]

Multimedia Appendix 3
Comprehensive impact of the immunotherapy course on medical students' knowledge and critical thinking. (A) Students strongly agreed that the course was instrumental in learning current immunotherapy techniques, providing up-to-date content and practical applications that prepared them well for real-world settings. (B) The new immunotherapy course significantly enhanced students' critical thinking in developing novel cancer treatments, with 19 out of 20 students affirming its impact. (C) Feedback from 20 medical students unanimously confirming that the content covered in the cancer immunotherapy course significantly filled gaps and deficiencies in their knowledge of immunotherapy. Each student reported agreement with the comprehensive nature of the course material, highlighting its effectiveness in addressing both fundamental concepts and advanced applications in the field of cancer treatment.
[JPG File, 127 KB - mededu_v12i1e71628_app3.jpg]

References

1. Abbott M, Ustoyev Y. Cancer and the immune system: the history and background of immunotherapy. Semin Oncol Nurs 2019 Oct;35(5):150923. [doi: 10.1016/j.soncn.2019.08.002]
2. Ginex PK, Brassil K, Ely B. Immunotherapy: exploring the state of the science. Clin J Oncol Nurs 2017 Apr 1;21(2 Suppl):9-12. [doi: 10.1188/17.CJON.S2.9-12] [Medline: 28315554]
3. Le I, Dhandayuthapani S, Chacon J, Eiring AM, Gadad SS. Harnessing the immune system with cancer vaccines: from prevention to therapeutics. Vaccines (Basel) 2022 May 21;10(5):816. [doi: 10.3390/vaccines10050816]
4. Kye Y, Nagineni L, Gadad S, et al. The identification and clinical applications of mutated antigens in the era of immunotherapy. Cancers (Basel) 2022 Aug 31;14(17):4255. [doi: 10.3390/cancers14174255]
5. Ramirez F, Zambrano A, Hennis R, Holland N, Lakshmanaswamy R, Chacon J. Sending a message: use of mRNA vaccines to target the tumor immune microenvironment. Vaccines (Basel) 2023 Sep 7;11(9):1465. [doi: 10.3390/vaccines11091465]
6. Ha MT, Siddiqui ZS. Understanding medical students' transition to clinical training: a qualitative study of transformative learning and professional identity formation. BMJ Open 2025 Jun 17;15(6):e098675. [doi: 10.1136/bmjopen-2024-098675] [Medline: 40527562]
7. Mirzadeh P, Pituskin E, Au I, Sneath S, Buick CJ. Cancer immunotherapy: the role of nursing in patient education, assessment, monitoring, and support. Curr Oncol 2025 Jul 9;32(7):7. [doi: 10.3390/curroncol32070392] [Medline: 40710202]
8. Holtrop JS, Estabrooks PA, Gaglio B, et al. Understanding and applying the RE-AIM framework: clarifications and resources. J Clin Transl Sci 2021;5(1):e126. [doi: 10.1017/cts.2021.789] [Medline: 34367671]
9. Glasgow RE, Battaglia C, McCreight M, et al. Use of the reach, effectiveness, adoption, implementation, and maintenance (RE-AIM) framework to guide iterative adaptations: applications, lessons learned, and future directions. Front Health Serv 2022;2:959565. [doi: 10.3389/frhs.2022.959565] [Medline: 36925843]

XSL•FO
RenderX

10. Rouse DN. Employing Kirkpatrick's evaluation framework to determine the effectiveness of health information management courses and programs. Perspect Health Inf Manag 2011 Apr 1;8(Spring):1c. [Medline: 21464860]
11. Chacon JA, Cervantes JL, Perry CN, Pfarr CM, Ayoubieh H. Personalized medicine in undergraduate medical education: a spiral learning model. Med Sci Educ 2020 Dec;30(4):1741-1744. [doi: 10.1007/s40670-020-01066-0] [Medline: 34457840]
12. DeSilets LD. An update on Kirkpatrick's model of evaluation: part two. J Contin Educ Nurs 2018 Jul 1;49(7):292-293. [doi: 10.3928/00220124-20180613-02] [Medline: 29939374]
13. Johnston S, Coyer FM, Nash R. Kirkpatrick's evaluation of simulation and debriefing in health care education: a systematic review. J Nurs Educ 2018 Jul 1;57(7):393-398. [doi: 10.3928/01484834-20180618-03] [Medline: 29958308]
14. Whitcomb ME. Medical education research. Acad Med 2004 Oct;79(10):907. [doi: 10.1097/00001888-200410000-00001] [Medline: 15383344]
15. Hankir AK, Northall A, Zaman R. Stigma and mental health challenges in medical students. BMJ Case Rep 2014 Sep 2;2014:bcr2014205226. [doi: 10.1136/bcr-2014-205226] [Medline: 25183806]
16. Jupina M, Sidle MW, Rehmeyer Caudill CJ. Medical student mental health during the COVID-19 pandemic. Clin Teach 2022 Oct;19(5):e13518. [doi: 10.1111/tct.13518] [Medline: 35909320]
17. Pacheco JP, Giacomin HT, Tam WW, et al. Mental health problems among medical students in Brazil: a systematic review and meta-analysis. Braz J Psychiatry 2017;39(4):369-378. [doi: 10.1590/1516-4446-2017-2223] [Medline: 28876408]
18. Ayoubieh H, Perry C, Cervantes J, Pfarr C, Chacon J. Development of an online SARS-CoV-2/COVID-19 elective course for fourth-year medical students. Med Sci Educ 2023 Feb;33(1):21-25. [doi: 10.1007/s40670-023-01726-x] [Medline: 36713276]
19. Guo MZ, Allen J, Sakumoto M, Pahwa A, Santhosh L. Reimagining Undergraduate Medical Education in a Post-COVID-19 Landscape. J Gen Intern Med 2022 Jul;37(9):2297-2301. [doi: 10.1007/s11606-022-07503-7] [Medline: 35710661]
20. Kim KJ, Kim G, Kang Y. Faculty perceptions and use of e-learning resources for medical education and future predictions. Korean J Med Educ 2023 Dec;35(4):325-334. [doi: 10.3946/kjme.2023.270] [Medline: 38062680]
21. Tosto SA, Alyahya J, Espinoza V, McCarthy K, Tcherni-Buzzeo M. Online learning in the wake of the COVID-19 pandemic: mixed methods analysis of student views by demographic group. Soc Sci Humanit Open 2023;8(1):100598. [doi: 10.1016/j.ssaho.2023.100598] [Medline: 37366390]
22. Harden RM. The integration ladder: a tool for curriculum planning and evaluation. Med Educ 2000 Jul;34(7):551-557. [doi: 10.1046/j.1365-2923.2000.00697.x] [Medline: 10886638]
23. Dennick R. Constructivism: reflections on twenty five years teaching the constructivist approach in medical education. Int J Med Educ 2016 Jun 25;7:200-205. [doi: 10.5116/ijme.5763.de11] [Medline: 27344115]
24. Modo M, Kinchin I. A conceptual framework for interdisciplinary curriculum design: a case study in neuroscience. J Undergrad Neurosci Educ 2011;10(1):A71-A79. [Medline: 23626496]
25. Chaffey L, de Leeuw EJ, Finnigan G. Facilitating students′ reflective practice in a medical course: literature review. Educ Health 2012;25(3):198. [doi: 10.4103/1357-6283.109787]
26. Rogers SL, Priddis LE, Michels N, Tieman M, Van Winkle LJ. Applications of the reflective practice questionnaire in medical education. BMC Med Educ 2019 Feb 7;19(1):47. [doi: 10.1186/s12909-019-1481-6] [Medline: 30732611]
27. Sawyer T, White M, Zaveri P, et al. Learn, see, practice, prove, do, maintain: an evidence-based pedagogical framework for procedural skill training in medicine. Acad Med 2015 Aug;90(8):1025-1033. [doi: 10.1097/ACM.0000000000000734] [Medline: 25881645]
28. Engqvist Boman L, Sundberg K, Petersson LM, Backman M, Silén C. A pedagogical model to enhance nurses' ability to support patient learning: an educational design research study. Int J Med Educ 2022 Jul 29;13:176-186. [doi: 10.5116/ijme.62c2.b9c4]
29. Manninen K, Henriksson EW, Scheja M, Silén C. Patients' approaches to students' learning at a clinical education ward—an ethnographic study. BMC Med Educ 2014 Jul 2;14:131. [doi: 10.1186/1472-6920-14-131] [Medline: 24989155]
30. Haidet P, Jarecke J, Adams NE, et al. A guiding framework to maximise the power of the arts in medical education: a systematic review and metasynthesis. Med Educ 2016 Mar;50(3):320-331. [doi: 10.1111/medu.12925] [Medline: 26896017]
31. Alharbi NS. Evaluating competency-based medical education: a systematized review of current practices. BMC Med Educ 2024 Jun 3;24(1):612. [doi: 10.1186/s12909-024-05609-6] [Medline: 38831271]
32. Lee GB, Chiu AM. Assessment and feedback methods in competency-based medical education. Ann Allergy Asthma Immunol 2022 Mar;128(3):256-262. [doi: 10.1016/j.anai.2021.12.010] [Medline: 34929390]
33. Levine MF, Shorten G. Competency-based medical education: its time has arrived. Can J Anaesth 2016 Jul;63(7):802-806. [doi: 10.1007/s12630-016-0638-6] [Medline: 27030130]
34. Postow MA, Callahan MK, Wolchok JD. Immune checkpoint blockade in cancer therapy. J Clin Oncol 2015 Jun 10;33(17):1974-1982. [doi: 10.1200/JCO.2014.59.4358] [Medline: 25605845]
35. Wiley K, LeFebvre KB, Wall L, et al. Immunotherapy administration: oncology nursing society recommendations. Clin J Oncol Nurs 2017 Apr 1;21(2 Suppl):5-7. [doi: 10.1188/17.CJON.S2.5-7] [Medline: 28315551]
36. Galioto M, Mucenski J. Immunotherapy summit: proceedings and identified priorities for safe administration and care. Clin J Oncol Nurs 2019 Jun 1;23(3):E60-E65. [doi: 10.1188/19.CJON.E60-E65] [Medline: 31099794]

XSL•FO
RenderX

37. Lim AJS, Hong DZ, Pisupati A, et al. Portfolio use in postgraduate medical education: a systematic scoping review. Postgrad Med J 2023 Jul 21;99(1174):913-927. [doi: 10.1093/postmj/qgac007] [Medline: 36961214]
38. Fulton-Ward T, Bain R, Khoury EG, et al. Benefits of mentoring in oncology education for mentors and mentees: pre-post interventional study of the British Oncology Network for Undergraduate Societies' national oncology mentorship scheme. JMIR Med Educ 2023 Sep 11;9:e48263. [doi: 10.2196/48263] [Medline: 37695662]
39. Alkuran O, Al-Mehaisen L, Abu Mahfouz I, et al. Distance electronic learning strategy in medical teaching during the COVID-19 pandemic: cross-sectional survey study. JMIR Med Educ 2023 Dec 5;9:e42354. [doi: 10.2196/42354] [Medline: 38051556]
40. Kenjrawi Y, Dashash M. The first asynchronous online evidence-based medicine course for Syrian health workforce: effectiveness and feasibility pilot study. JMIR Form Res 2022 Oct 25;6(10):e36782. [doi: 10.2196/36782] [Medline: 36282556]
41. Emens LA, Romero PJ, Anderson AC, et al. Challenges and opportunities in cancer immunotherapy: a Society for Immunotherapy of Cancer (SITC) strategic vision. J Immunother Cancer 2024 Jun 19;12(6):6. [doi: 10.1136/jitc-2024-009063] [Medline: 38901879]
42. Gupta A, Omeogu C, Islam JY, et al. Socioeconomic disparities in immunotherapy use among advanced-stage non-small cell lung cancer patients: analysis of the National Cancer Database. Sci Rep 2023 May 20;13(1):8190. [doi: 10.1038/s41598-023-35216-2]
43. Vasquez VM Jr, McCabe M, McKee JC, et al. Transforming cancer care: a narrative review on leveraging artificial intelligence to advance immunotherapy in underserved communities. J Clin Med 2025 Jul 29;14(15):15. [doi: 10.3390/jcm14155346] [Medline: 40806968]
44. Ramirez F, Alvarado L, Dwivedi A, Chacon J. Immunotherapy treatment disparities: a Texas Cancer Registry analysis of patients with cutaneous melanoma. J Immigr Minor Health 2024 Feb;26(1):23-33. [doi: 10.1007/s10903-023-01536-8] [Medline: 37615821]
45. Ramirez F, Vasquez R, Gill J, et al. Understanding the association between social determinants of health and receipt of immunotherapy for melanoma. J Drugs Dermatol 2024 May 1;23(5):311-315. [doi: 10.36849/JDD.7803] [Medline: 38709695]

Abbreviations

CAR: chimeric antigen receptor
CTLA-4: cytotoxic T-lymphocyte antigen 4
FDA: Food and Drug Administration
IRB: institutional review board
PD-1: programmed death-1
PD-L1: programmed death ligand-1
RE-AIM: Reach, Effectiveness, Adoption, Implementation, and Maintenance
TTUHSC: Texas Tech University Health Sciences Center

Edited by B Lesselroth; submitted 22.Jan.2025; peer-reviewed by AN Abbasi, RR Kameny, S Manivasakan; revised version received 01.Dec.2025; accepted 03.Dec.2025; published 21.Jan.2026.

Please cite as:
Raynor M, Hock R, Godinich B, Maharaj S, Ayoubieh H, Perry C, Chacon J
Implementation and Evaluation of a Cancer Immunotherapy Elective for Medical Students: Mixed Methods Descriptive Study
JMIR Med Educ 2026;12:e71628
URL: https://mededu.jmir.org/2026/1/e71628
doi:10.2196/71628

© Mark Raynor, Rivers Hock, Brandon Godinich, Satish Maharaj, Houriya Ayoubieh, Cynthia Perry, Jessica Chacon. Originally published in JMIR Medical Education (https://mededu.jmir.org), 21.Jan.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Education, is properly cited. The complete bibliographic information, a link to the original publication on https://mededu.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Effect of an Online Continuing Professional Development Course on Physicians' Intention to Approach a Colleague in Difficulty: Mixed Methods Convergent Study

Florence Lizotte[1], MD, MSc; Martin Tremblay[2], PhD; Caroline Biron[3,4], PhD; Éloi Lachance[1], BA; Souleymane Gadio[3], MSc; Roberta de Carvalho Corôa[3], MSc, PhD; Claude Bernard Uwizeye[3], MSc; Sam J Daniel[2,5], MD, MS; France Légaré[1,3,6], BArch, MD, MSc, PhD

[1]Faculty of Medicine, Université Laval, Québec, QC, Canada

[2]Fédération des médecins spécialistes du Québec, Montréal, QC, Canada

[3]VITAM – Centre de recherche en santé durable, Université Laval, Québec, QC, Canada

[4]Department of Management, Université Laval, Québec, QC, Canada

[5]Faculty of Medicine, McGill University, Montreal, QC, Canada

[6]Centre de recherche du CHU de Québec, Université Laval, Québec, QC, Canada

Corresponding Author:
France Légaré, BArch, MD, MSc, PhD
VITAM – Centre de recherche en santé durable
Université Laval
2480 Chemin de la Canardière
Québec, QC, G1J 2G1
Canada
Phone: 1 (418) 663 5313 ext 12286
Email: France.Legare@fmed.ulaval.ca

Abstract

Background: Burnout and psychological distress are prevalent among physicians. Peer support appears to play a protective role, yet little is known about training interventions that motivate physicians to approach peers in difficulty, as such effects are often overlooked or assessed using nonvalidated tools.

Objective: We evaluated the effects of an online continuing professional development (CPD) course designed to increase physicians' intention to approach a colleague in difficulty.

Methods: Physicians who completed a 1-hour asynchronous online CPD course between March 2022 and May 2024 were invited to participate in this mixed methods convergent study. The e-learning course aimed to increase physicians' confidence in approaching colleagues in difficulty by recognizing signs of psychological distress, offering support, and referring them to appropriate resources. Participant characteristics were collected, and behavioral intention to approach a colleague in difficulty along with its determinants were measured pre- and postcourse using the validated CPD-REACTION tool. Differences in mean pre-post intention scores were assessed using 2-tailed paired *t* tests (n=466) and generalized estimating equations. Factors associated with postcourse intention were examined using multivariate analysis (n=466). Four months later, the proportion of physicians reporting adoption of the behavior was calculated (n=61). Qualitative responses to open-ended questions were analyzed thematically using behavior change models, and behavior change techniques used in the course were identified. Quantitative and qualitative results were triangulated. We reported results following STROBE (Strengthening the Reporting of Observational Studies in Epidemiology) and SRQR (Standards for Reporting Qualitative Research) guidelines for quantitative and qualitative analyses, respectively.

Results: Among 792 participating physicians, 466 (58.8%) completed online questionnaires pre- and postcourse. The average participant age was 48 (SD 12.4) years; 43.5% (332/762) were women, and 86% (655/762) were specialists. The average precourse intention score was 3.88 (SD 1.73) and average postcourse intention score was 4.92 (SD 1.40), for an adjusted mean difference of 1.06 (95% CI 0.93-1.20; P<.001). Factors associated with postcourse intention were beliefs about capabilities (β=0.52; P<.001), social influences (β=0.27; P<.001), and moral norm (β=0.26; P=.03; R^2=0.22). Four months later, 41% (25/61; 95% CI 28.6%-54.3%) of participants reported having approached a colleague in difficulty. Frequently reported reasons for intention to

XSL•FO
RenderX

adopt behavior were beliefs about capabilities, beliefs about consequences, and knowledge. Quantitative and qualitative results converged on beliefs about capabilities but diverged regarding beliefs about consequences. A total of 7 behavioral change techniques were identified in the CPD course: goal setting, increasing competence, planning, persuasive communication, behavior-related information, modeling, and behavioral experiments.

Conclusions: This online CPD course increased physicians' intention to approach a colleague in difficulty. The results highlight beliefs about capabilities as a key determinant of this behavioral intention. The study suggests that online learning has strong potential to raise awareness about peer support and ultimately build a culture of care among health care workers.

(JMIR Med Educ 2026;12:e80199) doi:10.2196/80199

KEYWORDS

burnout; continuing medical education; continuing professional development; distance education; helping behavior; occupational stress; peer support; physicians; psychological well-being; social support

Introduction

Burnout is prevalent among physicians [1-4]. In 2021, the Canadian Medical Association reported that half of Canadian physicians experienced at least 1 symptom of burnout on a weekly basis or more frequently [1]. Physicians experience greater social isolation compared to workers in other sectors [5]. Social isolation is linked to higher levels of burnout in physicians, but the direction of this relationship—whether it is a cause, a consequence, or bidirectional—remains unclear [5]. Furthermore, the "second victim" phenomenon, or health care providers' distress after a patient adverse event, can lead to burnout [6,7]. A patient adverse event is an unexpected but often preventable event which results in consequences for the patient of varying seriousness, including prolonging treatment, causing discomfort, disability, or death [7]. Moral distress was also associated with burnout in health care workers during the COVID-19 pandemic [8]. Moral distress is experienced when a health care worker "feels unable to act in accordance with core values and obligations, or attempted actions fail to achieve the desired outcome" [9]. Burnout impacts physicians' health and can have consequences on patient safety, satisfaction, and physician retention [10-12]. Improving physician well-being is important and the well-being of health care and social service workers is one element of the Quintuple Aim framework, whose goals are to improve population health, enhance care experiences, reduce costs, improve clinician well-being, and improve health equity [13].

A 2022 scoping review showed that peer support, informally or formally provided by health care providers with similar work conditions and experiences, can help health care providers after a distressing patient adverse event [14]. Furthermore, physicians themselves appear to find it the best form of support [15,16]. A 2024 trial found a peer support program protected health care providers aged 30 years or younger from psychological distress, although the effect was not significant for all age categories combined [17]. In another study, a peer support program reduced psychological distress among anesthesia professionals [18].

Components of a peer support conversation include initial outreach, listening, reflecting, reframing, eliciting the peer's personal coping mechanisms, closing the discussion, and referring the peer to available resources [19]. Experts recommend that peer supporters should receive training on listening skills, psychological first aid, and information about referral options to fulfill their role adequately [20]. Many formal peer-support programs have been implemented in health care organizations and include dedicated training courses for peer supporters [15,17,21-49]. Some of these training programs are inspired by (1) the Stress First Aid model, initially developed for high-risk occupations such as the military service [45,47-49]; (2) the psychological first aid training developed by George Everly [22,44]; (3) the peer-support approach developed by Shapiro and Galowitz [27,32,35,37,39] and (4) the Scott Three-Tier Intervention Model of peer support [15,21,23,28,30,41,42]. However, evaluation of the effect of courses about peer support is often omitted in program evaluations or assessed using cross-sectional surveys with nonvalidated measures [15,22,26,27,30,34,38-40,43]. Furthermore, most courses studied are in-person workshops rather than online continuing professional development (CPD) courses [15,22,26,27,30,34,38-40,43]. Therefore, we sought to assess the effect of an online CPD course on the intention of physicians to approach a colleague in difficulty using a mixed methods convergent research design and validated measures. We also analyzed why participants considered approaching a colleague in difficulty by conducting thematic analyses of qualitative data and triangulating these findings with quantitative results.

Methods

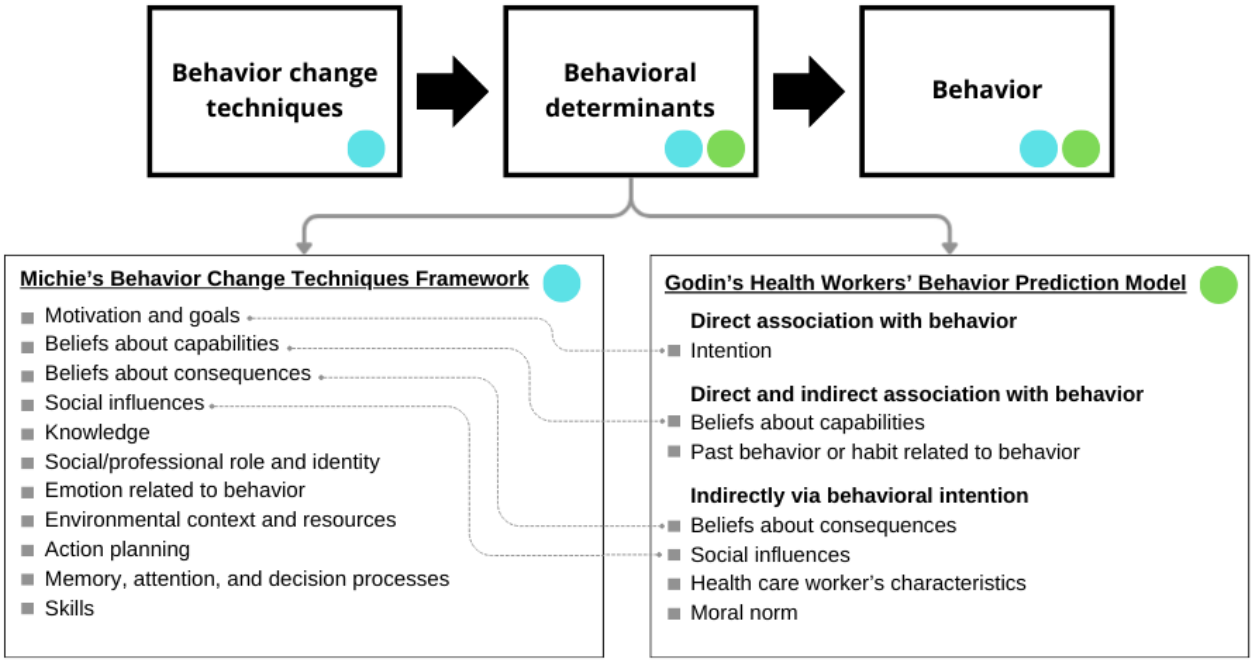
Study Design

We used a mixed methods convergent research design to triangulate results obtained from quantitative and qualitative methods [50]. Quantitative analysis involved a quasi-experimental before-and-after study design to evaluate the effect of an online CPD course that is part of the CPD program of the Fédération des médecins spécialistes du Québec (FMSQ). In this design, measurements before the CPD course served as comparators for the same measurements taken after the CPD course for the same participants [50]. We analyzed responses to open questions with a qualitative thematic deductive approach to understand why participants considered integrating the targeted behavior into their professional practice [50]. For both quantitative and qualitative analyses, we used Godin's integrated model for predicting health professionals' behavior as a conceptual framework [51] and we used Michie's conceptual framework on behavior change techniques (BCTs) to explore our qualitative data in detail [52]. As shown in Figure

XSL•FO
RenderX

1 [52,53], these frameworks share the concept that behavioral determinants lead to behavior adoption [51-53]. Both originate from Ajzen's Theory of Planned Behavior and have many behavioral determinants in common [51-53]. Michie et al [52] propose that BCTs can influence and target behavioral determinants, which in turn translate into behavior change. Possible behavioral change techniques include planning, persuasive communication and others [52]. Details on these 2 conceptual frameworks are available in Multimedia Appendix 1 [51-54]. We reported results following the STROBE (Strengthening the Reporting of Observational Studies in Epidemiology) guideline for quantitative results (Multimedia Appendix 2) [55] and used the SRQR (Standards for Reporting Qualitative Research) for qualitative results (Multimedia Appendix 3) [56].

Figure 1. Framework combining Godin's integrated model for predicting health professionals' behavior and Michie's framework on behavior change techniques [52,53].



Participants and Recruitment

The study took place in Quebec, Canada, from March 1, 2022, to June 20, 2024. Targeted participants were mainly specialists from the FMSQ membership. However, general practitioners who subscribe to the FMSQ's online learning platform MÉDUSE could also participate in the study. The study population was physicians who registered for the course and completed an online consent form. After registering, participants were informed about the study by a system-generated message displayed at the beginning of the online course. The CPD course was not mandatory but was recognized by the Royal College of Physicians and Surgeons of Canada as a 1-hour group learning credit (Section 1) [57]. This convenience sampling was based on time constraints and access to MÉDUSE. Inclusion criteria were being a licensed physician in Quebec and participating in the online CPD course. There were no exclusion criteria.

Intervention

We describe the intervention according to the Template for Intervention Description and Replication (TIDieR) guidelines [58]. The intervention was a 1-hour asynchronous online CPD course developed by the FMSQ to help physicians approach colleagues in difficulty (eg, burnout, psychological distress, second victim symptoms). Thus, this course focused on a form of peer support. The term "colleague" included anyone in the health and social services sector, including administrative staff. The scientific committee comprised 3 physicians, 2 of whom had peer support experience through their work with the Québec Physicians' Health Program (QPHP). The QPHP organization also helped with course design. A scenario-based video, filmed with professional actors, illustrates the QPHP's 5-Step Caring Approach: (1) becoming aware of context and stressors, (2) recognizing the warning signs, (3) initiating a conversation with the colleague, (4) providing compassionate support, and (5) respecting your own limits and practicing self-care [59]. Approaching a colleague in difficulty corresponds to Tier 1 of the Scott Three-Tier Model of Support, which includes immediate team members of a health care worker experiencing distress [21].

A CPD expert with more than 10 years of experience (MT) used Godin's integrated model for predicting health professionals' behavior [51] for instructional design to increase participants' willingness to approach a colleague in difficulty. For instance, the model was used to increase participants' beliefs in the consequences of approaching a colleague in difficulty. In the first module of the course, viewers are shown how a compassionate approach, which enables a colleague to share emotions related to stressful situations, can help mitigate the impact of these sources of stress. In the second module, physicians are shown the negative consequences that may occur if a colleague in difficulty is not helped or if help is delayed.

XSL•FO
RenderX

Participants completed an intervention plan during the CPD course to use when approaching a colleague at their workplace. Other educational strategies include narrated vignettes, knowledge-check questions, additional readings, and reflective writing, which often refers back to the scenario shown in the video. During the online course, physicians are also provided with a self-evaluation tool to identify when they or a colleague may need help. The tool provides individual strategies to maintain or improve mental health at each step along the continuum from being healthy to having important mental health problems [60].

During the course, participants have the option of downloading 10 summaries (ranging from 1 to 3 pages) of course content or supplementary materials [59-67]. Bloom's taxonomy defined the course's learning objectives [68]. The CanMEDS role targeted by this course was the professional role, specifically with the objective of emphasizing the commitment to physicians' health and well-being, which ultimately promotes optimal patient care [69]. The CanMEDS framework highlights skills that are essential throughout medical practice, providing guidance for medical education and ongoing professional development for practicing physicians, all within a competency-based approach [69]. The course was made available on MÉDUSE, the FMSQ's online CPD platform [70]. Physicians completing the course were sent an automated reminder after 2 weeks of inactivity, and the unfinished course remained available on the physicians' dashboard on MÉDUSE as a reminder to complete it.

Outcomes

The primary outcome was the intention of physicians to approach a colleague in difficulty. Godin defined intention, from the work of Ajzen and Fishbein on the Theory of Reasoned Action, as the motivation to adopt a specific behavior [51]. In this study, this targeted behavior was approaching a colleague in difficulty.

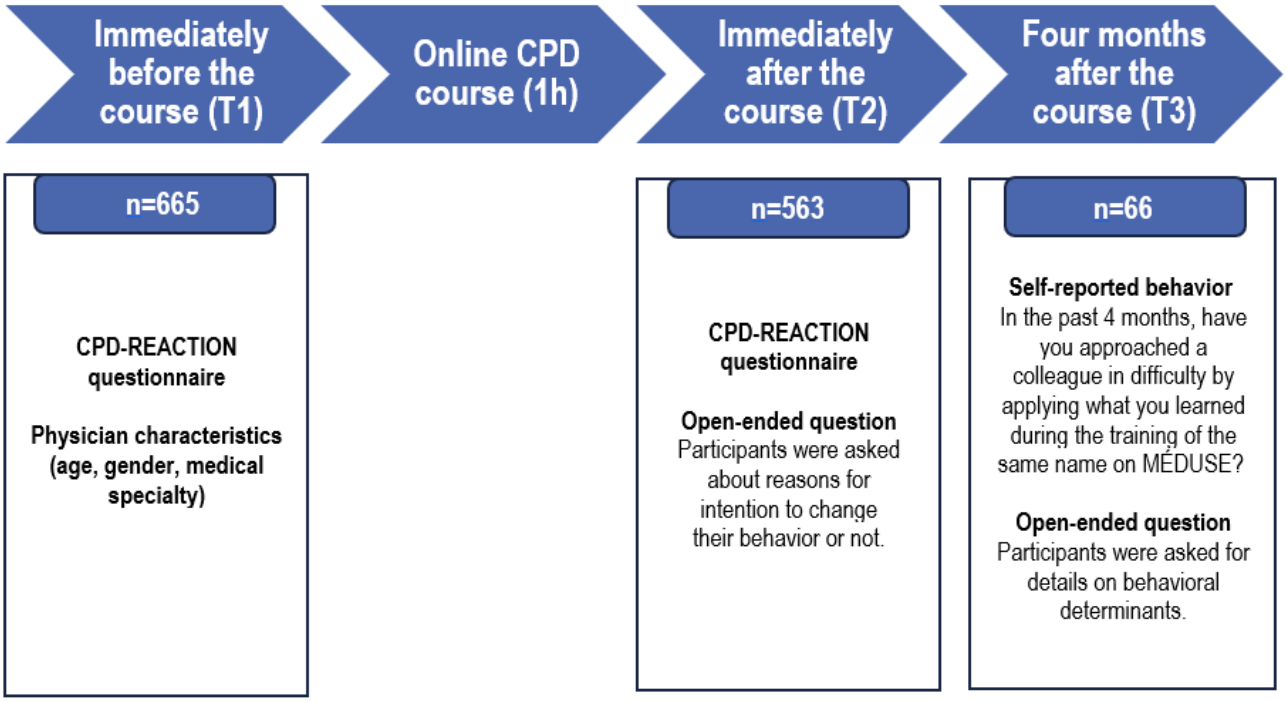
Intention to adopt a behavior is a precursor of behavior adoption in the integrated model for predicting health professionals' behavior [51]. Although intention is not a perfect proxy for future behavior, directly observing peer support behavior among health care workers is neither feasible nor ethical, as these conversations are sensitive and confidential [71]. Using self-reported behavioral intention as a proxy allowed for a practical and more timely evaluation of the CPD course [71]. Self-reported behavior 4 months after course completion was also measured but not used as a primary outcome because there is often a loss to follow-up when measures are not taken soon after course completion [72].

Data Sources and Measurements

Overview

The FMSQ collected physicians' responses on the MÉDUSE online learning platform from March 1, 2022, to June 20, 2024. Data were obtained from three sources: (1) quantitative data, (2) qualitative data, and (3) course content from the MÉDUSE platform (Figure 2). Quantitative data and qualitative data were collected using a questionnaire. The original version administered in French is available in Multimedia Appendix 4 [73,74] and the English translation is available in Multimedia Appendix 5 [73,74]. Physicians completed these online self-administered questionnaires at three timepoints: immediately before the course (T1), immediately after the course (T2), and 4 months after the course (T3) (Figure 2). Questionnaires at T1 and at T2 were administered by a system-generated message displayed at the beginning of the online course, and the questionnaire at T3 was sent automatically by email by the MÉDUSE online platform.

Figure 2. Data sources and collection. CPD: continuing professional development.



XSL•FO
RenderX

Quantitative Data Sources and Measurements

Before the CPD course, we collected data on the age, gender, and medical specialty of participants using an online questionnaire developed by the FMSQ. Before and immediately after the course (T1 and T2), we used the French version of CPD-REACTION to measure physicians' intention to approach a colleague in difficulty (dependent variable) and the determinants of this intention. Finally, we measured their self-reported behavior of approaching a colleague in difficulty 4 months after course completion (T3).

The CPD-REACTION is a measuring tool whose constructs were developed through a systematic review and whose items were validated using factor analysis [73]. Based on Godin's integrated model for predicting health professionals' behavior, the instrument measures intention to adopt a behavior and 4 determinants of intention [53,54,73]. It comprises 12 items grouped into 5 constructs: intention to adopt a behavior (2 items) and four determinants of intention: (1) moral norm (2 items), (2) beliefs about capabilities (3 items), (3) beliefs about consequences (2 items), and (4) social influence (3 items) [73]. Definitions of these determinants are available in Multimedia Appendix 1. Each construct is determined by averaging 2 or 3 items from the instrument [73]. The Cronbach α coefficients for each construct range from 0.79 to 0.89 [73], indicating good reliability [75].

Qualitative Data Sources

Overall, participants were asked 8 open-ended questions. Through 2 open-ended questions, physicians were asked about their reasons for intending to adopt the new behavior as well as factors influencing nonadoption of the behavior 4 months later. The remaining 6 questions, which assessed participants' appreciation of the CPD course, fall outside the scope of this work.

Course Content in the MÉDUSE platform

Two reviewers (F Lizotte and EL) were given access to all course content described in the intervention section above. The methodology for the analysis of this data is detailed in the section title "Identifying Behavioral Change Techniques Present in the Course."

Data Analysis

For both quantitative and qualitative analyses, we used Godin's integrated model for predicting health professionals' behavior as a conceptual framework [51]. In addition, we used Michie's conceptual framework on BCTs to analyze qualitative responses in detail as well as course content [52]. We analyzed all quantitative data with SAS software (version 9.4; SAS Institute Inc). We performed verification with nonparametric tests using R software (R Foundation for Statistical Computing). The NVivo 15 software (Lumivero) was used for qualitative analyses.

Participants' Intention and Its Determinants to Approach a Colleague in Difficulty, Pre-, and Postcourse

First, we performed descriptive statistics for all variables, including participant characteristics. Then, we calculated the difference between the pre- and postintention mean for each participant using 2-tailed paired *t* tests, assuming normality given our large sample size [76]. To confirm the robustness of the results, we also calculated the statistical significance of intention differences using a nonparametric Wilcoxon signed-rank test [77]. We did the same for each of the 4 determinants of intention to adopt the new behavior (social influences, beliefs about capabilities, beliefs about consequences, and moral norm). Finally, we verified that the results did not differ according to participants' survey response pattern or participant characteristics (gender, age, and domain of practice). We adopted the age categories used by the Canadian Medical Association in its survey on the wellness of Canadian physicians, as these categories reflect different stages of a medical career [78]. To describe effect size, we calculated Cohen d_z for within-subjects designs by dividing the mean of the within-person differences with the SD of those differences [79].

Estimation of the Effect of the CPD Course on the Intention to Approach a Colleague in Difficulty

We used generalized estimating equations (GEE) with a Gaussian distribution and identity link to estimate the marginal effect of the training program on intention scores over time (pre- vs postcourse [80]). The model included time (pre or post) as the main predictor, with an exchangeable working correlation structure to account for within-individual correlation, and was adjusted for age, sex, domain of practice, and baseline determinants of intention. We included age as a categorical variable since that variable did not respect the postulate of linearity.

Predicting Intention Postcourse

To predict intention to approach a colleague in difficulty after the course, we built a multiple linear regression model [81]. Behavioral intention was the dependent variable, while the 4 determinants of this intention and 3 sociodemographic characteristics (age, gender, and domain of practice) were independent variables. A full model including all candidate variables from Godin's framework was first fitted [51]. Using the backward elimination method, variables with a coefficient *P* value greater than .05 were successively eliminated for a more parsimonious model [82]. Sensitivity analyses were performed at critical α values of .10 or .20 to verify whether different choices changed the predictive performance of our model or the variables included in the final model. We presented the full and the fitted models.

Intention and its determinants were measured as aggregate multiple ordered responses of Likert items from 1 to 7 [73]. Thus, these composite scores approximate continuous latent variables and are commonly analyzed with linear models [83,84]. Simulation and methodological work have shown that treating such multi-item Likert scales (with ≥5 response options) as continuous yields valid estimates and tests in reasonably sized samples [83,84]. We checked for collinearity, normality of residuals, homoscedasticity, and linearity of residuals.

Power Analysis

Most participants were already recruited, so the sample size could not be changed. Therefore, we performed a post hoc power analysis. Based on 2 studies examining the effect of CPD courses on physicians' intentions to adopt new behaviors, we

XSL•FO
RenderX

should have been able to detect a mean difference of at least 0.5 in intention before vs after the course [72,85]. The statistical power of our paired *t* test for the primary outcome (mean difference in intention) was 99.9% at a significance level of α=.05 to detect this mean difference of at least 0.5. This indicates that the sample size used in this study was sufficiently large to detect a meaningful effect. We analyzed all quantitative data with SAS software (version 9.4; SAS Institute Inc). We performed verification with nonparametric tests using R software (R Foundation for Statistical Computing).

Physicians' Reasons for Intention to Adopt the New Behavior From Qualitative Data

Through 2 open-ended questions, physicians were asked about reasons for intention to adopt the new behavior as well as factors influencing nonadoption of the behavior 4 months later. The answers were analyzed using a thematic deductive approach [50]. We created an initial codebook by integrating 2 conceptual frameworks: Godin's integrated model for predicting health professionals' behavior [51] and Michie's BCTs with their respective behavioral domains [52]. Two researchers (F Lizotte and EL) organized 10% of the data into themes. Emergent themes were then discussed and added inductively to the framework. We documented these choices in a journal held by F Lizotte. The 2 reviewers (F Lizotte and EL) then coded data independently for greater credibility of the results and compared their analyses [50,86]. During qualitative analysis, F Lizotte, a medical resident, reflected critically about certain themes with EL, who has a background in psychology, and RCC with training in sociology, to improve the trustworthiness of the results. EL was completely blinded to the quantitative analysis process. To obtain consensus on results, any coding differences between the 2 reviewers (F Lizotte and EL) were resolved through discussion with RCC, whose main expertise was qualitative analysis. We described intercoder reliability of thematic analysis of open-ended questions using percentage agreement [87].

Triangulating Quantitative and Qualitative Data

We triangulated quantitative and qualitative data to enhance the credibility and validity of the research findings and offer practical theory-based recommendations for CPD course development on peer support [50,88]. We compared 4 determinants of intention, measured with CPD-REACTION and obtained using quantitative methods, with the reasons to adopt the new behavior obtained using qualitative methods [50,88]. We observed where results converged, diverged, or provided additional information [88].

Self-Reported Behavior Adoption 4 Months After CPD Course

We calculated the proportion of physicians who self-reported approaching a colleague in difficulty 4 months after the online CPD course. We compared the mean intention of those who adopted the behavior with those who did not using Wilcoxon rank tests because of the small sample sizes of the data at T3 and verified the results using paired *t* tests.

Identifying Behavioral Change Techniques Present in the Course

We developed an observation grid (Multimedia Appendix 6 [51,89,90]) from the conceptual framework by Michie et al [52]. Using this grid, 2 independent reviewers (F Lizotte and EL) identified BCTs present in the online CPD course. We analyzed whether the behavioral domains targeted by the BCTs identified in the course were the same as the determinants of intention that had significantly changed after the CPD course. This analysis explored which components of the course might be the active ingredients in changing physicians' behavioral intentions. Using the entirety of the interrater data, we calculated Cohen kappa to describe interrater reliability for BCT identification [91]. We chose Cohen kappa as the interrater reliability statistic, as it has the advantage of taking into account agreement by chance and can be applied to categorial nominal variables of the observation grid [87].

Handling of Missing Data

Item-level missing data were handled by following the CPD-REACTION Questionnaire User Manuel [74]. A nonresponse to 1 of the 2 items measuring a construct of the CPD-REACTION resulted in missing data for the observation of the measured variable [74]. A missing response to an item in the 3-item measure of the determinants of intention was handled with mean imputation [74]. Only 0.25% of responses to items on the CPD-REACTION questionnaire administered at T1 were missing, with item-level missingness ranging from 0.00% to 0.90% (Multimedia Appendix 7). Similarly, approximately 0.27% of responses were missing for items on the CPD-REACTION questionnaire administered at T2, with a range of 0.00% to 1.07% depending on the item (Multimedia Appendix 7). There were no missing sociodemographic characteristics of physicians who completed the CPD-REACTION questionnaire before and after the CPD course.

For wave-level missing data, the 466 participants with data at T1 and T2 were included in the analysis (ie, perfect sets). Sensitivity analyses were conducted to verify if results differed if all data, regardless of data pairing, were analyzed. A complete case analysis was performed in all models.

Ethical Considerations

We obtained ethical exemption from a hospital research ethics committee (Comité d'éthique de la recherche sectorielle en santé des populations et première ligne—Centre intégré universitaire de santé et de services sociaux de la Capitale-Nationale [CIUSSSCN]; project #2025-3127). The research project participants had given their free and informed consent to take part in the study. Before each questionnaire, a consent form detailed the objective and intended use of the collected data for research (Multimedia Appendix 4 contains the French version and Multimedia Appendix 5 contains the English version). The database was deidentified to protect the confidentiality of participants. A password was required to access the data saved in an Excel file, which has been stored on a secure server. Data analysis was carried out on a CIUSSSCN computer that was also password-protected. Participants experiencing emotional

XSL•FO
RenderX

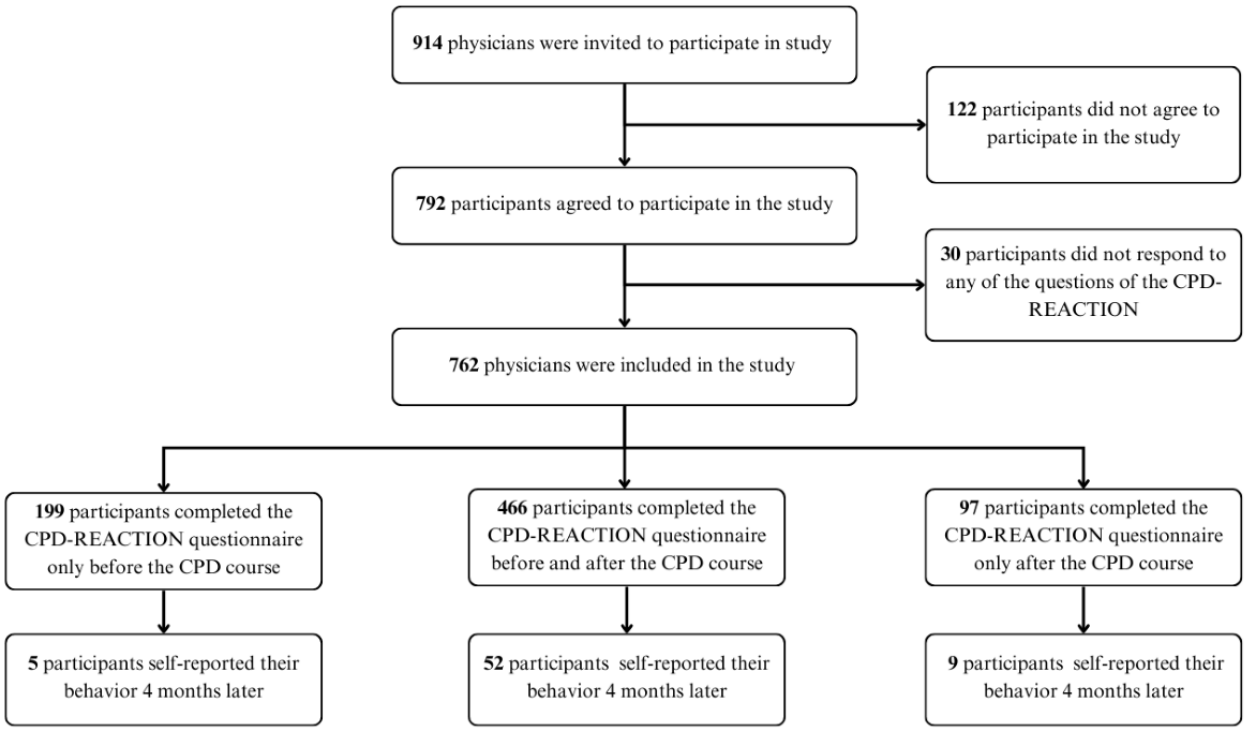
difficulties could speak with peer supporters from the QPHP. Physicians volunteered to participate and received no financial compensation.

Results

Participant Flow and Characteristics

During the 28-month study period, 914 physicians were invited, of whom 792 (87%) agreed to participate (Figure 3). A total of 30 participants were excluded because they did not answer a single CPD-REACTION questionnaire item. We performed analyses on 61% (466/762) of participants who completed the CPD-REACTION questionnaires before and after the course (T1 and T2). Finally, the frequency of self-reported behavior was calculated on 8% (61/762) of participants who completed both the self-reported behavior change questionnaire 4 months after the course (T3) and the CPD-REACTION questionnaire immediately after the CPD course (T2). We excluded 5 participants who did not complete the postcourse survey (T2) from the analysis of data at T3, as it was only at T2 that we could verify their completion of the course.

Figure 3. Study flowchart. CPD: continuing professional development.



Among the participants, the woman-to-man ratio was approximately 1:1, and the majority of physicians (655/762, or 86%) were specialists (Table 1). The average age of participants was 48 (SD 12.4) years, 43.5% (332/762) were women, and 86% (655/762) practiced in a specialty. Physicians who responded both before and after the course (n=466) had similar sociodemographic characteristics. We did not have sociodemographic characteristics for the 97 physicians who responded only to the T2 questionnaire without responding to the T1 questionnaire.

XSL•FO
RenderX

Table 1. Characteristics of participants.

Sociodemographic characteristics of physicians	Value (n=762)	Participants who completed the CPD-REACTION questionnaire before and after the CPD[a] course (n=466)
Gender, n (%)		
Woman	332 (43.5)	241 (51.7)
Man	325 (42.7)	222 (47.6)
Other, nonbinary, or prefer not to disclose	8 (1.1)	3 (0.7)
Missing	97 (12.7)	0 (0)
Domain of practice, n (%)		
Surgical specialty	293 (38.5)	208 (44.7)
Laboratory specialty	102 (13.4)	71 (15.2)
Medical specialty	246 (32.3)	169 (36.3)
General practitioner	10 (1.3)	9 (1.9)
Nonspecified specialty	14 (1.8)	9 (1.9)
Missing	97 (12.7)	0 (0)
Age (in years), n; mean (SD)	665; 48.1 (12.4)	466; 48.6 (12.3)

[a]CPD: continuing professional development.

Participants' Intention (and its Determinants) to Approach a Colleague in Difficulty, Pre- and Postcourse

Intention means increased from 3.88 (SD 1.73) before to 4.92 (SD 1.40) after the course, as measured on a scale of 1 to 7. This mean difference increase of 1.04 (P<.001) is statistically significant (Table 2). This 1-point increase was also clinically significant, with a Cohen d_z value of 0.72 representing a moderate effect size. Bivariate analyses of changes in the 4 determinants of intention (social influences, beliefs about capabilities, beliefs about consequences, and moral norm) were all statistically significant at an α level of .05. Determinants of intention (beliefs about capabilities, social influences, and beliefs about consequences) increased by statistically significant mean differences ranging from 0.14 to 0.85. Moral norm decreased by a statistically significant mean difference of 0.30 (P<.001). Lower extreme values after the CPD course did not explain the decrease in moral norm. The effect size on the increase in beliefs in capacities and on the increase in social influences was moderate. The effect size on the decrease in moral norm was small, and the effect size on the increase in beliefs about consequences was not clinically significant (Table 2). Mean intention and determinants of intention were similar whether participants had responded to both questionnaires (T1 and T2) or to only 1 questionnaire (Table S2 in Multimedia Appendix 8 [73,74]).

Table 2. Comparison of behavioral intention before and after the continuing professional development (CPD) course.

CPD-REACTION constructs[a]	Frequency, n	Before CPD course, mean (SD)	After CPD course, mean (SD)	Before vs after CPD course			
				Mean difference (CSD[b])	95% CI	P value[c]	Cohen d_z[d]
Intention	455	3.88 (1.72)	4.92 (1.40)	1.04 (1.45)	0.91 to 1.17	<.001	.72
Beliefs about capabilities	459	4.65 (1.17)	5.49 (0.74)	0.85 (1.11)	0.74 to 0.95	<.001	.76
Moral norm	462	6.09 (1.04)	5.79 (0.77)	–0.30 (1.06)	–0.40 to –0.21	<.001	.29
Social influences	464	3.36 (0.99)	3.81 (0.99)	0.45 (0.74)	0.38 to 0.52	<.001	.61
Beliefs about consequences	465	5.53 (1.04)	5.67 (0.65)	0.14 (0.90)	0.06 to 0.22	<.001	.16

[a]Response scale of the CPD-REACTION questionnaire ranged from 1 to 7 [73,74].

[b]CSD: combined SD.

[c]P values were calculated using 2-tailed paired t test.

[d]Cohen d_z estimates effect size (<0.2=not significant, 0.2-0.5=small, 0.5 to 0.8 medium, >0.8=large) [91].

XSL•FO
RenderX

Subgroup Analyses of Principal Outcome: Difference of Intention to Approach a Colleague in Difficulty

The increase in the mean difference of intention was approximately 1 point and was statistically significant for all categories of gender, age, and domain of practice (laboratory, surgical, and medical domains) that included at least 10 participants (Multimedia Appendix 9). We did not interpret subcategories containing fewer than 10 participants (eg, general practitioners). Results of the multiple linear regression model confirmed that the intention difference was not modified by gender, age, or domain of practice when all variables were included in the same model (Multimedia Appendix 9).

Effect of the CPD Course on the Intention to Approach a Colleague in Difficulty

In the GEE model, participation in the CPD course was associated with a higher intention to approach a colleague in difficulty. On average, intention scores were 1.06 points higher in participants after the CPD course compared to before the CPD course (β=1.06; *P*<.001), after adjustment for determinants and for sociodemographic characteristics (Table S1 in Multimedia Appendix 8). Beliefs about capabilities (β=0.37; *P*<.001), social influence (β=0.22; *P*=.002), and beliefs about consequences (β=0.15; *P*=.03) before the CPD course showed significant positive associations with mean intention, whereas moral norm was not significant. No sociodemographic characteristics were independently associated with intention to approach a colleague in difficulty (Table S1 in Multimedia Appendix 8).

Postcourse Determinants of Intention

In the final multivariate regression analysis model, beliefs about capabilities (β=0.52; *P*<.001), social influences (β=0.27; *P*<.001), and moral norm (β=0.26; *P*<.03) predicted physicians' intention to approach a colleague in difficulty (Table 3). Our model accounted for 22% of the variance of intention after the course. Sensitivity analyses verified the stability of the other variables and showed consistent results if we also included physicians who had not responded to the precourse questionnaire (Multimedia Appendix 10), except for the variable belief in consequences, whose statistical significance changed when we modelized physicians who had not responded to the precourse questionnaire.

Table 3. Predictive factors of physicians' intention to approach a colleague experiencing difficulties.

Variables[a]	Full multivariate model[b] (n=457)		Fitted multivariate model[c] (n=457)	
	β (95% CI)	*P* value	β (95% CI)	*P* value
Beliefs about capabilities	0.39 (0.12 to 0.67)	.005	0.52 (0.28 to 0.75)	<.001
Social influence	0.25 (0.12 to 0.37)	<.001	0.27 (0.15 to 0.39)	<.001
Moral norm	0.26 (0.03 to 0.49)	.02	0.26 (0.03 to 0.48)	.03
Beliefs about consequences	0.22 (–0.01 to 0.45)	.06	N/A[d]	N/A
Age (in years; reference: 35-54)		.39[e]		
<35	–0.22 (–0.56 to 0.13)	.21	N/A	N/A
>54	–0.11 (–0.37 to 0.14)	.38	N/A	N/A
Domain (reference: surgical)		.25[e]		
Laboratory	0.23 (–0.12 to 0.57)	.20	N/A	N/A
Medical	0.29 (0.03 to 0.55)	.03	N/A	N/A
Family medicine	0.39 (–0.45 to 1.23)	.36	N/A	N/A
Other	0.30 (–0.54 to 1.13)	.48	N/A	N/A
Gender (reference: women)		.54[e]		
Men	–0.12 (–0.37 to 0.12)	.33	N/A	N/A
Other, nonbinary, or prefer not to disclose	–0.44 (–1.86 to 0.97)	.54	N/A	N/A

[a]We included psychosocial variables measured after the course.

[b]R^2=0.24.

[c]R^2=0.22.

[d]N/A: not applicable, as variables with a coefficient *P* value greater than .05 were successively removed using the backward elimination method.

[e]Global *P* value.

Postulates of independence of observations and homoscedasticity were respected. Postulates of normality and linearity of residuals were not respected. We verified the correlations between the independent variables and found they did not explain our model's low predictive capacity (Multimedia Appendix 10).

XSL•FO
RenderX

Physicians' Reasons for Intention to Adopt the New Behavior From Qualitative Data

Of the 251 open-ended responses related to behavioral intentions (T2), 153 verbatims expressed clear behavioral intentions (eg, "I will help my peer"), 10 showed hesitation (eg, "I will try to"), 89 described reasons for intention to approach a colleague in difficulty (eg, "I realize more fully my ability to act"), and 8 referred to other related behaviors. The above categories are not mutually exclusive. Physicians most frequently intended to (1) be more attentive to their colleagues' signs of distress, (2) approach a colleague with kindness or help or support a colleague, and (3) be more available or listen more to colleagues (Multimedia Appendix 11). The reasons for intention to change behavior physicians most frequently cited were beliefs about capabilities, beliefs about consequences, and knowledge of the behavior (Table 4). For each of the 10 categories of reasons for wanting to adopt the new behavior, intercoder percentage agreement on thematic analysis varied between 98.1% and 99.5%.

Table 4. Reasons for intention to adopt new behavior of approaching a colleague in difficulty.

Reasons for intention to adopt new behavior	Representative excerpts of verbatim	Agreement, n (%)
Beliefs about capabilities	• "I realize more fully my ability to act." • "I have more tools to approach a colleague in difficulty." • "I feel more comfortable doing it."	41 (98.4)
Beliefs about consequences	• "This limits the consequences for the colleague and the team." • "To improve the collegiality of my department." • "I think that the current context of healthcare depends on the support we provide to our colleagues, so it is essential to do it."	16 (99.2)
Knowledge	• "Better knowledge of how to detect a colleague's distress and how to intervene in such cases." • "Better understanding of the approach." • "I will better assess where my colleague stands on the distress scale."	11 (96)
Social/professional role and identity	• "Necessary with my hospital duties." • "Feeling more legitimized to approach a colleague in difficulty." • "Better understanding and responsibility of collegiality."	6 (98.5)
Moral norm	• "I like the idea that doing nothing is not an option." • "Not acting is not an option." • "…understanding that some approaches may be better than others, but anything is better than doing nothing."	5 (97.9)
Past behavior or habit related to behavior	• "It has provided me with new tools to continue supporting colleagues in difficulty." • "I already apply many similar principles." • "Continue to approach my colleague when they seem to be having difficulties."	4 (99)
Emotion related to behavior	• "Being more willing to get uncomfortable." • "It reassures me." • "I feel more equipped and reassured about my reflections."	3 (99.5)
Environmental context and resources	• "Colleague to help." • "Frequent problem."	2 (98.9)
Action planning	• "Do not procrastinate." • "Do not hesitate to create opportunities outside of work."	2 (98.1)
Memory, attention, and decision processes	• "Because I realize that often, I get too absorbed in my professional or personal concerns to look around me. Being a witness allows for perspective and a new outlook."	1 (99)

The reasons physicians gave for not wanting to change their practice were mostly related to their past behavior, for example, that they were already helping a peer and thus had already adopted the behavior (Multimedia Appendix 11). Other reasons for not intending to change behavior were their environmental context (eg, "I'm now in solo practice") and that helping a peer was not relevant to their particular medical practice. Four months after the CPD course, the behavioral determinants most frequently given for not approaching a colleague in difficulty were having had no occasion to adopt the behavior, having no colleagues in difficulty, and not being actively in practice or not being in full-time practice (Multimedia Appendix 11).

Triangulated Quantitative and Qualitative Data

Qualitative and quantitative results showed most physicians were motivated to approach a colleague in difficulty following

XSL•FO
RenderX

the online CPD course. Both qualitative and quantitative results pointed to beliefs about capabilities as being an important determinant of intention. Beliefs about consequences was an important theme in the qualitative data, but it was not an important determinant of postcourse intention to approach a colleague in difficulty. Sensitivity analysis of the 251 physicians who provided open-ended responses related to behavioral intentions (T2) also showed a divergence regarding beliefs in consequences between qualitative and quantitative results.

Self-Reported Behavior Adoption 4 Months After CPD Course

Among the 66 physicians who completed the self-administered questionnaire at T3, 61 had completed CPD-REACTION immediately after the course. Of these, 41% (25/61; 95% CI 28.6%-54.3%) reported approaching a colleague in difficulty in the 4 months following the CPD course. Those who adopted the behavior had a mean intention after the CPD course of 5.40 (SD 1.01), and those who did not adopt the behavior had a mean intention of 5.00 (SD 1.40). This 0.4 (SD 1.26; *P*=.24) mean behavioral intention difference is not statistically significant.

Behavioral Change Techniques Identified as Present in the Course

A total of 7 BCTs were found in the online CPD course: goal setting, increasing competence, planning, persuasive communication, behavior-related information, modeling, and behavioral experiments (Table 5). Cohen kappa measuring interrater reliability was 0.53 (95% CI 0.14-0.91) for evaluation of BCTs identified the course, which represents a moderate strength of agreement [92]. Of these 7 BCTs, 4 were related to behavioral intentions, 3 involved competence, 2 involved behavioral regulation, and 2 were related to beliefs about consequences (Multimedia Appendix 11). Beliefs about capabilities, knowledge, social influences, and memory, attention, and decision processes were each targeted by 1 BCT.

Table 5. Behavioral change techniques (BCTs) identified in the continuing professional development (CPD) course.

BCT	Definition	Domains targeted
Goal setting	Clearly indicate the targeted behavior.	Competence, behavioral intention, behavioral regulation
Increasing competence	Enhance skills through problem-solving, decision-making, and goal setting.	Competence, beliefs about capabilities, behavioral intention
Action planning	Identify the components of the behavior and establish a plan to execute each component, including when and where the behavior will be performed (ie, scheduling behaviors).	Memory, attention, and decision processes, behavioral regulation
Persuasive communication	A credible source presents arguments in favor of the behavior. There must be a presentation of arguments; general pro-behavior communication does not qualify.	Beliefs about consequences, behavioral intention
Behavior-related information	Provide information about the antecedents or consequences of the behavior, the connections between them, or relevant BCTs.	Knowledge, beliefs about consequences, behavioral intention
Modeling	Observe the behavior of others as a model for action.	Competence, social influences
Behavioral experiments	Ask individuals to test hypotheses about the behavior, its causes, and its consequences by collecting and interpreting data, including personal experiences.	No consensus on targeted domains

Discussion

Principal Findings

We assessed the effect of an online CPD course on physicians' intention to approach a colleague in difficulty and analyzed why participants considered integrating this peer support approach or not into their practice. Mean behavioral intention increased significantly after the course. Postcourse, significant determinants of intention were social influences, beliefs about capabilities, and moral norm. Most frequently stated reasons for intention to change behavior were beliefs about capabilities, beliefs about consequences, and knowledge of behavior. Triangulation showed an alignment of quantitative and qualitative results on beliefs about capabilities as a key factor but diverged on beliefs about consequences. In the 4 months following the CPD course, 41% (25/61) of physicians reported approaching a colleague in difficulty. There was no significant difference in the intention to approach a colleague in difficulty between physicians who adopted this behavior at the 4-month follow-up and those who did not. Those who did adopt the behavior had a nonstatistically significant higher intention after the CPD course than those who did not. Finally, we identified 7 BCTs in the online CPD course: goal setting, increasing competence, planning, persuasive communication, behavior-related information, modeling, and behavioral experiments. These results lead us to make the following observations.

Comparison of Participants' Behavioral Intention Pre- and Postcourse With Prior Work

First, we found that, on average, physicians reported a stronger intention to approach a colleague in difficulty after completing the CPD course compared to before. This 1-point intention increase was statistically significant and also clinically significant, having a moderate effect size. This increase was observed in bivariate analysis and confirmed with multivariate analysis. This suggests that the CPD course motivated physicians to approach colleagues in difficulty. Enhancing peer support among physicians through CPD courses is particularly important

XSL•FO
RenderX

as it can play a protective role regarding psychological distress in physicians [17,18].

Also, to the best of our knowledge, this 1-point intention increase represents a greater increase in intention than has been observed in most other studies using the CPD-REACTION to evaluate changes in health professionals' intentions following an intervention [72,85,93-102]. The before-after difference of intention was more than double the difference observed in a study by Ayivi-Vinz et al [85] who evaluated intention before and after 9 online CPD courses taken by Quebec physicians, all on topics other than this one. The 95% CIs of difference in intention between their study and ours do not overlap [85]. Thus, the observed difference between the 2 studies is statistically significant. An important difference is that Godin's integrated model for predicting health professionals' behavior was used by the senior instructional designer (MT) to design the course, while the 9 courses in the study by Ayivi-Vinz et al [85] were developed by physicians who may not have used behavior change models in course design. This may give support to the importance of using theory to produce CPD courses. Another explanation is that preintervention intention among the Quebec physicians in the study by Ayivi-Vinz et al [85] was already high at 5.5 out of 7, while in ours it was 3.9 out of 7, and thus there was more room for improvement. Overall, this result suggests that there are benefits of using behavior change theories, especially for CPD courses targeting behaviors that physicians are initially less inclined to adopt. It also suggests that it is possible to increase behavioral intention in physicians using online training formats.

Change in Determinants of the Intention to Approach a Colleague in Difficulty

Social influences, beliefs about capabilities, and beliefs about consequences increased. An unexpected postcourse decrease in moral norm occurred. Moral norm had the highest mean among the 4 determinants of intention both before and after the intervention, indicating that participants already felt a moral obligation to adopt the targeted behavior and thus there was little room for improvement. Also, the apparent decrease had a small effect size, which may not translate into clinical significance. One possible explanation for this unexpected decrease in moral norm is that participants may have given more nuanced answers after the CPD course as they better understood the ethical and deontological implications of approaching a colleague in difficulty. In the precourse questionnaire, 38.3% (177/462) of the physicians answered with the maximum score of 7/7 for moral norm, compared to none after the intervention. Nevertheless, further qualitative data analysis is needed to explore this result. Analyzing which BCTs target moral norm could also be important, as this is not studied by Michie et al [52].

No Modifying Effect of Age, Gender, or Domain of Practice on Primary Outcome

Second, we also examined whether participant characteristics modified the impact of the course on their behavioral intention. The before–after intention difference was the same regardless of age, gender, and domain of practice (laboratory, surgical, and medical domains). Thus, the course successfully reached physicians regardless of their sociodemographic characteristics and does not need to be adapted for a specific population. For example, a previous study found that surgeons were the least likely physicians to use peer support [34], hence the importance of our finding that surgeons also responded favorably to the online CPD course.

Comparison of Determinants of Intention Pre- and Postcourse With Prior Work

Third, we found that determinants of intention to approach a colleague in difficulty after the CPD course included social influences, beliefs about capabilities, and moral norm. Beliefs about capabilities was the strongest predictive variable of the intention to approach a colleague in difficulty. This means improving physicians' perception of their self-efficacy in adopting a behavior, as well as their perceived control over performing that behavior, is important to motivate them to approach a colleague in difficulty. Thus, CPD course developers should target beliefs about capabilities through BCTs for peer support training. Beliefs about capabilities was also the most significant determinant of intention in Godin's systematic review on health care workers' behaviors [53]. BCTs that are known to be effective for this include: (1) increasing skills, (2) coping strategies, (3) repetition or practice of the behavior, (4) feedback, (5) planned self-affirmations, (6) motivational interviewing, and (7) graded tasks (starting with tasks that are easy to accomplish and gradually making them more difficult). In previous studies, beliefs about capabilities, moral norm, and beliefs about consequences were all predictive of intention to adopt behaviors related to patient care (eg, "Prescribe opioids in a safe manner when managing perioperative pain in my patients") postcourse [72,85]. Our results differ slightly from these, with social influence also being predictive of this intention. Approaching a colleague in difficulty is a social behavior that can be adopted by a whole departmental team [38]. Our results suggest that the social context may be a more important influence on team-based behaviors than on clinical behaviors, which are associated more with individual clinicians.

Comparing Predictive Capacity Postcourse With Prior Work

Our model had a low predictive capacity (R^2=0.22), explaining only 22% of the variance in intention. In comparison, previous literature showed that the predictive capacity of models analyzing 1 single behavior varied between 19% and 62%, depending on the type of behavior [53]. Bakwa et al's [72] study of multiple behaviors related to patient care among Quebec physicians explained 82% of the variance in physicians' intention to adopt the clinical behavior. However, Bakwa et al [72] included multiple behaviors in the same model, while we studied only 1 behavior, and moreover one that is specific to the well-being of physicians rather than geared to a clinical task. Indeed, existing studies mainly predict behaviors related to patient care [53,72]. While variables of Godin's integrated model for predicting health professionals' behavior are sufficient to predict such behaviors [51], approaching a colleague is not a clinical behavior and may be influenced by factors not measured by Godin's model such as lack of time, lack of privacy, and organizational culture [36,103]. This is significant

XSL•FO
RenderX

because the Quintuple Aim—adopted by many health care systems—now emphasizes 5 key outcomes, and not all of them clinical: patient outcomes, patient experience, efficiency, clinician well-being, and equity [13]. Therefore, further development is needed regarding fostering behaviors related to clinician well-being.

Triangulated Quantitative and Qualitative Data

Fourth, triangulating quantitative with qualitative results confirmed that beliefs about capabilities was a key determinant of intention to approach a colleague in difficulty. Interestingly, results regarding beliefs about consequences diverged slightly: while it was the second most frequent reason for intention to change behavior stated in physicians' open-ended responses, it was not a statistically significant determinant of intention in our quantitative results. One explanation is that statistical significance varied according to the other variables included in the model, perhaps because the *P* value for beliefs about consequences was so close to the critical α value of .05. Another possible explanation is that beliefs about consequences, as described by Godin [51], seems to refer to the consequences for oneself. Items for measuring beliefs about consequences quantitatively were worded as follows "Overall, I think that *for me* approaching a colleague in difficulty would be...." [73]. In contrast, in our qualitative analysis, beliefs about the consequences of adopting the behavior could also refer to its consequences for the colleague being approached, the team, the department, or for society (Table 5). Indeed, many physicians taking part in this study perceived that the consequences of approaching a colleague in difficulty would be beneficial for the team and the department, a theme often mentioned in the open-ended responses. Thus the perception of the consequences of team-focused care may include organizational benefits, team benefits, patient benefits, and benefits to individual team members [104]. Also, approaching a colleague in difficulty is a team-focused behavior that takes place between 2 health care workers, while behaviors related to patient care occur between a clinician and a patient (eg, prescribing medication). We hypothesize that such team-focused behaviors may depend on a wider variety of determinants, such as organizational culture, than do individual behaviors related to patient care [36,104,105]. Whether peer support strategies can be applied without fear of judgment or retaliation within the organization may depend on its climate of psychosocial safety, or the perception among employees that their psychological health is a priority for the management [106,107]. A study by Tolin et al [38] reported that physicians' perception of institutional support had a positive impact on implementation of a physician-focused peer support program. Broadening the item statements on beliefs about consequences to include effects on others and adding questions with regards to role and identity may be a good starting point for improving measuring tools regarding team-based behaviors.

Other determinants to measure in future studies should include social and professional role or identity and past behavior or habits related to the behavior, both mentioned by participants as determinants of intention and present in Godin's integrated model for predicting health professionals' behavior (Table 4) [51]. Our qualitative analysis identified the following further potential determinants: (1) knowledge; (2) action planning; (3) memory, attention, and decision processes; (4) environmental context and resources; and (5) emotion related to behavior. The first 3 of these determinants are also behavioral domains targeted by the BCTs we identified in the CPD course.

Self-Reported Behavior Adoption 4 Months After CPD Course

Fifth, 4 of 10 physicians reported having approached a colleague in difficulty in the 4 months following the CPD course, and those who did so had a higher intention to adopt the behavior after the CPD course. No statistically significant difference in intention to approach a colleague in difficulty was found between those who did and those who did not adopt the behavior (5.4 vs 5.0; *P*=.24). This unexpected finding is possibly due to the limited statistical power of the analysis, estimated at 23%. Also, environmental factors, such as practicing solo, were the barriers most mentioned by physicians. Some of these physicians who might have been willing to approach a colleague in difficulty were simply not in the right environment to be able to do so. Other studies report on peer support courses that are part of comprehensive peer support programs that include an organization-wide referral system [15,22,26,27,30,34,38-40,43]. Thus, a physician could be the only physician in a department and still participate in a peer support program thanks to the referral system. It would be valuable to further evaluate whether formalizing peer support through these programs helps overcome certain environmental barriers or, conversely, introduces new barriers, compared to informal peer support outside of a program, as occurred in our study.

Behavioral Change Techniques Present in the Course

Lastly, we identified 7 BCTs in the online CPD course. This is the first time that BCTs have been identified in an FMSQ course and triangulated with before-and-after measures of CPD-REACTION. The results suggest that the BCTs effectively targeted determinants of intention, such as beliefs about capabilities and social influence, which had changed after the course. However, this study, which analyzed only 1 course, cannot make recommendations on which BCTs to use for which behaviors. Nevertheless, our results offer significant methodological progress in evaluating CPD course content, which could lead to innovations in CPD course development. Further studies on the use of theory-informed BCTs in CPD course design to enhance clinician adoption of targeted behaviors could further our understanding of effective course development strategies.

Limitations

First, the quasi-experimental before-and-after study design is vulnerable to confounding concomitant events that could influence study outcomes because the design comprises only a single group [50]. However, the short duration of the online CPD course (1 hour) made it less likely for other events to introduce bias. We chose this before-and-after study design as it is better suited to the constraints of practice settings than randomized control trials and better reflects the real-world CPD context [50]. Furthermore, we confirmed bivariate analysis with the GEE model to better measure the effect of the CPD course.

XSL•FO
RenderX

Triangulation of quantitative with qualitative methods improved the internal validity of the study [50,108].

Second, our results mainly apply to specialist physicians who chose to follow a nonmandatory CPD course and thus had some interest in the subject. The effect of the CPD course could have differed if it were mandatory for all physicians, as seen in other education programs [109]. Furthermore, general practitioners were underrepresented in this study (n=9). Thus, we do not know how general practitioners would respond to the CPD course but have no reason to think that they would behave differently than specialists. In addition, the well-being of all health care professionals is paramount, and it is not clear if the online CPD course would have any impact on other types of health professionals.

Third, there was a possible social desirability bias whereby physicians overrated their intention to approach a colleague in difficulty [50]. Confidentiality and anonymity were ensured to limit this bias. Of note, we measured intention using the same instrument used by Bakwa et al [72] and by Ayivi-Vinz et al [85] in their studies, so the desirability bias would have applied to their results as well and does not explain the higher before-and-after intention difference we observed in this study.

Fourth, this study does not allow us to confirm the reasons for which results regarding beliefs about consequences diverged, reasons to change behavior that were prominent in qualitative results but unstable or nonsignificant in the multiple linear regression. Although hypotheses are detailed in the Discussion, further studies would be needed to validate them.

Fifth, another limitation is that we used a backward elimination procedure, which can increase the risk of overfitting and biased effect estimates. However, results in the full model and fitted model were similar.

Sixth, we lacked statistical power to measure the difference in intention to approach a colleague in difficulty between those who did and those who did not adopt the behavior.

Lastly, self-reporting of behavior may also overestimate behavior adoption, as participants may not have used the approach taught in the course and may still report having adopted the behavior. Furthermore, only a small proportion of participants completed the self-reported behavior questionnaire 4 months after the course. This may indicate a selection bias in favor of the most motivated physicians, which could overestimate the proportion of those adopting the behavior. The sensitivity analysis in Multimedia Appendix 8 shows that this potential selection bias on number of questionnaires answered does not affect the primary outcome (mean difference in intention to approach a colleague in difficulty) but only the secondary outcome of self-reported behavior 4 months after the training.

Conclusions

To the best of our knowledge, this is the first study that uses validated before-and-after intervention measures to examine the effect of a CPD course on physicians' intention to approach a colleague in difficulty (experiencing distress, burnout, or second-victim symptoms after a patient adverse event). The CPD course produced an increase in physicians' intention regardless of age, gender, or domain of practice (laboratory, surgical, and medical domains). Beliefs about capabilities was an important determinant of intention to approach a colleague in difficulty. Although the CPD course increased physicians' intention to approach a colleague in difficulty, there is still a gap in understanding what other factors predict this intention. Within 4 months of the CPD course, 41% (25/61) of the physicians reported approaching a colleague in difficulty using strategies taught in the online CPD course. Identifying BCTs in the CPD course design was a methodological innovation for FMSQ course evaluation.

Although the CPD course demonstrated a positive effect on physicians' intention to approach a colleague in difficulty, more research is needed to better understand the mechanisms through which the intervention works. This could be done by translating the course into English, making it available to English-speaking physicians in Quebec and other Canadian provinces or internationally, and measuring the effect of the CPD course on approaching a colleague in difficulty in other contexts. Furthermore, instructional designers should focus on enhancing physicians' beliefs about their capabilities to facilitate their adoption of this peer support behavior. Instructional designers could consider using BCTs, especially when targeted behaviors are less widely accepted by clinicians or perceived as more difficult to adopt. Evaluating the effect of BCTs on behavioral intention could deepen our understanding of interventions aimed at changing clinicians' behaviors.

Our results suggest that peer support courses help build a culture of care among physicians. By enhancing our knowledge of interventions aimed at improving the peer support capabilities of physicians, this project indirectly aimed to improve the well-being of health care professionals. Peer-supported professionals could then be better able to contribute to resilient health care systems and to providing quality care to patients.

Acknowledgments

We would like to highlight the contributions of the scientific committee to the development of this training program, including Dr. Pierre Laramée, Cardiologist in the Department of Medicine at the Centre hospitalier de l'Université de Montréal; Dr. Anne Magnan, General Practitioner and former Medical Advisor and Executive Director at the Quebec Physicians' Assistance Program; and Dr. Yves Tremblay, Pediatrician specializing in Developmental Pediatrics, Public Health, and Preventive Medicine, and Medical Advisor at the Quebec Physicians' Assistance Program. Furthermore, we underline the contributions of several collaborators at various stages of this research project: Louisa Blair (Editor) for editing this paper; Sabrina Guay-Belanger (PhD) for her contribution to designing the protocol and submitting it to the ethics committee; Georgina Suelene Dofara (MSc) for her assistance

XSL•FO
RenderX

with the ethics committee application and review of the paper; and Janyce Gnanvi, who carried out diagnostic evaluations of the linear regression model and programmed the generalized estimating equations (GEE) approach to corroborate findings from bivariate analyses.

The authors declare the use of generative artificial intelligence (GAI) in the research and writing process. According to the Generative Artificial Intelligence–Driven Education Taxonomy (GAIDeT, 2025), optimization of less than 10% of code, the proofreading and editing of 25% of manuscript, translation of questionnaires included in the Multimedia Appendices, and reformatting of certain sentences were delegated to GAI tools under full human supervision. The GAI tools used were Copilot (Microsoft) and ChatGPT (OpenAI). The suggestions made by the GAI tools were verified, modified, and approved by the authors, who bear responsibility for results presented in this study.

Funding

The work was supported by the Fédération des médecins spécialistes du Québec (FMSQ). France Légaré is supported by a Canada Research Chair in Shared Decision Making and Knowledge Mobilization.

Data Availability

The datasets generated or analyzed during this study are available from the corresponding author on reasonable request.

Authors' Contributions

All authors contributed to the conceptualization of the study. MT and SJD carried out the investigation. RCC, EL, and F Lizotte developed the methodology for qualitative data analysis and RCC supervised qualitative analysis. F Lizotte and EL analyzed the qualitative data. F Lizotte performed data curation and conducted quantitative data analysis. SG, as biostatistician, validated all quantitative data analyses except for the diagnostic evaluations of the linear regression model and the programming of the GEE approach, which were carried out entirely by JG. CBU contributed to project administration and participated in drafting the original manuscript. F Légaré and CB supervised the overall research activities. All authors were involved in reviewing and editing the manuscript and have approved the final submitted version.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Conceptual framework.
[DOCX File , 216 KB - mededu_v12i1e80199_app1.docx]

Multimedia Appendix 2
Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) Checklist.
[DOCX File , 117 KB - mededu_v12i1e80199_app2.docx]

Multimedia Appendix 3
Standards for Reporting Qualitative Research (SRQR) Checklist.
[DOCX File , 113 KB - mededu_v12i1e80199_app3.docx]

Multimedia Appendix 4
Complete Questionnaires and Consent Forms (French version).
[DOCX File , 122 KB - mededu_v12i1e80199_app4.docx]

Multimedia Appendix 5
Complete Questionnaires and Consent Forms (English version).
[DOCX File , 120 KB - mededu_v12i1e80199_app5.docx]

Multimedia Appendix 6
Grid for evaluating the CPD course "Approaching a Colleague in Difficulty" offered by the Fédération des médecins spécialistes du Québec (FMSQ) according to Michie’s Behavior Change Techniques.
[DOCX File , 115 KB - mededu_v12i1e80199_app6.docx]

Multimedia Appendix 7

XSL•FO
RenderX

Number and percentage of participants with missing data for each variable.
[DOCX File , 124 KB - mededu_v12i1e80199_app7.docx]

Multimedia Appendix 8
Assessing Robustness in Bivariate Comparisons.
[DOCX File , 126 KB - mededu_v12i1e80199_app8.docx]

Multimedia Appendix 9
Analysis of mean intention difference by age, gender and by domain of medical specialty.
[DOCX File , 114 KB - mededu_v12i1e80199_app9.docx]

Multimedia Appendix 10
Sensitivity analysis.
[DOCX File , 115 KB - mededu_v12i1e80199_app10.docx]

Multimedia Appendix 11
Supplementary qualitative results.
[DOCX File , 120 KB - mededu_v12i1e80199_app11.docx]

References

1. Physician wellness: new 2021 National Physician Health Survey findings — burnout, short-staffing and an overburdened system take their toll. Canadian Medical Association. URL: https://www.cma.ca/physician-wellness-hub/content/physician-wellness-new-2021-national-physician-health-survey [accessed 2023-01-26]
2. Shanafelt TD, West CP, Sinsky C, Trockel M, Tutty M, Satele DV. Changes in burnout and satisfaction with work-life integration in physicians and the general US working population between 2011 and 2017. Mayo Clin Proc 2019;94(9):1681-1694 [FREE Full text] [doi: 10.1016/j.mayocp.2018.10.023] [Medline: 30803733]
3. Shanafelt TD, West CP, Dyrbye LN, Trockel M, Tutty M, Wang H. Changes in burnout and satisfaction with work-life integration in physicians during the first 2 years of the COVID-19 pandemic. Mayo Clin Proc 2022;97(12):2248-2258 [FREE Full text] [doi: 10.1016/j.mayocp.2022.09.002] [Medline: 36229269]
4. Rotenstein LS, Torre M, Ramos MA, Rosales RC, Guille C, Sen S. Prevalence of burnout among physicians: a systematic review. JAMA 2018;320(11):1131-1150 [FREE Full text] [doi: 10.1001/jama.2018.12777] [Medline: 30326495]
5. West CP, Dyrbye LN, Sinsky CA, Trockel M, Tutty M, Satele D, et al. Social isolation and burnout, professional fulfillment, and suicidal ideation among US physicians. Mayo Clin Proc 2025;100(12):2115-2127 [FREE Full text] [doi: 10.1016/j.mayocp.2025.07.025] [Medline: 41175133]
6. Miller SC, Scott SD, Beck M. Second victims and mindfulness: A systematic review. J Patient Saf Risk Manag 2019;24(3):108-117. [doi: 10.1177/2516043519838176]
7. Seys D, Wu AW, Van Gerven E, Vleugels A, Euwema M, Panella M. Health care professionals as second victims after adverse events: a systematic review. Eval Health Prof 2013;36(2):135-162. [doi: 10.1177/0163278712458918] [Medline: 22976126]
8. Whitehead PB, Haisch CE, Hankey MS, Mutcheson RB, Dewitt SA, Stewart CA. Studying moral distress and moral injury among inpatient and outpatient healthcare professionals during the COVID-19 pandemic. Int J Psychiatry Med 2024;59(4):469-486. [doi: 10.1177/00912174231205660] [Medline: 37807925]
9. Hamric AB. A case study of moral distress. J Hosp Palliat Nurs 2014;16(8):457-463. [doi: 10.1097/njh.0000000000000104]
10. Dewa CS, Loong D, Bonato S, Trojanowski L. The relationship between physician burnout and quality of healthcare in terms of safety and acceptability: a systematic review. BMJ Open 2017;7(6):e015141 [FREE Full text] [doi: 10.1136/bmjopen-2016-015141] [Medline: 28637730]
11. Hodkinson A, Zhou A, Johnson J, Geraghty K, Riley R, Zhou A. Associations of physician burnout with career engagement and quality of patient care: systematic review and meta-analysis. BMJ 2022;378:e070442 [FREE Full text] [doi: 10.1136/bmj-2022-070442] [Medline: 36104064]
12. Dewa CS, Jacobs P, Thanh NX, Loong D. An estimate of the cost of burnout on early retirement and reduction in clinical hours of practicing physicians in Canada. BMC Health Serv Res 2014;14:254 [FREE Full text] [doi: 10.1186/1472-6963-14-254] [Medline: 24927847]
13. Nundy S, Cooper LA, Mate KS. The quintuple aim for health care improvement: a new imperative to advance health equity. JAMA 2022;327(6):521-522. [doi: 10.1001/jama.2021.25181] [Medline: 35061006]
14. Carbone R, Ferrari S, Callegarin S, Casotti F, Turina L, Artioli G. Peer support between healthcare workers in hospital and out-of-hospital settings: a scoping review. Acta Biomed 2022;93(5):e2022308 [FREE Full text] [doi: 10.23750/abm.v93i5.13729] [Medline: 36300208]

XSL•FO
RenderX

15. Finney RE, Czinski S, Fjerstad K, Arteaga GM, Weaver AL, Riggan KA. Evaluation of a second victim peer support program on perceptions of second victim experiences and supportive resources in pediatric clinical specialties using the Second Victim Experience and Support Tool (SVEST). J Pediatr Nurs 2021;61:312-317. [doi: 10.1016/j.pedn.2021.08.023] [Medline: 34500175]
16. Burlison JD, Scott SD, Browne EK, Thompson SG, Hoffman JM. The second victim experience and support tool: validation of an organizational resource for assessing second victim effects and the quality of support resources. J Patient Saf 2017;13(2):93-102 [FREE Full text] [doi: 10.1097/PTS.0000000000000129] [Medline: 25162208]
17. Meredith LS, Ahluwalia S, Chen PG, Dong L, Farmer CM, Bouskill KE. Testing an intervention to improve health care worker well-being during the COVID-19 pandemic: a cluster randomized clinical trial. JAMA Netw Open 2024;7(4):e244192 [FREE Full text] [doi: 10.1001/jamanetworkopen.2024.4192] [Medline: 38687482]
18. Pelikan M, Finney RE, Jacob A. Use of the Second Victim Experience and Support Tool (SVEST) to assess the impact of a departmental peer support program on anesthesia professionals' second victim experiences and perceptions of support two years after implementation. AANA J 2023;91(5):371-379. [Medline: 37788179]
19. Shapiro J, Galowitz P. Peer support for clinicians: a programmatic approach. Acad Med 2016;91(9):1200-1204. [doi: 10.1097/ACM.0000000000001297] [Medline: 27355784]
20. Creamer MC, Varker T, Bisson J, Darte K, Greenberg N, Lau W. Guidelines for peer support in high-risk organizations: an international consensus study using the Delphi method. J Trauma Stress 2012;25(2):134-141. [doi: 10.1002/jts.21685] [Medline: 22522726]
21. Scott SD, Hirschinger LE, Cox KR, McCoig M, Hahn-Cover K, Epperly KM. Caring for our own: deploying a systemwide second victim rapid response team. Jt Comm J Qual Patient Saf 2010;36(5):233-240. [doi: 10.1016/s1553-7250(10)36038-7] [Medline: 20480757]
22. Edrees H, Connors C, Paine L, Norvell M, Taylor H, Wu AW. Implementing the RISE second victim support programme at the Johns Hopkins Hospital: a case study. BMJ Open 2016;6(9):e011708 [FREE Full text] [doi: 10.1136/bmjopen-2016-011708] [Medline: 27694486]
23. Davidson JE, Graham P, Montross-Thomas L, Norcross W, Zerbi G. Code Lavender: cultivating intentional acts of kindness in response to stressful work situations. Explore (NY) 2017;13(3):181-185. [doi: 10.1016/j.explore.2017.02.005] [Medline: 28668136]
24. Merandi J, Winning A, Liao N, Rogers E, Lewe D, Gerhardt C. Implementation of a second victim program in the neonatal intensive care unit: an interim analysis of employee satisfaction. J Patient Saf Risk Manag 2018;23(6):231-238 [FREE Full text] [doi: 10.1177/2516043518809457]
25. Lane MA, Newman BM, Taylor MZ, O Neill M, Ghetti C, Woltman RM. Supporting clinicians after adverse events: development of a clinician peer support program. J Patient Saf 2018;14(3):e56-e60 [FREE Full text] [doi: 10.1097/PTS.0000000000000508] [Medline: 29878948]
26. El Hechi MW, Bohnen JD, Westfal M, Han K, Cauley C, Wright C. Design and impact of a novel surgery-specific second victim peer support program. J Am Coll Surg 2020;230(6):926-933. [doi: 10.1016/j.jamcollsurg.2019.10.015] [Medline: 31857209]
27. Keyser EA, Weir LF, Valdez MM, Aden JK, Matos RI. Extending peer support across the Military Health System to decrease clinician burnout. Mil Med 2021;186(Suppl 1):153-159. [doi: 10.1093/milmed/usaa225] [Medline: 32830273]
28. Klatt TE, Sachs JF, Huang CC, Pilarski AM. Building a program of expanded peer support for the entire health care team: no one left behind. Jt Comm J Qual Patient Saf 2021;47(12):759-767. [doi: 10.1016/j.jcjq.2021.08.010] [Medline: 34580016]
29. Schuster MA. Creating the Hematology/Oncology/Stem Cell Transplant Advancing Resiliency Team: a nurse-led support program for hematology/oncology/stem cell transplant staff. J Pediatr Oncol Nurs 2021;38(5):331-341. [doi: 10.1177/10434542211011046] [Medline: 33988484]
30. Wolf M, Smith K, Basu M, Heiss K. The prevalence of second victim syndrome and emotional distress in pediatric intensive care providers. J Pediatr Intensive Care 2023;12(2):125-130 [FREE Full text] [doi: 10.1055/s-0041-1731666] [Medline: 37082466]
31. Hinzmann D, Forster A, Koll-Krüsmann M, Schießl A, Schneider F, Sigl-Erkel T. Calling for help—peer-based psychosocial support for medical staff by telephone: a best practice example from Germany. Int J Environ Res Public Health 2022;19(23) [FREE Full text] [doi: 10.3390/ijerph192315453] [Medline: 36497526]
32. Wilkie T, Tajirian T, Stergiopoulos V. Advancing physician wellness, engagement and excellence in a mental health setting: a Canadian perspective. Health Promot Int 2022;37(1). [doi: 10.1093/heapro/daab061] [Medline: 33942064]
33. Barghouth U, Gianakos A, Verran D, Klein SE, Butler KL, Miller P. Peer support helps physicians navigate workplace conflict. J Gen Intern Med 2023;38(7):1768-1770 [FREE Full text] [doi: 10.1007/s11606-023-08092-9] [Medline: 36917432]
34. Doyle K, Murray T, Fong IC, Chavez A, Rounds G, Linenberger M. Building a culture of support at a pediatric surgery center through multidisciplinary peer support. J Surg Res 2023;291:90-96 [FREE Full text] [doi: 10.1016/j.jss.2023.04.029] [Medline: 37354705]
35. Liu JJW, Lum PA, Foxcroft L, Lim R, Richardson JD. Faculty-wide peer-support program during the COVID-19 pandemic: design and preliminary results. JMIR Form Res 2023;7:e37527 [FREE Full text] [doi: 10.2196/37527] [Medline: 36862464]

XSL•FO
RenderX

36. Simms L, Ottman KE, Griffith JL, Knight MG, Norris L, Karakcheyeva V. Psychosocial peer support to address mental health and burnout of health care workers affected by COVID-19: a qualitative evaluation. Int J Environ Res Public Health 2023;20(5) [FREE Full text] [doi: 10.3390/ijerph20054536] [Medline: 36901545]
37. Simpson SL, Khan S, Schiferl LM, Boehl L, Horewitz D, Hausfeld J. Implementation of a peer-to-peer support program in a quaternary pediatric medical center. Acad Pediatr 2023;23(8):1481-1488 [FREE Full text] [doi: 10.1016/j.acap.2023.07.004] [Medline: 37482296]
38. Tolins ML, Rana JS, Lippert S, LeMaster C, Kimura YF, Sax DR. Implementation and effectiveness of a physician-focused peer support program. PLoS One 2023;18(11):e0292917 [FREE Full text] [doi: 10.1371/journal.pone.0292917] [Medline: 37910457]
39. Fall F, Hu YY, Walker S, Baertschiger R, Gaffar I, Saltzman D. Peer support to promote surgeon well-being: the APSA program experience. J Pediatr Surg 2024;59(9):1665-1671. [doi: 10.1016/j.jpedsurg.2023.12.022] [Medline: 38272766]
40. Carne B, Furyk J. Supporting clinicians post exposure to potentially traumatic events: emergency department peer support program evaluation. Emerg Med Australas 2025;37(1):e14518. [doi: 10.1111/1742-6723.14518] [Medline: 39397272]
41. High AE, Forest S. Peer support targeting the second victim phenomenon: implementation and outcomes. Cureus 2025;17(2):e78854. [doi: 10.7759/cureus.78854] [Medline: 39949643]
42. Graham P, Zerbi G, Norcross W, Montross-Thomas L, Lobbestael L, Davidson J. Testing of a caregiver support team. Explore (NY) 2019;15(1):19-26. [doi: 10.1016/j.explore.2018.07.004] [Medline: 30424993]
43. Finney RE, Jacob A, Johnson J, Messner H, Pulos B, Sviggum H. Implementation of a second victim peer support program in a large anesthesia department. AANA J 2021;89(3):235-244. [Medline: 34042575]
44. Katzman JW, Tomedi LE, McCoy-Hayes S, Richardson K, Romero E, Rosenbaum N. The Project ECHO first responder resiliency program: curriculum development, listening groups and lessons learned to support providers virtually during a pandemic. Int Rev Psychiatry 2021;33(8):682-690 [FREE Full text] [doi: 10.1080/09540261.2021.2021867] [Medline: 35412429]
45. Bellehsen MH, Cook HM, Shaam P, Burns D, D'Amico P, Goldberg A. Adapting the Stress First Aid model for frontline healthcare workers during COVID-19. Int J Environ Res Public Health 2024;21(2):497-506 [FREE Full text] [doi: 10.3390/ijerph21020171] [Medline: 38397662]
46. Knudsen T, Abrahamsen C, Jørgensen JS, Schrøder K. Validation of the Danish version of the Second Victim Experience and Support Tool. Scand J Public Health 2022;50(4):497-506 [FREE Full text] [doi: 10.1177/14034948211004801] [Medline: 33860696]
47. McLean CP, Betsworth D, Bihday C, Daman MC, Davis CA, Kaysen D. Helping the helpers: adaptation and evaluation of Stress First Aid for healthcare workers in the Veterans Health Administration during the COVID-19 pandemic. Workplace Health Saf 2023;71(4):162-171 [FREE Full text] [doi: 10.1177/21650799221148650] [Medline: 36726298]
48. Gidengil C, Meredith LS, Ahluwalia S, Chen PG, Dong L, Farmer C. Protecting the well-being of healthcare workers during the COVID-19 pandemic—the COVER-HCW project. Patient-Centered Outcomes Research Institute (PCORI) 2024 [FREE Full text]
49. Timmins G, Williamson S, Cassells A, Davis K, Dong L, Tobin JN. Health care worker experiences with a brief peer support and well-being intervention during the COVID-19 pandemic. BMC Health Serv Res 2025;25(1):1253 [FREE Full text] [doi: 10.1186/s12913-025-13268-6] [Medline: 41029307]
50. Fortin MF, Gagnon J. Fondements et étapes du processus de recherche: Méthodes quantitatives et qualitatives. 4th Edition. Montréal: Chenelière Éducation; 2022:496.
51. Godin G. Les comportements dans le domaine de la santé: comprendre pour mieux intervenir. Montréal: Presses de l'Université de Montréal; 2012.
52. Michie S, Johnston M, Francis J, Hardeman W, Eccles M. From theory to intervention: mapping theoretically derived behavioural determinants to behaviour change techniques. Appl Psychol 2008;57(4):660-680 [FREE Full text] [doi: 10.1111/j.1464-0597.2008.00341.x] [Medline: 18631386]
53. Godin G, Bélanger-Gravel A, Eccles M, Grimshaw J. Healthcare professionals' intentions and behaviours: a systematic review of studies based on social cognitive theories. Implement Sci 2008;3(6):36 [FREE Full text] [doi: 10.1186/1748-5908-3-36] [Medline: 18631386]
54. Légaré F, Freitas A, Turcotte S, Borduas F, Jacques A, Luconi F. Responsiveness of a simple tool for assessing change in behavioral intention after continuing professional development activities. PLoS One 2017;12(5):e0176678 [FREE Full text] [doi: 10.1371/journal.pone.0176678] [Medline: 28459836]
55. Vandenbroucke JP, von Elm E, Altman DG, Gøtzsche PC, Mulrow CD, Pocock SJ. Strengthening the reporting of observational studies in epidemiology (STROBE): explanation and elaboration. Epidemiology 2007;18(6):805-835. [doi: 10.1097/EDE.0b013e3181577511] [Medline: 18049195]
56. O'Brien BC, Harris IB, Beckman TJ, Reed DA, Cook DA. Standards for reporting qualitative research: a synthesis of recommendations. Acad Med 2014;89(9):1245-1251 [FREE Full text] [doi: 10.1097/ACM.0000000000000388] [Medline: 24979285]
57. About CPD accreditation. Royal College of Physicians and Surgeons of Canada. URL: https://news.royalcollege.ca/en/standards-and-accreditation/cpd-accreditation [accessed 2025-03-29]

XSL•FO
RenderX

58. Hoffmann TC, Glasziou PP, Boutron I, Milne R, Perera R, Moher D. Better reporting of interventions: template for intervention description and replication (TIDieR) checklist and guide. BMJ 2014;348:g1687 [FREE Full text] [doi: 10.1136/bmj.g1687] [Medline: 24609605]
59. Novick JS. Six conseils pour soutenir un collègue. Fédération des médecins omnipraticiens du Québec (FMOQ). 2021. URL: https://www.pamq.org/wp-content/uploads/2021/10/Ecouter_6-conseils-pour-soutenir-un-collegue_fmoq.pdf [accessed 2025-02-27]
60. Prendre ses signes vitaux psychologiques. Québec Physicians' Health Program. 2021. URL: https://www.pamq.org/wp-content/uploads/2021/08/Prendre-ses-signes-vitaux-psychologiques.pdf [accessed 2025-02-27]
61. Événements personnels pouvant avoir un impact psychologique sur les médecins qui les vivent. Québec Physicians' Health Program. 2021. URL: https://www.pamq.org/wp-content/uploads/2021/10/Evenements-personnels-pouvant-avoir-un-impact.pdf [accessed 2025-02-27]
62. Situations pouvant amener un médecin à intervenir auprès d'un collègue. Québec Physicians' Health Program. 2021. URL: https://www.pamq.org/wp-content/uploads/2021/10/Situations-pouvant-amener-un-medecin-a-intervenir.pdf [accessed 2026-02-27]
63. Practice transformation series: preventing physician distress and suicide: sample scripting for approaching distressed physicians. American Medical Association. URL: https://tinyurl.com/572v9ubw [accessed 2026-01-20]
64. L'approche bienveillante axée sur le rétablissement. Québec Physicians' Health Program. 2021. URL: https://www.pamq.org/wp-content/uploads/2022/12/LApproche-bienveillante-axee-sur-le-retablissement.pdf [accessed 2024-05-05]
65. Événements professionnels stressants en médecine. Québec Physicians' Health Program. 2021. URL: https://www.pamq.org/wp-content/uploads/2021/10/Evenements-professionnels-stressants.pdf [accessed 2025-02-27]
66. L'identité rétrécie. Québec Physicians' Health Program. 2021. URL: https://www.pamq.org/wp-content/uploads/2021/10/Identite-retrecie.pdf [accessed 2025-02-27]
67. Stratégies individuelles pour traverser une période difficile. Québec Physicians' Health Program. 2021. URL: https://www.pamq.org/wp-content/uploads/2021/10/Strategies-individuelles-pour-traverser-une-periode-difficile.pdf [accessed 2025-02-27]
68. Bloom BS. Taxonomie des objectifs pédagogiques. 2nd Edition. Montréal: Presses de l'Université du Québec; 1977.
69. Frank JR, Snell L, Sherbino J, Boucher A. Référentiel de compétences CanMEDS 2015 pour les médecins. Ottawa: Collège royal des médecins et chirurgiens du Canada; 2015.
70. Specialists for you: FMSQ. Federation of Medical Specialists of Quebec. 2025. URL: https://www.fmsq.org/en [accessed 2026-01-20]
71. Hrisos S, Eccles MP, Francis JJ, Dickinson HO, Kaner EFS, Beyer F. Are there valid proxy measures of clinical behaviour? A systematic review. Implement Sci 2009;4(17):37-41 [FREE Full text] [doi: 10.1186/1748-5908-4-37] [Medline: 19575790]
72. Bakwa Kanyinga F, Gogovor A, Dofara SG, Gadio S, Tremblay M, Daniel SJ. Evaluating the impact of continuing professional development courses on physician behavioral intention: a pre-post study with follow-up at six months. BMC Med Educ 2023;23(1):629 [FREE Full text] [doi: 10.1186/s12909-023-04597-3] [Medline: 37661265]
73. Légaré F, Borduas F, Freitas A, Jacques A, Godin G, Luconi F. Development of a simple 12-item theory-based instrument to assess the impact of continuing professional development on clinical behavioral intentions. PLoS One 2014;9(3):e91013 [FREE Full text] [doi: 10.1371/journal.pone.0091013] [Medline: 24643173]
74. Légaré F, Borduas F, Freitas A, Turcotte S. User Manual – The Continuing Professional Development (CPD) Reaction Questionnaire. 2015. URL: https://www.dropbox.com/scl/fi/npjqz7cbdyjoc5f3ixvxm/CPD-Reaction_User-Manual.pdf?rlkey=5gw4tjicr3su0mof5aun4w8pj&e=1&dl=0 [accessed 2026-01-20]
75. Tavakol M, Dennick R. Making sense of Cronbach's alpha. Int J Med Educ 2011;2:53-55 [FREE Full text] [doi: 10.5116/ijme.4dfb.8dfd] [Medline: 28029643]
76. Lumley T, Diehr P, Emerson S, Chen L. The importance of the normality assumption in large public health data sets. Annu Rev Public Health 2002;23:151-169. [doi: 10.1146/annurev.publhealth.23.100901.140546] [Medline: 11910059]
77. Ancelle T. Statistique épidémiologie. 5th Edition. Paris: Maloine; 2023.
78. From data to action: understanding the drivers of physician wellness. Regression analyses – CMA 2021 National Physician Health Survey results. Canadian Medical Association. 2022. URL: https://www.cma.ca/sites/default/files/2022-11/NPHS-regression-analysis-FR.pdf [accessed 2024-03-18]
79. Lakens D. Calculating and reporting effect sizes to facilitate cumulative science: a practical primer for t-tests and ANOVAs. Front Psychol 2013;4:863 [FREE Full text] [doi: 10.3389/fpsyg.2013.00863] [Medline: 24324449]
80. Ballinger GA. Using generalized estimating equations for longitudinal data analysis. Organ Res Methods 2004;7(2):127-150. [doi: 10.1177/1094428104263672] [Medline: 10]
81. Vittinghoff E. Regression Methods in Biostatistics: Linear, Logistic, Survival, and Repeated Measures Models. New York: Springer; 2012.
82. Hastie T, Tibshirani R, Friedman JH. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. New York: Springer; 2009.
83. Huh I, Gim J. Exploration of Likert scale in terms of continuous variable with parametric statistical methods. BMC Med Res Methodol 2025;25(1):218 [FREE Full text] [doi: 10.1186/s12874-025-02668-1] [Medline: 41023822]

XSL•FO
RenderX

84. Rhemtulla M, Brosseau-Liard P, Savalei V. When can categorical variables be treated as continuous? A comparison of robust continuous and categorical SEM estimation methods under suboptimal conditions. Psychol Methods 2012;17(3):354-373. [doi: 10.1037/a0029315] [Medline: 22799625]
85. Ayivi-Vinz G, Tremblay M, Gadio S, Dofara SG, Daniel SJ, Talbot D. Impact of educational activity formats, online or in-person, on the intention of medical specialists to adopt a clinical behaviour: a comparative study. J CME 2024;13(1):2363550 [FREE Full text] [doi: 10.1080/28338073.2024.2363550] [Medline: 38873619]
86. Kivits JL, Balard FDR, Fournier CC, Winance M, Kivits J, Balard F. Les recherches qualitatives en santé. Malakoff: Armand Colin; 2016.
87. Cheung K, Tai K. The use of intercoder reliability in qualitative interview data analysis in science education. Res Sci Technol Educ 2021;41(3):1155-1175. [doi: 10.1080/02635143.2021.1993179]
88. Östlund U, Kidd L, Wengström Y, Rowa-Dewar N. Combining qualitative and quantitative research within mixed method research designs: a methodological review. Int J Nurs Stud 2011;48(3):369-383 [FREE Full text] [doi: 10.1016/j.ijnurstu.2010.10.005] [Medline: 21084086]
89. Sutton S. Interventions très brèves sur le changement de comportement dans des environnements de soins de santé. Practical Health Psychology 2017 [FREE Full text]
90. Reed SK. In: Chipman SEF, editor. The Oxford Handbook of Cognitive Science. London: Oxford University Press; 2017:231-248.
91. Cohen J. A Coefficient of Agreement for Nominal Scales. Educ Psychol Meas 1960;20(1):37-46 [FREE Full text] [doi: 10.1177/001316446002000104]
92. Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics 1977;33(1):159 [FREE Full text] [doi: 10.2307/2529310] [Medline: 38589863]
93. Bos-van den Hoek DW, van Laarhoven HWM, Ali R, Bakker SD, Goosens A, Hendriks MP. Blended online learning for oncologists to improve skills in shared decision making about palliative chemotherapy: a pre-posttest evaluation. Support Care Cancer 2023;31(3):184 [FREE Full text] [doi: 10.1007/s00520-023-07625-6] [Medline: 36820944]
94. Taqif H, Adisso L, Gomes Souza L, Dofara SG, Ghio SC, Rivest LP. Changes in intention to use an interprofessional approach to decision-making following training: a cluster before-and-after study. BMC Health Serv Res 2024;24(1):437-450 [FREE Full text] [doi: 10.1186/s12913-024-10899-z] [Medline: 38589863]
95. van de Water LF, Kuijper SC, Henselmans I, van Alphen EN, Kooij ES, Calff MM. Effect of a prediction tool and communication skills training on communication of treatment outcomes: a multicenter stepped wedge clinical trial (the SOURCE trial). EClinicalMedicine 2023;64:102244 [FREE Full text] [doi: 10.1016/j.eclinm.2023.102244] [Medline: 37781156]
96. Terrill BN, Pearce A, Chau A, Young MA. Navigating genomic testing: evaluation of an e-learning module with general practitioners. Focus Health Prof Educ 2024;25(1):37-50. [doi: 10.11157/fohpe.v25i1.630]
97. Latif A, Waring J, Chen LC, Pollock K, Solomon J, Gulzar N. Supporting the provision of pharmacy medication reviews to marginalised (medically underserved) groups: a before/after questionnaire study investigating the impact of a patient-professional co-produced digital educational intervention. BMJ Open 2019;9(9):e031548 [FREE Full text] [doi: 10.1136/bmjopen-2019-031548] [Medline: 31530620]
98. Yu C, Choi D, Bruno BA, Thorpe KE, Straus SE, Cantarutti P. Impact of MyDiabetesPlan, a web-based patient decision aid on decisional conflict, diabetes distress, quality of life, and chronic illness care in patients with diabetes: cluster randomized controlled trial. J Med Internet Res 2020;22(9):e16984 [FREE Full text] [doi: 10.2196/16984] [Medline: 32996893]
99. Lawani MA, Côté L, Morin M, Witteman HO, Caron D, Kroger E. Professional training on shared decision making with older adults living with neurocognitive disorders: a mixed-methods implementation study. BMC Med Inform Decis Mak 2020;20(1):189. [doi: 10.1186/s12911-020-01197-9]
100. Urban K, Wright PB, Hester AL, Curran G, Rojo M, Tsai PF. Evaluation of an education strategy versus usual care to implement the STEADI algorithm in primary care clinics in an academic medical center. Clin Interv Aging 2020;15(6):1059-1066 [FREE Full text] [doi: 10.2147/CIA.S256416] [Medline: 32753856]
101. McWilliams S. Addressing compassion fatigue of a family caregiver of adolescents in a mental health setting. Int J Nurs Health Care Res 2020;3(11):1206 [FREE Full text] [doi: 10.29011/2688-9501.101206]
102. Cheung IW, Bland VR, Castro-Noriega E, Silva DS, Miller KJ, Michalak EE. Pilot of a training format for clinicians' continuing professional development: open house with rotating short segments. J Contin Educ Health Prof 2022;42(1):e92-e95 [FREE Full text] [doi: 10.1097/CEH.0000000000000366] [Medline: 34108385]
103. Rivera-Chiauzzi EY, Smith HA, Moore-Murray T, Lee C, Goffman D, Bernstein PS. Healing our own: a randomized trial to assess benefits of peer support. J Patient Saf 2022;18(1):e308-e314. [doi: 10.1097/PTS.0000000000000771] [Medline: 32925571]
104. Babiker A, El Husseini M, Al Nemri A, Al Frayh A, Al Juryyan N, Faki MO. Health care professional development: working as a team to improve patient care. Sudan J Paediatr 2014;14(2):9-16 [FREE Full text] [Medline: 27493399]

XSL•FO
RenderX

105. Kruper A, Domeyer-Klenske A, Treat R, Pilarski A, Kaljo K. Secondary traumatic stress in Ob-Gyn: a mixed methods analysis assessing physician impact and needs. J Surg Educ 2021;78(3):1024-1034. [doi: 10.1016/j.jsurg.2020.08.038] [Medline: 32948508]
106. Dollard MF, Bakker AB. Psychosocial safety climate as a precursor to conducive work environments, psychological health problems, and employee engagement. J Occup Organ Psychol 2010;83(3):579-599 [FREE Full text] [doi: 10.1348/096317909X470690] [Medline: 17570838]
107. Loh MY, Idris MA, Dollard MF, Isahak M. Psychosocial safety climate as a moderator of the moderators: Contextualizing JDR models and emotional demands effects. J Occup Organ Psychol 2018;91(3):620-644. [doi: 10.1111/joop.12211]
108. O'Cathain A, Murphy E, Nicholl J. Why, and how, mixed methods research is undertaken in health services research in England: a mixed methods study. BMC Health Serv Res 2007;7:85. [doi: 10.1186/1472-6963-7-85]
109. Pawson R, Tilley N. Realistic Evaluation. Thousand Oaks: Sage Publications; 1997:231-247.

Abbreviations

BCT: behavior change technique
CIUSSSCN: Centre intégré universitaire de santé et de services sociaux de la Capitale-Nationale
CPD: continuing professional development
FMSQ: Fédération des médecins spécialistes du Québec
GEE: generalized estimating equations
QPHP: Québec Physicians' Health Program
SRQR: Standards for Reporting Qualitative Research
STROBE: Strengthening the Reporting of Observational Studies in Epidemiology
TIDieR: Template for Intervention Description and Replication

Edited by A Stone, T Leung; submitted 29.Aug.2025; peer-reviewed by P Whitehead, H Kwon; comments to author 05.Nov.2025; revised version received 10.Dec.2025; accepted 31.Dec.2025; published 05.Feb.2026.

Please cite as:
Lizotte F, Tremblay M, Biron C, Lachance É, Gadio S, de Carvalho Corôa R, Uwizeye CB, Daniel SJ, Légaré F
Effect of an Online Continuing Professional Development Course on Physicians' Intention to Approach a Colleague in Difficulty: Mixed Methods Convergent Study
JMIR Med Educ 2026;12:e80199
URL: https://mededu.jmir.org/2026/1/e80199
doi:10.2196/80199
PMID:41474517

©Florence Lizotte, Martin Tremblay, Caroline Biron, Éloi Lachance, Souleymane Gadio, Roberta de Carvalho Corôa, Claude Bernard Uwizeye, Sam J Daniel, France Légaré. Originally published in JMIR Medical Education (https://mededu.jmir.org), 05.Feb.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Education, is properly cited. The complete bibliographic information, a link to the original publication on https://mededu.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Impact of Hands-On Pediatric Emergency Training on Wound Suturing, Airway Management, and Orthopedic Casting Skills: Prospective Pre-Post Interventional Study

Saeed Alzahrani[1], MD; Nasser Assiri[1], MD; Fayez Alharthi[1], MD; Tariq Al Garni[1], MD; Asim Alamri[1], MD; Bader Asiri[1], MD; Maher Alkuwaiti[1], MD; Raed Habib[1], MD

Pediatric Emergency Department, Prince Sultan Military Medical City, Riyadh, Saudi Arabia

Corresponding Author:
Saeed Alzahrani, MD
Pediatric Emergency Department
Prince Sultan Military Medical City
Al Qilaa street, Dhahrat Laban Dist, 3117
Riyadh, 13753
Saudi Arabia
Phone: 966 507679144
Email: dr.saeeed.za@gmail.com

Abstract

Background: Opportunities to practice low-frequency skills for high-acuity pediatric emergency procedures in clinical settings are limited. The Basic Emergency Skills Techniques 3N1 (BEST 3N1) workshop was designed as a structured, hands-on training program to support procedural skill development among pediatric emergency care practitioners.

Objective: This study aimed to evaluate preworkshop to postworkshop changes in procedural performance and knowledge after a 3-day hands-on pediatric emergency skills workshop.

Methods: A quasi-experimental prospective pre-post interventional study was conducted during the 8th BEST 3N1 workshop, held from February 11 to 13, 2025, at the Simulation and Skills Development Centre, Princess Nourah bint Abdulrahman University, Riyadh, Saudi Arabia. The workshop included 24 scheduled continuing medical education hours in 3 domains: wound suturing, pediatric airway management, and pediatric orthopedic emergencies and casting. Procedural performance was assessed using complete paired preworkshop and postworkshop data for each assessment. Day 1 wound suturing and day 2 airway management were assessed using 4-item checklists scored from 0 to 8. Day 3 orthopedic casting competence was assessed using a keyword-based rubric covering equipment, procedural steps, and postapplication care, with a total score range of 0 to 30. Written knowledge was assessed using a 30-item multiple-choice test.

Results: The number of paired observations differed in assessments because the analysis was based on complete matched preworkshop and postworkshop data for each outcome. Day 1 wound suturing checklist mean scores increased from 4.14 (SD 1.72) to 7.72 (SD 0.68) among 88 paired participants, with a mean change of 3.58 (SD 1.71; P<.001) points. Day 2 airway management checklist scores increased from 5.47 (SD 1.88) to 7.81 (SD 0.56) among 77 paired participants, with a mean change of 2.34 (SD 1.96; P<.001) points. Day 3 orthopedic casting total competence scores increased from 4.20 to 6.39 among 77 paired participants, with a mean change of 2.19 (SD 1.72; 95% CI 1.80-2.58; P<.001). Written knowledge scores increased from 19.04 (SD 3.77) to 22.32 (SD 3.23) among 94 paired participants, with a mean change of 3.28 (95% CI 2.16-4.39; P<.001) points.

Conclusions: The BEST 3N1 workshop was associated with improved checklist-based procedural performance, orthopedic casting competence, and written knowledge among pediatric emergency care practitioners. These findings support the use of structured hands-on workshops for high-acuity, low-frequency pediatric emergency skills.

(JMIR Med Educ 2026;12:e86349) doi:10.2196/86349

KEYWORDS

pediatrics; emergency medicine; simulation training; clinical competence; airway management; surgical casts; continuing medical education; sutures

XSL•FO
RenderX

Introduction

Maintaining procedural competence is a challenge in pediatric emergency medicine, as essential interventions are performed infrequently. In a high-volume pediatric emergency department, critical procedures represent a small proportion of encounters, and fellows and faculty members often have limited annual exposure [1]. According to a multinational survey, pediatric emergency clinicians believe that most critical procedures should be practiced annually through simulation and supervised training [2]. Exposure is constrained by a variable case mix, competition for procedures, patient safety considerations, and the need for senior clinicians to intervene during unstable cases.

Simulation-based education addresses this gap by allowing repeated practice without exposing patients to preventable harm. A 2023 systematic review found that pediatric simulation programs have been used in diverse clinical scenarios to develop knowledge, technical skills, communication, and team performance, although assessment methods and validation are heterogeneous [3]. Hands-on workshops can combine demonstration, deliberate practice, immediate feedback, and reassessment in a controlled environment. Recent multidisciplinary emergency preparedness training in otolaryngology found that procedure-focused simulation improves learners' preparedness in technical domains [4].

Pediatric airway management exemplifies a high-risk area in which infrequent exposure may threaten readiness. A rapid-cycle deliberate-practice curriculum for pediatric emergency medicine fellows is associated with improved simulated airway performance and better first-attempt intubation performance in clinical practice [5]. Simulation-based airway education also improves knowledge and procedural performance among medical interns [6]. In 2025, a structured pediatric simulation course was found to improve the timing and coordination of advanced airway management during simulated resuscitations [7]. An earlier randomized study found that video-enhanced instruction improves aspects of pediatric intubation performance [8].

Comparable educational deficiencies affect wound closure and orthopedic casting. A systematic review of suturing education identified variation in curricula, teaching approaches, assessment instruments, and opportunities for supervised practice [9]. A 2023 hands-on laboratory course for emergency physicians and trainees reported improvements in suturing performance and perceived educational usefulness after structured instruction [10]. Casting is dependent on equipment selection, sequential technique, molding, and postapplication assessment. In a randomized simulation study, stepwise video-assisted deconstruction produced better immediate forearm-casting performance than continuous demonstration, although performance declined at longer follow-up [11].

Multidisciplinary programs integrate suturing, pediatric airway management, and orthopedic casting for pediatric emergency practitioners and assess procedural performance and knowledge using paired preworkshop and postworkshop measures. Evaluating the 3-day Basic Emergency Skills Techniques 3N1 (BEST 3N1) workshop was necessary to determine whether an integrated, hands-on course would improve several low-frequency, clinically important pediatric emergency skills.

Methods

Study Design and Setting

We conducted a prospective pre-post interventional study during the 8th BEST 3N1 workshop. The workshop was conducted from February 11 to February 13, 2025, at the Simulation and Skills Development Centre, Princess Nourah bint Abdulrahman University, Riyadh, Saudi Arabia. The course was organized under the oversight of the Prince Sultan Military Medical City (PSMMC) BEST 3N1 Scientific Committee.

Participants and Recruitment

The available course registration dataset contained 99 candidate records. All eligible participants attending the workshop and working in pediatric emergency care settings were invited to participate in the study. The registration dataset contained a variable labeled "gender," with "male" and "female" as the available recorded categories, in addition to nationality, physician specialty, physician level, and training center. Sex assigned at birth and additional gender identities were not collected separately. Age and years of clinical experience were not collected. Personal identifiers, including names, mobile numbers, email addresses, and Saudi Commission for Health Specialties registration details, were used only for administrative registration or data matching and are not reported in the manuscript.

Analysis was conducted using complete matched preworkshop and postworkshop data for each assessment. The number of paired observations differed in outcomes because some participants withdrew and did not complete all workshop assessment components. Participants with incomplete or unmatched preworkshop and postworkshop data for a specific assessment were excluded from the specific analysis. The final paired analytic samples were 88 participants (88.9%) for the day 1 wound suturing checklist, 77 participants (77.8%) for the day 2 airway management checklist, 77 participants (77.8%) for the day 3 orthopedic casting competence assessment, and 94 participants (94.9%) for the written knowledge test. All 99 eligible candidates initially agreed to participate, and none declined participation at recruitment.

Intervention

BEST 3N1 is a standardized hands-on educational program introduced by PSMMC in 2017. The workshop evaluated in this study included 3 procedural domains: basic emergency suture techniques, pediatric airway management, and pediatric orthopedic emergencies and casting. The course program listed 24 continuing medical education hours, corresponding to 8 hours of scheduled educational activities per day during 3 consecutive days.

Day 1 focused on basic emergency suture techniques and included registration, welcoming activities, a pretraining competency assessment, introduction to instruments and sutures, supervised hands-on practice, candidate self-training with close feedback, a posttraining competency assessment, and a closing

XSL•FO
RenderX

discussion. The structured suturing content included simple interrupted sutures, horizontal and vertical mattress sutures, buried and subcuticular sutures, hand injuries, facial injuries, and use of surgical glue and staplers.

Day 2 focused on pediatric airway management and included registration, a pretraining competency assessment, group rotation through airway workshop stations, candidate self-training with close feedback, a posttraining competency assessment, a multiple-choice knowledge examination, and a closing discussion. The airway stations included surgical cricothyrotomy; needle cricothyrotomy; pediatric ear, nose, and throat emergencies; simulation; direct laryngoscopy and supraglottic airways; and video laryngoscopy.

Day 3 focused on pediatric orthopedic emergencies and casting and included registration, a pretraining competency assessment, principles of orthopedic casting, group rotation through orthopedic workshop stations, candidate self-training with close feedback, a posttraining competency assessment, and a closing discussion. The orthopedic stations included hand and wrist casting; humerus, elbow, and clavicle casting; and foot, ankle, and knee casting.

Teaching was delivered by a multidisciplinary faculty. Faculty specialties represented in the course materials included pediatric emergency medicine; anesthesiology; pediatric surgery; plastic and maxillofacial surgery; ear, nose, and throat surgery; and orthopedic surgery, with technical support for relevant procedural stations.

Assessment Tools and Outcomes

Participants completed preworkshop and postworkshop assessments of procedural performance and knowledge. The primary outcomes were changes in checklist-based procedural performance, orthopedic casting competence assessed using a keyword-based rubric, and written knowledge test scores.

Day 1 and Day 2 Procedural Checklists

Day 1 wound suturing performance and day 2 airway management performance were assessed using station-based procedural checklists. Each checklist contained 4 items, and each item was scored from 0 to 2, where 0 indicated not done, 1 indicated partially done, and 2 indicated done well. The total checklist scores therefore ranged from 0 to 8 for each domain.

For day 1, the assessed suturing microskills were holding the instrument, grasping the needle, entering at 90°, and performing the knot-tying technique. Day 2 used the same 0 to 8 checklist-scoring structure for airway management performance. These checklists were program-specific assessment tools aligned with the workshop's learning objectives and station content.

Day 3 Orthopedic Casting Keyword Rubric

Day 3 orthopedic casting competence was assessed using a predefined, program-specific keyword-based rubric covering equipment preparation, procedural steps, and postapplication care. Each domain was scored from 0 to 10, resulting in a total score range of 0 to 30. The free-text responses were converted into scores using the predefined rubric. Separate independent ratings of the same free-text responses by multiple raters were not available, and inter-rater reliability could not be assessed.

Written Knowledge Test

Knowledge was assessed using a 30-item multiple-choice test administered before and after the workshop. Each item was scored as correct or incorrect, and the total score ranged from 0 to 30. Total knowledge scores were analyzed in the main manuscript.

Data Management

Assessment data were exported from the workshop assessment files and organized by assessment type. Participant identifiers were used to match preworkshop and postworkshop assessments. After matching, analyses were performed using only paired records. Participants with incomplete or unmatched data were excluded from the analysis of the affected outcome but were retained in other analyses when complete paired data were available. The course registration dataset was used to summarize available learner characteristics. Personally identifiable information was not included in the analysis tables or reported in the manuscript.

Statistical Analysis

Continuous paired outcomes were summarized using means and SDs, and medians and IQRs were reported where appropriate. Normality of the paired difference scores (postworkshop minus preworkshop) was assessed using the Shapiro-Wilk test and visual inspection of *Q-Q* plots. Day 1 wound suturing and day 2 airway management difference scores showed significant departures from normality (Shapiro-Wilk P=.01 and P=.02, respectively); these outcomes were analyzed using the Wilcoxon signed-rank test. P values were reported as the primary inferential result. Day 3 orthopedic casting competence scores and total written knowledge scores were analyzed using 2-tailed paired t tests. Within-participant effect sizes were reported using Cohen d_z where applicable.

Ethical Considerations

This study was reviewed and approved by the Institutional Review Board of PSMMC, Riyadh, Saudi Arabia (approval E-2598). Written informed consent was obtained from all participants for the use of their workshop assessment data. Participants were informed that study participation was voluntary, their data would be kept confidential, and they could withdraw at any time without consequences. Participants received no financial or nonfinancial compensation for their participation in the study.

Results

Course registration data were available for 99 candidates (Figure 1 and Table 1). Among the 99 participants, the registration gender field recorded 56 (56.6%) participants as male and 43 (43.4%) as female. Most candidates were Saudi (n=94, 95%), and the remaining candidates were Egyptian (n=3, 3%), Omani (n=1, 1%), and Bahraini (n=1, 1%). Most candidates were pediatric emergency fellows (n=93, 93.9%); the remaining participants were pediatric emergency specialists (n=4, 4%), a pediatric senior registrar (n=1, 1%), and a pediatric resident

XSL•FO
RenderX

(n=1, 1%). By physician level, candidates were mainly fellow F1 (n=89, 89.9%), followed by specialists (n=5, 5.1%), fellow F2 (n=4, 4%), and resident R4 (n=1, 1%). The registration data represented 22 training centers.

Analyses of training outcomes were conducted using complete paired preworkshop and postworkshop data available for each assessment. The number of paired observations differed in outcomes; 88 participants were included in the day 1 wound suturing checklist analysis, 77 in the day 2 airway management checklist analysis, 77 in the day 3 orthopedic casting competence analysis, and 94 in the written knowledge test analysis. The paired sample sizes differed because some participants withdrew and did not complete all components of the workshop assessment. Participants with incomplete or unmatched preworkshop and postworkshop data for a specific outcome were excluded.

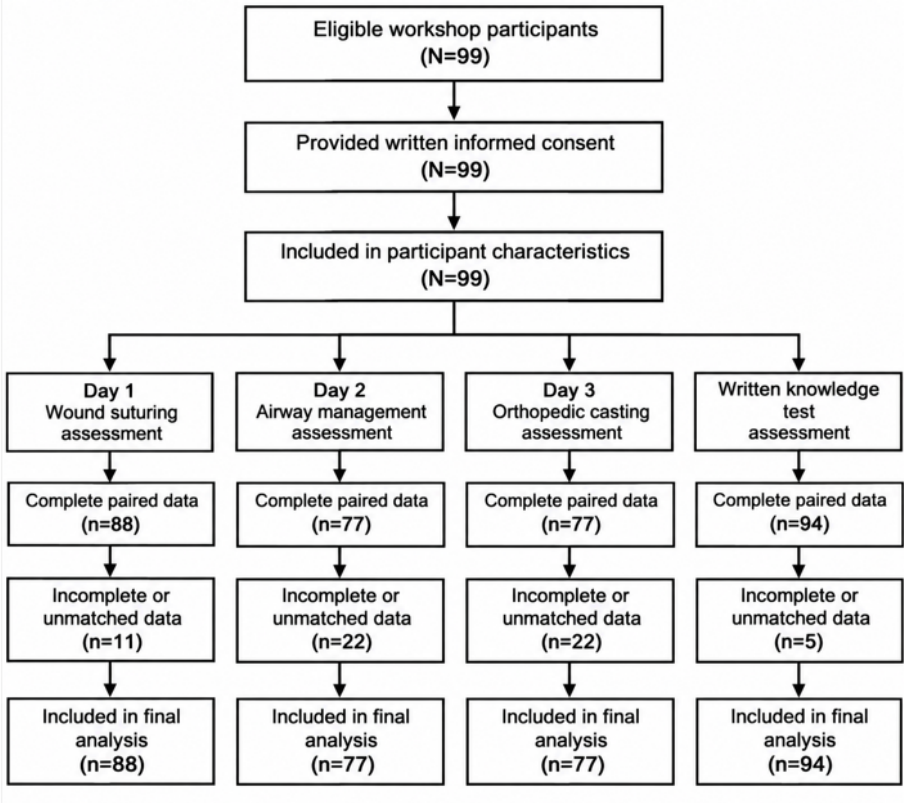
Day 1 focused on wound suturing skills, and checklist-based procedural performance improved after the workshop (Table 2). The mean total checklist score increased from 4.14 (SD 1.72) before training to 7.72 (SD 0.68) after training on a 0 to 8 scale. The mean change was 3.58 (SD 1.71) points, and the median change was 4.00 (IQR 2.75-5.00) points. Among the 88 paired participants, 83 improved, 4 had no change, and 1 had a lower postworkshop score. The Wilcoxon signed-rank test showed a statistically significant pre-post improvement (P<.001).

Day 2 focused on airway management skills, and checklist performance improved after the workshop (Table 2). The mean total score increased from 5.47 (SD 1.88) before training to 7.81 (SD 0.56) after training on the 0 to 8 scale. The mean change was 2.34 (SD 1.96) points, and the median change was 2.00 (IQR 1.00-4.00) points. Of the 77 paired participants, 63 improved, 12 had no change, and 2 had lower scores after the workshop. The Wilcoxon signed-rank test showed a significant improvement (P<.001). Compared with day 1, the mean change was smaller, while postworkshop performance remained high.

Day 3 assessed orthopedic casting competence using a keyword-based covering equipment, procedural steps, and postapplication care (Table 3). All components improved after training. The equipment score increased from 1.87 to 2.61, the procedural steps score increased from 1.49 to 2.43, and the postapplication care score increased from 0.84 to 1.35. The total competence score increased from 4.20 to 6.39 on a 0 to 30 scale, with a mean change of 2.19 (95% CI 1.80-2.58) points. The 2-tailed paired t test showed a significant improvement in the total day 3 score (P<.001), with a large within-participant effect size (Cohen d_z=1.27).

Among 94 participants with paired knowledge test data, the mean score increased from 19.04 (SD 3.77) before training to 22.32 (SD 3.23) after training on a 0 to 30 scale. The mean change was 3.28 (95% CI 2.16-4.39) points, and the 2-tailed paired t test showed a statistically significant improvement (P<.001), with a moderate effect size (Cohen d_z=0.60).

Figure 1. Participant flow.



XSL•FO
RenderX

Table 1. Candidate characteristics from the course registration dataset (N=99)[a].

Characteristics	Values
Gender, n (%)	
Male	56 (56.6)
Female	43 (43.4)
Nationality, n (%)	
Saudi	94 (95)
Egyptian	3 (3)
Omani	1 (1)
Bahraini	1 (1)
Physician specialty, n (%)	
Pediatric emergency fellow	93 (93.9)
Pediatric emergency specialist	4 (4)
Pediatrics senior registrar	1 (1)
Pediatric resident	1 (1)
Physician level, n (%)	
Fellow F1	89 (89.9)
Fellow F2	4 (4)
Specialist	5 (5.1)
Resident R4	1 (1)

[a]The registration data represents 22 training centers.

Table 2. Pre-post procedural checklist performance by domain.

Workshop domain	Score range	Paired participants, n (%)	Preworkshop score, mean (SD)	Preworkshop score, median (IQR)	Postworkshop score, mean (SD)	Postworkshop score, median (IQR)	Mean change, mean (SD)	Median change (IQR)	Improved score, n (%), n	No change, n (%)	Worsened score, n (%)	Primary test	*P* value
Day 1: wound suturing checklist	0-8	88 (88.9)	4.14 (1.72)	4.00 (3.00-5.00)	7.72 (0.68)	8.00 (8.00-8.00)	3.58 (1.71)	4.00 (2.75-5.00)	83 (94.3)	4 (4.5)	1 (1.1)	Wilcoxon signed-rank test	<.001
Day 2: airway management checklist	0-8	77 (77.8)	5.47 (1.88)	6.00 (4.00-7.00)	7.81 (0.56)	8.00 (8.00-8.00)	2.34 (1.96)	2.00 (1.00-4.00)	63 (81.8)	12 (15.6)	2 (2.6)	Wilcoxon signed-rank test	<.001

Table 3. Pre-post orthopedic casting competence using the day 3 keyword rubric.

Competence domain	Score range	Paired participants, n (%)	Preworkshop score, mean (SD)	Postworkshop score, mean (SD)	Mean change (95% CI)	Effect size (Cohen d_z)	Primary test	*P* value
Equipment score	0-10	77 (77.8)	1.87 (0.91)	2.61 (0.88)	0.74 (0.54-0.93)	0.85	2-tailed paired *t* test	<.001
Procedural steps score	0-10	77 (77.8)	1.49 (1.02)	2.43 (0.97)	0.94 (0.72-1.15)	0.99	2-tailed paired *t* test	<.001
Postapplication care score	0-10	77 (77.8)	0.84 (0.48)	1.35 (0.52)	0.51 (0.42-0.61)	1.18	2-tailed paired *t* test	<.001
Total competence score	0-30	77 (77.8)	4.20 (1.85)	6.39 (1.68)	2.19 (1.80-2.58)	1.27	2-tailed paired *t* test	<.001

XSL•FO
RenderX

Discussion

Principal Findings

The BEST 3N1 workshop was associated with immediate improvement in wound suturing, pediatric airway management, orthopedic casting competence, and written knowledge. Improvements across all domains indicate that a concentrated course combining demonstration, supervised practice, feedback, and reassessment strengthens performance in low-frequency pediatric emergency procedures. Differences between domains should not be interpreted as evidence that one component was superior because the assessments used different scales, methods, and baseline levels. Waters et al [12] reported that multidisciplinary in situ simulation improved team performance and use of a rapid sequence intubation checklist in a pediatric emergency department. Mustafa et al [13] also found that a collaborative simulation-based intervention improved pediatric advanced airway management processes in general emergency departments, and their findings support structured airway practice.

The suturing findings are consistent with research emphasizing active practice and feedback. Tejos et al [14] found that video-guided practice without peer or expert feedback was insufficient compared with feedback-supported approaches. Tieken et al [15] showed that feedback-containing instruction improved suturing and knot-tying performance, while Hanada et al [16] found improvement in objectively measured suturing quality after extended simulation practice. The supervised BEST 3N1 stations have supported learning by enabling correction of instrument handling, needle positioning, tissue entry, and knot-tying techniques.

The casting results are supported by orthopedic simulation research. Moktar et al [17] developed a cast-application simulator and objective checklist that produced scores reflecting training level and showed high rater reliability. Seeley et al [18] also validated a pediatric distal-radius reduction and casting model using objective procedural and radiographic outcomes. These studies support structured casting assessment and highlight the need to establish the reliability and validity of the program.

The improvement in written knowledge complements the procedural findings and indicates that hands-on instruction reinforces cognitive and psychomotor learning simultaneously. Lammers et al [19] found that simulation and procedural laboratory approaches improved selected pediatric emergency skills, although limited exposure did not consistently improve overall performance. Thim et al [20] concluded that pediatric emergency simulation improves time-critical performance and adherence to expected tasks, while noting variation in interventions and outcomes. A distinctive contribution is the integration of wound suturing, pediatric airway management, and orthopedic casting within one 3-day multidisciplinary program. This approach provides an efficient model for maintaining readiness in several procedural domains rather than organizing separate single-skill courses. Refresher sessions and standardized assessments could improve durability and reproducibility.

Limitations

This study had several limitations, and the findings should be interpreted in the context of its single-center, uncontrolled pre-post design; variable paired sample sizes; unblinded assessments; and program-specific assessment instruments. The day 3 free-text responses were not scored by multiple raters, and interrater reliability was not assessed. The orthopedic casting scores have been affected by subjective interpretation and rater-related measurement error. Long-term retention of procedural skills and transfer of workshop performance to clinical practice or patient outcomes were not assessed. Future evaluations should use independent raters, validated tools, retention testing, and clinical performance indicators.

Conclusions

The 3-day BEST 3N1 workshop was associated with immediate improvements in wound suturing, pediatric airway management, orthopedic casting competence, and written knowledge among pediatric emergency care practitioners. By integrating 3 low-frequency procedural domains within 1 multidisciplinary, hands-on program, the course offers a practical model for structured skills training in pediatric emergency medicine. Long-term retention, transfer of skills to clinical practice, and patient outcomes were not evaluated. Future studies should use validated assessment instruments, independent raters, longer follow-up, and multicenter controlled designs.

Funding

The authors declared that no financial support was received for this work.

Data Availability

The datasets generated and analyzed during this study are not publicly available because they contain participant-level information and data collected within an institutional educational program. Data are available from the corresponding author on request.

Generative AI Use

Generative AI, specifically ChatGPT (GPT-5; OpenAI), was used during manuscript revision to assist with language editing, grammatical correction, and the organization and clarity of selected sections of the manuscript. It was not used to collect or generate study data, conduct statistical analyses, determine the study findings, or make scientific or clinical decisions. All

XSL•FO
RenderX

AI-assisted content was reviewed and revised by the authors and checked against the original study data and cited literature. The authors take full responsibility for the accuracy, integrity, and final content of the manuscript.

Authors' Contributions

Conceptualization: SA, NA
Methodology: SA, NA, FA, TAG, AA, BA, MA, RH
Investigation: BA, MA, RH
Data curation: NA, FA, TAG, AA, BA
Formal analysis: SA, NA
Project administration: SA, NA
Resources: FA, TAG, AA, BA, MA
Supervision: SA
Visualization: NA, FA, TAG, AA
Writing—original draft: SA, NA, FA, TAG, AA, BA, MA, RH
Writing—review & editing: SA, NA, FA, TAG, ASA, BA, MA, RH
All authors reviewed and approved the final version of the manuscript and agreed to be accountable for the accuracy and integrity of the work.

Conflicts of Interest

None declared.

References

1. Mittiga MR, Geis GL, Kerrey BT, Rinderknecht AS. The spectrum and frequency of critical procedures performed in a pediatric emergency department: implications of a provider-level view. Ann Emerg Med 2013 Mar;61(3):263-270. [doi: 10.1016/j.annemergmed.2012.06.021] [Medline: 22841174]
2. Craig SS, Auerbach M, Cheek JA, Babl FE, Oakley E, Nguyen L, et al. Preferred learning modalities and practice for critical skills: a global survey of paediatric emergency medicine clinicians. Emerg Med J 2019 May;36(5):273-280. [doi: 10.1136/emermed-2017-207384] [Medline: 30327413]
3. Kim E, Song S, Kim S. Development of pediatric simulation-based education - a systematic review. BMC Nurs 2023 Aug 28;22(1):291 [FREE Full text] [doi: 10.1186/s12912-023-01458-8] [Medline: 37641090]
4. La Monte OA, Lee JH, Soliman SI, Saddawi-Konefka R, Harris JP, Coffey CS, et al. Simulation-based workshop for emergency preparedness in otolaryngology. Laryngoscope Investig Otolaryngol 2023 Aug 24;8(5):1159-1168 [FREE Full text] [doi: 10.1002/lio2.1139] [Medline: 37899850]
5. Edmunds KJ, Shah A, Geis GL, Kerrey BT, Klein G, DeBra R, et al. Rapid cycle deliberate practice to improve airway skills and performance of trainees in a pediatric emergency department. AEM Educ Train 2024 Jan 16;8(1):10.1002/aet2.10928. [doi: 10.1002/aet2.10928] [Medline: 38235393]
6. Basnet S, Shrestha SP, Shrestha R, Shrestha AP, Shrestha A, Sahu S, et al. Effect of simulation-based emergency airway management education on the knowledge, skills and perceived confidence of medical interns. Ann Med Surg (Lond) 2024 Aug 7;86(9):5191-5198. [doi: 10.1097/MS9.0000000000002376] [Medline: 39239053]
7. Donath C, Leonhardt A, Stibane T, Weber S, Mand N. To intubate or to resuscitate: the effect of simulation-based training on advanced airway management during simulated paediatric resuscitations. Adv Simul (Lond) 2025 Jan 06;10(1):1 [FREE Full text] [doi: 10.1186/s41077-024-00326-y] [Medline: 39762959]
8. Miller KA, Monuteaux MC, Aftab S, Lynn A, Hillier D, Nagler J. A randomized controlled trial of a video-enhanced advanced airway curriculum for pediatric residents. Acad Med 2018 Dec;93(12):1858-1864. [doi: 10.1097/ACM.0000000000002392] [Medline: 30095451]
9. Emmanuel T, Nicolaides M, Theodoulou I, Yoong W, Lymperopoulos N, Sideris M. Suturing skills for medical students: a systematic review. In Vivo 2021;35(1):1-12 [FREE Full text] [doi: 10.21873/invivo.12226] [Medline: 33402444]
10. Manno A, Pascale MM, Santoro MC, Saviano A, Benicchi A, Gabrielli M, et al. Teaching suturing skill to emergency physicians: the experience of a hands-on practical training in a laboratory simulation-based setting. Emerg Care J 2023 Jun 23;19(2). [doi: 10.4081/ecj.2023.11417]
11. Bouthors C, Veil R, Auregan JC, Molina V, Blanié A, Court C, et al. Deconstructing forearm casting task by videos with step-by-step simulation teaching improved performance of medical students: is making working student's memory work better similar to a process of artificial intelligence or just an improvement of the prefrontal cortex homunculus? Int Orthop 2023 Feb;47(2):467-477. [doi: 10.1007/s00264-022-05626-4] [Medline: 36370162]
12. Waters KM, Hwu R, Kulkarni M, Okonye J, Zamor R, Chaudhary S, et al. Use of in situ simulation to improve team performance and utilization of a rapid sequence intubation checklist. AEM Educ Train 2024 Nov 11;8(6):e11039. [doi: 10.1002/aet2.11039] [Medline: 39534112]

XSL•FO
RenderX

13. Mustafa M, Lutfi R, Alsaedi H, Castelluccio P, Pearson KJ, Montgomery EE, et al. Improvement of pediatric advanced airway management in general emergency departments after a collaborative intervention program. Respir Care 2021 Dec;66(12):1866-1875. [doi: 10.4187/respcare.09250] [Medline: 34670860]
14. Tejos R, Crovari F, Achurra P, Avila R, Inzunza M, Jarry C, et al. Video-based guided simulation without peer or expert feedback is not enough: a randomized controlled trial of simulation-based training for medical students. World J Surg 2021 Jan;45(1):57-65. [doi: 10.1007/s00268-020-05766-x] [Medline: 32892271]
15. Tieken KR, Kelly G, Maxwell J, Visenio MR, Reynolds J, Fingeret AL. Feedback versus compliments versus both in suturing and knot tying simulation: a randomized controlled trial. J Surg Res 2024 Feb;294:99-105. [doi: 10.1016/j.jss.2023.09.037] [Medline: 37866070]
16. Hanada K, Hoshina K, Tsuyuki S, Miyahara K, Suhara M, Matsukura M, et al. Ten-hour simulation training improved the suturing performance of medical students. Ann Vasc Surg 2022 Aug;84:163-168 [FREE Full text] [doi: 10.1016/j.avsg.2021.12.076] [Medline: 34995745]
17. Moktar J, Popkin CA, Howard A, Murnaghan ML. Development of a cast application simulator and evaluation of objective measures of performance. J Bone Joint Surg Am 2014 May 07;96(9):e76. [doi: 10.2106/JBJS.L.01266] [Medline: 24806022]
18. Seeley MA, Fabricant PD, Lawrence JT. Teaching the basics: development and validation of a distal radius reduction and casting model. Clin Orthop Relat Res 2017 Sep;475(9):2298-2305 [FREE Full text] [doi: 10.1007/s11999-017-5336-3] [Medline: 28374350]
19. Lammers RL, Willoughby-Byrwa MJ, Vos DG, Fales WD. Comparison of four methods of paramedic continuing education in the management of pediatric emergencies. Prehosp Emerg Care 2022;26(4):463-475. [doi: 10.1080/10903127.2021.1916140] [Medline: 33872104]
20. Thim S, Henriksen TB, Laursen H, Schram AL, Paltved C, Lindhard MS. Simulation-based emergency team training in pediatrics: a systematic review. Pediatrics 2022 Apr 01;149(4):e2021054305. [doi: 10.1542/peds.2021-054305] [Medline: 35237809]

Abbreviations

BEST 3N1: Basic Emergency Skills Techniques 3N1
PSMMC: Prince Sultan Military Medical City

Edited by A Stone; submitted 22.Oct.2025; peer-reviewed by J Nickerson, Y Han; comments to author 29.May.2026; revised version received 20.Oct.2026; accepted 20.Oct.2026; published 13.Nov.2026.

Please cite as:
Alzahrani S, Assiri N, Alharthi F, Al Garni T, Alamri A, Asiri B, Alkuwaiti M, Habib R
Impact of Hands-On Pediatric Emergency Training on Wound Suturing, Airway Management, and Orthopedic Casting Skills: Prospective Pre-Post Interventional Study
JMIR Med Educ 2026;12:e86349
URL: https://mededu.jmir.org/2026/1/e86349
doi:10.2196/86349
PMID:

©Saeed Alzahrani, Nasser Assiri, Fayez Alharthi, Tariq Al Garni, Asim Alamri, Bader Asiri, Maher Alkuwaiti, Raed Habib. Originally published in JMIR Medical Education (https://mededu.jmir.org), 13.Nov.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Education, is properly cited. The complete bibliographic information, a link to the original publication on https://mededu.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Developing a Realistic and Cost-Effective Training Model (MaiSurge) for Laparoscopic Hysterectomies to Train and Assess Surgical Skill: Prospective Nonrandomized Controlled Trial

Anna Maria Brechter[1]; Roxana Schwab[1], PD, Dr med; Christoph Dold[1]; Christine Skala[1], PD, Dr med; Maria Schröder[1], Dr med; Lina Schiestl[1], Dr med; Katharina Gillen[1,2], Dr med; Walburgis Brenner[1], Prof Dr; Annette Hasenburg[1], MHBA, Prof Dr Med; Mona Wanda Schmidt[1], Dr med, MSc

[1]Department of Obstetrics and Gynecology, University Medical Center of the Johannes Gutenberg University Mainz, Mainz, Germany

[2]Department of Gynecology and Gynecologic Oncology, Diakonie Hospital Jung-Stilling, Siegen, Germany

Corresponding Author:
Mona Wanda Schmidt, Dr med, MSc
Department of Obstetrics and Gynecology, University Medical Center of the Johannes Gutenberg University Mainz, Mainz, Germany

Abstract

Background: Laparoscopic surgery has a flatter learning curve compared to traditional open surgery. Therefore, structured programs and realistic training models are imperative to ensure patients' safety. However, commercially available models are often too expensive or technically unrealistic for continuous surgical training.

Objective: The aim of this trial was to develop a cost-efficient and highly realistic uterus model to perform a total laparoscopic hysterectomy (TLH) and evaluate its applicability.

Methods: A training model (MaiSurge) for a TLH with salpingectomy or adenectomy was developed using a 3D printer and different cast materials. Polyvinyl alcohol was used to allow for the use of electrosurgery. To gather the first validity evidence, novice and expert gynecologists performed a TLH on the model. Operative time as well as surgical performance scores (Hysterectomy–Objective Structured Assessment of Technical Skills) were compared between both groups.

Results: A total of 12 participants in the novice group and 18 participants in the expert group completed the simulation. The experts obtained significantly better modified Hysterectomy–Objective Structured Assessment of Technical Skills scores (mean 74.0, SD 12.9 vs mean 60.3, SD 14.9; P=.049) and performed significantly faster (median 69.5, IQR 49.5-74.3 minutes vs median 37.5, IQR 30.5-38.8 minutes; P<.001). An excellent interrater reliability was observed (intraclass correlation coefficient=0.91). Approximately 92% (11/12) of novices felt that they had improved their surgical performance after training on the MaiSurge uterus model. Overall, all participants agreed that the new MaiSurge uterus model should be integrated into training curricula to improve the performance of residents on TLHs.

Conclusions: A new highly realistic and cost-effective training model (MaiSurge) to perform a TLH was developed. The model distinguishes between good and poor laparoscopic performances and, thus, can be used in training as well as assessment of surgical skills. The possibility of simulating even complex laparoscopic procedures in a realistic environment may be an opportunity to train a future generation of gynecologists without compromising patient safety or exhausting the limited availability of operating room time.

Trial Registration: Deutsche Register für Klinische Studien DRKS00031825; https://drks.de/search/en/trial/DRKS00031825/details

(JMIR Med Educ 2026;12:e66369) doi:10.2196/66369

KEYWORDS

surgical education; laparoscopic hysterectomy; training model; surgical skill; assessment; surgery; hysterectomy; cost-effective; training; laparoscopic surgery; salpingectomy; adenectomy; electrosurgery; gynecologists; gynecology; patient safety; controlled trial

Introduction

A hysterectomy is a common procedure in gynecology in which the uterus and, in the case of an additional salpingectomy, the fallopian tubes are removed [1]. In 2016, the prevalence of hysterectomies was approximately 21% of approximately 600,000 hysterectomies performed overall each year in the United States [2,3]. The procedure can be performed

XSL•FO
RenderX

laparoscopically, vaginally, or per open surgery. While intraoperative complications seem to be comparable between those approaches [4,5], the loss of depth perception, the indirect view on the operating field, and the fulcrum effect including the inversion and scaling of movements in laparoscopic surgery are associated with prolonged learning curves [6]. Training residents in surgical fields is challenging while simultaneously trying to maintain patient safety and keep operating costs low in light of the increasing financial pressure on hospitals. Recent studies have shown that the complication rate and outcomes of total laparoscopic hysterectomies (TLHs) performed by residents are not worse compared to those of TLHs performed by senior physicians, but that the operating time is significantly longer (16%-32% increase in operative time) [7,8], which drastically increases costs for the hospitals and anesthesia times for patients. Currently, gynecology residents in Germany are still learning how to perform laparoscopic surgeries directly in the operating theater on the patient. Simulation training is rarely included within structured residency training programs, and there are no criteria that a resident must meet before performing the first procedure on a patient. These prolonged operative times can result in economic damages to the hospital and disadvantages for the patient due to the longer anesthesia time. Therefore, shortening the learning curve through skill laboratory training can positively impact not only patients but also the training of residents and financial aspects for hospitals.

Apart from numerous basic laparoscopic skill training models, which have been shown to improve surgical performance [9], there is a lack of effective, feasible, and cost-efficient training models for more complex skills. The critical steps of a surgical procedure, as well as full surgical procedures themselves, need to be trained [10]. Currently, there are only a few inanimate models of the uterus on the market, but these are expensive and mostly nonconductive and, thus, unsuitable to use with electrocautery. Animal cadavers, on the other hand, require specialized facilities, and the anatomical differences limit the transferability to humans. To simulate hysterectomies for continuous practice, some researchers have built very simple models from materials such as balloons and plastic tubes [11]. However, due to the materials used and their simplicity, these models are not realistic enough to simulate all surgical steps. Many surgical specialties are faced with similar challenges. To overcome them, some have used 3D printing technologies to create training models, for example, in pancreatic surgery [12]. However, no comparable models are available for laparoscopic hysterectomies.

Thus, the aim of this study was to develop a feasible uterus model to practice laparoscopic hysterectomies outside of the operating room and that can reflect the current surgical skill of a gynecological resident. A main focus was to create a model that addressed the aforementioned shortcomings of the available models by being realistic and cost-effective and allowing for electrosurgery to be used. This study focused on the question of whether the newly developed uterus model can differentiate between good and bad surgical performances and whether it is perceived to adequately reflect a real TLH on a patient. To evaluate this, we conducted a prospective, nonrandomized controlled trial. We hypothesized that the newly created uterus model would convince gynecological experts of its realism and be able to differentiate between levels of surgical expertise based on procedure time and surgical performance scores. If proven true, the newly developed uterus model could be used for continuous surgical education of gynecological residents worldwide, allowing inexperienced surgeons to have their first experience with a TLH and also allowing for their skill level to be assessed and the learning curve to be observed. Especially in areas where minimally invasive surgery is not performed every day due to restricted availability, cost-effective training models such as this uterus model could impact patient safety immensely.

Methods

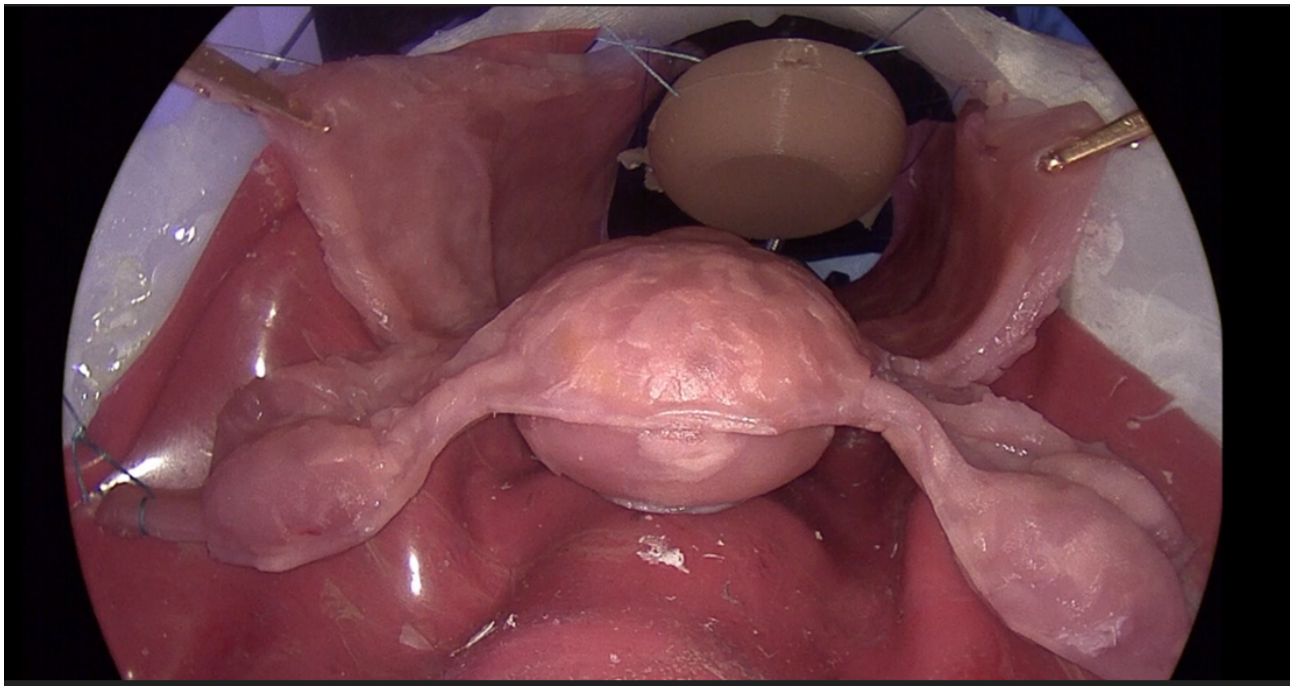
Development of the MaiSurge Uterus Model

Different uterus models were developed using 3D-printed molds, which were then cast with silicone and biotissue material. The uterus model was designed to meet the following criteria: similar texture to human tissue, applicability for training, similar optics to human situs, and realistic surgical procedure performance. Different materials and molds were tested throughout the development stage and adapted based on expert feedback.

The molds were created using the 3D software programs Blender (Blender Foundation) and Autodesk Fusion 360 (Autodesk). For dimension and shaping, pictures and videos, as well as data from textbooks, were used. The 3D datasets of the molds were then printed using the Original Prusa i3 MK3S+ (Prusa Research) 3D printer and injected with silicone or polyvinyl alcohol (PVA). Two different degrees of hardness were used to simulate the layers of the tissue in the silicone model. Models made of PVA underwent several frosting and defrosting cycles until they reached the desired consistency. The uterus model consists of the corpus and the cervix of the uterus with the vagina, as well as 2 tubes with fimbria; 2 ovaries with the utero-ovarian ligaments, the round ligaments, and the broad ligaments; and a layer of peritoneum and the uterine arteries. The corpus and the cervix of the uterus and the vagina are hollow inside so that an adapter can be inserted and the hysterectomy can be performed on it. The newly developed MaiSurge (PVA) uterus model (Figure 1) was used for further evaluation in the trial as it offers the opportunity to work on it using electrocautery during the surgical procedure. The model was fabricated in 3 steps: one for the inner layer, one for the outer layer and the adnexa, and one to add the ventral part of the visceral peritoneum.

XSL•FO
RenderX

Figure 1. MaiSurge model—newly developed uterus model made from polyvinyl alcohol.



Setting and Participants

This study was conducted in the training center for minimally invasive surgery of the Department of Obstetrics and Gynecology at the University Medical Center of Johannes Gutenberg University Mainz. Residents and attending gynecologists of the Department of Obstetrics and Gynecology at the University Medical Center of Johannes Gutenberg University Mainz were recruited for the trial and categorized according to their prior experience with TLH in the last 5 years (experts: ≥10; novices: <10).

Ethical Considerations

This study was registered at the German Clinical Trials Register (DRKS00031825), and ethics approval was granted by the ethics committee of the Rheinland-Pfalz medical association (2022-16451_1). Written informed consent was obtained from all participants before the start of the trial. Participants received no compensation for participating in this trial All data was recorded pseudonymised, with only the participant being able to identify their pseudonym retrospectivley.

Study Design

This was a prospective, nonrandomized, rater-blinded controlled trial. A flowchart of the study design can be found in Multimedia Appendix 1. All participants performed 3 Z-sutures on a vaginal cuff model (with or without prior video explanation) before being shown a video of the TLH procedure to be performed. Subsequently, all participants performed a TLH on the newly developed uterus model with the camera and instrument assistance of a trained tutor. The tutor only acted on direct instruction of the trainee. All performances were video recorded and evaluated by 2 blinded raters.

Surgical Procedure

As the first step of the surgery, the trocars were inserted in the holes of the training box. Starting with the coagulation of the utero-ovarian ligaments, the salpingectomy was performed. Coagulation was achieved through a bipolar grasper and cutting with simple nonelectrical scissors. The round ligament was coagulated completely and then cut. The ventral and dorsal layers of the broad ligament were separated to expose the uterine artery, which appeared after approximately 2 to 3 cm. The uterine artery was then coagulated and cut. After the identification of the vesico-uterine peritoneum, a horizontal incision was used to create a bladder flap and dissect the bladder from the lower uterine segment and cervix. To detach the uterus from the vagina, a colpotomy was performed along the vaginal adapter using a monopolar hook. Because, in this simulation, the extraction through the vagina was not possible, the uterus was removed from the top. The vaginal cuff was closed using Z-sutures. No specific number of sutures on the vaginal cuff was required; rather, the participants were asked to place as many as needed to ensure full closure of the vaginal cuff.

Outcomes

The primary outcome was the quality of the surgical performance, assessed using a modified Hysterectomy–Objective Structured Assessment of Technical Skills (H-OSATS) [13] score, which was developed specifically for the laparoscopic hysterectomy. This score assesses the complete performance of the operation, including positioning of the patient; access to the abdominal cavity; inspection of the organs; insertion of the trocars; and all steps of the operation, such as cutting of the ligaments, separation of the uterus, suturing, and the entire withdrawal of the uterus. In this case, 1 to 5 points (Likert scale) can be awarded depending on how well the surgical step was performed. To fit the simulated procedure, the original H-OSATS score [13] was modified by removing several items, such as the identification of some anatomical structures in the pelvis, which were not included in the model. The modified version can be found in Multimedia Appendix 1. A total of 21 items were included, leading to a

XSL•FO
RenderX

maximum of 105 points. Furthermore, a commonly used Objective Structured Assessment of Technical Skills (OSATS) score (global rating checklist, 7 items on a 5-point Likert scale) by Martin et al [14] was used to assess procedure-independent laparoscopic skills. The H-OSATS score was assessed by 2 specifically trained, blinded raters based on video recordings of the procedure. The operating time was evaluated as a secondary outcome. Furthermore, a 5-point Likert scale was used to evaluate the MaiSurge (PVA) model, the silicone model, and a commercially available uterus model with regard to texture, simulation quality, and realism. The baseline suturing and knot-tying attempts were scored based on the time taken to perform the sutures, as well as using the suturing and knot-tying–specific OSATS score by Chang et al [15]. Among other things, the score assesses how the needle is clamped, how it is guided through the tissue, how the knots are made, and whether the needle remains in the camera's field of vision. In addition, the handling of the instruments and the tissue, as well as the independent performance of the procedure, is assessed.

Statistical Analysis

Descriptive analysis and graphical illustrations (created using SPSS version 27; IBM Corp) are reported where appropriate. Categorical data are presented in absolute and relative frequencies, and continuous data are presented as medians and IQRs or medians and SDs where appropriate. The evaluation of the uterus model was performed using a 5-point Likert scale, which is reported as median and IQR or as a percentage of agreement (4 or 5 points on the Likert scale). The Shapiro-Wilk test was used to test for normality. Primary and secondary outcomes were compared between experts and novices using the Mann-Whitney *U* test (nonparametric) or an independent 2-tailed *t* test (parametric). Interrater reliability was assessed using the intraclass correlation coefficient (ICC) for the modified H-OSATS score and the OSATS score. ICCs of less than 0.5 were regarded as poor, ICCs between 0.5 and 0.75 were regarded as moderate, ICCs between 0.75 and 0.9 were regarded as good, and ICCs of >0.9 were regarded as excellent [16]. A 2-sided *P* value of less than .05 was considered statistically significant. Statistical analysis was performed using Stata/BE (version 17.5.0; StataCorp) and SPSS (version 27).

Results

Study Population

A detailed summary of the baseline characteristics is shown in Table 1.

In addition to the baseline data, the prequestionnaire asked to what extent the participants were confident that they could perform a TLH without major errors. This question was answered with a median of 1 (IQR 1-1) ("disagree") in the novice group and a median of 5 (IQR 4-5) ("agree") in the expert group using a 5-point Likert scale.

Table . Baseline characteristics.

	Novices (n=12)	Experts (n=8)
Age (y), median (IQR)	30 (28.4-33.8)	42 (33.2-49.8)
Gender, n (%)		
Woman	10 (83.3)	5 (62.5)
Man	2 (16.7)	3 (37.5)
Educational level, n (%)		
Resident	8 (66.7)	0 (0.0)
Specialist	3 (25.0)	2 (25.0)
Attending	1 (8.3)	6 (75.0)
Work experience (y), median (IQR)	3.75 (0.5-7.3)	15 (6.25-23.3)
Prior training on surgical simulator, n (%)	7 (58.3)	6 (75.0)
Prior training course in MIS[a], n (%)	3 (25.0)	6 (75.0)
Number of laparoscopic procedures seen, median (IQR)	50 (28.75-250)	590 (300-2500)
Number of laparoscopic procedures assisted, median (IQR)	30 (5.25-87.5)	300 (200-1700)
Number of laparoscopic procedures performed, median (IQR)	7.5 (0-50)	225 (125-1675)
Number of TLHs[b] performed overall, median (IQR)	0 (0-0.75)	75 (11.5-150)
Number of TLHs performed in the last 5 years, median (IQR)	0 (0-0.75)	40 (10.5-100)

[a]MIS: minimally invasive surgery.

[b]TLH: total laparoscopic hysterectomy.

XSL•FO
RenderX

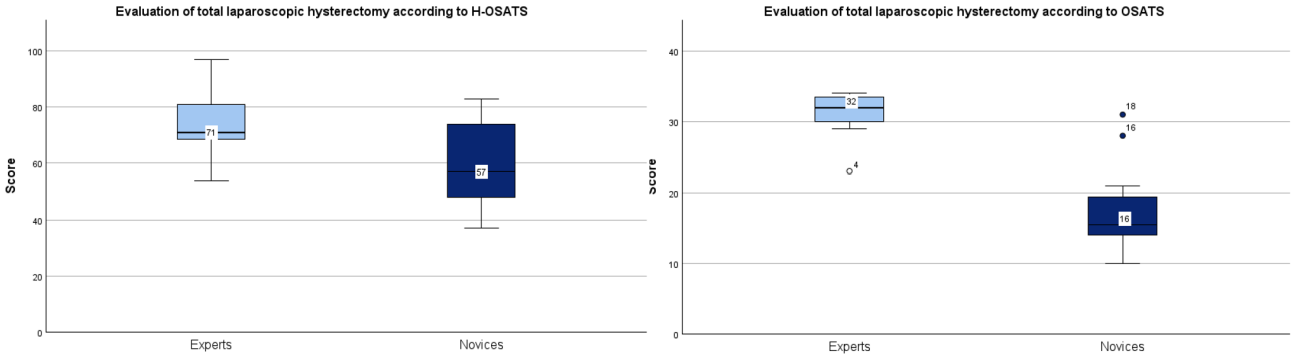
Suturing and Knot Tying on the Vaginal Cuff Model

The 3 suture and knot attempts on the vaginal tube model were assessed using the OSATS score [15]. The experts performed significantly faster (P<.001) and scored significantly better (P<.001; for details, see Multimedia Appendix 1).

Primary Outcome

The expert group obtained significantly higher procedure-specific modified H-OSATS scores compared to the novices (mean 74.0, SD 12.9 vs mean 60.3, SD 14.9; P=.049). Similar results were observed for the general OSATS score (P<.001), with a median of 71 (IQR 68.5-81) versus 57 (IQR 48-74; Figure 2). An excellent interrater reliability was observed for the H-OSATS score, with an ICC of 0.91 (P<.001), and a good interrater reliability was observed for the OSATS score, with an ICC of 0.81 (P<.001).

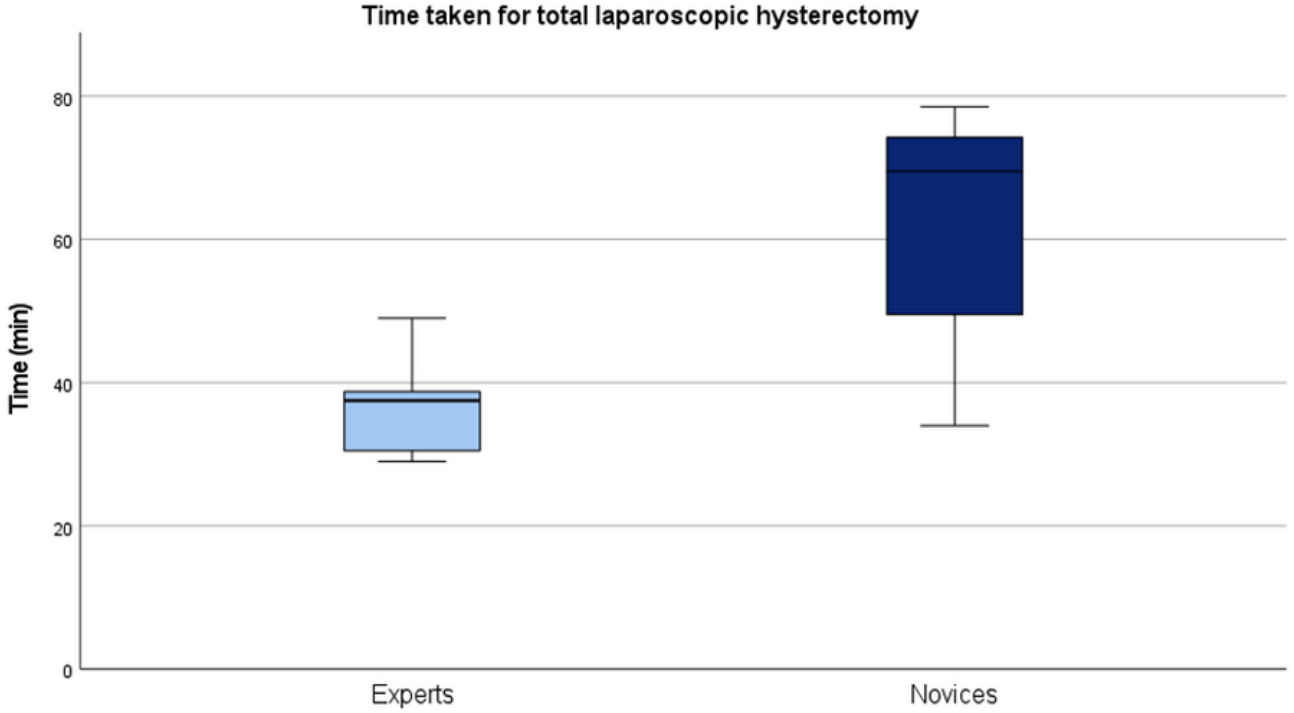
Figure 2. Comparison of Hysterectomy–Objective Structured Assessment of Technical Skills (H-OSATS) score and general Objective Structured Assessment of Technical Skills (OSATS) score between experts and novices.



Operating Time

Significantly prolonged operating times were recorded in the novice group compared to the expert group, with an 85.3% increase in operating time (novices: median 69.5, IQR 49.5-74.3 minutes; experts: median 37.5, IQR 30.5-38.8 minutes; P<.001; Figure 3).

Figure 3. Comparison of operating time for a total laparoscopic hysterectomy between experts and novices on the MaiSurge model.



Subjective Evaluation of the Uterus Model

The newly developed MaiSurge uterus model made of PVA was compared to the same model made from silicone, as well as a commercial uterus model from MedicFX (uterus and ovaries; Figure 4). Overall, the newly developed model was rated better or equally well compared to the silicone model and the commercial uterus model with regard to realism, color, consistency, size, thickness, tensile strength, proportions, and suitability for training (Figure 5).

XSL•FO
RenderX

Figure 4. Evaluated training models.

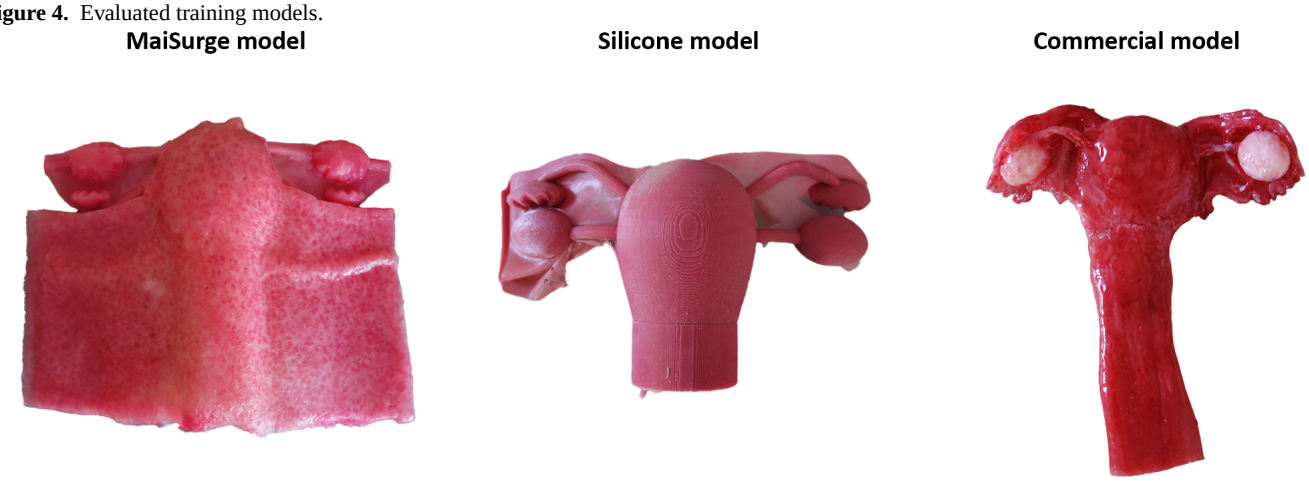
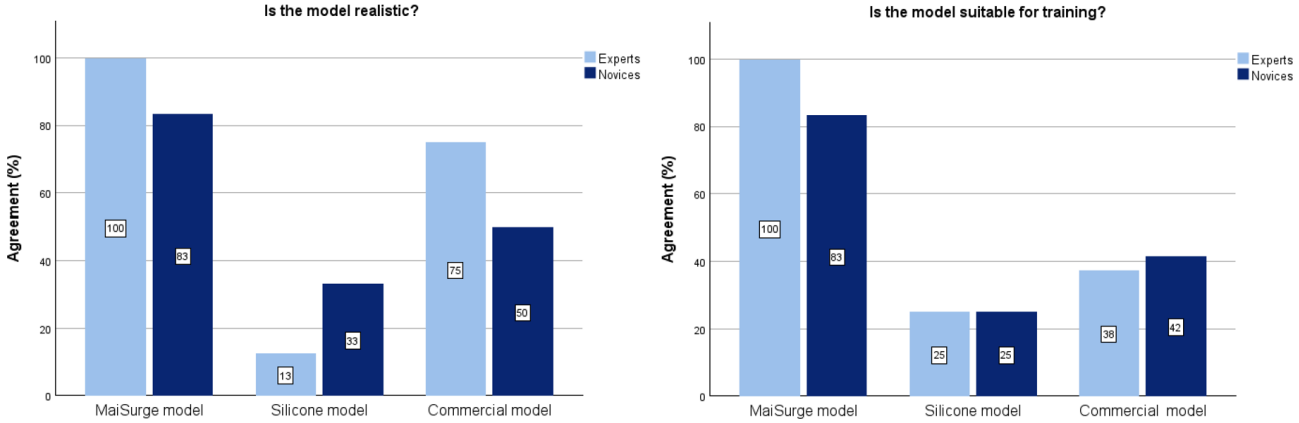


Figure 5. Comparison of different training models in terms of realism and suitability for training.

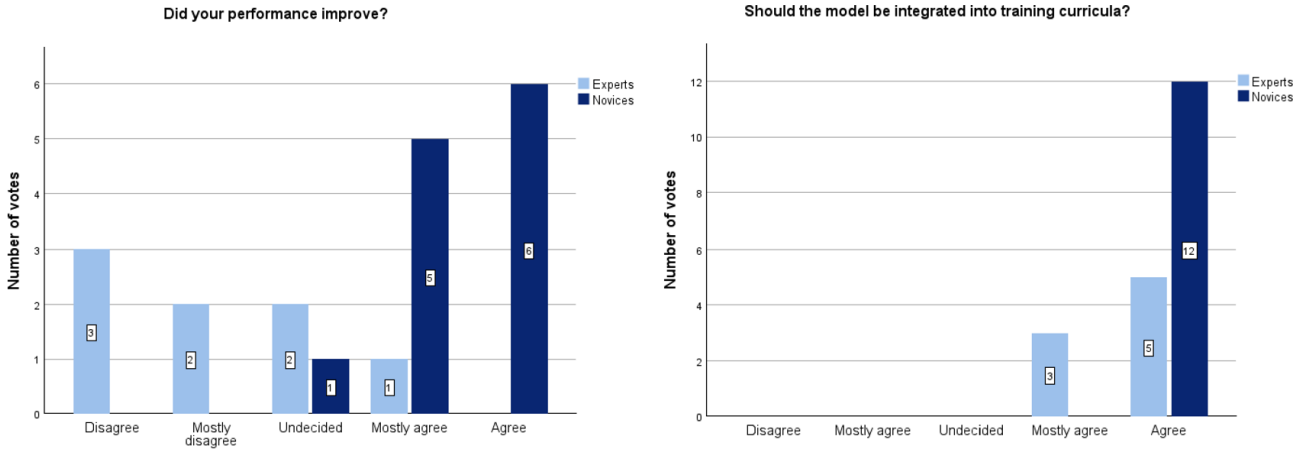


In terms of consistency, experts in particular identified areas for potential improvement of the MaiSurge model to increase the realism of the texture as it was often perceived to have less tensile strength compared to the real uterus tissue. The experts rated the model with a median of 3.5 (IQR 3.0-4.0) points on the Likert scale. In this regard, the commercial model scored slightly better, with a median of 4.0 (IQR 3.0-4.0) points. The novices gave the MaiSurge model a median score of 3.5 (IQR 3.0-4.0) points, with the other 2 models scoring lower.

Experts awarded the MaiSurge model a median score of 4.0 (IQR 2.25-4.0) points for tear resistance and, therefore, rated the model better than the other 2. Novices awarded the MaiSurge model a median score of 2.5 (IQR 1.25-4.0) points, which was considerably lower than the score from the experts.

While only 12.5% (1/8) of the experts reported improvement in their performance, approximately 92% (11/12) of the novices acknowledged that they had improved their surgical performance after training on the MaiSurge model. Overall, all participants agreed that the new MaiSurge uterus model should be integrated into training curricula to improve the performance of residents on TLHs (Figure 6).

Figure 6. Evaluation of subjective perception of skill improvement and use of the MaiSurge model for future training.



XSL•FO
RenderX

Discussion

In this study, the MaiSurge uterus model was developed and evaluated to realistically simulate a TLH.

Principal Findings

On the newly developed MaiSurge model, experts not only performed the TLH significantly faster than residents and fellows (approximately 85% longer operative times for the latter) but also achieved significantly higher performance scores as measured using the modified H-OSATS (approximately 23% higher scores) and general OSATS (approximately 24% higher scores) scales. Furthermore, the MaiSurge model can be used with conventional mechanical surgery as well as with electrosurgery and, thus, was rated as highly realistic by experienced gynecological surgeons and gynecological residents, especially when compared to widely available silicone models. All participants who completed the simulation TLH on the MaiSurge model agreed that the model was realistic in terms of color, size, thickness, proportions, and suitability for training. For inexperienced surgeons, the brittleness of the material posed a difficulty when suturing the vaginal stump as the model tended to tear out during "rough" movements. In addition, the thickness of the stump was an obstacle as it was difficult to push the needle through the material. However, despite the obstacles, the expert group was able to perform this surgical step without major problems. Furthermore, participants unanimously found the model to be a valuable tool to train residents, demonstrating its value and need in today's education of a future generation of gynecological surgeons.

Implications of the Findings

The number of gynecology inpatient procedures in Germany has declined in recent years, and the trend is likely to continue [17]. In 2023, the surgical catalog of outpatient procedures in Germany was adjusted to include more procedures that should be performed as outpatient procedures [18]. As a result, fewer procedures suitable for acquiring basic surgical skills will be performed in hospitals. Thus, especially in university hospitals, the number of training procedures per resident will decrease as the remaining surgical procedures are often too complex to be considered training procedures. Combined with the decreasing working hours due to improved labor laws, it will be difficult to fully guarantee adequate surgical training [17,18]. With the MaiSurge model, a realistic and affordable training tool was created to facilitate the surgical education of gynecological residents and fellows outside of the operating room. Regarding training model validity, evidence is needed before its implementation in training curricula. In this trial, gynecological experts outperformed novices with regard to time and performance scores, indicating that surgical skills when performing a TLH are adequately reflected through the use of the MaiSurge model. This provides attending physicians and program directors with the possibility to assess the current skill level of a gynecological resident. Furthermore, due to its ability to reflect surgical skill and its realism, the MaiSurge model could potentially be used as a test before residents are allowed to operate on a human patient. However, to achieve this, a benchmark with regard to time and performance scores needs to be set, which will require further research to evaluate a meaningful cutoff. We strongly believe that the MaiSurge model will be a necessary and feasible training tool to address the changes in surgical education ahead of us.

As expected, novices and experts found the model to be useful for the training of a TLH outside of the operating room. Novices reported a noticeable improvement in their performance after using the MaiSurge model. In contrast, experts did not report the same improvement as expected. However, this highlights that the MaiSurge model should be used at the beginning of a trainee's surgical learning curve rather than when a certain competency is reached so that they can benefit the most from the simulation. Since 2014, dry laboratory training is required by the European Union of Medical Specialists in every hospital offering laparoscopic procedures [19]. However, this is not yet a reality, which could be due to the required financial investments and lack of advanced training models. Thus, the MaiSurge uterus model is a good step toward meaningful and feasible dry laboratory training. As demonstrated by Klapdor et al [20], a structured skill laboratory training curriculum for gynecological residents showed a 16% decrease in operative time for a TLH after only 1 year. However, in the latter trial, a general laparoscopic skill training curriculum was used rather than a realistic uterus model, which leaves us imagining the potential impact that a realistic and procedure-specific training (eg, using the MaiSurge model) can have. This assumption is supported by the results of the study by Yang et al [10], who showed that a specific model for a pancreatoduodenectomy and training is better than just basic laparoscopic skill training.

Comparison to the Literature and Other Training Models

Basic laparoscopic skills can be taught through a variety of available models and trainers, which can be purchased for approximately €4000 (US $4697.44) (eg, the Lübeck Toolbox) [21], or by building one's own [22]. However, for more complex skills, advanced training models are needed. While some models for advanced skills, such as the closure of a vaginal cuff, can be cost-efficiently created from widely available building materials [23-25], they often lack realism and, thus, limit the extent of transferable skills to acquire. Purchasable options are often more realistic. However, they come at a price per vaginal cuff model that is often not suitable for large-scale training, such as those by LifeLike BioTissue (US $50) [26,27], Faux Medical (US $45) [28], or Medarchitect (US $40) [29]. Although it makes sense to train the surgical step of closing the vaginal cuff separately, a training model is needed that can cover as many surgical steps as possible to increase cost-efficiency. Only a few models for TLHs are available for purchase, such as a model from the company Oniko [30]. In addition to the high cost, this model is made of silicone and cannot be used with electrosurgery, which greatly impacts the realism of training and, thus, can potentially impact the skill transfer to the real operating room. To our knowledge, there is only 1 purchasable model from Limbs & Things that allows for electrosurgery during a TLH, which has an even higher cost of US $640 per uterus [31] and, thus, is not feasible for continuous training as only a few residency programs can provide each resident with multiple models at this cost.

XSL•FO
RenderX

With the MaiSurge model, some of these disadvantages of simulation models have been addressed. The MaiSurge model allows for the use of electrocautery, provides haptic resistance, and is relatively inexpensive to produce. The MaiSurge model developed in this study costs approximately €30 (US $35.23) (50 g of PVA per model) for the material plus 5 hours of work time (2 hours for boiling, 1 hour for casting the first layer, 1 hour for casting the second layer, and 1 hour for the peritoneum and vessels (boiling and casting); possibly less if mass produced). Once mass production is established, those costs will go down even further. Therefore, this model could be used in training courses, as well as for consistent training in residency programs.

Strengths and Limitations

The strengths of this study include the prospective, controlled trial design. Furthermore, 2 trained and blinded raters evaluated the primary outcome via video recordings, thus reducing the risk of assessment bias. A high interrater reliability was found, demonstrating the reproducibility of our results. With 75% (18/24) of the eligible physicians at the facility participating in the trial, including residents from every stage of training [32], we believe that the data presented in this study are representative of a university hospital in Germany and clearly reflect the potential use of the MaiSurge model to differentiate between experienced and inexperienced surgeons. However, it should be noted that the sample size of this trial was rather small, posing a risk of bias, but overall, the sample size is comparable to those of other studies developing new training models [23,33,34]. A follow-up multicenter trial is planned to assess transferability to the operating room. While the realism of the simulation of a TLH on the new MaiSurge model was rated highly, there are aspects in which the realism and, thus, the training experience could be further increased. This includes the simulation of blood flow in the model, as well as the possibility of removing the uterus vaginally (eg, if necessary with morcellation). Ultimately, the performance of the expert group could have been compromised by the few less realistic features of the model (eg, the traction on the tissue, which might have taken time to adjust to). While not explicitly assessed in this trial, a transfer of skills to the operating room can be expected based on current evidence of the transferability of surgical skills acquired in a skill laboratory and the high rating on realism and suitability from experienced gynecologists in this trial. Future randomized controlled trials could not only focus on the effect of training using the MaiSurge uterus model on intraoperative performances but also compare the training effects of the MaiSurge uterus model with those of commercially available and more expensive models such as the model by Limbs & Things [31]. While we were able to demonstrate the value of the MaiSurge uterus model in this trial, no objective comparison to similar and more expensive uterus models was drawn due to financial limitations.

One major advantage of the MaiSurge uterus model is the possibility to perform salpingectomies, adnexectomies, and TLHs on the same model, allowing inexperienced and experienced trainees to share a model to address their respective stages of training. Currently, the model is being developed further to include simulated bleedings, as well as uterine fibroids for laparoscopic and hysteroscopic resection. This will enhance the usefulness of the MaiSurge uterus model for more advanced training (eg, during fellowships or early stages of being an attending physician). Potentially individualized uterus models can be created based on real magnetic resonance imaging scans to simulate extremely difficult procedures and train a specific case before performing the same procedure on the actual patient. While this trial offers the validity evidence needed to use the MaiSurge model to assess a trainee's skill level, other questions remain unanswered. Future studies will need to focus on the learning curve of trainees on the MaiSurge model, the evaluation of other surgical procedures (eg, myomectomy or adnexectomy), clear benchmarks with regard to performance scores and time per procedure, and the transfer of the acquired skills to the operating room to not only confirm the model's great potential in surgical education but also offer a more detailed view on how to best incorporate the MaiSurge model into routine training.

Conclusions

Surgical simulation training is of growing importance to meet the requirements of patient safety and adequate surgical skill development during residency while addressing the growing financial pressure on hospitals. The newly developed MaiSurge uterus model presents a cost-effective, realistic, and feasible training option for TLHs, thus allowing for the simulation of even advanced surgical skills outside of the operating room during gynecological residency. Furthermore, the first validity evidence for the use of the model to assess the skill level of trainees was provided in this trial.

Acknowledgments

The authors would like to thank Susanne Gebhardt for her constant support in acquiring new materials for development.

Funding

No funding was received for the conduct of this trial.

Authors' Contributions

Conceptualization: MWS, AMB, CS; data curation MWS and AMB; formal analysis: AMB and MWS; investigation: AMB, CD, MWS, MS, and LS; methodology: MWS, AMB, CD, WB, RS, AH, and KG; project administration: MWS and AMB; resources: MWS, AH, and WB; supervision: MWS, AH, RS, WB, CS; validation: MWS, AMB, and CD; visualization: AMB and MWS; writing - original draft: AMB and MWS; writing: review & editing: CD, WB, RS, AH, KG, LS, CS.

XSL•FO
RenderX

Conflicts of Interest

RS reports honoraria and expenses from Roche Pharma AG, Sanofi, and AstraZeneca. LS reports honoraria and expenses from AstraZeneca. KG reports paid lectures from Clovis Oncology, AstraZeneca, Pharmamar, MSD, GSK, and Eisai. AH reports honoraria and expenses from AstraZeneca, Celgen, GSK, LEO Pharma, MedConcept, med update, Medicultus, Pfizer, Promedicis, Softconsult, Streamed Up, Roche Pharma AG, and Tesaro Bio Germany, as well as work as a consultant to AstraZeneca, GSK, MSD, Pharmamar, Promedicis, Roche Pharma AG, and Tesaro Bio Germany. None were related to this trial. All other authors declare no other conflicts of interest.

Multimedia Appendix 1
Supplementary material, including study flowchart, modified H-OSATS description and score, and comparison of suturing time and OSATS score between novices and experts. H-OSATS: Hysterectomy–Objective Structured Assessment of Technical Skills Score; OSATS: Objective Structured Assessment of Technical Skills Score.
[DOCX File, 233 KB - mededu_v12i1e66369_app1.docx]

References

1. Prütz F, Knopf H, von der Lippe E, Scheidt-Nave C, Starker A, Fuchs J. Prevalence of hysterectomy in women 18 to 79 years old: results of the German health interview and examination survey for adults (DEGS1) [Article in German]. Bundesgesundheitsblatt Gesundheitsforschung Gesundheitsschutz 2013 May;56(5-6):716-722. [doi: 10.1007/s00103-012-1660-7] [Medline: 23703490]
2. Harvey SV, Pfeiffer RM, Landy R, Wentzensen N, Clarke MA. Trends and predictors of hysterectomy prevalence among women in the United States. Am J Obstet Gynecol 2022 Oct;227(4):611. [doi: 10.1016/j.ajog.2022.06.028]
3. Wu JM, Wechter ME, Geller EJ, Nguyen TV, Visco AG. Hysterectomy rates in the United States, 2003. Obstetric Gynecol 2007;110(5):1091-1095. [doi: 10.1097/01.AOG.0000285997.38553.4b]
4. Marchand G, Masoud AT, Abdelsattar A, et al. Meta-analysis of laparoscopic radical hysterectomy, excluding robotic assisted versus open radical hysterectomy for early stage cervical cancer. Sci Rep 2023 Jan 6;13(1):273. [doi: 10.1038/s41598-023-27430-9] [Medline: 36609438]
5. Tantitamit T, Huang KG, Lee CL. Laparoscopic versus open radical hysterectomy in women with early stage cervical cancer: a systematic review and meta-analysis. Taiwan J Obstet Gynecol 2020 Jul;59(4):481-488. [doi: 10.1016/j.tjog.2020.05.003] [Medline: 32653117]
6. Peláez Mata D, Herrero Álvarez S, Gómez Sánchez A, Pérez Egido L, Corona Bellostas C, de Agustín Asensio JC. Laparoscopic learning curves. Cir Pediatr 2021 Jan 1;34(1):20-27. [Medline: 33507640]
7. Igwe E, Hernandez E, Rose S, Uppal S. Resident participation in laparoscopic hysterectomy: impact of trainee involvement on operative times and surgical outcomes. Am J Obstet Gynecol 2014 Nov;211(5):484. [doi: 10.1016/j.ajog.2014.06.024] [Medline: 24949539]
8. Jokinen E, Brummer T, Jalkanen J, et al. Hysterectomies in Finland in 1990-2012: comparison of outcomes between trainees and specialists. Acta Obstet Gynecol Scand 2015 Jul;94(7):701-707. [doi: 10.1111/aogs.12654] [Medline: 25867490]
9. Dawe SR, Windsor JA, Broeders J, Cregan PC, Hewett PJ, Maddern GJ. A systematic review of surgical skills transfer after simulation-based training: laparoscopic cholecystectomy and endoscopy. Ann Surg 2014 Feb;259(2):236-248. [doi: 10.1097/SLA.0000000000000245] [Medline: 24100339]
10. Yang J, Luo P, Wang Z, Shen J. Simulation training of laparoscopic pancreaticojejunostomy and stepwise training program on a 3D-printed model. Int J Surg 2022 Nov;107:106958. [doi: 10.1016/j.ijsu.2022.106958] [Medline: 36283653]
11. Burkett LS, Makin J, Ackenbom M, Artsen A, Bradley M. Validation of transvaginal hysterectomy surgical model – modification of the flowerpot model to improve vesicovaginal plane simulation. J Minim Invasive Gynecol 2021 Aug;28(8):1526-1530. [doi: 10.1016/j.jmig.2020.12.022] [Medline: 33359289]
12. Karadza E, Haney CM, Limen EF, et al. Development of biotissue training models for anastomotic suturing in pancreatic surgery. HPB (Oxford) 2023 Jun;25(6):625-635. [doi: 10.1016/j.hpb.2023.02.002] [Medline: 36828741]
13. Knight S, Aggarwal R, Agostini A, Loundou A, Berdah S, Crochet P. Development of an objective assessment tool for total laparoscopic hysterectomy: a Delphi method among experts and evaluation on a virtual reality simulator. PLoS ONE 2018;13(1):e0190580. [doi: 10.1371/journal.pone.0190580] [Medline: 29293635]
14. Martin JA, Regehr G, Reznick R, et al. Objective structured assessment of technical skill (OSATS) for surgical residents. Br J Surg 1997 Feb;84(2):273-278. [doi: 10.1046/j.1365-2168.1997.02502.x] [Medline: 9052454]
15. Chang OH, King LP, Modest AM, Hur HC. Developing an objective structured assessment of technical skills for laparoscopic suturing and intracorporeal knot tying. J Surg Educ 2016;73(2):258-263. [doi: 10.1016/j.jsurg.2015.10.006] [Medline: 26597729]
16. Koo TK, Li MY. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. J Chiropr Med 2016 Jun;15(2):155-163. [doi: 10.1016/j.jcm.2016.02.012] [Medline: 27330520]
17. Augurzky B, Kreienberg R, Mennicken R. Zukunft der stationären Versorgung in Gynäkologie und Geburtshilfe [Article in German]. Gynäkologe 2015 Jul;48(7):495-500. [doi: 10.1007/s00129-015-3722-z]

XSL•FO
RenderX

18. Stoff-Ahnis S. Ambulantisierungsreform: Nächster Schritt ist geschafft [Web page in German]. GKV-Spitzenverband. 2023. URL: https://www.gkv-spitzenverband.de/gkv_spitzenverband/presse/pressemitteilungen_und_statements/pressemitteilung_1727488.jsp [accessed 2024-02-10]
19. UEMS European training requirements. European Union of Medical Specialists. 2018. URL: https://www.uems.eu/european-training-requirements [accessed 2023-11-23]
20. Klapdor R, et al. StepOPgyn: Ein Curriculum für die strukturierte operative Ausbildung [Article in German]. Geburtshilfe Frauenheilkunde 2023;82(10):e104-e105. [doi: 10.1055/s-0042-1756898]
21. Thinggaard E, Bjerrum F, Strandbygaard J, Gögenur I, Konge L. Validity of a cross-specialty test in basic laparoscopic techniques (TABLT). Br J Surg 2015 Aug;102(9):1106-1113. [doi: 10.1002/bjs.9857] [Medline: 26095788]
22. Lee M, Savage J, Dias M, Bergersen P, Winter M. Box, cable and smartphone: a simple laparoscopic trainer. Clin Teach 2015 Dec;12(6):384-388. [doi: 10.1111/tct.12380] [Medline: 26135242]
23. Tunitsky-Bitton E, Propst K, Muffly T. Development and validation of a laparoscopic hysterectomy cuff closure simulation model for surgical training. Am J Obstet Gynecol 2016 Mar;214(3):392. [doi: 10.1016/j.ajog.2015.11.023] [Medline: 26640072]
24. Fuchs Weizman N, Manoucheri E, Vitonis AF, Hicks GJ, Einarsson JI, Cohen SL. Design and validation of a novel assessment tool for laparoscopic suturing of the vaginal cuff during hysterectomy. J Surg Educ 2015;72(2):212-219. [doi: 10.1016/j.jsurg.2014.08.015] [Medline: 25439178]
25. King CR, Donnellan N, Guido R, Ecker A, Althouse AD, Mansuria S. Development and validation of a laparoscopic simulation model for suturing the vaginal cuff. Obstetric Gynecol 2015;126(Supplement 4):27S-35S. [doi: 10.1097/AOG.0000000000001053]
26. Vaginal cuff. LifeLike BioTissue. 2023. URL: https://lifelikebiotissue.com/shop/obgyn-urology/vaginal-cuff [accessed 2023-10-19]
27. Vaginal cuff holder. LifeLike BioTissue. URL: https://lifelikebiotissue.com/shop/obgyn-urology/vaginal-cuff%2019.10.23%2016:36%20Uhr%20https:/lifelikebiotissue.com/shop/obgyn/vaginal-cuff-holder [accessed 2023-10-19]
28. Faux vaginal cuff model - laparoscopic training. Faux Medical. URL: https://www.fauxmedical.com/products/faux-vaginal-cuff-model-laparoscopic-training [accessed 2023-10-19]
29. Silicone vaginal cuff model laparoscopic suture pad simulator practice kit for doctors, medical students, veterinarians. Medarchitect. URL: https://mediarchitect.net/en-de/products/silicone-vaginal-cuff-model-laparoscopic-suture-pad-simulator-practice-kit-for-doctors-medical-students-veterinarians [accessed 2023-10-19]
30. Surgical female pelvic trainer. Oniko. URL: https://oniko.ua/en/products/simulators-and-trainers/hirurgiya/laparoskopiya/model-ginochogo-tazu-dlya-trenuvan-hirurgichnih-navichok.html#callform [accessed 2023-11-10]
31. SFPT mk 2 uterus. Limbs & Things. URL: https://limbsandthings.com/us/products/60252/60252-sfpt-mk-2-uterus#tab1 [accessed 2023-11-10]
32. Browning CM, Cloutier R, Rich TC, Leavesley SJ. Endoscopy lifetime systems architecture: scoping out the past to diagnose the future technology. Systems (Basel) 2022 Oct;10(5):189. [doi: 10.3390/systems10050189] [Medline: 36330206]
33. Greer JA, Segal S, Salva CR, Arya LA. Development and validation of simulation training for vaginal hysterectomy. J Minim Invasive Gynecol 2014;21(1):74-82. [doi: 10.1016/j.jmig.2013.06.006] [Medline: 23850966]
34. Miyazaki D, Matthews CA, Kia MV, El Haraki AS, Miyazaki N, Chen CCG. Validation of an educational simulation model for vaginal hysterectomy training: a pilot study. Int Urogynecol J 2019 Aug;30(8):1329-1336. [doi: 10.1007/s00192-018-3761-9] [Medline: 30191250]

Abbreviations

H-OSATS: Hysterectomy–Objective Structured Assessment of Technical Skills Score
ICC: intraclass correlation coefficient
OSATS: Objective Structured Assessment of Technical Skills Score
PVA: polyvinyl alcohol
TLH: total laparoscopic hysterectomy

XSL•FO
RenderX

Edited by B Lesselroth; submitted 19.Sep.2024; peer-reviewed by JP Bereuter, J Wang, W Isono; revised version received 05.May.2025; accepted 22.Sep.2025; published 12.Feb.2026.

Please cite as:
Brechter AM, Schwab R, Dold C, Skala C, Schröder M, Schiestl L, Gillen K, Brenner W, Hasenburg A, Schmidt MW
Developing a Realistic and Cost-Effective Training Model (MaiSurge) for Laparoscopic Hysterectomies to Train and Assess Surgical Skill: Prospective Nonrandomized Controlled Trial
JMIR Med Educ 2026;12:e66369
URL: https://mededu.jmir.org/2026/1/e66369
doi:10.2196/66369

© Anna Maria Brechter, Roxana Schwab, Christoph Dold, Christine Skala, Maria Schröder, Lina Schiestl, Katharina Gillen, Walburgis Brenner, Annette Hasenburg, Mona Wanda Schmidt. Originally published in JMIR Medical Education (https://mededu.jmir.org), 12.Feb.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Education, is properly cited. The complete bibliographic information, a link to the original publication on https://mededu.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Targeted Educational Intervention Through Game-Based Learning to Promote Rational Antimicrobial Use Among Health Care Learners: Prospective Interventional Study

Sumana MN, MD; Supreeta R Shettar, MSc; Yogeesh D Maheshwarappa, MSc; GK Megha, MSc; Veerabhadra Swamy GS, MSc; Chinchana Shylaja Eshwar, MSc; Shruthi Shree SC, MSc

Department of Microbiology, JSS Medical College, JSS Academy of Higher Education and Research, Mysuru, India

Corresponding Author:
GK Megha, MSc
Department of Microbiology, JSS Medical College, JSS Academy of Higher Education and Research, Mysuru, India

Abstract

Background: Antimicrobial resistance (AMR) is a global problem. Training health care professionals in the rational use of antimicrobials is essential to curb AMR.

Objective: To support efforts to reduce antibiotic resistance, this study assesses how well a gamified educational intervention might improve health care professionals' and students' understanding and use of appropriate antibiotics.

Methods: This is a prospective interventional study conducted for clinical practitioners, undergraduates (MBBS and interns), postgraduates, and pharmacy students. A total of 60 participants were included in the study. Innovative games were administered to support the management of infections across multiple body systems, in accordance with the 2022 Indian Council of Medical Research treatment guidelines and the latest Infectious Diseases Society of America guidelines, incorporating multiple instructional components. Pretest and posttest questionnaires were administered and evaluated.

Results: After the intervention, participants' ability to differentiate between bacterial and viral symptoms in respiratory tract infections and gastroenteritis improved from 48% to 94%. The practice of selecting the appropriate empirical antimicrobial at the correct dose, route, and duration also demonstrated significant improvement, reflecting enhanced adherence to principles of rational antimicrobial use.

Conclusions: The gamified intervention successfully improved participants' knowledge and awareness of rational antimicrobial use. Substantial improvements across all the assessed components highlight the positive impact of the intervention in promoting optimal antimicrobial use and curbing AMR. Innovative gamified interventions may foster better and longer-lasting awareness, supporting appropriate antimicrobial use.

(JMIR Med Educ 2026;12:e72236) doi:10.2196/72236

KEYWORDS

gamified intervention; antimicrobial resistance; rational antimicrobial prescribing; antimicrobial stewardship; Indian Council of Medical Research antimicrobial guidelines 2022; medical education; experiential learning; infection management; game-based learning; clinical decision-making

Introduction

Antimicrobial resistance (AMR) is one of this century's most alarming challenges to human health. With the AMR crisis peaking, experts warn that the era of ineffective antibiotics is imminent. The World Health Organization estimates that about half of antimicrobial prescriptions are inappropriate and that about half of patients do not complete their prescribed treatments [1,2].

The rise in AMR is particularly detrimental to low- and middle-income countries, undermining health and development policies aimed at eliminating severe poverty by 2030. According to the World Bank, the financial impact of AMR on health care could range from US $300 billion to over US $1 trillion by 2050. In India, the National Action Plan on Antimicrobial Resistance, launched in April 2017, outlines priorities that are in alignment with the World Health Organization Global Action Plan. The first strategy of the action plan focuses on improving awareness and understanding of AMR through effective communication, education, and training [3,4].

A significant challenge in antimicrobial stewardship (AMS) programs is the extensive knowledge students must quickly acquire. AMR remains a concept frequently misunderstood; several health care professionals, including pharmacists and dentists, incorrectly assume that resistance develops in individuals instead of microorganisms. The current knowledge

XSL•FO
RenderX

of optimal antimicrobial use among medical students and practitioners is insufficient, presenting a challenge for educators to develop effective teaching strategies for AMR and AMS [5,6].

In this context, gamification—the incorporation of game features in non-game environments—has emerged as a viable educational instrument in health professions education. Rooted in Kolb's experiential learning theory, gamification fosters active engagement by enabling learners to apply knowledge in realistic, consequence-driven settings. In addition, Bloom's taxonomy suggests that game-based learning may enhance higher-order cognitive abilities, including application, analysis, and assessment. Through the incorporation of competition, problem-solving, and feedback, games align with self-determination theory, which highlights autonomy, competence, and relatedness as fundamental motivators for learning [7-9].

Recently, games have been increasingly used in AMS programs, showing greater effectiveness compared with traditional teaching methods [10-12]. Games can enhance learners' interest in the subject, reinforce prior knowledge, and create a competitive and interactive learning environment. Consequently, games are seen as a valuable addition to current health care education efforts, addressing the shortcomings of conventional teaching techniques [13,14]. This study aims to explore how games can improve knowledge about antibiotic prescription practices by providing an engaging and interactive learning environment.

Methods

Overview

This prospective interventional study was conducted for clinical practitioners, undergraduates (MBBS and interns), and postgraduate medical and pharmacy students of Mysore Medical College, Mysuru, South India. Participants were selected using convenience sampling based on availability and willingness to participate. A total of 60 individuals were included and randomly allocated to 2 equal teams for activity-based learning; however, randomization was not stratified by discipline or prior knowledge. The aim was to assess the efficacy of a gamified intervention on AMS knowledge. The pretest questionnaire was first administered to each participant and took over 20 minutes to complete.

The intervention began with an introduction to the rational use of antibiotics and the importance of AMS interventions in a tertiary hospital setting, provided through a brief orally delivered educational intervention. This was followed by the administration of innovative games designed for the management of infections of the respiratory tract, gastrointestinal tract, urinary tract, skin and soft tissue, acute undifferentiated fever, and sepsis, in accordance with the 2022 Indian Council of Medical Research antimicrobial treatment guidelines and the latest Infectious Disease Society of America treatment guidelines. The games were designed to help learners differentiate between bacterial and viral features in common infections, such as respiratory tract infections and gastroenteritis, as most often these infections are caused by viruses yet are treated with antibiotics; to understand and select appropriate empirical antimicrobial therapy, including choosing the correct agent, dose, and duration based on the severity of the infection; to promote the rational use of combination antibiotic therapy; and to recognize antibiotics that are ineffective at certain sites of infection based on their pharmacodynamic and pharmacokinetic properties.

A posttest questionnaire was administered to assess the effectiveness of the intervention. The normality of pretest and posttest score distributions were determined using the Shapiro-Wilk test. All components demonstrated approximately normal distributions, permitting the use of paired *t* tests.

Questionnaire

A 30-item multiple-choice questionnaire covering systemic infections was designed, with a total score of 150 points (5 points per item with 25 points per component). The questions focused on appropriate empirical antimicrobial selection, including choice of agent, dose, and duration for specific infections; antibiotics to be avoided in certain conditions based on pharmacokinetic and pharmacodynamic properties; and intrinsic resistance. Participants completed the questionnaire independently over a 20-minute period without the use of reference materials, notes, or assistance, thereby allowing assessment of baseline knowledge (Multimedia Appendix 1).

Game Design

The gamified educational intervention comprised a series of structured games designed to address key principles of rational antimicrobial use. Games were used to differentiate between bacterial and viral infections of the respiratory and gastrointestinal tracts. Differentiation in approximately 30% to 50% of cases can help avoid unnecessary antibiotic treatment, which is a step forward in addressing AMR. For this purpose, Basketing the Ball and Monkeying with Donkey games were used (Multimedia Appendix 2). To select the appropriate antibiotic, the games carom, darts, Where in Venn, and musical chairs were used. The rational use of 2 or more antibiotics was taught using the Death by Double Strike game. Selection of antimicrobials from lower- to higher-spectrum agents based on infection severity was taught using the games Hierarchy by Severity, Build your Burj Khalifa, and Lagori. Selecting appropriate antimicrobials based on the site of infection was taught using the games Drop the Dropout and Push to Pandora. To understand the appropriate dosage and duration of antibiotics, the games Fun with Numbers, Pick Your Partner Right, and Monkeying with Donkey were used.

Respiratory Tract Infection

To understand the rational use of antimicrobials for respiratory tract infections, Basketing the Ball was played to differentiate the features of viral and bacterial pharyngitis, sinusitis, and lower respiratory tract infection (LRTI). Fun With Numbers was played to select the appropriate antibiotic dose and duration of treatment for bacterial pharyngitis and LRTI. Death by Double Strike was played to select the appropriate combination of antibiotics for community-acquired pneumonia treatment. Hierarchy was played to select the appropriate antibiotics for LRTI treatment based on infection severity. Darts were played

XSL•FO
RenderX

to understand which antibiotics should be used and which should be avoided in LRTI.

Gastrointestinal and Intra-Abdominal Infections

To understand the rational use of antimicrobials in gastroenteritis, Monkeying with Donkey was played to segregate clinical features of gastroenteritis across different clinical scenarios to prevent unnecessary antimicrobial use. Pick Your Partner Right was played to select the appropriate antibiotic, dose, and duration for the treatment of bacillary dysentery, cholera, amoebiasis, and giardiasis. Build Your Burj Khalifa was played to select appropriate antibiotics for the treatment of mild, moderate, and severe community-acquired intra-abdominal infections and hospital-acquired intra-abdominal infections progressing from lower- to higher-spectrum antimicrobials. This approach could prevent the use of higher-spectrum antimicrobials for milder infections and lower-spectrum antibiotics for severe infections, thereby improving patient outcomes and cost-effectiveness. It may also slow the development of evolving AMR.

Urinary Tract Infections

To understand the rational use of antimicrobials in urinary tract infections, picture puzzles were played to understand the treatment of uncomplicated cystitis. Mnemonics were created for treatment options of pyelonephritis, prostatitis, and epididymo-orchitis to facilitate easy recall of empirical antibiotic choices during clinical decision-making. For pyelonephritis, the mnemonic AIM-PE corresponds to amikacin, imipenem, meropenem, piperacillin–tazobactam, and ertapenem—agents commonly used in moderate to severe infections and when resistant gram-negative organisms are suspected. For epididymo-orchitis, the mnemonic COLD corresponds to ceftriaxone, ofloxacin, levofloxacin, and doxycycline, covering likely sexually transmitted and enteric pathogens depending on patient profile. These mnemonics were integrated into interactive educational games to enhance retention, promote rational antimicrobial selection, and strengthen applied learning among health care students. The Fun with Numbers game was used to understand the dosage and duration of antibiotics used for uncomplicated cystitis and pyelonephritis.

Acute Undifferentiated Fever

To understand the rational use of antimicrobials in acute undifferentiated fever, the Loop Game was played to understand the causative agents of fever with jaundice, fever with sore throat, and fever with rash in acute undifferentiated fever. The Where in Venn game was played to select the appropriate antibiotic for enteric fever, rickettsial fever, and leptospirosis.

Ethical Considerations

The study protocol was reviewed and approved by the Institutional Ethical Committee, JSS Medical College, Mysuru (JSS/MC/PG/192). Participant privacy and confidentiality were strictly maintained throughout the study by deidentifying data and using unique codes in place of personal identifiers. Only study investigators had access to the coded dataset, and all records were stored securely in accordance with institutional policies. No financial compensation was provided to participants for their involvement in the study. All procedures were conducted in accordance with the ethical standards of the Declaration of Helsinki and applicable regulatory requirements.

Results

In this interventional study involving 60 participants, including clinical practitioners, undergraduates (MBBS and interns), postgraduate doctors, and pharmacy students, we implemented a structured approach consisting of pretest and posttest assessments flanking an educational intervention. The intervention used innovative educational games designed for this study. These games were designed to cover a comprehensive array of topics crucial to antimicrobial therapy. They encompassed empiric and specific antimicrobial treatments, the intricate dynamics of pharmacokinetics and pharmacodynamics of antibiotics, optimal dosages and durations for antibiotic regimens, intrinsic antibiotic resistance, effective antifungal strategies, combined antibiotic therapies, and the use of biomarkers specific to various infections such as those of the respiratory tract, gastrointestinal tract, urinary tract, skin and soft tissues, and bloodstream (including sepsis), and acute undifferentiated fever.

Pretest analysis revealed that, of the 60 participants, the proportion of participants that answered at least one question correctly was 41.7% (n=25) for respiratory tract infections, 40% (n=24) for gastrointestinal tract infections, 35% (n=21) for urinary tract infections, 26.7% (n=16) for skin and soft tissue infections, 18.3% (n=11) for bloodstream infections (including sepsis), and 36.7% (n=22) for acute undifferentiated fever. Posttest analysis showed significant improvements, with the proportion of participants that answered at least one question correctly for respiratory tract infections increased to 93.3% (n=56), for gastrointestinal tract infections increased to 80% (n=48), for urinary tract infections increased to 71.7% (n=43), for skin and soft tissue infections increased to 66.7% (n=40), for bloodstream infections (including sepsis) increased to 73.3% (n=44), and for acute undifferentiated fever increased to 81.6% (n=49).

The following 6 preintervention and postintervention questionnaire components were assessed:

1. Respiratory tract infections
2. Gastrointestinal and intra-abdominal infections
3. Urinary tract infections
4. Skin and soft tissue infections
5. Bloodstream infections (including sepsis)
6. Acute undifferentiated fever

Table 1 shows the study results assessing changes in the 6 components from pretest to posttest. Each component demonstrated an increase in the proportion of participants who answered at least one question correctly after the intervention. In component 1, the posttest mean score of 10.87 (SD 7.63) is slightly lower than the pretest mean score of 11.62 (SD 9.02), but the improvement in participants with adequate knowledge (from n=25, 41.7% to n=56, 93.3%) was statistically significant (P=.02). Component 2 shows a more pronounced increase in participants with adequate knowledge (from n=24, 40% to n=48, 80%), with mean scores decreasing from 8.56 (SD 5.62) to 7.56

XSL•FO
RenderX

(SD 2.51); a highly significant *P* value less than .001 indicates a substantial difference. Component 3 also saw improvement in the number of participants with adequate knowledge (from n=21, 35% to n=43, 71.7%), which was significant (*P*=.007), although the mean score slightly decreased from 14.56 (SD 8.51) to 12.37 (SD 10.42). Component 4 demonstrated an increase in the number of participants with adequate knowledge (from n=16, 26.7% to n=40, 66.7%), with a highly significant *P* value less than .001, despite a decrease in the mean score from 18.50 (SD 25.55) to 15.5 (SD 14.34). Component 5 saw a notable increase in participant percentage score (from n=11, 18.3% to n=44, 73.3%), with a slight mean score increase from 21.37 (SD 18.56) to 22.43 (SD 17.60) and a significant *P* value of .02. Finally, component 6 showed a participant percentage score increase (from n=22, 36.7% to n=49, 81.6%), with a marginal mean increase (from 19.89, SD 17.51 to 20.17, SD 1.27) and a highly significant *P* value less than .001.

Table . Changes in the 6 components from pretest to posttest.

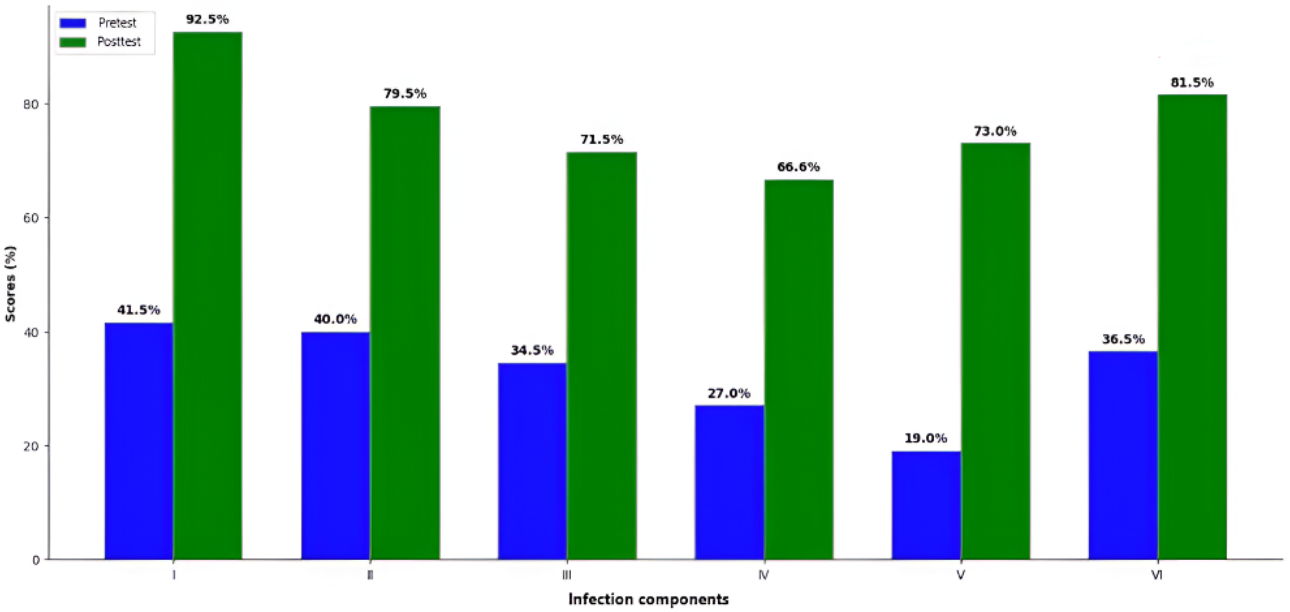
Component	Score range	Pretest		Posttest		*P* value	Cohen *d*	Significance
		Participants, n (%)[a]	Score, mean (SD; 95% CI)	Participants, n (%)[a]	Score, mean (SD; 95% CI)			
1	0-25	25 (41.7)	11.62 (9.02; 9.3-13.9)	56 (93.3)	10.87 (7.63; 8.9-12.8)	.02	0.56	Significant
2	0-25	24 (40.0)	8.56 (5.62; 7.1-10.0)	48 (80.0)	7.56 (2.51; 6.9-8.2)	<.001	0.66	Highly significant
3	0-25	21 (35.0)	14.56 (8.51; 12.3-16.8)	43 (71.7)	12.37 (10.42; 9.6-15.1)	.007	0.42	Significant
4	0-25	16 (26.7)	18.50 (25.55; 12.1-24.9)	40 (66.7)	15.5 (14.34; 11.8-19.2)	<.001	0.49	Highly significant
5	0-25	11 (18.3)	21.37 (18.56; 16.6-26.1)	44 (73.3)	22.43 (17.60; 17.9-26.9)	.02	0.58	Significant
6	0-25	22 (36.7)	19.89 (17.51; 15.2-24.5)	49 (81.6)	20.17 (1.27; 19.8-20.6)	<.001	0.33	Highly significant

[a]Participants who answered at least one question correctly in each component.

Overall, these results suggest that the intervention was effective across components, with statistically significant improvements in participant scores and response rates. The test improvements are represented in Figure 1. Paired *t* tests were used to compare pretest and posttest results for each component. A *P* value less than .05 was judged as statistically significant. The mean differences were determined using a 95% CI. To quantify the magnitude of change, effect sizes were calculated using Cohen *d*. Even though posttest success rates increased dramatically, minor decreases in mean scores in specific components (eg, components 1 and 3) might be attributable to a greater number of participants scoring near the passing threshold rather than achieving consistently high scores. This pattern suggests a broader distribution of correct responses, but with considerable variety in depth of knowledge. This difference might be ascribed to the variation in question difficulty and the distribution of incomplete tries among several items within a component's section. Although more participants selected correct answers overall, individual scores may have been lower because scoring required completely correct answers (no partial credit), and a small number of low posttest outliers may have influenced the mean.

XSL•FO
RenderX

Figure 1. Pretest and posttest improvements in the number of participants who answered at least one question correctly.



Participants reported heightened engagement and interest in the subject matter due to the interactive and competitive nature of the games, favoring this approach over traditional lectures for retaining complex information. This study underscores that innovative games effectively enhance knowledge and understanding of antimicrobial therapy among medical students and faculty. This method shows promise as a valuable supplement to traditional educational techniques, addressing the limitations of passive learning strategies. The substantial improvements in posttest scores across all infection categories highlight the effectiveness of this interactive approach in enhancing comprehension and retention of complex AMR concepts.

Discussion

The findings of this interventional study demonstrate the efficacy of game-based learning tools in improving AMS knowledge and promoting rational antibiotic use across diverse clinical contexts. The gamified intervention was developed according to experiential learning theory (Kolb), encouraging participants to apply clinical decision-making skills in virtual environments. This approach incorporates competition, peer interaction, and immediate feedback, and draws on self-determination theory to promote autonomy, engagement, and motivation. Learning goals were aligned with the Bloom taxonomy, focusing on the application, analysis, and assessment stages of cognition. A total of 60 participants, including undergraduate students, interns, postgraduate physicians, clinical practitioners, and pharmacy students, completed pretest and posttest assessments for this educational intervention that used purpose-designed instructional games.

There were significant improvements across all 6 infection types. For respiratory tract infections, the percentage of participants with appropriate antibiotic knowledge increased from 41.7% (25/60) in the pretest to 93.3% (56/60) in the posttest. Although the mean score decreased from 11.62 (SD 9.02) to 10.87 (SD 7.63), this change was statistically significant (P=.02), indicating it had a meaningful educational effect. This study supports prior data promoting gamification as an effective approach for teaching complex treatment concepts and facilitating active memory [13-15].

In urinary tract infections, the proportion of participants with appropriate knowledge increased from 35% (21/60) to 71.7% (43/60), whereas the mean score decreased from 14.56 (SD 8.51) to 12.37 (SD 10.42). The change remained statistically significant (P=.007), indicating effective knowledge translation. In skin and soft tissue infections, similar trends were seen: the proportion of correct responses went from 26.7% (16/60) to 66.7% (40/60), and the average score went from 18.50 (SD 25.55) to 15.5 (SD 14.34; P<.001) [16-18].

There was a significant improvement in bloodstream infections, with the proportion of participants with appropriate knowledge increasing from 18.3% (11/60) to 73.3% (44/60). The average score went up a little, from 21.37 (SD 18.56) to 22.43 (SD 17.60; P=.02). This finding supports the conclusion that participants not only learned to identify appropriate antibiotic regimens but also retained their quantitative knowledge. These results highlight how important interactive teaching methods are for conveying AMS care fundamentals [19,20].

In cases of acute undifferentiated fever, the proportion of participants with appropriate knowledge increased from 36.7% (22/60) to 81.6% (49/60). There was a statistically significant change (P<.001) in the mean scores, which increased from 19.89 (SD 13.24) to 20.17 (SD 12.75). The games incorporated key concepts that are not often covered in traditional classes, such as biomarker-guided treatment (eg, procalcitonin), empiric therapeutic techniques, and pathogen-specific approaches. Previous research has shown that active learning methods improve retention of multilayered decision-making frameworks [21,22].

Although the mean test scores differed somewhat, the consistent and statistically significant increase in the proportion of

XSL•FO
RenderX

participants across all 6 infection groups indicates that the intervention had a favorable effect. Game-based learning aligns with global guidelines emphasizing the need for new, interesting, and competency-based teaching approaches to improve antibiotics prescribing and reduce AMR [12,23,24]. The observed disparity between mean score trends and percentage increases across components indicates that group-level accuracy improved, while individual score distributions remained variable. This is most likely attributable to scoring standards and question design rather than a decline in knowledge.

Participants also reported improved engagement, motivation, and understanding of AMS concepts, consistent with research showing that game-based methods result in greater learner satisfaction and better preparation for clinical practice than passive teaching methods.

This study has several limitations. Reliance on convenience sampling and the absence of a control or comparator group limit the ability to establish causation between the intervention and improved knowledge outcomes. Second, participants were not stratified by discipline (eg, medicine vs pharmacy) or baseline knowledge, which may have affected the intervention's efficacy. Third, the absence of randomization or blinding in group allocation introduces the potential for selection and performance bias. The limited sample size and single-institution context may affect the generalizability of the results.

We advocate for the use of more rigorous methodologies for future research, such as randomized controlled trials or stepped-wedge cluster trials, including stratification based on educational background and baseline knowledge. Incorporating a comparison group (eg, conventional lecture-based instruction) would facilitate assessment of the true impact of gamified learning on AMS knowledge and retention.

With a heightened emphasis on education about AMR and AMS, numerous challenges will persist. However, these challenges could be alleviated through more immersive and interactive educational techniques. Gamified education presents a promising solution to address many of these issues. It is increasingly important to educate a diverse range of health care professionals nationally and globally within a short time frame to counter the looming threat of AMR.

Acknowledgments

The JSS Academy of Higher Education and Research, Department of Microbiology, and Mysore Medical College administrations are acknowledged by the authors for their assistance in making this study possible. Their support was crucial in supplying the resources and infrastructure required to successfully complete this study. This research received partial support from the *Journal of Medical Internet Research* (*JMIR*) through a 50% discount on the article processing fee granted during the publication process.

Funding

All authors declared that they had insufficient or no funding to support open access publication of this manuscript, including from affiliated organizations or institutions, funding agencies, or other organizations. JMIR Publications provided APF support for the publication of this article.

Data Availability

The datasets generated or analyzed during this study are not publicly available due to institutional privacy restrictions and participant confidentiality but are available from the corresponding author upon reasonable request.

Authors' Contributions

SMN conceptualized and designed the study, supervised the project, and critically revised the manuscript for important intellectual content. SRS contributed to the study design, performed data analysis, and assisted in manuscript drafting. YDM contributed to methodology development and data interpretation. GKM was involved in data collection, data analysis, and preparation of the initial manuscript draft. VSGS assisted in data collection and validation of results. CSE contributed to data compilation and manuscript editing. SSSC assisted in data collection and literature review. All authors reviewed, edited, and approved the final version of the manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Pre- and posttest questionnaire.
[DOCX File, 32 KB - mededu_v12i1e72236_app1.docx]

Multimedia Appendix 2
Game modules.
[DOCX File, 252 KB - mededu_v12i1e72236_app2.docx]

XSL•FO
RenderX

References

1. Majumder MAA, Rahman S, Cohall D, et al. Antimicrobial stewardship: fighting antimicrobial resistance and protecting global public health. Infect Drug Resist 2020;13:4713-4738. [doi: 10.2147/IDR.S290835] [Medline: 33402841]
2. Antimicrobial resistance. World Health Organization. 2023 Nov 21. URL: https://www.who.int/news-room/fact-sheets/detail/antimicrobial-resistance [accessed 2026-02-18]
3. About antimicrobial resistance. Centers for Disease Control and Prevention. 2025 Jan 31. URL: https://www.cdc.gov/antimicrobial-resistance/about/index.html [accessed 2026-02-18]
4. Prestinaci F, Pezzotti P, Pantosti A. Antimicrobial resistance: a global multifaceted phenomenon. Pathog Glob Health 2015;109(7):309-318. [doi: 10.1179/2047773215Y.0000000030] [Medline: 26343252]
5. Azevedo MM, Pinheiro C, Yaphe J, Baltazar F. Assessing the impact of a school intervention to promote students' knowledge and practices on correct antibiotic use. Int J Environ Res Public Health 2013 Jul 15;10(7):2920-2931. [doi: 10.3390/ijerph10072920] [Medline: 23860183]
6. Minen MT, Duquaine D, Marx MA, Weiss D. A survey of knowledge, attitudes, and beliefs of medical students concerning antimicrobial use and resistance. Microb Drug Resist 2010 Dec;16(4):285-289. [doi: 10.1089/mdr.2010.0009] [Medline: 20624097]
7. Kolb DA. Experiential Learning: Experience as the Source of Learning and Development: FT Press; 2014. URL: https://books.google.co.in/books?id=jpbeBQAAQBAJ [accessed 2024-05-24]
8. Deci EL, Ryan RM. Self-determination theory. In: Van Lange PAM, Kruglanski AW, Higgins ET, editors. Handbook of Theories of Social Psychology 2012:416-436. [doi: 10.4135/9781446249215.n21]
9. Anderson LW, Krathwohl DR. A Taxonomy for Learning, Teaching, and Assessing: A Revision of Bloom's Taxonomy of Educational Objectives: Complete Edition: Addison Wesley Longman, Inc; 2001. URL: https://eduq.info/xmlui/handle/11515/18824 [accessed 2024-05-24]
10. Aburahma MH, Mohamed HM. Educational games as a teaching tool in pharmacy curriculum. Am J Pharm Educ 2015 May 25;79(4):59. [doi: 10.5688/ajpe79459] [Medline: 26089568]
11. Sumana MN, Shettar SR, Maheshwarappa YD, et al. Gamified interventions to improve the knowledge, awareness and practice on rational use of antibiotics among school children in Mysuru, South India to curb the growing antimicrobial resistance (AMR). Front Public Health 2025;13:1574647. [doi: 10.3389/fpubh.2025.1574647] [Medline: 40963659]
12. Nowbuth AA, Asombang AW, Alaboud K, et al. Gamification as an educational tool to address antimicrobial resistance: a systematic review. JAC Antimicrob Resist 2023 Dec;5(6):dlad130. [doi: 10.1093/jacamr/dlad130] [Medline: 38089458]
13. Gentry SV, Gauthier A, L'Estrade Ehrstrom B, et al. Serious gaming and gamification education in health professions: systematic review. J Med Internet Res 2019 Mar 28;21(3):e12994. [doi: 10.2196/12994] [Medline: 30920375]
14. Gudadappanavar AM, Benni JM, Javali SB. Effectiveness of the game-based learning over traditional teaching-learning strategy to instruct pharmacology for phase II medical students. J Educ Health Promot 2021;10:91. [doi: 10.4103/jehp.jehp_624_20] [Medline: 34084838]
15. Oestreich JH, Guy JW. Game-based learning in pharmacy education. Pharmacy (Basel) 2022 Jan 6;10(1):11. [doi: 10.3390/pharmacy10010011] [Medline: 35076609]
16. Calik A, Cakmak B, Kapucu S, Inkaya B. The effectiveness of serious games designed for infection prevention and promotion of safe behaviors of senior nursing students during the COVID-19 pandemic. Am J Infect Control 2022 Dec;50(12):1360-1367. [doi: 10.1016/j.ajic.2022.02.025] [Medline: 35231565]
17. Wutzke SE, Artist MA, Kehoe LA, Fletcher M, Mackson JM, Weekes LM. Evaluation of a national programme to reduce inappropriate use of antibiotics for upper respiratory tract infections: effects on consumer awareness, beliefs, attitudes and behaviour in Australia. Health Promot Int 2007 Mar;22(1):53-64. [doi: 10.1093/heapro/dal034] [Medline: 17046966]
18. Yabrodi M, Abulebda K, Pearson KJ, et al. Using a serious game to teach central line care in pediatric critical care nursing. Am J Nurs 2025 Feb 1;125(2):36-42. [doi: 10.1097/AJN.0000000000000008] [Medline: 39844230]
19. Xu M, Luo Y, Zhang Y, Xia R, Qian H, Zou X. Game-based learning in medical education. Front Public Health 2023;11:1113682. [doi: 10.3389/fpubh.2023.1113682] [Medline: 36935696]
20. Peri AM, Chatfield MD, Ling W, Furuya-Kanamori L, Harris PNA, Paterson DL. Rapid diagnostic tests and antimicrobial stewardship programs for the management of bloodstream infection: what is their relative contribution to improving clinical outcomes? A systematic review and network meta-analysis. Clin Infect Dis 2024 Aug 16;79(2):502-515. [doi: 10.1093/cid/ciae234] [Medline: 38676943]
21. Jääskä E, Aaltonen K. Teachers' experiences of using game-based learning methods in project management higher education. Project Leadership and Society 2022 Dec;3:100041. [doi: 10.1016/j.plas.2022.100041]
22. Pesare E, Roselli T, Corriero N, Rossano V. Game-based learning and gamification to promote engagement and motivation in medical learning contexts. Smart Learn Environ 2016 Dec;3(1):1-21. [doi: 10.1186/s40561-016-0028-0]
23. Mahalmani V, Angadi NB, Pattanshetti S. Breaking the monotony: exploring the approach of game-based learning in students' understanding of antimicrobials and antimicrobial resistance. IJPP 2025;0:1-5. [doi: 10.25259/IJPP_100_2024]
24. Jones JS, Tincher L, Odeng-Otu E, Herdman M. An educational board game to assist PharmD students in learning autonomic nervous system pharmacology. Am J Pharm Educ 2015 Oct 25;79(8):114. [doi: 10.5688/ajpe798114] [Medline: 26689278]

XSL•FO
RenderX

Abbreviations

AMR: antimicrobial resistance
AMS: antimicrobial stewardship
LRTI: lower respiratory tract infection

Edited by LT Car; submitted 07.Feb.2025; peer-reviewed by C Br, E Castro-Sanchez, J Azai; revised version received 03.Nov.2025; accepted 06.Nov.2025; published 10.Mar.2026.

Please cite as:
MN S, Shettar SR, Maheshwarappa YD, Megha GK, GS VS, Eshwar CS, SC SS
Targeted Educational Intervention Through Game-Based Learning to Promote Rational Antimicrobial Use Among Health Care Learners: Prospective Interventional Study
JMIR Med Educ 2026;12:e72236
URL: https://mededu.jmir.org/2026/1/e72236
doi:10.2196/72236

© Sumana M N, Supreeta R Shettar, Yogeesh D Maheshwarappa, G K Megha, Veerabhadra Swamy GS, Chinchana Shylaja Eshwar, Shruthi Shree S C. Originally published in JMIR Medical Education (https://mededu.jmir.org), 10.Mar.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Education, is properly cited. The complete bibliographic information, a link to the original publication on https://mededu.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Enhancing Therapist Training in the Delivery of Exposure Therapy for Individuals with Anxiety Disorders Using Virtual Reality Simulation: Randomized Feasibility Trial

Joshua Kemp[1,2*], PhD; Ariella Rosen[2*], BA; Hyungjin Kim[1*], MD; Margo Adams Larsen[3*], PhD; Kristen Benito[1,2*], PhD; Jennifer Freeman[1,2*], PhD; Jason Machan[4*], PhD; Peter Tuerk[3*], PhD

[1]Department of Psychiatry and Human Behavior, Alpert Medical School, Brown University, Providence, RI, United States

[2]Pediatric Anxiety Research Center, Bradley Hospital, Brown University Health, 1011 Veterans Memorial Pkwy, Riverside, RI, United States

[3]Virtually Better Inc, Decatur, GA, United States

[4]Department of Biostatistics, Brown University Health, Providence, RI, United States

[*]all authors contributed equally

Corresponding Author:
Joshua Kemp, PhD
Department of Psychiatry and Human Behavior, Alpert Medical School, Brown University, Providence, RI, United States

Abstract

Background: Exposure-based cognitive behavioral therapy is among the least used evidence-based practices for anxiety disorders in routine care. Providers' negative beliefs about exposure (eg, fears of harm or intolerability) are a major barrier. Experiential methods can reduce these beliefs but are limited by accessibility, standardization, and fidelity. Virtual reality (VR) offers a scalable way to deliver standardized experiential practice. Guided by an "exposure to exposure" (E2E) framework, we conceptualized VR training as an exposure intervention targeting therapists' own anxious beliefs about exposure.

Objective: This feasibility study examined a VR-based exposure training program (SET-VR) (1) to evaluate usability and effects on therapist learning targets (knowledge, self-efficacy, attitudes) and (2) to test whether a high-immersion head-mounted display (HMD) format provides added benefit over a lower-immersion desktop format. Eligibility included holding an active caseload.

Methods: Eligible clinicians (ie, aged >21 years with an active caseload; n=41) completed a 4-hour didactic workshop on exposure and were randomized (1:1, blinded) to the desktop or HMD condition. In the experiential phase, therapists delivered 3 rounds of exposure with a virtual patient. They titrated exposure intensity (increase, decrease, continue as is) at fixed decision points based on state-dependent visual (character animations) and auditory (prerecorded verbalizations) cues reflecting the patient's distress. Exposure knowledge, self-efficacy, and beliefs about exposure were measured at baseline, post-didactic, post-experiential, and follow-up. Participants also rated the acceptability, usability, and authenticity of the program.

Results: Both groups (desktop and HMD) showed significant improvement in exposure knowledge (d=0.52, P=.006; d=0.58, P=.002), self-efficacy (d=0.88, P<.001; d=1.36, P<.001), and beliefs (d=0.61, P=.001; d=1.05, P<.001) from baseline to post-didactic training using binomial generalized estimating equations. There were no significant differences between the low- and high-immersion groups on any measure after didactics. Both groups demonstrated significant improvement in exposure self-efficacy (d=0.66, P<.001; d=0.93, P<.001) and beliefs (d=0.46, P<.01; d=0.66, P<.001) from post-didactic to post-experiential. Both groups gave positive ratings for acceptability, usability, and authenticity. No adverse events or side effects were reported.

Conclusions: In this feasibility randomized controlled trial, an E2E-guided VR training program produced promising improvements in therapists' self-efficacy and negative beliefs about exposure beyond gains from didactic training alone. This work is innovative in testing immersion as a dose parameter while also applying an explicit framework (E2E) to target a key mechanism (ie, therapist beliefs) in the underuse of exposure therapy. Compared to prior VR training studies focused on skills and knowledge acquisition, our findings support the standardization of an emotionally engaging exposure practice context that shifts therapist-level mechanisms linked to actual delivery. The lack of clear advantages for HMD over desktop VR suggests that lower-immersion, more scalable implementations may provide a sufficient experiential "dose." Larger, more diverse trials are needed to confirm effectiveness and determine the real-world impact of VR-based exposure training on access to evidence-based care.

Trial Registration: ClinicalTrials.gov NCT06706245; https://clinicaltrials.gov/study/NCT06706245

(JMIR Med Educ 2026;12:e80087) doi:10.2196/80087

XSL•FO
RenderX

KEYWORDS

virtual reality; exposure therapy; anxiety disorders; training; simulation

Introduction

Anxiety disorders are among the most common mental health conditions in children and youth, impacting 1 in 5 children since the COVID-19 pandemic [1]. Symptoms in youth can manifest as uncontrollable worry, fear, and hyperarousal, leading to a cascade of developmental consequences and pervasive disability in adulthood [2,3]. Cognitive behavioral therapy (CBT) is a highly effective treatment for addressing anxiety in children and is regarded as the first-line treatment [4]. Within CBT for anxiety, exposure therapy has gained extensive empirical support as a frontline intervention across anxiety disorders, including obsessive-compulsive disorder and posttraumatic stress disorder [5].

Despite robust empirical support, exposure-based CBT remains one of the least utilized evidence-based practices in typical practice settings, with as few as 7% of patients reporting receiving exposure therapy for their anxiety disorder [6]. A leading barrier to its dissemination is provider access to training, with only 12% to 28% of therapists reporting having received training to provide exposure therapy [7]. Another barrier is therapists' willingness to use exposure therapy after receiving training. Surveys suggest that a majority of therapists who have received training seldom use exposures in practice [7,8]. The current "gold standard" for exposure training involves a workshop, an accompanying manual, and clinical supervision [9,10]. Unfortunately, research suggests that this training approach may not be sufficient to prepare therapists to deliver exposure therapy. One study randomizing therapists to different "doses" of training showed that only 54% of clinicians receiving the highest "dose" of training reached proficiency levels [10]. Most training approaches are purely didactic in nature, with the goal of transferring knowledge about exposure and its delivery process; however, research consistently demonstrates that exposure knowledge is not predictive of uptake and quality delivery [9,11,12]. This suggests that current training approaches lack the effectiveness to promote the necessary levels of proficiency and confidence for providers to put exposure therapy into practice. Training innovations focused on the accessibility and methodology of exposure therapy training are necessary to overcome this glaring research-to-practice gap.

A key area unaddressed in conventional training that may explain the underutilization of exposure following training is therapists' negative beliefs about exposure. Researchers have found that therapists' decision to not use exposure therapy is linked to beliefs about its dangerousness or risk of negative events [13]. Such beliefs include concerns about symptom exacerbation [14], patient decompensation [15], direct patient harm [16], and eventual treatment dropout. However, serious negative consequences from exposure therapy are incredibly rare and comparable to those of other therapeutic modalities [17]. These negative beliefs are also associated with suboptimal delivery when exposure is used (ie, deviation from the prolonged and intense delivery recommended by exposure theorists and treatment manuals) [18]. This includes the introduction of breathing strategies [19], allowing the use of safety behaviors and premature termination of exposures [20], all of which negatively impact the efficacy of exposure therapy [21]. Given the well-documented role of negative beliefs in stifling both the use and optimal delivery of exposure therapy, ideal training in exposure therapy must include strategies to target this barrier.

Training studies of exposure therapy have shown that experiential learning (ie, partner- and self-exposure) can be an effective way to reduce therapists' negative beliefs [22]. This is consistent with the benefits of experiential learning observed in other domains of medical training [23]. Given the similarities between therapists' negative beliefs about using exposure and clients' anxious beliefs about engaging with feared items, and the well-established efficacy of exposure as a process for reducing anxious beliefs, it is logical to approach experiential tasks during exposure training as E2E to elicit and ameliorate therapists' negative beliefs [24]. Based on this logic, studies have demonstrated the benefits of incorporating experiential activities, such as role-playing, to enhance training [25]. While this is a step toward improving the quality of exposure therapy training, it faces challenges in accessibility, standardization, and fidelity to real-life experiences that limit optimization. Building on this E2E framework [24], we conceptualized the training exercises themselves as a form of exposure intervention directed at therapists' own anxious beliefs about exposure. In this view, a key design question is how closely the training context must approximate real exposure sessions with distressed patients for therapists to experience corrective learning about the safety and tolerability of exposure. VR-based simulations provide a way to systematically manipulate this "experiential dose" of E2E by varying the level of immersion while holding didactic content constant.

Virtual environments (VEs) and virtual reality (VR) offer a promising solution to further enhance exposure training via experiential learning. VEs are computer-created environments with objects that represent real-life counterparts and allow users to interact with these simulations in real time as though they were real [26]. VR involves the display of computer-generated environments in a 3D manner, which allows the user to experience immersive, multisensory interactions in real time [26]. VR has unique advantages in its representativeness of real clinical situations, individualization, replicability, and economic scalability—all of which currently limit the efficacy and reach of conventional role-playing and other forms of experiential learning [27]. Leveraging these strengths, the medical education field has recently made efforts to integrate VR into training with positive outcomes [23]. Emerging evidence suggests that VR may also be an effective tool for improving skills and attitudes pertinent to the treatment of mental health concerns. A study of mental health nursing trainees demonstrated VR's effectiveness in improving knowledge, skills, and attitudes [28]. A recent systematic review of VR in psychiatric education found several studies showing improved knowledge and clinical skills for participants across a range of conditions [29]. VEs and VR show great promise for solving exposure training's dual

XSL•FO
RenderX

challenges of accessibility and provider-level barriers via E2E for clinicians.

There has been growing research into how different aspects of a learner's experience within the virtual environment impact their learning outcomes. Presence, the feeling of "being there" in the virtual setting [30], has been identified as a moderator of increased engagement [31], emotional activation [30], and social attitude change [32]. Embodiment, the sense of occupying the virtual avatar controlled by the user [33], has been shown to generally improve engagement and outcomes especially for retention [34]. Generally, interacting with virtual environments via hardware, such as head-mounted display (HMD), has shown higher levels of presence and embodiment than use through computer monitors or hand-held screens [35]. Because the research on VR-based delivery of mental health training is in its nascency, studies have focused on whether incorporating VR enhances learning outcomes compared to conventional nondigital methods [27,36], and this study team found no studies that directly compare across different immersive experience conditions. This limits our ability to understand how VR-specific variables may impact learning outcomes. Thus, further investigation into how these variables impact psychotherapy training will be a key step in optimizing the effectiveness and implementation of virtually delivered mental health training. The goal of this study was to evaluate the feasibility of a theory-driven, simulation-based E2E training program and to obtain initial evidence that supplementing didactic exposure training with experiential VR can shift key therapist-level learning targets. Specifically, we tested the recently developed Simulated Exposure Trainer (SET-VR) in a randomized feasibility trial that compared 2 delivery modalities differing in levels of immersion (low: desktop computer; high: HMD), conceptualized as 2 experiential doses of E2E practice. Our primary questions were (1) whether SET-VR would be usable and acceptable to clinicians and produce additional improvements in exposure knowledge, perceived self-efficacy, and therapist beliefs about exposure beyond a standard didactic workshop, and (2) whether high-immersion HMD VR would produce any incremental benefit over lower-immersion desktop delivery on these mechanism-based outcomes and perceived realism. It was hypothesized that both modalities would meet an a priori benchmark for usability.

Methods

Participants

This National Institute of Mental Health–funded (R41MH131229) study aimed to recruit a final sample of 40 clinicians. Of the 42 participants enrolled in the therapist training trial, one withdrew after consenting due to scheduling difficulties. The remaining clinicians (N=41) completed all stages of the training trial, including follow-up. Clinicians were recruited from a variety of settings between November 2023 and April 2024. Recruitment methods included convenience sampling (recruiting from the pool of clinicians at Bradley Hospital), offering training to clinicians on a waitlist to participate in nonexperimental CBT training at Bradley led by the first author, and targeted email advertising to social workers, private-practice clinicians, and providers affiliated with a local mental health practice. The eligibility criteria for the study included a bachelor's level of education (or higher), the ability to come to Bradley Hospital for in-person experiential training, and the ability to deliver clinical care to patients in a professional capacity.

The majority of the therapist participants (37/41, 90%) were White and non-Hispanic. Four out of 41 (10%) therapists reported Hispanic or non-White identities. A majority (35/41, 85%) also identified as female. Participants had a variety of degree types, including doctorate (n=5), master's (n=27), and bachelor's-level (n=9) degrees. Licensure types include clinical psychologists (n=4), mental health counselors (n=13), social workers (n=13), and either graduate students or other (non-licensed) clinical roles (n=11).

Measures

System Usability Scale

The System Usability Scale (SUS) is a 10-item questionnaire that asks participants to rate a website or program across areas, including ease of use, intuitiveness, functionality, and user confidence on a scale of 1 ("strongly disagree") to 5 ("strongly agree") [37]. Raw total scores are converted to a 0 - 100 scale, with values >70 indicating average or better usability and >80 indicating excellent usability.

Virtual Patient Evaluation

The Virtual Patient Evaluation (VP Eval) consisted of 12 items rated on a 5-point Likert scale ("strongly disagree" to "strongly agree"), as well as 3 open-text responses [38]. The first 6 items focused on the authenticity of the therapist's engagement with the virtual patient (eg, "While working with the virtual patient, I felt I had to make the same decisions a therapist would make in real life"), while items 7 to 12 assessed SET-VR as a training tool and the extent to which it helped the therapists feel ready to engage with an actual patient.

Therapist Negative Beliefs about Exposure Scale

The Therapist Negative Beliefs about Exposure Scale (TBES) assesses the extent to which therapists agree with 21 negative beliefs about exposure therapy (eg, "most clients have difficulty tolerating the distress exposure therapy evokes"). Items are rated on a 5-point scale from 0 ("disagree strongly") to 4 ("agree strongly") [19]. Possible scores range from 0 to 84, with a higher score indicating more negative beliefs about exposure. The TBES is the primary outcome measure for assessing target engagement.

Exposure Knowledge

We condensed the original 49-item measure into 12 multiple-choice items that best fit the didactic content of the training (eg, "Why is it important to block avoidance during exposure tasks?") [20]. Items assessing procedures or content not included in this training were removed, leaving a subset aligned with the material taught in this trial. The same subset of items was used in the preceding study to this trial [24]. The total score is the percentage of correct answers out of 100.

XSL•FO
RenderX

Exposure Self-Efficacy

This is a 27-item measure of therapists' confidence in delivering exposure therapy [20]. The first 8 items focus on the therapist's ability to help patients learn skills related to exposure (eg, "I feel confident in my ability to help my clients identify how avoidance is maintaining their fear"), and the remaining 19 items assess the therapist's confidence with implementing various aspects of exposure (eg, "I feel confident in my ability to conduct imaginal exposure"). Items are rated on a scale ranging from 1 ("not confident") to 5 ("very confident"). This measure has demonstrated high internal consistency and predictive validity in determining the frequency of self-reported clinical use of exposure therapy.

Training Acceptability Rating Scale

The Training Acceptability Rating Scale (TARS) is a measure of therapists' satisfaction [39]. The first 6 items, rated on a scale of 1 ("strongly disagree") to 6 ("strongly agree"), assess the acceptability of the training program, and the following 9 items ask therapists to rate the training's perceived utility on a scale of 1 ("not at all") to 4 ("a great deal"). Item means for acceptability and effectiveness subscales are calculated separately. This measure has good psychometric properties.

Procedures

Overview

The total training duration was 6 hours, comprising a 4-hour didactic learning course delivered over Zoom and a 2-hour in-person experiential training that took place on-site at Bradley Hospital. All clinicians (n=41) received the same didactic training. After the online training workshop, participants were scheduled for the second portion of training and randomly assigned to either the VR (high immersion; n=20) or computer (low immersion; n=21) experiential learning conditions. Participants were allocated to the conditions in a 1:1 ratio using random number generation and the randomization tool in REDCap (Research Electronic Data Capture). The randomization of each participant took place as closely as possible to the start of experiential training, which prevented the research team from having advance knowledge of the assignments. All 3 learning measures (ie, exposure knowledge, self-efficacy, and negative beliefs) were administered at baseline (T0), post-didactic (T1), post-experiential (T2), and 1-month follow-up (T3) to allow a direct comparison of changes in learning over time (Table 1).

Table . Key outcome measures and timepoints for delivery in the SET-VR[a] therapist training trial.

Measure	Baseline (T0)	Post-didactic workshop (T1)	Post-experiential learning with SET-VR (T2)	Follow-up (T3)
Demographics	✓			
TBES[b]	✓	✓	✓	✓
Exposure knowledge	✓	✓	✓	✓
Self-efficacy	✓	✓	✓	✓
System usability scale			✓	
VP[c] evaluation			✓	
TARS[d]			✓	

[a]SET-VR: Simulated Exposure Trainer-Virtual Reality.

[b]TBES: Therapist Negative Beliefs about Exposure Scale.

[c]VP: virtual patient.

[d]TARS :Training Acceptability Rating Scale.

Didactic Workshop

The online workshop took place before randomization and consisted of several modules outlining the main components of exposure therapy and its delivery. The training modules included content areas such as diagnostic review, psychoeducation, functional assessment, exposure design and selection, exposure delivery, and exposure therapy troubleshooting. All explicit exposure-related knowledge (eg, key concepts, principles, and procedures) was delivered during this workshop and was identical for both conditions; the subsequent experiential component did not introduce any new didactic content. To accommodate therapists' schedules and the staggered nature of recruitment, a total of 8 didactic trainings were held, consisting of approximately 5 individuals per round. All didactic trainings were led by Dr Kemp, and each cohort covered the same course material.

Experiential Component (SET-VR Training Scenario)

Overview

Experiential training occurred 1 to 2 weeks after completing the didactic portion. Participants were blinded to their randomized condition until the start of their experiential training, at which point trainers disclosed whether they would be using a computer or a VR headset for the task. Given the nature of the task, further blinding was not possible. Participants in both conditions experienced the same SET-VR simulated training scenario (Table 1); only the delivery modality (VR vs computer) differed. Experiential training was conducted one-on-one with a member of the research team (JK or AR). In all, the SET-VR training experience consisted of an introductory sequence

XSL•FO
RenderX

orienting users to the program structure and navigation, followed by simulated exposure therapy delivery that consisted of 9 separate decision points and then a feedback module that provided targeted feedback about therapists' decisions during delivery. The 9 decision points were implemented using a simple branching decision tree, such that the exposure options available at each point depended on the therapist's immediately preceding choice. Participants repeated the full training scenario 3 times within the 2-hour sitting but were allowed to fast forward through the introductory sequence after their first time using the program. The VR component was designed as a purely experiential practice phase in which therapists applied the exposure procedures learned in the workshop with a simulated patient; no additional exposure-related knowledge content was provided during this phase. A video overview of the SET-VR program can be obtained by contacting the corresponding author.

Introductory Sequence

The program began with a 9-slide introductory sequence that welcomed the user to SET-VR, provided relevant background information to orient the therapist to the virtual patient's presenting concerns, introduced the Subjective Units of Distress Scale (SUDS), and demonstrated how to use "in-game" controls to ask for SUDS ratings and modify exposure difficulty. The sequence concluded with therapists selecting an initial exposure task for the virtual patient from a preconstructed hierarchy of 10 exposure tasks, which was ordered from the most difficult task at the top (eg, "touch shoe to face") to the least difficult at the bottom (eg, "look at shoe"). The initial exposure task could be selected from anywhere on the hierarchy.

Simulated Treatment Delivery

After initiating the first exposure task, the therapist progressed through a fixed sequence of 9 decision points (Figure 1). Beginning with the initial task selection on the hierarchy, the simulation implemented simple branching logic. At each subsequent decision point (occurring every 45 seconds), therapists chose whether to increase exposure difficulty (select an item higher on the hierarchy), decrease the difficulty (select a lower item), or keep it the same (continue with the same exposure task). When raising or lowering the difficulty level of an exposure, available options were restricted to tasks within 2 hierarchy positions above or below the current task. This constraint prevented unrealistic, large jumps in exposure intensity and subtly guided learners toward the gradual titration that characterizes optimal exposure delivery. Throughout the experience, prerecorded verbalizations from the virtual patient played every 15 seconds, and both verbalizations and patient animations were keyed to the patient's current distress state. The underlying subjective units of distress (SUDS; 0‐10) value was automatically updated by the program, decreasing by 1 point approximately every 2 minutes of continuous exposure to approximate the natural process of habituation during effective exposure sessions rather than being manually adjusted by an instructor. Therapists could sample the virtual patient's distress level throughout the experience by clicking the SUDS button in the bottom right corner of the screen.

Figure 1. Screen capture of the virtual patient guided by a therapist learner during SET-VR training.



System Implementation

The SET-VR app was developed in Unity (C#) and deployed in both immersive VR and desktop configurations using the same core application logic. In the high-immersion condition, participants used a Meta Quest 2 HMD; in the low-immersion condition, they interacted with the identical simulation on a laptop computer. The virtual patient is controlled by a centralized state machine that manages transitions between predefined exposure and distress states. Transitions between these states trigger synchronized auditory cues (prerecorded

XSL•FO
RenderX

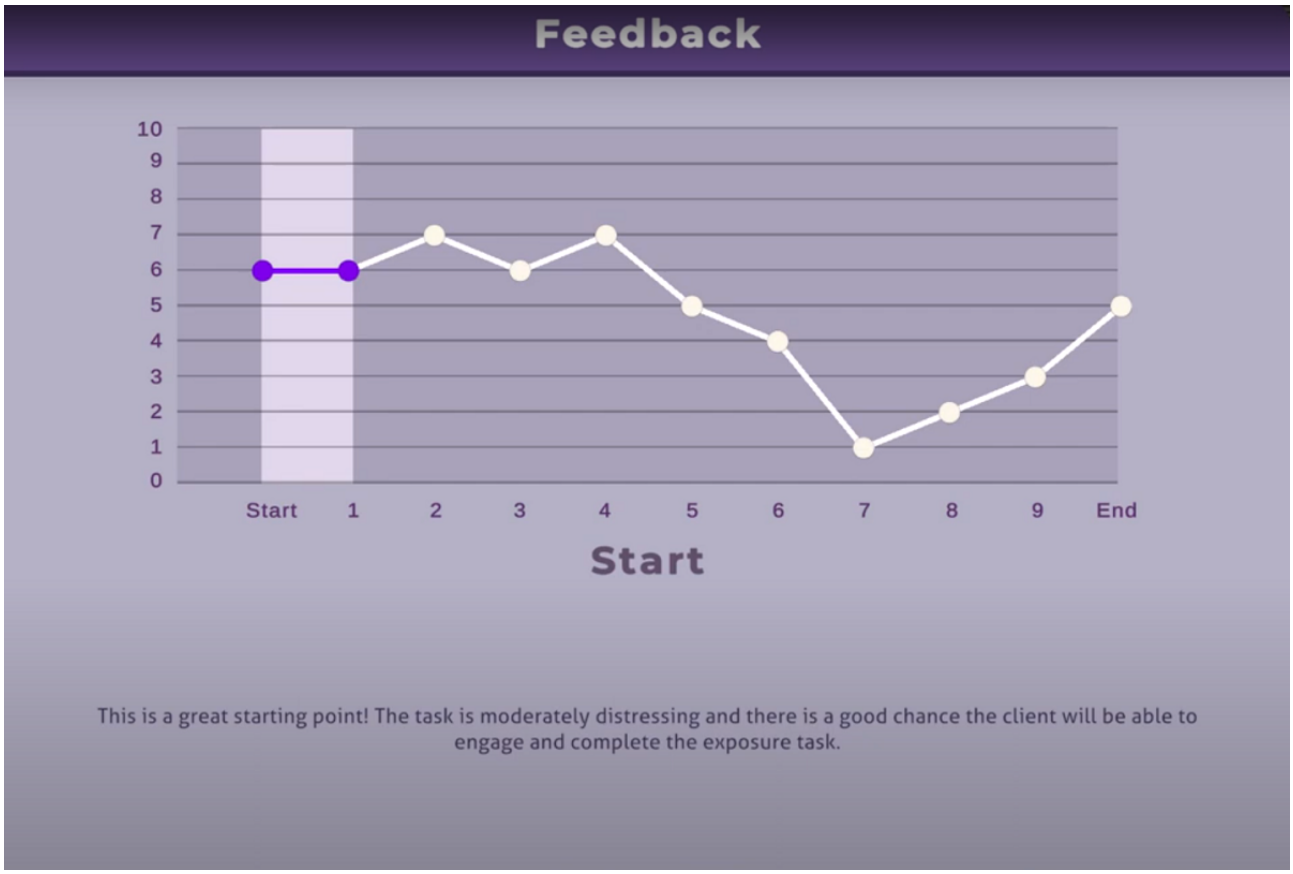
patient voice responses) and visual cues (state-specific 3D character animations), which together constitute the visual and auditory feedback that therapists used to guide their decisions. All interaction events (eg, exposure item selections, SUDS checks), exposure transitions, and SUDS values are logged by a persistent session manager and stored in an authentication-protected Firebase cloud database.

Treatment Delivery Feedback

After completing the ninth and final exposure task, the clinician was shown a graph of all their decision points. The feedback screen illustrated the virtual patient's SUDS levels throughout the experience and at each therapist decision point. As the clinician progressed through the feedback graph, they received specific feedback about each of the decisions they had made (Figure 2).

The program was reset after each use, giving clinicians a blank slate to make new decisions, as if it were the start of a new session (the virtual patient did not "remember" the previous session). The trainer stayed in the room with the participant throughout all 3 repetitions of the SET-VR program to orient the therapist to the headset and controllers and provide technical assistance, if needed. After completing the experiential training, participants received the end-of-study measures, which were completed electronically in the presence of the trainer.

Figure 2. Screen capture of the learning feedback screen provided to therapist learners after completing a SET-VR exposure scenario.



Ethical Considerations

Study procedures and informed consent documentation were approved by the Brown University Health Institutional Review Board (IRB00004624) in July 2022. All participants provided written consent after meeting with study staff to review the consent documents. As compensation for time spent engaged in study activities, participants received continuing education credits and payment in the form of either cash or an Amazon gift card. Both continuing education credits and monetary compensation (US $300 total) were awarded to participants after they completed all parts of the study. To protect the privacy and confidentiality of participant information, identifying information from the consent process was stored on a secure password-protected server, and all analyses were conducted using a fully deidentified dataset. This is not a secondary analysis, and consent was provided for all reported data. There is no identifying information presented in the analyses or supporting tables and figures. No adverse events or side effects were reported at any point during or after the trial. Study details are reported using CONSORT (Consolidated Standards of Reporting Trials) 2025 (Checklist 1) and CONSORT-EHEALTH (Consolidated Standards of Reporting Trials of Electronic and Mobile Health Applications and Online Telehealth) guidelines [40] (Checklist 2).

Analytic Plan

Statistical analyses were carried out using SAS version 9.4 1M7. Knowledge, self-efficacy, and TBES were assessed longitudinally and analyzed via binomial generalized estimating equations (GEE). SUS, VP Eval, and TARS were assessed only at the conclusion of the study and analyzed using a binomial generalized linear model. Both sets of analyses used the binomial distribution to constrain the model space to the range

XSL•FO
RenderX

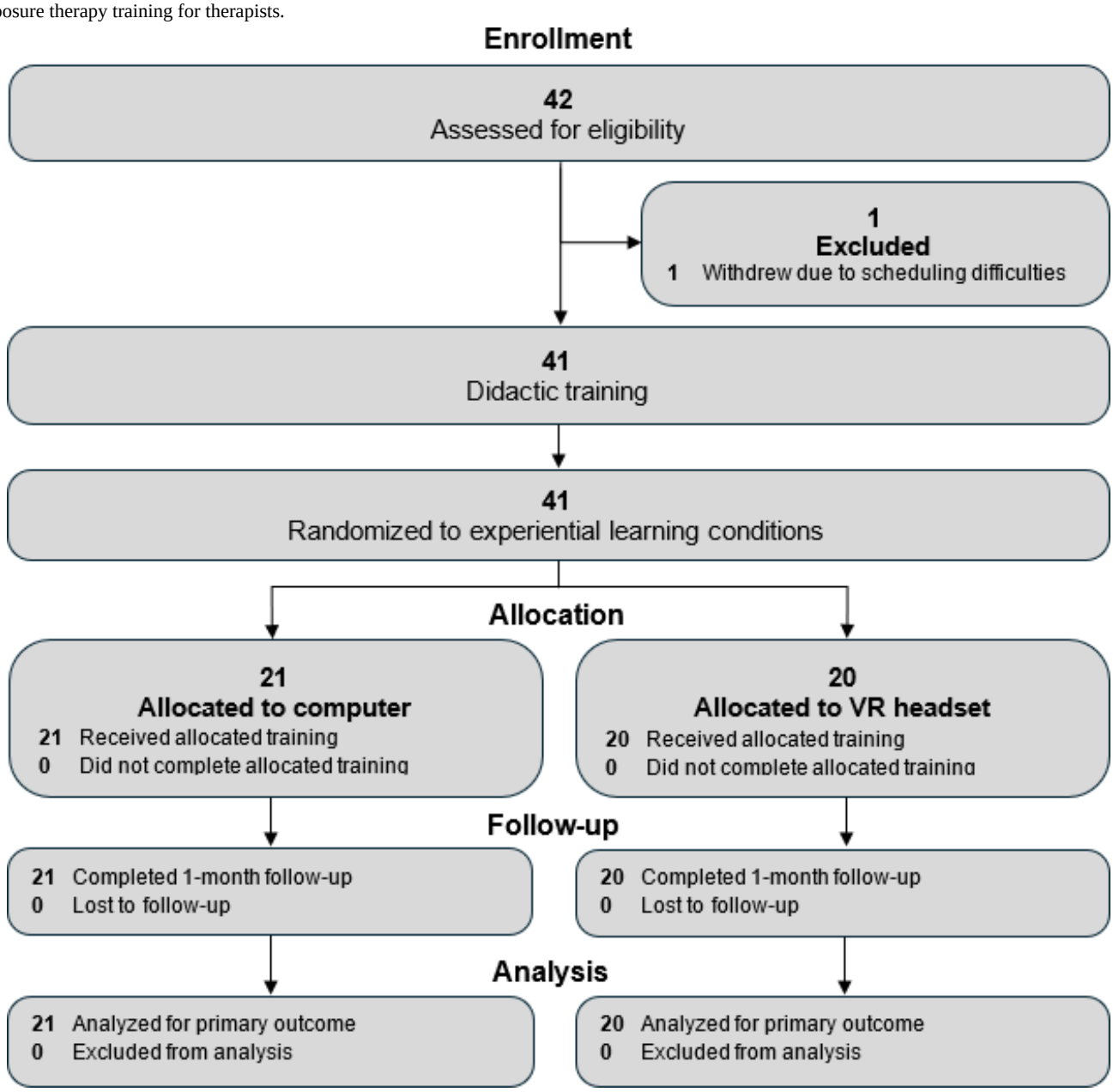
of measures being used. The scores were first shifted to have a minimum score of zero by subtracting the minimum score possible. This was used as the binomial number of "successes," and the range of the score was used as the number of "trials." Classical sandwich estimation was then used to better incorporate any differences in empirical variances and covariances. Both raw and adjusted (where appropriate) *P* values were presented, as well as unadjusted 95% CIs. For GEEs, 16 different hypotheses were tested using orthogonal linear estimates: (1) changes from baseline conditions in each condition (8 total), (2) high- and low-immersion training conditions compared at each timepoint cross-sectionally (4), and (3) condition comparisons in changes from baseline for each follow-up (4). The Holm test was used to calculate adjusted *P* values, maintaining familywise alpha at .05 across all 16 hypothesis tests in GEEs. For all other models, only 1 hypothesis was tested comparing the conditions. There was no attrition after randomization and participants were required to provide values for all survey items except open-text responses; imputation techniques were not used given the lack of missing data.

Results

Baseline Group Comparison (T0)

The high- and low-immersion training conditions did not exhibit statistically significant differences on measures of exposure knowledge (t_{117}=1.75, *P*=.08), exposure self-efficacy (t_{117}=1.90, *P*=.06), and attitudes about exposure therapy (t_{117}=0.72, *P*=.47) before initiating the workshop training, indicating both groups had similar ratings on key variables prior to initiating the training intervention (Figure 3) [40].

Figure 3. CONSORT (Consolidated Standards of Reporting Trials) diagram of participant flow through a randomized feasibility trial evaluating exposure therapy training for therapists.



XSL•FO
RenderX

Post-Didactic Between-Group Differences (T1)

There were no statistically significant differences between the groups on measures of exposure knowledge (t_{117}=0.93, P=0.35), exposure self-efficacy (t_{117}=1.40, P=.17), and attitudes about exposure (t_{117}=0.00, P=.998) following the completion of the didactic portion of training, which suggests both groups demonstrated similar amounts of change in these key variables from baseline to postdidactic timepoints.

Within-Group Changes From Baseline to Post-Didactic (T0 to T1)

The headset and computer groups exhibited significant improvement in exposure knowledge (t_{117}=3.16, P=.002 and t_{117}=2.81, P=.006, respectively), exposure self-efficacy (t_{117}=7.37, P<.001 and t_{117}=4.75, P<.001, respectively), and attitudes about exposure (t_{117}=5.68, P<.001 and t_{117}=3.32, P=.001, respectively) from baseline to postdidactic timepoints.

Post-Experiential Between-Group Differences (T2)

There were no significant differences in exposure knowledge (t_{117}=1.41, P=.16), exposure self-efficacy (t_{117}=1.38, P=.17), and attitudes about exposure (t_{117}=0.11, P=.92) at the end of the workshop training.

Within-Group Changes From Post-Didactic to Post-Experiential (T1 to T2)

The headset and computer groups demonstrated significant improvement in exposure self-efficacy (t_{117}=5.03, P<.001 and t_{117}=3.58, P<.001, respectively) and attitudes about exposure therapy (t_{117}=3.56, P<.001 and t_{117}=2.5, P=.01, respectively) from post-didactic to the end of the workshop timepoints. Neither group demonstrated significant change in exposure knowledge from post-didactic to the end of workshop (t_{117}=0.19, P=.853 and t_{117}=0.33, P=.74, respectively). Taken together, these findings indicate that the experiential portion of training significantly improved therapists' perceptions of self-efficacy in delivering exposure therapy and their attitudes toward exposure therapy, but it did not significantly improve their knowledge of exposure therapy beyond ratings taken after the completion of didactic activities. Mean knowledge scores at T1 and T2 (Table 2) remained well below the maximum possible score of 12, indicating that the absence of additional knowledge gains from post-didactic to post-experiential was not attributable to a ceiling effect. Instead, this pattern is consistent with the design of the training, in which no new didactic exposure content was introduced during the VR phase (see Figures 4-6).

XSL•FO
RenderX

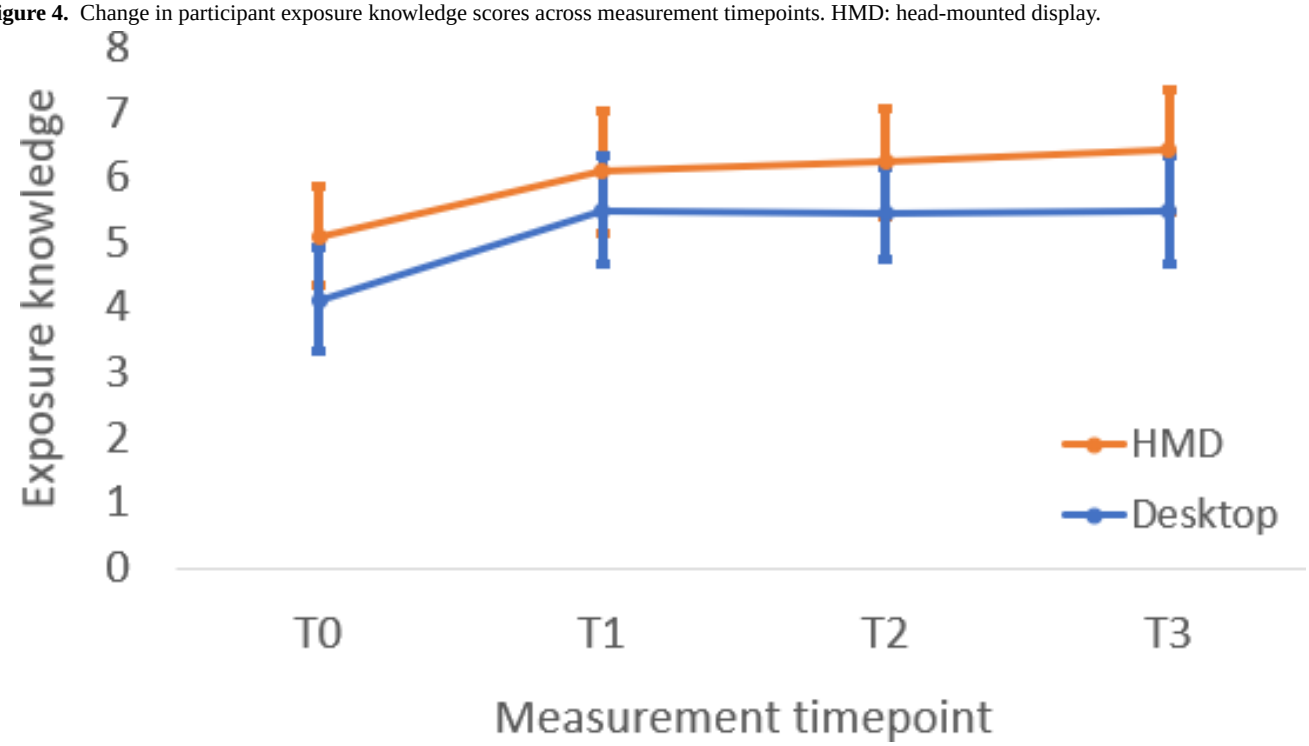
Table . Group means and CIs by timepoint and effect sizes for within-group changes across timepoints for primary and secondary outcome measures.

		Within-subject change			
Group and timepoint	Cross-sectional, mean (SD; 95% CI)	*P* value	Adjusted *P* value	Cohen *d*	From
Exposure knowledge					
Headset					
Baseline	34.17 (15.51; 27.37 - 40.97)				
Post-didactic	45.83 (16.11; 38.77 - 52.89)	.002	.03	0.58	Baseline
Post-experiment	45.42 (13.91; 39.32 - 51.52)	.85	.99	−0.04	Post-didactic
Follow-up	45.83 (16.11; 38.77 - 52.89)	.99	.99	0.00	Post-didactic
Desktop					
Baseline	42.46 (15.34; 35.90 - 49.02)				
Post-didactic	50.79 (18.80; 42.75 - 58.83)	.006	.06	0.52	Baseline
Post-experiment	51.98 (16.44; 44.95 - 59.01)	.74	.99	0.06	Post-didactic
Follow-up	53.57 (18.92; 45.48 - 61.66)	.30	.99	0.19	Post-didactic
Exposure self-efficacy					
Headset					
Baseline	72.35 (23.21; 62.77 - 82.53)				
Post-didactic	100.45 (19.22; 91.71 - 108.16)	<.001	<.001	1.36	Baseline
Post-experiment	110.70 (15.33; 103.48 - 116.67)	<.001	<.001	0.93	Post-didactic
Follow-up	114.00 (19.11; 104.56 - 121.04)	<.001	.001	0.74	Post-didactic
Desktop					
Baseline	87.52 (28.45; 75.43 - 98.99)				
Post-didactic	108.52 (18.41; 100.04 - 115.50)	<.001	<.001	0.88	Baseline
Post-experiment	117.29 (15.44; 109.81 - 122.87)	<.001	.004	0.66	Post-didactic
Follow-up	120.38 (11.65; 114.76 - 124.63)	<.001	<.001	0.88	Post-didactic
Therapist negative beliefs about exposure scale					
Headset					
Baseline	30.90 (10.49; 26.54 - 35.54)				
Post-didactic	21.20 (10.03; 17.18 - 25.80)	<.001	<.001	−1.05	Baseline
Post-experiment	16.65 (10.50; 12.58 - 21.63)	<.001	.005	−0.66	Post-didactic
Follow-up	15.95 (10.61; 11.87 - 21.02)	<.001	.001	−0.75	Post-didactic

XSL•FO
RenderX

Group and timepoint	Cross-sectional, mean (SD; 95% CI)	Within-subject change			
		P value	Adjusted *P* value	Cohen *d*	From
Desktop					
Baseline	28.76 (8.96; 25.12 - 32.64)				
Post-didactic	21.19 (10.94; 16.92 - 26.11)	.001	.01	−0.61	Baseline
Post-experiment	16.33 (9.06; 12.86 - 20.48)	.01	.10	−0.46	Post-didactic
Follow-up	14.71 (10.23; 10.89 - 19.52)	.002	.02	−0.57	Post-didactic

Figure 4. Change in participant exposure knowledge scores across measurement timepoints. HMD: head-mounted display.



XSL•FO
RenderX

Figure 5. Change in participant exposure self-efficacy scores across measurement timepoints. HMD: head-mounted display.

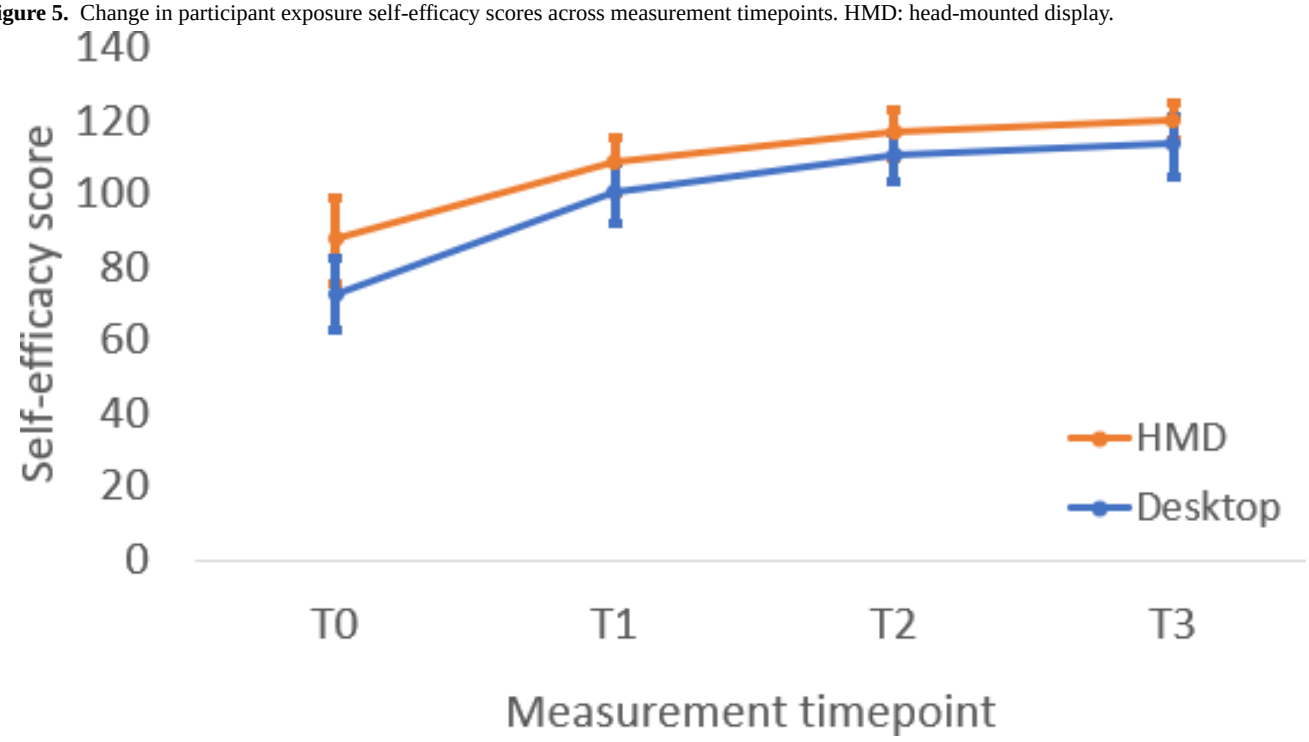
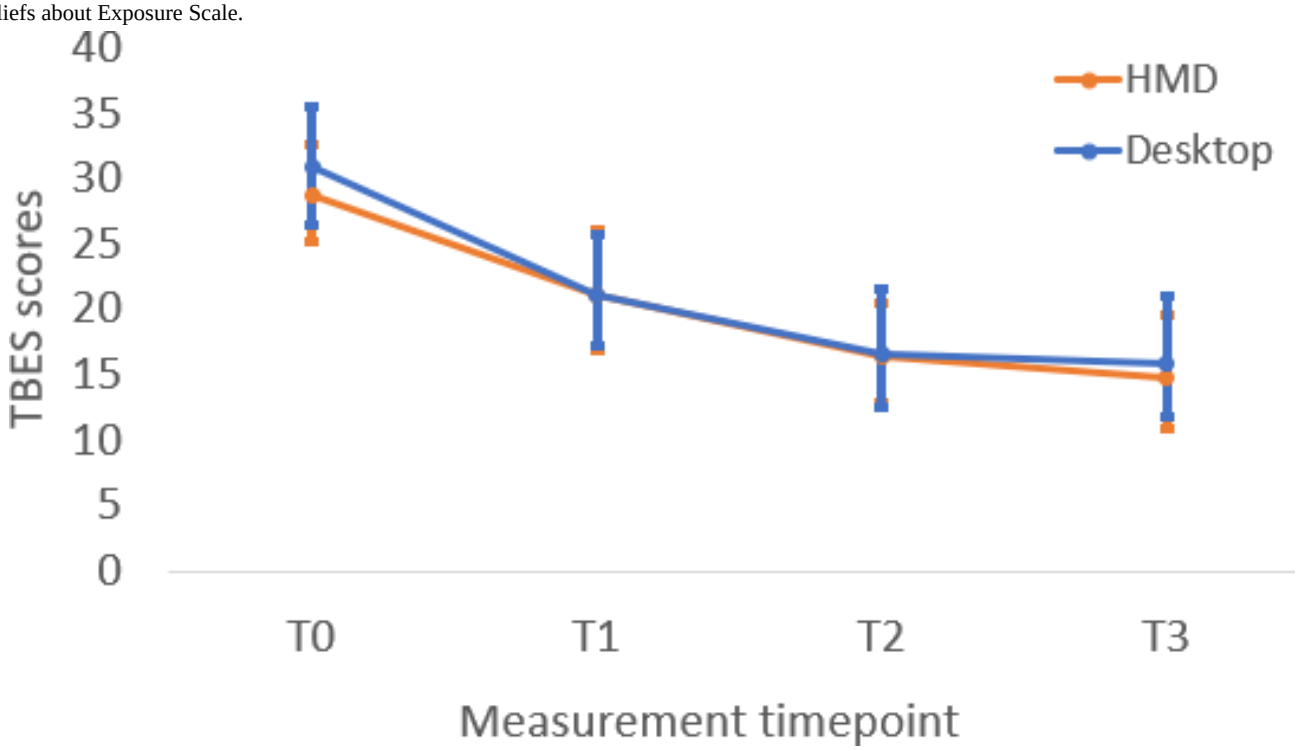


Figure 6. Change in participant beliefs about exposure therapy across measurement timepoints. HMD: head-mounted display; TBES: Therapist Negative Beliefs about Exposure Scale.



Follow-up Between-Group Differences (T3)

There were no significant differences in exposure knowledge (t_{117}=1.44; *P*=.15), exposure self-efficacy (t_{117}=1.38; *P*=.17), and attitudes about exposure (t_{117}=0.39; *P*=.70) after a 1-month follow-up period, indicating comparable maintenance of learning between conditions.

Within-Group Changes From Post-Experiential to Follow-up (T2 to T3)

The headset and computer groups exhibited durable maintenance of gains during the 1-month follow-up period on measures of exposure knowledge (t_{117}=0.17; *P*=.87 and t_{117}=0.49; *P*=.63, respectively), self-efficacy (t_{117}=1.26; *P*=.21 and t_{117}=1.50; *P*=.14, respectively), and attitudes about exposure (t_{117}=0.89; *P*=.38 and t_{117}=0.94; *P*=.35, respectively).

XSL•FO
RenderX

Perceptions of the SET-VR Training Simulation (T2)

Ratings of system usability using the SUS measure were statistically significantly higher for the computer group relative to the headset group (t_{39}=−2.87, P=.007); however, usability for both groups (HMD: mean 86.2, SD 9.1, 95% CI 81.68-89.75; desktop: mean 93.4, SD 6.7, 95% CI 89.87-95.80) was rated highly, with both modalities ranking in the upper 96th percentile of usability according to established norms [41]. Therapist perceptions of the virtual patient's authenticity, degree of engagement, and overall utility as a learning experience, as assessed by the VP Eval measure, were similarly positive in both groups (t_{39}=0.08; P=.93).

Overall Perceptions of Training (T2)

Ratings of the training processes and their acceptability using the TARS measure were positive. The groups exhibited nearly identical scores on ratings of the training's acceptability, which assess general acceptability, perceived effectiveness, and appropriateness of the training; however, the computer group reported significantly more positive ratings of training processes (t_{39}=−2.02; P=.05) relative to the headset group, which assessed the extent to which the teaching approach felt helpful.

Discussion

Principal Findings

This study tested the feasibility and preliminary effectiveness of a novel exposure therapy training program applying VR technology within an E2E framework. Two delivery modalities with varying degrees of immersion were piloted as different experiential doses of E2E practice to evaluate each method's usability and impact on key training targets. Clinicians were randomized to training programs of low immersion (desktop computer) and high immersion (HMD) and assessed for changes in exposure knowledge, self-efficacy, and attitudes about exposure therapy.

Our primary outcomes demonstrate that virtual technology is highly usable and effective in engaging relevant therapist-level learning variables during exposure therapy training. Participants rated both virtual programs highly on the SUS, which measures the programs' ease of use, intuitiveness, and functionality. Likewise, the programs received high scores on the VP evaluation scale, indicating that the virtual patient felt realistic to the therapists. This is especially promising as one challenge in conventional training is the limited ability of role-play to capture the real experience of delivering exposures [12].

Two of the primary training metrics in this study were therapists' perceptions of their efficacy in delivering exposure therapy and their attitudes about the safety and tolerability of exposure for their clients. Both metrics have been previously identified as therapist-level barriers that limit the frequency and quality of exposure delivery in typical practice [18]. Participants showed significant improvements in the self-efficacy measure, which has been shown to have positive predictive validity for future exposure use [20]. The improvements in TBES scores between T1 (after didactic training) and T2 (after SET-VR) demonstrate that virtual experiential learning further reduced negative beliefs and increased comfort with performing exposures. Notably, these findings were sustained at the 4-week follow-up, indicating potential long-term gains from virtual training. Participants did not show changes in levels of knowledge between T1 and T2. This is to be expected as the SET-VR program focuses on experiential learning with the assumption that the therapist has the necessary theoretical underpinnings. However, our results support virtual experiential learning as a promising option for further development to address key barriers in the current models of exposure therapy training. In contrast, exposure knowledge was included primarily as a control-like outcome to verify that both groups benefited similarly from the didactic workshop and entered the VR phase with comparable background knowledge. This approach is consistent with prior work showing that knowledge about exposure is necessary but far from sufficient to change exposure use or quality, whereas self-efficacy and negative beliefs represent more central barriers to implementation [9].

Interpreted through the E2E lens, these findings suggest that VR-based experiential practice can function as an exposure intervention directed at therapists' own anxious beliefs about exposure. The additional reductions in negative beliefs and gains in self-efficacy from T1 (post-didactic) to T2 (post-experiential) indicate that repeatedly "doing" exposure with a distressed virtual patient can provide corrective learning beyond what is achievable through didactic instruction alone, even when knowledge is already adequate. This psychotherapy-specific mechanism focus—targeting therapist beliefs that uniquely impede exposure implementation—was central to the rationale for using VR in this trial.

The results from the direct comparison of the 2 delivery conditions addressed our central question about how much immersion is needed for E2E training and did not support our initial hypothesis that the high-immersion condition (HMD) would outperform the low-immersion condition (desktop computer). Participants in both conditions reported comparably high values for the replication of real-life scenarios on VP Eval and satisfaction with the training on the TARS. While both groups gave high ratings on usability (SUS), the low immersion group's rating was higher at a statistically significant level. As for training outcomes, both conditions improved self-efficacy and negative attitudes without statistically significant between-group differences. Taken together, this finding shows that, at least within this feasibility study, the use of HMD to deliver the virtual training was not superior to the use of a desktop computer in addressing key therapist-level targets, and the lower immersion method may have higher usability. From an E2E perspective, this pattern suggests that, in the context of a brief single-session training, a lower-immersion desktop may already provide sufficient presence and emotional engagement with a distressed virtual patient to facilitate learning about the safety and tolerability of exposure. Higher-immersion HMD may yield incremental benefits only under conditions of greater training "dose" (eg, more repetitions, more complex scenarios) or in subgroups of learners; this is a question that will require a larger, mechanism-focused trial.

In the context of the larger medical education literature, there may also be a task-technology mismatch between virtual

XSL•FO
RenderX

exposure therapy delivery and high-immersion VR headset. Prior studies demonstrating a positive correlation between immersion factors and medical education learning outcomes have shown that the effect is most significant in spatial and procedural tasks, such as surgical procedures [42] and physical examinations [43]. Conducting exposure therapy does have procedural elements, like the sequential progression of actions based on branching decision points. However, our participants' tasks are different from physical procedures in that they do not need a high degree of spatial awareness or physical interaction with the virtual environment, which would be the main advantages of an HMD [44]. Thus, it is possible that visual and auditory stimuli through a desktop computer provided the same necessary input for experiential learning of delivering exposures.

The incremental benefits to learning provided by higher immersion could have been negated by the increased cognitive demands that arise from operating an HMD. Task-irrelevant sensory information or the mental effort required to navigate the less familiar HMD interface could have been distractors from the key experiential task. This would be consistent with the cognitive load theory framework, which states that extraneous cognitive demands reduce mental resources needed for optimal learning [45,46]. Additionally, methodological factors, such as participation selection bias and sample size, may have confounded the detection of between-group differences. One of the most well-studied benefits of high immersion VR is increased user engagement with the material [44]. Given that participants are a self-selected group who are motivated in exposure therapy training, this effect may have been overshadowed. It was anecdotally observed by the study team that participants in the HMD group demonstrated more consistent engagement with the exercise. It is possible that over a longer duration of training, differences in engagement could impact learner outcomes.

Overall, our findings present a basis for further investigation into how immersion functions as an experiential dose parameter within E2E-based psychotherapy training, as well as an opportunity for flexible implementation of VR technology based on factors such as technology literacy, cost-effectiveness, and learner preferences. Most existing studies of VR training in mental health focus on skills, knowledge, and attitudes for specific diagnoses (eg, depression, psychosis) [29]. In contrast, while this study involves a virtual patient with anxiety disorders, its primary aim is to enhance learner outcomes with a specific behavioral health technique (exposure therapy). In this regard, very little research has been done to evaluate procedural learning in VR for behavioral therapy. One notable review identified technology-based methods for training behavioral counseling skills [27]. It found that a mix of virtually created patients and filmed actors was being used. A majority of available outcomes were related to using simulations as opportunities for the repetition of single skills (affirmation, reflective listening, change talk, psychoeducation) [27,47,48], rather than aiming to deliver experiential learning for high-stress clinical scenarios or to address provider beliefs about a treatment modality.

To our knowledge, this is the first study to leverage virtual technology to augment exposure therapy training. The current norm of didactics-focused training is effective in instilling knowledge and, to some extent, improving clinician self-efficacy and negative beliefs [49], as demonstrated even in our single-session training intervention. However, the literature also shows that provider-level barriers continue to limit the optimal dissemination of exposure therapy [13], despite it being one of the most effective tools in psychiatry [18]. Thus, our finding that the supplementation of didactic learning with virtual experiential learning further improves self-efficacy and negative beliefs about exposure therapy suggests a promising area for further investigation to optimize the standardization and scalability of exposure training. This is particularly notable as VR technology in itself offers logistical advantages (accessibility, scalability, reproducibility), the ability to replicate rare or high-risk situations in a safe training space, and near-limitless customizability [50].

Our study design contains notable limitations. There were clear imbalances in the participant demographics. A large majority (90%) of participants were White, non-Hispanic, and female (85%). The gender imbalance may be particularly pertinent, as there have been heterogeneous results regarding the relationship between gender and VR experiences such as immersion, emotional response, and task performance [51]. There is also the potential for selection bias in the recruited participants. While technology-based training interventions are generally well accepted [27], there may be clinicians who prefer other modalities. Those who opted to join this study may be more interested in innovation or technology, which could have influenced their experience or learning outcomes. Additionally, participants were self-selected based on their motivation for exposure therapy training. Although it was not observed in our study, it is possible that cyber sickness and the initial learning curve of new digital platforms may present barriers to engagement in a more general clinician population.

This feasibility trial is designed to assess the effectiveness of SET-VR in addition to didactic training; thus, it is not possible to draw direct comparisons between SET-VR and standard didactic training alone. While a variety of carefully selected metrics were used to assess training efficacy, there are no direct measures of how the quality or frequency of exposures delivered by clinicians changed after SET-VR. Our most reliable proxy is the improvements in exposure self-efficacy, which has previously been shown to have predictive validity in the future likelihood of using exposure therapy [20]. Finally, the current version of SET-VR does not use novel algorithms or adaptive rendering techniques; instead, it uses a controlled, pre-programmed decision structure to standardize the training experience. While this limits technological novelty, it was appropriate for a feasibility trial focused on mechanism-relevant learning outcomes.

This study offers directions for further investigations. In the current trial, the SET-VR architecture intentionally relied on a relatively simple, preprogrammed decision tree and automated SUDS trajectory, rather than more complex adaptive algorithms, to maximize standardization and internal validity for testing immersion and belief change. Future iterations are planned to incorporate more complex interaction logic, including natural language processing–based virtual patient responses and richer affective behaviors. Future iterations will also include a longer

XSL•FO
RenderX

training course to see if empirical observations of differences in engagement between immersion levels lead to significant differences in learning. As with other training research, it will be important to evaluate how performance in the exposure simulation sessions correlates with behaviors in actual clinical practice. Noteworthy outcomes could include an increased likelihood of using exposures overall, as well as higher fidelity to the principles of effective exposure therapy [18]. To that end, a larger training trial with practicing clinicians is planned, which will compare decisions during the simulation against coded video tapes of exposure delivery with patients after training to evaluate how performance during training aligns with future delivery.

Conclusions

The adaptation of VR technology into clinical training offers new opportunities to overcome longstanding limitations in the accessibility, standardization, and experiential quality of exposure training. This study is innovative in applying an explicit "E2E" framework to VR, using simulated exposure sessions to directly target therapists' anxious beliefs about exposure, and in its examination of immersion as a dose parameter rather than assuming that "more immersive" is inherently better. In contrast to much of the existing VR training literature, which has focused primarily on knowledge and procedural skill acquisition, our findings highlight that both desktop and HMD implementations can produce meaningful changes in self-efficacy and negative beliefs, which are mechanisms that are uniquely tied to exposure underuse in routine care. The lack of clear incremental benefit for HMD over desktop VR suggests that lower-immersion, lower-cost formats may be sufficient to deliver an effective experiential dose of exposure practice, which has important implications for scalability and implementation in community and academic settings. Taken together, these results provide preliminary but encouraging evidence that E2E-guided VR training can enhance conventional didactic training and may be a promising next step in optimizing the standardization and effectiveness of exposure training.

Acknowledgments

This manuscript is solely the work of the authors; no generative artificial intelligence was used to create any portion of the content.

Funding

This study was funded by the National Institute of Mental Health (R41MH131229).

Data Availability

Study data have been shared with the National Institute of Mental Health Data Archive to facilitate data sharing. For additional access to data from this study, contact the corresponding author directly.

Conflicts of Interest

None declared.

Checklist 1
CONSORT 2025 checklist.
[PDF File, 97 KB - mededu_v12i1e80087_app1.pdf]

Checklist 2
CONSORT-EHEALTH checklist.
[PDF File, 1050 KB - mededu_v12i1e80087_app2.pdf]

References

1. Racine N, McArthur BA, Cooke JE, Eirich R, Zhu J, Madigan S. Global prevalence of depressive and anxiety symptoms in children and adolescents during COVID-19: a meta-analysis. JAMA Pediatr 2021 Nov 1;175(11):1142-1150. [doi: 10.1001/jamapediatrics.2021.2482] [Medline: 34369987]
2. Bittner A, Egger HL, Erkanli A, Jane Costello E, Foley DL, Angold A. What do childhood anxiety disorders predict? J Child Psychol Psychiatry 2007 Dec;48(12):1174-1183. [doi: 10.1111/j.1469-7610.2007.01812.x] [Medline: 18093022]
3. Ezpeleta L, Keeler G, Erkanli A, Costello EJ, Angold A. Epidemiology of psychiatric disability in childhood and adolescence. J Child Psychol Psychiatry 2001 Oct;42(7):901-914. [doi: 10.1111/1469-7610.00786] [Medline: 11693585]
4. Crowe K, McKay D. Efficacy of cognitive-behavioral therapy for childhood anxiety and depression. J Anxiety Disord 2017 Jun;49:76-87. [doi: 10.1016/j.janxdis.2017.04.001] [Medline: 28460329]
5. Sars D, van Minnen A. On the use of exposure therapy in the treatment of anxiety disorders: a survey among cognitive behavioural therapists in the Netherlands. BMC Psychol 2015;3(1):26. [doi: 10.1186/s40359-015-0083-2] [Medline: 26246900]

XSL•FO
RenderX

6. Marcks BA, Weisberg RB, Keller MB. Psychiatric treatment received by primary care patients with panic disorder with and without agoraphobia. Psychiatr Serv 2009 Jun;60(6):823-830. [doi: 10.1176/ps.2009.60.6.823] [Medline: 19487354]
7. Becker CB, Zayfert C, Anderson E. A survey of psychologists' attitudes towards and utilization of exposure therapy for PTSD. Behav Res Ther 2004 Mar;42(3):277-292. [doi: 10.1016/S0005-7967(03)00138-4] [Medline: 14975770]
8. Freiheit SR, Vye C, Swan R, Cady M. Cognitive-behavioral therapy for anxiety: is dissemination working? Behav Ther (N Y N Y) 2004;27(2):25-32 [FREE Full text]
9. Beidas RS, Kendall PC. Training therapists in evidence-based practice: a critical review of studies from a systems-contextual perspective. Clin Psychol (New York) 2010 Mar;17(1):1-30. [doi: 10.1111/j.1468-2850.2009.01187.x] [Medline: 20877441]
10. Sholomskas DE, Syracuse-Siewert G, Rounsaville BJ, Ball SA, Nuro KF, Carroll KM. We don't train in vain: a dissemination trial of three strategies of training clinicians in cognitive-behavioral therapy. J Consult Clin Psychol 2005 Feb;73(1):106-115. [doi: 10.1037/0022-006X.73.1.106] [Medline: 15709837]
11. Beidas RS, Cross W, Dorsey S. Show me, don't tell me: behavioral rehearsal as a training and analogue fidelity tool. Cogn Behav Pract 2014 Feb;21(1):1-11. [doi: 10.1016/j.cbpra.2013.04.002] [Medline: 25382963]
12. Farrell NR, Kemp JJ, Blakey SM, Meyer JM, Deacon BJ. Targeting clinician concerns about exposure therapy: a pilot study comparing standard vs. enhanced training. Behav Res Ther 2016 Oct;85:53-59. [doi: 10.1016/j.brat.2016.08.011] [Medline: 27567972]
13. van Minnen A, Hendriks L, Olff M. When do trauma experts choose exposure therapy for PTSD patients? A controlled study of therapist and patient factors. Behav Res Ther 2010 Apr;48(4):312-320. [doi: 10.1016/j.brat.2009.12.003] [Medline: 20056195]
14. Olatunji BO, Deacon BJ, Abramowitz JS. The cruelest cure? Ethical issues in the implementation of exposure-based treatments. Cogn Behav Pract 2009 May;16(2):172-180. [doi: 10.1016/j.cbpra.2008.07.003]
15. Rosqvist J. Exposure Treatments for Anxiety Disorders: A Practitioner's Guide to Concepts, Methods, and Evidence-Based Practice: Routledge; 2006.
16. Richard DCS, Gloster AT. Exposure therapy has a public relations problem: a dearth of litigation amid a wealth of concern. In: Richard DCS, Lauterbach D, editors. Handbook of Exposure Therapies: Academic Press; 2007:409-425. [doi: 10.1016/B978-012587421-2/50019-3]
17. Gola JA, Beidas RS, Antinoro-Burke D, Kratz HE, Fingerhut R. Ethical considerations in exposure therapy with children. Cogn Behav Pract 2016 May;23(2):184-193. [doi: 10.1016/j.cbpra.2015.04.003] [Medline: 27688681]
18. Abramowitz JS, Deacon BJ, Whiteside SPH. Exposure Therapy for Anxiety Principles and Practice, 2nd edition: Guilford Press; 2014.
19. Deacon BJ, Lickel JJ, Farrell NR, Kemp JJ, Hipol LJ. Therapist perceptions and delivery of interoceptive exposure for panic disorder. J Anxiety Disord 2013 Mar;27(2):259-264. [doi: 10.1016/j.janxdis.2013.02.004] [Medline: 23549110]
20. Harned MS, Dimeff LA, Woodcock EA, Contreras I. Predicting adoption of exposure therapy in a randomized controlled dissemination trial. J Anxiety Disord 2013 Dec;27(8):754-762. [doi: 10.1016/j.janxdis.2013.02.006] [Medline: 23538148]
21. Blakey SM, Abramowitz JS. The effects of safety behaviors during exposure therapy for anxiety: critical analysis from an inhibitory learning perspective. Clin Psychol Rev 2016 Nov;49:1-15. [doi: 10.1016/j.cpr.2016.07.002] [Medline: 27475477]
22. Goisman RM, Warshaw MG, Keller MB. Psychosocial treatment prescriptions for generalized anxiety disorder, panic disorder, and social phobia, 1991-1996. Am J Psychiatry 1999 Nov;156(11):1819-1821. [doi: 10.1176/ajp.156.11.1819] [Medline: 10553751]
23. Dhar E, Upadhyay U, Huang Y, et al. A scoping review to assess the effects of virtual reality in medical education and clinical care. Digit Health 2023;9:20552076231158022. [doi: 10.1177/20552076231158022] [Medline: 36865772]
24. Kemp J, Benito K, Herren J, Brown Z, Frank HE, Freeman J. Exposure to exposure: a protocol for leveraging exposure principles during training to address therapist-level barriers to exposure implementation. Front Psychiatry 2023;14:1096259. [doi: 10.3389/fpsyt.2023.1096259] [Medline: 36873204]
25. Frank HE, Becker-Haimes EM, Rifkin LS, et al. Training with tarantulas: a randomized feasibility and acceptability study using experiential learning to enhance exposure therapy training. J Anxiety Disord 2020 Dec;76:102308. [doi: 10.1016/j.janxdis.2020.102308] [Medline: 32992268]
26. Blade RA, Padgett ML. Virtual environments standards and terminology. In: Handbook of Virtual Environments: Design, Implementation, and Applications: CRC Press; 2002:55-68.
27. Magill M, Mastroleo NR, Martino S. Technology-based methods for training counseling skills in behavioral health: a scoping review. J Technol Behav Sci 2022;7(3):325-336. [doi: 10.1007/s41347-022-00252-8] [Medline: 35402697]
28. Kim GM, Lim JY, Kim EJ, Yeom M. Impact of virtual reality mental health nursing simulation on nursing students' competence. J Multidiscip Healthc 2024;17:191-202. [doi: 10.2147/JMDH.S435986] [Medline: 38226028]
29. Steen CW, Söderström K, Stensrud B, Nylund IB, Siqveland J. The effectiveness of virtual reality training on knowledge, skills and attitudes of health care professionals and students in assessing and treating mental health disorders: a systematic review. BMC Med Educ 2024 May 1;24(1):480. [doi: 10.1186/s12909-024-05423-0] [Medline: 38693509]
30. Slater M, Khanna P, Mortensen J, Yu I. Visual realism enhances realistic response in an immersive virtual environment. IEEE Comput Graph Appl 2009;29(3):76-84. [doi: 10.1109/mcg.2009.55] [Medline: 19642617]

XSL•FO
RenderX

31. Persky S, Kaphingst KA, McCall C, Lachance C, Beall AC, Blascovich J. Presence relates to distinct outcomes in two virtual environments employing different learning modalities. Cyberpsychol Behav 2009 Jun;12(3):263-268. [doi: 10.1089/cpb.2008.0262] [Medline: 19366319]
32. Aitamurto T, Zhou S, Sakshuwong S, Saldivar J, Sadeghi Y, Tran A. Sense of presence, attitude change, perspective-taking and usability in first-person split-sphere 360° video. Presented at: Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems; Apr 21-26, 2018. [doi: 10.1145/3173574.3174119]
33. Xu X, Kang J, Yan L. Understanding embodied immersion in technology-enabled embodied learning environments. Comput Assist Learn 2022 Feb;38(1):103-119. [doi: 10.1111/jcal.12594]
34. Johnson-Glenberg MC, Megowan-Romanowicz C, Birchfield DA, Savio-Ramos C. Effects of embodied learning and digital platform on the retention of physics content: centripetal force. Front Psychol 2016;7:1819. [doi: 10.3389/fpsyg.2016.01819] [Medline: 27933009]
35. Wenk N, Penalver-Andres J, Buetler KA, Nef T, Müri RM, Marchal-Crespo L. Effect of immersive visualization technologies on cognitive load, motivation, usability, and embodiment. Virtual Real 2023;27(1):307-331. [doi: 10.1007/s10055-021-00565-8] [Medline: 36915633]
36. March M, Zackoff M, Fleck J, et al. A randomized trial of virtual reality training to improve firearm safety counseling skills. Acad Pediatr 2025;25(1):102560. [doi: 10.1016/j.acap.2024.08.005] [Medline: 39128562]
37. Brooke J. SUS: a 'quick and dirty' usability scale. In: Jordan PW, Thomas B, Weerdmeester B, McClelland IL, editors. Usability Evaluation In Industry: CRC Press; 1996:189-194.
38. Huwendiek S, de Leng BA. Virtual patient design and curricular integration evaluation toolkit. Med Educ 2010 May;44(5):519-519. [doi: 10.1111/j.1365-2923.2010.03665.x] [Medline: 20519003]
39. Davis JR, Rawana EP, Capponi DR. Acceptability of behavioral staff management techniques. Behav Interv 1989 Jan;4(1):23-44. [doi: 10.1002/bin.2360040104]
40. Hopewell S, Chan AW, Collins GS, et al. CONSORT 2025 statement: updated guideline for reporting randomised trials. Lancet 2025 Apr 14;405(10489):P1633-P1640. [doi: 10.1016/S0140-6736(25)00672-5] [Medline: 40245901]
41. Lewis JR, Sauro J. Item benchmarks for the system usability scale. J Usability Stud 2018;13(3):158-167 [FREE Full text]
42. Dorey T, Nicholas J, Leydorf SD, et al. Assessing immersive virtual reality as learning tool for surgical trainees. JSLS 2025;29(2):e2025.00021. [doi: 10.4293/JSLS.2025.00021] [Medline: 40342609]
43. Chao YP, Kang CJ, Chuang HH, et al. Comparison of the effect of 360° versus two-dimensional virtual reality video on history taking and physical examination skills learning among undergraduate medical students: a randomized controlled trial. Virtual Real 2023;27(2):637-650. [doi: 10.1007/s10055-022-00664-0] [Medline: 35992202]
44. Stracke CM, Bothe P, Adler S, et al. Immersive virtual reality in higher education: a systematic review of the scientific literature. Virtual Real 2025;29(2):1-21. [doi: 10.1007/s10055-025-01136-x]
45. Fraser KL, Ayres P, Sweller J. Cognitive load theory for the design of medical simulations. Simul Healthc 2015 Oct;10(5):295-307. [doi: 10.1097/SIH.0000000000000097] [Medline: 26154251]
46. Andersen SAW, Mikkelsen PT, Konge L, Cayé-Thomasen P, Sørensen MS. The effect of implementing cognitive load theory-based design principles in virtual reality simulation training of surgical skills: a randomized controlled trial. Adv Simul (Lond) 2016;1:20. [doi: 10.1186/s41077-016-0022-1] [Medline: 29449989]
47. Washburn M, Parrish DE, Bordnick PS. Virtual patient simulations for brief assessment of mental health disorders in integrated care settings. Soc Work Ment Health 2020;18(2):121-148. [doi: 10.1080/15332985.2017.1336743] [Medline: 32952451]
48. Reger GM, Norr AM, Rizzo AS, et al. Virtual standardized patients vs academic training for learning motivational interviewing skills in the US Department of Veterans Affairs and the US military: a randomized trial. JAMA Netw Open 2020 Oct 1;3(10):e2017348. [doi: 10.1001/jamanetworkopen.2020.17348] [Medline: 33057643]
49. Whiteside SPH, Deacon BJ, Benito K, Stewart E. Factors associated with practitioners' use of exposure therapy for childhood anxiety disorders. J Anxiety Disord 2016 May;40:29-36. [doi: 10.1016/j.janxdis.2016.04.001] [Medline: 27085463]
50. Bell IH, Nicholas J, Alvarez-Jimenez M, Thompson A, Valmaggia L. Virtual reality as a clinical tool in mental health research and practice. Dialogues Clin Neurosci 2020 Jun;22(2):169-177. [doi: 10.31887/DCNS.2020.22.2/lvalmaggia] [Medline: 32699517]
51. Grassini S, Laumann K. Are modern head-mounted displays sexist? A systematic review on gender differences in HMD-mediated virtual reality. Front Psychol 2020;11:1604. [doi: 10.3389/fpsyg.2020.01604] [Medline: 32903791]

Abbreviations

CBT: cognitive behavioral therapy
CONSORT: Consolidated Standards of Reporting Trials
CONSORT-EHEALTH: Consolidated Standards of Reporting Trials of Electronic and Mobile Health Applications and Online Telehealth
E2E: exposure to exposure
GEE: generalized estimating equation

XSL•FO
RenderX

HMD: head-mounted display
REDCap: Research Electronic Data Capture
SUDS: Subjective Units of Distress Scale
SUS: System Usability Scale
TARS: Training Acceptability Rating Scale
TBES: Therapist Negative Beliefs about Exposure Scale
VE: virtual environment
VP Eval: Virtual Patient Evaluation
VR: virtual reality

Edited by S Brini, T Leung; submitted 03.Jul.2025; peer-reviewed by AY Saleh, T Mühling; revised version received 07.Jan.2026; accepted 17.Jan.2026; published 21.May.2026.

Please cite as:
Kemp J, Rosen A, Kim H, Adams Larsen M, Benito K, Freeman J, Machan J, Tuerk P
Enhancing Therapist Training in the Delivery of Exposure Therapy for Individuals with Anxiety Disorders Using Virtual Reality Simulation: Randomized Feasibility Trial
JMIR Med Educ 2026;12:e80087
URL: https://mededu.jmir.org/2026/1/e80087
doi:10.2196/80087

© Joshua Kemp, Ariella Rosen, Hyungjin Kim, Margo Adams Larsen, Kristen Benito, Jennifer Freeman, Jason Machan, Peter Tuerk. Originally published in JMIR Medical Education (https://mededu.jmir.org), 21.May.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Education, is properly cited. The complete bibliographic information, a link to the original publication on https://mededu.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

“AACHEN” e-Learning Tool in Augmentative and Alternative Communication for Medical Students in Germany: Cross-Sectional Evaluation Study

Jessica Büchs[1], MA; Christiane Neuschaefer-Rube[2], MD

[1]Clinic for Otorhinolaryngology, Phoniatrics and Pedaudiology, University Hospital and Medical Faculty, RWTH Aachen University, Pauwelsstraße 30, Aachen, Germany

[2]Phoniatrics and Pedaudiology, Medical Faculty, RWTH Aachen University, Aachen, Germany

Corresponding Author:
Jessica Büchs, MA
Clinic for Otorhinolaryngology, Phoniatrics and Pedaudiology, University Hospital and Medical Faculty, RWTH Aachen University, Pauwelsstraße 30, Aachen, Germany

Abstract

We developed an e-learning tool in augmentative and alternative communication (AAC) for medical students in Germany and tested it at Rhenish-Westphalian Technical University, Aachen. Our cross-sectional evaluation study is an innovative approach since the topic is yet to be implemented in medical education. Other AAC tools do not target medical students. Our study underlines the importance of AAC for medical doctors. Universities should include AAC in their lectures to prepare medical students for their clinical practice. The “*AAC*HEN” tool is a first step.

(JMIR Med Educ 2026;12:e88173) doi:10.2196/88173

KEYWORDS

medical students; e-learning; digital learning; e-learning tool; augmentative and alternative communication; AAC; teaching; phoniatrics

Introduction

Augmentative and alternative communication (AAC) is the study of communicative strategies for patients who cannot speak [1], including the use of body language, symbol-boards, or technological devices [2,3]. Patients who use AAC are most common in medical fields related to speech and language such as phoniatrics, pediatrics, neurology, and otorhinolaryngology. However, every medical doctor may treat patients with complex communication needs. Therefore, knowledge in AAC is relevant in all medical specialties. Hurtig et al [4] show a lack of training in the field when implementing AAC in intensive care. What tools are available to meet the need for instruction? In a previous study, we gained insight into existing AAC tools for medical students and professionals [5]. No free-access, audio-visual, German online course with a knowledge quiz was found. Now, we fill this gap with the “*AAC*HEN” tool. The objectives of this study are to (1) underline the importance of AAC for medical doctors, (2) assess the students’ evaluations of our tool, and (3) give recommendations for future application in medical education.

Methods

Ethical Considerations

This study involved human participants and was reviewed and approved by the ethics committee of the Medical Faculty at Rheinisch-Westfälische Technische Hochschule (RWTH; Rhenish-Westphalian Technical University in English) Aachen University Hospital (no EK 24 - 022; registration no CTC-A 24 - 035). The students were informed about the methods and objectives of this study through a brief presentation. An information sheet was available in paper form and as a digital document. Following the provision of information, participants signed a consent form about the collection and analysis of the data. A full data protection impact assessment was not required since the feedback form was anonymous. Participation was voluntary, not compensated, and could be ended at any time. Identification of individual participants is not possible in any parts of this manuscript or supplementary material.

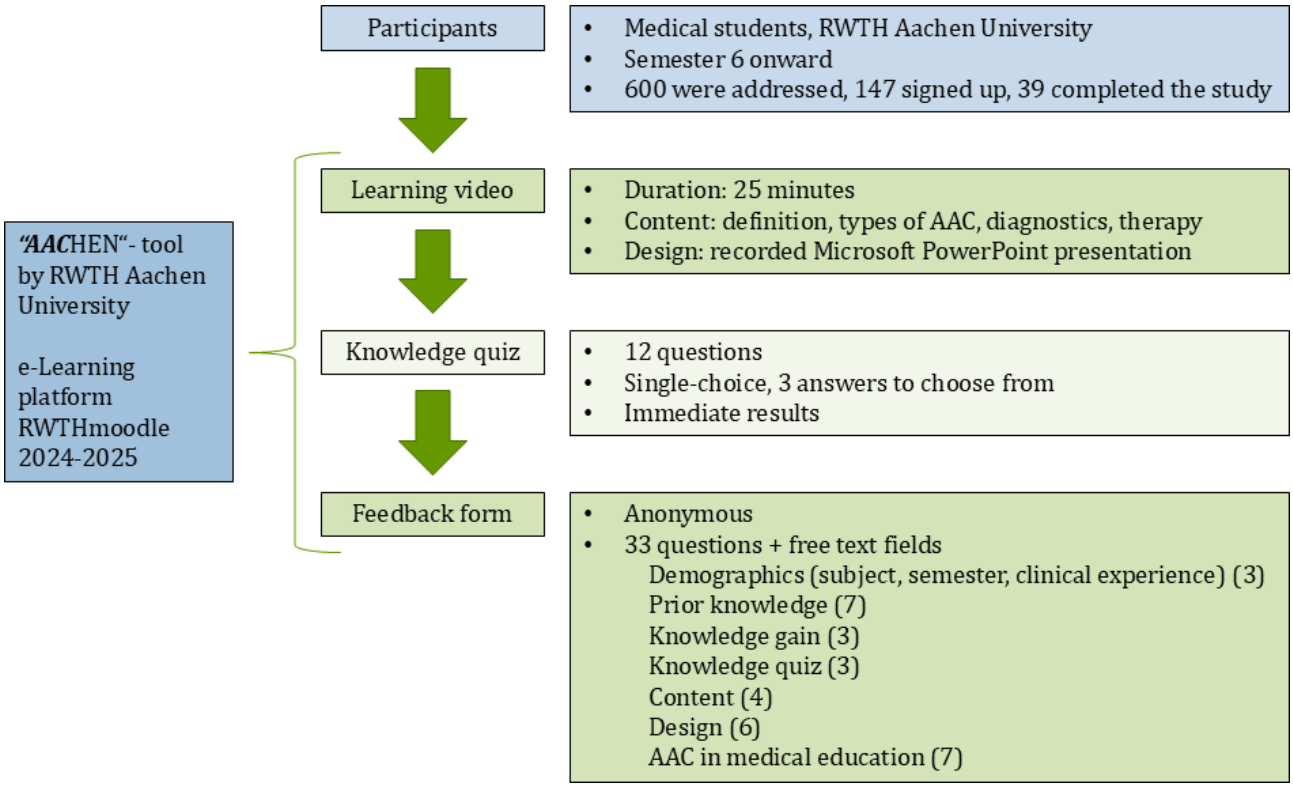
Study Design: Setting, Participants, Recruitment, and Sample Size

The tool was developed at the University Hospital and Medical Faculty RWTH Aachen from October 2023 until May 2024. Recruitment and evaluation lasted from June 2024 until August 2025. The study design was cross-sectional as students gave feedback after having worked through the tool. The inclusion

XSL•FO
RenderX

criterion was the enrollment in the medical study course from semester 6 onward. We used convenience sampling for recruiting (ie, announcing in lectures). Approximately 600 students were addressed, 147 signed up, and 39 completed the study (Figure 1 and Multimedia Appendix 1).

Figure 1. Overview of this study. The numbers in parentheses represent the number of questions. AAC: augmentative and alternative communication; RWTH: Rheinisch-Westfälische Technische Hochschule.



The "*AAC*HEN" Tool: Development, Review, Duration, Content, and Design

The tool consists of a learning video and a knowledge quiz (Multimedia Appendix 2). It was developed by author JB after she studied the literature and reviewed by medical doctors and speech-language pathologists. The video is a presentation of 25 minutes. The knowledge quiz is announced at the beginning. The presentation covers types of AAC, patients who could benefit from AAC, and which patient could benefit from what type of AAC (Multimedia Appendix 3).

Data Assessment, Measurement, and Statistical Analysis

The students' evaluations were assessed with a feedback form on RWTHmoodle (Multimedia Appendix 4). Levels of agreement to statements pertaining to the importance for medical doctors, inclusion in the curriculum, knowledge gain, and benefit were measured on Likert scales. Content, design, and prior knowledge were measured using the German grading system. Effects of the knowledge quiz were assessed using a multiple-choice format. Knowledge quiz results and feedback form data were analyzed. Means and confidence intervals were calculated using R software (version 4.5.2; R Foundation for Statistical Computing) (Multimedia Appendix 1).

Results

Almost all students (38/39, 97.4%; 95% CI 86.5% - 99.9%) agreed to AAC being an important topic for medical doctors and the majority agreed that it should be included in their curriculum (30/39, 76.9%; 95% CI 60.7% - 88.9%). Almost none of the students had ever heard of AAC or had an idea of it (38/39, 97.4%; 95% CI 86.5% - 99.9%). Almost all gained knowledge from our tool (38/39, 97.4%; 95% CI 86.5% - 99.9%) and perceived this knowledge to be beneficial for their future clinical practice (37/39, 94.9%; 95% CI 82.7% - 99.4%). The announcement of the knowledge quiz had positive effects; students paid more attention (27/39, 69.2%; 95% CI 52.4% - 83%) and memorized the content better (21/39, 53.9%; 95% CI 37.2% - 69.9%). Most of the students stated that the "*AAC*HEN" tool should be a fixed part in their curriculum (29/39, 74.4%; 95% CI 57.9% - 87%). The "*AAC*HEN" tool received "very good" grades in terms of content (mean value 1.3, 95% CI 1.1 - 1.6) and design (mean value 1.2, 95% CI 1 - 1.3). The mean value of correct results in the knowledge quiz was 92.3% (95% CI 90.2% - 94.3%) (Table 1, Table 2, and Multimedia Appendix 5).

XSL•FO
RenderX

Table . Students' answers in the feedback form (N=39).

	Agreement		Neutral, n (%)	Disagreement		Proportion agreement, n (%)	95 % CI (%)
	Yes, n (%)	Rather yes, n (%)		Rather no, n (%)	No, n (%)		
Importance for MD[a]	26 (66.7)	12 (30.8)	1 (2.6)	0 (0)	0 (0)	38 (97.4)	86.5 - 99.9
AAC[b] in curriculum	14 (35.9)	16 (41.0)	9 (23.1)	0 (0)	0 (0)	30 (76.9)	60.7 - 88.9
Knowledge gain "AACHEN"	34 (87.2)	4 (10.3)	1 (2.6)	0 (0)	0 (0)	38 (97.4)	86.5 - 99.9
Knowledge gain benefit MD	21 (53.9)	16 (41.0)	2 (5.1)	0 (0)	0 (0)	37 (94.9)	82.7 - 99.4
"AACHEN" in curriculum	14 (35.9)	15 (38.5)	8 (20.5)	2 (5.1)	0 (0)	29 (74.4)	57.9 - 87

[a]MD: medical doctors.

[b]AAC: augmentative and alternative communication.

Table . Students' answers in the feedback form (N=39).

	1=Very good, n (%)	2=Good, n (%)	3=Satisfactory, n (%)	4=Sufficient, n (%)	5=Poor, n (%)	6=Deficient, n (%)	Value of grades, 1-6, mean (95% CI)
Prior knowledge in AAC	1 (2.6)	4 (10.3)	6 (15.4)	14 (35.9)	9 (23.1)	5 (12.8)	4.1 (3.6-4.5)
Content of "AACHEN"	30 (76.9)	8 (20.5)	0 (0)	0 (0)	0 (0)	1 (2.6)	1.3 (1.1-1.6)
Design of "AACHEN"	33 (84.6)	6 (15.4)	0 (0)	0 (0)	0 (0)	0 (0)	1.2 (1-1.3)

[a]AAC: augmentative and alternative communication.

Discussion

Principal Findings

We are not surprised that most of the students were not familiar with AAC since the topic is yet to be implemented in medical education. It is gratifying to note that the feedback about our tool was very positive and that the students gained knowledge from it. Although our study had a small sample size, it underlines the importance of AAC for medical doctors. While our research focuses on the testing of an e-learning tool, other studies address the importance of including the voices of AAC users in the instructional process [6,7]. We support a previously stated interest of learners in first-hand experience with AAC users [8] (Multimedia Appendix 6).

Future Directions

We plan to combine our "*AAC*HEN" tool with an in-person lecture, where different types of AAC can be explored in cooperation with AAC users. The "*AAC*HEN" tool will be added to the RWTH Aachen toolbox app [9]. In addition, it could be added to an AAC online platform for German-speaking learners [10]. We would like to provide nationwide access to our tool and collaborate with other medical faculties.

Conclusion

Our study underlines the importance of AAC in medicine and the lack of students' knowledge in the field. AAC is not yet taught to medical students, and appropriate e-learning tools do not yet exist. Our "*AAC*HEN" tool fills a significant gap in the teaching of AAC. It can be included in lectures and be combined with practical instructions to improve medical education.

Funding

Open access funding provided by the Open Access Publishing Fund of RWTH Aachen University.

Data Availability

All data discussed in this article are provided in Multimedia Appendix 5.

XSL•FO
RenderX

Conflicts of Interest

None declared.

Multimedia Appendix 1
Detailed description of methods.
[DOCX File, 23 KB - mededu_v12i1e88173_app1.docx]

Multimedia Appendix 2
Knowledge quiz (original German version).
[DOCX File, 24 KB - mededu_v12i1e88173_app2.docx]

Multimedia Appendix 3
Detailed description of the "*AAC*HEN" tool.
[DOCX File, 24 KB - mededu_v12i1e88173_app3.docx]

Multimedia Appendix 4
Feedback form (original German version).
[PDF File, 198 KB - mededu_v12i1e88173_app4.pdf]

Multimedia Appendix 5
Detailed results.
[DOCX File, 47 KB - mededu_v12i1e88173_app5.docx]

Multimedia Appendix 6
Detailed discussion.
[DOCX File, 24 KB - mededu_v12i1e88173_app6.docx]

References

1. Zinkevich A, Uthoff SAK, Boenisch J, Sachse SK, Bernasconi T, Ansmann L. Complex intervention in augmentative and alternative communication (AAC) care in Germany: a study protocol of an evaluation study with a controlled mixed-methods design. BMJ Open 2019 Aug 28;9(8):e029469. [doi: 10.1136/bmjopen-2019-029469] [Medline: 31467052]
2. Griffiths T, Clarke M, Price K. Augmentative and alternative communication for children with speech, language and communication needs. Paediatr Child Health (Oxford) 2022 Aug;32(8):277-281. [doi: 10.1016/j.paed.2022.05.001]
3. Williams K, Holyfield C. Future of AAC technologies: priorities for inclusive innovation. Augment Altern Commun 2025 Sep;41(3):211-214. [doi: 10.1080/07434618.2025.2513906] [Medline: 40515715]
4. Hurtig RR, Alper RM, Bryant KNT, Davidson KR, Bilskemper C. Improving patient safety and patient–provider communication. Perspect ASHA SIGs 2019 Oct 31;4(5):1017-1027. [doi: 10.1044/2019_PERS-SIG12-2019-0021]
5. Büchs J, Neuschaefer-Rube C. e-Learning in phoniatrics and speech-language pathology: exploratory analysis of free access tools in augmentative and alternative communication. JMIR Med Educ 2025 Jun 27;11:e63392. [doi: 10.2196/63392] [Medline: 40577000]
6. Blasko G, Light J, McNaughton D, Williams B, Zimmerman J. Nothing about AAC users without AAC users: a call for meaningful inclusion in research, technology development, and professional training. Augment Altern Commun 2025 Sep;41(3):184-194. [doi: 10.1080/07434618.2025.2514748] [Medline: 40511843]
7. Balandin S, Hines M. The involvement of people with lifelong disability and communication impairment in lecturing to speech-language pathology students. Int J Speech Lang Pathol 2011 Oct;13(5):436-445. [doi: 10.3109/17549507.2011.513738] [Medline: 20964515]
8. Goldman A, Jackson D, Kannan K, et al. Training needs in augmentative and alternative communication: a virtual roundtable discussion. Assist Technol Outcomes Benefits 2021;15(1):1-28 [FREE Full text]
9. Lin Y, Lemos M, Neuschaefer-Rube C. Digital health and learning in speech-language pathology, phoniatrics, and otolaryngology: survey study for designing a digital learning toolbox app. JMIR Med Educ 2022 Apr 27;8(2):e34042. [doi: 10.2196/34042] [Medline: 35475980]
10. Burgio NM. E-Learning im Fort- und Weiterbildungsbereich Unterstützte Kommunikation (UK): Empfehlungen für die Gestaltung einer Onlineplattform UK und der Ausbildung zum E-Teacher UK [Article in German]. Behindertenpädagogik 2020;59(4):394-403. [doi: 10.30820/0341-7301-2020-4-394]

XSL•FO
RenderX

Abbreviations

AAC: augmentative and alternative communication
RWTH: Rheinisch-Westfälische Technische Hochschule

Edited by S Brini; submitted 20.Nov.2025; peer-reviewed by D McNaughton, S Ganesh; revised version received 22.Feb.2026; accepted 24.Feb.2026; published 29.Apr.2026.

Please cite as:
Büchs J, Neuschaefer-Rube C
"AACHEN" e-Learning Tool in Augmentative and Alternative Communication for Medical Students in Germany: Cross-Sectional Evaluation Study
JMIR Med Educ 2026;12:e88173
URL: https://mededu.jmir.org/2026/1/e88173
doi:10.2196/88173

© Jessica Büchs, Christiane Neuschaefer-Rube. Originally published in JMIR Medical Education (https://mededu.jmir.org), 29.Apr.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Education, is properly cited. The complete bibliographic information, a link to the original publication on https://mededu.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Speech- and Text-Based Emotion Recognition in Anesthesiology Residents During Critical Incident Simulation Training: Exploratory Observational Study

Sapir Gershov[1,2], PhD; Itay Bentov[3], MD, PhD; Fadi Mahameed[4,5], MD; Aeyal Raz[4,6], MD, PhD; Shlomi Laufer[2,5], PhD

[1]Department of Psychiatry, NYU Grossman School of Medicine, 1 Park Ave., 8th Floor, New York, NY, United States

[2]Technion Autonomous Systems Program, Technion – Israel Institute of Technology, Haifa, Israel

[3]Department of Anesthesiology and Pain Medicine, Harborview Medical Center, Seattle, WA, United States

[4]Department of Anesthesiology, Rambam Health Care Campus, Haifa, Israel

[5]Faculty of Data and Decision Sciences, Technion – Israel Institute of Technology, Haifa, Israel

[6]Rappaport Faculty of Medicine, Technion – Israel Institute of Technology, Haifa, Israel

Corresponding Author:
Sapir Gershov, PhD
Department of Psychiatry, NYU Grossman School of Medicine, 1 Park Ave., 8th Floor, New York, NY, United States

Abstract

Background: Communication during crisis management involves not only the exchange of clinical information but also the expression and regulation of emotional tone and sentiment, which may reflect clinician performance during a critical incident. In simulation-based medical education, these emotional dynamics are rarely measured objectively, limiting the ability to capture aspects of performance relevant to competency assessment.

Objective: This study aimed to examine whether automatically extracted speech- and text-based emotion signals varied across predefined phases of critical incident simulation training and whether these signals were associated with expert-rated performance among anesthesiology residents.

Methods: In this exploratory observational study, we analyzed 164 simulated crisis scenarios performed by 90 anesthesiology residents from 17 hospitals during a national board preparation workshop. Each high-fidelity simulation comprised 4 predefined phases: initial condition, extreme deterioration, advanced cardiovascular life support (ACLS) management, and recovery. Resident speech was isolated and analyzed using a speech-based emotion recognition (SER) model that estimated 8 emotion categories per utterance. Speech was also automatically transcribed and analyzed using text-based emotion recognition (TER), which estimated 9 emotion categories. Expert anesthesiologists rated performance on a 1-5 scale. ANOVA tested associations between emotion categories and performance, as well as variation across simulation phases. Exploratory mixed-effects sensitivity analyses accounted for repeated simulations and scenario-level clustering.

Results: After 7 recordings were excluded because of technical issues, 65 of 164 (39.6%) simulations were classified as poor performance (score ≤2), 72 (43.9%) as intermediate, and 27 (16.5%) as excellent (score=5). SER categories differed across performance groups ($F_{2,161}$=4.61; P=.02): poor performance showed more fear and surprise, whereas excellent performance showed more calm, happiness, and neutrality. In mixed-effects analyses, performance group remained associated with happy, angry, fearful, surprise, calm, and neutral speech-emotion proportions after false discovery rate correction. The largest excellent vs poor differences were observed for neutral (+4.24 percentage points), happy (+4.11), and surprise (−3.17) speech affect. TER showed a directionally similar but nonsignificant omnibus association ($F_{2,161}$=2.57; P=.08); mixed-effects analyses identified associations with happy and sad text-emotion proportions after correction. During extreme deterioration and ACLS management, fear and surprise together accounted for 64.1% and 64.6% of classified speech emotions, respectively, among poor performances. During ACLS management, neutral and calm emotions accounted for 49.1% and 37.7%, respectively, among excellent performances.

Conclusions: Automated speech- and text-based emotion recognition captured phase-dependent affective communication patterns during critical incident simulations. Speech-derived features were more consistently associated with expert-rated performance, whereas text-derived findings were more exploratory and emotion-specific. These markers may complement simulation debriefing and formative feedback, but validation using multirater outcomes, preregistered adjusted models, and larger multicenter datasets is needed before they inform competency assessment.

XSL•FO
RenderX

(JMIR Med Educ 2026;12:e96628) doi:10.2196/96628

KEYWORDS

simulation-based medical education; critical incident simulation; affective communication; speech-based emotion recognition; text-based emotion recognition; anesthesiology residents; formative feedback

Introduction

Affective Communication in Critical Care

Effective communication is central to safe care in high acuity clinical environments, such as in the operating room (OR) and intensive care unit (ICU). During critical events, clinicians must rapidly interpret evolving information, issue clear instructions, coordinate team responses, and maintain situational awareness under pressure. These exchanges are not purely informational. They also convey affective signals, including urgency, confidence, calmness, hesitation, and distress, which may influence team coordination, decision-making, and clinical execution [1,2]. Prior work has shown that both communication quality [3] and emotional processes [4] are relevant to performance in acute care settings.

In anesthesiology, these issues are especially complex because anesthesiologists operate under sustained cognitive load. Stress and arousal are expected components of medical emergencies, but when poorly regulated, they may impair attention, communication, and decision-making [5-7]. This is particularly challenging in perioperative crises, which often elicit strong emotional responses and a need for structured debriefing or recovery time [6].

Emotion and communication are intertwined rather than interchangeable. Well-regulated affect can facilitate clarity, closed-loop confirmation, and team coordination; dysregulated affect may coincide with interruptions, hesitations, or missed confirmations [7-9]. The key question is therefore not whether emotion is present, but how affective patterns unfold during critical situations and whether they align with effective communication behaviors and clinical performance.

Simulation-Based Assessment and the Need for Objective Communication Markers

Simulation-based crisis incident management training has been extensively used by anesthesiologists for over half a century [10,11], as it is effective for skill acquisition [12] and knowledge retention [13]. Critical incident simulation training is commonly used in high acuity training [14-16] and can also provide an opportunity to improve patient safety by exploring providers' responses to stressful situations [17,18]. Simulation-based assessment (SBA) is now a widely accepted method for evaluating clinical competency and is incorporated into board certification examinations by professional societies, including the Israel Society of Anesthesiologists [18]. However, current SBA approaches emphasize observable behavior scored by expert evaluators, which can be vulnerable to bias [19,20], while the emotional dimensions of communication remain largely unexamined. This creates a need for complementary, objective measures that can enrich traditional assessments without replacing expert evaluation.

Speech- and Text-Based Emotion Recognition

Recent advances in natural language processing (NLP) and speech analysis have created new opportunities to quantify emotional content in spoken clinical communication. Text-based emotion recognition (TER) [21,22] can infer emotional meaning from transcribed language, whereas speech-based emotion recognition (SER) [23,24] captures paralinguistic features such as tone, intensity, and vocal affect. The integration of these techniques provides a multidimensional view of communication, extending analysis beyond spoken words to encompass tone, urgency, and affect [25,26]. At the same time, applying TER and SER in healthcare poses challenges due to the specialized vocabulary and context-dependent nature of clinical communication [27]. Recent work has also emphasized the broader movement toward technology-enhanced assessment of nontechnical skills in high acuity procedural environments. For example, digital approaches to assessing nontechnical skills in the operating room have been proposed to provide more objective, scalable, and timely feedback, although validation and interpretability remain major challenges [7]. Similarly, recent reviews of SER [25] and textual emotion detection [28] in healthcare highlight the promise of automated affective analysis while emphasizing the need for careful domain adaptation, transparent evaluation, and validation in clinically meaningful settings. These considerations are particularly relevant in simulation-based anesthesiology training, where communication is brief, protocol-driven, emotionally charged, and shaped by clinical context. Accordingly, SER and TER should be viewed as potential complementary markers of affective communication rather than direct measures of clinical competency.

The contribution of the present study is the integration of existing speech- and text-based emotion recognition methods within a phase-structured, high-fidelity anesthesiology simulation setting. Specifically, we aligned resident-specific affective communication signals with predefined crisis phases and expert-rated simulation performance. This design allowed us to examine whether automatically derived emotion markers provide complementary information about affective communication during simulation-based training.

Study Aim and Hypotheses

In this study, we conceptualized affective communication as a behavioral signal that may reflect several overlapping processes relevant to crisis management, including cognitive load, stress regulation, communication clarity, situational awareness, and individual communication style. These processes are related to, but not equivalent to, clinical competency. Consistent with this conceptual framing, emotion recognition outputs were treated as exploratory markers of affective communication rather than direct measures of clinical competency. Specifically, our aim was to automatically extract emotion signals from speech and

XSL•FO
RenderX

transcribe text associated with expert-rated performance during simulation-based critical incident training among anesthesiology residents. Using audio recordings from simulated ICU scenarios, we analyzed SER and text-based emotion recognition outputs across predefined phases of each simulation. We hypothesized that higher-performing residents would display a greater level of regulated affective patterns, reflected in calmer or more positive emotional signals, whereas lower-performing residents would display a greater level of negative and high arousal emotional patterns. We also examined whether these emotional patterns differed across simulation phases, particularly during the most demanding portions of the scenario.

Methods

Study Design

This exploratory observational study examined whether automatically extracted emotion signals from residents' speech and transcribed language were associated with expert-rated performance during simulation-based critical incident training in anesthesiology.

Setting and Participants

Anesthesiology residency in Israel, as in other countries [29], includes staged qualifying examinations throughout training. In the first stage, residents must pass a national written board exam before advancing to the senior phase of residency. To become an attending anesthesiologist, residents must demonstrate knowledge and competency through oral board examinations and simulated scenarios. The Israel Society of Anesthesiologists organizes a biannual, 2-day workshop to prepare senior anesthesiology residents for their final board examinations. This workshop includes hands-on practice with high-fidelity simulations followed by expert feedback from board-certified anesthesiologists.

For this study, we analyzed all the residents who participated in the workshops conducted in 2022 and 2023. During enrollment, residents completed a demographic questionnaire that included their level of clinical experience, measured in years of residency, and the number of prior SBAs they had completed.

Ethical Considerations

The Rambam Health Care Campus Institutional Review Board approved this study (0482‐20-RMB). All participants provided written informed consent before participation.

Simulation Procedures and Performance Ratings

An experienced anesthesiologist and a medical simulation specialist developed 7 clinical simulation scenarios in accordance with advanced cardiovascular life support (ACLS) and advanced trauma life support (ATLS) principles: (1) severe anaphylaxis reaction, (2) postoperative severe bradycardia, (3) postoperative opioid overdose, (4) postoperative hypoglycemia, (5) postoperative residual paralysis, (6) postoperative supraventricular tachycardia, and (7) patient with head trauma.

Each resident was randomly assigned to 2 simulations. During each simulation, 2 members of the research team played the roles of a nurse and a medical intern, respectively. A board-certified anesthesiologist evaluated the resident's clinical performance using a task-specific checklist. The simulation setup included a Laerdal full-body manikin to create a high-fidelity, realistic environment (see Figure 1). The simulation area was recorded from multiple angles, and to ensure clear audio recordings, all participants (residents and team members) were equipped with wireless lavalier microphones. The data was recorded and synchronized using StreamPix digital video recording software from NorPix Inc.

To better capture participants' expressed emotions, we divided the simulation timeline into 4 chronologically distinct phases:

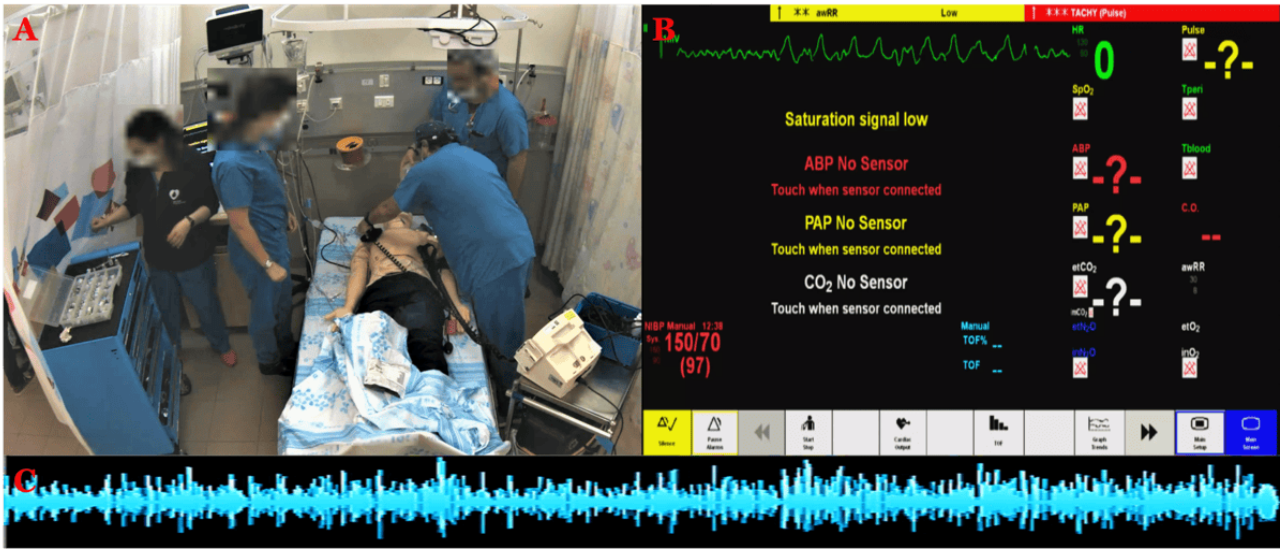
1. Initial condition: Participants understand the patient's status by examining the patient and communicating with the nurse.
2. Extreme deterioration: The patient's condition has deteriorated significantly (eg, dropping blood pressure, rising heart rate). Participants must manage this crisis and respond appropriately.
3. ACLS management: Participants must execute procedures quickly and accurately. Additionally, participants must demonstrate confidence and familiarity with the protocols. In the case of simulation 7 (patient with a head trauma), residents performed and managed ATLS.
4. Recovery: Participants update the team on the patient's status and coordinate postcrisis care.

This phase-based structure enabled us to examine whether emotional expression varied across different levels of clinical demand within the scenario. Phases were controlled by the simulator operator, each bounded by start and end timestamps and observable clinical cues. These included changes displayed on the patient monitor, such as alterations in vital signs and alarm activations, and, in some scenarios, physical indicators on the manikin itself (eg, breathing sounds, airway obstruction, chest movement, pupil size, etc).

At the end of each simulation, the anesthesiologist evaluator provided an overall performance rating on a Likert scale of 1 to 5, with 1 indicating poor performance and 5 indicating excellent performance. Based on these scores, we divided our sample into the following groups: poor performance (evaluation scores ≤2), intermediate performance (evaluation scores 3‐4), and excellent performance (evaluation scores=5).

XSL•FO
RenderX

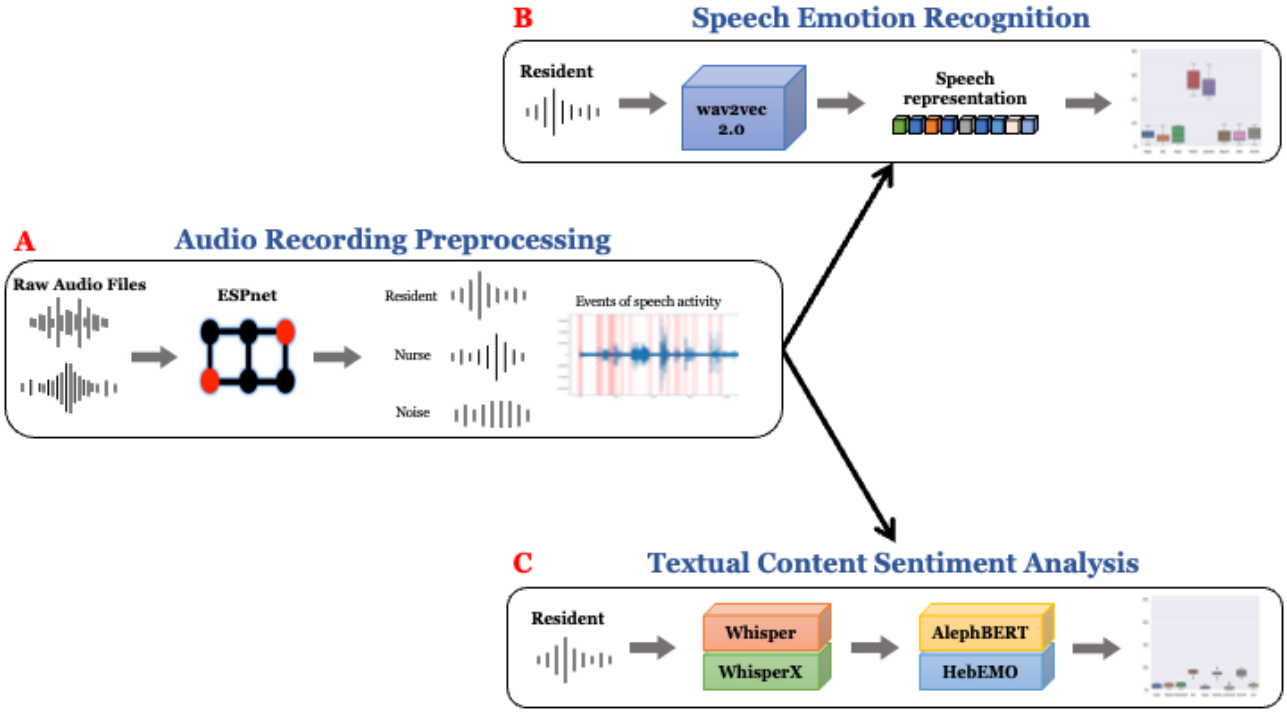
Figure 1. Simulation environment setup. (A) General overview of the simulation area. The faces have been manually blurred; (B) patient monitor display; (C) illustration of the lavalier microphone audio levels.



Audio Collection and Preprocessing

To minimize the effect of mixed audio recordings, participants (resident, nurse, and intern) were recorded on separate audio channels. However, participants' speech overlap and background noise (eg, patient monitor alarms) disrupted speech coherence, prompting us to integrate a speech-cleaning framework called end-to-end neural diarization with speaker subspace (EEND-SS) [30]. This framework performs several speech processing procedures (speaker diarization, speech separation, and speaker counting) and produces 2 types of audio files from each raw audio channel, (1) single-speaker files for each participant and (2) background noise files devoid of any traces of speech. These procedures eliminate disturbances that could have affected speech analysis. Additionally, the EEND-SS framework generates segments of "silence" areas in all processed audio files; silence thresholds are fine-tuned to optimize the results (see Figure 2A).

Figure 2. Illustration of the text-based emotion recognition (TER) and speech-based emotion recognition (SER) pipeline. (A) Preprocessing of the raw audio channels via ESPnet framework to extract intervals of the resident verbal expressions; (B) based on the extracted resident speech intervals, we executed SER for every sentence; (C) based on the extracted resident speech intervals, we transcribed each sentence and executed TER.



Speech-Based Emotion Recognition

Each EEND-SS single-speaker file is labeled according to the audio source (ie, anesthesia resident, nurse, and medical intern). The following stages apply only to the anesthesia resident's single-speaker audio file. First, feature extraction was performed using the SpeechBrain toolkit [31]. Specifically, we applied the wav2vec 2.0 model [32], which was pretrained via self

XSL•FO
RenderX

supervision. Then, we utilized Pepino et al's [33] model, which generates emotion category probabilities for 8 classes: happy, sad, angry, fearful, surprised, disgusted, calm, and neutral (see Figure 2B). During inference, the predicted emotion category for each segment was defined as the class with the highest probability. Segment-level predictions were then aggregated within each simulation, performance group, and simulation phase to estimate the relative frequency of each emotion category. These aggregated proportions were used in the statistical analyses.

Text-Based Emotion Recognition

The processed anesthesia resident audio file was transcribed to Hebrew using a state-of-the-art, large-scale, weakly supervised speech recognition model, Whisper [34]. In addition, since Whisper only provides timestamps for each speech utterance, we used the work of Bain et al [35] to generate a word-level timestamped transcription. Afterward, we performed word- and sentence-level TER using 2 language models trained specifically on Hebrew text: AlephBERT [36] and HebEMO [37] (see Figure 2C). As with the speech-based model, the dominant text-based emotion category was defined as the class with the highest model score for a given utterance. To improve compatibility with clinical terminology, we applied the medical domain adaptation procedure described in our previous work [38], which involved adapting the text-processing pipeline to medical language. We did not fine-tune the emotion recognition models on the current performance labels, thereby avoiding leakage between performance ratings and emotion classification outputs.

Implementation Details

All analyses were conducted on a computing cluster with 2 NVIDIA RTX A6000 48 GB graphics processing units (GPUs). The software environment consisted of Ubuntu 20.04 LTS, Python (version 3.9), and PyTorch (version 2.00). The EEND-SS framework used 6 transformer encoders and Conv-TasNet [39], which incorporates 8 temporal convolutional network blocks. The wav2vec 2.0 model had a feature encoder with 7 convolutional neural network layers and a transformer with 12 encoder layers. Both AlephBERT and HebEMO were used with their default parameters.

Statistical Analysis

To evaluate whether the distribution of emotions differed across performance groups, we first used one-way ANOVA as an exploratory omnibus test. All statistical tests were 2-sided. Statistical significance was defined as $P<.05$, and exact P values are reported unless $P<.001$. Because residents could contribute more than 1 simulation and simulations differed by scenario, we then conducted exploratory mixed-effects sensitivity analyses. Emotion proportions were modeled separately for each emotion category. The performance group was included as the primary fixed effect, with random intercepts for resident and simulation scenario. Residency year and sex were included as covariates when available. Global performance-group effects were evaluated using likelihood ratio tests comparing models with and without performance group. False discovery rate (FDR) correction was applied across emotion-level tests within each modality. These sensitivity analyses were intended to assess whether the observed associations persisted after accounting for repeated observations and scenario-level clustering.

We also performed phase-stratified analyses to examine whether the relationship between emotional expression and performance varied across phases of the simulation scenario. All statistical analyses were executed using SciPy (version 1.16.1) and statsmodels (version 0.14.6).

Results

Sample Characteristics

We recruited 90 senior anesthesiology residents from 17 different hospitals, representing a diverse range of clinical backgrounds (see Table 1). On average, each resident participated in 2 different SBAs, yielding 171 recordings overall. Seven simulation recordings were excluded due to technical issues, including incomplete audio or video capture and severe audio artifacts, resulting in the final analytic sample of 164 simulations (see Table 2).

Based on expert assessors' ratings, 65 out of 164 simulations (39.6%) were classified as poor performance (scores ≤2), 72 out of 164 (43.9%) as intermediate performance (scores 3‐4), and 27 out of 164 (16.5%) as excellent performance (score=5). The final dataset included recordings from all 7 simulation scenarios, with the largest proportions contributed by postoperative residual paralysis (42 out of 164 simulations, 25.6%) and postoperative severe hypoglycemia (36 out of 164 simulations, 22.0%) scenarios. It is worth noting that although participants communicated in Hebrew (with some English terms), the EEND-SS framework successfully processed our participants' speech.

XSL•FO
RenderX

Table . Demographic questionnaire summary.

Characteristic	Residents (N=90), n (%)	Years of residency, mean (SD)
Sex		
Male	58 (64)	5.48 (1.32)
Female	32 (36)	5.06 (3.51)
Residency year		
4	26 (29)	—[a]
5	32 (36)	—
6	19 (21)	—
≥7	13 (14)	—

[a]Not applicable.

Table . Simulation dataset summary.

Simulation type	Simulations (N=164), n (%)	Performance score, n				
		1	2	3	4	5
Patient with a severe anaphylaxis reaction	24 (14.6)	2	5	7	6	4
Postoperative patient with severe bradycardia	26 (15.9)	5	4	4	10	3
Postoperative patient with opioid overdose	12 (7.3)	1	2	2	4	3
Postoperative patient with severe hypoglycemia	36 (22.0)	7	7	6	8	8
Postoperative patient with residual paralysis	42 (25.6)	7	12	9	7	7
Postoperative patient with severe supraventricular tachycardia	14 (8.5)	5	2	4	2	1
Patient with head trauma	10 (6.1)	4	2	2	1	1

Affective Communication Across Performance Groups

Across performance groups, SER emotion proportions differed significantly ($F_{2,161}$=4.61; P=.02; Figure 3). Poor performances (scores ≤2) showed higher levels of fear and surprise, whereas excellent performances (score=5) showed higher levels of calm and happiness. TER emotion proportion showed a directionally similar but nonsignificant omnibus effect ($F_{2,161}$=2.57; P=.08; Figure 4), with poorer performance skewing toward more negative emotions and excellent performance toward more positive and neutral emotions. Cross-modal correspondence is evident: groups with higher negative emotions in SER also tended to have lower positive emotions in TER.

Across both modalities, the poor and excellent performance groups differed significantly for nearly all emotions.

XSL•FO
RenderX

Figure 3. Speech-based emotion recognition (SER): emotion classification of participants' speech and transcription according to their performance score group. The comparison evaluation is conducted between the poor and excellent groups.

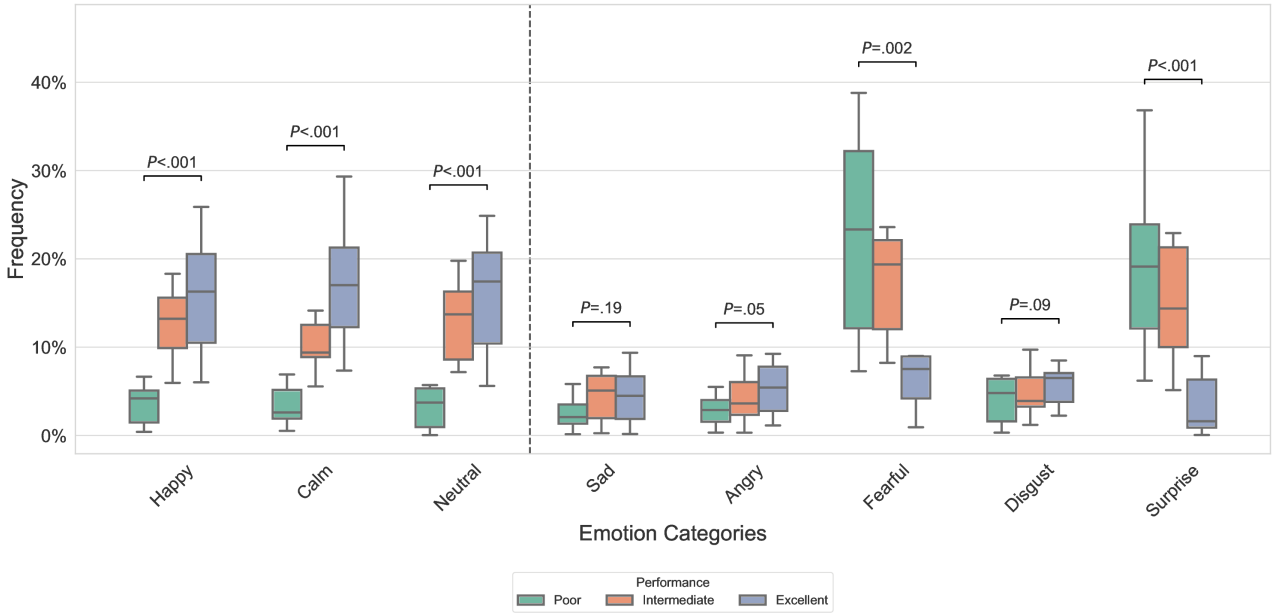
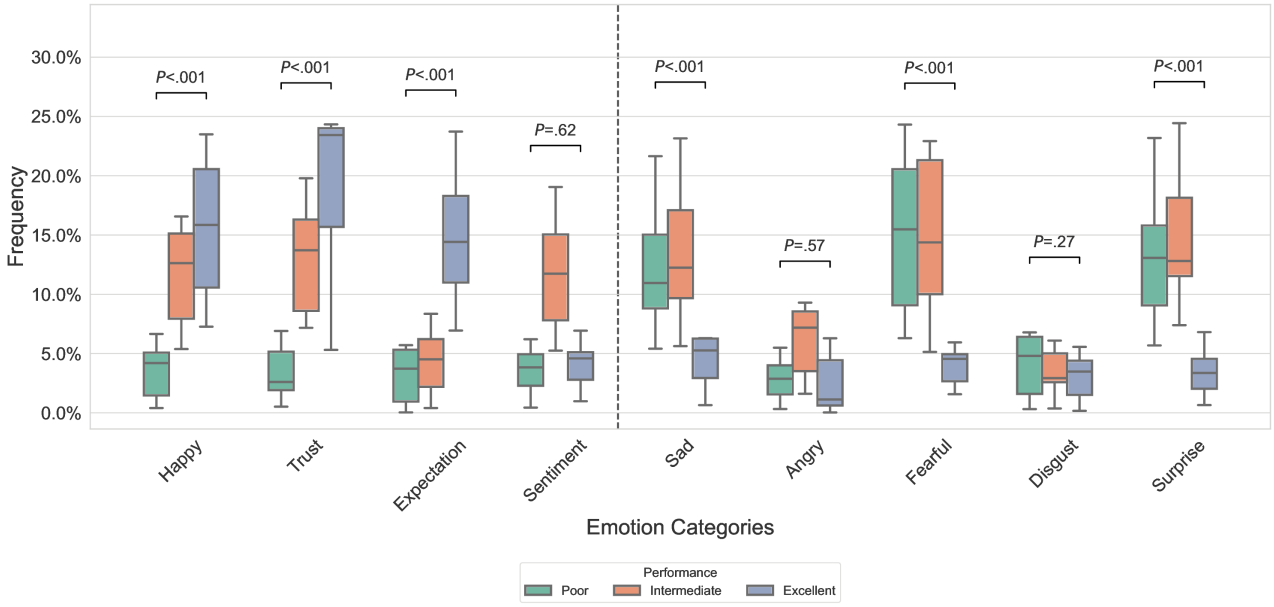


Figure 4. Text-based emotion recognition (TER): emotion classification of participants' speech and transcription according to their performance group. The comparison evaluation is conducted between the poor and excellent groups.



Sensitivity Analysis Using Mixed-Effects Models

Mixed-effects model results for speech- and text-based emotion proportions across performance groups are presented in Table 3. For SER, performance group remained significantly associated with happy, angry, fearful, surprise, calm, and neutral speech-emotion proportions after FDR correction. Sad and disgust were not significantly associated with performance group after correction. The largest excellent vs poor differences were observed for neutral speech affect, which was 4.24 percentage points higher in excellent performances; happy speech affect, which was 4.11 percentage points higher in excellent performances; and surprise speech affect, which was 3.17 percentage points lower in excellent performances.

XSL•FO
RenderX

Table . Mixed-effects sensitivity analysis of speech- and text-based emotion proportions across performance groups. Models were estimated separately for each emotion category, and emotion proportions were logit-transformed and modeled as the outcome, with performance group as the primary fixed effect. All models included random intercepts for resident and scenario and were adjusted for residency year and sex. Global performance group effects were evaluated using likelihood ratio tests comparing models with and without performance group. False discovery rate correction was applied across emotion-level tests within each modality.

Modality and emotion	LR[a] $\chi^2(2)$	*P* value	FDR[b]–adjusted *P*	Excellent-poor difference
SER[c]				
Happy	22.03	<.001	<.001	4.11
Sad	2.19	.34	.38	−0.83
Angry	11.32	.003	.005	−1.66
Fearful	19.76	<.001	<.001	−2.31
Surprise	21.88	<.001	<.001	−3.17
Disgust	0.67	.72	.72	0.00
Calm	15.53	<.001	<.001	2.21
Neutral	25.71	<.001	<.001	4.24
TER[d]				
Happy	23.96	<.001	<.001	8.66
Sad	14.39	<.001	.003	−5.33
Angry	1.82	.40	.52	−0.89
Fearful	6.52	.04	.09	−2.91
Surprise	7.71	.02	.06	−4.13
Disgust	1.15	.56	.63	−0.15
Trust	3.67	.16	.24	2.06
Expectation	0.88	.64	.64	1.15
Sentiment	4.27	.12	.21	1.58

[a]LR: likelihood ratio.

[b]FDR: false discovery rate.

[c]SER: speech-based emotion recognition.

[d]TER: text-based emotion recognition.

For text-based emotion recognition, performance group remained significantly associated with happy and sad text-emotion proportions after FDR correction. Fearful and surprise text-emotion proportions showed nominal associations with performance group but did not remain significant after correction. Angry, disgust, trust, expectation, and sentiment were not significantly associated with performance group. The largest excellent vs poor differences were observed for happy text affect, which was 8.66 percentage points higher in excellent performances; sad text affect, which was 5.33 percentage points lower in excellent performances; and surprise text affect, which was 4.13 percentage points lower in excellent performances.

Together, these sensitivity analyses support the overall direction of the primary findings: higher-rated simulations showed higher positive, calm, or neutral affective patterns, whereas lower-rated simulations showed higher fearful or surprise-related affective patterns. However, the text-based findings were weaker and more emotion-specific than the speech-based findings, supporting the interpretation that speech- and text-based emotion recognition provide complementary but differently sensitive measures of affective communication.

Affective Communication Across Simulation Phases

For this experiment, we compared the alignment (ie, agreement) between TER and SER across the simulation phase. For each phase, we allocated the 2 most prevalent emotion categories; if the highest-scoring emotion exceeded 50%, it was considered “dominant.” Phase-specific analyses revealed cross-modal concordance: negative SER and TER categories peaked during extreme deterioration and ACLS management, whereas calm and positive emotions predominated during the initial condition and recovery phases. Table 4 summarizes the top 2 categories per modality and phase; differences are most pronounced between the excellent and poor performance groups during high stress phases.

XSL•FO
RenderX

Table . Participants' text-based emotion recognition (TER) and speech-based emotion recognition (SER) results with respect to the simulation phases. The percentage indicates the prevalence of emotions.

	Initial condition		Extreme deterioration		ACLS[a] management		Recovery	
	Top 2 vocal emotions (%)	Top 2 textual emotions (%)	Top 2 vocal emotions (%)	Top 2 textual emotions (%)	Top 2 vocal emotions (%)	Top 2 textual emotions (%)	Top 2 vocal emotions (%)	Top 2 textual emotions (%)
Poor performance	• *Fearful (50.03)*[b] • Angry (31.78)	• Fearful (36.70) • Sad (30.22)	• *Surprise (64.12)* • Fearful (18.95)	• *Surprise (59.81)* • Fearful (14.08)	• Angry (11.89) • *Fearful (64.57)*	• *Surprise (55.78)* • Fearful (23.08)	• Sad (30.73) • Angry (24.67)	• *Sad (50.89)* • Surprise (26.84)
Intermediate performance	• Calm (40.32) • Neutral (16.70)	• *Happy (52.13)* • Trust (33.41)	• *Calm (53.25)* • Sad (23.97)	• *Sad (50.22)* • Sentiment (20.69)	• Calm (44.19) • Neutral (35.67)	• Surprise (34.89) • Expectation (32.22)	• Happy (37.21) • Calm (34.18)	• *Happy (53.29)* • Trust (30.62)
Excellent performance	• *Calm (57.10)* • Happy (27.85)	• *Expectation (66.78)* • Sentiment (11.67)	• *Surprise (53.06)* • Calm (22.65)	• *Surprise (63.88)* • Expectation (9.23)	• Neutral (49.10) • Calm (37.66)	• *Expectation (50.77)* • Happy (31.05)	• *Neutral (53.28)* • Happy (26.85)	• *Happy (64.92)* • Fearful (14.78)

[a]ACLS: advanced cardiovascular life support.

[b]Cases in which one of the emotions has a prevalence of over 50% (ie, "dominant") are indicated in italics.

Across phases, 3 broad patterns emerged. First, differences between performance groups were most pronounced during the clinically demanding phases of extreme deterioration and ACLS management. Second, poor performance simulations showed a greater concentration of high-arousal negative emotion labels, particularly fear and surprise, during these phases. Third, excellent performance simulations more often showed calm, neutral, or positive affective labels during phases requiring coordinated action and recovery. Thus, the phase-based results suggest that affective communication differences were not uniform across the simulation but were most evident during moments of acute clinical demand.

Discussion

Principal Findings

In this prospective, simulation-based study of senior anesthesia residents, we observed systematic differences in model-derived affective communication patterns across performance groups. In the primary omnibus analysis, SER profiles differed across performance groups ($F_{2,161}$=4.61; P=.02): residents in the poor performance group (scores ≤2) expressed predominantly fear and surprise, whereas those in the excellent performance group (score=5) showed higher proportions of calm, happy, and neutral speech affect. TER profiles showed a directionally similar but nonsignificant omnibus effect ($F_{2,161}$=2.57; P=.08). Exploratory mixed-effects sensitivity analyses supported the overall direction of the speech-based findings. Performance group remained associated with happy, angry, fearful, surprise, calm, and neutral speech-emotion proportions after FDR correction. The largest excellent vs poor differences were observed for neutral speech affect (+4.24 percentage points), happy speech affect (+4.11 percentage points), and surprise speech affect (−3.17 percentage points). For text-based emotion recognition, performance group remained associated with happy and sad text-emotion proportions after correction, whereas fearful and surprise text-emotion proportions showed nominal associations that did not remain significant after correction. These findings suggest that speech-derived affective markers were more consistently associated with expert-rated performance than text-derived markers, although both modalities showed patterns consistent with the primary results.

Interpretation of Affective Communication Patterns

The present study does not establish that emotion recognition outputs improve the accuracy, reliability, or predictive validity of simulation-based competency assessment. The findings show associations between model-derived affective communication patterns and expert-rated simulation performance. Therefore, these markers should be interpreted as exploratory and complementary signals that may inform formative feedback and future research, not as standalone assessment tools or substitutes for expert evaluation.

Phase-stratified summaries indicated that differences were most pronounced during extreme deterioration and ACLS management, when residents were required to recognize deterioration, coordinate the team, and execute time-sensitive actions. In these phases, higher-rated residents maintained calmer/neutral speech affect and more positive/neutral text tone. These associations suggest that affective communication patterns—how emotion is expressed and regulated—vary across performance groups and across the simulation phase. We interpret these findings as associations rather than causal effects. Elevated expression of emotions such as fear and surprise may co-occur with uncertainty or cognitive overload, whereas calm and neutral affect may co-occur with clearer commands, closed-loop confirmation, and anticipatory planning.

XSL•FO
RenderX

Modality-Specific Findings: Speech vs Text

The discrepancy between SER and TER findings warrants careful interpretation. In the primary omnibus analysis, SER showed a statistically significant association with performance group, whereas TER showed a directionally similar but nonsignificant trend. This suggests that speech-derived affective information may have been more robustly captured than text-derived affective information in the present dataset. One explanation is that SER captures paralinguistic features, including pitch, intensity, tempo, pauses, hesitation, and vocal arousal, which may be especially informative during high-stress simulation. TER, in contrast, depends on a multistep pipeline; thus, errors or biases introduced at any stage may attenuate the estimates.

Several modality-specific limitations may have contributed to the weaker primary TER finding. Residents' speech during crisis management was often brief, imperative, protocol-driven, and interspersed with Hebrew-English clinical terminology. Such utterances may contain limited lexical emotional information, even when vocal delivery conveys strong affective cues. Automatic transcription errors, word-level timestamping errors, code-switching, medical abbreviations, and domain-specific terminology may also introduce noise. In addition, TER models may misclassify clinically appropriate urgency, direct commands, or technical language as negative affect, whereas SER models may misinterpret raised volume or rapid speech as distress even when these features reflect appropriate leadership. Therefore, neither modality should be treated as a direct measure of psychological state or competency.

The mixed-effects sensitivity analyses suggested that some TER categories, particularly happy and sad, varied across performance groups after adjustment for repeated simulations and scenario-level clustering. However, these analyses were exploratory and emotion-specific; therefore, they should be interpreted as hypothesis-generating. Overall, the findings suggest that SER and TER provide complementary but imperfect views of affective communication, with different sources of noise, bias, and context dependence.

Educational Implications

Clinically, the phase specificity matters. The clearest separation between excellent and poor performances emerged precisely when the team, specifically the resident, transitioned from recognizing deterioration to taking decisive action. That observation generates practical hypotheses for training: debriefings might explicitly review affective communication during these windows; phase-targeted coaching could rehearse calm and neutral delivery of closed-loop commands; and formative feedback could pair technical checklists with brief, objective summaries of affective patterns.

From an educational perspective, the most appropriate near-term use of these metrics is formative rather than summative. SER and TER outputs should not be used to make independent pass-fail decisions, rank residents, or replace expert faculty assessment. Instead, they may be used as debriefing aids that help educators identify specific moments in a simulation where affective communication changed. For example, an educator could review a timeline showing increases in high-arousal vocal affect, pauses, or shifts toward calmer communication and use these moments to guide reflective discussion about the clinical context, communication clarity, use of closed-loop communication, and whether the emotional tone supported or interfered with team coordination.

Limitations and Future Directions

Several important limitations should be considered when interpreting these findings. First, this was an exploratory observational study; therefore, the results cannot determine whether the observed affective patterns reflect competency-relevant communication, cognitive load, stress response, individual communication style, or other latent constructs. Future studies should combine automated emotion recognition with expert-coded communication behaviors, physiological stress measures, and multi-rater performance assessments to clarify these mechanisms. Second, the primary analyses were univariate and did not fully adjust for potential confounders, including sex, postgraduate year, scenario, site, and language-related factors. Although we added exploratory mixed-effects sensitivity analyses, these models were not prespecified and should be interpreted as robustness checks rather than definitive confirmatory analyses. Future studies should use preregistered analytic plans, fully powered hierarchical models, and richer covariate structures to account for repeated simulations, scenario-level clustering, hospital affiliation, and other potential confounders.

Third, performance ratings were primarily based on single-rater expert assessments. Therefore, these ratings should be viewed as expert-rated indicators of simulation performance rather than definitive reference standards for clinical competency. Future work should incorporate multiple independent raters, standardized rating procedures, rater training, and reliability analyses to strengthen the validity of the performance outcome.

Fourth, SER may be affected by background alarms, overlapping speech, microphone placement, speech separation errors, speaker-specific vocal characteristics, and cultural or language-specific differences in vocal affect. TER may be affected by transcription errors, segmentation errors, code-switching, short utterance length, clinical abbreviations, and limited lexical emotional content in protocol-driven commands. These sources of noise may differentially affect SER and TER, partly explaining why the primary SER findings were stronger than the primary TER findings.

Additionally, the models operated on Hebrew speech with Hebrew-English code-switching, which may have amplified speech-to-text and emotion classification errors. While we accounted for domain adaptation in medical discourse, a known challenge given specialized vocabulary and context-dependent meaning [27], we did not specifically adapt the models to the emotional and communicative context of anesthesiology crisis simulation.

Finally, time-aligned analyses could further examine whether short surges in high arousal negative affect precede delayed or missed actions, and whether calm or neutral affect during extreme deterioration and ACLS management is associated with

XSL•FO
RenderX

more timely execution after adjustment for resident-, site-, and scenario-level factors.

Conclusions

In conclusion, this exploratory observational study suggests that automated speech- and text-based emotion recognition can capture phase-specific affective communication patterns during anesthesiology critical incident simulation training. Speech-derived emotion markers, and to a lesser extent, text-derived markers, were associated with expert-rated performance groups. These findings should not be interpreted as evidence that emotion recognition can independently assess competency or improve assessment accuracy. Rather, they support the feasibility of using affective communication markers as complementary tools for simulation debriefing, formative feedback, and future research on communication under crisis conditions. Further validation using multirater outcomes, adjusted statistical models, and larger multicenter datasets is needed before such measures can inform competency assessment.

Acknowledgments

The authors thank the Technion Autonomous Systems Program for institutional support of this work.

The authors used ChatGPT (OpenAI) to assist with language editing and improvement of manuscript readability. All scientific content, analytic decisions, interpretations, and final manuscript revisions were reviewed and approved by the authors, who take full responsibility for the content of the manuscript.

Funding

No specific funding was received for this study.

Data Availability

The data generated and analyzed during this study are not publicly available because they include identifiable or potentially sensitive simulation audio and video recordings. Deidentified derived data may be made available from the corresponding author on reasonable request, subject to institutional review board approval and data-sharing restrictions. To improve reproducibility, the analysis scripts used to process model outputs, compute emotion category proportions, perform statistical analysis, and generate figures are available from the corresponding author on reasonable request.

Authors' Contributions

Conceptualization: SG, IB, AR, SL
Data curation: SG
Formal analysis: SG
Funding acquisition: SL, AR
Investigation: FM
Methodology: SG, IB, SL
Project administration: AR, FM
Resources: SL, AR
Software: SG
Supervision: AR, SL
Validation: SG, SL
Visualization: SG
Writing – original draft: SG
Writing – review & editing: SG, IB, FM, AR, SL
All authors reviewed and approved the final manuscript.

Conflicts of Interest

AR served as a consultant and received research support from Medtronic. He gave seminars for MSD (Merck & Co) and for Medtechnica (Massimo). All other authors have no conflicts of interest.

References

1. Reader TW, Flin R, Mearns K, Cuthbertson BH. Developing a team performance framework for the intensive care unit. Crit Care Med 2009 May;37(5):1787-1793. [doi: 10.1097/CCM.0b013e31819f0451] [Medline: 19325474]
2. Marcum JA. The role of emotions in clinical reasoning and decision making. J Med Philos 2013 Oct;38(5):501-519. [doi: 10.1093/jmp/jht040] [Medline: 23975905]
3. Eisenberg EM, Murphy AG, Sutcliffe K, et al. Communication in emergency medicine: implications for patient safety. Commun Monogr 2005;72(4):390-413. [doi: 10.1080/03637750500322602]

XSL•FO
RenderX

4. Isbell LM, Boudreaux ED, Chimowitz H, Liu G, Cyr E, Kimball E. What do emergency department physicians and nurses feel? a qualitative study of emotions, triggers, regulation strategies, and effects on patient care. BMJ Qual Saf 2020 Oct;29(10):1-2. [doi: 10.1136/bmjqs-2019-010179] [Medline: 31941799]
5. Gurman GM, Klein M, Weksler N. Professional stress in anesthesiology: a review. J Clin Monit Comput 2012 Aug;26(4):329-335. [doi: 10.1007/s10877-011-9328-7] [Medline: 22180163]
6. Gazoni FM, Amato PE, Malik ZM, Durieux ME. The impact of perioperative catastrophes on anesthesiologists: results of a national survey. Anesth Analg 2012 Mar;114(3):596-603. [doi: 10.1213/ANE.0b013e318227524e] [Medline: 21737706]
7. Howie EE, Ambler O, Gunn EGM, et al. Surgical sabermetrics: a scoping review of technology-enhanced assessment of nontechnical skills in the operating room. Ann Surg 2024 Jun 1;279(6):973-984. [doi: 10.1097/SLA.0000000000006211] [Medline: 38258573]
8. Hu YY, Arriaga AF, Peyre SE, Corso KA, Roth EM, Greenberg CC. Deconstructing intraoperative communication failures. J Surg Res 2012 Sep;177(1):37-42. [doi: 10.1016/j.jss.2012.04.029] [Medline: 22591922]
9. Hall A, Kawai K, Graber K, et al. Acoustic analysis of surgeons' voices to assess change in the stress response during surgical in situ simulation. BMJ Simul Technol Enhanc Learn 2021 Apr 13;7(6):471-477. [doi: 10.1136/bmjstel-2020-000727] [Medline: 35520977]
10. Ross AJ, Kodate N, Anderson JE, Thomas L, Jaye P. Review of simulation studies in anaesthesia journals, 2001-2010: mapping and content analysis. Br J Anaesth 2012 Jul;109(1):99-109. [doi: 10.1093/bja/aes184] [Medline: 22696559]
11. Reynolds T, Kong ML. Shifting the learning curve. BMJ 2010 Dec 2;341:c6260. [doi: 10.1136/bmj.c6260] [Medline: 21127119]
12. Grantcharov TP, Kristiansen VB, Bendix J, Bardram L, Rosenberg J, Funch-Jensen P. Randomized clinical trial of virtual reality simulation for laparoscopic skills training. Br J Surg 2004 Feb;91(2):146-150. [doi: 10.1002/bjs.4407] [Medline: 14760660]
13. Cecilio-Fernandes D, Brandão CFS, de Oliveira DLC, Fernandes G, Tio RA. Additional simulation training: does it affect students' knowledge acquisition and retention? BMJ Simul Technol Enhanc Learn 2018 Jun 22;5(3):140-143. [doi: 10.1136/bmjstel-2018-000312] [Medline: 35514946]
14. Weile J, Nebsbjerg MA, Ovesen SH, Paltved C, Ingeman ML. Simulation-based team training in time-critical clinical presentations in emergency medicine and critical care: a review of the literature. Adv Simul (Lond) 2021 Jan 20;6(1):3. [doi: 10.1186/s41077-021-00154-4] [Medline: 33472706]
15. Robertson JM, Dias RD, Yule S, Smink DS. Operating room team training with simulation: a systematic review. J Laparoendosc Adv Surg Tech A 2017 May;27(5):475-480. [doi: 10.1089/lap.2017.0043] [Medline: 28294695]
16. Bienstock J, Heuer A, Zhang Y. Simulation-based training and its use amongst practicing paramedics and emergency medical technicians: an evidence-based systematic review. Intl J Paramedicine 2023 Jan 9;1:12-28. [doi: 10.56068/VWHV8080]
17. McGaghie WC, Issenberg SB, Petrusa ER, Scalese RJ. A critical review of simulation-based medical education research: 2003-2009. Med Educ 2010 Jan;44(1):50-63. [doi: 10.1111/j.1365-2923.2009.03547.x] [Medline: 20078756]
18. Ziv A, Rubin O, Sidi A, Berkenstadt H. Credentialing and certifying with simulation. Anesthesiol Clin 2007 Jun;25(2):261-269. [doi: 10.1016/j.anclin.2007.03.002] [Medline: 17574189]
19. Seehusen DA, Kleinheksel AJ, Huang H, Harrison Z, Ledford CJW. The power of one word to paint a halo or a horn: demonstrating the halo effect in learner handover and subsequent evaluation. Acad Med 2023 Aug 1;98(8):929-933. [doi: 10.1097/ACM.0000000000005161] [Medline: 36724305]
20. Humphrey-Murto S, Shaw T, Touchie C, Pugh D, Cowley L, Wood TJ. Are raters influenced by prior information about a learner? a review of assimilation and contrast effects in assessment. Adv Health Sci Educ Theory Pract 2021 Aug;26(3):1133-1156. [doi: 10.1007/s10459-021-10032-3] [Medline: 33566199]
21. Adoma AF, Henry NM, Chen W. Comparative analyses of BERT, RoBERTa, DistilBERT, and XLNet for text-based emotion recognition. Presented at: 2020 17th International Computer Conference on Wavelet Active Media Technology and Information Processing (ICCWAMTIP); Dec 18-20, 2020 URL: https://ieeexplore.ieee.org/document/9317379 [accessed 2026-08-07] [doi: 10.1109/ICCWAMTIP51612.2020.9317379]
22. Acheampong FA, Wenyu C, Nunoo‐Mensah H. Text‐based emotion detection: advances, challenges, and opportunities. Eng Rep 2020 Jul;2(7):e12189. [doi: 10.1002/eng2.12189]
23. Khalil RA, Jones E, Babar MI, Jan T, Zafar MH, Alhussain T. Speech emotion recognition using deep learning techniques: a review. IEEE Access 2019;7:117327-117345. [doi: 10.1109/ACCESS.2019.2936124]
24. Wani TM, Gunawan TS, Qadri SAA, Kartiwi M, Ambikairajah E. A comprehensive review of speech emotion recognition systems. IEEE Access 2021;9:47795-47814. [doi: 10.1109/ACCESS.2021.3068045]
25. Latif S, Qadir J, Qayyum A, Usama M, Younis S. Speech technology for healthcare: opportunities, challenges, and state of the art. IEEE Rev Biomed Eng 2021;14:342-356. [doi: 10.1109/RBME.2020.3006860] [Medline: 32746367]
26. Birjali M, Kasri M, Beni-Hssane A. A comprehensive survey on sentiment analysis: approaches, challenges and trends. Knowl Based Syst 2021 Aug 17;226:107134. [doi: 10.1016/j.knosys.2021.107134]
27. Holderness E, Cawkwell P, Bolton K, Pustejovsky J, Hall MH. Distinguishing clinical sentiment: the importance of domain adaptation in psychiatric patient health records. In: Proceedings of the 2nd Clinical Natural Language Processing Workshop:

XSL•FO
RenderX

Association for Computational Linguistics; 2019:117-123 URL: https://aclanthology.org/W19-1915/ [accessed 2026-08-07] [doi: 10.18653/v1/W19-1915]
28. Ahmed T, Gopala Krishnan C. A comprehensive study on emotion recognition in healthcare by applying machine learning and deep learning techniques. Presented at: 2024 International Conference on Knowledge Engineering and Communication Systems (ICKECS); Apr 18-19, 2024 URL: https://ieeexplore.ieee.org/document/10617024 [accessed 2026-08-07] [doi: 10.1109/ICKECS61492.2024.10617024]
29. Yamamoto S, Tanaka P, Madsen MV, Macario A. Comparing anesthesiology residency training structure and requirements in seven different countries on three continents. Cureus 2017 Feb 26;9(2):e1060. [doi: 10.7759/cureus.1060] [Medline: 28367396]
30. Maiti S, Ueda Y, Watanabe S, et al. EEND-SS: joint end-to-end neural speaker diarization and speech separation for flexible number of speakers. Presented at: 2022 IEEE Spoken Language Technology Workshop (SLT); Jan 9-12, 2023 URL: https://ieeexplore.ieee.org/document/10022924 [accessed 2026-08-07] [doi: 10.1109/SLT54892.2023.10022924]
31. Ravanelli M, Parcollet T, Moumen A, et al. Open-source conversational AI with SpeechBrain 1.0. J Mach Learn Res 2024;25(333):1-11 [FREE Full text]
32. Baevski A, Zhou H, Mohamed A, Auli M. Wav2vec 2.0: a framework for self-supervised learning of speech representations. In: NIPS'20: Proceedings of the 34th International Conference on Neural Information Processing Systems: Neural Information Processing Systems Foundation; 2020:12449-12460 URL: https://dl.acm.org/doi/abs/10.5555/3495724.3496768 [accessed 2026-08-07]
33. Pepino L, Riera P, Ferrer L. Emotion recognition from speech using wav2vec 2.0 embeddings. 2021 Presented at: Interspeech 2021; Aug 30 to Sep 3, 2021; Brno, Czechia p. 3400-3404 URL: https://www.isca-archive.org/interspeech_2021/pepino21_interspeech.html [accessed 2026-08-07] [doi: 10.21437/Interspeech.2021-703]
34. Radford A, Kim JW, Xu T, Brockman G, McLeavey C, Sutskever I. Robust speech recognition via large-scale weak supervision. In: Proceedings of the 40th International Conference on Machine Learning: PLMR; 2023. URL: https://proceedings.mlr.press/v202/radford23a.html [accessed 2026-08-07]
35. Bain M, Huh J, Han T, Zisserman A. WhisperX: time-accurate speech transcription of long-form audio. 2023 Presented at: Interspeech 2023; Aug 20-24, 2023; Dublin, Ireland p. 4489-4493 URL: https://www.isca-archive.org/interspeech_2023/bain23_interspeech.html [accessed 2026-08-07] [doi: 10.21437/Interspeech.2023-78]
36. Seker A, Bandel E, Bareket D, Brusilovsky I, Greenfeld R, Tsarfaty R. AlephBERT: language model pre-training and evaluation from sub-word to sentence level. In: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers): Association for Computational Linguistics; 2022:46-56 URL: https://aclanthology.org/2022.acl-long [accessed 2026-08-07] [doi: 10.18653/v1/2022.acl-long.4]
37. Chriqui A, Yahav I. HeBERT and HebEMO: A Hebrew BERT model and a tool for polarity analysis and emotion recognition. INFORMS J Dat Sci 2022 Apr;1(1):81-95. [doi: 10.1287/ijds.2022.0016]
38. Gershov S, Braunold D, Spektor R, Ioscovich A, Raz A, Laufer S. Automating medical simulations. J Biomed Inform 2023 Aug;144:104446. [doi: 10.1016/j.jbi.2023.104446] [Medline: 37467836]
39. Luo Y, Mesgarani N. Conv-TasNet: surpassing ideal time-frequency magnitude masking for speech separation. IEEE/ACM Trans Audio Speech Lang Process 2019 Aug;27(8):1256-1266. [doi: 10.1109/TASLP.2019.2915167] [Medline: 31485462]

Abbreviations

ACLS: advanced cardiovascular life support
ATLS: advanced trauma life support
EEND-SS: end-to-end neural diarization with speaker subspace
FDR: false discovery rate
GPU: graphics processing unit
ICU: intensive care unit
NLP: natural language processing
OR: operating room
SBA: simulation-based assessment
SER: speech-based emotion recognition
TER: text-based emotion recognition

XSL•FO
RenderX

Edited by LT Car; submitted 30.Mar.2026; peer-reviewed by A Turky, D Winterton; revised version received 05.Jun.2026; accepted 22.Jul.2026; published 21.Aug.2026.

Please cite as:
Gershov S, Bentov I, Mahameed F, Raz A, Laufer S
Speech- and Text-Based Emotion Recognition in Anesthesiology Residents During Critical Incident Simulation Training: Exploratory Observational Study
JMIR Med Educ 2026;12:e96628
URL: https://mededu.jmir.org/2026/1/e96628
doi:10.2196/96628

© Sapir Gershov, Itay Bentov, Fadi Mahameed, Aeyal Raz, Shlomi Laufer. Originally published in JMIR Medical Education (https://mededu.jmir.org), 21.Aug.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Education, is properly cited. The complete bibliographic information, a link to the original publication on https://mededu.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Attitudes Toward Digital Planetary Health Before and After a Lecture Series on Environmental Awareness and Future Visions: Repeated Cross-Sectional Survey Study Regarding Twin Transformation

Julia Nitsche[1*], Dr rer medic; Theresa Sophie Busse[2*], Prof Dr rer medic; Sonja Schmalen[3], MBA; Ina Otte[4], Prof Dr sc med; Jan Peter Ehlers[1], Prof Dr med vet

[1]Department of Didactics and Educational Research in Health Science, Faculty of Health, Witten/Herdecke University, Alfred-Herrhausen-Straße 50, Witten, North Rhine-Westphalia, Germany

[2]Junior Professorship for Digital Health, Faculty of Health, Witten/Herdecke University, Witten, North Rhine-Westphalia, Germany

[3]Health for Future, Deutsche Allianz für Klimawandel und Gesundheit, Berlin, Germany

[4]Department for Health Care Research, Institute for Diversity Medicine, Ruhr University Bochum, Bochum, North Rhine-Westphalia, Germany

[*]these authors contributed equally

Corresponding Author:
Julia Nitsche, Dr rer medic
Department of Didactics and Educational Research in Health Science, Faculty of Health, Witten/Herdecke University, Alfred-Herrhausen-Straße 50, Witten, North Rhine-Westphalia, Germany

Related Article:

This is a corrected version. See correction statement: https://mededu.jmir.org/2026/1/e111087

Abstract

Background: The focus on sustainability in university teaching and education about the climate catastrophe is constantly increasing and is essential for creating change. The chances offered by digital technological innovations are decisive in spotlighting planetary health education. Alongside welfare economies and social movements, they are among the areas with great potential to drive sustainable development.

Objective: This study aimed to evaluate attitudes toward digital planetary health before and after a fully digital planetary health lecture series. This included assessing environmental health awareness, climate-related health risks, and perceptions of the need to integrate digital and planetary health education into university curricula.

Methods: A repeated, pseudonymous, nonexperimental, cross-sectional online survey was conducted before (T1) and after (T2) a fully digital lecture series on digital planetary health during the 2022-2023 winter term. Because participant matching across time points was unsuccessful, descriptive analyses were conducted at the group level.

Results: A total of 180 participants completed the T1 survey, and 79 participants completed the T2 survey. The mean age in the T1 survey was 27.92 (SD 10.95) years, with a minimum age of 18 years and a maximum age of 77 years. The mean age in T2 was 29.22 (SD 12.78) years, with a minimum age of 18 years and a maximum age of 78 years. In the T2 survey, 70.9% (56/79) of the participants rated it as "(very) important" (mean 5.38, SD 1.15) to incorporate climate-related health impacts and digital health issues into their studies. In the T2 survey, 51.9% (41/79) of the participants stated that the range of educational events on digital health at universities was inadequate. Qualitative responses highlighted perceived benefits of digitalization for collaboration, knowledge dissemination, and health care access, while concerns regarding resource consumption and ethical challenges were also described.

Conclusions: Participants who attended the lecture series perceived digital planetary health as highly relevant to university education. The findings provide exploratory insights into attitudes toward the integration of planetary health and digital health topics into higher education curricula.

(JMIR Med Educ 2026;12:e89259) doi:10.2196/89259

KEYWORDS

education; digital medicine; technology-enhanced learning; digital competencies; technology; planetary health

XSL•FO
RenderX

Introduction

The focus on sustainability in university teaching as well as education about the climate catastrophe is steadily increasing [1]. In this context, Rossa-Roccor et al [2] further see academics as having an obligation to take an active role in the sphere of knowledge and action. They call for a commitment to policy change through an increased focus on planetary health; in short, to highlight climate change as a human health problem [2].

A meta-analysis of systematic reviews by Rocque et al [3] on the health impacts of climate catastrophe included 94 articles. They found meteorological and extreme weather events as the most common climate impacts. The most common health impacts were (1) infectious disease; (2) mortality; and (3) respiratory, cardiovascular, or neurological diseases [3].

However, this knowledge is not strongly disseminated among health care professionals (HCPs). Hathaway and Maibach [4] found that while there is a perception among HCPs toward the potential harm of the climate crisis on health, the self-assessed knowledge of HCPs is low. Societal perceptions of the hazards of the climate crisis are also high [4]. This supports the need for education in society at large, and especially for students (in the health professions), who can shape tomorrow's society and drive individual and political change.

Decisive for the particular focus on planetary health education are the opportunities presented by digital technological innovations. Along with economic prosperity and social movements, these innovations are one of the areas with great potential to drive sustainable development [5]. The joint realization of a transformation toward a society that is both sustainable and digital is often referred to as a twin transformation. It is important that these processes influence each other positively (an integrated path) and do not compete with one another [6].

Against this background, the lecture series *Digital Medicine Goes Planetary Health* (DMgPH) was developed in the 2022-2023 winter semester, focusing on the connections between digital health care and planetary health. Lecturers from Witten/Herdecke University (UW/H; JN and JPE), German Climate Change and Health Alliance (SSch), and Ruhr University Bochum (RUB; TSB) organized and moderated weekly 90-minute sessions in which they themselves, as well as guest lecturers from various areas of digital planetary health, spoke.

Although there has been a clear trend toward activating, participatory, and competence-oriented teaching formats in the discussion on university teaching in recent years, traditional formats such as lectures retain their didactic relevance under certain conditions [7]. The much-discussed change in basic assumptions from teacher-centered to learner-centered formats [8,9] has led to a meaningful shift in focus. Nevertheless, studies show that lectures—especially in digital and interactive formats—can make an independent contribution to higher education [6,10]. The online lecture series used in this project offers a concise example of this: digital access has made it possible to reach a broad, transdisciplinary audience and involve many external experts who would not have been available for small face-to-face seminars [11]. Lecture series not only promote the dissemination of current scientific findings but also enable a multiperspective examination of the complex challenges in the context of digital planetary health [2,5]. The focus here is on encouraging further engagement with the topic and highlighting its relevance [12]. Lectures are the most effective way of achieving this with a large audience, as they help to develop a mutual understanding [13]. A more in-depth examination of the topic should then take place in more learner-centered formats [14].

The aim of this study was to survey the attitudes of participants before and after participation in DMgPH regarding environmental awareness and visions of the future. The following research questions were considered:

- What attitudes and self-reported perceptions regarding climate-related health impacts and digital planetary health were reported by participants before the lecture series?
- How do group-level responses before and after the lecture series differ regarding environmental awareness and perceptions of digital planetary health?

Methods

Ethical Considerations

Ethical approval was obtained from the ethics committee of UW/H, Germany (approval code: S-198/2022). Participants did not receive compensation for their participation.

All participants provided informed consent prior to participation. For the online survey conducted via LimeSurvey, participants were presented with detailed study information and were required to confirm that they had read and understood the information and agreed to participate before accessing the questionnaire. Participation was voluntary, and respondents could discontinue the survey at any time without disadvantage.

For the qualitative data collection, participants were likewise informed that their responses would be anonymized and used exclusively for research purposes. They were instructed to provide qualitative responses only if they agreed to the anonymized use of their data for scientific analysis.

Setting and Lecture Series

The lecture series was also held in conjunction with a lecture series on transformative education in Germany, organized by the Planetary Health Academy, which, however, did not focus on digitalization [14]. DMgPH wanted to include the themes in the curriculum and move away from focusing on the personal footprint in relation to the climate crisis and toward increasing the handprint [15]. The handprint is a concept that focuses on actions and the possibility of changing things as an individual [16].

DMgPH was delivered digitally using Zoom meeting software (Zoom Communications, Inc) during the winter term of 2022-2023. Students at UW/H and medical students at RUB were eligible to receive 2 European Credit Transfer and Accumulation System credits in the DMgPH. This was tracked via logins using personalized links; therefore, there was no need

XSL•FO
RenderX

to register in advance with real names on Zoom. Students from other universities were given the opportunity to have their participation in the lecture series confirmed and to submit this to their respective examination offices for credit.

Sessions 1 to 11 were characterized by 1 or 2 keynotes by experts from various fields (eg, hospitals and nongovernmental organizations). Each session was followed by a subsequent discussion (chat or video). Work in breakout sessions, surveys, and encouragement to join discussions were used to improve attendee participation. The lecturers moderated the discussions in turn.

At the 12th session, a bar camp was held. Bar camps are characterized by questions from the participants or concrete input [17] and designed to create learning spaces together with the participants [18] where they meet as equals [19]. In DMgPH, the participants were asked in advance to indicate possible areas of interest on Moodle (Moodle Pty Ltd) or to offer their own sessions in the lecture series. Six sessions were offered in which both the lecturers and other experts presented certain topics with a short keynote speech, and discussion arose through joint work assignments (Table 1).

The final session was designed as a quiz in which questions were asked about all previous dates in the lecture series. The 207 participants had 30 seconds to answer each question. Three winners (for the quickest right answers) were awarded prizes as a pleasant incentive (eg, stickers, socks, and a hip bag).

Table . An overview of the *Digital Medicine Goes Planetary Health* program in WT22/23[a].

Session	Topic	Lecturers
1	Introduction to planetary health and digital medicine	JN, TSB, SSch, JPE, and Head of German Climate Change and Health Alliance (KLUG e.V.)
2	Digitization, sustainability and global health	Professor of Witten/Herdecke University and a collaborator of the CO:DINA project [20]
3	The German Sustainability Code and its implementation at the University Medical Centre Hamburg-Eppendorf	Employee of the German Sustainability Code and an employee of the Sustainability Department of the University Medical Center Hamburg-Eppendorf
4	Sustainability and telemedicine Green Health – what saves the world?	Director of the Department for Trauma Surgery, Orthopedics and Sports Medicine at Helios University Hospital Wuppertal Employee from eternalHealth [21]
5	Gamification for planetary health	Staff of the German Foreign Office and Leuphana University
6	Democratization of health and value-based healthcare	Oxford University collaborator and Patients4Digital contributor
7	Legal perspective on sexuality as a human right in a digitalized world	Employee of the University of Hanover
8	Transformation of two clinics—moving from knowledge to action	Employee of the specialist clinic Gaißach, employee of the Charité Berlin
9	Duisburg-Essen University Hospital—Smart Hospital goes Green Hospital	Two employees of the Duisburg-Essen University Hospital
10	Digitization in agriculture and planetary health	JPE, employee from Health for Future
11	Ethical consideration of digital medicine and interoperability and equity in the context of digitization	Professor from Charité Berlin, professor from FOM University
12	Bar camp	Six sessions from various internal and external lecturers
13	Final quiz using Kahoot! [22]	JN, TSB, and JPE

[a]WT22/23: 2022-2023 winter term.

Participants

Overview

Participation in the lecture series was open to everyone who was interested. Participants were made aware of the event through the curriculum, the social media channels of the university, the lecturers and experts, and various newsletters. A total of 697 participants were assessed for eligibility, since they had registered for the lecture series throughout the whole semester (including lecturers). An average of 257 (SD 33) participants attended each session and could also be invited to participate. About 91% (634/697) of all registered participants were students from 94 different universities and 98 different fields of study (eg, medicine, dentistry, psychology, and nursing, as well as theology, law, and business administration).

XSL•FO
RenderX

Sample Size

No formal a priori sample size calculation was conducted because the study was exploratory, and participation depended on voluntary attendance at the lecture series.

Recruitment

Recruitment materials were distributed through university mailing lists, social media channels, newsletters, and Zoom's registration system. All individuals who registered for the lecture series via Zoom received email invitations to participate in the surveys, irrespective of university affiliation. Moodle announcements were also used for participants affiliated with the 2 participating universities. The procedure remained the same in both the T1 and T2 surveys.

Participants were recruited for the T1 survey from October 13, 2022, to October 28, 2022. The research project was introduced to all participants during the first session of DMgPH.

The link to the T1 survey was distributed twice via chat to all participants in the first session and, as described above, via email.

For the T2 survey, participants were recruited from January 19, 2023, to January 29, 2023. The T2-survey questionnaire was sent out on the penultimate and last days of the lecture series via the chat and, as described above, via email.

Eligibility Criteria

Participants were only able to take part in the survey if they answered that they were planning to attend the lecture series regularly (T1) and if they had attended the lecture series at least 11 of 13 times (T2). Additionally, they had to be of legal age (T1 and T2).

Survey Instrument

Overview

An online questionnaire (Multimedia Appendix 1) was developed using the SoSciSurvey platform [23]. It was a nonexperimental, cross-sectional survey to explore group-level differences before and after the lecture series. All questions were to be answered voluntarily. During the T1 survey, participants were asked to generate a pseudonym, which they were asked to provide again during the T2 survey to merge T1 and T2 data. The pseudonym was derived by the participants from specific letters of personal information (first letter of the mother's first name, last letter of the place of birth, birthday as a 2-digit number, first letter of the first name, and house number of the first address) to generate a test subject code (Multimedia Appendix 1).

The questionnaire itself consists of 4 sections. Section A assesses sociodemographic characteristics. Section B assesses ecological attitudes using the New Ecological Paradigm (NEP) scale, a validated instrument for measuring environmental worldviews and attitudes, which was used in its German translation. Sections C and D include items assessing both knowledge-related perceptions and attitudes toward climate-related health impacts, digitalization, and digital planetary health. These items were adapted from previous studies and translated using forward-backward translation where applicable. The feedback section includes a mix of knowledge- and attitude-related open-text responses regarding the lecture series and the integration of digital planetary health into university curricula.

The lecture series and questionnaire were informed by transformative learning theory, which provides a conceptual framework for understanding how learners engage with complex and value-laden topics such as digital planetary health. From this perspective, learning is not limited to the acquisition of factual knowledge but also involves critical reflection on existing assumptions and the integration of new perspectives into one's understanding of the world [24,25]. However, whether such reflection or attitude-related differences translate into actual behavior change remains to be examined in future research.

The questionnaires at T1 and T2 were almost identical. Any discrepancies are explicitly noted below.

Sociodemographic Data

Sociodemographic information, including sex, age, size of household, school-leaving certificate, relationship to medical care, education by university, study program, and semester, was recorded as contextual data. In the T1 and T2 surveys, commitment to climate activist groups was assessed.

Relationship Between Humans and the Environment

This section examines the extent to which participants have an ecologically friendly and environmentally friendly worldview toward the health of the planet and how this is changing, on a scale ranging from 1 to 6 (1="strongly disagree" to 6="strongly agree"). The NEP scale, which is a validated instrument for assessing ecological worldviews [26], was used in its German translation [27].

Health Problem Assessment

The assessment of the participants toward digital planetary health is surveyed in this section. The survey is based on the list of various health problems used in other studies [28,29]. For the survey, the list was adapted and translated via forward-backward translation. The list was used to survey first whether participants are already experiencing effects from the climate catastrophe and second whether they suspect that this will be the case in 20 years (response options: "Yes"/"No"/"I don't know").

Attitude Toward Digitalization

In this section, the participants were asked for their opinions in free-text responses on how digitization can positively or negatively impact the climate crisis.

Attitude Toward Digital Planetary Health at the University

In T2, participants were asked to rate statements about the importance of addressing the climate catastrophe, its impact on health, and digital aspects in their own study program or other study programs, as well as their impressions of the scope of educational opportunities in the field of digital planetary health, which they rated on a scale from 1 to 6 (1="do not agree at all" to 6="completely agree").

XSL•FO
RenderX

Participants were also asked to write down free-text responses on how topics related to digital planetary health could best be integrated into the curricula of study programs.

Feedback on the Lecture Series

Participants were also asked to give qualitative feedback on the lecture series at T2 only, using free-text responses (Textbox 1).

Textbox 1. Content of the T1 and T2 surveys.

- Sociodemographic data
- Sex, age in years, size of household, school-leaving certificate, relationship to medical care (T1), education by university, study program, semester, commitment to climate activism (T1)
- Relationship between humans and the environment
- Statements regarding the relationship between humans and the environment, based on the New Ecological Paradigm scale by Schleyer-Lindenmann et al [27] (15 statements ranging from 1 to 6)
- Health problem assessment
- Rating of 12 health problems on whether they have increased at the current time, and rating of 12 health problems on whether they will increase in the next 20 years (response options: “Yes”/“No”/“I don’t know”)
- Attitude toward digitalization
- Qualitative feedback on the influence of digitalization on digital planetary health
- Attitude toward digital planetary health at university
- Importance of digital planetary health and perceptions of the sufficiency of existing educational opportunities of digital planetary health within participants’ own study programs or those of others (2 statements rated on a scale from 1 to 6), and qualitative feedback on opportunities to integrate digital planetary health into curricula
- Feedback on the lecture series
- Qualitative feedback on the lecture series

Data Analysis

Quantitative Analysis

The data analysis was primarily descriptive. First, datasets containing missing information were excluded from the analysis. A pseudonym-based comparison of responses between measurement points T1 and T2 was unsuccessful, meaning that no reliable longitudinal analyses could be carried out at the individual level. For this reason, the data from both measurement points were analyzed separately. The interpretation was based on the descriptive statistics obtained. Data analysis was performed using SPSS (version 29; IBM Corp).

Qualitative Analysis

The answers to the qualitative question in Section D were evaluated by content analysis [30] using MAXQDA Plus 2022 (VERBI Software GmbH) involving 2 researchers (TSB and IO). To begin with, the 2 researchers analyzed the entire dataset including answers to (1) attitude toward digitalization (T1 and T2), (2) attitude toward digital planetary health at universities (T2), and (3) feedback on the lecture series (T2).

Both researchers analyzed the data separately and inductively developed categories for sections A to C, resulting in 3 codebooks, one for each section. The 2 researchers met to discuss their findings and merged their coding systems. No conflicts arose that required the involvement of a third party. Intercoder agreement was not assessed. The quotes in the manuscript were selected to include particularly striking examples offering unique perspectives, as well as those reflecting a broad consensus and recurring frequently in a similar vein. Care was also taken to ensure a balanced representation of students and nonstudents, as well as across genders and degree programs to highlight the diversity of perspectives.

Results

Participant Profile

A total of 215 participants took part in the T1 survey, and 105 participants took part in the T2 survey. To ensure data quality, all incomplete data records were excluded during the data cleaning process. Only data records containing complete information on the relevant variables were included. A total of 180 participants from T1 and 79 participants from T2 were included in the final analysis (Table 2). The mean age in the T1 survey was 27.92 (SD 10.95) years, with a minimum age of 18 years and a maximum age of 77 years. The mean age in the T2 survey was 29.22 (SD 12.78) years, with a minimum age of 18 years and a maximum age of 78 years.

XSL•FO
RenderX

Table . Sociodemographic data of participants in the T1 and T2 surveys.

Characteristics	T1 (n=180)	T2 (n=79)
Age (y), mean (SD; range)	27.92 (10.95; 18-77)	29.22 (12.78; 18-78)
Gender, n (%)		
Women	137 (76.1)	52 (65.8)
Men	43 (23.9)	27 (34.2)
Student status, n (%)		
Student	155 (86.1)	68 (86.1)
Nonstudent	25 (13.9)	11 (13.9)
Household size, n (%)		
1	50 (27.8)	22 (27.8)
2	68 (37.8)	32 (40.5)
3	32 (17.8)	13 (16.5)
4	23 (12.8)	10 (12.7)
5	7 (3.9)	2 (2.5)
Currently working in the health care sector (y), n (%)		
Yes	83 (46.1)[a]	40 (50.6)[b]
No	97 (53.9)	39 (49.4)
Previously worked in the health care sector (y), n (%)		
Yes	35 (19.4)[c]	16 (20.3)[d]
No	145 (80.6)	63 (79.8)
Volunteering in climate protection, n (%)		
Little to no engagement	117 (65.0)	47 (59.5)
Moderate engagement	41 (22.8)	20 (25.3)
High engagement	22 (12.2)	12 (15.2)

[a]Mean 7.78 (SD 7.98; range 1-45).

[b]Mean 10.17 (SD 10.37; range 0.5-32).

[c]Mean 4.83 (SD 6.4; range 0.5-27).

[d]Mean 4.16 (SD 3.54; range 0.5-14).

The majority of participants were women. In the T1 survey (n=180), 137 (76.1%) participants were women and 43 (23.9%) participants were men. In the T2 survey (n=79), 52 (65.8%) participants were women and 27 (34.2%) participants were men. The “nonbinary” option was not selected.

In the T1 survey, 155 (86.1%) participants were students. Of those, 76.8% (119/155) were students at UW/H and 4% (6/155) were students at RUB. The remaining 19.4% (30/150) were from 13 other national and international universities. Twenty-five (13.9%) participants in the T1 survey were nonstudents.

In the T2 survey, 68 (86.1%) participants were students. About 86.8% (59/68) of them were at UW/H, 2.9% (2/68) at RUB, and 10.3% (7/68) at 3 other national universities (Table 2). In the T2 survey, 11 (13.9%) participants were nonstudents.

In terms of household size, most participants lived in 2-person households, with 68 (37.8%) participants in T1 and 32 (40.5%) participants in T2. Single-person households were the second most common category, comprising 50 (27.8%) participants in T1 and 22 (27.8%) participants in T2. Three-person households comprised 32 (17.8%) participants in T1 and 13 (16.5%) participants in T2. Four-person households were reported by 23 (12.8%) participants in T1 and 10 (12.7%) participants in T2. Households with 5 participants were only marginally represented, with 7 (3.9%) participants in T1 and 2 (2.5%) participants in T2.

At the time of the survey, 83 (46.1%) participants in T1 and 40 (50.6%) participants in T2 were working in the health care sector. In contrast, 97 (53.9%) participants in T1 and 39 (49.4%) participants in T2 were not working in the health care sector. The average length of professional experience in the health care sector among those currently employed was 7.78 (SD 7.98) years in T1 and 10.17 (SD 10.37) years in T2. In addition, 35 (19.4%) participants in T1 and 16 (20.3%) participants in T2 stated that they had previously worked in the health care sector.

As part of the T1 measurement, participants were asked whether they were already actively involved in climate protection (eg,

XSL•FO
RenderX

in a Campus for Future activist group). This made it possible to estimate the participants' handprints. One hundred seventeen (65%) participants in T1 and 47 (59.5%) participants in T2 reported low to no engagement. A moderate level of engagement was reported by 41 (22.8%) participants in T1 and 20 (25.3%) participants in T2. A high level of engagement was reported by 22 (12.2%) participants in T1 and 12 (15.2%) participants in T2.

Relationship Between Humans and the Environment

The results of the T1 and T2 surveys using the NEP scale can be found in Table 3.

To examine differences between T1 and T2, descriptive statistics in the form of means and SDs were reported for individual items of the NEP scale (Table 3) and for the assessment of health problems (Table 4). Across most items, there were only minor differences between the 2 measurement points, with the means generally falling within comparable ranges and similar variation in responses.

Table . Content items regarding the New Ecological Paradigm scale in the T1 and T2 surveys[a].

Item	T1 (n=180), mean (SD)	T2 (n=79), mean (SD)
We are approaching the limit of the number of people the Earth can support.	4.62 (1.32)	4.59 (1.45)
Humans have the right to modify the natural environment to suit their needs.	2.93 (1.14)	2.71 (1.14)
When humans interfere with nature it often produces disastrous consequences.	4.78 (1.10)	4.97 (1.05)
Human ingenuity will ensure that we do not make the Earth unlivable.	3.50 (1.18)	3.24 (1.22)
Humans are seriously abusing the environment.	5.22 (1.03)	5.34 (1.05)
The Earth has plenty of natural resources if we just learn how to develop them.	4.48 (1.29)	4.58 (1.31)
Plants and animals have as much right as humans to exist.	5.02 (1.20)	5.13 (1.23)
The balance of nature is strong enough to cope with the impacts of modern industrial nations.	2.15 (1.23)	2.03 (1.07)
Despite our special abilities, humans are still subject to the laws of nature.	5.04 (1.15)	5.03 (1.30)
The so-called "ecological crisis" facing humankind has been greatly exaggerated.	1.55 (0.97)	1.41 (0.91)
The Earth is like a spaceship with very limited room and resources.	3.58 (1.42)	3.59 (1.58)
Humans were meant to rule over the rest of nature.	1.72 (1.10)	1.53 (0.99)
The balance of nature is very delicate and easily upset.	4.57 (1.19)	4.53 (1.40)
Humans will eventually learn enough about how nature works to be able to control it.	2.90 (1.17)	2.76 (1.26)
If things continue on their present course, we will soon experience a major ecological catastrophe.	5.33 (1.03)	5.41 (1.01)

[a]Likert scale (1="do not agree at all" and 6="completely agree").

XSL•FO
RenderX

Table . Content items regarding health problems at present and in 20 years in the T1 and T2 surveys[a].

Item	T1 (n=180)		T2 (n=79)	
	Content items regarding health problems at present, mean (SD)	Content items regarding health problems in 20 years, mean (SD)	Content items regarding health problems at present, mean (SD)	Content items regarding health problems in 20 years, mean (SD)
Vector-borne infectious disease	0.99 (0.54)	0.91 (0.52)	1.03 (0.59)	0.86 (0.74)
Flooding-related displacement	0.94 (0.48)	0.98 (0.30)	0.91 (0.47)	0.99 (0.26)
Mental health conditions	1.00 (0.33)	0.88 (0.52)	0.99 (0.38)	1.00 (0.28)
Air quality–related illness	0.93 (0.67)	0.91 (0.55)	0.86 (0.63)	0.93 (0.50)
Food-borne disease	0.87 (0.92)	0.82 (0.78)	0.83 (0.92)	0.97 (0.67)
Disruption of health care services during extreme weather events	0.71 (0.96)	0.91 (0.58)	1.11 (0.70)	0.96 (0.60)
Water-borne infectious disease	0.66 (1.02)	0.89 (0.63)	0.84 (1.01)	0.91 (0.68)
Heat-related illness	0.88 (0.65)	0.91 (0.44)	0.93 (0.64)	0.96 (0.45)
Cold-related illness	0.88 (1.14)	0.69 (1.00)	0.86 (1.09)	0.92 (0.94)
Water-availability illness	0.83 (0.78)	0.95 (0.45)	0.91 (0.68)	0.93 (0.50)
Malnutrition	1.01 (0.67)	0.99 (0.50)	1.09 (0.57)	0.95 (0.57)
Other climate change health-related impacts	0.74 (0.75)	0.79 (0.63)	0.74 (0.74)	0.92 (0.48)

[a]Response options: “Yes (1)”/“No (0)”/“I don’t know” (2).

The descriptive statistics in Table 3 showed a broadly similar pattern of responses at both measurement points. Particularly high mean scores were found for the items “Humans are seriously abusing the environment” (T1: mean 5.22, SD 1.03; T2: mean 5.34, SD 1.05) and “If things continue on their present course, we will soon experience a major ecological catastrophe” (T1: mean 5.33, SD 1.03; T2: mean 5.41, SD 1.01). Comparatively high scores were also observed for “Plants and animals have as much right as humans to exist” (T1: mean 5.02, SD 1.20; T2: mean 5.13, SD 1.23) and for “Despite our special abilities, humans are still subject to the laws of nature” (T1: mean 5.04, SD 1.15; T2: mean 5.03, SD 1.33).

In contrast, the mean scores were lower for statements formulated in a more anthropocentric or techno-optimistic manner. This applied, for example, to “Human ingenuity will ensure that we do not make the Earth uninhabitable” (T1: mean 3.50, SD 1.19; T2: mean 3.24, SD 1.22) and “Humans were meant to rule over the rest of nature” (T1: mean 1.72, SD 1.10; T2: mean 1.54, SD 1.00). The item “The so-called ecological crisis facing humankind has been greatly exaggerated” also showed comparatively low levels of agreement (T1: mean 1.55, SD 0.97; T2: mean 1.41, SD 0.91).

The descriptive statistics in Table 4 showed a broadly similar pattern of responses for both T1 and T2. For the assessment of current health problems, the mean scores at both measurement points were predominantly in the range of approximately 0.66 to 1.01. Particularly high values were observed, for example, for mental health conditions, with mean 1.00 (SD 0.33) in T1 and mean 0.99 (SD 0.38) in T2, as well as for malnutrition, with mean 1.01 (SD 0.67) in T1 and mean 1.09 (SD 0.57) in T2. Comparatively lower mean values were reported, among other things, for water-borne infectious diseases, which stood at mean 0.66 (SD 1.02) in T1 and mean 0.84, SD (1.01) in T2.

Similar values were also observed across both measurement points for the assessment of health problems in 20 years. Here, the mean values generally ranged between 0.79 and 0.99. For example, the mean for mental health conditions was 0.88 (SD 0.52) in T1 and mean 1.00 (SD 0.28) in T2. Malnutrition also showed relatively high values, with mean 0.99 (SD 0.50) in T1 and mean 0.95 (SD 0.57) in T2. By contrast, slightly lower mean values were observed for food-borne disease, with mean 0.82 (SD 0.78) in T1 and mean 0.97 (SD 0.67) in T2, as well as for other climate change–related health impacts, with mean 0.79 (SD 0.63) in T1 and mean 0.92 (SD 0.48) in T2.

Overall, the descriptive values suggested that respondents assessed the various health impacts of climate change to a similar extent, both currently and in the future. The SDs varied across the items but were moderate overall, indicating a certain degree of variation within the sample. Descriptively, there were only minor differences between T1 and T2.

Health Problems

The results from T1 and T2 regarding the assessment of health problems at the time of the survey and in 20 years’ time are shown in Table 4.

XSL•FO
RenderX

Attitude Toward Digitalization

During the qualitative evaluation of the answers to this section, 290 codes were identified, with positive evaluations spanning 197 codes and negative evaluations of digitalization spanning 53 codes. In addition, ambiguous evaluations (21 codes) and demands (19 codes) were identified.

Participants described digitalization predominantly as a tool with transformative potential for addressing the climate crisis and strengthening planetary health. Responses frequently reflected a perceived tension between the transformative potential of digital technologies and concerns regarding their environmental and societal consequences.

A central theme concerned the ability of digital technologies to facilitate communication, collaboration, and access to information. Participants associated digitalization with faster dissemination of knowledge, improved networking across disciplines and regions, and lower barriers to participation, particularly for people in rural areas or with limited access to educational resources. Participants further described digital technologies as supporting organizational efficiency, crisis management, and broader transformation processes within society and health care.

Within the health care context, participants particularly emphasized the potential of digital technologies to improve access to care independent of location. In addition, they referred to opportunities for more efficient collaboration, prevention, diagnostics, and therapy, as well as the reduction of duplicated services and unnecessary resource use. Several participants further emphasized the potential of digital technologies to support public education and global engagement in planetary health.

At the same time, participants stressed that technological innovation alone would not be sufficient and repeatedly linked digitalization to broader societal transformation processes and normative questions of responsibility. One participant stated, “We need to transform our society(s) and change our own attitudes in the process” (T1: student of sustainability management).

Participants also emphasized that the societal benefits of digitalization depend on ethical and sustainable implementation. One participant stated, “However, this requires prudent planning, honesty and innovation for the benefit of the people and not greed for profit” (T2: psychology student).

Alongside the opportunities, participants critically reflected on the ecological and social risks associated with digitalization. Most prominently, they referred to increasing resource consumption, including electricity, rare-earth extraction, and electronic waste. In addition, participants raised serious concerns regarding negative impacts on mental health, dependency on technology, social isolation, and ethical risks such as discrimination by AI.

Digital Planetary Health in Study Programs

Participants largely regarded digital planetary health as an important topic for higher education and emphasized the need for stronger curricular integration across study programs. While respondents perceived the topic as highly relevant, they simultaneously described existing educational opportunities in this field as insufficient.

The results of the survey among T2 participants on the relevance of teaching digital planetary health are presented in Table 5.

Table . Attitudes toward digital planetary health at university in the T2 survey (n=79)[a].

Item	Mean (SD)	Participants with partial to full agreement, n (%)
It is important to address climate catastrophe, health impacts, and digital issues in the curriculum.	5.38 (1.16)	56 (70.9)
I have the impression that students (at my university) receive sufficient educational offerings in the area of digital planetary health.	3.49 (1.47)	38 (48.1)

[a]Likert scale (1=“do not agree at all” to 6=“completely agree”).

In addition to the quantitative query, survey participants were asked at T2 to describe how topics related to digital planetary health could best be integrated into their own studies and/or courses with which they were familiar. A total of 60 codes were identified in the qualitative analysis.

In the qualitative responses, participants repeatedly expressed the wish that educational offerings on digital planetary health should become more broadly accessible across universities and disciplines. Many participants advocated permanent integration into curricula, including both mandatory and elective formats. Mandatory teaching was perceived as particularly necessary to ensure that the topic reaches students beyond those already interested in sustainability-related issues. One participant stated, “[...] otherwise only the engaged ones learn it, which are only a fraction of all students” (T2: medical student).

Participants further described digital planetary health education as an opportunity to develop new perspectives on familiar topics and to strengthen their perceived ability to contribute to societal change (“handprint”). Practical relevance and applicability to future professional contexts were repeatedly emphasized.

Moreover, respondents highlighted the value of multiprofessional collaboration and interactive teaching formats. They emphasized that the lecturers should possess strong expertise in the field and communicate content in an evidence-based yet emotionally engaging manner. Participants also expressed interest in more specialized content, including preventive medicine and health psychology.

XSL•FO
RenderX

Feedback

A total of 78 codes were identified in the qualitative analysis of this section.

Overall, participants described the lecture series positively and highlighted both the interdisciplinary perspective and the digital format as particular strengths.

Participants frequently characterized the sessions as interactive, engaging, and informative. The atmosphere and moderation were described as pleasant, while the invited speakers were perceived as competent and knowledgeable. Participants particularly appreciated the combination of digital health and planetary health perspectives, which many had not previously encountered in an integrated educational format:

> *It was super interesting, the linkages between Planetary Health and Digital Medicine have never really been presented in such a combined way. Thank you for offering this lecture series.* [T2: medical student]

The digital format was also perceived as increasing flexibility and enabling exchange among participants and experts from diverse backgrounds. Participants valued the broad overview of the field and described the lecture series as successfully connecting prior knowledge with new perspectives.

Beyond knowledge acquisition, several participants reported that the lecture series stimulated personal reflection and discussions within their social environments regarding the climate crisis and environmental responsibility:

> *[…] the event made me reflect on my own environmental behavior and engage in discourse and exchange with friends/family on such topics. […]* [T2: health science student]

Critical feedback mainly focused on the desire for greater inclusion of topics related to animals and plants, more opportunities for participants' own research activities, and, in some cases, a more neutral stance from lecturers.

Participants also suggested expanding interactive and research-oriented elements in future iterations of the lecture series.

Discussion

Principal Findings

This study explored participants' perceptions, attitudes, and knowledge related to digital planetary health before and after participation in a fully digital lecture series. Overall, the findings suggest a high perceived relevance of integrating digital planetary health into higher education curricula. Participants particularly emphasized the importance of addressing climate-related health impacts within university education and highlighted the potential of digital technologies. The survey responses from T1 and T2 further indicate that participation in the lecture series may have contributed to increased awareness of specific climate-related health risks, particularly regarding water- and food-borne diseases. At the same time, qualitative findings revealed a differentiated understanding of digitalization, as participants described both opportunities and challenges associated with digital transformation in relation to sustainability.

In their feedback on the lecture series, the participants stated that the lecture series was positive and encouraged them to reflect on their own behavior and talk about it with friends and family. This finding is particularly important in the training of researchers, as the body of research on the twin transformation, sustainability, and digitalization in the health care sector is rather limited [31]. The survey results underline the relevance of interdisciplinary educational approaches that connect planetary health, digital health, and societal transformation. Universities do not merely have a duty to highlight problems; they must also make their own efforts, for example, by operating within planetary boundaries [32]. McGeown and Barry [33] put it in strong terms: "As producers and gatekeepers of knowledge, and as providers of education and training, our universities play a key role in the reproduction of unsustainability." To overcome this, it seems sensible to develop new and open formats that bring students and society together to discuss problems and solutions.

The results of the T1 survey show that the participants started the lecture series with a high level of knowledge and awareness of the climate crisis. This self-selection bias has been identified in various studies and must be explicitly addressed and overcome through curriculum planning [34]. For future similar formats, it is of great relevance to reach people who are also active outside this focus and who have been less involved in the field of planetary health to date, for example, by incorporating it directly into the compulsory curriculum of the degree programs. This is consistent with the data collected on engagement in climate protection among the participants in the lecture series: an extrapolation to the whole population in Germany aged 14 years and older, based on a survey of 23,524 participants, reveals that 23% of the population is engaged in climate protection [35]. Of the 180 participants, 63 (35%) in the T1 survey stated that they are engaged in climate protection groups at a moderate or high level.

Another challenge here is that some prior knowledge was necessary to be able to follow the event fully. Both nationally in Germany [36] and internationally [37,38], there are already various educational offerings in the field of planetary health for HCPs. It is now crucial to network these programs, create synergies, and channel existing activism. This requires strategically oriented education, supported by integration into curricula, to increase awareness and reach people who are not yet aware of the issue. It could also be worth considering advertising the event on social media, in print media (daily newspapers), and on the radio. Particularly at a time when AI is being used more and more, the call for a digital planetary health perspective is becoming increasingly urgent [39]. It is desirable that this be implemented jointly by academia and society in the form of transformative science [40].

With regard to the selection of guest lecturers, it should be noted that they were recruited from the lecturers' network. The involvement of guest experts in lectures ensures a good balance between theory and practice, thereby fostering both a love of

XSL•FO
RenderX

learning and the application of knowledge [41]. While there was a lack of remuneration, lecturers gained value through interaction with the participants. To prevent excessive subjectification, a meeting with the lecturers and guest lecturers took place beforehand, and a guideline on scientific working methods and gender-sensitive formulations was provided. The possibility of offering a wide variety by inviting experts in combination with its own sessions was judged to outweigh the risk of subjectification in the previous consideration. In addition, the discussion offered the opportunity to organize the content at any time. It has also become apparent in other fields that the multiperspective approach offered by various guest lecturers leads to holistic learning success among lecture attendees [42].

From a behavioral psychology perspective, the changes observed in participants following a lecture series such as this can be explained by several mechanisms. For instance, a review by Dolcos et al [43] shows that emotions, or emotionally charged content, can significantly enhance both attention and episodic long-term memory. Building on this, Immordino-Young and Damasio [44] emphasize that the close link between emotion and cognition, which is continuously at work in the processing of external stimuli, forms a central basis for moral and ethical judgment.

Against this backdrop, the lecture series focused thematically on future prospects and the risks of the climate crisis. At the same time, the program was deliberately designed to highlight not only problem areas but also concrete solutions and individual opportunities for action to strengthen a sense of collective efficacy. This approach aligns with the guiding principles of Education for Sustainable Development, which aim to empower people to actively shape sustainable transformation processes. The results indicate possible starting points for these changes, but they should be viewed as trends.

A clear example of the link described above is provided by the third session, in which the German Sustainability Code and its practical implementation were presented. Overall, the lecture series thus aims to foster students' moral judgment and values through engagement with ethical issues. At the same time, they are encouraged to develop their own positions and to critically engage with different perspectives [45].

In addition to the individual options for action identified, a dynamic also developed during the lecture series that led to further synergies. Several ideas for qualification projects have emerged and are being actively worked on today [46]. There have been various invitations and opportunities at conferences to present the lecture series and engage in discussions about it [47]. In December 2023, the team received the 1st Humboldt^{n} Prize for Sustainability (1. Humboldt^{n}-Preis für Nachhaltigkeit) for the lecture series itself. A concept for a lecture series on human-animal relationships was developed based on course discussions and implemented in subsequent semesters to revisit the many facets of this topic. Accompanying research was also carried out, focusing particularly on changes in participants' attitudes and knowledge [48,49].

Limitations

When looking at the results, several limitations should be considered. First, it is important to note that 86.1% (155/180) of the participants in the T1 survey were students. Opening up the event and promoting it to the wider community proved challenging during implementation. While students from various universities were attracted, the proportion of citizens as participants was very low. In addition, more people from the field of health care were reached. The target group of the lecture series—(future) HCPs—consists of trusted contact persons for their patients and communities. They are, therefore, an important and critical target group to inform and educate about the complex dynamics of climate change [50]. Second, participation in the lecture series was voluntary, so self-selection bias is likely, as individuals already interested in planetary health, sustainability, or digital medicine may have been more likely to participate. Third, reliable longitudinal matching between T1 and T2 was not possible, thereby limiting the assessment of a causal effect of the lecture series. Moreover, the study lacked a control group, which limits causal interpretation. Possible ceiling effects due to high baseline awareness, as well as social desirability in self-reported responses, may also have influenced the findings.

Future surveys should track whether students also take advantage of other learning opportunities on the topics of planetary health and digital medicine outside of the lecture series during the survey period.

Conclusions

This study shows that attendees of the digital lecture series perceived digital planetary health as highly relevant to university education. The lecture series was positively received, particularly regarding its interdisciplinary perspective, digital accessibility, and the integration of planetary health and digital medicine. The comparison of T1 and T2 responses showed exploratory group-level differences in selected perceptions of climate-related health impacts. These findings support the importance of further integrating digital planetary health into university teaching.

Acknowledgments

DeepL was used solely as a translation tool. All translated content was carefully reviewed, edited, and verified by the authors, who take full responsibility for the final wording and content of the manuscript.

Funding

The authors declared no financial support was received for this work.

XSL•FO
RenderX

Conflicts of Interest

None declared.

Multimedia Appendix 1
Online questionnaire (English translation): T1 and T2 surveys.
[DOCX File, 58 KB - mededu_v12i1e89259_app1.docx]

References

1. Leal Filho W, Sima M, Sharifi A, et al. Handling climate change education at universities: an overview. Environ Sci Eur 2021;33(1):109. [doi: 10.1186/s12302-021-00552-5] [Medline: 34603904]
2. Rossa-Roccor V, Giang A, Kershaw P. Framing climate change as a human health issue: enough to tip the scale in climate policy? Lancet Planet Health 2021 Aug;5(8):e553-e559. [doi: 10.1016/S2542-5196(21)00113-3] [Medline: 34390673]
3. Rocque RJ, Beaudoin C, Ndjaboue R, et al. Health effects of climate change: an overview of systematic reviews. BMJ Open 2021 Jun 9;11(6):e046333. [doi: 10.1136/bmjopen-2020-046333] [Medline: 34108165]
4. Hathaway J, Maibach EW. Health implications of climate change: a review of the literature about the perception of the public and health professionals. Curr Environ Health Rep 2018 Mar;5(1):197-204. [doi: 10.1007/s40572-018-0190-3] [Medline: 29423661]
5. de León EA, Shriwise A, Tomson G, et al. Beyond building back better: imagining a future for human and planetary health. Lancet Planet Health 2021 Nov;5(11):e827-e839. [doi: 10.1016/S2542-5196(21)00262-X] [Medline: 34774123]
6. Barth M, Gossen M, Lang DJ, Santarius T. Sustainable digitalization - fostering the twin transformation in a transdisciplinary way. GAIA 2023 Mar 13;32(1):6-9. [doi: 10.14512/gaia.32.S1.3]
7. Ardalan K. The philosophical foundation of the lecture - versus - case controversy: its implications for course goals, objectives, and contents. Int J Soc Econ 2008;35(1/2):15-34. [doi: 10.1108/03068290810843819]
8. Biggs JB, Tang CS. Teaching for Quality Learning at University: What the Student Does, 4th edition: Open University Press; 2011.
9. Hattie J. Visible Learning: A Synthesis of over 800 Meta-Analyses Relating to Achievement: Routledge; 2009.
10. Bligh DA. What's the Use of Lectures?, 1st edition: Jossey-Bass Publishers; 2000. URL: https://search.worldcat.org/title/What%27s-the-use-of-lectures/oclc/972837966 [accessed 2026-07-21]
11. Seymour-Walsh AE, Weber A, Bell A. Pedagogical foundations to online lectures in health professions education. Rural Remote Health 2020 May;20(2):6038. [doi: 10.22605/RRH6038] [Medline: 32466654]
12. Patrick BC, Hisley J, Kempler T. "What's everybody so excited about?": the effects of teacher enthusiasm on student intrinsic motivation and vitality. J Exp Educ 2000 Jan;68(3):217-236. [doi: 10.1080/00220970009600093]
13. AbiGhannam N, Kahlor L, Dudo A, Liang MC, Rosenthal S, Banner JL. Expectancies and motivations to attend an informal science lecture series. Int J Sci Educ B 2016 Jul 2;6(3):215-238. [doi: 10.1080/21548455.2015.1039468]
14. Torio H, Jimenez A, Golba M. Interdisciplinary teaching in the field of resilient energy systems: experiences with expert lecture series combined with workshops. Int Educ Res 2020;3(1):1. [doi: 10.30560/ier.v3n1p1]
15. Husgafvel R. Exploring social sustainability handprint—part 2: sustainable development and sustainability. Sustainability 2021;13(19):11051. [doi: 10.3390/su131911051]
16. Transformation gestalten lernen [Article in German]. Handprint Hub. 2025. URL: https://www.handprint-hub.de/ [accessed 2026-07-21]
17. Sammet J, Wolf J. Informal learning. In: From Trainer to Agile Learning Facilitator: How Teaching and Learning Works in Digital Times: Springer; 2022:103-118. [doi: 10.1007/978-3-662-65909-0_6]
18. Knackstedt J, Frische B, Jantschek O. Empowerment, Aushandlungsprozesse und der Mut zu Fehlern?! [Article in German]. Journal für politische Bildung 2023 Aug 10;13(3):50-53. [doi: 10.46499/2234.2779]
19. Lisi S. Wie das Edu-Barcamp Diversität kreiert – eine Anleitung. In: Miglbauer M, editor. Lehre 2022: Was Geht? Was Bleibt? Tagungsband Zur 5 Online-Tagung Hochschule [Book in German]: Private Pädagogische Hochschule Burgenland; 2022:75-81. [doi: 10.26041/fhnw-4802]
20. CO: DINA-Transformationsroadmap Digitalisierung und Nachhaltigkeit - Die Digitalisierung für die sozial-ökologische Transformation nutzen [Article in German]. Zukunft – Umwelt – Gesellschaft (ZUG). 2026. URL: https://www.z-u-g.org/foerderung/ki-leuchttuerme-fuer-umwelt-klima-natur-und-ressourcen/projekte/codina/ [accessed 2026-07-21]
21. eternal Health. 2026. URL: https://www.eternalhealth.com/ [accessed 2026-07-21]
22. kahoot. URL: https://kahoot.it/ [accessed 2026-07-21]
23. SoSci Survey. URL: https://www.soscisurvey.de [accessed 2026-07-21]
24. Christie M, Carey MD, Robertson A, Grainger PR. Putting transformative learning theory into practice. Aust J Adult Learn 2015;55(1):10-27 [FREE Full text]
25. Mezirow J. Transformative learning theory. In: Illeris K, editor. Contemporary Theories of Learning: Learning Theorists … In Their Own Words: Routledge; 2018:114-128. [doi: 10.4324/9781315147277-8]

XSL•FO
RenderX

26. Dunlap RE, Van Liere KD, Mertig AG, Jones RE. New trends in measuring environmental attitudes: measuring endorsement of the New Ecological Paradigm: a revised NEP scale. J Soc Issues 2000 Jan;56(3):425-442. [doi: 10.1111/0022-4537.00176]
27. Schleyer-Lindenmann A, Ittner H, Dauvier B, Piolat M. Die NEP-Skala – hinter den (deutschen) Kulissen des Umweltbewusstseins [Article in German]. Diagnostica 2018 Jul;64(3):156-167. [doi: 10.1026/0012-1924/a000202]
28. Felicilda-Reynaldo RFD, Cruz JP, Alshammari F, et al. Knowledge of and attitudes toward climate change and its effects on health among nursing students: a multi-Arab country study. Nurs Forum 2018 Apr;53(2):179-189. [doi: 10.1111/nuf.12240] [Medline: 28950412]
29. Polivka BJ, Chaudry RV, Mac Crawford J. Public health nurses' knowledge and attitudes regarding climate change. Environ Health Perspect 2012 Mar;120(3):321-325. [doi: 10.1289/ehp.1104025] [Medline: 22128069]
30. Kuckartz U, Rädiker S. Analyzing Qualitative Data with MAXQDA: Springer International Publishing; 2019. [doi: 10.1007/978-3-030-15671-8]
31. Berger M, Ehlers JP, Nitsche J. Aligning with the goals of the planetary health concept regarding ecological sustainability and digital health: scoping review. J Med Internet Res 2025 May 28;27:e71795. [doi: 10.2196/71795] [Medline: 40435494]
32. Thierry A, Horn L, von Hellermann P, Gardner CJ. "No research on a dead planet": preserving the socio-ecological conditions for academia. Front Educ 2023;8:1237076. [doi: 10.3389/feduc.2023.1237076]
33. McGeown C, Barry J. Agents of (un)sustainability: democratising universities for the planetary crisis. Front Sustain 2023 May 15;4. [doi: 10.3389/frsus.2023.1166642]
34. Spitzer S. Biases in health expectancies due to educational differences in survey participation of older Europeans: it's worth weighting for. Eur J Health Econ 2020 Jun;21(4):573-605. [doi: 10.1007/s10198-019-01152-0] [Medline: 31989388]
35. Anzahl der personen in deutschland, die sich für natur- und umweltschutz einsetzen, von 2019 bis 2025 [Article in German]. Statista. 2025. URL: https://de.statista.com/statistik/daten/studie/264282/umfrage/lebenseinstellung-einsatz-fuer-natur-und-umweltschutz/ [accessed 2026-07-21]
36. Gepp S, Jung L, Wabnitz K, et al. The Planetary Health Academy-a virtual lecture series for transformative education in Germany. Lancet Planet Health 2023 Jan;7(1):e68-e76. [doi: 10.1016/S2542-5196(22)00253-4] [Medline: 36608952]
37. Asaduzzaman M, Ara R, Afrin S, Meiring JE, Saif-Ur-Rahman KM. Planetary health education and capacity building for healthcare professionals in a global context: current opportunities, gaps and future directions. Int J Environ Res Public Health 2022 Sep 18;19(18):11786. [doi: 10.3390/ijerph191811786] [Medline: 36142057]
38. Leal Filho W, Eustachio JHPP, Paucar-Caceres A, et al. Planetary Health and health education in Brazil: towards better trained future health professionals. Int J Environ Res Public Health 2022 Aug 15;19(16):10041. [doi: 10.3390/ijerph191610041] [Medline: 36011675]
39. Lupton D. Towards a digital planetary health perspective: generative AI and the digital determinants of health. Health Promot Int 2025 Sep 3;40(5):daaf153. [doi: 10.1093/heapro/daaf153] [Medline: 41001915]
40. Schneidewind U, Singer-Brodowski M. Transformative Wissenschaft: Klimawandel Im Deutschen Wissenschafts- Und Hochschulsystem, 2nd edition: Metropolis Verlag; 2014.
41. Alshahrani ALI. The use of guest speakers in higher education and implications for pharmaceutical teaching: a systematic review of literature. F1000Res 2024;13:862. [doi: 10.12688/f1000research.153968.1]
42. Zou P, Sun W, Hallowell SG, Luo Y, Lee C, Ge L. Use of guest speakers in nursing education: an integrative review of multidisciplinary literature. Adv Med Educ Pract 2019;10:175-189. [doi: 10.2147/AMEP.S196456] [Medline: 31118860]
43. Dolcos F, Iordan AD, Dolcos S. Neural correlates of emotion-cognition interactions: a review of evidence from brain imaging investigations. J Cogn Psychol (Hove) 2011 Sep;23(6):669-694. [doi: 10.1080/20445911.2011.594433] [Medline: 22059115]
44. Immordino‐Yang MH, Damasio A. We feel, therefore we learn: the relevance of affective and social neuroscience to education. Mind Brain Educ 2007 Mar;1(1):3-10. [doi: 10.1111/j.1751-228X.2007.00004.x]
45. Wainwright T, Lind G. How to teach morality: promoting deliberation and discussion, reducing violence and deceit. Clin Psychol Forum 2017 May;1(293):51. [doi: 10.53841/bpscpf.2017.1.293.51]
46. International center for sustainable and just transformation [tra:ce] [Article in German]. Universität Witten/Herdecke. 2025 Jun 6. URL: https://www.uni-wh.de/ueber-uns/institute-zentren-und-ags/international-center-for-sustainable-and-just-transformation-trace#forschungsprojekte [accessed 2026-07-21]
47. Nitsche J, Schmalen S, Kernbeck S, et al. Digitale Planetare Gesundheit als Ringvorlesung: Relevanz für die Teilnehmenden und ihren persönlichen Handabdruck [Abstract] [Article in German]. Presented at: 24 Jahrestagung des Netzwerks Evidenzbasierte Medizin e V; Mar 22-24, 2023. [doi: 10.3205/23ebm038]
48. Widmann K, Busse TS, Ehlers JP, Nitsche J. Transformative learning in digital bioethics education: an interdisciplinary lecture series on human-animal relationships. Animals (Basel) 2025 Nov 25;15(23):3398. [doi: 10.3390/ani15233398] [Medline: 41375457]
49. Widmann K, Busse TS, Ehlers J, Nitsche J. From animal ethics to climate change: integrating human-animal-studies and one health into medical education. One Health 2026 Jun;22:101399. [doi: 10.1016/j.onehlt.2026.101399] [Medline: 41969637]
50. Wheat S, Katzman J, Herring D, Tomedi L, Sood N, Sorensen C. Climate-health tele-education as a force multiplier: a train-the-trainer ECHO course series. J Clim Chang Health 2023 Nov;14:100282. [doi: 10.1016/j.joclim.2023.100282]

XSL•FO
RenderX

Abbreviations

DMgPH: Digital Medicine goes Planetary Health
HCP: health care professional
NEP: New Ecological Paradigm
RUB: Ruhr University Bochum
UW/H: Witten/Herdecke University

Edited by T Gladman; submitted 10.Dec.2025; peer-reviewed by D Dinh, W Lu; revised version received 07.Jul.2026; accepted 12.Jul.2026; published 04.Aug.2026.

Please cite as:
Nitsche J, Busse TS, Schmalen S, Otte I, Ehlers JP
Attitudes Toward Digital Planetary Health Before and After a Lecture Series on Environmental Awareness and Future Visions: Repeated Cross-Sectional Survey Study Regarding Twin Transformation
JMIR Med Educ 2026;12:e89259
URL: https://mededu.jmir.org/2026/1/e89259
doi:10.2196/89259

© Julia Nitsche, Theresa Sophie Busse, Sonja Schmalen, Ina Otte, Jan Peter Ehlers. Originally published in JMIR Medical Education (https://mededu.jmir.org), 4.Aug.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Education, is properly cited. The complete bibliographic information, a link to the original publication on https://mededu.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

The Historical Evolution of Microlearning in Health Professions Education: Bibliometric Analysis

Jennie C De Gagne[1], PhD; Paige S Randall[2], PhD; Elizabeth R Blackwood[3], MSLS; Katherine A Carlson[3], MLiS; Hee Sun Kang[4], PhD

[1]School of Nursing, Duke University, Durham, NC, United States

[2]School of Nursing, University of North Carolina, Chapel Hill, NC, United States

[3]Medical Center Library & Archives, Duke University, Durham, NC, United States

[4]Red Cross College of Nursing, Chung-Ang University, 84, Heukseok-Ro, Dongjak-Gu, Seoul, Republic of Korea

Corresponding Author:
Hee Sun Kang, PhD
Red Cross College of Nursing, Chung-Ang University, 84, Heukseok-Ro, Dongjak-Gu, Seoul, Republic of Korea

Abstract

Background: Microlearning has emerged as a flexible, technology-enabled instructional approach in health professions education, yet its longitudinal evolution, collaboration patterns, and thematic development remain underexplored.

Objective: This study aimed to map publication trends, research impact, collaboration networks, and thematic evolution of microlearning research within health professions education.

Methods: A bibliometric analysis of the Web of Science Core Collection (1977 - 2025) was conducted. Records were retrieved, screened, cleaned, and analyzed using established bibliometric techniques to examine publication trends, coauthorship networks, and keyword co-occurrence patterns.

Results: A total of 560 publications were included in the analysis. Publications demonstrated substantial growth beginning in 2015, with the greatest expansion after 2020. Research output was concentrated in English-speaking regions, particularly the United States. Coauthorship networks showed moderate collaboration within several established academic hubs. Keyword analyses revealed 5 thematic domains reflecting the field's progression from instructional efficiency toward mobile, social, simulation-enhanced, and crisis-responsive microlearning. These themes suggest increasing attention to experiential, accessible, and adaptive learning approaches within health professions education.

Conclusions: Microlearning has evolved into a multidimensional scholarly area shaped by digital transformation, shifting pedagogical expectations, and increased demand for flexible learning formats. Ongoing development should prioritize global representation, theoretical grounding, and rigorous evaluation of emerging technologies. The findings may inform educators, curriculum designers, and institutional leaders seeking to implement evidence-based, digitally enabled microlearning strategies that support learners across diverse educational and clinical environments.

(JMIR Med Educ 2026;12:e91616) doi:10.2196/91616

KEYWORDS

artificial intelligence; AI; bibliometrics; health professions education; microlearning; mobile learning

Introduction

The integration of technology into health professions education (HPE) has fundamentally reshaped how knowledge is delivered and acquired, aligning with broader transitions toward flexible, learner-centered educational paradigms [1,2]. Within this digital transformation, microlearning—characterized by concise, focused, technology-mediated content—has emerged as an effective strategy to enrich HPE [3,4]. Defined as "bite-sized," "just-in-time," or "on-demand" learning, microlearning leverages digital platforms to provide efficient, self-directed training for students, educators, and practitioners, meeting the field's growing demand for rapid knowledge updates and sustained clinical competency [3]. Artificial intelligence (AI) further enhances the feasibility of microlearning by streamlining design, development, and implementation processes [5], positioning microlearning as a dynamic and scalable component of online learning to foster positive cognitive, behavioral, and affective outcomes [6].

Despite growing interest, the scope, evolution, and scholarly influence of microlearning in HPE remain underexamined. A foundational scoping review by De Gagne et al [4] demonstrated microlearning's potential across multiple health professions. That review highlighted microlearning's role in improving learner engagement and retention through succinct, interactive formats, particularly when integrated with digital or adaptive

XSL•FO
RenderX

technologies. Cronin and Durham [3] further conceptualized microlearning as a "self-directed, on-demand teaching strategy" characterized by focus, brevity, and multimodal delivery, although they emphasized the absence of definitional consistency across studies.

Bibliometric investigations in HPE have increasingly mapped research on technology-enhanced education, including online learning [7-9], mobile learning [10], and simulation [11-14]. Simulation-based learning has received significant scholarly attention [15], and social media–enabled learning is gaining traction as an engagement-driven modality [16]. Within this expanding digital landscape, microlearning appears frequently as an instructional strategy that delivers short, targeted content across formats such as videos, quizzes, podcasts, and social media clips. However, existing reviews are largely narrative or scoping in nature and often limited in focus. For example, one systematic review evaluated simulation-based microlearning for health care professionals [17], another focused on mobile microlearning in nursing professional development [18], and a literature review examined microlearning in combination with social media across select disciplines [19]. These studies demonstrate microlearning's versatility and growing relevance, yet they do not provide a comprehensive, quantitative overview of its development, adoption patterns, and intellectual structure within HPE.

Given the rapid expansion of microlearning-related methods, there is a need to systematically map its scholarly trajectory. Bibliometric analysis offers a quantitative approach for examining research productivity, citation impact, collaborative structures, and thematic evolution over time [20]. Accordingly, this study conducted a bibliometric analysis of microlearning in HPE to (1) analyze publication trends and citation impact (eg, annual output, journal distribution, geographic output, and influential references); (2) examine collaboration networks at the author, institutional, and national level; and (3) explore keyword co-occurrences to identify prevailing themes and emerging areas of focus.

Methods

Overview

This study used a longitudinal bibliometric analysis to examine the evolution and characteristics of microlearning within HPE. Bibliometric analysis enables quantitative evaluation of scholarly activity by generating publication, citation, and collaboration metrics through specialized software [21]. The study protocol was registered on the Open Science Framework to promote transparency and reproducibility [22].

Ethical Considerations

Ethics approval was not required as only publicly available bibliometric data were analyzed.

Data Collection and Cleaning

Data were retrieved from the Web of Science (WoS) Core Collection due to its comprehensive coverage of high-impact scholarly publications. A search strategy developed by professional medical librarians in collaboration with the author team was executed on June 11, 2025. Search filters were applied in the WoS user interface to limit results to the document types "article" and "review article," yielding 9552 records. The search incorporated terms related to HPE, specific health professions, and microlearning. To ensure comprehensive retrieval, we operationalized microlearning and microlearning-adjacent terms (eg, "microteaching," "nano-learning," and "just-in-time learning") as instructional approaches involving brief, focused learning units often delivered through digital or modular formats. These terms were treated as conceptually adjacent or antecedent instructional approaches rather than direct equivalents of contemporary microlearning. Studies using any of these terms were included only when the described intervention aligned with our operational definition of microlearning and met all other eligibility criteria. The complete search strategy is available in Table S1 in Multimedia Appendix 1.

Records were imported into Covidence (Veritas Health Innovation) [23] for deduplication and title and abstract screening. Screening was conducted independently by 2 reviewers using predefined eligibility criteria (Table 1), with discrepancies resolved by a third reviewer. Microlearning alignment was determined during title and abstract screening based on evidence that the instructional approach was brief, focused, and delivered in modular or segmented formats consistent with microlearning principles. Because this study was designed as a bibliometric analysis, screening focused on relevance and eligibility based on bibliographic metadata, titles, abstracts, and indexed record information instead of the conventional full-text appraisal used in evidence synthesis reviews. To restore any bibliometric metadata not retained during Covidence import, full citation details were re-extracted from WoS using accession numbers. Data were then imported into OpenRefine [24] for metadata standardization. Author names, institutional affiliations, and keywords were cleaned and consolidated to prevent duplication and ensure analytical consistency. No additional records were excluded during this phase; however, document types were verified and corrected as necessary. The final dataset comprised 560 studies for analysis. A flowchart based on the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) guidelines [25] is provided in Figure S1 in Multimedia Appendix 1. The eligibility criteria table presents the full set of prespecified inclusion and exclusion criteria, whereas the flow diagram summarizes the major exclusion categories applied during title and abstract screening. Individual exclusion reasons were not itemized for each excluded record.

XSL•FO
RenderX

Table . Eligibility criteria.

Domain	Inclusion criteria	Exclusion criteria
Population	Students, trainees, or licensed professionals engaged in HPE[a] (eg, nursing, medicine, pharmacy, public health, allied health, dentistry, physical therapy, and occupational therapy)	Populations not relevant to microlearning in HPE: studies focused on patient or caregiver education or other nonlearner populations, health-related training outside formal educational or professional development settings (eg, community outreach or corporate wellness), and veterinary education
Intervention or exposure	Studies focused on microlearning or microlearning-adjacent instructional strategies (eg, bite-sized learning, just-in-time learning, mobile microcontent, or nano-learning)	Studies lacking a clear microlearning or microlearning-adjacent instructional focus
Context	HPE settings, including academic, clinical, and continuing professional development environments	—[b]
Study characteristics	Scholarly publications retrieved from WoS[c] and initially filtered as "article" or "review article" published in any language and year	Non–peer-reviewed publication types (eg, editorials, letters, or commentaries), study protocols, or preregistrations

[a]HPE: health professions education.

[b]Not applicable.

[c]WoS: Web of Science.

Data Analysis

Bibliometric analyses were conducted using Biblioshiny (K-Synth Srl) [26] and VOSviewer (Centre for Science and Technology Studies, Leiden University) [27] to evaluate publication trends, citation impact, geographic output, collaboration networks, and thematic structures. Descriptive statistics, including annual publication counts and citation summaries, were compiled in Microsoft Excel for supplementary reporting. Biblioshiny, a web-based interface of the *bibliometrix* R package [26], facilitated descriptive analyses of publication trajectories, citation metrics, global contributions, influential authors and institutions, and cocitation networks to reveal the intellectual structure of microlearning research in HPE. In VOSviewer, collaboration networks were visualized at the author, institutional, and national levels, and keyword co-occurrence analysis using author-assigned keywords was conducted to identify thematic clusters and emerging areas of focus [28,29]. Full counting and association strength normalization were used to construct the maps. Thresholds were applied to reduce network complexity and improve interpretability, with minimum occurrence and document thresholds set according to the unit of analysis, and only connected items were retained in the final visualizations. The resulting maps displayed cluster membership, number of links, and total link strength (TLS) for included items. Entity labels in VOSviewer figures reflect the indexed source metadata and software parsing conventions used at the time of analysis. Additional analytical details are provided in Table S2 in Multimedia Appendix 1.

Results

Publication Trends and Citation Impact

The final dataset comprised 560 publications indexed between 1977 and 2025 authored by 2604 unique contributors and citing 13,229 references, with an average of 4.86 (SD 2.80) coauthors per publication. Nearly all publications (n=549, >98%) were in English, with only a small number in other languages. Most documents (n=514, 91.8%) were full research articles, followed by review articles (n=41, 7.3%), proceedings papers (n=4, 0.7%), and book chapters (n=1, 0.2%). Although the WoS filter was limited to "article" and "review article," document type metadata in the final dataset included a small number of records labeled as proceedings papers or a book chapter following export and metadata verification. Metadata completeness was high, although inconsistencies in author keywords (n=115, 20.5%) and KeyWords Plus (n=136, 24.3%) warrant caution when interpreting keyword-based trends (Table 2).

Publication activity was minimal prior to 2005, after which output increased gradually until 2014 (1 - 11 publications per year). A stronger upward trajectory was observed between 2015 and 2019 (16 - 29 publications annually), followed by a marked surge from 2020 onward, peaking at 93 publications in 2024. As of the search date, 7.3% (41/560) of the publications had already been indexed in 2025, indicating sustained momentum. The calculated annual growth rate was 8.04%. Citation concentrations appeared to cluster around publications from years with higher output, although these patterns should be interpreted cautiously because citation counts are cumulative and time dependent.

XSL•FO
RenderX

Table . Summary of dataset characteristics from Biblioshiny.

Category	Values
Total documents, n	560
Time span	1977-2025
Sources (eg, journals), n	266
Annual growth rate (%)	8.04
Document age (y), mean (SD)	4.81 (4.69)
References cited, n	13,229
Authors, n	2604
Single-authored documents, n	31
Coauthors per document, mean (SD)	4.86
Keywords (DE[a])[b], n	1312
KeyWords Plus (ID)[c], n	638

[a]DE: descriptors (author keywords).

[b]Provided by study authors.

[c]Generated/indexed keywords from databases.

Journal and Source Distribution

Among the 560 publications, 514 (91.8%) research articles were distributed across 266 journals, indicating broad and growing interest in microlearning across diverse HPE contexts. The most productive journal was *BMC Medical Education* (43/514, 8.4%), followed by *Anatomical Sciences Education* (17/514, 3.3%), *Nurse Education Today* (15/514, 2.9%), *Medical Science Educator* (13/514, 2.5%), and *Nurse Education in Practice* (11/514, 2.1%). Three journals—*Journal of Dental Education*, *Journal of Nursing Education*, and *Journal of Surgical Education*—contributed 1.9% (10/514) of the publications each. Additional frequently publishing journals included *Currents in Pharmacy Teaching and Learning* (9/514, 1.8%), *Medical Teacher* (9/514, 1.8%), *Nurse Educator* (9/514, 1.8%), and *Academic Medicine* (8/514, 1.6%). Citation trends mirrored these publication patterns, with the most frequently cited articles often appearing in the journals and years with peak research activity.

Geographic and Institutional Productivity

Microlearning research demonstrated broad international representation, although output was concentrated in high-resource regions. The United States was the leading contributor (219/560, 39.1%), followed by the United Kingdom (46/560, 8.2%), China (38/560, 6.8%), Canada (32/560, 5.7%), and India (28/560, 5%). Additional contributions from multiple regions suggest expanding global engagement with microlearning in digital and flexible health professions training environments. Institutional productivity was similarly concentrated among leading academic centers. Harvard University (43/560, 7.7%), the University of Toronto (40/560, 7.1%), and the University System of Ohio (35/560, 6.3%) were the most prolific, followed by the University of California system (22/560, 3.9%) and Johns Hopkins University (21/560, 3.8%). This pattern indicates that established academic hubs have driven early scholarship, with increasing participation emerging worldwide across additional institutions.

Intellectual Influence: Most Cited References

Citation impact analysis identified seminal works that shaped early discourse on microlearning and digitally mediated learning. The most frequently cited reference was the work by De Gagne et al [4] (n=36 citations), a scoping review synthesizing microlearning modalities and outcomes. Cho et al [30] (n=29 citations) examined the educational value of podcasts, whereas Mallin et al [31] (n=23 citations) documented the adoption of asynchronous digital learning resources in medical residency education. Additional influential publications included those by Cheston et al [32] and Riddell et al [33] (n=22 citations each), which reinforced the role of mobile, social, and podcast-based learning. Collectively, these works demonstrated that early scholarship positioned microlearning within broader digital, mobile, and networked learning ecosystems.

Coauthorship Networks: Research Collaboration Networks

Overview

Coauthorship networks were mapped using VOSviewer with full counting and association strength normalization. All networks were generated directly from the study dataset and reflect bibliometric relationships within the included publications. Thresholds were applied, and only connected items were retained in the final visualizations.

Author-Level Collaboration

Authors with at least one document were eligible for inclusion in the coauthorship analysis. Thresholds were applied, and only connected items were retained in the final visualization, resulting in 32 authors. Three distinct coauthorship clusters emerged, corresponding to major collaborative teams within the field (cluster 1=red; cluster 2=green; cluster 3=blue; Figure S2 in Multimedia Appendix 1). TLS quantified the intensity of

XSL•FO
RenderX

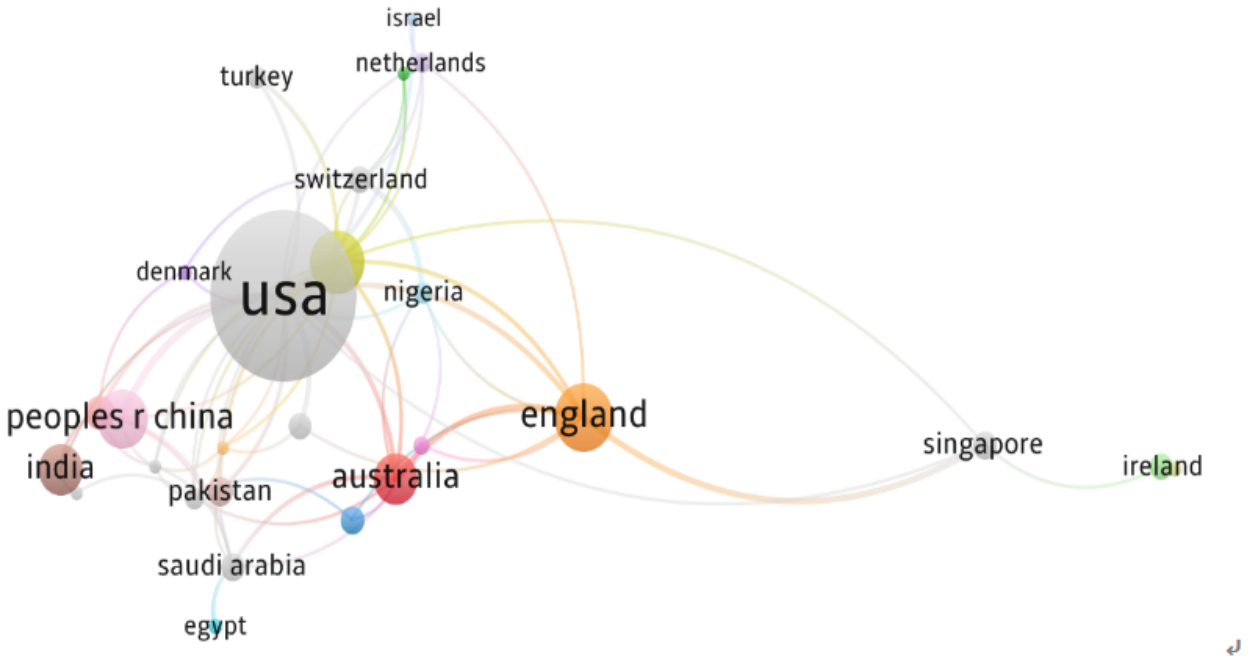
coauthorship connections among authors. Nematollahi exhibited the highest TLS [21], followed by Walker and Nolan (TLS=13 each), serving as key connectors within and between their respective clusters. Overall, collaboration intensity was moderate, with most publications produced by small- to mid-sized author groups rather than large-scale, multi-institutional teams (Figure S2 in Multimedia Appendix 1).

Institutional and National Collaboration

The final institutional coauthorship visualization displayed 39 institutions. Six distributed collaboration clusters emerged, indicating several institution-level partnership groupings within the field. The University of Michigan showed the strongest institutional linkage (TLS=14), followed by Johns Hopkins University; the University of California, San Francisco; and the University of Alabama at Birmingham (TLS=13 each). Institutions such as Emory University (TLS=12) and the University of Maryland (TLS=10) contributed to cross-cluster linkages. The network structure reflects multiple research hubs rather than single-institution dominance (Figure S3 in Multimedia Appendix 1).

At the national level, the final coauthorship visualization displayed 27 countries. The country network showed several interconnected collaboration groupings, with a small number of countries serving as major hubs. The United States served as the central hub (46 links; TLS=239), with strong collaborative ties to Canada, the United Kingdom, Australia, and China. Canada (TLS=28) and the United Kingdom (TLS=18) functioned as secondary anchors bridging networks across North America, Europe, and Asia-Pacific regions. Other countries contributed with lower TLS values (≤3), indicating emerging rather than sustained collaboration. Overall, microlearning research demonstrated a predominantly US-centered network with emerging global participation (Figure 1).

Figure 1. National coauthorship network in microlearning research within health professions education. Node size reflects total publications, and line thickness indicates collaboration intensity (total link strength). Proximity between nodes represents coauthorship frequency, and color clusters denote distinct regional or collaborative groupings. Thresholds were applied to reduce network complexity, and only connected items were retained in the visualization. Country labels in the VOSviewer map reflect source metadata and software-based parsing conventions and, therefore, may differ slightly from the standardized geopolitical naming used in the manuscript text.



Keyword Co-Occurrence and Thematic Trends

Keyword frequency analysis identified recurrent concepts, with “education” (n=157), “social media” (n=94), “medical education” (n=73), “students” (n=56), and “podcast” (n=36) among the most common (Figure 2). The VOSviewer co-occurrence network revealed 7 keyword clusters, which were synthesized into five higher-order thematic domains to enhance interpretability: (1) digital and mobile microlearning in medical education (clusters 1 and 7), (2) social media–driven learning and engagement (cluster 2), (3) pedagogical strategies and assessment (cluster 3), (4) simulation-based skill development in nursing and allied health (cluster 4), and (5) crisis-driven learning adaptation and performance optimization (clusters 5 and 6; Figure 3). Temporal mapping demonstrated an evolution from foundational instructional concepts (eg, lecture, microteaching, and just-in-time learning) toward more digitally integrated modalities (eg, podcasts, simulation, mobile learning, gamification, and TikTok; Figure 4).

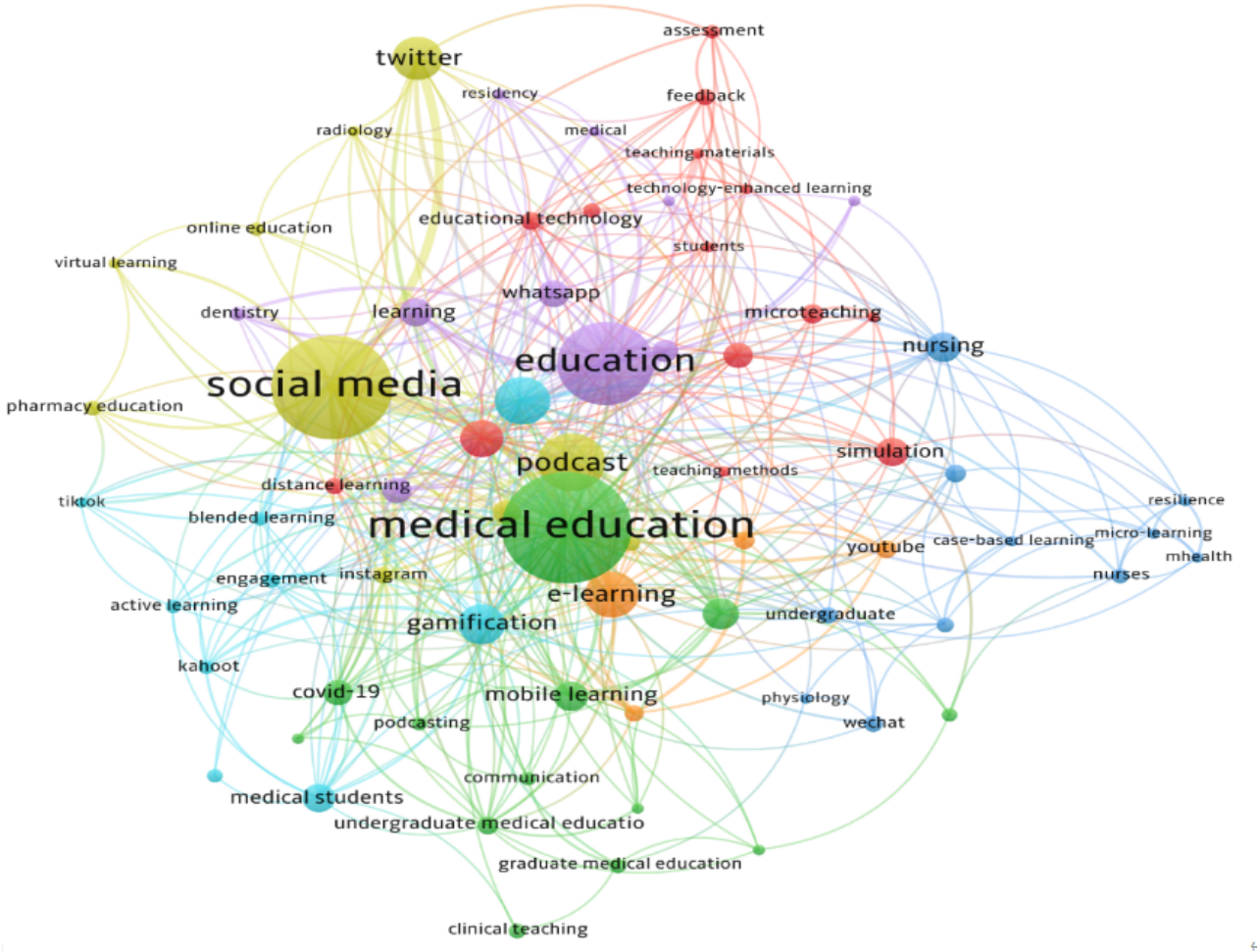
XSL•FO
RenderX

Figure 2. Tree map of the most frequent keywords in the microlearning literature within health professions education. Rectangle size is proportional to keyword frequency, with larger blocks indicating more frequently occurring terms in the dataset.



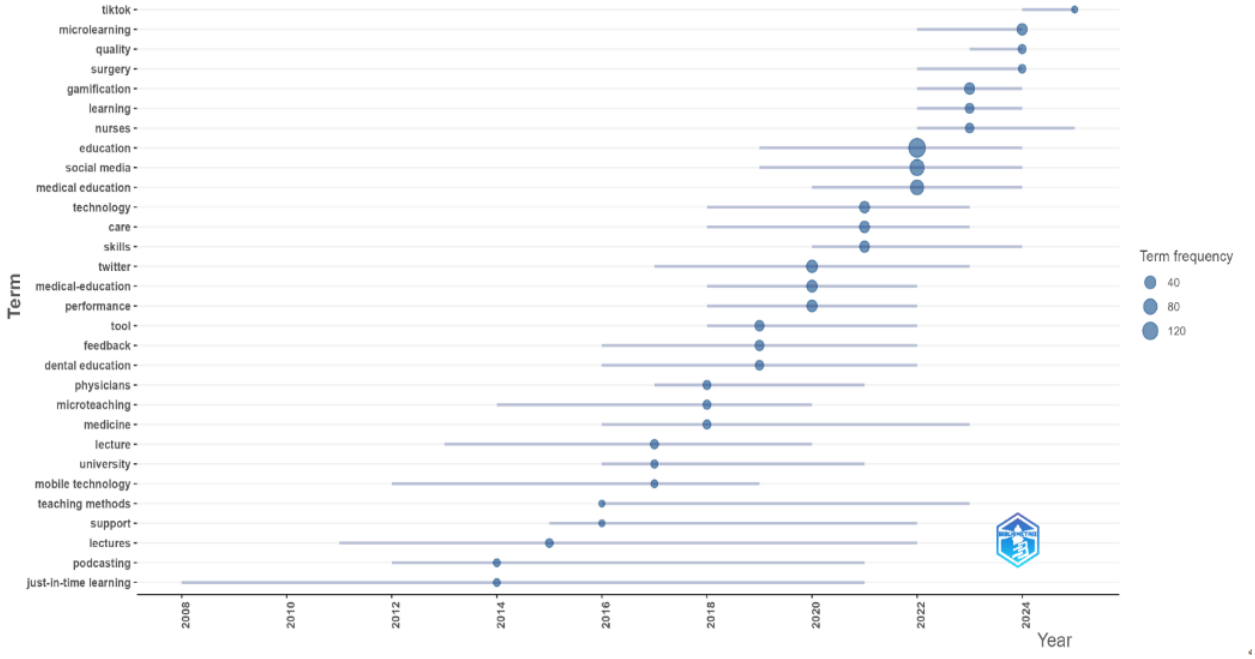
XSL•FO
RenderX

Figure 3. Keyword co-occurrence network map of microlearning research in health professions education. Nodes represent keywords, with larger nodes indicating higher frequency of occurrence. Lines represent co-occurrence relationships, with thicker lines indicating stronger connections. Colors denote the 7 distinct keyword clusters identified by VOSviewer. Thresholds were applied to reduce network complexity, and the final visualization displays 67 connected author keywords, which were subsequently synthesized into 5 higher-order thematic domains for interpretation in the text.



XSL•FO
RenderX

Figure 4. Temporal distribution of trending keywords in microlearning research within health professions education. Each bubble represents a keyword, positioned according to its average year of occurrence in the dataset. Bubble size reflects term frequency, and horizontal lines indicate the span of years during which each term appeared, highlighting the shift from foundational instructional concepts to more digitally integrated learning modalities over time.



Discussion

Growth and Maturation of Microlearning

Microlearning in HPE has evolved significantly over nearly 5 decades, progressing from sporadic experimentation to a widely recognized pedagogical approach. The earliest publications identified in this dataset appeared in 1977 (Crosby [34]) and 1989 (Hargie and Morrow [35]), reflecting early instructional approaches such as microteaching and microtraining that share features with contemporary microlearning. These studies can be interpreted as antecedent or conceptually related forms rather than direct representations of modern microlearning as currently defined in digitally mediated educational contexts. Growth accelerated after 2015 and surged following 2020, reflecting broader shifts toward flexible, competency-based digital learning and the need for rapid, time-efficient knowledge acquisition during pandemic-driven transitions. These patterns suggest that microlearning has become increasingly embedded within contemporary expectations for modular, accessible, and self-directed learning. Because citation counts accumulate over time, comparisons across publication years should be interpreted cautiously and understood as descriptive indicators rather than direct measures of relative scholarly influence.

The dominance of English-language publications and concentration of research in high-resource regions indicate that knowledge production remains unevenly distributed. While this pattern likely reflects existing academic infrastructures for digital pedagogy, it also points to inequities in representation and visibility for low- and middle-income contexts, where innovation may be occurring but less frequently indexed. Increasing citation activity during peak publication years may suggest growing scholarly attention and thematic convergence within the field, although bibliometric indicators alone cannot confirm scholarly consolidation. Foundational work on digitally mediated learning appears to have supported subsequent research on the design, implementation, and evaluation of microlearning across diverse HPE settings. Collectively, these patterns suggest that microlearning is becoming an increasingly established area of inquiry shaped by evolving technological ecosystems and the demand for scalable, adaptive instructional strategies.

Collaboration Patterns and Knowledge Flow

Collaboration patterns provide insights into how microlearning knowledge is generated and disseminated worldwide. The author-level network demonstrated moderate collaboration intensity, indicating that research is largely conducted within small, localized teams rather than large, internationally integrated consortia, similar to recent literature [7,36]. A limited number of recurring authors served as central connectors, shaping conceptual trajectories, supporting methodological consistency, and facilitating knowledge diffusion across groups.

Institutional collaboration patterns reflect a distributed, multi-hub research structure rather than dominance by a single academic center. Several well-established North American institutions functioned as collaborative anchors, with additional but more modest engagement from European universities. This distribution suggests both a regional concentration of leadership and the gradual development of research capacity within secondary academic settings.

At the national level, the United States emerged as the central hub of global collaboration, with the United Kingdom, Canada, Australia, and China functioning as secondary contributors. Emerging participation from regions such as the Middle East, South Asia, and parts of Europe suggests increasing geographic diversification. As the field continues to mature, broader inclusion of diverse institutional and cultural perspectives may

XSL•FO
RenderX

enhance contextual relevance and expand the adaptability of microlearning models across varied educational environments.

Thematic Evolution and Pedagogical Directions

The thematic progression of microlearning research indicates a clear pedagogical evolution within HPE. Early clusters associated with instructional efficiency suggest that microlearning initially emerged as a strategy to fragment content delivery for rapid recall and time-constrained knowledge acquisition. As mobile and digital platforms gained prominence, the focus shifted toward learner autonomy, self-regulation, and continuous access, reflecting broader trends in competency-based and personalized education [37].

A subsequent phase emphasized social media–driven microlearning, in which learning appeared increasingly peer mediated, networked, and embedded in informal digital ecosystems. This pattern may reflect growing attention to learner engagement, identity formation, and community participation as potentially important dimensions of microlearning. More recently, crisis-responsive and resilience-oriented themes appear more frequently in the literature, particularly during the COVID-19 pandemic [38,39]. Recent keyword patterns also indicate growing attention to themes such as resilience and emotional regulation, although these observations should be interpreted as descriptive patterns in term use rather than definitive evidence of thematic shift [4,40]. Collectively, these patterns suggest a broadening of microlearning scholarship beyond content reduction toward a more multidimensional pedagogical framing that includes accessibility, engagement, adaptability, and affective support.

The thematic domains identified in this study relate to established learning theories or frameworks. For example, digital and mobile microlearning delivers short, focused information segments that may minimize cognitive load, aligning with cognitive load theory. Social media–driven learning may enhance autonomy, allowing learners to engage in and regulate their learning while interacting with peers, consistent with self-regulated learning principles. In addition, simulation-based microlearning may support confidence development by enabling learners to practice tasks in simulated scenarios. Crisis-driven learning may reflect principles of situated learning by supporting knowledge development through contextually grounded participation in authentic or practice-relevant environments. These approaches may also provide contextualized practice opportunities for skill development aligned with competency-based education frameworks [6,41-44]. These connections suggest that microlearning functions not only as a technological format but also as a pedagogically grounded learning approach.

Implications and Future Directions

The maturation of microlearning in HPE presents key implications for curriculum design, faculty development, and learner engagement. As competency-based and outcome-driven models increasingly prioritize rapid knowledge acquisition and performance reinforcement, microlearning offers modular, time-efficient formats that can be integrated across didactic, simulation-based, and workplace learning [45]. In competency-based pathways, microlearning may support adaptive progression by targeting individual performance gaps and sequencing content responsively.

The rise of mobile, social, and simulation-enhanced microlearning underscores its utility for experiential, just-in-time learning beyond classroom settings. Scenario-based formats can reinforce clinical reasoning and decision-making while supporting continuing professional development and rapid dissemination of practice updates [17]. However, meaningful integration requires pedagogical alignment; poorly structured microcontent risks fragmented knowledge acquisition rather than deep learning. Therefore, faculty development must extend beyond content creation to include sequencing, scaffolding, reflection, and alignment with assessment and competency frameworks. Moreover, as socially driven and resilience-oriented microlearning increases, concerns emerge regarding digital credibility, professionalism, cognitive overload, and learner well-being [40]. Thus, microlearning should be viewed not only as an instructional method but also as part of broader discussions on digital citizenship and how learners construct professional identity within networked environments.

Research on AI-driven adaptive microlearning represents a key frontier. Generative AI may enable dynamic scenario generation, real-time feedback, and performance-based content sequencing. Investigating its effectiveness, cognitive implications, and ethical risks will be crucial as adoption increases [46]. Similarly, immersive microlearning through virtual or augmented reality warrants exploration, particularly for supporting emotional regulation, identity formation, and resilience in demanding clinical environments. Finally, studies examining faculty readiness, institutional integration, and sustainability will support scalable adoption. Cross-disciplinary collaboration among educators, technologists, cognitive scientists, and AI experts will be essential to fully realize microlearning's potential as a transformative and equitable educational modality in HPE.

These findings have implications across multiple levels of HPE. At the curricular level, microlearning can be incorporated into existing courses through short, focused learning modules (eg, brief videos, quizzes, or case-based microcontent) to support learner engagement and knowledge retention [4,47]. In faculty development, educators may benefit from training in designing and implementing microlearning activities and in using digital technologies effectively [48]. At the institutional level, leaders can support implementation by providing microlearning platforms or integrating them into existing learning management systems while also evaluating learning outcomes and supporting continuous improvement based on feedback.

Strengths and Limitations

This study's major strength lies in its comprehensive and systematically executed search strategy, which initially identified nearly 10,000 records through the WoS Core Collection. Collaboration with professional medical librarians ensured methodological rigor, transparent inclusion criteria, and reproducibility throughout the review process.

Several limitations should be acknowledged. Reliance on a single database may have restricted the retrieval of relevant

XSL•FO
RenderX

studies indexed elsewhere, particularly those outside English-language or biomedical education outlets. Although the inclusion criteria did not impose language restrictions, the final dataset was overwhelmingly in English, reflecting both indexing practices and broader publication inequities. Other major databases offer complementary coverage. For instance, Scopus provides broader international journal representation, PubMed extensively indexes biomedical literature, and ERIC focuses on education-related publications. The exclusion of these databases may have influenced geographic representation, disciplinary breadth, and the inclusion of education-focused journals not indexed in WoS.

Additionally, metadata inconsistencies in author keywords and KeyWords Plus fields, which are common in bibliometric datasets, may limit the interpretive depth of keyword co-occurrence analyses. Potential inconsistencies between author keywords and KeyWords Plus may have influenced theme emergence, clustering stability, and the interpretation of temporal trends. For example, evolving topics may have been labeled differently over time, which could affect how related studies were grouped and how thematic shifts were interpreted. The inclusion of microlearning-adjacent or antecedent terms (eg, “microteaching” and “microtraining”) may also have influenced the historical framing of the field by capturing earlier instructional approaches that share structural similarities with contemporary microlearning but differ in technological context and pedagogical framing. A sensitivity check excluding the 2 antecedent records from 1977 and 1989 showed no meaningful changes in keyword co-occurrence patterns or thematic interpretation, supporting the robustness of the findings. The search strategy also included a broad set of mobile- and social media–related instructional terms to ensure comprehensive retrieval. Although screening retained only studies aligned with the operational definition of microlearning, this approach may have broadened the thematic scope of the dataset, contributing to the prominence of social media–related keywords in the results. These factors should be considered when interpreting the findings. In addition, the bibliometric approach captures quantitative trends rather than the pedagogical quality or contextual nuance of the included studies, underscoring the need for complementary qualitative and theoretical reviews.

Conclusions

Microlearning in HPE has evolved from a fragmented instructional strategy into a multidimensional and increasingly globalized area of inquiry. This bibliometric analysis demonstrates a rapidly expanding field shaped by digital transformation; collaborative networks; and shifting pedagogical priorities aligned with mobility, social engagement, emotional resilience, and adaptive learning. As microlearning matures, future research should prioritize theoretical integration, global inclusivity, faculty readiness, and the pedagogical potential of AI-driven adaptive technologies. These directions will be essential for sustaining microlearning’s impact in preparing current and future health professionals for dynamic, digitally mediated clinical environments.

Funding

The authors declared no financial support was received for this work.

Data Availability

All data generated or analyzed during this study are included in this published article (and its supplementary information files).

Authors' Contributions

JCDG contributed to conceptualization, methodology, investigation, formal analysis, visualization, writing the original draft, review and editing, and supervision. HSK contributed to conceptualization, writing the original draft, and review and editing. ERB contributed to methodology, data curation, resources, and review and editing. KAC contributed to conceptualization, methodology, data curation, and review and editing. PSR contributed to conceptualization, investigation, formal analysis, project administration, and review and editing. All authors read and approved the final manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Search strategy, bibliometric procedures, and coauthorship networks.
[DOCX File, 704 KB - mededu_v12i1e91616_app1.docx]

References

1. Baako I, Abroampa WK. Research trends on ICT integration in education: a bibliometric analysis. Cogent Educ 2023;10(2):2281162. [doi: 10.1080/2331186X.2023.2281162]
2. De Gagne JC. Teaching in online learning environments. In: De Gagne JC, Oermann MH, editors. Teaching in Nursing and Role of the Educator: The Complete Guide to Best Practice in Teaching, Evaluation, and Curriculum Development, 4th edition: Springer; 2024:143-162. [doi: 10.1891/9780826188922]

XSL•FO
RenderX

3. Cronin J, Durham ML. Microlearning: a concept analysis. Comput Inform Nurs 2024 Jun 1;42(6):413-420. [doi: 10.1097/CIN.0000000000001122] [Medline: 38483251]
4. De Gagne JC, Park HK, Hall K, Woodward A, Yamane S, Kim SS. Microlearning in health professions education: scoping review. JMIR Med Educ 2019 Jul 23;5(2):e13997. [doi: 10.2196/13997] [Medline: 31339105]
5. De Gagne JC. Integrating technology in education. In: Oermann MH, De Gagne JC, editors. Teaching in Nursing and Role of the Educator: The Complete Guide to Best Practice in Teaching, Evaluation, and Curriculum Development, 4th edition: Springer; 2025:123-141. [doi: 10.1891/9780826188922]
6. Monib WK, Qazi A, Apong RA. Microlearning beyond boundaries: a systematic review and a novel framework for improving learning outcomes. Heliyon 2024;11(2):e41413. [doi: 10.1016/j.heliyon.2024.e41413] [Medline: 39882484]
7. De Gagne JC, Randall PS, Koppel PD, Cho E, Blackwood ER, Kang HS. Online learning in nursing education: a 21st century bibliometric analysis. Nurse Educ Today 2025 Aug;151:106740. [doi: 10.1016/j.nedt.2025.106740] [Medline: 40222324]
8. Oluwadele D, Singh Y, Adeliyi TT. E-learning performance evaluation in medical education-a bibliometric and visualization analysis. Healthcare (Basel) 2023 Jan 12;11(2):232. [doi: 10.3390/healthcare11020232] [Medline: 36673600]
9. Zhang M, Xiao G. Bibliometric analysis of online learning emotions from 2006 to 2023. Int J Emerg Technol Learn 2023;18(13):220-233. [doi: 10.3991/ijet.v18i13.39919]
10. Yalcinkaya T, Cinar Yucel S. Mobile learning in nursing education: a bibliometric analysis and visualization. Nurse Educ Pract 2023 Aug;71:103714. [doi: 10.1016/j.nepr.2023.103714] [Medline: 37552905]
11. Ba H, Zhang L, He X, Li S. Knowledge mapping and global trends in simulation in medical education: bibliometric and visual analysis. JMIR Med Educ 2025 Mar 26;11:e71844. [doi: 10.2196/71844] [Medline: 40139212]
12. Korkmaz EK, Ateş E. Standardized patient simulation in nursing education: a bibliometric analysis and visualization from 2002 to 2024. Nurse Educ Pract 2025 Mar;84:104322. [doi: 10.1016/j.nepr.2025.104322] [Medline: 40058329]
13. Mun M, Kim M, Woo K. Advancements in simulation-based nursing education: insights from a bibliometric analysis of temporal trends. Nurse Educ Today 2025 Aug;151:106719. [doi: 10.1016/j.nedt.2025.106719]
14. Wang Y, Li X, Liu Y, Shi B. Mapping the research hotspots and theme trends of simulation in nursing education: a bibliometric analysis from 2005 to 2019. Nurse Educ Today 2022 Sep;116:105426. [doi: 10.1016/j.nedt.2022.105426] [Medline: 35679627]
15. Oermann MH, Bailey HE, Carter-Templeton H, Condy C, Wrigley J, Ledbetter LS. Bibliometric analysis of nursing education reviews. Nurse Educ Today 2024 Aug;139:106217. [doi: 10.1016/j.nedt.2024.106217] [Medline: 38676961]
16. O'Connor S, De Gagne JC, Lee JJ, Harris R, Malone M, Booth RG. Social media in nursing and midwifery education: a 20-year bibliometric analysis. Nurse Educ Pract 2025 Jul;86:104411. [doi: 10.1016/j.nepr.2025.104411] [Medline: 40451061]
17. Patocka C, Pandya A, Brennan E, et al. The impact of just-in-time simulation training for healthcare professionals on learning and performance outcomes: a systematic review. Simul Healthc 2024 Jan 1;19(1S):S32-S40. [doi: 10.1097/SIH.0000000000000764] [Medline: 38240616]
18. Hilsmann N, Dodson C. Mobile microlearning in continuing professional development for nursing: a scoping review. J Contin Educ Nurs 2025 Feb;56(2):53-62. [doi: 10.3928/00220124-20250121-05] [Medline: 39880017]
19. Denojean-Mairet M, López-Pernas S, Agbo FJ, Tedre M. A literature review on the integration of microlearning and social media. Smart Learn Environ 2024;11:46. [doi: 10.1186/s40561-024-00334-5]
20. Ellegaard O, Wallin JA. The bibliometric analysis of scholarly production: how great is the impact? Scientometrics 2015;105(3):1809-1831. [doi: 10.1007/s11192-015-1645-z] [Medline: 26594073]
21. Kokol P, Blažun Vošner H. Historical, descriptive and exploratory analysis of application of bibliometrics in nursing research. Nurs Outlook 2019;67(6):680-695. [doi: 10.1016/j.outlook.2019.04.009] [Medline: 31204025]
22. Microlearning in health professions education: bibliometric analysis. Open Science Framework. 2025. URL: https://osf.io/eg4hc/overview [accessed 2026-05-28]
23. Covidence. URL: https://www.covidence.org/ [accessed 2025-06-11]
24. OpenRefine. URL: https://openrefine.org [accessed 2025-06-11]
25. Page MJ, McKenzie JE, Bossuyt PM, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ 2021 Mar 29;372:n71. [doi: 10.1136/bmj.n71] [Medline: 33782057]
26. Aria M, Cuccurullo C. bibliometrix: an R-tool for comprehensive science mapping analysis. J Informetr 2017 Nov;11(4):959-975. [doi: 10.1016/j.joi.2017.08.007]
27. VOSviewer. Leiden University. URL: https://www.vosviewer.com [accessed 2025-06-11]
28. Van Raan AF. Advanced bibliometric methods in the analysis of research performance and scientific developments: a contribution to science policy in transition countries. In: Dyker DA, Radosevic S, editors. Innovation and Structural Change in Post-Socialist Countries: A Quantitative Approach: Springer; 1999:89-105. [doi: 10.1007/978-94-011-4463-6_5]
29. van Eck NJ, Waltman L. Software survey: VOSviewer, a computer program for bibliometric mapping. Scientometrics 2010 Aug;84(2):523-538. [doi: 10.1007/s11192-009-0146-3] [Medline: 20585380]
30. Cho D, Cosimini M, Espinoza J. Podcasting in medical education: a review of the literature. Korean J Med Educ 2017 Dec;29(4):229-239. [doi: 10.3946/kjme.2017.69] [Medline: 29207454]

XSL•FO
RenderX

31. Mallin M, Schlein S, Doctor S, Stroud S, Dawson M, Fix M. A survey of the current utilization of asynchronous education among emergency medicine residents in the United States. Acad Med 2014 Apr;89(4):598-601. [doi: 10.1097/ACM.0000000000000170] [Medline: 24556776]
32. Cheston CC, Flickinger TE, Chisolm MS. Social media use in medical education: a systematic review. Acad Med 2013 Jun;88(6):893-901. [doi: 10.1097/ACM.0b013e31828ffc23] [Medline: 23619071]
33. Riddell JC, Robins L, Sherbino J, Brown A, Ilgen J. Residents' perceptions of effective features of educational podcasts. West J Emerg Med 2020 Dec 10;22(1):26-32. [doi: 10.5811/westjem.2020.10.49135] [Medline: 33439799]
34. Crosby MH. Teaching strategies: a microteaching project for nurses in Virginia. Nurs Res 1977;26(2):144-147. [doi: 10.1097/00006199-197703000-00021] [Medline: 584379]
35. Hargie OD, Morrow NC. The effectiveness of microtraining in developing pharmacists communication skills: a study of personality and attitudes. Med Teach 1989;11(2):195-203. [doi: 10.3109/01421598909146323]
36. Djeki E, Dégila J, Bondiombouy C, Alhassan MH. E-learning bibliometric analysis from 2015 to 2020. J Comput Educ 2022;9(4):727-754. [doi: 10.1007/s40692-021-00218-4]
37. Fidan M. The effects of microlearning-supported flipped classroom on pre-service teachers' learning performance, motivation and engagement. Educ Inf Technol 2023;28(10):12687-12714. [doi: 10.1007/s10639-023-11639-2]
38. Boutros P, Kassem N, Nieder J, et al. Education and training adaptations for health workers during the COVID-19 pandemic: a scoping review of lessons learned and innovations. Healthcare (Basel) 2023 Nov 4;11(21):2902. [doi: 10.3390/healthcare11212902] [Medline: 37958046]
39. Levidze M. Mapping the research landscape: a bibliometric analysis of e-learning during the COVID-19 pandemic. Heliyon 2024;10(13):e33875. [doi: 10.1016/j.heliyon.2024.e33875] [Medline: 39027537]
40. Abbasalizadeh M, Farsi Z, Sajadi SA, Atashi A, Fournier A. The effect of resilience training with mHealth application based on micro-learning method on the stress and anxiety of nurses working in intensive care units: a randomized controlled trial. BMC Med Educ 2024 Apr 24;24(1):442. [doi: 10.1186/s12909-024-05427-w] [Medline: 38658914]
41. Hochstrasser K, Stoddard HA. Use of cognitive load theory to deploy instructional technology for undergraduate medical education: a scoping review. Med Sci Educ 2022;32(2):553-559. [doi: 10.1007/s40670-021-01499-1] [Medline: 35528294]
42. Manning KD, Spicer JO, Golub L, Akbashev M, Klein R. The micro revolution: effect of Bite-Sized Teaching (BST) on learner engagement and learning in postgraduate medical education. BMC Med Educ 2021 Jan 21;21(1):69. [doi: 10.1186/s12909-021-02496-z] [Medline: 33478475]
43. Sweller J, van Merriënboer JJ, Paas F. Cognitive architecture and instructional design: 20 years later. Educ Psychol Rev 2019 Jun;31:261-292. [doi: 10.1007/s10648-019-09465-5]
44. Elendu C, Amaechi DC, Okatta AU, et al. The impact of simulation-based training in medical education: a review. Medicine (Baltimore) 2024 Jul 5;103(27):e38813. [doi: 10.1097/MD.0000000000038813] [Medline: 38968472]
45. Zhang J, West RE. Designing microlearning instruction for professional development through a competency based approach. TechTrends 2020 Mar;64:310-318. [doi: 10.1007/s11528-019-00449-4]
46. Boumalek K, El Mezouary A, Hmedna B, Bakki A. Transforming microlearning with generative AI: current advances and future challenges. In: Lahby M, Maleh Y, Bucchiarone A, Schaeffer SE, editors. General Aspects of Applying Generative AI in Higher Education: Opportunities and Challenges: Springer; 2024:241-262. [doi: 10.1007/978-3-031-65691-0_13]
47. Zarshenas L, Mehrabi M, Karamdar L, Keshavarzi MH, Keshtkaran Z. The effect of micro-learning on learning and self-efficacy of nursing students: an interventional study. BMC Med Educ 2022 Sep 7;22(1):664. [doi: 10.1186/s12909-022-03726-8] [Medline: 36071456]
48. Mitra NK, Htet Htet N, Rao Avupati V, et al. Faculty perceptions of microlearning in health professions education: a mixed method analysis of implementation factors. Int J Med Educ 2025 Oct 31;16:172-180. [doi: 10.5116/ijme.68ff.c63e] [Medline: 41196802]

Abbreviations

AI: artificial intelligence
HPE: health professions education
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
TLS: total link strength
WoS: Web of Science

XSL•FO
RenderX

Edited by J Gentges; submitted 16.Jan.2026; peer-reviewed by J Penaflor, MA Jahid; revised version received 15.Apr.2026; accepted 21.Apr.2026; published 01.Jun.2026.

Please cite as:
De Gagne JC, Randall PS, Blackwood ER, Carlson KA, Kang HS
The Historical Evolution of Microlearning in Health Professions Education: Bibliometric Analysis
JMIR Med Educ 2026;12:e91616
URL: https://mededu.jmir.org/2026/1/e91616
doi:10.2196/91616

© Jennie C De Gagne, Paige S Randall, Elizabeth R Blackwood, Katherine A Carlson, Hee Sun Kang. Originally published in JMIR Medical Education (https://mededu.jmir.org), 1.Jun.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Education, is properly cited. The complete bibliographic information, a link to the original publication on https://mededu.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Digital Training for Lay Health Care Workers' Knowledge and Skills in HIV Index Case Testing: Cluster Randomized Trial

Tapiwa A Tembo[1], PhD; Nora E Rosenberg[2], PhD; Katie Mollan[2,3], PhD; Maria Kim[1,4], MD; Sarah E Rutstein[2,3], MD, PhD; Angella M Mkandawire[1], BA; Mike J Chitani[1], BA; Caroline Kumbuyo[1], BSoc; Duncan Phiri[1], MA; Mtisunge Mphande[1], BSc; Elijah Kavuta[1], BSc; Samuel Chilala[1], BA; Jiayu Wang[2], MPA, MPH; Saeed Ahmed[1,4], MD; Katherine Simon[1,4], MD; Linda-Gail Bekker[5], MD, PhD

[1]Baylor College of Medicine Children's Foundation Malawi, Off Mzimba Road, P/Bag B-397, Lilongwe, Malawi

[2]Gillings School of Global Public Health, University of North Carolina at Chapel Hill, Chapel Hill, NC, United States

[3]School of Medicine, University of North Carolina at Chapel Hill, Chapel Hill, NC, United States

[4]Department of Pediatrics, Global Health Division, Baylor College of Medicine, Houston, TX, United States

[5]The Desmond Tutu HIV Centre, Institute of Infectious Disease and Molecular Medicine, Faculty of Health Sciences, University of Cape Town, Cape Town, Western Cape, South Africa

Corresponding Author:
Tapiwa A Tembo, PhD
Baylor College of Medicine Children's Foundation Malawi, Off Mzimba Road, P/Bag B-397, Lilongwe, Malawi

Abstract

Background: Task shifting in low-resource settings requires lay health care workers (HCWs) to provide a variety of health services, such as HIV index case testing, whereby sexual partners and family members of people living with HIV are offered HIV testing. For this, lay HCWs require adequate specialized training. Digital technologies hold promise for training lay HCWs in low-resource settings, but their impacts on improving knowledge, attitudes, and skills are not understood.

Objective: This study evaluates the impact of digital training on lay HCWs' knowledge, attitudes, and skills to provide HIV index case testing.

Methods: We recruited lay HCWs from 34 health facilities in Malawi. We conducted a 2-arm cluster randomized controlled trial from 2022 to 2023, evaluating the impact of a digital training approach. Health facilities (clusters) were randomized 1:2 to the enhanced or standard arms. Lay HCWs in both arms received the standard in-person index case testing training. In addition, lay HCWs in the enhanced arm received tablet-guided training. Knowledge acquisition was measured using multiple-choice questionnaires administered before and after training. Attitudinal gains were assessed through a questionnaire with Likert scale responses before and after training. Between-arm mean differences were evaluated using generalized estimating equations. Skills (fidelity to index and contact testing protocols) were measured using 15-item checklists. Fidelity scores were compared between the enhanced and standard arms by estimating mean differences and 95% CIs using generalized estimating equations.

Results: We enrolled 306 lay HCWs, with 125 (40.8%) in the enhanced arm and 181 (59.2%) in the standard arm. In the enhanced arm, there was 100% completion of the digital portion, 98% (123/125) completion of the face-to-face tablet-guided portion, and 81% (101/125) to 93% (116/125) completion of quality improvement sessions. Knowledge improved by 4.4% (95% CI 0.7%-8.2%) more in the enhanced arm than in the standard arm (P=.02). Attitudes toward digital training improved by 4.5% (95% CI 0.3%-7.0%) more in the enhanced arm than in the standard arm (P=.03). Lay HCWs' fidelity to index client counseling protocols was 30.5 (95% CI 26.0%-35.0%; P<.001) percentage points higher in the enhanced arm than in the standard arm. Fidelity to contact client counseling protocols was 24.0 (95% CI 20.6%-27.3%; P<.001) percentage points higher in the enhanced arm than in the standard arm.

Conclusions: Digital training improved lay HCWs' knowledge, attitudes, and skills surrounding index case testing counseling. These findings support digital training as a useful strategy for strengthening the capacity of lay HCWs in low-resource contexts.

Trial Registration: ClinicalTrials.gov NCT05343390; https://clinicaltrials.gov/study/NCT05343390

(JMIR Med Educ 2026;12:e89942) doi:10.2196/89942

KEYWORDS

index case testing; implementation fidelity; health care workers; HCWs; lay health workers; blended learning; digital training; e-learning; low- and middle-income countries; LMICs; Malawi

XSL•FO
RenderX

Introduction

The deficit of health professionals in low- and middle-income countries (LMICs) has laid the foundation for task shifting—delegation of medical and health service responsibilities from certified professionals to lay health care workers (HCWs) [1,2]. Lay HCWs are people working at the interface between primary health care and communities to provide services [3]. Lay HCWs have contributed to improved health outcomes in maternal and child health, HIV, and noncommunicable diseases [4-6]. Adequate and regular training facilitates the acquisition of sustained knowledge, attitudes, and skills to provide quality health care services and is needed for lay HCWs to be effective [3,7,8].

Traditional synchronous centralized classroom face-to-face training has been a longstanding method of training HCWs, and it offers numerous benefits, such as face-to-face interaction and immediate feedback [9]. However, many challenges surround its provision, including a shortage of skilled trainers and infrastructure, breaks in care, poor fidelity to training content, and costs [10]. Additionally, one-time face-to-face training does not ensure retention of knowledge and skills over time [11]. Notably, failure of lay HCW programs has often been due to inadequate training and lack of regular refresher training [12,13].

Given advances in digital technology, the World Health Organization recommends digital training to expand access, and digital training may also contribute to the effective training of lay HCWs [1]. Digital training allows learning to transcend boundaries of space and time [4], limiting the time health workers spend away from work while still providing knowledge and skills [14-16]. Digital training approaches may overcome costs associated with travel, lodging, arranging centralized classroom space, and requirements for skilled trainers [7,17].

Despite the promise of digital training for lay HCWs, the impacts of digital training on reactions, knowledge, attitudes, and practices are not well understood. Although studies have evaluated the impact of digital training on knowledge, these studies were mainly descriptive, small, and not rigorously conducted [18,19]. Additionally, fewer studies have reported the impact on attitudinal gains and skills attainment in LMICs. To address these gaps, we developed and evaluated a digital training program to improve lay HCWs' knowledge and skills in HIV index case testing [15]. This analysis is part of the PRACTICE (Package of Resources for Assisted Contact Tracing: Implementation, Costs, and Effectiveness) study, which examined the implementation and impacts of a digital training approach compared to the standard of care for Malawi's HIV index case testing program. This analysis evaluates the impact of digital training on lay HCWs' knowledge, attitudes, and skills.

Methods

Study Setting: Malawi's Index Case Testing Program

The study was conducted in Malawi, a country in southeastern Africa with an HIV prevalence of 8.9% [20]. Nearly 60% of Malawians cannot operate a computer or internet-based mobile device [21]. Malawi is among the countries with the lowest number of clinicians per capita [22]. Hence, to address its HIV burden with limited trained human resources, Malawi has task-shifted HIV testing services to lay HCWs [23,24].

The Ministry of Health (MOH), in collaboration with United States Government partners, implemented a national lay HCW program to address the HIV epidemic. The primary responsibilities of lay HCWs include conducting HIV testing and counseling, linking persons to treatment and prevention services, tracing clients with missed clinic appointments, and collecting dried blood spot samples for viral load testing. In 2011, the role of lay HCWs expanded to implement index case testing, an evidence-based intervention where persons with HIV are counseled about recruiting their sexual partners, biological children, and other family members for HIV testing.

PRACTICE was conducted in 34 health facilities in 2 districts: Machinga and Balaka. These facilities were operated by the MOH or a faith-based organization called the Christian Association of Malawi.

The Tingathe program supported MOH and Christian Association of Malawi health facilities in the 2 districts with the provision of HIV testing services according to MOH guidelines. The Tingathe program, a PEPFAR implementing partner, is part of the Baylor College of Medicine Children's Foundation Malawi, providing free technical assistance and support for HIV testing and treatment services [5,25]. The program employed and supervised over 95% of the HCWs providing HIV testing and care services in the study districts.

Lay HCWs at the 34 health facilities conducted HIV index case testing according to Malawi MOH guidelines [26]. Index case testing has been offered as a choice in routine HIV testing services since 2011, with concerted implementation since 2017 with United States Government implementing partners' support [26-28]. Lay HCWs elicit contacts (sexual partners and biological children) from index clients (people living with HIV) and then provide index clients with a choice of passive referral, provider referral, contract referral, or dual referral [29].

Synchronous trainer-led face-to-face training had typically been provided for lay HCWs in Malawi's index case testing program. However, these training modalities posed challenges, including scheduling, service disruption, high costs due to travel, lodging, and allowances, inconsistent training fidelity, and limited time for practice or feedback to reinforce skills. The PRACTICE study was designed to evaluate whether a decentralized digital training could overcome some of these challenges.

Study Design and Population

Two of the 34 nearby small facilities were combined to form a cluster to make a total of 33 clusters. We conducted a cluster randomized controlled trial with 33 clusters (health facilities) randomly assigned in a 1:2 ratio to the standard of care plus the digital training implementation strategy (enhanced arm) or standard of care (standard arm), respectively. The digital training occurred from August to September 2022. The primary data collection period was from October 2022 to September 2023.

XSL•FO
RenderX

Research staff recruited lay HCWs between May and July 2022 who were aged 18 years or older, worked at 1 of the 34 facilities, and provided index case testing services. All lay HCWs at the facility who met the eligibility criteria were offered the opportunity to consent and enroll. Enrolled lay HCWs participated in the following activities relevant to this analysis: (1) surveys conducted before and after the training period and (2) fidelity assessments conducted after the training period. Social harms were evaluated during the study implementation.

Implementation Strategies

Standard Arm

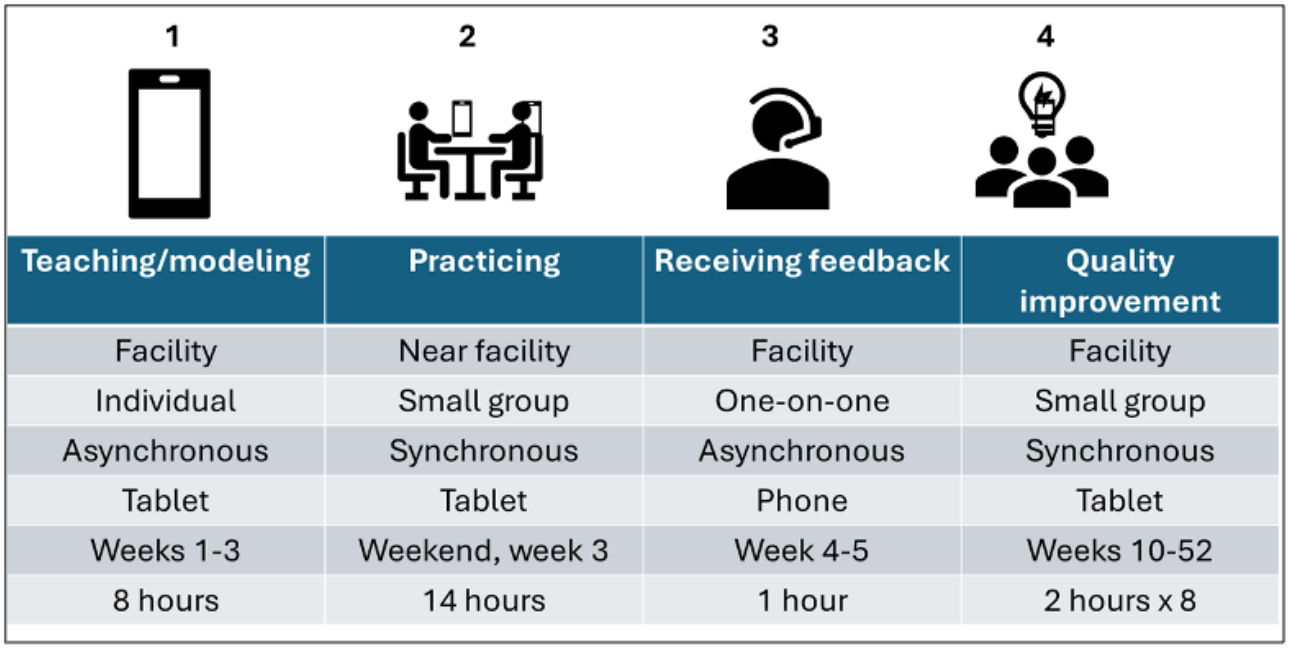
Prior to the start of the study, all lay HCWs in both arms received a 2-hour standard index case testing training, which was part of a larger in-person Tingathe training on all job responsibilities. The training included an overview of index case testing processes, referral methods, and counseling concepts. This was supplemented by routine in-person monthly facility mentorship visits conducted as part of standard implementation. The study did not alter any part of the existing standard strategies, and lay HCWs in both arms received this in October 2021.

Enhanced Arm

The enhanced arm had 4 components: (1) teaching, (2) practice, (3) feedback, and (4) quality improvement (Figure 1) [30].

1. Teaching/modeling: The 5 self-paced digital training sessions, lasting approximately 8-hours over a 3-week period, were accessible offline and focused on teaching and modeling pedagogies. The training videos included narration of index case testing concepts and culturally appropriate dramatic scenes modeling counseling sessions of index and contact clients. The videos centered on a fictional woman newly diagnosed with HIV and her contact clients (sexual partners, children, and household members). The video production was supported by a professional videographer.
2. Practice: The 8 small group face-to-face training sessions took place at the facility or a nearby venue during the weekend for nearly 14 hours, focusing on practicing that occurred at the facility or a nearby venue.
3. Feedback: One virtual phone feedback session at the facility lasting for approximately an hour.
4. Quality improvement: Six 2-hour tablet-guided monthly quality improvement sessions at the facility that focused on the facility's index case testing indicators to identify gaps and actionable solutions [31]. This digital training package was designed based on formative research [32], integrated social cognitive theory [33], and the theory of expertise [34]. The training was preloaded on a tablet distributed to the enhanced arm facilities and was accessible offline.

Figure 1. Enhanced implementation package.



Evaluation Framework and Outcomes

We used the Kirkpatrick training evaluation model [35] to guide this analysis. The framework has 4 levels of evaluation: reaction (level 1), which pertains to how participants perceived the training; learning (level 2), which relates to knowledge, attitudes, and skills (fidelity to index and contact testing protocols) based on the training; behavior (level 3), which measures the degree to which participants demonstrated a behavior change; and results as impact from training (level 4), which evaluates learning transfer to the clinical setting attributable to the training. We assessed the first 2 levels of the framework in this analysis: reactions and learning (Table 1). Levels 3 and 4 have been reported previously [30,36].

XSL•FO
RenderX

Table . Kirkpatrick framework training outcomes.

Outcome	Definition	Study arms	Timing	Measurement
Reaction (level 1)				
Engagement	• Completion of the self-paced digital training portion	Enhanced arm	Posttraining	• The study team monitored the completion of the self-directed digital training.
Satisfaction	• Participants' affective reactions to training	Enhanced arm	Posttraining	• Participants completed a 9-item training evaluation mainly focusing on training quality, technical level, duration, and usefulness.
Perceived usefulness	• Participants' belief that the training improves their performance and was easy to use	Enhanced arm	Posttraining	• Participants completed a 2-item training evaluation survey after each session using Likert scale response options on helpfulness and ease of use.
Attitude toward technology	• Confidence in using a digital training platform	Enhanced and standard arms	Pretraining and posttraining	• Attitudes toward digital training acceptability were adapted from Chao [37], with Likert scale response options.
Learning (level 2)				
Knowledge	• Information on how to conduct index case testing	Enhanced and standard arms	Pretraining and posttraining	• Twenty-one multiple-choice questions were developed related to the content of the digital training to measure understanding. The format was vignette-based questions developed by the study team. • Twenty-five multiple-choice questions were developed related to the content of the Tingathe program training to measure understanding. The format was type A and vignette-based questions developed by the Tingathe program.
Skills	• Fidelity to index client counseling protocols • Fidelity to contact client counseling protocols	Enhanced and standard arms	Posttraining	• Lay health care workers were observed by study staff counseling a mock index client based on a 15-item protocol. • Lay health care workers were observed by study staff counseling a mock contact client based on a 15-item protocol.

XSL•FO
RenderX

Reaction was assessed through engagement, satisfaction, perceived usefulness, and attitude toward digital training. Engagement was examined through monitoring the completion of the self-paced digital training. Satisfaction included (1) perceived training quality, measured using a 5-point rating scale with response options and free comments after each training session; (2) technical level of the material covered during the training, measured using a 3-point "Just-About-Right" scale with response options; and (3) length of material covered in the training, measured using a 3-point "Just-About-Right" scale with response options (Multimedia Appendix 1). Perceived usefulness was evaluated through participants completing 2 questions as part of a feedback evaluation survey at the end of each training session on (1) how helpful they found the training session using a 5-point Likert scale with response options used to measure helpfulness and (2) did they find the tablet easy to use during the training session using a 5-point Likert-type scale used to measure difficulty. Attitude toward digital technology was measured through self-rated questions adapted from Chao [37], with Likert scale response options.

Learning was assessed through knowledge and skills acquisition after the training. Knowledge was examined using 2 multiple-choice questionnaires. Knowledge scores were calculated on a scale from 0 to 100%.

Skills acquisition was evaluated using two 15-item fidelity checklists developed by the study team, one for index client counseling and the other for contact client counseling (Table S1 in Multimedia Appendix 2). Each lay HCW participated in simulated fidelity assessments with mock index and contact clients. These sessions were standardized counseling encounters, with study staff simulating the role of an index or contact client. All lay HCWs took part in 2 audio-recorded fidelity assessments, one with an index client and the other with a contact client, after the training. Research staff audio-recorded the simulated sessions.

Two assessors blinded to the study arm then listened to each session and provided a score on each of the 15 items on the checklist: needs improvement (0), satisfactory (1), or excellent (2). Total scores (0 - 30 points) were then calculated as percentages. A scoring criterion was developed for the assessors, and they were thoroughly trained to ensure consistency. The 2 assessors' scores were averaged.

Data Analysis

The data were exported into Stata software (version 19.0; StataCorp) for cleaning and analysis. Descriptive statistics were used to summarize baseline characteristics of the lay HCWs and study clusters. Additionally, descriptive analyses were performed to assess engagement, satisfaction, and perceived usefulness, as these were only ascertained in the enhanced arm.

Analyses of pretraining and posttraining knowledge and attitude questions included participants who responded to the same question at both time points. The scores were evaluated using the difference (enhanced−standard) in differences (Post-Pre within person) approach to adjust for small baseline differences between arms. The units of analysis for knowledge and attitude were the mean changes in per-item scores before training and post training. Between-arm (enhanced vs standard arm) mean differences were evaluated using generalized estimating equations with an identity link and normal distribution to account for potential correlation between observations at the cluster level.

Skill acquisition posttraining scores were used to assess agreement and/or reliability. Cohen κ with 95% CIs was calculated to evaluate scoring consistency (reliability) between the 2 assessors per checklist item. Landis and Koch [38] threshold values were used to interpret Cohen κ results. The κ values ranged from 0 (absence of agreement) to 1 (perfect agreement). Bland and Altman method [39] was used to evaluate the reliability of assessors' scoring. Two-sample Welch *t* tests were used to compare the mean fidelity scores per item between the enhanced and standard arms. Fidelity scores with 95% CIs were compared between enhanced and standard arms using generalized estimating equations with an identity link and normal distribution to account for potential correlation between observations at the cluster level. Fidelity scores for index and contact clients were analyzed separately.

Ethical Considerations

The study was reviewed and approved by the National Health Sciences Research Committee in Malawi (20/06/2566), the University of North Carolina at Chapel Hill Institutional Review Board (20-1810), and the Baylor College of Medicine Institutional Review Board (H-48800). Written informed consent was obtained from all study participants (Checklist 1). The trial was registered at ClinicalTrials.gov (NCT05343390). Participants received US $10 per study activity including: completing surveys, and three times for digital training activities.

Results

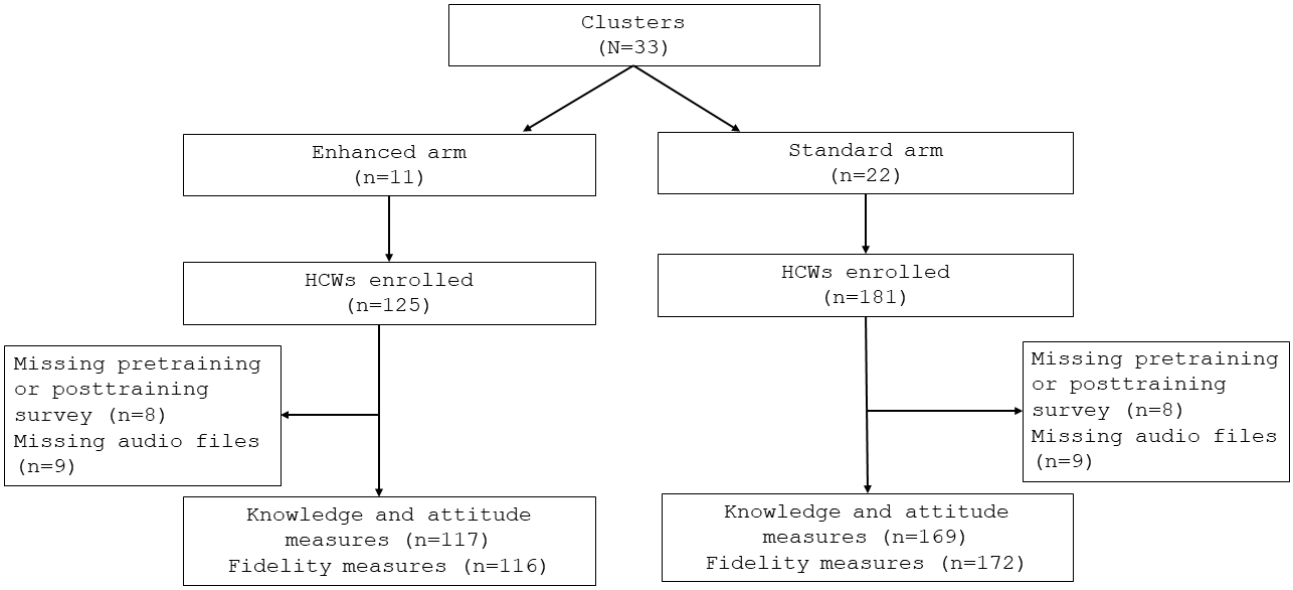
Population Characteristics

The study was conducted in 13 clusters in Balaka and 20 clusters in Machinga. There were 33 clusters, with 11 in the enhanced arm and 22 in the standard arm (Figure 2). The majority of clusters, 26 (79%), were in rural settings. Most clusters, 27 (82%), were health centers, and a few were secondary-level hospitals or small dispensaries. The median staffing per facility was 7 (IQR 6 - 9).

All the lay HCWs we approached at the facilities consented to join the study. They were all lay HCWs. We enrolled 306 lay HCWs: 125 (40.8%) in the enhanced arm and 181 (59.2%) in the standard arm. One lay HCW from the standard arm did not complete the baseline survey, and data for 305 lay HCWs are presented. The median age of the lay HCWs was 34 (IQR 27 - 39) years, and the majority, 173 (57%), completed secondary education. The median time in the lay HCW role was 3 (IQR 1 - 5) years. Most lay HCWs, 242 (80%), owned their own smartphone, and half had received prior tablet-based training (Table 2).

XSL•FO
RenderX

Figure 2. Flowchart of study participants. HCWs: health care workers.



XSL•FO
RenderX

Table . Health care worker characteristics by study arm.

Characteristic	Enhanced arm (n=125), n (%)	Standard arm (n=180)[a], n (%)
Age (y)		
21 - 30	48 (38)	80 (44)
31 - 40	59 (47)	54 (30)
41+	18 (14)	46 (26)
Sex		
Male	66 (52)	93 (51)
Female	59 (47)	87 (48)
Cadre		
CHW[b]	37 (30)	55 (31)
CHW/counselor	63 (50)	84 (47)
HIV diagnostic assistant	9 (7)	13 (7)
Tingathe site supervisor	12 (10)	22 (12)
Other health worker	4 (4)	6 (3)
Education		
Some secondary education	16 (13)	27 (15)
Completed secondary education	73 (58)	102 (56)
Tertiary education	36 (29)	51 (28)
Work experience (y)		
<1	24 (19)	32 (18)
1 - 3	45 (36)	62 (34)
4 - 6	40 (32)	65 (36)
7+	16 (13)	21 (12)
Own a smartphone		
No	19 (15)	33 (18)
Yes, mine	102 (82)	142 (79)
Yes, shared	4 (3)	5 (3)
Trained using a mobile device before		
No	64 (51)	91 (51)
Yes	61 (48)	89 (49)
Prior use of a tablet for work		
No	84 (67)	126 (70)
Yes	41 (33)	54 (30)
Prior index case testing training[c]		
No	29 (23)	42 (23)
Yes	95 (76)	137 (76)

[a]One lay health care worker left before completing the baseline survey.

[b]CHW: community health worker.

[c]Two lay health care workers chose not to respond.

Summary of Key Findings

Knowledge about index testing and attitudes toward digital training improved more in the enhanced arm than in the standard arm. At endline, fidelity measures were higher in the enhanced arm than in the standard arm. We detail these findings below.

XSL•FO
RenderX

Engagement

In the enhanced arm, 100% of the lay HCWs completed 8 hours of the self-paced digital training portion within the allotted 3 weeks. Uptake of the small group face-to-face training sessions was 98%, and the virtual phone feedback session was 100%. Quality improvement sessions ranged from 81% to 93%.

Experience of Digital Training

Most participants expressed satisfaction with self-paced training sessions (107/125, 85%) and small-group face-to-face training sessions (112/125, 90%). Additionally, the majority of participants (108/125, 86%) agreed that the digital training and face-to-face training content technical levels were just right and provided content relevant to their practice. Furthermore, 86% (108/125) of the participants perceived the length of the course to be adequate.

The results of the training evaluation after each session in the enhanced arm, on a scale of 1 to 5, with 5 being the highest, showed that overall HCWs found the training very helpful, with a mean of 4.3 (SD 0.6). They also found it very easy to use, with a mean of 4.8 (SD 0.6).

Attitudes Toward Any Digital Training

There were 94% (286/305) lay HCWs who completed the surveys pretraining and posttraining. On average, digital technology acceptability scores increased by 1.9%, from 67.0% pretraining to 68.9% posttraining in the enhanced arm, while in the standard arm, there was a 2.6% reduction in acceptability scores, from 67.9% pretraining to 65.2% posttraining. The between-arm difference in pre-to-post differences comparing the enhanced vs standard arms was 4.5% (95% CI 0.3%, 7.0%; *P*=.03).

Knowledge

There were 94% (286/305) lay HCWs who completed the questions pretraining and posttraining. Knowledge improved from 63.3% pretraining to 70.4% posttraining in the enhanced arm. It was relatively consistent in the standard arm: 63.4% to 66.2%. The between-arm difference was 4.4% (95% CI 0.7%-8.2%) percentage points higher in the enhanced arm (*P*=.02). Correct answers for the Tingathe program's index case testing examination increased from 68.2% pretraining to 72.8% in posttraining in the enhanced arm compared to no improvement in the standard arm (68.9% pretraining and 69.4% posttraining; Table 3). The between-arm difference in scores was 4.0% (95% CI 1.4%-6.7%; *P*=.002).

Table . Comparison of health care worker knowledge percentage scores and attitude scores between the enhanced and standard arms.

Item	Enhanced arm (n=117)			Standard arm (n=169)			Between-arm MD[a]	
	Pretraining	Posttraining	MD (95% CI)	Pretraining	Posttraining	MD (95% CI)	MD (95% CI)	*P* value
Knowledge: 21 multiple-choice questions developed by the study team	63.4	70.4	7.0 (4.5 to 9.5)	63.4	66.2	2.6 (0.7 to 4.4)	4.4 (0.7 to 8.2)	.02
Knowledge: 25 multiple-choice questions developed by the Tingathe program	68.2	72.8	4.6 (2.7 to 6.5)	68.9	69.4	0.6 (−0.9 to 2.0)	4.0 (1.4 to 6.7)	.002
Digital technology acceptability	67.0	68.9	1.9 (−0.5 to 4.3)	67.9	65.2	−2.6 (−5.0 to −0.3)	4.5 (0.3 to 7.0)	.03

[a]MD: mean difference.

Skill Acquisition

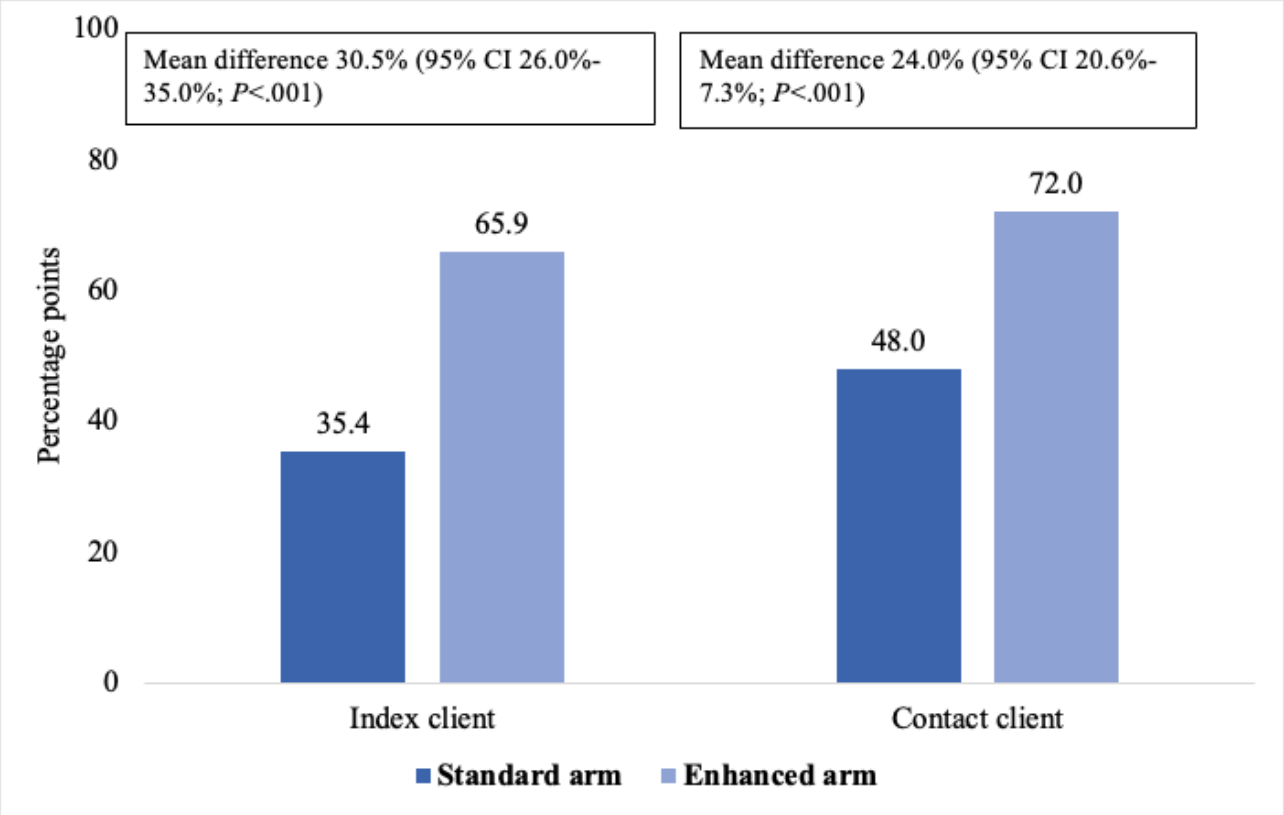
For the index client-HCW encounters, interassessor reliability with Cohen κ ranged from 0.56 to 0.95. For the contact client-HCW interassessor reliability with Cohen κ ranged from 0.69 to 0.97 (Table S1 in Multimedia Appendix 2). The Bland-Altman plots demonstrated good agreement between the assessors (Figure S1A,B in Multimedia Appendix 3).

There were better scores across each of the 15 items in the enhanced arm compared to the standard arm for the index fidelity checklist (*P*<.001). Similarly, there were better scores across 13 of the 15 items in the enhanced arm compared to the standard arm on the contact fidelity checklist (Table S1 in Multimedia Appendix 2).

The mean fidelity score for index client-HCW encounters was higher in the enhanced arm (65.9%, 95% CI 62.1%-69.8%), compared to the standard arm (35.4%, 95% CI 33.1%-37.7%). Fidelity was 30.5 percentage points higher in the enhanced arm (mean difference 30.5%, 95% CI 26.0%-35.0%; *P*<.001). The mean fidelity score for contact client-HCW encounters was 72.0% (95% CI 69.4%-74.5%) in the enhanced arm and 48.0% (95% CI 45.7%-50.1%) in the standard arm. Fidelity was 24.0 percentage points higher in the enhanced arm (mean difference 24.0%, 95% CI 20.6%-27.3%; *P*<.001; Figure 3).

XSL•FO
RenderX

Figure 3. Fidelity scores between arms.



Discussion

We developed and rolled out a digital training for lay HCWs in Malawi. Lay HCWs completed nearly the entire training; most were satisfied with it and found it easy to use. This training modestly improved knowledge (<5%) of index case testing concepts and attitudes toward digital training. Further, the training substantially improved (>20%) lay HCWs' skills to conduct index case testing counseling.

The PRACTICE study is among the first randomized trials assessing the impact of digital training using an experimental design in a low-resource setting. Notably, previous studies were small, descriptive, and nonexperimental. A scoping review that focused on digital training for rural professional cadres [18] identified few experimental studies (n=5) and focused mainly on knowledge acquisition. Similarly, a systematic review that evaluated digital learning for medical education in LMICs [40] found only 4 randomized controlled trials. Most prior work focused on the impact of digital training on knowledge. An increase in knowledge after digital training has been found in other LMICs [4,41] and high-income countries [10,42]. Our digital training also modestly improved knowledge of index case testing protocols and principles.

This study is among the first to provide evidence for enhanced skills. Our results are consistent with a similar study from Bangladesh reporting that digital training improved HCWs' counseling skills in the context of family planning and maternal and child health [43]. The substantial skill enhancement, coupled with only a modest increase in knowledge, suggests that knowledge was not the only ingredient needed to uplift HCWs' skills. Although our enhanced training included some didactic knowledge, it focused on modeling, practicing, and offering feedback on counseling skills. These pedagogical elements likely drove the improvements in skills. Similarly, acceptance of technology may have been a necessary ingredient for completing training, but it was not the main driver of skill improvement. In addition, we also observed an impact on behavioral and clinical outcomes in the PRACTICE study [30,36]. Perhaps the improvement in skills is what led to improvements in behavioral and clinical outcomes.

Overall, the lay HCWs reported a high level (>85%) of engagement and satisfaction. All participants completed the digital component. Other reports from Kenya [44] and Rwanda [6] similarly found digital training acceptable and feasible due to flexibility and content relevance. This was a clear advantage over centralized training, where lay HCWs would need to schedule time off work and potentially travel to an off-site venue. The inclusive methods used to develop the training and customized attention to culture and local context may have enhanced the levels of engagement and satisfaction [45-47].

HCWs in the enhanced arm had improved acceptability toward digital training modalities compared to those in the standard arm. Our findings are consistent with other research in low-resource settings, which shows that digital training interventions to improve lay HCWs' quality of health services can be acceptable [7,45]. Additionally, positive attitudinal gains after digital training have been reported in other studies in Malawi [48] and other LMICs [49], demonstrating that technology-based training for lay HCWs in LMICs can improve confidence and attitudes to use the technology.

XSL•FO
RenderX

We designed our training with feasibility considerations in mind. First, our choice of an *offline* platform took into consideration the limited network infrastructure in this setting. Second, we oriented the facility supervisors on how to support training for their sites. Third, we used a training platform that was easy to navigate. Finally, the allocated time to complete the training was appropriate.

One limitation of the digital training acceptability and satisfaction measures is social desirability. HCWs may have sought to please trainers who were requesting feedback, thereby providing more positive responses than they might otherwise. To minimize this, lay HCWs were encouraged to respond honestly and to provide their feedback privately.

It is not known whether the results of this study would generalize to government-employed lay HCWs, who provide services in facilities that lack the support of implementing partners. Government- and Tingathe-supported lay HCWs may differ in their scope of work, supervision, management, funding source, and compensation. Given that digital training has the potential to meet the training needs of lay HCWs with minimal disruption compared to centralized training, it is important to understand the uptake of digital training strategies in other contexts as the next step.

Contamination was a small potential risk in this study. In this study, contamination referred to HCWs in the standard arm having access to digital training. Notably, the lay HCWs were recruited from real-world settings; therefore, it was not possible to eliminate contamination. We minimized this by controlling access to the training tablets. We also tracked all staff transfers from one facility to another, and this was minimal (<1% of lay HCWs).

Digital training can be an acceptable and practical tool to enhance the competence of lay HCWs to conduct index case testing counseling in a resource-constrained setting. The training strengthened HCWs' reaction and confidence in digital training. It led to improvement in knowledge and counseling skills, which may strengthen the provision of quality health care services. In light of reductions in development assistance for health, there is an increasing need for decentralized, cost-effective, minimal-interruption training for HCWs in LMICs, and our digital training package shows promise in supporting this need. Digital training approaches can address training needs for lay health workers in many low-income settings.

Acknowledgments

The authors acknowledge and thank the Malawi Ministry of Health, study participants, research team, Tingathe Outreach Program, and Baylor College of Medicine Children's Foundation Malawi. The authors confirm that no artificial intelligence tools were used in the conduct of the research or the preparation of this manuscript.

Funding

This research was supported by R01 MH124526. TAT was supported by the Fogarty International Center of the National Institutes of Health under award number D43 TW010060. KM and NER were supported in 2019 by P30 AI050410. The funders had no role in trial design, data collection and analysis, decision to publish, or preparation of the manuscript.

Data Availability

The data that support the findings of this study are available from the corresponding author upon reasonable request.

Authors' Contributions

TAT, NER, KM, KS, SER, MK, SA, and LGB conceptualized the study. TAT, KS, MJC, MM, EK, MK, SA, and NER designed the training package. TAT, AMM, MJC, CK, and DP facilitated the training. MJC and MM were responsible for data coding. KM provided statistical guidance for the data analyses. TAT wrote the first draft of the manuscript and analyzed the data. All authors critically reviewed, edited, and approved the final manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Index case testing training evaluation form.
[DOCX File, 19 KB - mededu_v12i1e89942_app1.docx]

Multimedia Appendix 2
Mean scores and Cohen κ values per item on the simulated fidelity checklists.
[DOCX File, 19 KB - mededu_v12i1e89942_app2.docx]

Multimedia Appendix 3

XSL•FO
RenderX

Bland and Altman plot of (A) index and (B) contact clients–health care worker simulated fidelity score.
[DOCX File, 72 KB - mededu_v12i1e89942_app3.docx]

Checklist 1
CONSORT-EHEALTH checklist (V 1.6.1).
[PDF File, 1102 KB - mededu_v12i1e89942_app4.pdf]

References

1. WHO guideline on health policy and system support to optimize community health worker programmes. : World Health Organization; 2018 Aug 6 URL: https://www.who.int/publications/i/item/9789241550369 [accessed 2026-06-19]
2. Pallas SW, Minhas D, Pérez-Escamilla R, Taylor L, Curry L, Bradley EH. Community health workers in low- and middle-income countries: what do we know about scaling up and sustainability? Am J Public Health 2013 Jul;103(7):e74-e82. [doi: 10.2105/AJPH.2012.301102] [Medline: 23678926]
3. Lewin SA, Dick J, Pond P, et al. Lay health workers in primary and community health care. Cochrane Database Syst Rev 2005 Jan 25;2005(1):CD004015. [doi: 10.1002/14651858.CD004015.pub2] [Medline: 15674924]
4. Sangwa Y, Ndaruhutse V, Radeny S, et al. Using eLearning to improve and retain the knowledge of community health workers in maternal and neonatal health in Rwanda: a cohort study. Public Health Chall 2024 Jun;3(2):e174. [doi: 10.1002/puh2.174] [Medline: 40496154]
5. Kim MH, Ahmed S, Buck WC, et al. The Tingathe programme: a pilot intervention using community health workers to create a continuum of care in the prevention of mother to child transmission of HIV (PMTCT) cascade of services in Malawi. J Int AIDS Soc 2012 Jul 11;15 Suppl 2(Suppl 2):17389. [doi: 10.7448/IAS.15.4.17389] [Medline: 22789644]
6. Sranacharoenpong K, Hanning RM. Developing a diabetes prevention education programme for community health-care workers in Thailand: formative findings. Prim Health Care Res Dev 2011 Oct;12(4):357-369. [doi: 10.1017/S146342361100020X] [Medline: 22284949]
7. O'Donovan J, Kabali K, Taylor C, et al. The use of low-cost Android tablets to train community health workers in Mukono, Uganda, in the recognition, treatment and prevention of pneumonia in children under five: a pilot randomised controlled trial. Hum Resour Health 2018 Sep 19;16(1):49. [doi: 10.1186/s12960-018-0315-7] [Medline: 30231894]
8. O'Donovan J, O'Donovan C, Kuhn I, Sachs SE, Winters N. Ongoing training of community health workers in low-income and middle-income countries: a systematic scoping review of the literature. BMJ Open 2018 Apr 28;8(4):e021467. [doi: 10.1136/bmjopen-2017-021467] [Medline: 29705769]
9. Budiman R, Efendi Y, Riyanto A. Benefits, challenges, and opportunities of taking face-to-face tutorials in a blended learning environment: students' view. Presented at: ICEEL '18: Proceedings of the 2018 2nd International Conference on Education and E-Learning; Nov 5-7, 2018. [doi: 10.1145/3291078.3291116]
10. Vallée A, Blacher J, Cariou A, Sorbets E. Blended learning compared to traditional learning in medical education: systematic review and meta-analysis. J Med Internet Res 2020 Aug 10;22(8):e16504. [doi: 10.2196/16504] [Medline: 32773378]
11. Willcox M, LeFevre A, Mwebaza E, Nabukeera J, Conecker G, Johnson P. Cost analysis and provider preferences of low-dose, high-frequency approach to in-service training programs in Uganda. J Glob Health 2019 Jun;9(1):010416. [doi: 10.7189/jogh.09.010416] [Medline: 30774942]
12. Condo J, Mugeni C, Naughton B, et al. Rwanda's evolving community health worker system: a qualitative assessment of client and provider perspectives. Hum Resour Health 2014 Dec 13;12:71. [doi: 10.1186/1478-4491-12-71] [Medline: 25495237]
13. Scott K, Beckham SW, Gross M, et al. What do we know about community-based health worker programs? A systematic review of existing reviews on community health workers. Hum Resour Health 2018 Aug 16;16(1):39. [doi: 10.1186/s12960-018-0304-x] [Medline: 30115074]
14. Pamfilie R, Onete B, Maiorescu I, Pleşea D. E-Learning as an alternative solution for sustainable lifelong education. Procedia Soc Behav Sci 2012;46:4026-4030. [doi: 10.1016/j.sbspro.2012.06.191]
15. Wang W, Zhang HB, Liu JM, et al. Variations, effectiveness and its associated factors of a nationwide web-based hypertension management training project in China: insights from a government-led campaign for 1.2 million lay health workers. J Geriatr Cardiol 2024 Jul 28;21(7):733-750. [doi: 10.26599/1671-5411.2024.07.010] [Medline: 39183951]
16. Ladur AN, Egere U, Ravit M, et al. A blended learning approach for capacity strengthening to improve the quality of integrated HIV, TB, and malaria services during antenatal and postnatal care in LMICs: a feasibility study. BMC Med Educ 2025 Jan 8;25(1):35. [doi: 10.1186/s12909-024-06633-2] [Medline: 39780210]
17. Feroz A, Jabeen R, Saleem S. Using mobile phones to improve community health workers performance in low-and-middle-income countries. BMC Public Health 2020 Jan 13;20(1):49. [doi: 10.1186/s12889-020-8173-3] [Medline: 31931773]
18. Woods L, Martin P, Khor J, Guthrie L, Sullivan C. The right care in the right place: a scoping review of digital health education and training for rural healthcare workers. BMC Health Serv Res 2024 Sep 2;24(1):1011. [doi: 10.1186/s12913-024-11313-4] [Medline: 39223581]

XSL•FO
RenderX

19. Aryee GFB, Amoadu M, Obeng P, et al. Effectiveness of eLearning programme for capacity building of healthcare professionals: a systematic review. Hum Resour Health 2024 Sep 2;22(1):60. [doi: 10.1186/s12960-024-00924-x] [Medline: 39223555]
20. Malawi population-based HIV impact assessment 2020-2021(MPHIA 2020–2021): final report. : Ministry of Health (MOH); 2022 Nov URL: https://phia.icap.columbia.edu/wp-content/uploads/2022/12/241122_Mphia_Foreword.pdf [accessed 2026-06-19]
21. Digital skills ecosystem and gap assessment in Malawi. : Ernst & Young LLP (EY); 2021 URL: https://digmap.pppc.mw/wp-content/uploads/2022/01/23092021PPPC-Digital-Skills-Ecosystem-and-Gap-Assessment_Malawi_Final-Report-v1.pdf [accessed 2026-06-19]
22. Medical doctors (per 10,000 population). World Health Organization. URL: https://www.who.int/data/gho/data/indicators/indicator-details/GHO/medical-doctors-(per-10-000-population) [accessed 2026-06-19]
23. Kim MH, Ahmed S, Preidis GA, et al. Low rates of mother-to-child HIV transmission in a routine programmatic setting in Lilongwe, Malawi. PLoS ONE 2013;8(5):e64979. [doi: 10.1371/journal.pone.0064979] [Medline: 23741437]
24. McCollum ED, Preidis GA, Kabue MM, et al. Task shifting routine inpatient pediatric HIV testing improves program outcomes in urban Malawi: a retrospective observational study. PLoS One 2010 Mar 10;5(3):e9626. [doi: 10.1371/journal.pone.0009626] [Medline: 20224782]
25. Ahmed S, Kim MH, Dave AC, et al. Improved identification and enrolment into care of HIV-exposed and -infected infants and children following a community health worker intervention in Lilongwe, Malawi. J Int AIDS Soc 2015;18(1):19305. [doi: 10.7448/IAS.18.1.19305] [Medline: 25571857]
26. Ministry of Health and Sanitation. URL: https://www.health.gov.mw/ [accessed 2026-06-26]
27. Tembo TA, Kim MH, Simon KR, et al. Enhancing an HIV index case testing passive referral model through a behavioural skills-building training for healthcare providers: a pre-/post-assessment in Mangochi District, Malawi. J Int AIDS Soc 2019 Jul;22(Suppl 3):e25292. [doi: 10.1002/jia2.25292] [Medline: 31321917]
28. Malawi HIV policy updates: the government of Malawi's commitment and plan to reach HIV epidemic control through 90-90-90 in Malawi by 2020. : Ministry of Health and Population; 2019 URL: https://hivpreventioncoalition.unaids.org/en/resources/malawi-hiv-policy-updates-ref-no-adm-11-government-malawis-commitment-and-plan-reach?utm_source=chatgpt.com [accessed 2026-06-19]
29. Ferreira A, Young T, Mathews C, Zunza M, Low N. Strategies for partner notification for sexually transmitted infections, including HIV. Cochrane Database Syst Rev 2013 Oct 3;2013(10):CD002843. [doi: 10.1002/14651858.CD002843.pub2] [Medline: 24092529]
30. Rosenberg N, Mollan K, Wang J, et al. Effect of a digital training package on clinical outcomes in Malawi's index case testing programme: a cluster randomised controlled trial. BMJ Glob Health 2026 Mar 13;11(3):e022563. [doi: 10.1136/bmjgh-2025-022563] [Medline: 41825916]
31. Tembo TA, Mollan K, Simon K, et al. Does a blended learning implementation package enhance HIV index case testing in Malawi? A protocol for a cluster randomised controlled trial. BMJ Open 2024 Jan 22;14(1):e077706. [doi: 10.1136/bmjopen-2023-077706] [Medline: 38253452]
32. Rosenberg NE, Tembo TA, Simon KR, et al. Development of a blended learning approach to delivering HIV-assisted contact tracing in Malawi: applied theory and formative research. JMIR Form Res 2022 Apr 19;6(4):e32899. [doi: 10.2196/32899] [Medline: 35438644]
33. Bandura A. Human agency in social cognitive theory. Am Psychol 1989 Sep;44(9):1175-1184. [doi: 10.1037/0003-066x.44.9.1175] [Medline: 2782727]
34. Ericsson KA. Deliberate practice and acquisition of expert performance: a general overview. Acad Emerg Med 2008 Nov;15(11):988-994. [doi: 10.1111/j.1553-2712.2008.00227.x] [Medline: 18778378]
35. Yardley S, Dornan T. Kirkpatrick's levels and education "evidence". Med Educ 2012 Jan;46(1):97-106. [doi: 10.1111/j.1365-2923.2011.04076.x] [Medline: 22150201]
36. Tembo T, Mollan K, Chitani M, et al. Assessing fidelity to index case testing protocol in a cluster randomized controlled trial to improve HIV case finding [Abstract]. Presented at: The 16th AIDSImpact Conference (AIDSImpact 2025); May 26-28, 2025 URL: https://www.aidsimpact2025.com/programme/ [accessed 2025-06-26]
37. Chao CM. Factors determining the behavioral intention to use mobile learning: an application and extension of the UTAUT model. Front Psychol 2019;10:1652. [doi: 10.3389/fpsyg.2019.01652] [Medline: 31379679]
38. Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics 1977 Mar;33(1):159-174. [doi: 10.2307/2529310] [Medline: 843571]
39. Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. Lancet 1986 Feb 8;1(8476):307-310. [Medline: 2868172]
40. Barteit S, Guzek D, Jahn A, Bärnighausen T, Jorge MM, Neuhann F. Evaluation of e-learning for medical education in low- and middle-income countries: a systematic review. Comput Educ 2020 Feb;145:103726. [doi: 10.1016/j.compedu.2019.103726] [Medline: 32565611]

XSL•FO
RenderX

41. Hicks JP, Allsop MJ, Akaba GO, et al. Acceptability and potential effectiveness of eHealth tools for training primary health workers from Nigeria at scale: mixed methods, uncontrolled before-and-after study. JMIR Mhealth Uhealth 2021 Sep 16;9(9):e24182. [doi: 10.2196/24182] [Medline: 34528891]
42. Hurley DA, Keogh A, Mc Ardle D, et al. Evaluation of an e-learning training program to support implementation of a group-based, theory-driven, self-management intervention for osteoarthritis and low-back pain: pre-post study. J Med Internet Res 2019 Mar 7;21(3):e11123. [doi: 10.2196/11123] [Medline: 30843863]
43. Limaye RJ, Ballard Sara A, Ahmed N, et al. Enhancing the knowledge and behaviors of fieldworkers to promote family planning and maternal, newborn, and child health in Bangladesh through a digital health training package: results from a pilot study. Int Q Community Health Educ 2020 Jan;40(2):143-149. [doi: 10.1177/0272684X19861866] [Medline: 31274369]
44. Allott H, Smith A, White S, et al. Improving capacity for advanced training in obstetric surgery: evaluation of a blended learning approach. BMC Med Educ 2025 Jan 17;25(1):80. [doi: 10.1186/s12909-025-06660-7] [Medline: 39819576]
45. Muke SS, Shrivastava RD, Mitchell L, et al. Acceptability and feasibility of digital technology for training community health workers to deliver brief psychological treatment for depression in rural India. Asian J Psychiatr 2019 Oct;45:99-106. [doi: 10.1016/j.ajp.2019.09.006] [Medline: 31557716]
46. Khan A, Shrivastava R, Tugnawat D, et al. Design and development of a digital program for training non-specialist health workers to deliver an evidence-based psychological treatment for depression in primary care in India. J Technol Behav Sci 2020 Dec;5(4):402-415. [doi: 10.1007/s41347-020-00154-7] [Medline: 34150992]
47. Rahman A, Akhtar P, Hamdani SU, et al. Using technology to scale-up training and supervision of community health workers in the psychosocial management of perinatal depression: a non-inferiority, randomized controlled trial. Glob Ment Health (Camb) 2019;6:e8. [doi: 10.1017/gmh.2019.7] [Medline: 31157115]
48. Mastellos N, Tran T, Dharmayat K, et al. Training community healthcare workers on the use of information and communication technologies: a randomised controlled trial of traditional versus blended learning in Malawi, Africa. BMC Med Educ 2018 Apr 2;18(1):61. [doi: 10.1186/s12909-018-1175-5] [Medline: 29609596]
49. Willems A, Iyamuremye JD, Misage CN, Smith-Swintosky V, Kayiteshonga Y. Co-creation and evaluation of nationwide remote training service for mental health education of community health workers in Rwanda. Front Public Health 2021;9:632793. [doi: 10.3389/fpubh.2021.632793] [Medline: 34504821]

Abbreviations

CONSORT-EHEALTH: Consolidated Standards of Reporting Trials of Electronic and Mobile Health Applications and Online Telehealth
HCW : health care worker
LMICs: low- and middle-income countries
MOH : Ministry of Health
PRACTICE: Package of Resources for Assisted Contact Tracing: Implementation, Costs, and Effectiveness

Edited by LT Car; submitted 19.Dec.2025; peer-reviewed by G Semjonova, M Medich; revised version received 31.Mar.2026; accepted 30.May.2026; published 15.Jul.2026.

Please cite as:
Tembo TA, Rosenberg NE, Mollan K, Kim M, Rutstein SE, Mkandawire AM, Chitani MJ, Kumbuyo C, Phiri D, Mphande M, Kavuta E, Chilala S, Wang J, Ahmed S, Simon K, Bekker LG
Digital Training for Lay Health Care Workers' Knowledge and Skills in HIV Index Case Testing: Cluster Randomized Trial
JMIR Med Educ 2026;12:e89942
URL: https://mededu.jmir.org/2026/1/e89942
doi:10.2196/89942

© Tapiwa A Tembo, Nora E Rosenberg, Katie Mollan, Maria Kim, Sarah E Rutstein, Angella M Mkandawire, Mike J Chitani, Caroline Kumbuyo, Duncan Phiri, Mtisunge Mphande, Elijah Kavuta, Samuel Chilala, Jiayu Wang, Saeed Ahmed, Katherine Simon, Linda-Gail Bekker. Originally published in JMIR Medical Education (https://mededu.jmir.org), 15.Jul.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Education, is properly cited. The complete bibliographic information, a link to the original publication on https://mededu.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Digital Simulation–Based Ultrasound Training for Physiotherapy Students: Prospective Randomized Controlled Trial

Samuel Fernández-Carnero[1], PhD; Belen Diaz-Pulido[1], PhD; Jorge Méndez-Rodriguez[2], MSc; Daniel Pecos-Martin[1], PhD; Santiago Garcia-Miguel[1], PhD; Alexander Achalandabaso-Ochoa[3], PhD; Nicolas Cuenca-Zaldívar[4], PhD; Fermín Naranjo-Cinto Sr[5], PhD

[1]Universidad de Alcalá, Facultad de Medicina y Ciencias de la Salud,, Departamento de Enfermería y Fisioterapia, Grupo de Investigación Fisioterapia y Dolor, 28801 Alcalá de Henares, Spain

[2]Development and Programming Department, Amazonic Global, s.l, Madrid, Madrid, Spain

[3]Department of Health Sciences, Unversidad de Jaen, Jaen, Spain

[4]Research Group in Nursing and Health Care, Puerta de Hierro Health Research Institute—Segovia de Arana (IDIPHISA), C. Joaquín Rodrigo, 1, Majadahonda, Madrid, Spain

[5]Facultad de Medicina, Benemerita Universidad Autonoma de Puebla, Puebla, Mexico

Corresponding Author:
Nicolas Cuenca-Zaldívar, PhD
Research Group in Nursing and Health Care, Puerta de Hierro Health Research Institute—Segovia de Arana (IDIPHISA), C. Joaquín Rodrigo, 1, Majadahonda, Madrid, Spain

Abstract

Background: Ultrasound education has traditionally relied on on-site training, but scalable digital solutions are increasingly needed.

Objective: This study aimed to evaluate the effectiveness of a simulation-based online ultrasound platform compared with traditional on-site training.

Methods: A prospective randomized blinded study was conducted with 68 physiotherapy students (n=34 per group). Participants were assigned to a simulation-based online training platform (WAZO) or traditional on-site instruction. Both groups completed identical theoretical and practical assessments. Item response theory using a Rasch model was applied to evaluate item difficulty and student ability.

Results: No significant differences were found between the online and on-site groups in theoretical scores (mean 4.94, SD 1.47 vs mean 5.08, SD 1.14; *P*=.65) or practical performance variables, including probe handling (26/34, 76.5% vs 28/34, 82.4%; *P*=.37) and structure identification (24/34, 70.6% vs 22/34, 64.7%; *P*=.19). Measurement outcomes also showed no significant differences, including structure diameter (mean 3.78, SD 0.79 mm vs mean 3.98, SD 1.27 mm; *P*=.46) and structure surface distance (mean 3.94, SD 1.97 mm vs mean 3.24, SD 0.64 mm; *P*=.06). Item response theory analysis identified image optimization, structure diameter, and structure surface distance as the most difficult items, while probe handling and structure identification were the most informative. The model demonstrated high discriminative performance (area under the curve=0.93), with sensitivity of 87% and specificity of 80%.

Conclusions: Simulation-based online ultrasound training provides comparable theoretical and practical outcomes to traditional on-site instruction, supporting its use as a scalable and accessible educational alternative.

(JMIR Med Educ 2026;12:e87897) doi:10.2196/87897

KEYWORDS

ultrasound education; simulation-based training; physiotherapy education; digital health education; randomized controlled trial; item response theory

Introduction

Training in ultrasound for health science professionals for clinical use is based on the traditional classroom model [1] due to the empirical need to learn this tool, which has demonstrated its usefulness and suitability. Innovation in higher education is an opportunity to provide better educational experiences that allow health care professionals to provide a greater benefit to society [2]. Point-of-care ultrasound has demonstrated high reliability in soft-tissue scanning, being completely harmless, low cost, and high value for bedside use [3,4]. These conditions mean that ultrasound enjoys greater consideration than other imaging tests because the clinician can extract relevant information that can be directly correlated with clinical history,

XSL•FO
RenderX

making diagnosis more powerful, because imaging tests are not always indicative of dysfunction or pain [5]. The training pathway in some health sciences disciplines includes ultrasound competencies from the second year of training [6], an approach that has also been adopted by other disciplines to better prepare professionals for clinical practice.

The use of simulators in certain disciplines such as aeronautics [7], automotive engineering, or robotic surgeries is widely accepted and proven to be a fundamental basis for a lower carbon footprint. Online training systems in the health sciences have been shown to enable professionals to achieve higher success rates, lower costs, and greater efficiency and to work more quickly [8].

Online and digital materials have been previously used to explore the possibility of ultrasound education in different specialties, such as obstetrics [9], anesthesia residents for lung assessment [10], physiotherapy [11,12], postdegree physician echocardiography [13], and other allied health sciences, demonstrating the potential of these resources and the need to improve this line of education to achieve the necessary skills and knowledge for clinical settings.

To the best of our knowledge, the current models of training in ultrasound are based on videos of varying complexity, leaving aside simulation and training in aspects of image optimization or the use of tools such as gain, dual image, or calipers.

Item response theory (IRT) was used to evaluate the psychometric properties of the assessment. Detailed methodological assumptions and model specifications are provided in Multimedia Appendix 1 [14,15].

One of the main arguments in favor of IRT methods is that they do not use, as in classical test theory, the standard error of measurement (SEM) of the tested population to estimate the parameters, so the calculated parameters are not influenced by the characteristics of the population. This allows the creation of questionnaires whose properties are invariant regardless of the population to which they are applied [16].

The use of the Rasch analysis requires the fulfillment of several assumptions: unidimensionality, that is, that all the items measure the same construct; local independence, that is, the items are not related to each other; and invariance of the parameters, that is, the items present the same item characteristic curves across subsamples of the participants. Finally, the existence of differential item functioning must be verified, that is, whether students with the same level of θ belonging to different subgroups of the sample have the same probability of responding correctly to the items [17].

The purpose of this study was to evaluate whether online training with a simulator platform called WAZO (WAZO Technologies Lab, sl) [18] (“the platform” hereafter) could be a viable alternative to traditional training, given that point-of-care ultrasound is becoming essential, and health care professionals need the knowledge necessary to provide better health care.

Methods

Overview

A prospective longitudinal experimental study was designed in which all second-year students enrolled in the Assessment in Physiotherapy course of the e Degree in Physiotherapy program at the University of Alcalá (Spain) were randomized in 2 groups: online (experimental) and on-site (control).

The study design followed the CONSORT (Consolidated Standards of Reporting Trials) standard [19]. This study was not prospectively registered in a clinical trials registry, as it was conducted as an educational intervention within an academic setting.

The inclusion criterion was successful completion of previous subjects related to this course (such as anatomy), and the exclusion criterion was the presence of vision and/or mobility difficulties in the upper limbs.

The on-site group received the basic musculoskeletal ultrasound training provided each year as part of the General Procedures II and Assessment in Physiotherapy courses through face-to-face lectures followed by practical sessions using ultrasound scanners at a ratio of 1 ultrasound scanner for every 2 students, with teacher support provided in small groups of 12 students. The online group was given access to the online ultrasound training platform without the opportunity to ask questions to instructors or handle an ultrasound scanner until 2 hours before the examination, when they were allowed contact with the device.

For the evaluation of the competencies acquired, all students took the same tests. The study variables were extracted from these tests, consisting of two parts: (1) theoretical (10 multiple-choice questions with 4 answers, only 1 of which was valid), and (2) practical examination to identify an anatomical structure (selected at random) using ultrasound in a partner (selected at random) where the examiner collected the following items on a rubric (success or failure, time to recognize it, position of the assessor, position of the patient, skill in handling the probe, and image optimization). Measurements in a training phantom (Venipuncture IV Injection Training Pad Model, VTurboWay, Ltd) of the intragel structures (structure diameter and structure surface distance) were taken with the ultrasound calipers.

The examiners who assessed the students were different from the investigator who performed the first randomization for online and on-site assignment to avoid bias at this point and differed from the instructor who taught the face-to-face classes. The phantom measurements were taken by an ultrasound expert on the same day prior to the student assessments to provide a reference standard for subsequent comparative analysis.

Statistical analysis was performed with R software (version 4.1.3: R Foundation for Statistical Computing). The significance level was set at $P<.05$. The distribution of the quantitative variables was tested with the Shapiro-Wilk test. Quantitative variables are expressed as mean (SD) and qualitative variables as absolute and relative values (%).

XSL•FO
RenderX

The presence of significant differences in the responses between the online and on-site groups was verified using Fisher exact test in the case of qualitative variables, and with the Mann-Whitney *U* test or 2-tailed Student *t* test for independent samples (after checking the assumption of homogeneity of variances with the Levene test) in the quantitative variables based on their distribution.

For structure diameter and structure surface distance measurements, the intraclass correlation coefficient (ICC [3,1]) was calculated between student measurements and the reference values using a 2-way mixed single-measure model (consistency and absolute agreement). Relative reliability was defined as poor (<0.5), moderate (0.5-0.75), good (0.75-0.9), or excellent (>0.9). As absolute reliability, the SEM was calculated. The minimum detectable change was also calculated.

IRT analysis using a Rasch model was conducted to estimate item difficulty and student ability. Detailed assumptions, model fit procedures, and statistical tests are provided in Multimedia Appendix 1.

To minimize potential bias related to conflicts of interest, several safeguards were implemented. Group allocation was performed by an independent researcher not involved in teaching or assessment, ensuring allocation concealment. Examiners responsible for evaluating both theoretical and practical performance were independent of the intervention and blinded to group assignment. These examiners were not involved in the development of the platform and had no access to allocation data.

All assessments were conducted under standardized conditions using identical evaluation criteria for both groups. Although participants could not be blinded due to the nature of the intervention, no explicit indicators of group allocation were provided to examiners during the evaluation process.

Ethical Considerations

The study was approved by the Vicerrectorado de Innovación Docente from the Universidad de Alcalá (UAHEV/1434). An informed consent form was signed by all participants. In addition, data were analyzed after anonymization to further reduce the risk of bias.

Results

Overview

The sample consisted of 68 students, mostly women (n=42, 61.8%). No significant differences were observed in the demographic characteristics and examination outcomes between the groups (Table 1).

In the analysis of the structure diameter and structure surface distance vs reference measurements, the ICC (3,1) was not significant with poor values and high and precise SEMs (Table 2).

XSL•FO
RenderX

Table . Demographic characteristics and examination outcomes of the participants.

Variables	Overall (n=68)	Online (n=34)	On-site (n=34)	*P* value
Sex, n (%)				>.99
Female	42 (61.8)	21 (61.8)	21 (61.8)	
Male	26 (38.2)	13 (38.2)	13 (38.2)	
Patient position, n (%)				.37
Correct	56 (82.4)	27 (79.4)	29 (85.3)	
Incorrect	7 (10.3)	3 (8.8)	4 (11.8)	
Missing	5 (7.4)	4 (11.8)	1 (2.9)	
Sonographer position, n (%)				.24
Correct	62 (91.2)	30 (88.2)	32 (94.1)	
Incorrect	1 (1.5)	0 (0)	1 (2.9)	
Missing	5 (7.4)	4 (11.8)	1 (2.9)	
Probe handling, n (%)				.37
Correct	54 (79.4)	26 (76.5)	28 (82.4)	
Incorrect	9 (13.2)	4 (11.8)	5 (14.7)	
Missing	5 (7.4)	4 (11.8)	1 (2.9)	
Image optimization, n (%)				.31
Correct	17 (25)	7 (20.6)	10 (29.4)	
Incorrect	46 (67.6)	23 (67.6)	23 (67.6)	
Missing	5 (7.4)	4 (11.8)	1 (2.9)	
Structure identification, n (%)				.19
Correct	46 (67.6)	24 (70.6)	22 (64.7)	
Incorrect	17 (25)	6 (17.6)	11 (32.4)	
Missing	5 (7.4)	4 (11.8)	1 (2.9)	
Test duration (seconds), mean (SD)	101.44 (66.41)	102.52 (61.55)	100.46 (71.48)	.90
Structure diameter (mm), mean (SD)	3.88 (1.06)	3.78 (0.79)	3.98 (1.27)	.46
Structure surface distance (mm), mean (SD)	3.57 (1.47)	3.94 (1.97)	3.24 (0.64)	.06
Theoretical exam score, mean (SD)	5.01 (1.30)	4.94 (1.47)	5.08 (1.14)	.66

XSL•FO
RenderX

Table . Interobserver reliability between student and reference measurements.

Group and variable	ICC[a] (95% CI)	Mean (SD; 95% CI)	SEM[b] (95% CI)	MDC[c]
Online				
Structure diameter (mm)	0.253 (−0.052 to 0.514)	3.771 (0.606; 3.59-3.951)	0.537 (0.408-0.666)	1.488
Structure surface distance (mm)	0 (−0.301 to 0.301)	4.304 (1.453; 3.923-4.684)	1.457 (1.057-1.856)	4.038
On-site				
Structure diameter (mm)	0.009 (−0.278 to 0.295)	3.863 (0.923; 3.629-4.097)	0.923 (0.712-1.134)	2.558
Structure surface distance (mm)	0 (−0.287 to 0.287)	3.959 (0.921; 3.84-4.078)	0.67 (0.454-0.885)	1.857

[a]ICC: intraclass correlation coefficient.

[b]SEM: SE of measurement.

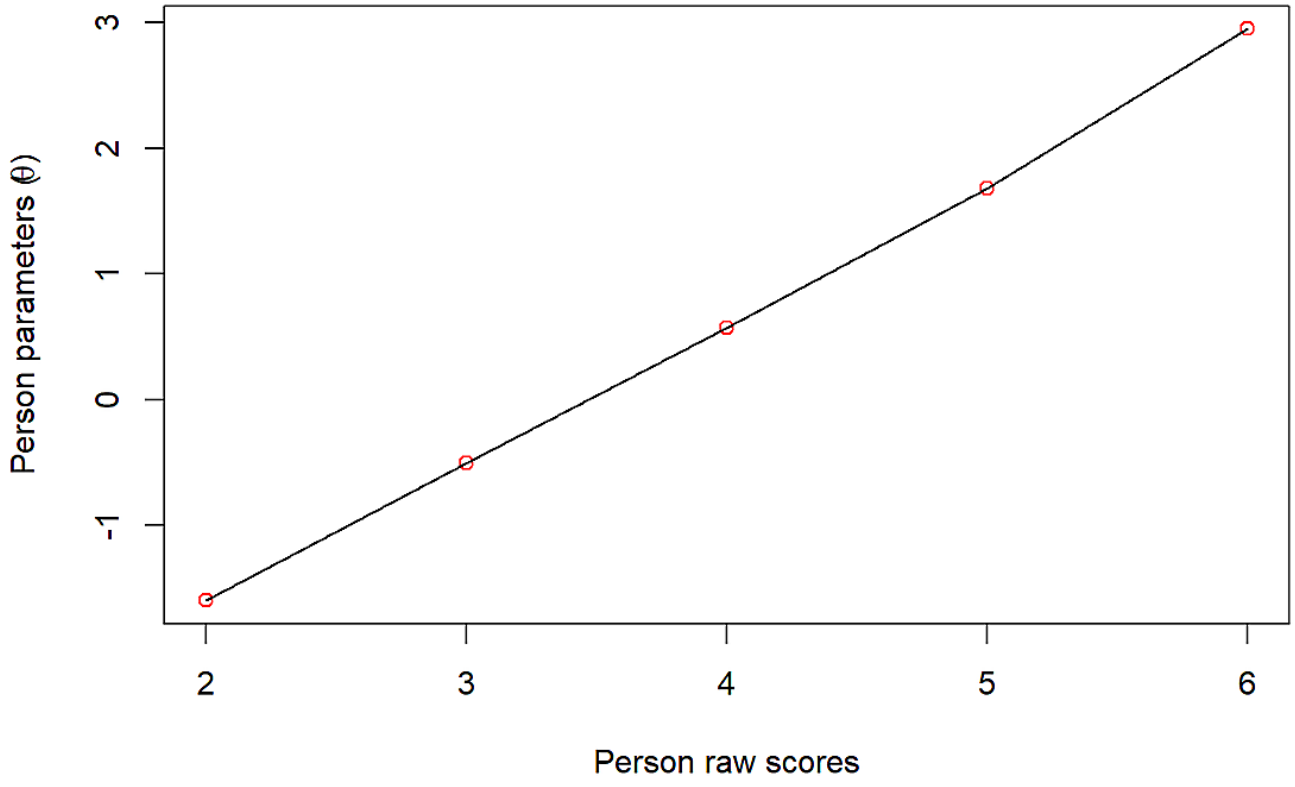
[c]MDC: minimal detectable change.

IRT Analysis

Overview

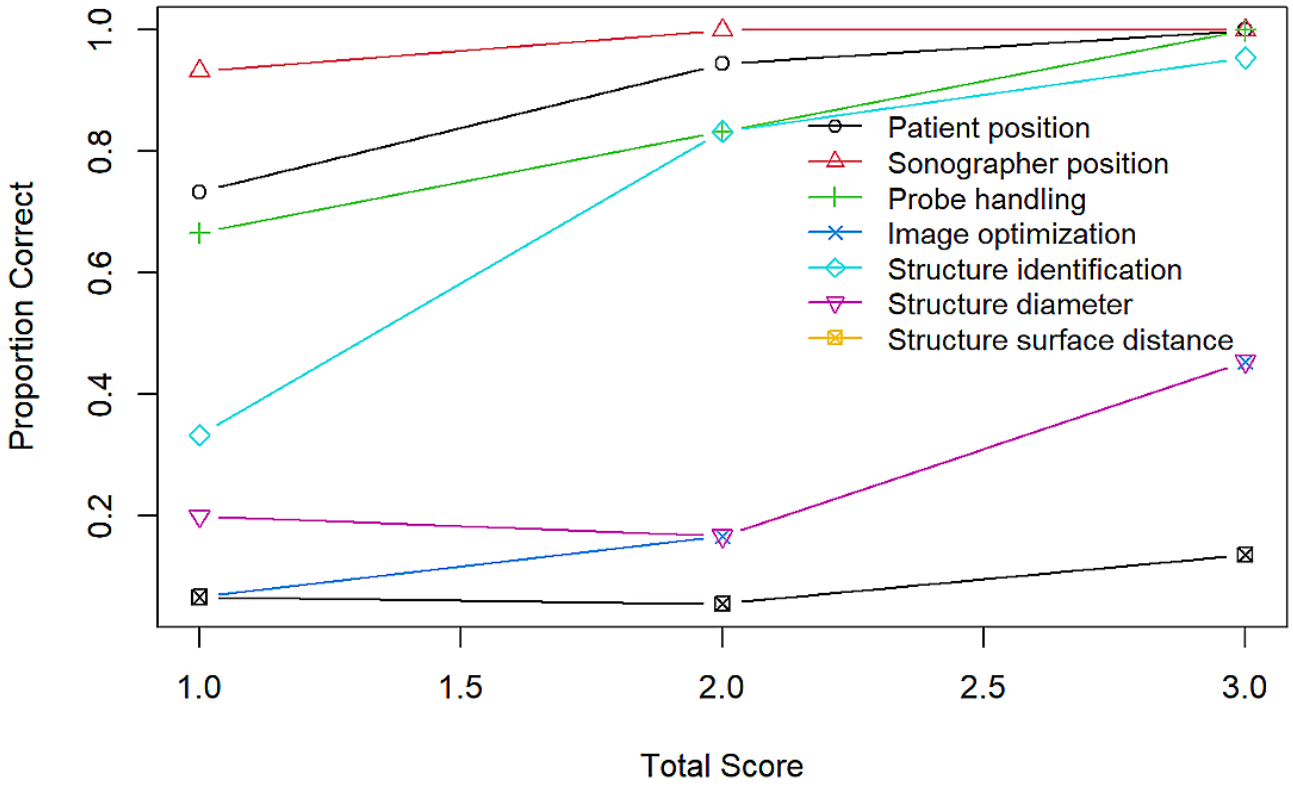
The average level of success of the items was 59.637%, with the questions on image optimization, structure diameter, and structure surface distance having a hit rate of less than 50%. A total of 15 students answered 3 questions correctly, 18 answered 4 questions correctly, and 22 answered 5 questions correctly (Figure 1). Only 5 students answered 6 questions correctly and 3 students answered 2 questions correctly; there were no students who answered all the questions correctly or who failed all of them. It can be seen that in the questions on patient position, sonographer position, probe handling, image optimization, and structure identification, the proportion of correct answers increases as the final score obtained increased, while in the questions on structure diameter and structure surface distance, the proportion of hits decreased up to a score of 2 and then increased slightly up to the maximum score. The probability of correct responses increased consistently with total score for most items, supporting the expected hierarchical structure of the assessment (Figure 2).

Figure 1. Plot of person parameters showing the monotonic mapping of total raw scores to the estimated latent ability parameters. Higher raw scores correspond to higher levels of latent competence, providing a linear-like transformation within the observed score range (2-6).



XSL•FO
RenderX

Figure 2. Test information plot showing the proportion of correct responses for each assessment criterion across overall score levels (1.0-3.0). Steeper lines (eg, structure identification) indicate criteria with high discriminative power. Lines near the top (eg, sonographer position) represent easier tasks, whereas lines near the bottom (eg, structure surface distance) represent highly difficult tasks across all proficiency levels.



Model Assumptions

The nonsignificant T1m statistic in all pairs of questions (P>.05) indicated that the practical examination was 1D. Pairwise analysis of item responses (Table S1 in Multimedia Appendix 2) showed no significant violations of unidimensionality, with only a minimal proportion of item pairs displaying residual dependence. The distribution of person fit statistics also showed acceptable fit across participants (Figure 3). The nonsignificant Andersen likelihood ratio test (χ^2_3=1.2; P=.75) confirmed the invariance of the model parameters (Figure 4). Regarding local independence, Ponocny T1 statistic revealed that only the pair "structure identification–patient position" showed a significant residual correlation (P=.02). This dependency represented only 4.76% of the total item pairs, a figure within the expected margin of error (<5%) that did not compromise the model's unidimensionality. This specific correlation was likely due to the procedural nature of the ultrasound examination, in which proper patient positioning serves as a technical prerequisite for successful structure identification. Therefore, this minor violation was considered clinically coherent and statistically negligible for the overall validity of the trait estimates (θ).

XSL•FO
RenderX

Figure 3. Person vs infit t-statistic plot displaying person ability parameters on the vertical axis and their respective infit t-statistics on the horizontal axis. The green vertical lines at −2 and 2 represent the thresholds for acceptable item fit.

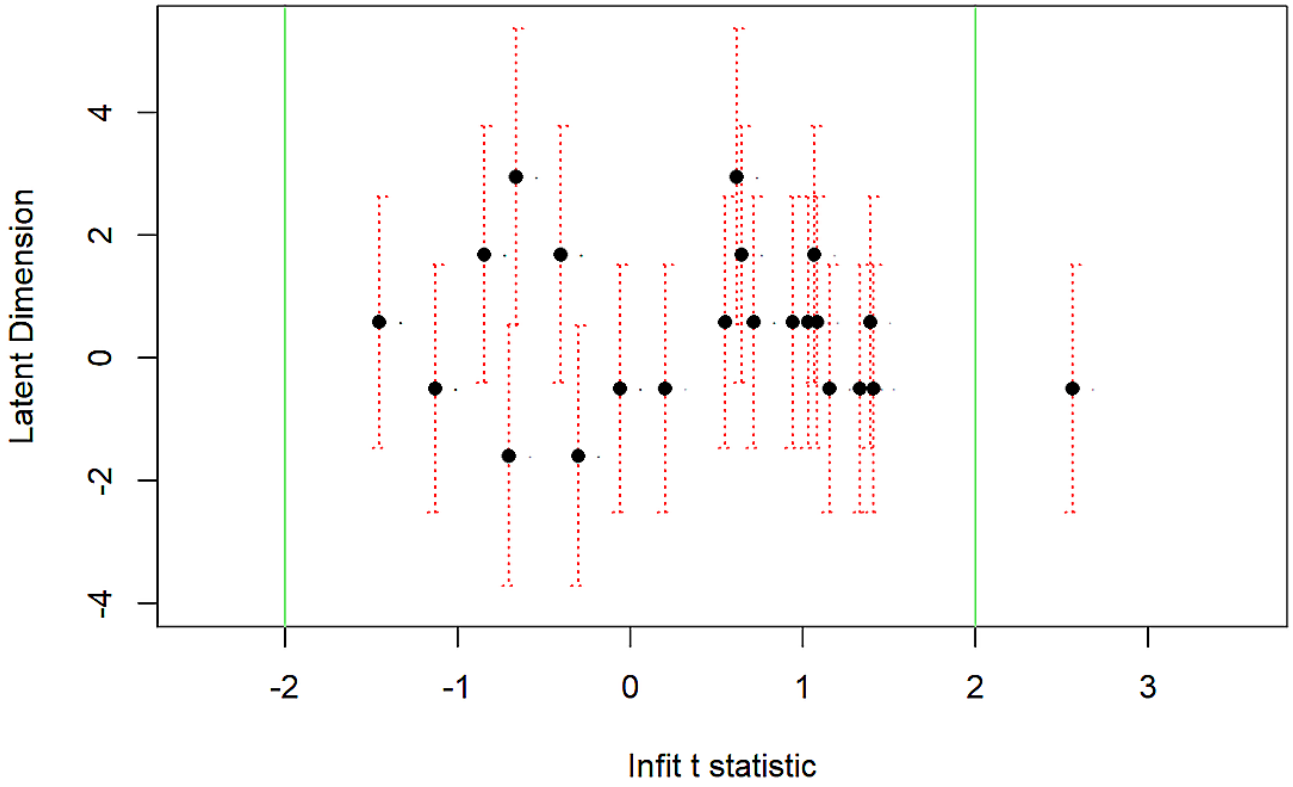
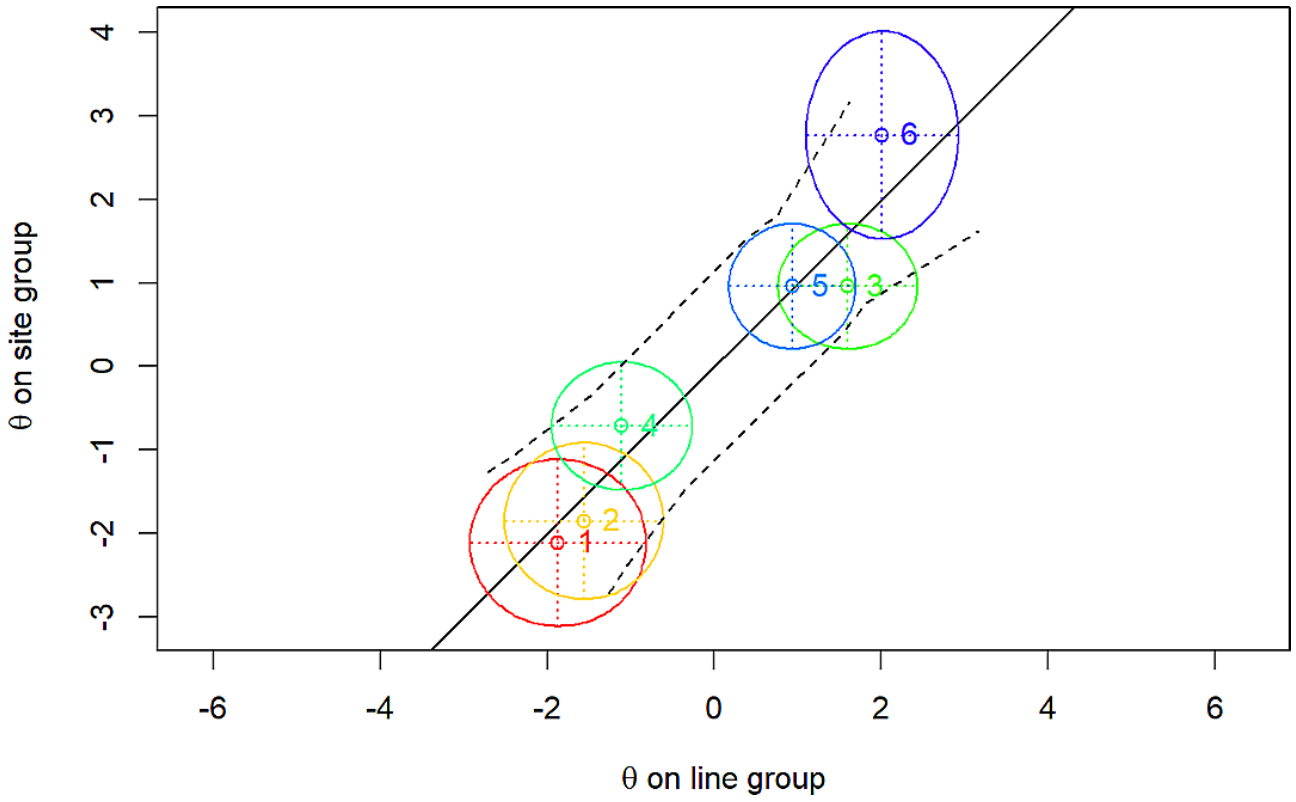


Figure 4. Goodness of fit plot. The numbered colored circles (1-6) represent the location parameters (difficulty) of the 6 items, with circle sizes reflecting the confidence ellipses. The solid diagonal line represents perfect invariance (equal difficulty across groups). The dashed reference lines indicate the 95% confidence bounds. Because all item circles either cross or fall within these dashed boundaries, the item parameters are considered invariant, indicating no significant differential item functioning between the online and onsite student groups.



XSL•FO
RenderX

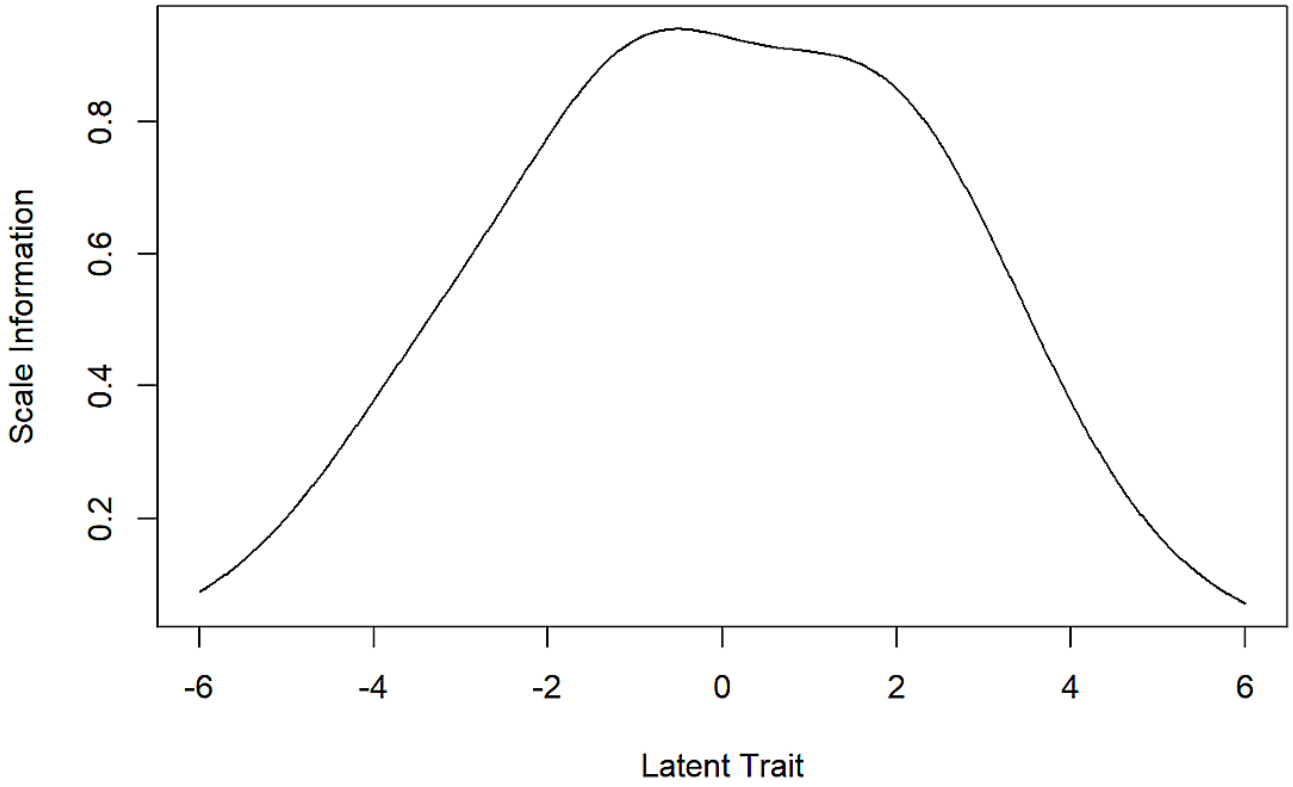
Item Parameters

The examination presented an average level of difficulty δ of 0 (SD 2.213) with a higher difficulty level in the questions on image optimization, structure diameter, and structure surface distance and lower for the questions on patient position, sonographer position, and probe handling (Table 3). It was observed that the test provided the maximum information with the minimum error at levels of θ around −2 to 2 (Figure 5). A clear division of the items was observed according to the level of difficulty; the easiest questions were on patient position and sonographer position; the questions with an intermediate level of difficulty were on probe handling and structure identification; and the most difficult questions were on image optimization, structure diameter, and structure surface distance. The Wright map (Figure 6) illustrated a clear separation between lower- and higher-difficulty items, confirming the hierarchical progression of skill acquisition. The item characteristic curves for the seven assessment criteria are presented in Figure 7, showing the observed response patterns and the model-predicted probabilities across the latent ability continuum.

Table . Item difficulty parameters.

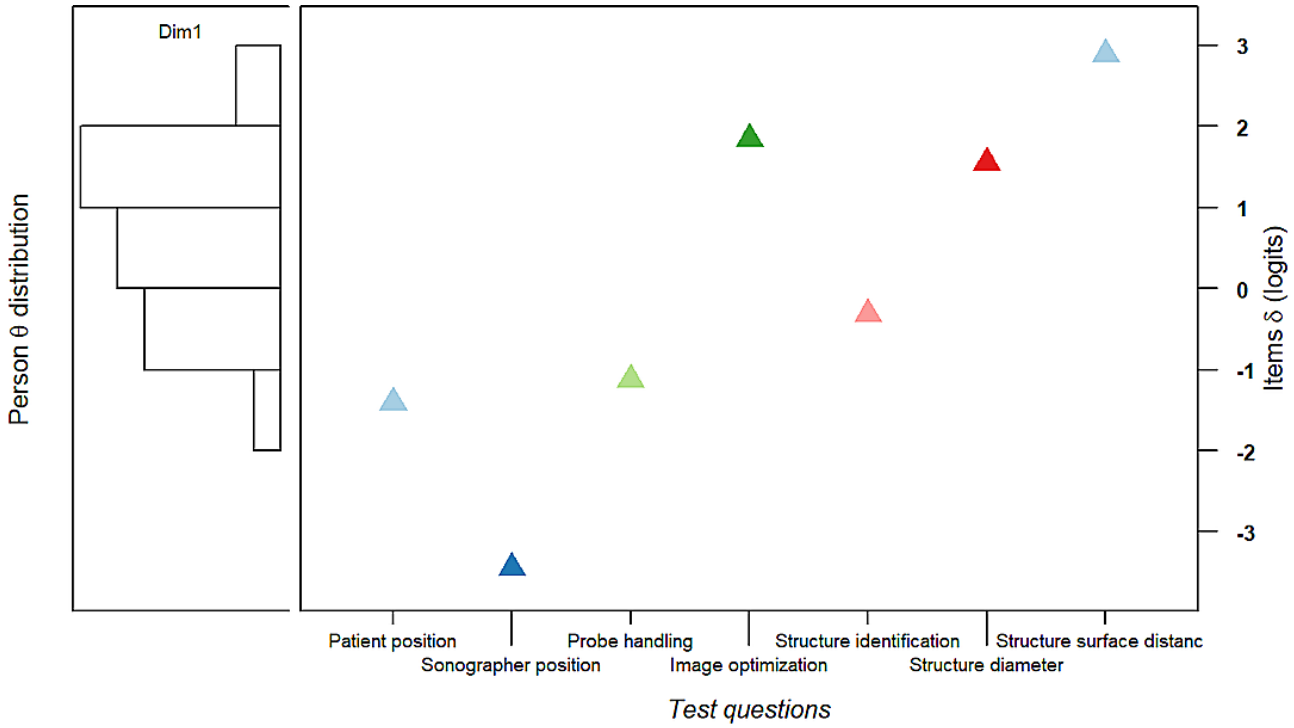
Item	Difficulty (SE; 95% CI)
Patient position	−1.407 (0.395; −2.181 to −0.634)
Sonographer position	−3.446 (0.876; −5.163 to −1.729)
Probe handling	−1.121 (0.366; −1.837 to −0.405)
Image optimization	1.845 (0.319; 1.22 to 2.469)
Structure identification	−0.312 (0.314; −0.928 to 0.303)
Structure diameter	1.548 (0.308; 0.944 to 2.152)
Structure surface distance	2.894 (0.399; 2.113 to 3.675)

Figure 5. Test information plot illustrating the amount of psychometric information provided by the test across different levels of studentability (latent trait). The curve peaks and remains high between approximately −2 and +2, indicating that the assessment instrument achievesmaximum measurement precision and reliability when evaluating students with average or moderate proficiency levels.



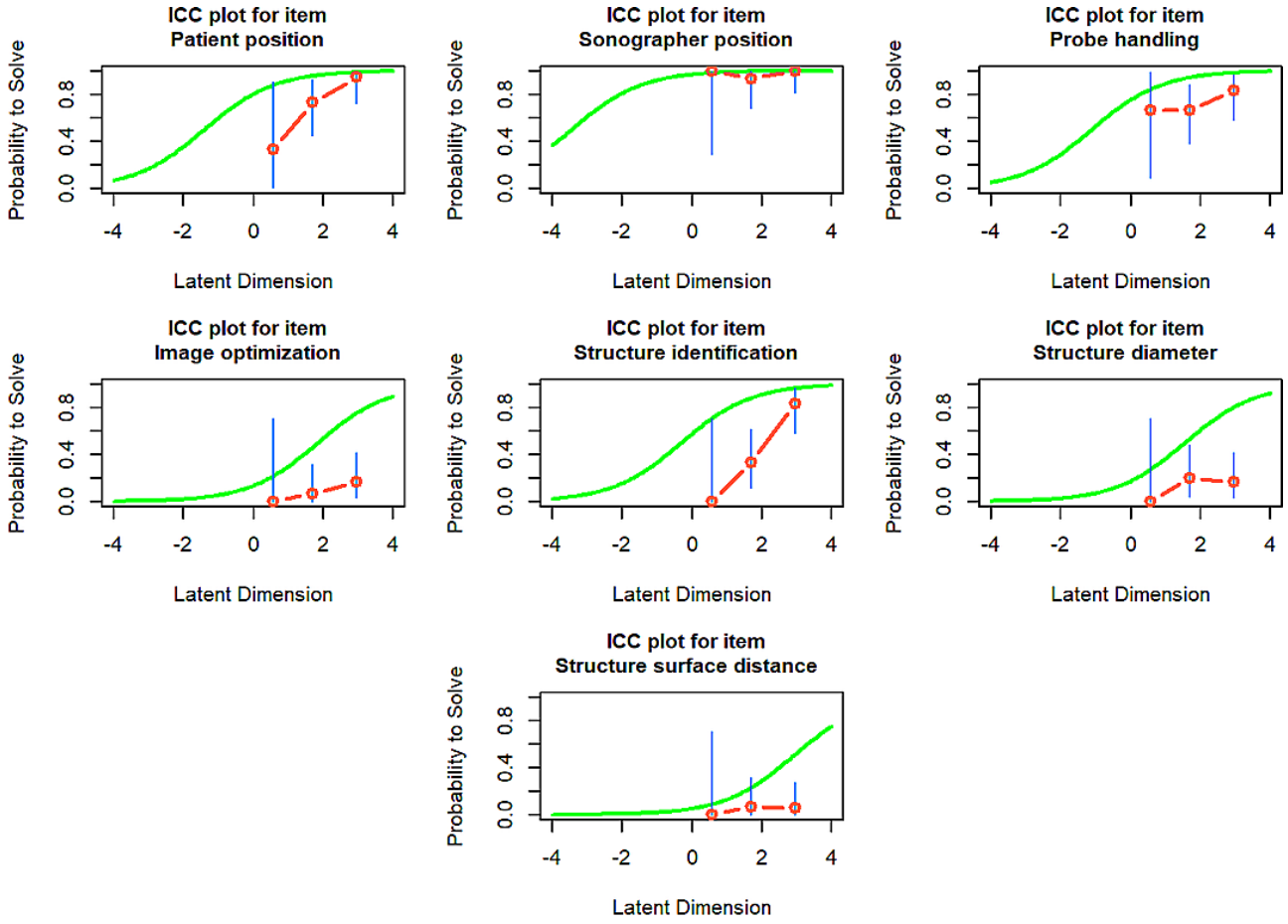
XSL•FO
RenderX

Figure 6. Wright map. The left panel displays a histogram of the participants' estimated latent proficiency (θ distribution) along the primary latentdimension (Dim1). The right panel plots the difficulty parameters (δ in logits) of the 7 assessment criteria, where each colored triangle representsa specific test question or criterion. Both panels share the central logit scale. Items located higher on the vertical axis (eg, "structure surfacedistance") represent greater difficulty, whereas triangles near the bottom (eg, "sonographer position") indicate easier tasks. This alignment allows direct evaluation of targeting, demonstrating that the test items cover a broad range of the students' proficiency distribution.



XSL•FO
RenderX

Figure 7. Item characteristic curves. The green solid lines represent the theoretical intraclass correlation coefficient (ICCs; model-predicted probability of a correct response across the latent dimension). The red squares and dashed lines show the observed empirical proportions for each score group, with blue vertical lines indicating the 95% CIs. Closer alignment between the empirical red points and the theoretical green curve reflects better fit of the item to the measurement model.



Discussion

Principal Findings

This randomized study found that simulation-based online ultrasound training produced comparable theoretical and practical outcomes to traditional on-site instruction. IRT analysis revealed a structured hierarchy of item difficulty and identified key competencies for assessment, while both training modalities showed limited reliability relative to expert performance, reflecting the early stage of skill acquisition.

The theoretical and practical assessment outcomes revealed comparable knowledge acquisition across both groups, with no statistically significant differences between the online and on-site training modalities. Notably, the online group demonstrated nearly half the error rate in image identification compared with the on-site cohort, suggesting enhanced visual-cognitive integration when using simulation-based e-learning.

A noteworthy finding was the poor absolute reliability (ICC) and the relatively high SEM and minimal detectable change values observed in both groups. We contextualize these results as a transparent reflection of the learning curve and the expert-novice gap inherent in high-precision technical measurements.

Because the ICC compared student performance against a gold standard expert, these low values underscore that a single training session—regardless of the modality (online or on-site)—is insufficient for achieving clinical-grade proficiency. These data are crucial for future pedagogical designs, as they highlight that while online training might slightly reduce measurement error (SEM), both formats require extended supervised practice to transition from conceptual understanding to technical mastery.

The IRT analysis enabled a detailed understanding of item difficulty within the practical assessment. The practical implications of these findings are 2-fold. First, the item difficulty hierarchy ( ) provides a roadmap for curriculum design: training should prioritize “image optimization” and “metric precision” (diameter and distance), as these were identified as the most challenging competencies to acquire (>1.5). Second, the test information function reveals that the current assessment is highly reliable for discriminating between novice and intermediate students (latent trait between −2 and +2) but may require more challenging items to effectively assess advanced practitioners. Finally, the disparity between observed responses and the ICC curves (Figure 7) in high-difficulty items suggests that metric accuracy is a specific technical bottleneck that persists despite general procedural knowledge, highlighting the need for targeted feedback in ultrasound-guided measurements.

XSL•FO
RenderX

The insights provided by the IRT model allow for a data-driven optimization of the training program, focusing on 3 key areas.

First, regarding curriculum design, because "sonographer position" and "patient position" are low-difficulty items (<−1.4), they should be introduced as foundational self-learning modules. This allows classroom or simulation time to be shifted toward high-difficulty tasks such as "structure surface distance" (=2.89), which require expert-led, hands-on guidance.

Second, assessment weighting and IRT difficulty parameters justify a nonlinear scoring system. We suggest that "image optimization" and "metric measurements" should carry higher weight in final assessments, as they are the primary discriminators of technical proficiency, whereas "patient positioning" serves only as a prerequisite.

Third, regarding targeted skill remediation, the item characteristic curves identified specific failure points. For instance, students who failed "structure identification" should not proceed to "structure diameter" training, as the model confirmed a hierarchical dependency. Remediation can thus be personalized based on each student's position on the latent trait (θ) scale.

No significant differences were found between online and on-site groups in item difficulty patterns. Although ICCs between the 2 groups were nonsignificant and associated with relatively high SEMs, the online group demonstrated slightly better reliability, a finding consistent with previous research [12].

Prior studies have investigated the feasibility of online ultrasound education across various health professions, including obstetrics [9], anesthesia [10], physiotherapy [11,12], and postgraduate echocardiography [13]. These works generally reported equivalent or superior outcomes for knowledge acquisition and image interpretation in comparison to traditional face-to-face methods. These findings are consistent with this body of evidence, supporting the educational validity of simulation-based e-learning for ultrasound training.

However, few previous investigations have objectively assessed psychomotor and image acquisition skills using psychometrically robust frameworks. To our knowledge, no earlier ultrasound education studies have used an IRT model to examine item difficulty and learner ability distribution. This methodological innovation provides a degree of measurement precision rarely applied outside high-stakes simulation domains, such as aviation [20]. Consequently, our results extend the current understanding of how simulation-based training can be objectively evaluated and standardized in medical education.

The findings suggest that digital simulation tools such as the platform can democratize ultrasound training by increasing accessibility and scalability, enabling a broader range of health care professionals to achieve comparable levels of competence. As highlighted by Einstein's well-known maxim, "doing the same thing and expecting different results" limits innovation [21]; the integration of digital solutions into medical curricula represents a necessary evolution aligned with modern health care and societal needs.

IRT-derived ability mapping indicated that most students clustered at ability levels 1 to 2 within a –3 to +3 range , a valuable reference point for designing adaptive assessment systems. Future studies may leverage these psychometric profiles to personalize training progression and feedback in simulation-based e-learning environments.

Finally, previous evidence demonstrated that online experiential learning models offer significant advantages in efficiency and cost reduction [8]. Combined with comparable learning outcomes, these results reinforce simulation-based digital education as a sustainable and effective approach for future ultrasound and clinical skills training.

Because this study was not prospectively registered, transparency may be limited. Future studies should ensure prior trial registration, use a multi-institution design, and examine a broader cohort.

Conclusions

Online ultrasound training provides the same theoretical and practical results as face-to-face training, with all the economic, academic, and professional benefits that this entails. The most recommendable questions ("probe handling" and "structure identification") for the evaluation of this training have been identified. It will be interesting in future studies to evaluate the capacity achieved according to the number of hours, as well as to conduct a cost-effectiveness study.

This study provides a preliminary validation of a practical assessment tool for ultrasound-guided procedures using IRT. The application of the 1-parameter Rasch model proved to be a robust methodological choice for this cohort (n=68); its mathematical parsimony allowed for stable and reliable estimates of item difficulty and student proficiency even in samples of modest size, where more complex models might falter [22,23]. The nonsignificant Andersen likelihood ratio test and the low SEs further reinforced the internal validity of the findings within this specific educational context. However, given the single-site nature of the study and the specific sample size, these results should be interpreted as a foundational pilot validation. While the identified difficulty hierarchy offers a clear road map for curriculum design and targeted remediation, the generalizability and global scalability of the instrument remain to be confirmed. Future research should focus on multicenter studies with larger, more diverse populations to establish definitive normative data and ensure cross-cultural parameter invariance. In conclusion, this work represents a significant step toward a data-driven, standardized evaluation in ultrasound training, providing a reliable framework for initial competency assessment.

XSL•FO
RenderX

Acknowledgments

The authors thank the participating physiotherapy students and the faculty members who facilitated data collection. The authors also thank Dr Josue Fernández-Carnero from King Juan Carlos University for his helpful contributions to this publication.

Funding

This research received no external funding. The study was conducted without financial support from public agencies, commercial entities, or nonprofit organizations.

Authors' Contributions

Conceptualization: SF-C
Data curation: BD-P, DP-M, SG-M, AA-O, NC-Z, FN-C Sr
Formal analysis: BD-P, DP-M, SG-M, AA-O, NC-Z, FN-C Sr
Funding acquisition: SF-C, BD-P, JM-R, DP-M, SG-M, AA-O, NC-Z, FN-C Sr
Investigation: BD-P, DP-M, SG-M, AA-O, NC-Z, FN-C Sr
Methodology: SF-C, BD-P, JM-R, DP-M, SG-M, AA-O, NC-Z, FN-C Sr
Project administration: SF-C, BD-P, JM-R, DP-M, SG-M, AA-O, NC-Z, FN-C Sr
Resources: SF-C, BD-P, JM-R, DP-M, SG-M, AA-O, NC-Z, FN-C Sr
Software: SF-C, BD-P, JM-R, DP-M, SG-M, AA-O, NC-Z, FN-C Sr
Supervision: SF-C, BD-P, JM-R, DP-M, SG-M, AA-O, NC-Z, FN-C Sr
Validation: SF-C, BD-P, JM-R, DP-M, SG-M, AA-O, NC-Z, FN-C Sr
Visualization: SF-C, BD-P, JM-R, DP-M, SG-M, AA-O, NC-Z, FN-C Sr
Writing—original draft: BD-P, DP-M, SG-M, AA-O, NC-Z, FN-C Sr
Writing—review and editing: SF-C, BD-P, JM-R, DP-M, SG-M, AA-O, NC-Z, FN-C Sr
All authors read and approved the published version of the manuscript.

Conflicts of Interest

The authors declare no commercial, financial, or personal relationships that could be perceived as potential conflicts of interest. The platform provided access to the digital ultrasound simulation platform used in this study; however, the company had no role in the study design, data collection, statistical analysis, interpretation of the results, or the decision to submit the manuscript for publication. No author received financial compensation for participating in this research. Several authors participated in the development of the ultrasound simulation tool. To minimize bias, all researchers involved in its design or programming were blinded to group allocation, performance metrics, and outcome data until all analyses were completed.

Multimedia Appendix 1
Detailed item response theory analysis and model diagnostics.
[DOCX File, 24 KB - mededu_v12i1e87897_app1.docx]

Multimedia Appendix 2
Psychometric analyses and graphical outputs from the item response theory (IRT) Rasch model, including item characteristic curves, Wright maps, goodness-of-fit plots, and test information functions comparing online and on-site ultrasound training groups.
[DOCX File, 31 KB - mededu_v12i1e87897_app2.docx]

References

1. Minimum training requirements for the practice of medical ultrasound in Europe. European Federation of Societies for Ultrasound in Medicine and Biology. URL: http://s834315022.websitehome.co.uk/wp-content/uploads/2020/12/2009-04-14apx12.pdf [accessed 2026-5-27]
2. Montesinos L, Salinas-Navarro DE, Santos-Diaz A. Transdisciplinary experiential learning in biomedical engineering education for healthcare systems improvement. BMC Med Educ 2023 Apr 3;23(1):207. [doi: 10.1186/s12909-023-04171-x] [Medline: 37013525]
3. Andersen CA, Holden S, Vela J, Rathleff MS, Jensen MB. Point-of-care ultrasound in general practice: a systematic review. Ann Fam Med 2019 Jan;17(1):61-69. [doi: 10.1370/afm.2330] [Medline: 30670398]
4. Chawla TP, Cresswell M, Dhillon S, et al. Canadian Association of Radiologists position statement on point-of-care ultrasound. Can Assoc Radiol J 2019 Aug;70(3):219-225. [doi: 10.1016/j.carj.2019.06.001] [Medline: 31376885]

XSL•FO
RenderX

5. Nakashima H, Yukawa Y, Suda K, Yamagata M, Ueta T, Kato F. Abnormal findings on magnetic resonance images of the cervical spines in 1211 asymptomatic subjects. Spine (Phila Pa 1976) 2015 Mar 15;40(6):392-398. [doi: 10.1097/BRS.0000000000000775] [Medline: 25584950]
6. Bacon DR, Cowles K, Thapa D, et al. Creating an ultrasound scholarly concentration program for medical students. Adv Med Educ Pract 2021;12:1103-1110. [doi: 10.2147/AMEP.S330771] [Medline: 34594147]
7. Sommer KJ. Pilot training: what can surgeons learn from it? Arab J Urol 2014 Mar;12(1):32-35. [doi: 10.1016/j.aju.2013.08.011] [Medline: 26019919]
8. Maloney S, Haas R, Keating JL, et al. Breakeven, cost benefit, cost effectiveness, and willingness to pay for web-based versus face-to-face education delivery for health professionals. J Med Internet Res 2012 Apr 2;14(2):e47. [doi: 10.2196/jmir.2040] [Medline: 22469659]
9. Shaw‐Battista J, Young‐Lin N, Bearman S, Dau K, Vargas J. Interprofessional obstetric ultrasound education: successful development of online learning modules; case‐based seminars; and skills labs for registered and advanced practice nurses, midwives, physicians, and trainees. J Midwife Womens Health 2015 Dec;60(6):727-734. [doi: 10.1111/jmwh.12395] [Medline: 26769384]
10. Krishnan S, Kuhl T, Ahmed W, Togashi K, Ueda K. Efficacy of an online education program for ultrasound diagnosis of pneumothorax. Anesthesiology 2013 Mar;118(3):715-721. [doi: 10.1097/ALN.0b013e31827f0979] [Medline: 23291625]
11. Fernández-Lao C, Cantarero-Villanueva I, Galiano-Castillo N, Caro-Morán E, Díaz-Rodríguez L, Arroyo-Morales M. The effectiveness of a mobile application for the development of palpation and ultrasound imaging skills to supplement the traditional learning of physiotherapy students. BMC Med Educ 2016 Oct 19;16(1):274. [doi: 10.1186/s12909-016-0775-1] [Medline: 27756288]
12. Fernández-Carnero S, Cuenca-Zaldivar JN, Pecos-Martin D, Achalandabaso-Ochoa A, Ferragut-Garcias A, Gallego-Izquierdo T. Validity and reliability of methods for sonography education in physiotherapy: onsite vs. online. Appl Sci 2020;10(18):6456. [doi: 10.3390/app10186456]
13. Vegas A, Meineri M, Jerath A, Corrin M, Silversides C, Tait G. Impact of online transesophageal echocardiographic simulation on learning to navigate the 20 standard views. J Cardiothorac Vasc Anesth 2013 Jun;27(3):531-535. [doi: 10.1053/j.jvca.2012.09.006] [Medline: 23127696]
14. DeMars C. Item Response Theory: Oxford University Press; 2010.
15. Kruyen PM. Using short tests and questionnaires for making decisions about individuals: when is short too short [PhD thesis]. : Tilburg University; 2012 URL: https://research.tilburguniversity.edu/en/publications/using-short-tests-and-questionnaires-for-making-decisions-about-i/ [accessed 2026-05-14]
16. Hambleton RK. Item response theory: introduction and bibliography. Psicothema 1990;2:97-107 [FREE Full text]
17. Ponocny I. Nonparametric goodness-of-fit tests for the Rasch model. Psychometrika 2001 Sep;66:437-459. [doi: 10.1007/BF02294444]
18. WazO. URL: https://wazotechnologies.com/ [accessed 2023-07-05]
19. Moher D, Schulz KF, Altman DG, Lepage L. The CONSORT statement: revised recommendations for improving the quality of reports of parallel-group randomised trials. Lancet 2001 Apr 14;357(9263):1191-1194. [Medline: 11323066]
20. Arjoni DH, de Souza Rehder I, Pereira Figueira JM, Villani E. Augmented reality for training formation flights: an analysis of human factors. Heliyon 2023 Mar;9(3):e14181. [doi: 10.1016/j.heliyon.2023.e14181] [Medline: 36950645]
21. Noseworthy T. A flight simulator for careers in health services and policy research. Healthc Policy 2019 Oct;15(SP):6-9. [doi: 10.12927/hcpol.2019.25984] [Medline: 31755855]
22. Linacre JM. Sample size and item calibration [or person measure] stability. Rasch.org. URL: https://www.rasch.org/rmt/rmt74m.htm [accessed 2026-05-14]
23. Stone M, Yumoto F. The effect of sample size for estimating Rasch/IRT parameters with dichotomous items. J Appl Meas 2004;5(1):48-61. [Medline: 14757991]

Abbreviations

CONSORT: Consolidated Standards of Reporting Trials
ICC: intraclass correlation coefficient
IRT: item response theory
SEM: standard error of measurement

XSL•FO
RenderX

Edited by LT Car; submitted 17.Nov.2025; peer-reviewed by P Adedigba, R Tezza; revised version received 21.Mar.2026; accepted 22.Apr.2026; published 02.Jul.2026.

Please cite as:
Fernández-Carnero S, Diaz-Pulido B, Méndez-Rodriguez J, Pecos-Martin D, Garcia-Miguel S, Achalandabaso-Ochoa A, Cuenca-Zaldívar N, Naranjo-Cinto Sr F
Digital Simulation–Based Ultrasound Training for Physiotherapy Students: Prospective Randomized Controlled Trial
JMIR Med Educ 2026;12:e87897
URL: https://mededu.jmir.org/2026/1/e87897
doi:10.2196/87897

©Samuel Fernández-Carnero, Belen Diaz-Pulido, Jorge Méndez-Rodriguez, Daniel Pecos-Martin, Santiago Garcia-Miguel, Alexander Achalandabaso-Ochoa, Nicolas Cuenca-Zaldívar, Fermín Naranjo-Cinto Sr. Originally published in JMIR Medical Education (https://mededu.jmir.org), 2.Jul.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Education, is properly cited. The complete bibliographic information, a link to the original publication on https://mededu.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Combining Problem-Based Learning Methods With the WeChat Platform in Teaching Ophthalmology: Randomized Controlled Trial

Fang Fang, MD, PhD; Bing Bu, BM; Wenmin Jiang, MD, PhD

Department of Ophthalmology in the Second Xiangya Hospital, Central South University, 139 Renmin Road, Changsha, China

Corresponding Author:
Wenmin Jiang, MD, PhD
Department of Ophthalmology in the Second Xiangya Hospital, Central South University, 139 Renmin Road, Changsha, China

Abstract

Background: Ophthalmology poses distinct learning challenges for medical students due to the complex anatomy of the eye and the requirement of essential hands-on skills. Problem-based learning (PBL), a student-centered approach, fosters clinical reasoning and self-directed learning. To address the time and logistical constraints of traditional teaching methods, this study implemented a WeChat-based PBL model that leveraged the platform's efficiency and interactivity to enhance student engagement and skill acquisition in ophthalmology.

Objective: This study aimed to evaluate the effectiveness of WeChat-based PBL in ophthalmology education, focusing on its impact on students' self-perception of learning and clinical skills compared to traditional teaching methods.

Methods: This study involved 108 undergraduate students who successfully passed the Chinese National Entrance Examination. Among them, 54 (50%) were randomly selected to participate in the WeChat-based PBL, while the other 54 (50%) received traditional teaching. Students were placed into 6 groups (18 students for each group) using a random number table, and the new teaching methods were tested outside their regular class time. Three groups were randomly selected to receive PBL using WeChat as the platform, while the remaining 3 groups received conventional teaching.

Results: Our analysis indicated that although students in the WeChat-based PBL group scored marginally lower in memorization compared to their peers in the traditional teaching group (traditional group: mean 37.6, SD 2.8; WeChat group: mean 32.0, SD 4.1; *P*=.006; n=54), they exhibited markedly superior levels of understanding (traditional group: mean 24.1, SD 1.8; WeChat group: mean 28.0, SD 1.3; *P*=.007; n=54) and knowledge application (traditional group: mean 24.3, SD 1.9; WeChat group: mean 27.6, SD 1.3; *P*=.008; n=54). This suggests that the WeChat-based PBL method promotes deeper engagement, enabling students to better comprehend essential concepts, even with a diminished emphasis on rote learning. Additionally, students in the WeChat group reported increased collaboration (traditional group: mean 3.8889, SD 0.8393; WeChat group: mean 1.7222, SD 0.5961; *P*<.001); motivation (traditional group: mean 3.5471, SD 0.7915; WeChat group: mean 1.8333, SD 0.5746; *P*=.004); knowledge acquisition (traditional group: mean 3.6667, SD 0.7770; WeChat group: mean 1.8704, SD 0.7017; *P*<.001); self-learning ability (traditional group: mean 3.5741, SD 0.7673; WeChat group: mean 1.8519, SD 0.4917; *P*<.001); clinical reasoning (traditional group: mean 2.9444, SD 0.8777; WeChat group: mean 1.9630, SD 0.6132; *P*=.01); and problem-solving skills (traditional group: mean 3.2037, SD 0.6553; WeChat group: mean 1.8519, SD 0.5287; *P*=.001) than the students in the traditional group.

Conclusions: Integrating PBL methods with WeChat has been shown to improve ophthalmic education outcomes compared to traditional teaching, suggesting that this method may offer a superior alternative to conventional teaching.

(JMIR Med Educ 2026;12:e65279) doi:10.2196/65279

KEYWORDS

problem-based learning; WeChat; China; ophthalmology; undergraduate; medical students

Introduction

Medical education increasingly incorporates innovative strategies in undergraduate training to better prepare students for the challenges of clinical practice. Key strategies such as problem-based learning (PBL) [1,2], case-based learning [3], team-based learning [4], flipped classroom–based learning [5,6], and competency-based medical education [7] have been emphasized to bridge the gap between theoretical knowledge and clinical skills. These strategies are particularly relevant in fields such as ophthalmology, which represents a distinct and highly specialized discipline within clinical medicine [8]. The intricate anatomy of the eye and the nonintuitive nature of its structure necessitate the use of specialized equipment and tests for accurate diagnosis and management [9]. These factors contribute to the considerable hurdles medical students

XSL•FO
RenderX

encounter while attempting to master this subject. Therefore, effective teaching methodologies are vital for adequately preparing students for clinical practice and developing their competencies as future clinicians [10].

The conventional teacher-centered educational model often relies on the passive transmission of knowledge from instructors to students. While efficient for delivering core knowledge, especially in large classes common in China, this method can limit students' engagement and active learning [11]. As a result, students may struggle to operate ophthalmic instruments or perform fundus examinations, thereby diminishing their confidence in applying ophthalmology skills. Moreover, although textbooks provide detailed descriptions of the clinical manifestations and signs of ophthalmic diseases, students frequently encounter challenges when attempting to apply this theoretical knowledge to real-world clinical scenarios [12].

PBL, which inherently starts with a clinically based problem as its name, is an active learning strategy that fosters student-centered engagement [13]. Evidence shows that PBL enhances learning motivation, promotes self-guided learning, and improves the integration of theoretical knowledge with clinical practice [14,15]. Students exposed to PBL report greater satisfaction and participation compared with those in traditional learning settings [16,17]. In today's digital era, college students are adept at using the internet and mobile networks to access information, helping to overcome the temporal and spatial limitations of traditional PBL methods. With the advancement of mobile networks, WeChat has become a popular social networking platform in China [18]. It supports multiple modes of interaction—including texting, voice messages, images, and multimedia—similar to WhatsApp in Western countries [19,20]. Moreover, WeChat is widely used for disseminating knowledge to medical students due to its convenience and the extensive availability of information [21,22].

This study addresses the challenges faced in conventional ophthalmology education, where passive learning hinders the development of practical skills and confidence among undergraduate medical students. It aims to evaluate the effectiveness and acceptability of a WeChat-based PBL approach in enhancing students' self-perceptions of learning, initiative, organization, clinical skills, and knowledge mastery. The central research question guiding this inquiry is: How does the integration of WeChat in PBL influence the learning outcomes of medical students in ophthalmology? This randomized controlled trial will involve the comparison of student outcomes between those engaged in traditional teaching methods and those using the WeChat platform for online PBL, allowing for a robust assessment of the intervention's effectiveness. We hypothesize that the use of WeChat for online PBL will significantly improve student engagement and competency in ophthalmology. The findings of this study will provide valuable insights for medical educators, curriculum developers, and policymakers seeking to improve medical education methodologies in China and beyond.

Methods

Ethical Considerations

This study was conducted in accordance with the Declaration of Helsinki and received approval from the institutional review board and ethics committee of the Second Xiangya Hospital (Z0005-01). In accordance with institutional guidelines for educational research, the intervention was classified as minimal risk and considered part of the routine evaluation of teaching quality. All participants were informed of the study's purpose, procedures, and the planned use of the collected data, and all provided written informed consent before participation. Participant anonymity was strictly protected, with all data deidentified and analyzed only in aggregated form. No study procedures posed risks beyond those ordinarily encountered in standard educational activities.

Study Participants

A total of 108 fifth-year undergraduate students from Xiangya Medical College of Central South University, all of whom had passed the Chinese National Entrance Examination, participated in the study. The cohort consisted of 42 (38.9%) male and 66 (61.1%) female participants, aged 20 to 22 years. Participants were randomly assigned into 6 groups of 18 students each using Microsoft Excel to generate random numbers to ensure unbiased distribution. In this study, students were informed of their group assignments due to the nature of the teaching intervention, but evaluators responsible for assessing the examination results were blinded to group allocation to minimize assessment bias. The new teaching methods were implemented outside their regular class schedules, primarily during evenings and weekends. This approach presented unique challenges, including an increased workload for both students and instructors, as it required students to allocate additional time beyond their standard academic hours to participate in the study. Additionally, coordinating schedules among all faculty members and students involved posed challenges, which required careful planning and communication to ensure participation without conflicting with existing institutional educational policies. Both intervention and control groups received instruction in dedicated classrooms at the hospital, which facilitated a conducive learning environment (Table 1).

XSL•FO
RenderX

Table . Timeline of the WeChat-based problem-based learning activities.

Week	Activity	Duration (h)	Learning objectives
1	The instructor posts guiding questions and students form study groups	2	Students are randomly assigned to study groups of 10 and establish WeChat groups, through which group members communicate about the study content
2	Explore contributing factors	4	Identify and explain major factors and search for medical literature and books on your own to find the answers to the guiding questions provided by the instructor
3	Apply concepts to scenarios	4	Apply learned concepts to hypothetical scenarios and communicate with group members and the instructor through the WeChat group
4	Solve the guiding problems and present the key points	5	Students create relevant course Microsoft PowerPoint presentations according to the assigned groups, which cover all the answers to the instructor's guiding questions and the knowledge points obtained from self-study of the teaching content of this course
5	Evaluate proposed solutions	4	Evaluate the effectiveness of various solutions by presenting them to each other during the formal class

Research Methods

We selected primary angle-closure glaucoma (PACG) as the topic for applying the WeChat-based PBL approach in this study, as the diagnosis and treatment of PACG represent crucial skills that students must acquire. The textbook used for this study was *Ophthalmology*, 9th edition, published by the People's Medical Publishing House [23].

WeChat-Based PBL

Establishment of Learning Groups

Three groups (18 students for each group) were randomly selected to take part in the WeChat-based PBL program. In each group, 1 volunteer self-selected by students was responsible for recording each student's speech and the outcomes of the panel discussions.

Problem Introduction Phase

In alignment with the national syllabus and the ophthalmology textbook, 3 representative clinical cases were selected from the 2022-2023 glaucoma service database of the Second Xiangya Hospital. Each vignette depicted a distinct stage or mechanism of PACG (including asymptomatic primary angle closure, acute angle closure attack, and plateau-iris configuration) covering the essential diagnostic and therapeutic spectrum required for undergraduate training. To guide case selection and design, chapter-level learning objectives were first extracted, encompassing (1) anterior-chamber anatomy and angle structures, (2) the pathophysiology of pupillary-block versus nonpupillary-block mechanisms, (3) biometric risk factors, (4) gonioscopic and imaging diagnosis, (5) laser and surgical indications, and (6) emergency management of acute attacks. Each objective was then translated into a practical key decision point (eg, interpreting a gonioscopy record, determining when laser peripheral iridotomy is indicated, or choosing between phacoemulsification and trabeculectomy). All selected cases, provided by the instructors, were distributed to students 1 week prior to the lecture to facilitate preparatory discussion and independent learning.

Preclass Guidance Phase

Instructors provided curated readings including key sections drawn from the textbook, supplemented with selected review articles, clinical guidelines (eg, *Asia-Pacific Glaucoma Guidelines, 3rd Edition*), and representative case reports from the hospital's glaucoma database. Additionally, each vignette was accompanied by inquiry-based questions that directed students to identify risk factors, compare pathogenic mechanisms, and determine appropriate management strategies, encouraging active and evidence-based learning. These questions could direct students' literature searches toward key concepts and clinical decision-making points related to PACG. Each group participating in the WeChat-based PBL established its own discussion team on the WeChat platform, fostering an environment of active participation. To encourage full student engagement, instructors emphasized the importance of collaboration and implemented a system that recognized contributions, motivating students to prepare thoroughly before classes. Through this platform, instructors and students engaged in interactive discussions on the provided cases, sharing their diagnostic reasoning, proposing treatment plans, and exploring the latest global developments in disease management. For example, instructors could prompt a discussion by asking specific diagnostic questions, thereby guiding students to reflect

XSL•FO
RenderX

critically. During class sessions, instructors monitored and reviewed the discussions and the online outputs, often projecting key discussion points or summaries from WeChat onto a shared screen for collective evaluation and real-time feedback.

Discussion and Summarization Phase

Each WeChat-based PBL session lasted 2 hours and was supported by the platform's features for sharing text, images, voice recordings, videos, and documents, which enabled immediate and efficient communication. All instructors followed a structured facilitator's guide, identical across groups, and distributed 1 week before the session. The guide provided a minute-by-minute timetable covering the entire 2-hour process (10 min recap, 45 min group presentation, 30 min cross-groups challenge, 25 min instructor synthesis, and 10 min reflection), ensuring balanced pacing and consistent coverage across groups.

At the beginning of each session, a representative student from each group presented their understanding of the assigned disease and summarized unresolved problems identified during preclass discussions. This was followed by an open, interactive discussion in which students from other groups exchanged perspectives and clinical reasoning. To maintain alignment with curricular standards, the instructor's guide specified 5 "must-cover" learning objectives mapped to the national ophthalmology syllabus (eg, explain the mechanism of pupillary block, justify the timing of laser iridotomy, and interpret ultrasound biomicroscopy images). It also included prompt questions addressing common misconceptions identified during the pilot phase, and a 1-slide "take-home" summary that instructors were required to display at the end of the session.

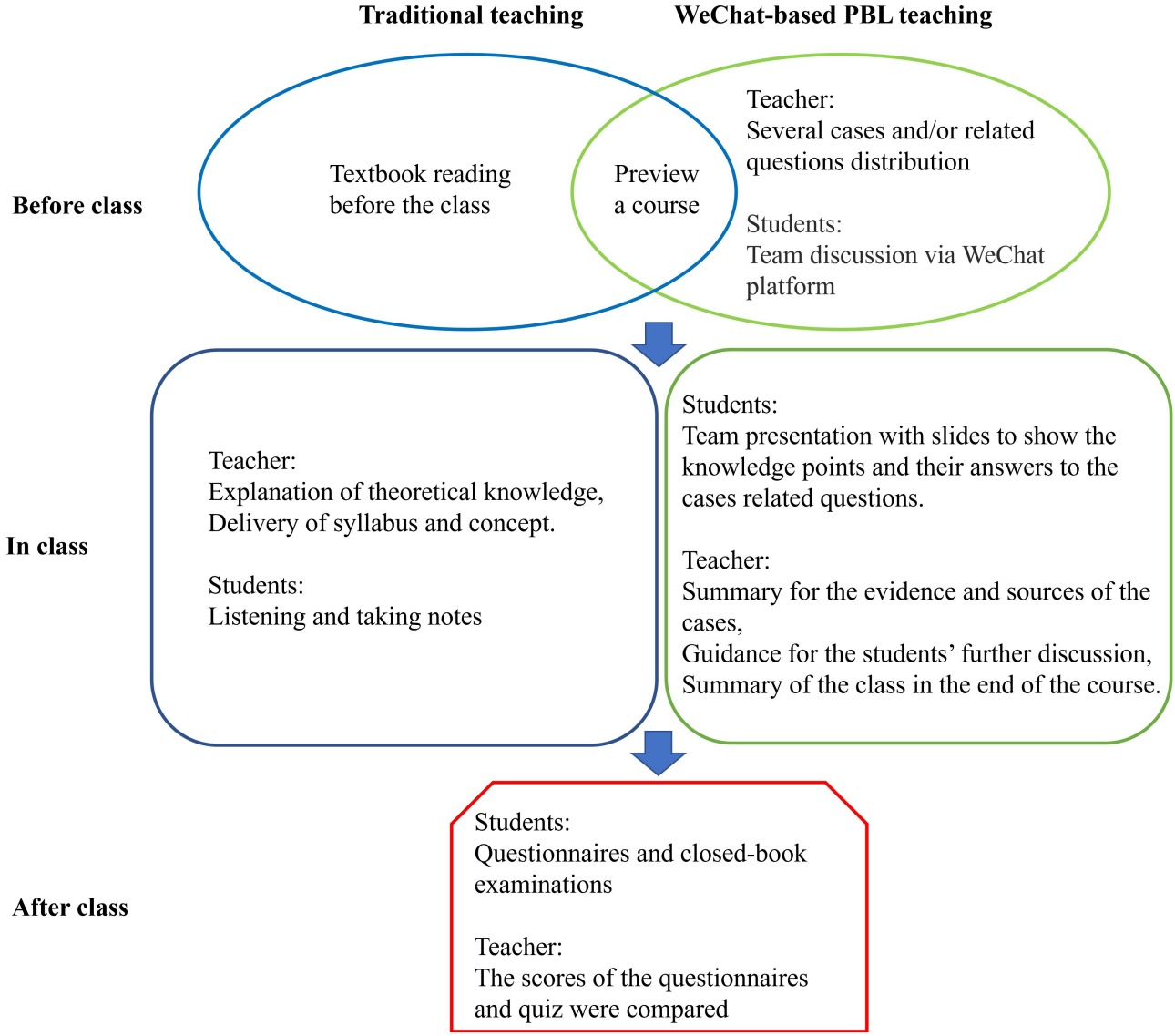
After the session, instructors completed a structured checklist to document which objectives had been achieved. They were instructed to intervene only when discussion drifted more than 2 minutes off task or when critical misconceptions arose; otherwise, they guided students by inquiry-based questioning to ensure that all checklist objectives were discussed. The structured facilitation process, combined with WeChat's interactive features, fostered active participation, peer collaboration, and efficient knowledge integration, greatly enhanced the PBL learning experience.

Traditional Teaching Group

The traditional teaching model involved the instructor predominantly delivering comprehensive explanations of foundational knowledge, adhering to a prescribed multimedia teaching outline. In this model, students were not divided into smaller groups and were not required to give formal presentations during the course. Both groups completed the same quizzes and questionnaires, and the results did not affect course grades or performance. The study design is summarized in Figure 1.

XSL•FO
RenderX

Figure 1. The flowchart shows the 2 different models of teaching and learning in ophthalmology for Chinese medical students. PBL: problem-based learning.



Teaching Efficacy Evaluation and Statistical Analysis

The efficacy and satisfaction of both teaching methods were evaluated through a combination of questionnaire surveys for subjective assessment and examinations for objective evaluation.

Subjective Evaluation

The questionnaire (Table 2) was designed and modified based on the Course Experience Questionnaire by Ramsden [24] and the Study Process Questionnaire by Biggs [25]. All the students completed questionnaires to assess their perceptions and self-evaluated competence at the end of the teaching course.

XSL•FO
RenderX

Table . Specific questions used in the questionnaire

Topic area	Questions
The course improves my learning motivation	• "Can you analyze the impact of the study findings on the field?"
The course helps me to acquire knowledge	• "Can you define the main terms used in the study?"
The course is helpful for passing the final examination	• "What are your thoughts on the effectiveness of the methods used in the study?"
The course encourages me to express my opinions	• "In this course, do you feel that your opinions are respected?" • "In this course, do you often take the initiative to express your opinions?"
This course occupies too much of my spare time	• "How much time do you spend in WeChat-based problem-based learning by using this method to communicate with teachers and group members?"
The course improves my communication skills	• "During discussions, does the instructor give you sufficient time to express your opinions?"
The course improves my self-learning skills	• "What are the key concepts discussed in the study material?"
The course improves my clinical thinking ability	• "How would you apply the concepts learned to a real-world scenario?"
The course improves my ability to analyze and solve problems	• "How much do you agree that this course has provided you with the tools and techniques to effectively solve problems?" • "How often have you used the analytical skills you learned in the course to solve real-world problems outside the classroom?

Objective Evaluation

Both groups then undertook an identical closed-book examination comprising essay questions, true or false questions, judgment questions, and case analyses to assess students' memory, comprehension, and clinical application abilities. The examination questions were formulated based on the 9th edition of *Ophthalmology*, an ophthalmology textbook published by the People's Medical Publishing House [23]. Notably, both the examination format and evaluation criteria were standardized across both groups. To maintain objectivity in scoring the theoretical examination, 3 instructors performed a blinded evaluation of the students' responses.

Statistical analyses were performed using SPSS (version 11.0; IBM Corp) software. Measurement data were expressed as means (SDs). Questionnaire data were analyzed using the Mann-Whitney *U* test, and examination scores across the two groups were compared using the independent samples 2-tailed *t* test. A *P* value <.05 was considered statistically significant.

Results

Participants' Demographic Data

Figure 2 shows a flowchart of participation for the 108 medical students in this study. Among them, 54 (50%) students were randomly selected to participate in the WeChat-based PBL, while the other 54 (50%) received traditional teaching. Among the participants, there were 42 (38.9%) male and 66 (61.1%) female individuals who were aged 20 to 22 years. The mean age of the WeChat-based group was 20.426 (SD 0.602) years, and the mean age of the traditional teaching group was 20.296 (SD 0.571) years. There was no significant difference in gender and average age between the two groups (Table 3).

XSL•FO
RenderX

Figure 2. Consolidated Standards of Reporting Trials (CONSORT) flowchart.

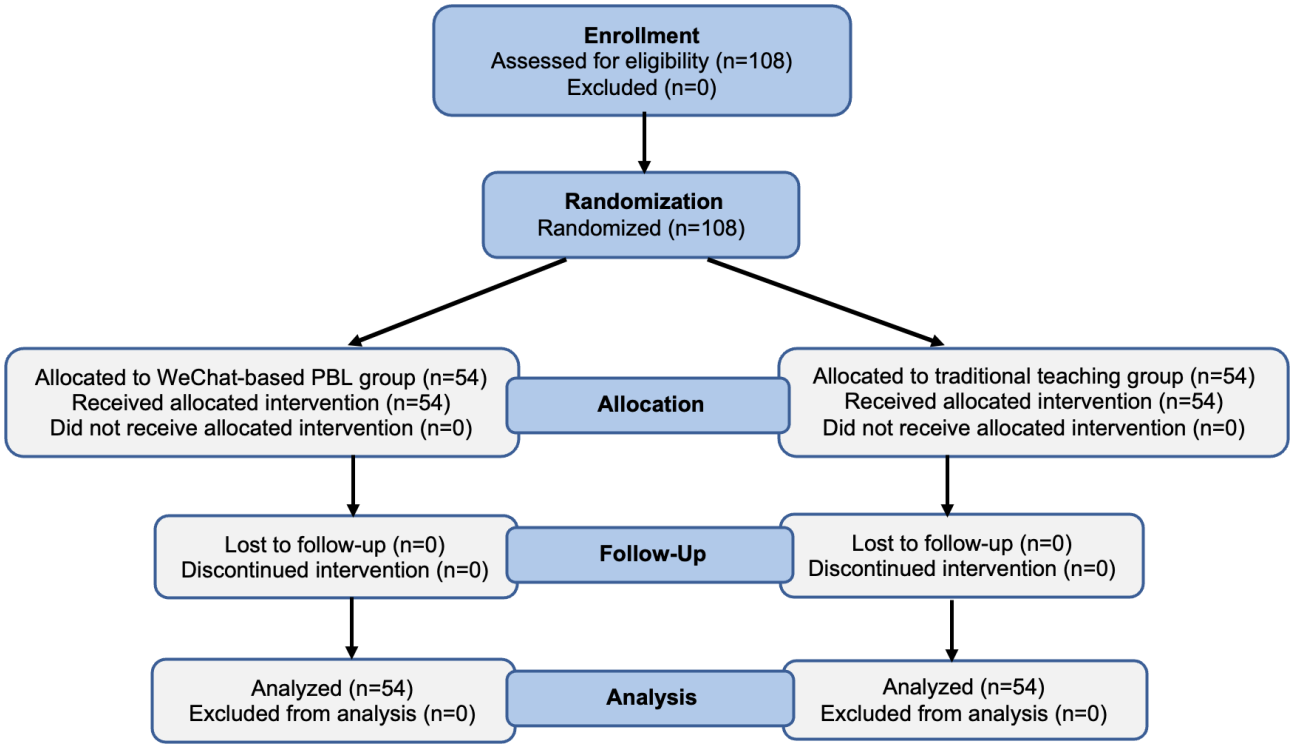


Table . Participants' demographic data of the Second Xiangya Medical College of WeChat-based problem-based learning (PBL) and traditional teaching groups.

Characteristics	WeChat-based PBL (n=54)	Traditional teaching (n=54)	*P* value
Age (y), mean (SD)	20.426 (0.602)	20.296 (0.571)	.12[a]
Sex, n (%)			.48[b]
Male	23 (42.6)	22 (40.7)	
Female	31 (57.4)	32 (59.3)	

[a]For age, independent sample t_{106}=1.188.

[b]For sex, χ^2_1=1.0.

Comparison of the WeChat-Based PBL Method and Traditional Teaching Method

The closed-book examination was administered 1 week after course completion to allow students sufficient time for review while minimizing long-term memory effects. This clarification has been included in the revised manuscript. Examination questions were designed by 3 instructors and included 6 types of questions: single-choice questions, multiple-choice questions, noun definitions, short-answer questions, essay questions, and comprehensive questions. The exam assessed 3 cognitive levels—memorization, understanding, and application, which accounted for 50%, 20%, and 30% of the total score, respectively. The memorization section evaluated students' ability to recall fundamental facts, definitions, and terminology. The understanding section assessed their capacity to interpret information, explain concepts in their own words, and demonstrate deeper comprehension beyond simple recall. The application section examined students' ability to apply their knowledge and understanding to new or practical situations, such as problem-solving, formula application, and case analysis. Notably, the sections on understanding and application emphasized students' flexible mastery of knowledge and its practical implementation. The goal was to enable students to apply theoretical knowledge in clinical contexts, thereby improving their diagnostic and treatment skills as well as their ability to analyze and solve clinical problems.

Our analysis indicated that although students in the WeChat-based PBL group scored marginally lower in memorization compared to their peers in the traditional teaching group (traditional vs WeChat groups: 37.6, SD 2.8 vs 32.0, SD 4.1; t_{106}=-8.426; *P*=.006; n=54), they exhibited markedly superior levels of understanding (traditional vs WeChat groups: 24.1, SD 1.8 vs 28.0, SD 1.3; t_{106}=12.753; *P*=.007; n=54) and knowledge application (traditional vs WeChat groups: 24.3, SD 1.9 vs 27.6, SD 1.3; t_{106}=10.899; *P*=.008; n=54). However, when considering the total score, there was no significant difference between the two groups (traditional vs WeChat: 85.8, SD 4.9 vs 87.4, SD 5.1; t_{106}=1.694; *P*=.60; n=54).

All 108 (100%) students responded to the questionnaire for the comparison of students' perceptions between the WeChat-based

XSL•FO
RenderX

PBL and lecture-based class (Table 4). More students believed (rather than did not believe) that the WeChat-based PBL class enhanced their learning motivation (traditional group: mean 3.5471, SD 0.7915; WeChat group: mean 1.8333, SD 0.5746; *P*=.004), helped them acquire knowledge (traditional group: mean 3.6667, SD 0.7770; WeChat group: mean 1.8704, SD 0.7017; *P*<.001), and improved their performance on the final examination (*P*=.001). They also believed that it encouraged them to express opinions (traditional group: mean 3.5556, SD 0.7181; WeChat group: mean 1.9630, SD 0.5482; *P*=.001), enhanced their communication skills (traditional group: mean 3.8889, SD 0.8393; WeChat group: mean 1.7222, SD 0.5961; *P*<.001), improved their self-learning ability (traditional group: mean 3.5741, SD 0.7673; WeChat group: mean 1.8519, SD 0.4917; *P*<.001), improved their clinical thinking (traditional group: mean 2.9444, SD 0.8777; WeChat group: mean 1.9630, SD 0.6132; *P*=.012), and enhanced their ability to analyze and resolve problems (traditional group: mean 3.2037, SD 0.6553; WeChat group: mean 1.8519, SD 0.5287; *P*=.001). However, a majority of students in the WeChat-based PBL class felt that the course occupied too much of their spare time (traditional group: mean 3.7963, SD 0.7618; WeChat group: mean 1.9630, SD 0.5482; *P*<.001).

Table . Comparison of students' perceptions and self-evaluation between WeChat-based problem-based learning (PBL) method and traditional teaching method. The results were ranked as a Likert scale from 1 to 5 (1=strongly agree, 2=agree, 3=neutral, 4=disagree, and 5=strongly disagree).

Questions	WeChat-based PBL, mean (SD)	Control (traditional teaching), mean (SD)	Mann-Whitney *U* test	*P* value	Effect size
The course improves my learning motivation	1.8333 (0.5746)	3.5741 (0.7915)	1012	.004	2.5169
The course helps me to acquire the knowledge	1.8704 (0.7017)	3.6667 (0.7770)	914	<.001	2.4264
The course is helpful for passing the final examination	2.7593 (0.7507)	3.2593 (0.8284)	975.5	<.001	0.6325
The course encourages me to express my opinions	1.9630 (0.5482)	3.5556 (0.7181)	963	.001	2.4930
This course occupies too much of my spare time	3.3148 (0.8201)	3.7963 (0.7618)	714	<.001	0.6083
The course improves my communication skills	1.7222 (0.5961)	3.8889 (0.8393)	725	<.001	2.9765
The course improves my self-learning skills	1.8519 (0.4917)	3.5741 (0.7673)	930.5	<.001	2.6725
The course improves my clinical thinking ability	1.9630 (0.6132)	2.9444 (0.8777)	1072	.012	1.2964
The course improves my ability to analyze and solve problems	1.8519 (0.5287)	3.2037 (0.6553)	932	.001	2.2706

Discussion

Principal Findings

This study aimed to assess the effectiveness of WeChat-based PBL in ophthalmic education compared with traditional teaching methods. Our analysis revealed that while students in the WeChat-based PBL group exhibited slightly lower scores in memorization than those in the traditional teaching group, they demonstrated significantly higher levels of understanding and application of knowledge. This suggests that the WeChat-based PBL approach fosters deeper engagement, allowing students to grasp critical concepts more effectively, despite having less focus on rote memorization. Moreover, students in the WeChat group reported enhanced collaboration, motivation, and willingness to participate actively in their learning process.

WeChat has been used as a platform for PBL because of its time efficiency and convenience [26-28]. Students were able to present questions and engage in discussions with instructors at any time and from any location via WeChat. Moreover, it allows the sharing of images, videos, links, and other related resources. This capability greatly aids discussion and ensures that all members can access case-related or problem-related resources promptly. Everyone can share their opinions and suggestions, receiving feedback from others anytime and anywhere. This timely and efficient communication via WeChat offers a better understanding among team members [26,27,29]. Successful collaboration guarantees the productivity of the team, and instructors can actively participate in students' discussions and debates, responding to questions and offering advice to enhance the efficiency of the learning process [26,27].

XSL•FO
RenderX

Although the primary benefit of WeChat-based PBL is time-saving and convenience, as traditional PBL can be a time-intensive approach [30], the perception that WeChat-based PBL occupied too much spare time underscores the need for careful consideration of time management and workload in educational program design. One of the reasons could be the additional time dedicated to self-learning [31] and discussion among the team members increased the pressure and burden on the students. Increased preparation time can lead to deeper understanding and better retention of the material as students engage with content more thoroughly before class discussions. This proactive approach enhances participation during PBL sessions and ultimately improves learning outcomes. To optimize learning efficiency while preventing overload, several refinements are warranted. Specifically, preclass materials should be streamlined to emphasize key concepts and reduce redundancy, and brief in-class reviews can help reinforce complex topics, minimizing postclass study demands. Periodic guided learning sessions can also be scheduled to help balance students' independent study and supervised learning time. These refinements aim to alleviate cognitive load while maintaining the effectiveness of the teaching model. Moreover, it is essential to refine the new teaching approach to effectively integrate the advantages of the PBL model with the interactive capabilities of the WeChat platform. Furthermore, instructors should also receive appropriate training to embrace this innovative teaching approach [32]. Dolmans and Wolfhagen [33] have highlighted that a tutor's actions significantly influence both the productivity and effectiveness of a PBL group's efforts. To enhance the effectiveness of tutors in PBL tutorials, ongoing training that encourages reflective practice on their development as educators is recommended [34,35]. These efforts can effectively strike a balance between optimal learning outcomes and manageable student workload.

At the same time, the increase in students' out-of-class learning hours also implied additional workload for instructors. Although instructor workload in asynchronous chat facilitation was not quantitatively measured, informal feedback indicated a moderate increase. Instructors reported spending an additional 15 to 20 minutes per day monitoring discussions and providing timely feedback. Despite this, they generally considered this extra time acceptable and valuable, as it enabled closer observation of students' reasoning and early identification of misconceptions. The main challenges involved balancing prompt responses with other duties and managing message overload during peak activity. Nonetheless, most instructors agreed that the improved engagement and learning outcomes justified the additional effort. A rotating instructor schedule for online facilitation and setting clearer expectations for response timelines may help balance instructional workload while maintaining effective student-teacher interaction.

Our statistical analysis revealed that the WeChat-based PBL group exhibited slightly lower scores in memorization than the traditional teaching group. However, their levels of understanding and application of knowledge were higher than the traditional teaching group. This discrepancy may be attributed to the engagement of students in the WeChat-based PBL group with a broader array of relevant information, extending beyond textbook content. Consequently, while the WeChat-based PBL group demonstrated a better performance in understanding and application of knowledge, their memorization of textbook-specific information was slightly weaker. Nevertheless, the WeChat-based PBL group showed advantages in mastering key knowledge points and achieving higher total scores.

However, several factors must be considered while implementing WeChat-based PBL. For instance, it is advisable to keep the number of students in each team small, as smaller groups facilitate easier administration and instruction. Teachers should contact students at least 1 week in advance to provide questions or case reports, ensuring both teachers and students are well prepared to achieve optimal educational outcomes. By adapting to real-time situations, teachers can enhance students' interest in learning, invigorate the classroom dynamics, and correct students' misconceptions. Teachers should act as a resource and facilitator, responding promptly to students' inquiries throughout their research. Overall, the WeChat-based PBL model marks a significant departure from the traditional didactic "cramming education" approach, fostering a more engaging and active learning environment. Therefore, WeChat-based PBL should be increasingly incorporated into future ophthalmology education to enable continuous identification and refinement of relevant issues [36].

Limitations

There are several limitations that should be acknowledged. First, while students covered the same core content on PACG, differences in teaching methodologies could have influenced memorization outcomes and overall comprehension. This research was performed targeting a single textbook chapter on PACG. Therefore, the conclusion requires further verification in other subject fields. Second, questionnaire results indicated that students in the WeChat-based PBL group spent more extracurricular time preparing course materials. Therefore, understanding their perceptions and attitudes toward this additional workload is particularly important. However, as the students involved in this study have already graduated, it was not possible to collect further feedback from them. This limitation prevents a more comprehensive understanding of their long-term perceptions of the learning burden and its potential impact on sustained engagement and clinical competency development. Future research will include longitudinal follow-up surveys or interviews to further evaluate students' attitudes, learning experiences, and workload perceptions. Third, the necessity for student participation in additional discussions and presentations may have affected their independent study time and overall academic performance. Fourth, the lack of blinding may have introduced biases, as awareness of the teaching method could have influenced both students' and instructors' expectations and engagement levels. Fifth, the generalizability of our findings is also limited by the single-institution context and the sample size of 108 fifth-year undergraduate medical students, which may not reflect broader educational settings or diverse student populations. Finally, in our study, we did not compare the effects of traditional PBL and WeChat-based PBL methods. Further research should compare the functions between these two approaches in future

XSL•FO
RenderX

studies. Moreover, future research should explore the long-term impacts of WeChat-based PBL on learning outcomes across different medical disciplines and institutions. Further empirical testing is necessary to assess the effectiveness of this approach in various contexts and to identify best practices for implementation. Optimizing digital communication tools such as WeChat holds great promise for enhancing educational experience and fostering collaborative learning in medical education.

Conclusions

In summary, compared to traditional teaching, integrating PBL methods combined with WeChat as a communication platform improved the effectiveness of ophthalmic education. Although students invested more time in preclass, they did not perceive it as a waste of their spare time. Instead, they recognized that the WeChat-based PBL mode facilitated more effective communication among students and between students and teachers. It enhanced their motivation to learn; promoted knowledge acquisition; encouraged expression of opinions; and improved self-learning abilities, clinical reasoning, and problem-solving. This innovative teaching mode may represent a superior alternative to conventional teaching methods. However, further exploration to optimize the students' spare time for effective online communication is warranted. This new method merits continued refinement and evaluation based on the findings of this study.

Funding

This work was supported by the 2024 Education and Teaching Reform Research Project of Central South University (grant 2024jy141), the Changsha Natural Science Foundation (grant kq2403090), the Hunan Province Health Commission (grant 202307027224), and the Natural Science Foundation of Hunan (grants 2023JJ10087 and 2025JJ50538).

Data Availability

The datasets used or analyzed during this study are available from the corresponding author on reasonable request.

Authors' Contributions

WJ and FF designed the experiments and drafted the manuscript. FF and BB collected the data and performed the statistical analysis. WJ and BB reviewed the manuscript. All authors read and approved of the final manuscript.

Conflicts of Interest

None declared.

References

1. Jin J, Bridges SM. Educational technologies in problem-based learning in health sciences education: a systematic review. J Med Internet Res 2014 Dec 10;16(12):e251. [doi: 10.2196/jmir.3240] [Medline: 25498126]
2. Wyer PC. Evidence-based medicine and problem based learning a critical re-evaluation. Adv Health Sci Educ Theory Pract 2019 Dec;24(5):865-878. [doi: 10.1007/s10459-019-09921-5] [Medline: 31617018]
3. Thistlethwaite JE, Davies D, Ekeocha S, et al. The effectiveness of case-based learning in health professional education. A BEME systematic review: BEME Guide No. 23. Med Teach 2012;34(6):e421-e444. [doi: 10.3109/0142159X.2012.680939] [Medline: 22578051]
4. Burgess A, Ayton T, Mellis C. Implementation of team-based learning in year 1 of a PBL based medical program: a pilot study. BMC Med Educ 2016 Feb 4;16:49. [doi: 10.1186/s12909-016-0550-3] [Medline: 26846425]
5. Han E, Klein KC. Pre-class learning methods for flipped classrooms. Am J Pharm Educ 2019 Feb;83(1):6922. [doi: 10.5688/ajpe6922] [Medline: 30894772]
6. Park KH, Park KH, Chae SJ. Experiences of medical teachers in flipped learning for medical students: a phenomenological study. Korean J Med Educ 2018 Jun;30(2):91-100. [doi: 10.3946/kjme.2018.84] [Medline: 29860775]
7. Succar T, McCluskey P, Grigg J. Enhancing medical student education by implementing a competency-based ophthalmology curriculum. Asia Pac J Ophthalmol (Phila) 2017;6(1):59-63. [doi: 10.22608/APO.2016102] [Medline: 28161917]
8. Gostimir M, Sharma RA, Bhatti A. Status of Canadian undergraduate medical education in ophthalmology. Can J Ophthalmol 2018 Oct;53(5):474-479. [doi: 10.1016/j.jcjo.2017.11.015] [Medline: 30340714]
9. Wang A, Xiao R, Zhang C, et al. Effectiveness of a combined problem-based learning and flipped classroom teaching method in ophthalmic clinical skill training. BMC Med Educ 2022 Jun 23;22(1):487. [doi: 10.1186/s12909-022-03538-w] [Medline: 35733187]
10. Schmidt HG. Problem-based learning: does it prepare medical students to become better doctors? Med J Aust 1998 May 4;168(9):429-430. [doi: 10.5694/j.1326-5377.1998.tb139020.x] [Medline: 9612452]
11. Tu MG, Yu CH, Wu LT, Li TC, Kwan CY. Dental and medical students' perspectives on early exposure to PBL in Taiwan. J Dent Educ 2012 Jun;76(6):746-751. [Medline: 22659703]
12. Al Wadani F, Khan AR. Problem-based learning in ophthalmology: a brief review. Oman J Ophthalmol 2014 Jan;7(1):1-2. [doi: 10.4103/0974-620X.127908] [Medline: 24799792]

XSL•FO
RenderX

13. Rui Z, Rong-Zheng Y, Hong-Yu Q, Jing Z, Xue-Hong W, Chuan Z. Preliminary investigation into application of problem-based learning in the practical teaching of diagnostics. Adv Med Educ Pract 2015;6:223-229. [doi: 10.2147/AMEP.S78893] [Medline: 25848334]
14. Tudor Car L, Kyaw BM, Dunleavy G, et al. Digital problem-based learning in health professions: systematic review and meta-analysis by the digital health education collaboration. J Med Internet Res 2019 Feb 28;21(2):e12945. [doi: 10.2196/12945] [Medline: 30816846]
15. Khaki AA, Tubbs RS, Zarrintan S, et al. The first year medical students' perception of and satisfaction from problem-based learning compared to traditional teaching in gross anatomy: introducing problem-based anatomy into a traditional curriculum in Iran. Int J Health Sci (Qassim) 2007 Jan;1(1):113-118. [Medline: 21475460]
16. Albanese MA, Mitchell S. Problem-based learning: a review of literature on its outcomes and implementation issues. Acad Med 1993 Jan;68(1):52-81. [doi: 10.1097/00001888-199301000-00012] [Medline: 8447896]
17. Groves M, Régo P, O'Rourke P. Tutoring in problem-based learning medical curricula: the influence of tutor background and style on effectiveness. BMC Med Educ 2005 Jun 7;5(1):20. [doi: 10.1186/1472-6920-5-20] [Medline: 15938758]
18. Sun ZJ, Zhu L, Liang M, Xu T, Lang JH. The usability of a WeChat-based electronic questionnaire for collecting participant-reported data in female pelvic floor disorders: a comparison with the traditional paper-administered format. Menopause 2016 Aug;23(8):856-862. [doi: 10.1097/GME.0000000000000690] [Medline: 27326820]
19. Hussain I, Khan RA, Mahboob U, Latif MZ, Waheed K. Comparison of effectiveness of WhatsApp and Facebook as learning tool for the students of a medical college. J Pak Med Assoc 2020 Oct;70(10):1705-1708. [doi: 10.5455/JPMA.35071] [Medline: 33159737]
20. Oyewole BK, Animasahun VJ, Chapman HJ. A survey on the effectiveness of WhatsApp for teaching doctors preparing for a licensing exam. PLoS One 2020;15(4):e0231148. [doi: 10.1371/journal.pone.0231148] [Medline: 32240259]
21. Zeng F, Deng G, Wang Z, Liu L. WeChat: a new clinical teaching tool for problem-based learning. Int J Med Educ 2016 Apr 25;7:119-121. [doi: 10.5116/ijme.5708.e5c4] [Medline: 27111920]
22. Wang J, Gao F, Li J, et al. The usability of WeChat as a mobile and interactive medium in student-centered medical teaching. Biochem Mol Biol Educ 2017 Sep;45(5):421-425. [doi: 10.1002/bmb.21065] [Medline: 28581156]
23. Peizeng Y, Xianqun F. Ophthalmology, 9th edition: People's Medical Publishing House; 2018.
24. Ramsden P. A performance indicator of teaching quality in higher education: the course experience questionnaire. High Educ Stud 1991 Jan;16(2):129-150. [doi: 10.1080/0307507911233138294]
25. Biggs JB. Study process questionnaire manual: student approaches to learning and studying. : Australian Council for Educational Research Ltd; 1987 URL: https://eric.ed.gov/?id=ED308200 [accessed 2025-10-05]
26. Zhang W, Li ZR, Li Z. WeChat as a platform for problem-based learning in a dental practical clerkship: feasibility study. J Med Internet Res 2019 Mar 19;21(3):e12127. [doi: 10.2196/12127] [Medline: 30888328]
27. Luo P, Pang W, Wang Y, et al. WeChat as a platform for problem-based learning among hematological postgraduates: feasibility and acceptability study. J Med Internet Res 2021 May 25;23(5):e16463. [doi: 10.2196/16463] [Medline: 34032573]
28. Tu S, Yan X, Jie K, Ying M, Huang C. WeChat: an applicable and flexible social app software for mobile teaching. Biochem Mol Biol Educ 2018 Sep;46(5):555-560. [doi: 10.1002/bmb.21170] [Medline: 30369037]
29. Zhou Y, Zhang D, Guan X, Pan Q, Deng S, Yu M. Application of WeChat-based flipped classroom on root canal filling teaching in a preclinical endodontic course. BMC Med Educ 2022 Mar 2;22(1):138. [doi: 10.1186/s12909-022-03189-x] [Medline: 35236332]
30. Seneviratne RD, Samarasekera DD, Karunathilake IM, Ponnamperuma GG. Students' perception of problem-based learning in the medical curriculum of the Faculty of Medicine, University of Colombo. Ann Acad Med Singap 2001 Jul;30(4):379-381. [Medline: 11503544]
31. Gillette C, Rudolph M, Kimble C, Rockich-Winston N, Smith L, Broedel-Zaugg K. A meta-analysis of outcomes comparing flipped classroom and lecture. Am J Pharm Educ 2018 Jun;82(5):6898. [doi: 10.5688/ajpe6898] [Medline: 30013248]
32. Gómez-Carrasco CJ, Monteagudo-Fernández J, Moreno-Vera JR, Sainz-Gómez M. Evaluation of a gamification and flipped-classroom program used in teacher training: perception of learning and outcome. PLoS ONE 2020;15(7):e0236083. [doi: 10.1371/journal.pone.0236083] [Medline: 32673373]
33. Dolmans DH, Wolfhagen IH. Complex interactions between tutor performance, tutorial group productivity and the effectiveness of PBL units as perceived by students. Adv Health Sci Educ Theory Pract 2005 Aug;10(3):253-261. [doi: 10.1007/s10459-005-0665-5] [Medline: 16193404]
34. Hendry GD. Problem-based learning tutors' conceptions of their development as tutors. Med Teach 2009 Feb;31(2):145-150. [doi: 10.1080/01421590802146026] [Medline: 19330673]
35. Bollela VR, Gabarra MHC, da Costa C, Lima RCP. Students and tutors' social representations of assessment in problem-based learning tutorials supporting change. BMC Med Educ 2009 Jun 7;9:30. [doi: 10.1186/1472-6920-9-30] [Medline: 19500408]
36. Hartling L, Spooner C, Tjosvold L, Oswald A. Problem-based learning in pre-clinical medical education: 22 years of outcome research. Med Teach 2010 Jan;32(1):28-35. [doi: 10.3109/01421590903200789] [Medline: 20095771]

XSL•FO
RenderX

Abbreviations

PACG: primary angle closure glaucoma
PBL: problem-based learning

Edited by B Lesselroth; submitted 16.Aug.2024; peer-reviewed by E Li, K Zhu; revised version received 14.Nov.2025; accepted 25.Nov.2025; published 05.Jan.2026.

Please cite as:
Fang F, Bu B, Jiang W
Combining Problem-Based Learning Methods With the WeChat Platform in Teaching Ophthalmology: Randomized Controlled Trial
JMIR Med Educ 2026;12:e65279
URL: https://mededu.jmir.org/2026/1/e65279
doi:10.2196/65279

© Fang Fang, Bing Bu, Wenmin Jiang. Originally published in JMIR Medical Education (https://mededu.jmir.org), 5.Jan.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Education, is properly cited. The complete bibliographic information, a link to the original publication on https://mededu.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Ultrasound-Guided Regional Anesthesia in a Resource-Limited Hospital: Prospective Pilot Study of a Hybrid Training Program

Jakob E Gamboa[1], MD; Inge Tamm-Daniels[1], MD; Roland Flores[1], MD; Nancy G Sarat Diaz[2], MD; Mario A Villasenor[1], MD; Mitchell A Gist[1], MD; Aidan B Hoie[1], MD; Christopher Kurinec[3], MS; Colby G Simmons[1], DO, MBA

[1]Department of Anesthesiology, University of Colorado School of Medicine, Aurora, CO, United States

[2]Department of Anesthesiology, Hospital Nacional de Coatepeque, Coatepeque, Quetzaltenango, Guatemala

[3]Department of Biostatistics and Informatics, Colorado School of Public Health, Aurora, CO, United States

Corresponding Author:
Jakob E Gamboa, MD
Department of Anesthesiology
University of Colorado School of Medicine
Mail Stop B113
12401 E. 17th Avenue
Aurora, CO, 80045
United States
Phone: 1 3034826929
Email: jakob.gamboa@cuanschutz.edu

Abstract

Background: Ultrasound-guided regional anesthesia (UGRA) remains underused in low- and middle-income countries due to barriers to training and equipment. Recent advances in portable ultrasound devices and international partnerships have expanded access to UGRA, enhancing patient safety and quality of care.

Objective: This study describes the development and outcomes of a hybrid UGRA training program for anesthesiologists at the Hospital Nacional de Coatepeque (HNC) in Guatemala.

Methods: An educational pilot program for UGRA was developed based on local needs and feedback, comprising 4 weeks of online modules, an in-person educational conference, and 1 month of supervised clinical practice. Evaluation followed the Kirkpatrick framework using preprogram and postprogram surveys adapted from the Global Regional Anesthesia Curricular Engagement model. Outcomes included participants' satisfaction, change in knowledge and skill, and procedural performance. Knowledge and skill assessments were compared before and after the training, and clinical data were recorded for 10 months. Nonparametric tests were used to assess changes and associations with performance outcomes.

Results: All 7 anesthesiologists at HNC completed the training program. Knowledge test scores improved by a median percentage increase of 20.8% (IQR 13.5%-28.1%; r=0.899; P=.02), and procedural skill rating scores increased by a median percentage of 147.1% (IQR 96.9%-197.3%; r=0.904; P=.03) at 1 month and 131.4% (IQR 90.5%-172.3%; r=0.909; P=.04) at 4 months after the program. Participants self-reported high satisfaction and substantial clinical improvement and motivation. A total of 54 peripheral nerve blocks were performed under direct supervision in the first month, with 187 blocks recorded over 10 months. The supraclavicular brachial plexus block was the most frequently used (66/187, 35.3%) and replaced the standard general anesthetic for upper extremity surgery in 70 patients. The procedure success rate was 96.3% (180/187), and there were no observed patient complications.

Conclusions: This hybrid curriculum enabled the successful implementation of UGRA at a public hospital in Guatemala, safely expanding clinical capabilities and reducing reliance on general anesthesia for upper extremity surgery. This practical training model provides a framework for implementing UGRA in similar resource-limited hospitals.

(JMIR Med Educ 2026;12:e84181) doi:10.2196/84181

KEYWORDS

health education; regional anesthesia; resource-limited settings; capacity building; global health

XSL•FO
RenderX

Introduction

Ultrasound-guided regional anesthesia (UGRA) has become the standard of care for safe and effective perioperative pain management in high-income countries. It is frequently used as the primary anesthetic in patients undergoing extremity surgeries [1]. Beyond the operating room, regional anesthesia is increasingly performed by nonanesthesia clinicians, including emergency physicians, to provide timely pain relief in acute care settings [2,3]. Ultrasound-guided peripheral nerve blocks (PNBs) are now used in emergency departments for conditions such as hip and rib fractures and wound management, demonstrating the growing interdisciplinary role of UGRA worldwide [4-6]. However, limited availability of ultrasound equipment and the lack of formal training programs contribute to the underuse of PNBs in low- and middle-income countries (LMICs) [7-10]. The emergence of portable, reliable, and cost-effective ultrasound devices offers new opportunities to expand the use of UGRA in diverse and resource-limited settings [11]. Advancing global regional anesthesia capacity can improve patient care by providing an alternative to general anesthesia, reducing cost and resource consumption, and enhancing patient comfort and safety [1,12].

Addressing the need for training opportunities is critical to the adoption of regional anesthesia techniques in LMICs. International partnerships help improve global capabilities by promoting educational activities and access to supplies. Although several initiatives have attempted to implement UGRA in diverse settings, few have been formally evaluated, and published evidence on their impact in LMICs remains limited [12]. Brouillette et al [13] described the successful development of an in-person educational program called the Global Regional Anesthesia Curricular Engagement (GRACE). In this program, anesthesia providers in Ghana were trained in UGRA techniques, demonstrating positive outcomes and increased procedure volume. Other collaborations have explored fully remote training models delivered via online communication platforms [14,15]. However, to our knowledge, the use of a hybrid training curriculum for UGRA, combining online didactic instruction with in-person, hands-on training, has not been described. Furthermore, most existing reports focus on programs based in African countries at large teaching hospitals with preexisting ultrasound capabilities and delivered in English. There is a paucity of evidence describing educational initiatives developed specifically for medical professionals in Latin America.

Our program was delivered in Guatemala, a Central American country where approximately 90% of the population relies on the public health sector, and approximately half of the country's 17 million inhabitants live in rural or impoverished communities [16]. Geographic and financial barriers limit access to advanced medical technologies such as ultrasound, which are largely concentrated in metropolitan areas [17]. Additionally, most physician training opportunities are offered in large urban hospitals near the capital of Guatemala City [17]. As a result, smaller public hospitals across the country have limited capacity and access to UGRA. Despite these challenges, there is substantial interest and growing demand for ultrasound-guided procedures nationwide.

In the southwestern region of Guatemala, the Hospital Nacional de Coatepeque (HNC) serves as the regional public hospital, facing high surgical demands amid notable resource constraints. In 2023, a needs assessment was performed through a partnership between the department of anesthesiology at the University of Colorado (CU) and HNC and identified a lack of UGRA capabilities in the region. In response, the department of anesthesiology at the HNC requested a formal training program to build local expertise and improve the quality of perioperative care. Our objective was to develop a tailored hybrid training program to establish a sustainable UGRA service at HNC. This evaluation assessed the outcomes and impact of this pilot program. We hypothesized that the implementation of this hybrid training model would increase clinical knowledge and confidence, enhance skill acquisition, and enable the successful adoption of PNBs in this resource-limited setting in Guatemala.

Methods

Overview

We conducted a prospective pilot study of a novel pilot training program for UGRA delivered at HNC in Coatepeque, Guatemala, from April 2024 through June 2024. The evaluation design and materials were adapted from the GRACE model developed by Brouillette et al [13], with the authors' permission.

Setting

HNC is the only public surgical hospital in the town of Coatepeque, Quetzaltenango, Guatemala, and serves as the primary referral center for an estimated 150,000 people in this region. A large proportion of patients come from underserved, agriculture-based communities that experience some of the highest rates of poverty and poorest health outcomes in the country [18]. This hospital has 2 main operating rooms and an additional operating room in the obstetric unit for cesarean deliveries. Based on internal hospital records, approximately 5300 surgeries are performed annually across general surgery, obstetrics and gynecology, and orthopedics and traumatology. Trauma and orthopedic extremity procedures represent a major portion of surgical volume due to the high incidence of motor vehicle accidents in this area. Neuraxial anesthesia is used preferentially when feasible, although general anesthesia is typically required for upper extremity surgery. However, due to limitations in supplies, monitoring equipment, and postoperative nursing capacity at this hospital, the use of long-acting opioid analgesics, deep sedation, and general anesthesia is avoided when possible to optimize patient safety [19].

Needs Assessment

The initial site needs assessment was conducted in February 2023 and identified both strong interest and need for UGRA. A follow-up visit with faculty from CU in August 2023 focused on developing a collaborative action plan, gathering feedback, taking inventory of available supplies, and obtaining approval from hospital leadership. The primary barrier to implementation was the absence of an ultrasound machine due to limited purchasing access. The anesthesia department already

XSL•FO
RenderX

maintained a supply of nerve block needles and local anesthetics for infrequent use of nerve stimulation for distal lower extremity blocks. Standard emergency medications were available, except for lipid emulsions. However, no procedure logs or patient consent forms were in place because a regional anesthesia workflow had not yet been established.

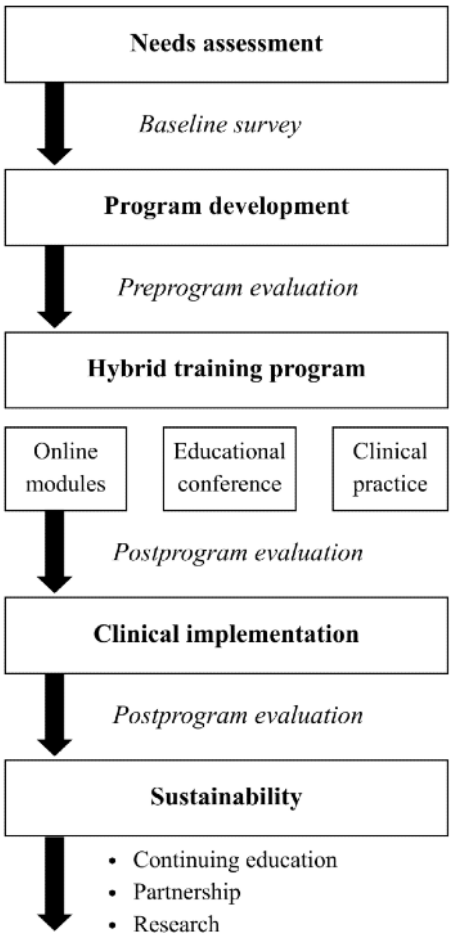
To inform curriculum design, an online survey was distributed to HNC anesthesiologists to assess baseline experience and learning preferences (Multimedia Appendix 1). Before training, participants were asked to complete the questionnaire to capture demographic information, self-reported practices, educational needs, and motivations for participation.

Program Design and Implementation

Curriculum Development

Information gathered from site assessments, planning meetings, and the online survey was used to develop a hybrid training curriculum. The curriculum was designed collaboratively with local stakeholders and regional anesthesia specialists from CU and subsequently revised based on feedback. All 7 physician anesthesiologists at HNC were invited to participate and provided written consent. The program consisted of three core components: (1) online self-directed learning modules, (2) an in-person educational conference, and (3) clinical practice with direct supervision and bedside teaching (Figure 1). Training was conducted over a 2-month period from April 2024 to May 2024, with a follow-up site visit 4 months after clinical implementation. To simplify the program and emphasize high-yield applications, a limited number of blocks were selected based on the needs assessment and team discussions. These included interscalene and supraclavicular blocks of the brachial plexus, femoral, saphenous (adductor canal), and popliteal blocks.

Figure 1. Flowchart of the program design and evaluation.



Online Modules

The online learning curriculum was delivered as 4 weekly learning modules on a web-based platform to provide open access to independent learning. As no existing materials met our program objectives for topics or in the Spanish language, we created a dedicated website to address this need [20]. Content for weekly modules was derived from open access resources, such as publications from the New York School of Regional Anesthesia (Multimedia Appendix 2). Additionally, the option for virtual chats or question-and-answer sessions was offered weekly on Fridays at the end of each module.

Educational Conference

The educational conference consisted of a 2-day, 16-hour in-person program that included lectures, clinical demonstrations, and hands-on workshops. Sessions were open to anesthesiologists, nurse anesthetists, nursing staff, and

XSL•FO
RenderX

surgeons (refer to Multimedia Appendix 3 for conference outline).

Clinical Practice

Clinical practice was initiated following the conference, with HNC participants performing nerve blocks during work hours under the supervision of CU preceptors. The program faculty comprised multiple physician anesthesiologists with fellowship training or proficiency in regional anesthesia, ensuring the presence of at least one supervising preceptor during a 4-week period for continuity. Following 1 month of direct supervision, ongoing training and support were offered remotely to address emerging needs. A week-long follow-up visit was completed 4 months after clinical implementation to provide additional support and evaluations.

Equipment Donation

To support program initiation, CU faculty donated essential supplies, including a portable ultrasound machine, tablet, nerve block needles, gel, lidocaine, bupivacaine, and lipid emulsion. In response to site visits and internal requests, hospital leadership committed to increasing local supplies and purchased additional nerve block needles and local anesthetics. Additionally, operating room schedules of the surgical services were adjusted to increase time allotted to orthopedic surgery and increase PNB volume during the clinical practice phase.

Evaluation Framework

Evaluation Approach

The evaluation framework and instruments were adapted from the GRACE model, with approval from the study authors [13]. This model used the method proposed by Kirkpatrick and Kirkpatrick [21], a globally recognized and validated approach for evaluating the effectiveness of educational and training programs across disciplines [22]. This model evaluates outcomes across four levels: (1) reaction, (2) learning, (3) behavior, and (4) outcomes.

Reaction

Participant reactions were assessed through postprogram satisfaction surveys to solicit perceptions and feedback related to the training. This survey was adapted from the original GRACE satisfaction survey, which contained 10 statements where trainees rate their level of agreement on a 5-point Likert scale and 3 open-ended questions for program feedback and suggestions. Two open-ended questions were also added to describe the most and least helpful components of the training (Multimedia Appendix 4).

Learning

Participant learning was evaluated by analyzing changes in knowledge and clinical skills. Knowledge tests were administered before the program and 1 month after clinical implementation of UGRA. The knowledge test consisted of 24 multiple-choice questions that were adapted from the original GRACE instrument. Modifications were made to remove content not taught in this program, provide clarifications, and include items addressing identified knowledge gaps (Multimedia Appendix 5). This test was administered through an online secure platform (Qualtrics XM Platform). Procedural skill acquisition was assessed using a validated global rating scale (GRS) for UGRA developed by Chuan et al [23]. This instrument evaluates 7 domains—preparation, respect for tissue, handling, time and motion, instrument handling, flow, knowledge, and overall performance—graded on a 5-point scale (Multimedia Appendix 6). Participants were observed during procedures and graded at program initiation, 1 month, and 4 months. To further capture participants' perceptions of the program, additional postsurvey questions were included to self-report changes in confidence, knowledge, and procedural skills using a 5-point Likert scale. All adapted GRACE evaluation instruments were translated into Spanish by 2 language-certified researchers and reviewed for cultural and linguistic accuracy by a local physician coinvestigator.

Behavior

Participants were prompted to report their level of interest, plans to continue performing PNBs, and perceived impact on patient care on the postprogram survey at program completion. The number of procedures performed by the anesthesiologist was recorded to assess adoption into practice.

Clinical Outcomes

Clinical outcomes were analyzed by total number and type of PNB, success and complication rate, and anesthesia method for 10 months after clinical implementation. Block success was defined a priori by the local department of anesthesiology at the HNC as a patient's ability to tolerate surgery without general anesthesia or an analgesic block with moderate to complete pain relief. To address the lack of workflow, a block documentation form and clinical log were developed in Spanish based on the New York School of Regional Anesthesia Universal Documentation Sheet for Peripheral Nerve Blocks to record and track procedure information, and patient consent forms were adapted from CU consent documentation (Multimedia Appendices 7 and 8) [24]. Additionally, a safety surveillance system was implemented to evaluate for complications and ensure resolution of the nerve block and lack of new neurological symptoms before discharge. Additional training was provided to nursing staff by a specialized CU postanesthesia care unit nurse to manage and identify complications in patients who received PNBs.

Statistical Analysis

Descriptive statistics were calculated and reported as frequencies, percentages, and means with SDs. Preprogram and postprogram knowledge test and GRS scores were summarized using medians and ranges. Changes in scores between preprogram and postprogram tests were assessed using paired Wilcoxon signed-rank tests. Likert scale outcomes across multiple time points were evaluated with the Friedman test, followed by pairwise Wilcoxon signed-rank tests for post hoc comparisons. Effect sizes for the Friedman test (Kendall W) and for Wilcoxon signed-rank tests were quantified with 95% CIs estimated using the percentile bootstrap method from 1000 resamples [25,26]. Standardized effect sizes were interpreted using Cohen general guidelines (0.2=weak, 0.5=medium, and 0.8=strong) while recognizing that these thresholds are context

XSL•FO
RenderX

dependent and should be interpreted cautiously. Interrater reliability (IRR) for preprogram GRS items, assessed across 2 fixed raters, was quantified using a 2-way mixed-effects intracorrelation coefficient and weighted κ, with 95% CIs obtained using the percentile bootstrap method from 1000 resamples. Two-sided *P* values <.05 were considered statistically significant, with post hoc comparisons corrected for false discovery rate. Given the small sample size, these results may not be generalizable and should be interpreted as exploratory. Statistical analyses were performed using R (version 4.5.1; R Foundation for Statistical Computing). The cohort sample size was determined by the number of anesthesiologists available to perform PNBs at HNC.

Ethical Considerations

This evaluation received approval from the local research ethics committee (001-2024) and was considered exempt by the Colorado Multiple Institutional Review Board (24-0212). All participants provided written informed consent before enrollment. To protect participant privacy, all data were deidentified and managed in accordance with institutional data security standards, with access restricted to authorized study personnel. Results are reported in aggregate to prevent individual identification. Participants did not receive financial or other compensation for participation in the study.

Results

Overview

All anesthesiologists from the HNC completed the training, with a 100% (7/7) response rate for surveys and evaluations. Baseline participant characteristics are provided in Table 1 (complete preprogram survey responses are provided in Multimedia Appendix 9). While all participants had previously performed PNBs using landmark or nerve stimulator techniques, none had previous experience with ultrasound guidance. Learning style preferences varied, although all participants favored workshops and hands-on clinical practice. Reported barriers included a lack of training opportunities and limited supplies.

Table 1. Baseline participant demographics (N=7).

Characteristic	Values
Participant demographics	
Participants, n (%)	7 (100)
Female, n (%)	4 (57.1)
Experience (y), mean (SD)	9.1 (5.1)
Have you used ultrasound previously? n (%)	
Yes	0 (0)
No	7 (100)
Do you have previous experience with blocks? n (%)	
Yes	7 (100)
No	0 (0)
What blocks have you performed?[a] n (%)	
Supraclavicular	2 (28.6)
Axillary	5 (71.4)
Peripheral nerve (radial, medial, and ulnar)	4 (57.1)
Ankle	5 (71.4)
Sciatic	2 (28.6)
Scalp	1 (14.2)
What is your primary motivation for participating in this program?[a] n (%)	
Learning new clinical skills	6 (85.7)
Patient outcomes	5 (71.4)
Work satisfaction	2 (28.6)
Patient satisfaction	4 (57.1)
Decreased resource use	3 (42.8)

[a]Multiple responses were allowed per participant.

Each participant completed the full online curriculum, attended both in-person workshops, and performed multiple supervised

XSL•FO
RenderX

UGRA procedures during the clinical practice period. The number of blocks performed by each participant varied widely in the 10 months following implementation (range 14-41).

Reaction Outcomes

The overall satisfaction scores for the program were high (Table 2). The most frequently cited strengths were the integration of theory with personalized hands-on practice. No component was identified as unhelpful; however, participants suggested increasing collaboration with surgical specialties and extending the duration of trainers on-site.

Table 2. Participants' reactions to the training program.

Survey question	Score[a], mean (SD)
Performing my own blocks during work hours was useful	4.71 (0.49)
The content of the program was satisfactory	4.86 (0.38)
The online study program helped prepare me for the in-person sessions	4.86 (0.38)
The conference presentations were useful	4.86 (0.38)
The program was applicable to my practice	5.00 (0)
The program should continue at HNC[b]	5.00 (0)
It was useful to practice with ultrasound during the workshops	5.00 (0)
The clinical teaching and supervision were helpful during procedures	5.00 (0)
The knowledge test was a fair evaluation of the course material	5.00 (0)
This program should be implemented in other hospitals	5.00 (0)

[a]Responses rated using a 5-point Likert scale (1=strongly disagree; 5=strongly agree).

[b]HNC: Hospital Nacional de Coatepeque.

Learning Outcomes

As shown in Figure 2, all participants demonstrated improved knowledge of UGRA. Knowledge test scores increased from a median of 45.8% (11/24; IQR 33.2%-54.0%) to 66.7% (16/24; IQR 58.3%-87.5%), representing a median absolute improvement of 20.8 (IQR 16.7%-41.7%; P=.02) percentage points with a large effect size (r=0.90, 95% CI 0.896-0.913). Similarly, all participants reported perceived gains in knowledge and confidence with UGRA (Table 3). Procedural skills improved following training across all domains, with median GRS scores increasing from 1.2 before the program to 3.4 at 1 month and 3.0 at 4 months (Figure 3). This corresponded to a median percentage increase of 147.1% (IQR 96.9%-197.3%; P=.03) at 1 month and 131.4% (IQR 90.5%-172.3%; P=.04) at 4 months, with large effect sizes at both time points (r=0.90, 95% CI 0.90-0.92 at 1 month; r=0.91, 95% CI 0.90-0.95 at 4 months). IRR for average preprogram scores showed moderate-to-high agreement with wide confidence intervals (intracorrelation coefficient=0.69, 95% CI 0.31-0.88; κ=0.63, 95% CI 0.47-0.93). The IRR for individual skills is provided in Multimedia Appendix 10, although the IRR could not be meaningfully calculated for some skills due to minimal variability. Only preprogram items had multiple raters due to the limited on-site availability of research personnel, providing preliminary evidence of measurement reliability. IRR was not assessed for postprogram items.

XSL•FO
RenderX

Figure 2. Participant knowledge test scores before training and 1 month after training.

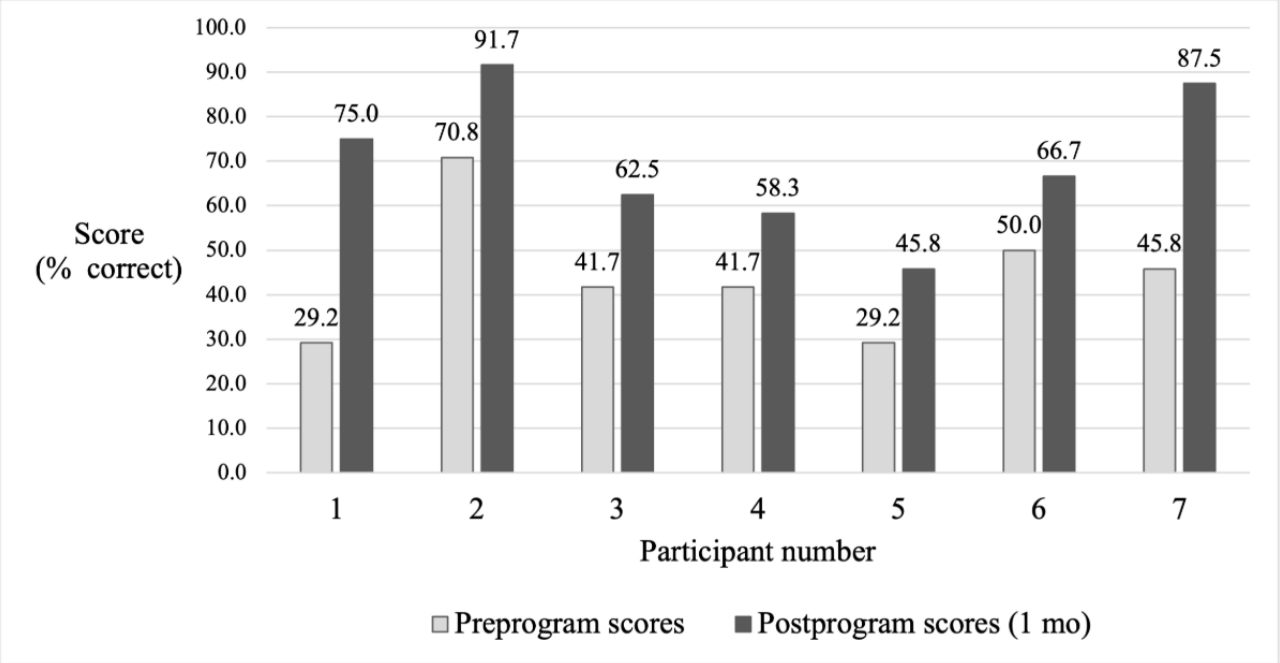
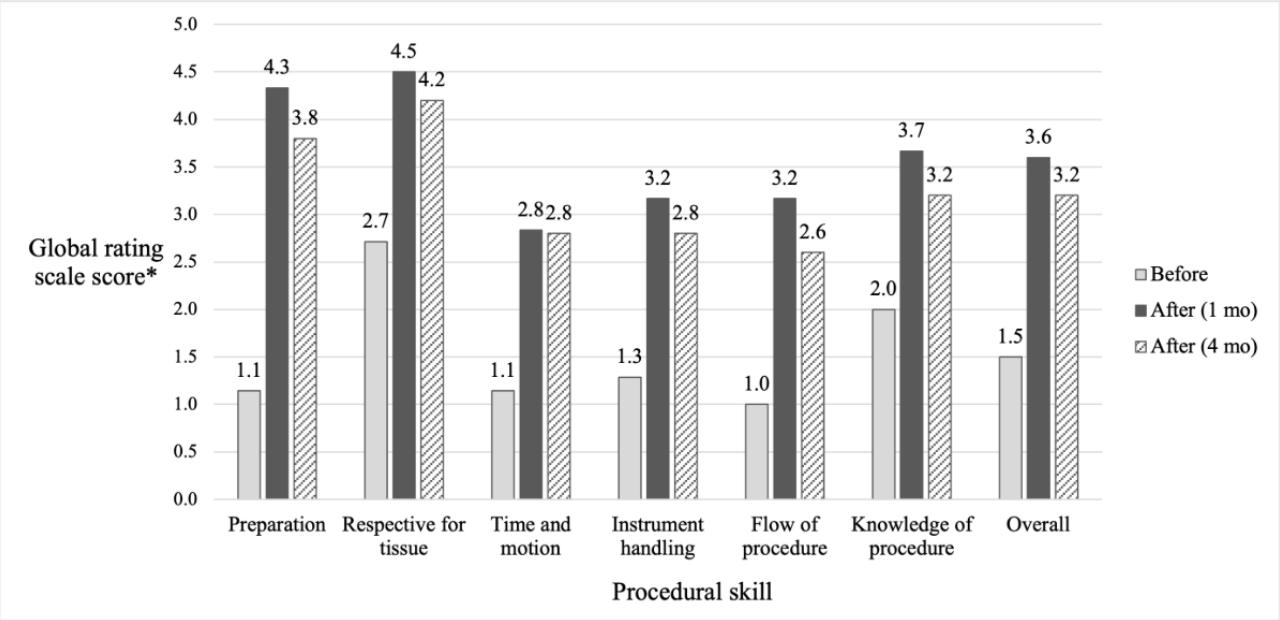


Figure 3. Change in procedural skills before and after the program at 1 and 4 months, measured by average global rating scale scores. *Using a 5-point scoring system (1=poor performance; 5=superior performance).



XSL•FO
RenderX

Table 3. Participants' self-reported changes after participating in the training program.

Survey question	Score[a], mean (SD)
I feel that I have improved my ability to perform blocks	4.29 (0.76)
My knowledge of PNBs[b] has increased	4.43 (0.79)
I feel more comfortable using and interpreting ultrasound	4.57 (0.53)
I have a stronger desire to implement blocks in my practice	4.86 (0.38)
I plan to continue using PNBs in my practice	4.57 (0.53)
I feel that PNBs improve patient care at HNC[c]	5.00 (0)

[a]Responses rated using a 5-point Likert scale (1=strongly disagree; 5=strongly agree).

[b]PNB: peripheral nerve blocks.

[c]HNC: Hospital Nacional de Coatepeque.

Behavior Outcomes

All participants reported high motivation to continue performing UGRA and intention to integrate PNBs into their practice (Table 3). All anesthesiologists continued to perform ultrasound-guided nerve blocks through the follow-up period after implementation.

Clinical Outcomes

A total of 54 blocks were performed in the first month under direct supervision, and 187 total procedures were recorded during the 10 months following implementation (Table 4). Patient demographic and case information are provided in Multimedia Appendix 11. The overall procedure success rate was 96.3% (180/187), with no reported patient complications. The supraclavicular block was the most frequently performed (66/187, 35.3%). General anesthesia was avoided in 95.9% (70/73) of the patients for upper limb surgery through successful brachial plexus blocks. On the postprogram survey, all 7 anesthesiologists strongly agreed that UGRA improves patient care at HNC.

Table 4. Ultrasound-guided nerve blocks performed during the first 10 months after program implementation by type (n=187).

Block type	Blocks, n (%)
Interscalene	10 (5.3)
Supraclavicular	66 (35.3)
Femoral	35 (18.7)
Adductor canal	25 (13.4)
Popliteal	40 (21.4)
Other	11 (5.9)

Discussion

Principal Findings

We describe the successful design and implementation of a hybrid training program at a resource-limited hospital in southwestern Guatemala. This program resulted in substantial growth in knowledge and technical skills, practice change, and enhanced patient care for a 10-month period. The educational model developed for this context offers a framework for introducing UGRA services in LMICs where clinical capabilities are lacking.

Educational Program Design

This program featured a unique educational setting, requiring input from local stakeholders to ensure alignment with daily clinical practice and local needs. Unlike previous initiatives, such as the original GRACE program, which aimed to expand preexisting UGRA capacity, anesthesiologists at HNC had no previous exposure to ultrasound or ultrasound-guided procedures. This lack of training presented challenges due to the steep learning curve and high-stakes environment in which these procedures are performed, reflecting the reality for much of the anesthesia workforce in LMICs [12]. Additionally, the paucity of available Spanish-language UGRA resources necessitated the translation or development of all surveys, lectures, and training materials in Spanish.

Multiple approaches have been used to deliver education in UGRA [27]. International programs have relied on either in-person training or predominantly online or virtual formats, with variable success [13,15,28-30]. Each method has inherent advantages and limitations. In-person training provides direct interaction with learners, hands-on skill development, real-time adaptation and feedback, and access to resources and equipment but is constrained by travel, time away from clinical duties, short program duration, and higher costs [31]. Conversely, online-only training offers flexibility, continuity, lower costs, and the ability to record and revisit content. However, it is limited by reliance on visual learning, time-zone variability, and internet connectivity and lacks in-person teaching and supervision. The duality of our methodology leverages the benefits of both approaches while minimizing the challenges and limitations.

XSL•FO
RenderX

Given the limited exposure to UGRA and trainer availability, a hybrid design was selected to maximize learning and skill development. Initial self-directed learning was used to establish foundational knowledge, particularly for theoretical concepts such as ultrasound physics, pharmacology, and management of complications. Concepts were then reinforced with classroom instruction and followed by workshops and clinical practice to support technical skill acquisition. This "flipped classroom" approach prepares learners for hands-on experiential learning and has become a standard in medical education [32]. Evidence suggests that combining self-directed learning with expert clinical instruction reduces procedural errors and improves performance [33,34]. Using this multifaceted approach, HNC anesthesiologists achieved proficiency and confidence with UGRA after 1 month and, importantly, maintained these skills even after direct trainer support had ended. Blended learning has been effectively integrated into other medical contexts [35-38]. Increasing access and familiarity with e-learning tools and resources will continue to expand opportunities for student-centered medical education worldwide.

Impact on Patient Care

A notable impact of UGRA in this setting was the avoidance of general anesthesia for upper limb surgery through ultrasound-guided brachial plexus blocks. This was achieved in 70 (95.9%) of the 73 patients over the 10-month follow-up period. Regional anesthesia has well-documented advantages over general anesthesia, such as fewer perioperative respiratory events, reduced opioid use, superior pain management, and improvement in some procedure-specific outcomes [39]. These benefits are more pronounced in resource-limited settings, where constraints in supplies, equipment, and staffing reduce safety margins. Large datasets from Médecins Sans Frontières (Doctors Without Borders) facilities in LMICs similarly reported lower mortality with regional anesthesia in comparison to general endotracheal anesthesia [40]. Analgesic blocks were also performed during training and clinical practice and were perceived by both patients and surgeons to provide superior pain control compared with the existing nonopioid regimens, although formal evaluation is warranted. Nevertheless, all anesthesia participants unanimously agreed on the postprogram survey that PNBs improved patient care at their institution.

Patient safety was a priority for this pilot program. Ultrasound-based interventions require careful technique to achieve consistent and safe outcomes. Given that improper ultrasound application or technique may lead to unintended physiological effects or nerve injury, hands-on training becomes essential to ensure safe use [41,42]. In our program, no complications were observed, and the block failure rate of 3.7% was consistent or superior compared to published rates ranging from 5% to 10% [43-45].

Lessons Learned

Implementing a regional anesthesia service at a resource-limited public hospital in Guatemala required adaptability and creative problem-solving. The preoperative workflow presented challenges, as patients typically arrived directly at the operative area with minimal preparation time, limiting the window for block placement and onset. Through pharmacological adjustments, the combination of the shorter-acting lidocaine with bupivacaine anecdotally accelerated the onset of blocks. Consent was typically not obtained by the anesthesia team; therefore, a new process was established to obtain authorization for procedures, especially parental consent in the case of minors, where parents were required to provide consent. In patients who were illiterate, we integrated the local protocol of reading the consent form information, allowing for questions, and then obtaining fingerprint stamps on forms. Technical limitations with portable ultrasound devices further complicated implementation. The prolonged procedure times during skill development led to probe overheating or battery depletion, and limited penetration of deeper structures hindered visualization in some blocks. The anesthesiologists' unique 24-hour schedule every sixth day also required careful planning to ensure consistent practice and skill reinforcement. Despite variation in the number of blocks performed by trainees, consistent improvements in knowledge and skills were observed across all participants. Finally, establishing this new service necessitated collaboration with departments beyond anesthesiology and surgery to address issues involving supply chain, documentation processes, emergency preparedness protocols, postoperative monitoring, and nursing education on the wards.

Multiple strategic decisions contributed to the program's success. First, the smaller cohort size, extended trainer presence on-site, and longer follow-up period increased direct interaction with experts, maximizing learning and enabling assessment of clinical practice and outcomes. Second, multidisciplinary coordination allowed for adjustment of the surgical schedule to dedicate one operating room to orthopedics, increasing block volume and creating greater opportunities for supervised practice. Third, limiting the focus to a few high-yield blocks proved effective for trainees with minimal baseline experience, as it allowed for focused application of learning and consolidation of skills. As proficiency grew with these limited blocks and ultrasound maneuvering, we then observed the adoption of new blocks among the group at HNC. Fourth, the development of data collection tools, such as the procedure log, assisted with real-time monitoring of progress, quality improvement, supply procurement, and research efforts. Finally, active engagement with hospital leadership and the Ministry of Public Health and Welfare, Guatemala not only reinforced the program's value but also generated public recognition, laying the foundation for potential expansion across the national health system.

Future Directions

While the initial results are encouraging, future efforts must continue to prioritize sustainability, ensuring progress is both durable and scalable. This includes implementing strategies for long-term follow-up, identifying and training local experts, and maintaining buy-in from local stakeholders [12]. Our group has engaged in regular communication and quarterly visits to reinforce relationships and provide mentorship. A long-term partnership with the CU Global Anesthesia program will include ongoing peer-to-peer support, faculty engagement, research collaborations, and resident rotations. Ultimately, the shared objective of this program is to develop local experts who can

XSL•FO
RenderX

then train and support colleagues in the region. This "train-the-trainer" model has the potential to foster a self-sustaining system of local expertise and training cascades to build long-term capacity within the health system [46,47]. Efforts are underway to expand UGRA training to residency programs, equipping anesthesiologists early in their careers and integrating these techniques into both public and private practice, thereby expanding the scope and consistency of anesthesia care nationwide. Additionally, given the growing role of UGRA, future training initiatives could include collaboration with emergency physicians at HNC to integrate ultrasound-guided nerve block techniques into emergency care, allowing patients to benefit from regional anesthesia earlier in their clinical course.

Collaboration with local stakeholders, including nursing staff, surgeons, and hospital administration, is essential to the long-term success of this program. Hospital leadership engagement is critical to ensure the purchasing of supplies and medications needed to maintain a successful and safe UGRA program. To sustain interdisciplinary commitment, research should prioritize patient satisfaction, clinical outcomes, system-level impact, and cost analyses, thereby demonstrating value in patient care and resource optimization [48].

These data can also support efforts to increase access to ultrasound technology. Although donation-based models have helped bridge initial gaps, there remains a need for improved purchasing mechanisms and sustainable supply chains for portable devices in Guatemala and other Latin American countries where demand is high. Advances in technology continue to create new opportunities for training and patient safety, and recent studies demonstrate that technology-assisted models can enhance clinical decision-making and expand training capacity across multiple medical disciplines, even in resource-limited settings [49]. Technology-driven clinical training models increase accessibility, standardization, and clinical accuracy. This hybrid UGRA program, combining structured digital content with hands-on practice, offers a pragmatic approach to overcoming training barriers in LMICs.

Limitations

There are multiple limitations to this study. First, training was delivered at a single hospital site with a small cohort of anesthesiologists. While this helped ensure quality control and exposure to practice with trainers, this limits generalizability to other departments, residency programs, or hospital systems. Second, there is a risk of information biases inherent in the observational study methods, such as recall or observer bias. Self-reporting may yield incomplete or inaccurate data; assessments of technical competency were performed by the training team with single raters for follow-up assessments, which may introduce expectancy and Hawthorne bias. Future studies should incorporate blinded, independent assessors or video adjudication to enhance objectivity and IRR. Additionally, although evaluation materials were adapted from validated sources, the translated GRACE instruments were not formally validated for cultural or psychometric equivalence, which may limit measurement accuracy. Program fidelity measures, such as detailed module completion analytics, were not available for all participants, which may limit precise quantification of exposure. However, attendance and active participation were confirmed for all anesthesiologists.

The costs of delivering in-person training can be prohibitive and should be considered when planning similar educational initiatives. Previous UGRA programs have reported costs up to US $5000 per trainer per trip [13]. This program minimized costs by integrating the hybrid educational model. Thus, costs were lower and estimated to be an average of US $1500 to US $2000 per trainer per trip. We believe that this model offers a pragmatic and cost-efficient approach for future international collaborations.

Conclusions

This evaluation demonstrates the successful implementation of a hybrid training model for UGRA at a public hospital in Guatemala, advancing perioperative pain management in a resource-limited setting. By combining online modules, in-person didactics, and supervised clinical practice, the program achieved substantial improvements in anesthesiologists' knowledge, procedural skills, and confidence. Despite a lack of previous capabilities, this program resulted in the adoption of UGRA, with sustained clinical practice observed over a 10-month follow-up period. These outcomes reflect both a tailored curriculum and strong local engagement, highlighting the model's value, safety, and potential scalability in LMICs. The avoidance of general anesthesia in upper limb surgeries underscores UGRA's potential to improve patient safety and optimize resources within the public health system. Future efforts should prioritize local expert development, interdisciplinary collaboration, and increased access to ultrasound devices to ensure the long-term sustainability and expansion of UGRA, thereby transforming anesthesiology care for underserved communities across Guatemala.

Acknowledgments

The authors thank Drs Rafael Menes and Juan Nájera, director and vice director of Hospital Nacional de Coatepeque, for their support in this program and Dr Kimberly Mendoza for her contributions. The authors also thank the Global Outreach Program of Butterfly Network, Inc, for their support.

Funding

No external financial support or grants were received from any public, commercial, or not-for-profit entities for the research, authorship, or publication of this paper.

XSL•FO
RenderX

Data Availability

The datasets generated or analyzed during this study are available from the corresponding author on reasonable request.

Authors' Contributions

Conceptualization: JEG (lead), NGSD (supporting), CGS (supporting)
Formal analysis: JEG (lead), IT-D (supporting), CGS (supporting)
Investigation: JEG (lead), NGSD (supporting), CGS (supporting)
Methodology: JEG, IT-D (supporting), NGSD (supporting), ABH (supporting)
Project administration: JEG (lead), NGSD (equal), CGS (supporting)
Writing—original draft: JEG (lead), IT-D (supporting), RF (supporting), MAV (supporting), MAG (supporting), ABH (supporting), CGS (supporting)
Writing—review and editing: JEG (lead), IT-D (supporting), CGS (supporting)

Conflicts of Interest

None declared.

Multimedia Appendix 1
Preprogram participant survey.
[DOCX File , 19 KB - mededu_v12i1e84181_app1.docx]

Multimedia Appendix 2
Curriculum outline of online learning modules.
[DOCX File , 18 KB - mededu_v12i1e84181_app2.docx]

Multimedia Appendix 3
Outline of educational conference topics.
[DOCX File , 17 KB - mededu_v12i1e84181_app3.docx]

Multimedia Appendix 4
Postprogram participant survey.
[DOCX File , 19 KB - mededu_v12i1e84181_app4.docx]

Multimedia Appendix 5
Ultrasound-guided regional anesthesia knowledge test.
[DOCX File , 8133 KB - mededu_v12i1e84181_app5.docx]

Multimedia Appendix 6
Global rating scale for procedural skills.
[DOCX File , 17 KB - mededu_v12i1e84181_app6.docx]

Multimedia Appendix 7
Procedure documentation form.
[PDF File (Adobe PDF File), 42 KB - mededu_v12i1e84181_app7.pdf]

Multimedia Appendix 8
Procedure consent form.
[DOCX File , 20 KB - mededu_v12i1e84181_app8.docx]

Multimedia Appendix 9
Results of preprogram participant survey.
[DOCX File , 19 KB - mededu_v12i1e84181_app9.docx]

Multimedia Appendix 10
Interrater reliability for global rating scale assessment scores.
[DOCX File , 16 KB - mededu_v12i1e84181_app10.docx]

XSL•FO
RenderX

Multimedia Appendix 11
Characteristics of patients and procedures performed during the ultrasound-guided regional anesthesia training program.
[DOCX File , 17 KB - mededu_v12i1e84181_app11.docx]

References

1. Albrecht E, Chin KJ. Advances in regional anaesthesia and acute pain management: a narrative review. Anaesthesia 2020 Jan;75 Suppl 1:e101-e110. [doi: 10.1111/anae.14868] [Medline: 31903582]
2. Resta F, Barcella B, Angeli V, Lago E, Santaniello A, Dedato AS, et al. Simulation-based training in ultrasound-guided regional anaesthesia for emergency physicians: insights from an Italian pre/post intervention study. BMC Med Educ 2024 Dec 20;24(1):1510 [FREE Full text] [doi: 10.1186/s12909-024-06500-0] [Medline: 39707369]
3. Tucker RV, Peterson WJ, Mink JT, Taylor LA, Leech SJ, Nagdev AD, et al. Defining an ultrasound-guided regional anesthesia curriculum for emergency medicine. AEM Educ Train 2021 Jul;5(3):e10557 [FREE Full text] [doi: 10.1002/aet2.10557] [Medline: 34124505]
4. Di Pietro S, Maffeis R, Jannelli E, Mascia B, Resta F, De Silvestri A, et al. Comparing the pericapsular nerve group block and fascia iliaca block for acute pain management in patients with hip fracture: a randomised clinical trial. Anaesthesia 2025 Dec;80(12):1484-1492. [doi: 10.1111/anae.16695] [Medline: 40727959]
5. Di Pietro S, Mascia B, Lo Bianco G, Perlini S, Iotti GA. Anterior cutaneous nerve block for analgesia in anterior chest trauma: is the parasternal approach necessary? Clin Exp Emerg Med 2020 Mar;7(1):67-70 [FREE Full text] [doi: 10.15441/ceem.18.089] [Medline: 32252136]
6. Di Pietro S, Caracciolo E, Barcella B, Perlini S, Regional Anaesthesia in Emergency Medicine Research Group. Superficial cervical plexus block in emergency departments: rationale for its use in incision and drainage of neck skin abscesses. Intern Emerg Med 2022 Aug;17(5):1533-1536. [doi: 10.1007/s11739-022-02992-0] [Medline: 35670893]
7. Schnittger T. Regional anaesthesia in developing countries. Anaesthesia 2007 Dec;62 Suppl 1:44-47 [FREE Full text] [doi: 10.1111/j.1365-2044.2007.05297.x] [Medline: 17937713]
8. Rukewe A, Fatiregun A. The use of regional anesthesia by anesthesiologists in Nigeria. Anesth Analg 2010 Jan 01;110(1):243-244. [doi: 10.1213/ANE.0b013e3181c0f00e] [Medline: 19933532]
9. Ho M, Livingston P, Bould MD, Nyandwi JD, Nizeyimana F, Uwineza JB, et al. Barriers and facilitators to implementing a regional anesthesia service in a low-income country: a qualitative study. Pan Afr Med J 2019;32:152 [FREE Full text] [doi: 10.11604/pamj.2019.32.152.17246] [Medline: 31303923]
10. Admassie BM, Admass BA, Melesse DY. Practice and challenges related to regional anesthesia in Amhara regional hospitals, Northwest-Ethiopia: a web-based survey study. BMC Anesthesiol 2024 Nov 04;24(1):398 [FREE Full text] [doi: 10.1186/s12871-024-02783-4] [Medline: 39491037]
11. Ranger BJ, Bradburn E, Chen Q, Kim M, Noble JA, Papageorghiou AT. Portable ultrasound devices for obstetric care in resource-constrained environments: mapping the landscape. Gates Open Res 2023;7:133 [FREE Full text] [doi: 10.12688/gatesopenres.15088.2] [Medline: 39935587]
12. Dohlman LE, Kwikiriza A, Ehie O. Benefits and barriers to increasing regional anesthesia in resource-limited settings. Local Reg Anesth 2020;13:147-158 [FREE Full text] [doi: 10.2147/LRA.S236550] [Medline: 33122941]
13. Brouillette MA, Aidoo AJ, Hondras MA, Boateng NA, Antwi-Kusi A, Addison W, et al. Regional anesthesia training model for resource-limited settings: a prospective single-center observational study with pre-post evaluations. Reg Anesth Pain Med 2020 Jul;45(7):528-535. [doi: 10.1136/rapm-2020-101550] [Medline: 32447288]
14. Dohlman LE, Thakkar N, Jivanelli B, Pakala S, Brouillette MA, Global Regional Anesthesia and Pain Management Research Group. Regional anesthesia global health collaborations- a scoping review of current intervention methods. Curr Opin Anaesthesiol 2022 Oct 01;35(5):647-653. [doi: 10.1097/ACO.0000000000001178] [Medline: 35942724]
15. Liu M, Salmon M, Zaidi R, Nagdev A, Debebe F, Muller MF, et al. Ultrasound-guided regional anesthesia: feasibility and effectiveness of teaching via telesimulation in Ethiopia. Reg Anesth Pain Med 2021 Aug;46(8):722-726. [doi: 10.1136/rapm-2020-102394] [Medline: 33903219]
16. Avila C, Bright R, Gutierrez J, Hoadley K, Manuel C, Romero N, et al. Guatemala health system assessment 2015. Health Finance & Governance Project. 2015. URL: https://internationalmedicalrelief.org/wp-content/uploads/2019/07/Guatemala-HSA-_ENG-FULL-REPORT-FINAL-APRIL-2016.pdf [accessed 2025-12-25]
17. Zha Y, Truché P, Izquierdo E, Zimmerman K, de Izquierdo S, Lipnick MS, et al. Assessment of anesthesia capacity in public surgical hospitals in Guatemala. Anesth Analg 2021 Feb 01;132(2):536-544. [doi: 10.1213/ANE.0000000000005297] [Medline: 33264116]
18. Gamboa JE, Bolaños AG, Simmons CG. Community perceptions of accessing surgical and anesthetic care in rural Guatemala: a cross-sectional survey. Cureus 2025 Jun;17(6):e85783. [doi: 10.7759/cureus.85783] [Medline: 40656399]
19. Turner RA, Simmons CG, Ramirez S, Gamboa JE. Hypoxemia and postoperative monitoring after anesthesia: a prospective observational study using portable pulse oximetry in a resource-limited setting in Guatemala. Cureus 2025 Jan;17(1):e78075. [doi: 10.7759/cureus.78075] [Medline: 40013179]
20. Programa de estudio. Regional anestesia Guatemala. URL: https://anestesiaregionalguatemala.com [accessed 2025-06-01]

XSL•FO
RenderX

21. Kirkpatrick JD, Kirkpatrick WK. Kirkpatrick's Four Levels of Training Evaluation. Alexandria, VA: Association for Talent Development; 2016.
22. Anderson LN, Merkebu J. The Kirkpatrick model: a tool for evaluating educational research. Fam Med 2024 Jun;56(6):403 [FREE Full text] [doi: 10.22454/FamMed.2024.161519] [Medline: 38996342]
23. Chuan A, Graham PL, Wong DM, Barrington MJ, Auyong DB, Cameron AJ, et al. Design and validation of the Regional Anaesthesia Procedural Skills Assessment Tool. Anaesthesia 2015 Dec;70(12):1401-1411 [FREE Full text] [doi: 10.1111/anae.13266] [Medline: 26558857]
24. Gandhi K, Patel V, Maliakal T, Xu D, Flisinski K. Universal documentation sheet for peripheral nerve blocks. J NY Sch Reg Anest 2009 May;12:23 [FREE Full text]
25. Rosenthal R. Meta-Analytic Procedures for Social Research. Thousand Oaks, CA: SAGE Publications; 1991.
26. Rosenthal R. Science and ethics in conducting, analyzing, and reporting psychological research. Psychol Sci 1994 May 01;5(3):127-134. [doi: 10.1111/j.1467-9280.1994.tb00646.x] [Medline: 11652978]
27. Nix CM, Margarido CB, Awad IT, Avila A, Cheung JJ, Dubrowski A, et al. A scoping review of the evidence for teaching ultrasound-guided regional anesthesia. Reg Anesth Pain Med 2013;38(6):471-480. [doi: 10.1097/AAP.0b013e3182a4ed7a] [Medline: 24108249]
28. Moll V, Schmidt PC, Amos C, Workneh RS, Tadesse M, Mulugeta E, et al. Building regional anesthesia capacity in limited-resource settings: a pilot study evaluating a 4-week curriculum. Pain Manag 2021 Jan;11(1):29-37. [doi: 10.2217/pmt-2020-0044] [Medline: 33073715]
29. Burckett-St Laurent DA, Cunningham MS, Abbas S, Chan VW, Okrainec A, Niazi AU. Teaching ultrasound-guided regional anesthesia remotely: a feasibility study. Acta Anaesthesiol Scand 2016 Aug;60(7):995-1002. [doi: 10.1111/aas.12695] [Medline: 26860837]
30. Pfenninger EG, Tugtekin I, Stahl W, Dinse A, Gorsewski G, Vicent O, et al. "Anesthesia-focused sonography": first analysis of transferring a training from Germany to Ghana. SDRP J Anesth Surg 2018;3(1):1-9 [FREE Full text] [doi: 10.25177/jas.3.1.3]
31. Maloney S, Nicklen P, Rivers G, Foo J, Ooi YY, Reeves S, et al. A cost-effectiveness analysis of blended versus face-to-face delivery of evidence-based medicine to medical students. J Med Internet Res 2015 Jul 21;17(7):e182 [FREE Full text] [doi: 10.2196/jmir.4346] [Medline: 26197801]
32. Phillips J, Wiesbauer F. The flipped classroom in medical education: a new standard in teaching. Trends Anaesth Crit Care 2022 Feb;42:4-8 [FREE Full text] [doi: 10.1016/j.tacc.2022.01.001] [Medline: 38620968]
33. de Oliveira Filho GR, Mettrau FA. The effect of high-frequency, structured expert feedback on the learning curves of basic interventional ultrasound skills applied to regional anesthesia. Anesth Analg 2018 Mar;126(3):1028-1034. [doi: 10.1213/ANE.0000000000002748] [Medline: 29261545]
34. Udani AD, Macario A, Nandagopal K, Tanaka MA, Tanaka PP. Simulation-based mastery learning with deliberate practice improves clinical performance in spinal anesthesia. Anesthesiol Res Pract 2014;2014:659160 [FREE Full text] [doi: 10.1155/2014/659160] [Medline: 25157263]
35. Koch R, Gassner L, Gerlach N, Festl-Wietek T, Hirt B, Joos S, et al. Integrated e-learning for shoulder anatomy and clinical examination skills in first-year medical students: randomized controlled trial. JMIR Med Educ 2025 Sep 17;11:e62666. [doi: 10.2196/62666] [Medline: 40961369]
36. Vallée A, Blacher J, Cariou A, Sorbets E. Blended learning compared to traditional learning in medical education: systematic review and meta-analysis. J Med Internet Res 2020 Aug 10;22(8):e16504 [FREE Full text] [doi: 10.2196/16504] [Medline: 32773378]
37. Tudor Car L, Kyaw BM, Dunleavy G, Smart NA, Semwal M, Rotgans JI, et al. Digital problem-based learning in health professions: systematic review and meta-analysis by the digital health education collaboration. J Med Internet Res 2019 Feb 28;21(2):e12945. [doi: 10.2196/12945] [Medline: 30816846]
38. Oftring ZS, Deutsch K, Tolks D, Jungmann F, Kuhn S. Novel blended learning on artificial intelligence for medical students: qualitative interview study. JMIR Med Educ 2025 May 26;11:e65220. [doi: 10.2196/65220] [Medline: 40418795]
39. Hughey S, Cole J, Drew B, Brust A, Stedjelarsen E. Regional anesthesia in resource-limited and disaster environments: a daring discourse. Reg Anesth Pain Med 2025 Aug 05;50(8):686-689. [doi: 10.1136/rapm-2024-105680] [Medline: 38942425]
40. Ariyo P, Trelles M, Helmand R, Amir Y, Hassani GH, Mftavyanka J, et al. Providing anesthesia care in resource-limited settings: a 6-year analysis of anesthesia services provided at Médecins Sans Frontières facilities. Anesthesiology 2016 Mar;124(3):561-569. [doi: 10.1097/ALN.0000000000000985] [Medline: 26881395]
41. Ozsoy U, Ogut E, Sekerci R, Hizay A, Rink S, Angelov DN. Effect of pulsed and continuous ultrasound therapy on the degree of collateral axonal branching at the lesion site, polyinnervation of motor end plates, and recovery of motor function after facial nerve reconstruction. Anat Rec (Hoboken) 2019 Aug;302(8):1314-1324 [FREE Full text] [doi: 10.1002/ar.24122] [Medline: 30950229]
42. El-Tallawy SN, Ahmed RS, Salem GI, Alzahrani TA, Haddara MM, Ahmed RH, et al. Neurological deficits following regional anesthesia and pain interventions: reviewing current standards of care. Pain Ther 2025 Jun;14(3):817-839. [doi: 10.1007/s40122-025-00726-6] [Medline: 40146371]

XSL•FO
RenderX

43. Barrington MJ, Watts SA, Gledhill SR, Thomas RD, Said SA, Snyder GL, et al. Preliminary results of the Australasian Regional Anaesthesia Collaboration: a prospective audit of more than 7000 peripheral nerve and plexus blocks for neurologic and other complications. Reg Anesth Pain Med 2009;34(6):534-541. [doi: 10.1097/aap.0b013e3181ae72e8] [Medline: 19916206]
44. Bottomley T, Gadsden J, West S. The failed peripheral nerve block. BJA Educ 2023 Mar;23(3):92-100 [FREE Full text] [doi: 10.1016/j.bjae.2022.12.002] [Medline: 36844443]
45. Cotter JT, Nielsen KC, Guller U, Steele SM, Klein SM, Greengrass RA, et al. Increased body mass index and ASA physical status IV are risk factors for block failure in ambulatory surgery - an analysis of 9,342 blocks. Can J Anaesth 2004 Oct;51(8):810-816. [doi: 10.1007/BF03018454] [Medline: 15470170]
46. Mormina M, Pinder S. A conceptual framework for training of trainers (ToT) interventions in global health. Global Health 2018 Oct 22;14(1):100 [FREE Full text] [doi: 10.1186/s12992-018-0420-3] [Medline: 30348183]
47. Anderson CR, Taira BR. The train the trainer model for the propagation of resuscitation knowledge in limited resource settings: a systematic review. Resuscitation 2018 Jun;127:1-7. [doi: 10.1016/j.resuscitation.2018.03.009] [Medline: 29545135]
48. Mohamed SS, Temu R, Komba LF, Kaino MM, Olotu FI, Ndebea AS, et al. Patient satisfaction with, and outcomes of, ultrasound-guided regional anesthesia at a referral hospital in Tanzania: a cross-sectional study. Anesth Analg 2024 Jan 01;138(1):180-186. [doi: 10.1213/ANE.0000000000006374] [Medline: 36727852]
49. Ogut E. Artificial intelligence in clinical medicine: challenges across diagnostic imaging, clinical decision support, surgery, pathology, and drug discovery. Clin Pract 2025 Sep 16;15(9):169. [doi: 10.3390/clinpract15090169] [Medline: 41002784]

Abbreviations

CU: University of Colorado
GRACE: Global Regional Anesthesia Curricular Engagement
HNC: Hospital Nacional de Coatepeque
IRR: interrater reliability
LMIC: low- and middle-income country
PNB: peripheral nerve block
UGRA: ultrasound-guided regional anesthesia

Edited by J Gentges; submitted 16.Sep.2025; peer-reviewed by S Di Pietro, E Ogut; comments to author 09.Oct.2025; revised version received 17.Dec.2025; accepted 22.Dec.2025; published 08.Jan.2026.

Please cite as:
Gamboa JE, Tamm-Daniels I, Flores R, Sarat Diaz NG, Villasenor MA, Gist MA, Hoie AB, Kurinec C, Simmons CG
Ultrasound-Guided Regional Anesthesia in a Resource-Limited Hospital: Prospective Pilot Study of a Hybrid Training Program
JMIR Med Educ 2026;12:e84181
URL: https://mededu.jmir.org/2026/1/e84181
doi:10.2196/84181
PMID:41428829

©Jakob E Gamboa, Inge Tamm-Daniels, Roland Flores, Nancy G Sarat Diaz, Mario A Villasenor, Mitchell A Gist, Aidan B Hoie, Christopher Kurinec, Colby G Simmons. Originally published in JMIR Medical Education (https://mededu.jmir.org), 08.Jan.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Education, is properly cited. The complete bibliographic information, a link to the original publication on https://mededu.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Development of Virtual Reality Scenarios Addressing Gender-Based Violence in Health Sciences Education: Qualitative Approach

Belén Valverde-Alirangues[1*], PhD; Marta Benet[2*], PhD; Mar Carrió[3*], PhD

[1]Health Sciences Education Research Group, Department of Medicine and Life Sciences (MELIS), Universitat Pompeu Fabra, Barcelona, Spain

[2]Sant Joan de Déu Research Foundation, Barcelona, Catalonia, Spain

[3]Department of Optics and Optometry, Faculty of Optics and Optometry, Universitat Politècnica de Catalunya, Carrer del Violinista Vellsola, 37, 08222 Terrassa, Barcelona, Catalonia, Spain

[*]all authors contributed equally

Corresponding Author:
Mar Carrió, PhD
Department of Optics and Optometry, Faculty of Optics and Optometry, Universitat Politècnica de Catalunya, Carrer del Violinista Vellsola, 37, 08222 Terrassa, Barcelona, Catalonia, Spain

Abstract

Background: Gender-based violence (GBV) is a public health issue affecting 1 in 3 women globally. Its impact on women's health is challenging, including physical, mental, and social consequences. Health care professionals have a unique opportunity in identifying and supporting GBV survivors, but there is a lack of adequate training.

Objective: This study aims to develop educational resources based on problem-based and experiential learning approaches using virtual reality (VR) scenarios for health sciences students to enhance their skills in addressing GBV.

Methods: A co-creation approach was adopted, encompassing 3 main strategies. First, a focus group was conducted with frontline professionals experienced in GBV. Second, co-creation workshops involved professionals from diverse fields, including higher education pedagogy, gender and public health, nursing and medical education, and immersive technology. Third, expert consultation with frontline professionals ensured coherence between the educational resources and real-world challenges. Following this phase, a first iteration of the materials was piloted with students to assess usability and relevance.

Results: The thematic analysis of the focus group content led to the identification of 9 categories illustrating the competencies and knowledge areas considered relevant to address GBV. As a result of the co-creation workshops, these categories were translated into 18 learning needs, and 4 use cases for the VR component were also identified. The VR scenarios were designed to cover critical GBV situations, fostering transversal skills, such as empathic communication, ethical decision-making, and interdisciplinary collaboration. Two didactic methodologies were proposed for each scenario: a problem-based learning sequence and a single experiential learning session approach, culminating in 4 VR videos and their methodological guides.

Conclusions: The grounding of these educational resources in real-world scenarios, in conjunction with the competencies identified by frontline health and social care professionals with expertise in GBV, ensured alignment with the challenges professionals face in their practice. This helped bridge the gap between theory and practice, offering an innovative approach to GBV education for students of health sciences.

(JMIR Med Educ 2026;12:e76098) doi:10.2196/76098

KEYWORDS

gender-based violence; women's health; virtual reality; professional education; problem-based learning

Introduction

Gender-based violence (GBV) has been identified as a prevalent and globally recognized public health problem, affecting 1 in 3 women and girls during their lifetimes [1]. The term "GBV" is an umbrella concept encompassing all forms of violence perpetrated against women, including physical, psychological, sexual, and other types. These harmful practices occur across various contexts and at different stages of women's lives. This study seeks to comprehend the phenomenon in its intricacy, which is deeply rooted in the prevailing gender inequality [2]. As intimate partner violence is the most prevalent form described, this study primarily focuses on this type due to its widespread occurrence and relevance in clinical settings. Furthermore, it represents a major concern for women's health, including not only physical injuries but also numerous mental and social problems for the survivors and their children [2].

XSL•FO
RenderX

There are multiple health problems related to GBV listed, such as sexually transmitted infections, intrauterine bleeding, gastrointestinal problems, neurological disorders, chronic pain, alcohol use, depression, anxiety, and posttraumatic stress disorder, as well as suicide and death by homicide [3]. Moreover, there are psychosocial issues intimately related to GBV survivors that are not rigorously captured within the symptomatology of mental disorders. These often include nonclinical and relational forms of distress associated with poverty, relatively weak legal systems, and protection issues, among others [4].

Therefore, because of its prevalence and high impact on women's health, addressing GBV must also be closely integrated into health care services, complementing the social care services. Health care services play a crucial role, as many GBV survivors seek care at these facilities without having their specific needs adequately addressed [5]. Health care professionals (HCPs) have a unique opportunity to detect and support survivors by promoting women's safety and helping them to prevent the worst consequences [6]. Despite acknowledging the importance of health care providers' capacity to respond effectively to cases of GBV, current scientific evidence highlights the scarcity of HCPs' abilities and skills to undertake this role [5,7].

In recent years, several educational programs on GBV have been developed for HCPs to improve their competency. Even though recent evidence on this topic suggests positive outcomes, Sprague et al [8] noted that involving professional experts and using specific resources were associated with positive program effectiveness. Numerous interventions have reported an overall positive assessment and significant effectiveness regarding knowledge and ability to address GBV in both health science students and working professionals [9-12]. However, it is challenging to preserve these skills over time [13]. Specific areas of improvement for these interventions have been recognized, such as time constraints, previous acknowledgment, and redundancy in content [14]. Therefore, before proposing any training to professionals, it is essential to determine their learning needs first, considering the sensitivity, complexity, and multifactorial nature of GBV.

Regarding the most suitable educational approach, it is essential to provide opportunities to develop skills in meaningful, real-world scenarios within a safe environment, incorporating active learning methods, such as problem-based learning (PBL), experiential learning, simulation [15], or immersive approaches. In this regard, a recent review highlights that learning is more effective when active educational strategies are used rather than relying solely on theory-based approaches [16]. Despite the documented benefits of active learning methods, it is important to acknowledge that they may also elicit a range of emotions and potentially increase stress levels among students as they engage with challenging concepts embedded in complex scenarios with a high emotional burden [17].

Virtual reality (VR) can contribute to active learning by offering immersive experiences that simulate professional situations, providing a realistic context for learning. The use of VR has been shown to be effective in enhancing learning outcomes while also significantly reducing the psychological pressure often experienced by learners [18]. Moreover, this technology has proven to be a pedagogically promising approach in terms of both productivity and learner satisfaction [18,19]. VR scenarios enable the recreation of situations that closely mirror real-life contexts, allowing for the in-depth analysis of subtle elements, such as character personality and attitudes, verbal and nonverbal communication, and contextual details within the scene.

Combining immersive VR scenarios with other active learning methodologies, such as PBL and experiential learning (EL), has the potential to enhance the learning process. In the PBL approach, the VR scenario can serve as a case study to initiate an inquiry process that leads to an in-depth exploration of causes, theoretical frameworks, and various approaches described in the literature for addressing cases of GBV. Traditionally, the PBL methodology involves a sequence of small-group tutorial sessions, concluding with a synthesis session in which the case findings are discussed and strategies for addressing the case are reflected upon [20].

Alternatively, VR scenarios can be integrated into a shorter sequence based on Kolb's experiential learning model, as revised by Morris [21], within a single session. This approach is grounded in the reflection and conceptualization of experience, where learning begins with a specific situation in which learners actively participate and which is embedded in a rich context. In this regard, VR serves as a powerful tool to introduce such experiences to students [22]. This is followed by critical reflective observation, during which learners review the experience, pose questions, challenge actions, and collaboratively analyze the situation [23]. The process then advances to abstract conceptualization, where new ideas are developed or existing ones refined [21].

Despite the numerous educational interventions described in the existing literature [24] and the encouraging outcomes frequently reported, uncertainty remains regarding what works and why. This is largely due to the heterogeneity in methods, duration, and outcome measures used across studies [25].

This study aims to develop innovative educational resources that combine active learning methodologies—including VR scenarios—to enhance the competencies of health sciences students in addressing GBV. The paper presents the co-creation process underlying the design of these resources, which involved identifying the training needs of health professionals, selecting the most relevant cases to be represented in VR, and developing both the narrative scripts and didactic guidelines to support their pedagogical integration. The following sections describe this process in detail and discuss its implications for health education and professional training.

Methods

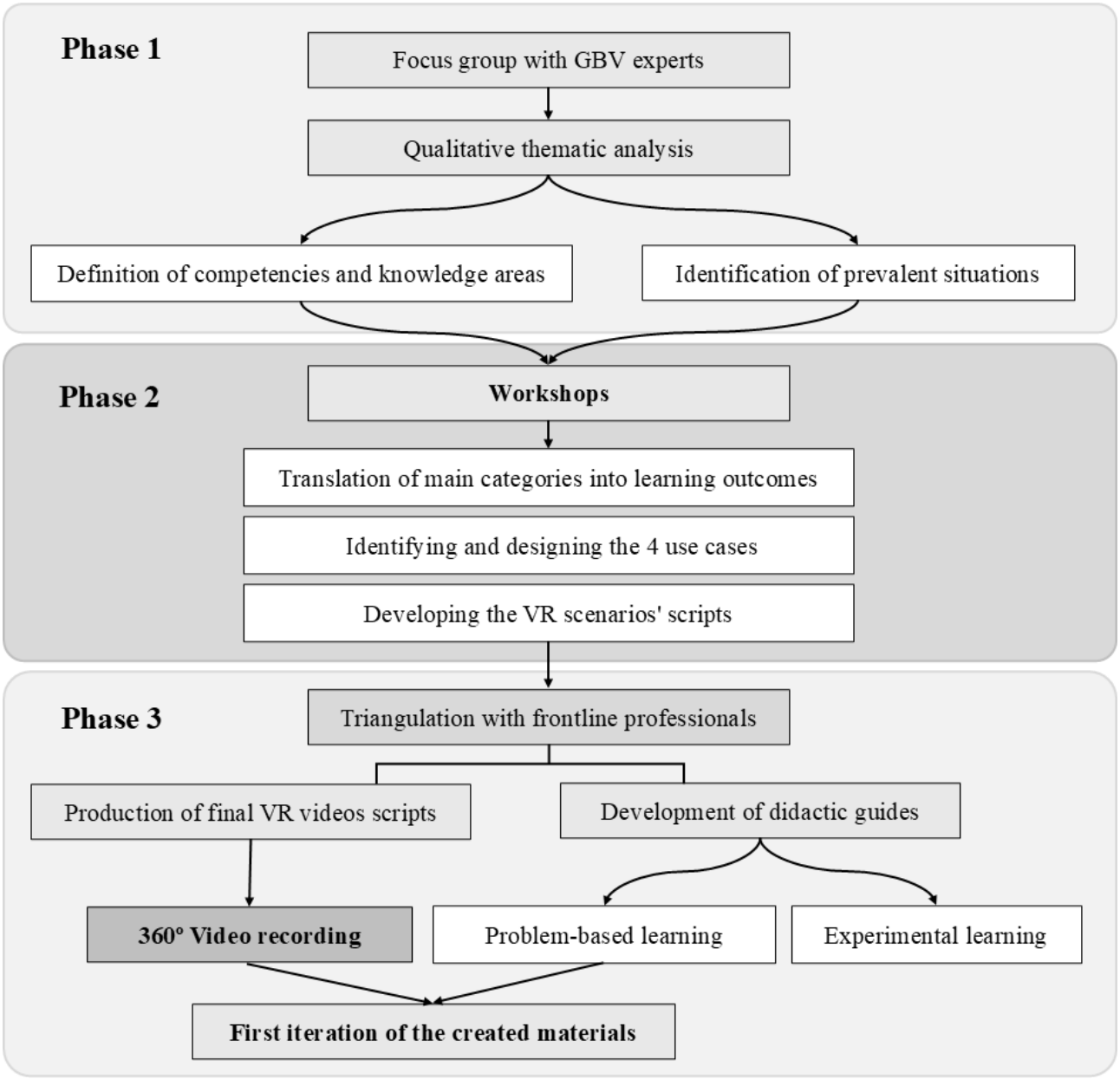
Overview

The development of the educational resources was grounded in a co-creation approach with the active involvement of frontline health and social care professionals trained and highly experienced in GBV, higher education pedagogues, lecturers, and people with expertise in immersive technologies throughout

XSL•FO
RenderX

all stages of the design process. The co-creation approach included 3 main strategies: focus group, co-creation workshops, and iterative feedback from frontline professionals. This approach was adopted to foster relevance, contextual accuracy, and practical applicability of the final materials. The overall study design was organized into 3 sequential phases, depicted in Figure 1.

Figure 1. Flowchart illustrating the co-creation process for developing the virtual reality (VR) scenarios. The figure outlines the stages through which the VR scenarios were conceptualized and validated by consulting the original experts, before ultimately being produced. GBV: gender-based violence.



The first phase consisted of a focus group to capture the day-to-day practices of the frontline professionals trained and highly experienced in GBV who take care of women who are or have been suffering from GBV. It included professionals with different profiles and was held with two main aims: (1) to establish the competencies and knowledge areas required for HCPs to address GBV and (2) to identify the key situations encountered in daily practice.

Following the focus group's results, the project moves on to its second phase, which consists of a series of co-creation workshops aimed at (1) translating the competencies and knowledge areas into learning outcomes, (2) identifying 4 use cases to cover the identified learning outcomes, and (3) developing the VR scenarios and the PBL and EL sequences.

Once the 4 use cases were established, the learning outcomes for each case were identified and used as a reference for designing the VR scenarios. This process involved scriptwriting, which included composing the scripts with dialogue, actions, and interactive elements. Additionally, the development of the characters featured in the scenarios was carried out, with key aspects of their physical and behavioral attributes defined to enhance intersectionality by incorporating additional axes of inequality, such as ethnicity, profession, migrant status, and age. To complete this second phase, a storyboard was created to visually represent the sequence of scenes. For the 2 distinct

XSL•FO
RenderX

methodologies selected to integrate the VR scenarios into educational practice—PBL and EL—an initial draft of the corresponding sequences was developed alongside the design of the VR scenarios.

In the third phase, the VR scenarios underwent expert validation by frontline experts, who were the professionals from the initial focus group. This process was undertaken to ensure rigor in terms of real-world applicability and relevance. Following the incorporation of their feedback, the final versions of the VR scenario scripts were meticulously refined to ensure maximum reliability, culminating in the recording of the scenes. Subsequently, teaching guides were developed based on the previously drafted PBL and EL sequences to facilitate and standardize the implementation of the educational resources. These guides provide a detailed description of both the pedagogical foundations for classroom application and the theoretical framework required for HCPs to address GBV. Their purpose is to serve as training materials for educators seeking to incorporate these resources into their teaching practice. Finally, a first iteration of the materials was developed to explore its efficacy and areas of improvement.

Phase 1: Exploration and Identification of Learning Needs and Use Cases

Focus Group Participants

Participants for the focus group were selected through intentional sampling, applying both homogeneity (fulfilled by all participants) and heterogeneity (to ensure diversity within the sample) criteria (Table 1). The homogeneity criterion required participants to be trained and highly experienced in GBV and to be currently working as frontline professionals in health or social care. The heterogeneity criteria included level and type of service (primary health care, hospital care, essential and specialized social services, community services for women), age (younger or older than 40 years), gender (female, male, and nonbinary), geographical area (urban or rural), training in cultural diversity, and involvement in GBV coordination bodies. The intentional sampling aimed to maximize diversity across these dimensions to ensure a broad range of perspectives.

Table . Distribution of the focus group participants according to their profiles, indicating the criteria used to assess homogeneity and heterogeneity among groups.

Participant code	Health care role								Sociodemographic axis						
	Primary health care			Hospital health care	Social services				Age (y)		Gender			Geographical area	
	Doctor or physician	Social worker	Expert on GBV[a]	Midwife	Basic	Social emergency services	Hosting services	Attention services for women	<40	>40	Female	Male	Nonbinary	Rural	Urban
P1			✓					✓		✓	✓			✓	
P2	✓		✓							✓	✓				✓
P3			✓	✓					✓		✓			✓	
P4		✓	✓		✓				✓		✓				✓
P5			✓			✓	✓			✓	✓				✓
P6		✓	✓		✓					✓	✓				✓

[a]GBV: gender-based violence.

The rationale for selecting highly experienced frontline professionals was based on the assumption that those with extensive experience in addressing GBV are best positioned to identify the competencies and knowledge areas required in daily practice. In contrast, professionals with limited experience may have a less comprehensive understanding of the skills needed to effectively manage situations they have not yet mastered. The selected participants represented a variety of organizations—including health care, social services, and community-based entities—and none were affiliated with the researchers' institutions, nor did they maintain any previous professional or personal relationship with the researchers.

Six people were selected and contacted by email, all of whom were women. The group included 2 HCPs (a doctor and a midwife), 2 social workers (one specialized in adult and child welfare, the other in addiction rehabilitation), 1 institutionally linked to the women's affairs organization, and, finally, the last participant belonged to a community-based organization that supports women affected by human trafficking.

Data Collection and Analysis

The focus group was conducted online using a semistructured script () that outlined the key issues to be explored and included a set of guiding questions for the moderator. The main topics addressed were competencies and knowledge areas relevant to

XSL•FO
RenderX

addressing GBV, how participants identify GBV situations in their daily work with women, and descriptions of typical cases to inform the design of characters (eg, survivor, perpetrator, relatives, professionals) and scenarios (eg, reasons for seeking services, nature of interactions, and service settings where detection occurs) for the VR simulations.

Two researchers were present during the focus group, which lasted 2 hours; one attended as a moderator, while the other observed. The observer documented observations and emergent themes using a predetermined template, supplementing the planned guide.

A transcription of the session was conducted to convert spoken and visual information into written data. This process used both written notes and video recordings to ensure accuracy and completeness.

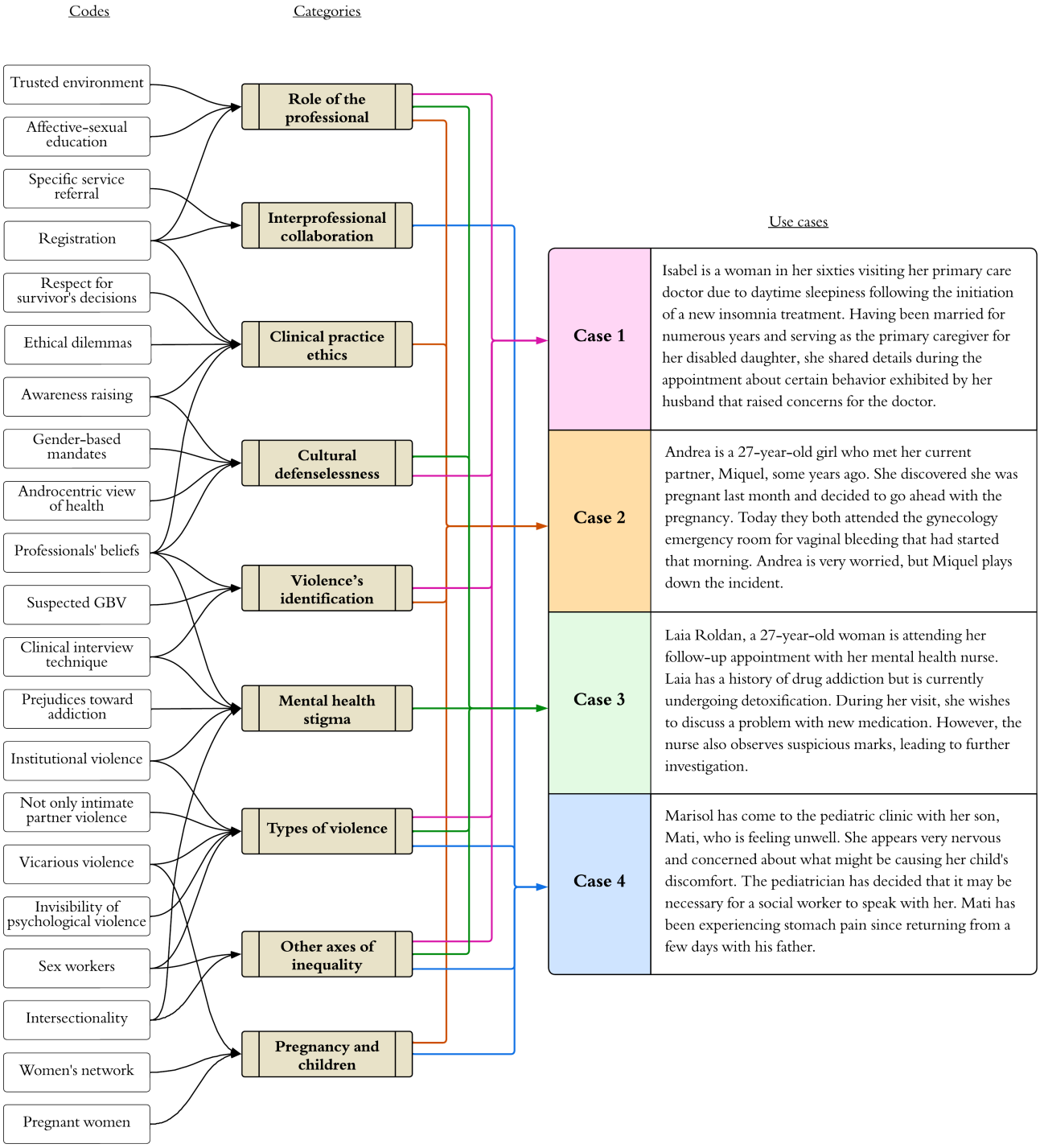
The qualitative thematic analysis of the focus group transcription was carried out primarily using an inductive approach, aiming to identify emerging competencies and knowledge areas needed to address GBV without relying on a predefined framework. The analysis was conducted independently by 3 researchers. Each researcher followed a structured process: (1) reading the transcript twice to gain a deep understanding of the content, (2) identifying relevant competencies and knowledge areas and assigning corresponding codes, (3) identifying relevant case examples and creating additional codes, and (4) reviewing the coding to ensure internal consistency.

The coding process was conducted using a Microsoft Word document, and the results were summarized in a comparative table. The findings were then discussed in a group meeting, applying researcher triangulation [26], a technique shown to reduce potential biases in data interpretation. This triangulation exercise also served to reach consensus on the final definition of categories.

As a result of this analysis, the most prevalent situations were identified and selected as potential use cases based on the case examples discussed, and 9 main categories of competencies and knowledge areas were defined. Figure 2 describes the categorization process and its application to each case.

XSL•FO
RenderX

Figure 2. Coding and categorization process. The diagram reflects how the codes and categories that emerged from the co-creation process evolved into the 4 use cases. GBV: gender-based violence.



Ethical Considerations

The project was approved by the Institutional Committee for Ethical Review of Projects at Universitat Pompeu Fabra under the approval number 202. The session was recorded with previous explicit and written participant informed consent. Opting for an online environment accommodated transportation and scheduling challenges. Video, alongside audio, was integral to capturing participants' body language, enriching the analysis and interpretation of their narratives [27]. During the focus group transcription, each participant was anonymized. Participation was voluntary, and no financial compensation or incentives were provided to the participants.

Phase 2: Workshops: Learning Outcomes and Virtual Reality Scenarios Design

A series of co-creation workshops was carried out with the aim of (1) translating the competencies and knowledge areas into learning outcomes, (2) identifying and designing 4 use cases to address the identified learning outcomes, and (3) developing the VR scenarios along with the PBL and EL sequences.

The working group responsible for analyzing and creating the scenarios comprised a multidisciplinary team of 2 researchers

XSL•FO
RenderX

with expertise in gender and health. These researchers contributed their specialized knowledge to understanding GBV and its impact on the health care environment. The group was joined by 2 specialists in health education and an educational innovation technician, whose contribution made it possible to contextualize the findings in relation to health education training and practice. Moreover, the working team included 2 technicians responsible for the production and development of the VR videos linked to the research. These technicians ensured the technical and didactic integration of the audiovisual material in the analysis of the results.

The first workshop served to translate the content of the 9 identified categories into 18 learning outcomes and to define 4 use cases, which would form the basis for the development of the educational resources.

Four additional co-creation sessions focused on designing the use cases, VR scenarios, and the first drafts of the PBL and EL sequences. The second workshop focused on distributing the learning outcomes across the 4 use cases and cross-checking that all were adequately covered. The key aspects of each use case were then defined. The remaining 3 sessions were dedicated to developing the VR scenarios for each use case and the corresponding methodological sequences. In particular, the first drafts of the scripts were produced, with special attention to the dialogues and actions that would best address the specific learning needs. A detailed description of the characters was also created at this stage. Additionally, an initial draft of the key elements for the PBL and EL sequences was developed, including relevant content and frameworks to support the teaching and learning process.

Phase 3: VR Scenarios' Recording and Educational Resources Development

To ensure alignment with the real-world situations that professionals face in their daily practice, the resulting scenarios were sent to the frontline professionals who had participated in the focus group. This step was considered a form of contextual validation to ensure appropriateness and relevance. In fact, 1 scenario was significantly revised to better reflect real-life conditions. Initially, the research team had proposed incorporating the stigma surrounding mental health and addictions in 1 use case. However, feedback from experts indicated that this approach might reduce the impact of the case, as instances of sexual exploitation are often closely linked to drug addiction, frequently facilitated by the survivor's partner. To maintain fidelity to real cases while ensuring accessibility for novice students, the script was accordingly revised to address this concern.

Finally, the scripts for recording the 360° scenarios were produced, and the didactic guides associated with each use case were simultaneously developed to facilitate their future implementation. These guides provide essential pedagogical support for anyone intending to implement the scenarios, offering structured guidance for educators, while both the guides and the accompanying videos are freely accessible, ensuring open access to the resources for broader educational use. Both the guides and the educational intervention rely on 2 didactic methodologies: a classic sequence of PBL sessions and single sessions of EL based on Morris' cycle [20]. These methodologies are designed to immerse students in real-world scenarios, fostering active engagement and critical thinking. The development process was further refined through an initial iteration of the materials using the classic PBL sequence, ensuring both clarity and pedagogical effectiveness of the resources. This initial iteration will also help identify areas for improvement and refine the use of the scenarios as well as the content of the didactic guides. The test included 64 students enrolled in a degree program in diagnostic imaging and nuclear medicine. A quasi-experimental pre-post design was used, using questionnaires that incorporated the Spanish-translated Perception Scale Partner Violence in Nursing Education (PSPV-NE) scale [28], which evaluates students' perceptions of GBV, including general training, survivor and aggressor identification, and professional roles and values in addressing the issue. The primary outcome was analyzed using the Student *t* test, with a significance threshold set at an α of .05.

Results

In each phase of the co-creation, certain results have been obtained, which have enabled the next phase to proceed.

Phase 1: Exploration and Identification of Learning Needs and Use Cases

As a result of the focus group analysis, 9 main categories were generated. In addition, the most common and critical situations of GBV 4 cases were identified: intimate partner violence during pregnancy, multiple types of violence against women in situations of drug addiction, psychological abuse in elderly women, and violence against the mother perpetrated through her children. The coding and categorization process is described in Figure 2.

In addition, there was consensus in the focus group on transversal aptitudes for addressing GBV. The main skills described that workers should have to deal in an appropriate manner with GBV situations were as follows: (1) a professional and careful interaction, empathic but not paternalist ("It is not just a conversation [...]. Words must be chosen carefully" [P3]); (2) establishing a climate of confidence and trust ("Generate trusting environments and relationships—so that they can explain, identify, and not to feel judged" [P1]); (3) preventing revictimization ("Women often arrive blocked, they want oblivion, they don't want to explain the same thing fifteen times" [P2]); and (4) social care and health care system awareness, knowledge about services, and its limitations ("It is important to confront the difficulty, accept the situation, take care of the victim and tell them that you will transfer them to a place where they can be looked after" [P5]. "It may be that there is a three-month wait for access to a resource, that it is far away [...], or that you do not have access because in your case it is 'just' psychological violence" [P4]).

Furthermore, the group underlined the importance of representing and taking into account other axes of inequality faced by the survivor and the professionals' willingness to set aside their own prejudices on the subject, in order to address the situation holistically.

XSL•FO
RenderX

> *Professionals have a lot of prejudices about these women, about their lives, about their potential to get out of the situation.* [P6]
>
> *There is structural violence and many culturally normalized behaviors [...]. There is not only intimate partner violence, other groups such as LGBTI community, sex workers or mental health issues must also be considered.* [P5]

Finally, a key point that emerged from the session was respect for the bioethical principle of autonomy and its dichotomy with respect to nonmaleficence. For professionals, this is a problematic issue, and they describe a tendency to solve it in the quickest possible way, sometimes reproducing patriarchal logic. However, it is not only key to respect women's decisions (protecting their safety) but also their time in the recovery process from violence.

> *It is important to avoid insisting on the denunciation as if everything ended with it [...] a very complicated process of reparation begins, which often starts with a psycho-emotional blockage.* [P5]
>
> *Health professionals have the tendency to tell others what to feel or do, but ultimately it is the individual who has to decide.* [P2]
>
> *Do not want to be women's saviors—they are the leading characters, and they are the ones who have to play an active role.* [P3]

Phase 2: Workshops: Learning Outcomes and Virtual Reality Scenarios Design

Based on the main categories that emerged in the focus group, 4 different VR scenarios have been created, as well as the learning outcomes related to each case. It was because of their frequency and didactic potential that these 4 key cases were selected.

Case 1: Doctor, These Pills Make Me Sleepy

The main topics related to this case were cultural defenselessness to GBV and effective and empathic communication (Multimedia Appendix 1). It shows psychological violence toward an elderly woman. Thereafter, specific learning outcomes were assigned according to the starting topics: (1) developing interview technique, active listening, and empathy; (2) applying different communicative and relational strategies based on the survivor's situation; (3) recognizing individual value systems and fostering a critical examination of these constructs; and (4) identifying GBV as a determinant of women's health.

Case 2: Pregnant Women's Spotting

In this case, principal topics related were clinical practice ethics, pregnancy as a vulnerability scenario, and HCP assessment and intervention (Multimedia Appendix 1). It shows both physical and sexual violence perpetrated by the intimate partner on a pregnant woman. Particularly in this case, 2 versions of the ending are available, depending on a student's decision, which will confirm or maintain the doubt that her situation is one of GBV. The specific related learning outcomes are (1) reflecting on ethical dilemmas arising when GBV is suspected and how to intervene as a professional; (2) developing an awareness and comprehensive understanding of the vulnerable circumstances faced by pregnant women, and acquiring the necessary skills to respond appropriately to their needs; (3) applying risk evaluation tools for women experiencing GBV; and (4) constructing a decision-making framework that incorporates complexity, consistently placing the needs of the woman at the forefront.

Case 3: Laia, Have You Returned to Drug Use?

This third case is associated with professionals' biases and drug addiction as a vulnerability scenario (Multimedia Appendix 1). It reveals sexual and institutional violence toward the survivor. The related learning outcomes are (1) recognizing indicators of GBV in complex contexts, including addiction scenarios, and implementing an appropriate intervention; (2) identifying stereotypes and beliefs that may impact HCP practices through the application of critical thinking and reflective practice; and (3) elucidating the concept of institutional violence and secondary revictimization, and discerning its manifestation within professional practices.

Case 4: Mati Has a Bellyache

Finally, the last case created is based on multidisciplinary collaboration and the inclusion of children as a vulnerability scenario (Multimedia Appendix 1). The last specific learning outcomes elaborated were (1) understanding the protocols for responding to situations of suspected GBV, particularly in cases involving minors; (2) identifying an instance of vicarious violence through the narrative and exploration of a child or their mother; and (3) recognizing the importance of documenting aspects involving GBV in the medical record and its advantages and disadvantages.

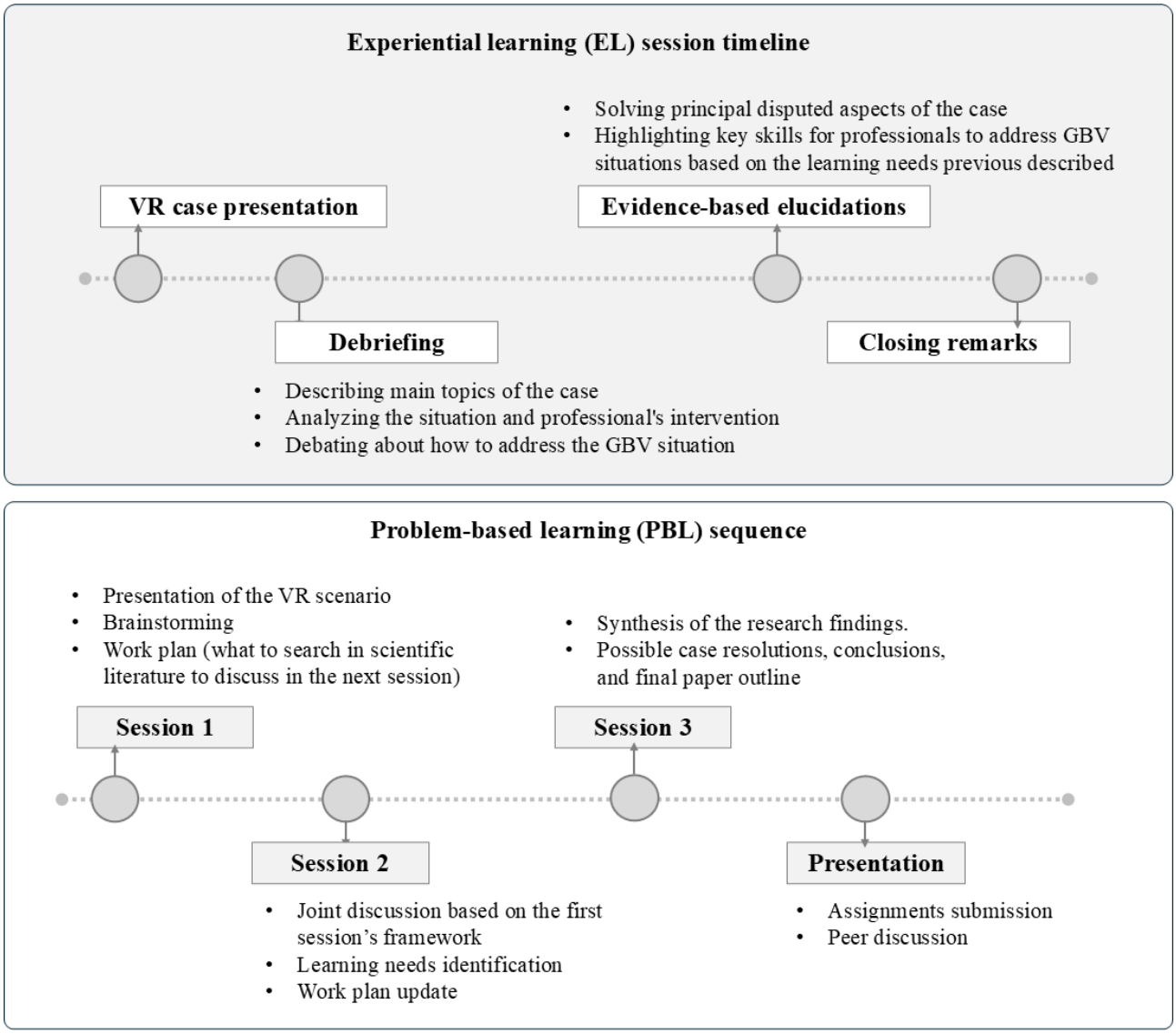
Phase 3: VR Scenarios' Recording and Educational Resources Development

The feedback received regarding the scripts following the expert consultation process was, in general, validating. The frontline professionals highlighted both the similarity of the cases to current clinical practice, as well as the importance of the learning outcomes proposed by the researchers. Finally, the scenarios outlined above were recorded using a 360° recording camera. Learning guides were also drafted for each case based on the learning outcomes obtained.

Two didactic methodologies were proposed for their application (Figure 3), depending on the objectives and logistical needs of the educational centers. For each of them, a didactic support guide was prepared containing the learning outcomes of each case, the scripts with the most noteworthy elements, a theoretical section concerning the essential concepts of the case, and finally, a list of useful resources available [29,30]. These guides will prove valuable for teachers who apply these interventions, not only in terms of content resources but also to unify criteria in the session's methodology and evaluation. Both the teaching guides and the 360° videos are available in open access format.

XSL•FO
RenderX

Figure 3. Proposed learning sequence for the implementation of virtual reality (VR) scenarios, using experiential learning teaching methodologies with a single 2-hour session and problem-based learning with four 2-hour sessions. GBV: gender-based violence.



The first methodology proposed was a PBL classical sequence. By integrating problem-solving methodology, students are encouraged to apply theoretical knowledge to practical situations, going deeper into the subject and skill development in addressing GBV within health care settings. The PBL sequence typically begins with the presentation of a case scenario for detailed analysis. In this study, VR was used as an alternative to traditional written cases. This technological innovation enhances student engagement by providing an immersive, first-hand experience of the situation, thereby creating a strong foundation for knowledge construction. The PBL method further reinforces content through multiple sessions, each lasting approximately 2 hours and conducted in small groups. These sessions facilitate in-depth case analysis and encourage students to reach the highest levels of Bloom's taxonomy, including the creation of novel products and the generation of alternative solutions [31]. The Bloom framework is particularly relevant in PBL contexts, as it informs the design of complex, real-world problems that require learners to use higher-order cognitive skills.

A single session of approximately 2 hours using EL was also proposed. The EL session is grounded in concrete reflection on the experience, providing a highly valuable didactic opportunity despite limited theoretical exploration. This approach allows students to engage deeply with the scenarios, fostering the specific learning outcomes through an extensive peer debriefing. This methodology is key to theorize and conceptualize professional activity based on a profound reflection of experience and to be able to analyze all the contextual elements of each scenario and its transferability to other contexts. This debriefing comprises a major part within the session, not only describing and analyzing what has been observed in the case but also culminating with a theory overview of relevant issues to address any doubts and concerns that have been raised in the debriefing.

Regarding the explorative iteration of the materials, the mean scores obtained on the PSPV-NE assessment were higher across all items following the intervention, reaching statistical significance (P<.05). Analysis of the 4 subscales similarly revealed significant improvements after the educational training, as did the participants' ability to correctly identify both the

XSL•FO
RenderX

survivor and the perpetrator. However, knowledge regarding the role of HCPs in addressing GBV, while showing a positive trend, did not reach statistical significance. These findings highlighted the need to reinforce the role of HCPs in addressing GBV within the educational materials, which was subsequently incorporated into the instructional guides.

Discussion

This study describes the design of VR-based educational materials developed through a co-creation methodology and discusses how this approach enhances the clinical relevance, realism, and educational alignment of GBV training in health sciences curricula. The collaborative involvement of HCPs, GBV experts, and specialists in educational innovation and immersive technologies resulted in the development of four immersive VR scenarios representing frequent and complex clinical situations: psychological violence within an elderly married couple, physical violence during pregnancy, GBV in a woman with substance use disorder, and vicarious violence. These scenarios were complemented by comprehensive teaching guides, enabling their direct integration into health sciences curricula and facilitating their use by university instructors.

The involvement of frontline HCPs with extensive experience in GBV was essential for identifying both the most pressing training needs of HCPs and the types of cases most commonly encountered in routine clinical practice. Their contributions informed the definition of 9 key categories that guided the selection of the 4 use cases and the identification of transversal competencies required for effective clinical responses to GBV. This methodological approach constitutes a significant strength of the study in terms of applicability and reliability. As noted by Van Weel-Baumgarten et al [32], co-creation processes enhance relevance and feasibility when experts are actively engaged throughout the design process.

The integration of GBV experts alongside academic and technical profiles further strengthened the educational alignment and implementation potential of the training program. This interdisciplinary collaboration is consistent with the recommendations of Surapaneni [33], who emphasized the importance of expert collaboration in the development of complex educational interventions. The use of focus groups facilitated in-depth and reflective dialogue [34], enabling the identification of critical professional skills, including empathic yet nonpaternalistic communication, the establishment of trust, prevention of revictimization, and awareness of health care and social system limitations. These competencies were explicitly embedded within the learning outcomes of each VR scenario, ensuring that the educational resources accurately reflect real-world clinical experiences.

In line with the work of Jiménez-Rodríguez et al [35], the use of videos simulating clinical scenarios may support a more effective alignment between theoretical knowledge and practical application, particularly when scenarios incorporate the ethical, emotional, and relational complexities identified by professionals themselves. While numerous educational initiatives addressing GBV have been described in the literature, substantial curricular limitations persist [5]. This study contributes to addressing these gaps by translating professional expertise into immersive VR scenarios that go beyond identification, incorporating ethical dilemmas, professional biases, and intersecting vulnerabilities such as pregnancy, addiction, older age, and vicarious violence.

Previous research has shown that educational interventions can improve knowledge, attitudes, skills, and behaviors related to GBV [36], with simulation-based and experiential learning approaches offering additional benefits [37]. The findings of Adánez-Martínez et al [38] further support the educational value of simulation and problem-solving videos, reinforcing the suitability of VR for representing complex and sensitive GBV cases. In this study, the pilot implementation indicated a positive educational potential, as reflected by increased mean PSPV-NE scores across all subscales and improvements in students' ability to identify both survivors and perpetrators of GBV.

These findings align with prior evidence suggesting that immersive VR experiences can promote behavioral change in educational contexts [39], particularly when combined with structured reflection and debriefing [40]. Although improvements in knowledge regarding the specific role of HCPs did not reach statistical significance, this result was used formatively to refine the instructional guides, placing greater emphasis on professional responsibility and system-level intervention. This iterative refinement responds to ongoing challenges in GBV education related to the use of robust evaluation strategies [5,41].

Beyond methodological innovation, this intervention also addresses content-related gaps in existing curricula. As noted by previous authors, medical education often prioritizes risk factors and the identification of GBV [42]. In contrast, this study integrates ethical considerations, institutional violence, and broader axes of inequality that shape women's health care experiences, reflecting the themes raised by expert participants. While the study makes a meaningful contribution to GBV education, further research using more robust and longitudinal evaluation designs is warranted.

From a practical standpoint, the open-access availability of the 360° VR videos and accompanying teaching guides supports dissemination and scalability across educational contexts. Overall, the educational resources developed represent a valuable contribution to the integration of GBV education into health sciences curricula. By grounding VR scenario design in professional expertise and combining immersive technology with active learning methodologies, this intervention offers a safe, experiential environment for addressing complex and sensitive clinical situations. The co-creation process presented may therefore serve as a transferable model for the development of educational interventions in other complex areas of health care education.

XSL•FO
RenderX

Funding

The authors declare that the research was conducted with financial support by Grants for Innovation Projects for the Improvement of Teaching Quality in the field of Genderbased Violence 2021 (INDOVIG 2021).

Authors' Contributions

Conceptualization: BV-A, MB, MC
Formal analysis: BV-A, MB, MC
Funding acquisition: MC
Investigation: BV-A, MB, MC
Methodology: BV-A, MB, MC
Project administration: MC
Resources: MB, MC
Supervision: MB, MC
Validation: BV-A, MB, MC
Writing – original draft and visualization: BV-A
Writing – review and editing: MB, MC

Conflicts of Interest

None declared.

Multimedia Appendix 1
Virtual reality scenarios and instructional guides for addressing gender-based violence in health sciences education.
[PDF File, 407 KB - mededu_v12i1e76098_app1.pdf]

References

1. Jatfors AK. Violence against women and women's economic empowerment. United Nations Economic and Social Commission for Asia and the Pacific. 2017 Feb 23. URL: https://www.unescap.org/sites/default/files/Session%203%20-%20Anna-Karin%20Jatfors.pdf [accessed 2026-03-02]
2. Global plan of action to strengthen the role of the health system within a national multisectoral response to address interpersonal violence, in particular against women and girls, and against children. : World Health Organization; 2016 URL: http://apps.who.int/iris/bitstream/handle/10665/252276/9789241511537-eng.pdf?sequence=1 [accessed 2026-03-02]
3. Global and regional estimates of violence against women: prevalence and health effects of intimate partner violence and nonpartner sexual violence. : World Health Organization; 2013 URL: https://iris.who.int/server/api/core/bitstreams/befdb859-99f5-4632-804d-04eb3fe8d940/content [accessed 2026-03-02]
4. Wessells MG, Kostelny K. The psychosocial impacts of intimate partner violence against women in LMIC contexts: toward a holistic approach. Int J Environ Res Public Health 2022 Nov 4;19(21):14488. [doi: 10.3390/ijerph192114488] [Medline: 36361364]
5. Ambikile JS, Leshabari S, Ohnishi M. Curricular limitations and recommendations for training health care providers to respond to intimate partner violence: an integrative literature review. Trauma Violence Abuse 2022 Oct;23(4):1262-1269. [doi: 10.1177/1524838021995951] [Medline: 33622184]
6. Hewitt LN. Intimate partner violence: the role of nurses in protection of patients. Crit Care Nurs Clin North Am 2015 Jun;27(2):271-275. [doi: 10.1016/j.cnc.2015.02.004] [Medline: 25981728]
7. Colombini M, Dockerty C, Mayhew SH. Barriers and facilitators to integrating health service responses to intimate partner violence in low- and middle-income countries: a comparative health systems and service analysis. Stud Fam Plann 2017 Jun;48(2):179-200. [doi: 10.1111/sifp.12021] [Medline: 28422291]
8. Sprague S, Swaminathan A, Slobogean GP, et al. A scoping review of intimate partner violence educational programs for health care professionals. Women Health 2018;58(10):1192-1206. [doi: 10.1080/03630242.2017.1388334] [Medline: 29252143]
9. Barnard M, Bynum LA, Bouldin A. Preparing student pharmacists to care for patients exposed to intimate partner violence. Curr Pharm Teach Learn 2021 Dec;13(12):1619-1626. [doi: 10.1016/j.cptl.2021.09.038] [Medline: 34895671]
10. Bermele C, Andresen PA, Urbanski S. Educating nurses to screen and intervene for intimate partner violence during pregnancy. Nurs Womens Health 2018 Feb;22(1):79-86. [doi: 10.1016/j.nwh.2017.12.006] [Medline: 29433702]
11. Gürkan Ö, Kömürcü N. The effect of a peer education program on combating violence against women: a randomized controlled study. Nurse Educ Today 2017 Oct;57:47-53. [doi: 10.1016/j.nedt.2017.07.003] [Medline: 28732209]
12. Gürkan Ö. The effect of a domestic violence course on nursing students' recognition of violence symptoms against women and their attitudes toward violence. J Psy Nurs 2020;11(1):57-63. [doi: 10.14744/phd.2019.82712]

XSL•FO
RenderX

13. Arora S, Rege S, Bhate-Deosthali P, et al. Knowledge, attitudes and practices of health care providers trained in responding to violence against women: a pre- and post-intervention study. BMC Public Health 2021 Nov 1;21(1):1973. [doi: 10.1186/s12889-021-12042-7] [Medline: 34724912]
14. Sprague S, EDUCATE Investigators. A qualitative evaluation of the implementation of an intimate partner violence education program in fracture clinics. J Fam Violence 2019;34(7):621-630. [doi: 10.1007/s10896-019-00052-4] [Medline: 31929680]
15. Manuel B, Valcke M, Keygnaert I, Roelens K. Improving medical students' communication competencies to deal with intimate partner violence using clinical simulations in Mozambique. BMC Med Educ 2021 Feb 23;21(1):126. [doi: 10.1186/s12909-021-02560-8] [Medline: 33622300]
16. Sammut D, Kuruppu J, Hegarty K, Bradbury-Jones C. Which violence against women educational strategies are effective for prequalifying health-care students?: A systematic review. Trauma Violence Abuse 2021 Apr;22(2):339-358. [doi: 10.1177/1524838019843198] [Medline: 31122182]
17. Cooper KM, Downing VR, Brownell SE. The influence of active learning practices on student anxiety in large-enrollment college science classrooms. Int J STEM Educ 2018;5(1):23. [doi: 10.1186/s40594-018-0123-6] [Medline: 30631713]
18. Ma CW, Cheng PS, Chan YS, Tipoe GL. Virtual reality: a technology to promote active learning of physiology for students across multiple disciplines. Adv Physiol Educ 2023 Sep 1;47(3):594-603. [doi: 10.1152/advan.00172.2022] [Medline: 37382503]
19. Barteit S, Lanfermann L, Bärnighausen T, Neuhann F, Beiersmann C. Augmented, mixed, and virtual reality-based head-mounted devices for medical education: systematic review. JMIR Serious Games 2021 Jul 8;9(3):e29080. [doi: 10.2196/29080] [Medline: 34255668]
20. Dolmans DHJM, Loyens SMM, Marcq H, Gijbels D. Deep and surface learning in problem-based learning: a review of the literature. Adv in Health Sci Educ 2016 Dec;21(5):1087-1112. [doi: 10.1007/s10459-015-9645-6]
21. Morris TH. Experiential learning—a systematic review and revision of Kolb's model. Interact Learn Environ 2020 Nov 16;28(8):1064-1077. [doi: 10.1080/10494820.2019.1570279]
22. Sharma S, Agada R, Ruffin J. Virtual reality classroom as an constructivist approach. 2013 Presented at: 2013 Proceedings of IEEE Southeastcon; Apr 4-7, 2013. [doi: 10.1109/SECON.2013.6567441]
23. Kolb AY, Kolb DA. Experiential learning theory: a dynamic, holistic approach to management learning, education and development. In: The SAGE Handbook of Management Learning, Education and Development: SAGE Publications; 2009:42-68. [doi: 10.4135/9780857021038.n3]
24. Kalra N, Hooker L, Reisenhofer S, Di Tanna GL, García-Moreno C. Training healthcare providers to respond to intimate partner violence against women. Cochrane Database Syst Rev 2021;5:CD012423. [doi: 10.1002/14651858.CD012423.pub2]
25. Zaher E, Keogh K, Ratnapalan S. Effect of domestic violence training: systematic review of randomized controlled trials. Can Fam Physician 2014 Jul;60(7):618-624. [Medline: 25022633]
26. Donkoh S, Mensah J. Application of triangulation in qualitative research. J Appl Biotechnol Bioeng 2023;10(1):6-9. [doi: 10.15406/jabb.2023.10.00319]
27. Santhosh L, Rojas JC, Lyons PG. Zooming into focus groups: strategies for qualitative research in the era of social distancing. ATS Sch 2021 Jun;2(2):176-184. [doi: 10.34197/ats-scholar.2020-0127PS] [Medline: 34409412]
28. Burjalés-Martí MD, Rigol-Cuadra MA, Anguiano-Carrasco C, et al. Scale for the study of nursing students' perception of intimate partner violence: adaptation and validation. Nurse Educ Today 2018 Jun;65(1):17-22. [doi: 10.1016/j.nedt.2018.02.025] [Medline: 29522962]
29. Valverde-Alirangues B, Carrió M, Benet M. Virtual reality resources to address gender-based violence in health science studies: a didactic guide for experiential learning. Repositori. 2023. URL: http://hdl.handle.net/10230/60832 [accessed 2026-03-03]
30. Valverde-Alirangues B, Carrió M, Benet M. Virtual reality resources to address gender-based violence in health science studies: a didactic guide for problem-based learning sessions. Repositori. 2023. URL: http://hdl.handle.net/10230/60834 [accessed 2026-03-03]
31. Chandio MT, Pandhiani SM, Iqbal S. Bloom's taxonomy: improving assessment and teaching-learning process. J Educ Educ Dev 2016;3(2):203. [doi: 10.22555/joeed.v3i2.1034]
32. van Weel-Baumgarten E, Benson J, Hoshi G, et al. Co-creation and collaboration: a promising approach towards successful implementation. Experience from an integrated communication and mental health skills training programme for Japanese general practice. Patient Educ Couns 2021 Oct;104(10):2386-2392. [doi: 10.1016/j.pec.2021.07.027] [Medline: 34330598]
33. Surapaneni KM. Key considerations associated with the implementation of virtual reality to enhance the learning experiences in medical education. Adv Physiol Educ 2023 Dec 1;47(4):746. [doi: 10.1152/advan.00135.2023] [Medline: 37703414]
34. Polit DF, Beck CT. Essentials of Nursing Research: Appraising Evidence for Nursing Practice, 8th edition: Wolters Kluwer Health; 2014. URL: https://www.ncbi.nlm.nih.gov/nlmcatalog/101586119 [accessed 2026-03-02]
35. Jiménez-Rodríguez D, Belmonte García MT, Santillán García A, Plaza Del Pino FJ, Ponce-Valencia A, Arrogante O. Nurse training in gender-based violence using simulated nursing video consultations during the COVID-19 pandemic: a qualitative study. Int J Environ Res Public Health 2020 Nov 21;17(22):8654. [doi: 10.3390/ijerph17228654] [Medline: 33233390]

XSL•FO
RenderX

36. Sawyer S, Coles J, Williams A, Williams B. A systematic review of intimate partner violence educational interventions delivered to allied health care practitioners. Med Educ 2016 Nov;50(11):1107-1121. [doi: 10.1111/medu.13108] [Medline: 27762036]
37. Wood L. Simulation as a training tool for intimate partner violence screenings. Clin Simul Nurs 2016 Jan;12(1):8-15. [doi: 10.1016/j.ecns.2015.11.001]
38. Adánez-Martínez MG, Pérez-Cánovas C, Gutiérrez-Muñoz I, et al. Impact of a training program on gender-based violence of medical students: a quasi-experimental simulation study. Clin Simul Nurs 2023 Nov;84:101458. [doi: 10.1016/j.ecns.2023.101458]
39. Fertleman C, Aubugeau-Williams P, Sher C, et al. A discussion of virtual reality as a new tool for training healthcare professionals. Front Public Health 2018;6:44. [doi: 10.3389/fpubh.2018.00044] [Medline: 29535997]
40. Jenney A, Koshan J, Ferreira C, et al. Developing virtual gaming simulations to promote interdisciplinary learning in addressing intimate partner and gender-based violence. J Soc Work Educ 2023 Dec 8;59(sup1):S76-S88. [doi: 10.1080/10437797.2023.2193597]
41. Crombie N, Hooker L, Reisenhofer S. Nurse and midwifery education and intimate partner violence: a scoping review. J Clin Nurs 2017 Aug;26:2100-2125. [doi: 10.1111/jocn.13376] [Medline: 27412048]
42. Ghaith S, Voleti SS, Ginsberg Z, Marks LA, Files JA, Kling JM. A scoping review of published intimate partner violence curricula for medical trainees. J Womens Health (Larchmt) 2022 Nov;31(11):1596-1613. [doi: 10.1089/jwh.2021.0345] [Medline: 35231186]

Abbreviations

EL: experiential learning
GBV: gender-based violence
HCP: health care professional
PBL: problem-based learning
PSPV-NE: Perception Scale Partner Violence in Nursing Education
VR: virtual reality

Edited by D Chartash; submitted 17.Apr.2025; peer-reviewed by AJ Santos, K Bilali; revised version received 19.Dec.2025; accepted 22.Dec.2025; published 18.Mar.2026.

Please cite as:
Valverde-Alirangues B, Benet M, Carrió M
Development of Virtual Reality Scenarios Addressing Gender-Based Violence in Health Sciences Education: Qualitative Approach
JMIR Med Educ 2026;12:e76098
URL: https://mededu.jmir.org/2026/1/e76098
doi:10.2196/76098

© Belén Valverde-Alirangues, Marta Benet, Mar Carrió. Originally published in JMIR Medical Education (https://mededu.jmir.org), 18.Mar.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Education, is properly cited. The complete bibliographic information, a link to the original publication on https://mededu.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Implementation of Emotional Connection Training in Pediatric Primary Care: Mixed Methods Study

Elizabeth Erickson[1], MD; J Blakely Amati[2], MD; Dani Dumitriu[3], MD, PhD; Anna Miller-Fitzwater[4], MPH, MD; Debra L Best[1], MD; Doreet Preiss[5], PhD; Elena Uejo[5], MSc; Adrian Brown[5], PhD; Jennifer E Crotty[6], MD; Leah Gillen[7], BA; Cherece Grier[4], MD; Kristina K Gustafson[8], MSCR, MD; Nicole Shearman[3], BA; Andréane Lavallée[3], PhD; Jennifer Warmingham[3], PhD; Elena Arduin[3], MPH; Nikki Shearman[7], PhD; Jackie Young[2], MS; Jessica Sperling[5], PhD

[1]Department of Pediatrics, Medical Center, Duke University, Durham, NC, United States

[2]Department of Pediatrics, Prisma Health, Greenville, SC, United States

[3]Department of Pediatrics, Center for Early Relational Health, Columbia University Irving Medical Center, 3 Columbus Circle, 11th floor, New York, NY, United States

[4]Department of Pediatrics, School of Medicine, Wake Forest University, Winston-Salem, NC, United States

[5]Social Science Research Institute, Duke University, Durham, NC, United States

[6]Department of Pediatrics, Brody School of Medicine, East Carolina University, Greenville, NC, United States

[7]Reach Out and Read, Boston, MA, United States

[8]Department of Pediatrics, Medical University of South Carolina, Charleston, SC, United States

Corresponding Author:
Dani Dumitriu, MD, PhD
Department of Pediatrics, Center for Early Relational Health, Columbia University Irving Medical Center, 3 Columbus Circle, 11th floor, New York, NY, United States

Abstract

Background: In 2021, the American Academy of Pediatrics released a policy statement spotlighting the health-promoting and stress-buffering effects of early relational health (ERH) and calling for universal ERH promotion in pediatric primary care. However, little educational content for the observation and promotion of ERH is available, highlighting the need for scalable ERH training modules.

Objective: This study aims to investigate the acceptability, feasibility, and impact of the "Lens of Emotional Connection," a self-paced, asynchronous, ~45-minute ERH training module codeveloped by Reach Out and Read and the Center for Early Relational Health at Columbia University. The module introduces practitioners to emotional connection, an observable component of ERH, through written and video didactic content and experiential rating of emotional connection in videos of parent-child dyads interacting face-to-face.

Methods: The evaluation was conducted by the Carolinas Collaborative. Pediatric providers across 8 clinical sites were invited to participate and responded to embedded pre-post surveys. Focus groups conducted with participants further examined the educational experience.

Results: Of 653 invited clinicians, 207 (31.7%) participated in the module. Individual survey responses were available for 44-75 participants, depending on the question. Of responders, 64 out of 69 (93%) reported the module was a good way to learn about emotional connection, and 63 out of 69 (91%) felt the module provided valuable knowledge. Overall, 60 out of 69 respondents (87%) reported satisfaction with the module length, and 36 out of 44 respondents (82%) reported they would recommend this training to other clinicians. Focus groups echoed these findings. Comparison of pre-post data showed the greatest changes were in familiarity with emotional connection (n=75, pre mean 54.20, SD 18.59; post mean 73.99, SD 14.73; Cohen d=1.14; P<.001) and confidence in observing the quality of the parent- or caregiver-infant relationship during well-child visits (n=75, pre mean 55.36, SD 18.49; post mean 74.20, SD 14.07; Cohen d=1.28; P<.001). Suggested areas for improvement included more thorough explanations of specific components of emotional connection identified in parent-child interaction videos, a desire for synchronous live training, and additional content addressing what to do if low emotional connection is identified.

Conclusions: In this evaluation of a training module designed to introduce pediatric practitioners to ERH and emotional connection, acceptability among participants was found to be high, with most responders reporting it as valuable and reporting they would recommend it. Statistically significant impact was noted in both perceptions of the importance of information about emotional connection and perceived knowledge acquisition. Feasibility of widespread implementation with voluntary participation, as here, was relatively low, with only a minority completing the module. Critically, Reach Out and Read's commitment to iterative

XSL•FO
RenderX

creation, validation, and eventual delivery of ERH training creates a scalable avenue for wide-scale implementation, given the organization's presence in >6500 clinics across the 50 states.

(JMIR Med Educ 2026;12:e81250) doi:10.2196/81250

KEYWORDS

medical education; community pediatrics; developmental-behavioral health; early relational health; resident education

Introduction

The intricate process of brain development occurs most rapidly in the first 3 years of life and depends on the complex interplay of a child's biological and emotional responses to both their physical environment and early relationships [1]. Early life exposure to prolonged, frequent, or severe stressors without the buffer of a supportive caregiver puts children at significant risk for poor health outcomes and developmental delays [2], including in social-emotional development, which the American Academy of Pediatrics (AAP) includes as a key domain for surveillance and screening in its Bright Futures Guidelines [3]. Strategies that promote positive parent-child interaction can mitigate these stressors and improve developmental outcomes [4,5]. Given the importance of these positive experiences, the AAP released a policy statement in 2021 on preventing childhood toxic stress that emphasized the importance of promoting early relational health (ERH) in pediatric primary care [5]. Contrasting this recommendation, little educational content for the observation and promotion of ERH is available, and consequently, most clinicians do not receive training in ERH in medical school or residency programs. This creates a clear gap in consistent assessment and anticipatory guidance in the domains of social-emotional development, which is exacerbated for children and families from minoritized populations [6,7].

There is therefore an urgent and critical need for the development and evaluation of scalable educational resources about the importance of ERH, particularly observable elements such as emotional connection, a key indicator of ERH associated with improved child outcomes [8,9]. Observing and engaging in this domain during clinical care can help providers gain insight into a young child's relationships, offer opportunities for light-touch interventions that strengthen those relationships (eg, encouraging talking, singing, playing, and reading), and help identify families who may need additional resources to support them and their young children [10]. Previous studies have shown that training in this area can improve residents' ability to identify different levels of emotional connection and improve their attitudes about the importance of emotional connection discussions during primary care visits [11].

Reach Out and Read as an avenue for wide-scale implementation of ERH promotion in pediatric primary care is specifically referred to in the AAP policy statement. Reach Out and Read is an evidence-based literacy intervention that builds on the unique relationship between caregivers and pediatricians to encourage interactions with children through early language and literacy to nurture strong relationships. Families participating in Reach Out and Read are more likely to read aloud to their children, leading to improved language scores compared to controls, an effect consistent across families of diverse backgrounds [12-15]. Reach Out and Read has consequently been identified as a primary component of pediatric primary care and endorsed as essential by the AAP [16], especially owing to the wealth of research demonstrating that shared reading is instrumental in promoting healthy social-emotional development [17-19]. Given the evidence supporting the literacy effects of Reach Out and Read and its impact on increasing the frequency of shared reading, layering in additional components that promote ERH into the Reach Out and Read curriculum is an especially promising avenue to support scalable implementation of ERH promotion in pediatric primary care.

Here, the Carolinas Collaborative, in partnership with Duke University, investigated the acceptability, feasibility, and impact of an early version of a self-paced, asynchronous, ~45-minute ERH training module codeveloped by Reach Out and Read and the Center for Early Relational Health at Columbia University. The module addresses the observation of emotional connection within the pediatric well-child visit as a core component of ERH and child development. The Carolinas Collaborative was established in 2016 as an AAP Community Pediatrics Training Initiative collaborative and is a coordinated collaboration between 8 academic institutions and health care systems serving as systems change infrastructure in North and South Carolina for innovation, training, research, and advocacy best practices to drive scalable child public health solutions. Pediatric clinicians from 8 clinical sites were invited to complete the module via email. Through pre-post surveys and focus groups, the research team investigated the impact of this new ERH module on perceived knowledge and attitudes about emotional connection and gained insight into additional training strategies to inform future design.

Methods

Training Module

Training content was based on a validated laboratory measure of the strength of emotional connection in parent-child dyads [8,20] adapted for use in the clinical setting through an iterative process that began with a learning collaborative between Reach Out and Read and the Center for Early Relational Health at Columbia University, and also included pediatric providers (spanning pediatricians, family physicians, and nurse practitioners) and 2 parent leaders. The training module was hosted on Reach Out and Read's Pathwright portal but was only available by special invitation. Training began with an experiential introduction to the "Lens of Emotional Connection" through four ~30-second videos of mother-infant dyads interacting with one another face-to-face, with varying levels of emotional connection across the videos. Participants were

XSL•FO
RenderX

asked to use their "gut feeling" to rate the strength of emotional connection using visual analog scales from 'not emotionally connected' to "emotionally connected" (range 1 - 100, but no numbers appearing for participants). Videos of these at-home interactions were recorded via Zoom as part of the longitudinal COMBO Initiative [21] and represent ethnically diverse mother-infant dyads at 4 - 6 months postpartum. Participants were then provided with written and video educational content of the principles of emotional connection, followed by the same 4 plus 3 new mother-infant videos for practice rating the strength of emotional connection using the same visual analog scales. Participants were informed that the scale acts as an educational tool and that they are not expected to use such a scale in their clinical practice. Finally, participants were shown 3 videos of parent-infant well-child visits with Reach Out and Read pediatricians trained in the observation and promotion of emotional connection for a real-life demonstration of using the new "lens" in a clinical setting. The training module was estimated to take 45 minutes to complete.

Mixed Methods Study Design

To evaluate the training module, this project used a sequential mixed methods design [22] with a pre-post survey process and follow-up focus groups. Surveys permitted a breadth of understanding across a larger number of participants; focus groups provided explanatory information to further understand and explain survey data as well as to identify relevant factors that were not captured in surveys. This mixed methods study is reported in accordance with both GRAMMS (Good Reporting of a Mixed Methods Study) and STROBE (Strengthening the Reporting of Observational Studies in Epidemiology) reporting guidelines to ensure proper reporting of both quantitative and qualitative aspects. The datasets generated and analyzed during this study (focus group interviews and survey data) are not publicly available, as the approved institutional review board (IRB) protocol did not allow for this. However, data can be made available by the corresponding author on reasonable request and pending appropriate IRB approval.

Survey

Pre-post surveys were developed across a series of recurring meetings held with an interdisciplinary team that included module developers, academic pediatricians, and social scientists with evaluation and measurement expertise. The pre- and postsurvey rationale was chosen to measure the effectiveness of the training in improving participants' perceived confidence and knowledge in ERH. Additional postsurvey items assessed module experience. Survey development and review by the diverse members of the planning team, which included experts in ERH, pediatricians with experiential engagement with pediatrics training, and training intervention developers, focused on content and face validity to ensure relevance, clarity, and alignment with module objectives. Surveys are available in Multimedia Appendix 1.

The presurvey questions were administered before participants engaged with module training content and assessed age, self-identified race and ethnicity, and gender, training level, and questions about previous ERH and emotional connection experience, attitude, and knowledge. The postsurvey questions were administered after participants completed all module training content and included parallel items addressing attitude and knowledge related to ERH and emotional connection (repeated from the presurvey), as well as items addressing module experience and perspective on the implementation of emotional connection-related practices in pediatric primary care. Pre- and postsurveys were programmed and hosted on Columbia University's Qualtrics and integrated directly into the Pathwright module experience. The combined length of the training module and pre-post surveys was ~45 minutes.

Enrollment

A total of 653 residents, advanced practice providers, and faculty at all 8 sites of the Carolinas Collaborative were recruited to participate in this study between June and October 2023 (Table 1). Inclusion criteria required participants to be either attending clinicians, resident clinicians, or faculty. Recruitment was the same across all sites, via email invitation with a link to the training. The module could be done on any computer and whenever they found or were allotted time. (Some clinical sites provided protected time for residents to complete the modules and surveys.) Participants were able to begin and finish at a later time as long as they used the same computer. It was made clear to participants that all aspects of this study were voluntary, and that participants could complete the training and choose to have their responses used in research or opt out of the use of their data for research purposes.

XSL•FO
RenderX

Table . Eligible participants invited to the study. Individuals invited to participate in the modules and study were largely residents and represented all 8 Carolinas Collaborative pediatrics programs.

Site	Faculty, n	Residents, n	Total, n
Prisma Health-Midlands	10	51	61
Prisma Health-Upstate	20	56	76
MUSC[a]	16	73	89
UNC[b]	19	70	89
Atrium Wake Forest Baptist Health	12	62	74
Atrium Health	15	38	53
ECU[c]	18	56	74
Duke	39	98	137
Total	149	504	653

[a]MUSC: Medical University of South Carolina.

[b]UNC: University of North Carolina.

[c]East Carolina University.

Data and Statistical Analysis

Survey data were not available for all participants who engaged with the module due to a combination of skipped survey items, some participants stopping before the postsurvey, and technical errors in the integration of data across Qualtrics and Pathwright. Sample sizes for available data for each survey item ranged between 44 and 75. Available data were analyzed using SAS software (version 9.4; SAS Institute, Inc.; copyright 2002 - 2012) and used descriptive statistics and paired 2-tailed *t* tests to address change for individuals with both pre- and postdata. Additional ordered logit and ordinary least squares regression models examined the potential effect of varied participant demographic characteristics and prior related experience; however, minimal effect of background characteristics was detected, and these results are thus not reported here.

Focus Groups

Individual Carolinas Collaborative sites coordinated with the social science research team at Duke University to schedule focus group opportunities with participants who completed the training between August and November 2023. Using a sequential mixed methods design [22], focus groups were constructed to provide additional context and explanation for survey-based quantitative results, in addition to providing added nuance more generally on experience with and effect of module engagement. Participants were recruited via email sent to the same distribution lists as during initial recruitment, and an incentive of a US $10 gift card was provided for participation in focus groups. As a result, 7 semistructured focus groups were conducted with 33 total participants across 4 Carolinas Collaborative sites. Focus groups were held virtually, lasted approximately 1 hour, and assessed initial knowledge and perspective on emotional connection, experience with the module, use, value of the module for their knowledge and practice, and opportunities for improvement; see Multimedia Appendix 2 for the focus group guide. Recruitment concluded after the interested participants identified by the end of the planned data collection period had been engaged. While the team had not definitively reached saturation across all possible points, results were revealing clear common themes regarding priority areas of experience and outcomes. Decisions against added waves of recruitment were based on project timeline considerations, likelihood of more limited participation with added recruitment efforts, and value in the existing scope data that had been collected (eg, clear common themes identifiable based on data already collected). Focus groups were recorded and transcribed. Data were analyzed using a rapid analysis framework [23], identifying key themes in responses to questions addressed. This used templated matrices structured based on primary areas of the focus group instrument (eg, aspects of module experience, distinct outcomes; see Multimedia Appendix 3 for matrix structure), review of transcripts and documentation of relevant content per item in each focus group, and cross-review among analysts (members of the social science team) to identify themes per topic. Qualitative and quantitative results were considered in tandem, with qualitative results providing an added explanatory lens on quantitative findings.

Ethical Considerations

The survey portion of this study was approved by the AAP's IRB (#22 ER 01), which served as the main approval body; the approval notice was shared with and reviewed by each constituent IRB within the Carolinas Collaborative (Atrium Health Levine Children's Hospital, Atrium Health Wake Forest Baptist, Duke University, East Carolina University, Medical University of South Carolina, Prisma Health Midlands, Prisma Health Upstate, and the University of North Carolina), some of which did not require further review and some which considered the study exempt following site-specific application. The Columbia University IRB also separately approved the study (#AAAU5749). Participant consent was waived by both the AAP and Columbia University IRBs because the data collected were anonymous and could not be linked to any identifying information. In lieu of consent, all participants were provided with an informational sheet (electronically before the survey) that provided them with all the information about the study and

XSL•FO
RenderX

offered them the opportunity to opt in or out of research participation. Participants could complete the training without participating in the research study. Participants did not receive compensation for their time completing the training (though, as noted above, a US $10 gift card was offered to the subset of participants who took part in the focus groups). The videos used in the training itself were collected in a previous study with participants who gave consent for their videos to be used for research and training purposes. The focus group portion of this study was approved by the Duke University IRB (#2023‐0300).

Results

Participants

Of those invited, 207 (31.7%) clinicians participated in the module, and of these participants, 117 respondents completed at least 1 survey (56.5% response rate) with representation across 8 Carolinas Collaborative sites (Table 2). Participants were mostly under 30 years old (n=71, 60.7%), self-identified as women (n=92, 78.6%), and self-identified as White or Caucasian race and ethnicity (n=92, 78.6%). Most participants were residents (n=38, 32.5% post graduate year (PGY)-1; n=33, 28.2% PGY-2; n=23, 19.7% PGY-3/4). While response rates could not be quantified among faculty versus residents due to the inherent lack of demographic data for participants who did not complete at least 1 survey, participation rates relative to invitations were relatively similar for the 2 groups among the 115 participants who reported on their clinical role: n=19 of 149 (12.8%) for faculty and n=94 of 504 (18.7%) for residents. While the majority (n=108, 93.2%) reported having completed the Reach Out and Read Core Module, only 26.4% (n=31) reported having moderate or significant prior experience with emotional connection.

XSL•FO
RenderX

Table . Study participant characteristics (n=117).

Characteristics[a]	Participants, n (%)
Age (years)	
Under 30 years old	71 (60.68)
30 years old or older	45 (38.46)
Missing response	1 (0.85)
Gender	
Man	22 (18.80)
Woman	92 (78.63)
Nonbinary	1 (0.85)
Prefer not to answer	2 (1.71)
Missing response	0 (0)
Race and ethnicity	
Asian or Pacific Islander	5 (4.27)
Black or African American	6 (5.13)
Latino or Latina	4 (3.42)
Multiracial	9 (7.69)
White or Caucasian	92 (78.63)
Missing response	1 (0.85)
Role	
AAP[b] (eg, NP[c] and PA[d])	2 (1.71)
PGY[e]-1	38 (32.48)
PGY-2	33 (28.21)
PGY-3/4	23 (19.66)
Physician - private panel	1 (0.85)
Physician - resident preceptor and private panel	11 (9.40)
Physician - resident preceptor only	7 (5.98)
Other	0 (0)
Missing response	2 (1.71)
Carolinas Collaborative locations	
Atrium Health - Levine	5 (4.27)
Atrium Health - Wake Forest University	23 (19.66)
Duke University	18 (15.38)
East Carolina University	12 (10.26)
Medical University of South Carolina	16 (13.68)
Prisma Health-Midlands	27 (23.08)
Prisma Health-Upstate	8 (6.84)
UNC[f] Chapel Hill	6 (5.13)
Missing	2 (1.71)
Amount of prior experience with emotional connection	
I have no experience with this	21 (17.95)
I have a little experience with this	65 (55.56)
I have a moderate amount of experience with this	26 (22.22)

XSL•FO
RenderX

Characteristics[a]	Participants, n (%)
I have a lot of experience with this	5 (4.27)
Missing response	0 (0)
Prior experience type (n=136[g])	
Involved in ROR[h] training development	2 (1.71)
Prior experience via Reach Out and Read Core module	108 (92.31)
Prior experience via Keystones of Development curriculum	19 (16.24)
Missing response	7 (5.98)

[a]All demographic and personal characteristics were self-identified by respondents.

[b]AAP: American Academy of Pediatrics.

[c]NP: nurse practitioner.

[d]PA: physician assistant.

[e]PGY: postgraduate year.

[f]UNC: University of North Carolina.

[g]Respondents were instructed to select "all that apply" for this question, resulting in 136 responses to this question.

[h]ROR: Reach Out and Read.

Module Experience

Survey data indicated that 64 out of 69 respondents (93%) reported that the module was a good way to learn about emotional connection (responses ranging between "somewhat agree" and "strongly agree"; Table 3). More than half of these respondents (n=40, 58%) reported they either "agree" or "strongly agree." Similarly, 63 out of 69 respondents (91%) reported that the training provided valuable knowledge (responses ranging from "somewhat agree" to "strongly agree"; Table 3), with 39 (57%) reporting agreement to some degree and 24 (35%) reporting strong agreement. In addition, 60 out of 69 respondents (87%) reported satisfaction with the module length. Focus group data echoed a positive experience with the module content and format, with focus group data indicating that the videos embedded within the modules provided valuable examples of what to look for when assessing emotional connection:

> *The content was easy to understand and very well laid out. Yeah, I think I particularly liked how interactive the videos were and the prompts after.*
>
> *…the online format was good because you got to see videos of actual children. [If] we just did like a role-playing thing or someone presented it then you might miss out on that aspect.*

A plurality of respondents reported that the training provided valuable knowledge, was a good way to learn about emotional connection, and presented strategies that could be implemented.

Suggested areas for improvement, reflected in both an open-ended survey question and in focus group data, frequently addressed aspects of content structure within the module. This included a desire for the incorporation of written summaries and/or reviews of the video content by experts either within or following module completion and clearer explanations of specific behaviors and examples of emotional connection within each video. Focus group data also indicated technical challenges with the log-in process, some challenges with the repetitive nature of the module videos, response fatigue, and a desire for added diversity in family structure within the videos (ie, the inclusion of more dyads beyond mother-baby):

> *I do remember it was initially kind of hard to figure out how to log in.*
>
> *When the same video came up I'm like okay, I saw this.*
>
> *I feel like maybe it would have been better to have the different types of caregivers present with them there, like instead of just mom, since that's like more what we see in real life, [the] different spectrum of caregivers.*

Focus group data additionally explored the experience of how the module was presented to participants and when respondents were expected to complete the modules. Some respondents indicated that their programs set aside a specific time for module completion, while others reported completing it independently, with benefits for each format:

> *Doing it with the dedicated time..., I was more focused because that's what everybody was doing.*
>
> *You can do it at like your own pace, like kind of a game, review the videos maybe a second time if you want to. Everyone has their own phase of like completing the things.*

XSL•FO
RenderX

Table . Value and acceptability of the training module.

Value and acceptability items	Respondents, n (%)
The training gave valuable knowledge (n=69)	
1 (strongly disagree)	1 (1.45)
2	0 (0)
3	1 (1.45)
4 (neither agree nor disagree)	4 (5.80)
5	21 (30.43)
6	18 (26.09)
7 (strongly agree)	24 (34.78)
The module was a good way to learn about emotional connection (n=69)	
1 (strongly disagree)	1 (1.45)
2	1 (1.45)
3	1 (1.45)
4 (neither agree nor disagree)	2 (2.90)
5	24 (34.78)
6	14 (20.29)
7 (strongly agree)	26 (37.68)
The strategies presented through this training could be implemented in respondents' practice (n=69)	
1 (strongly disagree)	1 (1.45)
2	0 (0)
3	0 (0)
4 (neither agree nor disagree)	2 (2.90)
5	20 (28.99)
6	21 (30.43)
7 (strongly agree)	25 (36.23)
Length of the training (n=69)	
Too short	0 (0)
Just about right	60 (86.96)
Too long	9 (13.04)
Would recommend training (n=44[a])	
No	2 (4.55)
Unsure	6 (13.64)
Yes	36 (81.82)
Recommendations on Incorporating into residency training (n=44[a])	
Don't know	0 (0)
Would not recommend in residency training	0 (0)
Recommend making available but not promoting	2 (4.55)
Recommend (not require)	24 (54.55)
Require for all residents	17 (38.64)
Other	1 (2.27)

[a]The smaller sample size represents attrition between survey initiation and survey completion.

XSL•FO
RenderX

However, challenges were identified in both approaches: set-aside time may be rushed or pulled from busy schedules, but finding one's own time could also be challenging. Distractions were experienced during module completion in both cases.

Regarding the module's future use, 36 out of 44 respondents (82%) reported they would recommend this training to other clinicians. In addition, most respondents suggested this module be required (n=17 out of 44, 39%) or recommended (n=24 out of 44, 55%) as part of residency training. An additional 2 out of 44 respondents (5%) indicated it should be available but not promoted, and 1 respondent who selected an "other" response indicated this should not exist as a stand-alone activity but rather should be "incorporated into a larger/longer learning session where the nuances of such strategies can be explored through discussion"; this was echoed in focus group data, which indicated a desire for synchronous or live training or an opportunity for more interpersonal engagement through the training:

> *It would have been cool if there was like an adaptive strategy to make it more of like a group-based activity so those of us who were in the same room could have done it more, in an engaged manner.*

Participant Effect and Implications for Practice

Additional survey data comparing responses on assessed outcomes before and after module completion (Table 4) indicated that the greatest changes were in familiarity with the concept of emotional connection (n=75, pre mean 54.20, SD 18.59; post mean 73.99, SD 14.73; Cohen *d*=1.14; *P*<.001) and confidence in observing the quality of the parent- or caregiver-infant relationship during well-child visits (n=75, pre mean 55.36, SD 18.49; post mean 74.20, SD 14.07; Cohen *d*=1.28; *P*<.001). In qualitative data from focus groups, respondents reported value in varied respects, including having a framework and model that provides guidance for the observation of emotional connection and that demonstrates the varied ways emotional connection can present and be observed:

> *It was useful because it gave like a practical framework on how to work with patients and also make a relationship with them in the office, especially with you maybe just first meeting these people and their baby in a more technical way.*

XSL•FO
RenderX

Table . Pre-post measures of module intended outcomes.

Outcome measures	Presurvey, mean (SD)	Postsurvey, mean (SD)	*t* test (*df*)	*P* value[a]	Effect size (Cohen *d*)[b]	Respondents, n[c]
Estimated minimum time to observe the quality of the relationship between a parent/caregiver and infant during a well-child visit (range 1 - 120 min)	13.14 (18.42)	8.11 (14.98)	1.98 (70)	.05	0.23	71
Familiarity, confidence, and beliefs about early relational health in well-child visits (scale=1 - 100)						
Familiarity with the concept of emotional connection	54.20 (18.59)	73.99 (14.73)	−9.9 (74)	<.001	1.14	75
Confidence in observing the quality of the parent- or caregiver-infant relationship during well-child visits	55.36 (18.49)	74.20 (14.07)	−11.09 (74)	<.001	1.28	75
Belief that understanding early relational health is essential to work as a clinician	80.41 (15.72)	86.00 (14.20)	−3.92 (64)	<.001	0.49	65
Belief that observing emotional connection in well-child visits will improve practice as a clinician	80.91 (15.73)	84.15 (14.47)	−2.23 (45)	.03	0.33	46
Belief that observing emotional connection in well-child visits will improve outcomes for patients and their families	78.34 (16.01)	82.68 (16.02)	−2.39 (46)	.02	0.35	47

[a]All *P* values are based on 2-tailed tests.

[b]Effect size (Cohen *d*): no effect=0.0 - 0.19, small=0.2 - 0.49, medium=0.5 - 0.79, large=0.8 and above; calculated as $d=t/\sqrt{n}$

[c]Analyses are subset to only include respondents with both pre and post responses.

> *I really liked having, you know, all the different examples of parent interaction…the model of what is good doesn't necessarily look exactly the same for everyone.*

Significant pre-to-post change was evidenced across all survey measures related to familiarity, confidence, and beliefs about ERH and emotional connection in well-child visits, and the expected estimated minimum time to observe the quality of the relationship between a parent-caregiver and infant during a well-child visit decreased.

Beliefs about ERH also exhibited a statistically significant change between the pre- and postsurvey. Specifically, participants exhibited a change in their beliefs that understanding ERH is essential to their work as a clinician (n=65, pre mean 80.41, SD 15.72; post mean 86, SD 14.20; Cohen *d*=0.49; *P*<.001), observing emotional connection in well-child visits will improve their practice as a clinician (n=46, pre mean 80.91, SD 15.73; post mean 84.15, SD 14.47; Cohen *d*=0.33; *P*=.03), and observing emotional connection in well-child visits will improve outcomes for patients and their families (n=47, pre

XSL•FO
RenderX

mean 78.34, SD 16.01; post mean 82.68, SD 16.02; Cohen *d*=0.35; *P*=.02). As one focus group respondent elaborated:

> *It makes me pay attention more even like with these little, very little babies that the emotional connection between them and their caregivers really does have such a huge impact.*

Of note, the baseline (pre) data indicated considerably greater agreement (mean ranges 78.34 - 80.91) with these assessed beliefs about emotional connection compared to assessed confidence in or familiarity with emotional connection (mean ranges 54.20 - 55.36) prior to module engagement. This suggests the training is filling a learning gap that the respondents desire to gain expertise in.

Barriers to Integrating Emotional Connection Into Practice

In survey data, 66 out of 69 respondents (95%) reported that the strategies presented could be implemented in clinical practice ("somewhat agree" to "strongly agree" that they could be implemented) (Table 3). When asked to consider primary challenges to implementing the tools taught in the module in practice, the time required to assess emotional connection during a clinical encounter was the most noted challenge (30 out of 188 survey responses [25.6%] on a "select up to 3 response" item) (Table 5). As one focus group participant explained:

> *Time is always a barrier. Time is always against you in clinic, you know, trying to get to the next patient, the next patient, the next patient…I'd love to spend as much time as possible like observing and talking about it, but a lot of times it's like I've got to tell them about their vaccines. I've got to talk to them about like lab testing, you know.*

Table . Primary barriers to integrating emotional connection into practice (N=188[a]).

Barriers	Respondents, n (%)
Clinic support or capacity for supporting families	9 (7.69)
Cultural differences between me and my patients and families	14 (11.97)
Families may not be open to talking about this	13 (11.11)
Knowing what to do if I identify a problem	19 (16.24)
Knowledge of how to do this	5 (4.27)
Language differences between me and my patients and families	15 (12.82)
The focus and energy required to do this	8 (6.84)
Time required to do this	30 (25.64)
Other	2 (1.71)
Did not respond to question	73 (62.39)

[a]Respondents were instructed to select up to 3 responses for this question; therefore, the total N of responses for this question is greater than the number of survey respondents.

Respondents most often reported implementation challenges associated with the time required to assess emotional connection, knowing what to do if a problem was identified, and language and cultural challenges.

In addition, 19 out of 188 survey responses (16.2%) pointed to potential challenges with determining follow-up action steps ('Knowing what to do if I identify a problem'), with focus group data echoing this:

> *I know how to realize good and bad. I know what to do with the good ones, but like what exactly should I do if I'm like 'you're not connecting to your child'?*

Language and cultural differences between respondents and their patients and families were also cited as potential challenges (15 out of 188 survey responses [12.8%] and 14 out of 188 survey responses [12%], respectively), in addition to family willingness to engage in relevant discussion (13 out of 188 survey responses [11.1%]) (Table 5). Focus group data additionally raised the concern that well-child visits may not provide a good example of family dynamics at home:

> *Taking off the clothes, and it's cold, and there are strangers…it feels a little bit harsh to judge what someone's like home interaction with their baby is like based on me walking in for five minutes.*

> *We see a lot of families that speak other languages and a lot of the times maybe the parents focusing on like the interpreter on the computer so we're not always getting a full kind of natural reaction between the parent and the kid…also if we have people coming in with maybe a grandparent or some other guardian who's not their parent or not even the person who they live with primarily, we're not always getting a feel for what the connection is with the people they live with.*

Discussion

Principal Findings

ERH and emotional connection have been identified as foundational components of healthy child development, and yet

XSL•FO
RenderX

there is a lack of scalable training resources to incorporate the observation of ERH into pediatric primary care practice [4,24-28]. Efforts to date have primarily focused on diagnosing and treating relational problems in children with relatively severe disturbances in behavior [29]. In addition to being narrowly targeted and costly, these efforts do not lend themselves to universal primary prevention [30]. Indeed, in a recent study exploring residency program directors' views on the adequacy of current ERH training, a clear gap in this area of medical education was identified [31].

Here, we tested the feasibility of a brief self-paced, asynchronous, interactive ERH training module in teaching pediatricians the observation of emotional connection, an observable component of ERH. We found that the large majority of respondents reported that the information provided in this module gave them valuable knowledge on emotional connection, and that this module was a good way to learn about emotional connection. Despite social-emotional development and healthy relationships being domains addressed in the AAP Bright Futures guidelines and policy statements [5], these results clearly show that this is an area of educational growth needed for clinicians who were initially unfamiliar with these concepts. In addition, specific results provide key areas to target. For instance, pretraining scores for knowledge and confidence (familiarity with emotional connection and confidence in observing parent and caregiver–infant relationships during well-child visits) were lower than scores for beliefs about emotional connection (eg, its importance to clinical work, practice improvement, and patient outcomes). This may explain the larger effect sizes for knowledge and confidence, and it suggests the training addresses a key gap: residents value ERH and emotional connection but have less knowledge and confidence in observing them. This is also consistent with prior literature indicating knowledge and confidence to be more responsive to short trainings than beliefs and attitudes [32,33].

In a field as complex as medicine, evolving medical information necessitates a feasible and acceptable method for efficiently disseminating new knowledge. In recent years, asynchronous modules have been thought of as an efficient way to reach learners in different locations [34,35]; additionally, with time constraints on faculty and residents, modules developed by topic experts can be a valuable way to add to clinician education without coordinating timing for in-person learning. Access to modules can decrease the burden on individual programs or faculty to find ways to accomplish educational and curricular goals in both training and continuing medical education. Modules and other e-learning strategies have long been considered effective in accomplishing this goal [36,37] and were widely adopted as a key component of medical and graduate medical education during the COVID-19 pandemic [38,39]. In the domain of ERH, there is a clear benefit to these modules, which enable multiple learners across multiple programs to be trained in a similar manner, allowing for shared knowledge and a common language to discuss social-emotional health and development.

However, there is significant variability in the content and quality of medical education training modules. Intentional development of modules through interdisciplinary collaborations like the one presented here can potentially increase access to high-quality educational opportunities, especially for new or emerging topics. The results of this pilot training suggest that a~45-minute training module is a feasible and acceptable way to provide education about ERH and specifically emotional connection, and that providers found the training valuable to improve their knowledge base.

In addition to the knowledge that these modules provide and an acceptable means of delivering novel educational content, it is also important to look at the challenges participants face. Notably, we discovered challenges with identifying time to complete modules, and the benefits and drawbacks of asynchronous delivery are critical to consider. Additionally, concerns noted by participants about how to incorporate learned skills into already full well-child visits are pertinent to many topics as more components are introduced into the well-child visit. Given these challenges, it will be essential to develop trainings that address how to easily integrate the content into the workflow and how to make changes or adjustments to clinical practice. These benefits and challenges are generalizable to a wide range of topics and disciplines that are interested in expanding educational opportunities via asynchronous modules.

Limitations and Future Directions

This study has several limitations, many of which can direct us to future opportunities for investigation. The first is that there may have been recruitment bias. Participants who engaged with the voluntary training module and study may have been interested in emotional connection or closely related topics that inspired them to participate. Response bias may also be a factor; specifically, social desirability in reporting knowledge or positive attitudes toward the subject or for residents who were asked to complete a module by their preceptors, regardless of the voluntary nature of the study [40]. Qualitative data provided some nuance on this possible bias. Although focus group participants provided an overall positive response to the module, they also displayed freedom to critique the training in this new feedback format, indicating that, indeed, the forced-choice survey data might be slightly inflated. Additionally, while there is value in the fact that these data represent 8 unique sites across the Carolinas Collaborative, they are all geographically located in the same area of the country, which limits generalizability. There was also a lack of diversity within the demographics of the participating clinicians. Additionally, adult learning styles were not considered for participants, and nuances in learning styles may make this training method more or less effective for individuals. Finally, this learning modality was not compared to a more traditional in-person experience, which was suggested by some participants as important and desired. We also do not have information from participants who started but failed to complete the modules. While we obtained some information from focus groups to understand improvements that could be made to the module, it would be important to better understand why some participants did not complete the training. Future studies will also need to evaluate if participation rates and satisfaction scores differ by sites that provide protected time for completion of the module versus those that do not.

XSL•FO
RenderX

Second, the study had a relatively low response rate. The timing of the study might have been a limiting factor in our successful recruitment of participants. Due to the timing of funding availability, modules were released at the beginning of June, and recruitment continued through October. Because the study primarily targeted residents and teaching faculty, timing may have contributed to the low response rate, as this is generally not an ideal time for change or new initiatives in academic centers. The voluntary nature of the training also likely contributed to the low response rate. This points to the necessity of embedding future iterations of such modules into already existing training infrastructure, such as the Reach Out and Read Core Curriculum, making it a mandatory component for certification.

In addition to these limitations, there are several areas that were not addressed by this study that call for future investigation. A concern mentioned by participants in both surveys and focus groups was that while there was thorough education about the importance of and strategies for identifying emotional connection, the modules did not address what to do when a lack of emotional connection was noted or an additional problem with social-emotional development was identified. Along this same theme, more investigation is needed to effectively address cultural and linguistic differences in parenting that could be perceived as "problems" without an appropriate appreciation for different cultural lenses to assess connection. Another future domain of study is how this education impacts clinicians' daily practice. This could be accomplished by direct observed exams, which are the gold-standard in resident education to assess trainees' competencies in various aspects of care [41,42]. Finally, we did not address the efficacy of the module, as defined in the ability to change how observers rate the video interactions compared to trained researchers.

Using the information learned here, we have created an updated iteration of this training module that is currently being piloted at Columbia University. This new version expands the training with a higher number of parent-child videos across a larger age range (0 - 3 years), incorporates real-time feedback of participants' ratings of emotional connection, and gives additional opportunities to score videos when the original score was out of range. The new module also includes several suggested actionable next steps if emotional connection is observed to be low, though participants are cautioned that light-touch intervention development targeting emotional connection is a current area under investigation.

Conclusions

ERH and emotional connection are foundational to healthy child development, and education about these topics delivered through asynchronous digital modules is highly valued by participants. While completion rates of voluntary training are generally low, including in this study, Reach Out and Read's commitment to creation, validation, and eventual delivery of ERH training creates a scalable avenue for wide-scale implementation, given the organization's presence in >6500 clinics across the 50 states. The data presented here inform next steps in the iterative design of medical education on ERH. Further research is needed to more fully understand how this training impacts clinician behaviors during clinical encounters, how families receiving care perceive these changes, the impact on child and family outcomes, and potential disparities in the perception of different cultural norms in parenting as demonstrative of strong or weak emotional connection, to avoid over- or underidentification of deficits in this area.

Acknowledgments

The authors attest that there was no use of generative artificial intelligence in the generation of text, figures, or other informational content of this manuscript.

Funding

This study was funded by Einhorn Collaborative.

Authors' Contributions

Conceptualization: EE, JBA, DD, Nikki Shearman, JW, AL
Data curation: DD, Nicole Shearman, EA, JW, AL
Formal analysis: JS, DP, EJ, AB
Funding acquisition: DD, Nikki Shearman
Investigation and data collection: EA, DD, Nicole Shearman, EA, JW, AL, DP, EJ, AB
Methodology: JS, DP
Project administration: Nikki Shearman, DD, EE
Resources: DD, EE
Software: DD, JS
Validation: JS, DP
Visualization: JS, DP
Writing – original draft: EE, JBA, JS, DD
Writing – revision & editing: EE, BA, DD, AM, DB, DP, EJ, AB, JC, LG, CG, KG, Nicole Shearman, AL, JW, EA, Nikki Shearman, JY, JS

XSL•FO
RenderX

Conflicts of Interest

EE reports she is a consultant for DynaMed Plus. The remaining authors declare no conflicts of interest relevant to this study.

Multimedia Appendix 1
Pre- and postsurveys.
[DOCX File, 36 KB - mededu_v12i1e81250_app1.docx]

Multimedia Appendix 2
Focus group guide.
[DOCX File, 17 KB - mededu_v12i1e81250_app2.docx]

Multimedia Appendix 3
Matrix structure for qualitative rapid analysis.
[DOCX File, 16 KB - mededu_v12i1e81250_app3.docx]

References

1. National Research Council (US) and Institute of Medicine (US) committee on integrating the science of early childhood development. In: Shonkoff JP, Phillips DA, editors. From Neurons to Neighborhoods: The Science of Early Childhood Development: National Academies Press; 2000.
2. Shonkoff JP, Garner AS, Committee on Psychosocial Aspects of Child and Family Health, Committee on Early Childhood, Adoption, and Dependent Care, Section on Developmental and Behavioral Pediatrics. The lifelong effects of early childhood adversity and toxic stress. Pediatrics 2012 Jan;129(1):e232-e246. [doi: 10.1542/peds.2011-2663] [Medline: 22201156]
3. Hagan JF. Bright Futures: Guidelines for Health Supervision of Infants, Children, and Adolescents, 4th edition: American Academy of Pediatrics; 2017.
4. Garner AS, Saul RA. Thinking Developmentally: Nurturing Wellness in Childhood to Promote Lifelong Health: American Academy of Pediatrics; 2018. [doi: 10.1542/9781610021531]
5. Garner A, Yogman M, Committee on Psychosocial Aspects of Child and Family Health, Section on Developmental and Behavioral Pediatrics, Council on Early Childhood. Preventing childhood toxic stress: partnering with families and communities to promote relational health. Pediatrics 2021 Aug;148(2):e2021052582. [doi: 10.1542/peds.2021-052582] [Medline: 34312296]
6. Zuckerman KE, Mattox KM, Sinche BK, Blaschke GS, Bethell C. Racial, ethnic, and language disparities in early childhood developmental/behavioral evaluations: a narrative review. Clin Pediatr (Phila) 2014 Jun;53(7):619-631. [doi: 10.1177/0009922813501378] [Medline: 24027231]
7. National Academies of Sciences, Engineering, and Medicine, Division of Behavioral and Social Sciences and Education, Board on Children, Youth, and Families, Committee on Exploring the Opportunity Gap for Young Children From Birth to Age Eight. In: Hutton R, Allen L, editors. Closing the Opportunity Gap for Young Children: National Academies Press; 2023. [doi: 10.17226/26743]
8. Frosch CA, Fagan MA, Lopez MA, et al. Validation study showed that ratings on the Welch Emotional Connection Screen at infant age six months are associated with child behavioural problems at age three years. Acta Paediatr 2019 May;108(5):889-895. [doi: 10.1111/apa.14731] [Medline: 30702768]
9. Warmingham J, Lavallée A, Curtin P, et al. The quality of early relational health modifies the effect of early life stress on child emerging psychopathology. medRxiv. Preprint posted online on Jan 24, 2026. [doi: 10.64898/2026.01.23.26344636] [Medline: 41646658]
10. Fagan MA, Frosch CA, Middlemiss W, et al. The practical utility of the Welch Emotional Connection Screen for rating parent-infant relational health. Infancy 2019 Nov;24(6):881-892. [doi: 10.1111/infa.12311] [Medline: 32677361]
11. O'Banion DD, Hane AA, Litsas D, Austin J, Welch MG. The Welch Emotional Connection Screen: adapting observational methods to pediatric primary care via resident training. Infant Behav Dev 2021 Nov;65:101629. [doi: 10.1016/j.infbeh.2021.101629] [Medline: 34425493]
12. Diener ML, Hobson-Rohrer W, Byington CL. Kindergarten readiness and performance of Latino children participating in reach out and read. J Community Med Health Edu 2012;2:133. [doi: 10.4172/jcmhe.1000133]
13. Klass P, Dreyer BP, Mendelsohn AL. Reach out and read: literacy promotion in pediatric primary care. Adv Pediatr 2009;56:11-27. [doi: 10.1016/j.yapd.2009.08.009] [Medline: 19968940]
14. Rikin S, Glatt K, Simpson P, Cao Y, Anene-Maidoh O, Willis E. Factors associated with increased reading frequency in children exposed to reach out and read. Acad Pediatr 2015;15(6):651-657. [doi: 10.1016/j.acap.2015.08.008] [Medline: 26409305]
15. Shonkoff JP. Building a new biodevelopmental framework to guide the future of early childhood policy. Child Dev 2010;81(1):357-367. [doi: 10.1111/j.1467-8624.2009.01399.x] [Medline: 20331672]

XSL•FO
RenderX

16. Council on Early Childhood, High PC, Klass P. Literacy promotion: an essential component of primary care pediatric practice. Pediatrics 2014 Aug;134(2):404-409. [doi: 10.1542/peds.2014-1384] [Medline: 24962987]
17. Martin KJ, Beck AF, Xu Y, et al. Shared reading and risk of social-emotional problems. Pediatrics 2022 Jan 1;149(1):e2020034876. [doi: 10.1542/peds.2020-034876] [Medline: 34889450]
18. Canfield CF, Miller EB, Shaw DS, Morris P, Alonso A, Mendelsohn AL. Beyond language: impacts of shared reading on parenting stress and early parent-child relational health. Dev Psychol 2020 Jul;56(7):1305-1315. [doi: 10.1037/dev0000940] [Medline: 32352828]
19. Dexter CA, Stacks AM. A preliminary investigation of the relationship between parenting, parent-child shared reading practices, and child development in low-income families. J Res Child Educ 2014 Jul 3;28(3):394-410. [doi: 10.1080/02568543.2014.913278]
20. Hane AA, LaCoursiere JN, Mitsuyama M, et al. The Welch Emotional Connection Screen: validation of a brief mother-infant relational health screen. Acta Paediatr 2019 Apr;108(4):615-625. [doi: 10.1111/apa.14483] [Medline: 29959878]
21. Center for Early Relational Health. Vagelos College of Physicians and Surgeons, Columbia University. 2025. URL: https://www.vagelos.columbia.edu/departments-centers/center-early-relational-health [accessed 2026-02-10]
22. Clark VLP, Creswell JW. The Mixed Methods Reader: Sage Publications; 2008.
23. Watkins DC. Rapid and rigorous qualitative data analysis: the "RADaR" technique for applied research. Int J Qual Methods 2017 Dec 1;16(1):1609406917712131. [doi: 10.1177/1609406917712131]
24. Early relational health: building foundations for child, family, and community well-being 2025. : The National Academies Press; 2025 URL: https://www.nationalacademies.org/publications/29234 [accessed 2026-04-06]
25. Willis DW, Condon MC, Moe V, Munson L, Smith L, Eddy JM. The context and development of the early relational health screen. Infant Ment Health J 2022 May;43(3):493-506. [doi: 10.1002/imhj.21986] [Medline: 35537064]
26. Williams RC. From ACEs to early relational health: implications for clinical practice. Paediatr Child Health 2023 Oct;28(6):377-393. [doi: 10.1093/pch/pxad025] [Medline: 37744756]
27. Pediatricians are essential supports for families. RAPID Survey Project, Stanford University. 2023. URL: https://rapidsurveyproject.com/our-research/pediatricians-are-essential-supports-for-families [accessed 2026-05-20]
28. Boudreau A, Hamling A, Pont E, et al. Pediatric primary health care: the central role of pediatricians in maintaining children's health in evolving health care models. Pediatrics 2022 Feb 1;149(2):e2021055553. [doi: 10.1542/peds.2021-055553] [Medline: 35104359]
29. Dozier M, Bernard K. Coaching Parents of Vulnerable Infants: The Attachment and Biobehavioral Catch-up Approach: Guilford Press; 2019.
30. Yogman M, Garner A, Hutchinson J, et al. The power of play: a pediatric role in enhancing development in young children. Pediatrics 2018 Sep;142(3):e20182058. [doi: 10.1542/peds.2018-2058] [Medline: 30126932]
31. Matheson KM, Sheridan N, Zarb S, Citron S, Ward M, Hammond B. Early relational health training in Canadian paediatric residency programs: a national program director survey. Paediatr Child Health 2026 Apr;31(3):200-206. [doi: 10.1093/pch/pxaf136] [Medline: 41994042]
32. Riess H, Kelley JM, Bailey RW, Dunn EJ, Phillips M. Empathy training for resident physicians: a randomized controlled trial of a neuroscience-informed curriculum. J Gen Intern Med 2012 Oct;27(10):1280-1286. [doi: 10.1007/s11606-012-2063-z] [Medline: 22549298]
33. Bylund CL, Michaels M, Weiss ES, et al. The impact of an online training program about cancer clinical trials on primary care physicians' knowledge, attitudes and beliefs, and behavior. J Cancer Educ 2021 Oct;36(5):1039-1044. [doi: 10.1007/s13187-020-01731-3] [Medline: 32157570]
34. Tischendorf J, Sharkey C, Westmas K, Zelenksi AB, Hollnagel F, Chapman E. Three-year outcomes of a longitudinal department of medicine fellow as medical educator training program pilot. WMJ 2026;125(1):221-224. [Medline: 41980169]
35. Krick JA, Jung E, Abuhamdeh S, Zheng B. Choose your own adventure! A novel approach in ethics teaching. Clin Teach 2026 Jun;23(3):e70424. [doi: 10.1111/tct.70424] [Medline: 41994960]
36. Wittich CM, Agrawal A, Cook DA, et al. E-learning in graduate medical education: survey of residency program directors. BMC Med Educ 2017 Jul 11;17(1):114. [doi: 10.1186/s12909-017-0953-9] [Medline: 28697744]
37. Cook DA, Levinson AJ, Garside S, Dupras DM, Erwin PJ, Montori VM. Internet-based learning in the health professions: a meta-analysis. JAMA 2008 Sep 10;300(10):1181-1196. [doi: 10.1001/jama.300.10.1181] [Medline: 18780847]
38. Rivera R, Smart J, Sakaria S, et al. Planning engaging, remote, synchronous didactics in the COVID-19 pandemic era. JMIR Med Educ 2021 May 11;7(2):e25213. [doi: 10.2196/25213] [Medline: 33872191]
39. Deepika V, Soundariya K, Karthikeyan K, Kalaiselvan G. "Learning from home": role of e-learning methodologies and tools during novel coronavirus pandemic outbreak. Postgrad Med J 2021 Sep;97(1151):590-597. [doi: 10.1136/postgradmedj-2020-137989] [Medline: 33154099]
40. Terry DL, Graham J, Terry CP, Telge M. Methods for studying bias in graduate medical education. PRiMER 2025;9(34):34. [doi: 10.22454/PRiMER.2025.759472] [Medline: 40727870]
41. Dabir S, Hoseinzadeh M, Mosaffa F, et al. The effect of repeated direct observation of procedural skills (R-DOPS) assessment method on the clinical skills of anesthesiology residents. Anesth Pain Med 2021 Jan 24;11(1):e111074. [doi: 10.5812/aapm.111074] [Medline: 34221940]

XSL•FO
RenderX

42. Clark BW, Lee YZJ, Niessen T, Desai SV, Garibaldi BT. Assessing physical examination skills using direct observation and volunteer patients. Diagnosis (Berl) 2021 Feb 23;8(1):101-110. [doi: 10.1515/dx-2019-0089] [Medline: 32167926]

Abbreviations:

AAP: American Academy of Pediatrics
ERH: early relational health
GRAMMS: Good Reporting of a Mixed Methods Study
IRB: institutional review board
STROBE: Strengthening the Reporting of Observational Studies in Epidemiology

Edited by A Stone; submitted 05.Aug.2025; peer-reviewed by C Crosh, J Guo; revised version received 04.May.2026; accepted 05.May.2026; published 16.Jun.2026.

Please cite as:
Erickson E, Amati JB, Dumitriu D, Miller-Fitzwater A, Best DL, Preiss D, Uejo E, Brown A, Crotty JE, Gillen L, Grier C, Gustafson KK, Shearman N, Lavallée A, Warmingham J, Arduin E, Shearman N, Young J, Sperling J
Implementation of Emotional Connection Training in Pediatric Primary Care: Mixed Methods Study
JMIR Med Educ 2026;12:e81250
URL: https://mededu.jmir.org/2026/1/e81250
doi:10.2196/81250

© Elizabeth Erickson, J Blakely Amati, Dani Dumitriu, Anna Miller-Fitzwater, Debra L Best, Doreet Preiss, Elena Uejo, Adrian Brown, Jennifer E Crotty, Leah Gillen, Cherece Grier, Kristina K Gustafson, Nicole Shearman, Andréane Lavallée, Jennifer Warmingham, Elena Arduin, Nikki Shearman, Jackie Young, Jessica Sperling. Originally published in JMIR Medical Education (https://mededu.jmir.org), 16.Jun.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Education, is properly cited. The complete bibliographic information, a link to the original publication on https://mededu.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Exploring the Role of Early Career Medical Professionals From a Digital-Oriented University in Germany in Promoting Digital Health in Professional Settings: Qualitative Interview Study

Christopher Koch[1,2]; Jan P Ehlers[2], Prof Dr med vet; Julia Nitsche[2*], Dr rer medic; Theresa Sophie Busse[1*], Prof Dr rer medic

[1]Faculty of Health, Witten/Herdecke University, Alfred-Herrhausen-Straße 50, 58455 Witten, Witten, Germany

[2]Department of Didactics and Educational Research in Health Science, Faculty of Health, Witten/Herdecke University, Witten, Germany

[*]these authors contributed equally

Corresponding Author:
Theresa Sophie Busse, Prof Dr rer medic
Faculty of Health, Witten/Herdecke University, Alfred-Herrhausen-Straße 50, 58455 Witten, Witten, Germany

Abstract

Background: To address care delivery gaps, the health care system must embrace innovative digital solutions. Additionally, the rising integration of digitalization as a topic into medical education is providing students with broader opportunities to engage with digitalization overall. As digital health becomes an increasingly integral component of medical education and health care practice, digitally affine early career medical professionals constitute a vital resource for advancing digitalization within the health care sector.

Objective: This study examines how early career medical professionals from Witten/Herdecke University in Germany, with various courses focusing on digitalization, acquire digital knowledge, apply it across diverse practical contexts—ranging from start-ups and corporate environments to traditional clinical settings—and how they contribute to the advancement of digitalization within the entire health care sector.

Methods: Using a qualitative approach, 19 interviews were conducted with early career medical professionals who graduated in the last 15 years at Witten/Herdecke University. Subsequently, the interviews were transcribed and analyzed using a deductive-inductive approach.

Results: The findings reveal that medical graduates often acquire digital skills through intensive self-study and learning on the job, integrating them in various ways into their professional lives. Moreover, while graduates recognize their high potential to make their own contribution to the advancement of digitalization, they also face significant barriers such as knowledge gaps, limited resources, and complex regulations, which hinder their ability to contribute to digitalization in a variety of professional settings. Medical graduates report that they face a pressing need for enhanced knowledge access, improved institutional frameworks, and supportive policy measures to maximize their potential in advancing digitalization initiatives.

Conclusions: Recent medical graduates represent an underused resource for health care digitalization. Unlocking this potential requires coordinated action across medical education, health care institutions, and policymaking to create appropriate conditions for graduates to actively drive digitalization.

(JMIR Med Educ 2026;12:e86107) doi:10.2196/86107

KEYWORDS

digital skills; digitalization; health care; medical education; medical graduates

Introduction

Background

The entire health care system in Germany is under large pressure to solve existing care delivery gaps and to maintain an appropriate and patient-oriented health care infrastructure [1,2]. Already today, health care providers and patients are confronted with many emergent changes and obstacles, preventing optimal medical care. In this context, one possible approach is the increased integration of digital solutions into the whole health care industry [3,4]. The provision of health care and well-being via technology itself is referred to by the umbrella term “digital health,” including various forms [5]. In general, 3 different types of digital offerings can be distinguished: the lowest level is “digitization,” where analog components are converted to digital data as an earlier technical process [6]. Followed by “digitalization” referring to “the integration of digital technologies into public health operations” [6], which affects all health care processes in a public health perspective.

XSL•FO
RenderX

Digitalization additionally refers to the use of digital technologies to optimize business processes and value creation [6]. This is clearly distinct from digital transformation as a multifaceted process including a change in the culture and form of health care service delivery as well as an extension beyond the health care sector [6].

Digitalization in Health Care

Therefore, as in many sectors, digitalization has also emerged as a pivotal strategy in health care [3,7,8], addressing care delivery gaps and offering innovative solutions to enhance the efficiency, quality, and accessibility of health care services [9]. Thereby, many stakeholders can benefit from digital opportunities [10,11]. Patients can use digital services to develop a deeper understanding of their illnesses, to save travel and waiting time by moving medical consultations to the digital space, or to integrate innovative treatment methods in the recovery process. Furthermore, digitalization enables patients to become an active and preventive part of their own care [10,12,13]. Additionally, medical practitioners can use digital services to reduce administrative work, to improve diagnostic quality, or to foster exchange with other health care institutions [10,14,15]. But also, health insurance companies can benefit from an overall efficiency potential, that is, through the reduction of double treatments or new emerging marketing opportunities [16]. To improve the actual practical standards in Germany, health care organizations must find ways to integrate more digital technologies—that is, the meanwhile better-known electronic patient record, the electronic prescription, the telematics infrastructure messenger, or telemedicine—into daily practice [17,18]. To realize this, health care institutions must solve different challenges; that is, they need to overcome institutional resistance or integrate digital approaches that align with data security and legal constraints [19-21]. Overall, there is further potential for improvement in the integration of technology and the design of pathways—especially regarding emerging technologies such as artificial intelligence (AI) [22-25]. Therefore, the health care sector must seek ways to integrate the full potential of digitalization into practice, while existing challenges may continue to slow down progress. Recent advances in AI and the growing prevalence of online misinformation increase the need for digital health literacy. Being able to evaluate the credibility of digital health information and services is essential to distinguish high-quality, evidence-based offers from misleading or potentially harmful content [26,27]. One way of supporting the digital advancement within the health care sector could be the consideration of early career medical professionals [28,29].

Early Career Medical Professionals and Digitalization

Looking at their individual imprints, these early career medical professionals have grown up in the digital age and have experienced the emergence of digitalization since their early childhood [30-32]. In doing so, they have learned to constantly adapt to digital innovations and to use them in daily life [33]. In this context, they are particularly open to digital solutions and can handle computer-based technologies in an appropriate way [33,34]. Consequently, these people already have a pronounced digital affinity [35]. Also, during their studies, those medical students want to understand the benefits of new digital apps, making them ready to shape their professional environment [28,29].

At the same time, overall, digital literacy is becoming one of the key skills on the job market [36]. These generations will also occupy a large part of the labor market in the coming years and thus shape the entire health care industry [37-39]. Consequently, after graduation, these cohorts can become pioneers in transferring digital knowledge, that is, from academia or their private learning to the health care practice [40,41].

However, extensive private use of digital media does not automatically ensure that students integrate these technologies into their studies effectively, highlighting the need for clear education training strategies [42-44]. Consequently, to make an appropriate practice transfer come true, universities must educate students based on their digital needs and set the right framework by promoting appropriate and supporting digital content [45,46].

Digitalization in Medical Studies

At least since the COVID-19 pandemic, increased digitalization is also affecting the study experience of medical students [47-49]. In recent years, more and more medical content has been taught in or transformed into digital formats [43,50,51]. Furthermore, the topic of digital health is also increasingly being integrated into the elective areas of various degree programs, showing the importance of digitalization in practice and consideration in medical education [52,53]. Thus, the medical curriculum is evolving, aiming to develop future-oriented digital skills and competencies [54]. In this context, medical lecturers as well as students classify the thematization of digitalization in study programs as an important part of modern medical education [55,56]. Nevertheless, as the National Competence-Based Learning Objectives Catalogue for Medicine in Germany has decided to allocate only a limited amount of time to digital content within the medical curriculum, many universities still consider a few digital medical topics throughout their studies, deriving an existing improvement potential in the full medical education market [57-59]. Regardless of the digitalization of study programs, students are also looking for new ways to optimize their studies with digital offerings, primarily focusing on innovative and digital resources to enhance learning outcomes and performance evaluation [60-62]. Here, digital learning opportunities give students an additional way to acquire and stabilize knowledge within their medical studies [63]. In addition to that, the student's environment also becomes more digital, as universities are digitizing their processes and structures [64,65]. All these developments together show that medical students are confronted with an increased number of digital elements in their day-to-day life while becoming a professional physician, highlighting at the same time the improvement potential by this cohort.

Research Question

It becomes clear that today's medical students are experiencing an increased digitalization in health care, which will shape their professional future. Based on their personal experiences, they

XSL•FO
RenderX

are probably open to digital innovations and new technologies. Exactly matching, increasing digitalization also offers an opportunity to improve the entire health care system and to solve existing care delivery gaps. Therefore, health care actors must find ways to better implement useful digital solutions into practice and discover problematic solutions or hurdles. As few existing studies have investigated the potential of early career medical professionals in shaping digitalization in health care practice, this study tries to connect these 2 strands and focuses on how, when, and why early career medical professionals can (not) transfer digital skills and knowledge into their professions, improving the practical standard—focusing on the perspective of this target group. Therefore, this study investigates early career medical professionals to understand the potential practical transfer, deriving implications for medical education and practice. Consequently, the research question is "How can early career medical professionals contribute to advance digitalization and digital transformation in practice?"

To answer this research question in full, 3 more subresearch questions were formulated:

- In which contexts do medical students learn digital skills, and how are these applied in practice?
- What potential do early career medical professionals have to contribute to advancing digitalization and digital transformation?
- What challenges does this target group face, and what possible solutions are there?

Methods

Ethical Considerations

This study, as well as its ethical and practical implications, has been approved by the ethics committee of Witten/Herdecke University (UWH), Faculty of Medicine (application number S-254/2024). All participants provided informed consent. Data were anonymized for the presentation in this paper. Participants did not receive compensation.

Reporting Guideline

The COREQ (Consolidated Criteria for Reporting Qualitative Research) checklist [66] was applied and can be found in Checklist 1.

Research Purpose

This qualitative study serves to provide insights into the practice transfer of digital skills by early career medical professionals. Thus, the research team wants to understand how and in which context early career medical professionals learn and apply digital skills, contributing to the advancement of digitalization in the health care sector. This research tries to contribute to this ambiguous problem by preparing fundamental content, which needs to be tackled with explorative research [67]. To gain these specific insights, a qualitative approach was adopted. Therefore, data were primarily collected for this specific research question by conducting exploratory interviews [68]. This research approach was deliberately chosen because qualitative interviews enable a deeper exploration of participants' experiences and further data interpretation. Furthermore, as the research team pursues a mixed methods approach, the results will serve as the basis for an upcoming quantitative study.

Sample and Recruitment

The participating interview partners were selected based on the following four criteria:

1. Only medical graduates of UWH as early career medical professionals were considered, as various optional courses on the basics of digitalization and digital health are offered at this university [69,70], and best practices should be identified.
2. Early career medical professionals from the 2 medical specialties offered at UWH were included: human medicine and dentistry.
3. Only early career medical professionals who have encountered digitalization in their studies or professional work in any form were considered (self-disclosure).
4. As the study focused on the role of early career medical professionals, only individuals who had graduated within the past 15 years were included. This broad time span was selected to encompass both individuals at the beginning of their careers and those with at least some distance to reflect on their experiences.

Participants were invited to take part in the study via multiple channels, including face-to-face meetings, telephone interviews, email, and the social media platform LinkedIn (LinkedIn Corporation). The participants were recruited from September 23, followed by a smooth transition to the interview phase. In this context, participants were contacted and interviewed until theoretical saturation was reached, meaning that further interviews no longer provided relevant new insights.

Interview Guide

The interview guide can be found in the Multimedia Appendix 1.

To ensure comparability across the interviews, a semistructured interview guide was developed. In this context, the SPSS method (Sammeln, Prüfen, Sortieren, Strukturieren [German]; collecting, checking, sorting, subsuming [English]) was used for collecting, checking, sorting, and subsuming the questions within the research team [71]. Furthermore, the existing literature was considered while developing the guide. In sum, this allowed us to explore 3 key aspects in depth: learning digital skills, practical perspective (current status, potential, and challenges), and overcoming challenges. Each topic was accompanied by main and potential follow-up questions, enabling the researchers to address up to 4 questions per category. At the beginning of each section, open questions were asked, which became more closed as the interview section progressed. In addition to that, a deliberate question formulation was integrated to avoid social desirability and to foster creative thinking [68,72]. During the interviews, participants were encouraged to speak freely and share their experiences. Overall, the guide served as a framework to structure the interviews, enabling possible adaptation to changing interview situations.

Prior to conducting the interviews, the practical applicability was successfully tested with a smaller pilot group (n=4), which

XSL•FO
RenderX

led to further refinement before finalizing the questions. The interview recordings from the pilot were neither transcribed nor included in the final analysis.

Data Collection

From October 1, 2024, to November 30, 2024, 19 interviews were conducted by CK either in person or in a digital space by Zoom (version 6.4.10; Zoom Communications Inc).

Before conducting the interviews, the participants were informed about the study's objective, design, and recording procedures. In this context, complete anonymity was assured throughout the evaluation and publication processes. All participants subsequently provided their consent to participate under the outlined conditions. For their participation, the interview partner received no remuneration. After the interviews, CK transcribed the data according to Przyborski and Wohlrab-Sahr [73].

Data Analysis

In the next step, the interviews were analyzed and coded using MAXQDA (version 24; VERBI Software), a software suitable for qualitative data analysis. This was done between December 1, 2024 and January 31, 2025.

A qualitative content analysis based on Kuckartz and Rädiker was conducted [74]. The analysis was initially carried out separately by CK and TSB for 2 interviews. They then met to discuss the interviews, and an initial coding system was developed inductively. CK continued to work with this coding system and refined it while analyzing the remaining interviews. Where there were questions regarding categorization, CK consulted TSB, and they reached a mutual consensus. Once CK had completed the analysis of all interviews, TSB, JN, and JPE reviewed the findings. This was followed by a joint meeting with CK, JN, JPE, and TSB to discuss and finalize the analysis and coding system. Any discrepancies were resolved jointly at this meeting.

The final category manual presents a detailed overview of the individual categories and subcategories. Here, the first 3 coding levels were reported, as these concisely summarize the most important findings. Furthermore, the reporting will focus exclusively on codes that are relevant to addressing the research question. All published quotes have been translated into the English language and have been adapted grammatically.

Results

Overview

In the following, the key findings are presented, structured by the final coding scheme. A total of 19 interviews were conducted, lasting between 20 and 48 minutes, with a total duration of 10.4 hours, equivalent to 624 minutes. On average, each interview lasted 31 minutes. The participants' characteristics are presented in Table 1.

Table . Participants' statistics.

Parameters	Values, n (%)
Participants	19 (100)
Sex	
Male	11 (58)
Female	8 (42)
Background	
Human Medicine	17 (89)
Dentistry	2 (11)
Position	
Physician	13 (68)
Entrepreneur	6 (32)
Age (years)	
Minimum	26
Maximum	40
Average	34
Interview duration (min)	
Minimum	21
Maximum	48
Average	33

Overall, a total of 1640 coded segments and 1957 distinct codes were obtained through the entire analysis. Here, 5 codes belong to the first order, 17 codes to the second order, and 58 codes to the third order. The code structure can be found in Table 2. In the following, these first-, second-, and third-level codes are described in more detail. The participants did not provide feedback on the research findings.

XSL•FO
RenderX

Table . Code structure.

First-order code	Second-order code	Third-order code
Occasion learning digital skills	Job	For concrete activities
		As a basic working method
	Private	Due to digital touchpoints
		Through self-study
	Study	Through teaching
		Through research projects as part of dissertations
		In self-study during exam preparation
Application purpose digital skills	Administrative	Digital office software
		Clinical programs
	Strategic	Foundations
		Business model development
		Process optimization
	Communicative	Training and further education of others
		Leadership and negotiation
Potentials of early career medical professionals in advancing digitalization	Organizational characteristics	Recognition
		Error acceptance
	Personality traits	Impartiality
		Motivation
		Load capacity
		Willingness to change
		Learning ability
		Pragmatism
		Naivety
	Professional competences	Affinity and intuition
		Knowledge
		Digital ideas and interest
		Positive professional attitude
		Study proximity
	Social skills	Willingness to compromise
		Critical perspective

XSL•FO
RenderX

First-order code	Second-order code	Third-order code
Challenges of early career medical professionals in advancing digitalization	Organizational	Role in the organization
		Diverging points of view
		Hierarchy
		Inadequate structures and processes
		Lack of incentives
		Lack of technical implementation options
		Established analog functionality
	Personal	No motivation
		Lack of personality traits
		Lack of knowledge
	Subject-specific	Regulation
		Data protection
		Personal burden
		Lack of networking
		Lack of investment and infrastructure
		High cost pressure
Success factors and possible solutions	Education	School
		Study
		Further training
	Alignment of the organization	Positioning for digitalization
		Support from management and leadership
		Change management
		Thinking through the system and culture
		Providing space and infrastructure
		Building digitalization expertise
	Politics and the state	Provide infrastructure
		Reviewing regulation
		Optimize data protection

Learning Digital Skills

Overview

Early career medical professionals report that they have enhanced their digitalization knowledge in various settings—particularly during their studies, in their professional roles, and in their personal lives.

Study

During their studies, early career medical professionals were introduced to digitalization through curriculum-based teaching, dissertation projects, and self-directed learning. However, digitalization is still relatively underrepresented in the medical curriculum—even at UWH. When addressed, it is typically limited to foundational courses within the selection area, offering just a broader overview of digitalization potential, digital apps, or basic technical tools.

> *I attended a series of lectures on digitalization in medicine, which focused on innovations and their potential applications in future medical practice.* [Interviewee No. 16, Human Medicine, Physician, Female]

Consequently, early career medical professionals receive limited formal preparation to increase their digital knowledge within their medical study programs. Furthermore, early career medical professionals report that they familiarize themselves with digital tools through intense self-studies, particularly during exam preparation. In addition to that, some respondents acquired their foundational digital skills through past high school activities.

Sometimes, digitalization is explored in medical research, especially during dissertation projects.

Job

In their professional roles, early career medical professionals develop digital skills through an inherently digital work

XSL•FO
RenderX

environment and related collaboration models. As a result, they frequently gained experience with digital programs, including advanced AI-based tools, which they encounter as part of their daily practice. Throughout this process, early career medical professionals passively acquired new digital skills, shaped by the demands and technologies prevalent in their specific environments. Additionally, early career medical professionals had to acquire deeper digital skills to perform specific job-related tasks and activities, especially to solve emergent entrepreneurial challenges. Therefore, the required skills were often developed through a hands-on learning-by-doing-approach.

> *Learning digital skills often happens naturally when encountering them in daily work, especially when solving an immediate problem. It's not something I actively sought out, thinking, "What could I learn about digitalization today?" Rather, it's something that gradually becomes part of the workflow.* [Interviewee No. 18, Human Medicine, Physician, Female]

Private

In their private lives, early career medical professionals report that they have developed digital skills through constant confrontation with digitalization over the years. Regular interaction with digital technologies in everyday activities—often since childhood—has fostered an intuitive and ongoing engagement with a variety of digital elements. Thus, this passive yet consistent immersion has naturally shaped their digital competencies over time. Beyond this passive acquisition, many early career medical professionals have actively developed digital skills through various personal initiatives. Here, some graduate students learned coding or programming during their leisure time, which broadened their technical and digital understanding in depth, enabling them to take on special professional tasks in practice later. Driven by personal interest, some early career medical professionals also pursued self-study through courses and training programs, deepening their engagement with overlapping digital megatrends.

> *I have also acquired some digital skills in my personal life out of genuine interest. For instance, I am currently learning to code. For me, this skill that is particularly relevant for understanding the growing prominence and impact of AI.* [Interviewee No. 6, Dentistry, Co-Founder, Male]

Application Purpose of Digital Skills

Overview

Early career medical professionals report that they have applied their digital skills in different ways—especially in administrative, strategic, or communicative paths. Consequently, there are different approaches to bringing their digital knowledge into practice and sharing it with others.

Administrative

An administrative application occurs when early career medical professionals leverage their digital skills to manage simple and existing tasks or processes. In their workplaces, most early career medical professionals use digital software, tools, or communication platforms that have already been integrated—both in everyday clinical practice and in office-related settings.

> *In the hospital setting, a significant amount of time is spent on computers due to the need to document all information using digital systems.* [Interviewee No. 19, Human Medicine, Physician, Female]

Due to their strong digital affinity, early career medical professionals report that they can intuitively understand and incorporate these tools into their daily routines. The overarching goal of these efforts from the early career medical professionals' perspective is to save time and ultimately improve the quality of work by enhancing workflow and coordination within their professional settings. In addition to that, some early career medical professionals take on other roles regarding digitalization, that is, acting as key users for a digital app.

Strategic

A strategic application occurs when early career medical professionals harness their digital skills to creatively and entrepreneurially advance digitalization initiatives. Thus, some digitally affine early career medical professionals even enhance internal processes by introducing new software, communication tools, or workflows, that is, digital lean management. Furthermore, early career medical professionals have also engaged in various entrepreneurial ventures, either independently or in the respective organization they are working at. In this context, they pursue a variety of different initiatives, such as developing digital business models or founding digital health-tech start-ups to address emerging market opportunities. As an example, some early career medical professionals have developed digital knowledge tools, providing easy access to relevant medical knowledge. Ultimately, these creative endeavors have enabled early career medical professionals to generate entrepreneurial value while actively contributing to the overall advancement of digitalization.

> *I would say that even before AI, we digitized quite a lot here in the company and simply looked into the processes, both on improvement potential on the patient side and in the processes towards employees.* [Interviewee No. 11, Dentistry, Managing Director, Male]

Communicative

A communicative application occurs when early career medical professionals share their digital skills with others through various means. In this context, especially early career medical professionals in a clinical practice pass on their digital expertise by organizing training sessions and engaging in dialogues with colleagues, where they inform them about emerging digital opportunities and recommend best practices. Consequently, some early career medical professionals also serve as multipliers, facilitating the adoption and integration of digital solutions. Furthermore, early career medical professionals also apply their digital skills in different management and negotiation contexts, using their expertise to enhance final work results.

> *It's also important to take other medical colleagues by the hand and say: Hey, watch out, do it like this.*

XSL•FO
RenderX

> *I would say that I'm very active in this respect, that I'm a bit of a multiplier* [Interviewee No. 8, Human Medicine, Physician, Male]

Potential of Early Career Medical Professionals in Driving Digitalization Forward

Overview

Early career medical professionals possess personal attributes and professional conditions that enable them to advance digitalization. These encompass organizational characteristics, personality traits, professional skills, and social skills.

Organizational Characteristics

Organizational characteristics encompass the advantages that early career medical professionals derive from their specific roles within their respective organizations. Therefore, early career medical professionals perceive that their needs are increasingly recognized as important by their employers, particularly considering demographic shifts that make these young professionals a valuable future resource. Aware of their value, some early career medical professionals report that they actively advocate for digital offers and, in some cases, even consider leaving employers who fail to meet their digital expectations. Additionally, early career professionals perceive themselves to benefit from a form of entry-level protection, where mistakes are more readily accepted. Furthermore, their perceived risks of failure remain relatively low. As a result, young early career medical professionals possess considerable potential to advance digitalization through their unique starting position.

> *You always enjoy a kind of "puppy protection" at the beginning. So, if you make mistakes, everyone tries to help you, and mistakes tend to be tolerated. That applies typically to the first six months.* [Interviewee No. 4, Human Medicine, Physician, Male]

Personality Traits

Personality traits encompass the personal characteristics that many early career medical professionals possess, supporting their ability to drive digitalization forward. In this context, early career medical professionals report that they bring an unbiased perspective, allowing them to notice dusty aspects, challenge organizational blind spots, and approach longstanding problems in a different way of doing. In this context, early career medical professionals perceive themselves as maintaining an outsider's perspective that enables them to tackle problems with their fresh insights.

> *A huge advantage that you should try to retain is that you usually look at old problems with a fresh eye and then you may notice certain problems, for example structures that are not running well.* [Interviewee No. 1, Human Medicine, Physician, Male]

Furthermore, their intrinsic motivation and genuine commitment to drive meaningful change in health care, combined with a strong enthusiasm for digitalization, position them—from their perspective—as key agents of digitalization. Moreover, early career medical professionals perceive themselves as highly resilient. In this context, many early career medical professionals report that, having not yet assumed family responsibilities, they face fewer constraints than established leaders within their organizations. This relative flexibility allows them to devote more time and energy to exploring additional topics, such as advancing digitalization. Moreover, they demonstrate a high willingness to implement new digital solutions, as they point out to perceive the high potential of digitalization for real practical improvements. On top of that, they report their adaptability, pragmatism, and a certain degree of naivety as further advantages helping them to contribute to the advancement of digitalization.

Professional Competencies

Professional competencies include the technical and subject-specific skills that many early career medical professionals possess, enabling them to advance digitalization. Therefore, early career medical professionals point out a strong affinity for and interest in digitalization, having mostly grown up in a digital environment where they learned to navigate through digital topics with ease. They possess foundational digital knowledge and, in some cases, also specialized expertise—especially in comparison to older generations. Their enthusiasm for digitalization extends beyond theory, as they describe themselves as actively identifying practical opportunities for digital improvements. Furthermore, early career medical professionals note that they approach data protection with a pragmatic mindset. Additionally, early career medical professionals indicate that their proximity to academic studies could allow them to be specifically trained in digitalization right before entering the workforce, enabling them to immediately apply what they have learned.

> *So, I would say that graduates today no longer must learn the digital language or being digital from scratch, but that for many it is internalized like a second mother language. And that they therefore feel much less of a barrier or fear of having to adapt or learn it.* [Interviewee No. 15, Human Medicine, Co-Founder, Female]

Social Skills

Professional competencies include the interpersonal skills that many early career medical professionals possess, enhancing their ability to drive digitalization forward. Here, early career medical professionals perceive themselves tending to question the status quo, seeking ways to improve or optimize existing digital structures. Consequently, early career medical professionals report that they often adopt a critical perspective, questioning whether things are truly effective or necessary. Additionally, they describe themselves as open to compromise, which enables them to collaboratively find solutions when implementing digital solutions.

> *In conclusion, I think it is simply important to remain critical. In other words, to always ask critical questions: What is the opportunity? What is the limit? Is it useful? And we can certainly contribute a lot to this because we have been brought up as critical minds in the broadest sense.* [Interviewee No. 18, Human Medicine, Physician, Female]

XSL•FO
RenderX

Challenges for Early Career Medical Professionals in Advancing Digitalization

Overview

But early career medical professionals also face general and target group-specific challenges—especially organizational, personal, and subject-specific hurdles.

Organizational

Organizational challenges encompass all obstacles that arise within the organization itself. As an example, some early career medical professionals state that they face challenges in gaining recognition within their organizations at all—especially in clinical settings. In this context, they report that they have limited influence and that driving digitalization is not perceived as one of their core responsibilities—both by their employers and by themselves. Thus, their initial priority is to establish their role and to integrate into existing structures, a process further complicated by frequent transitions between different workplaces. Consequently, early in their careers, they often focus on fulfilling assigned tasks and hesitate to step into a shaping role, fearing mistakes or drawing too much attention. Diverging perspectives within the workforce present additional obstacles that early career medical professionals describe. While some professionals advocate for digitalization, others—particularly senior staff—remain committed to analog practices, exhibiting limited openness to change. Hierarchical dynamics further exacerbate this issue, as decision-making power over digital initiatives rests primarily with established professionals, who—from the perspective of early career medical professionals—often resist transformation to maintain existing structures.

> *Of course, we are fresh, young and over-committed when it comes to some things. But that is simply not heard. [...] People don't take us that seriously. We young people come and say we want digital, and then the older ones say, yes, good for you.* [Interviewee No. 19, Human Medicine, Physician, Female]

Furthermore, early career medical professionals report that organizational constraints overall also impede digital progress. In this context, limited personnel resources, inflexible institutional frameworks, and the inability to disrupt ongoing health care operations create barriers to a successful digital improvement—especially in clinical settings. Moreover, early career medical professionals report a lack of incentives to engage in digitalization efforts, as neither formal support nor tangible rewards are provided. Likewise, technical limitations prevent the seamless implementation of digital solutions. Additionally, the existing functionality of analog systems diminishes the perceived urgency for digital changes.

Personal

Personal challenges encompass all obstacles stemming from the individual characteristics and circumstances of early career medical professionals. Here, not all early career medical professionals are intrinsically motivated to drive digitalization. While some prioritize efficiently fulfilling their true core responsibilities, others find digitalization less engaging or are reluctant to invest additional effort, particularly when it requires extra work or overtime. Additionally, some early career medical professionals do not feel as digitally inclined. Furthermore, they identify certain personality traits they feel they lack, which hinder their ability to contribute to digitalization. In this context, some exhibit a strong aversion to risk, some a lack of confidence, and some a lack of entrepreneurial mindset. Furthermore, early career medical professionals highlighted several knowledge gaps that hinder the adoption of digital tools, such as limited technical expertise, limited digital knowledge, and difficulties in navigating organizational processes.

> *There are few who are really motivated, so the term work-life balance comes up very early on. And to be honest, I also did these projects in my free time in overtimes.* [Interviewee No. 7, Human Medicine, Managing Director, Male]

Subject-Specific

Professional challenges include all obstacles that are rooted in specific content-related issues. First, even at the beginning of their career, health care professionals report that they are still overburdened, leaving little time and energy to engage in shaping digitalization. This was reported only by employees in a clinical setting.

> *What prevents young physicians from doing this is the time aspect. So you are overloaded with an extremely large number of medical and non-medical tasks.* [Interviewee No. 1, Human Medicine, Physician, Male]

Second, early career medical professionals argue that implementing digital solutions also requires significant financial investments, which are often hard to justify in the cost-sensitive health care sector. Furthermore, many early career medical professionals face bureaucratic processes and an outdated infrastructure. Third, early career medical professionals report that strict regulations prevent digital progress—especially due to strict data protection regulations. On top of that, early career medical professionals report that limited interconnectivity among health care institutions further impedes digital progress, as the lack of standardized digital collaboration systems prevents seamless data exchange.

Opportunities to Drive Digitalization Forward

Overview

Early career medical professionals point out potential solutions to solve existing challenges, focusing on enhanced education, improved practice conditions, and changed political regulations.

Education

The area of education encompasses potential education-oriented measures. In this context, early career medical professionals emphasized the need for more comprehensive digital education and continuous professional development—especially during their medical study programs and, in some cases, also during high school. They suggested that universities should not only provide foundational digital courses but also offer practical, hands-on experiences that reflect real-world challenges and best practices, preparing students to actively contribute to advancing digitalization in health care practice. This preparation should

XSL•FO
RenderX

focus on the specific content of digitalization, but also on the organizational challenges they will face, such as resistance to change. Therefore, universities should provide medical students with strategies on how to navigate these as a recent alumnus effectively. Furthermore, early career medical professionals report that fostering a digital and entrepreneurial mindset is crucial for early career medical professionals to solve a variety of challenges.

> *Digital skills, applications and the courage to integrate this innovation into must also be trained during the study program.* [Interviewee No. 4, Human Medicine, Physician, Male]

Thus, teaching students to think critically about how digitalization can be integrated into health care and different business models is a critical success factor from early career medical professionals' perspective. Furthermore, early career medical professionals state that continuous workplace training is also essential to keep professionals up to date with rapidly evolving digitalization.

> *An employer should send a clear signal to the outside world that it is willing and has interest in digital innovations. They should clearly say: We want to actively promote this.* [Interviewee No. 17, Human Medicine, Physician, Female]

Politics and the State

Additionally, from the participants' perspective, policy reforms are crucial to reduce bureaucratic hurdles, streamline processes, and provide more flexibility for digital initiatives. Therefore, reforms should aim to balance data security with the freedom to push digitalization, allowing for a safe but agile development of digital solutions. Furthermore, the state should establish a robust infrastructural framework to support the development and implementation of digital initiatives.

> *This is very top-down. That's typical with the EU directives, they always come from the top and are not coordinated with the market. And I would like to see these hurdles taken to realize the entire digital functionality.* [Interviewee No. 11, Dentistry, Managing Director, Male]

Discussion

Principal Findings

This qualitative interview study with 19 interviewees provides new insights into the role of early career medical professionals as potential drivers of digitalization within the health care sector. By exploring their digital competencies, application experiences, and perceptions of challenges, it highlights both the opportunities and limitations of this target group. The findings offer theoretical and practical implications for medical education, health care institutions, and policymaking.

Implications of the Findings for Research and Practice

Learning and Applying Digital Skills

Existing research emphasizes the lack of appropriate digital training in medical curricula [55,58]. This study confirms these findings, showing that even at progressive institutions such as UWH, students acquire digital skills mostly informally—through extracurricular activities or on-the-job learning. In this context, participants reported a clear gap between the digital demands of health care practice and the content of their former education, regardless of their year of graduation. Despite being digitally socialized and confident in the use of digital offers [31,32], students expressed frustration over the absence of a structured and practice-oriented digital education. These findings support the calls of the early career medical professionals for integrating digital competencies more and more into national frameworks such as the National Competence-Based Learning Objectives Catalogue for Medicine. At UWH, the demands of Neumann et al [55] and Sorg et al [59] also become evident: early career medical professionals seek a digital education, but this potential remains largely untapped. However, the actual conditions could also enable universities to go in new ways, setting their own digital priorities. This perspective aligns with the arguments of Khafizova et al [54] and Monteiro and Leite [75], who emphasize that universities could play a greater role in preparing students for the upcoming digital reality across various practical settings. This is particularly important because, to date, early career medical professionals appear to have developed their digital competencies primarily through informal means—such as extracurricular learning or on-the-job experiences, with an immediate application process. However, it is important to critically acknowledge that the majority of learning processes are occurring outside the traditional classroom, driven by experience and social interaction [76,77]. Nevertheless, as further research pointed out, universities could play an even more important role and make their own contribution to ensure that early career medical professionals enter the workplace digitally prepared [43,44,78]. In this context, digitalization has a transformative potential in education [79]. Consequently, in medical education, it is essential for universities to rethink their approach to teaching digital competencies and to prepare early career medical professionals for a changing job market [80,81]. In this context, universities should incorporate hands-on digital skills training that focuses not only on technical proficiency but also on how to apply these tools in real-world health care settings.

Potentials of Early Career Medical Professionals in Driving Digitalization Forward

This study has made evident the diverse potentials of early career medical professionals, which are likewise addressed in the existing literature.

First, the perceived relevance of recent early career medical professionals for the future of health care organizations reflects broader demographic and workforce trends. As experienced professionals retire and the "war for talent" intensifies, early-career physicians are increasingly regarded as a strategic asset [82,83]. The literature confirms this development, highlighting projected personnel shortages in the health care sector and the growing importance of attracting and retaining qualified talent—including early career medical professionals [84-86]. In light of these trends, health care organizations must adapt more deliberately to the expectations of this target group—for example, by providing the desired digital

XSL•FO
RenderX

infrastructure [87-89]. Taken together, these arguments suggest that the digital expectations of recent early career medical professionals alone can serve as a catalyst for change, particularly as organizations have to respond to these demands, keeping the skilled labor shortages in mind. This finding aligns with Burnes's perspective, which conceptualizes external stakeholder expectations as key drivers of organizational change [90]. In this context, both external pressures—such as the demographic shift—and internal resources—such as early career medical professionals' digital readiness and intrinsic motivation—interact to position early career medical professionals as important agents of digitalization. Building on Burnes's view, this implies that health care organizations must not only acknowledge these expectations but also actively respond by implementing targeted solutions for early career medical professionals. These may include investing in talent development programs, establishing supportive digital infrastructures, or fostering a work culture that empowers early career professionals to contribute meaningfully to digital initiatives.

Second, participants in this study emphasized that early career medical professionals possess a diverse set of professional competencies and personal attributes that position them well to contribute to the advancement of digitalization. However, these characteristics are not fixed. As Asselmann and Specht [91] and Pataki-Bittó and Kapusy [92] note, early career professionals continue to develop essential traits—such as conscientiousness, assertiveness, or team orientation—during their initial years in practice. While these studies support the notion of a developmental trajectory, they also highlight its variability, as not all early career medical professionals display the same levels of confidence, initiative, or digital affinity. It is therefore critical to avoid reductive assumptions based on generational stereotypes [93,94]. Instead, intragroup differences and individual circumstances must be acknowledged, particularly when designing support systems intended to empower recent early career medical professionals to lead digital initiatives. Here, early career medical professionals' individual strengths must be understood in relation to both their specific work environments and their individual characteristics.

Challenges for Early Career Medical Professionals in Advancing Digitalization and Opportunities to Solve

This study also surfaces substantial structural, personal, and regulatory barriers that restrict the actualization of early career medical professionals in making their own contribution to the advancement of digitalization. As in previous studies by Kakale [95] or Nawroth et al [96], this study also points out a lack of formal authority, hierarchical resistance, time constraints, or a fragmented regulatory environment. These findings resonate with broader critiques of the German health care system, particularly concerning overly strict data protection frameworks and their chilling effect on digitalization and innovation [97-99]. Furthermore, participants' references to bureaucratic overload, outdated infrastructure, and the difficulty of implementing digital solutions in analog-dominated settings align with the observations made by Wendling [100], who also emphasizes the systemic nature of these obstacles. It is therefore evident that early career medical professionals face still-documented and persistent challenges within the entire health care system, dampening their intrinsic motivation to act as key drivers of digitalization.

To address these limitations, this study also highlights potential enablers of digitalization and emerging solutions. In particular, structured mentorship, leadership support, and a receptive organizational culture are frequently identified as critical success factors [101], aligning with participants' calls for greater top-down encouragement. Such indirectly acting measures may contribute to creating an environment that increasingly supports the digital initiatives of early career medical professionals. Moreover, consistent with the change management frameworks proposed by Glasl and Lievegoed [102] and Mizrak [103], the findings emphasize the necessity of coordinating diverse stakeholder interests. Early career medical professionals alone cannot be expected to navigate the often conflicting expectations of all different stakeholders alone. Here, effective professional leadership seems to be essential. As Burnes [90] notes, change is not a linear process, but rather a dynamic and ongoing interaction between internal capabilities and external pressures. This perspective is also reflected in the views of the study participants, who call on their organizations to manage emerging change processes systematically in order to respond professionally to evolving and changing conditions. In this context, it may be beneficial for organizations to implement structured change management approaches, as suggested by Mento et al [104], and to adopt targeted interventions at the individual, team, and organizational levels [105]. Using such established reference models may help to meet the growing demand for structured and sustainable change processes, impacting, in the end, also the opportunities for early career medical professionals in advancing digitalization in practical settings. Finally, the political dimension of digitalization cannot be overlooked. The literature highlights a central dilemma: fostering innovation while ensuring robust data protection [106,107]. This tension is also frequently emphasized by participants in this study, underscoring the need for appropriate and balanced solutions by policymakers.

Professional Identity Development as an Explanatory Lens

The findings can be interpreted through professional identity development, which conceptualizes early career as a phase in which early career medical professionals internalize professional norms and negotiate who they are as practitioners [108,109]. In this transition, identity is shaped through participation in workplace communities; newcomers often remain in more peripheral positions until they gain legitimacy and influence [110,111]. This helps explain why participants—despite strong digital affinity—report limited recognition, restricted influence, and that advancing digitalization is not perceived as part of their core role, leading them to prioritize integrating into existing structures and avoiding a shaping role. Early career medical professionals' hesitation is further reinforced by hierarchical decision-making and fear of drawing attention, as well as workload and overtime constraints that crowd out discretionary innovation work. Recent research on digital implementation also suggests that professional identity threats, time pressure, workflow fit, and organizational structures shape clinicians'

XSL•FO
RenderX

willingness to engage with digital change [112]. From a professional identity development perspective, digital change agency may therefore require identity-supportive conditions: visible role models and mentoring, plus organizational legitimation (eg, protected time and formal roles) so that being a digital physician becomes a valued identity rather than an extracurricular add-on [109,113]. This aligns with participants' emphasis on mentorship, leadership support, and receptive culture as key enablers of a new role.

Overall, it becomes evident that a diverse range of stakeholders are playing a crucial role in enabling early career medical professionals to contribute meaningfully to the advancement of digitalization in practice. Considering this, the main recommendations for action are summarized below:

1. At the educational level, medical curricula should systematically integrate comprehensive knowledge about digitalization, digital health, or concrete digital tools. In addition, practical training should equip students with essential digital skills to prepare them for the demands of modern health care.
2. At the organizational level, it is crucial to create an environment that actively supports and promotes digital initiatives of early career medical professionals. This includes establishing mentorship programs, appropriate leadership structures, and incentive systems that genuinely value and reward digital innovation. Furthermore, young professionals should be given the necessary resources, time, and autonomy to develop and implement their own digital solutions.
3. At the political level, regulatory and legal barriers hindering digital progress should be carefully considered and, if necessary and possible, further developed.

Here, without coordinated reform across policy, education, and health care governance, the transformative potential of digitally competent early career medical professionals is likely to remain underused. In sum, a multilevel approach is therefore imperative; early career medical professionals alone cannot bear the burden of driving digital initiatives without supportive structures and an appropriate ecosystem.

Limitations

The results must be viewed with the following limitations in mind. First, the reliance on early career medical professionals limits the generalizability of the results. In this context, older health care professionals, stakeholders within the medical environment, or lecturers could also add valuable insights and perspectives to fully explore the research topic. Second, the study is based on self-reported data. In this context, the findings rely on participants' self-reported perceptions and there is no verification of actual contributions to digitalization efforts or objective assessment of digital competencies. Therefore, participants may overstate their digital abilities and contributions. A further limitation is the absence of a systematic interrater reliability assessment and the limited extent of double-coding, which may reduce the methodological rigor and reproducibility of the coding process. Overall, the findings must be tested with a higher number of participants to generalize and test the elaborated findings, consequently following a subsequent quantitative approach. Furthermore, there is also the option of investigating other geographical areas. Here, as the research team pursues a mixed methods approach, the quantitative study is still in elaboration.

Conclusion

To sum it up, this study revealed the perceived potential and challenges that early career medical professionals experience as they seek to advance digitalization in the health care sector. Early career medical professionals see themselves as well-positioned to drive digitalization, outlining their organizational, personal, professional, and social potential. However, they also perceive facing different organizational, personal, and professional barriers, hindering them from making an appropriate contribution to push digitalization. Consequently, early career medical professionals request support options, especially through improved digital education, organizational support, and changed regulations by the state. In conclusion, from their own perspective, early career medical professionals are an underused resource in the digitalization of health care, which is caught between high motivation and structural obstacles in the entire health care system.

Acknowledgments

Generative artificial intelligence tools were used to check the accuracy of the wording.

Funding

The authors declared no financial support was received for this work.

Data Availability

The datasets generated or analyzed during this study are available from the corresponding author on reasonable request.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Interview guide.

XSL•FO
RenderX

[DOCX File, 14 KB - mededu_v12i1e86107_app1.docx]

Checklist 1
COREQ checklist.
[DOCX File, 22 KB - mededu_v12i1e86107_app2.docx]

References

1. Möller C, Popa A. Herausforderung Zukunft: Das deutsche Gesundheitswesen im Wandel. In: Stich V, Schumann JH, Beverungen D, Gudergan G, Jussen P, editors. Digitale Dienstleistungsinnovationen: Springer Berlin Heidelberg; 2019:511-522. [doi: 10.1007/978-3-662-59517-6_25]
2. Schneider B. Ärzte beklagen Versorgungslücken [Article in German]. Orthop Rheuma 2017 Feb;20(1):44. [doi: 10.1007/s15002-017-1073-y]
3. Weidenauer U. Digitalisierung zum schließen der menschlichen versorgungslücke. In: Matusiewicz D, editor. Apotheke Der Zukunft: Springer Fachmedien Wiesbaden; 2024:47-55. [doi: 10.1007/978-3-658-42960-7_6]
4. Becker R. Telemedizin in der Geriatrie - Möglichkeiten und Grenzen [Article in German]. Geriatr Rep 2021 Mar;16(1):26-27. [doi: 10.1007/s42090-021-1226-7]
5. Fatehi F, Samadbeik M, Kazemi A. What is digital health? Review of definitions. Stud Health Technol Inform 2020 Nov 23;275(67-71):67-71. [doi: 10.3233/SHTI200696] [Medline: 33227742]
6. Iyamu I, Xu AXT, Gómez-Ramírez O, et al. Defining digital public health and the role of digitization, digitalization, and digital transformation: scoping review. JMIR Public Health Surveill 2021 Nov 26;7(11):e30399. [doi: 10.2196/30399] [Medline: 34842555]
7. Haggerty E. Healthcare and digital transformation. Netw Secur 2017 Aug;2017(8):7-11. [doi: 10.1016/S1353-4858(17)30081-8]
8. Dal Mas F, Massaro M, Rippa P, Secundo G. The challenges of digital transformation in healthcare: an interdisciplinary literature review, framework, and future research agenda. Technovation 2023 May;123:102716. [doi: 10.1016/j.technovation.2023.102716]
9. Horak I. Medizinische versorgung zu hause – wie virtuelle und persönliche betreuung für eine gesündere und nachhaltigere zukunft verschmelzen. In: Plugmann J, Plugmann P, editors. Innovationen in Der Gesundheitsversorgung: Springer Fachmedien Wiesbaden; 2023:311-325. [doi: 10.1007/978-3-658-41681-2_20]
10. Lux T, Breil B, Dörries M, et al. Digitalisierung im Gesundheitswesen — zwischen Datenschutz und moderner Medizinversorgung [Article in German]. Wirtschaftsdienst 2017 Oct;97(10):687-703. [doi: 10.1007/s10273-017-2200-8]
11. Dionisio M, de Souza Junior SJ, Paula F, Pellanda PC. The role of digital transformation in improving the efficacy of healthcare: a systematic review. J High Technol Manag Res 2023 May;34(1):100442. [doi: 10.1016/j.hitech.2022.100442]
12. Burkhart S, Hanser F. Einfluss globaler megatrends auf das digitale betriebliche gesundheitsmanagement. In: Matusiewicz D, Kaiser L, editors. Digitales Betriebliches Gesundheitsmanagement: Springer Fachmedien Wiesbaden; 2018:37-55. [doi: 10.1007/978-3-658-14550-7_2]
13. Bamarni A. Ärzte auf Social Media – Einfluss auf die Arzt-Patienten-Kommunikation [Article in German]. best practice onkologie 2025 Feb;20(1-2):40-43. [doi: 10.1007/s11654-024-00619-8]
14. Patrinos GP, Sarhangi N, Sarrami B, Khodayari N, Larijani B, Hasanzad M. Using ChatGPT to predict the future of personalized medicine. Pharmacogenomics J 2023 Nov;23(6):178-184. [doi: 10.1038/s41397-023-00316-9] [Medline: 37726551]
15. Zhang Y, Tan W, Lee E. Consumers' responses to personalized service from medical artificial intelligence and human doctors. Psychol Mark 2024 Jan;41(1):118-133. [doi: 10.1002/mar.21911]
16. Waldschmitt E. Digitalisierung in Der Gesetzlichen Krankenversicherung Oder: Wie Startups Körperschaften Beine Machen: CUVILLIER VERLAG; 2017.
17. Grösgen S, editor. Ökonomische Bewertung Digitalisierter Gesundheitsdaten: Springer Fachmedien Wiesbaden; 2023.
18. Stachwitz P, Debatin JF. Digitalization in healthcare: today and in the future. Bundesgesundheitsblatt Gesundheitsforschung Gesundheitsschutz 2023 Feb;66(2):105-113. [doi: 10.1007/s00103-022-03642-8] [Medline: 36648499]
19. Ilin I, Iliashenko VM, Dubgorn A, Esser M. Critical factors and challenges of healthcare digital transformation. In: Rudskoi A, Akaev A, Devezas T, editors. Digital Transformation and the World Economy: Springer International Publishing; 2022:205-220.
20. Garcia-Perez A, Cegarra-Navarro JG, Sallos MP, Martinez-Caro E, Chinnaswamy A. Resilience in healthcare systems: cyber security and digital transformation. Technovation 2023 Mar;121:102583. [doi: 10.1016/j.technovation.2022.102583]
21. Polacsek-Ernst R, Benda F, Bretschneider J, Pagel C. Digitales mindset im gesundheitswesen – chancen und herausforderungen digitaler technologien. In: Pfannstiel MA, editor. Technologien Und Technologiemanagement Im Gesundheitswesen: Springer Fachmedien Wiesbaden; 2024:89-111. [doi: 10.1007/978-3-658-43860-9_5]
22. Vatandsoost M, Litkouhi S. The future of healthcare facilities: how technology and medical advances may shape hospitals of the future. Hosp Pract Res 2019;4(1):1-11. [doi: 10.15171/hpr.2019.01]

XSL•FO
RenderX

23. Lee CH, Wang C, Fan X, Li F, Chen CH. Artificial intelligence-enabled digital transformation in elderly healthcare field: scoping review. Adv Eng Inform 2023 Jan;55:101874. [doi: 10.1016/j.aei.2023.101874]
24. Kleemann T, Beß A, Antic A. Digitalisierung im gesundheitswesen. In: Fend L, Hofmann J, editors. Digitalisierung in Industrie-, Handels- Und Dienstleistungsunternehmen: Springer Fachmedien Wiesbaden; 2024:637-673. [doi: 10.1007/978-3-658-43441-0_28]
25. Ehizogie Paul A, Chioma Anthonia O, Olumuyiwa Tolulope O. The role of big data in healthcare: a review of implications for patient outcomes and treatment personalization. World J Bio Pharm Health Sci 2024;17(3):198-204. [doi: 10.30574/wjbphs.2024.17.3.0133]
26. Stauch L, Renninger D, Rangnow P, et al. Digital health literacy of children and adolescents and its association with sociodemographic factors: representative study findings from Germany. J Med Internet Res 2025 May 5;27:e69170. [doi: 10.2196/69170] [Medline: 40324766]
27. Feng X, Hui V, Jiang J, Liu M, Li Y, Tian L. The Adult Inpatient eHealth Literacy Scale (AIPeHLS): development and validation study. J Med Internet Res 2025 Oct 14;27:e75657. [doi: 10.2196/75657] [Medline: 41086421]
28. Cretu I, Grigore M, Scripcariu IS. Get ready for Gen Z, our next generation of medical students. Rev Cercet si Interv Soc 2020 Jun 15;69:283-292. [doi: 10.33788/rcis.69.18]
29. Baumgartner M, Sauer C, Blagec K, Dorffner G. Digital health understanding and preparedness of medical students: a cross-sectional study. Med Educ Online 2022 Dec;27(1):2114851. [doi: 10.1080/10872981.2022.2114851] [Medline: 36036219]
30. Klaffke M. Millennials und generation z – charakteristika der nachrückenden arbeitnehmer-generationen. In: Klaffke M, editor. Generationen-Management: Springer Fachmedien Wiesbaden; 2014:57-82. [doi: 10.1007/978-3-658-02325-6_3]
31. Eckleberry-Hunt J, Lick D, Hunt R. Is medical education ready for generation Z? J Grad Med Educ 2018 Aug;10(4):378-381. [doi: 10.4300/JGME-D-18-00466.1] [Medline: 30154963]
32. Schenarts PJ. Now arriving: surgical trainees from generation Z. J Surg Educ 2020;77(2):246-253. [doi: 10.1016/j.jsurg.2019.09.004] [Medline: 31562032]
33. Jones S. The Internet Goes to College: How Students Are Living in the Future with Today's Technology: Pew Internet & American Life Project; 2002. URL: https://www.researchgate.net/publication/240317878_The_Internet_Goes_to_College_How_Students_Are_Living_in_the_Future_with_Today's_Technology [accessed 2026-06-08]
34. Schmidt C, Möller J, Windeck P. Arbeitsplatz Krankenhaus: Vier Generationen unter einem Dach [Article in German]. Dtsch Arztebl Int 2013;19:928-933 [FREE Full text]
35. Brinkmann S. Arbeitswerte Der Generation Y Im Internationalen Vergleich: Springer Fachmedien Wiesbaden; 2020.
36. Law NWY, Woo DJ, La Torre J, Wong KWG. A global framework of reference on digital literacy skills for indicator 4.4.2: information paper no.51. : UNESCO; 2018 URL: https://unesdoc.unesco.org/ark:/48223/pf0000265403.locale=en [accessed 2026-06-06]
37. Desy JR, Reed DA, Wolanskyj AP. Milestones and millennials: a perfect pairing—competency-based medical education and the learning preferences of generation Y. Mayo Clin Proc 2017 Feb;92(2):243-250. [doi: 10.1016/j.mayocp.2016.10.026] [Medline: 28160874]
38. Evans KH, Ozdalga E, Ahuja N. The medical education of generation Y. Acad Psychiatry 2016 Apr;40(2):382-385. [doi: 10.1007/s40596-015-0399-5] [Medline: 26283527]
39. Lourenco AP, Cronan JJ. Teaching and working with millennial trainees: impact on radiological education and work performance. J Am Coll Radiol 2017 Jan;14(1):92-95. [doi: 10.1016/j.jacr.2016.06.029] [Medline: 27554062]
40. Peterson MW, Rowat J, Kreiter C, Mandel J. Medical students' use of information resources: is the digital age dawning? Acad Med 2004 Jan;79(1):89-95. [doi: 10.1097/00001888-200401000-00019] [Medline: 14691004]
41. Machleid F, Kaczmarczyk R, Johann D, et al. Perceptions of digital health education among European medical students: mixed methods survey. J Med Internet Res 2020 Aug 14;22(8):e19827. [doi: 10.2196/19827] [Medline: 32667899]
42. Graham-Jones P, Jain SH, Friedman CP, Marcotte L, Blumenthal D. The need to incorporate health information technology into physicians' education and professional development. Health Aff (Millwood) 2012 Mar;31(3):481-487. [doi: 10.1377/hlthaff.2011.0423] [Medline: 22392658]
43. Kuhn S, Frankenhauser S, Tolks D. Digitale Lehr- und Lernangebote in der medizinischen Ausbildung: Schon am Ziel oder noch am Anfang? [Digital learning and teaching in medical education: Already there or still at the beginning?]. Bundesgesundheitsblatt Gesundheitsforschung Gesundheitsschutz 2018;61(2):201-209. [doi: 10.1007/s00103-017-2673-z] [Medline: 29234823]
44. Jarva E, Oikarinen A, Andersson J, Tomietto M, Kääriäinen M, Mikkonen K. Healthcare professionals' digital health competence and its core factors; development and psychometric testing of two instruments. Int J Med Inform 2023 Mar;171(104995):104995. [doi: 10.1016/j.ijmedinf.2023.104995] [Medline: 36689840]
45. Hege I, Tolks D, Kuhn S, Shiozawa T. Digital skills in healthcare. GMS J Med Educ 2020;37(6):Doc63. [doi: 10.3205/zma001356] [Medline: 33225055]
46. Hernandez-de-Menendez M, Escobar Díaz CA, Morales-Menendez R. Educational experiences with generation Z. Int J Interact Des Manuf 2020 Sep;14(3):847-859. [doi: 10.1007/s12008-020-00674-9]

XSL•FO
RenderX

47. Alfonso Beltran-Sanchez J, Marcela González-Treviño I, Dominguez A. Digital education in times of COVID-19: the experience of medical educators. In: Alfonso Beltran-Sanchez J, Marcela González-Treviño I, Dominguez A, editors. The 4th International Conference on Digital Technology in Education: ACM; 2020:26-31. [doi: 10.1145/3429630.3429633]
48. Egarter S, Mutschler A, Brass K. Impact of COVID-19 on digital medical education: compatibility of digital teaching and examinations with integrity and ethical principles. Int J Educ Integr 2021 Dec;17(1). [doi: 10.1007/s40979-021-00084-8]
49. Holzmann-Littig C, Zerban NL, Storm C, et al. One academic year under COVID-19 conditions: two multicenter cross-sectional evaluation studies among medical students in Bavarian medical schools, Germany students' needs, difficulties, and concerns about digital teaching and learning. BMC Med Educ 2022 Jun 10;22(1):450. [doi: 10.1186/s12909-022-03480-x] [Medline: 35681177]
50. Schick K, Reiser S, Mosene K, et al. How can communicative competence instruction in medical studies be improved through digitalization? GMS J Med Educ 2020;37(6):Doc57. [doi: 10.3205/zma001350] [Medline: 33225049]
51. Gottschalk M, Milch PM, Albert C, Werwick K, Braun-Dullaeus RC, Stieger P. Medical education during the Covid-19 pandemic long-term experiences of German clinical medical students. PLoS One 2023;18(6):e0286642. [doi: 10.1371/journal.pone.0286642] [Medline: 37279236]
52. Tudor Car L, Kyaw BM, Nannan Panday RS, et al. Digital health training programs for medical students: scoping review. JMIR Med Educ 2021 Jul 21;7(3):e28275. [doi: 10.2196/28275] [Medline: 34287206]
53. Kuhn S, Müller N, Kirchgässner E, Ulzheimer L, Deutsch KL. Digital skills for medical students - qualitative evaluation of the curriculum 4.0 "Medicine in the digital age". GMS J Med Educ 2020;37(6):Doc60. [doi: 10.3205/zma001353] [Medline: 33225052]
54. Khafizova AA, Galimov AM, Kharisova SR, Grebenshchikova LY, Yagudina RI, Smirnova LM. The impact of healthcare digitalization on the medical education curricula and programs: points of convergence and divergence. Contemp Educ Technol 2023;15(4):ep479. [doi: 10.30935/cedtech/13768]
55. Neumann M, Fehring L, Kinscher K, et al. Perspective of German medical faculties on digitization in the healthcare sector and its influence on the curriculum. GMS J Med Educ 2021;38(7):Doc124. [doi: 10.3205/zma001520] [Medline: 34957329]
56. Edirippulige S, Gong S, Hathurusinghe M, et al. Medical students' perceptions and expectations regarding digital health education and training: a qualitative study. J Telemed Telecare 2022 May;28(4):258-265. [doi: 10.1177/1357633X20932436] [Medline: 32571157]
57. Haag M, Igel C, Fischer MR, German Medical Education Society (GMA), Committee "Digitization – Technology-Assisted Learning and Teaching", Joint working group "Technology-enhanced Teaching and Learning in Medicine (TeLL)" of the German Association for Medical Informatics, Biometry and Epidemiology (gmds) and the German Informatics Society (GI). Digital teaching and digital medicine: a national initiative is needed. GMS J Med Educ 2018;35(3):Doc43. [doi: 10.3205/zma001189] [Medline: 30186953]
58. Aulenkamp J, Mikuteit M, Löffler T, Schmidt J. Overview of digital health teaching courses in medical education in Germany in 2020. GMS J Med Educ 2021;38(4):Doc80. [doi: 10.3205/zma001476] [Medline: 34056069]
59. Sorg H, Ehlers JP, Sorg CGG. Digitalization in medicine: are German medical students well prepared for the future? Int J Environ Res Public Health 2022 Jul 7;19(14):8308. [doi: 10.3390/ijerph19148308] [Medline: 35886156]
60. Maloney S, Chamberlain M, Morrison S, Kotsanas G, Keating JL, Ilic D. Health professional learner attitudes and use of digital learning resources. J Med Internet Res 2013 Jan 16;15(1):e7. [doi: 10.2196/jmir.2094] [Medline: 23324800]
61. Scott K, Morris A, Marais B. Medical student use of digital learning resources. Clin Teach 2018 Feb;15(1):29-33. [doi: 10.1111/tct.12630] [Medline: 28300343]
62. Lattouf OM. Impact of digital transformation on the future of medical education and practice. J Card Surg 2022 Sep;37(9):2799-2808. [doi: 10.1111/jocs.16642] [Medline: 35612355]
63. Hyll M, Schvarcz R, Manninen K. Exploring how medical students learn with the help of a digital presentation: a qualitative study. BMC Med Educ 2019 Jun 13;19(1):210. [doi: 10.1186/s12909-019-1569-z] [Medline: 31196111]
64. Gürtler L, Rietl J, Zinganell LM. Young Professionals in Publishing – Nachwuchskräfte Im Verlagswesen: Springer Fachmedien Wiesbaden; 2025.
65. Safiullin* MR, Akhmetshin EM. Digital transformation of a university as a factor of ensuring its competitiveness. Int J Eng Adv Technol 2019;9(1):7387-7390. [doi: 10.35940/ijeat.A3097.109119]
66. Tong A, Sainsbury P, Craig J. Consolidated Criteria for Reporting Qualitative Research (COREQ): a 32-item checklist for interviews and focus groups. Int J Qual Health Care 2007 Dec;19(6):349-357. [doi: 10.1093/intqhc/mzm042] [Medline: 17872937]
67. Sarstedt M, Mooi E. Introduction to market research. In: Sarstedt M, Mooi E, editors. A Concise Guide to Market Research: Springer Berlin Heidelberg; 2019:1-9.
68. Aghamanoukjan A, Buber R, Meyer M. Qualitative interviews. In: Buber R, Holzmüller HH, editors. Qualitative Marktforschung: Gabler; 2009:415-436. [doi: 10.1007/978-3-8349-9441-7_26]
69. Zupanic M, Ehlers JP, Fricke J, et al. Qualitative studies on implicit criteria during the individualized selection procedure for medical studies at Witten/Herdecke University (UW/H). GMS J Med Educ 2019;36(1):Doc3. [doi: 10.3205/zma001211] [Medline: 30828603]

XSL•FO
RenderX

70. Nitsche J, Busse TS, Ehlers JP. Teaching digital medicine in a virtual classroom: impacts on student mindset and competencies. Int J Environ Res Public Health 2023 Jan 22;20(3):2029. [doi: 10.3390/ijerph20032029] [Medline: 36767393]
71. Helfferich C. In: Helfferich C, editor. Interviewplanung Und Intervieworganisation: VS Verlag für Sozialwissenschaften; 2011:167-193. [doi: 10.1007/978-3-531-92076-4_6]
72. Iacobucci D, Churchill GA. Marketing Research: Methodological Foundations, 11th edition: Earlie Lite Books Inc; 2015.
73. Przyborski A, Wohlrab-Sahr M. Qualitative Sozialforschung: Ein Arbeitsbuch, 4th edition: DE GRUYTER; 2014.
74. Kuckartz U, Rädiker S. Qualitative Inhaltsanalyse: Methoden, Praxis, Umsetzung Mit Software Und Künstlicher Intelligenz, 6th edition: Belz Juventa; 2024.
75. Monteiro A, Leite C. Digital literacies in higher education. Red ;21(65). [doi: 10.6018/red.438721]
76. Falk JH, Moussouri T, Coulson D. The effect of visitors 'agendas on museum learning. Curator: Mus J 1998;41(2):107-120. [doi: 10.1111/j.2151-6952.1998.tb00822.x]
77. Falk J, Dierking L. The 95 percent solution. Am Sci 2010;98(6):486. [doi: 10.1511/2010.87.486]
78. Scuotto V, Morellato M. Entrepreneurial knowledge and digital competence: keys for a success of student entrepreneurship. J Knowl Econ 2013 Sep;4(3):293-303. [doi: 10.1007/s13132-013-0155-6]
79. Schmidt JT, Tang M. Digitalization in education: challenges, trends and transformative potential. In: Harwardt M, Niermann PJ, Schmutte AM, Steuernagel A, editors. Führen Und Managen in Der Digitalen Transformation: Springer Fachmedien Wiesbaden; 2020:287-312.
80. Mohamed Hashim MA, Tlemsani I, Matthews R. Higher education strategy in digital transformation. Educ Inf Technol (Dordr) 2022;27(3):3171-3195. [doi: 10.1007/s10639-021-10739-1] [Medline: 34539217]
81. Kopackova H, Simonova S, Reimannova I. Digital transformation leaders wanted: How to prepare students for the ever-changing demands of the labor market. Int J Manag Educ 2024 Mar;22(1):100943. [doi: 10.1016/j.ijme.2024.100943]
82. Liebenow D, Haase C, von BC, Nachtwei J. Bestehen im War for Talent: Methodische Qualität des Kompetenzmodells als Überlebensstrategie [Article in German]. Wirtschaftspsychologie 2014;16:25-38 [FREE Full text]
83. Tirrel H, Hohenhorst S, Winnen L. Digital corporate identity in the healthcare sector. In: KL T, AB M, editors. Organizational Risks, Challenges, and Barriers in Developing Sustainability Start-Ups: IGI Global Scientific Publishing; 2025:325-350.
84. Schwendemann H, Thiede M. Herausforderungen der babyboomer-generation – probleme und lösungsansätze im gesundheitswesen. In: Kreuzenbeck C, Schwendemann H, Thiede M, editors. Die Herausforderungen Der Generation Babyboomer Für Das Gesundheitswesen: Springer Berlin Heidelberg; 2023:3-11. [doi: 10.1007/978-3-662-67575-5_1]
85. Johna S. Medizinischer Fachkräftemangel als strukturelles Problem [Article in German]. Inn Med (Heidelb) 2024;65(9):857-864. [doi: 10.1007/s00108-024-01759-3] [Medline: 39115594]
86. Reuß S, Brohm M, Schubert M. Zukunft sichern: Wie werden Gesundheitsberufe wieder unwiderstehlich für den Nachwuchs? [Securing the future: How can healthcare professions become irresistible to young talent again?]. Urologie 2025;64(8):778-783. [doi: 10.1007/s00120-025-02612-7] [Medline: 40377709]
87. Vordenbäumen V, Pfannstiel MA. Auswirkungen digitaler technologien auf die arbeitswelt im krankenhaus. In: Pfannstiel MA, editor. Technologien Und Technologiemanagement Im Gesundheitswesen: Springer Fachmedien Wiesbaden; 2024:783-800. [doi: 10.1007/978-3-658-43860-9_39]
88. Winkelhake U. "Digital lifestyle" – zukünftige mitarbeiter und kunden. In: Winkelhake U, editor. Die Digitale Transformation Der Automobilindustrie: Springer Berlin Heidelberg; 2024:33-51. [doi: 10.1007/978-3-662-68794-9_3]
89. Burfien L, Busch E, Gürtler L, Rietl J, Weckert K. Ein attraktives arbeitsumfeld für junge fachkräfte schaffen. In: Gürtler L, Rietl J, Zinganell LM, editors. Young Professionals in Publishing – Nachwuchskräfte Im Verlagswesen 2025:97-113. [doi: 10.1007/978-3-658-45330-5_9]
90. Burnes B. Managing Change, 7th edition: Pearson Education; 2017.
91. Asselmann E, Specht J. Personality maturation and personality relaxation: differences of the Big Five personality traits in the years around the beginning and ending of working life. J Pers 2021 Dec;89(6):1126-1142. [doi: 10.1111/jopy.12640] [Medline: 33872401]
92. Pataki-Bittó F, Kapusy K. Work value orientations of career starters. Curr Psychol 2025 Jan;44(1):717-734. [doi: 10.1007/s12144-024-07168-x]
93. Jahn R, Nolten A, Weigand M. Generation X,Y,Z oder Alpha? Wir sind so verschieden, wie ihr es seid! [Article in German]. SV 2024 Dec;42(4):42-45. [doi: 10.30820/1431-7168-2024-4-42]
94. Hurrelmann K. Die lebenssituation der jungen generation. In: Bertelsmann E, editor. Zukunft in Die Schule Holen Lebensplanung, Arbeits- Und Berufsorientierung Tagungsdokumentation 2009:14-24. [doi: 10.3278/6004019w014]
95. Kakale MM. Of digital transformation in the healthcare (systematic review of the current state of the literature). Health Technol 2024 Jan;14(1):35-50. [doi: 10.1007/s12553-023-00803-w]
96. Nawroth M, Sabotka S, Fleßa S. Gründe unterschiedlicher diffusion von digital-health-innovationen. In: Nawroth M, Sabotka S, Fleßa S, editors. Digital-Health-Innovation Im Ostseeraum: Springer Fachmedien Wiesbaden; 2024:31-43. [doi: 10.1007/978-3-658-46404-2_5]
97. Čas J. Datenschutz bei Pervasive Computing im Gesundheitswesen [Article in German]. TATuP ;17(1):57-65. [doi: 10.14512/tatup.17.1.57]

XSL•FO
RenderX

98. Hessel S, Jung A. Datenschutzrechtliche Herausforderungen bei der Verwertung von Daten vernetzter Medizinprodukte [Article in German]. Datenschutz Datensich 2024 May;48(5):289-294. [doi: 10.1007/s11623-024-1932-z]
99. Kriegel J, Enzmann M, Wolf R. Medizinische digitale Zwillinge [Article in German]. Datenschutz Datensich 2024 Aug;48(8):532-537. [doi: 10.1007/s11623-024-1970-6]
100. Wendling M. Health data management im krankenhaus: spielball zwischen bund und ländern. In: Henke V, Hülsken G, Schneider H, Varghese J, editors. Health Data Manag: Springer Fachmedien Wiesbaden; 2024:167-173. [doi: 10.1007/978-3-658-43236-2_18]
101. Beer G, Benischek I, Dangl O, Plaimauer C, editors. Mentoring Im Berufseinstieg Eine Mehrperspektivische Betrachtung: LIT Verlag; 2014.
102. Glasl F, Lievegoed B. Dynamische Unternehmensentwicklung: Grundlagen Für Nachhaltiges Change Management: Haupt Verlag; 2016.
103. Mızrak F. Effective change management strategies. In: Giannoccaro I, Babulak E, editors. The Rise of Explainable and Generative AI-Driven Cyber and Information Security: IGI Global Scientific Publishing; 2026:135-162.
104. Mento A, Jones R, Dirndorfer W. A change management process: grounded in both theory and practice. J Change Manag 2002 Mar;3(1):45-59. [doi: 10.1080/714042520]
105. Cameron E, Green M. Making Sense of Change Management: A Complete Guide to the Models, Tools and Tech-Niques of Organizational: Kogan Page Publishers; 2019.
106. Pförringer D, Matusiewicz D, Tsitsilonis S, Gehlen T, Back DA. Ökonomische Aspekte der Digitalisierung in Orthopädie und Unfallchirurgie [Article in German]. Unfallchirurg 2020 Nov;123(11):856-861. [doi: 10.1007/s00113-020-00891-7]
107. Weichert T. Patientengeheimnis, Datenschutz und das Internet [Article in German]. MedR 2024;42(2):84-90. [doi: 10.1007/s00350-023-6668-5]
108. Koh EYH, Koh KK, Renganathan Y, Krishna L. Role modelling in professional identity formation: a systematic scoping review. BMC Med Educ 2023 Mar 29;23(1):194. [doi: 10.1186/s12909-023-04144-0] [Medline: 36991373]
109. Cruess RL, Cruess SR, Boudreau JD, Snell L, Steinert Y. Reframing medical education to support professional identity formation. Acad Med 2014 Nov;89(11):1446-1451. [doi: 10.1097/ACM.0000000000000427] [Medline: 25054423]
110. Wenger E. Communities of Practice: Learning, Meaning, and Identity: Cambridge University Press; 1998.
111. Lave J, Wenger E, editors. Situated Learning: Legitimate Peripheral Participation: Cambridge University Press; 1991.
112. Ackerhans S, Huynh T, Kaiser C, Schultz C. Exploring the role of professional identity in the implementation of clinical decision support systems-a narrative review. Implement Sci 2024 Feb 12;19(1):11. [doi: 10.1186/s13012-024-01339-x] [Medline: 38347525]
113. Hafferty FW. Beyond curriculum reform: confronting medicine's hidden curriculum. Acad Med 1998 Apr;73(4):403-407. [doi: 10.1097/00001888-199804000-00013] [Medline: 9580717]

Abbreviations

AI: artificial intelligence
COREQ: Consolidated Criteria for Reporting Qualitative Research
UWH: Witten/Herdecke University

Edited by A Stone; submitted 18.Oct.2025; peer-reviewed by C Anyaegbuna, M Lange; revised version received 10.May.2026; accepted 12.May.2026; published 24.Jun.2026.

Please cite as:
Koch C, Ehlers JP, Nitsche J, Busse TS
Exploring the Role of Early Career Medical Professionals From a Digital-Oriented University in Germany in Promoting Digital Health in Professional Settings: Qualitative Interview Study
JMIR Med Educ 2026;12:e86107
URL: https://mededu.jmir.org/2026/1/e86107
doi:10.2196/86107

© Christopher Koch, Jan P Ehlers, Julia Nitsche, Theresa Sophie Busse. Originally published in JMIR Medical Education (https://mededu.jmir.org), 24.Jun.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Education, is properly cited. The complete bibliographic information, a link to the original publication on https://mededu.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Principles for Responsible AI in Health Professions Education, Research, and Care: Health CARE-AI (Contextual, Accountable, Responsible, and Equitable Artificial Intelligence) Framework Delphi Consensus Study

Lyn K Sonnenberg[1,2*], BSc, MSc, MEd, EMBA, MD; David Wiljer[1,3,4*], BA, MA, PhD; Muhammad Mamdani[5,6], MA, MPH, PhD; Victor Do[1,2], MSc, MD; Brandon Tang[1,7], MD; Babar Haroon[8], MEd, MD; Bizav Jaffer[1,9], BSc; Jerry Maniate[1,9*], MEd, EMBA, MD

[1]Equity in Health Systems (EqHS) Lab, Bruyère Health Research Institute, 43 Bruyère Street, Ottawa, ON, Canada

[2]Department of Pediatrics, Faculty of Medicine and Dentistry, University of Alberta, Edmonton, AB, Canada

[3]Education Technology Innovation, Digital, University Health Network, Toronto, ON, Canada

[4]Department of Psychiatry, University of Toronto, Toronto, ON, Canada

[5]Department of Laboratory Medicine and Pathology, Temerty Faculty of Medicine, University of Toronto, Toronto, ON, Canada

[6]Leslie Dan Faculty of Pharmacy, University of Toronto, Toronto, ON, Canada

[7]Department of Medicine, University of Toronto, Toronto, ON, Canada

[8]Department of Critical Care, Faculty of Medicine, Dalhousie University, Halifax, NS, Canada

[9]Faculty of Medicine, University of Ottawa, Ottawa, ON, Canada

[*]these authors contributed equally

Corresponding Author:
Lyn K Sonnenberg, BSc, MSc, MEd, EMBA, MD
Equity in Health Systems (EqHS) Lab, Bruyère Health Research Institute, 43 Bruyère Street, Ottawa, ON, Canada

Abstract

Background: Artificial intelligence (AI) is rapidly integrating into health professions education and clinical practice, creating significant opportunities alongside new ethical challenges. Although current international and professional guidance establishes essential values, it offers limited direction for how clinicians, educators, learners, and institutions should act in routine educational, research, and clinical contexts. The CARE-AI (Contextual, Accountable, Responsible, and Equitable Artificial Intelligence) project responds to this practice-level gap by articulating guidance that moves beyond values toward professional accountability and equity, with explicit attention to educational, research, and clinical practice contexts.

Objective: The study objective was to develop and validate a consensus-based, actionable framework of principles to guide responsible AI use across health professions education, research, and clinical care.

Methods: We conducted a 3-phase modified Delphi consensus study, reported in accordance with the Accurate Consensus Reporting Document. Phase 1 involved 2 international professional meetings and 3 purposively sampled focus groups (AI or technology, health professions education, and ethics or professionalism) to adapt and refine draft principles using an exploratory qualitative approach. Phase 2 used an online survey with a 5-point importance scale and prespecified consensus criteria (inclusion ≥70%: high ratings; exclusion ≥70%: low ratings). Phase 3 used include, exclude, or undecided voting on revised principles. Quantitative thresholds determined consensus. Qualitative free-text comments informed iterative refinement.

Results: Participants represented diverse communities of practice across health professions education, clinical care, patient partners, ethics, and digital health, spanning multiple professional roles and training levels. Across all phases, 303 unique participants contributed to the study. Phase 1 focus groups (n=61) provided early insight and direction. In phase 2, the first Delphi survey round, 242 participants initiated the survey, with 120 (49.6%) participants completing it. In phase 3, the second Delphi survey round, 103 participants were invited based on expressed interest at the end of the first round; 78 participants initiated the survey and 75 completed it (75/78, 96.2% of starters). In phase 2, of the 61 statements, 58 (95%) met the inclusion criteria, and participants submitted 1887 comments (697 were content rich), prompting clearer accountability language, stronger equity commitments, and more usable wording. In phase 3, all 10 principles and their statements met the inclusion criteria. Participants contributed 224 comments (179 were content rich) that informed final refinements. Endorsement was near unanimous: 96% (72/75) agreed or strongly agreed that the framework clearly defined professionalism expectations for AI to meet educational, technological, and ethical needs in the health professions.

XSL•FO
RenderX

Conclusions: The Health CARE-AI Framework, with its preamble and 10 principles, articulates actionable, consensus-validated guidance that moves from values to competence, into professional accountability, and toward structural commitments to equity. Paired with a companion implementation guide and toolkit, the framework is intended to support use across education, research, and clinical settings.

(JMIR Med Educ 2026;12:e91626) doi:10.2196/91626

KEYWORDS

artificial intelligence; medical education; health professions education; professionalism; ethics; equity; consensus methods; Delphi technique; responsible artificial intelligence; responsible AI

Introduction

The integration of artificial intelligence (AI) into health care systems has ushered in a new era of medical practice and administration, a transformative development that is reshaping the landscape of health care. These technologies hold great promise for enhancing patient care, improving outcomes, and streamlining processes. However, the rapid adoption of these technologies also presents a unique set of challenges, often unintended, particularly in the realm of professional ethics and conduct. To navigate this transformation responsibly, health care professionals need clear guidance and practical frameworks for using AI in ways that safeguard patient welfare, uphold professional values, and nurture public trust [1-4].

Several international organizations have issued AI ethics frameworks that provide important high-level guidance. The United Nations Educational, Scientific, and Cultural Organization's (UNESCO) "Recommendation on the Ethics of AI" emphasizes human rights, fairness, transparency, and human oversight across sectors, including education, science, culture, public administration, and health [5]. The Organisation for Economic Co-operation and Development (OECD) AI Principles highlight human-centered values, transparency, robustness, and accountability, and have been adopted across government, industry, and public-sector governance frameworks since 2019 [6]. The World Health Organization (WHO) has outlined broad ethical principles for AI in health, including human autonomy, transparency, equity, and accountability [1], and its more recent regulatory considerations reinforce these priorities for national systems [7]. These frameworks provide important foundations but are not designed for use in busy daily clinical workflows or the specific needs of health professions education (HPE), including supervision, assessment, and the relational dimensions of care and learning, defined here as trust-based interactions among patients, learners, educators, and care teams [1,5-7].

Professional bodies and academic consortia have begun to interpret these principles for medicine and medical education. The American Medical Association's policy on augmented intelligence outlines expectations for safety, bias mitigation, transparency, and preservation of clinical autonomy [8]. The Association of American Medical Colleges (AAMC) has issued principles for the responsible use of AI in and for medical education, emphasizing transparency, equitable access, interdisciplinary collaboration, and human-centered focus in teaching and learning [9]. It has also released separate guidance for the use of AI in medical school and residency selection, focusing on issues of bias, validity, and equitable access [10]. While these initiatives highlight the urgency of addressing AI in education, they remain early consensus statements rather than frameworks that have been systematically validated in practice. A recent AAMC curriculum snapshot across Canada and the United States documents rapid AI uptake alongside variability and a lack of common standards [11].

Reviews and guidance documents on AI implementation in medical education consistently point to fragmented efforts and limited evidence. Peer-reviewed articles on generative AI highlight both opportunities for innovation and risks to integrity and professionalism, yet empirical evidence to guide implementation remains scarce [4]. Scoping reviews describe heterogeneous curricula with wide variation in how AI is introduced, a lack of common standards, and little evaluation of long-term outcomes [2]. Systematic reviews reinforce these concerns, identifying persistent gaps in validated competencies and measurable outcomes [12].

Together, these observations indicate that available frameworks explain why ethics matter and what values ought to guide AI, but they do not resolve how clinicians, educators, and learners should act in specific contexts or embed accountability, equity, and relational trust in day-to-day decision-making. This gap created the rationale for our study—Contextual, Accountable, Responsible, and Equitable Artificial Intelligence (CARE-AI)—which sought to develop principles tailored to HPE, research, and clinical care through a structured international consensus process.

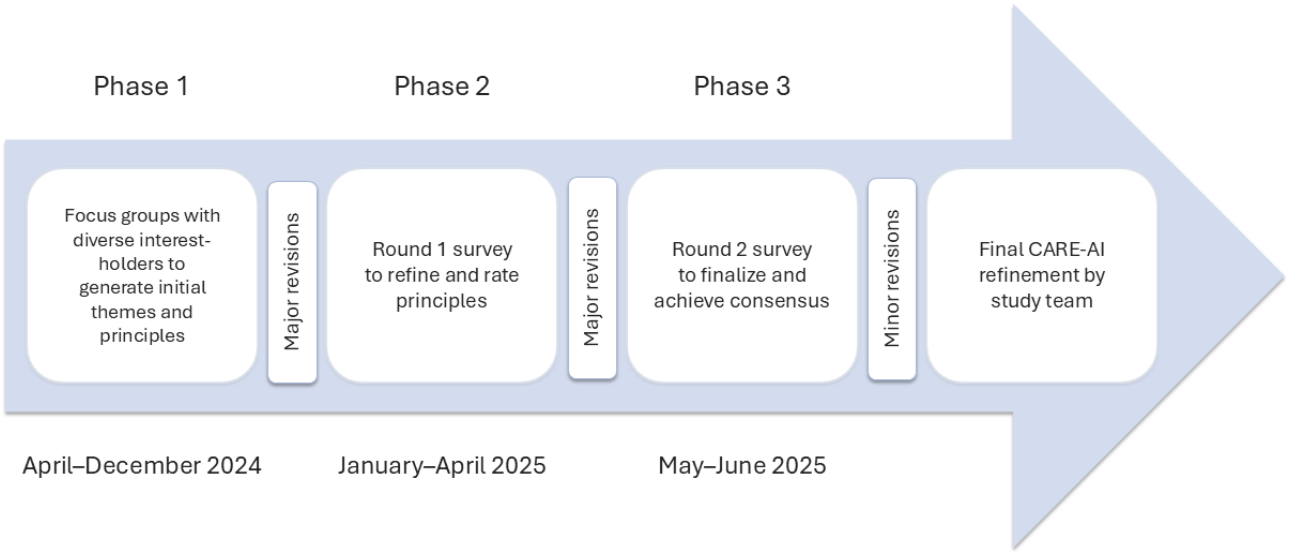
Methods

Design and Purpose

The Equity in Health Systems (EqHS) leadership team (LKS, JM, DW, VD, BT, BJ, MM, and BH) conducted a 3-phase, modified Delphi study to develop and validate principles for accountable and responsive AI ethics in health care and across the health professions learning continuum (Figure 1). A modified Delphi was selected to support structured, iterative consensus where definitive empirical standards are limited and expert judgment is warranted [13,14]. Reporting follows the Accurate Consensus Reporting Document guideline for consensus methods [15].

XSL•FO
RenderX

Figure 1. Project phases and timeline. CARE-AI: Contextual, Accountable, Responsible, and Equitable Artificial Intelligence.



Participants and Recruitment

Purposive sampling sought breadth across roles and perspectives. To ensure epistemic diversity, recruitment was organized around three predefined domains of expertise: (1) AI and technology, (2) HPE, and (3) ethics and professionalism. Participants included clinicians, educators, and learners, as well as patient partners and caregivers, who contributed lived-expertise perspectives alongside their additional expertise in AI, technology, or ethics; members of professionalism offices and institutional bodies, who had experience overseeing professional conduct and ethics; and health equity experts, who specialized in health care optimization and addressing systemic disparities in health systems. Eligibility required current or recent engagement with AI in health or education and willingness to participate in one or more phases. Invitations were sent via email by the research team, initially distributed through EqHS, the AAMC Group on Information Resources, the International Association for Health Professions Education Technology Enhanced Learning community, the Temerty Centre for Artificial Intelligence Research and Education in Medicine, and the Canadian Association for Medical Education technology working group, with snowballing encouraged [16].

Phase 1: Building the Foundation

The first phase began from a position of listening. Rather than drafting principles in isolation, the team drew on established social media professionalism guidelines and invited critique and reimagining in the context of AI [17]. Two professional meetings became early testing grounds for this work: the International Summit on Leadership Education for Physicians (2024; theme: *ethical integration of AI in health education*) and the Academy for Professionalism in Health Care (2024; theme: *enhancing professionalism through effective communication: from in-person to AI*).

Phase 1 used an exploratory qualitative approach to inform early principle development. Semistructured group discussions and focus groups with leaders across AI and technology, HPE, and ethics and professionalism were used to surface shared concerns, practical challenges, and gaps in existing guidance related to AI use. Discussions were oriented toward identifying areas of convergence across roles and contexts, with particular attention to where existing digital professionalism and social media guidelines failed to address AI-specific ethical issues. Analytic emphasis was placed on pragmatic, inductive pattern recognition and sense-making sufficient to identify areas of convergence, practical concerns, and implications for subsequent principle development and consensus testing rather than on formal thematic analysis [18].

Building on these insights, 3 focus groups, aligned with the domains of AI and technology, HPE, and ethics and professionalism, were convened. Using semistructured prompts, facilitators encouraged participants to identify priorities and surface risks and describe practical contexts in which the principles would need to be applied. Sessions were recorded and supported by detailed notes.

The study team reviewed recordings and notes using a pragmatic, inductive approach consistent with a constructivist descriptive framework, emphasizing close engagement with participant perspectives and the integration of lived and professional expertise. Analysis focused on identifying recurring concepts, areas of convergence, and practical challenges across groups. These insights were synthesized through iterative team discussion, where overlapping ideas were consolidated and translated into draft principle statements. The principles were refined for clarity and applicability across contexts, resulting in an initial set of draft principles that were carried forward into the modified Delphi process.

Instrument Development and Piloting

An online survey was built in Research Electronic Data Capture (REDCap; Vanderbilt University), an electronic data capture platform widely used in health research (Multimedia Appendix 1) [19,20]. Items asked participants to rate each draft principle on appropriateness and actionability and to provide free-text comments on clarity, scope, feasibility, and equity or relational considerations. The instrument was refined through cognitive interviewing with coauthors and a small group of collaborators to optimize wording, anchors, and flow prior to fielding [21].

XSL•FO
RenderX

Phase 2: Round 1 Survey

Round 1 opened with a brief preamble describing the aims and intended use of the framework, followed by item-level ratings and comment boxes for each principle. Participants rated importance on a 5-point Likert scale anchored from unimportant to extremely important and could identify gaps, justify ratings, and suggest wording refinements. A priori consensus was defined as follows: include if at least 70% of respondents rated 4 to 5 and fewer than 15% rated 1 to 2, exclude if at least 70% rated 1 to 2, and no consensus flagged for modification and rerating [13,14]. Free-text comments were reviewed descriptively and mapped to each principle. Even when inclusion thresholds were met, qualitative input guided wording changes to strengthen clarity and accountability.

Phase 3: Round 2 Survey

Round 2 presented the revised preamble and principles for focused judgment with 3 response options: include, exclude, or undecided. Items with at least 70% "include" were retained and items with at least 70% "exclude" were removed. Panelists received an anonymized summary of group feedback and revisions between rounds. Where participant feedback improved clarity or feasibility without altering intent, wording was adjusted.

Data Integration and Rigor

While quantitative thresholds determined provisional inclusion or exclusion, qualitative commentary was systematically integrated to sharpen scope, specify accountability, and strengthen equity and relational dimensions. Rigor was supported by *a priori* definitions of objectives, panel composition and recruitment strategy, scoring approach and consensus rules, number of rounds, and stopping criteria; controlled feedback between rounds; investigator review of borderline items; and documentation of wording changes across iterations, consistent with Accurate Consensus Reporting Document recommendations for transparent reporting of consensus methods [15].

Ethical Considerations

This study was reviewed and approved by the University of Alberta Research Ethics Board, Edmonton, Canada (level 2, study ID Pro00143413). Participation was voluntary and anonymous through survey rounds. Participants did not receive compensation for their involvement in the study.

Results

Phase 1: Building the Foundation

The CARE-AI principles were informed by social media professionalism guidelines [17]. Initial feedback from the two professional meetings made it clear that AI was not a continuation of social media best practices. These sessions engaged 24 participants, whose input guided the need for ethical, privacy, and bias considerations to be central to any AI framework. Three subsequent focus groups (n=37: AI, n=12; HPE, n=14; and ethics, n=11) broadened this base of understanding. Participants included clinicians, educators, administrators, ethicists, AI specialists, and patient and caregiver partners predominantly from Canada and the United States. Collectively, they generated 10 draft principle statements and provided extensive feedback on the framing preamble.

Feedback from these meetings underscored both the urgency of addressing AI's ethical challenges and the need for a framework that could speak across professions, training levels, and contexts. The early preamble emphasized individual judgment and institutional values, positioning CARE-AI as "professionalism principles with a focus on ethical, privacy, and bias challenges posed by AI technologies." By the end of phase 1, participants had urged a more expansive framing, one that clarified what "AI" means, recognized diverse roles, and emphasized collective responsibility. These suggestions set the stage for considerable refinement in later phases.

Phase 2: Delphi Round 1

A total of 242 participants initiated the round 1 survey, with 120 completing it (completion rate: 49.6%). Participants represented a wide spectrum of roles, including patient partners (Table 1). Additionally, participants were able to select multiple roles. Geographically, most were from Canada (68/120, 56.7%) and the United States (39/120, 32.5%), with smaller representation from Europe, Asia, Australia, the Caribbean, and Africa.

Of the 61 statements rated, 58 (95%) surpassed the prespecified consensus threshold and none required removal. Participants contributed 1887 comments, 697 of which were content-rich. These qualitative data prompted significant revisions: clarifying accountability language, embedding equity more explicitly, and adjusting tone to ensure accessibility. Accountability-related comments also highlighted expectations for responsible data use and stewardship. Each line item was modified based on the feedback provided. By the close of round 1, 89% (110/124) of respondents agreed or strongly agreed that the draft principles effectively outlined professionalism expectations for AI.

XSL•FO
RenderX

Table . Distribution of participant roles among respondents in Delphi rounds 1 and 2[a].

Role	Phase 2 participants (round 1), n (%)	Phase 3 participants (round 2), n (%)
AI[b] specialist	104 (43.7)	41 (54.7)
Clinician	107 (45)	41 (54.7)
Education leader	117 (49.2)	44 (58.7)
Education technology specialist	96 (40.3)	32 (42.7)
Health equity expert	58 (24.4)	24 (32)
Health professions educator	153 (64.3)	53 (70.7)
Patient or caregiver partner	49 (20.6)	19 (25.3)
Professionalism office	50 (21)	16 (21.3)
Researcher	121 (50.8)	45 (60)

[a]Roles were multiselect, so percentages exceed 100%.

[b]AI: artificial intelligence.

Phase 3: Delphi Round 2

In round 2, of the 103 participants invited, 78 initiated the survey, and 75 completed it (completion rate: 75/103, 72.8% of those invited, 75/78, 96.2% of starters). The panel reflected a similar distribution to phase 2. Participants were primarily from Canada (47/75, 62.7%) and the United States (22/75, 29.3%), with additional representation from the United Kingdom, Europe, Asia, Australia, and the Caribbean. All principles presented in round 2 achieved inclusion consensus (≥70% agreement). Participants submitted 224 comments, of which 179 were substantive, guiding final refinements. The preamble was expanded to include what we mean by AI, the need for iterative self-reflection, and the interconnectedness of AI.

Evolution of a Principle

The progression of principle 1 illustrates how participant input shaped the framework (Table 2). In the earliest draft developed after the conferences, the principle was expressed as “AI is a personal responsibility,” highlighting judicious use and the need for critical human review of AI outputs. By the end of phase 1, following focus group feedback, this was reframed as a matter of “professional duty,” urging individuals to handle information carefully, recognize how AI could perpetuate bias, and act as ethical stewards.

XSL•FO
RenderX

Table . Progression of principle 1 across study phases.

Study phase	Principle 1
Phase 1 (after conferences)	Curation of information is a personal responsibility. Handle information with care, especially when using AI[a] to train, generate, or process data. Understand that AI systems can unintentionally perpetuate and amplify existing biases in data. Apply this knowledge to guide your use of AI, ensuring that it promotes fairness and equity in health care settings.
Phase 1 (after focus groups)	Responsible AI is a professional duty. As a professional, it is your individual responsibility to maintain knowledge and ethical use of AI systems. Your role may span institutions and contexts; you are expected to align with their organizational policies related to AI. You remain an advocate for evaluating, selecting, and implementing appropriate AI tools. You are also responsible for refining their performance to align with the values of health care and accountability to patients, team members, and health care systems.
Phase 2 (after Delphi round 1)	Responsible AI use is both an individual and collective duty. Individuals, teams, and institutions contribute to the safe and equitable use of AI in health settings by establishing expectations, developing supportive policies, and regularly reviewing tools to ensure alignment with health care values. As professional contributions often span institutions and contexts, expectations and standards may vary. Stay informed about guidelines, apply best practices in AI use, and help refine policies as new use cases and concerns emerge. AI integration must remain grounded in a shared commitment to safe, responsible use, centered on patients[b], learners, care partners, and the broader health community.
Phase 2 (after Delphi round 2)	Responsible AI use is both an individual and collective duty. Individuals, teams or groups, and institutions contribute to the safe and equitable use of AI in education, research, and health care settings by establishing expectations, developing supportive policies, and regularly reviewing resources to ensure alignment with health care values. Stay informed about guidelines as they may vary between institutions and contexts. AI integration must be grounded in a shared commitment to safe and responsible use, centered on patients[b], learners, and the broader health community.

[a]AI: artificial intelligence.

[b]We use the term “patients” inclusively to encompass individuals receiving care as well as their caregivers and others directly involved in supporting their health and well-being.

In phase 2, participants stressed that responsibility extended beyond individual behavior to include institutional alignment, policy awareness, and accountability to patients, teams, and systems, shifting the wording to underscore shared responsibility. The final version emphasized that individuals, teams, and institutions collectively contribute to the safe and equitable use of AI in education, research, and health care, and that accountability must remain grounded in a shared commitment to patients, learners, and the broader health community.

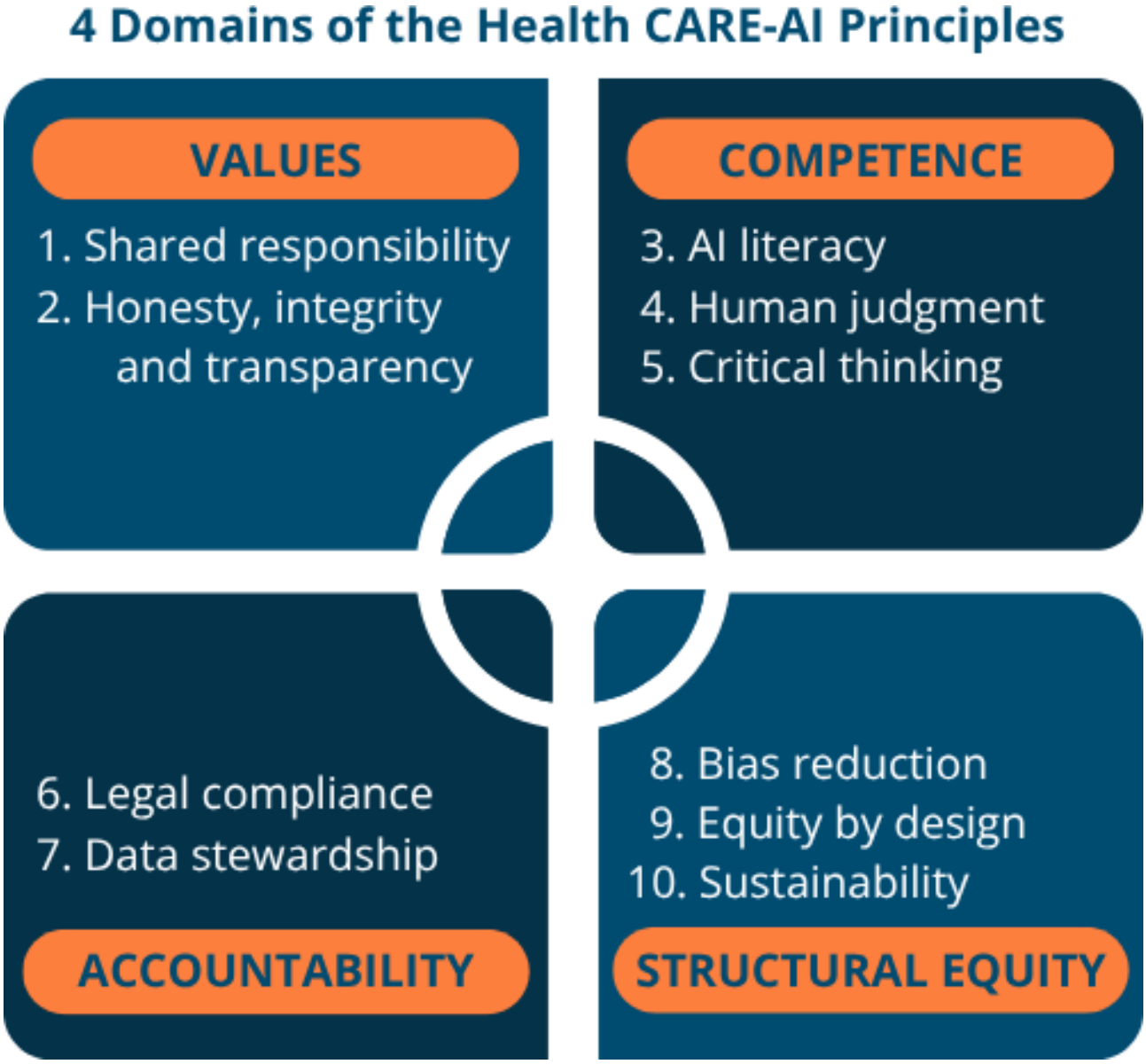
This trajectory demonstrates how most substantive shifts occurred between phases 1 and 2, when participants pressed for clearer professional and institutional accountability. By phase 3, refinements were more incremental, polishing language for clarity and inclusiveness.

Final Framework

Endorsement for the final framework was near unanimous: 96% (72/75) of participants agreed or strongly agreed that it clearly articulated the professionalism expectations required for AI to effectively address the educational, technological, and ethical needs within the health professions (agree: 43/75, 57.3%; strongly agree: 29/75, 38.7%). The Health CARE-AI Framework, shown in Figure 2, brings together a preamble and 10 guiding principles that move deliberately from foundational values to structural equity. Across all phases, participant feedback sharpened the language of both the preamble and the principles, reinforcing clarity, strengthening accountability, and embedding equity. The framework is organized into 4 interdependent domains (values, competence, accountability, and structural equity) that chart a trajectory from individual behaviors to systemic change. Figure 2 presents the Health CARE-AI principles as 4 interrelated domains rather than a sequential process; visual proximity reflects conceptual relationships rather than hierarchy or order of implementation. The full framework is found in Multimedia Appendix 1, with the high-level summary presented in Textbox 1.

XSL•FO
RenderX

Figure 2. Conceptual model of the Health Contextual, Accountable, and Responsible Ethics for Artificial Intelligence (CARE-AI) principles. AI: artificial intelligence.



XSL•FO
RenderX

Textbox 1. A high-level summary of the Health Contextual, Accountable, and Responsible Ethics for Artificial Intelligence Framework principles.

1. Responsible artificial intelligence (AI) use is both an individual and collective duty. This clarifies responsibilities at individual, team, and institutional levels; sets expectations for reviewing tools; and keeps patients and learners at the center.
2. Use AI with honesty and integrity. Be transparent about AI assistance, avoid deception or manipulation, and disclose use in learning and care in line with policy.
3. Build and maintain AI literacy. Commit to ongoing, role-appropriate learning, resource training accordingly, and feed experience back to improve local policies and practice.
4. Responsible AI use should complement, not replace, human judgment. Retain professional accountability and verify AI outputs against clinical and educational judgment.
5. Think critically before speaking, acting, or uploading with an AI system present. Treat AI as a present third party, anticipate persistence and misinterpretation, and choose words and uploads that protect relationships and trust.
6. Work within the law. Follow privacy, consent, and intellectual property requirements, and do not delegate legally reserved tasks to AI.
7. Use and share information ethically in AI-supported environments. Understand how data are collected, stored, and shared; obtain appropriate consent; and safeguard privacy and data sovereignty.
8. Use AI in ways that actively reduce bias and promote equity. Check outputs for bias, report concerns, support auditing and updates, and adapt use as models and data change.
9. Build equity into AI foundations. Embed equity in design and governance, ensure accessibility, use diverse data, and co-design with affected communities throughout development and evaluation.
10. Advance sustainable AI in health systems. Use AI in ways that are proportionate to benefit; mindful of impacts on patient care, the workforce, and the environment; and that strengthen rather than erode people and systems.

Discussion

Principal Findings

This study used a multiphase consensus process to develop and validate the Health CARE-AI Framework to bridge the gap between traditional health care ethics and the unique challenges posed by AI. Across conferences, focus groups, and modified Delphi rounds, participants refined a preamble and 10 principles that move from foundational values to structural commitments to equity. All 10 principles reached inclusion in round 2, with strong endorsement (72/75, 96% agreed or strongly agreed). Qualitative comments sharpened accountability language, strengthened equity, and improved clarity for education and clinical use.

The Health CARE-AI Framework is designed for integration into education and governance. Within curricula, educators can map existing professionalism and digital health teaching against the 10 principles. At the institutional level, the principles can guide policy language, supervision protocols, and oversight of AI initiatives. Barriers to implementation exist and include uneven faculty readiness, high workload, and policy lag. An integration toolbox is likely needed for success.

The framework aligns with major reports while addressing practice-level gaps. Relative to BEME Guide No. 84, which maps the evolving integration of AI in medical education, our work moves beyond descriptive synthesis toward actionable expectations for educators and learners across classroom and clinical contexts [2]. It similarly operationalizes AAMC principles by translating high-level commitments such as transparency, equity, and human-centered use into role-specific behaviors that can be explicitly taught, supervised, and assessed within training and practice environments [9]. In parallel, the framework renders foundational values articulated by UNESCO, OECD, and WHO, including human rights, accountability, robustness, and fairness, into concrete responsibilities that identify who is accountable, in which contexts, and under what conditions of oversight [5-7]. In doing so, CARE-AI bridges the gap between normative guidance and day-to-day professional practice, offering a consensus-based framework that supports implementation across diverse roles, settings, and stages of the health professions continuum.

Participants consistently emphasized 3 challenges: accountability, data stewardship, and equity. Principle 1 establishes responsibility at both individual and collective levels, while privacy and stewardship expectations are reinforced through commitments to ethical data use, transparency, and alignment with law and policy (principles 2, 6, and 7). Principles 8 and 9 address equity in complementary ways: users are called to mitigate bias in practice, and institutions are charged with embedding equity into system design. Some questioned whether principle 9 was necessary alongside principle 8, but consensus affirmed its importance for structural change rather than reliance on vigilant end users.

Participants raised environmental concerns, primarily related to the energy use and carbon footprint of large-scale AI models and infrastructure. This resulted in the development of principle 10 on sustainability, inclusive of workforce sustainability. Additionally, stewardship is embedded in shared responsibility (principle 1), lawful and ethical governance (principles 6 and 7), and equity by design (principle 9), with institutions considering sustainability in procurement and oversight.

Participants also emphasized how the principles required greater pregrounding. In response, we expanded the preamble to clarify scope, define AI, and underscore collective accountability. The preamble now explicitly addresses the environments in which the principles will be applied; recognizes the diverse interactions of patients, care partners, clinicians, educators, learners, researchers, developers, and administrators; and addresses the

XSL•FO
RenderX

need for inclusive dialogue, transparency, collaboration, and co-design in the implementation process.

Implications for Implementation and Practice

We offer 2 practical starting points for implementation. First, embedding the principles into existing professionalism expectations in HPE ensures learners encounter them early and repeatedly. Second, incorporating them into program and institutional governance provides a common standard for committees overseeing AI initiatives such as clinical decision support, admissions processes, and educational technologies.

To help educators and organizations embed these principles into contexts, we have created a practical implementation guide and toolkit, which includes 4 illustrative scenarios for use in teaching and governance, and targeted faculty development sessions [22]. This planning resource supports leaders in health professions to proactively assess their current state, engage their teams, and implement values-aligned AI adoption strategies. Without such support, integration risks remaining aspirational. Validation of this implementation guide and toolkit is completed within the contexts of HPE, research, and clinical care. Engagement with professional organizations (eg, medical education, regulatory, and accreditation bodies) will be important to support broader dissemination, alignment with standards, and update of the framework. Future work should explore formal endorsement and collaboration with these organizations as a pathway to implementation at scale.

Reflexivity and Limitations

While substantial qualitative feedback was collected across phases (eg, free-text survey comments and focus group discussions), we did not undertake a formal thematic analysis. Instead, qualitative data were synthesized using a pragmatic, inductive approach to inform iterative refinement of principles, with emphasis on clarity, accountability, and equity. As a result, the study does not provide a detailed qualitative account of participant perspectives or divergent viewpoints. This may limit insight into how different interest-holder groups interpreted or prioritized key issues and should be considered when interpreting the findings.

Most participants were from English-speaking, high-resource settings. Regions outside North America and contributors from remote locations were underrepresented. The study team was largely academic, though it included learners, clinical leaders, frontline clinicians, and researchers, all from different career stages, which strengthened reflexivity and interpretive breadth. Additionally, participants represented diverse roles (eg, clinicians, educators, AI specialists, ethicists, and patient partners), which strengthen the study but may also have introduced role-related biases. However, it is worth noting that representation of patient and caregiver partners had lower representation and therefore may influence the perspectives reflected in the final framework. Those in ethics or equity may have emphasized risks and safeguards, while those in AI-focused roles may have prioritized feasibility and implementation. Although the Delphi process aimed to balance these perspectives, role-based priorities may have influenced the framing and weighting of principles, resulting in a negotiated consensus across groups rather than uniform perspectives.

As with all consensus-based approaches, panel composition shapes the perspectives represented, and iterative rating processes may attenuate minority or dissenting views. Apparent agreement may therefore reflect convergence on least-objectionable wording rather than uniform strength of endorsement across participant groups. Findings should be interpreted as consensus among those who participated at this time and revisited as technologies, professional norms, and implementation contexts evolve.

Methodological and practical constraints also warrant consideration. Selection and attrition bias are possible, as approximately half of the participants who initiated round 1 completed the survey; reasons for noncompletion were not formally assessed. To preserve clarity and durability of the framework, suggestions to incorporate extensive cross-references or illustrative examples were intentionally deferred, with applied scenarios provided separately in the implementation toolkit. Institutional and policy constraints may also limit frontline agency, reinforcing participant emphasis on the need for clearly articulated responsibilities and accountability across individual, organizational, and system levels during implementation.

Conclusions

This study developed and validated the Health CARE-AI Framework through a multiphase consensus process shaped by diverse communities of practice. The final framework, comprising a preamble and 10 principles, received strong endorsement and reflects careful refinement across phases. Taken together, the principles emphasize equity and professional accountability as core considerations for AI use across the health professions, progressing from values to competence and toward structural commitments to equity.

The framework addresses professionalism expectations for AI use across the health professions in practical ways. Educators can anchor curricula and faculty development; learners can guide day-to-day decisions in study and supervised practice; and programs and health systems can align policies, governance, and quality improvement to safeguard patient welfare, uphold professional integrity, and nurture public trust.

Building on this work, we have developed a companion implementation guide and toolkit with scenario-based applications across education, research, and clinical contexts, which has undergone high-level validation prior to rollout. We invite programs to integrate the framework, evaluate its use, and share outcomes. Future work should examine uptake, fidelity, and impacts on equity and trust, and update the framework as technologies and contexts evolve.

XSL•FO
RenderX

Acknowledgments

The authors would like to thank those participants who gave freely of their time to this project and to Janet Corral, senior scientist at Equity in Health Systems (EqHS), for her conceptualization of this work and her assistance with data analysis. The authors have previously published a case-based educational resource designed to support facilitated discussion of ethical and professional considerations related to artificial intelligence (AI) in health professions education. The authors own full copyright of that material. That resource presents illustrative cases and practical guidance for use in educational settings but does not report empirical data, consensus methods, or validation findings. Preliminary versions of the CARE-AI principles and related cases have also been shared through conference presentations and open access educational materials. This paper reports the first peer-reviewed empirical development and validation of the Health CARE-AI Framework through a multiphase consensus study [22-26]. A generative AI tool (GPT-5.5 Pro) was used in a limited, supportive role during manuscript preparation to assist with language refinement and structural clarity. All substantive content, analysis, interpretation of findings, and conclusions were generated by the authors. The authors take full responsibility for the accuracy, integrity, and originality of the work.

Funding

This research was funded by the government of Canada under the Future Skills program for the project named Fostering Equity and Inclusion in Healthcare by Accelerating Responsible Use of Artificial Intelligence. (Le projet Promouvoir l'équité et l'inclusion dans les soins de santé en accélérant l'utilisation responsable de l'intelligence artificielle est financé par le gouvernement du Canada dans le cadre du programme Compétences futures.) This funding supported the broader Contextual, Accountable, Responsible, and Equitable Artificial Intelligence (CARE-AI) program of work, including participant engagement, coordination, and knowledge mobilization activities, but had no role in the design of this study, data collection, analysis, interpretation, or manuscript preparation.

Authors' Contributions

LKS, DW, MM, BT, VD, BH, and JM conceptualized the project. All authors contributed to data curation and analysis. DW acquired funding for the project. BJ handled project administration. LKS wrote the original draft. All authors approved the final manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Full framework of the Health CARE-AI (Contextual, Accountable, Responsible, and Equitable Artificial Intelligence) principles.
[PDF File, 562 KB - mededu_v12i1e91626_app1.pdf]

References

1. Ethics and governance of artificial intelligence for health: WHO guidance. World Health Organization. 2021. URL: https://www.who.int/publications/i/item/9789240029200 [accessed 2026-06-29]
2. Gordon M, Daniel M, Ajiboye A, et al. A scoping review of artificial intelligence in medical education: BEME guide no. 84. Med Teach 2024 Apr;46(4):446-470. [doi: 10.1080/0142159X.2024.2314198] [Medline: 38423127]
3. Smith H, Downer J, Ives J. Clinicians and AI use: where is the professional guidance? J Med Ethics 2024 Jun 21;50(7):437-441. [doi: 10.1136/jme-2022-108831] [Medline: 37607805]
4. Boscardin CK, Gin B, Golde PB, Hauer KE. ChatGPT and generative artificial intelligence for medical education: potential impact and opportunity. Acad Med 2024 Jan 1;99(1):22-27. [doi: 10.1097/ACM.0000000000005439] [Medline: 37651677]
5. Recommendation on the ethics of artificial intelligence. United Nations Educational, Scientific and Cultural Organization. 2021. URL: https://unesdoc.unesco.org/ark:/48223/pf0000380455 [accessed 2026-06-29]
6. Recommendation of the council on artificial intelligence. Organisation for Economic Co-operation and Development. 2019. URL: https://legalinstruments.oecd.org/en/instruments/oecd-legal-0449 [accessed 2025-09-23]
7. Regulatory considerations on artificial intelligence for health. World Health Organization. 2023. URL: https://www.who.int/publications/i/item/9789240078871 [accessed 2025-09-23]
8. Augmented intelligence in health care H-480.940. American Medical Association. URL: https://policysearch.ama-assn.org/policyfinder/detail/H-480.940?uri=%2FAMADoc%2FHOD.xml-H-480.940.xml [accessed 2025-09-23]
9. Principles for the responsible use of artificial intelligence in and for medical education. Association of American Medical Colleges. URL: https://www.aamc.org/about-us/mission-areas/medical-education/principles-ai-use [accessed 2025-09-23]
10. Principles for responsible AI in medical school and residency selection. Association of American Medical Colleges. 2025. URL: https://www.aamc.org/media/82971/download [accessed 2025-09-23]
11. Farmakidis AL, Singh A, Leaf K, Rossi A. Artificial intelligence curricula in U.S. and Canadian medical schools. : Association of American Medical Colleges; 2025 URL: https://www.aamc.org/media/84666/download [accessed 2026-06-29]

XSL•FO
RenderX

12. Pupic N, Ghaffari-Zadeh A, Hu R, et al. An evidence-based approach to artificial intelligence education for medical students: a systematic review. PLOS Digit Health 2023 Nov;2(11):e0000255. [doi: 10.1371/journal.pdig.0000255] [Medline: 38011214]
13. Sinha IP, Smyth RL, Williamson PR. Using the Delphi technique to determine which outcomes to measure in clinical trials: recommendations for the future based on a systematic review of existing studies. PLoS Med 2011 Jan 25;8(1):e1000393. [doi: 10.1371/journal.pmed.1000393] [Medline: 21283604]
14. Humphrey-Murto S, Wood TJ, Gonsalves C, Mascioli K, Varpio L. The Delphi method. Acad Med 2020 Jan;95(1):168. [doi: 10.1097/ACM.0000000000002887] [Medline: 31335812]
15. Gattrell WT, Logullo P, van Zuuren EJ, et al. ACCORD (ACcurate COnsensus Reporting Document): a reporting guideline for consensus methods in biomedicine developed via a modified Delphi. PLoS Med 2024 Jan;21(1):e1004326. [doi: 10.1371/journal.pmed.1004326] [Medline: 38261576]
16. Johnson TP. Snowball sampling: introduction. In: Balakrishnan N, Colton T, Everitt B, Piegorsch WW, Ruggeri F, Teugels JL, editors. Wiley StatsRef: Statistics Reference Online: John Wiley & Sons; 2014. [doi: 10.1002/9781118445112]
17. Ellaway R, Tworek J. Digital professionalism. Association of American Medical Colleges. URL: https://www.aamc.org/professional-development/affinity-groups/gir/viewpoint-digital-professionalism [accessed 2025-09-23]
18. Patton MQ. Qualitative Research & Evaluation Methods, 4th edition: SAGE Publications; 2015. URL: https://ia800500.us.archive.org/30/items/michael-quinn-patton-qualitative-research-evaluation-methods-integrating-theory-/Michael%20Quinn%20Patton%20-%20Qualitative%20Research%20&%20Evaluation%20Methods%20Integrating%20Theory%20and%20Practice-Sage%20Publications,%20Inc%20(2014).pdf [accessed 2026-06-29]
19. Harris PA, Taylor R, Minor BL, et al. The REDCap consortium: building an international community of software platform partners. J Biomed Inform 2019 Jul;95:103208. [doi: 10.1016/j.jbi.2019.103208] [Medline: 31078660]
20. Harris PA, Taylor R, Thielke R, Payne J, Gonzalez N, Conde JG. Research electronic data capture (REDCap)--a metadata-driven methodology and workflow process for providing translational research informatics support. J Biomed Inform 2009 Apr;42(2):377-381. [doi: 10.1016/j.jbi.2008.08.010] [Medline: 18929686]
21. Drennan J. Cognitive interviewing: verbal data in the design and pretesting of questionnaires. J Adv Nurs 2003 Apr;42(1):57-63. [doi: 10.1046/j.1365-2648.2003.02579.x] [Medline: 12641812]
22. CARE-AI implementation guide and toolkit 4.0. Equity in Health Systems. URL: https://EqHS.ca/CARE-AI [accessed 2026-07-03]
23. Corral JK, Mamdani M, Maniate JM, Sonnenberg LK. Guiding with Care-AI: A Practical Framework for Facilitating AI Conversations in Health Professions Education: Equity in Health Systems Publishing; 2025.
24. Sonnenberg LK, Corral J, Maniate J, Mamdani M. The CARE-AI study: creating accountable and responsible ethics for artificial intelligence for healthcare [Webinar]. Presented at: Artificial Intelligence (AI) in Academic Medicine Webinar Series; Apr 16, 2025 URL: https://www.aamc.org/media/82776/download?attachment [accessed 2026-06-30]
25. Sonnenberg L, Maniate J, Fleming S, Corral J, Wiljer D, Tang B. Unlocking AI's potential: harnessing AI responsibly to promote health for all. Presented at: World Federation for Medical Education World Conference 2025 Proceedings; May 25-28, 2025 URL: https://asmepublications.onlinelibrary.wiley.com/doi/10.1111/medu.70063 [accessed 2026-07-07]
26. Sonnenberg LK, Corral J, Maniate J, Mamdani M, Wiljer D, Haroon B. CARE-AI symposium and think shop with EqHS. EqHS webinar. 2025 Jun 9. URL: https://simulationcanada.ca/care-ai-series-symposium-think-shop-eqhs/ [accessed 2026-07-07]

Abbreviations

AAMC: Association of American Medical Colleges
AI: artificial intelligence
CARE-AI: Contextual, Accountable, and Responsible Ethics for Artificial Intelligence
EqHS: Equity in Health Systems
HPE: health professions education
OECD: Organisation for Economic Co-operation and Development
REDCap: Research Electronic Data Capture
UNESCO: United Nations Educational, Scientific, and Cultural Organization
WHO: World Health Organization

XSL•FO
RenderX

Edited by A Stone; submitted 17.Jan.2026; peer-reviewed by A Arora, D Verran; revised version received 05.Jun.2026; accepted 17.Jun.2026; published 22.Jul.2026.

Please cite as:
Sonnenberg LK, Wiljer D, Mamdani M, Do V, Tang B, Haroon B, Jaffer B, Maniate J
Principles for Responsible AI in Health Professions Education, Research, and Care: Health CARE-AI (Contextual, Accountable, Responsible, and Equitable Artificial Intelligence) Framework Delphi Consensus Study
JMIR Med Educ 2026;12:e91626
URL: https://mededu.jmir.org/2026/1/e91626
doi:10.2196/91626

© Lyn Kathryn Sonnenberg, David Wiljer, Muhammad Mamdani, Victor Do, Brandon Tang, Babar Haroon, Bizav Jaffer, Jerry Maniate. Originally published in JMIR Medical Education (https://mededu.jmir.org), 22.Jul.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Education, is properly cited. The complete bibliographic information, a link to the original publication on https://mededu.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Generative AI–Assisted Progressive-Disclosure Case-Based Learning for Clinical Reasoning in Occupational Medicine: Quasi-Experimental Study

Peng Su[1*], MD, PhD, Prof Dr; Min Hu[2*], PhD; Chengzhi Chen[1], PhD; Shangcheng Xu[1], PhD

[1]Department of Occupational and Environmental Health, School of Public Health, Chongqing Medical University, Chongqing, Chongqing, China

[2]School of Business, Chongqing College of Finance and Economics, Chongqing, Chongqing, China

[*]these authors contributed equally

Corresponding Author:
Peng Su, MD, PhD, Prof Dr
Department of Occupational and Environmental Health
School of Public Health
Chongqing Medical University
No.1, Yixueyuan Road, Yuzhong District
Chongqing, Chongqing, 400016
China
Phone: 86 13274909310
Fax: 86 02368485207
Email: 103335@cqmu.edu.cn

Abstract

Background: Case-based learning (CBL) promotes transfer of knowledge to practice, yet occupational health CBL must also develop exposure assessment and epidemiologic thinking. Static, single-session cases that disclose all information upfront can truncate iterative reasoning and encourage premature diagnostic closure. Generative AI (GenAI) can support the efficient development of high-fidelity, progressively disclosed cases, but hallucination risks require strict quality control.

Objective: This study aimed to develop and evaluate a multicomponent GenAI-assisted, 4-act progressive-disclosure CBL package for an occupational lead poisoning module, using a human-in-the-loop workflow to mitigate hallucination risk.

Methods: In a nonrandomized posttest controlled quasi-experimental study, 224 undergraduates were assigned by administrative class to an intervention group (n=114) or a control group (n=110). The control group received conventional static CBL; the intervention group received a GenAI-assisted progressive-disclosure CBL package. The primary outcome was the standardized individual case-analysis assignment score. Secondary outcomes were the delayed final examination score, 5 self-reported learning experience domains, and 3 video-derived behavioral engagement indicators. Subgroup analyses by academic major and group-by-major interaction tests were exploratory.

Results: Baseline characteristics were comparable between groups. The intervention group scored higher on case analysis (mean 88.41, SD 3.85 vs mean 82.51, SD 3.94; $P<.001$; Cohen $d=1.51$) and on the delayed final exam (mean 80.96, SD 8.47 vs mean 78.62, SD 6.63; $P=.02$; Cohen $d=0.31$). Significant improvements were observed in information gathering, hypothesis generation, and differential diagnosis (all $P<.001$ after Holm correction), while treatment/management planning did not differ ($P=.19$). The intervention group reported higher perceived difficulty, perceived improvement in clinical reasoning, engagement, transfer of self-efficacy, and evaluation of the course materials (all $P<.001$). Voluntary responses (mean 2.30, SD 1.11 vs mean 1.68, SD 0.92; $P=.008$), evidence-referencing statements (mean 1.90, SD 0.88 vs mean 1.45, SD 0.71; $P=.01$), and net group discussion time (mean 36.50, SD 6.88 vs mean 28.42, SD 7.02 minutes; $P<.001$) were all higher in the intervention group. The group-by-major interaction for the final examination was not statistically significant ($P=.41$). Audit logs showed that all AI-generated case drafts required expert correction.

Conclusions: A multicomponent GenAI-assisted, 4-act progressive-disclosure CBL package implemented with rigorous human-in-the-loop verification was associated with higher case-analysis performance, modestly higher delayed examination performance, and greater behavioral engagement in one nonrandomized cohort. Randomized and class-level multilevel designs with longer follow-up are needed to determine the durability and generalizability of these findings.

XSL•FO
RenderX

(JMIR Med Educ 2026;12:e103584) doi:10.2196/103584

KEYWORDS

generative AI; case-based learning; clinical reasoning; occupational medicine; medical education

Introduction

Case-based learning (CBL) has become a widely adopted pedagogical approach in higher medical education, valued for its capacity to bridge theoretical knowledge and authentic clinical practice and to cultivate clinical reasoning skills [1,2]. Within preventive medicine and occupational health curricula, occupational case discussion is a core instructional component that links classroom concepts to public health and clinical decision-making. Compared with case-based instruction in purely clinical disciplines, occupational health case teaching demands that learners simultaneously master clinical diagnostic competencies and develop capabilities in epidemiologic investigation and occupational or environmental exposure assessment. However, CBL in occupational health education is often delivered in a single-session, static format, in which comprehensive clinical and exposure-related information is disclosed upfront. Such synchronous information disclosure prematurely curtails the iterative reasoning cycle that characterizes authentic clinical practice, in which evidence is gathered progressively and hypotheses are repeatedly calibrated. When information is overly complete, learners may adopt an "omniscient" perspective and become susceptible to premature diagnostic closure. They may circumvent the deliberate practice of hypothesis generation, evidence seeking, and hypothesis testing and refinement, instead relying on keyword matching and knowledge recall to reach solutions [3]. This tendency to jump to conclusions reflects a limitation in analytical thinking and may, in turn, be reinforced by learners' drive for immediate certainty, encouraging reliance on mental shortcuts and avoidance of careful, deliberate reasoning [4]. Consequently, clinical reasoning may remain at the level of superficial information retrieval rather than deeper analytic inference, with insufficient training in identifying and prioritizing key information amid uncertainty and confounding cues. Traditional pedagogical models reliant on rote retrieval fall short in developing the higher-order clinical reasoning required in complex medical environments, underscoring the need to reform case-based pedagogy to enhance learners' critical thinking and reasoning skills.

Unfolding case studies (UCSs) use progressive disclosure to promote active reasoning under uncertainty, closely mirroring the incremental decision-making of real clinical practice [5]. Although widely adopted in nursing education, UCS application in clinical and preventive medicine remains limited [3]. Designing these dynamic cases—particularly segmenting information and managing missing data—is highly resource-intensive and demands significant clinical and pedagogical expertise from instructors [6]. To address these barriers, the present study aims to develop practical UCS models tailored to clinical and preventive medicine, seeking an optimal balance between instructional realism and operational efficiency.

Situated Learning Theory (SLT) emphasizes the construction of authentic or high-fidelity contexts to promote learners' cognitive processing and competence development within situated practice [7]. Although SLT has been incorporated into CBL to improve learning outcomes, several practical constraints remain [8]. From a cognitive load perspective, embedding plausible distractors into a case forces learners to stop passively absorbing information and start actively filtering evidence. This productive struggle generates germane cognitive load, which is essential for building robust clinical schemas [9]. However, in practical educational settings, medical contexts are inherently complex and uncertain, and teachers' scenario design may be constrained by experience and thus become overly directional [9]. To reduce organizational burden, scenarios may be simplified, preserving only the most salient cues, and standardized cases may not be readily adaptable to learners from different disciplinary backgrounds. Moreover, the design of high-quality "distractor cues" and the development of multiple scenario variants are time- and labor-intensive. Accordingly, how SLT can be effectively integrated into CBL while improving the fidelity of real-world simulation has become a key direction in contemporary educational reform [10].

In recent years, AI, particularly generative AI (GenAI), has advanced rapidly; large language models now demonstrate substantial capabilities in knowledge integration and natural language generation [11,12]. GenAI has been increasingly applied to medical education and has been used to support the development of CBL resources, such as knowledge restructuring, knowledge-graph generation, production of visual materials, and drafting of assessment items [13]. Often referred to as "Socratic AI," this approach uses targeted questioning to help students build clinical hypotheses and evaluate evidence, rather than simply handing them answers [14]. By guiding learners to articulate their reasoning step-by-step, it promotes much deeper cognitive engagement. Although these applications may reduce learners' cognitive load and improve memorization efficiency, their contribution to cultivating reasoning potential and higher-order cognitive training remains limited. Other studies have explored the use of GenAI for SLT-based scenario simulation [15]. As a scalable generative engine, GenAI can produce high-fidelity "cognitive friction" scenarios in bulk and dynamically introduce complex and personalized distractor cues aligned with instructional objectives, thereby shifting AI from a passive content producer toward an active facilitator of uncertainty and complexity. Furthermore, the cognitive friction introduced by GenAI acts as a "task-relevant resistance" that drives deeper learning [16]. Aligned with the concept of "desirable difficulties," this friction increases short-term effort but significantly improves long-term retention and clinical transfer. This productive friction inherently forces learners to abandon heuristics and systematically navigate diagnostic uncertainty. Nevertheless, GenAI is subject to hallucination, producing outputs that appear plausible but are factually inaccurate or fabricated. In the medical domain, erroneous

XSL•FO
RenderX

contextual information can mislead learners and undermine appropriate reasoning development. The primary operational challenge is integrating GenAI efficiently while maintaining rigorous guardrails for clinical accuracy and educational safety.

To address the limitations of existing case-based pedagogy, we propose a multicomponent GenAI-assisted progressive-disclosure CBL package, a scripted, 4-act CBL model aligned with the epistemic structure of occupational medicine. This approach uses an "act" structure to simulate complex clinical contexts while explicitly incorporating occupational and environmental exposure attribution into the scenario trajectory. Through progressive disclosure, clinical information is released in stages and supplemented with validated distractor cues to emulate real-world diagnostic resistance and to prompt repeated cycles of hypothesis generation and evidence verification. In addition, a human-in-the-loop (HITL) mechanism is implemented, combining prompt engineering with expert review to ensure rigorous quality control of AI-generated content and to mitigate hallucination risk. Using occupational lead poisoning as a paradigmatic module, our quasi-experimental findings suggest that the GenAI-assisted progressive-disclosure CBL was associated with improvements in learners' clinical reasoning and classroom engagement and with higher scores on selected reasoning processes theoretically related to premature diagnostic closure, thereby providing empirical support for a safety-focused pathway for pedagogical innovation.

Methods

Study Design and Participant Recruitment

This quasi-experimental study adopted a nonrandomized posttest design and was reported in accordance with the Transparent Reporting of Evaluations with Nonrandomized Designs (TREND) statement. The study was conducted at Chongqing Medical University during the 2024-2025 academic year and was approved by the Ethics Committee of Chongqing Medical University on September 14, 2024 (approval number: 2024-143). The study was conducted in accordance with the Declaration of Helsinki. All participants were informed of the study purpose, the voluntary nature of participation, and the confidentiality procedures and provided written informed consent prior to enrollment.

Participants were undergraduate students enrolled in the compulsory "occupational lead poisoning" teaching module. A total of 224 students from 3 administrative classes (Class of 2022) were included: clinical medicine (n=110), psychiatry (n=56), and medical laboratory science (n=58). Because curricular allocation followed the administrative class structure, allocation was implemented at the class and teaching-session level rather than at the individual level. Students in each intact teaching session were assigned to the same study arm, resulting in an intervention group (n=114; clinical medicine 56, psychiatry 28, and medical laboratory science 30) and a control group (n=110; clinical medicine 54, psychiatry 28, and medical laboratory science 28). Because each public health laboratory classroom accommodated a maximum of approximately 36 students, the clinical medicine class was taught in 2 parallel teaching sessions and the psychiatry and medical laboratory science classes in one session each within each study arm, yielding 8 teaching sessions in total (4 per arm). After entering the classroom, students were randomly organized into small groups, with a maximum of 6 students per group, for classroom discussion and group tasks. Baseline equivalence between groups was assessed for age, sex distribution, and grade point average (GPA) in the prerequisite course. The inclusion criteria were as follows: (1) undergraduate students in clinical medicine, psychiatry, or medical laboratory science; (2) completion of prerequisite theoretical teaching and full attendance in the case-based teaching sessions; and (3) voluntary participation with written informed consent. The exclusion criteria were as follows: (1) an absence rate >20%; (2) failure to complete key outcome assessments; and (3) missing or clearly implausible data that were unsuitable for analysis.

Because the module was compulsory, all scheduled students were invited to participate. No participant was excluded, and all participants completed the intervention and outcome assessments. Thus, the analytic sample was identical to the recruited sample (intervention: n=114; control: n=110).

To ensure adequate statistical power and to mitigate selection bias commonly encountered in nonrandomized educational studies, an a priori power analysis was conducted using G*Power (version 3.1; Heinrich-Heine-Universität Düsseldorf). Assuming a 2-tailed independent-samples t test with an α of .05, 80% power, and a moderate effect size (Cohen d=0.5), the minimum required sample size was 128. The final sample of 224 exceeded this requirement, providing adequate power to detect moderate-to-large intervention effects on the primary outcomes. Because students were allocated by administrative class and teaching session, this individual-level calculation may overestimate the effective sample size if class-level clustering is substantial; clustering-aware sensitivity analyses are described in the Statistical Analysis section.

Setting and Instructors

A total of 3 instructors with relevant teaching experience in the course and similar pedagogical backgrounds and experience in occupational medicine participated in the teaching sessions and taught across both study arms on a rotating basis. Before the teaching sessions, all instructors underwent standardized training and participated in collective lesson preparations to harmonize the learning objectives, core knowledge points, classroom tasks, and class duration. Apart from the specific components of the GenAI-assisted progressive-disclosure CBL package, including staged disclosure, structured reasoning tasks, and multimodal materials, the 2 groups received the same teaching topic, comparable class duration, and the same basic instructional requirements. We acknowledge that instructor-level differences and other unmeasured factors, such as prior exposure to CBL, digital literacy, learning motivation, and previous clinical-reasoning experience, may have influenced the results and are addressed as limitations.

Teaching Intervention

Each case discussion session lasted 3 class hours (45 minutes per class hour). Both groups covered the same learning

XSL•FO
RenderX

objectives, core content, and contact hours on occupational lead poisoning; only the case-delivery format differed between groups. The control group used a static case format, whereas the intervention group used a GenAI-assisted 4-act progressive-disclosure format while covering the same knowledge points.

Control Group

Students received conventional CBL, in which the full, static case record (including chief complaint, history, physical examination, laboratory findings, occupational exposure history, and workplace context) was disclosed at the beginning of the session. Classroom activities followed a linear workflow ("presentation → discussion → summary"), primarily emphasizing information extraction and diagnosis matching.

Intervention Group

Students received the GenAI-assisted progressive-disclosure CBL package [17,18], which used a scripted 4-act internal structure and progressively released key information together with teacher-verified distractor cues across acts.

The 4-act structure mapped onto the real-world workflow for recognizing and managing occupational poisoning and aligned with a staged progression of learners' cognitive processes from cue detection to evidence weighting, causal attribution, and intervention decision-making. To prevent premature "answer-matching" once information became sufficient, each act provided only the minimum information required for the current task, while other cues were deliberately retained as "to be elicited/to be verified." This approach prompted students to generate structured question lists, propose subsequent diagnostic tests or field investigations, and update their diagnostic hypotheses as new evidence became available.

With respect to scenario and competence design, Act 1 focused on problem framing and initial hypothesis generation under nonspecific symptoms. Act 2 emphasized occupational history elicitation and laboratory cues to train evidence identification, evidence grading, and diagnostic calibration. Act 3 shifted to field investigation and exposure attribution, training learners to integrate individual clinical information with workplace/environmental evidence to infer the "exposure–environment–health effect" causal chain. Act 4 addressed management and prevention, training learners to balance clinical treatment with occupational health interventions and to formulate action plans integrating individual therapy and population-level prevention, followed by reflective summarization and transfer to new contexts.

Across the acts, GenAI served as a "scenario driver" and "instructional scaffolding" tool. Specifically, GenAI was used to rapidly construct high-fidelity clinical encounter scenarios using multimodal materials (eg, images and audio) and to generate controllable nonspecific distractor cues in Act 1; to present structured dialogue scripts for occupational history probing and to externalize key evidence-threshold relationships through mind maps or knowledge graphs in Act 2; to provide visualizations and checklist-style prompts for workplace assessment and field investigation in Act 3; and to support synthesis of treatment and occupational health intervention points using knowledge graphs and decision frameworks in Act 4. Importantly, all GenAI-generated materials and cues were reviewed and validated by instructors before classroom deployment to ensure medical accuracy and instructional safety.

To minimize contamination bias (ie, cross-group sharing of case details), the 2 groups were scheduled in parallel time slots within the same week to reduce opportunities for information exchange. In addition, all students signed an academic integrity agreement not to disseminate the 4-act structure or distractor cues prior to study completion. During the intervention, GenAI was used solely on the instructor side for instructional material development. To assess students' independent clinical reasoning, the use of GenAI tools by students was prohibited during in-class discussions and the final case-analysis assessment. After the class, students were asked to limit any subsequent discussions to their general learning experiences and perceptions of the course and not to discuss specific case content, implementation procedures, or instructional methods.

Table 1. Scripted "4-act" case-based learning design for the lead poisoning case.

Act	Scenario	Key cues and prompts	Generative AI support	Targeted competencies
Act 1	Illness onset and clinical encounter	Nonspecific symptoms/signs (eg, fatigue, arthralgia, right upper-limb numbness, and abdominal pain)	Rapid construction of the encounter scenario (multimodal visualization via text-to-image/audio); generation of controllable distractor cues (eg, pallor and constipation)	Problem framing and initial hypothesis generation
Act 2	Occupational history and laboratory cues	Battery-factory exposure history; elevated blood/urine lead levels; signs such as a "lead line"	Dialogue-based scenario for targeted occupational-history probing; mind maps/knowledge graphs (eg, blood/urine lead grading and diagnostic thresholds)	Evidence identification and grading
Act 3	Field investigation and exposure attribution	Insufficient engineering controls, inadequate PPE, airborne lead exceeding limits, etc.	Visualization of the workplace/operational environment; scenario-based field investigation; structured presentation of sampling/testing indicators and cues	Exposure–health effect causal attribution
Act 4	Management and prevention decision-making	Standardized treatment, occupational health interventions, and tertiary prevention	Knowledge graphs/mind maps to support synthesis of key points and transfer-oriented summarization	Integrated decision-making and reflective practice

XSL•FO
RenderX

HITL Quality Control Protocol

To address the inherent risks of AI hallucinations, defined as the generation of plausible but clinically inaccurate or fabricated information, a formal HITL quality control protocol was established. Under this protocol, the GenAI tool was strictly confined to a "faculty-controlled development environment" (sandboxed environment). At no point were students granted direct, unmediated access to the raw model output.

All toxicological parameters or excerpts of real clinical records used to generate virtual patient data were deidentified in accordance with applicable privacy standards (eg, China's Personal Information Protection Law). In addition, the model's data retention option was disabled to ensure that the proprietary educational case scripts were not stored or used for subsequent model training. The case-generation process followed a 3-stage verification cycle:

1. Iterative prompt engineering: faculty members used structured prompts to generate the initial 4-act script, clinical distractors, and virtual patient dossiers. Prompt templates are provided in Multimedia Appendix 1.
2. Expert blind review: a panel of 3 senior occupational medicine specialists (PS, CC, and SX), blinded to the AI's initial drafts, reviewed all generated content for clinical accuracy, toxicological realism, and alignment with the National Medical Licensing Examination standards. The panel comprised occupational medicine faculty members, each with more than 3 years of teaching experience, who also performed the hallucination filtering and correction stage and recorded the audit metrics.
3. Hallucination filtering: instances of "hallucinated" data—such as incorrect laboratory reference ranges or illogical environmental exposure histories—were manually corrected or regenerated. The number of errors, normative edits, and review time per draft were recorded to enable transparency reporting in accordance with the GenAI in Medical Research (GAMER) statement.

This HITL framework intended to support the use of GenAI as an efficient pedagogical engine while remaining under strict human oversight, thereby safeguarding the educational and ethical integrity of the intervention.

GenAI Implementation and Prompt Engineering

Tools and Prompting

DeepSeek-V3 was used as the underlying model to generate multimodal case materials, patient personas, and stage-specific distractor information. The prompting template included explicit role specification (eg, "act as an expert in occupational toxicology"), clinical constraints (eg, "blood lead levels must comply with Chinese national diagnostic criteria for occupational chronic lead poisoning, GBZ 37-2015"), and a structured output format (see Multimedia Appendix 1). In addition, this study followed the GAMER reporting guidelines [19]. We explicitly detail the model specifications, prompt histories, human oversight workflows, and privacy measures. The completed GAMER checklist is provided in Multimedia Appendix 2.

Content Verification and Quality Control

To mitigate inaccuracies in AI-generated outputs, a standard operating procedure (SOP) was established [19]. Faculty members reviewed the generated materials for medical thresholds, logical consistency, terminology, and alignment with occupational health teaching objectives before classroom use. All AI drafts were independently reviewed by 3 senior instructors in occupational medicine (the same expert panel described in "HITL Quality Control Protocol") against current national occupational health standards and clinical guidelines. Critical thresholds (eg, lethal doses and absolute contraindications) were manually checked before any material was approved for classroom use.

Student-Use Policy

During the intervention, GenAI was used solely on the instructor side for instructional material development. To assess students' independent clinical reasoning, the use of GenAI tools by students was prohibited during in-class discussions and the final case-analysis assessment.

Outcome Measures

The outcome measures covered both learning outcomes and learning processes, including objective academic performance, subjective learning experience, and objective behavioral indicators of classroom engagement. To manage the multiplicity of statistical tests, a hierarchical analytic plan was defined. The case-analysis assignment score was the sole primary outcome; the delayed final examination score, the 4 reasoning-dimension scores, and the 5 learning experience domain ratings were secondary outcomes; and the 3 behavioral indicators and the subgroup analyses by major were exploratory process and generalizability outcomes.

Primary Outcome: Standardized Case-Analysis Assignment Score (0-100)

This measure was used to evaluate students' knowledge integration and knowledge-transfer performance in an authentic context. Each student completed the assignment independently. The assignment was uniformly designed and deidentified and was scored independently by 2 raters who were blinded to group allocation; interrater reliability was assessed using the intraclass correlation coefficient (ICC). To reduce subjectivity in scoring open-ended responses, an analytic rubric was used to standardize grading. Each question was decomposed into observable key scoring points, with partial-credit and deduction rules predefined. Prior to formal scoring, rater training and calibration were conducted, and a rescoring/adjudication rule was established (eg, third-party review was initiated when the discrepancy between the 2 raters exceeded a predefined threshold). In addition, the total score was further decomposed into 4 clinical reasoning dimensions (25 points each; total 100 points), corresponding to information gathering, hypothesis generation, differential diagnosis, and treatment/management planning, to examine the differential effects of the intervention across reasoning components. The case-analysis assignment used a new clinical case with novel contextual details and was

XSL•FO
RenderX

not identical to the case used in the teaching session; the rubric was developed by the teaching team before the intervention on the basis of the 4 target reasoning dimensions, and students did not have access to the scoring rubric during teaching. Interrater reliability for the case-analysis scoring was excellent (ICC=0.91).

Secondary Outcome: Final Examination Score (0-100)

The final examination was administered 3-4 weeks after completion of the case-based teaching module and served as a delayed posttest to assess short-term retention and transfer of the intervention effects several weeks later. The examination was a closed-book test with a total score of 100 points, consisting of 40 points for subjective items and 60 points for objective items. The subjective section included term explanations, short-answer questions, and case-analysis questions, whereas the objective section covered knowledge points from the preventive medicine curriculum. Within the examination, case analysis–related items accounted for 15 points, lead poisoning–related objective items accounted for 5 points, and reasoning-oriented objective items accounted for approximately 10 points. Students in both groups took the same course examination using versions A and B; the 2 versions contained the same item content but used different option orders to reduce potential examination-related interference. Item writing and grading were conducted while blinded to students' group allocation. The final examination scores were independently scored by 2 raters blinded to group allocation, with excellent interrater reliability (ICC=0.92). The reliability of the examination was verified according to the routine quality-control procedures for the course examination. Because of institutional examination-confidentiality requirements, the complete examination paper cannot be publicly released, and item-level analyses by question type could not be further conducted.

Secondary Outcome: Subjective Learning Experience

Students completed a structured self-report questionnaire using a 0-10 point scale to rate five dimensions: (1) perceived difficulty, (2) perceived improvement in clinical reasoning, (3) engagement, (4) self-efficacy (confidence in transferring the reasoning framework to new contexts), and (5) evaluation of the course materials, which included the GenAI-generated elements in the intervention group. Details of the instrument are provided in Multimedia Appendix 3. Each domain comprised 4 items (Multimedia Appendix 3), and the domain score was the mean of its 4 items. Internal consistency was assessed with Cronbach α within each domain (α=0.77-0.91); because the 5 domains measure conceptually distinct constructs, an overall coefficient across domains was not interpreted as evidence of scale reliability, and each domain was analyzed as its own rating. The ratings were exploratory and should not be interpreted as objective evidence of educational effectiveness.

Exploratory Process Outcomes: Objective Behavioral Indicators of Classroom Engagement

To objectively validate students' self-reported learning engagement and further capture behavioral changes during the classroom learning process, video recordings of the sessions were reviewed and analyzed. A behavioral test was performed by trained instructors using a standardized question list developed during collective lesson preparation (10 questions designed for individual responses and 6 questions designed for group discussion followed by representative responses). During the sessions, the instructor used a random-calling strategy to select individual students and representative groups to answer these structured questions. Consequently, a total of 80 individual responses (10 questions×8 teaching sessions; 40 per arm) and 48 group responses (6 questions×8 teaching sessions; 24 per arm) were captured in the video recordings and fully coded for behavioral metrics. Because of constraints related to course scheduling and teaching implementation, repeated measurements of these classroom behavioral indicators were not conducted. In addition, voluntary responses may have been influenced by the specific questions posed. Therefore, these indicators were reported descriptively as exploratory process measures and were used to support the interpretation of changes in student engagement behaviors.

Three classroom behavioral indicators were extracted during the case discussions: (1) the number of voluntary responses to questions, counted at the individual student level, (2) the number of evidence- or reference-based statements, counted at the individual student level, and (3) net group discussion duration, recorded at the group level in minutes per group. These indicators were used to assess, from a behavioral perspective, whether the teaching model promoted a shift in students' learning behaviors from passive reception to active inquiry.

A "voluntary response" was defined as a response initiated by a student after the teacher posed a question, without being called on or required to answer by the teacher. This included voluntarily raising a hand to answer or actively communicating with the teacher regarding the question. Responses resulting from teacher nomination, direct assignment, or mandatory questioning were not counted as voluntary responses; "The number of evidence- or reference-based statements" is defined as the number of times a student, when answering a question, cites in-class clinical materials or diagnostic standards to support their answer. "Net group discussion duration" was defined as the period during which more than 2 students within a group interacted with each other in the video recording; the discussion was considered to have ended when no student interaction was observed.

Exploratory Outcome: Subgroup Comparisons

To evaluate the applicability of the intervention across educational backgrounds, subgroup comparisons of final exam scores were conducted by major (clinical medicine, medical laboratory science, and psychiatry), together with group-by-major interaction tests, as exploratory outcomes.

AI-Generated Occupational Exposure Case Development and Expert Audit

To evaluate and control the quality of AI-generated occupational medicine cases, we developed a structured HITL audit protocol. The teaching content covered representative occupational hazards commonly included in undergraduate occupational medicine training, including lead poisoning, arsenic exposure,

XSL•FO
RenderX

hydrogen sulfide poisoning, and dust-related occupational lung disease.

For each teaching topic, a hazard-specific key prompt was constructed to guide the large language model in generating an occupational exposure case. The prompt required the model to produce a staged clinical scenario involving occupational history, exposure pathway, clinical manifestations, laboratory or environmental monitoring findings, differential diagnosis, and prevention or management measures. The model was instructed to align case details with relevant national occupational health standards and diagnostic criteria and to include plausible distractors that could support progressive disclosure and clinical reasoning training.

All AI-generated case drafts were subsequently reviewed by 3 occupational medicine faculty members, each with more than 3 years of teaching experience. The reviewers independently assessed each case according to national occupational disease diagnostic standards, occupational exposure limit criteria, and routine teaching requirements. During the review process, each teacher recorded:

1. The number of medical indicator or threshold errors identified and corrected
2. The number of logical inconsistencies or unreasonable plots identified and rectified
3. The number of terminology or expression revisions made for standardization
4. The total review and correction time in minutes

Medical indicator or threshold errors included incorrect diagnostic thresholds, inappropriate laboratory reference values, unreasonable exposure concentrations, or inconsistency with national diagnostic standards. Logical inconsistencies referred to contradictions between exposure history, clinical manifestations, laboratory findings, and disease progression. Terminology and expression revisions referred to modifications of nonstandard occupational health terminology, ambiguous descriptions, or wording inconsistent with teaching norms.

The final quality-control indicators were calculated at the case level. A total of 25 case drafts were generated (lead poisoning n=6, arsenic n=6, hydrogen sulfide n=7, and dust-related occupational lung disease n=6), and each draft was independently reviewed by 3 instructors, yielding 75 reviewer assessments. For each case, the mean number of errors or revisions identified by the 3 reviewers was calculated; descriptive statistics were then reported as mean (SD) across the 25 cases.

Statistical Analysis

Statistical analyses were conducted using R (version 4.5.1; R Core Team) software. Continuous variables are reported as mean (SD), and categorical variables as frequencies and percentages. Baseline characteristics were compared between groups using independent-samples Student *t* tests for continuous variables and chi-square tests for categorical variables; standardized mean differences (SMDs) were reported for continuous baseline variables. Effect sizes for continuous outcomes were quantified using Cohen *d* (with conventional thresholds: small ≥0.2, medium ≥0.5, large ≥0.8), and 95% CIs for *d* were computed from the noncentral *t* distribution. For between-group comparisons of the primary case-analysis score and the final examination score, we used independent-samples *t* tests. To evaluate the robustness of the main comparisons to baseline imbalance, analysis of covariance (ANCOVA) models adjusted for prerequisite-course GPA were fitted for both outcomes. The 4 clinical reasoning dimension scores, the 5 subjective-experience domains, and the 3 behavioral indicators were treated as secondary or exploratory outcomes, with the Holm-Bonferroni correction applied within each outcome family to control the family-wise type I error rate. The 3 subgroup comparisons by major were treated as exploratory; the formal test for differential treatment effects across majors was a group-by-major interaction term in a 2-way ANOVA, rather than the statistical significance of individual subgroup tests.

Interrater reliability for both case-analysis scoring and final examination scoring was assessed using a 2-way random-effects ICC with absolute agreement. Behavioral discussion-duration data were analyzed at the group level, consistent with their unit of measurement; voluntary responses and evidence-referencing statements were analyzed at the individual-student level. Missing data were handled by complete-case analysis because the rate of missingness was negligible (<1%). All tests were 2-tailed, and $P<.05$ was considered statistically significant unless otherwise specified.

Because allocation was implemented at the class and teaching-session level, students within a session cannot be regarded as fully independent observations. As a sensitivity analysis, we estimated ICCs using academic major as a proxy clustering variable; the estimated ICCs were approximately 0 for both objective outcomes. We also computed cluster-robust standard errors at the 6 group-by-major strata (3 majors×2 study arms), which corroborated the primary results. Administrative-class identifiers were not retained in the deidentified dataset, so a formal class-level multilevel model could not be fitted; this is acknowledged in the Limitations.

Results

Participant Flow and Baseline Equivalence

A total of 224 undergraduate students completed the study (intervention group, n=114; control group, n=110). No participants were excluded or withdrew, and no primary outcome data were missing; the analytic sample was identical to the enrolled sample (Figure 1). Baseline characteristics were compared prior to outcome comparisons. As shown in Table 2, there were no statistically significant between-group differences in age, sex distribution, or academic performance in the prerequisite occupational medicine course (all $P>.05$). These results suggest similar measured baseline characteristics on the measured variables; however, nonsignificant baseline comparisons do not establish equivalence, and unmeasured characteristics (eg, prior CBL exposure and motivation) may still differ between groups.

XSL•FO
RenderX

Figure 1. Flowchart of participation.

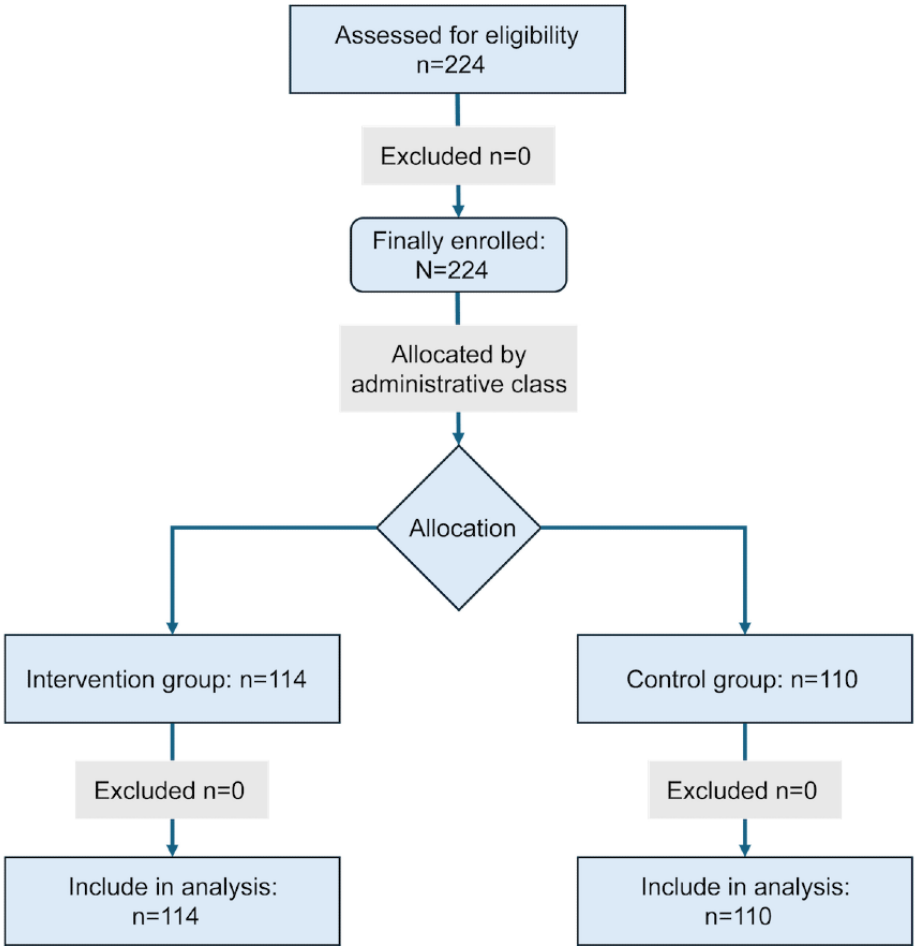


Table 2. Baseline characteristics of the participants

Characteristic	Intervention group (n=114)	Control group (n=110)	Test statistic	*P* value	Standardized difference
Age (years), mean (SD)	20.4 (1.2)	20.6 (1.1)	–1.33 (222)[a]	.20	–0.17[b]
Sex (male/female)	54/60	50/60	0.10 (1)[c]	.75	0.02[d]
Prior toxicology GPA[e], mean (SD)	78.41 (4.65)	78.55 (8.24)	–0.15 (222)[a]	.88	–0.02

[a]*t* test (*df*) value.

[b]Standardized mean difference.

[c]Chi-square (*df*) values.

[d]Cramér V.

[e]GPA: grade point average.

Objective Learning Outcomes

Objective learning outcomes were assessed using the postclass case-analysis score and the final examination score. Interrater reliability was good for the case-analysis grading (ICC=0.91) and for the final examination scoring (ICC=0.92). For the case analysis, the intervention group achieved significantly higher scores than the control group (mean 88.41, SD 3.85 vs mean 82.51, SD 3.94; *P*<.001), with a large effect size (Cohen *d*=1.51, 95% CI 1.23-1.83). Next, the final examination score was also compared, showing higher performance in the intervention group (mean 80.96, SD 8.47 vs mean 78.62, SD 6.63; *P*=.02), with a small-to-moderate effect size (Cohen *d*=0.31, 95% CI 0.04-0.57). The divergence between the case-analysis effect (Cohen *d*=1.51) and the final examination effect (Cohen *d*=0.31) suggests that the intervention most strongly reinforced the structured reasoning workflow practiced in the case-analysis task, while effects on broader knowledge consolidation were more modest. Cluster-robust sensitivity analyses at the 6 group-by-major strata corroborated these conclusions. Details are presented in Table 3.

XSL•FO
RenderX

Table 3. Comparison of objective learning outcomes between the 2 groups.

Outcome measure	Intervention group (n=114), mean (SD)	Control group (n=110), mean (SD)	Mean difference (95% CI)	*P* value	Cohen *d* (95% CI)
Case analysis score (0-100)	88.41 (3.85)	82.51 (3.94)	5.90 (4.87-6.93)	<.001	1.51 (1.23-1.83)
Final exam score (0-100)	80.96 (8.47)	78.62 (6.63)	2.33 (0.33-4.33)	.02	0.31 (0.04-0.57)

Subdimensional Clinical Reasoning Performance

To clarify the impact of the GenAI-assisted progressive-disclosure CBL package on domain-specific reasoning competencies, the total case-analysis score was decomposed into 4 core dimensions (25 points each). As shown in Table 4, the intervention group outperformed the control group in information gathering (mean 22.66, SD 1.87 vs mean 20.46, SD 2.13), hypothesis generation (mean 22.28, SD 1.99 vs mean 20.47, SD 2.06), and differential diagnosis (mean 22.34, SD 1.96 vs mean 20.72, SD 2.09), with all differences reaching statistical significance (all P<.001) and corresponding to large effect sizes (Cohen d≈1.10, 0.89, and 0.80, respectively). By contrast, the difference in treatment and management planning did not reach statistical significance (mean 21.13, SD 1.56 vs mean 20.86, SD 1.55; P=.19; Cohen d=0.18).

Table 4. Subdimensional scores for clinical reasoning in the case analysis worksheet.

Reasoning dimension (maximum 25 points each)	Intervention group (n=114), mean (SD)	Control group (n=110), mean (SD)	Mean difference (95% CI)	*P* value	Cohen *d*	Holm-adjusted *P* value
Information gathering	22.66 (1.87)	20.46 (2.13)	2.20 (1.67 to 2.73)	<.001	1.10	<.001
Hypothesis generation	22.28 (1.99)	20.47 (2.06)	1.80 (1.27 to 2.34)	<.001	0.89	<.001
Differential diagnosis	22.34 (1.96)	20.72 (2.09)	1.62 (1.09 to 2.16)	<.001	0.80	<.001
Treatment and management plan	21.13 (1.56)	20.86 (1.55)	0.27 (–0.14 to 0.68)	.19	0.18	.19

This pattern is consistent with the conceptual framework underlying the progressive-disclosure design. The progressive disclosure and verified distractor cues may be particularly effective for strengthening hypothesis-evidence calibration during exploratory reasoning, whereas treatment and management planning relies more heavily on retrieval and application of standardized guideline-based knowledge, which can also be adequately supported by conventional CBL. We interpret this differential pattern as consistent with—rather than direct evidence of—the hypothesized mechanism, rather than as a generic “more is better” effect on all reasoning dimensions.

Subgroup Analysis by Major

To explore whether the intervention effect varied across educational backgrounds of the GenAI-assisted progressive-disclosure CBL package across educational backgrounds, we conducted subgroup analyses by major and compared final examination scores (Table 5). The intervention group scored higher than the control group in all 3 majors; however, statistical significance was observed only among students in clinical medicine (mean 81.55, SD 8.63 vs mean 78.66, SD 6.27; P=.05). Differences were not statistically significant in psychiatry (mean 81.05, SD 8.22 vs mean 77.50, SD 6.39; P=.08) or medical laboratory science (mean 79.75, SD 8.55 vs mean 79.68, SD 7.57; P=.97). After Holm-Bonferroni adjustment across the 3 subgroup comparisons, none of the differences remained statistically significant (adjusted P=.14, .15, and .97, respectively). The group-by-major interaction for the final examination was not statistically significant ($F_{2,218}$=0.88; P=.41), indicating no reliable evidence that the intervention effect differed by major. Scores did not differ significantly across majors overall ($F_{2,221}$=0.23; P=.79). Overall, these exploratory findings suggest that the intervention effect was directionally larger in clinically oriented disciplines, but the limited subgroup sample sizes and the nonsignificant interaction test warrant cautious interpretation and validation in larger, multicenter studies.

Table 5. Subgroup analysis of final examination scores by academic major.

Academic major	Intervention group score	Control group score	Mean difference (95% CI)	Exploratory *P* value
Clinical medicine	81.55 (8.63); (n=56)	78.66 (6.27); (n=54)	2.90 (0.00 to 5.79)	.05
Psychiatry	81.05 (8.22); (n=28)	77.50 (6.39); (n=28)	3.55 (–0.42 to 7.53)	.08
Laboratory medicine	79.75 (8.55); (n=30)	79.68 (7.57); (n=28)	0.07 (–4.12 to 4.35)	.97

Subjective Learning Experience

To control the family-wise type I error arising from multiple comparisons among the 5 domains, a Holm-Bonferroni correction was applied. Internal consistency was acceptable to high across the 5 domains (Cronbach α: perceived difficulty 0.77; reasoning improvement 0.85; engagement 0.83; self-efficacy 0.84; course materials 0.91). As shown in Table 6, the intervention group reported significantly higher scores for perceived improvement in clinical reasoning (mean 8.55, SD 0.50 vs mean 5.96, SD 0.79; P<.001) and course engagement (mean 8.46, SD 0.50 vs mean 6.11, SD 0.81; P<.001) than the

XSL•FO
RenderX

control group. Self-efficacy for transferring the reasoning framework to new contexts was also higher in the intervention group (mean 8.58, SD 0.50 vs mean 6.13, SD 0.80; *P*<.001). Meanwhile, the intervention group perceived greater case difficulty (mean 7.53, SD 0.50 vs mean 5.98, SD 0.80; *P*<.001), suggesting that progressive disclosure and teacher-verified AI distractor cues increased cognitive challenge and demanded more rigorous hypothesis testing, consistent with the desirable-difficulty framing introduced in the Introduction. In addition, ratings for evaluation of course materials were substantially higher in the intervention group than in the control group (mean 8.54, SD 0.88 vs mean 5.44, SD 0.93; *P*<.001), suggesting a more positive evaluation of the course materials. All 5 comparisons remained statistically significant after the Holm-Bonferroni adjustment (adjusted *P*<.001).

Table 6. Student subjective evaluation of the learning experience (scale: 0-10).

Questionnaire domain	Intervention group (n=114)	Control group (n=110)	Mean difference (95% CI)	Unadjusted *P* value	Holm-adjusted *P* value
Perceived difficulty	7.53 (0.50)	5.98 (0.80)	1.54 (1.37-1.72)	<.001	<.001
Clinical reasoning enhancement	8.55 (0.50)	5.96 (0.79)	2.60 (2.42-2.77)	<.001	<.001
Course engagement	8.46 (0.50)	6.11 (0.81)	2.36 (2.18-2.53)	<.001	<.001
Self-efficacy	8.58 (0.50)	6.13 (0.80)	2.45 (2.28-2.63)	<.001	<.001
Evaluation of course materials	8.54 (0.88)	5.44 (0.93)	3.10 (2.86-3.34)	<.001	<.001

Objective Classroom Engagement and Behavioral Metrics

The self-reported increase in engagement was supported by objective behavioral indicators derived from classroom video analysis (Table 7). Compared with the control group, students in the intervention group responded voluntarily to questions more frequently (mean 2.30, SD 1.11 vs mean 1.68, SD 0.92; *P*=.008). The intervention group also produced significantly more evidence- or reference-based statements (mean 1.90, SD 0.88 vs mean 1.45, SD 0.71; *P*=.01). The progressive-disclosure structure also significantly increased net group discussion time (mean 36.50, SD 6.88 vs mean 28.42, SD 7.02 minutes; *P*<.001), suggesting more intensive interaction and deeper discussion. All 3 behavioral differences remained significant after the Holm-Bonferroni adjustment across the 3 indicators (adjusted *P*=.02, .02, and <.001, respectively). Although the frequency of evidence-referencing statements was higher in the intervention group, the absolute frequencies remained low in both groups, and the formal quality of the cited evidence was not assessed; this remains an area for further instructional development.

Table 7. Objective behavioral metrics of student engagement during case discussions.

Behavioral metric	Intervention group (n=114), mean (SD)	Control group (n=110), mean (SD)	*P* value
Voluntary responses	2.30 (1.11)	1.68 (0.92)	.008
Information/reference search frequency	1.90 (0.88)	1.45 (0.71)	.01
Net group discussion duration (minutes)	36.50 (6.88)	28.42 (7.02)	<.001

AI Hallucination Mitigation and Quality Control Metrics

In accordance with the GAMER statement on transparency in reporting AI use, we quantified the HITL quality-control process during case development. Across the generated case drafts (each independently reviewed by 3 occupational medicine faculty members; per-case values are the mean of the 3 reviewers), every draft required expert correction. Independent instructor audit logs showed that each AI-generated case contained, on average, 1.65 (SD 0.68) errors in medical indicators/thresholds and 2.43 (SD 1.08) instances of logical inconsistency. To standardize terminology and expression, instructors implemented an average of 3.12 (SD 1.12) normative edits per case. The mean time required for expert review and correction of one AI-generated draft was 18.72 (SD 2.89) minutes. These findings indicate that, prior to classroom deployment, domain-expert oversight remains indispensable for mitigating hallucination risk and ensuring instructional safety (Table 8).

Table 8. Quantification of the human-in-the-loop audit for AI-generated cases.

Quality control metric	Occurrences/time per case, mean (SD)
Medical indicators/threshold errors corrected	1.65 (0.68)
Logical inconsistencies/unreasonable plots rectified	2.43 (1.08)
Terminology/expression normative revisions	3.12 (1.12)
Average expert review and audit time (minutes)	18.72 (2.89)

XSL•FO
RenderX

Discussion

Principal Findings

Our findings indicate that the GenAI-assisted progressive-disclosure CBL package was associated with significantly higher clinical reasoning performance among learners, higher scores on reasoning dimensions theoretically related to premature closure, and, to some extent, greater willingness to engage in active classroom participation. The observed benefits also suggest a degree of cross-disciplinary applicability, although the nonsignificant group-by-major interaction indicates that differential effects across majors were not established. It should be emphasized that the proposed 4-act structure is not an entirely novel design; rather, it is conceptually grounded in the well-established tradition of UCS [5,20]. The core pedagogical mechanism of UCS lies in progressive disclosure, whereby information is released in stages so that learners repeatedly cycle through hypothesis generation, information seeking, and judgment recalibration under uncertainty—more closely approximating how evidence emerges in real clinical care and occupational health practice. A recent scoping review further suggests that UCS-based designs are associated with overall positive performance outcomes in nursing education, including improvements in critical thinking, knowledge acquisition, and decision-making performance in situated tasks [5,21]. Building on this foundation, our 4-act design further distills key reasoning steps from authentic practice and places greater emphasis on contextual dynamics and the process nature of reasoning. An increase in student-initiated questions better reflects a learner's active drive to seek information and test hypotheses under uncertainty, making it a highly process-sensitive measure of engagement. Furthermore, while the intervention yielded a large effect on case analysis scores (Cohen d=.51), this likely indicates improved "near transfer" of the specific reasoning workflow. Conversely, the smaller effect on the final examination (Cohen d=0.31) suggests only a modest gain in "far transfer"—the ability to integrate these skills into a broader knowledge base. The convergence between our results and the UCS literature provides additional support for the theoretical plausibility of progressive-disclosure CBL, while also indicating potential generalizability beyond a single discipline.

Regarding case construction, the "Scripted Remedies" approach proposed by Simms and Fox illustrates how GenAI can be leveraged to align narrative elements (eg, popular-culture motifs) with instructional goals to produce scripted cases with controllable complexity and narrative tension [22]. This is consistent with our design logic—namely, using AI to support efficient generation of scenarios and cues while maintaining instructor verification to safeguard accuracy. Consistent with the hypothesized mechanism, our dimension-level outcome analyses suggest that the intervention advantage was more pronounced in information gathering, hypothesis generation, and differential diagnosis, whereas differences in treatment and management planning were not statistically significant. This pattern implies that the scripted 4-act structure and verified distractor cues may preferentially strengthen exploratory reasoning under uncertainty, while treatment planning may rely more heavily on retrieval and application of established knowledge and guidelines—skills that can also be adequately supported by conventional teaching approaches. This suggests the intervention's scope is bounded by exploratory reasoning; it should not be regarded as a general clinical-reasoning enhancer. Process indicators likewise showed that learners in the intervention condition demonstrated higher levels of engagement, including a significantly higher frequency of explicit evidence-referencing statements, which had previously been regarded as a nonsignificant trend in an earlier analysis; however, the absolute frequencies remained low in both groups, and the quality of the cited evidence was not assessed, so the formal quality and appropriateness of evidence-based argumentation remain areas for further instructional development. In subgroup analyses stratified by major, the small sample sizes in the psychiatry and medical laboratory science groups limited the statistical power to detect moderate between-group differences, and the group-by-major interaction was not statistically significant. Future studies with larger cohorts are needed to robustly confirm the intervention's efficacy across these disciplines.

SLT has been widely applied in medical education [23]. Using Goffman's dramaturgical metaphor, Cantillon et al [24] described how professional roles and norms are co-constructed through everyday interactions, informal learning, and collaborative work within clinical teams. From this perspective, clinical reasoning is rarely a purely linear "paper-and-pencil" logic exercise; rather, it resembles an evolving "performance" and "detective work" that requires iterative cue tracking and judgment recalibration under incomplete information, distributed responsibilities, and contextual pressures. Accordingly, the GenAI-assisted "4-act" scenario in the present study may be interpreted as a classroom-based micro-model of clinical ecology: by combining staged disclosure with role-based tasks, it allows learners to rehearse professional participation in a safer instructional environment, which is closely aligned with SLT's emphasis on context-embedded learning. Related work on "Socratic AI" and adaptive tutoring for clinical CBL further proposes decomposing lengthy case narratives into discrete information nodes, releasing key cues only when learners ask the appropriate questions and complete necessary reasoning steps. This aligns with our "thresholded information disclosure" strategy and suggests that our intervention can be situated within an emerging family of adaptive, reasoning-oriented tutoring paradigms. In addition, questionnaire results indicated that learners perceived higher difficulty under the progressive-disclosure format, while simultaneously reporting higher engagement and self-efficacy. This pattern is consistent with Bjork's concepts of desirable difficulties and productive struggle: moderate cognitive friction may not constrain learners but may instead motivate continuous hypothesis calibration and evidence-chain reconstruction, thereby facilitating deeper processing, a stronger sense of mastery, and greater learning motivation.

The application space of GenAI in medical education continues to expand. Potter and Jefferies [25] emphasized that GenAI may move beyond static scripts by enabling more dynamic, natural-language interaction with "living" virtual patients,

XSL•FO
RenderX

potentially providing higher-fidelity contexts for communication and clinical reasoning training. In the present study, both process and outcome data suggest that the GenAI-assisted progressive-disclosure design can improve performance and promote engagement; the higher ratings for the course materials (which included GenAI-generated elements in the intervention group) also indicate an overall positive attitude toward integrating GenAI into the course [21,26]. Importantly, the value of GenAI in this study lies in its role as an efficient drafting tool whose outputs still required substantial expert verification; the available data do not allow us to claim reductions in total case-development time or cost, because no comparator based on conventional instructor-developed cases was included. Silvestri-Elmore and Burton [20] reported that AI and educational technologies can substantially alleviate the time and workload barriers that faculty face when designing complex case scripts (including UCS). Consistent with that view, the expert audit documented that review and correction of AI-generated drafts were required before classroom use; quantifying the net time saved relative to conventional case development is a task for future comparative studies.

Nevertheless, the capability boundaries and safety risks of GenAI in clinical reasoning should not be underestimated [27]. Although some models have been adapted using medical corpora and clinical knowledge [28,29], their reliability and safety remain contested [30]. Rao et al [31] reported that large language models may achieve high final-diagnosis accuracy when provided with complete case information (>90%), yet often perform poorly in early-stage differential diagnosis and reasoning navigation under information scarcity. This discontinuity has direct methodological implications for our instructional strategy: precisely because AI may be intrinsically less reliable in open-ended, information-incomplete reasoning phases, we deliberately avoided allowing learners to engage in unconstrained free-form dialogue with the model [32]. Instead, we adopted an expert-controlled, prescripted 4-act structure together with a human HITL verification and correction process. Consistent with this rationale, our audit data showed that AI-generated materials still required human correction of medical thresholds/values and logical inconsistencies, indicating that the unreviewed model outputs generated in this study were suitable for direct classroom deployment [33,34]. Accordingly, rather than advocating “AI replacing instructors,” our findings support a sustainable trajectory of human-AI collaboration, in which AI serves as a high-efficiency generator and scaffolding tool [35], whereas medical fact-checking, logical consistency review, and ethical/value judgments remain firmly human-led—particularly in domains such as clinical medicine, occupational medicine, and public health that depend heavily on standards, field evidence, and practice-based expertise. Our HITL workflow may therefore be viewed not only as an experimental control but also as a potentially transferable SOP for responsible integration of GenAI in educational settings. It should also be acknowledged that such a pathway increases demands on instructors. Educators require sufficient knowledge and practical experience to identify and correct AI errors in a timely manner.

This study has several limitations: (1) group assignment was based on administrative classes rather than individual-level randomization. Although baseline characteristics were comparable (SMD <0.2 for the measured variables), unobserved class-level factors could have influenced effect estimates, and the absence of class identifiers prevented a formal class-level multilevel analysis. (2) The intervention was a multicomponent package; progressive disclosure, the 4-act structure, distractor cues, multimodal materials, and differences in instructor facilitation were introduced together, so the independent contribution of GenAI cannot be isolated. (3) No preintervention baseline measure of clinical reasoning or the primary outcome was available; baseline similarity in demographic variables does not establish equivalence in reasoning ability, and unmeasured confounders (prior CBL exposure, motivation, digital literacy, and previous reasoning experience) cannot be excluded. (4) Teacher-level variables were not systematically measured or controlled. Delivering the GenAI-assisted progressive-disclosure CBL package typically requires stronger in-the-moment facilitation skills, including adaptive questioning, guidance under uncertainty, and fine-grained management of pacing and small-group interaction; implementation fidelity was not assessed with an independent quantitative rating instrument, and instructor enthusiasm and interaction style were not rated. (5) Students were not blinded to condition; favorable subjective ratings and behavioral engagement may partly reflect novelty effects, expectations, and greater instructor attention rather than pedagogical effectiveness per se. (6) Although parallel scheduling and academic-integrity agreements were used to reduce information leakage, cross-group contamination across groups could have occurred within the same school and peer networks, and no formal post hoc contamination checks (eg, social media monitoring or contamination-survey items) were conducted; this could attenuate between-group differences or introduce bias of uncertain direction. (7) As a single-center study, external validity is limited; multicenter replication is needed. (8) We assessed primarily short-term outcomes at module completion and 3-4 weeks thereafter; longer follow-up (eg, 3-6 months) is needed to determine the durability of intervention effects. (9) The case-analysis outcome was aligned with the taught reasoning workflow, so the large effect likely reflects near transfer; the smaller effect on the less-aligned final examination (Cohen d=0.31) suggests that far transfer remains uncertain. (10) Behavioral indicators were exploratory: they were coded by a single nonblinded coder without formal inter-rater reliability assessment, were available for a subsample of students and groups, and repeated observations within classes were not modeled. (11) Every audited GenAI draft required expert correction, and model specifications were only partially recorded, indicating that current GenAI outputs are not suitable for unsupervised classroom deployment and that reporting of implementation details should be improved in future work.

The findings provide preliminary evidence that a GenAI-assisted progressive-disclosure CBL package is feasible and acceptable in occupational health education and is associated with improved performance on a reasoning-aligned assessment and with greater behavioral engagement. Future studies should use a 3-arm design that separates conventional static CBL, instructor-created progressive-disclosure CBL, and GenAI-assisted

XSL•FO
RenderX

progressive-disclosure CBL to isolate the contribution of GenAI; retain class and discussion-group identifiers and use multilevel models or cluster-robust inference with a sufficient number of clusters; include a preintervention measure of clinical reasoning to support change-score or baseline-adjusted analyses; assess long-term retention (eg, at 3, 6, and 12 months) to characterize the retention curve; incorporate explicit measures of premature diagnostic closure, diagnostic switching, and hypothesis revision; and prompt learners in Act 2 to state the source, evidentiary weight, and contradiction check for each conclusion, so that, the frequency and quality of evidence-based justification can be increased and formally assessed.

Conclusion

This study suggests that integrating GenAI with progressive-disclosure CBL offers a feasible and pedagogically meaningful approach to occupational health education. Rather than serving merely as a technical aid, GenAI can support the construction of more authentic, dynamic, and learner-centered clinical scenarios when guided by clear instructional design principles and expert human oversight. This approach has the potential to strengthen students' engagement with complex occupational health problems and promote a more systematic, reflective, and evidence-informed mode of clinical reasoning. In line with the nonrandomized, multicomponent design, these findings should be interpreted as consistent with the hypothesized mechanism rather than as direct evidence of it.

More broadly, the findings highlight the value of human-AI collaboration in reimagining case-based medical education. The educational significance of GenAI lies not in replacing teachers or clinical expertise but in expanding the possibilities for scenario design, multimodal representation, and adaptive learning experiences. Future research should further clarify the mechanisms through which AI-supported instructional designs influence learning, examine their longer-term effects across diverse educational contexts, and develop robust frameworks for ethical, reliable, and discipline-specific implementation.

Acknowledgments

Generative AI tools, including the DeepSeek-V3 model, were used during the preparation of this manuscript for case generation, language polishing, and proofreading. All content, data, analyses, interpretations, references, and conclusions are the original work of the authors, who assume full responsibility for the accuracy and integrity of the entire manuscript. We also wish to thank the faculty and staff of the Department of Occupational and Environmental Health, as well as our graduate students, for their valuable help and support.

Funding

This research was supported by Chongqing Municipal Key Education Reform Project (number 252048), the Scientific Research Project of the Chongqing Association of Higher Education (number cqgj25314C), and the Education and Teaching Reform Project of Chongqing Medical University (number JY20250413).

Data Availability

Deidentified data, analysis code, scoring rubrics, prompt templates, and case materials are available from the corresponding author on reasonable request and will be deposited in a public repository upon acceptance.

Authors' Contributions

Conceptualization: PS
Data curation: MH
Formal analysis: PS
Funding acquisition: PS, MH, CC, SX
Investigation: PS
Methodology: PS
Resources: CC, SX
Supervision: CC, SX
Validation: CC, SX
Visualization: MH
Writing – original draft: PS
Writing – review and editing: MH, CC, SX

Conflicts of Interest

None declared.

Multimedia Appendix 1
Prompt templates.

XSL•FO
RenderX

[DOCX File , 32 KB - mededu_v12i1e103584_app1.docx]

Multimedia Appendix 2
GAMER checklist.
[DOCX File , 29 KB - mededu_v12i1e103584_app2.docx]

Multimedia Appendix 3
Structured self-report questionnaire.
[DOCX File , 20 KB - mededu_v12i1e103584_app3.docx]

References

1. Barrows H. Problem‐based learning in medicine and beyond: a brief overview. N Dir Teaching Learning 2006 Aug 17;1996(68):3-12. [doi: 10.1002/tl.37219966804]
2. Trowbridge RL, Rencic JJ, Durning SJ. Teaching Clinical Reasoning. Philadelphia, PA: American College of Physicians; 2015.
3. Daniel M, Rencic J, Durning SJ, Holmboe E, Santen SA, Lang V, et al. Clinical reasoning assessment methods: a scoping review and practical guidance. Acad Med 2019;94(6):902-912. [doi: 10.1097/ACM.0000000000002618] [Medline: 30720527]
4. Kruglanski AW, Webster DM. Motivated closing of the mind: "seizing" and "freezing". Psychol Rev 1996;103(2):263-283. [doi: 10.1037/0033-295x.103.2.263] [Medline: 8637961]
5. Kassabry M, Al-Kalaldeh M, Ayed A, Abu-Shosha G, Rn P, Salameh B. Impacts of unfolding case-study learning on nursing students' performance outcomes: a scoping review. J Adv Med Educ Prof 2026;14(1):1-20. [doi: 10.30476/jamp.2025.108065.2234] [Medline: 41623995]
6. Schmidt HG, Mamede S. How to improve the teaching of clinical reasoning: a narrative review and a proposal. Med Educ 2015;49(10):961-973. [doi: 10.1111/medu.12775] [Medline: 26383068]
7. van Merriënboer JJG, Sweller J. Cognitive load theory in health professional education: design principles and strategies. Med Educ 2010;44(1):85-93. [doi: 10.1111/j.1365-2923.2009.03498.x] [Medline: 20078759]
8. Abidi SH, Almazan J, Fabiyi O, Zehra F, Tariq M. AI-supported case-based learning in medical education: a comprehensive scoping review. Front Med (Lausanne) 2026;13:1798097 [FREE Full text] [doi: 10.3389/fmed.2026.1798097] [Medline: 41958555]
9. Chamberland M, St-Onge C, Setrakian J, Lanthier L, Bergeron L, Bourget A, et al. The influence of medical students' self-explanations on diagnostic performance. Med Educ 2011;45(7):688-695. [doi: 10.1111/j.1365-2923.2011.03933.x] [Medline: 21649701]
10. Qian C, Gao C, Park S, Gim H, Hou K, Cook B, et al. Use of large language models for rapid quantitative feedback in case-based learning: a pilot study. Med Sci Educ 2025;35(3):1169-1171. [doi: 10.1007/s40670-025-02343-6] [Medline: 40625928]
11. Preiksaitis C, Rose C. Opportunities, challenges, and future directions of generative artificial intelligence in medical education: scoping review. JMIR Med Educ 2023;9:e48785 [FREE Full text] [doi: 10.2196/48785] [Medline: 37862079]
12. Abd-Alrazaq A, AlSaad R, Alhuwail D, Ahmed A, Healy PM, Latifi S, et al. Large language models in medical education: opportunities, challenges, and future directions. JMIR Med Educ 2023;9:e48291 [FREE Full text] [doi: 10.2196/48291] [Medline: 37261894]
13. Gilson A, Safranek CW, Huang T, Socrates V, Chi L, Taylor RA, et al. How does ChatGPT perform on the United States medical licensing examination (USMLE)? The implications of large language models for medical education and knowledge assessment. JMIR Med Educ 2023;9:e45312 [FREE Full text] [doi: 10.2196/45312] [Medline: 36753318]
14. Li Y, Yorke J, Li J, He M, Dai Y, Zhao IY, et al. An innovative socratic method-based artificial intelligence platform for healthcare education: a quasi-experimental study. Nurse Educ Pract 2026;92:104770 [FREE Full text] [doi: 10.1016/j.nepr.2026.104770] [Medline: 41690156]
15. O'Brien BC, Battista A. Situated learning theory in health professions education research: a scoping review. Adv Health Sci Educ Theory Pract 2020;25(2):483-509. [doi: 10.1007/s10459-019-09900-w] [Medline: 31230163]
16. Cutrer WB, Miller B, Pusic MV, Mejicano G, Mangrulkar RS, Gruppen LD, et al. Fostering the development of master adaptive learners: a conceptual model to guide skill acquisition in medical education. Acad Med 2017;92(1):70-75. [doi: 10.1097/ACM.0000000000001323] [Medline: 27532867]
17. Jiang D, Huang D, Wan H, Fu W, Shi W, Li J, et al. Effect of integrated case-based and problem-based learning on clinical thinking skills of assistant general practitioner trainees: a randomized controlled trial. BMC Med Educ 2025;25(1):62 [FREE Full text] [doi: 10.1186/s12909-025-06634-9] [Medline: 39806361]
18. Zhang SL, Ren SJ, Zhu DM, Liu TY, Wang L, Zhao JH, et al. Which novel teaching strategy is most recommended in medical education? A systematic review and network meta-analysis. BMC Med Educ 2024;24(1):1342 [FREE Full text] [doi: 10.1186/s12909-024-06291-4] [Medline: 39574112]

XSL•FO
RenderX

19. Luo X, Tham YC, Giuffrè M, Ranisch R, Daher M, Lam K, GAMER Working Group. Reporting guideline for the use of generative artificial intelligence tools in mEdical research: the GAMER statement. BMJ Evid Based Med 2025;30(6):390-400 [FREE Full text] [doi: 10.1136/bmjebm-2025-113825] [Medline: 40360239]
20. Silvestri-Elmore A, Burton C. How can nursing faculty create case studies using AI and educational technology? Nurse Educ 2025;50(1):35-39. [doi: 10.1097/NNE.0000000000001734] [Medline: 39269731]
21. Turney J, Young TM, Chauhan DR, Beeharry R, Mahmud M. AI use for medical students: impact on clinical skill acquisition and retention. A systematic review. Adv Med Educ Pract 2026;17:583763 [FREE Full text] [doi: 10.2147/AMEP.S583763] [Medline: 42006163]
22. Cox Simms R, Fox AB. Scripted remedies: leveraging AI and pop culture in nursing pharmacology case studies. J Nurs Educ 2025;64(8):511-514. [doi: 10.3928/01484834-20250131-01] [Medline: 40378252]
23. Lave J, Wenger E. Situated Learning: Legitimate Peripheral Participation. United Kingdom: Cambridge University Press; 1991.
24. Cantillon P, De Grave W, Dornan T. Uncovering the ecology of clinical education: a dramaturgical study of informal learning in clinical teams. Adv Health Sci Educ Theory Pract 2021;26(2):417-435 [FREE Full text] [doi: 10.1007/s10459-020-09993-8] [Medline: 32951128]
25. Potter L, Jefferies C. Enhancing communication and clinical reasoning in medical education: building virtual patients with generative AI. Future Healthc J 2024;11:100043. [doi: 10.1016/j.fhj.2024.100043]
26. Hudon A, Phan V, Charlin B, Wittmer R. Teaching clinical reasoning in health care professions learners using AI-generated script concordance tests: mixed methods formative evaluation. JMIR Form Res 2025;9:e76618 [FREE Full text] [doi: 10.2196/76618] [Medline: 41264864]
27. Rodman A, Topol EJ. Is generative artificial intelligence capable of clinical reasoning? Lancet 2025;405(10480):689. [doi: 10.1016/S0140-6736(25)00348-4] [Medline: 40023638]
28. Güvel MC, Kıyak YS, Varan HD, Sezenöz B, Coşkun, Uluoğlu C. Generative AI vs. human expertise: a comparative analysis of case-based rational pharmacotherapy question generation. Eur J Clin Pharmacol 2025;81(6):875-883. [doi: 10.1007/s00228-025-03838-2] [Medline: 40205076]
29. Shalong W, Yi Z, Bin Z, Ganglei L, Jinyu Z, Yanwen Z, et al. Enhancing self-directed learning with custom GPT AI facilitation among medical students: a randomized controlled trial. Med Teach 2025;47(7):1126-1133. [doi: 10.1080/0142159X.2024.2413023] [Medline: 39425996]
30. Çiçek FE, Ülker M, Özer M, Kıyak YS. ChatGPT versus expert feedback on clinical reasoning questions and their effect on learning: a randomized controlled trial. Postgrad Med J 2025;101(1195):458-463 [FREE Full text] [doi: 10.1093/postmj/qgae170] [Medline: 39656920]
31. Rao AS, Esmail KP, Lee RS, Jiang S, Arraiza Carlo B, Gill J, et al. Large language model performance and clinical reasoning tasks. JAMA Netw Open 2026;9(4):e264003 [FREE Full text] [doi: 10.1001/jamanetworkopen.2026.4003] [Medline: 41973425]
32. Schwartzstein RM. Clinical reasoning and artificial intelligence: can AI really think? Trans Am Clin Climatol Assoc 2024;134:133-145. [Medline: 39135584]
33. Kavadella A, Dias da Silva MA, Kaklamanos E, Stamatopoulos V, Giannakopoulos K. Evaluation of ChatGPT's real-life implementation in undergraduate dental education: mixed methods study. JMIR Med Educ 2024;10:e51344 [FREE Full text] [doi: 10.2196/51344] [Medline: 38111256]
34. Han Z, Battaglia F, Udaiyar A, Fooks A, Terlecky SR. An explorative assessment of ChatGPT as an aid in medical education: use it with caution. Med Teach 2024;46(5):657-664. [doi: 10.1080/0142159X.2023.2271159] [Medline: 37862566]
35. Dekerlegand R, Bell A, Clancy MJ, Pletcher ER, Pollen T. Generative artificial intelligence in education: insights from rehabilitation sciences students. Educ Sci 2025;15(3):380. [doi: 10.3390/educsci15030380]

Abbreviations

ANCOVA: analysis of covariance
CBL: case-based learning
GAMER: Generative AI in Medical Research
GenAI: generative AI
GPA: grade point average
HITL: human-in-the-loop
ICC: intraclass correlation coefficient
LLM: large language model
OSHA: Occupational Safety and Health Administration
PIPL: Personal Information Protection Law
PPE: personal protective equipment
SLT: situated learning theory
SMD: standardized mean difference

XSL•FO
RenderX

SOP: standard operating procedure
TREND: Transparent Reporting of Evaluations with Nonrandomized Designs
UCS: unfolding case study

Edited by W Jerjes; submitted 04.Jun.2026; peer-reviewed by NL Ritter, J McMordie; comments to author 26.Jul.2026; revised version received 21.Aug.2026; accepted 26.Aug.2026; published 22.Sep.2026.

Please cite as:
Su P, Hu M, Chen C, Xu S
Generative AI–Assisted Progressive-Disclosure Case-Based Learning for Clinical Reasoning in Occupational Medicine: Quasi-Experimental Study
JMIR Med Educ 2026;12:e103584
URL: https://mededu.jmir.org/2026/1/e103584
doi:10.2196/103584
PMID:

©Peng Su, Min Hu, Chengzhi Chen, Shangcheng Xu. Originally published in JMIR Medical Education (https://mededu.jmir.org), 22.Sep.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Education, is properly cited. The complete bibliographic information, a link to the original publication on https://mededu.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Development of a Deep Learning–Based Feedback Model to Assist Medical Students Learning Renal Ultrasound Acquisition: Mixed Methods Study

Andy Cheuk Nam Hwang[1*], MMedSc; Rahul Singh[1*], PhD; Elizabeth Ann Barrett[2], PhD; Peng Cao[1], PhD; Varut Vardhanabhuti[1], PhD; Pauline Yeung Ng[3], MBBS; Gordon Tin Chun Wong[4], MD; Michael Tiong Hong Co[5], MS; Elaine Yuen-Phin Lee[1], MD

[1]Department of Diagnostic Radiology, School of Clinical Medicine, Li Ka Shing Faculty of Medicine, The University of Hong Kong, 21 Sassoon Rd, Pok Fu Lam, Hong Kong, China

[2]Academic Unit of Human Communication, Learning, and Development, Faculty of Education, The University of Hong Kong, Hong Kong, China

[3]Critical Care Medicine Unit, School of Clinical Medicine, Li Ka Shing Faculty of Medicine, The University of Hong Kong, Hong Kong, China

[4]Department of Anaesthesiology, School of Clinical Medicine, Li Ka Shing Faculty of Medicine, The University of Hong Kong, Hong Kong, China

[5]Department of Surgery, School of Clinical Medicine, Li Ka Shing Faculty of Medicine, The University of Hong Kong, Hong Kong, China

[*]these authors contributed equally

Corresponding Author:
Elaine Yuen-Phin Lee, MD
Department of Diagnostic Radiology, School of Clinical Medicine, Li Ka Shing Faculty of Medicine, The University of Hong Kong, 21 Sassoon Rd, Pok Fu Lam, Hong Kong, China

Abstract

Background: Point-of-care ultrasound training is being increasingly integrated into undergraduate medical education, leading to a substantial demand for trained faculty to provide instruction and feedback.

Objective: This study aimed to develop an adjunct tool, a deep learning–based feedback model, to facilitate student learning.

Methods: Renal ultrasound images (N=2807) were used to train a cascaded deep learning–based feedback model that classified images into three categories: optimal, suboptimal, and incorrect. Suboptimal images were further subcategorized as images with artifact, incorrect gain, and/or incorrect positioning. The model was deployed among year 5 medical students receiving bedside ultrasound training, who were invited to upload renal ultrasound images to an online platform for automated image quality grading and feedback. A mixed methods analysis was used to evaluate students' learning experience. Focus group interviews were organized to qualitatively analyze the successes and challenges of implementation. Quantitative analysis was based on responses to a 5-point Likert scale questionnaire and performance on the objective structured clinical examination (OSCE). Objective structured clinical examination scores were compared with mean OSCE scores from the 2 years preceding implementation of the deep learning–based feedback model.

Results: Focus group interviews identified that the deep learning–based feedback model encouraged self-regulated learning but also recognized that discordant curricular design and hardware limitations impeded its use. The 11-item online questionnaire had a response rate of 42.4% (98/231 students). Among respondents, 32% (31/98) to 48% (47/98) found the model helpful in assisting ultrasound training (Likert score of 4 - 5 for items 1-3), while 49% (48/98) to 76% (74/98) were satisfied with its usability and their interaction with the model (Likert score of 4 - 5 for items 4-11). The mean OSCE score was 9.73 (SD 0.76) out of 10, compared with mean scores of 9.35 (SD 1.03; *P*=.06) and 9.45 (SD 0.97; *P*=.15) out of 10 in the 2 individual years preceding implementation of the model.

Conclusions: A cascaded deep learning–based feedback model was developed and deployed among year 5 medical students receiving bedside ultrasound training, with positive learner responses and enhanced self-regulated learning. The innovation was associated with increased student engagement and improved ultrasound skill acquisition among novice learners.

(JMIR Med Educ 2026;12:e72110) doi:10.2196/72110

KEYWORDS

point-of-care ultrasound; renal ultrasound learning; deep learning; convolutional neural networks; automated feedback system

XSL•FO
RenderX

Introduction

Point-of-care ultrasound (POCUS) refers to the use of ultrasound imaging to facilitate clinical diagnosis and management while patients are being treated. Substantial evidence supports POCUS in aiding diagnosis and improving bedside procedures and clinical management [1-3]. To address this clinical need, the next generation of clinicians involved in acute care should master skills such as image acquisition, interpretation, and clinical integration of POCUS findings. Therefore, ultrasound training is increasingly being introduced and incorporated into undergraduate medical education (UME). In 2019, a survey conducted in the United States reported that 72.6% of 168 accredited medical schools included an ultrasound curriculum in their UME [4]. Cross-specialty POCUS training has been shown to augment physical examination skills among undergraduate learners and to lay the foundation for future postgraduate training [5,6]. However, the demand for trained faculty and tutors remains a major barrier to implementing ultrasound curricula in UME; 63% of medical schools in the United States reported that they did not have trained faculty for POCUS instruction [7].

Ultrasound imaging across different organs has variable learning curves, reflecting different rates of skill acquisition. Among these, renal ultrasound has a relatively longer learning curve [8], suggesting that it is moderately challenging to students. This underscores the need for additional practice and faculty guidance to support skill development. Different pedagogical approaches in renal and abdominal ultrasound training have demonstrated comparable improvements in learner proficiency [9,10]. This suggests that renal ultrasound skill acquisition is adaptable to different pedagogical approaches to support student learning.

Feedback is important to student learning because it facilitates self-reflection, understanding, and future improvement, particularly in skill acquisition and mastery [11]. Effective feedback positions students as active learners in the feedback process, empowering them to understand their performance and develop evaluative judgment to improve learning [12]. Feedback is most beneficial when immediate, external feedback on a specific task is provided [11]. Students benefit from multiple opportunities to engage with feedback from different sources [13,14]. However, feedback is often neither adequately provided nor delivered effectively [15]. This challenge is exacerbated in large-scale higher education [16,17]. A technology-enabled feedback process may streamline practice and address challenges associated with low instructor-to-student ratios and the inability to provide on-demand feedback [18]. It can customize students' learning by allowing them to determine when and where their learning occurs and to control their pace. It may support both blended and adaptive learning strategies and provide a flexible learning environment not limited to workshops and bedside teaching sessions [19-22]. This approach may also promote self-regulated learning [23], an important conceptual framework in education. Self-reflection is an invaluable step in preparing students for the next phase of the learning cycle [24].

A deep learning model based on convolutional neural networks (CNNs) is a promising approach for image classification [25]. A number of pretrained CNNs with strong general performance, such as ResNet [26] and SENet [27], have been developed. These pretrained CNNs can be fine-tuned and trained for classification tasks involving medical images. In this study, we hypothesized that a deep learning–based feedback model could enhance learner motivation and support self-regulated learning in renal ultrasound acquisition. Accordingly, the aims of this study were to develop a deep learning–based feedback model and evaluate its acceptance and impact on renal ultrasound acquisition skills.

Methods

Ethical Considerations

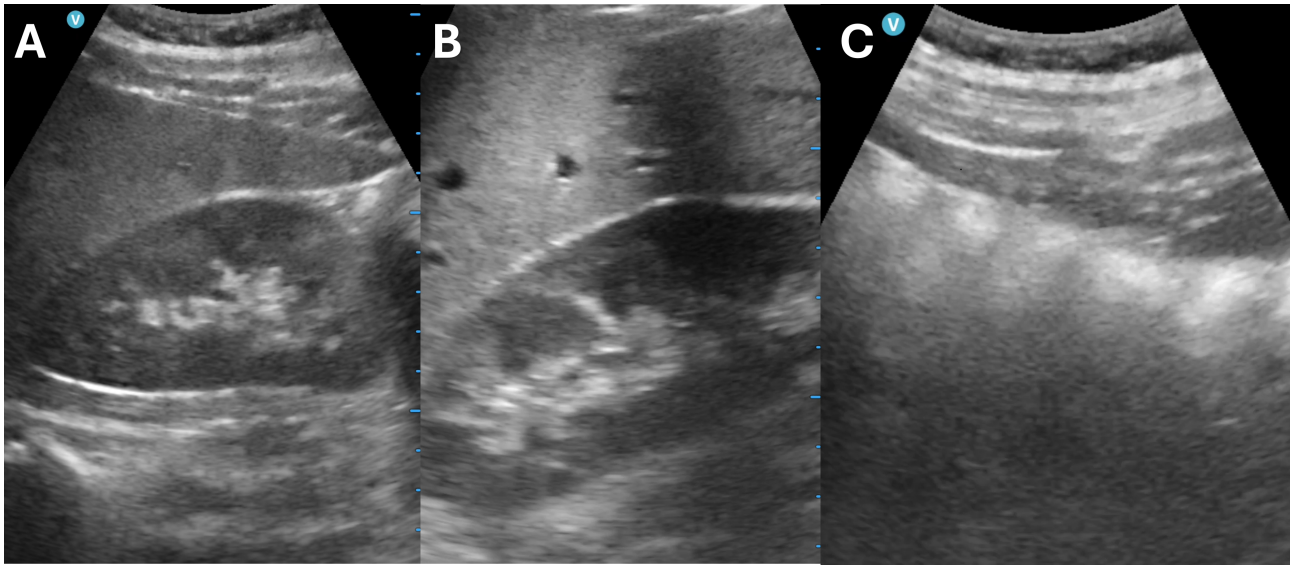
This study was approved by the Institutional Review Board of the University of Hong Kong/Hospital Authority Hong Kong West Cluster (HKU/HA HKW IRB; UW 22 - 797). This study was conducted in accordance with the Declaration of Helsinki and the International Council for Harmonisation Good Clinical Practice guidelines. Students were invited to complete the questionnaire voluntarily and anonymously. All data generated and analyzed in this study were anonymized. The requirement for informed consent was waived by the HKU/HA HKW Institutional Review Board. No compensation was provided.

Development of a Deep Learning Feedback Model

All renal ultrasound images (N=2807) in transverse and longitudinal views were retrospectively retrieved from the local radiology database, anonymized, and used to train the algorithm. All renal ultrasound images were classified into three main categories by a board-certified radiologist with more than 15 years of postfellowship experience. Images were classified as optimal when they were free of artifact and demonstrated appropriate brightness and positioning of the kidney (Figure 1A). Images were classified as suboptimal when they exhibited one or more of the following features: artifact (eg, acoustic shadowing or edge artifact) that obscured visualization of the kidney; incorrect gain (brightness), defined as improper amplification of the ultrasound signal, resulting in either very dark or very bright pixels and a degraded grayscale image; and/or incorrect positioning, in which the image of the kidney was truncated or its contour was unclear (Figure 1B). Images were classified as incorrect when the image showed an incorrect organ or when no kidney was visualized (Figure 1C).

XSL•FO
RenderX

Figure 1. Examples of renal ultrasound images for model training: (A) optimal, (B) suboptimal (artifact and incorrect positioning), and (C) incorrect.



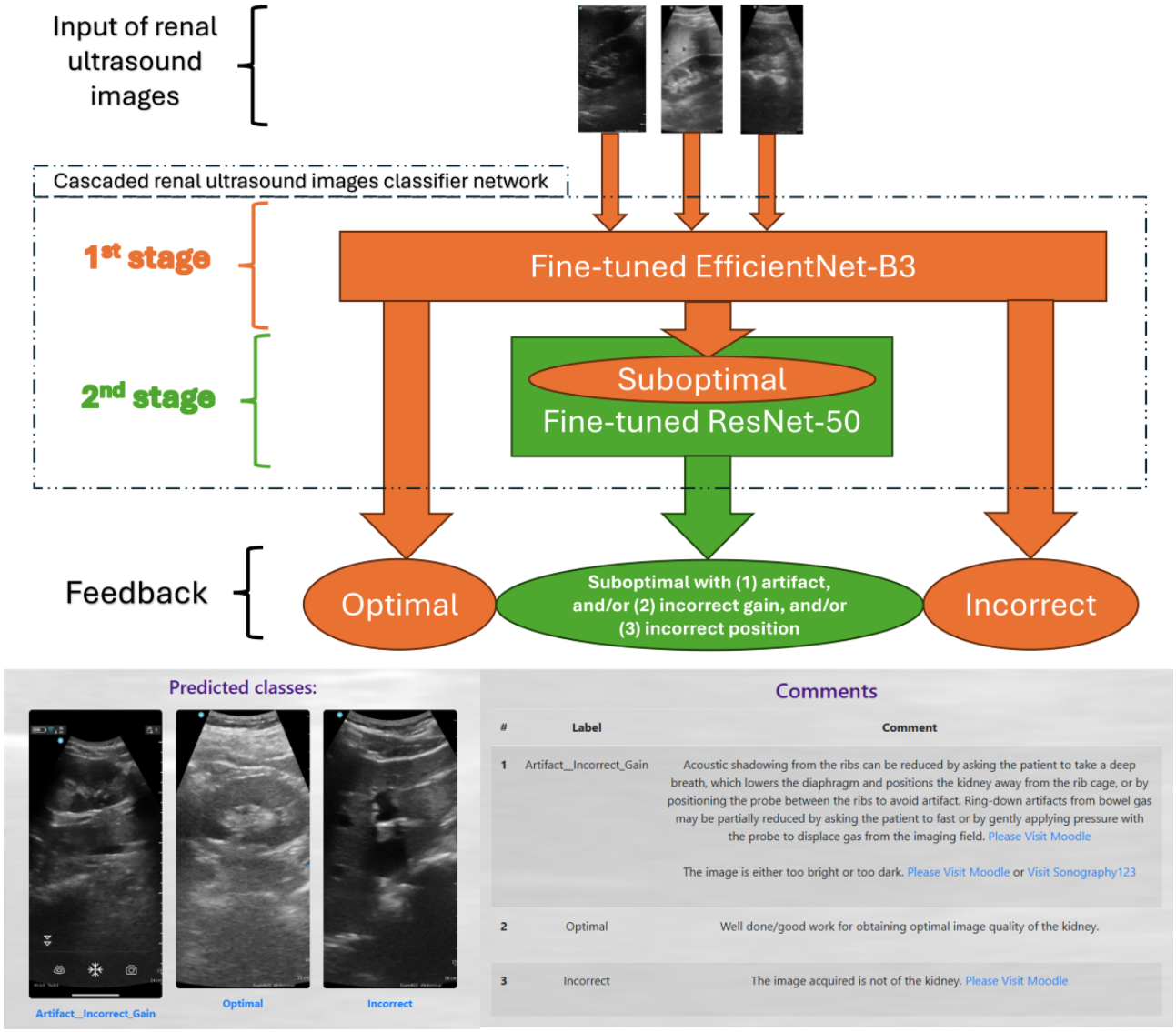
In the training dataset, 562 images were classified as optimal, 1288 as suboptimal, and 957 images as incorrect. Among the 1288 suboptimal images, 200 showed artifact alone, 256 showed incorrect gain alone, 202 showed incorrect position alone, 248 showed artifact and incorrect gain, 106 showed artifact and incorrect positioning, 164 showed incorrect gain and incorrect positioning, and 112 showed all 3 subcategories.

Two-Stage Cascaded Network

All images were preprocessed with pixel resizing (224 × 224 × 1) and *z* score normalization before being used to fine-tune the pretrained CNNs, with an 8:2 split between the training and test sets. The resulting model formed a deep learning–based, fully automated renal ultrasound image grading system (Figure 2). The cascaded classifier network was composed of two stages. In the first stage, a pretrained CNN, EfficientNet-B3 [28], was fine-tuned on all 2807 ultrasound images for the classification task, labeled according to the three previously defined categories. In the second stage, images classified as suboptimal were input into another pretrained CNN, ResNet-50 [26], to further subclassify them into the following subcategories: artifact, incorrect gain, and/or incorrect positioning.

XSL•FO
RenderX

Figure 2. Cascaded renal ultrasound image grading system with automated grading and feedback for students (demonstration video in Multimedia Appendix 1).



Both EfficientNet-B3 and ResNet-50 backbones pretrained on the ImageNet dataset [29] were used and fine-tuned on the renal ultrasound dataset via transfer learning. Subsequently, the trained model was hosted on a website through Microsoft Azure, the cloud computing platform (Figure 2).

Study Cohort and Interventions

The deep learning–based feedback model was deployed to a cohort of novice year 5 medical students receiving training in ultrasound imaging during the 2023-2024 academic year. The students received a 1-hour didactic lecture and a 3-hour face-to-face ultrasound training session with an experienced ultrasound instructor. Students were given access to individual ultrasound handheld devices for practice during a 6-week surgical rotation. Students were encouraged to practice ultrasound scanning with their peers and to submit renal ultrasound images to the online platform during the 6-week surgical rotation. Images captured during these practice sessions were uploaded to the online platform, where the model immediately analyzed the submitted images and provided instant feedback and grading. For suboptimal or incorrect images, the platform provided comments that were cross-referenced to the current teaching material on the e-learning platform (Figure 2). Accordingly, students were able to assess image quality and adjust their image scanning technique for subsequent scans. Students had free access to the platform, with no limit on the number of times they could access it or the number of images they could submit, to encourage use during training.

Mixed Methods Analysis: Qualitative Analysis

Qualitative evaluation was conducted through focus group interviews with selected subgroups within the cohort to gain insight into the successes and challenges of implementing a deep learning–based feedback model in ultrasound training. Two teaching associates were invited to attend the focus group interviews and took notes. These notes were subsequently summarized and circulated among the instructors and teaching associates to ensure crucial points were accurately captured. Thematic analysis was then performed to identify and analyze the pertinent points discussed during the focus group interviews.

XSL•FO
RenderX

Mixed Methods Analysis: Quantitative Analysis

Students were invited to complete an 11-item questionnaire based on a 5-point Likert scale (Multimedia Appendix 2). All items were selected from validated instruments, including the Objective Structured Assessment of Ultrasound Skills [30], the System Usability Scale [31], and the Client Satisfaction Questionnaire-8 [32]. The questionnaire aimed to evaluate the model's effectiveness in supporting ultrasound training (items 1 - 3), its usability (items 4 and 5), and students' experiences interacting with the model (items 6 - 11). Three experienced medical educators skilled in ultrasound teaching (PYN, MTHC, and EYPL) rated each questionnaire item for relevance to ultrasound learning. All items achieved an Item-Content Validity Index of 1, indicating high relevance to ultrasound education. For reliability testing, 30 medical students were invited to complete the questionnaire twice with a 1-week interval. Cronbach α was 0.965, and the intraclass correlation coefficient was 0.971. Both measures demonstrated high questionnaire reliability.

Objective Structured Clinical Examination

At the end of the surgical rotation, an objective structured clinical examination (OSCE) was conducted, which included a station assessing renal ultrasound acquisition skills. Students were evaluated at the standardized OSCE station and instructed to perform POCUS on a healthy volunteer to demonstrate normal renal anatomy. The skills assessed included (1) patient preparation, (2) ultrasound probe selection, (3) probe handling, (4) a systematic approach to examination, and (5) the acquisition of an optimal sonographic image. All students were assessed by a surgical specialist experienced in performing POCUS examinations.

Mean OSCE scores from the 2 academic years preceding implementation of the deep learning–based feedback model (2021 - 2022 and 2022 - 2023) were retrieved and compared with the mean OSCE score of this cohort of students receiving the deep learning–based feedback.

Results

Overview

The cohort of year 5 medical students (n=231) receiving ultrasound training was divided into 6 groups and enrolled between October 2023 and September 2024. A total of 786 renal ultrasound images were submitted to the platform for grading after exclusion of duplicate images. The mean number of images contributed per student was 3.40 (SD 3.32). The mean interval between the first submission and the OSCE was 15.6 (SD 6.04) days. Of the 786 images, 269 images were classified as optimal, 349 as suboptimal, and 168 as incorrect.

Thematic Analysis in Focus Group Interviews

Within the cohort, 2 subgroups were invited for focus group interviews (n=71). Through the interviews conducted in this study, highly motivated students who wanted to be more involved in the project were identified, and these students provided valuable suggestions for model refinement and implementation. The thematic framework identified three major themes: the positive impact of deep learning–based feedback, the challenges of implementation, and suggestions to enhance the feedback model.

Positive Impact

Participants generally agreed that the model motivated their learning and helped improved ultrasound skill acquisition. Students were satisfied with the availability of immediate feedback upon submitting images to the platform, which promoted and encouraged self-regulated learning. Representative participant quotations illustrating the positive impact of the model are provided below.

> *I can know the quality of my images in a short time.* [Participant #16, female]
>
> *The system instantly confirmed if my images were qualified.* [Participant #25, male]
>
> *Great that it can classify my images immediately.* [Participant #42, male]
>
> *It allows me to take more images during practice and upload them later.* [Participant #45, male]

Challenges

Two major challenges were identified: curricular design and hardware provision. Students reported infrequent use of the handheld ultrasound device due to limited tutor guidance at the bedside, which reduced the incentive to practice newly acquired skills. In addition, students found the ultrasound curriculum overwhelming in terms of knowledge load and skills mastery. Usability issues with the handheld devices were also reported. Representative participant quotations illustrating these challenges are presented below.

> *The handheld ultrasound device is different from the one I learned in the lesson.* [Participant #11, female]
>
> *I cannot set-up the software of the handheld ultrasound device on my smartphone.* [Participant #13, male]
>
> *The battery life of the handheld ultrasound device is not long enough.* [Participant #27, male]
>
> *The handheld device needs to be set-up first and it is a bit complicated.* [Participant #52, female]

Student Suggestions

Students suggested that integrating real-time feedback would enhance ease of use and potentially increase learning engagement. Representative participant statements are presented below.

> *Real-time feedback would be more convenient.* [Participant #5, female]
>
> *If it provides real-time feedback, then there is no need to take screenshots and upload images.* [Participant #27, male]

Questionnaire

A response rate of 42.4% was achieved, with 98 of 231 students completing the online questionnaire (Table 1).

XSL•FO
RenderX

Table . Questionnaire and results (n=98).

Questionnaire item	Respondent scores, n (%)				
	1	2	3	4	5
Q1. The grading system assists me in familiarizing myself with the handheld device and its function.[a]	0 (0)	12 (12)	50 (51)	28 (29)	8 (8)
Q2. The grading system assists me in optimizing image quality.[b]	0 (0)	25 (25)	42 (43)	27 (28)	4 (4)
Q3. The grading system assists me in presenting the renal image according to instruction.[c]	1 (1)	15 (15)	35 (36)	43 (44)	4 (4)
Q4. I would like to use this system frequently.[d]	3 (3)	7 (7)	34 (35)	44 (45)	10 (10)
Q5. The system was easy to use.[d]	1 (1)	2 (2)	21 (21)	53 (54)	21 (21)
Q6. The system was consistent with its grading.[d]	2 (2)	4 (4)	44 (45)	42 (43)	6 (6)
Q7. The system offered useful comments.[d]	1 (1)	9 (9)	36 (37)	44 (45)	8 (8)
Q8. The system met my needs.[d]	1 (1)	6 (6)	37 (38)	46 (47)	8 (8)
Q9. The system helped me learn and improve my ultrasound skills.[d]	1 (1)	6 (6)	26 (27)	56 (57)	9 (9)
Q10. I will recommend this system to other peers.[d]	1 (1)	5 (5)	32 (33)	51 (52)	9 (9)
Q11. I am satisfied with the system.[d]	1 (1)	2 (2)	35 (36)	54 (55)	6 (6)

[a]Question 1 response options: 1=unable to operate equipment, 2=limited ability to operate equipment, 3=operates with some experience, 4=confident in operating equipment, 5=familiar with operating equipment.

[b]Question 2 response options: 1=unable to optimize, 2=limited ability to optimize, 3=competent but optimization inconsistently done, 4=confident in optimization with minor inconsistencies, 5=consistent optimization.

[c]Question 3 response options: 1=unable to achieve, 2=occasionally achieve with difficulty, 3=partially achieve, 4=frequently achieve with some consistency, 5=consistently achieve.

[d]Questions 4-11 response options: 1=strongly disagree, 2=disagree, 3=neutral, 4=agree, 5=strongly agree.

When evaluating the feedback model for assisting ultrasound skill acquisition (questions 1 - 3), 32% (31/98) to 48% (47/98) of the respondents rated the items as 4 or higher, indicating that the model helped build confidence in acquiring new skills, including use of the handheld ultrasound device, image optimization, and image presentation according to instructions. A similar proportion of students (35/98, 36% to 50/98, 51%) rated questionnaire items 1 to 3 as 3, suggesting partial improvement in ultrasound acquisition skills through the feedback from the model (Table 1).

Among respondents, at least 55% (54/98) were satisfied with the usability of the model (scores ≥4 for questions 4 and 5) and more than 49% (48/98) had positive experiences interacting with the model on the cloud platform (scores ≥4 for questions 6 - 11). However, 6% (6/98) of students rated the model as inconsistent in its grading (scores 1 or 2 for question 6) and 45% (44/98) were neutral (score 3 for question 6) on this aspect (Table 1).

OSCE Score Comparison

The mean OSCE score in the current cohort was 9.73 (SD 0.76) out of 10, reflecting high performance in renal ultrasound acquisition skills. In the preceding 2 academic years, when a deep learning–based feedback model was not used, mean OCSE

XSL•FO
RenderX

scores were 9.35 (SD 1.03; *P*=.06) out of 10 in the 2021-2022 academic year and 9.45 (SD 0.97; *P*=.15) out of 10 in the 2022-2023 academic year. Although the mean score was higher in the current cohort, the difference was not statistically significant. This finding may reflect the model's role in sustaining students' interest in renal ultrasound learning by providing continuous access to the platform and encouraging active engagement during the clinical placement.

Discussion

Principal Findings

This study developed and deployed a deep learning–based feedback model to assist novice learners in mastering ultrasound acquisition skills. Artificial intelligence (AI) has already substantially influenced many aspects of society, including medical education [33,34]. Several POCUS studies have explored the use of AI in educational design. Medical students using AI-based tools demonstrated improved performance in acquiring cardiac views on echocardiography [35-38]. Artificial intelligence has also enhanced novice performance in measuring left ventricular ejection fraction, achieving diagnostic accuracy comparable to that of cardiologists [39]. The diagnostic performance of inexperienced medical residents or fellows in the evaluation of thyroid nodules was improved with an AI-based computer-assisted diagnostic system [40]. Integration of AI into 3-dimensional and 4-dimensional ultrasound analysis has enhanced fetal facial profiling, contributing to education in prenatal diagnosis [41]. Collectively, these studies highlight the importance of AI in ultrasound education. The findings of this study are consistent with this evidence, suggesting that a deep learning–based feedback model can serve as an effective adjunct to ultrasound learning by providing automated feedback that supports students' self-regulated learning.

The positive experiences students reported while interacting with the model were essential in sustaining continuous interest in learning ultrasound skills. According to the 4-phase model of interest development [42], learners progress from triggered situational interest to maintained situational interest and eventually to emerging and well-developed individual interest. The initial face-to-face ultrasound training may have triggered situational interest, whereas the feedback model may have contributed to maintaining that interest, thereby allowing for individual interest to develop.

The deep learning–generated grading was not intended as a final assessment of learning; rather, ambiguous images were encouraged to be reviewed and discussed with instructors during face-to-face sessions. Feedback model analytics (ie, the number of platform accesses and image submissions) indicated that students actively engaged with the model throughout the surgical rotation. These findings suggest that the feedback model promoted self-regulated learning and allowed students to develop ultrasound skills at their own pace [18].

Based on feedback model analytics, students frequently submitted more than 1 renal ultrasound image to the platform for evaluation, which may indicate that these students perceived the model as helpful in supporting their learning. Overall, sustained engagement with the feedback model suggests that it contributed to fostering novice learners' interest and motivation, and self-regulated their development of ultrasound skills.

Information gathered from the focus group interviews prompted further in-depth discussion among the course instructors and tutors to re-examine the current ultrasound curriculum. Options for streamlining the curriculum into more focused areas are being actively explored. The framework of load reduction [43] may be applied through instructional strategies such as increasing scaffolding and progressively guiding learners toward independent mastery.

To promote student engagement, the benefits of adopting the deep learning–based feedback model and its role in the broader ultrasound curriculum can be communicated to students. Course leaders may also develop reflective tasks or prompts based on principles of self-regulated learning [23] to guide students in reflecting on their learning (ie, self-efficacy, self-monitoring, self-evaluation, adaptive changes) [44]. These reflections can span both face-to-face ultrasound training and interactions with the deep learning–based feedback model during the 6-week surgical rotation. Such integrations may scaffold students' reflection on their ultrasound learning, help close the feedback loop, and encourage feedforward learning [45]. This refinement to the curriculum structure may enhance self-regulated learning and provide a sustainable, iterative process for students to develop their ultrasound skills through integration of the deep learning–based feedback model.

Furthermore, interview findings indicated a need for enhanced tutor guidance. In addition to scaling-up the recruitment of ultrasound instructors and strengthening local training support, a midrotation tutorial or workshop may provide support to students who may be struggling. This approach may enable them to make more effective use of the feedback model and feel empowered to practice their ultrasound skills more during the rotation. Integrating near-peer feedback during the midrotation tutorial may further support struggling students and help bridge the gaps between intensive face-to-face sessions [46].

As noted from the focus group interviews, full realization of the intervention's benefits would likely require real-time integration of the model into the hardware device. Real-time integration could address the need for immediate guidance while also providing instructional scaffolding. However, such development would be more resource intensive, requiring real-time image tracking and continuous feedback from the deep learning–based model. With further research, clinical validation, and implementation, this type of AI application could potentially be adopted in POCUS training and possibly incorporated into the OSCE.

There is also room for further improvement such as incorporating highlighted contours or bounding boxes to delineate the kidney and associated artifacts. This could be achieved by integrating additional deep learning–based segmentation or object detection models.

XSL•FO
RenderX

Limitations

Although the performance of the trained cascaded network was satisfactory for the classification task, there remains room for improvement in its accuracy and efficiency. We are currently prospectively collecting ultrasound data submitted by students to enable continuous model training and refinement, in order to enhance the model's accuracy and relevance. Second, the response rate to the questionnaire was low, introducing a risk of sampling bias [47]. Despite numerous reminders and encouragement from instructors, students who participated in the questionnaire were likely highly motivated and may have been more proficient in self-regulated learning. Future studies that include a broader range of students would provide a clearer understanding of how the feedback model supports both high- and low-achieving learners in developing ultrasound skills. Third, the lack of a preintervention questionnaire limits the strength of inferences regarding the impact of the deep learning–based model on ultrasound learning and precludes detailed assessment of change. However, given the low postintervention response rate, adding a preintervention questionnaire might have increased the risk of attrition bias.

Conclusion

A cascaded renal ultrasound image feedback model was successfully developed and deployed, personalizing the learning experience in medical education and providing on-demand feedback. It was well received by students and supported self-regulated learning. The innovation enhanced student engagement and improved ultrasound skill acquisition among novice learners.

Acknowledgments

No generative artificial intelligence was used in the preparation of this manuscript.

Funding

This research was funded by the 2023 Teaching Development Grant, University of Hong Kong.

Data Availability

The datasets generated or analyzed during this study are not publicly available due to patient and student privacy. Data supporting the study findings can be found in the translated quotations in the Results section.

Authors' Contributions

Co-first authors: ACNH (equal), RS (equal)
Co-corresponding authors: EYPL (equal; email: eyplee77@hku.hk), MTHC (equal; email: mcth@hku.hk), GTCW (equal; email: gordon@hku.hk)
Conceptualization: EYPL
Data curation: ACNH (equal), RS (equal)
Formal analysis: ACNH (equal), RS (equal), EYPL (equal)
Funding acquisition: EYPL (equal), PC (equal), VV (equal), PYN (equal), GTCW (equal), MTHC (equal)
Investigation: ACNH (equal), RS (equal), EYPL (equal)
Methodology: ACNH (equal), RS (equal), EYPL (equal), EAB (equal)
Project administration: ACNH
Resources: EYPL (lead), ACNH (supporting), RS (supporting)
Software: RS
Supervision: EYPL
Validation: ACNH (equal), RS (equal), EYPL (equal)
Visualization: ACNH (equal), RS (equal)
Writing – original draft: ACNH
Writing – review & editing: ACNH (equal), RS (equal), EYPL (equal), PC (supporting), VV (supporting), PYN (supporting), GTCW (supporting), MTHC (supporting), EAB (supporting)

Conflicts of Interest

None declared.

Multimedia Appendix 1
Demonstration video of the cascaded renal ultrasound image grading system.
[MP4 File, 22456 KB - mededu_v12i1e72110_app1.mp4]

Multimedia Appendix 2
Sample questionnaire.

XSL•FO
RenderX

[DOCX File, 19 KB - mededu_v12i1e72110_app2.docx]

References

1. Smallwood N, Dachsel M. Point-of-care ultrasound (POCUS): unnecessary gadgetry or evidence-based medicine? Clin Med (Northfield) 2018 Jun;18(3):219-224. [doi: 10.7861/clinmedicine.18-3-219]
2. Archer J, Beck S. Accuracy and clinical use of biliary point-of-care ultrasound: a retrospective cohort study. Emerg Med Australas 2023 Apr;35(2):218-224. [doi: 10.1111/1742-6723.14099] [Medline: 36196041]
3. Yoshida T, Yoshida T, Noma H, Nomura T, Suzuki A, Mihara T. Diagnostic accuracy of point-of-care ultrasound for shock: a systematic review and meta-analysis. Crit Care 2023 May 25;27(1):200. [doi: 10.1186/s13054-023-04495-6] [Medline: 37231510]
4. Nicholas E, Ly AA, Prince AM, Klawitter PF, Gaskin K, Prince LA. The current status of ultrasound education in United States medical schools. J Ultrasound Med 2021 Nov;40(11):2459-2465. [doi: 10.1002/jum.15633] [Medline: 33448471]
5. Wong CK, Hai J, Chan KYE, et al. Point-of-care ultrasound augments physical examination learning by undergraduate medical students. Postgrad Med J 2021 Jan 1;97(1143):10-15. [doi: 10.1136/postgradmedj-2020-137773]
6. Cheung ACK, Ng PY, Lam RPK, Wong GTC. Cross-specialty point-of-care ultrasound education in The University of Hong Kong. Hong Kong Med J 2024 Jun;30(3):255. [doi: 10.12809/hkmj2411666] [Medline: 38831748]
7. Russell FM, Zakeri B, Herbert A, Ferre RM, Leiser A, Wallach PM. The state of point-of-care ultrasound training in undergraduate medical education: findings from a national survey. Acad Med 2022 May 1;97(5):723-727. [doi: 10.1097/ACM.0000000000004512] [Medline: 34789665]
8. Breunig M, Hanson A, Huckabee M. Learning curves for point-of-care ultrasound image acquisition for novice learners in a longitudinal curriculum. Ultrasound J 2023 Jul 5;15(1):31. [doi: 10.1186/s13089-023-00329-2] [Medline: 37402989]
9. Moga T, Dancu GM, Cotrau R, et al. Learning curves in abdominal ultrasound in medical students. Med Ultrason 2024 Mar 27;26(1):21-25. [doi: 10.11152/mu-4235] [Medline: 38150697]
10. Alerhand S, Choi A, Ostrovsky I, et al. Integrating basic and clinical sciences using point-of-care renal ultrasound for preclerkship education. MedEdPORTAL 2020 Dec 9;16:11037. [doi: 10.15766/mep_2374-8265.11037] [Medline: 33324747]
11. Burgess A, van Diggele C, Roberts C, Mellis C. Feedback in the clinical setting. BMC Med Educ 2020 Dec 3;20(Suppl 2):460. [doi: 10.1186/s12909-020-02280-5] [Medline: 33272265]
12. Carless D. Excellence in University Assessment: Learning from Award-Winning Practice: Routledge; 2015.
13. Tai J, Ajjawi R, Boud D, Dawson P, Panadero E. Developing evaluative judgement: enabling students to make decisions about the quality of work. High Educ 2018 Sep;76(3):467-481. [doi: 10.1007/s10734-017-0220-3]
14. Bienstock JL, Katz NT, Cox SM, et al. To the point: medical education reviews--providing feedback. Am J Obstet Gynecol 2007 Jun;196(6):508-513. [doi: 10.1016/j.ajog.2006.08.021] [Medline: 17547874]
15. Reddy ST, Zegarek MH, Fromme HB, Ryan MS, Schumann SA, Harris IB. Barriers and facilitators to effective feedback: a qualitative analysis of data from multispecialty resident focus groups. J Grad Med Educ 2015 Jun;7(2):214-219. [doi: 10.4300/JGME-D-14-00461.1] [Medline: 26221437]
16. Henderson M, Phillips M, Ryan T, et al. Conditions that enable effective feedback. High Educ Res Dev 2019 Nov 10;38(7):1401-1416. [doi: 10.1080/07294360.2019.1657807]
17. Henderson M, Ryan T, Phillips M. The challenges of feedback in higher education. Assess Eval High Educ 2019 Nov 17;44(8):1237-1252. [doi: 10.1080/02602938.2019.1599815]
18. Fuller R, Goddard VCT, Nadarajah VD, et al. Technology enhanced assessment: Ottawa consensus statement and recommendations. Med Teach 2022 Aug;44(8):836-850. [doi: 10.1080/0142159X.2022.2083489] [Medline: 35771684]
19. Cutrer WB, Miller B, Pusic MV, et al. Fostering the development of master adaptive learners: a conceptual model to guide skill acquisition in medical education. Acad Med 2017;92(1):70-75. [doi: 10.1097/ACM.0000000000001323]
20. Pickering JD. Anatomy drawing screencasts: enabling flexible learning for medical students. Anat Sci Educ 2015;8(3):249-257. [doi: 10.1002/ase.1480] [Medline: 25091417]
21. Fitzgerald DA, Scott KM, Ryan MS. Blended and e-learning in pediatric education: harnessing lessons learned from the COVID-19 pandemic. Eur J Pediatr 2022 Feb;181(2):447-452. [doi: 10.1007/s00431-021-04149-1] [Medline: 34322730]
22. Challis M. AMEE medical education guide no. 19: personal learning plans. Med Teach 2000 Jan;22(3):225-236. [doi: 10.1080/01421590050006160]
23. Zimmerman BJ. Becoming a self-regulated learner: an overview. Theory Pract 2002 May;41(2):64-70. [doi: 10.1207/s15430421tip4102_2]
24. Panadero E. A review of self-regulated learning: six models and four directions for research. Front Psychol 2017;8(422):422. [doi: 10.3389/fpsyg.2017.00422] [Medline: 28503157]
25. Krizhevsky A, Sutskever I, Hinton GE. ImageNet classification with deep convolutional neural networks. Commun ACM 2017 May 24;60(6):84-90. [doi: 10.1145/3065386]
26. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. Presented at: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR); Jun 27-30, 2016. [doi: 10.1109/CVPR.2016.90]
27. Hu J, Shen L, Sun G. Squeeze-and-excitation networks. Presented at: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition; Jun 18-23, 2018. [doi: 10.1109/CVPR.2018.00745]

XSL•FO
RenderX

28. Tan M, Le QV. EfficientNet: rethinking model scaling for convolutional neural networks. Proc Machine Learning Res 2019;97:6105-6114 [FREE Full text]
29. Deng J, Dong W, Socher R, Li LJ. ImageNet: a large-scale hierarchical image database. Presented at: 2009 IEEE Conference on Computer Vision and Pattern Recognition; Jun 20-25, 2009. [doi: 10.1109/CVPR.2009.5206848]
30. Tolsgaard MG, Todsen T, Sorensen JL, et al. International multispecialty consensus on how to evaluate ultrasound competence: a Delphi consensus survey. PLoS One 2013;8(2):e57687. [doi: 10.1371/journal.pone.0057687] [Medline: 23469051]
31. Brooke J. SUS. SUS – a quick and dirty usability scale. In: Jordan PW, Thomas B, McClelland IL, Weerdmeester B, editors. Usability Evaluation in Industry, 1st edition 1996:189-194.
32. Larsen DL, Attkisson CC, Hargreaves WA, Nguyen TD. Assessment of client/patient satisfaction: development of a general scale. Eval Program Plann 1979;2(3):197-207. [doi: 10.1016/0149-7189(79)90094-6] [Medline: 10245370]
33. Tolsgaard MG, Pusic MV, Sebok-Syer SS, et al. The fundamentals of artificial intelligence in medical education research: AMEE Guide No. 156. Med Teach 2023 Jun;45(6):565-573. [doi: 10.1080/0142159X.2023.2180340] [Medline: 36862064]
34. Gordon M, Daniel M, Ajiboye A, et al. A scoping review of artificial intelligence in medical education: BEME guide no. 84. Med Teach 2024 Apr;46(4):446-470. [doi: 10.1080/0142159X.2024.2314198] [Medline: 38423127]
35. Gohar E, Herling A, Mazuz M, et al. Artificial intelligence (AI) versus POCUS expert: a validation study of three automatic AI-based, real-time, hemodynamic echocardiographic assessment tools. J Clin Med 2023 Feb 8;12(4):36835888. [doi: 10.3390/jcm12041352] [Medline: 36835888]
36. Aronovitz N, Hazan I, Jedwab R, et al. The effect of real-time EF automatic tool on cardiac ultrasound performance among medical students. PLoS One 2024;19(3):e0299461. [doi: 10.1371/journal.pone.0299461] [Medline: 38547257]
37. Baum E, Tandel MD, Ren C, et al. Acquisition of cardiac point-of-care ultrasound images with deep learning: a randomized trial for educational outcomes with novices. Chest Pulmonary 2023 Sep 23;1(3):100023. [doi: 10.1016/j.chpulm.2023.100023]
38. Soliman-Aboumarie H, Geers J, Lowcock D, et al. Artificial intelligence-assisted focused cardiac ultrasound training: a survey among undergraduate medical students. Ultrasound 2025 May;33(2):123-128. [doi: 10.1177/1742271X241287923] [Medline: 39555149]
39. Dadon Z, Orlev A, Butnaru A, et al. Empowering medical students: harnessing artificial intelligence for precision point-of-care echocardiography assessment of left ventricular ejection fraction. Int J Clin Pract 2023:5225872. [doi: 10.1155/2023/5225872] [Medline: 38078051]
40. Lee SE, Kim HJ, Jung HK, et al. Improving the diagnostic performance of inexperienced readers for thyroid nodules through digital self-learning and artificial intelligence assistance. Front Endocrinol (Lausanne) 2024;15:1372397. [doi: 10.3389/fendo.2024.1372397] [Medline: 39015174]
41. Bachnas MA, Andonotopo W, Dewantiningrum J, Adi Pramono MB, Stanojevic M, Kurjak A. The utilization of artificial intelligence in enhancing 3D/4D ultrasound analysis of fetal facial profiles. J Perinat Med 2024 Nov 26;52(9):899-913. [doi: 10.1515/jpm-2024-0347] [Medline: 39383043]
42. Hidi S, Renninger KA. The four-phase model of interest development. Educ Psychol 2006 Jun;41(2):111-127. [doi: 10.1207/s15326985ep4102_4]
43. Martin AJ, Evans P. Load reduction instruction: exploring a framework that assesses explicit instruction through to independent learning. Teach Teach Educ 2018 Jul;73:203-214. [doi: 10.1016/j.tate.2018.03.018]
44. Leggett H, Sandars J, Roberts T. Twelve tips on how to provide self-regulated learning (SRL) enhanced feedback on clinical performance. Med Teach 2019 Feb;41(2):147-151. [doi: 10.1080/0142159X.2017.1407868] [Medline: 29228830]
45. Fuentes-Cimma J, Sluijsmans D, Riquelme A, et al. Designing feedback processes in the workplace-based learning of undergraduate health professions education: a scoping review. BMC Med Educ 2024 Apr 23;24(1):440. [doi: 10.1186/s12909-024-05439-6] [Medline: 38654360]
46. Sader J, Cerutti B, Meynard L, et al. The pedagogical value of near-peer feedback in online OSCEs. BMC Med Educ 2022 Jul 25;22(1):572. [doi: 10.1186/s12909-022-03629-8] [Medline: 35879752]
47. Spooner M, Pawlikowska T. Feedback literacy as a model to explore how learners respond to feedback. Br J Hosp Med 2023 Jul 2;84(7):1-9. [doi: 10.12968/hmed.2022.0446]

Abbreviations

AI: artificial intelligence
CNN: convolutional neural network
OSCE: objective structured clinical examination
POCUS: point-of-care ultrasound
UME: undergraduate medical education

XSL•FO
RenderX

Edited by T Leung; submitted 04.Feb.2025; peer-reviewed by CC Jenq, C Ma, JY Meuwly, JMI Arockiasamy; accepted 31.Jan.2026; published 09.Mar.2026.

Please cite as:
Hwang ACN, Singh R, Barrett EA, Cao P, Vardhanabhuti V, Ng PY, Wong GTC, Co MTH, Lee EYP
Development of a Deep Learning–Based Feedback Model to Assist Medical Students Learning Renal Ultrasound Acquisition: Mixed Methods Study
JMIR Med Educ 2026;12:e72110
URL: https://mededu.jmir.org/2026/1/e72110
doi:10.2196/72110

© Andy Cheuk Nam Hwang, Rahul Singh, Elizabeth Ann Barrett, Peng Cao, Varut Vardhanabhuti, Pauline Yeung Ng, Gordon Tin Chun Wong, Michael Tiong Hong Co, Elaine Yuen-Phin Lee. Originally published in JMIR Medical Education (https://mededu.jmir.org), 9.Mar.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Education, is properly cited. The complete bibliographic information, a link to the original publication on https://mededu.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Mobile Learning in Medical Education: Quasi-Experimental Realist Evaluation of Usage, Context, and Examination Performance in a Curricular Setting

Jan Peter Herbert Holzhäuer[1]; Joy Backhaus[1], MSc; Thiên-Tri Lâm[2], MME, MD; Sarah König[1], MME, MD

[1]Institute of Medical Teaching and Medical Education Research, University Hospital Würzburg, Josef-Schneider-Str. 2/D6, Würzburg, Germany
[2]Institute for Hygiene and Microbiology, University of Würzburg, Würzburg, Germany

Corresponding Author:
Sarah König, MME, MD
Institute of Medical Teaching and Medical Education Research, University Hospital Würzburg, Josef-Schneider-Str. 2/D6, Würzburg, Germany

Abstract

Background: Mobile learning (mLearning) is widely used in medical education. Previous research has focused on technology acceptance and intervention effectiveness, but rarely on their integration. Using realist evaluation, this study investigated the conditions under which mLearning is adopted and associated with learning-related outcomes in an authentic curricular setting.

Objective: This study aimed to examine how learner context and engagement patterns shape mLearning use and outcomes, while secondarily contextualizing its association with examination performance.

Methods: A quasi-experimental study was conducted among fifth-semester undergraduate medical students at a German medical school across 2 consecutive summer semesters (2023 and 2024). Students were offered a voluntary, app-based mLearning course in microbiology, delivered via the eSquirrel platform. The course comprised interactive tasks, incorporating elements of gamification and spaced-repetition features. Data sources included nonreactive in-app usage logs, baseline academic performance data, demographic information, and postsemester questionnaire responses. Usage profiles were derived using cluster analysis. Context-mechanism-outcome patterns were explored by relating app usage status to academic performance and survey responses.

Results: Of the 245 eligible students, 220 (89.8%) participated in the study; 110 (50%) used the app. In 2024, app users (n=64, 58%) initially appeared to outperform nonusers (n=46, 42%) in the oral microbiology examination (mean grade 2.3, SD 1.1 vs 2.8, SD 1.3; $t_{63,0}$=1.90; 1-sided P=.03). After adjustment, these differences were largely explained by baseline academic performance, with only limited evidence of an independent intervention effect. Cluster analysis of app users identified 3 engagement profiles: continuous low-intensity use (n=60, 54%), increased use before the examination (n=31, 28%), and use at the beginning of the semester (n=19, 17%). Cluster 2 reported the greatest enjoyment, satisfaction, perceived learning gains, and examination performance in microbiology.

Conclusions: Nonreactive in-app usage data provided valuable insights into student engagement. The effectiveness of mLearning was not universal. Examination-oriented use, associated with more strategic and self-regulated study behavior, was linked to more favorable learning outcomes. Future research needs to address equity concerns, as higher-performing students tended to benefit most, as well as explore adaptive, context-sensitive approaches to support diverse learners.

(JMIR Med Educ 2026;12:e85892) doi:10.2196/85892

KEYWORDS

medical education; mobile learning; realist evaluation; usage patterns; digital learning tools; technology acceptance; student engagement; educational technology; context factors; equity; artificial intelligence; AI

Introduction

Mobile learning (mLearning) has become a prominent but also debated concept in technology-enhanced education. Definitions of mLearning remain inconsistent and have shifted alongside changing technological and pedagogical paradigms [1-6]. In the early 2000s, mLearning was mostly defined as “learning anytime and anywhere using mobile devices” [7-12]. This device-oriented view initially aimed to distinguish mobile from electronic learning and other stationary formats. With the widespread adoption of smartphones, early visions of an all-in-one mLearning device, always connected, interactive, and multimedia-capable, started to become a reality [13,14]. More recent definitions are increasingly theory-based approaches and adopt learner-centered perspectives. They emphasize context-awareness, personalization, and continuity between physical and digital learning spaces [4,13-24]. In this view,

XSL•FO
RenderX

mLearning is no longer merely a technological mode of content delivery, now reflecting the educational dynamics of learning in increasingly mobile, digitally mediated spaces [25-27]. Accordingly, mLearning describes how learners engage with educational content across contexts, locations, and temporal boundaries [20,28-31]. This aligns with contemporary, postpandemic developments in technology-enhanced education research, which increasingly emphasize the social, interactive, and relational dimensions of technology-learner interaction, including artificial intelligence (AI)–supported learning environments [32].

Notably, some scholars argue that mLearning also includes traditional learning activities such as reading a textbook while commuting [28,33-35]. At the same time, rapid technological innovation challenges the generalizability of empirical findings [35]. What was once counted as mLearning, for example receiving an SMS text message from a teacher [14], differs markedly from today's AI-powered mobile apps, augmented reality tools, or social robots [32]. The contexts in which mLearning has been evaluated are highly diverse [14,15,36]: Existing studies have made valuable contributions by examining a wide range of interventions across educational levels, settings, and technologies. However, the substantial heterogeneity in intervention types, instructional content, outcome measures, and study populations makes it difficult to draw generalizable conclusions. Although the COVID-19 pandemic marked a major surge in mLearning research, it also exposed persistent conceptual and methodological fragmentation in the field, as documented in recent bibliometric syntheses [37]. This fragmentation underscores the need for context-sensitive approaches that move beyond average effects to examine mLearning in authentic curricular settings.

Although much of the literature is optimistic [38-43], mLearning has not delivered the anticipated educational transformation [44]. Accordingly, more context-sensitive research is needed to understand how and under what conditions mLearning can be effectively integrated into higher education [45]. This challenge is particularly pressing in undergraduate medical education [46], in which curricula are shifting toward a hybrid and individualized format and in a broader move from knowledge transmission to competency-based learning [47,48]. As students are expected to develop professional competencies rather than merely acquire facts, a didactic gap emerges: foundational theoretical knowledge must be acquired through self-directed learning in informal contexts. mLearning may help bridge this gap by supporting learners' self-directed and autonomous learning, and by linking theoretical knowledge with the development of applied competencies [49-51]. By providing content access "anytime and anywhere," it can generate informal learning episodes that complement clinical training and enhance readiness for professional practice. Consistently, informal learning spaces are gaining importance [52]. Accordingly, mLearning can contribute to more integrated, context-sensitive learning trajectories in medical education [53].

Within mLearning research, 2 major lines of inquiry have received particular attention [54]: (1) studies examining the overall effectiveness of mLearning interventions and (2) studies investigating psychological and contextual factors influencing learners' uptake. However, these perspectives have rarely been integrated in authentic curricular settings. As a result, there is limited understanding of how learners engage with mobile tools, how contextual factors shape this engagement, and how engagement relates to academic outcomes. To address this gap, we applied the realist evaluation (RE) framework [55] to examine the complex interplay. Using the context-mechanism-outcome (CMO) model [56,57], we conceptualize mLearning as a contingent process in which learner-specific contexts (C) activate engagement mechanisms (M) that, in turn, have measurable outcomes (O).

Driven by an initial program theory, we assumed that heterogeneous curricular contexts shape how learners engage with the intervention. We hypothesized that structured mLearning resources activate different motivational and self-regulatory processes depending on prior academic performance and learner characteristics.

Accordingly, this study pursued 2 main objectives with a predominantly exploratory orientation: (1) to examine whether voluntary participation in mLearning was associated with examination performance within a routine medical curriculum and (2) to explore how learner contexts influenced engagement patterns by deriving usage profiles from in-app engagement data and relating them to contextual variables and academic outcomes.

Methods

Evaluation Setting and Methodical Approach

An RE design [53] was used to identify CMO patterns under routine curricular settings. Accordingly, no randomization or blinding was applied. The evaluation was conducted at a German medical school (6-year curriculum). Within the faculty-wide digitalization initiative WueMedica, a general mLearning module was developed and implemented across multiple disciplines. In this study, we focused on its use in the fifth-semester microbiology course, comprising bacteriology, parasitology, mycology, and virology. The subject-specific component of the module was delivered via the eSquirrel app (eSquirrel GmbH). In addition to a voluntary lecture series and mandatory weekly laboratory sessions, mLearning was offered as a supplementary learning resource. Teaching staff delivered the standard curriculum independently, and no direct contact occurred between the study team and students.

Intervention

The intervention consisted of a voluntary app-based course designed to support examination preparation alongside the curriculum. Content was developed by trained student assistants in collaboration with faculty experts and was based on official course materials, the Nationaler Kompetenzbasierter Lernzielkatalog Medizin (the German national competency-based learning objectives catalog) [58], and standard textbooks. The app-based course was organized into chapters with interactive tasks (such as multiple-choice, matching, and fill-in-the-blank formats) and supplemented by images and links to relevant textbook chapters. Gamification features such as point scoring, a leaderboard, push notifications,

XSL•FO
RenderX

and an adaptive spaced-repetition algorithm were included. The course comprised 88 chapters in 2023 and was expanded to 138 in 2024.

Ethical Considerations and Data Protection

The study was approved by the local independent review board (ref 20230217 01 and ref 202580-ka) and conducted in accordance with the European Union General Data Protection Regulation. All study data were stored on secure institutional servers of the University Hospital. The app complied with General Data Protection Regulation requirements and supported cross-platform and offline use.

Students were informed about the study during an in-person session at the beginning of the semester and provided written informed consent. Consent included permission to link the pseudonymized student identifiers with examination results and study data temporarily. Access to the mLearning app was provided via the Moodle-based institutional learning management system (Moodle Pty Ltd, West Perth, Australia) and required an official university email address. Student IDs were used solely for data linkage purposes and removed from the merged dataset prior to analysis. Data from students who used the app but did not consent to participation were excluded. It was determined in 2023 that multiple registrations were technically possible, but these were effectively prevented in 2024 through email verification and ID matching. Participation was voluntary and had no impact on students' courses, assessments, or academic progression.

Study Design and Data Sources

The evaluation spanned 2 consecutive summer semesters (2023 and 2024). Eligible participants were students enrolled in the microbiology course and taking the oral examination at the end of the course for the first time. Repeat examinees or students without formal registration were excluded.

Data were collected from four sources:

1. App usage data (nonreactive) as in-app tasks within chapters representing activity (chapters completed), task success (points earned), and learning efficiency (chapters per unit time).
2. An online questionnaire administered at the end of the semester via EvaSys (evasys GmbH, Lüneburg, Germany) assessing demographics (age and gender), app usage, and learning-related characteristics; 19 items adapted from past studies (the unified theory of acceptance and use of technology [59] and the Revised Two-Factor Study Process Questionnaire [60], as well as studies on technology-enabled formative assessment in medical education [61] and the use of unsupervised online quizzes in medical physiology [62]); and open-ended questions on user satisfaction. The questionnaire was originally developed in German and then translated into English.
3. End-of-course oral microbiology examination grades in 2024, obtained at the end of the fifth semester; grades ranged from 1 (very good) to 5 (failed).
4. Percentage scores from the written part of the German first national licensing examination (Staatsexamen M1) after the fourth semester, served as an indicator of prior academic performance.

In 2023, 117 students qualified for participation. On the basis of an expected 80% (94/117) participation rate, the study was powered to detect a moderate effect size (Cohen d=0.58) at α=.05 with 80% power.

Data Analysis

Standardized usage and demographic variables (z-scores, calculated as z=(X–M)/SD, X as individual values, means, and SDs) were entered into a 2-step cluster analysis (log-likelihood distance and the Bayesian information criterion) to identify user profiles. Students were categorized according to app usage status (nonusers vs app users). App users were further grouped into usage clusters based on engagement patterns. All groups were compared with respect to survey responses and performance outcomes.

Survey data were analyzed using principal component analysis with pairwise deletion for missing values. Components with eigenvalues >1 were extracted and subjected to oblique (Oblimin) rotation. Reliability was assessed with Cronbach α, and component scores were linearly rescaled to a 1-to-5 scale to ease interpretation. Variance contributions of the final scales were approximated using regression-based scores from the unrotated solution.

For the 2024 data, an analysis of covariance was conducted to test for any potential intervention effect. The written M1 examination score as a percentage was used as the academic baseline, and the analysis examined whether oral microbiology examination results deviated significantly from this baseline. Data from 2023 were not included due to methodological limitations. These included incomplete coverage of exam-relevant chapters in the app and the inability to identify app users clearly.

All analyses were conducted using SPSS Statistics (version 29.0.0.0; IBM Corp) at a 95% significance level, applying Welch t tests (2-tailed) for unequal variances [63,64] and standard t tests when variances were equal, with 2-sided P values reported unless otherwise specified.

Reporting followed the CONSORT-EHEALTH (Consolidated Standards of Reporting Trials of Electronic and Mobile Health Applications and Online Telehealth V.1.6.1; Checklist 1) checklist for mobile health interventions [65] and the RAMESES-II (Realist and Meta-narrative Evidence Syntheses: Evolving Standards-2; Checklist 2) standards for REs [66].

Results

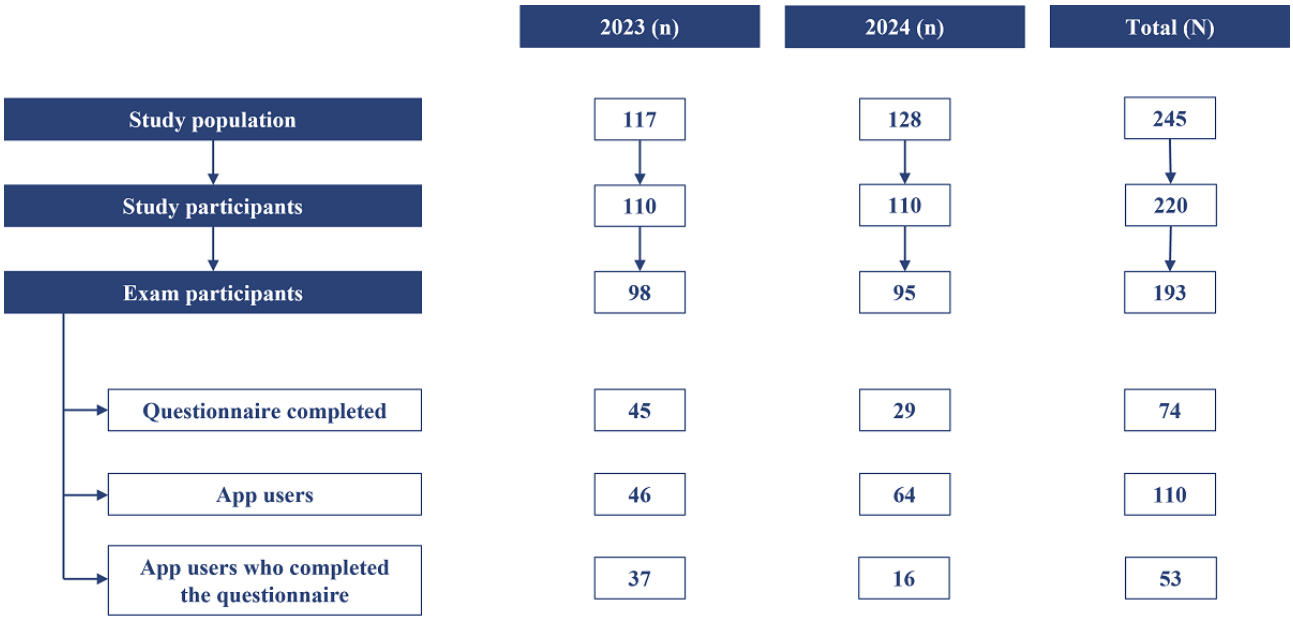
Participant Flow and Sample Characteristics

There were 245 eligible students, of whom 220 (89.8%) participated. Participant recruitment and data availability are summarized in Figure 1. Participants were categorized according to user status, with 110 app users and 110 nonusers. The survey produced 74 valid responses out of 220 participants, corresponding to an overall response rate of 33.6%. In total, 53 participants (24.1%) both used the app and completed the survey. Demographic characteristics were representative of the

XSL•FO
RenderX

student population (study participants: 70% female; mean age 23.1 years, SD 3.1 years and overall study population: 69% female; mean age 23.9 years, SD 3.3 years). No significant differences were observed between semesters. Baseline characteristics of participants are provided in Multimedia Appendix 1.

Figure 1. Flowchart of participant recruitment and data availability.



App Usage Clusters

Cluster analysis (average silhouette coefficient=0.40) of the 110 app users identified 3 usage clusters that differed primarily in the timing and intensity of engagement throughout the semester. Figure 2 illustrates the corresponding longitudinal z-standardized engagement trajectories, with user performance defined as successfully completed in-app tasks within chapters.

XSL•FO
RenderX

Figure 2. Longitudinal z-standardized engagement trajectories (completed in-app tasks within chapters) across app-user clusters throughout the semester.

XSL•FO
RenderX

Cluster analysis revealed three distinct usage patterns:

- Cluster 1 ("continuous low-intensity use," n=60) exhibited regular activity across the semester (active on 66.6% of days) but low content completion (26.5% of in-app tasks in chapters), with engagement largely confined to rare content repetitions (78.4% of app chapters repeated only once) and the lowest task success rate (11.6%).
- Cluster 2 ("increased use before the examination," n=31) combined a lower proportion of active days (48.1% of days) with high content completion (80.9%), a more balanced distribution across content repetitions, and moderate task success (45.4%), with engagement intensifying markedly shortly before the examination.
- Cluster 3 ("use at the beginning of the semester," n=19) demonstrated the highest overall engagement (active on 87.2% of days), substantial content completion (67.3%), and the highest content repetition intensity (35.6% of app chapters repeated 4 times), but engagement peaked early and declined approximately 4 weeks before the examination period.

No significant differences in age or gender were observed across clusters.

The bar charts illustrate changes in standardized performance (*z* scores) over the course of the semester. Positive values indicate above-average gains relative to all users; negative values indicate below-average gains. For example, between days 39 and 37 before the examination, cluster 3 users achieved gains around 0.75 SD above the mean.

Oral Examination Performance According to App Usage Status

For the 2024 cohort, end-of-course oral examination results were available for 95 participants; 15 students (13.6%) were excused for medical reasons. Grades ranged from 1 (excellent) to 5 (failed; Table 1). App users (n=64) achieved better examination results than nonusers (n=46; mean grade 2.3, SD 1.1 vs 2.8, SD 1.3; 1-sided Welch $t_{63,0}$=1.90; *P*=.03). Among app users, cluster 2 (n=20) presented the best performance (mean 2.1, SD 0.9; *P*=.008 vs nonusers), whereas cluster 1 (n=34; mean 2.4, SD 1.1; *P*=.09) and cluster 3 (n=10; mean 2.5, SD 1.3; *P*=.26) did not differ significantly from nonusers. A higher proportion of nonusers than users were excused from the examination (24% vs 6%). No statistically significant differences were observed between the 3 app-user clusters.

Table . Oral examination grades in microbiology stratified according to app usage status and usage clusters (N=110).

	Total	Nonusers (n=46)	App users			
			Total (n=64)	Cluster 1 (n=34)	Cluster 2 (n=20)	Cluster 3 (n=10)
Results, mean (SD)	2.5 (1.2)	2.8 (1.3)	2.3 (1.1)	2.4 (1.1)	2.1 (0.9)	2.5 (1.3)
Grade 1, n (%)	19 (17)	5 (11)	14 (22)	7 (21)	5 (25)	2 (20)
Grade 2, n (%)	36 (33)	12 (26)	24 (38)	12 (35)	8 (40)	4 (40)
Grade 3, n (%)	21 (19)	7 (15)	14 (22)	8 (24)	4 (20)	2 (20)
Grade 4, n (%)	12 (11)	7 (15)	5 (8)	3 (9)	1 (5)	1 (10)
Failed, n (%)	7 (6)	4 (9)	3 (5)	2 (6)	N/A[a]	1 (10)
Excused, n (%)	15 (14)	11 (24)	4 (6)	2 (6)	2 (10)	N/A

[a]N/A: not applicable.

Questionnaire Results: Psychometric Properties and Scale Structure

Sampling adequacy was acceptable (Kaiser-Meyer-Olkin test=0.750; Bartlett test *P*<.001). Principal component analysis initially identified 6 scales explaining 74.3% of the variance. Scales 3 to 5 were excluded due to low internal consistency (Cronbach α=0.46-0.09) and content overlap. Scale 1 was divided into *learning outcome* and *user experience* in line with the theoretical framework [67,68], with 2 items reassigned to enhance conceptual fit. Scale 2 was retained without modification. Three final scales were identified: *learning outcome* (perceived gains in learning efficiency, comprehension, and long-term retention), *user experience* (usability, enjoyment, motivational impact, feedback on progress, curricular relevance, and willingness to recommend), and *study motivation* (general motivation, avoidance of superficial learning, openness to engaging topics, and preference for self-testing) with item loadings depicted in Table 2. Together, the 3 scales required no further rotation and explained 50.2% of the total variance (Table 3).

XSL•FO
RenderX

Table . Principal component analysis of questionnaire data: item loadings and internal consistency.

Item	Item loading		
	Learning outcome	User experience	Study motivation
4.1: mLearning with WueMedica enhances my learning outcomes.	0.92	—[a]	—
4.2: mLearning with WueMedica makes my studying more effective.	0.92	—	—
4.3: mLearning with WueMedica improves my understanding of the learning topics.	0.80	—	—
4.4: mLearning with WueMedica helps me retain what I have learned for a longer period.	0.59	—	—
5.1: mLearning with WueMedica is intuitive.	—	0.57	—
5.2: mLearning with WueMedica excites me.	—	0.87	—
5.3: mLearning with WueMedica motivates me to study more than I did before.	—	0.60	—
5.4: mLearning with WueMedica is enjoyable.	—	0.70	—
5.5: mLearning with WueMedica provides me with feedback on my learning progress.	—	0.49	—
5.6: mLearning with WueMedica helps me identify my strengths and weaknesses in learning.	—	0.46	—
5.7: mLearning with WueMedica adds value to the curricular teaching.	—	0.82	—
6.1[b]: Studying sometimes gives me a sense of deep satisfaction.	—	—	–0.27
6.2: My goal is to pass the course with as little effort as possible.	—	—	0.53
6.3: I only study the topics that are relevant for the exam.	—	—	0.73
6.4: Any topic can become interesting to me once I am engaged with it.	—	—	–0.44
6.5: I learn some things by repeating them until I have memorized them, even if I do not understand them.	—	—	0.68
6.6: I believe that I can pass most exams by memorizing the key points rather than understanding them.	—	—	0.68
6.7: I like to test myself on the key topics until I fully understand them.	—	—	–0.65
6.8: I will recommend mLearning with WueMedica to others.	—	0.90	—

[a]Loadings below the reporting threshold.

XSL•FO
RenderX

Table . Cronbach α and explained variance of the principal component analysis–derived scales.

	Learning outcome	User experience	Study motivation
Cronbach α	0.90	0.82	0.67
Explained variance (%)	22.8	13.4	14.0

Detailed results according to usage status and usage clusters are reported in Multimedia Appendix 2.

Questionnaire Results: Scale Scores According to App Usage Status

We compared app users and nonusers at both the scale and item level. We observed no statistically significant differences at the scale level. At the item level, however, app users consistently rated more favorably, most notably with regard to enjoyment.

Cluster 1 was not found to have any advantage over nonusers on any scale. At the item level, enjoyment was the only aspect rated higher than by nonusers.

Cluster 2 exhibited the most favorable self-reported learning experience. Compared with nonusers, this cluster scored significantly higher on *learning outcome* and *user experience* scales. Furthermore, *learning outcome* scores were also higher than those of cluster 1. No differences were observed for *study motivation*. At the item level, cluster 2 exhibited consistently higher ratings across multiple items related to perceived learning gains, enjoyment, feedback on learning progress, satisfaction with studying, and willingness to recommend the app.

Cluster 3 did not differ from nonusers or other clusters for outcomes at the scale level. At the item level, enjoyment was again rated more favorably than by nonusers. Descriptive questionnaire results are provided in Multimedia Appendix 2.

Open-Ended Responses

Open-ended responses are provided as tabulated summaries in Multimedia Appendix 3. Among nonusers and cluster 1, the most frequently reported themes were time constraints and a perceived misalignment between the app content, predominantly consisting of selected-response formats, and the oral examination format. Participants from clusters 2 and 3 more frequently highlighted motivational aspects and emphasized the app's contribution to learning the subject of microbiology. In these clusters, positive experiences related to enjoyment, perceived progress, and engagement were more prominent. For example, there were frequent references to gamification features such as affirmation through comparative rankings.

Baseline-Controlled Examination Performance According to App Usage Status

Analyses focused on the summer semester of 2024, including 35 nonusers and 60 app users distributed across 3 usage clusters. The written M1 examination percentage score served as the primary standardized measure of prior academic ability (Table 4).

Table . Examination performance by app usage status and usage clusters[a] (N=210[b]).

	Students, n	Microbiology examination oral grade, mean (SD)	National licensing examination written score (%), mean (SD)
Eligible population	210	2.5 (1.2)	74.0 (8.5)
Total participants	189	2.5 (1.2)	73.9 (8.5)
Summer 2023	94	2.5 (1.2)	75.6 (8.8)
Summer 2024	95	2.5 (1.2)	72.2 (7.8)
Nonusers	35	2.8 (1.3)	70.2 (8.1)
Users	60	2.3 (1.1)	73.4 (7.5)
Cluster 1	32	2.4 (1.1)	72.3 (6.8)
Cluster 2	18	2.1 (0.9)	73.8 (8.2)
Cluster 3	10	2.5 (1.3)	76.6 (8.1)

[a]Grades for both microbiology and national licensing examination (M1) range from 1 (excellent) to 5 (failed). The written M1 examination results are reported as percentage scores, calculated as the proportion of points achieved relative to the maximum possible score.

[b]Data are presented for the 210 participants with available M1 scores; the total study sample comprised 220 individuals.

At this baseline, users had slightly higher written M1 examination scores than nonusers (73.4% vs 70.2%; *P*=.04). Comparing usage clusters, only cluster 3 scored significantly higher than nonusers (76.6% vs 70.2%; 1-sided *P*=.02), while cluster 1 scores did not differ significantly from those of nonusers (72.3% vs 70.2%; 1-sided *P*=.12). Cluster 2 exhibited a nonsignificant trend toward higher performance (73.8% vs 70.2%; 1-sided *P*=.06).

Within-group analyses of oral microbiology grades relative to the written M1 baseline revealed significant improvement for nonusers (*P*=.04), cluster 1 (*P*=.001), and cluster 2 (*P*=.01). In contrast, cluster 3, despite the highest written M1 examination score (76.6%), did not exhibit any significant improvement (*P*=.34). Among users, cluster 3 had the lowest oral microbiology grade (mean 2.5), whereas nonusers scored slightly lower (mean 2.8). A significant interaction effect between cluster

XSL•FO
RenderX

2 and nonusers was observed for oral microbiology grades (1-sided *P*=.04).

Written M1 examination scores predicted oral microbiology grades significantly (*P*=.006; η_p^2=0.08). After analysis of covariance adjustment, the group×examination interaction for users versus nonusers remained statistically significant (*P*=.03; η_p^2=0.05), whereas cluster interaction effects were not significant (*P*=.23; η_p^2=.05). Effect sizes were small.

Discussion

Integration of Findings Within Existing Literature

Established models of mLearning emphasize that learning outcomes emerge from the interaction between learning context, technological affordances, and individual learner characteristics, rather than from the technology alone [17,19,20]. However, evidence derived from authentic curricular settings remains limited. The primary aim of this study was not to demonstrate the effectiveness of a specific mLearning intervention. Instead, we examined how learner characteristics and contextual conditions shape engagement and learning outcomes when mLearning is offered as a voluntary supplement within a routine medical curriculum [16,17]. By integrating effectiveness-related outcomes with technology acceptance perspectives within a RE framework, we moved beyond average effects. This enabled us to identify CMO configurations that explain heterogeneity in engagement patterns and learning outcomes (Multimedia Appendix 4) [55].

A central finding of the study was that prior academic performance remained the strongest predictor of examination scores, while the independent effect of app use was small after statistical adjustment. Rather than constituting a limitation, this finding represents a key empirical result. Students with higher baseline academic performance were more likely to engage with the app, reported more favorable learning experiences, and achieved better outcomes. This pattern is consistent with previous research illustrating that academically stronger students tend to display greater self-regulatory capacity, stronger goal orientation, and a higher propensity to engage with optional learning resources [45,69-71]. Accordingly, we found no evidence that the mLearning intervention compensates for lower baseline academic performance. This raises important equity concerns when mLearning is offered as a voluntary supplement, as it may preferentially benefit academically stronger students and reinforce existing differences [72]. This observation aligns with broader discussions on digital inequities in health professions education, in which students with weaker baseline skills or fewer resources are at risk of being left behind [73-75].

Self-selection provides valuable insight in this setting and should not be regarded solely as a source of bias. However, it limits causal inference regarding any independent intervention effect. At the same time, it directly addresses the core question of this study: Who engages with mLearning under routine curricular settings? It also helps explain why mLearning tools may fail to benefit those learners who arguably need them most. This gap is reinforced by the fact that most existing studies report average effects without stratifying outcomes according to baseline performance, leaving equity-related questions largely unaddressed [42,76].

The cluster analysis further illustrated substantial heterogeneity in engagement trajectories, reinforcing evidence that mLearning use is highly individualized [77]. Neither sporadic and low-intensity use nor early but temporary engagement was associated with favorable outcomes. In contrast, students who integrated the app strategically into existing study routines and used it in an examination-oriented manner exhibited the most favorable profiles. These findings indicate that mLearning does not compensate for absent or inconsistent learning strategies, but rather amplifies existing ones [78]. Instructional features such as spaced-repetition, feedback, or gamification should not be interpreted as mechanisms in themselves, but as affordances that may activate motivational, metacognitive, and self-regulatory processes in specific learner groups [69,77]. Consistent with the RE perspective, mechanisms emerge from the interaction between learner dispositions and contextual conditions, which may render identical features effective for some learners while remaining largely ineffective for others [55,56,60].

In contrast to much of the prior technology acceptance research [59,79-81], which has predominantly emphasized perceived usefulness as the key driver of adoption, our findings indicate that enjoyment emerged as a more robust correlate of sustained engagement and positive learning experiences [76,80,82,83]. In addition, intensive and strategic users were consistently more satisfied with their learning experience. This suggests that affective components play a central role in learners' intention to engage with mLearning tools, particularly when use is voluntary. In this context, the prominent role of enjoyment highlights the importance of addressing affective and cognitive dimensions of learning [77,80,84].

Theoretical and Practical Implications

Our findings indicate that mLearning effectiveness is context-dependent rather than technology-driven. Favorable outcomes were observed linked to strategic integration into existing study routines. Accordingly, commonly used instructional elements such as spaced repetition [85-87], gamified feedback [88-90], just-in-time learning [91,92], and microlearning [93-95] should be understood as supportive conditions rather than universally effective components. For curriculum designers, mLearning needs to be embedded within structured learning pathways. Evidence suggests that didactic scaffolding (clear pedagogy, feedback, and structured learning activities) [96] and systemic monitoring of engagement patterns [97] can substantially strengthen mLearning. In addition, learning analytics or AI-supported personalization may help address learner heterogeneity by enabling more adaptive and context-sensitive learning pathways [98-101].

To further operationalize these principles, implementation strategies should target both initial uptake and sustained engagement. First, low-threshold access and structured onboarding processes (eg, seamless integration into learning management systems, single-sign-on, and visible entry points such as QR codes) are essential to reduce barriers to participation. Second, curricular alignment and active

XSL•FO
RenderX

implementation strategies**,** including faculty endorsement and in-app prompts such as push notifications, can reinforce learning routines and support sustained, self-regulated learning engagement [39,102]. Third, interventions should explicitly address both perceived usefulness and affective engagement, as both are critical determinants of adoption and continued use in voluntary learning contexts [103]. Fourth, content should follow microlearning principles, using small, manageable units; intuitive navigation; and adaptive difficulty to support learners with different prior knowledge and promote equitable participation [92,95]. Finally, gamification elements (eg, leaderboards, badges, and progress indicators) can support engagement by making learning progress visible, as implemented in the app used in this study.

Limitations

The quasi-experimental design and voluntary app use limit causal inference, as observed differences in academic outcomes cannot be attributed to the intervention alone. In addition, the absence of randomization and the reliance on observational data restrict conclusions regarding independent intervention effects. Self-selection by higher-performing students precludes attribution of observed differences in academic outcomes to the intervention alone. Within an explanatory framework, however, this self-selection reflects authentic adoption behavior and therefore constitutes a meaningful finding rather than a methodological flaw.

Further limitations include potential expectancy effects, the timing of the postexamination survey, unmeasured learner characteristics (eg, self-regulated learning strategies or time management skills), and limited statistical power in subgroup analyses. Exactly half of the 220 participants used the app, but only 33.6% (n=37) completed the questionnaire. This potential nonresponse bias limits the generalizability of conclusions linking self-reported evaluations to actual usage behavior. The exclusive reliance on quantitative data further constrained interpretive depth and limited insight into learners' subjective experiences and decision-making processes.

Future Directions

Future research should integrate summative academic outcomes with longitudinal analyses of engagement behavior to examine how learner characteristics and curricular embedding jointly shape mLearning use and effectiveness.

Prior academic performance proved to be the strongest predictor of engagement in our study. However, other potentially relevant factors were not assessed, such as digital literacy, personality traits, or interest in the subject, which may partly explain cluster formation and differential usage patterns [104-107]. These unmeasured characteristics may influence both students' interaction with mLearning interventions and the resulting learning outcomes, thereby contributing to the heterogeneous engagement patterns observed. Consequently, future research should systematically investigate these variables, ideally using mixed methods designs that combine quantitative usage tracking with qualitative insights into learners' intentions and experiences.

Learning analytics and AI-enabled personalization can help address differences between learners. One example of this is providing feedback on engagement and learning behavior to support more individualized learning [108,109] However, their potential to support learners with lower baseline academic performance requires careful empirical evaluation.

Overall, this study highlights the importance of examining mLearning as a situated educational practice, emphasizing the need for designs and evaluations that reflect the complexity of students' attitudes toward and actual use of mLearning within their learning contexts.

Acknowledgments

The authors gratefully acknowledge the students who participated in this study for their time, engagement, and valuable contributions to the research. They would like to acknowledge Marc Appel for his administrative coordination and support in implementing the WueMedica project. They thank Carsten Scheller for his support in reviewing the app content in WueMedica. They value the organizational support by Waltraud Flederer to support the successful execution of the project. Finally, they thank Andrew Entwistle for the linguistic and grammatical revision of the manuscript as well as for his valuable suggestions for semantic improvements. No generative artificial intelligence was used in writing the final manuscript. The manuscript was proofread for style and language by a native English speaker.

Funding

This study was conducted within the WueMedica project and supported by WueDive – Innovation in Education, an intramural funding initiative of the University of Würzburg (Germany).

Data Availability

The datasets generated or analyzed during this study are available from the corresponding author on reasonable request.

Authors' Contributions

Conceptualization: SK, JB, JPHH
Data curation: JB, JPHH
Formal analysis: JB, JPHH

XSL•FO
RenderX

Funding acquisition: SK
Investigation: JPHH
Methodology: SK, JB, JPHH
Project administration: SK
Resources: SK, TTL
Software: JB
Supervision: SK
Validation: SK, TTL, JB, JPHH
Visualization: JPHH
Writing – original draft: JPHH
Data curation: JB, JPHH

Conflicts of Interest

None declared.

Multimedia Appendix 1
Baseline characteristics of participants according to semester, including demographics, app usage, questionnaire response, and examination participation (N=220).
[DOCX File, 16 KB - mededu_v12i1e85892_app1.docx]

Multimedia Appendix 2
Questionnaire results: factor scores and item-level ratings according to usage status and usage clusters (N=74).
[DOCX File, 26 KB - mededu_v12i1e85892_app2.docx]

Multimedia Appendix 3
Open-ended responses.
[DOCX File, 27 KB - mededu_v12i1e85892_app3.docx]

Multimedia Appendix 4
Context-mechanism-outcome configurations.
[DOCX File, 32 KB - mededu_v12i1e85892_app4.docx]

Checklist 1
CONSORT-EHEALTH checklist.
[DOCX File, 16 KB - mededu_v12i1e85892_app5.docx]

Checklist 2
RAMESES-II checklist.
[DOCX File, 15 KB - mededu_v12i1e85892_app6.docx]

References

1. Grant MM. Difficulties in defining mobile learning: analysis, design characteristics, and implications. Educ Technol Res Dev 2019 Apr;67(2):361-388. [doi: 10.1007/s11423-018-09641-4]
2. Crompton H. A historical overview of mobile learning: toward learner-centered education. In: Berge ZL, Muilenburg L, editors. Handbook of Mobile Learning: Routledge; 2013:3-14.
3. Seipold J. Aus der Geschichte des mobilen Lernens: Strömungen, Trends und White Spaces. In: Handbuch Mobile Learning: Springer; 2018:13-41. [doi: 10.1007/978-3-658-19123-8_2]
4. Weiss T, Petzoldt T, Bannert M, Krems J. Calibration as side effect? Computer-based learning in driver education and the adequacy of driving-task-related self-assessments. Transp Res F Traf 2013 Feb;17:63-74. [doi: 10.1016/j.trf.2012.10.001]
5. Frohberg D. Mobile Learning: Der Wirtschaftswissenschaftlichen Fakultät der Universität Zürich [Dissertation]. 2008 URL: https://capuana.ifi.uzh.ch/publications/PDFs/2292_m-learning_frohberg_komprimiert.pdf [accessed 2026-05-09]
6. Winters N. What is mobile learning? Kaleidoscope. 2006. URL: https://www.cin.ufpe.br/~mlearning/intranet/m-learning/Big%20Issues%20in%20Mobile%20Learning.pdf [accessed 2026-05-07]
7. Witt C, Sieber A, editors. Mobile Learning: Potenziale, Einsatzszenarien und Perspektiven des Lernens mit Mobilen Endgeräten: Springer VS; 2013.
8. West M, Vosloo S. UNESCO Policy Guidelines for Mobile Learning: United Nations Educational, Scientific and Cultural Organization; 2013.

XSL•FO
RenderX

9. Ellaway RH, Masters K. AMEE Guide 32: e-Learning in Medical Education part 1: learning, teaching and assessment. Med Teach 2008 Jun;30(5):455-473. [doi: 10.1080/01421590802108331] [Medline: 18576185]
10. Ally M. Mobile Learning: Transforming the Delivery of Education and Training: Athabasca University Press; 2009.
11. Ally M, Lin F, McGreal R, Woo B, Li Q. An intelligent agent for adapting and delivering electronic course materials to mobile learners. Athabasca University. 2005. URL: https://auspace.athabascau.ca/handle/2149/216 [accessed 2026-5-14]
12. Traxler J. Defining mobile learning. In: Isaías P, Borg C, Kommers P, Bonanno P, editors. Mobile Learning 2005 Proceedings: IADIS; 2005:261-266.
13. Danish JA, Hmelo-Silver CE. On activities and affordances for mobile learning. Contemp Educ Psychol 2019;60:101829. [doi: 10.1016/j.cedpsych.2019.101829]
14. Kukulska‐Hulme A. Mobile learning as a catalyst for change. Open Learn J Open Distance e-Learn 2010;25(3):181-185. [doi: 10.1080/02680513.2010.511945]
15. Traxler J. Defining, discussing and evaluating mobile learning: the moving finger writes and having writ. Int Rev Res Open Distrib Learn 2007;8(2). [doi: 10.19173/irrodl.v8i2.346]
16. Sharples M, Taylor J, Vavoula GN. A theory of learning for the mobile age. In: Haythornthwaite C, Andrews R, Fransman J, Meyers EM, editors. The SAGE Handbook of E-Learning Research, 2nd edition: SAGE Publications; 2016:63-81. [doi: 10.4135/9781529716696.n4]
17. Koole ML. A model for framing mobile learning. In: Ally M, editor. Mobile Learning: Transforming the Delivery of Education and Training: Athabasca University Press; 2009:25-50.
18. Park Y. A pedagogical framework for mobile learning: categorizing educational applications of mobile technologies into four types. Int Rev Res Open Distrib Learn 2011;12(2):78-102. [doi: 10.19173/irrodl.v12i2.791]
19. Kearney M, Schuck S, Burden K, Aubusson P. Viewing mobile learning from a pedagogical perspective. Res Learn Technol 2012 Feb;20:14406. [doi: 10.3402/rlt.v20i0.14406]
20. Petit T, Santos GL. Mobile learning: an ergonomic alternative for long-awaited educational changes? In: Bayyurt Y, Kalz M, Specht M, editors. International Conference on Mobile and Contextual Learning, 1st edition: Springer International Publishing; 2014:1-14.
21. Parsons D, Palalas A, Nikou S, Rodulfo S. Mobile learning frameworks and pedagogy: a systematic review. Eur J Educ 2024 Jun;59(2):e12601. [doi: 10.1111/ejed.12601]
22. McQuiggan S, Kosturko L, McQuiggan J, Sabourin J. Mobile Learning: A Handbook for Developers, Educators, and Learners: SAS Institute Inc; 2015.
23. Kukulska-Hulme A, Traxler J, editors. Mobile Learning: A Handbook for Educators and Trainers: Routledge; 2005.
24. Sharples M, Taylor J, Vavoula G. Towards a theory of mobile learning. In: van der Merwe H, Brown T, editors. Mobile Technology: The Future of Learning in Your Hands: mLearn; 2005.
25. Dalloz X. With the Internet of Augmented Me, a new wave of innovation is coming, which will change everything. In: Duvaut P, Dalloz X, Menga D, Koehl F, Chriqui V, Brill J, editors. Internet of Augmented Me, IAM: Wiley-ISTE; 2020:1-36.
26. Kakihara M, Sorensen C. Mobility: an extended perspective. Presented at: Proceedings of the 35th Annual Hawaii International Conference on System Sciences; Jan 10, 2002. [doi: 10.1109/HICSS.2002.994088]
27. Traxler J. Current state of mobile learning. In: Ally M, editor. Mobile Learning: Transforming the Delivery of Education and Training: Athabasca University Press; 2009:9-24.
28. Stoller-Schai D. Mobiles Lernen: die Lernform des Homo Mobilis. In: Hohenstein A, Wilbers K, editors. Handbuch E-Learning: Expertenwissen Aus Wissenschaft Und Praxis - Strategien, Instrumente, Fallstudien: Wolters Kluwer; 2010.
29. Badillo PY, Tarrier F. Mobilité et Ubiquité Dans le Futur Vers le Nomadisme Numérique. Cahier N°1 de l'ANR: Agence Nationale de la Recherche; 2009.
30. Sharples M. Big issues in mobile learning. : University of Nottingham; 2006 URL: https://www.cin.ufpe.br/~mlearning/intranet/m-learning/Big%20Issues%20in%20Mobile%20Learning.pdf [accessed 2026-05-07]
31. Dewey J. Democracy and Education: An Introduction to the Philosophy of Education: Macmillan; 1916.
32. Lampropoulos G. Social robots in education: current trends and future perspectives. Information 2025;16(1):29. [doi: 10.3390/info16010029]
33. Bartelsen J. Mobile Learning: Ein Überblick. Nordakademie - Hochschule der Wirtschaft. 2011. URL: https://www.econstor.eu/bitstream/10419/67093/1/685361942.pdf [accessed 2026-05-07]
34. Laouris Y, Eteokleous N. We need an educationally relevant definition of mobile learning. Presented at: Proceedings of mLearn 2005; Oct 25-28, 2005 URL: https://www.cin.ufpe.br/~mlearning/intranet/m-learning/mlearn2005/We%20need%20an%20Educationally%20Relevant.pdf [accessed 2026-05-07]
35. O'Malley C, Vavoula G, Glew J, et al. Guidelines for learning/teaching/tutoring in a mobile environment. : HAL; 2005 URL: https://hal.science/hal-00696244 [accessed 2026-05-07]
36. Chat M, Zghida N, Ouaali AA, Youlyouz A, Amiri EM, Janati-Idrissi R. Towards a new definition of e-learning and m-learning. Multidiscip Rev 2025;8(9):e2025281. [doi: 10.31893/multirev.2025281]
37. Samala AD, Papadakis S, Rawas S. Global insights into mobile learning in higher education: a PRISMA-guided bibliometric analysis from 2007 to 2023. Int J Educ Reform 2025. [doi: 10.1177/10567879251341869]

XSL•FO
RenderX

38. Garzón J, Burgos D, Tlili A. Advantages and challenges associated with mobile learning in education: a systematic literature review. J Comput Educ 2025 Dec;12:1173-1205. [doi: 10.1007/s40692-024-00342-x]
39. Crompton H, Burke D. The use of mobile learning in higher education: a systematic review. Comput Educ 2018;123:53-64. [doi: 10.1016/j.compedu.2018.04.007]
40. Kho A, Henderson LE, Dressler DD, Kripalani S. Use of handheld computers in medical education. J Gen Intern Med 2006 May;21:531-537. [doi: 10.1111/j.1525-1497.2006.00444.x]
41. Fischer S, Stewart TE, Mehta S, Wax R, Lapinsky SE. Handheld computing in medicine. J Am Med Inform Assoc 2003;10(2):139-149. [doi: 10.1197/jamia.m1180] [Medline: 12595403]
42. Chandran VP, Balakrishnan A, Rashid M, et al. Mobile applications in medical education: a systematic review and meta-analysis. PLoS One 2022;17(3):e0265927. [doi: 10.1371/journal.pone.0265927] [Medline: 35324994]
43. Dimond R, Bullock A, Lovatt J, Stacey M. Mobile learning devices in the workplace: "as much a part of the junior doctors" kit as a stethoscope'? BMC Med Educ 2016 Aug 17;16(1):207. [doi: 10.1186/s12909-016-0732-z] [Medline: 27530343]
44. Traxler J. What killed the mobile learning dream? Medium. 2016. URL: https://medium.com/@Jisc/what-killed-the-mobile-learning-dream-8c97cf66dd3d [accessed 2025-10-06]
45. Karimi S. Do learners' characteristics matter? An exploration of mobile-learning adoption in self-directed learning. Comput Hum Behav 2016 Oct;63:769-776. [doi: 10.1016/j.chb.2016.06.014]
46. Lewis KO, Popov V, Fatima SS. From static web to metaverse: reinventing medical education in the post-pandemic era. Ann Med 2024 Dec;56(1):2305694. [doi: 10.1080/07853890.2024.2305694] [Medline: 38261592]
47. Car J, Ong QC, Erlikh Fox T, et al. The Digital Health Competencies in Medical Education Framework: an international consensus statement based on a Delphi study. JAMA Netw Open 2025 Jan 2;8(1):e2453131. [doi: 10.1001/jamanetworkopen.2024.53131] [Medline: 39888625]
48. Neustrukturierung des Medizinstudiums und Änderung der Approbationsordnung für Ärzte | Empfehlungen der Expertenkommission zum Masterplan Medizinstudium. Wissenschaftsrat. 2020. URL: https://www.wissenschaftsrat.de/download/archiv/7271-18 [accessed 2026-05-08]
49. Gloerfeld C. Mobile learning - was ist eigentlich der Kern? In: de Witt C, Gloerfeld C, editors. Handbuch Mobile Learning: Springer VS; 2018:257-282. [doi: 10.1007/978-3-658-19123-8_14]
50. Report to the European Commission on improving the quality of teaching and learning in Europe's higher education institutions. : Publications Office of the European Union; 2013 URL: http://search.library.nuigalway.ie/primo-explore/fulldisplay?vid=353GAL_VU1&tab=local&docid=dedupmg3167800&lang=en_US&context=L&query=sub,exact,%20Educational%20exchanges,AND&mode=advanced&offset=30 [accessed 2026-05-08]
51. Khurana MP, Raaschou-Pedersen DE, Kurtzhals J, Bardram JE, Ostrowski SR, Bundgaard JS. Digital health competencies in medical school education: a scoping review and Delphi method study. BMC Med Educ 2022 Feb 26;22(1):129. [doi: 10.1186/s12909-022-03163-7] [Medline: 35216611]
52. Barradell S, Bell A, Thomson K, Hughes J. Students' informal learning interactions in health professions education: insights from a qualitative synthesis. High Educ Res Dev 2025;44(2):322-337. [doi: 10.1080/07294360.2024.2393125]
53. Bashir F, Warraich NF, Malik A. Paradigm shift of health professions education through mobile devices: a literature review of reviews. Med Sci Educ 2025;35(2):1137-1147. [doi: 10.1007/s40670-025-02294-y] [Medline: 40352993]
54. Parsons D. A mobile learning overview by timeline and mind map. Int J Mob Blended Learn 2014 Oct;6(4):1-21. [doi: 10.4018/ijmbl.2014100101]
55. Pawson R, Tilley N. Realistic Evaluation: SAGE Publications; 1997.
56. Lemire S, Kwako A, Nielsen SB, Christie CA, Donaldson SI, Leeuw FL. What is this thing called a mechanism? Findings from a review of realist evaluations. New Dir Eval 2020;2020(167):73-86. [doi: 10.1002/ev.20428]
57. Nielsen SB, Lemire S. Nothing as practical as an analytical strategy in realist evaluation. In: Manzano A, Williams E, editors. Realist Evaluation: Principles and Practice: Routledge; 2024:40-56.
58. Nationaler Kompetenzbasierter Lernzielkatalog Medizin (NKLM). URL: https://nklm.de/ [accessed 2026-02-25]
59. Venkatesh V, Morris MG, Davis GB, Davis FD. User acceptance of information technology: toward a unified view. MIS Q 2003 Sep;27(3):425-478. [doi: 10.2307/30036540]
60. Biggs J, Kember D, Leung DY. The Revised Two-Factor Study Process Questionnaire: R-SPQ-2F. Br J Educ Psychol 2001 Mar;71(Pt 1):133-149. [doi: 10.1348/000709901158433] [Medline: 11307705]
61. Youhasan P, Raheem S. Technology enabled formative assessment in medical education: a pilot study through Kahoot. Educ Med J 2019;11(3):23-29. [doi: 10.21315/eimj2019.11.3.3]
62. Kibble J. Use of unsupervised online quizzes as formative assessment in a medical physiology course: effects of incentives on student participation and performance. Adv Physiol Educ 2007 Sep;31(3):253-260. [doi: 10.1152/advan.00027.2007] [Medline: 17848591]
63. Rasch D, Kubinger KD, Moder K. The two-sample t test: pre-testing its assumptions does not pay off. Stat Papers 2011 Feb;52:219-231. [doi: 10.1007/s00362-009-0224-x]
64. Rasch D, Teuscher F, Guiard V. How robust are tests for two independent samples? J Stat Plan Inference 2007 Aug;137(8):2706-2720. [doi: 10.1016/j.jspi.2006.04.011]

XSL•FO
RenderX

65. Eysenbach G, CONSORT-EHEALTH Group. CONSORT-EHEALTH: improving and standardizing evaluation reports of web-based and mobile health interventions. J Med Internet Res 2011 Dec 31;13(4):e126. [doi: 10.2196/jmir.1923] [Medline: 22209829]
66. Wong G, Westhorp G, Manzano A, Greenhalgh J, Jagosh J, Greenhalgh T. RAMESES II reporting standards for realist evaluations. BMC Med 2016 Jun 24;14(1):96. [doi: 10.1186/s12916-016-0643-1] [Medline: 27342217]
67. O'Rourke N, Hatcher L. A Step-by-Step Approach to Using SAS for Factor Analysis and Structural Equation Modeling, 2nd edition: SAS Press; 2013.
68. Savalei V, Huang M. Fit indices are insensitive to multiple minor violations of perfect simple structure in confirmatory factor analysis. Psychol Methods 2025 Feb 13. [doi: 10.1037/met0000718] [Medline: 39946625]
69. Zimmerman BJ. Becoming a self-regulated learner: an overview. Theory Pract 2002;41:64-70. [doi: 10.1207/s15430421tip4102_2]
70. Chi NY, Huyen NT. Self-regulation of learning in higher education mobile learning landscape: a review. VNU J Sci Educ Res 2025;41(4):66-77. [doi: 10.25073/2588-1159/vnuer.5207]
71. Cheon J, Lee S, Crooks SM, Song J. An investigation of mobile learning readiness in higher education based on the theory of planned behavior. Comput Educ 2012 Nov;59(3):1054-1064. [doi: 10.1016/j.compedu.2012.04.015]
72. Bhat PP, Rajashekar B, Kamath U. Perspectives on self-directed learning — the importance of attitudes and skills. Biosci Educ 2007;10(1):1-3. [doi: 10.3108/beej.10.c3]
73. Han SP, Kumwenda B. Bridging the digital divide: promoting equal access to online learning for health professions in an unequal world. Med Educ 2025 Jan;59(1):56-64. [doi: 10.1111/medu.15455] [Medline: 38860836]
74. König LM, Krukowski RA, Kuntsche E, et al. Reducing intervention- and research-induced inequalities to tackle the digital divide in health promotion. Int J Equity Health 2023 Dec 4;22(1):249. [doi: 10.1186/s12939-023-02055-6] [Medline: 38049789]
75. Pham Thi TD, Pham VK, Duong NT. Understanding m-learning app adoption: an integrated model for college students in Vietnam. Interact Learn Environ 2025;33(10):5847-5869. [doi: 10.1080/10494820.2025.2488144]
76. Golenhofen N, Heindl F, Grab-Kroll C, Messerer DA, Böckers TM, Böckers A. The use of a mobile learning tool by medical students in undergraduate anatomy and its effects on assessment outcomes. Anat Sci Educ 2020 Jan;13(1):8-18. [doi: 10.1002/ase.1878] [Medline: 30913369]
77. Terras MM, Ramsay J. The five central psychological challenges facing effective mobile learning. Br J Educ Technol 2012;43(5):820-832. [doi: 10.1111/j.1467-8535.2012.01362.x]
78. Baars M, Khare S, Ridderstap L. Exploring students' use of a mobile application to support their self-regulated learning processes. Front Psychol 2022;13:793002. [doi: 10.3389/fpsyg.2022.793002] [Medline: 35360578]
79. Davis FD. A technology acceptance model for empirically testing new end-user information systems: theory and results [PhD thesis]. : Massachusetts Institute of Technology, Sloan School of Management; 1986 URL: https://dspace.mit.edu/handle/1721.1/15192 [accessed 2026-05-08]
80. Nabipour Sanjebad N, Shrestha A, Shahid P. The impact of personality traits towards the intention to adopt mobile learning. In: Sharma SK, Dwivedi YK, Metri B, Rana NP, editors. Re-Imagining Diffusion and Adoption of Information Technology and Systems: A Continuing Conversation: Springer; 2020:182-193. [doi: 10.1007/978-3-030-64861-9_17]
81. Granić A, Marangunić N. Technology acceptance model in educational context: a systematic literature review. Br J Educ Technol 2019 Sep;50(5):2572-2593. [doi: 10.1111/bjet.12864]
82. Kucuk S, Baydas Onlu O, Kapakin S. A model for medical students' behavioral intention to use mobile learning. J Med Educ Curric Dev 2020;7:2382120520973222. [doi: 10.1177/2382120520973222] [Medline: 33313399]
83. Wu H, Que X, Pan L. Analysing students' aspiration factors that impact actual use of mobile learning systems: a two-stage SEM-ANN approach. Humanit Soc Sci Commun 2025;12:406. [doi: 10.1057/s41599-025-04717-y]
84. Ebbinghaus H. Über Das Gedächtnis Untersuchungen Zur Experimentellen Psychologie: Duncker & Humblot; 1885.
85. Cepeda NJ, Pashler H, Vul E, Wixted JT, Rohrer D. Distributed practice in verbal recall tasks: a review and quantitative synthesis. Psychol Bull 2006 May;132(3):354-380. [doi: 10.1037/0033-2909.132.3.354] [Medline: 16719566]
86. Schimanke F. The impact of spaced repetition learning on the learning success in mobile learning games. Presented at: 2021 IEEE International Symposium on Multimedia (ISM); Nov 29 to Dec 1, 2021. [doi: 10.1109/ISM52913.2021.00054]
87. Ratinho E, Martins C. The role of gamified learning strategies in student's motivation in high school and higher education: a systematic review. Heliyon 2023 Aug;9(8):e19033. [doi: 10.1016/j.heliyon.2023.e19033] [Medline: 37636393]
88. Krishnamurthy K, Selvaraj N, Gupta P, et al. Benefits of gamification in medical education. Clin Anat 2022 Sep;35(6):795-807. [doi: 10.1002/ca.23916] [Medline: 35637557]
89. Su CH, Cheng CH. A mobile gamification learning system for improving the learning motivation and achievements. J Comput Assist Learn 2015 Jun;31(3):268-286. [doi: 10.1111/jcal.12088]
90. Breitwieser J, Neubauer AB, Schmiedek F, Brod G. Realizing the potential of mobile interventions for education. NPJ Sci Learn 2024 Dec 12;9(1):76. [doi: 10.1038/s41539-024-00289-9] [Medline: 39668147]
91. Riel M. Education in the 21st century: just-in-time learning or learning communities. In: Education and the Arab World: Challenges of the Next Millennium: Emirates Center for Strategic Studies and Research; 2000:137-160.

XSL•FO
RenderX

92. HajiAbad FA, Keshmiri F, Jabinian F, Shiryazdi SM. Micro-learning and Mindtools in Mobile-Learning (M-Learning): effects on knowledge and self-efficacy in laparoscopy education for surgical nursing students. BMC Nurs 2025 Feb 21;24(1):202. [doi: 10.1186/s12912-025-02841-3] [Medline: 39984878]
93. Kim C, Pekrun R. Emotions and motivation in learning and performance. In: Spector J, Merrill M, Elen J, Bishop M, editors. Handbook of Research on Educational Communications and Technology: Springer; 2014:65-76. [doi: 10.1007/978-1-4614-3185-5_6]
94. Hug T. Mikrolernen: konzeptionelle Überlegungen und Anwendungsbeispiele. In: Herzig B, Meister DM, Moser H, Niesyto H, editors. Jahrbuch Medienpädagogik 8: VS Verlag für Sozialwissenschaften; 2010:221-238. [doi: 10.1007/978-3-531-92135-8_12]
95. Thillainadesan J, Le Couteur DG, Haq I, Wilkinson TJ. When I say … microlearning. Med Educ 2022 Aug;56(8):791-792. [doi: 10.1111/medu.14848] [Medline: 35654438]
96. Kalantarion M, Ahmady S, Kallestrup P, et al. How to create a successful mobile learning strategy for medical education during lockdowns? J Educ Health Promot 2024;13:70. [doi: 10.4103/jehp.jehp_167_23] [Medline: 38559490]
97. Zhao Y, Morris A, Marais BJ, Pardo A, Scott KM. Exploring how medical students learn during clinical rotations: a pilot study with a mobile application. Health Technol 2019 May;9:257-267. [doi: 10.1007/s12553-019-00305-8]
98. Feigerlova E, Hani H, Hothersall-Davies E. A systematic review of the impact of artificial intelligence on educational outcomes in health professions education. BMC Med Educ 2025 Jan 27;25:129. [doi: 10.1186/s12909-025-06719-5]
99. Zhu Y, Dai W, Kang Q. The analysis of artificial intelligence-based mobile learning in students' open teaching recommendation system based on deep learning. Sci Rep 2025;15:21927. [doi: 10.1038/s41598-025-08147-3]
100. Boscardin CK, Gin B, Golde PB, Hauer KE. ChatGPT and generative artificial intelligence for medical education: potential impact and opportunity. Acad Med 2024 Jan 1;99(1):22-27. [doi: 10.1097/ACM.0000000000005439] [Medline: 37651677]
101. Lee J, Wu AS, Li D, Kulasegaram KM. Artificial intelligence in undergraduate medical education: a scoping review. Acad Med 2021 Nov 1;96(11S):S62-S70. [doi: 10.1097/ACM.0000000000004291] [Medline: 34348374]
102. Palalas A, Wark N. The relationship between mobile learning and self-regulated learning: a systematic review. Australas J Educ Technol 2020;36(4):151-172. [doi: 10.14742/ajet.5650]
103. Li J, Jin M, Chen X. Understanding continued use of smart learning platforms: psychological wellbeing in an extended TAM-ISCM model. Front Psychol 2025;16:1521174. [doi: 10.3389/fpsyg.2025.1521174] [Medline: 40438757]
104. Getenet S, Cantle R, Redmond P, Albion P. Students' digital technology attitude, literacy and self-efficacy and their effect on online learning engagement. Int J Educ Technol High Educ 2024;21:3. [doi: 10.1186/s41239-023-00437-y]
105. Yaseen H, Mohammad AS, Ashal N, Abusaimeh H, Ali A, Sharabati AA. The impact of adaptive learning technologies, personalized feedback, and interactive AI tools on student engagement: the moderating role of digital literacy. Sustainability 2025;17(3):1133. [doi: 10.3390/su17031133]
106. Afini Normadhi NB, Shuib L, Md Nasir HN, Bimba A, Idris N, Balakrishnan V. Identification of personal traits in adaptive learning environment: systematic literature review. Comput Educ 2019;130:168-190. [doi: 10.1016/j.compedu.2018.11.005]
107. Klemmt C, Backhaus J, Jeske D, Koenig S. Psychometric properties and calibration of the SPOREEM (Students' Perception of the Operating Room Educational Environment Measure). J Surg Educ 2021;78(4):1151-1163. [doi: 10.1016/j.jsurg.2020.10.015] [Medline: 33168469]
108. Pishtari G, Rodríguez‐Triana MJ, Sarmiento‐Márquez EM, et al. Learning design and learning analytics in mobile and ubiquitous learning: a systematic review. Br J Educ Technol 2020;51(4):1078-1100. [doi: 10.1111/bjet.12944]
109. Deshpande S, Kumari D, Rokade R, Chatandar N, Komati P. AI based personalized e-learning system. Int J Sci Technol 2025;16(2). [doi: 10.71097/IJSAT.v16.i2.5955]

Abbreviations

AI: artificial intelligence
CMO: context-mechanism-outcome
mLearning: mobile learning
RE: realist evaluation

Edited by J Gentges; submitted 15.Oct.2025; peer-reviewed by A Afra, S Papadakis; revised version received 23.Mar.2026; accepted 31.Mar.2026; published 21.May.2026.

Please cite as:
Holzhäuer JPH, Backhaus J, Lâm TT, König S
Mobile Learning in Medical Education: Quasi-Experimental Realist Evaluation of Usage, Context, and Examination Performance in a Curricular Setting
JMIR Med Educ 2026;12:e85892
URL: https://mededu.jmir.org/2026/1/e85892
doi:10.2196/85892

XSL•FO
RenderX

© Jan Peter Herbert Holzhäuer, Joy Backhaus, Thiên-Trí Lâm, Sarah König. Originally published in JMIR Medical Education (https://mededu.jmir.org), 21.May.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Education, is properly cited. The complete bibliographic information, a link to the original publication on https://mededu.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Mobile Medical Education Apps as Digital Health Information Resources in China and the United States: Longitudinal App Store Metadata Analysis and Topic Modeling of User Review Narratives

Junhao Ma[1,2*], MSc; Tian Shen[1,2*], MSc; Peng Xiang[3,4*], MD; Yingjun Li[1*], PhD; Yuanshuai Chen[3,5], MD; Zhengyang Lu[6], BM; Jingkai Zeng[7], BM; Wei Dai[8], BM; Hanwen Pan[9,10,11,12], BM; Liping Yang[13,14], MD; Yunfan He[15], MS; Xingxi Li[16], MS; Joyce Wu[17,18], BSc; Qi Yu[19,20], MD; Jun Liang[1,2,3,6], MS; Jianbo Lei[21,22], MD, PhD

[1]Department of Epidemiology and Health Statistics, School of Public Health, Hangzhou Medical College, Hangzhou, China

[2]National Key Laboratory of Transvascular Implantable Devices, Second Affiliated Hospital, School of Medicine, Zhejiang University, Hangzhou, China

[3]Department of AI and IT, Second Affiliated Hospital, School of Medicine, Zhejiang University, Hangzhou, China

[4]Intelligent Medical Research Center, Zhejiang University Institute of Computer Innovation Technology, Hangzhou, China

[5]Medical Education Research Center, School of Medicine, Zhejiang University, Hangzhou, China

[6]Department of Preventive Medicine, School of Public Health, Zhejiang University, Hangzhou, China

[7]Department of Clinical Medicine, Taizhou University, Taizhou, China

[8]School of Medical Technology and Information Engineering, Zhejiang Chinese Medical University, Hangzhou, China

[9]Foshan Fosun Chancheng Hospital, Guangdong Medical University, Foshan, China

[10]Guangdong Provincial Hospital of Chinese Medicine, Guangzhou, China

[11]Foshan Clinical Medical School, Guangzhou University of Chinese Medicine, Foshan, China

[12]Zhonghe Guoyi Beijing International Medical Research Institute, Beijing, China

[13]College of Medical Engineering and Technology, Xinjiang Medical University, Urumqi, Xinjiang, China

[14]Institute of Medical Engineering Interdisciplinary Research, Xinjiang Medical University, Urumqi, China

[15]School of International Relations and Public Affairs, Fudan University, Shanghai, China

[16]Department of Industrial Engineering, Tsinghua University, Beijing, China

[17]Academy for Allied Health Sciences, Scotch Plains, NJ, United States

[18]New Jersey Medical School, Rutgers University, Newark, NJ, United States

[19]Shanxi Key Laboratory of AI-Enhanced Clinical Decision Support, Shanxi Medical University, Taiyuan, China

[20]Xinzhou Vocational and Technical College, Xinzhou, China

[21]Center for Medical Informatics, Institute of Advanced Clinical Medicine, Peking University, Beijing, China

[22]Affiliated Hospital of Southwest Medical University, Luzhou, China

[*]these authors contributed equally

Corresponding Author:
Jianbo Lei, MD, PhD
Center for Medical Informatics, Institute of Advanced Clinical Medicine
Peking University
38 Xueyuan Rd, Haidian District
Beijing, 100191
China
Phone: 86 10 8280 5901
Email: jblei@hsc.pku.edu.cn

Abstract

Background: Mobile learning apps are increasingly used in medical education, but cross-national evidence on their longitudinal development and rating-based satisfaction patterns remains limited.

Objective: This study aimed to examine the longitudinal evolution of medical education apps in China and the United States and identify app-related factors associated with positive and negative rating deviations among app store reviewers.

XSL•FO
RenderX

Methods: We conducted a longitudinal app store analysis of 467 medical education apps and 232,598 user reviews from major Chinese and US app stores. Joinpoint regression was used to characterize temporal trends in app development. Latent Dirichlet allocation (LDA) topic modeling, with large language model–assisted topic labeling, was used to identify themes in user reviews. Tobit regression, Wald tests, and Kano model classification were used to examine asymmetric associations with positive and negative rating deviations and to classify review topics into functional attributes.

Results: Among the 232,598 analyzed reviews, 219,105 were rated 4 stars or higher, corresponding to an overall rating-based satisfaction proxy rate of 94.20%. This rating-based satisfaction proxy rate was lower in China than in the United States (128,018/139,416, 91.82% vs 91,087/93,182, 97.75%). Within the 2025-observable analytic cohort, cumulative observable app counts increased significantly in both countries, with average annual percentage changes of 41.69% in China and 41.17% in the United States. The joinpoint analysis identified temporal inflection points within this observable cohort around 2017 and 2021-2022. User concerns also shifted over time. In China, the dominant review theme changed from "Clinical practicality" in 2009-2017 (2735/7385 reviews, 37.03%) to "Practice test function" in 2018-2025 (53,929/132,031 reviews, 40.85%). In the United States, user attention shifted from "Functional availability" in 2009-2017 (20,736/26,118 reviews, 79.39%) to "Cost" in 2018-2025 (27,401/67,064 reviews, 40.86%). Tobit regression and Kano classification showed that factors associated with positive and negative rating deviations differed by country and developmental stage, with essential and charm factors varying across temporal phases.

Conclusions: This longitudinal app store analysis revealed distinct but partially converging patterns of observable app development and rating-based satisfaction among app store reviewers in China and the United States. These findings suggest that future medical education apps should prioritize reliable content, technical stability, affordable exam preparation, and personalized learning support. They may also inform the design of future AI-supported medical education platforms while requiring further validation with learning process and outcome data.

(JMIR Med Educ 2026;12:e107560) doi:10.2196/107560

KEYWORDS

digital health informatics; medical education; mobile health apps; text mining; app store analysis; user-generated content; rating-based satisfaction; cross-national comparison

Introduction

With the rapid advancement of mobile internet, AI, and digital technologies, coupled with the growing demands of professional education for continuous learning and self-renewal capabilities, digital learning methods are gradually evolving from traditional teaching aids into essential infrastructure supporting lifelong learning. Medical education, as a field highly dependent on specialized knowledge, clinical skills, and situational decision-making abilities, demands rapid knowledge updates and navigates complex practical scenarios, placing heightened demands on learning flexibility and accessibility [1]. However, the long-standing training model centered on fixed classroom instruction and offline clinical rotations remains constrained in terms of scheduling, learning continuity, and personalized support. This vulnerability becomes particularly pronounced during public health emergencies [2,3]. The COVID-19 pandemic has further accelerated the shift toward digital, self-directed, and contextualized medical education models [4,5]. Against this backdrop, the development of mobile learning and online educational resources has enabled learning activities to be more flexibly integrated into daily teaching and clinical practice [6]. Medical education apps support diverse educational activities including coursework, clinical guideline access [7,8], case reasoning training, and formative assessment [9-13], providing on-demand learning support for medical students, residents, and practicing physicians [14].

Despite the rapid global adoption of medical education apps, their practical implementation faces numerous challenges. Existing research indicates that learners frequently encounter issues such as outdated content, insufficient system stability, and distractions from multifunctional mobile devices when using medical education apps. These factors negatively impact learning outcomes, user experience, and willingness for sustained use [15-18]. Given the high diversity in developers, target users, and application scenarios of medical education apps, quality issues do not manifest uniformly but exhibit distinct characteristics across different markets and cultural contexts [19]. Differences in medical education systems, learning approaches, and cultural backgrounds across countries and regions also lead to significant variations in learners' perceptions and demands regarding app functionality, content quality, and "user satisfaction" [20]. This implies that evaluations of medical education apps that overlook market and cultural context differences may fail to accurately reflect their true utility and areas for improvement.

These contextual differences are particularly salient in a China-US comparison. Both countries represent large digital app markets, but medical education apps are embedded in different institutional, educational, and sociocultural environments. Established technology adoption theories, including the technology acceptance model (TAM) and the unified theory of acceptance and use of technology, suggest that users' acceptance and continued use of digital technologies are shaped by perceived usefulness, ease of use, social influence, and facilitating conditions [21,22]. Cross-cultural technology acceptance research further indicates that national and organizational cultures may moderate how users perceive and adopt digital technologies [23]. In China, medical education app use is situated within standardized medical training

XSL•FO
RenderX

pathways, national medical education reform, and examination-oriented learning pressures within the medical education system [24]. In the United States, medical education app use occurs within a credentialing environment shaped by the United States Medical Licensing Examination (USMLE) and residency selection processes [25]. These differences provide a rationale for comparing how app store reviewers in the 2 countries evaluate content reliability, technical stability, cost, exam preparation, and personalized learning support. Therefore, this study examines medical education apps not only as digital market products but also as sociotechnical educational tools embedded in distinct policy, institutional, and learning contexts.

Nevertheless, current research on user experience and satisfaction with medical education apps still primarily relies on methods such as questionnaire surveys [26,27], standardized scales, and on-site usability testing [28]. Common tools such as the Mobile App Rating Scale (MARS), developed by Stoyanov et al [29], and the Mobile App Rubric for Learning (MARuL), developed by Gladman et al [28], are mainly used to quantify an app's functionality, learning support effectiveness, usability, and overall satisfaction [30,31]. Such studies typically rely on participants' subjective evaluations, feature limited sample sizes, and often focus on single institutions or regions. Research perspectives also tend to be narrow, frequently emphasizing app effectiveness [32], usability [33], or overall user experience [34], while insufficient attention is given to cross-national differences and large-scale user feedback in real-world usage scenarios [35]. A growing body of research indicates that dynamic analysis based on longitudinal data is more effective than static, single-point-in-time assessments for revealing patterns in user experience and satisfaction over time and across app development stages [36,37]. Overall, existing studies remain constrained by limitations in sample size, data types, and temporal dimensions, making it challenging to systematically characterize the evolutionary process of medical education apps and the dynamic shifts in user satisfaction needs across different markets and cultural contexts.

Beyond these methodological limitations, a related conceptual gap is that medical education apps are often evaluated as functional toolkits rather than as sociotechnical interventions. Digital tools in professional education may reshape users' learning behaviors, daily routines, role expectations, and professional identity. In medical education, mobile technologies are not merely static tools for accessing learning materials; their design and use are shaped by educational theory, preclinical and clinical training contexts, and the needs of institutions, teachers, and learners [38]. Similar role-related effects have been observed in other professional service settings. For example, research on smart technologies in nursing homes showed that digital implementation changed caregivers' role expectations, work routines, performance norms, and adaptation strategies [39]. This perspective suggests that medical education apps should be examined not only as functional toolkits, but also as digital interventions embedded in professional training environments and institutional workflows.

Given these limitations, it is necessary to examine the evolution of medical education apps and the app-related factors associated with rating-based satisfaction patterns among app store reviewers from a systematic and dynamic perspective. As technology-driven educational products, medical education apps are also closely tied to institutional and cultural contexts in their usage and development. Therefore, understanding the evolution and use of medical education apps requires consideration of multiple dimensions, including changes in app quantity and functionality, as well as user needs and evaluation feedback. This study innovatively proposes a bidirectional "supply-demand" analytical framework. By integrating market evolution with user reviews, it explores functional adjustments, demand evolution, and cross-national differences in rating-based satisfaction patterns among app store reviewers in China and the United States, providing insights for app evaluation and design optimization. Based on the aforementioned research background and analytical framework, this study poses the following research questions (RQs):

- RQ1: What are the temporal evolution characteristics of market supply for medical education apps?
- RQ2: What app-related factors are associated with positive rating deviations (PD) and negative rating deviations (ND) among reviewers of medical education apps, and do these associations exhibit asymmetric patterns?
- RQ3: How can asymmetric patterns in PD and ND be used to classify review topics into Kano functional attributes across country-specific samples and temporal phases?

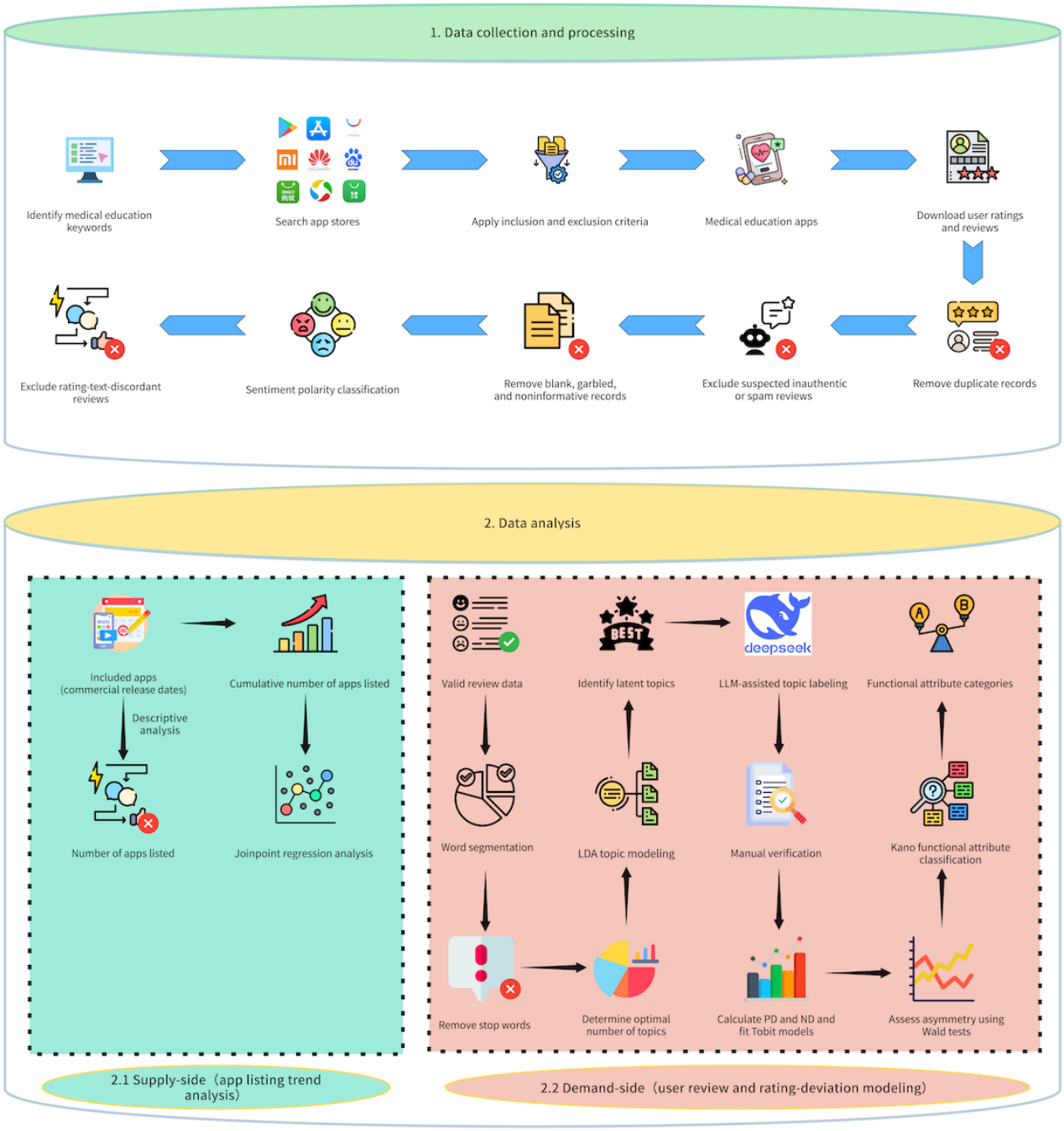
Methods

Overall Study Design and Analytical Framework

This study uses a longitudinal, cross-national mixed methods research design to systematically characterize the observable developmental features of medical education apps and rating-based satisfaction patterns among app store reviewers. The overall research approach introduces a bidirectional "supply-demand" analytical framework (Figure 1), integrating structural changes at the app market level with user experience feedback from real-world usage scenarios. This enables simultaneous capture of both the developmental trajectory of medical education apps and the dynamics of user perceptions. The analytical workflow comprised app identification and screening, extraction and preprocessing of ratings and reviews, and supply- and demand-side analyses.

XSL•FO
RenderX

Figure 1. Supply-demand analytical framework of medical education apps in China and the United States.



On the supply side, this study analyzes annual cumulative counts and temporal trends of observable and analytically eligible medical education apps from 2009 to 2025 based on app store data from China and the United States. Joinpoint regression models identify data-derived temporal segments and potential inflection points in observable app growth, thereby describing temporal patterns within the 2025-observable analytic cohort.

On the demand side, this study analyzes app store reviewers' primary concerns and rating patterns using large-scale user-generated review data. First, by combining latent Dirichlet allocation (LDA) topic modeling with large language model–assisted topic labeling using DeepSeek-R1, we identified key themes in app store reviews. Subsequently, Tobit regression models examined the associations between topic probabilities and PD and ND, while Wald tests assessed asymmetry between these associations. Building on these results, the Kano model was used to classify the identified topics into functional attributes according to their observed associations with rating deviations.

By integrating the supply-side analysis of observable app development patterns with the demand-side analysis of review topics and rating deviations, this study provides a unified framework for comparing medical education apps in China and the United States. This framework supports the comparison of data-derived temporal patterns and topic-rating deviation associations across the 2 country-specific samples, directly addressing the research questions posed in this study.

XSL•FO
RenderX

Protocol Registration and Deviations

This study was registered on the Open Science Framework [40]. The overall research objectives, cross-national comparative design, study scope, and supply-demand analytical framework remained consistent with the registered protocol. However, several deviations and methodological refinements were introduced during implementation because the availability, structure, and quality of the app store data could not be fully determined in advance. First, app identification was supplemented by screening related or recommended apps displayed by the app store platforms in addition to retrieval using the prespecified search terms. This refinement was introduced to reduce the risk of missing eligible apps that used atypical titles or descriptions. Second, the eligibility criteria and review preprocessing procedures were operationalized in greater detail, including the requirements concerning recent app updates, minimum review counts, duplicate and noninformative records (including developer replies), suspected inauthentic or spam reviews, and rating-text concordance. These procedures were specified to ensure that the final sample was observable, analytically eligible, and suitable for linking textual content with rating-based outcomes. Third, after LDA topic modeling, AI-assisted candidate topic labels were subjected to consensus-based review by 3 research team members and independent structured assessment by 2 domain experts. This additional human validation step was introduced to improve the interpretability and semantic validity of the topic labels. These deviations and refinements did not alter the study's primary objectives or overall analytical framework; they were made to improve app coverage, data quality, analytic feasibility, reproducibility, and methodological transparency.

Data Sources and Sample Construction

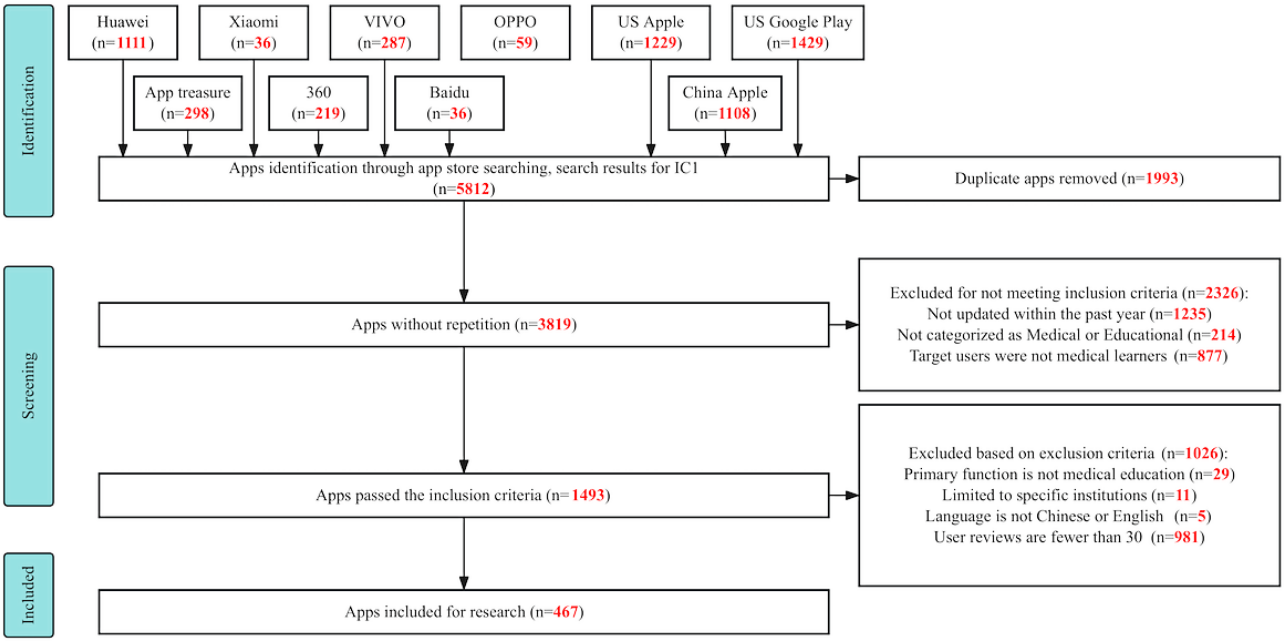
App Selection

To comprehensively cover medical education apps in China and the United States, this study systematically searched mainstream app stores in both countries in September 2025. In China, the search encompassed the Apple App Store and 7 major Android platforms (Huawei, Xiaomi, OPPO, VIVO, Baidu, 360, and App Treasure). In the United States, it covered the Apple App Store and Google Play. Searches used Chinese and English keywords (eg, "medical education," "medical examinations," "USMLE"; detailed strategy in Table S1 in Multimedia Appendix 1), initially identifying 5812 apps. This keyword-based app store search strategy was consistent with prior systematic searches of medical education apps, which used predefined search terms followed by title-and-description screening and eligibility assessment to identify relevant educational apps [31]. To reduce the risk of missing apps that used alternative descriptions, the keyword set was designed to include broad bilingual terms related to medical education, medical examinations, clinical learning, and professional training. In addition to keyword-based retrieval, related or recommended apps displayed by the app-store platforms were reviewed where available and screened using the same eligibility criteria. Retrieved apps were then screened according to app titles, descriptions, categories, and predefined eligibility criteria rather than keywords alone.

The predefined eligibility criteria were as follows. Apps were included if they (1) were identified through the prespecified Chinese or English search terms listed in Table S1 in Multimedia Appendix 1 or through related or recommended apps displayed by the app store platforms during the search process; (2) had an update record within the past year as shown in the app store at the September 2025 search cutoff; (3) were categorized under "Medical" or "Education"; and (4) primarily targeted medical learners, including medical students, residents, medical graduate school applicants, or users preparing for the National Medical Licensing Examination or the USMLE. Apps were excluded if they (1) primarily provided noneducational or nonsystematic medical or health-related functions, such as pharmaceutical electronic commerce, wellness or health preservation, news or information services, medical social networking, or doctor-patient communication; (2) were institution-restricted internal systems of hospitals or medical schools; (3) operated in languages other than Chinese or English; or (4) had fewer than 30 platform reviews as of the search cutoff. These criteria are also summarized in Table S2 in Multimedia Appendix 1. To ensure rigorous and objective screening, 2 researchers (JM and J Liang) independently conducted selection and evaluation based on these predefined criteria [41], with consistency verified via the kappa coefficient (Cohen κ=0.87). Disputed results were adjudicated by a third expert (YC). After removing 1993 duplicate apps, 3819 apps remained for screening. Of these, 2326 did not meet the inclusion criteria, leaving 1493 apps. A further 1026 apps were excluded based on the exclusion criteria. Ultimately, 467 medical education apps were included, comprising 192 apps from China and 275 apps from the United States (refer to Figure 2 for the complete screening process). This rigorous screening process ensured the systematic nature, comparability, and reproducibility of the sample. The final analytic sample was defined based on information available at the September 2025 search cutoff. Therefore, apps that had been discontinued before the search cutoff and were no longer searchable or had no retrievable metadata could not be systematically included in the historical denominator. The update-record and review-count criteria were used to define apps that were observable and analytically eligible at the time of data collection.

XSL•FO
RenderX

Figure 2. Flowchart of the medical education apps screening process. IC1: inclusion criterion 1; n: the number of apps.

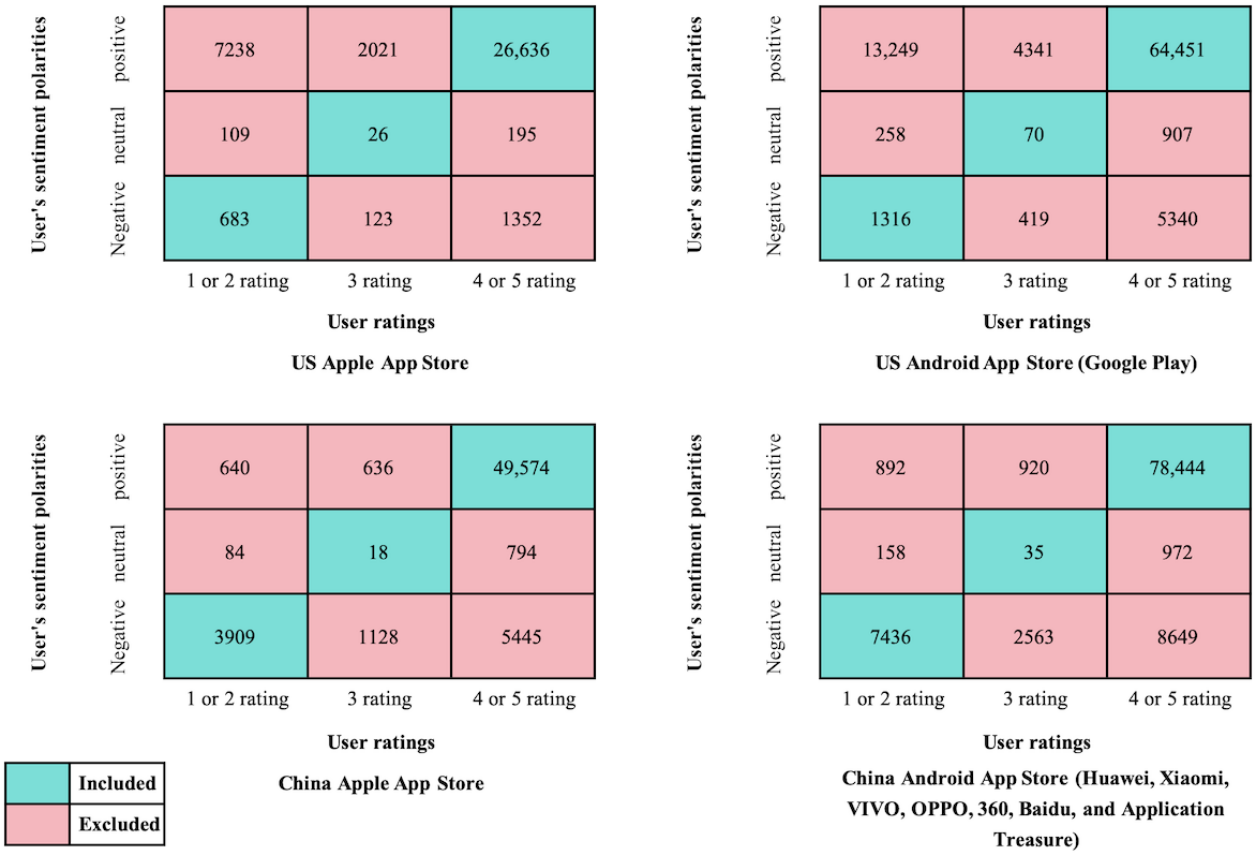


User-Generated Data

To obtain real-world user feedback, this study collected user ratings and review data for the 467 selected apps across 10 major app stores in China and the United States. As of October 2025, a total of 498,760 retrievable reviews were collected (China: 321,911; United States: 176,849). All reviews underwent standardized preprocessing, including removing duplicate records and records with missing or noninformative content (including official developer replies), as well as excluding suspected inauthentic or spam reviews [42]. Sentiment polarity was classified as negative, neutral, or positive using the Baidu AI Cloud Sentiment Analysis API [43]. Rating-text discordant reviews were then excluded as a quality-control step to construct a rating-text-concordant analytic subset (refer to Figure 3 and Note S1 in Multimedia Appendix 1). Figure 3 presents 4 platform groups: the Apple App Store in China, the 7 Chinese Android app stores combined, the Apple App Store in the United States, and Google Play in the United States. Rows represent negative, neutral, and positive sentiment, and columns represent ratings of 1-2 stars, 3 stars, and 4-5 stars. Some low-rated reviews in the Chinese Android app stores expressed positive sentiment, whereas some high-rated reviews in the US Apple App Store expressed negative sentiment. Rating-text discordant reviews were not interpreted as definitively invalid records, because prior online review studies have shown that numerical ratings and textual sentiment do not always align [44,45]. Such discordance may reflect mixed evaluations, feature-specific complaints, sarcasm, or rating noise. The detailed exclusion criteria, the number of reviews removed during each quantitative preprocessing step, and the proportions excluded by the rating-text concordance screen are provided in Note S1 in Multimedia Appendix 1. After preprocessing, the final rating-text-concordant analytic subset comprised 232,598 reviews (China: 139,416; United States: 93,182). Drawing from prior research [46], we operationalized the rating-based satisfaction proxy rate as the proportion of reviews with ratings of 4 stars or higher among app store reviewers. This proxy was not interpreted as a direct measure of satisfaction in the full user population. To eliminate rating scale differences across apps, user ratings were converted into deviation values relative to the corresponding app's overall rating. Based on this, PD and ND were calculated, both ranging from 0 to 4. To ensure analytic consistency, the final review-level dataset retained only records with valid app identifiers, review text, user ratings, and review dates. All subsequent review-level analyses, including topic modeling, Tobit regression, Wald tests, and Kano classification, were conducted using this rating-text-concordant analytic subset.

XSL•FO
RenderX

Figure 3. Matching between user sentiment polarities (rows) and user ratings (columns).



Sample Comparability

The Chinese and US samples maintained consistency in search time, platform coverage, inclusion criteria, and data processing procedures to ensure comparability across cultures. Through this methodology, this study constructed a 2-sided dataset that links observable app-market supply with app-store review-based demand signals, laying the foundation for subsequent supply-side trend analysis and demand-side rating-deviation modeling. The same search period, screening criteria, preprocessing workflow, and analytic procedures were applied to the Chinese and US samples to improve cross-national comparability. However, individual-level demographic information of app-store reviewers, such as age, sex, education level, professional background, and training stage, was not available from the app stores. Therefore, direct demographic adjustment was not possible. To reduce platform-related heterogeneity, we applied identical eligibility criteria, preprocessing procedures, rating-deviation construction, and analytic workflows across all app stores, and all rating-deviation models were estimated separately by country and developmental phase.

Supply-Side Analysis: Cumulative Observable App Counts and Temporal Trends

To characterize supply-side temporal patterns, this study used the cumulative count of apps that remained observable and analytically eligible at the September 2025 search cutoff as the core metric. In this study, "survival" was operationalized as an app being retrievable from the included app stores or Qimai Apps Data platform (Beijing QiMai Technology Co, Ltd) at the search cutoff, having an update record within the preceding year, and meeting the minimum review-count criterion. Accordingly, the annual cumulative survival counts should be interpreted as temporal patterns within the 2025-observable analytic cohort, rather than as a complete historical census of all medical education apps that had ever existed in each year. Therefore, the terms "survival" and "inflection points" in this study refer to app availability and trend shifts within the 2025-observable analytic cohort, not to the survival history of all medical education apps ever released. In time-series research, relevant literature has used joinpoint regression analysis to model trends in publicly available data (such as app counts, user interest, or technological attention) to identify significant inflection points in developmental trajectories and reveal phased market shifts [47]. This method avoids predefined time segmentation by identifying trend structure shifts directly from the data, effectively mitigating subjective classification biases. Drawing on this research paradigm and considering the potential phased development characteristics of the medical education app market, this study uses joinpoint regression to analyze the temporal trends in cumulative counts of observable and analytically eligible apps. This approach identifies data-derived temporal segments, providing methodological support for subsequent phase-based analysis grounded in user reviews.

Specifically, we used Joinpoint software (version 5.4.0; National Cancer Institute, November 2025) to construct regression models on a logarithmic scale. Using annual cumulative counts of observable and analytically eligible medical education apps as

XSL•FO
RenderX

the dependent variable, we estimated annual percentage changes (APCs) and average annual percentage changes (AAPCs) for each time segment, supplemented by pairwise tests to assess trend parallelism. The model allowed 0 to 3 joinpoints, selected based on data-driven Bayesian information criterion (BIC) with a global significance level of 0.05. Poisson SEs were used to estimate variability under the assumption of Poisson-distributed count data. Finally, the AAPC and its 95% CI for each trend segment were calculated from the slope of the log-transformed regression line, systematically revealing temporal dynamics.

Based on critical inflection points identified through joinpoint regression analysis, user review data underwent temporal segmentation. Analysis revealed key inflection points occurring in both China and the United States in 2017, dividing the data into 2 developmental phases: “2009-2017” and “2018-2025.” Although the first included Chinese app appeared in 2010, the first phase was labeled 2009-2017 to maintain a unified cross-national comparison window; the Chinese sample had no included apps in 2009. APC estimates that did not reach statistical significance were also considered in each phase to ensure that the subsequent analysis covered the entire temporal trajectory. Through this segmentation strategy, the study examined app-related factors associated with PD and ND and assessed how these associations varied across temporal phases.

Demand-Side Analysis: User Reviews and Rating-Deviation Modeling

Review Topic Identification (Text Processing and LDA Modeling)

To extract core themes from user reviews of medical education apps, the study first performed text preprocessing on Chinese and English reviews. This included tokenization using the jieba library [48] and the Natural Language Toolkit (NLTK), stopword removal using Chinese and English stopword lists, and cleaning numerical values, punctuation, and emoticons. Subsequently, the LDA topic model [49] was applied to analyze review data across different phases, identifying core themes and their distribution. The number of themes was determined comprehensively by balancing perplexity and coherence metrics with the interpretability of thematic clusters, ensuring themes were both statistically robust and understandable (Figures S1-S4 in Multimedia Appendix 1). This step structured review texts into quantifiable topic probability distributions, providing a quantitative foundation for the subsequent rating-deviation analysis. Visualization results, detailed text processing workflows, parameter settings, and thematic modeling specifics are documented in Figures S5-S8 and Note S2 in Multimedia Appendix 1. To address cross-linguistic differences, topic modeling was conducted separately for Chinese and English review corpora rather than by pooling translated texts. Chinese and English reviews were preprocessed using the same standardized preprocessing workflow, and topic labels were generated and validated separately within each country-phase corpus. For reporting and cross-national comparison, Chinese topic labels were translated into English after modeling and checked by bilingual researchers against the original high-probability keywords. Therefore, topics were compared at the level of conceptual domains, such as exam preparation, cost, technical stability, and learning support, rather than being treated as strictly one-to-one equivalent linguistic categories across languages.

Topic Naming and Semantic Validation (Assisted by DeepSeek-R1)

To enhance topic interpretability, DeepSeek-R1 was used to assist with the semantic labeling of topics generated by the LDA models. For each topic, the top 30 high-probability keywords were entered into the model to generate candidate topic names and brief descriptions (Table S3 in Multimedia Appendix 1). The model was accessed on October 26, 2025, with the temperature set to 0.6 and all other parameters kept at their default settings. After AI-assisted label generation, 3 members of the research team conducted a consensus-based semantic review by comparing each candidate topic name and description with the corresponding high-probability keywords. Disagreements were resolved through discussion, and the final topic labels were determined by human consensus. To provide additional verification, 2 domain experts independently assessed the finalized topic labels using a structured 5-point rubric for name accuracy and description interpretability (Table S4 in Multimedia Appendix 1). Because this validation task involved a small number of rubric-based semantic assessments of finalized topic names and descriptions, rather than independent categorical coding of individual reviews into topic categories, Cohen κ was not calculated. Instead, exact agreement between the two experts and expert validity scores were reported as descriptive indicators of topic-label validation quality. This human-in-the-loop approach was used to combine the efficiency of AI-assisted semantic labeling with the contextual judgment of human reviewers [50]. Generative AI was not used for data collection, review preprocessing, topic modeling, statistical analysis, result interpretation, or conclusion generation. To strengthen the computational validity and robustness of the text-mining procedure, we adopted a staged human-in-the-loop workflow. LDA was used as an unsupervised topic modeling approach to reduce high-dimensional review text into latent topic probability distributions, rather than to generate final conceptual categories directly. The number of topics was determined by combining perplexity, coherence, and human interpretability. AI-assisted candidate labels were subsequently reviewed through the consensus-based and expert-validation process described above. Similar natural language processing approaches have been used to analyze unstructured user-generated digital text in health-related contexts [51]. Therefore, the resulting topics should be interpreted as structured approximations of recurrent user-review themes rather than exhaustive or definitive categories of all user needs.

Asymmetric Modeling of Rating Deviations (Tobit Regression and Wald Test)

Guided by Herzberg two-factor theory, which distinguishes positive and negative evaluations, this study modeled PD and ND. PD and ND were used as dependent variables to represent the extent to which an individual review rating deviated above or below the corresponding app’s overall rating, while LDA topic probability distributions were used as independent variables to quantify the thematic composition of each review.

XSL•FO
RenderX

Because PD and ND are bounded dependent variables with a lower bound of zero, Tobit regression was used to examine the associations between review topics and rating deviations. Before model estimation, multicollinearity was assessed using variance inflation factors (VIFs); all VIF values were below 5, indicating no substantial multicollinearity [52] among the independent variables (Tables S5-S12 in Multimedia Appendix 1). Tobit regression is commonly used for limited dependent variables with restricted distributions [53,54]. Wald tests were subsequently conducted to test the null hypothesis that the PD coefficient for each topic equaled the sign-reversed coefficient from the corresponding ND model (H_0: $\beta_{PD,k} = -\beta_{ND,k}$), thereby evaluating asymmetry between the PD and ND associations. Detailed procedures for constructing PD and ND, model specification, and statistical testing are provided in Note S3 in Multimedia Appendix 1.

Functional Attribute Classification Using the Kano Model

Based on the asymmetric associations identified in the PD and ND models, each review topic was further classified into a functional attribute using the Kano model. The resulting categories included essential factors, expectation factors, charm factors, no differential factors, and reverse factors. The classification criteria integrated statistical characteristics from the PD and ND models, including coefficient significance, coefficient direction, and the significance of coefficient differences [55]. This classification was used to translate the observed rating-deviation patterns into potentially actionable categories for app design and optimization. Detailed classification criteria and procedures are provided in Figure S9, Table S13, and Note S4 in Multimedia Appendix 1.

Methodological Integration and Analytical Advantages

This study systematically analyzed reviews of medical education apps by integrating phase-specific topic modeling, asymmetric rating-deviation modeling, and Kano functional attribute classification. This methodological framework aligns the review analysis with the data-derived temporal segments identified from the observable app cohort, thereby supporting longitudinal comparisons. By modeling PD and ND separately, the analysis avoided collapsing the 2 directions of rating deviation into a single outcome. The Kano model then translated the observed topic-rating deviation associations into functional attribute categories that may inform the design and optimization of medical education apps.

Statistical Analysis

Data entry and organization for this study were performed using Microsoft Excel 2024. Data preprocessing, LDA topic modeling, and visualization were implemented using R 4.3.3 (R Core Team) and Python (*Gensim*, *PyLDAvis*, and other packages). The temporal trend analysis of cumulative counts of observable and analytically eligible medical education apps used the joinpoint regression method, executed using Joinpoint software (version 5.4.0, National Cancer Institute). Tobit regression and Wald tests of the associations between review topics and PD and ND were performed using Stata 18.0 (StataCorp LLC). All statistical tests were 2-tailed, with the significance level set at α=.05.

Ethical Considerations

All data used in this study were sourced from publicly accessible websites, and all user information was anonymized. Sensitive data were not collected for this study, and no compensation was provided for participants. All data relevant to this study were stored on a password-encrypted computer, and only the researchers had access to the data. Ethics approval was not required according to the ethical review by the ethics committee of the Second Affiliated Hospital Zhejiang University School of Medicine.

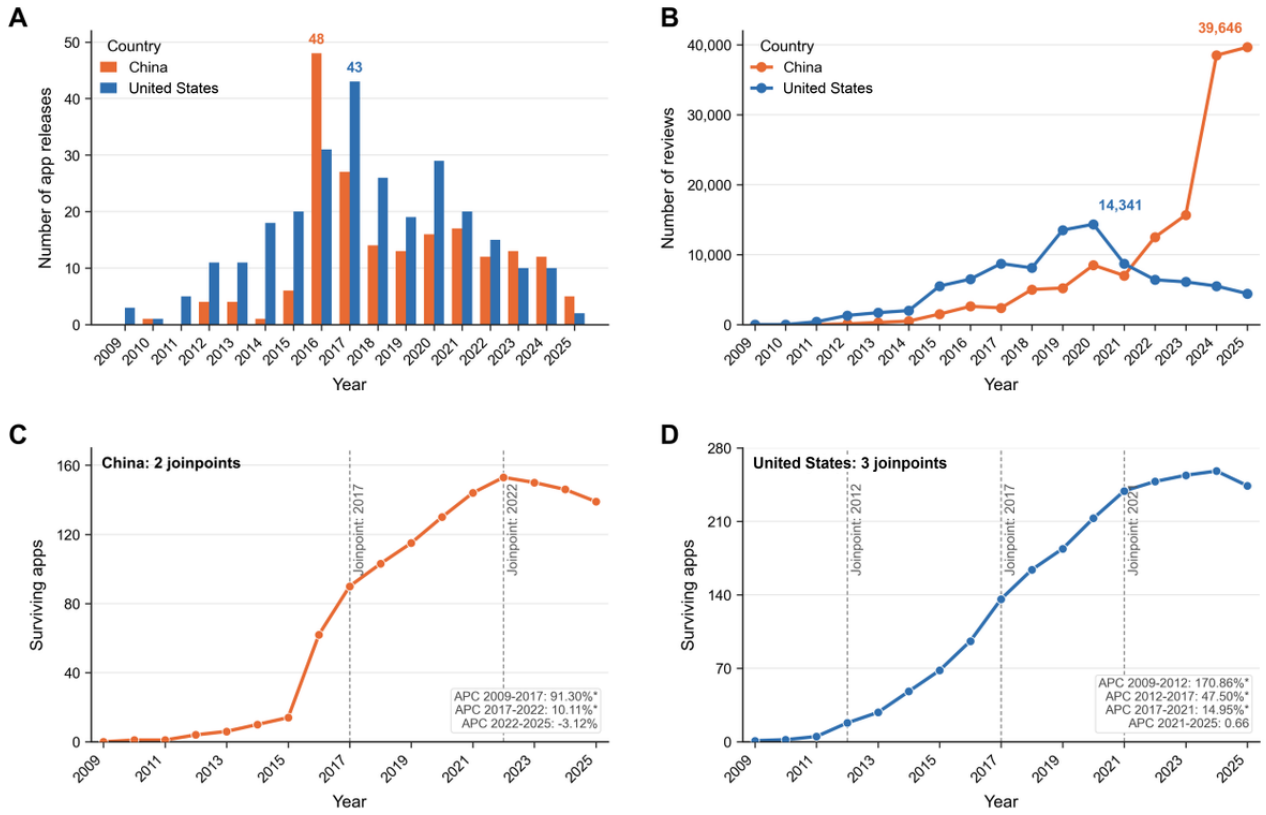
Results

Market Evolution and Phased Changes of Medical Education Apps in China and the United States

The final analytic sample included 467 medical education apps, comprising 192 apps from China and 275 apps from the United States. Among all reviews, 219,105 were rated 4 stars or higher, corresponding to an overall rating-based satisfaction proxy rate of 94.20%. By region, the rating-based satisfaction proxy rates for medical education apps in China and the United States were 91.82% (128,018/139,416) and 97.75% (91,087/93,182), respectively. This baseline difference in the rating-based satisfaction proxy rate was interpreted as a descriptive cross-national contrast rather than as direct evidence of higher or lower educational quality in either market. Because app store ratings are generated by self-selected reviewers, the observed difference may reflect multiple factors, including app quality, platform review moderation practices, reviewer composition, and culturally shaped rating behaviors. Therefore, we further examined whether the baseline rating difference was accompanied by distinct review topics and asymmetric rating-deviation patterns across countries and temporal phases. Cumulative downloads reached 361,367,013 for Chinese apps and 28,189,607 for US apps. Figure 4 presents the annual numbers of app releases (Figure 4A), annual user review counts (Figure 4B), and joinpoint regression analyses of annual cumulative observable app counts for China (Figure 4C) and the United States (Figure 4D) within the 2025-observable analytic cohort. Apps in the Chinese market first appeared in 2010, reaching an annual release peak of 48 in 2016, followed by a gradual decline in new app releases each year thereafter. The US market began in 2009, peaking in 2017 (43 apps), and also showed a declining trend thereafter. Therefore, the 2009-2017 phase for China should be understood as a harmonized comparative period, with no included Chinese apps observed in 2009. Regarding user reviews, the Chinese market exhibited sustained overall growth; the US market reached a review peak in 2020 (14,341 reviews) before gradually declining.

XSL•FO
RenderX

Figure 4. Comparison of development trends in medical education apps between China and the United States: annual number of app releases, user review volume, and joinpoint regression analysis results. APC: annual percentage change.



Joinpoint results indicate that, within the 2025-observable analytic cohort, the annual cumulative counts of observable and analytically eligible medical education apps in both China and the United States showed significant growth (P<.001) throughout the observation period. The AAPCs were 41.69% (95% CI 31.69-52.45) and 41.17% (95% CI 31.61-51.42) for China and the United States, respectively, indicating similar overall growth rates. Specifically, China's cumulative observable app counts exhibited distinct inflection points in 2017 and 2022, revealing a 3-phase evolutionary pattern: rapid growth from 2009 to 2017 (APC=91.30%, 95% CI 64.36-122.66; P<.001), decelerating growth from 2017 to 2022 (APC=10.11%, 95% CI 0.68-20.43; P=.04), and a nonsignificant decline after 2022 (APC=–3.12%, P=.59). The cumulative observable app counts in the United States exhibited turning points in 2012, 2017, and 2021, following a 4-phase trajectory: a steep rise from 2009 to 2012 (APC=170.86%, 95% CI 72.86-324.44; P=.002), followed by sustained high growth from 2012 to 2017 (APC=47.50%, 95% CI 38.78-56.77; P<.001), a further deceleration in growth from 2017 to 2021 (APC=14.95%, 95% CI 9.36-20.83; P<.001), and no significant change after 2021 (P=.58).

Topic Distribution in User Reviews of Medical Education Apps

Tables 1 and 2 illustrate the thematic distribution of user reviews for medical education apps in China and the United States across different developmental stages. Overall, user-focused topics exhibit phased evolution over time. In China, from 2009 to 2017, users primarily emphasized "Clinical practicality" and "Technical stability," whereas from 2018 to 2025, the focus shifted toward "Practice test function." In the United States, users primarily focused on "Functional availability" from 2009 to 2017, while from 2018 to 2025, attention shifted toward "Cost" and "Exam preparation and practice questions." Detailed analysis results are provided in Note S5 in Multimedia Appendix 1.

XSL•FO
RenderX

Table 1. Review topics and keywords identified using latent Dirichlet allocation topic modeling (China).

Stage and topics	Keywords	Reviews, n (%)
2009-2017		
Topic 1: Clinical practicality	clinical, practical, guide, job, function, content	2735 (37.03)
Topic 2: Professional learning community	platform, communication, discussion, share, recommended, medical students	1070 (14.49)
Topic 3: Technical stability	flashback, unable, solve, update, installation, professional version	1555 (21.06)
Topic 4: Medical knowledge question bank	medications, exam, e-book, complete, physician, instructions	694 (9.40)
Topic 5: Cost	charge, cheap, offline, version, time, use	203 (2.75)
Topic 6: Content reliability	professional, certification, medical school, course, case, bone	325 (4.40)
Topic 7: Learning and practice support	postgraduate entrance examination, medical record, patient, work, growth, companionship	511 (6.92)
Topic 8: Infant medication inquiry	baby, medicine, lactation, internet, department, not complete	292 (3.95)
2018-2025		
Topic 1: Practice test function	practice problems, title, analysis, past exam papers, question mistakes, postgraduate entrance exam	53,929 (40.85)
Topic 2: Audio-visual teaching function	explanation, course, video, lecture, teacher, interesting	11,197 (8.48)
Topic 3: Medical knowledge question bank	question bank, question type, content-rich, software, exam preparation, easy to understand	12,518 (9.48)
Topic 4: Functional usability	like, gospel, with heart, satisfied, treasure, value for money	10,309 (7.81)
Topic 5: Cost	free, advertisement, charges, purchase, unlock, money	10,860 (8.23)
Topic 6: Technical stability	crash, collapse, unable to open, operation, resolve, type	9230 (6.99)
Topic 7: Professional learning community	study, share, companionship, good helper, appreciate, skill	4980 (3.77)
Topic 8: Nurse qualification examination	nursing certification exam, occupation, nursing licensing examination, anatomy, attend lectures, subject	6250 (4.73)
Topic 9: User-perceived value	recommended, worth, premium, support, knowledge, anatomy	1837 (1.39)
Topic 10: Personalized learning	precision, relaxed, smoothly, test center, function, convenient and quick	1063 (0.81)
Topic 11: User service	customer service, update, refund, cancel registration, reply, store, answering questions	2331 (1.76)
Topic 12: Cross-screen adaptation function	tablet, landscape mode, full screen, adapt, switch, play	7527 (5.70)

XSL•FO
RenderX

Table 2. Review topics and keywords identified using latent Dirichlet allocation topic modeling (United States).

Stage and topics	Keywords	Reviews, n (%)
2009-2017		
Topic 1: Practice test function	questions, quiz, practice, exam, explanation, study	72 (0.28)
Topic 2: Usability	simple, guide, step, learn, fun, educational	1283 (4.91)
Topic 3: Functional availability	feature, add, search, offline, mobile, flashcard	20,736 (79.39)
Topic 4: Interface design and visual clarity	interface, picture, organize, easy, understand, quick	1083 (4.15)
Topic 5: Content accuracy and reliability	accurate, reliable, medical, resource, professional, student	677 (2.59)
Topic 6: Clinical teaching practice	doctor, surgery, patient, informative, teach, reference	957 (3.66)
Topic 7: Cost	free, pay, price, buy, purchase, offer	126 (0.48)
Topic 8: Learning management and note-taking	set, create, note, edit, delete, sync	1045 (4)
Topic 9: Technical stability	update, issue, download, load, Android, check	139 (0.53)
2018-2025		
Topic 1: Cost	free, pay, money, subscription, premium, purchase	27,401 (40.86)
Topic 2: Exam preparation and practice questions	quiz, exam, test, practice, pass, game	9000 (13.42)
Topic 3: Error analysis and memorization function	wrong, explanation, rationale, review, flash, revision	6000 (8.95)
Topic 4: Personalized learning services	feature, track, progress, option, set, enjoy	4413 (6.58)
Topic 5: Nursing qualification exam	nursing, preparation, aspirant, exercise, learn, knowledge	3948 (5.89)
Topic 6: Clinical teaching practice	anatomy, physiology, clinical, flashcard, skill, understanding	4137 (6.17)
Topic 7: Convenience	simple, quick, resource, reference, platform, recommend	2911 (4.34)
Topic 8: Teaching quality	easy, understand, explain, lecture, teacher, teach	1373 (2.05)
Topic 9: Learning outcomes and confidence building	grade, confident, effective, accurate, simply, ready	7156 (10.67)
Topic 10: Clinical decision support	patient, treatment, medication, quality, science, AI	725 (1.08)

Associations Between Review Topics and Rating Deviations

Tobit Regression Results for PD and ND

The Tobit models revealed country-specific and phase-specific patterns in the associations between review topics and rating deviations. In China, reviews from the early phase were more strongly characterized by clinical practicality, content reliability, technical stability, cost, and learning support, whereas reviews from the later phase were more concentrated around examination preparation, question banks, user service, technical stability, and cost-related experiences. In the United States, reviews from the early phase were more strongly characterized by functional availability, interface clarity, content reliability, and usability, whereas reviews from the later phase showed greater attention to cost, exam preparation, convenience, and learning support. These patterns indicate that rating deviations were associated with different app-related topics across countries and temporal phases. Tables 3 and 4 present the Tobit regression results for medical education apps in China and the United States. These tables report complete estimates for all LDA-derived topics before subsequent factor discrimination and Kano classification. Overall, both the PD and ND models showed that multiple review topics were significantly associated with rating deviations. In China, during Phase I (2009-2017), content reliability showed the strongest positive association with PD (regression coefficient β=0.488, P<.001), whereas cost showed the strongest negative association with PD (β=–0.459, P<.001). In the ND model, cost showed the strongest positive association with ND (β=2.704, P<.001), whereas technical stability showed the strongest negative association with ND (β=–4.464, P<.001). During Phase II (2018-2025), user service showed the strongest positive association with PD (β=0.196, P<.001), whereas user-perceived value showed the strongest negative association with PD (β=–0.134, P<.001). In the ND model, user-perceived value showed the strongest positive association with ND (β=3.368, P<.001), whereas the nurse qualification examination showed the strongest negative association with ND (β=–1.774, P<.001).

XSL•FO
RenderX

Table 3. Tobit regression results for rating deviations in China.

Stage and review topic	Model 1 (PD[a]) β	SE	*P* value	Model 2 (ND[b]) β	SE	*P* value
2009-2017						
Topic 1: Clinical practicality	0.254	0.034	<.001	–0.051	0.179	.77
Topic 2: Professional learning community	–0.445	0.036	<.001	2.432	0.170	<.001
Topic 3: Technical stability	0.180	0.032	<.001	–4.464	0.240	<.001
Topic 4: Medical knowledge question bank	0.380	0.047	<.001	–0.988	0.210	<.001
Topic 5: Cost	–0.459	0.071	<.001	2.704	0.226	<.001
Topic 6: Content reliability	0.488	0.070	<.001	0.779	0.232	.001
Topic 7: Learning and practice support	0.126	0.037	.001	–1.876	0.254	<.001
Topic 8: Infant medication inquiry	0.024	0.040	.55	–1.166	0.308	<.001
2018-2025						
Topic 1: Practice test function	0.031	0.004	<.001	–1.214	0.039	<.001
Topic 2: Audio-visual teaching function	–0.108	0.005	<.001	3.286	0.040	<.001
Topic 3: Medical knowledge question bank	–0.025	0.004	<.001	0.584	0.045	<.001
Topic 4: Functional usability	0.087	0.005	<.001	–1.503	0.025	<.001
Topic 5: Cost	0.042	0.004	<.001	–1.441	0.051	<.001
Topic 6: Technical stability	0.153	0.006	<.001	–1.127	0.054	<.001
Topic 7: Professional learning community	0.115	0.006	<.001	–1.672	0.067	<.001
Topic 8: Nurse qualification examination	0.097	0.005	<.001	–1.774	0.065	<.001
Topic 9: User-perceived value	–0.134	0.009	<.001	3.368	0.052	<.001
Topic 10: Personalized learning	0.021	0.009	.03	1.329	0.073	<.001
Topic 11: User service	0.196	0.008	<.001	–0.774	0.070	<.001
Topic 12: Cross-screen adaptation function	0.124	0.005	<.001	–0.626	0.049	<.001

[a]PD: positive rating deviations. The log likelihoods for Model 1 were –9967.3345 (2009-2017) and –10993.34 (2018-2025).

[b]ND: negative rating deviations. The log likelihoods for Model 2 were –5175.8596 (2009-2017) and –62667.978 (2018-2025).

XSL•FO
RenderX

Table 4. Tobit regression results for rating deviations in the United States.

Stage and review topic	Model 1 (PD[a]) β	SE	*P* value	Model 2 (ND[b]) β	SE	*P* value
2009-2017						
Topic 1: Practice test function	–0.883	0.017	<.001	0.271	0.064	<.001
Topic 2: Usability	0.127	0.013	<.001	–0.413	0.053	<.001
Topic 3: Functional availability	0.128	0.031	<.001	–0.342	0.107	.001
Topic 4: Interface design and visual clarity	0.216	0.013	<.001	–0.490	0.057	<.001
Topic 5: Content accuracy and reliability	0.185	0.014	<.001	–0.605	0.062	<.001
Topic 6: Clinical teaching practice	–0.018	0.013	.16	0.216	0.050	<.001
Topic 7: Cost	–0.072	0.019	<.001	1.537	0.069	<.001
Topic 8: Learning management and note-taking	–0.121	0.015	<.001	0.863	0.051	<.001
Topic 9: Technical stability	–0.017	0.015	.27	0.164	0.064	.01
2018-2025						
Topic 1: Cost	–0.033	0.004	<.001	0.382	0.027	<.001
Topic 2: Exam preparation and practice questions	0.066	0.005	<.001	–0.449	0.032	<.001
Topic 3: Error analysis and memorization function	0.125	0.005	<.001	–0.804	0.038	<.001
Topic 4: Personalized learning services	0.062	0.006	<.001	–0.041	0.035	.24
Topic 5: Nursing qualification exam	–0.188	0.006	<.001	1.157	0.034	<.001
Topic 6: Clinical teaching practice	–0.131	0.006	<.001	0.394	0.032	<.001
Topic 7: Convenience	0.168	0.006	<.001	–0.574	0.041	<.001
Topic 8: Teaching quality	–0.004	0.008	.64	0.112	0.045	.01
Topic 9: Learning outcomes and confidence building	–0.070	0.004	<.001	–0.888	0.034	<.001
Topic 10: Clinical decision support	–0.019	0.011	.08	0.286	0.057	<.001

[a]PD: positive rating deviations. The log likelihoods for Model 1 were –18560.334 (2009-2017) and –44306.823 (2018-2025).

[b]ND: negative rating deviations. The log likelihoods for Model 2 were –14143.986 (2009-2017) and –39305.998 (2018-2025).

In the United States, during Phase I (2009-2017), interface design and visual clarity showed the strongest positive association with PD (β=0.216, *P*<.001), whereas the practice test function showed the strongest negative association with PD (β=–0.883, *P*<.001). In the ND model, cost showed the strongest positive association with ND (β=1.537, *P*<.001), whereas content accuracy and reliability showed the strongest negative association with ND (β=–0.605, *P*<.001). During Phase II (2018-2025), convenience showed the strongest positive association with PD (β=0.168, *P*<.001), whereas the nursing qualification exam topic showed the strongest negative association with PD (β=–0.188, *P*<.001). In the ND model, the nursing qualification exam topic showed the strongest positive association with ND (β=1.157, *P*<.001), whereas learning outcomes and confidence building showed the strongest negative association with ND (β=–0.888, *P*<.001). Complete model results are provided in Note S6 in Multimedia Appendix 1.

Asymmetry Between PD and ND Coefficients

The Wald tests compared the PD coefficient for each review topic with the sign-reversed coefficient from the corresponding ND model. A significant difference in this comparison indicated asymmetry between the positive and negative rating-deviation associations. In the Chinese sample, during Phase I (2009-2017), all topics included in the Wald comparison except “Clinical practicality” showed significant differences. During Phase II (2018-2025), all included topics showed significant differences. In the US sample, all topics included in the Wald comparison showed significant differences during Phase I (2009-2017). During Phase II (2018-2025), all included topics except “Personalized learning services” showed significant differences. Detailed results are presented in Tables S14 and S15 and Note S7 in Multimedia Appendix 1.

Kano Classification of Review Topics

Furthermore, the Kano classifications of review topics differed across countries and temporal phases. In the Chinese sample, Phase I (2009-2017) included 1 essential factor (infant medication inquiry) and 6 charm factors, whereas Phase II (2018-2025) included 1 essential factor (user-perceived value) and 10 charm factors. In the US sample, Phase I (2009-2017) included 2 essential factors (clinical teaching practice and technical stability) and 7 charm factors, whereas Phase II (2018-2025) included 2 essential factors (teaching quality and clinical decision support) and 7 charm factors. Classification results and the corresponding classification indicators are presented in Figure S10 and Tables S16-S19 in Multimedia Appendix 1, respectively. Detailed analysis is provided in Note S8 in Multimedia Appendix 1. Taken together, the Kano classification suggested that app store reviewers in both

XSL•FO
RenderX

countries increasingly distinguished between baseline requirements and value-added features. Essential factors reflected functions whose absence was more likely to generate dissatisfaction, such as technical stability, clinical teaching practice, teaching quality, or clinical decision support, depending on country and phase. Charm factors reflected features that could increase satisfaction when present, particularly exam preparation, learning management, personalized learning, convenience, user service, and cross-screen adaptation. This distinction helped translate the Tobit regression results into more actionable insights for app design by separating foundational quality expectations from features that may create additional perceived value.

Discussion

Principal Findings

This study aimed to characterize the longitudinal evolution of medical education apps in China and the United States and to identify app-related factors associated with PD and ND among app store reviewers. Overall, within the 2025-observable analytic cohort, the findings showed phased development patterns, with notable data-derived inflection points around 2017 and 2021-2022, while country-specific trajectories differed in observable app availability, review activity, and user-evaluation patterns. On the demand side, Chinese app store reviewers showed stronger associations between rating deviations and exam preparation, cost, question banks, personalized learning, and user service, whereas US app store reviewers placed greater emphasis on technical stability, functional availability, content reliability, and clinical decision support. These findings address the study objectives by linking supply-side observable app development patterns with demand-side user review patterns in 2 distinct medical education contexts.

Phased Evolutionary Characteristics of Medical Education Apps in China and the United States

Phased Inflection Points and Synchronization

Based on joinpoint regression results, the 2025-observable medical education app cohorts in China and the United States exhibited similar data-derived temporal inflection points during the study period, primarily concentrated in 2017 and 2021-2022. The apparent similarity in these temporal patterns may be interpreted in relation to shared contextual developments, including the global diffusion of mobile internet technology, changes in digital medical education platforms, and shifts in the external environment. However, these factors were not directly measured, and the observed similarity should not be interpreted as evidence that they caused the temporal patterns. From a technological perspective, policy updates on the iOS platform may provide contextual support for interpreting the observed temporal patterns. Starting with iOS 11, all 32-bit apps were no longer supported, requiring developers to update their apps to a 64-bit architecture for operation [56]. As early as 2015, Apple mandated that new apps and updates support 64-bit architecture [57]. This policy may have contributed to the withdrawal of outdated or unmaintained applications by increasing technical requirements for app development and maintenance, and may therefore have influenced the composition of the observable app cohort during the study period. From a technology adoption perspective, while the TAM was introduced to medical education in 2010, it remained in an exploratory phase until 2016, with research primarily focused on electronic learning and mobile learning [58]. These developments in technology adoption may have created contextual conditions that supported growing awareness and use of mobile learning tools around 2017. Institutionally, China formally advanced its integrated clinical medical education system (“5+3” [5-year undergraduate + 3-year residency] and “3+2” [3-year specialty + 2-year assistant general practice training]) alongside medical education coordination reforms in 2017, emphasizing seamless integration between undergraduate education, residency training, and clinical practice [59]. This reform may have increased attention to standardized and practice-oriented clinical skills training and may have created an institutional context supportive of digital and mobile learning tools.

Changes in the external environment may also provide contextual explanations for the observed temporal patterns. Previous research has reported increased use of mobile medical education apps among medical students during COVID-19 lockdowns, with use remaining elevated after the pandemic [60]. This evidence may provide contextual support for interpreting the data-derived temporal changes observed around 2021-2022. More broadly, the increasing use of mobile learning may reflect its transition from a supplementary resource toward a more established educational approach.

Divergent Market Trajectories in China and the United States

Despite similar data-derived inflection points, the observable cohort-based trajectories in China and the United States showed notable differences. Within the 2025-observable analytic cohort, China showed rapid early expansion followed by slower growth, while user reviews continued to grow. This may suggest that review activity and expressed engagement increased among app-store reviewers even as observable app growth slowed. In contrast, the US sample showed declining app listings after 2017, with review volumes peaking around 2020 before gradually declining. This pattern may be consistent with product consolidation within the observable app cohort and changing review activity in a more mature platform environment. Overall, the similarity in inflection points may reflect shared pressures in the global digital transformation of medical education, whereas the divergent trajectories may be related to differences in market maturity, institutional environments, and platform competition dynamics.

Dynamic Differences in Rating-Based Satisfaction Patterns for Medical Education Apps in China and the United States

The observed pattern of rating-based satisfaction among Chinese app store reviewers showed distinct phased changes. During the early stage (2009-2017), reviewers emphasized clinical practicality, technical stability, and professional learning community. These patterns indicate that the directions and magnitudes of associations with PD and ND varied across review topics. In the later phase, reviewer attention was more

XSL•FO
RenderX

frequently directed toward exam preparation drills, question banks, personalized learning, user services, and cross-screen compatibility. These findings indicate that greater positive rating deviations were most strongly associated with user service, technical stability, and cross-screen adaptation, whereas greater NDs were most strongly associated with user-perceived value and audio-visual teaching function. This demand shift may be consistent with an assessment-driven learning context in high-stakes medical education. Because this study did not measure learning processes or outcomes, it cannot determine whether greater attention to question banks reflects rote memorization or is associated with superficial learning. As a broader educational design consideration, app-based exam preparation could be combined with activities such as problem-based learning (PBL) case discussions to support critical thinking and deep learning.

The observed pattern of rating-based satisfaction among US app store reviewers showed a somewhat different trajectory. In the early phase (2009-2017), reviewers appeared to prioritize functional availability, interface design, and content reliability, suggesting that these topics were associated with smaller negative rating deviations. In the later phase (2018-2025), greater PDs were associated with exam preparation and practice questions, error analysis and memorization, personalized learning services, and convenience, whereas greater NDs were associated with cost, nursing qualification exam, clinical teaching practice, teaching quality, and clinical decision support. These observed associations may be considered through self-determination theory as a possible interpretive lens [61]. From this perspective, stable technology and authoritative content could be perceived as supporting a trustworthy learning environment and a sense of competence, while access to clinical resources and guideline-search functions may be consistent with learner autonomy.

Cross-national comparisons indicated that app store reviewers in both countries showed a gradual shift in focus from core learning and technical functions toward exam preparation, question banks, and personalized learning. This observed pattern may be considered in relation to product iterations and the broader institutional contexts in which medical education apps are used. Country-specific patterns were also observed: among Chinese app store reviewers, rating deviations were more closely associated with service experience, cost, and exam-related and personalized features [62], whereas US app store reviewers showed stronger associations with the stability and professional reliability of core functions [63]. The shared attention to “personalized learning” among app store reviewers in both countries may suggest growing interest in adaptive learning models. Such systems could customize learning paths by adjusting content difficulty to learners’ ability levels, potentially helping manage cognitive load and learning efficiency. Future medical education apps could move beyond content aggregation toward intelligent tutoring systems that are responsive to learners’ real-time states, although such designs require further validation using learning process and outcome data. Developers could also consider optimizing exam support, personalized learning, and user service functions for different educational contexts, with the effects of such design changes requiring further evaluation.

From a functional classification perspective, the Kano model results indicated that review topics were primarily classified as essential or charm factors across country-phase models, whereas expectation factors showed no stable pattern. These classifications suggest that the associations between review topics and rating deviations varied across countries and temporal phases rather than following a single pattern. Essential factors differed across country-phase models, while exam support, learning management, and personalized features were more frequently classified as charm factors in particular models. This pattern suggests that functional innovation may contribute to competitive differentiation among medical education apps.

This study conducted a cross-national comparison of the observable development and rating-based satisfaction patterns of medical education apps in China and the United States. Within the 2025-observable analytic cohorts, data-derived temporal changes were identified around 2017 and 2021-2022. Although the timing of these changes was similar, the country-specific trajectories differed: the Chinese cohort showed rapid early expansion followed by slower growth while review activity continued to increase, whereas the US cohort showed declining app listings after 2017 and changing review activity after 2020. The similarity in timing may be considered in relation to shared contextual developments, including the global diffusion of mobile internet technology, institutional reforms, and shifts in the external environment, whereas the divergent trajectories may be considered in relation to differences in institutional environments, market maturity, and platform competition dynamics.

Regarding rating-based satisfaction patterns, this study identified phased shifts in app store reviewers’ priorities between China and the United States. In the Chinese sample, reviewers’ attention shifted from clinical utility and interactive learning in the early phase toward exam preparation, question banks, personalized learning, user service, and cross-screen compatibility in the later phase, while the directions of topic-rating-deviation associations varied across topics and phases. In the US sample, reviewers’ attention shifted from functional availability, interface design, and content reliability in the early phase toward cost, exam preparation, and practice questions, and convenience-related features in the later phase. Cross-national comparisons suggest that while functional evolution paths differed between countries, reviewers in both samples showed increasing attention to personalized learning, which may be consistent with growing interest in adaptive learning models in educational technology.

This study holds significant theoretical and practical implications. Theoretically, this study extends the cross-national evidence on mobile learning and rating-based satisfaction and situates the observed review patterns in relation to the TAM and educational evaluation theories. Educational systems and cultural contexts may provide useful interpretive context for the observed differences in user review patterns. Practically, the findings identify potential design priorities for medical education app developers, including core functional stability,

XSL•FO
RenderX

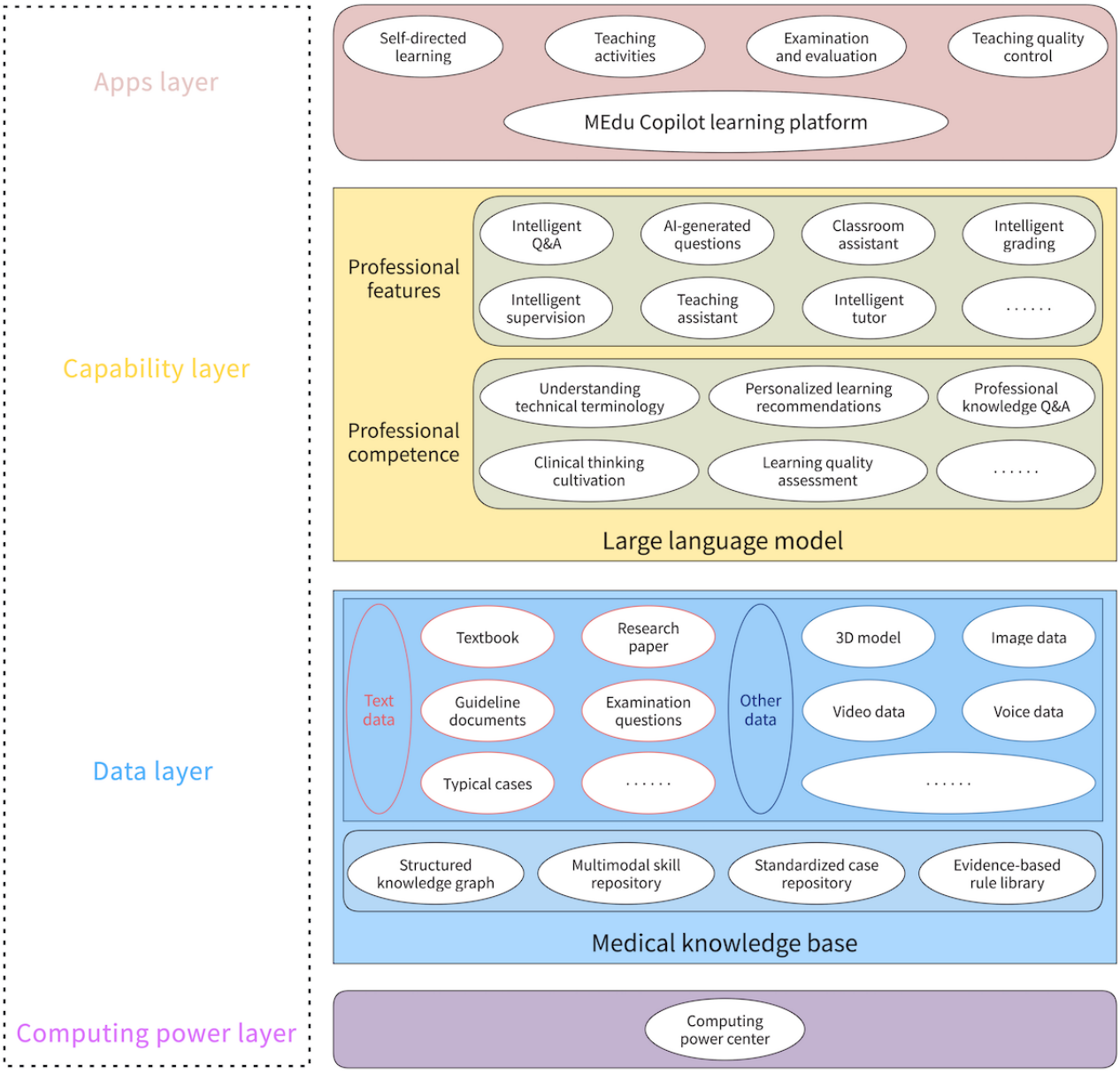
professional reliability, exam preparation support, personalized learning, and user service features. For educational administrators and policymakers, the observed patterns may inform the planning and evaluation of digital teaching resources in different institutional and educational contexts.

Design Implications for AI-Supported Medical Education Platforms

Before translating these findings into a design framework, it is important to interpret the observed cross-national differences within their institutional and educational contexts. In China, the stronger association of rating deviations with exam preparation, question banks, cost, and user service may be considered in relation to a standardized and assessment-oriented medical training environment, in which affordable and efficient exam-related support may be particularly salient. In the United States, the stronger associations with technical stability, functional availability, content reliability, and clinical decision support may be considered in relation to a credentialing and self-directed learning context, in which stable access to trustworthy resources may be particularly salient across clinical and educational settings.

On the basis of these empirically observed and contextually interpreted review and rating-deviation patterns, we proposed a conceptual design framework, termed "MEdu Copilot," to illustrate potential design implications for future AI-supported medical education platforms (Figure 5). This framework conceptually translates the user review patterns and topic-rating deviation associations identified in China and the United States into potential design strategies for intelligent recommendation. Building on these findings, it illustrates a goal-oriented design approach that could inform the development of intelligent assistance for knowledge acquisition, clinical reasoning, and professional competency development.

Figure 5. MEdu Copilot conceptual framework.



XSL•FO
RenderX

Figure 5 presents a proposed layered architecture for a future AI-supported medical education platform. The computing layer represents the computational infrastructure; the data layer represents a "Medical Knowledge Base" comprising multimodal and structured resources, such as textbooks, guidelines, clinical cases, 3D models, videos, images, and examination questions. The capability layer centers on a large language model that could support functions such as terminology assistance, clinical reasoning support, intelligent question answering, classroom assistance, and automated examination generation and grading. The application layer maps these potential capabilities to educational scenarios, including self-directed learning, instructional activities, assessment, and quality monitoring. The framework illustrates possible relationships among infrastructure, data, AI-supported capabilities, and educational apps; it does not represent a validated or implemented platform.

This conceptual framework follows a user-centered, goal-oriented approach that connects the user and technical design space through an iterative feedback mechanism. In a future implementation, user interaction data and learning outcomes could inform the adaptive refinement of learning resources, instructional support tools, and assessment strategies. Privacy protection, trust, and cultural sensitivity are proposed as guiding design principles that may help align future platform development with expectations for accountability and inclusivity in digital health interventions. Overall, the framework illustrates a potential pathway for the responsible integration of AI-supported capabilities into medical education.

Limitations

This study has several limitations. First, the analytic sample was defined at the September 2025 search cutoff. Although broad bilingual keywords, related-app screening where available, and title-description screening were used to improve app identification, apps discontinued earlier, no longer searchable, lacking retrievable metadata, or using atypical titles or descriptions may still have been missed. Therefore, supply-side trends should be interpreted as patterns within the 2025-observable analytic cohort, rather than as a complete historical census of medical education apps. Because the temporal segmentation was derived from cumulative counts within the 2025-observable analytic cohort, phase-based demand-side comparisons should be interpreted as exploratory comparisons across data-derived temporal segments rather than as definitive comparisons across complete historical market stages.

Second, app store reviews provide a self-selected and rating-based source of user feedback. These reviews may underrepresent "silent users" who used apps but did not leave ratings or comments. They may also overrepresent users with particularly positive experiences or strong technical frustrations. Because reviewer-level demographic, professional, training-stage, and app-use metadata were unavailable, we could not directly adjust for demographic differences across app stores or assess subgroup-specific satisfaction patterns. Accordingly, rating-based satisfaction metrics and topic patterns should be interpreted as rating- and review-based indicators among app store reviewers, rather than population-level estimates for all medical education app users or learners. In addition, all review-level analyses were conducted using the rating-text-concordant analytic subset. Although this concordance criterion was intended to improve consistency between numerical ratings and review text, it may also have excluded meaningful mixed, sarcastic, or feature-specific evaluations. Accordingly, the possibility of selection bias should be considered when interpreting the review-level findings.

Third, topic modeling and cross-linguistic comparison involve interpretive uncertainty. Although Chinese and English reviews were modeled separately and topic labels were checked through bilingual semantic validation, differences in sentiment expression and topic wording may still affect comparability across countries. The AI-assisted topic labels should therefore be understood as structured representations of review themes, rather than definitive categories of user needs.

Finally, the observational design limits causal interpretation. Reviews reflect subjective perceptions rather than actual learning outcomes, and the data do not contain individual-level information on educational status, examination participation, institutional affiliation, or cultural orientation. Interpretations related to national educational systems, institutional reforms, cultural learning behaviors, or learner psychology should therefore be viewed as contextual explanations for observed associations rather than causal inferences. Future studies should integrate learning logs, examination outcomes, user interviews, subgroup analyses, and longitudinal or experimental designs to evaluate how app features influence learning processes and professional competency development.

Conclusions

In summary, this longitudinal app store analysis identified phased evolution and country-specific rating-based satisfaction patterns among medical education apps in China and the United States. The findings suggest that app store reviewers' concerns shifted from basic functionality, clinical practicality, and technical stability toward exam preparation, cost, and personalized learning support. More broadly, this study highlights the need to understand medical education apps not only as digital products but also as sociotechnical educational tools situated within platform ecosystems, institutional contexts, and learner needs. The MEdu Copilot framework should therefore be interpreted as a conceptual synthesis of the empirical findings and contextual policy interpretations, rather than as a validated platform, and future studies should evaluate such designs using learning process and outcome data.

Acknowledgments

The authors thank all contributors who supported data collection, analysis, figure preparation, and manuscript revision.

XSL•FO
RenderX

Funding

This work was supported by the National Natural Science Foundation of China (grant number 72574006), the Noncommunicable Chronic Diseases-National Science and Technology Major Project (grant numbers 2023ZD0509800 and 2023ZD0509804), the Key Project of Higher Education Research of Zhejiang Higher Education Association in 2025 (grant number KT2025001), the Key Project of the 2024 Higher Education Science Research Plan of the Chinese Society of Higher Education (grant number 24BJ0311), and the 2025 Medical Education Research Project of the Medical Education Branch of the Chinese Medical Association and the National Medical Education Development Center (grant number 2025B349).

Data Availability

The deidentified and aggregated datasets supporting the main analyses of this study are available from the corresponding author on reasonable request, where permitted by applicable data-use restrictions. The analysis code and model specifications are publicly available in a GitHub repository [64]. App metadata and user review data were obtained from 10 major app stores through the Qimai Apps Data platform [65] and web scraping technology. Because the raw app-store review data may be subject to platform terms of service and may contain user-generated text, the full raw review dataset is not publicly redistributed. App-store searches were conducted in September 2025, and user review data were collected as of October 2025.

Generative AI Use

DeepSeek-R1 was used to assist with semantic labeling of latent Dirichlet allocation–generated topics by generating candidate topic names and brief descriptions from the top 30 high-probability keywords for each topic. All AI-assisted topic labels and descriptions were critically reviewed, edited, and verified by the authors through consensus-based semantic review and structured expert validation. The authors take full responsibility for the final topic labels, analyses, interpretations, and manuscript content. Generative AI was not used for data collection, review preprocessing, topic modeling, statistical analysis, result interpretation, or generation of scientific conclusions. It was used to assist with language editing and polishing of the manuscript, and all AI-assisted revisions were critically reviewed and approved by the authors.

Authors' Contributions

J Lei and J Liang conceived the study. JM and J Liang independently screened and evaluated the apps according to the predefined eligibility criteria, and YC adjudicated disagreements. ZL and JZ collected developer and user feedback. YH and XL performed the data analysis. YL and JW supervised and validated the qualitative analysis using large language models. TS and WD completed the manuscript translation. QY, LY, and HP were responsible for figure preparation and design. JM drafted the manuscript. J Liang and QY critically revised the manuscript, and PX reviewed it. QY, J Liang, and J Lei are co-corresponding authors of this work. All authors reviewed and approved the final manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Supplementary methods, tables, figures, and results on medical education app store reviews and Kano classification in China and the United States.
[DOCX File , 7069 KB - mededu_v12i1e107560_app1.docx]

References

1. Cernega A, Nicolescu DN, Mele canu Imre M, Ripszky Totan A, Arsene AL, erban RS, et al. Volatility, uncertainty, complexity, and ambiguity (VUCA) in healthcare. Healthcare (Basel) 2024;12(7):773 [FREE Full text] [doi: 10.3390/healthcare12070773] [Medline: 38610195]
2. Stojan J, Haas M, Thammasitboon S, Lander L, Evans S, Pawlik C, et al. Online learning developments in undergraduate medical education in response to the COVID-19 pandemic: a BEME systematic review: BEME Guide No. 69. Med Teach 2022;44(2):109-129. [doi: 10.1080/0142159x.2021.1992373]
3. Alrashed FA, Ahmad T, Almurdi MM, Alderaa AA, Alhammad SA, Serajuddin M, et al. Incorporating technology adoption in medical education: a qualitative study of medical students' perspectives. Adv Med Educ Pract 2024;15:615-625 [FREE Full text] [doi: 10.2147/AMEP.S464555] [Medline: 38975614]
4. Alsharif MA, Bahbah AA, Ghonaim MM, Omran JA, Rihan MM, Zayed MA, et al. Usage and attitude of medical students towards mobile medical applications during and after COVID-19 lockdown: repeated cross-sectional study. BMC Med Educ 2024;24(1):233 [FREE Full text] [doi: 10.1186/s12909-024-05216-5] [Medline: 38438883]

XSL•FO
RenderX

5. Khan HU, Ali Y, Azeem Akbar M, Khan F. A comprehensive survey on exploring and analyzing COVID-19 mobile apps: meta and exploratory analysis. Heliyon 2024;10(15):e35137 [FREE Full text] [doi: 10.1016/j.heliyon.2024.e35137] [Medline: 39170132]
6. Prober CG, Khan S. Medical education reimagined: a call to action. Acad Med 2013;88(10):1407-1410. [doi: 10.1097/ACM.0b013e3182a368bd] [Medline: 23969367]
7. Zhou Y, Yang Y, Liu L, Zeng Z. Effectiveness of mobile learning in medical education: a systematic review. Nan Fang Yi Ke Da Xue Xue Bao 2018;38(11):1395-1400. [doi: 10.12122/j.issn.1673-4254.2018.11.20] [Medline: 30514692]
8. Dornan T, Boshuizen H, King N, Scherpbier A. Experience-based learning: a model linking the processes and outcomes of medical students' workplace learning. Med Educ 2007;41(1):84-91. [doi: 10.1111/j.1365-2929.2006.02652.x] [Medline: 17209896]
9. Yeung AWK, Parvanov ED, Hribersek M, Eibensteiner F, Klager E, Kletecka-Pulker M, et al. Digital teaching in medical education: scientific literature landscape review. JMIR Med Educ 2022;8(1):e32747 [FREE Full text] [doi: 10.2196/32747] [Medline: 35138260]
10. Chandran VP, Balakrishnan A, Rashid M, Pai Kulyadi G, Khan S, Devi ES, et al. Mobile applications in medical education: a systematic review and meta-analysis. PLoS One 2022;17(3):e0265927 [FREE Full text] [doi: 10.1371/journal.pone.0265927] [Medline: 35324994]
11. Sadler J, Wright J, Vincent T, Kurka T, Howlett D. What is the impact of apps in medical education? A study of CAPSULE, a case-based learning app. BMJ Simul Technol Enhanc Learn 2021;7(5):293-296. [doi: 10.1136/bmjstel-2020-000593] [Medline: 35515743]
12. Samuel A, Durning S. From crisis to opportunity: reinventing medical education after COVID-19. Perspect Med Educ 2025;14(1):371-382 [FREE Full text] [doi: 10.5334/pme.1672] [Medline: 40586112]
13. Zhang Z, Ji T. Application and progress of digital techniques in head and neck oncology reconstructive surgery. Holist Integ Oncol 2025;4(1):16. [doi: 10.1007/s44178-024-00137-1]
14. Kyaw BM, Posadzki P, Paddock S, Car J, Campbell J, Tudor Car L. Effectiveness of digital education on communication skills among medical students: systematic review and meta-analysis by the digital health education collaboration. J Med Internet Res 2019;21(8):e12967 [FREE Full text] [doi: 10.2196/12967] [Medline: 31456579]
15. Almulhem JA, Aldekhyyel RN, Binkheder S. Evaluation of mobile apps used among medical students for learning and education: a mixed-method concurrent triangulation approach. Appl Clin Inform 2024;15(4):717-726 [FREE Full text] [doi: 10.1055/s-0044-1788332] [Medline: 39231539]
16. Xu Q, Hou X, Xiao T, Zhao W. Factors affecting medical students' continuance intention to use mobile health applications. J Multidiscip Healthc 2022;15:471-484 [FREE Full text] [doi: 10.2147/JMDH.S327347] [Medline: 35300179]
17. Johnson SG, Potrebny T, Larun L, Ciliska D, Olsen NR. Usability methods and attributes reported in usability studies of mobile apps for health care education: scoping review. JMIR Med Educ 2022;8(2):e38259 [FREE Full text] [doi: 10.2196/38259] [Medline: 35767323]
18. Walsh K. Mobile learning in medical education: review. Ethiop J Health Sci 2015;25(4):363-366 [FREE Full text] [doi: 10.4314/ejhs.v25i4.10] [Medline: 26949301]
19. Barteit S, Guzek D, Jahn A, Bärnighausen T, Jorge MM, Neuhann F. Evaluation of e-learning for medical education in low- and middle-income countries: a systematic review. Comput Educ 2020;145:103726 [FREE Full text] [doi: 10.1016/j.compedu.2019.103726] [Medline: 32565611]
20. Gimpel H, Manner-Romberg T, Schmied F, Winkler TJ. Understanding the evaluation of mHealth app features based on a cross-country Kano analysis. Electron Mark 2021;31(4):765-794 [FREE Full text] [doi: 10.1007/s12525-020-00455-y] [Medline: 35602116]
21. Davis FD. Perceived usefulness, perceived ease of use, and user acceptance of information technology. MIS Quarterly 1989;13(3):319-340. [doi: 10.2307/249008]
22. Venkatesh V, Morris MG, Davis GB, Davis FD. User acceptance of information technology: toward a unified view. MIS Quarterly 2003;27(3):425-478. [doi: 10.2307/30036540]
23. Jan J, Alshare KA, Lane PL. Hofstede's cultural dimensions in technology acceptance models: a meta-analysis. Univ Access Inf Soc 2024;23(2):717-741. [doi: 10.1007/s10209-022-00930-7]
24. Liu X, Feng J, Liu C, Chu R, Lv M, Zhong N, et al. Medical education systems in China: development, status, and evaluation. Acad Med 2023;98(1):43-49 [FREE Full text] [doi: 10.1097/ACM.0000000000004919] [Medline: 35947483]
25. Patel OU, Haynes WK, Burge KG, Yadav IS, Peterson T, Camino A, et al. Results of a national survey of program directors' perspectives on a pass/fail US Medical Licensing Examination Step 1. JAMA Netw Open 2022;5(6):e2219212 [FREE Full text] [doi: 10.1001/jamanetworkopen.2022.19212] [Medline: 35763299]
26. Singh K, Sarkar S, Gaur U, Gupta S, Adams OP, Sa B, et al. Smartphones and educational apps use among medical students of a smart university campus. Front Commun 2021;6:649102. [doi: 10.3389/fcomm.2021.649102]
27. Elhassan SE, Sajid MR, Syed AM, Fathima SA, Khan BS, Tamim H. Assessing familiarity, usage patterns, and attitudes of medical students toward ChatGPT and other chat-based AI apps in medical education: cross-sectional questionnaire study. JMIR Med Educ 2025;11:e63065 [FREE Full text] [doi: 10.2196/63065] [Medline: 39883912]

XSL•FO
RenderX

28. Gladman T, Tylee G, Gallagher S, Mair J, Rennie SC, Grainger R. A tool for rating the value of health education mobile apps to enhance student learning (MARuL): development and usability study. JMIR Mhealth Uhealth 2020;8(7):e18015 [FREE Full text] [doi: 10.2196/18015] [Medline: 32735228]
29. Stoyanov SR, Hides L, Kavanagh DJ, Zelenko O, Tjondronegoro D, Mani M. Mobile app rating scale: a new tool for assessing the quality of health mobile apps. JMIR Mhealth Uhealth 2015;3(1):e27 [FREE Full text] [doi: 10.2196/mhealth.3422] [Medline: 25760773]
30. Owrangi S, Zarifsanaiey N, Rezaei M. Mobile apps to enhance student learning in medical education: a systematic search in app stores and evaluation using the mobile app rating scale. Comput Biol Med 2025;196(Pt B):110740. [doi: 10.1016/j.compbiomed.2025.110740] [Medline: 40669285]
31. Gladman T, Tylee G, Gallagher S, Mair J, Grainger R. Measuring the quality of clinical skills mobile apps for student learning: systematic search, analysis, and comparison of two measurement scales. JMIR Mhealth Uhealth 2021;9(4):e25377 [FREE Full text] [doi: 10.2196/25377] [Medline: 33890859]
32. Briz-Ponce L, Juanes-Méndez JA, García-Peñalvo FJ. Description of a session to assess the efectiveness of apps in medical education. ResearchGate. URL: https://www.researchgate.net/publication/280114331_Description_of_a_session_to_assess_the_efectiveness_of_apps_in_medical_education [accessed 2026-07-16]
33. S S, Vedam VAV, Ghanta MK. An assessment of utility of educational apps in medical education among undergraduate medical students - a cross-sectional survey. Natl J Physiol Pharm Pharmacol 2024;14(1):182-187. [doi: 10.5455/njppp.2023.13.06317202322062023]
34. Darras KE, van Merriënboer JJG, Toom M, Roberson ND, de Bruin ABH, Nicolaou S, et al. Developing the evidence base for M-learning in undergraduate radiology education: identifying learner preferences for mobile apps. Can Assoc Radiol J 2019;70(3):320-326 [FREE Full text] [doi: 10.1016/j.carj.2019.03.007] [Medline: 31300315]
35. Zhao S, Cao Y, Cao H, Liu K, Lv X, Zhang J, et al. Chinese version of the mHealth app usability questionnaire: cross-cultural adaptation and validation. Front Psychol 2022;13:813309 [FREE Full text] [doi: 10.3389/fpsyg.2022.813309] [Medline: 35185732]
36. Jiao X, Jiang L, Zhao M, Ma J, Li Y, Shen T, et al. User satisfaction with pregnancy management apps in Mainland China: user-generated content analysis and text mining study. J Med Internet Res 2025;27:e78828 [FREE Full text] [doi: 10.2196/78828] [Medline: 40839869]
37. Liew MS, Zhang J, See J, Ong YL. Usability challenges for health and wellness mobile apps: mixed-methods study among mHealth experts and consumers. JMIR Mhealth Uhealth 2019;7(1):e12160 [FREE Full text] [doi: 10.2196/12160] [Medline: 30698528]
38. Masters K, Ellaway RH, Topps D, Archibald D, Hogue RJ. Mobile technologies in medical education: AMEE Guide No. 105. Med Teach 2016;38(6):537-549. [doi: 10.3109/0142159x.2016.1141190]
39. Yang F, Zheng J, Li W, Liu S, Zhou Y, Lu S. Digitalizing caregiving for what? Role dynamics of caregivers in nursing homes implementing smart technologies. J Technol Hum Serv 2025:1-28. [doi: 10.1080/15228835.2025.2585409]
40. Toward AI-enhanced medical education: cross-national evolution and user satisfaction in mobile learning apps. OSF Registries. URL: https://osf.io/hy2cj/ [accessed 2026-07-15]
41. Su Z, Figueiredo MC, Jo J, Zheng K, Chen Y. Analyzing description, user understanding and expectations of AI in mobile health applications. AMIA Annu Symp Proc 2020;2020:1170-1179 [FREE Full text] [Medline: 33936493]
42. Martens D, Maalej W. Towards understanding and detecting fake reviews in app stores. Empir Software Eng 2019;24(6):3316-3355. [doi: 10.1007/s10664-019-09706-9]
43. Baidu AI Cloud. Sentiment analysis. Baidu AI Cloud Documentation. URL: https://cloud.baidu.com/doc/NLP/s/zk6z52hds [accessed 2025-10-17]
44. Valdivia A, Luzon MV, Herrera F. Sentiment analysis in TripAdvisor. IEEE Intell Syst 2017;32(4):72-77. [doi: 10.1109/mis.2017.3121555]
45. Valdivia A, Hrabova E, Chaturvedi I, Luzón MV, Troiano L, Cambria E, et al. Inconsistencies on TripAdvisor reviews: a unified index between users and sentiment analysis methods. Neurocomputing 2019;353:3-16. [doi: 10.1016/j.neucom.2018.09.096]
46. He Y, Zhu W, Wang T, Chen H, Xin J, Liu Y, et al. Mining user reviews from hypertension management mobile health apps to explore factors influencing user satisfaction and their asymmetry: comparative study. JMIR Mhealth Uhealth 2024;12:e55199 [FREE Full text] [doi: 10.2196/55199] [Medline: 38547475]
47. Patel VR, Liu M, Jena AB. Public interest in an AI-enabled clinical decision support tool. JAMA Netw Open 2025;8(11):e2544672 [FREE Full text] [doi: 10.1001/jamanetworkopen.2025.44672] [Medline: 41264272]
48. Sun J. Jieba: Chinese words segmentation utilities. GitHub. URL: https://github.com/fxsjy/jieba [accessed 2026-07-15]
49. Blei DM, Ng AY, Jordan MI. Latent Dirichlet allocation. J Mach Learn Res 2003;3:993-1022 [FREE Full text]
50. Arora N, Chakraborty I, Nishimura Y. AI–human hybrids for marketing research: leveraging large language models (LLMs) as collaborators. J Mark 2025;89(2):43-70. [doi: 10.1177/00222429241276529]
51. Liu S, Zheng J, Yang F, Li W, Zhou Y, Lu S, et al. Internet altruistic behaviors and affective well-being among family caregivers: a natural language processing approach. J Gerontol B Psychol Sci Soc Sci 2026;81(1):gbaf230. [doi: 10.1093/geronb/gbaf230] [Medline: 41224706]

XSL•FO
RenderX

52. Wang T, Fan L, Zheng X, Wang W, Liang J, An K, et al. The impact of gamification-induced users' feelings on the continued use of mHealth apps: a structural equation model with the self-determination theory approach. J Med Internet Res 2021;23(8):e24546 [FREE Full text] [doi: 10.2196/24546] [Medline: 34387550]
53. Wang T, Zheng X, Liang J, An K, He Y, Nuo M, et al. Use of machine learning to mine user-generated content from mobile health apps for weight loss to assess factors correlated with user satisfaction. JAMA Netw Open 2022;5(5):e2215014 [FREE Full text] [doi: 10.1001/jamanetworkopen.2022.15014] [Medline: 35639374]
54. Nuo M, Zheng S, Wen Q, Fang H, Wang T, Liang J, et al. Mining the influencing factors and their asymmetrical effects of mHealth sleep app user satisfaction from real-world user-generated reviews: content analysis and topic modeling. J Med Internet Res 2023;25:e42856 [FREE Full text] [doi: 10.2196/42856] [Medline: 36719730]
55. Wang T, Wang W, Feng J, Fan X, Guo J, Lei J. A novel user-generated content-driven and Kano model focused framework to explore the impact mechanism of continuance intention to use mobile APPs. Comput Hum Behav 2024;157:108252. [doi: 10.1016/j.chb.2024.108252]
56. Barrett B. The iOS 11 32-bit appocalypse isn't actually that bad. Wired. URL: https://www.wired.com/story/the-ios-11-32-bit-appocalypse-isnt-actually-that-bad/ [accessed 2026-07-15]
57. 64-bit and iOS 8 requirements for app updates. Apple Developer News. URL: https://developer.apple.com/news/?id=12172014b [accessed 2026-07-15]
58. Lee JWY, Tan JY, Bello F. Technology acceptance model in medical education: systematic review. JMIR Med Educ 2025;11:e67873 [FREE Full text] [doi: 10.2196/67873] [Medline: 40668663]
59. The general office of the state council issued the opinions on deepening medical-education collaboration and further promoting the reform and development of medical education. National Health Commission of the People's Republic of China. URL: https://www.nhc.gov.cn/wjw/gwywj/201707/8620cd5834ec47bd966f2d0acb67d4d3.shtml [accessed 2026-07-15]
60. Izquierdo-Condoy JS, Arias-Intriago M, Nati-Castillo HA, Gollini-Mihalopoulos R, Cardozo-Espínola CD, Loaiza-Guevara V, et al. Exploring smartphone use and its applicability in academic training of medical students in latin America: a multicenter cross-sectional study. BMC Med Educ 2024;24(1):1401 [FREE Full text] [doi: 10.1186/s12909-024-06334-w] [Medline: 39616324]
61. Howard JL, Bradshaw EL, Ryan RM. Future directions for self-determination theory: introduction to the special issue. Motiv Emot 2026;50(1):1-12. [doi: 10.1007/s11031-025-10160-y]
62. Shen X, Wang Y, Zhao D, Li M. Medical licensing examination-based reform of medical microbiology teaching. Med Educ 2023;57(11):1122. [doi: 10.1111/medu.15206] [Medline: 37705373]
63. Liu L, Chachad N, Tadjalli A, Rajput V. Decoupling the United States Medical Licensing Examinations (USMLEs) from the medical curriculum to promote student well-being and professional identity development. Cureus 2025;17(5):e83335. [doi: 10.7759/cureus.83335] [Medline: 40458346]
64. The source code for the medical education app analysis. GitHub. URL: https://github.com/junhao-ma/medical-education-app-analysis [accessed 2026-09-28]
65. Qimai apps data platform. Qimai. URL: https://www.qimai.cn [accessed 2025-10-19]

Abbreviations

AAPC: average annual percentage change
APC: annual percentage change
BIC: Bayesian information criterion
LDA: latent Dirichlet allocation
MARS: Mobile App Rating Scale
MARuL: Mobile App Rubric for Learning
ND: negative rating deviation
NLTK: Natural Language Toolkit
PD: positive rating deviation
PBL: problem-based learning
RQ: research question
TAM: technology acceptance model
USMLE: United States Medical Licensing Examination
VIF: variance inflation factor

XSL•FO
RenderX

Edited by T de Azevedo Cardoso; submitted 20.Jul.2026; peer-reviewed by F Yang, A Wahbeh, F Cui; comments to author 19.Aug.2026; revised version received 29.Sep.2026; accepted 05.Oct.2026; published 03.Nov.2026.

Please cite as:
Ma J, Shen T, Xiang P, Li Y, Chen Y, Lu Z, Zeng J, Dai W, Pan H, Yang L, He Y, Li X, Wu J, Yu Q, Liang J, Lei J
Mobile Medical Education Apps as Digital Health Information Resources in China and the United States: Longitudinal App Store Metadata Analysis and Topic Modeling of User Review Narratives
JMIR Med Educ 2026;12:e107560
URL: https://mededu.jmir.org/2026/1/e107560
doi:10.2196/107560
PMID:

©Junhao Ma, Tian Shen, Peng Xiang, Yingjun Li, Yuanshuai Chen, Zhengyang Lu, Jingkai Zeng, Wei Dai, Hanwen Pan, Liping Yang, Yunfan He, Xingxi Li, Joyce Wu, Qi Yu, Jun Liang, Jianbo Lei. Originally published in JMIR Medical Education (https://mededu.jmir.org), 03.Nov.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Education, is properly cited. The complete bibliographic information, a link to the original publication on https://mededu.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Learning Participation and Performance Among Social and Academic Learners in a Nursing Massive Open Online Course: Retrospective Cohort Study

Jianqing Zeng[1*], MS; Kun Li[1*], PhD; Wenxuan Wang[1], MS; Lingjun Xiao[2], MS; Yingmin Wang[2], MS; Xinghui He[2], BS; Yonghong Qing[2], MS; Jingjing Xue[2], MS; Lirong Zhong[2], BS; Tiebin Yan[2], PhD; Xiuyuan Zheng[2], PhD

[1]School of Nursing, Sun Yat-sen University, Guangzhou, China

[2]Sun Yat-sen Memorial Hospital, Sun Yat-sen University, Guangzhou, China

[*]these authors contributed equally

Corresponding Author:
Xiuyuan Zheng, PhD
Sun Yat-sen Memorial Hospital, Sun Yat-sen University
107 Yanjiang West Road
Guangzhou, Guangdong, 510120
China
Phone: 86 02081332880
Email: zhengxy27@mail.sysu.edu.cn

Abstract

Background: Massive open online courses (MOOCs) are increasingly used in nursing education, attracting learners with diverse participation modes. Social learners (self-directed) enroll voluntarily based on personal interests, while academic learners (institution-directed) are mandated by nursing schools, with performance linked to grades. Existing literature shows that academic learners outperform social learners in MOOCs, but evidence is scarce on whether this performance gap stems from distinct participation patterns shaped by institutional requirements.

Objective: We sought to compare participation typologies between social and academic learners using latent class analysis (LCA) and assess their association with learning outcomes.

Methods: A retrospective analysis was conducted of 8222 students enrolled in a rehabilitation nursing MOOC from 2021 to 2022. Learners were classified by topic test participation frequency via LCA, and their distributions and scores were statistically compared.

Results: The cohort comprised 7683 (93.4%) academic and 539 (6.6%) social learners. LCA identified 3 participation types: participation-declining (93/8222, 1.1%), high-participation (6959/8222, 84.6%), and lowest-participation (1170/8222, 14.2%). The distributions of participation types differed significantly between groups (χ^2=2745.1; *df*=2; *P*<.001). Most social learners belonged to the lowest-participation group (482/539, 89.4%), whereas most academic learners belonged to the high-participation group (6923/7683, 90.1%). Among high-participation learners, academic learners significantly outperformed social learners (all *P*<.05; *r*=0.06-0.10). No significant between-group differences were observed for the other participation types (all *P*>.05).

Conclusions: Distinct participation patterns were associated with the performance advantage of academic learners. Our findings highlight the importance of sustained participation, suggesting that MOOC instructors should provide enhanced support for voluntary learners.

(JMIR Med Educ 2026;12:e90827) doi:10.2196/90827

KEYWORDS

massive open online courses; MOOCs; nursing education; learning participation; academic performance; latent class analysis

Introduction

Massive open online courses (MOOCs) have gained widespread attention in nursing education due to their openness and accessibility [1]. MOOCs provide unprecedented opportunities for lifelong learning and professional development and, over recent years, have been widely introduced into nursing education, becoming a key approach for promoting reform in

XSL•FO
RenderX

nursing teaching and enhancing the professional competence of nursing staff [2,3]. However, while providing convenience, MOOCs also face a common challenge relating to a high enrollment rate but a low completion rate, with significant variations in the participation of learners and their learning performance [4]. Studies have shown that the initial motivation of learners is a key factor influencing their online learning behaviors and final outcomes [5]. On the basis of motivation, learners are categorized as "social learners" (voluntary) and "academic learners" (institution-mandated), with their performance often linked to credits or final grades and driven by significant external motivations [6].

Previous studies indicate that academic learners often exhibit higher participation than social learners due to external motivation [7], which confirms the important role of external motivation in maintaining online learning behaviors. However, these studies have overlooked the heterogeneity and complexity within groups of different learners. For example, social or mixed MOOC learners were classified as committed, early dropout, and other learners [8]; Wang et al [9] used latent class analysis (LCA) to categorize MOOC learners into committed, negative, midterm dropout, and early dropout learner types. Not all academic learners are actively engaged, nor are all social learners passively inactive: both groups include learners with high participation, minimal effort, and midterm withdrawal. This heterogeneity within each group suggests that a simple division of social and academic learners cannot fully explain the performance differences in MOOCs [8].

Therefore, simply applying "social" or "academic" distinctions to explain differences in learning performance may obscure the real factors associated with this phenomenon. We need to consider whether the improved overall performance of academic learners due to the "academic" identity can improve their efficiency of learning or whether the proportion of "high-participation learners" within the academic learner group is much higher than that in the social learner group. If learners can be objectively classified based on their actual online learning behavioral characteristics, and then the distribution differences of learners with different backgrounds in these behavioral categories can be compared, it may be possible to more accurately identify the key factors associated with learning performance.

Consequently, in this study, we retrospectively analyzed data from a rehabilitation nursing MOOC and applied LCA to identify different participation types based on the participation behaviors of learners in topic tests. We analyzed and compared differences in the proportion of these participation types between social and academic learners and investigated how these differences in distribution might be associated with final learning performance. In this study, we aimed to comprehensively analyze the heterogeneity within learner groups to provide empirical evidence for nursing educators to design more targeted MOOC teaching strategies and implement refined teaching management, thereby effectively improving the quality and efficacy of online nursing education.

Methods

We conducted a retrospective study using data from a rehabilitation nursing MOOC.

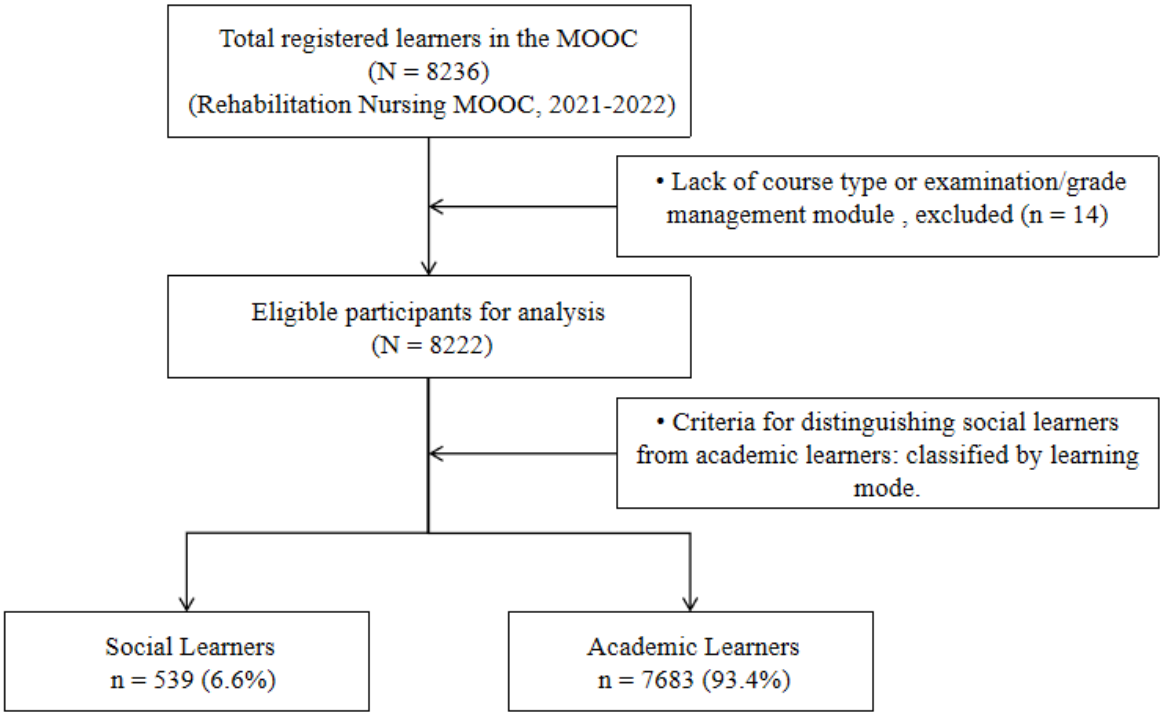
Participants

This MOOC featured a dual structure. First, the course was freely open to the public without prerequisites. Second, participants who registered and engaged voluntarily were defined as social learners. Learners assigned by university teachers, who received blended online and offline instruction, were defined as academic learners with relatively smaller class sizes.

Our dataset included all social and academic learners registered across 3 semesters (2021-2022) who had complete data on course type and examination and grade management modules. After excluding learners with missing information, the final sample comprised 8222 participants (Figure 1). This sample size far exceeds the minimum requirement of 300 for valid category analysis [10], ensuring sufficient statistical power. Notably, the online learning platform did not collect demographic characteristics (such as age, gender, and prior educational background) during learner registration, precluding control for these potential confounding variables.

XSL•FO
RenderX

Figure 1. Flowchart of study participant screening. MOOC: massive open online courses.



The Rehabilitation Nursing MOOC

The rehabilitation nursing MOOC was hosted on the Wisdom Tree online education platform [11]. The primary objective of this course was to enhance the comprehension of rehabilitation nursing theory through perceptual learning. The course welcomed all who registered via the Wisdom Tree platform without imposing any prerequisites. The course featured patient evaluation, rehabilitation nursing evaluation, rehabilitation treatment technology, rehabilitation nursing technology, and rehabilitation nursing for common diseases. Academic learners were provided with offline direct feedback, offline practicum activities, and online assessments. Offline direct feedback included face-to-face consultation between teachers and students and detailed feedback on assignments, provided every 2 weeks. Offline practical activities, arranged every 4 weeks in line with the teaching schedule, involved practical training related to the specialty. A diverse range of online assessments was used to evaluate the knowledge mastery of each learner. Diverse instructional resources, such as teaching videos, topic-specific assessments, and downloadable teaching materials, were used to deliver the content of each topic. Topic assessments comprised 5 to 17 questions that aimed to evaluate the understanding of each learner in that specific area.

Limitations in platform data collection restricted our analysis to topic test participation as the sole measure of learning participation; broader engagement metrics (eg, logins, video viewing duration, time-on-task, forum activity, and teaching material downloads) were not available for extraction and analysis.

Outcome Indicators

Learners who completed and submitted topic tests (including partial attempts and 0 scores) were treated as participants for each topic. The participation rate in each topic test was computed by dividing the number who submitted a test (achieved a score, including 0) by the total number enrolled. The learner types were then categorized via LCA based on data from 6 administered topic tests.

Learning performance was the secondary outcome indicator, quantified by individual topic test scores. Because the 6 tests had different numbers of questions, raw scores were standardized to a 0 to 100 scale by dividing the actual score by the maximum possible score and multiplying by 100.

Data Analysis

All statistical analyses were performed using SPSS software (version 25.0; IBM Corp) and Mplus software (version 8.3; Muthén & Muthén). Descriptive statistics were computed to describe the number of social and academic learners considering each topic test, including participation rates, frequencies, and percentages.

LCA quantifies the heterogeneity of a given population by classifying individuals based on their different response patterns across a series of observed indicators. For LCA in this study, we used binary indicator coding (1=submitted the topic test and 0=did not submit the topic test). Model estimation was conducted via maximum likelihood with full information maximum likelihood to handle missing data, adhering to the local independence assumption of LCA. We applied LCA to categorize learners based on their participation in each topic test. The distribution of social learners and academic learners across different learning participation types was computed, and

XSL•FO
RenderX

results were expressed as frequencies and percentages. Differences were then tested for significance using the chi-square test. Descriptive statistics were also computed for the scores acquired by social and academic learners in each topic test. Scores are represented as means and SDs. The significance of the disparities in performance between social learners and academic learners was evaluated using 2-tailed *t* tests (for normally distributed data) or Mann-Whitney *U* tests (for data that were not normally distributed). Because of the limited reliability and stability of statistical analyses, comparisons between subgroups with small sample sizes (n<5) and the other 2 groups were considered exploratory.

To ascertain the optimum number of latent classes, a single-class model was first prepared as a baseline for comparison. Subsequently, the number of latent classes was incrementally increased, and the fit indices of the resulting models were compared. The best model, with the optimum number of classes, was determined based on both statistical and practical significance [12]. We considered several indices of model fit, including the Akaike information criterion, Bayesian information criterion, and the adjusted Bayesian information criterion. For each of these indices, a smaller value indicates a better model fit. Lo-Mendell-Rubin and bootstrap likelihood ratio tests were also performed. A significant *P* value indicated that the k-class model outperformed the (k−1)-class model [10]. Entropy was used as an index of classification accuracy for a given model and ranged from 0 to 1. A value of ≥0.80 was considered to signify high classification accuracy.

Ethical Considerations

This retrospective study was conducted using fully anonymized learner data obtained from an online course platform. The study protocol was reviewed and approved by the Institutional Review Board of Sun Yat-sen University. Because of the use of deidentified retrospective data involving no more than minimal risk to participants, the study was granted exemption from full ethical review, and no specific approval number was assigned. Informed consent was waived in accordance with institutional guidelines for low-risk secondary research using anonymized data. No participants were recruited, and no compensation or incentives were provided for participation, as only anonymized, nonidentifiable secondary data were analyzed. The study adhered to the ethical principles of the Declaration of Helsinki and protected the privacy of research participants. Data were obtained with official authorization from the Wisdom Tree platform, and all learner data were anonymized and renumbered before analysis.

Results

Participation Rates

Academic learners maintained high participation rates across all topic tests (89.9%-91.2%), while social learners showed consistently low participation (6.7%-12.6%) and a progressive decline in engagement (Table 1).

Table 1. Participation rates in the different topic tests.

Topic test	Social learners (n=539), n (%)	Academic learners (n=7683), n (%)
Test 1	68 (12.6)	7005 (91.2)
Test 2	53 (9.8)	6989 (91)
Test 3	47 (8.7)	6951 (90.5)
Test 4	38 (7.1)	6937 (90.3)
Test 5	42 (7.8)	6954 (90.5)
Test 6	36 (6.7)	6910 (89.9)

Classification

LCA identified 3 distinct participation patterns based on 6 topic tests, and the 3-class model was selected as the optimal solution (Table 2). Although Lo-Mendell-Rubin and bootstrap likelihood ratio tests supported a 4-class model, the latter generated an overly specific category with only 0.1% (9/8222) of learners that had no practical interpretive value. In large samples, these tests can detect substantively trivial divisions, leading to overextraction [13]. The emergent fourth class contributed negligible improvement in model fit (ΔAkaike information criterion <25) and, given its small size, precluded meaningful between-group comparisons, while the 3-class model had high entropy (>0.8) and meaningful, interpretable participation patterns. Therefore, consistent with recommended practices [14], the 3-class solution was prioritized on grounds of parsimony, interpretability, and statistical stability. The final classification included 3 types: participation-declining (93/8222, 1.1%), high-participation (6959/8222, 84.6%), and lowest-participation (1170/8222, 14.2%).

XSL•FO
RenderX

Table 2. Comparison of latent class model fit indices (N=8222).

Model	AIC[a]	BIC[b]	aBIC[c]	LMR[d], *P* value	BLRT[e], *P* value	Entropy	Latent class distribution, n
1-class	41,367.193	41,409.280	41,390.213	—[f]	—	1.000	8222
2-class	9925.287	10,016.477	9975.165	<.001	<.001	1.000	6974/1248[g]
3-class	9252.108	9392.399	9328.843	<.001	<.001	0.998	93/6959/1170[g]
4-class	9227.600	9416.993	9331.192	<.01	<.001	0.997	9/6958/1172/83[g]

[a]AIC: Akaike information criterion.

[b]BIC: Bayesian information criterion.

[c]aBIC: adjusted Bayesian information criterion.

[d]LMR: Lo-Mendell-Rubin.

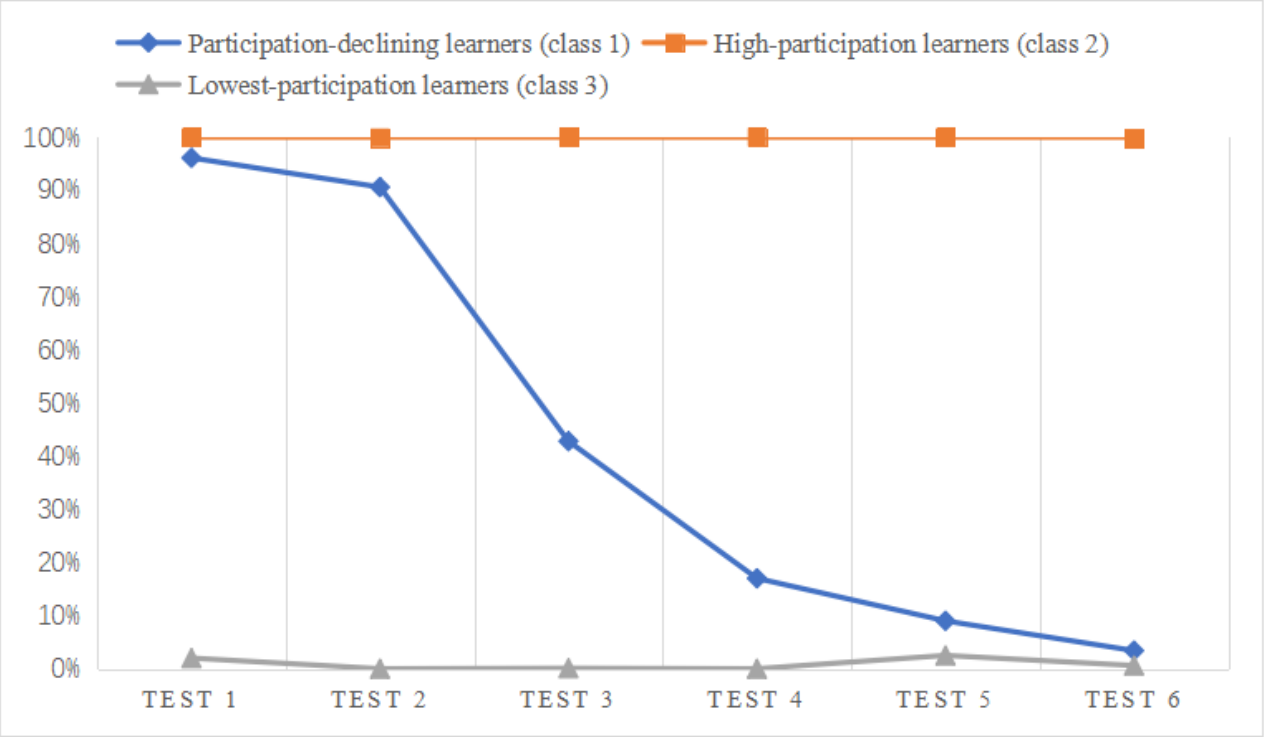
[e]BLRT: bootstrap likelihood ratio test.

[f]Not applicable.

[g]These separated numbers represent the sample sizes of the latent classes respectively. The slash symbols are only used to separate different group values and do not indicate fractions.

The participation probabilities trajectory of the 3 learner categories is shown in Figure 2. Participation-declining learners (class 1; 93/8222, 1.1%) exhibited gradually decreasing participation, with a notable decline after topic test 3. In contrast, high-participation learners (class 2; 6959/8222, 84.6%) maintained near-perfect participation rates across all tests, while lowest-participation learners (class 3; 1170/8222, 14.2%) showed minimal engagement throughout.

Figure 2. Participation probability of different types of total learners across 6 topic tests.



The distribution of learner types differed significantly between groups (χ^2=2745.1; *P*<.001; Table 3). Most social learners belonged to the lowest-participation group (482/539, 89.4%), with few belonging to the high-participation (36/539, 6.7%) or participation-declining (21/539, 3.9%) groups. In contrast, most academic learners belonged to the high-participation group (6923/7683, 90.1%), with smaller proportions belonging to the lowest-participation (688/7683, 9%) and participation-declining (72/7683, 0.9%) groups.

XSL•FO
RenderX

Table 3. Distribution of learner types by group.

Group	Social learners (n=539), n (%)	Academic learners(n=7683), n (%)	Chi-square(*df*)	*P* value
Learner type			2745.1 (2)	<.001
Participation-declining	21 (3.9)	72 (0.9)		
High-participation	36 (6.7)	6923 (90.1)		
Lowest-participation	482 (89.4)	688 (9)		

Learning Performance

Academic learners significantly outperformed social learners among high-participation learners (all $P<.05$; $r=0.06$-0.10), with no significant performance differences detected between the 2 groups in the other 2 participation types (Table 4). Comparisons of subgroups with small sample sizes (n<5) were considered exploratory only, with no statistical inference value because of limited reliability and stability.

Table 4. Comparison of test scores between social and academic learners by learner type.

Topic test and learner type	Social learners (n=539)			Academic learners (n=7683)			Mann-Whitney *U* test		
	Mean (SD)	95% CI	n (%)	Mean (SD)	95% CI	n (%)	*Z*	*P* value	Effect size (*r*)
Test 1									
Participation-declining	89.00 (3.40)	87.21-90.79	20 (3.7)	78.84 (2.97)	78.03-79.65	69 (0.9)	−1.541	.12	−0.16
High-participation	74.29 (6.74)	71.98-76.60	35 (6.5)	95.56 (0.16)	95.52-95.60	6922 (90.1)	5.125	<.001	0.06
Lowest-participation	86.15 (4.17)	80.23-92.07	13 (2.4)	78.57 (6.45)	69.81-87.33	14 (0.2)	−0.700	.48	−0.14
Test 2									
Participation-declining	70.00 (7.09)	66.41-73.59	18 (3.3)	75.71 (3.04)	74.94-76.48	70 (0.9)	0.721	.47	0.08
High-participation	69.71 (6.58)	67.49-71.93	35 (6.5)	94.63 (0.19)	94.59-94.67	6919 (90.1)	6.839	<.001	0.08
Lowest-participation	__[a]	0	0 (0)	—	0	0 (0)	—	—	—
Test 3									
Participation-declining	82.42 (3.95)	79.81-85.03	11 (2)	66.00 (5.37)	64.12-67.88	30 (0.4)	−1.349	.18	−0.21
High-participation	71.30 (6.36)	69.18-73.42	36 (6.7)	94.40 (0.18)	94.36-94.44	6920 (90.1)	6.360	<.001	0.08
Lowest-participation	—	0	0 (0)	—	0	1 (0)	—	—	—
Test 4									
Participation-declining	41.67 (41.67)	8.01-75.33	2 (0.4)	75.00 (7.41)	70.98-79.02	14 (0.2)	1.148	.25	0.29
High-participation	74.07 (6.63)	71.95-76.19	36 (6.7)	95.38 (0.18)	95.34-95.42	6923 (90.1)	5.192	<.001	0.06
Lowest-participation	—	0	0 (0)	—	—	0 (0)	—	—	—
Test 5									
Participation-declining	66.67 (6.67)	50.11-83.23	3 (0.6)	52.00 (14.97)	38.04-65.96	5 (0.1)	−0.800	.42	−0.28
High-participation	58.33 (6.34)	56.29-60.37	36 (6.7)	92.27 (0.22)	92.23-92.31	6922 (90.1)	8.561	<.001	0.10
Lowest-participation	53.33 (6.67)	40.11-66.55	3 (0.6)	50.37 (5.05)	48.19-52.55	27 (0.4)	0.000	>.99	0.00
Test 6									
Participation-declining	—	0	0 (0)	47.06 (8.99)	31.08-63.04	3 (0)	—	—	—
High-participation	57.35 (7.52)	54.51-60.19	32 (5.9)	94.57 (0.18)	94.53-94.61	6904 (89.9)	7.412	<.001	0.09
Lowest-participation	51.47 (17.04)	26.11-76.83	4 (0.7)	33.33 (13.73)	12.11-54.55	3 (0)	−0.535	.59	−0.20

[a]Insufficient data for analysis (n<5).

XSL•FO
RenderX

Discussion

Principal Findings

As an observational study, this research explored factors associated with performance disparities between social and academic learners in a rehabilitation nursing MOOC from the perspective of participation type distribution; causal relationships between the investigated variables cannot be established, and only associative correlations were identified. There were three key findings: (1) divergent participation type distributions between social and academic learners; (2) superior performance of academic learners in the high-participation subgroup; and (3) no significant performance differences in non–high-participation subgroups.

Most notably, the 2 groups showed starkly divergent participation type distributions—the core factor associated with their overall performance gap: more than 90% of academic learners belonged to the high-participation group, whereas nearly 90% of social learners belonged to the lowest-participation group. This underscores the pivotal impact of learner background and learning environment on participation patterns [15]. As nursing undergraduates, academic learners sustained high participation through teacher supervision, peer learning, and credit pressure—a robust external support system for continuous engagement, consistent with Uijl et al [16], who found interpersonal connections and structured learning arrangements to enhance learner interaction and participation. In contrast, social learners lacked such external constraints and institutional support despite greater autonomy, leaving them prone to low participation or midcourse dropout [17]. These patterns align with prior LCA of nursing MOOCs [9]. Wang et al [9] also identified committed learners as the highest-performing group, with no significant performance differences across noncommitted types. While Wang et al [9] distinguished 4 classes among course completers, this analysis of all registrants yielded 3, a discrepancy likely attributable to sample composition differences [18]. This work extends prior research by demonstrating that academic learners' performance advantage stems from their overrepresentation in the committed learner group.

In addition, academic learners' high-participation subgroup outperformed social counterparts across all 6 topic tests (all $P<.001$) with trivial to small effect sizes ($r=0.06$-0.10) [19]. This large-sample research characteristic arose because the sizable academic high-participation cohort (n≥6900) heightened the Mann-Whitney *U* test sensitivity, detecting minor mean differences as significant [20]. In contrast, participation-declining and lowest-participation subgroups (mostly n<50) showed no statistical score differences and similarly trivial to small effect sizes, with their disparities bearing neither statistical nor practical meaning. This confirms the high-participation group as the core carrier of performance differences between cohorts, validating our design centered on participation type distribution.

The far higher proportion of high-participation learners among academic groups, and their stable, sustained engagement, drives the superior overall performance of academic learners: such learners exhibit strong engagement and self-regulation, enabling systematic task completion, active course participation, and in-depth knowledge acquisition and transfer. Conversely, the low proportion of high-participation learners among social learners underpins their poorer performance. While prior research identified high-participation learners as the highest-performing subgroup [8,9], our findings further reveal that background-based disparities in their proportion decisively shape learning outcomes, highlighting participation composition as central to explaining performance differences. Consistent statistical significance in the high-participation group's gap confirms academic learners' stable advantage, a strength derived from institutional external support rather than inherent learning ability gaps [21].

Our analysis also showed that both academic and social learner groups featured a certain proportion of participation-declining and lowest-participation learners, although there were no significant differences between these learners in academic performance. This finding is consistent with the early classifications of MOOC learner types. For example, Reich [22] and Gilligan et al [23] identified low-participation groups, including bystander learners and early dropout learners. Subsequent studies have confirmed that these groups are widespread and achieve poor learning outcomes [22-24]. This may be due to the general lack of internal learning motivation and continuous engagement among such learners. This aligns with self-determination theory, which posits that when internal motivation is insufficient, the effect of external incentives is also greatly reduced, making it difficult to drive effective learning behaviors [5]. These data also indicate that once learners lack continuous engagement, it is difficult for their academic performance to reach the desired level regardless of their background. Therefore, future development and application of MOOCs should involve incentive mechanisms, emotional support, and learning analytics to increase student participation and reduce the proportion of participation-declining and lowest-participation learners.

The predominance of committed learners among academic students (6923/7683, 90.1% vs 36/539, 6.7% in social learners) suggests that participation type may serve as an intervening variable in the pathway between learner background and performance—a finding consistent with recent MOOC research demonstrating that engagement patterns link learner characteristics to outcomes [25]. However, this observational finding is hypothesis-generating only. Academic learners' superior performance may equally reflect offline institutional support absent for social learners, and unmeasured mediator-outcome confounding cannot be ruled out [26]. Future causal mediation studies with comprehensive covariate measurement are needed to test whether participation patterns represent a genuine pathway.

Limitations

This study has several limitations. First, the lack of demographic data (eg, age and prior educational background) from the online platform limited our ability to control for potential confounding variables in the analysis. Second, some subgroups had extremely small sample sizes (n<5), which reduced the statistical power

XSL•FO
RenderX

of relevant comparisons and limited their generalizability. These results should therefore be considered exploratory and require further validation in future studies with larger samples.

Conclusions

This study identified 3 distinct participation types (participation-declining, high-participation, and lowest-participation) among rehabilitation nursing MOOC learners. Academic learners predominantly belonged to the high-participation group, while social learners predominantly belonged to the lowest-participation group. The performance advantage of academic learners was primarily associated with the larger proportion of high-participation learners in their group, and academic learners significantly outperformed social learners within this subgroup. No significant performance differences were found between the 2 groups in the participation-declining and lowest-participation types, confirming that sustained participation is key to improving MOOC learning outcomes. These findings provide empirical evidence for nursing MOOC teaching design, suggesting that instructors should prioritize sustaining learner participation and implement tailored support strategies for voluntary social learners to enhance their engagement and persistence, thereby narrowing the performance gap between the 2 groups.

Acknowledgments

The authors are grateful to the Wisdom Tree online education platform for providing anonymized course data and technical support for this retrospective study. The authors also thank the research assistants at Sun Yat-sen Memorial Hospital for their assistance with preliminary data organization.

Data Availability

The data that support the findings of this study are available from the Wisdom Tree online education platform [11], although restrictions apply to the availability of these data, which were used under license for this study. Aggregated data (eg, participation rates by learner group, mean test scores by participation class, and latent class analysis model fit indices) are available from the corresponding author upon reasonable request and with permission from the Wisdom Tree online education platform. Raw learner log data cannot be shared due to platform data privacy policies.

Funding

The research was supported by the National Natural Science Foundation of China (grants 71974213 and 81902290), the Undergraduate Teaching Quality and Teaching Reform Project of Guangdong Province (grant [2023] number 4), and the Undergraduate Teaching Quality and Teaching Reform Project of Sun Yat-sen University (2018, 2019, and 2021; grant 81000-12220011).

Authors' Contributions

Conceptualization: KL, TY, XZ
Data curation: WW, LX, YW, XH, YQ, JX, LZ, TY, XZ
Formal analysis: JZ
Funding acquisition: KL, XZ
Investigation: JZ
Methodology: KL
Project administration: KL, TY, XZ
Resources: LX, YQ, JX, TY, XZ
Software: JZ, WW
Supervision: TY, XZ
Validation: JZ
Writing—original draft: JZ
Writing—review and editing: KL, WW, YW, XH, LZ, LX, YQ, JX, XZ

Conflicts of Interest

None declared.

References

1. Mahajan R, Gupta P, Singh T. Massive open online courses: concept and implications. Indian Pediatr 2019 Jun 15;56(6):489-495 [FREE Full text] [Medline: 31278230]
2. Longhini J, De Colle B, Rossettini G, Palese A. What knowledge is available on massive open online courses in nursing and academic healthcare sciences education? A rapid review. Nurse Educ Today 2021 Apr;99:104812 [FREE Full text] [doi: 10.1016/j.nedt.2021.104812] [Medline: 33610051]

XSL•FO
RenderX

3. Sermeus W, Aiken LH, Van den Heede K, Rafferty AM, Griffiths P, Moreno-Casbas MT, et al. Nurse forecasting in Europe (RN4CAST): rationale, design and methodology. BMC Nurs 2011 Apr 18;10:6 [FREE Full text] [doi: 10.1186/1472-6955-10-6] [Medline: 21501487]
4. Jordan K. Massive open online course completion rates revisited: assessment, length and attrition. Int Rev Res Open Distrib Learn 2015 Jun;16(3):341-358 [FREE Full text] [doi: 10.19173/irrodl.v16i3.2112]
5. Ryan RM, Deci EL. Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being. Am Psychol 2000 Jan;55(1):68-78. [doi: 10.1037//0003-066x.55.1.68] [Medline: 11392867]
6. Badali M, Hatami J, Banihashem SK, Rahimi E, Noroozi O, Eslami Z. The role of motivation in MOOCs' retention rates: a systematic literature review. Res Pract Technol Enhanc Learn 2022 Feb 05;17:5. [doi: 10.1186/s41039-022-00181-3]
7. Romero-Frías E, Arquero JL, del Barrio-García S. Exploring how student motivation relates to acceptance and participation in MOOCs. Interact Learn Environ 2020 Jul 28;31(1):480-496. [doi: 10.1080/10494820.2020.1799020]
8. Jia M, Gong D, Luo J, Zhao J, Zheng J, Li K. Who can benefit more from massive open online courses? A prospective cohort study. Nurse Educ Today 2019 May;76:96-102. [doi: 10.1016/j.nedt.2019.02.004] [Medline: 30776535]
9. Wang W, Zhao J, Cao X, Bai Y, Cheng L, Jin S, et al. Analysis of participation and performance of MOOC learners via latent class analysis: a retrospective study based on the data of a nursing MOOC from 2018 to 2022. Nurse Educ Today 2023 Sep;128:105888. [doi: 10.1016/j.nedt.2023.105888] [Medline: 37385148]
10. Nylund KL, Asparouhov T, Muthén BO. Deciding on the number of classes in latent class analysis and growth mixture modeling: a Monte Carlo simulation study. Struct Equ Model 2007 Dec 05;14(4):535-569. [doi: 10.1080/10705510701575396]
11. Rehabilitation nursing. Wisdom Tree Network. URL: https://coursehome.zhihuishu.com/courseHome/1000003742/127615/18#teachTeam [accessed 2026-03-26]
12. Dziak JJ, Lanza ST, Tan X. Effect size, statistical power and sample size requirements for the bootstrap likelihood ratio test in latent class analysis. Struct Equ Model 2014 Jan 01;21(4):534-552 [FREE Full text] [doi: 10.1080/10705511.2014.919819] [Medline: 25328371]
13. Depaoli S, Qiu M, Liu H, Jauregui M. Class selection in growth mixture models: comparing information criteria to nonparametric and parametric Bayesian approaches. Struct Equ Model 2025 Oct 21;33(1):64-80. [doi: 10.1080/10705511.2025.2563187]
14. Tein JY, Coxe S, Cham H. Statistical power to detect the correct number of classes in latent profile analysis. Struct Equ Model 2013 Oct 01;20(4):640-657 [FREE Full text] [doi: 10.1080/10705511.2013.824781] [Medline: 24489457]
15. Shi H, Zhou Y, Dennen VP, Hur J. Correction: From unsuccessful to successful learning: profiling behavior patterns and student clusters in massive open online courses. Educ Inf Technol 2023 Sep 01;29:9039. [doi: 10.1007/s10639-023-12191-9]
16. Uijl S, Filius R, Ten Cate O. Student interaction in small private online courses. Med Sci Educ 2017 Feb 14;27:237-242. [doi: 10.1007/s40670-017-0380-x]
17. Rawat S, Kumar D, Kumar P, Khattri C. A systematic analysis using classification machine learning algorithms to understand why learners drop out of MOOCs. Neural Comput Appl 2021 May 31;33:14823-14835. [doi: 10.1007/s00521-021-06122-3]
18. Kang IG. Heterogeneity of learners' behavioral patterns of watching videos and completing assessments in massive open online courses (MOOCs): a latent class analysis. Int Rev Res Open Distribut Learn 2020;21(4):222-237. [doi: 10.19173/irrodl.v21i4.4645]
19. Fritz CO, Morris PE, Richler JJ. Effect size estimates: current use, calculations, and interpretation. J Exp Psychol Gen 2012 Feb;141(1):2-18. [doi: 10.1037/a0024338] [Medline: 21823805]
20. Schmidt FL. Statistical significance testing and cumulative knowledge in psychology: implications for training of researchers. Psychol Methods 1996;1(2):115-129. [doi: 10.1037/1082-989x.1.2.115]
21. Salas-Pilco SZ, Yang Y, Zhang Z. Student engagement in online learning in Latin American higher education during the COVID-19 pandemic: a systematic review. Br J Educ Technol 2022 May;53(3):593-619 [FREE Full text] [doi: 10.1111/bjet.13190] [Medline: 35600418]
22. Reich J. MOOC completion and retention in the context of student intent. EDUCAUSE Rev 2014 Dec 8 [FREE Full text]
23. Gilligan J, Chen W, Darzentas J. Using MOOCs to promote digital accessibility and universal design, the MOOCAP experience. Stud Health Technol Inform 2018;256:78-86. [Medline: 30371462]
24. Kizilcec RF, Piech C, Schneider E. Deconstructing disengagement: analyzing learner subpopulations in massive open online courses. In: Proceedings of the Third International Conference on Learning Analytics and Knowledge. 2013 Presented at: LAK '13; April 8-13, 2013; Leuven, Belgium. [doi: 10.1145/2460296.2460330]
25. Wei X, Saab N, Admiraal W. What rationale would work? Unfolding the role of learners' attitudes and motivation in predicting learning engagement and perceived learning outcomes in MOOCs. Int J Educ Technol High Educ 2024 Jan 22;21:5. [doi: 10.1186/s41239-023-00433-2]
26. Loh WW. Unmeasured mediator-outcome confounding distorts both indirect and direct effects. Proc Natl Acad Sci U S A 2024 Nov 19;121(47):e2418979121 [FREE Full text] [doi: 10.1073/pnas.2418979121] [Medline: 39508758]

Abbreviations

LCA: latent class analysis

XSL•FO
RenderX

MOOC: massive open online course

Edited by T Gladman; submitted 04.Jan.2026; peer-reviewed by T Schlosser, S Funghetto; comments to author 05.Mar.2026; revised version received 25.Apr.2026; accepted 03.May.2026; published 02.Jun.2026.

Please cite as:
Zeng J, Li K, Wang W, Xiao L, Wang Y, He X, Qing Y, Xue J, Zhong L, Yan T, Zheng X
Learning Participation and Performance Among Social and Academic Learners in a Nursing Massive Open Online Course: Retrospective Cohort Study
JMIR Med Educ 2026;12:e90827
URL: https://mededu.jmir.org/2026/1/e90827
doi:10.2196/90827
PMID:

©Jianqing Zeng, Kun Li, Wenxuan Wang, Lingjun Xiao, Yingmin Wang, Xinghui He, Yonghong Qing, Jingjing Xue, Lirong Zhong, Tiebin Yan, Xiuyuan Zheng. Originally published in JMIR Medical Education (https://mededu.jmir.org), 02.Jun.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Education, is properly cited. The complete bibliographic information, a link to the original publication on https://mededu.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Adoption and Use of Social Media in Health Care Among Medical Residents: Cross-Sectional Study

Zineb Zouakia[1,2], MD; Aurélia Manns[1,2], MD; Thomas Meyre[1,2], MD; Anita Burgun[1,2], MD, PhD; Natacha Kadlub[3], MD, PhD; Rosy Tsopra[1,2,4], MD, PhD

[1]Clinical Bioinformatics Laboratory, Imagine Institute, Université Paris Cité, INSERM UMR1163, 24 Boulevard du Montparnasse, Paris, France

[2]Department of Medical Informatics, Hôpital Européen Georges Pompidou, Hôpital Necker Enfants Malades, AP-HP, Paris, France

[3]Department of Maxillofacial Surgery and Plastic Surgery, MAFACE Rare Diseases Reference Centre, Faculty of Medicine, Hôpital Necker Enfants Malades, APHP, Université Paris Cité, Paris, France

[4]Université Paris Cité, UPPERS, US 007, Paris, France

Corresponding Author:
Zineb Zouakia, MD
Clinical Bioinformatics Laboratory, Imagine Institute, Université Paris Cité, INSERM UMR1163, 24 Boulevard du Montparnasse, Paris, France

Abstract

Background: Social media apps are widely used by health care professionals despite security and regulatory risks. Identifying factors associated with this use is important for developing effective risk-reduction strategies.

Objective: This study aimed to investigate how medical residents use 6 popular social media apps in professional tasks and to identify factors influencing their adoption in health care, using the validated Unified Theory of Acceptance and Use of Technology 2 (UTAUT2) model.

Methods: An anonymous web-based survey was conducted between June 2024 and November 2024 among medical residents in France. Participants reported demographic characteristics, frequency, and professional contexts of use for 6 apps (Facebook, Instagram, LinkedIn, Messenger, TikTok, and WhatsApp) and completed UTAUT2-based items. The model was adapted by adding a technology trust construct. Descriptive analyses were performed for all apps. With a sufficient sample size, partial least squares structural equation modeling was conducted for WhatsApp to identify factors associated with behavioral intention and use behavior.

Results: A total of 137 residents (n=87, 63.5% female participants) across 40 specialties completed the survey. WhatsApp was the most widely and professionally used app (n=127, 92.7%), with 75.9% (n=104) using it at least many times per week. It was primarily used for patient care, including written transmissions (n=86, 62.8%), case discussions (n=76, 55.5%), and specialist advice (n=86, 62.8%), as well as for professional networking (n=62, 45.3%). Messenger was used by 46.7% (n=64) of participants for similar purposes. Facebook (n=35, 25.6%) and LinkedIn (n=20, 14.6%) were mainly used for education and networking, whereas Instagram (n=11, 8%) was rarely used, and TikTok was not used for professional purposes. Regarding adoption factors, WhatsApp had the highest overall scores, including the highest performance expectancy (mean 5.4, SD 1.12), behavioral intention (mean 5.28, SD 1.15), and use behavior (mean 5.91, SD 1.30), with high effort expectancy (mean 6.82, SD 0.55) and facilitating conditions (mean 6.07, SD 0.85). LinkedIn showed the highest social influence (mean 5.05, SD 1.06), whereas Instagram showed the highest hedonic motivation (mean 6.61, SD 0.51). Technology trust scores were low across all apps, ranging from a mean of 2.23 (SD 1.16) for Facebook to a mean of 3.72 (SD 1.39) for LinkedIn. In the partial least squares structural equation modeling analysis for WhatsApp, habit was the only significant predictor of behavioral intention (β=.53; P<.001) and use behavior (β=.45; P<.001).

Conclusions: WhatsApp dominates professional use among residents despite low trust in its security, and its use is mainly driven by habit. Secure alternatives with features similar to popular social media apps, supported by institutional policies and digital professionalism training, are needed to encourage physicians to better consider safety when using social media.

(JMIR Med Educ 2026;12:e83475) doi:10.2196/83475

KEYWORDS

digital health; social media; messaging; WhatsApp; Facebook; health communication; health care professionals; physicians; medical residency; clinical practice; technology acceptance; UTAUT2; Unified Theory of Acceptance and Use of Technology 2

XSL•FO
RenderX

Introduction

Social media, including messaging apps, are now common tools in health care [1]. Their use is boosted by the popularity of smartphones and quick communication, especially among young people [2,3]. For example, WhatsApp is used by more than 90% of physicians to share clinical information [4-7]; Facebook is used by more than 70% of health care professionals, primarily for continuing education [8]; LinkedIn is used by more than half for career development [9,10]; and Instagram is popular among the younger users [11], with 68% of medical residents and students using it daily [12]. These apps are used for a wide range of health care activities [1,2,13]: (1) clinical coordination (eg, shift rotation and clinical updates); (2) peer consultation (eg, case discussion and specialist advice); (3) interaction with patients (eg, responding to questions and providing test result updates), with more than 12% of residents having patients (or patients' family members) listed as social media "friends" [14]; (4) professional networking; and (5) continuing education (eg, access to the latest medical information).

However, the use of such apps raises significant concerns around patient confidentiality and data security, as they fail to comply with data protection standards such as the European General Data Protection Regulation [15] or the US Health Insurance Portability and Accountability Act [16], making them inappropriate for use in health care [17]. To guarantee the security of patient data, secure tools have been developed, such as institutional platforms for patient exchange (eg, intrahospital provider portals; regional or national secure messaging systems [18-20]); private solutions like Siilo [21,22], a secure messaging app for medical professionals; or Pixacare, designed for secure medical photography exchange [23]. Despite the availability of these secure tools, medical professionals continue to prefer using popular social media apps [4], raising questions about the underlying reasons for their continued use.

Several studies have explored the reasons behind the increasing use of social media apps in health care. For example, McGowan et al [24] found that physicians' use was driven by the ease of use and usefulness in achieving better outcomes, while Nikolic et al [6] pointed out factors such as being free, accessible, widely adopted by peers, and offering features like group chat and data sharing. Additionally, Youssef et al [11] reported that orthopedic surgeons viewed social media as a means to boost visibility, connect with patients, stay informed, and advance their careers. However, prior studies have limitations: (1) they focused on single apps (mostly WhatsApp [7,25-28]) or specific specialties (mostly surgery; eg, orthopedics [11,12,29-32]); (2) they reported general reasons for using social media without examining each app separately [5,6,8,9,12,24,31,33-36]; and (3) they did not use a validated theoretical model or relied on earlier models [24,25,35,37]. For example, the technology acceptance model [38] does not account for relevant factors like habit or social influence.

Here, we aimed to examine the use of social media apps in health care, by investigating how medical residents use 6 popular social media apps in professional contexts, and to identify factors associated with their adoption using the validated Unified Theory of Acceptance and Use of Technology 2 (UTAUT2) model [39].

Methods

Study Design

We conducted a cross-sectional study using an anonymous web-based survey to investigate the professional use of social media apps among medical residents and to identify the factors influencing their adoption in health care. The study was designed and reported in accordance with the CHERRIES (Checklist for Reporting Results of Internet E-Surveys) [40] (Checklist 1).

Participants and Recruitment

Eligible participants were medical residents currently enrolled in postgraduate training programs in France. Participants were recruited between June 25 and November 27, 2024. Recruitment was conducted using a nonprobability convenience sampling strategy. The survey link was disseminated via institutional mailing lists, university networks, and social media through medical residents' groups. Participants were also encouraged to share the survey with peers, enabling snowball sampling. The recruitment message included a brief description of the study objectives, eligibility criteria, estimated completion time, and a link to the survey. Participation was voluntary, and no incentives were provided.

Survey Development and Administration

The questionnaire was developed to assess the professional use of social media applications among medical residents and the factors influencing their adoption. It included both descriptive and theory-based components, with the latter based on the UTAUT2 model [39], adapted to the health care context. It was administered online using LimeSurvey and was designed to be completed in approximately 15 minutes. Before dissemination, it was pilot-tested on 3 medical residents to ensure clarity, relevance, and usability.

The survey was structured into 4 sections covering participant characteristics, personal and professional use of social media apps, frequency and contexts of professional use, and factors influencing adoption and use. Participants accessed the survey through a web link and could complete it using a computer or mobile device. They were able to review and modify their responses before submission. No randomization of items was implemented. Adaptive questioning was used based on the apps selected by participants to reduce the number and complexity of questions. To minimize duplicate entries, survey settings were configured to restrict multiple submissions from the same participant.

Measures

Selection of Social Media Apps

The social media apps included in this study were selected based on data from the Digital 2024 Global Overview Report by DataReportal [41], which provides global statistics on digital usage. Apps were chosen according to their number of active users worldwide and to represent different categories. We selected the 3 most widely used multipurpose social media

XSL•FO
RenderX

platforms (Facebook, Instagram, and TikTok), the 2 most widely used mobile messaging apps (WhatsApp and Messenger), and the most widely used professional networking platform (LinkedIn). WeChat, although among the most widely used applications globally, was excluded due to its limited use outside of China.

Participants' Characteristics

Participants were asked to provide demographic information, including age, gender, medical specialty, and whether they used each app in their personal lives.

Use of Social Media Apps in Health Care

For each app used, participants reported the frequency of use in their professional practice using a 5-point Likert scale ("Many times per day," "Many times per week," "Less than once a week," "Less than once a month," and "Never").

Participants also specified the professional contexts in which each app was used. These were assessed using a predefined list of 12 tasks covering 3 domains: patient care (eg, written transmission, case discussion, specialist advice), medical knowledge communication (eg, consumption or creation of educational content), and professional networking (eg, exchange of opportunities, relationship building). The list of tasks was developed empirically based on the authors' clinical experience as medical residents and doctors, informal discussions with peers, and iterative refinement during questionnaire development. This approach aimed to capture the most common real-world uses of social media apps in clinical practice.

Factors of Acceptance and Use in Health Care

To investigate the determinants of social media adoption and use in health care, we used the UTAUT2 model from Venkatesh et al [39], as it is a validated and widely used model for examining factors that influence technology adoption and use [42], including in health care research (eg, telemedicine systems [43] and mobile health apps [44]).

Adaptation and Use of the UTAUT2 Model

The original UTAUT2 model includes 7 key constructs (eg, performance expectancy) that predict 2 constructs representing technology use [39]: behavioral intention and use behavior (Table 1). Each construct contains 1 or more items (Table 1).

Table . Description of the constructs used in our model.

Constructs	Definition	Number of items	Source	Included in model
Performance expectancy	Perceived efficiency gain from using the technology	3	Venkatesh et al [39]	Yes
Effort expectancy	Ease of use	4	Venkatesh et al [39]	Yes
Social influence	Peer influence on the decision to use the technology	3	Venkatesh et al [39]	Yes
Facilitating conditions	Resources and support available to users	4	Venkatesh et al [39]	Yes
Hedonic motivation	Fun or pleasure derived from using the technology	3	Venkatesh et al [39]	Yes
Price value	Perceived value relative to the cost of use	3	Venkatesh et al [39]	No
Habit	Degree to which technology use becomes automatic	3	Venkatesh et al [39]	Yes
Technology trust	Degree to which technology is perceived reliable and secure	3	Casaló et al [45], Yee - Loong Chong et al [46], Lebrument et al [47]	Yes
Behavioral intention	Intention to use the technology	3	Venkatesh et al [39]	Yes
Use behavior	Actual use (frequency)	1	Venkatesh et al [39]	Yes

For the 7 constructs and behavioral intention, each item is measured using a 7-point Likert scale of agreement (ranging from 1 [strongly disagree] to 7 [strongly agree]), while for use behavior, it is measured using a 7-point Likert scale of frequency of use (from 1 [never] to 7 [many times per day]) [39].

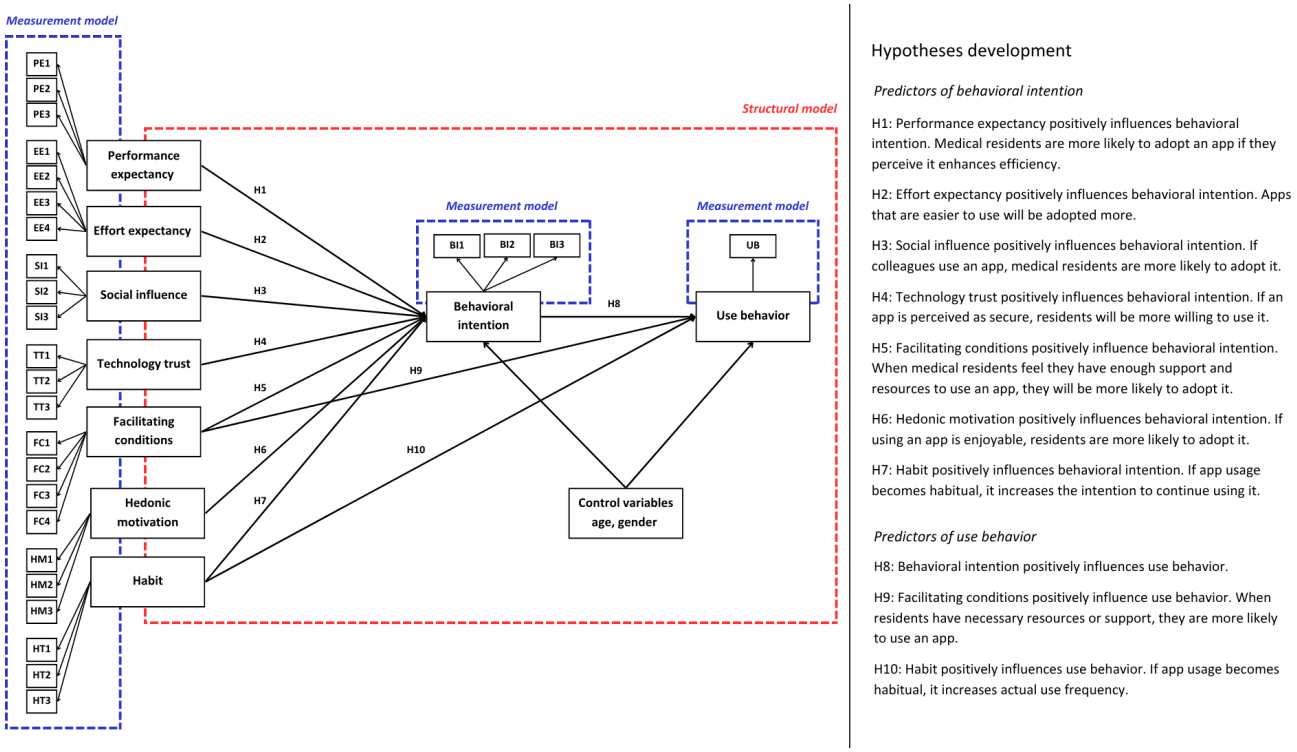
For the purposes of this study, the model was adapted in several ways (Table 1). First, the price value construct was excluded, as all apps assessed were free of charge. Second, a technology trust construct was added to account for perceived reliability and security, which are particularly relevant in the health care context, especially with the sharing of sensitive patient data. This construct and its associated items were added following the guidance of previous studies [45-47].

All items were translated into French based on previously validated translations [48]. While UTAUT2 constructs were measured using 7-point Likert scales, the frequency of use (use behavior) was initially assessed using a 5-point scale for clarity and interpretability. For the statistical analysis, these values were converted to a 7-point scale using linear interpolation to align with the other constructs.

XSL•FO
RenderX

Based on the original UTAUT2 hypothesis development [39], we built our model with the following hypotheses: (1) performance expectancy, effort expectancy, social influence, technology trust, facilitating conditions, hedonic motivation, and habit positively influence behavioral intention; (2) behavioral intention, facilitating conditions, and habit positively influence use behavior; (3) age and gender serve as control variables. The final model and its hypotheses are presented in Figure 1.

Figure 1. Theoretical framework and hypotheses development of our study based on the Unified Theory of Acceptance and Use of Technology 2 (UTAUT2) model. The measurement model, outlined in blue, illustrates the relationship between the items and their respective constructs, while the structural model, outlined in red, shows the relationships between the predictor constructs and the predicted constructs. BI: behavioral intention; EE: effort expectancy; FC: facilitating conditions; H: hypothesis; HM: hedonic motivation; HT: habit; PE: performance expectancy; SI: social influence; TT: technology trust; UB: use behavior.



Analysis of the Model According to the Partial Least Squares Structural Equation Modeling Method

To test the model hypotheses, we used partial least squares structural equation modeling (PLS-SEM), a variance-based approach commonly applied in causal modeling [49]. PLS-SEM comprises 2 components: a measurement model, which assesses the relationships between observed items and their corresponding constructs, and a structural model, which evaluates the relationships between constructs and represents the hypothesized paths (Figure 1). Analyses were conducted in RStudio (version 4.3.3; Posit Software, PBC) using the SEMinR package [50,51].

The analysis was performed in 3 steps. First, descriptive statistics were computed to summarize construct scores across apps, using means and SDs.

Second, the measurement model was assessed to evaluate the reliability and validity of the constructs. Prior to analysis, sample size adequacy was verified. The minimum required sample size, based on the number of predictor constructs, was 70 participants [49,52-54]. This threshold was met only for WhatsApp (n=127), leading to the exclusion of Messenger (n=64), Facebook (n=35), LinkedIn (n=20), and Instagram (n=11) from the PLS-SEM analysis. The “Other” gender category (n=2) was excluded due to its very small representation, resulting in a final analytical sample of 125 responses. Measurement model evaluation for WhatsApp included the assessment of internal consistency reliability, convergent validity, and discriminant validity, all of which met the recommended thresholds (Multimedia Appendix 1). Indicator reliability was satisfactory for all items except for 2 items within the facilitating conditions construct. However, they were retained because the construct as a whole demonstrated strong internal consistency and convergent validity, and the items captured distinct, theoretically relevant aspects of the construct, thus maintaining content validity [55].

Third, the structural model was assessed to test the hypothesized relationships between constructs. Collinearity was evaluated using variance inflation factors, which were below the recommended thresholds. Path coefficients (β) were estimated using a regression-based approach, and their significance was assessed through bootstrapping with 5000 resamples [49]. The explanatory power of the model was evaluated using the coefficient of determination (R^2) (Multimedia Appendix 1).

Free-Text Comments

In addition to the structured questionnaire items, participants were invited to provide optional free-text comments describing positive and negative perceptions of each app. Comments were analyzed using a thematic approach. Responses were reviewed

XSL•FO
RenderX

and grouped into recurring themes. The analysis was conducted by one author (ZZ) and reviewed by another author (RT). Any discrepancies were resolved through discussion until consensus was reached. The frequency of each theme was calculated, and illustrative quotes were selected (Multimedia Appendix 2). These responses were used to complement the quantitative results concerning factors influencing adoption and use (UTAUT2 analysis), providing additional insights into participants' experiences.

Statistical Analysis

Only fully completed questionnaires were included in the analysis. Incomplete responses were excluded to ensure data integrity. All statistical analyses were conducted using RStudio (version 4.3.3; Posit Software, PBC). Descriptive statistics were used to summarize participant characteristics, frequency and context of app use, and UTAUT2 construct scores. Categorical variables were reported as counts and percentages, and continuous variables as means and SDs. To investigate factors influencing technology adoption and use, PLS-SEM was conducted for WhatsApp as described earlier. A 2-sided significance level of $P<.05$ was considered statistically significant.

Ethical Considerations

This study was approved by the CERAP-HP Center Ethics Committee (IRB: IORG0010044). Participants were informed about the study objectives, eligibility criteria, and procedures on the survey's introductory page. Informed consent was obtained electronically, as completion of the questionnaire implied consent to use anonymized responses for research purposes. No personally identifiable information was collected. Data were anonymous, stored on secure servers, and scheduled for deletion after publication. No financial or material compensation was provided.

Results

Participants' Characteristics

A total of 137 medical residents were included in the study (4 were excluded from the analysis due to inconsistent responses). Most of them were women (n=87, 63.5%) and were aged between 26 and 40 years (n=102, 74.5%). A total of 40 distinct specialties were represented, with pediatrics (n=18, 13.1%), general medicine (n=12, 8.8%), and radiology and medical imaging (n=10, 7.3%) being the most represented (Multimedia Appendix 3).

In private life, the most used apps among the 137 medical residents were WhatsApp (n=133, 97.1%), Messenger (n=132, 96.4%), Facebook (n=120, 87.6%), and Instagram (n=105, 76.6%). LinkedIn (n=10, 7.3%) and TikTok (n=12, 8.8%) were less common.

Use of Social Media Apps in Health Care

Frequency of Use in Health Care

In professional life, the most used app among the 137 medical residents was WhatsApp (n=127, 92.7%). Messenger, Facebook, LinkedIn, and Instagram were used but very less often (n=64, 46.7%; n=35, 25.5%; n=20, 14.6%; n=11, 8%, respectively). TikTok was never used to achieve professional tasks. WhatsApp was used at least "many times per week" by 75.9% (n=104) of residents, while Messenger was used by 32.1% (n=44), Facebook by 10.2% (n=14), Instagram by 4.4% (n=6), and LinkedIn by 3% (n=4; Table 2).

Table . Frequency of professional use by app (N=137).

Frequency	Facebook, n (%)	Instagram, n (%)	LinkedIn, n (%)	Messenger, n (%)	TikTok, n (%)	WhatsApp, n (%)
Many times per day	3 (2.2)	1 (0.7)	2 (1.5)	15 (11)	0 (0)	64 (46.7)
Many times per week	11 (8)	5 (3.7)	2 (1.5)	29 (21.2)	0 (0)	40 (29.2)
Less than once a week	12 (8.8)	4 (2.9)	8 (5.8)	13 (9.5)	0 (0)	17 (12.4)
Less than once a month	9 (6.6)	1 (0.7)	8 (5.8)	7 (5.1)	0 (0)	6 (4.4)
Never	102 (74.4)	126 (92)	117 (85.4)	73 (53.2)	0 (0)	10 (7.3)

Medical Context of Use

The app use depended on context. For patient care, WhatsApp was the most used app among the 137 medical residents, with 62.8% (n=86) of residents using it for written transmission to colleagues and specialist advice, and 55.5% (n=76) for case discussions. Messenger was also used for these purposes but less often. Facebook, Instagram, and LinkedIn were rarely used for clinical tasks.

For medical knowledge communication, Facebook was the top app, mostly for consuming medical training content (n=26, 19.0%). All apps showed minimal use for sharing daily professional life and creating content.

For networking, WhatsApp was the most popular app for exchanging professional opportunities (n=62, 45.3%), followed by Messenger (n=27, 19.7%) and Facebook (n=25, 18.2%). WhatsApp was also the most used app for building professional relationships (n=33, 24.1%), while LinkedIn was used by 12.4% (n=17) of residents for this purpose (Table 3).

XSL•FO
RenderX

Table . Professional tasks by app (N=137).

Tasks	Facebook, n (%)	Instagram, n (%)	LinkedIn, n (%)	Messenger, n (%)	WhatsApp, n (%)
Patient care					
I make a written transmission to a colleague or another health care professional	2 (1.5)	1 (0.7)	2 (1.5)	32 (23.4)	86 (62.8)
I discuss a patient case with my colleagues or another health care professional	6 (4.4)	2 (1.5)	2 (1.5)	31 (22.6)	76 (55.5)
I request a specialized advice from a colleague or another health care professional	2 (1.5)	2 (1.5)	1 (0.7)	10 (7.3)	86 (62.8)
I discuss a complex case with a multidisciplinary group of experts	3 (2.2)	1 (1.4)	1 (0.7)	2 (1.5)	13 (9.5)
I discuss with a patient about his or her care management	0 (0)	0 (0)	0 (0)	5 (3.7)	12 (8.8)
Medical knowledge communication					
I consume medical training content	26 (19)	11 (8)	8 (5.8)	3 (2.2)	7 (5.1)
I share my daily life as a health care professional with the general public	0 (0)	1 (0.7)	1 (0.7)	4 (2.9)	2 (1.5)
I create training content for health care students	1 (0.7)	0 (0)	1 (0.7)	1 (0.7)	1 (0.7)
I create educational content intended for the general public	0 (0)	1 (0.7)	2 (1.5)	0 (0)	0 (0)
Networking					
I exchange professional opportunities with my peers (job, on-call duties, replacement, courses, etc)	25 (18.2)	4 (2.9)	9 (6.6)	27 (19.7)	62 (45.3)
I build professional relationships	10 (7.3)	3 (2.2)	17 (12.4)	16 (11.7)	33 (24.1)
I share advancements in my specialty with the scientific community or with my colleagues	2 (1.5)	2 (1.5)	4 (2.9)	2 (1.5)	8 (5.8)

Factors of Acceptance and Use in Health Care

Facebook

Facebook scored high in effort expectancy (mean 6.72, SD 0.48) and facilitating conditions (mean 6.11, SD 0.73), indicating ease of use and adequate support. Residents noted its ease of use for sharing information on shift exchanges, medical seminars, and knowledge within medical communities (Ex1, 2, and 4 in Multimedia Appendix 2). However, performance expectancy (mean 3.89, SD 1.53) was moderate, suggesting limited professional effectiveness. Some considered it outdated or inappropriate for professional use, citing misinformation and poor moderation (Ex6-8 in Multimedia Appendix 2). The technology trust score was the lowest among apps (mean 2.23, SD 1.16), reflecting concerns about its security and privacy (Ex10 in Multimedia Appendix 2). Some also criticized it as time-wasting and addictive (Ex9 in Multimedia Appendix 2) (Table 4).

XSL•FO
RenderX

Table . Scores of the model's constructs by app. Scores are reported in mean (SD).

	Facebook (n=35), mean (SD)	Instagram (n=11), mean (SD)	LinkedIn (n=20), mean (SD)	Messenger (n=64), mean (SD)	WhatsApp (n=127), mean (SD)
Performance expectancy	3.89 (1.53)	4.09 (1.45)	4.20 (1.30)	4.67 (1.26)	5.43 (1.12)
Effort expectancy	6.72 (0.48)	6.93 (0.16)	5.47 (1.45)	6.84 (0.36)	6.82 (0.55)
Social influence	3.61 (1.11)	3.70 (1.59)	5.05 (1.06)	3.71 (1.19)	4.19 (1.18)
Technology trust	2.23 (1.16)	2.27 (0.88)	3.72 (1.39)	2.41 (1.36)	3.39 (1.43)
Facilitating conditions	6.11 (0.73)	6.45 (0.56)	5.94 (0.76)	6.14 (1.00)	6.07 (0.85)
Hedonic motivation	4.57 (1.46)	6.61 (0.51)	4.48 (1.22)	4.28 (1.38)	4.01 (1.38)
Habit	3.10 (1.29)	2.42 (1.07)	2.80 (1.43)	3.54 (1.43)	3.97 (1.50)
Behavioral intention	4.05 (1.46)	4.39 (1.62)	5.18 (1.17)	4.36 (1.35)	5.28 (1.15)
Use behavior	4.34 (1.41)	4.82 (1.23)	3.85 (1.45)	5.22 (1.39)	5.91 (1.30)

Instagram

Instagram had the highest effort expectancy (mean 6.93, SD 0.16) and hedonic motivation (mean 6.61, SD 0.51) among apps, showing that residents found it very easy to use and enjoyed using it for educational content (Ex11 and 12 in Multimedia Appendix 2). However, performance expectancy (mean 4.09, SD 1.45) was moderate, indicating limited professional effectiveness. It was also viewed as addictive, inappropriate for professional use, and lacking relevance (Ex15-17 in Multimedia Appendix 2). The technology trust score (mean 2.27, SD 0.88) was low as well (Table 4).

LinkedIn

LinkedIn scored the highest in facilitating conditions (mean 5.94, SD 0.76), indicating adequate support and resources for professional use. Residents noted strengths in networking and career development (Ex19 and 20 in Multimedia Appendix 2). However, performance expectancy was moderate (mean 4.20, SD 1.30), and use behavior was the lowest among apps (mean 3.85, SD 1.45), indicating limited practical use despite its professional focus. Residents confirmed that LinkedIn had little relevance to their daily practice (Ex21 in Multimedia Appendix 2). Technology trust (mean 3.72, SD 1.39) was moderate but the highest among apps. Social influence (mean 5.05, SD 1.06) was also the highest, suggesting adoption was largely peer-driven (Table 4).

Messenger

Messenger scored the highest for effort expectancy (mean 6.84, SD 0.36) and facilitating conditions (mean 6.14, SD 1.00), reflecting speed, ease of use, and convenient contact access (Ex24-26 in Multimedia Appendix 2). However, the technology trust score (mean 2.41, SD 1.36) was low because of security concerns (Ex30 in Multimedia Appendix 2). Some residents considered it unsuitable for professional use, citing its link to Facebook, its outdated nature, and missing features found in WhatsApp such as message pinning (Ex28-31 in Multimedia Appendix 2 and Table 4).

WhatsApp

Average Scores of the Factors Influencing App Use

WhatsApp scored the highest in performance expectancy (mean 5.43, SD 1.12), behavioral intention (mean 5.28, SD 1.15), and use behavior (mean 5.91, SD 1.30) among apps, making it the most frequently used and effective app for professional use. Effort expectancy (mean 6.82, SD 0.55) and facilitating conditions (mean 6.07, SD 0.85) were also high, reflecting ease of use and practical features like phone number integration (Ex32 and 34 in Multimedia Appendix 2). Habit was moderate but the highest among apps (mean 3.97, SD 1.50), indicating WhatsApp's growing routine use, supported by its popularity and suitability for professional interactions (Ex33 and 37 in Multimedia Appendix 2). Technology trust was also moderate (mean 3.39, SD 1.43) but higher than Facebook, Instagram, and Messenger, with some residents viewing it as relatively more secure despite ongoing privacy concerns (Ex38 and 39 in Multimedia Appendix 2 and Table 4).

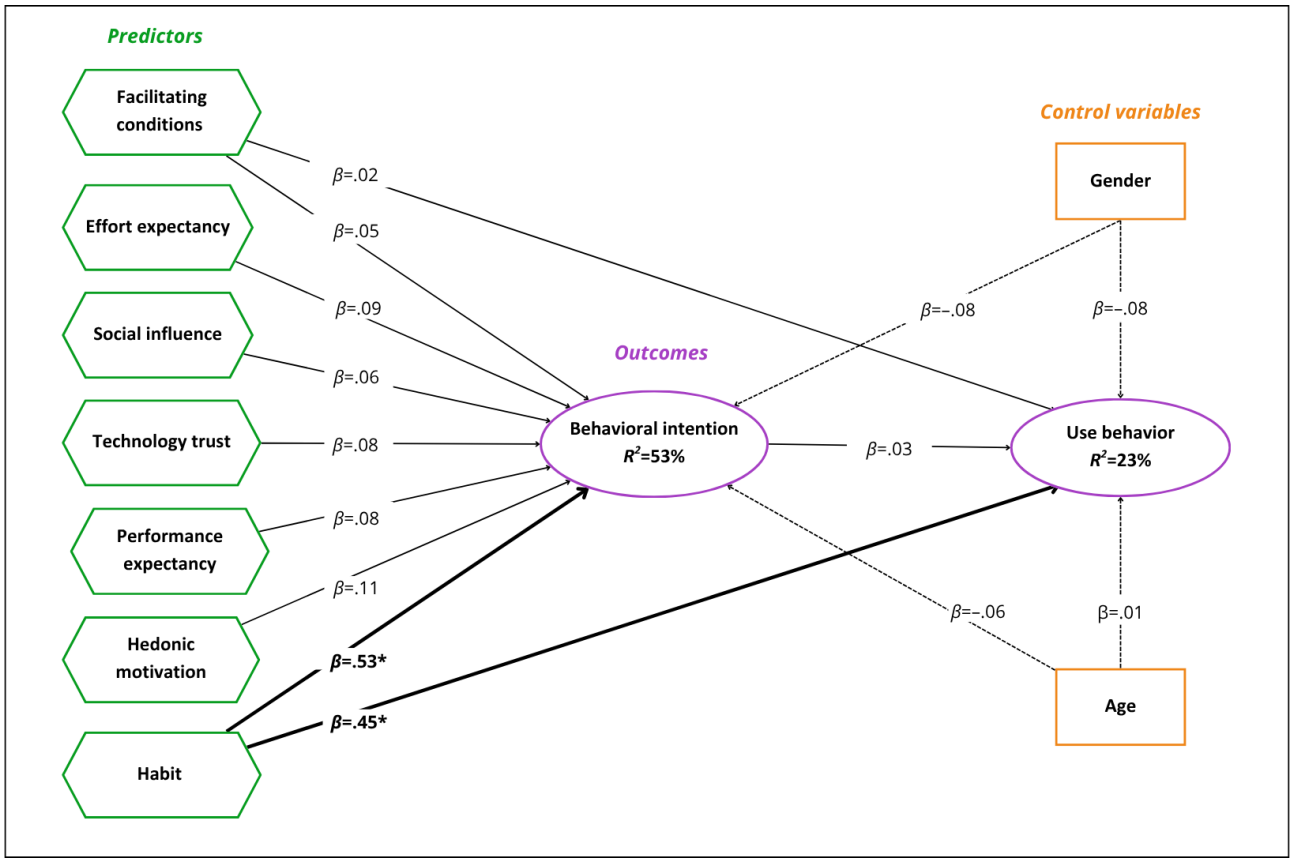
Factors Influencing Acceptance and Use According to PLS-SEM Analysis

Overall, the model explained 53% of the variance in behavioral intention and 23% of the variance in use behavior (Figure 2). The hypothesized relationships were tested (Table 5, Figure 2). Habit was the only construct with significant effects, showing a positive association with both behavioral intention (H7: β=.53, P<.001) and use behavior (H10: β=.45, P<.001).

All other hypothesized paths were not significant. The control variables (age, gender) also showed weak, nonsignificant effects on behavioral intention and use behavior.

XSL•FO
RenderX

Figure 2. Summary of the structural model results for WhatsApp. The continuous arrows represent the path coefficients (β) from the predictor constructs to the predicted constructs; the dotted arrows represent the β from the control variables to the predicted constructs; the arrows in bold represent the significant path coefficients; *P<.001; the coefficient of determination (R^2) represents the explanatory power of the model.



XSL•FO
RenderX

Table . Structural model results for WhatsApp, including path coefficients (β), SE, *t* values, and 95% CI for the tested hypotheses.

Path	β	*B* [a]	SE	*t* value	95% CI	Hypotheses (H)
BI[b]						
PE[c]→BI	.08	.09	0.08	1.03	–0.07 to 0.25	H1 (not supported)
EE[d]→BI	.09	.09	0.10	0.88	–0.12 to 0.27	H2 (not supported)
SI[e]→BI	.06	.06	0.07	0.83	–0.08 to 0.20	H3 (not supported)
TT[f]→BI	.08	.08	0.08	0.96	–0.08 to 0.24	H4 (not supported)
FC[g]→BI	.05	.07	0.08	0.65	–0.08 to 0.23	H5 (not supported)
HM[h]→BI	.11	.12	0.08	1.24	–0.05 to 0.28	H6 (not supported)
HT[i]→BI	.53[j]	.52	0.09	5.96	0.34 to 0.69	H7 (supported)
Age→BI	–.06	–.06	0.06	–1.00	–0.18 to 0.07	—[k]
Gender→BI	–.08	–.08	0.06	–1.28	–0.21 to 0.04	—
UB[l]						
BI→UB	.03	.02	0.11	0.31	–0.19 to 0.23	H8 (not supported)
FC→UB	.02	.04	0.13	0.12	–0.22 to 0.29	H9 (not supported)
HT→UB	.45[j]	.45	0.10	4.34	0.24 to 0.65	H10 (supported)
Age→UB	.01	.01	0.09	0.13	–0.15 to 0.18	—
Gender→UB	–.08	–.07	0.08	–0.95	–0.22 to 0.10	—

[a]*B*: bootstrap path coefficient.
[b]BI: behavioral intention.
[c]PE: performance expectancy.
[d]EE: effort expectancy.
[e]SI: social influence.
[f]TT: technology trust.
[g]FC: facilitating conditions.
[h]HM: hedonic motivation.
[i]HT: habit.
[j]*P*<.001.
[k]Not applicable.
[l]UB: use behavior.

Discussion

Principal Findings

This study examined the professional use of 6 popular social media platforms among 137 French medical residents across 40 different specialties. Regarding the context of professional use, over 90% of residents reported using WhatsApp professionally, particularly for written transmissions, patient case discussions, and seeking specialist advice. Messenger was used less often for similar purposes. Surprisingly, WhatsApp was also the top app for professional networking, far ahead of LinkedIn, which was used by less than 15% of residents. About 25% of residents used Facebook, mostly as the top app for medical training. Conversely, Instagram was rarely used, mainly for education, and TikTok was never used. Regarding the factors influencing adoption and use, WhatsApp ranked highest overall, leading in performance expectancy, behavioral intention, and use behavior, and showing high effort expectancy, indicating it was perceived as highly effective, easy to use, and likely to remain in use. LinkedIn ranked highest for social influence (peer effect), while Instagram ranked highest for hedonic motivation (entertainment value). Technology trust was low across apps, particularly for Facebook, Instagram, and Messenger, reflecting common security and privacy concerns. Notably, Habit was the only construct significantly associated with both the intention to use and actual use of WhatsApp, suggesting that routine use may outweigh security concerns. These results suggest that informal, noncompliant apps are widely used in clinical practice, mainly because of habit and convenience, even when trust in their security is low.

Comparison With Prior Work

Regarding the context of professional use, our study found WhatsApp to be the most used app among residents (127/137, 93%), a result that is similar to O'Sullivan et al [7], who reported 100% use among Irish hospital residents. In contrast, Van

XSL•FO
RenderX

Ravenswaay et al [8] found Facebook to be the most popular platform among health care professionals (71%) in a survey conducted by the Clinical Education Alliance, based in the United States. In our study, WhatsApp was mainly used for patient care, including transmissions and specialist advice (86/137, 63%) and case discussions (55%), which is similar to the findings in Australia, where 85% of physicians use it for clinical communication (eg, patient management details, imaging reports) [6]. Similarly, Cherrez-Ojeda et al [4] reported a 93% WhatsApp usage rate for physician communication in Ecuador, without detailing clinical contexts. Surprisingly, our study found WhatsApp to be the top app for professional networking, surpassing LinkedIn, which is specifically designed for this purpose. Youssef et al [11] also found that German surgeons preferred messenger apps (34%) over employment platforms (27%). Lastly, our results found limited and mostly passive use of social media for education, consistent with previous studies [8,11,12].

Regarding the factors influencing adoption and use, our study showed that Habit significantly influenced both the intention to use and the actual use of WhatsApp among medical residents, a finding consistent with another study using elements of the technology acceptance model and UTAUT2 [35]. Additionally, other factors such as perceived usefulness [24,35] and ease of use [24] have also been identified as predictors of social media use. However, comparing with other studies is challenging because they differ in terms of studied population (eg, physicians, pharmacists, allied employees [35]); theoretical models used (eg, technology acceptance model [24,25,35], UTAUT2 [35]); or factors examined (eg, habit is not systematically studied [24,25]). The absence of significant effects of other factors in our study may reflect our population: medical residents, a digitally fluent group likely in a late-stage adoption phase, where WhatsApp use is already routine. In such contexts, technology use often becomes automatic, and habit can outweigh the influence of early adoption factors such as perceived usefulness or ease of use, a pattern well documented in postadoption research [56-58].

Implications for the Future

Our study showed the widespread use of social media in health care, despite security and regulation concerns. Restrictive policies cannot be a solution to prevent security breaches and may even be counterproductive. Instead, various strategies should be adopted at different levels:

IT level: Secure health care apps should be developed with the desirable features of current social media apps while incorporating strong security safeguards. As shown in a prior study [59], required features should include usability (eg, easy and fast), interoperability with electronic health records, safety (eg, secure authentication and secure certified hosting), and functionalities, such as “instant messaging,” “group discussion,” “notification,” or “data storage in electronic health records” [59]. Additionally, the Medical Informatics community has recognized the importance of human factors in the safe and effective adoption of new technologies [60]. As suggested by Kushniruk and Kaufman [61] in 2024, this requires a multilevel perspective, from the individual user to the complex social and organizational context. In parallel, the American Medical Association stressed the need to integrate physicians’ perspectives throughout development so that digital solutions enhance care, fit into workflows, and address concerns such as liability, transparency, and trust [62].

Educational level: Training in digital health, including security, patient confidentiality [63-65], and regulations [66,67], should be implemented in medical curricula [59]. Digital professionalism (ie, responsible use of social media and professional integrity) should be taught as well [68]. Evidence shows that targeted digital professionalism programs improve knowledge of appropriate online behavior, professional conduct, and confidence in privacy and security practices [69]. Innovative approaches, such as interactive workshops reviewing social media content, can lead to meaningful behavior changes [70]. Likewise, promoting critical thinking in digital health [71] and interprofessional team-based training [72] can further strengthen competencies, enhance safety culture, and prepare physicians to practice effectively in a digital health care environment.

Institutional level: Hospitals and health institutions should establish clear, enforceable policies for digital communication, including acceptable tools [59]. They should also provide professional smartphones to separate personal and work use, thereby reducing security risks [3]. Additionally, they should implement and promote the adoption of secure apps, focusing on a bottom-up approach (eg, direct engagement with frontline clinicians and designation of an experienced colleague to lead adoption), which has been proven to increase adoption [73,74]. Finally, institutions should allocate resources to secure communication as part of broader efforts to improve patient safety, efficiency, and regulatory compliance. Financial investment is essential to enable practice change and technology adoption. For example, Sevick et al [75] demonstrated that implementing a digital tool not only improved provider and patient satisfaction but also reduced per-patient costs, illustrating that such investments can be both impactful and cost-effective.

Limitations

This work has several limitations. First, it focused on French medical residents, which may limit the generalizability to other professional groups, health care systems, or cultural contexts. However, residents are a relevant group as they are heavily involved in clinical work and represent the next generation of physicians likely to use social media in clinical practice. The sample size (N=137) is relatively small, but it spanned 40 specialties, enabling analysis across various disciplines. Due to the sample size, structural modeling (PLS-SEM) could not be performed for all apps [55]. However, we still conducted descriptive factor analysis for each app and used free-text responses to ensure a broad interpretative scope despite the analytical constraints.

Second, reliance on self-reported data may introduce “selection bias,” possibly overrepresenting residents more engaged with social media [76], which may have been further influenced by the use of social media as one of the recruitment channels, potentially favoring more active users. Responses may also be affected by “social desirability bias,” as residents may have described behaviors that do not fully reflect their actual practices

XSL•FO
RenderX

due to a desire to appear more responsible or avoid judgment [77]. To limit this bias, the survey was anonymous, which encouraged participants to answer honestly. Likewise, "recall bias" might occur [78], as participants may not remember exactly how they used each app. We mitigated this by designing a structured, detailed questionnaire, breaking down distinct dimensions of use (frequency, professional tasks, and adoption factors), encouraging residents to reflect more precisely on their practices.

Finally, we limited the analysis to 6 adoption factors (performance expectancy, effort expectancy, social influence, facilitating conditions, hedonic motivation, habit), as defined by the validated and widely used UTAUT2 model for technology acceptance and use behavior. We also added a technology trust factor to address health care–specific concerns regarding data privacy and security. However, other factors, such as institutional factors (eg, hospital management policies [25]), could also be considered in further studies.

Conclusion

This study found that medical residents overwhelmingly used WhatsApp for professional communication, particularly in patient care and networking, with Messenger serving similar functions at a lower frequency. Facebook, Instagram, and LinkedIn were used far less often and mainly for limited purposes, such as education and career development. The analysis of factors influencing adoption and use showed that *habit* was the only significant factor influencing WhatsApp's use, highlighting the dominance of routine over security concerns. Secure alternatives with similar features to popular social media apps, along with proactive institutional policies and digital professionalism training, are needed to help physicians better consider security when using social media apps.

Acknowledgments

While preparing this manuscript, the authors used ChatGPT-4o (OpenAI) for language correction, editing, and formatting to improve the clarity of the language. The authors reviewed and revised the content as needed and take full responsibility for the accuracy and integrity of the paper.

Funding

The authors declared that no financial support was received for this work.

Data Availability

The datasets generated and analyzed during this study are not publicly available due to restrictions imposed by the ethics committee to protect participants' confidentiality.

Authors' Contributions

Conceptualization: ZZ, AM, NK, RT
Data curation: ZZ
Formal analysis: ZZ, TM
Investigation: ZZ
Methodology: ZZ, AM, RT
Supervision: RT
Visualization: ZZ
Writing – original draft: ZZ
Writing – review & editing: ZZ, RT, NK, AB

Conflicts of Interest

None declared.

Multimedia Appendix 1
The measurement and structural model assessments of the WhatsApp partial least squares structural equation modeling (PLS-SEM) analysis.
[DOCX File, 35 KB - mededu_v12i1e83475_app1.docx]

Multimedia Appendix 2
Free-text comments analysis.
[DOCX File, 24 KB - mededu_v12i1e83475_app2.docx]

Multimedia Appendix 3
Participants' characteristics.

XSL•FO
RenderX

[DOCX File, 15 KB - mededu_v12i1e83475_app3.docx]

Checklist 1
CHERRIES checklist.
[DOCX File, 24 KB - mededu_v12i1e83475_app4.docx]

References

1. Chen J, Wang Y. Social media use for health purposes: systematic review. J Med Internet Res 2021 May 12;23(5):e17917. [doi: 10.2196/17917] [Medline: 33978589]
2. Lee M, Bin Mahmood ABS, Lee ES, Smith HE, Tudor Car L. Smartphone and mobile app use among physicians in clinical practice: scoping review. JMIR Mhealth Uhealth 2023 Mar 31;11:e44765. [doi: 10.2196/44765] [Medline: 37000498]
3. Kraushaar J, Bohnet-Joschko S. Smartphone use and security challenges in hospitals: a survey among resident physicians in Germany. Int J Environ Res Public Health 2022 Dec 9;19(24):16546. [doi: 10.3390/ijerph192416546] [Medline: 36554426]
4. Cherrez-Ojeda I, Vanegas E, Felix M, et al. Frequency of use, perceptions and barriers of information and communication technologies among Latin American physicians: an Ecuadorian cross-sectional study. J Multidiscip Healthc 2020;13:259-269. [doi: 10.2147/JMDH.S246253] [Medline: 32214821]
5. de Oliveira Figueiredo MJ, Cuesta A, Duncker D, Boveda S, Guerra F, Márquez MF. Use of instant messaging in electrophysiological clinical practice in Latin America: a LAHRS survey. EP Europace 2023 Feb 8;25(1):59-64. [doi: 10.1093/europace/euac080] [Medline: 35727727]
6. Nikolic A, Wickramasinghe N, Claydon-Platt D, Balakrishnan V, Smart P. The use of communication apps by medical staff in the Australian health care system: survey study on prevalence and use. JMIR Med Inform 2018 Feb 9;6(1):e9. [doi: 10.2196/medinform.9526] [Medline: 29426813]
7. O'Sullivan DM, O'Sullivan E, O'Connor M, Lyons D, McManus J. WhatsApp doc? BMJ Innov 2017 Oct;3(4):238-239. [doi: 10.1136/bmjinnov-2017-000239]
8. Van Ravenswaay L, Parnes A, Nisly SA. Clicks for credit: an analysis of healthcare professionals' social media use and potential for continuing professional development activities. Med Educ Online 2024 Dec 31;29(1):2316489. [doi: 10.1080/10872981.2024.2316489] [Medline: 38359156]
9. Pronk SA, Gorter SL, van Luijk SJ, Barnhoorn PC, Binkhorst B, van Mook WNKA. Perception of social media behaviour among medical students, residents and medical specialists. Perspect Med Educ 2021 Aug;10(4):215-221. [doi: 10.1007/s40037-021-00660-1] [Medline: 33826108]
10. Loeb S, Carrick T, Frey C, Titus T. Increasing social media use in urology: 2017 American Urological Association Survey. Eur Urol Focus 2020 May 15;6(3):605-608. [doi: 10.1016/j.euf.2019.07.004] [Medline: 31351900]
11. Youssef Y, Gehlen T, Ansorg J, Back DA, Scherer J. Professional social media use among orthopedic and trauma surgeons in Germany: cross-sectional questionnaire-based study. JMIR Form Res 2024 Apr 19;8:e53336. [doi: 10.2196/53336] [Medline: 38639987]
12. Schmidt S, Darwich A, Leutheuser S, Krahl D, Navas L. The use of social media in orthopedic and trauma surgery education: a cross-sectional survey of German-speaking residents and medical students. Healthcare (Basel) 2024 Oct 10;12(20):2016. [doi: 10.3390/healthcare12202016] [Medline: 39451431]
13. Chan WS, Leung AY. Use of social network sites for communication among health professionals: systematic review. J Med Internet Res 2018 Mar 28;20(3):e117. [doi: 10.2196/jmir.8382] [Medline: 29592845]
14. Azer SA, AlShiha LZ, AlSkait GA, et al. Use of social media and professional attitude and beliefs of medical students and interns: should social media use be part of professional assessment? Medicine (Baltimore) 2025 Jul 11;104(28):e43114. [doi: 10.1097/MD.0000000000043114] [Medline: 40660555]
15. Le règlement général sur la protection des données - RGPD [Web Page in French]. Commission Nationale de l'Informatique et des Libertés (CNIL). 2016 May 24. URL: https://www.cnil.fr/fr/reglement-europeen-protection-donnees [accessed 2025-07-18]
16. Health Insurance Portability and Accountability Act of 1996. : Office of the Assistant Secretary for Planning and Evaluation (ASPE), U.S. Department of Health and Human Services (HHS); 1996 URL: http://aspe.hhs.gov/reports/health-insurance-portability-accountability-act-1996 [accessed 2025-07-18]
17. Masoni M, Guelfi MR. WhatsApp and other messaging apps in medicine: opportunities and risks. Intern Emerg Med 2020 Mar;15(2):171-173. [doi: 10.1007/s11739-020-02292-5] [Medline: 32062746]
18. Esmaeilzadeh P, Sambasivan M. Health information exchange (HIE): a literature review, assimilation pattern and a proposed classification for a new policy approach. J Biomed Inform 2016 Dec;64:74-86. [doi: 10.1016/j.jbi.2016.09.011] [Medline: 27645322]
19. Chen M, Esmaeilzadeh P. Adoption and use of various health information exchange methods for sending inside health information in US hospitals. Int J Med Inform 2023 Sep;177:105156. [doi: 10.1016/j.ijmedinf.2023.105156] [Medline: 37487455]
20. Mon espace santé [Web Page in French]. URL: https://www.monespacesante.fr/ [accessed 2025-07-23]

XSL•FO
RenderX

21. Ezra O, Toren A, Tadmor O, Katorza E. Secure instant messaging application in prenatal care. J Med Syst 2020 Feb 22;44(4):73. [doi: 10.1007/s10916-020-01552-3] [Medline: 32086609]
22. Shahar T, Tadmor O, Dior U, et al. The role of secure instant messaging applications in medical education: evaluating student satisfaction in a case-based learning program using Siilo. Front Med (Lausanne) 2023;10:1139859. [doi: 10.3389/fmed.2023.1139859] [Medline: 36936245]
23. Kuster C, Ruffenach L, Dissaux C, Bruant-Rodier C, Bodin F. Medical photographs: time saving and data security thanks to a dedicated application [Article in French]. Ann Chir Plast Esthet 2021 Jun;66(3):217-222. [doi: 10.1016/j.anplas.2020.12.004] [Medline: 33431220]
24. McGowan BS, Wasko M, Vartabedian BS, Miller RS, Freiherr DD, Abdolrasulnia M. Understanding the factors that influence the adoption and meaningful use of social media by physicians to share medical information. J Med Internet Res 2012 Sep 24;14(5):e117. [doi: 10.2196/jmir.2138] [Medline: 23006336]
25. De Benedictis A, Lettieri E, Masella C, et al. WhatsApp in hospital? An empirical investigation of individual and organizational determinants to use. PLOS ONE 2019;14(1):e0209873. [doi: 10.1371/journal.pone.0209873] [Medline: 30633754]
26. Gulacti U, Lok U, Hatipoglu S, Polat H. An analysis of WhatsApp usage for communication between consulting and emergency physicians. J Med Syst 2016 Jun;40(6):130. [doi: 10.1007/s10916-016-0483-8] [Medline: 27083574]
27. Shaarani I, El-Kantar A, Hamzeh N, et al. Interprofessional communication of physicians using WhatsApp: physicians' perspective. Telemed J E Health 2020 Oct;26(10):1257-1264. [doi: 10.1089/tmj.2019.0216] [Medline: 32083515]
28. Ganasegeran K, Renganathan P, Rashid A, Al-Dubai SAR. The m-Health revolution: exploring perceived benefits of WhatsApp use in clinical practice. Int J Med Inform 2017 Jan;97:145-151. [doi: 10.1016/j.ijmedinf.2016.10.013] [Medline: 27919374]
29. Youssef Y, Scherer J, Niemann M, Ansorg J, Back DA, Gehlen T. Social media use among orthopedic and trauma surgeons in Germany: cross-sectional survey study. JMIR Form Res 2023 Sep 22;7:e45665. [doi: 10.2196/45665] [Medline: 37738084]
30. Schneider AM, Tate J, Murphy MP, Hamid K, Light TR, Schiff AP. The characterization of social media in orthopaedic surgery: a survey study of 312 residents and applicants. JB JS Open Access 2022;7(2):e21.00159. [doi: 10.2106/JBJS.OA.21.00159] [Medline: 35506017]
31. McDonald AC, Maling A, Puttick M. Mobile data security: instant messaging in orthopaedics. ANZ J Surg 2022 Apr;92(4):825-829. [doi: 10.1111/ans.17550] [Medline: 35191179]
32. Bellam K, Yakkanti R, Amaya A, Qiu MF, Conte B, Aiyer A. Orthopedic surgery residencies: the leap to social media. Orthopedics 2023;46(5):e281-e286. [doi: 10.3928/01477447-20230310-07] [Medline: 36921225]
33. Bhandoria GP, Jayraj AS, Tiwari S, et al. Use of social media for academic and professional purposes by gynecologic oncologists. Int J Gynecol Cancer 2025 Jun;35(6):101845. [doi: 10.1136/ijgc-2024-005573] [Medline: 38955373]
34. Guerra F, Linz D, Garcia R, et al. Use and misuse of instant messaging in clinical data sharing: the EHRA-SMS survey. Europace 2021 Aug 6;23(8):1326-1330. [doi: 10.1093/europace/euab063] [Medline: 33709102]
35. Hazzam J, Lahrech A. Health care professionals' social media behavior and the underlying factors of social media adoption and use: quantitative study. J Med Internet Res 2018 Nov 7;20(11):e12035. [doi: 10.2196/12035] [Medline: 30404773]
36. Pedersen MRV, Precht H, Jensen J, et al. Radiographers use of social media - SoMe in a Nordic perspective. Radiography (Lond) 2024 Mar;30(2):651-658. [doi: 10.1016/j.radi.2024.01.015] [Medline: 38341986]
37. Chiang KF, Wang HH, Hsieh HF. Factors influencing the mobile instant messaging application usage behaviours among home healthcare providers. J Clin Nurs 2021 Dec;30(23-24):3577-3589. [doi: 10.1111/jocn.15868] [Medline: 34189775]
38. Davis FD. Perceived usefulness, perceived ease of use, and user acceptance of information technology. MIS Q 1989 Sep 1;13(3):319-340. [doi: 10.2307/249008]
39. Venkatesh V, Thong JYL, Xu X. Consumer acceptance and use of information technology: extending the unified theory of acceptance and use of technology. MIS Q 2012 Mar 1;36(1):157-178. [doi: 10.2307/41410412]
40. Eysenbach G. Improving the quality of web surveys: the Checklist for Reporting Results of Internet E-Surveys (CHERRIES). J Med Internet Res 2004 Sep 29;6(3):e34. [doi: 10.2196/jmir.6.3.e34] [Medline: 15471760]
41. Digital 2024: Global Overview Report. : DataReportal – Global Digital Insights; 2024 URL: https://wearesocial.com/wp-content/uploads/2024/02/Digital_2024_Global.pdf [accessed 2026-05-20]
42. Tamilmani K, Rana NP, Wamba SF, Dwivedi R. The extended Unified Theory of Acceptance and Use of Technology (UTAUT2): a systematic literature review and theory evaluation. Int J Inf Manage 2021 Apr;57:102269. [doi: 10.1016/j.ijinfomgt.2020.102269]
43. Owusu Kwateng K, Darko-Larbi O, Amanor K. A modified UTAUT2 for the study of telemedicine adoption. Int J Healthc Manag 2023 Apr 3;16(2):207-223. [doi: 10.1080/20479700.2022.2088068]
44. Schomakers EM, Lidynia C, Vervier LS, Calero Valdez A, Ziefle M. Applying an extended UTAUT2 model to explain user acceptance of lifestyle and therapy mobile health apps: survey study. JMIR Mhealth Uhealth 2022 Jan 18;10(1):e27095. [doi: 10.2196/27095] [Medline: 35040801]
45. Casaló L, Flavián C, Guinalíu M. The impact of participation in virtual brand communities on consumer trust and loyalty: the case of free software. Online Inf Rev 2007 Nov 27;31(6):775-792. [doi: 10.1108/14684520710841766]

XSL•FO
RenderX

46. Yee‐Loong Chong A, Ooi KB, Lin B, Tan BI. Online banking adoption: an empirical analysis. Int J Bank Mark 2010 Jun 15;28(4):267-287. [doi: 10.1108/02652321011054963]
47. Lebrument N, Zumbo-Lebrument C, Rochette C. L'acceptation des applications mobiles de MaaS: une application de l'UTAUT2 au contexte des villes intelligentes françaises. Systèmes Inf Manag 2022 May 2;Volume 26(4):19-54. [doi: 10.3917/sim.214.0019]
48. Pagé I, Roos M, Collin O, et al. UTAUT2-based questionnaire: cross-cultural adaptation to Canadian French. Disabil Rehabil 2023 Feb;45(4):709-716. [doi: 10.1080/09638288.2022.2037746] [Medline: 35179430]
49. Hair Jr JF, Hult GTM, Ringle CM, Sarstedt M. A Primer on Partial Least Squares Structural Equation Modeling (PLS-SEM), 2nd edition: SAGE; 2017.
50. Ray S, Danks N, Calero Valdez A. SEMinR: domain-specific language for building, estimating, and visualizing structural equation models in R. SSRN Electron J 2021. [doi: 10.2139/ssrn.3900621]
51. Hair Jr JF, Hult GTM, Ringle CM, Sarstedt M, Danks NP, Ray S. Partial Least Squares Structural Equation Modeling (PLS-SEM) Using R: A Workbook: Springer International Publishing; 2021. [doi: 10.1007/978-3-030-80519-7]
52. Barclay D, Thompson R, Higgins C. The partial least squares (PLS) approach to causal modeling: personal computer adoption and use as an illustration. Technol Stud 1995;2:285-309 [FREE Full text]
53. Faul F, Erdfelder E, Buchner A, Lang AG. Statistical power analyses using G*Power 3.1: tests for correlation and regression analyses. Behav Res Methods 2009 Nov;41(4):1149-1160. [doi: 10.3758/BRM.41.4.1149] [Medline: 19897823]
54. Kock N, Hadaya P. Minimum sample size estimation in PLS‐SEM: the inverse square root and gamma‐exponential methods. Inf Syst J 2018 Jan;28(1):227-261. [doi: 10.1111/isj.12131]
55. Hair Jr JF, Hult GTM, Ringle CM, Sarstedt M. Partial Least Squares Structural Equation Modeling (PLS-SEM) Using R: A Workbook, 3rd edition: Springer Publications; 2022. [doi: 10.1007/978-3-030-80519-7]
56. Venkatesh V, Davis FD, Zhu Y. Competing roles of intention and habit in predicting behavior: a comprehensive literature review, synthesis, and longitudinal field study. Int J Inf Manage 2023 Aug;71:102644. [doi: 10.1016/j.ijinfomgt.2023.102644]
57. Nguyen DM, Chiu YTH, Vu HQ, Nguyen T. Toward a unified theory of technology continuance: effects of conscious versus unconscious factors. J Consum Behav 2022 Sep;21(5):1219-1235. [doi: 10.1002/cb.2073]
58. Ray S, Seo D. The interplay of conscious and automatic mechanisms in the context of routine use: an integrative and comparative study of contrasting mechanisms. Inf Manage 2013 Nov;50(7):523-539. [doi: 10.1016/j.im.2013.07.008]
59. Manns A, Pezziardi T, Kadlub N, Burgun A, Destrez A, Tsopra R. Enhancing security in patient medical information exchange: a qualitative study. Int J Med Inform 2025 May;197:105841. [doi: 10.1016/j.ijmedinf.2025.105841] [Medline: 39986124]
60. Kushniruk AW, Borycki EM. Human factors in healthcare IT: management considerations and trends. Healthc Manage Forum 2023 Mar;36(2):72-78. [doi: 10.1177/08404704221139219] [Medline: 36847593]
61. Kushniruk A, Kaufman D. Human factors and organizational issues in health informatics: review of recent developments and advances. Yearb Med Inform 2024 Aug;33(1):196-209. [doi: 10.1055/s-0044-1800744] [Medline: 40199306]
62. Ehrenfeld J. Physician leadership in the new era of AI and digital health tools. Intell-Based Med 2023;8:100109. [doi: 10.1016/j.ibmed.2023.100109]
63. Pears M, Konstantinidis S. Cybersecurity training in the healthcare workforce – utilization of the ADDIE model. Presented at: 2021 IEEE Global Engineering Education Conference (EDUCON); Apr 21-23, 2021. [doi: 10.1109/EDUCON46332.2021.9454062]
64. Alhuwail D, Al-Jafar E, Abdulsalam Y, AlDuaij S. Information security awareness and behaviors of health care professionals at public health care facilities. Appl Clin Inform 2021 Aug;12(4):924-932. [doi: 10.1055/s-0041-1735527] [Medline: 34587638]
65. Arain MA, Tarraf R, Ahmad A. Assessing staff awareness and effectiveness of educational training on IT security and privacy in a large healthcare organization. J Multidiscip Healthc 2019;12:73-81. [doi: 10.2147/JMDH.S183275] [Medline: 30666123]
66. Glueckauf RL, Maheu MM, Drude KP, et al. Survey of psychologists' telebehavioral health practices: technology use, ethical issues, and training needs. Prof Psychol Res Pract 2018;49(3):205-219. [doi: 10.1037/pro0000188]
67. Myers J, Frieden TR, Bherwani KM, Henning KJ. Ethics in public health research: privacy and public health at risk: public health confidentiality in the digital age. Am J Public Health 2008 May;98(5):793-801. [doi: 10.2105/AJPH.2006.107706] [Medline: 18382010]
68. Cain J, Romanelli F. E-professionalism: a new paradigm for a digital age. Curr Pharm Teach Learn 2009 Dec;1(2):66-70. [doi: 10.1016/j.cptl.2009.10.001]
69. O'Connor S, Zhang M, Honey M, Lee JJ. Digital professionalism on social media: a narrative review of the medical, nursing, and allied health education literature. Int J Med Inform 2021 Sep;153:104514. [doi: 10.1016/j.ijmedinf.2021.104514] [Medline: 34139621]
70. Gormley M, Collins L, Schofield S, Neville P. Exploring the impact of digital professionalism awareness training on dental undergraduate students. Eur J Dent Educ 2021 May;25(2):271-281. [doi: 10.1111/eje.12601] [Medline: 32949078]

XSL•FO
RenderX

71. Tsopra R, Sedki K, Courtine M, et al. Helping GPs to extrapolate guideline recommendations to patients for whom there are no explicit recommendations, through the visualization of drug properties. The example of AntibioHelp® in bacterial diseases. J Am Med Inform Assoc 2019 Oct 1;26(10):1010-1019. [doi: 10.1093/jamia/ocz057] [Medline: 31077275]
72. Schmidt J, Gambashidze N, Manser T, et al. Does interprofessional team-training affect nurses' and physicians' perceptions of safety culture and communication practices? Results of a pre-post survey study. BMC Health Serv Res 2021 Apr 14;21(1):341. [doi: 10.1186/s12913-021-06137-5] [Medline: 33853593]
73. Byrd 4th TF 4th, Speigel PS, Cameron KA, O'Leary KJ. Barriers to adoption of a secure text messaging system: a qualitative study of practicing clinicians. J Gen Intern Med 2023 Apr;38(5):1224-1231. [doi: 10.1007/s11606-022-07912-8] [Medline: 36376637]
74. Tsega S, Kalra A, Sevilla CT, Cho HJ. A bottom-up approach to encouraging sustained user adoption of a secure text messaging application. Appl Clin Inform 2019 Mar;10(2):326-330. [doi: 10.1055/s-0039-1688554] [Medline: 31091544]
75. Sevick LK, Santana MJ, Ghali WA, Clement F. Prospective economic evaluation of an electronic discharge communication tool: analysis of a randomised controlled trial. BMJ Open 2017 Dec 14;7(12):e019139. [doi: 10.1136/bmjopen-2017-019139] [Medline: 29247110]
76. Olsen R. Self-selection bias. In: Encyclopedia of Survey Research Methods: Sage Publications, Inc; 2008:809-809. [doi: 10.4135/9781412963947.n526]
77. Grimm P. Social desirability bias. In: Wiley International Encyclopedia of Marketing, 1st edition: John Wiley & Sons, Ltd.; 2010. [doi: 10.1002/9781444316568]
78. Coughlin SS. Recall bias in epidemiologic studies. J Clin Epidemiol 1990;43(1):87-91. [doi: 10.1016/0895-4356(90)90060-3] [Medline: 2319285]

Abbreviations

CHERRIES: Checklist for Reporting Results of Internet E-Surveys
PLS-SEM: partial least squares structural equation modeling
UTAUT2: Unified Theory of Acceptance and Use of Technology 2

Edited by A Stone; submitted 03.Sep.2025; peer-reviewed by S Ganesh; revised version received 17.Apr.2026; accepted 17.Apr.2026; published 05.Jun.2026.

Please cite as:
Zouakia Z, Manns A, Meyre T, Burgun A, Kadlub N, Tsopra R
Adoption and Use of Social Media in Health Care Among Medical Residents: Cross-Sectional Study
JMIR Med Educ 2026;12:e83475
URL: https://mededu.jmir.org/2026/1/e83475
doi:10.2196/83475

© Zineb Zouakia, Aurélia Manns, Thomas Meyre, Anita Burgun, Natacha Kadlub, Rosy Tsopra. Originally published in JMIR Medical Education (https://mededu.jmir.org), 5.Jun.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Education, is properly cited. The complete bibliographic information, a link to the original publication on https://mededu.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Instagram Video Engagement in Medical Education: Cross-Sectional Study

Aapeli Kemppainen[1], MD; Henrik Nuutinen[2], MD, PhD; Virve Koljonen[1], MD, PhD; Petri Kulmala[3,4], MD, PhD

[1]Department of Plastic Surgery, Helsinki University Hospital, University of Helsinki, Fabianinkatu 33, Helsinki, Finland

[2]Department of Plastic Surgery, Kuopio University Hospital, University of Eastern Finland, Kuopio, Finland

[3]Education Development and Service Unit, Faculty of Medicine, University of Oulu, Oulu, Finland

[4]Medical Research Center, Oulu University Hospital, Oulu, Finland

Corresponding Author:
Aapeli Kemppainen, MD
Department of Plastic Surgery, Helsinki University Hospital, University of Helsinki, Fabianinkatu 33, Helsinki, Finland

Abstract

Background: Social media (SM) has become an essential tool for delivering formal and informal educational content, including medical education. Instagram-based educational initiatives have shown strong engagement and improved learning outcomes in accounts created for research purposes. However, empirical evidence on engagement with real-life account content remains limited.

Objective: This study aims to identify engaging types of posts, determine the optimal video length for medical education, and provide evidence-based recommendations to medical educators on SM content.

Methods: A cross-sectional study analyzed real-life Instagram medical education account data from Instagram Insights, with a focus on video posts. Insights' post-related data—reached accounts, views, and watch time—were used. Video posts were categorized based on their type (theoretical, clinical, health promotional, or entertaining) and their implementation (use of animation, human models, or medical models). Video length was stratified into 3 categories: ≤60 seconds, 61-120 seconds, and >120 seconds. The Kruskal-Wallis test was used to determine whether video type, implementation, or length were associated with reach, views, or watch time.

Results: From May 26, 2020, to May 3, 2024, 125 video posts were analyzed. The median of reached accounts was 5317 (IQR 4007-6244) per video. The median (IQR) of views was 6533 (4708-8601) per video, and relative watch time was 19% (15%-30%). Health promotion and awareness videos had the greatest reach (median 7070, IQR 6093-11,340) compared to theory (median 5265, IQR 4476-5829; *P*=.006), clinical skills (median 4597, IQR 2001-5688; *P*=.003), and entertainment videos (median 5605, IQR 3887-6492; *P*=.04). Videos >120 seconds had higher reach (median 5774, IQR 4754-6636) and views (median 7455, IQR 5321-11,031) compared to videos that were 61-120 seconds (median reach 4514, IQR 2551-5615, *P*=.02; views 5987, IQR 3727-6970, *P*=.02). Similarly, videos that were ≤60 seconds had higher reach (median 5484, IQR 4299-6737) and views (median 7189, IQR 5182-10,609) than videos that were 61-120 seconds (median reach 4514, IQR 2551-5615, *P*=.02; views 5987, IQR 3727-6970, *P*=.02). However, videos that were ≤60 seconds had significantly higher relative watch time (median 29%, IQR 20%-40%) compared to videos that were 61-120 seconds (median 17%, IQR 14%-20%; *P*=.003) and videos that were >120 seconds (median 15%, IQR 11%-17%; *P*=.002). The implementation method did not significantly affect video engagement.

Conclusions: Unlike previous studies that have primarily explored the theoretical potential of Instagram or analyzed isolated posts, this study provides a longitudinal real-world analysis of 125 video posts published on an active medical education Instagram account. Our findings demonstrate that shorter video lengths and health promotion content are associated with greater engagement, while diverse implementation types may further support content effectiveness. By examining authentic educational content in SM, this study offers practical, data-driven insights for medical educators seeking to optimize their educational strategies. Future research should explore how different demographic groups engage, thereby enabling more targeted and effective medical education approaches.

(JMIR Med Educ 2026;12:e79124) doi:10.2196/79124

KEYWORDS

medical education; social media; Instagram; educational videos; learner engagement

XSL•FO
RenderX

Introduction

Since its launch in 2010, the image-sharing service Instagram has become one of the most popular social networks, with over 1 billion monthly active users [1]. Communication on Instagram relies on short videos, which are suitable for the current fast-paced lifestyle. Social media (SM) platforms have evolved into essential tools for the delivery of both formal and informal educational content [2]. A recent study revealed that health science students spend over 3 hours daily on SM, with 37% of this time dedicated to Instagram [3]. Additionally, a recent survey found that up to 90% of medical and health science students use SM for educational purposes [4].

The use of videos in medical education dates back to the 1960s [5]. Videos are an effective way of communicating knowledge, particularly in medical education [5,6]. Previous studies have demonstrated the effectiveness of videos as educational tools [7,8]. A systematic review found that adding video materials to existing higher education teaching methods significantly improved learning outcomes [9]. For example, compared to standard textbooks, watching a short instructional video before splinting led to faster learning and more successful procedural skills [10]. Video-based education has been shown to enhance knowledge acquisition in surgical training compared to traditional teaching methods [11].

Multimedia learning uses text, narration, pictures, or videos, engaging all senses and minimizing any extra words or pictures, and is suitable even for complex topics [12]. Microlearning involves acquiring knowledge or skills in smaller units [13]. Keeping with the multimedia learning principles and microlearning, 2 comparable eLearning programs on type 2 diabetes treatment were converted into media clips that were 30 to 180 seconds and distributed via SM [14]. The transition from eLearning programs to short SM clips increased participation in the programs by up to 100-fold compared to traditional eLearning programs [14]. An online application was created for emergency medicine guidelines covering 7 topics, for example, chest pain and headache, and was used during shifts by medical students [15]. With the application, they significantly enhanced the frequency and quality of emergency medicine education for undergraduate students in the emergency department [15]. From another standpoint, a family medicine program produced SM videos on evidence-based medicine topics to teach family medicine residents, resulting in significantly improved learner engagement and knowledge acquisition [16]. Although the principle of the cognitive theory of multimedia learning was established already in 2001 [17], it remains applicable to the current medical student Generation Z, as they favor visual and technology-enhanced learning methods [18].

As SM, and especially Instagram, is a widely used tool to share visual material, its educational potential has been recognized [4,19,20]. The use of SM in medical education [21] is a unique combination of open educational resources (OER); social networking; and with regard to videos, multimedia learning. Medical education on SM may use OER and enable just-in-time (JIT) learning to provide resources as needs arise rather than as prescheduled lessons [22]. Aligning with JIT's on-demand accessibility, SM is well suited for educational purposes via smart devices, either to learn new materials or to revisit what has been learned before. Although OER in medical education is recommended, there are still several obstacles to its widespread use. Such barriers include copyright issues and, especially in clinical contexts, concerns over the use of patient data, photographs, and videos.

Instagram has been instrumental in surgical education, facilitating the sharing of procedural videos and resources to a broad audience while fostering collaboration among students and professionals [23]. Pathologists, in particular, have been active in using Instagram and static images. In previous studies, Instagram accounts developed for histology education demonstrated high usability (98% of respondents), increased learner confidence (95%), and were associated with improved final examination performance [20,24]. Furthermore, an Instagram-based dermatopathology course outperformed a more traditional Canvas platform–based course among a group of 40 dermatology residents [19]. These examples underscore Instagram's potential to complement traditional learning methods.

SM has been described as promoting social isolation [25,26]. However, SM may also create a sense of communal belonging [27,28]. Interacting with other account followers and content creators by liking, asking, and answering questions in the comment section or through direct messaging is nowadays a new way of expressing sociability. Connectivism, a learning theory, explains learning and teaching in networks using modern technology [29]. It provides insights into using new technology to deliver knowledge, such as diversity, autonomy, openness, and connectivity [30], all of which seem to be fulfilled when using SM in medical education.

Systematic reviews have highlighted the benefits of integrating SM into medical education [31-33]. Despite the increase in research on SM as an educational resource, evidence remains limited regarding the types of content that maximize engagement on platforms like Instagram [34]. Previously, in studies on engagement, accounts have been created solely for study purposes [16,19,20,24]. There is a paucity of research on the engagement of educational SM posts in real-life medical education accounts.

Therefore, we analyzed engagement metrics for medical education posts from a real-world educational SM account using a cross-sectional design. We also aimed to identify the most engaging types of videos by analyzing several features, thereby providing real-life recommendations for effective medical education content on SM.

Methods

Ethical Considerations

The study protocol was reviewed and approved by the University of Helsinki Ethical Review Board in Humanities and Social and Behavioral Sciences (statement 65/2024). The study was conducted as a secondary analysis of aggregated SM analytics

XSL•FO
RenderX

data and did not involve direct interaction with human participants or the collection of identifiable personal data.

The data were obtained from Instagram Insights, which provides anonymized and aggregated post-level metrics, including reach, views, comments, and average watch time, as well as aggregated follower demographics, such as age group and gender distribution. Instagram Insight data have been used for scientific purposes in previous studies [20,35]. No individual-level data were accessible to the researchers, and individual users could not be identified from the dataset. Since the study analyzed, anonymized, and aggregated data without participant recruitment or intervention, individual informed consent was not collected. No compensation was provided.

Some video posts analyzed in this study included individuals appearing in publicly available content on the account. All individuals who appeared in the videos had provided permission for the publication of the content on the SM platform. No additional identifying information is included in the manuscript or supplementary materials, and no individuals can be identified from the reported data.

Instagram Account

Plastiikkaope (in English, *Plasticsteacher*) is a nonprofit Instagram account created and maintained by one of the authors (VK) to share surgical educational material through short videos and pictures. Targeting medical students, its content focuses on surgery. Posts adhere to multimedia learning principles, with written explanations, narratives, and key points included either in the videos or in the comments section. As a public account, it is accessible to everyone aged 16 years and older. The *Plasticsteacher* account was created in May 2020, with the first post published on May 26, 2020.

Data

Data were collected on July 9, 2024, by 1 author (AK) using Instagram's analytics tool, Instagram Insights. All posts published before data collection were included in the initial analysis. The first post was published on May 26, 2020, and the latest on May 3, 2024. The sampling of the posts was not conducted.

To study the popularity and further engagement of the posts, we collected reach, views, and watch time as post-related parameters from Insights. Reach refers to the number of unique accounts that viewed a post at least once after its release, while views indicate the number of times a video starts playing. Starting from September 27, 2023, updates in the Insights software enabled the separate analysis of initial plays and replays. Insights also provided watch time metrics such as total watch time and average watch time from September 2022 onwards. A relative watch time was calculated by dividing the average watch time by the length of a video. Not all post-related parameters were available for each post, not only because of updates on the Insights analytics application but also for technical reasons. For carousel posts, which contain multiple videos or pictures in 1 post, Insights provided reach, but views or watch time metrics were not provided. From September 2023 onwards, Instagram Insights divided views into initial plays and replays.

In addition to post-related parameters, we collected the account's follower metrics, including the followers' age range, gender, and geographic locations. The follower data were visible for only the previous 90 days in Insights, meaning that the follower data represented the demographics of followers between April 10, 2024, and July 9, 2024, limiting the analysis of follower development.

Collaborative posts with other nonprofit accounts and posts with pictures only were excluded from the analysis. Included posts were stratified into educational and noneducational (ie, entertainment) groups and further divided into subgroups by authors. Educational posts were further divided into three content categories: (1) theory, (2) clinical skills, and (3) health promotion and awareness. Educational posts were also classified into three video implementation types: (1) animated videos, (2) videos using medical models (phantoms), and (3) videos using human participants. All individuals appearing in the videos were healthy volunteers who gave their consent to be featured. Typically, these individuals included fellow doctors, nurses, and relatives of author VK. Videos were categorized into 3 groups based on their length: ≤60 seconds, 61-120 seconds, and >120 seconds.

The analysis focused on video posts. First, we examined the distribution of different videos over the period investigated. On the basis of these findings, videos were selected for a second analysis, in which statistical differences were tested between video categories, implementation types, and length groups under 3 parameters: reached accounts, video views, and relative watch time.

Statistical Analysis

Statistical analysis was conducted using IBM SPSS Statistics version 29.0.2.0. Data were tested for skewed distribution by visual inspection of histograms. To assess statistical significance between subgroups based on length, video content, and video implementation, the independent-samples Kruskal-Wallis test was applied, followed by the Dunn pairwise comparisons test. Benjamini-Hochberg post-hoc corrections were applied in pairwise comparisons since the nature of the study was exploratory, and multiple subgroups were tested. Post hoc corrections were calculated in Microsoft Excel version 16.103.3 based on pairwise *P* values. For *P* values reported by SPSS as $P<.001$, a value of .001 was used for multiple-comparison correction, resulting in conservative estimates. A *P* value of ≤.05 was considered statistically significant. For missing data, the Little missing completely at random (MCAR) test was performed.

Results

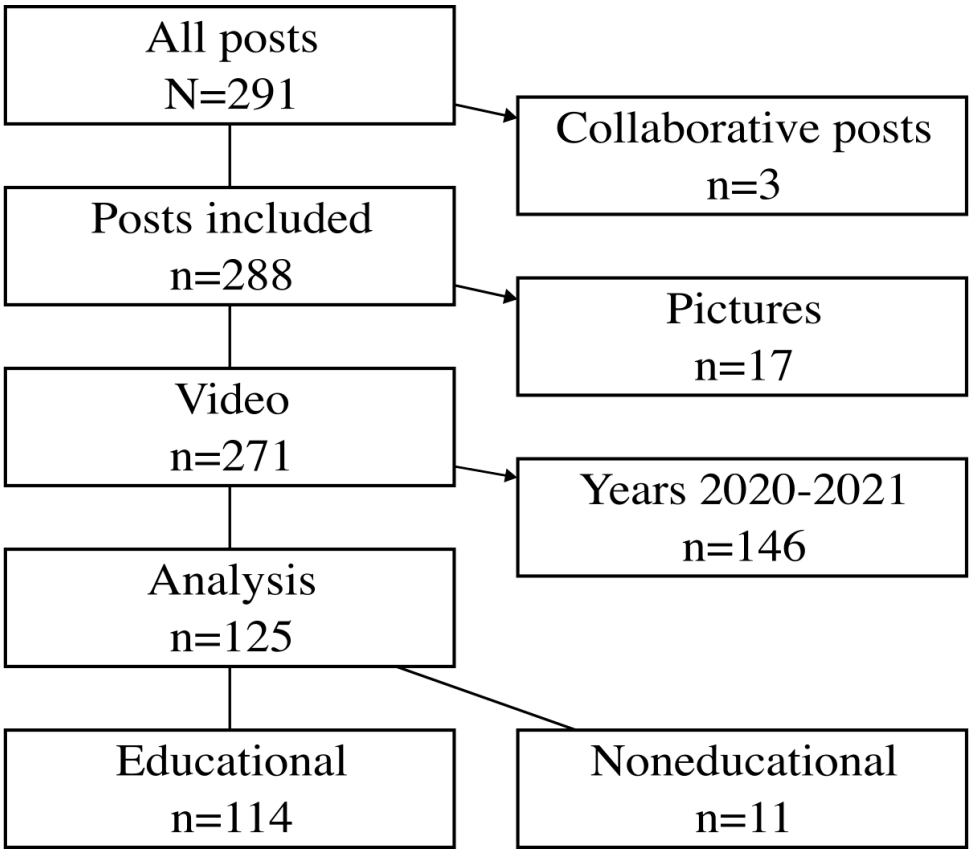
Overview

From May 26, 2020, to May 3, 2024, 291 posts were identified. Of these, 125 posts were included in the final analysis (Figure 1). Videos considered as health promotion and awareness and lasting ≤60 seconds were not published until 2022. The median values were lower for reached accounts (551, IQR 433-763 vs 5317, IQR 4010-6205) and views (1695, IQR 1415-2128 vs 6533, IQR 4756-8573) for videos posted between 2020 and

XSL•FO
RenderX

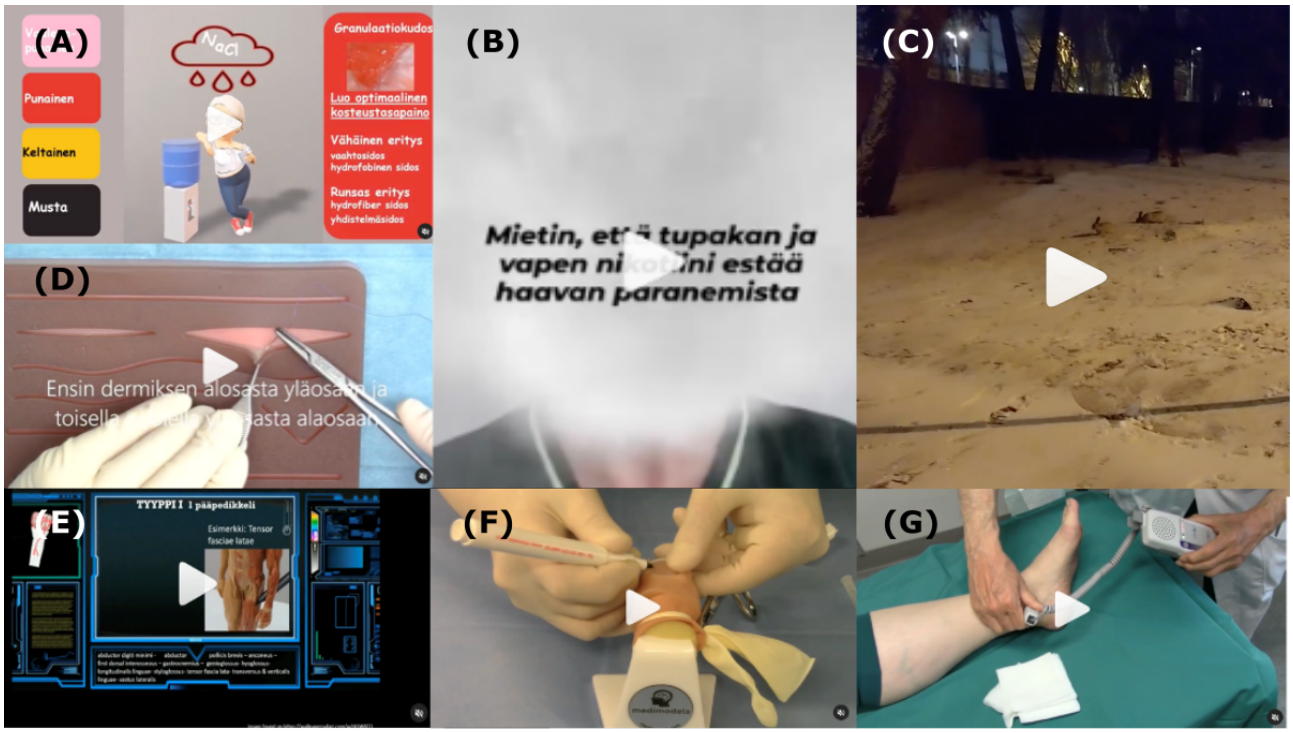
2021 compared to videos posted between 2022 and 2024. On the basis of these findings, the videos posted between 2020 and 2021 were excluded to avoid underestimating the median reach and views of other subgroups, including theory, clinical skills, entertainment, and videos lasting 61 - 120 and >120 seconds. The MCAR test was performed for all videos included in the final analysis (N=125), and missing data were not replaced. The percentage of missing data was low for reached accounts (0%) and views (2.4%), while the percentage was moderate for watch time (24%) and high for initial or replay views (76.8%), which is explained by developmental updates on Instagram Insights. Figure 2 illustrates categories under video content and video implementation type. A detailed description of the exclusion criteria for 2020 - 2021 posts and MCAR testing is provided in Multimedia Appendix 1.

Figure 1. Flowchart of study analysis. Collaborative posts, pictures, and posts during 2020 - 2021 were excluded from the final analysis. A more detailed description of the last phase of exclusion is shown in Multimedia Appendix 1.



XSL•FO
RenderX

Figure 2. Examples of different video categories and implementation types. (A) Theory wound bed classification using animation, (B) health promotion and awareness using a filter on how nicotine in tobacco and vape inhibits wound healing, (C) an entertainment video about rabbits from a nearby hospital, (D) clinical skills showing how to make an inverted dermal suture, (E) animation presenting classification of flaps, (F) using a medical model to demonstrate an ingrown toenail operation, and (G) using a healthy volunteer (author VK) to demonstrate examining pulse status in the lower extremity.



Follower Metrics

As of June 9, 2024, the account had 7538 followers, 83% of whom were females. Most followers were from Finland (98%). The largest age group was 25 - 34 years (44%), followed by 35 - 44 years (26%) and 45 - 54 years (13%). Among female followers, 41% were aged 25 - 34 years, while 60% of male followers were in this age group. Follower metrics were available in Instagram Insights for 90 days prior to the data collection date, so the development of follower demographics could not be analyzed.

Posts

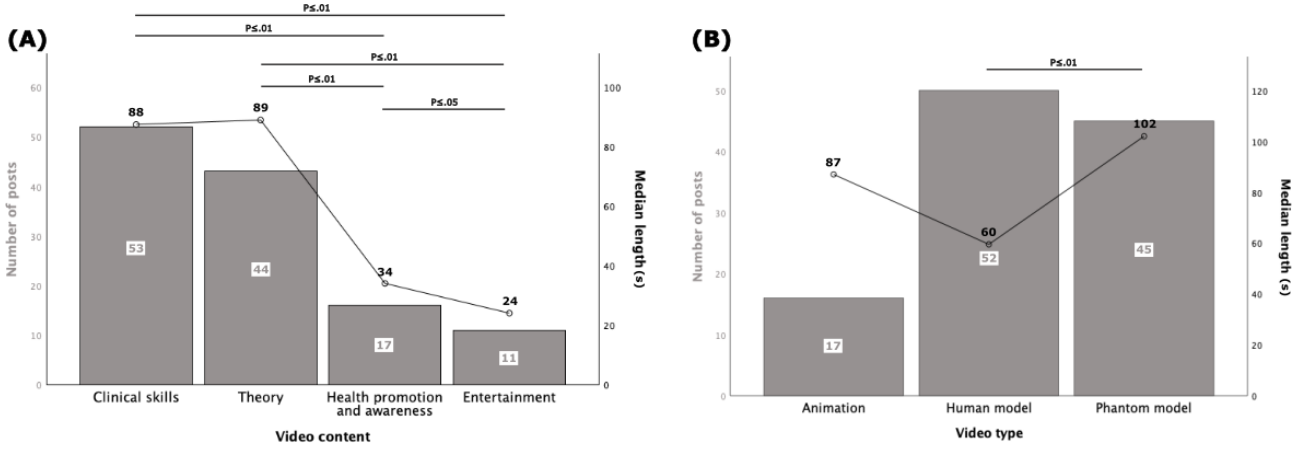
Of the 125 videos included, most (n=53, 42%) were clinical skills videos (Figure 3A). Among educational videos (n=114, 91%), human models were used most frequently (n=52, 46%; Figure 3B). Regarding video length, 40% (n=51) were ≤60 seconds, 41% (n=51) were 61 - 120 seconds, and 19% (n=23) were >120 seconds.

The median (IQR) video length of the analyzed videos (N=125) was 72 (42-109) seconds. The longest median (IQR) length was recorded as 89 (63-110) seconds for theory videos (n=43). Theory and clinical skills videos were significantly longer than health promotion and awareness videos and entertainment videos in pairwise comparisons (theory vs health promotion and awareness, *P*=.006; theory vs entertainment, *P*=.003; clinical skills vs health promotion and awareness, *P*=.002; clinical skills vs entertainment, *P*=.002; Figure 3A). Health promotion and awareness videos were longer than entertainment videos (*P*=.04). There was no statistically significant pairwise difference between clinical skills and theory videos.

By video implementation types, the longest median (IQR) video lengths were 102 (72-185) seconds for phantom model videos (n=45). Phantom model videos were significantly longer than human model videos (*P*=.003; Figure 3B). There was no statistically significant difference between other groups in pairwise comparisons.

XSL•FO
RenderX

Figure 3. Video classification based on video content, implementation type, and comparisons between video lengths of the categories. (A) Columns refer to the number of posts under different subgroups. Of the included 125 videos, 42% (n=53) were clinical skills videos, 35% (n=44) were theory videos, 14% (n=17) were health promotion videos, and 9% (n=11) were entertainment videos. Data points refer to the median video length and are shown with a connecting line for visualization purposes only. Median (IQR) video lengths were 89 (63-110) seconds for theory videos (n=43), 88 (60-138) seconds for clinical skills videos (n=52), 34 (23-53) seconds for health promotion and awareness videos (n=16), and 24 (16-29) seconds for entertainment videos (n=11). Theory and clinical skills videos were significantly longer than health promotion and awareness videos and entertainment videos pairwise. Health promotion and awareness videos were significantly longer than entertainment videos. There was no statistically significant pairwise difference between theory and clinical skills videos. (B) Among educational videos (n=114), human models were used in 46% (n=52), phantom models in 39% (n=45), and animations in 15% (n=17). Median (IQR) video lengths were 102 (72-185) seconds for phantom model videos (n=45), 87 (54-106) seconds for animated videos (n=16), and 60 (35-91) seconds for human model videos (n=50). Phantom model videos were significantly longer than human model videos. There was no statistically significant difference between other groups pairwise. Exact *P* values are listed in Multimedia Appendix 1.



Relative Watch Time

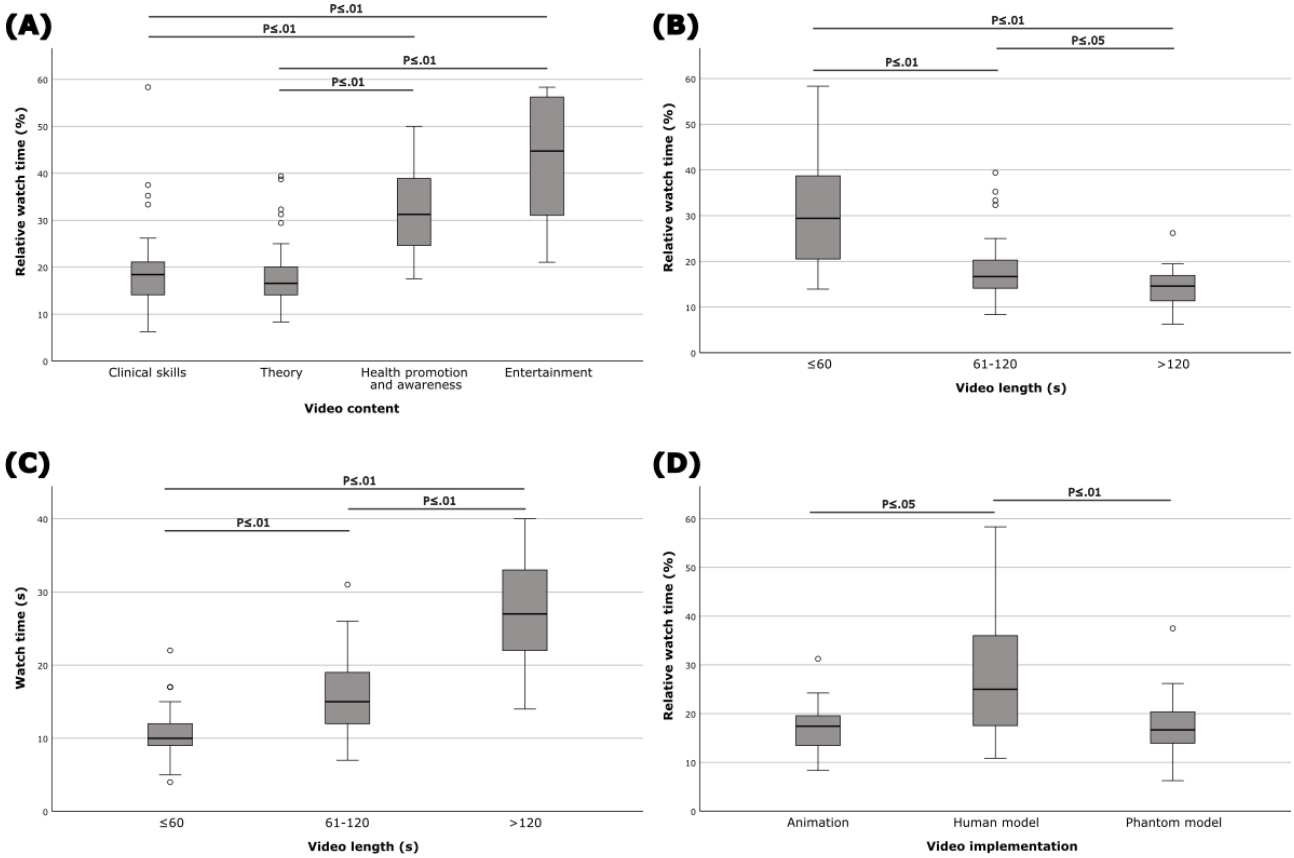
The median relative watch time for the 95 (76%) videos with available watch time data was 19% (IQR 15%-30%). Insights provided total watch time and average watch time from September 2022 onwards; therefore, data were not available for posts published earlier in 2022 (30 video posts). The highest median (IQR) relative watch time was 45% (26%-56%) for entertainment videos (n=9). Entertainment and health promotion and awareness videos were watched significantly further relative to their length compared to theory and clinical skills videos (entertainment vs theory, *P*=.003; entertainment vs clinical skills, *P*=.002; health promotion and awareness vs theory, *P*=.006; health promotion and awareness vs clinical skills, *P*=.002; Figure 4A).

When analyzed by video length, the highest median (IQR) relative watch time was 29% (20% - 40%) for videos 60 seconds or less (n=41). The shortest videos (≤60 seconds) were watched significantly more compared to other length groups (≤60 seconds vs 61-120 seconds, *P*=.003; ≤60 seconds vs >120 seconds, *P*=.002), followed by videos that were 61 to 120 seconds that were watched significantly more than videos that were more than 120 seconds (*P*=.04; Figure 4B). To complete the previous finding, an average watch time of the video length groups was analyzed. From the analysis, videos that were more than 120 seconds had statistically significantly the longest watch time, and videos that were 60 seconds or less had the shortest watch time (>120 seconds vs 61-120 seconds, *P*=.001; >120 seconds vs ≤60 seconds, *P*=.002; 61-120 seconds vs ≤60 seconds, *P*=.003; Figure 4C), meaning the longer the video, the longer the absolute watch time on average.

By video implementation type, the highest median (IQR) relative watch time was 25% (17% - 36%) for human model videos (n=37). Human model videos had statistically significantly higher relative watch time compared to other video types (human model vs phantom model, *P*=.003; human model vs animation, *P*=.01; Figure 4D).

XSL•FO
RenderX

Figure 4. Statistical analysis of relative watch time for different video content categories, video length groups, and video implementation types. (A) Health promotion and awareness videos and entertainment videos were significantly more watched than theory and clinical skills videos when compared pairwise, but there was no statistically significant difference between health promotion and awareness videos and entertainment or between theory and clinical skills videos. (B) The shortest length videos (≤60 s) were watched significantly more compared to other video length groups, followed by videos that were 61 - 120 seconds and videos that were >120 seconds; the difference was statistically significant between all groups when compared pairwise. (C) Videos longer than 120 seconds had the significantly longest watch time, followed by videos that were 61 - 120 seconds and videos ≤60 seconds; the difference was statistically significant between all groups when compared pairwise. (D) Human model videos had significantly higher relative watch time compared to animated and phantom model videos. Exact *P* values are listed in Multimedia Appendix 1.



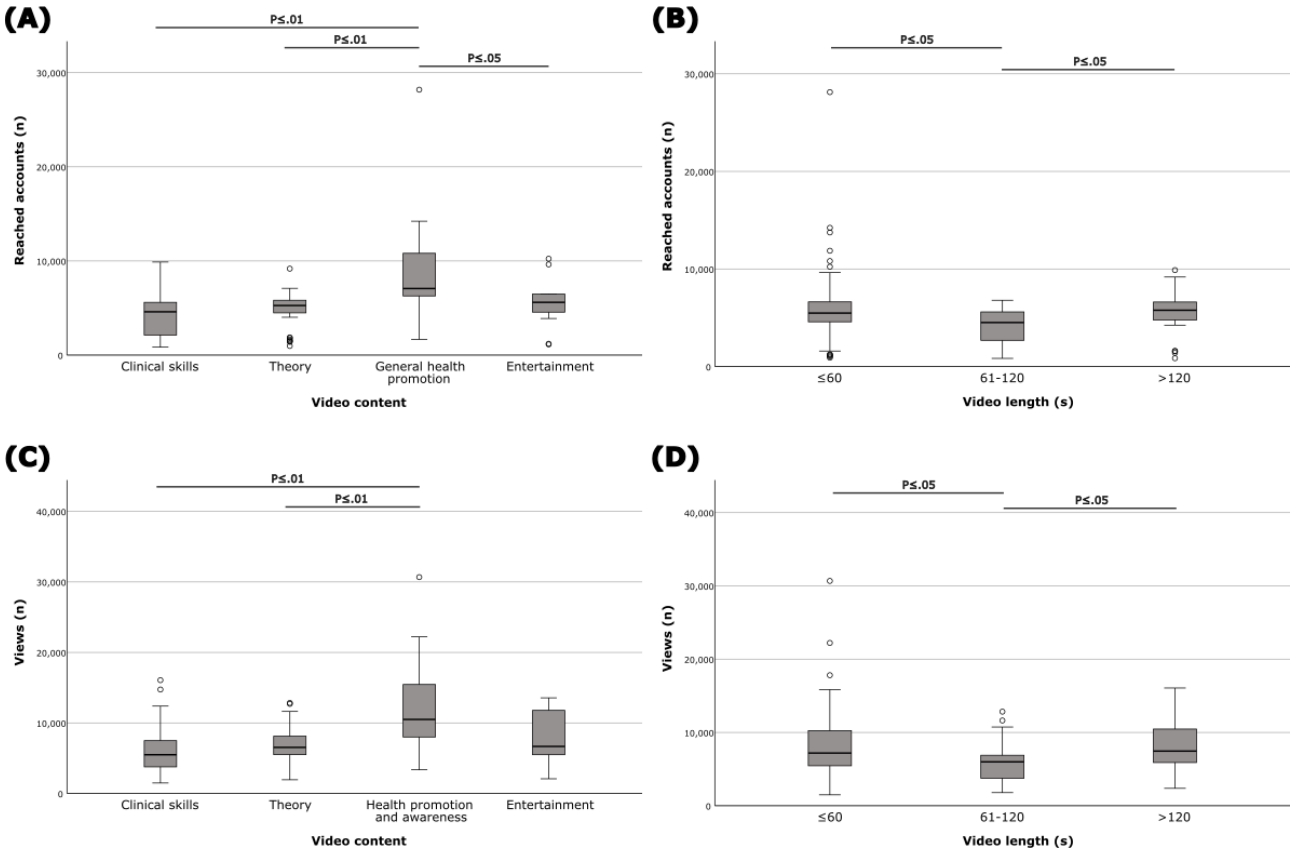
Reach

The reach demonstrates the number of unique users who saw the post, highlighting the audience size. The median (IQR) reach was 5317 (4007-6244) unique accounts per video for 125 analyzed videos. The most popular videos were health promotion and awareness videos (n=17), with a median (IQR) of 7070 (6093-11,340) reached accounts. Health promotion and awareness videos had a significantly higher reach compared to all other categories, while no statistically significant differences were observed between theory, clinical skills, and entertainment videos (health promotion and awareness vs clinical skills, *P*=.003; health promotion and awareness vs theory, *P*=.006; health promotion and awareness vs entertainment, *P*=.04; Figure 5A).

When analyzed by video length, the most popular were videos that were more than 120 seconds (n=23) with a median (IQR) of 5774 (4754-6636) reached accounts. Videos with the length of ≤60 seconds and >120 seconds had significantly higher reach compared to videos 61 - 120 seconds in length, but no statistically significant differences were found between the longest and shortest videos (≤60 seconds vs 61-120 seconds, *P*=.02; >120 seconds vs 61-120 seconds, *P*=.02; Figure 5B). Comparing video implementation types (animation, human, and phantom), there was no statistically significant difference.

XSL•FO
RenderX

Figure 5. Statistical analysis of reached accounts and views for different video content categories and video length groups. (A) Health promotion and awareness videos reached more accounts compared to all other categories, with a statistically significant difference. There were no significant differences between the other categories. (B) Videos ≤60 seconds and >120 seconds reached significantly more accounts compared to videos 61 - 120 seconds in length, but there was no significant difference between videos that were ≤60 seconds and videos that were >120 seconds. (C) Health promotion and awareness videos had significantly more views compared to theory and clinical skills videos. There were no statistically significant differences between the other categories. (D) Videos ≤60 seconds and >120 seconds were significantly more viewed compared to videos 61 - 120 seconds in length, but there were no significant differences between videos that were ≤60 seconds and videos that were 120 seconds. Exact *P* values are listed in Multimedia Appendix 1.



Views

The data were available for 122 (98%) videos. The median (IQR) views were 6533 (4708-8601) per video. The highest median (IQR) views were 10,499 (7792-15,650) for health promotion and awareness videos (n=16). Health promotion and awareness videos had significantly more views compared to theory and clinical skills videos (health promotion and awareness vs theory, *P*=.006; health promotion and awareness vs clinical skills, *P*=.003). There were no significant differences between other groups in pairwise comparisons (Figure 5C).

When analyzed by video length, the highest median (IQR) views were 7455 (5321-11,031) for videos that were more than 120 seconds (n=23). The longest videos (>120 seconds) and the shortest (≤60 seconds) reached more views compared to videos that were 61 to 120 seconds, but no statistically significant differences were observed between the longest and shortest videos (≤60 seconds vs 61-120 seconds, *P*=.02; >120 seconds vs 61-120 seconds, *P*=.02; Figure 5D). Comparing video implementation types (animation, human, phantom), no statistically significant differences were found.

From September 2023 onwards, Instagram Insights divided views into initial plays and replays. We analyzed the replays for videos published after September 2023 (n=29). The median (IQR) percentage of replays for the videos was 36% (35%-38%) of all views. When comparing video content, length, or implementation type, no statistically significant differences were found.

Discussion

In this study, we analyzed the content of a real-life SM account focused on medical education to determine which types of content were most well received by followers. To the best of our knowledge, no previous studies have performed a longitudinal real-life content analysis of a similar Instagram account [23,31-33]. By focusing exclusively on videos, this study provides new insights into their educational value. We specifically studied videos shared on Instagram, which is an applicable platform for short instructional educational videos applying the principles of multimedia learning and microlearning [4,17,36].

Health promotion and awareness posts were particularly successful in reaching audiences, as they reached most accounts (health promotion and awareness vs clinical skills, *P*=.003; health promotion and awareness vs theory, *P*=.006; health promotion and awareness vs entertainment, *P*=.04). As the World Health Organization (WHO) recognizes, SM platforms provide unique opportunities for promoting health awareness

XSL•FO
RenderX

and prevention [37]. Video-based educational interventions have previously been used to promote health awareness and prevention by professionals, researchers, and a variety of organizations [38,39]. For example, video interventions resulted in more targeted short-term health behaviors such as sunscreen adherence and breast self-examination, although the data are not uniform [39]. Our data support previous findings, as health promotional videos reached the widest audience. These videos could be integrated into broader organizational educational campaigns. *Plasticsteacher*'s health promotion and awareness videos were usually inspired by and produced to match SM trends, which could at least partially explain their popularity. There might be an opportunity to both educate students and spread medical knowledge by combining medical education with SM trends [21,29].

The median length of *Plasticsteacher* videos was 72 seconds, which is longer than the previously found average length of 33 seconds for medical posts on SM [40]. Health promotion and awareness videos, as well as entertainment videos, were significantly shorter than theory and clinical skills videos (health promotion and awareness vs theory, *P*=.006; health promotion and awareness vs clinical skills, *P*=.002; entertainment vs theory, *P*=.003; entertainment vs clinical skills, *P*=.002). We aimed to seek the optimal length for educational videos by assessing views, reaches, and relative watch time. Both the videos that were 60 seconds or less or more than 120 seconds were evenly reached and viewed, yet still statistically significantly better than videos that were 61 to 120 seconds (≤60 seconds vs 61-120 seconds, *P*=.02; >120 seconds vs 61-120 seconds, *P*=.02). However, the shorter videos (≤60 seconds) were watched significantly further when comparing the relative watch time (≤60 seconds vs 61-120 seconds, *P*=.003; ≤60 seconds vs >120 seconds, *P*=.002). Hence, the shortest videos (≤60 seconds) were most effective to deliver information, but longer videos (>120 seconds) were equally popular. Although previous qualitative and experimental studies have demonstrated that shorter videos were more engaging and university students were more inclined to rewatch videos [41], our study indicates that a variety of video lengths may be suitable for medical education on SM. The shortest video length in our study was 9 seconds, and the longest was 6 minutes and 16 seconds. Microlearning, which involves acquiring information in small units, serves as an emerging educational technique in medical education [36,42]. Depending on the literature, a microlearning lesson can be up to 5 minutes long or even up to 15 minutes in length, although students' focus tends to decrease after 9 minutes [36]. On the basis of our findings, short videos, adhering to the principles of microlearning, could be used to help students prepare for exams, such as reviewing the key steps of surgical procedures.

Although health promotion and awareness videos were the most popular, clinical skills, theory, and entertainment videos performed equally well, with no significant differences in reach or views between these categories. *Plasticsteacher* is targeting medical students on surgical topics, with views having an IQR of 4708 to 8601, which may indicate that videos of all categories are seen as meaningful. Interestingly, the video implementation type—whether using human models, phantom models, or animations—did not significantly impact video popularity, suggesting that producing a variety of video types may be beneficial when planning educational content. Different visualization methods, such as animations for explaining surgical techniques or human models for realistic demonstrations, can complement each other and enrich the learning experience [17]. However, the data on whether video implementation influences popularity or educational outcomes remain limited in the current literature [31-33]. For reference, in SM marketing, pictures with human faces and filters received more reactions compared to pictures without human faces or those that were unfiltered [43].

Plasticsteacher publishes posts in Finnish, which we see as an explanation for 98% of followers being from Finland. Eighty-three percent of the followers were females, and the largest age group was 25 to 34 years, with 44% representation. The primary target population for *Plasticsteacher* is medical students, with a secondary focus on other health care professionals. A survey was conducted among 4th to 6th year Finnish medical students, with a median age of 25 years and an age range of 21-42 years [44]. Similar results have been surveyed in Austria, where the mean age of medical students was 24.7 years, with an age range of 21 - 32 years [45]. Females have been over-represented in surveys of medical students, with 51% to 58% representation [44,45], but globally, as in Finland, professional care work is a female-dominated sector [46]. Based on the findings, we see that demographic data from Insights are in line with the followers *Plasticsteacher* is aiming to reach.

Videos on Instagram are easily accessible and provide a suitable platform for educational videos when needed [34]. The mobile app as a platform enables viewers to use videos for JIT learning to obtain information when needed [13]. It has been shown that a JIT learning–based mobile app enhanced the quality of medical education during on-call shifts in the emergency department [15]. Similarly, a short instructional video on splinting yielded better outcomes in terms of learning and skills [10]. Our analysis underlines the potential of Instagram videos as an educational tool, as 36% of views were replays, showing that, on average, more than half of the videos were watched again after the initial view.

Despite these strengths, this study has some limitations. The basis for parameter selection was that reached accounts and views could be seen as objective parameters for general popularity. Further, relative watch time may have indicated whether the video was considered beneficial by the viewer. The wide variability observed in reach, views, and watch times across categories may reflect the influence of Instagram's algorithms, which can cause posts to go viral unpredictably [43]. As in most retrospective studies, the data were not comprehensive and caused limitations for result interpretations. The development of Instagram Insights analytics during the study period also meant that some metrics were not consistently available for older posts. For example, Insights provided total watch time and average watch time beginning September 27, 2022. Nevertheless, these limitations were systematic, considering all types of posts similarly and did not compromise the overall findings.

Additionally, follower metrics including followers' age range, gender, and geographic data were visible for only the previous

XSL•FO
RenderX

90 days. A recent questionnaire study evaluated the engagement and impact of an educational veterinary Instagram account [47]. It demonstrated a high engagement rate and a positive impact on knowledge and practical skills among veterinary students [47]. Similar studies in the medical field would enable a more comprehensive analysis of followers, which was prevented by the Instagram Insight data limitation.

In conclusion, this study provides a unique analysis of video-based real-life educational content on Instagram. The study emphasizes the opportunity to produce high-quality medical content for the public without using actual patient data. The role of social media as a platform for medical education has been increasing [31-34], yet there remains a need to develop educational material that is evidence-based [2]. There is a clear gap in analytical research on this topic, in which this study stands out as a pioneering effort. Our findings highlight the effectiveness of short video length and health promotional content in engaging viewers, while also demonstrating the value of diverse video implementation types. We observed that demonstrative video content is well received by audiences. Health promotion content can serve as inspiration for medical professionals, and based on its founded effectiveness, it may be beneficial to incorporate such content into larger educational campaigns on SM. Short videos can be used in diverse ways on SM, such as summarizing key points on theoretical topics or reviewing key steps of procedures. Future studies should explore the impact of Instagram's algorithms on variations in video popularity and further examine how different demographic groups engage with educational content. Such studies would support the development of a more targeted and effective approach to medical education content creation.

Acknowledgments

The authors acknowledge the Biostatistics Unit's consulting service at the University of Helsinki for technical support with the data analysis. The authors declare the use of generative artificial intelligence (GAI) in the research and writing process. According to the GAIDeT (Generative AI Delegation Taxonomy; 2025), the following tasks were delegated to GAI tools under full human supervision: proofreading and editing, summarizing text, and translation. The GAI tool used was ChatGPT (GPT-4.0). Responsibility for the final manuscript lies entirely with the authors. GAI tools are not listed as authors and do not bear responsibility for the final outcomes. Artificial intelligence was not used in manuscript writing, data collection, generating figures or tables, or at any point in designing or carrying out the analysis.

Funding

The funding for this study was from departmental sources only.

Data Availability

The data are available from the corresponding author, AK, upon reasonable request.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Supplementary methodological analyses: exclusion of 2020-2021 posts, missing completely at random test, and adjusted *P* values (Benjamini-Hochberg corrections).
[DOCX File, 425 KB - mededu_v12i1e79124_app1.docx]

References

1. Maciuk K, Apollo M, Skorupa J, Jakubiak M, Wengel Y, Geary DC. Facebook is "for old people"—So why are we still studying it the most? A critical look at social media in science. Journal Media 2025;6(2):62. [doi: 10.3390/journalmedia6020062]
2. Katz M, Nandi N. Social Media and medical education in the context of the COVID-19 pandemic: scoping review. JMIR Med Educ 2021 Apr 12;7(2):e25892. [doi: 10.2196/25892] [Medline: 33755578]
3. Khattak O, Ganji KK, Agarwal A, et al. Student perception and preferences with social media for enhanced learning in health sciences following post-COVID-19 era: a cross-sectional study. Cureus 2023 Oct;15(10):e47390. [doi: 10.7759/cureus.47390] [Medline: 38022006]
4. Dash NR, Hasswan AA, Dias JM, et al. The educational use of social networking sites among medical and health sciences students: a cross campus interventional study. BMC Med Educ 2022 Jul 3;22(1):525. [doi: 10.1186/s12909-022-03569-3] [Medline: 35786406]
5. Hurtubise L, Martin B, Gilliland A, Mahan J. To play or not to play: leveraging video in medical education. J Grad Med Educ 2013 Mar;5(1):13-18. [doi: 10.4300/JGME-05-01-32] [Medline: 24404220]

XSL•FO
RenderX

6. Au-Yong-Oliveira M, Pinto Ferreira JJ. What if colorful images become more important than words? Visual representations as the basic building blocks of human communication and dynamic storytelling. World Futures Review 2014 Mar;6(1):48-54. [doi: 10.1177/1946756714528880]
7. Hsin W, Cigas J. Short videos improve student learning in online education. J Comput Sci Coll 2013;28(5):253-259 [FREE Full text]
8. Stockwell BR, Stockwell MS, Cennamo M, Jiang E. Blended learning improves science education. Cell 2015 Aug 27;162(5):933-936. [doi: 10.1016/j.cell.2015.08.009] [Medline: 26317458]
9. Noetel M, Griffith S, Delaney O, et al. Video improves learning in higher education: a systematic review. Rev Educ Res 2021 Apr;91(2):204-236. [doi: 10.3102/0034654321990713]
10. Cheng YT, Liu DR, Wang VJ. Teaching splinting techniques using a just-in-time training instructional video. Pediatr Emerg Care 2017 Mar;33(3):166-170. [doi: 10.1097/PEC.0000000000000390] [Medline: 25834963]
11. Ahmet A, Gamze K, Rustem M, Sezen KA. Is video-based education an effective method in surgical education? A systematic review. J Surg Educ 2018;75(5):1150-1158. [doi: 10.1016/j.jsurg.2018.01.014] [Medline: 29449162]
12. Mayer RE, Johnson CI. Revising the redundancy principle in multimedia learning. J Educ Psychol 2008;100(2):380-386. [doi: 10.1037/0022-0663.100.2.380]
13. De Gagne JC, Park HK, Hall K, Woodward A, Yamane S, Kim SS. Microlearning in health professions education: scoping review. JMIR Med Educ 2019 Jul 23;5(2):e13997. [doi: 10.2196/13997] [Medline: 31339105]
14. Bannister J, Neve M, Kolanko C. Increased educational reach through a microlearning approach: can higher participation translate to improved outcomes? J Eur CME 2020 Oct 20;9(1):1834761. [doi: 10.1080/21614083.2020.1834761] [Medline: 33354408]
15. Ginsburg J, Sande M, Ahmed A, Moak J, Sudhir A, Mutter MK. A smartphone resource for just-in-time medical student teaching by emergency medicine residents. Cureus 2024 Sep;16(9):e70117. [doi: 10.7759/cureus.70117] [Medline: 39449879]
16. Castelli G, Winters S, Taylor A, Ge Y. Using social media to teach about and engage residents in evidence-based medicine. Fam Med 2024 Oct;56(9):572-578. [doi: 10.22454/FamMed.2024.808735] [Medline: 39207790]
17. Mayer RE. Multimedia learning. Annu Rep Educ Psychol Jpn 2002;41:27-29. [doi: 10.5926/arepj1962.41.0_27]
18. Lewis LF, Rohloff LM. Organizations prepare: generation z has arrived with a new learning style. Int Educ Cult Stud 2021;1:57-61 [FREE Full text]
19. Malik AM, Welborn M, Gao H, Motaparthi K. Self-directed virtual educational strategies for dermatopathology: a prospective comparison of Instagram-based static images to whole slide images. J Cutan Pathol 2023 Aug;50(8):692-694. [doi: 10.1111/cup.14472] [Medline: 37272167]
20. Essig J, Watts M, Beck Dallaghan GL, Gilliland KO. InstaHisto: utilizing Instagram as a medium for disseminating visual educational resources. Med Sci Educ 2020 Sep;30(3):1035-1042. [doi: 10.1007/s40670-020-01010-2] [Medline: 34457765]
21. Madanick RD. Education becomes social: the intersection of social media and medical education. Gastroenterology 2015 Oct;149(4):844-847. [doi: 10.1053/j.gastro.2015.08.037] [Medline: 26311278]
22. Brandenburg DC, Ellinger AD. The future: just-in-time learning expectations and potential implications for human resource development. Adv Dev Hum Resour 2003 Aug;5(3):308-320. [doi: 10.1177/1523422303254629]
23. Yamamoto T, Goto K, Kitano S, et al. Current insights on social media as a tool for the dissemination of research and education in surgery: a narrative review. Surg Today 2024 Oct;54(10):1113-1123. [doi: 10.1007/s00595-024-02891-1] [Medline: 38980332]
24. Escamilla-Sanchez A, López-Villodres JA, Alba-Tercedor C, et al. Instagram as a tool to improve human histology learning in medical education: descriptive study. JMIR Med Educ 2025 Feb 19;11:e55861. [doi: 10.2196/55861] [Medline: 39970433]
25. Primack BA, Shensa A, Sidani JE, et al. Social media use and perceived social isolation among young adults in the U.S. Am J Prev Med 2017 Jul;53(1):1-8. [doi: 10.1016/j.amepre.2017.01.010] [Medline: 28279545]
26. O'Day EB, Heimberg RG. Social media use, social anxiety, and loneliness: a systematic review. Computers Hum Behav Rep 2021;3:100070. [doi: 10.1016/j.chbr.2021.100070]
27. Wong D, Amon KL, Keep M. Desire to belong affects instagram behavior and perceived social support. Cyberpsychol Behav Soc Netw 2019 Jul;22(7):465-471. [doi: 10.1089/cyber.2018.0533] [Medline: 31295026]
28. Lin H, Fan W, Chau PYK. Determinants of users' continuance of social networking sites: a self-regulation perspective. Inf Manage 2014 Jul;51(5):595-603. [doi: 10.1016/j.im.2014.03.010]
29. Goldie JGS. Connectivism: A knowledge learning theory for the digital age? Med Teach 2016 Oct;38(10):1064-1069. [doi: 10.3109/0142159X.2016.1173661] [Medline: 27128290]
30. Siemens G. Connectivism: a learning theory for the digital age. Int J Instr Technol Distance Learn 2005;2(1) [FREE Full text]
31. Cheston CC, Flickinger TE, Chisolm MS. Social media use in medical education: a systematic review. Acad Med 2013 Jun;88(6):893-901. [doi: 10.1097/ACM.0b013e31828ffc23] [Medline: 23619071]
32. Guckian J, Utukuri M, Asif A, et al. Social media in undergraduate medical education: a systematic review. Med Educ 2021 Nov;55(11):1227-1241. [doi: 10.1111/medu.14567] [Medline: 33988867]
33. Sterling M, Leung P, Wright D, Bishop TF. The use of social media in graduate medical education: a systematic review. Acad Med 2017 Jul;92(7):1043-1056. [doi: 10.1097/ACM.0000000000001617] [Medline: 28225466]

XSL•FO
RenderX

34. Sutherland S, Jalali A. Social media as an open-learning resource in medical education: current perspectives. Adv Med Educ Pract 2017;8:369-375. [doi: 10.2147/AMEP.S112594] [Medline: 28652840]
35. Muca E, Buonaiuto G, Lamanna M, et al. Reaching a wider audience: Instagram's role in dairy cow nutrition education and engagement. Animals (Basel) 2023 Nov 13;13(22):3503. [doi: 10.3390/ani13223503] [Medline: 38003121]
36. Denojean-Mairet M, López-Pernas S, Agbo FJ, Tedre M. A literature review on the integration of microlearning and social media. Smart Learn Environ 2024;11(1):46-46. [doi: 10.1186/s40561-024-00334-5]
37. Health 2020: a European policy framework supporting action across government and society for health and well-being. : World Health Organization Regional Office for Europe; 2013 URL: https://iris.who.int/server/api/core/bitstreams/2731a5d0-ec46-48ca-b828-973bb9b50919/content [accessed 2026-03-19]
38. Moorhead SA, Hazlett DE, Harrison L, Carroll JK, Irwin A, Hoving C. A new dimension of health care: systematic review of the uses, benefits, and limitations of social media for health communication. J Med Internet Res 2013 Apr 23;15(4):e85. [doi: 10.2196/jmir.1933] [Medline: 23615206]
39. Tuong W, Larsen ER, Armstrong AW. Videos to influence: A systematic review of effectiveness of video-based education in modifying health behaviors. J Behav Med 2014 Apr;37(2):218-233. [doi: 10.1007/s10865-012-9480-7] [Medline: 23188480]
40. Gurler D, Buyukceran I. Assessment of the medical reliability of videos on social media: detailed analysis of the quality and usability of four social media platforms (Facebook, Instagram, Twitter, and YouTube). Healthcare (Basel) 2022 Sep 22;10(10):1836. [doi: 10.3390/healthcare10101836] [Medline: 36292284]
41. Carmichael M, Reid AK, Karpicke JD. Assessing the impact of educational video on student engagement, critical thinking and learning: the current state of play. : SAGE Publishing; 2024 URL: https://www.sagepub.com/docs/default-source/corp-comms/white-papers/assessing-the-impact-of-educational-video-on-student-engagement-critical-thinking-and-learning-the-current-state-of-play.pdf?sfvrsn=7b013e6f_1 [accessed 2026-03-19]
42. Palmon I, Brown CS, Highet A, et al. Microlearning and social media: a novel approach to video-based learning and surgical education. J Grad Med Educ 2021 Jun;13(3):323-326. [doi: 10.4300/JGME-D-20-01562.1] [Medline: 34178254]
43. Jaakonmäki R, Müller O, vom Brocke J. The impact of content, context, and creator on user engagement in social media marketing. 2017 Presented at: Proceedings of the 50th Hawaii International Conference on System Sciences (HICSS); Jan 4-7, 2017. [doi: 10.24251/HICSS.2017.136]
44. Hannula V, Lehtovirta L, Liu H, Sumanen M. Medical students' perceptions of the best clinical teaching. Clin Teach 2025 Aug;22(4):e70119. [doi: 10.1111/tct.70119] [Medline: 40514192]
45. Siller H, Tauber G, Komlenac N, Hochleitner M. Gender differences and similarities in medical students' experiences of mistreatment by various groups of perpetrators. BMC Med Educ 2017 Aug 14;17(1):134. [doi: 10.1186/s12909-017-0974-4] [Medline: 28806943]
46. Hökkä M, Lehto JT, Kyngäs H, Pölkki T. Finnish nursing students' perceptions of the development needs in palliative care education and factors influencing learning in undergraduate nursing studies - a qualitative study. BMC Palliat Care 2022 Mar 23;21(1):40. [doi: 10.1186/s12904-022-00915-6] [Medline: 35321718]
47. Lamanna M, Muca E, Buonaiuto G, Formigoni A, Cavallini D. From posts to practice: Instagram's role in veterinary dairy cow nutrition education—How does the audience interact and apply knowledge? A survey study. J Dairy Sci 2025 Feb;108(2):1659-1671. [doi: 10.3168/jds.2024-25347] [Medline: 39662806]

Abbreviations

JIT: just-in-time
MCAR: missing completely at random
OER: open educational resources
SM: social media
WHO: World Health Organization

Edited by S Brini; submitted 24.Jun.2025; peer-reviewed by D Cavallini, MMDA Mascarenha; accepted 23.Feb.2026; published 15.Apr.2026.

Please cite as:
Kemppainen A, Nuutinen H, Koljonen V, Kulmala P
Instagram Video Engagement in Medical Education: Cross-Sectional Study
JMIR Med Educ 2026;12:e79124
URL: https://mededu.jmir.org/2026/1/e79124
doi:10.2196/79124

XSL•FO
RenderX

© Aapeli Kemppainen, Henrik Nuutinen, Virve Koljonen, Petri Kulmala. Originally published in JMIR Medical Education (https://mededu.jmir.org), 15.Apr.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Education, is properly cited. The complete bibliographic information, a link to the original publication on https://mededu.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Exploring the Icarus Paradox in Indonesia's Specialist Medical Education System Using the Public Perspective From Online Media: Convergent Mixed Methods Study

Faisal Binsar[1], DR (AM); Mohammad Hamsal[2], Prof Dr

[1]Management Department, Binus Online Learning, Binus University, Jakarta, Jakarta, Indonesia

[2]Binus University, Binus Business School Doctor of Research in Management, Binus University, Jakarta, Jakarta, Indonesia

Corresponding Author:
Faisal Binsar, DR (AM)
Management Department
Binus Online Learning
Binus University
Jl K H Syahdan No 9
Kemanggisan, Palmerah
Jakarta, Jakarta, 11480
Indonesia
Phone: 62 215345830
Fax: 62 215300244
Email: faisal.binsar@binus.ac.id

Related Articles:

Comment in: https://mededu.jmir.org/2026/1/e94825/

Comment in: https://mededu.jmir.org/2026/1/e97664/

Abstract

Background: The Icarus Paradox in health care refers to the tension between the ambition to succeed as a specialist doctor and the limitations of the medical education system. Indonesia aspires to produce quality doctors, yet limited infrastructure and resources hinder the educational journey of prospective specialists.

Objective: This study aimed to identify the Icarus Paradox in Indonesia's specialist medical education by examining prospective specialist medical students and the quality of health services and by analyzing how this paradox is reflected in society's perspectives.

Methods: Using a convergent mixed methods design, this study integrated quantitative content analysis of 5047 online reviews across multiple platforms with qualitative thematic and cognitive analysis using NVivo 14, combining sentiment classification and topic coding.

Results: Twitter contributed 573 (11.3%) of 5047 reviews, with 218 (38%) negative, 251 (43.8%) neutral, and 104 (18.2%) positive entries. TikTok generated 282 (5.6%) reviews, the majority being neutral (n=225, 79.5%). YouTube produced 96 (1.9%) reviews, with 89 (92.7%) neutral entries. News platforms exhibited the largest volume (n=3040, 60.2%) of reviews, with 2885 (94.9%) neutral, 105 (3.5%) positive, and 50 (1.6%) negative entries. Blogs and websites contributed 353 (7%) and 692 (11.3%) reviews, respectively, with neutral sentiment dominating (n=329, 93.2%, for blogs and n=599, 86.6%, for websites). Three cognitive perspectives demonstrated the Icarus Paradox in the Indonesian medical education system: education system, society's views of students, and health care services. Although there are aspirations to improve education and health care quality, these ambitions often collide with structural challenges, such as resource shortages, heavy workloads, and limited accessibility, which link directly to cognitive themes of stress, resilience, and ethical dilemmas. We proposed a conceptual model to illustrate these dynamics.

Conclusions: Our findings offer insights into the Icarus Paradox in Indonesia's medical education system, highlighting its complexity and reinforcing the need for systemic reform. Beyond academic relevance, the findings also emphasize the importance of strengthening student mental health support, ensuring equitable access to health care, and enhancing regulatory oversight of

XSL•FO
RenderX

training. This was not a clinical trial. Although limited by reliance on online reviews, the results underscore the urgent need for targeted policy interventions in medical education and health care services.

Trial Registration: ClinicalTrials.gov registration: NCT123456

(JMIR Med Educ 2026;12:e60452) doi:10.2196/60452

KEYWORDS

Icarus Paradox; aspirations; realities; medical education; student well-being; community perspective; online reviews

Introduction

Background

The Icarus Paradox, originally used in management studies to explain how strengths can become liabilities, offers a useful lens to examine paradoxical dynamics in specialist medical education [1]. Although such programs are designed to produce highly competent professionals, they often generate conditions that undermine performance and well-being. These paradoxes arise from three interconnected sources.

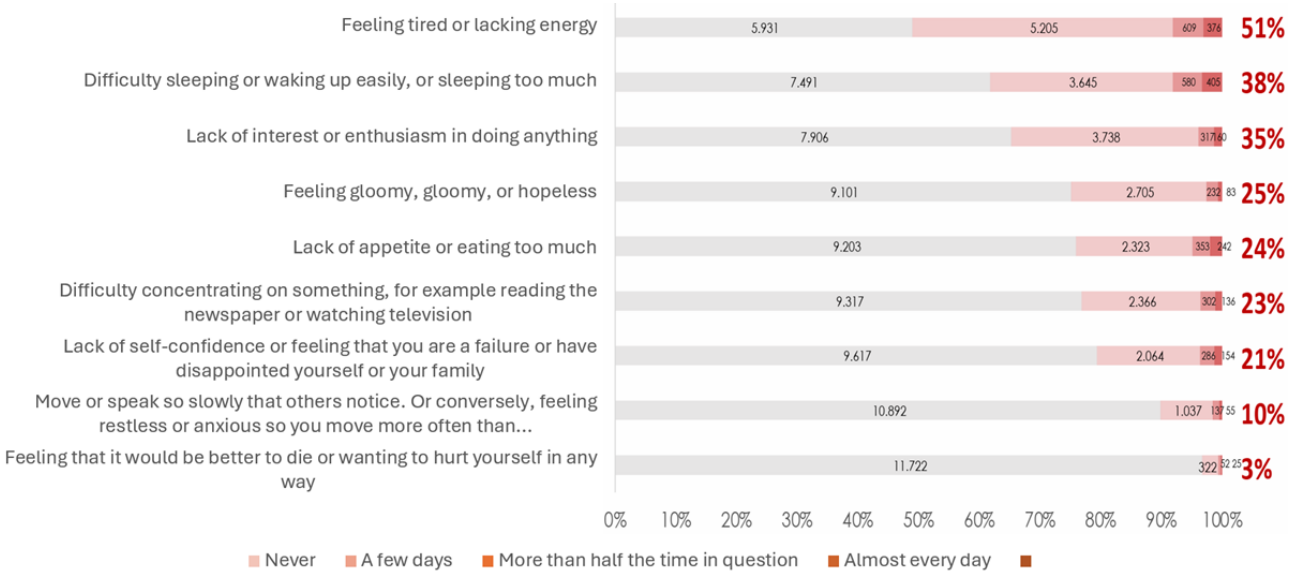
First, systemic educational design—marked by rigid curricula, competitive entry pathways, and limited infrastructure—restricts flexibility and adaptability. Second, entrenched institutional cultures characterized by hierarchy, conformity, and normalization of bullying shape learning environments where silence is often preferred over expression [2-6]. Third, the expectation that trainees endure prolonged working hours, intensive workloads, and psychological pressure with minimal support places substantial demands on individual resilience. These interactions illustrate how a system striving for excellence may inadvertently create vulnerabilities that undermine its goals.

Paradox theory underpins this study, emphasizing the simultaneous coexistence of mutually contradictory yet interdependent elements that persist over time [7,8]. A paradox is defined as “contradictory yet interrelated elements that exist simultaneously and persist over time” [9]. In the context of specialist medical education, these tensions emerge when academic rigor, resilience, and institutional pride—typically considered strengths—lead to burnout, intimidation, or compromised well-being.

Medical students and residents frequently face extreme academic pressure that may lead to burnout or depression [10]. Burnout is characterized by emotional exhaustion, depersonalization, and reduced personal accomplishment [11], while depression comprises persistent sadness, loss of interest, or suicidal ideation. Although overlapping, the two conditions are conceptually distinct. Resident doctors are particularly vulnerable to both, and their compromised well-being negatively affects patient care quality [11]. Media coverage frequently portrays emergency conditions in specialist medical education, illustrating the severity of burnout and distress among trainees [12-14]. Policymakers have expressed growing concern over declining patient care quality linked to these systemic issues [15].

A national survey conducted by the Ministry of Health of Indonesia in 2024 reported that among 12,121 participants in the *Program Pendidikan Dokter Spesialis* (PPDS; Specialist Doctor Education Program) at 28 vertical education hospitals, 2716 (22.4%) showed depressive symptoms [16]. Additionally, 6182 (51%) reported fatigue or low energy, 4606 (38%) experienced sleep disturbances, 4242 (35%) reported reduced interest in daily activities, and 400 (3.3%) indicated suicidal ideation or self-harm thoughts (Figure 1) [16]. These results underscore the significant mental health challenges facing trainees.

Figure 1. PPDS mental health screening at vertical education hospitals. PPDS: *Program Pendidikan Dokter Spesialis* (Specialist Doctor Education Program).



XSL•FO
RenderX

The same survey revealed that the specialties with the highest depressive symptoms were pediatrics, with 1672 (13.8%) trainees; orthopedics and traumatology, with 1491 (12.3%) trainees; and internal medicine, with 1091 (9.0%) trainees [16]. This situation raises concerns regarding both trainee well-being and patient care quality. Media reports frequently highlight psychological crises among specialist medical trainees, echoing public worries about inadequate institutional support. High academic expectations combined with insufficient psychosocial resources leave many trainees struggling with anxiety, exhaustion, and a profound sense of entrapment.

These challenges reflect worldwide patterns. In the United States, a prospective study found that depressive symptoms increase progressively across medical training, with fourth-year students having more than eight times the odds of emotional exhaustion compared to first-year students [17]. More than one-third of US family medicine residents report burnout, particularly women and those with high educational debt [18]. In Canada, over half of resident doctors experience burnout, especially those working over 80 hours per week [19]. A scoping review reported a wide variation in the prevalence of burnout and depression among physicians and residents, with the evidence consistently indicating greater vulnerability among women [20].

Regional data reveal similar patterns: Iranian medical students experience significant burnout associated with academic pressures [21]; Hong Kong students commonly report burnout linked to poor sleep and limited exercise [22]; and Thai clinical students report high emotional exhaustion and depersonalization, particularly among males and those with lower academic performance [23]. The British Medical Association highlights excessive demands and poor working conditions as major contributors to declining physician well-being worldwide [24].

Although these issues occur worldwide, Indonesia's case is particularly severe due to rigid systemic design, hierarchical cultures, and limited psychosocial support. These characteristics intensify paradoxical tensions consistent with broader paradox scholarship, which argues that unresolved tensions often generate both innovation and dysfunction [7,8].

Aims of the Study

This study had two primary aims. Theoretically, it used the Icarus Paradox to explain how exclusivity and academic excellence within Indonesia's specialist medical education system can inadvertently produce negative outcomes for trainee well-being and service quality. Practically, it examined public perspectives captured through online media to identify institutional gaps in responsiveness to societal expectations, health care market needs, regulatory dynamics [25-28], and medical technology advancement [29]. In doing so, the study asked how the Icarus Paradox manifests within Indonesia's specialist medical education system from the viewpoint of the public, and how these insights can guide policymakers and training institutions toward reforms that balance quality, responsiveness, and trainee well-being [29].

Theoretical Background

Paradoxical dynamics have become increasingly prominent in contemporary organizations as scholars examine how contradictory yet interrelated elements coexist and shape behavior over time. Such tensions surface when organizations must respond to opposing demands that are both legitimate, generating cycles of alignment, resistance, and adaptation. These dynamics appear across institutional logics, identity structures, and organizational processes, with senior leaders experiencing them as strategic contradictions and employees encountering them in daily routines and socioemotional relationships [30]. Although paradox scholarship has generated numerous influential studies [31-33], its rapid adoption also raises concerns about conceptual stretching, overconfidence in prevailing interpretations, and labeling practices that may reinforce dominant theoretical positions [34].

This study adopted the Icarus Paradox as its core theoretical lens. Rooted in Greek mythology, the metaphor describes how Icarus's success in flying ultimately led to failure when he ignored warnings and flew too close to the sun [35-37]. In strategic management, the Icarus Paradox explains how organizations' strengths can become vulnerabilities when past success fosters overconfidence, rigidity, or reluctance to adapt [35,38]. Institutions that excel may inadvertently undermine their own survival by persisting with established routines despite environmental shifts [39]. This phenomenon, often referred to as "vicious inertia," reflects a broader tension between stability and adaptation—an enduring hallmark of paradox theory. Comparable dynamics emerge in education, where aspirations for excellence collide with resource scarcity, rigid structures, and uneven institutional preparedness [40].

Paradoxical tensions are also evident in medical education. High academic standards and professional expectations are essential for producing competent physicians, yet they simultaneously generate psychological burden among trainees. Medical students frequently report stress, burnout, and depression [41,42], and Indonesia's specialist medical trainees experience particularly elevated levels of distress, with consequences for well-being and professional performance [10]. These tensions represent *strategic paradoxes*—institutions prioritizing excellence and productivity at the expense of trainee mental health. Identity paradoxes also emerge as students' aspirations to become empathetic, capable physicians conflict with the exhaustion, self-doubt, and emotional strain encountered during training. Institutional logic paradoxes further arise when cultures valorize endurance and sacrifice, while health systems increasingly demand empathy, patient-centeredness, and clinician well-being.

These psychological strains carry serious implications. Burnout and depressive symptoms reduce clinical performance, increase the likelihood of errors, and diminish empathy toward patients [43]. If unaddressed, these conditions often persist into residency and professional practice, reinforcing negative cycles within the health care workforce [11]. Medical education paradoxes therefore represent tangible, lived realities that directly influence patient safety and service quality.

Addressing these tensions requires intentional institutional support. Access to mental health services, resilience education,

XSL•FO
RenderX

and curricula integrating well-being are foundational [42,43]. Another paradox emerges from the gap between expected graduate competencies and actual system-level outcomes: training aspires to produce highly capable specialists, but health care performance does not always match societal expectations [44]. Continuous reform is therefore essential to align educational structures with evolving health care needs, while recognizing human limitations.

Social media has recently played a dual role in illuminating these paradoxes. On the one hand, it facilitates collaboration [27], resource sharing [45], knowledge exchange [46], and peer support that strengthens students' sense of belonging [26,47,48]. On the other hand, it can propagate misinformation, amplify stress, and distract from academic focus [49]. Online engagement also provides a space for public commentary [50,51], revealing lived experiences of medical trainees and societal expectations of the health care system. Thus, social media functions both as a catalyst and a mirror of the paradoxes embedded within specialist medical education.

Methods

Study Design

This study was registered at ClinicalTrials.gov (NCT123456). The research used a convergent mixed methods design [52], in which quantitative and qualitative approaches were conducted simultaneously and coalesced to provide a rich understanding of the Icarus Paradox within the specialist medical trainee education system in Indonesia. In this design, quantitative sentiment analysis and qualitative thematic analysis were conducted independently but merged at the interpretation stage to facilitate triangulation of findings.

Data Collection

Data collection was carried out over 30 days (March 24-April 23, 2024) using Brand24 [50,53,54], a social listening and monitoring tool that enables focused and comprehensive analysis of conversations on platforms such as Twitter [55], online news portals, blogs, videos, discussion forums, and websites [56,57]. Brand24 was selected over other monitoring tools because it provides reliable access to Indonesian language data, allows automated tracking of multiple keywords, and has been previously validated in academic studies [50].

The search was conducted using a set of keywords relevant to Indonesia's specialist medical education system, including "PPDS," "resident doctor," "specialist candidate," "collegium," "medical student," and "medical education." The data collected were stored in Microsoft Excel format, containing metadata, such as the posting time, ID or username, source, review content, sentiment category, number of replies, number of likes, number of reposts, number of followers, and influence score.

To ensure the relevance of the dataset, irrelevant or duplicate content was systematically removed. Posts were excluded if they were (1) purely promotional or advertisements, (2) unrelated to medical education despite containing a keyword, (3) spam or bot-generated content, or (4) duplicates of previously captured entries. This filtering step ensured that only meaningful and contextually relevant discussions were retained for analysis.

Data Analysis

The analysis process involved both quantitative sentiment analysis and qualitative thematic analysis.

For quantitative sentiment analysis, Python was used with text mining and natural language processing (NLP) libraries, such as *NLTK*, *spaCy*, *Pandas*, and *Scikit-learn* [58,59]. Sentiment classification grouped posts into three categories: positive, neutral, and hostile. Additional preprocessing steps included removing hashtags, emojis, hyperlinks, and user mentions, as these elements did not add meaningful semantic value for sentiment classification. The output of this analysis included statistics, graphs, proportions, and trends.

Since the dataset was originally in Indonesian, translation into English was necessary for subsequent qualitative coding. Translation was performed using Google Translate, combined with manual verification by bilingual members of the research team to ensure contextual and semantic accuracy. This step reduced the risk of misinterpretation and enhanced the reliability of findings.

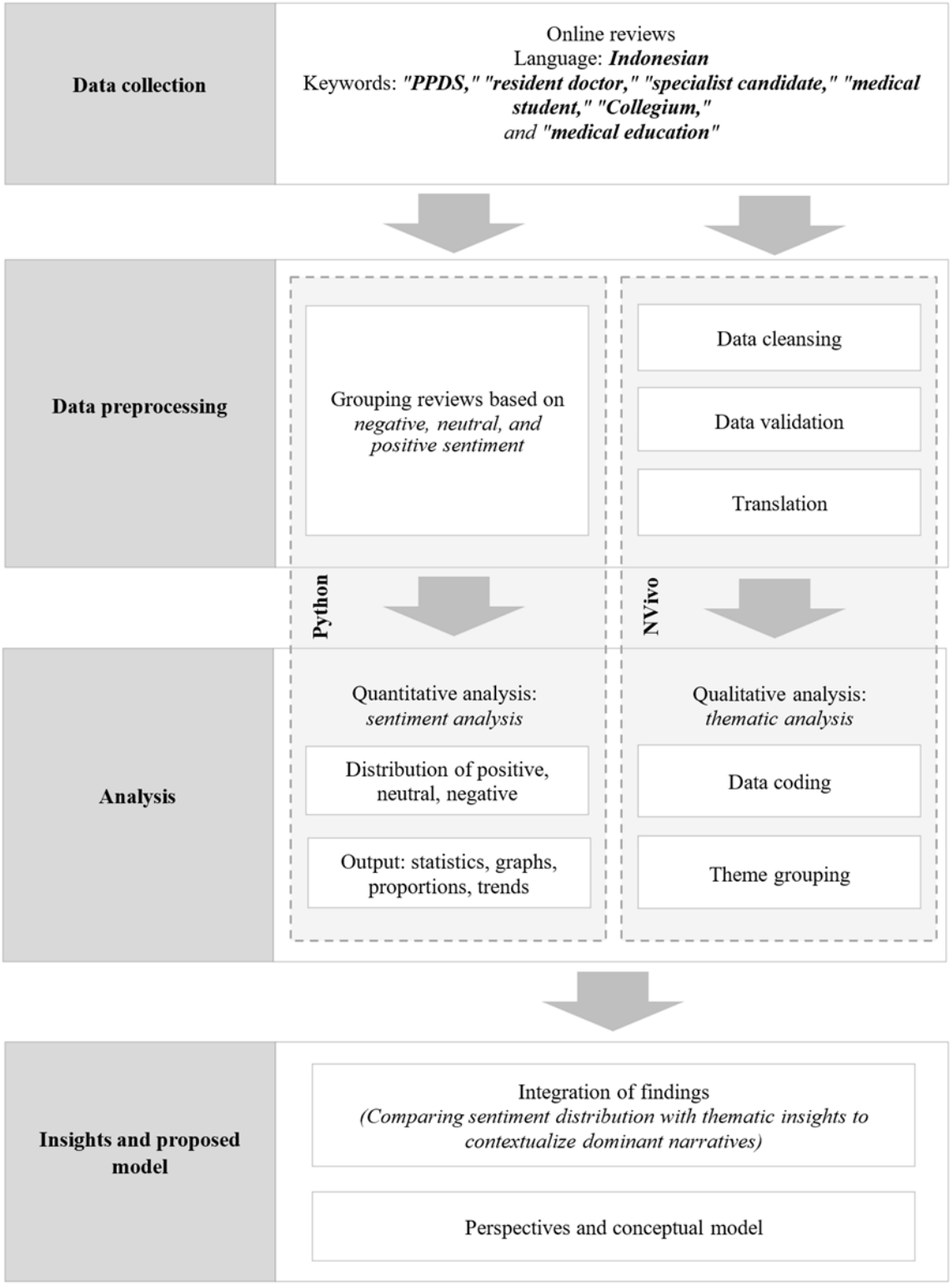
For qualitative thematic analysis, data were imported into NVivo 14 [60]. Coding was performed using a hybrid approach: inductive coding allowed themes to emerge directly from the data, while deductive coding was informed by paradox theory and the Icarus Paradox framework. To enhance intercoder reliability, multiple researchers were involved in the coding process, and discrepancies were resolved through discussion until a consensus was reached. Codes with similar content were grouped into broader themes to identify dominant narratives.

Integration of Findings

The final stage involved integrating the results from both strands of analysis. Quantitative sentiment trends were compared with qualitative thematic insights to contextualize dominant narratives about specialist medical education. This integration enabled the development of a perspectives-based conceptual model, summarizing how the Icarus Paradox manifests in Indonesia's specialist medical education system and how public perceptions highlight areas requiring reform. A summary of all stages of the method is shown in Figure 2.

XSL•FO
RenderX

Figure 2. Convergent mixed methods study design. PPDS: *Program Pendidikan Dokter Spesialis* (Specialist Doctor Education Program).



Ethical Considerations

This study used publicly available, nonidentifiable online data and did not involve interaction with human participants; therefore, Institutional Review Board (IRB) approval was not required. The authors adhered to applicable ethical standards for research using public online data and removed any potentially identifying information during analysis.

Results

Online Review Data

Data were collected for 30 days (March 24-April 23, 2024) using Brand24 as the primary social listening and analytics platform [50,53,54]. This tool was selected because of its ability to capture Indonesian language data across multiple platforms, provide sentiment classification, and export structured datasets

XSL•FO
RenderX

for further quantitative and qualitative analyses. In total, 5047 reviews were retrieved based on the specified keywords, distributed across digital platforms such as Twitter (X), TikTok, video-sharing sites, online news portals, podcasts, forums, blogs, and websites.

The platform with the highest number of reviews was news portals, with 3040 (60.2%) reviews; followed by websites, with 692 (13.7%), Twitter with 573 (11.3%), and blogs with 353 (7%) reviews. TikTok and video systems additionally contributed significantly, with 282 (5.6%) reviews and 96 (1.9%) opinions, respectively. Podcasts accounted for the best 6 (0.1%) evaluations, and forums contributed 5 (0.1%) reviews, making them the least represented platforms. These percentages provide a clearer picture of the dominance of such structures, with news portals rising to become the number one space for public discourse on professional clinical education in Indonesia. The variation highlights how exclusive audiences use more than one online channel to share their perspectives and worries, underscoring the subject's relevance in numerous segments of digital media.

Sentiment Analysis

Sentiment classification of the collected data was conducted using the Brand24 platform [61]. Brand24 integrates artificial intelligence (AI)–driven deep learning with rule-based lexicon techniques, enabling both high accuracy and methodological transparency. Its multilingual sentiment model supports analysis in more than 100 languages, including Indonesian, and its performance has been validated using the F_1-score metric. Based on evaluations of 50,000 mentions, overall accuracy increased [61].

The processed data reflected a mixture of negative, neutral, and positive sentiments across platforms, as summarized in Table 1. The online review dataset showed substantial variation in the volume and distribution of sentiments between media sources. Of the 573 (11.3%) reviews contributed by Twitter, 218 (38%) were categorized as negative, 251 (43.8%) as neutral, and 104 (18.2%) as positive entries. Of the 282 (5.6%) reviews generated by TikTok, the majority (n=225, 79.5%) were categorized as neutral. For video-based platforms, such as YouTube, which produced 96 (1.9%) reviews, 89 (92.7%) reviews were categorized as neutral.

Table 1. Distribution of review data by platform.

Platform	Negative reviews (n=312), n (%)	Neutral reviews (n=4389), n (%)	Positive reviews (n=346), n (%)	Total (N=5047), n (%)
Twitter (X)	218 (69.9)	251 (5.7)	104 (30.1)	573 (11.3)
TikTok	19 (6.1)	225 (5.1)	38 (11.0)	282 (5.6)
Videos	1 (0.3)	89 (2.0)	6 (1.7)	96 (1.9)
News portals	50 (16.0)	2885 (65.7)	105 (30.3)	3040 (60.2)
Podcasts	0	6 (0.1)	0	6 (0.1)
Forums	0	5 (0.1)	0	5 (0.1)
Blogs	11 (3.5)	329 (7.5)	13 (3.8)	353 (7.0)
Websites	13 (4.2)	599 (13.6)	80 (23.1)	692 (13.7)

News platforms exhibited the largest volume of data, amounting to 3040 (60.2%) reviews; however, most were neutral (n=2885, 94.9%), accompanied by 105 (3.5%) positive and 50 (1.6%) negative sentiments. Blogs and websites contributed 353 (7%) and 692 (11.3%) reviews, respectively, with neutral sentiment dominating both categories (n=329, 93.2%, for blogs and n=599, 86.6%, for websites).

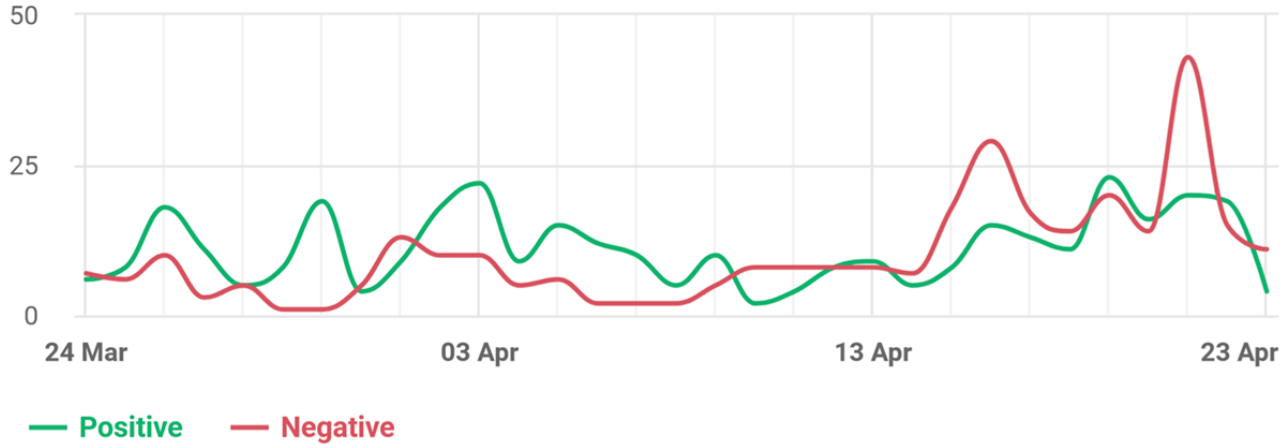
Overall, these findings indicated substantial cross-platform variation in sentiment distribution. Despite differences in volume, most platforms show a predominance of neutral sentiment, suggesting that public discourse on specialist medical education tends to be descriptive or informational rather than overtly evaluative.

These data show that Twitter has the highest proportion of negative sentiment, whereas TikTok has the highest proportion of positive sentiment relative to its volume. Meanwhile, news portals are overwhelmingly neutral, reflecting the reporting style of professional media outlets. Podcasts and forum platforms do not show any negative or positive sentiment, while video and web-based platforms have a low number of negative- and positive-sentiment reviews. Furthermore, news platforms show a fairly low number of negative- and positive-sentiment reviews, with neutral sentiment dominating. Overall, the percentage of negative sentiment of around 6.2% (n=312 reviews) is smaller than the percentage of positive sentiment of 6.9% (n=346 reviews).

Figure 3 shows the trend in daily positive and negative reviews for 1 month. At the beginning of the month, the number of positive and negative reviews was relatively low and stable. However, starting April 16, there was a significant increase in the number of positive and negative reviews. This is in line with reports in the mass media regarding the results of PPDS mental health screening [16] on that date [12], which may have triggered an increase in online activity related to the topics discussed in the research.

XSL•FO
RenderX

Figure 3. Sentiment-based review history for 1 month.



Coding Review Topics

After carrying out data cleaning, 292 (5.8%) reviews with negative sentiment and 317 (6.3%) reviews with positive sentiment were obtained. These data were then analyzed using NVivo 14.

Negative-Review Topic Coding

As the initial aim of this research, each sentiment in the review was coded to find review themes. We used the *autocode* feature provided by NVivo 14 to automatically recognize and mark text or other qualitative data based on predetermined rules. NVivo applies techniques such as text analysis and content analysis to identify patterns in text and tag or code data accordingly [62].

To ensure the credibility of the coding process, the results of autocode generation were followed by manual validation carried out by two independent coders. An intercoder agreement check was conducted, and discrepancies were resolved through discussion to strengthen the reliability of the coding.

The autocode results produced 28 negative-review topic codes, as presented in Table 2. To improve interpretability, these topics were subsequently grouped into six broader clusters: *medical workforce*, *educational system and training*, *health care institutions and patients*, *financial and resource constraints*, *media and public discourse*, and *other student life factors*.

The largest cluster was “educational system and training” (n=8, 28.6%, topics), followed by “medical workforce” (n=7, 25%, topics). This highlights that criticisms are not only about individuals (doctors, students, residents) but also about structural issues in programs and curricula. Smaller clusters, such as “media and public discourse” and “student life factors,” capture contextual influences that extend beyond institutional boundaries, underscoring the multifaceted nature of the challenges.

Discussion topic data generated by NVivo 14 related to negative community reviews revealed a variety of topics highlighted in these reviews. Of the 28 topics identified, based on the number of references and reviews (Figure 4a), the data showed that the topics of discussion tended to be spread out, with significant variations in the number of references and reviews. Most topics had an equal number of references and reviews, with some topics, such as “doctors” and “specialists,” having a slight difference in the number of references and reviews. However, when looking from a coverage perspective (Figure 4b), most topics had low scores, indicating that these topics were not widely discussed in the data. Nevertheless, it was still important to pay attention to them because some topics could affect the validity and representativeness of analysis results.

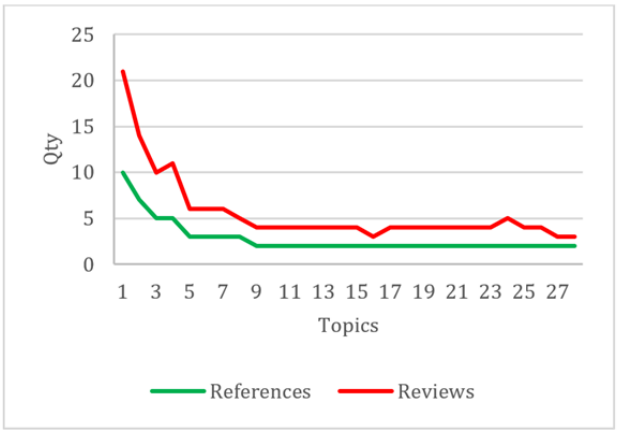
XSL•FO
RenderX

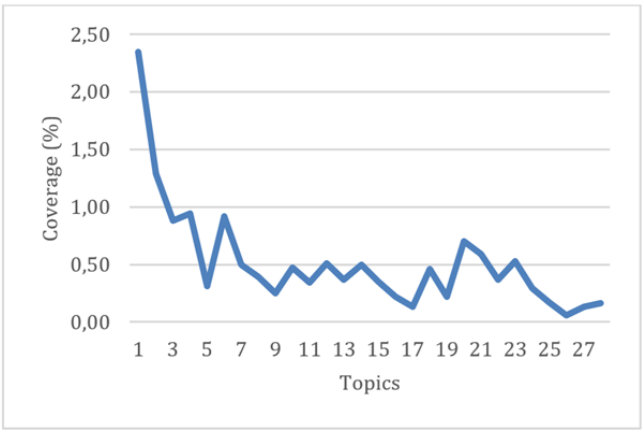
Table 2. Topics in negative reviews.

Topic number	Topic	Topic feature words	Cluster
1	Doctors	intern doctors, resident doctors, alone specialist doctors, emergency room doctors, previous doctors, procuring specialist doctors, prospective doctors, teachers, specialist doctors	Medical workforce
2	Specialists	lung specialist, alone specialist doctors, dermatology specialists, otolaryngology specialists, procuring specialist doctors, teachers, specialist doctors	Medical workforce
3	Students	medical student status, boarding student, dental students, nursing students	Educational system and training
4	Hospitals	vertical hospitals, hospital environment, hospital management, large hospitals	Health care institutions and patients
5	Teaching	teaching staff, teachers, specialist doctors	Medical workforce
6	Medicine	internal medicine, traditional medicine	Educational system and training
7	Programs	8th pregnancy program, expert master education program, study program	Educational system and training
8	Lungs	lung specialist, lung disease	Medical workforce
9	Nursing	nursing student, senior nurse	Medical workforce
10	Patients	sleep patients, trigger patients	Health care institutions and patients
11	Sleep	patient sleep, sleeping pills	Student life factors
12	Media	media briefings, social media	Media and public discourse
13	Vertical hospitals	vertical hospitals	Health care institutions and patients
14	Health	community health center, first-class health clinic	Health care institutions and patients
15	Education	expert master education program, medical education providers	Educational system and training
16	Expert master	expert master candidates, expert master education program	Educational system and training
17	Suitable	following suit, really suitable	Student life factors
18	Money	received money, scientific money	Financial and resource constraints
19	Medical student status	medical student status	Educational system and training
20	Internal doctors	internal doctors	Medical workforce
21	Results	good results, solid results	Educational system and training
22	Community	community health center, local community	Health care institutions and patients
23	Working	work strikes, working hours	Financial and resource constraints
24	Resident doctors	resident doctor	Medical workforce
25	Urine	urine bag, urine hunter	Financial and resource constraints
26	Smells	formalin smell, n't smell	Financial and resource constraints
27	Stages	stage compositor, stage rotation	Educational system and training
28	House	boarding house, halfway house	Financial and resource constraints

XSL•FO
RenderX

Figure 4. References and reviews (a) and coverage (b) for each negative-review topic.





Positive-Review Topic Coding

Autocoding carried out by NVivo 14 on positive reviews produced 17 topic codes, as summarized in Table 3.

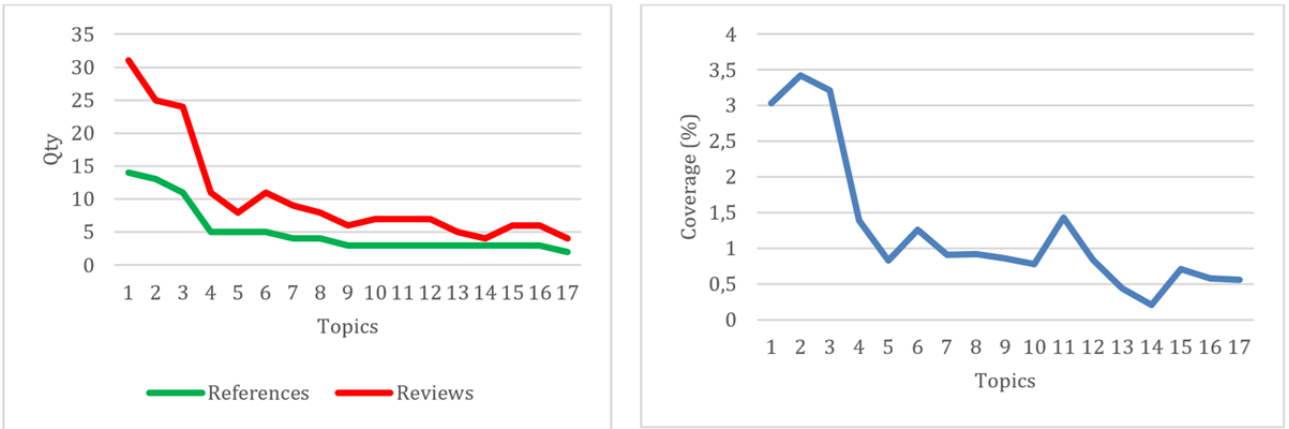
Based on the number of references and reviews (Figure 5a), the data showed significant variations in the number of references and reviews for each topic. Most topics had an equal number of references and reviews, but some had quite striking differences, such as the topics “doctors” and “services,” which had a higher number of reviews than the number of references. However, from a coverage perspective (Figure 5b), the majority of topics had quite high scores, indicating that these topics were widely discussed in the data. Topics such as “doctors” and “services” even had coverage values above 3% (17 reviews), indicating that they are the main focus of positive community reviews. Even though some topics had lower coverage values, the majority of them were still quite high, showing the variety of positive topics discussed in public conversations.

XSL•FO
RenderX

Table 3. Topics in positive reviews.

Topic number	Topic	Topic feature words	Operational definition
1	Doctors	professional doctor candidates, resident doctors, aspiring doctors, general doctors, eye specialists, online doctor apps	Mentions of doctors in a positive sense, reflecting professionalism, expertise, or accessibility
2	Services	superior service, pediatrician services, integrated services, rehab services	References to the quality, diversity, or efficiency of health care services
3	Students	medical students, students	Positive perception of medical students' role, contribution, or progress
4	Health	various health problems, health centers, health fields, health practitioners	General statements about health promotion, prevention, or community health benefits
5	Surgery	cardiac, cataract, orthopedic, preventive surgery	Praise for successful surgical procedures or advanced surgical options
6	Specialists	pediatric specialists, beauty specialists, eye specialists	Recognition of specialist expertise or availability in various fields
7	Treatment	beauty treatment, canal treatment, treating skin, treatment techniques	Appreciation of effective or innovative treatment modalities
8	Practitioners	general practitioner, health practitioner, health care practitioner	Positive portrayal of practitioners' skills and dedication
9	Facilities	equivalent facilities, sophisticated facilities, surgical facilities	Highlighting the adequacy, modernity, or comfort of health care facilities
10	Consultation	direct consultation, online doctor consultation application, psychological consultation services	Positive feedback on ease of access and quality of medical consultations
11	Superior service	superior service	Reinforcement of outstanding service quality in medical practice
12	Fields	health fields, medical fields, various fields	References to the diversity of medical and health disciplines
13	Beauty	beauty specialist, beauty treatment, well-maintained beauty	Positive associations with medical aesthetics and wellness
14	Rooms	physiotherapy rooms, reflection rooms, speech therapy rooms	Praise for the availability and functionality of medical or therapeutic rooms
15	Practice	specialist pediatrician practice services, practice schedule	Mentions of structured and accessible medical practices
16	Programs	23 study programs, male pregnancy program, professional program	Positive recognition of medical or educational programs
17	Choices	menu choices, the right choice	Appreciation of having diverse and appropriate health care or program options

Figure 5. References and reviews (a) and coverage (b) for each positive-review topic.



XSL•FO
RenderX

Cognitive Perspectives and Themes

Cognitive perspectives were established to help explain how society responds to the Icarus Paradox and how these perceptions influence their behavior regarding medical education. The cognitive perspective in the research context refers to the way individuals understand, interpret, and give meaning to information obtained from their environment. This includes how people process the information they obtain from online media, such as news articles, comments on social media, blogs, and discussion forums, as well as how their views are formed, influenced, and expressed through these platforms [27]. The combination of all topics in positive and negative reviews is then sorted to develop appropriate themes. Each corresponding theme is then grouped into a cognitive perspective. Based on all the reviews collected in this research, we defined four cognitive perspectives: *education system*, *policy*, *society's views of students*, and *health care services* (Table 4).

Table 4. Cognitive perspectives and themes by review topic.

Cognitive perspective and themes	Negative topics	Positive topics
Education system		
Professional challenges	• Topic 1: doctors • Topic 24: resident doctors • Topic 25: urine	• Topic 1: doctors • Topic 12: fields
Economic challenges	• Topic 18: money	—[a]
Workload	• Topic 19: medical student status • Topic 20: internal doctors	—
Medical understanding	• Topic 8: lungs	—
Teacher qualifications	• Topic 5: teaching • Topic 17: suitable	—
Medical education	• Topic 7: programs • Topic 13: vertical hospitals • Topic 15: education • Topic 16: expert master	• Topic 4: health • Topic 16: programs
Policy		
Government policy	• Topic 23: working • Topic 27: stages	—
Society's views of students		
About students	• Topic 3: students • Topic 28: houses	• Topic 3: students
Community involvement	• Topic 12: media	• Topic 6: specialists
Health care services		
Quality of health services	• Topic 2: specialists • Topic 4: hospitals • Topic 6: medicine • Topic 9: nursing • Topic 14: health • Topic 21: results • Topic 26: smells	• Topic 2: services • Topic 5: surgery • Topic 8: practitioners • Topic 9: facilities • Topic 11: superior service • Topic 13: beauty • Topic 15: practice
Patient experience	• Topic 10: patients • Topic 11: sleep • Topic 22: community	• Topic 7: treatment • Topic 10: consultation • Topic 14: rooms • Topic 17: choices

[a]Not applicable.

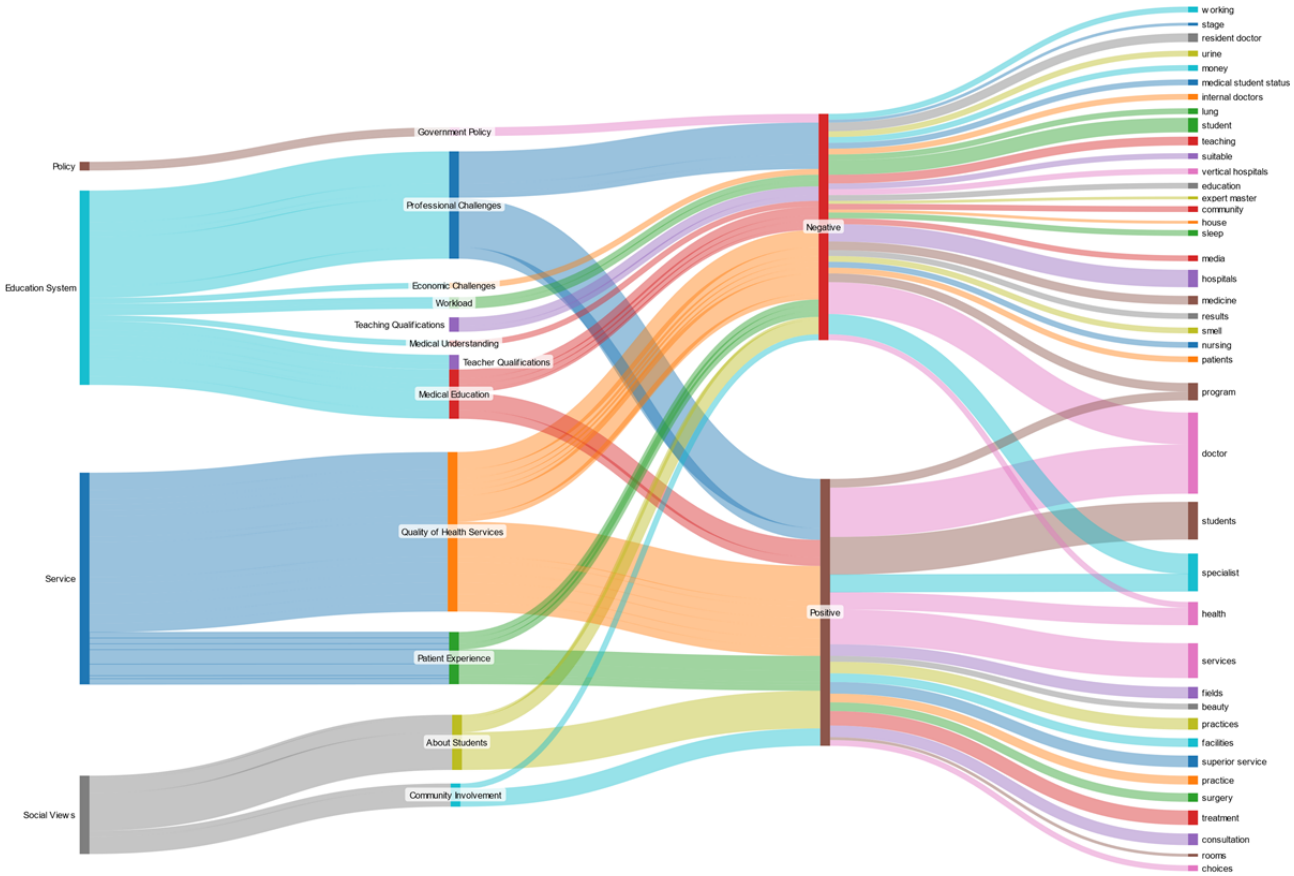
The "education system" perspective explains prospective doctors' direct experience with the medical education system and health facilities, which will shape their perceptions and assessments of the quality and effectiveness of the system. The information they obtain from online media can influence how they interpret and remember their own experiences. This perspective explains several themes, such as professional challenges, economic challenges, workload, medical

XSL•FO
RenderX

understanding, teaching qualifications, and medical education. The "policy" perspective explains the public's response to assessing the quality of medical education and health services, which is linked to the impact of government policy. Information about government policies, societal responses, and changes in the health system may influence the public's perceptions of the medical education system. This perspective focuses solely on the theme of government policy. Furthermore, the "society's views of students" perspective influences perceptions of medical education and the medical profession, shaped by the social support individuals receive from their environment. The perspectives of friends, family, and the medical community can influence how individuals understand and respond to the challenges they face in their medical education and careers. This perspective explains themes such as student views, as well as the themes of community engagement and public health. The "health care services" perspective explains public reviews regarding the experience of receiving services and the results obtained when visiting a doctor. This perspective explains themes such as the quality of health services and the patient experience. The entire cognitive perspective, along with its respective themes, is further explained in the coding process of the negative and positive reviews next.

Visualization of topic flow across cognitive perspectives, themes, sentiments, and issues was produced using a Sankey diagram (Figure 6). The diagram shows that five topics (*program*, *doctor*, *student*, *specialist*, and *health*) contained both positive and negative sentiments, indicating contradictory reviews and the presence of paradoxical dynamics. Three topics (*program*, *doctor*, *student*) displayed opposing sentiments within the same theme and perspective, whereas two topics (*specialist* and *health*) exhibited sentiment divergence across different themes or perspectives.

Figure 6. Sankey diagram of topic distribution for cognitive perspectives and themes.



The topic *program* centered on the theme of "medical education" within the "education system" perspective. Positive reviews emphasized the value of educational opportunities, such as clinical clerkships and diverse study programs offered by Indonesia's medical schools. These were seen as mechanisms to broaden competence and strengthen professional readiness. At the same time, negative sentiment highlighted recurring concerns, including bullying cases in the PPDS and the closure of certain specialties, such as radiology. These contradictory views—progressive ambition versus internal dysfunction—reflect the Icarus Paradox, where efforts to elevate educational excellence may simultaneously generate harmful systemic pressures [35,36].

The topic *doctors* related to "professional challenges" within the "education system" perspective. Positive sentiment acknowledged the achievements of doctors and medical students, recognition of expertise, and optimism surrounding innovations, such as online consultation platforms. However, negative sentiment pointed to structural barriers, such as shortages of specialist lecturers, harsh working conditions, low compensation for interns, and ethical concerns related to patient rights. These tensions illustrate the conflict between professional ideals and the realities imposed by institutional limitations.

The topic *students* was discussed under the "society's views of students" perspective. Positive sentiment highlighted recognition

XSL•FO
RenderX

of student accomplishments, the use of simulators and models in training, and hopes for producing compassionate and resilient physicians. Conversely, negative reviews drew attention to workload imbalance, ethical dilemmas when shifting roles, and perceptions of injustice in training environments. These competing narratives reflect the paradox of aspiring to excellence, while navigating a system that can undermine fairness and well-being.

Sentiment patterns for the topic *specialists* also revealed contrasting perspectives. Positive sentiment, categorized under "community involvement," underscored the indispensable role of specialists in delivering quality care, especially in underserved areas. Public comments highlighted trust in specialists' ability to accelerate recovery and address complex conditions. In contrast, negative sentiment, linked to the "health care services" perspective, emphasized limited specialist availability, particularly in remote regions, as well as funding and resource constraints. These shortages translate directly into service gaps, as illustrated by cases in which patients cannot obtain urgently needed specialist care.

Finally, the topic *health* contained both positive and negative sentiment across differing themes. Positive reviews, aligned with the "medical education" theme under the "education system" perspective, highlighted improvements in training quality, innovations in medical research, and the contribution of diverse specialists to system enhancement [63]. Negative sentiment, viewed from the "health care services" perspective, reflected dissatisfaction with access barriers, cost concerns, and perceived shortcomings in primary-level care. Public comments described challenges in securing treatment for family members and skepticism toward provider competence in community clinics.

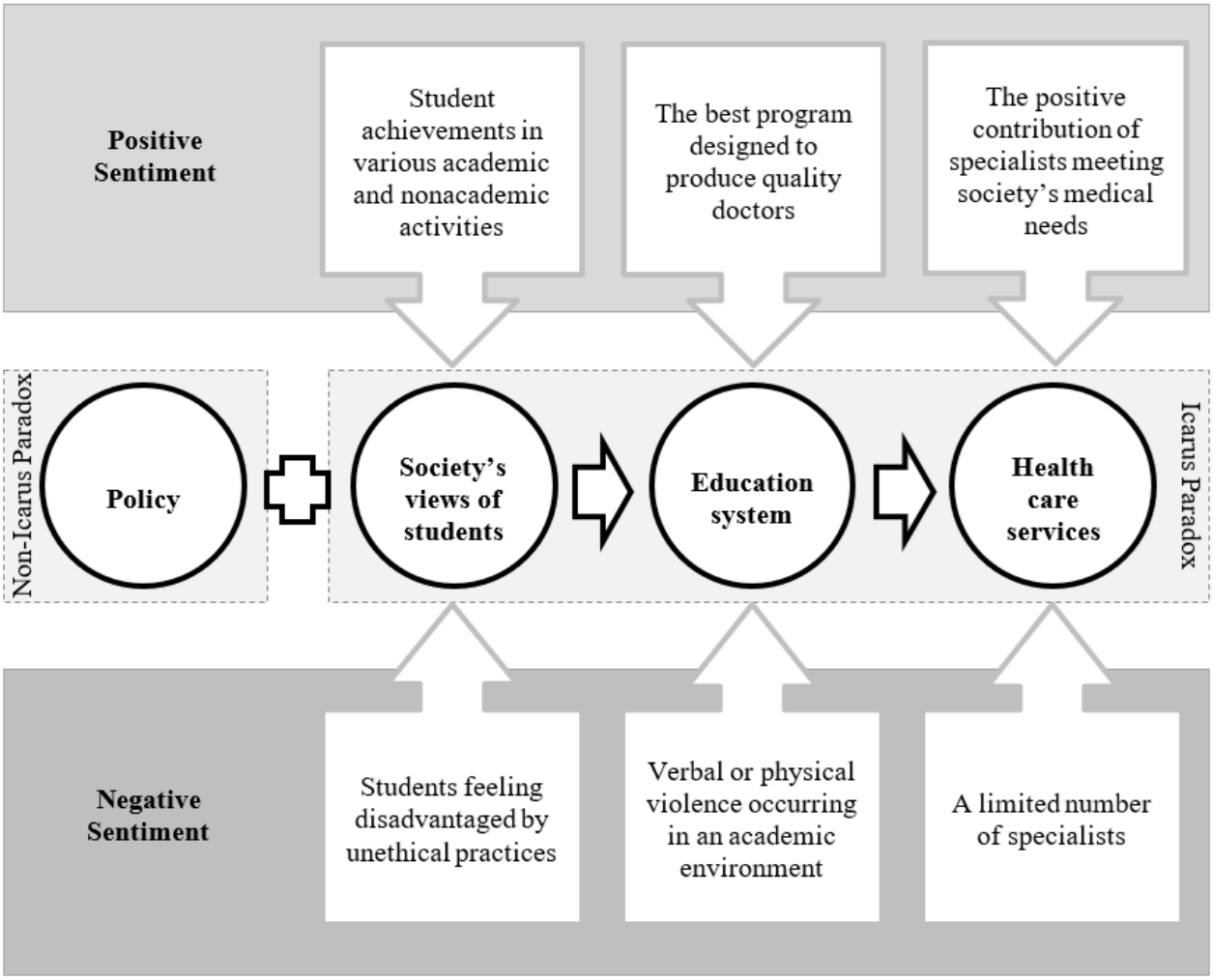
Collectively, these topic-based findings revealed how paradoxical tensions manifest simultaneously across multiple perspectives in Indonesia's specialist medical education ecosystem, providing critical insight into structural, educational, and service-related challenges.

Conceptual Model

Based on the results of our analysis, we proposed a conceptual model that links these four cognitive perspectives. This model can serve as a basis for formulating comprehensive policies and strategies to enhance medical education, improve public perception of the medical profession, and overall health services. With an integrated and holistic approach, it is hoped that the Icarus Paradox identified in this study can be overcome, thereby improving the overall quality of the health system. The proposed conceptual model is shown in Figure 7. Perspectives with white circles have the Icarus Paradox effect, while perspectives marked with gray circles do not have the Icarus Paradox effect. The "education system" perspective in this conceptual model is the primary component in producing quality health services, which is largely determined by the input and role of regulations, as well as "society's views of students" regarding prospective medical students.

XSL•FO
RenderX

Figure 7. Conceptual model linking the four cognitive perspectives in this study: (1) education system, (2) policy, (3) society's views of students, and (4) health care services.



Discussion

Principal Findings

This study identified three major cognitive perspectives—education system, health care services, and society's views of students—frequently reflected in public reviews, while policy issues appeared less prominently. Findings illustrate the Icarus Paradox as it emerges across these perspectives, where ambition and the pursuit of excellence coexist with constraints and systemic limitations. Rooted in classical interpretations of overreach and human limitation [37,38], the paradox becomes visible in how communities describe tensions within Indonesia's specialist medical education system. In contrast, the policy perspective generated comparatively uniform sentiment, suggesting limited public engagement with regulatory matters.

The Icarus Paradox in Specialist Medical Education

The Icarus Paradox describes how attributes that drive success can, when overstretched, evolve into vulnerabilities [7,37,38]. Indonesia's specialist medical education system reflects this dynamic. Programs such as clinical clerkships and PPDS training offer essential experiential learning and uphold high academic and professional standards [11,44,64]. These strengths, however, may escalate into risks when demands become excessive. Heavy workloads, hierarchical structures, and competitive environments contribute to burnout, intimidation, and uncertainty, exemplified by bullying cases and program closures, such as in radiology [10,11]. Public discourse captured in this study also shows admiration for educational excellence, alongside accounts of program instability and insufficient institutional support, illustrating how the system's strongest features can simultaneously generate tension.

Addressing these paradoxes requires structural and cultural reforms, including more inclusive learning environments, transparent governance, and stronger mentoring systems grounded in empathy and collaboration [65]. Updating curricula to align with evolving scientific and educational developments is equally vital. Such measures may help sustain academic quality, while minimizing the unintended consequences of overly rigid standards.

The Icarus Paradox in the Student Context

Among students, the paradox emerges when determination, resilience, and idealism—traits celebrated at the start of medical training—become sources of vulnerability under excessive pressure. Students often begin their education with strong

XSL•FO
RenderX

motivation and aspirations to contribute meaningfully to society [66]. Yet system demands, extended working hours, ethical dilemmas, and limited supervision due to dual lecturer-clinician roles can erode well-being and moral sensitivity [41]. Review data reveal disillusionment when expectations of professional fairness conflict with realities, such as inequitable pathways into practice or nonformal short-course alternatives.

Such tensions reflect how ambition without adequate support can lead to burnout and ethical distress. Strengthening mental health support, promoting fairness in training systems, and integrating ethical guidance are essential to prevent resilience from becoming a liability [65].

The Icarus Paradox in the Health Care Context

Paradoxes also materialize within health care delivery [67]. Advances in training, the expansion of simulations, and improvements in health facilities have contributed to better-prepared professionals and broader service coverage. Government initiatives, such as infrastructure development, *Jaminan Kesehatan Nasional* (JKN; National Health Insurance Program) implementation, and adoption of digital health and telemedicine, have improved access in many regions. However, unchecked growth and technological progress can introduce new weaknesses, including practitioner burnout, ethical compromises, or neglect of basic primary care, particularly in underserved communities [8,25].

Persistent disparities between urban and rural areas reflect this tension. Although capacity has increased nationally, specialist-to-population ratios and hospital bed availability remain considerably lower in remote regions [2,68-70]. These gaps demonstrate how reforms designed to elevate care quality can inadvertently reinforce inequality.

Addressing such paradoxes requires integrating community-based training, strengthening support for rural deployment, and ensuring equitable resource allocation [26]. Civil society involvement can help promote fairness and public accountability.

The “education system” perspective illustrates how efforts to sustain high academic standards and broaden learning opportunities in specialist training can inadvertently generate strain. Although diverse programs and research initiatives aim to enhance the quality and relevance of specialist medical education, these strengths may create vulnerabilities when demands exceed student capacity. Heavy academic loads, tight schedules, and competitive selection for placements or specialties can intensify stress and disrupt students’ balance between academic, social, and personal lives. Limited institutional attention to student well-being, insufficient support systems, and cultures that tolerate intimidation or academic mistreatment may further exacerbate these pressures, reinforcing the system’s paradoxical nature.

These dynamics reveal the tension between the ideal of academic excellence and the structural challenges embedded in training environments. Understanding this paradox is essential when evaluating how effectively the educational system prepares future doctors for real-world health care demands.

The “society’s views of students” perspective similarly reflects tensions between public expectations of medical professionalism and students’ lived realities. High academic pressure, limited access to qualified supervisors, financial burdens, and long working hours often conflict with students’ aspirations, contributing to fatigue and reduced well-being. These constraints show how ambition in medical training may collide with structural and ethical limitations, creating conditions consistent with the Icarus Paradox.

Although the “policy” perspective does not directly display paradoxical tension, it remains crucial in mitigating risks emerging from the other three perspectives. Well-designed regulations can support improvements in medical training quality, shape public perceptions of the profession, and promote equitable access to health services.

The “health care services” perspective represents the system’s final output and illustrates tensions between the need for high-quality health care and persistent resource constraints. Limited numbers and uneven distribution of specialists, especially in remote regions, hinder timely access to appropriate care. Funding shortages, inadequate human resources, and complex bureaucratic processes further complicate service delivery. These factors generate a paradox in which aspirations for excellent, equitable care conflict with systemic limitations [67].

Overall, the conceptual model highlights how progress within education, professional expectations, and health care delivery can unintentionally produce new vulnerabilities. Recognizing and managing these paradoxes is essential for creating a more resilient and equitable medical education and health care system.

Implications

The findings of this study carry significant implications for medical education, health service quality, and the role of specialist medical trainees [26]. First, within medical education, the Icarus Paradox, reflecting tensions between ambition and structural limitations, shapes student experiences and underscores the urgency of curriculum reform. Greater emphasis is needed on clinical skill development [71], communication, and leadership, which are essential for effective future practice. Educational institutions must also enhance academic and emotional support systems to help students manage stressors that arise during training.

Second, gaps in health service quality between urban and rural regions remain substantial. This highlights the need for increased investment in rural health infrastructure and the equitable placement of medical personnel. Elevating service standards nationwide requires consistent quality assurance, adequate staffing, and measures ensuring equal access to care.

Third, specialist medical trainees play a vital role as agents of change within both the education system and the broader health care landscape. Students must be equipped to navigate contemporary medical challenges and encouraged to engage in system improvement. Sufficient institutional support will enable them to contribute effectively to positive transformation in health services.

XSL•FO
RenderX

Fourth, the findings emphasize that oversight and governance in specialist medical education must be strengthened. Historically, weaknesses in program supervision have received attention only after issues were identified by the Ministry of Health. This implies a strong need for more rigorous monitoring by both government and universities.

Fifth, reliance solely on university-based specialist medical training may limit the number and diversity of specialists needed nationwide. Expanding hospital-based specialist medical training is therefore essential. Since the number of hospitals exceeds the number of medical schools, such an expansion could broaden access, improve specialist distribution to underserved regions, and reduce the training burden on universities. Increasing practitioner involvement in specialist medical education would also enrich training quality. Strengthening hospital-based pathways represents an important step toward addressing disparities and enhancing the national medical education system.

Overall, these implications highlight the need for systemic adjustments in medical education and health care delivery. Recognizing the presence of the Icarus Paradox helps identify areas for improvement and guides targeted reforms. Positive aspects, such as the growing role of hospital-based training, should be sustained and expanded, while structural weaknesses, such as inadequate governance and uneven specialist distribution, must be addressed. With joint commitment from the government, health care institutions, academics, and society, Indonesia can build a more inclusive, sustainable, and community-oriented health system [26], ultimately improving medical education and health service outcomes nationwide.

Limitations and Further Research

Despite offering valuable insights into the Icarus Paradox within medical education and health services, this study has several limitations. First, the findings rely heavily on online review data, which may not represent the full population. Individuals without internet access or those who do not use online review platforms may be underrepresented, and the anonymous nature of such reviews may affect data reliability.

Second, the study primarily captures community perceptions of the medical education system and specialist medical trainees, which introduces potential subjective bias. Different individuals may express varying experiences and expectations, which may influence sentiment data.

Third, this study does not directly measure clinical performance indicators or objective metrics of service quality. The data analyzed are predominantly qualitative and descriptive. To obtain a more comprehensive understanding of health service quality, future research should integrate quantitative measures with qualitative insights.

Further research can address these limitations in several ways. First, studies may incorporate broader respondent groups, including patients [25], medical practitioners, and students, through structured surveys or interviews to enhance representativeness. Second, advanced analytical techniques, such as deeper text mining or longitudinal sentiment tracking, can identify emerging patterns in digital public discourse. Third, expanding the study scope to examine interactions between variables, such as health care access, medical training quality, and patient satisfaction, over extended time periods (beyond 30 days) will enrich understanding.

Finally, empirical testing of the proposed conceptual model can offer deeper insights into the complex dynamics of Indonesia's health system. Such research can guide policymakers and educational institutions in designing interventions that enhance system effectiveness and mitigate paradoxical tensions. Ultimately, further investigation will contribute to a more comprehensive understanding of strategies needed to strengthen medical education and health care delivery in Indonesia.

Conclusion

Indonesia's specialist medical education system continues to face complex challenges that can be understood through the lens of the Icarus Paradox, where strengths that drive success may simultaneously give rise to failure. Ambition, resilience, and rigorous training enable medical students to pursue excellence, yet these same strengths can lead to burnout, intimidation, and psychological strain when not supported by adequate institutional resources. As Eriksson et al [7] describe, such dynamics reflect an "inverted Icarus" pattern in which progress carries inherent vulnerabilities. Prospective specialist medical students therefore stand at a critical juncture, symbolizing both the promise of advancement and the risks embedded in systemic pressure.

Health service quality also reflects this tension. Although national initiatives, such as infrastructure expansion and the JKN universal health coverage program, have increased access, disparities remain significant. For example, although the doctor-to-population ratio in Jakarta is approximately 1:1000, in some rural provinces, such as East Nusa Tenggara, the ratio drops below 1:7000 [72]. These inequalities illustrate how the push to expand services may exceed the capacity to ensure equitable quality of care. In education, national surveys indicate that more than half of PPDS students report mental health distress and excessive workloads, demonstrating that high standards often translate into substantial personal cost.

Addressing this paradox requires coordinated and targeted reforms. At the policy level, priority actions include (1) redistributing specialist doctors through structured incentives for underserved regions; (2) revising PPDS curricula to incorporate mental health resilience, mentorship, and leadership training; (3) expanding investment in simulation-based learning to reduce overreliance on hospital workloads; and (4) strengthening regulations against intimidation and unsafe learning environments.

At the same time, progress is attainable. Initiatives, such as scholarships for PPDS candidates from remote regions, the development of eLearning platforms in medical schools, and improvements in provincial teaching hospitals, indicate ongoing reform efforts. These programs show that although paradoxical tensions persist, Indonesia is gradually moving toward solutions that balance ambition with sustainable practice.

With collaborative efforts involving government, universities, hospitals, and society, the Icarus Paradox in medical education and health services can be mitigated. This will support

XSL•FO
RenderX

Indonesia's movement toward a more inclusive, equitable, and resilient health care system, one in which the aspiration to produce excellent medical professionals does not compromise their well-being or widen disparities in health service delivery.

Acknowledgments

The authors received no financial support for the research, authorship, or publication of this manuscript. This study did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Authors' Contributions

FB was responsible for conceptualization, methodology, data curation, formal analysis, writing—original draft, resources, supervision, and visualization and MH for conceptualization, validation, writing—review and editing, supervision, and formal analysis. All authors have read and approved the final manuscript.
Generative artificial intelligence (ChatGPT) was used only for language refinement and did not contribute to the analysis, interpretation, or generation of research data.

Conflicts of Interest

None declared.

References

1. Sulphey MM. How Icarus paradox doomed Kingfisher Airlines. Vision: J Bus Perspect 2020 Feb 28;24(1):118-124 [FREE Full text] [doi: 10.1177/0972262919897646]
2. Mahendradhata Y, Trisnantoro L, Listyadewi S, Soewondo P, Marthias T, Harimurti P, et al. The Republic of Indonesia health system review. In: Hort K, Patcharanarumol W, editors. Health Systems in Transition, Vol-7, Number-1. New Delhi: World Health Organization (Asia Pacific Observatory on Health Systems and Policies); 2017.
3. Wallin G. An introduction to R and Python for data analysis: a side-by-side approach. Am Stat 2024 Apr 17;78(2):265 [FREE Full text] [doi: 10.1080/00031305.2024.2320949]
4. Bondestam F, Lundqvist M. Sexual harassment in higher education – a systematic review. Eur J High Educ 2020 Feb 18;10(4):397-419 [FREE Full text] [doi: 10.1080/21568235.2020.1729833]
5. Heffernan T, Bosetti L. Incivility: the new type of bullying in higher education. Cambridge J Educ 2021 Mar 21;51(5):641-652 [FREE Full text] [doi: 10.1080/0305764x.2021.1897524]
6. Yousaf R, Schmiede R. Harassment Act implementation in higher education institutions. Open J Leadersh 2016;05(01):8-19 [FREE Full text] [doi: 10.4236/ojl.2016.51002]
7. Eriksson K, Lakomaa E, Nykvist R, Sandström C. Introducing the inverted Icarus paradox in business history – evidence from David and Goliath in the Swedish telecommunications industry 1981–1990. Bus Hist 2024 Mar 05;67(2):551-576 [FREE Full text] [doi: 10.1080/00076791.2023.2292134]
8. Desveaux L. Rethinking healthcare: why paradox science is core to the future of health and health leadership. JHL 2025 Feb 21;17:45-48 [FREE Full text] [doi: 10.2147/jhl.s512053]
9. Smith WK, Lewis MW. Toward a theory of paradox: a dynamic equilibrium model of organizing. Acad Manag Rev 2011 Apr;36(2):381-403 [FREE Full text] [doi: 10.5465/amr.2009.0223]
10. Lili R, Molodynski A, Farrell SM, Citraningtyas T, Kloping NA. Wellbeing and mental health among medical students in Indonesia: a descriptive study. Int J Soc Psychiatry 2022 Sep 25;68(6):1277-1282 [FREE Full text] [doi: 10.1177/00207640211057709] [Medline: 34823405]
11. Dyah Perwitasari V, Hidayat R. Depression level among neurology resident doctors in the Faculty of Medicine, Universitas Indonesia. Acta Neurol Indones 2024 Apr 17;2(01):60-71 [FREE Full text] [doi: 10.69868/ani.v2i01.25]
12. Arlinta D. Depresi, 3,3 Persen Calon Dokter Spesialis Ingin Akhiri Hidup atau Lukai Diri. Kompas. Jakarta: Kompas; 2024 Apr 15. URL: https://www.kompas.id/artikel/224-persen-mahasiswa-pendidikan-dokter-spesialis-alami-gejala-depresi-sebagian-merasa-lebih-baik-mengakhiri-hidup [accessed 2026-01-22]
13. Kesaksian calon dokter spesialis yang sempat berusaha bunuh diri – "perundungan dijustifikasi atas nama pendidikan mental. BBC News Indonesia. 2024 Apr. URL: https://www.bbc.com/indonesia/articles/cyj3mkp7jl0o [accessed 2026-01-20]
14. 2.716 Calon dokter spesialis mengalami gejala depresi, 3,3% ingin akhiri hidup. Metro TV News. 2024 Apr. URL: https://www.metrotvnews.com/play/b3JCr9d2-2-716-calon-dokter-spesialis-mengalami-gejala-depresi-3-3-ingin-akhiri-hidup [accessed 2026-01-20]
15. Widayanti AW, Green JA, Heydon S, Norris P. Health-seeking behavior of people in Indonesia: a narrative review. J Epidemiol Glob Health 2020 Mar;10(1):6-15 [FREE Full text] [doi: 10.2991/jegh.k.200102.001] [Medline: 32175705]
16. Martina Ws Nasrun. Depression Phenomenon in PPDS, Fact or Mirage? Kompas. Kompas: Kompas; 2024 May 10. URL: https://www.kompas.id/artikel/en-fenomena-depresi-di-ppds-fakta-atau-fatamorgana-1 [accessed 2026-01-21]

XSL•FO
RenderX

17. Ranasinghe PD, Owusu JT, Bertram A, Michtalik H, Yeh H, Cofrancesco J, et al. Depressive symptoms and burnout among medical students: a prospective study. J Gen Intern Med 2022 Jan 26;37(1):64-69 [FREE Full text] [doi: 10.1007/s11606-021-06765-x] [Medline: 34037922]
18. Doe S, Coutinho AJ, Weidner A, Cheng Y, Sanders K, Bazemore AW, et al. Prevalence and predictors of burnout among resident family physicians. Fam Med 2024 Mar 1;56(3):148-155 [FREE Full text] [doi: 10.22454/fammed.2024.875388]
19. Shalaby R, Oluwasina F, Eboreime E, El Gindi H, Agyapong B, Hrabok M, et al. Burnout among residents: prevalence and predictors of depersonalization, emotional exhaustion and professional unfulfillment among resident doctors in Canada. Int J Environ Res Public Health 2023 Feb 19;20(4):3677 [FREE Full text] [doi: 10.3390/ijerph20043677] [Medline: 36834373]
20. Obeng Nkrumah S, Adu MK, Agyapong B, da Luz Dias R, Agyapong VIO. Prevalence and correlates of depression, anxiety, and burnout among physicians and postgraduate medical trainees: a scoping review of recent literature. Front Public Health 2025 Jul 8;13:1537108 [FREE Full text] [doi: 10.3389/fpubh.2025.1537108] [Medline: 40697832]
21. Faghihzadeh E, Eghtesad A, Fawad M, Xu X. Exploring connections between mental health, burnout, and academic factors among medical students at an Iranian University: cross-sectional questionnaire study. JMIR Med Educ 2025 May 15;11:e58008 [FREE Full text] [doi: 10.2196/58008] [Medline: 40372966]
22. Lee KP, Yeung N, Wong C, Yip B, Luk LHF, Wong S. Prevalence of medical students' burnout and its associated demographics and lifestyle factors in Hong Kong. PLoS One 2020 Jul 10;15(7):e0235154 [FREE Full text] [doi: 10.1371/journal.pone.0235154] [Medline: 32649681]
23. Thamissarakul S, Hongkan W, Wannapaschaiyong P. Factors associated with burnout syndrome among clinical medical students at Chonburi Hospital, Thailand. J Health Sci Med Res 2024;42(3):1-6 [FREE Full text] [doi: 10.31584/jhsmr.20231020]
24. Caring for the mental health of the medical workforce. British Medical Association. 2019. URL: https://www.bma.org.uk/media/ckshvkzc/bma-mental-health-survey-report-september-2024.pdf [accessed 2026-01-20]
25. Berger S, Saut AM, Berssaneti FT. Using patient feedback to drive quality improvement in hospitals: a qualitative study. BMJ Open 2020 Oct 23;10(10):e037641 [FREE Full text] [doi: 10.1136/bmjopen-2020-037641] [Medline: 33099495]
26. Bismantara H, Ahern S, Teede HJ, Liew D. Academic health science centre models across the developing countries and lessons for implementation in Indonesia: a scoping review. BMJ Open 2022 Sep 06;12(9):e051937 [FREE Full text] [doi: 10.1136/bmjopen-2021-051937] [Medline: 36691121]
27. Abid A, Harrigan P, Roy S. A relationship marketing orientation in politics: young voters' perceptions of political brands' use of social media. Journal of Strategic Marketing 2020 Jun 12;29(4):359-374 [FREE Full text] [doi: 10.1080/0965254x.2020.1777457]
28. Ashraf MA, Khan MN, Chohan SR, Khan M, Rafique W, Farid MF, et al. Social media improves students' academic performance: exploring the role of social media adoption in the open learning environment among international medical students in China. Healthcare (Basel) 2021 Sep 26;9(10):1272 [FREE Full text] [doi: 10.3390/healthcare9101272] [Medline: 34682952]
29. Noya FC, Carr SE, Thompson SC. Attracting, recruiting, and retaining medical workforce: a case study in a remote province of Indonesia. Int J Environ Res Public Health 2023 Jan 12;20(2):1435 [FREE Full text] [doi: 10.3390/ijerph20021435] [Medline: 36674191]
30. Schad J, Lewis MW, Raisch S, Smith WK. Paradox research in management science: looking back to move forward. Acad Manag Ann 2016 Jan;10(1):5-64 [FREE Full text] [doi: 10.5465/19416520.2016.1162422]
31. West J, Gallagher S. Challenges of open innovation: the paradox of firm investment in open-source software. R D Manag 2006 Jun;36(3):319-331 [FREE Full text] [doi: 10.1111/j.1467-9310.2006.00436.x]
32. Oliveira P, e Cunha MP. Centralized decentralization, or distributed leadership as paradox: the case of the Patient Innovation's COVID-19 portal. J Chang Manag 2021 May 09;21(2):203-221 [FREE Full text] [doi: 10.1080/14697017.2021.1917493]
33. Klein SP, Spieth P, Söllner M. Employee acceptance of digital transformation strategies: a paradox perspective. J Prodt Innov Manag 2024 Jan 10;41(5):999-1021 [FREE Full text] [doi: 10.1111/jpim.12722]
34. Cunha MPE, Putnam LL. Paradox theory and the paradox of success. Strateg Organ 2017 Oct 31;17(1):95-106 [FREE Full text] [doi: 10.1177/1476127017739536]
35. Miller D. The Icarus paradox: how exceptional companies bring about their own downfall. Bus Horizons 1992 Jan;35(1):24-35 [FREE Full text] [doi: 10.1016/0007-6813(92)90112-M]
36. Dowds B. Depression and the Erosion of the Self in Late Modernity. London, New York: Routledge, Taylor and Francis; 2018.
37. Sherrer D, Franklin A, Kimatian S, Black I, Tsai M. The Icarus paradox and the future of anesthesiology. Anesth Analg 2023 Jan 01;136(1):185-189 [FREE Full text] [doi: 10.1213/ANE.0000000000006253] [Medline: 36534720]
38. Biçer C. The Icarus paradox in management: how to be a well-balanced leader? Nevşehir Hacı Bektaş Veli Üniversitesi SBE Derg 2021;11(4):1891-1905 [FREE Full text] [doi: 10.30783/nevsosbilen.892797]
39. Vermeulen F. Businesses and the Icarus paradox. Harvard Business Review. 2009 Mar 5. URL: https://hbr.org/2009/03/businesses-and-the-icarus-para [accessed 2026-01-20]

XSL•FO
RenderX

40. Zreik M. The paradox of educational inequality in Indonesia: socioeconomic implications and paths towards inclusion. Socio-Economic Implic Glob Educ Inequalities 2024:69-85 [FREE Full text] [doi: 10.4018/979-8-3693-0693-2.ch004]
41. Winzer R, Lindberg L, Guldbrandsson K, Sidorchuk A. Effects of mental health interventions for students in higher education are sustainable over time: a systematic review and meta-analysis of randomized controlled trials. PeerJ 2018;6:e4598 [FREE Full text] [doi: 10.7717/peerj.4598] [Medline: 29629247]
42. Rotenstein LS, Ramos MA, Torre M, Segal JB, Peluso MJ, Guille C, et al. Prevalence of depression, depressive symptoms, and suicidal ideation among medical students: a systematic review and meta-analysis. JAMA 2016 Dec 06;316(21):2214-2236 [FREE Full text] [doi: 10.1001/jama.2016.17324] [Medline: 27923088]
43. Shapiro J. The paradox of teaching empathy in medical education. In: Decety J, editor. Empathy: From Bench to Bedside. Cambridge, MA: MIT Press; 2011:275-290.
44. Brown T, Williams B, McKenna L, Palermo C, McCall L, Roller L, et al. Practice education learning environments: the mismatch between perceived and preferred expectations of undergraduate health science students. Nurse Educ Today 2011 Nov;31(8):e22-e28 [FREE Full text] [doi: 10.1016/j.nedt.2010.11.013] [Medline: 21131106]
45. Tess PA. The role of social media in higher education classes (real and virtual) – a literature review. Comput Hum Behav 2013 Sep;29(5):A60-A68 [FREE Full text] [doi: 10.1016/j.chb.2012.12.032] [Medline: 21131106]
46. Sivakumar A, Jayasingh S, Shaik S. Social media influence on students' knowledge sharing and learning: an empirical study. Educ Sci 2023 Jul 20;13(7):745 [FREE Full text] [doi: 10.3390/educsci13070745]
47. Wilson C, McDarby V. Social media and mental health. Clin Child Psychol Psychiatry 2023 Jan 28;28(1):157-160 [FREE Full text] [doi: 10.1177/13591045221144926] [Medline: 36579367]
48. Firdos S, Almulla S, Aldossary S, Al Hassan S, Aldhaif L. Exploring the attitudes of medical students towards social media and e-professionalism in Al-Ahsa, Saudi Arabia. Cureus 2023 Nov;15(11):e48718 [FREE Full text] [doi: 10.7759/cureus.48718] [Medline: 38094527]
49. Terry K, Yang F, Yao Q, Liu C. The role of social media in public health crises caused by infectious disease: a scoping review. BMJ Glob Health 2023 Dec 28;8(12):e013515 [FREE Full text] [doi: 10.1136/bmjgh-2023-013515] [Medline: 38154810]
50. Binsar F, Arief M, Tjhin VU, Susilowati I. Exploring consumer sentiments in telemedicine and telehealth services: towards an integrated framework for innovation. J Open Innov Technol Mark Complex 2025 Mar;11(1):100453 [FREE Full text] [doi: 10.1016/j.joitmc.2024.100453]
51. Kim D, Jung W, Jiang T, Zhu Y. An exploratory study of medical journal's Twitter use: metadata, networks, and content analyses. J Med Internet Res 2023 Jan 19;25:e43521 [FREE Full text] [doi: 10.2196/43521] [Medline: 36656626]
52. Creswell JW, Creswell JD. In: Felts DC, editor. Research Design: Qualitative, Quantitative, and Mixed Methods Approaches. Fifth Edition. Thousand Oaks, CA: Sage Publications; 2018.
53. Hutagalung SS, Kartika T, Suciska W. Media monitoring analysis of government image in infrastructure development in Indonesia. J Komun 2023 Jul 27;15(1):212-227 [FREE Full text] [doi: 10.24912/jk.v15i1.20605]
54. Hadi SP, Ibrahim MH, Prabawani B, Hamdani RS. Environmental dimension of pandemic COVID-19: case studies of Indonesia. IOP Conf Ser Earth Environ Sci 2021 Jan 01;623(1):012051 [FREE Full text] [doi: 10.1088/1755-1315/623/1/012051]
55. Punziano G, De Falco CC, Trezza D. Digital mixed content analysis for the study of digital platform social data: an illustration from the analysis of COVID-19 risk perception in the Italian Twittersphere. J Mix Methods Res 2023 Apr 21;17(2):143-170 [FREE Full text] [doi: 10.1177/15586898211067647] [Medline: 38603067]
56. Chhabra J, Pilkington V, Benakovic R, Wilson MJ, La Sala L, Seidler Z. Social media and youth mental health: scoping review of platform and policy recommendations. J Med Internet Res 2025 Jun 20;27:e72061 [FREE Full text] [doi: 10.2196/72061] [Medline: 40540734]
57. Zhou Z, Jin D, He J, Zhou S, Wu J, Wang S, et al. Digital health platform for improving the effect of the active health management of chronic diseases in the community: mixed methods exploratory study. J Med Internet Res 2024 Nov 18;26:e50959 [FREE Full text] [doi: 10.2196/50959] [Medline: 39556830]
58. Binsar F, Mauritsius T. Mining of social media on Covid-19 big data infodemic in Indonesia. J Comput Sci 2020 Nov 01;16(11):1598-1609 [FREE Full text] [doi: 10.3844/jcssp.2020.1598.1609]
59. Lee YU, Chung SH, Park JY. Online review analysis from a customer behavior observation perspective for product development. Sustainability 2024 Apr 24;16(9):3550 [FREE Full text] [doi: 10.3390/su16093550]
60. Limna P. The impact of NVivo in qualitative research: perspectives from graduate students. J Appl Learn Teach 2023 Aug 2;6(2):271-282 [FREE Full text] [doi: 10.37074/jalt.2023.6.2.17]
61. Dereń K, Rajda K. How to master AI-powered sentiment analysis in 2025? Brand24. 2025. URL: https://brand24.com/blog/sentiment-analysis/ [accessed 2025-08-19]
62. Jackson K, Bazeley P. Qualitative Data Analysis with NVivo. 3rd Edition. London, California, New Delhi, Singapore: Sage Publications; 2019:1-346.
63. Binsar F, Legowo N. Design of cloud computing outpatient registration model through SMS messages at hospitals using TOGAF ADM. Int J Recent Technol Eng 2020 Jan 30;8(5):3857-3865 [FREE Full text] [doi: 10.35940/ijrte.E4885.018520]

XSL•FO
RenderX

64. Binsar F, Kartono R, Bandur A, Kosasih W. Digital transformation of information fulfillment and patient engagement for health service safety. 2022 Jul 18 Presented at: 4th International Conference on Management Science and Industrial Engineering; April 28-30, 2022; Chiang Mai, Thailand p. 229-236. [doi: 10.1145/3535782.3535813]
65. Mustika R, Nishigori H, Ronokusumo S, Scherpbier A. The odyssey of medical education in Indonesia. Asia Pacific Sch 2019 Jan 2;4(1):4-8 [FREE Full text] [doi: 10.29060/taps.2019-4-1/gp1077]
66. Zechariah S, Ansa BE, Johnson SW, Gates AM, Leo GD. Interprofessional education and collaboration in healthcare: an exploratory study of the perspectives of medical students in the United States. Healthcare (Basel) 2019 Oct 15;7(4):117 [FREE Full text] [doi: 10.3390/healthcare7040117] [Medline: 31618920]
67. Hofmann B. The paradox of health care. Health Care Anal 2001 Dec;9(4):369-386 [FREE Full text] [doi: 10.1023/a:1013854030699]
68. Sutrisno MD. Shortages of medical doctors in indonesia, is it true? Asian J Heal Res 2023 Aug 11;2(2):2828-2829 [FREE Full text] [doi: 10.55561/ajhr.v2i2.121]
69. Risky YPA, Projo NWK. The effect of the digital economy on indecent work in Indonesia 2019. Jakarta; 2022 Jan 04 Presented at: 2021 International Conference on Data Science and Official Statistics (ICDSOS); November 13-14, 2021; Jakarta, Indonesia. [doi: 10.34123/icdsos.v2021i1.88]
70. Awofeso N, Rammohan A, Asmaripa A. Exploring Indonesia's "low hospital bed utilization-low bed occupancy-high disease burden" paradox. J Hosp Adm 2012 Nov 22;2(1):49 [FREE Full text] [doi: 10.5430/jha.v2n1p49]
71. Binsar F, Riantono I, Kartono R, Bandur A, Kosasih W. Gamification to grow motivation for interactive engagement of health nurses in using health information systems: a conceptual framework. 2023 Presented at: Asia Pacific Computer Systems Conference 2021; February 18, 2023; North Carolina, USA p. 99-120. [doi: 10.1007/978-981-19-7904-0_10]
72. Putri LP, Russell DJ, O'Sullivan BG, Kippen R. Factors associated with working in remote Indonesia: a national cross-sectional study of early-career doctors. Front Med (Lausanne) 2021 May 13;8:594695 [FREE Full text] [doi: 10.3389/fmed.2021.594695] [Medline: 34055819]

Abbreviations

JKN: Jaminan Kesehatan Nasional (National Health Insurance Program)
PPDS: *Program Pendidikan Dokter Spesialis* (Specialist Doctor Education Program)

Edited by J Gentges; submitted 11.May.2024; peer-reviewed by PH Sebong, NA Pambudi; comments to author 15.May.2025; accepted 17.Sep.2025; published 26.Jan.2026.

Please cite as:
Binsar F, Hamsal M
Exploring the Icarus Paradox in Indonesia's Specialist Medical Education System Using the Public Perspective From Online Media: Convergent Mixed Methods Study
JMIR Med Educ 2026;12:e60452
URL: https://mededu.jmir.org/2026/1/e60452
doi:10.2196/60452
PMID:

©Faisal Binsar, Mohammad Hamsal. Originally published in JMIR Medical Education (https://mededu.jmir.org), 26.Jan.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Education, is properly cited. The complete bibliographic information, a link to the original publication on https://mededu.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

AI-Enhanced Continuing Professional Development as an Evolving Sociotechnical System: Multimethod Theoretical Framework Development Study

Vjekoslav Hlede[1], DVM, PhD; Sofia Valanci[2], MD, PhD; G Robert D'Antuono[3], MHA; Heather Dow[4], BA; Ronan O'Beirne[5], EdD; Richard Wiggins[6], MD

[1]Learning Division, American Society of Anesthesiologists, 1061 American Lane, Schaumburg, IL, United States

[2]Office of Learning and Connecting, Royal College of Physicians and Surgeons of Canada, Ottawa, ON, Canada

[3]Office of Continuing Medical Education, NYU Grossman Long Island School of Medicine, New York, NY, United States

[4]Canadian Association of Physical Medicine and Rehabilitation, Kingston, ON, Canada

[5]Division of Continuing Medical Education, Marnix E Heersink School of Medicine, The University of Alabama at Birmingham, Birmingham, AL, United States

[6]Department of Continuing Medical Education, University of Utah, Salt Lake City, UT, United States

Corresponding Author:
Vjekoslav Hlede, DVM, PhD
Learning Division, American Society of Anesthesiologists, 1061 American Lane, Schaumburg, IL, United States

Abstract

Background: Artificial intelligence (AI) is changing continuing professional development (CPD) in health care and its interactions with the broader health care system. However, current scholarship lacks an integrated theoretical model that explains how AI impacts CPD as a complex sociotechnical system. Existing frameworks usually focus on isolated phenomena, such as ethics, literacy, or learning theory, leaving unaddressed the dynamics of how those phenomena interact in the complex sociotechnical AI-enhanced CPD system, as well as the new roles that AI-empowered patients and society play.

Objective: The objective of this study is to propose a comprehensive, theory-driven framework that provides insight into how AI transforms CPD systems. The goal was to integrate established AI constructs with Complexity Theory (CT) and Actor-Network Theory (ANT) to develop a model that guides practice, research, and policy.

Methods: We conducted a multimethod theory construction. The process started with identifying the AI-enhanced CPD as an established yet evolving phenomenon. Through a structured literature review, the main building blocks of AI-enhanced CPD were identified, as well as the ontological base (CT and ANT). The model was developed through iterative human-led and AI-assisted abductive analysis. The final model was abductively validated on a case study of a national organization pioneering AI use, demonstrating the theoretical model makes sense in practice. All conceptual decisions were reviewed collaboratively by the author group.

Results: The ALEERRT-CA framework is made of 6 pillars: AI literacy, explainability, ethics, readiness, reliability, and learning theories, and 2 theoretical lenses: CT and ANT. CT elucidates macro-level system behaviors in the AI-enhanced CPD system. Those behaviors include emergence, feedback loops, adaptation, and reality made of nested complex systems. ANT explains how localized interactions among human and nonhuman actors shape AI-enhanced CPD. Together, these lenses illustrate how AI redistributes agency, amplifies tensions, and generates emergent learning dynamics within CPD and the broader health care system.

Conclusions: This study presents a novel conceptual model of AI-enhanced CPD as a sociotechnical system. The integration of CT and ANT with AI constructs improves explanatory power of the ALEERRT-CA framework. Educators, program leaders, and policymakers can use the framework as a structured toolset to evaluate AI readiness, design responsible AI-enhanced CPD practices, and plan future empirical research. The framework provides a theoretical lens for observing the rapidly evolving field of AI-enhanced CPD and health care practice.

(JMIR Med Educ 2026;12:e69156) doi:10.2196/69156

KEYWORDS

artificial intelligence; AI; AI-enhanced CPD; AI explainability; AI readiness; AI reliability; learning theories; complexity; Actor-Network Theory; black box; continuing professional development

XSL•FO
RenderX

Introduction

AI as a Transformative Sociotechnical Force in Continuing Professional Development

Artificial intelligence (AI) applications are increasingly impacting all aspects of the health care system, from clinical practice (diagnosis, treatment, prevention) to research, communication, administration, and learning [1]. The super-connected, postdigital nature of our society [2] enables AI tools to access vast amounts of data and allows their impacts to spread quickly globally, amplifying AI's power. AI's rapid evolution is associated with ethical, legal, social, and professional opportunities and risks for health and continuing professional development (CPD) professionals, patients, and the broader society [3,4]. Therefore, we need to increase our capacity to analyze and improve AI-enhanced CPD and the broader sociotechnical systems with which it is associated [5].

AI tools are gaining transformative powers. They act as a mediating technology that changes how information flows, decisions are made, and how CPD occurs [6]. They are not just new tools improving existing CPD routines; they modify the relationships between clinician-learners, educators, patients, technologies, organizational structures, and regulators.

From an actor-network perspective, AI acts as a nonhuman actor that reorganizes agency, influences interactions, and supports transformations of established professional practices [7]. For example, the introduction of AI into CPD environments contributes to nonlinear, emergent system behavior characteristic of complex sociotechnical adaptive systems, in which small changes in one part of the network can propagate unpredictably across broader educational and clinical contexts [8].

That phenomenon of AI-empowered patients illustrates this transformation. It is reshaping the clinician-patient relationship, influencing decision-making processes, redefining the learning needs of health care professionals, the health system, and the public [9,10]. Those changes can significantly affect patient outcomes.

Gap, Aim, and Research Question

Despite the increasing importance of AI-enhanced CPD, we lack a framework for examining it as a sociotechnical system. Most published approaches focus on specific issues like ethics, literacy, or digital learning, while leaving out complex relationships between people and technology, as well as the feedback and system dynamics that influence CPD and its role in the broader health care ecosystem.

This study aims to address this by creating an accessible, theory-based framework to analyze how AI changes networks of people and technology, introduces new feedback loops, and transforms CPD systems. Using a multimethod approach based on Borsboom et al [11], we bring together Complexity Theory (CT) and Actor-Network Theory (ANT) to study AI-enhanced CPD. Our primary research question is: How can we best explain AI-enhanced CPD as a sociotechnical system, and what new system-level dynamics and relationships appear when we look at it through both CT and ANT?

This paper makes three contributions. First, it conceptualizes AI-enhanced CPD as a complex sociotechnical system rather than a set of isolated tools. Second, it integrates CT and ANT into a single explanatory framework (ALEERRT-CA). That connects macro-level system dynamics with micro-level human–AI interactions. Third, it provides a reusable analytic tool for educators, organizations, and researchers to evaluate AI readiness, governance, and learning design in CPD contexts.

Literature Review

AI-enhanced CPD does not emerge in isolation. Our increasingly networked society and health care systems are characterized by increasing interdependence, speed, and systemic complexity [12]. They create a context inseparable from AI-enhanced CPD. In this context, professional learning is no longer bounded by formal educational activities or institutional settings [13]. Instead, AI-enhanced CPD is interwoven with clinical workflows, digital infrastructures, regulatory environments, patient participation, and continuously evolving technologies [14].

This qualitatively changes how CPD functions. Traditional linear, tool-centered, or competency-based models are increasingly insufficient to explain how AI reshapes learning, decision-making, and professional practice across interconnected health care systems [15,16].

This literature review, therefore, approaches AI-enhanced CPD through the dual lenses of complex adaptive systems (CASs) [17,18] and ANT [19]. Rather than treating AI as an isolated educational innovation, we examine AI-enhanced CPD as a dynamic system embedded in a networked society—one in which relationships between human and nonhuman actors, feedback loops, and emergent behaviors are central. This perspective indicates the need for an integrative framework that explains AI's systemic influence on CPD.

Changing Complex Systems

CPD acts as an open, adaptive, complex system (CAS) [20,21]. CPD meets established CAS criteria not because of its educational content alone, but because (1) it is embedded within a complex health care system and (2) it tackles complex individual-, team-, and organization-wide learning processes. CPD is structurally connected to clinical workflows, organizational incentives, regulatory regimes, and patient outcomes. CPD is focused on addressing the contextual learning needs of individuals, teams, and organizations and is shaped by emerging technologies and clinical practices. AI adds additional layers of complexity [22]. It provides new variables to the CPD system and connects it more dynamically with the broader health care system.

Networked Society: A New Context

Initiatives using AI in education have a long history [23]. From teaching machines in the 1920s to expert systems in the 1970s to 1990s, AI tools existed as (semi)isolated centers, relying almost entirely on rules designed by human experts. Now, our society's super-connected, big-data nature creates a considerably different context. From (semi)isolated centers with limited input

XSL•FO
RenderX

of data, AI has become omnipresent and empowered by the global internet of knowledge [24].

Whether we talk about personalized search findings, text editor suggestions, AI-generated meeting minutes, health monitoring apps, 1357 AI-enhanced medical devices recognized by the US Food and Drug Administration [25], learning analytics, or health care quality improvement [26], AI tools are here, tightly embedded in our daily practices.

There Is Nothing So Practical as a Good Theory

Despite the close relationship between AI, individuals, and society, research on AI in education and CPD has been mostly technocentric and atheoretical [27,28]. Technocentric research may miss the complex interactions between AI, CPD, and broader health care contexts [29].

As we aim to understand how AI impacts CPD activities and the broader sociotechnical health care system, our focus must shift from individual phenomena of AI toward the interaction between AI and various phenomena and activities in the complex context of health care CPD [30]. Theoretical frameworks help us make that shift, allowing us to understand better and manage interactions between various elements of the complex system [31].

Kurt Lewin, one of the pioneers of organizational and social psychology, famously noted, “There is nothing so practical as a good theory” and “the best way to understand something is to try to change it” [32,33]. Guided by those maxims, we propose a theoretical framework to better comprehend and guide the integration of AI and CPD. This is a distinct change within the CPD domain; it creates an opportunity to better understand and improve CPD within the broader sociotechnical context [34,35].

Why an Integrative Framework Is Needed

Mouloudj et al [36] review of AI adoption in health care has revealed a fragmented landscape of success factors and barriers. Numerous studies describe the importance of AI literacy, explainability, trust, ethical safeguards, organizational readiness, workflow compatibility, social influence, and system reliability. However, these determinants are presented as isolated pillars rather than tightly interconnected components of an AI-enhanced sociotechnical system. In practice, these elements can interact very dynamically. For example, when AI enters a clinical CPD environment, it affects learning design, decision-making processes, roles, relationships, and system behavior.

Since CPD is a CAS inseparable from the complexity of the broader health care system, we need a lens that lets us observe the system and relationships in the system. Focusing on isolated fragments of the system will not provide the needed insight into the system and how AI reshapes it [30].

This creates a gap. We know what forces influence the adoption of AI and that, if properly used, AI can have a transformative positive impact [4]. However, we lack a coherent framework describing how these forces interact across individual, team, organizational, and system levels. AI is being transformed from a tool that supports existing practice into a set of tools that are transforming our sociotechnical system. It reorganizes workflows, amplifies feedback loops, and alters the learning system itself. Therefore, it is becoming increasingly important to have a theoretical lens that can track that transformation.

The ALEERRT-CA framework responds to this need. It brings together 6 foundational AI pillars and situates them within CT and ANT, providing a practical and theoretically grounded way to analyze how AI operates within CPD as a sociotechnical system. By articulating the “anatomy” of AI-enhanced CPD and offering tools to examine both macro-level patterns and micro-level interactions, ALEERRT-CA provides the conceptual infrastructure needed for responsible innovation, rigorous research, and evidence-informed CPD strategy.

Methods

Methodology: Theory Construction

This study adopts a theory construction methodology rather than an empirical evaluation design, with the goal of producing an explanatory framework suitable for subsequent empirical testing.

We employed the theory construction methodology outlined by Borsboom et al [11]. This process included the following 5 steps:

1. Identification of Phenomenon: We identified AI-enhanced CPD as a relevant, explainable, and reproducible phenomenon that is rapidly evolving but also stable in terms of its broad and continued impact on CPD [37,38].
2. Drafting Core Principles and Models: VH drafted core principles to explain the phenomenon and created multiple explanatory models using abductive reasoning (explanatory inference). ChatGPT assisted with initial brainstorming and concept review. After concepts were refined, the draft was shared with the author group (2023).
3. Model Development: After multiple additional iterations and contributions from all authors, the final model was developed.
4. Assessment of Explanatory Adequacy: As the fourth step, we assessed the adequacy of the framework and its ability to explain how AI-enhanced CPD can enhance learning.
5. Evaluation of Theoretical Framework: We concluded the process by evaluating the value of the constructed theoretical framework through 2 simulated scenarios and a real-world case study.

Complexity-Ready Lenses

Overview

AI-enhanced CPD occurs in the open, adaptive, complex sociotechnical health care system, where many elements are on the edge of chaos [21,30]. Working with complex, constantly evolving phenomena requires tolerance for ambiguity, a shift of focus from individual phenomena to interactions among multiple phenomena within the system, and the development of unique theoretical tools [30,39]. CT and ANT, 2 complementary theoretical lenses, were used to examine the sociotechnical health care system and the role of AI within it.

XSL•FO
RenderX

Selection of the 6 Foundational Pillars

The 6 AI pillars (Literacy, Explainability, Ethics, Readiness, Reliability, and Learning Theories) were selected through an iterative process of pillar identification (Textbox 1). Selection is based on 3 criteria: necessity, nonredundancy, and system relevance. The necessity criterion stipulates that each pillar addresses an important mechanism through which AI influences CPD and the broader health care system. Nonredundancy ensures that no pillar duplicates another pillar's explanatory function. System relevance requires that each pillar can meaningfully participate in macro-level system dynamics (CT) and micro-level actor interactions (ANT).

Textbox 1. Construct selection logic (ALEERRT-CA)

1. Literature frequency and significance: Pillars repeatedly cited across AI, health care, CPD, and education fields
2. Distinct explanatory role: Each pillar explains a unique facet of AI-CPD interaction
3. Alignment with CT and ANT: Must connect to system-level dynamics and actor-level mechanisms
4. Parsimony: Remove overlapping or derivative constructs; retain essential ones

Final Pillars: Literacy, Explainability, Ethics, Readiness, Reliability, and Learning Theories

Constructs frequently mentioned in literature but treated as subsets of broader domains (eg, trust, governance, data quality) were incorporated under more comprehensive theoretical pillars. The final 6 pillars represent the minimal set needed to explain AI's influence on CPD within a sociotechnical system while maintaining parsimony.

In ALEERRT-CA, learning theories serve a dual role. Epistemologically, they explain how learning occurs across different levels of the system. Functionally, they serve as mechanisms through which AI influences CPD design and outcomes. For this reason, learning theories are treated as both a foundational pillar and a mediating layer within the framework.

AI-Enhanced Authoring and Analytic Support

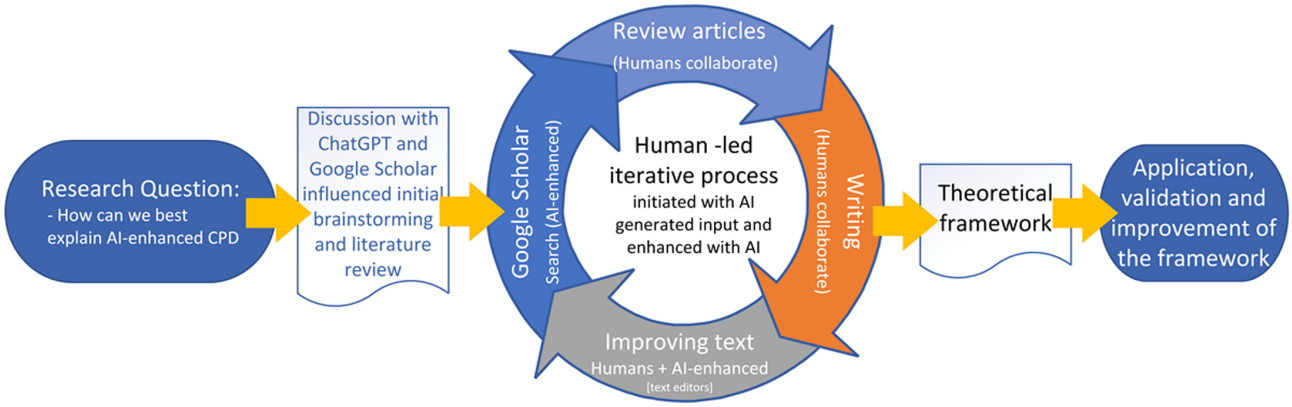
All parts of this article—except the practical examples of framework use—were created by human authors with AI support [40]. For example, ChatGPT was used for brainstorming and scenario setting. Google Scholar was used organically to search for literature supporting or expanding the topics discussed. AI-enhanced typing assistants (Grammarly, Microsoft, and Google) were used to improve the text's clarity, syntax, and concision (Figure 1).

All selected tools come with significant benefits and limitations—most noticeably ChatGPT. For example, ChatGPT output can be biased and lack deep domain-specific insight. ChatGPT cannot make ethical decisions or provide a transparent "thinking process." Additionally, ChatGPT can hallucinate [26,41]. To address these limitations, we used iterative, collaborative human evaluation and cross-verification with existing literature.

Abductive reasoning was used for framework validation. We tested the framework/hypothesis against real-world case studies.

Due to the complex, emerging nature of the investigated phenomena, inductive and deductive reasoning cannot work well for this task. Therefore, abductive reasoning is the most appropriate approach for this task.

Figure 1. The AI-enhanced writing process of this article. AI: artificial intelligence; CPD: continuing professional development.



Ethical Considerations

This manuscript is a conceptual paper. It does not involve human participants, human data, or identifiable personal information. Therefore, per applicable ethical guidelines, including COPE recommendations, ethics committee or institutional review board approval was not required.

Results

An Explanatory Framework for AI-Enhanced CPD

This study developed the ALEERRT-CA framework. It is a theory-driven model that provides insight into ways AI can transform CPD as a sociotechnical system. Our multimethod analysis identified 6 central AI pillars and demonstrated how their interactions become clearer when interpreted through CT

XSL•FO
RenderX

and ANT. The findings indicated that, in addition to new technological capabilities, AI serves as a transformative agent. It can change relationships, mediating artifacts, and system-level dynamics in CPD and the broader health care system. This study contributes to the conceptual clarity of how AI transforms CPD. It delivers an integrative explanatory model to an under-theorized domain, providing a foundation for future empirical, design-oriented, and policy-focused work in AI-enhanced CPD.

AI and Learning Theories

Learning theories are an essential building block of AI-enhanced learning environments. For example, Cognitive Load Theory (CLT) and Connectivism can provide valuable insight into using AI to enhance learning activities in different parts of the learning system.

CLT focuses on managing learners' cognitive load to enhance learning efficiency [42]. The goal is to decrease or eliminate extraneous load (eg, unnecessary examples), adjust the intrinsic load to the learner's skill level (eg, context appropriate to the skill level), and ensure that the remaining work memory capacity is focused on germane load (ie, cognitive learning processes). Examples of AI-enhanced and CLT-guided interventions are adaptive learning modules, in which AI adjusts difficulty based on learner performance [43,44], and multimedia learning, in which AI optimizes multimedia delivery to optimize cognitive load [45].

Connectivism emphasizes the role of networks and connections in the learning process [46,47]. AI and connectivism may facilitate expanding and managing learning networks in the health care sector [48]. For example, AI can aggregate and curate up-to-date information for health care professionals [49,50] and suggest learning paths based on individual goals and interests.

The examples above illustrate the interaction between learning theories and AI. Learning theories explain how learning happens, and AI can help enhance the learning processes described by learning theories. However, learning theories do not describe the broader framework of how AI, learning practices, and learning theories interact in the broader health care context.

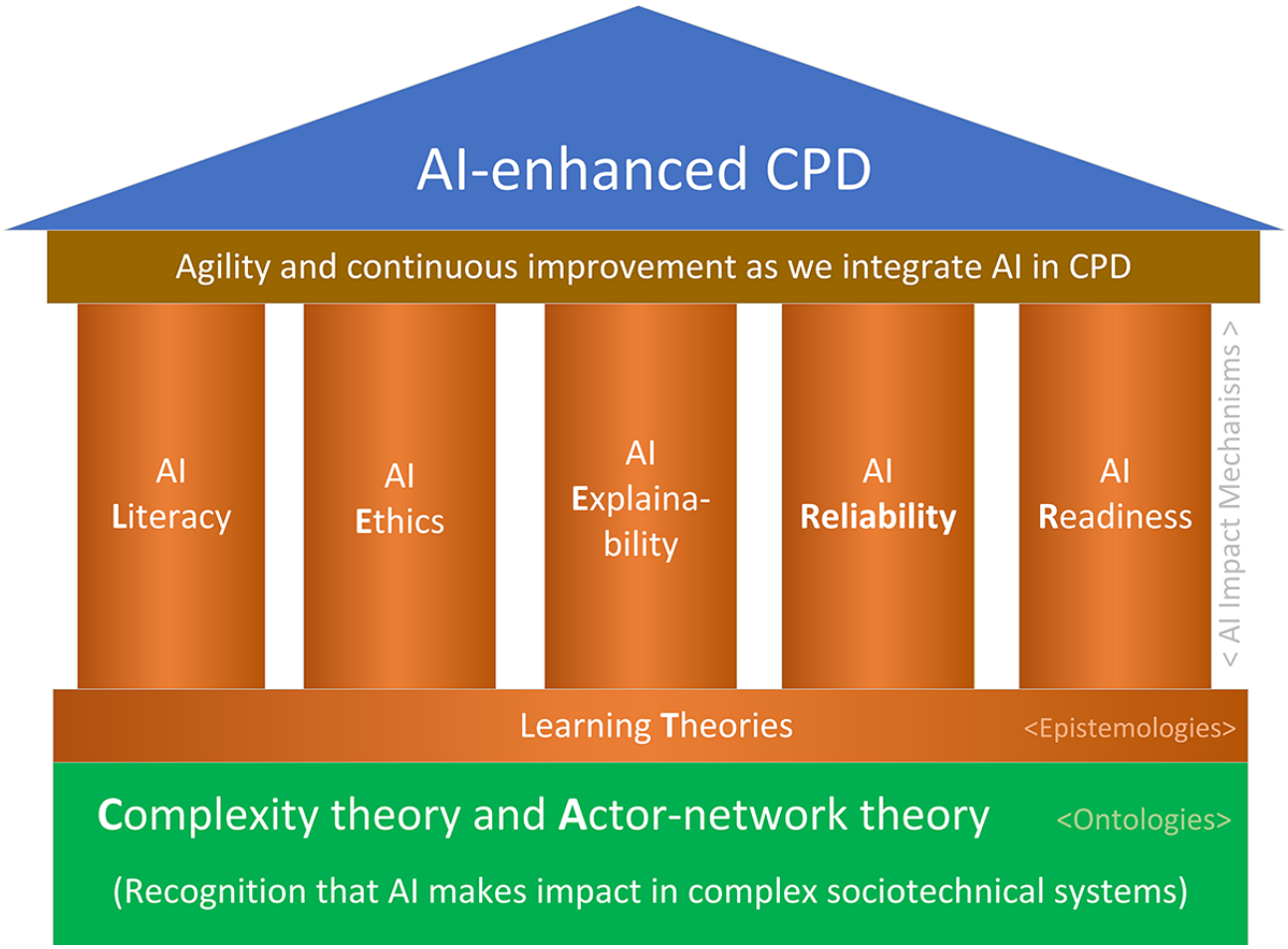
Framework for Integration of AI and CPD

The complex interaction between AI, learning theories, and CPD practices introduces the need for a framework to comprehend and guide the integration of AI with CPD for health professionals. We propose a framework made of the 6 foundational pillars: AI Literacy, Explainability, Ethics, Readiness, Reliability, and Learning Theories, and 2 complementary theoretical lenses: Complexity theory and Actor-network theory, ALEERRT-CA in short (Figure 2).

AI literacy describes our capacity to understand AI's basic concepts and principles, such as natural language processing, machine learning, computer vision, and deep data analytics [51]. It also involves the ability to use AI tools and applications in clinical practice, such as decision support systems, diagnostic tools, treatment recommendations, patient monitoring, and health education. AI literacy is essential for health and CPD professionals to leverage AI's potential for improving quality and efficiency.

XSL•FO
RenderX

Figure 2. The proposed framework for integrating AI in CPD—ALEERRT-CA is made of the 6 foundational pillars: AI Literacy, Explainability, Ethics, Readiness, Reliability, and Learning Theories, and 2 complementary theoretical lenses to observe system changes associated with AI: Complexity theory and Actor-network theory. AI: artificial intelligence; CPD: continuing professional development.



AI explainability (internal logic of how AI makes decisions) and the explainability of the impact AI makes on learning interventions (relationship between AI actions and broader sociotechnical CPD context) are enablers of AI literacy, readiness, and ethics—understanding how AI-enhanced systems work eases successful implementation of AI [52-54]. However, it seems that trusting that AI is reliable and performs well is more important than having a deep understanding of how AI algorithms work [55]. Very often, we tolerate AI-enhanced solutions as black boxes (Figure 3).

Black boxing is a common process associated with maturation, reliability, and wide acceptance of technology. While the inputs and outputs of the system are known, the user does not understand, or is not even aware of the processes inside the black box. Smartphones are a typical example of a black box [56]. As users, we are experienced in inputting and using smartphone outputs. Yet, the average user is minimally aware of the processes in smartphones and how networked and often AI-enhanced apps in the phone interact with external actors. We focus on the service it delivers to us, not on how it works. We trust it works well.

When failures occur or in attempts to improve the system, we need tools to open black boxes of sociotechnical systems that use AI to make processes visible and ideally understandable to humans [57]. It is an opportunity to examine human and nonhuman actors in the network of relationships between them and how their interactions deliver desired or, in some cases, erroneous outcomes.

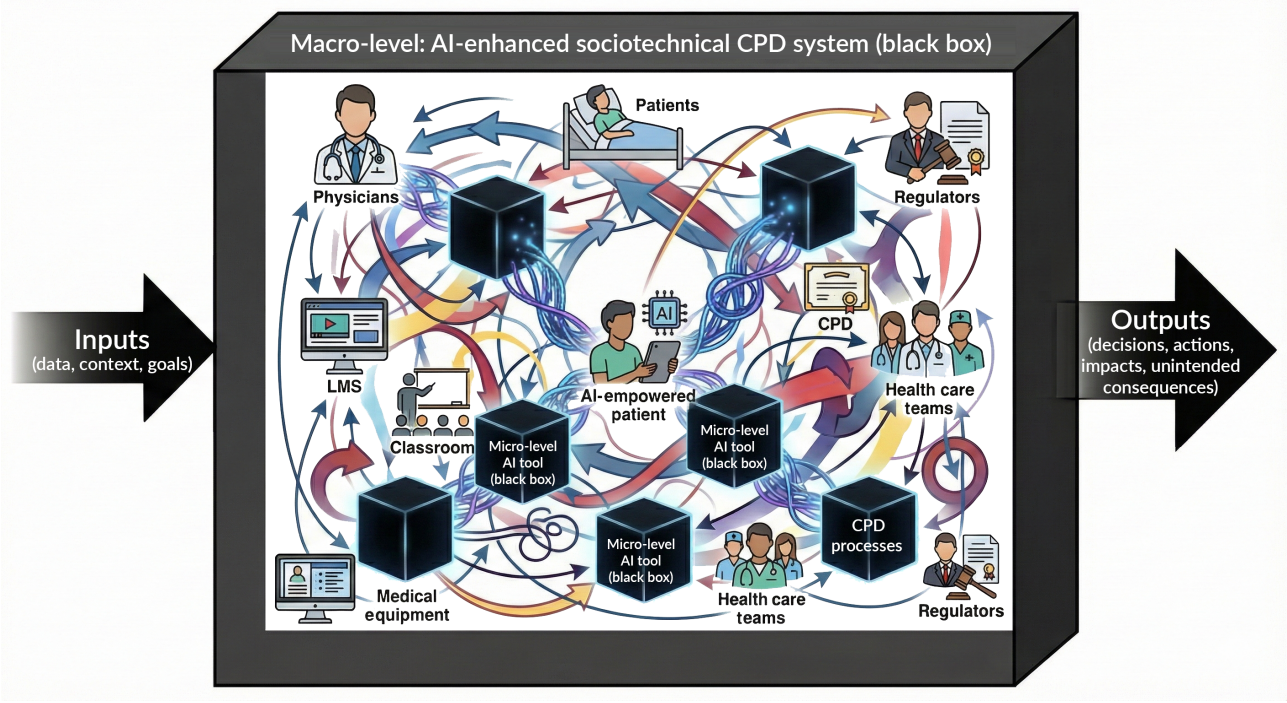
AI ethics refers to AI's ethical, legal, social, and professional implications for health care and CPD practice. It involves protecting patient privacy, obtaining informed consent, ensuring accountability, managing bias, ensuring fairness, protecting intellectual property, promoting transparency, and building trust [58]. AI ethics principles enable health professionals to use AI safely and responsibly.

AI readiness refers to the willingness and preparedness to adopt and implement AI technologies in learning and clinical practice [59]. It involves having a positive attitude and mindset towards AI, being open to learning from and with AI systems, and being able to cope with the changes and challenges that AI brings. AI readiness empowers health professionals to embrace AI as a partner in health care delivery and CPD. This readiness involves properly implementing AI-enhanced, research- and data-driven care by developing mental models, skill sets, and support systems for health care and CPD providers, their teams, and their organizations. Enterprise-wide AI readiness models can be considered to help organize and prioritize the organizational resources for the successful implementation of AI technologies [60-62].

XSL•FO
RenderX

AI reliability is the ability of an AI system to perform consistently and accurately under varying conditions [63]. It is crucial in high-stakes clinical and CPD applications. Yet, it comes with considerable plasticity [64]. AI systems, or humans using AI tools, must be more reliable than humans alone. For example, apps that flag at-risk students or problematic content posted by students may not be as reliable as humans, yet they may reduce the time-consuming task of reading posts and communicating findings, enabling humans to perform much better and faster than they would alone.

Figure 3. Opening the black box of AI-enhanced CPD. The image presents a conceptual model of an "AI-enhanced sociotechnical CPD system" as a macro-level "black box." The macro system is a dynamic network of interacting human actors (physicians, patients and their families, and CPD professionals), technical artifacts, and regulatory forces. Crucially, embedded within this chaotic environment are multiple "micro-level AI tools," presented as opaque black boxes. The diagram illustrates that understanding AI in CPD requires navigating a dual layer of complexity: the lack of explainability in individual AI tools and the unpredictable, emergent dynamics of the broader sociotechnical environment in which they operate. Image created with Google Gemini. AI: artificial intelligence; CPD: continuing professional development; LMS: learning management system.



Learning theories describe how learning happens and explain where and how AI can improve learning interventions.

CT and ANT are proposed as complementary theoretical lenses, where the CT lens is better suited to deliver a holistic view, while ANT can easily zoom in on a specific part of the system or a specific AI-enhanced app and deliver more actionable insight.

CT explains that our world is made of complex, constantly evolving systems. Those systems are open, and they adapt to changes in the context. Complex systems have emerging properties. Therefore, we cannot understand them simply by analyzing their parts. To analyze AI-enhanced systems, we should not focus solely on AI, but on the system and how the addition of AI is transforming the system.

Complex systems are connected, open, and nested. For example, an individual clinician is a system. Yet she is part of the operating room team—a supersystem. Above that, we have suprasystems such as hospital, national, and global health care systems [65,66]. On all those levels, AI can play a role [67]. Furthermore, AI-related improvement in one system, for example, in the operating room, will stimulate changes in external systems such as CPD, administration, and patient communication.

CT explains the need for multiple learning theories and the interaction between them. Learning theories observe the same phenomena—learning. However, they observe learning in different parts of nested hierarchies of CASs—that is, different contexts [67,68]. The CLT, for example, focuses on individual and learning that happens primarily internally—in a learner's brain. Connectivism, on the other hand, focuses on learning as a global, social, and technology and artifact-enhanced endeavor. AI can have an impact on all levels of our reality (from individual to global society). Therefore, it is fair to believe that the interaction between AI and established learning theories will be fruitful, allowing us to optimize CPD interventions at every level (individual, team, organization, population, state, and global society) [67].

ANT explains that our reality is shaped through evolving networks of relationships between human and nonhuman actors [69]. ANT posits that nonhuman actors, such as text, digital devices (smartphones or electronic health records), software programs, ideas, organizations, or AI tools, have the agency to shape our reality like humans.

ANT can serve as a magnifier for analyzing micro-level networks of nonhuman and human actors in a specific part of a complex sociotechnical system and at a particular time [69]. It is a good lens to analyze the use of a specific AI-enhanced

XSL•FO
RenderX

app or department using AI. CT is better suited to the holistic view of the system [66,70], such as how AI is changing CPD. While the picture created with CT is more inclusive, it is blurry. The macro system's complex, evolving nature does not allow us to capture all system elements. As a combination, ANT and CT enable us to observe the big picture and, when needed, zoom in and observe a specific part of the AI-enhanced learning health care system.

The framework provides tools that can help us open the black box and observe the "main anatomical structures of AI-enhanced CPD" through both a macro-level (CT) and micro-level (ANT) lenses (Figure 3).

The ALEERRT-CA framework is organized in three conceptual layers. CT and ANT serve as the ontological foundation by defining CPD as a nested, adaptive sociotechnical system made of interacting human and nonhuman actors. Learning theories serve as the epistemological layer, offering assumptions about how knowledge is constructed and how learning can be enhanced. The 6 AI pillars—Literacy, Explainability, Ethics, Readiness, Reliability, and Learning Theories—function as interactional mechanisms that shape how AI influences relationships, behaviors, and emergent patterns within the CPD system. Learning theories simultaneously serve as an AI impact mechanism that enables AI impact and an epistemological lens.

Discussion

Reframing AI-Enhanced CPD as a Sociotechnical System

This study presents a novel theoretical lens for analyzing the sociotechnical dynamics of AI-enhanced CPD. It provides insight into how AI reshapes relationships, redistributes agency, and generates emergent learning dynamics across CPD systems. The framework combines 6 AI pillars (AI literacy, explainability, ethics, readiness, reliability, and learning theories) with CT and ANT. The main finding is that complex interactions among system elements are the primary forces shaping AI-enhanced CPD, while isolated AI tools or adoption factors are of secondary importance. Due to its complex nature, AI-enhanced CPD often acts as a black box. The framework strengthens our capacity to analyze the complex sociotechnical nature of AI-enhanced CPD.

Adequacy of the Framework

The proposed framework provides a checklist of 6 building blocks of AI implementation and 2 theoretical lenses to observe system changes associated with AI. It is a tool set that can help us plan AI implementation and, when needed, open black boxes of AI-enhanced CPD. Therefore, it addresses the criteria of practical application and simplicity (parsimony) [71]. CT and ANT, as macro-level and micro-level theories, allow us to observe the system and large-scale structures (CT), and, when needed, focus on the interaction between a small network of human and nonhuman actors (ANT). The proper theoretical lens for a proper task model aligns with the principle of parsimony.

The framework is rooted in 2 established theories (CT and ANT) and the literature on AI implementation, which enhances its external consistency and ensures it fits with the broader theoretical landscape of AI in CPD. It appears to provide a good balance of practicality, simplicity, and theoretical robustness, suggesting it can effectively explain how successful AI-enhanced improvement of CPD can occur.

Value of the Constructed Theoretical Framework

The ALEERRT-CA allows us to open the black boxes of AI-enhanced CPD. The black boxes exist on at least 2 levels. The first level is AI algorithms. That level is not addressed in this paper. On the level above, where this framework is focused, we have AI-enhanced sociotechnical systems and need to understand hidden, complex interactions between AI, technical, and social elements that influence learning health system performance and outcomes. The framework describes the main parts—the anatomy—of the AI-enhanced CPD and 2 theoretical lenses used to analyze them.

AI implementation-focused guides, such as Rashidi Your AI Survival Guide [72], highlight many of the conditions necessary for successful AI adoption, including literacy, workflow redesign, trust, governance, readiness, and interprofessional collaboration. These themes map closely onto the ALEERRT-CA pillars. The proposed framework builds on that. ALEERRT-CA extends beyond operational guidance.

Whereas leadership manuals explain what organizational leaders must pay attention to, ALEERRT-CA explains how these elements interact within a complex adaptive learning ecosystem and why AI reshapes roles, routines, and system behaviors. By embedding AI constructs within CT and ANT, the framework provides a level of explanatory depth that practice-oriented texts do not have. The framework explains the dynamics through which AI mediates relationships, generates tensions, and produces emergent patterns in CPD. In this way, ALEERRT-CA transforms a collection of best practices into a coherent, theory-driven model capable of guiding both empirical research and strategic system design.

Interaction Between Pillars and Theoretical Lenses

The 6 pillars of ALEERRT-CA demonstrate distinct yet interdependent explanatory functions when analyzed through ANT and CT (Figure 4).

For example, ANT shows that explainability is not a feature strictly associated with AI algorithms. Reality is more complex. Explainability is negotiated across a network of users, artifacts, interfaces, and organizational structures. Therefore, explainability is relational, contextual, and dynamic. What counts as "a good explanation" varies with the actor making the assessment, purpose, location within the system, and level of risk.

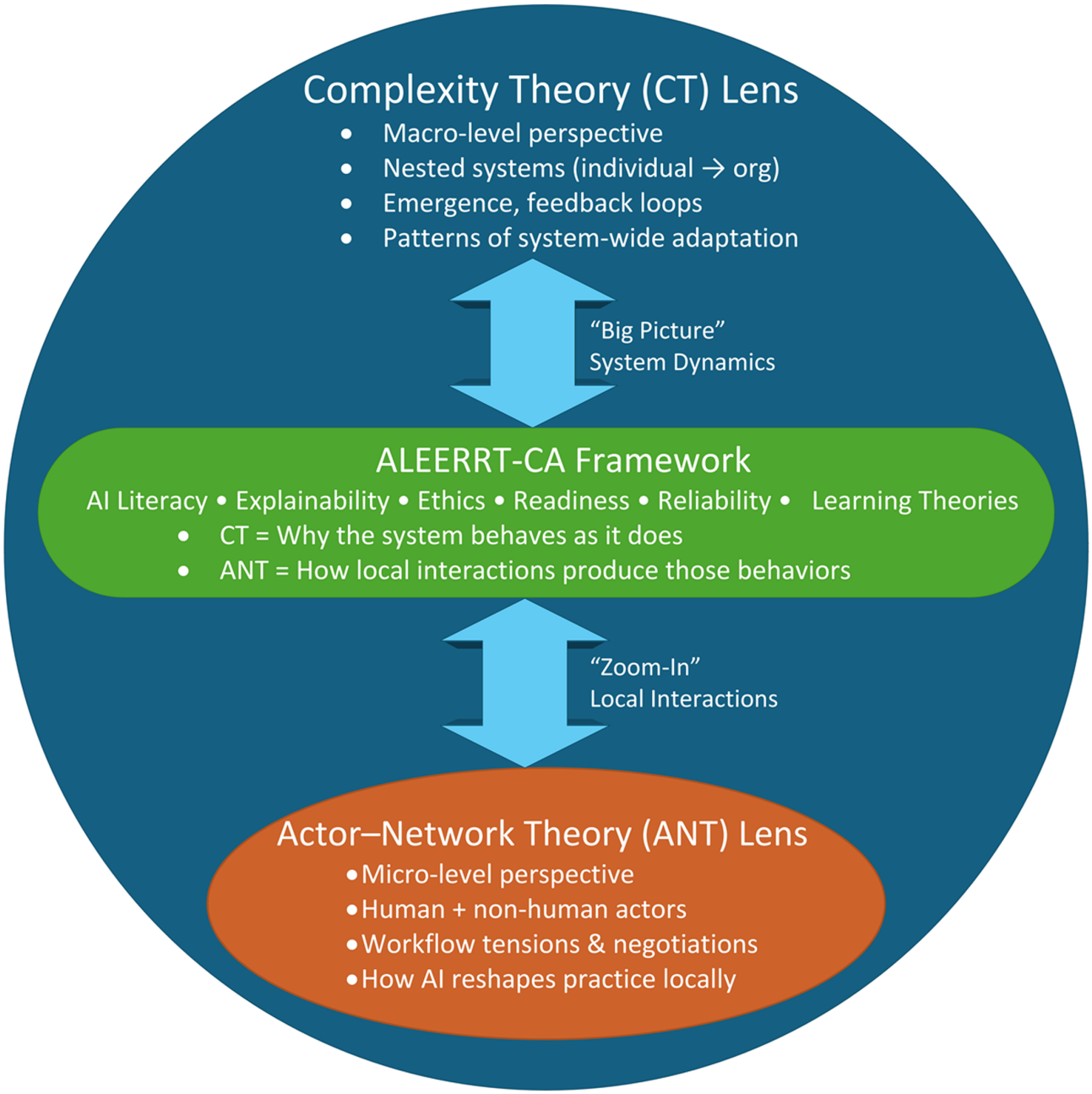
ANT informs AI ethics by showcasing how agency, accountability, and power are redistributed when nonhuman actors participate in decisions. System-wide ethical outcomes are created through interactions among designers, clinicians, learners, data, models, platforms, policies, and workflows. ANT enables CPD leaders to evaluate not only whether AI is ethical,

XSL•FO
RenderX

but how ethical behavior is produced, constrained, or undermined by the network.

Yang et al [73] provide a practical example. Explainability and perceived reliability during peer reviews are shaped by the social "advice-taking phenomenon." Physicians who openly use AI tools during decision-making usually receive lower competence ratings, despite shared belief that AI is helpful. Positioning AI as a verification tool only moderately reduces that negative effect. The probable reason is that we have a significant reconfiguration of the actor network. Epistemic authority, agency, and accountability are redistributed and reshaped by AI. That change reduces trust and perceived reliability, despite improved performance.

Figure 4. Complexity Theory (CT) and Actor-Network Theory (ANT) serve as complementary analytical lenses within the ALEERRT-CA framework. CT provides a macro-level view of continuing professional development (CPD). Through that lens, CPD is seen as a nested, adaptive system shaped by emergence, feedback loops, and system-wide interactions. ANT, on the other hand, provides a micro-level view. Through that lens, CPD is seen as a network of human and nonhuman actors (clinicians, educators, learners, AI tools, platforms, and policies). We can see how they interact and negotiate new practices. Together, CT and ANT enable ALEERRT-CA to explain both broad system transformation and the specific mechanisms through which AI reshapes CPD.



As illustrated in Table 1, ANT enables micro-level analytic tracing of how AI reshapes agency, accountability, and epistemic authority through human–nonhuman associations. CT complements this by explaining how localized interactions among these actors influence emergent system trajectories across nested sociotechnical strata. Together, the dual-lens positioning of ALEERRT-CA reframes AI-enhanced CPD not as a linear implementation challenge but as a dynamic process of relational negotiation and systemic adaptation. This dual-theoretical

XSL•FO
RenderX

integration provides explanatory depth and design utility, capturing both mechanism and emergence.

Table . Theoretical interaction of ALEERRT-CA pillars with Actor-Network Theory and Complexity Theory.

ALEERRT-CA pillar	ANT[a] contribution, a micro-level analytical mechanism	CT[b] contribution, a macro-level system mechanism	Theoretical implication
AI[c] literacy	Positions literacy as a prerequisite for relational/contextual agency. Actors with greater literacy have disproportionate influence over problem definition, meaning-making, and artifact interpretation.	Literacy accelerates or constrains system adaptation; uneven literacy becomes a systemic bottleneck spreading nonlinearly across teams and institutions.	AI literacy shapes epistemic power and mediates how knowledge circulates within sociotechnical networks.
Explainability	ANT frames explainability as an outcome of interaction between various human and nonhuman actors and ideas; what is "understandable" is rooted in actor identity, status/position in the networks, network structures, and interpretive frames.	Changes in the level of explainability can amplify or suppress feedback loops, affecting trust, adoption, and the rate of transformation. Explainability is needed at multiple levels: from the work of AI algorithms, individual-focused AI-enhanced CPD[d], team- and organization-focused learning, and integration between learning and clinical practice.	Explainability is relational, negotiated, and context-dependent rather than a static technical property. Explainability of processes (eg, AI algorithms) on one level influences explainability on other levels (eg, clinical practice)
Ethics	Shows how responsibility and agency are redistributed when nonhuman actors coproduce decisions and outcomes.	Ethical disruptions propagate via feedback loops, enabling amplification of harms or benefits at scale. Ethics is important on all levels: algorithm, individual, team, organization, society	Ethics becomes a governance outcome of network configuration, not solely adherence to level-specific principles.
Readiness	Readiness depends on network alignment. Misaligned goals, competencies, and affordances produce friction and resistance.	Initial conditions and small deviations shape emergent trajectories ("sensitivity to initial states"). Readiness is multilayered. At different levels of system aggregation (individual, team, organization, society), we can observe considerable differences in readiness.	Readiness is dynamic and coevolving, requiring continual recalibration across nested systems.
Reliability	Reliability emerges from situated interaction between human and nonhuman actors; failure modes are relational.	Failures or successes can trigger rapid changes in trust in AI-enhanced systems. That can impact behavior across the system(s).	Reliability is contingent and contextual, negotiated by the system (and between systems) rather than guaranteed by the artifact.
Learning theories	ANT[a] reveals how AI acquires pedagogical agency (content generation, evaluation, orchestration) and shapes relationships in a network of actors.	CT[b] positions AI-enhanced learning as emergent across individual, social, and organizational strata.	CPD[d] must align AI-enhanced learning design with both cognitive and sociotechnical learning dynamics. AI can enhance individual, team, and system learning.

[a]ANT: Actor-Network Theory.

[b]CT: Complexity Theory.

[c]AI: artificial intelligence.

[d]CPD: continuing professional development.

To illustrate the value of the ALEERRT-CA framework, we offer an explanatory case.

Explanatory Case: The American Society of Anesthesiologists BeaconBot and the Emerging AI-Enhanced CPD Ecosystem

The American Society of Anesthesiologists (ASA) is coproducing an AI-enhanced learning ecosystem that tackles clinical practice, quality improvement, education, research dissemination, and workforce development. A central AI-enhanced component is ASA BeaconBot, a generative AI knowledge assistant available to ASA members [74]. BeaconBot synthesizes information from ASA's public materials, member-only resources, *Anesthesiology® journal*, committee documents, and ASA Standards, Guidelines, and Statements. BeaconBot functions as a real-time mediator of organizational knowledge, supporting clinicians, educators, and staff in accessing evidence, policies, and practice recommendations more efficiently. It serves as a librarian, quickly providing answers to complicated questions with citations and links to sources that informed the answers.

XSL•FO
RenderX

In parallel, the Anesthesia Quality Institute is advancing a multistage integration strategy linking (1) Epic electronic health record data, (2) the National Anesthesia Clinical Outcomes Registry [75], and (3) ASA's association management system. Public descriptions of this work emphasize using high-volume clinical data to improve benchmarking and quality improvement across anesthesia practices.

A long-term, aspirational goal, currently conceptual rather than operational, is to explore whether abstracted, nonpatient-identified insights about clinicians' practice patterns (eg, procedural focus, strengths, and learning needs) could be generated in the future. At present, legal, regulatory, and data privacy constraints limit such use.

These insights could eventually inform ASA's learning management system and learning experience platform, enabling AI-supported personalized CPD recommendations aligned with actual clinical practice [76].

ASA operates across multiple educational layers: individual clinicians and residents, perioperative teams, organizational purchasers of CPD, national public education and legislative advocacy, and international patient-safety leadership. Across these layers, AI is already entangled with everyday practice: faculty and members use AI tools for literature synthesis, communication, research workflows, and meeting documentation. Many committee processes and reports are generated or refined with AI-assisted writing tools, and several e-learning materials, including SCORM (Shareable Content Object Reference Model) e-learning modules, are produced with AI-supported content development [77].

Internally, ASA is implementing an empowered product team model [78]. The traditional hierarchical and often siloed, feature-focused waterfall project management practices are being replaced with an agile continuous discovery model built around the active involvement of all stakeholders [79].

The empowered product teams model aligns with both CT and ANT, as it operates as an adaptive, feedback-driven hub capable of responding to emergent interactions among human and nonhuman actors. In contrast, waterfall models assume stability, linearity, and fixed requirements—assumptions that are poorly aligned with the aim of innovation in AI-enhanced CPD systems.

Viewed through the ALEERRT-CA framework, this evolving ecosystem illustrates:

- Literacy: Heterogeneous AI literacy across clinicians, faculty, learners, and staff creates both capability gaps and opportunities for targeted support.
- Explainability: BeaconBot's responses and any future recommendation logic require transparent reasoning pathways for clinicians, educators, and regulators. In addition to how recommendations are created, we want to know how they help clinicians improve their practice.
- Ethics: Registry-to-learning integration raises governance questions around data use, privacy, and boundaries between performance data and educational personalization.
- Readiness: Organizational readiness fluctuates across departments as product teams, educators, IT, and leadership adapt to new workflows and AI-mediated artifacts.
- Reliability: Perceived reliability depends on both technical accuracy and user trust, particularly when AI synthesizes standards or summarizes clinical knowledge.
- Learning theories: Learning shifts from content-centric modules toward system-wide, data-driven adaptive learning, where clinical practice informs CPD and CPD feeds back into practice.

From a CT perspective, ASA's system shows nonlinear interactions: small changes (eg, improved data quality, AI-generated insights, or revised governance policies) can propagate across clinical, educational, and organizational layers. From an ANT lens, AI agents (BeaconBot, analytics engines, transcription models, generative AI) operate as nonhuman actors that mediate relationships among clinicians, educators, committees, systems, and CPD infrastructures. Those actions empower CPD to act as a learning health ecosystem, characterized by evolving feedback loops, distributed agency, and emergent patterns of adaptation.

While results are promising, strong tensions around data governance, ethics, explainability, change fatigue, and the need for AI literacy among clinicians, faculty, and staff impede progress.

This case illustrates how AI-enhanced CPD cannot be understood as a discrete educational intervention but must be analyzed as a system-level transformation involving data flows, governance, professional identity, and learning infrastructure—precisely the analytic space addressed by ALEERRT-CA. Thus, it is equally beneficial for organizations and CPD professionals.

As a relatively simple but inclusive and complexity-ready model, it can serve as a toolset that will help us with the integration of AI in CPD. It can support interdisciplinary collaboration between medical, engineering, social, ethical, and legal domains and help us shape human-centric AI-enhanced CPD.

As with physical toolsets, it is not necessary to use all tools in the box for every situation. Some analytic tasks require only a subset of the framework's components. The framework is also extensible. While CT and ANT are well-suited for explaining macro-level system dynamics and micro-level sociotechnical interactions in AI-enhanced CPD, other theoretically compatible lenses could be incorporated if they better match a specific analytic need or reflect the expertise of a given research team.

Limitations

The complex, constantly evolving, and rapid nature of AI in our society requires an agile, continuous-improvement mindset. Therefore, we propose this framework as a work in progress and a position on the direction we, as the health care CPD community empowered with AI tools, may take. The framework does not claim completeness or finality but is intended as a theoretically grounded starting point for iterative refinement through empirical studies.

The iterative nature of the framework allows us to improve both our use of it and the framework itself. Similarly, repeated empirical validation will help confirm its value.

XSL•FO
RenderX

Conclusion

Integration of AI into health care CPD is reshaping a complex system, demanding a deeper understanding and strategic application of available resources.

This work introduces ALEERRT-CA, a theory-driven framework that provides a lens for observing dynamic interactions in AI-enhanced CPD. The framework explains how AI reorganizes relationships, redistributes agency, and generates emergent learning across health care and CPD systems.

ALEERRT-CA can serve as an analytic tool for practice. Educators and organizational leaders can use the tool to assess AI-enhanced learning at multiple levels, from individual learners to organizations and broader health care systems. The paired macro and micro lenses support both strategic "big picture planning" and focused analysis of local AI-enhanced CPD initiatives.

Lessons for Practice

As AI becomes increasingly embedded in healthcare education and practice, CPD professionals are faced with practical decisions about how to design, implement, and govern AI-enhanced learning. The following lessons for practice distill key insights from this study:

- AI-enhanced tools have become part of our daily lives. It is not whether we use AI, but how effectively we use it to enhance our CPD practices.
- AI makes an impact on complex sociotechnical systems. Therefore, AI-enhanced CPD interventions should be designed to foster beneficial interactions between AI tools, health care professionals, and the broader health care system.
- Theoretical tools, such as ALEERRT-CA, can increase our ability to open the black boxes of AI-enhanced CPD sociotechnical systems and understand and improve how AI is used in CPD, ultimately improving the impact of CPD.

Acknowledgments

We thank our colleagues and leaders at SACME, especially Todd Dorman, MD; Heather MacNeill, MD; David Wiljer, PhD; Simon Kitto, PhD; Olivier Petinaux, MS; and Marianna Shershneva, MD, PhD, for their helpful feedback, encouragement, and support.

AI tools, such as ChatGPT 3.5 (starting in 2023) to 5.1 (2025), Google Gemini, and AI-based writing assistants, supported idea generation, summarization, and text refinement. Human authors conducted all initial brainstorming, conceptual development, theoretical interpretation, and final editing. All AI outputs were reviewed and verified.

Data Availability

No datasets were generated or analyzed during the current study. This work is conceptual and theory-building in nature.

Authors' Contributions

Conceptualization: VH (lead)
Methodology: VH (lead), SV (supporting), GRD (supporting), HD (supporting), ROB (supporting), RW (supporting)
Formal analysis: VH (lead)
Validation: SV (equal), GRD (equal), HD (equal), ROB (equal), RW (equal)
Visualization: VH (lead), GRD (supporting)
Writing – original draft: VH (lead)
Writing – review & editing: VH (lead), SV (supporting), GRD (supporting), HD (supporting), ROB (supporting), RW (supporting)
All authors contributed to the study design, iterative framework refinement, critical manuscript review, and approval of the final version.

Conflicts of Interest

None declared.

References

1. Davenport T, Kalakota R. The potential for artificial intelligence in healthcare. Future Healthc J 2019 Jun;6(2):94-98. [doi: 10.7861/futurehosp.6-2-94] [Medline: 31363513]
2. Jandrić P, Knox J, Besley T, Ryberg T, Suoranta J, Hayes S. Postdigital science and education. Educ Philos Theory 2018;50(10):893-899. [doi: 10.1080/00131857.2018.1454000]
3. Federspiel F, Mitchell R, Asokan A, Umana C, McCoy D. Threats by artificial intelligence to human health and human existence. BMJ Glob Health 2023 May;8(5):e010435. [doi: 10.1136/bmjgh-2022-010435] [Medline: 37160371]
4. Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. Nat Med 2019 Jan;25(1):44-56. [doi: 10.1038/s41591-018-0300-7] [Medline: 30617339]
5. Salwei ME, Carayon P. A sociotechnical systems framework for the application of artificial intelligence in health care delivery. J Cogn Eng Decis Mak 2022 Dec;16(4):194-206. [doi: 10.1177/15553434221097357] [Medline: 36704421]

XSL•FO
RenderX

6. Ensign D, Nisly SA, Pardo CO. The future of generative AI in continuing professional development (CPD): crowdsourcing the Alliance community. J CME 2024 Dec 9;13(1):2437288. [doi: 10.1080/28338073.2024.2437288] [Medline: 39664116]
7. Mastrianni A, Kmetz-Cutrone P, Chang K, Stein JY, Sarcevic A. Beyond decision making: considering collaboration and agency in the design of AI-based decision-support systems for fast-response medical teams. 2025 Presented at: 28th ACM Conference on Computer-Supported Cooperative Work and Social Computing (ACM CSCW 2025); Oct 18-22, 2025. [doi: 10.1145/3757457]
8. Greenhalgh T, Papoutsi C. Studying complexity in health services research: desperately seeking an overdue paradigm shift. BMC Med 2018 Jun 20;16(1):95. [doi: 10.1186/s12916-018-1089-4] [Medline: 29921272]
9. Mesko B, deBronkart D, Dhunnoo P, Arvai N, Katonai G, Riggare S. The evolution of patient empowerment and its impact on health care's future. J Med Internet Res 2025 May 1;27:e60562. [doi: 10.2196/60562] [Medline: 40311140]
10. Bachina L, Kanagala A. Health revolution: AI-powered patient engagement. Int J Chem Biochem Sci 2023;24(5):722-731 [FREE Full text]
11. Borsboom D, van der Maas HLJ, Dalege J, Kievit RA, Haig BD. Theory construction methodology: a practical framework for building theories in psychology. Perspect Psychol Sci 2021 Jul;16(4):756-766. [doi: 10.1177/1745691620969647] [Medline: 33593167]
12. Chen Y, Lehmann CU, Malin B. Digital information ecosystems in modern care coordination and patient care pathways and the challenges and opportunities for AI solutions. J Med Internet Res 2024 Dec 2;26:e60258. [doi: 10.2196/60258] [Medline: 39622048]
13. Doraiswamy PM, Blease C, Bodner K. Artificial intelligence and the future of psychiatry: insights from a global physician survey. Artif Intell Med 2020 Jan;102:101753. [doi: 10.1016/j.artmed.2019.101753] [Medline: 31980092]
14. Ning Y, Ong JCL, Cheng H, et al. How can artificial intelligence transform the training of medical students and physicians? Lancet Digit Health 2025 Oct;7(10):100900. [doi: 10.1016/j.landig.2025.100900] [Medline: 41047321]
15. Wartman SA, Combs CD. Reimagining medical education in the age of AI. AMA J Ethics 2019 Feb 1;21(2):E146-E152. [doi: 10.1001/amajethics.2019.146] [Medline: 30794124]
16. Mosch L, Fürstenau D, Brandt J, et al. The medical profession transformed by artificial intelligence: qualitative study. Digit Health 2022;8:20552076221143903. [doi: 10.1177/20552076221143903] [Medline: 36532112]
17. Lipsitz LA. Understanding health care as a complex system: the foundation for unintended consequences. JAMA 2012 Jul 18;308(3):243-244. [doi: 10.1001/jama.2012.7551] [Medline: 22797640]
18. Ogden K, Kilpatrick S, Elmer S. Examining the nexus between medical education and complexity: a systematic review to inform practice and research. BMC Med Educ 2023 Jul 5;23(1):494. [doi: 10.1186/s12909-023-04471-2] [Medline: 37408005]
19. Mitchell B. Engaging with Actor-Network Theory as a Methodology in Medical Education Research: Routledge; 2021.
20. Woodruff JN. Accounting for complexity in medical education: a model of adaptive behaviour in medicine. Med Educ 2019 Sep;53(9):861-873. [doi: 10.1111/medu.13905] [Medline: 31106901]
21. Smith CS. Complex adaptive systems. In: Foundations of Interprofessional Health Education: An Ecological Theory: Springer; 2023:89-97. [doi: 10.1007/978-3-031-33414-6_18]
22. Alami H, Lehoux P, Auclair Y, et al. Artificial intelligence and health technology assessment: anticipating a new level of complexity. J Med Internet Res 2020 Jul 7;22(7):e17707. [doi: 10.2196/17707] [Medline: 32406850]
23. Randhawa GK, Jackson M. The role of artificial intelligence in learning and professional development for healthcare professionals. Healthc Manage Forum 2020 Jan;33(1):19-24. [doi: 10.1177/0840470419869032] [Medline: 31802725]
24. Dazzi P. The internet of AI agents (IAIA): a new frontier in networked and distributed intelligence. Int J Netw Distrib Comput 2025 Jun;13:16. [doi: 10.1007/s44227-025-00057-0]
25. Artificial intelligence and machine learning (AI/ML)-enabled medical devices. Software as a Medical Device (SaMD). : US Food and Drug Administration; 2025 URL: https://www.fda.gov/medical-devices/digital-health-center-excellence/software-medical-device-samd [accessed 2026-01-22]
26. Nilsen P, Svedberg P, Neher M, et al. A framework to guide implementation of AI in health care: protocol for a cocreation research project. JMIR Res Protoc 2023 Nov 8;12:e50216. [doi: 10.2196/50216] [Medline: 37938896]
27. Zawacki-Richter O, Marín VI, Bond M, Gouverneur F. Systematic review of research on artificial intelligence applications in higher education – where are the educators? Int J Educ Technol High Educ 2019;16:39. [doi: 10.1186/s41239-019-0171-0]
28. Papa R, et al. Discourse analysis on learning theories and AI. 2020 Presented at: Intelligent Computing: Proceedings of the 2020 Computing Conference; Jul 16-17, 2020. [doi: 10.1007/978-3-030-52243-8_50]
29. Xu W, Ouyang F. A systematic review of AI role in the educational system based on a proposed conceptual framework. Educ Inf Technol 2022;27(3):4195-4223. [doi: 10.1007/s10639-021-10774-y]
30. Bleakley A, Cleland J. Sticking with messy realities: how 'thinking with complexity' can inform health professions education research. In: Cleland J, Durning SJ, editors. Researching Medical Education 2022:197-208. [doi: 10.1002/9781119839446]
31. Turner JR, Baker RM. Complexity theory: an overview with potential applications for the social sciences. Systems 2019;7(1):4. [doi: 10.3390/systems7010004]
32. Lewin K. Problems of research in social psychology (1943-44). In: Lewin GW, editor. Resolving Social Conflicts and Field Theory in Social Science: American Psychological Association; 1997:279-288. [doi: 10.1037/10269-021]

XSL•FO
RenderX

33. Greenwood DJ, Levin M. Introduction to Action Research: Social Research for Social Change: SAGE Publications; 2006. URL: https://methods.sagepub.com/book/introduction-to-action-research [accessed 2026-01-22] [doi: 10.4135/9781412984614]
34. Makridakis S. The forthcoming artificial intelligence (AI) revolution: its impact on society and firms. Futures 2017 Jun;90:46-60. [doi: 10.1016/j.futures.2017.03.006]
35. Zhang W, Cai M, Lee HJ, Evans R, Zhu C, Ming C. AI in medical education: global situation, effects and challenges. Educ Inf Technol 2024;29(4):4611-4633. [doi: 10.1007/s10639-023-12009-8]
36. Mouloudj K, et al. Adopting artificial intelligence in healthcare: a narrative review. In: Teixeira S, Remondes J, editors. The Use of Artificial Intelligence in Digital Marketing: Competitive Strategies and Tactics: IGI Global Scientific Publishing; 2024:1-20. [doi: 10.4018/978-1-6684-9324-3.ch001]
37. Rajpurkar P, Lungren MP. The current and future state of AI interpretation of medical images. N Engl J Med 2023 May 25;388(21):1981-1990. [doi: 10.1056/NEJMra2301725] [Medline: 37224199]
38. Preiksaitis C, Rose C. Opportunities, challenges, and future directions of generative artificial intelligence in medical education: scoping review. JMIR Med Educ 2023 Oct 20;9:e48785. [doi: 10.2196/48785] [Medline: 37862079]
39. Allen PM, Varga L. Complexity: the co-evolution of epistemology, axiology and ontology. Nonlinear Dynamics Psychol Life Sci 2007 Jan;11(1):19-50. [Medline: 17173728]
40. Salvagno M, Taccone FS, Gerli AG. Can artificial intelligence help for scientific writing? Crit Care 2023 Feb 25;27(1):75. [doi: 10.1186/s13054-023-04380-2] [Medline: 36841840]
41. Dave T, Athaluri SA, Singh S. ChatGPT in medicine: an overview of its applications, advantages, limitations, future prospects, and ethical considerations. Front Artif Intell 2023 May 4;6:1169595. [doi: 10.3389/frai.2023.1169595] [Medline: 37215063]
42. Young JQ, Van Merrienboer J, Durning S, Ten Cate O. Cognitive load theory: implications for medical education: AMEE Guide No. 86. Med Teach 2014 May;36(5):371-384. [doi: 10.3109/0142159X.2014.889290] [Medline: 24593808]
43. Triola MM, Burk-Rafel J. Precision medical education. Acad Med 2023 Jul 1;98(7):775-781. [doi: 10.1097/ACM.0000000000005227] [Medline: 37027222]
44. Narang A, Velagapudi P, Rajagopalan B, et al. A new educational framework to improve lifelong learning for cardiologists. J Am Coll Cardiol 2018 Jan 30;71(4):454-462. [doi: 10.1016/j.jacc.2017.11.045] [Medline: 29389363]
45. Bays HE, Fitch A, Cuda S, et al. Artificial intelligence and obesity management: an obesity medicine association (OMA) clinical practice statement (CPS) 2023. Obes Pillars 2023 Apr;6:100065. [doi: 10.1016/j.obpill.2023.100065] [Medline: 37990659]
46. MacNeill H, Masters K, Nemethy K, Correia R. Online learning in health professions education. Part 1: teaching and learning in online environments: AMEE Guide No. 161. Med Teach 2024 Jan;46(1):4-17. [doi: 10.1080/0142159X.2023.2197135] [Medline: 37094079]
47. Goldie JGS. Connectivism: a knowledge learning theory for the digital age? Med Teach 2016 Oct;38(10):1064-1069. [doi: 10.3109/0142159X.2016.1173661] [Medline: 27128290]
48. Chaudhry S, Dhawan S. AI-based recommendation system for social networking. Presented at: Proceedings of the International Conference on Soft Computing: Theories and Applications (SoCTA 2018); Dec 21-23, 2018. [doi: 10.1007/978-981-15-0751-9_58]
49. Stamatis D, Kefalas P, Kargidis T. A multi‐agent framework to assist networked learning. J Comput Assist Learn 1999 Sep;15(3):201-210. [doi: 10.1046/j.1365-2729.1999.153094.x]
50. Eggmann N. Not plug-and-play: successful adoption of an AI-based learning experience platform. In: Ifenthaler D, Seufert S, editors. Artificial Intelligence Education in the Context of Work: Springer International Publishing; 2022:215-226. [doi: 10.1007/978-3-031-14489-9_13]
51. Laupichler MC, Aster A, Schirch J, Raupach T. Artificial intelligence literacy in higher and adult education: a scoping literature review. Comput Educ Artif Intell 2022;3:100101. [doi: 10.1016/j.caeai.2022.100101]
52. Amann J, Blasimme A, Vayena E, Frey D, Madai VI, Precise4Q consortium. Explainability for artificial intelligence in healthcare: a multidisciplinary perspective. BMC Med Inform Decis Mak 2020 Nov 30;20(1). [doi: 10.1186/s12911-020-01332-6] [Medline: 33256715]
53. Barredo Arrieta A, Díaz-Rodríguez N, Del Ser J, et al. Explainable artificial intelligence (XAI): concepts, taxonomies, opportunities and challenges toward responsible AI. Inf Fusion 2020 Jun;58:82-115. [doi: 10.1016/j.inffus.2019.12.012]
54. Dwivedi R, Dave D, Naik H, et al. Explainable AI (XAI): core ideas, techniques, and solutions. ACM Comput Surv 2023;55(9):1-33. [doi: 10.1145/3561048]
55. McCoy LG, Brenna CTA, Chen SS, Vold K, Das S. Believing in black boxes: machine learning for healthcare does not need explainability to be evidence-based. J Clin Epidemiol 2022 Feb;142:252-257. [doi: 10.1016/j.jclinepi.2021.11.001] [Medline: 34748907]
56. Sage D, Vitry C, Dainty A. Exploring the organizational proliferation of new technologies: an affective actor-network theory. Organ Stud 2020 Mar;41(3):345-363. [doi: 10.1177/0170840618815524]
57. Guidotti R, Monreale A, Ruggieri S, Turini F, Giannotti F, Pedreschi D. A survey of methods for explaining black box models. ACM Comput Surv 2018;51(5):1-42. [doi: 10.1145/3236009]

XSL•FO
RenderX

58. Murphy K, Di Ruggiero E, Upshur R, et al. Artificial intelligence for good health: a scoping review of the ethics literature. BMC Med Ethics 2021 Feb 15;22(1):14. [doi: 10.1186/s12910-021-00577-8] [Medline: 33588803]
59. Karaca O, Çalışkan SA, Demir K. Medical artificial intelligence readiness scale for medical students (MAIRS-MS) - development, validity and reliability study. BMC Med Educ 2021 Feb 18;21(1):112. [doi: 10.1186/s12909-021-02546-6] [Medline: 33602196]
60. Uba C, Lewandowski T, Böhmann T. The AI-based transformation of organizations: the 3D-model for guiding enterprise-wide AI change. Presented at: Proceedings of the 56th Hawaii International Conference on System Sciences; Jan 3-7, 2023 URL: https://scholarspace.manoa.hawaii.edu/items/5f90237d-7ad2-444c-9cba-66c95c0ac1e0 [accessed 2026-01-22] [doi: 10.24251/HICSS.2023.743]
61. Jöhnk J, Weißert M, Wyrtki K. Ready or not, AI comes— an interview study of organizational AI readiness factors. Bus Inf Syst Eng 2021;63:5-20. [doi: 10.1007/s12599-020-00676-7]
62. Wiljer D, Salhia M, Dolatabadi E, et al. Accelerating the appropriate adoption of artificial intelligence in health care: protocol for a multistepped approach. JMIR Res Protoc 2021 Oct 6;10(10):e30940. [doi: 10.2196/30940] [Medline: 34612839]
63. Asan O, Bayrak AE, Choudhury A. Artificial intelligence and human trust in healthcare: focus on clinicians. J Med Internet Res 2020 Jun 19;22(6):e15154. [doi: 10.2196/15154] [Medline: 32558657]
64. Appelganc K, Rieger T, Roesler E, Manzey D. How much reliability is enough? A context-specific view on human interaction with (artificial) agents from different perspectives. J Cogn Eng Decis Mak 2022;16(4):207-221. [doi: 10.1177/15553434221104615]
65. Begun JW, Zimmerman B, Dooley K. Health care organizations as complex adaptive systems. In: Mick SM, Wyttenbach M, editors. Advances in Health Care Organization Theory: Jossey-Bass; 2003:253-288 URL: https://citeseerx.ist.psu.edu/document?doi=1f1da1cc58471e929b11090811479441057fff6f&repid=rep1&type=pdf [accessed 2026-01-22]
66. Borghi J, Ismail S, Hollway J, et al. Viewing the global health system as a complex adaptive system - implications for research and practice. F1000Res 2022 Oct 7;11:1147. [doi: 10.12688/f1000research.126201.1] [Medline: 37600221]
67. Gibson D, Kovanovic V, Ifenthaler D, Dexter S, Feng S. Learning theories for artificial intelligence promoting learning processes. Br J Educ Technol 2023;54(5):1125-1146. [doi: 10.1111/bjet.13341]
68. Mukhalalati BA, Taylor A. Adult learning theories in context: a quick guide for healthcare professional educators. J Med Educ Curric Dev 2019 Apr 10;6:2382120519840332. [doi: 10.1177/2382120519840332] [Medline: 31008257]
69. Cresswell KM, Worth A, Sheikh A. Actor-Network Theory and its role in understanding the implementation of information technology developments in healthcare. BMC Med Inform Decis Mak 2010 Nov 1;10(1):67. [doi: 10.1186/1472-6947-10-67] [Medline: 21040575]
70. Pourbohloul B, Kieny MP. Complex systems analysis: towards holistic approaches to health systems planning and policy. Bull World Health Organ 2011 Apr 1;89(4):242-242. [doi: 10.2471/BLT.11.087544] [Medline: 21479083]
71. Sober E. The principle of parsimony. Br J Philos Sci 1981 Jun 1;32(2):145-156. [doi: 10.1093/bjps/32.2.145]
72. Rashidi S. Your AI survival guide: scraped knees, bruised elbows, and lessons learned from real-world AI deployments: John Wiley & Sons; 2024.
73. Yang H, Dai T, Mathioudakis N, Knight AM, Nakayasu Y, Wolf RM. Peer perceptions of clinicians using generative AI in medical decision-making. NPJ Digit Med 2025 Aug 18;8(1):530. [doi: 10.1038/s41746-025-01901-x] [Medline: 40826224]
74. Betty platform. URL: www.asahq.org/beaconbotpublic [accessed 2026-01-23]
75. American society of anesthesiologists announces first-of-its-kind medical specialty society collaboration with Epic to create anesthesia community registry. American Society of Anesthesiologists (ASA). 2025 May. URL: https://www.asahq.org/about-asa/newsroom/news-releases/2025/05/asa-announces-medical-specialty-society-collaboration-with-epic-to-create-anesthesia-community-registry [accessed 2026-01-22]
76. Pusic MV, Birnbaum RJ, Thoma B, et al. Frameworks for integrating learning analytics with the electronic health record. J Contin Educ Health Prof 2023 Jan 1;43(1):52-59. [doi: 10.1097/CEH.0000000000000444] [Medline: 36849429]
77. Guidance on the responsible use of artificial intelligence (AI) in accredited continuing education (CE). : Accreditation Council for Continuing Medical Education (ACCME); 2025 URL: https://accme.org/wp-content/uploads/2025/12/1098_20251125_Guidance_on_Artificial_Intelligence_in_Accredited_CE_ACCME.pdf [accessed 2026-01-22]
78. Product operating model: the healthcare IT approach that can unleash transformational value. Chartis. 2023. URL: https://www.chartis.com/insights/product-operating-model-healthcare-it-approach-can-unleash-transformational-value [accessed 2026-01-22]
79. Cagan M, et al. Transformed: Moving to the Product Operating Model, 1st edition: Silicon Valley Product Group; 2024. URL: https://www.wiley.com/en-us/Transformed%3A+Moving+to+the+Product+Operating+Model-p-9781119697398

Abbreviations

AI: artificial intelligence

XSL•FO
RenderX

ALEERRT-CA: AI Literacy, Explainability, Ethics, Readiness, Reliability, and learning Theories-Complexity theory and Actor-network theory
ANT: Actor-Network Theory
ASA: American Society of Anesthesiologists
CAS: complex adaptive system
CLT: Cognitive Load Theory
CPD: continuing professional development
CT: Complexity Theory

Edited by B Lesselroth; submitted 23.Nov.2024; peer-reviewed by A Raz, JMI Arockiasamy, K Mouloudj, PH Liao; revised version received 23.Dec.2025; accepted 06.Jan.2026; published 12.Feb.2026.

Please cite as:
Hlede V, Valanci S, D'Antuono GR, Dow H, O'Beirne R, Wiggins R
AI-Enhanced Continuing Professional Development as an Evolving Sociotechnical System: Multimethod Theoretical Framework Development Study
JMIR Med Educ 2026;12:e69156
URL: https://mededu.jmir.org/2026/1/e69156
doi:10.2196/69156

© Vjekoslav Hlede, Sofia Valanci, G Robert D'Antuono, Heather Dow, Ronan O'Beirne, Richard Wiggins. Originally published in JMIR Medical Education (https://mededu.jmir.org), 12.Feb.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Education, is properly cited. The complete bibliographic information, a link to the original publication on https://mededu.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

The Relationship Between Sense of Community and Collaborative Learning: Quantitative Study in Medical Education

Harald Knof[1,2], MD; Markus Berndt[3], PhD; Anne Herrmann-Werner[1], MME, Prof Dr Med; Bernhard Hirt[2], Prof Dr Med; Thomas Shiozawa[2], MME, Dr med

[1]TIME – Tuebingen Institute for Medical Education, Faculty of Medicine, University of Tuebingen, Tuebingen, Germany

[2]Institute of Clinical Anatomy and Cell Analysis, Faculty of Medicine, University of Tuebingen, Elfriede-Aulhorn-Straße 8, Tuebingen, Germany

[3]Institute of Medical Education, LMU University Hospital, LMU Munich, Munich, Germany

Corresponding Author:
Thomas Shiozawa, MME, Dr med
Institute of Clinical Anatomy and Cell Analysis, Faculty of Medicine, University of Tuebingen, Elfriede-Aulhorn-Straße 8, Tuebingen, Germany

Abstract

Background: Medical education has shifted from an individual, teacher-led process to an interactive, group-oriented approach, fostering clinical reasoning and teamwork. Sense of Community (SoC) appears to be a key factor in this process; however, its link to collaborative learning and academic success is underexplored.

Objective: This study investigates this relationship, focusing on SoC dimensions (*connectedness* and *learning*) as well as the different forms of *initiative collaborative learning* and *subordinate collaborative learning*.

Methods: The German form of the Classroom Community Scale (CCS-D) was used to assess SoC. The extent of collaborative learning was measured using the Learning Strategies in Study questionnaire. A total of 331 first-year medical students participated. Data analysis included exploratory factor analysis, correlation analysis, and regression analysis.

Results: SoC showed a moderate positive correlation with the use of collaborative learning strategies (r=0.466; P<.001). *Connectedness* emerged as a significant predictor of collaborative learning (R^2=0.257, adjusted R^2=0.255; $F_{1,329}$=113.990; P<.001). Both *initiative collaborative learning* and *subordinate collaborative learning* are based more strongly on feelings of *connectedness*. The feeling that members of the course depend on oneself had the strongest predictive value for collaborative learning (R^2=0.188, adjusted R^2=0.185; $F_{1,329}$=75.956; P<.001).

Conclusions: A strong SoC promotes the use of collaborative learning strategies, particularly through social *connectedness*. The lesser importance of the *learning* dimension suggests that social bonds appear more influential than shared academic goals, particularly during help seeking, where trust and psychological safety are critical. Targeted support of learning communities may enhance didactic approaches and foster both *initiative collaborative learning* and *subordinate collaborative learning*. Cultivating connectedness and responsibility may support academic adaptation and emotional well-being in the critical early phase of medical education.

(JMIR Med Educ 2026;12:e86893) doi:10.2196/86893

KEYWORDS

medical education; students; sense of community; collaborative learning; interpersonal relations; social identification; social environment; survey; regression analysis; factor analysis

Introduction

Background

Learning has shifted from an individual and teacher-directed process to a more interactive, group-oriented approach [1]. Knowledge is therefore constructed collaboratively to enhance knowledge acquisition, clinical reasoning, and teamwork skills [2-4]. The theoretical roots can be traced back to Piagetian and Vygotskian perspectives, both emphasizing the role of social interactions in learning [5,6]. In the Piagetian theory, sociocognitive conflicts drive learning by exposing individuals to alternative problem-solving strategies [5,7-10]. These conflicts encourage cognitive restructuring, deeper understanding, and skill development [11,12]. The Vygotskian perspective highlights the so-called *zone of proximal development*, where learning occurs through guided social interaction and peer support until learners achieve independence [6,7,10,13].

Contemporary educational research further highlights the importance of leadership, learning culture, and socially

XSL•FO
RenderX

supportive educational environments in shaping successful learning processes [14-18]. Within modern "leadership for learning" approaches, educators are increasingly understood not merely as instructors but as mediators who create psychologically safe, collaborative, and student-centered learning environments [14-17]. More recent perspectives, such as socially shared regulation of learning, conceptualize learning as a process in which learners manage motivation, cognition, and learning activities through interaction, feedback, and shared responsibility within groups [19-21]. Similarly, the communities of practice theory emphasizes that learning and professional development occur through active participation in socially and professionally meaningful communities [22,23].

These developments are also reflected in competency-based medical education frameworks such as entrustable professional activities, which underline the acquisition of professional responsibility, self-regulated learning, feedback, and collaborative practice within authentic clinical contexts [24]. Together, these modern perspectives increasingly conceptualize learning not only as knowledge acquisition but as a socially embedded, collaborative, and professionally oriented developmental process.

Collaborative Learning

Collaborative learning is a widely studied but inconsistently defined concept, often used interchangeably with terms such as *group work* or *tutorial learning* [25]. It refers to situations where 2 or more individuals learn together by joint problem-solving and pursuing shared goals [26,27]. Two forms can be distinguished: *initiative collaborative learning*, which is proactive, and *subordinate collaborative learning*, which is reactive and occurs when students seek help after encountering difficulties [28]. Collaborative learning offers numerous advantages across multiple layers and levels [29]. Cognitively, it promotes argumentation, questioning, and explanation, deepening understanding for both the explainer and the recipient [30-32]. Culturally, it fosters a learning environment that values professional exchange [33]. It encourages the identification of knowledge gaps, explanation, questioning, and clarification, leading to deeper learning [34-37]. Professionally, it supports communication and teamwork skills essential in fields such as medicine, where complex knowledge must be conveyed clearly and collaboratively [38,39].

In addition, collaborative learning influences cognitive, emotional, and motivational regulation, as conceptualized in socially shared regulation of learning [19-21]. Learners engage in self-shared, coshared, and socially shared regulation to refine strategies and sustain engagement [19,21,40]. These shared responsibilities and mutual support create a Sense of Community (SoC), which plays a key role in sustaining motivation and fostering a positive learning environment [1,41].

Sense of Community

Despite extensive research, there is no universally accepted definition of the term "Sense of Community" [42-45]. Some definitions emphasize belonging and emotional connection, while others focus on interdependence and shared identity [46,47]. The study by McMillan and Chavis [44] describes SoC as a feeling of belonging, mutual significance, and confidence that members' needs will be met through group commitment [45]. Core elements include trust, responsibility, shared goals, and mutual interdependence [42,48-51]. The expression of SoC, however, varies by context, particularly in educational settings [48,52,53].

In classrooms, community takes a distinct form shaped by learning goals, structured interactions, and time-limited membership [54]. It consists of 2 main components: *connectedness*, reflecting belonging and trust, and *learning*, representing shared academic goals [42,55]. Learners who experience a high sense of classroom community are better able to coordinate working together to achieve the group's collective goals [56]. A strong classroom community promotes engagement, collaboration, and resilience [57,58]. Weak community ties instead increase the risk of disengagement and dropout [59].

Belonging appears to be a fundamental psychological need that directly influences motivation, engagement, and persistence in educational contexts [60,61]. This perspective is further extended by the concept of communities of practice, which frames learning as participation in a shared professional culture and the gradual development of identity through social interaction [22,23]. Educators therefore play a crucial role in fostering connectedness and collective learning to enhance student satisfaction and success [62,63].

Both SoC and collaborative learning contribute to learning and may lead to academic success. These concepts share commonalities, may be linked, or build a communal foundation for social learning. Till date, there is no research connecting these 2 concepts or investigating both constructs in the same study population. This study aims to unveil the interrelations between SoC and collaborative learning based on 2 validated questionnaires [64-66]. The research questions (RQs) can be formulated as follows:

- RQ1: What is the relationship between SoC and collaborative learning?
- RQ2: How do connectedness and learning aspects influence collaborative learning?
- RQ3: How does the SoC differentially influence *initiative collaborative learning* and *subordinate collaborative learning*?
- RQ4: Which items of the German form of the Classroom Community Scale (CCS-D) are key predictors of collaborative learning?

Methods

Data Collection

First-semester medical students from the winter semester 2021-2022 and summer semester 2022 cohorts were invited to participate in a paper-based survey exploring their learning strategies and perceived SoC. By this time, participants had completed all courses for the first semester of their medical studies. The initial invitation to participate was extended personally by the first author. Students received a €5 (€1=US

XSL•FO
RenderX

$1.15 as of June 8, 2026) expense allowance for their participation.

Ethical Considerations

Prior to the study, written informed consent was obtained from all participants, who were thoroughly informed about the study's opportunities, risks, rights, obligations, and voluntary nature. Data collection was conducted in a pseudonymized format, with all participants consenting to anonymized data publication. Students retained the right to withdraw consent at any time without facing any disadvantages. Ethics approval for the study was granted by the ethics committee at Eberhard Karls University of Tuebingen (reference number 086/2022BO2). The study was carried out in accordance with the Declaration of Helsinki.

Measures

German Version of the Classroom Community Scale

To assess the perceived SoC, the CCS-D was used [64] (Table 1). On the basis of Rovai's original Classroom Community Scale (CCS), this version has been validated for use in German-speaking contexts [64,67].

Table . German form of the Classroom Community Scale.

Item	Subscale	Prompt
1	Connectedness	I feel that students in this course care about each other.
2	Learning	I feel that I am encouraged to ask questions.
3	Connectedness	I feel connected to others in this course.
4i	Learning	I feel that it is hard to get help when I have a question.
5i	Connectedness	I do not feel a spirit of community.
6	Learning	I feel that I receive timely feedback.
7	Connectedness	I feel that this course is like a family.
8i	Learning	I feel uneasy exposing gaps in my understanding.
9i	Connectedness	I feel isolated in this course.
10i	Learning	I feel reluctant to speak openly.
11	Connectedness	I trust others in this course.
12i	Learning	I feel that this course results in only modest learning.
13	Connectedness	I feel that I can rely on others in this course.
14i	Learning	I feel that other students do not help me learn.
15	Connectedness	I feel that members of this course depend on me.
16	Learning	I feel that I am given ample opportunities to learn.
17i	Connectedness	I feel uncertain about others in this course.
18i	Learning	I feel that my educational needs are not being met.
19	Connectedness	I feel confident that others will support me.
20i	Learning	I feel that this course does not promote a desire to learn.

The CCS consists of 20 items divided into 2 subscales: *connectedness* and *learning*. The *connectedness* subscale focuses on feelings of connection, while the *learning* subscale addresses perceptions of interaction within the learning community. Participants rate each item on a 5-point Likert scale. The scoring for items 1, 2, 3, 6, 7, 11, 13, 15, 16, and 19 is as follows: *strongly agree=4, agree=3; neutral=2, disagree=1, and strongly disagree=0*. For items 4, 5, 8, 9, 10, 12, 14, 17, 18, and 20, the scoring scale is reversed: *strongly agree=0, agree=1, neutral=2, disagree=3, and strongly disagree=4* [67]. For further analysis, these item scores must be inverted. These items are labeled with an "i" in the tables and figures of this paper.

To calculate the total CCS score (maximum score=80), the values of all 20 items are summed. Each subscale has a maximum score of 40. Higher scores on the total CCS score reflect a stronger sense of classroom community, while lower scores indicate a weaker sense of classroom community [67].

Learning Strategies in Study Questionnaire

To assess the use of collaborative learning strategies, the Learning Strategies in Study (LIST) questionnaire found in the

XSL•FO
RenderX

study by Wild et al [65,66] was used (Table 2). In this survey, only the *learning with fellow students* subscale of the LIST questionnaire was used. Items 1 to 4 form the subscale for *initiative collaborative learning*, while items 5 to 7 form the subscale for *subordinate collaborative learning* [28]. Participants rate the frequency of their use of each learning strategy on a 5-point Likert scale (*1=very seldom*, *2=seldom*, *3=sometimes*, *4=often*, and *5=very often*) [65]. The extent of the use of collaborative learning strategies was determined by calculating the scale mean, with higher values indicating greater use.

Table . Learning Strategies in Study scale—learning with fellow students.

Item	Prompt
1	I work on texts or assignments together with my fellow students.
2	I take time to discuss the material with fellow students.
3	I compare my lecture notes with those of my fellow students.
4	I have a fellow student quiz me and ask him questions about the material as well.
5	I enlist the help of others when I have serious comprehension problems.
6	If something is not clear to me, I ask a fellow student for advice.
7	If I discover major gaps in my records, I contact my fellow students.

Data Analysis

Following a descriptive analysis, Pearson correlation coefficients were calculated, and linear regression analyses were conducted. Subsequently, a principal component analysis (PCA) with varimax rotation and Kaiser normalization was conducted to verify the underlying structure and confirm the division into *subordinate collaborative learning* and *initiative collaborative learning* of the *learning with fellow students* scale. The PCA used listwise deletion, and items with factor loadings ≤ |0.30| on one or both dimensions were retained [68,69]. Prior to the PCA, the Kaiser-Meyer-Olkin measure of sampling adequacy and Bartlett test of sphericity were assessed. The PCA results served as the foundation for subsequent stepwise regression analyses. The significance level for evaluating results was set at .05.

SPSS Statistics (version 29; IBM Corp) was used for collecting questionnaire responses, data extraction, and statistical analysis.

Results

Sample

A total of 344 first-semester medical students were surveyed. With a total number of 210 students per semester, a maximum of 420 participants could have been included, resulting in a response rate of 81.9% (344/420). Participants with missing data were excluded, resulting in a final sample of 331 respondents. The study sample reflected the gender distribution of the cohort, with 68.9% (n=228) of participants being female, and the ages of the participants ranged from 18 to 35 years, with a mean age of 20.72 (SD 2.72) years. The majority of participants were medical students (320/331, 96.7%), followed by dentistry students (10/331, 3%) and students from molecular medicine (1/331, 0.3%).

Scale Characteristics and Internal Consistency

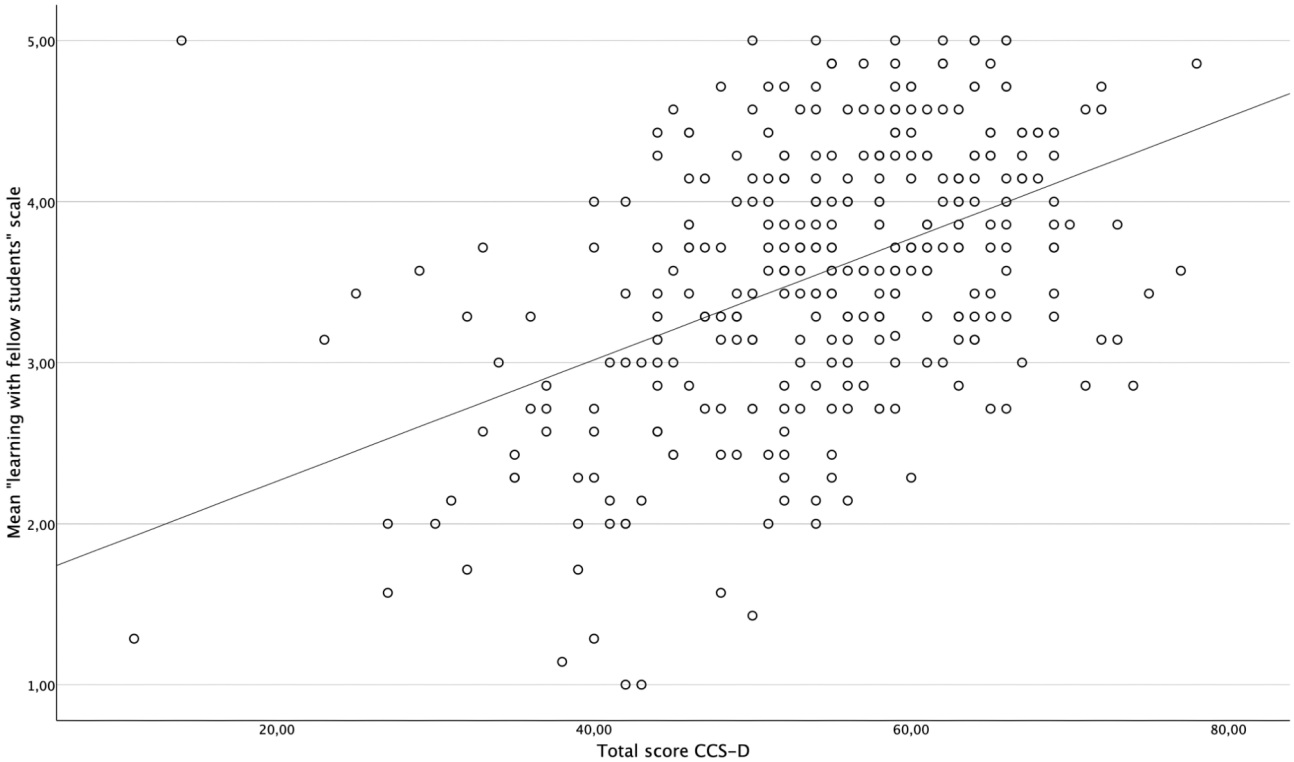
Descriptive statistics for the individual items, as well as the total CCS-D score and its subscale scores, are presented in Table S1 in Multimedia Appendix 1. The total CCS-D demonstrated good internal consistency (Cronbach α=0.87) as a measure for reliability. The *connectedness* subscale (Cronbach α=0.84) also showed good internal consistency, while the *learning* subscale (Cronbach α=0.77) showed acceptable internal consistency. Similarly, Table S2 in Multimedia Appendix 1) presents descriptive statistics of the LIST questionnaire along with the total score on the *learning with fellow students* scale and its subscales. The total *learning with fellow students* scale showed good internal consistency (Cronbach α=0.86). The subscales for *initiative collaborative learning* (Cronbach α=0.82) and *subordinate collaborative learning* (Cronbach α=0.84) demonstrated high reliability.

RQ1: Relationship Between SoC and Collaborative Learning

To measure the relationship between SoC and collaborative learning, Pearson correlation was conducted. The correlation coefficient for the relationship between the SoC and the *learning with fellow students* scale was found to be r=0.466; P<.001. Higher levels of SoC were associated with greater use of collaborative learning strategies (Figure 1).

XSL•FO
RenderX

Figure 1. Scatterplot of the total score of the German form of the Classroom Community Scale (CCS-D) and the *learning with fellow students* subscale of the Learning Strategies in Study questionnaire (LIST). Each point represents an individual student (N=331).



To further examine this relationship, the next step was to examine the association between SoC and collaborative learning through linear regression analysis. The regression model explained 21.7% of the variance in the use of collaborative learning strategies (R^2=0.217, adjusted R^2=0.214; $F_{1,329}$=91.020; P<.001). The unstandardized coefficient was B=0.038 and the standardized coefficient was ß=0.466, both statistically significant at P<.001. SoC was significantly associated with the use of collaborative learning strategies.

RQ2: Influence of *Connectedness* and *Learning* on Collaborative Learning

To determine whether *connectedness* or *learning* is more strongly associated with collaborative learning, a stepwise regression analysis was conducted using CCS-D subscales *connectedness* and *learning* as predictors, with the *learning with fellow students* scale as the dependent variable (Table S3 in Multimedia Appendix 1).

The model including only *connectedness* yielded significant results: R^2=0.257, adjusted R^2=0.255; $F_{1,329}$=113.990; P<.001. The unstandardized coefficient B=0.070 and the standardized coefficient ß=0.507 were both significant at P<.001. The *learning* subscale was excluded from the model (β=0.023; 2-tailed t_{329}=0.382; r=0.021, T=0.651; P=.70).

RQ3: Differences in the Influence of SoC on Initiative Collaborative Learning and Subordinate Collaborative Learning

Principal Component Analysis

To verify the division between *initiative collaborative learning* and *subordinate collaborative learning*, a PCA was conducted. The Kaiser-Meyer-Olkin measure was 0.844, representing a relatively good factor analysis. The Bartlett test of sphericity was significant (approximately χ^2=1120.7; P<.001), suggesting that the analyzed data do not result in an identity matrix and are therefore suitable for factor analysis. Only factors with eigenvalues ≥1 were considered [70,71]. On the basis of the Kaiser criteria and the scree plot, 2 factors with eigenvalues ≥1 were retained, accounting for 70.95% of the total variance. In the varimax-rotated 2-factor model (Table 3), the underlying structure of *subordinate collaborative learning* and *initiative collaborative learning* can be verified, as the items 1 to 4 are loading on factor 1 and the items 5 to 7 on factor 2 [28].

XSL•FO
RenderX

Table . Principal component analysis of *learning with fellow students*[a,b].

Rotated component matrix[c]	Component	
	1	2
Item 1	0.840	—[d]
Item 2	0.838	—
Item 3	0.641	—
Item 4	0.776	—
Item 5	—	0.854
Item 6	—	0.852
Item 7	—	0.808

[a]Extraction method: principal component analysis. Rotation method: varimax with Kaiser normalization.

[b]Factor loadings <.35 are not reported.

[c]Rotation converged in 3 iterations.

[d]Not applicable.

Stepwise Regression

To investigate which factor was more strongly related to *initiative collaborative learning* and *subordinate collaborative learning* strategies, stepwise regression analyses were conducted using the CCS-D subscales *connectedness* and *learning* as predictors, with the subscales *initiative collaborative learning* and *subordinate collaborative learning* as outcome variables.

For *initiative collaborative learning*, the model including only *connectedness* yielded significant results, with R^2=0.157, adjusted R^2=0.155, $F_{1,329}$=61.419; P<.001 (Table S4 in Multimedia Appendix 1). The unstandardized coefficient was B=0.065 and the standardized coefficient was ß=0.397, both statistically significant at P<.001. The *learning* subscale was not retained from the model (β=–0.013; t_{329}=−0.208; r=−0.012; T=0.651; P=.83).

For *subordinate collaborative learning*, the model including *connectedness* also showed significant results, with R^2=0.277, adjusted R^2=0.275; $F_{1,329}$=125.976; P<.001 (Table S5 in Multimedia Appendix 1). The unstandardized coefficient was B=0.077 and the standardized coefficient was ß=0.526, both statistically significant at P<.001. The *learning* subscale was again not retained from the model (β=0.071; t_{329}=1.214; r=0.067; T=0.651; P=.22).

RQ4: CCS-D Items Associated With Collaborative Learning

To determine which items are most strongly associated with collaborative learning, a stepwise regression analysis was conducted using the CCS-D items as independent variables, with the scale *learning with fellow students* as the dependent variable.

The final model, which included item 15, item 3, item 14i, item 9i, and item 19, yielded significant results: R^2=0.338, adjusted R^2=0.328; $F_{5,325}$=33.145; P<.001. The unstandardized coefficients were B=0.211 (P<.001) for item 15, B=0.130 (P=.007) for item 3, B=0.131 (P=.002) for item 14i, B=0.128 (P=.009) for item 9i, and B=0.138 (P=.01) for item 19. The standardized coefficients ß ranged from ß=0.139 to ß=0.244, indicating that item 15 (“I feel that members of this course depend on me”) showed the strongest association with collaborative learning (ß=0.244), followed by item 14i (ß=0.156) and item 19 (ß=0.139; Table S6 in Multimedia Appendix 1).

Discussion

Principal Findings

This study aimed to explore the relationship between SoC and collaborative learning among first-semester medical students. The results of this study show that perceived SoC is significantly associated with the use of collaborative learning strategies. The underlying structure of *subordinate collaborative learning* and *initiative collaborative learning* was confirmed. *Connectedness* showed a stronger association with collaborative learning than the *learning* dimension. *Connectedness* was more strongly associated with *subordinate collaborative learning* than with *initiative collaborative learning*. Item 15 of the CCS-D (“I feel that the members of this course depend on me”) showed the strongest association with collaborative learning.

RQ1: Relationship Between SoC and Collaborative Learning

The findings resonate with theoretical perspectives that link social and emotional bonds to cooperative learning [5,6]. A stronger SoC is associated with a greater engagement in cooperative learning, supporting prior research highlighting the importance of social belonging in promoting student engagement [42,45].

This is particularly relevant for first-semester medical students who must navigate both academic and social transitions [72]. Up to this point of their studies, many students have little experience with self-organized group learning and must first familiarize themselves with new learning demands and environments [73]. School-based learning strategies must be adapted to the new university requirements at the beginning of studies [74]. In this transitional phase, social connections can help by reducing uncertainties and lowering the threshold for

XSL•FO
RenderX

interaction [75]. Numerous studies show that both learning environment and interaction with fellow students play a central role in the selection and adaptation of learning strategies [76-79]. Collaborative learning helps reduce exam anxiety and related stress, which are known to hinder academic success [80,81]. Students who feel well-prepared are more likely to remain motivated and persist through academic challenges [82,83].

RQ2: Influence of *Connectedness* and *Learning* on Collaborative Learning

In this study, only *connectedness* was significantly associated with collaborative learning, while *learning* was not retained from the analytical models. This suggests that interpersonal connectedness is more strongly associated with collaborative learning behavior than sharing common academic goals. One possible explanation for this lies in the specific structure of medical studies, which is more characterized by external learning objectives (eg, exams and clinical competencies). These standardized and homogeneous goals may be less relevant for collaborative behavior. Studies further show that SoC serves as a mediator between motivation and student contribution during collaborative learning, with the mediation effect being stronger for intrinsic motivation than for extrinsic motivation [84]. Therefore, interpersonal connectedness not only appears to facilitate social integration but also strengthens intrinsic motivation, which in turn significantly contributes to active participation in group learning [84]. The first term of medical school is a phase in which students need to orient themselves both academically and socially as written above. In this orientation phase, feelings of belonging and trust are often more relevant than clearly defined common learning goals. It is therefore plausible that interactions in early study phases primarily serve social exchange and network building, while learning goals gain secondary importance.

Another interpretation is that social distancing during the COVID-19 pandemic may have increased the importance of interpersonal connection over content motivation in developing collaborative learning. During the pandemic, opportunities to build personal relationships and engage in face-to-face interactions were severely restricted, as universities and schools shifted to digital learning environments [85-89]. The participants in this study had to complete high school during that time under remote learning conditions [90]. Many students reported missing in-person interaction and expressed a need to meet and learn together to stay motivated [91,92]. These prior experiences may have intensified the value of social connectedness as a foundational support in their early university learning. In this context, feeling connected to peers may serve not only as a motivator for collaborative learning but also as a compensatory mechanism for previously unmet social and educational needs.

RQ3: Differences in the Influence of SoC on Initiative Collaborative Learning and Subordinate Collaborative Learning

Interestingly, *connectedness* is more strongly associated with *subordinate collaborative learning* than *initiative collaborative learning*. This difference can be understood by considering their distinct natures. *Initiative collaborative learning* is proactive when students actively engage with peers to work through tasks, discuss lecture content, and question one another to strengthen understanding [28]. In contrast, *subordinate collaborative learning* is reactive, arising when students encounter challenges such as unclear concepts or incomplete notes and turn to peers for support to fill those gaps and improve comprehension [28]. This differentiation and our findings suggest that social factors are particularly relevant when students rely on support, whereas initiative is more likely driven by individual traits such as motivation or self-efficacy [93-95].

In highly selective programs such as medicine in Germany, where all students have been admitted based on outstanding academic achievement, asking for help is not only a question of competence but also social safety [96]. When all peers may be considered intellectually capable, students may hesitate to seek help from someone if they fear being judged or feeling help seeking as evidence of incompetence [96-99]. In this context, interpersonal trust becomes a crucial factor for starting collaborative behaviors [100]. Particularly in *subordinate collaborative learning*, students are maybe more likely to turn to peers who create a nonthreatening environment and signal mutual respect. This supports the finding that social connectedness, more than academic motivations and shared learning goals, is associated with collaborative learning strategies. Trust creates a psychologically safe space in which students can take academic risks without fear of judgment [101]. Feeling embedded in a supportive peer community may lower the psychological threshold for admitting uncertainty and promotes help seeking without fear of negative evaluation [100,101].

RQ4: CCS-D Items Associated With Collaborative Learning

Particularly noteworthy is the role of item 15 of the CCS-D (“I feel that the members of this course depend on me”), which showed the strongest association with collaborative learning. This highlights the importance of perceived responsibility within a learning community and aligns with the concept of a “Sense of Community Responsibility” in collaborative environments [50,51].

Accordingly, psychological experiences of responsibility and commitment among members of a community are significantly linked to their attitudes and behaviors within collaborative processes [102]. It has also been demonstrated that collaborative learning fosters a sense of accountability, as students feel responsible not only for their own performance but also for that of their peers [103]. Research has shown that individuals tend to emphasize group cohesion when the value of the group is threatened, with such reactions being most pronounced by those members who already feel a strong sense of commitment toward other group members [102,104]. Feeling that others depend on one’s contribution may strengthen identification with the group, which in turn increases the motivation to participate actively and contribute meaningfully. This aligns with social identity theory, which posits that individuals derive part of their self-concept from group membership [105]. Within learning contexts, a sense of connectedness and perceived responsibility toward peers can foster intrinsic motivation and engagement, thereby promoting collaborative behaviors and support [106].

XSL•FO
RenderX

It is also worth discussing whether the present finding may, in part, reflect a form of "helper syndrome," in which individuals derive a sense of self-worth from being needed by others and perceive their peers as dependent on their contributions [107]. In the context of health care professionals and medical education, where altruism, care, and accountability are core professional values, this perceived responsibility may be reinforced by internalized norms of helping and supporting others. The statement "they depend on me" likely captures a positive perception of having a meaningful role within the group, which can be intrinsically motivating for collaborative learning as individuals experience social worth when they feel needed [108,109]. While such attitudes can strengthen engagement and cooperative behaviors, they may also be associated with a tendency to prioritize others' needs over one's own well-being, or to underestimate personal limits, which are characteristic aspects of helper syndrome [110].

Strengths and Limitations

This study has several strengths. Data were collected across multiple cohorts, providing a broader and more generalizable view of first-semester medical students. The high response rate reduces selection bias, and the use of validated instruments strengthens construct validity for both SoC and collaborative learning measures. Robust statistical analyses, including principal component and regression analyses, further enhance methodological rigor.

However, some limitations should be noted. The LIST questionnaire may not fully capture the range of collaborative learning strategies actually used, and self-reported data are susceptible to recall and response bias. Furthermore, all students participated in fixed-group laboratory courses and mentoring programs that fostered collaboration and community, which may have elevated CCS-D scores. Still, as all participants experienced the same conditions, a consistent baseline across the sample can be assumed. A moderate proportion of variance in collaborative learning behavior was explained by the models, while a substantial proportion remained unexplained. This may be attributed to the use of stepwise regression analyses, which are prone to chance findings, thereby potentially compromising the stability and generalizability of the identified predictors. Similarly, the cross-sectional design does not allow for causal inferences, meaning that the observed relationships can only be interpreted as associations. The results should therefore be interpreted with appropriate caution.

Conclusions

A strong SoC appears to be closely associated with collaborative learning in medical education. Social *connectedness* showed a stronger association with collaborative learning than shared academic goals. Both *initiative collaborative learning* and *subordinate collaborative learning* benefited from stronger community ties, with the effect being most pronounced for reactive peer support. The feeling that course members depend on one's own contributions showed the strongest association for collaborative learning.

These findings suggest that deliberately cultivating SoC through structured group work, peer networks, and mentoring may support not only learning outcomes but also teamwork skills and collaboration. Beyond the academic context, fostering a strong sense of belonging and mutual support may better prepare future health care professionals for the collaborative demands of clinical practice, where trust, communication, and cooperation are essential.

Acknowledgments

The authors would like to express their gratitude to the participating students. The authors would also like to acknowledge the support from the Open Access Publication Fund of the University of Tuebingen.

Excerpts from this manuscript were presented at the Annual Conference of the DACH Association for Medical Education in Düsseldorf 2025 [110].

The authors used DeepL Write to check individual sentences for grammar and spelling. They did not check entire paragraphs nor did they use artificial intelligence to generate text. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the final content of the publication.

Funding

This study was funded by the Faculty of Medicine Tuebingen.

Data Availability

The datasets used and analyzed during this study are available from the corresponding author on reasonable request.

Authors' Contributions

Material preparation, data collection, and analysis were performed by HK. The manuscript was written by HK. MB, AH-W, BH, and TS contributed to the manuscript in revising it critically for important intellectual content. AH-W, MB, and BH made substantial contributions to the conception of the study. TS oversaw the overall direction of the study. All authors approved the final version of the manuscript to be published and are accountable for all aspects of the work.

XSL•FO
RenderX

Conflicts of Interest

None declared.

Multimedia Appendix 1
Descriptive statistics and stepwise regression analyses.
[DOCX File, 41 KB - mededu_v12i1e86893_app1.docx]

References

1. Järvelä S, Kirschner PA, Panadero E, et al. Enhancing socially shared regulation in collaborative learning groups: designing for CSCL regulation tools. Educ Technol Res Dev 2015 Feb;63:125-142. [doi: 10.1007/s11423-014-9358-1]
2. Joshi T, Budhathoki P, Adhikari A, Poudel A, Raut S, Shrestha DB. Team-based learning among health care professionals: a systematic review. Cureus 2022 Jan;14(1):e21252. [doi: 10.7759/cureus.21252] [Medline: 35178311]
3. Chawla S, Pawlina W, Hafferty F, Lachman N. The hidden learning environment: effect of informal student networks on team performance in a gross anatomy course. FASEB J 2015 Apr;29(S1). [doi: 10.1096/fasebj.29.1_supplement.344.2]
4. Bergman EM, Sieben JM, Smailbegovic I, de Bruin AB, Scherpbier A, van der Vleuten CP. Constructive, collaborative, contextual, and self-directed learning in surface anatomy education. Anat Sci Educ 2013;6(2):114-124. [doi: 10.1002/ase.1306] [Medline: 22899567]
5. Piaget J. Equilibration of Cognitive Structures: The Central Problem of Intellectual Development: University of Chicago Press; 1985. URL: https://archive.org/details/equilibrationofc0000piag/page/n21/mode/2up [accessed 2026-06-19]
6. Vygotsky LS, Cole M, editors. Mind in Society: Development of Higher Psychological Processes: Harvard University Press; 1978. URL: https://books.google.co.in/books/about/Mind_in_Society.html?id=RxjjUefze_oC&redir_esc=y [accessed 2026-06-19]
7. Brown AL, Palincsar AS. Guided, cooperative learning and individual knowledge acquisition. In: Resnick LB, editor. Knowing, Learning, and Instruction: Routledge; 1989:393-451. [doi: 10.4324/9781315044408]
8. Doise W. The development of individual competencies through social interaction. In: Foot HC, Morgan MJ, Shute RH, editors. Children Helping Children: John Wiley & Sons; 1990:43-64.
9. Doise W, Mugny G. The Social Development of the Intellect: Pergamon Press; 1984. URL: https://api.pageplace.de/preview/DT0400.9781483286105_A23884450/preview-9781483286105_A23884450.pdf [accessed 2026-06-19]
10. Palincsar AS. Social constructivist perspectives on teaching and learning. Annu Rev Psychol 1998;49:345-375. [doi: 10.1146/annurev.psych.49.1.345] [Medline: 15012472]
11. Perret-Clermont AN. In: Tajfel H, editor. Social Interaction and Cognitive Development in Children: Academic Press; 1980. URL: https://libra.unine.ch/entities/publication/088ab9c9-febd-4108-ab13-b0861b0ab93a [accessed 2026-06-19]
12. Hatano G, Inagaki K. Sharing cognition through collective comprehension activity. In: Resnick LB, Levine JM, Teasley SD, editors. Perspectives on Socially Shared Cognition, 1st edition: American Psychological Association; 1991:331-348. [doi: 10.1037/10096-014]
13. Foot HC, Howe C. The psycho-educational basis of peer-assisted learning. In: Topping K, Ehly S, editors. Peer-Assisted Learning, 1st edition: Lawrence Erlbaum; 1998:27-43.
14. Papadakis S, Kanadlı S, Kardas A, Tülübaş T, Karakose T, Polat H. Investigating the relationship between leadership for learning and student achievement through the mediation of teacher performance: a meta-analytic structural equation modeling (MASEM) approach. Educ Sci 2024;14(12):1320. [doi: 10.3390/educsci14121320]
15. Harden RM, Crosby J. AMEE guide no 20: the good teacher is more than a lecturer - the twelve roles of the teacher. Med Teach 2000 Jan;22(4):334-347. [doi: 10.1080/014215900409429]
16. Daniëls E, Hondeghem A, Dochy F. A review on leadership and leadership development in educational settings. Educ Res Rev 2019 Jun;27:110-125. [doi: 10.1016/j.edurev.2019.02.003]
17. Järvelä S, Hadwin AF. New frontiers: regulating learning in CSCL. Educ Psychol 2013 Jan;48(1):25-39. [doi: 10.1080/00461520.2012.748006]
18. Leithwood K, Louis KS, Anderson S, Wahlstrom K. How leadership influences student learning. The Wallace Foundation. 2004. URL: https://wallacefoundation.org/sites/default/files/2023-07/How-Leadership-Influences-Student-Learning.pdf [accessed 2026-06-19]
19. Hadwin A, Järvelä S, Miller M. Self-regulation, co-regulation, and shared regulation in collaborative learning environments. In: Schunk DH, Greene JA, editors. Handbook of Self-Regulation of Learning and Performance, 2nd edition: Routledge; 2017. [doi: 10.4324/9781315697048]
20. Järvenoja H, Järvelä S, Malmberg J. Supporting groups' emotion and motivation regulation during collaborative learning. Learn Instr 2020 Dec;70:101090. [doi: 10.1016/j.learninstruc.2017.11.004]
21. Wenger E. Communities of practice and social learning systems: the career of a concept. In: Blackmore C, editor. Social Learning Systems and Communities of Practice: Springer; 2010:179-198. [doi: 10.1007/978-1-84996-133-2_11]
22. Wenger E. Communities of Practice: Learning, Meaning, and Identity: Cambridge University Press; 1998. [doi: 10.1017/CBO9780511803932]

XSL•FO
RenderX

23. ten Cate O. Entrustability of professional activities and competency-based training. Med Educ 2005 Dec;39(12):1176-1177. [doi: 10.1111/j.1365-2929.2005.02341.x] [Medline: 16313574]
24. Pauli C, Reusser K. Zur Rolle der Lehrperson beim kooperativen Lernen. Swiss J Educ Res 2000;22(3):421-442. [doi: 10.24452/sjer.22.3.4585]
25. Dillenbourg P. What do you mean by collaborative learning? In: Dillenbourg P, editor. Collaborative Learning: Cognitive and Computational Approaches: Elsevier; 1999:1-19 URL: https://telearn.hal.science/hal-00190240v1/document [accessed 2026-06-08]
26. Roschelle J, Teasley SD. The construction of shared knowledge in collaborative problem solving. In: O'Malley C, editor. Computer Supported Collaborative Learning: Springer; 1995:69-97. [doi: 10.1007/978-3-642-85098-1_5]
27. Knof H, Berndt M, Shiozawa T. The influence of collaborative learning and self-organisation on medical students' academic performance in anatomy. Ann Anat 2024 Jan;251:152182. [doi: 10.1016/j.aanat.2023.152182] [Medline: 37879501]
28. Nokes-Malach TJ, Richey JE, Gadgil S. When is it better to learn together? Insights from research on collaborative learning. Educ Psychol Rev 2015 Dec;27:645-656. [doi: 10.1007/s10648-015-9312-8]
29. Bereiter C, Scardamalia M. Learning to work creatively with knowledge. In: De Corte E, Verschaffel L, Entwistle N, van Merrienboer J, editors. Powerful Learning Environments: Unravelling Basic Components and Dimensions: Elsevier Science BV; 2003:55-68 URL: https://books.google.co.in/books/about/Powerful_Learning_Environments.html?id=atTUfmfJVMAC&source=kp_book_description&redir_esc=y [accessed 2026-06-19]
30. Vuopala E, Näykki P, Isohätälä J, Järvelä S. Knowledge co-construction activities and task-related monitoring in scripted collaborative learning. Learn Cult Soc Interact 2019 Jun;21:234-249. [doi: 10.1016/j.lcsi.2019.03.011]
31. Baker MJ. A model for negotiation in teaching-learning dialogues. J Artif Intell Educ 1994;5(2):199-254 [FREE Full text]
32. Renkl A, Mandl H. Kooperatives Lernen: die Frage nach dem Notwendigen und dem Ersetzbaren [Article in German]. Unterrichtswissenschaft 1995;23(4):292-300 [FREE Full text]
33. Webb NM. Peer interaction and learning in small groups. Int J Educ Res 1989 Jan;13:21-39. [doi: 10.1016/0883-0355(89)90014-1]
34. Webb NM, Troper JD, Fall R. Constructive activity and learning in collaborative small groups. J Educ Psychol 1995;87(3):406-423. [doi: 10.1037/0022-0663.87.3.406]
35. Nieder GL, Parmelee DX, Stolfi A, Hudes PD. Team-based learning in a medical gross anatomy and embryology course. Clin Anat 2005 Jan;18(1):56-63. [doi: 10.1002/ca.20040] [Medline: 15597377]
36. Offir B, Lev Y, Bezalel R. Surface and deep learning processes in distance education: synchronous versus asynchronous systems. Comput Educ 2008 Nov;51(3):1172-1183. [doi: 10.1016/j.compedu.2007.10.009]
37. Kemp K, Baxa D, Cortes C. Exploration of a collaborative self-directed learning model in medical education. Med Sci Educ 2022 Jan;32(1):195-207. [doi: 10.1007/s40670-021-01493-7] [Medline: 35003877]
38. Langewitz W. Physician-patient communication in medical education: can it be learned? [Article in German]. Bundesgesundheitsblatt Gesundheitsforschung Gesundheitsschutz 2012 Sep;55(9):1176-1182. [doi: 10.1007/s00103-012-1533-0] [Medline: 22936486]
39. Hadwin AF, Bakhtiar A, Miller M. Challenges in online collaboration: effects of scripting shared task perceptions. Intern J Comput Support Collab Learn 2018 Sep;13:301-329. [doi: 10.1007/s11412-018-9279-9]
40. Reeves T, Gomm P. Community and contribution: factors motivating students to participate in an extra-curricular online activity and implications for learning. E-Learn Digit Media 2015;12(3-4):391-409. [doi: 10.1177/2042753015571828]
41. Rovai AP. Sense of community, perceived cognitive learning, and persistence in asynchronous learning networks. Internet High Educ 2002;5(4):319-332. [doi: 10.1016/S1096-7516(02)00130-6]
42. Westheimer J, Kahne J. Building school communities: an experience-based model. Phi Delta Kappan 1993;75(4):324-328 [FREE Full text]
43. Graves LN. Cooperative learning communities: context for a new vision of education and society. J Educ 1992 Apr;174(2):57-79. [doi: 10.1177/002205749217400205]
44. McMillan DW, Chavis DM. Sense of community: a definition and theory. J Community Psychol 1986 Jan;14(1):6-23. [doi: 10.1002/1520-6629(198601)14:1<6::AID-JCOP2290140103>3.0.CO;2-I]
45. Sarason SB. The Psychological Sense of Community: Prospects for a Community Psychology: Jossey-Bass; 1974. URL: https://books.google.co.in/books/about/The_Psychological_Sense_of_Community.html?id=-DK3AAAAIAAJ&redir_esc=y [accessed 2026-06-19]
46. Unger DG, Wandersman A. The importance of neighbors: the social, cognitive, and affective components of neighboring. Am J Commun Psychol 1985 Apr;13:139-169. [doi: 10.1007/BF00905726]
47. Hill JL. Psychological sense of community: suggestions for future research. J Community Psychol 1996 Oct;24(4):431-438. [doi: 10.1002/(SICI)1520-6629(199610)24:4<431::AID-JCOP10>3.0.CO;2-T]
48. Shaffer CR, Anundsen K. Creating Community Anywhere: Finding Support and Connection in a Fragmented World: Tarcher; 1993. URL: https://scholar.google.com/scholar_lookup?title=Creating%20community%20anywhere%3A%20finding%20support%20and%20connection%20in%20a%20fragmented%20world&publication_year=1993&author=Shaffer%2CC.R&author=Anundsen%2CK [accessed 2026-06-19]

XSL•FO
RenderX

49. Nowell B, Boyd NM. Sense of community responsibility in community collaboratives: advancing a theory of community as resource and responsibility. Am J Community Psychol 2014 Dec;54(3-4):229-242. [doi: 10.1007/s10464-014-9667-x] [Medline: 25047759]
50. Nowell B, Boyd N. Viewing community as responsibility as well as resource: deconstructing the theoretical roots of psychological sense of community. J Community Psychol 2010 Sep;38(7):828-841. [doi: 10.1002/jcop.20398]
51. Rheingold H. The Virtual Community: Finding Connection in a Computerized World, 1st edition: Secker & Warburg; 1993. URL: https://books.google.co.in/books/about/The_Virtual_Community.html?id=0XyYEV9Ev8gC&redir_esc=y [accessed 2026-06-19]
52. Strayhorn TL. College Students' Sense of Belonging: A Key to Educational Success for All Students, 2nd edition: Routledge; 2018. [doi: 10.4324/9781315297293]
53. Rovai AP. Building and sustaining community in asynchronous learning networks. Internet High Educ 2000;3(4):285-297. [doi: 10.1016/S1096-7516(01)00037-9]
54. McMillan DW. Sense of community. J Community Psychol 1996 Oct;24(4):315-325. [doi: 10.1002/(SICI)1520-6629(199610)24:4<315::AID-JCOP2>3.0.CO;2-T]
55. Watkins C. Classrooms as Learning Communities: What's in It for Schools?: Routledge; 2005. [doi: 10.4324/9780203390719]
56. Tinto V. Classrooms as communities: exploring the educational character of student persistence. J High Educ 1997;68(6):599-623. [doi: 10.1080/00221546.1997.11779003]
57. Keyes TS, Heath RD. The association between teaching practices and students' perceptions of being in a classroom community of engaged learners. Sch Community J 2023;33(1):251-278 [FREE Full text]
58. Sheets MF. Characteristics of adult education students and factors which determine course completion: a review. New Horiz Adult Educ Hum Resour Dev 1992;6(1):3-18. [doi: 10.1002/nha3.10034]
59. Yeager DS, Walton GM. Social-psychological interventions in education: they're not magic. Rev Educ Res 2011 Jun;81(2):267-301. [doi: 10.3102/0034654311405999]
60. Baumeister RF, Leary MR. The need to belong: desire for interpersonal attachments as a fundamental human motivation. Psychol Bull 1995 May;117(3):497-529. [doi: 10.1037/0033-2909.117.3.497] [Medline: 7777651]
61. Bruffee KA. Collaborative Learning: Higher Education, Interdependence, and the Authority of Knowledge: Johns Hopkins University Press; 1993. URL: https://books.google.co.in/books/about/Collaborative_Learning.html?id=1cedAAAAMAAJ&redir_esc=y [accessed 2026-06-19]
62. Dede C. The evolution of distance education: emerging technologies and distributed learning. Am J Distance Educ 1996;10(2):4-36. [doi: 10.1080/08923649609526919]
63. Knof H, Shiozawa T. Validation of the German form of the Classroom Community Scale (CCS-D). GMS J Med Educ 2025;42(1):Doc13. [doi: 10.3205/zma001737] [Medline: 40395962]
64. Wild KP, Schiefele U. Lernstrategien im Studium: Ergebnisse zur Faktorenstruktur und Reliabilität eines neuen Fragebogens [Article in German]. Zeitschrift Differentielle Diagnostische Psychologie 1994;15(4):185-200 [FREE Full text]
65. Wild KP, Schiefele U, Winteler A. List: Ein Verfahren Zur Erfassung von Lernstrategien Im Studium [Book in German]: Institut für Erziehungswissenschaft und Pädagogische Psychologie, Universität der Bundeswehr München; 1992. URL: https://www.fachportal-paedagogik.de/literatur/vollanzeige.html?FId=2248205 [accessed 2026-06-09]
66. Rovai AP. Development of an instrument to measure classroom community. Internet High Educ 2002 Sep;5(3):197-211. [doi: 10.1016/S1096-7516(02)00102-1]
67. Comrey AL, Lee HB. A First Course in Factor Analysis: Psychology Press; 1992. [doi: 10.4324/9781315827506]
68. Kline RB. Principles and Practice of Structural Equation Modeling: Guilford Publications; 1998. URL: https://books.google.co.in/books/about/Principles_and_Practice_of_Structural_Eq.html?id=JGVuQgAACAAJ&redir_esc=y [accessed 2026-06-19]
69. Kaiser HF. The application of electronic computers to factor analysis. Educ Psychol Meas 1960 Apr;20(1):141-151. [doi: 10.1177/001316446002000116]
70. Guttman L. Some necessary conditions for common-factor analysis. Psychometrika 1954 Jun;19:149-161. [doi: 10.1007/BF02289162]
71. Picton A, Greenfield S, Parry J. Why do students struggle in their first year of medical school? A qualitative study of student voices. BMC Med Educ 2022 Feb 16;22(1):100. [doi: 10.1186/s12909-022-03158-4] [Medline: 35172820]
72. Fabry G, Giesler M. Novice medical students: individual patterns in the use of learning strategies and how they change during the first academic year. GMS Z Med Ausbild 2012;29(4):Doc56. [doi: 10.3205/zma000826] [Medline: 22916082]
73. Schiefele U, Streblow L, Ermgassen U, Moschner B. Lernmotivation und Lernstrategien als Bedingungen der Studienleistung [Article in German]. Zeitschrift Pädagogische Psychologie 2003;17(3/4):185-198. [doi: 10.1024/1010-0652.17.3.185]
74. Worsley JD, Harrison P, Corcoran R. Bridging the gap: exploring the unique transition from home, school or college into university. Front Public Health 2021;9:634285. [doi: 10.3389/fpubh.2021.634285] [Medline: 33816421]
75. Entwistle NJ. Approaches to learning and forms of understanding. In: Dart B, Boulton-Lewis G, editors. Teaching and Learning in Higher Education: From Theory to Practice: Australian Council for Educational Research; 1998:72-101 URL: https://books.google.co.in/books/about/Teaching_and_Learning_in_Higher_Educatio.html?id=8Rs3-TNRGccC&redir_esc=y [accessed 2026-06-19]

XSL•FO
RenderX

76. Entwistle N, Ramsden P. Understanding Student Learning: Routledge; 1983. [doi: 10.4324/9781315718637]
77. Wierstra RF, Kanselaar G, van der Linden JL, Lodewijks H, Vermunt JD. The impact of the university context on European students' learning approaches and learning environment preferences. High Educ 2003 Jun;45:503-523. [doi: 10.1023/A:1023981025796]
78. Trigwell K, Prosser M. Relating approaches to study and quality of learning outcomes at the course level. Br J Educ Psychol 1991 Nov;61(3):265-275. [doi: 10.1111/j.2044-8279.1991.tb00984.x]
79. Laakkonen J, Muukkonen H. Fostering students' collaborative learning competencies and professional conduct in the context of two gross anatomy courses in veterinary medicine. Anat Sci Educ 2019 Mar;12(2):154-163. [doi: 10.1002/ase.1811] [Medline: 30053330]
80. Pintrich PR, De Groot EV. Motivational and self-regulated learning components of classroom academic performance. J Educ Psychol 1990;82(1):33-40. [doi: 10.1037/0022-0663.82.1.33]
81. Ames C, Archer J. Achievement goals in the classroom: students' learning strategies and motivation processes. J Educ Psychol 1988;80(3):260-267. [doi: 10.1037/0022-0663.80.3.260]
82. Smit K, de Brabander CJ, Boekaerts M, Martens RL. The self-regulation of motivation: motivational strategies as mediator between motivational beliefs and engagement for learning. Int J Educ Res 2017;82:124-134. [doi: 10.1016/j.ijer.2017.01.006]
83. Han J, Jiang Y, Mentzer N, Kelley T. The role of sense of community and motivation in the collaborative learning: an examination of the first-year design course. Int J Technol Des Educ 2022 Jul;32:1837-1852. [doi: 10.1007/s10798-021-09658-6]
84. Ordinance of the State Government on infection protection measures against the spread of the SARS-CoV-2 virus (Corona Ordinance - "CoronaVO"). Ministry of Social Affairs, Health and Integration Baden-Württemberg. 2020. URL: https://sozialministerium.baden-wuerttemberg.de/fileadmin/redaktion/m-sm/intern/downloads/Downloads_Gesundheitsschutz/200502_CoronaVO_konsolidiert_Stand_040520_englisch.pdf [accessed 2026-06-19]
85. Ordinance of the State Government on infection protection measures against the spread of the SARS-CoV-2 virus (Corona Ordinance – "CoronaVO"). Ministry of Social Affairs, Health and Integration Baden-Württemberg. 2022. URL: https://www.baden-wuerttemberg.de/fileadmin/redaktion/m-sm/intern/downloads/Downloads_Gesundheitsschutz/220531_CoronaV_konsolidiert_en.pdf [accessed 2026-06-09]
86. Ferrel MN, Ryan JJ. The impact of COVID-19 on medical education. Cureus 2020 Mar 31;12(3):e7492. [doi: 10.7759/cureus.7492] [Medline: 32368424]
87. Braun T, Qekaj A, Rapp S, König S, Loewe S, Boehringer D. Anregungen für eine Gestaltung des Studieneinstiegs unter Pandemiebedingung: Blaupausen für das digitale Wintersemester 2020/21. Zentrum für Lehre und Weiterbildung. 2020. URL: https://elib.uni-stuttgart.de/server/api/core/bitstreams/4671534c-b60e-4e75-bdb8-c956414295e6/content [accessed 2026-06-09]
88. Andresen S, Lips A, Möller R, Rusack T, Schröer W, Thomas S, et al. Erfahrungen und Perspektiven von jungen Menschen während der Corona-Maßnahmen: erste Ergebnisse der bundesweiten Studie JuCo. : Universitätsverlag Hildesheim; 2020 URL: https://hilpub.uni-hildesheim.de/server/api/core/bitstreams/dcfebb6e-eecb-4da2-af9e-138d9655ebef/content [accessed 2026-06-09]
89. Schneider R, Sachse KA, Schipolowski S, Enke F. Teaching in times of COVID-19: the evaluation of distance teaching in elementary and secondary schools in Germany. Front Educ 2021;6. [doi: 10.3389/feduc.2021.702406]
90. Boulos AN. Evaluation of the effectiveness of online education in anatomy for medical students during the COVID-19 pandemic. Ann Anat 2022 Oct;244:151973. [doi: 10.1016/j.aanat.2022.151973] [Medline: 35738314]
91. Kalmar E, Aarts T, Bosman E, et al. The COVID-19 paradox of online collaborative education: when you cannot physically meet, you need more social interactions. Heliyon 2022 Jan;8(1):e08823. [doi: 10.1016/j.heliyon.2022.e08823] [Medline: 35128108]
92. Zimmerman BJ. Becoming a self-regulated learner: an overview. Theory Pract 2002 May;41(2):64-70. [doi: 10.1207/s15430421tip4102_2]
93. Bandura A. Self-efficacy: toward a unifying theory of behavioral change. Psychol Rev 1977 Mar;84(2):191-215. [doi: 10.1037//0033-295x.84.2.191] [Medline: 847061]
94. Schunk DH, DiBenedetto MK. Motivation and social cognitive theory. Contemp Educ Psychol 2020 Jan;60:101832. [doi: 10.1016/j.cedpsych.2019.101832]
95. Ryan AM, Gheen MH, Midgley C. Why do some students avoid asking for help? An examination of the interplay among students' academic efficacy, teachers' social–emotional role, and the classroom goal structure. J Educ Psychol 1998;90(3):528-535. [doi: 10.1037/0022-0663.90.3.528]
96. Newman RS, Goldin L. Children's reluctance to seek help with schoolwork. J Educ Psychol 1990;82(1):92-100. [doi: 10.1037/0022-0663.82.1.92]
97. Butler R. Determinants of help seeking: relations between perceived reasons for classroom help-avoidance and help-seeking behaviors in an experimental context. J Educ Psychol 1998;90(4):630-643. [doi: 10.1037/0022-0663.90.4.630]
98. Karabenick SA, Knapp JR. Relationship of academic help seeking to the use of learning strategies and other instrumental achievement behavior in college students. J Educ Psychol 1991;83(2):221-230. [doi: 10.1037/0022-0663.83.2.221]

XSL•FO
RenderX

99. Qayyum A. Student help-seeking attitudes and behaviors in a digital era. Int J Educ Technol High Educ 2018 Dec;15(1):17. [doi: 10.1186/s41239-018-0100-7]
100. Edmondson A. Psychological safety and learning behavior in work teams. Adm Sci Q 1999;44(2):350-383. [doi: 10.2307/2666999]
101. Ellemers N, Spears R, Doosje B. Self and social identity. Annu Rev Psychol 2002;53:161-186. [doi: 10.1146/annurev.psych.53.100901.135228] [Medline: 11752483]
102. VanLeuven AJ, Szymik BG, Ramsey LM, Hesse DW. A multi‐year evaluation of medical student performance on and perceptions of collaborative gross anatomy laboratory examinations. Anat Sci Educ 2023;16(3):557-566. [doi: 10.1002/ase.2223] [Medline: 36102481]
103. Turner JC, Hogg MA, Turner PJ, Smith PM. Failure and defeat as determinants of group cohesiveness. Br J Soc Psychol 1984 Jun;23(2):97-111. [doi: 10.1111/j.2044-8309.1984.tb00619.x]
104. Tajfel H, Turner JC. The social identity theory of intergroup behavior. In: Jost JT, Sidanius J, editors. Political Psychology: Key Readings: Psychology Press; 2004. [doi: 10.4324/9780203505984]
105. Ryan RM, Deci EL. Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being. Am Psychol 2000 Jan;55(1):68-78. [doi: 10.1037//0003-066x.55.1.68] [Medline: 11392867]
106. Schmidbauer W. Die hilflosen Helfer: U¨ber die seelische Problematik der helfenden Berufe (German Edition): Rowohlt; 1977.
107. McAdams DP, de St. Aubin E. A theory of generativity and its assessment through self-report, behavioral acts, and narrative themes in autobiography. J Pers Soc Psychol 1992;62(6):1003-1015. [doi: 10.1037/0022-3514.62.6.1003]
108. Grant AM, Gino F. A little thanks goes a long way: explaining why gratitude expressions motivate prosocial behavior. J Pers Soc Psychol 2010 Jun;98(6):946-955. [doi: 10.1037/a0017935] [Medline: 20515249]
109. Maringgele VE, Scherr M, Aichhorn W, Kaiser AK. Helper Syndrome and Pathological Altruism in nurses - a study in times of the COVID-19 pandemic. Front Psychol 2023;14:1150150. [doi: 10.3389/fpsyg.2023.1150150] [Medline: 37901070]
110. Knof H, Berndt M, Shiozawa T. The influence of community spirit on collaborative learning in medical education. In: Annual Meeting of the Society for Medical Education (GMA) [Book in German]: German Medical Science GMS Publishing House; 2025. [doi: 10.3205/25gma135]

Abbreviations

CCS: Classroom Community Scale
CCS-D: German form of the Classroom Community Scale
LIST: Learning Strategies in Study
PCA: principal component analysis
RQ: research question
SoC: Sense of Community

Edited by W Jerjes; submitted 31.Oct.2025; peer-reviewed by JI Shin, JW Li, S Papadakis, T Ungar; revised version received 29.May.2026; accepted 30.May.2026; published 06.Jul.2026.

Please cite as:
Knof H, Berndt M, Herrmann-Werner A, Hirt B, Shiozawa T
The Relationship Between Sense of Community and Collaborative Learning: Quantitative Study in Medical Education
JMIR Med Educ 2026;12:e86893
URL: https://mededu.jmir.org/2026/1/e86893
doi:10.2196/86893

©Harald Knof, Markus Berndt, Anne Herrmann-Werner, Bernhard Hirt, Thomas Shiozawa. Originally published in JMIR Medical Education (https://mededu.jmir.org), 6.Jul.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Education, is properly cited. The complete bibliographic information, a link to the original publication on https://mededu.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Design and Evaluation of a Faculty Development Workshop Series on Integrating Generative Artificial Intelligence in Medical Education: Mixed Methods Pilot Study

Rajalakshmi Anand[1,2*], MBA, EdD; Nicole Bowers[2*], PhD; Mange Festo Manyama[3*], MD, PhD

[1]Institute for Population Health, Weill Cornell Medical College-Qatar, Qatar Foundation—Education City, Doha, Baladīyat ad Daw ah, Qatar

[2]Mary Lou Fulton College for Teaching and Learning Innovation, Arizona State University, Tempe, AZ, United States

[3]Division of Medical Education, Weill Cornell Medical College-Qatar, Doha, Qatar

[*]all authors contributed equally

Corresponding Author:
Rajalakshmi Anand, MBA, EdD
Institute for Population Health, Weill Cornell Medical College-Qatar, Qatar Foundation—Education City, Doha, Baladīyat ad Daw ah, Qatar

Abstract

Background: Generative artificial intelligence (GenAI) tools are being increasingly applied to teaching and learning in medical education creating both instructional opportunities and pedagogical challenges. While GenAI offers potential to enhance teaching, assessment, and curriculum design, many medical faculty lack structured guidance on how to integrate these tools ethically and pedagogically within discipline-specific, high-stakes educational contexts.

Objective: This study aimed to design, implement, and evaluate a faculty development workshop series for ethical and pedagogical integration of GenAI in medical education teaching.

Methods: A mixed methods pilot study was conducted to design, implement, and evaluate a faculty development workshop series "Professional Development in Generative Artificial Intelligence for Pedagogy" at Weill Cornell Medicine-Qatar, a US medical school in Qatar. The program consisted of five 1-hour synchronous online workshops grounded in Experiential Learning Theory and the Technological Pedagogical Content Knowledge framework. Ten medical faculty from multiple disciplines participated. Quantitative data were collected through an online preintervention survey, an online postintervention survey with open-ended questions, and an online 2-week follow-up survey. Surveys consisted of 5-point Likert scale items capturing perceptions of workshop quality, confidence, and intended application. Qualitative data included full workshop transcripts, facilitator theoretical notes, and facilitator memos. Descriptive statistics summarized quantitative findings, while qualitative data were analyzed using a combination of deductive and inductive coding, alongside narrative analysis. Findings were integrated to generate convergent interpretations.

Results: Qualitative analysis of workshop transcripts suggested evolving engagement with GenAI, with participants describing movement from exploratory use toward more intentional pedagogical application. Postintervention survey results indicated high satisfaction with program content, organization, relevance, and overall quality. Two-week follow-up survey responses (n=5) suggested increased self-reported confidence in applying GenAI tools, and perceived shifts in how participants conceptualized teaching with GenAI. Faculty described intended strategies for integrating GenAI into lesson planning, assessment design, visualization of learning materials, and case-based instruction, while emphasizing the importance of human oversight, critical appraisal, and ethical judgment. Findings highlighted the perceived value of hands-on experimentation, reflective discussion, and adaptive facilitation in supporting early faculty engagement.

Conclusions: This pilot study provides early evidence that an experiential, theory-informed, and adaptively facilitated faculty development workshop series may support medical faculty in developing self-reported confidence, awareness, and initial strategies for responsible GenAI integration. Findings are exploratory and limited by a small sample size, a single institution, and reliance on self-reported data. Nevertheless, the Professional Development in Generative Artificial Intelligence for Pedagogy workshop series presents a flexible and theory-informed faculty development approach that may inform future faculty development initiatives in medical education as GenAI technologies continue to evolve.

(JMIR Med Educ 2026;12:e89815) doi:10.2196/89815

XSL•FO
RenderX

KEYWORDS

generative artificial intelligence; medical education; faculty development; experiential learning; TPACK; ethical AI; responsible AI

Introduction

Background

The emergence of generative artificial intelligence (GenAI) tools such as ChatGPT in November 2022 sparked widespread concern and curiosity among medical faculty, many of whom at that time lacked clarity on the capabilities, limitations, and ethical implications of the use of GenAI in medical education (ME) pedagogy. ChatGPT became one of the fastest growing consumer applications in history [1,2]. Unlike static instructional technologies, GenAI tools are protean, generative, opaque, unstable, and context-sensitive [3]. While these characteristics enable novel forms of instructional support, they also introduce uncertainty, particularly when GenAI tools are adopted faster than educators can develop responsible and theory-informed faculty development approaches for their use [4].

At Weill Cornell Medicine-Qatar, a US medical school in Doha, Qatar, the rapid emergence of GenAI technologies highlighted the need to proactively support educators in navigating this emerging landscape [5]. This institutional experience reflected a broader concern of faculty in academic medicine during 2023 - 2024. Although research from related fields highlights the value of ongoing professional development (PD) for building GenAI competence [1], existing PD approaches do not adequately meet the needs of experienced medical educators working under intense time constraints. ME educators are disciplinary experts whose instructional decisions carry professional accountability and implications for patient care [5]. Without structured, contextually relevant learning experiences, GenAI may risk being adopted primarily for efficiency, with limited attention to pedagogical alignment or ethical considerations [6,7]. Inaccuracies, oversimplifications, or uncritical reliance on automated GenAI outputs may shape learners' clinical reasoning and professional formation [8,9]. Additionally, reliance on automated GenAI outputs also raises the risk of cognitive deskilling, whereby higher-order pedagogical reasoning and instructional decision-making may be diminished through overautomation rather than being exercised and refined [10,11].

While studies have demonstrated growing interest in GenAI's potential, few have empirically examined its ethical and pedagogical use within ME settings [9,12,13]. Emerging scholarship has raised concerns that frequent delegation of complex cognitive tasks to GenAI systems may shift educators from active decision-making toward uncritical or nonpurposeful reliance on generated content, with implications for teaching quality and professional expertise [14-16]. These risks underscore the need for faculty development approaches that explicitly position GenAI as a tool for critique and reflection rather than a substitute for pedagogical reasoning. These risks are amplified for experienced educators who are pressed for time and navigating rapidly evolving technologies without structured opportunities for critical examination.

A growing body of ME scholarship has begun to describe institutional readiness and emerging approaches to GenAI training and guidance for faculty and learners. For example, recent work has reported variability in GenAI policy and training availability across ME institutions and highlighted the need for structured support for faculty implementation [17,18]. Other studies have described targeted faculty development workshops aimed at building practical GenAI literacy and teaching-oriented application, underscoring both interest and variability in preparedness [19,20]. This emerging evidence base supports the need for context-specific faculty development interventions that move beyond general awareness toward educator-directed and pedagogically aligned use [21-23]. Unlike many other educational contexts, ME is embedded within frameworks of professional accountability and ethical responsibility. As a result, integrating GenAI into teaching requires approaches that preserve the central role of human expertise and judgment. Faculty development initiatives must therefore move beyond tool familiarization to support reflective, context-sensitive engagement with GenAI outputs, positioning these tools as assistive resources rather than substitutes for pedagogical reasoning.

Guided by these considerations, we designed the Professional Development in Generative Artificial Intelligence for Pedagogy (PDGenAI-P) workshop series as a faculty development intervention for medical educators [5]. Grounded in the Technological Pedagogical Content Knowledge (TPACK) [3,24] framework and Experiential Learning Theory (ELT) [25,26], PDGenAI-P was structured to support hands-on experimentation with GenAI tools alongside guided critique and structured reflection on pedagogical alignment and ethical considerations. This study did not evaluate the effectiveness of GenAI tools, nor did it compare instructional modalities or measure changes in teaching performance. Instead, it describes the design, implementation, and pilot evaluation of a theory-informed PD approach that supported a small medical faculty cohort (n=10) in critically engaging with GenAI within their pedagogical contexts. By centering disciplinary expertise, reflective judgment, and ethical awareness, PDGenAI-P offers an exploratory contribution to emerging scholarship on faculty development for GenAI integration in ME.

Program Design

PDGenAI-P [5] was designed to address a gap in context-specific PD that recognizes medical educators' disciplinary expertise while supporting pedagogical and ethical engagement with GenAI. The program consisted of 5 live, online 1-hour workshops delivered via Zoom (Zoom Communications, Inc). Two scheduling formats were offered to accommodate faculty schedules: (1) a 5-day format with 1-hour workshops or sessions each day (July 2024), and (2) a condensed 2-day format with multiple workshops or sessions per day (August 2024). Each workshop session was approximately 1 hour in duration in the 5-day format; in the condensed format, multiple 1-hour sessions were delivered across 2 days. Total instructional

XSL•FO
RenderX

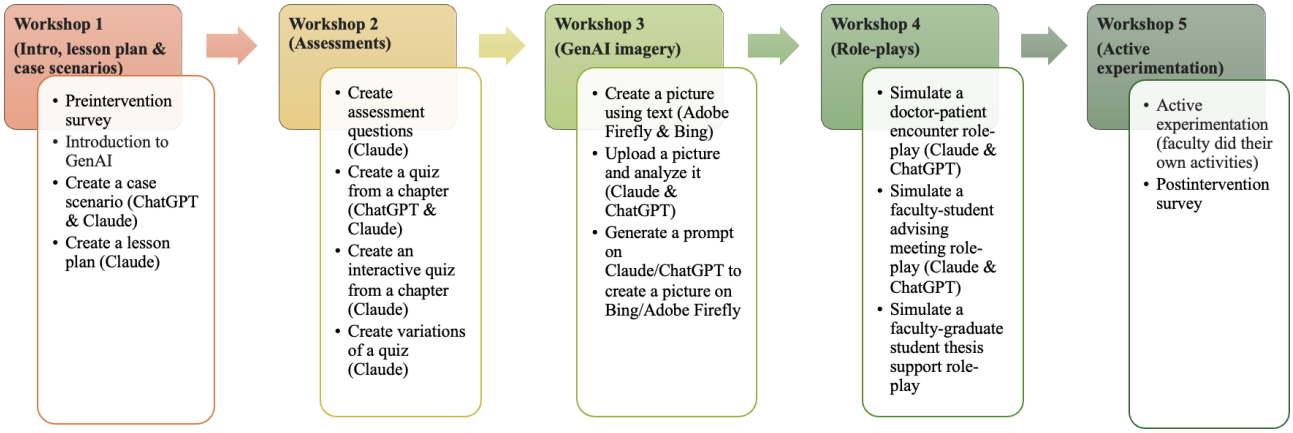
time across the 5 workshops was approximately 5 hours. Each workshop built upon the previous one, moving from foundational exploration of GenAI tools toward context-specific pedagogical application.

The program's learning objectives were to support medical educators to (1) analyze key characteristics, capabilities, and limitations of GenAI within ME contexts; (2) evaluate transparent and responsible GenAI practices that foreground human expertise and educator decision-making; (3) design pedagogically sound learning activities by integrating GenAI with medical educators' disciplinary and instructional expertise; (4) evaluate ethical and pedagogical challenges associated with GenAI use in ME teaching and exercise professional judgment in mitigating them; and (5) value human expertise, confidence, and professional judgment when engaging with GenAI as an assistive, rather than substitutive, pedagogical resource.

These learning objectives guided the overall program design and facilitation but were not explicitly revisited as assessment criteria during workshops. The PDGenAI-P workshop series was developed as part of a doctoral dissertation and represents the second iteration of an evolving faculty development initiative [5]. The initial iteration was delivered in a fully self-paced format. Between October 2023 and April 2024, the program design was iteratively refined through informal general and targeted needs assessment, including review of feedback from the first self-paced iteration, consultation with faculty colleagues, and reflection on emerging pedagogical and ethical concerns related to GenAI use in ME. The 5-workshop structure and pedagogical focus areas were finalized in consultation with various medical, GenAI, and other faculty experts by April 2024 to support institutional review board approval. Following delivery of the first live iteration in July 2024, the program was further reviewed and modestly adapted based on participant feedback prior to the subsequent August iteration. This iterative design process was intended to ensure that the program remained contextually relevant, needs-sensitive, and responsive to medical educators' evolving concerns regarding GenAI integration.

An overview of the PDGenAI-P workshops and their pedagogical content focus is shown in Figure 1. Workshops focused on lesson planning and case scenario design (workshop 1), assessments (workshop 2), GenAI imagery (workshop 3), role-play development (workshop 4), and independent experimentation (workshop 5). Learning activities were fully scaffolded through structured prompts provided by the facilitator, with limited flexibility intentionally built in to allow participants to adapt outputs to their own teaching contexts. Activities followed an experiential learning cycle, beginning with facilitator-led demonstration, followed by individual experimentation, and culminating in structured group reflection and abstract conceptualization.

Figure 1. Overview and content focus of the Professional Development in Generative Artificial Intelligence for Pedagogy workshops.



An overview of each workshop's focus and examples of scaffolded GenAI prompts used during activities is available in Multimedia Appendix 1. During workshops, participants generated tangible pedagogical artifacts, such as draft lesson plans, assessment items, role-play scripts, and image-based materials. While activity worksheets were provided as optional scaffolds to guide prompt use and documentation, artifacts were not collected, graded, or formally analyzed. To mitigate risks of cognitive deskilling associated with automated GenAI use, the program design intentionally emphasized critique and judgment over correctness. At the start of the reflection phase, participants were informed that activities were intended to surface pedagogical reasoning, limitations, and ethical considerations rather than to produce optimal outputs. Reflection was facilitated using a structured yet adaptive approach, with participants discussing their generated artifacts in a group setting and articulating considerations related to accuracy, pedagogical alignment, contextual constraints, and responsible implementation. This design foregrounded educator judgment and critical appraisal, positioning GenAI as an assistive resource rather than a substitute for pedagogical decision-making. Faculty completed an online preintervention survey (PreIS) at workshop 1, postintervention survey (PostIS) at workshop 5, and a follow-up survey (FUS) 2 weeks later to capture self-reported learning and intended implementation of GenAI-supported teaching practices.

The workshops were designed as applied and interactive sessions, with learning occurring primarily through guided exploration, structured activities, and facilitated reflection. During each workshop, participants were asked to complete structured activity worksheets to document GenAI prompts, generated artifacts, and reflections on their learning. No required readings or formal homework assignments were assigned between sessions, although participants were encouraged to

XSL•FO
RenderX

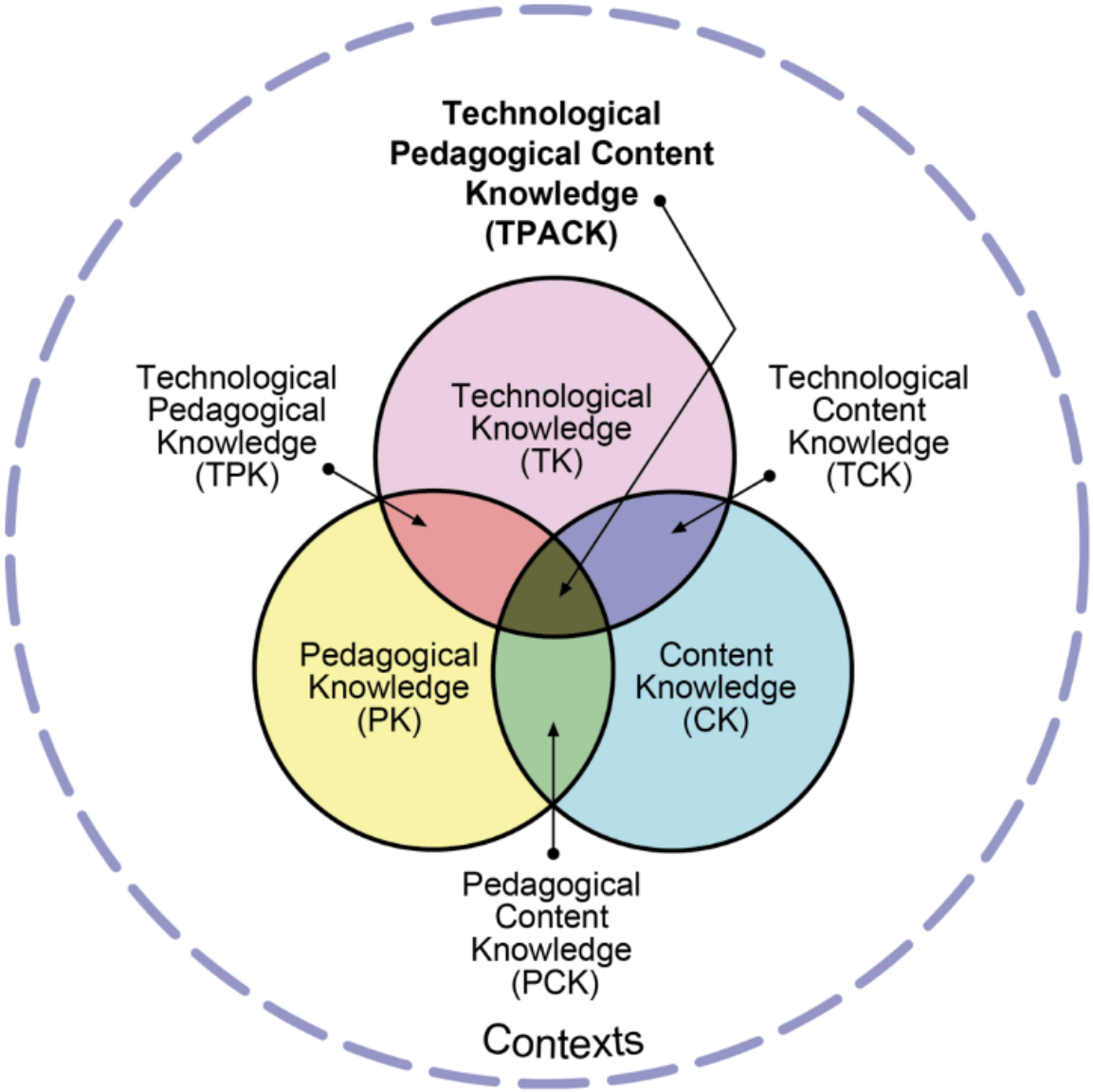
independently experiment with GenAI tools in their teaching contexts.

Theoretical Framework

The PDGenAI-P program was grounded in ELT [25,26] and the TPACK [3,24] framework. ELT emphasizes learning through iterative cycles of experience, reflection, conceptualization, and experimentation, supporting active engagement and reflective sense-making in professional learning contexts. ELT progresses from experience to action with 4 phases ordered as experiencing (Concrete Experience), reflecting (Reflective Observation), thinking (Abstract Conceptualization), and acting (Active Experimentation).

TPACK identifies the intersection of technological, pedagogical, and content knowledge essential for effective technology integration in education [24]. The TPACK framework posits that effective technology integration requires a synergistic understanding of 3 primary forms of knowledge, that is, Content Knowledge, Pedagogical Knowledge, and Technological Knowledge. The TPACK framework is built on 7 domains of educator knowledge (Figure 2) that form the basis of good teaching [24]. These are Technological Knowledge, Content Knowledge, Pedagogical Knowledge, Pedagogical Content Knowledge, Technological Content Knowledge, and Technological Pedagogical Knowledge, with the central idea being TPACK.

Figure 2. The Technological Pedagogical Content Knowledge framework and its knowledge components (reproduced with permission from the publisher, © 2012 by tpack.org).



Although TPACK was originally developed in the context of school-based education [24], subsequent scholarship has extended its application to higher education [27] and health professions education [28], including faculty development and clinical teaching contexts [29-31]. In these settings, TPACK has been used to examine how disciplinary expertise, pedagogical reasoning, and technological fluency intersect in professional education environments where instructional decisions carry high-stakes implications. This broader

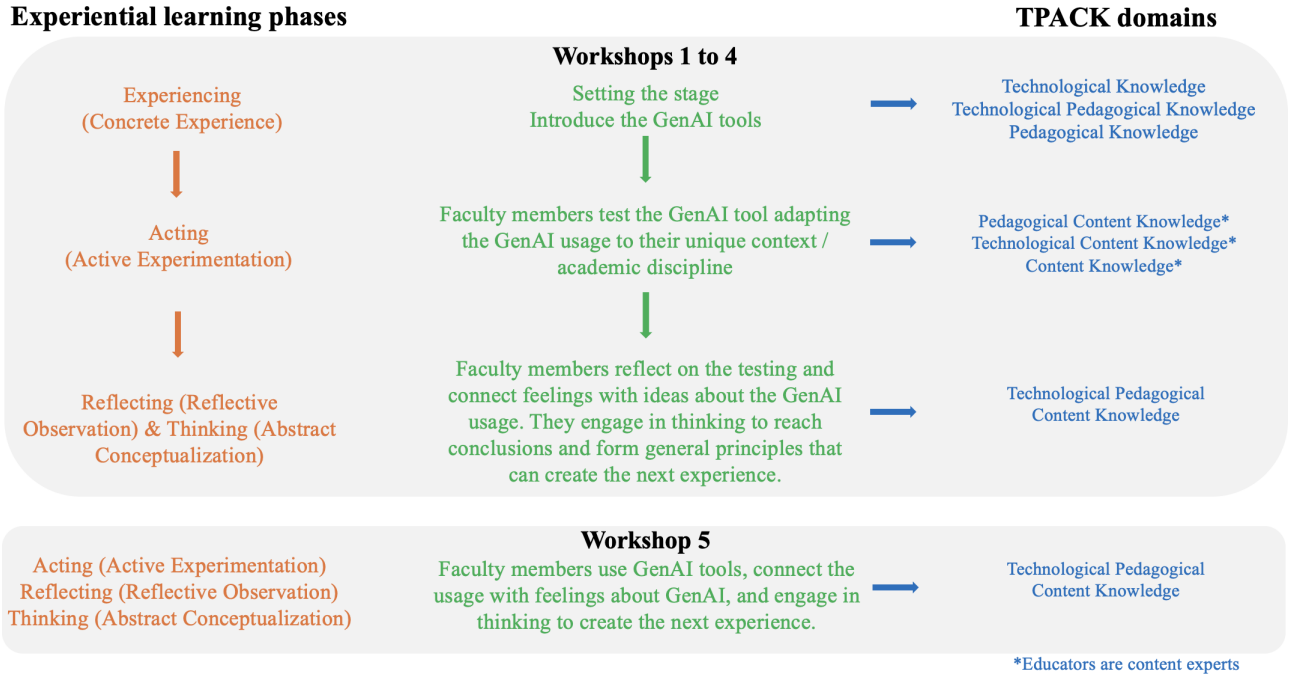
XSL•FO
RenderX

application supports its relevance for medical faculty navigating integration of emerging technologies such as GenAI.

Figure 3 illustrates the structure and describes the mapping of ELT phases and TPACK domains in the PDGenAI-P workshops. Each workshop embedded an experiential learning cycle, beginning with facilitator-led examples (concrete experience), followed by hands-on experimentation with GenAI tools, and structured group reflection and conceptualization focused on pedagogical alignment, ethical considerations, and contextual constraints. Although ELT traditionally progresses from experience to action with 4 phases ordered as experiencing (Concrete Experience), reflecting (Reflective Observation), thinking (Abstract Conceptualization), and acting (Active Experimentation), the cycle was intentionally restructured in the PDGenAI-P so that action followed immediately after experience to support rapid testing and critique of GenAI outputs.

This design created a dynamic interplay between experimentation, reflection, and conceptual understanding, enabling faculty to iteratively examine, contextualize, and refine GenAI-supported teaching practices within their own disciplinary settings [5]. Participants used a range of GenAI tools across workshops, including ChatGPT, Claude, Microsoft Copilot, Adobe Firefly, and Bing (at the time of delivery). Tool selection was participant-dependent and varied based on access and activity type; most participants primarily used ChatGPT or Claude, with other tools used for specific activities such as image-based tasks. During each workshop, participants were provided with structured activity worksheets designed to scaffold GenAI prompt development, artifact generation, and reflective analysis. The worksheets guided participants to document prompts used, generated outputs, contextual adaptations, and reflections on pedagogical alignment, limitations, and ethical considerations. These worksheets functioned as learning scaffolds rather than assessment tools and were optional. While participants generated tangible pedagogical artifacts (eg, draft lesson plans, assessment items, role-play scripts, and image-based materials), these artifacts and worksheets were not collected for grading, formally evaluated, or systematically analyzed as outcome data.

Figure 3. Professional Development in Generative Artificial Intelligence for Pedagogy design describing the mapping of experiential learning phases and Technological Pedagogical Content Knowledge domains across workshops. GenAI: generative artificial intelligence.



Methods

Study Design and Intervention Setting

This study examined the design, implementation, and short-term outcomes of a faculty PD workshop series—PDGenAI-P at Weill Cornell Medicine-Qatar, a US medical school in Doha, Qatar. The institution offers a 6-year Doctor of Medicine program comprising 2 premedical years followed by 4 years of ME. At the time of the study, the institution employed 83 full-time faculty members [32]. The PDGenAI-P was delivered synchronously online and included faculty from multiple premedical and ME disciplines. Two iterations of the program were implemented between July and August 2024. A mixed methods design was used, with quantitative and qualitative data collected concurrently to assess faculty experiences, learning processes, and intended near-term application of GenAI tools. This approach enabled the integration of descriptive quantitative findings with qualitative analysis of workshop transcriptions, facilitator memos, and facilitator theoretical notes to contextualize how faculty engaged with GenAI within their pedagogical settings.

Participants

Participants were recruited from the faculty body at Weill Cornell Medicine-Qatar.. The institution delivers a US medical curriculum and employs a diverse, multinational faculty engaged in premedical and medical education and research. Recruitment

XSL•FO
RenderX

took place within this institutional setting, targeting faculty members involved in teaching across preclinical and clinical phases of the curriculum. Ten faculty members from Weill Cornell Medicine-Qatar participated in PDGenAI-P. As an exploratory pilot faculty development intervention, the study focused on feasibility and short-term self-reported learning and intended application rather than statistical generalizability.

Recruitment was voluntary and open to all teaching faculty at the institution. Faculty across multiple disciplines were invited via targeted email outreach to ensure representation from a range of premedical and ME subspecialities. The invitation email or recruitment letter outlined the purpose of the study, participation requirements, and the voluntary nature of involvement. Participation was not incentivized. Participants therefore represent a self-selected sample of faculty interested in GenAI and PD, rather than a statistically representative sample of all faculty at the institution. Written informed consent was obtained prior to participation in the workshops.

Data Collection

Quantitative and qualitative data were collected concurrently during the PDGenAI-P [33,34]. Quantitative data were collected from an online PreIS and online PostIS with open-ended questions administered at the start and end of the PDGenAI-P, and an online FUS which was sent via email, 2 weeks after the PDGenAI-P was completed. Surveys were administered electronically using Qualtrics survey software (Qualtrics).

The PreIS primarily focused on faculty TPACK for GenAI as part of a dissertation study; only demographic information and baseline self-reported data on GenAI familiarity were used to contextualize participant experience levels for this manuscript. Detailed TPACK analyses are reported elsewhere and are outside the scope of this manuscript. Participants' familiarity with GenAI tools was identified through workshop self-descriptions and facilitator observations. It was then confirmed using a single self-rated familiarity item in the PreIS, in which participants selected 1 of 5 levels ranging from "no knowledge" to "expert level," based on their understanding and use of GenAI tools. Participants were also invited to provide open-text examples of GenAI use.

PostIS and FUS included both quantitative items and open-ended questions inviting reflections on learning experiences, challenges, and recommendations. Likert-scale items in the PostIS assessed perceptions of program quality (content, organization, relevance, and overall quality), and in the FUS, assessed self-reported confidence in applying GenAI concepts to teaching practice and perceived shifts in perspective. Open-ended questions captured reflections on learning experiences, challenges, and intended application of GenAI in teaching. Confidence and perspective change were each measured using single-item 5-point ordinal scales. The study did not include objective competency-based assessments or validated multi-item scales for these constructs. A 2-week follow-up interval was selected to assess early retention, continued experimentation, and intended near-term application of GenAI following program completion. This interval was chosen to balance feasibility with the opportunity for participants to independently engage with GenAI in their teaching context shortly after the intervention. Follow-up results are therefore interpreted as short-term indicators rather than evidence of sustained integration or long-term behavior change. The survey instruments are provided in Multimedia Appendix 2.

Qualitative data sources included transcriptions of workshop discussions. Analytic sensibilities were informed by ongoing reflexive memoing and maintenance of theoretical notes throughout the design, facilitation, and implementation of the PDGenAI-P workshops. These notes were used to document emerging pedagogical considerations, participant responses, and facilitator reflections.

Quantitative Data Analysis

Descriptive statistics were used to summarize quantitative data. Given the exploratory pilot design and small sample size, quantitative findings are reported descriptively and interpreted as indicators of participant perceptions rather than inferential outcomes. In the absence of a comparison group, the study was not intended to isolate the effects of specific program components or establish causal attribution. Findings reflect participant-reported perceptions and analytically derived patterns across data sources rather than demonstrable causal impact. While PreIS and PostIS were administered, these instruments primarily captured self-reported perceptions rather than objective pretest or posttest measurement of pedagogical skill. Quintile rankings reflect ordinal self-assessment and should not be interpreted as validated measures of learning gain.

Qualitative Data Analysis

Qualitative analysis followed a combination of deductive and inductive coding [33,35] drawing on constructs from TPACK and ELT while allowing themes to emerge. Initial deductive coding using the TPACK framework was conducted to sensitize analysis to technological, pedagogical, and disciplinary considerations relevant; however, detailed TPACK analyses are outside the scope of this manuscript. Inductive coding was subsequently used to capture patterns emerging from participant reflections and workshop interactions. While thematic analysis was initially conducted, a narrative analytic approach was ultimately used to more authentically represent the experiential and temporal progression of the workshops and participants' articulated learning and intentions [36]. All qualitative analysis was conducted by the corresponding author using Dedoose (SocioCultural Research Consultants, LLC), a cloud application for managing, analyzing, and presenting qualitative and mixed methods research data [5].

Rigor and trustworthiness were supported through reflexive memoing and theoretical note-taking across design, implementation, and analysis; iterative cycling between data, interpretations, and analytic approaches; triangulation across coding, thematic, and narrative analyses; and member checking of interpretive summaries with the principal investigator and site principal investigator to enhance credibility.

Data Integration

Findings were derived through an integrated analysis in which quantitative and qualitative data were collected concurrently and analyzed separately. Quantitative data were analyzed

XSL•FO
RenderX

descriptively, and qualitative data were analyzed using deductive and inductive approaches, alongside narrative analysis. Integration occurred at the interpretation stage, where quantitative results contextualized participant-reported perceptions, while qualitative data provided depth, context, and explanatory insights into how participants engaged with GenAI across workshops. Both data strands were interpreted together to identify areas of convergence, complementarity, and divergence. These were synthesized to generate insights on faculty experiences, learning processes, and intended application. Facilitator memos and theoretical notes were reviewed to triangulate findings, enrich interpretation, and capture iterative changes in faculty engagement.

Ethical Considerations

The study received approval from the institutional review boards (IRBs) of both Arizona State University (IRB# STUDY00020020) and Weill Cornell Medicine-Qatar (IRB# 24 - 00027) and was conducted as part of the corresponding author's doctoral research. All participants received written information about the study's purpose, procedures, voluntary nature, and data confidentiality. Informed consent was obtained prior to participation. Surveys were anonymous. All participant responses were confidential. Given the study's focus on GenAI tools, additional measures were taken to address ethical considerations specific to GenAI use. Faculty were informed about the limitations of GenAI tools, including potential hallucinations, bias in training data, and risks associated with overreliance on AI-generated content. During workshops, participants were encouraged to critically evaluate GenAI outputs and apply disciplinary expertise to verify accuracy. No student data or patient-related information was used at any point. The study adhered to ethical guidelines for educational research and maintained compliance with institutional data governance policies throughout the project.

Results

Participants Characteristics

Ten faculty members participated in the PDGenAI-P workshop series. Participants represented a range of academic disciplines, including anatomy, biochemistry, biology, biophysics, biostatistics, chemistry, medical humanities, pathology, public health, research methodology, and physics. In terms of professional experience, 5 (50%) participants identified as senior faculty with more than 10 years of professional experience, 2 (20%) were midcareer (5 - 10 years), and 3 (30%) were junior faculty with fewer than 5 years of experience. Six (60%) participants had been employed at the institution for more than 10 years. Participants reported varying levels of prior familiarity with GenAI tools. This distribution reflects participation from faculty at varying career stages and levels of teaching experience. This diversity enhanced disciplinary breadth but limited claims of representativeness or homogeneity.

Workshop-Level Progression

This subsection presents findings derived from narrative analysis of workshop transcripts, facilitator memos, and theoretical notes from the PDGenAI-P. Across the 5 PDGenAI-P workshops, narrative analysis indicated a developmental progression in faculty engagement with GenAI tools, moving from exploratory use toward more intentional pedagogical application and early integration within teaching contexts. In early sessions, faculty experimented with text-based GenAI tools with curiosity but expressed uncertainty regarding reliability and accuracy. Across early workshop discussions and reflections, faculty described GenAI outputs as often generic or incomplete without disciplinary oversight, prompting recalibration of expectations about GenAI's capabilities and limitations.

As workshops progressed, narrative comparison across sessions showed that faculty applied GenAI to specific pedagogical tasks, including lesson planning and case scenario design (ChatGPT/Claude/Copilot), assessment design (ChatGPT/Claude), image generation (ChatGPT/Adobe Firefly/Bing), and simulation-based role-play (ChatGPT/Claude/Copilot). Narrative analysis of participant reflections indicated increasing confidence with prompt refinement and a more nuanced understanding of tool limitations, especially related to hallucinations, contextual misalignment, and training data bias. Across successive workshop transcripts, faculty discourse shifted from tool operation to deeper pedagogical reasoning as they considered when, why, and how GenAI could meaningfully enhance learning. Cross-session analysis showed faculty applying GenAI tools (primarily ChatGPT and Claude) to address discipline-specific instructional challenges. Workshop transcripts showed greater fluency in faculty adapting GenAI outputs, openly discussed ethical considerations, and describing GenAI as a potential "copilot" to augment rather than replace, educator expertise. Narrative analysis of workshop discussions showed ethical discussions centered on tool hallucinations, data bias, plagiarism, and the importance of pedagogical intent. They also reinforced the need for authenticity, bias mitigation, and human oversight. This phase reflected consolidation of learning, identified through synthesis of participant reflections and application-focused discussions, as faculty connected GenAI capabilities with their pedagogical goals. Across all workshops, faculty reflections illustrated an evolving mindset from exploratory use to critical, discipline-specific integration. Facilitator memos corroborated analytically derived patterns, documenting similar shifts in faculty discourse from tool operation to pedagogical alignment and ethical reflection. Findings in this subsection reflect qualitative interpretation of participant discussions and reflections rather than objective measures of behavior change.

Workshop Experiences and Perspectives (PostIS)

Quantitative PostIS results are reported descriptively to summarize participant perceptions of program quality and relevance. All participants (n=10) completed the PostIS. Faculty rated workshop content, organization, relevance, and overall quality of the program as either above average (n=4, 40%) or excellent (n=6, 60%), with no responses below these categories (Table 1). Relevance was rated as above average by 5 (50%) participants or excellent by 5 (50%) participants. Qualitative feedback further supported these ratings. The faculty described workshops as "directly responding to our needs" and praised the logical sequencing and hands-on structure. Faculty reported

XSL•FO
RenderX

that immediate experimentation facilitated meaningful learning, and several highlighted newly discovered applications for GenAI in their teaching and research. One faculty member commented, "This workshop opened my eyes to all the resources in GenAI. I will definitely use GenAI to enhance my teaching and research." Constructive suggestions included allowing more time for reflection and offering temporary access to premium GenAI tools to mitigate limitations of freeware versions.

Table . Participant ratings of the PDGenAI-P[a] workshops (PostIS[b]; N=10)[c].

Workshop domain	Very poor, n (%)	Below average, n (%)	Average, n (%)	Above average, n (%)	Excellent, n (%)
Workshop content	0 (0)	0 (0)	0 (0)	4 (40)	6 (60)
Workshop organization	0 (0)	0 (0)	0 (0)	4 (40)	6 (60)
Workshop relevance	0 (0)	0 (0)	0 (0)	5 (50)	5 (50)
Overall workshop quality	0 (0)	0 (0)	0 (0)	4 (40)	6 (60)

[a]PDGenAI-P: Professional Development in Generative Artificial Intelligence for Pedagogy.

[b]PostIS: postintervention survey.

[c]Ratings were based on a 5-point Likert scale (1=Very poor to 5=Excellent).

Knowledge and Confidence (FUS)

Given the exploratory nature of the FUS and small sample size, results are reported as descriptive indicators of early confidence rather than inferential outcomes. Five (50%) faculty members completed the FUS 2 weeks after PDGenAI-P. This construct assessed self-reported readiness to apply GenAI-related skills in teaching practice. As shown in Table 2, faculty reported varying levels of confidence in applying workshop concepts to their teaching: 2 (40%) were "very confident," 2 (40%) were "moderately confident," and 1 (20%) reported being "slightly confident." No faculty selected "not confident." In open-ended responses, faculty attributed increased confidence to hands-on experimentation, iterative prompt refinement, and opportunities for guided reflection. One faculty member commented, "The workshop introduced us to different and very interesting GenAI tools. I am experimenting with these tools and learning to apply them in my daily teaching," highlighting the importance of continued practice.

Table . Self-reported confidence in applying PDGenAI-P[a] concepts or skills in pedagogical practice (FUS[b]; N=5)[c].

Confidence level	Values, n (%)
Not at all confident (I do not feel prepared to apply the concepts or skills in my teaching practice).	0 (0)
Slightly confident (I feel somewhat prepared but would need more guidance or practice before applying these concepts or skills in my pedagogical practice).	1 (20)
Moderately confident (I feel somewhat prepared and might try to apply some of the concepts or skills in my pedagogical practice).	2 (40)
Confident (I feel prepared and confident in my ability to apply most concepts or skills learned in my pedagogical practice).	0 (0)
Very confident (I feel fully prepared and very confident in my ability to seamlessly integrate all concepts or skills learned in my pedagogical practice).	2 (40)

[a]PDGenAI-P: Professional Development in Generative Artificial Intelligence for Pedagogy.

[b]FUS: follow-up survey.

[c]Confidence was assessed on a 5-point Likert scale ranging from 1=Not at all confident to 5=Very confident.

Shift in Perspectives on GenAI in Teaching and Learning (FUS)

This FUS construct asked participants to self-rate perceived changes in their understanding of GenAI's role in teaching and learning. Among respondents (n=5), 3 (60%) participants reported a moderate perceived change, 1 (20%) participant reported a slight change, and 1 (20%) participant reported a substantial perceived change (Table 3). Given the small number of responses and the self-reported nature of the item, these results reflect subjective perceptions rather than objectively measured change. Qualitative analysis of open-ended follow-up responses provided additional context for these participants' self-reported perceived changes shifts. Faculty who indicated moderate perceived change described gaining new insights into the pedagogical applications of GenAI beyond text generation. One respondent who indicated a substantial perceived change described beginning to use GenAI in planning small-group activities and problem-based learning sessions. The faculty who reported complete transformation noted adopting GenAI to redesign small-group activities, laboratories, and problem-based learning sessions. The faculty member observed, "After this

XSL•FO
RenderX

workshop, I'm using GenAI to plan PBL sessions and syllabus design. It has changed how I think about lesson preparation." These comments reflect participant-reported experiences rather than independently verified changes in instructional practice.

Table . Self-reported perceived change in faculty perspectives on teaching and learning with GenAI[a] tools (FUS[b]; N=5)[c].

Change level	Values, n (%)
No change at all (My perspectives have not changed after attending the workshops. I maintain the same views as I did before).	0 (0)
Slight change (There has been a minor shift in my perspectives. I see a few things differently but largely hold the same views).	1 (20)
Moderate change (I have gained new insights that have changed my views to some extent).	3 (60)
Considerable change (I have adopted many new views and approaches).	0 (0)
Complete transformation (The workshops have changed my approach and understanding of GenAI[d] in teaching and learning).	1 (20)

[a]GenAI: generative artificial intelligence.

[b]FUS: follow-up survey.

[c] Perceived change was assessed on a 5-point scale ranging from 1=No change to 5=Complete transformation.

[d]GenAI: generative artificial intelligence.

Intended Implementation Strategies and Good Practices (FUS)

This FUS construct item asked faculty to indicate how they intended to use GenAI in teaching and administrative workflows. Among respondents (n=5), all 5 (100%) participants reported intentions to use GenAI to support lesson planning, instructional design, and assessment development, including multiple-choice questions and United States Medical Licensing Examination–style questions (Table 4). Four participants (80%) indicated plans to use GenAI to develop case scenarios for project-based learning and to visualize complex learning material. Fewer respondents reported intentions to define responsible use practices (n=2, 40%), initiate ethical discussions (n=1, 20%), or adapt materials for personalized learning (n=1, 20%). These results reflect short-term self-reported intentions and should be interpreted as exploratory indicators of anticipated application rather than evidence of implementation.

Table . Faculty-intended utilization of GenAI[a] in teaching and learning activities (FUS[b]; N=5).

Implementation strategies	Values, n (%)
Create case scenarios to facilitate project-based learning.	4 (80)
Create a lesson plan, interactive activities and a rubric using a backward design model.	5 (100)
Visualize learning material.	4 (80)
Create various types of assessments (MCQs[c] USMLE[d] style, formative etc) based on the lesson plan and rubric.	5 (100)
Adapt learning material to create personalized learning experiences.	1 (20)
Critically analyze content GenAI.	1 (20)
Define responsible use of GenAI in your settings.	2 (40)
Initiate discussions of responsible and ethical use of GenAI in your classroom.	1 (20)
Manage administrative tasks.	1 (20)

[a]GenAI: generative artificial intelligence.

[b]FUS: follow-up survey.

[c]MCQs: multiple-choice questions.

[d]USMLE: United States Medical Licensing Examination.

Anticipated Concerns and Support Needs (FUS)

The FUS also asked participants to describe anticipated challenges and additional support required for GenAI implementation (Table 5). Among respondents (n=5), 3 (60%) described minor anticipated obstacles including limited institutional guidance, access constraints, or the need for continued practice. One (20%) reported no anticipated obstacles, and 1 (20%) anticipated more substantial challenges requiring additional training or structural support. These results reflect

XSL•FO
RenderX

perceived expectations rather than verified implementation barriers. They provide preliminary insight into areas where ongoing faculty development or institutional support structures may warrant consideration.

Table . Faculty levels of concern regarding application of GenAI[a] tools (FUS[b]; N=5)[c].

Level of concern	Values, n (%)
Very concerned (I am highly concerned about several major obstacles and feel I need additional support to apply my learning).	0 (0)
Fairly concerned (I foresee several obstacles and will need to seek solutions to overcome them).	1 (20)
Moderately concerned (I am worried about potential obstacles and am considering strategies to address them).	0 (0)
Mildly concerned (I have slight concerns that there might be minor obstacles, but I feel well equipped to manage them).	3 (60)
Not concerned at all (I am completely confident in my ability to apply the learning and resolve any obstacles that I might encounter).	1 (20)

[a]GenAI: generative artificial intelligence.

[b]FUS: follow-up survey.

[c]Level of concern was assessed on a 5-point scale ranging from 1 = Very concerned to 5 = Not concerned at all.

Faculty also identified areas of support that they perceived would enhance their ability to apply PDGenAI-P strategies in practice (Table 6). Among respondents (n=5), 4 (80%) indicated interest in consulting with GenAI experts and participating in peer discussion forums, and 3 (60%) expressed interest in additional workshops to deepen understanding. Participants also requested curated guides (n=3, 60%), institutional navigation support (n=3, 60%), and step-by-step instructions for integrating GenAI into teaching (n=2, 40%). These results should be interpreted as exploratory indicators of anticipated support requirements rather than definitive implementation gaps.

Table . Faculty-perceived support needs (FUS[a]; N=5).

Description of support	Values, n (%)
Step-by-step instructions for integrating PDGenAI-P[b] strategies into teaching practices.	2 (40)
Opportunities to discuss specific challenges and questions with GenAI[c] experts.	4 (80)
Forums for exchanging ideas and experiences with other educators applying GenAI.	4 (80)
Additional workshops or training sessions to deepen understanding and address challenges.	3 (60)
A customized collection of online guides and resources for further exploration.	3 (60)
Guidance on navigating institutional policies and obtaining support.	3 (60)

[a]FUS: follow-up survey.

[b]PDGenAI-P: Professional Development in Generative Artificial Intelligence for Pedagogy.

[c]GenAI: generative artificial intelligence.

Summary of Findings

The integrated quantitative and qualitative findings reflect high participant-reported satisfaction, perceived increases in confidence, and articulated intentions to integrate GenAI into teaching contexts. These results derive from a small pilot cohort and rely on self-reported perceptions rather than objective measures of instructional change. As such, they should be interpreted as exploratory indicators of early engagement. The findings suggest that a theory-informed, experiential workshop series may serve as a feasible entry point for supporting faculty reflection and initial GenAI experimentation in ME, while underscoring the need for longitudinal evaluation and institutional reinforcement.

Discussion

Principal Findings

This pilot mixed methods study explored the design and short-term participant-reported outcomes of PDGenAI-P, a faculty development workshop series focused on integrating GenAI into ME teaching. Findings suggest that a theory-informed, experiential workshop format may support medical educators in developing the initial knowledge, perceived

XSL•FO
RenderX

confidence, and reflection needed to integrate GenAI tools responsibly into their teaching. Qualitative analysis suggested that structured, hands-on engagement combined with facilitated reflection may support medical educators in moving from exploratory tool use toward more intentional pedagogical application. Participants reported increased self-reported confidence and articulated intended strategies for using GenAI in lesson planning, assessment design, visualization, and case-based instruction. These findings reflect perceived learning and engagement rather than independently verified changes in teaching practice. They offer preliminary insight into how short, structured faculty development may initiate reflective engagement with GenAI in ME contexts.

In relation to scholarship on emerging educational technologies in education and health professions training, several patterns warrant consideration. The observed engagement aligns with literature suggesting that faculty benefit from experiential, context-sensitive approaches when adopting rapidly evolving educational technologies [37-39]. Structured opportunities for experimentation, guided critique, and peer dialogue appear important in supporting thoughtful integration of digital tools [40,41]. At the same time, the imbalance observed between strong intentions for content creation and comparatively lower intention for explicit ethical operationalization mirrors concerns in the broader GenAI in-education literature that early adoption may prioritize efficiency gains over critical or reflexive practice [42,43]. These patterns underscore the importance of longitudinal scaffolding beyond initial exposure.

Participants' reported satisfaction and articulated applications suggest that short, theory-informed workshop series may create space for educators to reflect on how GenAI intersects with disciplinary expertise and instructional decision-making. Rather than positioning GenAI as a substitute for educator judgment, the workshop design emphasized structured experimentation and facilitated reflection in supporting intentional, context-sensitive integration. Participants frequently attributed their learning to the experiential, hands-on structure of the workshop series, particularly the opportunity to test GenAI outputs within their own disciplinary contexts. Their reflections suggest that iterative experimentation combined with facilitated dialogue may help educators critically appraise both the affordances and limitations of GenAI tools.

The workshop series was informed by ELT and the TPACK framework, and the observed patterns of engagement were consistent with the principles underlying these models [3,24-26]. Structuring activities around cycles of experimentation and reflection, while explicitly connecting technological tools to pedagogical and disciplinary considerations, appeared to create opportunities for faculty to examine how GenAI might intersect with their existing expertise. Rather than emphasizing tool proficiency alone, the approach situated GenAI within broader pedagogical reasoning, reinforcing the centrality of professional judgment in ME.

Facilitator reflections across the 2 PDGenAI-P iterations suggest that adaptive facilitation played an important role in shaping participant engagement. Modifications such as shortening introductory content, increasing time for hands-on exploration, and shifting from structured guidance to flexible experimentation were implemented directly in response to faculty feedback. These adjustments were associated with more active participation and extended reflective dialogue during subsequent sessions. Given that participants' diverse disciplinary priorities and limited time, responsiveness to context appeared important. However, the extent to which facilitation strategies alone can influence sustained pedagogical change remains uncertain, reinforcing the need for continued institutional and longitudinal support.

The PostIS and FUS results reflect short-term, self-reported perceptions rather than objective evidence of sustained pedagogical change. Faculty described increased confidence in applying GenAI tools and perceived shifts in how they conceptualized and planned instructional activities. However, these responses were collected 2 weeks postintervention and were based on a limited subset of participants (n=5) warranting cautious interpretation. Reported intended uses of GenAI focused on assessment design, lesson planning, visualization of learning materials, and the creation of case scenarios, suggesting primarily applied and discipline-specific integration. Ethical considerations, such as bias and hallucinations, were acknowledged in reflections; suggesting awareness of responsible GenAI use, although intentions to operationalize these practices were less consistently articulated. This highlighted a potential gap between awareness and explicit implementation planning. These patterns should be interpreted as early indicators of engagement rather than a confirmation of durable transformation in teaching practice. Intention to apply GenAI and confidence in doing so do not necessarily translate into consistent implementation, nor do they guarantee long-term integration. Longitudinal evaluation and behavioral measures would be necessary to determine whether reported shifts are sustained over time and embedded into routine pedagogical practice.

An imbalance was observed in the FUS responses, with faculty reporting strong intentions to use GenAI for content creation and assessment design but lower intention to engage in critical evaluation of GenAI outputs, initiate ethical discussions, or define responsible use. Although based on a small sample (n=5), this pattern aligns with broader educational technology literature suggesting that early adoption is often characterized by instrumental, efficiency-oriented uses before more critical, reflective, and ethically grounded practices are operationalized [23,43,44]. In ME, where ethical reasoning and professional judgment are central, this finding underscores the importance of continued scaffolding and longitudinal faculty development to support translation of ethical awareness into explicit pedagogical practice, rather than relying on initial exposure alone [6,45,46]. Awareness of risks such as bias and hallucinations does not necessarily translate into structured pedagogical practices that foreground responsible use [7,43], suggesting that faculty development initiatives should move beyond exposure and skill-building toward explicit modeling and reinforcement of critical appraisal and responsible use.

The integration of GenAI in ME raises important risks, including overreliance on artificial intelligence–generated outputs, automation bias, and potential cognitive deskilling, if learners

XSL•FO
RenderX

or educators defer uncritically to generated content. Automation bias, well described in clinical decision support contexts, can lead individuals to privilege automated recommendations even when flawed, contributing to errors of commission or omission and reducing vigilance in information seeking [47]. Parallel concerns have emerged in ME scholarship where uncritical or habitual reliance on emerging technology such as GenAI may promote cognitive offloading and undermine cognitive autonomy, particularly when use is not accompanied by explicit pedagogical scaffolding and reflective appraisal [11,48]. Within the workshop series, risks such as hallucinations, bias, and overreliance were explicitly discussed and participants were encouraged to verify outputs against disciplinary standards and position GenAI as a "copilot" rather than a substitute for professional judgment. Nevertheless, short-term faculty development cannot eliminate structural risks associated with automation bias or cognitive offloading. In ME, where diagnostic reasoning, ethical judgment, and patient safety are foundational [49,50], sustained institutional alignment and longitudinal reinforcement are necessary to ensure that GenAI augments rather than diminishes professional competence [7,43].

In summary, these findings suggest that faculty development for GenAI integration should move beyond technical proficiency and efficiency oriented use toward structured opportunities for experimentation, critical and ethical evaluation, and discipline-specific pedagogical alignment. Early indicators of satisfaction and self-reported confidence should not be interpreted as comprehensive integration but rather as initial stages in a longer developmental process. Particularly in ME, where ethical judgment and professional accountability are central, short-term faculty development may initiate experimentation and reflection but cannot complete the work of embedding GenAI into teaching in critically reflective and ethically explicit ways.

Implications for Faculty Development

Several implications emerge for the design and scaling of GenAI-focused faculty development in ME. Iterative hands-on experimentation supported by structured reflection appears important for enabling faculty to evaluate GenAI tools within their disciplinary contexts. Adaptive facilitation grounded in learning theory may enhance relevance and responsiveness to participant needs [26]. Longitudinal structures progress from foundational prompting strategies to more advanced applications, such as simulation design, or personalized learning development may further support pedagogical fluency [16,17]. Institutional infrastructure, including curated repositories of vetted prompts, exemplar lesson plans, and structured use cases, may reduce implementation barriers and promote consistent practice [16,17,43]. Equally important are communities of practice and peer-mentoring structures that foster sustained dialogue, shared problem-solving, and ethical reflection as GenAI tools evolve [51]. Attention to responsible use, including bias detection, output verification, and transparency, remains essential to align GenAI integration with professional standards in ME.

Limitations

This study has several limitations. First, reported outcomes are based on self-reported confidence, perceived shifts in thinking, and intended implementation strategies rather than direct observation of enacted teaching practice. Self-reported measures are also susceptible to social desirability bias and overestimation of competence, particularly immediately following a PD intervention. The study did not include classroom observations, external evaluation of instructional artifacts, structured rubrics, performance-based assessments, behavioral indicators of pedagogical change, or longitudinal follow-up beyond 2 weeks. Ethical and critical dimensions of GenAI use were explicitly discussed throughout the workshops but were not assessed as measurable competencies. Consequently, expressed awareness of ethical concerns should not be interpreted as demonstrated competence in ethical enactment.

Second, the study design limits causal inference. The intervention was not evaluated against alternative faculty development approaches, and no comparison or control group was included. Because component-level analyses were not conducted, it is not possible to determine which elements of the workshop series may have contributed to reported changes, nor to attribute observed patterns solely to the intervention itself. Although a PreIS was administered, the study did not use objective pretest or posttest assessments of specific GenAI pedagogical competencies, nor did it include a comparison group. Outcomes therefore reflect descriptive, self-reported perceptions rather than verified changes in teaching practice. Instructional artifacts generated during workshops were not formally evaluated, and teaching practices were not observed. The quantitative tables represent descriptive summaries of participant perceptions and intended use. The sample included faculty from multiple disciplines and career stages, which enhanced diversity of perspectives but limited homogeneity and statistical generalizability.

Third, the follow-up period was limited to 2 weeks, and only half of the participants (n=5) completed the FUS. This reduces the breadth of longitudinal perspectives available and restricts insight into sustained integration and introduces potential response bias. Confidence and intention to use GenAI do not necessarily translate into consistent implementation, or guarantee long-term pedagogical transformation. Finally, the small, self-selected sample (n=10) and single-institution context limit transferability to other ME settings. Although baseline GenAI familiarity was captured, confidence in GenAI use was not independently measured prior to the intervention. Quantitative findings are based on descriptive measures only, and no inferential statistical analyses were conducted given the pilot design.

Future Work

This pilot study was designed to examine feasibility and early indicators of engagement rather than to establish effectiveness or long-term impact. Future research should incorporate longitudinal designs; objective, performance-based measures; multisite implementation; and structured evaluation of ethical competence to determine whether reported shifts translate into sustained and observable transformation in ME practice. Despite

XSL•FO
RenderX

these limitations, the iterative design and adaptive facilitation of PDGenAI-P provided preliminary insight into how an experiential, theory-informed GenAI faculty development workshop series may support early engagement and critical reflection. Future research with larger samples, multisite collaborations, and longitudinal evaluation will be important to assess broader applicability and sustained impact on instructional practice. Future work should examine how PDGenAI-P might evolve beyond a pilot workshop series into a scalable and institutionally integrated faculty development initiative. Offering the program as an accredited Continuing Medical Education or Continuing Professional Development activity may acknowledge the depth of pedagogical and ethical engagement required for responsible GenAI use in ME. A longitudinal structure that progresses from foundational prompting strategies to advanced applications such as GenAI-supported simulation design may further develop faculty fluency and pedagogical judgment. Developing a structured implementation guide could facilitate cross-institutional adoption while supporting contextual adaptation rather than replication.

Sustainable integration will likely depend on institutional infrastructure. Centralized repositories of vetted prompts, exemplar lesson plans, and instructional resources, alongside integration within existing learning management systems, may reduce logistical barriers. Equally important are social learning structures, such as interdisciplinary communities of practice and peer-mentoring structures, which may support ongoing critical dialogue, shared problem-solving, and ethical reflection, reducing the sense of isolation that often accompanies rapid technological change. Further empirical evaluation will be necessary to determine how such structures influence sustained and responsible GenAI integration in ME contexts.

Conclusions

This exploratory pilot study provides preliminary insight into the design and short-term reception of a theory-informed faculty development workshop series PDGenAI-P, focused on GenAI integration in ME. Findings reflect participant-reported engagement, perceived confidence, and early intentions rather than verified changes in teaching practice. While the workshop series appeared to initiate experimentation and reflection, sustained pedagogical and ethical integration will likely require institutional alignment, longitudinal reinforcement, and explicit modeling of responsible GenAI facilitation. Future research should incorporate objective measures of instructional change, including evaluation of implemented teaching artifacts, and long-term follow-up to determine whether early shifts translate into durable transformation in ME practice.

Acknowledgments

The authors thank Dr James Dunnigan (Arizona State University) for contributions to program design, Dr Melissa Warr (New Mexico State University) and Dr Punya Mishra (Arizona State University) for input on select workshop activities, and Dr Danah Henriksen (Arizona State University) for guidance on the application of Experiential Learning Theory. Their input meaningfully informed the development of this faculty development initiative. The authors declare the use of generative artificial intelligence (GenAI) in the research and writing process. According to the GAIDeT taxonomy (2025), the following tasks were delegated to GenAI tools under full human supervision: visualization, proofreading and editing, and summarizing text. The GAI tool used was ChatGPT 5 versions. Responsibility for the final manuscript lies entirely with the authors. GAI tools are not listed as authors and do not bear responsibility for the final outcomes.

Funding

The authors declared no financial support was received for this work. The publication of this paper was funded by the Weill Cornell Medicine-Qatar Health Sciences Library.

Data Availability

The datasets generated and analyzed during this study are not publicly available due to the conditions of institutional review board approval and participant consent, which restrict data sharing.

Authors' Contributions

RA conceptualized and designed the study, conducted data collection and analysis, and led manuscript writing. NB contributed to study design, supervised and contributed to qualitative analysis, and critically revised the manuscript. MFM contributed to study design, validated coding, and critically revised the manuscript. All authors reviewed and approved the final manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Overview of Professional Development in Generative Artificial Intelligence for Pedagogy workshops and scaffolded generative artificial intelligence prompts.
[DOCX File, 32 KB - mededu_v12i1e89815_app1.docx]

XSL•FO
RenderX

Multimedia Appendix 2
Preintervention, postintervention, and follow-up survey instruments.
[PDF File, 658 KB - mededu_v12i1e89815_app2.pdf]

References

1. Atlas S. ChatGPT for higher education and professional development: a guide to conversational AI. College of Business Faculty Publications. 2023 Jan 1. URL: https://digitalcommons.uri.edu/cba_facpubs/548 [accessed 2026-04-22]
2. Eysenbach G. The role of ChatGPT, generative language models, and artificial intelligence in medical education: a conversation with ChatGPT and a call for papers. JMIR Med Educ 2023 Mar 6;9:e46885. [doi: 10.2196/46885] [Medline: 36863937]
3. Mishra P, Warr M, Islam R. TPACK in the age of ChatGPT and generative AI. J Digit Learn Teach Educ 2023 Oct 2;39(4):235-251. [doi: 10.1080/21532974.2023.2247480]
4. Miao F, Holmes W. Guidance for Generative AI in Education and Research: UNESCO Digital Library; 2023:92 URL: https://unesdoc.unesco.org/ark:/48223/pf0000386693 [accessed 2024-02-08]
5. Anand R. Integrating Generative Artificial Intelligence in Medical Education Pedagogy with TPACK and Experiential Learning [Doctoral Dissertation]. : Arizona State University; 2025 URL: https://www.proquest.com/dissertations-theses/integrating-generative-artificial-intelligence/docview/3194408959 [accessed 2026-04-22]
6. Nguyen KV. The use of generative AI tools in higher education: ethical and pedagogical principles. J Acad Ethics 2025 Sep;23(3):1435-1455. [doi: 10.1007/s10805-025-09607-1]
7. Belkina M, Daniel S, Nikolic S, et al. Implementing generative AI (GenAI) in higher education: a systematic review of case studies. Comput Educ Artif Intell 2025 Jun;8:100407. [doi: 10.1016/j.caeai.2025.100407]
8. Mehta N, Mehta S, Rubenstein A, Wood SK. Not replaced, but reinvented: AI education pathways to prepare future physicians to lead healthcare transformation. Perspect Med Educ 2025;14(1):849-859. [doi: 10.5334/pme.2233] [Medline: 41278578]
9. Preiksaitis C, Rose C. Opportunities, challenges, and future directions of generative artificial intelligence in medical education: scoping review. JMIR Med Educ 2023 Oct 20;9:e48785. [doi: 10.2196/48785] [Medline: 37862079]
10. Monteith S, Glenn T, Geddes JR, et al. Artificial intelligence and deskilling in medicine. Br J Psychiatry 2026 Jan 8:1-3. [doi: 10.1192/bjp.2025.10496]
11. Izquierdo-Condoy JS, Arias-Intriago M, Tello-De-la-Torre A, Busch F, Ortiz-Prado E. Generative artificial intelligence in medical education: enhancing critical thinking or undermining cognitive autonomy? J Med Internet Res 2025 Nov 3;27:e76340. [doi: 10.2196/76340] [Medline: 41183320]
12. Pham TD, Karunaratne N, Exintaris B, et al. The impact of generative AI on health professional education: a systematic review in the context of student learning. Med Educ (Chicago Ill) 2025 Dec;59(12):1280-1289. [doi: 10.1111/medu.15746]
13. Kang J, Ahn J. Technologies, opportunities, challenges, and future directions for integrating generative artificial intelligence into medical education: a narrative review. Ewha Med J 2025 Oct;48(4):e53. [doi: 10.12771/emj.2025.00787]
14. Naqvi WM, Ganjoo R, Rowe M, Pashine AA, Mishra GV. Critical thinking in the age of generative AI: implications for health sciences education. Front Artif Intell 2025 May 21;8. [doi: 10.3389/frai.2025.1571527]
15. Dong C, Aw DCW, Lee DWC, Low SC, Yan CC. Generative AI in healthcare: insights from health professions educators and students. IME 2025;4(2):11. [doi: 10.3390/ime4020011]
16. Roveta A, Castello LM, Massarino C, Francese A, Ugo F, Maconi A. Artificial intelligence in medical education: a narrative review on implementation, evaluation, and methodological challenges. AI 2025 Sep 11;6(9):227. [doi: 10.3390/ai6090227]
17. Rincón EHH, Jimenez D, Aguilar LAC, Flórez JMP, Tapia Á, Peñuela CLJ. Mapping the use of artificial intelligence in medical education: a scoping review. BMC Med Educ 2025 Apr 12;25(1):526. [doi: 10.1186/s12909-025-07089-8] [Medline: 40221725]
18. Reddy S. Generative AI in healthcare: an implementation science informed translational path on application, integration and governance. Implementation Sci 2024 Mar 15;19(1):27. [doi: 10.1186/s13012-024-01357-9]
19. Hou SI. Enhancing interactive teaching for the next generation of nurses: generative-AI-assisted design of a full-day professional development workshop. Informatics (MDPI) 2026 Jan 15;13(1):11. [doi: 10.3390/informatics13010011]
20. Khamis N, Chen B, Egan C, Gaglani S, Tackett S. More intelligent faculty development: integrating GenAI in curriculum development programs. Med Teach 2025 Nov;47(11):1739-1741. [doi: 10.1080/0142159X.2025.2473606] [Medline: 40029714]
21. Chadha N, Popil E, Gregory J, Armstrong-Davies L, Justin G. How do we teach generative artificial intelligence to medical educators? Pilot of a faculty development workshop using ChatGPT. Med Teach 2025 Jan;47(1):160-162. [doi: 10.1080/0142159X.2024.2341806] [Medline: 38648540]
22. Ichikawa T, Olsen E, Vinod A, et al. Generative artificial intelligence in medical education-policies and training at US osteopathic medical schools: descriptive cross-sectional survey. JMIR Med Educ 2025 Feb 11;11:e58766. [doi: 10.2196/58766] [Medline: 39934984]
23. Kasneci E, Sessler K, Küchemann S, et al. ChatGPT for good? On opportunities and challenges of large language models for education. Learn Individ Differ 2023 Apr;103:102274. [doi: 10.1016/j.lindif.2023.102274]

XSL•FO
RenderX

24. Mishra P, Koehler MJ. Technological pedagogical content knowledge: a framework for teacher knowledge. Teach Coll Rec (1970) 2006 Jun;108(6):1017-1054. [doi: 10.1177/016146810610800610]
25. Kolb D. Experiential Learning: Experience as the Source of Learning and Development: FT Press; 1984. URL: https://books.google.com.qa/books?hl=en&lr=&id=jpbeBQAAQBAJ&oi=fnd&pg=PR7&dq=kolb%27s ISBN:0-13-389250-6 [accessed 2026-04-22]
26. Kolb AY, Kolb DA. Experiential Learning Theory as a guide for experiential educators in higher education. ELTHE 2017 Jun 1;1(1):38. [doi: 10.46787/elthe.v1i1.3362]
27. Maor D. Using TPACK to develop digital pedagogues: a higher education experience. J Comput Educ 2017 Mar;4(1):71-86. [doi: 10.1007/s40692-016-0055-4]
28. Ait Ali D, El Meniari A, El Filali S, Morabite O, Senhaji F, Khabbache H. Empirical research on Technological Pedagogical Content Knowledge (TPACK) framework in health professions education: a literature review. Med Sci Educ 2023 Jun;33(3):791-803. [doi: 10.1007/s40670-023-01786-z] [Medline: 37501808]
29. Karadag E. Turkish medical educators' TPACK components and characters: an analysis within the framework of simulation-based medical education. BMC Med Educ 2019 Jun 25;19(1):229. [doi: 10.1186/s12909-019-1664-1] [Medline: 31238933]
30. Celik I. Towards Intelligent-TPACK: an empirical study on teachers' professional knowledge to ethically integrate artificial intelligence (AI)-based tools into education. Comput Human Behav 2023 Jan;138:107468. [doi: 10.1016/j.chb.2022.107468]
31. Dewi NR, Rusilowati A, Saptono S, et al. Technological, pedagogical, content Knowledge (TPACK) research trends: A systematic literature review publish on 2010 to 2020. J Turkish Sci Educ 2021;18(4):589-604. [doi: 10.36681/tused.2021.92]
32. Fact sheet. Weill Cornell Medicine-Qatar. 2023. URL: https://qatar-weill.cornell.edu/media-and-news/fact-sheet [accessed 2024-01-31]
33. Ivankova NV. Mixed Methods Applications in Action Research: From Methods to Community Action: SAGE Publications; 2014.
34. Clark VLP, Creswell JW. Understanding Research: A Consumer's Guide: Pearson Higher Ed; 2014.
35. Saldana J. The Coding Manual for Qualitative Researchers: Sage Publications Ltd; 2021. URL: https://sk.sagepub.com/book/mono/the-coding-manual-for-qualitative-researchers-5e/toc [accessed 2024-4-30] [doi: 10.4135/9781036235611]
36. Chase SE. Narrative inquiry: still a field in the making. In: Denzin NK, Lincoln YS, editors. Collecting and Interpreting Qualitative Materials, 4th Ed: Sage Publications; 2013:55-83 URL: https://uk.sagepub.com/en-gb/eur/collecting-and-interpreting-qualitative-materials/book237870 [accessed 2024-11-24]
37. Huynh L, Le T, Dang B, An BT, Vu CT, Nguyen A. Mindset matters: fostering teachers' responsible AI use through professional development. J New Approaches Educ Res 2025 Oct 20;14(1):23. [doi: 10.1007/s44322-025-00043-y]
38. Ali M, Ma M, Muneeb M, Wong GKW. Mapping contemporary AI-education intersections and developing an integrated convergence framework: a bibliometric-driven and inductive content analysis. Metrics 2025 Nov 3;2(4):23. [doi: 10.3390/metrics2040023]
39. Du J(, Kohnke L. Repositioning teachers in the age of AI: a cross-disciplinary study of perceptions and practices in China. Teach Educ 2025 Nov 21;0:1-25. [doi: 10.1080/10476210.2025.2590694]
40. Rider EA, Chou C, Abraham C, et al. Longitudinal faculty development to improve interprofessional collaboration and practice: a multisite qualitative study at five US academic health centres. BMJ Open 2023 Apr 18;13(4):e069466. [doi: 10.1136/bmjopen-2022-069466] [Medline: 37076167]
41. Sullivan R(, Neu V, Yang F. Faculty development to promote effective instructional technology integration: a qualitative examination of reflections in an online community. OLJ 2018;22(4). [doi: 10.24059/olj.v22i4.1373]
42. Jin Y, Yan L, Echeverria V, Gašević D, Martinez-Maldonado R. Generative AI in higher education: a global perspective of institutional adoption policies and guidelines. Comput Educ Artif Intell 2025 Jun;8:100348. [doi: 10.1016/j.caeai.2024.100348]
43. Batista J, Mesquita A, Carnaz G. Generative AI and higher education: trends, challenges, and future directions from a systematic literature review. Information 2024 Oct 28;15(11):676. [doi: 10.3390/info15110676]
44. Zawacki-Richter O, Marín VI, Bond M, Gouverneur F. Systematic review of research on artificial intelligence applications in higher education – where are the educators? Int J Educ Technol High Educ 2019 Dec;16(1):39. [doi: 10.1186/s41239-019-0171-0]
45. Al-kfairy M, Mustafa D, Kshetri N, Insiew M, Alfandi O. Ethical challenges and solutions of generative AI: an interdisciplinary perspective. Informatics (MDPI) 2024 Sep;11(3):58. [doi: 10.3390/informatics11030058]
46. Rubeis G. Ethics of Medical AI: Springer International Publishing; 2024. [doi: 10.1007/978-3-031-55744-6]
47. Goddard K, Roudsari A, Wyatt JC. Automation bias: a systematic review of frequency, effect mediators, and mitigators. J Am Med Inform Assoc 2012;19(1):121-127. [doi: 10.1136/amiajnl-2011-000089] [Medline: 21685142]
48. Mancinetti M, Guttormsen S, Berendonk C. Cognitive load in internal medicine: what every clinical teacher should know about cognitive load theory. Eur J Intern Med 2019 Feb;60:4-8. [doi: 10.1016/j.ejim.2018.08.013] [Medline: 30181017]
49. Croskerry P. A universal model of diagnostic reasoning. Acad Med 2009 Aug;84(8):1022-1028. [doi: 10.1097/ACM.0b013e3181ace703] [Medline: 19638766]

XSL•FO
RenderX

50. Norman G. Research in clinical reasoning: past history and current trends. Med Educ 2005 Apr;39(4):418-427. [doi: 10.1111/j.1365-2929.2005.02127.x] [Medline: 15813765]
51. Wenger E. Communities of Practice: Learning, Meaning, and Identity: Cambridge University Press; 1999.

Abbreviations

ELT: Experiential Learning Theory
FUS: follow-up survey
GenAI: generative artificial intelligence
IRB: institutional review board
ME: medical education
PD: professional development
PDGenAI-P: Professional Development in Generative Artificial Intelligence for Pedagogy
PostIS: postintervention survey
PreIS: preintervention survey
TPACK: Technological Pedagogical Content Knowledge

Edited by A Stone, Z Khlaif; submitted 18.Dec.2025; peer-reviewed by D Hilty, N Khamis; revised version received 24.Mar.2026; accepted 27.Mar.2026; published 06.May.2026.

Please cite as:
Anand R, Bowers N, Manyama MF
Design and Evaluation of a Faculty Development Workshop Series on Integrating Generative Artificial Intelligence in Medical Education: Mixed Methods Pilot Study
JMIR Med Educ 2026;12:e89815
URL: https://mededu.jmir.org/2026/1/e89815
doi:10.2196/89815

© Rajalakshmi Anand, Nicole Bowers, Mange Festo Manyama. Originally published in JMIR Medical Education (https://mededu.jmir.org), 6.May.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Education, is properly cited. The complete bibliographic information, a link to the original publication on https://mededu.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

A Competency Framework for Medical AI Education: Mixed Methods Study

Chang Cai[1,2], MAS; Gaoxia Zhu[3], PhD; Shang-Ming Zhou[4], PhD; Olivia Ng[1,2], PhD; Jamie Duell[5,6], PhD; Weng Kin Ho[3], PhD; Daisy Minghui Chen[1], MAS; Bernett Lee[1,7,8], PhD; Fang Li[9], PhD; Siyuan Liu[9], PhD; Vidya Sudarshan[9], PhD; Li Rong Wang[9,10], BSc; Chanwoo Choi[9], MAS; Xiuiyi Fan[1,2,9,11], PhD

[1]Lee Kong Chian School of Medicine, Nanyang Technological University, 50 Nanyang Avenue, Singapore, Singapore
[10]Centre for Frontier AI Research, Agency for Science, Technology and Research (A*STAR), Singapore, Singapore
[11]Centre of AI in Medicine, Nanyang Technological University, Singapore, Singapore
[2]National Health Group, Singapore, Singapore
[3]National Institute of Education, Nanyang Technological University, Singapore, Singapore
[4]Center for Health Technology, Faculty of Health, University of Plymouth, Plymouth, United Kingdom
[5]School of Computing and Digital Technologies, Sheffield Hallam University, Sheffield, United Kingdom
[6]Centre of Excellence in AI and Robotics, Sheffield Hallam University, Sheffield, United Kingdom
[7]Singapore Immunology Network (SIgN), Agency for Science, Technology and Research (A*STAR), Singapore, Singapore
[8]Infectious Disease Labs (ID Labs), Agency for Science, Technology and Research (A*STAR), Singapore, Singapore
[9]College of Computing and Data Science, Nanyang Technological University, Singapore, Singapore

Corresponding Author:
Xiuiyi Fan, PhD
Lee Kong Chian School of Medicine, Nanyang Technological University, 50 Nanyang Avenue, Singapore, Singapore

Abstract

Background: Although artificial intelligence (AI) is increasingly adopted in health care, clinicians face barriers, including insufficient understanding, limited trust, and challenges in interpreting AI outputs. Existing frameworks, such as the United Nations Educational, Scientific and Cultural Organization (UNESCO) AI competency framework, lack clinical specificity. Additionally, there remains limited evidence on framework-based training programs for medical professionals.

Objective: This study aimed to (1) develop a medical AI competency framework and (2) apply the framework to design an AI training program and pilot 1 module.

Methods: We conducted a mixed methods study comprising 2 phases. In Study 1, we developed a medical AI competency framework by integrating the UNESCO AI framework with the Miller pyramid model. The framework was refined through expert input from 24 stakeholders (6 hospital administrators, 8 medical professionals, and 10 university instructors). Expert responses were analyzed using deductive content analysis based on predefined codebooks. In study 2, we designed an AI training program based on the framework, and evaluated it using a 2-round Delphi process. Nine educators with expertise in instructional design, medical education, and AI participated in the Delphi study. Consensus in round 1 was defined as an IQR≤1, agreement score proportion >75%, and full-score frequency >49.23%; in round 2, consensus was defined as agreement score proportion >80%. A pilot workshop with 28 participants and 4 instructors assessed the feasibility of 1 module using self-reported measures of satisfaction, engagement, and confidence.

Results: A 6D 4-level medical AI competency framework was developed. Among the 24 experts, 19 (79.17%) mentioned competencies related to AI foundations, and 23 (95.83%) mentioned competencies related to application skills. The framework was translated into a 5-module training program, covering patient-centered and ethical AI, privacy and security, medical AI applications, bias and health equity, and generative AI in health care. Each module included 5 elements: content, learning goals, teaching activities, learning resources, and assessment. The Delphi process achieved complete consensus across all 25 elements of the training program. The pilot workshop indicated high participant satisfaction (mean 4.00, SD 0.52), good engagement (mean 3.80, SD 0.71 - mean 4.05, SD 0.51), and moderate self-reported confidence (mean 3.63, SD 0.53). These findings suggest that the module was feasible, although outcomes should be interpreted cautiously given the self-reported and short-term nature of the evaluation.

Conclusions: The framework provides a structured reference for AI training program design in medical education. The workshop findings provide preliminary support for the feasibility of the program's technical module, while also highlighting the need for broader and longer-term evaluations. Future work should expand the framework and training program to new regions and delivery formats (eg, semester-long courses and continuing medical education) and evaluate their long-term impact.

XSL•FO
RenderX

(JMIR Med Educ 2026;12:e91116) doi:10.2196/91116

KEYWORDS

artificial intelligence; AI; AI competency framework; AI literacy; AI training; medical education; medical professionals; United Nations Educational, Scientific and Cultural Organization; UNESCO

Introduction

Artificial intelligence (AI) is transforming health care by enabling early diagnosis of diseases, real-time health monitoring, and personalized health care [1-4]. It also contributes to health care management by enhancing patient satisfaction and optimizing hospital resources [5]. Nevertheless, clinicians encounter multiple barriers to AI adoption, including risks of misuse, insufficient understanding, limited trust in AI, and challenges in interpreting AI outputs [6,7].

These challenges underscore the need for clinicians to acquire AI competencies, as "AI is not meant to replace medical professionals, but the ones using AI will probably replace those who don't" [8, page 4]. Additionally, responsible AI adoption depends not only on external safeguards [9] but also on the capacities of medical professionals to interpret, evaluate, and oversee AI use in health care [10-16].

However, a dedicated AI competency framework tailored to medical professionals remains lacking. Global initiatives, such as the United Nations Educational, Scientific and Cultural Organization (UNESCO)'s AI competency framework [17], offer general guidelines for AI literacy but do not address the specific requirements of clinical practice. AI in health care requires unique considerations, including patient-centered decision-making, data governance, and ethical responsibility [11,12,18-20]. For example, clinicians need to explain AI-supported recommendations to patients, assess the reliability and limitations of AI outputs, and ensure that the use of AI aligns with patient values [10-12]. Moreover, proposed health care–related AI competencies are often limited to specific specialties, educational levels, or traditional AI technologies [13,15]. Emerging technologies, such as generative AI (GenAI), further highlight the insufficiency of existing proposed AI competencies. Although GenAI has considerable potential for health care [21], it also introduces challenges, including hallucinations and source uncertainty [22,23]. These challenges differ from those of traditional AI systems and require new competencies, such as verifying AI-generated content and ensuring clinician oversight [21,24]. However, few studies have examined the specific competencies required for medical professionals to use GenAI safely and effectively.

Furthermore, the absence of a dedicated competency framework has implications for medical education. Existing medical AI education initiatives, such as extracurricular workshops, online courses, and specialized lectures [25-29], often focus on specific topics rather than a shared foundational competency structure. As a result, there is limited guidance on how AI skills should be developed, assessed, and integrated into curricula. Recent reviews have highlighted this gap, noting the lack of framework-guided curriculum design in medical AI education [30,31]. More broadly, AI education should not be limited to technical content coverage alone. It also requires structured pathways and guidance to help learners address application risks and support responsible implementation [32].

These gaps highlight the need for a comprehensive, clinically grounded, and future-oriented medical AI competency framework. Such a framework should be directly applicable to curriculum development. To address these issues, this study aims to answer the following research questions:

- What core AI competencies should medical professionals possess to meet the evolving demands of AI-enabled health care?
- How can we use an AI competency framework to build a training program?

This study makes 3 contributions. First, we develop a medical AI competency framework, aligning AI competencies with clinical demands. Second, we showed the practical application of the framework by designing an AI training program. Third, we provide preliminary evidence of feasibility through pilot implementation, offering insights into the design and delivery of AI curricula in medicine. This research may contribute to the integration of AI into medical education and practice.

A preliminary version of this work was presented in the Artificial Intelligence in Education 2025 conference proceedings [33]. That paper reported on the development and refinement of the medical AI competency framework. This study substantially extends that foundational work in several important ways. First, it provides a more detailed and rigorous description of the framework development and expert-informed refinement process. Second, it extends the framework by incorporating observable behaviors, making it more actionable for teaching and assessment. Third, building on the framework, a training program for medical AI education was developed and evaluated through a 2-round Delphi process. Finally, a pilot study of 1 module of the program was implemented, and student feedback was collected to examine its feasibility. Overall, this study extends the preliminary work by providing a more actionable framework, showing its application in curriculum development, and offering preliminary evidence of implementation feasibility through a pilot module.

Methods

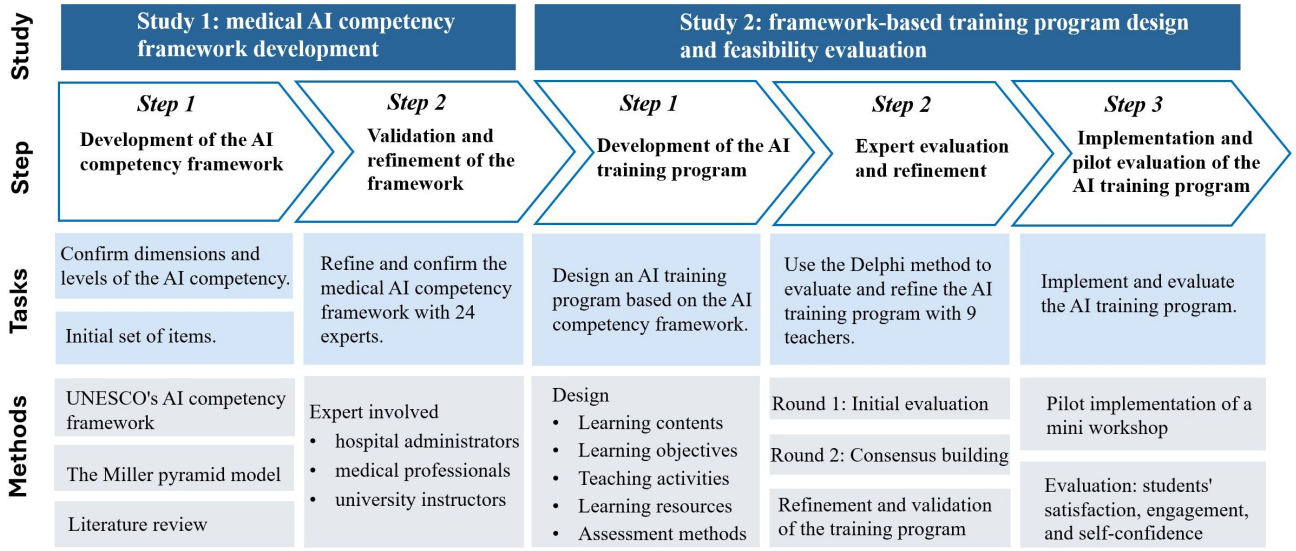
Overview

This research consists of 2 studies, as shown in Figure 1. Study 1 focused on the development of the medical AI competency framework. An initial version of study 1 was reported in our Artificial Intelligence in Education 2025 conference paper [33]; in this paper, this component is described in greater methodological detail, while study 2 is newly reported. Study 2 showed one practical application of this framework. We first designed a training program grounded in the competency

XSL•FO
RenderX

framework. The program was subsequently evaluated and refined through a 2-round Delphi process. Finally, a mini workshop was carried out to examine the program's feasibility.

Figure 1. Five-step development of the medical artificial intelligence (AI) competency framework and training program. UNESCO: United Nations Educational, Scientific and Cultural Organization.



Study 1: Development of the Medical AI Competency Framework

Step 1: Initial Medical AI Competency Framework Conceptualization

This step establishes the theoretical foundation for the initial medical AI competency framework. The framework was developed based on (1) UNESCO's AI competency framework [17], (2) the Miller pyramid [34], and (3) prior literature on AI competencies in medicine [11,12,15,35,36].

Theoretical Foundation

To conceptualize medical AI competence, we first drew on UNESCO's AI competency framework [17], a general guideline for AI literacy development. It outlines four core domains [17]: (1) human-centered mindset (developing a human-focused perspective on AI's benefits, risks, and proportionality in addressing societal and environmental needs), (2) ethics of AI (acquiring social and ethical competencies to guide the responsible use of AI throughout its lifecycle), (3) AI techniques and applications (combining conceptual understanding with practical skills through the real-world application of AI tools), and (4) AI system design (developing technical proficiency in designing, implementing, and refining AI systems).

Building on these domains and prior literature [11,12,15,17,35,36], we derived 6 competency dimensions tailored to the needs of health care practice. From UNESCO's *human-centered mindset*, we introduced *patient-centered AI in health care* to capture clinicians' ability to interpret and apply AI-supported decisions while preserving autonomy and maintaining human oversight. Extending UNESCO's *ethics of AI*, we proposed *ethics and transparency in clinical AI*, highlighting ethical principles in medical AI implementation. From UNESCO's *AI techniques and applications*, we derived *technical proficiency in medical AI applications*, emphasizing practical AI use by clinicians. Integrating UNESCO's *AI system design* with a focus on health care equity [37,38], we introduced *bias mitigation and health equity in AI design*. Considering the importance of data privacy and governance in medicine [19,20], we incorporated *data privacy, security, and compliance in health care AI*. Finally, recognizing the expanding role and challenges of GenAI in health care [21-23], we proposed *GenAI for health care*.

Proficiency Levels

Rather than UNESCO's 3-level progression (understand, apply, and create), we adopted the Miller pyramid model [34] to guide competency development. Unlike generic progression models, the Miller pyramid provides a clinically oriented structure that differentiates knowledge, application, simulated performance, and real clinical practice [34], making it well-suited for medical professionals. Widely used in medical education to assess clinical skills [39,40], it defines stepwise progression for competency development. The Miller pyramid includes the following levels [34]:

1. Know (knowledge): acquisition of foundational knowledge.
2. Know how (competency): the ability to apply knowledge in context.
3. Show how (performance): demonstration of skills in simulated settings.
4. Do (action): application of skills in real clinical practice.

These 6 dimensions and 4 levels jointly constituted the initial medical AI competency framework (Table S1 in Multimedia Appendix 1).

Step 2: Expert-Informed Framework Refinement

In this step, the initial framework was used to build deductive codebooks (Tables S2-S7 in Multimedia Appendix 1). The codebook was used to analyze experts' responses to the open-ended questions.

XSL•FO
RenderX

Expert Panel

To examine whether the initial framework aligned with expert views and to refine it where needed, a total of 24 experts from Singapore were selected, including 6 hospital administrators, 8 medical professionals, and 10 university instructors (Table 1). These diverse stakeholders ensured alignment of the framework with health care system needs and pedagogical principles. The selection of Singaporean experts was due to logistical constraints, and future studies will expand to include international perspectives.

Table . Demographics and selection criteria of framework validation experts.

Experts	Number of experts	Department	Selection criteria
Hospital administrators	6	Hospitals and government agencies	Leaders from hospitals or government agencies with more than 5 years of experience in policymaking and management.
University instructors	10	A university and health care organization	Teachers from universities or health care organizations with more than 5 years of teaching experience in AI[a] or medicine.
Medical professionals	8	Hospitals	Medical professionals with more than 5 years of clinical experience and prior exposure to AI applications.

[a]AI: artificial intelligence.

Data Collection

Experts were invited via email between August 2024 and December 2024 and completed an anonymous online survey with open-ended questions. Responses were collected anonymously to ensure participant confidentiality. Participants had the opportunity to review and modify their responses before submission.

The survey included the following questions for all experts:

- What AI-related skills do you think medical professionals should acquire?
- What topics in AI for medicine, do you consider essential for medical professionals to learn?
- For each AI-related skill mentioned above, to what extent do you think medical professionals should possess it (eg, basic awareness, practical knowledge, high-level skills, or advanced expertise)?

For medical professionals, additional questions included the following:

- Based on your own work experience, how can AI complement medical practices?
- What are the main barriers you have encountered when using or attempting to use AI?
- What challenges have you faced, or do you anticipate facing, when co-designing and co-developing a medical AI product?

Data Analysis

Expert feedback was analyzed through deductive content analysis using predefined codebooks (Tables S2-S7 in Multimedia Appendix 1). The codebooks were developed based on the initial medical AI competency framework (Table S1 in Multimedia Appendix 1), which was derived from our prior work [33]. Two researchers independently coded one-third of the responses using the codebooks to assess inter-rater reliability. The κ coefficient was 0.81, indicating strong agreement [41]. Researchers assigned each response to the most relevant competency category. Comments that did not fit existing definitions or suggested a new competency were flagged, discussed iteratively, and refined. In this study, we did not identify expert comments that justified adding a new dimension. However, several code definitions were refined to better reflect the expert comments while preserving the broader 6D structure. Discrepancies were resolved through iterative discussion until full consensus was reached. After establishing reliability, 1 researcher coded the remaining responses using the codebooks.

Study 2: Framework-Based Training Program Design and Feasibility Evaluation

Step 1: Framework-Based AI Training Program Development

We developed an AI training program as an application of the framework. Prior AI in Medicine initiatives suggested that combining lectures with case-based learning, hands-on activities, and problem-based learning supports the application of AI concepts in clinical contexts and enhances learner engagement [25,26,42,43]. Additionally, flipped classrooms have been extensively adopted to improve learning outcomes [25,44]. Integration of online and offline resources can support teaching design by aligning preclass learning, in-class interaction, and practice [45]. To support experiential and reflective learning, the Kolb experiential learning cycle and the Gibbs reflective cycle were introduced [46,47]. These pedagogical and theoretical principles informed the design of the program's learning activities and assessment strategies.

The program design included 5 core components: content, learning objectives, teaching activities, learning resources, and assessment methods. Modules were aligned with the competency framework, and learning objectives were mapped to the Miller pyramid. We also adopted teaching activities aligned with the

XSL•FO
RenderX

Miller pyramid. Lectures build foundational knowledge (know); case studies, discussions, and hands-on activities enable application and demonstration (know how and show how); and problem-based learning fosters real-world problem-solving (do). Various resources were provided: (1) curated datasets for hands-on data analysis, (2) case studies on AI's clinical application, and (3) online platforms for coding.

Step 2: Evaluation and Refinement of the Training Program

To validate the AI training program, we used the Delphi method, inviting 9 educators with expertise in AI, medical education, and instructional design.

Course Evaluation Panel

The panel was selected based on three criteria: (1) participants have designed at least 1 course, (2) participants have teaching experience in higher education, and (3) participants are from the fields of education, medicine, or AI.

As detailed in Table 2, the course evaluation panel included 2 education specialists from the education department of a public university in Singapore, 3 AI teachers from the computer science department of a public university in Singapore, and 4 medical teachers from the medical departments of public universities in Singapore and the United Kingdom. This multidisciplinary composition ensured a balanced integration of pedagogical, AI, and clinical expertise, providing a comprehensive evaluation of the training program.

Table . Demographics of the course evaluation panel participant.

Participant	Number of experts	Department	Nationality
Education experts	2	Educational department	Singapore
AI[a] educators	3	Computer science department	Singapore
Medicine educators	4	Medicine department	Singapore and the United Kingdom

[a]AI: artificial intelligence.

Delphi Method

The Delphi method is a well-established, anonymous group process designed to achieve consensus among participants [48,49]. This method has been widely used in previous studies to support and validate course design [50-52]. We conducted a 2-round Delphi process over 1 month to reach consensus on the AI training program design. No consultation on methods took place. Participants were invited by email and completed online questionnaires. The process followed a quasi-anonymous design: panelists were not informed of each other's identities, and no direct interaction occurred between participants. Responses were deidentified prior to analysis.

In round 1, participants reviewed the proposed 5-module training program and rated 25 items (covering content, learning objectives, teaching activities, learning resources, and assessment) using a 5-point Likert scale (1="strongly disagree" to 5="strongly agree"). Open-ended comments were also collected to inform revisions. Qualitative feedback from round 1 was synthesized by 1 researcher by deduplicating and grouping similar suggestions. The grouped suggestions were then summarized into specific revisions (eg, module, element, and corresponding suggestions). A second researcher reviewed the synthesized feedback, and discrepancies were resolved through discussion. The aggregated feedback was then provided to the same panel in round 2. Participants re-evaluated the revised items based on the summarized feedback from round 1 using the same rating scale. The round 1 and round 2 surveys can be accessed in Multimedia Appendix 2.

Data Analysis and Consensus

Panel ratings were analyzed using the IQR, agreement score (AS) proportion (proportion of ratings ≥4), and full-score (FS) frequency (proportion of ratings=5), consistent with previous studies [52-57]. All consensus thresholds were prespecified. In round 1, an item was considered to have reached consensus if it met all 3 criteria: IQR ≤1, AS >75%, and FS above a predefined threshold. Following previous work [58], the threshold for FS was calculated as mean–SD, which yielded 49.23% (63.11%-13.88%). In this study, FS>49.23% corresponds to more than half of the panel assigning the highest score. Items that did not meet these criteria were carried forward to round 2. In round 2, consensus was defined as achieving AS>80% [55-57].

Step 3: Pilot Implementation of a Mini Workshop

To assess the feasibility of the designed AI training program, we conducted a pilot mini workshop. Due to time constraints, the workshop focused primarily on module 3 (medical AI applications), while the remaining modules remained at the design stage and were not fully implemented. Module 3 was selected because it represented the core technical component of the program. A 1-day in-person workshop was held in July 2025. A total of 28 participants and 4 instructors attended. During in-class sessions, participants worked in small groups to analyze real-world AI use cases in health care. The teaching activities included lectures, hands-on sessions, group presentations, and reflective discussions.

After the workshop, participants were invited to complete a voluntary and anonymous online postclass survey (Table 3). The 30 items were adapted from validated instruments: the Student Satisfaction and Self-Confidence in Learning scale and the Student Engagement scale [59,60], covering satisfaction, self-confidence, and engagement in learning. We also collected instructors' insights through a survey with 6 open-ended questions. Full surveys are available in Multimedia Appendix 2. Participants were able to review and modify their responses before submission. Descriptive statistics were used to summarize questionnaire responses. Open-ended responses were analyzed using the 6-step thematic analysis approach, developed by Braun

XSL•FO
RenderX

and Clarke [61]. One researcher read all responses repeatedly and generated initial codes. These codes were then grouped into preliminary themes. A second researcher subsequently reviewed the coding and the preliminary thematic structure. Any discrepancies in code interpretation or theme grouping were discussed between the 2 researchers until consensus was reached. Through this process, the final themes were named and synthesized.

Table . Structure of the postworkshop survey.

Participants, dimension, and subdimension	Number of items	Type
Students		
Satisfaction	5	5-point Likert scale
Engagement		
Behavioral engagement	5	5-point Likert scale
Emotional engagement	4	5-point Likert scale
Cognitive engagement	8	5-point Likert scale
Self-confidence	8	5-point Likert scale
Overall feedback	2	Open-ended questions
Instructors		
Overall feedback	6	Open-ended questions

Ethical Considerations

The study was reviewed and approved by the institutional review board of Nanyang Technological University (IRB-2024‐928). Written informed consent was waived by the review board, as the study posed minimal risk and involved anonymous surveys conducted online. Online informed consent was provided to all participants prior to data collection, and participation was entirely voluntary. All survey data were collected anonymously with no personally identifiable information recorded, and access was restricted to the research team. Participants did not receive any compensation.

Results

Results of Study 1: Medical AI Competency Framework Development

Expert Validation and Framework Refinement

Most experts highlighted the importance of “AI foundations” (n=19) and “application skills” (n=23). These findings highlight the need for core AI knowledge and the ability to learn, apply, and co-develop AI solutions within medical workflows. As a medical professional emphasized, “gaining basic AI knowledge, gaining data analysis skills, and applying AI in clinical settings to improve patient care” were seen as essential competencies. Similarly, a hospital administrator noted that expected competencies include being “able to identify problems that can be solved by AI, understanding the process of deployment and evaluation of AI projects, applying AI knowledge in innovation, and possessing data analysis skills.”

In addition to technical competencies, 16 experts highlighted the importance of “critical reflections on AI,” emphasizing the ethical and safe use of AI in medicine. As a medical professional pointed out, “AI ethics concepts are similar to medical ethics concepts ... [professionals should] learn more about AI ethics and governance.” Moreover, 5 experts emphasized the integration of GenAI into clinical workflows. For example, a medical professional stated, “Using ChatGPT to conceptualize brief background of certain issues.” Based on expert feedback for greater clarity, “application skills” and “co-design equitable AI systems” were redefined to emphasize collaboration with engineers. For example, 1 hospital administrator highlighted:

> *Not to become a coder/developer/engineer, but to get skillset to be able to comprehensively lead teams and run AI projects. Some technical + system thinking + appreciation of broader landscape.*

Furthermore, the definition of “real-world compliance” was expanded to stress the proper handling of both sensitive and nonsensitive data in clinical contexts. Detailed results are provided in Table S8 in Multimedia Appendix 3.

Final Medical AI Competency Framework

Table 4 and Table 5 show the final framework, which comprises 6 dimensions across 4 levels, ranging from foundational knowledge (know) to clinical practice (do).

XSL•FO
RenderX

Table . Medical artificial intelligence (AI) competency framework (part 1; dimensions 1 - 3).

Aspects	Progression			
	Know	Know how	Show how	Do
Patient-centered AI[a] in health care	• Basic understanding: understand the foundational impact of AI on patient rights and autonomy • Observable behaviors: identification of patient rights affected by AI use; recognition of patient autonomy concerns in AI-supported care; identification of patient-centered risks in AI use	• Human advancement: learn how to apply AI to enhance treatment planning, automate care processes, and personalize health care for individual patients • Observable behaviors: explanation of how AI supports treatment and personalization; identification of appropriate AI use in care processes	• Patient-centered decision-making: use AI effectively to support decision-making, ensuring patient well-being is prioritized • Observable behaviors: interpretation of AI outputs; comparison of AI recommendations with patient needs; justification of AI decisions based on patient well-being	• Patient-centered practice: integrate AI tools into clinical practice, taking responsibility for incorporating them into patient care workflows to improve care quality and outcomes • Observable behaviors: communication with patients about AI outputs in routine care workflows; integration of AI into patient care workflows with clinician oversight; accountable use of AI to support care quality
Ethics and transparency in clinical AI	• Critical reflections on AI: comprehend the basic ethical principles guiding the use of AI in health care settings • Observable behaviors: identification of core ethical principles in clinical AI; recognition of the importance of transparency in health care AI; recognition of ethical concerns in AI-supported care	• Safe and responsible use: know how to use AI in ways that ensure ethical considerations and transparency • Observable behaviors: explanation of ethical principles in health care AI use; consideration of transparency when planning AI-supported decisions; identification of ethical risks in planned AI use	• Ethical decision-making: demonstrate ethical considerations during AI-enabled decision-making in simulated settings • Observable behaviors: appraisal of ethical issues in simulated AI-supported cases; justification of decisions based on ethical and transparency considerations	• Ethics by design: implement ethical standards in real-world AI applications, evaluating and improving systems for ethical compliance • Observable behaviors: adherence to ethical standards in clinical AI practice; monitoring of AI use for ethical compliance; improvement of AI systems for ethical and transparent practice; documentation of ethical concerns in practice
Data privacy, security, and compliance in health care AI	• Privacy awareness: remember regulations and the importance of safeguarding patient data • Observable behaviors: awareness of privacy and data governance regulations; recognition of the importance of patient data protection	• Regulatory adherence: know how to handle patient data securely, using techniques like anonymization and governance • Observable behaviors: explanation of secure data handling practices; identification of data anonymization approaches	• Secure data practices: manage patient data in simulated research projects, adhering to privacy laws • Observable behaviors: application of data protection measures in structured tasks; management of data under privacy constraints in projects	• Real-world compliance: process clinical data under specific conditions, ensuring all data handling and AI applications comply with privacy and security standards in clinical and research settings • Observable behaviors: secure handling of clinical data in routine workflows; responsible data use in clinical and research settings

[a]AI: artificial intelligence.

XSL•FO
RenderX

Table . Medical artificial intelligence (AI) competency framework (part 2; dimensions 4-6).

Aspects	Progression			
	Know	Know how	Show how	Do
Technical proficiency in medical AI[a] applications	• AI foundations: acquire foundational knowledge of AI concepts and techniques • Observable behaviors: awareness of core AI concepts; recognition of basic AI methods and models; understanding of algorithms in health care AI	• Health care AI implementation: learn how to apply AI methods effectively in medical contexts • Observable behaviors: explanation of AI applications in medical contexts; explanation of AI support for clinical decision-making; identification of appropriate clinical use cases for AI	• Application skills: grasp the ability to perform and cocreate basic AI-related programming or data analysis tasks • Observable behaviors: performance of basic AI-related programming tasks; performance of basic data analysis tasks	• Evaluate and create AI: critically assess existing AI systems and develop solutions tailored to specific health care needs • Observable behaviors: evaluation of AI systems for clinical relevance; development of AI solutions for health care needs; implementation of AI tools in real-world settings
Bias mitigation and health equity in AI design	• Foundational understanding of bias: Identify how AI impacts health equity and its potential to either mitigate or exacerbate disparities • Observable behaviors: awareness of bias in health care AI; recognition of AI-related health disparities; understanding of AI's impact on health equity	• Bias detection and analysis: know how to identify, evaluate, and address bias in AI systems within health care contexts • Observable behaviors: explanation of bias sources in health care AI; evaluation of bias risks in AI use; identification of bias risks in data and AI systems	• Equity-oriented practices: select and apply AI tools that minimize bias and enhance fairness, ensuring health equity in controlled design or assessment tasks • Observable behaviors: selection of AI tools with fairness considerations; application of bias mitigation approaches in structured tasks	• Co-design equitable AI systems: cocreate and refine AI solutions to enhance health equity, using iterative feedback to address systemic disparities • Observable behaviors: co-design of AI solutions for health equity; refinement of AI systems to reduce bias; iterative improvement of AI for equitable practice
GenAI[b] for health care	• GenAI basics: identify the basic principles, limitations, and benefits of GenAI • Observable behaviors: identification of core GenAI concepts; understanding of potential clinical uses of GenAI; recognition of GenAI limitations in health care	• Interpretative proficiency: know how to interpret AI outputs, adapt data, and identify risks like hallucinations. • Observable behaviors: Explanation of strategies for using GenAI tools; identification of hallucination and reliability risks	• Skillful development: use GenAI, validate outputs, and communicate effectively in clinical scenarios • Observable behaviors: validation of GenAI outputs in structured clinical tasks; writing of safe prompts; documentation of output limitations and need for human review	• Integration in practice: independently integrates, evaluates, and oversees GenAI in real clinical workflows • Observable behaviors: Integration of GenAI into real clinical workflows; verification of GenAI outputs before use in patient care; disclosure of limitations and maintenance of clinician oversight

[a]AI: artificial intelligence.

[b]GenAI: generative artificial intelligence.

Translation of the Study 1 Findings Into the Study 2 Training Program

The final competency framework developed in study 1 served as the reference for the training program design in study 2. Specifically, 6 competency dimensions informed the learning topics and content of the program, while 4 Miller-aligned levels guided the design of learning objectives and assessment tasks. Learning activities were further developed in alignment with both the competency framework and educational theories. As shown in Table 6, the findings from study 1 were translated into the training program by mapping 6 dimensions of the medical AI competency framework into 5 modules. "Patient-centered AI in health care" and "ethics and transparency in clinical AI" were combined into a single module because of their close conceptual relationship in clinical practice. In this way, study 2 represents one practical application of the competency framework developed in study 1.

XSL•FO
RenderX

Table . Mapping of the medical artificial intelligence (AI) competency framework to the AI training program.

Module	Competency dimension	Learning focus
Module 1: patient-centered and ethical AI in Health Care	• Patient-centered AI • Ethics and transparency in clinical AI	Ethical principles of AI use in health care, patient autonomy, fairness, transparency, and responsible clinical decision-making
Module 2: AI privacy and security in health care	• Data privacy, security, and compliance in health care AI	Data protection, regulatory frameworks (eg, HIPAA[a] and GDPR[b]), and secure handling of health care data in AI applications
Module 3: medical AI applications	• Technical proficiency in medical AI applications	Fundamental AI concepts, model implementation, and interpretation of AI outputs in health care contexts
Module 4: AI bias and health equity	• Bias mitigation and health equity in AI design	Identification of bias in AI systems, fairness considerations, and strategies to promote equitable health care outcomes
Module 5: generative AI in Health Care	• Generative AI for health care	Safe use of generative AI, including prompt design, output evaluation, and responsible integration into clinical workflows

[a]HIPAA: Health Insurance Portability and Accountability Act.

[b]GDPR: General Data Protection Regulation.

Results of Study 2: Framework-Based Training Program Design and Feasibility Evaluation

Delphi Round 1 and Round 2 Results

Nine experts participated in both Delphi rounds. After round 1, consensus was reached on 16 of the 25 (64%) training program elements. The remaining 9 elements were carried forward to round 2 because they did not meet at least one of the predefined round 1 consensus criteria. In round 2, agreement was reached on the remaining 9 elements. Table 7 shows the results of the 2-round survey. Table S9 in Multimedia Appendix 3 shows the full results of the 2-round survey.

Table . Results of the 2-round Delphi study.

Element of program and module	Round 1[a]			Round 2[b]
	Median (IQR)	AS[c] (%)	FS[d] (%)	AS (%)
Teaching activities				
Module 3	4 (1)	88.89	44.44	100
Learning resources				
Module 1	5 (2)	66.67	66.67	100
Module 2	5 (2)	66.67	66.67	88.89
Module 5	5 (2)	66.67	55.56	88.89
Learning assessments				
Module 1	4 (1)	88.89	44.44	100
Module 2	4 (1)	88.89	44.44	100
Module 3	4 (1)	88.89	33.33	100
Module 4	4 (1)	88.89	44.44	100
Module 5	4 (1)	88.89	44.44	100

[a]In round 1, consensus was defined as meeting all the 3 criteria: IQR≤1, AS >75%, and FS >49.23%.

[b]In round 2, consensus was defined as achieving AS>80%.

[c]AS: agreement score proportion, the proportion of responses within a defined score range (eg, 4 and 5 on a 5-point scale) [53].

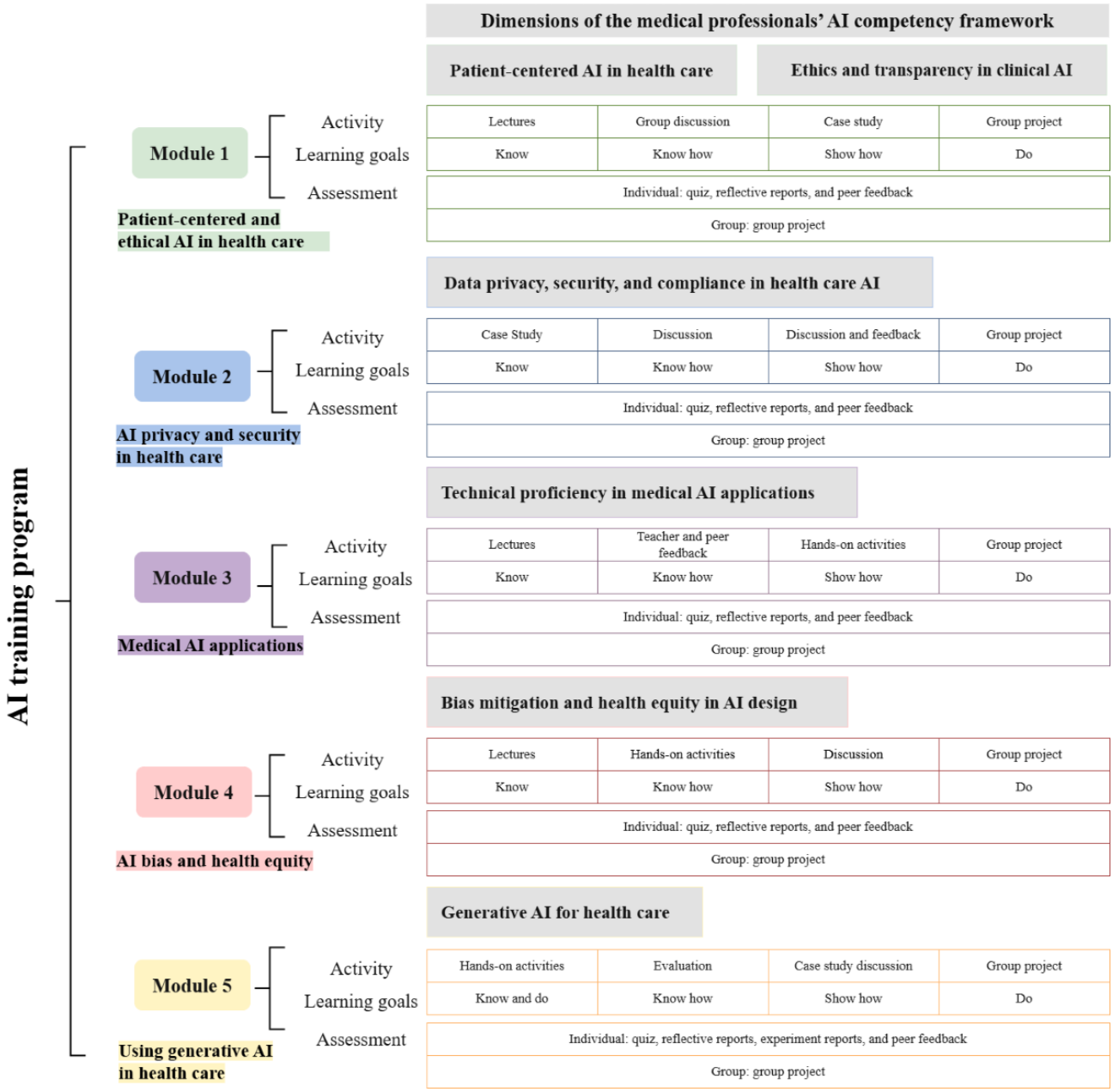
[d]FS: full score frequency, the proportion of the highest scores [54].

Final Framework–Based AI Training Program

Figure 2 shows the overview structure of the 5-module AI training program. Multimedia Appendix 4 shows the final program design.

XSL•FO
RenderX

Figure 2. Structure of the artificial intelligence (AI) training program.



Module 1: Patient-Centered and Ethical AI in Health Care

This module aligns with the framework domains of "patient-centered AI in health care" and "ethics and transparency in clinical AI." It introduces the ethical use of AI in health care, with a focus on patient-centered care, the impact of AI on patient rights, and the principles of transparency in clinical AI applications. Medical professionals are guided to balance ethical decision-making with AI integration, ensuring patient autonomy and the protection of patient rights. For example, Table 8 shows example learning tasks and assessments aligned with the "show how" and "do" levels of the Miller pyramid.

XSL•FO
RenderX

Table . Example alignment of learning tasks and assessment methods.

Module and level	Tasks	Assessment methods
Module 1		
Show how	Analyze a health care case involving AI[a]-supported diagnosis; identify issues related to patient autonomy, fairness, transparency, and human oversight; justify how patient interests should be protected.	Case analysis report or presentation with rubric-based assessment of ethical analysis, patient-centered health care, and justification of recommendations.
Do	Propose a proposal for an AI tool in health care that respects patient rights, supports fair decision-making, and includes clear human oversight in clinical use.	Group project with rubric-based assessment of patient-centered design, ethical safeguards, transparency, and practical integration into health care workflow.
Module 2		
Show how	Analyze privacy and security case studies involving improper disclosure, disposal, or handling of patient data.	Case-based task with structured rubric assessing identification of privacy and security risks.
Do	Design a health care AI project involving patient data; classify sensitive and nonsensitive data and propose protection measures.	Group project with rubric-based assessment of appropriateness of privacy and security measures and alignment with health care data governance requirements.
Module 3		
Show how	Run a preconfigured AI model using health care data; interpret performance metrics such as accuracy, F_1-scores, precision, and recall; modify code or model settings (eg, hyperparameters).	Hands-on coding task with rubric-based assessment of code execution, interpretation of model outputs, and appropriateness of modifications.
Do	Adapt a basic AI model to a health care problem and present how the model could support clinical tasks, including its strengths and limitations.	Group project with rubric-based assessment of technical implementation, performance improvement, relevance to the health care problem, and critical interpretation of model limitations.
Module 4		
Show how	Use crafted imbalanced datasets to identify bias in model development; examine biased predictions using fairness metrics such as false positive rates.	Written reports assessed with structured rubric (bias detection, fairness analysis, and model interpretation).
Do	Design an AI model using a crafted dataset, apply bias mitigation methods such as oversampling, and evaluate whether the revised model improves fairness and equity.	Group project with rubric-based assessment of model design, bias mitigation, and fairness outcomes.
Module 5		
Show how	Write safe and effective prompts for clinical tasks; identify hallucinated or unsupported claims; analyze cases in which GenAI[b] produces incorrect medical advice; document limitations explicitly.	Case-based task with a structured rubric assessing prompt quality, error detection, justification of validation steps, and documentation of limitations; individual report evaluating AI-generated outputs.
Do	Design a GenAI-supported solution based on a health care task (eg, documentation or literature synthesis); implement validation steps; specify when human review is required.	Group project with rubric-based assessment of use-case appropriateness, output reliability, validation strategy, transparency of limitations, and workflow integration.

[a]AI: artificial intelligence.

[b]GenAI: generative artificial intelligence.

Module 2: AI Privacy and Security in Health Care

This module corresponds to the framework domain of "data privacy, security, and compliance in health care AI." It covers knowledge of data privacy, security, and regulatory compliance in medical AI applications. Medical professionals will learn key regulations (eg, General Data Protection Regulation and Health Insurance Portability and Accountability Act [HIPAA]) and ensure compliance in clinical settings. For example, Table 8 presents example learning tasks and assessments aligned with the "show how" and "do" levels of the Miller pyramid.

XSL•FO
RenderX

Module 3: Medical AI Applications

Aligned with the framework domain of "technical proficiency in medical AI applications," this module provides knowledge of AI applications in health care. Medical professionals will learn the foundational principles of medical AI, explore AI applications, and gain hands-on experience in building and evaluating AI models. The module emphasizes application skills as well as the ability to evaluate and create clinically meaningful AI solutions. For example, Table 8 shows example learning activities and assessments aligned with the "show how" and "do" levels of the Miller pyramid.

Module 4: AI Bias and Health Equity

This module corresponds to the framework domain of "bias mitigation and health equity in AI design." It develops medical professionals' foundational understanding of algorithmic bias and its unequal impact on diverse patient populations. Through hands-on work with real datasets and explainable AI tools, they will examine how bias arises in data and affects models. The module further emphasizes bias detection, equity-oriented practices, and the co-design of fair and inclusive AI systems. For example, Table 8 shows example learning activities and assessments aligned with the "show how" and "do" levels of the Miller pyramid.

Module 5: Using Generative AI in Health Care

Aligned with the framework domain of "generative AI for health care," this module introduces medical professionals to the principles and clinical applications of GenAI. They will acquire knowledge of generative models, learn prompt engineering, and gain practical experience applying GenAI. The module emphasizes the safe, effective, and responsible integration of GenAI into clinical practice. This emphasis is particularly important in health-related large language model applications, where outputs may be unverifiable or clinically incorrect, highlighting the need for validation, source grounding, and clinician oversight [21,62]. For example, Table 8 presents example learning activities and assessments aligned with the "show how" and "do" levels of the Miller pyramid.

Pilot Implementation (Mini Workshop) Results

Overview

Of the 28 participants, 15 reported their learning experiences (Table 9). All questionnaires had high internal consistency, with Cronbach α values greater than 0.8. The mean scores were 4 (SD 0.52) for satisfaction, 3.8 (SD 0.71) for behavioral engagement, 4.05 (SD 0.51) for emotional engagement, 3.96 (SD 0.57) for cognitive engagement, and 3.63 (SD 0.53) for self-confidence (Table 10). Figure 3 illustrates the distribution of participant responses across survey dimensions.

Table . Participant descriptions and demographics.

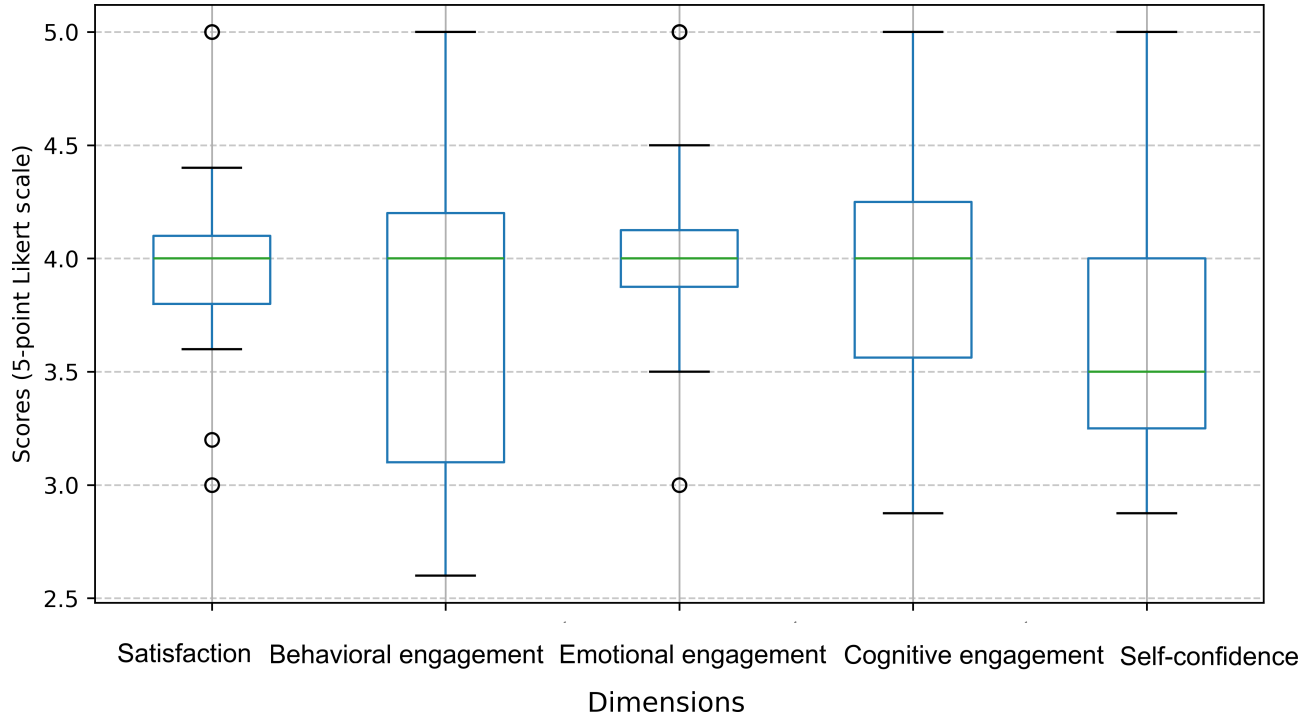
Characteristics and categories	Participants (n=28), n (%)	Survey respondents (n=15), n (%)
Roles		
Year 2 medical student	12 (42.9)	6 (40)
Year 3 medical student	4 (14.3)	2 (13.3)
Year 4 medical student	7 (25)	4 (26.7)
Year 5 medical student	5 (17.9)	3 (20)
Training stage		
Preclinical (year 1-2)	12 (42.9)	6 (40)
Clinical (year 3-5)	16 (57.1)	9 (60)
Motivation for attending		
Voluntary	28 (100)	15 (100)
Others	0 (0)	0 (0)

Table . Descriptive statistics of post–workshop survey dimension (n=15).

Dimension	Mean (SD)	Median (IQR)	Min-max
Satisfaction	4 (0.52)	4 (5-3)	3.00-5.00
Behavioral engagement	3.80 (0.71)	4 (5-2.60)	2.60-5.00
Emotional engagement	4.05 (0.51)	4 (5-3)	3.00-5.00
Cognitive engagement	3.96 (0.57)	4 (5-2.88)	2.88-5.00
Self-confidence	3.63 (0.53)	3.50 (5-2.88)	2.88-5.00

XSL•FO
RenderX

Figure 3. Distribution of 15 participant responses across the post–workshop survey dimension.



Qualitative findings from open-ended responses revealed 3 main themes that reflected participants' learning experiences and perceptions of the workshop.

Perceived Understanding of Coding and AI-Related Concepts

Participants reported learning coding basics, AI applications in medicine, and fundamental machine learning concepts. For instance, a participant described acquiring "a basic overview on AI, prompt generation, and machine learning." Several participants also emphasized exposure to specific machine learning techniques, such as "oversampling, undersampling, and Shapley Additive Explanations plots."

Development of Ethical and Reflective Awareness in AI Use

Beyond technical skills, participants demonstrated an increased awareness of ethical considerations and the reflective use of AI tools. Some participants mentioned learning about "AI ethics in medicine." In addition, they reflected on how AI could support ongoing learning, with a participant noting learning "how to use AI tools to help with producing workable code and what concepts I need to be reading about for future self-learning."

Learning Experience and Suggestions

Overall, students expressed positive perceptions of the learning environment, particularly valuing the instructors' support and interactive teaching approach. Participants also offered constructive suggestions for improvement, such as covering "practical aspects of AI that can be applied to research projects."

All instructors indicated that the "AI in Medicine" training workshop was well received, engaged students, and largely achieved its learning objectives, although some complex content required additional time. While participants enjoyed the workshop, their confidence in applying AI remained limited. For example, an instructor noted:

> *Participants demonstrated strong engagement and interest throughout the workshops. They consistently asked questions, ... , they found some of the machine learning content challenging. They asked numerous questions related to the depth of the ML material, which suggests that time constraints limited their ability to fully absorb these concepts. Several participants expressed a desire for additional time to gain a deeper understanding of the machine learning and explainable AI methods covered in the sessions.*

Instructors also offered several suggestions for future implementations, including providing hands-on and practical sessions (emphasizing experiential learning and group activities, as these are the most engaging and effective), simplifying and allocating more time (simplifying technical content for nontechnical participants and providing more time and clearer instructions for practical exercises), and changing content depth and focus (narrowing the range of machine learning methods and exploring key topics in greater depth).

These self-reported findings align with several topics identified in study 1's framework, particularly the basic understanding of AI concepts and their application in health care.

Discussion

Principal Results

This study addresses a critical gap in global medical education: the absence of standardized medical AI curriculum frameworks [30,31]. By integrating the UNESCO AI competency framework with the Miller pyramid, we translated high-level AI literacy principles into clinically actionable competencies across 6 dimensions. The 6D framework provides a structured reference

XSL•FO
RenderX

for competency development, spanning foundational AI understanding to ethical, patient-centered, and practice-integrated application.

Building on this framework, we translated these competencies into a structured training program and piloted a workshop. The pilot implementation of module 3 provided initial evidence of the program's feasibility, although the self-reported learning experience and confidence required further evaluation.

Comparison With Prior Work

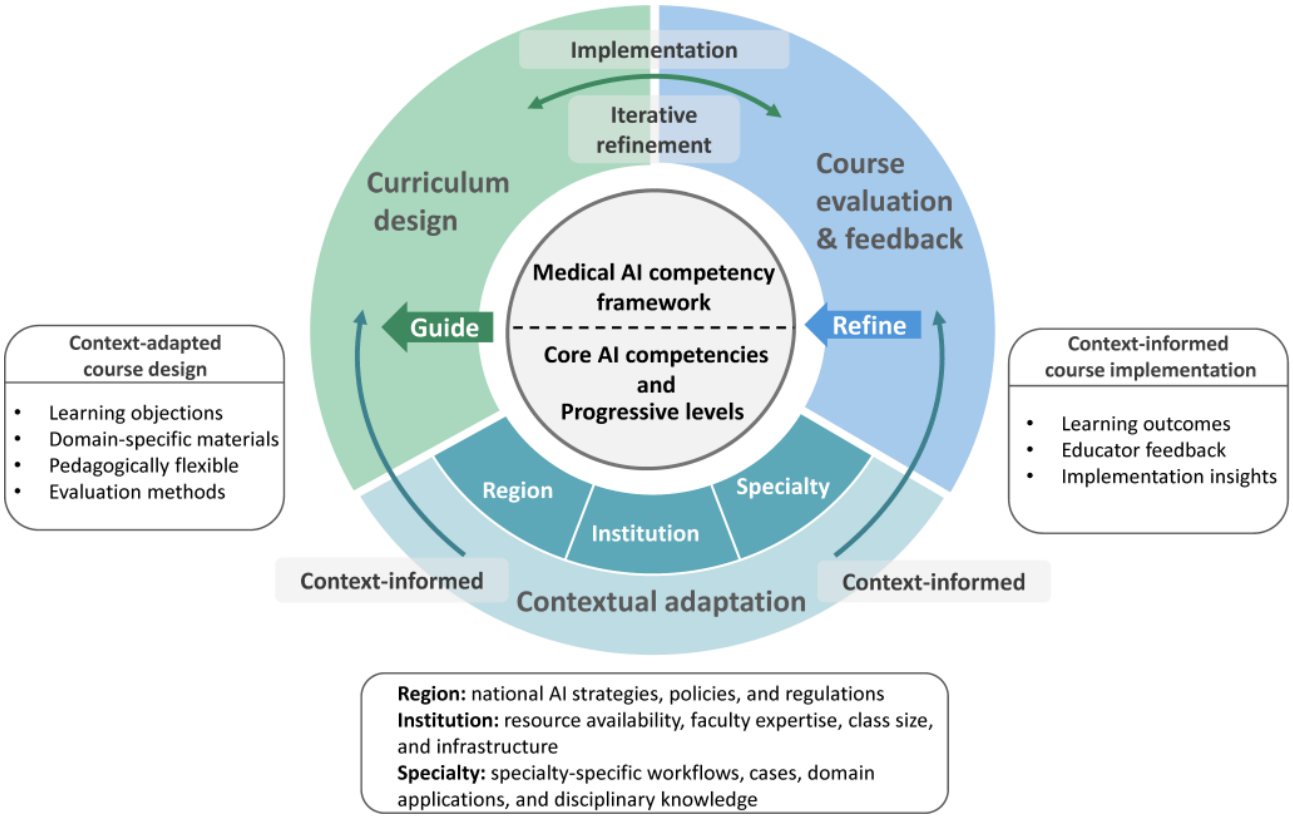
Extending UNESCO's AI competency framework to the medical domain [17], we incorporated domain-specific needs, such as patient-centered decision-making, ethical considerations in clinical AI, and health equity. Consistent with prior studies, our findings reinforce the importance of integrating both technical (eg, AI applications) and nontechnical competencies (eg, ethics) [11,12,15,35,36]. Importantly, this study responds to the growing impact of GenAI in health care by integrating competencies for its responsible use. Although recent literature has discussed both the applications and risks of GenAI in clinical contexts [21-23], few studies have articulated the competencies required for medical professionals to use these technologies safely and effectively. The inclusion of GenAI competencies provides a forward-looking perspective that is currently insufficiently examined in existing research.

Prior studies on AI education in medicine have been largely domain-specific, focusing on areas such as machine learning fundamentals or programming [26,27,29,42]. For example, some programs emphasized conceptual understanding of AI methods through lectures [26,29], while others focused on developing basic programming skills [25]. Although these initiatives contribute to increased awareness and technical competence, they are typically organized around discrete topics rather than being guided by a unified framework. Furthermore, the literature has highlighted the absence of unified competency frameworks to guide these efforts [30,31]. In contrast, our study presents a framework-guided curriculum design approach that translates competency domains into training programs.

Theoretical and Educational Implications

This study provides both theoretical and practical implications. Theoretically, it proposes a medical AI competency framework that defines shared foundational AI competencies for medical education. It focuses on how medical professionals interact with, evaluate, communicate, and apply AI in health care, rather than replacing discipline-specific knowledge. We, therefore, view it as a living framework (Figure 4). While it defines core competencies, its implementation should be tailored to regional, institutional, and specialty-specific contexts. Such adaptation occurs mainly at the curriculum level. For example, different specialties may require different cases, data, and learning materials. Feedback from course implementation can then support the iterative refinement of both the curriculum and the framework.

Figure 4. The medical artificial intelligence (AI) competency framework as a living framework.



Practically, the framework serves as a reference for curriculum design. By linking competency domains with the Miller levels, this framework supports curriculum mapping, the design of learning objectives and assessments, and the identification of training gaps. Additionally, prior work suggests that AI governance cannot rely solely on external constraints [9]. In this context, our framework supports the development of competencies for responsible AI use and may contribute to a

XSL•FO
RenderX

broader governance-oriented approach to AI adoption in health care. This study further presents a competency-based AI training program that may serve as a model for other institutions. In addition, the explicit integration of GenAI into the competency framework and curriculum is particularly timely, given its rapid adoption in health care, education, and research.

Limitations and Future Directions

There are several limitations to this study. First, although the experts in this study represented diverse backgrounds, the majority were based in Singapore. As cultural attitudes toward AI adoption differ significantly across regions [63], the framework may mainly reflect Singapore's specific health care and educational context. The framework should be viewed as a living framework, open to ongoing refinement. Future research can validate and extend it using more diverse expert samples and combined deductive-inductive approaches. Future work may also extend the framework to specialty-specific applications (eg, course design and implementation) in areas such as radiology, surgery, and public health, using locally relevant cases, datasets, and assessment methods.

Second, the pilot was conducted as a voluntary workshop in Singapore, which may limit generalizability. Participants may have had greater prior interest in AI, introducing self-selection bias and potentially inflating the reported results. In addition, the pilot evaluation was based only on self-reported measures, covered only 1 module, and had a modest response rate (15/28), which may introduce nonresponse bias. Accordingly, these findings should be interpreted as preliminary evidence of feasibility. Future studies should evaluate the full program in broader settings using preassessments and postassessments of competency gains.

Third, while our program emphasizes real-world applicability, its short-term nature limits the direct assessment of AI competencies in clinical practice. Future work should collaborate with health care institutions to integrate training into clinical workflows and examine the longer-term impact across multiple levels, for example, using the Kirkpatrick model [64].

Conclusions

We developed a theoretically grounded medical AI competency framework. By translating competencies into a structured training program, it provides a practical application for integrating AI into medical education. Ultimately, this work aims to contribute to the development of an AI-ready health care workforce capable of using AI safely, ethically, and effectively in clinical practice.

Acknowledgments

The authors used ChatGPT for language polishing during manuscript preparation. All content was reviewed and verified by the authors. Responsibility for the final manuscript rests entirely with the authors. Generative artificial intelligence tools are not listed as authors and bear no responsibility for the final content.

Funding

This research is supported by the Ministry of Education, Singapore (grant: RG22/23, RS15/23, and LKCMedicine Start Up Grant).

Authors' Contributions

Conceptualization: C Cai
Data curation: C Cai, C Choi, LRW, XF
Formal analysis: C Cai, XF
Funding acquisition: XF
Investigation: C Cai
Methodolgy: C Cai
Resources: C Choi, LRW, GZ, SMZ, ON, JD, WKH, DMC, BTKL, FL, SL, VS
Supervision: XF
Visualization: C Cai
Writing—original draft: C Cai
Writing—review and editing: C Cai, XF, GZ, SMZ, ON, JD, WKH, DMC, BTKL, FL, SL, VS

The authors used ChatGPT for language polishing during manuscript preparation. All content was reviewed and verified by the authors. Responsibility for the final manuscript rests entirely with the authors. Generative AI tools are not listed as authors and bear no responsibility for the final content.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Initial medical artificial intelligence (AI) competency framework and codebook for deductive content analysis.

XSL•FO
RenderX

[DOCX File, 26 KB - mededu_v12i1e91116_app1.docx]

Multimedia Appendix 2
Course evaluation surveys and post–workshop surveys.
[DOCX File, 14 KB - mededu_v12i1e91116_app2.docx]

Multimedia Appendix 3
Deductive coding and Delphi results.
[DOCX File, 23 KB - mededu_v12i1e91116_app3.docx]

Multimedia Appendix 4
Final artificial intelligence (AI) training program.
[DOCX File, 14 KB - mededu_v12i1e91116_app4.docx]

References

1. Singh P. Transforming healthcare through AI: enhancing patient outcomes and bridging accessibility gaps. SSRN. Preprint posted online on Jul 17, 2024. [doi: 10.2139/ssrn.5115767]
2. Kumar Y, Koul A, Singla R, Ijaz MF. Artificial intelligence in disease diagnosis: a systematic literature review, synthesizing framework and future research agenda. J Ambient Intell Humaniz Comput 2023;14(7):8459-8486. [doi: 10.1007/s12652-021-03612-z] [Medline: 35039756]
3. Johnson KB, Wei WQ, Weeraratne D, et al. Precision medicine, AI, and the future of personalized health care. Clin Transl Sci 2021 Jan;14(1):86-93. [doi: 10.1111/cts.12884] [Medline: 32961010]
4. Schork NJ. Artificial intelligence and personalized medicine. Cancer Treat Res 2019;178:265-283. [doi: 10.1007/978-3-030-16391-4_11] [Medline: 31209850]
5. Pillai AS. AI-enabled hospital management systems for modern healthcare: an analysis of system components and interdependencies. J Adv Anal Healthc Manag 2023;7(1):212-228 [FREE Full text]
6. Alanazi A. Clinicians' views on using artificial intelligence in healthcare: opportunities, challenges, and beyond. Cureus 2023;15(9):e45255. [doi: 10.7759/cureus.45255]
7. Esmaeilzadeh P. Challenges and strategies for wide-scale artificial intelligence (AI) deployment in healthcare practices: a perspective for healthcare organizations. Artif Intell Med 2024 May;151:102861. [doi: 10.1016/j.artmed.2024.102861] [Medline: 38555850]
8. Meskó B, Hetényi G, Győrffy Z. Will artificial intelligence solve the human resource crisis in healthcare? BMC Health Serv Res 2018 Jul 13;18(1):545. [doi: 10.1186/s12913-018-3359-4] [Medline: 30001717]
9. Chen L. Beyond external constraints: the missing dimension of AI governance. SSRN. Preprint posted online on Mar 21, 2026. [doi: 10.2139/ssrn.6449738]
10. Aslam TM, Hoyle DC. Translating the machine: skills that human clinicians must develop in the era of artificial intelligence. Ophthalmol Ther 2022 Feb;11(1):69-80. [doi: 10.1007/s40123-021-00430-6] [Medline: 34807411]
11. Charow R, Jeyakumar T, Younus S, et al. Artificial intelligence education programs for health care professionals: scoping review. JMIR Med Educ 2021 Dec 13;7(4):e31043. [doi: 10.2196/31043] [Medline: 34898458]
12. Lee J, Wu AS, Li D, Kulasegaram KM. Artificial intelligence in undergraduate medical education: a scoping review. Acad Med 2021 Nov 1;96(11S):S62-S70. [doi: 10.1097/ACM.0000000000004291] [Medline: 34348374]
13. Liaw W, Kueper JK, Lin S, Bazemore A, Kakadiaris I. Competencies for the use of artificial intelligence in primary care. Ann Fam Med 2022;20(6):559-563. [doi: 10.1370/afm.2887] [Medline: 36443071]
14. Sand M, Durán JM, Jongsma KR. Responsibility beyond design: Physicians' requirements for ethical medical AI. Bioethics 2022 Feb;36(2):162-169. [doi: 10.1111/bioe.12887] [Medline: 34089625]
15. Singla R, Pupic N, Ghaffarizadeh SA, et al. Developing a Canadian artificial intelligence medical curriculum using a Delphi study. NPJ Digit Med 2024 Nov 18;7(1):323. [doi: 10.1038/s41746-024-01307-1] [Medline: 39557985]
16. Weidener L, Fischer M. Artificial intelligence in medicine: cross-sectional study among medical students on application, education, and ethical aspects. JMIR Med Educ 2024 Jan 5;10(1):e51247. [doi: 10.2196/51247] [Medline: 38180787]
17. AI competency framework for students. UNESCO. 2024. URL: https://unesdoc.unesco.org/ark:/48223/pf0000391105 [accessed 2025-02-01]
18. Mennella C, Maniscalco U, De Pietro G, Esposito M. Ethical and regulatory challenges of AI technologies in healthcare: a narrative review. Heliyon 2024 Feb 29;10(4):e26297. [doi: 10.1016/j.heliyon.2024.e26297] [Medline: 38384518]
19. Nasir S, Khan RA, Bai S. Ethical framework for harnessing the power of AI in healthcare and beyond. IEEE Access 2024;12:31014-31035. [doi: 10.1109/ACCESS.2024.3369912]
20. Reddy S, Allan S, Coghlan S, Cooper P. A governance model for the application of AI in health care. J Am Med Inform Assoc 2020 Mar 1;27(3):491-497. [doi: 10.1093/jamia/ocz192] [Medline: 31682262]

XSL•FO
RenderX

21. Chen A, Liu L, Zhu T. Advancing the democratization of generative artificial intelligence in healthcare: a narrative review. J Hosp Manag Health Policy 2024 Jan 1;8:12-12. [doi: 10.21037/jhmhp-24-54]
22. Moulaei K, Yadegari A, Baharestani M, Farzanbakhsh S, Sabet B, Reza Afrash M. Generative artificial intelligence in healthcare: a scoping review on benefits, challenges and applications. Int J Med Inform 2024 Aug;188:105474. [doi: 10.1016/j.ijmedinf.2024.105474] [Medline: 38733640]
23. Zhang P, Kamel Boulos MN. Generative AI in medicine and healthcare: promises, opportunities and challenges. Future Internet 2023;15(9):286. [doi: 10.3390/fi15090286]
24. Annapureddy R, Fornaroli A, Gatica-Perez D. Generative AI literacy: twelve defining competencies. Digit Gov Res Pract 2025 Mar 31;6(1):1-21. [doi: 10.1145/3685680]
25. Hu R, Fan KY, Pandey P, et al. Insights from teaching artificial intelligence to medical students in Canada. Commun Med (Lond) 2022;2(1):63. [doi: 10.1038/s43856-022-00125-4] [Medline: 35668847]
26. Lindqwister AL, Hassanpour S, Lewis PJ, Sin JM. AI-RADS: an artificial intelligence curriculum for residents. Acad Radiol 2021 Dec;28(12):1810-1816. [doi: 10.1016/j.acra.2020.09.017] [Medline: 33071185]
27. AI for healthcare. NUS. 2024. URL: https://medicine.nus.edu.sg/continuing-education/course-catalogue/ai-for-healthcare [accessed 2024-01-20]
28. Oftring ZS, Deutsch K, Tolks D, Jungmann F, Kuhn S. Novel blended learning on artificial intelligence for medical students: qualitative interview study. JMIR Med Educ 2025 May 26;11:e65220. [doi: 10.2196/65220] [Medline: 40418795]
29. Fundamentals of machine learning for healthcare. Stanford. URL: https://online.stanford.edu/courses/som-xche0010-fundamentals-machine-learning-healthcare [accessed 2024-01-20]
30. Rincón EHH, Jimenez D, Aguilar LAC, Flórez JMP, Tapia Á, Peñuela CLJ. Mapping the use of artificial intelligence in medical education: a scoping review. BMC Med Educ 2025 Apr 12;25(1):526. [doi: 10.1186/s12909-025-07089-8] [Medline: 40221725]
31. Tolentino R, Baradaran A, Gore G, Pluye P, Abbasgholizadeh-Rahimi S. Curriculum frameworks and educational programs in AI for medical students, residents, and practicing physicians: scoping review. JMIR Med Educ 2024 Jul 18;10(1):e54793. [doi: 10.2196/54793] [Medline: 39023999]
32. Yao J. Research on the application risk and path of artificial intelligence in general education. Adv Educ 2025;15(8):44-51. [doi: 10.12677/ae.2025.1581400]
33. Cai C, Duell J, Chen DM, et al. Advancing AI literacy in medical education: a medical AI competency framework development. Presented at: Proceedings of the International Conference on Artificial Intelligence in Education; Jul 22-26, 2025. [doi: 10.1007/978-3-031-98462-4_15]
34. Miller GE. The assessment of clinical skills/competence/performance. Acad Med 1990 Sep;65(9 Suppl):S63-S67. [doi: 10.1097/00001888-199009000-00045] [Medline: 2400509]
35. Paranjape K, Schinkel M, Nannan Panday R, Car J, Nanayakkara P. Introducing artificial intelligence training in medical education. JMIR Med Educ 2019 Dec 3;5(2):e16048. [doi: 10.2196/16048] [Medline: 31793895]
36. Pupic N, Ghaffari-Zadeh A, Hu R, et al. An evidence-based approach to artificial intelligence education for medical students: a systematic review. PLOS Digit Health 2023 Nov;2(11):e0000255. [doi: 10.1371/journal.pdig.0000255] [Medline: 38011214]
37. Abràmoff MD, Tarver ME, Loyo-Berrios N, et al. Considerations for addressing bias in artificial intelligence for health equity. NPJ Digit Med 2023 Sep 12;6(1):170. [doi: 10.1038/s41746-023-00913-9] [Medline: 37700029]
38. Thomasian NM, Eickhoff C, Adashi EY. Advancing health equity with artificial intelligence. J Public Health Policy 2021 Dec;42(4):602-611. [doi: 10.1057/s41271-021-00319-5] [Medline: 34811466]
39. Sim JH, Abdul Aziz YF, Mansor A, Vijayananthan A, Foong CC, Vadivelu J. Students' performance in the different clinical skills assessed in OSCE: what does it reveal? Med Educ Online 2015;20(1):26185. [doi: 10.3402/meo.v20.26185] [Medline: 25697602]
40. Witheridge A, Ferns G, Scott-Smith W. Revisiting Miller's pyramid in medical education: the gap between traditional assessment and diagnostic reasoning. Int J Med Educ 2019 Oct 25;10:191-192. [doi: 10.5116/ijme.5d9b.0c37] [Medline: 31655795]
41. Emam KE. Benchmarking Kappa: interrater agreement in software process assessments. Empir Softw Eng 1999 Jun;4(2):113-133. [doi: 10.1023/A:1009820201126]
42. Abid A, Murugan A, Banerjee I, Purkayastha S, Trivedi H, Gichoya J. AI education for fourth-year medical students: two-year experience of a web-based, self-guided curriculum and mixed methods study. JMIR Med Educ 2024 Feb 20;10:e46500. [doi: 10.2196/46500] [Medline: 38376896]
43. Huang CJ, Wu T, Lu JT, et al. Developing a medical artificial intelligence course for high school students. Presented at: International Forum on Medical Imaging in Asia 2021; Jan 24-27, 2021; Taipei, Taiwan p. 103-108. [doi: 10.1117/12.2590769]
44. Laupichler MC, Hadizadeh DR, Wintergerst MWM, et al. Effect of a flipped classroom course to foster medical students' AI literacy with a focus on medical imaging: a single group pre-and post-test study. BMC Med Educ 2022 Nov 18;22(1):803. [doi: 10.1186/s12909-022-03866-x] [Medline: 36397110]

XSL•FO
RenderX

45. Yao J, Li M, Yang B. Blended learning into database courses: a path to integrating online and offline resources. Presented at: 2025 16th International Conference on E-Education, E-Business, E-Management and E-Learning (IC4e); Jan 17-19, 2025; Tokyo, Japan p. 377-381. [doi: 10.1109/IC4e65071.2025.11075421]
46. Kolb DA. Experiential Learning: Experience as the Source of Learning and Development: Prentice Hall; 1984.
47. Gibbs G. Learning by Doing: A Guide to Teaching and Learning Methods: Further Education Unit; 1988.
48. Turoff M. The Delphi Method: Techniques and Applications: Addison-Wesley; 1975.
49. Shariff NJ. Utilizing the Delphi survey approach: a review. J Nurs Care 2015;04(3):2167-1168. [doi: 10.4172/2167-1168.1000246]
50. Reeves G, Jauch LR. Curriculum development through Delphi. Res High Educ 1978;8(2):157-168. [doi: 10.1007/BF00992116]
51. Sitlington H, Coetzer A. Using the Delphi technique to support curriculum development. Educ + Train 2015 Apr 13;57(3):306-321. [doi: 10.1108/ET-02-2014-0010]
52. York CS, Ertmer PA. Towards an understanding of instructional design heuristics: an exploratory Delphi study. Educ Technol Res Dev 2011 Dec;59(6):841-863. [doi: 10.1007/s11423-011-9209-2]
53. Taylor E. We agree, don't we? The Delphi method for health environments research. HERD 2020 Jan;13(1):11-23. [doi: 10.1177/1937586719887709] [Medline: 31887097]
54. Zheng J, Lou L, Xie Y, et al. Model construction of medical endoscope service evaluation system-based on the analysis of Delphi method. BMC Health Serv Res 2020 Jul 9;20(1):629. [doi: 10.1186/s12913-020-05486-x] [Medline: 32646429]
55. Banayan J, Blood A, Park YS, Shahul S, Scavone BM. A modified Delphi method to create a scoring system for assessing team performance during maternal cardiopulmonary arrest. Hypertens Pregnancy 2015;34(3):314-331. [doi: 10.3109/10641955.2015.1033926] [Medline: 25954824]
56. Eubank BH, Mohtadi NG, Lafave MR, et al. Using the modified Delphi method to establish clinical consensus for the diagnosis and treatment of patients with rotator cuff pathology. BMC Med Res Methodol 2016 May 20;16:56. [doi: 10.1186/s12874-016-0165-8] [Medline: 27206853]
57. Jacob E, Duffield C, Jacob D. A protocol for the development of a critical thinking assessment tool for nurses using a Delphi technique. J Adv Nurs 2017 Aug;73(8):1982-1988. [doi: 10.1111/jan.13306] [Medline: 28329414]
58. Hou YF, Liu D, Zhou SG, et al. The capacity for psychotherapy process scale: development and psychometric validation of patient self-reports. Psychother Res 2021 Feb;31(2):236-246. [doi: 10.1080/10503307.2020.1769218] [Medline: 32496976]
59. Franklin AE, Burns P, Lee CS. Psychometric testing on the NLN student satisfaction and self-confidence in learning, simulation design scale, and educational practices questionnaire using a sample of pre-licensure novice nurses. Nurse Educ Today 2014 Oct;34(10):1298-1304. [doi: 10.1016/j.nedt.2014.06.011] [Medline: 25066650]
60. Reeve J, Tseng CM. Agency as a fourth aspect of students' engagement during learning activities. Contemp Educ Psychol 2011 Oct;36(4):257-267. [doi: 10.1016/j.cedpsych.2011.05.002]
61. Braun V, Clarke V. Using thematic analysis in psychology. Qual Res Psychol 2006 Jan;3(2):77-101. [doi: 10.1191/1478088706qp063oa]
62. Hu Y. Toward retrieval-grounded evaluation for conversational large language model-based risk assessment. JMIR AI 2026 Mar 12;5:e90759. [doi: 10.2196/90759] [Medline: 41818631]
63. Yam KC, Tan T, Jackson JC, Shariff A, Gray K. Cultural differences in people's reactions and applications of robots, algorithms, and artificial intelligence. Manag Organ Rev 2023 Oct;19(5):859-875. [doi: 10.1017/mor.2023.21]
64. Introduction to the New World Kirkpatrick® model. : Kirkpatrick Partners; 2015 URL: https://www.kirkpatrickpartners.com/wp-content/uploads/2021/11/Introduction-to-The-New-World-Kirkpatrick®-Model-1.pdf [accessed 2026-05-06]

Abbreviations

AI: artificial intelligence
AS: agreement score
FS: full score
GenAI: generative artificial intelligence
HIPAA: Health Insurance Portability and Accountability Act
UNESCO: United Nations Educational, Scientific and Cultural Organization

XSL•FO
RenderX

Edited by A Stone; submitted 09.Jan.2026; peer-reviewed by M Rohaim, M Roni, Y Hu; revised version received 20.Apr.2026; accepted 22.Apr.2026; published 20.May.2026.

Please cite as:
Cai C, Zhu G, Zhou SM, Ng O, Duell J, Ho WK, Chen DM, Lee B, Li F, Liu S, Sudarshan V, Wang LR, Choi C, Fan X
A Competency Framework for Medical AI Education: Mixed Methods Study
JMIR Med Educ 2026;12:e91116
URL: https://mededu.jmir.org/2026/1/e91116
doi:10.2196/91116

© Chang Cai, Gaoxia Zhu, Shang-Ming Zhou, Olivia Ng, Jamie Duell, Weng Kin Ho, Daisy Minghui Chen, Bernett Lee, Fang Li, Siyuan Liu, Vidya Sudarshan, Li Rong Wang, Chanwoo Choi, Xiuiyi Fan. Originally published in JMIR Medical Education (https://mededu.jmir.org), 20.May.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Education, is properly cited. The complete bibliographic information, a link to the original publication on https://mededu.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Systemic Barriers to Curriculum Adaptation for Rapidly Changing Knowledge in Medical Education: Qualitative Study

Martin Baumgartner[1], MSc, MLS; Matthias Stadler[2], MD; Matthias Samwald[3], PhD; Martin R Fischer[1,2], MD, MME

[1]Teaching Center, Medical University of Vienna, Spitalgasse 23, Vienna, Austria

[2]Institute of Medical Education, LMU University Hospital, LMU Munich, Munich, Germany

[3]Institute of Artificial Intelligence, Center for Medical Data Science, Medical University of Vienna, Vienna, Austria

Corresponding Author:
Martin Baumgartner, MSc, MLS
Teaching Center, Medical University of Vienna, Spitalgasse 23, Vienna, Austria

Abstract

Background: For this study, digitalization in medicine was used as an illustrative case to investigate how medical curricula respond to rapidly changing knowledge. Digitalization is transforming the way medicine and health care are provided and experienced. Experts have suggested various topics for medical curricula to keep pace with rapidly evolving knowledge. However, adapting these curricula remains a lengthy process that often lacks an interdisciplinary approach.

Objective: The perspectives of curriculum governing bodies and the boards responsible for curriculum operations at 2 medical universities were examined regarding the need for curriculum changes to account for digitalization in medicine, as well as the difficulties in adapting the curriculum to ever-growing knowledge. We identify and suggest ways to achieve more agile curriculum development.

Methods: This study consists of a qualitative analysis of governing policy frameworks and a qualitative study involving 14 video interviews. The interviews were performed with members of university curriculum governing bodies and the boards responsible for curriculum operations.

Results: All participants agreed that digitalization will reshape the medical profession by reducing physical contact, enhancing data-driven communication, and streamlining administrative processes. They highlighted the need for graduates to acquire digital literacy, critical evaluation skills, and a basic understanding of data and statistics. Yet, despite being designed as an integrated program, most participants noted that curricula have become fragmented over time due to a lack of coordination between curriculum modules. Furthermore, current processes lead to a siloed perspective, where limited coordination between modules makes it difficult to implement new knowledge holistically. This lack of intermodule alignment emerged as a key barrier to coherent curricular change. Learning objectives were identified as a promising but underutilized tool for monitoring content, aligning modules, and ensuring that emerging topics such as digitalization are integrated consistently.

Conclusions: Most participants agreed that current processes for monitoring and updating curricula are not efficiently designed, tending to be too static and focusing primarily on the advancement of subject-specific medical knowledge. To prepare current and future students for a rapidly changing world, curriculum processes should evolve from static, fragmented structures to more agile, integrated systems. By mapping the survey results to the curriculum development frameworks of Kern and Harden, we find that the challenge lies not so much in adding new content, but rather in designing curriculum processes that achieve a holistic overview. Strengthening the use of learning objectives as a dynamic monitoring and alignment tool offers a concrete opportunity to integrate rapidly changing knowledge holistically.

(JMIR Med Educ 2026;12:e96244) doi:10.2196/96244

KEYWORDS

curriculum; artificial intelligence; program development; digitalization; curriculum governance; medical education

Introduction

The European Commission emphasizes the pivotal role of universities in the “Europe of knowledge” [1]. However, the scientific literature on the development, monitoring, and quality assurance of curricula has so far only insufficiently addressed how quickly disruptive knowledge should be integrated into university curricula [2,3]. In the medical field, where safety and quality are of high importance, there is hardly any literature on this topic. The goal of our study was to better understand how quickly the medical curriculum can and should adapt to rapidly changing knowledge, using digitalization in medicine as an example of a presumably rapidly evolving domain [4].

XSL•FO
RenderX

This research is grounded in the curriculum development frameworks of Kern [5] and Harden [6]. Kern addresses curriculum monitoring as part of an iterative approach to curriculum development [5], which begins by identifying a health problem or societal need that the curriculum should address. In the "Goals and Objectives" step, broad educational goals and measurable learning objectives are defined. At the program level, these broad goals ensure that the various modules are aligned with overarching educational needs. The evaluation step assesses the achievement of goals and learning objectives for both students and the program itself, allowing for continuous improvement after each course iteration.

Harden [7] pioneered the concept of curriculum mapping in medical education. In his model, overarching learning outcomes serve as thematic anchors under which specific learning objectives are organized. The resulting curriculum map enables the identification of gaps and overlaps while visualizing dependencies between learning outcomes. The concepts of Kern and Harden are partly intertwined. Kern is more relevant to the development and renewal of specific courses, while Harden takes a strategic view of an entire program. Consequently, reviewing a curriculum map is a high-level strategic process that often involves a curriculum committee and longer review cycles.

Since the COVID-19 pandemic, attention to digital applications in medicine has increased exponentially, and this trend continues [8-11]. This has resulted in a growing number of publications calling for additional content within existing medical curricula [12-14]. Numerous studies focus on introducing new elective courses [15-17] to provide undergraduate medical students with digital competencies. However, because digitalization in medicine lacks a single "home" specialty that could serve as a natural anchor for structured content development, these initiatives tend to emphasize current technologies and specialty-specific applications, such as medical imaging in radiology or dermatology [18,19] or digital therapeutics in psychiatry [20]. Sumner et al [21] argue that "the lack of consensus in the literature on digital health content, skills and pedagogies for medical curricula may lead to overloaded frameworks with excessive content, hindering implementation." A coordinated, broad foundation, including general topics such as telemedicine, electronic health records, or administrative applications such as automated history taking, is currently missing.

In addition, many of these initiatives are developed within existing, rather static curricular frameworks based on current knowledge. No general approach has been developed to ensure the long-term relevance of these curricula. In particular, the fundamental question of how quickly new, disruptive content can be identified and systematically integrated into medical curricula remains unaddressed.

Governance structures may differ across countries, but the underlying challenge of integrating fast-moving knowledge into traditionally slow-changing curricula appears to be universal. Therefore, this paper addresses 2 research questions. First, it examines the intentions of government and governing bodies regarding curriculum monitoring and the implementation of new content through a qualitative analysis of the relevant policy framework. Second, it explores current curricular adaptation practices in Austria by examining 2 medical universities. A qualitative study was performed at the largest public medical university in Austria, the Medical University of Vienna (MUW), with around 8900 students, and the smallest public medical university, the Medical University of Innsbruck (MUI), with around 3600 students. Both MUW and MUI emerged as independent universities from their respective medical faculties (at the University of Vienna and the Leopold-Franzens University of Innsbruck, respectively) in 2004.

To contextualize and interpret our findings, the results of this study were compared with the Kern 6-Step Approach to curriculum development [5] and the Harden concepts of the spiral curriculum and curriculum integration [6,22,23].

Methods

Method 1: Qualitative Policy Framework Analysis

To address the first research question, we identified the requirements for curriculum monitoring and change processes within the policy frameworks governing Austrian medical universities. We conducted a qualitative analysis of the most recent, publicly available higher education policy documents. The sample included national legislative frameworks retrieved from the Federal Legal Information System (RIS database), university-specific statutes, as well as quality audit reports and white papers sourced from the universities' websites. A complete list of the included documents is provided in Table S1 in Multimedia Appendix 1.

The documents were systematically searched using the keywords "monitoring," "quality," "curriculum," and "learning objectives." Data analysis followed the qualitative content analysis framework proposed by Kuckartz and Rädiker [24]. The national policy and university-specific documents provided a foundational, saturated perspective, which was further triangulated with insights from quality reports and white papers. To ensure interpretive rigor, the analysis was performed by a researcher holding a master's degree in business law.

Method 2: Qualitative Study of the Curriculum Change Process

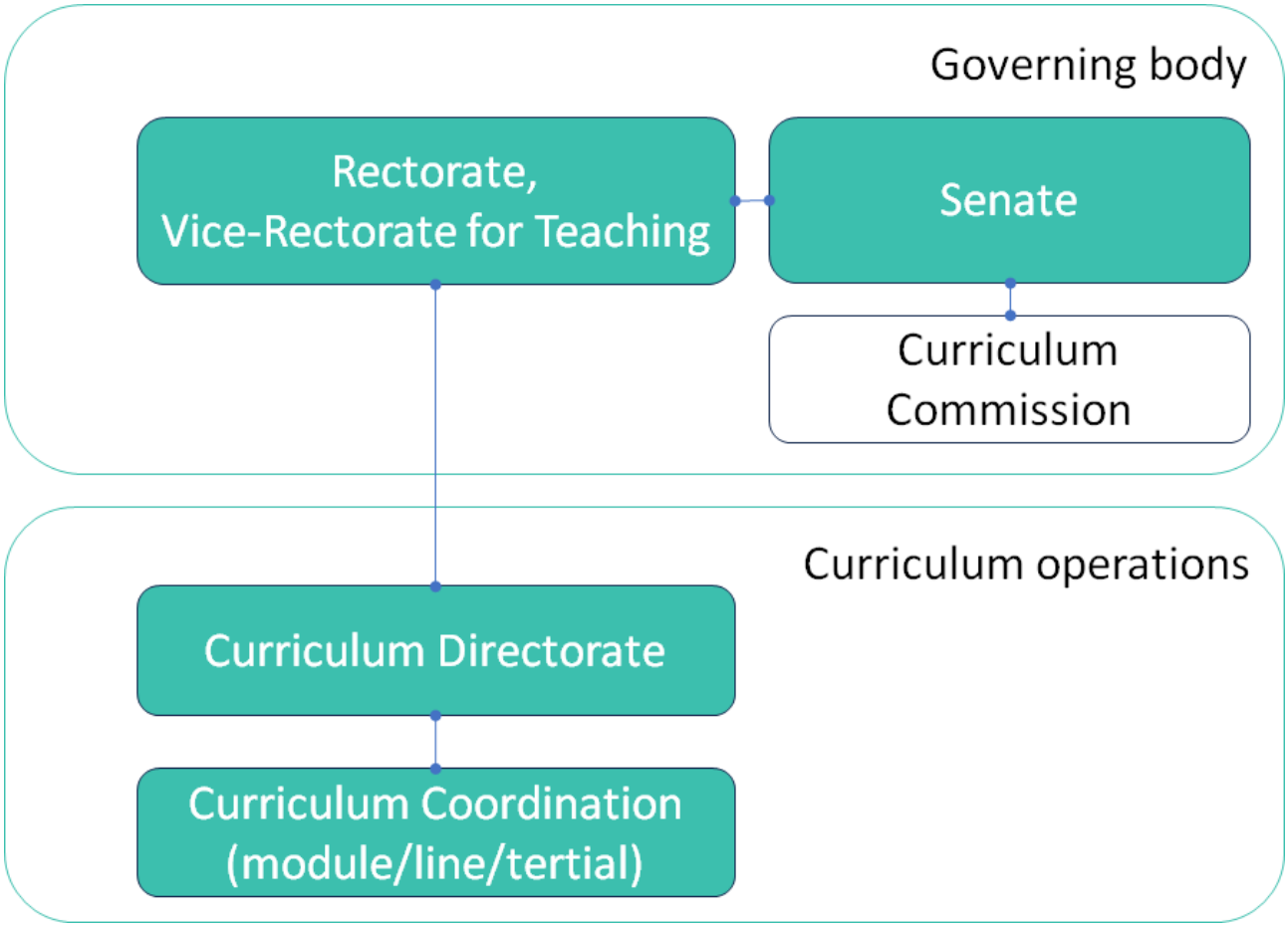
For the second research question, the COREQ (Consolidated Criteria for Reporting Qualitative Research) checklist was applied [25].

Inclusion Criteria and Recruitment

Inclusion criteria targeted academic and clinical staff from MUW and MUI with significant involvement in the planning and operation of the medical curriculum. Specifically, individuals serving on either the curriculum governing body or the curriculum operations boards were eligible for participation (Figures 1 and 2). Other stakeholders, such as students, policymakers, and administrative staff, were excluded because they typically lacked in-depth knowledge in both areas: digitalization and curricular processes.

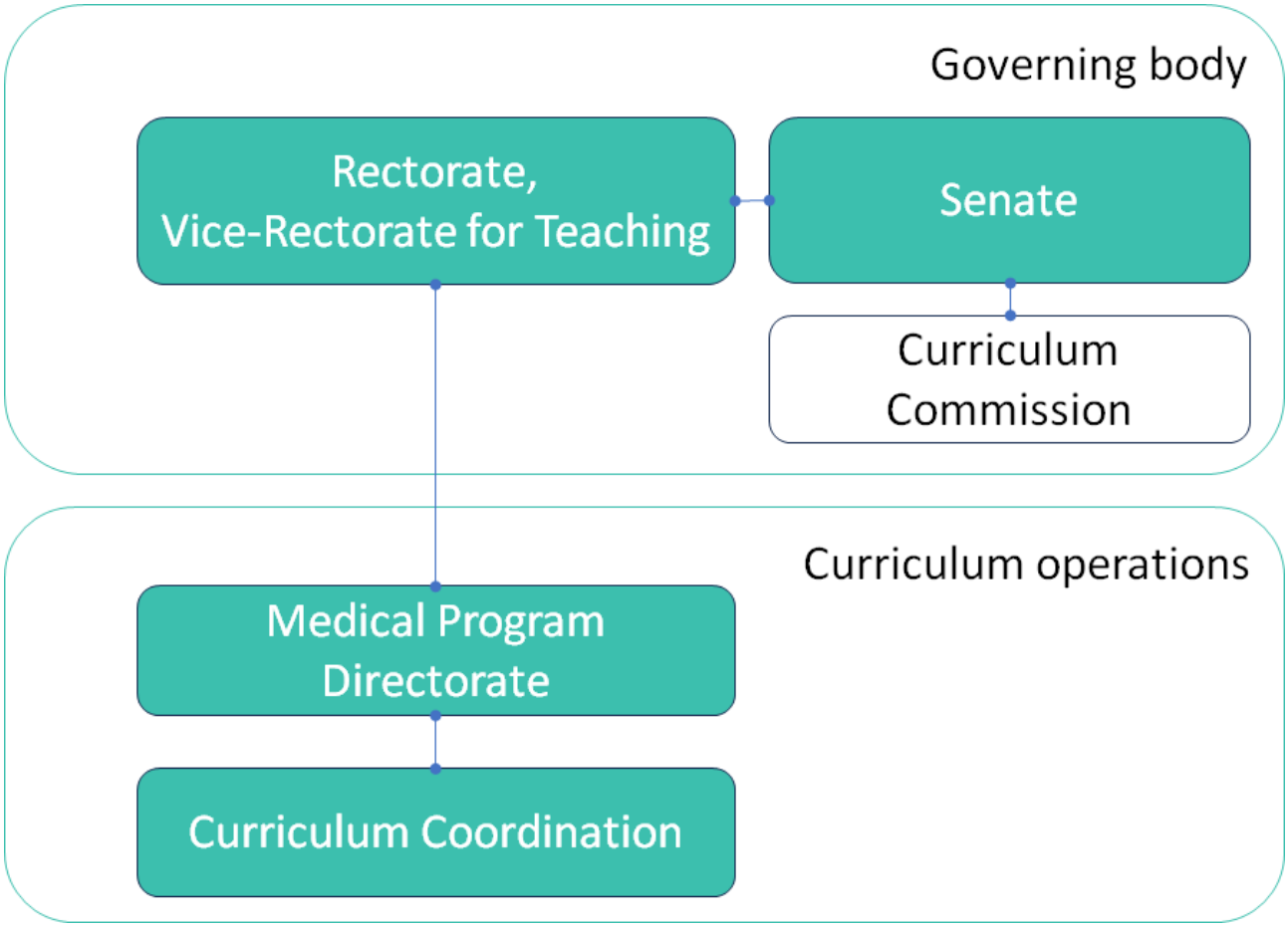
XSL•FO
RenderX

Figure 1. Curriculum process responsibilities at the Medical University of Vienna.



XSL•FO
RenderX

Figure 2. Curriculum process responsibilities at the Medical University of Innsbruck.



A modified snowball sampling approach was used for recruitment. Senior key individuals with long-standing experience in curriculum development and operations were identified and asked to nominate additional potential participants. To ensure a broad range of perspectives, the final sample was balanced across university governing bodies versus boards responsible for curriculum operations (Figures 1 and 2), as well as preclinical versus clinical teaching. Gender was also taken into account to ensure a balanced sample.

Study Design

A qualitative study was conducted based on semistructured interviews with the following 4 key questions:

- How are the demands placed on doctors changing due to the possibilities of digitalization?
- What are the potential future scenarios for the medical profession over the next 10 to 20 years?
- What processes and challenges are involved in updating the curriculum?
- Necessity and speed of curriculum updating: how dynamic does a curriculum need to be?

The semistructured interview guide was developed based on previous studies by Baumgartner et al [12,26] and Car et al [27] to gain a more in-depth understanding of both formal and informal processes involved in curriculum updating.

Data Collection

Before use, the interview guide was tested with a member of the research team who serves as Vice Dean for Clinical Education in Human Medicine at a university not involved in the study. All interviews were conducted in German using the video conferencing tool Webex (version 45.9.0.33069; Cisco Systems), utilizing its built-in transcription function. Individual video meetings were scheduled between each participant and the researcher. Interviews were planned to last approximately 50 minutes, beginning with the participant's professional profile followed by the interview questions. Data analysis followed the framework proposed by Kuckartz and Rädiker [24]. As described in the framework, data collection and data analysis are interwoven. Thus, a first draft of inductive categories was developed and refined alongside the interviews.

All interviews were recorded and automatically transcribed. While data saturation was reached after 12 participants, a total of 14 interviews were conducted (7 per university) to satisfy the study's preapproved sampling protocol.

Data Analysis

The researcher who conducted the interviews also reviewed the automatic transcription against the video recordings; spelling errors and unclear words were corrected. The profiles of the interviewees were saved separately, and all data were anonymized. Each interviewee's name was replaced with "Interviewee" plus a randomly assigned number and the abbreviation of the university (MUW for Medical University

XSL•FO
RenderX

of Vienna and MUI for Medical University of Innsbruck). Text passages by the interviewer were marked with "Moderator." To improve readability, repetitions and filler words (eg, "um" and "hm") were removed.

For data management and analysis, the transcripts were loaded into the qualitative data analysis software ATLAS.ti for Windows (version 25.0.1; ATLAS.ti Scientific Software Development GmbH). The anonymized transcripts were reread, and the inductive categories were developed (Table S2 in Multimedia Appendix 2). Based on these categories, case summaries [24] were created for 3 randomly selected interviews. A second rater, not involved in the study, independently analyzed the interviews and developed additional case summaries. These summaries were compared, and discrepancies were discussed and resolved before the remaining case summaries were produced. Finally, all case summaries were consolidated and reorganized according to the established categories.

Ethical Considerations

The ethics committee of the MUW deemed this study exempt from requiring a formal decision. Approval was granted by the data security board of the MUW. All participants received a formal invitation letter by email, including detailed information about the study purpose and a declaration of informed consent. Participation was voluntary.

Results

Qualitative Policy Framework Analysis for the Curriculum Change Process

To understand the policy framework of the universities for curriculum monitoring and change processes, the relevant regulations were analyzed. Austrian universities are regulated by the Universities Act (Universitätsgesetz [UG]) [28], the Higher Education Quality Assurance Act (Hochschul-Qualitätssicherungsgesetz) [29], and the Austrian University Development Plan (Gesamtösterreichischer Entwicklungsplan) [30]. Specifically, § 2 para 1 UG references Art 17 of the Basic Law on the General Rights of Citizens, which guarantees the "Freedom of sciences and teaching" [31]. According to the common interpretation, this freedom encompasses both the content to be taught and the manner of teaching. §14 UG [28] defines the need for a university internal quality management system and continuous evaluations, while § 2a (2) Hochschul-Qualitätssicherungsgesetz [29] emphasizes that teaching must comply with the current state of knowledge within each respective discipline. Although these requirements remain relatively broad, they structurally favor siloed curriculum models where educational content is primarily viewed from the perspective of individual subject areas.

Based on these documents, the medical universities developed their institutional frameworks. Overall, the MUW documents primarily focus on ensuring that scientific lecturers align their teaching with current scientific knowledge [32]. However, certain peripheral annotations suggest an awareness of structural challenges. The White Paper on Teaching at the MUW [33] cites an external audit report recommending "Increased coordination of curriculum content and learning objectives." The MUI emphasizes within their documents "Curricula are continuously updated and adapted to current development" [34,35]. They acknowledge "Intensity and speed of knowledge acquisition will continue to increase" and that "Innovation will be much more strongly orientated towards the boundaries between the established disciplines than before" [36].

Consequently, the institutional documents reveal distinct strategic orientations between the institutions. The MUI states its commitment to a proactive approach by regularly monitoring curricular content to adapt to rapidly changing knowledge and to strengthen intracurricular alignment [35]. The MUW follows a more reactive approach that relies on students and lecturer evaluations while focusing on the knowledge development within specific medical specialties [37].

Overall, the policy framework provides relatively loose guidelines for the universities. While it mandates a quality management system, it specifies neither the relevant topics and key performance indicators nor the concrete curricular renewal processes. Crucially, this lack of a rigid top-down prescription presents a distinct opportunity for universities, granting them the institutional autonomy required to develop their frameworks toward an agile curriculum change process. However, despite this flexibility, the institutional frameworks of universities remain predominantly focused on specialty-specific improvements, treating general curricular adaptability merely as a secondary consideration. These policy findings serve as background information for the development of the interview guide.

Qualitative Study of the Curriculum Change Process

The interviews were conducted between March and July 2025 and lasted an average of 53:45 (SD 14:37) minutes. The participants included 7 participants from the MUW and 7 participants from the MUI. Seven participants identified as female and 7 as male. Twelve participants held a habilitation (the postdoctoral lecture qualification "venia legendi"), and 2 were medical residents. The participants were also balanced according to the following characteristics: university governing body versus boards for curriculum operations and preclinical teaching versus clinical teaching. To ensure anonymity, the profiles were not described in more detail. The key themes were identified and summarized as follows: society's evolving needs for the health care system, impact on the doctor-patient relationship, curriculum governance and adaptation process, dynamics and challenges of curriculum adaptation, and current usage of learning objectives.

Society's Evolving Needs for the Health Care System

Impact of Policies on Health Care Delivery

In 2023, the health care policy "digital before outpatient before inpatient" [38] was introduced as a strategic directive for the Austrian health care system. To date, the Austrian health care system does not have a formal gatekeeping mechanism. This means patients can freely choose whether they want to see a general practitioner, a specialist, or a hospital directly. The introduction of "digital before outpatient before inpatient" is

XSL•FO
RenderX

not just a step toward a digital patient journey; it is also a first step toward a gatekeeper system.

> *So, I believe that this is also a re-education process somewhere in society, that you cannot just go to the university hospital and then get everything, but that there are a few places beforehand.* [Interviewee3_MUI]

All participants were positive about the new strategy; they acknowledged that the implementation of digital first will reduce the waiting time for patients and improve the management of patient flow for specialists and hospitals. However, some participants noted that they missed a detailed implementation plan.

Impact of Policies on Medical Education

The introduction of the policy "digital before outpatient before inpatient" will have implications for doctors and for medical education. Patients will be much more efficiently triaged to the right specialist and health care level. There will be a need to use telemedicine, digital monitoring, remote patient management, and integrated health care platforms. The data from the nationwide electronic health record system, which covers 97% of the population and is currently viewed primarily as a data repository, will enable a holistic approach to treatment.

> *It will also have an impact on the job profile; I am one hundred per cent sure of that. In my opinion, we do not have it in the curriculum yet.* [Interviewee3_MUI]

Most participants confirmed that this will have an impact on the medical profession but also recognized that it is not yet considered in current medical training.

Impact on the Doctor-Patient Relationship

Changing Communication Needs

Participants were unanimously sure that digitalization will influence communication between doctor and patient. They mentioned that physical contact will decrease and that the importance of dialogue will increase. Some participants were concerned that reduced physical contact could negatively affect the doctor-patient relationship.

> *As there is less contact, it becomes more impersonal. As there is less time available, there may be less trust between doctor and patient.* [Interviewee7_MUI]

Some participants raised concerns regarding attempts to replace interpersonal medical interaction for patients with mental illnesses with digital triage systems or algorithmic symptom assessments. A few participants mentioned that digitalization would help to visualize medical diagnoses and improve the quality of communication.

> *If artificial intelligence can create a visualization for me …, then I can have a much better, more focused conversation.* [Interviewee4_MUW]

Most participants emphasized that, due to reduced physical contact with patients, communication skills are becoming even more important for doctors. Furthermore, patients will confront doctors with information from internet search engines, artificial intelligence (AI) tools, and TikTok influencers, which makes communication even more challenging.

> *Doctors will be confronted with many more questions, as patients have already formed an opinion via Google and ChatGPT. This will take up more time again. This will change medical communication; medical students need to be prepared for this to be a positive thing.* [Interviewee6_MUW]

Several participants discussed the growing role of digital communication channels such as video consultations, messaging systems, and remote consultations. While these tools were considered useful supplements, they stressed that digital communication cannot fully replace direct physician-patient interaction. At the same time, some participants described increasing communication challenges, including unrealistic expectations regarding physician availability, immediate digital responses, and constant accessibility.

About half of the participants mentioned that they are afraid that current students are too focused on diagnostic tools and are less capable of conducting physical examinations. This concern was consistent with the general view that the teaching of existing core medical competencies must be ensured.

> *The biggest challenge for doctors will be not to lose touch with the patient. I think patient-orientation and working with patients is what students are least familiar with. The students are already very apparatus-orientated and use multidisciplinary clarification at an early stage before they even speak to the patient.* [Interviewee11_MUI]

Impact of Telemonitoring and Wearables on the Doctor-Patient Relationship

Most participants highlighted the potential of patient monitoring and the resulting diverse communication needs. This includes the use of data from medically certified applications, data from nonmedically certified applications, and digitally implemented questionnaires (eg, quality-of-life questionnaires).

> *… patients do not send the data from their wearables to the software provider, but to their doctor, who then proactively recognizes changes and calls the patient in or makes a video consultation.* [Interviewee6_MUW]

These tools were perceived as particularly useful for chronic disease management, preventive medicine, and long-term follow-up care, provided that the collected data are discussed with the patient to initiate clinical or behavioral changes.

Impact on the Administrative Burden for Doctors

One of the strongest consensual themes concerned the expectation that digitalization should reduce administrative workload. All participants repeatedly criticized inefficient documentation systems, duplicate data entry, fragmented hospital information systems, and excessive bureaucracy.

> *That preliminary results and questionnaires can be uploaded by the patient and automatically*

XSL•FO
RenderX

> *summarized in preparation for the first appointment.* [Interviewee6_MUW]
>
> *We do the patient information by phone; a doctor sits there and asks the questions and listens to the YES or NO. You could also clarify the information with a good video.* [Interviewee2_MUW]

Participants observed many options, such as uploading preliminary diagnostic findings via the internet, and emphasized the capabilities of chatbots to perform anamneses, video-supported patient information, coding of diagnoses, and the preparation of doctor's letters. For internal processes, the participants hoped for a seamless integration of the different platforms.

Curriculum Governance and Adaptation Process

Description of Current Curriculum Change Processes

The curriculum of both universities follows an integrated approach [39]. The organization around the development and operations of the curriculum differs. The MUW participants described that the governing body consists of the rectorate and the curriculum commission, a board of the senate. In addition, the Curriculum Directorate has a strategic operational focus, as also described in the statutes (section III, para 4 (1) 2) [40]. This includes "creating guidelines for coordinating curriculum content and specifying the learning and training objectives of each curriculum based on suggestions from curriculum coordinators," whereby curriculum coordinators coordinate a specific curriculum module. Most participants described the curriculum adaptation processes as highly complex and politically negotiated. There is a common agreement that changes are first discussed in the Curriculum Directorate, where proposals for changes are prepared. These are then submitted to the Vice-Rectorate for Teaching and the Rectorate. Finally, the Senate votes on the changes and approves them. Some participants stated the Vice-Rectorate for Teaching can intervene, but this is always seen as invasive and aggressive.

The MUI participants stated that they do not have a Curriculum Directorate. Instead, their Medical Program Directorate focuses on monitoring educational quality and introducing recommendations for future curriculum optimization. As a result, the Curriculum Commission and the Rectorate are much more closely involved in the operational strategic development of the curriculum. All MUI participants agreed that the Vice-Rectorate for Teaching very often develops new initiatives. For major structural changes, the Curriculum Commission determines which subjects are included at the degree program level. Additionally, some participants mentioned that a special educational advisory board, consisting of international experts, has been established to strengthen strategic capabilities.

Initiators of Curriculum Changes

Most MUW participants described that at MUW, great emphasis is placed on "freedom of teaching," and successful curriculum innovation often depends on motivated individuals rather than systematic institutional processes. Unfortunately, this bottom-up approach also leads to several participants emphasizing a lack of central strategic coordination.

> *Changes to the curriculum are primarily initiated by teachers and tertial or module coordinators, who try to keep improving the curriculum.* [Interviewee12_MUW]

If new subjects emerge, there will be an extensive discussion about the reallocation of curriculum space. Some participants stated that the Curriculum Directorate will coordinate the discussion, and influence from the governing body will be seen as intrusive. This decision process can last 1 to 2 years.

Most MUI participants described that the Vice-Rectorate for Teaching very often initiates curriculum changes and cooperates intensively with the Curriculum Commission. The resulting top-down approach is appreciated by survey participants. Thereby, the Curriculum Commission determines which subjects are included at the degree program level and, together with the Vice-Rectorate of Teaching, also recommends the redistribution of curriculum hours.

This is also reflected in participants' statements:

> *There is very good cooperation between the Vice-Rector for Teaching and the Curriculum Commission. Many of the activities we undertake originate from the Vice-Rector for Teaching.* [Interviewee9_MUI]
>
> *When it comes to redistributing hours, it would really require changes from above.* [Interviewee11_MUI]

Aligning New Content Within the Existing Curriculum

Most participants from both universities indicated that new and emerging technologies within their specialty are identified by the module coordinator and lecturers and implemented into the existing curriculum space. The module coordinator ensures that the content from the different lecturers is aligned within that module. However, all participants agreed that there is no formalized intermodule coordination process in place to ensure alignment between the various modules. It is up to the personal interests of the specific lecturer to check the slides of adjacent lectures.

> *Well, I have never experienced anyone telling me to look. If I do not sit down and look to see if there are slides from other lecturers, then it will not happen.* [Interviewee6_MUW]

Especially, the MUW participants mentioned that it would be difficult to discuss overlaps, gaps, or outdated content with lecturers from other modules. Some MUI participants added that there is an open end-of-semester meeting with students to discuss whether they experienced any overlap, gap, or outdated content. The findings suggest that the comprehensive longitudinal integration of digitalization remains organizationally difficult.

Dynamics and Challenges of Curriculum Adaptation

Need to Include Digitalization Within the Curriculum

All participants agreed that digitalization should be part of the medical curriculum, but they noted that the current curriculum insufficiently prepares medical students for the digital transformation. Most participants also repeatedly observed that

XSL•FO
RenderX

students are often less digitally competent than assumed, particularly regarding critical AI literacy. Many participants view elective courses as a starting point but favor longitudinal implementation across all specialties.

All participants also agreed that, for undergraduate students, the focus should be on learning how to use current digital applications in conjunction with critical thinking to distinguish reliable from unreliable information. Students should also develop a basic understanding of statistics and the meaning of "data" to better understand the strengths and weaknesses of digital tools. In this context, some participants mentioned that the underlying data for digital tools are not always adapted to the local population profile and local medical guidelines.

The participants also unanimously agreed that IT basics, working with big data, data management, and pattern recognition with AI should be trained in postgraduate and specialty-specific courses.

The main consensus among participants was that the medical curriculum needs to ensure that doctors have a broad overview of digital tools, can advise patients on finding the right tool, and can explain the results and their implications for further treatment.

> *If the patient comes to the doctor and the doctor says: I am sorry, I have never seen that before, then it will not go down well. And if he has no idea about digital tools, then he will not be able to use them for treatment.* [Interviewee10_MUW]

Challenges for Curriculum Changes

Several participants explained that the formal curriculum primarily regulates structural aspects such as the number of teaching hours, module organization, assessment structures, and semester distribution. Most participants recognized that the process of change within the curriculum is slow. At the MUW, the reprioritization of existing curriculum subjects emerges as the most prominent stumbling block. There are no clear processes for reprioritizing curriculum content; instead, it is a matter of give and take. As a result, newly emerging subjects often lead to long-lasting political discussions. At the MUI, the reprioritization process is guided by the Vice-Rector for Teaching and the Curriculum Commission and is therefore more efficient. Several participants from both universities criticized the absence of systematic longitudinal coordination across modules and semesters.

There is a general tendency across several participants to believe that curriculum changes do not necessarily have to follow the state of digital development quickly; rather, it makes sense to wait until the development stabilizes.

> *I do not believe that curriculum changes necessarily have to follow the state of digital development quickly. Things change so quickly in the digital world that it makes sense to take a little time to see where the journey is really going. A few things sprinkled in at the beginning certainly cannot hurt. I think the curriculum is flexible enough for that.* [Interviewee11_MUI]

This approach of waiting for robust evidence also delays the implementation of rapidly evolving knowledge. In society, this is perceived as a desperate clinging to old ways of behaving, which directly contradicts the previously stated requirement that doctors need to be able to advise the patient on the appropriate tool for treatment or prevention.

Current Usage of Learning Objectives

Formal Use of Learning Objectives

All participants confirmed that both universities have formally established learning objectives as a framework for the curriculum. Learning objectives are required for the approval of a new lecture but remain an inconsistently operationalized tool. Some participants view them more as a bureaucratic necessity and questioned whether learning objectives are systematically reviewed, updated, or evaluated over time.

> *For the main lectures, the learning objectives are kept relatively general. We have not spelled them out in detail. It is probably quite reasonable to reconsider what the average student should know.* [Interviewee3_MUI]

In this context, several participants referred to the evaluation process. They reported that evaluations focus primarily on organizational and didactic aspects rather than the actual alignment between teaching activities and intended learning outcomes. The findings suggest that learning objectives often function as formal accreditation requirements rather than dynamic curriculum–steering instruments.

Learning Objectives as Curriculum Monitoring and Coordination Tools

All participants unanimously described deficits in systematic, curriculum-wide coordination of learning objectives. Neither university uses learning objectives to monitor the curriculum or to provide a holistic overview. Consequently, coordination between modules frequently depends on informal communication and individual initiative. As a result, several participants highlighted redundancies across modules, missing competencies, and a lack of monitoring regarding the longitudinal progression of content.

> [There is no evaluation and monitoring of the curriculum in terms of learning objectives and integrative character. Interviewee8_MUW]

> *Learning objectives are an absolute must for us; they are always present. How they are monitored is a good question. I do not know.* [Interviewee1_MUI]

Several participants view central, learning objective–based monitoring as an opportunity to elevate discussions to a more substantive level and to achieve faster curriculum adaptation. At the same time, participants noted tensions between standardization and academic freedom. Some participants expressed concerns that stronger central regulation of learning objectives might restrict lecturers' autonomy. Overall, the findings indicate that learning objectives are currently not utilized as active curriculum monitoring and coordination instruments.

XSL•FO
RenderX

Discussion

Findings

Digitalization in medicine is a relatively new field that has not emerged from traditional fields of medicine but rather from society's need to streamline processes in the health care system and optimize health care spending. Furthermore, it does not affect only one medical specialty but is a cross-cutting technology. The implementation of such cross-cutting and fast-evolving content challenges existing decision-making and implementation processes in medical universities. Structures originally designed to accommodate the gradual growth of specialty-specific knowledge are increasingly overwhelmed by the influx of new, externally driven developments [41].

This study uses digitalization in medicine as an example to examine how efficiently existing processes for curriculum change and renewal can cope with the current pace of change in technology, information and data dissemination, and health care. To determine the current status, the results of a qualitative study with faculty members at 2 Austrian medical universities were compared with the curriculum development frameworks of Kern [5] and Harden [6]. Both frameworks stress the importance of clearly defined educational goals, learning objectives, and learning outcomes to hold medical educators accountable and to provide inputs for reforming and managing medical education [5,23]. Both also stress the importance of monitoring: monitoring the existing curriculum, as well as monitoring the continuous development of medical science, technology, and society's needs. Kern [5] uses learning objectives to assess whether students achieve the expected outcomes, whereas Harden [7] organizes learning outcomes into curriculum maps to identify overlaps and gaps.

Study participants acknowledged the significant impact of digitalization on the medical profession. In addition to new tools that will change the way doctors do their work, digitalization is expected to transform health care delivery and the doctor-patient relationship. However, participants also agreed that these changes are not yet systematically reflected in the current curriculum.

The universities examined mainly rely on processes tailored to gradual, discipline-specific curriculum development. MUW's bottom-up approach encourages incremental improvements and aligns well with past periods of slow curricular change. Hamsal [42] described this organizational behavior as the Icarus Paradox. In addition, MUW places a high value on teaching autonomy. This emphasis on academic freedom inadvertently reinforces resistance to reform, directly enabling the phenomenon of "curriculum hoarding" described by Romanelli [43] (see also Bearman et al [44]). In contrast, MUI applies a combined top-down and bottom-up approach, allowing for somewhat more flexible responses to external developments. Nevertheless, both universities depend primarily on external triggers for curricular change rather than on systematic internal review mechanisms.

Existing curriculum monitoring is established with a focus on student satisfaction and their learning experiences. While such evaluations fulfill formal requirements [28], they do not provide sufficient data to improve the organization of curricular content or the overall curriculum structure. Participants reported that learning objectives are mandatory for the approval of new modules; however, no procedures exist to monitor them thereafter. Although both universities originally developed integrated curricula and adopted learning objectives as a guiding principle, the lack of structured monitoring has led to growing gaps and overlaps over time. In slowly evolving environments, such shifts may pose limited issues. However, in fields characterized by rapid innovation, these inconsistencies can lead to curriculum fragmentation, as described by Harden [22], and to integrated modules without an integrated curriculum.

Although some module coordinators actively use learning objectives within their responsibilities, participants agreed that the absence of formal intermodule coordination contributes to divergence between modules. Without consistently monitoring learning objectives and learning outcomes, the university relies solely on student assessments to verify whether overarching educational goals are achieved. Consequently, curriculum mapping [7], which would make learning objectives and their interdependencies visible across the curriculum, is not implemented at the universities examined. Instead of a structured, evidence-informed discussion guided by shared priorities, the integration of new cross-cutting technologies, such as digitalization, becomes an exhausting negotiation for space in a packed curriculum.

In his work "The Integration Ladder," Harden [22] discusses the importance of making interdependencies between modules visible and understandable. Visualizing these interdependencies helps to maintain curricular coherence and ensures that content is aligned with the overall educational goals. It also clarifies which modules need to be informed or adapted when changes are introduced elsewhere in the curriculum. Existing models illustrate potential approaches: for example, Theurich et al [45] compared the learning objectives of their curriculum against a national catalog to validate curricular relevance. Naylor and Torres [46] used student surveys to assess the implementation of learning objectives in 2 courses. In contrast, participants from both universities reported intensive intramodule coordination but no formal process for inter-module coordination. Participants are even reluctant to approach colleagues, as this could be perceived as intrusive.

This study shows that curriculum processes suitable for slow-evolving, specialty-specific development are not able to keep pace with today's changes in technology, information and data dissemination, and health care delivery [47]. This mismatch generates a widening gap between current medical training and the competencies required by future health care systems. To reduce this gap, medical universities need to transform their curriculum governance and operational organization [48-51]. This transformation needs to address the tension between academic freedom and coordinated curriculum development [52]. While lecturers should maintain autonomy over teaching methods and disciplinary depth, aligning their content with overarching program competencies must become a centralized function. Therefore, the study indicates that modern curriculum governance must balance decentralized expertise with

XSL•FO
RenderX

centralized coordination mechanisms capable of maintaining curricular coherence.

To operationalize this balance, the Kern needs-based curriculum development concept and the Harden systematic curriculum monitoring and integration concept provide a solid theoretical foundation. Building on these models, institutions can draw on established change management frameworks tailored to medical education. For instance, Gale and Grant (AMEE Medical Education Guide Number 10) [53] provide structured guidelines for organizational change in a medical context, while Odiaga et al [51] demonstrate how the Kotter change model can successfully drive the required cultural transformation. Implementing these frameworks requires raising curriculum planning quality by providing faculty development on learning objectives, curriculum integration, and curriculum monitoring and mapping. As a result, the curriculum development process will move from reactive, module-centered change toward proactive, program-level governance.

Regulatory bodies can further support this transition by requiring continuous, forward-looking monitoring of learning objectives and educational goals to meet the needs of society, patients, medical students, and health care providers in the 21st century.

Limitations

This study evaluates existing curriculum processes by mapping them onto the frameworks of Kern and Harden. Since the analysis is grounded in these specific frameworks, the application of alternative frameworks might reveal different perspectives on the identified challenges. Although the results represent a snapshot of curriculum processes at only 2 of the 4 Austrian public medical universities, the findings may not be fully transferable to medical universities operating under different legal, cultural, or accreditation frameworks.

Participant recruitment followed a snowball sampling approach, where senior key individuals nominated additional participants. This strategy may have introduced selection bias. Furthermore, as attitudes toward new technologies often vary with age, a potential age-related bias among participants cannot be excluded.

Because this study required in-depth knowledge of curriculum change processes, other stakeholders, such as students, policymakers, and administrative staff, were excluded. While these groups were omitted due to the study's specific focus, customized research targeting these cohorts could provide a valuable external perspective on curriculum change processes.

Finally, this study did not include an analysis of existing documented curriculum processes, teaching materials, or learning outcomes. Consequently, no conclusions can be drawn regarding the actual impact of the observed curricular structures and processes on student learning.

Conclusions

In this study, digitalization in medicine serves as an example of how existing curriculum processes are challenged by emerging topics. The challenge lies not primarily in the rapid addition of new content but in the development of agile, integrated processes that support strategically coordinated curriculum design and enable timely, future-oriented adaptation. Curriculum management driven predominantly by operational considerations risks adding new content without systematically removing less relevant material. Integrating established curriculum development frameworks into curriculum management processes can support prioritization and help maintain curricula that are both agile and future-oriented. As students already provide feedback via web-based evaluation, one simple step forward is to include the successful achievement of individual learning objectives within this evaluation. This approach would provide immediate feedback to lecturer and curriculum coordinators.

Learning objectives, in particular, provide a powerful tool for tracking curriculum content, coordinating modules, and embedding cross-cutting topics such as digitalization throughout the curriculum. In addition, strengthening mechanisms for cross-module coordination and systematic curriculum monitoring is crucial. Through such approaches, medical education can ensure that graduates are not only familiar with digital tools but also resilient, critical, and adaptable to continuous change. While governmental initiatives strongly support the development of digital skills in general, more comprehensive and legally binding guidelines for agile curriculum processes could further facilitate the integration of rapidly evolving knowledge into medical education.

Even though the study was conducted in Austria and the specific governance structures may differ across countries, the underlying challenge of integrating fast-moving knowledge into traditionally slow-changing curricula appears universal. In highly regulated systems, curriculum adaptation may be constrained by accreditation standards or national competency frameworks. In less regulated or resource-constrained systems, challenges may include limited digital infrastructure, faculty expertise, or institutional capacity. Here, structured learning objectives provide relatively low-cost mechanisms for alignment, prioritization, and continuous adaptation.

This study contributes to the growing body of knowledge on curriculum development and renewal and highlights areas for future research. Further studies should include qualitative interviews with curriculum planners to explore challenges in implementing curriculum renewal in greater depth. This research aims to provide curriculum planners with a practical toolbox for the efficient and effective management of curriculum processes, enabling even resource-limited medical degree programs to adopt a future-oriented approach. Ultimately, this field of research is essential for a holistic understanding of medical curricula and for strengthening the shared sense of responsibility among medical educators.

XSL•FO
RenderX

Acknowledgments

The authors would like to thank all supporting colleagues and survey participants for their contributions. The authors also declare that they used the generative artificial intelligence tool ChatGPT to improve the grammar and readability of the text.

Funding

The authors declared no financial support was received for this work.

Data Availability

The datasets generated and analyzed during this study are not publicly available due to obligations related to data security.

Authors' Contributions

Conceptualization: MB, MRF, M Stadler, M Samwald
Investigation: MB
Methodology: MB, MRF, M Stadler, M Samwald
Project administration: MB, MRF
Supervision: MB, MRF
Validation: MB, researcher not involved in the project
Writing – original draft: MB
Writing – review and editing: MB, MRF, M Stadler, M Samwald

Conflicts of Interest

M Stadler holds associate editor role for *JMIR Medical Education* at the time of this publication. All other authors declare no conflicts of interest.

Multimedia Appendix 1
Analyzed university regulations.
[DOCX File, 32 KB - mededu_v12i1e96244_app1.docx]

Multimedia Appendix 2
Deductive and inductive grading categories.
[DOCX File, 34 KB - mededu_v12i1e96244_app2.docx]

References

1. The role of universities in the Europe of knowledge. European Union. 2003. URL: https://eur-lex.europa.eu/EN/legal-content/summary/the-role-of-universities-in-the-europe-of-knowledge.html [accessed 2024-09-03]
2. Bajada C, Kandlbinder P, Trayler R. A general framework for cultivating innovations in higher education curriculum. High Educ Res Dev 2019 Apr 16;38(3):465-478. [doi: 10.1080/07294360.2019.1572715]
3. Bleiklie I, Powell WW. Universities and the production of knowledge—introduction. High Educ 2005 Jan;49(1-2):1-8. [doi: 10.1007/s10734-004-2909-3]
4. Győrffy Z, Radó N, Mesko B. Digitally engaged physicians about the digital health transition. PLoS One 2020;15(9):e0238658. [doi: 10.1371/journal.pone.0238658] [Medline: 32986733]
5. Thomas PA, Kern DE, Hughes MT, Chen BY, editors. Curriculum Development for Medical Education: A Six-Step Approach, 3rd edition: Johns Hopkins University Press; 2016. URL: https://muse.jhu.edu/book/44600 [accessed 2026-06-23]
6. Harden RM. What is a spiral curriculum? Med Teach 1999;21(2):141-143. [doi: 10.1080/01421599979752] [Medline: 21275727]
7. Harden RM. AMEE Guide No. 21: curriculum mapping: a tool for transparent and authentic teaching and learning. Med Teach 2001 Mar;23(2):123-137. [doi: 10.1080/01421590120036547] [Medline: 11371288]
8. Cuff A. The evolution of digital health and its continuing challenges. BMC Digit Health 2023;1(1):3. [doi: 10.1186/s44247-022-00004-x] [Medline: 38014367]
9. Yeung AWK, Torkamani A, Butte AJ, et al. The promise of digital healthcare technologies. Front Public Health 2023;11:1196596. [doi: 10.3389/fpubh.2023.1196596] [Medline: 37822534]
10. Stoumpos AI, Kitsios F, Talias MA. Digital transformation in healthcare: technology acceptance and its applications. Int J Environ Res Public Health 2023 Feb 15;20(4):3407. [doi: 10.3390/ijerph20043407] [Medline: 36834105]
11. Odunlami WO, Li E, Greenfield G, et al. Factors influencing primary care physicians recommending patients to use digital health technologies for self-management: a cross-sectional study across 20 countries. Eur J Gen Pract 2025 Dec;31(1):2555819. [doi: 10.1080/13814788.2025.2555819] [Medline: 40948455]

XSL•FO
RenderX

12. Baumgartner M, Wagner-Menghin M, Vajda C, Lecaks G, Redzic A, Dorffner G. Analysis of the impact of a university distance learning course on digitalization in medicine on students and healthcare professionals. Wien Klin Wochenschr 2025 Jul;137(13-14):412-418. [doi: 10.1007/s00508-024-02393-7] [Medline: 38985177]
13. Veikkolainen P, Tuovinen T, Kulmala P, et al. The evolution of medical student competencies and attitudes in digital health between 2016 and 2022: comparative cross-sectional study. JMIR Med Educ 2025 Jul 31;11:e67423. [doi: 10.2196/67423] [Medline: 40743520]
14. Khurana MP, Raaschou-Pedersen DE, Kurtzhals J, Bardram JE, Ostrowski SR, Bundgaard JS. Digital health competencies in medical school education: a scoping review and Delphi method study. BMC Med Educ 2022 Feb 26;22(1):129. [doi: 10.1186/s12909-022-03163-7] [Medline: 35216611]
15. Tudor Car L, Kyaw BM, Nannan Panday RS, et al. Digital health training programs for medical students: scoping review. JMIR Med Educ 2021 Jul 21;7(3):e28275. [doi: 10.2196/28275] [Medline: 34287206]
16. Aulenkamp J, Mikuteit M, Löffler T, Schmidt J. Overview of digital health teaching courses in medical education in Germany in 2020. GMS J Med Educ 2021;38(4):Doc80. [doi: 10.3205/zma001476] [Medline: 34056069]
17. Zainal H, Xiaohui X, Thumboo J, Kok Yong F. Digital competencies for Singapore's national medical school curriculum: a qualitative study. Med Educ Online 2023 Dec;28(1):2211820. [doi: 10.1080/10872981.2023.2211820] [Medline: 37186901]
18. Lindqwister AL, Hassanpour S, Levy J, Sin JM. AI-RADS: successes and challenges of a novel artificial intelligence curriculum for radiologists across different delivery formats. Front Med Technol 2022;4:1007708. [doi: 10.3389/fmedt.2022.1007708] [Medline: 36688145]
19. Clark Lambert W, Grzybowski A. Dermatology and artificial intelligence. Clin Dermatol 2024;42(3):207-209. [doi: 10.1016/j.clindermatol.2023.12.014] [Medline: 38181889]
20. Vogelmann U, Stadler M, Soldini A, et al. A comparative analysis of technical data: at-home vs. in-clinic application of transcranial direct current stimulation in depression. Brain Stimul 2025;18(4):1097-1105. [doi: 10.1016/j.brs.2025.05.103] [Medline: 40374109]
21. Sumner B, Martin R, Gladman T, Wilkinson TJ, Grainger R. Understanding the gap: a balanced multi-perspective approach to defining essential digital health competencies for medical graduates. BMC Med Educ 2025 May 9;25(1):682. [doi: 10.1186/s12909-025-07194-8] [Medline: 40346629]
22. Harden RM. The integration ladder: a tool for curriculum planning and evaluation. Med Educ 2000 Jul;34(7):551-557. [doi: 10.1046/j.1365-2923.2000.00697.x] [Medline: 10886638]
23. Harden RM, Crosby JR, Davis MH, Friedman M. AMEE Guide No. 14: outcome-based education: part 5-from competency to meta-competency: a model for the specification of learning outcomes. Med Teach 1999 Jan;21(6):546-552. [doi: 10.1080/01421599978951] [Medline: 21281173]
24. Kuckartz U, Rädiker S. Qualitative Inhaltsanalyse Methoden, Praxis, Umsetzung Mit Software Und Künstlicher Intelligenz [Book in German], 6th edition: Beltz Juventa; 2024.
25. Tong A, Sainsbury P, Craig J. Consolidated Criteria for Reporting Qualitative Research (COREQ): a 32-item checklist for interviews and focus groups. Int J Qual Health Care 2007 Dec;19(6):349-357. [doi: 10.1093/intqhc/mzm042] [Medline: 17872937]
26. Baumgartner M, Sauer C, Blagec K, Dorffner G. Digital health understanding and preparedness of medical students: a cross-sectional study. Med Educ Online 2022 Dec;27(1):2114851. [doi: 10.1080/10872981.2022.2114851] [Medline: 36036219]
27. Car J, Ong QC, Erlikh Fox T, et al. The digital health competencies in medical education framework: an international consensus statement based on a Delphi Study. JAMA Netw Open 2025 Jan 2;8(1):e2453131. [doi: 10.1001/jamanetworkopen.2024.53131] [Medline: 39888625]
28. Bundesrecht konsolidiert: Gesamte Rechtsvorschrift für Universitätsgesetz 2002, Fassung vom 23.06.2026) [Article in German]. Rechtsinformationssystem des Bundes. URL: https://www.ris.bka.gv.at/GeltendeFassung.wxe?Abfrage=Bundesnormen&Gesetzesnummer=20002128 [accessed 2025-05-20]
29. Bundesrecht konsolidiert: gesamte rechtsvorschrift für hochschul-qualitätssicherungsgesetz, fassung vom 23.06.2026 [Article in German]. Rechtsinformationssystem des Bundes (RIS). URL: https://www.ris.bka.gv.at/GeltendeFassung.wxe?Abfrage=Bundesnormen&Gesetzesnummer=20007384 [accessed 2025-10-05]
30. Der Gesamtösterreichische Universitätsentwicklungsplan 2025-2030 [Report in German]. : Bundesministerium für Bildung, Wissenschaft und Forschung; 2023 URL: https://pubshop.bmbwf.gv.at/index.php?rex_media_type=pubshop_download&rex_media_file=guep_2025_2030_231031.pdf [accessed 2026-06-23]
31. Bundesrecht konsolidiert: Gesamte Rechtsvorschrift für Staatsgrundgesetz über die allgemeinen Rechte der Staatsbürger, Fassung vom 27.06.2026 [Article in German]. Rechtsinformationssystem des Bundes. URL: https://www.ris.bka.gv.at/GeltendeFassung.wxe?Abfrage=Bundesnormen&Gesetzesnummer=10000006 [accessed 2025-10-05]
32. Entwicklungsplan der Medizinischen Universität Wien 2025–2030 [Report in German]. : Medizinische Universität Wien; 2023 URL: https://www.meduniwien.ac.at/web/fileadmin/content/serviceeinrichtungen/rechtsabteilung/mitteilungsblaetter_2023-24/4_Mitteilungsblatt_06122023_Entwicklungsplan_MedUni_Wien_2025-2030_01.pdf [accessed 2025-07-02]

XSL•FO
RenderX

33. White paper lehre [Report in German]. : Medizinische Universität Wien; 2025 URL: https://www.meduniwien.ac.at/web/fileadmin/content/serviceeinrichtungen/studienabteilung/lehrende/MedUni_WhitePaper_Lehre.pdf [accessed 2025-05-20]
34. Wissensbilanz der Medizinischen Universität Innsbruck für das Jahr 2024 [Report in German]. : Medizinische Universität Innsbruck; 2025 URL: https://www.i-med.ac.at/mitteilungsblatt/2024/39.pdf [accessed 2025-05-26]
35. Leistungsvereinbarung 2022 - 2024 - 5. Erganzung (Abdeckung der Gehaltserhohungen im Ärztebereich an der Medizinischen Universität Innsbruck) [Report in German]. : Medizinische Universität Innsbruck; 2024 URL: https://www.i-med.ac.at/mitteilungsblatt/2024/19.pdf [accessed 2025-10-05]
36. Entwicklungsplan der Medizinischen Universität Innsbruck 2025 - 2030 [Report in German]. : Medizinische Universität Innsbruck; 2023 URL: https://www.i-med.ac.at/universitaet/docs/Entwicklungsplan_Med-Uni-IBK_2025-2030.pdf [accessed 2025-06-25]
37. Auditbericht: Medizinische Universität Wien (MedUni Wien) [Report in German]. : AHPGS Akkreditierung gGmbH; 2022 URL: https://www.meduniwien.ac.at/web/fileadmin/content/serviceeinrichtungen/evaluierung/Bericht_Med_Uni_Wien_final_August_2022.pdf [accessed 2025-06-17]
38. Vereinbarung gemäß Art. 15a B-VG über die Organisation und Finanzierung des Gesundheitswesens (BGBl. I Nr. 2/2025) [Report in German]. : Bundeskanzleramt der Republik Österreich; 2025 URL: https://www.ris.bka.gv.at/Dokumente/BgblAuth/BGBLA_2025_I_2/BGBLA_2025_I_2.pdf [accessed 2026-06-23]
39. Brauer DG, Ferguson KJ. The integrated curriculum in medical education: AMEE Guide No. 96. Med Teach 2015 Apr;37(4):312-322. [doi: 10.3109/0142159X.2014.970998] [Medline: 25319403]
40. Satzung der Medizinischen Universität Wien [Report in German]. : Medizinische Universität Wien; 2025 Feb 5 URL: https://www.meduniwien.ac.at/web/fileadmin/content/serviceeinrichtungen/rechtsabteilung/satzung/satzung.pdf [accessed 2025-07-02]
41. Thomas A, Ellaway RH. Rethinking implementation science for health professions education: a manifesto for change. Perspect Med Educ 2021 Dec;10(6):362-368. [doi: 10.1007/s40037-021-00688-3] [Medline: 34757538]
42. Hamsal M. Exploring the Icarus Paradox in Indonesia's specialist medical education system using the public perspective from online media: convergent mixed methods study. JMIR Med Educ 2026 Jan 26;12:e60452. [doi: 10.2196/60452] [Medline: 1587450]
43. Romanelli F. Curricular hoarding. Am J Pharm Educ 2020 Jan;84(1):847714. [doi: 10.5688/ajpe847714] [Medline: 32292200]
44. Bearman M, Chandir H, Mahoney P, Partridge H. Sustaining teaching and learning innovations: a scoping review. High Educ Res Dev 2024 Oct 2;43(7):1495-1510. [doi: 10.1080/07294360.2024.2364096]
45. Theurich T, Holzhausen Y, Ahlers O, Peters H. Mapping the undergraduate medical curriculum of the Charité Berlin against the National Competence-Based Catalogue of Learning Objectives (NKLM 2.0). GMS J Med Educ 2025;42(4):Doc46. [doi: 10.3205/zma001770] [Medline: 41103404]
46. Naylor KA, Torres KC. Translation of learning objectives in medical education using high-and low-fidelity simulation: learners' perspectives. J Taibah Univ Med Sci 2019 Dec;14(6):481-487. [doi: 10.1016/j.jtumed.2019.10.006] [Medline: 31908634]
47. Koechli L. From isolation to innovation: how higher ed is embracing cross-departmental synergy. The EvoLLLution. 2025 Mar 3. URL: https://evolllution.com/from-isolation-to-innovation-how-higher-ed-is-embracing-cross-departmental-synergy [accessed 2026-05-17]
48. McVitty D. Institutional silos are making it harder to build learning environments for student success. Wonkhe. 2024 Nov 12. URL: https://wonkhe.com/blogs/institutional-silos-are-making-it-harder-to-build-learning-environments-for-student-success/ [accessed 2026-05-17]
49. Operationalizing the breaking of silos in universities: a step-by-step guide. Equiss Accelerator. 2024 Jun 11. URL: https://www.equiss.org/post/operationalizing-the-breaking-of-silos-in-universities-a-step-by-step-guide [accessed 2026-05-17]
50. Operational alignment in higher ed: why departments can't work in silos anymore. EDUTECHLoft. 2026 Feb 10. URL: https://www.edutechloft.com/en/resources/blog/operational-alignment-in-higher-ed-why-departments-cant-work-in-silos-anymore/ [accessed 2026-05-17]
51. Odiaga J, Guglielmo MJ, Catrambone C, et al. Kotter's change model in higher education: transforming siloed education to a culture of interprofessionalism. J Organ Cult Commun Confl 2021 Mar;25(2):1-7 [FREE Full text]
52. Alzahmi RA, Syed RT, Singh D, Arshi TA, Kutty SV. Organizational change in higher education institutions: thematic mapping of the literature and future research agenda. Humanit Soc Sci Commun 2025 Aug 8;12(1):1282. [doi: 10.1057/s41599-025-05650-w]
53. Gale R, Grant J. AMEE Medical Education Guide No. 10: managing change in a medical context: guidelines for action. Med Teach 1997 Jan;19(4):239-249. [doi: 10.3109/01421599709034200]

Abbreviations

AI: artificial intelligence
COREQ: Consolidated Criteria for Reporting Qualitative Research
MUI: Medical University of Innsbruck

XSL•FO
RenderX

MUW: Medical University of Vienna
UG: Universitätsgesetz

Edited by W Jerjes; submitted 26.Mar.2026; peer-reviewed by F Binsar, L Silsand, L Ilodigwe; revised version received 29.May.2026; accepted 11.Jun.2026; published 15.Jul.2026.

Please cite as:
Baumgartner M, Stadler M, Samwald M, Fischer MR
Systemic Barriers to Curriculum Adaptation for Rapidly Changing Knowledge in Medical Education: Qualitative Study
JMIR Med Educ 2026;12:e96244
URL: https://mededu.jmir.org/2026/1/e96244
doi:10.2196/96244

© Martin Baumgartner, Matthias Stadler, Matthias Samwald, Martin R Fischer. Originally published in JMIR Medical Education (https://mededu.jmir.org), 15.Jul.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Education, is properly cited. The complete bibliographic information, a link to the original publication on https://mededu.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

An AI-Integrated Relational Curriculum Mapping Framework for Alignment and Transparency in Medical Education: Implementation Study

Robert M Badeau[1,2], PhD, MSc, PMI-CPMAI; Jennifer Laury[1], MA; Danielle Dickey[1,3], EdD

[1]Office of Academic Affairs, Naresh K. Vashisht College of Medicine, Texas A&M University, Bryan, TX, United States

[2]Department of Humanities in Medicine, Naresh K. Vashisht College of Medicine, Texas A&M University, Bryan, TX, United States

[3]Department of Medical Education, Naresh K. Vashisht College of Medicine, Texas A&M University, Bryan, TX, United States

Corresponding Author:
Robert M Badeau, PhD, MSc, PMI-CPMAI
Office of Academic Affairs
Naresh K. Vashisht College of Medicine
Texas A&M University
8447 John Sharp Pkwy
Bryan, TX, 77807
United States
Phone: 1 979 436 0621
Email: rmbadeau@tamu.edu

Abstract

Background: Medical school curricula are managed through fragmented, static tools that prevent faculty and curriculum committees from seeing a unified view of content scope, sequence, and licensing examination alignment. Governance relies on manual reporting that is neither timely nor consistent, and no widely adopted platform incorporates a generative AI agent that queries a relational curriculum database to produce tailored, on-demand reports.

Objective: This study aimed to design, implement, and evaluate a 4-phase, AI-integrated relational curriculum mapping framework providing faculty self-directed curriculum search; cross-course alignment reports classifying content as redundant, complementary, or a gap, with recommendations; and alignment with United States Medical Licensing Examination (USMLE) content and accreditation requirements.

Methods: We implemented the framework during academic year 2025-2026. It captures session learning objectives (SLOs) annually; maps them to program objectives, course objectives, USMLE codes, Association of American Medical Colleges keywords, and Bloom's levels using a large language model with expert verification; centralizes verified data in a relational database (Airtable) with faculty read-only search; and uses a second AI agent via an API to generate on-demand reports on scope, sequence, redundancy, gaps, and USMLE alignment across all 16 preclerkship courses and all clinical clerkship rotations.

Results: Course directors gained self-directed curriculum access through Airtable Extensions search. Cross-course vertical alignment reports identified redundant, complementary, and gap findings and generated course-level recommendations. An analysis of 2 preclerkship courses identified 50 matched SLO pairs across 3 USMLE pathology domains within minutes. A USMLE Step 1 alignment report across all 16 courses identified a curriculum-wide pathology coverage deficit (42% vs a target of 45%-55%) that was descriptively consistent with institutional performance data (not formally tested). An accreditation query identified all preclerkship and clerkship content on a specified topic and generated Data Collection Instrument-ready documentation within seconds. A Medical Education Program Objective coverage analysis across all 9 clinical clerkships, generated within minutes, found documented coverage for 24 of 25 (96%) program objectives and 1 objective with no matching SLO in any clerkship, a gap not previously tracked. In a blinded validation study, 3 independent course directors each coded a 30-SLO sample; agreement with the database (Fleiss κ, 3 raters plus database) was substantial at the exact-code level (κ=0.67, 95% CI 0.55-0.78) and almost perfect at the USMLE domain level (κ=0.94, 95% CI 0.88-1.00).

Conclusions: An AI-integrated relational curriculum database with a faculty-facing search interface can convert curriculum governance from a periodic, retrospective activity into a continuous, self-service one, producing alignment and gap analyses within minutes rather than weeks. Workload effects have not been formally measured. Extension to assessment items and prospective correlation with licensing examination outcomes are planned.

XSL•FO
RenderX

(JMIR Med Educ 2026;12:e107727) doi:10.2196/107727

KEYWORDS

curriculum mapping; AI; education, medical; learning objectives; curriculum development; databases, relational; program evaluation; faculty

Introduction

Medical school curricula encompass thousands of session learning objectives (SLOs) distributed across preclerkship and clerkship phases; yet, the tools most institutions use to manage this content have not kept pace with the complexity they must represent. Most rely on flat file formats, spreadsheets, disconnected syllabi, and separate assessment records that cannot capture the relational architecture of a modern curriculum [1,2]. The consequence is "curriculum opacity": course directors and faculty cannot readily determine what is taught in adjacent courses, at what cognitive level, or whether it is assessed [1].

The practical costs are significant. When faculty lack cross-course visibility, unintended content redundancy accumulates; course directors reteach concepts already addressed elsewhere because they have no systematic way to know otherwise. Barde et al [3] documented that medical students and faculty identified redundant theoretical content in more than 30 of every 100 surveyed topics, with redundancy directly associated with perceptions of curriculum inefficiency and reduced student engagement. Without cross-course visibility, the distinction between introducing a concept and teaching it to mastery is operationally unenforceable. Youm et al [4] proposed a 6-degree model of curriculum integration encompassing content, assessment, faculty, and community dimensions, each of which depends on a system that makes the full scope and sequence of the curriculum visible to those responsible for designing it. Without that system, integration remains aspirational.

Curriculum committees face a parallel challenge: answering governance and accreditation questions about scope and coverage requires manual analyses that take weeks and may be outdated on delivery [1,2]. Existing commercial platforms, including Elentra, Entrada, One45, and VidaNovaVLE, have improved on spreadsheet approaches, but critical limitations persist: most require significant staff time to input and maintain data. Faculty remain passive consumers of typically uninterpretable staff-generated summary reports rather than active participants in a living curriculum system [2,5]. While computational approaches such as topic modeling have demonstrated the analytical potential of curriculum data, these have been implemented as one-time research exercises rather than persistent operational infrastructure [6]. Large language models are increasingly recognized as capable of natural-language classification and synthesis tasks at a scale not previously feasible; yet, their application in medical education has concentrated on student-facing tutoring and content generation rather than institutional governance infrastructure [7]. No widely adopted platform incorporates a generative AI agent capable of faculty and instructors querying the live database and producing tailored, on-demand reports in response to specific committee queries.

Previously, we described and validated a granular, hierarchical coding library derived from the USMLE (United States Medical Licensing Examination) Content Outline, together with a proof-of-concept application mapping SLOs and examination items to that library [8]. That earlier work established the granularity and machine-readability of the coding structure but did not address institutional-scale operationalization, faculty search access, or accreditation linkage. The present study builds on that infrastructure, extending it into a complete framework: a relational database populated using the coding library, a faculty-facing search interface, and an AI-agent reporting layer generating alignment and accreditation documentation across an entire preclerkship curriculum.

Methods

Setting and Development

The Naresh K. Vashisht College of Medicine Relational Curriculum Mapping Framework was developed and implemented during academic year (AY) 2025-2026, transitioning from annual, manually assembled Microsoft Excel–based curriculum snapshots to a centralized, AI-integrated relational architecture that connects learning objectives to program outcomes in a single queryable system. The framework spans 16 preclerkship courses and all clinical clerkship rotations. It is grounded in constructive alignment theory [9]. The framework comprises 4 sequential phases (Table 1 and Figure 1).

XSL•FO
RenderX

Table 1. Framework components, tools, inputs, outputs, and Liaison Committee on Medical Education (LCME) elements addressed by each phase[a].

Phase	Tool or platform	Input	Output	LCME element
1. Session capture	Excel intake form	Session titles and SLOs[b] from course directors	Version-controlled session data; prior-year database reference distributed to CDs[c]	6.1, 8.3
2. AI-agent mapping	Claude Sonnet 5 + human expert review	SLOs, USMLE[d] code library (18 domains, step 1 and step 2) [8], AAMC[e] keywords, Bloom's levels	AI-proposed tags verified by subject-matter experts before database entry	7.1, 8.3, 8.7
3. Database and faculty access	Airtable cloud relational database	Verified SLOs, USMLE codes, AAMC keywords, Bloom's levels, CLOs[f], MEPOs[g]	Faculty read-only searchable database; Airtable Extensions search for cross-course curriculum discovery	6.3, 8.3, 8.8
4. AI-agent API reporting	Claude API connected to Airtable live database via REST[h] API	Natural language queries from curriculum committee or mapping unit	USMLE alignment dashboards, redundancy and gap reports with actionable recommendations by SLO, CLO, MEPO, and USMLE domain	6.3, 8.3, 9.9

[a]Standards cited from LCME Functions and Structure of a Medical School, 2022 edition.

[b]SLO: session learning objective.

[c]CD: course director.

[d]USMLE: United States Medical Licensing Examination.

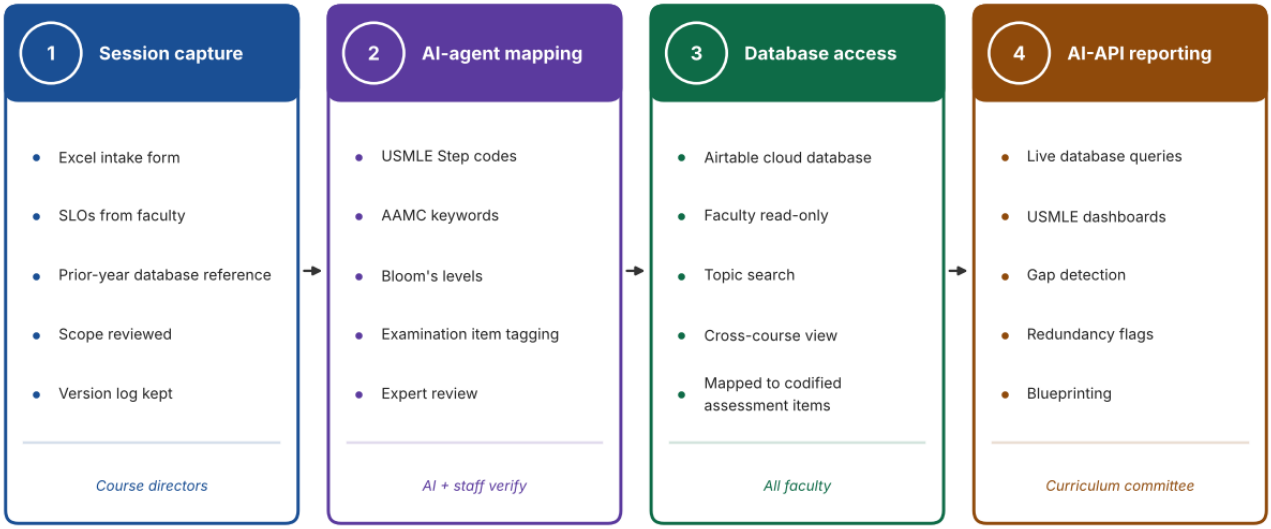
[e]AAMC: Association of American Medical Colleges.

[f]CLO: course learning objective.

[g]MEPO: Medical Education Program Objective.

[h]REST: Representational State Transfer.

Figure 1. Four-phase AI-integrated curriculum mapping workflow. Phase 1: annual session data capture via Microsoft Excel intake. Phase 2: AI-agent mapping of SLOs to USMLE STEP codes, AAMC keywords, and Bloom's Taxonomy levels with expert verification. Phase 3: centralization in the Airtable cloud relational database, with faculty read-only access, Extensions search for cross-course discovery, and preclerkship examination items mapped to codified assessment items linked to SLOs. Phase 4: AI-agent API reporting generating alignment dashboards, gap analyses, and redundancy reports by SLO, CLO, MEPO, and USMLE domain. AAMC: Association of American Medical Colleges; CLO: course learning objective; MEPO: Medical Education Program Objective; SLO: session learning objective; USMLE: United States Medical Licensing Examination.



Phase 1: Session Data Capture

Course directors submit SLOs annually via a standardized Excel intake form. Only SLOs are requested from faculty. Before finalizing SLOs for the incoming year, each course director receives read-only access to the prior AY's relational Airtable database, enabling intentional scope and sequence decisions grounded in direct visibility of what adjacent courses already teach, as well as review of the USMLE content coding, Association of American Medical Colleges (AAMC) keywords, and Bloom's Taxonomy levels assigned to their own course's SLOs from the prior cycle.

A well-functioning curriculum map also documents curricular changes systematically across AYs. Using an AI-API interface with Airtable, each annual submission cycle is version-controlled, with additions, deletions, and modifications to session sequence and SLO content logged against prior archived data. This longitudinal record supports Liaison Committee on Medical Education (LCME) Standard 8

XSL•FO
RenderX

expectations for curricular management: rather than reconstructing history from scattered records at a site visit, the curriculum committee has a continuously updated, queryable archive of how the curriculum has evolved.

Phase 2: AI-Agent Mapping

Each SLO is mapped to USMLE Content Outline codes, AAMC keywords, and Bloom's Taxonomy levels using structured, single-turn prompts submitted to Claude Sonnet 5 (Anthropic PBC; model identifier claude-sonnet-5, accessed May through September 2026) via its public API. Each prompt supplies one SLO together with the full USMLE coding library and a fixed set of classification rules and returns a single proposed code with a brief rationale; the workflow does not use retrieval-augmented generation or a multistep agentic pipeline at this stage, since the coding library is supplied directly in the prompt context rather than retrieved. Representative prompt templates, the classification rules used to select a code, and example inputs and outputs are provided in Multimedia Appendix 1. The framework is model-agnostic: any comparable large language model accessible via API can be substituted. All AI-generated tags are reviewed and verified by subject-matter experts [7].

The USMLE coding library and tagging methodology underlying this process were described and validated in the previous study reporting the construction and preclerkship application of the coding library [8].

AI-Tagging Validation

To quantify the accuracy of phase 2, a blinded validation study was conducted. A random sample of 30 SLOs was drawn across all 16 tagged courses, stripped of their AI-proposed and staff-verified USMLE codes, and distributed with only the session title and SLO text to 3 course directors with USMLE-content expertise who had not been involved in the original tagging of the sampled items. Each rater independently assigned a USMLE content code to every SLO using the same coding library, without sight of the database's existing code or of the other raters' responses. Rater-assigned codes were compared against the corresponding database codes and against one another, classified at 2 levels of granularity: an exact match (identical code) and a domain-level match (same top-level USMLE domain but a different specific subcode). Items a rater could not confidently code from the text alone were excluded from that rater's scored comparisons rather than treated as a mismatch. Agreement across the 3 raters and the database was summarized with Fleiss κ, computed separately at the exact-code and domain levels; 95% CIs were obtained by bootstrap resampling of the 30 items (3000 resamples). Pairwise Cohen κ between each pair of raters and between each rater and the database was computed as a secondary, descriptive comparison.

Phase 3: Relational Database and Faculty Access

Subject-matter expert–verified data are uploaded into Airtable, a cloud-based relational database. The framework is not limited to Airtable: any cloud-based relational database with API access can serve as the infrastructure backbone.

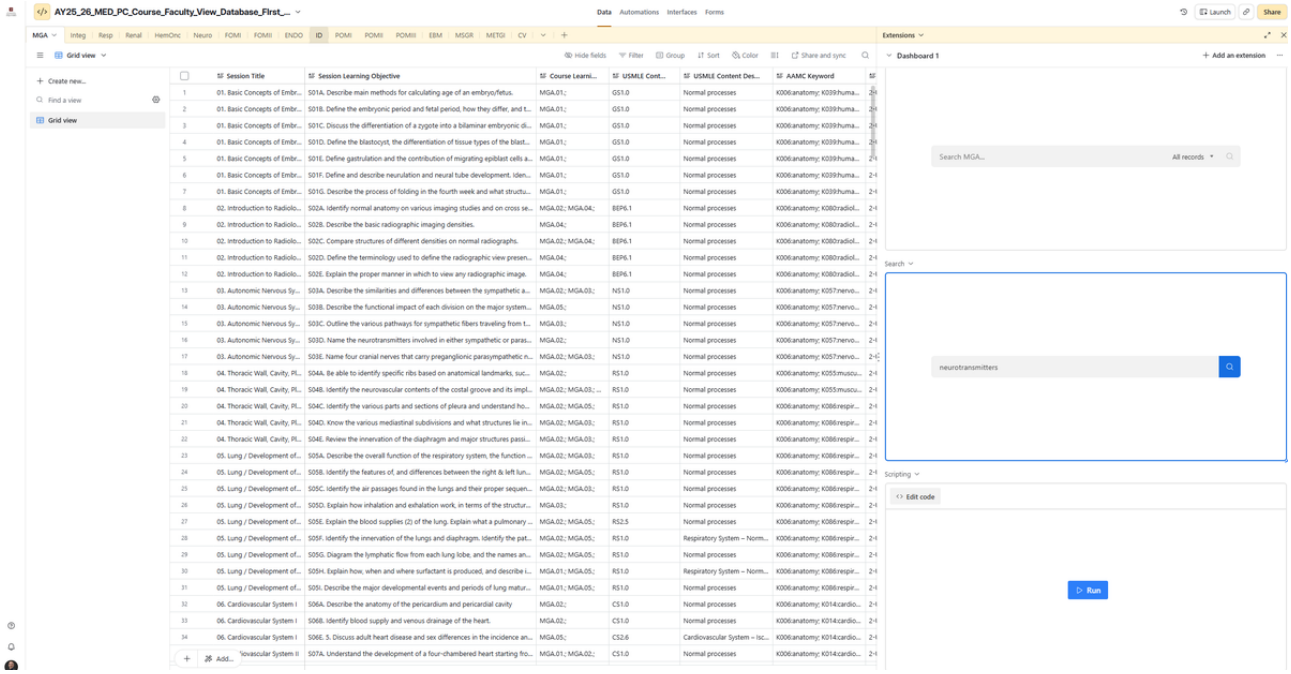
A read-only shared link is distributed to all course directors and faculty (Figures 2 and 3). The interface is deliberately simple: all 16 preclerkship courses are laid out as tabs across the top of a single screen, visible to any user immediately upon opening the database, without submenus, additional logins, or navigation beyond a single click. The key feature enabling cross-course discovery is Airtable's Extensions search, which allows any faculty member to query the entire curriculum database using free-text search and filters without technical training. A course director types a clinical concept and sees, within seconds, every SLO, CLO, Medical Education Program Objective (MEPO), and Bloom's Taxonomy level across all 16 courses addressing that concept.

The preclerkship database is mirrored by a clerkship database aligned to the clerkship core objectives, the Accreditation Council for Graduate Medical Education core competencies, clerkship rotation–specific learning objectives, institutional program objectives (MEPOs), and the USMLE Step 2 Clinical Knowledge framework (Multimedia Appendix 2). This cross-phase architecture operationalizes vertical curriculum integration [10].

The database contains curriculum content only: session titles, learning objectives, coding tags, and course-level metadata submitted by course directors. It contains no student records, grades, enrollment data, or other information covered by the Family Educational Rights and Privacy Act (FERPA), and access is accordingly governed by institutional data-use policy rather than FERPA-specific controls. Access is restricted to a read-only link distributed to faculty and curriculum staff through the institution's existing single-sign-on authentication; no public or student-facing access exists at this stage. Phase 2 tagging submits only deidentified SLO text, containing no patient or student information, to the external AI vendor's API; the vendor's standard data-retention and training-opt-out terms for API traffic apply, consistent with the institution's existing data-processing agreements for other cloud services.

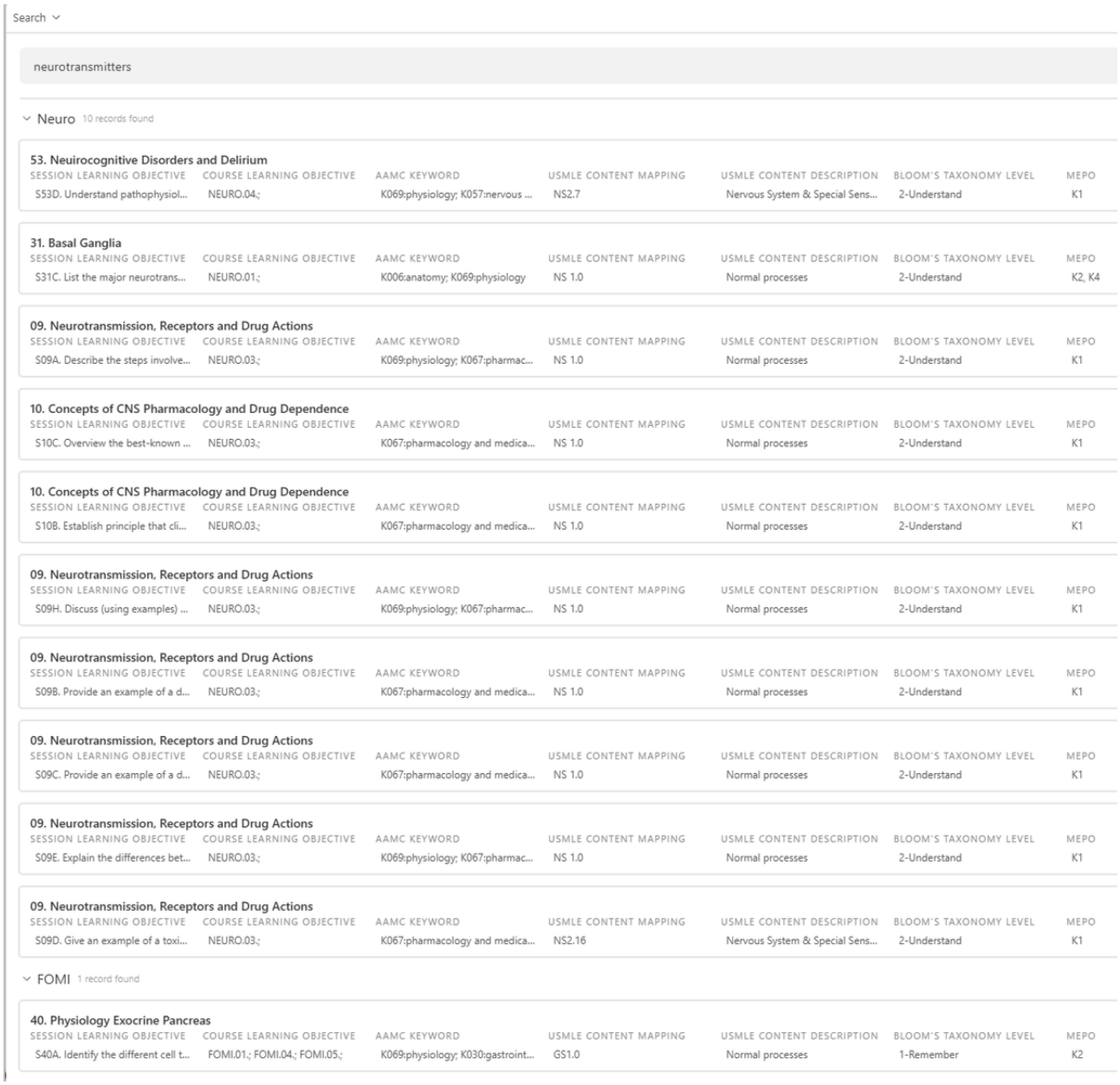
XSL•FO
RenderX

Figure 2. Screenshot of the faculty-facing read-only Airtable preclerkship curriculum database layout. All 16 preclerkship courses appear as tabs across the top of a single screen, immediately visible without scrolling or navigating a menu; the course view beneath displays session titles, session learning objectives, and mapped USMLE codes, AAMC keywords, and Bloom's Taxonomy levels for the selected course. AAMC: Association of American Medical Colleges; USMLE: United States Medical Licensing Examination.



XSL•FO
RenderX

Figure 3. Screenshot of the faculty-facing read-only Airtable preclerkship curriculum database. The result of an Extensions search for the term "neurotransmitters," returning every matching session learning objective across all 16 courses, organized by course, with the corresponding USMLE code, AAMC keyword, and Bloom's Taxonomy level displayed for each match. Faculty search across all 16 preclerkship courses by session topic, USMLE code, AAMC keyword, or Bloom's Taxonomy level without staff mediation. AAMC: Association of American Medical Colleges; USMLE: United States Medical Licensing Examination.



Phase 4: AI-Agent API Reporting

An AI agent interfaces directly with the live database via the Airtable Representational State Transfer API, accepting natural language queries from curriculum staff or the committee. This stage is a retrieval-based workflow rather than a multistep autonomous agentic pipeline: the query retrieves the relevant matched records directly from the relational database, and the retrieved records, not a separate search or planning step, are passed to Claude Sonnet 5 to be classified and written into a report. Report generation takes minutes; the complete pipeline including human expert review and committee-ready delivery is completed within minutes. No institution-level AI infrastructure is required.

A central phase 4 output is the longitudinal alignment report: a query comparing session titles and SLO text across a multicourse sequence that classifies each matched pair as redundant, complementary, or a gap, and generates a specific recommendation for each finding, such as consolidating redundant sessions, adopting cross-course numbering or cross-referencing conventions, or designating a single owning course for a given concept. Each report closes with an explicit caution that classifications involve interpretation and should be verified by course directors before curricular changes are made.

USMLE Step 1 Alignment Analysis

For each of the 16 preclerkship courses, discipline coverage percentages were calculated as the count of that course's SLOs coded to a pathology, physiology, or pharmacology USMLE

XSL•FO
RenderX

domain, divided by the total number of SLOs in that course. An SLO carrying more than 1 code was counted once toward each distinct discipline its codes mapped to, so a course's 3 discipline percentages do not necessarily sum to 100%; SLOs coded to a domain outside these 3 disciplines contributed to the denominator but not to any of the 3 numerators. Categorical status labels (on, below, above, and exceeds) were assigned by comparing each course's discipline percentage to the corresponding USMLE Step 1 Content Outline target range: below or above indicates the percentage fell outside the target range by less than 10 percentage points, and low gap or exceeds indicates a deviation of 10 percentage points or more. The composite alignment score (0-100) was calculated as 100 minus the sum, across the 3 disciplines, of each discipline's absolute percentage-point deviation from the nearest edge of its target range, floored at 0; a course exactly within all 3 target ranges receives a score of 100, and larger or more numerous deviations reduce the score proportionally. Evidence-Based Medicine and Medical Student Grand Rounds were scored against only the discipline targets applicable to their content, given their methodological and integrative focus rather than organ-system or disease content.

Ethical Considerations

Curriculum mapping and database development activities described in this manuscript constitute quality improvement and do not meet the regulatory definition of human subjects research [11].

Results

During AY 2025-2026, the framework was fully operational across all 16 preclerkship courses and all clinical clerkship rotations.

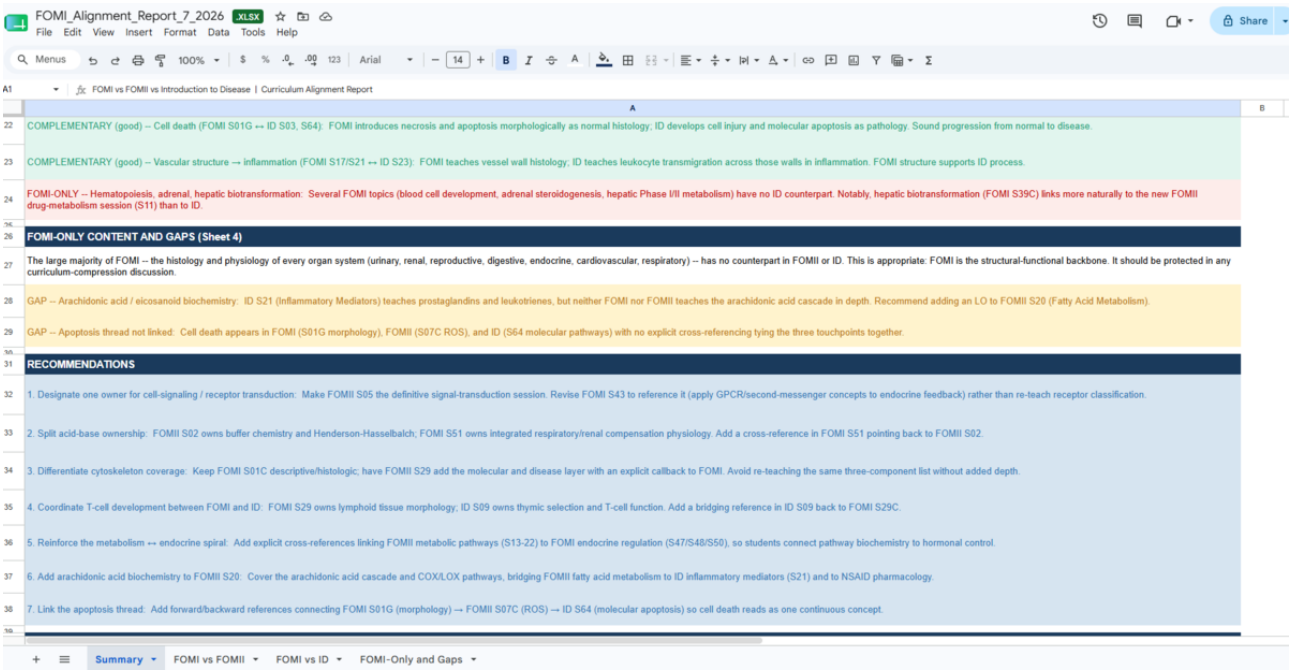
Faculty Access and Preventing Unintended Content Repetition

Read-only database access was distributed to all 16 preclerkship course directors and all clerkship discipline directors. Faculty used the Extensions search to verify topic coverage before finalizing annual SLOs, replacing informal colleague communication as the primary mechanism for cross-course coordination. Faculty also received AI-generated alignment reports classifying their SLOs against related courses as complementary, redundant, or a gap, with recommendations for revising session content to improve horizontal alignment; an example of one such report is shown in Figure 4.

The practical impact of cross-course visibility was demonstrated when the Curriculum Mapping Unit compared Foundations of Medicine I (FOMI, semester 1, AY 2025-2026) and Introduction to Disease (ID, semester 2, AY 2025-2026). A structured AI query identified 50 matched SLO pairs across 14 FOMI and 15 ID sessions spanning 3 USMLE pathology domains: cell injury and apoptosis, inflammation and repair, and neoplasia. Delivered within minutes, the report showed for the first time that FOMI's histology and cell biology sessions systematically introduce the structural substrate ID later formalizes as pathology; the FOMI cell organelles session alone connects to 10 downstream ID pathology SLOs. Course directors can now scaffold this progression explicitly rather than leaving vertical connections invisible to students, one practical response to the reteaching problem that curriculum opacity can create.

XSL•FO
RenderX

Figure 4. Screenshot of a longitudinal alignment report comparing FOMI, the revised FOMII, and Introduction to Disease, exported from the framework's database as a structured spreadsheet. Each row is labeled with 1 of 4 classifications: COMPLEMENTARY, indicating sound scaffolding from normal structure to disease process, illustrated here by FOMI's morphological introduction of cell death (necrosis and apoptosis) preceding Introduction to Disease's treatment of molecular apoptosis as pathology; GAP, identifying content taught in neither course, such as arachidonic acid and eicosanoid biochemistry; FOMI-ONLY, denoting foundational content, such as organ-system histology and physiology, that appropriately has no downstream counterpart and should be protected from curriculum compression; and RECOMMENDATIONS, a dedicated set of rows translating each finding into a specific action, such as designating FOMII as the definitive session for cell-signaling and receptor transduction with FOMI referencing rather than reteaching it or adding a learning objective to an existing session to close an identified gap. Each row includes the specific session codes involved and a brief rationale, providing course directors with an audit trail from finding to recommended action. Rows are color-coded by finding type. Green indicates a COMPLEMENTARY finding (sound scaffolding between courses), red indicates a GAP or FOMI-ONLY finding (content taught in neither course, or foundational content appropriately unique to FOMI), dark blue denotes section headers separating report categories, and unshaded rows under the RECOMMENDATIONS header list denote the specific corrective action for each numbered finding above it. FOMI: Foundations of Medicine I; FOMII: Foundations of Medicine II.



Alignment Reports With Actionable Recommendations for Course Directors

A parallel longitudinal alignment report compared the 3-course Practice of Medicine sequence (Practice of Medicine I [POMI], Practice of Medicine II [POMII], and Practice of Medicine III), finding that patient-centered interviewing was taught in overlapping form at least 5 times across POMI and POMII, and that POMI and POMII independently assign the same session number to 3 unrelated sessions, creating scheduling and assessment-tracking ambiguity distinct from content overlap. For each finding, the report generated a specific recommendation rather than a bare classification: consolidating redundant sessions, adopting a unified cross-course numbering scheme, revising physical examination objectives toward clinical reasoning rather than repeating skill-acquisition language, and adding sessions on telehealth and addiction medicine to address gaps identified through AAMC keyword tags with no corresponding taught session.

The FOMI-to-ID report similarly moved beyond classification to recommendation (Figure 4). For redundant content, such as cell-signaling receptor transduction and acid-base physiology taught in both FOMI and Foundations of Medicine II (FOMII), the report proposed designating a single owning course for each concept, with the nonowning course referencing rather than reteaching it, and for acid-base physiology, it proposed splitting ownership so that FOMII retains buffer chemistry and Henderson-Hasselbalch calculations, while FOMI develops integrated respiratory and renal compensation. For a gap in arachidonic acid and eicosanoid biochemistry, a prerequisite for inflammatory mediator content taught later in ID, it proposed a specific session and learning objective addition. A second gap, the apoptosis thread, appeared disconnected across all 3 courses: FOMI introduces its morphology, the revised FOMII covers its molecular mechanism, and ID addresses its pathological consequences, but no session in any course explicitly cross-references the other 2, and the report recommended adding forward and backward references so the 3 treatments read as one continuous concept rather than 3 unrelated topics. Course directors received the underlying matched-objective data alongside each recommendation and were advised to verify findings against current materials before making curricular changes.

USMLE Step 1 Alignment Analysis

A full USMLE Step 1 content alignment report spanning all 16 preclerkship courses was generated through a single natural language query and delivered to the curriculum committee within minutes (Table 2).

XSL•FO
RenderX

Table 2. Academic year 2025-2026 United States Medical Licensing Examination (USMLE) Step 1 discipline coverage and composite alignment scores across all 16 preclerkship courses[a].

Code	Course	Pathology (%)	Physiology (%)	Pharmacology (%)	USMLE status			Score out of 100
					Pathology	Physiology	Pharmacology	
POMII	POM II	44	30	18	On	On	On	99
METGI	Metabolism/Gastrointestinal	46	41	16	On	Above	On	99
RENAL	Renal	55	28	19	On	Below	On	97
POMIII	POM III	43	29	16	Below	Below	On	96
POMI	POM I	40	36	14	Below	On	On	93
ENDO	Endocrinology/Reproductive	51	31	28	On	On	Above	89
NEURO	Neuroscience	58	26	21	Above	Below	Above	89
FOMII	Foundations of Medicine II	38	42	12	Below	Above	On	88
RESP	Respiratory	49	52	12	On	Exceeds	On	84
MGA	Medical Gross Anatomy	37	44	6	Below	Above	Below	79
HEMEONC	Hematology/Oncology	62	22	22	Exceeds	Below	Above	77
FOMI	Foundations of Medicine I	33	48	10	Low	Above	On	73
ID	Introduction to Disease	54	14	26	On	Low gap	Above	71
INTEG	Integumentary System	68	18	10	Exceeds	Low gap	On	67
MSGR	Medical Student Grand Rounds	23	19	15	Low	Low	On	56
EBM	Evidence-Based Medicine	11	16	8	Low	Low	Below	45
AVG	All 16 courses	42	33	14	Below	On	On	—[b]
Target	USMLE Step 1 Blueprint	45-55	30-40	10-20	Reference ranges	Reference ranges	Reference ranges	—

[a]Generated via AI-agent API query against the Airtable curriculum database, May 2026. Composite alignment score (0-100) reflects simultaneous proximity to all 3 discipline targets. EBM and MSGR scored against applicable USMLE competencies only, given their methodological and integrative nature.

[b]Not applicable.

The report identified a curriculum-wide pathology coverage average of 42%, 3 percentage points below the USMLE Step 1 target of 45%-55%, with 6 of 16 courses below 40%. Physiology (33%) and pharmacology (14%) were both within target ranges. These findings were descriptively compared with institutional USMLE Step 1 performance data (National Board of Medical Examiners Step 1 Performance Profile, 2023-2025 cohorts) (unpublished institutional data, NBME Step 1 Performance Profile, 2023-2025 cohorts): a total of 34 of every 100 examinees scored below the national mean in pathology, compared with 33 of every 100 in physiology and 29 of every 100 in pharmacology. This pattern was only partly consistent with the coverage gap: pathology's below-target coverage coincided with the largest performance gap, but physiology showed a comparably large performance gap despite coverage within its target range, indicating that coverage alignment alone does not fully account for relative discipline performance. No formal statistical test of any of these associations was conducted. The report delivered prioritized, course-level recommendations alongside a composite alignment scorecard.

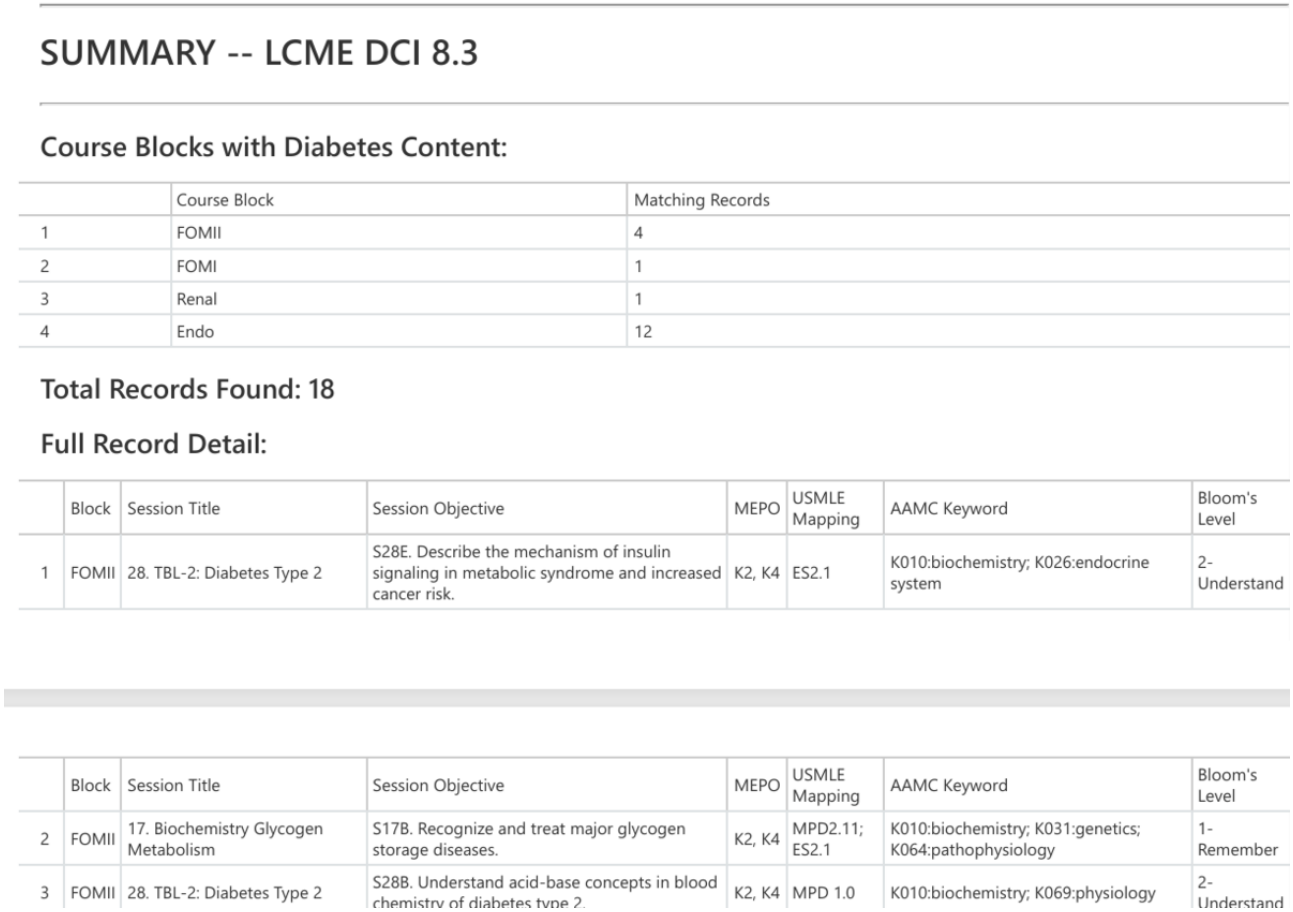
Accreditation Support

The relational database explicitly links every SLO to its parent course learning objective (CLO) and program objectives, addressing LCME Element 8.3, which requires faculty to define competencies and learning objectives for each curricular segment and link them to program objectives [12]. The read-only faculty interface evidences the LCME expectation that faculty can navigate and use the curriculum map. For Element 8.3, within seconds, staff generated a report, using a scripting extension in Airtable, to determine where a specified topic (diabetes) appeared across preclerkship and clerkship content for Data Collection Instrument reporting (Figure 5). The database supports accreditation readiness and gives the curriculum committee a tool for strategic curriculum planning: data-informed decisions about scope, sequence, and content alignment that previously required weeks of manual analysis by staff are now available rapidly and on demand, enabling the committee to identify gaps proactively. The framework was developed and is sustained without external funding or a dedicated commercial curriculum-mapping platform, relying instead on general-purpose, subscription-based tools (Airtable

XSL•FO
RenderX

and the Claude API) that are themselves proprietary but require no institution-specific licensing or custom development.

Figure 5. Example of an AI agent–generated LCME Standard 8.3 documentation report showing all preclerkship and clerkship course blocks containing diabetes-related content, including matching course blocks, record counts, and session-level detail with mapped MEPOs, USMLE codes, and AAMC keywords. Generated in seconds using an Airtable scripting extension from a single free-text query. AAMC: Association of American Medical Colleges; LCME: Liaison Committee on Medical Education; MEPO: Medical Education Program Objective; USMLE: United States Medical Licensing Examination.



SUMMARY -- LCME DCI 8.3

Course Blocks with Diabetes Content:

	Course Block	Matching Records
1	FOMII	4
2	FOMI	1
3	Renal	1
4	Endo	12

Total Records Found: 18

Full Record Detail:

	Block	Session Title	Session Objective	MEPO	USMLE Mapping	AAMC Keyword	Bloom's Level
1	FOMII	28. TBL-2: Diabetes Type 2	S28E. Describe the mechanism of insulin signaling in metabolic syndrome and increased cancer risk.	K2, K4	ES2.1	K010:biochemistry; K026:endocrine system	2-Understand
2	FOMII	17. Biochemistry Glycogen Metabolism	S17B. Recognize and treat major glycogen storage diseases.	K2, K4	MPD2.11; ES2.1	K010:biochemistry; K031:genetics; K064:pathophysiology	1-Remember
3	FOMII	28. TBL-2: Diabetes Type 2	S28B. Understand acid-base concepts in blood chemistry of diabetes type 2	K2, K4	MPD 1.0	K010:biochemistry; K069:physiology	2-Understand

Clerkship MEPO Coverage Analysis

To evaluate whether the framework's diagnostic capability extends to the clerkship phase, an AI-agent query generated a full coverage report tracing all 25 institutional MEPOs against SLOs across all 9 clinical clerkships (emergency medicine, internal medicine, family medicine, pediatrics, psychiatry, radiology, neurology, obstetrics and gynecology, and surgery). The report was generated within minutes from the relational clerkship curriculum database (Table 3).

Of the 25 MEPOs evaluated, 24 (96%) had documented SLO coverage in at least 1 clerkship, and 20 (80%) were covered across 5 or more of the 9 clerkships, indicating broad, distributed reinforcement across the clinical curriculum. In total, 1 (4%) MEPO had zero matching SLOs identified in any clerkship, a coverage gap that had not been previously identified through the institution's existing reporting processes. A total of 3 (12%) additional MEPOs were matched to SLOs in only a single clerkship, indicating narrow rather than absent coverage that may still warrant curricular attention. This analysis was limited to SLOs; assessment items had not yet been mapped to MEPOs in the clerkship database at the time of this report, consistent with the assessment-tagging extension described in the Future Directions section.

XSL•FO
RenderX

Table 3. Clerkship Medical Education Program Objective (MEPO) coverage report across all 9 clinical clerkships, generated via AI-agent query against the Airtable clerkship curriculum database, July 2026[a].

MEPO code	Matching clerkship SLOs[b], n	Clerkships covering (of 9)	Status
K1	97	9	Covered
K2	1	1	Single-clerkship only[c]
K3	97	9	Covered
K4	1	1	Single-clerkship only
K5	13	9	Covered
S1	16	8	Covered
S2	26	8	Covered
S3	14	8	Covered
S4	1	1	Single-clerkship only
S5	52	9	Covered
S6	97	9	Covered
S7	32	9	Covered
S8	13	9	Covered
S9	103	9	Covered
S10	22	7	Covered
S11	6	5	Covered
S12	11	9	Covered
S13	2	2	Covered
S14	6	5	Covered
AB1	23	9	Covered
AB2	10	8	Covered
AB3	0	0	Gap—no coverage[d]
AB5	11	9	Covered
AB6	13	9	Covered
AB7	22	9	Covered

[a]Full matched SLO codes by clerkship are available from the corresponding author.

[b]SLO: session learning objective.

[c]"Single-clerkship only" indicates the MEPO was matched to SLOs in exactly 1 of 9 clerkships.

[d]"Gap—no coverage" indicates no matching SLO was identified in any clerkship.

AI-Tagging Validation

Three independent course directors completed the blinded coding task; the number of items an individual rater excluded as not confidently codable from the SLO text alone ranged from 0 to 3 of 30. Agreement across the 3 raters and the database, summarized with Fleiss κ, was substantial at the exact-code level (κ=0.67, 95% CI 0.55-0.78) and almost perfect at the domain level (κ=0.94, 95% CI 0.88-1.00); the domain-level CI did not cross below the almost-perfect threshold at either bound. Pairwise comparison showed that this aggregate figure reflected 2 raters converging closely with each other and with the database (exact-code: κ=0.82-0.89; domain-level: κ=0.96-1.00) and a third rater whose exact-code agreement with the other 2 was lower (κ=0.47-0.55) despite comparable domain-level agreement (κ≥0.88), consistent with that rater's own report of difficulty completing portions of the task. Review of the residual exact-code disagreements showed most reflected genuine differences in expert framing at the subcode level for content plausibly spanning 2 adjacent categories (eg, whether a laboratory-test-interpretation objective is coded to the organ-system domain or to the biostatistics and epidemiology domain covering test characteristics) rather than database error; one item, on which the database initially diverged from all 3 raters, was corrected to match the converging rater consensus. This blinded, independent comparison is distinct from, and addresses the limitation of, the routine phase 2 verification described in the Methods section, which is performed by curriculum staff rather than independent raters and was not separately quantified.

XSL•FO
RenderX

Discussion

Principal Findings

These findings offer preliminary evidence that AI-integrated relational architecture can convert curriculum governance from a periodic, retrospective activity into a continuous, self-service one. The FOMI-to-ID vertical alignment analysis and the 16-course USMLE Step 1 report demonstrate that questions requiring days to weeks of manual staff effort can be answered within a single working day, and that resulting reports are specific enough to inform course-level action, including concrete recommendations such as session consolidation and content-ownership assignment, rather than generic classifications alone. The blinded validation study also offers an independent answer to a question the framework's design left open: how closely AI-proposed tags track subject-matter expert judgment once verified. Almost-perfect domain-level agreement (Fleiss κ=0.94) across 3 independent course directors and the database suggests the phase 2 workflow is not merely producing plausible-looking output that staff rubber-stamp, but output that substantively tracks how subject-matter experts would classify the same material; substantial, rather than almost-perfect, exact-code agreement (κ=0.67) is consistent with genuine ambiguity in content that spans 2 adjacent domains rather than with systematic AI error, a pattern reinforced by the fact that the 2 most consistent raters agreed with each other about as often as either agreed with the database. The narrower panel size (3 raters) means these κ estimates, while their CIs did not cross into lower agreement categories, should be interpreted as an initial rather than definitive estimate of population-level agreement. Regardless of this level of agreement, we want to be explicit that these findings are not a basis for reducing human review of AI-proposed codes. The validation study measures how closely a verified database matches independent expert judgment; it does not establish that any individual AI-proposed code can be trusted without the course-director confirmation step described in the Methods section, since even the raters who most closely agreed with the database and each other still diverged on a meaningful minority of items, and the courses represented in the discrepant items were not predictable in advance. We view the verification step as a permanent feature of this framework rather than a transitional safeguard to be relaxed as confidence in the underlying model grows.

The curriculum-wide pathology coverage deficit is consistent with the pattern described by Mahzari et al [2], in which foundational courses often underweight pathology relative to physiology and pharmacology. One plausible explanation is that faculty teaching in early preclerkship courses are often drawn from basic science rather than pathology backgrounds, and SLOs in these courses establish normal structure and function before disease processes are introduced elsewhere. However, the discipline-level performance comparison in the Results section complicates that a simple coverage explains performance account: physiology showed a below-national-mean performance gap nearly as large as pathology's despite coverage within its target range, so coverage alignment by itself does not fully explain relative discipline performance, and factors such as question difficulty, item exposure, or student study allocation likely also contribute. This descriptive association does not establish a causal or formally tested statistical relationship; a larger, prospective sample linking alignment scores to individual outcomes, controlling for other contributors such as prior preparation, is needed before stronger claims can be made. We return to this point in the Limitations section.

Comparison With Prior Work

This study builds directly on our previously published study describing the underlying USMLE coding library [8]. The present study reports the operational deployment of that library within a complete relational database and reporting framework, including faculty-facing search, cross-course redundancy detection, and accreditation-linked reporting not described in the earlier work. It also extends the theoretical foundation established by Harden [1] and the integration framework proposed by Youm et al [4] by operationalizing curriculum visibility as a queryable, faculty-facing technology rather than a static document. Consistent with the growing body of literature on large language models in medical education published in this journal, including a review of opportunities and challenges for large language models in medical education by Abd-Alrazaq et al [7], this study demonstrates a specific institutional governance application of generative AI that extends beyond the student-facing tutoring and content-generation applications more commonly reported to date. Unlike prior curriculum mapping efforts that relied on one-time computational analyses [6], the architecture described here is persistent operational infrastructure, generating new reports on demand as curriculum data evolve.

Positioned against a formal program-evaluation framework such as Kirkpatrick's 4-level model or the context, input, process, product model, the evidence reported here sits primarily at the process level: it demonstrates that the framework functions as designed, produces specific and actionable outputs, and yields tagging that agrees closely with independent expert judgment, but it does not yet include reaction-level data from faculty users, learning- or behavior-level data on whether curriculum decisions changed as a result of the reports, or product-level data linking the framework to downstream outcomes such as licensing examination performance. Kirkpatrick level 1 (reaction, via faculty usability data) and level 4 (results, via the planned outcome correlation described below) are both identified as priorities for subsequent work rather than claimed here.

Limitations

This study has several limitations. First, routine phase 2 verification of AI-proposed mappings by the Curriculum Mapping Unit was not accompanied by a formally logged acceptance/correction rate at the time of initial deployment; a retrospective, blinded validation study was subsequently conducted to address this gap and is reported in the Results section and discussed further below, but it postdates the original tagging decisions rather than having been built into the original workflow. That validation study also used a modest rater panel (3 course directors) and a single 30-item sample; the resulting κ estimates, while their CIs remained within their respective agreement categories, would benefit from a larger panel and sample in future work. Second, the pathology coverage and

XSL•FO
RenderX

Step 1 performance association is descriptive, not a formal statistical test controlling for other contributors to performance and should be read as hypothesis-generating rather than confirmatory. Third, the generative AI models used are proprietary and nondeterministic, constraining exact reproducibility; human verification mitigates but does not eliminate this constraint. Fourth, the framework was implemented at a single institution during a single AY, and findings on adoption, workload effects, and report quality may not generalize to institutions with different curricular structures, governance processes, or database platforms. Fifth, formal faculty usability and adoption data, such as search-query volume, satisfaction ratings, or structured usability testing, were not systematically collected during this reporting period; faculty engagement is described narratively in the Results section based on curriculum-committee and course-director feedback, and a structured usability survey is planned as near-term work. Finally, the accuracy of the searching and reporting depends on the conciseness and granularity of the underlying session and clerkship learning objectives; faculty are encouraged to use a staff-assisted AI prompt for writing granular objectives.

Future Directions

A near-term extension of the phase 2 tagging workflow will map all assessment items that are deidentified codes systematically to their corresponding SLOs, completing the bidirectional chain from examination item to SLO, CLO, MEPO, and USMLE content code across all preclerkship courses. Once in place, the AI agent will be able to generate blueprinting reports showing which program objectives any assessment evaluates, at what Bloom's Taxonomy level, and how that distribution aligns with USMLE discipline targets. This assessment-tagging extension had not yet been implemented at the time of this report and is planned as the focus of a subsequent publication.

Two further extensions are planned once assessment-item tagging is complete. First, codifying assessment items and linking them explicitly to their corresponding SLOs, described earlier, would allow both students and curriculum administrators, not faculty alone, to view which objectives are assessed, at what cognitive level, and how thoroughly, extending the framework's transparency function beyond its current faculty-only scope. Second, a student-facing version of the database is planned that would allow learners to search the curriculum by clinical concept and see, for each session, the corresponding USMLE content domain, providing the same transparent, self-directed curriculum visibility described for faculty to students preparing for licensing examinations.

A recurring challenge described by clerkship leadership at a peer institution illustrates the framework's potential for rapid-cycle improvement. In their current workflow, gap and redundancy reports are generated using an existing reporting system, but the same objective coverage issues recur across successive academic cycles because identified changes are not implemented before the clerkship resumes: with reporting turnaround measured in weeks, the narrow window, often as short as 2 weeks, between rotation blocks in which curricular revisions can practically be made closes before the relevant report becomes available. Because the present framework's AI-agent reporting layer returns gap and redundancy analyses within minutes rather than weeks, a report generated at the close of one rotation block can be reviewed and acted on well within the following block's revision window. The same leadership described a related diagnostic gap: their system can confirm that a fixed number of institutional program objectives are formally defined but can successfully trace curricular coverage for only a subset of them, with no mechanism to identify which objective is missing or why. Because the present framework tags every SLO explicitly to specific MEPOs at the point of data entry, an AI-agent query can trace the chain of linked SLOs to determine where in curriculum design a gap originates. This diagnostic capability was demonstrated directly within our own institution's clerkship phase (see Clerkship MEPO Coverage Analysis section; Table 3), where a single query traced all 25 MEPOs against clerkship SLOs across 9 rotations and identified 1 MEPO with no matching SLO in any clerkship, a gap not previously visible through existing reporting processes. Extending this tagging discipline to assessment items would enable the equivalent diagnostic capability for assessment coverage specifically.

The most institutionally significant planned application of this infrastructure is expected to be its support for LCME accreditation preparation, particularly for Standard 6, which requires schools to document that MEPOs are reflected throughout the curriculum and demonstrably map to curricular content and, eventually, to assessments. Airtable's native Claude AI integration offers a no-configuration pathway to the reporting capabilities described here, and future work will evaluate whether this native integration reproduces the analytical quality of structured-prompt API methods. This framework can be built at any medical school regardless of size or infrastructure, requiring no proprietary curriculum management platform and no institution-level AI servers; ongoing costs, limited to a cloud database subscription and AI-agent query fees, typically total under several hundred dollars per AY at current pricing. This cost structure is a source of risk as well as accessibility: because the framework depends on 2 external commercial services rather than institution-owned infrastructure, a substantial vendor price increase, a change in API terms, or discontinuation of either service could require migration effort or increased recurring cost. The framework's database-agnostic and model-agnostic design, described in the Methods section, is intended to mitigate this risk by allowing either component to be replaced without redesigning the underlying data structure, but this substitutability has not been tested in practice, and institutions adopting this approach should budget for the possibility of a migration.

The present framework was built around the USMLE Content Outline and LCME accreditation standards, which apply specifically to US MD-granting programs. Adaptation to US osteopathic medical education would require substituting a Comprehensive Osteopathic Medical Licensing Examination (COMLEX)–aligned content outline for the USMLE library described here; because COMLEX and USMLE differ in content organization and osteopathic-specific domains, this would require rebuilding the coding library rather than simply relabeling the existing one, though the relational database

XSL•FO
RenderX

architecture and AI-tagging workflow themselves would not need to change. International medical schools present a larger adaptation challenge, since many national or regional licensing examinations do not publish a content outline with the same structure, granularity, or public availability as the USMLE Content Outline, and some jurisdictions have no equivalent standardized examination at all; institutions in those settings would likely need to construct a comparable local coding library from their own accreditation or licensing body's standards before this framework's tagging and reporting methodology could be applied. Adaptation for nursing, pharmacy, dental, and postgraduate training is comparatively more straightforward wherever a curriculum aligns to a standardized external examination with a published content outline. We have not tested the framework in any of these adjacent contexts and present this discussion as an assessment of adaptation requirements rather than a demonstrated capability.

Conclusions

An AI-integrated relational curriculum database with a faculty-facing search interface and an AI-agent reporting layer can make curriculum content, redundancy, and USMLE alignment visible and actionable on a timescale of minutes rather than weeks, producing specific recommendations rather than aggregate metrics alone. This implementation study reports the first phase of a larger program of work; extension to assessment-item tagging and prospective validation against licensing examination outcomes are planned as subsequent contributions.

Acknowledgments

The authors thank Fabiola Saldivar, MD, MSHS; Anthony Hackett, DO, FACEP, FAAEM, FRSM; and Anitha Shenoy, PhD for independently coding session learning objectives against the USMLE (United States Medical Licensing Examination) Content Outline as part of the blinded validation study, contributing the rater data underlying the interrater reliability analysis reported in the Results section. The authors also thank Bridget Lighthall, MEd; Dustin Dubois, PhD; and the Office of Academic Affairs team for their support and encouragement. Anthropic Claude Sonnet 5, accessed via API, was used as described in the Methods section to propose USMLE content code, Association of American Medical Colleges (AAMC) keyword, and Bloom's Taxonomy tags for session learning objectives; all AI-generated tags were reviewed and verified by the Curriculum Mapping Unit before database entry. An AI agent connected to the study database via API was used to generate the alignment, redundancy, gap, and accreditation reports described in the Results section. Elements of this work were presented at the Southern Group on Educational Affairs 2026 Annual Meeting and submitted to the AAMC 2026 Staff Track Call for Proposals.

Funding

The authors declared no financial support was received for this work.

Data Availability

The USMLE (United States Medical Licensing Examination) Content Outline coding library and the Airtable database template generated during this study are available from the corresponding author on reasonable request.

Generative AI Use

The authors declare that generative AI was not used in the preparation of this manuscript.

Authors' Contributions

RMB wrote the manuscript and developed the innovation. JL assisted RB with data collection and processing. DD provided senior supervision and oversight.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Representative prompt templates and classification decision rules used for phase 2 AI-agent mapping of session learning objectives to USMLE (United States Medical Licensing Examination) Content Outline codes, including the exact prompt structure, the rules governing code selection when an objective could plausibly map to more than 1 domain, and example inputs and outputs.
[PDF File (Adobe PDF File), 166 KB - mededu_v12i1e107727_app1.pdf]

Multimedia Appendix 2
Screenshot of the faculty-facing read-only Airtable clerkship curriculum database (AY25_26_Clerkship_Map_USMLE_NBME), shown for the Emergency Medicine (EM) rotation (academic year 2025-2026). Rotation tabs and a dedicated Medical Education

XSL•FO
RenderX

Program Objective (MEPO) Reference tab provide direct access to all clerkship disciplines. Each record links a clerkship core objective to its rotation-specific session learning objective, mapped MEPO, United States Medical Licensing Examination Step 2 content topic, and corresponding National Board of Medical Examiners (NBME) content category, extending the preclerkship database architecture to include direct NBME mapping.
[PNG File , 162 KB - mededu_v12i1e107727_app2.png]

References

1. Harden R. AMEE Guide No. 21: curriculum mapping: a tool for transparent and authentic teaching and learning. Med Teach 2001;23(2):123-137. [doi: 10.1080/01421590120036547] [Medline: 11371288]
2. Mahzari M, AlNahedh T, Ahmed AA, Al Rumyyan A, Shaban S, Magzoub ME. Practical guide to undergraduate medical curriculum alignment and mapping. Adv Med Educ Pract 2023;14:1001-1012 [FREE Full text] [doi: 10.2147/AMEP.S424815] [Medline: 37745032]
3. Barde PB, Parmar N, Chitturi V, Sharma G, Kathrotia R, Joshi K, et al. Curricular redundancy in medical undergraduate course: critical insights from various stakeholders in medical institutes in India. Avicenna J Med 2024;14(3):167-174 [FREE Full text] [doi: 10.1055/s-0044-1791843] [Medline: 39584165]
4. Youm J, Christner J, Hittle K, Ko P, Stone C, Blood AD, et al. The 6 degrees of curriculum integration in medical education in the United States. J Educ Eval Health Prof 2024;21:15 [FREE Full text] [doi: 10.3352/jeehp.2024.21.15] [Medline: 38872249]
5. Al-Eyd G, Achike F, Agarwal M, Atamna H, Atapattu DN, Castro L, et al. Curriculum mapping as a tool to facilitate curriculum development: a new school of medicine experience. BMC Med Educ 2018;18(1):185 [FREE Full text] [doi: 10.1186/s12909-018-1289-9] [Medline: 30081897]
6. Takizawa PA. Using a topic model to map and analyze a large curriculum. PLoS One 2023;18(4):e0284513 [FREE Full text] [doi: 10.1371/journal.pone.0284513] [Medline: 37079546]
7. Abd-Alrazaq A, AlSaad R, Alhuwail D, Ahmed A, Healy PM, Latifi S, et al. Large language models in medical education: opportunities, challenges, and future directions. JMIR Med Educ 2023;9:e48291 [FREE Full text] [doi: 10.2196/48291] [Medline: 37261894]
8. Badeau RM, Mallini T, Laury J, VanBuren V, Dickey D. From parallel to aligned: a United States Medical Licensing Examination (USMLE) mapping and assessment-tagging framework for medical education. BMC Med Educ 2026;26(1):1606 [FREE Full text] [doi: 10.1186/s12909-026-09912-2] [Medline: 42449355]
9. Biggs J, Tang C. Teaching for Quality Learning at University. 4th Edition. London, United Kingdom: Open University Press; 2011.
10. Wijnen-Meijer M, van den Broek S, Koens F, Ten Cate O. Vertical integration in medical education: the broader perspective. BMC Med Educ 2020;20(1):509 [FREE Full text] [doi: 10.1186/s12909-020-02433-6] [Medline: 33317495]
11. Title 45, Part 46, Section 46.102: Definitions for purposes of this policy. Code of Federal Regulations. URL: https://www.ecfr.gov/current/title-45/subtitle-A/subchapter-A/part-46/subpart-A/section-46.102 [accessed 2026-10-29]
12. Functions and structure of a medical school. Liaison Committee on Medical Education. 2026. URL: https://lcme.org/standards-and-rules/ [accessed 2026-10-22]

Abbreviations

AAMC: Association of American Medical Colleges
AY: academic year
COMLEX: Comprehensive Osteopathic Medical Licensing Examination
CLO: course learning objective
FERPA: Family Educational Rights and Privacy Act
FOMI: Foundations of Medicine I
FOMII: Foundations of Medicine II
ID: Introduction to Disease
LCME: Liaison Committee on Medical Education
MEPO: Medical Education Program Objective
POMI: Practice of Medicine I
POMII: Practice of Medicine II
SLO: session learning objective
USMLE: United States Medical Licensing Examination

XSL•FO
RenderX

Edited by D Chartash; submitted 22.Jul.2026; peer-reviewed by SSR Anugu, A M, Y Lim; comments to author 10.Sep.2026; revised version received 25.Sep.2026; accepted 14.Oct.2026; published 12.Nov.2026.

Please cite as:
Badeau RM, Laury J, Dickey D
An AI-Integrated Relational Curriculum Mapping Framework for Alignment and Transparency in Medical Education: Implementation Study
JMIR Med Educ 2026;12:e107727
URL: https://mededu.jmir.org/2026/1/e107727
doi:10.2196/107727
PMID:

©Robert M Badeau, Jennifer Laury, Danielle Dickey. Originally published in JMIR Medical Education (https://mededu.jmir.org), 12.Nov.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Education, is properly cited. The complete bibliographic information, a link to the original publication on https://mededu.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Simulation-Based Radiology Training for Medical Technical Assistant Trainees: Quasi-Experimental Study

Matthias J Witti[1], Dr rer biol hum; Joséphine-Charlotte Coleman[1]; Constanze Richters[1], PhD; Johanna Schönberger[2], PhD; Mary-Anne de Ruijter[3]; Alina Stortz[1]; Melanie Bauer[1], BSc; Raisa Kiriakidou[1]; Matthias Stadler[1], Prof Dr

[1]Institute of Medical Education, LMU University Hospital, LMU Medizin, Ludwig-Maximilians-Universität München, Pettenkoferstr. 8a, Munich, Germany

[2]medverse GmbH, Munich, Germany

[3]Bayer Medical Care B.V., Radiology, Maastricht, The Netherlands

Corresponding Author:
Matthias J Witti, Dr rer biol hum
Institute of Medical Education, LMU University Hospital, LMU Medizin, Ludwig-Maximilians-Universität München, Pettenkoferstr. 8a, Munich, Germany

Abstract

Background: Simulation-based education is increasingly used in medical training to safely develop clinical skills. However, the perceived relevance of a simulation ("utility value") to a learner's future work may influence its effectiveness.

Objective: This study evaluated a digital radiology simulation module for medical technical assistant trainees in Germany, examining learning gains, group differences in motivational variables, and the moderating role of utility value.

Methods: Fifty German medical technical assistant trainees (including 24 specializing in radiology and 26 focused on laboratory medicine) completed a pretest and posttest requiring them to arrange 12 steps of a contrast injection procedure in the correct order. All participants then engaged in an interactive computer simulation of the procedure. After the simulation, trainees rated their self-efficacy, as well as the intrinsic value and perceived utility value of the simulation. Learning gains were analyzed using a mixed-design ANOVA on the Spearman footrule distance composite score, and group differences in motivational variables were tested using independent-sample 2-tailed *t* tests. The moderating role of utility value on learning gains was examined using a mixed-design analysis of covariance.

Results: Trainees showed significant improvement in procedural sequencing from pretest to posttest ($F_{1,48}$=22.05; P<.001; η^2=0.103), indicating effective learning across both groups. Radiology-specialized trainees reported higher utility value for the training than laboratory-focused trainees (t_{46}=4.43; P<.001; Cohen d=1.28), whereas no group differences emerged in intrinsic value or self-efficacy. Contrary to our hypothesis, perceived utility value did not significantly moderate learning gains ($F_{1,46}$=0.92; P=.34).

Conclusions: The simulation produced robust learning gains regardless of learner specialization or perceived relevance. While utility value perception differed markedly between groups, it did not translate into differential learning outcomes, suggesting that well-designed simulations can be effective across learners with varying levels of task relevance.

(JMIR Med Educ 2026;12:e97489) doi:10.2196/97489

KEYWORDS

simulation-based education; utility value; motivation; radiology training; medical technical assistants

Introduction

Simulation-based learning has become a cornerstone of medical and allied health education, offering learners a safe and controlled environment to practice clinical skills and procedures. Numerous studies have demonstrated that high-fidelity medical simulations can significantly improve the knowledge and procedural skills of learners [1]. Well-designed simulations lead to effective learning by allowing deliberate practice with feedback, and technology-enhanced simulation training has positive effects on learning outcomes in health professions education [2], including improved skill performance and knowledge retention [3,4]. Through virtual scenarios, learners can engage in complex procedures without risk to patients, building competence and confidence for real clinical settings. For example, virtual radiography simulations have enabled learners to practice technical skills (eg, equipment setup and protocol execution) in a safe environment, leading to increased confidence and better preparedness for clinical practice [1]. Medical radiation trainees also appreciate that simulation allows

XSL•FO
RenderX

them to practice techniques and skills that may not be possible while on placement due to safety or access limitations [5].

While the general efficacy of simulation-based education is well established, an important question is what factors might further enhance or hinder its impact on learning. One such factor is the perceived relevance or utility value of the simulation to the learner. According to the expectancy-value theory of motivation [6,7], individuals are more motivated and perform better when they expect to succeed and perceive value in a task. Two important components of task value are intrinsic value, which refers to the enjoyment that a learner derives from engaging in a task, and utility value, which reflects the perceived usefulness of a task for one's future or current goals. Both intrinsic and utility value can strongly promote interest and engagement, yet they do so in different ways: intrinsic value captures the inherent enjoyment and interest experienced during a task, whereas utility value captures the learner's perception of the task's relevance to personal or professional goals outside the task itself. Research has shown that learners who perceive high utility value tend to invest more effort, show greater interest, and achieve higher performance [8]. Conversely, if learners view an activity as low in relevance to their goals, their engagement and learning may suffer even if the instructional design is sound. This principle suggests that the same simulation could yield different outcomes depending on the learner's specialization or career intentions, which shape the utility value they attach to the training.

In the context of medical technical assistant (MTA) education in Germany, trainees typically specialize in distinct fields such as radiology or laboratory medicine during their training. This presents a natural contrast in utility value: a radiology-focused simulation would be directly relevant to trainees specializing in radiological technology, medical technologists for radiology (MTRAs), but might appear less pertinent to those on a laboratory technology track, medical technologists for laboratory analytics (MTLAs). Expectancy-value theory predicts that radiology trainees (high-utility group) may be more engaged and learn more from the simulation than laboratory trainees (low-utility group) even when both groups receive identical training. However, this hypothesis requires empirical examination as other motivational factors and the engaging nature of the simulation itself could also influence the outcome.

This study reports an experimental evaluation of a simulation-based radiology training module for MTA trainees, with a focus on how trainees' specialization and the associated utility value they derive affect learning and motivational outcomes. The simulation is a digital training module for a computed tomography (CT) contrast injection system (MEDRAD Centargo), developed in collaboration with a medical technology provider (Bayer MEDRAD). The module immerses trainees in the step-by-step process of using the CT injector, from initial setup to performing a contrast-enhanced scan procedure to postprocedure cleanup, organized into 7 sequential chapters reflecting the real-world clinical workflow. By the end of the module, trainees should have a clear mental model of the entire contrast injection workflow and be able to recall the correct sequence of steps involved in using the CT injector safely and efficiently.

Using this simulation as an intervention, the present study addressed 3 main research questions (RQs):

- RQ 1: Does completing the radiology simulation lead to significant learning gains in procedural knowledge from before to after training, and do these gains differ between radiology-specialized and laboratory-specialized trainees?
- RQ 2: Do radiology-specialized trainees report higher intrinsic value, self-efficacy, and utility value of the simulation than laboratory-specialized trainees?
- RQ 3: Does perceived utility value moderate learning gains (ie, do trainees who perceive the training as more useful show larger improvements)?

We hypothesized that, through the simulation activity, trainees in both specializations would improve their sequencing of the steps in the procedure, reflecting effective procedural learning. We further expected that radiology track trainees would rate the simulation as more useful than laboratory track trainees given its direct relevance to radiology practice, whereas we did not formulate directional expectations regarding group differences in intrinsic value or self-efficacy. Finally, based on expectancy-value theory, we anticipated that higher perceived utility value would be associated with greater learning improvement.

Methods

Participants and Design

Participants were 50 MTA trainees (second to third trainee year) enrolled in an MTA training program in Germany. All participants had completed basic coursework in their respective tracks but had limited or no prior hands-on experience with the specific radiology procedure featured in the simulation. The sample included 24 trainees specializing in radiology (MTRA track) and 26 specializing in laboratory medicine (MTLA track). This grouping served as a quasi-experimental factor to represent different levels of inherent utility value for the radiology simulation. The majority in both groups were female: MTRA (70%) and MTLA (73%), which is consistent with the typical gender distribution in MTA training programs. Mean age overall was 23.1 (SD 6.93) years, with no significant difference between groups (P=.30).

This study used a pretest-posttest design, with all trainees undergoing the same simulation intervention. There was no separate control group as our primary interest was in within-subject learning gains and between-group differences linked to specialization. Trainees first completed a pretest, then the simulation, and immediately afterward a posttest and posttraining questionnaire. Specialization was treated as an independent grouping variable in the analysis. Procedure and testing were identical for both groups, and the study was conducted in a single session per participant.

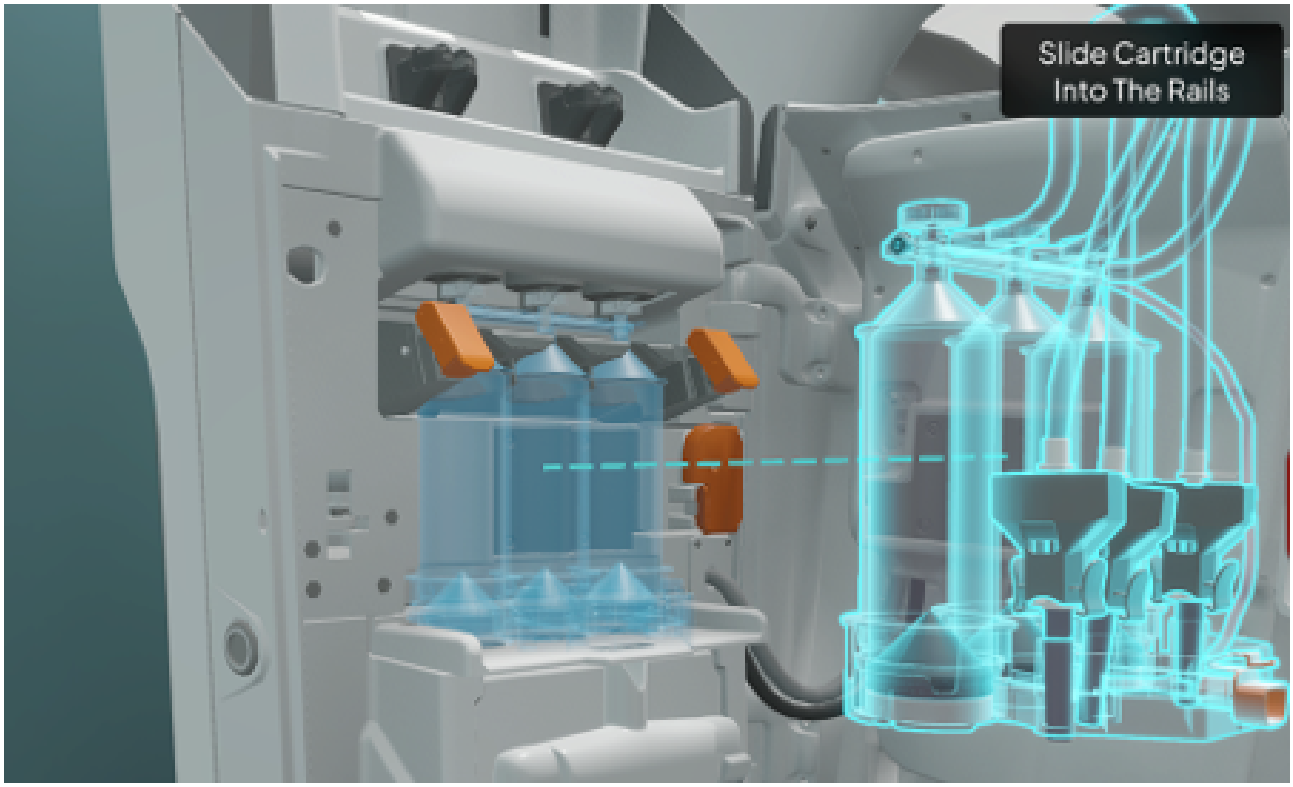
Simulation Training System Intervention

The training system was a custom digital simulation module teaching the operation of the MEDRAD Centargo CT power injection system. This system is used in CT radiology suites to administer contrast media and saline during CT scans. The simulation was delivered via a computer, required trainees to

XSL•FO
RenderX

perform virtual actions in a clinically realistic sequence, and was structured into 7 chapters corresponding to major phases of the procedure (see Figure 1 for a screenshot of the simulation interface).

Figure 1. Screenshot of the simulation interface ("Install the daily set").



Measures

Procedural Knowledge Test (Pretest and Posttest)

The primary learning outcome was the ability to correctly sequence the steps of the radiology procedure following the approach by Bauer et al [9]. The task consisted of a list of 12 key steps involved in the CT injector workflow presented in a randomized order. Trainees were instructed to arrange these steps into the correct chronological order. The pretest assessed baseline knowledge before the simulation; the posttest used the same steps (rerandomized by position) immediately after the simulation to measure learning. For scoring, each participant's sequence was compared to the expert-established correct order using the Spearman footrule distance, computed as the sum of absolute deviations between each step's assigned position and its correct position. Lower values indicate greater accuracy.

Postsimulation Questionnaire

After the posttest, trainees completed a questionnaire assessing intrinsic value [10], self-efficacy [11], and utility value [10]. Each scale used a Likert-type format from 1 ("strongly disagree") to 5 ("strongly agree"). The intrinsic value scale (3 items) evaluated enjoyment derived from the simulation. The self-efficacy scale (3 items) measured confidence in performing radiology procedures after training. The utility value scale (3 items) assessed perceived relevance and usefulness of the simulation for the trainee's future work. Internal consistency (McDonald ω) was high for all 3 scales (intrinsic value: ω=0.95; self-efficacy: ω=0.77; utility value: ω=0.96).

Procedure

Study sessions were conducted at the MTA training institutes. Trainees were scheduled in regular classes and provided with individual computers and headphones. Two or three researchers were present to ensure standardized administration. Trainees were informed of the general purpose of the session but not told about the specific hypotheses regarding specialization and utility value.

Data Analysis

We used Jamovi (version 2.3) for all statistical analyses [12]. Following the structure of our 3 RQs, we conducted 3 separate analyses, each using the Spearman footrule distance as the aggregate outcome measure. This composite avoids the dependency inherent in analyzing the 12 individual step positions separately as these constitute a permutation and are therefore not statistically independent. Raw, nontransformed distance scores were used throughout; log-transformation would only be warranted in cases of severe nonnormality.

For RQ 1, we conducted a 2 × 2 mixed-design ANOVA with time (pretest vs posttest) as the within-subject factor and group (radiology vs laboratory) as the between-subject factor. This model tests both the overall learning gain (main effect of time) and whether that gain differed between the 2 trainee groups (time × group interaction).

For RQ 2, we compared the 2 specialization groups on intrinsic value, self-efficacy, and utility value using independent-sample *t* tests. Where the Levene test indicated a violation of the equal variance assumption, the Welch correction was applied.

XSL•FO
RenderX

For RQ 3, we examined whether perceived utility value moderated learning gains using a mixed-design analysis of covariance with time as the within-subject factor and utility value as a continuous covariate. The time × utility value interaction tests whether trainees who rated the simulation as more useful showed larger pre- to posttest improvements.

Effect sizes are reported as η^2 for ANOVA-based analyses and Cohen *d* for *t* tests. Statistical significance was set at α=.05 for all tests.

Ethical Considerations

This study was deemed ethically unobjectionable following consultation with the Ethics Committee at the Faculty of Medicine of Ludwig-Maximilians-University Munich (26-0009-KB) and was conducted in accordance with the ethical principles of the Declaration of Helsinki (World Medical Association). Participation was voluntary, and informed consent was obtained.

Results

Learning Gains and Group Differences (RQ 1)

Table 1 shows the average position assigned to each procedural step at pretest and posttest alongside the deviation from the correct position. As can be observed, deviations were considerably reduced from pretest to posttest across most steps, most notably for step 9 (“empty or eject the daily set”), where trainees’ placement improved substantially. Figure 2 visualizes this pattern.

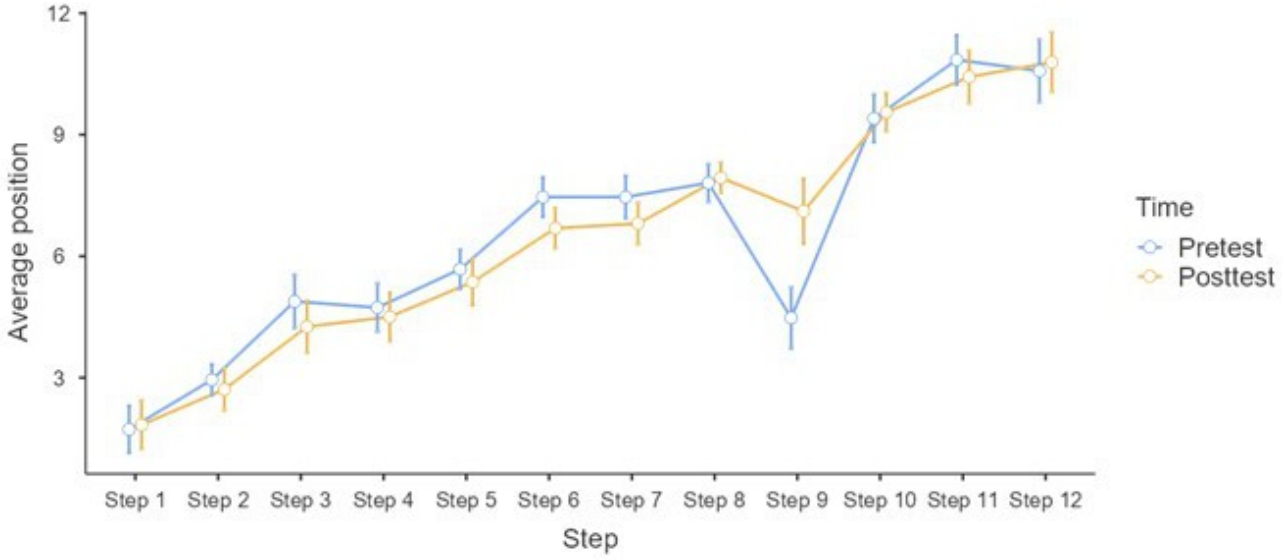
Table . Average step position and deviation from the correct position at pretest and posttest.[a]

Step	Pretest		Posttest	
	Position, mean (SD)	Deviation from correct position	Position, mean (SD)	Deviation from correct position
1. Perform injection	1.74 (2.05)	0.74	1.82 (2.13)	0.82
2. Insert contrast agent and saline containers	2.96 (1.35)	0.96	2.72 (1.84)	0.72
3. Check patient tubing for air bubbles	4.90 (2.32)	1.90	4.26 (2.23)	1.26
4. Prime patient tubing	4.72 (2.09)	0.72	4.50 (2.08)	0.50
5. Dispose of patient tubing	5.68 (1.68)	0.68	5.36 (1.99)	0.36
6. Install patient tubing	7.46 (1.69)	1.46	6.70 (1.74)	0.70
7. Connect patient-side tubing to patient	7.46 (1.18)	0.46	6.82 (1.83)	−0.18
8. Switch on the system	7.80 (1.64)	−0.20	7.94 (1.30)	−0.06
9. Empty or eject the daily set	4.46 (2.67)	−4.54	7.10 (2.81)	−1.90
10. Install the daily set	9.38 (2.10)	−0.62	9.56 (1.64)	−0.44
11. Fill the daily set with fluids	10.86 (2.14)	−0.14	10.44 (2.31)	−0.56
12. Check (test) protocol status	10.58 (2.69)	−1.42	10.78 (2.57)	−1.22

[a]Deviation=assigned position – correct position; values closer to 0 indicate greater accuracy.

XSL•FO
RenderX

Figure 2. Average position of the 12 procedural steps at pretest and posttest.



Descriptively, the footrule composite score decreased from pretest to posttest in both groups (radiology: mean 16.7, SD 11.5 to 14.3, SD 17.7; laboratory: mean 20.9, SD 11.4 to 13.6, SD 13.7), confirming the direction of the observed main effect of time. The pretest-posttest correlation for the composite score was significant and positive in both groups (radiology: r=0.544, 95% CI 0.180-0.777 and P=.006; laboratory: r=0.417, 95% CI 0.035-0.693 and P=.03).

The mixed-design ANOVA showed a significant and medium-sized main effect of time ($F_{1,48}$=22.05; P<.001; η^2=0.103), confirming that the simulation effectively improved procedural sequencing accuracy across both groups. The time × group interaction was not significant ($F_{1,48}$=2.53; P=.12; η^2=0.012), indicating that learning gains did not differ meaningfully between radiology-specialized and laboratory-specialized trainees. The between-subject main effect of group was also nonsignificant ($F_{1,48}$=0.41; P=.53; η^2=0.006), confirming comparable overall performance levels across the 2 tracks. In sum, both groups benefited substantially and equally from the simulation.

Group Differences in Motivational Variables (RQ 2)

Descriptive statistics for all self-report measures are shown in Table 2.

Table . Descriptive statistics and intercorrelations for self-report measures.

	Score (1-5), mean (SD)	Intrinsic value	Self-efficacy	Utility value
Intrinsic value	3.79 (1.10)	0.95[a]	0.49[b]	0.59[b]
Self-efficacy	4.07 (0.65)	0.49[b]	0.77[a]	0.42[b]
Utility value	3.48 (1.38)	0.59[b]	0.42[b]	0.96[a]

[a]Internal consistency (McDonald ω).

[b]P<.001.

Radiology- and laboratory-specialized trainees did not differ significantly in intrinsic value (t_{45}=1.49; P=.14; Cohen d=0.44) or self-efficacy (t_{44}=0.97; P=.34; Cohen d=0.29). As expected, the groups differed substantially in utility value (t_{46}=4.43; P<.001; Cohen d=1.28), with the Levene test indicating unequal variances and the Welch correction applied accordingly. Radiology-specialized trainees rated the simulation as considerably more useful and relevant to their future work (mean 4.29, SD 0.74) than laboratory-specialized trainees (mean 2.79, SD 1.43). The large effect size confirms the effectiveness of the specialization grouping in creating differential utility perceptions.

Utility Value as a Moderator of Learning Gains (RQ 3)

The mixed-design analysis of covariance with utility value as a continuous covariate confirmed the overall learning gain with a significant main effect of time ($F_{1,46}$=6.79; P=.01; η^2=0.035). However, the time × utility value interaction was not significant ($F_{1,46}$=0.92; P=.34; η^2=0.005), indicating that perceived utility value did not moderate the degree of learning improvement. The between-subject effect of utility value was similarly nonsignificant ($F_{1,46}$=0.32; P=.58; η^2=0.005). Trainees who found the simulation more relevant did not show larger learning gains than those who found it less relevant. Descriptively, the direction of the covariate effect was positive. Trainees with higher utility value scores showed marginally larger pre- to

XSL•FO
RenderX

posttest gains (r=0.096, 95% CI –0.194 to 0.370; P=.52), consistent with, though far weaker than, theoretical expectations. Raw, unadjusted bivariate correlations between utility value and the footrule composite score were similarly weak and nonsignificant at both pretest (r=0.085, 95% CI –0.204 to 0.361; P=.57) and posttest (r=–0.024, 95% CI −0.306 to 0.262; P=.87), confirming that utility value was not meaningfully associated with performance at either time point.

Discussion

Principal Findings

This study investigated the effectiveness of a simulation-based radiology training module for MTA trainees and examined how perceived utility value influenced learning outcomes. The results demonstrated that the simulation was an effective learning tool: trainees significantly improved their knowledge of the procedural sequence from pretest to posttest, confirming that even a single simulation session can yield meaningful procedural knowledge gains. Importantly, these gains were consistent across both radiology-specialized and laboratory-specialized trainees, suggesting that the simulation was equally effective regardless of specialization background. Contrary to our hypothesis, however, perceived utility value did not significantly moderate learning gains. Trainees who found the simulation more relevant to their future work did not show measurably larger improvements than those who perceived it as less relevant.

Theoretical Implications: Expectancy Value in Simulation Learning

Our findings offer a partial, rather than full endorsement of expectancy-value theory in this simulation context. The group difference in utility value was large and consistent with theoretical predictions: radiology-specialized trainees rated the simulation as substantially more relevant than their laboratory-specialized peers. However, this difference in perceived relevance did not translate into differential learning outcomes either at the group level or when utility value was modeled as a continuous predictor of individual learning gains. This null moderation finding suggests that utility value, while a meaningful dimension of task motivation, does not uniformly amplify learning in simulation-based contexts, particularly when the simulation itself is well designed, engaging, and instructionally scaffolded.

One possible explanation is that the simulation's interactive design and immediate feedback structure created sufficient conditions for effective procedural learning across all trainees, effectively compensating for lower utility perceptions in the laboratory group. When the learning environment is highly structured and cognitively supportive, motivational differences between learners may matter less than in less scaffolded settings [3,4]. Alternatively, the sequencing task may have been learnable by most trainees regardless of motivation, and utility value effects might emerge more clearly in longer or more cognitively demanding interventions. Future research using more complex procedural tasks or longer training durations may be better equipped to detect motivational moderation effects. Additionally, immediate posttests may not fully capture the influence of utility value as its effects may be more apparent over time than after a single training session. If utility value promotes deeper engagement or more effortful processing during learning, its benefits may only become apparent once the initial performance gains from immediate feedback have faded. Therefore, delayed retention measures may offer a more sensitive test of motivational moderation than performance assessed directly after training.

Interestingly, intrinsic value and self-efficacy were uniformly high across both groups and did not differentiate learning outcomes. The simulation appeared to capture trainees' interest and confidence regardless of their specialization, suggesting that well-crafted interactive simulations can engage diverse learner populations. Authors and journals alike have a responsibility to transparently report null findings to prevent publication bias, and the null moderation result reported in this study contributes important empirical evidence on the boundary conditions of expectancy-value theory in simulation-based medical education.

Practical Implications: Designing and Targeting Simulations for Maximum Impact

From a practical standpoint, our findings offer nuanced guidance for educators and simulation designers. First, the consistent learning gains across both groups suggest that high-quality, feedback-rich simulations can be effective equalizers, supporting procedural knowledge acquisition even among learners who perceive the content as less directly relevant to their career. For heterogeneous cohorts, this is encouraging: simulation-based training need not be restricted to those with the most obvious career alignment.

Second, while utility value did not moderate learning outcomes in this study, the marked group difference in utility perceptions (Cohen d=1.28) is itself noteworthy. In longer interventions, more sustained training programs, or contexts in which motivation may play a greater role over time, utility value could yet prove consequential. Framing interventions that highlight the broader relevance of procedural skills, such as structured reflection exercises connecting the simulation to trainees' future professional roles [13], remain a promising avenue for instructional design even if their effect on short-term learning gains was not confirmed in this study.

Third, the uniformly high intrinsic value and self-efficacy scores attest to the simulation's success in creating an engaging and confidence-building learning experience. Features likely contributing to this include interactive hands-on tasks, immediate feedback, and clinical authenticity. These qualities align with established principles of effective simulation design [4,14,15] and should continue to guide the development of training modules across specializations.

Limitations

Several limitations warrant consideration. The sample consisted exclusively of German MTA trainees, limiting generalizability to other educational systems or more advanced learner populations. Specialization was not randomly assigned, introducing potential confounds between group membership and other learner characteristics. The study assessed only

XSL•FO
RenderX

short-term procedural learning; longer-term retention and clinical transfer remain beyond its scope. The outcome measure, a sequencing task, captures procedural knowledge but not complex decision-making or interpersonal skills, where motivational moderation might operate differently.

Finally, the self-report scales, while adapted from established instruments, were short (3 items each) and administered only once after training, which may have limited sensitivity to within-person motivational variation. Future studies could complement self-reports with behavioral engagement measures such as time on task or exploration frequency within the simulation to capture motivation more dynamically.

Conclusions

This study demonstrates that simulation-based training effectively supports the acquisition of complex procedural knowledge among MTA trainees. The simulation produced robust and equivalent learning gains across radiology- and laboratory-specialized trainees, indicating that well-designed interactive simulations can be effective regardless of learners' perceived task relevance. While utility value differed substantially between groups, as expected, it did not moderate individual learning gains, providing important evidence on the boundary conditions of expectancy-value theory in simulation-based medical education. Future research should examine whether motivational moderation emerges in more complex, longer, or retention-focused learning scenarios and should consider experimental manipulation of utility value to test its causal role more directly.

Acknowledgments

The authors would like to thank the trainees for their participation in this study. During the preparation of this manuscript, the authors used ChatGPT (OpenAI; 5.6 Luna) and DeepL Write (DeepL SE; DeepL Translator and Write) to enhance readability and language. After using these tools, the authors reviewed and edited the output as needed and take full responsibility for the content of the publication.

Funding

This study was funded by medverse GmbH.

The authors report there are no competing interests to declare.

Conflicts of Interest

MS is the Associate Editor of *JMIR Medical Education* at the time of this publication. MS had no involvement in the editorial review and processing of this manuscript. All other authors had no conflicts to declare.

References

1. Wertz CI, Wagner J, Ward TM, Mickelsen W. Using simulation in radiographic science education. In: Teaching, Learning, and Leading With Computer Simulations: IGI Global; 2020:37-67. [doi: 10.4018/978-1-7998-0004-0.ch002]
2. Vestbøstad M, Karlgren K, Olsen NR. Research on simulation in radiography education: a scoping review protocol. Syst Rev 2020 Nov 21;9(1):263. [doi: 10.1186/s13643-020-01531-2] [Medline: 33220713]
3. Chernikova O, Heitzmann N, Stadler M, Holzberger D, Seidel T, Fischer F. Simulation-based learning in higher education: a meta-analysis. Rev Educ Res 2020;90(4):499-541. [doi: 10.3102/0034654320933544]
4. Issenberg SB, McGaghie WC, Petrusa ER, Lee Gordon D, Scalese RJ. Features and uses of high-fidelity medical simulations that lead to effective learning: a BEME systematic review. Med Teach 2005 Jan;27(1):10-28. [doi: 10.1080/01421590500046924] [Medline: 16147767]
5. Chau M, Arruzza E, Johnson N. Simulation-based education for medical radiation students: a scoping review. J Med Radiat Sci 2022 Sep;69(3):367-381. [doi: 10.1002/jmrs.572] [Medline: 35178870]
6. Wigfield A, Eccles JS. Expectancy-value theory of achievement motivation. Contemp Educ Psychol 2000 Jan;25(1):68-81. [doi: 10.1006/ceps.1999.1015] [Medline: 10620382]
7. Eccles JS, Wigfield A. Motivational beliefs, values, and goals. Annu Rev Psychol 2002;53:109-132. [doi: 10.1146/annurev.psych.53.100901.135153] [Medline: 11752481]
8. Priniski SJ, Rosenzweig EQ, Canning EA, et al. The benefits of combining value for the self and others in utility-value interventions. J Educ Psychol 2019 Nov;111(8):1478-1497. [doi: 10.1037/edu0000343] [Medline: 31772414]
9. Bauer M, Witti MJ, Stadler M, Weiß A, Fischer MR, Richters C. Interprofessional training in virtual reality for health care: experimental study on procedural knowledge and willingness to collaborate. JMIR Med Educ 2026 May 27;12:e85139. [doi: 10.2196/85139] [Medline: 42202475]
10. Gaspard H, Häfner I, Parrisius C, Trautwein U, Nagengast B. Assessing task values in five subjects during secondary school: measurement structure and mean level differences across grade level, gender, and academic subject. Contemp Educ Psychol 2017;48:67-84. [doi: 10.1016/j.cedpsych.2016.09.003]

XSL•FO
RenderX

11. Chernikova O, Stadler M, Sommerhoff D, et al. The relation between learners' experience in simulations and diagnostic accuracy: generalizability across medical and teacher education. Comput Hum Behav Rep 2024 Aug;15:100454. [doi: 10.1016/j.chbr.2024.100454]
12. jamovi. URL: https://www.jamovi.org [accessed 2026-09-22]
13. Hulleman CS, Godes O, Hendricks BL, Harackiewicz JM. Enhancing interest and performance with a utility value intervention. J Educ Psychol 2010;102(4):880-895. [doi: 10.1037/a0019506]
14. Kim J, Park JH, Shin S. Effectiveness of simulation-based nursing education depending on fidelity: a meta-analysis. BMC Med Educ 2016 May 23;16:152. [doi: 10.1186/s12909-016-0672-7] [Medline: 27215280]
15. Lee J, Kim H, Kim KH, Jung D, Jowsey T, Webster CS. Effective virtual patient simulators for medical communication training: a systematic review. Med Educ 2020 Sep;54(9):786-795. [doi: 10.1111/medu.14152] [Medline: 32162355]

Abbreviations

CT: computed tomography
MTA: medical technical assistant
MTLA: medical technologist for laboratory analytics
MTRA: medical technologist for radiology
RQ: research question

Edited by S Guinez-Molinos; submitted 07.Apr.2026; peer-reviewed by JC Nesbit, S Prakash; revised version received 20.Aug.2026; accepted 03.Sep.2026; published 13.Oct.2026.

Please cite as:
Witti MJ, Coleman JC, Richters C, Schönberger J, de Ruijter MA, Stortz A, Bauer M, Kiriakidou R, Stadler M
Simulation-Based Radiology Training for Medical Technical Assistant Trainees: Quasi-Experimental Study
JMIR Med Educ 2026;12:e97489
URL: https://mededu.jmir.org/2026/1/e97489
doi:10.2196/97489

© Matthias J Witti, Joséphine-Charlotte Coleman, Constanze Richters, Johanna Schönberger, Mary-Anne de Ruijter, Alina Stortz, Melanie Bauer, Raisa Kiriakidou, Matthias Stadler. Originally published in JMIR Medical Education (https://mededu.jmir.org), 13.Oct.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Education, is properly cited. The complete bibliographic information, a link to the original publication on https://mededu.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Immersive Virtual Reality Training to Improve Novice Physicians' Emergency Response Skills: Randomized Controlled Trial

Yeon-Ju Huh[1,2*], MD, PhD; Ju Whi Kim[1,3*], MD, PhD; Narae Yoon[1,4], MD, PhD; Seoi Jeong[5,6,7], MS; Hyoun-Joong Kong[5,7,8], PhD; Sun Jung Myung[1,9], MD, PhD

[1]Office of Medical Education, Seoul National University College of Medicine, Education building, 2nd Fl., 103, Daehak-ro, Jongno-gu, Seoul, Republic of Korea

[2]Department of Surgery, Seoul National University Hospital, Seoul, Republic of Korea

[3]Department of Pediatrics, Seoul National University Hospital, Seoul, Republic of Korea

[4]Department of Psychiatry, Seoul National University Hospital, Seoul, Republic of Korea

[5]Department of Transdisciplinary Medicine, Seoul National University Hospital, Seoul, Republic of Korea

[6]Department of Clinical Medical Sciences, Seoul National University College of Medicine, Seoul, Republic of Korea

[7]Institute of Convergence Medicine with Innovative Technology, Seoul National University Hospital, Seoul, Republic of Korea

[8]Department of Medicine, Seoul National University College of Medicine, Seoul, Republic of Korea

[9]Department of Internal Medicine, Seoul National University Hospital, Seoul, Republic of Korea

[*]these authors contributed equally

Corresponding Author:
Sun Jung Myung, MD, PhD
Office of Medical Education, Seoul National University College of Medicine, Education building, 2nd Fl., 103, Daehak-ro, Jongno-gu, Seoul, Republic of Korea

Abstract

Background: Simulation-based training is essential for preparing medical interns to manage high-stakes emergencies. Although virtual reality (VR)-based simulation has been rapidly integrated into medical education, there remains limited evidence directly assessing its effectiveness relative to established high-fidelity simulation (HFS) methodologies.

Objective: This study aimed to assess the perceived educational effectiveness of VR and HFS in enhancing novice physicians' confidence, satisfaction, and perceived preparedness for managing acute oxygen desaturation.

Methods: A randomized controlled trial was conducted with 168 medical interns from Seoul National University Hospital. Participants were randomly assigned to VR group (n=81) or HFS group (n=87). Overall, 4 participants were excluded due to incomplete surveys, leaving 164 for analysis (VR: 79 and HFS: 85). Both groups were trained to manage simulated patients with low oxygen saturation. Confidence (10-point Likert scale) and satisfaction (7-point Likert scale) were measured using pre and posttraining surveys. Usability was assessed with the User Experience Questionnaire-Short. Between-group comparisons were conducted using *t* tests and chi-square tests, while within-group confidence changes were analyzed using paired *t* tests and repeated-measures analysis of variance. To account for correlated data and estimate effect sizes, generalized estimating equations were applied, with statistical significance set at $P<.05$. Focus group interviews at 1 and 5 months posttraining explored real-world application and behavior transfer. Transcripts were independently reviewed by 2 researchers (YJH and SJM) and thematically analyzed to identify recurring patterns and insights related to clinical behavior.

Results: Confidence in managing oxygen desaturation significantly improved from a mean 3.78 (SD 2.12) to mean 6.20 (SD 2.02) across VR and HFS groups ($t_{163}=-14.04$; $P<.001$), with no significant difference between groups ($F_{1,162}=3.28$; $P=.07$). Satisfaction was high overall mean 6.07 (SD 1.02), but significantly greater in the HFS group than in the VR group (mean 6.23, SD 0.92 vs mean 5.89, SD 1.10; $t_{162}=2.29$; $P=.02$). HFS participants rated tutor guidance (mean 6.49, SD 0.86 vs mean 6.10, SD 1.02; $P=.008$) and authenticity (mean 6.24, SD 1.05 vs mean 5.77, SD 1.15; $P=.006$) higher, whereas both groups scored usability above 5 on all items. Qualitative analyses revealed complementary strengths. Interns valued VR for its immersive environment, focused repetition, and reduced distractions that facilitated stepwise problem-solving. HFS was praised for palpable realism, hands-on practice with equipment, and immediate feedback that reinforced team communication and role clarity. Across follow-up interviews, interns reported improved recognition of desaturation, more structured initial responses (airway assessment, oxygen delivery adjustments, and escalation), and greater willingness to act promptly under pressure—suggesting perceived transfer of learning to clinical practice beyond the simulation lab.

XSL•FO
RenderX

Conclusions: VR may complement HFS in emergency response training. Both modalities were associated with improvements in interns' self-reported confidence and perceived preparedness. The scalability and accessibility of VR suggest its potential value in diverse training contexts.

Trial Registration: ClinicalTrials.gov NCT06295887; https://clinicaltrials.gov/study/NCT06295887

(JMIR Med Educ 2026;12:e71455) doi:10.2196/71455

KEYWORDS

virtual reality; high-fidelity simulation; simulation-based education; medical Interns; emergency response; clinical decision-making; confidence; satisfaction; educational technology; randomized controlled trial

Introduction

The ability to respond effectively to emergencies is a critical competency for medical professionals. Ensuring that interns and new practitioners are adequately prepared for these high-stakes scenarios is essential for improving patient outcomes and reducing errors [1]. A rapid decline in oxygen saturation represents one of the most critical and life-threatening challenges in medical practice, demanding immediate intervention to avert severe complications or death. This emergency is particularly prevalent in high-stakes settings, such as intensive care units, emergency departments, and operating rooms, where the ability of physicians to respond swiftly and effectively is crucial for patient survival and recovery. However, for novice doctors with limited clinical experience, managing such crises can be particularly daunting.

Traditional educational approaches for emergency response training, such as didactic lectures and case-based discussion, remain the predominant methods used in many medical curricula [2,3]. While foundational, these approaches may offer limited opportunities for immersive, hands-on experience necessary to equip learners fully for real-world challenges. They rarely replicate the stress, urgency, and unpredictability of real-life emergencies. Simulation-based education helps address this gap by providing a safe and controlled environment for developing and practicing skills [4,5]. High-fidelity simulation (HFS) using advanced mannequins to mimic realistic patient responses provides an interactive learning experience with immediate feedback without patient risk, but it is resource-intensive, requiring substantial investment and logistical support [6,7]. Recent advances in technology have introduced virtual reality (VR) as a promising alternative to traditional HFS [8]. VR simulation leverages immersive, computer-generated environments to recreate highly realistic clinical scenarios. VR offers several benefits, including lower cost, greater accessibility, and the ability to simulate a wide range of scenarios that might be impractical or impossible to replicate with physical mannequins [9]. However, the comparative effectiveness of VR and HFS in training medical interns for emergency response remains underexplored [10]. In addition, most medical VR systems focus on step-by-step procedures training rather than dynamic decision-making under pressure [11-13].

This study compared VR and HFS in training novice physicians for emergency oxygen desaturation management. Beyond immediate satisfaction and confidence (Kirkpatrick Levels 1 and 2), we explored the long-term perceived clinical transfer (Kirkpatrick Level 3) by examining learners' reflections on their responses to similar clinical situations 1 and 5 months posttraining [14]. This longitudinal perspective aligns with Knowles' adult learning theory [15], emphasizing the transition from cognitive readiness to the practical application of experiential knowledge. We hypothesized that VR would demonstrate educational outcomes comparable to HFS in terms of both immediate psychological readiness and perceptions of learning transfer, while offering a distinct learner experience.

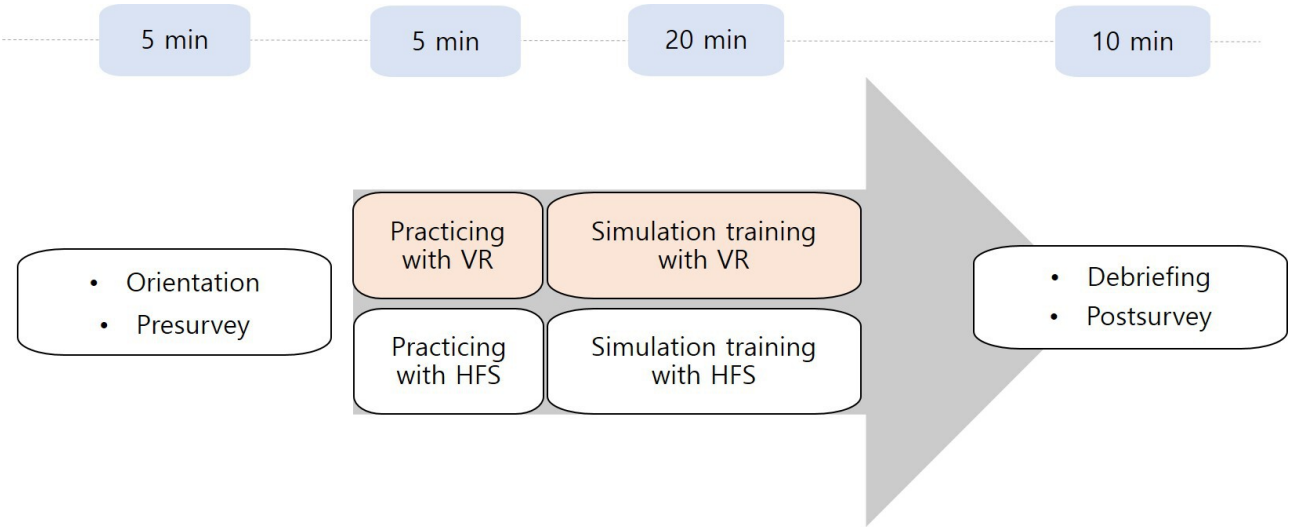
Methods

Study Design

A 2-arm parallel randomized controlled trial was conducted to compare the educational effectiveness of VR and HFS methods. The sample size for this study was based on a prior study using G*Power 3.1.2 software (Heinrich Heine University Düsseldorf) for an independent *t* test, assuming a type I error (α) of .05, statistical power ($1-\beta$) of .80, and an effect size (Cohen d) of 0.80, which indicated that a minimum of 52 participants would be required [16]. The random allocation sequence was generated using the clinical trial randomization tool [17] by an independent third party unrelated to the study, ensuring reproducibility. To ensure allocation concealment, the randomization list was stored on a password-protected server accessible only to the independent administrator. All participant data were anonymized, and the trial administrators and personnel analyzing the data were blinded to the participants' identities and group assignments. Participants simulated scenarios involving desaturation based on their previous knowledge using the specific training method for their group, followed by a debriefing session to reinforce learning. Following completion of the simulation scenarios, participants in both groups engaged in an in-person debriefing session facilitated by faculty instructors (Figure 1). Debriefing sessions were conducted using a structured format with comparable group size and time allocation across the VR and HFS groups. Tutors were faculty physicians with clinical experience in acute care and prior experience in simulation-based medical education. All tutors had previously facilitated HFS sessions and received orientation to the study scenarios and debriefing framework before the intervention. Tutors facilitated the simulation sessions and led the postscenario debriefing. The findings were reported following the CONSORT (Consolidated Standards of Reporting Trials) guidelines, [18] the checklist and the trial protocol can be accessed in Checklist 1.

XSL•FO
RenderX

Figure 1. Study flow chart. HFS: high-fidelity simulation; VR: virtual reality.



Participants and Recruitment

Participants were postgraduate year 1 medical interns who had recently graduated from medical school and were about to begin their internship training at a tertiary teaching hospital. Eligible interns were identified from the annual intern cohort and were invited to participate in February 2024 through email invitations and in-person announcements during scheduled educational sessions. Interns who declined participation or were unable to attend the scheduled simulation sessions were excluded. Participation was voluntary, and interns were explicitly informed that participation or nonparticipation would not affect their training, evaluation, or professional standing. To minimize potential coercion, recruitment and consent procedures were conducted separately from routine educational assessments, and participation status was not disclosed to supervising faculty involved in intern evaluation. Written informed consent was obtained from all participants prior to enrollment. After consent, participants were assigned identification numbers and randomly allocated to either the VR or HFS training group using simple randomization.

Development of an Educational Program

Scenario Development

The goal of the program is to enable participants to quickly recognize when a patient's oxygen saturation is dropping, identify the critical factors that need to be assessed, take appropriate action to manage the patient, and promptly recognize when it is necessary to seek additional help.

The simulation program was designed to alter the patient's condition based on the interns' expected responses to critical aspects, such as a disconnected monitor or oxygen line, low oxygen flow or low blood pressure, and when oxygen saturation dropped. To address the need for flexible problem-solving in emergency situations, we developed a scenario-based VR system that allows nonlinear decision-making. Unlike procedural-focused VR simulations, our system presented participants with complex, evolving emergency scenarios that required rapid assessment and adaptable responses. The simulation scenario was refined and enhanced through research team meetings to ensure that it was properly aligned with the educational objectives.

The program included simple and complex scenarios to promote learning through experience. The first scenario involved a patient whose condition was stable, with the primary issue being a disconnected oxygen line resulting in a lower oxygen saturation level. Resolving this issue by reconnecting the oxygen line restored the patient's oxygen saturation to the normal level. The second scenario presented a more challenging situation where both the oxygen line and saturation probe were initially disconnected. Despite reconnecting the oxygen line and saturation probe, there was no improvement in the patient's heart rate or consciousness. This lack of clinical improvement signaled ongoing deterioration despite correction of equipment-related issues, reflecting a critical condition potentially related to worsening of underlying medical problem. In this scenario, the patient remained unstable and required further escalation of care, including activation of the cardiopulmonary resuscitation team.

Development of the VR Simulation Training Program

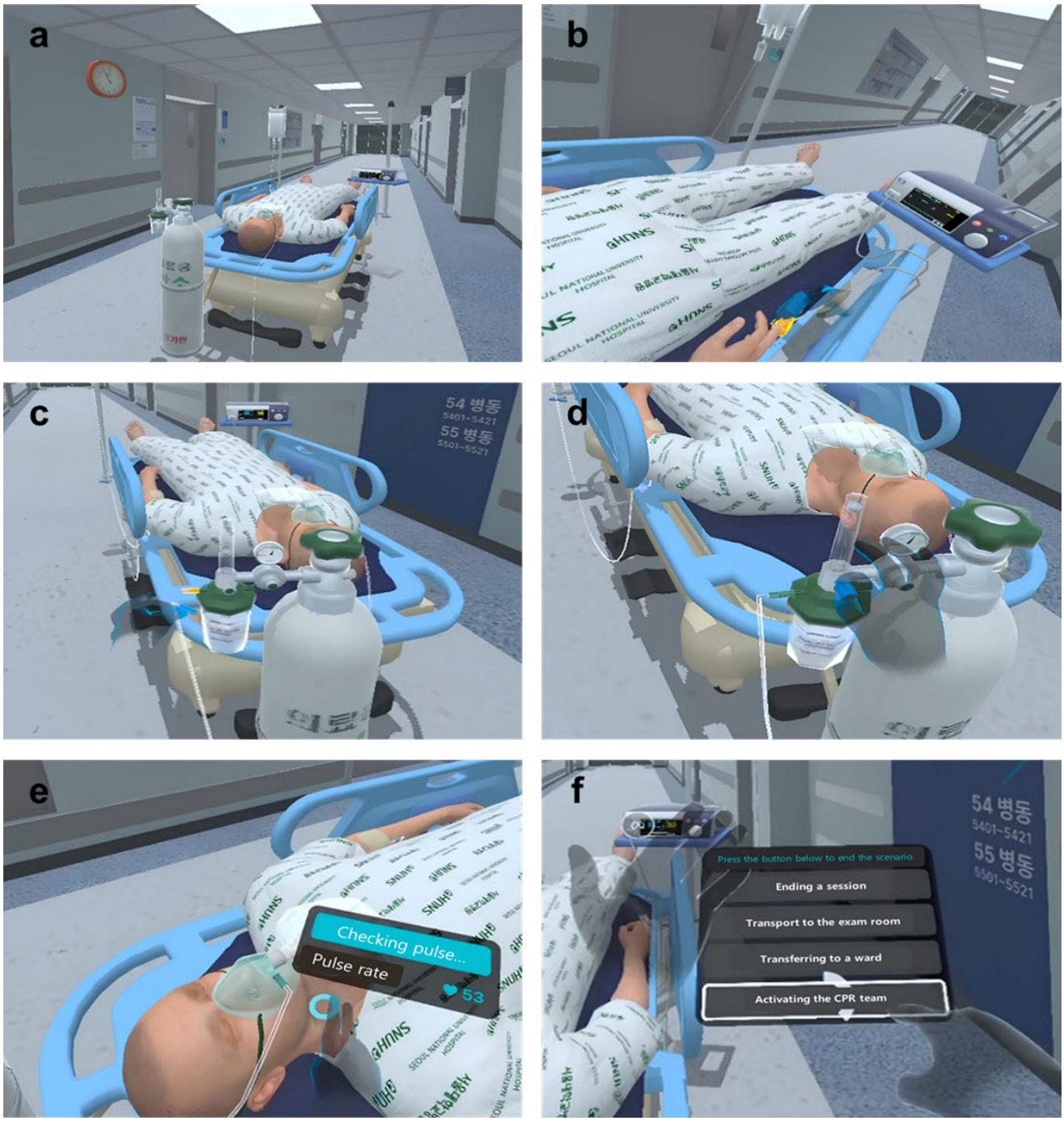
The VR program was developed using the Unity Technologies platform (Unity Software, Inc.) and used immersive VR devices, including the Oculus Quest Pro, the Oculus Quest 3, and the Oculus Quest 2 (Meta Platforms, Inc.). The virtual environment consisted of a hospital hallway, the patients, oxygen tanks, monitors, and saturation probes, which were meticulously crafted using real-time 3D rendering and 3D modeling techniques. This setup provided a first-person perspective, allowing participants to engage in realistic interactions and auditory experiences as if they were physically present in the environment.

The VR scenario focused on training participants to manage patients with suddenly decreasing oxygen saturation, particularly when immediate assistance was not readily available. Participants were required to prioritize checks (eg, oxygen line connectivity, oxygen flow rate, and the patient's level of consciousness) and manage the situation accordingly. The system was designed to facilitate direct interaction, enabling participants to identify and resolve these issues effectively.

XSL•FO
RenderX

Once the participants believed they had completed all of the necessary actions, they were prompted to determine the next appropriate step, such as transferring the patient to the laboratory, transferring them to the ward, or calling the cardiopulmonary resuscitation team (Figure 2).

Figure 2. Scenes captured from a virtual reality simulation program showing how to manage a patient in desaturation. (a) The participant encounters an urgent situation where a monitor alarm sounds during a patient transfer; (b) after confirming that no numbers are displayed on the monitor, the participant places the saturation probe back on securely; (c) if oxygen saturation does not increase, the participant securely attaches the disconnected oxygen line; (d) the participant maximizes the oxygen flow rate; (e) the participant checks the patient's level of consciousness and pulse rate; and (f) the participant chooses a plan depending on whether the patient is recovering, such as calling for help (eg, activating a cardiopulmonary resuscitation team) or proceeding to the exam room as scheduled. CPR: cardiopulmonary resuscitation.



Conventional SimMan ALS Simulation Program

An educational program based on the scenarios developed was programmed into the HFS SimMan ALS simulator (Laerdal Medical). Situations involving an intentionally disconnected O_2 line and saturation monitor probe were manually staged. Vital signs corresponding to these scenarios were programmed and displayed through the SimMan ALS simulator. The intervention required participants to implement appropriate responses to restore the patient's vital signs, which was considered a successful intervention.

XSL•FO
RenderX

Program Evaluation

To evaluate the educational impact and learner experience, we used a combination of standardized and study-specific instruments focusing on 3 primary constructs: self-efficacy (confidence), learner satisfaction, and perceived authenticity.

Self-Efficacy (Confidence)

The primary outcome was the change in self-reported confidence in managing oxygen desaturation, measured on a 10-point Likert scale [19] To ensure the specificity of the training effect, we included "control questions" regarding unrelated emergencies (eg, seizures and anaphylaxis). This allowed us to distinguish between a general increase in confidence and the specific cognitive gains from the simulation.

Study-Specific Questionnaires

Secondary outcomes, including learner satisfaction and perceived educational experience (tutor guidance, training duration), were assessed using questionnaires developed by the research team. These items were adapted from routinely administered institutional course evaluation protocols used in our tertiary teaching hospital to support content relevance.

Perceived Authenticity and Usability

Perceived authenticity was measured using a 7-point Likert scale [20] to evaluate how well each modality mirrored clinical reality—a particularly relevant factor for novice learners. Usability and VR-specific symptoms (eg, motion sickness) were assessed using the User Experience Questionnaire-Short (UEQ-S), a validated 7-point scale [21].

Design and Validity

All study-specific instruments were reviewed by a panel of 3 senior medical education experts (YJH, JWK, and SJM) to support face and content validity regarding the target clinical scenarios. While formal psychometric factor analysis was not performed, the items were designed to descriptively capture the immediate psychological and educational reactions of the interns, consistent with Kirkpatrick Level 1 and 2 evaluations.

To further evaluate the effect of the program, qualitative data were collected through follow-up focus group interviews (FGIs), as detailed below.

Qualitative Methods and Mixed-Methods Integration

We used an embedded qualitative approach within the randomized controlled trial to explore participants' perceptions of learning transfer to clinical practice (Kirkpatrick level 3) [14] and to contextualize the quantitative outcomes. Reporting was guided by established qualitative reporting standards (eg, COREQ), with adaptations appropriate for an embedded mixed-methods design.

Sampling and Recruitment

Interns from both trial arms and tutors who facilitated the sessions were eligible. We invited eligible participants after the training and again at follow-up time points; participation was voluntary. FGIs were conducted at 1 month (interns and tutors) and 5 months (interns only) posttraining. At 1 month, 5 interns (VR=3; HFS=2) and 4 tutors participated; at 5 months, 4 interns (VR=2; HFS=2) participated. The number of individuals invited versus those who participated was not systematically recorded in the original study logs.

Data Collection

Semi-structured interview guides were developed by the research team based on the study objectives, prior simulation literature, and the Kirkpatrick framework. Questions explored perceived preparedness, examples of applying the training in real clinical settings, and perceived strengths and limitations of each modality. FGIs were moderated by a facilitator not involved in participant evaluation, audio-recorded, transcribed verbatim, and deidentified.

Analysis

Overall, 2 researchers (YJH and SJM) independently coded transcripts using a concise framework-based thematic analysis. An initial coding framework was informed by Kirkpatrick level 3 constructs (eg, recognition of desaturation, initial actions, escalation, and communication), with additional inductive codes added for emergent themes. Coders met to reconcile differences and refine themes through iterative discussion until consensus was achieved. Themes were then mapped to the Kirkpatrick framework to support interpretation as perceived learning transfer, rather than as objectively verified behavior change; given the small and relatively homogeneous sample, the qualitative findings were presented descriptively to illustrate how the training translated into clinical practice, rather than to draw broadly generalizable conclusions.

Statistical Analysis

The answers from all questionnaires were compared between the VR and HFS groups. The primary outcome was the pre- to posttraining change in self-reported confidence for managing oxygen desaturation. The primary between-group analysis tested the group-by-time interaction for desaturation confidence. Secondary outcomes included overall satisfaction and subdomains, UEQ-S usability scores, and confidence changes for nontargeted scenarios (seizure, anaphylaxis, general care, and general skills). Exploratory analyses included preference data, VR head-mounted display (HMD) log metrics, and the qualitative focus group findings, which were used to contextualize and interpret the quantitative results. Statistical analysis was performed using the SPSS (version 25; IBM Corp) statistical package. Continuous variables were analyzed with the *t* test, and proportions were compared with the chi-square test. Changes in confidence before and after training, as well as in confidence according to the training method, were analyzed using a paired *t* test, repeated-measures analysis of variance, and generalized estimating equations. Effect sizes were calculated using the generalized estimating equations and Cohen *d* values. A *P* value <.05 was considered significant.

Ethical Considerations

This study received approval from the Seoul National University Hospital Institutional Review Board (IRB no. H2402-076-1511) before commencement and was registered at Clinical Trials official.gov website (no. NCT06295887) on 13 February 2024 [22]. This study was conducted following the Declaration of Helsinki, and informed consent was obtained from all eligible

XSL•FO
RenderX

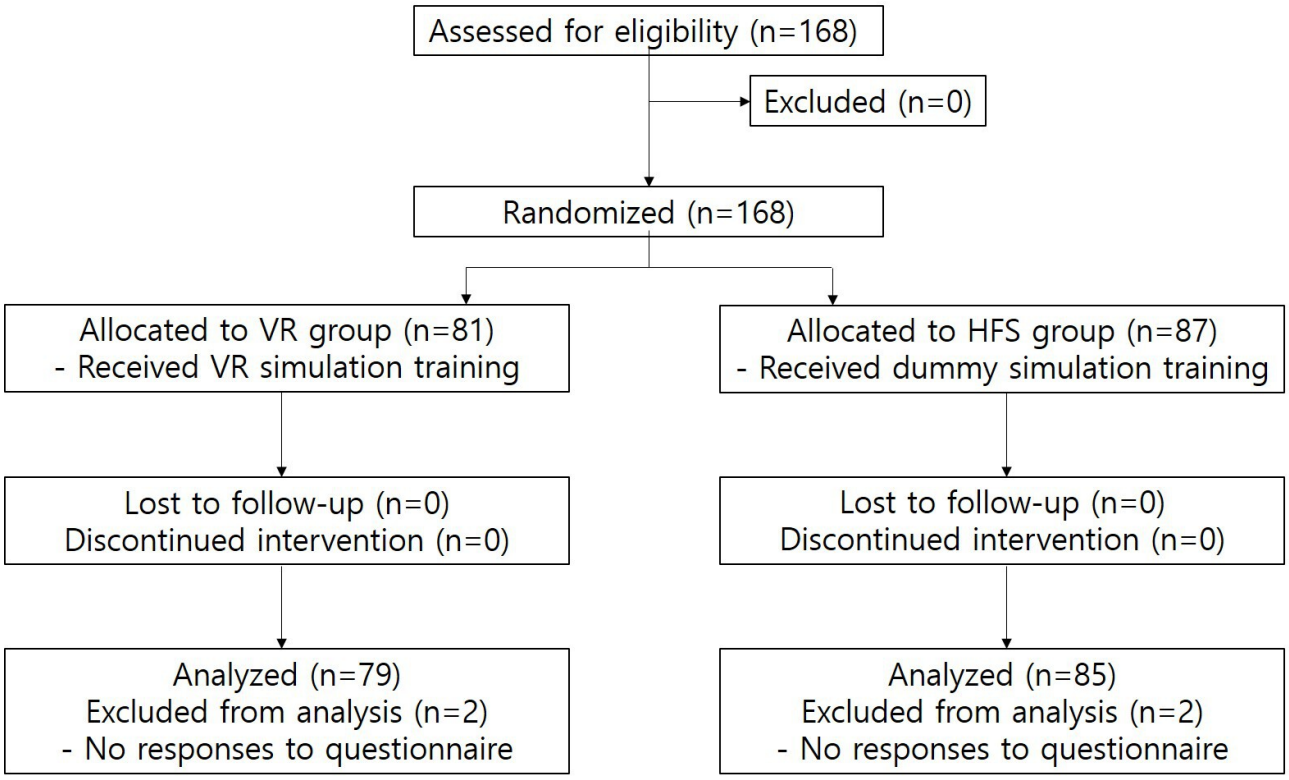
participants. All participant data were anonymized prior to analysis to protect privacy. Confidentiality was maintained through the use of identification codes and secure data storage. No financial compensation was provided for participation.

Results

Participants and Baseline Characteristics

This study included 168 prospective interns who were randomly assigned to either the VR group (n=81) or the HFS group (n=87). No dropouts were reported before randomization. Data from 4 participants who did not complete the presurvey were excluded from the final analysis, resulting in 79 participants in the VR group and 85 in the HFS group (Figure 3). The sex ratio (male:female) was 1.9:1 in the HFS group and 1.4:1 in the VR group, with no significant difference between the groups (P=.34). The mean age did not differ between the groups (P=.29) (Multimedia Appendix 1). All interns had passed the Korean medical licensing examination, meeting a common minimum competence standard, and were selected using comparable pre-entry criteria (eg, medical school grade point average and Korean medical licensing examination written scores).

Figure 3. Flow diagram of the two-group parallel randomized trial. HFS: high-fidelity simulation; VR: virtual reality.



Changes in Confidence Before and After Training

On a 10-point scale, the mean (SD) confidence score for managing oxygen desaturation increased significantly from mean 3.78 (SD 2.12) before training to mean 6.20 (SD 2.02) after training (t_{163}=−14.04; P<.001). Notably, confidence in handling unrelated scenarios, such as seizures, anaphylaxis, general care, and basic skills, also improved, suggesting a broader educational impact. However, the increase in confidence for managing oxygen desaturation was significantly higher than other situations (t_{163}=−13.87; P<.001; Cohen d=−1.08).

As shown in Table 1, both groups demonstrated significant improvements: from mean 3.82 (SD 2.31) to mean 6.54 (SD 1.91) in the HFS group (t_{84}=−10.79; P<.001) and from mean 3.72 (SD 1.91) to mean 5.82 (SD 2.08) in the VR group (t_{78}=−9.15; P<.001). The training method did not significantly affect confidence ($F_{1, 162}$=3.28; P=.07), indicating that we did not detect a statistically significant difference between groups in confidence change within the precision of this study; however, this trial was not designed or powered as a noninferiority study and no noninferiority margin was prespecified. Similarly, confidence in other nontargeted situations also improved in both groups without significant between-group differences.

XSL•FO
RenderX

Table . Differences in self-confidence between training methods.

Confidence domain	Pretraining (N=164), mean (SD)		Posttraining (N=164), mean (SD)		Sources	*P* value
	HFS[a] (n=85)	VR[b] (n=79)	HFS (n=85)	VR (n=79)		
Seizure	3.67 (2.38)	3.38 (1.89)	4.49 (2.02)	4.32 (2.07)	Group	.54
					Time	<.001
					Group*Time	.53
Desaturation[cd]	3.82 (2.31)	3.72 (1.91)	6.54 (1.91)	5.82 (2.08)	Group	.19
					Time	<.001
					Group*Time	.07
Anaphylaxis	4.85 (2.58)	4.67 (2.26)	5.32 (2.08)	5.05 (2.10)	Group	.62
					Time	.004
					Group*Time	.75
General care	4.24 (2.29)	4.18 (1.92)	5.33 (1.88)	5 (1.96)	Group	.59
					Time	<.001
					Group*Time	.41
General skills	4.47 (2.23)	4.14 (1.81)	5.28 (1.87)	4.96 (1.88)	Group	.31
					Time	<.001
					Group*Time	.85

[a]HFS: high-fidelity simulation.

[b]VR: virtual reality.

[c]Rated on a 10-point Likert scale ranging from 1 (strongly disagree) to 10 (strongly agree).

[d]Result of the simulation training offered.

VR simulation was the most frequently preferred training method (63/139, 45.3%), followed by HFS (54/139, 38.8%) and lectures (22/139, 15.8%). Notably, participants' preferences tended to correspond to the training method they were assigned, indicating that postassignment exposure may have shaped their reported preferences rather than reflecting preexisting inclinations. However, concordance between a participant's assigned and preferred training method was not associated with satisfaction (t_{137}=0.21, *P*=.13) or perceived improvement in confidence following training ($F_{1,137}$=0.08, *P*=.45).

Satisfaction With the Educational Program

Participants in both groups reported high levels of satisfaction with the educational program, with an overall mean (SD) score of 6.07 (1.02) out of 7. As shown in Table 2, the HFS group was more satisfied overall than was the VR group (mean 6.23, SD 0.92 vs mean 5.89, SD 1.10; t_{162}=2.29, *P*=.02), particularly regarding tutor guidance (mean 6.49, SD 0.86 vs mean 6.10, SD 1.02; *P*=.008) and authenticity of the educational experience (mean 6.24, SD 1.05 vs mean 5.77, SD 1.15; *P*=.006). However, the overall difference between the groups was minimal, and both groups reported very high levels of overall satisfaction.

Table . Student satisfaction with the simulation education.

Themes[a]	HFS[b] (n=85), mean (SD)	VR[c] (n=79), mean (SD)	*P* value
Overall satisfaction	6.23 (0.92)	5.89 (1.10)	.02
Training duration	6.06 (1.21)	5.81 (1.25)	.20
Tutor's guidance	6.49 (0.86)	6.10 (1.02)	.008
Authenticity	6.24 (1.05)	5.77 (1.15)	.006

[a]Rated on a 7-point Likert scale ranging from 1 (strongly disagree) to 7 (strongly agree).

[b]HFS: high-fidelity simulation.

[c]VR: virtual reality.

The usability was evaluated by the UEQ-S. As shown in Table 3, the analysis of 8 items on a 7-point scale revealed significant differences between the 2 groups on 4 items. The HFS group rated the training as more supportive, easy, efficient, and clear than did the VR group. No significant differences were detected between the groups in terms of excitement, interest, inventiveness, or being cutting-edge. All items had scores above 5, indicating the usability of both training methods.

XSL•FO
RenderX

Table . Results of User Experience Questionnaire-Short version.

Negative[a] (1)	HFS[b] (n=85), mean (SD)	VR[c] (n=79), mean (SD)	*P* value	Positive[a] (7)
Obstructive	5.97 (0.96)	5.62 (1.09)	.03	Supportive
Complicated	5.45 (1.13)	5.05 (1.27)	.04	Easy
Inefficient	5.83 (1.00)	5.42 (1.24)	.02	Efficient
Confusing	5.71 (1.31)	5.00 (1.42)	.001	Clear
Boring	6.00 (1.23)	5.96 (1.18)	.84	Exciting
Not interesting	5.99 (1.23)	5.99 (1.20)	.99	Interesting
Conventional	5.33 (1.18)	5.69 (1.30)	.06	Inventive
Usual	5.31 (1.17)	5.67 (1.41)	.08	Leading edge

[a]Rated on a 7-point Likert scale ranging from 1 (strongly negative) to 7 (strongly positive).

[b]HFS: high-fidelity simulation.

[c]VR: virtual reality.

Analysis of the VR Group Log Data

Further analysis was conducted using the log data recorded from the VR group on a HMD. Usable data were available for 32 simple and 39 complex cases, with 29 participants having both. The logs captured task duration, completion of key actions (eg, checking oxygen connections, confirming pulse oximeter clip placement, adjusting oxygen flow rate, assessing consciousness, and checking pulse), and final decision time.

Participants spent an average (SD) of 151.05 (60.94) seconds on the simple case and 132.95 (SD 67.58) seconds on the complex case, showing no significant difference (t_{69}=1.17, *P*=.11). The average number of tasks completed out of the 6 subtasks was 3.19 in the simple case and 4.33 in the complex case (t_{69}=−3.02, *P*=.002). Correlation and multiple regression analysis revealed no significant relationships between task duration, number of tasks completed, satisfaction, or confidence predicted posttraining confidence except pretraining confidence (Multimedia Appendix 2).

Focus Group Interview

Simulation-based education aims not only to enhance performance within controlled scenarios but also to improve the transfer of skills to actual clinical practice. To explore this transfer, FGIs were conducted with the participants at 1 and 5 months after the training. In addition, FGIs were also conducted with the tutors at 1 month after the program to gather feedback on the program's implementation and operation (Table 4).

XSL•FO
RenderX

Table . Summary of the focus group interviews on simulation education.

Group and category	Illustrative quotations
Intern	
Advantages of the education program	• "It provided a realistic and authentic experience closely resembling what to do in actual clinical situations, such as when a patient's condition changes suddenly." (VR[a] and HFS[b]) • "I was really panicked, and it felt like a real situation because the alarms were going off and I felt like I was the only one there." (VR and HFS) • "The detailed and realistic depiction of clinical changes, such as pallor as oxygen saturation changes, were immersive." (VR) • "It was helpful to learn step by step on what to check and do first in an emergency situation." (VR) • "Interactions were only created for the items that needed to be checked, and having this option made learning the procedure fast and efficient." (VR) • "I could repeat the learning as many times as I needed." (VR) • "We were able to learn at our own pace." (VR) • "With no fixed answers, the learning felt more like real situations and allowed for greater freedom." (HFS) • "I found the interactive approach to be more helpful than the lectures." (HFS) • "Being able to train with real equipment allowed for experiential learning." (HFS)
Limitations and areas for improvement	• "Because the simulations lacked realistic manipulation, I think additional training with real equipment is necessary." (VR) • "The preset choices limited interaction and reduced the flexibility of the learning process." (VR) • "The system's responsiveness was a bit low, making it difficult to perform certain tasks." (VR) • "My HMD[c] device malfunctions caused inconvenience." (VR) • "I think mannequins are limited in their capability to fully depict detailed scenarios." (HFS) • "The high cost and limited availability of equipment made it hard to access when needed." (HFS) • "I couldn't progress at my own pace during the training." (HFS) • "Not all interns were able to participate directly, so it would be great if there were more opportunities for them to freely join and experience the training. To achieve this, the tutor should minimize direct involvement and focus primarily on a supervising role." (HFS) • "Scenarios with various approaches, where the participant's actions determine the patient's recovery, are necessary." (VR and HFS)

XSL•FO
RenderX

Group and category	Illustrative quotations
The usefulness of the education program in real clinical situations	• <One month after the training> “In real situations where a patient’s oxygen saturation was dropping, I knew what to check and was able to assess the critical factors quickly.” (VR and HFS) • “I developed the habit of checking the oxygen tank before patient transport.” (HFS) • “While transporting a patient, the patient’s oxygen saturation suddenly dropped, and instead of blindly calling for CPR[d], I was able to determine that the oxygen tank had run out of oxygen and resolve it by transporting the patient directly to the nearest ward.” (VR) • “I felt more confident and less anxious in real situations.” (VR and HFS) • <Five months after the training> “While transporting a patient with a ventilator, the ventilator alarm kept going off, and when I checked the patient’s condition and the line connection, I found that the ventilator’s line was connected incorrectly, so I fixed the misconnection, and the problem was resolved.” (VR) • “When the patient’s oxygen saturation suddenly dropped, a basic check of the line connections and oxygen levels showed no abnormalities, leading me to believe it was an exacerbation of the patient’s underlying condition, and I was able to take action to improve the patient’s condition, first by maximizing the oxygen flow and then by reporting to a supervisor to insert a percutaneous drainage into the lungs.” (VR) • “I was transporting a critically ill pediatric patient to the MRI[e] room when his oxygen levels began to drop. Despite the many connected lines, I noticed the oxygen line was disconnected. After reconnecting it, his condition improved.” (HFS)
Tutor	

XSL•FO
RenderX

Group and category	Illustrative quotations
Advantages of the education program	• "It was an engaging teaching method that resulted in high tutor satisfaction with the overall training experience." (VR and HFS) • "The training allowed the interns to successfully achieve most of their learning objectives, demonstrating the educational effectiveness of these new approaches." (VR and HFS) • "The 2-stage training program, based on difficulty, helped the interns achieve their learning objectives." (VR and HFS) • "Real-time observation of intern performance was possible." (HFS) • "We were able to provide specific and immediate feedback." (HFS) • "I was able to deliver the same training to a large group of interns simultaneously." (VR) • "Following predeveloped educational materials for individual learning made things easier for me as an instructor." (VR)
Limitations and areas for improvement	• "Since tutors were unable to observe each interns' performance in real time, we had to provide collective feedback to interns at the end of the training." (VR) • "We need a system for real-time monitoring, or automated checklist systems to easily keep an eye on how well the interns are performing." (VR) • "We need a multiplayer platform to make peer learning possible." (VR) • "It was challenging to provide repeated learning and deliver training tailored to the intern's level." (HFS) • "The current system, which requires tutors to manually change system settings that change as the scenario progresses, needs improvement." (HFS) • "Some of the interns actually participated in the simulated role-plays, while others remained as observers, so we need to make sure that the training is delivered in a way that engages as many interns as possible." (HFS) • "We need dedicated scenarios to fully use the diverse functions of HFS." (HFS) • "The training methods should be adjusted based on the content being taught." (VR and HFS)

[a]VR: virtual reality.

[b]HFS: high-fidelity simulation.

[c]HMD: head-mounted display.

[d]CPR: cardiopulmonary resuscitation.

[e]MRI: magnetic resonance imaging.

FGI at 1 Month

In total, 5 interns (3 in VR and 2 in HFS) and 4 tutors participated. Interns from both groups described the distinct benefits of their respective training methodologies in preparing them for real-world clinical practice. They also highlighted the value of training in highly controlled yet realistic environments, which appeared to alleviate anxiety and deepen their comprehension of critical desaturation management protocols. VR participants noted the unparalleled immersive and focused nature of the environment, whereas HFS participants appreciated the hands-on realism and flexibility that mirrored real clinical uncertainty. Both approaches were seen as complementary: VR appeared to foster concentration and repeated practice, while HFS seemed to promote situational adaptability.

Regarding the participant's comment on the efficiency of focused interactions in VR, this reflects the contingent feedback mechanism inherent in both training modalities. In both VR and HFS, the simulation was designed so that when a learner identified a clinical issue and implemented the appropriate intervention, the system provided an immediate physiological reaction, such as stabilization of vital signs or a change in the patient's status. This equivalence in responsiveness suggests that, regardless of the medium, the learner experiences a consistent "action-reaction" loop.

The tutors expressed overall satisfaction with the training outcomes and process, observing that both the VR and HFS programs effectively supported interns in achieving their learning goals. They praised VR for scalability and standardized materials, though they pointed out the absence of real-time monitoring and immediate, personalized feedback. HFS enabled direct observation and instant feedback but was limited by resource demands and unequal engagement among interns. Tutors suggested the future integration of automated checklists and multiplayer platforms to enhance feedback and encourage peer learning.

FGI at 5 Months

A follow-up FGI with 4 interns (2 in VR and 2 in HFS) explored how the participants perceived their ability to handle

XSL•FO
RenderX

emergencies in real clinical settings, as part of the behavior change evaluation at the third level of the Kirkpatrick model [14]. The interns reported that they were able to effectively manage oxygen desaturation events by systematically assessing oxygen delivery and identifying underlying causes, including complex cases such as ventilator-related desaturation. They reported sustained improvements in their perceived ability to handle desaturation emergencies months after the training (median self-rating scores on a 10-point scale: 0.75 (range 0 - 2) before training, 3.25 (range 2 - 5) at 1 month, and 7.25 (range 5.5 - 8) at 5 month), suggesting potential behavioral transfer of learning to practice.

Discussion

Principal Findings

This study compared the educational effectiveness of VR and HFS training in preparing medical interns to manage oxygen desaturation. Both methods significantly improved interns' self-reported confidence and perceived preparedness, with no substantial difference in overall confidence change between groups. Pretraining preference favored VR education, whereas posttraining satisfaction and perceived authenticity were higher in the HFS group.

Although this study primarily assessed confidence rather than objective competence, confidence in managing acute emergencies can reflect psychological readiness and decision-making under pressure. In high-stress situations such as hypoxia, the ability to apply existing knowledge swiftly often determines performance more than technical skill alone. The focus on oxygen desaturation was intentional, emphasizing cognitive and situational judgment rather than procedural execution. VR simulation aimed to help learners practice rapid assessment and response under stress, strengthening cognitive readiness and composure rather than manual skills. Although increased confidence does not equate to competence, it may reasonably reflect perceived preparedness to act effectively in real clinical contexts.

Comparison With the Literature

Consistent with prior simulation-based education literature [23-25], the significant increase in self-confidence observed in the VR and HFS training groups underscores the value of simulation-based education. In contrast, satisfaction and perceived authenticity were greater with HFS. Although VR more easily reproduces challenging clinical scenarios [26-28], participants perceived higher realism when using actual medical equipment in HFS. Qualitative feedback provided deeper insights into these findings: interns reported that VR training enhanced their composure and independent decision-making in real situations, as immersive warning sounds and the requirement to manage scenarios alone closely mirrored real-life experiences. This perception may reflect differences in how learners experienced the interaction structure, which was implemented similarly in both VR and HFS but may have been more salient in the VR experience. However, the participants noted that the experience would be enhanced by the ability to physically interact with actual equipment, suggesting opportunities to improve VR realism through mixed or augmented reality and haptic feedback.

Implications of the Findings

VR education engages learners in an active and central role, but it also brings challenges during debriefing and feedback, key components of simulation-based education. Although the overall structure of debriefing was comparable across groups, learners experienced feedback differently during the simulation. In the HFS group, instructors could directly observe learners' actions and errors in real time, which may have enabled more specific and context-rich feedback during debriefing, representing one of the major strengths of HFS. In contrast, real-time mirroring was not possible in the VR sessions due to the use of HMDs. Instead, postscenario feedback was provided through structured summaries and predefined prompts, which offered flexibility but may have provided less depth than informed by direct instructor observation in HFS settings. Interestingly, learners did not express dissatisfaction with VR debriefing, suggesting that immersion and autonomy may have compensated for limited instructor feedback. Because both groups received structured debriefing as part of the educational intervention, this shared component may have reduced observable differences between VR and HFS, potentially diluting the treatment effect specific to the simulation modality itself.

Future VR models should consider incorporating hybrid approaches, such as AI-assisted feedback, real-time instructor input via integrated communication tools, or structured peer discussions, to enhance reflective learning and clinical reasoning. Furthermore, it is important to recognize that cultural factors may influence the effectiveness and format of debriefing; learners in high-power distance cultures may hesitate to participate openly unless psychological safety is deliberately structured. As noted in prior studies [29,30], culturally adapted debriefing frameworks could help optimize the educational impact of VR-based simulation across diverse settings. Further research is required to clarify the impact of debriefing methods in VR-based training, as current evidence remains inconclusive [31].

Limitations of the Study

This study had several limitations. First, the trial was conducted at a single tertiary academic center within a specific institutional and cultural context, and participants were a relatively homogeneous cohort of incoming interns, which limits the generalizability of the findings. Because this study was not designed as a noninferiority or equivalence trial, the absence of statistically significant between-group differences should not be interpreted as evidence of equivalence between VR and HFS. Additionally, practical considerations, such as cost of VR implementation, hardware maintenance, and potential usability challenges in high-pressure clinical environments, may constrain the broader adoption of VR-based training. Multicenter studies across diverse training environments and learner populations will be necessary to validate and extend these findings.

Second, the primary quantitative outcomes focused on self-reported confidence and satisfaction rather than objectively measured performance or long-term retention. Although an

XSL•FO
RenderX

objective structured clinical examination was initially planned to evaluate posttraining performance, it was not feasible during the study period [32]. Accordingly, the qualitative findings should be interpreted as perceived learning transfer rather than objectively verified behavior change, and their interpretive value is limited by the small, exploratory sample. While this approach provided insight into learners' perceived preparedness in relatively straightforward emergency scenarios, it limits conclusions regarding actual performance improvement. Future studies should incorporate objective performance measures, structured observer-rated assessments, longer-term follow-up, and mixed-method designs to more rigorously evaluate learning outcomes and behavioral change over time.

Third, while debriefing occurred after scenario completion in both groups, differences in instructors' ability to observe learners during the simulation may have influenced the depth of feedback provided during debriefing. Although both modalities shared a similar action–reaction feedback structure, the structured interaction design of the VR environment may have shaped how learners perceived guidance and autonomy, which should be considered when interpreting differences in learner experience. Direct, real-time feedback during the simulation was limited in the VR group, which could be addressed in future studies by integrating mirroring or hybrid feedback methods. Finally, participants' previous experience with VR was not formally measured; collecting these data would improve the accuracy and interpretation of the results.

Conclusions

In conclusion, both VR and HFS improved interns' self-reported confidence and perceived preparedness for managing acute oxygen desaturation. Although satisfaction was higher with HFS, we did not detect a statistically significant between-group difference in confidence change. VR may complement HFS as a scalable educational option in settings where access to mannequin-based simulation is limited. Future work should evaluate objective performance outcomes, longer-term retention, implementation feasibility, and generalizability across diverse institutions and learner groups.

Acknowledgments

We would like to thank the research team from the Department of Convergence Medicine at Seoul National University Hospital for leading the simulation development. This manuscript was assisted by ChatGPT (OpenAI) for language editing and grammar refinement. All artificial intelligence-generated suggestions were reviewed and edited by the authors. The use of generative artificial intelligence was limited to language-related tasks and did not involve data analysis, interpretation of results, or decision-making regarding the manuscript.

Funding

This study was supported by the Seoul National University Hospital Research Fund (to S.J.M., Grant No. 04-2024-0440).

Data Availability

The data that support the findings of this study are available from the corresponding author upon reasonable request.

Authors' Contributions

Conceptualization and supervision: HJK and SJM
Study design, study execution, data analysis, and discussion of results: YJH, JWK, NY, SJ, HJK, and SJM
Writing–original draft and review and editing: YJH, JWK, and SJM
Figures and tables: YJH
Virtual reality program development and writing–methodology: SJ and HJG

Yeon-Ju Huh and Ju Whi Kim contributed equally as co–first authors. Sun Jung Myung and Hyoun-Joong Kong contributed as co–corresponding authors.
Hyoun-Joong Kong, Department of Transdisciplinary Medicine and Innovative Medical Technology Research, Seoul National University Hospital, Seoul, Republic of Korea, Department of Medicine, Seoul National University College of Medicine, Seoul, Republic of Korea, Tel:82-2-2072-4492, Fax: 82-2-6072-5304, E-mail: gongcop7@snu.ac.kr

Conflicts of Interest

None declared.

Multimedia Appendix 1
Baseline characteristics.
[DOC File, 67 KB - mededu_v12i1e71455_app1.doc]

Multimedia Appendix 2
Correlation of time spent and task resolution scores by case extracted from log data with posttraining confidence change and satisfaction with training.

XSL•FO
RenderX

[DOC File, 69 KB - mededu_v12i1e71455_app2.doc]

Checklist 1
CONSORT-eHEALTH checklist v1.6.1 and Clinical trial protocol.
[PDF File, 2173 KB - mededu_v12i1e71455_app3.pdf]

References

1. John TP, Shirley CS, Deborah DG, Laura SB. Comprehensive Healthcare Simulation: InterProfessional Team Training and Simulation: Springer International Publishing; 2020. [doi: 10.1007/978-3-030-28845-7]
2. Issenberg SB, McGaghie WC, Petrusa ER, Lee Gordon D, Scalese RJ. Features and uses of high-fidelity medical simulations that lead to effective learning: a BEME systematic review. Med Teach 2005 Jan;27(1):10-28. [doi: 10.1080/01421590500046924] [Medline: 16147767]
3. Littlewood KE, Shilling AM, Stemland CJ, Wright EB, Kirk MA. High-fidelity simulation is superior to case-based discussion in teaching the management of shock. Med Teach 2013;35(3):e1003-e1010. [doi: 10.3109/0142159X.2012.733043] [Medline: 23126242]
4. DCosta S, Zadow G, Reidlinger DP, et al. The impact of moulage on learners' experience in simulation-based education and training: systematic review. BMC Med Educ 2024 Jan 3;24(1):6. [doi: 10.1186/s12909-023-04976-w] [Medline: 38172859]
5. Sørensen JL, Østergaard D, LeBlanc V, et al. Design of simulation-based medical education and advantages and disadvantages of in situ simulation versus off-site simulation. BMC Med Educ 2017 Jan 21;17(1):20. [doi: 10.1186/s12909-016-0838-3] [Medline: 28109296]
6. Tacgin Z. Enhanced Learning Environments: Technology and Innovation Location: Cambridge Scholars Publishing; 2022. [doi: 10.65325/EB10151]
7. Eppich W, Reedy G. Advancing healthcare simulation research: innovations in theory, methodology, and method. Adv Simul (Lond) 2022 Jul 27;7(1):23. [doi: 10.1186/s41077-022-00219-y] [Medline: 35897062]
8. Gupta S, Wilcocks K, Matava C, Wiegelmann J, Kaustov L, Alam F. Creating a successful virtual reality-based medical simulation environment: tutorial. JMIR Med Educ 2023 Feb 14;9:e41090. [doi: 10.2196/41090] [Medline: 36787169]
9. Pottle J. Virtual reality and the transformation of medical education. Future Healthc J 2019 Oct;6(3):181-185. [doi: 10.7861/fhj.2019-0036] [Medline: 31660522]
10. Liaw SY, Chan SWC, Chen FG, Hooi SC, Siau C. Comparison of virtual patient simulation with mannequin-based simulation for improving clinical performances in assessing and managing clinical deterioration: randomized controlled trial. J Med Internet Res 2014 Sep 17;16(9):e214. [doi: 10.2196/jmir.3322] [Medline: 25230684]
11. Kim HY, Kim EY. Effects of medical education program using virtual reality: a systematic review and meta-analysis. Int J Environ Res Public Health 2023 Feb 22;20(5):3895. [doi: 10.3390/ijerph20053895] [Medline: 36900904]
12. Mergen M, Graf N, Meyerheim M. Reviewing the current state of virtual reality integration in medical education - a scoping review. BMC Med Educ 2024 Jul 23;24(1):788. [doi: 10.1186/s12909-024-05777-5] [Medline: 39044186]
13. Tursø-Finnich T, Jensen RO, Jensen LX, Konge L, Thinggaard E. Virtual reality head-mounted displays in medical education: a systematic review. Simul Healthc 2023 Feb 1;18(1):42-50. [doi: 10.1097/SIH.0000000000000636] [Medline: 35136005]
14. Kirkpatrick DL, Kirkpatrick JD. Evaluating Training Programs: The Four Levels, 3rd edition 2006.
15. Taylor DCM, Hamdy H. Adult learning theories: implications for learning and teaching in medical education: AMEE Guide No. 83. Med Teach 2013 Nov;35(11):e1561-e1572. [doi: 10.3109/0142159X.2013.828153] [Medline: 24004029]
16. Park S, Hur HK, Chung C. Learning effects of virtual versus high-fidelity simulations in nursing students: a crossover comparison. BMC Nurs 2022 Apr 27;21(1):100. [doi: 10.1186/s12912-022-00878-2] [Medline: 35473614]
17. Clinical trial randomization tool. National Cancer Institute (NIH). URL: https://ctrandomization.cancer.gov/tool/ [accessed 2026-02-13]
18. Schulz KF, Altman DG, Moher D, CONSORT Group. CONSORT 2010 statement: updated guidelines for reporting parallel group randomised trials. BMJ 2010 Mar 23;340:c332. [doi: 10.1136/bmj.c332] [Medline: 20332509]
19. Miller KA, Monuteaux MC, Roussin C, Nagler J. Self-confidence in endotracheal intubation among pediatric interns: associations with gender, experience, and performance. Acad Pediatr 2019;19(7):822-827. [doi: 10.1016/j.acap.2019.06.013] [Medline: 31229695]
20. Ho IMK, Cheong KY, Weldon A. Predicting student satisfaction of emergency remote learning in higher education during COVID-19 using machine learning techniques. PLoS ONE 2021;16(4):e0249423. [doi: 10.1371/journal.pone.0249423] [Medline: 33798204]
21. Schrepp M, Hinderks A, Thomaschewski J. Design and Evaluation of a Short Version of the User Experience Questionnaire (UEQ-S). IJIMAI 2017;4(6):103-108. [doi: 10.9781/ijimai.2017.09.001]
22. ClinicalTrials.gov. US National Library of Medicine (NLM). URL: https://clinicaltrials.gov/study/NCT06295887 [accessed 2026-02-13]

XSL•FO
RenderX

23. Swamy M, Sawdon M, Chaytor A, Cox D, Barbaro-Brown J, McLachlan J. A study to investigate the effectiveness of SimMan® as an adjunct in teaching preclinical skills to medical students. BMC Med Educ 2014 Nov 19;14:231. [doi: 10.1186/1472-6920-14-231] [Medline: 25410815]
24. Ten Eyck RP, Tews M, Ballester JM. Improved medical student satisfaction and test performance with a simulation-based emergency medicine curriculum: a randomized controlled trial. Ann Emerg Med 2009 Nov;54(5):684-691. [doi: 10.1016/j.annemergmed.2009.03.025] [Medline: 19394113]
25. Steadman RH, Coates WC, Huang YM, et al. Simulation-based training is superior to problem-based learning for the acquisition of critical assessment and management skills. Crit Care Med 2006 Jan;34(1):151-157. [doi: 10.1097/01.ccm.0000190619.42013.94] [Medline: 16374169]
26. Coffey F, Tsuchiya K, Timmons S, Baxendale B, Adolphs S, Atkins S. Simulated patients versus manikins in acute-care scenarios. Clin Teach 2016 Aug;13(4):257-261. [doi: 10.1111/tct.12425] [Medline: 26597094]
27. McFetrich J. A structured literature review on the use of high fidelity patient simulators for teaching in emergency medicine. Emerg Med J 2006 Jul;23(7):509-511. [doi: 10.1136/emj.2005.030544] [Medline: 16794089]
28. Quilici AP, Bicudo AM, Gianotto-Oliveira R, Timerman S, Gutierrez F, Abrão KC. Faculty perceptions of simulation programs in healthcare education. Int J Med Educ 2015 Nov 22;6:166-171. [doi: 10.5116/ijme.5641.0dc7] [Medline: 26590951]
29. Cheng A, Eppich W, Grant V, Sherbino J, Zendejas B, Cook DA. Debriefing for technology-enhanced simulation: a systematic review and meta-analysis. Med Educ 2014 Jul;48(7):657-666. [doi: 10.1111/medu.12432] [Medline: 24909527]
30. Rudolph JW, Simon R, Dufresne RL, Raemer DB. There's no such thing as "nonjudgmental" debriefing: a theory and method for debriefing with good judgment. Simul Healthc 2006;1(1):49-55. [doi: 10.1097/01266021-200600110-00006] [Medline: 19088574]
31. Brown KM, Swoboda SM, Gilbert GE, Horvath C, Sullivan N. Curricular integration of virtual reality in nursing education. J Nurs Educ 2023 Jun;62(6):364-373. [doi: 10.3928/01484834-20230110-01] [Medline: 36701128]
32. Yoon JH, Kwon IH, Park HW. The South Korean health-care system in crisis. Lancet 2024 Jun 15;403(10444):2589. [doi: 10.1016/S0140-6736(24)00766-9] [Medline: 38879246]

Abbreviations

CONSORT: Consolidated Standards of Reporting Trials
FGI: focus group interview
HFS: high-fidelity simulation
HMD: head-mounted display
UEQ-S: User Experience Questionnaire-Short
VR: virtual reality

Edited by B Lesselroth; submitted 18.Jan.2025; peer-reviewed by FH Leung, P Spronk; revised version received 29.Jan.2026; accepted 02.Feb.2026; published 19.Mar.2026.

Please cite as:
Huh YJ, Kim JW, Yoon N, Jeong S, Kong HJ, Myung SJ
Immersive Virtual Reality Training to Improve Novice Physicians' Emergency Response Skills: Randomized Controlled Trial
JMIR Med Educ 2026;12:e71455
URL: https://mededu.jmir.org/2026/1/e71455
doi:10.2196/71455

© Yeon-Ju Huh, Ju Whi Kim, Narae Yoon, Seoi Jeong, Hyoun-Joong Kong, Sun Jung Myung. Originally published in JMIR Medical Education (https://mededu.jmir.org), 19.Mar.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Education, is properly cited. The complete bibliographic information, a link to the original publication on https://mededu.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Application of Mixed Reality for Ophthalmic Clinical Skills and Diagnosis: Prospective Study

Chun Jin Marcus Tan[1,2*], MBBS, MMed, MTech; Wei Wei Dayna Yong[1*], MBBS; Hui'En Hazel Anne Lin[1], MBBS, MMed; Jaslyn Oh[1], MBBS; How Sheng Rubin Yong[2], BSc, PhD; Fang Mei Jayme Khew[1], BSc; Liang Shen[3], MSc, PhD; Yujia Gao[4], MBBS, MRCS, DFD, MMed; Wei Chieh Alfred Kow[4], MBBS, MMed, MRCS; Yih Chung Tham[2], PhD; Dianbo Liu[2], PhD; Ching-Yu Cheng[2], MD, MPH, PhD; Kee Yuan Ngiam[5], MBBS, MRCS, MMed; Yew Sen Yuen[1], MBBS, MMed; Ray Manotosh[1], MBBS, MD; Eng Tat Khoo[6], PhD; Teck Chang Victor Koh[1,2*], MBBS, MSc, MRCS, MMed; Woon Teck Clement Tan[1*], MBBS, MMed

[1]Department of Ophthalmology, National University Hospital, 5 Lower Kent Ridge Road, Singapore, Singapore

[2]Centre for Innovation and Precision Eye Health, Yong Loo Lin School of Medicine, National University of Singapore, Singapore, Singapore

[3]Biostatistics Unit, Yong Loo Lin School of Medicine, National University of Singapore, Singapore, Singapore

[4]Division of Hepatobiliary & Pancreatic Surgery, National University Hospital, Singapore, Singapore

[5]Division of General Surgery (Thyroid & Endocrine Surgery), National University Hospital, Singapore, Singapore

[6]Engineering Design and Innovation Centre, National University of Singapore, Singapore, Singapore

[*]these authors contributed equally

Corresponding Author:
Chun Jin Marcus Tan, MBBS, MMed, MTech
Department of Ophthalmology, National University Hospital, 5 Lower Kent Ridge Road, Singapore, Singapore

Abstract

Background: Mixed reality has the potential to transform delivery of medical education. With tools such as HoloLens 2, educators can create immersive, interactive simulations that enable students to practice and engage with real-world scenarios in a controlled environment.

Objective: We postulated that a hybrid ophthalmology curriculum incorporating EyelearnMR (a simulation application) would be noninferior to traditional teaching. We compared learning outcomes and obtained user feedback.

Methods: This was a single-blind, cluster-randomized prospective study. Fourth-year medical students were organized into batches and then assigned to 2 groups: EyelearnMR and control arms. We used a quasi-randomized design with alternation allocation based on clinical grouping. The intervention group had an additional 2 hours of practice with the EyelearnMR devices. During the second week of their posting, a video assessment (5 scenarios with 17 multiple-choice questions) was conducted for both groups—mid-posting for the intervention group and at the end of the posting for the control group. The rationale for assessing the intervention group earlier, in addition to setting a higher bar for EyelearnMR, was to allow for the provision of outcomes showing noninferiority between both groups. In the event of noninferiority, we could demonstrate that EyelearnMR can replace some degree of traditional clinical teaching, even with a shorter total clinical exposure time. Students in the control group were allowed to experience the Eyelearn MR modules for 2 hours at the end of the posting. Both groups were asked to complete the User Experience Questionnaire.

Results: This study was funded in February 2023, and recruitment took place from July 2023 to January 2024. A total of 54 students were recruited—24 (44.4%) in the control arm and 30 (55.6%) in the EyelearnMR arm. The EyelearnMR group performed significantly better than the control group (median scores of 16, IQR 15-17, and 15, IQR 14-15, respectively; P=.03; Mann-Whitney U test). A total of 100% (30/30) of the students in the EyelearnMR group scored full marks (3/3) for the technique portion, compared to 70.8% (17/24) of the students in the control group (P=.002). There was no statistically significant difference between the groups for the examination (P=.13) and pathology (P=.33) portions. This was despite the EyelearnMR group having a reduced clinical time of 7 days compared to 10 days in the control group. The User Experience Questionnaire showed positive evaluations for attractiveness (mean 1.413, SD 0.969), efficiency (mean 0.822, SD 1.068), dependability (mean 1.087, SD 0.801), stimulation (mean 1.577, SD 0.845), and novelty (mean 1.606, SD 0.967).

Conclusions: EyelearnMR with traditional teaching was noninferior to traditional teaching alone. It provided a comparable experience and supported learning objectives equally. It is an effective supplementary teaching tool in ophthalmic education and may confer additional learning benefits beyond a traditional clinical posting.

XSL•FO
RenderX

(JMIR Med Educ 2026;12:e71338) doi:10.2196/71338

KEYWORDS

mixed reality; medical education; HoloLens; ocular motility; ophthalmology teaching

Introduction

Mixed reality (MR) is a new concept that merges the real and virtual worlds [1,2], having the potential to transform the delivery of medical education [3]. This technology could address challenges faced in the delivery of high-quality medical education, including accessibility, consistency, quality, and cost [4,5]. The Microsoft HoloLens 2 is a commercially available MR headset that allows for bidirectional communication with multiple remote users via video, voice, and MR composites [6]. This technology has been used previously in various clinical and educational scenarios, including anatomical teaching, perioperative planning, and surgical training [4,5,7-16]. This technology has also been used to augment and supplement medical student clinical skill teaching and examination [17-19], allowing for the creation of immersive multisensory (audio, visual, and tactile) content to replicate real-world scenarios. MR allowed students to practice techniques and refine their skills in a controlled environment, leading to fewer errors and increased confidence in real-world scenarios.

The traditional methods of teaching students how to perform an eye movement examination and detect abnormalities of ocular motility include clinical demonstrations, slide and video presentations, and case-based teaching in the wards and clinics. The last scenario, while the most realistic, is also opportunistic and inconsistent.

EyelearnMR is an MR simulation application designed to teach and assess ocular examination via a wearable headset. The application consists of a hologram 3D patient, a Lang fixation cube for ocular motility examination, and an animation logic that models eye movement when the patient's eyes are fixating on the cube. It is deployed on the Microsoft HoloLens 2 device and offers a 3D superimposition of simulated patients in a physical space (Figure 1). MR simulation enables technology-enhanced learning, which provides opportunities for trainees to acquire, develop, and maintain knowledge, skills, and behavior via experiential learning. MR simulation in medical education has several advantages, including precise tracking of quantifiable performance parameters in a clinical examination and consistent reproduction of clinical signs without fatiguing a volunteer patient. It can also be readily deployable in large class sizes outside of a traditional classroom with fewer physical constraints.

XSL•FO
RenderX

Figure 1. Screen capture of (A) the instructor's view from EyelearnMR and (B) the user's view from EyelearnMR.



A review by Cook et al [20] showed that simulation-based learning can achieve comparable outcomes to those of traditional methods in other fields of medical education. As such, we came up with a study that allowed us to evaluate the feasibility and reception of an MR-based ophthalmology curriculum in honing ophthalmic skills, providing evidence to support the wider adoption of MR in medical curricula. Moreover, the introduction of the MR-based ophthalmology curriculum could help address challenges in ophthalmology training posed by an increasing number of students, limited faculty tutors, and reduced patient exposure.

In this study, we aimed to compare students' performance on a bespoke video-based assessment between a curriculum incorporating EyelearnMR and a traditional clinical teaching–only curriculum. The assessment evaluated 3 key dimensions of ocular motility learning: examination technique, recognition of clinical signs, and diagnostic reasoning. We postulated that a hybrid ophthalmology curriculum incorporating EyelearnMR would be noninferior to traditional teaching. In addition, we evaluated student perceptions of the MR platform through the validated User Experience Questionnaire (UEQ).

Methods

Overview

This was a single-blind, cluster randomized prospective study. We included undergraduate medical students (fourth year and 10 clinical groups in total; age range 23 to 26 years) posted to the Department of Ophthalmology, National University Hospital, Singapore, for their 2-week ophthalmology rotation. The clinical groups (each group comprised 6-7 students) were randomized into 2 arms (5 clinical groups randomized into the EyelearnMR arm and the other 5 clinical groups randomized into the control arm). We used the quasi-randomized design using alternation allocation. For feasibility, clinical groups were randomized, and all students in the same clinical group who agreed to participate in the study were in the same arm. The EyelearnMR learning module consists of 6 different scenarios of ocular motility simulated in a model patient, which the students were free to peruse within the practice session of 2 hours. The module was presented to the students in a structured and guided fashion in accordance with pedagogical best practices.

Both groups underwent the usual 2-week ophthalmology rotation with 6 preset tutorials for various topics and 9 days of clinical exposure. The intervention group had an additional 2 hours of

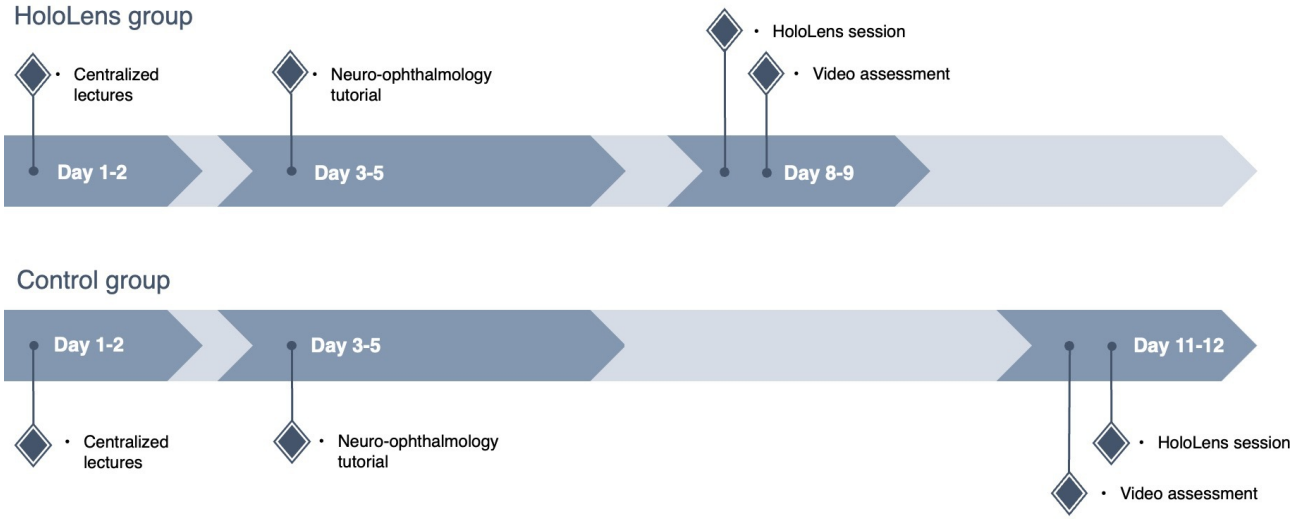
XSL•FO
RenderX

practice with the EyelearnMR devices. Six HoloLens units were available for the study. During the second week of their posting, a 30-minute assessment (5 videos were played, and students had to answer 17 multiple-choice questions related to ocular motility in the domains of examination technique, pathology, and examination signs) was conducted for both the control and intervention groups. To our knowledge, no standardized assessments using psychometric data and benchmarks are available. As such, a bespoke assessment was created. We used "remember," "understand," and "apply" from Bloom's taxonomy hierarchy to come up with the questions for the scenarios. The clinical assessment was intentionally designed to focus on core competencies relevant to ocular motility—specifically, examination technique, recognition of clinical signs, and diagnostic reasoning. A total of 17 structured multiple-choice questions were administered across 5 video-based clinical scenarios, each designed to assess different facets of performance. The questions were vetted by 2 senior neuro-ophthalmologists in clinical practice, ensuring clinical validity and educational relevance. The following is a breakdown of the scenarios:

- Scenario 1: normal patient to test ocular motility assessment technique
- Scenario 2: patient with left partial cranial nerve 3 palsy
- Scenario 3: patient with right cranial nerve 6 palsy
- Scenario 4: patient with thyroid eye disease (lid retraction and ophthalmoplegia)
- Scenario 5: patient with left internuclear ophthalmoplegia

The assessment was conducted mid-posting for the intervention group and at the end of the posting for the control group. The assessment questions were not shared with the groups beforehand. The rationale for assessing the intervention group earlier, in addition to setting a higher bar for EyelearnMR, was to allow for the provision of outcomes showing noninferiority between both groups. In the event of noninferiority, we could demonstrate that EyelearnMR is able to replace some degree of traditional clinical teaching even with a shorter total clinical exposure time. For the EyelearnMR group, the 2-hour EyelearnMR session and the 30-minute video-based assessment were scheduled on days 8 and 9 of the 2-week posting (Figure 2). For the control group, the same assessment was conducted on days 11 and 12, coinciding with the usual end-of-posting evaluation. This schedule reflected how EyelearnMR was intended to be implemented in practice: as a mid-rotation consolidation session that could be integrated into the existing timetable without displacing core teaching activities or overloading the final days of the posting.

Figure 2. User study timeline.



Students in both groups were asked to fill out the UEQ [21]. Students in the control group were given the opportunity to experience the EyelearnMR modules for 2 hours at the end of the posting, following which they were also asked to fill out the UEQ.

A sample of the assessment questions and the UEQ can be found in Multimedia Appendix 1.

Statistical Analysis

Data were analyzed descriptively first. Means and SDs or medians and ranges were reported for the numerical variables, whereas frequencies and percentages were reported for the categorical or ordinal variables. Comparison of scores between groups was analyzed using the Mann-Whitney *U* test. The Fisher exact test was used to compare the test results on technique, examination, and pathology-related knowledge. *P* values of <.05 were considered statistically significant. Results from the UEQ were evaluated using the data analysis tools provided by the UEQ team [22]. The range of the scales is between −3 ("horribly bad") and +3 ("extremely good"). Values between −0.8 and 0.8 represent a neutral evaluation of the corresponding scale, values of >0.8 represent a positive evaluation, and values of <−0.8 represent a negative evaluation [21]. Statistical analysis was carried out using SPSS (version 29.0.0; IBM Corp).

Ethical Considerations

This study was approved by the National University of Singapore Learning and Analytics Committee on Ethics (L2022-09-02). Informed consent was obtained and participants' confidentiality was maintained by not having any personal data recorded.

XSL•FO
RenderX

Results

Recruitment was conducted from July 2023 to January 2024 and was completed as of submission of this manuscript.

A total of 54 fourth-year medical students (age range 23-26 years) were recruited for this study: 24 (44.4%) in the control arm and 30 (55.6%) in the EyelearnMR arm. The median assessment score was 15 (IQR 14-15) for the control group and 16 (IQR 15-17) for the EyelearnMR group (*P*=.03). The assessment questionnaire consisted of items about testing techniques, examination, and pathology-related knowledge (as shown in Multimedia Appendix 1). As shown in Table 1, a total of 100% (30/30) of the students scored full marks (3/3) for the technique portion in the EyelearnMR group compared to 70.8% (17/24) of the students in the control group, and this was statistically significant (*P*=.002). There was no statistically significant difference between the groups for the examination and pathology portions.

Table . Fisher exact test results according to subgroup analysis.

Question	Control group (n=24), n (%)	MR[a] group (n=30), n (%)	Fisher exact test *P* value
Technique (out of 3 points)			.002
2 points	7 (29.2)	0 (0.0)	
3 points	17 (70.8)	30 (100.0)	
Examination (out of 8 points)			.13
4 points	1 (4.2)	1 (3.3)	
5 points	3 (12.5)	3 (10.0)	
6 points	9 (37.5)	9 (30.0)	
7 points	9 (37.5)	11 (36.7)	
8 points	2 (8.3)	11 (36.7)	
Pathology (out of 6 points)			.33
4 points	1 (4.2)	1 (3.3)	
5 points	7 (29.2)	4 (13.3)	
6 points	16 (66.7)	25 (83.3)	

[a]MR: mixed reality.

The results from the UEQ are shown in Tables 2 and 3, as well as in Figure 3. Regarding the combined evaluation, attractiveness, efficiency, dependability, stimulation, and novelty received a positive evaluation. Perspicuity received a neutral evaluation. Statistical analysis using a simple 2-tailed *t* test showed no statistically significant difference between the HoloLens and control groups for attractiveness, perspicuity, efficiency, dependability, stimulation, and novelty (*P*=.37, *P*=.26, *P*=.57, *P*=.38, *P*=.79, and *P*=.12, respectively).

XSL•FO
RenderX

Table . Mean scores and variances of the User Experience Questionnaire items.

Item	Score (–3 to +3), mean (SD)	Variance	Lower bound	Upper bound	Scale
1[a]	1.0 (1.6)	2.6	“Annoying”	“Enjoyable”	Attractiveness
2[a]	1.6 (1.0)	0.9	“Not understandable”	“Understandable”	Perspicuity
3[a]	1.6 (1.6)	2.6	“Creative”	“Dull”	Novelty
4[b]	0.5 (1.6)	2.4	“Easy to learn”	“Difficult to learn”	Perspicuity
5[a]	1.5 (1.2)	1.4	“Valuable”	“Inferior”	Stimulation
6[a]	1.7 (1.0)	1.0	“Boring”	“Exciting”	Stimulation
7[a]	1.9 (1.0)	1.0	“Not interesting”	“Interesting”	Stimulation
8[b]	0.2 (1.4)	2.0	“Unpredictable”	“Predictable”	Dependability
9[b]	0.0 (1.5)	2.3	“Fast”	“Slow”	Efficiency
10[a]	2.0 (1.0)	1.1	“Inventive”	“Conventional”	Novelty
11[a]	1.6 (1.1)	1.1	“Obstructive”	“Supportive”	Dependability
12[a]	1.8 (1.0)	1.0	“Good”	“Bad”	Attractiveness
13[b]	0.0 (1.4)	1.9	“Complicated”	“Easy”	Perspicuity
14[a]	1.3 (1.1)	1.3	“Unlikable”	“Pleasing”	Attractiveness
15[a]	1.3 (1.2)	1.5	“Usual”	“Leading edge”	Novelty
16[a]	1.2 (1.2)	1.4	“Unpleasant”	“Pleasant”	Attractiveness
17[a]	1.0 (1.2)	1.5	“Secure”	“Not secure”	Dependability
18[a]	1.2 (1.2)	1.4	“Motivating”	“Demotivating”	Stimulation
19[a]	1.5 (1.1)	1.2	“Meets expectations”	“Does not meet expectations”	Dependability
20[b]	0.7 (1.5)	2.2	“Inefficient”	“Efficient”	Efficiency
21[a]	0.8 (1.6)	2.6	“Clear”	“Confusing”	Perspicuity
22[a]	1.1 (1.3)	1.7	“Impractical”	“Practical”	Efficiency
23[a]	1.5 (1.1)	1.2	“Organized”	“Cluttered”	Efficiency
24[a]	1.6 (1.0)	1.1	“Attractive”	“Unattractive”	Attractiveness
25[a]	1.6 (1.0)	0.9	“Friendly”	“Unfriendly”	Attractiveness
26[a]	1.5 (1.4)	1.9	“Conservative”	“Innovative”	Novelty

[a]Positive evaluation.

[b]Neutral evaluation.

XSL•FO
RenderX

Table . Mean scores and variances of the User Experience Questionnaire (UEQ) scales overall, for the HoloLens group, and for the control group.

UEQ scale	Score (–3 to +3), mean (SD)	Variance	Cronbach α coefficient
Overall			
Attractiveness[a]	1.413 (0.969)	0.94	0.92
Perspicuity[b]	0.736 (1.094)	1.2	0.8
Efficiency[a]	0.822 (1.068)	1.14	0.8
Dependability[a]	1.087 (0.801)	0.64	0.61
Stimulation[a]	1.577 (0.845)	0.71	0.78
Novelty[a]	1.606 (0.967)	0.93	0.72
sEyelearnMR group			
Attractiveness[a]	1.311 (1.01)	1.02	0.92
Perspicuity[b]	0.592 (1.15)	1.31	0.82
Efficiency[b]	0.75 (1.13)	1.28	0.86
Dependability[a]	1 (0.77)	0.6	0.6
Stimulation[a]	1.55 (0.86)	0.74	0.76
Novelty[a]	1.433 (1.04)	1.09	0.72
Control group			
Attractiveness[a]	1.553 (0.915)	0.84	0.93
Perspicuity[a]	0.932 (1.012)	1.02	0.76
Efficiency[a]	0.92 (0.992)	0.98	0.71
Dependability[a]	1.205 (0.84)	0.71	0.61
Stimulation[a]	1.614 (0.841)	0.71	0.83
Novelty[a]	1.841 (0.818)	0.67	0.71

[a]Positive evaluation.

[b]Neutral evaluation.

XSL•FO
RenderX

Figure 3. Graphs of mean User Experience Questionnaire scales: (A) combined control and EyelearnMR groups; (B) EyelearnMR group; and (C) control group.

XSL•FO
RenderX

Discussion

Principal Findings

In this noninferiority study, a curriculum incorporating EyelearnMR performed at least as well as the traditional clinical teaching–only curriculum and overall supported comparable learning objectives. Our study demonstrates that EyelearnMR on the HoloLens 2 is an effective teaching tool for ocular motility examination. Students who used EyelearnMR achieved significantly higher scores on the video-based assessment compared with those who underwent traditional clinical teaching alone (median score 16 vs 15 out of 17; *P*=.03). Although the difference was statistically significant, the absolute gap of 1 point on a 17-point scale is small, and its practical or clinical relevance should be interpreted in context. There was a particularly marked difference in examination technique (*P*=.002). This was despite the EyelearnMR group having reduced clinical exposure time at the time of testing. The improved performance in the EyelearnMR group likely reflects the combined impact of structured deliberate practice, standardized exposure to key motility disorders, and consistent reproduction of clinical signs that are not guaranteed in opportunistic encounters with real patients.

Positive evaluations were received for attractiveness, efficiency, dependability, stimulation, and novelty. Efficiency, dependability, stimulation, and perspicuity are parameters that are relevant for learning, and positive evaluations were received for the first 3. However, neutral evaluations were received for perspicuity (quality of being clear and easy to understand). It is postulated that many students had difficulty with the pinching gesture required to select specific options when operating the HoloLens. Despite demonstrating the technique to the students a few times, a few of them were unable to grasp it. This was deemed to be a technical limitation of the device and is expected to improve in subsequent updates. Very positive evaluations were received for novelty, which is not surprising in view of the fact that students are not exposed to MR in their usual clinical rotations. When comparing the UEQ results of the 2 groups, there was no statistically significant difference in the evaluation. As such, we can conclude that there was no translatable interaction between the study design and perceptions of the software.

Implications of the Findings

Deliberate Practice and Experiential Learning

EyelearnMR, being a type of simulation learning, facilitates deliberate practice in a controlled environment and allows for learning through problem-solving [23]. A systematic review by Cook et al [20] found that simulation training in health profession education was consistently associated with large positive effects on outcomes of knowledge, skills, and behaviors and moderate positive effects for patient-related outcomes, cementing its role in health care education.

In the EyelearnMR group, the learning module was presented to the students in a structured and guided fashion in accordance with pedagogical best practices, and they were allowed to practice their examinations at their own pace, repeating specific cases if necessary. Conversely, the control group only had access to traditional clinical skill tutorials and clinical sessions, which typically use either simulated or real patients. These patients may not allow for enough deliberate practice time due to fatigue or time constraints. In a clinical tutorial group setting, the teacher-to-student ratio is typically 1:7, which limits the number of attempts that students may be able to practice [24].

Reproducibility of Clinical Signs With a Comprehensive Selection of Conditions

Making a diagnosis from physical examination involves applying a set of clinical examination skills consistently, then evaluating the findings and reaching a conclusion. Repeated deliberate practice of the examination steps improves consistency, and exposure to a variety of conditions builds a knowledge base for appropriate diagnosis.

In the control group (and traditional clinical teaching environments), the range of conditions that a student may encounter is unequal across the student population. Simulated patients (often volunteers) are not able to fully replicate critical, life-threatening clinical signs for effective skill transfer, especially for a topic such as ocular examination. The alternative is to have students examine real patients with the actual condition. While this is done opportunistically in the clinic for all students, it is impossible for them to be exposed to the same conditions with the same degree of significant findings. However, in the EyelearnMR group, every student was given the opportunity to examine a defined and comprehensive set of conditions. They were able to fully experience and learn from "patients" displaying consistent and reproducible clinical signs [25]. This facilitated a broad knowledge base on which to build their diagnostic competencies [26]. This is in contrast with the control group, which had to rely on chance encounters with such patients in the clinical setting, likely with varying physical findings.

The ability of the EyelearnMR module to reproduce the same clinical signs allows for truly experiential learning and deliberate practice for the students, increasing the effectiveness of the teaching and knowledge retention. This forms a strong foundation for the understanding of the underlying concepts, which translated to better scores on the final assessment.

Comparison to the Literature

There are existing virtual and augmented reality applications in ophthalmology, but these are mainly on surgical simulation such as cataract surgery training or anatomical visualization [27]. However, EyelearnMR targets the ocular motility clinical examination—a skill set often neglected in current platforms [28]. It enables users to interact with life-sized, holographic patients with varied and reproducible motility disorders. This provides students with the opportunity to practice their examination techniques repeatedly in a realistic and reproducible environment and may help build their confidence and develop diagnostic skills in a more consistent and reliable way.

Strengths and Limitations

This is a novel study that investigates the use of MR to teach ocular motility testing. To our knowledge, no existing simulator

XSL•FO
RenderX

can perform this task. A scoping review by Krutsinger et al [29] on virtual reality–based medical education in ophthalmology summarized its use mainly for surgical training; clinical applications in diagnosis; counseling; and physical examination, including simulated pupil examination, simulation-based slit lamp training, and ophthalmoscopy [23]. As such, this study addresses an important gap by examining the use of MR to teach ocular motility testing, an area not covered by current simulators or virtual reality platforms. By combining real and virtual environments, MR enables learners to appreciate and interact with extraocular movements in 3 dimensions, supporting a deeper understanding of spatial relationships and examination techniques that are difficult to convey through traditional or screen-based simulations.

The technical limitations of the HoloLens 2 headset include limited field of view and low battery life. The device has a field of view of 43° horizontally compared to the human's field of view of 135°. Therefore, users must adjust their head angles to ensure that the targets are within the device's field of view. On average, battery life can only be sustained for 2 hours of active use. Inclusive of setup time, a device would usually be near the end of its battery life at the end of the assessment. Student feedback highlighted issues with tracking accuracy despite the device's capabilities for both hand and eye tracking. Cost is also a limitation as each HoloLens 2 headset can be costly and adequate funding is required to support such deployment. Finally, the EyelearnMR evaluated a single examination skill (ocular motility) with tightly defined parameters. This was done purposefully to maintain the scientific rigor in the study and reduce other potential confounders. Therefore, the results of this study cannot be extrapolated to other forms of clinical skill examination at this point in time. Future studies will be useful for exploring broader applications of MR in medical education.

MR can also create a potentially dangerous environment if users are not careful; for example, users may not be aware of their surroundings and could trip, fall, or walk into objects. This is more of a concern in virtual reality, where the user is unable to see the actual surroundings, whereas in MR, the user can observe their real environment.

Our cohort consisted of fourth-year medical students from a single medical school rotation, and participation was opt-in. This introduces 2 potential issues. First, the sample size was modest, and group allocation by clinical batch may mean that unmeasured differences between groups (eg, baseline motivation, prior exposure to neuro-ophthalmology, or varying tutorial quality) could influence outcomes. Second, because participation was voluntary, there may have been self-selection bias; students who agreed to participate could differ systematically from those who declined, particularly in interest or confidence in ophthalmology.

In terms of study design, as this study was on an opt-in basis for the students who rotated to the hospital, the recruited sample size was not balanced between the 2 groups: 24 in the control arm and 30 in the EyelearnMR arm. This imbalance in sample size could cause selection bias and potentially reduce statistical power and precision. Although the completed questionnaires were retrieved from the students, a limitation of this study was that it assumed academic integrity, trusting that students would refrain from sharing the scenarios or questions with future student groups rotating through the ophthalmology posting.

The assessment was administered at different time points for the 2 groups (days 8 and 9 vs days 11 and 12). Although this reflected how EyelearnMR would realistically be integrated into the rotation, we cannot exclude a potential recency effect. However, the assessment targeted application of examination technique and diagnostic reasoning rather than simple recall, and the control group had continuous clinical exposure and tutorials up to the end of the posting.

Although video-based assessments standardize stimuli, they may not fully represent the variability observed in real patients. Subtle motility deficits may appear different when captured on video rather than during a face-to-face examination. Additionally, the bespoke nature of the assessment means that it has not yet undergone formal psychometric validation (eg, reliability testing and item difficulty calibration). This may affect reproducibility if used in different settings or with different cohorts.

It is also important to acknowledge that the observed improvement cannot be attributed solely to the MR software itself. EyelearnMR was delivered during a structured teaching session that incorporated pedagogical best practices, including guided demonstration, self-paced exploration, and opportunities for repeated practice with immediate visual feedback. Any of these components may have contributed to the improved performance. This study was not designed to isolate the individual effects of software features, instructional approach, or practice opportunities, and therefore, the greatest effect on learning cannot be determined from our data alone.

Conclusions

On the basis of this study, EyelearnMR with traditional teaching is noninferior to traditional teaching alone. It provided a comparable experience and equally supported learning objectives. It is an effective supplementary teaching tool in ophthalmic education and may confer additional learning benefits to those of a traditional clinical posting, especially in the field of clinical examination technique.

Funding

This study received funding support from the Centre for Development of Teaching and Learning, National University of Singapore, via a Teaching Enhancement Grant. The funding source was not involved in the study design; collection, analysis, and interpretation of the data; writing of the report; or decision to submit the paper for publication.

XSL•FO
RenderX

Data Availability

The datasets generated or analyzed during this study are available from the corresponding author on reasonable request.

Authors' Contributions

Marcus and Dayna contributed equally as first authors. Victor and Clement contributed equally as last authors.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Video assessment questions and the user experience questionnaire.
[DOCX File, 192 KB - mededu_v12i1e71338_app1.docx]

References

1. Gerup J, Soerensen CB, Dieckmann P. Augmented reality and mixed reality for healthcare education beyond surgery: an integrative review. Int J Med Educ 2020 Jan 18;11:1-18. [doi: 10.5116/ijme.5e01.eb1a] [Medline: 31955150]
2. Bogomolova K, Sam AH, Misky AT, et al. Development of a virtual three‐dimensional assessment scenario for anatomical education. Anat Sci Educ 2021 May;14(3):385-393. [doi: 10.1002/ase.2055] [Medline: 33465814]
3. Pennefather P, Krebs C. Exploring the role of xR in visualisations for use in medical education. Adv Exp Med Biol 2019;1171:15-23. [doi: 10.1007/978-3-030-24281-7_2] [Medline: 31823236]
4. Wish-Baratz S, Crofton AR, Gutierrez J, Henninger E, Griswold MA. Assessment of mixed-reality technology use in remote online anatomy education. JAMA Netw Open 2020 Sep 1;3(9):e2016271. [doi: 10.1001/jamanetworkopen.2020.16271] [Medline: 32940677]
5. Ruthberg JS, Tingle G, Tan L, et al. Mixed reality as a time-efficient alternative to cadaveric dissection. Med Teach 2020 Aug;42(8):896-901. [doi: 10.1080/0142159X.2020.1762032] [Medline: 32401090]
6. Martin G, Koizia L, Kooner A, et al. Use of the HoloLens2 mixed reality headset for protecting health care workers during the COVID-19 pandemic: prospective, observational evaluation. J Med Internet Res 2020 Aug 14;22(8):e21486. [doi: 10.2196/21486] [Medline: 32730222]
7. Moro C, Štromberga Z, Raikos A, Stirling A. The effectiveness of virtual and augmented reality in health sciences and medical anatomy. Anat Sci Educ 2017 Nov;10(6):549-559. [doi: 10.1002/ase.1696] [Medline: 28419750]
8. Tepper OM, Rudy HL, Lefkowitz A, et al. Mixed reality with HoloLens: where virtual reality meets augmented reality in the operating room. Plast Reconstr Surg 2017 Nov;140(5):1066-1070. [doi: 10.1097/PRS.0000000000003802] [Medline: 29068946]
9. Al Janabi HF, Aydin A, Palaneer S, et al. Effectiveness of the HoloLens mixed-reality headset in minimally invasive surgery: a simulation-based feasibility study. Surg Endosc 2020 Mar;34(3):1143-1149. [doi: 10.1007/s00464-019-06862-3] [Medline: 31214807]
10. Zuo Y, Jiang T, Dou J, et al. A novel evaluation model for a mixed-reality surgical navigation system: where Microsoft HoloLens meets the operating room. Surg Innov 2020 Apr;27(2):193-202. [doi: 10.1177/1553350619893236] [Medline: 31920155]
11. Pratt P, Ives M, Lawton G, et al. Through the HoloLens™ looking glass: augmented reality for extremity reconstruction surgery using 3D vascular models with perforating vessels. Eur Radiol Exp 2018;2(1):2. [doi: 10.1186/s41747-017-0033-2] [Medline: 29708204]
12. Ramesh PV, Joshua T, Ray P, et al. Holographic elysium of a 4D ophthalmic anatomical and pathological metaverse with extended reality/mixed reality. Indian J Ophthalmol 2022 Aug;70(8):3116-3121. [doi: 10.4103/ijo.IJO_120_22] [Medline: 35918983]
13. Kumar N, Pandey S, Rahman E. A novel three-dimensional interactive virtual face to facilitate facial anatomy teaching using Microsoft HoloLens. Aesthetic Plast Surg 2021 Jun;45(3):1005-1011. [doi: 10.1007/s00266-020-02110-5] [Medline: 33469701]
14. Maniam P, Schnell P, Dan L, et al. Exploration of temporal bone anatomy using mixed reality (HoloLens): development of a mixed reality anatomy teaching resource prototype. J Vis Commun Med 2020 Jan;43(1):17-26. [doi: 10.1080/17453054.2019.1671813] [Medline: 31645155]
15. Robinson BL, Mitchell TR, Brenseke BM. Evaluating the use of mixed reality to teach gross and microscopic respiratory anatomy. Med Sci Educ 2020 Aug 18;30(4):1745-1748. [doi: 10.1007/s40670-020-01064-2] [Medline: 32837799]
16. Silvero Isidre A, Friederichs H, Müther M, Gallus M, Stummer W, Holling M. Mixed reality as a teaching tool for medical students in neurosurgery. Medicina (Kaunas) 2023 Sep 26;59(10):1720. [doi: 10.3390/medicina59101720] [Medline: 37893438]

XSL•FO
RenderX

17. Muangpoon T, Haghighi Osgouei R, Escobar-Castillejos D, Kontovounisios C, Bello F. Augmented reality system for digital rectal examination training and assessment: system validation. J Med Internet Res 2020 Aug 13;22(8):e18637. [doi: 10.2196/18637] [Medline: 32788146]
18. Schoeb DS, Schwarz J, Hein S, et al. Mixed reality for teaching catheter placement to medical students: a randomized single-blinded, prospective trial. BMC Med Educ 2020 Dec 16;20(1):510. [doi: 10.1186/s12909-020-02450-5] [Medline: 33327963]
19. Minty I, Lawson J, Guha P, et al. The use of mixed reality technology for the objective assessment of clinical skills: a validation study. BMC Med Educ 2022 Aug 23;22(1):639. [doi: 10.1186/s12909-022-03701-3] [Medline: 35999532]
20. Cook DA, Hatala R, Brydges R, et al. Technology-enhanced simulation for health professions education: a systematic review and meta-analysis. JAMA 2011 Sep 7;306(9):978-988. [doi: 10.1001/jama.2011.1234] [Medline: 21900138]
21. Laugwitz B, Held T, Schrepp M. Construction and evaluation of a user experience questionnaire. Presented at: Proceedings of the 4th Symposium of the Workgroup Human-Computer Interaction and Usability Engineering of the Austrian Computer Society; Nov 20-21, 2008. [doi: 10.1007/978-3-540-89350-9_6]
22. Hinderks A, Schrepp M, Domínguez Mayo FJ, Escalona MJ, Thomaschewski J. Developing a UX KPI based on the user experience questionnaire. Comput Stand Interfaces 2019 Jul;65:38-44. [doi: 10.1016/j.csi.2019.01.007]
23. McGaghie WC, Issenberg SB, Cohen ER, Barsuk JH, Wayne DB. Does simulation-based medical education with deliberate practice yield better results than traditional clinical education? A meta-analytic comparative review of the evidence. Acad Med 2011 Jun;86(6):706-711. [doi: 10.1097/ACM.0b013e318217e119] [Medline: 21512370]
24. Yardley S, Teunissen PW, Dornan T. Experiential learning: transforming theory into practice. Med Teach 2012;34(2):161-164. [doi: 10.3109/0142159X.2012.643264] [Medline: 22288996]
25. Issenberg SB, McGaghie WC, Petrusa ER, Lee Gordon D, Scalese RJ. Features and uses of high-fidelity medical simulations that lead to effective learning: a BEME systematic review. Med Teach 2005 Jan;27(1):10-28. [doi: 10.1080/01421590500046924] [Medline: 16147767]
26. Botezatu M, Hult H, Tessma MK, Fors U. Virtual patient simulation: knowledge gain or knowledge loss? Med Teach 2010;32(7):562-568. [doi: 10.3109/01421590903514630] [Medline: 20653378]
27. Ong CW, Tan MC, Lam M, Koh VT. Applications of extended reality in ophthalmology: systematic review. J Med Internet Res 2021 Aug 19;23(8):e24152. [doi: 10.2196/24152] [Medline: 34420929]
28. Iskander M, Ogunsola T, Ramachandran R, McGowan R, Al-Aswad LA. Virtual reality and augmented reality in ophthalmology: a contemporary prospective. Asia Pac J Ophthalmol (Phila) 2021;10(3):244-252. [doi: 10.1097/APO.0000000000000409] [Medline: 34383716]
29. Krutsinger BL, Moore JC. Virtual reality-based medical education in ophthalmology: a scoping review. Cureus 2025 May;17(5):e83845. [doi: 10.7759/cureus.83845] [Medline: 40497222]

Abbreviations

MR: mixed reality
UEQ: User Experience Questionnaire

Edited by B Lesselroth; submitted 15.Jan.2025; peer-reviewed by B Afolabi, R Scherer; revised version received 13.Nov.2025; accepted 25.Nov.2025; published 03.Mar.2026.

Please cite as:
Tan CJM, Yong WWD, Lin HHA, Oh J, Yong HSR, Khew FMJ, Shen L, Gao Y, Kow WCA, Tham YC, Liu D, Cheng CY, Ngiam KY, Yuen YS, Manotosh R, Khoo ET, Koh TCV, Tan WTC
Application of Mixed Reality for Ophthalmic Clinical Skills and Diagnosis: Prospective Study
JMIR Med Educ 2026;12:e71338
URL: https://mededu.jmir.org/2026/1/e71338
doi:10.2196/71338

© Chun Jin Marcus Tan, Wei Wei Dayna Yong, Hui'En Hazel Anne Lin, Jaslyn Oh, How Sheng Rubin Yong, Fang Mei Jayme Khew, Liang Shen, Yujia Gao, Wei Chieh Alfred Kow, Yih Chung Tham, Dianbo Liu, Ching-Yu Cheng, Kee Yuan Ngiam, Yew Sen Yuen, Ray Manotosh, Eng Tat Khoo, Teck Chang Victor Koh, Woon Teck Clement Tan. Originally published in JMIR Medical Education (https://mededu.jmir.org), 3.Mar.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Education, is properly cited. The complete bibliographic information, a link to the original publication on https://mededu.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Cost-Effectiveness of a Blended Learning Training Approach for Health Care Workers Conducting HIV Index Case Testing in Malawi: Secondary Economic Analysis of a Cluster-Randomized Controlled Trial

Poonam Rawat[1], MPH; Nora E Rosenberg[1], PhD; Katie R Mollan[1,2], PhD; Tapiwa A Tembo[3], PhD; Dhrutika Vansia[3], MPH; Jiayu Wang[1], MPA, MPH; Mike J Chitani[3], BA; Maria H Kim[3,4], MD; Ireen Chirombo[3], BSoc; Caroline Kumbuyo[3], BSoc; Victor Mwapasa[5], PhD; Katherine R Simon[3,4], MD; Saeed Ahmed[3,4], MD; Sarah E Rutstein[2], MD, PhD

[1]Gillings School of Global Public Health, University of North Carolina at Chapel Hill, Chapel Hill, NC, United States

[2]School of Medicine, University of North Carolina at Chapel Hill, Chapel Hill, NC, United States

[3]Baylor College of Medicine Children's Foundation, Lilongwe, Malawi

[4]Department of Pediatrics, Baylor College of Medicine, Houston, TX, United States

[5]Department of Epidemiology and Biostatistics, Kamuzu University of Health Sciences, Blantyre, Malawi

Corresponding Author:
Jiayu Wang, MPA, MPH
Gillings School of Global Public Health
University of North Carolina at Chapel Hill
135 Dauer Drive,
Chapel Hill, NC, 27599-7400
United States
Phone: 1 9199662536
Email: mjw0@unc.edu

Abstract

Background: Index case testing (ICT) is an effective strategy for HIV case finding, but implementation in low- and middle-income countries (LMICs) is often limited by cost and logistical challenges. Traditional ICT training—centralized and in-person—is costly, disrupts service delivery, and varies in quality.

Objective: This study evaluates the cost-effectiveness of a blended learning (BL) implementation package designed to build health care worker capacity for ICT, combining tablet-guided teaching and practice sessions, phone-based feedback, and tablet-guided continuous quality improvement, compared with standard of care (SOC) training.

Methods: The Package of Resources for Assisted Contact Tracing: Implementation, Costs, and Effectiveness (PRACTICE) cluster-randomized controlled trial included 33 clusters in southern Malawi from May 2022 to September 2023, randomized 2:1 to SOC (n=22) or SOC + BL implementation package (n=11). Our cost-effectiveness analysis, from the health system perspective, used microcosting, time-and-motion assessments, and observed trial outcomes. The decision tree model estimated total program costs, contact testing outcomes, and incremental cost-effectiveness ratios (ICERs) per contact tested and per person diagnosed with HIV across 1 year of implementation. Sensitivity analyses assessed parameter uncertainty, and a scenario analysis modeled a nationwide scale-up under a decentralized, Ministry of Health–led approach.

Results: Our model simulated 100,000 index clients eligible for contact elicitation over 1 year across 2 districts (50,000 per arm). The BL implementation package arm yielded 891 additional contacts tested and 54 more HIV diagnoses. The ICERs were US $125 per contact tested and US $2045 per person diagnosed with HIV; excluding training development costs reduced these to US $69 and US $1136, respectively. Nationwide scale-up under a Ministry of Health–led model further reduced ICERs to US $43 per contact tested and US $698 per person diagnosed with HIV. Probabilistic sensitivity analyses showed the BL implementation package was cost-effective in most simulations.

Conclusions: The BL implementation package improved ICT delivery and HIV case finding. Scenario analyses suggested that a decentralized, government-led scale-up could substantially reduce costs and may represent an efficient case-finding strategy,

XSL•FO
RenderX

particularly in Malawi's mature epidemic. BL implementation packages provide a scalable, system-integrated, cost-effective approach that could strengthen health care worker capacity across other service delivery areas in low-resource settings.

International Registered Report Identifier (IRRID): RR2-10.1136/bmjopen-2023-077706

(JMIR Med Educ 2026;12:e95450) doi:10.2196/95450

KEYWORDS

ACT; assisted contact tracing; blended learning; CEA; cost-effectiveness analysis; ICER; ICT; incremental cost-effectiveness ratio; index case testing; LMIC; low- and middle-income country

Introduction

Blended learning (BL), an educational approach combining in-person instruction with digital components, offers a promising and often more effective solution for strengthening health care worker (HCW) capacity in low- and middle-income countries (LMICs) [1-5]. It standardizes training content, enhances accessibility, and reduces the logistical burdens associated with traditional in-person training [3,6-11] and is supported by the World Health Organization (WHO) as a complement to face-to-face instruction [12,13].

One opportunity for BL is in the implementation of index case testing (ICT) for HIV case finding. Also known as voluntary assisted partner notification or contact tracing, ICT is an evidence-based intervention in which persons living with HIV, referred to as index clients, are supported in recruiting their contacts, including sexual partners and children, for HIV testing [14-17]. When implemented effectively, ICT plays a critical role in identifying persons living with HIV who need treatment services and HIV-uninfected individuals who may benefit from prevention services [18-21]. The WHO recommends ICT as a core component of comprehensive HIV testing and care [17]. Although Malawi and many other countries have adopted ICT into their guidelines [22,23], outcomes often fall short of those observed in clinical trials. In Malawi, limited HCW capacity and challenges with clinical coordination have hindered effective ICT implementation [15,18,22,24].

One key pillar for strengthening HCW capacity in LMICs is continuous training and quality improvement [25]. Historically, ICT capacity building has involved in-person training, role-playing, hands-on practice, and feedback. However, centralized, synchronous, face-to-face training poses significant challenges, including high costs related to travel, lodging, and per diems; disruptions to clinical services when multiple providers are away; limited opportunities for individualized practice or feedback; and variable quality due to differences in trainers' teaching abilities [26-28]. Posttraining support, such as continuous quality improvement (CQI) activities, is often constrained by shortages of central staff, high travel costs, and difficulty monitoring progress across multiple facilities [14]. These barriers undermine the sustainability and consistency of ICT delivery.

To address these challenges, we developed a BL ICT implementation package that integrates digitally guided individual and small-group training to strengthen HCW capacity and improve ICT program fidelity [15,29]. The parent study, Package of Resources for Assisted Contact Tracing: Implementation, Costs, and Effectiveness (PRACTICE), assessed the feasibility, effectiveness, and cost-effectiveness of a BL implementation package layered on top of the standard of care (SOC) for Malawi's ICT program, compared with SOC alone (hereafter referred to as SOC and SOC + BL, respectively). The study demonstrated improved fidelity to ICT protocols and more effective ICT outcomes [29]. Here, we examine the costs and cost-effectiveness of the BL implementation package for ICT implementation. We hypothesize that by improving HCW fidelity to ICT protocols and enhancing the effectiveness of ICT for reaching contacts of index clients, the approach may offer a cost-effective strategy for strengthening HIV prevention and care in Malawi.

Methods

Study Design and Setting

The PRACTICE study was developed using implementation science theory and formative qualitative research. Intervention design was informed by the Consolidated Framework for Implementation Research (CFIR) to identify barriers and facilitators to ICT delivery, while educational and behavioral theories and human-centered design principles guided the development and digitization of the BL implementation package. The resulting package was evaluated in a pragmatic cluster-randomized controlled trial (RCT) [15,16,30].

PRACTICE was a cluster-RCT conducted in the Machinga and Balaka districts of southern Malawi from May 2022 to September 2023. PRACTICE was described in detail previously [16]. Briefly, the study took place in 34 health care facilities, 14 in Balaka and 20 in Machinga, all of which routinely offer HIV testing and treatment services. These facilities included hospitals (n=3), health centers (n=27), and dispensaries (n=4), the latter being smaller facilities that provide basic primary health care. All facilities were operated by either the Malawi Ministry of Health (MOH) or the Christian Health Association of Malawi (CHAM), a faith-based organization. Both the MOH and CHAM provide HIV services free of charge under the national program. The selected sites received additional implementation support from the Tingathe Program, a US President's Emergency Plan for AIDS Relief (PEPFAR)–supported initiative led by Baylor College of Medicine Children's Foundation Malawi.

For randomization purposes, the 34 health care facilities were grouped into 33 clusters, with 2 neighboring facilities with similar baseline characteristics combined into a single cluster to allow for evenly sized allocation groups. Clusters were stratified by district, facility type, and the baseline number of

XSL•FO
RenderX

potential index clients, then randomized in a 2:1 ratio to receive either SOC or SOC + BL.

ICT Approach

ICT is an evidence-based approach used to identify individuals who may have been exposed to HIV. In Malawi, HCWs support persons living with HIV in identifying and testing their sexual partners and family members. This study considered index clients to be persons living with HIV who were either newly diagnosed through HIV testing services (HTS) at a facility or previously diagnosed and receiving antiretroviral therapy (ART). Potential contacts include the index clients' sexual partners, biological children, household members, and social contacts [16].

At the time of the study, ICT in Malawi was implemented through three referral methods: (1) active or assisted referral, in which an HCW confidentially notifies and traces contacts for testing; (2) passive referral without self-test kits, in which the index client independently encourages contacts to test using a referral slip; and (3) passive referral with HIV self-test kits, in which the index client provides contacts with both a referral slip and an HIV self-test kit. Effective ICT implementation relies on trained lay HCWs to provide counseling and organize appropriate logistics for tracing contacts.

SOC Implementation

The SOC consisted of 1 centralized, in-person, facilitator-led, 2-hour ICT training for HCWs, covering key ICT concepts, procedures, and counseling approaches. This one-time training was part of a larger Tingathe training that covered a range of areas beyond ICT. The SOC implementation was also supplemented by regular in-person facility mentorship visits conducted across all 33 Tingathe-supported clusters.

BL Implementation Package for ICT

As previously described [16,29], clusters randomly assigned to the BL implementation package arm (n=11) received SOC + BL, which combined digital and in-person training with monthly CQI sessions. The intervention was structured into 4 phases. First, all HCWs completed an 8-hour asynchronous digital learning module over weeks 1 to 3 using tablets at their facilities. Following week 3, a 2-day (14-hour), in-person synchronous training was held over a weekend at or near their facility. These small-group, tablet-guided sessions included role-plays, individual feedback, and group discussions focused on ICT protocols. In weeks 4 to 5, all HCWs received a 1-hour phone feedback session based on simulated ICT encounters with study staff portraying index and contact clients; study staff provided feedback, and HCWs who struggled to master the material were given opportunities to repeat the exercises until their performance improved. Finally, from weeks 10 to 52, HCWs participated in a total of six 2-hour facility-based CQI sessions to review implementation, troubleshoot challenges, and refine strategies. In total, the BL components layered onto SOC accounted for approximately 35 additional hours of training, in addition to the 2 hours provided through SOC.

Base Case

The base case represents the cluster-RCT scenario, modeling the estimated costs and outcomes for index clients enrolled in the 33 clusters participating in the PRACTICE study over 1 year [16,29]. The trial observed approximately 50,000 potential index clients receiving services at the 33 clusters per year. Potential index clients were defined as the sum of individuals enrolled in the ART program and those newly diagnosed through the HTS program over 1 year. To address the asymmetry of the 2:1 randomization, we simulated all 33 clusters receiving the intervention package (50,000 potential index clients) compared with all 33 clusters receiving the SOC (50,000 potential index clients), for a total of 100,000 index clients modeled.

Both study arms received standard ICT training and programming; thus, costs associated with the BL components were treated as incremental to SOC. SOC-related costs were assumed to be equivalent across arms and were not estimated separately.

Consistent with Malawi's public health system, analyses were conducted from a health system perspective, incorporating system-level costs related to intervention implementation, tracing, and testing contacts of index cases living with HIV. We estimated the total incremental costs associated with developing and implementing BL over 1 year of operation. All costs and effectiveness outcomes were annualized to reflect this 1-year analytic time horizon. Our primary cost-effectiveness outcomes included cost per client tested and cost per person diagnosed with HIV, estimated using incremental cost-effectiveness ratios (ICERs).

Probability Inputs

Event probabilities were estimated using observed clinical trial outcomes when available and supplemented by clinic register review and published literature. Plausible ranges were derived from reported 95% CIs, the minimum and maximum values found across relevant studies, or, when ranges were not explicitly reported, from the spread of values across comparable published estimates. When no published or empirical reference was available to inform parameter uncertainty, ranges were defined by applying a ±20% variation around the base-case value (Table 1 and Figure 1). The model was constructed to reflect the trajectory of index and contact client engagement with the health system and affiliated HCWs, from the point of index presentation to contact testing and HIV test outcomes.

XSL•FO
RenderX

Table 1. Model input parameters.

Parameter (decision node)[a]	Event probability value (plausible range), %		References
	SOC[b]	SOC + BL[c]	
Index identification[d]			
Index identified via HTS[e] (A1)	6.6 (6.6-18.2)	Same as SOC	Trial [31,32]
Index identified via ART[f] (A2)	93.3	Same as SOC	Trial, assumed
Index case testing participation			
HTS group (B1)	92.9 (74.4-100)	97.1 (77.6-100)	Trial, assumed
ART group (C1)	44.7 (36-54)	73.1 (58.4-87.6)	Trial, assumed
Preferred contact elicitation method: HTS group			
Active referral (B2)	32.8 (31-100)	33.4 (26.4-39.6)	Trial [24], assumed
Passive referral with self-test kits (B3)	7.4 (5.6-8.4)	7.5 (5.6-8.4)	Trial, assumed
Passive referral (B4)	59.8 (30-100)	59.1 (47.2-70.8)	Trial, [24]
Preferred contact elicitation method: ART group			
Active referral	37.7 (30-46)	47.7 (38.4-57.6)	Trial, assumed
Passive referral with self-test kits	13.2 (10-16)	13.2 (10.5-15.8)	Trial, assumed
Passive referral	49 (39-59)	39.1 (31.2-46.8)	Trial, assumed
Tracing and testing[g]			
Tracing (D1)	100 (70-100)	Same as SOC	Assumed
Testing: active referral (D2)	49 (40-62)	Same as SOC	Clinical record review [31]
Testing: passive referral plus self-test kits (D3)	44 (16-71)	Same as SOC	Clinical record review [31,32]
Testing: passive referral (D4)	15 (15-34)	Same as SOC	Clinical record review [31]
HIV Positivity			
HTS group (B5)	9.6 (9.6-73.7)	9.1 (9.1-73.7)	Trial [31]
ART group	2.9 (2.9-16)	4.2 (4.2-16)	Trial [31]

[a]The letter labels correspond to the specific decision nodes shown in Figure 1, connecting the probabilities in this table to their location within the decision tree.

[b]SOC: standard of care.

[c]SOC + BL: standard of care + blended learning.

[d]The probability of index client identification via ART is complementary to HTS (ie, 1 – HTS) and was assumed to be independent of intervention arm; therefore, overall trial proportions were applied to both study arms rather than arm-specific estimates.

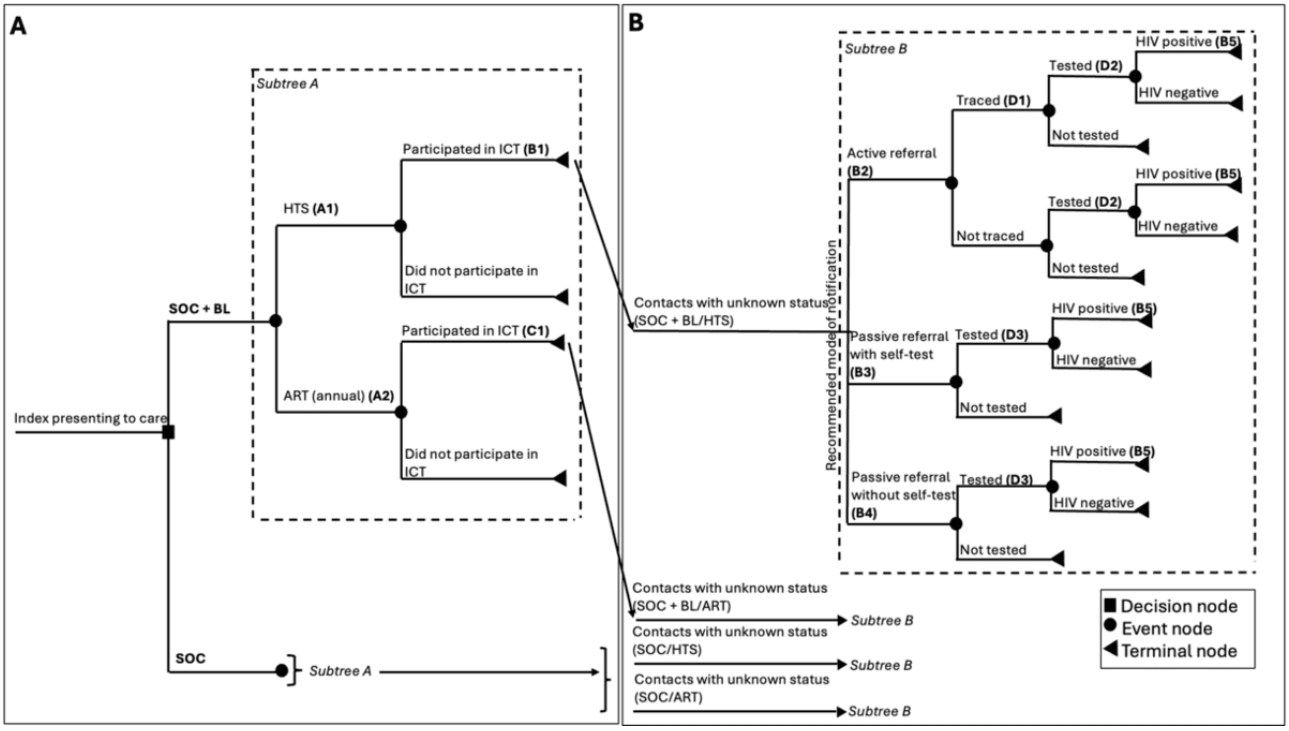
[e]HTS: HIV testing services.

[f]ART: antiretroviral therapy.

[g]Probabilities assumed to apply across both HTS and ART groups.

XSL•FO
RenderX

Figure 1. Decision analytic model for (A) index case testing pathways and (B) contact testing and HIV diagnosis following index elicitation. Standard of care (SOC) indicates standard index case testing (ICT) training and programming; SOC + blended learning (SOC + BL) indicates SOC supplemented with a BL implementation package. The letter labels (A, B, C, etc) on the nodes correspond to the probability values presented in Table 1. Each decision node represents a branching point where the sum of probabilities for all possible outcomes equals 1. For instance, the probability that an index client identified through HIV testing services (HTS) did not participate in ICT equals 1 minus the probability of participation (1 − B1). ART: antiretroviral therapy.



Cost Inputs

We included costs related to both developing and implementing the BL implementation package, including all 4 BL components described earlier (Multimedia Appendix 1). Costs were further designated as strategy development or implementation costs (Table 2).

XSL•FO
RenderX

Table 2. Cost inputs by activity and level (2023 US $).

Cost inputs	Level of cost incurrence	Total cost across 33 clusters (US $)
Strategy development		
Content development	N/A[a]	22,000
Video production	N/A	10,719
Capital purchases[b]	District	14,951
Office supplies	Facility	1428
Miscellaneous[c]	N/A	386
Implementation		
Site-supervisor orientation (face-to-face)	District	3876
Individual tablet-guided learning sessions	HCW[d]	7839
Small-group face-to-face practice sessions	HCW or facility	31,950
One-on-one phone-based feedback	Facility	2013
Continuous quality improvement orientation (face-to-face and remote)	Facility	3394
Continuous quality improvement sessions and remote debrief (6 sessions)	Facility	11,927
Total (fixed costs)	N/A	110,483
Tracing cost (per contact indicated for active referral)[e]	HCW	SOC + BL[f]: 0.90; SOC: 1.09

[a]Not applicable.

[b]Capital purchases include tablets and chargers.

[c]Miscellaneous includes data and time for preparing tablets for field use.

[d]HCW: health care worker.

[e]Tracing costs not included in fixed total. Applied dynamically in the model based on number of contacts traced in each arm.

[f]SOC + BL: standard of care + blended learning.

Cost data were obtained through microcosting and time-and-motion assessments. Microcosting relied on program receipts, program accounts, administrative records, and human resource records. Time-and-motion data were gathered through 3 distinct methods: self-reports and timed assessments from HCWs conducting field-based tracing (Multimedia Appendix 2), direct observations by study staff for phone tracing, and tablet-based paradata tracking of the time spent on digital learning tasks. Time estimates from each method were converted to costs using relevant HCW cadre salaries. In cases in which daily subsistence allowances were provided for a specific activity, we used the daily subsistence allowances as the cost input and did not double-count the associated salary. Costs were categorized at the district, facility, and HCW levels to inform scale-up projections (Multimedia Appendix 2).

Costs were calculated in 2023 Malawian Kwacha (MWK) and 2023 US $. The exchange rate used during the bulk of the trial period was US $1=MWK 1036 [33].

Model Structure

We developed a decision tree model to simulate costs, contact tracing and contact testing outcomes, and estimate the ICERs comparing SOC with SOC + BL (Figure 1). Based on annual visit estimates for individuals presenting for HTS or ART care at 33 participating clusters, our base-case model reflects 50,000 potential index clients.

Our decision tree structure reflects the intentional separation between index and contact registers, in which index clients and contact clients are not directly linked to maintain confidentiality. As such, the model is presented in 2 subparts. Figure 1A reflects the movement of presenting index clients, beginning with a decision node (black square) assigning index clients to either the SOC or SOC + BL, using a 50:50 allocation. Within each arm, a chance node (black circle) determines whether the index client was identified via HTS or ART registers. This is followed by another chance node that models whether the individual participates in ICT, defined as naming contacts for tracing and testing.

Named contacts elicited from participating index clients then serve as the basis for the population moving through Figure 1B. The number of contacts elicited per index client was derived from trial data and varied by arm and source (SOC + BL/ART, SOC + BL/HTS, SOC/ART, and SOC/HTS). These estimates were applied externally by multiplying the number of index clients who participated in ICT (from Figure 1A) by the average number of contacts elicited per participating index client.

Once elicited, contact clients are categorized by HIV status as either unknown (requiring HIV testing) or known HIV-positive but not on ART. Contacts with unknown status proceed through 1 of 3 tracing methods according to the referring index client's preference: active referral, passive referral with self-test kits, or passive referral without self-test kits. A series of chance nodes determine whether the contact is successfully traced, whether

XSL•FO
RenderX

they undergo HIV testing, and, if tested, whether they receive a positive or negative result.

Key Assumptions

We made several key assumptions that influence the model structure and potentially the outcomes. In this 2:1 cluster-RCT, only 11 clusters were exposed to the intervention. In our model, we present outcomes and associated costs as if all 33 clusters were exposed to the intervention, scaling district, facility, and HCW inputs to reflect this assumption. This is then compared with the SOC, again reflecting a scenario in which all 33 clusters received only standard services.

Further, we assumed that the proportion of index clients identified through either HTS or ART was independent of the intervention arm (Table 1). Finally, as the trial did not collect data on the proportion of named contacts who were traced, we used a base-case assumption that all named contacts in the active referral pathway were traced. The trial also did not collect testing rates by notification approach among named contacts; these estimates were obtained retrospectively through manual review of HIV testing registers at participating health facilities. All inputs were varied over a plausible range and tested in univariate and probabilistic sensitivity analyses to examine their relative importance, as described below.

Analyses

We calculated ICERs by dividing the difference in costs by the difference in each outcome for the SOC + BL ICT implementation approach compared with the SOC approach.

We built the model in Microsoft Excel, version 16.97.2 (25052611), and used Crystal Ball (version 11.1.2.4.900; Oracle) to conduct probabilistic sensitivity analyses across 1000 simulations drawn from defined individual parameter distributions.

Sensitivity and Scenario Analyses

To assess the robustness of the model, we conducted both 1-way (univariate) and probabilistic sensitivity analyses. The 1-way sensitivity analyses examined the impact of varying individual parameters on key outcomes to identify which inputs most influenced model results [34]. The probabilistic uncertainty analyses varied all parameters simultaneously based on their distributions, capturing the combined effect of parameter uncertainty on model outcomes [35]. We also conducted a sensitivity analysis of the US $–MWK exchange rate to assess the impact of fluctuations during the study period, which spanned October 2022 to September 2023.

We conducted 2 scenario analyses to explore the cost implications of alternative implementation approaches for the BL implementation package. First, we estimated implementation costs excluding one-time development costs, such as strategy design, content creation, and training materials, to simulate annual implementation across the original 33 clusters. This provided insights into the cost-effectiveness of a realistic implementation scenario in which development costs would not be duplicated, as these costs are otherwise included in the base-case approach.

Second, we modeled a nationwide rollout of the BL implementation package across all facilities providing HIV testing and ART services in Malawi over a 1-year implementation horizon. This scenario reflected a pragmatic, government-led implementation approach incorporating key efficiencies expected under MOH leadership by assuming integration within existing health system structures and personnel tasks. We used 2024 national data on new ART initiations and ongoing ART use to estimate the eligible index testing population, defining “initiations” as persons newly diagnosed with HIV (ie, HTS in our model). To estimate HCW needs, we scaled the number of HCWs per facility by client volume, using the base-case average (n=9) as a reference.

Implementation was envisioned through a decentralized cascade model. A 4-day central training of trainers (TOT) was modeled after existing programs for similar activities. The TOT was intended for district-level leadership, including 1 district health officer, ART coordinator, HTS coordinator, and driver from each district, along with 4 MOH staff members and 1 national trainer. Under this approach, efficiencies were expected through reduced reliance on external facilitators, use of existing MOH training spaces, and integration of ICT duties into the routine responsibilities of district and facility staff. Development costs were excluded, assuming use of already developed materials for national rollout. Cost categories were scaled based on facility, district, and personnel counts, with recurring costs (training delivery, travel, venues, and refreshments) adjusted to reflect district-led implementation. Fuel and venue costs were substantially reduced compared with the base case, consistent with local delivery and MOH-owned or subsidized training spaces. Detailed operational assumptions, including the composition of the TOT cohort, district-level training logistics, per diem and travel policies, and CQI delivery procedures, are provided in Multimedia Appendix 3.

To account for the possibility that training materials may require periodic updating or adaptation over time (eg, in response to evolving guidelines or contextual needs), we examined an additional scenario incorporating partial reuse of content development and video production materials (Multimedia Appendix 4). This scenario reflects reduced development costs associated with maintaining and updating existing materials rather than full redevelopment.

Ethical Considerations

The PRACTICE trial was approved by the Malawi National Health Science Research Committee (#20/06/2566), the University of North Carolina at Chapel Hill Institutional Review Board (20-1810), and the Baylor College of Medicine Institutional Review Board (H-48800). All participating HCWs provided written informed consent prior to engaging in any study-related activities, including training sessions and program implementation.

Results

Overview

In the base case, we evaluated 50,000 potential index clients receiving services in the trial-included clinics over 1 year,

XSL•FO
RenderX

modeling this cohort of 50,000 clients receiving SOC + BL compared with the same cohort receiving SOC. SOC + BL yielded a total of 9374 named contacts, compared with 7122 named contacts under SOC. Among these, we estimated that 3026 contacts would be tested and 183 individuals would be newly diagnosed with HIV under SOC + BL, compared with 2135 contacts tested and 129 individuals newly diagnosed under SOC. SOC + BL therefore resulted in 891 additional contacts tested and 54 more persons diagnosed with HIV (per 50,000) compared with SOC, representing approximately a 42% relative increase in both contact testing and HIV diagnoses. The total cost of implementing the BL strategy in the 33 clusters across 2 districts was US $111,258, resulting in an ICER of US $125 per additional contact tested and US $2045 per additional person diagnosed with HIV (Table 3).

Table 3. Cost-effectiveness outcomes under base-case and scenario assumptions.

Implementation strategy	Total cost of implementation (US $)	Eligible index clients over 1 year, n	Total contacts tested, n	Total persons diagnosed with HIV, n	Incremental contacts tested, n	Incremental persons diagnosed with HIV, n	ICER (US $ per contact tested)	ICER (US $ per person diagnosed with HIV
SOC[a] (base case)	—	50,000	2135	129	—	—	—	—
SOC + BL[b] (base case)	111,258	50,000	3026	183	891	54	125	2045
SOC + BL (excluding strategy development costs)	61,774	50,000	3026	183	891	54	69	1136
SOC (nationwide expansion)	—	957,654	40,897	2466	—	—	—	—
SOC + BL (nationwide expansion)	727,113	957,654	57,962	3508	17,065	1042	43	698

[a]SOC: standard of care.

[b]SOC + BL: standard of care + blended learning.

Sensitivity Analyses

One-way sensitivity analyses identified the key drivers of uncertainty in the cost-effectiveness model. For cost per contact tested, ICERs were most sensitive to probabilities within the SOC condition, specifically, the probability that index clients identified through ART services would engage in ICT and the probability that index clients identified through HTS (ie, newly diagnosed) would opt for passive referral for their named contacts. Variations in these parameters shifted the estimated ICER for contact testing by up to ±US $34, indicating substantial leverage on cost-effectiveness.

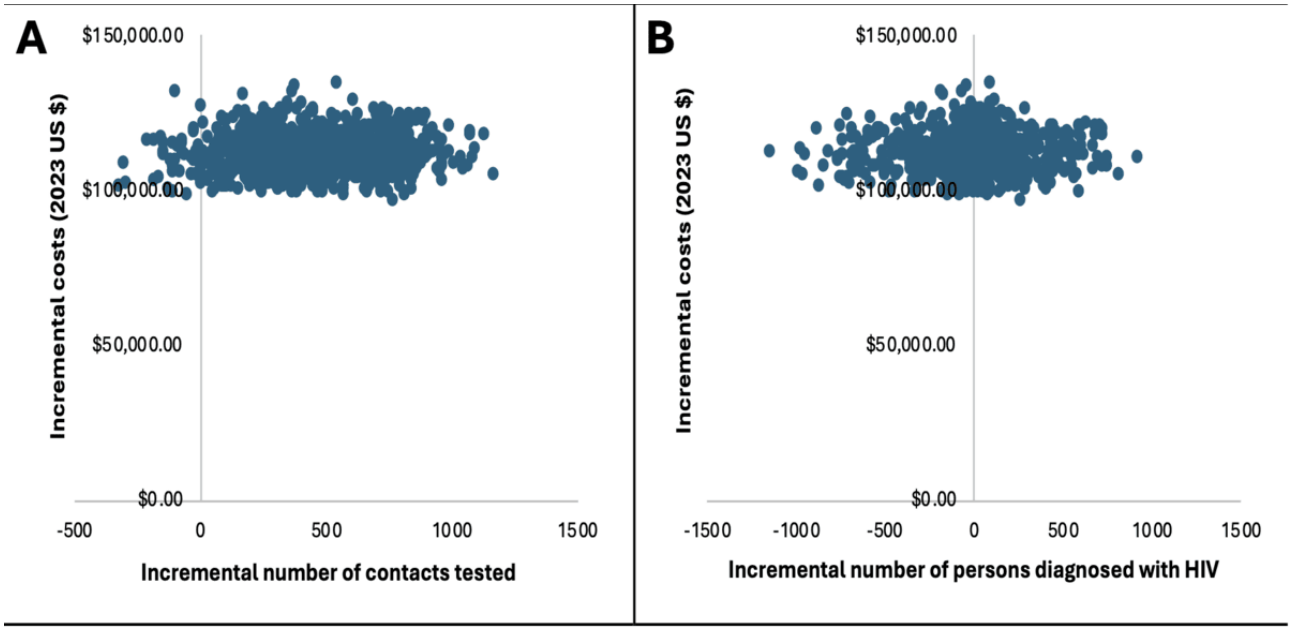
For cost per person diagnosed with HIV, ICERs were again primarily driven by probabilities estimated in the SOC condition, specifically, the probability that an index client identified through HTS would choose passive referral and the probability of ICT enrollment among individuals identified through HTS. These parameters produced the largest ICER swings for this outcome, shifting the estimated cost per person diagnosed with HIV by ±US $850. Together, these findings indicate that both cost per contact tested and cost per person diagnosed with HIV are far more sensitive to programmatic outcomes than to associated implementation costs. Overall, these findings highlight that the robustness of the cost-effectiveness estimates depends on accurate empirical data for referral pathways and ICT uptake.

The full table of 1-way sensitivity analysis results comparing SOC and SOC + BL, along with tornado diagrams showing percentage changes in ICERs across plausible parameter ranges, are provided in Multimedia Appendices 5 and 6.

Figure 2 presents cost-effectiveness planes from the probabilistic sensitivity analysis, showing the joint distribution of incremental costs and effects for SOC compared with SOC + BL. Figure 2A plots incremental cost against the number of contacts tested, while Figure 2B plots incremental cost against the number of persons diagnosed with HIV. Each point represents 1 simulation drawn from the parameter distributions.

XSL•FO
RenderX

Figure 2. Incremental cost-effectiveness ratio (ICER) plane (base-case) showing (A) incremental cost vs contact tested and (B) incremental cost vs persons diagnosed with HIV. The 4 quadrants on the ICER plane indicate different cost-effectiveness outcomes: northeast (more costly, more effective), northwest (more costly, less effective), southeast (less costly, more effective), and southwest (less costly, less effective). This figure presents the base-case scenario summarized in Table 3.



For the cost per contact tested, nearly all (966/1000, 96.6%) simulations fell in the northeast quadrant, indicating that the BL implementation package was both more effective and more costly than SOC. For the cost per person diagnosed with HIV, 58% (580/1000) of simulations were in the northeast quadrant, while 42% (420/1000) were in the northwest. These findings suggest that although the BL implementation package strategy typically improves health outcomes, its cost-effectiveness is sensitive to variations in key input parameters.

Exchange Rate Sensitivity Analysis

To assess the impact of exchange rate fluctuations during the study period, we conducted a sensitivity analysis comparing the base-case exchange rate of MWK 1036=US $1 (reflecting November 2023) with a lower exchange rate scenario of MWK 811=US $1 (reflecting November 2021). These time points correspond to the start and end of the PRACTICE study implementation period. Under the MWK 811=US $1 scenario, total program costs increased from US $111,258 to US $141,910, and cost-effectiveness estimates increased from US $125 to US $159 per contact tested and from US $2045 to US $2609 per person diagnosed with HIV. These findings indicate a moderate effect of exchange rate variability on program costs and cost-effectiveness. Multimedia Appendix 7 [36] provides additional context and assumptions.

Scenario Analyses

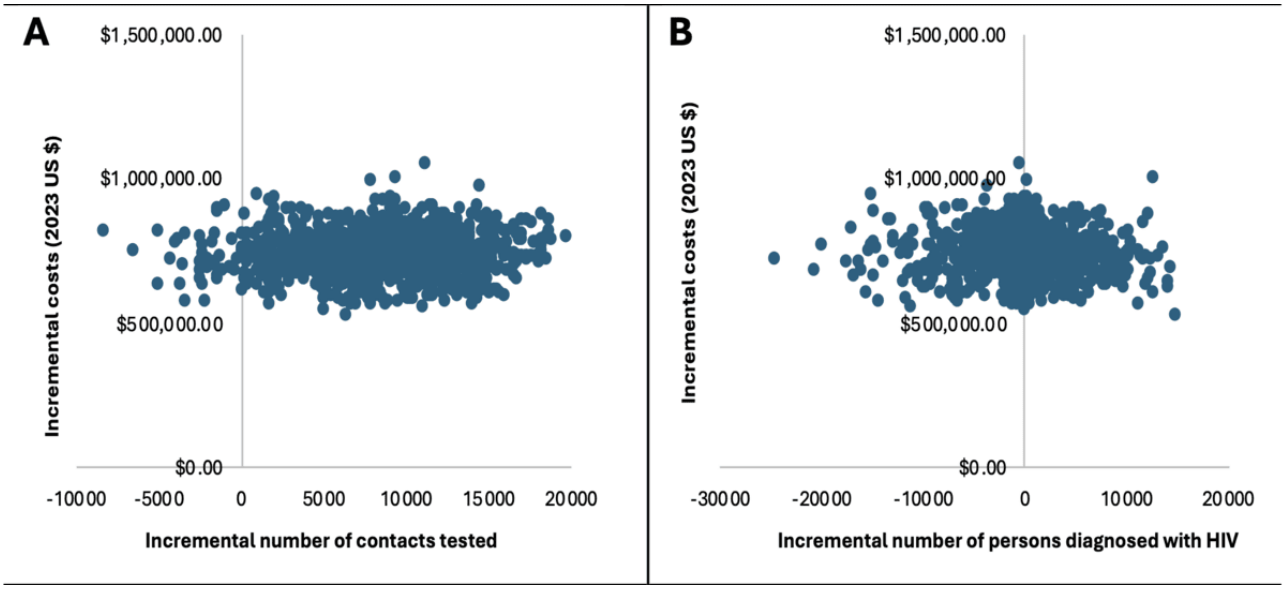
Removing strategy development costs reduced the total program cost from US $111,258 in the base case to US $61,774 (Table 3). This led to a substantial decrease in the estimated ICER, with the cost per contact tested falling by US $56 (from US $125 to US $69) and the cost per person diagnosed with HIV declining by US $909 (from US $2045 to US $1136).

In the pared-down, government-led, nationwide scale-up scenario, annual program costs increased to US $727,113, while the cost per contact tested decreased to US $43 and the cost per person diagnosed with HIV declined to US $698.

Figure 3 shows the results of the probabilistic sensitivity analysis for the nationwide expansion scenario, using the same ICER plane structure described above. For the cost per contact tested, 96.4% (964/1000) of simulations fell in the northeast quadrant, indicating that SOC + BL was more effective and more costly than SOC. For the cost per person diagnosed with HIV, just over half (552/1000, 55.2%) of simulations were in the northeast quadrant and 44.8% (448/1000) were in the northwest quadrant. No simulations fell in the southeast quadrant, indicating minimal potential for scenarios in which SOC + BL would result in both cost savings and improved outcomes relative to SOC.

XSL•FO
RenderX

Figure 3. Incremental cost-effectiveness ratio (ICER) plane (nationwide expansion scenario) showing (A) incremental cost vs contact tested and (B) incremental cost vs persons diagnosed with HIV. The 4 quadrants on the ICER plane indicate different cost-effectiveness outcomes: northeast (more costly, more effective), northwest (more costly, less effective), southeast (less costly, more effective), and southwest (less costly, less effective). This figure presents the nationwide expansion scenario summarized in Table 3.



Discussion

Building on the important clinical outcomes that improved with the BL implementation package [29,30,37], our model suggests that SOC + BL is a cost-effective strategy relative to SOC for improving contact testing and HIV case finding. National expansion scenarios support additional efficiencies of scale. While similar training approaches have been assessed in other contexts [11,38,39], our findings provide unique evidence of the economic value of integrating BL into existing HCW training for HIV testing and case-finding services. The results suggest that BL training can strengthen ICT delivery at a modest cost increase compared with the current SOC.

This study is the first to examine the cost-effectiveness of a digital or blended training intervention on patient-level outcomes. BL has been implemented across various health training contexts and has generally been found to be a less expensive strategy than traditional face-to-face training [6,38], with additional cost savings observed with multiple rounds of training [39]. However, these studies focused on proxy measures of effectiveness, such as provider competency, without directly measuring patient-level benefits. We observed that cost-effectiveness was strongly influenced by participation in ICT and uptake of active referral methods, both of which were enhanced with the training provided in the BL strategy.

We included development costs to accurately portray all costs associated with this strategy [40]. Cost-effectiveness improves when these fixed, one-time investments are either spread across broader implementation contexts or excluded from ongoing programmatic analyses. Excluding one-time development costs reduced total program costs by nearly half and substantially improved cost-effectiveness, with ICERs falling from US $125 to US $69 per contact tested and from US $2045 to US $1136 per person diagnosed with HIV. Although we did not model multiple years of implementation, we observed that these effectiveness outcomes were maintained in the second year of follow-up. Therefore, we expect that the observed ICERs would decrease when subsequent years of program implementation are modeled.

Scenario analysis provided additional insight into the potential cost dynamics and scalability of SOC + BL. A pragmatic, government-led nationwide scale-up scenario demonstrated the potential for further efficiencies through integration within existing MOH structures. As expected, expanding to nearly 900 ART facilities in all 29 districts increased total program costs for 1 year of operation. However, ICERs decreased further to US $43 per contact tested and US $698 per person diagnosed with HIV. These results suggest that decentralizing training delivery and leveraging existing staff and facilities could substantially reduce per-unit costs. However, it is important to note that our trial was conducted at sites that had support from the implementing partner, Tingathe. This represents a limitation to the generalizability of our findings, as the local implementation context may change as the funding environment from external donors shifts [41-43].

Our findings are sensitive to the epidemiologic context and related parameters and can be contextualized using benchmarks from other HIV testing and partner services interventions. A 2013 provider-led partner notification study in Malawi estimated costs of US $19 per partner tested and US $36 per new diagnosis identified [44]. However, this study was conducted a decade earlier, when a much larger share of the population was living with HIV and unaware of their status. In higher-prevalence or higher-incidence settings, or settings with less mature HIV programs in which populations are expected to have greater unmet testing needs, the cost per diagnosis would likely be lower. Notably, as donor funding supporting countries with some of the world's highest HIV burdens decreases, negative impacts on HIV services across the testing, prevention, and treatment cascades may change the cost-effectiveness of this BL implementation package.

XSL•FO
RenderX

One strength of our approach was that our effectiveness estimates were directly measured through a rigorous cluster-RCT, and this same research infrastructure facilitated robust microcosting for program elements. However, a limitation of the trial is that outcomes did not distinguish the type of elicited contact, namely sexual partners, children, or other household members, which may influence testing behaviors and outcomes. It is possible that targeting elicitation to more vulnerable contact populations could improve efficiency, particularly in Malawi's mature HIV epidemic. Ultimately, our approach reflects how ICT data are tracked in programmatic settings and provides a useful starting point for future evaluations that may explore subgroup-specific outcomes. Another limitation is that the trial did not collect data on the proportion of named contacts who were successfully traced following notification. Our base-case assumption of a 100% tracing rate among named contacts in the active referral pathway may overestimate the effectiveness of active referral if tracing rates are lower in routine practice, as fewer contacts would subsequently be tested. However, when we varied tracing-related parameters over a plausible range (70%-100%) in both univariate and probabilistic sensitivity analyses, we observed that tracing assumptions had minimal influence on overall cost-effectiveness estimates.

Our findings may generalize to other HCW training opportunities in Malawi, given the use of nationally representative input parameters and staffing data. However, a limitation regarding external validity is that other LMICs may not have the same staffing models or costs, which could affect both program effectiveness and cost-effectiveness estimates. Regardless, this study offers a replicable cost collection matrix and evaluation framework to inform cost evaluations in other countries. Like all cost-effectiveness analyses, our outcomes and the potential generalizability of the observed ICERs depend on the primary effectiveness outcome. Another limitation affecting generalizability is that all PRACTICE trial sites were supported by PEPFAR implementing partners and may have benefited from additional infrastructure support. We observed tremendous fidelity to the BL implementation package, which was expected to influence ICT effectiveness, but may not be achieved in other cadres or settings. Adaptation and evaluation will be needed in contexts with different health system supports.

Unlike targeted HIV testing interventions designed for short-term service delivery, our BL implementation package was explicitly developed as a health system strengthening intervention to support sustained ICT delivery within routine MOH structures [40]. While alternative strategies to improve HIV testing uptake, such as secondary distribution of HIV self-testing kits among men who have sex with men in Uganda, have reported lower costs per new HIV diagnosis (eg, US $325 per new diagnosis and US $147 per additional diagnosis relative to SOC) [45], these interventions were not evaluated under national scale-up scenarios. In contrast, our intervention includes comprehensive provider training and infrastructure intended to support sustainable national scale-up, offering long-term value for health system strengthening.

The BL platform may generate additional efficiencies that were not captured in our analysis. Digital training materials and devices, such as tablets, could be reused across multiple modules or adapted for different cadres of HCWs and scaled with minimal marginal cost. Even accounting for expected tablet loss or malfunction, these strategies may offer cross-training efficiencies. While our model did not account for these downstream gains, the infrastructure established for ICT could be leveraged for other clinical areas, thereby improving the return on investment. Future evaluations could examine the cost-effectiveness of BL when integrated more broadly across health service delivery programs.

The digitally supported "spaced" training model offers pedagogical advantages with potential cost implications. Traditional week-long, centralized training requires substantial expenditures for travel, accommodation, per diem, and venue costs, as well as multiple consecutive days of staff time off-site—upward of US $500 per participant, not accounting for any gaps in clinic staffing. Spaced training also lowers the marginal cost of follow-up sessions, as digital "top-ups" can be delivered with minimal additional resources and service disruption once materials are developed. Evidence from adult learning research consistently shows that distributed practice and spaced repetition improve retention and skill consolidation relative to one-off intensive training [46-48]. While we did not model these downstream effects, noting them provides context for why blended training approaches may be economically attractive when implemented at scale, particularly in the current global health funding environment, in which less development assistance for health is available.

Ultimately, our findings support the cost-effectiveness of a decentralized, system-integrated scale-up of BL to facilitate improved ICT delivery and HIV case finding, strengthening HCW capacity for service delivery in Malawi and likely other similarly low-resource settings.

Acknowledgments

The authors thank the health care workers and study participants in Malawi, as well as the Malawi Ministry of Health and Tingathe, for their collaboration and support. The authors also thank the implementation research team and study coordinators for their assistance with site-based data collection.

During the preparation of this work, PR used ChatGPT 5.1 (OpenAI) to assist with minor text editing (eg, improving sentence flow and grammar). All suggestions were reviewed for accuracy, and the author takes full responsibility for the content of the final manuscript.

XSL•FO
RenderX

Data Availability

The datasets generated or analyzed during this study are available from the corresponding author on reasonable request. Primary study data from the parent ICT implementation trial are not publicly available due to data sharing agreements and participant confidentiality restrictions but are available from the principal investigator on reasonable request and subject to appropriate data sharing agreements.

Funding

This work was supported by the National Institutes of Health (NIH; grant number R01MH124526). The funding body had no role in the design of the study, data collection, analysis, interpretation, or writing of the manuscript.

Authors' Contributions

PR and SER conceptualized the cost-effectiveness model. DV, MJC, IC, CK, and TAT supported site-level cost data extraction and review. PR led the extraction of cost data and conducted the economic analysis, with oversight from SER and NER. SER provided methodological guidance and contributed to the study design. KS provided national antiretroviral therapy data to model the nationwide expansion scenario. NER, KS, TT, VM, SA, KRM, MJC, JW, and MK conceptualized the parent trial and helped contextualize the findings. SER and PR led the drafting of the manuscript. All authors reviewed and approved the final version of the manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Summary of cost categories and components.
[DOCX File , 15 KB - mededu_v12i1e95450_app1.docx]

Multimedia Appendix 2
Contact tracing cost and workforce estimation methods.
[DOCX File , 20 KB - mededu_v12i1e95450_app2.docx]

Multimedia Appendix 3
Nationwide scenario: expanded methods.
[DOCX File , 15 KB - mededu_v12i1e95450_app3.docx]

Multimedia Appendix 4
Scenario analysis: partial reuse of content development and video production materials.
[DOCX File , 14 KB - mededu_v12i1e95450_app4.docx]

Multimedia Appendix 5
One-way sensitivity analyses comparing standard and blended learning index case testing implementation strategies.
[DOCX File , 22 KB - mededu_v12i1e95450_app5.docx]

Multimedia Appendix 6
One-way sensitivity analysis.
[DOCX File , 27 KB - mededu_v12i1e95450_app6.docx]

Multimedia Appendix 7
Exchange rate variation sensitivity analysis.
[DOCX File , 15 KB - mededu_v12i1e95450_app7.docx]

References

1. Ahsen NF, Batul SA, Ahmed AN, Imam SZ, Iqbal H, Shamshair K, et al. Developing counseling skills through pre-recorded videos and role play: a pre- and post-intervention study in a Pakistani medical school. BMC Med Educ 2010;10:7 [FREE Full text] [doi: 10.1186/1472-6920-10-7] [Medline: 20102626]
2. Gartmeier M, Bauer J, Fischer MR, Hoppe-Seyler T, Karsten G, Kiessling C, et al. Fostering professional communication skills of future physicians and teachers: effects of e-learning with video cases and role-play. Instr Sci 2015;43(4):443-462. [doi: 10.1007/s11251-014-9341-6]

XSL•FO
RenderX

3. Liu Q, Peng W, Zhang F, Hu R, Li Y, Yan W. The effectiveness of blended learning in health professions: systematic review and meta analysis. J Med Internet Res 2016;18(1):e2 [FREE Full text] [doi: 10.2196/jmir.4807] [Medline: 26729058]
4. Martino S. Strategies for training counselors in evidence-based treatments. Addict Sci Clin Pract 2010;5(2):30-39 [FREE Full text] [Medline: 22002451]
5. Nyblade L, Srinivasan K, Mazur A, Raj T, Patil DS, Devadass D, et al. HIV stigma reduction for health facility staff: development of a blended-learning intervention. Front Public Health 2018;6:165 [FREE Full text] [doi: 10.3389/fpubh.2018.00165] [Medline: 29977887]
6. Agarwal S, Perry HB, Long L, Labrique AB. Evidence on feasibility and effective use of mHealth strategies by frontline health workers in developing countries: systematic review. Trop Med Int Health 2015;20(8):1003-1014 [FREE Full text] [doi: 10.1111/tmi.12525] [Medline: 25881735]
7. Braun R, Catalani C, Wimbush J, Israelski D. Community health workers and mobile technology: a systematic review of the literature. PLoS One 2013;8(6):e65772 [FREE Full text] [doi: 10.1371/journal.pone.0065772] [Medline: 23776544]
8. Källander K, Tibenderana JK, Akpogheneta OJ, Strachan DL, Hill Z, ten Asbroek AHA, et al. Mobile health (mHealth) approaches and lessons for increased performance and retention of community health workers in low- and middle-income countries: a review. J Med Internet Res 2013;15(1):e17 [FREE Full text] [doi: 10.2196/jmir.2130] [Medline: 23353680]
9. Long LA, Pariyo G, Kallander K. Digital technologies for health workforce development in low- and middle-income countries: a scoping review. Glob Health Sci Pract 2018;6(Suppl 1):S41-S48 [FREE Full text] [doi: 10.9745/GHSP-D-18-00167] [Medline: 30305338]
10. Winters N, Langer L, Nduku P, Robson J, O'Donovan J, Maulik P, et al. Using mobile technologies to support the training of community health workers in low-income and middle-income countries: mapping the evidence. BMJ Glob Health 2019;4(4):e001421 [FREE Full text] [doi: 10.1136/bmjgh-2019-001421] [Medline: 31413872]
11. Ladur AN, Egere U, Ravit M, Mgawadere F, Murray C, White SA, et al. A blended learning approach for capacity strengthening to improve the quality of integrated HIV, TB, and malaria services during antenatal and postnatal care in LMICs: a feasibility study. BMC Med Educ 2025;25(1):35. [doi: 10.1186/s12909-024-06633-2] [Medline: 39780210]
12. Cometto G, Ford N, Pfaffman-Zambruni J, Akl EA, Lehmann U, McPake B, et al. Health policy and system support to optimise community health worker programmes: an abridged WHO guideline. Lancet Glob Health 2018;6(12):e1397-e1404 [FREE Full text] [doi: 10.1016/S2214-109X(18)30482-0] [Medline: 30430994]
13. Recommendations on digital interventions for health system strengthening. World Health Organization. 2019. URL: https://www.who.int/publications/i/item/9789241550505 [accessed 2025-05-27]
14. Han H, Myers S, Mboh Khan E, Masyuko SJ, Paredes Z, Chimoun FT, et al. Assisted HIV partner services training in three sub-saharan African countries: facilitators and barriers to sustainable approaches. J Int AIDS Soc 2019;22 Suppl 3(Suppl Suppl 3):e25307 [FREE Full text] [doi: 10.1002/jia2.25307] [Medline: 31321889]
15. Rosenberg NE, Tembo TA, Simon KR, Mollan K, Rutstein SE, Mwapasa V, et al. Development of a blended learning approach to delivering HIV-assisted contact tracing in Malawi: applied theory and formative research. JMIR Form Res 2022;6(4):e32899 [FREE Full text] [doi: 10.2196/32899] [Medline: 35438644]
16. Tembo TA, Mollan K, Simon K, Rutstein S, Chitani MJ, Saha PT, et al. Does a blended learning implementation package enhance HIV index case testing in Malawi? A protocol for a cluster randomised controlled trial. BMJ Open 2024;14(1):e077706 [FREE Full text] [doi: 10.1136/bmjopen-2023-077706] [Medline: 38253452]
17. World Health Organization. Guidelines on HIV Self-Testing and Partner Notification: Supplement to Consolidated Guidelines on HIV Testing Services. Geneva: World Health Organization; 2016:104.
18. Kariithi E, Sharma M, Kemunto E, Lagat H, Otieno G, Wamuti BM, et al. Using assisted partner services for HIV testing and the treatment of males and their female sexual partners: protocol for an implementation science study. JMIR Res Protoc 2021;10(5):e27262 [FREE Full text] [doi: 10.2196/27262] [Medline: 34014172]
19. Kariuki RM, Rithaa GK, Oyugi EO, Gachathi DM. What is the level of uptake of partner notification services in HIV testing in selected health facilities in Gatanga Sub County, Muranga County–Kenya: a retrospective study. BMC Infect Dis 2020;20(1):432 [FREE Full text] [doi: 10.1186/s12879-020-05146-9] [Medline: 32571230]
20. Sharma M, Ying R, Tarr G, Barnabas R. Systematic review and meta-analysis of community and facility-based HIV testing to address linkage to care gaps in sub-Saharan Africa. Nature 2015;528(7580):S77-S85 [FREE Full text] [doi: 10.1038/nature16044] [Medline: 26633769]
21. Tih PM, Temgbait Chimoun F, Mboh Khan E, Nshom E, Nambu W, Shields R, et al. Assisted HIV partner notification services in resource-limited settings: experiences and achievements from Cameroon. J Int AIDS Soc 2019;22 Suppl 3(Suppl Suppl 3):e25310 [FREE Full text] [doi: 10.1002/jia2.25310] [Medline: 31321902]
22. Lasry A, Medley A, Behel S, Mujawar MI, Cain M, Diekman ST, et al. Scaling up testing for human immunodeficiency virus infection among contacts of index patients - 20 countries, 2016-2018. MMWR Morb Mortal Wkly Rep 2019;68(21):474-477 [FREE Full text] [doi: 10.15585/mmwr.mm6821a2] [Medline: 31145718]
23. Watts H. Why PEPFAR is going all in on partner notification services. Enhancing HIV Testing: The Impact of Partner Notification Services. 2025 Jan 04. URL: https://www.slideserve.com/janicef/why-pepfar-is-going-all-in-on-partner-notification-services-powerpoint-ppt-presentation

XSL•FO
RenderX

24. Tembo TA, Simon KR, Kim MH, Chikoti C, Huffstetler HE, Ahmed S, et al. Pilot-testing a blended learning package for health care workers to improve index testing services in southern Malawi: an implementation science study. J Acquir Immune Defic Syndr 2021;88(5):470-476 [FREE Full text] [doi: 10.1097/QAI.0000000000002796] [Medline: 34483296]
25. Rowe AK, Rowe SY, Peters DH, Holloway KA, Chalker J, Ross-Degnan D. Effectiveness of strategies to improve health-care provider practices in low-income and middle-income countries: a systematic review. Lancet Glob Health 2018;6(11):e1163-e1175 [FREE Full text] [doi: 10.1016/S2214-109X(18)30398-X] [Medline: 30309799]
26. Kemp N, Grieve R. Face-to-face or face-to-screen? Undergraduates' opinions and test performance in classroom vs. online learning. Front Psychol 2014;5:1278 [FREE Full text] [doi: 10.3389/fpsyg.2014.01278] [Medline: 25429276]
27. O'Donovan J, O'Donovan C, Kuhn I, Sachs SE, Winters N. Ongoing training of community health workers in low-income andmiddle-income countries: a systematic scoping review of the literature. BMJ Open 2018;8(4):e021467 [FREE Full text] [doi: 10.1136/bmjopen-2017-021467] [Medline: 29705769]
28. Vallée A, Blacher J, Cariou A, Sorbets E. Blended learning compared to traditional learning in medical education: systematic review and meta-analysis. J Med Internet Res 2020;22(8):e16504 [FREE Full text] [doi: 10.2196/16504] [Medline: 32773378]
29. Rosenberg N, Mollan K, Wang J, Tembo T, Chitani M, Wetzel E, et al. Effect of a digital training package on clinical outcomes in Malawi's index case testing programme: a cluster randomised controlled trial. BMJ Glob Health 2026;11(3):e022563 [FREE Full text] [doi: 10.1136/bmjgh-2025-022563] [Medline: 41825916]
30. Mbeya-Munkhondya TE, Meek CJ, Mphande M, Tembo TA, Chitani MJ, Jean-Baptiste M, et al. Anticipated acceptability of blended learning among lay health care workers in Malawi: qualitative analysis guided by the technology acceptance model. JMIR Form Res 2025;9:e62741 [FREE Full text] [doi: 10.2196/62741] [Medline: 40194303]
31. Heller T, Damba D, Kumwenda T, Huwa J, Kamamia C, Nhlema A, et al. Implementing advanced HIV disease care for inpatients in a referral hospital in Malawi - demand, results and cost implications. Ann Glob Health 2022;88(1):16 [FREE Full text] [doi: 10.5334/aogh.3532] [Medline: 35433287]
32. Matoga MM, Rosenberg NE, Stanley CC, LaCourse S, Munthali CK, Nsona DP, et al. Inpatient mortality rates during an era of increased access to HIV testing and ART: a prospective observational study in Lilongwe, Malawi. PLoS One 2018;13(2):e0191944 [FREE Full text] [doi: 10.1371/journal.pone.0191944] [Medline: 29415015]
33. Brown LB, Miller WC, Kamanga G, Nyirenda N, Mmodzi P, Pettifor A, et al. HIV partner notification is effective and feasible in sub-Saharan Africa: opportunities for HIV treatment and prevention. J Acquir Immune Defic Syndr 2011;56(5):437-442 [FREE Full text] [doi: 10.1097/qai.0b013e318202bf7d] [Medline: 22046601]
34. Adamiak G. Methods for the economic evaluation of health care programmes, 3rd ed. J Epidemiol Community Health 2006;41(4):822-823 [FREE Full text]
35. Aalabaf-Sabaghi M. Decision modelling for health economic evaluation. J Epidemiol Community Health 2007;61(9):839 [FREE Full text]
36. Currency exchange rates. ExchangeRate.com. 2025. URL: https://www.exchangerate.com/Currency_Rates_D_April_23_2025.html [accessed 2026-06-25]
37. Meek CJ, Munkhondya TEM, Mphande M, Tembo TA, Chitani M, Jean-Baptiste M, et al. Examining the feasibility of assisted index case testing for HIV case-finding: a qualitative analysis of barriers and facilitators to implementation in Malawi. BMC Health Serv Res 2024;24(1):606 [FREE Full text] [doi: 10.1186/s12913-024-10988-z] [Medline: 38720312]
38. Yigzaw M, Tebekaw Y, Kim YM, Kols A, Ayalew F, Eyassu G. Comparing the effectiveness of a blended learning approach with a conventional learning approach for basic emergency obstetric and newborn care training in Ethiopia. Midwifery 2019;78:42-49. [doi: 10.1016/j.midw.2019.07.014] [Medline: 31349183]
39. Maloney S, Nicklen P, Rivers G, Foo J, Ooi YY, Reeves S, et al. A cost-effectiveness analysis of blended versus face-to-face delivery of evidence-based medicine to medical students. J Med Internet Res 2015;17(7):e182 [FREE Full text] [doi: 10.2196/jmir.4346] [Medline: 26197801]
40. Sohn H, Tucker A, Ferguson O, Gomes I, Dowdy D. Costing the implementation of public health interventions in resource-limited settings: a conceptual framework. Implement Sci 2020;15(1):86 [FREE Full text] [doi: 10.1186/s13012-020-01047-2] [Medline: 32993713]
41. Impact of US funding cuts on HIV programmes in Malawi. UNAIDS. 2025. URL: https://www.unaids.org/en/resources/presscentre/featurestories/2025/march/20250327_Malawi_fs [accessed 2025-12-17]
42. Chewere L. Lessons from Malawi on reprioritizing HIV Services amidst funding cuts and ensuring a sustainable response. International AIDS Society. URL: https://www.iasociety.org/sites/default/files/EduFund/meeting-material/mozambique/Lessons_from_Malawi_on_reprioritizing_HIV_Services_amidst_funding_cuts_and_ensuring_a_sustainable_response.pdf [accessed 2025-12-17]
43. Lankiewicz E, Sharp A, Drake P, Sherwood J, Macharia B, Ighodaro M, et al. Early impacts of the PEPFAR stop-work order: a rapid assessment. J Int AIDS Soc 2025;28(2):e26423. [doi: 10.1002/jia2.26423] [Medline: 39964153]
44. Rutstein SE, Brown LB, Biddle AK, Wheeler SB, Kamanga G, Mmodzi P, et al. Cost-effectiveness of provider-based HIV partner notification in urban Malawi. Health Policy Plan 2014;29(1):115-126 [FREE Full text] [doi: 10.1093/heapol/czs140] [Medline: 23325584]

XSL•FO
RenderX

45. Okoboi S, Castelnuovo B, Van Geertruyden J, Lazarus O, Vu L, Kalibala S, et al. Cost-effectiveness of peer-delivered HIV self-tests for MSM in Uganda. Front Public Health 2021;9:651325 [FREE Full text] [doi: 10.3389/fpubh.2021.651325] [Medline: 33816426]
46. Ericsson KA. Deliberate practice and the acquisition and maintenance of expert performance in medicine and related domains. Acad Med 2004;79(10 Suppl):S70-S81. [doi: 10.1097/00001888-200410001-00022] [Medline: 15383395]
47. Roediger HL, Karpicke JD. Test-enhanced learning: taking memory tests improves long-term retention. Psychol Sci 2006;17(3):249-255. [doi: 10.1111/j.1467-9280.2006.01693.x] [Medline: 16507066]
48. Kang SHK. Spaced repetition promotes efficient and effective learning. Policy Insights Behav Brain Sci 2016;3(1):12-19. [doi: 10.1177/2372732215624708]

Abbreviations

ART: antiretroviral therapy
BL: blended learning
CFIR: Consolidated Framework for Implementation Research
CHAM: Christian Health Association of Malawi
CQI: continuous quality improvement
HCW: health care worker
HTS: HIV testing services
ICER: incremental cost-effectiveness ratio
ICT: index case testing
LMIC: low- and middle-income country
MOH: Ministry of Health
MWK: Malawian Kwacha
PEPFAR: US President's Emergency Plan for AIDS Relief
PRACTICE: Package of Resources for Assisted Contact Tracing: Implementation, Costs, and Effectiveness
RCT: randomized controlled trial
SOC: standard of care
TOT: training of trainers
WHO: World Health Organization

Edited by A Stone; submitted 19.Mar.2026; peer-reviewed by N Puttkammer, SD Ernest; comments to author 11.May.2026; revised version received 09.Jun.2026; accepted 19.Jun.2026; published 27.Aug.2026.

Please cite as:
Rawat P, Rosenberg NE, Mollan KR, Tembo TA, Vansia D, Wang J, Chitani MJ, Kim MH, Chirombo I, Kumbuyo C, Mwapasa V, Simon KR, Ahmed S, Rutstein SE
Cost-Effectiveness of a Blended Learning Training Approach for Health Care Workers Conducting HIV Index Case Testing in Malawi: Secondary Economic Analysis of a Cluster-Randomized Controlled Trial
JMIR Med Educ 2026;12:e95450
URL: https://mededu.jmir.org/2026/1/e95450
doi:10.2196/95450
PMID:

©Poonam Rawat, Nora E Rosenberg, Katie R Mollan, Tapiwa A Tembo, Dhrutika Vansia, Jiayu Wang, Mike J Chitani, Maria H Kim, Ireen Chirombo, Caroline Kumbuyo, Victor Mwapasa, Katherine R Simon, Saeed Ahmed, Sarah E Rutstein. Originally published in JMIR Medical Education (https://mededu.jmir.org), 27.Aug.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Education, is properly cited. The complete bibliographic information, a link to the original publication on https://mededu.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Effects of an Immersive 180° Video–Based Virtual Reality Instructional Module on Learning Satisfaction, Technology Acceptance, Motivation, and Quiz-Based Learning Achievement in Undergraduate Physical Therapy Students: Randomized Controlled Trial

Hyoshil Yoon[1], MS, PT; Yongwoo Lee[2], PT, PhD, DC

[1]Department of Physical Therapy, Graduate School, Sahmyook University, Seoul, Republic of Korea

[2]Department of Physical Therapy, Sahmyook University, 815 Hwarang-ro, Nowon-gu, Seoul, Republic of Korea

Corresponding Author:
Yongwoo Lee, PT, PhD, DC
Department of Physical Therapy, Sahmyook University, 815 Hwarang-ro, Nowon-gu, Seoul, Republic of Korea

Abstract

Background: Virtual reality (VR) offers immersive learning opportunities in health professions education, yet its effects on learner experience and short-term learning outcomes in education on execution of musculoskeletal special tests remain underexplored.

Objective: This randomized controlled trial evaluated the effects of an immersive 180° video–based VR instruction, compared with conventional 2D video–based instruction, on undergraduate physical therapy students' learning satisfaction, technology acceptance, learning motivation, and quiz-based learning achievement, reflecting short-term procedural knowledge.

Methods: Undergraduate physical therapy students were randomly assigned to either the immersive 180° video–based VR instruction group (VR group) or the conventional 2D video–based instruction group (control group). Both groups received identical instructional content on 4 musculoskeletal special tests (pronator teres, Hawkins-Kennedy, Yergason, and Neer tests). The VR group participated in an immersive 180° video–based instructional module delivered via a head-mounted display, whereas the control group viewed the same content on a standard monitor. Participants completed 4 instructional sessions within a single day, each lasting approximately 3 to 4 minutes. The primary outcome was learning satisfaction. Secondary outcomes were technology acceptance, learning motivation, and quiz-based learning achievement. These outcomes were assessed before and after the intervention using validated questionnaires and a 12-item quiz. Between-group posttest differences were primarily analyzed using analysis of covariance (ANCOVA) with pretest scores as covariates. Supplementary within-group and unadjusted between-group analyses were also performed.

Results: Fifty-two students were randomly assigned to receive either immersive 180° video–based VR instruction (VR group; n=28) or conventional 2D video–based instruction (control group; n=24) on 4 musculoskeletal special tests, with all 4 instructional sessions completed within a single day. After adjustment for baseline scores, the VR group showed significantly higher learning satisfaction (adjusted mean 107.04 vs 93.28; $F_{1,49}$=13.710; P=.001; ηp^2=0.219) and technology acceptance (adjusted mean 74.44 vs 68.03; $F_{1,49}$=4.402; P=.04; ηp^2=0.082) than the control group. Adjusted between-group differences were not significant for quiz-based learning achievement ($F_{1,49}$=1.175; P=.28). For learning motivation, the homogeneity of regression slopes assumption was violated; therefore, the adjusted estimate was interpreted descriptively and did not indicate a clear between-group advantage ($F_{1,48}$=1.729; P=.20).

Conclusions: Compared with conventional 2D video–based instruction, immersive 180° video–based VR instruction showed a clear advantage in learning satisfaction, while technology acceptance also favored the VR group in exploratory secondary analysis, among undergraduate physical therapy students. Findings for learning motivation and quiz-based learning achievement were less conclusive. These results support immersive video as a supplementary educational approach rather than as evidence of superior psychomotor skills training.

Trial Registration: ClinicalTrials.gov NCT07003217; https://clinicaltrials.gov/study/NCT07003217

(JMIR Med Educ 2026;12:e94122) doi:10.2196/94122

XSL•FO
RenderX

KEYWORDS

virtual reality; physical therapy; physiotherapy; learning satisfaction; technology acceptance; learning motivation; undergraduate students

Introduction

With the rapid advancement of digital technologies, technology-enhanced learning has emerged as a transformative approach in health professions education [1]. In particular, virtual reality (VR) has gained recognition as a useful educational tool, creating immersive and learner-centered learning experiences for both instructors and students [2-4].

The use of immersive technologies in health professions education has expanded substantially, with applications ranging from procedural simulation to skills training. Recent developments in head-mounted displays have increased accessibility and instructional flexibility in these learning environments [4-7].

Several studies have found that VR-enhanced learning environments increase learner enjoyment and engagement, particularly in anatomy education [7]. According to self-determination theory, such enjoyable and autonomous learning contexts can foster intrinsic motivation and improve educational outcomes [8].

In physical therapy education, where hands-on clinical skills are essential, immersive digital media may support early familiarization with procedural steps by allowing learners to observe anatomical landmarks, examiner hand placement, patient positioning, and movement sequence in a more engaging format [9-12]. In particular, musculoskeletal special tests—used to diagnose orthopedic conditions—are fundamental skills for physical therapists in clinical settings [13]. Given the complex anatomical relationships involved in these assessments, learners must be able to understand when and how these tests are performed with both accuracy and confidence. Emerging evidence suggests that immersive video–based and other VR-supported educational approaches may be useful in physiotherapy education, although their effects may vary according to the level of interactivity and instructional design [10,14]. Rather than serving as a comprehensive practical skills curriculum, immersive 180° video–based VR may be particularly useful as a brief supplementary instructional approach for early familiarization with clinical procedures. By providing a first-person view of anatomical landmarks, examiner hand placement, patient positioning, and procedural sequence, such immersive video may support concrete observational exposure and learner engagement, which are broadly consistent with experiential learning principles [15]. In addition, a standardized immersive format may help novice learners focus on key procedural elements before hands-on practice [16,17]. From a contemporary instructional design perspective, immersive 360° or spherical video should be distinguished from fully interactive VR simulation, as immersion alone does not provide the same level of interactivity, feedback, or practice opportunity [18]. Within this fidelity spectrum, immersive 180° video–based VR can be positioned between conventional 2D video and fully interactive VR simulation, offering greater visual immersion than standard video but without task manipulation, haptic feedback, or real-time performance feedback. Thus, the present intervention was conceptualized as an immersive video–based instructional module rather than as fully interactive VR simulation and was designed to support early familiarization with musculoskeletal special tests rather than to provide comprehensive psychomotor skills training.

Although numerous VR studies have explored its application in rehabilitation for patients with stroke or degenerative arthritis [19,20], evidence regarding immersive 180° video–based VR instruction for undergraduate physical therapy students remains limited [10,16]. The pedagogical potential of immersive 180° video–based VR delivered through a head-mounted display, as distinct from fully interactive VR simulation, remains underexplored in musculoskeletal special tests education. Further evidence is needed to support its curricular integration.

Therefore, this study aimed to compare the effects of an immersive 180° video–based VR instructional module vs conventional 2D video–based instruction on learning satisfaction, technology acceptance, learning motivation, and quiz-based learning achievement, reflecting short-term procedural knowledge, among undergraduate physical therapy students receiving education on musculoskeletal special tests.

Methods

Study Design

This study adopted a single-blinded, parallel-group randomized controlled trial design to compare the effects of an immersive 180° video–based VR instructional module versus conventional 2D video–based instruction on key learning outcomes in undergraduate physical therapy students. No important changes to the trial design, eligibility criteria, intervention procedures, or outcome assessment procedures were made after trial commencement. The study protocol was registered at ClinicalTrials.gov (NCT07003217) on June 3, 2025, after participant recruitment and data collection had been completed; therefore, the trial registration was retrospective. The delayed registration occurred because the study was initially conducted as a single-institution educational intervention study and was not prospectively registered before recruitment; public trial registration was completed later during manuscript preparation. ClinicalTrials.gov is a primary registry of the World Health Organization International Clinical Trials Registry Platform and is recognized by the International Committee of Medical Journal Editors. The study was reported in accordance with the CONSORT-EHEALTH (Consolidated Standards of Reporting Trials of Electronic and Mobile Health Applications and Online Telehealth) V1.6 checklist [21], and the completed checklist is provided as Checklist 1.

Ethical Considerations

This study was conducted in accordance with the ethical principles of the Declaration of Helsinki. Ethical approval was

XSL•FO
RenderX

obtained from the institutional review board of Sahmyook University (SYU 2023-11-001-002). All participants were informed about the purpose and procedures of the study and provided written informed consent prior to participation.

Participants

A total of 53 undergraduate physical therapy students were recruited. Participants were recruited from third- and fourth-year undergraduate physical therapy students because these students had already received theoretical instruction on the target musculoskeletal special tests. Inclusion criteria were (1) enrollment in the physical therapy program, (2) prior theoretical instruction on the 4 target musculoskeletal special tests, and (3) no "severe" symptoms on the Simulator Sickness Questionnaire (SSQ; Multimedia Appendix 1). Exclusion criteria included neurological or musculoskeletal conditions limiting VR use, pregnancy, or refusal to participate.

The required sample size was estimated using G*Power (version 3.1.9.7; Heinrich Heine University). Because prior effect size estimates specific to immersive 180° video–based VR instruction in undergraduate physical therapy education were limited, a moderate-to-large effect size (Cohen d=0.70) was selected for the sample size calculation. Using a 2-tailed test with a statistical power of 0.80 and a significance level of .05, the analysis indicated that a minimum of 52 participants was required.

Randomization and Blinding

Participants were randomly assigned to either the immersive 180° video–based VR instruction group (VR group) or the conventional 2D video–based instruction group (control group) using a computer-generated randomization sequence. Allocation was performed by an independent researcher not involved in data collection. Outcome assessors were blinded to group allocation.

Educational Content: Musculoskeletal Special Tests

The intervention targeted 4 clinically relevant upper extremity special tests commonly used in orthopedic physical therapy. These were selected based on their diagnostic utility and inclusion in standard clinical curricula.

- Pronator teres test: used to assess compression of the median nerve by the pronator teres muscle, distinguishing pronator teres syndrome from carpal tunnel syndrome.
- Hawkins-Kennedy test: a provocative test for subacromial impingement, conducted through passive internal rotation of the shoulder at 90° of flexion.
- Yergason test: evaluates the integrity of the transverse humeral ligament and identifies pathology involving the long head of the biceps tendon or the glenoid labrum.
- Neer test: detects impingement involving the supraspinatus or long head of the biceps brachii through passive forward flexion of the internally or externally rotated arm.

Educational modules emphasized anatomical landmarks, examiner hand placement, patient positioning, proper test execution, and interpretation of positive signs. The 4 musculoskeletal special tests were established clinical examination procedures; however, the instructional video content used was custom-produced for the study. The procedures were recorded based on standard musculoskeletal examination methods and then produced as immersive 180° VR video content by a professional video production company. The instructional videos were not commercially available educational materials. Instructional content was standardized across both groups to ensure consistency. This intervention was designed as a brief supplementary instructional module rather than a comprehensive practical skills curriculum. Its educational purpose was to provide standardized immersive observational exposure to key procedural elements of each special test, including anatomical landmarks, examiner hand placement, patient positioning, movement sequence, and interpretation of positive signs. Repeated hands-on practice, performance feedback, and formal retention assessment were not incorporated into the intervention design.

Intervention Procedures

Participants completed 4 special test instructional sessions within a single day. To accommodate all participants, data collection was conducted over 4 consecutive days. Each session included 1 test and lasted approximately 3 to 4 minutes. The single-day format was selected to align with the study aim of evaluating immediate learner experience and short-term procedural knowledge after a brief supplementary familiarization module, while minimizing variability in exposure conditions across participants. Interventions were conducted in a laboratory setting.

Immersive 180° Video–Based VR Instruction (VR Group)

Participants wore a head-mounted display (Pico 4 Enterprise [Pico Technology]) and used handheld controllers to access prerecorded immersive 180° VR videos for each special test. All participants were instructed on device use, hygiene, and safety protocols prior to the sessions. The immersive video content was custom-produced for this study by a professional video production company and was preloaded onto the head-mounted displays before each session. Within the headset environment, participants used the handheld controllers to adjust the viewing direction (up, down, left, and right), zoom in and out, pause and replay the videos, and move to the next segment. Thus, the intervention allowed limited user control over viewpoint and playback, but it did not include virtual object manipulation, hand tracking, haptic feedback, task-performance tracking, or real-time performance feedback. Detailed production specifications, including camera model, recording resolution, frame rate, stabilization software, and audio-production characteristics, were not available to the authors because the videos were externally produced by a professional vendor. Device-level playback characteristics such as latency and rendering-pipeline parameters were not independently measured in this study.

Conventional 2D Video–Based Instruction (Control Group)

Participants received the same instructional content via 2D video on a standard monitor while seated in a classroom setting. Session duration and frequency were identical to those of the VR group.

XSL•FO
RenderX

Accordingly, the 4 short sessions were intended to support early procedural familiarization in a standardized format rather than to replace hands-on skills practice. A summary of the intervention is shown in Table 1.

Table . Purpose and procedures of musculoskeletal special tests included in VR[a] group and control group.

Special test	Clinical purpose	Testing procedure description
Pronator teres test	To detect median nerve compression by the pronator teres muscle (pronator teres syndrome)	With the elbow flexed to 90°, the patient performs forearm pronation while the examiner applies counter-resistance.
Hawkins-Kennedy test	To identify subacromial impingement syndrome	With the shoulder and elbow flexed to 90°, the examiner passively internally rotates the shoulder.
Yergason test	To assess stability of the biceps tendon and integrity of the glenoid labrum	With the elbow flexed to 90° and adducted, the patient attempts to supinate and externally rotate the forearm against resistance.
Neer test	To detect impingement of the supraspinatus or long head of the biceps brachii	The examiner passively flexes the shoulder with the arm internally or externally rotated, noting pain provocation.

[a]VR: virtual reality.

Outcome Measures

All outcomes were assessed before (pretest) and immediately after (posttest) the single-day intervention per participant using validated instruments, as described in Multimedia Appendices 2-5. All outcome assessments were conducted face-to-face in the laboratory or classroom setting and were not administered through an online platform. The primary outcome was learning satisfaction. Secondary outcomes were technology acceptance, learning motivation, and quiz-based learning achievement.

- Learning satisfaction: measured by a 19-item scale evaluating system experience, emotional response, and learning satisfaction (Cronbach α=0.91), using a 7-point Likert scale [1,22]
- Technology acceptance: assessed with 13 items (6 for usefulness, 7 for ease of use), validated in prior studies (Cronbach α=0.95 and 0.94), using a 7-point Likert scale [23]
- Learning motivation: evaluated through 7 items measuring interest, perceived importance, and intention to learn (Cronbach α=0.79), also on a 7-point Likert scale [23]
- Quiz-based learning achievement: assessed with a 12-item quiz (short-answer and multiple-choice), covering all 4 tests. Questions were reviewed for content validity by 2 physical therapy education experts. Identical quizzes were used for pre- and posttesting. This measure was intended to assess short-term procedural knowledge related to the 4 special tests rather than direct psychomotor performance. No delayed posttest or retention assessment was conducted

Participant Feedback

Participant feedback was collected immediately after the intervention using open-ended written response items included in the postintervention questionnaire. The feedback was optional and was intended to provide descriptive contextual information regarding participants' experiences with the instructional method. Participants were asked to comment on perceived advantages, disadvantages, discomfort, and overall impressions of the instructional format. Responses were anonymized before analysis. Two researchers independently reviewed the written responses, identified recurring meaning units, and grouped similar responses into descriptive themes. Any differences in theme grouping were resolved through discussion. Because the feedback was collected through brief written responses rather than interviews or focus groups, the findings were summarized descriptively and were not intended to represent a full qualitative study.

Statistical Analysis

Data were analyzed using SPSS (version 22.0; IBM Corp), with statistical significance set at $P<.05$. Baseline homogeneity was assessed using independent *t* tests or Mann-Whitney *U* tests for continuous variables and χ^2 tests for categorical variables, as appropriate. The primary between-group analysis was conducted using analysis of covariance (ANCOVA) on posttest scores, with the corresponding pretest scores entered as covariates, after verifying the assumptions of homogeneity of variances and homogeneity of regression slopes. Partial eta-squared (ηp^2) was reported as an effect size for the ANCOVA results. Supplementary analyses included paired *t* tests (or Wilcoxon signed rank tests for nonnormal data) for within-group pre-post changes and independent *t* tests (or Mann-Whitney *U* tests for nonnormal data) for unadjusted between-group comparisons. Because learning satisfaction was specified as the primary outcome and the remaining outcomes were secondary, no formal multiplicity adjustment was applied. The secondary outcomes were considered exploratory and were interpreted more cautiously than the primary outcome.

Results

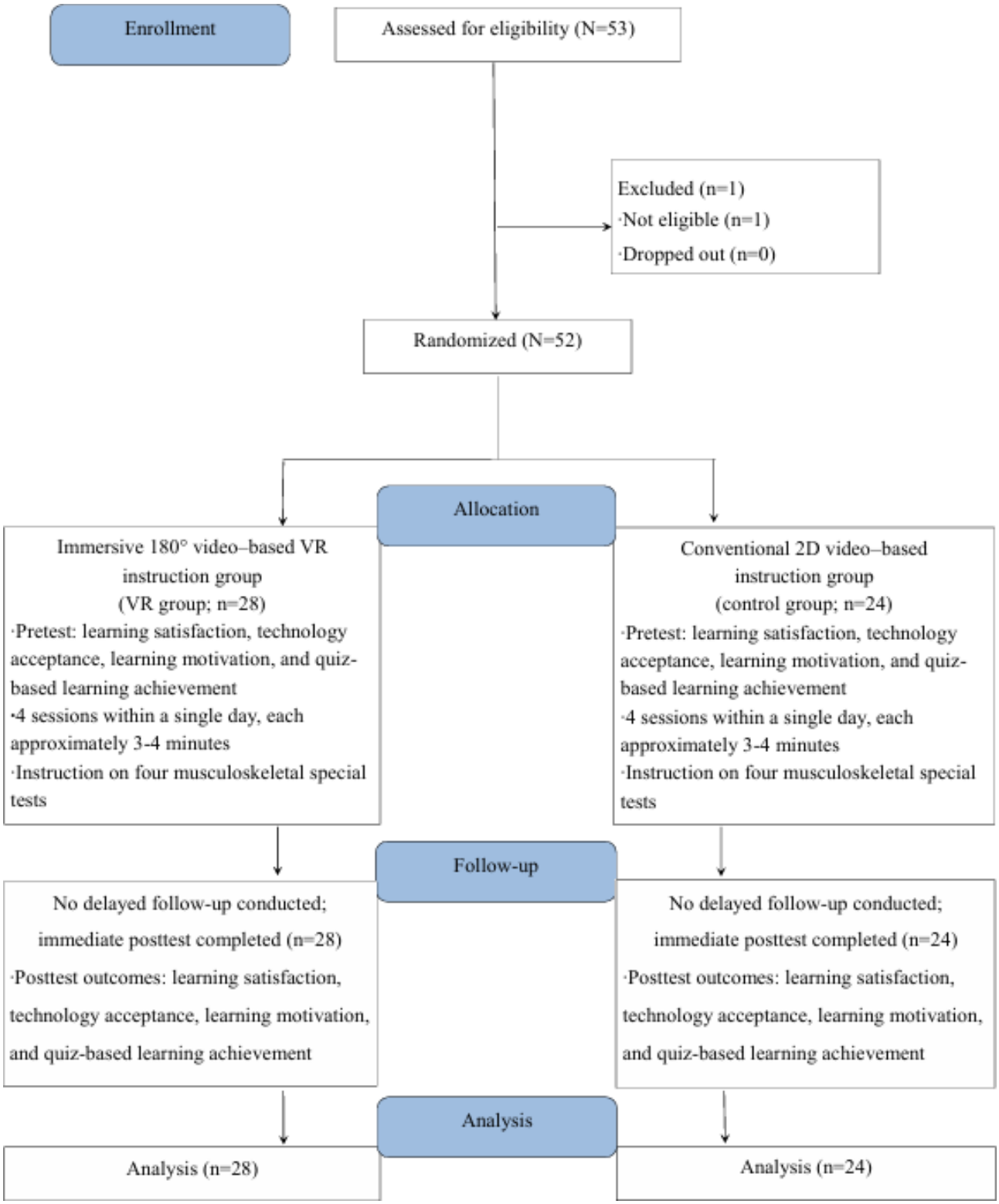
Participant Characteristics and Baseline Comparisons

Of 53 students recruited, 1 was excluded because the eligibility criteria were not met, and 52 were randomly assigned to the VR group (n=28; 10 males and 18 females) or the control group (n=24; 13 males and 11 females). The CONSORT participant flow diagram is presented in Figure 1. All randomized participants received the assigned intervention and were included

XSL•FO
RenderX

in the primary outcome analysis. Baseline demographic characteristics showed a significant between-group difference in age, but no significant differences in height, weight, SSQ scores, or sex distribution (age: *P*=.02; height: *P*=.32; weight: *P*=.94; SSQ total score: *P*=.61; sex: *P*=.18) (Table 2).

Figure 1. CONSORT (Consolidated Standards of Reporting Trials) participant flow diagram.



XSL•FO
RenderX

Table . Baseline characteristics of participants (N=52). Between-group comparisons for continuous variables were conducted using independent *t* tests.

Variable	Virtual reality group (n=28)	Control group (n=24)	*P* value
Age (years), mean (SD)	21.68 (1.59)	24.00 (4.91)	.02
Height (cm), mean (SD)	166.82 (8.25)	169.29 (8.79)	.32
Weight (kg), mean (SD)	62.07 (12.30)	62.33 (12.22)	.94
SSQ[a] total score, mean (SD)	4.67 (6.77)	3.71 (5.60)	.61
Male:female (ratio)	10:18	13:11	.18[b]

[a]SSQ: Simulator Sickness Questionnaire.

[b]*P* value from χ^2 test for sex distribution.

Baseline Equivalence of Outcome Measures

Preintervention scores for learning satisfaction (*P*=.30), technology acceptance (*P*=.16), learning motivation (*P*=.68), and quiz-based learning achievement (*P*=.26) were statistically equivalent between the 2 groups, confirming baseline homogeneity (Table 3).

Table . Baseline equivalence of outcome measures between groups (N=52).

Outcome variable	Virtual reality group, mean (SD)	Control group, mean (SD)	*P* value
Learning satisfaction	89.96 (12.30)	83.54 (11.32)	.30
Technology acceptance	63.43 (11.84)	59.33 (8.22)	.16
Learning motivation[a]	40.29 (5.30)	40.72 (3.46)	.68
Quiz-based learning achievement	0.68 (0.98)	0.98 (0.91)	.26

[a]For learning motivation, the unequal variances *t* test result was used because Levene test indicated unequal variances.

Learning Satisfaction

An ANCOVA on posttest scores, controlling for baseline, revealed a significant effect for group. The VR group (adjusted mean 107.04) scored significantly higher than the control group (adjusted mean 93.28; $F_{1,49}$=13.710; *P*=.001; ηp^2=0.219), with an adjusted mean difference of 13.762 (95% CI 6.293-21.231). An unadjusted *t* test on posttest scores confirmed this significant difference (*P*<.001) (Tables 4 and 5).

Table . Primary analysis: analysis of covariance (ANCOVA) results for posttest scores (N=52).

Variable and group	Adjusted mean[a]	Adjusted mean difference (95% CI)	Between-group ANCOVA, *F* test (*df*)	*P* value	Partial eta-squared (ηp^2)
Learning satisfaction		13.762 (6.293 to 21.231)	13.710 (1, 49)	.001	0.219
Virtual reality	107.04				
Control	93.28				
Technology acceptance		6.410 (0.270 to 12.549)	4.402 (1, 49)	.04	0.082
Virtual reality	74.44				
Control	68.03				
Learning motivation		1.688 (−1.131 to 4.508)[b]	1.729 (1, 48)[b]	.20[b]	0.035
Virtual reality	43.66				
Control	41.98				
Quiz-based learning achievement		0.833 (−0.712 to 2.378)	1.175 (1, 49)	.28	0.023
Virtual reality	5.94				
Control	5.11				

[a]Adjusted mean: adjusted posttest mean.

[b]Homogeneity of regression slopes was violated for learning motivation (interaction: $F_{1,48}$=4.328; *P*=.04; ηp^2=0.083); therefore, this result should be interpreted with caution and was retained as a descriptive adjusted estimate rather than a confirmatory ANCOVA finding.

XSL•FO
RenderX

Table . Supplementary analysis: unadjusted within-group and between-group comparisons (N=52).

Variable and group	Pretest	Posttest	Within-group comparison[a] (pre- vs posttest), *t* test (*df*)	Within-group comparison[a] (pre- vs posttest), *P* value	Between-group comparison[b] (posttest), *t* test (*df*)	Between-group comparison[b] (posttest), *P* value
Learning satisfaction, mean (SD)					4.508 (50)	<.001
Virtual reality	89.96 (12.30)	107.82 (14.19)	−6.398 (27)	<.001		
Control	83.54 (11.32)	91.25 (11.97)	−3.286 (23)	.003		
Technology acceptance, mean (SD)					2.828 (50)	.007
Virtual reality	63.43 (11.84)	75.29 (10.86)	−4.053 (27)	<.001		
Control	59.33 (8.22)	66.75 (10.84)	−3.798 (23)	.001		
Learning motivation, mean (SD)					1.721 (50)	.09
Virtual reality	40.29 (5.30)	44.07 (5.02)	−2.444 (27)	.02		
Control	40.72 (3.46)	41.71 (4.83)	−1.399 (23)	.18		
Quiz-based learning achievement[c], median (IQR)					267.500	.21[d]
Virtual reality	0.00 (0.00 - 2.00)	5.75 (3.63 - 7.88)	−4.625[e]	<.001		
Control	1.00 (0.00 - 2.00)	5.00 (2.63-8.00)	−3.682[e]	<.001		

[a]Within-group comparison: paired *t* test or Wilcoxon signed rank test.

[b]Between-group comparison: independent *t* test.

[c]Nonparametric tests were used for quiz-based learning achievement due to nonnormality.

[d]Between-group comparison: Mann-Whitney *U* test.

[e]*z* value from Wilcoxon signed rank test.

Technology Acceptance

After adjusting for pretest scores, the VR group (adjusted mean 74.44) showed significantly higher acceptance than the control group (adjusted mean 68.03; $F_{1,49}$=4.402; *P*=.04; ηp^2=0.082) with an adjusted mean difference of 6.41 (95% CI 0.270-12.549). This finding was consistent with the unadjusted *t* test on posttest scores (*P*=.007) (Tables 4 and 5).

Learning Motivation

The ANCOVA assumption of homogeneity of regression slopes was violated (interaction effect: $F_{1,48}$=4.328; *P*=.04). At the mean baseline, the adjusted difference between the VR group (adjusted mean 43.66) and the control group (adjusted mean 41.98) was not significant ($F_{1,48}$=1.729; *P*=.20; ηp^2=0.035) with an adjusted mean difference of 1.688 (95% CI −1.131 to 4.508), a finding consistent with the unadjusted *t* test result (*P*=.09) (Tables 4 and 5).

Quiz-Based Learning Achievement

There was no significant difference in quiz-based learning achievement between the VR group (adjusted mean 5.94) and the control group (adjusted mean 5.11) after controlling for pretest scores ($F_{1,49}$=1.175; *P*=.28; ηp^2=0.023) with an adjusted mean difference of 0.833 (95% CI −0.712 to 2.378). An unadjusted Mann-Whitney *U* test on posttest scores was also not statistically significant (*P*=.21) (Tables 4 and 5).

Participant Feedback on Instructional Methods

Written feedback was collected from 22 of 28 (79%) participants in the VR group and 10 of 24 (42%) in the control group. VR group responses emphasized increased engagement, realism, and learning interest. However, several participants noted visual discomfort or fatigue after prolonged use of the head-mounted display. Control group participants found the 2D video format familiar and easy to follow but reported perceived limitations in motivation and interactivity. A descriptive summary of participant responses is provided in Table 6.

Table . Participants' feedback on instructional methods (n=52).

Group	Respondents, n (%)	Themes identified
Virtual reality group (n=28)	22 (79)	Enhanced realism, 3D learning, personalized viewpoints, close observation, positive engagement. Some discomfort (eg, dizziness, visual fatigue, headset weight)
Control group (n=24)	10 (42)	Familiar video format, easy to follow. Reported limitations: lack of interactivity, shared viewing, perceived monotony

XSL•FO
RenderX

Discussion

This study compared a low-interactivity immersive 180° video–based VR instruction with conventional 2D video instruction for musculoskeletal special test education in undergraduate physical therapy students. The immersive video module was associated with higher learning satisfaction than 2D video, and technology acceptance also favored the VR group in exploratory secondary analysis, whereas between-group differences were not significant for learning motivation or quiz-based learning achievement.

First, the enhanced learning satisfaction observed in the VR group highlights the pedagogical value of immersive viewing environments, which may help learners more intuitively understand anatomical landmarks, examiner hand placement, patient positioning, and procedural sequence. These findings align with principles of experiential learning theory, which emphasize the role of concrete experience and reflective observation in knowledge construction [15]. This interpretation is also broadly consistent with previous studies suggesting that immersive learning environments can increase learner engagement and perceived educational value [6,16].

Second, the higher technology acceptance observed in the VR group may reflect the immersive first-person viewing experience, focused attention, and perceived novelty afforded by the head-mounted display environment rather than fully interactive simulation features [5,6]. From a self-determination theory perspective, immersive and autonomy-supportive learning environments may enhance learner interest and engagement [8]. However, because the present intervention consisted of prerecorded 180° videos with limited interactivity, these findings should not be interpreted as evidence for the effects of fully interactive VR simulation.

Third, although the VR group showed a significant within-group increase in learning motivation, the between-group difference was not statistically significant. Similarly, both groups showed within-group increases in quiz-based learning achievement, but the adjusted between-group difference was not statistically significant. Importantly, the measured outcomes did not include direct assessment of clinical skill performance, such as Objective Structured Clinical Examination–based evaluation, checklist-based observation, or other performance-based assessments. Accordingly, the present findings more strongly support the value of immersive 180° video–based VR for learner experience outcomes than for superior short-term quiz performance. Because quiz-based learning achievement was assessed using an identical pre- and posttest administered within a single day, the observed gains should be interpreted cautiously, as they may partly reflect test-retest or recall effects rather than differences attributable solely to instructional modality. In addition, because multiple outcomes were examined and adjusted between-group differences were not significant for all secondary outcomes, findings for learning motivation and quiz-based learning achievement should be interpreted cautiously.

Despite these strengths, several limitations should be considered. First, the present VR intervention used prerecorded immersive 180° videos with limited interactivity and no haptic feedback. Therefore, it was conceptually distinct from fully interactive simulation-based VR and was limited in its ability to support psychomotor skill development, task manipulation, or real-time performance feedback [10,24]. Detailed production specifications for the externally produced immersive videos were not available to the authors; therefore, some technical parameters of content capture and postproduction could not be fully reported. Future studies should investigate whether more interactive systems, including hand tracking, gesture-based interaction, or haptic feedback, produce different educational effects [24,25].

Second, the intervention was brief, with 4 instructional sessions completed within a single day. As a result, the study did not examine repeated practice, skill transfer, or longer-term retention. Future research should incorporate longitudinal designs, repeated exposure, and performance-based assessments to better evaluate the role of immersive technologies in musculoskeletal special test education [10,16].

Third, the primary and several secondary outcomes were based on self-reported questionnaires. Although outcome assessors were blinded to group allocation, participant blinding was not feasible because learners could clearly recognize whether they were using a head-mounted display or viewing a conventional 2D video. Therefore, the higher learning satisfaction and technology acceptance observed in the VR group may have been influenced by novelty effects, expectancy effects, or differences in perceived technological sophistication. Future studies should include additional objective or performance-based outcomes and, where possible, comparator interventions that better control for novelty and attention.

Fourth, this study did not control for potential differences in baseline academic ability, spatial cognition, or prior clinical exposure, which may have influenced the observed outcomes. Future research should consider these learner characteristics and explore whether specific subgroups benefit more from immersive video–based instruction [24].

Fifth, this was a single-institution study conducted with undergraduate physical therapy students in Korea, and a significant baseline age difference was observed between groups. Although the primary ANCOVA adjusted for baseline outcome scores, age was not included as a covariate; therefore, residual confounding related to age cannot be fully excluded. Therefore, the findings should be generalized cautiously to other countries, institutions, or educational systems, as well as to programs with different curricula or different levels of prior VR exposure. Future multicenter studies across diverse educational contexts are needed to examine the broader applicability of immersive 180° video–based VR instruction.

Nevertheless, this study provides preliminary evidence that an immersive 180° video–based VR format may serve as a useful supplementary educational tool in physical therapy education [10,16]. In the present study, the most robust advantage was observed for learning satisfaction, while technology acceptance also favored the VR group in exploratory secondary analysis. This type of immersive video–based instruction may be particularly useful in situations where opportunities for direct

XSL•FO
RenderX

observation are limited by resource constraints or accessibility barriers, while still being conceptually distinct from fully interactive simulation-based VR [10].

To maximize the instructional potential of immersive technologies in health professions education, future studies should also examine extended reality and spatial computing applications that integrate immersive visualization with real-time interaction, spatial mapping, and task-specific feedback, as well as repeated practice and performance-based assessments.

This randomized controlled trial demonstrated that immersive 180° video–based VR instruction may serve as a useful supplementary educational approach in undergraduate physical therapy education. Compared to conventional 2D video instruction, immersive 180° video–based VR instruction showed a clear advantage in learning satisfaction, while an exploratory advantage was observed for the VR group in technology acceptance. Findings for learning motivation and quiz-based learning achievement were less conclusive.

These findings contribute to the growing body of evidence supporting the pedagogical utility of immersive digital tools in health professions education. In the present study, the findings are most appropriately interpreted as support for learner experience outcomes rather than as evidence of superior psychomotor skills training. In this study, the immersive 180° video–based VR environment provided a safe and engaging context for students to observe procedural content related to musculoskeletal special tests; however, it did not include interactive simulation features such as hand tracking, haptic feedback, task-performance tracking, or real-time feedback.

Because this study used pre-recorded 180° VR content without tactile or interactive functionality, future research should examine whether more advanced VR systems, including haptic feedback, gesture recognition, and AI-guided instruction, produce additional benefits for psychomotor performance, skill transfer, or retention. However, because the intervention was brief and low in interactivity, and quiz-based learning achievement was assessed using a short-term quiz rather than direct performance-based assessment, the present findings should not be interpreted as evidence of improved psychomotor learning or skill transferability to clinical practice.

As physical therapy education continues to embrace digital transformation, immersive 180° video–based VR instruction may be integrated as a supplementary modality to reinforce procedural knowledge within similar undergraduate physical therapy education contexts. However, because this study was conducted at a single institution in Korea, the findings should be generalized cautiously to other countries or educational systems. Future research should explore the long-term effects of immersive 180° video–based VR instruction, using spaced practice, Objective Structured Clinical Examination–based or other performance-based assessments, more interactive VR modalities, and longer-term retention measures.

Funding

This paper was supported by the Sahmyook University Research Fund in 2025.

Data Availability

The datasets generated and/or analyzed during the current study are available from the corresponding author on reasonable request.

Authors' Contributions

Conceptualization: YL
Methodology: YL
Investigation: HY, YL
Formal analysis: HY
Writing – original draft: YL, HY
Writing – review & editing: YL, HY

Conflicts of Interest

None declared.

Multimedia Appendix 1
Simulator Sickness Questionnaire (SSQ) items and severity scores.
[DOCX File, 13 KB - mededu_v12i1e94122_app1.docx]

Multimedia Appendix 2
Learning satisfaction questionnaire.
[DOCX File, 19 KB - mededu_v12i1e94122_app2.docx]

Multimedia Appendix 3
Technology Acceptance Questionnaire.
[DOCX File, 17 KB - mededu_v12i1e94122_app3.docx]

XSL•FO
RenderX

Multimedia Appendix 4
Learning motivation questionnaire.
[DOCX File, 14 KB - mededu_v12i1e94122_app4.docx]

Multimedia Appendix 5
Quiz-based learning achievement test.
[DOCX File, 15 KB - mededu_v12i1e94122_app5.docx]

Checklist 1
CONSORT-EHEALTH Checklist.
[PDF File, 1418 KB - mededu_v12i1e94122_app6.pdf]

References

1. Chang CY, Hwang GJ. Trends in digital game-based learning in the mobile era: a systematic review of journal publications from 2007 to 2016. Int J Mobile Learning Organisation 2019;13(1):68-90. [doi: 10.1504/IJMLO.2019.096468]
2. Gunn T, Jones L, Bridge P, Rowntree P, Nissen L. The use of virtual reality simulation to improve technical skill in the undergraduate medical imaging student. Interact Learn Environ 2018 Jul 4;26(5):613-620. [doi: 10.1080/10494820.2017.1374981]
3. Geng J, Luk ETH, Jong MSY. Teachers' concerns about adopting interactive spherical video-based virtual reality. Presented at: International Conference on Computers in Education; Dec 4-8, 2017 URL: https://research.cuhk.edu.hk/en/publications/teachers-concerns-about-adopting-interactive-spherical-video-base-2/ [accessed 2026-07-24]
4. Tursø-Finnich T, Jensen RO, Jensen LX, Konge L, Thinggaard E. Virtual reality head-mounted displays in medical education: a systematic review. Simul Healthc 2023 Feb 1;18(1):42-50. [doi: 10.1097/SIH.0000000000000636] [Medline: 35136005]
5. Plotzky C, Lindwedel U, Bejan A, König P, Kunze C. Virtual reality in healthcare skills training: the effects of presence on acceptance and increase of knowledge. I-com 2021 Apr 22;20(1):73-83. [doi: 10.1515/icom-2021-0008]
6. Kurul R, Ögün MN, Neriman Narin A, Avci Ş, Yazgan B. An alternative method for anatomy training: immersive virtual reality. Anat Sci Educ 2020 Sep;13(5):648-656. [doi: 10.1002/ase.1959] [Medline: 32163659]
7. Moro C, Štromberga Z, Raikos A, Stirling A. The effectiveness of virtual and augmented reality in health sciences and medical anatomy. Anat Sci Educ 2017 Nov;10(6):549-559. [doi: 10.1002/ase.1696] [Medline: 28419750]
8. Ryan RM, Deci EL. Intrinsic and extrinsic motivation from a self-determination theory perspective: definitions, theory, practices, and future directions. Contemp Educ Psychol 2020 Apr;61:101860. [doi: 10.1016/j.cedpsych.2020.101860]
9. Cooper JB, Taqueti VR. A brief history of the development of mannequin simulators for clinical education and training. Postgrad Med J 2008 Nov;84(997):563-570. [doi: 10.1136/qshc.2004.009886] [Medline: 19103813]
10. Lucena-Anton D, Fernandez-Lopez JC, Pacheco-Serrano AI, Garcia-Munoz C, Moral-Munoz JA. Virtual and augmented reality versus traditional methods for teaching physiotherapy: a systematic review. Eur J Investig Health Psychol Educ 2022 Dec 2;12(12):1780-1792. [doi: 10.3390/ejihpe12120125] [Medline: 36547026]
11. Hartstein AJ, Zimney K, Verkuyl M, Yockey J, Berg-Poppe P. Virtual reality instructional design in orthopedic physical therapy education: a randomized controlled trial. JOPTE 2022 Jun;36(2):176-184. [doi: 10.1097/JTE.0000000000000221]
12. Marougkas A, Troussas C, Krouska A, Sgouropoulou C. Virtual reality in education: a review of learning theories, approaches and methodologies for the last decade. Electronics 2023 Jun 26;12(13):2832. [doi: 10.3390/electronics12132832]
13. Cyriax J. Textbook of Orthopaedic Medicine, Vol 1: Diagnosis of Soft Tissue Lesions, 8th edition: Baillière Tindall; 1982.
14. Foronda CL, Gonzalez L, Meese MM, et al. A comparison of virtual reality to traditional simulation in health professions education: a systematic review. Simul Healthc 2024 Jan 1;19(1S):S90-S97. [doi: 10.1097/SIH.0000000000000745] [Medline: 37651101]
15. Kolb DA. Experiential Learning: Experience as the Source of Learning and Development, 2nd edition: Pearson FT Press; 2014.
16. Hartstein AJ, Verkuyl M, Zimney K, Yockey J, Berg-Poppe P. Virtual reality instructional design in orthopedic physical therapy education: a mixed-methods usability test. Simul Gaming 2022 Apr;53(2):111-134. [doi: 10.1177/10468781211073646]
17. Naef AC, Jeitziner MM, Jakob SM, Müri RM, Nef T. Creating custom immersive 360-degree videos for use in clinical and nonclinical settings: tutorial. JMIR Med Educ 2023 Sep 14;9(1):e42154. [doi: 10.2196/42154] [Medline: 37707883]
18. Rosendahl P, Wagner I. 360° videos in education – a systematic literature review on application areas and future potentials. Educ Inf Technol 2024 Feb;29(2):1319-1355. [doi: 10.1007/s10639-022-11549-9]
19. Hong S. A study on the effectiveness of home-based virtual reality rehabilitation training content for elderly individuals with degenerative arthritis [Article in Korean]. Phys Ther Rehabil Sci 2023 Dec 30;12(4):522-528. [doi: 10.14474/ptrs.2023.12.4.522]

XSL•FO
RenderX

20. Kim JW, Kim JH, Lee BH. Effects of virtual reality-based core stabilization exercise on upper extremity function, postural control, and depression in persons with stroke. Phys Ther Rehabil Sci 2020 Sep 30;9(3):131-139. [doi: 10.14474/ptrs.2020.9.3.131]
21. Eysenbach G, CONSORT-EHEALTH Group. CONSORT-EHEALTH: improving and standardizing evaluation reports of web-based and mobile health interventions. J Med Internet Res 2011 Dec 31;13(4):e126. [doi: 10.2196/jmir.1923] [Medline: 22209829]
22. Chu HC, Hwang GJ, Tsai CC, Tseng JCR. A two-tier test approach to developing location-aware mobile learning systems for natural science courses. Comput Educ 2010 Dec;55(4):1618-1627. [doi: 10.1016/j.compedu.2010.07.004]
23. Hwang GJ, Yang LH, Wang SY. A concept map-embedded educational computer game for improving students' learning performance in natural science courses. Comput Educ 2013 Nov;69:121-130. [doi: 10.1016/j.compedu.2013.07.008]
24. Stavroulia KE, Christofi M, Zarraonandia T, Michael-Grigoriou D, Lanitis A. Virtual reality environments (VREs) for training and learning. In: Díaz P, Ioannou A, Bhagat K, Spector J, editors. Learning in a Digital World: Perspective on Interactive Technologies for Formal and Informal Education: Springer; 2019:195-211. [doi: 10.1007/978-981-13-8265-9_10]
25. Chang CC, Hwang GJ. An experiential learning-based virtual reality approach to fostering problem-resolving competence in professional training. Interact Learn Environ 2023 Nov 17;31(8):4713-4728. [doi: 10.1080/10494820.2021.1979049]

Abbreviations

ANCOVA: analysis of covariance
CONSORT-EHEALTH: Consolidated Standards of Reporting Trials of Electronic and Mobile Health Applications and Online Telehealth
SSQ: Simulator Sickness Questionnaire
VR: virtual reality

Edited by S Guinez-Molinos; submitted 25.Feb.2026; peer-reviewed by F Ng, LA Lee, S Chaichulee; revised version received 15.Jun.2026; accepted 08.Jul.2026; published 14.Aug.2026.

Please cite as:
Yoon H, Lee Y
Effects of an Immersive 180° Video–Based Virtual Reality Instructional Module on Learning Satisfaction, Technology Acceptance, Motivation, and Quiz-Based Learning Achievement in Undergraduate Physical Therapy Students: Randomized Controlled Trial
JMIR Med Educ 2026;12:e94122
URL: https://mededu.jmir.org/2026/1/e94122
doi:10.2196/94122

© Hyoshil Yoon, Yongwoo Lee. Originally published in JMIR Medical Education (https://mededu.jmir.org), 14.Aug.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Education, is properly cited. The complete bibliographic information, a link to the original publication on https://mededu.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Augmented Reality–Based Training for a Rapid Blood Transfusion Device Among Emergency Nurses: Randomized Controlled Trial

Wooyoung Jang[1], RN; Wooree Chai[1], RN; Yeji Kim[1], RN; Youjung Moon[1], RN; Suyoung Yoo[2], PhD; Sun Gyoung Na[1], RN; Soyeon Shin[1], RN; Meong Hi Son[3], MD, PhD

[1]Department of Nursing, Samsung Medical Center, Seoul, Seoul, Republic of Korea

[2]Research Institute for Future Medicine, Samsung Medical Center, Seoul, Seoul, Republic of Korea

[3]Department of Emergency Medicine, Samsung Medical Center, 81, Irwon-ro, Gangnam-gu, Seoul, Seoul, Republic of Korea

Corresponding Author:
Meong Hi Son, MD, PhD
Department of Emergency Medicine, Samsung Medical Center, 81, Irwon-ro, Gangnam-gu, Seoul, Seoul, Republic of Korea

Abstract

Background: Emergency nurses must be proficient in operating the Level-1 rapid infusion system to manage hypovolemic shock effectively. However, training opportunities for this infrequently used but life-critical device remain scarce, owing to resource constraints and limited access to equipment. Augmented reality (AR) has emerged as a promising educational technology that provides immersive, hands-on learning experiences without compromising patient safety; yet its application to specialized medical device training in nursing has not been rigorously evaluated.

Objective: This study aimed to evaluate the effects of an AR-based training program using Microsoft HoloLens 2 on emergency nurses' clinical competency, self-efficacy, and educational satisfaction in operating the Level-1 rapid infusion system, compared with traditional guideline-based self-directed learning.

Methods: A posttest-only randomized controlled trial was conducted at Samsung Medical Center in Seoul, Republic of Korea. Between July 17 and July 20, 2023, 42 registered nurses with no prior Level-1 experience were enrolled and randomly assigned in a 1:1 ratio to an experimental group receiving AR-based training on a single HoloLens 2 device (n=21) or a control group performing self-directed learning from a printed manual (n=21). Clinical competency was assessed by time (learning and performance), accuracy (a manufacturer-aligned checklist scored out of 100, and an expert-validated 22-step pass or fail evaluation), and the number of assistance requests. Self-efficacy (6-item scale; Cronbach α=0.80) and educational satisfaction (4-item scale; Cronbach α=0.87) were measured by questionnaire. Because most outcomes were non-normally distributed, groups were compared using the Mann-Whitney *U* test, with data reported as medians and IQRs.

Results: Learning time was longer in the experimental group (median 18.20, IQR 15.48 - 21.67 vs 8.98, IQR 5.85 - 11.68 min; *P*<.001), but device setup time was markedly shorter (3.67, IQR 2.90 - 4.63 vs 9.85, IQR 8.03 - 11.35 min; *P*<.001). The experimental group achieved higher median device operation competency scores (90.00, IQR 80.00 - 100.00 vs 70.00, IQR 50.00 - 75.00 of 100; *P*<.001), passed more of the 22 evaluation steps (20.00, IQR 20.00 - 22.00 vs 16.00, IQR 13.00 - 18.00; *P*<.001), and required fewer assistance requests (0.00, IQR 0.00 - 1.00 vs 2.00, IQR 2.00 - 3.00; *P*<.001). Self-efficacy (20.00, IQR 18.00 - 25.00 vs 16.50, IQR 13.00 - 20.00; *P*=.003) and educational satisfaction (18.00, IQR 16.00 - 18.00 vs 12.00, IQR 9.75 - 15.00; *P*<.001) were also significantly higher in the experimental group. Effect sizes for the principal competency outcomes were large (Cohen *d*=1.2 - 3.1).

Conclusions: AR-based training significantly improved emergency nurses' clinical competency, self-efficacy, and educational satisfaction in operating the Level-1 rapid infusion system compared with traditional self-directed learning. Despite requiring longer initial learning time, AR training produced faster device setup, greater accuracy, and enhanced learner independence. These findings suggest that AR technology can serve as an effective and scalable training solution for infrequently used but critically important medical devices in emergency care settings.

Trial Registration: ClinicalTrials.gov NCT06506851; https://clinicaltrials.gov/study/NCT06506851

(JMIR Med Educ 2026;12:e98073) doi:10.2196/98073

KEYWORDS

augmented reality; nursing education; clinical competency; self-efficacy; educational satisfaction; rapid infusion; medical device training; randomized controlled trial

XSL•FO
RenderX

Introduction

The rapid evolution of medical technology presents a persistent challenge for nursing education. New medical devices are introduced frequently, each requiring dedicated training for safe and effective operation [1,2]. Maintaining clinical competency across a growing inventory of devices is demanding, particularly for equipment that is infrequently used yet essential during emergencies [3,4]. Research consistently demonstrates that hands-on practice is the most effective approach for developing procedural proficiency [5]; however, providing individualized, device-specific training at scale remains logistically and financially prohibitive in most clinical settings.

Massive transfusion using the Level-1 rapid infusion system exemplifies this training dilemma. Although massive transfusion is performed infrequently, it is a lifesaving procedure required in urgent situations such as hypovolemic shock, severe trauma, and major surgical hemorrhage. Emergency nurses must be able to assemble, prime, and operate the Level-1 device swiftly and accurately under extreme time pressure. In practice, however, limited hands-on exposure leads to wide variability in skill levels and increased dependence on senior team members during critical resuscitation events, potentially compromising patient outcomes [6,7].

Augmented reality (AR) has emerged as a promising educational technology that addresses these constraints by overlaying digital instructional content onto the physical environment, allowing learners to practice with real equipment under guided conditions without risk to patients. A recent systematic mapping review of 156 studies published between 2010 and 2025 confirmed that AR is an effective educational tool in health sciences, improving hands-on skill acquisition, clinical competencies, and learner engagement while offering scalable, cost-effective solutions even in resource-limited settings [8]. In the specific context of medical device training, randomized trials have demonstrated that AR-based instruction for cardiopulmonary resuscitation and intubation significantly improves clinical performance compared with traditional methods [9]. Furthermore, AR supports the standardization of training by delivering consistent instructional content regardless of instructor expertise, which is particularly advantageous for devices that individual instructors may themselves rarely operate [10].

Despite this growing body of evidence, no prior study has specifically evaluated AR-based training for the Level-1 rapid infusion system. More broadly, the application of AR to training for infrequently used but critically important medical devices in nursing remains underexplored. Beyond individual skill development, AR-based training has implications for nursing management; standardized digital training programs can reduce dependence on instructor availability, ensure uniform skill acquisition across nursing teams, and support continuous professional development in resource-constrained environments [11-13]. Understanding whether AR can effectively address these dual objectives, enhancing individual competency while supporting organizational training goals, is essential for informed adoption decisions.

This study aimed to evaluate the effects of an AR-based training program using Microsoft HoloLens 2 on emergency nurses' clinical competency, self-efficacy, and educational satisfaction in operating the Level-1 rapid infusion system. We hypothesized that AR-based training would produce superior outcomes across all 3 domains compared with traditional guideline-based self-directed learning.

Methods

Study Design

This study used a randomized controlled trial (RCT) with a posttest-only design.

Participants and Randomization

Registered nurses working in the emergency department of Samsung Medical Center in Seoul, Republic of Korea, were recruited between July 17 and July 20, 2023. The required sample size of 42 participants (21 per group) was determined a priori using G*Power 3.1, based on an effect size (Cohen *d*) of 0.9, a significance level (α) of .05, and a statistical power ($1-\beta$) of .80.

Eligibility criteria included (1) voluntary participation with informed consent and (2) no prior experience operating the Level-1 device. Participants were randomly assigned to the experimental or control group through concealed allocation: each participant drew a slip of paper from an opaque box indicating group assignment in a 1:1 ratio. Both groups completed training and evaluation at the same location with 1:1 supervision.

Interventions

AR-Based Training (Experimental Group)

The experimental group received AR-based training using Microsoft HoloLens 2, a head-mounted mixed reality device that projects interactive holographic content onto the user's real-world field of view. The AR training module was developed using Microsoft Dynamics 365 Guides [14], an authoring application for creating step-by-step holographic work instructions, and was delivered on a single HoloLens 2 device. The training program guided participants through the Level-1 setup and operation process in a step-by-step sequence. Holographic overlays provided visual instructions, material checklists, finger-shaped pointing guides for button locations, and short procedural videos for complex steps (Figure 1). A physical anchor affixed to the device served as a spatial reference point for hologram positioning. This design allowed participants to interact directly with the actual equipment while receiving real-time digital guidance.

XSL•FO
RenderX

Figure 1. Trainee's view during augmented reality–based training for the Level-1 rapid infusion system. (A) The anchor affixed to the device retrieves prestored instructional content and serves as a spatial reference for hologram positioning. (B) Finger-shaped holograms guide users to button locations. (C) Required materials for each step are displayed in advance. (D) Learning objectives appear in a blue instructional window. (E) Short instructional videos are played as holograms for complex operations. (F) The trainer's hand movements are displayed for procedures that require visual demonstration. (G) The step navigation interface: a practice card that advances to the next step when the trainee gazes at the arrow (left) and the expanded step list for the Level-1 setup procedure (right).



Traditional Training (Control Group)

The control group completed self-directed learning using the Level-1 nursing practice guidelines, a standardized printed manual developed by emergency nurses with more than 10 years of clinical experience. The manual describes the complete procedure from device power-on through system setup and rapid infusion, with step-by-step tasks and model answers. Participants studied the manual independently and then performed the procedure.

XSL•FO
RenderX

Outcome Measures

Clinical Competency

Clinical competency was assessed across 3 domains. Time was measured as learning time (time to complete the educational content) and performance time, further segmented into device setup, fluid infusion, and blood transfusion times. Accuracy was evaluated using two instruments: (1) a manufacturer-aligned Level-1 checklist comprising 4 items scored out of 100 and (2) an expert-validated 22-step evaluation form derived from Level-1 nursing practice guidelines, with each step scored as pass or fail (Multimedia Appendix 1). Because the 2 accuracy instruments were intended to capture the same underlying construct, their convergent validity was examined using the Spearman rank correlation. Independence was measured by the number of assistance requests made during the performance evaluation.

Self-Efficacy

Self-efficacy was measured using a 6-item modified scale adapted from Compeau and Higgins [15] and Abdul Hamid et al [10], with 1 reverse-scored item on a 5-point Likert scale (total score range 6‐30; higher scores indicate greater self-efficacy). Expert validation yielded a content validity index of 1.0, and the Cronbach α was 0.80 in this study (Multimedia Appendix 2 [10,15,16]).

Educational Satisfaction

Educational satisfaction was assessed using 4 items on a 5-point Likert scale measuring confidence, perceived benefit, and willingness to recommend the training to colleagues (total score range 4‐20; higher scores indicate greater satisfaction) [16]. The Cronbach α was 0.87 in this study.

Statistical Analysis

Data were analyzed using SPSS (version 18.0; IBM Corp). Normality was assessed using the Shapiro-Wilk test together with skewness values. Most continuous outcomes were nonnormally distributed, and the timed and count variables were markedly right-skewed (skewness ranging from +0.50 to +6.09). Accordingly, between-group differences in clinical competency, self-efficacy, and educational satisfaction were compared using the Mann-Whitney *U* test, and the corresponding data are summarized as medians with IQRs. Baseline homogeneity was assessed using the independent-samples *t* test. Convergent validity between the 2 accuracy instruments was evaluated using the Spearman rank correlation. All tests were 2-tailed, and statistical significance was set at $P<.05$.

Ethical Considerations

This study was approved by the Samsung Medical Center Institutional Review Board (2023-06-084). All participants provided written informed consent prior to enrollment. Participation was voluntary, and participants could withdraw at any time without consequence. To ensure ethical fairness, participants in the control group were offered the opportunity to receive HoloLens-based training after the study was completed, if they wished. The clinical trial was registered at ClinicalTrials.gov (NCT06506851).

Results

Participant Characteristics

A total of 42 participants (n=21, 50% per group) were enrolled and completed the study; there were no dropouts (Figure 2). The mean age was 28.00 (SD 3.93) years in the experimental group and 26.48 (SD 2.04) years in the control group. Clinical experience averaged 3.82 (SD 3.98) years and 2.32 (SD 2.17) years, respectively, with emergency department experience of 1.68 (SD 1.23) years and 1.75 (SD 1.09) years, respectively. No significant differences were found between groups in age, clinical experience, emergency department experience, or sex distribution, confirming baseline homogeneity (Table 1). One control-group participant did not complete the posttraining self-report questionnaire; therefore, the self-efficacy and educational satisfaction analyses include 20 control participants (Table 2), whereas all clinical competency analyses include the full sample (N=42).

XSL•FO
RenderX

Figure 2. Flowchart of participant recruitment, randomization, and study procedures. AR: augmented reality.

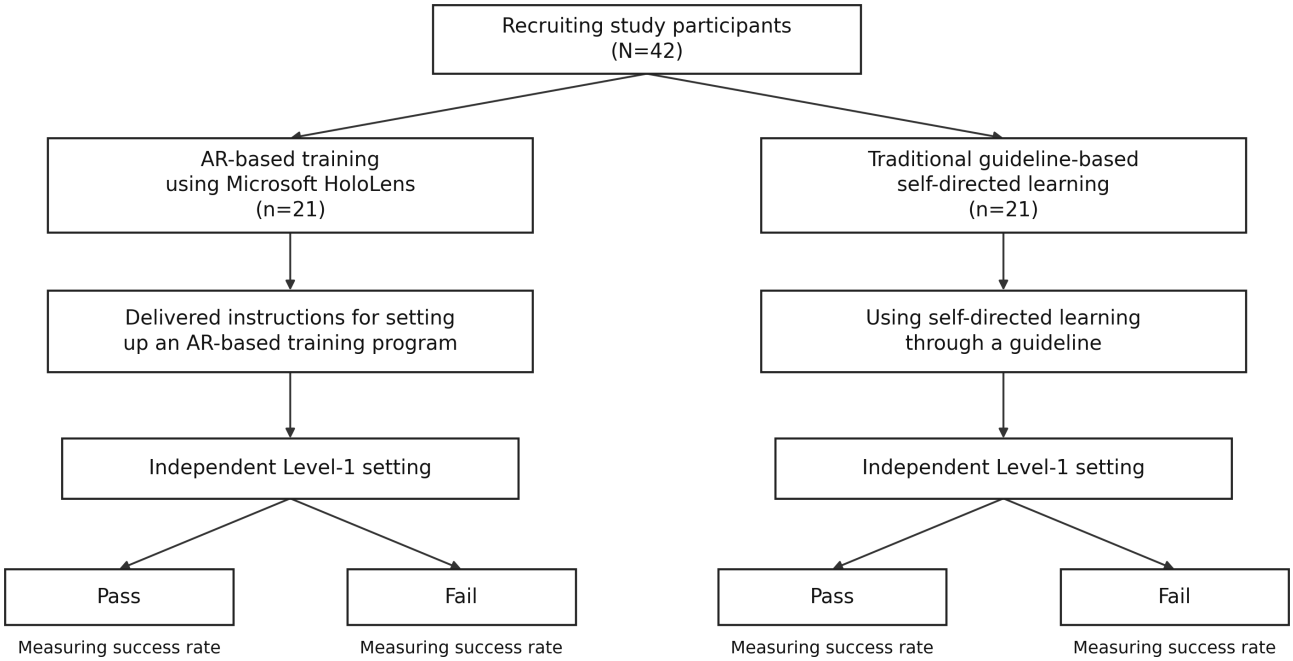


Table . Baseline characteristics of participants.[a]

Characteristics	Experimental (n=21)	Control (n=21)	*t* test (*df*)[a]	*P* value
Age (y), mean (SD)	28.00 (3.93)	26.48 (2.04)	−1.575 (40)	.12
Clinical experience (y), mean (SD)	3.82 (3.98)	2.32 (2.17)	−1.514 (40)	.14
Emergency department experience (y), mean (SD)	1.68 (1.23)	1.75 (1.09)	0.198 (40)	.84
Sex (male), n (%)	3 (14)	5 (24)	−0.773 (40)	.44

[a]All between-group comparisons were made using the independent-samples *t* test.

Table . Self-efficacy scores[a].

Items	Experimental (n=21), median (IQR)	Control (n=20)[b], median (IQR)	Mann-Whitney *U* test	*P* value
Confident advising others	4.00 (3.00 - 5.00)	2.00 (1.75 - 3.00)	345.0	<.001
Confident troubleshooting problems	3.00 (3.00 - 4.00)	2.00 (1.00 - 3.00)	329.0	.001
Confident learning advanced skills	4.00 (3.00 - 4.00)	3.00 (3.00 - 4.00)	304.5	.008
Confident understanding terminology	3.00 (3.00 - 4.00)	3.00 (2.75 - 3.00)	256.0	.21
Confident describing functions	4.00 (3.00 - 4.00)	3.00 (2.00 - 3.00)	309.0	.007
Pressured when observed (reverse scored)	3.00 (2.00 - 4.00)	3.00 (2.00 - 3.25)	220.5	.80
Total	20.00 (18.00 - 25.00)	16.50 (13.00 - 20.00)	322.5	.003

[a]Data are presented as median (IQR); between-group comparisons used the Mann-Whitney *U* test.

[b]One control participant did not complete the posttraining questionnaire.

XSL•FO
RenderX

Clinical Competency

Table 3 presents the clinical competency outcomes. Learning time was significantly longer in the experimental group (median 18.20, IQR 15.48 - 21.67 min vs median 8.98, IQR 5.85 - 11.68 min; Mann-Whitney *U*=377.0; *P*<.001), reflecting the more structured, step-by-step nature of the AR-guided program. However, this additional learning time translated into substantially faster performance: device setup time was markedly shorter in the experimental group (median 3.67, IQR 2.90 - 4.63 min vs median 9.85, IQR 8.03 - 11.35 min; Mann-Whitney *U*=3.5; *P*<.001), and total performance time was likewise shorter (median 5.83, IQR 4.80 - 6.68 min vs median 13.63, IQR 10.33 - 15.17 min; Mann-Whitney *U*=26.0; *P*<.001). Of note, the between-group difference in total performance time was not detected by a parametric test (independent-samples *t* test: *P*=.06) but was highly significant under the Mann-Whitney *U* test, underscoring the importance of using distribution-appropriate methods for these skewed data.

Table . Clinical competency outcomes[a].

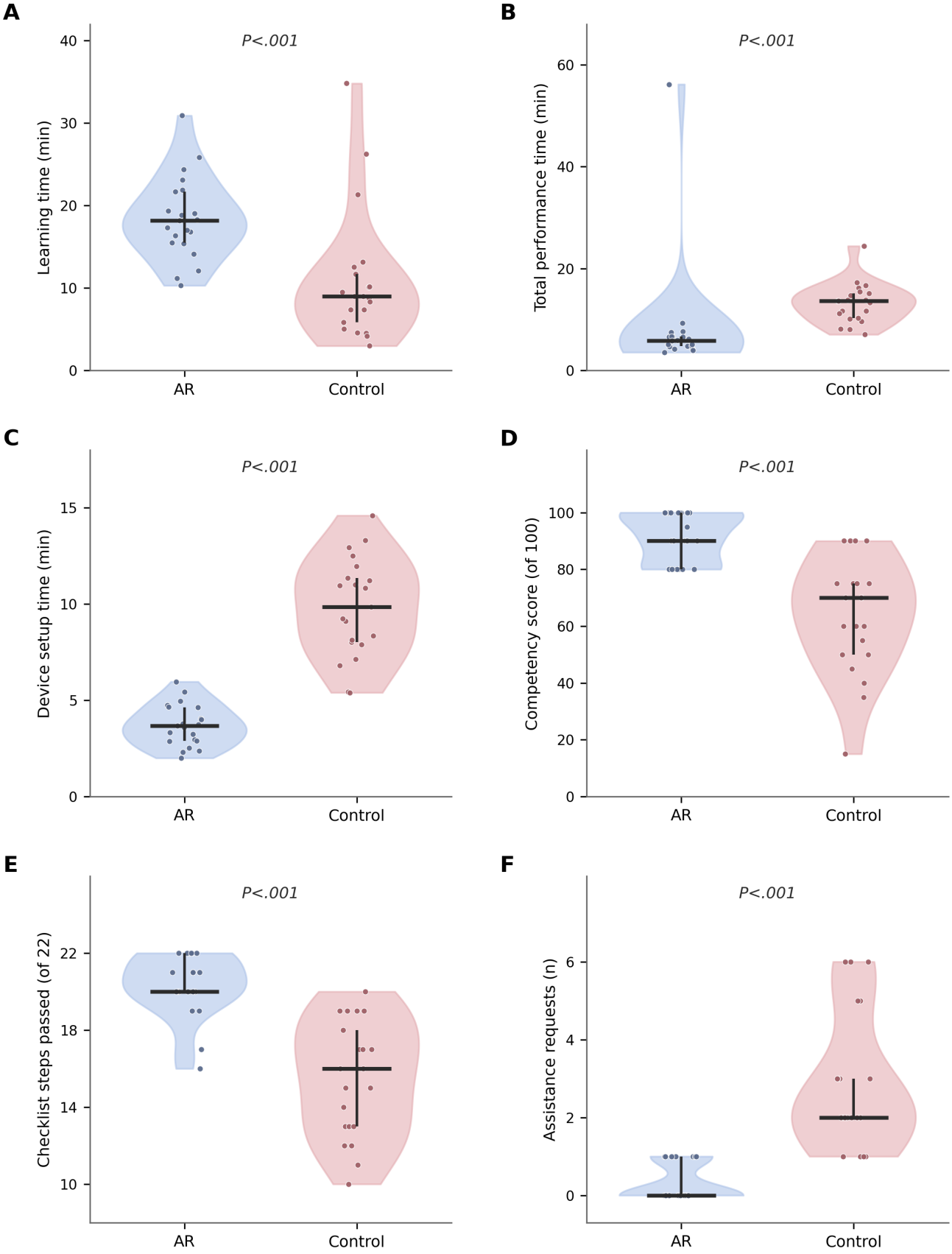
Variables	Experimental (n=21), median (IQR)	Control (n=21), median (IQR)	Mann-Whitney *U* test	*P* value
Learning time (min)	18.20 (15.48 - 21.67)	8.98 (5.85 - 11.68)	377.0	<.001
Total performance time (min)	5.83 (4.80 - 6.68)	13.63 (10.33 - 15.17)	26.0	<.001
Device setup time (min)	3.67 (2.90 - 4.63)	9.85 (8.03 - 11.35)	3.5	<.001
Fluid infusion time (min)	0.50 (0.37 - 0.62)	0.65 (0.33 - 0.92)	188.5	.43
Blood transfusion time (min)	1.50 (1.30 - 1.85)	2.05 (1.37 - 2.48)	147.0	.08
Device operation competency score	90.00 (80.00 - 100.00)	70.00 (50.00 - 75.00)	407.0	<.001
Checklist completion (22 items passed)	20.00 (20.00 - 22.00)	16.00 (13.00 - 18.00)	414.0	<.001
Device setup tasks (15 items)	14.00 (13.00 - 15.00)	10.00 (9.00 - 12.00)	410.5	<.001
Fluid infusion tasks (2 items)	2.00 (2.00 - 2.00)	2.00 (2.00 - 2.00)	252.5	.17
Blood transfusion tasks (5 items)	5.00 (5.00 - 5.00)	3.00 (2.00 - 5.00)	309.5	.02
Assistance requests, n	0.00 (0.00 - 1.00)	2.00 (2.00 - 3.00)	14.0	<.001

[a]Data are presented as median (IQR); between-group comparisons used the Mann-Whitney *U* test.

The experimental group also demonstrated superior accuracy, with a median device operation competency score of 90.00 (IQR 80.00 - 100.00) compared with a median of 70.00 (IQR 50.00 - 75.00) in the control group (Mann-Whitney *U*=407.0; *P*<.001). On the 22-step checklist, the experimental group passed a median of 20.00 (IQR 20.00 - 22.00) steps compared with a median of 16.00 (IQR 13.00 - 18.00) steps in the control group (Mann-Whitney *U*=414.0; *P*<.001; Multimedia Appendix 3). The 2 accuracy instruments were strongly correlated (Spearman ρ=0.86; *P*<.001), supporting their convergent validity. Notably, the experimental group required significantly fewer assistance requests (median 0.00, IQR 0.00 - 1.00 vs median 2.00, IQR 2.00 - 3.00; Mann-Whitney *U*=14.0; *P*<.001), indicating that AR training fostered greater independence during device operation. The distributions of the principal competency outcomes are shown in Figure 3.

XSL•FO
RenderX

Figure 3. Distribution of the principal clinical competency outcomes by group (N=42). (A) Learning time, (B) total performance time, (C) device setup time, (D) device operation competency score, (E) checklist steps passed of 22, and (F) number of assistance requests. Violin plots show the kernel density of each outcome; overlaid points represent individual participants, and horizontal bars denote medians with IQRs. *P* values are from the Mann-Whitney *U* test. AR: augmented reality.



XSL•FO
RenderX

Self-Efficacy

The experimental group reported significantly higher total self-efficacy scores (median 20.00, IQR 18.00 - 25.00) than the control group (median 16.50, IQR 13.00 - 20.00; Mann-Whitney *U*=322.5; *P*=.003). The largest between-group differences were observed in confidence in advising others about Level-1 operation (Mann-Whitney *U*=345.0; *P*<.001) and confidence in troubleshooting device problems (Mann-Whitney *U*=329.0; *P*=.001). The reverse-scored item measuring performance pressure when observed by others showed no significant difference between groups (*P*=.80; Table 2).

Educational Satisfaction

Educational satisfaction was significantly higher in the experimental group across all items (Table 4). The total score was higher in the experimental group (median 18.00, IQR 16.00 - 18.00) than in the control group (median 12.00, IQR 9.75 - 15.00; Mann-Whitney *U*=385.0; *P*<.001). Participants in the experimental group reported significantly greater confidence in performing rapid Level-1 setups independently (median 4.00, IQR 3.00 - 4.00 vs median 2.00, IQR 1.75 - 3.00; Mann-Whitney *U*=369.5; *P*<.001) and higher perceived benefit from the training (median 5.00, IQR 5.00 - 5.00 vs median 4.00, IQR 3.00 - 4.00; Mann-Whitney *U*=354.0; *P*<.001).

Table . Educational satisfaction scores[a].

Items	Experimental (n=21), median (IQR)	Control (n=20)[b], median (IQR)	Mann-Whitney *U* test	*P* value
Confidence in Level-1 setup	4.00 (3.00 - 4.00)	2.00 (1.75 - 3.00)	369.5	<.001
Confidence in rapid infusion	4.00 (3.00 - 4.00)	3.00 (1.75 - 3.25)	326.0	.002
Perceived benefit from training	5.00 (5.00 - 5.00)	4.00 (3.00 - 4.00)	354.0	<.001
Willingness to recommend training	5.00 (5.00 - 5.00)	3.50 (2.00 - 5.00)	343.0	<.001
Total	18.00 (16.00 - 18.00)	12.00 (9.75 - 15.00)	385.0	<.001

[a]Data are presented as median (IQR); between-group comparisons used the Mann-Whitney *U* test.

[b]One control participant did not complete the posttraining questionnaire.

Discussion

Principal Findings

This study provides the first randomized controlled evidence that AR-based training significantly improves emergency nurses' clinical competency, self-efficacy, and educational satisfaction in operating the Level-1 rapid infusion system compared with traditional guideline-based self-directed learning. The experimental group showed a 62% reduction in median device setup time (3.67, IQR 2.90-4.63 min vs 9.85, IQR 8.03-11.35 min), a 20-point higher median device operation competency score (90 vs 70 out of 100), and a reduction in the median number of assistance requests from 2 to 0. These results suggest that AR technology can effectively address a critical gap in nursing education: the challenge of training for infrequently used but life-critical medical devices.

Comparison With Prior Work

Our findings align with and extend the broader evidence base supporting AR in health professional education. A systematic mapping review of 156 studies (2010 - 2025) confirmed that AR effectively supports hands-on skill acquisition and clinical competency development across health sciences disciplines [8], and an umbrella review of virtual reality (VR) and AR in medical education similarly concluded that these technologies enhance specific professional competencies compared with traditional methods [17]. In the nursing domain specifically, a 2025 scoping review of 41 AR studies in nursing found that head-mounted displays were the most commonly used AR device type, with more than 90% of studies published since 2020 and more than half using RCT designs [18,19]. Recent randomized trials have demonstrated that AR-based training improves procedural performance in specific clinical tasks: Sun et al [5] showed that an AR app significantly improved nurses' advanced cardiac life support knowledge and crash cart skills, and Othman et al [13] reported that combining AR with gamification significantly enhanced critical care nursing students' self-efficacy and motivation. A 2025 RCT of VR and AR in emergency health care training further confirmed that immersive technologies improve emergency skill acquisition compared with traditional instruction [20]. Our study adds to this literature by demonstrating AR's efficacy for a specialized emergency medical device, the Level-1 rapid infuser, which poses unique training challenges owing to its infrequent use and high-stakes application context.

The effect sizes observed in this study were notably large. Expressed as Cohen *d*, the between-group differences were approximately 3.1 for device setup time, 2.0 for both the 22-step checklist and the number of assistance requests, 1.8 for the competency score, and 1.2 for learning time, substantially larger than the moderate-to-large effects typically reported in AR or immersive nursing education trials, which generally fall in the range of Cohen *d*≈0.5 to 1.0 (eg, the improvements in knowledge, skills, self-efficacy, and motivation reported by Sun et al [5] and Othman et al [13]). Several design factors most plausibly account for this magnitude. First, the comparator was self-directed learning from a printed manual without any hands-on guidance, a deliberately conservative, real-world control that maximizes the contrast with real-time, in situ AR guidance. Second, all participants were novices with no prior

XSL•FO
RenderX

Level-1 experience, leaving substantial room for improvement and limiting ceiling effects in the AR group. Third, the AR program delivered step-by-step holographic instruction anchored to the actual device (finger-pointing cues, material checklists, and embedded video), tightly coupling instruction to the physical task and reducing cognitive load during a procedurally complex, low-frequency procedure. Fourth, the single-device focus permitted concentrated, uninterrupted practice. These instructional design features, rather than the technology alone, are the most likely drivers of the large effects, and the magnitude should therefore be interpreted in the context of this conservative comparator. Cohen *d* is reported here for comparability with the exemplar studies; the corresponding nonparametric effect sizes (rank-biserial correlation) were of similar magnitude (r=0.71 to 0.98).

The finding that AR training required significantly more learning time yet produced faster performance is consistent with prior work on immersive training modalities [21]. This pattern suggests that AR's step-by-step, interactive format promotes deeper encoding of procedural knowledge during the learning phase, which subsequently translates into more efficient and accurate performance. A similar trade-off was observed in a controlled trial of AR-assisted intubation training, in which the AR group took longer to complete the training but demonstrated greater adherence to evidence-based procedures [22]. A systematic review and meta-analysis of immersive technology–based nursing education using the GRADE (Grading of Recommendations, Assessment, Development, and Evaluation) approach further supports that immersive technologies improve practical skill outcomes, although with varying effect sizes depending on the specific clinical domain [3].

The significant improvement in self-efficacy, particularly in confidence-related domains such as advising colleagues and troubleshooting equipment, is consistent with findings from studies showing that immersive learning environments foster greater confidence through realistic practice opportunities [21,23]. Notably, there was no significant between-group difference in the reverse-scored item measuring performance anxiety, suggesting that AR training enhanced positive confidence without increasing stress. A recent review of the HoloLens platform for health care simulation confirmed that mixed reality training supports competency development across diverse medical and procedural domains, with learners consistently reporting increased confidence [24]. The substantial improvement in educational satisfaction echoes findings from our prior mixed methods study evaluating AR adoption for nurses in the intensive care unit [4], reinforcing the acceptability and user experience advantages of AR-based educational approaches.

Implications for Nursing Education and Management

These findings carry practical implications for nursing management. The Level-1 rapid infuser represents a category of medical devices that are critical during emergencies but used too infrequently for conventional training programs to maintain competency effectively. AR-based training offers a scalable solution: once developed, the digital content can be deployed across multiple sites with consistent quality, reducing dependence on expert instructors and physical training resources [11,25,26]. The significant reduction in assistance requests observed in the experimental group suggests that AR training may improve team efficiency during actual resuscitation events by reducing the need for on-the-job guidance from senior staff. For nursing managers seeking to implement digital transformation initiatives [13,27,28], AR-based device training represents a concrete, evidence-supported application with measurable outcomes.

Limitations

This study has several limitations that should be considered when interpreting the results. First, outcomes were assessed immediately after the intervention, and the long-term retention of skills and knowledge was not evaluated. Future studies should incorporate follow-up assessments at defined intervals (eg, 1, 3, and 6 months) to determine the durability of training effects and the need for refresher training. Second, this was a single-center study conducted within one emergency department, which limits the generalizability of the findings. Multicenter trials involving nurses from diverse clinical settings, including intensive care units and surgical departments, would provide a broader understanding of AR training's applicability. Third, the study assessed individual performance metrics only; team-based outcomes such as communication, coordination, and collective decision-making during simulated resuscitation scenarios were not evaluated. Fourth, blinding was not feasible owing to the nature of the intervention, introducing the possibility of performance bias. Finally, cost-effectiveness was not assessed; future research should analyze the financial feasibility of AR training programs, including equipment acquisition, content development, and maintenance costs relative to improved clinical outcomes.

Conclusions

This RCT demonstrates that AR-based training using Microsoft HoloLens 2 significantly enhances emergency nurses' clinical competency, self-efficacy, and educational satisfaction in operating the Level-1 rapid infusion system. Despite requiring additional initial learning time, AR training produced faster, more accurate, and more independent device operation compared with traditional self-directed learning. These findings position AR technology as a viable and effective educational tool for medical devices that are infrequently used but critically important in emergency care. Future research should evaluate long-term skill retention, cost-effectiveness, and the applicability of AR-based training across a broader range of medical devices and clinical settings.

XSL•FO
RenderX

Acknowledgments

The authors thank the Nursing Staff Development Team at Samsung Medical Center for their encouragement and support, the Office of Education and Human Resources Development for facilitating access to HoloLens equipment, and the Strategy and Business Development Team for their administrative support and assistance throughout the project. The authors also thank KM Min for invaluable assistance throughout the study.

Funding

This study was supported by the research fund of the Department of Nursing, Samsung Medical Center, in 2023 (SMC-NSD-2023-02). This research was also supported by a grant from the Korea Health Technology R&D Project through the Korea Health Industry Development Institute, funded by the Ministry of Health and Welfare, Republic of Korea (grant RS-2023-KH135335).

Authors' Contributions

WJ and WC conceptualized the study and developed the training program. WJ, WC, YK, YM, SGN, and SS collected the data. SY performed the data analysis and drafted the initial manuscript. MHS supervised the study and provided critical revisions. All authors reviewed and approved the final manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
The 22-step clinical competency evaluation form for Level-1 rapid infusion system operation.
[DOCX File, 18 KB - mededu_v12i1e98073_app1.docx]

Multimedia Appendix 2
Self-efficacy and Educational Satisfaction Scale items with scoring instructions.
[DOCX File, 18 KB - mededu_v12i1e98073_app2.docx]

Multimedia Appendix 3
Item-level pass rate analysis of the 22-step evaluation by procedural phase and group.
[DOCX File, 17 KB - mededu_v12i1e98073_app3.docx]

Checklist 1
CONSORT (Consolidated Standards of Reporting Trials) checklist.
[PDF File, 66 KB - mededu_v12i1e98073_app4.pdf]

References

1. Kanschik D, Bruno RR, Wolff G, Kelm M, Jung C. Virtual and augmented reality in intensive care medicine: a systematic review. Ann Intensive Care 2023 Sep 11;13(1):81. [doi: 10.1186/s13613-023-01176-z] [Medline: 37695464]
2. Ortega-Lapiedra R, Barrado-Narvión MJ, Bernués-Oliván J. Acquisition of competencies of nurses: improving the performance of the healthcare system. Int J Environ Res Public Health 2023 Mar 3;20(5):4510. [doi: 10.3390/ijerph20054510] [Medline: 36901519]
3. Park S, Shin HJ, Kwak H, Lee HJ. Effects of immersive technology-based education for undergraduate nursing students: systematic review and meta-analysis using the Grading of Recommendations, Assessment, Development, and Evaluation (GRADE) approach. J Med Internet Res 2024 Jul 24;26:e57566. [doi: 10.2196/57566] [Medline: 38978483]
4. Yoo S, Heo S, Song S, et al. Adoption of augmented reality in educational programs for nurses in intensive care units of tertiary academic hospitals: mixed methods study. JMIR Serious Games 2024 May 23;12:e54188. [doi: 10.2196/54188] [Medline: 38780998]
5. Sun WN, Hsieh MC, Wang WF. Nurses' knowledge and skills after use of an augmented reality app for advanced cardiac life support training: randomized controlled trial. J Med Internet Res 2024 Dec 5;26:e57327. [doi: 10.2196/57327] [Medline: 39636667]
6. Niles D, Sutton RM, Donoghue A, et al. "Rolling Refreshers": a novel approach to maintain CPR psychomotor skill competence. Resuscitation 2009 Aug;80(8):909-912. [doi: 10.1016/j.resuscitation.2009.04.021] [Medline: 19467759]
7. Yang CW, Yen ZS, McGowan JE, et al. A systematic review of retention of adult advanced life support knowledge and skills in healthcare providers. Resuscitation 2012 Sep;83(9):1055-1060. [doi: 10.1016/j.resuscitation.2012.02.027] [Medline: 22391016]

XSL•FO
RenderX

8. Lampropoulos G, Fernández-Arias P, Del Bosque A, Vergara D. Augmented reality in health education: transforming nursing, healthcare, and medical education and training. Nurs Rep 2025 Aug 8;15(8):289. [doi: 10.3390/nursrep15080289] [Medline: 40863676]
9. Cheng A, Fijacko N, Lockey A, et al. Use of augmented and virtual reality in resuscitation training: a systematic review. Resusc Plus 2024 Jun;18:100643. [doi: 10.1016/j.resplu.2024.100643] [Medline: 38681058]
10. Abdul Hamid R, Ismail IM, Wan Yahaya WA. Augmented reality for skill training from TVET instructors' perspective. Interact Learn Environ 2024;32(4):1291-1299. [doi: 10.1080/10494820.2022.2118788]
11. Huter K, Krick T, Domhoff D, Seibert K, Wolf-Ostermann K, Rothgang H. Effectiveness of digital technologies to support nursing care: results of a scoping review. J Multidiscip Healthc 2020;13:1905-1926. [doi: 10.2147/JMDH.S286193] [Medline: 33328736]
12. Mlambo M, Silén C, McGrath C. Lifelong learning and nurses' continuing professional development, a metasynthesis of the literature. BMC Nurs 2021 Apr 14;20(1):62. [doi: 10.1186/s12912-021-00579-2] [Medline: 33853599]
13. Othman SY, Ghallab E, Eltaybani S, Mohamed AM. Effect of using gamification and augmented reality in mechanical ventilation unit of critical care nursing on nurse students' knowledge, motivation, and self-efficacy: a randomized controlled trial. Nurse Educ Today 2024 Nov;142:106329. [doi: 10.1016/j.nedt.2024.106329] [Medline: 39116661]
14. Microsoft Dynamics 365 Guides documentation. Microsoft. URL: https://learn.microsoft.com/en-us/dynamics365/mixed-reality/guides/ [accessed 2026-06-24]
15. Compeau DR, Higgins CA. Computer self-efficacy: development of a measure and initial test. MIS Q 1995 Jun 1;19(2):189-211. [doi: 10.2307/249688]
16. Heo JY, Lim CH, Kim KJ. Scales for measuring mobile service quality: a literature review and identification of key dimensions. Int J Serv Oper Manag 2017;27(4):524-528. [doi: 10.1504/IJSOM.2017.085214]
17. Tene T, Vique López DF, Valverde Aguirre PE, Orna Puente LM, Vacacela Gomez C. Virtual reality and augmented reality in medical education: an umbrella review. Front Digit Health 2024;6:1365345. [doi: 10.3389/fdgth.2024.1365345] [Medline: 38550715]
18. Wei W, Gao J, Luo X, Gong X, Dong L. A scoping review of the applications of augmented reality in nursing. BMC Nurs 2025 Aug 23;24(1):1105. [doi: 10.1186/s12912-025-03750-1] [Medline: 40849629]
19. Mendez KJ, Piasecki RJ, Hudson K, et al. Virtual and augmented reality: implications for the future of nursing education. Nurse Educ Today 2020 Oct;93:104531. [doi: 10.1016/j.nedt.2020.104531] [Medline: 32711132]
20. Castillo-Rodríguez JM, Gómez-Urquiza JL, García-Oliva S, Suleiman-Martos N. Effectiveness of virtual and augmented reality for emergency healthcare training: a randomized controlled trial. Healthcare (Basel) 2025 Apr 30;13(9):1034. [doi: 10.3390/healthcare13091034] [Medline: 40361812]
21. Makransky G, Borre‐Gude S, Mayer RE. Motivational and cognitive benefits of training in immersive virtual reality based on multiple assessments. J Comput Assist Learn 2019 Dec;35(6):691-707. [doi: 10.1111/jcal.12375]
22. Iqbal AI, Aamir A, Hammad A, et al. Immersive technologies in healthcare: an in-depth exploration of virtual reality and augmented reality in enhancing patient care, medical education, and training paradigms. J Prim Care Community Health 2024;15:21501319241293311. [doi: 10.1177/21501319241293311] [Medline: 39439304]
23. Dhar P, Rocks T, Samarasinghe RM, Stephenson G, Smith C. Augmented reality in medical education: students' experiences and learning outcomes. Med Educ Online 2021 Dec;26(1):1953953. [doi: 10.1080/10872981.2021.1953953] [Medline: 34259122]
24. Tătaru OS, Ferro M, Marchioni M, et al. HoloLens® platform for healthcare professionals simulation training, teaching, and its urological applications: an up-to-date review. Ther Adv Urol 2024;16:17562872241297554. [doi: 10.1177/17562872241297554] [Medline: 39654822]
25. Cachata D, Costa M, Magalhães T, Gaspar F. The integration of information technology in the management and organization of nursing care in a hospital environment: a scoping review. Int J Environ Res Public Health 2024 Jul 24;21(8):968. [doi: 10.3390/ijerph21080968] [Medline: 39200579]
26. Gause G, Mokgaola IO, Rakhudu MA. Technology usage for teaching and learning in nursing education: an integrative review. Curationis 2022 Jun 15;45(1):e1-e9. [doi: 10.4102/curationis.v45i1.2261] [Medline: 35792609]
27. Navarro Martínez O, Leyva-Moral JM. Digital transformation led by nurses and nursing managers' priorities: a qualitative study. J Nurs Manag 2024;2024:8873127. [doi: 10.1155/2024/8873127] [Medline: 40224899]
28. Takemura Y, Inoue M, Ichikawa N, et al. Key strategies for managing nursing care under the COVID-19 pandemic: a multiple-case study of nursing directors. J Nurs Manag 2022 Nov;30(8):4042-4053. [doi: 10.1111/jonm.13844] [Medline: 36194470]

Abbreviations

AR: augmented reality
GRADE: Grading of Recommendations, Assessment, Development, and Evaluation
RCT: randomized controlled trial
VR: virtual reality

XSL•FO
RenderX

Edited by T Mühling; submitted 13.Apr.2026; peer-reviewed by MA Bathish, Q Zhou; revised version received 25.Jun.2026; accepted 29.Jun.2026; published 13.Aug.2026.

Please cite as:
Jang W, Chai W, Kim Y, Moon Y, Yoo S, Na SG, Shin S, Son MH
Augmented Reality–Based Training for a Rapid Blood Transfusion Device Among Emergency Nurses: Randomized Controlled Trial
JMIR Med Educ 2026;12:e98073
URL: https://mededu.jmir.org/2026/1/e98073
doi:10.2196/98073

© Wooyoung Jang, Wooree Chai, Yeji Kim, Youjung Moon, Suyoung Yoo, Sun Gyoung Na, Soyeon Shin, Meong Hi Son. Originally published in JMIR Medical Education (https://mededu.jmir.org), 13.Aug.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Education, is properly cited. The complete bibliographic information, a link to the original publication on https://mededu.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Development and Educational Evaluation of a Radiation Visualization System in a Virtual Reality Training Module for Radiopharmaceutical Administration: Mixed Methods Study of Students' Learning Experiences

Akihiro Kakimoto[1,2*], PhD; Daisuke Fujise[3*], PhD; Shota Watanabe[1], PhD; Yutaka Otaka[1], PhD; Sakura Minemoto[1]; Ayaka Yasumatsu[1]; Shin Hasegawa[1,3], PhD

[1]Department of Radiological Sciences, Faculty of Medical Science Technology, Morinomiya University of Medical Sciences, 1-26-16 Nankokita, Suminoe-ku, Osaka-shi, Osaka, Osaka, Japan

[2]Department of Biofunctional Imaging, Preeminent Medical Photonics Education and Research Center, Hamamatsu University School of Medicine, Hamamatsu-shi, Shizuoka, Japan

[3]IT Planning Section, Department of Information Technology and Management, National Institutes for Quantum Science and Technology, Chiba-shi, Chiba, Japan

[*]these authors contributed equally

Corresponding Author:
Akihiro Kakimoto, PhD
Department of Radiological Sciences, Faculty of Medical Science Technology, Morinomiya University of Medical Sciences, 1-26-16 Nankokita, Suminoe-ku, Osaka-shi, Osaka, Osaka, Japan

Abstract

Background: Our research group previously developed a virtual reality (VR) training module for radiopharmaceutical administration, focusing on procedural skills and patient interactions. The module's educational effectiveness was validated with objective measures, such as electroencephalography and mood assessments.

Objective: The authors aimed to extend the module by integrating a radiation visualization system that renders otherwise invisible radiation fields within an immersive VR environment, and to compare students' learning experiences associated with the module with those associated with conventional hands-on training.

Methods: Twenty students enrolled in a radiological technician training program participated in this study. Students' learning experiences were evaluated using Likert-scale–based questionnaires, together with qualitative analysis of open-ended responses.

Results: VR-based training was associated with significantly higher ratings for students' conceptual understanding of radiation behavior, radiation safety awareness, learning satisfaction, and self-assessed competency than conventional training. Conversely, perceived realism and immersion were rated significantly higher following conventional training. No significant differences were observed for operational understanding, technical skills, or system usability.

Conclusions: These findings suggest that the developed VR module offers a safe and repeatable training platform that may support positively perceived learning experiences in radiopharmaceutical administration and radiation safety education.

(JMIR Med Educ 2026;12:e104397) doi:10.2196/104397

KEYWORDS

education; radiological technologist students; virtual reality; radiological protection; radiopharmaceutical administration

Introduction

Virtual reality (VR)–based training is increasingly recognized as a valuable modality in health professions education, offering immersive, safe, and controlled environments for learning complex clinical skills that are difficult to replicate in traditional clinical training settings [1]. Systematic reviews have indicated that VR yields greater engagement, competence, and confidence across diverse medical and radiology training contexts than traditional teaching methods [1,2]. In radiology education, VR has been shown to improve proficiency in patient positioning, equipment handling, and radiographic techniques, with students demonstrating improved perceived competence and satisfaction. Moreover, immersive technologies such as VR and mixed reality (MR) have demonstrated significant improvements in learning outcomes, performance, and engagement in medical and radiation physics education [3].

Beyond general clinical skills training, VR applications are increasingly being explored in radiation safety education. 3D VR simulations have demonstrated an ability to enhance

XSL•FO
RenderX

students' understanding of radiation protection principles and promote active engagement with safety practices, with learners reporting positive perceptions of VR as an educational tool [4]. Comparative studies have further suggested that VR-based training can outperform conventional lecture-based radiation safety training, leading to improved knowledge retention and reduced simulated occupational exposure among health care professionals [5].

Our research group previously developed a VR-based training module for radiopharmaceutical administration focusing on procedural performance and simulated patient interactions. This module was evaluated using physiological and psychological measures, including electroencephalography and mood-state assessments, and yielded promising educational outcomes [6]. Although the previous system included basic visualization features, the representation of radiation fields had limited fidelity and was not fully grounded in clinical measurement data. Therefore, in this study, we developed a specialized radiation visualization system designed to represent radiation distribution and exposure risk with enhanced realism and clinical accuracy. Spatial dose measurements acquired from clinical settings were used to construct a 3D map of radiation fields within the VR environment, enabling the reproduction of radiation distribution patterns that closely approximate real clinical conditions. By integrating clinically measured dose data into an immersive VR space, we aimed to examine students' learning experiences associated with data-driven, high-fidelity visualization of radiation, including conceptual understanding, radiation-safety awareness, perceived realism and immersion, learning satisfaction, and self-perceived competence, compared with conventional hands-on training.

The primary objective of this study was to systematically evaluate students' learning experiences associated with perceptual radiation visualization in a clinically grounded VR-based training environment. By integrating 3D radiation field mapping with spatial dose measurements acquired from real clinical settings, a high-fidelity, data-driven simulation platform was developed. Using subjective evaluations and qualitative feedback, the study systematically examined students' perceptions of their understanding of radiation, radiation safety awareness, and learning experiences associated with the VR environment. Through this approach, we aimed to provide insights into the potential role of perceptual visualization in radiation safety education and the application of experiential learning in radiopharmaceutical administration training.

Methods

Study Design

This study adopted a prospective, within-participants, comparative, quasi-experimental, mixed methods design to compare students' learning experiences associated with a VR-based radiopharmaceutical administration training module incorporating a radiation visualization system (VR-based training) and conventional radiopharmaceutical administration training (conventional training), using both quantitative and qualitative methods.

Ethical Considerations

This study was conducted in accordance with the Declaration of Helsinki and was approved by the Institutional Ethics Committee of Morinomiya University of Medical Sciences (approval number: 2025 - 069). All eligible students were provided with a detailed explanation of the study's purpose, procedures, and data management processes, and electronic informed consent was obtained via Microsoft Forms prior to enrollment. Participation was entirely voluntary, and students were explicitly informed that participation or nonparticipation would not influence their academic evaluations. All data were anonymized and securely stored to prevent individual identification. Data collection was initiated only after the release of final grades for the relevant academic terms to eliminate any potential influence on student evaluations. Study information sheets, consent forms, and withdrawal forms were distributed in paper format, and consent records were managed using Microsoft Forms.

The study period spanned from April 1, 2025, to January 31, 2026, and volunteer-based measurements were performed between December 1, 2025, and December 23, 2025.

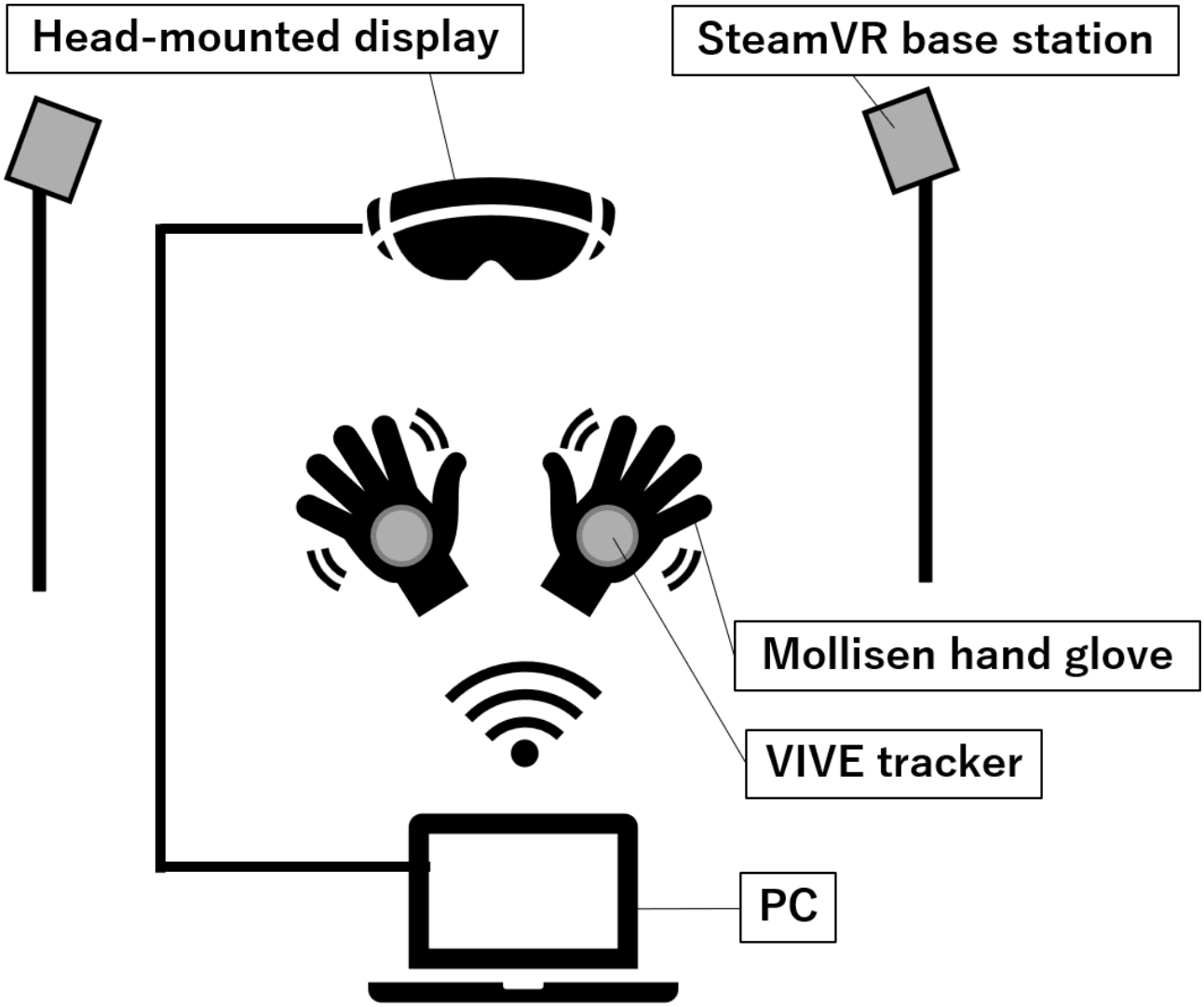
VR System Development and Visualization Design

The VR-based training system was built on a PC running Windows 10 Pro (64-bit) equipped with an NVIDIA GeForce RTX 2080 graphics card. The hardware configuration consisted of a VIVE Pro Eye head-mounted display (HMD), VIVE Tracker 3.0 devices, and the SteamVR Base Station 2.0 (High Tech Computer Corp) for motion tracking, and Vib+[L] VR gloves (Mollisen) for precise hand interaction. The software environment was constructed using Unity (version 2023.3.7f1; Unity Technologies), supported by the SteamVR Plugin (version 2.7.4; software development kit [SDK] 1.14.15), SteamVR (version 2.2.3), and the Mollisen SDK.

The system configuration is illustrated in Figure 1. The HMD was connected to the PC via a wired connection, while the VR gloves communicated wirelessly. Each glove was equipped with a dedicated tracker for spatial tracking. Using the SteamVR Base Station system, the position and orientation of the HMD and gloves were continuously monitored and transmitted to the VR application, enabling real-time synchronization of user movements within the virtual environment. The VR-based training software was executed on the Unity platform. By wearing an HMD and VR gloves, users could interact with objects in the virtual space. The finger motion data captured from the gloves were integrated using the Mollisen SDK, enabling accurate reproduction of the user's finger movements within the VR-based training environment.

XSL•FO
RenderX

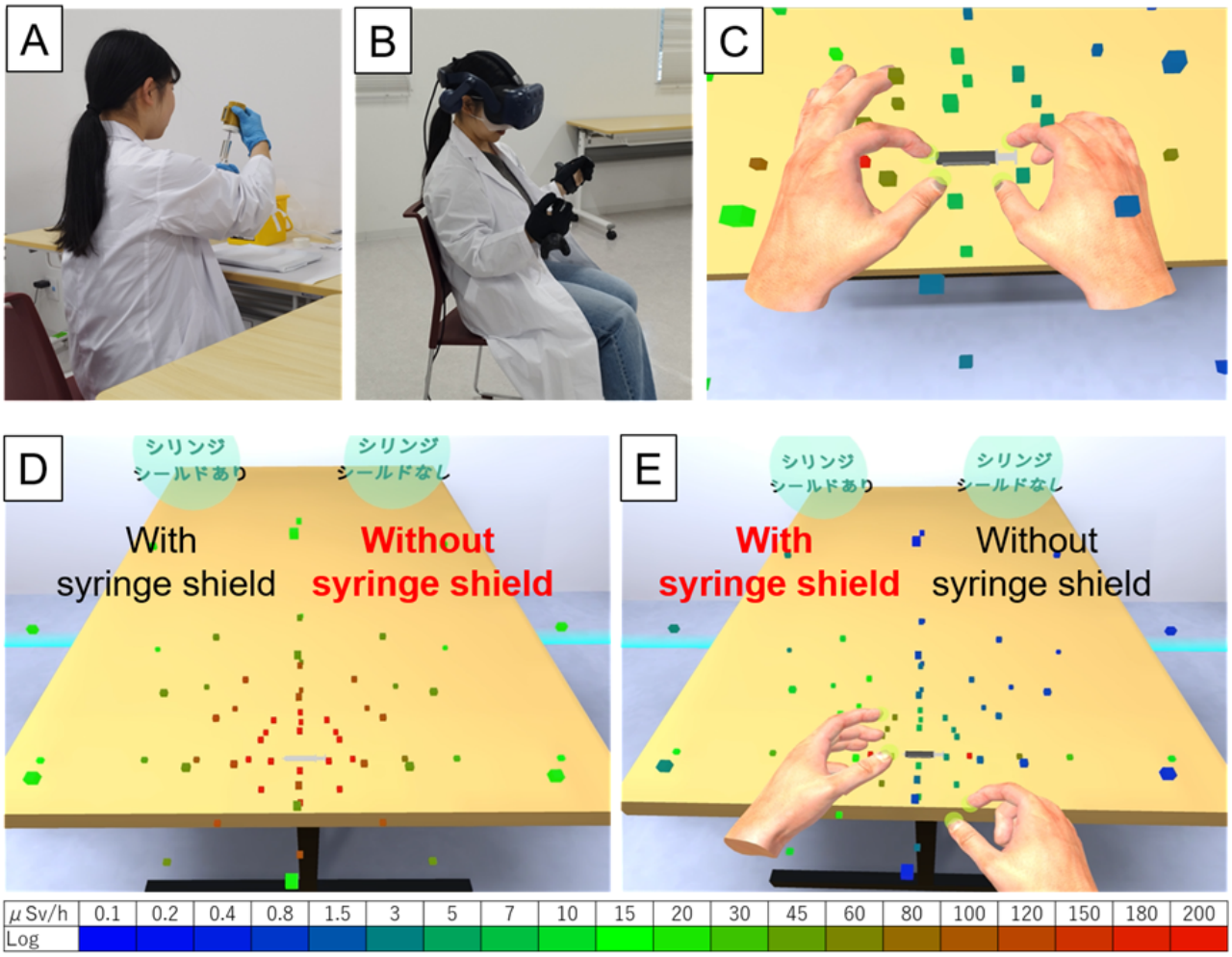
Figure 1. Virtual reality (VR) system configuration diagram.



The developed VR module is presented in Figure 2. Figure 2A shows conventional training, whereas Figure 2B depicts VR-based training using an HMD and dedicated VR gloves. Figure 2C shows learners' viewpoints within the VR-based training environment. To support the visualization of radiation exposure during training, radiation field visualization data were incorporated into the VR environment. Figure 2D and E present the visualized dose-rate distributions without and with a syringe shield, respectively. These visualizations can be activated or deactivated via interface buttons, allowing learners to directly compare shielding effects during task performance.

XSL•FO
RenderX

Figure 2. Radiopharmaceutical administration training: conventional and virtual reality (VR)–based methods. (A) Conventional training conducted in a real clinical environment. (B) VR-based training performed using a head-mounted display and VR gloves. (C) First-person perspective of the learner's hands within the VR-based training environment. (D) Radiation field visualization in the absence of a syringe shield. (E) Radiation field visualization with a syringe shield in place.



Radiation field visualization was constructed from empirical measurements obtained using Technetium-99m (Tc), the radionuclide most commonly used in nuclear medicine, to more closely represent radiation distributions encountered in clinical practice. Radioactivity was measured using a curiemeter (IGC-7F; ALOKA Co, Ltd), and spatial dose rates were measured using a NaI(Tl) scintillation survey meter (NHC6; Fuji Electric Co, Ltd), which was calibrated annually by an accredited external service provider. Dose rates were measured at predefined locations surrounding the vial and syringe (10, 20, 30, and 50 cm at predetermined horizontal and vertical angles), and the mean values from 3 repeated measurements were used to generate the spatial dose distribution. Although the measurements were performed using 68.6 MBq of Tc, which is lower than activities commonly administered in clinical practice, this activity was selected to minimize radiation exposure during the prolonged measurement procedure and to remain within the operating range of the survey meter. Because the objective was to reproduce the spatial distribution of the radiation field rather than absolute dose rates, the measured dose distribution was normalized and implemented as transparent fixed objects in the VR environment. As shown in Figure 2, the measured data were visualized as a discrete spatial distribution in the VR environment. Dose rate values were presented using a conventional blue-green-red rainbow color scale and were rendered on a logarithmic scale rather than a linear scale to enhance perceptual discrimination of radiation intensity gradients (Multimedia Appendix 1). As the air dose rate from a radiation source decreases with distance according to the inverse-square law [7], variations in the low-dose region become difficult to distinguish when visualized on a linear scale. Therefore, in this study, dose values were color-mapped using a logarithmic scale to maintain perceptual contrast across a wide dynamic range of radiation intensities [8,9].

The color bar indicates the corresponding dose-rate values used in the radiation visualization; detailed color maps are provided in Multimedia Appendix 1.

Participants and Procedure

A total of 20 third-year undergraduate radiological technology students volunteered to participate in this study (n=10 male and n=10 female participants; mean age 20.9, SD 0.45, range: 20-22 years). All participants had completed the required technical skills and patient communication training before participation in this study. In addition, they had prior experience with both conventional cold-run training and supervised hot-run training

XSL•FO
RenderX

involving unsealed radioactive sources during their nuclear medicine practical training approximately 5 months before the study. Therefore, they were not novices with respect to radiopharmaceutical administration procedures.

The study procedure was as follows: first, participants completed conventional training using vials, syringes, and physiological saline as surrogates for unsealed radioactive sources (cold run), which involved no actual radiation exposure. Next, participants wore an HMD and VR gloves to complete the same tasks in the VR-based training environment, which closely simulated real-world clinical conditions and incorporated radiation-field visualization. Each session consisted of brief tasks lasting approximately 5 minutes. Finally, participants completed a web-based questionnaire to report their learning experiences and perceptions.

Quantitative Evaluation

A quantitative evaluation was conducted using a self-administered questionnaire based on a 5-point Likert scale (1="strongly disagree," 2="disagree," 3="neutral," 4="agree," and 5="strongly agree") [10]. The questionnaire consisted of 7 primary domains and 35 secondary items, designed to reflect hierarchical learning progression from basic understanding to independent performance and the ability to explain or teach others (Multimedia Appendix 2). The questionnaire was used to assess participants' self-reported perceptions and learning experiences associated with the training modules.

A questionnaire previously developed and validated by our research group for evaluating learning experiences in nuclear medicine practical training was adapted for the present study [11]. The original instrument underwent expert review by 3 faculty members specializing in nuclear medicine education and radiological technology, followed by pilot testing involving 10 students and 6 faculty members who were not involved in the questionnaire's development. The validation study demonstrated excellent internal consistency, as assessed using Cronbach α of 0.91 [11,12], with domain-specific Cronbach α values ranging from 0.68 to 0.90. Preliminary construct validity was supported through exploratory factor analysis using the Kaiser-Meyer-Olkin (KMO) measure [13] and Bartlett test of sphericity [14], which demonstrated acceptable sampling adequacy (KMO=0.85 and 0.81 for cold- and hot-run datasets, respectively; Bartlett test of sphericity, $P<.001$ for both datasets [11]). The extracted factor structure was generally consistent with the intended educational domains, providing preliminary support for the construct validity of the instrument [11]. For this study, the questionnaire structure was retained and modified to include items related to radiation visualization, VR usability, and learning experiences associated with the VR-based training environment.

The 7 primary domains were as follows: (A) radiation understanding and visualization, (B) operational understanding and technical skills (vial and shielding manipulations), (C) perceived realism and immersion, (D) medical safety awareness and radiation protection, (E) system evaluation (usability and performance), (F) learning satisfaction, and (G) self-evaluation.

The radiation understanding and visualization domain (A) assessed intuitive comprehension of spatial radiation distribution, understanding of shielding effects (with and without syringe shields), perception of radiation intensity and attenuation through color mapping, improvement in safety awareness, and the ability to communicate spatial dose distribution concepts to peers.

The operational understanding and technical skills domain (B) evaluated the comprehension of equipment structure, appropriate use of syringe shields, perceived realism of operational sensations (eg, weight and spatial positioning), conscious integration of radiation protection principles during task execution, and confidence in explaining these procedures to others.

The perceived realism and immersion domain (C) investigated perceived realism, correspondence with real clinical working environments, naturalness of hand and vial manipulation, the ability to maintain concentration, and the capacity to articulate task-related tension. Item C2, which assessed a sense of presence similar to a real working environment, was interpreted as an indicator of perceived realism.

The medical safety awareness and radiation protection domain (D) examined understanding of safety considerations when handling unsealed radioactive sources; awareness of dose-reduction strategies; recognition of the principles of distance, shielding, and time; the perceived importance of radiation protection; and confidence in instructing others.

The system evaluation domain (E) evaluated interface clarity, operational smoothness, absence of operational discomfort, perceived value for repeated practice, and perceived educational significance for integration into future training curricula.

The learning satisfaction domain (F) assessed perceived acquisition of new knowledge, a deeper understanding of radiation and safety management, satisfaction with the training experience, perceived applicability to future professional practice, and willingness to continue using similar educational tools.

The self-evaluation domain (G) assessed confidence in newly acquired knowledge and skills, perceived ability to independently perform comparable procedures, confidence in executing procedures safely in future clinical practice, heightened awareness of radiation handling, and willingness to recommend this learning approach to peers.

Primary outcome scores were estimated as the mean of the 5 corresponding secondary items within each domain. Secondary outcome items were analyzed separately to identify specific aspects of participants' learning experiences. The mean scores were compared between conventional (cold-run) training and VR-based training.

Statistical analyses were conducted using the Wilcoxon signed-rank test because the questionnaire responses were measured on a 5-point Likert scale. Effect sizes (r) were also calculated to evaluate the magnitude of the observed differences. Statistical significance was initially defined as $P<.05$. To account for multiple comparisons, Bonferroni correction was

XSL•FO
RenderX

applied separately to the primary outcomes (7 domains; adjusted α=.0071) and secondary outcomes (35 items; adjusted α=.0014). All quantitative analyses were performed using Microsoft Excel.

Qualitative Evaluation

In this study, open-ended responses corresponding to H1 to H3 were analyzed using a qualitative descriptive approach based on Kawakita Jiro method (KJ method), a systematic technique for organizing and interpreting qualitative data [15], supplemented by established qualitative analysis frameworks [16-18]. To ensure analytical neutrality, the analyses were conducted by 2 researchers who were not involved in the educational instruction or VR-based training.

The open-ended survey items analyzed were as follows:

- H1: Please describe any impressions or insights gained from the radiation field visualization experience.
- H2: Please describe any perceived differences between performing tasks in a real-world environment and in a VR environment.
- H3: Please describe any desired improvements or additional functions you would recommend for the system.

The analysis commenced with repeated readings of all responses to familiarize the researchers with the data and to identify meaningful statements related to radiation field visualization experiences, operational differences between conventional and VR-based training, and suggestions for system improvement. These statements were divided into discrete meaning units, grouped according to content similarity, and assigned descriptive labels. The 2 researchers independently reviewed and coded the responses before collaboratively comparing and refining the resulting codes, categories, and themes through discussion until consensus was reached. Following the principles of the KJ method, related groups were systematically compared and integrated into higher-order categories, from which overarching themes reflecting learners' subjective perceptions were extracted. This consensus-based process was used to enhance the credibility and consistency of the qualitative analysis.

This approach enabled the systematic extraction of learners' subjective experiences and insights into the educational impact that could not be adequately quantified using a Likert-scale–based quantitative approach [10].

Results

Quantitative Evaluation

Quantitative analyses were performed to examine differences in participants' self-reported learning experiences and perceptions between conventional and VR-based training. The 7 primary evaluation categories described in the "Methods" section were assessed using a Likert scale.

Primary Outcomes

The primary outcomes of this study are presented in Table 1. All 20 participants completed both conventional and VR-based training sessions and submitted questionnaires for each type of training. The 7 primary outcome domains demonstrated good internal consistency, with Cronbach α values of 0.822 and 0.850 for the conventional and VR-based training questionnaires, respectively.

Table . Comparison of primary outcomes between conventional training and virtual reality (VR)–based training.

Primary outcomes	Conventional training, mean (SD)	VR-based training, mean (SD)	Wilcoxon signed-rank test	*P* value	Effect size (*r*)	Significance (α=.0071)
(A) Radiation understanding and visualization experience	1.85 (0.45)	4.70 (0.25)	0	<.001	0.878	__[a]
(B) Operational understanding and technical skills (vial and shielding manipulation)	3.60 (0.52)	3.55 (0.66)	80.5	.83	0.049	n.s.[b]
(C) Perceived realism and immersion	4.44 (0.52)	3.48 (0.70)	9	<.001	0.774	__[a]
(D) Medical safety awareness and radiation protection	3.07 (0.57)	4.35 (0.46)	0	<.001	0.856	__[a]
(E) System evaluation (usability and performance)	4.24 (0.65)	4.05 (0.65)	54	.29	0.239	n.s.
(F) Learning satisfaction	3.88 (0.47)	4.52 (0.46)	0	<.001	0.835	__[a]
(G) Self-evaluation	3.81 (0.64)	4.48 (0.44)	0	<.001	0.811	__[a]

[a]Indicates statistical significance.

[b]n.s.: nonsignificant.

XSL•FO
RenderX

For the (A) radiation understanding and visualization domain, the mean score for conventional training was 1.85 (SD 0.45), whereas VR-based training yielded a significantly higher mean score of 4.70 (SD 0.25). The (B) operational understanding and technical skills (vial and shielding manipulation) domain showed comparable results across both training modalities, with mean scores of 3.60 (SD 0.52) for conventional training and 3.55 (SD 0.66) for VR-based training. In the (C) perceived realism and immersion domain, conventional training was rated higher (mean 4.44, SD 0.52) than VR-based training (mean 3.48, SD 0.70).

The (D) medical safety awareness and radiation protection domain yielded mean scores of 3.07 (SD 0.57) for conventional training and 4.35 (SD 0.46) for VR-based training. The (E) system usability and performance domain demonstrated comparable scores for conventional training and VR-based training (mean 4.24, SD 0.65 vs mean 4.05, SD 0.65). The (F) learning satisfaction domain yielded mean scores of 3.88 (SD 0.47) for conventional training and 4.52 (SD 0.46) for VR-based training, while (G) self-evaluation domain yielded mean scores of 3.81 (SD 0.64) for conventional training and 4.48 (SD 0.44) for VR-based training, with higher ratings reported for VR-based training in both domains.

Wilcoxon signed-rank tests revealed that participants reported significantly higher ratings for categories (A), (D), (F), and (G) following VR-based training than following conventional training, whereas higher ratings were reported for conventional training in category (C). No significant differences were observed in categories (B) and (E). The significant differences were generally associated with large effect sizes.

Secondary Outcomes

The secondary outcomes for each evaluation category are presented in Table 2, which shows the mean score (SD) for each item in both conventional and VR-based training. All 20 participants completed questionnaires for both training modalities. The questionnaire revised for this study, including additional items related to radiation visualization and VR-based training, demonstrated good internal consistency, with Cronbach α values of 0.893 and 0.914 for the conventional and VR-based training questionnaires, respectively.

XSL•FO
RenderX

Table . Comparison of secondary outcomes between conventional training and virtual reality (VR)–based training.

Secondary outcomes	Conventional training, mean (SD)	VR-based training, mean (SD)	Wilcoxon signed-rank test	*P* value	Effect size (*r*)	Significance (α=.0014)
A1	1.55 (0.83)	4.70 (0.47)	0	<.001	0.890	__[a]
A2	1.80 (0.77)	4.70 (0.47)	0	<.001	0.893	__[a]
A3	1.20 (0.52)	5.00 (0)	0	<.001	0.946	__[a]
A4	2.75 (0.72)	4.70 (0.47)	0	<.001	0.888	__[a]
A5	1.95 (0.76)	4.40 (0.50)	0	<.001	0.886	__[a]
B1	3.80 (1.06)	3.55 (0.94)	63.0	.51	0.146	n.s.[b]
B2	3.25 (0.72)	3.90 (0.97)	37.5	.06	0.427	n.s.
B3	4.60 (0.60)	2.40 (1.23)	0	<.001	0.794	__[a]
B4	3.05 (0.83)	4.00 (0.86)	5.5	.001324[c]	0.718	__[a]
B5	3.30 (0.98)	3.90 (0.64)	0	.002282[c]	0.682	n.s.
C1	4.20 (1.01)	3.45 (0.83)	38.0	.04	0.473	n.s.
C2	4.75 (0.44)	3.20 (0.95)	0	<.001	0.821	__[a]
C3	4.85 (0.37)	2.85 (1.04)	0	<.001	0.822	__[a]
C4	4.40 (0.82)	4.10 (0.97)	9	.19	0.289	n.s.
C5	4.00 (0.79)	3.80 (1.01)	41.5	.48	0.160	n.s.
D1	3.25 (0.64)	4.40 (0.68)	5.5	<.001	0.774	__[a]
D2	3.15 (0.67)	4.30 (0.80)	0	<.001	0.783	__[a]
D3	2.95 (0.76)	4.55 (0.76)	0	<.001	0.800	__[a]
D4	3.05 (0.83)	4.75 (0.44)	0	<.001	0.837	__[a]
D5	2.95 (1.00)	3.75 (0.97)	0	.001491[c]	0.710	n.s.
E1	3.75 (1.37)	4.05 (0.94)	29.5	.45	0.170	n.s.
E2	4.45 (0.76)	4.10 (0.91)	23.5	.21	0.282	n.s.
E3	4.50 (0.76)	3.80 (0.89)	23.0	.03	0.487	n.s.
E4	4.20 (0.89)	4.05 (1.10)	50.5	.58	0.122	n.s.
E5	4.30 (0.98)	4.25 (0.72)	25.0	.79	0.058	n.s.
F1	3.25 (0.91)	4.70 (0.47)	0	<.001	0.826	__[a]
F2	3.15 (0.75)	4.70 (0.47)	0	<.001	0.850	__[a]
F3	3.90 (0.72)	4.35 (0.59)	9	.05	0.439	n.s.
F4	4.60 (0.60)	4.35 (0.81)	10.5	.27	0.246	n.s.
F5	4.50 (0.69)	4.50 (0.69)	27.5	>.99	0	n.s.
G1	3.20 (1.20)	4.60 (0.60)	0	<.001	0.804	__[a]
G2	3.90 (1.12)	3.85 (1.04)	20.0	.76	0.068	n.s.
G3	4.10 (0.72)	4.65 (0.49)	0	.008	0.589	n.s.
G4	3.55 (0.83)	4.75 (0.44)	6	<.001	0.799	__[a]
G5	4.30 (0.66)	4.55 (0.51)	20.0	.22	0.275	n.s.

[a]Indicates statistical significance.

[b]n.s.: nonsignificant.

[c]*P* values close to the Bonferroni-adjusted significance threshold (α=.0014), requiring additional precision to distinguish statistically significant from

XSL•FO
RenderX

nonsignificant results. B4 is statistically significant, whereas B5 and D5 are nonsignificant.

For (A) radiation understanding and visualization domain, all 5 subitems (A1-A5) demonstrated markedly higher scores with VR-based training than with conventional training. For example, A1 (“intuitive understanding of the spatial extent and distribution of radiation”) had a mean score of 1.55 (SD 0.83) for conventional training and 4.70 (SD 0.47) for VR-based training, and A3 (“visual perception of radiation intensity and attenuation through color changes”) increased from 1.20 (SD 0.52) to 5.00 (SD 0).

For (B) operational understanding and technical skills (vial and shielding manipulation) domain, results varied across items. B3 (“sufficiently experienced realistic operational sensations, such as weight and spatial positioning”) was higher in conventional training (mean 4.60, SD 0.60) than in VR-based training (mean 2.40, SD 1.23), whereas B4 (“performed tasks while consciously applying radiation protection principles”) increased from a mean score of 3.05 (SD 0.83) in conventional training to 4.00 (SD 0.86) in VR-based training. The remaining items showed no substantial differences between the 2 training modalities.

For (C) perceived realism and immersion domain, conventional training scored higher on some items, such as C2 (“sense of presence similar to an actual working environment”; mean 4.75, SD 0.44 vs mean 3.20, SD 0.95); however, the remaining items were comparable between the 2 methods.

For (D) safety awareness and radiation protection domain, VR-based training yielded higher scores for most items. For instance, the mean score of D3 (“recognition of the importance of distance, shielding, and time principles through the experience”) increased from 2.95 (SD 0.76) in conventional training to 4.55 (SD 0.76) in VR-based training; however, similar patterns were observed for D1, D2, and D4.

For (E) system usability and performance domain, similar scores were observed across all items; however, no statistically significant differences were observed between conventional and VR-based training for any item in this domain.

For (F) learning satisfaction domain, “higher ratings were observed for VR-based training in” F1 (mean 3.25, SD 0.91 vs mean 4.70, SD 0.47) and F2 (mean 3.15, SD 0.75 vs mean 4.70, SD 0.47); however, the remaining items were comparable between the 2 methods.

Finally, in (G) self-evaluation domain, VR-based training yielded higher scores for G1 (mean 3.20, SD 1.20 vs mean 4.60, SD 0.60) and G4 (mean 3.55, SD 0.83] vs mean 4.75, SD 0.44) than conventional training. The remaining items did not differ substantially.

Overall, participants reported higher ratings for VR-based training primarily in domains related to radiation understanding and visualization, medical safety awareness and radiation protection, learning satisfaction, and self-evaluation, whereas conventional training received higher ratings for selected items related to physical and operational realism.

Qualitative Findings: Comparison Between Real and VR Environments

Overview

Open-ended responses from 20 participants were analyzed using a qualitative descriptive approach based on the KJ method [15]. The frequencies presented below are intended to provide a descriptive summary of the occurrence of participant comments and should not be interpreted as quantitative outcome measures. A summary of the identified qualitative themes and the number of participants contributing to each theme is presented in Table 3. Following this methodology, the following five key themes were identified from the participants’ responses:

1. Radiation field visualization in VR: Participants described how the radiation distribution, invisible during conventional training, became perceptible within the VR-based training environment.
2. Radiation safety awareness in VR: Participants reported heightened attention to radiation protection while operating in the VR-based training environment.
3. Operational differences between conventional and VR-based training: Participants compared tactile sensations, weight perception, and procedural manipulation across both training modalities.
4. Practical skill acquisition in conventional training: Participants emphasized that hands-on practice in the conventional training environment remained important for procedural skill development.
5. VR-based training usability and suggestions for improvement: Participants reported operational difficulties and proposed recommendations for enhancing the VR-based training system.

XSL•FO
RenderX

Table . Summary of qualitative themes identified from participants' open-ended responses (N=20).

Theme	Participants[a], n
Radiation field visualization in VR[b]	20
Radiation safety awareness in VR	15
Operational differences between conventional and VR-based training	20
Practical skill acquisition in conventional training	8
VR-based training usability and suggestions for improvement	10

[a]"Participants" indicates the number of participants contributing to each theme. Individual participants could contribute to more than one theme; therefore, the totals exceed 20.

[b]VR: virtual reality.

The following subsections summarize participants' responses according to these 5 themes.

Radiation Field Visualization in VR-Based Training

All participants (20/20) reported that radiation, which was not directly observable during conventional training, became perceptible in the VR-based training environment. Representative statements included the following:

> *Radiation could not be observed during conventional training; however, in the VR-based training environment, the radiation field became visible.*
>
> *Visualizing this field allowed me to identify areas of high exposure risk.*
>
> *Observing the depth and spread of the radiation made the associated risks easier to understand.*

Radiation Safety Awareness in VR

Fifteen participants stated that VR-based training heightened their radiation safety awareness. Representative statements included the following:

> *In the VR-based training environment, I was able to perform the procedure while simultaneously considering radiation exposure.*
>
> *Visualizing the radiation field heightened my attention to safety management.*
>
> *I became more conscious of shielding and maintaining distance.*

The participants reported that during conventional training, attention was primarily directed toward procedural execution, whereas the VR-based training environment allowed simultaneous consideration of radiation exposure and safety measures.

Operability Differences Between Conventional and VR Environments

All participants (20/20) reported differences in operability between the conventional and VR-based training environments. Common observations included the following:

- The presence of tactile feedback and weight perception in the conventional training environment.
- Absence of haptic feedback in the VR-based training environment.
- Challenges in accurately judging spatial distance within the VR-based training environment.
- Difficulties in manipulating interface controls during VR-based training.

Representative comments included the following:

> *In the conventional training environment, I could feel the weight of the syringe.*
>
> *In the VR-based training environment, it was difficult to perceive the distance.*
>
> *Operating the interface buttons in VR-based training was more challenging than handling real objects.*

Participants frequently compared tactile feedback, weight perception, spatial distance perception, and interface operability between the conventional and VR-based training environments.

Practical Skill Acquisition in Conventional Training

Eight participants emphasized that procedural skills could only be reliably developed in a conventional training environment. Representative statements included the following:

> *In the conventional training environment, I was able to practice actual manipulations.*
>
> *Handling real objects deepened my understanding of their physical structure and weight.*
>
> *The conventional training environment was superior for procedural skill development.*

Participants emphasized the importance of hands-on practice using real objects for procedural skill acquisition.

VR-Based Training Usability and Improvement Suggestions

Ten participants reported usability limitations in the VR-based training and proposed the following targeted system improvements:

- Difficulty in accurately judging spatial distances within the VR-based training environment.
- Visual focus misalignment experienced during VR-based training.
- Operational difficulties with VR interface controls.
- Requests for more realistic procedural tasks.

Representative comments included the following:

XSL•FO
RenderX

> *Visual focus did not align accurately within the VR-based training environment, and determining the distance to interface buttons was challenging. It would be better if actual injection procedures could be performed during the VR-based training.*

Participants frequently reported technical limitations related to spatial perception, visual alignment, interface operation, and procedural realism in the VR-based training environment.

Discussion

Principal Findings

Participants reported higher ratings for radiation field understanding and visualization, radiation safety awareness, learning satisfaction, and self-evaluation following VR-based training than following conventional training. The improvements in radiation field understanding and visualization were expected because visualization of otherwise invisible radiation was a core design feature of the VR system and should therefore be interpreted as confirmation of the intended system functionality rather than independent evidence of educational superiority. By contrast, no significant differences were observed in operational understanding, technical skills, or system usability. Notably, perceived realism and immersion scores were significantly higher in the conventional training environment than in the VR-based training. Because all participants had already completed both conventional cold-run training and supervised hot-run training approximately 5 months before this study, these findings should be interpreted as reflecting perceptions formed after prior practical experience rather than those of novice learners.

These quantitative findings were supported by qualitative analysis. Participants emphasized the value of radiation field visualization in VR-based training, whereas the higher ratings for realism-related items were considered to reflect inherent characteristics of the respective training environments. The qualitative findings also suggested that the lower realism and immersion ratings for the VR-based training were primarily attributable to limitations in tactile feedback and object manipulation rather than to the educational content itself. Future advances in VR hardware, including improvements in haptic feedback, hand tracking, and display technologies, may further enhance realism and immersion in VR-based training. Collectively, these results suggest that VR-based training may support positive perceived learning experiences related to cognitive and conceptual learning domains, whereas conventional training retains advantages in experiential immersion and psychomotor realism. Beyond the expected benefit of radiation field visualization, the observed improvements in radiation safety awareness, learning satisfaction, and self-evaluation suggest that the educational value of VR-based training extends beyond visualization alone.

Improvement in Radiation Visualization and Safety Awareness

The most substantial effect of VR-based training was observed in radiation field understanding and visualization ($P<.001$). Because radiation is inherently invisible, conventional training based on static diagrams and theoretical explanations may limit learners' intuitive understanding of spatial dose distribution and shielding effects.

Previous research has highlighted the limitations of lecture-based radiation safety education and the need for practical, perceptually enriched training strategies [19]. Previous VR studies have used interactive radiation field visualization to support learners' understanding of spatial dose distributions, shielding, exposure patterns, and protective positioning [20,21]. The current findings are consistent with previous reports showing that VR-based training was associated with higher ratings for radiation field visualization and radiation safety awareness. These findings suggest that immersive radiation visualization may help bridge the gap between abstract radiation concepts and practical radiation safety education.

Learning Satisfaction and Self-Evaluation

Learning satisfaction ($P<.001$) and self-evaluation ($P<.001$) were significantly higher following VR-based training. These gains are likely attributable to the interactive nature, immediate visual feedback, and immersive simulation afforded by the VR-based training environment. However, these findings should be interpreted with caution because the novelty of the immersive VR environment and the inability to blind participants may have contributed to higher satisfaction and self-evaluation ratings.

Previous meta-analyses in health care education have demonstrated that immersive VR-based training yielded greater improvements in learner satisfaction than conventional training [22]. Immersive environments foster emotional engagement and perceived educational value, positively influencing the motivational dimensions of learning [23]. VR has the potential to promote engagement and active participation, which are important factors influencing educational satisfaction [24]. In our previous investigation of VR-based training for radiopharmaceutical administration, higher levels of a sense of presence during VR-based training were associated with improved perceived educational outcomes [6]. The qualitative responses in this study similarly indicated that participants perceived VR-based training as an educationally meaningful tool. Collectively, these findings suggest that VR-based training may support both the conceptual and motivational dimensions of participants' perceived learning experiences.

Operability and Technical Skills

No significant differences were observed in operational understanding or technical skills ($P=.83$), indicating that the current VR-based training implementation did not outperform conventional training in psychomotor skill acquisition. Although conceptual comprehension improved, procedural performance remained comparable across both training modalities.

Qualitative analysis revealed frequent references to operability differences and the importance of tactile feedback in conventional training environments. Participants consistently emphasized that physical resistance, fine motor control, and authentic material manipulation were critical contributors to procedural mastery. Likewise, limitations in haptic fidelity and usability have been recognized as barriers to effective

XSL•FO
RenderX

psychomotor skill transfer in VR-based training [25,26], which is consistent with the qualitative findings of the present study.

Technical constraints related to finger tracking and standardized VR glove sizing may have reduced fine motor precision and perceived operability, particularly when the glove dimensions did not adequately match participants' hand sizes. These factors may partly explain why VR-based training did not outperform conventional training in perceived psychomotor skills.

These findings suggest that limitations in haptic realism, ergonomic adaptability, and interface customization may constrain perceived technical skill development. Adaptive haptic systems and customizable interfaces may help address these limitations.

Perceived Realism and Immersion

Notably, ratings for perceived realism and immersion were significantly higher ($P<.001$) in conventional training than in VR-based training. Participants may have rated the conventional training environment as more realistic and immersive because of multisensory integration, tactile feedback, and environmental complexity.

In this study, the marked differences in object weight (3.86 g and 13.3 g without shielding vs 114 g and 602 g with shielding) likely contributed to this perceptual discrepancy. These differences provided proprioceptive, kinesthetic, and tactile cues that were absent in the current VR-based training environment, where virtual objects had no physical weight. Such a fundamental incongruence between actual force feedback and the visual representation may reduce the perceived physical realism and immersion in VR-based training, particularly during fine motor tasks that require accurate weight discrimination and resistance perception.

Previous research has demonstrated that incorporating haptic feedback into VR-based training enhances realism and learner confidence in procedural tasks [27,28]. Consistent with these previous findings, participants in the present study reported lower ratings for perceived realism and immersion in the absence of realistic weight perception and tactile feedback. These findings suggest that perceived immersion may depend not only on visual fidelity but also on congruent multisensory inputs.

Emerging MR approaches may help address this sensory incongruence by allowing interaction with real objects while overlaying digital guidance, thereby preserving the authentic physical properties of weight and resistance [29]. Moreover, advances in tactile technologies for VR-based training in radiation education suggest that integrating sensor-based or vibrotactile feedback into VR glove systems could further strengthen perceived realism, immersion, fine-motor realism, and overall perceived operability [30]. Collectively, these findings suggest that the 2 training modalities engage different experiential and sensorimotor mechanisms. Integrating MR and advanced haptic systems may therefore combine the visual advantages of VR-based training with the tactile authenticity of conventional training.

Educational Implications

The findings indicate that VR-based training may be particularly well-suited to undergraduate radiological education, where students are still developing core procedural competencies and foundational radiation protection awareness. VR-based training transforms abstract and imperceptible radiation phenomena into visually and spatially interpretable representations, which may facilitate conceptual understanding of radiation and enhance perceived learning experiences. Accordingly, these findings are most applicable to undergraduate radiological technology students who have already acquired fundamental procedural skills through conventional practical training.

Conventional training remains important for tactile engagement and realistic procedural experience but does not provide direct visualization of spatial radiation distributions. These findings support a complementary framework in which VR-based training reinforces conceptual understanding and radiation protection awareness alongside conventional practical training.

Limitations

This study has some limitations that should be considered when interpreting the findings.

First, all participants completed the conventional training before the VR-based training. Therefore, potential order, practice, fatigue, and recency effects could not be separated from the effects of the training modality. Although all participants had previously completed both conventional cold-run training and supervised hot-run training and therefore were not novices, the influence of the fixed-order design cannot be excluded. Consequently, the observed differences should be interpreted as associations rather than evidence of causal educational effects. Future studies using randomized or counterbalanced study designs are warranted to better isolate the educational effects of VR-based training. Furthermore, this study evaluated participants' perceived learning experiences using self-reported questionnaires and qualitative feedback and did not include objective assessments of knowledge acquisition or procedural performance. Future studies incorporating objective educational outcome measures, such as knowledge tests and performance-based assessments, are warranted to provide a more comprehensive evaluation of the educational effectiveness of VR-based training.

Second, this study was conducted at a single institution with a relatively small sample size consisting of undergraduate students, which may restrict the generalizability to broader educational contexts or to experienced health care professionals. As the participants were students in the early stages of training, the results primarily reflect educational effects at a foundational developmental level and cannot be extrapolated to established clinical or professional competence.

Third, the study relied exclusively on self-reported perceptual measures and lacked objective assessments of operational or procedural skills, such as objective structured clinical examination–based evaluations or performance-based metrics. Therefore, conclusions regarding operational skills or procedural competence should be interpreted cautiously.

XSL•FO
RenderX

Fourth, although the questionnaire was adapted from a previously validated instrument, several items were modified to address radiation visualization and VR-based training experiences. Therefore, the psychometric properties of the modified questionnaire should be further examined in future studies.

Fifth, participants were recruited on a voluntary basis, which may have introduced selection bias toward learners with a greater interest in educational technologies.

Sixth, participants completed the familiar conventional training before experiencing the novel VR-based training environment. Therefore, the higher ratings for learning satisfaction and self-evaluations may have been influenced not only by the characteristics of the VR-based training itself but also by novelty effects and demand characteristics. Because blinding was not feasible in this educational intervention, these potential influences cannot be excluded.

Seventh, only immediate posttraining outcomes were assessed; long-term knowledge retention and transfer to authentic clinical practice remain entirely unexamined. The absence of advanced haptic feedback in the VR-based training system, as well as potential variability in VR glove fit, particularly for students whose hand size did not adequately match the glove dimensions, may also have affected the perceived operability, realism, and immersion during VR-based training. Although the VR system was calibrated using dedicated sensors before each training session, and no obvious misalignment between the visualized objects and users' hand movements was perceived during operation with the HMD and VR gloves, the spatial accuracy of the visualization was not quantitatively evaluated.

Furthermore, because this study included only undergraduate radiological technology students from a single institution who had already completed conventional practical training, the findings may not be generalizable to novice learners, experienced health care professionals, or other educational settings. Future investigations incorporating objective skill assessment, longitudinal follow-up to evaluate knowledge retention, quantitative validation of spatial registration and visualization accuracy, enhanced haptic interfaces and MR technologies in VR-based training, and multicenter validation across different learner populations would strengthen the evidence base for the systematic integration of VR-based training into radiological education.

Conclusions

This study found that participants reported significantly higher ratings for radiation visualization, radiation protection awareness, learner satisfaction, and self-evaluation following VR-based training, as evidenced by both quantitative and qualitative findings. These findings suggest that VR-based training may provide positive perceived learning experiences related to understanding invisible radiation phenomena and radiation protection awareness. Conversely, conventional training was associated with higher ratings for experiential immersion and realistic operational sensation. Rather than functioning as a substitute for established training methods, VR-based training should be strategically positioned as a complementary educational modality that supports positive perceived learning experiences related to conceptual understanding of radiation and radiation protection awareness.

Acknowledgments

The authors sincerely thank the students in the Radiological Technologist Training Program for their participation in this study. The authors also thank Dr Kazuhiro Ooe and colleagues at the Radioisotope Research Center, Osaka University, for providing facilities and equipment for the measurements used to map spatial radiation dose rates. Their invaluable support was essential to the successful completion of this study. Google Translate was partly used to assist with the translation of the authors' original work into English for inclusion in the manuscript. We would like to thank Editage for English-language editing. Generative AI was not used at any stage of the planning, VR development, implementation, or data analysis of this study.

Funding

AK was supported by the Japan Society for the Promotion of Science through a Grant-in-Aid for Early-Career Scientists (grant number: 22K13770), which funded this study. The funder played no role in the study design, data collection and analysis, decision to publish, or manuscript preparation. The other authors received no funding.

Data Availability

The datasets generated and analyzed in this study contain sensitive information from human participants and are not publicly available to protect participants' privacy. Data may be made available upon reasonable request to the corresponding author, subject to approval by the institutional ethics committee and in accordance with applicable ethical and privacy regulations.

Authors' Contributions

AK and DF conceived and designed the study and were responsible for its overall implementation. DF and SH developed the VR-based training module and analyzed the associated data. SM and AY measured the spatial radiation dose rates and performed the quantitative evaluations. SW and YO conducted the experimental measurements and performed both quantitative and qualitative data analyses. All authors contributed to the interpretation of the results and critically revised the manuscript.

XSL•FO
RenderX

Conflicts of Interest

None declared.

Multimedia Appendix 1
Color maps of spatial dose rate distributions.
[DOCX File, 297 KB - mededu_v12i1e104397_app1.docx]

Multimedia Appendix 2
Questionnaire domains and items used for quantitative and qualitative evaluation.
[DOCX File, 17 KB - mededu_v12i1e104397_app2.docx]

References

1. Gårdling J, Viseu C, Hettinger E, Jildenstål P, Augustinsson A. The effects of virtual reality (VR) on clinical skills training in undergraduate radiography education: a systematic review. Radiography (Lond) 2025 May;31(3):102911. [doi: 10.1016/j.radi.2025.102911] [Medline: 40048819]
2. Shetty S, Bhat S, Al Bayatti S, et al. The scope of virtual reality simulators in radiology education: systematic literature review. JMIR Med Educ 2024 May 8;10:e52953. [doi: 10.2196/52953] [Medline: 38722205]
3. Tene T, Bonilla García N, Coello-Fiallos D, Borja M, Vacacela Gomez C. A systematic review of immersive educational technologies in medical physics and radiation physics. Front Med (Lausanne) 2024;11:1384799. [doi: 10.3389/fmed.2024.1384799] [Medline: 39450107]
4. Rainford L, Tcacenco A, Potocnik J, et al. Student perceptions of the use of three-dimensional (3-D) virtual reality (VR) simulation in the delivery of radiation protection training for radiography and medical students. Radiography (Lond) 2023 Jul;29(4):777-785. [doi: 10.1016/j.radi.2023.05.009] [Medline: 37244141]
5. Takahashi H, Nakamura Y, Fujiwara A. Virtual reality vs traditional lecture-based methods in radiation safety continuing medical education. Radiol Technol 2025;97(2):72-81. [Medline: 41348632]
6. Kakimoto A, Fujise D, Hasegawa S, et al. Virtual reality-based training for radiopharmaceutical administration: development and educational effectiveness. PLOS ONE 2025;20(3):e0321101. [doi: 10.1371/journal.pone.0321101] [Medline: 40163490]
7. Siegel JA, Marcus CS, Sparks RB. Calculating the absorbed dose from radioactive patients: the line-source versus point-source model. J Nucl Med 2002 Sep;43(9):1241-1244. [Medline: 12215565]
8. Bushberg JT, Seibert JA, Leidholdt EM, Boone JM. The Essential Physics of Medical Imaging, 3rd edition: Lippincott Williams & Wilkins; 2011.
9. Mahesh M. The essential physics of medical imaging, Third Edition. Med Phys 2013 Jul;40(7). [doi: 10.1118/1.4811156] [Medline: 28524933]
10. Likert R. A technique for the measurement of attitudes. Arch Psychol 1932;22(140):1-55 [FREE Full text]
11. Kakimoto A, Watanabe S, Otaka Y, et al. Educational experiences associated with cold- and hot-run practical training on the handling of unsealed radioactive sources by radiological technologist students. BMC Med Educ 2026 Jun 5;26(1):1301. [doi: 10.1186/s12909-026-09594-w] [Medline: 42243921]
12. Cronbach LJ. Coefficient alpha and the internal structure of tests. Psychometrika 1951 Sep;16(3):297-334. [doi: 10.1007/BF02310555]
13. Kaiser HF. An index of factorial simplicity. Psychometrika 1974 Mar;39(1):31-36. [doi: 10.1007/BF02291575]
14. Bartlett MS. A note on the multiplying factors for various χ2 approximations. J R Stat Soc B 1954 Jul;16(2):296-298. [doi: 10.1111/j.2517-6161.1954.tb00174.x]
15. Scupin R. The KJ method: a technique for analyzing data derived from Japanese ethnology. Hum Organ 1997;56(2):233-237. [doi: 10.17730/humo.56.2.x335923511444655]
16. Sandelowski M. Whatever happened to qualitative description? Res Nurs Health 2000 Aug;23(4):334-340. [doi: 10.1002/1098-240x(200008)23:4<334::aid-nur9>3.0.co;2-g] [Medline: 10940958]
17. Braun V, Clarke V. Using thematic analysis in psychology. Qual Res Psychol 2006 Jan;3(2):77-101. [doi: 10.1191/1478088706qp063oa]
18. Elo S, Kyngäs H. The qualitative content analysis process. J Adv Nurs 2008 Apr;62(1):107-115. [doi: 10.1111/j.1365-2648.2007.04569.x] [Medline: 18352969]
19. Faggioni L, Paolicchi F, Bastiani L, Guido D, Caramella D. Awareness of radiation protection and dose levels of imaging procedures among medical students, radiography students, and radiology residents at an academic hospital: results of a comprehensive survey. Eur J Radiol 2017 Jan;86:135-142. [doi: 10.1016/j.ejrad.2016.10.033] [Medline: 28027740]
20. Takata T, Kondo H, Yamamoto M, et al. Immersive radiation experience for interventional radiology with virtual reality radiation dose visualization using fast Monte Carlo dose estimation. Interv Radiol (Higashimatsuyama) 2020;5(2):58-66. [doi: 10.22575/interventionalradiology.2019-0007] [Medline: 36284664]
21. Fujibuchi T. Radiation protection education using virtual reality for the visualisation of scattered distributions during radiological examinations. J Radiol Prot 2021 Sep 27;41(4). [doi: 10.1088/1361-6498/ac16b1] [Medline: 34289461]

XSL•FO
RenderX

22. Chen FQ, Leng YF, Ge JF, et al. Effectiveness of virtual reality in nursing education: meta-analysis. J Med Internet Res 2020 Sep 15;22(9):e18290. [doi: 10.2196/18290] [Medline: 32930664]
23. Makransky G, Lilleholt L. A structural equation modeling investigation of the emotional value of immersive virtual reality in education. Education Tech Research Dev 2018;66(5):1141-1164. [doi: 10.1007/s11423-018-9581-2]
24. Pottle J. Virtual reality and the transformation of medical education. Future Healthc J 2019 Oct;6(3):181-185. [doi: 10.7861/fhj.2019-0036] [Medline: 31660522]
25. Seymour NE, Gallagher AG, Roman SA, et al. Virtual reality training improves operating room performance: results of a randomized, double-blinded study. Ann Surg 2002 Oct;236(4):458-463. [doi: 10.1097/00000658-200210000-00008] [Medline: 12368674]
26. Alaker M, Wynn GR, Arulampalam T. Virtual reality training in laparoscopic surgery: a systematic review & meta-analysis. Int J Surg 2016 May;29:85-94. [doi: 10.1016/j.ijsu.2016.03.034] [Medline: 26992652]
27. Sapkaroski D, Baird M, McInerney J, Dimmock MR. The implementation of a haptic feedback virtual reality simulation clinic with dynamic patient interaction and communication for medical imaging students. J Med Radiat Sci 2018 Sep;65(3):218-225. [doi: 10.1002/jmrs.288] [Medline: 30006966]
28. Collaço E, Kira E, Sallaberry LH, et al. Immersion and haptic feedback impacts on dental anesthesia technical skills virtual reality training. J Dent Educ 2021 Apr;85(4):589-598. [doi: 10.1002/jdd.12503] [Medline: 33274441]
29. Jarzembak JM, Biggs J, James A, et al. Immersive learning in nursing: a mixed reality approach to IV simulation with bimanual haptic feedback. Nurse Educ 2025;50(6):335-340. [doi: 10.1097/NNE.0000000000001941] [Medline: 40705668]
30. Fujise D, Koba Y, Hasegawa S, Negishi T. Development and evaluation of educational systems using airborne tactile technology for medical radiation protection. Sens Mater 2025;37(1):63-74. [doi: 10.18494/SAM5418]

Abbreviations

HMD: head-mounted display
KJ method: Kawakita Jiro method
KMO: Kaiser-Meyer-Olkin
MR: mixed reality
SDK: software development kit
VR: virtual reality
Tc: Technetium-99m

Edited by M Stadler; submitted 11.Jun.2026; peer-reviewed by B Patel, M Siraz; revised version received 22.Aug.2026; accepted 03.Sep.2026; published 30.Sep.2026.

Please cite as:
Kakimoto A, Fujise D, Watanabe S, Otaka Y, Minemoto S, Yasumatsu A, Hasegawa S
Development and Educational Evaluation of a Radiation Visualization System in a Virtual Reality Training Module for Radiopharmaceutical Administration: Mixed Methods Study of Students' Learning Experiences
JMIR Med Educ 2026;12:e104397
URL: https://mededu.jmir.org/2026/1/e104397
doi:10.2196/104397

© Akihiro Kakimoto, Daisuke Fujise, Shota Watanabe, Yutaka Otaka, Sakura Minemoto, Ayaka Yasumatsu, Shin Hasegawa. Originally published in JMIR Medical Education (https://mededu.jmir.org), 30.Sep.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Education, is properly cited. The complete bibliographic information, a link to the original publication on https://mededu.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Refining the Immersive Technology Evaluation Measure for Virtual Reality Clinical Skills Among Native Arabic-Speaking Medical Students: Qualitative Study

Afaf Sulaiman Alblooshi[1*], MD, PhD; Falah Mohammed Almarzooqi[2*], MBchB, CCFP; Taleb M Almansoori[3], MBBS, FRCPC; Anmol Punn[1], MD, MPH; Marwa Gaffar Alameen[1], MD, MScHPE; Alexander Kieu[2], MD; Faten Abdullah AlRadini[4], MD

[1]Department of Medical Education, College of Medicine and Health Sciences, United Arab Emirates University, Al Ain, Abu Dhabi, United Arab Emirates

[2]Department of Family Medicine, College of Medicine and Health Sciences, United Arab Emirates University, Al Ain, Abu Dhabi, United Arab Emirates

[3]Department of Radiology, College of Medicine and Health Sciences, United Arab Emirates University, Sheikh Khalifa Street, Al Ain, Abu Dhabi, United Arab Emirates

[4]Department of Family and Community Medicine, College of Medicine, Princess Nourah bint Abdulrahman University, Riyadh, Riyadh Region, Saudi Arabia

[*]these authors contributed equally

Corresponding Author:
Taleb M Almansoori, MBBS, FRCPC
Department of Radiology, College of Medicine and Health Sciences, United Arab Emirates University, Sheikh Khalifa Street, Al Ain, Abu Dhabi, United Arab Emirates

Abstract

Background: Extended reality technologies, including virtual reality (VR), augmented reality, and mixed reality, are increasingly used in medical education to create immersive and interactive learning environments. As these modalities expand, validated instruments are needed to measure learners' experiences accurately. The Immersive Technology Evaluation Measure (ITEM) is a multidomain questionnaire assessing immersion, motivation, cognitive load, usability, and debriefing. Although cognitive interviewing informed its original development, less is known about how ITEM questions function when used in a different linguistic and educational context.

Objective: This study aimed to evaluate the clarity, comprehensibility, and response processes of ITEM among native Arabic-speaking medical students enrolled in an English-medium medical program. We also sought to identify linguistic, referential, and contextual sources of comprehension difficulty and use these findings to inform proposed item-level refinements.

Methods: We conducted a qualitative cognitive interviewing study with 10 third-year and fourth-year medical students at United Arab Emirates University following VR-based clinical skills activities. Using concurrent think-aloud and verbal probing techniques, participants explained how they interpreted each ITEM question and arrived at their responses. Interview transcripts, recordings, and interviewer notes were independently reviewed using a descriptive, item-focused approach. Participant feedback was examined for recurring comprehension problems, and proposed item-level decisions were reviewed through research team consensus. Item-level saturation was reached after 8 interviews and confirmed with 2 additional interviews.

Results: Participants identified comprehension or response-process problems in 47.5% (19/40) of ITEM questions. Difficulties occurred across all 5 domains and were most frequent in immersion (6/9, 66.7%) and usability (6/10, 60%), followed by debriefing (4/5, 80%), motivation (2/10, 20%), and cognitive load (1/6, 16.7%). Common difficulties involved nonspecific references such as "activity" and "technology," unfamiliar or abstract terminology, negative wording, and unclear temporal or contextual framing. For example, "concern" was interpreted by some participants as worry rather than focus, and "mentally demanding" was interpreted in relation to mental health rather than cognitive effort. Cognitive interviewing informed different item-level decisions rather than a uniform revision: 11 problematic questions received proposed wording or contextual revisions, 2 received presentation-only modifications, and 6 were retained unchanged where clarification could introduce a meaning not clearly established in the original item. Six additional questions received limited contextual or terminology-standardizing changes for consistency. No items were removed, and the original 5-domain structure was retained.

Conclusions: Cognitive interviewing identified response-process difficulties that were not apparent from the questionnaire wording alone and provided a systematic basis for determining when clarification was warranted and when the original wording should be preserved. The findings extend response-process evidence for ITEM and illustrate the value of examining established

XSL•FO
RenderX

educational measures in new linguistic and educational settings. The proposed refinements provide a foundation for further cognitive testing and psychometric evaluation across immersive learning contexts.

(JMIR Med Educ 2026;12:e95904) doi:10.2196/95904

KEYWORDS

interview; extended reality; virtual reality; immersive technology; interactive learning environments; medical education

Introduction

Extended reality (XR) technologies, including virtual reality (VR), augmented reality (AR), and mixed reality (MR), are increasingly used in medical education to create immersive and interactive learning environments [1-4]. These modalities allow learners to engage with simulated clinical scenarios that support decision-making, procedural skills practice, and teamwork in a safe and engaging learning environment [5,6]. As immersive learning platforms expand in medical education, there is a parallel need for validated instruments that can reliably assess learners' experiences and outcomes [7].

Several established questionnaires assess individual components of immersive learning; however, these tools are often applied independently and may fail to capture the broader, multidimensional nature of immersive learning experiences. For example, usability and technology acceptance are commonly assessed using the System Usability Scale (SUS) [8] and the technology acceptance model [9], presence and immersion using the iGroup Presence Questionnaire [10], perceived cognitive workload using the NASA Task Load Index (NASA-TLX) [11], and learner motivation using the intrinsic motivation inventory [12]. This limitation is particularly relevant in XR-based education, where learners' perceptions may be shaped simultaneously by the level of immersion, system usability, motivation, cognitive demands, and opportunities for structured reflection. A comprehensive measure integrating complementary constructs can therefore support a more complete evaluation of learner experience in immersive technology-enhanced learning (TEL).

To address the need for a multidomain evaluation instrument, Jacobs et al [7] developed the Immersive Technology Evaluation Measure (ITEM), a 40-item self-report questionnaire designed to assess key dimensions of the learner's experience in immersive environments. ITEM integrates 5 established measures and frameworks: the Adapted Immersion Experience Questionnaire (AIEQ) [13,14], abridged intrinsic motivation inventory (AIMI) [14,15], NASA-TLX [11], SUS [8], and PEARLS (Prompts for Engaging and Reflective Learning in Simulation) debriefing tool [16]. Together, these domains assess immersion, intrinsic motivation, cognitive load, system usability, and debriefing. Cognitive interviewing was included in the original development of ITEM to assess participants' understanding of its questions [7]. However, cognitive evaluation during initial instrument development does not necessarily establish that questions will be interpreted similarly when an instrument is subsequently used in different linguistic, cultural, or educational contexts.

Questionnaires are widely used in medical education research to evaluate learners' perceptions, attitudes, and experiences [17-20]. Notably, even survey items that appear straightforward to researchers can be interpreted differently by respondents, potentially introducing misunderstanding and response error [21-23]. Cognitive interviewing is a qualitative questionnaire-evaluation approach that examines how respondents understand questions and formulate responses [22-24]. The survey response process has commonly been conceptualized in terms of comprehension, retrieval of relevant information, judgment, and response formulation [24]. Cognitive interviewing uses techniques such as think-aloud (TA) and verbal probing (VP) to explore these processes and identify problems involving terminology, ambiguous wording, question complexity, and contextual interpretation [22,25-28]. It has also been used to examine whether questionnaire items function similarly across cultural and linguistic contexts [28-30].

This consideration is particularly relevant when respondents complete questionnaires in a language that is not their first language. Previous research has demonstrated that language proficiency and linguistic characteristics of survey questions, including complex syntax, ambiguous wording, and less familiar terminology, warrant consideration when questionnaires are administered across linguistic groups [31,32]. Such issues are relevant to medical education in the United Arab Emirates, where medical training may be delivered in English within a predominantly Arabic-speaking clinical and social environment [33]. Research involving Arabic-speaking university students has also demonstrated differences in knowledge and attitude scores when the same health-related questionnaire was administered in Arabic and English [34]. These findings do not imply inadequate English proficiency among learners studying in English-medium programs; rather, they highlight the potential value of examining how specific questionnaire terms and expressions are interpreted when the survey language differs from respondents' first language.

Although cognitive interviewing was incorporated into the original development of ITEM [7], limited evidence is available regarding how its established English-language questions are interpreted by learners whose first language is not English when the instrument is administered without translation. Accordingly, this study aimed to evaluate the clarity, comprehensibility, and response processes of ITEM among native Arabic-speaking medical students enrolled in an English-medium medical program in the United Arab Emirates after VR-based clinical skills activities. Using cognitive interviewing with TA and VP techniques, we examined how participants interpreted individual ITEM questions and identified linguistic, referential, and contextual sources of comprehension difficulty. The findings were used to inform proposed linguistic or contextual

XSL•FO
RenderX

clarifications where appropriate, while retaining the original wording when a modification could not be made confidently without potentially altering the intended meaning.

Methods

Study Design

This qualitative study used cognitive interviewing to evaluate the clarity, comprehensibility, and response processes of ITEM among native Arabic-speaking medical students enrolled in an English-medium medical program. We selected cognitive interviewing as an evidence-based questionnaire evaluation method to identify potential sources of misunderstanding and response error prior to broader administration of the instrument. Reporting was guided by the Cognitive Interviewing Reporting Framework (CIRF) [28] (Checklist 1).

Setting

We conducted the study in September 2025 at the Simulation Center (iSTAR), College of Medicine and Health Sciences, United Arab Emirates University (UAEU), in the United Arab Emirates. The undergraduate medical curriculum is delivered in English, while Arabic is the first language of the enrolled United Arab Emirates–national student population. The UAEU medical program is a 6-year curriculum with premedical, preclinical, and clinical phases; we recruited participants from years 3 and 4 because these preclinical years represent a transition into structured clinical skills training supported by simulation-based education.

Ethical Considerations

The Social Sciences Ethics Committee at the UAEU approved the study (ERSC_2025_8044). Participation was voluntary. Participants received no financial compensation or other incentives for participation. We obtained written informed consent electronically prior to participation using REDCap, a secure, web-based data capture platform designed to support research data collection and management [35,36]. Participants consented to audiovisual recording for transcription and analysis and were informed that they could decline to answer any question or withdraw at any time without penalty. We stored data securely with access restricted to the research team. We removed identifying information during transcript processing and analyzed anonymized transcripts to protect participant privacy and confidentiality.

Instrument

ITEM is a 40-item multidomain questionnaire designed to capture user experience in immersive TEL environments [7] (Multimedia Appendix 1). Participants completed the original English-language ITEM wording during the cognitive interviews. The questionnaire integrates 5 domains derived from established frameworks and validated measures: AIEQ [13,14] measures immersion and system fidelity, AIMI [14,15] assesses learner motivation and perceived educational value, NASA-TLX [11] evaluates cognitive and physical load associated with task performance, SUS [8] assesses confidence and perceived accessibility of the technology, and PEARLS [16] explores reflective learning and debriefing processes.

Proposed wording decisions arising from the cognitive interviews are presented separately and were not administered in a second round of cognitive testing.

Participants and Recruitment

Participants were undergraduate medical students in years 3 and 4 of the UAEU undergraduate medical program who had completed the preliminary VR clinical skills activity. Twenty students participated in the preliminary activity and were informed verbally about the cognitive interview study; all were subsequently invited by email. Ten students volunteered to participate. Interviews were intended to continue until no new item-level comprehension issues emerged rather than to achieve a predetermined fixed sample size. Saturation was identified after 8 interviews through ongoing review and team discussion, and 2 additional interviews were conducted to confirm that no new relevant item-level issues emerged. All 10 participants were United Arab Emirates nationals and native Arabic speakers. English-language proficiency and prior language of schooling were not formally assessed or collected.

VR Experience (Context for Questionnaire Evaluation)

Prior to the cognitive interviews, participants completed a curriculum-aligned VR clinical skills session. Participants also completed a presession demographic and XR-background questionnaire in REDCap (Multimedia Appendix 2). Facilitators first provided a technical orientation explaining the VR equipment, headset controls, and activity sequence. All participants used the same VR platform and followed the same instructional process, although the clinical module differed according to year-level curricular content. Year 3 students were randomly allocated to 1 of 2 modules (vital signs, n=3; and intravenous cannula insertion, n=1), and year 4 students completed a nasogastric tube insertion module (n=6). Each student completed the guided procedure individually. After the VR activity, participants provided verbal reflections organized broadly around strengths, weaknesses, opportunities, and threats. This activity was intended to elicit general feedback on usability, engagement, and instructional value and did not constitute a formal structured simulation debrief or use the PEARLS framework. The study did not compare cognitive interview findings across VR modules. Figure 1 provides an overview of the study process.

XSL•FO
RenderX

Figure 1. Workflow of the structured virtual reality (VR) clinical skills session and subsequent cognitive interview evaluation of the Immersive Technology Evaluation Measure (ITEM). IV: intravenous; SWOT: strengths, weaknesses, opportunities, and threats.

XSL•FO
RenderX

Cognitive Interview Procedure

Following the VR learning sessions, year 3 (n=4) and year 4 (n=6) students participated in cognitive interviews. Two faculty members (FMA and AK) conducted the interviews online via Microsoft Teams (version 25240.1603.3956.3103). Both interviewers were experienced in medical education and familiar with the use of cognitive interviewing procedures. Neither interviewer participated in the students' VR teaching sessions or assessment. With participant permission, we recorded interviews and transcribed them verbatim. Interviews were scheduled for 30 minutes and lasted 20 to 45 minutes depending on participant engagement. Participants were asked to keep their cameras on to support observation of nonverbal cues such as facial expressions or hesitation. Interviewers used a combination of concurrent and retrospective cognitive interviewing techniques [22]: TA (concurrent) [26], whereby participants were prompted to think aloud while reading each survey item aloud, verbalizing their thoughts, interpretations, and reasoning as they answered the questions; VP (concurrent) [27], whereby interviewers used a combination of prescripted and spontaneous probes to clarify interpretation, meaning, and response selection; and retrospective debriefing [22], whereby, after completing the questionnaire, participants were asked to reflect on overall clarity, confusing terms, and any culturally unfamiliar wording.

Interviewers used a combination of prescripted probes and spontaneous follow-up probes when participants hesitated, requested clarification, or provided an interpretation requiring further exploration.

Data Management and Item-Level Analysis

We transcribed recordings using Otter AI (version 3.93.0; Otter.ai, Inc) and manually checked each transcript against the recording for accuracy. Identifying information was removed prior to analysis. Two researchers (AP and MGA) independently reviewed all transcripts, recordings, and interviewer notes. Participant comments were organized in Microsoft Excel by ITEM question, together with observations of hesitation, requests for clarification, and alternative interpretations.

The analysis was descriptive and item-focused rather than a formal thematic analysis. For each ITEM question, the researchers first summarized the specific comprehension or response-process difficulties raised by participants. These item-level observations were then compared across questions to identify recurring types of difficulty. Recurring problems were organized into descriptive categories comprising ambiguous referents, unfamiliar or abstract terminology, negative wording, temporal ambiguity, and contextual relevance. These categories were used to organize and summarize the item-level findings; they were not intended to represent a formal qualitative coding framework or thematic analysis. An item was considered problematic when participants expressed confusion, requested clarification, hesitated because of uncertainty about meaning, or provided differing interpretations that indicated potential ambiguity. Item-level findings, participant-reported problems, modification decisions, and the rationale for retaining or revising each question were documented in an item-level audit (Multimedia Appendix 3).

AP and MGA compared their item-level assessments and assignment of descriptive problem categories after independent review and did not identify substantive disagreements. The wider research team subsequently reviewed the participant comments and proposed item-level decisions. Any differences during team discussion were resolved through consensus. No specialized qualitative analysis software or interrater agreement statistic was used.

Proposed wording changes were kept deliberately conservative. Where participant feedback indicated a comprehension problem that could be clarified without intentionally changing the item meaning, the team considered a brief linguistic or contextual clarification. Where the intended scope of an item could not be established confidently from the original wording or source measure, the original wording was retained rather than imposing a new interpretation. Multimedia Appendix 4 presents the proposed wording and presentation modifications. A native English-speaking language expert from the United States reviewed the proposed wording for linguistic clarity. The proposed wording was not subjected to a second round of cognitive interviews in this study.

Rigor and Reflexivity

To support consistency and transparency, 2 researchers independently reviewed the interview transcripts, recordings, and interviewer notes, and item-level findings were subsequently discussed with the wider research team. Neither cognitive interviewer participated in the students' VR teaching sessions or assessment. Because several members of the research team were involved in immersive TEL implementation, the team considered how their familiarity with VR and ITEM could influence interpretation of participant comments. Proposed item-level decisions were therefore discussed collectively and grounded in participants' statements, observed comprehension difficulties, and interviewer notes, with any differences in interpretation resolved through consensus.

Results

Participants

Ten undergraduate medical students participated in the cognitive interviews (n=4 in year 3 and n=6 in year 4). The mean age was 20 (SD 0.5; range 20 - 21) years for year 3 students and 21 (SD 0.9; range 20 - 22) years for year 4 students. Nine participants were female and 1 was male. All participants were United Arab Emirates nationals and native Arabic speakers enrolled in an English-medium medical program. All participants completed a VR-based clinical skills session prior to the interviews.

Cognitive Interview Duration and Item-Level Saturation

Interviews were scheduled for 30 minutes and lasted 20 to 45 minutes, with a mean duration of approximately 33 minutes. Item-level saturation was reached after 8 interviews, with no new comprehension problems identified. Two additional interviews were conducted to confirm saturation, and no new relevant item-level issues emerged.

XSL•FO
RenderX

Overview of Item-Level Findings

Across the 40 questions of ITEM, participants identified 47.5% (19/40) as presenting a comprehension or response-process problem. Problems occurred across all 5 domains: immersion (6/9, 66.7%), usability (6/10, 60%), debriefing (4/5, 80%), motivation (2/10, 20%), and cognitive load (1/6, 16.7%). The most common difficulties involved ambiguous or nonspecific referents, unfamiliar or abstract terminology, negative wording, temporal ambiguity, and contextual relevance.

Of the 19 participant-identified problematic questions, 11 (57.9%) received proposed wording or contextual revisions, 2 (10.5%) received presentation-only modifications without alteration of the original wording, and 6 (31.6%) were retained in their original wording. An additional 6 questions that were not independently identified as problematic received limited contextual or terminology-standardizing changes for consistency. The complete item-level findings and decision rationale are provided in Multimedia Appendix 3. Illustrative examples of revisions made following participant feedback are presented in Table 1.

Table . Illustrative cognitive interview findings and proposed item-level decisions for the Immersive Technology Evaluation Measure (ITEM)[a].

Original question (Q)	Problem identified	Representative participant comment	Item-level decision
Section 1: immersion			
Q1: I was interested in seeing how the activity would progress	Ambiguous reference to "activity"	"Which activity are they talking about here?" "What activity do you mean? So, the VR in general, like the whole study, just my experience with the VR? So, I think specifying the activity would be good."	Proposed contextual clarification: "I was interested in seeing how the activity (simulation session using VR) would progress."
Q6: At the time, the activity was my only concern	"Concern" interpreted as worry rather than focus	"Why concern? It was not a concern. It was something nice."	Proposed plain-language clarification: "At the time, the activity was my only concern (focus)."
Q7: I wanted to learn more about the outcome following the activity	Unclear "outcome" referent.	"Outcome of what?" "Like, would it be applied or introduced in our university? Is that what they mean?"	Original wording retained. Participants interpreted "outcome" differently, and the intended referent could not be established confidently without imposing a new interpretation.
Section 2: motivation			
Q13: This activity did not hold my attention at all	Negative wording and response directionality	"What do they mean by this question?" "Okay? No, I was like immersed in it, but I am still confused actually with the DID NOT HOLD."	No wording revision. "DID NOT" was capitalized as a typographic emphasis to draw attention to the negative construction: "This activity DID NOT hold my attention at all."
Section 3: cognitive load			
Q20: How mentally demanding was the task?	Misinterpretation of the phrase "mentally demanding"	"How much I had focused into the task, or did I understand it wrong?" "When I read mentally demanding, I thought about mental health and not in a thinking way."	Proposed plain-language clarification: "How mentally demanding (needs a lot of focus) was the task?"
Section 4: usability			
Q27: I found the technology unnecessarily complex	Difficulty interpreting "unnecessarily complex"	"So, it is not complex, or it is complex?" "This word. Do they mean I found the technology complex?"	Proposed clarification and contextual specification: "I found the VR technology unnecessarily complex (more complicated than necessary)."
Q29: I think that I would need the support of a technical person to be able to use this technology	Unclear temporal scope (initial vs ongoing support)	"Do you mean every time I use it, or only the first time?" "I think the first time, I think the first time we need someone, but after that, I don't think we need someone. So, I don't know what to put my answer."	Original wording retained. The source item does not specify a temporal frame; therefore, no clarification was introduced.

[a]Examples are illustrative; complete item-level findings and decision rationales for all 40 ITEM questions are provided in Multimedia Appendix 3. Proposed wording and presentation modifications were not cognitively retested in this study.

XSL•FO
RenderX

Patterns of Comprehension Difficulty

A recurring issue across domains was the use of nonspecific references such as "activity" and "technology." Participants variously interpreted these as referring to the VR software, the clinical procedure, or the overall simulation session. Where appropriate, proposed contextual clarifications specified the VR activity or VR technology to reduce uncertainty regarding the referent.

Participants also reported difficulty with abstract or unfamiliar terms. For example, some were uncertain about the meaning of "immersion," and one participant interpreted "mentally demanding" as referring to mental health rather than the amount of cognitive effort required. Negatively worded statements also created response uncertainty. In response, typographic emphasis was proposed for "DID NOT" in question 13 without altering the wording of the item.

Not all identified problems resulted in wording changes. For question 7, participants differed in their interpretation of "outcome," and for question 29, they questioned whether technical support referred to initial or ongoing use. Because these meanings could not be established confidently from the original wording, the original questions were retained rather than introducing a clarification that could alter their intended scope.

Debriefing Domain

Participants identified comprehension difficulties in 4 of the 5 debriefing-domain questions. However, participants had not undergone a formal structured postactivity debrief; the brief strengths, weaknesses, opportunities, and threats–style reflection following the VR activity was not based on the PEARLS framework. Consequently, these difficulties could not be attributed confidently to questionnaire wording alone and may have reflected limited contextual applicability. The original wording of all debriefing-domain questions was therefore retained.

General Feedback

Across interviews, participants described some questions as broad, formal, or unfamiliar and indicated that brief contextual cues or plain-language explanations could support comprehension of selected terms. Most identified issues concerned wording or contextual interpretation rather than the overall structure of ITEM. No participant feedback indicated a need to remove an ITEM domain or change the questionnaire's overall multidomain structure.

Discussion

Principal Findings

This study used cognitive interviewing to evaluate the clarity, comprehensibility, and response processes of the ITEM among native Arabic-speaking medical students enrolled in an English-medium medical program. Participants identified comprehension or response-process difficulties across all 5 ITEM domains, including ambiguous referents, unfamiliar or abstract terminology, negative wording, temporal ambiguity, and contextual relevance.

The findings demonstrated that response-process difficulties can emerge when an established English-language instrument is used in a different linguistic and educational context, even when learners study in English. Importantly, cognitive interviewing informed not only where clarification might be useful but also where retaining the original wording was more appropriate because the intended meaning could not be established confidently. The resulting item-level decisions therefore included proposed linguistic clarifications, contextual specifications, presentation changes, and deliberate retention of original wording. These findings provide response-process evidence for ITEM in this context but do not demonstrate that the proposed revisions themselves improved interpretability.

Comparison With Prior Work

Our findings are consistent with cognitive interviewing literature showing that respondents may interpret apparently straightforward questionnaire items differently according to wording, context, and linguistic experience [22,23,28,29]. Cross-cultural cognitive interviewing research similarly emphasizes that equivalent interpretation should not be assumed solely because respondents can complete an instrument in the language in which it was developed [29,30].

Research on survey response across linguistic groups also supports consideration of language proficiency and linguistic complexity when questionnaires are administered across languages [31,32]. In an Arabic-speaking context, differences have been reported when the same health-related questionnaire was administered in Arabic and English [34]. Although these studies differ from the present study in design and population, they support examining response processes when questionnaire language differs from respondents' first language. The present study extends this work by examining an established English-language immersive technology measure without translating it, reflecting how ITEM would be used in an English-medium medical education setting in the United Arab Emirates [33].

Domain-Specific and Contextual Findings

Comprehension problems occurred across all ITEM domains but were most frequent in the immersion and usability domains. Several of these questions contained abstract terminology or broad referents such as "activity," "technology," and "outcome," which participants interpreted in different ways. Where clarification could be introduced without intentionally changing the item scope, brief contextual or plain-language explanations were proposed.

Findings from the debriefing domain required separate interpretation. Participants identified difficulties with 4 of the 5 debriefing questions; however, they had not undergone a formal structured postactivity debrief. The brief strengths, weaknesses, opportunities, and threats–style reflection was not a PEARLS-based debriefing process. Because PEARLS was developed specifically for structured simulation debriefing [7,16], these difficulties may have reflected limited contextual applicability rather than wording alone. The original wording of all debriefing questions was therefore retained.

XSL•FO
RenderX

The cognitive interview findings also highlighted the importance of avoiding overcorrection. For question 7, the meaning of "outcome" remained uncertain, and for question 29, the original item did not specify whether technical support referred to initial or ongoing use. Rather than introducing a meaning not clearly established in the source items, we retained the original wording. This conservative approach recognizes that cognitive interviewing can identify ambiguity without necessarily providing sufficient evidence to redefine an item [28,29], and even apparently minor questionnaire adaptations should be considered carefully because they may affect item functioning [37].

Implications for Medical Education and Instrument Use

These findings support cognitive interviewing as a practical method for examining response-process evidence when established educational measures are applied beyond their original development context [22,23,28,29]. For ITEM, this study provides an item-level account of where participants encountered uncertainty and documents the rationale for proposed linguistic, contextual, or presentation modifications. Because the proposed wording was not cognitively retested, these modifications should be considered provisional rather than validated improvements.

The addition of "VR" to selected proposed clarifications does not make ITEM a VR-only instrument. ITEM was developed for immersive TEL and was originally evaluated in both VR and AR contexts [7]. In this study, "VR" was added only where participants found the generic term "technology" insufficiently specific. In other immersive settings, the original technology-neutral wording could be retained or the relevant modality, such as AR or MR, could be specified where needed.

More broadly, studying medicine in English does not necessarily mean that all English-language questionnaire expressions will be interpreted uniformly. Response-process evaluation may therefore complement psychometric testing when established instruments are used in different linguistic or educational settings.

Limitations

This study has several limitations. First, the proposed wording was not cognitively retested; therefore, the study identifies response-process difficulties and proposes potential refinements but cannot establish that the revisions resolved the identified comprehension problems. Second, English-language proficiency and prior language of schooling were not formally assessed, so the findings cannot determine whether the observed comprehension difficulties varied according to individual language proficiency or educational language background. Third, only 1 male student participated, which limited our ability to examine whether interpretation differed by sex. The sample was also drawn from a single institution. These characteristics constrain subgroup comparisons and broader generalizability; however, the purpose of cognitive interviewing was to identify and characterize item-level comprehension problems rather than estimate their prevalence in the wider student population. Item-level saturation was reached after 8 interviews and confirmed with 2 additional interviews, supporting the adequacy of the sample for the study's primary cognitive interviewing objective. Fourth, participants did not undergo a formal structured debrief; therefore, difficulties with the PEARLS-derived items may reflect limited contextual applicability rather than questionnaire wording alone. Finally, interviews were conducted online via Microsoft Teams. In some cases, screen sharing disabled participants' cameras, a common constraint when using tablets or iPads, which limited observation of nonverbal cues such as facial expressions and hesitations. Online interviewing may also reduce opportunities for rapport building compared with in-person interviews, which could have affected the depth of spontaneous verbalizations during the TA process.

Future Directions

Future research should cognitively retest the proposed item-level modifications to determine whether they resolve the identified problems without introducing new interpretations. Subsequent studies should evaluate reliability, construct validity, and measurement performance in larger and more diverse samples, including learners with different linguistic backgrounds and levels of English proficiency. Further investigations across VR, AR, and MR settings and in activities involving formal structured debriefing would also help determine when modality-specific contextual wording is appropriate.

Conclusions

Cognitive interviewing identified linguistic, referential, and contextual response-process difficulties in several ITEM questions among native Arabic-speaking medical students enrolled in an English-medium medical program. The findings informed proposed item-level clarifications while also identifying questions for which retaining the original wording was more appropriate because further modification could impose an interpretation not clearly established in the source item.

These findings provide additional response-process evidence for ITEM in a different linguistic and educational context but do not establish that the proposed wording changes improve interpretability. Further cognitive testing and subsequent psychometric evaluation are needed before the proposed refinements can be considered confirmed.

Acknowledgments

The authors gratefully acknowledge Dr Chris Jacobs for granting permission to use the Immersive Technology Evaluation Measure (ITEM) in this study.

The authors acknowledge support from the Princess Nourah bint Abdulrahman University Researchers Supporting Project (PNURSP2026R290), Princess Nourah bint Abdulrahman University, Riyadh, Saudi Arabia.

XSL•FO
RenderX

The authors also acknowledge MediSim VR for providing access to the VR software used in this study and Dr Pradeesh Sathyan, Senior Consultant for Medical Technology, Arrownex LLC, for technical and content-related support.

The authors acknowledge Ms Jane Koester (College of Medicine and Health Sciences, United Arab Emirates University [UAEU]) for support with linguistic clarification and contextual refinement of selected questionnaire items to enhance clarity for the local context. The authors acknowledge the contributions of the medical students Latifa Mohammed Alderei, Aryam Muhsen Albreiki, Raghad Salem Alharthi, and Shaikha Ahmed Alzaabi (College of Medicine and Health Sciences, UAEU) for assistance with participant recruitment. The authors are especially appreciative of all the medical students who volunteered their time to take part in this study.

The authors acknowledge the use of OpenAI's ChatGPT (GPT 5.2) to support language editing and improve clarity and readability of the manuscript. All content was critically reviewed, verified, and approved by the authors. No generative AI tools were used to generate, analyze, or interpret study data.

Funding

This research was supported by a research grant from the United Arab Emirates University (grant 12M225).

Data Availability

The qualitative interview transcripts are not publicly available because of the sensitive nature of the interview data and the potential risk of participant identification. Deidentified excerpts supporting the findings are included in the manuscript, and complete item-level findings and decision rationales are provided in Multimedia Appendix 4. Additional information may be made available by the corresponding author (TMA) upon reasonable request and subject to institutional ethical approval.

Authors' Contributions

Conceptualization: ASA
Data curation: FMA, AK
Formal analysis: AP, MGA
Investigation: FMA, AK
Methodology: ASA, FAA, TMA
Validation: ASA, FMA, AP, MGA
Writing—original draft: ASA, AP
Writing—review and editing: ASA, AP, FMA, MGA, AK, TMA, FAA
All authors reviewed and approved the final manuscript and agree to be accountable for all aspects of the work.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Immersive Technology Evaluation Measure questionnaire administered during cognitive interviews (original wording).
[DOCX File, 26 KB - mededu_v12i1e95904_app1.docx]

Multimedia Appendix 2
Presession participant questionnaire capturing demographic and background information.
[DOCX File, 18 KB - mededu_v12i1e95904_app2.docx]

Multimedia Appendix 3
Immersive Technology Evaluation Measure item-level cognitive interview audit.
[XLSX File, 24 KB - mededu_v12i1e95904_app3.xlsx]

Multimedia Appendix 4
Proposed modified Immersive Technology Evaluation Measure following cognitive interview review.
[DOCX File, 27 KB - mededu_v12i1e95904_app4.docx]

Checklist 1
Cognitive Interviewing Reporting Framework (CIRF) checklist.
[DOCX File, 21 KB - mededu_v12i1e95904_app5.docx]

XSL•FO
RenderX

References

1. Mergen M, Graf N, Meyerheim M. Reviewing the current state of virtual reality integration in medical education - a scoping review. BMC Med Educ 2024 Jul 23;24(1):788. [doi: 10.1186/s12909-024-05777-5] [Medline: 39044186]
2. Tene T, Vique López DF, Valverde Aguirre PE, Orna Puente LM, Vacacela Gomez C. Virtual reality and augmented reality in medical education: an umbrella review. Front Digit Health 2024;6:1365345. [doi: 10.3389/fdgth.2024.1365345] [Medline: 38550715]
3. Sung H, Kim M, Park J, Shin N, Han Y. Effectiveness of virtual reality in healthcare education: systematic review and meta-analysis. Sustainability 2024;16(19):8520. [doi: 10.3390/su16198520]
4. Rojas-Sánchez MA, Palos-Sánchez PR, Folgado-Fernández JA. Systematic literature review and bibliometric analysis on virtual reality and education. Educ Inf Technol 2023 Jan;28(1):155-192. [doi: 10.1007/s10639-022-11167-5]
5. Forrest K, McKimm J. Healthcare Simulation at a Glance, 1st edition: John Wiley & Sons; 2019.
6. Ryan GV, Callaghan S, Rafferty A, Higgins MF, Mangina E, McAuliffe F. Learning outcomes of immersive technologies in health care student education: systematic review of the literature. J Med Internet Res 2022 Feb 1;24(2):e30082. [doi: 10.2196/30082] [Medline: 35103607]
7. Jacobs C, Wheeler J, Williams M, Joiner R. Cognitive interviewing as a method to inform questionnaire design and validity - Immersive Technology Evaluation Measure (ITEM) for healthcare education. Comput Educ X Real 2023;2:100027. [doi: 10.1016/j.cexr.2023.100027]
8. Brooke J. SUS: a 'quick and dirty' usability scale. In: Jordan PW, Thomas B, McClelland IL, Weerdmeester B, editors. Usability Evaluation in Industry: CRC Press; 1996. [doi: 10.1201/9781498710411]
9. Davis FD. Perceived usefulness, perceived ease of use, and user acceptance of information technology. MIS Q 1989;13(3):319-340. [doi: 10.2307/249008]
10. Schubert T, Friedmann F, Regenbrecht H. The experience of presence: factor analytic insights. Presence Teleoperators Virtual Environ 2001 Jun;10(3):266-281. [doi: 10.1162/105474601300343603]
11. Hart SG. NASA-task load index (NASA-TLX); 20 years later. Proc Hum Factors Ergon Soc Annu Meet 2006 Oct;50(9):904-908. [doi: 10.1177/154193120605000909]
12. McAuley E, Duncan T, Tammen VV. Psychometric properties of the intrinsic motivation inventory in a competitive sport setting: a confirmatory factor analysis. Res Q Exerc Sport 1989 Mar;60(1):48-58. [doi: 10.1080/02701367.1989.10607413] [Medline: 2489825]
13. Jennett C, Cox AL, Cairns P, et al. Measuring and defining the experience of immersion in games. Int J Hum Comput Stud 2008 Sep;66(9):641-661. [doi: 10.1016/j.ijhcs.2008.04.004]
14. Jacobs C, Rigby JM. Developing measures of immersion and motivation for learning technologies in healthcare simulation: a pilot study. J Adv Med Educ Prof 2022 Jul;10(3):163-171. [doi: 10.30476/JAMP.2022.95226.1632] [Medline: 35910517]
15. Ryan RM. Control and information in the intrapersonal sphere: an extension of cognitive evaluation theory. J Pers Soc Psychol 1982;43(3):450-461. [doi: 10.1037/0022-3514.43.3.450]
16. Eppich W, Cheng A. Promoting Excellence and Reflective Learning in Simulation (PEARLS): development and rationale for a blended approach to health care simulation debriefing. Simul Healthc 2015 Apr;10(2):106-115. [doi: 10.1097/SIH.0000000000000072] [Medline: 25710312]
17. Roff S, McAleer S, Harden RM, et al. Development and validation of the Dundee Ready Education Environment Measure (DREEM). Med Teach 1997;19(4):295-299. [doi: 10.3109/01421599709034208]
18. Marshall RE. Measuring the medical school learning environment. J Med Educ 1978 Feb;53(2):98-104. [doi: 10.1097/00001888-197802000-00003] [Medline: 633337]
19. Wilson KL, Lizzio A, Ramsden P. The development, validation and application of the Course Experience Questionnaire. Stud High Educ 1997;22(1):33-53. [doi: 10.1080/03075079712331381121]
20. AlHaqwi AI, Kuntze J, van der Molen HT. Development of the Clinical Learning Evaluation Questionnaire for undergraduate clinical education: factor structure, validity, and reliability study. BMC Med Educ 2014 Mar 4;14:44. [doi: 10.1186/1472-6920-14-44] [Medline: 24592913]
21. Tourangeau R, Rips LJ, Rasinski K. The Psychology of Survey Response: Cambridge University Press; 2000. [doi: 10.1017/CBO9780511819322]
22. Willis GB. Cognitive Interviewing: A Tool for Improving Questionnaire Design: SAGE Publications; 2004.
23. Willis GB, Artino ARJ. What do our respondents think we're asking? Using cognitive interviewing to improve medical education surveys. J Grad Med Educ 2013 Sep;5(3):353-356. [doi: 10.4300/JGME-D-13-00154.1] [Medline: 24404294]
24. Tourangeau R. Cognitive sciences and survey methods. In: Jabine TB, Straf ML, Tanur JM, Tourangeau R, editors. Cognitive Aspects of Survey Methodology: Building a Bridge between Disciplines: National Academies Press; 1984:73-100.
25. Beatty PC, Willis GB. Research synthesis: the practice of cognitive interviewing. Public Opin Q 2007;71(2):287-311. [doi: 10.1093/poq/nfm006]
26. Ericsson KA, Simon HA. Verbal reports as data. Psychol Rev 1980;87(3):215-251. [doi: 10.1037/0033-295X.87.3.215]
27. Forsyth BH, Lessler JT. Cognitive laboratory methods: a taxonomy. In: Biemer PP, Groves RM, Lyberg LE, Mathiowetz NA, Sudman S, editors. Measurement Errors in Surveys: John Wiley & Sons; 2004:393-418. [doi: 10.1002/9781118150382]

XSL•FO
RenderX

28. Boeije H, Willis G. The Cognitive Interviewing Reporting Framework (CIRF): towards the harmonization of cognitive testing reports. Methodology 2013;9(3):87-95. [doi: 10.1027/1614-2241/a000075]
29. Willis GB, Miller K. Cross-cultural cognitive interviewing: seeking comparability and enhancing understanding. Field Methods 2011;23(4):331-341. [doi: 10.1177/1525822X11416092]
30. Schildmann EK, Groeneveld EI, Denzel J, et al. Discovering the hidden benefits of cognitive interviewing in two languages: the first phase of a validation study of the Integrated Palliative care Outcome Scale. Palliat Med 2016 Jun;30(6):599-610. [doi: 10.1177/0269216315608348] [Medline: 26415736]
31. Park H, Sha MM, Willis G. Influence of English-language proficiency on the cognitive processing of survey questions. Field Methods 2016;28(4):415-430. [doi: 10.1177/1525822X16630262]
32. Wenz A, Al Baghal T, Gaia A. Language proficiency among respondents: implications for data quality in a longitudinal face-to-face survey. J Surv Stat Methodol 2021;9(1):73-93. [doi: 10.1093/jssam/smz045]
33. Ismaiel S, AlGhafari D, Ibrahim H. Promoting physician-patient language concordance in undergraduate medical education: a peer assisted learning approach. BMC Med Educ 2023 Jan 3;23(1):1. [doi: 10.1186/s12909-022-03986-4] [Medline: 36593450]
34. Gazzaz ZJ, Baig M, Albarakati M, Alfalig HA, Jameel T. Language barriers in understanding healthcare information: Arabic-speaking students' comprehension of diabetic questionnaires in Arabic and English languages. Cureus 2023 Oct;15(10):e46777. [doi: 10.7759/cureus.46777] [Medline: 37954810]
35. Harris PA, Taylor R, Thielke R, Payne J, Gonzalez N, Conde JG. Research electronic data capture (REDCap)--a metadata-driven methodology and workflow process for providing translational research informatics support. J Biomed Inform 2009 Apr;42(2):377-381. [doi: 10.1016/j.jbi.2008.08.010] [Medline: 18929686]
36. Harris PA, Taylor R, Minor BL, et al. The REDCap consortium: building an international community of software platform partners. J Biomed Inform 2019 Jul;95:103208. [doi: 10.1016/j.jbi.2019.103208] [Medline: 31078660]
37. Sousa VE, Matson J, Dunn Lopez K. Questionnaire adapting: little changes mean a lot. West J Nurs Res 2017 Sep;39(9):1289-1300. [doi: 10.1177/0193945916678212] [Medline: 28322671]

Abbreviations

AIEQ: Adapted Immersion Experience Questionnaire
AIMI: abridged intrinsic motivation inventory
AR: augmented reality
ITEM: Immersive Technology Evaluation Measure
MR: mixed reality
NASA-TLX: NASA Task Load Index
PEARLS: Prompts for Engaging and Reflective Learning in Simulation
SUS: System Usability Scale
TA: think-aloud
TEL: technology-enhanced learning
UAEU: United Arab Emirates University
VP: verbal probing
VR: virtual reality
XR: extended reality

Edited by A Stone; submitted 23.Mar.2026; peer-reviewed by A Clark, Y Yang; revised version received 26.Aug.2026; accepted 26.Aug.2026; published 28.Sep.2026.

Please cite as:
Alblooshi AS, Almarzooqi FM, Almansoori TM, Punn A, Alameen MG, Kieu A, AlRadini FA
Refining the Immersive Technology Evaluation Measure for Virtual Reality Clinical Skills Among Native Arabic-Speaking Medical Students: Qualitative Study
JMIR Med Educ 2026;12:e95904
URL: https://mededu.jmir.org/2026/1/e95904
doi:10.2196/95904

© Afaf Sulaiman Alblooshi, Falah Mohammed Almarzooqi, Taleb Mohamed Almansoori, Anmol Punn, Marwa Gaffar Alameen, Alexander Kieu, Faten Abdullah AlRadini. Originally published in JMIR Medical Education (https://mededu.jmir.org), 28.Sep.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium,

XSL•FO
RenderX

provided the original work, first published in JMIR Medical Education, is properly cited. The complete bibliographic information, a link to the original publication on https://mededu.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Virtual Reality–Enhanced Training for Trauma-Informed Care Among Residential and Child Mental Health Professionals: Pre-Post Evaluation Study

Kostadin Kostadinov[1], MPH, MD, PhD; John Goodwin[2], BSc, BA, PhD; Gunter Groen[3], PhD; Astrid Jörns-Presentati[3], BA, MA, PhD; Sinéad Heffernan[4], DN; Áine O'Donovan[2], BNS, MSc, PhD; Maria O'Malley[2], BSc, MSc, DN; Ryan Biskupovic Goulding[2], BSc; James O'Mahony[2], BSc, MSc, MBA, PhD; Stephanie Allen[5], BSc, MSc; Margaret Curtin[2], MSc, PhD; Satu Haapalainen[6], RN, MNSc; Jami Aho[7], MEng; Gergana Petrova[8], MPH, PhD; Svetla Ivanova[8]; Valentina Lalova[8], PhD; Joonas Korhonen[6], RN, MNSc, PhD

[1]Department of Social Medicine and Public Health, Faculty of Public Health, Medical University of Plovdiv, Bul. Vasil Aprilov 15A, Plovdiv, Bulgaria
[2]School of Nursing & Midwifery, University College Cork, Cork, Ireland
[3]Department of Social Work, Hamburg University of Applied Sciences (HAW Hamburg), Hamburg, Germany
[4]West Cork Women Against Violence Project CLG, West Cork Beacon, Cork, Ireland
[5]School of Nursing and Midwifery, Trinity College Dublin, Dublin, Ireland
[6]Health and Well-Being, Turku University of Applied Science, Turku, Finland
[7]Turku University of Applied Sciences, Turku, Finland
[8]Department of Nursing Care, Faculty of Public Health, Medical University of Plovdiv, Plovdiv, Bulgaria

Corresponding Author:
Kostadin Kostadinov, MPH, MD, PhD
Department of Social Medicine and Public Health, Faculty of Public Health, Medical University of Plovdiv, Bul. Vasil Aprilov 15A, Plovdiv, Bulgaria

Abstract

Background: Trauma-informed care (TIC) is a framework that embeds safety, trust, choice, collaboration, and empowerment into practice. Although training improves TIC attitudes, implementation in European residential child and adolescent settings remains inconsistent, and immersive technologies such as virtual reality (VR) are underevaluated for this purpose.

Objective: This study aims to evaluate whether a VR-enhanced training program improves TIC attitudes among staff and trainees in residential child and adolescent care across European sites and to examine site-level heterogeneity.

Methods: We conducted a multisite pre-post evaluation within the European Union–co-funded Safe4Child project. Participants completed a standardized online TIC module followed by a mentor-facilitated VR simulation. Attitudes were measured using the ARTIC-10 (10-item version of the Attitudes Related to Trauma-Informed Care; 7-point) scale immediately before the online module and immediately after the VR simulation; the design, therefore, evaluates the combined program rather than VR in isolation. Analyses included Wilcoxon signed-rank tests, multivariable regression, and Bayesian models with skeptical priors.

Results: Among 79 matched participants from Bulgaria, Finland, and Germany, mean (SD) ARTIC-10 scores increased from 5.33 (1.05) to 5.57 (1.20; $P<.001$; Cohen $d=0.22$). Effects varied by site. Germany showed significant improvement (Cohen $d=0.56$), whereas Bulgaria and Finland did not reach significance. Bayesian estimation yielded a mean change of 0.24 points (95% CI 0.11 - 0.37) with $P(\Delta>0) \approx 1.00$. Prior trauma-related training showed a probable but uncertain additional benefit.

Conclusions: The VR-enhanced TIC program produced a small overall improvement in attitudes, with a moderate effect in Germany, suggesting that immersive training can strengthen TIC learning when aligned with local contexts. The absence of a control group precludes attributing effects specifically to VR. Larger comparative trials are needed to determine whether VR confers advantages over conventional training approaches.

(JMIR Med Educ 2026;12:e86543) doi:10.2196/86543

KEYWORDS

trauma-informed care; virtual reality; professional education; child welfare; residential care

XSL•FO
RenderX

Introduction

Background

Trauma-informed care (TIC) is recognized as an organizational and clinical framework that acknowledges the widespread prevalence and significant impact of trauma, especially among vulnerable groups such as children and adolescents in residential care settings [1]. Core principles including safety, trustworthiness, choice, collaboration, and empowerment are integrated into routine practices through this approach, which acknowledges the link between intergenerational and childhood trauma to wellness and therefore represents a fundamental shift from inquiring, "What is wrong with you?" to "What happened to you?" [2]. Within residential programs, TIC has been associated with notable outcomes, including reductions in behavioral incidents, enhanced placement stability, and improved relationships between carers and children [3]. Furthermore, the adoption of TIC principles has been linked to decreased reliance on restrictive practices, such as seclusion and restraint, which are often used in managing behaviors that challenge but can exacerbate trauma responses if not handled sensitively [4]. TIC not only mitigates immediate risks but also fosters long-term resilience by addressing underlying trauma histories, potentially reducing the cycle of retraumatization in residential environments [5].

Challenges in Implementing TIC

Despite its established benefits, the implementation of TIC is hindered by various barriers. Organizational resistance to cultural shifts, insufficient staff training and awareness, burnout among personnel, and systemic limitations such as funding shortages or rigid policies have been identified in systematic reviews as key obstacles [6,7]. These challenges often create a gap between what policies promise and what happens in practice. To close this gap, targeted strategies are needed. In Europe, there are additional barriers, such as differences between national health care systems, language and cultural diversity, and varying regulations, all of which can make it harder to apply TIC in the same way everywhere [8]. For instance, in European Union member states, resource allocation disparities and a lack of interdisciplinary collaboration have been noted as exacerbating factors, particularly in residential settings serving diverse migrant populations where trauma experiences may intersect with cultural stigma or access inequities [9]. These barriers underscore the importance of viewing TIC implementation as a policy shift, where overcoming them requires not only training but also structural reforms to align incentives. For example, progress in TIC could be tracked with specific measures that are linked to how residential facilities receive funding or how staff performance is evaluated [7].

The Role of Training in TIC Adoption

Training is regarded as essential for TIC adoption [10], given that staff attitudes and beliefs significantly influence practice efficacy. Such training is crucial for aligning daily practices with TIC principles, thereby addressing the policy-practice disconnect [11]. Effective training should incorporate experiential elements to challenge implicit biases, particularly in managing behaviors that challenge where reactive responses might prevail without adequate preparation [12]. In child and adolescent settings, this could extend to scenario-based learning that simulates real-world interactions, promoting proactive rather than punitive approaches [13].

Immersive Technologies and Virtual Reality

Substantial progress has been made in recent years regarding immersive technologies in health care [14], with specific applications in mental health. Augmented reality, for example, has been used effectively for assessing and treating contamination-based obsessive-compulsive disorder [15], and haptic feedback has been explored as a novel method for anxiety management [16]. A considerable body of literature focuses on virtual reality (VR), which has demonstrated efficacy in alleviating suicidal ideation [17], reducing symptoms of depression, anxiety, and stress [18], and improving attitudes toward schizophrenia [19]. VR's role in health care education and behavioral training is supported by evidence indicating its capacity to enhance empathy and perspective-taking through simulations [20,21]. Studies have shown that VR simulations for health care students foster empathy for individuals experiencing psychosis or schizophrenia [22,23], and improvements in empathy domains among medical students have been noted [24]. Although VR applications for trauma survivors exist [25,26], the use of VR in educating staff about TIC remains underexplored.

VR provides distinct advantages for TIC training, such as promoting empathy, facilitating perspective-taking, and enabling experiential learning via realistic scenarios [27]. Its cost-effectiveness, repeatability, and standardization have been highlighted in scoping reviews of medical education [28], positioning it as a viable tool for TIC. Emerging applications include immersive simulations that deepen understanding of TIC principles and enhance application skills [29].

This study underscores the potential of VR to safely recreate trauma-informed environments, allowing staff to practice de-escalation without real-world risks. This is particularly important for managing risk behaviors in children, which is a key factor in specialized childcare [30]. Recent advancements, such as VR programs for trauma-informed de-escalation in hospitals and simulations for understanding adverse childhood experiences, illustrate how VR can transform training by immersing users in scenarios that highlight the impact of space design or behavioral triggers on trauma responses [31,32]. In residential settings, VR could further address aggressive behaviors by enabling trainee teachers or carers to rehearse interventions in virtual classrooms or homes, potentially reducing restrictive practices through better-prepared staff [33].

The Need for Evaluation in European Residential Settings

Given the barriers to TIC implementation and the potential of VR, there is a critical need to develop and evaluate VR-enhanced TIC training in diverse care environments, including residential child and adolescent settings across multiple European contexts. This study addresses this gap through a multicenter pilot conducted in residential facilities and a child psychiatric ward across three European Union countries.

XSL•FO
RenderX

The primary objective of this study was to examine whether participation in a VR-based training program leads to changes in attitudes toward TIC among staff and trainees in residential child and adolescent care settings. Secondary objectives included exploring variability in outcomes across different national contexts and mentor teams, as well as examining participant characteristics associated with attitudes toward TIC.

Methods

This evaluation is reported in line with the TREND (Transparent Reporting of Evaluations with Non-Randomized Designs) statement (Checklist 1) [34], where applicable, and the intervention is described using TIDieR (Template for Intervention Description and Replication; Checklist 2) [35].

Study Design

A multisite pre-post evaluation design was used to evaluate the short-term impact of a VR-enhanced training program on professionals' attitudes toward TIC. This within-subject design enabled each participant to serve as their own control, minimizing confounding due to interindividual variability and enhancing sensitivity to changes in attitudes [36].

Safe4Child Project

This study was conducted as part of the European Union–co-funded project Safe4Child—Caring Violent Child Safely in Child Psychiatric and Residential Units [37]. The project aimed to strengthen frontline workers' competencies in managing behaviors that challenge in children's residential and psychiatric care settings across four European countries (Finland, Germany, Bulgaria, and Ireland). A VR environment and an online continuing education course were developed to support this goal. The project addressed the underlying causes of challenging behaviors and specifically sought to reduce the use of restrictive practices while promoting trauma-informed and child-centered approaches. It also aimed to enhance staff resilience and well-being, improve safety for both children and professionals, and foster more sustainable and ethically grounded practices. All materials produced within the project are open access and available to health and social care professionals on the Safe4Child website.

Setting and Participants

The study was initiated across 4 European Union countries in collaboration with Hamburg University of Applied Sciences (HAW, Germany), Turku University of Applied Sciences (TUAS, Finland), University College Cork (Ireland), and the Medical University of Plovdiv (MUP, Bulgaria). Participants were recruited through convenience sampling coordinated by local research partners within the Safe4Child consortium. However, the postintervention assessment was completed only at 3 sites—Germany, Finland, and Bulgaria. As a result, data from University College Cork comprised preintervention measurements only and were excluded from pre-post change analyses; all inferential results therefore reflect the 3 sites with matched observations.

Eligible participants included professionals and trainees in health care and social work currently employed in, or preparing for, roles in residential and child and adolescent care settings or child psychiatric wards. A total of 120 participants were enrolled, with 30 from each country, consistent with the pilot phase requirements of the Safe4Child grant agreement. Across study sites, participants represented both trainees and practicing professionals with varying degrees of prior exposure to child welfare and mental health care. In Germany, the sample consisted of undergraduate students in social work at HAW, most of whom had completed field placements, primarily in child welfare settings, though only a small subset had professional roles in residential childcare. In Finland and Bulgaria, the cohorts included a broader mix of professionals and trainees working in or preparing for roles in child welfare and mental health contexts.

Participation was voluntary, and informed consent was obtained in accordance with ethical standards and institutional review board approvals in each country.

Intervention

The intervention comprised 2 sequential components designed to build foundational knowledge and practical skills in TIC. The first component was a standardized, self-paced online training module (5 ECTS), delivered in English, covering TIC principles, including the neurobiological effects of trauma, relational dynamics in care, and strategies for managing distress and behavioral escalation in residential and psychiatric settings. Interactive self-evaluation quizzes were embedded to reinforce learning, with the module intended for completion within 2 weeks.

Following the online modules, participants engaged in a VR simulation, delivered individually and uniformly across all study sites. The 15 to 25-minute session was mentor-facilitated and used a standardized scenario depicting a distressed child in a residential care setting; full intervention details are provided in Multimedia Appendix 1 and are documented using the TIDieR checklist [35].

The scenario and mentoring protocol were not culturally customized but were developed to be universally applicable, informed by thematic analysis of case descriptions and semistructured interviews with frontline professionals across the 3 countries [38,39]. This ensured the scenario reflected common patterns of behavioral escalation and staff-child interactions, remaining relevant across national contexts. Mentors provided real-time guidance, and each session concluded with a structured debriefing to promote critical reflection and reinforce TIC approaches.

Outcome Measures

The primary outcome was the shift in participants' attitudes toward TIC, assessed using the ARTIC-10 (10-item version of the Attitudes Related to Trauma-Informed Care) scale [40]. The ARTIC scale, comprising 45 items, is commonly used to evaluate these attitudes, while its 10-item abbreviated version exhibits robust internal consistency (Cronbach α=0.89) and responsiveness to training-induced changes [40]. Improvements in ARTIC scores posttraining reflect enhanced trauma-informed perspectives among staff [41-43]. This validated instrument, designed for rapid administration, maintains strong psychometric

XSL•FO
RenderX

properties [44,45]. Responses were captured on a 7-point Likert scale, with composite scores calculated by averaging all items after reverse-scoring items 2, 4, 6, 8, and 9, per standard procedures. Higher scores indicate stronger alignment with TIC principles. Although the ARTIC-10 demonstrates acceptable internal consistency and responsiveness to training, recent psychometric evaluations have identified limitations regarding factor structure stability and cultural invariance across diverse populations [46]. These properties warrant caution when comparing scores across national or cultural contexts. The ARTIC-10 was administered at 2 time points via a secure, General Data Protection Regulation (GDPR)-compliant, self-administered online survey system: immediately before commencement of the online module (preintervention) and immediately after completion of the VR simulation (postintervention). Data collection occurred between January 2024 and March 2024. This design evaluates the effect of the combined training program (online module plus VR); however, the contribution of the VR component cannot be isolated from that of the preceding didactic content.

Data Analysis

All analyses were performed using R version 4.5.1 (R Foundation for Statistical Computing). Descriptive statistics summarized participant demographics and ARTIC-10 scores. Continuous variables were reported as means and SDs for normally distributed data or medians and IQRs for non-normally distributed data. Categorical variables were expressed as frequencies and percentages. Computational reproducibility materials, including data-processing scripts, model code, and figure-generation routines, are provided in Multimedia Appendix 2 (Code Reproduction).

Only participants with complete, matched preintervention and postintervention responses were included (N=79). Matching was achieved using anonymized identifiers assigned by local investigators to ensure within-subject comparability. To assess measurement reliability, we evaluated the internal consistency of the ARTIC-10 at preintervention and postintervention, overall and stratified by country. The Cronbach α was estimated from the 10-item responses, with its asymptotic SE, and Wald-type 95% CIs.

To evaluate the intervention's effect on TIC attitudes, a multilayered analytical approach was adopted. The Wilcoxon signed-rank test was used to assess within-subject differences in ARTIC-10 scores, accounting for non-normal distribution. A binomial test evaluated the proportion of participants with increased ARTIC-10 scores postintervention, using a 50% threshold for significance. Effect size was estimated using Cohen *d*, with thresholds of 0.2, 0.5, and 0.8 for small, medium, and large effects, respectively.

To explore contextual and demographic influences, multivariable regression modeling was applied. Linear regression modeled the ARTIC-10 score difference (score_diff = artic_post - artic_pre), and logistic regression modeled the probability of an increase (increase_fct, postintervention > preintervention). Predictors included country, sex, age, prior health care experience, years of experience with children and adolescents, education level, previous trauma-related training, and type of employing organization. Data on field placement duration and specific focus on child welfare were collected only for the German cohort, as these variables were relevant to participants' undergraduate Social Work curriculum at HAW; equivalent structured placement data were not systematically recorded at the Finnish and Bulgarian sites. These predictors were therefore examined in exploratory site-specific analyses for Germany only, and results should be interpreted accordingly. A 2-sided significance level of $P<.05$ was used for all frequentist tests.

In addition to the prespecified frequentist tests, we used a Bayesian framework to quantify the magnitude and uncertainty of pre–post change and to model site-level heterogeneity under explicit priors. For each participant with matched observations (N=79), the paired change was computed as $\Delta i=ARTICpost,i-ARTICpre,i$ after standard scoring.

We first fit an intercept-only model:

$$\Delta i \sim \text{Student-t}(\mu,\sigma,\nu), \mu=\alpha,$$

with a skeptical prior $\alpha \sim N(0,0.3)$ to encode the expectation of little to no average change on the 1 - 7 ARTIC scale, and a robust residual prior $\sigma \sim t3(0,0.7)$ to mitigate the influence of outliers. The 2-tailed Student *t* likelihood provides additional robustness to occasional extreme differences.

To examine contextual moderation, we then specified:

$$\Delta i \sim \text{Student-t}(\mu i,\sigma,\nu), \mu i=\alpha+\beta \text{country}[i]+\gamma 1\{\text{prior training}i=\text{Yes}\}$$

treating country as a factor with Bulgaria as the reference (default alphabetical ordering in R; the choice of reference category does not affect pairwise contrasts or substantive interpretation) and adding an indicator for prior trauma-related training. Regression coefficients received weakly informative priors $\beta,\gamma \sim N(0,0.3)$; the same priors on σ and ν and the same likelihood as above were used.

Models were fit in Bayesian regression models using Stan (Stan Development Team) [47] (Hamiltonian Monte Carlo via Stan) with 4 chains, 4000 iterations per chain (1000 warm-up), adapt_delta 0.98 - 0.99, and maximum treedepth 12. Convergence was assessed using R^≈1.00 and large effective sample sizes, and model adequacy was inspected with posterior predictive checks.

Inference focused on posterior means, 95% credible intervals (CrI), and probabilities for decision-relevant events. Specifically, we report P(Δ\gt 0) as the probability of any improvement, P(Δ\gt 0.20) as a benchmark for a practically larger lift, and the share of posterior mass within a region of practical equivalence (ROPE). The ROPE formalizes "negligible" change: we prespecified ±0.10 ARTIC points, which is approximately 1.4% of the total scale range and smaller than the observed within-person dispersion, making it a conservative threshold. When most posterior mass lies inside the ROPE, we interpret the effect as practically null; when most mass lies above both zero and the ROPE (and, where relevant, beyond +0.20), we interpret it as a practically meaningful improvement. Country-specific posteriors and contrasts were obtained by transforming the linear predictor, propagating full uncertainty

XSL•FO
RenderX

from all parameters. Reproducible code for all Bayesian analyses is provided in Multimedia Appendix 2.

Ethical Considerations

This study was conducted in accordance with the ethical principles established in the Declaration of Helsinki (World Medical Association, 2013) [48]. Ethical approval was granted by the Human Sciences Ethics Committee of Turku University of Applied Sciences (Statement LP06_2022, dated September 5, 2022), covering all participating study sites. All participants received a comprehensive information sheet describing the study objectives, procedures, voluntary nature of participation, and their right to withdraw at any time without consequence. Written informed consent was obtained from all participants prior to data collection.

All data were collected, processed, and stored in accordance with the GDPR (Regulation [EU] 2016/679). A GDPR-compliant privacy notice was provided to all participants alongside the study information materials. Participant confidentiality was maintained throughout; data were anonymized prior to analysis and no identifying information is reported in this manuscript. Participants received no financial compensation for their participation in this study.

Results

Participant Characteristics

Only participants who completed both the preintervention and postintervention questionnaires were included in the analysis, reducing the final analytical sample from the original 120 respondents to 79. The overall response rate was estimated at 66%. The sample was predominantly female (n=71, 90%), with the highest proportion of male participants coming from Finland and Germany (13% each, Table 1).

Table . Participant characteristics by country (N=79).

Characteristic	Overall (N=79)	Bulgaria (n=25)	Finland (n=24)	Germany (n=30)
Sex (male), n/N (%)	8/79 (10)	1/25 (4)	3/24 (13)	4/30 (13)
Age (y), mean (SD)	29.8 (10.7)	34.1 (14.5)	26.3 (6.0)	29.1 (8.9)
Education[a], n/N (%)				
Student	58/79 (73)	12/25 (48)	21/24 (88)	25/30 (83)
Diploma	3/79 (3.8)	2/25 (8)	—[b]	1/30 (3.3)
Bachelor's degree	6/79 (7.6)	—	3/24 (13)	3/30 (10)
Master's degree	7/79 (8.9)	6/25 (24)	—	1/30 (3.3)
Doctorate	5/79 (6.3)	5/25 (20)	—	—
Experience				
Health care experience (y), mean (SD)	6.7 (8.9)	12.6 (13.5)	3.8 (3.3)	3.9 (2.9)
Specialized experience (y), mean (SD)[a]	3.6 (4.8)	4.6 (6.9)	1.7 (2.7)	4.0 (3.3)
No working experience, n/N (%)	11/79 (14)	3/25 (12)	6/24 (25)	2/30 (6.7)
Previous TIC[c] training (yes), n/N (%)	17/79 (22)	9/25 (36)	4/24 (17)	4/30 (13)
Type of organization[a]				
None	62/79 (78)	24/25 (96)	18/24 (75)	20/30 (67)
Child welfare (residential)	8/79 (10)	—	1/24 (4.2)	7/30 (23)
Mental health	8/79 (10)	—	5/24 (21)	3/30 (10)
Pediatric clinics or units	1/79 (1.3)	1/25 (4)	—	—

[a]Indicates statistically significant differences between sites (P<.05).

[b]Not applicable.

[c]TIC: trauma-informed care.

The mean age of participants was 29.8 (SD 10.7) years, with Bulgarian participants being older on average (34.1, SD 14.5 y) compared to those from Finland (26.3, SD 6.0 y) and Germany (29.1, SD 8.9 y). Educational backgrounds varied, with 73% of participants identifying as students. Among countries, Finland (88%) and Germany (83%) had the highest proportions of student participants, while Bulgaria showed a more diverse distribution, including individuals with doctoral degrees (20%) and master's degrees (24%).

XSL•FO
RenderX

On average, participants had 6.7 (SD 8.9) years of health care experience, with Bulgarian participants reporting the most extensive experience (mean 12.6, SD 13.5 y). Specialized experience averaged 3.6 (SD 4.8) years across all participants. Between-site comparisons revealed statistically significant differences in educational background (χ^2_8=28.66; *P*<.001), type of employing organization (χ^2_6=17.50; *P*=.008), and years of experience working with children and adolescents (Kruskal-Wallis χ^2_2=8.23; *P*=.016). No significant differences were observed for sex (χ^2_2=1.52; *P*=.47), age (Kruskal-Wallis χ^2_2=1.20; *P*=.55), years of health care experience (Kruskal-Wallis χ^2_2=1.06; *P*=.59), or prior trauma-related training (Fisher exact test, *P*=.13). These baseline differences reflect the heterogeneous composition of cohorts across sites—predominantly undergraduate students in Germany, with more diverse professional backgrounds in Bulgaria and Finland—and were accounted for in multivariable regression models.

Overall, 22% of participants reported previous training related to TIC, with the highest prevalence among Bulgarian participants (36%). The majority of participants (78%) reported no affiliation with specific organizational types, though a smaller subset identified with child welfare or mental health organizations. Notably, 23% of German participants were affiliated with child welfare organizations, while 21% of Finnish participants worked in mental health settings.

Internal Consistency (Cronbach α) of the ARTIC-10

Internal consistency of the ARTIC-10 was good at baseline and improved after the intervention. Overall, Cronbach α increased from 0.839 (95% CI 0.792 - 0.886) preintervention to 0.894 (95% CI 0.863 - 0.926) postintervention. By country, reliability was acceptable to good at baseline: Bulgaria 0.703 (95% CI 0.578 - 0.827), Finland 0.810 (95% CI 0.703 - 0.918), and Germany 0.809 (95% CI 0.708 - 0.910)—and generally stable or higher postintervention: Bulgaria 0.836 (95% CI 0.761 - 0.911), Finland 0.843 (95% CI 0.740 - 0.946). Germany showed a modest decrease to 0.765 (95% CI 0.621 - 0.909) postintervention, remaining within an acceptable range for group comparisons.

These values indicate that the instrument demonstrated adequate internal consistency across sites and time points, with the strongest reliability observed in the pooled postintervention data.

Changes in ARTIC-10 Scores Following the Intervention

Analysis of ARTIC-10 scores revealed a significant overall improvement in TIC attitudes following the intervention. The mean score increased from 5.33 (SD 1.05; range 2.3 - 7.0) at baseline to 5.57 (SD 1.20; range 2.1 - 7.0) postintervention. A Wilcoxon signed-rank test confirmed that this change was statistically significant (*P*<.001), with a small-to-moderate effect size (Cohen *d*=0.22).

To explore potential heterogeneity in outcomes across implementation contexts, we conducted center-specific analyses. A significant improvement in ARTIC-10 scores was observed among participants in Germany (HAW; *P*<.001), with a moderate effect size (Cohen *d*=0.56). In contrast, no significant changes were detected among participants from Bulgaria (MUP; *P*=.796) or Finland (TUAS; *P*=.19).

The proportion of participants exhibiting an increase in ARTIC-10 scores (defined as postintervention > preintervention) also varied by country. In Germany, 80% of participants demonstrated improvement (95% CI 61.4% - 92.3%), compared with 60% in Bulgaria (95% CI 38.7% - 78.9%) and 54.2% in Finland (95% CI 32.8% - 74.4%; Figure 1).

XSL•FO
RenderX

Figure 1. Changes in ARTIC-10 (10-item version of the Attitudes Related to Trauma-Informed Care) scores by country. Panel (A) displays the distribution of change scores (post–pre intervention) in ARTIC-10 for each country. Eye plots show the density of individual scores, with the mean and 95% CIs indicated at the center. Panel (B) presents the proportion of participants in each country who demonstrated an increase in ARTIC-10 scores following the intervention, also with 95% CIs. In both panels, countries are color-coded for visual distinction.

XSL•FO
RenderX

To further examine differences in outcomes, participants were stratified based on the direction of change in their ARTIC-10 scores: those with an increase (Δ>0) and those with a decrease or no change (Table 2). The increase group reported significantly higher postintervention scores (mean 5.9, SD 0.9) compared to the decrease group (mean 5.0, SD 1.4; *P*=.005). The average relative change in ARTIC-10 scores across the full sample was 4.9% (SD 14.1%). The increase group experienced a mean improvement of 11.6% (SD 10.6%), while the decrease group exhibited a mean reduction of −8.0% (SD 10.6%; *P*<.001).

Table . Descriptive statistics of participants by ARTIC-10 (10-item version of the Attitudes Related to Trauma-Informed Care) score change group.

Characteristic	Overall (N=79)	Decrease (n=27)	Increase (n=52)	*P* value
ARTIC-10 scores, mean (SD)				
Preintervention	5.3 (1.1)	5.4 (1.2)	5.3 (1.0)	.70
Postintervention	5.6 (1.2)	5.0 (1.4)	5.9 (0.9)	.005[a]
Relative change (%)	4.9 (14.1)	−8.0 (10.6)	11.6 (10.6)	<.001[b]
Country, n/N (%)				
Bulgaria	25/79 (32)	10/27 (37)	15/52 (29)	.50
Finland	24/79 (30)	11/27 (41)	13/52 (25)	.15
Germany	30/79 (38)	6/27 (22)	24/52 (46)	.04[c]
Type of organization, n/N (%)				
None	62/79 (78)	21/27 (78)	41/52 (79)	>.90
Child welfare (residential)	8/79 (10)	3/27 (11)	5/52 (9.6)	>.90
Mental health	8/79 (10)	3/27 (11)	5/52 (9.6)	>.90
Pediatric	1/79 (1.3)	0/27 (0)	1/52 (1.9)	>.90
Demographics and experience				
Sex (male), n/N (%)	8/79 (10)	4/27 (15)	4/52 (7.7)	.40
Age (y), mean (SD)	29.8 (10.7)	31.9 (11.4)	28.8 (10.3)	.20
Health care experience (y), mean (SD)	6.7 (8.9)	8.2 (8.4)	5.9 (9.2)	.07
No prior experience, n/N (%)	11/79 (14)	5/27 (19)	6/52 (12)	.50
Prior TIC[d] training, n/N (%)	17/79 (22)	4/27 (15)	13/52 (25)	.30

[a]*P*<.01.

[b]*P*<.001.

[c]*P*<.05.

[d]TIC: trauma-informed care.

A significantly higher proportion of participants in the increase group were from Germany (46% vs 22% in the decrease group; *P*=.04). No significant differences were observed by country of origin for Bulgaria (*P*=.50) or Finland (*P*=.15). Prior trauma-related training was more common in the increase group (25% vs 15%), although the difference was not statistically significant (*P*=.30). Similarly, the two groups did not differ significantly by sex, age, or organization type.

Participants in the decrease group tended to be slightly older (mean 31.9 y, SD 11.4 y) than those in the increase group (mean 28.8, SD 10.3 y), though this difference did not reach statistical significance (*P*=.20). Those in the decrease group also reported more years of health care experience (mean 8.2, SD 8.4 y) compared to the increase group (mean 5.9, SD 9.2 y; *P*=.07). Prior health care experience was marginally more common in the decrease group (19% vs 12%; *P*=.50).

Multivariable Regression Analysis

A linear regression model was used to identify predictors of change in ARTIC-10 scores, defined as the difference between post- and preintervention scores. As shown in Figure 2, the univariate analysis revealed that participants from HAW (Germany) experienced a significantly greater improvement in ARTIC-10 scores compared to those from MUP (Bulgaria), which served as the reference category (β=0.461, 95% CI 0.132 - 0.790; *P*=.007). The effect for TUAS, Finland, also compared to Bulgaria, was not statistically significant (β=0.241, 95% CI −0.106 to 0.589; *P*=.170). Younger age was marginally associated with greater score improvements (β=−0.012, 95% CI −0.025 to 0.002; *P*=.083), suggesting that younger participants tended to benefit more from the intervention.

XSL•FO
RenderX

Figure 2. Predictors of change in ARTIC-10 (10-item version of the Attitudes Related to Trauma-Informed Care) scores. (A) Linear regression results shown as a forest plot of β coefficients with 95% CIs. Red bars represent univariate estimates; blue bars indicate multivariable model estimates. The vertical dashed line at 0 denotes the null effect. (B) Logistic regression results displayed as odds ratios (ORs) with 95% CIs for the likelihood of a positive change in ARTIC-10 scores. The vertical dashed line at 1 represents the null value (no effect). Bulgaria serves as the reference category for country comparisons. Asterisks indicate statistical significance: *P<.05; ***P<.001.

XSL•FO
RenderX

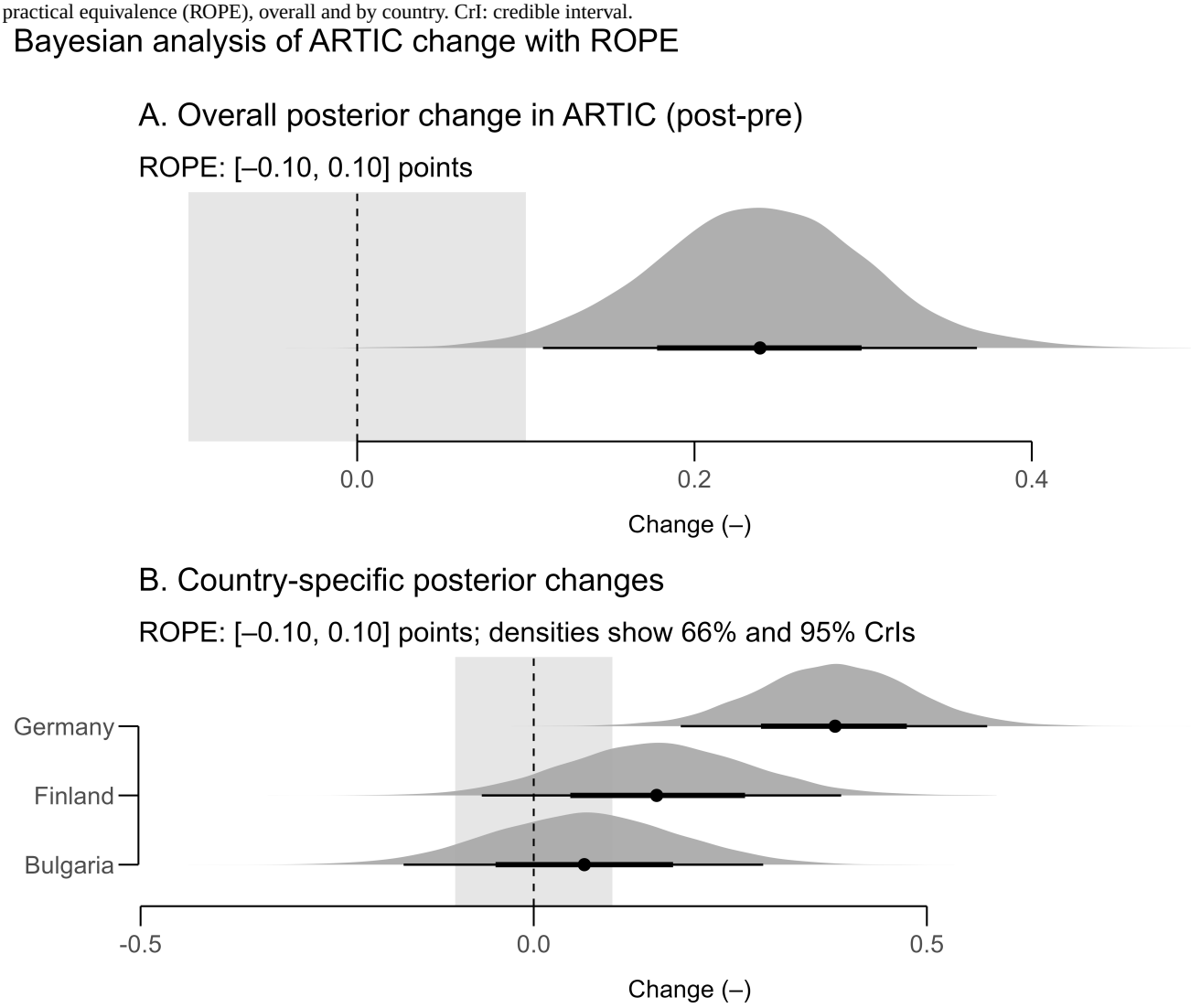
In the multivariable linear regression, the effect of the program delivery center at HAW, Germany, remained strong and statistically significant (β=0.464, 95% CI 0.111 - 0.816; *P*=.01), while the association with age was not significant (β=−0.009, 95% CI −0.023 to 0.004; *P*=.18). These findings indicate that country-level differences in score change were not fully accounted for by individual-level covariates, such as age or prior experience.

To examine the likelihood of any positive change in ARTIC-10 scores, a multivariable logistic regression model was fitted. As depicted in Figure 2, both the center delivering the intervention and prior trauma-related training emerged as predictors. Participants at HAW, Germany, showed a trend toward higher odds of positive change compared to those in Bulgaria, though this did not reach statistical significance (odds ratio 3.39, 95% CI 0.90 - 14.00; *P*=.077). Similarly, participants who had received previous trauma-related training showed a nonsignificant tendency toward a greater likelihood of improvement (odds ratio 2.75, 95% CI 0.73 - 12.60; *P*=.16). While neither association attained conventional significance at α=.05, the direction and magnitude of effects suggest potential roles for site-related factors and foundational knowledge in facilitating intervention responsiveness, warranting further investigation in adequately powered studies.

Bayesian Analysis

A Bayesian analysis was undertaken to quantify the pre-post change in ARTIC-10 while incorporating a skeptical prior centered at zero (Figure 3). The paired change was defined for each participant as Δ=ARTICpost−ARTICpre. Under a normal (0, 0.3) prior on the mean change and a robust Student 2-tailed *t* test likelihood, the posterior mean Δ was estimated at 0.24 points on the 1 - 7 scale, with a 95% CrI from 0.11 to 0.37. The posterior probability that the true mean change exceeded zero was effectively 1.00, whereas only 1.8% of posterior mass lay within a region of practical equivalence of ±0.10 points. The probability that Δ exceeded a practically relevant threshold of 0.20 points was 0.73.

Figure 3. Posterior distributions of ARTIC-10 (10-item version of the Attitudes Related to Trauma-Informed Care) change (Δ=post-pre) with the region of practical equivalence (ROPE), overall and by country. CrI: credible interval.



Heterogeneity by site was examined by adding country and prior training as additive predictors of Δ. Posterior country-specific means were 0.06 points for Bulgaria (95% CrI −0.17 to 0.29; *P*[Δ>0]=0.71; 54% within ±0.10), 0.16 points for Finland (95% CrI −0.07 to 0.38; *P*[Δ>0]=0.91; 30% within ±0.10), and 0.38 points for Germany (95% CrI 0.19 to 0.58;

XSL•FO
RenderX

$P[\Delta>0]\approx1.00$; 0.3% within ±0.10). These posteriors support a clear positive change in Germany, a likely but smaller positive change in Finland, and an uncertain change centered near zero in Bulgaria, mirroring the pattern observed in the frequentist analyses.

Prior trauma-related training was associated with an additional mean increase of 0.09 points (95% CrI −0.20 to 0.39), with a posterior probability of 0.73 that the training effect exceeded zero, suggesting probable, though uncertain, amplification among previously trained participants.

Discussion

Principal Findings

In our study, an improvement in attitudes related to TIC, as measured by the ARTIC-10 scale, was observed among students and professionals in mental health and child-welfare settings following a VR-enhanced TIC training program. The overall mean (SD) ARTIC-10 score increased from 5.33 (1.05) to 5.57 (1.20), with statistical significance ($P<.001$) and a small-to-moderate effect size. The center delivering the intervention and prior trauma-related training emerged as predictors, with participants from Germany showing greater improvement and those with prior training exhibiting higher odds of positive change. Heterogeneity across countries was noted, with significant effects limited to the German cohort, potentially influenced by their academic background and recent field placement experiences in child welfare.

Comparison With Other Interventions

The magnitude of attitudinal change observed in this study aligns with the broader literature on TIC training interventions. A recent scoping review encompassing 157 studies found that ARTIC-based evaluations consistently demonstrate small-to-moderate improvements following diverse training modalities, including didactic curricula, experiential workshops, web-based modules, and blended approaches [49]. Across studies using pre-post designs in health care and child welfare settings, effect sizes have typically ranged from 0.2 to 0.5 SDs, with absolute score increases of approximately 0.15 to 0.30 points on the 7-point ARTIC scale [41,42,50,51]. Our overall effect ($d=0.22$; mean change 0.24 points) falls within this range, while the German cohort's larger effect ($d=0.56$) exceeds typical benchmarks.

What distinguishes VR-enhanced training from traditional approaches is the mode of experiential engagement rather than the magnitude of effect. Conventional TIC interventions rely on seminars, case discussions, and self-directed learning; VR adds immersive, embodied simulation that may facilitate empathy development and perspective-taking through direct experiential exposure [52,53]. The comparable effect sizes suggest VR can achieve attitudinal gains similar to more time-intensive traditional formats, potentially offering efficiency advantages. The amplifying role of prior trauma-related training observed here mirrors findings elsewhere, suggesting that foundational knowledge enhances receptivity to experiential learning regardless of delivery modality [42].

VR-based interventions were evaluated as demonstrating unique advantages over non-VR methods in enhancing TIC attitudes, primarily through immersion, empathy-building, and emotional engagement. A scoping review by Tay et al [52] of 16 studies found that over half reported improvements in knowledge and attitudes toward mental illnesses, including trauma-related conditions, with VR consistently enhancing empathy in all seven relevant studies. Compared to controls, VR often yielded superior outcomes in empathy and stigma reduction, though not always in attitudes. Thorough consideration suggests VR facilitates active, embodied learning with immediate feedback, promoting retention of attitudinal skills that are difficult to achieve via passive methods [53]. In social work education, VR simulations for substance-engaged clients have built competencies in trauma-sensitive approaches, underscoring VR's complementary role rather than its replacement for traditional training [54].

The magnitude of improvement in ARTIC-10 scores observed in this study is broadly comparable to those reported in prior ARTIC-based interventions. For instance, structured seminars and web-based approaches by Stewart et al [42] and Loomis et al [51] achieved similarly modest but statistically significant gains, demonstrating that attitudinal change toward TIC is generally incremental. What distinguishes the present VR-enhanced approach is the efficiency with which changes were achieved, as immersive simulations created embodied learning experiences without requiring extensive seminar series or multiple module completion. These findings support VR's role as a complementary modality, capable of delivering experiential benefits similar to traditional approaches but in potentially shorter, more engaging formats.

Strengths and Limitations

Strengths of this study include the use of a pre-post intervention design with multivariable regression analyses to control for confounders, providing robust evidence of intervention effects. The diverse sample from multiple countries (Bulgaria, Finland, and Germany) enhanced the cross-cultural perspective, while the use of the validated ARTIC-10 scale ensured reliable outcome assessment.

However, this study has several limitations that warrant consideration when interpreting its findings. The reliance on the ARTIC-10 as the sole outcome measure introduces vulnerability to self-report bias; moreover, documented psychometric limitations of the scale, particularly regarding factor structure stability and cultural invariance [46], warrant caution when interpreting cross-national comparisons. The lack of a control group prevents disentangling VR-specific effects from novelty or facilitator-driven influences. The sample size and heterogeneous participant backgrounds reduce generalizability, especially across national contexts. Practical barriers related to VR technology, including cost, accessibility, and cybersickness, may limit scalability and create inequities in who can access such training. These issues warrant caution in extrapolating the findings beyond the study's specific settings.

This study used a single-group pre-post design with immediate posttraining measurement, which introduces several threats to internal validity. First, testing and demand effects may have

XSL•FO
RenderX

occurred: participants might respond differently at postassessment because they have just completed the training or wish to align with perceived expectations. Second, history, maturation, and regression to the mean cannot be excluded without a control group. Third, social desirability and self-report bias are inherent to attitudinal scales and may inflate apparent improvements. Fourth, novelty and facilitator effects—variation in mentors across sites and the novelty of VR technology itself—could influence responses independently of the training content.

The relatively small sample size may have constrained statistical power, particularly for detecting smaller effect sizes or differences within subgroups, such as those based on program delivery center or professional experience. While the case-crossover design mitigated interindividual variability, a larger sample could enhance the precision and robustness of findings, especially for exploring contextual or demographic influences. The small proportion of male participants (n=8, 10%) also limits the generalizability of findings across sexes.

Second, the generalizability of the results may be limited due to the study's focus on specific countries, centers (Germany, Finland, and Bulgaria), and settings (residential child and adolescent care and child psychiatric wards). Although the VR intervention was designed to be universally applicable, cultural, organizational, or systemic differences across countries and care settings may restrict the applicability of the findings to other contexts, such as nonresidential care or regions with distinct health care infrastructures.

Third, potential biases may have influenced the outcomes. Selection bias is a concern, as convenience sampling may have resulted in a participant pool that was more motivated or receptive to VR-based training, potentially inflating observed improvements in ARTIC-10 scores. Additionally, social desirability bias may have affected self-reported ARTIC-10 scores, particularly in postintervention assessments, as participants might have provided responses aligned with trauma-informed principles due to the training context or mentor presence.

Finally, limitations related to VR technology itself must be acknowledged. Accessibility issues, such as limited availability, technological variability of VR equipment, or inadequate technical infrastructure in some settings, may hinder the scalability and equitable implementation of the intervention. Furthermore, variability in participants' comfort with VR technology, prior exposure, or technical challenges (eg, motion sickness and interface difficulties) may have introduced variability in the intervention's effectiveness, potentially affecting the consistency of outcomes across participants. Finally, without long-term follow-up, it remains unclear whether the gains observed represent lasting attitudinal change or short-term responses to the intervention context.

Implications for Practice and Future Research

The critical next step for this line of research is comparative evaluation. Future trials should randomize participants to VR-enhanced versus standard (non-VR) TIC training conditions, matched for contact time and content, to isolate the specific contribution of immersive simulation. Such designs would address whether VR adds value beyond conventional approaches or simply represents an alternative delivery mode with equivalent outcomes.

Beyond comparative effectiveness, several methodological advances are warranted. First, longitudinal follow-up assessments at 3, 6, and 12 months postintervention would establish whether attitudinal gains are durable or decay over time. Second, behavioral outcome measures—such as direct observation of staff-child interactions, documented use of de-escalation techniques, or rates of restrictive practice use—would determine whether attitudinal improvements translate into practice change. Third, nested qualitative components (eg, semistructured interviews and focus groups) could illuminate mechanisms of effect and identify barriers to skill transfer. Fourth, process evaluation embedded within future trials should capture implementation fidelity, participant engagement metrics, and contextual factors that may moderate effectiveness.

The observed site-level heterogeneity underscores the need for studies that are adequately powered to examine moderators of intervention response, including baseline TIC exposure, professional background, and organizational readiness. Multilevel designs that account for clustering within sites and mentors would strengthen causal inference. Finally, cost-effectiveness analyses comparing VR-enhanced training to conventional approaches would inform resource allocation decisions for workforce development in child welfare and mental health settings.

Conclusions

This multisite evaluation provides preliminary evidence that a VR-enhanced TIC training program can improve attitudes toward TIC among emerging and practicing professionals in child and adolescent residential and mental health settings. The relatively small, predominantly female sample and the potential for selection and self-report bias warrant caution in interpretation, and the heterogeneity of results across countries highlights the need for approaches tailored to cultural, professional, and systemic contexts.

Nevertheless, the significant improvement observed in Germany (d=0.56) and the amplifying effect of prior trauma-related training suggest that VR-enhanced approaches may be most effective when integrated into scaffolded curricula that build upon learners' existing knowledge. These findings support the role of immersive, experiential methods as a complementary tool within broader TIC education frameworks. Comparative trials with non-VR controls are now needed to determine whether VR confers specific advantages and to guide evidence-based workforce development in TIC.

XSL•FO
RenderX

Funding

This work was supported, in part, by the European Union's Erasmus+ Programme (KA220-HED—Cooperation Partnerships in Higher Education) for project activities only. The analysis and writing of this manuscript received no specific funding from public, commercial, or not-for-profit sectors. The sponsor had no role in the study design; in the collection, analysis, and interpretation of data; in the writing of the manuscript; or in the decision to submit for publication.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Safe4Child Intervention: curriculum, VR simulation, and delivery details.
[DOCX File, 2101 KB - mededu_v12i1e86543_app1.docx]

Multimedia Appendix 2
Reproducible code with simulated data.
[DOCX File, 61 KB - mededu_v12i1e86543_app2.docx]

Checklist 1
TREND Checklist.
[PDF File, 451 KB - mededu_v12i1e86543_app3.pdf]

Checklist 2
TiDieR Checklist.
[DOCX File, 19 KB - mededu_v12i1e86543_app4.docx]

References

1. Chapter 3: understanding the impact of trauma. In: Trauma-Informed Care in Behavioral Health Services: Substance Abuse; Mental Health Services Administration (US); 2014. URL: https://www.ncbi.nlm.nih.gov/books/NBK207191/ [accessed 2026-03-20]
2. Bloom SL, Farragher B. Destroying sanctuary: the crisis in human service delivery systems. In: 3 When Terror Becomes a Way of Life: Oxford University Press; 2010:91-129. [doi: 10.1093/acprof:oso/9780195374803.003.0004]
3. Zhang S, Conner A, Lim Y, Lefmann T. Trauma-informed care for children involved with the child welfare system: a meta-analysis. Child Abuse Negl 2021 Dec;122:105296. [doi: 10.1016/j.chiabu.2021.105296] [Medline: 34478999]
4. Kelly P, Saab MM, Hurley EJ, et al. Trauma informed interventions to reduce seclusion, restraint and restrictive practices amongst staff caring for children and adolescents with challenging behaviours: a systematic review. J Child Adolesc Trauma 2023 Sep;16(3):629-647. [doi: 10.1007/s40653-023-00524-2] [Medline: 37593061]
5. Resnick B, Galik E, Wells CL, Boltz M, Holtzman L. Optimizing physical activity among older adults post trauma: overcoming system and patient challenges. Int J Orthop Trauma Nurs 2015 Nov;19(4):194-206. [doi: 10.1016/j.ijotn.2015.03.007] [Medline: 26547682]
6. Huo Y, Couzner L, Windsor T, Laver K, Dissanayaka NN, Cations M. Barriers and enablers for the implementation of trauma-informed care in healthcare settings: a systematic review. Implement Sci Commun 2023 May 5;4(1):49. [doi: 10.1186/s43058-023-00428-0] [Medline: 37147695]
7. MacRae A, Berkovic D, Ryan J, Hatzikiriakidis K, Ayton D. Trauma-informed care within residential aged care settings: a systematic scoping review. Int J Geriatr Psychiatry 2023 Dec;38(12):e6028. [doi: 10.1002/gps.6028] [Medline: 38038652]
8. Graham N, Whitaker L, Smith G, Hurley J. Trauma-informed care in acute adult public mental health settings: a scoping study examining implementation. Issues Ment Health Nurs 2024 Feb;45(2):217-231. [doi: 10.1080/01612840.2024.2308543] [Medline: 38466388]
9. Greenbaum J, Caffo E. Migrant trauma and the role of healthcare professionals and systems in enhancing wellbeing. Med Res Arch 2024;12(1). [doi: 10.18103/mra.v12i1.5037]
10. Şimşek Z, Uğur B. Trauma-informed healthcare systems: an evaluation of trauma-informed care training for hospital-based healthcare professionals in the aftermath of the 2023 earthquakes in Türkiye. Health Policy Plan 2025 Feb 6;40(2):234-243. [doi: 10.1093/heapol/czae118] [Medline: 39658005]
11. Morra C, Nguyen K, Sieracki R, Pavlic A, Barry C. Trauma-informed care training in trauma and emergency medicine: a review of the existing curricula. West J Emerg Med 2024 May;25(3):423-430. [doi: 10.5811/westjem.18394] [Medline: 38801050]
12. Lee CH, Santos CD, Brown T, Ashworth H, Lewis JJ. Trauma-informed care for acute care settings: a novel simulation training for medical students. MedEdPORTAL 2023;19:11327. [doi: 10.15766/mep_2374-8265.11327] [Medline: 37520013]

XSL•FO
RenderX

13. Creswell Báez J, Lichlyter-Klein M, O'Connell M. Trauma-informed care. In: Social Work: Oxford University Press; 2023. [doi: 10.1093/obo/9780195389678-0323]
14. Petrakis A, Koumakis L, Kazantzaki E, Kondylakis H. Transforming healthcare: a comprehensive review of augmented and virtual reality interventions. Sensors (Basel) 2025 Jun 15;25(12):3748. [doi: 10.3390/s25123748] [Medline: 40573635]
15. Ji Y, Luo Y, Chen R. Cleanscape: an augmented reality intervention for assessing and treating contamination-based obsessive-compulsive disorder. Presence: Virtual Augment Realit 2025 Mar 13;34:119-133. [doi: 10.1162/pres_a_00443]
16. Mishra S, Seth S, Jain S, et al. An emotionally intelligent haptic system—an efficient solution for anxiety detection and mitigation. Comput Methods Programs Biomed 2025 Mar;260:108590. [doi: 10.1016/j.cmpb.2025.108590] [Medline: 39778397]
17. Ronaghi MH, Ronaghi M. How does virtual reality technology affect suicidal ideation in society? Int J Ment Health Nurs 2025 Feb;34(1):e13443. [doi: 10.1111/inm.13443] [Medline: 39394638]
18. Yang J, Li Y, Gao D, Xie X, Ji W, Gao J. Effects of virtual reality technology on anxiety and depression in older adults with chronic diseases: a systematic review and meta-analysis of randomized controlled trials. Worldviews Evid Based Nurs 2025 Feb;22(1):e12763. [doi: 10.1111/wvn.12763] [Medline: 39697092]
19. Mason O, Westhead M. Senua's psychosis in virtual reality: effects on mental health stigma. Behav Inf Technol 2025 Jun 15;44(10):2134-2139. [doi: 10.1080/0144929X.2023.2265516]
20. Baidoo K, Adu C. Virtual reality training to enhance clinical competence and student engagement in Ghana. Br J Midwifery 2025 Jan 2;33(1):6-12. [doi: 10.12968/bjom.2024.0037]
21. Tan CP, Howes DC, Manikam A, Lim SM. Role of a dementia virtual reality app in developing situated empathy, attitude and person-centred care—a qualitative approach. Int J Older People Nurs 2025 Mar;20(2):e70012. [doi: 10.1111/opn.70012] [Medline: 39895420]
22. Pina-Thomas D, Zohn JH, Phillips L. A comparison of virtual reality and traditional audio-recorded hearing voices simulations: their impact on empathy in health care students. Nurs Educ Perspect 2025;46(2):116-118. [doi: 10.1097/01.NEP.0000000000001271] [Medline: 38682864]
23. Marques AJ, Gomes Veloso P, Araújo M, et al. Impact of a virtual reality-based simulation on empathy and attitudes toward schizophrenia. Front Psychol 2022;13:814984. [doi: 10.3389/fpsyg.2022.814984] [Medline: 35602736]
24. Lin HL, Wang YC, Huang ML, et al. Can virtual reality technology be used for empathy education in medical students: a randomized case-control study. BMC Med Educ 2024 Nov 5;24(1):1254. [doi: 10.1186/s12909-024-06009-6] [Medline: 39497110]
25. Loucks L, Rizzo A, Rothbaum BO. Virtual reality exposure for treating PTSD due to military sexual trauma. J Clin Psychol 2025 Feb;81(2):81-92. [doi: 10.1002/jclp.23750] [Medline: 39526899]
26. Mistry D, Zhu J, Tremblay P, et al. Meditating in virtual reality: proof-of-concept intervention for posttraumatic stress. Psychol Trauma 2020 Nov;12(8):847-858. [doi: 10.1037/tra0000959] [Medline: 32897088]
27. Pottle J. Virtual reality and the transformation of medical education. Future Healthc J 2019 Oct;6(3):181-185. [doi: 10.7861/fhj.2019-0036] [Medline: 31660522]
28. Mergen M, Graf N, Meyerheim M. Reviewing the current state of virtual reality integration in medical education—a scoping review. BMC Med Educ 2024 Jul 23;24(1):788. [doi: 10.1186/s12909-024-05777-5] [Medline: 39044186]
29. Riches S, Iannelli H, Reynolds L, Fisher HL, Cross S, Attoe C. Virtual reality-based training for mental health staff: a novel approach to increase empathy, compassion, and subjective understanding of service user experience. Adv Simul (Lond) 2022 Jul 19;7(1):19. [doi: 10.1186/s41077-022-00217-0] [Medline: 35854343]
30. Hsu R, McCabe C, Ellison C, Howells M, Loetscher T. Bridging the gap: virtual reality as a tool for de-escalation training in disability support settings. Multimodal Technol Interact 2024;8:100. [doi: 10.3390/mti8110100]
31. Ahmad M, Tran MT, Ahmad B, Kavallieros K, Shalhoub J, Davies AH. Virtual reality versus simulation in the management of trauma based scenarios—a systematic review. Health Sci Rep 2024 Dec;7(12):e70216. [doi: 10.1002/hsr2.70216] [Medline: 39633831]
32. Miron A. Virtual reality (VR) as medical prevention of adverse childhood experiences (aces). In: Research Anthology on Pediatric and Adolescent Medicine: IGI Global; 2022:450-467. [doi: 10.4018/978-1-6684-5360-5.ch024]
33. Kim YK, Lee D, Oh SC, Park JH. The applicability of virtual reality and augmented reality to develop intervention programs to reduce aggressive behaviors in children based on a literature review. Korean J Child Psychother 2024 Mar 31;19(1):17-33. [doi: 10.23931/kacp.2024.19.1.17]
34. Des Jarlais DC, Lyles C, Crepaz N, TREND Group. Improving the reporting quality of nonrandomized evaluations of behavioral and public health interventions: the TREND statement. Am J Public Health 2004 Mar;94(3):361-366. [doi: 10.2105/ajph.94.3.361] [Medline: 14998794]
35. Hoffmann TC, Glasziou PP, Boutron I, et al. Better reporting of interventions: Template for intervention description and replication (TIDieR) checklist and guide. BMJ 2014 Mar 7;348:g1687. [doi: 10.1136/bmj.g1687] [Medline: 24609605]
36. Marshall RJ, Jackson RT. Analysis of case-crossover designs. Stat Med 1993 Dec 30;12(24):2333-2341. [doi: 10.1002/sim.4780122409] [Medline: 8134736]
37. Safe4Child. URL: https://safe4child.turkuamk.fi/ [accessed 2026-04-02]

XSL•FO
RenderX

38. Groen G, Jörns-Presentati A, Goodwin J, et al. Safe4Child: Trauma-informed approaches to support staff working with children and adolescents exhibiting behaviours that challenge. Standp Sozial 2025;35:1-13. [doi: 10.15460/spsoz.2025.35.1.240]
39. Goodwin J, O'Malley M, O'Donovan A, et al. Staff experiences of working with children and adolescents engaging in behaviours that challenge in mental health and paediatric inpatient environments: a qualitative exploratory study across four European countries. Child Care in Practice 2025 Jul 3;31(3):380-400. [doi: 10.1080/13575279.2024.2407085]
40. Baker CN, Brown SM, Wilcox PD, Overstreet S, Arora P. Development and psychometric evaluation of the attitudes related to trauma-informed care (ARTIC) scale. School Ment Health 2016 Mar;8(1):61-76. [doi: 10.1007/s12310-015-9161-0]
41. Cerny S, Berg-Poppe P, Anis M, Wesner C, Merrigan M, LaPlante K. Outcomes from an interprofessional curriculum on trauma-informed care among pediatric service providers. J Interprof Care 2023;37(2):288-299. [doi: 10.1080/13561820.2022.2070142] [Medline: 35687015]
42. Stewart SL, Cloutier S, King G, Withers A. Evaluating a trauma-informed care training program for mental health clinicians. J Child Adolesc Trauma 2024 Sep;17(3):981-998. [doi: 10.1007/s40653-024-00639-0]
43. Wholeben M, Castro Y, Salazar G, Field C. Impact of trauma-informed care training on attitudes among emergency department personnel, staff advocates, and nursing students. J Trauma Nurs 2023;30(5):261-270. [doi: 10.1097/JTN.0000000000000740] [Medline: 37702727]
44. Mendez A, Bosk EA, Keller A, et al. Expanding the trauma-informed care measurement toolkit: an evaluation of the attitudes related to trauma-informed care (ARTIC-45) scale with SUD workers in PIMH. Behav Sci (Basel) 2023 Jun 5;13(6):471. [doi: 10.3390/bs13060471] [Medline: 37366724]
45. Keesler JM, Wu W, Presnell J, Fukui S. Trauma-informed climate scale-10 and attitudes related to trauma-informed care-10: an examination of scale psychometrics using data from disability service providers. Disabil Health J 2024 Jul;17(3):101583. [doi: 10.1016/j.dhjo.2024.101583] [Medline: 38278724]
46. Keesler JM, Samways B, McNally P. Trauma-informed care training: a systemwide process to transform the act of caring. Adv Neurodev Disord 2025 Sep;9(3):494-506. [doi: 10.1007/s41252-025-00449-x]
47. Bürkner PC. brms: an R package for Bayesian multilevel models using Stan. J Stat Softw 2017;80(1):1-28. [doi: 10.18637/jss.v080.i01]
48. World Medical Association Declaration of Helsinki: ethical principles for medical research involving human subjects. JAMA 2013 Nov 27;310(20):2191-2194. [doi: 10.1001/jama.2013.281053] [Medline: 24141714]
49. Berring LL, Holm T, Hansen JP, Delcomyn CL, Søndergaard R, Hvidhjelm J. Implementing trauma-informed care-settings, definitions, interventions, measures, and implementation across settings: a scoping review. Healthcare (Basel) 2024 Apr 27;12(9):908. [doi: 10.3390/healthcare12090908] [Medline: 38727465]
50. Davis PA, Hubbard D, Gladdis T, et al. Improving attitudes toward trauma-informed care in the neonatal intensive care unit through comprehensive multi-disciplinary education. J Perinatol 2024 May;44(5):650-658. [doi: 10.1038/s41372-024-01897-4] [Medline: 38383575]
51. Loomis AM, Musson Rose D, Mitchell J. Closing gaps in care: effects of CARE training on early childhood provider trauma-informed attitudes and collaboration. J Child Fam Stud 2024 Mar;33(3):924-938. [doi: 10.1007/s10826-023-02720-y]
52. Tay JL, Xie H, Sim K. Effectiveness of augmented and virtual reality-based interventions in improving knowledge, attitudes, empathy and stigma regarding people with mental illnesses-a scoping review. J Pers Med 2023 Jan 4;13(1):112. [doi: 10.3390/jpm13010112] [Medline: 36675773]
53. Holla M, van den Berg M. Virtual reality techniques for trauma education. Injury 2022 Nov;53:S64-S68. [doi: 10.1016/j.injury.2022.08.067] [Medline: 36109249]
54. Matto HC, Ihara ES, Cieslowski B, Hines DA, Booth J. Virtual reality case simulation to build skill competencies in working with substance-engaged clients. Soc Work Educ 2024 Dec 20;43(9):2774-2786. [doi: 10.1080/02615479.2023.2289452]

Abbreviations

ARTIC-10: 10-item version of the Attitudes Related to Trauma-Informed Care
GDPR: General Data Protection Regulation
HAW: Hamburg University of Applied Sciences
MUP: Medical University of Plovdiv
ROPE: region of practical equivalence
TIC: trauma-informed care
TIDieR: Template for Intervention Description and Replication
TREND: Transparent Reporting of Evaluations with Non-Randomized Designs
TUAS: Turku University of Applied Sciences
VR: virtual reality

XSL•FO
RenderX

Edited by T Gladman; submitted 29.Oct.2025; peer-reviewed by C Bausha, MC Runyon; revised version received 05.Mar.2026; accepted 05.Mar.2026; published 17.Apr.2026.

Please cite as:
Kostadinov K, Goodwin J, Groen G, Jörns-Presentati A, Heffernan S, O'Donovan Á, O'Malley M, Biskupovic Goulding R, O'Mahony J, Allen S, Curtin M, Haapalainen S, Aho J, Petrova G, Ivanova S, Lalova V, Korhonen J
Virtual Reality–Enhanced Training for Trauma-Informed Care Among Residential and Child Mental Health Professionals: Pre-Post Evaluation Study
JMIR Med Educ 2026;12:e86543
URL: https://mededu.jmir.org/2026/1/e86543
doi:10.2196/86543

© Kostadin Kostadinov, John Goodwin, Gunter Groen, Astrid Jörns-Presentati, Sinéad Heffernan, Áine O'Donovan, Maria O'Malley, Ryan Biskupovic Goulding, James O'Mahony, Stephanie Allen, Margaret Curtin, Satu Haapalainen, Jami Aho, Gergana Petrova, Svetla Ivanova, Valentina Lalova, Joonas Korhonen. Originally published in JMIR Medical Education (https://mededu.jmir.org), 17.Apr.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Education, is properly cited. The complete bibliographic information, a link to the original publication on https://mededu.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Perspectives of Premedical Students at a Medical University College in Malaysia Regarding Augmented Reality Integration in Anatomy Education: Cross-Sectional Study

Fadhilah Idris[1], BSc, MSc; Nur Hidayah Mohd Noor[1], BSc, MSc; Faidruz Azura Jam[2], BSc, MSc, PhD; Nurul Huda Mohd Nor[3], MD, PhD; Siti Fadziyah Mohamad Asri[4], MD, PhD

[1]Centre for Foundation and General Studies, Manipal University College Malaysia, Persimpangan Batu Hampar, Bukit Baru, Malacca, Malaysia

[2]Department of Biochemistry, Faculty of Medicine, Manipal University College Malaysia, Malacca, Malaysia

[3]Department of Human Anatomy, Faculty of Medicine and Health Sciences, Universiti Putra Malaysia, Selangor, Malaysia

[4]Department of Anatomy and Physiology, Faculty of Medicine, Universiti Sultan Zainal Abidin, Terengganu, Malaysia

Corresponding Author:
Fadhilah Idris, BSc, MSc
Centre for Foundation and General Studies, Manipal University College Malaysia, Persimpangan Batu Hampar, Bukit Baru, Malacca, Malaysia

Abstract

Background: For novice anatomy learners, studying human anatomy using textbooks, 2D learning materials, and static anatomical models frequently causes challenges in understanding complex anatomical structures. Since access to dissected human donor bodies is limited in many premedical programs, researchers are concerned with exploring novel supplementary approaches to anatomy learning. This research explores the effectiveness of an augmented reality (AR) app in enhancing the anatomy learning experiences of premedical students.

Objective: This cross-sectional study was conducted to evaluate premedical students' perceptions of AR integration in early anatomy learning and to examine whether these perceptions differ according to the students' demographic variables.

Methods: An AR app was introduced to premedical students during anatomy practical sessions. Participants were divided into groups of 4, and a volunteer from each group wore a garment that displayed dynamic anatomical visuals when scanned with a smartphone or tablet. Participants then evaluated their anatomy learning experiences using a survey questionnaire after using the AR app for 1 hour.

Results: A total of 284 participants were included in the study. Based on the participants' ratings of the anatomy learning app with AR, the overall perception score was high (median 4.92, IQR 4.17 - 5.00). A statistically significant difference in perception score was observed between male and female students (ρ=0.212; P<.001), with female students reporting more favorable opinions of the AR app. However, no statistically significant associations were observed between perception and age (ρ=0.034; P=.57), student batch number (ρ=-0.009; P=.88), or monthly household income (ρ=−0.041; P=.49).

Conclusions: By providing interactive learning experiences, AR is positively perceived among diverse premedical student backgrounds; thus, it has broad potential to serve as a valuable supplementary tool to complement conventional teaching practices for novice anatomy learners.

(JMIR Med Educ 2026;12:e64402) doi:10.2196/64402

KEYWORDS

premedical students; augmented reality; AR; interactive learning; anatomy education

Introduction

Human anatomy education is an essential component of medical training, providing the foundational knowledge necessary for clinical practice and diagnosis. However, traditional methods of teaching anatomy, primarily reliant on 2D textbook images, photographic projections of lecture slides, static anatomical models, and flashcards, often fail to effectively convey the complex 3D relationships between structures [1-3]. While human donor body dissection remains the gold standard for anatomy training in medical schools [4,5], donor body dissection has several drawbacks, such as restricted access [6-8], time constraints [9], financial limitations [9,10], and ethical concerns [10-12]. These challenges are particularly prominent in the context of foundation-level students in Malaysia, where public medical schools face constraints in providing adequate dissection facilities and incorporating this method into early anatomy courses [6]. Although private institutions may have more

XSL•FO
RenderX

resources, sourcing human donor bodies and integrating dissection into premedical curricula remains difficult in many cases.

In the last few years, many educational institutions have progressively explored technology-enhanced learning approaches to supplement traditional anatomy instruction. Following the COVID-19 pandemic, digital learning tools have gained prominence with the accelerated adoption of technology in higher educational institutions [13,14]. Even though online platforms such as learning management systems and video-conferencing tools like Zoom (Zoom Communications, Inc), Microsoft Teams, and Google Meet enabled the continuation of teaching and learning sessions during the COVID-19 pandemic [15,16], as well as postpandemic periods, these platforms mostly supported content delivery rather than interactive visualization of anatomical structures. Consequently, educators have sought advanced digital tools that can improve spatial learning and enhance student engagement in anatomical education.

Among the myriad of emerging technologies, augmented reality (AR) has stood out as a technology that enhances students' engagement and understanding of complex scientific concepts. AR incorporates computer-generated content into real-world elements, permitting users to experience immersive learning environments that overcome the limitations of traditional approaches [17]. In the context of anatomy education, AR allows students to visualize 3D representations of various human body systems and internal organs without the requirement for physical specimens. This capability enables students to explore anatomical structures dynamically through multiple functions such as zooming, rotation, and layered visualization, potentially enhancing spatial understanding and conceptual comprehension [18].

Another significant benefit of AR technology over other extended reality technologies, such as virtual reality and mixed reality, is its accessibility through personal devices such as smartphones and tablets [19]. This permits AR-based learning approaches to be incorporated not only into classroom-based but also into self-directed learning practices [20], without requiring expensive, specialized, or tethered headsets, which are crucial for virtual reality and mixed reality technologies. Previous studies have reported that AR-based learning apps can improve student engagement, interest, and understanding of complex scientific concepts [19,21,22]. Consequently, AR is progressively considered an effective supplementary tool that can support traditional anatomy teaching methods.

However, the effectiveness of AR in enhancing anatomy education, especially in the context of premedical students, remains underexplored. A conceptual framework grounded in educational theories for AR-powered teaching is essential to assess its effectiveness in improving student learning outcomes. This study is informed by the perspectives of Kirkwood and Price's [23] review, which conceptualizes enhancement in relation to the nature of technology intervention, how improvement in learning is defined (operational, quantitative, or qualitative), and the level of evidence used to support such claims. Within this framework, AR is considered to enhance learning primarily through qualitative improvements in the learner experience, including improved spatial visualization, interactivity, and engagement with anatomical content. Accordingly, this study focuses on the learner experience dimension and operationalizes enhancement through students' perceived educational value of AR, treated as a unified construct reflecting usability, engagement, and perceived learning support. The use of self-reported perception is consistent with prior technology-enhanced learning research as an appropriate early-stage evaluative approach, particularly for interventions that supplement existing teaching practices. However, this represents a lower level of evidence and is subject to inherent limitations, including potential response bias and the absence of direct measurement of learning outcomes.

Although AR is increasingly used in medical education, there is still limited understanding of how premedical students experience and perceive its integration into early anatomy learning, particularly in resource-constrained settings where traditional hands-on approaches are limited. Hence, this study aimed to evaluate premedical students' perceptions of AR integration in early anatomy learning, conceptualized as a single perception construct. Specifically, we asked whether students view AR as an acceptable and useful aid tool for learning anatomy and whether this perception varies across demographic variables. This was examined using a cross-sectional study with a structured and validated questionnaire. We hypothesized that students would generally report a positive perception of AR integration regardless of demographic background. The findings from this study may inform curriculum design and guide evidence-based decisions for educators, curriculum designers, and academic administrators regarding the integration of AR in anatomy education, particularly in settings with limited access to conventional anatomical teaching resources.

Methods

Ethical Considerations

Ethical approval was obtained from the Research Ethics Committee of Manipal University College Malaysia (MUCM; 035/2023) prior to the commencement of the research study. This study met the criteria given its educational nature with no potential risks to participants. To ensure the highest standards of ethical research in educational contexts, the study incorporated comprehensive measures in accordance with established guidelines. Informed consent was acquired from the participating students or their parents or legal guardians to ensure their understanding of the purpose of the study, its procedures, any side effects, and any benefits of being in this study. To further ensure the confidentiality and anonymity of the participants, no personal identifiers were collected in this study. All data were recorded anonymously, and access to the collected dataset was restricted to authorized members of the research team for data management and data analysis purposes. Participation in this study was voluntary, and participants had the right to refuse to answer any questions in the questionnaire. Participants also had the right to withdraw from this study at any time, and refusal to participate or withdrawal from this study did not affect any benefits to which they were otherwise

XSL•FO
RenderX

entitled. By implementing these ethical considerations, the research study has not only protected the rights, safety, and well-being of the participants but also upheld the credibility and integrity of the research findings.

Participants

This research study was conducted on premedical students attending a Basic Human Anatomy course in semester 2 of the Foundation in Science (FiS) program at MUCM. The study population comprised students enrolled in the FiS program at the medical university college during the 2023/2024 academic year, including FiS Batch 27, FiS Batch 28, and FiS Batch 29. The term "Batch" refers to the student cohort within the premedical program, typically categorized according to the intake period (ie, entry group). A total of 3 student cohorts or batches from the 2023/2024 academic year were included in this study, whereby FiS Batch 27, FiS Batch 28, and FiS Batch 29 enrolled in the premedical program in April 2023, July 2023, and August 2023, respectively. Upon completion of this 1-year FiS program, the students will pursue their undergraduate studies in MBBS in the upcoming academic year. Before the students' enrollment in this FiS program, these students either completed the International General Certificate of Secondary Education or the Malaysian Higher School Certificate, and thus, they had basic science knowledge of biology, chemistry, and physics.

One of the compulsory courses in semester 2 of the FiS program is Basic Human Anatomy, which is designed to enhance the students' basic knowledge of the human body. The topics covered in this course include an introduction to anatomy and the human body, the introduction of different types of human body tissues and their histological features, and various topics in systemic anatomy, such as the skeletal, nervous, respiratory, cardiovascular, digestive, and urinary systems. The methods of content delivery were mainly didactic lectures, tutorials, and practical sessions facilitated by qualified instructors.

During week 3 of the Basic Human Anatomy course, after the introduction to anatomy and the human body, as well as the introduction of different types of human body tissues and their histological features, topics had been delivered, the participants were introduced to the Virtuali-Tee AR app in focused sessions. These were formal practical sessions, and the participants were guided by qualified instructors to interact with the AR app in groups for 1 hour to promote collaborative learning.

All the FiS student batches received the same curriculum and teaching content and participated in the same AR-based practical session, delivered using a standardized protocol, ensuring consistency and comparability across student cohorts despite variations in their intake periods.

Sample Size

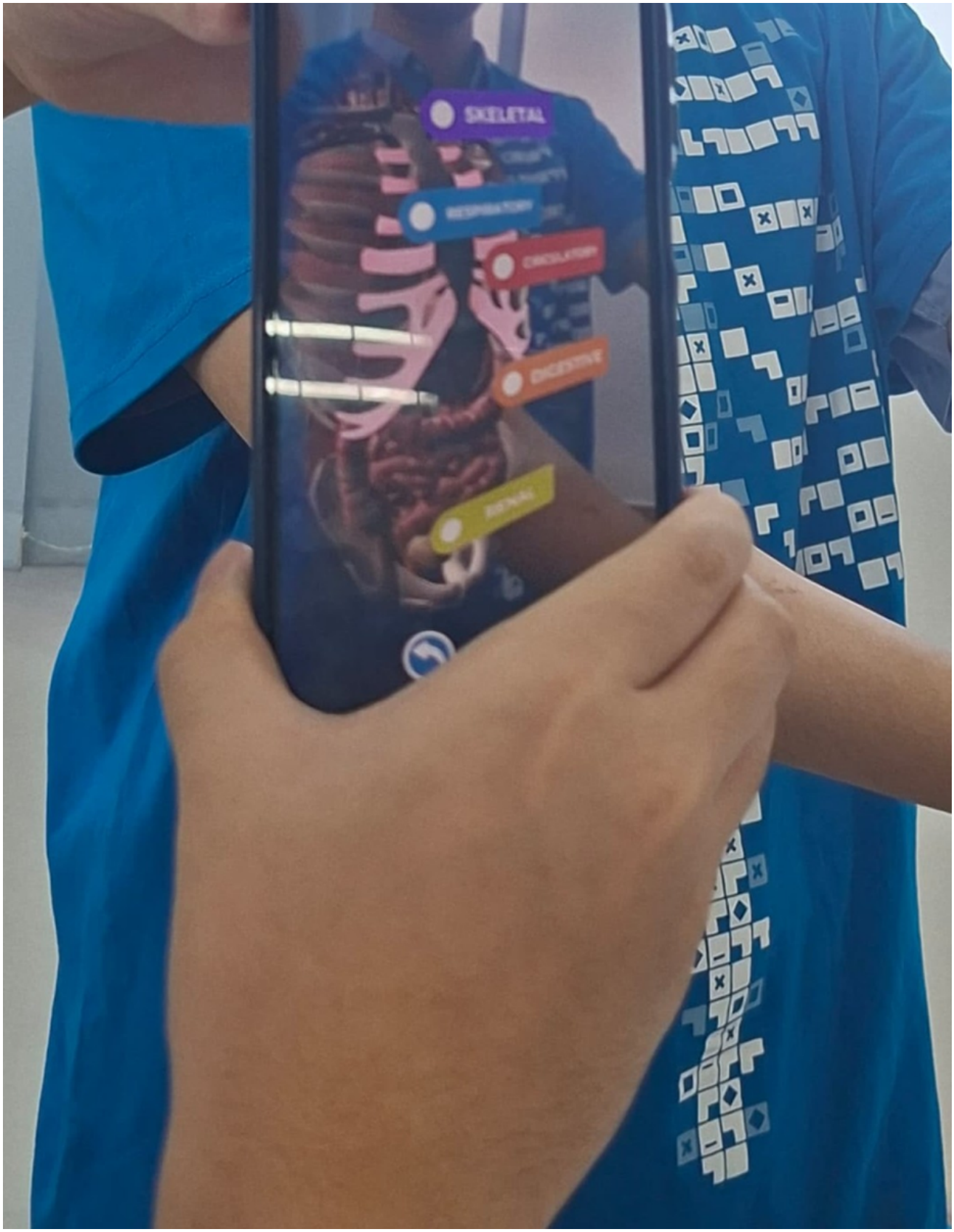
This cross-sectional study was conducted among premedical students enrolled at MUCM during the 2023/2024 academic year, with a total eligible population of 293 students. The minimum required sample size was determined using the formula for estimating finite population proportion with an assumption of a 95% CI, a margin of error of 5%, and a maximum response distribution of 50%. This calculation yielded a required sample size of 167 respondents. The researchers anticipated a 30% nonresponse rate; therefore, the final adjusted sample size was set at 239 students. In this study, a total of 284 students voluntarily participated, exceeding the sample size requirement and thereby ensuring adequate statistical power for data analysis. The larger sample size also allows for more robust subgroup analyses across demographic variables, including age, gender, student batch number, and monthly household income.

Study Design

This study involved researchers, who are experienced and qualified instructors, in introducing premedical students to the Virtuali-Tee AR app, which was developed by Curiscope (Figure 1). This app has been purchased by MUCM and is available on the App Store and Google Play Store. The researchers instructed the participants to install the AR app on their smartphones or tablets during the AR-based practical sessions. These sessions were carried out with structured learning objectives, including: (1) identification of basic anatomical structures, (2) visualization of 3D anatomical relationships, and (3) familiarization with the AR app's functional features (eg, rotation, zooming, and organ description). These standardized sessions commenced with a brief orientation given by the researchers, followed by a guided demonstration of the AR app's functionality. The researchers then supervised the participants as they engaged in hands-on exploration of the AR app within a 1-hour time frame. During this time frame, a few volunteers among the participants wore a piece of garment that, when scanned with smartphones or tablets, superimposed anatomical graphics and visuals on their bodies. The researchers were present throughout the sessions to guide the participants and ensure alignment with the learning objectives of the sessions. For the Virtuali-Tee AR app, the garment integrated a unique pattern of a ribcage as a marker, enabling the app to track participants' body positions and orientations accurately. This ensured that the overlaid 3D anatomical visuals remained precisely aligned with the participant's body movements in real-time, while providing a unique and immersive educational experience.

XSL•FO
RenderX

Figure 1. Upon scanning the garment with smartphones or tablets, the Virtuali-Tee augmented reality (AR) app reveals interactive anatomical graphics and visuals.



In the Virtuali-Tee AR app settings, participants had the option to adjust sound volume, activate subtitles, and choose their preferred speech and text language (Figure 2). Participants were able to interact with various features within the app, allowing them to navigate and explore multiple systems of the human body, namely the skeletal, respiratory, circulatory, digestive, and urinary systems (Figure 3), by using their personal devices. This app enabled participants to locate the positions of internal organs (Figure 4) and perform gross observations of the organs from different angles. By clicking on the 3D organ models

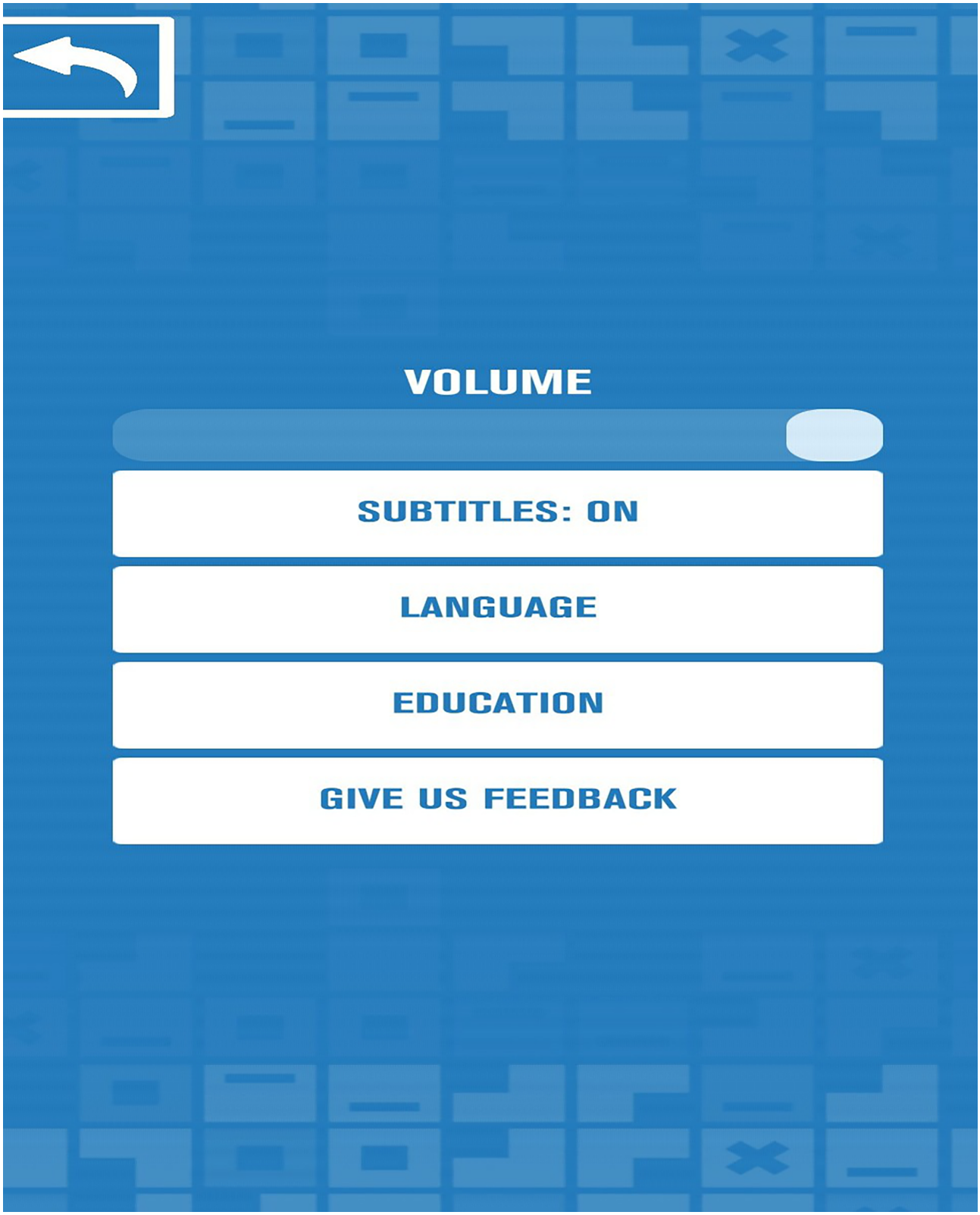
XSL•FO
RenderX

featured in the app, comprehensive information about the organs appeared. The 360° view function embedded in the app enabled participants to explore detailed structures and experience immersive virtual learning as they examined the internal wall of the small intestine (Figure 5), the circulation of blood (Figure 6), and the surface of alveoli at the microscopic level.

The Virtuali-Tee AR content was designed to provide precise depictions of internal anatomical structures by focusing primarily on the thoracic and abdominal regions while avoiding inappropriate details in other sensitive regions of the body. Additionally, this app required active participation from all participants, with a volunteer from each group choosing to wear the garment and providing informed consent. These measures aligned with ethical guidelines in educational technology, ensuring that the use of the Virtuali-Tee AR garment respected participants' dignity and promoted a safe teaching and learning environment. Following the 1-hour AR-guided practical sessions, the researchers distributed an online survey questionnaire to the participants to evaluate their anatomy learning experiences with the AR app.

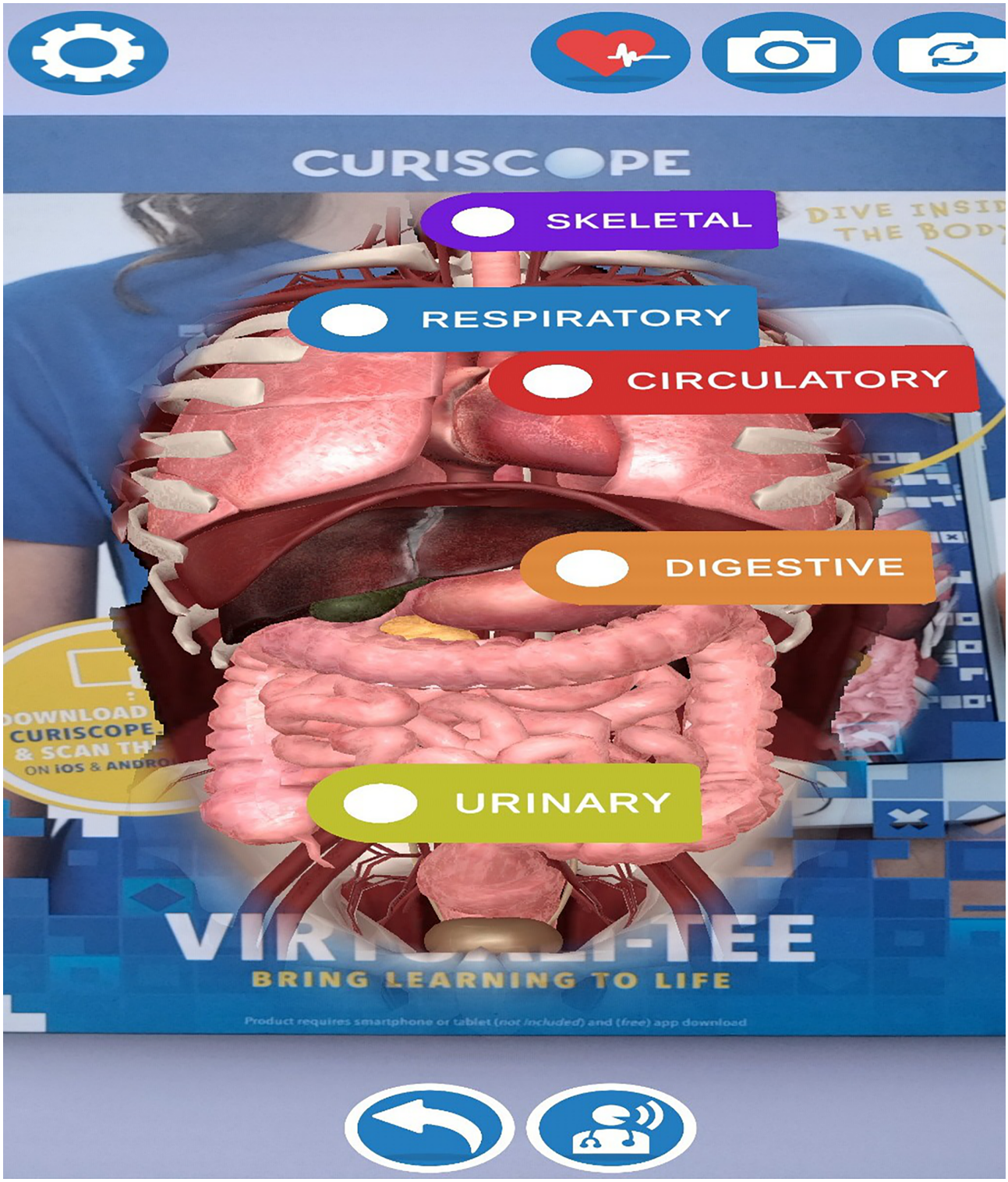
XSL•FO
RenderX

Figure 2. The settings section of the app enabled the participants to adjust the sound volume, activate subtitles, and choose their preferred language.



XSL•FO
RenderX

Figure 3. The 5 different human body systems are further detailed as participants clicked on the various interactive functions.



XSL•FO
RenderX

Figure 4. Various organs of the circulatory system were further explained as participants clicked on the 3D organ models featured in the app.



XSL•FO
RenderX

Figure 5. The 360-degree view function enabled the participants to visualize the inner wall of the small intestine at the microscopic level.



XSL•FO
RenderX

Figure 6. Participants had the opportunity to virtually dive into the blood circulation, enabling them to observe different blood components and formed elements.



Data Collection

The data were collected through an online survey questionnaire administered to premedical students who participated in the research study. The survey questionnaire, hosted on Google Forms, was shared with participants electronically. Responses were collected online and stored in a secure online database. No manual data entry was performed. Both questionnaire items and response options were randomized using the built-in shuffle functions in Google Forms. There were 2 sections in the questionnaire, which consisted of 5 pages. The first section gathered demographic information such as age, gender, student batch number, and monthly household income. In the second section, a questionnaire adapted from a previous study [1] was used. Data completeness was assessed both during and after survey submission. As a dynamic completeness check, each question required a response before the respondent could proceed to the next page. While responding to the questionnaire,

XSL•FO
RenderX

the back button and response editing were enabled. The survey was open from September 6, 2023, to January 9, 2024. Participation was voluntary, and participants could exit the survey at any point before submission.

Online Survey Administration

The online survey was reported in accordance with the CHERRIES (Checklist for Reporting Results of Internet E-Surveys) guidelines (Checklist 1) [24]. The questionnaire was administered using Google Forms as a closed survey and was accessible only to eligible premedical students through an electronic invitation sent by the researchers. Prior to accessing the questionnaire, participants were presented with an information page describing the study's purpose, voluntary participation, confidentiality of responses, and estimated completion time. Proceeding with the survey implied informed consent. No incentives were offered for participation. To prevent multiple submissions, responses were restricted to a single entry per participant.

Data Collection Tool

The questionnaire used in this study was adapted from a previously published survey instrument [1]. The questionnaire comprised 12 items rated on a 5-point Likert scale, ranging from 1 ("strongly disagree") to 5 ("strongly agree"). The researchers measured the construct validity and internal consistency reliability of the adapted questionnaire to assess its psychometric properties within this study population.

The construct validity was evaluated using exploratory factor analysis. The exploratory factor analysis demonstrated that all items in the survey instrument had high communality values (>0.80), suggesting that a substantial proportion of the variance in each item was shared with the extracted factor. This important finding indicated that the items in the survey instrument predominantly measured a single latent construct rather than multiple distinct domains. Based on these important findings, the survey instrument was treated as a single construct in the subsequent data analyses.

On the other hand, the researchers evaluated the internal consistency reliability using Cronbach α. The overall Cronbach α value was 0.96, indicating high internal consistency and reliability, and further supporting the interpretation that the questionnaire items collectively measured a 1D construct.

Additional forms of validity testing, including predictive and discriminant validity, were not performed. The researchers decided this because the survey instrument was intended to capture students' self-reported perceptions rather than to predict objective learning outcomes or differentiate between theoretically distinct constructs.

Data Analysis

The researchers included only the fully completed questionnaires in the final analysis. Of the 293 eligible students, 284 completed the online questionnaire, yielding a completion rate of 96.9%. Data analysis was conducted using IBM SPSS Statistics version 29.0. For the first section of the questionnaire, descriptive statistics were performed on the demographic data of the study participants, and the data were reported as frequencies and percentages. In the second section, a Likert scale ranging from 1 ("strongly disagree") to 5 ("strongly agree") was used to assess participants' perceptions of anatomy learning with the AR app. Participants indicated their level of agreement or disagreement with the statements using this scale. The researchers conducted descriptive analysis to calculate the median and IQR as the data were derived from Likert-scale responses and were not normally distributed. An overall perception score was also computed by averaging individual item responses, from which the median and IQR were calculated. Then, inferential analysis was conducted using the Spearman rank-order correlation to explore the association between participants' demographic variables and the overall perception score of the AR learning experiences. Spearman ρ was applied for all correlations since this test handles nonnormally distributed and ordinal data better, particularly in Likert scale-based studies. All the statistical tests were 2-sided, and a significance level of $P<.05$ was adopted.

Results

Participants' Demographic

The results of the first section of the questionnaire, which gathered demographic information on the participants, are presented in Table 1. A total of 284 premedical students participated in the study, with 148 (52.1%) participants aged older than 18 years, 192 (67.6%) being female, and 60 (21%) participants having a monthly household income between MYR 1000 and MYR 3000 (US $251.95-US $755.86). Furthermore, 151 (53.2%) participants were from FiS Batch 29, who enrolled in the medical university college in August 2023.

XSL•FO
RenderX

Table . Participants' demographic information (N=284).

Variable	Value, n (%)
Age (y)	
15 - 16	2 (0.7)
17 - 18	134 (47.2)
>18	148 (52.1)
Gender	
Male	92 (32.4)
Female	192 (67.6)
Student batch number	
FiS[a] Batch 27	36 (12.7)
FiS Batch 28	97 (34.2)
FiS Batch 29	151 (53.2)
Monthly household income (MYR)	
<1000 (<US $251.95)	38 (13.4)
1000-3000 (US $251.95-US $755.86)	60 (21.1)
3001-5000 (US $755.87-US $1259.76)	42 (14.8)
5001-7000 (US $1259.77-US $1763.67)	53 (18.7)
7001-10,000 (US $1763.68-US $2519.53)	48 (16.9)
>10,000 (>US $2519.53)	43 (15.1)

[a]FiS: Foundation in Science.

Evaluation Outcomes

An analysis of the results gathered from the second section of the questionnaire is presented in Table 2. The researchers analyzed the questionnaire items as a unified construct, and due to the ordinal nature and nonnormal distribution of the data, the participants' responses were summarized using the median and IQR. For all items, the median score was 5.00 (IQR 4.00 - 5.00), suggesting generally positive responses among respondents. For each participant, the composite mean of all Likert-scale items was used to calculate the overall perception score through the calculation of the median and IQR. The computed overall perception score reflected a high level of agreement, with a median of 4.92 (IQR 4.17 - 5.00). These important findings suggest that the premedical students perceived the AR app as an interactive, engaging, and accessible supplementary tool for anatomy learning.

XSL•FO
RenderX

Table . Evaluation of outcomes of the Likert scale[a].

Number	Items	Value, median (IQR)
1	I think the "anatomy learning app with AR[b]" is very interesting.	5.00 (4.00 - 5.00)
2	I think "anatomy learning app with AR" is very helpful in learning human anatomy materials.	5.00 (4.00 - 5.00)
3	In my opinion, "anatomy learning app with AR" is very usable in other fields.	5.00 (4.00 - 5.00)
4	I think "anatomy learning app with AR" is helpful in visualizing the anatomy of the human body.	5.00 (4.25 - 5.00)
5	I think "anatomy learning app with AR" is helpful in understanding more deeply about the anatomy of the human body.	5.00 (4.00 - 5.00)
6	I think the "anatomy learning app with AR" is faster in understanding the anatomy of the human body.	5.00 (4.00 - 5.00)
7	I think the function of "anatomy learning app with AR" to see the description of each organ is very helpful and interesting.	5.00 (4.00 - 5.00)
8	I think the function of "anatomy learning app with AR" to rotate 3D model is very beneficial in anatomical observation.	5.00 (4.00 - 5.00)
9	I think the function of "anatomy learning app with AR" to enlarge 3D model is very beneficial to see anatomy details.	5.00 (4.00 - 5.00)
10	I think learning by "anatomy learning app with AR" is very practical in anatomy class.	5.00 (4.00 - 5.00)
11	I am very interested in using "anatomy learning app with AR" anytime.	5.00 (4.00 - 5.00)
12	Overall "anatomy learning app with AR" greatly aided the learning process.	5.00 (4.00 - 5.00)

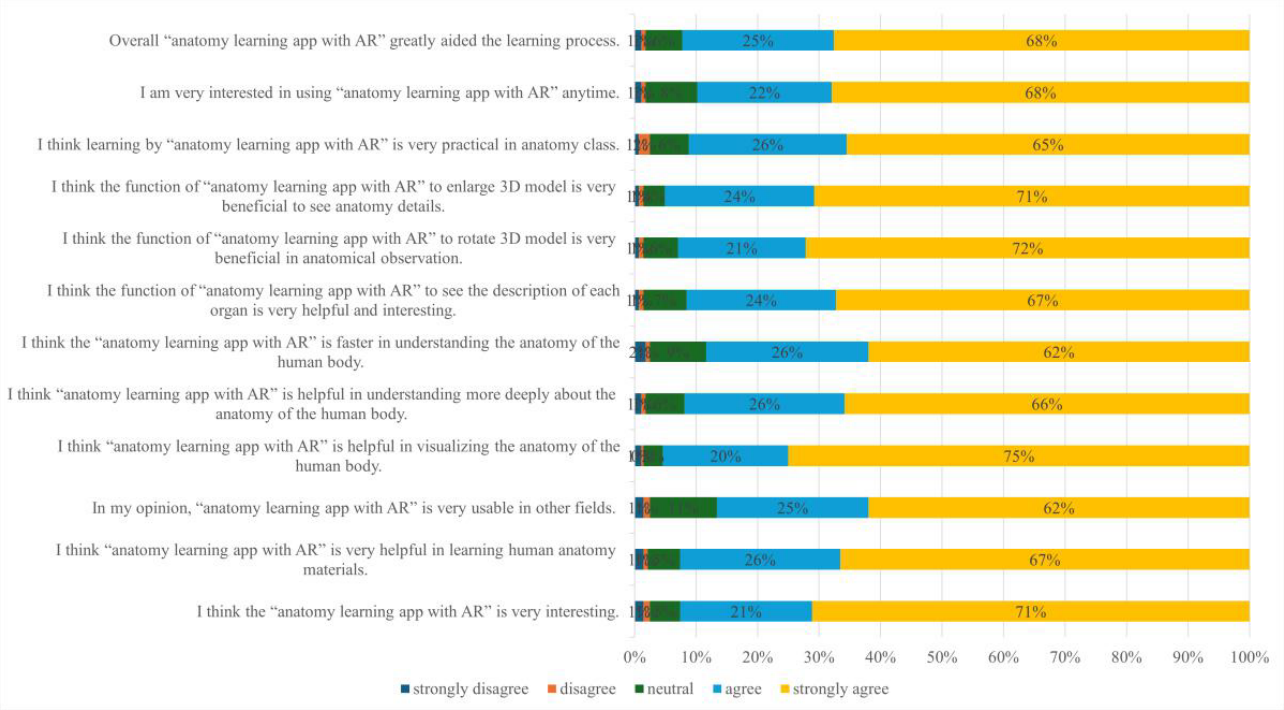
[a]Overall median perception score was 4.92 (IQR 4.17 - 5.00).

[b]AR: augmented reality.

Figure 7 encapsulates the results from the 12 explicit statements which were responded to using the 5-point Likert scale. Overall, participants evaluated the anatomy learning experiences with AR positively. From a total of 284 participants, 202 (71%) strongly agreed, and 61 (21%) agreed that the anatomy learning app with AR is very interesting. Additionally, 213 (75%) participants strongly agreed, and 58 (20%) agreed that anatomy learning apps with AR are helpful in visualizing the anatomy of the human body. Furthermore, 205 (72%) participants strongly agreed, and 59 (21%) agreed that the function of the anatomy learning app with AR to rotate 3D models is very beneficial for anatomical observation. Finally, 201 (71%) participants strongly agreed, and 69 (24%) agreed that the function of the anatomy learning app with AR to enlarge 3D models is very beneficial for observing anatomical details.

XSL•FO
RenderX

Figure 7. Participants' perceptions of the anatomy learning app with augmented reality (AR). Different colors in the bars represent each distinct item on the Likert scale (1="strongly disagree," 2="disagree," 3="neutral," 4="agree," 5="strongly agree").



Based on participants' responses to the Likert scale, it was noted that both the median scores across individual questionnaire items and the overall perception score were at the upper end of the scale (item: median 5.00, IQR 4.00 - 5.00; overall: median 4.92, IQR 4.17 - 5.00), demonstrating that most participants selected option 4 ("agree") and option 5 ("strongly agree") in their responses.

Most participants agreed that the experiences of learning anatomy with AR were interesting and helpful in understanding anatomy materials and in visualizing the human body's anatomy. Participants also rated items related to perceived usefulness and engagement favorably, indicating positive impressions of the AR app in anatomy learning. Besides, participants appreciated the ability to rotate the models for observation from multiple angles and the zooming function of the 3D models. Participants also agreed that the AR app's practicality significantly assisted their overall learning process in the complex course of anatomy. These findings reflect participants' immediate, self-reported perception feedback following a short-term exposure to the AR app. It is vital to highlight that these findings do not represent objective measures of learning outcomes, efficiency, or the long-term educational impact of the AR app in anatomy education.

Correlations Between Demographic Variables and Overall Perception Scores

This section contains an analysis of the results obtained from the inferential statistics. Due to the ordinal nature of several demographic variables and the nonparametric distribution of the Likert scale data, the Spearman rank-order correlation was chosen for the inferential statistics. Table 3 presents a statistically significant correlation between gender and overall perception (ρ=0.212; *P*<.001), with female students reporting more favorable opinions of the AR app compared to male students. However, no statistically significant correlations were found between the overall perception score and age (ρ=0.034; *P*=.57), student batch number (ρ=−0.009; *P*=.88), or monthly household income (ρ=-0.041; *P*=.49), demonstrating that participants' perceptions of AR-based learning were not influenced by their demographic variables.

These results indicated that while participants' perception ratings of AR-based learning were mostly high across various demographic groups, female students had slightly more favorable responses compared to their male counterparts. Since no strong correlations existed with other demographic variables, this supports the comprehensive potential of AR-based learning in anatomy education.

Table . Correlation analysis (Spearman ρ and 2-tailed *P* value) between the overall perception scores and demographic information (N=284).

Variables	ρ	*P* value
Age (y)	0.034	.57
Gender	0.212[a]	<.001
Student batch number	−0.009	.88
Monthly household income (MYR)	−0.041	.49

[a]The correlation is significant at a significance level of .01 level (2-tailed).

XSL•FO
RenderX

Discussion

Principal Findings

Our findings demonstrate preliminary evidence that the integration of AR in early anatomy learning is positively perceived by premedical students. Following a structured 1-hour session, participants reported a favorable perception of the AR app as a potentially beneficial supplementary learning tool for anatomy education. These findings, however, are based on a single, short-term exposure and reflect students' perceived value rather than objective measures of learning outcomes or sustained educational effectiveness. Demographic correlation analysis revealed that the overall perception score was not significantly associated with age, student batch number, or monthly household income, suggesting that AR's positive reception was consistent across diverse students' academic backgrounds and socioeconomic statuses. Although more than 50% of respondents were from a single batch (ie, FiS Batch 29), this distribution reflected student cohort size and participation rates, and no significant association was observed between student batch number and overall perception score, indicating that this imbalance did not substantially influence the findings. It also reflects that variations in the intake period did not influence students' evaluation of the AR learning experience. However, it was noted that a statistically significant association existed between gender and perception, highlighting slight variations in how different gender subgroups interact with AR technology. Overall, these findings have answered the research question and supported the study hypothesis that premedical students perceive AR technology as a beneficial and acceptable pedagogical adjunct for anatomy learning, and this positive perception is universally shared and not strongly influenced by the premedical students' demographic factors.

Implications of Findings

From an educational viewpoint, this preliminary evidence supports the integration of AR as a supplementary tool rather than a substitution for traditional anatomy teaching methods such as lectures, textbook images, and static anatomical models. AR may be a particularly valuable approach in resource-constrained settings where access to human body donors or advanced laboratory facilities is limited, offering students an interactive means to explore the relationships between anatomical structures. Even though this study evaluated AR within structured learning sessions, the accessibility of AR via personal mobile devices suggests potential for self-directed and distance learning, enabling students to engage with interactive 3D anatomical models outside the classroom. However, optimal use may still benefit from facilitated and in-person integration, particularly for learners at an early stage of anatomy education. Collectively, these findings indicate that AR functions as a versatile and flexible digital tool that complements conventional teaching approaches while also offering opportunities to extend learning beyond traditional classroom settings.

Comparison With Literature

This study's results reinforce the potential of AR in anatomy education. The high perception ratings evident in this study align with earlier findings reporting favorable student opinions toward AR-based anatomy learning tools, particularly in terms of perceived usefulness, support for visualization of 3D structures, and increased student engagement [25-27]. However, like many preliminary assessments, these findings are based on brief exposure and student self-reported perceptions rather than objective measures of learning outcomes. In contrast to studies involving prolonged or curriculum-integrated use of AR, the present findings demonstrate initial impressions following a structured, short-term AR-based anatomy learning session.

AR can facilitate interactive, dynamic, and spatially immersive learning by transforming traditional educational approaches that have long relied on static 2D representations and conventional modeling techniques. AR's interactive nature significantly enhances visual learning experiences by allowing students to observe intricate processes in detail [28]. AR provides clear depictions of complex anatomical structures by offering detailed visualizations in 3D and enhancing their layout and functions. This capability aligns with the need for comprehensive teaching and learning materials that offer in-depth anatomical descriptions [29]. Students expressed confidence in the effectiveness of learning through 3D models, including those involving AR, for grasping anatomical concepts [30]. The high perception ratings in this study reflect AR's potential to make learning more appealing, thereby boosting students' enthusiasm and motivation. As a result, students who engage with AR apps are more likely to attain improved knowledge retention and practical competency [31,32], which are critical for ensuring safe clinical practice.

One of the standout benefits of using AR as an educational tool is its multimedia features. Traditional anatomy educational methods frequently struggle to efficiently convey the intricate spatial relationships [33], often leaving students with difficulties understanding 3D inter-relationships between anatomical structures [30,34,35]. AR addresses this challenge by improving spatial reasoning, thus allowing students to visualize abstract scientific phenomena [36], that are otherwise challenging to replicate in conventional settings [37]. Historically, anatomy educators relied on verbal explanations and static illustrations from textbooks and atlases, expecting students to imagine these structures in real-time. AR allows students to scrutinize intricate structures such as nerves and blood vessels in isolation, which is a task challenging to achieve with traditional human donor body dissections, as these structures often lose their form when separated from supporting tissues [38].

The interactive capabilities of AR foster deeper engagement and enhance visual and spatial thinking [39]. The ability to display animated 3D models offers a significant advantage over traditional text and static illustrations. The interactive nature of AR supports real-time exploration and detailed examinations, which is particularly beneficial for premedical students who need to memorize complex anatomical information. AR enables learners to visualize and manipulate intricate concepts in real time by overlaying digital content onto the physical environment [40-43]. AR apps, designed with features like assistive touch, enable students to enlarge or rotate specific organs in the human body. AR also facilitates the acquisition of knowledge about numerous anatomical structures by allowing students to choose

XSL•FO
RenderX

specific regions or structures of interest [44]. A study used a mobile AR app developed by the authors to enable users to dynamically explore various parts of the human body in 3D, enhancing their overall educational experience [45]. Likewise, a mobile app with AR designed to enhance learning of the musculoskeletal system received positive evaluations from users. The findings demonstrated that the anatomy system could efficiently improve retention of anatomical concepts and students' engagement [46].

One of the key strengths of AR for anatomy education is its practicability. The ability to use personal devices for learning promotes self-paced and flexible learning, which is crucial in today's digital learning environment [38]. The incorporation of virtual and real environments offers intuitive interaction mechanisms and unparalleled realism, which is largely supported by advancements in smart devices [47,48]. These advancements allow students to retrieve anatomical information anytime and anywhere. AR technology's accessibility extends to various personal devices such as iPads, mobile phones, laptops, and smartwatches, enhancing its affordability and ease of use [49]. This accessibility and flexibility are particularly advantageous for understanding complex systems that involve multiple muscle groups, such as eye movement, which students can explore from a student's desk or home [38]. The ability to control the app and engage in self-practice promotes active learning and self-assessment [50]. As AR decreases cognitive load and enhances cognitive, meta-cognitive, and affective outcomes, it is gaining popularity in higher educational institutions [51]. Furthermore, university students frequently express confidence in embracing AR as an alternative to traditional learning methods, largely due to their high digital literacy and familiarity with tablets and smartphones [52]. Integrating AR into an anatomy curriculum may provide students with an engaging, flexible tool that enhances their overall learning experiences, preparing them for future professional challenges in the field of medicine.

Strengths and Limitations

This study has several strengths that contribute to the credibility and relevance of its findings. A major strength is the large sample size, which exceeded the minimum requirement and included students from multiple cohorts and diverse socioeconomic backgrounds. This allowed us to evaluate the perception of AR across a broad premedical student population. In addition, the questionnaire demonstrated excellent internal consistency and strong construct validity, supporting its suitability for capturing students' overall perception of AR integration.

Several limitations should also be acknowledged. First, the study focused exclusively on students' self-reported perceptions and did not assess objective learning outcomes, such as examination performance, knowledge retention, or skill acquisition. While a positive perception is an important indicator of educational acceptability, it does not necessarily equate to improved learning. Future studies should incorporate preintervention and postintervention assessments or comparative designs to better evaluate the educational effectiveness of AR.

Second, the cross-sectional design and single, brief exposure to the AR app limit conclusions about its long-term impact. Students' positive responses may have been influenced by the novelty of the technology, and it remains unclear whether similar levels of engagement and perceived benefits would be sustained with repeated use over time. Longitudinal studies examining continued AR use across multiple anatomy topics or semesters would provide valuable insights into its durability as a learning tool.

Third, this study was conducted at a single institution using a commercially available AR app focused on basic anatomy content. Therefore, the findings may not be fully generalizable to other institutions, curricular structures, or learner groups, particularly students at more advanced stages of medical training. Differences in instructional design, technological infrastructure, and faculty facilitation may influence how AR is experienced and perceived.

Fourth, although the survey instrument demonstrated strong internal consistency, other forms of validity, such as predictive or discriminant validity, were not examined. The exclusive use of quantitative survey data precluded deeper exploration of students' cognitive processes, challenges, or usability concerns during AR engagement. Incorporating qualitative methods, such as focus group discussions or interviews, could provide deeper insights into how AR supports or potentially hinders anatomy learning.

Finally, practical considerations related to implementation should be considered. Although AR apps are becoming increasingly accessible through personal devices, issues such as licensing costs, device compatibility, and classroom logistics may impact scalability and equitable access, particularly in larger cohorts or resource-limited settings.

Overall, this study provides early, learner-centered evidence that AR is positively received as a supplementary tool in premedical anatomy education. Future research incorporating objective learning outcomes, longer follow-up periods, and multi-institutional designs will be important to further define the educational value of AR and guide its sustainable integration into medical education curricula.

Conclusions

This study provided preliminary evidence that AR is consistently perceived as an interactive and engaging supplementary tool to traditional approaches in anatomy learning by premedical students. Students expressed high levels of receptivity to AR across various demographics, regardless of their age, previous educational background, or individual socioeconomic status. These findings support the incorporation of AR as a comprehensive tool in early anatomy learning. Importantly, this work provides preliminary evidence and serves as a foundation for larger-scale, multi-institutional studies. The app used in this study offers multifaceted benefits, including the enhancement of visualization and the enrichment of interest in anatomy learning. Future research involving multiple institutions, various learner cohorts, and incorporating objective assessments is warranted to validate and extend these findings and to further evaluate AR's pedagogical efficacy in anatomy education.

XSL•FO
RenderX

Acknowledgments

The authors express their sincerest appreciation to Adinegara Lutfi Abas, Prof Dr, Pro Vice Chancellor, Manipal University College Malaysia; Dr Chiong Kai Shing, Director of the Centre for Foundation and General Studies, Manipal University College Malaysia; all Centre for Foundation and General Studies faculty members; and all premedical students who participated in this research study.

Funding

No external financial support was received for the conduct of this study. However, institutional support from Manipal University College Malaysia was provided for the implementation of the instructional session and the article processing fee for publication.

Authors' Contributions

FI conceived the presented idea and designed the study. FI and N Hidayah Mohd Noor recruited the participants, conducted the intervention at Manipal University College Malaysia, Malacca, Malaysia, and analyzed the data. FI, N Hidayah Mohd Noor, and FAJ drafted the manuscript. FAJ, N Huda Mohd Nor, and SFMA reviewed the manuscript critically for intellectual content. All authors revised the manuscript and approved the final version to be published.

Conflicts of Interest

None declared.

Checklist 1
CHERRIES checklist.
[PDF File, 57 KB - mededu_v12i1e64402_app1.pdf]

References

1. Kurniawan MH, Witjaksono G. Human anatomy learning systems using augmented reality on mobile application. Procedia Comput Sci 2018;135:80-88. [doi: 10.1016/j.procs.2018.08.152]
2. Azer SA, Azer S. 3D anatomy models and impact on learning: a review of the quality of the literature. Health Prof Educ 2016 Dec;2(2):80-98. [doi: 10.1016/j.hpe.2016.05.002]
3. Trelease RB. From chalkboard, slides, and paper to e-learning: how computing technologies have transformed anatomical sciences education. Anat Sci Educ 2016 Nov;9(6):583-602. [doi: 10.1002/ase.1620] [Medline: 27163170]
4. Bhattarai L, Kc S, Shrivastava AK, Sah RP. Importance of cadaveric dissection in learning anatomy for medical students: a review. MedPhoenix J Natl Med Coll 2022;6(2):68-72. [doi: 10.3126/medphoenix.v6i2.38744]
5. Cercenelli L, De Stefano A, Billi AM, et al. AEducaAR, anatomical education in augmented reality: a pilot experience of an innovative educational tool combining AR technology and 3D printing. Int J Environ Res Public Health 2022 Jan 18;19(3):1024. [doi: 10.3390/ijerph19031024] [Medline: 35162049]
6. Syed Abd Halim SA, Yusoff MSB, Yaman MN, et al. Clinical students' reflections on the preclinical anatomy learning experience. J Taibah Univ Med Sci 2022 Dec;18(4):757-770. [doi: 10.1016/j.jtumed.2022.12.007] [Medline: 36852241]
7. Duarte ML, Santos LR, Guimarães Júnior JB, Peccin MS. Learning anatomy by virtual reality and augmented reality. A scope review. Morphologie 2020 Dec;104(347):254-266. [doi: 10.1016/j.morpho.2020.08.004] [Medline: 32972816]
8. Oliveira PS, Lucena JD, da Silva FRO, et al. The teaching of human anatomy in Brazil: the reality of northeast Brazil. Int J Anat Res 2017 Nov 1;5(4.2):4623-4629. [doi: 10.16965/ijar.2017.423]
9. Kumar R, Singh R. Model pedagogy of human anatomy in medical education. Surg Radiol Anat 2020 Mar;42(3):355-365. [doi: 10.1007/s00276-019-02331-7] [Medline: 31612273]
10. Rivera García GE, Cervantes López MJ, Ramírez Vázquez JC, Llanes Castillo A, Cruz Casados J. Reviewing mobile apps for teaching human anatomy: search and quality evaluation study. JMIR Med Educ 2025 Feb 14;11:e64550. [doi: 10.2196/64550] [Medline: 39951706]
11. Naidoo N, Akhras A, Banerjee Y. Confronting the challenges of anatomy education in a competency-based medical curriculum during normal and unprecedented times (COVID-19 pandemic): pedagogical framework development and implementation. JMIR Med Educ 2020 Oct 7;6(2):e21701. [doi: 10.2196/21701] [Medline: 32873536]
12. Wickramasinghe N, Thompson BR, Xiao J. The opportunities and challenges of digital anatomy for medical sciences: narrative review. JMIR Med Educ 2022 May 20;8(2):e34687. [doi: 10.2196/34687] [Medline: 35594064]
13. Jiang H, Vimalesvaran S, Wang JK, Lim KB, Mogali SR, Car LT. Virtual reality in medical students' education: scoping review. JMIR Med Educ 2022 Feb 2;8(1):e34860. [doi: 10.2196/34860] [Medline: 35107421]
14. Sinou N, Sinou N, Filippou D. Virtual reality and augmented reality in anatomy education during COVID-19 pandemic. Cureus 2023 Feb;15(2):e35170. [doi: 10.7759/cureus.35170] [Medline: 36949987]

XSL•FO
RenderX

15. Kang B. How the COVID-19 pandemic is reshaping the education service. In: Lee J, Han SH, editors. The Future of Service Post-COVID-19 Pandemic, Volume 1 Rapid Adoption of Digital Service Technology: Springer:15-36. [doi: 10.1007/978-981-33-4126-5_2]
16. Rashid AA, Rashid MRA, Yaman MN, Mohamad I. Teaching medicine online during the COVID-19 pandemic: a Malaysian perspective. Bangladesh J Med Sci 2020 Jul 20;19(Special issue):S77-S81. [doi: 10.3329/bjms.v19i0.48170]
17. Pallavicini F, Anesa P. A narrative review on augmented reality in education. Educ Sci 2026;16(2):261. [doi: 10.3390/educsci16020261]
18. Abhyankar A, Thakur M, Vaity A, et al. AR based human internal organs for education. Int J Sci Res Sci Technol 2025;12(2):938-942. [doi: 10.32628/IJSRST251222647]
19. Jantanukul W. Immersive reality in education: transforming teaching and learning through AR, VR, and mixed reality technologies. J Educ Learn Rev 2024;1(2):51-62. [doi: 10.60027/jelr.2024.750]
20. Pani D, Maharana N, Pradhan SK, Das S. Integrating effect of AR and VR in sustained learning outcomes in self-directed learning environments. In: Impacts of AI on Students and Teachers in Education 5.0: IGI Global; 2024:317-348. [doi: 10.4018/979-8-3693-8191-5.ch013]
21. Pandya K. Augmented reality for education: enhancing STEM learning through interactive visualization. Int J Multidiscip Res 2025 Oct 24;7(5). [doi: 10.36948/ijfmr.2025.v07i05.58352]
22. Simok AA, Mohd Ismail ZI, Mat Zin NA, et al. Harnessing 3D technology to transform anatomy education. Educ Med J 2024;16(4):171-179. [doi: 10.21315/eimj2024.16.4.13]
23. Kirkwood A, Price L. Technology-enhanced learning and teaching in higher education: what is 'enhanced' and how do we know? A critical literature review. Learn Media Technol 2014 Jan 2;39(1):6-36. [doi: 10.1080/17439884.2013.770404]
24. Eysenbach G. Improving the quality of Web surveys: the Checklist for Reporting Results of Internet E-Surveys (CHERRIES). J Med Internet Res 2004 Sep 29;6(3):e34. [doi: 10.2196/jmir.6.3.e34] [Medline: 15471760]
25. García-Robles P, Cortés-Pérez I, Nieto-Escámez FA, García-López H, Obrero-Gaitán E, Osuna-Pérez MC. Immersive virtual reality and augmented reality in anatomy education: a systematic review and meta-analysis. Anat Sci Educ 2024;17(3):514-528. [doi: 10.1002/ase.2397] [Medline: 38344900]
26. Chytas D, Johnson EO, Piagkou M, et al. The role of augmented reality in anatomical education: an overview. Ann Anat 2020 May;229:151463. [doi: 10.1016/j.aanat.2020.151463] [Medline: 31978568]
27. Iparraguirre-Villanueva O, Andia-Alcarraz J, Saba-Estela F, Epifanía-Huerta A. Mobile application with augmented reality as a support tool for learning human anatomy. Int J Eng Ped 2024 Jan 30;14(1):82-95. [doi: 10.3991/ijep.v14i1.46845]
28. Ateş H, Polat M. Leveraging augmented reality and gamification for enhanced self-regulation in science education. Educ Inf Technol 2025 Aug;30(12):17079-17110. [doi: 10.1007/s10639-025-13481-0]
29. Al Hamidy Hazidar RS. Visualization cardiac human anatomy using augmented reality mobile application. Int J Electron Commun Comput Eng 2014 May;5(3):2278-4209 [FREE Full text]
30. Allen LK, Eagleson R, de Ribaupierre S. Evaluation of an online three-dimensional interactive resource for undergraduate neuroanatomy education. Anat Sci Educ 2016 Oct;9(5):431-439. [doi: 10.1002/ase.1604] [Medline: 26990135]
31. Garg V, Soni L. Augmented reality enhanced clinical training: implications for skill acquisition and retention in medical education. Presented at: 2024 2nd DMIHER International Conference on Artificial Intelligence in Healthcare, Education and Industry (IDICAIEI); Nov 29-30, 2024. [doi: 10.1109/IDICAIEI61867.2024.10842884]
32. Gutiérrez-Jara JK, Malpica-Rodríguez ME, Pérez-Aguilar DA, Pérez-Aguilar JM. Mobile application with augmented reality and its impact on anatomy learning in human medicine students. Int J Interact Mob Technol 2025 Jan 27;19(2):38-52. [doi: 10.3991/ijim.v19i02.52841]
33. Niu S, Zhang J, Lin J, Wang B, Yan J. Enhancing anatomy education with virtual reality: integrating three-dimensional models for improved learning efficiency and student satisfaction. Front Med (Lausanne) 2025;12:1555053. [doi: 10.3389/fmed.2025.1555053] [Medline: 40534689]
34. Yammine K, Violato C. The effectiveness of physical models in teaching anatomy: a meta-analysis of comparative studies. Adv Health Sci Educ Theory Pract 2016 Oct;21(4):883-895. [doi: 10.1007/s10459-015-9644-7] [Medline: 26459329]
35. Wainman B, Wolak L, Pukas G, Zheng E, Norman GR. The superiority of three-dimensional physical models to two-dimensional computer presentations in anatomy learning. Med Educ 2018 Nov;52(11):1138-1146. [doi: 10.1111/medu.13683] [Medline: 30345680]
36. Arici F, Yildirim P, Caliklar Ş, Yilmaz RM. Research trends in the use of augmented reality in science education: content and bibliometric mapping analysis. Comput Educ 2019 Dec;142:103647. [doi: 10.1016/j.compedu.2019.103647]
37. Mokmin NAM, Hanjun S, Jing C, Qi S. Impact of an AR-based learning approach on the learning achievement, motivation, and cognitive load of students on a design course. J Comput Educ 2024 Jun;11(2):557-574. [doi: 10.1007/s40692-023-00270-2]
38. Dhar P, Rocks T, Samarasinghe RM, Stephenson G, Smith C. Augmented reality in medical education: students' experiences and learning outcomes. Med Educ Online 2021 Dec;26(1):1953953. [doi: 10.1080/10872981.2021.1953953] [Medline: 34259122]

XSL•FO
RenderX

39. Chien CH, Chen CH, Jeng TS. An interactive augmented reality system for learning anatomy structure. Presented at: The International MultiConference of Engineers and Computer Scientists 2010, IMECS 2010; Mar 17-19, 2010 URL: https://www.iaeng.org/publication/IMECS2010/IMECS2010_pp370-375.pdf [accessed 2026-05-19]
40. Ateş H, Garzón J. An integrated model for examining teachers' intentions to use augmented reality in science courses. Educ Inf Technol 2023 Feb;28(2):1299-1321. [doi: 10.1007/s10639-022-11239-6]
41. Ateş H. Integrating augmented reality into intelligent tutoring systems to enhance science education outcomes. Educ Inf Technol 2025 Mar;30(4):4435-4470. [doi: 10.1007/s10639-024-12970-y]
42. Koumpouros Y. Revealing the true potential and prospects of augmented reality in education. Smart Learn Environ 2024;11(1):2. [doi: 10.1186/s40561-023-00288-0]
43. Martín-Valero R, Vega-Morales A Sr, Martín-Vega FJ, Rodriguez-Huguet M, Rodríguez-Martínez MC, Vinolo-Gil MJ. Effectiveness of augmented reality in the teaching of health university students: quasi-experimental study. JMIR Serious Games 2025 Mar 27;13:e54312. [doi: 10.2196/54312] [Medline: 40146914]
44. Huang HM, Rauch U, Liaw SS. Investigating learners' attitudes toward virtual reality learning environments: based on a constructivist approach. Comput Educ 2010 Nov;55(3):1171-1182. [doi: 10.1016/j.compedu.2010.05.014]
45. Kashinath KA, Ambadas KB, Rajendra DR, Sanjay AS, Shimant SV. Augmented reality in anatomy. Int J Eng Res Technol 2022 May 13;11(5):78-82. [doi: 10.5281/zenodo.18446997]
46. Boomgaard A, Fritz KA, Isafiade OE, et al. A novel immersive anatomy education system (Anat_Hub): redefining blended learning for the musculoskeletal system. Appl Sci (Basel) 2022 Jun 3;12(11):5694. [doi: 10.3390/app12115694]
47. Omar K, Fakhouri H, Zraqou J, Marx Gómez J. Usability heuristics for metaverse. Computers 2024 Sep 6;13(9):222. [doi: 10.3390/computers13090222]
48. Wong JY, Azam AB, Cao Q, et al. Evaluations of virtual and augmented reality technology–enhanced learning for higher education. Electronics (Basel) 2024;13(8):1549. [doi: 10.3390/electronics13081549]
49. Albrecht UV, Folta-Schoofs K, Behrends M, von Jan U. Effects of mobile augmented reality learning compared to textbook learning on medical students: randomized controlled pilot study. J Med Internet Res 2013 Aug 20;15(8):e182. [doi: 10.2196/jmir.2497] [Medline: 23963306]
50. Nagayo Y, Saito T, Oyama H. A novel suture training system for open surgery replicating procedures performed by experts using augmented reality. J Med Syst 2021 Apr 7;45(5):60. [doi: 10.1007/s10916-021-01735-6] [Medline: 33829327]
51. Grübel J, Chatain J, Schmid C, et al. Applying augmented reality to convey medical knowledge on osteoclasts to users of a serious game: vignette experiment. JMIR Serious Games 2025 Jun 16;13:e64751. [doi: 10.2196/64751] [Medline: 40523213]
52. Sharmin N, Chow AK. Augmented reality application to develop a learning tool for students: transforming cellphones into flashcards. Healthc Inform Res 2020 Jul;26(3):238-242. [doi: 10.4258/hir.2020.26.3.238] [Medline: 32819042]

Abbreviations

AR: augmented reality
CHERRIES: Checklist for Reporting Results of Internet E-Surveys
FiS: Foundation in Science
MUCM: Manipal University College Malaysia

Edited by B Lesselroth; submitted 16.Jul.2024; peer-reviewed by NHM Zainal, S Alkahtany; revised version received 08.Apr.2026; accepted 21.Apr.2026; published 08.Jun.2026.

Please cite as:
Idris F, Mohd Noor NH, Jam FA, Mohd Nor NH, Mohamad Asri SF
Perspectives of Premedical Students at a Medical University College in Malaysia Regarding Augmented Reality Integration in Anatomy Education: Cross-Sectional Study
JMIR Med Educ 2026;12:e64402
URL: https://mededu.jmir.org/2026/1/e64402
doi:10.2196/64402

© Fadhilah Idris, Nur Hidayah Mohd Noor, Faidruz Azura Jam, Nurul Huda Mohd Nor, Siti Fadziyah Mohamad Asri. Originally published in JMIR Medical Education (https://mededu.jmir.org), 8.Jun.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Education, is properly cited. The complete bibliographic information, a link to the original publication on https://mededu.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Virtual Reality Versus In-Person Simulation of Sepsis for Medical Students: Randomized Comparative Pilot Study

Lauren Medwell[1*], MBBS, MRes; Tim Old[1*], MB BChir, MSc; Awais Ahmed[1], MBBS; Victoria Holloway[1], MB ChB; Lauren McTaggart[1], MB ChB; Dafydd Morgan[1], MB ChB, MSc; Callum O'Keeffe[1], MBBS; Jemaima Olori[1], MB ChB; Ellie Payne[1], MB ChB; Christopher Rainforth[1], MB ChB; Namritha Ramanujam[1], MB ChB; Kriti Vaidya[1], MBBS; Chris Jacobs[1,2], MB BChir, MRes, MD(Res)

[1]Undergraduate Medical Education Team, Great Western Hospitals NHS Foundation Trust, The Academy, Marlborough Road, Swindon, United Kingdom

[2]Department of Psychology, University of Bath, Bath, United Kingdom

[*]these authors contributed equally

Corresponding Author:
Chris Jacobs, MB BChir, MRes, MD(Res)
Undergraduate Medical Education Team, Great Western Hospitals NHS Foundation Trust, The Academy, Marlborough Road, Swindon, United Kingdom

Abstract

Background: Virtual reality (VR) simulation—using head-mounted displays to present a computer-generated, 3D, interactive environment—may be a cost-effective alternative to in-person (IP) medical simulation training. However, studies directly comparing learning outcomes have demonstrated mixed results and mainly focused on knowledge or skill acquisition rather than integrated practice.

Objective: This randomized comparative pilot study aimed to evaluate the effectiveness of VR versus IP simulation in developing sepsis management skills among final-year medical students, addressing a gap in medical education evidence.

Methods: Final-year medical students at Great Western Hospital, United Kingdom, participated in both IP and VR simulation sessions featuring sepsis scenarios. Session order was randomized, determining study group assignment. Participants underwent an additional video-recorded "assessment" IP simulation of septic shock management either between or after both scheduled sessions. Questionnaires were completed between scenario completion and debriefing across all sessions. Performance was evaluated using a modified Queen's Simulation Assessment Tool (mQSAT) by facilitating study authors across all sessions, with the assessment simulation additionally evaluated by blinded assessors. The primary outcomes included mQSAT scores, recognition of septic shock, and identification of critical care needs. Analysis of covariance was conducted to detect differences in mQSAT scores between the groups, with simulation modality as the independent variable and the number of simulations or debriefs prior to assessment as the covariate. Binary outcomes between the groups were analyzed using binomial tests.

Results: A total of 32 participants were recruited and allocated to 1 of 4 groups based on completed simulation sessions prior to assessment: IP only (IP-Assess, n=10), VR only (VR-Assess, n=6), IP then VR (IP-VR-Assess, n=11), and VR then IP (VR-IP-Assess, n=5). No statistically significant differences in mQSAT scores were detected between any groups for any domain nor for the recognition of septic shock. For the recognition of need for critical care, participants who completed VR simulation only prior to assessment were more likely to recognize need for critical care than those who completed IP simulation only (3/5 vs 1/7; *P*=.01).

Conclusions: This study demonstrates the feasibility of the proposed trial method and provides insight into likely effect sizes for the design of further studies. The measured learning outcomes were similar across the groups, regardless of which simulation modalities were used prior to assessment. Our study found no statistically significant differences for VR simulation versus IP simulation for the measured educational outcomes, which is reassuring for the ethical conduct of further studies comparing VR and IP simulation.

(JMIR Med Educ 2026;12:e80316) doi:10.2196/80316

KEYWORDS

virtual reality; simulation; undergraduate; technology-enhanced learning; medical student education; high-fidelity simulation; extended reality

XSL•FO
RenderX

Introduction

Virtual reality (VR) is a method of representing an immersive visual and sound sensory experience, which has the potential to mirror complex human interactions. This, coupled with its scope of user interactivity, has given researchers and educationalists significant interest in VR health care training [1,2]. VR can recreate realistic clinical scenarios [3] that promote learning in Bloom's taxonomy of clinical knowledge [4,5], psychomotor skills [6,7], and affective attitudes [8,9]. These overlapping domains form crucial areas of learning outcomes for simulation-based education. Simulation-based education is embedded in both undergraduate and graduate curriculums across a wide range of health care professions, and the integration of technological advances has included higher fidelity manikins and immersive digital technologies [10]. Simulation combines educational theory and clinical competency within traditional and emergent technological–led practices, ultimately becoming a multifaceted pedagogy that complements teaching strategies to reduce harm to patients [11].

Despite the promise of VR training for health care professions [12], there remains a lack of robust evidence directly comparing its effectiveness to traditional simulation methods [13]. While VR offers immersive, scalable, and cost-effective training opportunities, the number of experimental studies assessing its impact on clinical skills, decision-making, and patient outcomes is limited. Many existing studies focus on user experience and engagement rather than objective performance measures, highlighting the need for more rigorous trials to establish VR's true educational value in health care training [2]. One area of focus has been cardiopulmonary respiratory training, where VR training has been reported to be noninferior to traditional training methods of in-person (IP) simulation [14]. However, there remains heterogeneity in studies that compare VR to a control condition, whereby the content of the control differs from the intervention. Hence, this directly impacts the outcomes and interpretations.

The main objective of this study was to pilot a comparative study to evaluate the feasibility of conducting a randomized control trial comparing VR simulation of a medical emergency with traditional IP simulation methods, ensuring the alignment of educational content and incorporating an objective assessment score as the outcome measure. Second, we aimed to develop a method that would allow the comparison of these 2 differing simulation modalities, accounting for the acquisition of both knowledge and skills. Finally, we sought to quantify the effects of both IP and VR simulation on learning outcomes, inform the design of future studies, and ensure VR is not obviously inferior to IP simulation training.

Methods

Trial Design

As a pilot study, the full protocol was not entirely prespecified in advance. The outline methodology received local ethical approval prior to commencement of other study activities; external prospective trial registration was not pursued as this pilot study involved an intervention for health care professionals with solely education outcomes, which does not meet the International Committee of Medical Journal Editors definition of a clinical trial requiring preregistration [15]. The study was originally designed as a 1:1 randomized parallel groups trial, with an assessment simulation testing performance pre- and post-testing either IP or VR simulation teaching. A pragmatic adaptation was subsequently undertaken as follows:

We adopted a 4-parallel-arm design, with students cluster randomized to undergo an assessed simulation scenario at 1 of 4 time points relative to taught simulation sessions:

1. Assessment after IP simulation teaching only (IP-Assess group)
2. Assessment after VR simulation teaching only (VR-Assess group)
3. IP simulation teaching, then VR simulation teaching, then assessment (IP-VR-Assess group)
4. VR simulation teaching, then IP simulation teaching, then assessment (VR-IP-Assess group)

We intended to capture the added benefit of an additional session of each modality by comparing performance across arms. For example, the additional learning effect of an IP simulation could be quantified by comparing assessment simulation scores between the VR-Assess group and scores from the VR-IP-Assess group. Similarly, the additional learning effect of a VR simulation could be quantified by comparing assessment scores between the IP-Assess and IP-VR-Assess groups. Given the efficacy of VR simulation teaching is less well established than IP simulation, we targeted a ratio of 2:1 participants favoring arms informing the estimation of VR simulation efficacy (ie, the IP-Assess and IP-VR-Assess groups).

This cluster randomization process introduced incongruence between units of randomization (simulation groups) and analysis (individual participants), thus complicating the interpretation of subsequent results. However, cluster randomization is necessary when the intervention is delivered at the level of the cluster, and such incongruence is considered acceptably pragmatic research practice, particularly to facilitate research embedded within usual practice as we have done [16]. It was deemed both logistically infeasible and poorly reflective of real-world practice to deliver simulation teaching on an individual level; the statistical handling of this compromise is described in detail in the *Analytical Methods* section. This study was reported according to the Consolidated Standards of Reporting Trials (CONSORT) 2025 guidelines (Checklist 1).

Participants

The inclusion criteria were final-year medical students on placement at Great Western Hospital (GWH). These students were time-tabled both 1 VR and 1 IP simulation session in their placement, regardless of study enrollment. Three separate cohorts, each of up to 16 students, attended during the study period for 4 weeks at a time between September and December 2023.

All students were invited to participate in this study at their placement induction and again at the beginning of their first simulation session. Students were assured that their study participation status would not affect their course progress or

XSL•FO
RenderX

academic record. Participant information sheets were provided, any questions were answered, and formal written consent was taken; information sheets and consent forms are presented in our OSF repository (see [17]).

Interventions

VR simulation sessions followed a format piloted and refined on 2 previous cohorts of final-year medical students. Three VR simulation scenarios developed by Goggleminds were available (Sepsis, v.SE1.100009; Anaphylaxis, v.AN1.100006; and Asthma, v.AS1.100006), accessed using Meta Quest 2 head-mounted displays. The sepsis simulation used was codeveloped between Goggleminds and several National Health Service training providers and has been endorsed by the Sepsis Trust [18]; it has been studied previously to review instructional design and investigate user experience, where it was demonstrated to elicit high intrinsic motivation, perceived learning, and immersion [19]. Sessions lasted 2 hours, including session prebrief and orientation to the virtual environment, 2 to 3 scenarios completed by individual students alone with other students watching via screencast to a video display, and whole group debrief following each scenario. Where 4 students attended, the final scenario was completed by 2 students concurrently, though in separate virtual environments, and with only 1 screencast. No significant software updates occurred during the study period.

IP simulation sessions were administered as per established local practice. Four to six students underwent high-fidelity simulation training in a dedicated suite, using a mix of simulated patients (played by clinical teaching fellows) and high-fidelity manikins (Ares Emergency Care Manikin, CAE). Scenarios were taken from a preapproved bank developed in our institute, all written with oversight from consultant-grade doctors from a mix of acute specialties. Sessions lasted 3 to 4 hours, including session prebrief and orientation to the simulation suite, 2 to 3 scenarios completed by students in pairs with other students watching via video link, and whole group debrief following each scenario.

Both VR and IP simulation sessions included sepsis scenarios, with adjustments made to align both as far as possible in terms of patient factors, room setup, and equipment available. Both sessions were debriefed by the same 2 faculty (LM and TO) for all participants, and debriefs included discussion of septic shock recognition and management.

Outcomes

Learning effects of each modality were quantified using an assessment simulation, which was performed in addition to students' usual time-tabled simulation sessions. The assessment simulation presented a simulated patient in septic shock secondary to cellulitis; full scenario details can be found in Multimedia Appendix 1.

To the authors' knowledge, there exists no simulation performance assessment tool validated across both IP and VR modalities. To quantify differences in simulation performance across modalities, we identified the modified Queen's Simulation Assessment Tool (mQSAT) [20]. The mQSAT uses 5 Likert-style scales ranging from inferior (1) to superior (5) performance, presented together with guide statements; 4 scales rate specific domains of primary assessment, diagnostic actions, therapeutic actions, and communication, with the final scale being a global rating. The mQSAT is validated for use in the assessment of IP simulation performance of medical students with good interrater reliability [21] and was thought prima facie to be applicable to the assessment of VR simulation.

To quantify and help account for baseline differences in participant ability, mQSAT assessments of performance in time-tabled VR and IP simulation teaching were also recorded. Two authors (LM and TO) marked all simulations, teaching and assessment, in an unblinded fashion. The assessment simulations were additionally video-recorded and double-marked by 2 blinded assessors, randomly allocated from the other study authors.

To quantify specific knowledge–based learning around the recognition and management of septic shock separate to assessment of skills or simulation performance, a questionnaire was presented to participants after the assessment simulation. The questionnaire comprised a free-text question concerning overall diagnosis; a checklist of standard actions during an A-E assessment, where participants were asked to select the actions most relevant to the scenario; and a multiple-choice question regarding the level of care required (eg, ward based vs intensive care) with justification. The full postscenario questionnaire can be found in Multimedia Appendix 1. To familiarize participants with the format prior to the assessment, the questionnaire was also provided to each participant after each scenario in time-tabled IP and VR assessment simulations, regardless of whether they had been performing the scenario themselves or watching one of their peers. The questionnaire was agreed upon between 2 of the authors (LM and TO) and designed to be usable across all scenarios across both modalities without undue prompting, yet to elicit the explicit recognition of septic shock and need for critical care input in the assessment scenario where this was achieved.

The primary outcomes were differences in the following between intervention groups:

1. mQSAT scores
2. Recognition of septic shock
3. Recognition of need for critical care

Secondary outcomes were differences in the following between intervention groups:

1. Completion of the Sepsis 6

Qualitative outcome data, including measures of participant experience and questionnaire free-text responses, were also collected and will be analyzed and reported separately.

Sample Size

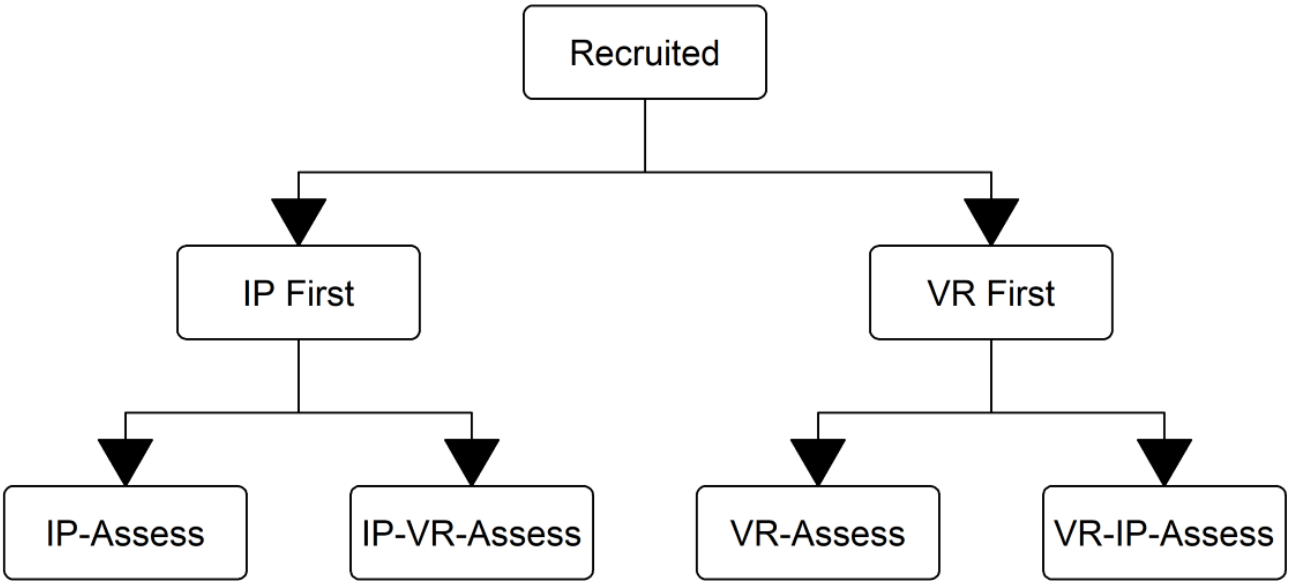
This was a pilot trial, designed in part to generate plausible effect sizes for any differences in the efficacy of IP versus VR simulation teaching; it was therefore not possible to calculate a formal sample size a priori. The decision to stop recruiting was pragmatic due to the end of the academic year.

XSL•FO
RenderX

Randomization

The eligible student cohort was chosen, as the existing timetable template allowed both IP and VR simulation sessions for all students in a given placement. Students attending each placement block were given a letter from A to P by their university. Students A to F were allocated to simulation group 1, students G to K to simulation group 2, and students L to P to simulation group 3. The choice of group size was linked to the capacity of the simulation suite and is reflective of typical class size for simulation teaching at the study site. The subsequent cluster randomization process and relevant R code is available in our OFS repository [17]. Briefly, room availability created natural variation between whether IP or VR simulation was time-tabled first for a given simulation group. A random number generator was used to determine the order in which available simulation slots were allocated to simulation groups, thereby randomizing simulation groups to start with either IP or VR simulation. Extra assessment simulation slots were booked either between or after both time-tabled IP and VR simulation sessions (Figure 1). All allocations were completed prior to the arrival of students at GWH, with separate authors performing allocation of random numbers (TO) and students (LM) to each simulation group.

Figure 1. Group allocation flow diagram. IP: in-person; VR: virtual reality.



Analytical Methods

Analyses were finalized after the pragmatic adaptation to a 4-arm design but were fully prespecified prior to participant recruitment or data collection. Differences between mQSAT scores in the assessment simulation were statistically assessed by one-way analysis of covariance (ANCOVA), with assessment mQSAT score as the dependent variable, simulation modality being assessed as the independent variable, and number of simulations or debriefs prior to assessment as the covariate. Differences in recognition of septic shock and need for critical care were statistically assessed using binomial tests. Completion of the Sepsis 6 was compared using the Fisher exact test. Statistical testing used a significance threshold of α=.05; no formal correction for multiple comparisons was made. Given the nature of this study as a small-scale pilot, complete case analysis without imputation was undertaken for all analyses performed.

Pre-specified sensitivity analyses were performed to ensure study findings were robust to potential violations of assumptions. To ensure no baseline differences in students' simulation performance, ANCOVA was repeated with baseline simulation mQSAT score (either for VR or IP training simulation session) as the covariate. Reanalysis using baseline scores as the covariate was also intended to account for issues arising from clustering effects. Versus individual-level randomization, cluster randomization inflates the variance of effect estimates in proportion to the intracluster correlation coefficient (representing how strongly individuals within clusters are related to each other), thereby decreasing statistical power [22]. Unadjusted use of methods assuming individual randomization, including ANCOVA, therefore risks an inflated type 1 error rate. ANCOVA analysis of cluster means using baseline measurements has been shown to produce similar treatment effect estimates versus the gold-standard method of mixed multilevel regression, though it is conceptually and computationally simpler [23]; given available resources, this approach was adopted as a pragmatic method of accounting for clustering effects. Finally, as all mQSAT scores for nonassessment simulations were undertaken by nonblinded study authors (LM and TO), to assess bias in scoring from nonblinded versus blinded mQSAT scoring, interrater agreement was quantified by the Krippendorff α.

Ethical Considerations

The study protocol was granted ethical approval by the Swindon Academy Medical Education Research Committee (reference: LMTO0823). An outline methodology was drafted and submitted to peer review as part of obtaining local ethical approval; relevant documentation is available via our OSF repository [17]. All students were time-tabled for both VR and IP simulations regardless of study participation. Students were assured that their study participation status would not affect

XSL•FO
RenderX

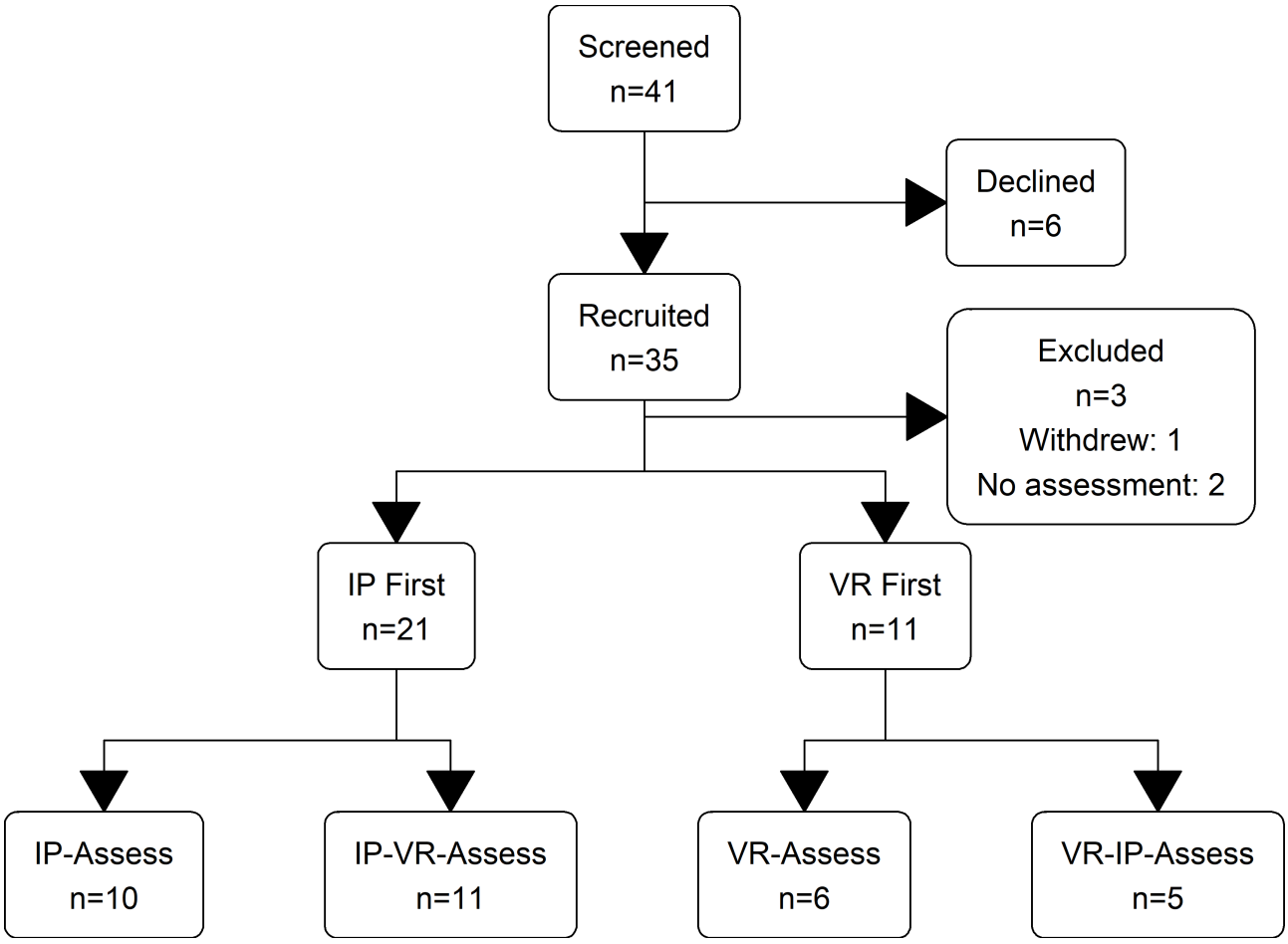
their course progress or academic record and that they could withdraw at any time; participation was entirely voluntary, with no financial or other compensation offered. Participant information sheets were provided, any questions were answered, and formal written consent was taken; information sheets and consent forms are presented in our OSF repository [17]. Potential risks and mitigations were considered and communicated to participants within written information and verbal prebriefs prior to simulations. Video recordings were performed using Microsoft Teams linked to a secure Trust account of one author (TO) and were only available to the research team otherwise through sharing secure links to individual videos on a named-person basis to relevant study staff as required. All digital participant data including video recordings were held on secure Trust computer systems, with physical forms stored in a secure office; all data were held for the duration of the study and destroyed on study completion.

Results

Participant Recruitment and Flow

A total of 41 students were approached, and 35 participants were recruited between September 25 and December 22, 2023. One participant withdrew, and 2 participants did not attend the assessment simulation due to illness. There were therefore 32 participants in the final analysis for each of the primary and secondary outcomes; participant flow is summarized in Figure 2. All 32 participants otherwise completed all simulation sessions as scheduled. Recruitment was stopped due to the end of eligible final-year student placements at GWH for the academic year.

Figure 2. CONSORT (Consolidated Standards of Reporting Trials) flow diagram. IP: in-person; VR: virtual reality.



Baseline Data

Baseline characteristics are summarized in Table 1, and the baseline experience of simulation training and VR use is summarized in Multimedia Appendix 2 in Tables S1-S3. Baseline characteristics were broadly comparable across the groups with respect to age, sex, proportion of graduate-entry students, previous simulation experience, and previous VR experience.

XSL•FO
RenderX

Table . Baseline characteristics—age and proportion of graduate-entry students.

Study group	Participants, n	Age (y), mean (SD)	Graduate-entry students, n (%)	Female, n (%)
IP-VR-Assess[a,b]	11	25.6 (2.7)	2 (18)	5 (46)
VR-IP-Assess	5	24 (1.4)	1 (20)	1 (20)
IP-Assess	10	23.4 (0.5)	0 (0)	5 (50)
VR-Assess	6	25 (2.8)	2 (33)	4 (67)

[a]IP: in-person.

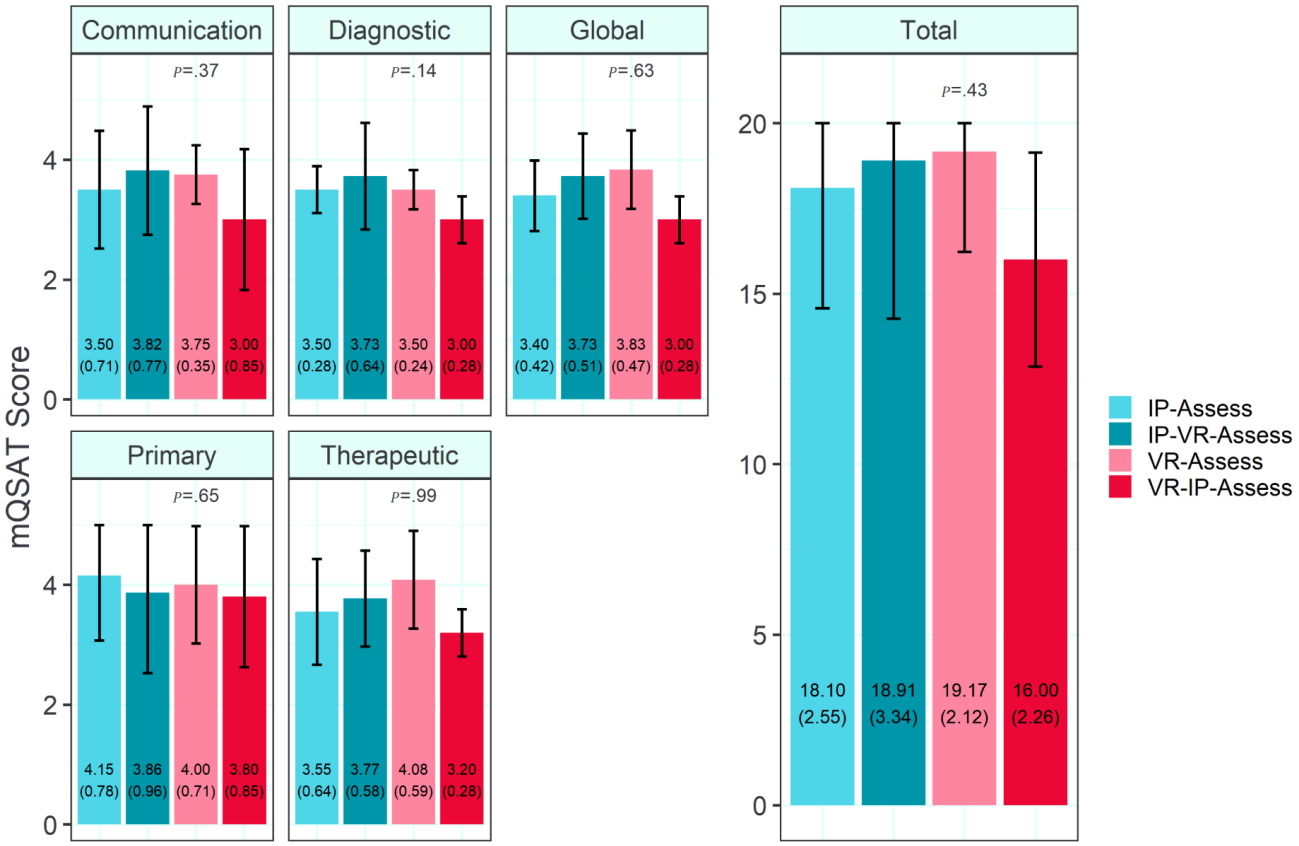
[b]VR: virtual reality.

Outcomes and Estimation

Primary Analyses

For mQSAT scores (n=32), no statistically significant differences were observed between any of the study groups for any domain by ANCOVA analysis. The results are summarized in Figure 3.

Figure 3. Mean(SD) Modified Queen's Simulation Assessment Tool (mQSAT) scores by study group. Error bars are 95% confidence interval. IP: in-person; VR: virtual reality.



For the recognition of septic shock (n=32), no statistically significant differences were observed between groups by binomial test. Results are summarized in Table 2.

XSL•FO
RenderX

Table . Recognition of septic shock and need for critical care referral by study group.

Study group	Participants, n	Prior debriefs, n	Septic shock recognized[b], n	Septic shock *P* value[a]	Critical care recognized[b], n	Critical care *P* value[a]
IP-Assess[c]	10	1	7	.23	1	.01
VR-Assess[d]	6	1	5	—[e]	3	—
IP-VR-Assess	11	2	7	.70	2	>.99
VR-IP-Assess	5	2	3	—	0	—

[a]*P* value by binomial test.

[b]Agreed by both blinded reviewers.

[c]IP: in-person.

[d]VR: virtual reality.

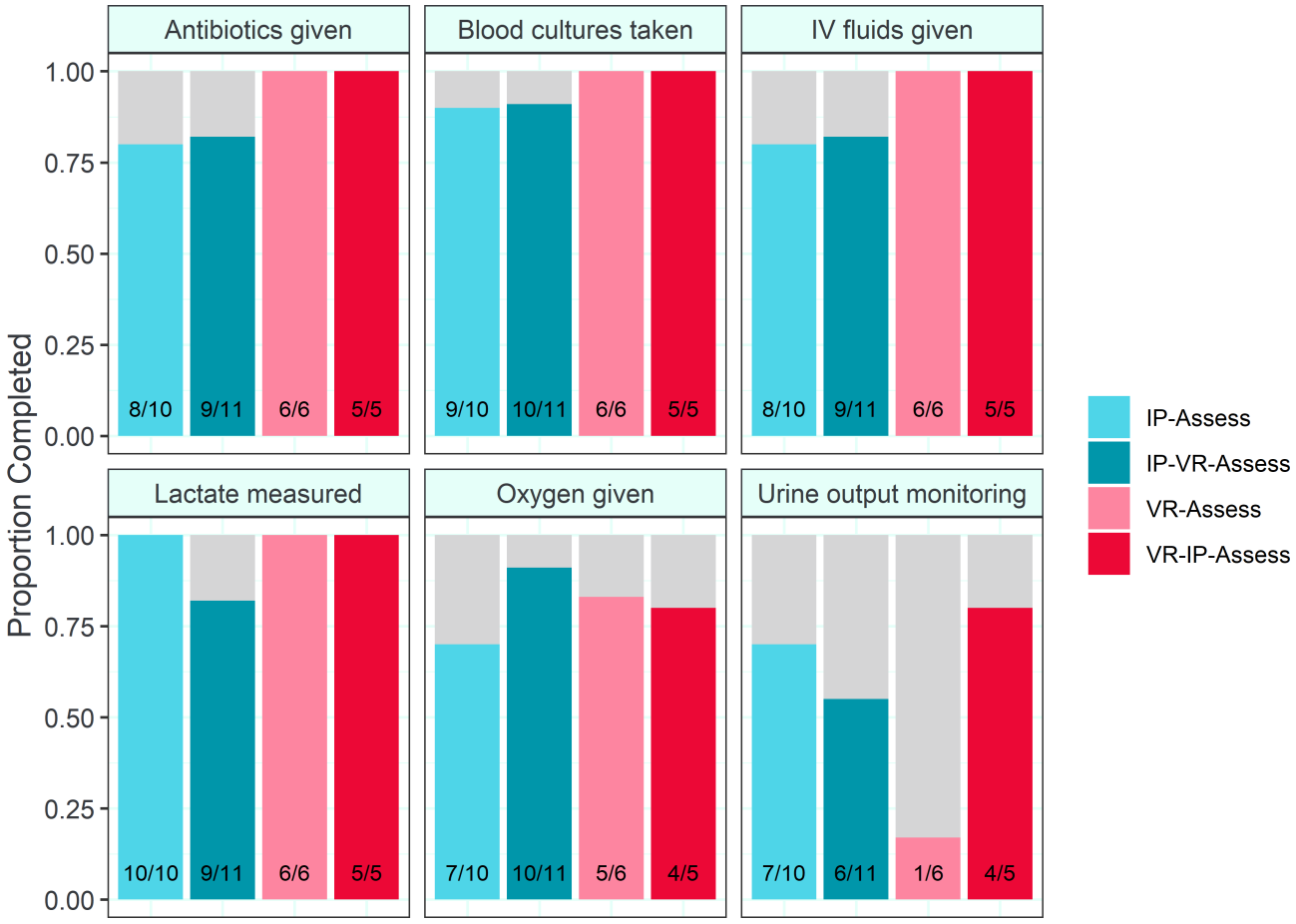
[e]Not applicable.

For recognition of need for critical care (n=32), a statistically significant difference was noted between the groups in participants who had received only 1 debrief (ie, 1 simulation session) prior to assessment. Participants who completed only 1 VR simulation and debrief prior to assessment were more likely to recognize need for critical care than those who completed only IP simulation and debrief (3/5 vs 1/7; *P*=.01). The results are summarized in Table 2.

Secondary Analyses

For the completion of the Sepsis 6 (n=32), no differences in performance were observed between any groups for any component by the Fisher exact test. The results are summarized in Figure 4.

Figure 4. Proportion of Sepsis 6 components completed by study group. IP: in-person; VR: virtual reality.



Ancillary Analyses

The results of prespecified sensitivity analyses are reported in full in Multimedia Appendix 3.

Using first mQSAT scores as the covariate in ANCOVA found no significant differences between the study groups (n=32; Table S1 in Multimedia Appendix 3).

XSL•FO
RenderX

Using either unblinded reviewers' assessment mQSAT scores or all 4 reviewers' assessment mQSAT scores did not change conclusions, finding no significant differences between the study groups (n=32; Tables S2 and S3 in Multimedia Appendix 3).

Using a more lenient criterion of any of 4 reviewers marking the recognition of septic shock or need for critical care did not change conclusions, finding no significant differences between the study groups (n=32; Table S4 in Multimedia Appendix 3).

The Krippendorff α indicated slight-to-moderate agreement between reviewers (n=4; range −0.155 to 0.491; Table S5 i-iv in Multimedia Appendix 3).

Harms

No unintended harms, data breaches, or other unintended effects were noted during the study. Qualitative feedback was collected from participants, which will be fully reported separately.

Some noted motion or cybersickness, which is a known occurrence in VR settings. This was considered in our risk assessment and communicated during participant prebriefings. Two participants reported motion sickness during sessions, leading to a brief pause and the option to stop completely. With small adjustments, such as completing the simulation in a seated rather than ambulatory position, participants were able to continue and complete the full simulation.

Discussion

Principal Findings

This pilot study found no statistically significant differences in overall simulation performance between medical students who received VR simulation training versus those who received traditional IP simulation training for sepsis management. The objective assessment mQSAT scores across all domains (primary assessment, diagnostic actions, therapeutic actions, communication, and global rating) showed comparable performance regardless of the simulation modality used prior to assessment. The only statistically significant finding was that students who completed a single VR simulation session prior to assessment were more likely to recognize the need for critical care compared to those who completed a single IP simulation session (3/5 vs 1/7; *P*=.01). These findings showed no clear differences between VR simulation and traditional IP simulation for teaching sepsis recognition and management to final-year medical students. While this is encouraging, this study was neither designed nor powered to conclude noninferiority with certainty. This exploratory pilot study does suggest feasibility of an experimental design and may be beneficial to the design of a larger piece of research to consider noninferiority.

Comparison With Prior Work

Our findings align with the growing evidence base suggesting VR may be a viable alternative to traditional simulation methods in health care teaching. A systematic review and meta-analysis of VR applications in health care education found that VR use could achieve similar or better knowledge acquisition and skills scores as well as producing greater satisfaction and confidence compared to conventional teaching methods across multiple specialties [12]. The review identified wide heterogeneity and need for more studies using objective evaluation tools. Our study piloted directly comparing performance outcomes between modalities using a validated assessment tool, although formal validation of this tool for use in VR would be advisable prior to use in a larger study.

Studies have shown mixed results when comparing VR to traditional simulation with insufficient evidence to endorse 1 modality. A systematic review reported on 15 randomized control trials, with 2 demonstrating traditional simulation superiority, 4 showing VR simulation was superior, and 8 with comparable outcomes [13]. However, most of these studies focused on learning reactions or knowledge tests, arguably not the main intended learning outcome of simulation. Focusing on the "shows how" level of the pyramid of Miller [24], several studies demonstrated comparable outcomes between simulation modalities. A randomized control trial on managing status epilepticus found equivalent times to critical actions for VR and traditional simulation training [25]. The noninferiority of VR simulation was also found for nursing students learning ABCDE (Airway Breathing Circulation Disability Exposure) approach examinations [26] and for learning appropriate triage of casualties within mass casualty incidents [27]. Additionally, a multisimulation study comparing a package of 4 VR versus 4 traditional simulations for nursing students found that the VR group had significantly better knowledge outcomes and equivalent practical test (Objective Structured Clinical Examination) scores [28]. Our findings add to the growing body of evidence that VR simulation may provide equivalent outcomes to traditional simulation when teaching these integrated performance skills.

Much research comparing VR with traditional teaching methods using objective findings has primarily been on surgical skills acquisition [29] and resuscitation [14]. Systematic reviews on VR simulation to teach resuscitation found promising overall benefit for health care professions [14] and laypeople [30]. For non–health care professionals, a study demonstrated significantly better cardiopulmonary resuscitation training across several metrics within the VR group with learning retained 12 months post-training [31]. A systematic review of VR in surgical training found that VR showed benefits in multiple areas, including procedural times and postintervention scoring [32]. Our findings on sepsis management extend these results to another critical clinical scenario, suggesting that VR may effectively teach the recognition and management of septic shock and also the procedural steps required for the correct management and escalation of care.

Regarding specific outcomes for sepsis training, it has been shown for VR simulation-based education that sepsis management can be learned, with improvements in knowledge and interpersonal skills development [33]. Similarly, earlier work on VR sepsis simulation development evidence high perceived learning levels and optimum cognitive load [19]. Our study builds on this by demonstrating that VR might achieve similar recognition of sepsis and critical care needs. This is particularly relevant to VR's potential cost-effectiveness and scalability compared to traditional simulation methods [34-36].

XSL•FO
RenderX

The slightly higher recognition of critical care needs in the VR-only group compared to the IP-only group is an interesting finding that warrants further investigation. It is possible that this is a false-positive signal due to multiple comparisons made, though plausible explanations for a potentially causal relationship are also apparent. VR's immersive environment may enhance situational awareness and clinical decision-making, particularly for time-critical scenarios [37]. Indeed, exposure to a VR simulation in pediatrics significantly increased the recognition of respiratory distress [38] with a subsequent study demonstrating most participants found that the modality accurately depicted a decompensating patient [39]. Alternatively, the standardized nature of VR scenarios might provide more consistent exposure to specific clinical cues compared to IP simulations, which can vary based on facilitator and simulated patient performance [40]. A randomized control study highlighted repeatability as one of the positives of VR, which may have contributed to VR's positive study outcomes [28]. In the first instance, however, this finding should be interpreted as hypothesis-generating only, and further study would be needed to fully investigate the differential recognition of critical care need between these simulation modalities.

Our finding of only slight-to-moderate interrater reliability (Krippendorff α=−.155 to 0.491) for mQSAT scores is consistent with existing literature on simulation assessment challenges [41], although some studies report higher levels of interrater reliability [42]. A validity assessment of score inferences benefits from scoring and item-total correlations. While the mQSAT has demonstrated good reliability in previous studies with IP simulation [21], our findings suggest that assessment tools may need further validation across different simulation modalities.

Regarding procedural skills, we found no significant differences in Sepsis 6 bundle completion rates between the groups. Research has demonstrated that procedural skills can be effectively taught through various simulation modalities when coupled with structured debriefing [43,44]. This highlights the importance of the educational approach surrounding the technology rather than the technology itself, a finding echoed within medical education literature [45,46]. Research regarding virtual simulation specifically notes the importance of debrief regardless of the modality used, as it is central to learning outcomes [34,47]. Our findings reinforce this concept, suggesting that the structured debriefing process used in both IP and VR modalities influences the outcomes. This aligns with pedagogical frameworks emphasizing that technology serves merely as a vehicle for instruction, while learning outcomes are primarily determined by instructional design, facilitation quality, and debriefing approaches. The comparable performance across modalities in our study, despite the substantial technological differences between VR and IP simulation, underscores this principle.

Limitations

Our study had several limitations. The small sample size (n=32) across the 4 groups limited statistical power for detecting subtle differences, with imbalanced group sizes further constraining our analysis. The use of unequal group sizes would not be expected to bias effect estimates but reduces the precision of estimates for the smaller groups [48]. In our study, smaller sizes of arms estimating the learning effects of IP simulation may reduce our ability to distinguish learning differences between our cohort and those previously reported in the literature, impacting external validity. While pragmatically necessary, the cluster randomization method of a small sample potentially introduced selection bias despite broadly comparable baseline characteristics between the groups. The pilot nature of this comparative study, while useful for developing future research, does mean only limited conclusions can be drawn.

Our study did not use a classical pretest-posttest design, which may impact the ability to ascribe observed outcomes directly to the interventions studied. It has previously been suggested that a pretest may be undesirable due to, for example, influencing learning during the intervention, or familiarity with the assessment; however, it may be warranted when pretesting is an integral part of the intervention, when using a nonrandomized design, or for small sample sizes of less than 40 per interventional arm, in order to account for baseline differences between study groups [49]. We anticipate that our cluster randomization should have at least reduced, even if not entirely eliminated, baseline differences between the groups. Our sensitivity analysis using First mQSAT scores as the covariate in ANCOVA was designed to account for any residual baseline differences in simulation performance despite cluster randomization; this analysis did not change conclusions.

The mQSAT tool, though validated for IP simulation assessment, has not been previously validated for VR simulation scenarios. Modest interrater reliability suggests that assessment tools designed specifically for cross-modality evaluation may be needed. The validation of this tool within VR should be undertaken before use in larger studies to ensure it is appropriate, especially regarding areas that might be approached differently in VR versus IP simulation, such as communication. Additionally, our assessment focused on immediate performance without longitudinal follow-up, preventing conclusions about knowledge retention or transfer to clinical practice. The potential novelty effect of VR technology may have influenced student engagement and performance, which may not be found on follow-up assessment. However, there is initial evidence that some retention is found on longitudinal follow-up of VR-taught clinical skills 2 weeks [50] and 1 year later [51].

As a single-center study with a homogeneous population of final-year medical students, our findings have limited generalizability to different learner populations, institutions, or health care systems with varying simulation resources and curricula.

Implications for Practice

Despite these limitations, our findings have several important implications for medical education. The comparable performance between VR and IP simulation groups adds to the evidence base that VR may be a viable alternative for clinical management when traditional simulation resources are limited or inaccessible. This could be particularly valuable for institutions with geographic, financial, or staffing constraints that limit access to high-fidelity simulation centers.

XSL•FO
RenderX

The higher recognition of critical care needs in the VR-only group is an interesting finding requiring further investigation but could represent the more life-like appearance of a sick patient constructed in VR compared to the normally well-looking IP simulation representative. Educational designers might consider using VR specifically for scenarios that emphasize the recognition of deterioration and escalation of care. The similar completion rates of Sepsis 6 components across groups suggest that procedural aspects of care can be taught effectively with either modality. However, institutions should consider the specific educational objectives, available resources, and learner characteristics when selecting between simulation approaches.

Future Research

This pilot study provides valuable data for designing future larger-scale trials. A fully powered noninferiority trial with a larger sample size would allow more definitive conclusions about the relative effectiveness of VR versus IP simulation for clinical education. Such a trial should include longitudinal follow-up to assess knowledge retention and, ideally, impact on clinical practice.

Future research should also explore the optimal integration of VR and IP simulation within a comprehensive curriculum. Rather than viewing these as competing modalities, a blended approach might leverage the strengths of each method for different learning objectives or at different stages of training. VR may be particularly valuable for independent practice and reinforcement after initial IP simulation training. Finally, cost-effectiveness analyses comparing VR and IP simulation would further inform educational policy and resource allocation decisions. While VR has potential scalability advantages, the initial development and technology costs should be weighed against the long-term benefits.

Conclusions

This pilot study found no significant differences in sepsis management performance between medical students trained using VR versus IP simulation, with the exception of slightly better recognition of critical care needs in the VR-only group. These findings indicate that VR simulation performed similarly to IP simulation for teaching sepsis management to final-year medical students, though larger studies are needed for definitive conclusions. As simulation technology continues to evolve, educational approaches that strategically integrate various modalities based on specific learning objectives, available resources, and learner needs are likely to be most effective. VR simulation shows promise as a complementary or alternative approach to traditional simulation methods, potentially expanding access to high-quality simulation education across diverse educational settings.

Funding

Virtual reality simulation scenarios and headsets were provided free of charge to GWH Academy by Goggleminds as part of beta testing and codevelopment, including the sepsis simulation used in this study. Goggleminds remains the primary developer of the software used; Goggleminds retains intellectual property rights for the scenarios, and the study authors or GWH Academy hold no financial interests in Goggleminds or its products. Goggleminds were not involved in the inception, design, performance, analysis, or reporting of this study. There are no other links, financial or otherwise, between the study authors and Goggleminds.

Data Availability

The datasets generated or analyzed during this study are not publicly available due to limited sample size and timeframe of collections risking the identification of individual participants but are available from the corresponding author on reasonable request.

The OSF repository for this study can be found at [17].

Conflicts of Interest

None declared.

Multimedia Appendix 1
Scenarios and questionnaires.
[DOCX File, 426 KB - mededu_v12i1e80316_app1.docx]

Multimedia Appendix 2
Baseline testing.
[DOCX File, 29 KB - mededu_v12i1e80316_app2.docx]

Multimedia Appendix 3
Sensitivity analyses.
[DOCX File, 28 KB - mededu_v12i1e80316_app3.docx]

Checklist 1

XSL•FO
RenderX

CONSORT checklist.
[PDF File, 706 KB - mededu_v12i1e80316_app4.pdf]

References

1. Liu K, Zhang W, Li W, Wang T, Zheng Y. Effectiveness of virtual reality in nursing education: a systematic review and meta-analysis. BMC Med Educ 2023 Sep 28;23(1):710. [doi: 10.1186/s12909-023-04662-x] [Medline: 37770884]
2. Curran VR, Xu X, Aydin MY, Meruvia-Pastor O. Use of extended reality in medical education: an integrative review. Med Sci Educ 2022 Feb;33(1):275-286. [doi: 10.1007/s40670-022-01698-4] [Medline: 36569366]
3. Walls R, Nageswaran P, Cowell A, et al. Virtual reality as an engaging and enjoyable method for delivering emergency clinical simulation training: a prospective, interventional study of medical undergraduates. BMC Med 2024 Jun 3;22(1):222. [doi: 10.1186/s12916-024-03433-9] [Medline: 38831293]
4. Ropponen P, Tomietto M, Pramila-Savukoski S, et al. Impacts of VR simulation on nursing students' competence, confidence, and satisfaction: a systematic review and meta-analysis of randomised controlled trials. Nurse Educ Today 2025 Sep;152:106756. [doi: 10.1016/j.nedt.2025.106756] [Medline: 40279771]
5. Bray K, Bray L, Sathyapalan T, Rigby AS, Hepburn D. A pilot study comparing immersive virtual reality simulation and computerized virtual patient simulation in undergraduate medical education. Int J Healthc Simul 2023;10. [doi: 10.54531/rxca9513]
6. Mok TN, Chen J, Pan J, et al. Use of a virtual reality simulator for tendon repair training: randomized controlled trial. JMIR Serious Games 2021 Jul 12;9(3):e27544. [doi: 10.2196/27544] [Medline: 34255649]
7. Hu KC, Salcedo D, Kang YN, et al. Impact of virtual reality anatomy training on ultrasound competency development: a randomized controlled trial. PLoS One 2020;15(11):e0242731. [doi: 10.1371/journal.pone.0242731] [Medline: 33227037]
8. Dubovi I. Cognitive and emotional engagement while learning with VR: the perspective of multimodal methodology. Comput Educ 2022 Jul;183:104495. [doi: 10.1016/j.compedu.2022.104495]
9. Gillespie GL, Farra S, Regan SL, Brammer SV. Impact of immersive virtual reality simulations for changing knowledge, attitudes, and behaviors. Nurse Educ Today 2021 Oct;105:105025. [doi: 10.1016/j.nedt.2021.105025] [Medline: 34175566]
10. Elendu C, Amaechi DC, Okatta AU, et al. The impact of simulation-based training in medical education: a review. Medicine (Baltimore) 2024 Jul 5;103(27):e38813. [doi: 10.1097/MD.0000000000038813] [Medline: 38968472]
11. Burton R, Hope A. Simulation based education and expansive learning in health professional education: a discussion. J Appl Learn Teach 2018 Jun 1;1(1):25-34. [doi: 10.37074/jalt.2018.1.1.4]
12. Sung H, Kim M, Park J, Shin N, Han Y. Effectiveness of virtual reality in healthcare education: systematic review and meta-analysis. Sustainability 2024 Jan;16(19):8520. [doi: 10.3390/su16198520]
13. Foronda CL, Gonzalez L, Meese MM, et al. A comparison of virtual reality to traditional simulation in health professions education: a systematic review. Simul Healthc 2024 Jan 1;19(1S):S90-S97. [doi: 10.1097/SIH.0000000000000745] [Medline: 37651101]
14. Trevi R, Chiappinotto S, Palese A, Galazzi A. Virtual reality for cardiopulmonary resuscitation healthcare professionals training: a systematic review. J Med Syst 2024 May 15;48(1):50. [doi: 10.1007/s10916-024-02063-1] [Medline: 38748244]
15. Clinical trials registration. ICMJE. URL: https://www.icmje.org/about-icmje/faqs/clinical-trials-registration/ [accessed 2026-02-07]
16. Hemming K, Taljaard M. Key considerations for designing, conducting and analysing a cluster randomized trial. Int J Epidemiol 2023 Oct 5;52(5):1648-1658. [doi: 10.1093/ije/dyad064] [Medline: 37203433]
17. GWH VR vs IP simulation. OSF. URL: https://osf.io/wujdn/?view_only=203a688fa9c04548a40d6a5a2e934cee [accessed 2026-02-24]
18. Start of virtual reality training at Great Western Hospitals NHS Foundation Trust to help spot signs of sepsis. Great Western Hospital. URL: https://www.gwh.nhs.uk/news/posts/2024/january/start-of-virtual-reality-training-at-great-western-hospitals-nhs-foundation-trust-to-help-spot-signs-of-sepsis/ [accessed 2026-02-23]
19. Jacobs C, Vaidya K, Medwell L, Old T, Joiner R. Case study of virtual reality sepsis management-instructional design and ITEM outcomes. J Vis Commun Med 2023 Jul;46(3):168-177. [doi: 10.1080/17453054.2023.2280611] [Medline: 37990908]
20. Hall AK, Dagnone JD, Lacroix L, Pickett W, Klinger DA. Queen's Simulation Assessment Tool: development and validation of an assessment tool for resuscitation objective structured clinical examination stations in emergency medicine. Simul Healthc 2015 Apr;10(2):98-105. [doi: 10.1097/SIH.0000000000000076] [Medline: 25710317]
21. Yu JH, Lee MJ, Kim SS, et al. Assessment of medical students' clinical performance using high-fidelity simulation: comparison of peer and instructor assessment. BMC Med Educ 2021 Sep 25;21(1):506. [doi: 10.1186/s12909-021-02952-w] [Medline: 34563180]
22. Hemming K, Girling AJ, Sitch AJ, Marsh J, Lilford RJ. Sample size calculations for cluster randomised controlled trials with a fixed number of clusters. BMC Med Res Methodol 2011 Jun 30;11(1):102. [doi: 10.1186/1471-2288-11-102] [Medline: 21718530]

XSL•FO
RenderX

23. Van Breukelen GJP. Cluster randomized trials with a pretest and posttest: equivalence of three-, two-and one-level analyses, and sample size calculation. Multivariate Behav Res 2024;59(2):206-228. [doi: 10.1080/00273171.2023.2240779] [Medline: 37590444]
24. Miller GE. The assessment of clinical skills/competence/performance. Acad Med 1990 Sep;65(9 Suppl):S63-S67. [doi: 10.1097/00001888-199009000-00045] [Medline: 2400509]
25. Abulfaraj MM, Jeffers JM, Tackett S, Chang T. Virtual reality vs. high-fidelity mannequin-based simulation: a pilot randomized trial evaluating learner performance. Cureus 2021 Aug;13(8):e17091. [doi: 10.7759/cureus.17091] [Medline: 34527478]
26. Berg H, Steinsbekk A. Is individual practice in an immersive and interactive virtual reality application non-inferior to practicing with traditional equipment in learning systematic clinical observation? A randomized controlled trial. BMC Med Educ 2020 Apr 22;20(1):123. [doi: 10.1186/s12909-020-02030-7] [Medline: 32326948]
27. Ferrandini Price M, Escribano Tortosa D, Nieto Fernandez-Pacheco A, et al. Comparative study of a simulated incident with multiple victims and immersive virtual reality. Nurse Educ Today 2018 Dec;71:48-53. [doi: 10.1016/j.nedt.2018.09.006] [Medline: 30241022]
28. Kiegaldie D, Shaw L. Virtual reality simulation for nursing education: effectiveness and feasibility. BMC Nurs 2023 Dec 19;22(1):488. [doi: 10.1186/s12912-023-01639-5] [Medline: 38114940]
29. Tang YM, Chau KY, Kwok APK, Zhu T, Ma X. A systematic review of immersive technology applications for medical practice and education—trends, application areas, recipients, teaching contents, evaluation methods, and performance. Educ Res Rev 2022 Feb;35:100429. [doi: 10.1016/j.edurev.2021.100429]
30. Alcázar Artero PM, Pardo Rios M, Greif R, et al. Efficiency of virtual reality for cardiopulmonary resuscitation training of adult laypersons: a systematic review. Medicine (Baltimore) 2023 Jan 27;102(4):e32736. [doi: 10.1097/MD.0000000000032736] [Medline: 36705392]
31. Zhang N, Liu YJ, Yang C, et al. Long-term effect of intelligent virtual reality first-aid training on cardiopulmonary resuscitation skill proficiency. Prehosp Emerg Care 2026;30(1):24-30. [doi: 10.1080/10903127.2025.2465718] [Medline: 39937876]
32. Mao RQ, Lan L, Kay J, et al. Immersive virtual reality for surgical training: a systematic review. J Surg Res 2021 Dec;268:40-58. [doi: 10.1016/j.jss.2021.06.045] [Medline: 34284320]
33. Liaw SY, Tan JZ, Lim S, et al. Artificial intelligence in virtual reality simulation for interprofessional communication training: mixed method study. Nurse Educ Today 2023 Mar;122:105718. [doi: 10.1016/j.nedt.2023.105718] [Medline: 36669304]
34. Pottle J. Virtual reality and the transformation of medical education. Future Healthc J 2019 Oct;6(3):181-185. [doi: 10.7861/fhj.2019-0036] [Medline: 31660522]
35. Plotzky C, Lindwedel U, Sorber M, et al. Virtual reality simulations in nurse education: a systematic mapping review. Nurse Educ Today 2021 Jun;101:104868. [doi: 10.1016/j.nedt.2021.104868] [Medline: 33798987]
36. Bumbach MD, Culross BA, Datta SK. Assessing the financial sustainability of high-fidelity and virtual reality simulation for nursing education: a retrospective case analysis. Comput Inform Nurs 2022 Sep 1;40(9):615-623. [doi: 10.1097/CIN.0000000000000916] [Medline: 35524349]
37. Chandanani M, Laidlaw A, Brown C. Extended reality and computer-based simulation for teaching situational awareness in undergraduate health professions education: a scoping review. Adv Simul (Lond) 2025 Apr 2;10(1):18. [doi: 10.1186/s41077-025-00343-5] [Medline: 40176179]
38. Zackoff MW, Real FJ, Sahay RD, et al. Impact of an immersive virtual reality curriculum on medical students' clinical assessment of infants with respiratory distress. Pediatr Crit Care Med 2020 May;21(5):477-485. [doi: 10.1097/PCC.0000000000002249] [Medline: 32106189]
39. Zackoff MW, Cruse B, Sahay RD, et al. Development and implementation of augmented reality enhanced high-fidelity simulation for recognition of patient decompensation. Simul Healthc 2021 Jun 1;16(3):221-230. [doi: 10.1097/SIH.0000000000000486] [Medline: 32910102]
40. Yauger SJ, Konopasky A, Battista A. Reliability in healthcare simulation setting: a definitional review. Cureus 2020 May 14;12(5):e8111. [doi: 10.7759/cureus.8111] [Medline: 32542164]
41. Tavares W, Brydges R, Myre P, et al. Applying Kane's validity framework to a simulation based assessment of clinical competence. Adv Health Sci Educ Theory Pract 2018 May;23(2):323-338. [doi: 10.1007/s10459-017-9800-3]
42. Devitt JH, Kurrek MM, Cohen MM, et al. Testing the raters: inter-rater reliability of standardized anaesthesia simulator performance. Can J Anaesth 1997 Sep;44(9):924-928. [doi: 10.1007/BF03011962] [Medline: 9305554]
43. Huang J, Tang Y, Tang J, et al. Educational efficacy of high-fidelity simulation in neonatal resuscitation training: a systematic review and meta-analysis. BMC Med Educ 2019 Aug 29;19(1):323. [doi: 10.1186/s12909-019-1763-z] [Medline: 31464614]
44. Lindamood KE, Weinstock P. Application of high-fidelity simulation training to the neonatal resuscitation and pediatric advanced life support programs. Newborn Infant Nurs Rev 2011 Mar;11(1):23-27. [doi: 10.1053/j.nainr.2010.12.010]
45. Sandars J. It appeared to be a good idea at the time but … A few steps closer to understanding how technology can enhance teaching and learning in medical education. Med Teach 2011;33(4):265-267. [doi: 10.3109/0142159X.2011.550972] [Medline: 21456982]

XSL•FO
RenderX

46. Sandars J, Patel RS, Goh PS, Kokatailo PK, Lafferty N. The importance of educational theories for facilitating learning when using technology in medical education. Med Teach 2015;37(11):1039-1042. [doi: 10.3109/0142159X.2015.1019438] [Medline: 25776228]
47. Luctkar-Flude M, Tyerman J, Verkuyl M, et al. Effectiveness of debriefing methods for virtual simulation: a systematic review. Clin Simul Nurs 2021 Aug;57:18-30. [doi: 10.1016/j.ecns.2021.04.009]
48. Elkins M, Soh SE. Research Note: unequal randomisation in randomised trials. J Physiother 2026 Jan;72(1):75-78. [doi: 10.1016/j.jphys.2025.11.009] [Medline: 41419406]
49. Cook DA, Beckman TJ. Reflections on experimental research in medical education. Adv Health Sci Educ Theory Pract 2010 Aug;15(3):455-464. [doi: 10.1007/s10459-008-9117-3] [Medline: 18427941]
50. Butt AL, Kardong-Edgren S, Ellertson A. Using game-based virtual reality with haptics for skill acquisition. Clin Simul Nurs 2018 Mar;16:25-32. [doi: 10.1016/j.ecns.2017.09.010]
51. Gan W, Mok TN, Chen J, et al. Researching the application of virtual reality in medical education: one-year follow-up of a randomized trial. BMC Med Educ 2023 Jan 3;23(1):3. [doi: 10.1186/s12909-022-03992-6] [Medline: 36597093]

Abbreviations

ABCDE: Airway Breathing Circulation Disability Exposure
ANCOVA: analysis of covariance
CONSORT: Consolidated Standards of Reporting Trials
GWH: Great Western Hospital
IP: in-person
mQSAT: Queen's Simulation Assessment Tool
VR: virtual reality

Edited by T Gladman; submitted 18.Jul.2025; peer-reviewed by D Yoo, X Liang; revised version received 08.Feb.2026; accepted 09.Feb.2026; published 30.Mar.2026.

Please cite as:
Medwell L, Old T, Ahmed A, Holloway V, McTaggart L, Morgan D, O'Keeffe C, Olori J, Payne E, Rainforth C, Ramanujam N, Vaidya K, Jacobs C
Virtual Reality Versus In-Person Simulation of Sepsis for Medical Students: Randomized Comparative Pilot Study
JMIR Med Educ 2026;12:e80316
URL: https://mededu.jmir.org/2026/1/e80316
doi:10.2196/80316

© Lauren Medwell, Tim Old, Awais Ahmed, Victoria Holloway, Lauren McTaggart, Dafydd Morgan, Callum O'Keeffe, Jemaima Olori, Ellie Payne, Christopher Rainforth, Namritha Ramanujam, Kriti Vaidya, and Chris Jacobs. Originally published in JMIR Medical Education (https://mededu.jmir.org), 30.Mar.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Education, is properly cited. The complete bibliographic information, a link to the original publication on https://mededu.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Remote Augmented Reality Versus Traditional Simulation for Team Leader Assessment in a Cardiac Arrest Scenario: Noninferiority Randomized Controlled Trial

Renan Gianotto-Oliveira[1*], PhD; Marcos Rojas[2*], MD; Maria Queiroz[1], MSc; Flávia Zanchetta[1], PhD; Anabel Ferrari[1], MD; Lucas Kojima[1], MD; Alexandre Paula[1], MD; Bruno Campos[1], MD; Dario Cecilio-Fernandes[1,3], PhD; Thomas J Caruso[2], PhD

[1]School of Medical Sciences, University of Campinas, Campinas, Brazil

[2]Department of Anesthesiology, Perioperative, and Pain Medicine, Stanford School of Medicine, 453 Quarry Road, office 422A, Palo Alto, CA, United States

[3]Erasmus MC University Medical Centre Rotterdam, Institute of Medical Education Research Rotterdam, Rotterdam, The Netherlands

[*]these authors contributed equally

Corresponding Author:
Marcos Rojas, MD
Department of Anesthesiology, Perioperative, and Pain Medicine, Stanford School of Medicine, 453 Quarry Road, office 422A, Palo Alto, CA, United States

Abstract

Background: Simulation-based education is crucial for training health care professionals in advanced cardiac life support. However, access to high-fidelity in-person simulation is frequently limited by geographic, logistical, and financial constraints. Augmented reality (AR) offers the potential to deliver remote, immersive training experiences that may overcome these barriers, but its effectiveness compared with traditional simulation remains uncertain.

Objective: This study aimed to determine whether remote AR simulation is noninferior to traditional in-person simulation for assessing team leader performance during a ventricular fibrillation cardiac arrest scenario.

Methods: This noninferiority randomized trial enrolled participants at the State University of Campinas (UNICAMP), Brazil, and used cross-continental remote instruction from Stanford University (in the United States) for the AR arm. A total of 50 health care professionals were randomized to either remote AR simulation with a geographically distant instructor (n=25) or traditional in-person simulation (n=25). All participants completed an identical ventricular fibrillation cardiac arrest case as team leaders. Leader performance was assessed using an adapted, validated checklist-based instrument for cognitive leadership and an observational behavioral measure (Behaviorally Anchored Rating Scale). Secondary outcomes included AR participants' evaluations of usability and ergonomics.

Results: A total of 42 participants fully completed the study procedures (remote AR group: n=22; traditional in-person group: n=20). The AR group demonstrated noninferior performance compared to the traditional group across all outcomes. The mean checklist scores were 41.6 (SD 6.2) and 42.6 (SD 5.8) in the remote AR group and traditional in-person group, respectively. The AR group's 95% CI (38.9 - 44.4) was above the 20% noninferiority threshold of 34.1. Usability and ergonomics were favorably reported by most participants.

Conclusions: Participants in the remote AR simulation demonstrated noninferior team leader decision-making and behavioral performance compared with those in traditional in-person simulation. These findings suggest that remote AR may be a viable strategy to expand access to scenario-based assessment of cardiac arrest leadership, particularly in resource-limited settings. AR participants also reported high usability and low ergonomic burden, indicating comfortable headset use.

Trial Registration: ClinicalTrials.gov NCT06326450; https://clinicaltrials.gov/study/NCT06326450

(JMIR Med Educ 2026;12:e84367) doi:10.2196/84367

KEYWORDS

augmented reality; AR; simulation-based education; advanced cardiac life support; resuscitation; digital health education; randomized controlled trial

XSL•FO
RenderX

Introduction

In an advanced cardiac life support (ACLS), the performance of the team leader is a critical determinant of resuscitation quality and patient outcomes. Effective leadership requires rapid clinical decision-making, clear communication, and coordinated task delegation to optimize team performance under pressure [1]. Deficiencies in leadership behaviors can compromise adherence to resuscitation algorithms and reduce overall team efficiency, even when individual team members possess strong technical skills [2,3]. Simulation-based assessments provide structured environments to evaluate these competencies objectively, integrating both technical and nontechnical skills essential for safe patient care [4,5].

Despite its educational value, traditional in-person simulation faces persistent barriers. High costs, logistical complexity, and the requirement for physical facilities and expert facilitators limit its scalability, particularly in resource-constrained or geographically dispersed settings [6-8]. These challenges highlight the need for accessible strategies capable of delivering high-fidelity assessment without the infrastructure demands of traditional simulation.

Immersive technologies offer innovative pathways to mitigate these limitations. Augmented reality (AR) simulation has gained traction in health professions education as a way to deliver interactive scenarios for learners. In contrast to traditional virtual reality (VR), which replaces the surrounding environment with a fully virtual setting, AR augments the real world with digital information while keeping learners embedded in a shared physical space. Newer VR headsets now use a video "pass through" mode, replacing fully computer-generated imagery in VR with the natural world, simulating an AR experience. This distinction matters for clinical training that depends on direct interaction with teammates, natural face-to-face communication, and the manipulation of authentic equipment and task trainers, features that can be harder to preserve when the environment is fully virtual. By overlaying interactive digital elements onto real-world settings through head-mounted displays (HMDs), AR enables learners to engage in realistic, team-based scenarios while preserving face-to-face communication and allowing real-time remote facilitation [9]. This ability to blend virtual overlays with task trainers and clinical equipment enables mixed reality experiences that can mirror key features of conventional simulation. Despite growing adoption, existing studies have less often examined AR through an instructional design lens or clearly demonstrated what educational benefit AR provides beyond other simulation approaches [10,11]. AR preserves the essential elements of conventional simulation, such as realism, interactivity, and teamwork, while providing a practical alternative to in-person assessments.

Because this modality requires minimal equipment and setup time, AR can be implemented flexibly in multipurpose clinical spaces, reducing scheduling constraints that often restrict access to dedicated simulation centers. Prior studies suggest that it enhances learner engagement and reproduces essential elements of pediatric advanced life support and ACLS, including leadership under pressure and team coordination [12]. Participants report high usability, and AR simulations can support complex task execution [13,14].

Remote instruction and assessment through AR simulations offers promising potential for providing standardized evaluation of clinical performance in settings where traditional simulation is costly, difficult to scale, or inaccessible. To address this gap, we conducted an international, randomized noninferiority trial evaluating team leader performance in a ventricular fibrillation (VF) cardiac arrest scenario. The primary outcome was to determine whether AR-based assessment yielded noninferior leadership performance compared with traditional in-person simulation. Secondary outcomes measured the AR participants' evaluation of usability and ergonomics.

Methods

Design

This international trial used a parallel, noninferiority design with a 1:1 allocation ratio to compare team leader performance in a simulated VF cardiac arrest scenario. All participants were recruited and completed study procedures at the State University of Campinas (UNICAMP), Brazil. Stanford University functioned as the remote instructor site for the AR arm. Participants were randomized to complete either a remote instructor–led AR simulation or a traditional in-person simulation. Recruitment occurred between June 20, 2024, and August 2, 2024, across 2 academic centers. The study was prospectively registered at ClinicalTrials.gov (NCT06326450; March 15, 2024).

Following trial commencement, no significant modifications were made to the study protocol, including eligibility criteria or protocol design, ensuring uniformity throughout the investigation. Participants received consistent instructions, regardless of their group. Minor logistical adjustments to ensure smooth operation during data collection did not affect the trial's scientific integrity. This manuscript adheres to the CONSORT (Consolidated Standards of Reporting Trials) guidelines for noninferiority trials (Checklist 1) [15].

Ethical Considerations

The trial was conducted in accordance with the ethical standards of the participating institutions and with the principles of the Declaration of Helsinki. Approval was obtained from the Institutional Review Board at UNICAMP, Brazil (CAAE 79474024.7.0000.5404; approval: 6.893.573; June 18, 2024). Institutional approval was also secured at the collaborating site in the United States prior to participant enrollment. All participants provided informed consent before participation and were assured of confidentiality and the voluntary nature of their involvement. They did not receive any financial compensation or material incentives for their participation in the study.

Participants

A faculty member not involved in data collection recruited residents of internal medicine and emergency medicine at UNICAMP because they require routine ACLS training. The recruitment of internal medicine and emergency medicine residents from the same institution resulted in a relatively

XSL•FO
RenderX

homogeneous participant population, with comparable baseline exposure to ACLS training across groups, consistent with the study's focus on comparing assessment modalities. Participants in both the traditional in-person group and remote AR group received instruction from UNICAMP faculty instructors, although the AR group's UNICAMP instructor was remotely located at Stanford's School of Medicine. Participants were excluded if they had a history of severe motion sickness, current nausea, a history of seizures, or required corrective glasses that would interfere with the use of AR HMD. Participants requiring corrective glasses that interfered with the safe fitting of the AR headset were excluded from the remote AR group due to device compatibility constraints. This exclusion was applied after randomization when headset fitting was attempted. Potential harms, such as physical discomfort, disorientation, or adverse reactions to the AR headset, were predefined and monitored, with participants instructed to report any symptoms immediately during or after the simulation. Recruitment and invitations to participate were conducted through direct oral communication. Data collection occurred in the UNICAMP simulation laboratory.

Randomization and Blinding

Participants were randomly assigned to either the traditional in-person group or the remote AR group, using a computer-generated randomization sequence (GraphPad Software, Inc). Randomization was performed in a 1:1 allocation ratio to ensure an equal distribution of participants between the 2 groups, without matching or stratification by specialty, year of training, or prior clinical experience. To maintain allocation concealment and minimize selection bias, the randomization sequence was generated by an independent researcher who was not involved in the recruitment, enrollment, or data analysis processes. The independent researcher had no access to participant identities, preserving impartiality in group assignment. Raters of participants' performance were not blinded to group allocation because ratings were performed live in the simulation setting and the assigned modality was visually apparent. Raters were aware of the study procedures and instruments but were not informed of the expected direction of results. The researchers conducting the statistical analyses were blinded to group assignments, using coded datasets to prevent bias while interpreting results.

Interventions

After screening, trained research assistants provided informed consent to participants, who then completed demographic questionnaires. Participants were randomly assigned to either the traditional, in-person simulation or the remote-instructor AR simulation. Hereafter, these will be referred to as the traditional in-person group and the remote AR group, respectively. Participants were informed of their group assignments after signing the informed consent form.

To ensure a comparable baseline skill set while preserving the assessment-focused nature of the study, all participants viewed the same brief, scripted 10-minute instructional video immediately before the simulation. This video reviewed core principles of effective team leadership, closed-loop communication, and the VF algorithm, based on the 2020 American Heart Association Guidelines [16]. It did not include hands-on practice, rehearsal, or individualized feedback. This minimal standardization was intended to reduce variability related to prior exposure while avoiding training for the test. Then, each simulation featured 1 participant serving as the team leader, coordinating 3 actors assigned to specific resuscitation roles: chest compressions, ventilation, and defibrillation or medications. These actors were nurses and physicians who acted in scripted supporting roles consistent with standard resuscitation team structures. This design simulated a realistic interprofessional team, consistent with contemporary best practices in resuscitation training [17]. Although the evaluation instrument includes items referring to actions performed by resuscitation team members, these elements were used to operationalize assessment of team leader performance rather than to evaluate individual technical skills. All hands-on tasks were executed by trained actor-participants following predefined scripts, while the team leader was responsible for recognizing deficiencies, issuing corrective instructions, coordinating task allocation, and ensuring adherence to the resuscitation algorithm. The same actors participated in both the AR and traditional scenarios.

The simulation followed a standard, scripted sequence of events simulating an unexpected cardiac arrest (Multimedia Appendix 1) [18]. In both groups, the participant's role was to lead the resuscitation, including defibrillation and administration of medications such as epinephrine and amiodarone. The instructor facilitated the simulation by controlling the progression of the scenario and updating clinical parameters in real time, in response to the participant's decisions. The instructor did not provide coaching, feedback, or suggestions regarding clinical actions during the scenario.

In the traditional group, the simulation was facilitated by an instructor who was physically present with the participant in the simulation room. The equipment used included an adult manikin (MegaCode; Laerdal Medical, Inc), a manual defibrillator, a medication cart, a bag-valve-mask, advanced airway equipment, and a stopwatch.

In the remote AR group, research assistants fitted participants with a Magic Leap One (ML1) AR HMD (Magic Leap, Inc). The ML1 used the Chariot AR Medical simulation software (Stanford Chariot Program) which integrates real-time communication into a portable AR medical simulator featuring a holographic patient, monitor, and other medical equipment. The holographic patient's chest was aligned directly over a real compression torso to provide tactile feedback for the actors. The software has effectively delivered medical simulations in different contexts [12-14,19,20]. Research assistants oriented participants to the ML1 and Chariot AR medical simulation software functionalities before the simulation. Orientation to the AR system required approximately 5 to 10 minutes and included headset fitting, calibration, and a brief walkthrough of the Chariot AR interface and communication features. Each remote AR session used 4 HMDs: 1 ML1 worn by the remote instructor at Stanford and 3 ML1 devices at UNICAMP worn by the team leader and two of the actor-participants. Room requirements were consistent with standard resuscitation simulation: a simulation room with adequate space for the team

XSL•FO
RenderX

and equipment, a bed or table to simulate a stretcher, and reliable broadband internet connectivity to support real-time remote facilitation. The participants experienced the same clinical scenario as the in-person group. However, instead of a mannequin, they interacted with a holographic patient and a holographic monitor (Figure 1). In addition, the instructor performed this facilitation role remotely from Stanford, using a headset to interact in real time with participants located in Brazil.

In both groups, the instructor adjusted vital signs, pulses, and other clinical cues in real time according to the participant's actions. In both cases, the simulation concluded with the return of spontaneous circulation, evidenced by the presence of a pulse and breathing. The scenario had an average duration of 12 (SD 1.1) minutes. At the end, all participants received a structured debriefing and completed the postsimulation questionnaires.

The primary risk associated with the ML1 headset was minor physical discomfort, such as pressure or irritation from wearing the device. To minimize this risk, trained staff ensured proper fitting and adjustment for each participant. The ML1 is a commercially available headset that has undergone manufacturer testing and regulatory review to ensure safety and compliance. Overall, the study posed minimal risk to participants. All participants were informed of their right to withdraw at any time without penalty, preserving their autonomy and comfort throughout the trial.

Figure 1. Example view of the augmented reality (AR) simulation environment as seen through the Magic Leap One headset, showing the holographic patient and monitor used during training.



Outcomes

Demographic data, collected via Google Forms prior to study initiation, included age, gender, race, level of training, prior exposure to in-person simulations, and prior exposure to AR. The primary outcomes were assessed using two validated instruments: (1) an adapted checklist-based tool, derived from the Evaluation of Advanced Life Support Performance instrument, measuring cognitive leadership and adherence to resuscitation algorithms; and (2) the Behaviorally Anchored Rating Scale (BARS), evaluating observable nontechnical behaviors.

For the adapted checklist, we calculated 2 types of scores for each participant—objective and subjective—based on the original instrument's 26-item checklist and global rating scale (Multimedia Appendix 1) [4]. Although originally developed to assess both technical and nontechnical skills, in this study, the instrument was adapted specifically to evaluate cognitive performance and scenario leadership, such as clinical decision-making, adherence to resuscitation protocols, and effective team coordination. No psychomotor or procedural technical skills were assessed, as all hands-on resuscitation tasks were performed by trained health care professional actors under the direction of the participant. Accordingly, we excluded

XSL•FO
RenderX

checklist elements that require direct evaluation of team members' psychomotor execution (eg, chest compression and ventilation technique or quality metrics). We retained items that can be attributed to the team leader's cognition and directives, including recognition and prioritization of actions, algorithm adherence (eg, rhythm recognition or defibrillation and medication timing), situational awareness, and team coordination or closed-loop communication. The checklist items, each scored from –2 to +2, covered two domains: (1) adherence to cardiopulmonary resuscitation guidelines and (2) clinical decision making. The objective score was obtained by summing the ratings across the 26 items, reflecting the participant's ability to manage the crisis effectively. The subjective score was a global rating from 0 to 10, assigned by the rater at the end of the simulation to summarize overall performance. These 2 scores were independently assigned by 2 trained raters during each session, and final scores were calculated as the average of both raters' assessments.

BARS is a tool designed to assess observable behavioral competencies and combines qualitative and quantitative assessment by linking specific behavioral examples to numerical ratings. The four domains of this instrument are: (1) vigilance situation-aware, (2) decision-making, (3) communication, and (4) teamwork (Multimedia Appendix 1) [21]. BARS rates these domains as "poor," "average," or "excellent" and assigns a score ranging from 1 to 9. Two trained raters assessed behavioral competencies using the BARS scale in real time during and immediately after each simulation session, offering a complementary perspective on participants' nontechnical performance. The BARS domain scores were prespecified as secondary noninferiority outcomes to provide supportive evidence on nontechnical leadership performance.

Assessments were completed live during each simulation session by 2 trained raters, and no video was recorded. Prior to data collection, raters completed 3 calibration sessions (trial runs) in which they practiced scoring using the adapted checklist and BARS, reviewed the intent of each item and domain, and discussed discrepancies with the research team to standardize scoring.

The secondary outcomes included usability and ergonomic perceptions of the AR system for participants in the AR group. Usability was assessed using the system usability scale (SUS; Multimedia Appendix 1) [22]. The SUS includes 10 items; each rated on a 5-point Likert scale from 1 ("Completely Disagree") to 5 ("Completely Agree"). Ergonomics were measured using the International Organization for Standardization (ISO) 9241 - 400 scale (Multimedia Appendix 1) which is composed of 6 items, each rated on a 5-point Likert scale from 1 ("Totally Disagree") to 5 ("Totally Agree"). AR group participants completed these questionnaires at the end of the simulation.

Sample Size

A typical mean score in the Evaluation of Advanced Life Support Performance instrument during reasonable simulation performance is 7.875, with an IQR from 7 to 8.75 [18]. Importantly, this estimate was derived from the instrument's 0 to 10 subjective or global rating scale, which was retained unchanged in our adapted version; thus, the sample size calculation does not depend on the adapted objective checklist score (0 - 52). Given a noninferiority margin of 20% and assuming a SE of 1.3, 16 participants (8 in each group) were needed to achieve 80% power with a 1-sided 95% CI. To account for incomplete data and dropouts, a total of 30 participants (15 in each group) was set as the minimal enrollment target. However, given that participants were recruited from a single institution's residency program, no upper limit was imposed on enrollment. All interested volunteers were allowed to participate, even after the minimum sample size required for statistical power had been reached.

Statistical Analysis

The primary analysis was per-protocol and included all randomized participants who completed the assigned simulation session with a fully evaluable checklist and BARS outcomes. Participants who did not attend the scheduled session had no outcome measurements and were not included in outcome analyses.

We conducted a noninferiority analysis using a 20% noninferiority margin based on the in-person group's mean score. Because there is no established minimally important educational difference for the adapted checklist or BARS in ACLS leadership assessment, we prespecified a 20% noninferiority margin as a pragmatic threshold, based on consensus among the study team's ACLS instructors, to define acceptable similarity to the in-person benchmark. Noninferiority was established if the lower bound of the 95% CI for the AR group's mean score exceeded this threshold. Performance scores from 2 independent raters were averaged for analysis, and interrater reliability (IRR) was assessed using the intraclass correlation coefficient (ICC). ICC values between 0.50 and 0.75 were considered moderate, 0.75 to 0.90 as good, and values above 0.90 as excellent [23]. We did not compute internal consistency indices (eg, Cronbach α) because the checklist is a structured performance rubric spanning heterogeneous actions. We therefore report IRR and coherence between objective and global scores. BARS domain scores were computed at the participant level by averaging the 2 raters' ratings. Noninferiority for each domain was evaluated using the prespecified 20% margin relative to the traditional group mean, based on 95% CIs. We did not perform superiority hypothesis testing (no *P* values or standardized effect sizes). Usability and ergonomics ratings were analyzed using descriptive statistics for the AR group only. All analyses were performed in RStudio (RStudio Team). No ancillary analyses were conducted.

Results

Participant Flow and Baseline Characteristics

We invited 50 residents to participate. Of these, 42 met the inclusion criteria, agreed to participate, and completed all stages and questionnaires of the study. We randomized participants into 2 groups: 20 (48%) in the traditional group and 22 (52%) in the remote AR group. The flow diagram of the participants is in Figure 2. Participants in both groups were at comparable stages of postgraduate training, as reflected by similar distributions of specialty and recent ACLS exposure (Table 1).

XSL•FO
RenderX

Figure 2. CONSORT (Consolidated Standards of Reporting Trials) flow diagram. AR: augmented reality; HMD: head-mounted display.

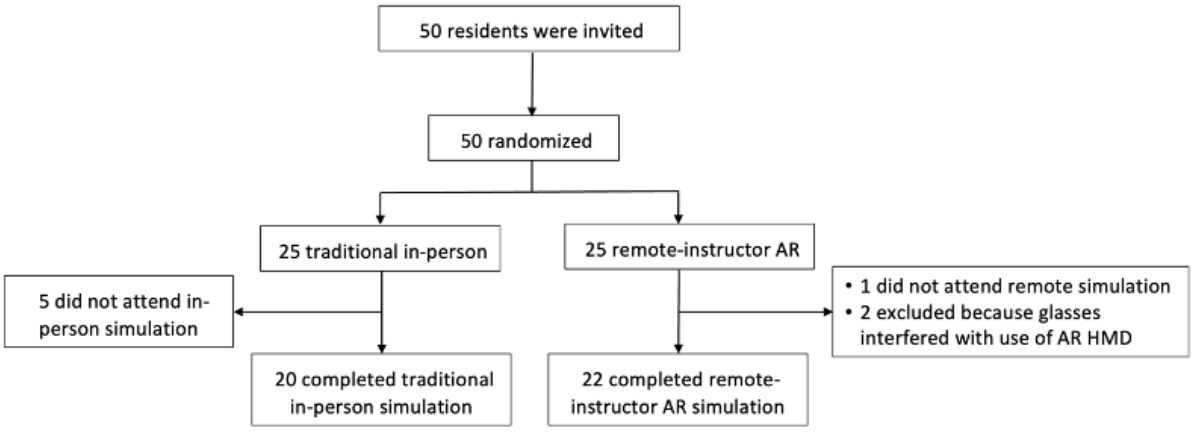


Table . Baseline demographic characteristics of participants in the traditional in-person group and remote augmented reality (AR) group.

Characteristics	Traditional (n=20)	Remote AR (n=22)
Demographics[a]		
Age, mean (SD)	26.8 (1.9)	27.1 (2.3)
Female, n (%)	11 (55)	11 (50)
Specialty, n (%)		
Internal medicine	18 (90)	18 (81.8)
Emergency medicine	2 (10)	4 (18.2)
Time since last ACLS[b] course, n (%)		
Never taken	0 (0)	1 (4.5)
<1 year	5 (25)	6 (27.3)
<2 years	8 (40)	7 (31.8)
<3 years	4 (20)	4 (18.2)
>3 years	3 (15)	4 (18.2)
How much do you work in emergency settings (1="little/not at all" to 5="a lot/routinely"), mean (SD)	3.8 (1.5)	3.9 (1.2)
Have you ever used augmented reality in medical education? (yes=1 and no=0), n (%)	0 (0)	0 (0)

[a]For continuous variables, values are presented as mean (SD); for categorical variables, values are presented as n (%).

[b]ACLS: advanced cardiac life support.

Primary Outcome: Team Leader Performance

The remote AR group demonstrated comparable performance to the traditional group in cognitive leadership tasks, in the objective and subjective scores (Table 2). In all cases, the lower bound of the 95% CI for the AR group exceeded the predefined noninferiority margin, confirming that this group performed noninferiorly to the traditional in-person simulation. IRR was good, with an ICC of 0.84 for the objective score and 0.76 for the subjective score. Additionally, the correlation between the 2 types of scores was 0.855 (95% CI 0.745 - 0.920), indicating good internal coherence between checklist-based performance and global assessments.

XSL•FO
RenderX

Table . Scores from the adapted checklist from Evaluation of Advanced Life Support Performance, compared across simulation modalities with noninferiority thresholds.

Outcome	Traditional, mean (SD)	Remote AR[a], mean (SD)	Noninferiority margin[b]	95% CI (AR)
Objective score[c]	42.6 (5.8)	41.6 (6.2)	34.1	38.9-44.4
Subjective score[d]	7.9 (1.0)	8.1 (1.2)	6.3	7.6-8.6

[a]AR: augmented reality.

[b]Noninferiority cutoffs were defined as 20% below the in-person group mean for each outcome.

[c]Objective checklist score range = −52 to +52 (26 items scored −2 to +2). Higher scores indicate better performance. Subjective/global score range = 0-10. Higher scores indicate better overall performance. Noninferiority cutoffs were defined as 20% below the in-person group mean for each outcome.

[d]Subjective/global score range = 0-10. Higher scores indicate better overall performance.

The remote AR group demonstrated performance comparable to the in-person group using BARS (Table 3). For each BARS domain, the lower bound of the 95% CI for the AR group exceeded the predefined noninferiority margin, indicating noninferior performance relative to traditional in-person simulation.

Table . Leader behavior performance assessed with the Behaviorally Anchored Rating Scale (BARS), compared across simulation modalities with noninferiority thresholds[a].

Outcome	Traditional, mean (SD)	Remote AR[b], mean (SD)	Noninferiority margin	95% CI (AR)
Vigilance (BARS)	8.2 (0.7)	8.3 (0.7)	6.5	8.0-8.6
Decision-making (BARS)	8.2 (0.8)	8.3 (0.9)	6.5	7.9-8.7
Communication (BARS)	8.2 (0.8)	8.1 (0.9)	6.6	7.8-8.5
Teamwork (BARS)	8.3 (0.7)	8.4 (0.6)	6.6	8.1-8.7

[a]BARS domain scores range from 1 to 9; higher scores indicate better performance. Noninferiority cutoffs were defined as 20% below the in-person group mean for each outcome.

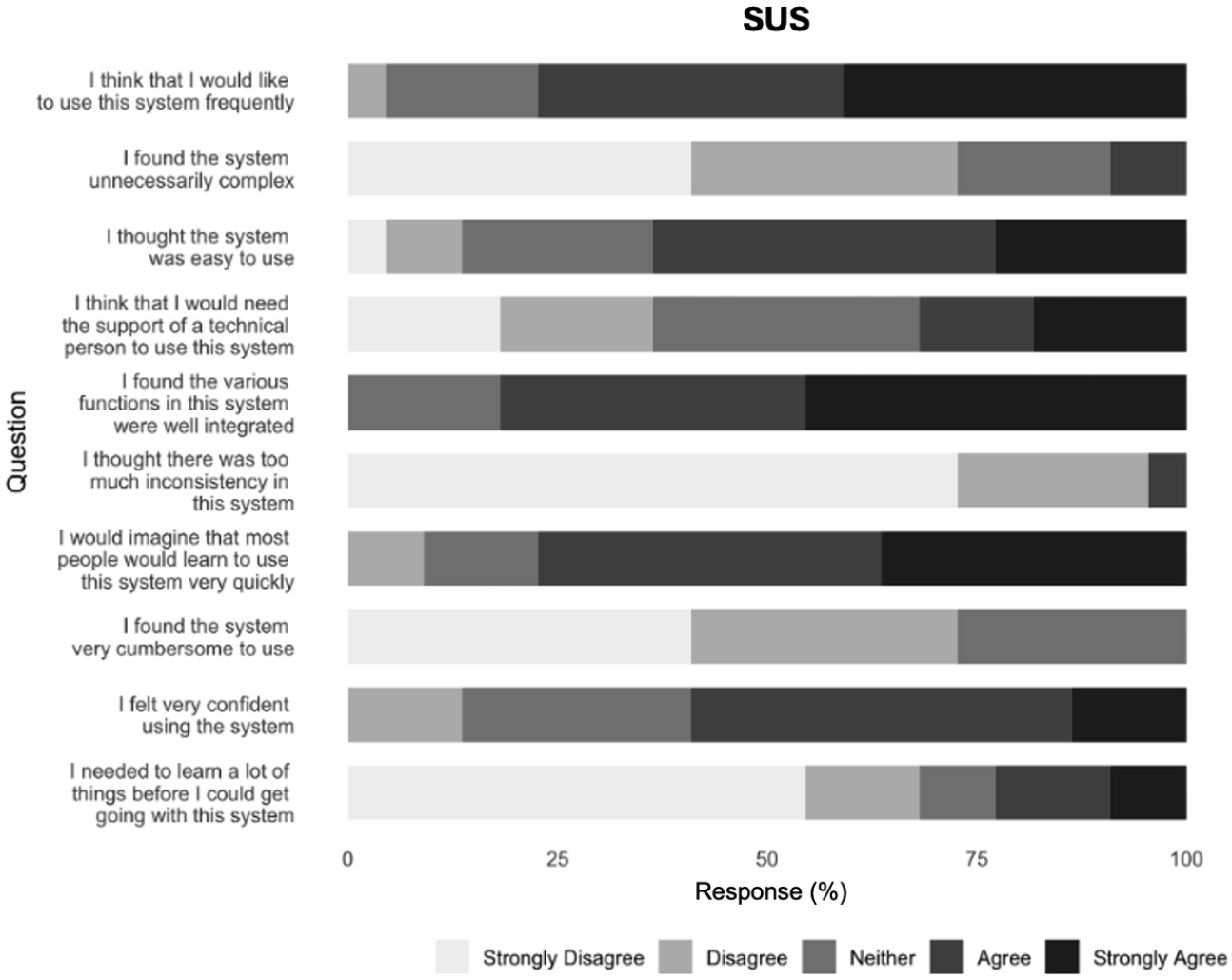
[b]AR: augmented reality.

Secondary Outcome: Usability

We computed the average overall SUS score (0 - 100), which was 73.8 (SD 13.4) among AR participants (n=22), consistent with good or acceptable usability. Most participants agreed or strongly agreed that the AR system was easy to use, well-integrated, and worth frequent use, while they disagreed with statements about complexity, inconsistency, and cumbersomeness (Figure 3). Only the item regarding the need for technical support drew mixed opinions. Overall, the distribution indicates high perceived usability of the platform.

XSL•FO
RenderX

Figure 3. System usability scale (SUS) responses for the augmented reality (AR) simulation system (n=22).



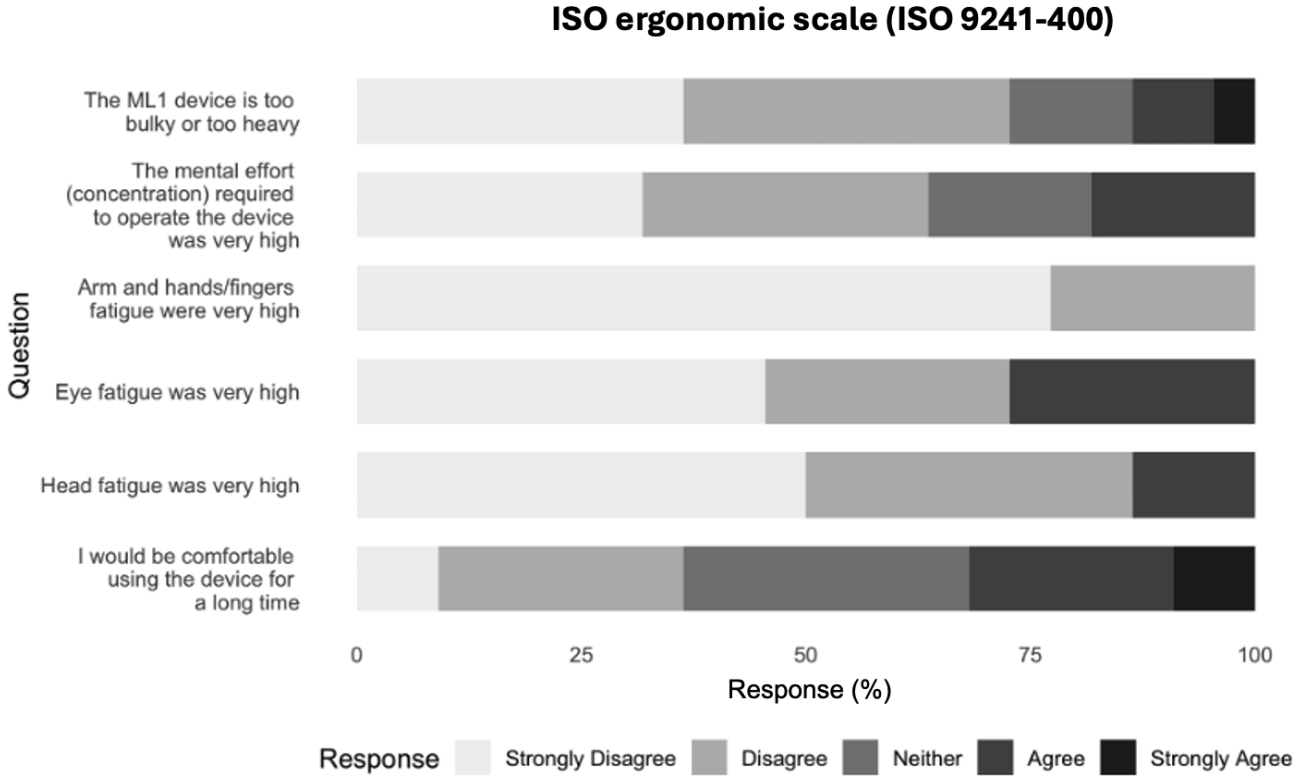
Secondary Outcome: Ergonomics

Regarding ergonomics, a majority rejected statements about excessive bulk, fatigue, or high mental effort, and most agreed the headset would be comfortable for extended use (Figure 4). The responses support minimal physical or cognitive burden during simulation.

No AR-related adverse effects (eg, nausea, dizziness, or eye strain) were observed or reported during or immediately after the simulation sessions, and no participant discontinued the session due to headset discomfort. Minor connectivity lag was noted at the beginning of some remote AR sessions, but no session required termination or rescheduling, and all simulations were completed as planned.

XSL•FO
RenderX

Figure 4. International Organization for Standardization (ISO) 9241 - 400 ergonomic responses for the augmented reality (AR) headset (n=22). ML1: Magic Leap One.



Discussion

Principal Findings

These findings suggest that remote AR simulation may serve as a feasible alternative to traditional in-person simulation for assessing team leader performance during cardiac arrest scenarios. Beyond demonstrating noninferior outcomes, this study highlights the practicality of AR for remote assessment, characterized by high usability, minimal technical barriers, and reliable real-time instruction. Our findings align with evidence from systematic reviews and meta-analyses demonstrating the feasibility of AR and VR simulations in medical education [9,24,25]. While much of this literature focuses on training applications, our study extends these findings by showing that AR can also serve as a standardized assessment modality, producing performance outcomes comparable to conventional in-person simulation.

Evidence suggests that AR can preserve key elements of team communication, including unbroken eye contact, while enabling performance outcomes comparable to those seen in conventional manikin-based simulations [12,26,27]. In addition, AR can deliver real-time clinical prompts and integrate simulated patient data to support decision-making during assessment scenarios [28]. While current evidence does not consistently demonstrate superiority over traditional methods of basic and advanced life support training, our findings contribute to the growing recognition of immersive technologies as promising and scalable alternatives for resuscitation education and assessment across diverse contexts [24].

Video conferencing has also enabled remotely instructed simulations. For example, remote neonatal training showed noninferior technical skills at 2 months among inexperienced medical students, supporting its use in resource-limited settings [29]. A key strength of our study was its cross-continental, parallel group, noninferior randomized design testing instructor-led AR simulation for team leader performance in a simulated cardiac arrest scenario. By delivering cross-continental remote facilitation from Stanford while conducting all learner assessments at UNICAMP, this study serves as a proof of concept for international remote AR assessment. Broader generalizability across settings and learner populations will require replication in additional sites. The required setup (4 headsets, a compression torso, and broadband internet) directly addressed logistical barriers noted in implementation reviews [5]. This portability makes AR simulation especially relevant for rural hospitals, humanitarian settings, and low- and middle-income countries with limited access to simulation centers [9,24].

While this study supports the feasibility of using remote AR simulation for assessing team leader performance, it is important to critically consider whether this level of technological immersion is necessary for the intended evaluation goals. Desktop simulations, video-enhanced medical scenarios, and computer-based training modules are well-established, low-cost alternatives that can effectively assess decision-making and communication in certain contexts, particularly for single-provider scenarios [30-33]. These approaches are scalable, accessible in low-resource environments, and do not require specialized hardware or technical support. In settings where only cognitive rehearsal or basic assessment is required, they may represent more pragmatic solutions. Our study introduces an additional modality of simulation-based assessment, focusing specifically on team leader performance.

XSL•FO
RenderX

Such an approach could be particularly relevant for certification processes in which demonstration of leadership competence is a required component.

The rationale for using AR in this study was to replicate not only the cognitive demands but also the spatial awareness and real-time team coordination involved in resuscitation leadership. AR can integrate contextual cues, spatial orientation, and immersive team interactions that are more representative of the clinical environment than screen-based tools. Unlike video-enhanced medical scenarios, which are typically linear and observer-based, AR allows for active, embodied participation in real-time decision-making within a simulated physical space, thereby enhancing the ecological validity of assessment. This immersive approach supports evaluation of leadership behaviors such as spatial coordination, situational awareness, and communication under time pressure. Moreover, because only the team leader, defibrillation operator, and remote instructor require HMDs, the same hardware can serve larger cohorts through staggered sessions or observer roles, lowering per learner costs compared to setups using multiple high-fidelity mannequins. After initial capital costs, software and headset upkeep are more affordable than ongoing maintenance of complex mannequins [34]. As broadband access expands, AR-enabled simulation provides a scalable option for delivering standardized clinical assessments in underserved settings [9].

Participants generally rated ergonomics and usability as adequate. This was encouraging, as prior studies have reported nausea, dizziness, and visual discomfort after short VR or AR sessions [26,27,35]. Unlike VR, AR reduces motion sickness by allowing natural visual grounding [35]. In our study, scenarios lasted around 12 minutes, likely minimizing adverse effects. Continued improvements in headset design may also have contributed.

Although AR is less established than in-person or VR simulation, it offers potential educational benefits by integrating digital content with the physical environment, enabling learners to interact with both holographic and real clinical tools. This combination may enhance immersion and support knowledge transfer, particularly in scenarios requiring spatial awareness and team coordination. As AR technologies become more accessible, they could serve as a complementary option within a spectrum of simulation-based assessment modalities, particularly where portability and cost efficiency are priorities. Continued research is needed to define its role, establish validity evidence, and determine how AR-based assessments relate to long-term clinical performance and patient outcomes.

Limitations

This study had several limitations. First, recruitment at a single institution limits external validity, and replication at other sites is needed. Second, enrolling only residents restricts generalizability to other learner groups, especially medical students. Third, usability and ergonomic data were self-reported, which may introduce response bias despite the use of validated scales. Fourth, the study evaluated performance at a single time point and therefore does not provide information on longitudinal consistency or reproducibility of assessment outcomes. Fifth, 2 participants were excluded after randomization due to incompatibility between corrective glasses and the AR headset. This exclusion applied only to the AR group and reflects a technical device limitation rather than participant characteristics. Sixth, cardiopulmonary resuscitation technical quality metrics (eg, compression depth, rate, and recoil) were not measured; therefore, our findings speak to team leader cognition and nontechnical behaviors (decision-making, communication, and coordination) rather than hands-on resuscitation technical performance. Seventh, the study used a single VF scenario, which may not capture the full range of skills needed for real-world cardiac arrest. Finally, because assessments were conducted live and the simulation modality was visually apparent, raters were not blinded to group allocation, which may have introduced observer bias in favor of or against the AR modality.

While this study focused on team leader performance, future research should incorporate additional outcomes such as chest compression quality and other key resuscitation tasks. Expanding to diverse scenarios would strengthen the ecological validity of AR-based assessment. While acknowledging that role rotation, common in ACLS training, was not implemented, further studies could examine whether AR platforms are suitable for assessing competence across multiple roles. Future studies should explore how AR platforms can provide real-time, objective metrics to support standardized, instructor-independent evaluations. These innovations may enable scalable, competency-based credentialing in settings with limited faculty resources. Further research is also needed to determine whether these immersive features confer measurable advantages over simpler screen-based approaches when the primary goal is leadership-focused assessment.

Conclusions

This international noninferiority randomized trial found that remote AR simulation produced team leader performance comparable to traditional in-person simulation in a VF arrest scenario. Usability scores were high and ergonomic burden was low, indicating that learners operated the headset comfortably and with minimal technical support. These findings suggest that remote instructor-led AR simulation may be a feasible option for delivering standardized, team-based assessment of cardiac arrest leadership in contexts where geographic, logistical, or financial constraints limit access to conventional simulation.

Acknowledgments

The authors would like to acknowledge the Stanford Chariot Program for the development of immersive technologies, specifically in augmented reality medical simulation.

XSL•FO
RenderX

Funding

The study was funded by the Stanford Chariot Program.

Data Availability

Data are available from the corresponding author upon reasonable request.

Authors' Contributions

RG-O and MR: conceived and designed the work, led the acquisition of the data and its analysis, and drafted the work for publication; drafted the initial manuscript and subsequent revisions of the manuscript; final approval and agreement to be accountable for all aspects of the work. MQ, FZ, AF, LK, and AP: substantial contributions on the design of the work, data acquisition, and the writing of the manuscript. BC: substantial contributions to the data analysis or interpretation and the writing of the manuscript. DC-F and TJC: conceived and designed the work, analyzed and interpreted the data, and critically revised the manuscript. All authors have read and approved the final manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Standardized simulation script for cardiac arrest scenario.
[DOCX File, 1544 KB - mededu_v12i1e84367_app1.docx]

Checklist 1
CONSORT checklist.
[PDF File, 213 KB - mededu_v12i1e84367_app2.pdf]

References

1. Fernandez Castelao E, Russo SG, Riethmüller M, Boos M. Effects of team coordination during cardiopulmonary resuscitation: a systematic review of the literature. J Crit Care 2013 Aug;28(4):504-521. [doi: 10.1016/j.jcrc.2013.01.005] [Medline: 23602030]
2. Cooper S, Wakelam A. Leadership of resuscitation teams: "Lighthouse Leadership". Resuscitation 1999 Sep;42(1):27-45. [doi: 10.1016/s0300-9572(99)00080-5] [Medline: 10524729]
3. Andersen PO, Jensen MK, Lippert A, Østergaard D. Identifying non-technical skills and barriers for improvement of teamwork in cardiac arrest teams. Resuscitation 2010 Jun;81(6):695-702. [doi: 10.1016/j.resuscitation.2010.01.024] [Medline: 20304547]
4. Peltonen LM, Peltonen V, Salanterä S, Tommila M. Development of an instrument for the evaluation of advanced life support performance. Acta Anaesthesiol Scand 2017 Oct;61(9):1215-1231. [doi: 10.1111/aas.12960] [Medline: 28832902]
5. Cheng A, Lockey A, Bhanji F, Lin Y, Hunt EA, Lang E. The use of high-fidelity manikins for advanced life support training—a systematic review and meta-analysis. Resuscitation 2015 Aug;93:142-149. [doi: 10.1016/j.resuscitation.2015.04.004] [Medline: 25888241]
6. Cusimano MD, Cohen R, Tucker W, Murnaghan J, Kodama R, Reznick R. A comparative analysis of the costs of administration of an OSCE (Objective Structured Clinical Examination). Acad Med 1994 Jul;69(7):571-576. [doi: 10.1097/00001888-199407000-00014] [Medline: 8018269]
7. Walsh K, Jaye P. Cost and value in medical education. Educ Prim Care 2013 Sep;24(6):391-393. [doi: 10.1080/14739879.2013.11494206] [Medline: 24196592]
8. Brown C, Ross S, Cleland J, Walsh K. Money makes the (medical assessment) world go round: the cost of components of a summative final year Objective Structured Clinical Examination (OSCE). Med Teach 2015 Jul;37(7):653-659. [doi: 10.3109/0142159X.2015.1033389] [Medline: 25923233]
9. Barteit S, Lanfermann L, Bärnighausen T, Neuhann F, Beiersmann C. Augmented, mixed, and virtual reality-based head-mounted devices for medical education: systematic review. JMIR Serious Games 2021 Jul 8;9(3):e29080. [doi: 10.2196/29080] [Medline: 34255668]
10. Kassutto SM, Baston C, Clancy C. Virtual, augmented, and alternate reality in medical education: socially distanced but fully immersed. ATS Sch 2021 Dec;2(4):651-664. [doi: 10.34197/ats-scholar.2021-0002RE] [Medline: 35079743]
11. Asoodar M, Janesarvatan F, Yu H, de Jong N. Theoretical foundations and implications of augmented reality, virtual reality, and mixed reality for immersive learning in health professions education. Adv Simul (Lond) 2024 Sep 9;9(1):36. [doi: 10.1186/s41077-024-00311-5] [Medline: 39252139]

XSL•FO
RenderX

12. Hess O, Qian J, Bruce J, et al. Communication skills training using remote augmented reality medical simulation: a feasibility and acceptability qualitative study. Med Sci Educ 2022 Oct;32(5):1005-1014. [doi: 10.1007/s40670-022-01598-7] [Medline: 35966166]
13. Tsai A, Bodmer N, Hong T, et al. Participant perceptions of augmented reality simulation for cardiac anesthesiology training: a prospective, mixed-methods study. J Educ Perioper Med 2023;25(3):E712. [doi: 10.46374/volxxv_issue3_Tsai] [Medline: 37720369]
14. Friedman N, Zuniga-Hernandez M, Titzler J, et al. Prehospital pediatric emergency training using augmented reality simulation: a prospective, mixed methods study. Prehosp Emerg Care 2024;28(2):271-281. [doi: 10.1080/10903127.2023.2224876] [Medline: 37318845]
15. Piaggio G, Elbourne DR, Altman DG, Pocock SJ, Evans SJW, CONSORT Group. Reporting of noninferiority and equivalence randomized trials: an extension of the CONSORT statement. JAMA 2006 Mar 8;295(10):1152-1160. [doi: 10.1001/jama.295.10.1152] [Medline: 16522836]
16. Panchal AR, Bartos JA, Cabañas JG, et al. Part 3: adult basic and advanced life support: 2020 American Heart Association guidelines for cardiopulmonary resuscitation and emergency cardiovascular care. Circulation 2020 Oct 20;142(16_suppl_2):S366-S468. [doi: 10.1161/CIR.0000000000000916] [Medline: 33081529]
17. Bochatay N, Ju M, O'Brien BC, van Schaik SM. A scoping review of interprofessional simulation-based team training programs. Simul Healthc 2025 Feb 1;20(1):33-41. [doi: 10.1097/SIH.0000000000000792] [Medline: 38526045]
18. Peltonen V, Peltonen LM, Rantanen M, et al. Randomized controlled trial comparing pit crew resuscitation model against standard advanced life support training. J Am Coll Emerg Physicians Open 2022 Jun;3(3):e12721. [doi: 10.1002/emp2.12721] [Medline: 35601649]
19. Caruso TJ, Hess O, Roy K, et al. Integrated eye tracking on Magic Leap One during augmented reality medical simulation: a technical report. BMJ Simul Technol Enhanc Learn 2021;7(5):431-434. [doi: 10.1136/bmjstel-2020-000782] [Medline: 35515734]
20. Qian J, Rama A, Wang E, et al. Assessing pediatric life support skills using augmented reality medical simulation with eye tracking: a pilot study. J Educ Perioper Med 2022;24(3):E691. [doi: 10.46374/volxxiv_issue3_qian] [Medline: 36274998]
21. Watkins SC, Roberts DA, Boulet JR, McEvoy MD, Weinger MB. Evaluation of a simpler tool to assess nontechnical skills during simulated critical events. Simul Healthc 2017 Apr;12(2):69-75. [doi: 10.1097/SIH.0000000000000199] [Medline: 28704284]
22. Brooke J. SUS: a "quick and dirty" usability scale. In: Usability Evaluation In Industry, 1st edition: CRC Press; 1996:6.
23. Koo TK, Li MY. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. J Chiropr Med 2016 Jun;15(2):155-163. [doi: 10.1016/j.jcm.2016.02.012] [Medline: 27330520]
24. Cheng A, Fijacko N, Lockey A, et al. Use of augmented and virtual reality in resuscitation training: a systematic review. Resusc Plus 2024 Jun;18:100643. [doi: 10.1016/j.resplu.2024.100643] [Medline: 38681058]
25. Sun R, Wang Y, Wu Q, et al. Effectiveness of virtual and augmented reality for cardiopulmonary resuscitation training: a systematic review and meta-analysis. BMC Med Educ 2024 Jul 5;24(1):730. [doi: 10.1186/s12909-024-05720-8] [Medline: 38970090]
26. Sun WN, Hsieh MC, Wang WF. Nurses' knowledge and skills after use of an augmented reality app for advanced cardiac life support training: randomized controlled trial. J Med Internet Res 2024 Dec 5;26:e57327. [doi: 10.2196/57327] [Medline: 39636667]
27. Toohey S, Wray A, Hunter J, et al. Comparing the psychological effects of manikin-based and augmented reality–based simulation training: within-subjects crossover study. JMIR Med Educ 2022 Aug 1;8(3):e36447. [doi: 10.2196/36447] [Medline: 35916706]
28. Zhu E, Hadadgar A, Masiello I, Zary N. Augmented reality in healthcare education: an integrative review. PeerJ 2014;2:e469. [doi: 10.7717/peerj.469] [Medline: 25071992]
29. Mediratta RP, Clary MK, Liang JW, et al. Remote versus in-person pre-service neonatal resuscitation training: a noninferiority randomized controlled trial in Ethiopia. Resuscitation 2025 Apr;209:110556. [doi: 10.1016/j.resuscitation.2025.110556] [Medline: 39986344]
30. Shahrezaei A, Sohani M, Taherkhani S, Zarghami SY. The impact of surgical simulation and training technologies on general surgery education. BMC Med Educ 2024 Nov 13;24(1):1297. [doi: 10.1186/s12909-024-06299-w] [Medline: 39538209]
31. Liaw SY, Chua WL, et al. Desktop virtual reality versus face-to-face simulation for team-training on stress levels and performance in clinical deterioration: a randomised controlled trial. J Gen Intern Med 2023 Jan;38(1):67-73. [doi: 10.1007/s11606-022-07557-7] [Medline: 35501626]
32. Dubovi I. Online computer-based clinical simulations: the role of visualizations. Clin Simul Nurs 2019 Aug;33:35-41. [doi: 10.1016/j.ecns.2019.04.009]
33. Kononowicz AA, Woodham LA, Edelbring S, et al. Virtual patient simulations in health professions education: systematic review and meta-analysis by the Digital Health Education Collaboration. J Med Internet Res 2019 Jul 2;21(7):e14676. [doi: 10.2196/14676] [Medline: 31267981]

XSL•FO
RenderX

34. Hippe DS, Umoren RA, McGee A, Bucher SL, Bresnahan BW. A targeted systematic review of cost analyses for implementation of simulation-based education in healthcare. SAGE Open Med 2020;8:2050312120913451. [doi: 10.1177/2050312120913451] [Medline: 32231781]
35. Rochlen LR, Levine R, Tait AR. First-person point-of-view-augmented reality for central line insertion training: a usability and feasibility study. Simul Healthc 2017 Feb;12(1):57-62. [doi: 10.1097/SIH.0000000000000185] [Medline: 27930431]

Abbreviations

ACLS: advanced cardiac life support
AR: augmented reality
BARS: Behavioral Assessment Rating Scale
CONSORT: Consolidated Standards of Reporting Trials
HMD: head-mounted display
ICC: intraclass correlation coefficient
IRR: interrater reliability
ISO: International Organization for Standardization
SUS: system usability scale
VF: ventricular fibrillation
VR: virtual reality

Edited by S Guinez-Molinos; submitted 19.Sep.2025; peer-reviewed by GJ Noordergraaf, T Tangpaisarn, Y Yang; revised version received 27.Dec.2025; accepted 15.Jan.2026; published 16.Mar.2026.

Please cite as:
Gianotto-Oliveira R, Rojas M, Queiroz M, Zanchetta F, Ferrari A, Kojima L, Paula A, Campos B, Cecilio-Fernandes D, Caruso TJ Remote Augmented Reality Versus Traditional Simulation for Team Leader Assessment in a Cardiac Arrest Scenario: Noninferiority Randomized Controlled Trial
JMIR Med Educ 2026;12:e84367
URL: https://mededu.jmir.org/2026/1/e84367
doi:10.2196/84367

© Renan Gianotto-Oliveira, Marcos Rojas, Maria Queiroz, Flávia Zanchetta, Anabel Ferrari, Lucas Kojima, Alexandre Paula, Bruno Campos, Dario Cecilio-Fernandes, Thomas J Caruso. Originally published in JMIR Medical Education (https://mededu.jmir.org), 16.Mar.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Education, is properly cited. The complete bibliographic information, a link to the original publication on https://mededu.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Investigating the Impact of a Virtual Reality Experience on Medical Student Empathy: Mixed Methods Study

Allen G Mundok[1*], BS; Vivian N Ho[1*], BS; Lauren A Fowler[2*], PhD; Ann Blair Kennedy[1*], DrPH; Shannon Stark-Taylor[1*], PhD

[1]Department of Biomedical Sciences, University of South Carolina School of Medicine Greenville, 607 Grove Rd, Greenville, SC, United States

[2]Department of Translational Neuroscience, Wake Forest University School of Medicine Charlotte, Charlotte, NC, United States

[*]all authors contributed equally

Corresponding Author:
Ann Blair Kennedy, DrPH
Department of Biomedical Sciences, University of South Carolina School of Medicine Greenville, 607 Grove Rd, Greenville, SC, United States

Abstract

Background: Physician empathy is important not only for improving patient satisfaction and health outcomes but also for increasing physician job satisfaction and protecting against burnout. However, amid concerns over declining empathy levels in medical education, there is a need for innovative teaching approaches that address the empathy gap, a critical element in patient-centered care.

Objective: This study aimed to use a mixed-methods analysis to explore the effectiveness of a virtual reality (VR) intervention versus traditional lecture methods in enhancing empathy among medical students.

Methods: Overall, 50 first- and second-year medical students were randomized to either a VR intervention, which simulated patient experiences, or a control group receiving traditional empathy lectures. Both groups watch 2 videos with reflections gathered after each video to capture students' experiential learning. Empathy was measured using the Jefferson Scale of Empathy-Student Version before and after the intervention.

Results: Quantitative analysis revealed significant increases in empathy scores post intervention for both groups (lecture group: mean increase 4.71, SD 11.01; VR group: mean increase 5.6, SD 10.02; $P<.001$), indicating that both interventions enhanced empathy. The VR group exhibited a significant difference in qualitative empathy coding after the second video ($U=165.5$; $P<.001$) compared to the lecture group. Qualitative feedback from the VR group emphasized a more profound emotional and cognitive engagement with the patient perspective than the lecture group.

Conclusions: This study supports the integration of VR into medical education as a complementary approach to traditional teaching methods for empathy training. VR immersion provides a valuable platform for students to develop a deeper, more nuanced understanding of empathy. These findings advocate for further exploration into VR's long-term impact on empathy in clinical practice.

(JMIR Med Educ 2026;12:e76504) doi:10.2196/76504

KEYWORDS

medical education; virtual reality; medical students; mixed-methods; empathy

Introduction

Empathy can be interpreted as the emotional engagement of the observer with the circumstances of another [1]. In medical settings, higher physician empathy has been associated with improved patient outcomes, including greater patient satisfaction and increased adherence to medical recommendations and treatments [2]. For example, individuals with diabetes whose physicians score higher on measures of empathy demonstrate better glucose and cholesterol control than those whose physicians score lower, controlling for other physician and patient factors [3]. For physicians, empathy has also been linked to higher supervisor ratings of clinical competence and fewer malpractice claims [4,5]. Despite the known benefits of empathy, research also shows that empathy tends to decline during medical school, particularly during clinical training, with male students showing larger decreases [5]. These studies highlight the importance of understanding how empathy develops and changes throughout the course of one's medical education and suggest a need for interventions to support its maintenance.

Studies have demonstrated the effectiveness of many different modes of empathy training for increasing empathy ratings in medical students [6,7]. To train individuals in empathy, traditional interventions have often taken the form of

XSL•FO
RenderX

perspective-taking. These interventions involve having participants imagine another's situation and put themselves "in that person's shoes." Perspective-taking has also been shown to promote empathy in a range of nonmedical contexts [8-10].

Compared to traditional perspective-taking, virtual reality (VR) interventions offer a fully immersive experience, giving the user the visual and emotional feeling of being present for the interaction and embodying another [11,12]. One study compared traditional perspective-taking to a VR perspective-taking intervention using a virtual environment to promote increased empathy for homeless persons. Participants who underwent the VR intervention showed more positive, longer-lasting attitudes toward the homeless and signed a petition supporting the homeless at a significantly higher rate than those who engaged in a traditional perspective-taking task [11]. Studies also suggest VR interventions are less cognitively taxing than imagining another's perspective and allow for improved methodology as all participants engage in the same exercise [13-15].

Despite this evidence, few studies have applied VR technology to teaching medical students and physicians. With medical students increasingly using technology resources to enhance their classroom learning [16,17], finding ways to integrate technology into empathy training may be an effective method to engage medical students [18,19]. Given the benefits of immersive perspective-taking in other contexts, we designed an innovative VR experience to facilitate patient perspective-taking among medical students and compared it to traditional teaching methods (a video lecture). Based on prior findings promoting empathy for several different human conditions in lay populations and the theoretical advantages of VR for perspective-taking, we hypothesize that engaging in patient perspective-taking through VR will result in greater increases in empathy among medical students than the traditional teaching method.

Methods

Study Design and Video Development

To compare the efficacy of a VR experience to traditional lecture learning, we used a randomized controlled mixed-methods convergent parallel design. In this study design, both qualitative (structured reflection responses) and quantitative data (Jefferson Scale of Empathy for medical students [JSE-S] pre- and posttest scores) were collected simultaneously in a single phase. Each type of data were analyzed separately, then merged to identify points of convergence [20]. This design allows us to integrate the strengths of both approaches, providing a more comprehensive understanding of how the VR intervention influenced empathy, while also allowing the qualitative data to contextualize and enrich the quantitative findings [20].

The lecture videos were PowerPoint lectures designed and narrated by a biomedical sciences professor and research team member (LAF) with expertise in empathy. Lecture video 1 (V1) described the science behind empathy and its importance in health care. Lecture video 2 (V2) provided strategies for students to enhance empathy in clinical practice.

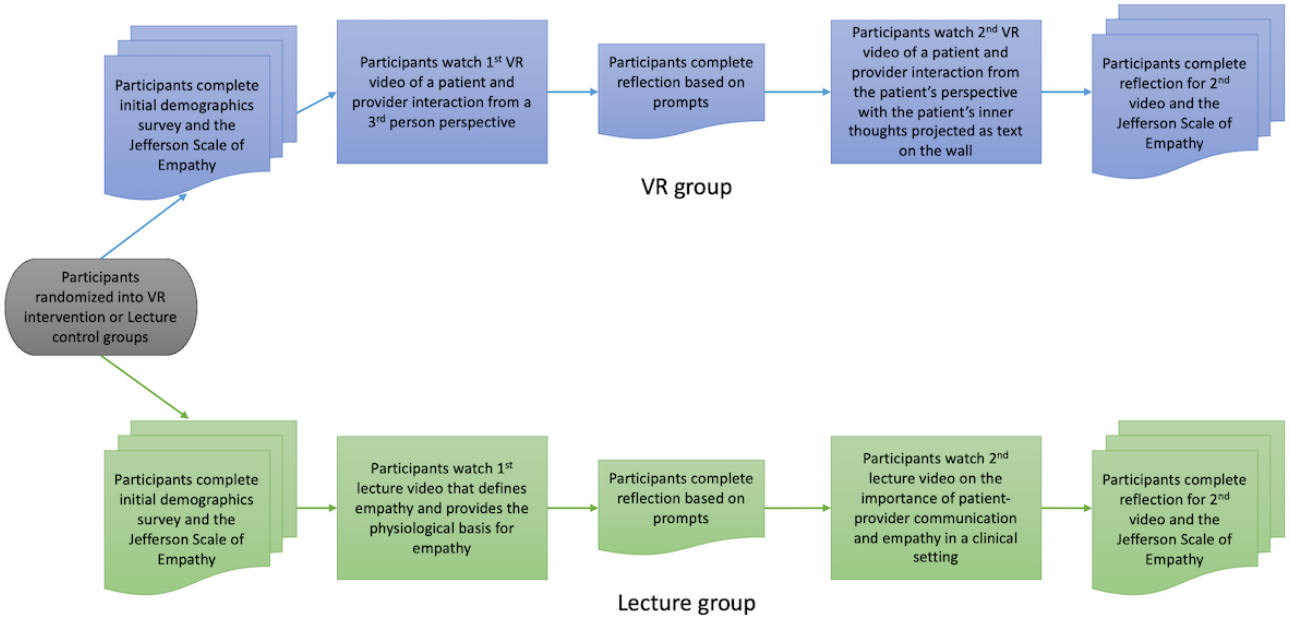
To make the VR videos and reflection questions as authentic as possible, these items were presented to a panel of patients and providers trained in giving feedback to researchers. The panel members made suggestions for script and question improvements. After incorporating the received feedback into the revised script, the VR videos were filmed using a Garmin VIRB Ultra 360 camera (Garmin, Ltd.) with hired actors playing the roles of a patient and physician. VR Video 1 presented a third-person perspective of a "normal" patient-physician interaction during a routine office visit. VR Video 2 showed the same interaction but through the first-person perspective of the patient, and the patient's inner thoughts were projected as text on the screen for students to read. An example of the patient's thoughts is, "How am I going to afford a dietician? Will my insurance pay for that?" Videos for the lecture and VR group were designed to be the same length to control for time.

Data Collection

Using a random number generator, students were randomly assigned to either the lecture group (n=25) or VR group (n=25). In the lecture group, students were given a tablet to watch the 2 PowerPoint lecture videos on empathy. In the VR group, students wore a VR headset to watch the 2 VR videos depicting a patient-physician interaction. Prior to watching their respective videos, students completed an initial demographics survey and the Jefferson Scale of Empathy for medical students (JSE-S). After watching the first video, students answered 3 open-ended reflection questions on their learning experience. Students then watched the second video and again completed the 3 reflection prompts and the JSE-S. Figure 1 depicts this methodology.

XSL•FO
RenderX

Figure 1. Study design overview. Flowchart depicting participant randomization into virtual reality or lecture video group and timeline of assessment (demographics survey, Jefferson Scale of Empathy pre- and postvideos, and reflection prompts). VR: virtual reality.



Throughout their viewing of the videos, students in both groups were also equipped with Empatica E4 wristbands (Empatica, Inc.) to measure physiological markers of empathy such as galvanic skin response and heart rate variability. While physiological markers have been used in prior research to complement self-reported empathy measures, technical artifacts in our dataset rendered the recordings uninterpretable; these data were therefore not included in our final results [21].

Instruments

The validity of the JSE in evaluating empathy in physicians and medical students has been well-established [22,23]. We used the student version (JSE-S), which was developed to measure medical students' attitudes toward empathy in patient-physician interactions. The JSE comprises 20 items, rated on a 7-point Likert-type scale (1=strongly disagree, 7=strongly agree), with higher total scores indicating greater empathy (maximum score=140). An example of a typical item is, "Patients feel better when their health care providers understand their feelings." The JSE-S has demonstrated strong internal consistency, construct validity, and predictive associations with clinical competence and patient-centered outcomes [24,25]. Although the JSE-S primarily assesses attitudes toward empathy rather than direct empathic behavior, it is widely accepted as a valid proxy for empathy in medical education research [26]. Empathy in clinical contexts is understood as a multidimensional construct encompassing cognitive, affective, and attitudinal components [22]. The attitudinal dimension captured by the JSE-S reflects physicians' and students' willingness and value orientation toward understanding patients' experiences, which are key determinants of empathetic engagement in practice. Thus, while the JSE-S measures empathy-related attitudes, its theoretical foundation and empirical validation support its continued use as a reliable instrument for assessing empathy development in medical students.

To complement the JSE-S and provide qualitative insight into students' experiences, participants also responded to structured reflection questions. These served as our qualitative data collection tool, analogous to an interview guide. The lecture group questions asked students to describe the video, to describe how the video made them feel, and to explain how they would apply the information in future practice. The VR group questions asked students to describe the video, to describe how the video made them feel, and to explain how they thought the patient felt during the interaction. While these reflections are self-reported and may be influenced by social desirability or awareness of study aims, they provide rich, contextual data that complement the quantitative JSE-S measures.

Statistical Analysis

We conducted qualitative analysis based on the participants' responses to the reflection prompts using a thematic analysis approach as outlined by Braun and Clarke [27-29]. Written responses were imported into Microsoft Excel for data management. Given the dataset in this study was relatively small (n=50 participants, 100 reflections), Microsoft Excel was sufficient for systematically organizing codes and themes, while analytic rigor was ensured through team-based coding and consensus review. Initial codes were created by reading the text provided by the participants and assigning a code identifier. Coding was an iterative process in which each instance of a code was compared to previous instances to confirm or modify the code and its definition. Once all the data were coded, emergent themes were abstracted by grouping similar codes together. Codes and themes were each reviewed by members of the research team (AM, VH, ABK, and SST). Any discrepancies were resolved through discussion until consensus was reached. This process represents a "modified" application of thematic analysis, and codes were later used to derive quantitative empathy scores as described below. Major themes and illustrative quotes from the participants' responses are reported in the results section to provide further context and to support our findings.

XSL•FO
RenderX

We then determined quantitative scores based on the level of empathy present in the participants' responses. We assigned empathy scores to each participant based on their responses to video 1 (V1) and responses to video 2 (V2), and we also assigned an overall empathy score based on their responses to both videos 1 and 2 (V3). Empathy was scored on a 3-point scale (1=no signs of empathy, 2=some signs of empathy, 3=high signs of empathy). We used our codes to assign these empathy scores. High signs of empathy were indicated by codes such as "patient focused" and "emotive words," while low signs of empathy were indicated by codes like "surface level" and "misperception." We also used a 3-point scale to assign each participant a score for level of change in assigned empathy scores from video 1 to video 2 (1=negative change, 2=no change, and 3=positive change).

The data were quantitatively analyzed by LAF using IBM SPSS Statistics (version 29.0). Descriptive statistics (counts and frequencies for categorical data; means and SD values for continuous data) were performed. We used Mann-Whitney tests to compare the assigned empathy scores at V1, V2, and V3 between the lecture and VR groups. JSE-S scores pre- and posttest were analyzed using paired-sample 2-tailed *t* tests to determine whether JSE-S scores changed over time for all participants. A factorial analysis of variance assessed whether time (pre or posttest) and group (lecture or VR) affected JSE-S scores. To confirm a relationship between the JSE-S and our assigned empathy scores, we ran a Spearman correlation between posttest JSE-S score and V3 for both lecture and VR groups. We considered $P<.05$ as statistically significant.

Ethical Considerations

From January to February of 2020, 50 first- and second-year medical students at the University of South Carolina School of Medicine Greenville participated in our study. The study was reviewed and approved by the University of South Carolina Office of Research Compliance (Institutional Review Board approval number: Pro00089391). Participation was voluntary, and students received a $30 gift card as compensation. All participants provided written informed consent prior to participation. Upon arrival, study personnel explained the study procedures and reviewed the informed consent document with each participant, answering all questions before consent was obtained. To minimize response bias, recruitment materials and the consent process described the study as an investigation of the role of virtual reality in medical education, which was accurate but did not fully disclose that empathy was a primary outcome of interest. Participants were verbally debriefed immediately following completion of their participation and informed of the study's full purpose. The study posed minimal risk to participants. Potential risks included mild skin irritation from the polyurethane E4 wristband and nausea or headache associated with use of the virtual reality headset. Participants were informed they could discontinue use of either device or withdraw from the study at any time without penalty. Identifying information was collected solely for recruitment purposes and was not linked to study data. Recruitment information was stored on a password-protected Microsoft Teams site accessible only to study personnel. All study data were deidentified, assigned a unique participant identification number, and stored in a HIPAA-compliant REDCap database on a secure network accessible only to authorized members of the research team.

Results

Participant Characteristics

Of the 50 preclinical medical students that participated in our study, the overall mean age was 24.14 (SD 2.66) years. The majority of participants were female (37/50, 74%) and non-Hispanic White (31/50, 62%). There were similar numbers of participants in each class year, with 24 of 50 (48%) M1 students and 26 of 50 (52%) M2 students. Eight of 50 (16%) participants identified as first-generation students, and 14 of 50 (28%) participants were from rural backgrounds. The specialty of interest varied among the participants, but the highest number of students (11/50, 22%) chose obstetrics and gynecology. Additional demographic information is depicted in Table 1.

XSL•FO
RenderX

Table . Participant demographic data.

Variables	Lecture, (n=25)	VR[a], (n=25)	Total, (N=50)
Age (years), mean (SD)	23.92 (1.53)	24.36 (3.46)	24.14 (2.66)
Sex, n (%)			
Male	8 (32)	4 (16)	12 (24)
Female	16 (64)	21 (84)	37 (74)
Transgender	1 (4)	0 (0)	1 (2)
Race and ethnicity[b], n (%)			
Asian	2 (8)	7 (28)	9 (18)
Black or African American	4 (16)	5 (20)	9 (18)
Hispanic or Latino	1 (4)	0 (0)	1 (2)
Non-Hispanic White	18 (72)	13 (52)	31 (62)
Another or not listed	2 (7.7)	0 (0)	2 (4)
Decline to answer	0 (0)	1 (4)	1 (2)
Class year, n (%)			
M1	13 (52)	11 (44)	24 (48)
M2	12 (48)	14 (56)	26 (52)
Marital status, n (%)			
Single or never married	23 (92)	21 (84)	44 (88)
Married	2 (8)	4 (16)	6 (12)
First generation[c], n (%)	4 (16)	4 (16)	8 (16)
Rural, n (%)	7 (28)	7 (28)	14 (28)
Specialty, n (%)			
Dermatology	0 (0)	2 (8)	2 (4)
Emergency medicine	2 (8)	1 (4)	3 (6)
Family medicine	1 (4)	2 (8)	3 (6)
General surgery	6 (24)	2 (8)	8 (16)
Internal medicine	2 (8)	1 (4)	3 (6)
Internal medicine-pediatrics	1 (4)	0 (0)	1 (2)
Interventional radiology	1 (4)	0 (0)	1 (2)
Obstetrics and Gynecology	5 (20)	6 (24)	11 (22)
Ophthalmology	1 (4)	0 (0)	1 (2)
Orthopedic surgery	2 (8)	2 (8)	4 (8)
Otolaryngology	1 (4)	1 (4)	2 (4)
Pediatrics	1 (4)	5 (20)	6 (12)
Undecided	2 (8)	3 (12)	5 (10)

[a]VR: virtual reality.

[b]Response options allowed respondents to check more than 1 option.

[c]In our study, defined as a student whose parents or guardians did not complete a four-year degree.

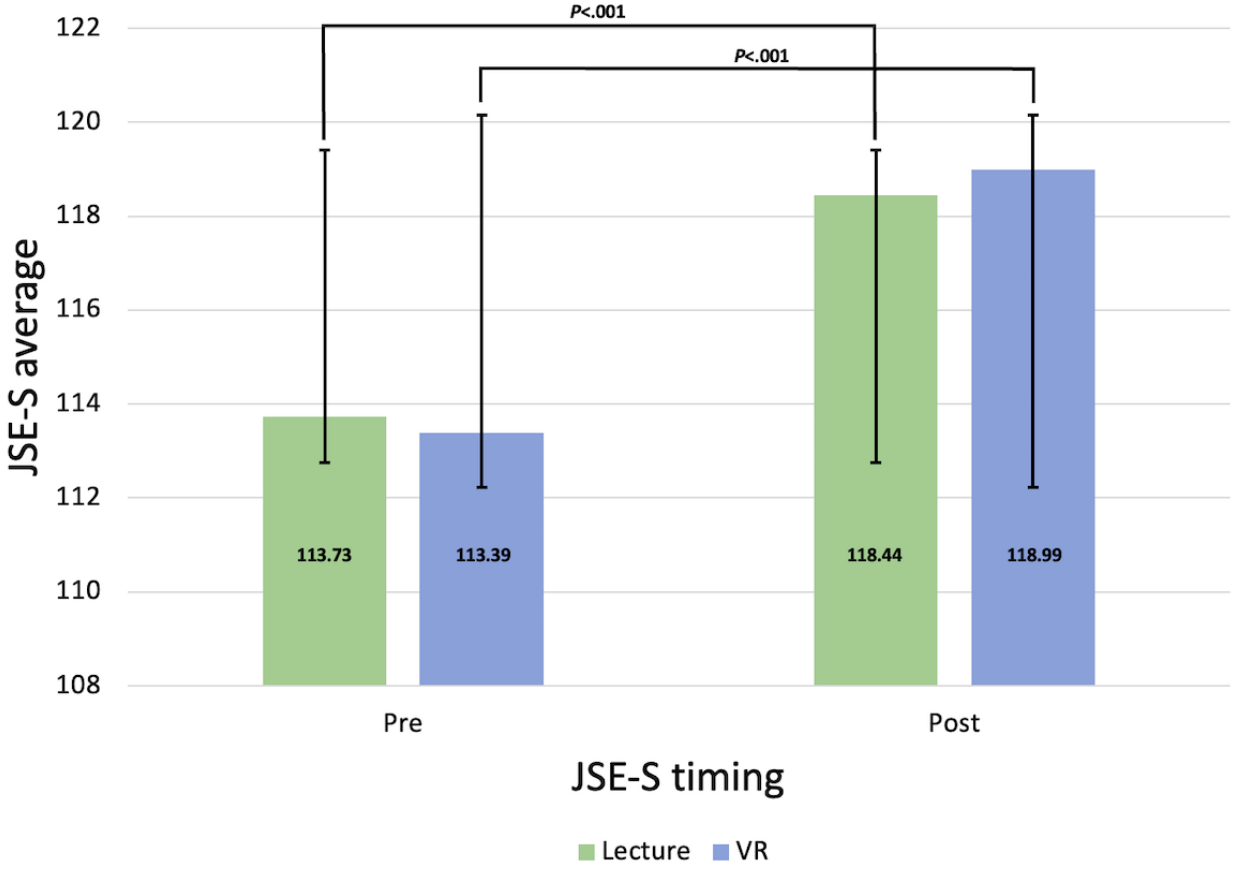
Quantitative Analysis: JSE-S

JSE-S scores were assessed prior to and following each testing session (pre or posttest) for both the lecture and VR groups. Prevideo JSE-S scores were not significantly different between the groups, with the lecture group averaging 113.73 (SD 10.57) and the VR group averaging 113.39 (SD 10.24). Postvideo averages of the JSE scores between the 2 groups were again not significantly different, showing that the lecture group averaged 118.44 (SD 11.01) and the VR group averaged 118.99 (SD 10.02). A paired-samples *t* test was used to determine whether JSE-S scores changed across time for all participants. Results

XSL•FO
RenderX

indicate that JSE scores significantly increased after the empathy training sessions compared to before the sessions (t_{49}=7.28; P<.001). A factorial analysis of variance assessed whether time (pre or post) and group (lecture or VR) affected JSE-S scores, with results indicating no significant effect of either ($F_{1,48}$=0.17; P=.68). These results indicate that while there was no statistically significant difference in JSE-S scores observed between the lecture and VR groups, both groups were effective in significantly increasing empathy scores. Refer to Figure 2 for comparison of JSE-S scores across time for both groups.

Figure 2. Average scores across time (pre or post) by group (lecture or VR). The Jefferson Scale of Empathy-Student Version was administered to participants prior to viewing both videos (pre) and after viewing both videos (post). Jefferson Scale of Empathy-Student Version score was compared across time in lecture versus VR groups using a paired-sample *t* test (***P<.001). SD bars are shown. JSE-S: Jefferson Scale of Empathy-Student Version; VR: virtual reality.

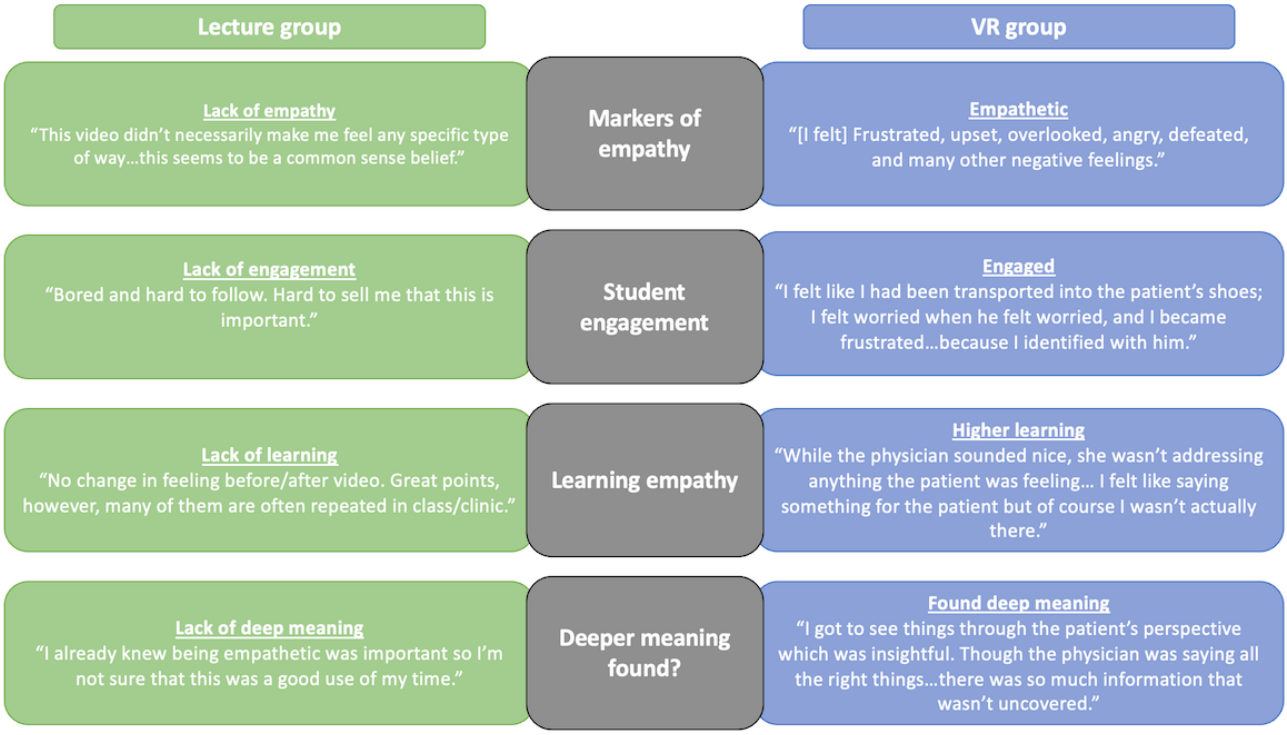


Qualitative Analysis of Reflection Responses

Qualitative analysis of participants' reflection responses in the lecture group versus the VR group revealed 4 major themes relating to empathy and its learning that diametrically opposed each other: (1) markers of empathy, (2) student engagement, (3) learning empathy, and (4) deeper meaning found (Figure 3). The lecture group responses overall demonstrated lack of empathy, lack of engagement, lack of learning, and lack of deeper meaning. While lecture group students seemed to cognitively acknowledge the importance of empathy, their responses were generally sarcastic and demonstrated low levels of empathy. For example, when asked how the lecture video made them feel, 1 student responded simply with, "bored." Another stated, "this video made me feel sleepy…this is all stuff we have heard before." When asked how they would apply the lecture content to future clinical practice, 1 student responded with "be empathetic like I was supposed to be before?" In contrast, the VR group responses showed empathy, engagement, higher learning, and deeper meaning. VR group students experienced a range of empathetic emotions after the VR experience, with 1 student stating that they felt "frustrated, upset, overlooked, angry, defeated, and many other negative feelings" on behalf of the patient. The VR group participants were highly engaged, commenting, "I felt like stepping in and asking the patient for input," and "I felt like I wanted to be Sam's advocate and ask questions…the provider should've asked." Because the VR videos promoted perspective-taking, the students found deeper meaning. One individual stated, "I think the patient…needed to talk about his feelings just as much, if not more than his abdominal pain." The VR group also demonstrated learned empathy in their responses. After watching the first video, 1 participant commented, "I felt that the physician was doing a good job being pleasant." After the second video, the same participant stated, "I felt like I had been transported into the patient's shoes; I felt worried when he felt worried, and I became frustrated when the doctor didn't give him any feasible solutions because I identified with him at that point."

XSL•FO
RenderX

Figure 3. Major themes and representative quotations. Qualitative coding of participant reflection responses revealed 4 major themes. Example quotations representative of each theme are shown to compare lecture versus VR groups. VR: virtual reality.

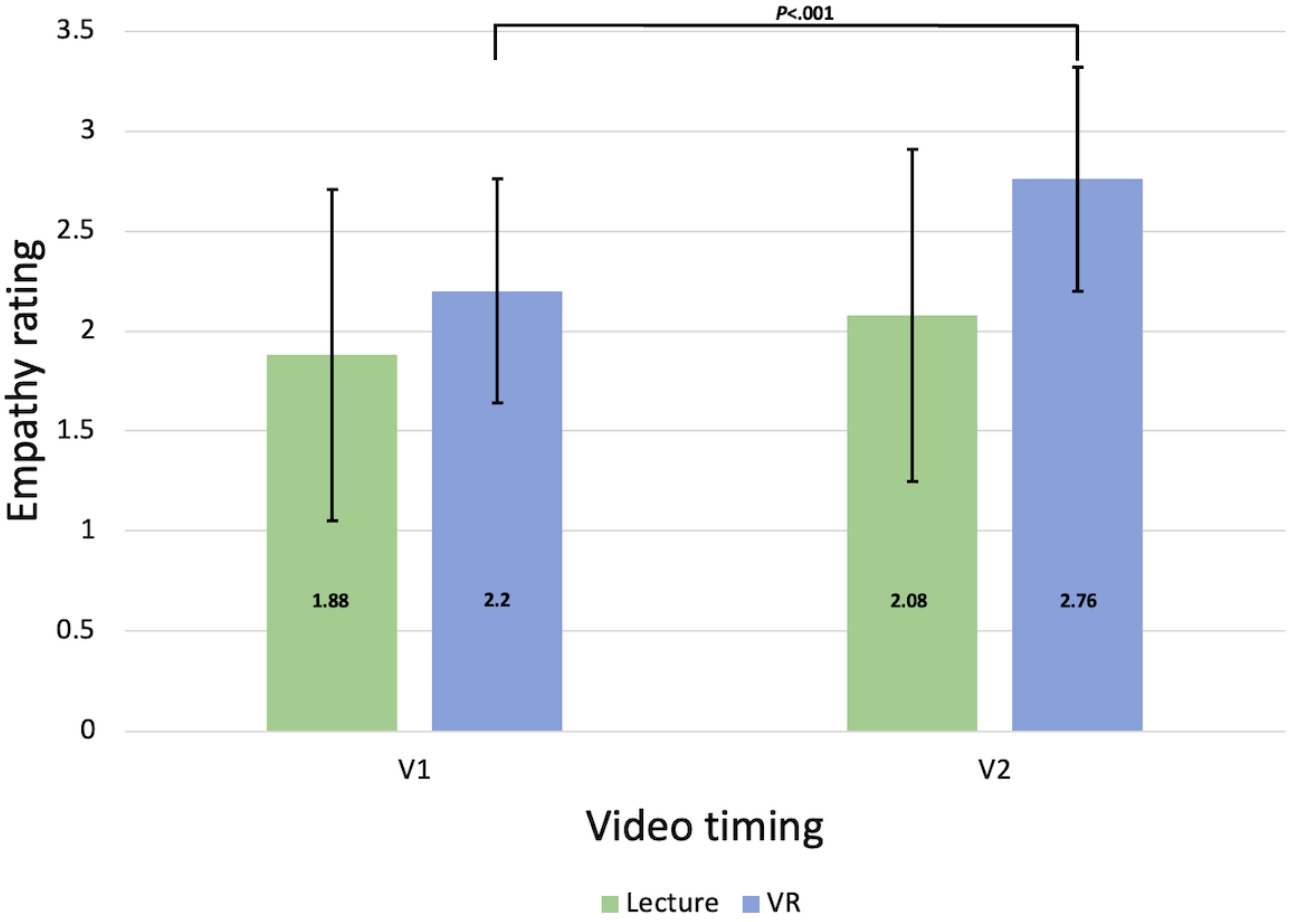


Quantitative Scoring of Qualitative Coding and Comparison to JSE-S Scores

A Mann-Whitney analysis was used to compare the assigned empathy scores at V1 (after video 1) and V2 (after video 2) between the lecture and VR groups. Results indicate that there was no significant difference in qualitative coding for empathy scores between the 2 groups after the first video (V1). However, after the second video (V2), there was a significant difference between qualitative scores, with those in the VR group having significantly higher empathy coding scores than the lecture group (U=165.5; P<.001). A Mann-Whitney analysis also showed that overall qualitative empathy coding for the participants (V3) was significantly higher for those in the VR group as compared to those in the lecture group (U=202.5; P=.03), based on empathy scores of 1.96 and 2.52 in the lecture and VR groups, respectively.. The overall qualitative coding for empathy (V3) was assessed with the post-JSE-S scores to determine if the 2 scores were related to each other (a quantitative rating of empathy in the JSE-S compared to the quantitative scoring of the qualitative coding). Spearman correlation between posttest JSE-S and V3 for both the lecture and VR groups was significantly, positively correlated (ρ_{49}=0.29; P=.03). Refer to Figure 4 for more information.

XSL•FO
RenderX

Figure 4. Qualitative scoring of comments after video 1 (V1) and video 2 (V2). V1 is defined as participants' assigned empathy score after viewing Video 1. V2 is defined as participants' assigned empathy score after viewing Video 2. A Mann-Whitney analysis compared the assigned empathy scores between V1 and V2 in the lecture and VR groups (***P<.001). SD error bars are shown. VR: virtual reality.



Discussion

Principal Findings

This study presents early findings that the VR experience uniquely engaged students and elicited more empathetic responses compared to the traditional lecture. In our mixed methods evaluation, the intervention or VR group demonstrated qualitatively higher markers of empathy, more student engagement, found deeper meaning, and demonstrated more learning of empathy than the lecture group. The intervention or VR group also showed significantly higher overall assigned empathy scores than the lecture group. These findings indicate that immersive VR experiences may enhance empathy development beyond conventional lecture-based methods.

Quantitative and Qualitative Findings

The JSE-S is widely used and has been validated by numerous empirical studies as an instrument to measure empathy [22,30]. Our quantitative analysis conducted using this tool demonstrated a significant increase in JSE-S scores after both the VR and traditional lecture training sessions compared to before the sessions. Seeing an increase in JSE-S scores in both groups posttest is not a surprising result considering students were primed on the topic of physician empathy while completing the questionnaire and since traditional teaching methods have already been proven effective in teaching empathy to some degree [6,31,32]. However, the VR intervention's efficacy in fostering empathy demonstrates noninferiority to traditional lecture-based learning and underscores the importance of experiential learning, as outlined in adult learning theory, and provides a new dimension to empathy theory in medical education. Adult learning theory emphasizes the significance of self-directed, experiential learning processes, where adults learn best when engaged in activities that relate closely to real-life scenarios or challenges they might face [33]. The immersive nature of VR, by providing medical students with a first-person perspective of patient experiences, leverages these principles effectively, offering a profound, emotive learning experience that traditional didactic methods may lack. This approach aligns with the constructs of empathy theory, which posits that empathy involves not just cognitive understanding but also affective sharing and emotional resonance with others [34].

Qualitative analysis helped us better understand the experiential differences between the groups. Themes from the qualitative responses relate clearly to the literature on empathy and adult learning. Individuals in the VR group described experiencing a variety of emotions while watching the VR video, while the lecture group described their experience using more cognitive language. This suggests that the VR intervention met our intent to promote perspective taking. Furthermore, adult learning theorists emphasize the role emotions can play in helping learners connect a construct with prior knowledge and past experiences [35]. Beyond enriching their learning through

XSL•FO
RenderX

perspective taking, the VR intervention appeared to uniquely engage students more than the lecture. The theme of deeper meaning found in the VR group's reflections aligns with the concept of "presence" in VR environments, the true feeling of "being there" in the virtual space. This sense of presence is critical for generating impactful, empathetic responses, as it bridges the gap between knowing about another's experience and feeling with them. Literature on VR and empathy underscores the role of presence in enhancing the emotional impact of VR experiences, suggesting that it is this immersion that enables learners to connect with content on a deeper emotional level, fostering a genuine understanding and appreciation for the patient's perspective [36].

By connecting the quantitative increases in JSE-S scores with the rich qualitative reflections, a clear pattern emerges: immersive VR experiences not only improve self-reported empathy but also foster deeper emotional engagement, perspective-taking, and understanding of patient experiences. This combined evidence highlights the main contribution of the study, showing that VR interventions can meaningfully complement and enhance traditional teaching methods.

Implications for Medical Education

For this topic in particular, embodied and applied experience seems to provide much more nuance to the concept of empathy versus the cognitive acknowledgment that empathy is important in practice. Through the comments, the intervention appeared to help many students more fully understand how empathy can promote better health care. Furthermore, this study's findings resonate with previous research indicating the potential of VR to overcome some of the limitations associated with traditional empathy training methods. For instance, studies have demonstrated VR's effectiveness in reducing cognitive load and allowing for a more standardized and immersive learning experience, characteristics that are particularly advantageous in the context of medical training where the cognitive demand is high and the need for consistent, reproducible training experiences is paramount [11].

Limitations and Future Directions

Our review consisted of a single-institution study at 1 academic center, meaning more work is needed to make more widespread application of our results possible. We note that our participants were predominantly female (37/50, 74%). This reflects the voluntary nature of our participant recruitment, as a greater number of female students elected to enroll compared to their male counterparts. Prior literature has documented sex and gender differences in empathy [37]; thus, we acknowledge the overrepresentation of female participants in our study as a factor that may have influenced the observed results. Furthermore, we did not follow up with participants after the end of the study to assess empathy during clinical rotations and in practice. We feel this would have provided a better understanding of the long-term impact of our empathy training intervention on participants' empathy throughout their medical training. We also recognize some limitations in our use of the JSE-S, including the fact that we may have applied the scale too early following our intervention, thus potentially limiting its reliability. Compared to other interventions to promote empathy in health care learners, this intervention was quite brief; a systematic review of such interventions found that the average length was 10 hours [38].

While traditional teaching methods have been shown to help increase empathy in medical students, the profound differences noted in our qualitative data, however, lead us to question whether full immersion through a VR experience may be a helpful additional tool in allowing students to develop a deeper understanding. Future directions warrant testing a combined approach of traditional lecture alongside a VR intervention to see if a combination provides an even greater increase in empathy than either one alone. We would also like to assess the long-term impact of these interventions on students as they progress through their clinical years.

Although the JSE-S is widely used and validated, it primarily assesses attitudes toward empathy rather than direct empathetic behavior, and participants may respond in a socially desirable manner. No filler items were included to mask the study's purpose, but participants were not explicitly informed of the study hypotheses to minimize expectance effects. We acknowledge the Hawthorne effect, in which participants modify behavior because they know they are being observed, could have influenced responses. However, the randomized controlled design, balanced group assignments, and inclusion of both qualitative and quantitative data help mitigate potential bias and strengthen the validity of the findings. Future studies might incorporate validated infrequency or social desirability scales, such as the Infrequency Scale from the Zuckerman-Kuhlman Personality Questionnaire to further account for response bias.

Conclusions

Our study underscores the transformative potential of VR interventions in enhancing empathy among medical students. By allowing learners to experience the patient's perspective firsthand, VR offers an emotionally engaging, experiential form of learning that complements traditional didactic methods. Our findings, demonstrating higher JSE-S scores and richer qualitative reflections in the VR group, indicate that immersive learning environments can promote deeper perspective taking and maintain empathic growth. These findings illuminate the importance of adopting innovative, technology-enhanced educational strategies to cultivate a more empathetic health care workforce. They also suggest that technology-enhanced approaches may help counter the well-documented decline in empathy during medical training. Integrating VR into medical curricula could provide a scalable, standardized way to strengthen the humanistic foundations of physician education while aligning with contemporary adult learning principles. Future research should examine the long-term effects of such interventions and explore combinations of VR with reflective or clinical teaching to reinforce empathetic skills in practice. Our work lays a foundational step toward reimagining empathy training in medical education by leveraging immersive technology to cultivate emotional understanding and advance a more holistic model of physician training that integrates scientific expertise with the empathy essential to compassionate, patient-centered care.

XSL•FO
RenderX

Acknowledgments

Special thanks to our student research assistants: Kurestin Miller, Dan Strat, Meghan Dubose, Sarah Barto, and Anne Kirby; our actors Jenn Maness and Chris Freeland; and the Patients from the University of South Carolina's Patient Engagement Studio who reviewed the project and provided feedback for improvement on the patient/provider scenario for the video. Additionally, a game of Rochambeau (rock, paper, and scissors) was used to determine senior and corresponding author positions.

Funding

Partial financial support was received from a Prisma Health-Upstate seed grant, The University of South Carolina Center for Teaching Excellence teaching innovation grant, and the University of South Carolina School of Medicine SOARinG initiative.

Conflicts of Interest

None declared.

References

1. Singer T, Lamm C. The social neuroscience of empathy. Ann N Y Acad Sci 2009 Mar;1156(1):81-96. [doi: 10.1111/j.1749-6632.2009.04418.x] [Medline: 19338504]
2. Kim SS, Kaplowitz S, Johnston MV. The effects of physician empathy on patient satisfaction and compliance. Eval Health Prof 2004 Sep;27(3):237-251. [doi: 10.1177/0163278704267037] [Medline: 15312283]
3. Hojat M, Louis DZ, Markham FW, Wender R, Rabinowitz C, Gonnella JS. Physicians' empathy and clinical outcomes for diabetic patients. Acad Med 2011 Mar;86(3):359-364. [doi: 10.1097/ACM.0b013e3182086fe1] [Medline: 21248604]
4. Hojat M, Gonnella JS, Nasca TJ, Mangione S, Vergare M, Magee M. Physician empathy: definition, components, measurement, and relationship to gender and specialty. Am J Psychiatry 2002 Sep;159(9):1563-1569. [doi: 10.1176/appi.ajp.159.9.1563] [Medline: 12202278]
5. Hojat M, Mangione S, Nasca TJ, et al. An empirical study of decline in empathy in medical school. Med Educ 2004 Sep;38(9):934-941. [doi: 10.1111/j.1365-2929.2004.01911.x] [Medline: 15327674]
6. Batt-Rawden SA, Chisolm MS, Anton B, Flickinger TE. Teaching empathy to medical students: an updated, systematic review. Acad Med 2013 Aug;88(8):1171-1177. [doi: 10.1097/ACM.0b013e318299f3e3] [Medline: 23807099]
7. Wündrich M, Schwartz C, Feige B, Lemper D, Nissen C, Voderholzer U. Empathy training in medical students - a randomized controlled trial. Med Teach 2017 Oct;39(10):1096-1098. [doi: 10.1080/0142159X.2017.1355451] [Medline: 28749198]
8. Batson CD, Eklund JH, Chermok VL, Hoyt JL, Ortiz BG. An additional antecedent of empathic concern: valuing the welfare of the person in need. J Pers Soc Psychol 2007 Jul;93(1):65-74. [doi: 10.1037/0022-3514.93.1.65] [Medline: 17605589]
9. Vescio TK, Sechrist GB, Paolucci MP. Perspective taking and prejudice reduction: the mediational role of empathy arousal and situational attributions. Euro J Social Psych 2003 Jul;33(4):455-472. [doi: 10.1002/ejsp.163]
10. Galinsky AD, Ku G. The effects of perspective-taking on prejudice: the moderating role of self-evaluation. Pers Soc Psychol Bull 2004 May;30(5):594-604. [doi: 10.1177/0146167203262802] [Medline: 15107159]
11. Herrera F, Bailenson J, Weisz E, Ogle E, Zaki J. Building long-term empathy: a large-scale comparison of traditional and virtual reality perspective-taking. PLoS ONE 2018;13(10):e0204494. [doi: 10.1371/journal.pone.0204494] [Medline: 30332407]
12. Ahn SJG, Bostick J, Ogle E, Nowak KL, McGillicuddy KT, Bailenson JN. Experiencing nature: embodying animals in immersive virtual environments increases inclusion of nature in self and involvement with nature. J Comput-Mediat Comm 2016 Nov;21(6):399-419. [doi: 10.1111/jcc4.12173]
13. Zaki J. Empathy: a motivated account. Psychol Bull 2014 Nov;140(6):1608-1647. [doi: 10.1037/a0037679] [Medline: 25347133]
14. Macrae CN, Bodenhausen GV. Social cognition: thinking categorically about others. Annu Rev Psychol 2000;51:93-120. [doi: 10.1146/annurev.psych.51.1.93] [Medline: 10751966]
15. Oh SY, Bailenson J, Weisz E, Zaki J. Virtually old: embodied perspective taking and the reduction of ageism under threat. Comput Human Behav 2016 Jul;60:398-410. [doi: 10.1016/j.chb.2016.02.007]
16. Sirkkunen E, Väätäjä H, Uskali T, Rezaei PP. Journalism in virtual reality. Presented at: AcademicMindtrek'16; Oct 17-18, 2016; Tampere, Finland p. 297-303. [doi: 10.1145/2994310.2994353]
17. Benbassat J, Baumal R. What is empathy, and how can it be promoted during clinical clerkships? Acad Med 2004 Sep;79(9):832-839. [doi: 10.1097/00001888-200409000-00004] [Medline: 15326005]
18. Gallop R, Lancee WJ, Garfinkel PE. The empathic process and its mediators. a heuristic model. J Nerv Ment Dis 1990 Oct;178(10):649-654. [doi: 10.1097/00005053-199010000-00006] [Medline: 2230750]
19. Coll MP, Viding E, Rütgen M, et al. Are we really measuring empathy? Proposal for a new measurement framework. Neurosci Biobehav Rev 2017 Dec;83:132-139. [doi: 10.1016/j.neubiorev.2017.10.009] [Medline: 29032087]

XSL•FO
RenderX

20. Creswell JW, Plano Clark VL. Designing and Conducting Mixed Methods Research, 3rd edition: Sage Publications; 2017. URL: https://collegepublishing.sagepub.com/products/designing-and-conducting-mixed-methods-research-3-241842 [accessed 2025-12-30]
21. Deuter CE, Nowacki J, Wingenfeld K, et al. The role of physiological arousal for self-reported emotional empathy. Auton Neurosci 2018 Nov;214:9-14. [doi: 10.1016/j.autneu.2018.07.002] [Medline: 30104144]
22. Hojat M, DeSantis J, Shannon SC, et al. The Jefferson Scale of Empathy: a nationwide study of measurement properties, underlying components, latent variable structure, and national norms in medical students. Adv Health Sci Educ Theory Pract 2018 Dec;23(5):899-920. [doi: 10.1007/s10459-018-9839-9] [Medline: 29968006]
23. Hojat M, Gonnella JS, Nasca TJ, Mangione S, Veloksi JJ, Magee M. The Jefferson Scale of Physician Empathy: further psychometric data and differences by gender and specialty at item level. Acad Med 2002 Oct;77(10 Suppl):S58-S60. [doi: 10.1097/00001888-200210001-00019] [Medline: 12377706]
24. Yu J, Ganesh S, Lancastle D. Psychometric evaluation of the 'Jefferson Scale of Empathy' in a sample of nursing students in the United Kingdom. Nurs Open 2024 Jan;11(1):e2054. [doi: 10.1002/nop2.2054]
25. Piumatti G, Abbiati M, Baroffio A, Gerbase MW. Empathy trajectories throughout medical school: relationships with personality and motives for studying medicine. Adv Health Sci Educ Theory Pract 2020 Dec;25(5):1227-1242. [doi: 10.1007/s10459-020-09965-y] [Medline: 32095990]
26. Hojat M, Shannon SC, DeSantis J, Speicher MR, Bragan L, Calabrese LH. Empathy in medicine National norms for the Jefferson Scale of Empathy: a nationwide project in osteopathic medical education and empathy (POMEE). J Am Osteopath Assoc 2019 Aug 1;119(8):520-532. [doi: 10.7556/jaoa.2019.091] [Medline: 31355891]
27. Braun V, Clarke V, Terry G, Hayfield N. Thematic analysis. In: Liamputtong P, editor. Handbook of Research Methods in Health and Social Sciences: Springer; 2019:843-860. [doi: 10.1007/978-981-10-5251-4_103]
28. Braun V, Clarke V. Successful Qualitative Research: A Practical Guide for Beginners: SAGE Publications; 2013. URL: https://uk.sagepub.com/en-gb/eur/successful-qualitative-research/book233059 [accessed 2025-12-17]
29. Braun V, Clarke V. Using thematic analysis in psychology. Qual Res Psychol 2006 Jan;3(2):77-101. [doi: 10.1191/1478088706qp063oa]
30. Tavakol S, Dennick R, Tavakol M. Psychometric properties and confirmatory factor analysis of the Jefferson Scale of Physician Empathy. BMC Med Educ 2011 Aug 2;11:54. [doi: 10.1186/1472-6920-11-54] [Medline: 21810268]
31. Doyen S, Klein O, Simons DJ, Cleeremans A. On the other side of the mirror: priming in cognitive and social psychology. Soc Cogn 2014 Jun;32:12-32. [doi: 10.1521/soco.2014.32.supp.12]
32. Chartrand TL, Bargh JA. Automatic activation of impression formation and memorization goals: nonconscious goal priming reproduces effects of explicit task instructions. J Pers Soc Psychol 1996;71(3):464-478. [doi: 10.1037//0022-3514.71.3.464]
33. Knowles MS, Holton EF, Swanson RA. The Adult Learner, 8th edition: Routledge; 2015. [doi: 10.4324/9781315816951]
34. Decety J, Jackson PL. The functional architecture of human empathy. Behav Cogn Neurosci Rev 2004 Jun;3(2):71-100. [doi: 10.1177/1534582304267187] [Medline: 15537986]
35. Merriam SB. Adult learning theory for the twenty‐first century. New Dir Adult Contin Educ 2008 Sep;2008(119):93-98. [doi: 10.1002/ace.309]
36. Riva G, Wiederhold BK, Mantovani F. Neuroscience of virtual reality: from virtual exposure to embodied medicine. Cyberpsychol Behav Soc Netw 2019 Jan;22(1):82-96. [doi: 10.1089/cyber.2017.29099.gri] [Medline: 30183347]
37. Christov-Moore L, Simpson EA, Coudé G, Grigaityte K, Iacoboni M, Ferrari PF. Empathy: gender effects in brain and behavior. Neurosci Biobehav Rev 2014 Oct;46 Pt 4(Pt 4):604-627. [doi: 10.1016/j.neubiorev.2014.09.001] [Medline: 25236781]
38. Winter R, Issa E, Roberts N, Norman RI, Howick J. Assessing the effect of empathy-enhancing interventions in health education and training: a systematic review of randomised controlled trials. BMJ Open 2020 Sep 25;10(9):e036471. [doi: 10.1136/bmjopen-2019-036471] [Medline: 32978187]

Abbreviations

JSE-S: Jefferson Scale of Empathy for medical students
V1: lecture video 1
V2: lecture video 2
V3: lecture video 3
VR: virtual reality

XSL•FO
RenderX

Edited by J Eriksen; submitted 24.Apr.2025; peer-reviewed by R Aro, S Bidmon; accepted 30.Nov.2025; published 04.Feb.2026.

Please cite as:
Mundok AG, Ho VN, Fowler LA, Kennedy AB, Stark-Taylor S
Investigating the Impact of a Virtual Reality Experience on Medical Student Empathy: Mixed Methods Study
JMIR Med Educ 2026;12:e76504
URL: https://mededu.jmir.org/2026/1/e76504
doi:10.2196/76504

© Allen G Mundok, Vivian N Ho, Lauren A Fowler, Ann Blair Kennedy, Shannon Stark-Taylor. Originally published in JMIR Medical Education (https://mededu.jmir.org), 4.Feb.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Education, is properly cited. The complete bibliographic information, a link to the original publication on https://mededu.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Enhancing Interprofessional Team Performance to Prevent Medication Errors in Emergency Care: Quasi-Experimental Study Using Multimodal Virtual Simulation-Based Interprofessional Education

Ora-In Chu[1], BSc, BBA; Phanupong Phutrakool[2], PhD; Khrongwong Musikatavorn[3,4], MD; Thitiporn Kongchim[5], MNS; Lapol Herabat[3,4], MD; Jiraphan Ritsamdang[6], BPharm; Krittin Bunditanukul[6], PharmD; Kanittha Triamamornwooth[7], MScPharm; Khuansiri Narajeenron[3,4], MD, MSc, MHPE, CHSE

[1]Faculty of Medicine, Chulalongkorn University, Bangkok, Thailand

[2]Chula Data Management Center, Faculty of Medicine, Chulalongkorn University, Bangkok, Thailand

[3]Department of Emergency Medicine, Faculty of Medicine, Chulalongkorn University, Bangkok, Thailand

[4]Department of Emergency Medicine, King Chulalongkorn Memorial Hospital, Thai Red Cross Society, Bangkok, Thailand

[5]Emergency Department of the Nurse Department, King Chulalongkorn Memorial Hospital, Thai Red Cross Society, Bangkok, Thailand

[6]Department of Pharmacy Practice, Faculty of Pharmaceutical Sciences, Chulalongkorn University, Bangkok, Thailand

[7]Clinical Pharmacy Services, Pharmacy Division, King Chulalongkorn Memorial Hospital, Thai Red Cross Society, Bangkok, Thailand

Corresponding Author:
Khuansiri Narajeenron, MD, MSc, MHPE, CHSE
Department of Emergency Medicine, Faculty of Medicine
Chulalongkorn University
1873 MDCU Faculty, Rama IV Road, Pathumwan
Bangkok, 10330
Thailand
Phone: 66 022564000 ext 83058
Email: khuansiri.n@chula.ac.th

Abstract

Background: Effective interprofessional collaboration (IPC) is essential for patient safety; yet, poor teamwork and communication remain key challenges in high-pressure settings like the emergency department (ED), contributing to medication errors. Although Team Strategies and Tools to Enhance Performance and Patient Safety (TeamSTEPPS)–based interprofessional education addresses these issues, adaptation in clinical settings remains difficult. To bridge this gap, we developed Emergency Room Virtual Simulation-Based Interprofessional Education (ER-VIPE), a multimodal, TeamSTEPPS-integrated intervention designed to enhance IPC and reduce medication errors.

Objective: The aim of the study is to evaluate the effectiveness of ER-VIPE in enhancing IPC performance among emergency physicians, nurses, and pharmacists and in reducing medication errors. The primary objective is to assess changes in IPC performance in both real-world ED settings and in computer-based simulations. The secondary objective is to examine the intervention's impact on medication error rates in the ED.

Methods: This quasi-experimental study involved 15 interprofessional teams (each comprising 1 physician, 1 pharmacist, and 2 nurses), undergoing the ER-VIPE training. This multimodal intervention included 2 medical films, a massive open online course on TeamSTEPPS and IPC, and a computer-based simulation session on acute chest pain and cardiac arrest scenarios via the simulation-based interprofessional education (SIMBIE) platform. Co-debriefings were provided as a complement to the SIMBIE session, guiding participants through positive feedback and areas of improvement. TeamSTEPPS performance was measured using the Modified TeamSTEPPS and Team Performance Observation Tool (mTPOT) in both simulation and real-world ED settings. Generalized estimating equations with a Gaussian family, identity link, and exchangeable correlation structure were used to evaluate IPC score changes. Chi-square and Fisher exact tests were applied to compare near-miss and actual medication errors before and after the intervention. A 2-tailed *P* value <.05 was considered statistically significant.

XSL•FO
RenderX

Results: The study was conducted from November 2023 to January 2024 at a university hospital with 60 participants. Following the co-debriefing session in the simulation, overall mTPOT scores increased by 2.00 points (P<.001), with the greatest improvement among physicians (+2.70), followed by nurses (+1.75) and pharmacists (+1.56). In the ED, most mTPOT domains improved significantly across all professions 2 months after the intervention (P<.001). Although no significant reduction in harmful medication errors was observed, reporting of near-miss prescription errors increased significantly (P=.01).

Conclusions: ER-VIPE enhanced IPC among ED physicians, nurses, and pharmacists, with sustained effects observed up to 2 months in real-world settings. The combination of medical films and massive open online courses provided accessible foundational knowledge, while computer-based virtual SIMBIE with co-debriefing reinforced practical communication and teamwork. Increased near-miss reporting suggests improved situational awareness and a more transparent safety culture. This multimodal training model shows promise for advancing collaboration and patient safety in emergency care.

(JMIR Med Educ 2026;12:e66999) doi:10.2196/66999

KEYWORDS

interprofessional collaboration; interprofessional education; emergency medicine; TeamSTEPPS; Team Strategies and Tools to Enhance Performance and Patient Safety; computer-based simulation; virtual simulation-based interprofessional education; medical movie; massive open online courses; co-debriefing; medication error

Introduction

Interprofessional collaboration (IPC), as defined by the World Health Organization, is the practice of multiple health workers from different professional backgrounds working together with patients, families, caregivers, and communities to deliver the highest quality of care [1]. IPC plays a critical role in improving patient care quality, safety, and outcomes by facilitating coordination in identifying needs, solving problems, and making joint decisions [2,3]. IPC is a key outcome for ensuring patient safety, and interprofessional education serves as a strategic educational approach to promote and strengthen IPC. Recognizing the global need for effective IPC, the Interprofessional Education Collaborative (IPEC) has established core competencies for interprofessional collaborative practice. These competencies are categorized into 4 major pillars: teams and teamwork, values and ethics, roles and responsibilities, and communication [4].

Despite the benefits of IPC, barriers to collaboration persist [5], contributing to medication errors, particularly medication-related harm [6]. Estimates suggest that medication errors may represent up to half of all preventable harm within health care, and these errors can be serious or even life-threatening [7]. Out of all medication error events, prescribing errors account for more than 50% of all medication errors; yet, only 30% to 70% are identified by pharmacists and nurses before administration [8-10]. Poor teamwork and interprofessional communication are major contributors to medication errors. These errors often stem from incomplete medication histories and inadequate engagement of patients and families in the care process. Ineffective team communication—such as lack of clarification, failure to verify, or neglecting to cross-check medication orders—further increases the risk, particularly during patient handoffs and in high-pressure clinical settings such as the emergency department (ED) [7,9]. The ED possesses the greatest need of IPC, where rapidly changing health care teams are required to make rapid decisions [11], often under conditions of uncertainty and resource constraints [12].

One of the most well-established frameworks for teaching IPC is Team Strategies and Tools to Enhance Performance and Patient Safety (TeamSTEPPS), developed by the US Agency for Healthcare Research and Quality (AHRQ) [13]. TeamSTEPPS provides structured training in team structure, leadership, situation monitoring, mutual support, and communication, all of which are part of the core competencies of IPC set by the IPEC. This framework is essential for enhancing collaboration and reducing errors [14]. Multiple studies have demonstrated the positive impact of TeamSTEPPS training on patient outcomes, team efficiency, and safety culture [15]. Traditional methods of TeamSTEPPS training, like e-learning and classroom-based instruction, often lack the ability to immerse learners in the high-pressure, multifactorial scenarios typical of emergency settings. Consequently, the application of TeamSTEPPS principles during hospital crises is compromised. Additionally, while TeamSTEPPS has advanced training practices, its lengthy sessions can place a heavy load on facilitators and diminish learners' engagement.

To overcome traditional IPC TeamSTEPPS training limitations, this study introduces Emergency Room Virtual Simulation-Based Interprofessional Education (ER-VIPE) [16], a Thailand's pioneering compact, scalable, and cost-effective virtual reality platform for IPC in emergency settings. ER-VIPE combines medical-themed movies, massive open online courses (MOOCs), and simulation-based interprofessional education (SIMBIE) with co-debriefing, and is designed to foster a no-blame culture, strengthen teamwork and communication, including patient and family engagement, and work under high-pressure conditions. The pretraining movie and MOOC components are designed to build foundational knowledge and attitudes related to the TeamSTEPPS framework, while reducing instructional burden through a self-paced, learner-directed format aligned with the flipped classroom model [17]. Following this preparatory phase, the SIMBIE component—combined with co-debriefing—offers learners the opportunity to apply these concepts in practice, grounded in experiential learning theory [18-24].

This study aims to evaluate the effectiveness of ER-VIPE in enhancing IPC performance and reducing medication errors through the application of the TeamSTEPPS framework. The primary objective is to quantitatively assess changes in IPC

XSL•FO
RenderX

performance using the Modified TeamSTEPPS and Team Performance Observation Tool (mTPOT), evaluating participants both in real-world ED settings and in computer-based virtual SIMBIE sessions, following full implementation of the TeamSTEPPS-integrated ER-VIPE intervention. The secondary objective is to evaluate the intervention's impact on medication error rates in the ED using retrospective data from hospital records collected before and after implementation of ER-VIPE among ED health care teams. The a priori hypothesis is that the implementation of ER-VIPE will improve IPC performance and will be associated with a reduction in medication error rates in the ED.

Methods

Research Design

Due to the rotational nature of clinical duties in the ED and logistical constraints within the real-world training environment, random assignment of participants was not feasible. Therefore, a quasi-experimental pre- and posttest design was used to evaluate the effectiveness of the multimodal TeamSTEPPS-based intervention under practical clinical conditions. This is a quasi-experimental pre- and posttest study designed to evaluate the impact of ER-VIPE integrated with the TeamSTEPPS framework as an intervention to train IPC for interprofessional health care teams and its efficacy in the reduction of medication error rates. The study is conducted in the setting of the ED at a 1500-bed teaching hospital in Thailand.

Theoretical Framework

Each component of the ER-VIPE intervention was grounded in established educational and behavioral theories that justify its design for teaching TeamSTEPPS principles to reduce medication errors in the ED. The foundation of ER-VIPE integrates IPC with TeamSTEPPS, emphasizing team structure, communication, leadership, mutual support, and situation monitoring. These competencies are well recognized as essential for reducing prescribing, dispensing, and administration errors by strengthening shared mental models, clarifying team roles, and improving coordination among physicians, nurses, and pharmacists.

Cinemeducation and social learning theory [25] further inform the design through emotionally engaging films depicting both safe and unsafe management of high-alert drugs (HADs) and adverse drug reactions (ADRs). Observing modeled behaviors and their consequences enhances learners' recognition of error-prone conditions and reinforces medication safety principles [26]. This foundation is complemented by experiential learning theory [27], which underpins the SIMBIE platform. SIMBIE provides deliberate, repetitive practice of medication-related teamwork behaviors within controlled scenarios, enabling learners to develop adaptive responses, strengthen role clarity, and improve communication processes essential for preventing medication errors.

Finally, the co-debriefing component is informed by social constructivism [28] and transformative learning theory [29]. Through guided reflection, dialogue, and facilitated meaning-making, learners critically examine assumptions underlying their medication practices, identify latent safety threats, and reconstruct safer mental models for IPC. This reflective process supports both individual and collective transformation, consolidating learning into long-term behavioral change [30]. Together, these interconnected frameworks provide a cohesive theoretical rationale for how ER-VIPE enhances cognitive, emotional, and team-based capacities necessary to reduce medication errors in the ED.

Components of the ER-VIPE Training

The ER-VIPE teaching course is a multimodal virtual simulation-based interprofessional education bundle designed specifically by the research team as an innovation to teach interprofessional health care teams in team working skills to achieve the main outcome of reducing medication errors in the ED. The bundle consists of 3 major components including the medical movies, IPC course in the MOOC platform, and SIMBIE with a co-debriefing process.

Medical Movies

There were a total of 2 in-house production including a 48-minute movie called "Small mistakes, big consequences." This movie covers the topic of HAD do's and don'ts and ADR do's and don'ts. The second movie is a 75-minute film titled "Emergency love." Both movies were designed to demonstrate both effective and ineffective examples of teamwork in ED settings. Participants who completed the course and passed both the quiz and the final assessment had received certification prior to proceeding to the next stage of training.

Medical movies, as part of the ER-VIPE intervention, had specific learning objectives, in which participants were expected to recognize the value of IPC and TeamSTEPPS principles in improving medication safety, particularly in relation to HADs and ADRs. They were also expected to describe strategies for engaging patients and families; identify methods to prevent medication errors across prescribing, dispensing, and administration stages; and apply effective monitoring and reassessment techniques. Additionally, the course encouraged the application of metacognitive skills to help participants manage their emotions and cognitive biases during high-pressure situations.

Massive Open Online Courses

The MOOC platform, a publicly available, subscription-based online teaching initiative by Chulalongkorn University, aims to promote lifelong learning and public education and provides a scalable educational solution, enabling expansion of the learner base while maintaining educational quality and effectiveness through a robust online technological infrastructure. It is also a generalizable solution, applicable across diverse learner populations and educational settings [31,32].

The 1.45-hour self-paced online course on IPC for patient safety is made available through the MOOC platform. The contents were developed by contributions from experts and representative lecturers from the Faculty of Medicine, the Faculty of Nursing, the Faculty of Allied Health Sciences, the Faculty of Pharmaceutical Sciences, the Faculty of Psychology, the Faculty

XSL•FO
RenderX

of Arts, the Healthcare Accreditation Institute of Thailand, and aviation training professionals.

This short course is divided into 7 soft skills modules, covering IPEC's core competencies for IPC, TeamSTEPPS, critical team thinking for diagnosis, coping strategies for IPC, communication for IPC, 2P safety goals (patient and personnel safety), and ethical principles for IPC. The participant must complete a quiz after each module to receive a certificate of completion.

Computer-Based Virtual SIMBIE and Co-Debriefing Process

SIMBIE, a key component of the ER-VIPE, is an interprofessional simulation platform designed and developed by the ER-VIPE Study Group [33] in collaboration with Chulalongkorn University. It is designed using virtual simulation and 3D computer-based simulation technology [33] for virtual access through commonly available devices such as computers, enabling accessibility across institutions and adaptability to a wide range of educational contexts—thereby enhancing the platform's generalizability. Each SIMBIE scenario is developed as a multiplayer simulation, supporting scalability for simultaneous multiuser participation. This approach ensures consistent educational quality and learner engagement, facilitated by a reliable and well-integrated online technological infrastructure [34]. The platform presents a scenario where a patient is admitted to the ED with ischemic heart disease and acute heart failure, progressing to cardiogenic shock and sudden cardiac arrest with ventricular fibrillation. Each team, consisting of a physician, a pharmacist, and 2 nurses, is responsible for managing the patient collaboratively. Players communicate via simulated telephones when avatars are located in separate areas within the game (eg, a physician in the ED and a pharmacist in the pharmacy) or through direct verbal communication when avatars are situated in the same simulated location (eg, both the physician and nurses at the patient's bedside). Players can interact with in-game items, such as syringes and medications, by selecting them within the simulation and are expected to adhere to standard hospital protocols, including patient identification procedures using identification wristbands in the ED. A detailed description of the SIMBIE user experience has been previously reported by Narajeenron et al [16].

While managing the case, participants are evaluated by a trainer based on their respective professions, followed by a personalized co-debrief session. In this context, personalization refers to the tailored feedback and reflective discussion provided to each participant, considering their specific role, decision-making, and contributions during the scenario. This individualized approach ensures that feedback aligns with the TeamSTEPPS framework, thereby enhancing the relevance and effectiveness of the learning experience. Structured co-debriefing is conducted by 3 trained TeamSTEPPS facilitators—a physician, a nurse, and a pharmacist educator [35]. These pretrained facilitators were to observe and document key behaviors of learners during their simulation case according to the TeamSTEPPS objectives and were each assigned to lead specific TeamSTEPPS topics to provide actionable strategies for future similar situations. These strategies were coordinated by a debriefing leader to coordinate pre-debriefing, debriefing, and post-debriefing activities among the facilitators [16]. The debriefing process was delivered using the gather, analyze, summarize model [36], which promotes critical self-reflection, recognition of effective practices, and identification of areas for improvement.

The gather, analyze, summarize model supports transformative learning and fosters a growth mindset by guiding participants through 3 key phases. In the "gather" phase, participants share their experiences and emotions from the simulation, promoting self-awareness and engagement. In the "analyze" phase, facilitators lead discussions on the application of the 5 TeamSTEPPS domains in crisis situations, encouraging critical evaluation of both effective and suboptimal behaviors. The final "summarize" phase concludes with reflection on key takeaways and actionable strategies, helping participants identify behaviors to continue, modify, or discontinue in real-life situations.

After the co-debriefing, participants repeat the simulation, applying the feedback received to reinforce their learning. Trainers then re-evaluate their performance, assessing improvements and ensuring that critical concepts have been embedded into their practice.

Participants

Participants were recruited from among 3 health care professionals working in the ED during the period of this study. Recruitment was conducted through departmental announcements and direct invitations by research assistants. Participation was voluntary and had no impact on job performance evaluations outside of this study. Informed consent was obtained from all participants prior to enrollment.

The final sample comprised 15 emergency medicine residents, 15 pharmacists (comprising 5 ED pharmacists and 10 pharmacy students on ED clerkship rotations functioning as clinical pharmacists), and 30 level 6 registered nurses (nurses who have worked in the emergency unit for over 6 years). All recruited participants are qualified to treat Emergency Severity Index level 1 and 2 patients in the ED.

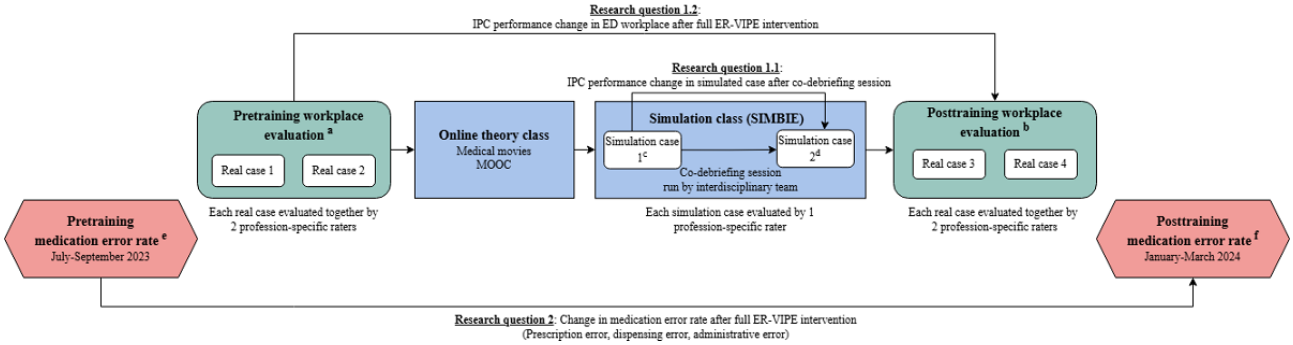
Outcome Measurements

Overview

Participants were assigned to undergo the ER-VIPE training in the sequence shown in Figure 1, with participants' IPC performance based on the TeamSTEPPS framework being evaluated at 4 time points (time points a and b as workplace-based pre- and posttraining evaluation, and time points c and d as simulation-based pre- and post–co-debriefing session evaluation) using an evaluation form. Medication error rates were measured at 2 time points (time points e and f) using retrospective data.

XSL•FO
RenderX

Figure 1. Research design and intervention overview. ED: emergency department; ER-VIPE: Emergency Room Virtual Simulation-Based Interprofessional Education; ESI: Emergency Severity Index; IPC: interprofessional collaboration; MOOC: massive open online course. aPretraining workplace IPC performance evaluations by 2 raters from the corresponding profession for 2 ESI level 1 or level 2 real cases in the ED. The assessments are conducted within 2 weeks prior to the ER-VIPE training. bPosttraining workplace IPC performance evaluations by 2 raters from the corresponding profession for 2 ESI level 1 or level 2 real cases in the ED. The assessment takes place between 10 days and 2 months after the ER-VIPE training. cPre–co-debriefing IPC performance evaluations of a simulated case by a rater (also acting as trainer) for an ESI level 2 patient with chest pain presenting with acute coronary syndrome and new-onset heart failure. dPost–co-debriefing IPC performance evaluations of a simulated case by a rater (also acting as trainer) for an ESI level 2 patient with chest pain presenting with acute coronary syndrome and new-onset heart failure. ePre–ER-VIPE intervention medication error rate in ED, recorded in a 3-month period prior to ER-VIPE training of participants. fPost–ER-VIPE intervention medication error rate in ED, recorded in a 3-month period after ER-VIPE training of participants.



To ensure an accurate assessment of typical workplace performance, participants were notified only 24 hours before the evaluation period. Workplace performance evaluations were scheduled according to participants' preassigned shift schedules in the ED to minimize disruption to staffing and clinical operations.

Evaluation Tool for IPC Performance Based on TeamSTEPPS Framework

To assess the IPC performance of participants, we used a TeamSTEPPS and Team Performance Observation Tool (TPOT). The TPOT was published in 2004 by the AHRQ [13], in accordance with the TeamSTEPPS framework, and is an evaluation tool used to assess the quality of IPC skills according to the TeamSTEPPS framework. A Thai version of the original tool was developed and validated prior to this study [37].

Our research team had elaborated on the details of each item in the TPOT into a modified version of TPOT called mTPOT, with details of each measuring item designed to assist evaluators in assessing specific behaviors of participants, particularly those that contribute to preventing medication errors. Each item within the tool focuses on critical aspects of care, ensuring that the health care professionals perform actions that promote patient safety and reduce the risk of errors. Three sets of mTPOT were developed, each designed for physicians, nurses, and pharmacists. Each tool was specifically adapted to measure IPC behaviors expected of each professional group to support medication error reduction.

An example of the nurse evaluation within the communication domain of the TeamSTEPPS framework focuses on assessing the nurse's ability to use the check-back technique when verifying information with the health care team, in order to ensure patient safety and accuracy during medication administration. The evaluation includes the use of patient identification techniques, requiring at least 2 of the following identifiers: patient's name, surname, hospital number, or date of birth. Additionally, it assesses medication verification practices, including (1) correct medication name, (2) precise dosage with units (eg, 200 mg), (3) route of administration (eg, intravenous, intramuscular, and oral), (4) confirmation of diluent and volume if dilution is required, (5) frequency of administration (eg, immediately [STAT] and every 6 hours), (6) rate of administration (eg, slow intravenous push or infusion over 1 hour), and (7) duration of treatment (eg, total number of tablets or intravenous drip time).

To assess the quality of the check-back technique used by nurses, a 5-point Likert scale was used. Evaluators assigned scores based on predefined criteria reflecting the completeness and consistency of information verification. A full score of 5 was given when the nurse's verification process is consistently performed, and complete information is received every time. A score of 4 is given when the nurse tries to verify information every time, but sometimes receives incomplete information. A score of 3 is given when verification is sometimes performed, and the nurse sometimes gets complete information. A score of 2 means that the nurse performs verification sometimes and never receives complete information each time. Finally, the lowest score of 1 is given when no verification happens when information is communicated.

The evaluation form consists of 5 sections aligned with the 5 TeamSTEPPS domains, incorporating specialty-specific behavioral indicators for each item. The team structure domain includes assessment of patient and family involvement in decision-making. The communication domain evaluates clarity, conciseness, accuracy of information exchange, use of verification techniques, and adherence to Identification, Situation, Background, Assessment, Recommendation handover protocols. The leadership domain assesses goal-setting, resource use workload distribution, and delegation. The situation monitoring domain examines the quality and frequency of patient monitoring, safety checks, and treatment follow-ups. The mutual support domain evaluates the ability to provide feedback, issue warnings, and foster a culture of teamwork.

All items in the mTPOT are scored using a 5-point Likert scale (1=very poorly done and 5=excellent, consistently done), with observers rating each participant's behavior against

XSL•FO
RenderX

specialty-specific performance standards. Higher scores reflect more frequent and proficient teamwork behaviors. Trained evaluators completed the form through direct observation, as participants managed each patient case.

An example of the TeamSTEPPS scoring tool (mTPOT) used for evaluating nurses is included in Multimedia Appendix 1 to demonstrate the types of interaction metrics assessed. Due to intellectual property restrictions, the complete set of evaluation tools for all professional groups cannot be made publicly available.

Content validation was conducted by a panel of 3 experts from each professional group to evaluate the profession-specific versions of the mTPOT. Content Validity Index (CVI) scores above 0.8 indicated strong agreement among the expert panel, confirming that the behavioral anchors demonstrated a high degree of content validity before the final version was established for use [38].

Medication Error Rate

Overview

Retrospective secondary data on medication errors, including near-miss events and miss events at all levels of ADRs in the ED, were requested from the hospital's data center. The data were collected over a 3-month period (August to October 2023) before the ER-VIPE intervention and a 3-month period (January to March 2024) after the intervention. Medication errors were further categorized as prescription errors (PEs) by physicians, dispensing errors (DEs) by pharmacists, and administration errors (AEs) by nurses. Medication error incidents are voluntarily reported by health care team members to their supervisors or department heads for formal reporting and follow-up. Medication error data will be reported as an error rate through additional comparison with monthly ED visits and monthly prescription orders sent to the ED pharmacy.

In reporting, medication errors were categorized into 3 levels of severity: near misses that do not reach the patient, no-harm incidents that reach the patient but cause no harm, and harmful incidents that cause patient harm. Due to the voluntary nature of reporting, any increase in near-miss incidents could also correlate with a heightened detection of medication errors and a safety culture that enables medical professionals to detect errors before it reaches patients.

Collective and Validated Data on PE and DE

Each pharmacy unit within the ED independently collects data on PE and DE and submits these reports to the Risk Management (RM) Committee. The pharmacy department consolidates these monthly reports for validation and further analysis. The RM Committee systematically reviews the validated data to identify trends, implement preventive measures, and recommend corrective actions to mitigate the recurrence of such incidents.

Collective and Validated Data on AE

Each nurse in the ED independently collects incident reports on AE and submits them to the head nurse for initial review. The head nurse assesses the report, determining necessary follow-up actions, such as work review, supervisory interventions, and procedural modifications, to address the issue. Following this review, the report undergoes further validation and investigation by the supervisory inspector, who may coordinate with the pharmacy department or the patient care team as needed. The chief nursing officer then reviews and verifies the findings before submission to the RM Committee for final evaluation. The RM Committee consolidates and systematically analyzes the validated data monthly, identifying trends, implementing preventive measures, and recommending corrective actions to mitigate recurrence.

Statistical Analysis

Overview

Data analysis was performed using Stata (version 15; StataCorp LLC). Each stage of the analysis was conducted according to the following methods: (1) The content validity of the mTPOT assessment form was evaluated using the CVI prior to its implementation as a performance measurement tool. (2) Participants' demographic data (sex, age, and job experience) and baseline team performance scores, as measured by the mTPOT, were summarized using descriptive statistics. Continuous variables were reported as means with SDs and medians with IQRs, while categorical variables were presented as frequencies and percentages. (3) Baseline team performance was measured using the mTPOT. One-way ANOVA was used to compare normally distributed baseline mTPOT scores across professional groups in the initial simulated case (Multimedia Appendix 2). The Kruskal-Wallis *H* test was applied to compare nonnormally distributed workplace mTPOT scores (Multimedia Appendix 3) as well as demographic variables such as age and job experience across the 3 professions. (4) Generalized estimating equations with a Gaussian family, an identity link function, and an exchangeable correlation structure were used to measure changes in participants' TeamSTEPPS performance (measured by mTPOT) in the ED before and after the intervention, as well as changes in TeamSTEPPS performance (measured by mTPOT) in simulation cases before and after debriefing. The average score for each domain is calculated using the total scores of each item in that section divided by the number of items assessed. The full score for each domain is 5, and the lowest score is 1. (5) Interrater reliability for each assessment was calculated using the Cohen κ coefficient, with quadratic weights to ensure the impartiality of the results. (6) Near-miss and miss events of medication errors in the ED before and after the intervention were assessed using the chi-square test and Fisher exact test. (7) A 2-tailed *P* value of <.05 was considered statistically significant for all analyses.

Sample Size Calculation

The sample size was determined based on a power analysis using a 2-tailed paired *t* test to compare mean differences between 2 dependent groups in a pre- and postintervention design. A total of 60 participants were required to detect an effect size of 0.37 with a significance level .05 and 80% power. This effect size, larger than a small effect (0.2) but below a medium effect (0.5), remains clinically and practically significant in this study.

XSL•FO
RenderX

Ethical Considerations

This research was approved by the Ethical Review Committee for Research Involving Human Research Subjects, Faculty of Medicine, Chulalongkorn University (approval COA 1508/2023). All participants were informed that their participation was voluntary and that they could withdraw from the study at any time without academic or professional consequences. Written informed consent was obtained prior to data collection. Participants' privacy and confidentiality were strictly protected; all data were anonymized and stored securely, accessible only to the research team. As compensation for their time, participants received approximately US $27 after completing the study activities. No additional incentives were provided.

Results

The Study Cohort

The study was conducted at a teaching hospital from November 2023 to January 2024, involving 60 interdisciplinary participants. The cohort (Table 1) comprised 15 physicians (8 first-year and 7 second-year residents), 15 pharmacists (5 ED pharmacists and 10 final-year pharmacy students rotating through the ED), and 30 ED nurses. Of the total workforce in the ED, this cohort represents 62.5% (15/24) of emergency medicine residents, 71.4% (5/7) of pharmacists in the ED pharmacy, 100% (10/10) of pharmacy students in the ED, and 29.1% (30/103) of ED nurses. The mean age of participants is 30.55 (SD 6.37) years, with 73.3% (44/60) being female and having an average experience in the ED of 6.07 (SD 7.04) years. The minimum working experience in the ED is 0 years for pharmacists, and the maximum is 28 years for nurses. Of all participants, 70% (42/60) reported no prior TeamSTEPPS training, 85% (51/60) were unfamiliar with the IPEC core competencies, and 75% (45/60) had not received any medication error prevention training.

Table 1. Demographic of participants by health care profession.

Factor	Overall (N=60)	Physician (n=15)	Nurse (n=30)	Pharmacist (n=15)	*P* value
Sex, n (%)					.17[a]
Female	44 (73.3)	8 (53.3)	24 (80)	12 (80)	
Male	16 (26.7)	7 (46.7)	6 (20)	3 (20)	
Age (years)					
Mean (SD)	30.55 (6.37)	27.53 (1.25)	34.27 (6.86)	26.13 (3.42)	—[b]
Median (IQR)	28.00 (27.00-32.50)	27.00 (27.00-28.00)	32.50 (28.00-40.00)	24.00 (23.00-29.00)	<.001[c]
Job experience (years)					
Mean (SD)	6.07 (7.04)	1.45 (1.03)	11.37 (6.46)	0.08 (0.32)	–
Median (IQR)	3.75 (0.42-10.00)	1.42 (0.50-2.50)	10.00 (5.42-17.00)	0.00 (0.00-0.00)	<.001[c]
Previous training, n (%)					
TeamSTEPPS[d] (all)	18 (30)	2 (13.3)	10 (33.3)	6 (40)	.24[a]
TeamSTEPPS (Sim[e])	8 (13.3)	2 (13.3)	6 (20)	0 (0)	.21[a]
IPEC[f] (all)	9 (15)	2 (13.3)	2 (6.7)	5 (33.3)	.06[a]
IPEC (Sim)	2 (3.3)	1 (6.7)	1 (3.3)	0 (0)	>.99[a]
Medication error (all)	15 (25)	3 (20)	5 (16.7)	7 (46.7)	.11[a]
Medication error (Sim)	3 (5)	1 (6.7)	2 (6.7)	0 (0)	.80[a]

[a]Fisher exact test.

[b]Not available.

[c]Kruskal-Wallis *H* test.

[d]TeamSTEPPS: Team Strategies and Tools to Enhance Performance and Patient Safety.

[e]Sim: simulation as a training modality.

[f]IPEC: Interprofessional Education Collaborative.

XSL•FO
RenderX

Key Highlights

Our findings are consistent with our a priori hypothesis that the ER-VIPE intervention significantly improves IPC performance across all professional groups in both simulation case and in the ED workplace, as demonstrated by increased mTPOT scores (P<.001). This improvement was accompanied by an increase in the reporting of near-miss PEs (P=.01), likely representing clinically meaningful improvements, whereby enhanced team performance translates into safer medication practices.

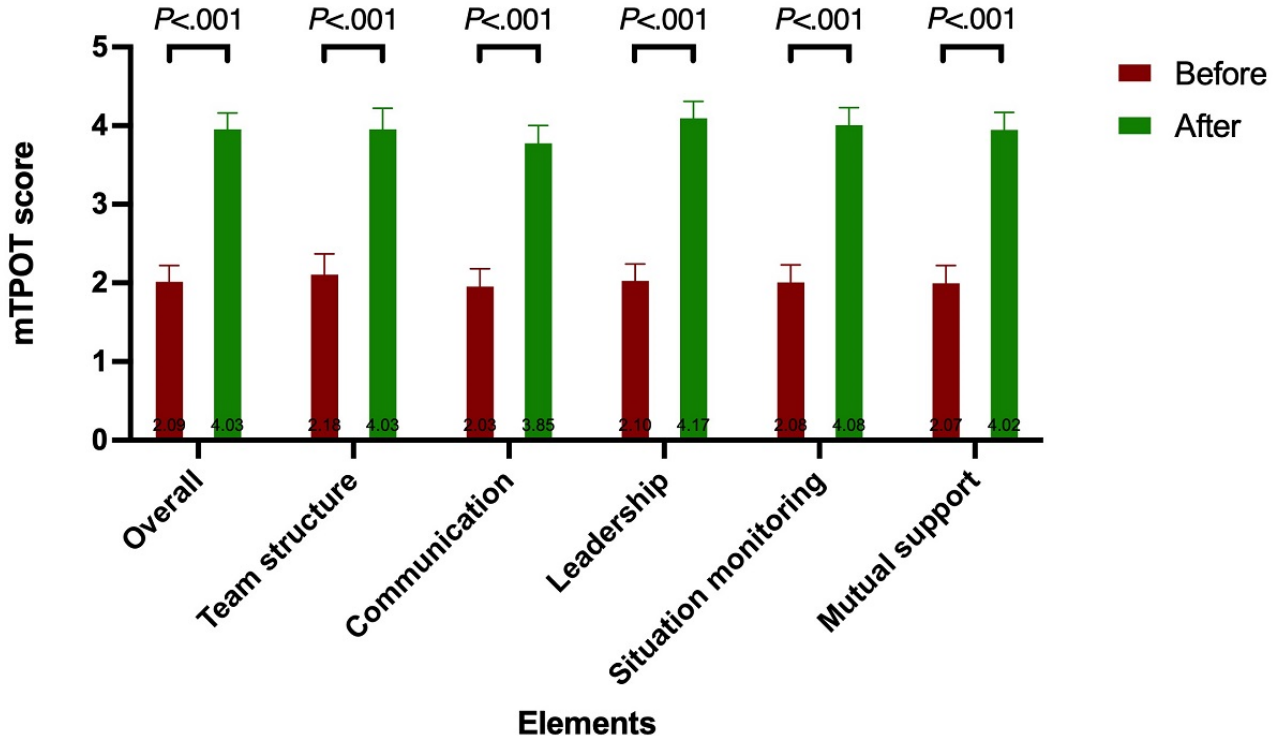
Changes in Team Performance

TeamSTEPPS IPC Performance Outcome in Simulated Scenarios

With 5 as the maximum score and 1 as the minimum score for each domain of TeamSTEPPS, the baseline mTPOT scores across all disciplines (Multimedia Appendix 2) averaged 2.09, with physicians scoring the lowest (1.74) and pharmacists the highest (2.49). In the mTPOT baseline scoring, the overall strength of the 15 multidisciplinary teams is team structure, while communication is the weakest area. Regarding individual specialties, physicians score lowest on mutual support (1.62), while nurses perform worst in team structure (2.00). Leadership skills are the strongest area for both physicians and nurses, scoring 1.90 and 2.13, respectively. Conversely, pharmacists scored lowest in leadership (2.24), while achieving the highest score in team structure (3.07).

Following the co-debriefing session after SIMBIE, the overall improvement in TeamSTEPPS performance as measured by mTPOT (Multimedia Appendix 4) was +2.00 points (P<.001), with physicians showing the most significant increase (+2.70) and pharmacists the least (+1.56). Nurses showed a modest increase, with an average of +1.75 points. Of the 5 domains evaluated, all showed significant improvement (P<.001), with leadership skills improving the most, followed by situation monitoring, mutual support, team structure, and communication (Figure 2).

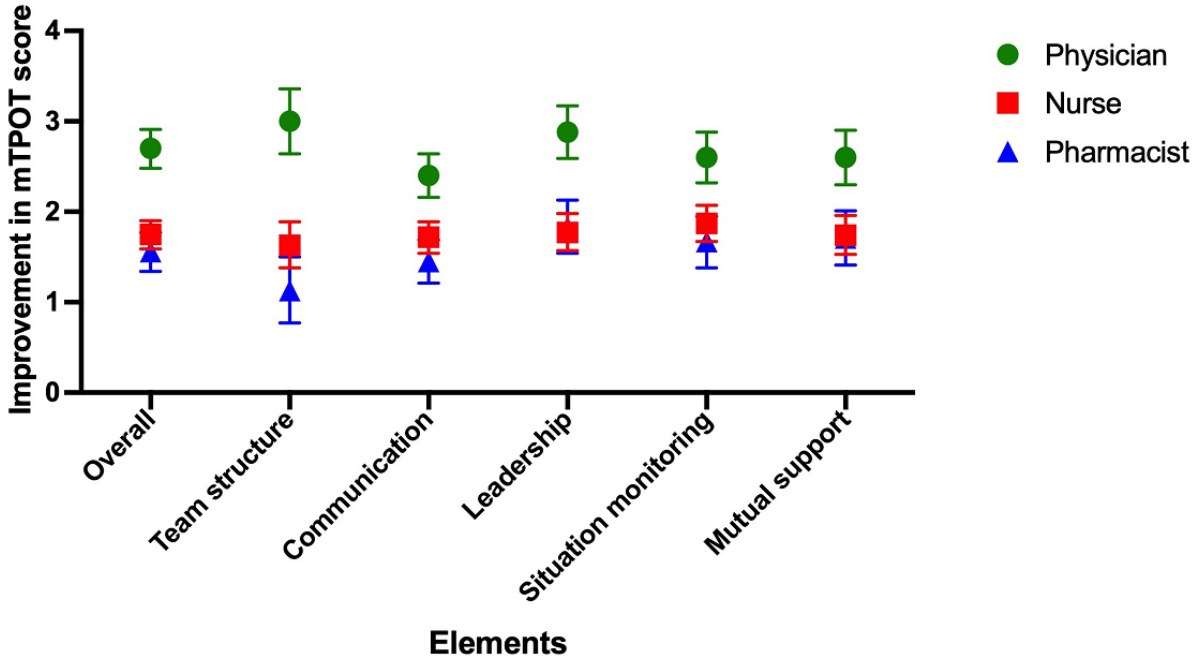
Figure 2. Pre- and post–co-debriefing mTPOT performance outcome in simulation scenario: overall changes across disciplines. mTPOT: Modified TeamSTEPPS and Team Performance Observation Tool.



Regarding individual specialties (Multimedia Appendix 4 and Figure 3), physicians show the greatest improvement in team structure (+3.00), followed by leadership (+2.88), a tie between situation monitoring and mutual support (+2.60), and the least improvement in communication (+2.40). Nurses showed the greatest improvement in situation monitoring (+1.87), followed by leadership (+1.77), mutual support (+1.74), communication (+1.72), and finally, team structure (+1.63). Pharmacists showed the greatest improvement in leadership, which was their weakest area during the baseline assessment (+1.83), followed by mutual support (+1.71), situation monitoring (+1.67), communication (+1.45), and the least improvement in team structure (+1.13).

XSL•FO
RenderX

Figure 3. Pre- and post–co-debriefing mTPOT performance outcome in simulation scenario: breakdown by health care profession. mTPOT: Modified TeamSTEPPS and Team Performance Observation Tool.



TeamSTEPPS Team Performance Outcome in the ED Workplace

When evaluating team performance in actual emergency situations using the same mTPOT, this section aims to identify any correlation between IPC improvement in simulation and in the ED. At baseline (Multimedia Appendix 3), before any intervention, pharmacists scored lowest on the mTPOT, with a score of 2.08, while physicians achieved the highest overall score of 3.15. Prior to any intervention, all specialties score lowest in leadership (with nurses tying for the lowest score in leadership and teamwork). Among the 5 domains, physicians and pharmacists performed best in team structure, while nurses received the highest scores in situational awareness.

Overall, all specialties showed a small but significant (P<.001) improvement in IPC performance in the ED workplace after the intervention (Figure 4). In contrast to the baseline score (Multimedia Appendix 5), pharmacists show the greatest improvement of +0.64, while physicians show the least, with a score difference of +0.45. The individual specialties show slightly different improvement patterns (Figure 5), with leadership being the area of greatest improvement for physicians, while communication tops the list for nurses and pharmacists. Mutual support shows the least improvement for nurses and pharmacists, while situation monitoring shows the least improvement for physicians. Of the 5 domains assessed (Multimedia Appendix 5), communication skills showed the greatest improvement (+0.77), followed by leadership and situation monitoring tied (+0.58), team structure (+0.49), and mutual support (P<.001).

XSL•FO
RenderX

Figure 4. Pre- and post–ER-VIPE mTPOT performance outcome in the emergency department workplace: overall changes across disciplines. ER-VIPE: Emergency Room Virtual Simulation-Based Interprofessional Education; mTPOT: Modified TeamSTEPPS and Team Performance Observation Tool.

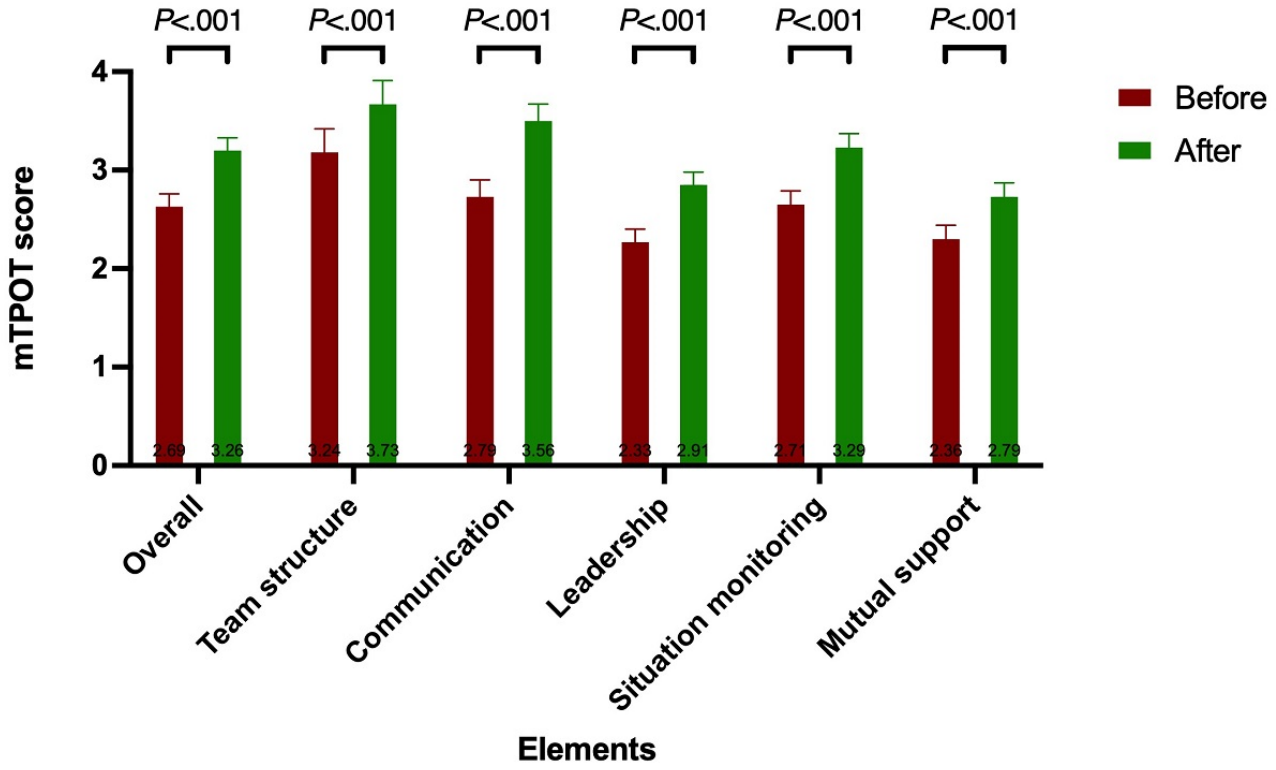
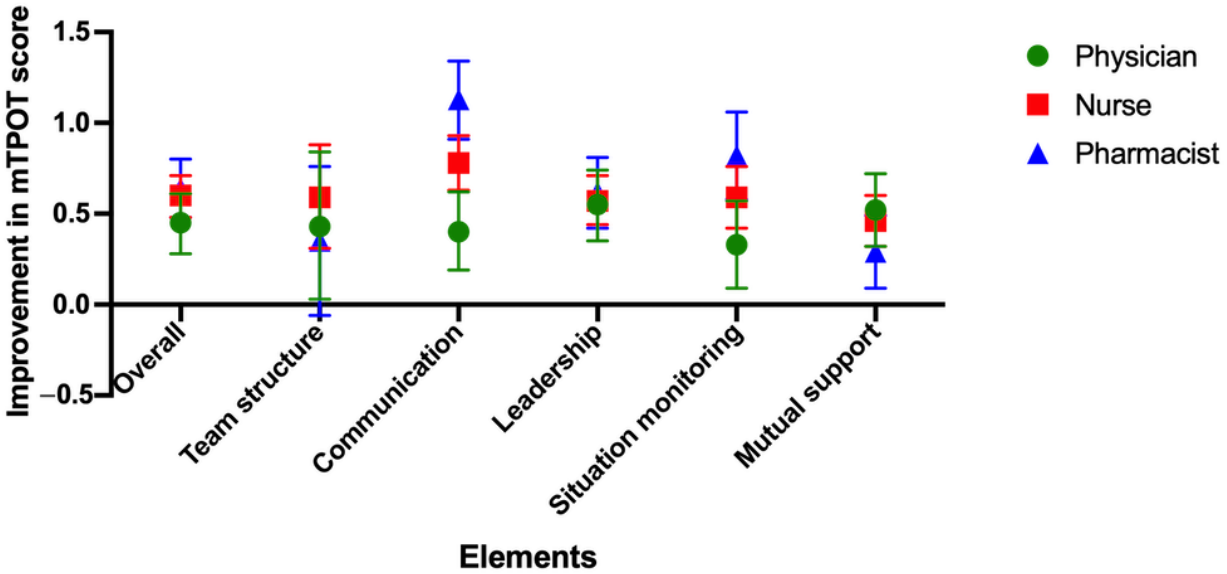


Figure 5. Pre- and post–ER-VIPE mTPOT performance outcome in the emergency department workplace: breakdown by health care profession. ER-VIPE: Emergency Room Virtual Simulation-Based Interprofessional Education; mTPOT: Modified TeamSTEPPS and Team Performance Observation Tool.



mTPOT scores in the ED were calculated based on TeamSTEPPS IPC evaluations of 2 cases by 2 independent evaluators, before and after the ER-VIPE intervention. Based on the Cohen κ coefficient [39] (Multimedia Appendix 6), each of the specialties including physician, nurse, and pharmacist demonstrates a moderate κ score of 0.53, 0.43, and 0.54, respectively, with a high percent agreement of 97.9%, 95.5%, and 87.9%, respectively.

Although the team's performance improvement pattern in actual Emergency Severity Index level 1 or level 2 emergency cases does not mirror the simulation pattern, physicians and nurses showed significant improvement across all 5 TeamSTEPPS domains in both simulated and real-world scenarios. Pharmacists showed significant improvement across most simulation domains, with only 1 domain (team structure) showing minimal improvement.

XSL•FO
RenderX

Changes in Miss and Near-Miss Medication Error Incidents in the ED

Retrospective data on medication errors were collected from the hospital database from August to October 2023 as preintervention baseline data and from January to March 2024 as postintervention data. Errors in each specialty were recorded separately as PE from physicians, AE from nurses, and DE from pharmacists. PE rate and AE rate were calculated in relation to the number of ED visits (preintervention: n=14,798 and postintervention: n=14,431), while the DE rate was calculated in relation to the number of prescriptions (preintervention: n=21,388 and postintervention: n=20,152; Table 2).

Table 2. Changes in medication error incident rate in the emergency department before (2023) and after (2024) interventions on interprofessional education.

	2023, n (%)	2024, n (%)	*P* value
Physician (PE[a])	n=14,798[b]	n=14,431[c]	
Near miss	113 (0.8)	151 (1.1)	.01[d]
No harm incident	0 (0)	0 (0)	—[e]
Harmful incident	0 (0)	1 (0)	.49[f]
Nurse (AE[g])	n=14,798[b]	n=14,431[c]	
Near miss	4 (0)	3 (0)	>.99[f]
No harm incident	2 (0)	8 (0.1)	.06[f]
Harmful incident	1 (0)	0 (0)	>.99[f]
Pharmacist (DE[h])	n=21,388[i]	n=20,152[j]	
Near miss	4 (0)	3 (0)	>.99[f]
No harm incident	9 (0)	9 (0)	>.99[f]
Harmful incident	0 (0)	0 (0)	—

[a]PE: prescription error.

[b]Number of emergency department (ED) visits (during August-October 2023).

[c]Number of ED visits (during January-March 2023).

[d]Chi-square test.

[e]Not available.

[f]Fisher exact test.

[g]AE: administration error.

[h]DE: dispensing error.

[i]Number of prescription orders from the ED to ED pharmacists (during August-October 2023).

[j]Number of prescription orders from the ED to ED pharmacists (during January-March 2023).

Without significant changes in staffing or ED policies, the number of near-miss incidents due to PE increased significantly from 0.8% to 1.1% (*P*=.01). This finding indicates an increase in the reporting of near-miss PEs that were identified and intercepted by the health care team before reaching patients. There is no significant difference in other categories of medication errors across other specialties (Table 2).

Discussion

Principal Findings

This study demonstrates a significant improvement in TeamSTEPPS performance, as measured by mTPOT scores, in both simulated and real-world ED settings. However, despite a high percentage agreement, the moderate interrater reliability observed in the real-world ED setting warrants cautious interpretation, as variability in rater judgment may have affected score consistency. Among the 5 TeamSTEPPS domains measured in the ED setting, communication showed the most substantial improvement, followed by situation monitoring and leadership. While evaluating IPC performance over a 10- to 63-day period after ER-VIPE, a consistent enhancement in teamwork behaviors was recorded by mTPOT, alongside an increase in near-miss prescribing errors, indicating improved error detection and reporting.

In-Depth Evaluation by Profession

Physician: Enhanced Leadership Skills

Effective communication with colleagues and strong leadership skills are crucial aspects of professionalism for emergency physicians [40]. For physicians, the mTPOT scores reflected a marked improvement in the leadership domain in the TeamSTEPPS framework. A key scoring criterion in this domain measures the physician's ability to consult or notify other professions, such as the pharmacist in the use of HAD for high-risk patients to reduce medication errors, and was able to

XSL•FO
RenderX

delegate tasks to others, such as pharmacists for medication preparation, medication reconciliation, allergy history collection, drug interaction monitoring, and dose adjustments based on patient factors such as weight, liver function, and kidney function. These skills were particularly beneficial in minimizing medication-related risk in high-risk patient groups, including children, pregnant women, and older patients [41].

Another scoring criterion in the leadership domain measures the use of "brief, huddle, and debrief" in team huddles. Physicians, often as leaders of the health care team, engaged in discussions with other health care professionals, ensuring that the entire team had a shared understanding of the situation, goals, and management plan. This ongoing communication helps maintain team alignment and enable real-time adaptive decision-making throughout the service process. This action also enables health care teams to more efficiently and effectively detect potential medication errors before they occur [42].

Nurses and Pharmacists: Strengthened Communication

Workplace-based evaluation of both nurses and pharmacists shows significantly improved mTPOT scores in communication skills. In ER-VIPE, TeamSTEPPS communication skills including closed-loop communication, use of verification techniques, handover techniques, and the correct use of information sources were taught to participants. The improved mTPOT scores in the workplace thus reflect that these skills were effectively enhanced in the workplace after the intervention, where pharmacists and nurses were able to communicate more effectively, with greater clarity, conciseness, and accuracy. In the team setting, the mTPOT scores also reflect that while performed minimally prior to the ER-VIPE intervention, nurses and pharmacists showed an increased ability to help the health care team in gathering information from various sources, including direct communication with patients and family members, and double-checking medication information through an external data source. The strengthened communication skills do not only lead to increased efficiency and timeliness in the completion of emergency tasks [43] but also allow the interprofessional team to support each other in gathering patients' health information, such as drug history, and acting synergistically in cross-checking on the correct use of medications to prevent medication errors [44].

Situation Monitoring Impact on Clinical Practice and Patient Safety

For nurses and pharmacists, situation monitoring is the second most improved domain in the workplace. This improvement, along with enhanced communication, has been shown to contribute to better clinical outcomes by preventing medication errors and identifying patient safety threats [45]. An example of this can be seen in the mTPOT evaluation rubrics where nurses demonstrated increased awareness of changes in patients' condition before administering medication.

After the ER-VIPE intervention, mTPOT scores of the nurses reflect an increased awareness of events where colleagues might be at risk of making unsafe decisions, such as preparing to administer medication without performing an independent double-check using the "7 Rights" of medication administration (right patient, right drug, right dose, right route, right time, right reason, and right documentation) [46]. This shared situational awareness enabled team members to intervene before the occurrence of medication errors, enhancing the overall safety of medication administration processes [47].

Shifting away from a culture of underreporting—driven by cognitive biases, frustration with limited changes, or fear of embarrassment and professional repercussions [7,48], the increase in reported near-miss PEs following ER-VIPE training reflects a growing awareness among health care staff, particularly in nurses and pharmacists, of the importance of shared situational awareness through near-miss medication error reporting. By maintaining a blame-free culture for a safe medication error reporting environment, the transparency in near-miss data can further help authorities in identifying systemic issues and address root causes of errors, leading to overall improvements in patient safety [49].

Finally, the practice of using "thinking aloud" for shared situational awareness in the TeamSTEPPS framework allows the interprofessional team to create a more cohesive understanding of the patient's condition and care strategies. By ensuring all team members were "on the same page," the teams were better equipped to prevent medication errors due to incomplete history such as drug allergies, identification of strengths and weaknesses in clinical decisions, and other potential risks, contributing to safer and more efficient patient care [50].

Implications: Enhancing ED Pharmacists' Skills and Fostering a Culture of IPC for Medication Error Reduction Through ER-VIPE

This study presents an innovative multimodal virtual simulation-based interprofessional training approach in emergency settings with TeamSTEPPS principles. It further enhances our understanding of how virtual simulation, particularly in emergency settings, plays a crucial role in developing the skills of ED pharmacists. Previous studies by Bunditanukul et al [51] have highlighted the effectiveness of simulation in fostering positive teamwork attitudes among ED pharmacists. Our findings build on this, showing that training not only improves attitudes but also significantly enhances pharmacists' performance in both simulated and real-world ED settings.

In the simulated environment, the pharmacist subgroup showed a modest improvement. We anticipated that this was a result of the limited potential magnitude of change with a relatively high baseline mTPOT score. Despite the smaller margin of improvement, their final scores remained higher than those of the nurse subgroup (Multimedia Appendix 4). Conversely, in the real-world ED observations, pharmacists demonstrated a lower baseline team performance, which provided greater room for growth. In this setting, the pharmacist group achieved a marked improvement, showing the largest gain among all professional groups (Multimedia Appendix 5).

We speculate that the higher baseline mTPOT scores in the simulation compared with the real-world workplace setting may be explained by 2 factors. First, the existing siloed work culture

XSL•FO
RenderX

within the ED may have contributed to a fixed professional mindset, limiting IPC in daily practice. Second, participants had completed the medical movie and MOOC modules prior to the baseline simulation assessment. These preparatory learning activities likely served as a supportive primer, fostering a more positive mindset toward teamwork and communication, thereby elevating their baseline performance during the simulation.

Taken together, the consistently high posttraining mTPOT scores across both simulation and real-world settings highlight the strong responsiveness of pharmacists to the ER-VIPE training. In the current clinical context—where the role of clinical pharmacists in the ED remains limited—these results underscore ER-VIPE's potential to strengthen pharmacists' competencies and support their integration into interprofessional emergency care teams.

The role of ED pharmacists is an emerging field that can greatly benefit from enhanced IPC training. This study serves as a model for expanding TeamSTEPPS IPC training applications, particularly in training clinical pharmacists to support ED physicians and core nursing teams. Emerging ED pharmacists are expected to play a key role in identifying and addressing medication discrepancies that may currently be overlooked [52,53]. ER-VIPE can be used to train them to work confidently alongside other health care professionals in the ED [16], especially since pharmacists often report a lack of experience or familiarity in the clinical setting [54].

Given the dynamic nature of clinical cases and team dynamics in emergency care, effective TeamSTEPPS IPC training should involve ongoing practice and peer feedback. This approach not only enhances individual skills but also fosters a culture of teamwork and ongoing improvement among the team. By implementing regular feedback mechanisms, we can ensure that the trained multiprofessional ED members, including pharmacists, are well-equipped to handle evolving challenges and contribute to improving patient outcomes.

Comparison With Prior Work

Compared to previous research, there is evidence supporting the link between educational interventions and improved patient outcomes, as shown in the literature. However, these studies typically used conventional simulation training methods rather than virtual approaches. For instance, studies conducted in China by Shi et al [14] have shown a positive impact of patient safety education programs and their return on investment. However, these studies did not leverage the scalability and generalizability offered by virtual simulation, which is a distinct advantage of the ER-VIPE intervention [15].

This study is one of the first to track behavioral changes in the ED setting following IPC training, rather than confining observations to classroom simulations alone. We designed the data collection process to include evaluation of multiple patient interactions per assessment point, with each expert evaluator observing more than 1 clinical encounter per study session, enhancing the robustness of our findings. Another factor that enhanced this reliability was our team's development of a tailored mTPOT tool for the Thai language context, customized to meet the needs of the respective professional groups. This tool underwent comprehensive validation, achieving a CVI above 0.8 prior to its implementation in our study.

These results not only validate the effectiveness of the ER-VIPE training as a scalable and cost-effective model for continuous professional development in health care [15] but also demonstrate its applicability as a general educational tool for improving workplace performance in the ED (Kirkpatrick Evaluation Level 3). Furthermore, these findings highlight the importance of IPC in enhancing patient safety outcomes in real-world clinical settings (Kirkpatrick Evaluation Level 4) [55,56]. This is a step further from the previous study where TeamSTEPPS training had been shown to improve team members' perception and attitudes about teamwork (Kirkpatrick Evaluation Level 2-3) [57].

The ER-VIPE training aligns with the latest TeamSTEPPS 3.0 model, launched by AHRQ in 2024, which introduces key updates such as greater patient involvement in medical decisions and enhanced active learning strategies, including discussions, online group exercises, and video-based simulations [58]. ER-VIPE builds on this new version by incorporating a local context—for example, fostering a no-blame culture in co-debriefing sessions to address the root causes of medication errors. Additionally, with local trainers, we tailored the training to cultural nuances, such as tone of voice and precise word choice, to enhance participant engagement and understanding.

Limitations

Several limitations should be acknowledged. First, variations in emergency patient cases—such as severity, urgency, and staffing—may have influenced team dynamics and communication, potentially affecting mTPOT scores regardless of the intervention.

Although a high percentage of agreement, the moderate reliability suggests that while the mTPOT provides a structured framework for assessment, future refinements could enhance consistency. Strategies such as enhanced rater training, standardized scoring protocols, or double-checking ratings with additional raters could be considered to improve agreement.

The study design also limits generalizability. The absence of randomization and a control group introduces potential self-selection bias and restricts the ability to attribute improvements solely to the intervention. Additionally, the voluntary nature of participation may have led to self-selection bias, as more motivated individuals were likely to enroll and perform better.

Due to the nature of voluntary reporting, this study could not rule out the risk of underreporting. However, to the best of our knowledge, we had observed a significant increase in reporting in the workplace of near-miss PEs, which we are hopeful that our effort has helped minimize potential harmful medication errors in the ED. In addition, the cohort of our study consisted of 62.5% (15/24) emergency medicine residents, 71.4% (5/7) ED pharmacists, and only 29.1% (30/103) ED nurses. The smaller proportion of nurses who underwent the ER-VIPE intervention may have contributed to fewer reported incidents of medication errors in both AE and DE categories, which, in

XSL•FO
RenderX

real-world practice, are often identified through nursing surveillance and reporting.

Finally, the potential influence of the Hawthorne effect should be considered. Although participants may have altered their behavior due to being observed, this effect still reflects improved awareness and understanding of appropriate teamwork practices—aligning with the principle of "fake it until you make it" in the development of professional behaviors.

Future Studies

Future research should use more rigorous designs, such as randomized controlled trials or matched control groups, to strengthen causal inferences. Longitudinal studies with repeated measures and blinding could also help minimize the Hawthorne effect and better assess sustained behavioral change and its impact on medication error reduction.

Further investigation is needed to determine the optimal frequency of retraining to maintain IPC skills and to reinforce a culture of safety and open communication. Expanding the sample size and tracking missed or near-miss events over time would provide a clearer picture of long-term impact and sustainability.

Although this study did not demonstrate a significant reduction in harmful medication errors, it revealed meaningful improvement in IPC performance, even with limited participation among ED professionals, particularly nurses (30/103, 29.1%). This finding underscores the need for future studies to increase the proportion of nurse participants and to incorporate targeted training aimed at enhancing nurses' ability to detect and report adverse and near-miss medication events. Collectively, these directions could guide the development of scalable, evidence-based IPC and TeamSTEPPS programs to strengthen teamwork, cross-disciplinary communication, and patient safety in emergency care.

Conclusions

The multimodal virtual SIMBIE tool (ER-VIPE), integrated with TeamSTEPPS, effectively improved IPC among ED residents, nurses, and pharmacists, with sustained impact up to 2 months. The increase in near-miss medication error reporting suggests heightened awareness and a stronger safety culture. This approach shows promise in promoting team communication, error reporting, and developing dedicated ED teams to address medication errors at their root causes.

Acknowledgments

The authors extend their gratitude to Chulalongkorn Healthcare Advanced Multi-Profession Simulation Center, Mr Surachai Pianpetchlert, and all research assistants: Natasha Sahapattana for their assistance in IT support in our virtual simulation laboratory and institutional review board submission process. The authors also appreciate the help of Dr Tanawat Tarapan, Jiraporn Prompila, Pataraporn Kheawwan, Chanya Thanomlikhit, Rattana Nakchan, Rung Samabhunn, Tipparat Pholin, Chawika Nareuponjirakul, Kanyarat Susantitaphong Suthida Chanchenchop, Thanapoorn Poohihunkul, Adisak Weerasaksanti, and Weasarath Chittiwan, who facilitated and assessed Team Strategies and Tools to Enhance Performance and Patient Safety performance. Finally, the authors extend their gratitude to Dr Zarina Sadad and Sunisa Seepom for her contribution to rater training and rating modified Team Strategies and Tools to Enhance Performance and Patient Safety and Team Performance Observation Tool scores. Preliminary findings were abstract presented at the International Association for Health Professions Education 2024 conference.

Funding

This project received support from the Second Century Fund, Chulalongkorn University.

Authors' Contributions

TK, OIC, PP, KM, LH, JR, KB, KT, and KN each played important roles in the project. The conceptualization was led by KN, while the methodology was developed collaboratively by KN, OIC, PP, and KM. Software development was managed by KN and JR, and validation of the work was performed by KN, OIC, JR, KB, and KM. Formal analysis was carried out by KN, PP, and LH, with the investigation being conducted by KN, TK, OIC, KM, LH, JR, KB, and KT. Resources were provided by KN, TK, JR, and KT, and data curation was handled by KN, OIC, PP, and TK. The original draft of the manuscript was written by KN, TK, OIC, PP, and LH, while all authors participated in reviewing and editing the final version, which was approved for submission. Visualization efforts were contributed by KN, OIC, PP, and LH, and the supervision of the project was overseen by KN. Project administration was the responsibility of KN and OIC, while KN also managed the acquisition of funding for the project.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Modified Team Performance Observation Tool for nurses.
[PDF File (Adobe PDF File), 164 KB - mededu_v12i1e66999_app1.pdf]

XSL•FO
RenderX

Multimedia Appendix 2
Baseline simulated scenario Modified Team Performance Observation Tool score in each health care discipline.
[DOCX File , 17 KB - mededu_v12i1e66999_app2.docx]

Multimedia Appendix 3
Baseline emergency department workplace case Modified Team Performance Observation Tool score in each health care discipline.
[DOCX File , 17 KB - mededu_v12i1e66999_app3.docx]

Multimedia Appendix 4
Simulated scenario Modified Team Performance Observation Tool score change in each health care discipline after co-debriefing session.
[DOCX File , 18 KB - mededu_v12i1e66999_app4.docx]

Multimedia Appendix 5
Emergency department workplace case Modified Team Performance Observation Tool score change in each health care discipline after Emergency Room Virtual Simulation-Based Interprofessional Education intervention.
[DOCX File , 18 KB - mededu_v12i1e66999_app5.docx]

Multimedia Appendix 6
Interrater reliability by specialty.
[DOCX File , 15 KB - mededu_v12i1e66999_app6.docx]

References

1. Framework for action on interprofessional education and collaborative practice. World Health Organization. 2010. URL: https://iris.who.int/handle/10665/70185 [accessed 2023-12-10]
2. Didier A, Dzemaili S, Perrenoud B, Campbell J, Gachoud D, Serex M, et al. Patients' perspectives on interprofessional collaboration between health care professionals during hospitalization: a qualitative systematic review. JBI Evid Synth 2020;18(6):1208-1270. [doi: 10.11124/JBISRIR-D-19-00121] [Medline: 32813373]
3. Manias E. Effects of interdisciplinary collaboration in hospitals on medication errors: an integrative review. Expert Opin Drug Saf 2018;17(3):259-275. [doi: 10.1080/14740338.2018.1424830] [Medline: 29303376]
4. IPEC core competencies for interprofessional collaborative practice: version 3. Interprofessional Education Collaborative. 2023. URL: https://www.ipecollaborative.org/2021-2023-core-competencies-revision [accessed 2023-12-10]
5. Wilson AJ, Palmer L, Levett-Jones T, Gilligan C, Outram S. Interprofessional collaborative practice for medication safety: nursing, pharmacy, and medical graduates' experiences and perspectives. J Interprof Care 2016;30(5):649-654. [doi: 10.1080/13561820.2016.1191450] [Medline: 27351385]
6. Verspuy M, Van BP. Interprofessional collaboration and communication. In: Van Bogaert P, Clarke S, editors. The Organizational Context of Nursing Practice. Cham: Springer; 2018:259-278.
7. Global burden of preventable medication-related harm in health care: a systematic review. World Health Organization. 2023. URL: https://www.who.int/publications/i/item/9789240088887 [accessed 2025-02-13]
8. Chua SS, Tea MH, Rahman MHA. An observational study of drug administration errors in a Malaysian hospital (study of drug administration errors). J Clin Pharm Ther 2009;34(2):215-223. [doi: 10.1111/j.1365-2710.2008.00997.x] [Medline: 19250142]
9. Top 10 patient safety issues Thailand 2023. The Healthcare Accreditation Institute (Public Organization). 2023. URL: https://backend.ha.or.th/fileupload/DOCUMENT/00535/fae6d157-644a-4493-83e8-a8afb34bce0a.pdf [accessed 2026-03-02]
10. Tariq RA, Vashisht R, Sinha A, Scherbak Y. Medication Dispensing Errors and Prevention. Treasure Island, FL: StatPearls Publishing; 2025.
11. Olde Bekkink M, Farrell SE, Takayesu JK. Interprofessional communication in the emergency department: residents' perceptions and implications for medical education. Int J Med Educ 2018;9:262-270 [FREE Full text] [doi: 10.5116/ijme.5bb5.c111] [Medline: 30368487]
12. Yadav S, Rawal G, Jeyaraman M. Decision fatigue in emergency medicine: an exploration of its validity. Cureus 2023;15(12):e51267 [FREE Full text] [doi: 10.7759/cureus.51267] [Medline: 38288179]
13. Team Strategies and Tools to Enhance Performance and Patient Safety (TeamSTEPPS®) 2.0. Agency for Healthcare Research and Quality. 2017. URL: https://www.ahrq.gov/teamstepps/instructor/tools.html [accessed 2022-09-17]
14. Shi Y, Miao S, Fu Y, Sun C, Wang H, Zhai X. TeamSTEPPS improves patient safety. BMJ Open Qual 2024;13(2):e002669 [FREE Full text] [doi: 10.1136/bmjoq-2023-002669] [Medline: 38670556]
15. Bogne Kamdem V, Daelemans C, Englert Y, Morin F, Sansregret A. Using simulation team training with human's factors components in obstetrics to improve patient outcome: a review of the literature. Eur J Obstet Gynecol Reprod Biol 2021;260:159-165. [doi: 10.1016/j.ejogrb.2021.03.015] [Medline: 33784580]

XSL•FO
RenderX

16. Narajeenron K, Chintakovid T, Phutrakool P, Ritsamdang J, Viriyopase A, Musikatavorn K, et al. Enhancing team strategies and tools to enhance performance and patient safety performance through medical movies, massive open online courses, and 3D virtual simulation-based interprofessional education: mixed methods double-blind quasi-experimental study. J Med Internet Res 2025;27:e67001 [FREE Full text] [doi: 10.2196/67001] [Medline: 40921059]
17. Özbay Ö, Çınar S. Effectiveness of flipped classroom teaching models in nursing education: a systematic review. Nurse Educ Today 2021;102:104922. [doi: 10.1016/j.nedt.2021.104922] [Medline: 33940481]
18. Fewster-Thuente L, Batteson TJ. Kolb's experiential learning theory as a theoretical underpinning for interprofessional education. J Allied Health 2018;47(1):3-8. [Medline: 29504014]
19. Kolb DA, Boyatzis RE, MainemelisC. Experiential learning theory: previous research and new directions. In: Sternberg RJ, Zhang LF, editors. Perspectives on Thinking, Learning, and Cognitive Styles. New York, NY: Lawrence Erlbaum Associates; 2021:227-247.
20. Aldriwesh MG, Alyousif SM, Alharbi NS. Undergraduate-level teaching and learning approaches for interprofessional education in the health professions: a systematic review. BMC Med Educ 2022;22(1):13 [FREE Full text] [doi: 10.1186/s12909-021-03073-0] [Medline: 34980083]
21. Chen H, O'Donnell JM, Chiu Y, Chen Y, Kang Y, Tuan Y, et al. Comparison of learning outcomes of interprofessional education simulation with traditional single-profession education simulation: a mixed-methods study. BMC Med Educ 2022;22(1):651 [FREE Full text] [doi: 10.1186/s12909-022-03640-z] [Medline: 36042449]
22. Costello M, Huddleston J, Atinaja-Faller J, Prelack K, Wood A, Barden J, et al. Simulation as an effective strategy for interprofessional education. Clin Simul Nurs 2017;13(12):624-627 [FREE Full text] [doi: 10.1016/j.ecns.2017.07.008]
23. Mahmood LS, Mohammed CA, Gilbert JH. Interprofessional simulation education to enhance teamwork and communication skills among medical and nursing undergraduates using the TeamSTEPPS® framework. Med J Armed Forces India 2021;77(Suppl 1):S42-S48 [FREE Full text] [doi: 10.1016/j.mjafi.2020.10.026] [Medline: 33612931]
24. Samarasekera DD, Nyoni CN, Amaral E, Grant J. Challenges and opportunities in interprofessional education and practice. Lancet 2022;400(10362):1495-1497 [FREE Full text] [doi: 10.1016/S0140-6736(22)02086-4] [Medline: 36522199]
25. Bandura A, Walters RH. Social Learning Theory. Englewood Cliffs, NJ: Prentice Hall; 1977.
26. Salajegheh M, Sohrabpour AA, Mohammadi E. Exploring medical students' perceptions of empathy after cinemeducation based on Vygotsky's theory. BMC Med Educ 2024;24(1):94 [FREE Full text] [doi: 10.1186/s12909-024-05084-z] [Medline: 38287370]
27. Kolb DA. Experiential Learning: Experience as the Source of Learning and Development. New Jersey: FT Press; 2014.
28. Adams P. Exploring social constructivism: theories and practicalities. Education 3-13 2006;34(3):243-257. [doi: 10.1080/03004270600898893]
29. Mezirow J. Learning as Transformation: Critical Perspectives on a Theory in Progress. In: The Jossey-Bass Higher and Adult Education Series. Washington, DC: ERIC; 2000.
30. Mezirow J. How critical reflection triggers transformative learning. In: Fostering Critical Reflection in Adulthood. New York, NY: Jossey-Bass Publishers; 1990:1-20.
31. Vivian R, Falkner K, Falkner N. Addressing the challenges of a new digital technologies curriculum: MOOCs as a scalable solution for teacher professional development. Res Learn Technol 2014;22:24691. [doi: 10.3402/rlt.v22.24691]
32. Kidzinski L, Sharma K, Boroujeni MS, Dillenbourg P. On generalizability of MOOC models. 2016 Presented at: Proceedings of the 9th International Conference on Educational Data Mining; June 29-July 2, 2016; Raleigh, NC, United States URL: https://www.educationaldatamining.org/EDM2016/proceedings/paper_85.pdf
33. Dhanakoses K, Pavavimol T, Issarasak S. The codesign process of virtual simulation games in medical education: a case study of the ER-VIPE platform. CoDesign 2025;21(4):745-764. [doi: 10.1080/15710882.2025.2515222]
34. Wu Q, Wang Y, Lu L, Chen Y, Long H, Wang J. Virtual simulation in undergraduate medical education: a scoping review of recent practice. Front Med (Lausanne) 2022;9:855403 [FREE Full text] [doi: 10.3389/fmed.2022.855403] [Medline: 35433717]
35. Cheng A, Palaganas J, Eppich W, Rudolph J, Robinson T, Grant V. Co-debriefing for simulation-based education: a primer for facilitators. Simul Healthc 2015;10(2):69-75. [doi: 10.1097/SIH.0000000000000077] [Medline: 25710318]
36. Levine AI, DeMaria S, Schwartz AD, Sim AJ. The Comprehensive Textbook of Healthcare Simulation. New York, NY: Springer; 2013.
37. Kheawwan P, Thanomlikhit C, Narajeeenron K, Rojnawee S. Translation and psychometric validation of the Thai version of TeamSTEPPS® Team Performance Observation Tool. J Interprof Care 2024;38(3):573-582 [FREE Full text] [doi: 10.1080/13561820.2024.2307547] [Medline: 38343289]
38. Downing SM. Validity: on meaningful interpretation of assessment data. Med Educ 2003;37(9):830-837. [doi: 10.1046/j.1365-2923.2003.01594.x] [Medline: 14506816]
39. Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics 1977;33(1):159-174. [Medline: 843571]
40. Narajeenron K, Tarapan T, Dukovic J, Sutham K, Imsuwan I, Dadeh A, et al. Perceptions of emergency physician professionalism among healthcare providers and patients: a multicenter study in thailand and the US. Chula Med J 2024;68(2):67-77. [doi: 10.56808/2673-060x.5396]

XSL•FO
RenderX

41. tefan CS, Nechita A, Dragostin OM, Fulga A, Lisă E, Vatcu R, et al. Drugs associated with adverse effects in vulnerable groups of patients. Clin Pract 2024;14(3):1010-1020 [FREE Full text] [doi: 10.3390/clinpract14030080] [Medline: 38921258]
42. Lin SP, Chang CW, Wu CY, Chin CS, Lin CH, Shiu SI, et al. The effectiveness of multidisciplinary team huddles in healthcare hospital-based setting. J Multidiscip Healthc 2022;15:2241-2247 [FREE Full text] [doi: 10.2147/JMDH.S384554] [Medline: 36225857]
43. El-Shafy IA, Delgado J, Akerman M, Bullaro F, Christopherson NAM, Prince JM. Closed-loop communication improves task completion in pediatric trauma resuscitation. J Surg Educ 2018;75(1):58-64. [doi: 10.1016/j.jsurg.2017.06.025] [Medline: 28780315]
44. Fitzgerald RJ. Medication errors: the importance of an accurate drug history. Br J Clin Pharmacol 2009;67(6):671-675 [FREE Full text] [doi: 10.1111/j.1365-2125.2009.03424.x] [Medline: 19594536]
45. Avalos J, Roy D, Asan O, Zhang Y. The influential factors on nurses' situation awareness in inpatient settings: a literature review. Hum Factors Healthcare 2021;1:100006. [doi: 10.1016/j.hfh.2022.100006]
46. Smeulers M, Verweij L, Maaskant JM, de Boer M, Krediet CTP, Nieveen van Dijkum EJM, et al. Quality indicators for safe medication preparation and administration: a systematic review. PLoS One 2015;10(4):e0122695 [FREE Full text] [doi: 10.1371/journal.pone.0122695] [Medline: 25884623]
47. Sitterding MC, Ebright P, Broome M, Patterson ES, Wuchner S. Situation awareness and interruption handling during medication administration. West J Nurs Res 2014;36(7):891-916. [doi: 10.1177/0193945914533426] [Medline: 24823968]
48. Lee J. Understanding nurses' experiences with near-miss error reporting omissions in large hospitals. Nurs Open 2021;8(5):2696-2704 [FREE Full text] [doi: 10.1002/nop2.827] [Medline: 33655710]
49. Sheikhtaheri A. Near misses and their importance for improving patient safety. Iran J Public Health 2014;43(6):853-854 [FREE Full text] [Medline: 26110160]
50. Siddiqui S. 'Think-aloud' protocol for ICU rounds: an assessment of information assimilation and rational thinking among trainees. Med Educ Online 2014;19:25783 [FREE Full text] [doi: 10.3402/meo.v19.25783] [Medline: 25326045]
51. Bunditanukul K, Narajeenron K, Worasilchai N, Saepow S, Nontakityothin N, Ritsamdang J. Evaluating pharmacy students' teamwork attitudes in virtual COVID-19 emergency department simulations: a pilot study. Am J Pharm Educ 2024;88(7):100716. [doi: 10.1016/j.ajpe.2024.100716] [Medline: 38729614]
52. Almazrou S, Alnaim L, Al-Kofide H. Perceptions, expectations and barriers of physicians towards working with clinical pharmacists in Saudi Arabia. J Sci Res Rep 2015;6(5):404-415. [doi: 10.9734/jsrr/2015/11637]
53. Judkins S. Clear guidance on improving care in emergency departments: clinical pharmacists as part of the team. J Pharm Pract Res 2019;49(6):504-505. [doi: 10.1002/jppr.1630]
54. Girvin B, Wilson D. Clinical skills training for pharmacists in general practice. Prescriber 2018;29(12):19-25. [doi: 10.1002/psb.1724]
55. Alsabri M, Boudi Z, Lauque D, Dias RD, Whelan JS, Östlundh L, et al. Impact of teamwork and communication training interventions on safety culture and patient safety in emergency departments: a systematic review. J Patient Saf 2022;18(1):e351-e361. [doi: 10.1097/PTS.0000000000000782] [Medline: 33890752]
56. Mohsen MM, Gab Allah AR, Amer NA, Rashed AB, Shokr EA. Team strategies and tools to enhance performance and patient safety at primary healthcare units: effect on patients' outcomes. Nurs Forum 2021;56(4):849-859. [doi: 10.1111/nuf.12627] [Medline: 34196002]
57. Obenrader C, Broome ME, Yap TL, Jamison F. Changing team member perceptions by implementing TeamSTEPPS in an emergency department. J Emerg Nurs 2019;45(1):31-37. [doi: 10.1016/j.jen.2018.08.006] [Medline: 30322675]
58. TeamSTEPPS 3.0. Agency for Healthcare Research and Quality. 2025. URL: https://www.ahrq.gov/teamstepps-program/index.html [accessed 2025-03-20]

Abbreviations

ADR: adverse drug reaction
AE: administration error
AHRQ: Agency for Healthcare Research and Quality
CVI: Content Validity Index
DE: dispensing error
ED: emergency department
ER-VIPE: Emergency Room Virtual Simulation-Based Interprofessional Education
HAD: high-alert drug
IPC: interprofessional collaboration
IPEC: Interprofessional Education Collaborative
MOOC: massive open online course
mTPOT: Modified TeamSTEPPS and Team Performance Observation Tool
PE: prescription error
RM: Risk Management

XSL•FO
RenderX

SIMBIE: simulation-based interprofessional education
TeamSTEPPS: Team Strategies and Tools to Enhance Performance and Patient Safety
TPOT: Team Performance Observation Tool

Edited by B Lesselroth; submitted 30.Sep.2024; peer-reviewed by KL Chen, LM Mazur; comments to author 27.Jan.2025; revised version received 12.Jan.2026; accepted 02.Feb.2026; published 13.Mar.2026.

Please cite as:
Chu OI, Phutrakool P, Musikatavorn K, Kongchim T, Herabat L, Ritsamdang J, Bunditanukul K, Triamamornwooth K, Narajeenron K
Enhancing Interprofessional Team Performance to Prevent Medication Errors in Emergency Care: Quasi-Experimental Study Using Multimodal Virtual Simulation-Based Interprofessional Education
JMIR Med Educ 2026;12:e66999
URL: https://mededu.jmir.org/2026/1/e66999
doi:10.2196/66999
PMID:

©Ora-In Chu, Phanupong Phutrakool, Khrongwong Musikatavorn, Thitiporn Kongchim, Lapol Herabat, Jiraphan Ritsamdang, Krittin Bunditanukul, Kanittha Triamamornwooth, Khuansiri Narajeenron. Originally published in JMIR Medical Education (https://mededu.jmir.org), 13.Mar.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Education, is properly cited. The complete bibliographic information, a link to the original publication on https://mededu.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Development and Initial Evaluation of Specific Immersive Competence in Virtual Reality–Based Medical Assessments: Exploratory Observational Study

Verena Schreiner[1], MD; Joy Backhaus[1], MSc; Marco Lindner[1], MD; Melina Heinisch[2], MSc; Sarah König[1], MD, MME; Sebastian Oberdörfer[2], PhD; Tobias Mühling[1], MD

[1]Institute of Medical Teaching and Medical Education Research, University Hospital Würzburg, Würzburg, Germany

[2]Chair for Human-Computer Interaction, Julius Maximilians University of Würzburg, Würzburg, Germany

Corresponding Author:
Tobias Mühling, MD
Institute of Medical Teaching and Medical Education Research
University Hospital Würzburg
Joseph-Schneider-Strasse 2
Würzburg, 97080
Germany
Phone: 49 931 201 70627
Email: muehling_t@ukw.de

Abstract

Background: Virtual reality (VR) is increasingly used in medical education for training and examination purposes; yet, learners' performance in VR-based assessments may be influenced by more than clinical competence alone. Immersive competence (IC) has been proposed as a relevant factor in VR-based performance assessment. While general IC captures application-independent VR interaction skills, domain-specific applications may require an additional construct, specific immersive competence (specific IC), reflecting context-dependent interaction proficiency.

Objective: This study aimed to develop and examine the initial psychometric characteristics of a newly developed in situ checklist for assessing specific IC, to explore the relationship between specific and general IC as well as related human abilities and characteristics, and to capture preliminary associations between both IC measures and medical performance in a VR-based objective structured clinical examination (OSCE) station.

Methods: In this observational follow-up study, 21 final-year medical students who had previously completed a curricular OSCE including a VR-based emergency medicine station were recruited. General IC was assessed using the VR competence app, and specific IC using a checklist embedded in the original VR simulation. Additional measures included self-reported technological affinity, spatial ability, and OSCE performance scores. Analyses focused on descriptive statistics and exploratory associations, including item difficulty, item-total correlations, and internal consistency for the specific IC checklist. Pearson *r* was used to examine associations among variables.

Results: The final 13-item specific IC checklist demonstrated acceptable internal consistency for an early-stage instrument (Cronbach α=0.79) and balanced item difficulty (mean P 0.56, SD 0.28). Specific IC showed a strong exploratory association with general IC (r=0.56; P=.008) and with prior 3D application experience (r=0.57; P=.007), but no relevant association with spatial ability. Both general and specific IC showed borderline, moderate associations with VR-OSCE performance (r=0.41; P=.06 and r=0.37; P=.09, respectively), while neither was related to overall analog OSCE performance.

Conclusions: In this pilot sample, specific IC emerged as a psychometrically accessible construct aimed to capture context-sensitive VR interaction skills. The proposed in situ approach offers a feasible and scalable method to assess this construct within domain-specific VR applications. Although associations with medical performance were exploratory and limited by sample size, the findings suggest the relevance of IC as a potential source of construct-irrelevant variance in VR-based assessments and support further investigation in larger, confirmatory studies.

(JMIR Med Educ 2026;12:e82136) doi:10.2196/82136

XSL•FO
RenderX

KEYWORDS

virtual reality; immersive competence; OSCE; objective structured clinical examinations; medical education; performance assessment; exploratory study; pilot study

Introduction

Virtual reality (VR) is increasingly being used in medical education, providing students with immersive and interactive environments that support skill acquisition, deliberate practice, and feedback-driven learning [1]. VR-based training formats can support the development of both technical skills, such as those required in laparoscopic surgery [2], as well as procedural competencies, including clinical reasoning [3], teamwork [4], and the management of rarely encountered scenarios such as medical emergencies [5] or brain death diagnostics [6]. Beyond its role in simulation-based training, VR is progressively being integrated into performance assessment, offering authentic and interactive scenarios for the structured evaluation of clinical competencies [7-9]. A growing number of studies have explored the use of VR for high-stakes evaluations, such as virtual objective structured clinical examinations (OSCEs), in which learners' clinical decisions and actions are assessed within predefined clinical scenarios in an immersive environment [10-13]. Such formats can facilitate precise logging of user behavior, consistent scenario delivery across candidates, and automated scoring options. Moreover, VR-based assessments suggest advantages in terms of scalability, objectivity, and ability to replicate complex clinical environments that are challenging or resource-intensive to recreate in conventional OSCE settings [8,10-13].

As the boundaries between learning and testing environments become increasingly fluid, the question emerges as to whether learners' performance in VR accurately reflects their actual clinical competence, or whether it is confounded by interface-related factors. This concern is reinforced by the fact that VR users must adapt to abstract and often nonstandardized metaphorical representations [14], such as administering medication via 2D menus or receiving feedback through controller vibrations. Even more recent approaches using hand-tracking rely on abstracted gestures that must be learned and mastered by the user [11]. Steed et al [15] introduced the concept of immersive competence (IC), defined as a user's familiarity and skill with immersive controls. IC thus represents the user readiness required for interaction tasks in VR applications, the absence of which can lead to difficulties with controls, interactions, or navigation within the VR environment. IC increases with prior VR experience [16,17] but also appears to be trainable [18], thereby reducing cognitive load and enhancing multitasking performance in VR environments [18]. From a validity perspective, IC may constitute a source of construct-irrelevant variance [19], as it can influence performance in VR-based assessments without being part of the intended construct of clinical competence. Indeed, marked differences in item discrimination parameters between VR-based and content-equivalent analog OSCE stations already suggested a systematic bias in VR-based examinations, potentially driven by differences in IC [12]. To capture the construct of IC, based on previous approaches [20,21], a novel assessment tool was developed: the VR competence app [17]. This application presents users with a range of prototypical interactions commonly required in immersive environments, such as picking up and placing objects, spatial navigation, and the activation of user interface (UI) elements. Importantly, it is content-agnostic and does not reference any specific educational or clinical scenario. As such, it aims to provide a standardized and domain-independent estimation of a user's general interactional proficiency in VR (general IC).

However, VR interfaces vary considerably across applications and hardware platforms [22]. This heterogeneity can be confusing for users and hinder the effective implementation of this technology [23,24]. Differences include customized controller mappings, unique interface layouts, and spatial or logistical features specific to a simulation [22,25]. For instance, general familiarity with UI activation in VR environments does not necessarily translate into proficiency with context-specific tasks such as adjusting infusion rates or navigating electronic menus in a VR-based emergency room simulation. The importance of distinguishing between general (medium-related) and specific (content- or task-related) operational competences has been highlighted by studies from other fields: For example, general internet skills like navigating websites or conducting basic research have been shown to be a poor predictor of more specific internet skills such as selecting appropriate websites for information retrieval or critically evaluating information sources [26]. Similarly, principles derived from the acquisition of simple motor tasks often fail to generalize to complex, context-rich skill environments [27]. While usability measures focus on the system's design quality [28] and general IC captures abstract interaction skills [17], neither adequately reflects a user's ability to operate a specific VR interface under real-world constraints. Technological and digital literacy are typically self-reported and too general to capture the specific, context-bound interaction skills required in VR-based assessments [29,30]. This underscores the need for a complementary measure that reflects application-specific motor, cognitive, and interface-related skills. Here, we introduce the concept of Specific Immersive Competence (specific IC)—a context-sensitive construct that reflects a user's ability to perform interactional tasks within a particular, domain-relevant VR application. Specific IC may ideally be measured directly in situ, that is, directly within the target application, for example, by using guided, nonclinical tasks designed to isolate interactional skills from domain-specific knowledge. This approach aims to enhance contextual congruence between training, task, and testing environments, in line with best practices in simulation-based assessment [31,32]. In the implementation presented here, users are guided through tasks composed of exemplary interaction steps (eg, connecting an IV line or confirming a medication dose via UI). Task completion within a given time frame is logged using an objective checklist format. This enables a targeted assessment of users' interactional readiness for high-stakes VR-based training and assessments.

XSL•FO
RenderX

To explore the relationship between specific and general IC and their association with students' performance, as well as to gather initial validity evidence for the specific IC construct, we conducted a follow-up study within an existing curricular assessment framework. Specifically, we recruited participants from a cohort of advanced clinical-phase medical students who had previously completed a VR-based station as part of their OSCE [12]. This post hoc recruitment strategy allowed us to link newly collected data on both general and specific IC with previously obtained medical performance scores—both for the VR station and for the OSCE as a whole. In doing so, this pilot study pursued three objectives: (1) we first assessed the initial psychometric characteristics of the newly developed in situ checklist for assessing specific IC, including item difficulty, discrimination (via item-total correlations), and internal consistency; (2) second, we explored patterns of association between specific and general IC as well as established human abilities and characteristics to explore whether both measures capture potentially distinct facets of VR interaction competence; and (3) finally, we analyzed the preliminary associations between both IC measures and students' medical performance in a VR-based OSCE station as well as the overall OSCE score, to provide initial evidence of IC's potential impact on VR-based performance assessment.

Methods

Apparatus

The VR competence app (developed by the Chair for Human-Computer-Interaction, University of Wuerzburg, Germany) was used for the assessment of general IC. STEP-VR (Simulation-based Training of Medical Emergencies for Physicians Using Virtual Reality; version 0.11b; ThreeDee GmbH) was used as a VR simulation of complex emergencies for in situ assessment of specific IC [33]. The application places users in a virtual emergency department, where they are required to diagnose and treat patients using a variety of medical procedures. STEP-VR is a joint development between a 3D visualization company and the University Hospital of Würzburg and had previously been used for the VR-OSCE station [12].

The hardware setup consisted of a Schenker XMG Core 15 Laptop (Schenker Technologies GmbH; chipset: Intel Core i7-9750H, 6×2.6 GHz; graphics adapter: Nvidia GeForce GTX 1650, 4 GB GDDR6 VRAM) and a Meta Quest III VR head-mounted display. Corresponding Meta Quest III controllers were used, which provide standard buttons, thumbsticks, and haptic feedback.

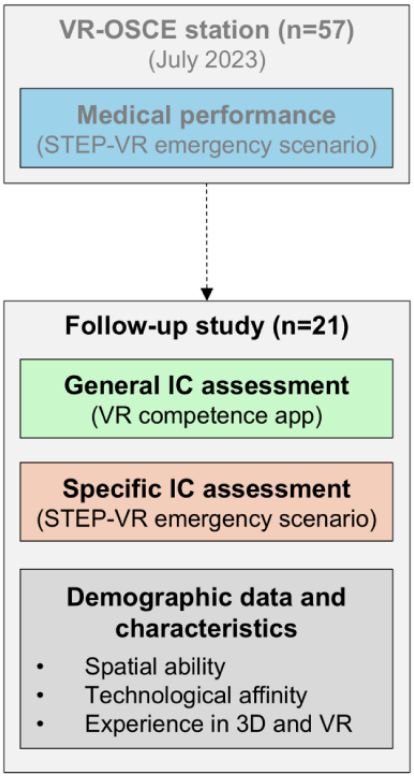
Study Design

This exploratory observational follow-up study was designed to develop and operationalize the construct of specific IC and to generate preliminary psychometric and correlational evidence. The study was conducted at the skills lab of a German medical school with a 6-year medicine degree curriculum between July 2024 and May 2025. Participants were retrospectively recruited from a cohort of advanced clinical-phase medical students who had completed a curricular OSCE in July 2023, including a VR-based emergency medicine station (VR-OSCE station). Recruitment was through semester mailing lists. The study appointments took place outside regular teaching hours in the form of individual sessions. To examine the relationship between general and specific IC and students' medical performance, we performed individual post hoc sessions with students who had previously completed the VR-OSCE station. As the study aimed to recruit the complete eligible cohort of students who had previously participated in the VR-based OSCE, a formal a priori sample size calculation was not performed, as it was determined by participant availability. However, at the time of the study, participants were in their final year of medical training and thus dispersed across various clinical rotations and institutions, making recontact and scheduling particularly challenging. From the original cohort of 57 students, a total of 21 (37%) could be recruited for the follow-up. As in the original VR-OSCE study, no a priori exclusion criteria were applied, and all enrolled participants completed the study protocol in full.

Each session lasted approximately 30 minutes. After the study procedure had been explained and informed consent was given, participants received a short basic introduction to the operation of the VR headset and controllers. Following, general IC was assessed using the VR competence app (10 minutes). After an organizational break of approximately 2 minutes, the specific IC was tested using the domain-specific application (6 minutes), which was already used during the original VR-OSCE station. At last, participants completed a questionnaire assessing demographic data (including age, gender, and cumulative experience with 3D and VR applications excluding the previously conducted VR-OSCE), technological affinity, and spatial ability (10 minutes). Figure 1 illustrates the study procedure and data collection.

XSL•FO
RenderX

Figure 1. Study Procedure and Data Collection. This study was conducted with participants who completed the virtual reality–objective structured clinical examination station in July 2023. IC: immersive competence; OSCE: objective structured clinical examination; VR: virtual reality.



General IC Assessment

General IC was measured essentially as previously described [17] using the VR competence app. This app comprises 8 standardized challenges reflecting common VR interactions from the categories of object manipulation, spatial navigation, and UI activation. For 1 minute, the user is asked to complete as many repetitions of the respective interaction as possible while task difficulty increases. Notably, this application is content-agnostic and independent of any specific educational or clinical scenario. Each task was scored automatically and summed into a total IC score. A visual example is provided in Figure 2.

XSL•FO
RenderX

Figure 2. Exemplary interaction steps illustrating general and specific immersive competence domains across different levels of abstraction. For each domain, abstract interaction tasks from the general immersive competence assessment (virtual reality competence app) are displayed above their concrete, domain-specific implementation within STEP-VR (virtual emergency department). A/D: Object manipulation (specific immersive competence: picking up blood culture bottles). B/E: Navigation (specific immersive competence: teleporting to the nurse station). C/F: User interface activation (specific immersive competence: selecting diagnostics and consults in the virtual computer menu). IC: immersive competence; UI: user interface.



Specific IC Assessment

Specific IC was assessed using a checklist-based performance measure embedded within the VR emergency room scenario previously used for summative OSCE assessment. This approach ensured high ecological validity and close alignment between IC measurement and the application-specific interaction demands encountered in high-stakes VR-based settings. To isolate immersive competence from medical knowledge, all tasks were deliberately designed to be independent of clinical reasoning. Participants received standardized verbal instructions (eg, “Measure oxygen saturation”), which were decomposed into discrete, technically feasible interaction steps (eg, “Pick up and attach SpO_2 clip” and “Read out SpO_2 value”). Each step was classified as object manipulation, spatial navigation, or UI activation, in accordance with established VR interaction typologies [16] (Figure 2). Steps involving more than one category were assigned to all relevant categories for analytical purposes.

Initially, a pool of 20 interaction steps was defined based on expert walkthroughs of the VR scenario and a systematic mapping to the application’s core control logic. Steps were selected to reflect a range of difficulty levels and to ensure coverage of all three interaction categories. This initial item pool (Multimedia Appendix 1) was reviewed by one VR developer (SO), one clinician with specialized qualification in medical education (SK), and one clinician experienced in simulation-based training (TM) to confirm technical feasibility and independence from medical knowledge. During pilot testing with five medical students recruited from the skills lab tutor pool (seventh to ninth semester, without specific prior VR experience and not part of the VR OSCE cohort), 6 steps were consistently completed by all participants and thus showed no variance. These were excluded from further analysis, as they lacked discrimination. One additional item (selecting “Surgical Consult”) showed a negative item-total correlation (r=–0.06) in our data, likely due to a software usability issue in menu navigation that was identified retrospectively and was therefore removed. The final IC checklist included 13 dichotomously scored steps (0=incorrect or incomplete; 1=correctly completed within time limit), with a maximum raw score of 13 points (Table 1). Time limits for each step were calibrated based on internal pilot runs to reflect escalating difficulty, ensuring that early steps were accessible for users with low IC, while later steps were aimed to challenge higher proficiency levels.

XSL•FO
RenderX

Table 1. Tasks, instructions, and individual interaction steps with their corresponding time limits, assigned interaction categories.

Task	Instruction	Time (s)	Step	Interaction step	Interaction category	P[a]	r'[b]
Measuring SpO_2[c]	Take the SpO_2 clip from the desk and attach it to the patient`s right hand. Then read the SpO_2 value from the vital signs monitor aloud.	10					
			1	Attach SpO_2 clip to patient	O[d], N[e]	0.71	0.14
			2	Read out SpO_2 value	U[f]	0.52	0.31
Administering saline	Take an IV catheter out of the box in the labelled drawer and place it on the patient`s arm. Then pick up a 1 liter IV fluid bag from the drawer and hang it on the IV pole. Increase the rate of the infusion to 1000 ml per hour.	30					
			3	Pick up IV[g] catheter	O, N	0.90	0.29
			4	Pick up IV fluid bag	O, N	0.90	0.22
			5	Set desired infusion rate	U	0.48	0.48
Taking blood cultures	Take a blood culture bottle from the "Microbiology" drawer and connect it to a cannula. Use this blood culture bottle to draw blood from one of the patient's arms. Then place it on the desk.	15					
			6	Pick up blood culture bottle	O, N	0.90	0.14
			7	Connect bottles to cannula	O	0.57	0.48
			8	Draw blood	O	0.43	0.54
			9	Place filled bottles on desk	N	0.05	0.23
Administering oral medication	Take a box of clopidogrel from the "oral medication" drawer. Drop the medication onto the patient's mouth.	10					
			10	Pick up oral medication	O, N	0.86	0.22
			11	Drop medication on patient's mouth	N	0.48	0.55
Administering IV medication	Take a box of aspirin from the "IV medication" drawer. Fill it into a 50ml bottle and confirm the dose. Then hang the infusion on the IV pole.	15					
			12	Prepare IV push and confirm dose	U	0.33	0.59
			13	Hang infusion on IV pole	O, N	0.19	0.60

[a]P: resulting difficulty; mean 0.56, SD 0.28.

[b]r':discrimination values; mean 0.37, SD 0.17.

[c]SpO_2: peripheral capillary oxygen saturation.

[d]O: object manipulation.

[e]N: navigation.

[f]U: user interface activation.

[g]IV: intravenous.

XSL•FO
RenderX

Technological Affinity

Participants' attitudes toward and confidence with technology were assessed using 2 subscales of the validated German Affinity for Technology- Electronic Gadgets (TA-EG) questionnaire [30], namely "enthusiasm" (eg, "I enjoy trying out new technical devices") and "competence" (eg, "I am confident using complex devices without help"). Each scale comprised 5 items rated on a 5-point Likert scale (1=strongly disagree to 5=strongly agree). Scores were averaged separately for each scale.

Spatial Ability

Spatial ability was assessed via a German version of a paper-based mental rotation task used in psychometric aptitude testing [34] following a similar principle as the original version [35]. Participants were shown 10 3D tube-shaped figures and asked to identify matching rotated versions from four options. Each correct response was awarded one point; thus, scores ranged from 0 to 10.

Medical Performance

Students' medical performance data were extracted from the original OSCE scores [12]. Two performance indicators were analyzed: The VR-OSCE station score (range 0-100%) and the overall OSCE score (calculated as the mean of all 9 other OSCE stations, range 0-100%).

Statistical Analysis

Given the exploratory nature and limited sample size, analyses focused on descriptive statistics, effect size patterns, and correlation coefficients rather than confirmatory hypothesis testing. Mean (SDs) and difficulties (P) as well as item discrimination (r'), were used to characterize the sample and items of the specific IC assessment. Given the small sample size, which precluded common factor analysis, we followed suggestions to enhance the validation of education instruments [36]. Item discrimination (r') representing item-total correlation for the dichotomous items of the specific IC assessment was calculated using point-biserial correlations.

Normality of data distributions was initially assessed using the D'Agostino–Pearson test, which did not indicate relevant deviations from normality. Accordingly, Pearson correlation coefficients are reported. In response to reviewer feedback, additional sensitivity analyses using Spearman rank correlations were conducted to assess the robustness of the findings against potential violations of normality assumptions.

Point-biserial correlations were calculated in R (version 4.3.1; RRID:SCR_001905; R Core Team), and all other statistical analyses were conducted in GraphPad Prism (Version 10.5.0, RRID:SCR_002798; GraphPad Software Inc). A significance level of α=.05 was used throughout.

Ethical Considerations

Ethics Approval

The local institutional review and ethics board (Ethikkommission der Universität Würzburg) judged the project as not to constitute medical or epidemiological research on human participants and as such adopted a simplified assessment protocol. The project was approved without any reservation under the proposal number 20240208-01.

Transparency Statement

The authors confirm that all measures, conditions, and data exclusions associated with this study have been fully reported in the manuscript and the supplementary data (Multimedia Appendices 1-Multimedia Appendices 3). No additional experimental conditions or outcome measures were collected, but not reported.

Informed Consent and Compensation Details

Students were informed about the study, and their participation was voluntary. Informed consent was obtained from all participants, who were also provided with information on data processing for the analysis and the publication of results. Contact details were supplied for participants wishing to withdraw their consent to data processing. The decision to participate or not had no consequences on the students' academic progress. A €30 (US $33.96) book voucher was given to the participants at the end of the study appointment.

Privacy and Confidentiality

Data were stored in pseudonymized form on university servers with restricted access, and no personal identifiers were retained beyond the matching of OSCE ID codes. Survey data from the questionnaires were collected anonymously using the EvaSys platform (Evasys GmbH). Data were processed and stored in accordance with local data protection laws.

Results

Descriptive Statistics

The final sample included 21 medical students (age: mean 26.5, SD 2.6 years), with 10 identifying as female and 11 as male. Both exposure to 3D applications and prior experience with VR applications apart from the previously conducted VR OSCE were very limited (Table 2). No symptoms of simulation sickness were reported, and all participants were able to complete the study session in full. In reference to the scale midpoint (=3), participants reported moderate levels of technology enthusiasm (mean 2.38, SD 0.87) and above-average self-rated technology competence (mean 3.46, SD 0.68, both on a 5-point Likert scale) on the TA-EG. Spatial ability, measured via a standardized mental rotation test, showed a mean score of 5.48 (SD 2.29) out of 10. General immersive competence, assessed via the VR competence app, yielded a mean performance score of 57.27 (SD 7.38), while specific immersive competence, measured in the VR emergency room simulation, showed a mean of 7.33 (SD 2.62) out of 13. Medical task performance among participants in the original VR OSCE station averaged at 67.73% (SD 14.62), and the overall OSCE score across all stations was mean 80.30% (SD 4.83).

XSL•FO
RenderX

Table 2. Participant characteristics, including age, gender, and experience with 3D and virtual reality applications (total n=21).

Characteristics	Values
Age (years), mean (SD)	26.5 (2.6)
Sex, n (%)	
Female	10 (48)
Male	11 (52)
Diverse	0 (0)
Frequency of using desktop 3D-applications, n (%)	
Never	14 (67)
Rarely (<1/month)	4 (19)
Sometimes (<1/week)	3 (14)
Frequently (>1/week)	0 (0)
Daily	0 (0)
Cumulative experience with VR [a] applications, n (%)	
None	4 (19)
0-1 hours	11 (52)
1-10 hours	6 (29)
10-50 hours	0 (0)
>50 hours	0 (0)

[a]VR: virtual reality.

Performance Checklist and Psychometric Properties

Overall, item completion rates ranged from 0.05 to 0.90 (mean 0.56, SD 0.28), indicating a broad range of item difficulty (P). While several basic object manipulation tasks (eg, picking up the oxygen saturation clip or retrieving an infusion bottle) were completed by nearly all participants, more complex UI-dependent steps, such as setting the exact infusion rate, showed substantially lower success rates. Discrimination ranged from 0.14 to 0.60 (Mean 0.37, SD 0.17), with a subset of UI-related and navigation-heavy tasks displaying particularly good discrimination. In contrast, items with extreme difficulty (either too easy or too hard) showed limited ability to differentiate between participants. All performance data are displayed in Table 1.

The internal consistency across all 13 items, calculated using Cronbach's alpha, was α=0.79, suggesting good reliability for an early-stage checklist instrument. Removing low-discrimination items only slightly improved internal consistency (maximum α=0.80), supporting the checklist's robustness in its current form. For interaction categories Cronbach alpha was α=0.60 (n=8) for navigation, α=0.64 for object manipulation (n=8), and α=0.65 for UI activation (n=3). Solely for navigation removal of the item #1 "Attach SpO_2 clip to patient" would markedly improve reliability to 0.64.

Correlation Analyses

Specific IC showed a strong correlation with general IC (r=0.56; P=.008), providing preliminary evidence for their conceptual relatedness. In contrast, correlations between general and specific IC within corresponding interaction categories were low and nonsignificant for object manipulation (r=0.17; P=.50) and navigation (r=0.17; P=.14). The highest correspondence was observed for UI activation (r=0.41; P=.07).

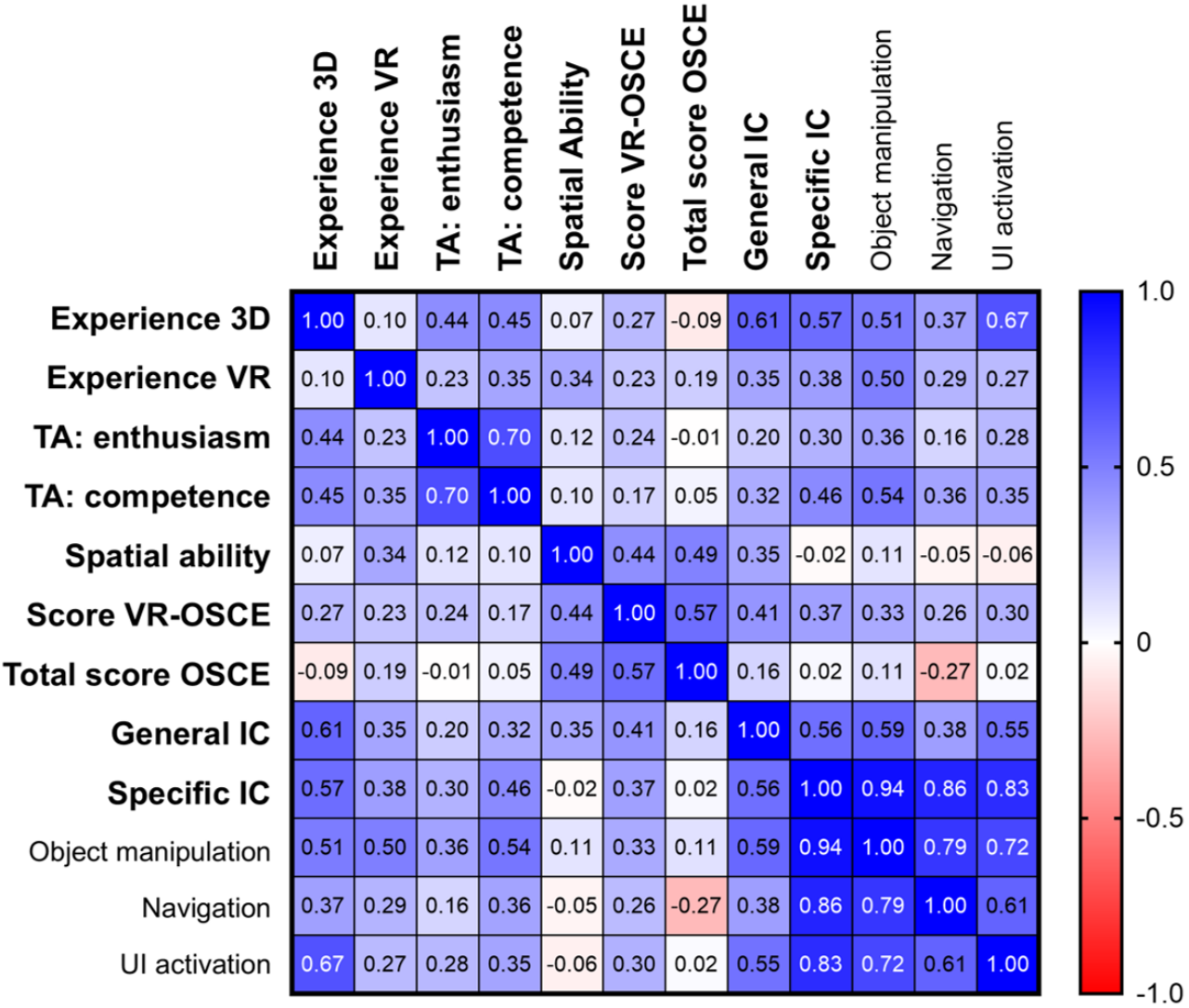
Moreover, specific IC correlated moderately to strongly with prior 3D experience (r=0.57; P=.007) and self-rated technical competence (r=0.46; P=.04). While both general and specific IC were moderately associated with performance in the VR-OSCE station at a borderline level of significance (r=0.41; P=.06 and r=0.37; P=.09, respectively), neither correlated meaningfully with overall OSCE performance (r=0.16; P=.48 and r=0.02; P=.92, respectively). Similarly, no relevant correlations were found between specific IC and technology enthusiasm or spatial ability.

When decomposing specific IC into interaction categories, object manipulation correlated most strongly with the total specific IC score (r=0.94, P<.001). Object manipulation and—most notably—UI activation also showed preliminary associations with 3D experience (r=0.51; P=.02 and r=0.67; P=.001) and technical competence (r=0.54; P=.01 and r=0.35; P=.12).

All Pearson correlations are listed in Figure 3, and their significance values are listed in Multimedia Appendix 2. Importantly, Pearson and Spearman correlation analyses yielded comparable results in terms of direction and magnitude across all key variables, indicating that the reported associations are robust to potential violations of normality assumptions. All Spearman analyses, along with their significance values, are also reported in Multimedia Appendix 2.

XSL•FO
RenderX

Figure 3. Correlation matrix between prior experience in 3D and virtual reality applications, technical affinity, spatial ability, and students' performance scores of the virtual reality-objective structured clinical examination station and the total objective structured clinical examination as well as general and specific immersive competence (the latter subdivided into object manipulation, navigation and user interface activation). The heatmap scale (right panel) indicates the color coding for positive (blue) and negative (red) correlations. The more intense the color, the higher the correlation coefficient. Corresponding Pearson significance values are listed in Multimedia Appendix 2. IC: immersive competence, OSCE: objective structured clinical examination, TA: technical affinity questionnaire, UI: user interface, VR: virtual reality.



Discussion

Principal Findings

This pilot study introduces and empirically examines the construct of specific IC as a context-sensitive dimension of VR user readiness, reflecting the ability to successfully perform application-specific interaction tasks independent of domain knowledge. Moreover, it provides initial data to inform further exploration of its potential role in predicting performance in VR-based medical assessments. Building on prior theoretical definitions of IC [15] and on established approaches for assessing broad, application-independent VR interaction skills such as object manipulation, spatial navigation, and UI activation [17,20,21], we argue that domain-specific VR systems may demand an additional, situationally anchored level of interaction proficiency. In response, we evaluated a simple, objective, and reproducible approach for measuring this construct, referred to here as "specific IC," directly within a domain-specific VR application.

Our first objective addressed the initial psychometric evaluation of our new method for measuring specific IC. Within the limited sample of this study, the developed checklist demonstrated satisfactory psychometric properties for an experimental prototype: overall internal consistency was acceptable (Cronbach α=0.79), meeting common thresholds for exploratory research, where α values above 0.70 are typically considered adequate for group-level comparisons [37]. However, internal consistency for subscales measured by Cronbach α requires further exploration. Since the α coefficient is based solely on interitem correlation, it is possible that lower values were caused by subgroups, which could not be investigated due to the small sample. Overall difficulty levels were balanced and largely within the desired range, which is typically between 0.4 and 0.8 [38]. Most items demonstrated good item-total correlations (>0.3), indicating that they were able to differentiate well

XSL•FO
RenderX

between participants with varying levels of specific IC. An item discrimination above 0.3 is usually considered indicative of a useful item, especially in early test development [38]. One exception was item #6, which displayed poor discrimination and an item difficulty of 0.90. Since our small sample size did not allow conclusions about the specific IC range of the general population, we retained this item to avoid potential floor effects in future studies with different populations. Future refinements may benefit from larger, more representative samples and from user interviews, for example, to identify subtle misunderstandings of task instructions.

Our second research objective was to explore its relationship with general IC and related human abilities to determine whether the 2 constructs reflect distinct aspects of VR readiness. We observed a moderate-to-strong and significant correlation between general and specific IC scores, while their subscales did not correlate. This suggests that while both constructs may partially share underlying cognitive resources, they also reflect distinct aspects of VR user readiness, similar to findings from other domains [26,27]. Beyond the limited evidence of this small correlational study (which may not have the statistical power to effectively detect smaller correlations), future studies could further investigate the distinctness of general and specific IC using targeted training interventions to manipulate each construct independently, as well as latent-variable approaches (eg, confirmatory factor analysis).

Both general and specific IC showed significant preliminary associations with participants' prior exposure to 3D applications, which is consistent with previous findings on (general) IC that describe these abilities as (formally or informally) trainable [16-18]. At first glance, this may raise the question of whether IC represents a redundant construct. However, due to the heterogeneity of VR experiences (eg, gaming vs educational use, passive vs interactive applications), as well as differences in preexisting motor and perceptual abilities, two individuals with similar amounts of VR experience may differ substantially in their IC. This individual variability and empirical decoupling of IC from mere experience is further supported by the multidimensional structure of IC, with separable interaction domains showing heterogeneous associations with prior experience and only moderate correlations with VR and 3D exposure in this study. Moreover, unlike experience, IC constitutes a target parameter that can be systematically assessed, compared across users, and potentially trained to enable fair conditions for immersive assessments. While familiarization or training in VR is likely to increase immersive competence on average, the extent of this effect, as well as the contributions of prior VR experience or preexisting motor and perceptual abilities, remains unclear without explicit measurement of IC. This underscores the need to assess IC directly rather than assuming VR user readiness based on exposure alone. The comparatively stronger effect of 3D experience over prior VR exposure in this study is likely due to the limited variance in the latter, as most participants had minimal VR experience, thereby reducing its explanatory power. Importantly, reported *P* values should be interpreted cautiously in light of the study's limited statistical power.

While there is some conceptual overlap with constructs such as self-reported technological competence, correlations with these measures remained weak and nonsignificant - both in this and in previous studies [17]. In our sample, participants' self-rated technology competence was slightly above average compared with normative data from a large German validation study of the TA-EG [30], whereas enthusiasm was somewhat lower, providing additional context for the observed weak associations. The absence of a significant correlation with spatial ability in the present work is noteworthy, especially given that similar—although not identical—mental rotation tasks were used in the earlier study [17]. One possible explanation could be the different disciplinary background of the present sample and participants' potential familiarity with the mental rotation test, which, for some students, forms part of the medical school aptitude test, potentially leading to biased results.

At the level of interaction categories, object manipulation and UI activation showed strong associations with general and specific IC as well as prior 3D experience, while spatial navigation appeared less indicative. One plausible explanation is the simplicity and uniformity of the navigation tasks in this study, which relied on basic teleportation in a confined virtual space. In more complex simulations requiring advanced spatial orientation or locomotion under time pressure, navigation may pose high cognitive demands [39,40] and could be more substantially associated with IC or related human abilities.

With regard to the third research objective, both general and specific IC showed moderate, borderline-significant correlations with students' medical performance as measured by the VR-OSCE station. Beyond the relatively small sample size limiting statistical power, several methodological and contextual factors may have limited the observable correlation between IC and medical performance. First, the original VR-OSCE station allowed for technical assistance from faculty members. Such instructor-mediated compensation strategies can mitigate usability barriers, particularly for users with lower digital competence, as noted in digital learning contexts [41]. However, this likely reduced the influence of IC on medical performance in our setting. Furthermore, the original medical checklist in the VR-OSCE station also included two verbal questions (about the correct diagnosis and the correct sequence of actions), which could be answered without interaction (and thus without engaging IC) and may therefore have weakened the observed correlations. Although the correlations did not reach conventional significance thresholds, the consistent direction and convergence of effects across both types of IC—despite differing operationalizations—support the construct relevance of IC in performance-based VR assessment. This is in line with arguments emphasizing that near-significant effects, particularly when consistent across measures, may still reflect meaningful relationships [42]. The lack of correlation between both types of IC and overall OSCE medical performance suggests that IC is unlikely to be a relevant factor in traditional analog assessment settings. Apparently, IC as a technical, practical skill forms a psychometrically distinct construct from analytical or academic intelligence - a distinction supported by seminal intelligence theories [43]. In line with this, empirical studies have shown that practical intelligence correlates poorly with

XSL•FO
RenderX

general cognitive ability, yet predicts job-relevant outcomes, even after controlling for general intelligence [44]. Taken together, our preliminary findings underscore the need for systematic assessment in larger cohorts to better understand the role of IC in VR-based assessment of medical performance: From a psychometric perspective, addressing potential construct-irrelevant variance is essential to uphold the validity and interpretability of performance assessments [19], particularly as immersive technologies are scaled for high-stakes examinations. From an equity standpoint, ignoring individual differences in IC may unintentionally privilege those with prior exposure to similar technologies—such as through gaming—thus reinforcing a second-order digital divide [45]. Therefore, integrating specific IC measures in VR-based examinations may help distinguish true deficits in knowledge and/or practical skills from challenges in VR interaction. This could inform both targeted training interventions and adjustments to assessment design to ensure fairness.

Limitations

Several limitations of this exploratory study should be noted. First, the sample size was small and relatively homogeneous, which limits the generalizability of the findings. Accordingly, statistical power to detect associations between specific immersive competence and performance was limited, and effect estimates may be unstable or not replicate in a larger sample.

Second, although the checklist for specific IC showed promising internal consistency and item-level characteristics, external validation by independent raters is still pending. Since the measure was assessed using a single domain-specific application (STEP-VR), the generalizability of a checklist-based approach to other VR systems or medical domains also remains to be established.

Third, retrospective recruitment may have introduced selection bias, as participants who volunteered for follow-up testing may have had above-average motivation or technical interest.

Fourth, participants had previously been exposed to the VR-OSCE station and preparatory training. While we assumed that exposure effects had subsided after six months, residual familiarity with the application might still have facilitated specific IC performance, potentially inflating scores. Similarly, frequent interim use of VR could have led to changes in IC since the original performance assessment. However, this appears unlikely considering the low levels of VR exposure reported by participants.

Fifth, specific IC scores may be influenced by system-related factors, such as software constraints or hardware-related issues (eg, controller tracking stability or attachment problems), which were not systematically assessed in this study. While no major technical failures occurred, such factors could have introduced additional variance unrelated to the intended construct.

Sixth, IC may manifest differently depending on the type of immersive system and interaction technique used (eg, controller-based VR, hand-tracking VR, or mixed reality), each relying on distinct interaction metaphors and motor demands. As a result, the present findings are specific to the controller-based VR setup used and may not directly generalize to other immersive systems.

Finally, the correlational design precludes causal conclusions, which future studies could address by manipulating IC (eg, through targeted training) and evaluating its impact on medical performance.

Conclusions

Despite its limitations, this pilot study provides initial evidence that general and specific IC may be distinct, meaningful, and psychometrically accessible constructs in the context of VR-based medical assessments. The proposed in situ approach—via guided, domain-relevant interaction tasks—demonstrated adequate item characteristics for early-stage evaluation and may provide a practical approach to gauge VR user readiness. While correlations with medical performance in the current sample were limited, this may reflect limited statistical power, ceiling effects, instructor compensation, or the independence of certain verbally addressed OSCE items from IC. Nevertheless, the increasing adoption of immersive technologies in high-stakes training and assessment underscores the potential relevance of IC as a determinant of user performance.

Future research should refine the item design, validate IC measures in larger and more diverse populations, and conduct interventional studies to clarify causal pathways. Such work may help inform strategies to mitigate performance inequities related to operational readiness and potentially enhance the fairness and validity of VR-based assessments in medical education.

Data Availability

The datasets generated and analyzed during this study are fully provided in the Supplement (Multimedia Appendices 1-3). The VR competence app used in this study was designed by the Chair of Human-Computer-Interaction at the University of Würzburg. Requests regarding its use for research purposes should be directed to the corresponding author. The software STEP-VR (Simulation-based Training of Medical Emergencies for Physicians Using Virtual Reality) used in this study is licensed by ThreeDee GmbH. Requests regarding its use for research purposes should be directed to the company.

Funding

The authors declared no financial support was received for this work.

XSL•FO
RenderX

Authors' Contributions

TM conceived the study, acquired funding, supervised the project, performed data analysis, and wrote the original draft. VS contributed to the study design and conducted the investigation. ML supported participant recruitment and assisted with data collection. JB performed the statistical analyses. SO and MH contributed to the study conceptualization and contributed to the review and editing of the manuscript. SK also contributed to the study design and participated in writing, reviewing, and editing the manuscript.

Conflicts of Interest

TM was involved in the software development of STEP-VR (Simulation-based Training of Medical Emergencies for Physicians Using Virtual Reality). All other authors declare no conflicts of interest.

Multimedia Appendix 1
Original item set for specific IC.
[XLSX File (Microsoft Excel File), 12 KB - mededu_v12i1e82136_app1.xlsx]

Multimedia Appendix 2
Significance values for Pearson correlation data from Figure 3 as well as non-parametric Spearman correlation analyses.
[XLSX File (Microsoft Excel File), 15 KB - mededu_v12i1e82136_app2.xlsx]

Multimedia Appendix 3
Raw data generated and analyzed during the study.
[XLSX File (Microsoft Excel File), 19 KB - mededu_v12i1e82136_app3.xlsx]

References

1. Mergen M, Graf N, Meyerheim M. Reviewing the current state of virtual reality integration in medical education - a scoping review. BMC Med Educ 2024;24(1):788 [FREE Full text] [doi: 10.1186/s12909-024-05777-5] [Medline: 39044186]
2. Alaker M, Wynn GR, Arulampalam T. Int J Surg 2016;29:85-94 [FREE Full text] [doi: 10.1016/j.ijsu.2016.03.034] [Medline: 26992652]
3. Jans C, Bogossian F, Andersen P, Levett-Jones T. Examining the impact of virtual reality on clinical decision making - an integrative review. Nurse Educ Today 2023;125:105767 [FREE Full text] [doi: 10.1016/j.nedt.2023.105767] [Medline: 36906980]
4. Wespi R, Schwendimann L, Neher A, Birrenbach T, Schauber SK, Manser T, et al. TEAMs go VR-validating the TEAM in a virtual reality (VR) medical team training. Adv Simul (Lond) 2024;9(1):38 [FREE Full text] [doi: 10.1186/s41077-024-00309-z] [Medline: 39261889]
5. Lindner M, Leutritz T, Backhaus J, König S, Mühling T. Knowledge gain and the impact of stress in a fully immersive virtual reality-based medical emergencies training with automated feedback: randomized controlled trial. J Med Internet Res 2025;27:e67412 [FREE Full text] [doi: 10.2196/67412] [Medline: 40465566]
6. Junga A, Kockwelp P, Valkov D, Schulze H, Bozdere P, Hätscher O, et al. Teach the unteachable with a virtual reality (VR) brain death scenario - 800 students and 3 years of experience. Perspect Med Educ 2025;14(1):44-54 [FREE Full text] [doi: 10.5334/pme.1427] [Medline: 39897616]
7. Neher AN, Bühlmann F, Müller M, Berendonk C, Sauter TC, Birrenbach T. Virtual reality for assessment in undergraduate nursing and medical education - a systematic review. BMC Med Educ 2025;25(1):292 [FREE Full text] [doi: 10.1186/s12909-025-06867-8] [Medline: 39987099]
8. Keicher F, Backhaus J, König S, Mühling T. Virtual reality for assessing emergency medical competencies in junior doctors - a pilot study. Int J Emerg Med 2024;17(1):125 [FREE Full text] [doi: 10.1186/s12245-024-00721-2] [Medline: 39333858]
9. Kman NE, Panchal AR, Patterson J, Danforth D, Mani A, Babbitt D, et al. Virtual reality simulation for assessment of hemorrhage control and SALT triage accuracy in first responders. Disaster med. public health prep 2024;18. [doi: 10.1017/dmp.2024.233]
10. Knudsen MH, Breindahl N, Dalsgaard T, Isbye D, Mølbak AG, Tiwald G, et al. Using virtual reality head-mounted displays to assess skills in emergency medicine: validity study. J Med Internet Res 2023;25:e45210 [FREE Full text] [doi: 10.2196/45210] [Medline: 37279049]
11. Rodríguez-Matesanz M, Guzmán-García C, Oropesa I, Rubio-Bolivar J, Quintana-Díaz M, Sánchez-González P. A new immersive virtual reality station for cardiopulmonary resuscitation objective structured clinical exam evaluation. Sensors (Basel) 2022;22(13):4913 [FREE Full text] [doi: 10.3390/s22134913] [Medline: 35808422]
12. Mühling T, Schreiner V, Appel M, Leutritz T, König S. Comparing virtual reality-based and traditional physical objective structured clinical examination (OSCE) stations for clinical competency assessments: randomized controlled trial. J Med Internet Res 2025;27:e55066 [FREE Full text] [doi: 10.2196/55066] [Medline: 39793025]

XSL•FO
RenderX

13. Neher AN, Berendonk C, Neubauer FB, Delahaye M, Kannape OA, Sauter TC, et al. Immersive virtual reality versus in-person performance assessment in undergraduate medical education – an exploratory comparison. Educ Inf Technol 2025;31(2):487-503. [doi: 10.1007/s10639-025-13810-3]
14. LaViola JJ, Kruijff E, McMahan RP, Bowman D, Poupyrev IP. 3D User Interfaces: Theory and Practice. Boston: Addison-Wesley; 2017.
15. Steed A, Archer D, Izzouzi L, Numan N, Shapiro K, Swapp D, et al. Immersive competence and immersive literacy: exploring how users learn about immersive experiences. Front. Virtual Real 2023;4. [doi: 10.3389/frvir.2023.1129242]
16. Junga A, Schulze H, Scherzer S, Hätscher O, Bozdere P, Schmidle P, et al. Immersive learning in medical education: analyzing behavioral insights to shape the future of VR-based courses. BMC Med Educ 2024;24(1):1413 [FREE Full text] [doi: 10.1186/s12909-024-06337-7] [Medline: 39627755]
17. Oberdörfer S, Heinisch M, Mühling T, Schreiner V, König S, Latoschik ME. Ready for VR? Assessing VR competence and exploring the role of human abilities and characteristics. Front Virtual Real 2025;6. [doi: 10.3389/frvir.2025.1608593]
18. Marsh WE, Hantel T, Zetzsche C, Schill K. Is the user trained? Assessing performance and cognitive resource demands in the virtusphere. 2013 Presented at: Proceedings of the IEEE Symposium on 3D User Interfaces (3DUI); March 16-17, 2013; Orlando, FL, USA p. 15-22. [doi: 10.1109/3dui.2013.6550191]
19. Messick S. The psychology of educational measurement. J Educational Measurement 2005;21(3):215-237. [doi: 10.1111/j.1745-3984.1984.tb01030.x]
20. Griffiths G, Sharples (nee Nichols) S, Wilson JR. Performance of new participants in virtual environments: the Nottingham tool for assessment of interaction in virtual environments (NAÏVE). International Journal of Human-Computer Studies 2006;64(3):240-250. [doi: 10.1016/j.ijhcs.2005.08.008]
21. Lampton DR, Knerr BW, Goldberg SL, Bliss JP, Moshell JM, Blau BS. The virtual environment performance assessment battery (VEPAB): development and evaluation. Presence: Teleoperators & Virtual Environments 1994;3(2):145-157. [doi: 10.1162/pres.1994.3.2.145]
22. Tusher HM, Mallam S, Nazir S. A systematic review of virtual reality features for skill training. Tech Know Learn 2024;29(2):843-878. [doi: 10.1007/s10758-023-09713-2]
23. Lie SS, Helle N, Sletteland NV, Vikman MD, Bonsaksen T. Implementation of virtual reality in health professions education: scoping review. JMIR Med Educ 2023;9:e41589 [FREE Full text] [doi: 10.2196/41589] [Medline: 36692934]
24. Mühling T, Backhaus J, Demmler L, König S. How personality and affective responses are associated with skepticism towards virtual reality in medical training-a pre-post intervention study. Cyberpsychol Behav Soc Netw 2025;28(5):335-341. [doi: 10.1089/cyber.2024.0567] [Medline: 40219582]
25. Chen MX, Hu H, Yao R, Qiu L, Li D. A survey on the design of virtual reality interaction interfaces. Sensors (Basel) 2024;24(19):6204 [FREE Full text] [doi: 10.3390/s24196204] [Medline: 39409244]
26. van Deursen A, van Dijk J. Internet skills and the digital divide. New Media & Society 2010;13(6):893-911. [doi: 10.1177/1461444810386774]
27. Wulf G, Shea CH. Principles derived from the study of simple skills do not generalize to complex skill learning. Psychon Bull Rev 2002;9(2):185-211. [doi: 10.3758/bf03196276] [Medline: 12120783]
28. Hornbæk K. Current practice in measuring usability: challenges to usability studies and research. International Journal of Human-Computer Studies 2006;64(2):79-102. [doi: 10.1016/j.ijhcs.2005.06.002]
29. Yeşilyurt E, Vezne R. Digital literacy, technological literacy, and internet literacy as predictors of attitude toward applying computer-supported education. Educ Inf Technol (Dordr) 2023:1-27 [FREE Full text] [doi: 10.1007/s10639-022-11311-1] [Medline: 36688220]
30. Karrer-Gauß K, Roesler E, Siebert FW. Neuauflage des TAEG Fragebogens: Technikaffinität valide und multidimensional mit einer kurz- oder langversion erfassen [Article in German]. Z Arb Wiss 2024;78(3):387-406. [doi: 10.1007/s41449-024-00427-4]
31. Hamstra SJ, Brydges R, Hatala R, Zendejas B, Cook DA. Reconsidering fidelity in simulation-based training. Acad Med 2014;89(3):387-392. [doi: 10.1097/ACM.0000000000000130] [Medline: 24448038]
32. Motola I, Devine LA, Chung HS, Sullivan JE, Issenberg SB. Simulation in healthcare education: a best evidence practical guide. AMEE Guide No. 82. Med Teach 2013;35(10):e1511-e1530. [doi: 10.3109/0142159X.2013.818632] [Medline: 23941678]
33. Mühling T, Späth I, Backhaus J, Milke N, Oberdörfer S, Meining A, et al. Virtual reality in medical emergencies training: benefits, perceived stress, and learning success. Multimedia Systems 2023;29(4):2239-2252. [doi: 10.1007/s00530-023-01102-0]
34. Zentrum für Testentwicklung und Diagnostik. Test für Medizinische Studiengänge und Eignungstest für das Medizinstudium III. Göttingen: Hogrefe; 2019.
35. Vandenberg SG, Kuse AR. Mental rotations, a group test of three-dimensional spatial visualization. Percept Mot Skills 1978;47(2):599-604. [doi: 10.2466/pms.1978.47.2.599] [Medline: 724398]
36. Schönrock-Adema J, Heijne-Penninga M, Van Hell EA, Cohen-Schotanus J. Necessary steps in factor analysis: enhancing validation studies of educational instruments. The PHEEM applied to clerks as an example. Med Teach 2009;31(6):e226-e232. [doi: 10.1080/01421590802516756] [Medline: 19089728]

XSL•FO
RenderX

37. Tavakol M, Dennick R. Making sense of cronbach's alpha. Int J Med Educ 2011;2:53-55 [FREE Full text] [doi: 10.5116/ijme.4dfb.8dfd] [Medline: 28029643]
38. Möltner A, Schellberg D, Jünger J. Grundlegende quantitative analysen medizinischer prüfungen - basic quantitative analyses of medical examinations [Article in German]. GMS Z Med Ausbild 2006;23(3):Doc53 [FREE Full text]
39. Kalantari S, Xu BT, Mostafavi AS, Lee A, Yang Q. Comparing spatial navigation and human environment interaction in virtual reality vs. identical real environments across the adult lifespan. arXiv:2308.15774 2023 [FREE Full text] [doi: 10.1016/j.chb.2024.108210]
40. Liang H, Lu F, Shi Y, Nanjappan V, Papangelis K. Evaluating the effects of collaboration and competition in navigation tasks and spatial knowledge acquisition within virtual reality environments. Future Generation Computer Systems 2019;95:855-866. [doi: 10.1016/j.future.2018.02.029]
41. Bond M, Marín VI, Dolch C, Bedenlier S, Zawacki-Richter O. Digital transformation in German higher education: student and teacher perceptions and usage of digital media. Int J Educ Technol High Educ 2018;15(1):48. [doi: 10.1186/s41239-018-0130-1]
42. Gelman A, Stern H. The difference between “Significant” and “Not Significant” is not itself statistically significant. The American Statistician 2006;60(4):328-331. [doi: 10.1198/000313006x152649]
43. Sternberg RJ. A triarchic view of intelligence in cross-cultural perspective. In: Irvine SH, Berry JW, editors. Human Abilities in Cultural Context. Cambridge: Cambridge University Press; 1988:60-85.
44. Case R, Demetriou A, Platsidou M, Kazi S. Integrating concepts and tests of intelligence from the differential and developmental traditions. Intelligence 2001;29(4):307-336. [doi: 10.1016/s0160-2896(00)00057-x]
45. Riggins F, Dewan S. The digital divide: current and future research directions. JAIS 2005;6(12):298-337. [doi: 10.17705/1jais.00074]

Abbreviations

IC: immersive competence
OSCE: objective structured clinical examination
STEP-VR: Simulation-based Training of Medical Emergencies for Physicians Using Virtual Reality
TA-EG: Affinity for Technology- Electronic Gadgets
UI: user interface
VR: virtual reality

Edited by L Tudor Car; submitted 10.Aug.2025; peer-reviewed by J Bauer, A Neher; comments to author 14.Jan.2026; accepted 24.Feb.2026; published 08.Apr.2026.

Please cite as:
Schreiner V, Backhaus J, Lindner M, Heinisch M, König S, Oberdörfer S, Mühling T
Development and Initial Evaluation of Specific Immersive Competence in Virtual Reality–Based Medical Assessments: Exploratory Observational Study
JMIR Med Educ 2026;12:e82136
URL: https://mededu.jmir.org/2026/1/e82136
doi:10.2196/82136
PMID:

©Verena Schreiner, Joy Backhaus, Marco Lindner, Melina Heinisch, Sarah König, Sebastian Oberdörfer, Tobias Mühling. Originally published in JMIR Medical Education (https://mededu.jmir.org), 08.Apr.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Education, is properly cited. The complete bibliographic information, a link to the original publication on https://mededu.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Use of 3D-Printed Models and Augmented Reality in Medical Student Education of Congenital Heart Disease: Randomized Controlled Trial

Tyler Langenfeld[1*], MD; Gabriel N Bahrami[2*], DO; Yea-Lyn Pak[3], BS; Susanne Wish-Baratz[3], PhD; Yuxi Zhu[1,3], PhD; Arpit Agarwal[2], MD

[1]Department of Pediatrics, Rainbow Babies & Children's Hospital, Cleveland, OH, United States

[2]Department of Pediatric Cardiology, Rainbow Babies & Children's Hospital, 2101 Adelbert rd, Cleveland, OH, United States

[3]School of Medicine, Case Western Reserve University, Cleveland, OH, United States

[*]these authors contributed equally

Corresponding Author:
Arpit Agarwal, MD
Department of Pediatric Cardiology, Rainbow Babies & Children's Hospital, 2101 Adelbert rd, Cleveland, OH, United States

Abstract

Background: Three-dimensional modalities are increasingly being used as adjuncts for medical trainees learning about complex anatomical concepts, such as congenital heart disease.

Objective: This study aimed to evaluate the use of 2 such modalities, 3D-printed models, and augmented reality (AR), in improving medical students' understanding and knowledge retention of congenital heart disease when compared to traditional teaching methods.

Methods: A prospective cohort pilot study was performed with 26 first-year medical students. Students were randomly assigned to receive a 30-minute teaching session using traditional slide-based lecture, 3D-printed model, or AR. Participants completed a 16-question pretest consisting of 4 basic general cardiology questions and 6 questions each regarding the anatomy and physiology of tetralogy of Fallot and hypoplastic left heart syndrome. Participants completed a posttest immediately following the teaching session, as well as a delayed posttest 3 weeks later.

Results: When comparing overall and subsection posttest scores, the AR group obtained perfect immediate posttest scores at a significantly increased rate compared to the lecture and 3D model groups (6/9, 67% vs 1/8, 13% and 1/9, 11%, respectively; large effect size Cramér V=0.57; *P*=.02). Participants in the lecture group reported difficulty understanding cardiac anatomy and physiology using only 2D diagrams, whereas those in the 3D-printed model and AR groups almost unanimously reported improved visualization of complex cardiac defects, which enhanced their understanding.

Conclusions: Due to the visuospatial benefits of 3D-printed models and AR, there is potential for use in medical education to improve students' knowledge of complex anatomical and physiological concepts. Students who received teaching using 3D-printed models or AR overwhelmingly reported improved 3D visualization of congenital cardiac defects compared to those who were taught via lecture. Additionally, AR and 3D-printed models offer practical opportunities for implementation into medical education curricula as both adjunct and stand-alone teaching modalities.

Trial Registration: ClinicalTrials.gov NCT07295730; https://clinicaltrials.gov/study/NCT07295730

(JMIR Med Educ 2026;12:e85967) doi:10.2196/85967

KEYWORDS

medical education; augmented reality; 3D models; congenital heart disease; teaching modalities

Introduction

Determining the best methods of content delivery has become a hallmark of education research in recent years. Aligning teaching methods with students' preferred "learning styles," such as through visual, aural, read or write, and kinesthetic assessment, is popular among educators, though it has not been shown to improve learning [1,2]. The use of "flipped classrooms" is also growing in popularity, with some estimates showing over half of US colleges use this model [3]. However, it has failed to show empirical evidence that it improves learning [4]. The primary aim across these approaches is to foster active learning [4].

Medical education is no stranger to these alternative learning models, but it carries the added complexity of encompassing concepts that are both highly visual (eg, anatomy) and rather

XSL•FO
RenderX

abstract (eg, immunologic processes). Therefore, medical school education routinely implements a variety of methods and modalities to best deliver a given topic [5]. Gross anatomy is traditionally taught with cadaveric dissection and 2D visual representations, while topics such as physiology and biochemistry rely more heavily on textbook readings and lectures. Strong visuospatial skills are required to understand complex human anatomy. Students who face difficulties learning anatomy often do so because they struggle to visualize the structures in 3D [6]. Therefore, students who develop spatial abilities are better able to learn anatomy [6]. With the goal of enhancing students' visuospatial skills in mind, various modalities and methods have been explored to augment traditional slide-based lectures. Two modalities that can provide 3D context to anatomy lectures and improve visual conceptualization are 3D-printed models and augmented reality (AR).

Slide-based lectures have been the standard for teaching modalities in higher education for many years. Its popularity is in part due to its ability to condense and organize large amounts of information as well as add visuals such as graphs, charts, and animations to supplement the lecture. Creating a slide-based presentation offers a relatively small learning curve, making it accessible to a wide majority of instructors. Furthermore, the slides are easily stored and disseminated to students, allowing for review of the material at any time. However, this method has several drawbacks. This style of lecture often promotes passive learning and leads to less engaged learners [7]. Slide presentations are also limited to 2D visuals and are thus not ideal for topics that require understanding of complex 3D structures.

3D-printed models are one of the potential solutions aimed at bridging this learning gap. With recent technological improvements, it is now possible to produce life-sized, flexible 3D replicas of real human organs. This provides learners with hands-on experience with complex structures and enables them to manipulate the models in a way that is not possible with modalities other than cadavers. Despite the technological advancements, the cost of producing 3D models is still high. It can cost as much as US $65,000 for a printer, US $15,000 for the software, and another US $400 per hour to obtain the requisite computed tomography scans from which the models will be based [8]. Accurate use of the printers requires trained personnel, which further limits their use. Despite their portability, the excessive cost of the models limits student access outside the classroom, reducing opportunities for review.

AR is among the latest technologies being explored in medical education. This modality allows learners to visualize complex anatomy in 3D but does not offer the tangible element of 3D-printed models. AR offers an immersive and interactive experience that allows users to interact with the projected images without the physical constraints imposed by a 3D model. With the appropriate software, learners can "take apart," zoom-in, and overlay models as they see fit. Being able to effectively use AR requires a comfortable level with its operation for all involved, which leads to a relatively steep learning curve when compared to other modalities. It requires personnel able to create the digital files of the teaching material and instructors who are familiar with the technology. Students unfamiliar with AR due to its novelty may experience a learning curve with the controls before using it as a learning method. Finally, AR can induce several side effects, such as headaches and motion sickness, which can adversely affect the learning experience [9].

Each learning modality has clear benefits and limitations, and with increasing availability, each has the potential to play a role in the future of medical education in some capacity [10-17]. This study aims to evaluate and compare the impact of traditional slide-based lecture, 3D-printed models, and AR on undergraduate medical students' understanding of complex anatomical and physiologic concepts as well as their subjective experiences with the modalities.

Methods

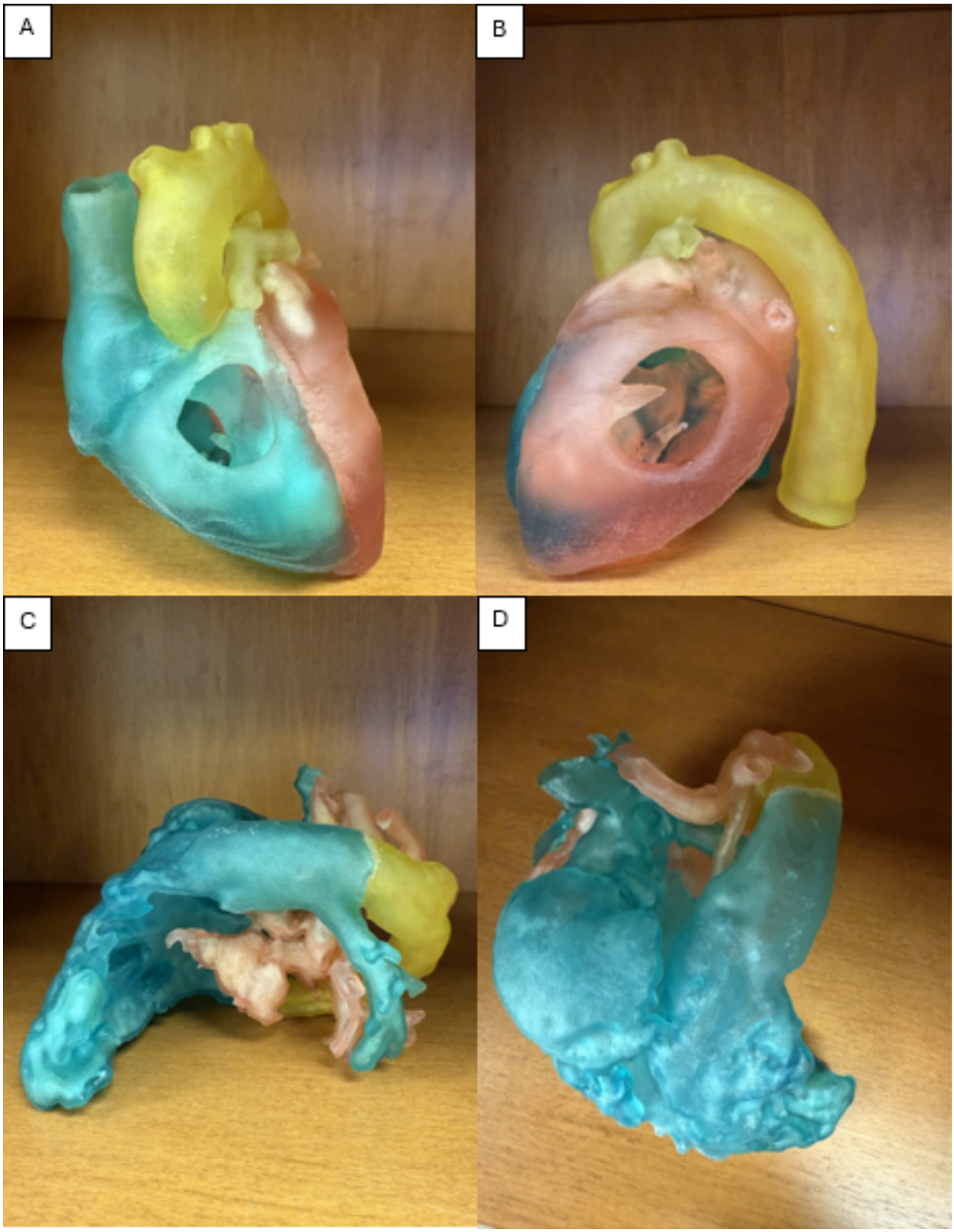
Overview

A prospective randomized controlled pilot study was conducted with first-year medical students at Case Western Reserve University in Cleveland, Ohio. A standardized curriculum was developed to ensure equivalent subject matter was covered in each group. A slide-based lecture was created using Microsoft PowerPoint and included text, static 2D figures, and basic 2D animations but did not include any videos or other media. A 16-question test was created consisting of 4 basic general cardiology questions and 6 questions each regarding the anatomy and physiology of tetralogy of Fallot (TOF) and hypoplastic left heart syndrome (HLHS). The test questions were reviewed by 3 pediatric cardiologists for technical accuracy. The questions were evaluated for face validity by 5 peers and pilot-tested on 10 participants, comprising medical residents and upper-level medical students. Other psychometric properties, including reliability and standardized administration, were reviewed with an expert in medical education and survey-based research.

Using the Unity development platform, a new AR application was developed to allow students an immersive experience using the HoloLens (Microsoft) headset. Images were sourced from cardiac computed tomography angiography images of deidentified patients with normal heart anatomy, TOF, and HLHS. The computed tomography angiography images were segmented using Elucis virtual reality software (Realize Medical) to create high-fidelity 3D anatomic models. The models were printed in 3 colors to better represent and differentiate the cardiac anatomy: blue to represent right-sided heart structures, red for left-sided structures, and yellow for pathologic structures such as a patent ductus arteriosus (Figure 1). The virtual 3D anatomic models were then converted to compatible file formats and uploaded to the Unity app and subsequently downloaded on Microsoft HoloLens to deliver the AR teaching (Figure 2). The same files were used to create a 3D-printed model using a Stratasys Medical 3D printer.

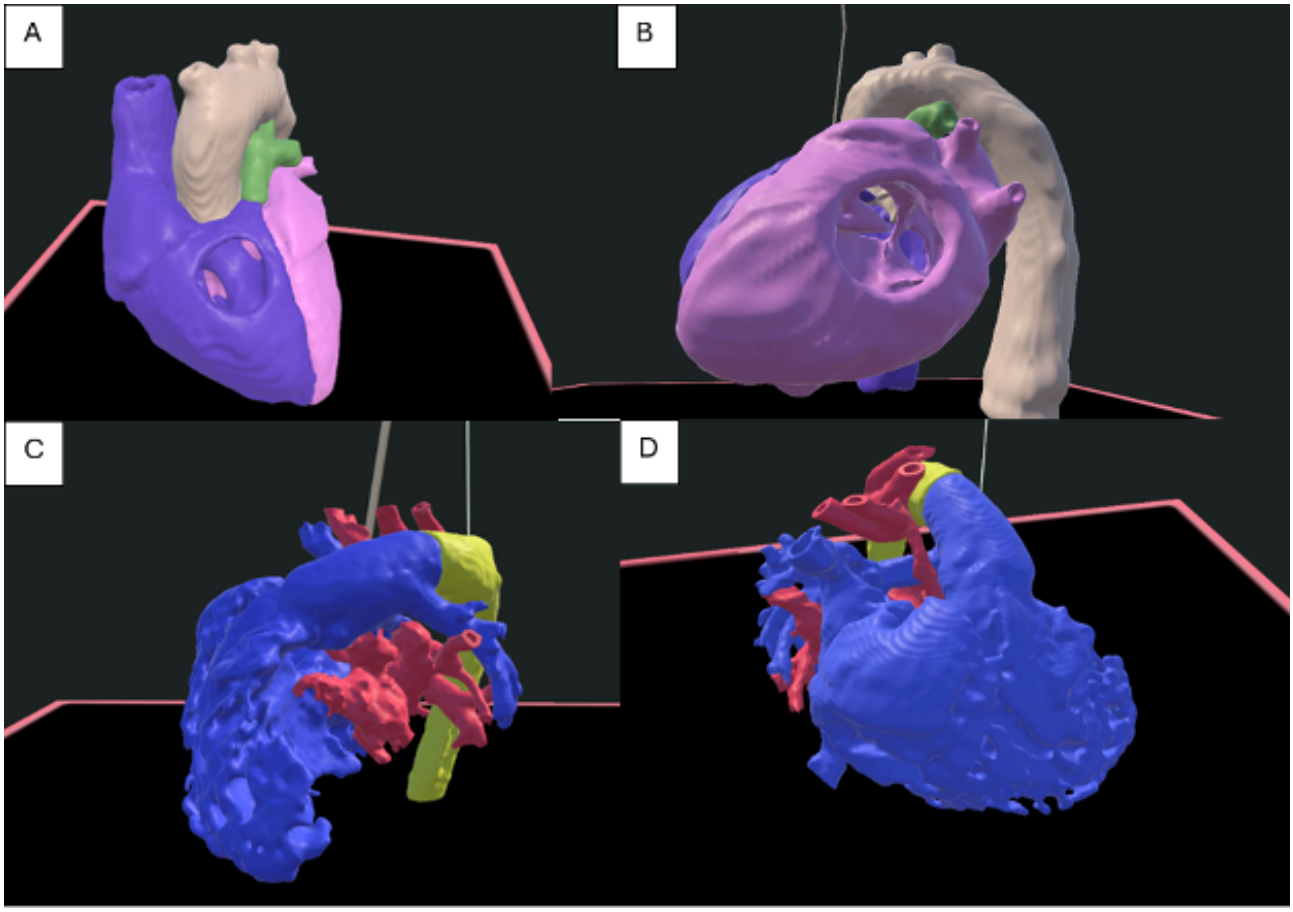
XSL•FO
RenderX

Figure 1. Three-dimensional–printed heart models: (A) right-sided view of tetralogy of Fallot (TOF), (B) left-sided view of TOF, (C) left-sided view of hypoplastic left heart syndrome (HLHS), and (D) overhead view of HLHS.



XSL•FO
RenderX

Figure 2. Augmented reality heart models: (A) right-sided view of tetralogy of Fallot (TOF), (B) left-sided view of TOF, (C) left-sided view of hypoplastic left heart syndrome (HLHS), and (D) right-sided view of HLHS.



Participants were randomly assigned to one of three groups: (1) traditional slide-based teaching, (2) hands-on instruction with 3D-printed heart models, and (3) AR instruction using HoloLens headsets. Simple randomization was performed, with participants assigned a number from 1 to 3 corresponding to their respective study groups. Numbers were given sequentially in order of participant arrival to the intervention to ensure similar size among the groups. The study was nonblinded as study team members provided the teaching sessions, and participants knew which modality they were receiving. However, the study team member who graded the questionnaires was blind to which modality was received.

Participants completed the 16-question pretest prior to taking part in a 30-minute teaching session using their assigned modality. Sessions were completed simultaneously, and each group used only the modality to which they were assigned (ie, the 3D model and AR groups did not provide corresponding teaching slides). Each group received teaching on the same topics, which included basic anatomy and physiology of the normal heart, TOF, and HLHS. The study team members who conducted the 3D model and AR teaching tailored their sessions based on the prepared slides to ensure similar content was being taught across all 3 groups. Each session was reviewed by a group of 3 cardiologists during a practice session to ensure each group received an equal amount of information.

Upon completion of the teaching session, participants completed a posttest questionnaire consisting of the same 16 questions. A delayed posttest was conducted 3 weeks later with the same questionnaire to evaluate the participants' knowledge retention. The delayed posttest also contained qualitative questions regarding their experience with the teaching session and use of the assigned modality. Confidentiality was ensured by not collecting any personally identifiable information from participants. Pretests and posttests were linked using participant-created codes that did not contain personal identifiers.

The test was divided into 3 sections: general cardiology, TOF, and HLHS. For each participant, 4 scores (general score, TOF score, HLHS score, and total score) were recorded at the pretest, posttest, and delayed posttest. Improvement scores were calculated by subtracting the pretest score from the posttest and delayed posttest scores. A perfect score was defined as achieving the full total score (16 points) on the posttest and was recorded for each modality.

Descriptive statistics were used to summarize the distribution of participants' scores. Medians with IQRs were reported for continuous variables, and frequencies with proportions were reported for categorical variables. To compare scores across modalities, the Kruskal-Wallis rank-sum test and Fisher exact test were applied to continuous and categorical variables, respectively.

To enhance the power to detect differences in score improvements from the pretest to the posttest and delayed posttest across the 3 modalities, a difference-in-differences

XSL•FO
RenderX

(DID) analysis was performed. This approach assessed the impact of each modality by comparing score changes over time. The DID model included an interaction term between modality and time point. To account for participant-level clustering, a random intercept was included using a mixed-effects model. All statistical analyses were conducted using R (version 4.3.3; R Foundation for Statistical Computing).

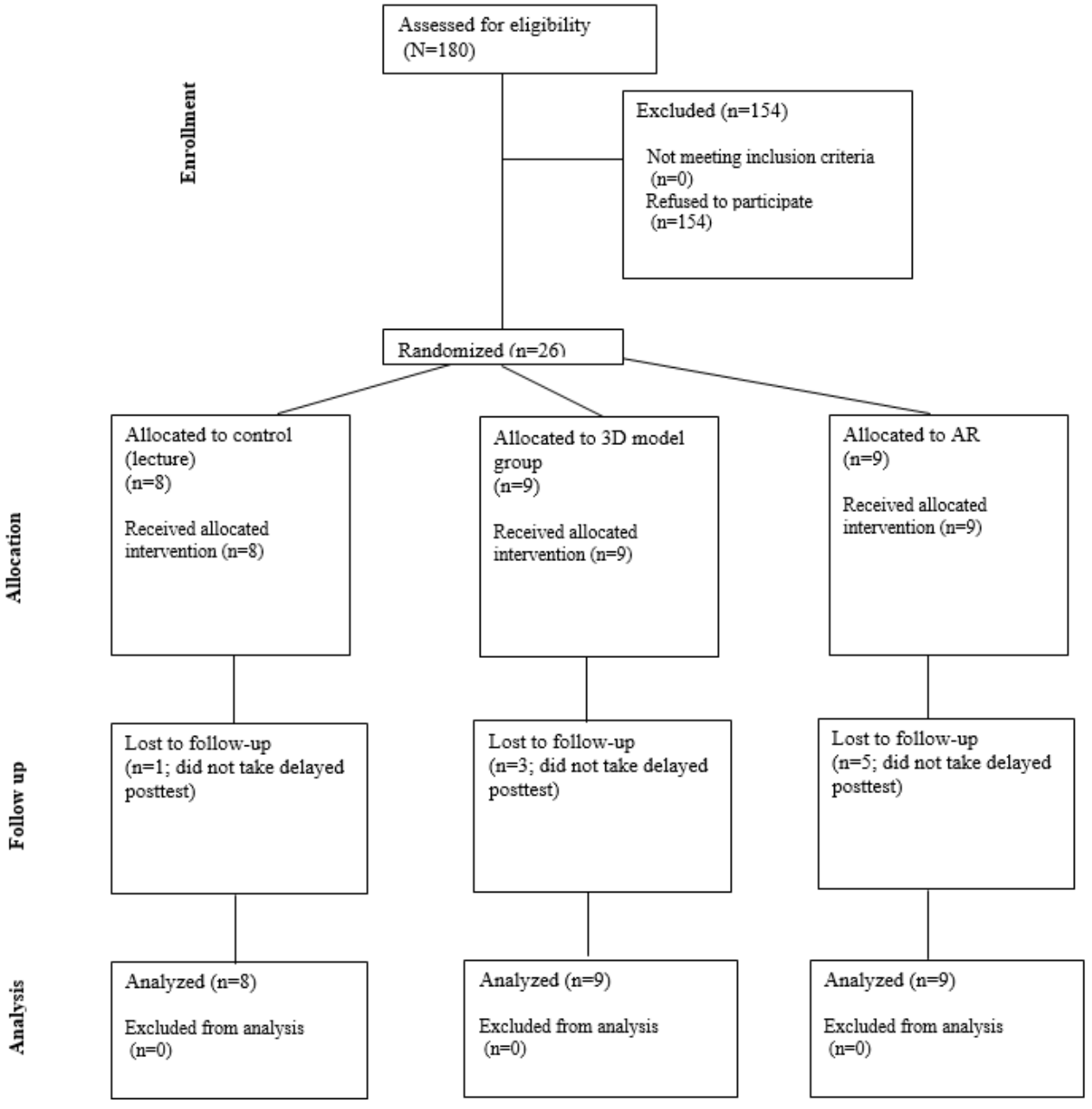
Ethical Considerations

Institutional review board approval was obtained from Case Western Reserve University as STUDY20240627. Students were recruited to participate in the study through emails and in-person announcements. Participants were provided with written and verbal explanations of the study procedures. Informed consent was obtained prior to the initial teaching session, and participants were informed of their ability to opt out of the study at any point. Participation was voluntary, and participants did not receive financial compensation. Confidentiality was ensured by not collecting any personally identifiable information from participants. Pretests and posttest were linked using participant-created codes that did not contain personal identifiers.

Results

A total of 26 first-year medical students participated in the study. In total, 8 (31%) participants were instructed by slide-based lecture, 9 (35%) were taught using 3D printed models, and 9 (35%) used the HoloLens AR goggles (Figure 3). Of the 26 participants, 17 (65%) completed the delayed posttest. Table 1 shows the median total and individual category scores, as well as the IQRs for each modality. Figure 4 shows violin plots of these scores. Figure 5 shows the improvement in posttest and delayed posttest scores compared to pretest scores for each modality.

Figure 3. CONSORT (Consolidated Standards of Reporting Trials) diagram. AR: augmented reality.



XSL•FO
RenderX

Table . Median (IQR) scores by modality.

Questions	Lecture, median (IQR)	3D model, median (IQR)	AR[a], median (IQR)	*P* value
Pretest				
General	3 (2.75 to 4)	4 (3 to 4)	3 (3 to 4)	.20
TOF[b]	3.5 (2.75 to 5.25)	4 (4 to 5)	3 (3 to 4)	.40
HLHS[c]	3 (3 to 3)	3 (2 to 3)	3 (2 to 4)	.40
Total	9.5 (7.75 to 12)	11 (10 to 11)	10 (9 to 10)	.60
Posttest				
General	4 (4 to 4)	4 (4 to 4)	4 (4 to 4)	.80
TOF	5.5 (5 to 6)	6 (4 to 6)	6 (5 to 6)	.50
HLHS	5 (5 to 5)	4 (4 to 6)	6 (6 to 6)	.05
Total	14 (14 to 15)	14 (13 to 15)	16 (15 to 16)	.10
Improvement from pretest				
General	1 (0 to 1)	0 (0 to 0)	1 (0 to 1)	.13
TOF	2 (0 to 3)	1 (0 to 1)	2 (1 to 3)	.12
HLHS	2 (2 to 2.25)	2 (0 to 3)	3 (3 to 4)	.80
Total	5 (2.75 to 6.25)	3 (3 to 5)	5 (4 to 6)	.20
Delayed posttest				
General	4 (3.5 to 4)	3.5 (3 to 4)	4 (4 to 4)	.30
TOF	6 (5 to 6)	4.5 (4 to 5.75)	5.5 (5 to 6)	.50
HLHS	5 (5 to 5)	5.5 (4.25 to 6)	5 (4.75 to 5.25)	>.99
Total	15 (13.5 to 15)	13 (12.25 to 13.75)	14.5 (13.75 to 15.25)	.40
Delayed posttest improvement				
General	1 (0 to 1)	0 (−0.75 to 0.75)	0.5 (0 to 1.25)	.30
TOF	1 (0.5 to 3)	0.5 (0 to 1)	1 (0.75 to 1.50)	.30
HLHS	2 (1.5 to 3)	3 (2.25 to 3)	3 (2.5 to 3)	.40
Total	5 (3 to 6)	3 (3 to 3.75)	4.5 (3.25 to 5.75)	.50

[a]AR: augmented reality.

[b]TOF: tetralogy of Fallot.

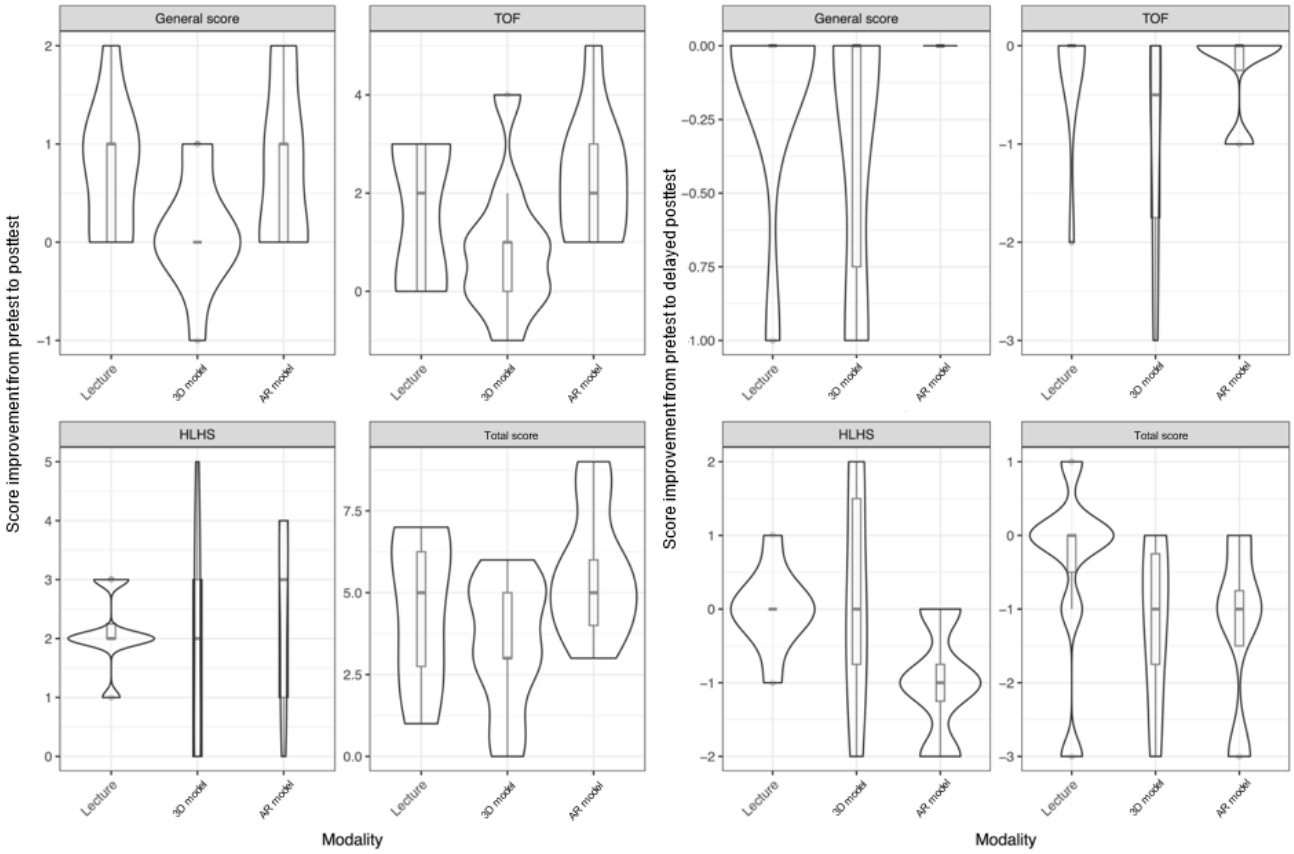
[c]HLHS: hypoplastic left heart syndrome.

XSL•FO
RenderX

Figure 4. in Figure Score distribution at 3 test time points for each modality. AR: augmented reality.



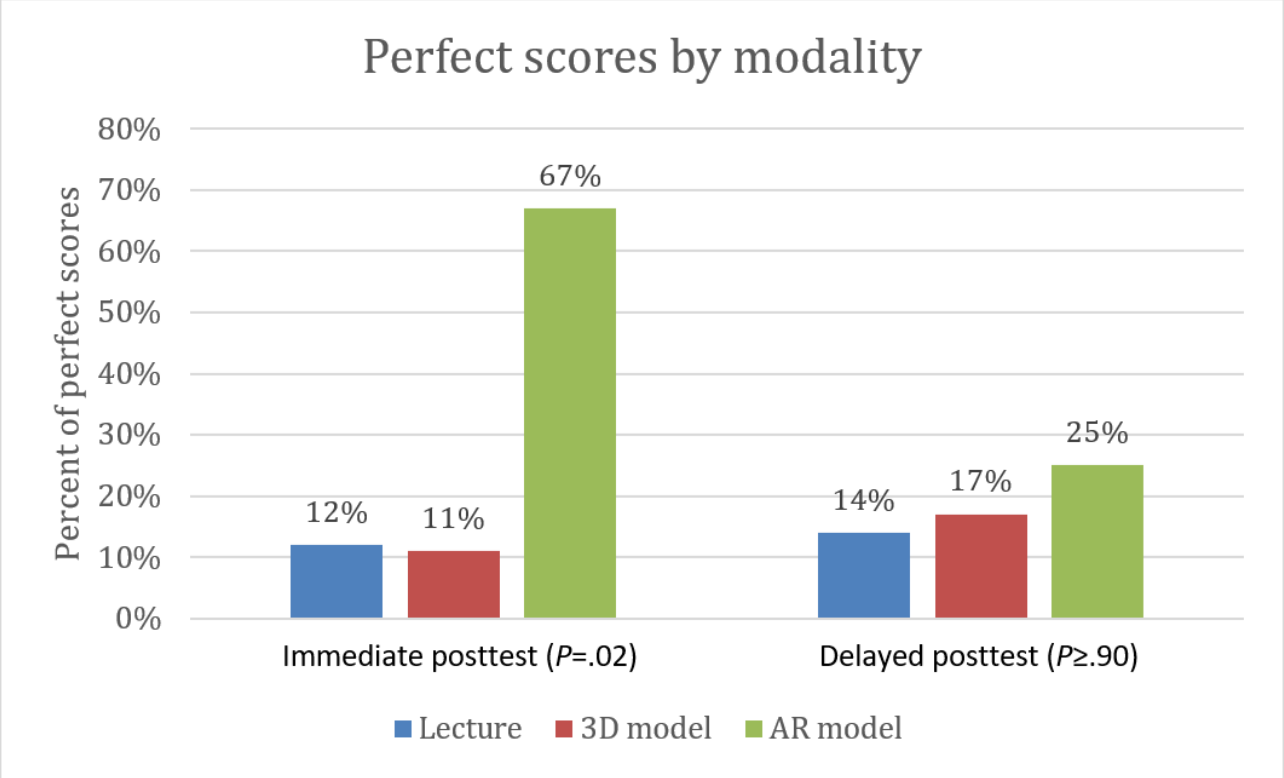
Figure 5. Difference in immediate and delayed posttest scores from pretest for each modality. AR: augmented reality; HLHS: hypoplastic left heart syndrome; TOF: tetralogy of Fallot.



XSL•FO
RenderX

There was no statistically significant difference in total median scores of pretest and posttest between the 3 modalities. There were also no statistically significant differences in subsection scores between modalities. The AR group (median 6, IQR 6 - 6) demonstrated higher immediate posttest scores on the HLHS questions compared with both the lecture model (median 5, IQR 5 - 5) and the 3D model (median 4, IQR 4 - 6; Kruskal-Wallis χ^2_2= 5.9; *P*=.05). Participants exposed to AR teaching obtained a perfect immediate posttest score at a significantly increased rate (6/9, 67%) compared to the lecture (1/8, 13%) and 3D model (1/9, 11%) groups (large effect size Cramér V=0.57; *P*=.02), though this benefit was not seen in the delayed posttest (Figure 6).

Figure 6. Perfect scores obtained on posttests by modality. AR: augmented reality.



The results of the DID analysis are shown in Tables 2 and 3. A statistically significant difference in total score improvement from pretest to posttest was observed between the AR and 3D models (mean difference 2.11, 95% CI 0.34-3.88; *P*=.02), indicating that the AR model led to a greater improvement in total scores compared to the 3D model. For the general category scores, participants in the lecture group showed a 0.64-point greater improvement than those in the 3D group (95% CI −0.02 to 1.30; *P*=.06). Additionally, participants in the AR group demonstrated a statistically significant 0.67-point greater improvement compared to the 3D group (95% CI 0.03-1.31; *P*=.05).

XSL•FO
RenderX

Table . Comparison of score improvements from pretest to posttest and delayed posttest across 3 modalities, using the 3D model as the reference.

	Lecture model		AR[a] model	
	Mean difference (95% CI)	*P* value	Mean difference (95% CI)	*P* value
Improvement in scores from posttest to pretest				
General	0.64 (−0.02 to 1.3)	.06	0.67 (0.03 to 1.31)	.05
TOF[b]	0.74 (−0.45 to 1.92)	.23	1.44 (0.29 to 2.60)	.02
HLHS[c]	0.12 (−1.04 to 1.29)	.83	0.44 (−0.68 to 1.57)	.44
Total score	1.06 (−0.77 to 2.88)	.26	2.11 (0.34 to 3.88)	.02
Improvement in scores from delayed posttest to pretest				
General	0.76 (0.05 to 1.48)	.04	0.99 (0.21 to 1.78)	.02
TOF	1.13 (−0.17 to 2.43)	.10	1.58 (0.13 to 3.02)	.04
HLHS	−0.47 (−1.72 to 0.78)	.46	−0.83 (−2.18 to 0.52)	.23
Total score	1.33 (−0.66 to 3.32)	.20	1.79 (−0.40 to 3.99)	.12

[a]AR: augmented reality.

[b]TOF: tetralogy of Fallot.

[c]HLHS: hypoplastic left heart syndrome.

Table . Comparison of score improvements from pretest to posttest and delayed posttest across 3 modalities, using the lecture model as the reference.

	3D model		AR[a] model	
	Mean difference (95% CI)	*P* value	Mean difference (95% CI)	*P* value
Posttest to pretest				
General	−0.64 (−1.3 to 0.02)	.06	0.03 (−0.63 to 0.69)	.93
TOF[b]	−0.74 (−1.92 to 0.45)	.23	0.71 (−0.48 to 1.9)	.25
HLHS[c]	−0.13 (−1.29 to 1.04)	.83	0.32 (−0.84 to 1.48)	.59
Total score	−1.06 (−2.88 to 0.77)	.26	1.06 (−0.77 to 2.88)	.26
Delayed posttest to pretest				
General	−0.76 (−1.48 to −0.05)	.04	0.23 (−0.54 to 1.01)	.56
TOF	−1.13 (−2.43 to 0.17)	.10	0.44 (−0.98 to 1.86)	.54
HLHS	0.47 (−0.78 to 1.72)	.46	−0.37 (−1.7 to 0.97)	.60
Total score	−1.33 (−3.32 to 0.66)	.20	0.47 (−1.7 to 2.63)	.68

[a]AR: augmented reality.

[b]TOF: tetralogy of Fallot.

[c]HLHS: hypoplastic left heart syndrome.

As shown in Figure 7 (individual) and Figure 8 (overall), the AR model demonstrated greater score improvement at the posttest compared to both the lecture group (mean difference 1.06, 95% CI –0.77 to 2.88; *P*=.26) and the 3D model group (mean difference 2.11, 95% CI 0.34-3.88; *P*=.02).

XSL•FO
RenderX

Figure 7. Individual trajectories of test scores across pretest, posttest, and delayed posttest. AR: augmented reality.

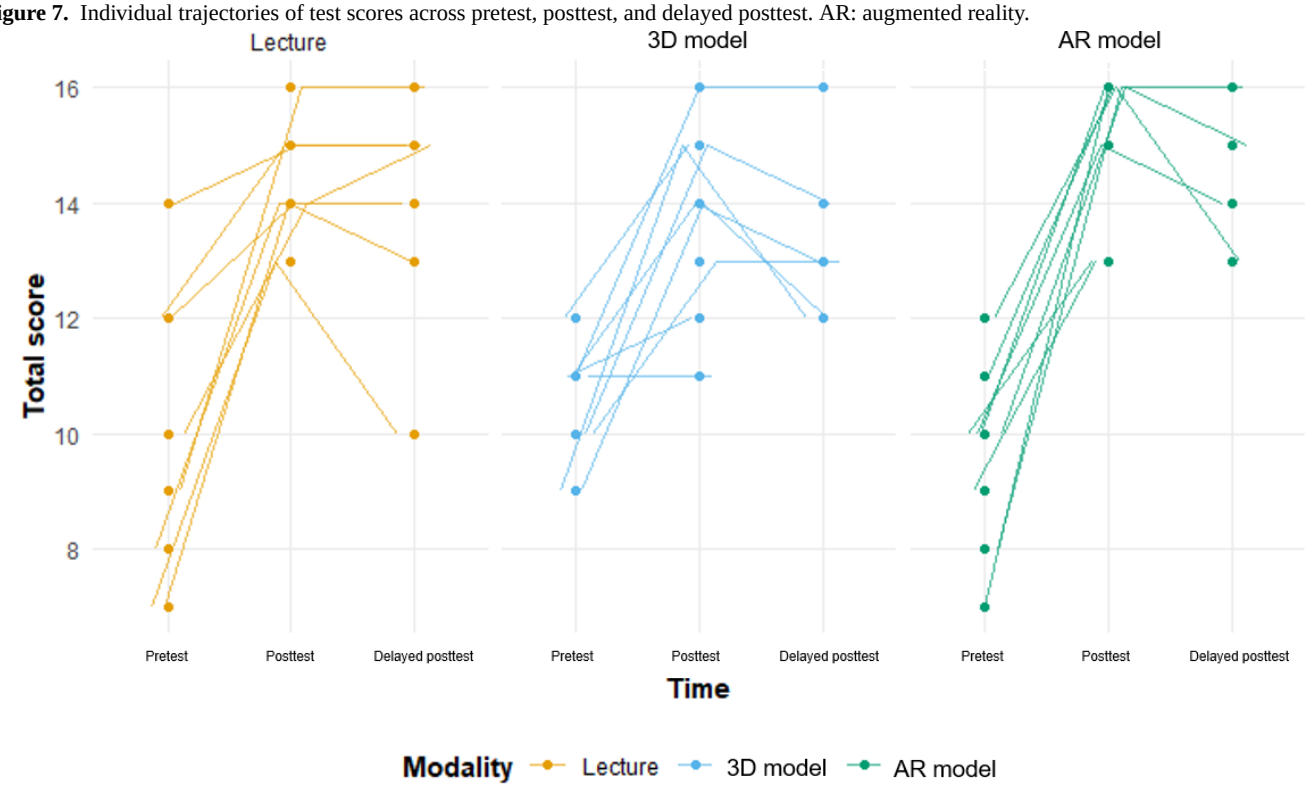
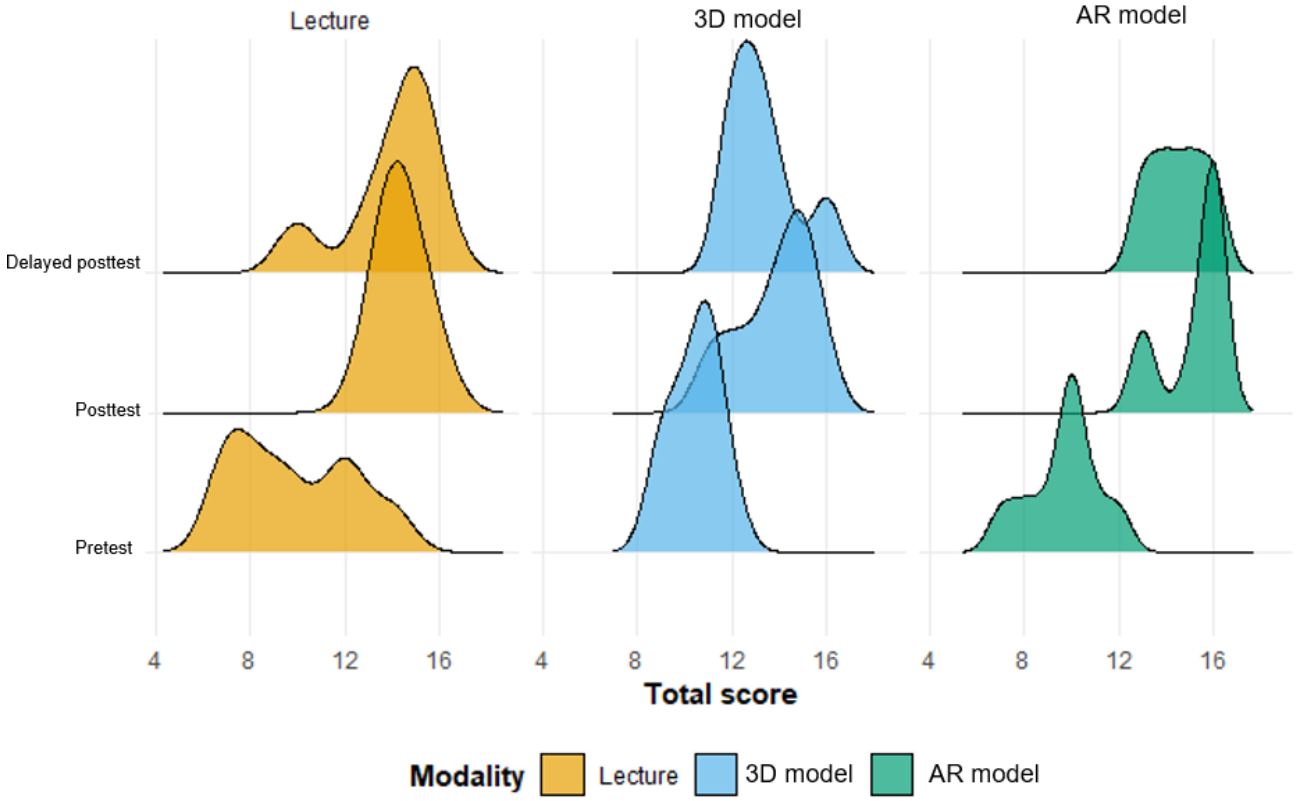


Figure 8. Temporal changes in score distributions across educational modalities. AR: augmented reality.



However, both the AR model (mean difference –0.59, 95% CI –2.75 to 1.57; *P*=.60) and the 3D-printed model (mean change –0.27, 95% CI –2.26 to 1.72; *P*=.79) showed a greater decline after the posttest when compared to the lecture group.

Participants' subjective experiences were categorized by modality. Those who participated in the slide-based lecture reported the presence of diagrams and familiarity with the method as being most helpful in understanding the material. Participants in both the 3D-printed model and AR groups reported that improved visualization of cardiac defects benefited their learning, with the ability to physically handle structures providing an additional advantage for the 3D-printed models. A common drawback reported by lecture participants was a lack

XSL•FO
RenderX

of 3D or hands-on experience. Multiple participants expressed difficulty grasping the anatomy and physiology using only 2D diagrams. Those in the 3D-printed model group noted difficulty manipulating the model to visualize internal structures. Some also reported a learning curve due to little or no prior experience with physical models. The primary downside noted by the AR group was intermittent glitches in the software.

Discussion

Principal Findings

In this pilot study of first-year medical students, we compared traditional slide-based lectures, 3D-printed models, and AR for teaching congenital heart disease. While we found no statistically significant data to suggest superiority of any one modality over another in relation to overall test scores, several important trends emerged. Students in the AR group achieved perfect immediate posttest scores at a significantly higher rate (P=.02) than those in the lecture or 3D model groups and demonstrated greater score improvements from pretest to posttest, particularly in total scores and TOF questions. DID analysis confirmed a significant advantage of AR over 3D models in immediate knowledge gain, though this did not persist with the delayed posttest. This finding may be partially explained by the small sample size, as discussed later, but it may also suggest that a single session using 3D models and AR is insufficient to produce significant improvements in retention. Subjectively, the 3D-printed model and AR groups reported that 3D visualization improved their understanding, while lecture participants reported greater difficulty conceptualizing structures using 2D diagrams.

Slide-based lectures remain a foundational component of medical education due to their efficiency and accessibility. However, their 2D format falls short when it comes to concepts that require a high degree of visuospatial understanding. This has led to the introduction of modalities such as 3D-printed models and AR into medical education curricula. In our study, students reported improved visualization, supporting the notion that 3D modalities may help address the gaps left by traditional teaching methods. Previous studies have compared 3D-printed models or AR to traditional methods [10,12,13,17]. Studies prospectively comparing all 3 modalities are, however, very limited.

Although the 3D-printed model and lecture groups showed no significant differences in total scores or score improvement, the students exposed to AR demonstrated statistically significant score improvement in multiple areas. Participants using the 3D-printed models and AR almost unanimously reported improved visualization of complex cardiac defects as a key benefit of the modality, whereas those in the lecture group felt that the lack of a 3D component hindered their understanding. This suggests that students recognize the importance of 3D modalities in conceptualizing complex anatomy and physiology.

Participants in this study experienced teaching exclusively on their assigned modality, that is, those who were taught using 3D-printed models or AR did not receive any supplementary slide-based presentation. White et al [10] performed a similar study comparing 3D-printed models to lectures in pediatric residents' understanding of ventricular septal defect and TOF, demonstrating improved understanding of TOF in the 3D-printed model group. However, in their study, both groups were given the same lecture, with the 3D-printed model incorporated into the intervention group's lecture. Importantly, even without accompanying slide-based lectures, the 3D-printed model and AR groups in this study performed comparably, or even superiorly, to the traditional teaching methods. These modalities are unlikely to fully replace traditional teaching methods; however, these findings suggest they can serve as a powerful adjunct. 3D modalities can fill visuospatial gaps left by traditional teaching, leading to more complete learning of complex anatomical and physiologic concepts such as congenital heart disease. Further investigation is therefore needed to evaluate the efficacy of incorporating these 3D modalities into traditional teaching methods.

AR and 3D-printed models can be feasibly incorporated into medical school curricula in several ways, as each has features that make it suitable for both classroom and laboratory settings. In the classroom setting, 3D-printed models can serve as a reference during traditional lectures. Students would be able to augment what they are learning from words, charts, and diagrams with a greater visuospatial component. These models can also serve as an adjunct in the cadaver laboratory as they can be compared to normal and diseased specimens as well as show anatomic variation within specific pathologies. Use of 3D-printed models allows for exposure to a wider variety of pathologies and structural anomalies than what can be feasibly achieved with cadavers. This adjunctive benefit is supported by a study by Aytaç et al [18], which demonstrated improved cardiac anatomy test scores in students taught using "plastinated" heart models (a process of dehydrating and fixing specimens) in addition to traditional formalin-fixed specimens, compared with those who received only formalin-fixed specimens.

Exposure to cardiac pathology specimens is usually limited by available cadavers or access to a specimen registry. 3D-printed models greatly increase the accessibility of physical representations of cardiac pathologies to students and allow for more hands-on learning. Furthermore, where cadaveric dissection is limited to sessions within the laboratory, 3D-printed models have no location restriction, which allows students to spend more time with the content material. Despite high initial costs, 3D-printed models can ultimately be more cost-effective than cadavers due to their durability and ability to be reused multiple times [8]. AR offers similar benefits to 3D-printed models, including improved access to visualization of cardiac pathologies and the ability to be used outside traditional laboratory or classroom settings. While 3D-printed models can be used simultaneously with traditional lecture, using AR simultaneously may be visually overstimulating and make it difficult to gain the desired effect. However, AR has several advantages over 3D-printed models. The first is the greater amount of information that can be conveyed in the AR space. AR cannot only display cardiac anatomy but also convey dynamic information such as blood flow, oxygen saturation, and pressures, which are essential for understanding key

XSL•FO
RenderX

physiological concepts. The second is the lack of physical constraints inherent in 3D-printed models. Students are able to view cardiac anatomy in multiple ways, including interior structures that can be difficult to assess with physical models. They are also better able to view small structures, due to the ability to zoom in, with AR than in 3D-printed models or cadaveric specimens. Because of these qualities, using AR in stand-alone sessions can be a practical supplementation following traditional lectures.

The benefits of supplemental, interactive learning approaches have previously been explored. A recent study by Ogut et al [19] provided medical students with a supplemental course on cross-sectional anatomy of various organ systems with clinical case correlates using computed tomography and assessed their subjective experience. Students reported an overall positive experience with the course and believed it improved their learning, primarily with regard to integrating anatomical knowledge with clinical practice. Similar to the aforementioned study, our findings suggest that students favor interactive, visual methods for learning anatomical concepts.

Limitations

There are several limitations that affect this study's reliability and generalizability. The questionnaire used as the study's assessment tool has only been pilot-tested and has therefore not been validated through peer review. The statistical power of the study was limited by a small sample size (n=26) and a high attrition rate (9/26, 34.6%). This is evident in the lack of significance in the posttest scores, where previous improvements in the 3D model and AR group appeared to diminish. The follow-up time of this study was 3 weeks, which limits conclusions about retention over longer periods that would be useful in medical education, such as board exam preparation.

The study population may present generalizability limitations as well. Participating medical students were from a single class at a single medical school. Students in this study were accustomed to AR technology as it is used routinely in their core curriculum. Therefore, they were not inhibited by the steep learning curve often associated with these devices. Conversely, they had little to no experience with 3D-printed models as a learning method. This introduces the possibility of familiarity bias affecting both objective and subjective outcomes. It is not surprising that students may gain relatively less understanding of a new and complex topic when it is presented in an unfamiliar format. Students appeared highly engaged with this modality, as many chose to spend additional time interacting with the models after completing the study. Therefore, the relatively smaller improvement following teaching with 3D models may be more a result of comfortability with the modality than the efficacy of the modality itself. Further studies involving participants who use 3D models as part of their standard curriculum and those unfamiliar with AR are warranted. Finally, this study focused on a narrow content topic of anatomy and physiology of 2 cardiac disease processes. This, however, does not limit the generalizability of this study's findings, as cross-sectional imaging of any structure can be converted into 3D models and used for teaching purposes. AR, for instance, has been studied in numerous anatomical topics from skull to kidneys [16,17,20]. Therefore, advanced 3D visualization modalities can be applied to any body part or structural disease process and used for teaching complex anatomy to medical students.

Conclusions

With advances in technology and the increased accessibility of these resources, we are seeing a paradigm shift toward the increasing use of 3D modalities in medical education and training. Students who received teaching using 3D-printed models or AR overwhelmingly reported improved 3D visualization of congenital cardiac defects compared to those who were taught via lecture. Owing to their visuospatial benefits, these modalities may serve as effective tools in medical education to improve student comprehension of complex anatomical and physiologic concepts. This study provides insight into the feasibility, benefits, and limitations of 3D modalities in medical education and emphasizes the importance of larger-scale studies to guide their integration into medical education and maximize educational impact.

Acknowledgments

The authors would like to acknowledge Case Western Reserve University School of Medicine for providing the physical space for this project's intervention and RediMinds for their role in developing the application used with the HoloLens headsets.

Funding

No external funding was received for this study.

Authors' Contributions

Conceptualization: AA (lead), TL (equal), GB (equal), YLP (supporting), SW-B (supporting)
Data curation: TL
Formal analysis: YZ
Funding acquisition: AA
Investigation: TL (lead), GB (equal), AA (equal), YLP (supporting)
Methodology: GB (lead), TL (equal), AA (equal), YZ (supporting)
Project administration: AA (lead), SW-B (equal), YLP (equal), GB (equal), TL (supporting)

XSL•FO
RenderX

Resources: GB (lead), AA (equal), SW-B (equal)
Software: AA
Supervision: AA
Validation: AA (lead), GB (equal), TL (equal)
Visualization: TL (lead), GB (equal), YZ (equal), AA (supporting), SW-B (supporting)
Writing—original draft: TL (lead), GB (equal), YZ (supporting)
Writing—review and editing: GB (lead), TL (equal), SW-B (equal), AA (equal)

Conflicts of Interest

None declared.

Checklist 1
CONSORT-eHEALTH checklist (V 1.6.1).
[PDF File, 1108 KB - mededu_v12i1e85967_app1.pdf]

References

1. Hattie J, O'Leary T. Learning styles, preferences, or strategies? An explanation for the resurgence of styles across many meta-analyses. Educ Psychol Rev 2025 Jun;37(2). [doi: 10.1007/s10648-025-10002-w]
2. Fallace T. The long origins of the visual, auditory, and kinesthetic learning style typology, 1921-2001. Hist Psychol 2023 Nov;26(4):334-354. [doi: 10.1037/hop0000240] [Medline: 37471036]
3. Schaffhauser D, Rhea K. 55 percent of faculty are flipping the classroom. Campus Technology. URL: https://campustechnology.com/articles/2016/10/12/55-percent-of-faculty-are-flipping-the-classroom.aspx [accessed 2025-07-24]
4. Kapur M, Hattie J, Grossman I, Sinha T. Fail, flip, fix, and feed – rethinking flipped learning: a review of meta-analyses and a subsequent meta-analysis. Front Educ 2022;7:956416. [doi: 10.3389/feduc.2022.956416]
5. Verma N, Yui JC, Record JD, Hueppchen NA, Naik RP. The changing landscape of the Preclinical Medical School curriculum: results from a nationwide survey of United States medical school curriculum deans. Am J Med 2024 Feb;137(2):178-184. [doi: 10.1016/j.amjmed.2023.10.021] [Medline: 37890571]
6. Nokovitch L, El Hajj H, Deneuve S, De Andrade V, Gagnayre R, Margat A. Visual-spatial abilities enhancement and spatial anatomy learning: a systematic review. Med Educ 2025 Dec;59(12):1322-1332. [doi: 10.1111/medu.70022] [Medline: 40830038]
7. Klemm WR. Computer slide shows: a trap for bad teaching. College Teaching 2007 Jul;55(3):121-124. [doi: 10.3200/CTCH.55.3.121-124]
8. Bui I, Bhattacharya A, Wong SH, Singh HR, Agarwal A. Role of three-dimensional visualization modalities in medical education. Front Pediatr 2021;9:760363. [doi: 10.3389/fped.2021.760363] [Medline: 34950617]
9. Park S, Lee G. Full-immersion virtual reality: adverse effects related to static balance. Neurosci Lett 2020 Aug 10;733:134974. [doi: 10.1016/j.neulet.2020.134974] [Medline: 32294492]
10. White SC, Sedler J, Jones TW, Seckeler M. Utility of three-dimensional models in resident education on simple and complex intracardiac congenital heart defects. Congenit Heart Dis 2018 Nov;13(6):1045-1049. [doi: 10.1111/chd.12673] [Medline: 30230245]
11. Smerling J, Marboe CC, Lefkowitch JH, et al. Utility of 3D printed cardiac models for medical student education in congenital heart disease: across a spectrum of disease severity. Pediatr Cardiol 2019 Aug;40(6):1258-1265. [doi: 10.1007/s00246-019-02146-8] [Medline: 31240370]
12. Karsenty C, Guitarte A, Dulac Y, et al. The usefulness of 3D printed heart models for medical student education in congenital heart disease. BMC Med Educ 2021 Sep 8;21(1):480. [doi: 10.1186/s12909-021-02917-z] [Medline: 34496844]
13. Loke YH, Harahsheh AS, Krieger A, Olivieri LJ. Usage of 3D models of tetralogy of Fallot for medical education: impact on learning congenital heart disease. BMC Med Educ 2017 Mar 11;17(1):54. [doi: 10.1186/s12909-017-0889-0] [Medline: 28284205]
14. Hanna MG, Ahmed I, Nine J, Prajapati S, Pantanowitz L. Augmented reality technology using Microsoft HoloLens in anatomic pathology. Arch Pathol Lab Med 2018 May;142(5):638-644. [doi: 10.5858/arpa.2017-0189-OA] [Medline: 29384690]
15. Maresky HS, Oikonomou A, Ali I, Ditkofsky N, Pakkal M, Ballyk B. Virtual reality and cardiac anatomy: exploring immersive three-dimensional cardiac imaging, a pilot study in undergraduate medical anatomy education. Clin Anat 2019 Mar;32(2):238-243. [doi: 10.1002/ca.23292] [Medline: 30295333]
16. Moro C, Štromberga Z, Raikos A, Stirling A. The effectiveness of virtual and augmented reality in health sciences and medical anatomy. Anat Sci Educ 2017 Nov;10(6):549-559. [doi: 10.1002/ase.1696] [Medline: 28419750]
17. Moro C, Birt J, Stromberga Z, et al. Virtual and augmented reality enhancements to medical and science student physiology and anatomy test performance: a systematic review and meta-analysis. Anat Sci Educ 2021 May;14(3):368-376. [doi: 10.1002/ase.2049] [Medline: 33378557]

XSL•FO
RenderX

18. Aytaç G, Öğüt E, Şekerci R, Gürçay S, Oğuz N, Sindel M. The contribution of plastinates to teaching complex anatomy of the heart. Anatomy 2021;15(1):69-75. [doi: 10.2399/ana.21.857788]
19. Ogut E, Yildirim FB, Senol Y, Senol AU. Comprehensive evaluation of the educational impact and effectiveness of specialized study modules in cross-sectional anatomy: a study on student engagement and learning outcomes. BMC Med Educ 2025 Apr 10;25(1):514. [doi: 10.1186/s12909-025-07050-9] [Medline: 40211255]
20. Rubat Baleuri F, Pattou M, Jaffredo M, et al. From visualization to education: the role of 3D-printed and virtual kidney models in training for renal cancer surgery, a systematic review. J Robot Surg 2025 Dec 5;20(1):52. [doi: 10.1007/s11701-025-02961-7] [Medline: 41345321]

Abbreviations

AR: augmented reality
DID: difference-in-differences
HLHS: hypoplastic left heart syndrome
TOF: tetralogy of Fallot

Edited by J Moen; submitted 15.Oct.2025; peer-reviewed by E Ogut, JH Kim; revised version received 26.Dec.2025; accepted 29.Dec.2025; published 05.May.2026.

Please cite as:
Langenfeld T, Bahrami GN, Pak YL, Wish-Baratz S, Zhu Y, Agarwal A
Use of 3D-Printed Models and Augmented Reality in Medical Student Education of Congenital Heart Disease: Randomized Controlled Trial
JMIR Med Educ 2026;12:e85967
URL: https://mededu.jmir.org/2026/1/e85967
doi:10.2196/85967

© Tyler Langenfeld, Gabriel N Bahrami, Yea-Lyn Pak, Susanne Wish-Baratz, Yuxi Zhu, Arpit Agarwal. Originally published in JMIR Medical Education (https://mededu.jmir.org), 5.May.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Education, is properly cited. The complete bibliographic information, a link to the original publication on https://mededu.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Blended Learning Compared With Face-to-Face Learning Among Family Medicine Residents: Randomized Controlled Trial

Pierre-Yves Meunier[1,2,3], MD; Sophie Schlatter[2,4], PhD; Juliette Macabrey[2,3,5,6], MD; Frédéric Zorzi[3], MD; Thomas Colleony[3,7], MD; Rémy Boussageon[3,8], MD, PhD; Hubert Maisonneuve[3,9,10], MD, MHPE; Marion Lamort-Bouché[2,3], MD, PhD

[1]Maison de santé pluriprofessionnelle et universitaire Beauvisage, Lyon, France

[10]Mermoz Primary Health Center, Lyon, France

[2]Research on Healthcare Performance RESHAPE, Inserm U1290, Université Claude Bernard Lyon 1, Lyon, France

[3]Collège universitaire de médecine générale, Université Claude Bernard Lyon 1, 8 avenue Rockefeller, Lyon, France

[4]Healthcare Simulation Center, SIMULYON, Hospices Civils de Lyon, Université Claude Bernard Lyon 1, Lyon, France

[5]Maison de santé pluriprofessionnelle et universitaire Bel Air, Saint Priest, France

[6]Faculty of Education, University of Ottawa, Ottawa, ON, Canada

[7]French Military Medical School, Bron, France

[8]Laboratoire de biométrie et biologie évolutive UMR CNRS 5558, Université Claude Bernard Lyon 1, Université de Lyon, Lyon, France

[9]University Institute for Primary Care (IuMFE), Faculty of Medicine, University of Geneva, Geneva, Switzerland

Corresponding Author:
Pierre-Yves Meunier, MD
Maison de santé pluriprofessionnelle et universitaire Beauvisage, Lyon, France

Abstract

Background: The medical education of French family medicine residents involves active, socioconstructivist-inspired small-group courses useful for skill acquisition. This is challenged by the increasing gap between the growing number of residents and the limited number of teachers. Blended courses have the potential to address this issue by reducing the duration of face-to-face sessions while preserving small-group courses.

Objective: This study aimed to compare the effects of blended vs traditional, face-to-face, active, socioconstructivist learning on the acquisition of knowledge and skills by family medicine residents.

Methods: We conducted a randomized controlled trial to compare a blended course and a traditional course. The blended course involved 2.5 hours of asynchronous e-learning and a 3-hour face-to-face session. The traditional course involved 5.5 hours of face-to-face teaching. Both courses were grounded in socioconstructivist principles and actively engaged residents. The primary outcome was residents' self-assessment of knowledge and skills. Secondary outcomes included satisfaction with knowledge- or skill-related learning objectives and academic achievement at 6 months.

Results: We included 155 family medicine residents (n=78, 50.3% in the blended course and n=77, 49.7% in the traditional course). There was no significant difference between groups regarding the primary outcome (mean difference 0.40 [maximum mean difference 20] points, 95% CI −0.21 to 1.02; *P*=.19; Cohen *d*=0.21). No significant differences were observed for the secondary outcomes except for knowledge self-assessment, which was higher in the blended course but not educationally meaningful (mean difference 0.40 [maximum possible 10] points, 95% CI 0.07-0.71; *P*=.02; Cohen *d*=0.39).

Conclusions: Blended courses can help sustain socioconstructivist small-group teaching methods while accommodating a growing family medicine resident population, with no deleterious impact on knowledge and skill self-assessments.

Trial Registration: ClinicalTrials.gov NCT06409273; https://clinicaltrials.gov/ct2/show/NCT06409273

(JMIR Med Educ 2026;12:e86387) doi:10.2196/86387

KEYWORDS

primary health care; family practice; education; medical; graduate; internship and residency; distance; computer-assisted instruction; educational technology; self-assessment; educational measurement; randomized controlled trial

XSL•FO
RenderX

Introduction

The World Health Organization emphasizes the critical role of efficient primary care systems in ensuring population health [1]. However, many countries face a shortage of family physicians, which limits access to primary care services [2-4]. To address this issue, some countries, including France, have recently increased the number of medical students [5-7]. In France, family medicine residents are medical graduates enrolled in a 4-year family medicine residency program [8]. Their curriculum is competency-based, guided by a national framework [9]. The constructive alignment of competency-based educational objectives, learning activities, and assessment implies, among other things, the provision of active, socioconstructivist-inspired courses [10,11]. Social constructivism emphasizes the role of social interactions and collaboration for effective learning [12]. As a result, residents engage in a variety of collaborative learning activities with 3 to 25 participants per group. Reports of these activities, individual narrative medicine work, and assessments are collected in a portfolio. However, from the perspective of health profession educators, it is difficult to maintain feasibility with increasing class sizes [13]. This difficulty is particularly acute for family medicine residents because their growing number has outpaced the availability of teaching hours, threatening the quality of their education.

Face-to-face traditional learning relies on nondigital materials and human interaction [14]. Blended learning (BL) combines face-to-face education with asynchronous or synchronous e-learning [15]. BL is often used to shift from a teacher-centered, behaviorist instructional model to a student-centered, socioconstructivist instructional model [16]. In blended courses, face-to-face sessions reinforce knowledge and develop skills through the collective mobilization of knowledge acquired during e-learning [17]. Two meta-analyses found that BL had a more positive effect on knowledge acquisition compared to traditional education in health professions (including medical students) but with large heterogeneity in the results [18,19]. This heterogeneity can be explained by differences in student populations and missing details about the instructional approach. Moreover, these studies did not evaluate the effect of BL on skill acquisition. While BL appears to improve satisfaction among both students and teachers, no meta-analysis has specifically evaluated this outcome. Teachers have reported that BL has the potential to increase interactions with students [20] and allow educators to develop more content [21] and they felt greater job satisfaction with this modality [22]. Students have generally been satisfied [23].

While existing studies suggest that BL is generally effective, its effectiveness compared with an already active, socioconstructivist, face-to-face course remains unexplored. Moreover, its effect on skill acquisition among postgraduate medical trainees (residents) remains underexplored. In this study, we compared the effects of a blended course (asynchronous active e-learning with shorter socioconstructivist face-to-face sessions) vs a fully face-to-face socioconstructivist course on the acquisition of knowledge and skills by family medicine residents. While implementing BL could mitigate the shortage of teachers by shortening face-to-face sessions, reduced in-person interaction may limit feedback and engagement, potentially leading to lower knowledge and skill acquisition.

Methods

Study Design

We conducted a single-blind randomized controlled trial with 2 parallel arms. The experimental arm received a blended course comprising an asynchronous e-learning module (2.5 hours; 45% of the course) followed by a face-to-face session (3 hours; 55% of the course). The control arm received the usual face-to-face traditional course (5.5 hours; 100% of the course). The trial was conducted between May 2024 and November 2024 at Université Claude Bernard Lyon 1 (France). It followed the CONSORT (Consolidated Standards of Reporting Trials) guidelines [24] and was prospectively registered on ClinicalTrials.gov (NCT06409273).

Sample and Setting

All first-year family medicine residents at Université Claude Bernard Lyon 1 were administratively enrolled, with no exclusion criteria. The course was delivered over a single day. Residents were assigned to 1 of 2 available dates. Participants were told that the study compared different teaching methods, with no further detail. They were block randomized using random sequences of block sizes of 4 and 8 in a 1:1 allocation ratio. Randomization was performed by the first author using random permuted blocks generated online [25]. On each teaching date, 4 groups were formed (2 per arm), with a maximum of 23 participants per group. Classrooms were configured similarly, with tables arranged in a U shape and spaces reserved for breakout work.

The 5 teachers were randomized via coin toss to 1 of the 2 study arms. Teachers who taught on both dates crossed over to the other arm on the second day. For each arm, we developed detailed, time-stamped scripts to ensure instructional fidelity and equivalence of content and duration.

Educational Intervention and Conceptual Framework

The course prepared residents to complete the thesis required for full qualification in France. This is a professionally oriented doctoral thesis publicly defended before an examination board. It is required for the award of the national Doctor of Medicine degree, which permits registration with the French Medical Council for independent practice. Distinct from a PhD, it is a concise, supervised scholarly work demonstrating the candidate's ability to design, conduct, and communicate research relevant to medical practice. We specified 5 process-oriented learning objectives: 3 for knowledge acquisition and 2 for skill development (Table 1).

XSL•FO
RenderX

Table . Learning objectives across study arms.

Learning objective	Study arm	
	Blended course	Traditional course
Knowledge oriented		
Identify the practical organization and regulatory requirements of the medical thesis	e-Learning	Face-to-face
Understand the different steps of a research project	e-Learning	Face-to-face
Summarize the medical thesis application form requirements	e-Learning	Face-to-face
Skill oriented		
Frame the research question for a thesis project	Face-to-face	Face-to-face
Assess the feasibility of a medical thesis project	Face-to-face	Face-to-face

The blended course sequenced active knowledge construction online with supervised skill development in person. We made the alignment between the 2 sessions explicit to residents to underscore the coherence between the learning objectives and the learning methods. The blended course arm began with an e-learning session grounded in self-regulated learning (SRL). SRL has been defined as self-generated thoughts, feelings, and actions that are planned and cyclically adapted to the attainment of personal goals [26]. Process-oriented learning objectives (Table 1) for knowledge acquisition were designed to activate the SRL forethought phase (goal setting and strategy selection) [27]. We aimed to help residents identify immediate next steps (securing a supervisor, conducting a literature review, and framing a research question), whereas strategy selection focused on learning how to choose an appropriate research method. Formative checkpoints and immediate feedback were embedded through interactive HTML5 Package activities—videos with decision points (“crossroads”), multiple-choice questions, drag-the-words items, and navigation hot spots. This design supported ongoing self-monitoring and adaptive control (comparison to standards and adjustment), thereby reducing discrepancies between current understanding and target performance. The accompanying explanatory text and curated external links to websites or documents enabled resource management and self-paced elaboration, further supporting SRL. The e-learning module was implemented online in the course management system Moodle (Moodle HQ; Multimedia Appendix 1).

The traditional course arm began with the collaborative construction of a knowledge base anchored in the knowledge-oriented learning objectives. Teachers acted as facilitators, providing key information and externalizing group thinking in mind maps that made success criteria and progress visible for monitoring.

Skill-oriented learning objectives were addressed identically in both arms during a face-to-face session. For the socioconstructivist activities, peer evaluation of thesis projects was organized in breakout groups, which discussed the research question and feasibility. These activities provided external feedback according to standards explicitly set by teachers, closing the control loop and supporting the SRL reflection phase (self-evaluation and calibration). To stimulate motivational processes and goal setting, we highlighted the timeline leading up to the thesis [28].

Data Collection and Outcomes

Data were collected at the end of the teaching day via individualized links to an online form (Multimedia Appendix 2). Participants provided demographics (gender, age, research experience, and progress stage of the thesis project before teaching) and completed two 7-point Likert questionnaires: self-assessed learning (3 - 8 items per objective) and satisfaction (1 item per objective). To ensure content validity and constructive alignment, all items were copied verbatim from prespecified intended learning outcomes (Multimedia Appendix 3), yielding a criterion-referenced measure. Item order was randomized.

The primary outcome was a composite score of 20 points reflecting residents’ self-assessment of knowledge and skills. It was the sum of 2 secondary outcomes: a knowledge self-assessment score (10 points) and a skill self-assessment score (10 points).

The secondary outcomes were as follows:

1. Overall satisfaction (20 points); sum of secondary outcomes 4 and 5
2. Self-assessment of knowledge acquisition (10 points), with equal weight given to each knowledge-oriented learning objective
3. Self-assessment of skill acquisition (10 points), with equal weight given to each skill-oriented learning objective
4. Satisfaction with knowledge-oriented learning objectives (10 points)
5. Satisfaction with skill-oriented learning objectives (10 points)
6. Academic achievement (submission of a thesis proposal form at 6 months; yes or no)

Statistical Analysis

Analysis was performed using R (version 4.4.3; R Foundation for Statistical Computing) in RStudio (version 2024.12.1; Posit

XSL•FO
RenderX

PBC) [29]. Because the primary outcome was assessed only at the end of the intervention, participants with missing outcome data were excluded from the analysis, resulting in a modified intention-to-treat approach with a complete-case analysis. Missing outcomes were not imputed, and reasons for missing data were documented. Multiplicity across secondary outcomes was controlled using a hierarchical fixed-sequence procedure that followed the prespecified order listed above (1 to 6). Each hypothesis was tested 2 sided at α=.05, and testing stopped at the first nonsignificant result (subsequent analyses were exploratory).

We modeled the primary outcome using multivariable linear regression, estimating the arm effect (blended vs traditional) while adjusting for a prespecified set of the following predictor variables. Precourse framing of a research question and prior research experience were included as proxies for baseline research preparedness. Full completion of another Moodle e-learning course was included as a proxy for familiarity with digital learning environments. Unblinding was included to account for potential expectancy or performance effects linked to perceived allocation. Teacher and course day were included to account for contextual instructional and organizational variability. Age and gender were included as standard demographic covariates. Model assumptions were checked (Multimedia Appendix 4). We report β coefficients with 95% CIs and adjusted R^2. For power, a sample size of 144 (72 per arm) was planned to provide 85% power to detect a 2-point difference on a 20-point scale (SD 4; 2-sided α=.05). The SD was conservatively inflated given limited prior data. Internal consistency per objective was estimated using the Cronbach α.

Ethical Considerations

This trial received ethics approval from the Comité d'Ethique de la Recherche du Collège Universitaire de Médecine Générale (CUMG-IRB 2024-04-30-02). Participants received oral and written information about the study and provided written informed consent. The data collected complied with the European Union's General Data Protection Regulation and with the data management policy of Université Claude Bernard Lyon 1 as confirmed by the university's data protection officer. No compensation was provided to participants.

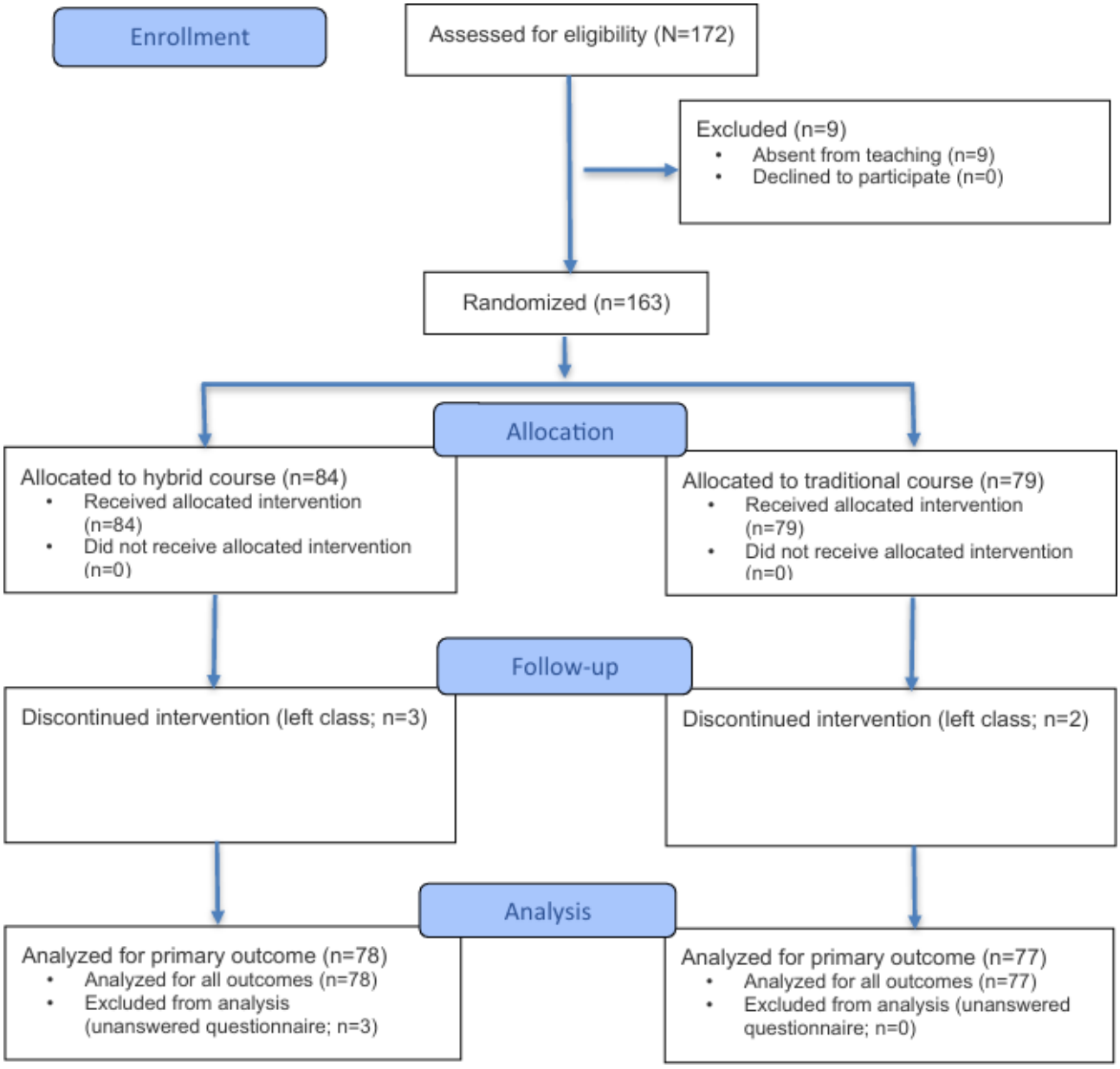
Results

Overview

We included 163 family medicine residents randomized to the blended (n=84, 51.5%) and the traditional (n=79, 48.5%) course. A total of 155 participants completed the online form at the end of the day: 78 (50.3%) in the blended course, and 77 (49.7%) in the traditional course (Figure 1). Table 2 presents the participants' characteristics. The mean age of the participants was 25.7 (SD 3.54) years, and most (112/155, 72.3%) were women. Most participants (114/155, 73.5%) had an idea for their thesis topic. A minority had framed their research question (24/155, 15.5%) or already had a thesis supervisor (22/155, 14.2%). In total, 29.7% (46/155) of the participants reported having research experience, and 3.9% (6/155) had completed a full master's degree or a PhD in addition to their medical studies. Less than a quarter of the participants (32/155, 20.6%) had already completed a full e-learning course on the university's digital learning platform (Moodle). Nearly a third of the participants (48/155, 31.0%) were unblinded at the end of the course. Five teachers offered the course, and 3 of them taught in both arms.

XSL•FO
RenderX

Figure 1. CONSORT (Consolidated Standards of Reporting Trials) flow diagram.



XSL•FO
RenderX

Table . Baseline participant demographics.

	Blended course (n=78)	Traditional course (n=77)
Age (y), mean (SD)	25.8 (4.1)	25.6 (2.8)
Gender, n (%)		
Men	19 (24.4)	22 (28.6)
Women	58 (74.4)	54 (70.1)
Nonbinary	1 (1.3)	1 (1.3)
Research experience (master's or PhD), n (%)		
None	45 (57.7)	57 (74.0)
Master's—first year	28 (35.9)	18 (23.4)
Master's—full	4 (5.1)	1 (1.3)
PhD	1 (1.3)	0 (0.0)
Achieved before the course, n (%)		
Idea for thesis subject	62 (79.5)	52 (67.5)
Research question formulated	10 (12.8)	14 (18.2)
Found a thesis supervisor	10 (12.8)	12 (15.6)
Other e-learning course completed on Moodle	21 (26.9)	11 (14.3)
Unblinded[a]	23 (29.5)	25 (32.5)

[a]Aware of group allocation.

The proportion of postrandomization exclusions was low (8/163, 4.9%) and similar in both group allocations (Multimedia Appendix 5). It was due to missing outcome data, mostly because of work-related early departure.

Primary Outcome

Knowledge and skill self-assessment did not differ significantly between the blended and traditional courses (mean difference 0.40 [maximum possible 20] points, 95% CI −0.21 to 1.02; *P*=.19; Cohen *d*=0.21; Table 3). This result remained not statistically significant after adjustment (Table 4). The most significant predictor of the primary outcome was having formulated a research question before the course, with a β coefficient of 1.87 (95% CI 1.06-2.68; *P*<.001). Unblinding during the day and teacher identity were not significantly associated with the primary outcome (*P*=.13).

XSL•FO
RenderX

Table . Education outcomes.

Outcome measure (KP[a] level)	Blended course (n=78)	Traditional course (n=77)	Mean difference (95% CI)	Cohen *d*	*P* value
Self-assessment of knowledge and skills (20 points; KP level: 2), mean (SD)	14.66 (1.92)	14.26 (1.94)	0.4 (–0.21 to 1.01)	0.21	.19[b]
Overall satisfaction (20 points; KP level: 1), mean (SD)	14.93 (2.80)	14.77 (2.55)	0.16 (–0.68 to 1.01)	0.06	.70[b,c]
Self-assessment of knowledge acquisition (10 points; KP level: 2), mean (SD)	7.49 (0.98)	7.09 (1.04)	0.4 (0.07 to 0.71)	0.38	.01[b,c]
Self-assessment of skill acquisition (10 points; KP level: 2), mean (SD)	7.18 (1.20)	7.17 (1.17)	0.01 (–0.27 to 0.27)	0.009	.76[c,d]
Satisfaction with knowledge-oriented educational objectives (10 points; KP level: 1), mean (SD)	7.54 (1.53)	7.42 (1.30)	0.12 (–0.00005 to 0.55)	0.08	.34[c]
Satisfaction with skill-oriented educational objectives (10 points; KP level: 1), mean (SD)	7.40 (1.48)	7.36 (1.46)	0.04 (–0.00007 to 0.83)	0.03	.64[c,e]
Submission of a thesis application form at 6 months (KP level: 3), n (%)	11 (14.1)	10 (13.0)	1.09 (0.39 to 3.1)[f]	—[g]	>.99[d]

[a]KP: Kirkpatrick [30].

[b]Student *t* test (2 sided).

[c]Once a hypothesis was tested and found not to be significantly different from the null hypothesis, all subsequent tests and results were considered exploratory. The secondary objectives described earlier were ranked according to this procedure.

[d]Fisher exact test (2-sided).

[e]Mann-Whitney *U* test.

[f]Odds ratio reported for this value.

[g]Not applicable.

XSL•FO
RenderX

Table . Effect of the study arm (blended or traditional course) on the primary outcome while adjusting for prespecified predictor variables.[a]

Predictor variable	β[b] (95% CI)	*P* value
Research question framed before the course (yes)	1.87 (1.06 to 2.68)	<.001
Research experience (yes)	−0.27 (–0.92 to 0.37)	.40
Unblinding (yes)	−0.52 (–1.20 to 0.16)	.13
Teacher 1	0.69 (–0.48 to 1.87)	.25
Teacher 2	0.91 (0.06 to 1.87)	.04
Teacher 3	−0.13 (–1.33 to 1.05)	.82
Teacher 4	−0.16 (–1.01 to 0.69)	.71
Another e-learning course completed on Moodle (yes)	0.09 (–0.63 to 0.82)	.79
Age	−0.08 (–0.16 to 0.01)	.07
Gender—man	0.26 (–0.40 to 0.93)	.43
Gender—nonbinary	1.41 (–1.20 to 4.03)	.28
Course day (day 2)	0.02 (–0.72 to 0.76)	.94
Arm (traditional course)	−0.23 (–0.99 to 0.53)	.55

[a]After adjustment for prespecified predictor variables, there were no significant differences between the blended and traditional courses regarding the primary outcome (*P*=.55). Given the exploratory nature of the variable selection procedure, adjusted estimates should be interpreted cautiously. R^2 is the proportion of the variance explained by the model (multiple R^2=0.21; adjusted R^2=0.13; *P*=.001 for the model).

[b]Coefficient representing the expected change in the primary outcome per 1-unit increase in the predictor variable.

Secondary Outcomes

As an exploratory result, the knowledge self-assessment was significantly higher in the blended course (mean difference 0.40 [maximum possible 10], 95% CI 0.07-0.71; *P*=.01; Cohen *d*=0.38).

The skill self-assessment by residents was not significantly different between the traditional and blended courses (*P*=.76). Their overall satisfaction or specific satisfaction regarding knowledge- or skill-oriented learning objectives did not differ between the traditional and blended courses. The submission of a thesis proposal form was also not found to be different between the 2 groups (11/78, 14.1% in the BL arm and 10/77, 13.0% in the traditional course arm).

Internal Consistency

Items assessing each educational objective showed acceptable Cronbach α values (>0.70) for all but one educational objective, which exhibited lower internal consistency (Cronbach α=0.47; Multimedia Appendix 6).

Discussion

Principal Findings

Among French family medicine residents, self-assessed knowledge and skills did not differ between the blended and traditional courses. The between-arm mean difference was small, and the 95% CI was narrow, indicating low uncertainty around the estimated difference. At 6 months, academic achievement (thesis proposal submission) did not differ between arms. The only statistically significant finding—an exploratory, slightly higher knowledge score in the blended arm—was not educationally meaningful.

Comparison With Other Studies

To our knowledge, no randomized controlled trial has evaluated whether partially replacing an already active, socioconstructivist face-to-face course with a blended format affects residents' self-reported knowledge and skills. A mixed methods randomized controlled trial compared a blended evidence-based medicine course with a lecture-based (behavioral) traditional course and found no significant difference in objectively measured knowledge and skills. However, the blended group reported higher self-efficacy, attitudes, and self-reported behaviors. Focus groups suggested a clear learner preference for the blended, integrated approach [31]. In health profession education, a meta-analysis comparing flipped classrooms with traditional, lecture-based (behavioral) courses found modest improvements in learning [32]. Flipped classrooms are a common BL configuration comparable to our study intervention combining preclass online preparation with active in-class work. Other observational studies have demonstrated the relevance of BL in medical education, most notably since the COVID-19 pandemic [33,34]. A recent meta-analysis of systematic reviews recommended further research to determine the relative benefits of BL in each individual context. It also confirmed our exploratory finding of improved knowledge acquisition with BL [35]. Moreover, a randomized controlled trial with undergraduate medical students demonstrated the noninferiority of BL vs traditional education in technical skill performance (life-saving trauma skills) and retention [36]. A randomized controlled trial of postgraduate medical students was conducted to compare BL to traditional education, and it showed significantly greater efficacy of BL for developing defibrillator

XSL•FO
RenderX

technical skills [37]. Our trial extends the evidence to academic skills related to developing a research project (thesis); it suggests that shortening face-to-face time through a structured e-learning component can yield comparable educational outcomes when the face-to-face component remains active and socioconstructivist.

More broadly, syntheses of digital education in the health professions consistently report large benefits vs no intervention [38] but little or no differences vs nondigital instruction [39].

Implications

This study may reassure teaching teams that BL can produce short-term educational outcomes comparable to those of traditional active and socioconstructivist education among residents, particularly in settings with constrained teaching capacity. Still, the implementation of BL does not address all the issues associated with the growing disparity between the number of teachers and residents. First, BL may not always be appropriate, either because of a misalignment between educational objectives aimed at developing complex competencies and the learning modality or because of the substantial resources required to develop online courses that effectively address such competencies [40]. For example, a course focused on professional communication competencies might be better suited to face-to-face relational simulation as the sole learning modality. Second, monitoring residents' acquisition of generic knowledge, skills, and competencies based on a national framework still necessitates frequent, individualized human contact. Third, strategies facilitating students' SRL may imply their supervision at the asynchronous e-learning phase, which would not allow for a significant reduction in teaching time.

The transition from traditional to blended education represents a substantial organizational shift. Teachers must allocate time to review asynchronous assessments and prepare collective feedback for the face-to-face session. It is important to estimate the time required for residents to complete all e-learning modules to avoid unintentionally increased workload. Strategies to promote SRL are also needed to optimize educational outcomes [41,42]. For example, providing dedicated time and space at the university to complete the e-learning module may help residents allocate time for it. Furthermore, working independently at home may increase feelings of stress and anxiety, as was observed during the COVID-19 pandemic [43].

Limitations and Strengths

One limitation of this study was the reliance on self-assessment rather than objective measures. Incorporating objective performance measures would enhance validity and mitigate self-report bias, for example, through standardized pretest-posttest knowledge tests and rubric-based scoring of residents' performance on key skills (framing a research question and assessing feasibility) rated by trained assessors. Still, formative self-assessment is integral to self-regulation, enabling feedback on progress relative to intended learning outcomes [44,45]. Moreover, the e-learning module was an initial version. A well-established, iteratively refined module might have produced higher scores among residents randomized to the BL arm. The e-learning module was completed immediately before the face-to-face session, which constrained residents' autonomy to plan their study and is partly misaligned with SRL. This design ensured full completion of the e-learning module, reduced the risk of early unblinding or contamination between groups before the face-to-face session, and standardized learning conditions (same setting and time on task). However, it did not capture real-life environmental variability, and the immediate sequencing in BL may also have facilitated knowledge activation for the subsequent face-to-face session. Although a high proportion of participants were unblinded by the end of the teaching day—a potential limitation—this variable was not associated with the primary outcome in multivariate analysis. The complete-case analysis may have introduced attrition bias; however, the proportion of missing outcome data remained low (8/163, 4.9%) and similar in both group allocations.

Regarding strengths, the narrow 95% CI suggests that an educationally meaningful difference of ≥10% (≥2/20 points) between groups would have been detected even though the study was not designed to test equivalence or noninferiority. A formal noninferiority trial is still needed. Furthermore, the lack of significant teacher effects on the primary outcome indicates that the standardized, timed script effectively minimized instructor-specific variability.

Conclusions

Self-assessment of knowledge and skills by family practice residents did not differ between blended and traditional courses. This study can reassure educators that BL may produce short-term educational outcomes (knowledge and academic skills) comparable to those of traditional active and socioconstructivist education among residents, particularly in settings with constrained teaching capacity.

Acknowledgments

The authors would like to thank Nathalie Gras (school administrator) for her help in organizing the courses (resident registration and scheduling).

Funding

No external financial support or grants were received from any public, commercial, or not-for-profit entities for the research, authorship, or publication of this paper.

XSL•FO
RenderX

Data Availability

The datasets generated or analyzed during this study are available from the corresponding author on reasonable request.

Authors' Contributions

Conceptualization: PYM (lead), FZ (supporting), TC (supporting), RB (supporting), ML-B (lead)
Data curation: PYM
Formal analysis: PYM (equal), SS (equal)
Investigation: PYM (equal), FZ (equal), TC (equal), RB (equal), ML-B (equal)
Methodology: PYM (lead), SS (supporting), FZ (supporting), TC (supporting), RB (supporting), HM (supporting), ML-B (supporting)
Project administration: PYM (equal), ML-B (equal)
Resources: PYM
Supervision: ML-B (lead), SS (supporting), HM (supporting)
Validation: SS (equal), HM (equal), ML-B (equal)
Visualization: PYM
Writing—original draft: PYM
Writing—review and editing: PYM (equal), SS (equal), JM (equal), FZ (equal), TC (equal), RB (equal), ML-B (equal)
All authors meet all the International Committee of Medical Journal Editors criteria, approved the final version of the manuscript, and agree to be accountable for all aspects of the work.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Screenshots of the e-learning module.
[DOCX File, 1677 KB - mededu_v12i1e86387_app1.docx]

Multimedia Appendix 2
Data collection form.
[DOCX File, 994 KB - mededu_v12i1e86387_app2.docx]

Multimedia Appendix 3
Questionnaire items for self-assessment and satisfaction (verbatim wording from intended learning outcomes).
[DOCX File, 21 KB - mededu_v12i1e86387_app3.docx]

Multimedia Appendix 4
Multiple linear regression model: checking assumptions.
[DOCX File, 86 KB - mededu_v12i1e86387_app4.docx]

Multimedia Appendix 5
Reasons for postrandomization exclusion in complete-case analysis.
[DOCX File, 14 KB - mededu_v12i1e86387_app5.docx]

Multimedia Appendix 6
Internal consistency of family medicine residents' self-assessment: Cronbach α values for each educational objective.
[DOCX File, 15 KB - mededu_v12i1e86387_app6.docx]

Checklist 1
CONSORT checklist.
[PDF File, 84 KB - mededu_v12i1e86387_app7.pdf]

References

1. Atun R. What are the advantages and disadvantages of restructuring a health care system to be more focused on primary care services? World Health Organization. 2004. URL: https://iris.who.int/handle/10665/363801 [accessed 2024-08-20]

XSL•FO
RenderX

2. Russo G, Perelman J, Zapata T, Šantrić-Milićević M. The layered crisis of the primary care medical workforce in the European region: what evidence do we need to identify causes and solutions? Hum Resour Health 2023 Jul 14;21(1):55. [doi: 10.1186/s12960-023-00842-4] [Medline: 37443059]
3. Majeed A. Shortage of general practitioners in the NHS. BMJ 2017 Jul 10;358:j3191. [doi: 10.1136/bmj.j3191] [Medline: 28694250]
4. Dall T, Reynolds R, Chakrabarti R, Ruttinger C, Zarek P, Parker O. The complexities of physician supply and demand: projections from 2021 to 2036. : Association of American Medical Colleges; 2024 URL: https://www.aamc.org/media/75236/download?attachment [accessed 2026-01-31]
5. Framework for action on the health and care workforce in the WHO European region 2023–2030. World Health Organization, European Region. 2023. URL: https://iris.who.int/bitstream/handle/10665/372563/73wd08e-HealthCareWorkforce-230575.pdf?sequence=5 [accessed 2024-08-15]
6. Petterson SM, Liaw WR, Tran C, Bazemore AW. Estimating the residency expansion required to avoid projected primary care physician shortages by 2035. Ann Fam Med 2015 Mar;13(2):107-114. [doi: 10.1370/afm.1760] [Medline: 25755031]
7. Chevillard G, Lucas-Gabrielli V, Mousques J. « Déserts médicaux » en France: état des lieux et perspectives de recherches. L'Espace géographique 2018 Dec 18;Tome 47(4):362-380. [doi: 10.3917/eg.474.0362]
8. Mouthon L, Clavelou P. Réforme du troisième cycle des études médicales en France. Presse Méd Form 2024 Apr;5(2):97-107. [doi: 10.1016/j.lpmfor.2024.02.001]
9. Compagnon L, Bail P, Huez JF, et al. Définitions et descriptions des compétences en médecine générale. Exercer 2013;24(108):148-155 [FREE Full text]
10. Kulasegaram K, Mylopoulos M, Tonin P, et al. The alignment imperative in curriculum renewal. Med Teach 2018 May;40(5):443-448. [doi: 10.1080/0142159X.2018.1435858] [Medline: 29490525]
11. Bay E, Bagceci B, Cetin B. The effects of social constructivist approach on the learners' problem solving and metacognitive levels. J Soc Sci 2012;8(3):343-349. [doi: 10.3844/jssp.2012.343.349]
12. Vygotsky L, Cole M, John-Steiner V, Scribner S, Souberman E. Mind in Society: The Development of Higher Psychological Processes: Harvard University Press; 1978.
13. Heck AJ, Cross CE, Tatum VY, Chase AJ. Active learning among health professions' educators: perceptions, barriers, and use. Med Sci Educ 2023 May 2;33(3):719-727. [doi: 10.1007/s40670-023-01793-0] [Medline: 37501807]
14. Tudor Car L, Poon S, Kyaw BM, et al. Digital education for health professionals: an evidence map, conceptual framework, and research agenda. J Med Internet Res 2022 Mar 17;24(3):e31977. [doi: 10.2196/31977] [Medline: 35297767]
15. Bonk CJ, Graham CR. The Handbook of Blended Learning: Global Perspectives, Local Designs: John Wiley & Sons; 2012.
16. Saichaie K. Blended, flipped, and hybrid learning: definitions, developments, and directions. New Direct Teach Learn 2020;2020(164):95-104. [doi: 10.1002/tl.20428]
17. Hurtubise L, Hall E, Sheridan L, Han H. The flipped classroom in medical education: engaging students to build competency. J Med Educ Curric Dev 2015;2:JMECD.S23895. [doi: 10.4137/JMECD.S23895] [Medline: 35187252]
18. Liu Q, Peng W, Zhang F, Hu R, Li Y, Yan W. The effectiveness of blended learning in health professions: systematic review and meta-analysis. J Med Internet Res 2016 Jan 4;18(1):e2. [doi: 10.2196/jmir.4807] [Medline: 26729058]
19. Vallée A, Blacher J, Cariou A, Sorbets E. Blended learning compared to traditional learning in medical education: systematic review and meta-analysis. J Med Internet Res 2020 Aug 10;22(8):e16504. [doi: 10.2196/16504] [Medline: 32773378]
20. Moskal PD. Evaluating the outcomes and impact of hybrid courses. New Direct Teach Learn 2017;2017(149):19-26. [doi: 10.1002/tl.20223]
21. Mason GS, Shuman TR, Cook KE. Comparing the effectiveness of an inverted classroom to a traditional classroom in an upper-division engineering course. IEEE Trans Educ 2013 Nov;56(4):430-435. [doi: 10.1109/TE.2013.2249066]
22. Lage MJ, Platt GJ, Treglia M. Inverting the classroom: a gateway to creating an inclusive learning environment. J Econ Educ 2000 Jan;31(1):30-43. [doi: 10.1080/00220480009596759]
23. AlSheikh MH, Iqbal MZ. Student perceptions regarding group learning activities in a hybrid medical curriculum. Med Sci Educ 2019 Sep 16;29(4):1221-1228. [doi: 10.1007/s40670-019-00817-y] [Medline: 34457606]
24. Hopewell S, Chan AW, Collins GS, et al. CONSORT 2025 statement: updated guideline for reporting randomised trials. BMJ 2025 Apr 14;389:e081123. [doi: 10.1136/bmj-2024-081123] [Medline: 40228833]
25. Create a blocked randomisation list. Sealed Envelope. 2024. URL: https://www.sealedenvelope.com/simple-randomiser/v1/lists [accessed 2024-05-14]
26. Schunk DH, Zimmerman BJ. Self-Regulation of Learning and Performance: Issues and Educational Applications: Routledge; 1994.
27. Zimmerman BJ. Becoming a self-regulated learner: an overview. Theory Pract 2002;41(2):64-70. [doi: 10.1207/s15430421tip4102_2]
28. Gavarkovs AG, Finan E, Jensen RD, Brydges R. When I say … active learning. Med Educ 2024 Aug;58(8):896-897. [doi: 10.1111/medu.15383] [Medline: 38503311]
29. The data science code editor. Posit Software. URL: http://www.posit.co/ [accessed 2026-01-24]
30. Kirkpatrick DL, Kirkpatrick JD. Evaluating Training Programs: The Four Levels: Berrett-Koehler Publishers; 2006.

XSL•FO
RenderX

31. Ilic D, Nordin RB, Glasziou P, Tilson JK, Villanueva E. A randomised controlled trial of a blended learning education intervention for teaching evidence-based medicine. BMC Med Educ 2015 Mar 10;15:39. [doi: 10.1186/s12909-015-0321-6] [Medline: 25884717]
32. Hew KF, Lo CK. Flipped classroom improves student learning in health professions education: a meta-analysis. BMC Med Educ 2018 Mar 15;18(1):38. [doi: 10.1186/s12909-018-1144-z] [Medline: 29544495]
33. Atwa H, Shehata MH, Al-Ansari A, et al. Online, face-to-face, or blended learning? Faculty and medical students' perceptions during the COVID-19 pandemic: a mixed-method study. Front Med (Lausanne) 2022;9:791352. [doi: 10.3389/fmed.2022.791352] [Medline: 35186989]
34. Abdull Mutalib AA, Md Akim A, Jaafar MH. A systematic review of health sciences students' online learning during the COVID-19 pandemic. BMC Med Educ 2022 Jul 3;22(1):524. [doi: 10.1186/s12909-022-03579-1] [Medline: 35786374]
35. Lockey A, Bland A, Stephenson J, Bray J, Astin F. Blended learning in health care education: an overview and overarching meta-analysis of systematic reviews. J Contin Educ Health Prof 2022 Oct 1;42(4):256-264. [doi: 10.1097/CEH.0000000000000455] [Medline: 36070399]
36. Bergmans E, Billington A, Thies KC. From tradition to innovation: a comparison of the traditional 4-step approach versus a blended learning modification for technical skills teaching. Scand J Trauma Resusc Emerg Med 2023 Nov 14;31(1):80. [doi: 10.1186/s13049-023-01127-4] [Medline: 37964361]
37. Siebert JN, Glangetas A, Grange M, Haddad K, Courvoisier DS, Lacroix L. Impact of blended learning on manual defibrillator's use: a simulation-based randomized trial. Nurs Crit Care 2022 Jul;27(4):501-511. [doi: 10.1111/nicc.12713] [Medline: 34519140]
38. Cook DA, Levinson AJ, Garside S, Dupras DM, Erwin PJ, Montori VM. Internet-based learning in the health professions: a meta-analysis. JAMA 2008 Sep 10;300(10):1181-1196. [doi: 10.1001/jama.300.10.1181] [Medline: 18780847]
39. Vaona A, Banzi R, Kwag KH, et al. E-learning for health professionals. Cochrane Database Syst Rev 2018 Jan 21;1(1):CD011736. [doi: 10.1002/14651858.CD011736.pub2] [Medline: 29355907]
40. Wurth S, Maisonneuve H, Moussa MA, et al. Development of complex pedagogical competencies and reflexivity in clinical teachers via distance learning: a mixed methods study. Med Educ Online 2023 Dec;28(1):2265163. [doi: 10.1080/10872981.2023.2265163] [Medline: 37818594]
41. Azevedo R. Using hypermedia as a metacognitive tool for enhancing student learning? The role of self-regulated learning. Educ Psychol 2005;40(4):199-209. [doi: 10.1207/s15326985ep4004_2]
42. Schumacher C, Ifenthaler D. Features students really expect from learning analytics. Comput Human Behav 2018 Jan;78:397-407. [doi: 10.1016/j.chb.2017.06.030]
43. Patte KA, Battista K, Ferro MA, et al. School learning modes during the COVID-19 response and pre- to during pandemic mental health changes in a prospective cohort of Canadian adolescents. Soc Psychiatry Psychiatr Epidemiol 2024 Jan;59(1):137-150. [doi: 10.1007/s00127-023-02557-2] [Medline: 37668673]
44. Brown GT, Harris LR. The future of self-assessment in classroom practice: reframing self- assessment as a core competency. Frontline Learn Res 2014 Apr;3:22-30. [doi: 10.14786/flr.v2i1.24]
45. Andrade HL. Students as the definitive source of formative assessment: academic self-assessment and the self-regulation of learning. In: Handbook of Formative Assessment: Routledge; 2010.

Abbreviations

BL: blended learning
CONSORT: Consolidated Standards of Reporting Trials
SRL: self-regulated learning

Edited by LT Car; submitted 23.Oct.2025; peer-reviewed by J Wilkinson, M Arsenault; accepted 12.Jan.2026; published 04.Feb.2026.

Please cite as:
Meunier PY, Schlatter S, Macabrey J, Zorzi F, Colleony T, Boussageon R, Maisonneuve H, Lamort-Bouché M
Blended Learning Compared With Face-to-Face Learning Among Family Medicine Residents: Randomized Controlled Trial
JMIR Med Educ 2026;12:e86387
URL: https://mededu.jmir.org/2026/1/e86387
doi:10.2196/86387

© Pierre-Yves Meunier, Sophie Schlatter, Juliette Macabrey, Frédéric Zorzi, Thomas Colleony, Rémy Boussageon, Hubert Maisonneuve, Marion Lamort-Bouché. Originally published in JMIR Medical Education (https://mededu.jmir.org), 4.Feb.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium,

XSL•FO
RenderX

provided the original work, first published in JMIR Medical Education, is properly cited. The complete bibliographic information, a link to the original publication on https://mededu.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Students' Learning Outcomes in a Virtual Microscopic Online and Face-to-Face Anatomy Course: Cross-Over Comparative Study

Katharina Sternecker[1], PhD; Christoph Schmitz[1], MD, PhD; Markus Berndt[2], PhD

[1]Department of Anatomy II, LMU Medizin, Ludwig-Maximilians-Universität München, Pettenkoferstr. 11, Munich, Germany

[2]Institute of Medical Education - LMU University Hospital, LMU Medizin, Ludwig-Maximilians-Universität München, Munich, Germany

Corresponding Author:
Katharina Sternecker, PhD
Department of Anatomy II, LMU Medizin, Ludwig-Maximilians-Universität München, Pettenkoferstr. 11, Munich, Germany

Abstract

Background: Since 2021, the histology of organs in the microscopic anatomy course at the Department of Anatomy II of the Ludwig-Maximilians-Universität Munich has been held as a hybrid course using a 3D virtual microscopy system. Evidence comparing synchronous online versus face-to-face instruction using an advanced 3D virtual slide system remains limited. In particular, it is unclear whether learning outcomes are equivalent across instructional modes, and whether individual student characteristics influence learning outcomes under these conditions.

Objective: This study systematically (1) compares the learning outcomes of preclinic students attending a microscopic anatomy course either face-to-face or synchronously online using a 3D virtual microscopy slide box generated from fully scanned histological glass slides previously used in a traditional course, and (2) investigates whether selected demographic characteristics and personal skills predict learning outcomes in either instructional mode. We hypothesized that learning outcomes would not differ between online and face-to-face participation when instructional conditions are identical.

Methods: The students attended 2 consecutive course days on 2 topics (kidney and liver) in a crossover design. Half of the students attended the first day face-to-face, and the second day online, while the other half attended in the reverse order. Demographic characteristics (eg, self-organization and affinity for technology interaction) and learning outcomes were collected via survey and simulated examination on the day after the second course day. To minimize potential confounding factors (eg, learning by means of external resources) and to ensure that the results reflected the learning outcomes within the scope of this study, 2 particularly challenging course days were intentionally selected, and the time interval before the simulated examination was kept short. Learning outcomes were analyzed using a linear mixed-effects model, and potential predictors were explored using stepwise linear regression.

Results: Overall, 740 student questionnaires and examinations (740/864, 86% of the total student cohort) were used to assess the items' quality of the simulated examination. Inferential analyses included data from 415 students who completed the crossover design as intended. No significant main effect of participation mode on learning outcomes was found ($F_{1,\ 413}$=0.16; P=.69). Although a significant effect of examination topic was observed ($F_{1,\ 413}$=25.56; P<.001), there was no significant interaction between participation mode and examination topic ($F_{1,\ 413}$=0.13; P=.72). Regression analyses showed low predictive power overall (R^2≥0.12), with self-organization emerging as a weak but consistent predictor across both instructional modes.

Conclusions: The results of this study suggest that the microscopic anatomy learning outcomes of students under controlled short-term conditions are comparable in synchronous online and face-to-face teaching. However, it should be noted that the study was limited to 2 topics and 2 consecutive course days. Further research should address additional topics, long-term learning outcomes, and broader student populations.

(JMIR Med Educ 2026;12:e82012) doi:10.2196/82012

KEYWORDS

virtual slides; hybrid course; learning outcomes; demographics; blended learning

Introduction

In recent years, 2 developments have taken place in Germany that were challenging for both learners and instructors at universities: (1) the continuous advancement of digital teaching and (2) the growing heterogeneity of students. Students at universities in Germany are becoming increasingly heterogeneous regarding demographic factors such as socioeconomic background, ethnic origin, financing of studies, economic situation, employment, housing situation, or personal and study-related factors such as prior educational background, digital literacy, time commitments (eg, part-time work and

XSL•FO
RenderX

family responsibilities), study experience, and learning strategies [1-3]. There is a steadily growing demand for distance learning offerings at universities worldwide, particularly if there are in-person class locations that can only be reached with considerable time effort by students, for example, when locations are hard to access because of long travel distances or densely scheduled timetables [4-6]. This growing diversity poses substantial challenges for curriculum design, particularly in demanding and content-dense disciplines such as anatomy.

During the COVID-19 pandemic, medical schools worldwide switched to remote learning in the preclinical phase, and many subjects, including microscopic anatomy, were taught online [7]. As it could be shown that the online formats were well received by students and instructors, they were partially retained. Overall, constantly evolving technological possibilities have led to an increased use of virtual slides (ie, digital copies of glass slides) for teaching microscopic anatomy in the laboratory [8,9], and it has even been observed that virtual microscopy (VM) has a slight advantage over optical microscopy (OM) in teaching microscopic anatomy [10,11]. Presumably, this is due to the greater accessibility of the technology for students compared to conventional microscopy. Furthermore, the shared access to identical specimens supports collaborative learning among students. Virtual 2D microscopy is well accepted and widely used in microscopic anatomy courses, as it shows good results regarding implementation, relevance, and educational effectiveness of microscopic anatomy [12,13], but virtual 3D microscopy is generally not used because of the considerable effort required to produce virtual 3D slides. Moreover, online teaching is viewed with skepticism, as negative experiences regarding learning success in anatomy have been reported in online teaching of anatomy [14]. Beyond technical challenges, insufficient peer interaction and the limited use of online platforms for communication were identified. Holland et al [15] showed that students who attended fewer face-to-face classes than their peers were also less active in online offerings. In this study, student attendance was correlated with their learning success. The online offerings were supplementary to the face-to-face classes, meaning it was not examined whether students would have shown a greater willingness to participate in the lessons if the entire traditional face-to-face course had been offered online. This is important because time spent on microscopy of histological specimens is essential for the later transfer of knowledge [16].

It can be seen that in such far-reaching changes like switching from face-to-face to online teaching, oftentimes insufficient attention is directed towards the existence of specific prerequisites or personality traits of students that enable them to adapt to the new conditions effectively. For example, it is not clear whether online teaching is used equally by all students or can be used by all students, and how the individual skills and personal environment of students affect their learning success in online teaching. It has been demonstrated that the social environment outside the university and the personal abilities of students can influence their learning success and, consequently, their academic success [17]. So far, there is limited knowledge about which groups of students require additional learning opportunities in addition to the offered online teaching, as students may not have the same skills and/or comparable social or technical prerequisites. There is insufficient understanding of students whose learning success may improve through online teaching (eg, students who have to work alongside their studies to support themselves), so that they can be adequately supported through appropriate instruction. It has been found, for example, that academically weaker students are more affected by a reduction in the histology curriculum than stronger students [18]. Therefore, the question arises whether the instructional mode, online or face-to-face, influences learning success depending on sociodemographic characteristics and personal factors. In addition to that, existing studies investigated teaching in online courses or additional online offerings alongside the face-to-face course in microscopic anatomy by examining entire courses conducted over several months. While such designs provide valuable insights into overall course effectiveness, they also introduce numerous potential confounders, including prolonged self-study, intensive use of external learning resources, formation of peer learning groups, and cumulative assessment effects. As a result, it remains difficult to isolate the specific contribution of instructional modality (online vs face-to-face) to observed learning outcomes.

Theoretical frameworks such as aptitude-treatment interaction (ATI) studies emphasize that learning outcomes are not only determined by instructional methods, but by interactions between learner aptitudes and instructional conditions. ATI is related to complex patterns on its own but also depends on the desired learning content and sociodemographic characteristics [19]. In this context, it is generally assumed that weaker and anxious students are positively influenced in their learning success by a well-organized learning environment, whereas stronger, nonanxious students may be hindered by it. From this perspective, these studies focus on connections between learning success and sociodemographic characteristics or personal skills. In relation to hybrid classes, it is important to know whether any of the implemented formats advantage or disadvantage particular student groups and whether additional support mechanisms are required to ensure equitable learning opportunities.

At Ludwig-Maximilians-Universität München (Ludwig Maximilian University [LMU], Munich), since 2021, the course on microscopic anatomy has been conducted in a hybrid format (ie, face-to-face and synchronously streamed online participation options) from course days 13 to 26 of the microscopic anatomy course. This approach ensures that students who followed the online option received the same instructions as those attending face-to-face. Online students can communicate with the instructor and/or assistants via chat or audio; students attending face-to-face can talk to the instructors directly. However, it remained unclear whether the students' learning outcomes in the online mode are comparable to face-to-face participation in the microscopic anatomy course using a virtual 3D slide box and whether individual demographic characteristics or personal skills meaningfully predict learning outcomes in either modality.

To address this gap, the aim of this study was (1) to investigate the learning outcomes of the students comparing the online participation with face-to-face attendance using Histologi@ as a virtual 3D slide box in microscopic anatomy and (2) to identify

XSL•FO
RenderX

demographic characteristics and personal skills (eg, affinity for technology interaction) that may be able to predict the learning outcomes of medical students attending the microscopic anatomy course face-to-face or online. By focusing on a short, well-defined instructional period to control for confounding factors and by explicitly examining demographic characteristics and personal skills within an ATI framework, this study aimed to contribute robust empirical evidence to the ongoing discussion about the role of online and hybrid teaching in microscopic anatomy education.

Methods

Study Design

The study took place at the end of the microscopic anatomy course in the study year 2023/2024 at the Department of Anatomy of LMU Munich and consisted of 2 course days (28 and 29) and one day for data collection. The course days were spread over several days; the data collection took place in one day (Table 1).

Table . Cross-over design of this study: Class A-F represents the 6 classes of the students' cohort taking the microscopic anatomy course during data collection; the students participated on 2 course days with different topics: course day 28 (topic 1=kidney) and course day 29 (topic 2=liver).

Course day and participation mode	Class A	Class B	Class C	Class D	Class E	Class F
Course day 28 (online)	A[a]	B[b]	A	B	A	B
Course day 28 (face-to-face)	B	A	B	A	B	A
Course day 29 (online)	B	A	B	A	B	A
Course day 29 (face-to-face)	A	B	A	B	A	B

[a]A: the students participated online first and then face-to-face afterwards.

[b]B: the students participated face-to-face first and then online afterwards.

The students in each class were divided into 2 groups. On course day 28, one group (Group A) participated online, while the other group (Group B) attended face-to-face. On course day 29, the groups switched formats. Group A and Group B were always instructed synchronously. Since the student cohort was split up to 6 classes, the course days 28 and 29, as well as the data collection (evaluation and simulated examination), were repeated 6 times.

Course days 28 (topic 1: kidney) and 29 (topic 2: liver) were conducted by trained instructors (one instructor per course day) and 2 to 3 additional teaching staff members per online or face-to-face condition, using the sandwich model to ensure well-defined teaching [20]. In the sandwich model, short interactive learning phases (active elements) are embedded between classic lecture segments to enhance attention and learning effectiveness. The content of each section is precisely scheduled with specified time allocations. In 10 sessions of the 2 course days, the first author served as the instructor, while in 2 sessions, the lead was taken by 2 instructors with experience in teaching microscopic anatomy. All instructors received a detailed briefing on the sandwich model from KS and used the same course materials on the course days.

The course days were held in the microscopic laboratories but were broadcast via Zoom (Zoom Communications, Inc). Consequently, the students received identical instructions, regardless of the type of participation. The class allocation was adopted from the dean's office and was not modified for the study. Students independently chose in isolated cases to participate in a different group. Therefore, class attendance and the mode of participation were additionally assessed in the evaluation.

Data Collection

The data collection took place face-to-face one day after course day 29 in the microscopic laboratories, and the instructor, as well as teaching staff members, monitored the data collection (course day 28 took place 3 or 4 days before course day 29, depending on the schedule). The students were not kept from independently using the Histologi@ system or engaging with the course materials in the period between the course days and the simulated examination and the evaluation, as data collection took place during the regular semester.

The evaluation survey and examination were conducted using EvaSys (evasys GmbH), and the students could participate digitally via a QR code. On average, students took about 7.5 (SD 0.4) min to complete the evaluation, immediately followed by a 30-minute time slot to answer the examination questions.

Simulated Examination and Instruments

To each single-choice question (5 answer options), a carefully selected 2D image section from the virtual 3D slides shown in both course days was assigned and projected onto the board for one minute. During that one minute, the students were instructed to submit their response digitally via EvaSys. In order for the students to answer the question correctly, the students first had to identify the image, and then answer a question related to that specific image section. The students were not informed in advance which image sections from the virtual 3D slides would be used for the simulated examination.

The simulated examination consisted of 30 single-choice questions, 15 referring to course day 28 (kidney) and 15 referring to course day 29 (liver). The questions corresponded to the standard of the final examination for the course in

XSL•FO
RenderX

microscopic anatomy at the Department of Anatomy. All examination items were developed by subject matter experts through group consensus.

The evaluation assessed demographic characteristics (3 items on age, commuting time, and working hours to finance studies), learning preferences (7 items on, eg, self-organization, preference for practical work, and attitudes toward group work), and affinity for technology interaction (9 items), all on a 5-point Likert scale (1=not correct at all; 5=completely correct). Detailed evaluation items are provided in Table S1 in Multimedia Appendix 1.

Commuting time was considered particularly relevant because medical education at LMU does not take place on a centralized campus; instead, students often have to travel between different teaching locations and may spend several hours per day commuting in order to attend various classes. As digital tools are playing an ever-growing role in teaching, and students' technical affinity may influence their learning outcomes, and as online participation with virtual microscopy may therefore present a challenge, this item was included as well. The affinity for technology interaction was calculated according to Franke et al [21].

Only fully completed evaluations were included in the analysis. The survey items were reviewed by the ethics committee and revised where necessary.

Microscopic Anatomy Course at the Department of Anatomy at LMU Munich

864 students attended the microscopic anatomy course in the study year 2023/2024. Due to the high number of students, the cohort was split up into 6 classes (A-F), each with approximately 144 students. The allocation of the students into the classes was carried out randomly at the beginning of the study year by the dean's office of studies. Sixty-eight workstations in the main microscopic anatomy laboratory and 21 workstations in a second smaller microscopic anatomy laboratory were provided at the Department of Anatomy for face-to-face participation in the microscopic anatomy course for students who wanted to attend the microscopic anatomy course face-to-face. At each workstation, 1 - 2 students can use the virtual 3D slide box of Histologi@ with a Mac-computer (M1 Mac and iMac, Apple Inc) with macOS High Sierra or macOS Monterey (Apple Inc) for virtual 3D microscopy.

The microscopic anatomy course at the Department of Anatomy of the LMU Munich consists of 29 course days and includes cell biology and tissue science (course days 1 - 15), as well as the histology of the organs (course days 16 - 29). The students attended the course days 1 - 15 exclusively in face-to-face mode. From course day 15 until course day 29, the course was conducted in a hybrid format (ie, face-to-face and online). For students who wished to participate online, the face-to-face classes were streamed synchronously via Zoom from the main microscopic laboratory, and they could use Histologi@ with Biolucida Viewer (MBF Bioscience) on their own personal devices. This approach ensured that students who chose the online mode received the same instruction as those attending face-to-face. Online students had the option to communicate with the instructor and/or additional teaching staff members via chat or audio, which allows the teaching staff an interactive supervision as is common in the microscopic anatomy laboratory.

Virtual Microscopy System

The students have access to the virtual 3D slide box of Histologi@ via MBF server software, which reads the 3D-files and compiles the data stream necessary for the visualization in the Biolucida Cloud Viewer 2020.1.0 (MBF Bioscience) [22].

Via cloud (Leibniz-Rechenzentrum der Bayerischen Akademie der Wissenschaften, Garching, Germany), students can access Histologi@ not only at the workstations, but also with their own personal devices, and they study the virtual 3D slide box location-independently.

The students used Histologi@ to learn microscopic anatomy by studying 197 virtual 3D slides. A distinctive feature of the program is that the glass slide collection was digitized with a research-grade microscope, including not only the x- and y-axis information but also the z-axis (ie, several overlying focus planes of the same slide), resulting in file sizes of the individual 3D virtual slides. Therefore, students can directly magnify the virtual slides and change the focus plane, as was previously common practice with analog microscopy.

All 3D virtual slides of Histologi@ can be obtained for free from the authors of this study. The content-related use of the virtual slides and the theoretical content of the course is summarized in lecture notes available free of charge and in a book [23,24].

Ethical Considerations

The study was reviewed and approved by the Ethics Committee of LMU Munich, Munich, Germany (Project reference number 22 - 0152 KB, February 14, 2022). The committee confirmed that the study raised no ethical or legal concerns in accordance with the Declaration of Helsinki, § 15 of the Professional Code for Physicians in Bavaria, and applicable faculty regulations.

Data collection for both the evaluation and the examination was blinded and entirely anonymous, with no personal identifiers recorded, thereby precluding any possibility of tracing responses to individual participants or influencing course completion outcomes. However, students readily accepted participation in the study as a means to support the creation of scientific evidence for future teaching formats and as a possibility to receive formative feedback on their learning progress. They had to confirm this through their active consent in the survey prior to the start of data collection. The students did not receive any monetary or other form of compensation for their participation.

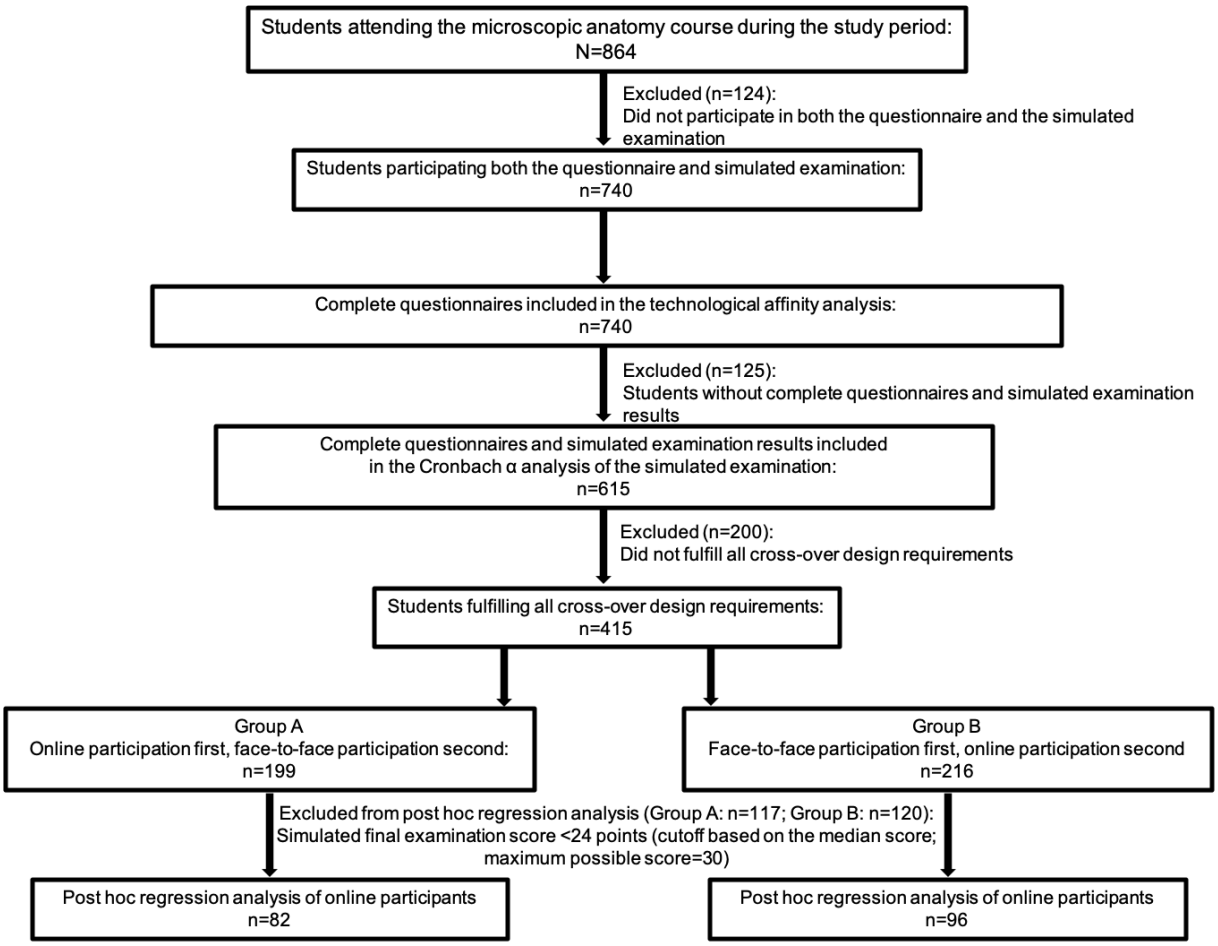
Sample Characteristics and Data Inclusion

Only questionnaires were used in which the respective sections were fully completed. A total of 740 students completed both the questionnaire and the simulated examination, corresponding to approximately 86% (740/864) of the total cohort. Of these, 415 students fulfilled all requirements of the cross-over design and were included in the inferential statistical analysis. The

XSL•FO
RenderX

selection of data included in the final analyses is illustrated in Figure 1.

Figure 1. Flowchart of participant inclusion and analysis.



Among the analyzed sample, 199 students (Group A) participated online first and face-to-face second, while 216 students (Group B) followed the reverse order.

Internal consistencies for the examination (30 items) and the technological affinity (9 items) were calculated using Cronbach α.

To calculate Cronbach α for the simulated examination, 615 questionnaires were included. For the calculation of the level of technological affinity, 740 questionnaires were used.

Statistical Analysis

All calculations and graphics were performed using SPSS (IBM Corp). Cronbach α on testing the questions 1 - 15 (topic 1=kidney) was 0.75, on testing the questions 16 - 30 it was 0.75 (topic 2=liver), and overall (questions 1 - 30) 0.85 (n=615). Therefore, all questions of the examination to evaluate the learning outcomes and all items to determine the level of technological affinity were used in the analysis.

Cronbach α on testing the 9 items to determine the level of technological affinity was 0.87 (n=740), signifying sufficient internal consistency.

Influence of the demographic characters and the personal skills were calculated using linear stepwise regressions. The stepwise regression analysis to predict the influence of demographic characteristics or personal skills was performed at four different levels: (1) across all learning outcomes; (2) across the learning outcomes for online and face-to-face participation on both course days; (3) across the learning outcomes of Group A, online and face-to-face participation, separately; and (4) across the learning outcomes of Group B, online and face-to-face participation, separately.

Additionally, a post hoc analysis of the online participants subgroup was conducted to allow for better identification of potential factors associated with lower outcomes after online instructions in greater detail. This analysis was considered exploratory and was performed to ensure that potential associations between demographic characteristics, personal skills, and learning outcomes in students who achieved lower scores in the simulated final examination after online instruction were not missed.

A cutoff of <24 points was applied because it corresponds to the median score achieved in the simulated final examination (maximum possible score=30). This approach was selected to differentiate between lower- and higher-performing students, as the median represents an objective cutoff and ensures balanced subgroup sizes.

Furthermore, separate regressions were conducted for this subgroup for course day 28 (n=82) and course day 29 (n=96),

XSL•FO
RenderX

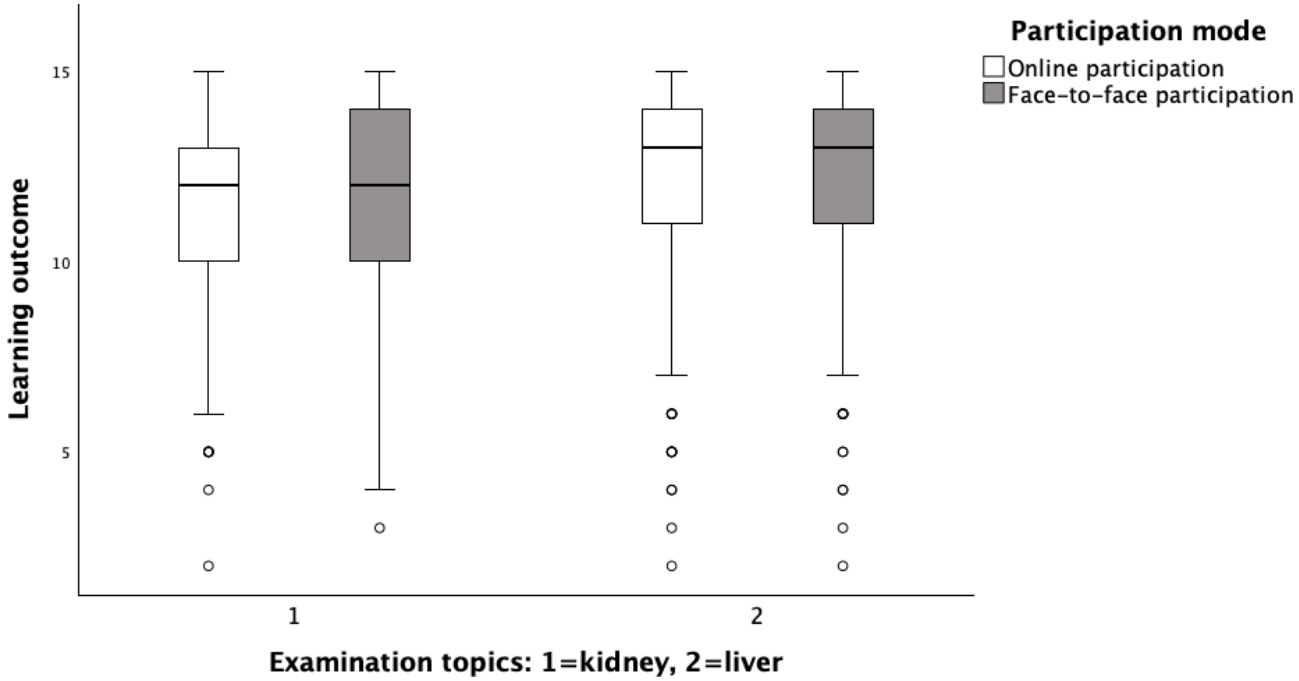
respectively. The comparison between the learning outcomes of online and face-to-face participation was analyzed using a linear mixed-effects model with restricted maximum likelihood estimation and an unstructured covariance matrix. Examination results were specified as the dependent variable. Participation mode (online vs face-to-face), course topic (liver vs kidney), and their interaction were included as fixed effects. The individual student was specified as the subject variable, and course topic day was modeled as the repeated factor.

Results

Effect of Participation Mode and Examination Topic

The linear mixed-effects model, including participation mode (online vs face-to-face) and course topic (liver vs. kidney) as fixed effects and examination results as dependent variable, revealed no significant main effect of participation mode ($F_{1, 413}$=0.16; P=.69) and a significant main effect of the examination topic ($F_{1, 413}$=25.56; P<.001). No significant interaction between participation mode and examination topic was found ($F_{1, 413}$=0.13; P=.72; Figure 2).

Figure 2. Box plots of medical students' learning outcomes after online or face-to-face participation on course day 28 (kidney; online: n=199, face-to-face: n=216) and course day 29 (liver; online: n=216, face-to-face: n=199). Learning outcomes differed significantly between the 2 topics (linear mixed-effects model with an unstructured covariance matrix: $F_{1, 413}$=25.56; P<.001); points indicate outliers; whiskers represent IQR.



The means and SDs for the learning outcomes were 11.4 (SD 2.7) for course day 28 and 11.9 (SD 2.6) for course day 29. When comparing learning outcomes for online and face-to-face participation, the mean values were 11.6 (SD 2.7) for online and 11.7 (SD 2.6) for face-to-face participation.

Predictors of Learning Outcomes

The model with the highest predictive probability was at level 4 (Table 2): the model in which the data from course day 28 of the online participants were used achieved a predictive probability of 12% with 3 predictors (commuting time, self-organization, and working hours per week; P<.001). The predictive probability was lower in the 3 other models at this level: for course day 28 with in-person participants, the predictive probability was 4% (P<.001); for course day 29 with online participants, it was 9% (P<.001); and for course day 29 with in-person participants, it was 7% (P<.001). In all 3 models, self-organization was also a predictor. For Group A (n=199), working hours per week was an additional predictor. Only in the model referring to course day 29 with online participants was intensive reviewing and reflective learning also identified as a predictor.

In every calculated model on the 4 levels, the ability for self-organization gave an indication of how great the learning outcomes can be, but the prediction probability for the model was low (detailed results of the models at 4 different levels are provided in Table S2 in Multimedia Appendix 2). It is remarkable that the influence was present regardless of whether the students participated online or face to face. Other personal skills that also had an influence on the prediction probability of the learning outcomes of the students could not be assigned to a participation mode. Students living nearer to the Department of Anatomy seem to have an advantage regarding the learning outcomes, regardless of the participation mode, as indicated by the results of the linear regressions analysis that were conducted. Commuting time as a predictor was identified in models that considered online participation as well as in those that considered face-to-face participation (Table S2 in Multimedia Appendix 2). However, the influence was minor for all the variables.

The analysis of the subgroup of low-performing participants showed that when only the results for online participation were considered, the model yielded an R^2 of 0.15 (P<.001; B=9.47,

XSL•FO
RenderX

SE 0.84; 2-tailed t_{174}=11.26). The significant predictors were self-organization (higher the value, better the performance; B=0.75, SE 0.18; β=.30; t_{174}=4.19; P<.001), commuting time (longer the commute time, worse the performance; B=−0.03, SE 0.01; β=−.25; t_{174}=−3.48; P<.001), and agreement with the statement "I prefer intensive reviewing and reflecting for learning" (lower the agreement, higher the scores; B=−0.47, SE 0.17; β=−.20; t_{174}=−2.82; P=.005).

Table . Linear stepwise regression models to predict the learning outcomes: the model with the best predictability for the subgroup is shown. (1)="I am very well self-organized when it comes to learning"; (2)="What is your average travel time in minutes to the Anatomical Institute?"; (3)="How old are you?"; (4)="I prefer intensive reviewing and reflecting for learning."; (5)="I prefer practical tasks for learning" (6)="How many hours per week do you work in addition to your studies to support yourself?"

Course days, participation mode, and predictor	B[a]	SE	β[b]	*t* test (*df*)	*P* value	R^2
Course day 28						
Online (n=199)						
—[c]	10.99	0.75	—	14.71 (195)	<.001	—
(2)	−0.04	0.01	−0.25	−3.53 (195)	<.001	0.12
(1)	0.48	0.18	0.18	2.70 (195)	.008	—
(6)	−0.06	0.03	−0.14	−2.01 (195)	.046	—
Face-to-face (n=216)						
—	9.46	0.65		14.65 (214)	<.001	—
(1)	0.50	0.16	0.20	3.05 (214)	.003	0.04
Course day 29						
Online (n=216)						
—	10.54	0.78	—	13.57 (213)	<.001	—
(1)	0.76	0.17	0.31	4.54 (213)	<.001	0.10
(4)	−0.43	0.16	−0.18	-2.65 (213)	.009	—
Face-to-face (n=199)						
—	10.40	0.68		15.29 (196)	<.001	—
(6)	−0.08	0.03	−0.20	−2.90 (196)	.004	0.07
(1)	0.49	0.17	0.20	2.86 (196)	.005	—

[a]B: unstandardized coefficient.

[b]β: standardized coefficient.

[c]Not applicable.

For students who attended only course day 28 online (model-yielded R^2=.06; P<.001; B=9.75, SE 0.47; t_{80}=20.73), commuting time remained the only significant predictor (B=−0.02, SE 0.01; β=−.24; t_{80}=−2.24; P=.03).

For students who attended only course day 29 online, self-organization (B=1.52, SE 0.25, β=.46; t_{91}=4.95; P<.001) and agreement with the statement "I prefer intensive reviewing and reflecting for learning" (B=−0.72, SE 0.22; β=−.29; t_{91}=−3.25; P=.002) were also significant (again negatively correlated), and agreement with the statement "I prefer practical tasks for learning" (B=0.62, SE 0.26; β=.22; t_{91}=2.44; P=.02) and commuting time (B=−0.03, SE 0.01; β=−.20; t_{91}=−2.20; P=.03) additionally emerged as significant positive predictors. The model explained more variance in this group (R^2=.28; P<.001; B=6.43, SE 1.69; t_{91}=3.81).

Technological affinity was retained as a predictor in the linear regression models, because no association was detected between students' technological affinity and their learning outcomes.

Discussion

Principal Findings

This study shows that there is no difference in the learning outcomes of the students using a virtual 3D slide box (Histologi@) and participating in the 2 anatomy course days face-to-face vs students using Histologi@ and participating in the course online via a common broadcast system.

Interpretation of the Topic Effect

The difference in learning outcomes between the 2 course days may be due to the fact that the course day on which students achieved higher learning success was closer to the simulated examination.

XSL•FO
RenderX

This assumption was supported by the finding that, in the analysis of students with lower outcomes who attended course day 29 online, agreement with the statement "I prefer intensive reviewing and reflecting for learning" emerged as a significant predictor with a negative correlation. This was not the case for course day 28, which had taken place earlier. This suggests that, for the students of this subgroup, having more time for reflection and self-directed learning may have led to better results in this instance. To verify this assumption, the learning outcome would have needed to be reassessed at a later point in time.

The 2 topics, liver and kidney, were selected to ensure a content scope that was comparable. However, it is conceivable that differences in conceptual complexity between the 2 topics led to students performing better on the second topic, "liver." For the purpose of addressing the research question, it is important that there was no significant difference between the instructional modes. Both groups of students, those who participated online on the respective day and those who attended in person, had similar learning outcomes.

Individual Characteristics and ATI Framework

In the study, neither individual demographic characteristics nor individual skill variables were identified as significant predictors with high predictive power of the outcomes. The regression analyses showed that there were no prominent variables referring to demographic characteristics or personal skills analyzed in this study that had a significant influence on the learning outcomes in either condition, online versus face-to-face. The study clearly shows that the learning preferences (auditory, visual, or kinesthetic) that were surveyed by the questions of the learning preferences are not related to the prediction of the learning outcomes.

This may be partly attributable to the fact that the students had already been exposed to both online and face-to-face learning formats for an extended period of time. Moreover, they had been able to practice using the digital 3D system online for several months prior to the study.

Because the students were already familiar with the Histologi@ technical system and had the opportunity to practice with it, it is possible that eg, the learner characteristic of technological affinity was no longer sufficiently differentiated within the framework of ATI studies, and therefore the different instructional methods did not lead to significant effects. This suggests that it would be necessary to examine beginners who have not yet had the opportunity to practice with the system. In such a group, larger effects might become observable.

ATI-studies assume that certain environmental characteristics or personal abilities influence a person's learning success in a specific learning environment [19]. ATI theory could be helpful at this point in guiding the development of personalized online interventions that support beginners in becoming familiar with the system. For example, within the ATI framework, it could also be investigated how long students with different aptitudes require to effectively engage with and follow the instructions online.

Another reason for the lack of detected differences may also be due to the fact that admission to medical school is subject to very strict selection criteria. For example, the final grade in the Abitur (German university entrance qualification) still plays a major role, as do the results in other qualifications that are relevant for admission to medical studies.

However, some variables can predict the learning outcome in microscopic anatomy to a small extent (eg, the assessment of self-organization). Since the probability of prediction is so small and especially since the variables make no general differences between the participation modes, it suggests that no major measures are necessary to adjust future online teaching accordingly to improve the learning conditions of individual groups with special characteristics tested in this study.

It is well known that self-organization, which includes self-regulated learning, is consistently associated with learning outcomes among medical students in particular [25]. Its predictive power was relatively modest in our study, and self-regulation did not differentiate between instructional modes.

Since self-regulation plays an especially important role in academic success during the clinical years, it should already be fostered in the preclinical phase, eg, by promoting autonomous learning structures or adapting the learning environment accordingly. Since, in our study, technical affinity did not emerge as a factor that led to a deterioration of students' learning outcomes in the online mode, this may represent an opportunity to foster an environment that supports self-organization. Naturally, students' level of experience must be considered for all measures taken.

Comparison With Prior Work

The results of this study indicate that the decision to implement the possibility of online participation with Histologi@ via broadcast system could provide a good alternative for students in the microscopic anatomy course. Participation in the microscopic anatomy course with Histologi@, no matter whether online or face-to-face, can be equally effective for students' learning progress, and for students of all levels of technological affinity. The use of Histologi@ under controlled short-term conditions is just as effective for medical students with their personal devices as it is in the microscopic laboratory at the provided workstations. Thus, our results confirm evidence that instruction in microscopic anatomy using virtual microscopy can be as effective as face-to-face instruction [26,27]. A distinctive feature of our study is that, within a clearly defined setting (comprising 2 topics covered over 2 course days followed by an immediate simulation examination) and compared with long-term studies a high control over external factors (eg, external study resources and group learning effects), we were able to show in a cross-over-designed study that learning outcomes did not differ between students attending online and those attending face-to-face. Previous studies compared the results of microscopic anatomy courses that were conducted over several months [26]. During this extended period, it is common for students to additionally study the course content using other resources (eg, asynchronous digital learning materials) and in learning groups.

The translation from a traditional microscopic anatomy course to VM was well planned and carefully prepared, both in terms

XSL•FO
RenderX

of technology and in terms of having experienced or trained educators, as this is crucial for a successful transformation [28-30]. Short-term shifts from traditional analog to online VM, as it has oftentimes been the case during the COVID-19 pandemic, may not always yield the same outcomes, especially in subjects like anatomy, which rely on describing and understanding 3D structures [31,32]. In contrast to many COVID-era studies investigating emergency remote instruction, our study examined structured instructional modes under controlled conditions and did not have to be introduced abruptly due to a political decision. It had already been implemented over several course cycles, allowing instructors to become more familiar with face-to-face and online synchronous instructions than was the case during the COVID-19 pandemic. Students were not required to participate online but did so voluntarily. It can therefore be assumed that psychosocial stress, which also plays a role in academic performance, was greater during the COVID-19 pandemic.

Relation to Broader Instructional Strategies

The curriculum for medical studies must be adapted to the changing requirements of medical education, ensuring the appropriate knowledge transfer to the students [18,33]. For example, the content of microscopic anatomy is increasingly integrated with clinical subjects. In order for students to engage with the study of microscopic anatomy regardless of the location of a microscopic laboratory, online instruction using VM is a good alternative to traditional face-to-face teaching. Being location-independent in teaching microscopic anatomy facilitates cooperation with clinical subjects, a key trend for future curriculum development [34,35]. Special consideration should be given to equipment and the learning environment in online teaching [36-38], and it should be carefully planned and not just be seen as a quick fix.

Educational Implications

Although online teaching resulted in the same learning outcomes as face-to-face instruction in our study, one should not assume that this would automatically be true for all types of anatomy teaching. Other important factors in medical training that may not be experienced during online teaching have to be trained face-to-face, a teaching mode that is still deemed necessary by students for a profound medical education [39].

The Sandwich Model [20], for example, alternates between collective teaching phases and individual or small-group active learning phases, thereby maintaining attention and promoting cognitive processing. Structured Sequence Models, Team-based learning emphasizes structured collaboration, accountability, and application of knowledge in teams, while the Flipped Classroom model shifts content acquisition to preclass preparation and uses classroom time for higher-order application tasks and is also applied in preclinical anatomy education [40-42].

In contrast to these pedagogical approaches, which primarily focus on restructuring in-class interaction and collaborative learning, this study did not investigate a specific active-learning intervention but rather examined instructional modes within an established Sandwich Model framework. Thus, the study's primary focus was not on comparing distinct didactic strategies but on evaluating learning outcomes across instructional delivery formats. However, it should also be reconsidered and further investigated to what extent instructional formats such as Problem-based learning or Team-based learning, particularly when implemented in an online setting, are suitable for the acquisition of foundational knowledge in the preclinical phase. In addition to teaching formats, diverse learning styles should be considered because it can also increase the effectiveness of learning units and should be examined in future studies [43].

Strengths and Limitations of the Study

A high number of participants was reached, and the study design was implemented under controlled circumstances, resulting in high statistical power.

To restrict external influences such as courses and examinations in other subjects or extended holiday breaks, the study was limited to 2 course days. In addition, efforts were made to minimize disruption to the regular teaching schedule.

For the study, the organization and the traditional structure of the microscopic anatomy course had to be adjusted. Timely information and transparent communication (eg, via email or presentation) were offered to minimize reservations from instructors, organizers, and students.

The students already had several weeks of experience applying histology in the histology laboratory when the study took place at the end of the course. Therefore, the study does not necessarily apply to students without experience in the use of virtual microscopy.

An important limitation of this study is that, although the students were encouraged to attend either the face-to-face or the online session on their assigned course day, due to the predefined framework conditions under which the study was conducted, the students ultimately had the option to participate online on their preferred day. As a result, a potential bias due to carry-over effects cannot be excluded. Nevertheless, since the students had already been working with the program for several weeks prior to this study, it can be assumed that the self-selected order of online and face-to-face participation did not introduce substantial bias. However, in future studies, it would be advisable to enforce a strictly randomized allocation of the participation mode and its sequence.

Additionally, potential confounding variables such as academic performance, motivation, and external study time were not examined in this study. These variables may also have influenced the learning outcomes.

Conclusions

With careful planning and the right equipment, online teaching in microscopic anatomy using virtual 3D slides can be comparable to face-to-face teaching in terms of digital microscopy trained student learning outcomes. The medical students at the Ludwig-Maximilians-Universität bring the necessary prerequisites with them at the end of the microscopic anatomy course to achieve similar learning outcomes in both online and face-to-face instruction using Histologi@. This suggests that students are able to follow the content of digital

XSL•FO
RenderX

instruction in microscopic anatomy as effectively as face-to-face instruction. However, it should also be recognized that students do not always have the opportunity to gain extensive experience with virtual microscopy in a face-to-face setting before transitioning to online formats. Investigations of digital teaching formats involving novice students may help to better differentiate associations between demographic characteristics and personal skills and learning outcomes. In further studies on online teaching, different instructional formats and learning preferences should also be considered to determine whether these factors may influence learning outcomes differently across instructional modes.

Acknowledgments

We thank Nirav Barapatre and Leon Kaub for having taught a class each within the framework of this study. Our special thanks go to the students who volunteered to participate in this study. We would also like to thank Dr. Stefan Milz (MD) for his technical support during the use of Histologi@.

Generative AI tools were used solely for language editing and to improve clarity in individual sentences. The authors are fully responsible for all content, analyses, and interpretations presented in the manuscript.

Funding

No external financial support or grants were received from any public, commercial, or not-for-profit entities for the research, authorship, or publication of this article.

Data Availability

The datasets analyzed during the current study are available from the corresponding author upon reasonable request.

Authors' Contributions

Conceptualization: KS, MB
Methodology: KS, MB
Investigation: KS
Formal analysis: KS
Data curation: KS
Writing – original draft: KS
Writing – review & editing: KS, MB
Supervision: MB
Resources: CS
Project administration: CS
All authors read and approved the final version of the manuscript for submission.

Conflicts of Interest

CS served as consultant for MBF Bioscience (Williston, VT, USA) until 12/2017. All other authors have no conflicts to declare.

Multimedia Appendix 1
Items of the evaluation that related to demographic characteristics of diversity and personal skills were shown.
[DOCX File, 34 KB - mededu_v12i1e82012_app1.docx]

Multimedia Appendix 2
Linear stepwise regression models to predict the learning outcomes: The model with the best predictability for the subgroup is shown.
[DOCX File, 38 KB - mededu_v12i1e82012_app2.docx]

References

1. Middendorff E. Growing heterogeneity among students? empirical findings to examine a postulated trend [Chapter in German]. In: Schaeper H, Kerst C, editors. Differentiation in the Higher Education System: National and International Developments and Challenges: Waxmann Verlag; 2015:261-278 URL: https://tinyurl.com/2kxnaet9 [accessed 2026-08-05]
2. Brändle T, Ordemann J. Same same but different? non-traditional students and alumni in Germany. Studia paedagogica 2020;25(4):35. [doi: 10.5817/SP2020-4-2]
3. El Hag S. Digital inequality and two levels of the digital divide in online learning: a mixed methods study of underserved college students. J Inf Syst Educ 2024;35(3):377-389. [doi: 10.62273/SSIF6302]

XSL•FO
RenderX

4. Rusatira JC, Tomaszewski B, Dusabejambo V, et al. Enabling access to medical and health education in Rwanda using mobile technology: needs assessment for the development of mobile medical educator apps. JMIR Med Educ 2016 Jun 1;2(1):e7. [doi: 10.2196/mededu.5336] [Medline: 27731861]
5. Judd T, Elliott K. Selection and use of online learning resources by first-year medical students: cross-sectional study. JMIR Med Educ 2017 Oct 2;3(2):e17. [doi: 10.2196/mededu.7382] [Medline: 28970187]
6. Escamilla-Sanchez A, López-Villodres JA, Alba-Tercedor C, et al. Instagram as a tool to improve human histology learning in medical education: descriptive study. JMIR Med Educ 2025 Feb 19;11(1):e55861. [doi: 10.2196/55861] [Medline: 39970433]
7. Remtulla R. The present and future applications of technology in adapting medical education amidst the COVID-19 pandemic. JMIR Med Educ 2020 Jul 17;6(2):e20190. [doi: 10.2196/20190] [Medline: 32634107]
8. Blake CA, Lavoie HA, Millette CF. Teaching medical histology at the University of South Carolina School of Medicine: transition to virtual slides and virtual microscopes. Anat Rec B New Anat 2003 Dec;275(1):196-206. [doi: 10.1002/ar.b.10037] [Medline: 14628320]
9. Vainer B, Mortensen NW, Poulsen SS, et al. Turning microscopy in the medical curriculum digital: experiences from the Faculty of Health and Medical Sciences at University of Copenhagen. J Pathol Inform 2017;8:11. [doi: 10.4103/2153-3539.201919] [Medline: 28382225]
10. Husmann PR, O'Loughlin VD, Braun MW. Quantitative and qualitative changes in teaching histology by means of virtual microscopy in an introductory course in human anatomy. Anat Sci Educ 2009 Oct;2(5):218-226. [doi: 10.1002/ase.105] [Medline: 19743410]
11. Wilson AB, Taylor MA, Klein BA, Sugrue MK, Whipple EC, Brokaw JJ. Meta-analysis and review of learner performance and preference: virtual versus optical microscopy. Med Educ 2016 Apr;50(4):428-440. [doi: 10.1111/medu.12944] [Medline: 26995482]
12. Lee LMJ, Goldman HM, Hortsch M. The virtual microscopy database-sharing digital microscope images for research and education. Anat Sci Educ 2018 Sep;11(5):510-515. [doi: 10.1002/ase.1774] [Medline: 29444388]
13. Somera Dos Santos F, Osako MK, Perdoná GDS, Alves MG, Sales KU. Virtual microscopy as a learning tool in Brazilian medical education. Anat Sci Educ 2021 Jul;14(4):408-416. [doi: 10.1002/ase.2072] [Medline: 33720510]
14. Attardi SM, Barbeau ML, Rogers KA. Improving online interactions: lessons from an online anatomy course with a laboratory for undergraduate students. Anat Sci Educ 2018 Nov;11(6):592-604. [doi: 10.1002/ase.1776] [Medline: 29493909]
15. Holland J, Clarke E, Glynn M. Out of sight, out of mind: do repeating students overlook online course components? Anat Sci Educ 2016 Nov;9(6):555-564. [doi: 10.1002/ase.1613] [Medline: 27096280]
16. Darici D, Flägel K, Sternecker K, Missler M. Transfer of learning in histology: Insights from a longitudinal study. Anat Sci Educ 2024 Mar;17(2):274-286. [doi: 10.1002/ase.2363] [Medline: 38158384]
17. Röwert R, Lah W, Dahms K, et al. Diversity and academic success – study-relevant heterogeneity characteristics at universities and universities of applied sciences and their influence on academic success – a quantitative study [Report in German]. : CHE Centrum für Hochschulentwicklung; 2017 URL: https://www.che.de/download/che_ap_198_diversitaet_und_studienerfolg-pdf/ [accessed 2026-08-05]
18. Gribbin W, Wilson EA, McTaggart S, Hortsch M. Histology education in an integrated, time-restricted medical curriculum: academic outcomes and students' study adaptations. Anat Sci Educ 2022 Jul;15(4):671-684. [doi: 10.1002/ase.2127] [Medline: 34363740]
19. Cronbach LJ, Snow RE. Individual differences in learning ability as a function of instructional variables (final report). : Stanford University School of Education; 1969.
20. Frank HG, Sternecker K. Histologi@: Cytology, Histology and Microscopic Anatomy Script and Annotated Learning Objective Catalog for Histological Courses for Students of Dental and Human Medicine [Book in German]: Anatomische Anstalt der LMU München; 2021. URL: https://epub.ub.uni-muenchen.de/76443/ [accessed 2026-08-05] [doi: 10.5282/ubm/epub.76443]
21. Frank HG. Histology for Dummies [Book in German]: Wiley-VCH; 2021. URL: https://www.wiley.com/en-br/shop/general-introductory-medical-science/histologie-f%C3%BCr-dummies-p-9783527827886 [accessed 2026-08-05]
22. Kadmon M, Strittmatter-Haubold V, Greifeneder R, Ehlail F, Lammerding-Köppel M. The sandwich principle--introduction to learner-centred teaching/learning methods in medicine [Article in German]. Z Evid Fortbild Qual Gesundhwes 2008;102(10):628-633. [doi: 10.1016/j.zefq.2008.11.018] [Medline: 19402349]
23. Franke T, Attig C, Wessel D. A personal resource for technology interaction: development and validation of the Affinity for Technology Interaction (ATI) Scale. Int J Hum-Comput Interact 2019 Apr 3;35(6):456-467. [doi: 10.1080/10447318.2018.1456150]
24. Biolucida® medical education. MBF Bioscience. URL: https://www.mbfbioscience.com/products/biolucida-medical-education [accessed 2026-08-27]
25. Cho KK, Marjadi B, Langendyk V, Hu W. The self-regulated learning of medical students in the clinical environment - a scoping review. BMC Med Educ 2017 Jul 10;17(1):112. [doi: 10.1186/s12909-017-0956-6] [Medline: 28693468]
26. Barbeau ML, Johnson M, Gibson C, Rogers KA. The development and assessment of an online microscopic anatomy laboratory course. Anat Sci Educ 2013;6(4):246-256. [doi: 10.1002/ase.1347] [Medline: 23364905]

XSL•FO
RenderX

27. Darici D, Reissner C, Brockhaus J, Missler M. Implementation of a fully digital histology course in the anatomical teaching curriculum during COVID-19 pandemic. Ann Anat 2021 Jul;236:151718. [doi: 10.1016/j.aanat.2021.151718] [Medline: 33675948]
28. Helle L, Nivala M, Kronqvist P. More technology, better learning resources, better learning? lessons from adopting virtual microscopy in undergraduate medical education. Anat Sci Educ 2013;6(2):73-80. [doi: 10.1002/ase.1302] [Medline: 22930425]
29. Bauler LD, Lesciotto KM, Lackey-Cornelison W. Factors impacting the rapid transition of anatomy curricula to an online environment in response to COVID-19. Anat Sci Educ 2022 Mar;15(2):221-232. [doi: 10.1002/ase.2163] [Medline: 34898018]
30. Albalushi H, Al Mushaiqri M, Sirasanagandla SR, Das S. Students' performance in face-to-face, online, and hybrid methods of teaching and assessment in anatomy. Int J Environ Res Public Health 2022 Oct 15;19(20):13318. [doi: 10.3390/ijerph192013318] [Medline: 36293903]
31. Evans DJR, Bay BH, Wilson TD, Smith CF, Lachman N, Pawlina W. Going virtual to support anatomy education: a STOPGAP in the midst of the COVID-19 pandemic. Anat Sci Educ 2020 May;13(3):279-283. [doi: 10.1002/ase.1963] [Medline: 32277598]
32. Wilcha RJ. Effectiveness of virtual medical teaching during the COVID-19 crisis: systematic review. JMIR Med Educ 2020 Nov 18;6(2):e20963. [doi: 10.2196/20963] [Medline: 33106227]
33. Mbagwu SI, Edem EE, Adegbilero‐Iwari OE, et al. Medical school students' perspectives on virtual anatomy education in a developing country. Anatomical Sciences Ed 2024 Sep;17(6):1345-1357. [doi: 10.1002/ase.2486]
34. Kelsey A, McCulloch V, Gillingwater TH, Findlater GS, Paxton JZ. Anatomical sciences at the University of Edinburgh: initial experiences of teaching anatomy online. Transl Res Anat 2020 Jun;19:100065. [doi: 10.1016/j.tria.2020.100065]
35. Eng-Tat A, Liu P, R J, et al. The challenges of histology education. Clin Anat 2023 Mar;36(2):308-314. [doi: 10.1002/ca.23989] [Medline: 36494890]
36. Mathiowetz V, Yu CH, Quake-Rapp C. Comparison of a gross anatomy laboratory to online anatomy software for teaching anatomy. Anat Sci Educ 2016;9(1):52-59. [doi: 10.1002/ase.1528] [Medline: 25903289]
37. Mione S, Valcke M, Cornelissen M. Remote histology learning from static versus dynamic microscopic images. Anat Sci Educ 2016 May 6;9(3):222-230. [doi: 10.1002/ase.1572] [Medline: 26441000]
38. Longhurst GJ, Stone DM, Dulohery K, Scully D, Campbell T, Smith CF. Strength, weakness, opportunity, threat (SWOT) analysis of the adaptations to anatomical education in the United Kingdom and Republic of Ireland in response to the COVID-19 pandemic. Anat Sci Educ 2020 May;13(3):301-311. [doi: 10.1002/ase.1967] [Medline: 32306550]
39. Abualadas HM, Xu L. Achievement of learning outcomes in non-traditional (online) versus traditional (face-to-face) anatomy teaching in medical schools: a mixed method systematic review. Clin Anat 2023 Jan;36(1):50-76. [doi: 10.1002/ca.23942] [Medline: 35969356]
40. Ogut E, Yildirim FB, Senol Y, Senol AU. Comprehensive evaluation of the educational impact and effectiveness of specialized study modules in cross-sectional anatomy: a study on student engagement and learning outcomes. BMC Med Educ 2025 Apr 10;25(1):514. [doi: 10.1186/s12909-025-07050-9] [Medline: 40211255]
41. Xiao J. Improving student learning and performance through a continuous flipped anatomy teaching model. Anat Sci Educ 2024 Mar;17(2):219-226. [doi: 10.1002/ase.2351] [Medline: 37916536]
42. Bergman EM, de Bruin ABH, Herrler A, Verheijen IWH, Scherpbier A, van der Vleuten CP. Students' perceptions of anatomy across the undergraduate problem-based learning medical curriculum: a phenomenographical study. BMC Med Educ 2013 Nov 19;13:152. [doi: 10.1186/1472-6920-13-152] [Medline: 24252155]
43. Ogut E, Senol Y, Yildirim FB. Do learning styles affect study duration and academic success? Eur J Anat 2017;21(3):235-240 [FREE Full text]

Abbreviations

ATI: aptitude-treatment interaction
LMU: Ludwig Maximilian University
OM: optical microscopy
VM: virtual microscopy

XSL•FO
RenderX

Edited by A Stone; submitted 07.Aug.2025; peer-reviewed by A Taylor, E Ogut; revised version received 07.Jul.2026; accepted 07.Jul.2026; published 16.Sep.2026.

Please cite as:
Sternecker K, Schmitz C, Berndt M
Students' Learning Outcomes in a Virtual Microscopic Online and Face-to-Face Anatomy Course: Cross-Over Comparative Study
JMIR Med Educ 2026;12:e82012
URL: https://mededu.jmir.org/2026/1/e82012
doi:10.2196/82012

© Katharina Sternecker, Christoph Schmitz, Markus Berndt. Originally published in JMIR Medical Education (https://mededu.jmir.org), 16.Sep.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Education, is properly cited. The complete bibliographic information, a link to the original publication on https://mededu.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Synchronous Remote Teaching for Neurological Examination Training in Fifth-Year Medical Students: Quasi-Experimental Study

Chih-Shan Huang[1,2,3], MD, MSc; Kung-Pei Tang[4*], PhD; Ming-Shun Wu[5,6*], MD, PhD

[1]Department of Neurology, Wan Fang Hospital, Taipei Medical University, Taipei, Taiwan

[2]Department of Neurology, School of Medicine, College of Medicine, Taipei Medical University, Taipei, Taiwan

[3]Taipei Neuroscience Institute, Taipei Medical University, Taipei, Taiwan

[4]Department of Early Childhood and Family Education, College of Education, National Taipei University of Education, Taipei, Taiwan

[5]Division of Gastroenterology, Department of Internal Medicine, Wan Fang Hospital, Taipei Medical University, No 111, Sec 3, Xinglong Rd, Wenshan Dist, Taipei, Taiwan

[6]Division of Gastroenterology and Hepatology, Department of Internal Medicine, School of Medicine, College of Medicine, Taipei Medical University, Taipei, Taiwan

[*]these authors contributed equally

Corresponding Author:
Ming-Shun Wu, MD, PhD
Division of Gastroenterology, Department of Internal Medicine, Wan Fang Hospital, Taipei Medical University, No 111, Sec 3, Xinglong Rd, Wenshan Dist, Taipei, Taiwan

Abstract

Background: The use of remote teaching in medical education has increased since the COVID-19 pandemic. However, the effectiveness of synchronous remote teaching for specific psychomotor components of the neurological examination, such as tendon reflex assessment, remains underexplored.

Objective: This study aimed to evaluate the short-term effectiveness of synchronous remote teaching compared with traditional in-person teaching for training medical students in tendon reflex examination skills and related neurological examination knowledge.

Methods: This quasi-experimental study enrolled 110 fifth-year medical students between August 2022 and April 2024, who were assigned to either synchronous remote teaching (n=46) or in-person teaching (n=64). The participants completed pre- and postcourse knowledge tests and an objective structured clinical examination (OSCE) assessing tendon reflex examination skills. Motivational beliefs related to self-regulated learning and technology acceptance were measured at baseline. Group differences were analyzed using *t* tests, Mann-Whitney *U* tests, and correlational analyses.

Results: Posttest knowledge scores were higher in the synchronous remote teaching group than in the in-person group (mean difference 0.86, 95% CI 0.21 - 1.51; Cohen *d*=0.50; *P*=.01), indicating a modest effect size. Both groups demonstrated substantial improvements in tendon reflex OSCE performance; however, the between-group difference in OSCE score improvement was not statistically significant (mean difference −3.64; 95% CI −9.36 to 2.08; Cohen *d*=−0.26; *P*=.21). Within the synchronous remote teaching group, higher motivational beliefs were positively correlated with OSCE score gains (Pearson *r*=0.34; *P*=.04; n=46), suggesting a potential role of self-regulated learning–related motivational processes. These exploratory findings should be interpreted with caution.

Conclusions: Under the conditions of this study, synchronous remote teaching was associated with modest improvements in knowledge outcomes and with OSCE-assessed tendon reflex examination skill gains that were comparable to those of traditional in-person teaching, without statistically significant between-group differences in psychomotor performance. These findings do not establish durable skill acquisition and should not be interpreted as evidence of equivalence or superiority of synchronous remote teaching over in-person instruction. Unlike many prior remote medical education studies that primarily focused on asynchronous learning or self-reported outcomes, this study incorporated synchronous interactive instruction with objective OSCE-based psychomotor assessment. These findings contribute to the emerging literature on synchronous remote clinical skills education, although future randomized and longitudinal studies are needed to evaluate long-term skill retention and broader applicability.

(JMIR Med Educ 2026;12:e77034) doi:10.2196/77034

XSL•FO
RenderX

KEYWORDS

synchronous remote teaching; neurological examination; medical education; quasi-experimental study; self-regulated learning; technology acceptance

Introduction

The COVID-19 pandemic has accelerated the adoption of remote teaching in medical education [1-3]. Compared with traditional face-to-face instruction, remote teaching offers flexibility by reducing geographical and physical constraints [4]. It is commonly categorized into asynchronous and synchronous formats [5]. While asynchronous learning allows self-paced study, synchronous teaching enables real-time interaction and feedback, which may be particularly important for training psychomotor clinical skills [6]. Psychomotor skills require not only knowledge acquisition but also supervised practice, immediate feedback, and performance correction. These requirements present challenges for remote teaching environments, where the absence of physical presence may limit instructors' ability to observe and guide learners effectively [7]. Neurological examination is a core clinical competency used to localize lesions and guide diagnostic reasoning [8]. Many components, such as tendon reflex assessment, are psychomotor skills that require precise technique, appropriate force application, and timing [9]. These skills are therefore highly dependent on direct observation, guided practice, and feedback, making them particularly relevant for evaluating the effectiveness of synchronous remote teaching [10].

Recent studies have further demonstrated that online and synchronous learning modalities can effectively support medical education, particularly when interactive elements are incorporated [11-13]. Synchronous remote teaching may address some limitations of asynchronous formats by enabling real-time supervision and interaction [14]. Prior educational studies have suggested that synchronous online learning environments may enhance learner engagement and cognitive participation through immediate interaction and feedback mechanisms [15]. With video-based platforms, instructors can observe learners' performance, provide immediate feedback, and correct errors during skill execution [16-18]. However, whether such real-time interaction can translate into objectively measurable improvements in clinical performance remains unclear. From a theoretical perspective, learner engagement in synchronous remote environments may be understood through the technology acceptance model (TAM) [19], which emphasizes perceived usefulness and perceived ease of use as key determinants of technology adoption [20]. Recent bibliometric analyses further suggest that TAM remains a dominant theoretical framework for understanding learner engagement and behavioral intention in contemporary online education environments, particularly in technology-mediated interactive learning settings [21]. In addition, self-regulated learning (SRL) theory highlights the role of learners' active monitoring, goal setting, and self-efficacy in skill acquisition [22]. Core SRL components relevant to psychomotor skill acquisition include goal setting, self-monitoring of performance against criteria, and self-efficacy, which influences effort, persistence, and responsiveness to feedback [23]. These frameworks suggest that both technological factors and learner-related processes may influence learning outcomes in remote clinical skills training.

Despite these theoretical considerations, empirical evidence regarding the effectiveness of synchronous online teaching for psychomotor clinical skills, particularly in relation to objectively assessed neurological examination performance, remains limited and inconsistent [24]. Prior studies have largely focused on asynchronous learning or knowledge-based outcomes, with relatively few directly comparing synchronous remote teaching with traditional in-person instruction using standardized performance-based assessments such as objective structured clinical examinations (OSCEs) [25-28]. Evidence specifically addressing discrete psychomotor tasks, such as tendon reflex examination, remains particularly scarce. To address this gap, we conducted a quasi-experimental study to compare the short-term effectiveness of synchronous remote teaching and traditional in-person teaching among fifth-year medical students. The primary outcome was OSCE-assessed tendon reflex examination performance, and the secondary outcome was neurological examination knowledge. We hypothesized that synchronous remote teaching would achieve similar short-term improvements in both OSCE performance and knowledge compared with in-person teaching.

Methods

Study Design

This study used a neurological examination course integrated into the physical examination curriculum at a medical center affiliated with a university in Taipei. This study used a nonrandomized, open-label, active-control, parallel-group design with within-participant repeated measurements (pretest-posttest). Data were collected at the following two predefined time points: baseline before the course (T0) and immediately after course completion (T1), with motivation and technology acceptance measured once at baseline.

Setting

The study was conducted within the neurological examination curriculum for fifth-year medical students at a university-affiliated medical center in Taipei, Taiwan. Participant recruitment and data collection were conducted from August 29, 2022, to April 22, 2024.

Participant Recruitment

Fifth-year medical students (clerkship 1) rotating through internal medicine and neurology were recruited during the study period. Because the assessments and questionnaires were embedded within the formal curriculum, participation was considered quasi-voluntary. The students participated in the regular basic clinical skills neurological examination course, with each class including approximately 8‐10 students. Because different clerkship blocks varied in size and instructional modality across the study period, the number of

XSL•FO
RenderX

participants in each teaching group was not identical. Students were informed that questionnaire completion was optional and that declining participation would not affect course grades, clinical evaluations, or academic standing. Survey responses were anonymized and deidentified before analysis.

Inclusion and Exclusion Criteria

Inclusion criteria included fifth-year medical students enrolled in the neurological examination course during the study period. Exclusion criteria included inability to complete the pre- and posttests, failure to complete course procedures, or missing any major component of the study assessments.

Sample Size and Group Allocation

This study was designed as a pragmatic educational evaluation embedded within the existing clerkship curriculum. Therefore, a formal a priori sample size calculation based on a prespecified effect size, α level, and statistical power was not performed. Participant recruitment and group assignment were determined by cohort availability and the preexisting teaching schedule during the study period, rather than by an optimal allocation strategy (eg, Neyman allocation). As a result, group sizes were unequal between the synchronous remote teaching and the in-person teaching groups.

Specifically, the control group consisted of students who received standard in-person teaching as part of the routine curriculum. The experimental group comprised students from the same academic year who were additionally invited to participate in synchronous remote teaching. Both groups received comparable instructional content and opportunities for practice and feedback, differing only in the mode of delivery. As participation in the remote group was based on additional invitation and voluntary agreement, group allocation reflects convenience sampling without randomization; therefore, selection bias and baseline imbalance cannot be fully excluded.

Teaching Intervention and Instructional Procedures

The experimental group used the Remote Clinical Presence System (RCPS), an enhanced telepresence setup developed for real-time remote clinical teaching, with permission granted for its use in this study. RCPS was designed to support synchronous remote teaching in clinical settings by preserving real-time interaction between instructors and learners [29]. Google Chrome was used as the web interface, and Google Meet was used for audiovisual communication.

RCPS enabled bidirectional, real-time audiovisual interaction, allowing students to observe neurological examination demonstrations and receive immediate feedback while performing examination tasks remotely. To facilitate transmission of fine motor actions and examination techniques—such as tendon localization, patient positioning, and reflex hammer handling—the camera setup provided a stable and close-up view of the examination field. Viewing angles could be adjusted during teaching sessions based on instructional needs, and instructors were able to request repositioning of the camera or repeat demonstrations to ensure adequate visualization of key procedural steps.

Before each teaching session, audiovisual connectivity was checked to ensure stable transmission. In the event of transient technical interruptions, instruction resumed after reconnection, and critical examination steps were repeated as needed to maintain instructional continuity. This instructional approach is consistent with prior reports on enhanced telepresence systems and synchronous digital platforms, which emphasize real-time interaction, demonstration, and feedback as essential elements for skill-based clinical education [29-32].

This study used a quasi-experimental design of “nonequivalent experimental group, control group with pre- and posttests” (nonrandomized controlled trial). The control group consisted of students who received in-person teaching in the classroom. These students observed the instructor’s interactions with standardized patients, demonstrations of neurological examination procedures, and the results. The instructor interacted with the students, observed how they performed the procedures, and provided feedback.

The experimental group participated in synchronous remote video learning using RCPS. These students also observed interactions, demonstrations, and results via video. The instructor interacted with them through video conferencing, observed how they performed the procedures, and provided feedback.

The neurological examination course was conducted over 2 weeks, with each session lasting about 60 minutes, using 2 standardized patient scenarios. One scenario covered the motor and sensory systems, while the other covered the cranial nerves and cerebellar systems. The goal was to train students to focus on appropriate examination strategies based on patient symptoms. The instructor demonstrated techniques, interacted with students, pointed out common errors, and allowed time for students to practice performing the neurological examination. The instructor monitored both in-person and remote students, offering real-time feedback.

Measures and Covariates

The participants completed assessments electronically via a QR code–linked Google questionnaire at two time points (T0 baseline and T1 immediately after the course). At T0 (pretest), they completed background questionnaires, knowledge tests, self-reported confidence items, and baseline measures of motivation and technology acceptance using 5-point Likert scales based on the Motivated Strategies for Learning Questionnaire (MSLQ) [33] and the TAM [34]. At T1 (posttest), they repeated the knowledge and confidence assessments. In this study, self-reported confidence referred to their perceived comfort and readiness to perform the examination, rather than an objectively observed psychomotor skill. Self-reported confidence was measured using a 5-point Likert scale (1=not confident at all; 5=very confident), administered at both T0 and T1.

The course evaluation questionnaire and knowledge tests were reviewed by a panel of neurology educators to ensure content relevance, clarity, and alignment with course objectives. Revisions were made based on expert consensus before administration. The knowledge assessment was designed to

XSL•FO
RenderX

cover key neurological domains and cognitive processes relevant to the course and was divided into pre- and posttests with comparable content and difficulty [35].

OSCE Assessment Procedures

The participants also underwent an OSCE administered at both T0 (pretest) and T1 (posttest) using the same standardized checklist and scoring rubric. They conducted tendon reflex assessments on standardized patients at the following five specific sites: biceps brachii, brachioradialis, triceps brachii, knee jerk, and ankle jerk. The evaluation criteria for this test included the following components:

1. Instruction: explaining the examination procedure to the patient.
2. Preparation: assisting the patient to adopt the correct position, identifying the correct tendon location, and ensuring the examined area was relaxed.
3. Technique: holding the reflex hammer correctly and striking the precise tendon location.
4. Response: eliciting a tendon reflex and comparing the response on both sides.

Neurologists, trained as examiners, observed the students and graded their performance using a standardized checklist and scoring rubric. Examiners were blinded to participants' group allocation to minimize assessment bias and were not involved in the instructional sessions. Before the study, all examiners participated in a consensus meeting to standardize the interpretation of scoring criteria and ensure consistency in evaluation. Interrater reliability was assessed before study initiation, with 2 examiners independently rating 5 students; the resulting Spearman correlation coefficient was 0.70, indicating acceptable agreement. In this study, the only directly observed psychomotor skill assessment was the tendon reflex examination OSCE, which served as a focused proxy for selected neurological examination skills. This focused approach was chosen to ensure feasibility and consistency within the course context.

Statistical Analysis

Motivation and technology acceptance were treated as baseline learner characteristics, whereas knowledge scores, self-reported confidence, and OSCE performance were analyzed as repeated-measures outcomes. Before inferential analyses, normality of continuous variables and their change scores (T1-T0) was assessed using the Shapiro-Wilk test, supplemented by visual inspection of histograms and Q-Q plots.

Because outcomes were measured at two time points (T0 and T1), within-group pre-post changes were examined using paired-sample *t* tests when normality assumptions were met; otherwise, Wilcoxon signed-rank tests were applied. Between-group differences in change scores were examined using independent-sample *t* tests or Mann-Whitney *U* tests, as appropriate based on distributional assumptions.

Motivation and technology acceptance were further examined using Pearson correlation analyses to assess their associations with improvements in knowledge and OSCE performance.

To address baseline imbalances inherent to the nonrandomized design, regression-based, covariate-adjusted analyses were conducted using an analysis of covariance (ANCOVA) framework, with posttest outcomes specified as dependent variables, teaching modality as the fixed factor, and baseline performance and prior neurology rotation experience included as covariates. To assess whether prior neurology rotation experience moderated the effect of teaching modality, a teaching modality × prior neurology rotation interaction term was additionally included in the ANCOVA models, and *P* values for interaction were reported. Adjusted estimates with 95% CIs were reported to aid interpretation of the magnitude and precision of observed effects.

Subgroup analyses by tendon reflex examination steps, anatomical sites, and prior neurology rotation experience, as well as correlation analyses, were conducted as exploratory, hypothesis-generating analyses. Given their exploratory nature, no formal correction for multiple comparisons (eg, Bonferroni adjustment) was applied; therefore, the potential risk of type I error inflation was acknowledged, and findings were interpreted cautiously with emphasis on effect sizes and CIs.

Given the low proportion of missing data, complete-case analysis was performed, and no imputation procedures were applied. Study data were organized using Microsoft Excel (Microsoft Corp) and analyzed with SPSS (version 22; IBM Inc). All statistical tests were 2-tailed, with a 95% CI, and statistical significance was set at $P<.05$.

Ethical Considerations

This study was approved by the Taipei Medical University Joint Institutional Review Board (approval N202201046). Written informed consent was obtained from all participants before study participation. Participation in questionnaire completion was voluntary. Students were informed that declining participation or skipping questionnaire items would not affect course grades, clinical evaluations, or academic standing. Survey responses and study data were anonymized and deidentified before analysis. All participants received a convenience store gift voucher valued at 100 New Taiwan Dollars (approximately US $3.17) as compensation for study participation. No identifiable participant images or personal information were included in the paper.

Results

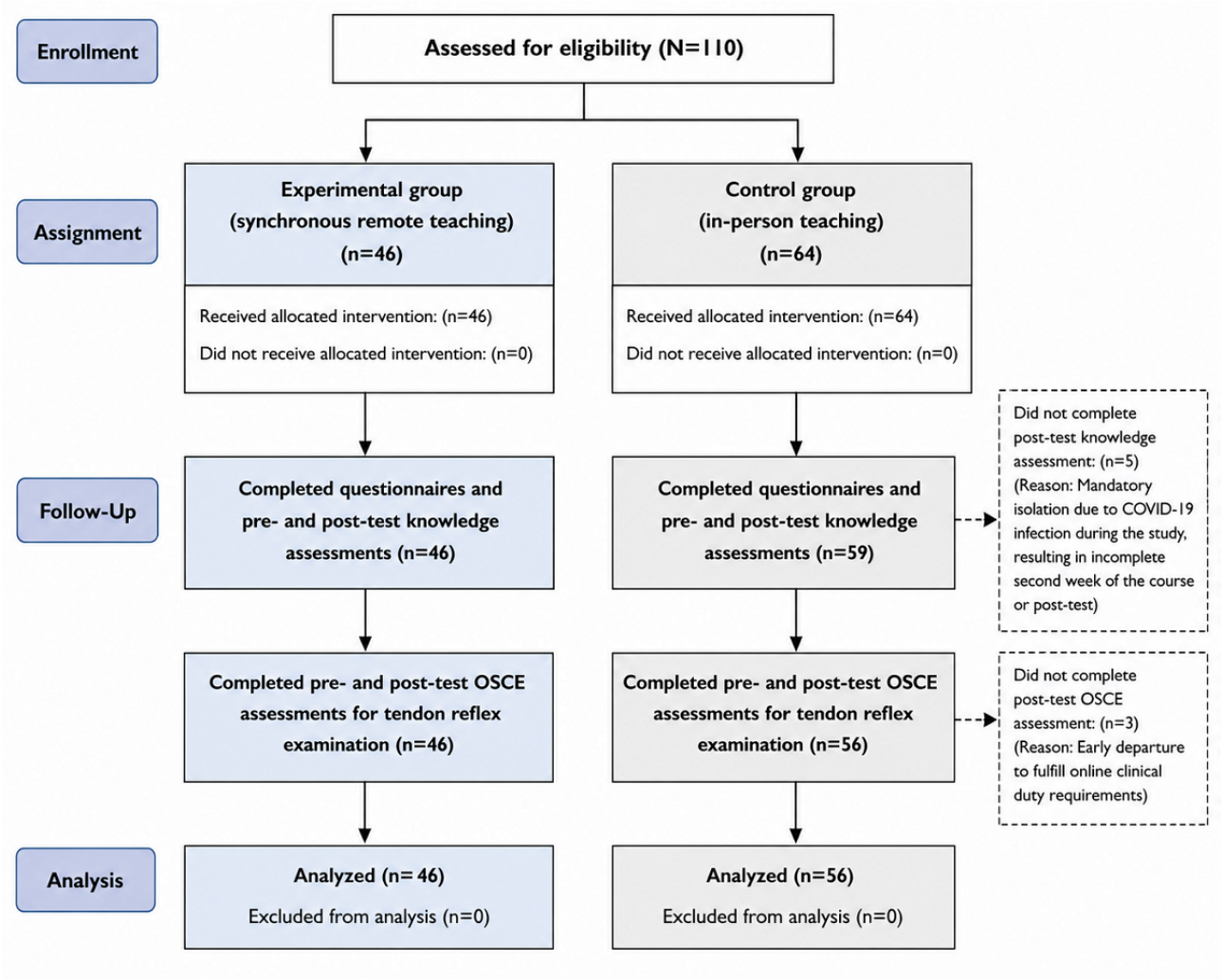
The course evaluation questionnaire and knowledge assessment demonstrated acceptable content coverage and internal consistency for the intended educational context.

This study recruited a total of 110 participants, of whom 46 (21 female, 45.7%; mean age 22.9, SD 1.3 years) were assigned to the experimental group (synchronous remote teaching). The control group (in-person teaching) included 64 participants (19 female, 29.7%; mean age 23.2, SD 1.9 years). No significant differences were found between the groups in terms of sex ($P=.11$) or mean age ($P=.31$). All participants completed a course evaluation survey, a pretest on neurological examination knowledge, and a pretest for the OSCE on tendon reflexes. In the control group, 5 participants did not complete the second

XSL•FO
RenderX

week of the course or posttest because they were under mandatory isolation following COVID-19 infection during the study period, leaving 59 participants for the posttest. In addition, 3 participants were unable to complete the OSCE posttest because they had to leave early to fulfill online clinical duty requirements, reducing the final sample size to 56. The recruitment process is shown in Figure 1.

Figure 1. Participant flowchart of the quasi-experimental study comparing synchronous remote teaching and traditional in-person teaching for neurological examination training among fifth-year medical students at a university-affiliated medical center in Taipei, Taiwan (August 2022-April 2024). OSCE: objective structured clinical examination.



Of the 110 participants, 105 completed both the pre- and posttests for the knowledge assessment and course evaluations. The Cronbach α for the MSLQ was 0.882, indicating good reliability, while the overall Cronbach α for the TAM was 0.937, indicating excellent reliability. The knowledge test items demonstrated a moderate difficulty index of 0.64 and an acceptable discrimination index of 0.23, supporting their suitability for educational assessment. Within-group analyses showed that the participants in the experimental group (n=46) had a significant increase in knowledge scores (mean 9.15, SD 1.42 vs mean 8.25, SD 1.63; $P<.001$), whereas the control group (n=59) had a smaller and nonsignificant improvement (mean 8.29, SD 1.95 vs mean 7.69, SD 1.79; $P=.08$). Between-group comparison indicated that the posttest knowledge scores were higher in the experimental group than in the control group (mean 9.15, SD 1.42 vs mean 8.29, SD 1.95; $P=.01$; Cohen $d=0.50$), corresponding to a moderate effect size. However, the absolute difference was less than 1 point on a 13-point scale, suggesting that the magnitude of this difference may be of limited practical educational significance. Notably, there was a baseline imbalance in prior neurology rotation experience between groups (42/46, 91.3% vs 43/59, 72.9%; $P=.02$), which may have partially contributed to the observed between-group difference in posttest scores. Importantly, no significant between-group differences were observed in pretest scores or in the magnitude of improvement from pre- to posttest, as shown in Table 1 and Figure 2, indicating that the overall learning gains were comparable between the 2 teaching modalities.

XSL•FO
RenderX

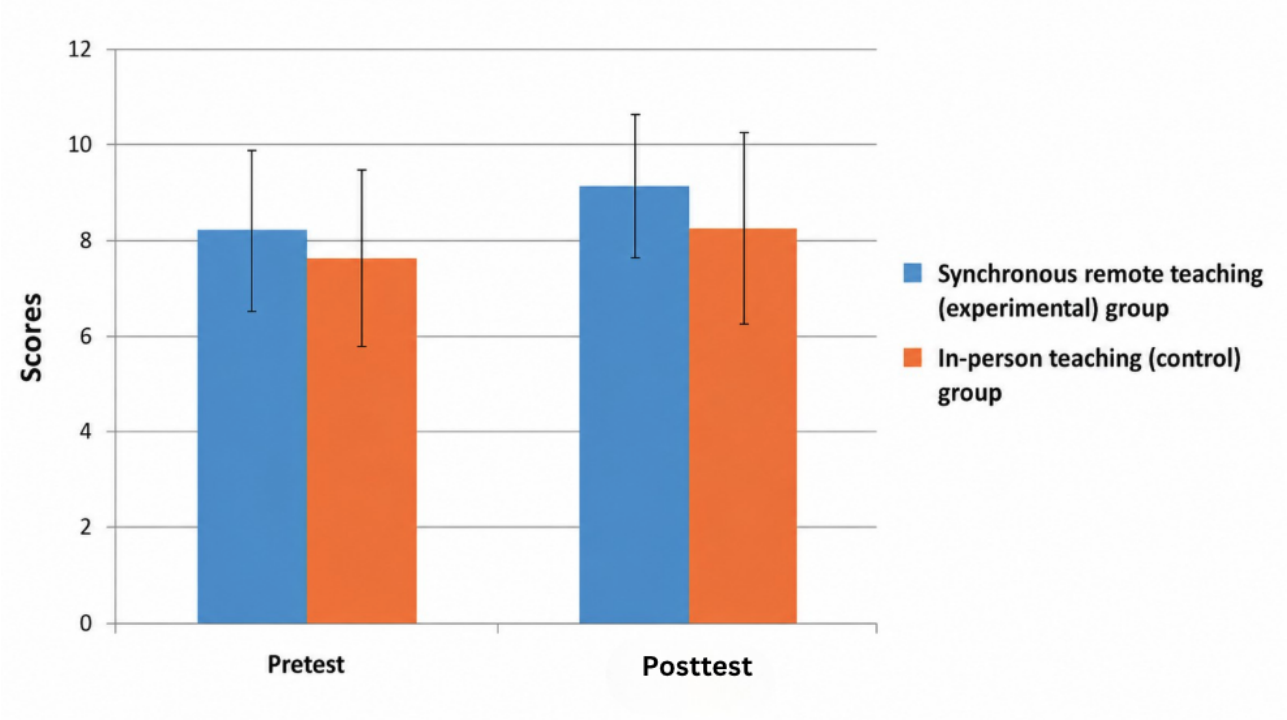
Table . Comparison of neurological examination knowledge between the experimental and control groups.

Characteristic	Experimental group (n=46)	Control group (n=59)	*t* test (*df*)	*P* value
Female, n (%)	21 (45.7%)	16 (27.1)	—[a]	.06
Age (years), mean (SD)	22.85 (1.28)	23.20 (1.94)	1.08 (103)	.29
Completed neurology rotation, n (%)	42 (91.3)	43 (72.9)	—	.02[b]
Pretest score (0 - 13), mean (SD)	8.25 (1.63)	7.69 (1.79)	–1.65 (103)	.10
Posttest score (0 - 13), mean (SD)	9.15 (1.42)	8.29 (1.95)	–2.61 (103)	.01[b]
Score improvement (pre-post), mean (SD)	0.90 (1.73)	0.60 (2.60)	–0.70 (103)	.48

[a]Not available.

[b]Significant at $P<.05$.

Figure 2. Comparison of pretest and posttest scores for the neurological examination knowledge assessment between the synchronous remote teaching (experimental) group and the in-person teaching (control) group.



Among the 105 participants who completed the knowledge test, 102 also completed the pretests and posttests for the tendon reflex neurological examination OSCE. The OSCE showed a moderate difficulty index of 0.55 and a good discrimination index of 0.43, indicating appropriate assessment performance for distinguishing learner ability. Within-group analyses showed that both the experimental group (n=46) and the control group (n=56) achieved substantial improvements in self-reported confidence and OSCE performance (all $P<.001$). These findings indicate that both instructional modalities were associated with marked short-term gains in psychomotor examination skills. However, regarding the primary outcome, the between-group difference in OSCE score improvement (pre-post change) was not statistically significant (mean 23.07, SD 16.53 vs mean 26.71, SD 11.31; $P=.21$), corresponding to a small effect size (Cohen $d=-0.26$; 95% CI −0.65 to 0.13). The CI included zero and spanned both potential benefits and disadvantages, indicating uncertainty in both the direction and magnitude of the effect. Overall, these findings suggest that the difference in OSCE-assessed psychomotor skill acquisition between the two teaching modalities was small and not educationally meaningful within the context of this study. Notably, a higher proportion of participants in the experimental group had prior neurology rotation experience (42/46, 91.3% vs 41/56, 73.2%; $P=.02$), which may have introduced baseline imbalance; however, this did not translate into a measurable advantage in OSCE improvement (Table 2 and Figure 3).

XSL•FO
RenderX

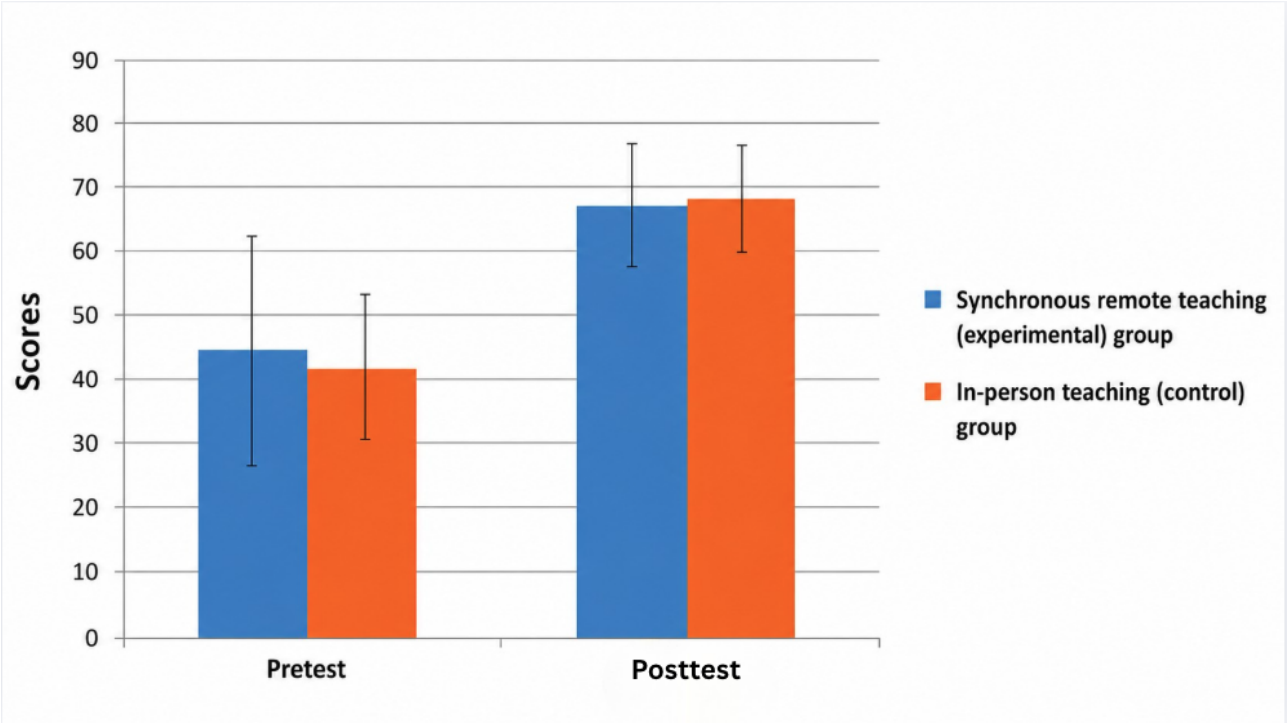
Table . Comparison of neurological tendon reflex objective structured clinical examination scores between the experimental and control groups.

Characteristic	Experimental group (n=46)	Control group (n=56)	*t* test (*df*)	*P* value	Cohen *D* (95% CI)
Female (%), n (%)	21 (45.7)	15 (26.8)	—[a]	.06	—
Age (years), mean (SD)	22.85 (1.28)	23.13 (1.82)	0.87 (100)	.39	—
Completed neurology rotation (%)	91.3 (42/46)	73.2 (41/56)	—	.02[b]	—
Frequency of practicing reflex examinations, mean (SD)	7.24 (14.93)	4.59 (13.16)	−0.95 (100)	.34	—
Preclass self-reported confidence in reflex examinations (1-5), mean (SD)	2.89 (0.95)	2.73 (1.12)	−0.77 (100)	.45	—
Postclass self-reported confidence in reflex examinations (1-5), mean (SD)	4.20 (0.78)	4.25 (0.75)	0.36 (100)	.72	—
Pretest total score (0 - 80), mean (SD)	44.35 (17.83)	41.66 (11.18)	−0.89 (100)	.38	0.18 (−0.21 to 0.58)
Posttest total score (0 - 80), mean (SD)	67.41 (9.47)	68.38 (8.31)	0.55 (100)	.59	−0.11 (−0.50 to 0.28)
Score Improvement (pre-post), mean (SD)	23.07 (16.53)	26.71 (11.31)	1.27 (100)	.21	−0.26 (−0.65 to 0.13)

[a]Not available.

[b]Significant at $P<.05$.

Figure 3. Comparison of pretest and posttest scores for the tendon reflex neurological examination objective structured clinical examination between the synchronous remote teaching (experimental) group and the in-person teaching (control) group.



Cohen *d* was calculated for between-group differences using pooled SDs (experimental minus control); 95% CIs were calculated using the SE of *d*.

Further subgroup analyses based on specific examination sites and procedural steps did not show statistically significant between-group differences in pre- to posttest changes, as shown in Tables 3 and 4. Importantly, the magnitude of changes across

XSL•FO
RenderX

subcomponents was consistently small, and the direction of effects was not systematic, further supporting the absence of a meaningful modality-specific advantage.

Table . Subgroup analysis of tendon reflex objective structured clinical examination scores by examination site comparing the experimental and control groups.

Examination site and assessment	Experimental group (n=46), mean (SD)	Control group (n=56), mean (SD)	*t* test (*df*)	*P* value
Biceps brachii				
Pretest (0 - 16)	9.13 (3.90)	8.79 (3.00)	−0.50 (100)	.62
Posttest (0 - 16)	13.61 (1.89)	13.54 (2.08)	−0.18 (100)	.86
Improvement	4.48 (3.93)	4.75 (2.58)	0.40 (100)	.69
Brachioradialis				
Pretest (0 - 16)	8.17 (4.72)	6.84 (3.95)	−1.56 (100)	.12
Posttest (0 - 16)	13.48 (2.43)	13.54 (2.26)	0.12 (100)	.90
Improvement	5.30 (4.20)	6.70 (4.27)	1.65 (100)	.10
Triceps brachii				
Pretest (0 - 16)	8.20 (4.18)	8.80 (2.50)	0.87 (100)	.39
Posttest (0 - 16)	13.17 (2.89)	13.50 (2.56)	0.60 (100)	.55
Improvement	4.98 (4.47)	4.70 (3.35)	−0.35 (100)	.72
Knee jerk				
Pretest (0 - 16)	10.33 (3.33)	9.52 (2.76)	−1.34 (100)	.18
Posttest (0 - 16)	13.72 (2.18)	13.96 (1.80)	0.63 (100)	.53
Improvement	3.39 (3.69)	4.45 (2.78)	1.65 (100)	.10
Ankle jerk				
Pretest (0 - 16)	8.52 (4.24)	7.71 (2.67)	−1.12 (100)	.27
Posttest (0 - 16)	13.43 (2.46)	13.84 (2.12)	0.89 (100)	.37
Improvement	4.91 (4.06)	6.13 (3.02)	1.68 (100)	.10

XSL•FO
RenderX

Table . Subgroup analysis of tendon reflex objective structured clinical examination by examination step comparing the experimental and control groups.

Examination step and assessment	Experimental group (n=46), mean (SD)	Control group (n=56), mean (SD)	*t* test (*df*)	*P* value
Step 1: instruction				
Pretest (0 - 10)	5.50 (3.21)	5.23 (3.06)	−0.43 (100)	.67
Posttest (0 - 10)	6.78 (3.12)	7.04 (3.38)	0.39 (100)	.70
Improvement	1.28 (4.28)	1.80 (4.11)	0.63 (100)	.53
Step 2: preparation				
Pretest (0 - 30)	17.41 (7.75)	16.29 (4.50)	−0.87 (100)	.39
Posttest (0 - 30)	26.46 (5.25)	26.95 (3.39)	0.55 (100)	.59
Improvement	9.04 (7.29)	10.66 (5.15)	1.27 (100)	.21
Step 3: technique				
Pretest (0 - 20)	10.98 (5.52)	10.52 (3.18)	−0.50 (100)	.62
Posttest (0 - 20)	18.26 (2.28)	18.32 (2.26)	0.13 (100)	.89
Improvement	7.28 (5.26)	7.80 (3.34)	0.58 (100)	.56
Step 4: response				
Pretest (0 - 20)	10.46 (5.20)	9.63 (4.67)	−0.85 (100)	.40
Posttest (0 - 20)	15.91 (3.27)	16.07 (3.43)	0.24 (100)	.81
Improvement	5.46 (5.09)	6.45 (4.79)	1.01 (100)	.32

To further address potential baseline imbalances between groups, additional covariate-adjusted analyses were conducted for both knowledge and tendon reflex OSCE outcomes using an ANCOVA framework. Posttest outcomes were specified as dependent variables, with teaching modality as the fixed factor, and baseline performance and prior neurology rotation experience included as covariates.

For the knowledge assessment, covariate adjustment for baseline knowledge scores and prior neurology rotation experience did not materially alter the primary findings. The adjusted posttest knowledge score was numerically higher in the synchronous remote teaching group than in the in-person group (adjusted mean difference=0.94 points; 95% CI −0.10 to 1.98), corresponding to a small effect size ($F(1,100)=3.23$; $P=.08$; partial $\eta^2=0.031$). Neither baseline knowledge performance nor prior neurology rotation experience demonstrated a significant independent association with post-test knowledge outcomes, and no significant interaction was observed between teaching modality and prior rotation experience.

Similarly, for the tendon reflex OSCE, covariate-adjusted analysis revealed a small, nonsignificant adjusted mean difference in posttest OSCE performance between the synchronous remote teaching and in-person teaching groups (−1.28 points; 95% CI −6.27 to 3.71), corresponding to a negligible effect size ($F(1,98)=0.26$; $P=.61$; partial $\eta^2=0.003$). Baseline OSCE performance was a significant predictor of posttest OSCE scores ($P<.001$), whereas prior neurology rotation experience was not. No interaction effect between teaching modality and prior rotation experience was observed ($P=.97$), indicating that the effectiveness of synchronous remote teaching was consistent regardless of prior clinical exposure.

In view of the potential impact of prior neurology rotation experience on learning outcomes, additional analyses were conducted. Among the 105 participants, 85 had completed a neurology rotation, while 20 had not. The participants with prior rotations demonstrated slightly higher baseline knowledge scores; however, this difference was not statistically significant. No meaningful differences were observed in posttest knowledge scores or in pre-post score improvement between those with and without prior rotations (Table 5).

XSL•FO
RenderX

Table . Comparison of knowledge outcomes between participants with and without prior neurology rotations[a].

Outcome measure	With prior neurology rotation (n=85), mean (SD)	Without prior neurology rotation (n=20), mean (SD)	*U* value	*z* value	*P* value
Pretest score (0 - 13)	8.03 (1.68)	7.54 (1.98)	756.5	−0.76	.45
Posttest score (0 - 13)	8.69 (1.82)	8.57 (1.62)	843.5	−0.05	.96
Score improvement (pre-post)	0.66 (2.11)	1.03 (2.85)	822.0	−0.23	.82

[a]Significant at *P*<.05. Subgroup analyses are exploratory. A formal interaction test (teaching modality × prior neurology rotation experience) was performed using a covariate-adjusted analysis of covariance framework; no significant interaction was observed.

Exploratory analyses suggested a numerically greater improvement in knowledge test scores among participants without prior neurology rotations, although this difference did not reach statistical significance. To formally evaluate whether prior neurology rotation experience moderated the effect of teaching modality on knowledge outcomes, we conducted a covariate-adjusted ANCOVA including baseline knowledge score, teaching modality, prior rotation experience, and their interaction. The interaction between teaching modality and prior neurology rotation experience was not statistically significant, indicating no evidence that prior clinical exposure moderated knowledge outcomes.

Among the 102 participants who completed the tendon reflex OSCE, those with prior neurology rotations (n=83) reported more frequent prior practice, higher precourse self-reported confidence, and higher pretest OSCE scores than those without rotations (n=19). Despite these baseline differences, postcourse self-reported confidence and posttest OSCE performance were similar between groups (Table 6). Exploratory subgroup analyses indicated a greater magnitude of OSCE score improvement among the participants without prior neurology rotations. However, this pattern should be interpreted cautiously, as the observed differences were small, derived from post hoc comparisons, and subject to potential type I error due to multiple testing.

Table . Comparison of tendon reflex objective structured clinical examination scores between participants with and without prior neurology rotations.

Outcome measure	With prior neurology rotation (n=83), mean (SD)	Without prior neurology rotation (n=19), mean (SD)	*U* value	*z* value	*P* value
Frequency of practicing reflex exams	6.53 (15.38)	2.53 (2.27)	508.0	−2.46	.01[a]
Preclass self-reported confidence in reflex exams (1-5)	2.90 (1.01)	2.37 (1.12)	557.5	−2.10	.04[a]
Postclass self-reported confidence in reflex exams (1-5)	4.23 (0.79)	4.21 (0.63)	744.0	−0.43	.67
Pretest total score (0 - 80)	44.46 (14.91)	35.95 (10.58)	488.0	−2.58	.01[a]
Posttest total score (0 - 80)	67.86 (8.70)	68.32 (9.59)	747.5	−0.35	.72
Score improvement (pre-post)	23.40 (14.17)	32.37 (10.43)	490.0	−2.57	.01[a]

[a]Significant at *P*<.05. Subgroup analyses are exploratory. A formal interaction test (teaching modality × prior neurology rotation experience) was performed using an analysis of covariance framework to evaluate effect modification (*P* for interaction=.97).

To assess whether prior neurology rotation experience moderated the effect of teaching modality on OSCE performance, a covariate-adjusted ANCOVA was performed with posttest OSCE score as the dependent variable, baseline OSCE score as a covariate, and teaching modality, prior rotation experience, and their interaction included as fixed factors. The interaction between teaching modality and prior neurology rotation experience was not statistically significant (*P* for interaction=.97), indicating that prior clinical exposure did not meaningfully influence the effect of teaching modality on psychomotor performance.

Subgroup analyses based on specific examination sites and procedural steps indicated that the participants with prior neurology rotations demonstrated higher baseline OSCE scores across several reflex sites and examination steps (Tables 7 and 8). Postcourse improvements were numerically greater among the participants without prior rotations in several components. These site- and step-based analyses involved multiple comparisons and were conducted for exploratory purposes only; therefore, the reported *P* values were not adjusted for multiple testing and should be interpreted cautiously.

XSL•FO
RenderX

Table . Subgroup analysis of tendon reflex examination, objective structured clinical examination scores by examination site among participants with and without prior neurology rotations.

Examination site and assessment	With prior neurology rotation (n=83), mean (SD)	Without prior neurology rotation (n=19), mean (SD)	*U* value	*z* value	*P* value
Biceps brachii					
Pretest (0 - 16)	9.30 (3.44)	7.37 (2.93)	485.0	−2.62	.009[a]
Posttest (0 - 16)	13.66 (1.88)	13.16 (2.43)	715.0	−0.64	.52
Improvement	4.36 (3.30)	5.79 (2.76)	570.5	−1.89	.06
Brachioradialis					
Pretest (0 - 16)	7.76 (4.51)	6.05 (3.29)	590.5	−1.71	.08
Posttest (0 - 16)	13.48 (2.29)	13.63 (2.54)	717.5	−0.62	.54
Improvement	5.72 (4.40)	7.58 (3.39)	591.5	−1.70	.09
Triceps brachii					
Pretest (0 - 16)	8.89 (3.46)	6.95 (2.37)	487.5	−2.61	.009[a]
Posttest (0 - 16)	13.24 (2.85)	13.84 (1.92)	712.0	−0.67	.50
Improvement	4.35 (4.00)	6.89 (2.40)	432.5	−3.07	.002[a]
Knee jerk					
Pretest (0 - 16)	10.06 (3.06)	9.11 (2.90)	666.0	−1.06	.29
Posttest (0 - 16)	13.89 (1.78)	13.68 (2.71)	747.5	−0.36	.72
Improvement	3.83 (3.29)	4.58 (3.06)	674.0	−0.99	.32
Ankle jerk					
Pretest (0 - 16)	8.45 (3.51)	6.47 (2.84)	490.5	−2.57	.01[b]
Posttest (0 - 16)	13.58 (2.28)	14.00 (2.29)	669.0	−1.05	.30
Improvement	5.13 (3.53)	7.53 (3.08)	476.5	−2.69	.007[a]

[a]Significant at $P<.01$.

[b]Significant at $P<.05$.

XSL•FO
RenderX

Table . Subgroup analysis of tendon reflex examination objective structured clinical examination scores by examination step among participants with and without prior neurology rotations.

Examination step and assessment	With prior neurology rotation (n=83), mean (SD)	Without prior neurology rotation (n=19), mean (SD)	*U* value	*z* value	*P* value
Step 1: instruction					
Pretest (0 - 10)	5.35 (3.03)	5.37 (3.55)	786.5	−0.02	.99
Posttest (0 - 10)	6.81 (3.34)	7.42 (2.89)	720.5	−0.61	.54
Improvement	1.46 (4.20)	2.05 (4.12)	731.0	−0.50	.62
Step 2: preparation					
Pretest (0 - 30)	17.46 (6.28)	13.89 (SD 4.84)	523.5	−2.28	.02[a]
Posttest (0 - 30)	26.77 (SD 4.30)	26.53 (SD 4.46)	782.0	−0.06	.95
Improvement	9.31 (SD 6.07)	12.63 (SD 6.34)	571.5	−1.87	.06
Step 3: technique					
Pretest (0 - 20)	11.18 (SD 4.52)	8.74 (SD 3.07)	474.0	−2.72	.007[b]
Posttest (0 - 20)	18.23 (SD 2.29)	18.58 (SD 2.12)	745.5	−0.40	.69
Improvement	7.05 (SD 4.39)	9.84 (SD 3.06)	451.0	−2.92	.004[b]
Step 4: response					
Pretest (0 - 20)	10.47 (SD 4.99)	7.95 (SD 4.03)	552.5	−2.03	.04[a]
Posttest (0 - 20)	16.05 (SD 3.14)	15.79 (SD 4.21)	769.5	−0.17	.87
Improvement	5.58 (SD 4.97)	7.84 (SD 4.39)	575.0	−1.84	.07

[a]Significant at $P<.05$.

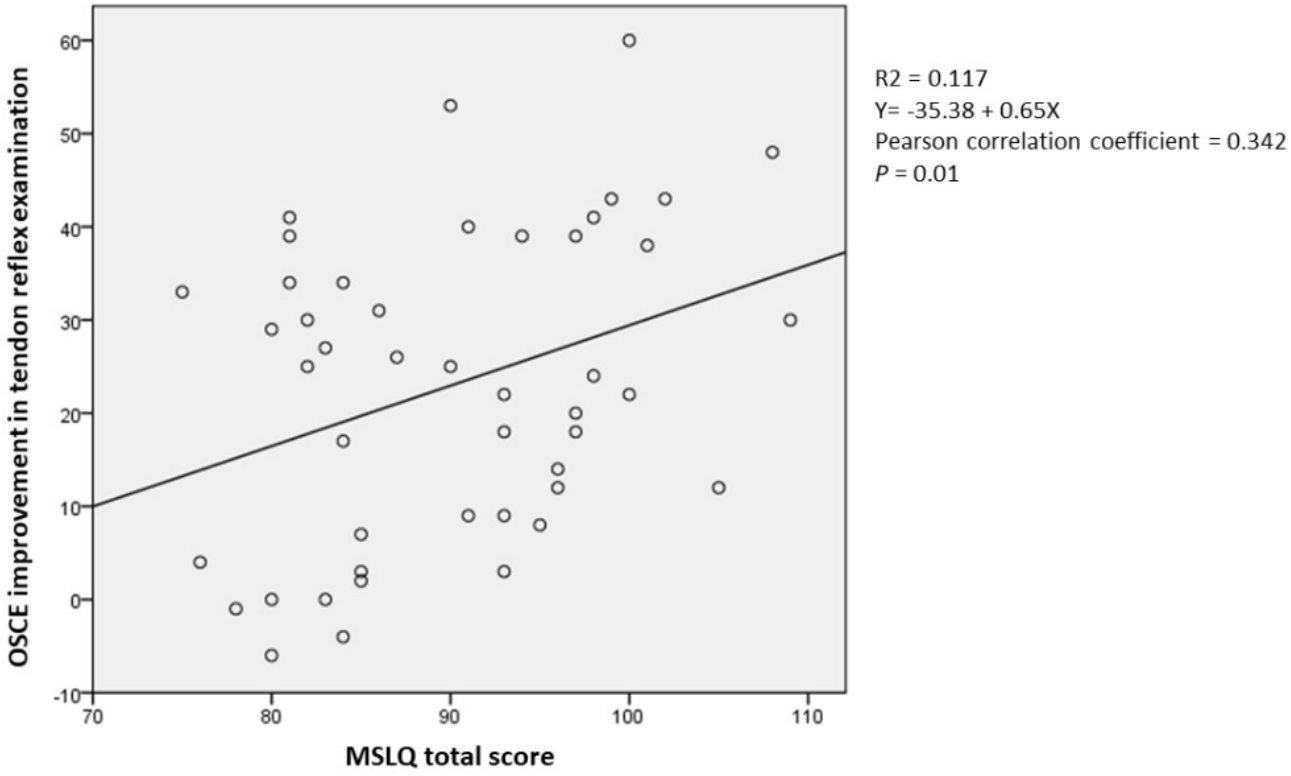
[b]Significant at $P<.01$.

All participants completed a pretest course evaluation survey assessing motivational beliefs regarding SRL (MSLQ) and technology acceptance of remote teaching products (TAM). Baseline scores were similar between the experimental group (n=46) and the control group (n=56) for both MSLQ (mean 90.17, SD 8.72 vs mean 89.79, SD 10.49; $P=.84$) and TAM (mean 60.54, SD 11.34 vs mean 59.75, SD 11.47; $P=.73$), indicating comparable baseline learner characteristics.

To explore whether learning outcomes were associated with motivational beliefs, participants were stratified into high- and low-scoring groups based on the median MSLQ score. These analyses were conducted as exploratory and hypothesis-generating, and no formal adjustment for multiple comparisons was applied. Participants with higher motivational belief scores demonstrated higher posttest knowledge scores (mean 9.06, SD 1.77 vs mean 8.35, SD 1.71; $P=.04$) and greater OSCE improvement (mean 27.75, SD 13.72 vs mean 22.16, SD 13.76; $P=.04$) compared with those with lower scores. Within the experimental group, higher motivational scores were associated with higher post-test knowledge scores (mean 9.55, SD 1.42 vs mean 8.68, SD 1.30; $P=.04$) and greater OSCE improvement (mean 27.60, SD 15.77 vs mean 17.67, SD 16.13; $P=.04$). However, these subgroup findings remain exploratory and should not be interpreted as evidence of a causal relationship. Correlation analyses showed moderate positive associations between self-efficacy ($r=0.367$) and overall MSLQ score ($r=0.342$) with OSCE improvement, suggesting that higher motivational beliefs may be related to greater performance gains. However, the strength of these associations was moderate, and regression analysis indicated that motivational beliefs explained only a limited proportion of the variance in OSCE improvement ($R^2=0.117$), as shown in Figure 4. These findings suggest that, while motivational factors may contribute to learning outcomes, they represent only one of multiple influencing factors.

XSL•FO
RenderX

Figure 4. Linear regression curve between the Motivated Strategies for Learning Questionnaire (MSLQ) total score and improvement in objective structured clinical examination (OSCE) scores for tendon reflex examination in the experimental group.



We further examined whether learning outcomes differed according to levels of technology acceptance. Participants were stratified into high- and low-scoring groups based on the median TAM score. No significant differences were observed in knowledge improvement (mean 0.75, SD 2.02 vs mean 0.75, SD 2.49; *P*>.99) or OSCE performance (mean 24.87, SD 15.56 vs mean 25.28, SD 12.23; *P*=.88) between groups. These findings suggest that, within the context of this study, variation in technology acceptance was not associated with differences in short-term learning outcomes, although this may reflect a threshold effect whereby basic usability was sufficient for effective participation.

Discussion

Principal Findings

This study evaluated the short-term effectiveness of synchronous remote teaching compared with traditional in-person teaching for neurological examination training among fifth-year medical students. Both teaching modalities were associated with improvements in neurological examination knowledge, tendon reflex OSCE performance, and self-reported confidence. Although the synchronous remote teaching group demonstrated modestly higher postcourse knowledge scores, no statistically significant between-group differences were observed in OSCE improvement after covariate adjustment. Exploratory analyses further suggested that motivational beliefs related to SRL may be associated with greater psychomotor skill improvement in synchronous remote learning environments.

The absence of meaningful between-group differences in tendon reflex OSCE improvement suggests that synchronous remote teaching may support short-term acquisition of selected psychomotor neurological examination skills under structured instructional conditions. One possible explanation is that tendon reflex assessment relies on standardized procedural sequences, visual observation, and repeated guided practice, all of which can be effectively demonstrated through synchronous video-based interaction [10,36]. In addition, the RCPS system enabled close-up visualization of tendon localization, examination technique, and patient positioning, while instructors provided immediate corrective feedback during skill performance. These findings are consistent with prior literature suggesting that synchronous remote teaching environments may support psychomotor clinical skills when real-time interaction and structured feedback are incorporated [17,36,37]. However, unlike many previous studies that relied primarily on self-reported outcomes or asynchronous formats, this study evaluated performance using an OSCE-based assessment focused on a specific neurological examination task.

As this study used a nonrandomized design, baseline imbalance in prior neurology rotation experience required careful interpretation. Participants with prior rotations demonstrated higher baseline self-reported confidence and greater OSCE performance, suggesting that prior clinical exposure may contribute to foundational procedural skills [38]. However, covariate-adjusted analyses and formal interaction testing did not demonstrate evidence that prior neurology rotation experience modified the effect of teaching modality on either knowledge or OSCE outcomes. These findings suggest that the observed learning outcomes associated with synchronous remote teaching appeared consistent across different baseline experience levels, although residual confounding cannot be fully excluded.

XSL•FO
RenderX

Theoretical Implications: Interpreting Findings Through the Lenses of SRL Theory and TAM

The exploratory association between motivational beliefs and OSCE improvement may be interpreted through the lens of SRL theory. In particular, self-efficacy and motivational regulation may support learners' engagement, persistence, and responsiveness to feedback during psychomotor skill acquisition [39-41]. These associations appeared more pronounced in the synchronous remote teaching group, suggesting that remote interactive environments may place greater emphasis on learners' active self-monitoring and self-regulatory capacity [38,42]. However, because these analyses were exploratory and this study was not designed to directly evaluate SRL mechanisms, these findings should be interpreted cautiously and regarded as hypothesis-generating.

In contrast, technology acceptance was not significantly associated with knowledge or OSCE outcomes. One possible interpretation is that, within highly interactive and instructor-guided synchronous environments, technology acceptance may function primarily as a prerequisite for participation rather than a direct determinant of learning effectiveness [21,43]. This distinction may help contextualize the role of TAM in synchronous psychomotor clinical skills training, although the present findings remain exploratory and do not establish theoretical mechanisms.

Collectively, these findings suggest that the effectiveness of synchronous remote psychomotor skills training may depend less on technology acceptance itself and more on how learners actively engage with feedback and regulate their own learning processes. From a theoretical perspective, this observation is broadly consistent with SRL models, which emphasize self-monitoring, self-efficacy, and strategic regulation as important determinants of learning performance in complex educational environments [22,23,42]. In contrast, TAM may be more useful for explaining learners' willingness to adopt and engage with educational technologies than for directly predicting skill acquisition outcomes once active participation has been established [19-21]. Although causal mechanisms cannot be inferred from this study, these findings contribute to the growing literature exploring how learner characteristics interact with technology-mediated clinical skills education and may help inform future theory-driven investigations of remote psychomotor skills training.

Practical and Real-World Implications

From a practical perspective, synchronous remote teaching may provide a flexible option for delivering neurological examination training when face-to-face instruction is constrained by geographical distance, scheduling limitations, infection control requirements, or limited access to specialist instructors. Enhanced telepresence systems and synchronous video-based educational environments may facilitate real-time demonstration, observation, and feedback while expanding access to clinical skills education across distributed training sites [17,29,36]. Although further evidence is required regarding scalability and long-term effectiveness, the present findings suggest that structured synchronous remote teaching may serve as a useful supplementary instructional approach for medical schools seeking to maintain interactive clinical skills training while addressing logistical challenges in contemporary medical education [36,37].

Limitations

Several limitations should be considered when interpreting the present findings. First, this study used a quasi-experimental design without randomization, which introduces potential selection bias and residual confounding. In particular, baseline imbalance in prior neurology rotation experience may have influenced learning outcomes despite the use of covariate-adjusted analyses and formal interaction testing. Therefore, the findings should be interpreted as associative rather than strictly causal. In addition, participants' awareness of being studied may have introduced a Hawthorne effect [44], potentially increasing motivation and engagement across both groups. Several subgroup and exploratory analyses were conducted without formal correction for multiple testing, which increases the potential risk of type I error; accordingly, these exploratory findings should be regarded as hypothesis-generating rather than confirmatory. Furthermore, self-reported confidence represents a subjective learner perception and may not fully correspond to objectively measured psychomotor performance.

Second, the study evaluated only immediate posttest outcomes and, therefore, does not establish long-term retention or durable skill acquisition. In addition, the directly observed psychomotor assessment focused specifically on tendon reflex examination skills, which represent only one component of the neurological examination; therefore, the findings should not be generalized to broader neurological examination competencies without further study. The pragmatic design also did not include a formal a priori sample size calculation, and unequal group sizes resulted from cohort availability and scheduling constraints, which may have limited statistical power to detect small between-group differences. Finally, the synchronous remote teaching sessions were conducted in relatively small groups with substantial instructor interaction and individualized feedback; therefore, the scalability and effectiveness of this instructional approach in larger educational settings remain uncertain.

Conclusions

Under the conditions of this quasi-experimental study, synchronous remote teaching was associated with short-term improvements in tendon reflex examination skills, neurological examination–related knowledge, and self-reported confidence among fifth-year medical students. Although modestly higher postcourse knowledge scores were observed in the synchronous remote teaching group, no statistically significant between-group differences were identified in OSCE-assessed psychomotor performance after covariate adjustment. These findings suggest that structured synchronous remote teaching may support short-term acquisition of selected neurological examination skills when real-time interaction, guided practice, and immediate feedback are incorporated into instructional design.

To our knowledge, this study is among the first to evaluate synchronous remote teaching for tendon reflex examination training using OSCE-based psychomotor assessment and

XSL•FO
RenderX

theory-informed interpretation grounded in SRL and technology acceptance frameworks. These findings contribute to the growing literature on synchronous digital clinical skills education by extending prior research that has primarily focused on asynchronous learning or subjective learner perceptions. However, because this study assessed only immediate outcomes and focused on tendon reflex examination skills, the findings should not be interpreted as evidence of long-term skill retention or generalized effectiveness across the full spectrum of neurological examination competencies. Future research should incorporate randomized designs, larger multicenter cohorts, and longitudinal follow-up to evaluate scalability, long-term retention, transferability to other neurological examination domains, and strategies to support SRL in remote clinical skills education.

Acknowledgments

The authors thank Daniel Salcedo for granting permission to use the Remote Clinical Presence System (RCPS) in this study. The authors also thank ATS Medical Editing for providing English-language editing support. The authors declare the use of generative AI (GAI) in the research and writing process. According to the GAI Delegation Taxonomy (GAIDeT) taxonomy (2025) [45], the following tasks were delegated to GAI tools under full human supervision: proofreading, editing, and adapting and adjusting the emotional tone. The GAI tool used was ChatGPT 5.5 (OpenAI). Responsibility for the final paper lies entirely with the authors. GAI tools are not listed as authors and do not bear responsibility for the final outcomes. The declaration submitted by the authors. MSW and KPT are co-corresponding authors on this work, and the latter can be reached by email at kptang@mail.ntue.edu.tw.

Funding

This study was funded by the 2023 Key Research Program of Taipei Municipal Wan Fang Hospital (approval 112-WF-EVA-35).

Authors' Contributions

CSH drafted the paper and prepared the tables and figures. KPT and MSW critically revised the paper for important intellectual content.

Conflicts of Interest

None declared.

References

1. Hall AK, Nousiainen MT, Campisi P, et al. Training disrupted: practical tips for supporting competency-based medical education during the COVID-19 pandemic. Med Teach 2020 Jul;42(7):756-761. [doi: 10.1080/0142159X.2020.1766669] [Medline: 32450049]
2. Sandars J, Correia R, Dankbaar M, et al. Twelve tips for rapidly migrating to online learning during the COVID-19 pandemic. MedEdPublish (2016) 2020;9(82):82. [doi: 10.15694/mep.2020.000082.1] [Medline: 38058947]
3. Rose S. Medical student education in the time of COVID-19. JAMA 2020 Jun 2;323(21):2131-2132. [doi: 10.1001/jama.2020.5227] [Medline: 32232420]
4. Pei L, Wu H. Does online learning work better than offline learning in undergraduate medical education? A systematic review and meta-analysis. Med Educ Online 2019 Dec;24(1):1666538. [doi: 10.1080/10872981.2019.1666538] [Medline: 31526248]
5. Stöhr C, Demazière C, Adawi T. The polarizing effect of the online flipped classroom. Comput Educ 2020 Apr;147:103789. [doi: 10.1016/j.compedu.2019.103789]
6. Alzahrani HA, Shati AA, Bawahab MA, et al. Students' perception of asynchronous versus synchronous distance learning during COVID-19 pandemic in a medical college, southwestern region of Saudi Arabia. BMC Med Educ 2023 Jan 23;23(1):53. [doi: 10.1186/s12909-023-04034-5] [Medline: 36691003]
7. Regmi K, Jones L. A systematic review of the factors - enablers and barriers - affecting e-learning in health sciences education. BMC Med Educ 2020 Mar 30;20(1):91. [doi: 10.1186/s12909-020-02007-6] [Medline: 32228560]
8. Hughes T. Neurology and the clinical anatomist. Medicine (Abingdon) 2020 Aug;48(8):497-501. [doi: 10.1016/j.mpmed.2020.05.006] [Medline: 32834734]
9. Campbell WW, Barohn RJ. DeJong's The Neurologic Examination, 8th edition: Wolters Kluwer; 2020.
10. Peyton JWR. Teaching & Learning in Medical Practice: Manticore Europe Ltd; 1998.
11. Chick RC, Clifton GT, Peace KM, et al. Using technology to maintain the education of residents during the COVID-19 pandemic. J Surg Educ 2020;77(4):729-732. [doi: 10.1016/j.jsurg.2020.03.018] [Medline: 32253133]
12. Khalil R, Mansour AE, Fadda WA, et al. The sudden transition to synchronized online learning during the COVID-19 pandemic in Saudi Arabia: a qualitative study exploring medical students' perspectives. BMC Med Educ 2020 Aug 28;20(1):285. [doi: 10.1186/s12909-020-02208-z] [Medline: 32859188]

XSL•FO
RenderX

13. Wilcha RJ. Effectiveness of virtual medical teaching during the COVID-19 crisis: systematic review. JMIR Med Educ 2020 Nov 18;6(2):e20963. [doi: 10.2196/20963] [Medline: 33106227]
14. Khan RA, Atta K, Sajjad M, Jawaid M. Twelve tips to enhance student engagement in synchronous online teaching and learning. Med Teach 2022 Jun;44(6):601-606. [doi: 10.1080/0142159X.2021.1912310] [Medline: 33877950]
15. Dinh CT. Impact of synchronous online learning environment on students' cognitive engagement and learning outcomes. TOJDE 2023;24(3):21-38. [doi: 10.17718/tojde.1165209]
16. Iancu AM, Kemp MT, Alam HB. Unmuting medical students' education: utilizing telemedicine during the COVID-19 pandemic and beyond. J Med Internet Res 2020 Jul 20;22(7):e19667. [doi: 10.2196/19667] [Medline: 32614776]
17. Belt ES, Lowenthal PR. Synchronous video-based communication and online learning: an exploration of instructors' perceptions and experiences. Educ Inf Technol (Dordr) 2023;28(5):4941-4964. [doi: 10.1007/s10639-022-11360-6] [Medline: 36320822]
18. Zhang RN, Bi NC, Mercado T. Do Zoom meetings really help? A comparative analysis of synchronous and asynchronous online learning during COVID-19 pandemic. J Comput Assist Learn 2022 Sep 4;39(1):210-217. [doi: 10.1111/jcal.12740] [Medline: 36247930]
19. Davis FD. Perceived usefulness, perceived ease of use, and user acceptance of information technology. MIS Q 1989 Sep 1;13(3):319-340. [doi: 10.2307/249008]
20. Venkatesh V, Davis FD. A theoretical extension of the technology acceptance model: four longitudinal field studies. Manage Sci 2000 Feb;46(2):186-204. [doi: 10.1287/mnsc.46.2.186.11926]
21. Jiang S, Li H, Gan D. Technology acceptance model for online education: identifying interdisciplinary topics and their evolution based on BERTopic model. Social Sciences & Humanities Open 2025;12:101831. [doi: 10.1016/j.ssaho.2025.101831]
22. Zimmerman BJ. Becoming a self-regulated learner: an overview. Theory Pract 2002 May;41(2):64-70. [doi: 10.1207/s15430421tip4102_2]
23. Pintrich PR. The Role of Goal Orientation in Self-Regulated Learning Handbook of Self-Regulation: Academic Press; 2000:451-502. [doi: 10.1016/B978-012109890-2/50043-3]
24. Gordon M, Patricio M, Horne L, et al. Developments in medical education in response to the COVID-19 pandemic: a rapid BEME systematic review: BEME Guide No. 63. Med Teach 2020 Nov;42(11):1202-1215. [doi: 10.1080/0142159X.2020.1807484] [Medline: 32847456]
25. Lau KHV, Hamlyn E, Williams TJ, et al. The effects of video instruction on neuroscience intensive care unit nursing skills in case presentations and neurological examinations. J Neurosci Nurs 2021 Jun 1;53(3):129-133. [doi: 10.1097/JNN.0000000000000591] [Medline: 33840806]
26. Rutherford-Hemming T, Kelsey NC, Grenig DL, Feliciano M, Simko L, Henrich CM. Multisite single-blinded randomized control study of transfer and retention of knowledge and skill between nurses using simulation and online self-study module. Simul Healthc 2016 Aug;11(4):264-270. [doi: 10.1097/SIH.0000000000000168] [Medline: 27388860]
27. Shin JY, Issenberg SB, Roh YS. The effects of neurologic assessment e-learning in nurses. Nurse Educ Today 2017 Oct;57:60-64. [doi: 10.1016/j.nedt.2017.07.007] [Medline: 28734188]
28. Harth C, Xue Y, Yang J. Internet teaching of the neurologic examination and OSCE performance in the neurology clerkship (s19.006). Neurology 2015 Apr 6;84(14_supplement). [doi: 10.1212/WNL.84.14_supplement.s19.006]
29. Salcedo D, Huang WC, Suk FM, Lin CW. Enhanced telepresence for clinical training in high-risk environments. Med Educ 2020 Nov;54(11):1063-1064. [doi: 10.1111/medu.14330] [Medline: 32885459]
30. Gegenfurtner A, Ebner C. Webinars in higher education and professional training: a meta-analysis and systematic review of randomized controlled trials. Educational Research Review 2019 Nov;28:100293. [doi: 10.1016/j.edurev.2019.100293]
31. Ødegaard NB, Myrhaug HT, Dahl-Michelsen T, Røe Y. Digital learning designs in physiotherapy education: a systematic review and meta-analysis. BMC Med Educ 2021 Jan 13;21(1):48. [doi: 10.1186/s12909-020-02483-w] [Medline: 33441140]
32. Oster C, Farhood I, Klebe S, Kleinschnitz C, Peters L. Neurological examination course in an interactive webinar as a solution during a pandemic. An overview of the implementation, optimization as well as critical considerations. GMS J Med Educ 2021;38(1):Doc9. [doi: 10.3205/zma001405] [Medline: 33659614]
33. Pintrich PR, Smith DAF, García T, McKeachie WJ. A Manual for the Use of the Motivated Strategies for Learning Questionnaire (MSLQ): National Center for Research to Improve Postsecondary Teaching and Learning, University of Michigan; 1991. URL: https://eric.ed.gov/?id=ED338122 [accessed 2026-06-30]
34. Davis FD. User acceptance of information technology: system characteristics, user perceptions and behavioral impacts. Int J Man Mach Stud 1993 Mar;38(3):475-487. [doi: 10.1006/imms.1993.1022]
35. Rovinelli RJ, Hambleton RK. On the use of content specialists in the assessment of criterion-referenced test item validity. Tijdschr Onderwijs Res 1977;2(2):49-60 [FREE Full text]
36. He L, Yang N, Xu L, et al. Synchronous distance education vs traditional education for health science students: a systematic review and meta-analysis. Med Educ 2021 Mar;55(3):293-308. [doi: 10.1111/medu.14364] [Medline: 32881047]
37. Martin F, Sun T, Turk M, Ritzhaupt A. A meta-analysis on the effects of synchronous online learning on cognitive and affective educational outcomes. IRRODL 2021;22(3):205-242. [doi: 10.19173/irrodl.v22i3.5263]

XSL•FO
RenderX

38. Turan S, Konan A. Self-regulated learning strategies used in surgical clerkship and the relationship with clinical achievement. J Surg Educ 2012;69(2):218-225. [doi: 10.1016/j.jsurg.2011.09.003] [Medline: 22365869]
39. Adesope OO, Zhou MM, Nesbit JC. Achievement goal orientations and self-reported study strategies as predictors of online studying activities. J Educ Comput Res 2015 Dec;53(3):436-458. [doi: 10.1177/0735633115603989]
40. Quesada-Pallarès C, Sánchez-Martí A, Ciraso-Calí A, Pineda-Herrero P. Online vs. classroom learning: examining motivational and self-regulated learning strategies among vocational education and training students. Front Psychol 2019;10(2795):2795. [doi: 10.3389/fpsyg.2019.02795] [Medline: 31920839]
41. Stegers-Jager KM, Cohen-Schotanus J, Themmen APN. Motivation, learning strategies, participation and medical school performance. Med Educ 2012 Jul;46(7):678-688. [doi: 10.1111/j.1365-2923.2012.04284.x] [Medline: 22691147]
42. Zheng B, Sun T. Self-regulated learning and learning outcomes in undergraduate and graduate medical education: a meta-analysis. Eval Health Prof 2025 Dec;48(4):430-450. [doi: 10.1177/01632787241288849] [Medline: 39361881]
43. Wong G, Greenhalgh T, Pawson R. Internet-based medical education: a realist review of what works, for whom and in what circumstances. BMC Med Educ 2010 Feb 2;10(12):12. [doi: 10.1186/1472-6920-10-12] [Medline: 20122253]
44. Merrett F. Reflections on the Hawthorne effect. Educ Psychol (Lond) 2006 Jan;26(1):143-146. [doi: 10.1080/01443410500341080]
45. Suchikova Y, Tsybuliak N, Teixeira da Silva JA, Nazarovets S. GAIDeT (Generative AI Delegation Taxonomy): a taxonomy for humans to delegate tasks to generative artificial intelligence in scientific research and publishing. Account Res 2026 Apr;33(3):2544331. [doi: 10.1080/08989621.2025.2544331] [Medline: 40781729]

Abbreviations

ANCOVA: analysis of covariance
MSLQ: Motivated Strategies for Learning Questionnaire
OSCE: objective structured clinical examination
RCPS: Remote Clinical Presence System
SRL: self-regulated learning
TAM: technology acceptance model

Edited by S Brini; submitted 07.May.2025; peer-reviewed by T Peng, V Reis; revised version received 10.Jun.2026; accepted 11.Jun.2026; published 15.Jul.2026.

Please cite as:
Huang CS, Tang KP, Wu MS
Synchronous Remote Teaching for Neurological Examination Training in Fifth-Year Medical Students: Quasi-Experimental Study
JMIR Med Educ 2026;12:e77034
URL: https://mededu.jmir.org/2026/1/e77034
doi:10.2196/77034

© Chih-Shan Huang, Kung-Pei Tang, Ming-Shun Wu. Originally published in JMIR Medical Education (https://mededu.jmir.org), 15.Jul.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Education, is properly cited. The complete bibliographic information, a link to the original publication on https://mededu.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Fostering Inductive and Deductive Learning in Oral Microbiology and Immunology With a Dual-Role Duel Card Game: Explanatory Sequential Mixed Methods Study

Kawin Sipiyaruk[1], DDS, PhD; Ratchapin Laovanitch Srisatjaluk[2], PhD; Ponpentip Chaisiri[3]; Kanokwan Jantawatchai[3]; Jidapha Srihatai[3]; Chutikarn Sonyaem[3]; Thanaphat Traikorakot[3]; Suphasuta Pookboonmee[4]

[1]Department of Orthodontics, Faculty of Dentistry, Mahidol University, 6 Yothi Road, Ratchathewi, Bangkok, Thailand

[2]Mahidol International Dental School, Faculty of Dentistry, Mahidol University, Bangkok, Thailand

[3]Doctor of Dental Surgery Program, Faculty of Dentistry, Mahidol University, Bangkok, Thailand

[4]Medical Education Technology Center, Faculty of Medicine Siriraj Hospital, Mahidol University, Bangkok, Thailand

Corresponding Author:
Kawin Sipiyaruk, DDS, PhD
Department of Orthodontics, Faculty of Dentistry, Mahidol University, 6 Yothi Road, Ratchathewi, Bangkok, Thailand

Abstract

Background: Game-based learning has emerged as an effective learning strategy in health care education. However, no games have been specifically designed to support cognitive improvement for diverse learning styles in oral microbiology and immunology.

Objective: This study aimed to develop and evaluate an educational card game designed to support diverse learning styles in oral microbiology and immunology, using a duel-style format.

Methods: An explanatory sequential mixed methods study was conducted with 40 third-year dental students, where half of them were assigned to the first group, starting as the host, while those in the other group began as the microbe. Participants alternated between the microbe and host roles during gameplay. Active engagement through playing as the microbe facilitated knowledge acquisition through observation, supporting inductive learning. On the other hand, the host role aimed to promote the application of knowledge for decision-making, facilitating deductive learning. Quantitative data were collected using pre- and postknowledge assessments and satisfaction questionnaires. Qualitative insights were obtained through semistructured interviews exploring learning experiences when playing as the microbe compared to the host.

Results: Students demonstrated significant improvements in knowledge scores across the 3 assessments (P<.001), with no difference between groups (P=.85). They also perceived the game positively in all 3 aspects (usefulness, ease of use, and enjoyment). Qualitative findings revealed that role variation supported both inductive and deductive learning processes. Participants valued the combination of pedagogical and entertaining components, leading to the game's motivation and engagement. A conceptual framework demonstrated key emerging themes relevant to the game design and implementation, including learner profile, learning setting, game design, learning process, and learning outcomes.

Conclusions: The card game demonstrated its potential in enhancing knowledge acquisition and student engagement in oral microbiology and immunology. Role-switching between the host and microbe was perceived by participants to facilitate different learning experiences. Further research is recommended to investigate long-term retention and broader practicality.

(JMIR Med Educ 2026;12:e80048) doi:10.2196/80048

KEYWORDS

active learning; card game; educational game; game-based learning; microbiology; immunology

Introduction

Oral microbiology provides essential insights into bacteria, viruses, and fungi in the oral cavity, while immunology focuses on the mechanisms by which the human body defends itself against these pathogens. This knowledge enables health care professionals, including dental practitioners, to understand pathogen transmission, immune interactions, and strategies for disease management and prevention [1-3]. The COVID-19 pandemic, for example, significantly impacted dental practices by introducing new protocols to reduce viral transmission through respiratory droplets [4,5]. Studying microbiology and immunology is therefore important for understanding infectious diseases, and it has proven essential for health care education.

Traditional teaching methods have served as the main method of education worldwide. While this approach can reduce anxiety by minimizing the need for active student participation, it can limit engagement due to the predominantly 1-way

XSL•FO
RenderX

communication [6,7]. Disengagement in traditional classrooms has been shown to negatively impact academic performance [8-10]. On the other hand, active learning strategies, which promote critical thinking and student interaction, are associated with better engagement [11,12]. In addition, not all students have the same learning preferences [13]. Some of them prefer direct participation and decision-making, while others may learn more effectively by observing patterns. Educational tools designed for both learning preferences may support diverse learning needs.

Game-based learning has been considered an innovative strategy to address the limitations of traditional teaching methods by incorporating game mechanics into educational settings. Game-based learning has been applied in medical and dental education [14-17], including microbiology and immunology [18,19]. This approach is widely recognized due to its ability to enhance student engagement and motivation in various disciplines of dental education, such as dental public health [20], human anatomy [21], oral diagnosis [22,23], orthodontics [24], endodontics [25], patient safety [26], clinical administration [27], and teledentistry [28]. These games can be implemented through various formats, ranging from digital simulations to physical formats such as board and card games.

Despite the growing number of educational games, most of them are not designed to accommodate a range of learning preferences or gameplay experience levels. In other words, they may not explicitly support both inductive (structured observation) and deductive learning (active problem-solving), which are 2 distinct approaches through which students engage with and make sense of content. This study aimed to develop and evaluate an educational card game designed to support diverse learning needs in oral microbiology and immunology, using a duel-style format in which players alternate between the roles of microbe and host to simulate infection and immune response dynamics. Specifically, the objectives were to assess the game's effectiveness in improving knowledge acquisition and to investigate how learners engaged differently when beginning as the microbe compared to the host.

Methods

Game Development

Game Concept

The "Invasion" game was designed to support teaching and learning in oral microbiology and immunology, where players alternate roles between microbes and hosts. This game served as a novel educational tool to engage dental students more deeply with the learning content. The game was designed to align with the expected learning outcomes in oral microbiology and immunology, requiring learners to apply their knowledge of microbial pathogenesis and human immune responses to infection. The role alternation within the game provides different forms of engagement, encouraging learners to interact with the content either through structured observation as microbes (inductive learning) or through direct decision-making as hosts (deductive learning). A physical card format was chosen because it encourages face-to-face interaction, peer discussion, and active participation, which strengthens the learning experience when students switch between roles. In addition, compared with a digital format, this design helps students remain focused on the learning activity itself rather than being distracted by technological navigation.

Instructional Design and Learning Tasks

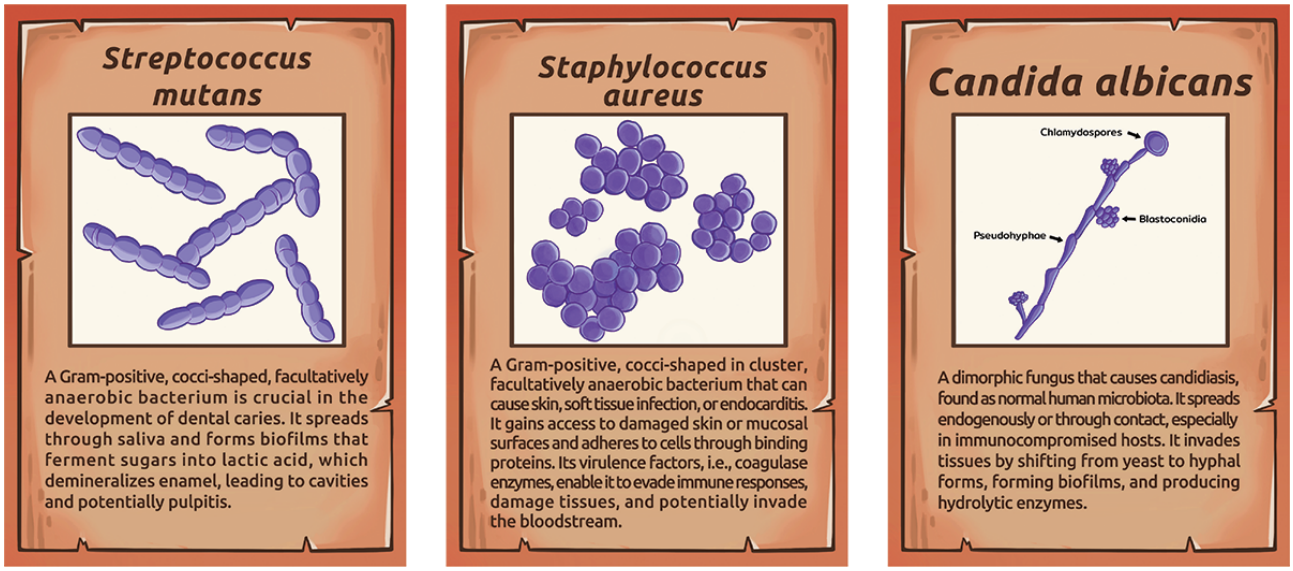
The pedagogical framework of this game is based on the constructivist learning theory, where learners are expected to actively construct new knowledge through problem-solving and critical reflection, based on their existing understanding. Engagement with the game elements can allow students to develop a deeper understanding of microbiological and immunological concepts. The dual-role format supports different types of cognitive engagement. Playing as the microbe emphasizes reflective observation and the construction of clues (inductive learning), while the host role requires active problem-solving and the application of knowledge to identify pathogens and apply appropriate responses (deductive learning). Another important component of the game is the feedback mechanism. Each action taken by the player provided an opportunity for formative assessment, allowing for real-time correction and learning from mistakes. The feedback mechanism can support learners to adjust their strategies and understandings during the gameplay. Through this process of testing, feedback, and adjustment, students can engage in a learning cycle that mirrors scientific reasoning and clinical decision-making.

Game Interface and Interaction

According to the gameplay mechanics, the game begins with the microbe player selecting 3 "Microbe" cards, along with their corresponding "Key" cards and 7 "Hint" cards (Figure 1). The host, in turn, starts the game with 5 heart points and a set of 7 "Protection" cards. During gameplay, the microbe presents "Hint" cards, prompting the host to select 1 of 3 actions: play up to 3 "Protection" cards in response to the clues, discard 3 "Protection" cards in exchange for a "Key" card offering more detailed information, or use an "Attack" card to guess the identity of a microbe. A correct guess leads to the revelation of that microbe, while an incorrect guess results in a loss of heart points calculated as the number of "Hint" cards played minus the number of correctly placed "Protection" cards. Game completion occurs when the host successfully identifies 3 microbes, resulting in a host victory, or when all heart points or "Attack" cards are used, which means the microbe player wins.

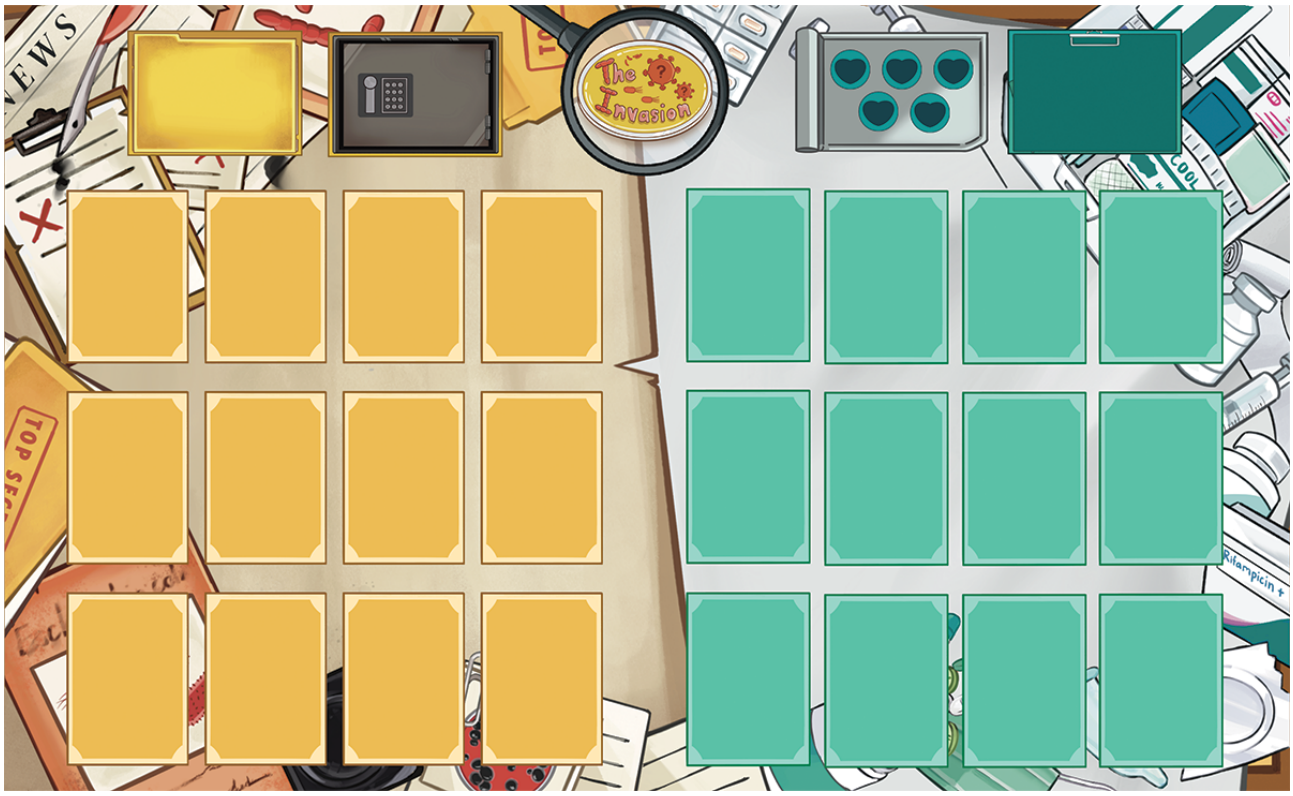
XSL•FO
RenderX

Figure 1. Examples of "Microbe" cards used in the "Invasion" game.



To further support gameplay, a cardboard was designed as an organizing component (Figure 2). The cardboard provided designated spaces for different types of cards (eg, microbe, host, or support actions), guiding players in organizing card placement throughout the game. It was intended to reduce confusion during gameplay, thereby enhancing usability and the overall learning experience. Its purpose was to minimize the cognitive load related to card management, allowing players to focus more on the learning tasks.

Figure 2. The cardboard with designated spaces for card placement in the "Invasion" game.



Effective gameplay requires the microbe player to present hints in a way that is strategically challenging to test the host's interpretation skills. In turn, the host must analyze the clues, make careful decisions about resource use, and minimize incorrect guesses in order to preserve their heart points. These interactive dynamics create a learning environment that reflects the cognitive and analytical challenges involved in diagnosing and managing microbial infections in clinical practice. By placing students in alternating roles with distinct cognitive demands, the game supports the development of microbiological understanding, analytical thinking, and adaptability, which are skills essential for clinical reasoning and for meeting the varied learning needs of students. This gameplay structure is designed to engage students in different types of thinking, depending on their assigned role. A summary of the game components is provided in Table 1.

XSL•FO
RenderX

Table . Description of card types used in the game.

Card type	Used by	Description
Microbe cards	Microbe	Represent various microbes selected for their clinical relevance in dentistry. These cards define the core challenges posed by the microbe player during gameplay.
Protection cards	Host	Provide the host with strategies to prevent or reduce damage from incorrect microbial identification. Each card contains information on treatments, immune responses, or behavioral changes.
Hint cards	Microbe	Contain clues related to the microbe's characteristics, such as virulence factors, clinical manifestations, and transmission routes. They support progressive disclosure of information to the host.
Key cards	Microbe	Offer more specific and detailed information about a microbe than "Hint" cards, enhancing the host's ability to make accurate identifications. Obtained through card exchange.
Attack cards	Host	Allow the host to make an informed guess about the identity of a microbe based on accumulated clues. Limited to 5 per game, requiring careful use.

Research Design

This study used an explanatory mixed methods design to evaluate the "Invasion" game and to examine differences in learning experiences and outcomes when students engaged in gameplay as either a microbe or a host.

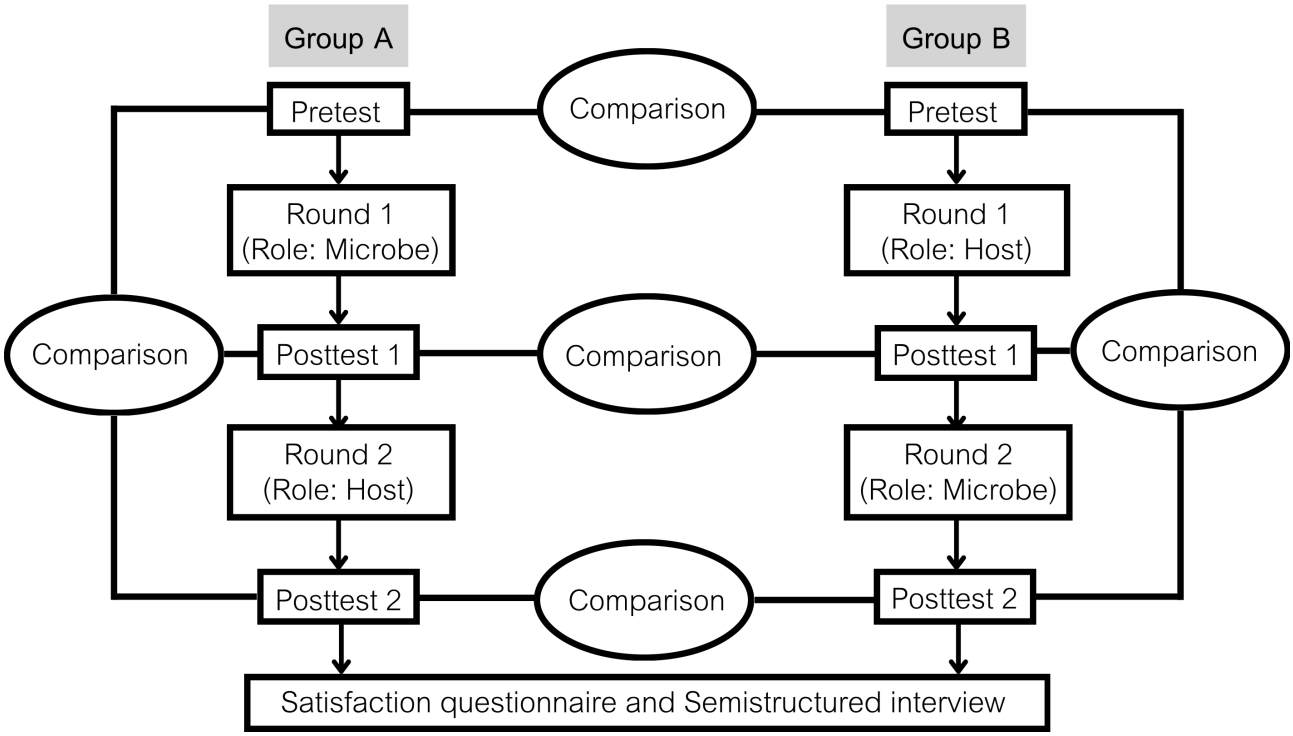
The quantitative phase used an experimental design with partial randomization to assess knowledge acquisition and compare learning outcomes based on gameplay role order. A total of 40 participants were allowed to self-select their partners, forming pairs based on personal preference to reflect natural peer interactions, commonly found in educational settings [21]. Within each pair, roles were randomly assigned using a coin toss method to determine which player would begin as the host or the microbe. Prior to gameplay, participants completed a preknowledge test (Pretest), followed by 1 round of the "Invasion" game in their assigned role. A postknowledge test (Posttest 1) was administered immediately after the first round. In the second round, participants switched roles and completed another postknowledge test (Posttest 2). A satisfaction questionnaire was also completed at the end of the second round. All test items, game materials, and instructions were provided in English.

For the qualitative phase, all participants took part in short semistructured interviews to explore their experiences and learning processes while engaging with both gameplay roles. This approach provided clarification on how the game's dual-role structure facilitated distinct forms of engagement, with students tending to adopt a more reflective and observational approach when playing as the microbe, and a more analytical and decision-driven approach when playing as the host. All 40 participants were interviewed to ensure full representation of perspectives across both gameplay roles and to enhance the integration of qualitative insights with quantitative findings. This decision was made because the approach was intended not only to achieve thematic saturation but also to ensure that the qualitative data contextualized the quantitative results, while the brief and focused nature of the interviews made it feasible to include all participants. The semistructured format allowed for a consistent set of guiding questions while providing flexibility to probe deeper into participant reflections. Interviews were conducted immediately after the quantitative phase to ensure relevance and recall.

All procedures were completed in a single session, where each game round (host or microbe) lasted approximately 25 - 30 minutes. All participants completed their activities under the same conditions, and knowledge assessments were completed independently within a fixed time frame. This standardized approach ensured consistency and comparability of learning experiences across groups. The full study protocol is illustrated in Figure 3.

XSL•FO
RenderX

Figure 3. Diagram presenting the outline of the study design.



Research Participants

The research population consisted of third-year undergraduate dental students enrolled in the 2024 academic year at the Faculty of Dentistry, Mahidol University. They were eligible if they had successfully completed 3 prerequisite courses: Fundamentals of Medical Microbiology for Dental Students, Essential Immunology for Dental Students, and Oral Microbiology and Immunology. Completion of these courses ensured a baseline understanding of core concepts relevant to the game content. While all game materials and knowledge assessments were designed in English, students with insufficient English proficiency to understand the game instructions were excluded from the study to ensure that language barriers did not affect gameplay engagement or performance. This exclusion was determined through an orientation session in which students were asked to demonstrate comprehension of the English instructions. Additionally, any participant who failed to complete a required task within the allocated time was withdrawn.

The required sample size was determined using G*Power software (version 3.1; Heinrich-Heine-Universität Düsseldorf). According to a previous study [21], the mean and SD of the pretest score were 9.37 and 2.40, respectively, while those of the posttest score were 11.37 and 1.96. With a statistical power of 95% and a significance level of P=.85, the required sample size per group was 18, resulting in a minimum total of 36 participants. To compensate for a potential 10% dropout rate, the sample size was set at 40 participants (20 per group).

Participant allocation followed a 2-step process. Students first formed pairs through voluntary matching. Within each pair, a coin toss was used to randomly assign 1 participant to begin as the host and the other as the microbe. This randomization step helped mitigate potential bias introduced by voluntary pairing. Allowing students to choose their own partners made the activity feel more natural, while randomly assigning gameplay roles ensured a fair comparison between the host and microbe roles. By assigning roles randomly, the study preserved experimental control and enabled the qualitative phase to capture the experiences of participants who may have been placed in roles misaligned with their learning preferences. This role mismatch provided valuable insight into the cognitive and emotional challenges encountered by different learners, offering a more comprehensive understanding of the game's effectiveness and highlighting both the strengths and limitations of its design across diverse learner profiles.

Data Collection Tools

Knowledge Assessments

To evaluate knowledge acquisition, 3 structured assessments were administered at different stages of the study: Pretest, Posttest 1, and Posttest 2. Pretest aimed to assess students' baseline understanding and help identify individual learning gaps prior to the intervention. Posttest 1 and Posttest 2 were used to measure knowledge gains following participation in the learning activities. Each assessment consisted of the same 20 multiple-choice questions to ensure consistency in content and difficulty. However, to minimize the influence of recall bias, both the order of the questions and the arrangement of answer choices were randomized across the assessments.

Content validity was confirmed through expert review by 3 microbiologists who were not involved in the study. These experts evaluated each question for its relevance and alignment with the expected learning outcomes. The Index of Item-Objective Congruence (IOC) was used, with items scoring below 0.5 being revised based on expert feedback. Face validity was assessed by a group of fourth-year dental students who had completed the same microbiology and immunology courses required for inclusion in the study. Involving these senior

XSL•FO
RenderX

students helped ensure that the test items were appropriate and clearly worded, and this approach prevented contamination of the research participants.

Satisfaction Questionnaire

To investigate user perceptions of the learning experience, a self-administered questionnaire comprising 18 items was used. Each item was rated on a 5-point Likert scale, ranging from 1 ("Strongly disagree") to 5 ("Strongly agree"), and adapted from previously validated instruments [22,27,28]. The questionnaire assessed 3 key dimensions commonly used in the evaluation of serious games: perceived usefulness, ease of use, and enjoyment. Content validity was established through expert review by 3 dental educators, who evaluated each item for relevance and alignment with the intended constructs. Items with an IOC below 0.5 were revised through iterative feedback until acceptable agreement was achieved.

Semistructured Interviews

To complement the quantitative findings on how different gameplay roles supported varied learning approaches, semistructured interviews were conducted using a topic guide developed by the research team. The topic guide comprised 5 main areas, with flexibility to probe further based on participants' responses. These questions examined students' reflections on the learning experience, the influence of assigned roles (microbe or host) on learning and engagement, and suggestions for improvement. Interview responses were audio-recorded with permission and transcribed using a verbatim technique. This process ensured that the learning experiences were captured with depth and fidelity.

Data Analysis

Quantitative data were analyzed using IBM SPSS Statistics for Windows, version 29.0. Descriptive statistics were used to summarize participant characteristics and assessment scores. A 2-way repeated-measures ANOVA was conducted to examine the effects of assigned role order (host and microbe) and time (Pretest, Posttest 1, and Posttest 2). Bonferroni-adjusted post hoc tests were used to explore pairwise differences across time and between groups. Assumptions of normality and sphericity were verified prior to analysis to ensure that the requirements for repeated-measures ANOVA were met. Data from the satisfaction questionnaire were analyzed using means and SDs. Statistical significance was set at $P<.05$.

Qualitative data were analyzed using thematic analysis. Two researchers independently reviewed the transcripts to identify recurring patterns, followed by discussion and consolidation of themes with the corresponding author. The analysis specifically examined how the 2 gameplay roles (host and microbe) contributed to learners' knowledge improvement. This focus aligned with the research objective of supporting diverse learner profiles through role variation. Credibility, consistency, and transparency were prioritized throughout the analytic process to ensure the rigor of interpretation.

Ethical Considerations

This study received ethical approval from the Institutional Review Board of the Faculty of Dentistry and the Faculty of Pharmacy, Mahidol University (MU-MOU CoA 2025/013.1002). Before data collection, all participants were provided with a detailed information sheet and given written informed consent. To protect confidentiality, all data were anonymized, and no personally identifiable information was available during analysis. No compensation was provided to participants.

Results

Research Participants

A total of 40 third-year undergraduate dental students participated in the study. Participants were randomly assigned to 2 groups of 20 students each. Group A consisted of students who began as the microbe (6 male and 14 female), while Group B included those who started as the host (3 male and 17 female). All participants completed the assigned tasks, and no dropouts occurred during the study.

Knowledge Assessments

Participants in both groups demonstrated progressive improvements in knowledge scores across the 3 assessments (Table 2). For Group A, the mean score increased from 9.85 (SD 3.72) at Pretest to 12.4 (SD 3.65) at Posttest 1, and further to 14.55 (SD 3.38) at Posttest 2. Similarly, Group B improved from 9.55 (SD 2.44) at Pretest to 11.8 (SD 3.07) at Posttest 1, and then to 15.9 (SD 2.05) at Posttest 2. These results suggest that both groups experienced consistent knowledge gains following each round of gameplay.

Table . Knowledge assessments across time points by group.

Group	Assessment	Mean score (SD)	95% CI
A	Pretest	9.85 (3.72)	8.43 to 11.27
	Posttest 1	12.4 (3.65)	10.87 to 13.93
	Posttest 2	14.55 (3.38)	13.29 to 15.82
B	Pretest	9.55 (2.44)	8.13 to 10.97
	Posttest 1	11.8 (3.07)	10.27 to 13.33
	Posttest 2	15.9 (2.05)	14.64 to 17.17

XSL•FO
RenderX

Effects of Time and Group (A 2-Way Repeated-Measures ANOVA)

A 2-way repeated-measures ANOVA was conducted to examine improvements in knowledge scores over time of assessments and whether there were differences between the 2 groups (host-first vs microbe-first). According to Table 3, the analysis demonstrated a significant main effect of time ($F_{2,76}$=54.02; P<.001), indicating that there were improvements in knowledge scores across the 3 assessments. However, the interaction between time and group was not statistically significant ($F_{2,76}$=1.94; P=.15), suggesting that both groups demonstrated a similar pattern of improvement over time. The between-subjects analysis showed no significant main effect of group on knowledge scores ($F_{1,38}$=0.038; P=.85), indicating that initial group assignment did not influence overall performance (Table 4).

Table . Results of 2-way repeated-measures ANOVA examining within-subjects effects.

Source	Type III sum of squares	Mean square	*F* test (*df*)	*P* value	ηp2
Time	614.017	307.008	54.02 (2, 76)	<.001	.587
Time × group	22.050	11.025	1.94 (2, 76)	.15	.049
Error (time)	431.933	5.683	—[a]	—	—

[a]Not applicable.

Table . Results of 2-way repeated-measures ANOVA examining between-subjects effects.

Source	Type III sum of squares	Mean square	*F* test (*df*)	*P* value	ηp2
Group	0.675	0.675	0.04 (1, 38)	.85	.001
Error (group)	672.317	17.693	—[a]	—	—

[a]Not applicable.

Pairwise Comparisons of Assessment Time Points

Given the absence of a significant interaction between time and group and no main effect of group, Bonferroni-adjusted pairwise comparisons were conducted by combining data from both groups. As shown in Table 5, knowledge scores increased significantly between each assessment point: from Pretest to Posttest 1 (P<.001), from Posttest 1 to Posttest 2 (P<.001), and from Pretest to Posttest 2 (P<.001). These findings reflect a cumulative learning effect over time. To further explore learning progression within each role-assigned group, Bonferroni-adjusted pairwise comparisons were performed separately within each group (Table 6).

Table . Bonferroni-adjusted pairwise comparisons of knowledge scores across assessment time points.

Comparison		Mean difference (95% CI)	SE	*P* value
Pretest	Posttest 1	−2.4 (−3.696 to −1.104)	0.517	<.001
Pretest	Posttest 2	−5.525 (−6.889 to −4.161)	0.545	<.001
Posttest 1	Posttest 2	−3.125 (−4.47 to −1.78)	0.537	<.001

XSL•FO
RenderX

Table . Bonferroni-adjusted pairwise comparisons of knowledge scores across assessment time points within each group.

Group	Comparison		Mean difference (95% CI)	SE	*P* value
A	Pretest	Posttest 1	−2.55 (−4.382 to −0.718)	0.732	.004
	Pretest	Posttest 2	−4.70 (−6.629 to −2.771)	0.770	<.001
	Posttest 1	Posttest 2	−2.15 (−4.052 to −0.248)	0.759	.02
B	Pretest	Posttest 1	−2.25 (−4.082 to −0.418)	0.732	.01
	Pretest	Posttest 2	−6.35 (−8.279 to −4.421)	0.770	<.001
	Posttest 1	Posttest 2	−4.10 (−6.002 to −2.198)	0.759	<.001

User Satisfaction

Participants from both groups perceived the game as a positive experience in all aspects: perceived usefulness, ease of use, and enjoyment (Table 7). Although Group B demonstrated slightly higher scores across all categories, there were no statistically significant differences between the groups (*P*>.05).

Table . Comparison of user satisfaction between the 2 groups.

Perceptions	Group A, mean (SD)	Group B, mean (SD)	Mean difference (95% CI)	*P* value
Perceived usefulness	4.51 (0.50)	4.53 (0.40)	0.02 (−0.31 to 0.27)	.91
Perceived ease of use	4.58 (0.51)	4.64 (0.34)	0.07 (−0.34 to 0.21)	.63
Perceived enjoyment	4.63 (0.55)	4.67 (0.55)	0.03 (−0.39 to 0.32)	.85

Learner Experiences When Engaging Within the Gameplay

Semistructured interviews were conducted to gain in-depth information about the gameplay experiences of participants, exploring how their initial roles (host-first or microbe-first) influenced their learning engagement and perceptions of the game. The following key themes emerged from the thematic analysis, highlighting factors that shaped learners' experiences during gameplay.

Theme 1: Learner Profile

Background Knowledge

Participants reported that their background knowledge significantly influenced their engagement and effectiveness during gameplay. Several students described challenges when their prior understanding was limited, particularly when they were required to make decisions. To manage these difficulties, participants strategically chose when to read detailed card descriptions, focusing more deeply on unfamiliar or difficult topics.

> *I read the description on the 'Microbe' and 'Hint' cards in detail only if I had limited knowledge about that card. If I was familiar with that microbe, I just read it briefly and sometimes just the heading.* [Participant 15, Group A, Female]
>
> *When I was the host, it was challenging to identify the microbes, because my microbiology knowledge wasn't that strong. Imagining just one microbe out of many possibilities felt difficult.* [Participant 32, Group B, Male]

Personal Preferences

Personal preferences toward the game can be considered as having a key role in shaping their engagement and emotional responses. Preferences for specific roles were likely to be influenced by individual perceptions of stress, decision-making enjoyment, and familiarity with game-based learning. Those who preferred less cognitive demand tended to favor the microbe role, while others were more motivated by the host role's active decision-making component. Their preference for card games also contributed to positive attitudes, enhancing their overall willingness to engage and learn.

> *I prefer learning in a more relaxed setting, so playing as the microbe suited me, as it didn't require as much intense thinking.* [Participant 12, Group A, Male]
>
> *I enjoy making decisions and being in control, so the host role felt more engaging and aligned with how I like to learn.* [Participant 31, Group B, Female]
>
> *I've always loved card games, so learning microbiology and immunology through one just made sense to me. It didn't feel like studying, as I was having fun, but I could tell I was actually learning something useful at the same time.* [Participant 32, Group B, Male]

XSL•FO
RenderX

Theme 2: Learning Process

Participants demonstrated differing patterns of engagement depending on whether they began the game as the microbe or the host, with each sequence influencing how they processed and applied microbiology and immunology content.

Host Before Microbe

Participants who began as the host described the experience as cognitively demanding but motivating. Without access to structured content at the outset, learners were required to draw on prior knowledge, make inferences, and formulate hypotheses, which align with higher-order thinking. This initial challenge appeared to increase engagement and stimulate curiosity.

> *I liked being the host first—it was challenging because I had no idea which microbes were involved. I had to gather information from both sides and think it through before attacking.* [Participant 32, Group B, Male]

> *Guessing what the microbe was felt exciting. I was genuinely happy when I got it.* [Participant 28, Group B, Female]

In subsequent rounds, these learners reported reflecting on their earlier performance, identifying knowledge gaps, and actively revisiting content to improve.

> *When I became the microbe, I went back and reviewed the Matching Key to see if I had used the right 'Protection' card before. I also wanted to know what other protections were possible because I had made mistakes in the last round.* [Participant 36, Group B, Female]

Microbe Before Host

Students who started as the microbe described this role as a preparatory stage, providing them with a broad overview of key concepts. This role encouraged exposure to content in a lower-pressure context, which later supported more confident and efficient decision-making when switching to the host role. The structure of the game allowed them to gradually transition from recognition-based learning to recall and application.

> *Playing as the microbe first gave me an advantage. It helped me remember the content for each microbe. So, when I switched to the host role, I could recall those features more quickly.* [Participant 15, Group A, Female]

Some participants also reported using short-term memorization techniques during the microbe phase, which helped build a basic conceptual scaffold for identifying pathogens in the host round. These findings suggest that beginning as the microbe may support initial content familiarization, enabling learners to make more informed and confident decisions when transitioning to the more cognitively demanding host role.

> *Since I started as the microbe, I tried to memorize the Matching Key. It helped me get a rough idea of which microbes stood out, so when I switched to host, I could answer based on what I remembered from the last round.* [Participant 5, Group A, Female]

Theme 3: Game Design

The design of the game played an important role in shaping both the cognitive and emotional aspects of learning. Participants reflected on how the game's structure, feedback mechanisms, visuals, and interactive elements supported their engagement and understanding.

Pedagogical Components

Learning Content

Participants reported they could learn about microbes and immunology from the information on the cards, whether by reviewing their own cards or observing those played by their opponents. They also noted that repeatedly experiencing the cards during gameplay deepened their understanding and helped them retain key concepts.

> *The cards had clear explanations. Sometimes I learned more from looking at my opponent's cards.* [Participant 14, Group A, Female]

Level of Difficulty

Participants generally perceived the level of difficulty as appropriate for their academic background. The content was challenging enough to stimulate interest, but it was not too complex to create frustration. This balance helped sustain motivation throughout gameplay. However, some noted that repeated play led to familiarity with answers, reducing difficulty over time.

> *The content was just the right level of difficulty; it wasn't too hard or too easy.* [Participant 8, Group A, Female]

> *The game was fun, but when I started to remember the answers, it became easier, so I'd love to see more variety in the microbes to keep it challenging.* [Participant 36, Group B, Female]

Feedback Mechanism

The review phase where players checked their "Protection" cards against the correct answers was consistently mentioned as a key learning moment. This immediate feedback helped participants reflect on their decisions, correct misunderstandings, and reinforce correct associations between microbes and immune responses.

> *I learned a lot from checking whether the Host's 'Protection' card matched the microbe. Even if it was wrong, I still got to learn from that mistake.* [Participant 13, Group A, Male]

Entertaining Components

Visual Engagement

The visual design of the game, including colorful illustrations and a playful layout, added to its appeal and enhanced the learning atmosphere. Many participants described the cards as cute or attractive, which positively influenced their willingness to engage.

XSL•FO
RenderX

> *The game design was really well done. The pictures were cute and made the whole experience more inviting.* [Participant 21, Group B, Female]

Active Decision-Making Elements

Gameplay was structured to allow active involvement from both players. While the host engaged in decision-making, the microbe observed and predicted responses. This dynamic encouraged deeper analysis and reflection on the game's logic and strategy.

> *Watching my friend play gave me time to see how the game worked overall. I was thinking about the clues they used and whether the microbe I had in mind matched what they played.* [Participant 4, Group A, Female]

Social Interaction

Interactions between players, including discussion, guessing, and bluffing, contributed to enjoyment and emotional investment. Participants appreciated the playful tension between roles, which made learning more immersive and memorable.

> *I liked playing as the microbe because I could watch my friend guess and have fun trying to throw them off with my clues. I didn't enjoy guessing as much, but I loved being the one giving hints.* [Participant 10, Group A, Female]

Theme 4: Learning Setting

The physical and contextual conditions under which gameplay occurred played a notable role in shaping students' comfort, focus, and willingness to engage with the learning material. Participants appreciated the autonomy offered by the flexible setting and timing, as well as the availability of technical support when needed.

Learning Atmosphere

Allowing participants to choose their preferred learning environment fostered a greater sense of comfort and autonomy. Many students reported feeling less anxious and more willing to take intellectual risks when playing outside a formal classroom setting. The absence of an authoritative figure reduced pressure and created space for experimentation and open guessing.

> *I liked that we could choose where to play … even if I wasn't sure about an answer, I felt okay taking a guess without worrying that someone would judge me for getting it wrong.* [Participant 22, Group B, Male]

Flexibility in Learning Schedule

The flexibility to decide when to engage with the game contributed to a more focused and self-directed learning experience. Participants appreciated that they could play when they felt most ready, which made learning feel more like a voluntary activity than a scheduled obligation.

> *The game didn't force us to play at a specific time. I could just pick it up whenever I wanted, and that made it feel like something I chose to do … not something I was made to do.* [Participant 15, Group A, Female]

Real-Time Gameplay support

Access to facilitators or staff members during gameplay reassured participants and minimized concerns about game mechanics. This allowed them to devote more attention to the educational content rather than the rules. The ability to clarify doubts in real-time improved confidence and supported sustained engagement.

> *Having staff there to answer questions made me less worried about the rules, so I could actually focus on the content and not get stuck on how to play.* [Participant 27, Group B, Female]

Theme 5: Learning Outcomes

Gameplay provided opportunities for cognitive engagement across multiple dimensions, from foundational knowledge acquisition to higher-order reasoning (Table 8). Participants described improvements in knowledge, critical and strategic thinking, and conceptual understanding, highlighting the game's educational value beyond passive content delivery.

Table . Frequency of reported learning outcomes achieved from the game.

Learning outcomes	Frequency
Knowledge recall	40
Knowledge acquisition	34
Critical thinking	30
Strategic thinking	26
Knowledge application	22

Knowledge Recall

As students alternated roles and revisited familiar content, their ability to retrieve specific microbial characteristics improved. The act of seeing clues repeatedly triggered memory recognition, reinforcing their confidence and retention.

> *Seeing the clues helped trigger my memory when I looked at the 'Key' card, the name of the microbe felt familiar. And when it was revealed, it confirmed that I remembered it right.* [Participant 22, Group B, Male]

Knowledge Acquisition

The game facilitated the acquisition of microbiology and immunology knowledge through its structured and interactive format. Repeated exposure to microbial features and host responses during gameplay helped build a stronger conceptual

XSL•FO
RenderX

foundation. Many participants reported gaining new information simply by observing their opponent's moves.

> *Reading the details on the cards played by my opponent gave me new information and helped me figure out the microbe.* [Participant 8, Group A, Female]

Critical Thinking

Participants engaged in critical thinking by interpreting incomplete information and making evidence-based guesses. Particularly in the host role, students described the need to synthesize clues and evaluate possible answers, moving beyond memorization to diagnostic reasoning.

> *When I played as the host, I had no idea what the microbe was at first. I had to slowly piece things together using the information I had.* [Participant 17, Group A, Male]

Strategic Thinking

The game structure encouraged participants to develop and adjust strategies based on observations of their opponent. This adaptive thinking extended beyond factual recall to prediction, planning, and real-time decision-making.

> *My opponent rarely played 'Protection' cards. He used a different strategy, like discarding cards to draw a keyword because it gave them a better advantage.* [Participant 18, Group A, Female]

Knowledge Application

Participants not only recalled learning content but also applied their knowledge during gameplay. They demonstrated an ability to connect microbiological information to real-life decision-making, such as selecting appropriate treatments or interventions.

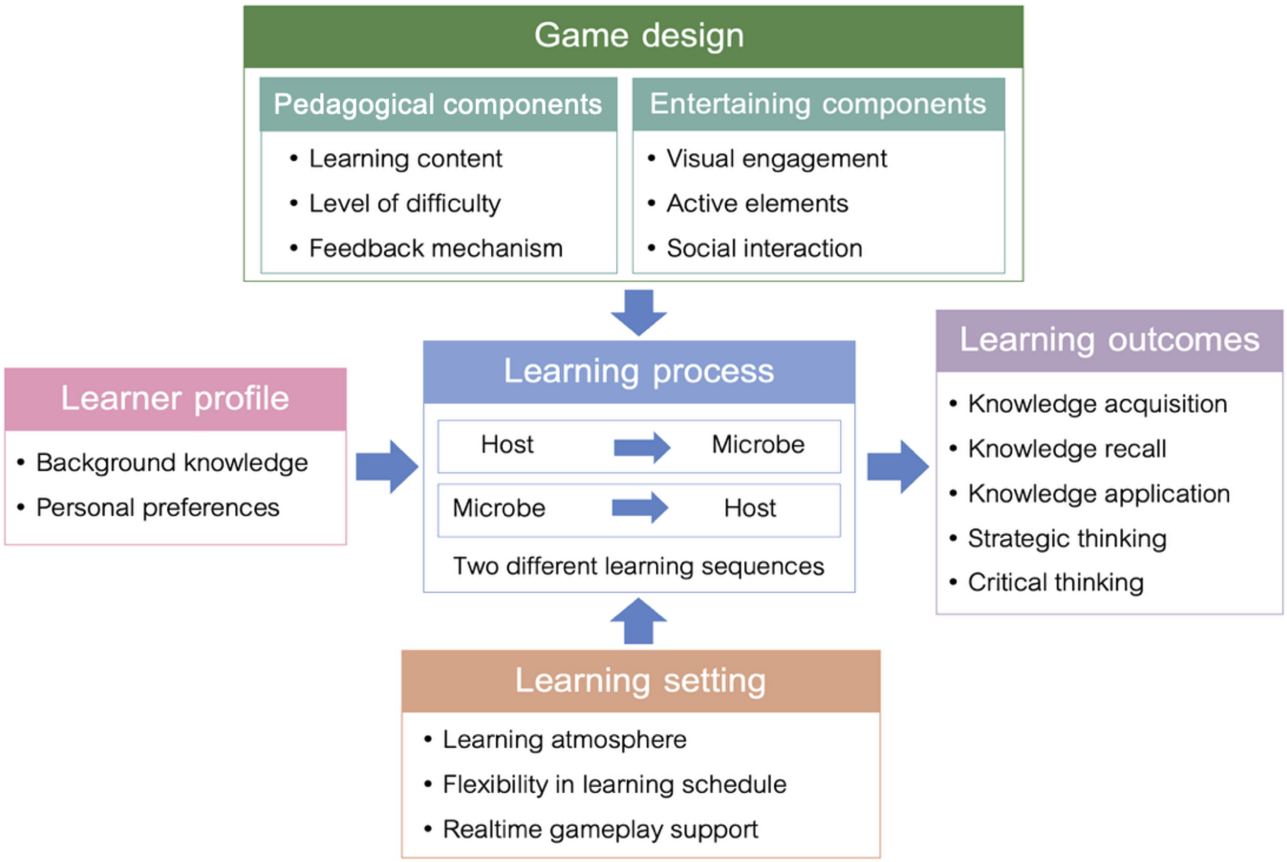
> *I learned which medications go with which types of microbes.* [Participant 9, Group A, Female]

Conceptual Framework for the Implementation of the Game in Dental Education

This conceptual framework explains how the emerging themes show the connections between the core elements of game design, learner characteristics, and the learning setting in supporting students to achieve expected learning outcomes (Figure 4). The game should be designed with a balance between pedagogical and entertaining components, supporting students in understanding key concepts while maintaining their motivation throughout the learning process. The learning process is at the center of the framework, where students move between playing as the host and the microbe. By switching roles, students encounter different kinds of challenges, which encourages them to think in new ways and gradually build their knowledge through the game. The game's rules allow for repeated practice and reflection, helping students remember what they learn and apply it in various situations. The framework highlights the importance of learners' background knowledge and personal preferences, as well as the influence of the learning settings. Factors such as the learning atmosphere, flexible scheduling, and support during gameplay all play a role in shaping the learning experience. These elements work together with the game design to affect how students participate, interact with each other, and respond to challenges in the game. When these factors are well aligned, the game not only helps students understand the subject content but also supports critical thinking and strategic decision-making skills.

XSL•FO
RenderX

Figure 4. Conceptual framework for the implementation of a dual-role duel card game.



Discussion

Effectiveness of Game-Based Learning on Knowledge Enhancement

The "Invasion" game demonstrated its potential as a supplementary learning tool for enhancing knowledge in oral microbiology and immunology. The progressive improvement in scores from Pretest to Posttest 1, and subsequently to Posttest 2, suggests that repeated gameplay contributed to knowledge enhancement. As students moved through the rounds, they were exposed to more diverse sets of microbes, with some organisms appearing more than once. This repetition likely reinforced recall, aligning with established principles of spaced learning and retrieval practice [29]. Participants reported varied learning experiences; some were able to recall prior coursework, others acquired new information, particularly related to antimicrobial treatments, and many developed a broader understanding of how microbiological knowledge applies within clinical contexts. The game was reported by participants to encourage strategic thinking and decision-making skills, which are important for developing clinical reasoning [30-32]. Participants perceived that the dual-role structure of the game, alternating between structured observation and active problem-solving, provided different learning experiences. Students learned not only through success but also through failure perspectives [33,34]. Lost heart points from incorrect guesses can be considered as immediate feedback, prompting them to reflect and adjust their approach. This process reflects the role of failure theory, where knowledge is built through doing, failing, and refining strategies. These findings align with previous research [21,23,28], showing that game-based learning can enhance understanding through engagement, feedback, and active participation.

Impact of Role Sequence on Cognitive Processing

Although knowledge improvement between groups did not differ significantly, qualitative findings suggest that the sequence in which students engaged with gameplay roles experienced the learning process differently. Students who began as the microbe interacted with the material in a more observational manner, generating clues and observing how their partner responded. This role offered an opportunity to reflect on microbial characteristics without the immediate demand for decision-making. This process aligns with inductive learning, where knowledge is built from observed patterns and structured clues. In contrast, those who started as the host were required to analyze clues and make diagnostic judgments from the beginning, enhancing deeper cognitive processing [35,36]. This approach reflects deductive reasoning, in which learners begin with a hypothesis and test it through gameplay. The effectiveness of each sequence may depend on individual learner traits such as cognitive readiness, prior knowledge, or preferred learning approach [37,38]. While some students appeared to benefit from first observing how the content functioned in context, others were more engaged when challenged to act at an early stage. Across both pathways, students reported cognitive improvement through the interactive and engaging environment, highlighting the importance of role variation in supporting diverse learning styles.

XSL•FO
RenderX

Influence of Learner Diversity on Gameplay Experiences

In addition to role assignment, other factors also influenced how students experienced the game and developed cognitive skills through play. Background knowledge and prior experience of learners appear to be important factors in shaping their engagement with the game [28]. Those with a stronger background knowledge tended to participate more confidently throughout the activity [39]. It is possible that students who felt less familiar with the content were more likely to prefer the microbe role, where they could rely more on the structured clues provided by the cards. In addition, individual learning preferences can shape students' approach to gameplay [40-42]; some preferred roles that felt less cognitively demanding, while others were more engaged by strategic decision-making and competitive challenge. The learning context, including peer interaction and minimal technical barriers, can further support student engagement with the game tasks [20,21]. These findings highlight the importance of designing educational games that incorporate role variation and provide flexibility to support learners across a range of backgrounds and preferences.

Pedagogical Implications and Curriculum Integration

The "Invasion" game was developed to align with the pedagogical goals of dental education by accommodating diverse learning preferences and cognitive styles. Its dual-role format allows students to alternate between reflective observation as the microbe and active problem-solving as the host, supporting both inductive and deductive reasoning [43,44]. This structured variation enables learners to engage with microbiology and immunology content through multiple perspectives, promoting deeper understanding and flexible thinking. Within the dental curriculum, the "Invasion" game can be considered a supplemental learning tool alongside didactic lectures to reinforce basic biomedical knowledge, similar to other game-based strategies successfully applied in preclinical dental education [21]. This is consistent with the argument that game-based learning should not be implemented as a mandatory learning method for learners [45]. The adaptable structure of the "Invasion" game allows flexible integration into tutorials, self-directed study, or flipped classrooms, while its emphasis on peer interaction and strategic decision-making fosters motivation, collaboration, and active learning. Beyond dental education, the dual-role format could also be applied in other health education contexts, such as medical, nursing, and pharmacy curricula, where collaborative learning and clinical decision-making are equally essential, although the present content is more directly aligned with oral microbiology and immunology.

Limitations and Recommendations for Future Research

While this mixed methods study provided a comprehensive integration of quantitative and qualitative insights into student learning, certain limitations should be acknowledged. The study was conducted at a single institution with a relatively small sample size, which may limit the generalizability of the findings to broader educational contexts. While the test items were randomized in sequence, the repeated use of the same items across all assessment points may have introduced a memory effect, enhancing posttest performance. Future studies could minimize this issue by using different but equivalent test items for pre- and postassessments, supported by analyses of item difficulty and parallel forms reliability. The short duration of the intervention did not allow for evaluation of long-term retention or application to real-world clinical tasks. Future research should incorporate diverse settings, larger cohorts, varied assessment formats, multiple gameplay sessions over time, and delayed postknowledge assessments to better evaluate the long-term educational impact of serious games.

Conclusions

This study demonstrated that a dual-role educational card game can effectively support diverse learning preferences in oral microbiology and immunology by offering both structured observation and active problem-solving. The game not only enhanced knowledge acquisition but was also perceived by participants to foster strategic thinking and learner engagement through its dynamic role-based format. By alternating between the roles of microbe and host, students engaged with content in different ways, enabling both inductive and deductive learning processes. These findings highlight the value of incorporating flexible, learner-centered tools into the dental curriculum to bridge theoretical science and clinical reasoning. Future research is recommended to investigate the adaptability of this game in broader health education contexts and to evaluate long-term knowledge retention through repeated gameplay sessions over time and follow-up assessments.

Acknowledgments

We would like to express our appreciation to Dr Sophita Suwuttho, Chairman of the Medical Education Technology Center, Faculty of Medicine, Siriraj Hospital, Mahidol University, for her helpful guidance in the development of the game. Furthermore, we are thankful to Dr Khanathat Nuntachurat, Dr Alana Khorana, Dr Prarthana Ahuja, and Dr Apitchaya Saetung for their contribution to the pilot testing of the game concept and to the experts who offered their insightful suggestions during the validation process of the data collection tools. Finally, we acknowledge all research participants for their valuable contributions to this study.

The authors used ChatGPT-5 (OpenAI) to assist with grammar checking and language refinement during manuscript preparation, including minor linguistic edits of interview quotes to improve readability. All artificial intelligence–assisted edits were critically reviewed by the authors to ensure accuracy, preserve the original meanings of quotes, and maintain the integrity of the manuscript.

XSL•FO
RenderX

Funding

This research project was supported by Mahidol University.

Data Availability

The datasets generated or analyzed during this study are not publicly available due to information that could compromise the privacy of research participants but are available from the corresponding author on reasonable request.

Authors' Contributions

Conceptualization: KS, RLS
Data curation: KS, RLS, PC, KJ, JS, CS, TT
Formal analysis: KS, RLS, PC, KJ, JS, CS, TT
Investigation: KS, PC, KJ, JS, CS, TT
Methodology: KS, RLS
Project administration: KS
Resources: KS, RLS
Validation: KS, RLS
Visualization: KS, RLS, PC, KJ, JS, CS, TT, SP
Writing – original draft: KS, RLS
Writing – review & editing: KS, RLS, PC, KJ, JS, CS, TT, SP

Conflicts of Interest

The authors declare no conflicts of interest. Although the "Invasion" game has been designed and developed by the research team, its primary purpose is exclusively for academic purposes, with no commercial objectives associated with it.

References

1. Peterson LR, Hamilton JD, Baron EJ, et al. Role of clinical microbiology laboratories in the management and control of infectious diseases and the delivery of health care. Clin Infect Dis 2001 Feb 15;32(4):605-611. [doi: 10.1086/318725] [Medline: 11181125]
2. Cox JL, Simpson MD, Letts W, Cavanagh HM. Re-thinking microbiology/infection control education to enhance the practice-readiness of health professional students: More than just a curriculum issue. J Learn Des 2015 Mar;8(1):55-67. [doi: 10.5204/jld.v8i1.224]
3. Soni J, Sinha S, Pandey R. Understanding bacterial pathogenicity: a closer look at the journey of harmful microbes. Front Microbiol 2024;15:1370818. [doi: 10.3389/fmicb.2024.1370818] [Medline: 38444801]
4. Banakar M, Bagheri Lankarani K, Jafarpour D, Moayedi S, Banakar MH, MohammadSadeghi A. COVID-19 transmission risk and protective protocols in dentistry: a systematic review. BMC Oral Health 2020 Oct 8;20(1):275. [doi: 10.1186/s12903-020-01270-9] [Medline: 33032593]
5. Rahman S, Montero MTV, Rowe K, Kirton R, Kunik F Jr. Epidemiology, pathogenesis, clinical presentations, diagnosis and treatment of COVID-19: a review of current evidence. Expert Rev Clin Pharmacol 2021 May;14(5):601-621. [doi: 10.1080/17512433.2021.1902303] [Medline: 33705239]
6. Prince M. Does active learning work? A review of the research. J Eng Edu 2004 Jul;93(3):223-231. [doi: 10.1002/j.2168-9830.2004.tb00809.x]
7. Dong JJ, Hwang WY, Shadiev R, Chen GY. Pausing the classroom lecture: the use of clickers to facilitate student engagement. Act Learn High Educ 2017 Jul;18(2):157-172. [doi: 10.1177/1469787417707617]
8. Schnitzler K, Holzberger D, Seidel T. All better than being disengaged: student engagement patterns and their relations to academic self-concept and achievement. Eur J Psychol Educ 2021 Sep;36(3):627-652. [doi: 10.1007/s10212-020-00500-6]
9. Anokye Effah NA, Nkwantabisa AO. The influence of academic engagement on academic performance of university accounting students in Ghana. S Afr J Account Res 2022 May 4;36(2):105-122. [doi: 10.1080/10291954.2021.1988204]
10. Acosta-Gonzaga E. The effects of self-esteem and academic engagement on university students' performance. Behav Sci (Basel) 2023 Apr 21;13(4):348. [doi: 10.3390/bs13040348] [Medline: 37102862]
11. Munna AS, Kalam MA. Impact of active learning strategy on the student engagement. Gnosi: Interdiscip J Human Theol Praxis 2021 Apr;4(2):96-114 [FREE Full text]
12. Venton BJ, Pompano RR. Strategies for enhancing remote student engagement through active learning. Anal Bioanal Chem 2021 Mar;413(6):1507-1512. [doi: 10.1007/s00216-021-03159-0] [Medline: 33458779]
13. Rohani N, Sowa S, Manataki A. Identifying learning preferences and strategies in health data science courses: systematic review. JMIR Med Educ 2024 Aug 12;10:e50667. [doi: 10.2196/50667] [Medline: 39133909]
14. Gentry SV, Gauthier A, L'Estrade Ehrstrom B, et al. Serious gaming and gamification education in health professions: systematic review. J Med Internet Res 2019 Mar 28;21(3):e12994. [doi: 10.2196/12994] [Medline: 30920375]

XSL•FO
RenderX

15. Xu M, Luo Y, Zhang Y, Xia R, Qian H, Zou X. Game-based learning in medical education. Front Public Health 2023;11:1113682. [doi: 10.3389/fpubh.2023.1113682] [Medline: 36935696]
16. Sipiyaruk K, Reynolds PA, Nakornnoi T, Santiwong P, Chintavalakorn R. Game-based learning in orthodontic education: a systematic review. BDJ Open 2024 Jul 4;10(1):56. [doi: 10.1038/s41405-024-00218-3] [Medline: 38965234]
17. Zaror C, Mariño R, Atala-Acevedo C. Current state of serious games in dentistry: a scoping review. Games Health J 2021 Apr;10(2):95-108. [doi: 10.1089/g4h.2020.0042] [Medline: 33818135]
18. Efthimiou G, Tucker NP. Microbes Against Humanity, a workshop game for horrible students: using a creative card game in higher education microbiology teaching. Access Microbiol 2021 Feb;3(2):000186. [doi: 10.1099/acmi.0.000186] [Medline: 34151143]
19. Imwattana K, Methawirune A, Voralakkulchai I, Ngamskulrungroj P. BactoBattle: a game-based learning companion for medical bacteriology. Access Microbiol 2023;5(6):acmi000608.v3. [doi: 10.1099/acmi.0.000608.v3] [Medline: 37424568]
20. Sipiyaruk K, Hatzipanagos S, Vichayanrat T, Reynolds PA, Gallagher JE. Evaluating a dental public-health game across two learning contexts. Educ Sci 2022;12(8):517. [doi: 10.3390/educsci12080517]
21. Arayapisit T, Pojmonpiti D, Dansirisomboon K, Jitverananrangsri K, Poosontipong D, Sipiyaruk K. An educational board game for learning orofacial spaces: an experimental study comparing collaborative and competitive approaches. Anat Sci Educ 2023;16(4):666-676. [doi: 10.1002/ase.2257] [Medline: 36695535]
22. Buajeeb W, Chokpipatkun J, Achalanan N, Kriwattanawong N, Sipiyaruk K. The development of an online serious game for oral diagnosis and treatment planning: evaluation of knowledge acquisition and retention. BMC Med Educ 2023 Nov 3;23(1):830. [doi: 10.1186/s12909-023-04789-x] [Medline: 37924052]
23. Buajeeb W, Reynolds PA, Boontub H, Tangmanpuwadol Y, Sipiyaruk K. Comparison of the effectiveness of a serious game and pre-recorded lecture in diagnosis and treatment planning of oral lesions for dental students. Sci Rep 2024 Dec 27;14(1):30641. [doi: 10.1038/s41598-024-83433-0] [Medline: 39730901]
24. Chintavalakorn R, Chanwanichkulchai C, Buranasing N, Parivisutt N, Patchasri N, Sipiyaruk K. Development and evaluation of a case-based serious game for diagnosis and treatment planning in orthodontic education: quasi-experimental study. JMIR Serious Games 2025 Aug 27;13:e73956. [doi: 10.2196/73956] [Medline: 40864887]
25. Fernandez-Garcia I, Riaza P, Arias A. Game-based learning: a pilot educational strategy to foster the learning of endodontic terminology. Int Endod J 2025 Jul 18;n/a(n/a). [doi: 10.1111/iej.70001] [Medline: 40682132]
26. Wu JH, Su PH, Wu HY, Hsin YM, Lin CH, Lee CY. Educational board game for training dental and dental hygiene students in patient safety issues. BMC Med Educ 2025 Apr 11;25(1):518. [doi: 10.1186/s12909-025-07115-9] [Medline: 40217221]
27. Peanchitlertkajorn S, Reynolds PA, Chaisawas M, Krungthong Y, Boonkaew B, Sipiyaruk K. The impact of an online gamified virtual tour on cognitive enhancement in dental practice management. Sci Rep 2024 Nov 6;14(1):26975. [doi: 10.1038/s41598-024-75128-3] [Medline: 39505891]
28. Teerawongpairoj C, Tantipoj C, Sipiyaruk K. The design and evaluation of gamified online role-play as a telehealth training strategy in dental education: an explanatory sequential mixed-methods study. Sci Rep 2024 Apr 22;14(1):9216. [doi: 10.1038/s41598-024-58425-9] [Medline: 38649451]
29. Wiklund-Hörnqvist C, Jonsson B, Nyberg L. Strengthening concept learning by repeated testing. Scand J Psychol 2014 Feb;55(1):10-16. [doi: 10.1111/sjop.12093] [Medline: 24313425]
30. Chon SH, Timmermann F, Dratsch T, et al. Serious games in surgical medical education: a virtual emergency department as a tool for teaching clinical reasoning to medical students. JMIR Serious Games 2019 Mar 5;7(1):e13028. [doi: 10.2196/13028] [Medline: 30835239]
31. Wong JYH, Ko J, Nam S, et al. Virtual ER, a serious game for interprofessional education to enhance teamwork in medical and nursing undergraduates: development and evaluation study. JMIR Serious Games 2022 Jul 14;10(3):e35269. [doi: 10.2196/35269] [Medline: 35834309]
32. Koelewijn G, Hennus MP, Kort HSM, Frenkel J, van Houwelingen T. Games to support teaching clinical reasoning in health professions education: a scoping review. Med Educ Online 2024 Dec 31;29(1):2316971. [doi: 10.1080/10872981.2024.2316971] [Medline: 38394053]
33. Anderson CG, Dalsen J, Kumar V, Berland M, Steinkuehler C. Failing up: how failure in a game environment promotes learning through discourse. Think Skills Creat 2018 Dec;30:135-144. [doi: 10.1016/j.tsc.2018.03.002]
34. Hefkaluk N, Linehan C, Trace A. Fail, fail again, fail better: how players who enjoy challenging games persist after failure in “Celeste”. Int J Hum Comput Stud 2024 Mar;183:103199. [doi: 10.1016/j.ijhcs.2023.103199]
35. Rossi IV, de Lima JD, Sabatke B, Nunes MAF, Ramirez GE, Ramirez MI. Active learning tools improve the learning outcomes, scientific attitude, and critical thinking in higher education: experiences in an online course during the COVID-19 pandemic. Biochem Mol Biol Educ 2021 Nov;49(6):888-903. [doi: 10.1002/bmb.21574] [Medline: 34652877]
36. Ghezzi J, Higa EFR, Lemes MA, Marin MJS. Strategies of active learning methodologies in nursing education: an integrative literature review. Rev Bras Enferm 2021;74(1):e20200130. [doi: 10.1590/0034-7167-2020-0130] [Medline: 33787786]
37. Gu X, Tan Y, Wu X, et al. Active versus passive strategy in online creativity training: how to best promote creativity of students with different cognitive styles? Think Skills Creat 2022 Jun;44:101021. [doi: 10.1016/j.tsc.2022.101021]
38. Aga K. Comfort in active learning spaces – students’ perceptions and preferences. Eur J Engi Educ 2024 Jul 3;49(4):785-806. [doi: 10.1080/03043797.2024.2341756]

XSL•FO
RenderX

39. Shukla J, Singh RM. Knowledge exploration among students: role of feedback, feeling of confidence, and academic motivation. Metacogn Learn 2024 Dec;19(3):1137-1159. [doi: 10.1007/s11409-024-09404-y]
40. Khenissi MA, Essalmi F, Jemni M, Graf S, Chen NS. Relationship between learning styles and genres of games. Comput Educ 2016 Oct;101:1-14. [doi: 10.1016/j.compedu.2016.05.005]
41. Bontchev BP, Terzieva V, Paunova-Hubenova E. Personalization of serious games for learning. Interact Technol Smart Educ 2021 May 19;18(1):50-68. [doi: 10.1108/ITSE-05-2020-0069]
42. Hwang GJ, Sung HY, Hung CM, Huang I, Tsai CC. Development of a personalized educational computer game based on students' learning styles. Education Tech Research Dev 2012 Aug;60(4):623-638. [doi: 10.1007/s11423-012-9241-x]
43. Prince MJ, Felder RM. Inductive teaching and learning methods: definitions, comparisons, and research bases. J Eng Educ 2006 Apr;95(2):123-138 [FREE Full text] [doi: 10.1002/j.2168-9830.2006.tb00884.x]
44. Pashler H, McDaniel M, Rohrer D, Bjork R. Learning styles: concepts and evidence. Psychol Sci Publ Int 2008 Dec;9(3):105-119. [doi: 10.1111/j.1539-6053.2009.01038.x] [Medline: 26162104]
45. Van Gaalen AEJ, Jaarsma ADC, Georgiadis JR. Medical students' perceptions of play and learning: qualitative study with focus groups and thematic analysis. JMIR Serious Games 2021 Jul 28;9(3):e25637. [doi: 10.2196/25637] [Medline: 34319237]

Abbreviations

IOC: The Index of Item-Objective Congruence

Edited by J Gentges; submitted 03.Jul.2025; peer-reviewed by A Allam, B Bunt, S Mukhida; accepted 09.Oct.2025; published 21.Jan.2026.

Please cite as:
Sipiyaruk K, Srisatjaluk RL, Chaisiri P, Jantawatchai K, Srihatai J, Sonyaem C, Traikorakot T, Pookboonmee S
Fostering Inductive and Deductive Learning in Oral Microbiology and Immunology With a Dual-Role Duel Card Game: Explanatory Sequential Mixed Methods Study
JMIR Med Educ 2026;12:e80048
URL: https://mededu.jmir.org/2026/1/e80048
doi:10.2196/80048

© Kawin Sipiyaruk, Ratchapin Laovanitch Srisatjaluk, Ponpentip Chaisiri, Kanokwan Jantawatchai, Jidapha Srihatai, Chutikarn Sonyaem, Thanaphat Traikorakot, Suphasuta Pookboonmee. Originally published in JMIR Medical Education (https://mededu.jmir.org), 21.Jan.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Education, is properly cited. The complete bibliographic information, a link to the original publication on https://mededu.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Multilevel Determinants of Engagement in Lifelong Learning in Orthodontics From Formal to AI-Supported Approaches: Development of a Conceptual Framework Through a Qualitative Study

Theerasak Nakornnoi, DDS, PhD; Jitpitchaya Wimonchaijit, DDS; Rochaya Chintavalakorn, DDS, PhD; Kawin Sipiyaruk, DDS, PhD

Department of Orthodontics, Faculty of Dentistry, Mahidol University, 6 Yothi Road, Ratchathewi, Bangkok, Thailand

Corresponding Author:
Kawin Sipiyaruk, DDS, PhD
Department of Orthodontics, Faculty of Dentistry, Mahidol University, 6 Yothi Road, Ratchathewi, Bangkok, Thailand

Abstract

Background: Lifelong learning (LLL) is increasingly important for health care professionals, particularly within the field of orthodontics, driven by emerging technologies, updated treatment techniques, and rising patient expectations. To maintain competence in current practice, orthodontists are expected to engage in continuous professional development throughout their careers. While the value of LLL is widely acknowledged, engagement levels among orthodontic professionals can vary due to a number of factors.

Objective: This study aims to investigate how orthodontic professionals perceive and engage with LLL by examining their learning strategies, along with the barriers and enablers. In addition, this study integrates these determinants of professional development to construct a conceptual framework to explain LLL engagement among orthodontic professionals.

Methods: A qualitative study using semistructured interviews with a topic guide was conducted among 16 orthodontic professionals who were recruited using purposive sampling to ensure diversity in settings and experience levels. Interviews were audio recorded and transcribed verbatim until data saturation was achieved. Data were analyzed using framework analysis informed by the theory of planned behavior. To ensure analytic rigor and credibility, an iterative coding process was performed independently by 2 researchers, with any discrepancies resolved through consensus among the research team.

Results: All participants believed that LLL was important for clinical improvement and personal fulfillment. However, they reported several challenges, including time constraints, financial restrictions, and limited access to learning resources. Social influences, institutional contexts, technological advancements, and patient expectations were considered key determinants in shaping LLL engagement. The results also revealed a pedagogical shift in LLL activities from formal short course training to the use of generative artificial intelligence for self-directed learning. A conceptual framework illustrating key determinants of LLL engagement was developed.

Conclusions: LLL engagement was perceived as essential for clinical competence and personal development, which could be influenced by individual, social, institutional, and external factors. The conceptual framework developed in this study demonstrates the key determinants of LLL in orthodontic practice and suggests that multilevel strategies are required to sustain long-term engagement.

(JMIR Med Educ 2026;12:e85634) doi:10.2196/85634

KEYWORDS

continuing education; dental education; lifelong learning; orthodontics; qualitative research

Introduction

Lifelong learning (LLL) appears to be an essential skill for health care professionals. This competency could be considered a required characteristic in the field of orthodontics, as the recent environment of orthodontic practice has become increasingly unpredictable and complex [1]. There is an increasing use of technologies, ranging from advanced imaging techniques to digital treatment planning and customized appliances, resulting in the improvement of treatment outcomes when compared with conventional approaches [2-5]. These technologies are perceived as significant by orthodontic professionals in clinical workflow [6-9]. Furthermore, the demographics of orthodontic patients underscore the importance of adopting patient-centric approaches and interdisciplinary collaboration to address their diverse needs and expectations of patients [10]. In addition,

XSL•FO
RenderX

health care professionals with a strong commitment to LLL can perform better in navigating and managing clinical challenges [11]. Consequently, LLL has become fundamental to maintaining professional standards in modern orthodontics.

While a number of learning methods have been used to update knowledge and skills for continuous improvement in patient care [12], evidence demonstrates that orthodontic professionals may have unique barriers to engaging in LLL activities. Interestingly, orthodontic professionals appear to exhibit lower levels of engagement in continuing education activities compared to other dental specialties [13]. Common challenges appear to be busy schedules of clinical work, which make it difficult to balance professional and personal tasks [14,15]. The high cost of courses and training seems to be another significant barrier, in particular for those with financial constraints [16,17]. Furthermore, unequal access to technology can limit learning opportunities in the digital age [17-19], while the large amount of online information can leave professionals feeling overwhelmed and struggling to identify trustworthy sources [20]. These barriers can negatively impact professional development, highlighting the need to understand LLL behaviors and attitudes within orthodontic professionals to better support their ongoing development.

The LLL engagement in orthodontic education can be considered complicated due to a number of unique factors. Evidence reveals a generation gap, where younger and older orthodontists perceive barriers toward evidence-based practice differently [21]. This finding is consistent with broader trends observed in medical education, where residents frequently get new information by self-directed learning from internet searches, while experienced clinicians tend to draw upon their past experiences [22]. Furthermore, a study confirms that natural cognitive changes associated with aging can influence learning preferences over a long career [23]. These approaches create an educational landscape, highlighting the need to move beyond a one-size-fits-all model and investigate how orthodontic professionals actually perceive and engage with LLL.

Although available evidence has explored LLL in a variety of health care disciplines, little research has been performed on the unique learning behaviors of orthodontic professionals. To enhance the quality of orthodontic care, it is necessary to understand how they engage in LLL activities. Therefore, this study aimed to explore how orthodontic professionals perceive LLL and to investigate its relevant determinants. By identifying these factors, this study seeks to provide insights into supporting dental educators to effectively develop more personalized training approaches, ensuring safe and high-quality patient care within orthodontic practice. To achieve this aim, the following research questions were formulated**:**

1. How do orthodontic professionals currently engage in LLL activities, ranging from formal education to artificial intelligence–supported approaches?
2. How do orthodontic professionals perceive the value of LLL in their practice?
3. What are the key enablers and barriers affecting their engagement in LLL activities?
4. How can these determinants be integrated into a conceptual framework to explain LLL engagement in orthodontics?

Methods

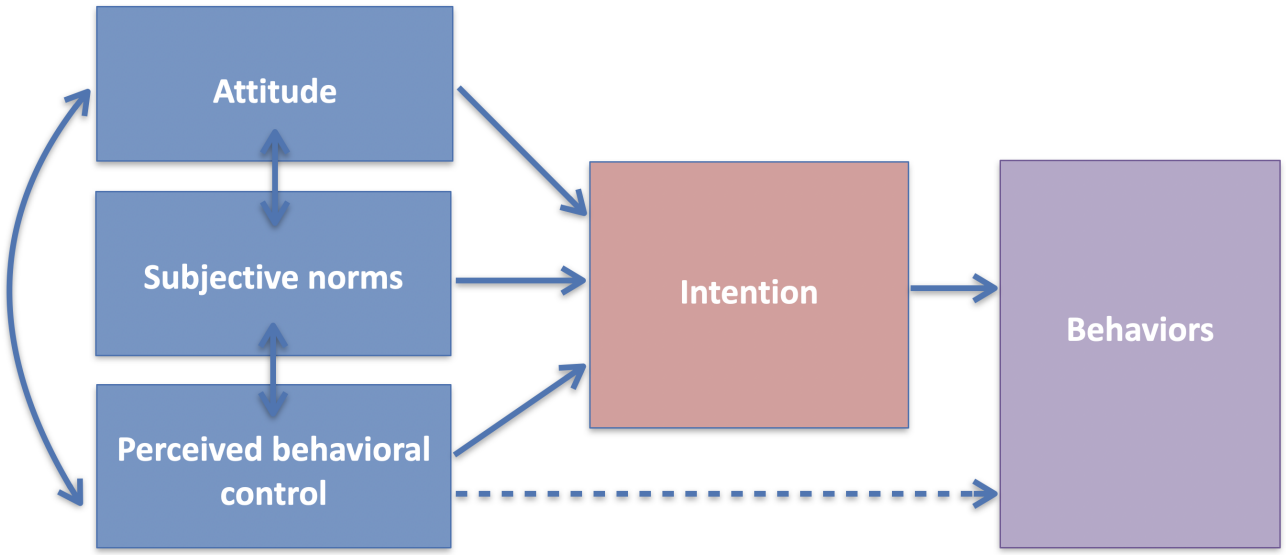
Theoretical Framework

The theory of planned behavior (TPB) was selected as the underpinning theoretical framework for this study (Figure 1), as LLL can be considered a complex behavior rather than a purely cognitive phenomenon. According to TPB, behavioral intentions are shaped by 3 core constructs, which are attitudes, subjective norms, and perceived behavioral control [24]. According to this theory, individuals are more likely to engage in a behavior when they hold positive attitudes, perceive support from influential others, and feel confident in their ability to perform it.

Existing research evidence in health care education has successfully applied TPB to explain engagement in professional activities. For instance, attitude and perceived behavioral control were found to be significant predictors of intentions to adopt technology-enhanced learning in continuing medical education among general practitioners [25]. In addition, there is scientific evidence demonstrating that stronger intentions to engage in targeted professional practices are associated with more positive attitudes, supportive social norms, and greater perceived control [26]. Therefore, TPB was considered a suitable framework to serve as a guide without rigidly constraining the analysis in this study. In other words, while the theory provided an initial structure for the analytical framework, the qualitative nature of this research allowed the researchers to uncover additional context-specific determinants from the data, further enriching the understanding of engagement in LLL activities.

XSL•FO
RenderX

Figure 1. Theory of planned behavior.



Research Design

This study used a qualitative research design using semistructured interviews to gather in-depth information from orthodontic professionals. The qualitative approach was chosen, as it could effectively capture the personal perspectives and insights of orthodontists regarding LLL that would be difficult to quantify through quantitative research [27]. As LLL can be influenced by individual attitudes, professional community and culture, personal challenges, and other contextual factors, qualitative research should provide the most appropriate way for exploration. With multilevel determinants and multifaceted factors, this study adopted a hybrid deductive-inductive approach. This qualitative study was conducted within an interpretivist paradigm and used a framework analysis approach informed by TPB. The deductive component used TPB to structure the inquiry around established psychological constructs, while the inductive component allowed for the emergence of novel themes and context-specific insights directly from the participants' experiences.

Research Participants

The research population included orthodontic professionals who were either enrolled in or had graduated from postgraduate programs in orthodontics at Mahidol University. However, they were excluded if they were no longer active in orthodontic practice. A purposive sampling technique was used to select participants for the interviews to ensure that diverse and information-rich participants were selected to provide their insights [28]. The sampling criteria were considered based on sex (female, male), age (26 - 75 years), years of orthodontic experience (categorized as 1 - 3 years, 3-10 years, 11 - 20 years, and over 20 years), and practice setting (including postgraduate residents, university instructors, private practitioners, and private practice owners). Potential participants were approached via professional networks. No individuals who were approached declined participation, and no interviews were withdrawn or terminated prematurely. The final sample size comprised 16 participants, determined by data saturation [29], where there were no new themes or subthemes that emerged from subsequent interviews.

Data Collection Procedures and Topic Guide Design

Semistructured interviews with a topic guide were performed to collect qualitative data from research participants. The topic guide was developed based on the TPB concept [24]. Accordingly, the topic guide covered 5 main areas, which were: "Professional background," "Perceived importance of LLL (Attitudes)," "Professional community and social expectations (Subjective norms)," "Perceived control over and barriers to learning opportunities (Perceived behavioral control)," "Behavioral intentions," and "Actual learning practices." The interviews were conducted in a private room, where only the participant and the assigned researcher (JW) were present, to ensure privacy and confidentiality. All interviews were conducted in Thai to preserve the authenticity of perceptions and expressions captured from research participants. Each interview lasted approximately 40 to 50 minutes and was audio-recorded after obtaining informed consent from each participant. Field notes and analytic memos were documented during and after interviews. The recordings were then transcribed verbatim using a confidential transcription service. All transcriptions were then anonymized before the analysis process.

Data Analysis

Data analysis was conducted using framework analysis [30], where a preliminary thematic framework was based on the TPB model [24]. This data analytic approach allowed researchers to integrate new themes and subthemes that emerged from the interview data into the original framework. All interviews were conducted in Thai, but they were translated into English using a double-translation method to confirm the accuracy of the translation. Firstly, the interview transcripts were translated from Thai to English by an external translator. Another independent translator then back-translated them into Thai. The original and back-translated versions were then compared by the research team to evaluate whether the English-translated quotes accurately reflected the original meaning. Any

XSL•FO
RenderX

discrepancies in the translations of the transcripts were revised to ensure accuracy before analysis. Therefore, the data analysis was performed on the translated English transcripts.

The verified English transcripts were imported into *Delve*, a qualitative data analysis tool, to facilitate systematic coding and transparent management. To enhance the trustworthiness of the analysis, all procedures were performed and documented with transparency, allowing them to be reviewed and validated. The coding process followed an iterative procedure to ensure the analytic rigor of the study. First, an initial coding framework was independently developed based on the core TPB constructs (attitudes, subjective norms, perceived behavioral control, intention, and behavior) by 2 researchers (JW and KS). After the initial round of coding, the research team reviewed and compared their decisions, during which any identified discrepancies in the development of themes and subthemes were discussed until a consensus was achieved to ensure consistency and credibility in the analysis. The revised coding framework was then indexed to the entire dataset by JW. During this stage, the coding process was regularly reviewed by KS and TN to ensure consistency. Regular discussions were arranged to maintain reflexivity throughout the analysis process, with identified codes being adjusted to ensure conceptual clarity and coherence across all transcripts. Finally, to ensure confirmability and accuracy, the themes and the conceptual framework were reviewed and validated by an external expert in dental education.

Ethical Considerations

This study protocol was reviewed and approved by the Institutional Review Board of the Faculty of Dentistry and the Faculty of Pharmacy, Mahidol University (certificate of approval number MU-DT/PY-IRB 2024/040.1207). All methods were performed in accordance with the relevant guidelines and regulations. Informed consent was obtained from all participants. To protect confidentiality, all data were anonymized, and no personally identifiable information was available during analysis. No compensation was provided to participants.

Results

Research Participants

The research participants included 16 orthodontic professionals, comprising 5 postgraduate residents, 4 university instructors, 5 private practitioners, and 2 dental practice owners. They consisted of 6 male and 10 female participants. The age range of the participants was 29 to 75 years. Their information is presented in Table 1.

Table . Participants included in this study.

Participant	Sex	Age range (years)	Practice setting	Years of orthodontic experience
1	Female	26 - 30	Postgraduate resident	1 - 3
2	Male	26 - 30	Postgraduate resident	1 - 3
3	Female	26 - 30	Postgraduate resident	1 - 3
4	Female	31 - 35	Postgraduate resident	1 - 3
5	Male	31 - 35	Postgraduate resident	1 - 3
6	Female	31 - 35	University instructor	3 - 10
7	Female	31 - 35	University instructor	3 - 10
8	Male	41 - 45	University instructor	11 - 20
9	Female	46 - 50	University instructor	11 - 20
10	Female	31 - 35	Private practitioner	3 - 10
11	Female	36 - 40	Private practitioner	3 - 10
12	Female	61 - 65	Private practitioner	Over 20
13	Male	61 - 65	Private practitioner	Over 20
14	Male	71 - 75	Private practitioner	Over 20
15	Female	36 - 40	Private practice owner	3 - 10
16	Male	41 - 45	Private practice owner	11 - 20

Themes and Subthemes Emerged from the Framework Analysis

Following the framework analysis, 7 themes were identified: (1) attitude: impact of LLL on professional practice, (2) subjective norms: social influence on LLL in orthodontics, (3) perceived behavioral control, (4) intention to engage in LLL, (5) learning behaviors and methods, (6) institutional context: organizational and workplace environment, and (7) external drivers of LLL. Although these themes are interconnected, each represents a distinct dimension of how orthodontists experience and engage in LLL, from personal attitudes and social influences to institutional contexts and external drivers, such as technological changes and patient expectations.

XSL•FO
RenderX

Theme 1: Impact of LLL on Professional Practice (Attitude)

Participants consistently considered LLL to be very important for their clinical practice and personal development. They perceived LLL as a mechanism to enhance patient care, increase treatment efficiency, and ensure their knowledge remains up-to-date. Learning was also linked to personal fulfillment.

Improvement in Clinical Practice

Most participants emphasized that LLL was essential for adapting to a field where knowledge and technology evolve continuously. With current technologies, orthodontists can simplify complex procedures and improve clinical efficiency. As a participant stated, "[It] helps us do our work more easily [and] reduces chair time and patient visits" [Participant 16]. Participants also noted that failing to update knowledge on evolving technologies can hinder the ability to optimize their clinical outcomes. For instance, a participant explained, "We need to adapt and apply knowledge ... like with clear aligners ... now the lab design has improved, ... If we do not update our knowledge, we would not know how to simplify things" [Participant 9].

Personal Fulfillment

Participants considered LLL as a way to fulfill intrinsic motivation, describing it as a mechanism to grow professionally and maintain personal relevance in a dynamic field. Learning was often associated with self-satisfaction, intellectual stimulation, and the desire to become better each day, not just for their patients, but for their own sense of purpose. For instance, a participant expressed a personal passion for LLL, stating, "I enjoy learning ... I feel motivated to keep learning and expanding my knowledge" [Participant 11]. Furthermore, participants believed that LLL provided intellectual stimulation, particularly for those who sought to validate new trends or treatment techniques through research rather than accepting them blindly. As 1 participant explained, "[I want] to see if it really works. ... I enjoy doing research and learning to verify whether claims are true or not" [Participant 7].

Theme 2: Social Influence on LLL in Orthodontics (Subjective Norms)

Participants reported that LLL could be influenced by perceived subjective norms, particularly from peers and mentors, whose behaviors and expectations affect participants' motivation to engage in continuous professional development.

Peer Influence

Participants highlighted that peers played a role in motivating them to explore new knowledge and concepts, and that social comparison often reinforced their intention to remain up-to-date. Attending academic conferences with friends made the experience "more enjoyable" [Participant 1]. In addition, social comparison could reinforce the intention to remain up-to-date, as 1 orthodontic resident stated that seeing peers advance created a competitive motivation: "If my friends are attending something, I want to go and learn with them. If my friends know more, I sometimes want to know too" [Participant 4].

Mentor Influence

Mentors and senior colleagues were perceived as role models in guiding the direction of learning, especially during the early years. Their recommendations could shape learning priorities and introduce new areas of interest or clinical trends. For instance, 1 resident highlighted how a mentor influenced her to engage in LLL activities, noting that her professor "advised me to regularly attend conferences ... [and] to present my works at least once a year" [Participant 1].

Theme 3: Perceived Behavioral Control

Participants identified multiple factors that influenced their ability to engage in LLL, which they described in terms of both challenges that created barriers and enablers that supported greater access and engagement.

Challenges to Participating in LLL Activities

Time Restrictions and Workloads

Time constraints and heavy workloads were frequently cited as barriers by participants, particularly for those in private practice or busy hospitals. Participants described being overwhelmed by clinical hours and administrative duties, which left little opportunity for professional development. For example, 1 participant noted that, while they understood concepts like 3D printing, they lacked the "time to learn the manual process [and] study the back-end details" [Participant 16]. Similarly, others reported reducing their attendance at orthodontic conferences and events because of "clinical duties and family responsibilities" [Participant 8], even though they considered these events very helpful.

Financial Constraints

High registration fees for courses and conferences were perceived as a major barrier. Additional expenses for travel and accommodation, especially for events held abroad or outside participants' towns, further discouraged attendance. Participants often weighed costs against potential benefits, noting that "if I feel a course is too costly for what I will gain, that becomes a barrier" [Participant 7]. This challenge was more emphasized for orthodontists in provincial or rural areas. For them, traveling to major cities meant not only paying for transport and accommodation but also losing income from their routine clinical work. As a participant explained, "That means closing my clinic for a few days, so it is not just the ticket, but it is the income I lose" [Participant 15].

Limited Access to Learning Resources

Limited access to high-quality academic resources was another challenge, particularly for those in private practice without institutional subscriptions. Participants mentioned that accessing leading journals was difficult because "some journals still require payment" [Participant 16].

Resistance to Change

This barrier was also observed, often stemming from a reliance on existing techniques or a lack of confidence in adopting new technologies. Although participants did not share this view themselves, they observed that some colleagues might avoid innovation because "they feel they cannot keep up" [Participant

XSL•FO
RenderX

15]. Furthermore, if practitioners feel they can "manage all cases effectively using their existing methods" [Participant 12], they may perceive no need for further learning.

Enablers to Engage in LLL Activities

Beyond the challenges, participants highlighted several factors that made it easier for them to stay engaged in LLL activities. One of the most frequently mentioned enablers was the growing accessibility of digital tools and online platforms. Resources such as Google, YouTube, and online conferences were described as convenient ways to check ideas, review concepts, or explore unfamiliar topics. In addition, modern tools like ChatGPT were noted for making it "much easier to find information in real time," albeit with a caution that users "still have to be careful" regarding accuracy [Participant 8]. Similarly, the transition to digital platforms has reduced the need to rely solely on physical copies or in-person events, offering a "much greater variety of resources available today" [Participant 9].

Theme 4: Intention to Engage in LLL Activities

This theme reflects a forward-looking commitment to professional growth, which was described in different ways. While participants were keen to explore new technologies like AI, they maintained a critical perspective, noting that currently, "marketing tends to come out before the science" [Participant 8]. Consequently, they expressed a strong intention to verify "reliable data" [Participant 8] and investigate "if it [AI] can actually help us" [Participant 9] before adoption. Furthermore, specific learning goals often drove this engagement. As 1 participant stated, "There are still areas I would like to explore ... improving my digital workflow ... [and] understanding how to use AI in clinical decision-making" [Participant 12].

Theme 5: Learning Behaviors and Methods

This theme demonstrates how participants engage in LLL activities through formal and informal methods. Although all participants had undergone structured orthodontic training, such as residency or postgraduate programs, they used a variety of strategies, ranging from formal short-course training to the use of generative AI for LLL engagement. This reflects a shifting paradigm in orthodontic education, where convenience, interactivity, cost, and trust influence individual choices.

Formal Short-Course Training

Several participants reported that continuing education provided opportunities to update knowledge and skills. Participants demonstrated a strong preference for in-person sessions over online formats, as these activities allowed for more engagement and real-time responses to their questions. These courses reflect a traditional educational model that is still highly valued for hands-on guidance. As a participant noted regarding their training in clear aligners, "It is easier to have somebody teach you rather than learning by yourself" [Participant 9], highlighting a preference for direct instruction over self-study for complex clinical skills.

Scientific Evidence

Scientific evidence from trusted sources, such as academic articles, was considered the most convenient and reliable method reported by most participants. Online access to these articles provided flexibility, allowing them to validate clinical knowledge at any time. One participant explained that this is their initial step: "When I want to know something new, first I read papers from top tier journals ... [and] try to find evidence to support" [Participant 5]. However, some participants noted that this approach was time-consuming. Another participant remarked that while they seek evidence-based support, "working in a full-time private practice, it takes me too long to understand all of this" [Participant 15].

Online Learning

Online learning sessions, recorded lectures, and on-demand webinars were popular among participants because they offered flexibility, cost-effectiveness, and the elimination of travel burdens. These digital formats represented a shift toward learner-controlled approaches, allowing busy clinicians to access content from reliable sources at their convenience. As a participant explained regarding the benefits of online learning materials like webinars, "I can replay it anytime, and when I forget something, these recordings help me recall the information easily" [Participant 12]. However, participants acknowledged that online learning could make them feel less engaged, as this approach lacked the interactivity of in-person education. In addition, while online platforms were convenient for studying theoretical concepts, participants found them insufficient for complex clinical tasks. One participant noted that despite attending many web-based sessions, "online learning alone is not enough ... I still need real cases and experienced professors to guide me" [Participant 3], emphasizing that expert mentorship remains essential for consolidating practical skills.

Conference-Based Learning

Conferences were considered by many participants to support professional learning. They were described as high-value events, particularly when featuring well-known speakers, real case presentations, or hands-on workshops. Compared to online formats, conferences offered networking activities, and participants were able to focus on learning while being away from daily clinical demands. Although expense and time were acknowledged as constraints, attendance was considered worthwhile if the content aligned with their interests and featured authoritative experts. As a participant emphasized, "a good speaker makes you really learn ... [and] if that person has solid foundational knowledge, then it is worth" [Participant 9], noting that the decision to attend often depended on the speaker's reputation regardless of the location.

Peer Learning

Participants noted that peer discussion played an important role in everyday learning. This method was valued because it was practical, trustworthy, and highly contextualized, particularly when advice came from experienced colleagues. For instance, these platforms provide a supportive setting to "share clinical experiences and review real-world cases from senior colleagues" [Participant 8]. Furthermore, the immediacy of digital communication allows for rapid problem-solving; as 1 participant described regarding a messenger group, colleagues can "post a photo with a question, and others can quickly provid responses" [Participant 9]. This highlights how learning has

XSL•FO
RenderX

extended beyond classrooms and conferences, moving from formal toward more informal activities and community-driven exchanges through professional communication groups.

Generative AI

Generative AI tools were increasingly considered emerging supplementary resources, valued for their speed and accessibility compared to formal courses. These tools facilitate on-demand exploration, allowing clinicians to "learn at [their] convenience" [Participant 1] and obtain immediate overviews. However, trust in these tools was tempered by caution regarding accuracy, particularly for clinical details. Participants emphasized that while generative AI is useful for quick inquiries, it is "often not right if you ask for references" [Participant 1]. Therefore, it served as a preliminary tool rather than a definitive source, with participants stressing the need to "double check the information with other sources" [Participant 9] to ensure validity. Consequently, generative AI was integrated not as a replacement for traditional approaches but as a supplementary aid, signaling a potential paradigm shift in orthodontic education toward AI-driven support integrated into everyday learning.

Theme 6: Organizational and Workplace Environment (Institutional Context)

In addition to personal-level barriers and enablers (addressed in Theme 3), this theme explores factors at the organizational level that shaped how orthodontists were engaged in LLL activities, including institutional policies and workplace culture.

Institutional Policies as Barriers

For participants in public hospitals and large organizations, operational requirements often outweighed personal motivation. Significant structural barriers, particularly strict leave policies and staff shortages, frequently prevented attendance at training events, even when participants possessed a strong willingness to learn. One participant described this conflict, citing instances where they could not attend "because there were not enough staff [or they] had already used up [their] leave" [Participant 7], noting that such logistical hurdles were a defining and persistent challenge during their tenure in the hospital system.

Supportive Workplace Culture

Academic institutions and teaching hospitals fostered an environment where LLL was a professional expectation. This culture was reinforced by support such as funding and flexible scheduling. One participant highlighted how this comprehensive support directly influenced their decision to engage in learning activities: "The university covered my registration fee, hotel and flight tickets to attend the meeting, so I thought ... why not go?" [Participant 2]. Similarly, the specific role of an educator created an intrinsic motivation to maintain high educational standards. As another participant explained, "Being part of an academic institution pushes you to stay sharp. When you are expected to teach or supervise, you cannot let your knowledge get stale" [Participant 6].

Theme 7: External Drivers of LLL

This theme examines the external forces that influence LLL engagement, extending beyond personal motivation and institutional context. Participants identified 2 primary external drivers: technological advancements in orthodontics and patient expectations.

Technological Advancements in Orthodontics

Participants considered advances in orthodontic technology as an external influence on their commitment to LLL. The increasing integration of digital innovations, ranging from in-house aligner production and 3D scanning to AI-assisted treatment planning, has transformed routine clinical workflows, necessitating continuous skill acquisition to maintain competency. One participant highlighted clear aligner therapy as a significant market force that compelled them to adapt, stating: "The most recent disruption would be the clear aligner ... The new materials have reached a good level. … I have to learn it" [Participant 8]. This reflection underscores the professional imperative to master emerging technologies once they reach a level of maturity that impacts standard practice.

Patient Expectations

Patient expectations significantly shaped LLL engagement among orthodontic professionals. As patients increasingly accessed dental information online, they often asked about new treatment techniques, prompting orthodontists to recognize the need to maintain professional authority. As 1 participant noted, "Patients are more informed ... If we do not know, how can we explain or treat?" [Participant 12].

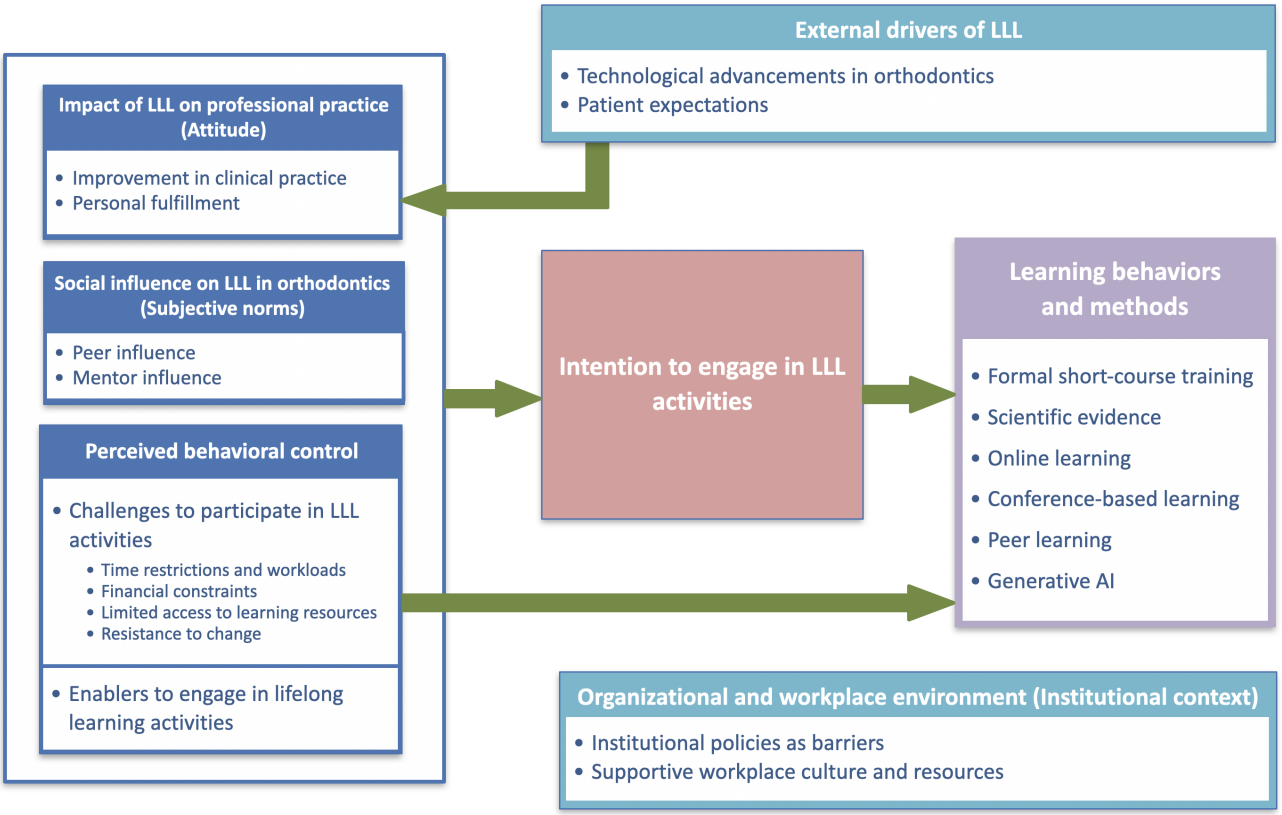
Additionally, patient demand often dictated the adoption of new techniques. One participant admitted that specific inquiries drove their skill acquisition, stating: "Patients requested self-ligating brackets, ... [and] their requests pushed me to learn more. … The same happened with aligners" [Participant 16].

A Conceptual Framework of Key Determinants of LLL Engagement Among Orthodontic Professionals

A conceptual framework was developed through framework analysis of qualitative data retrieved from semistructured interviews (Figure 2), demonstrating the interrelationships among the thematic elements. The framework provides an understanding of how key determinants could shape LLL engagement among orthodontic professionals through TPB, shifting from formal short-course training to more flexible approaches such as the use of generative AI. The framework highlights that the engagement of orthodontic professionals in LLL can be sustained not only by personal motivation, social influence, and perceived behavioral control but also by institutional factors and external pressures. This suggests that fostering LLL engagement among orthodontic professionals requires multilevel strategies, ranging from cultivating positive individual attitudes and creating supportive peer and mentor networks to leveraging external drivers to reinforce the value of continuous learning.

XSL•FO
RenderX

Figure 2. Conceptual framework of key determinants of lifelong learning engagement among orthodontic professionals. AI: artificial intelligence; LLL: lifelong learning.



Discussion

Key Findings and Conceptual Framework

This study aimed to explore how orthodontic professionals perceived and engaged in LLL, identify the strategies they adopt, and examine the determinants influencing such engagement using a TPB-informed framework. The findings revealed that although LLL was widely perceived by orthodontic professionals as essential for clinical efficiency and personal fulfillment, actual engagement (learning behaviors and methods) can be shaped by the interaction between individual attitudes, social influence (mentors and peers), perceived behavioral control, and institutional contexts. These factors were further driven by external pressures, specifically technological disruptions and patient expectations. These determinants were integrated into the conceptual framework, informing how orthodontic professionals consider potential enablers and barriers to make strategic decisions regarding their engagement in LLL.

Professional Impact and Personal Fulfillment

The important theme that emerged was the impact of LLL on professional practice, which aligned with the "Attitude" construct of the TPB model [24]. LLL is a necessary skill for ensuring patient safety and maintaining a high level of orthodontic competence, which is consistent with earlier research demonstrating that it can improve treatment outcomes and strengthen evidence-based practice [10,31]. In addition to the professional aspect, this research highlights that LLL engagement could be driven by personal fulfillment. This reflects intrinsic motivation, where they are satisfied by acquiring new skills or staying updated with rapid technology advancements, as a key driver of adult learning engagement [32]. They perceive LLL as a fundamental part of being a professional rather than just a mandatory requirement [33]. Therefore, both intrinsic and extrinsic motivations are required to engage orthodontists in LLL activities.

Barriers and Enablers to LLL Engagement

When considering "Perceived behavioral control," common barriers to LLL engagement included time restrictions, heavy workloads, financial concerns, and resistance to change, consistent with other studies [13-15,34]. Work-family conflicts and institutional barriers can also be cited as major barriers to continued professional development [35]. At the same time, new learning-aid technologies such as asynchronous learning platforms, webinars, and AI-driven tools have been acknowledged as important facilitators, especially for younger practitioners [36]. With immediate feedback offered by these educational technologies, orthodontic professionals can engage in self-directed learning as a part of their LLL activities [37]. Consequently, despite a number of challenges, the implementation of technologies can facilitate orthodontists to engage in LLL activities.

Mentors and Peers as Social Influence

"Social influence on LLL in orthodontics," corresponding to the "Subjective norms" construct in the TPB model, emerges as another key theme in this research. Mentors and peers can significantly impact LLL engagement, providing support for LLL. Mentorship can offer opportunities for knowledge sharing and skill development to support professional growth [38]. Peers

XSL•FO
RenderX

can also have a positive impact on LLL engagement through informal encouragement, such as inviting colleagues to conferences, sharing new tools and techniques, and discussing orthodontic cases. Collaborative peer learning can positively influence academic growth and offer emotional support, leading to the enhancement of learning engagement [39]. These informal situations can offer learning opportunities, where social influence can encourage orthodontists to engage in LLL activities.

Institutional Contexts and External Drivers

In addition to the constructs available in the TPB model, organizational contexts and external drivers can impact LLL engagement among orthodontic professionals. Academic or hospital-based practitioners could benefit from educational time allowances and formal faculty development programs, where a learning culture can be cultivated [40]. Meanwhile, private practitioners are usually required to finance their learning activities on their own, resulting in difficulties in balancing educational pursuits with business needs [15]. Furthermore, the increasing number of patients who expect advanced technologies can drive practitioners to continue their education [10,41]. These contextual and external influences demonstrate that engagement in LLL can extend beyond the constructs of the TPB model.

Generational Shifts in Learning Preferences

The intention to engage in LLL is reflected in a wide range of learning methods, revealing a significant generational transition. Rather than formal training, younger professionals tend to prefer more flexible learning approaches [42-44]. Based on the interview data, they were likely to engage in LLL activities using online learning communities and generative AI. These results align with other studies, showing an increasing trend of digital literacy in health care education [45,46]. In addition, younger orthodontic professionals tend to use generative AI in supporting their learning and decision-making in complex tasks [47]. These strategies highlight a paradigm shift in LLL engagement, where younger generations of orthodontic professionals tend to prefer more flexible approaches in supporting their professional and career growth.

Implications for Orthodontic Education

Another important consideration is to use the conceptual framework to inform the implications of these findings for enhancing LLL communities among orthodontic professionals. Institutions and professional bodies should consider flexible learning communities or funding for continuing education. Simulations and serious games can also be used as asynchronous interactive learning platforms to support LLL engagement for dental professionals, including orthodontists [48-51]. In addition, postgraduate orthodontics programs should integrate digital and AI literacy, enabling orthodontists to critically appraise AI outputs and identify potential biases [52-54]. These strategies highlight the need for more flexible LLL platforms, ensuring that orthodontic professionals remain competent in their practice.

Study Limitations and Future Recommendations

Although this qualitative study used a robust research design to ensure credibility and validity, several limitations should be acknowledged. Firstly, this study included only orthodontic professionals in Thailand, which means the generalizability of the findings to broader contexts may be limited. As a qualitative study, the findings are context-specific and are not intended to be statistically generalizable. Future quantitative studies could build on these findings by testing the identified themes and relationships in larger samples, thereby enhancing generalizability and applicability. However, purposive sampling was used to recruit orthodontic professionals from diverse practice settings and with varying levels of working experience. In addition, while this cross-sectional qualitative research allowed the researchers to capture the perceptions of orthodontic professionals at a single point in time, further longitudinal research would be helpful to explore how these patterns change throughout their careers, given that LLL activities are dynamic and subject to temporal changes. Finally, the conceptual framework constructed in this research should be further validated through a quantitative study with a robust research design to strengthen its practical application in dental education.

Conclusions

LLL engagement among orthodontic professionals represents a complex dynamic that cannot be solely explained by individual motivation. While widely recognized as essential for professional competence, learning strategies are undergoing a significant shift from traditional formal training toward flexible, technology-driven methods such as online communities and generative AI. However, LLL engagement depends on how well these learning expectations align with the realities of clinical practice. The main challenge is not a lack of interest or willingness to learn, but rather a mismatch between professional demands and the time, resources, and institutional support available for ongoing learning. While TPB serves as the analytical framework for this study, enabling a critical examination of how attitudes, social influences, and perceived behavioral control shape intentions at the individual level, this research further highlights that institutional contexts and external influences play key roles in determining whether learning intentions are translated into action. The conceptual framework developed in this research integrates individual intention within broader professional and organizational contexts. These findings emphasize the need for multilevel strategies and validate the proposed framework as a tool for supporting LLL beyond individual responsibility, with implications for future research and educational policy.

Acknowledgments

The authors express their gratitude to Mahidol University for supporting this research. We also sincerely thank all participants for their involvement. Furthermore, we are thankful to the experts who provided their insightful suggestions during the validation process of this research.

XSL•FO
RenderX

The authors used ChatGPT (OpenAI; GPT-5.1) for grammar checking and language refinement during the manuscript preparation process, including minor linguistic edits to interview quotes to improve readability. All artificial intelligence-assisted edits were critically reviewed by the authors to ensure accuracy, preserve the original meanings of quotes, and maintain the integrity of the manuscript.

Funding

This research was supported by Mahidol University.

Data Availability

The data that support the findings of this study are available from the corresponding author upon reasonable request. The data are not publicly available due to information that could compromise the privacy of research participants.

Authors' Contributions

Conceptualization: TN, JW, RC, KS
Data curation: TN, JW, KS
Formal analysis: TN, JW, RC, KS
Investigation: TN, JW, KS
Methodology: TN, JW, RC, KS
Resources: TN, JW, RC, KS
Validation: TN, JW, RC, KS
Writing-original draft preparation: TN, JW, RC, KS
Writing-review and editing: TN, JW, KS
All authors have read and agreed to the published version of the manuscript.

Conflicts of Interest

None declared.

References

1. Callahan J. Supporting our members and the specialty-building an AAO for all practice modalities. Semin Orthod 2022 Mar;28(1):32-36. [doi: 10.1053/j.sodo.2022.02.009]
2. Alford TJ, Roberts WE, Hartsfield JK, Eckert GJ, Snyder RJ. Clinical outcomes for patients finished with the SureSmileTM method compared with conventional fixed orthodontic therapy. Angle Orthod 2011 May;81(3):383-388. [doi: 10.2319/071810-413.1] [Medline: 21261488]
3. Vigneshkumar V, Deepak ER, Ferreira V, et al. Accuracy and stability of clear aligner treatment: a systematic review. J Pharm Bioallied Sci 2024 Dec;16(Suppl 4):S3077-S3079. [doi: 10.4103/jpbs.jpbs_1048_24] [Medline: 39926927]
4. Ziaei S, Samani D, Behjati M, et al. Accuracy of artificial intelligence in orthodontic extraction treatment planning: a systematic review and meta analysis. BMC Oral Health 2025 Oct 9;25(1):1576. [doi: 10.1186/s12903-025-06880-9] [Medline: 41068732]
5. Dipalma G, Inchingolo AD, Inchingolo AM, et al. Artificial intelligence and its clinical applications in orthodontics: a systematic review. Diagnostics (Basel) 2023 Dec 15;13(24):3677. [doi: 10.3390/diagnostics13243677] [Medline: 38132261]
6. Vaid NR, Adel SM. Contemporary orthodontic workflows: a panacea for efficiency? Semin Orthod 2023 Mar;29(1):1-3. [doi: 10.1053/j.sodo.2023.02.002]
7. Nakornnoi T, Chantakao C, Luangaram N, Janbamrung T, Thitasomakul T, Sipiyaruk K. Perceptions of orthodontic residents toward the implementation of dental technologies in postgraduate curriculum. BMC Oral Health 2023 Sep 1;23(1):625. [doi: 10.1186/s12903-023-03327-x] [Medline: 37658317]
8. Panayi NC, Efstathiou S, Christopoulou I, Kotantoula G, Tsolakis IA. Digital orthodontics: present and future. AJO-DO Clin Compan 2024 Feb;4(1):14-25. [doi: 10.1016/j.xaor.2023.12.001]
9. Peanchitlertkajorn S, Ngamdachakij C, Wongwatjana B, Jongpatranichpunth W, Sipiyaruk K. The implementation of teledentistry training in orthodontic practice: an explanatory sequential mixed-methods study. BMC Med Educ 2024 Dec 30;24(1):1555. [doi: 10.1186/s12909-024-06509-5] [Medline: 39736567]
10. Murdoch AIK, Blum J, Chen J, et al. Determinants of clinical decision making under uncertainty in dentistry: a scoping review. Diagnostics (Basel) 2023 Mar 13;13(6):1076. [doi: 10.3390/diagnostics13061076] [Medline: 36980383]
11. Maini A, Saravanan Y, Singh TA, Fyfe M. Coaching skills for medical education in a VUCA world. Med Teach 2020 Nov;42(11):1308-1309. [doi: 10.1080/0142159X.2020.1788713] [Medline: 32657666]
12. Taylor LE. How to Develop Your Healthcare Career: A Guide to Employability and Professional Development: John Wiley & Sons; 2016.

XSL•FO
RenderX

13. Nazir M, Al-Ansari A, Alabdulaziz M, AlNasrallah Y, Alzain M. Reasons for and barriers to attending continuing education activities and priorities for different dental specialties. Open Access Maced J Med Sci 2018 Sep 25;6(9):1716-1721. [doi: 10.3889/oamjms.2018.373] [Medline: 30337997]
14. Cochrane LJ, Olson CA, Murray S, Dupuis M, Tooman T, Hayes S. Gaps between knowing and doing: understanding and assessing the barriers to optimal health care. J Contin Educ Health Prof 2007;27(2):94-102. [doi: 10.1002/chp.106] [Medline: 17576625]
15. Nayak P, Prasad KVV, Jyothi C, Roopa G, Sanga R. Preferences and barriers for continuing professional development among dental practitioners in the twin cities of Hubli-Dharwad, India. J Indian Assoc Public Health Dent 2015;13(4):429. [doi: 10.4103/2319-5932.171190]
16. Bullock A, Firmstone V, Frame J, Thomas H. Using dentistry as a case study to examine continuing education and its impact on practice. Oxf Rev Educ 2010 Feb;36(1):79-95. [doi: 10.1080/03054981003593548]
17. Conesa J, Alsina MG, Busquets JMB, et al. A vision about lifelong learning and its barriers. Int J Grid Util Comput 2023;14(1):62-71. [doi: 10.1504/IJGUC.2023.129706]
18. Saeed SA, Masters RM. Disparities in health care and the digital divide. Curr Psychiatry Rep 2021 Jul 23;23(9):61. [doi: 10.1007/s11920-021-01274-4] [Medline: 34297202]
19. Gebhardt JS, Harth V, Groneberg DA, Mache S. Job demands and resources perceived by dentists in a digital dental workplace and perceived effects on job satisfaction and stress: a qualitative study. Clin Pract 2025 May 12;15(5):92. [doi: 10.3390/clinpract15050092] [Medline: 40422273]
20. Eliades T. A potentially hazardous shift in the orthodontic continuing education model: a crowd queuing up to educate us. Am J Orthod Dentofacial Orthop 2020 Jul;158(1):1-3. [doi: 10.1016/j.ajodo.2020.03.017] [Medline: 32600746]
21. Madhavji A, Araujo EA, Kim KB, Buschang PH. Attitudes, awareness, and barriers toward evidence-based practice in orthodontics. Am J Orthod Dentofacial Orthop 2011 Sep;140(3):309-316. [doi: 10.1016/j.ajodo.2010.05.023] [Medline: 21889075]
22. Peterson MW, Rowat J, Kreiter C, Mandel J. Medical students' use of information resources: is the digital age dawning? Acad Med 2004 Jan;79(1):89-95. [doi: 10.1097/00001888-200401000-00019] [Medline: 14691004]
23. Miche M, Brothers A, Diehl M, Wahl HW. The role of subjective aging within the changing ecologies of aging: perspectives for research and practice. Annu Rev Gerontol Geriatr 2015 Feb 1;35(1):211-245. [doi: 10.1891/0198-8794.35.211]
24. Ajzen I. The theory of planned behavior. Organ Behav Hum Decis Process 1991 Dec;50(2):179-211. [doi: 10.1016/0749-5978(91)90020-T]
25. Hadadgar A, Changiz T, Masiello I, Dehghani Z, Mirshahzadeh N, Zary N. Applicability of the theory of planned behavior in explaining the general practitioners elearning use in continuing medical education. BMC Med Educ 2016 Aug 22;16(1):215. [doi: 10.1186/s12909-016-0738-6] [Medline: 27549190]
26. Rich A, Medisauskaite A, Potts HWW, Griffin A. A theory-based study of doctors' intentions to engage in professional behaviours. BMC Med Educ 2020 Feb 10;20(1):44. [doi: 10.1186/s12909-020-1961-8] [Medline: 32041599]
27. Creswell JW. Research Design: Qualitative, Quantitative, and Mixed Methods Approaches, 3rd edition: SAGE Publications; 2009. URL: https://www.ucg.ac.me/skladiste/blog_609332/objava_105202/fajlovi/Creswell.pdf [accessed 2026-04-18]
28. Ahmad M, Wilkins S. Purposive sampling in qualitative research: a framework for the entire journey. Qual Quant 2025 Apr;59(2):1461-1479. [doi: 10.1007/s11135-024-02022-5]
29. Hennink M, Kaiser BN. Sample sizes for saturation in qualitative research: a systematic review of empirical tests. Soc Sci Med 2022 Jan;292:114523. [doi: 10.1016/j.socscimed.2021.114523] [Medline: 34785096]
30. Ritchie J, Spencer L, O'Connor W. Carrying out qualitative analysis. In: Qualitative Research Practice: A Guide for Social Science Students and Researchers: SAGE Publications; 2003:219-262 URL: https://books.google.co.in/books?id=IZ3fJlD5x8gC&redir_esc=y [accessed 2026-04-18]
31. Allareddy V, Rengasamy Venugopalan S, Nalliah RP, Caplin JL, Lee MK, Allareddy V. Orthodontics in the era of big data analytics. Orthod Craniofac Res 2019 May;22:8-13. [doi: 10.1111/ocr.12279] [Medline: 31074158]
32. Montasem A, Brown SL, Harris R. Subjective well-being in dentists: the role of intrinsic aspirations. Community Dent Oral Epidemiol 2014 Jun;42(3):279-288. [doi: 10.1111/cdoe.12084] [Medline: 24949514]
33. Papakostopoulou M, Kučera J, Tycová H. Understanding orthodontists' decision making in relation to innovations from encounter to implementation: a qualitative study. Am J Orthod Dentofacial Orthop 2022 Apr;161(4):e316-e335. [doi: 10.1016/j.ajodo.2021.10.012] [Medline: 35086738]
34. Koh SSL, Manias E, Hutchinson AM, Donath S, Johnston L. Nurses' perceived barriers to the implementation of a Fall Prevention Clinical Practice Guideline in Singapore hospitals. BMC Health Serv Res 2008 May 18;8:105. [doi: 10.1186/1472-6963-8-105] [Medline: 18485235]
35. Mlambo M, Silén C, McGrath C. Lifelong learning and nurses' continuing professional development, a metasynthesis of the literature. BMC Nurs 2021 Apr 14;20(1):62. [doi: 10.1186/s12912-021-00579-2] [Medline: 33853599]
36. Khanagar SB, Al-Ehaideb A, Maganur PC, et al. Developments, application, and performance of artificial intelligence in dentistry - a systematic review. J Dent Sci 2021 Jan;16(1):508-522. [doi: 10.1016/j.jds.2020.06.019] [Medline: 33384840]

XSL•FO
RenderX

37. Jacox LA, Bocklage C, Edwards T, Mihas P, Lin FC, Ko CC. Understanding technology adoption by orthodontists: a quantitative study. Am J Orthod Dentofacial Orthop 2022 Mar;161(3):364-374. [doi: 10.1016/j.ajodo.2020.08.024] [Medline: 34690040]
38. Daher-Armache G, Armache J, Ismail HN. Leadership development in U.S. higher education: strategies for lifelong learning and upskilling. J Innov Knowl 2025 Jul;10(4):100754. [doi: 10.1016/j.jik.2025.100754]
39. Li H. Impact of collaborative learning on student engagement in college English programs: mediating effect of peer support and moderating role of group size. Front Psychol 2025;16:1525192. [doi: 10.3389/fpsyg.2025.1525192]
40. Dapaah-Afriyie K, Rice LB. Developing a learning culture in hospital medicine. J Brown Hosp Med 2025;4(2):15-18. [doi: 10.56305/001c.132274] [Medline: 40191703]
41. Tan ELY, Song YL, Zhang Z, et al. Orthodontic treatment in National Dental Centre of Singapore: trends toward higher proportion of adult patients. APOS Trends Orthod 2019;9:89-93. [doi: 10.25259/APOS-13-2019]
42. Chang TY, Hong G, Paganelli C, et al. Innovation of dental education during COVID-19 pandemic. J Dent Sci 2021 Jan;16(1):15-20. [doi: 10.1016/j.jds.2020.07.011] [Medline: 32839668]
43. Motevasel H, Helms LR, Eckert GJ, Stewart KT, Albright DA. The impact of the COVID-19 pandemic on U.S. orthodontic practices in 2020. Am J Orthod Dentofacial Orthop 2022 Feb;161(2):198-207. [doi: 10.1016/j.ajodo.2020.11.040] [Medline: 34688518]
44. Mohammad-Rahimi H, Nadimi M, Ghalyanchi-Langeroudi A, Taheri M, Ghafouri-Fard S. Application of machine learning in diagnosis of COVID-19 through X-ray and CT images: a scoping review. Front Cardiovasc Med 2021;8:638011. [doi: 10.3389/fcvm.2021.638011] [Medline: 33842563]
45. Schlenz MA, Schmidt A, Wöstmann B, Krämer N, Schulz-Weidner N. Students' and lecturers' perspective on the implementation of online learning in dental education due to SARS-CoV-2 (COVID-19): a cross-sectional study. BMC Med Educ 2020 Oct 9;20(1):354. [doi: 10.1186/s12909-020-02266-3] [Medline: 33036592]
46. Hakami Z. Comparison between virtual and traditional learning methods for orthodontic knowledge and skills in dental students: a quasi-experimental study. Healthcare (Basel) 2021 Aug 24;9(9):1092. [doi: 10.3390/healthcare9091092] [Medline: 34574866]
47. Bichu YM, Hansa I, Bichu AY, Premjani P, Flores-Mir C, Vaid NR. Applications of artificial intelligence and machine learning in orthodontics: a scoping review. Prog Orthod 2021 Jul 5;22(1):18. [doi: 10.1186/s40510-021-00361-9] [Medline: 34219198]
48. Peanchitlertkajorn S, Reynolds PA, Chaisawas M, Krungthong Y, Boonkaew B, Sipiyaruk K. The impact of an online gamified virtual tour on cognitive enhancement in dental practice management. Sci Rep 2024 Nov 6;14(1):26975. [doi: 10.1038/s41598-024-75128-3] [Medline: 39505891]
49. Sipiyaruk K, Reynolds PA, Nakornnoi T, Santiwong P, Chintavalakorn R. Game-based learning in orthodontic education: a systematic review. BDJ Open 2024 Jul 4;10(1):56. [doi: 10.1038/s41405-024-00218-3] [Medline: 38965234]
50. Chintavalakorn R, Chanwanichkulchai C, Buranasing N, Parivisutt N, Patchasri N, Sipiyaruk K. Development and evaluation of a case-based serious game for diagnosis and treatment planning in orthodontic education: quasi-experimental study. JMIR Serious Games 2025 Aug 27;13:e73956. [doi: 10.2196/73956] [Medline: 40864887]
51. Marya A, Selvaraj S, Okazaki K, Wang DH, Kanetaka H, Porntaveetus T. The role of extended reality in orthodontic treatment planning and simulation-a scoping review. Int Dent J 2025 Dec;75(6):103855. [doi: 10.1016/j.identj.2025.103855] [Medline: 40885124]
52. Kim CS, Samaniego CS, Sousa Melo SL, Brachvogel WA, Baskaran K, Rulli D. Artificial intelligence (A.I.) in dental curricula: ethics and responsible integration. J Dent Educ 2023 Nov;87(11):1570-1573. [doi: 10.1002/jdd.13337] [Medline: 37489621]
53. Krishnan V. Integrating artificial intelligence: a step forward in orthodontic education. J World Fed Orthod 2024 Aug;13(4):153-154. [doi: 10.1016/j.ejwf.2024.07.002] [Medline: 39095128]
54. Hanenkrath J, Park JH, Bay C. Training, use, and modifications related to artificial intelligence in postgraduate orthodontic programs in North America. Am J Orthod Dentofacial Orthop 2025 Jan;167(1):89-94. [doi: 10.1016/j.ajodo.2024.09.008] [Medline: 39520421]

Abbreviations

AI: artificial intelligence
LLL: lifelong learning
TPB: theory of planned behavior

XSL•FO
RenderX

Edited by A Stone, T Leung; submitted 10.Oct.2025; peer-reviewed by D Karanth, MV Wyk; revised version received 28.Feb.2026; accepted 20.Mar.2026; published 14.May.2026.

Please cite as:
Nakornnoi T, Wimonchaijit J, Chintavalakorn R, Sipiyaruk K
Multilevel Determinants of Engagement in Lifelong Learning in Orthodontics From Formal to AI-Supported Approaches: Development of a Conceptual Framework Through a Qualitative Study
JMIR Med Educ 2026;12:e85634
URL: https://mededu.jmir.org/2026/1/e85634
doi:10.2196/85634

© Theerasak Nakornnoi, Jitpitchaya Wimonchaijit, Rochaya Chintavalakorn, Kawin Sipiyaruk. Originally published in JMIR Medical Education (https://mededu.jmir.org), 14.May.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Education, is properly cited. The complete bibliographic information, a link to the original publication on https://mededu.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Psychomotor Performance in Blended Versus In-Person Emergency Medical Technician Training: Retrospective Cohort Study

Cheng-Heng Liu[1,2,3], MD; Ming-Ju Hsieh[2], MD, PhD; Yu-Ting Chen[4,5], PhD; Chih-Wei Yang[1,2,5], MD, PhD

[1]Department of Medical Education, National Taiwan University Hospital, Taipei City, Taiwan

[2]Department of Emergency Medicine, National Taiwan University Hospital, Taipei City, Taiwan

[3]Institute of Health Policy and Management, National Taiwan University College of Public Health, Taipei City, Taiwan

[4]Department of Living Sciences, National Open University, New Taipei City, Taiwan

[5]Graduate Institute of Medical Education and Bioethics, National Taiwan University College of Medicine, Taipei City, Taiwan

Corresponding Author:
Chih-Wei Yang, MD, PhD
Department of Medical Education
National Taiwan University Hospital
No. 7, Zhongshan South Road, Zhongzheng District
Taipei City, 100
Taiwan
Phone: 886 223123456 ext 261426
Email: cwyang0413@gmail.com

Abstract

Background: Blended learning is common in health professions education, but aggregate comparisons may obscure differences among knowledge, procedural skill, and integrated performance. Whether these domains differ after program-level changes under a common assessment system remains uncertain.

Objective: We aimed to compare knowledge, psychomotor performance, and integrated scenario competence between blended and traditional in-person emergency medical technician level 1 certification training, and, as an exploratory secondary aim, to describe digital engagement and technology acceptance among blended-cohort participants who completed the survey.

Methods: We compared routinely collected certification records from traditional in-person (2020; n=603) and blended (2023; n=581) emergency medical technician level 1 cohorts. The primary outcome was psychomotor performance; secondary outcomes were knowledge, integrated scenario performance, and competency-threshold attainment. We estimated between-cohort differences with 95% CIs and adjusted linear regression. Equivalence was assessed using two 1-sided tests with expert-derived educational importance margins established through an instructor consensus exercise before the panel was provided with the between-cohort results (±5 points for knowledge, ±3 points for psychomotor performance, and ±4 points for integrated scenario performance). These margins represent educational consensus thresholds rather than externally validated clinical thresholds. Blended-cohort trainees were additionally offered a questionnaire on digital engagement and technology acceptance, analyzed by multiple linear regression on intention to use e-learning again.

Results: Psychomotor performance was lower in the blended cohort (80.60, SD 9.70, vs 84.44, SD 10.80; mean difference 3.84, in-person minus blended, 95% CI 2.67-5.01; d=0.37; P<.001); the adjusted coefficient was −4.04 (95% CI −5.15 to −2.92; P<.001). Equivalence within the prespecified ±3-point margin was not demonstrated. Knowledge (difference −0.30, 95% CI −1.68 to 1.08) and integrated scenario performance (−0.94, 95% CI −1.89 to 0.00) fell within the prespecified ±5- and ±4-point margins, respectively. Psychomotor-threshold (521/581, 89.7%, vs 557/603, 92.4%; P=.10) and overall-certification rates (459/581, 79%, vs 474/603, 78.6%; P=.87) did not differ detectably. Knowledge-threshold attainment was higher in the blended cohort (516/581, 88.8%, vs 503/603, 83.4%; P=.007). Among the 236 blended-cohort participants who completed the survey, perceived benefits and positive attitudes toward digital learning were positively associated with intention to use e-learning again.

Conclusions: The blended cohort had lower early psychomotor scores, whereas knowledge and integrated scenario differences fell within the prespecified margins. Programs adopting comparable models should assess psychomotor outcomes directly rather than infer them from knowledge or overall certification outcomes.

XSL•FO
RenderX

(JMIR Med Educ 2026;12:e94950) doi:10.2196/94950

KEYWORDS

blended learning; emergency medical services; health professions education; psychomotor skills; retrospective cohort study; technology acceptance

Introduction

Background

Blended learning is now a standing design choice in health professions education rather than a contingency, and the relevant question has shifted from whether programs can deliver content online to how different competencies perform after a program-level transition [1]. Pandemic-era transitions that prompted widespread adoption [2,3] retain methodological interest because they produced externally mandated changes applied to whole programs, a configuration difficult to obtain deliberately. Syntheses in health professions education report that blended learning yields knowledge outcomes at least comparable, and in some settings superior, to traditional instruction [1,4]. That evidence base, however, is dominated by aggregate comparisons of overall effectiveness. Whether a bundled program-level transition is associated differently with knowledge, procedural skill, and integrated performance under a common assessment system remains uncertain.

Health professions education is inherently multidimensional, requiring the integration of knowledge, psychomotor skills, and complex situational judgment. Traditional in-person health professions education has historically emphasized extensive hands-on practice, immediate instructor feedback, and scenario-based learning to support these competencies. Blended learning redistributes activities across digital and in-person environments [5,6]. In the program studied here, theoretical instruction moved online and scheduled practical hours were consolidated into fewer onsite course days. Evidence on psychomotor performance under such designs is inconsistent and often limited by small sample sizes, heterogeneous assessments, or reliance on self-reported competence [7]. Large-scale comparative data on standardized procedural and complex integrated performance outcomes remain sparse. A comprehensive evidence map of health professions digital education identified modality comparison, skills performance assessment, and learner-context interactions as areas with substantial remaining research gaps [8].

Another gap in the health professions education literature concerns learner digital readiness and technology acceptance, which may shape engagement with blended learning environments. The technology acceptance model and related frameworks hold that perceived usefulness and ease of use shape attitudes and intentions toward digital learning tools [9,10]. A systematic review of these applications in medical education found growing use alongside a need for context-specific validation across learner populations [11]. Digital literacy and self-efficacy are associated with engagement and performance in online learning [12,13], and inequalities in digital readiness have been documented in online education settings [14]. Emergency medical technician (EMT) trainees vary widely in age, educational background, and technological familiarity, and how these characteristics relate to engagement with blended learning has rarely been described in this population. The ancillary survey reported here addresses that descriptive gap among blended-cohort participants who completed the survey. It does not measure the certification outcomes and cannot account for any difference between the cohorts.

Theoretical Framework

Blended learning can be understood as a form of instructional restructuring that alters when and how learning activities occur across competency domains. Three bodies of literature are summarized below, and they do not serve the same purpose. Cognitive load theory [15] and mastery learning and deliberate practice [16,17] provide a rationale for reporting the competency domains separately. Technology acceptance [9,10] does not bear on the comparison between domains and is used only for the exploratory secondary aim concerning digital engagement among blended-cohort participants who completed the survey.

Cognitive load theory describes why early procedural learning is demanding: declarative knowledge and motor execution must be coordinated while both are still unfamiliar [15]. It provides a rationale for reporting the 3 competency domains separately rather than as a composite, but it does not specify the direction of any particular change in sequencing.

Distributed access to theoretical content may alter opportunities for self-paced review and retrieval practice, which can strengthen retention [18]. Whether these opportunities translate into measurable differences in simulated clinical scenarios has not been established.

Mastery learning and deliberate practice frameworks bear on how instructional restructuring may influence procedural competence [16,17]. These models hold that expertise develops through repeated, feedback-rich practice until predefined standards are met. Where a program alters the timing or density of practice, initial performance and later retention may be affected, although the direction is not obvious a priori. A systematic review reported advantages for distributed rather than massed surgical-skills training in several outcomes [19]. A small quasi-experimental study compared 3-month retention after combined spacing and testing with retention after single-session training without testing [20]; because both spacing and testing differed, it does not isolate the contribution of either component. In the present setting, theoretical instruction was separated from practical sessions and the scheduled practical hours were distributed across fewer course days; these bundled changes cannot be distinguished by this study. Whether the consolidated onsite schedule changed how procedural practice was distributed cannot be determined from the available records.

Together, these perspectives justify domain-specific reporting but do not predict a specific ordering across assessment scales.

XSL•FO
RenderX

The present study does not test the instructional mechanisms they propose.

Learner engagement with blended environments is also shaped by technology acceptance. Perceived usefulness and perceived ease of use are constructs of the technology acceptance model [10,21], while the facilitating-conditions construct comes from the unified theory of acceptance and use of technology [9]; digital literacy, self-efficacy, and experience further shape engagement behavior and perceived instructional value [12,13,22,23]. These constructs supplied the questionnaire items and organized the descriptive results reported for the exploratory secondary aim. They do not explain the certification outcomes because no acceptance measure was used as a predictor of assessment performance and the questionnaire was not administered to the traditional cohort.

The COVID-19–related transition from fully in-person to blended emergency medical technician level 1 (EMT-1) training at our institution created an opportunity for a retrospective historical cohort comparison of certification outcomes under a single national certification system. We compared knowledge, psychomotor performance, and integrated scenario competence between the traditional and blended cohorts, asking whether the difference between cohorts was of similar size across the 3 domains. As an exploratory secondary aim, we described digital engagement and technology acceptance among blended-cohort participants who completed the survey and examined which participant-reported factors were associated with their intention to use e-learning again; this second aim is confined to those participants and is not used to explain the between-cohort differences addressed by the first. The comparison is at the level of outcomes. It does not measure the learning processes that might explain any difference observed.

Methods

Study Design

We conducted a retrospective historical cohort comparison following an institutional change in instructional delivery related to the COVID-19 pandemic. The transition restructured EMT-1 training from a fully in-person format to a blended model. The curriculum framework, competency standards, and assessment procedures in both periods were governed by the same national certification system. The delivery format was fixed by program period and applied to every trainee enrolled in that period; there was no within-period choice between formats. This does not eliminate selection into course timing, attendance-related selection, or secular differences between the historically separated cohorts. This study analyzed routinely collected administrative training records from a regional EMT-1 training center; no prospective participant enrollment was conducted.

Two cohort periods were included: a pretransition cohort trained between January and December 2020 (traditional format) and a posttransition cohort trained between January and December 2023 (blended format). Cohorts trained in 2021 and 2022 were excluded because public health restrictions varied within and between course offerings, enrollment was reduced, and delivery format was not consistently defined. The 2023 cohort represents the earliest stable implementation of the blended curriculum, so the estimate may differ from one obtained after instructors had accumulated more experience with the format. This historical cohort design enabled comparison under one certification framework while retaining the limitations of a nonrandomized, historically separated cohort comparison.

Setting and Educational Context

This study was conducted at a regional emergency medical services (EMS) education center accredited by Taiwan's Ministry of Health and Welfare to deliver EMT-1 certification training. Taiwan's national EMS training system follows a 3-tier structure (EMT-1, emergency medical technician level 2 [EMT-2], and EMT-Paramedic), with EMT-1 representing foundational training in basic life support, scene safety, patient assessment, and initial emergency interventions. The EMT-1 curriculum comprises 40 hours of standardized instruction, including didactic teaching, hands-on psychomotor skills training, and scenario-based simulation. Accredited training centers are required to implement a uniform national curriculum, adhere to predefined competency requirements, and apply standardized assessment tools. These requirements provide a common framework but do not by themselves establish comparable realized instruction across cohorts.

Participants

This study's population comprised trainee records available in the retained EMT-1 certification database for the designated study periods. Study eligibility criteria were age ≥20 years, completion of the required EMT-1 training components, attendance at ≥90% of scheduled instructional sessions, no prior EMT certification at any level, and recorded completion of the certification assessments. The retained administrative dataset contained 1251 trainee records after program-completion and attendance requirements had already been applied. Original course-enrollment rosters were unavailable, so the numbers initially enrolled and excluded before assembly of the retained dataset could not be determined for either cohort. Sixty-six records were subsequently excluded for age <20 years, leaving 1185 age-eligible records. Analyses then used a complete-case set: a trainee was included only if a score was recorded for all 3 certification assessments. One traditional-cohort record lacked a recorded psychomotor score and was excluded, leaving 1184 for analysis, comprising 603 in the traditional cohort and 581 in the blended cohort. The attendance criterion identifies trainees who completed within their own cohort. Under program regulations in force in both periods, trainees who missed scheduled sessions with approved leave could be deferred to a subsequent cohort, where they completed missed instruction and assessments within a 6-month window. In the blended program, attendance at live online sessions was verified by requiring cameras to remain on with the participant's face identifiable; a designated assistant recorded attendance and captured periodic screenshots, and the live sessions were recorded. Trainees who missed live online hours were directed to the program's make-up procedure before the onsite practical days. The available administrative records do not show how often this procedure was used and do not separate exclusions attributable to attendance, incomplete assessment, or prior

XSL•FO
RenderX

certification, so cohort-specific exclusion counts and selection effects cannot be assessed.

Instructional Conditions

Traditional In-Person Program (Pretransition Cohort)

The traditional cohort completed a 40-hour course delivered across 5 days and entirely in person. The schedule comprised 18 hours of didactic instruction and 22 hours of practical activities, including psychomotor skills practice and integrated scenarios. The supplied period-specific records did not independently document a 16-hour/6-hour subdivision of the practical component.

All sessions were conducted onsite. Instructor qualifications, competency standards, and assessment processes were governed by national EMT-1 requirements.

Blended Learning Program (Posttransition Cohort)

The blended cohort completed the same 40 scheduled instructional hours, reorganized so that the didactic component was delivered online and the practical component onsite. Eighteen hours of theoretical instruction were delivered online, comprising 4 hours of prerecorded material released before the course and 14 hours of live synchronous teaching via a videoconferencing platform with interactive elements. Session recordings were available for asynchronous review. The remaining 22 scheduled hours comprised face-to-face practical activities across 3 onsite course days, including psychomotor skills practice and scenario-based simulation; the supplied period-specific records did not independently document a 16-hour/6-hour subdivision. The recognized schedule in both periods totaled 45 hours, of which 5 were allocated to written and practical assessment; the 40 instructional hours reported throughout this paper exclude assessment time. The online component preceded the onsite days, but published course notices for 2023 show intervals ranging from 1 day to approximately 3 weeks across offerings. The records do not identify which offering each trainee attended.

Documented Features of the Practical Training Component

The practical-training features are summarized in Table 1. The accredited provider had previously reported that, as a standing practice in both formats, learners practiced in pairs, classes were divided between concurrent skills stations, and instructors were assigned under a ceiling rule of 1 instructor per 15 learners. Because of rounding, the nominal ratio could range from approximately 1:10 to 1:15. The authors additionally confirmed from direct program knowledge that the onsite practical-training model, teaching materials, and instructor pool did not change between periods. These are program-level confirmations rather than period-specific session records. The records reviewed did not document realized group sizes, equipment used, cohort-specific scenario allocation, repetition counts, feedback or debriefing procedures, assessment-station configuration, or session-level instructor calibration.

Taken together, the provider's standing-practice account and the authors' direct program knowledge indicate that the onsite practical-training model, teaching materials, and instructor pool did not change between periods. The records reviewed did not permit a session-level audit of realized group size, equipment use, scenario allocation, repetition counts, feedback, debriefing, or hands-on exposure. The program transition nevertheless bundled online theory with consolidation of practical sessions into 3 onsite days rather than 5 course days. Observed differences therefore cannot be attributed to a single component.

The knowledge assessment (TR1), psychomotor skills assessment (TR2), and integrated scenario assessment (TR3) were administered at the conclusion of the training program, after all instructional components had been completed.

XSL•FO
RenderX

Table 1. Documented course and staffing information for the traditional and blended programs.

Parameter	Traditional in-person program (2020)	Blended program (2023)
Recognized schedule	45 hours (40 instructional, 5 assessment)	45 hours (40 instructional, 5 assessment)
Didactic instruction	18 hours, in person	18 hours, online (4 prerecorded, 14 live synchronous)
Scheduled practical instruction	22 hours	22 hours
Skills/scenario allocation	Not documented in supplied period-specific records	Not documented in supplied period-specific records
Delivery pattern	Five course days, with didactic and practical content partially interleaved within the same day	Online component preceding 3 onsite practical days; interval between components varied across offerings
Practice grouping	Learners in pairs; class divided between concurrent stations	Learners in pairs; class divided between concurrent stations
Instructor staffing	Same instructor pool; ceiling rule of 1 instructor per 15 learners	Same instructor pool; ceiling rule of 1 instructor per 15 learners
Facility, manikins, equipment	Same study center; period-specific room, manikin, and equipment records were not available for this audit	Same study center; period-specific room, manikin, and equipment records were not available for this audit
Scenario cases, repetition, and feedback	Teaching materials unchanged; cohort-specific scenario allocation, repetition counts, and feedback/debriefing records were not available for this audit	Teaching materials unchanged; cohort-specific scenario allocation, repetition counts, and feedback/debriefing records were not available for this audit
Assessment	Knowledge assessment (TR1[a]), psychomotor skills assessment (TR2[b]), and integrated scenario assessment (TR3[c]), under national protocols	Knowledge assessment (TR1), psychomotor skills assessment (TR2), and integrated scenario assessment (TR3), under national protocols

[a]TR1: knowledge assessment.

[b]TR2: psychomotor assessment.

[c]TR3: integrated scenario assessment.

Outcome Measures

Primary Outcome

The primary outcome was psychomotor performance, assessed via TR2: a standardized skills assessment station evaluating out-of-hospital cardiac arrest management, including cardiopulmonary resuscitation quality metrics, airway management, and defibrillation procedures. The passing threshold was ≥70%.

Secondary Outcomes

Exploratory between-cohort comparisons were conducted for 2 additional certification assessments:

- TR1 (Knowledge): a multiple-choice examination scored out of 100 points, covering anatomy, physiology, pharmacology, and clinical decision-making (passing threshold: ≥60%).
- TR3 (Integrated Scenario Performance): a high-fidelity simulation requiring integration of cognitive and psychomotor competencies in randomly assigned scenarios selected from a standardized case bank (eg, acute myocardial infarction, pulmonary embolism, intracerebral hemorrhage, and prehospital out-of-hospital cardiac arrest triage and transport decisions; passing threshold: ≥70%).

All assessments were administered under the same national certification framework and the standardized operational procedures in effect during each study period. Historical interrater agreement estimates for the operational scoring rubrics were available from a separate association instructor calibration exercise conducted on September 28, 2020. During that exercise, evaluators independently scored the same recorded trainee performances, yielding κ=0.85 for TR2 and κ=0.82 for TR3. These estimates were not derived from duplicate ratings of the assessments analyzed in this study, which were routinely scored by a single evaluator. Because the original calibration forms and complete information on the numbers of raters and performances were unavailable, the estimates could not be recomputed, and their applicability to the assessments administered during the study periods could not be fully verified.

Ancillary Survey Instrument and Analysis

Digital learning experiences were assessed using a theory-informed, self-administered questionnaire offered to the blended cohort. The instrument measured digital engagement across 4 dimensions (skills, emotional, participation, and performance engagement) using 12 items adapted from Fredricks et al [24]. Perceived challenges (8 items), perceived benefits (5 items), and technology acceptance (4 items, based on technology acceptance model constructs) were also assessed, together with demographic characteristics and prior digital learning experience. Items adapted from English-language instruments were translated into Chinese and reviewed by the author team. The instrument showed high internal consistency across subscales (Cronbach α=.928-.941) and content validity established by expert review (content validity index=0.92). These survey outcomes are exploratory and are reported as such throughout. The questionnaire analysis included the 236 blended-cohort participants who completed the survey.

Sample Size Calculation

Under equal group sizes, a 2-sided independent-samples *t* test with an α of .05 requires 176 participants per group to provide

XSL•FO
RenderX

80% power to detect a standardized mean difference (*d*) of 0.30. Both analyzed cohorts exceeded this requirement. Observed power is not reported because it is a deterministic function of the *P* values already given and adds nothing to the CIs. Nonsignificant results were not interpreted as evidence of equivalence [25]; equivalence testing against the prespecified margins is described in the Statistical Analysis subsection.

Statistical Analysis

Statistical analyses were performed using SPSS (version 28.0; IBM Corp). Shapiro-Wilk tests rejected exact normality in all 6 cohort-by-outcome distributions (all $P<.001$). Given the large group sizes, between-cohort mean comparisons, education-stratified TR2 comparisons, and equivalence tests used pooled-variance (equal-variances-assumed) independent-samples t procedures and were interpreted using effect estimates and CIs. Effect sizes are reported as absolute Cohen d values; mean differences are reported with 95% CIs. Knowledge variances differed between cohorts (SD 13.30 vs 10.75); Welch *t* test for that comparison gave the same conclusion (difference −0.30, 95% CI −1.67 to 1.08; $P=.67$). The primary analyses used the same complete-case sample for all 3 outcomes (n=603 in the traditional cohort and n=581 in the blended cohort). An additional sensitivity analysis of TR3 included all available scenario scores (n=604 and n=581, respectively).

A nonsignificant difference is not evidence of equivalence [25]. Each outcome was therefore also assessed by two 1-sided tests against its prespecified educational importance margin, with equivalence declared only when both one-sided tests reached $P<.05$. Educational importance thresholds of 5 points for TR1, 3 points for TR2, and 4 points for TR3 were established through an in-person consensus discussion among 12 EMS instructors at the association's instructor consensus camp on September 12, 2025. These thresholds were applied as symmetric equivalence margins of ±5, ±3, and ±4 points, respectively. The margins were established before the between-cohort analyses were conducted and before any study results were disclosed to the panel. Participants were asked to identify score differences that they considered educationally important for the knowledge, psychomotor, and integrated scenario assessments, and the discussion continued until consensus was reached. Based on the approximate scoring structure of the assessments, the margins were broadly comparable to the scores assigned to 2 knowledge-test items for TR1, 1 psychomotor checklist item for TR2, and either 1 major item or 2 minor items for TR3. These margins represent expert-derived educational thresholds rather than externally validated clinical thresholds (Multimedia Appendix 1). Competency-threshold attainment, defined as the proportion of trainees meeting the passing standard for each assessment and the proportion certified overall, was compared using Pearson chi-square tests without continuity correction. To assess whether the between-cohort imbalance in education level accounted for the psychomotor difference, that outcome was additionally examined within each education stratum. Sex differences were examined with 2-way ANOVA, fitted separately for each outcome with instructional modality, sex, and their interaction as factors and Type III sums of squares. The ANOVA interaction term tested whether sex differences varied by modality. Pearson correlation coefficients were calculated for age with each outcome within each cohort and for education level with each outcome, with education coded ordinally as high school or below, college or university, and graduate or higher. To assess the robustness of between-group differences, multivariable linear regression models were fitted for the primary outcome (TR2) and secondary outcomes (TR1, TR3) with instructional modality as the independent variable, adjusting for age, sex, and education level, using heteroscedasticity-consistent SEs (HC3 estimator). Education entered these models as a single ordinal term with the same 3 ordered categories. Regression coefficients (*b*) are unstandardized and are expressed in assessment score points. The survey data were analyzed separately from the certification outcomes, with the participant as the unit of analysis. Among the 236 blended-cohort participants who completed the survey, intention to use e-learning for future training was regressed on perceived benefits, positive attitudes toward digital learning, emotional engagement, performance engagement, and age in a single ordinary least squares model, with multicollinearity assessed using variance inflation factors. This model is exploratory and does not describe the full blended cohort.

Statistical significance was set at $P<.05$ for the primary outcome (TR2). TR1 and TR3 were treated as secondary exploratory comparisons; *P* values are reported descriptively without multiplicity correction.

Ethical Considerations

This study's protocol was reviewed and approved by the Research Ethics Committee B, National Taiwan University Hospital (NTUH-REC No. 202601212RINB), in accordance with the Declaration of Helsinki. Because this study involved retrospective analysis of deidentified educational records, the committee approved a waiver of informed consent. Data were deidentified at the source institution and stored in encrypted, password-protected databases in compliance with applicable data protection regulations.

Results

Participant Characteristics

The complete-case analysis included 1184 participants: 603 from the traditional cohort (January-December 2020) and 581 from the blended cohort (January-December 2023). All 3 assessments were administered as mandatory components of the certification program. Table 2 presents the baseline participant characteristics.

No statistically significant differences were detected in sex distribution ($P=.19$) or age ($P=.54$). Education level differed ($P=.005$), with a higher proportion of graduate-level trainees in the blended cohort (97/581, 16.7%, vs 68/603, 11.3%).

XSL•FO
RenderX

Table 2. Baseline participant characteristics.

Characteristic	In-person (n=603)	Blended (n=581)	*P* value
Male, n (%)	325 (53.9)	291 (50.1)	.19
Age (years), mean (SD)	31.2 (8.7)	31.5 (9.0)	.54
Education level, n (%)			.005
High school or below	141 (23.4)	104 (17.9)	
College or university	394 (65.3)	380 (65.4)	
Graduate or higher	68 (11.3)	97 (16.7)	

Primary Outcome: Psychomotor Performance

Table 3 summarizes performance on the 3 assessments together with the equivalence tests against the prespecified margins.

Mean differences are calculated as in-person minus blended, and Cohen d values are absolute. Equivalence testing was conducted against the prespecified educational importance margins. "Met" indicates that both 1-sided tests reached a *P* value less than .05 under the stated numerical bound; it does not establish clinical equivalence. Because the 3 raw-score margins corresponded to different standardized widths (approximately 0.41, 0.29, and 0.48 SD for TR1, TR2, and TR3, respectively), the pattern of "met" and "not met" partly reflects margin width relative to score variability and should not be interpreted as a ranking of domain sensitivity.

Psychomotor performance was lower in the blended cohort (80.60, SD 9.70, vs 84.44, SD 10.80; mean difference 3.84, 95% CI 2.67-5.01; *P*<.001), with a small-to-medium effect size (*d*=0.37). The psychomotor outcome did not meet the prespecified equivalence criterion (upper 1-sided *P*=.92).

In a multivariable model adjusting for age, sex, and education level, the blended-cohort indicator was associated with lower psychomotor scores (*b*=−4.04, 95% CI −5.15 to −2.92; *P*<.001). The adjusted coefficient was slightly larger in magnitude than the unadjusted difference; the nonrandomized design does not permit a causal interpretation.

Because education level differed between cohorts, with a higher proportion of graduate-level trainees in the blended cohort, psychomotor performance was examined within each education stratum. The difference favoring the traditional cohort was present in all 3 strata. Values were 2.78 points (95% CI −0.32 to 5.89; *P*=.08) among trainees with high school education or below (n=141 vs 104), 4.15 points (95% CI 2.76-5.54; *P*<.001) among those with college or university education (n=394 vs 380), and 4.40 points (95% CI 1.55-7.25; *P*=.003) among those with graduate education (n=68 vs 97). The estimate was imprecise in the smallest stratum.

Table 3. Competency assessment outcomes by training modality.

Outcome	In-person, mean (SD)	Blended, mean (SD)	Difference (95% CI)	|*d*|	*P* value	Equivalence within margin
TR1[a], knowledge	69.23 (13.30)	69.53 (10.75)	−0.30 (−1.68 to 1.08)	0.02	.67	Met (±5 points)
TR2[b], psychomotor	84.44 (10.80)	80.60 (9.70)	3.84 (2.67 to 5.01)	0.37	<.001	Not met (±3 points)
TR3[c], integrated scenario	85.59 (8.64)	86.53 (7.85)	−0.94 (−1.89 to 0.00)	0.11	.05	Met (±4 points)

[a]TR1: knowledge assessment.

[b]TR2: psychomotor assessment.

[c]TR3: integrated scenario assessment.

Secondary Outcomes: Knowledge and Integrated Scenario Performance

Knowledge scores differed by −0.30 points between cohorts (95% CI −1.68 to 1.08; *d*=0.02; *P*=.67), and the difference fell within the prespecified ±5-point equivalence margin (both 1-sided *P*<.001). Integrated scenario scores differed by −0.94 points (95% CI −1.89 to 0.00; *d*=0.11; *P*=.05), and the difference fell within the prespecified ±4-point margin (both 1-sided *P*<.001). Using every record with a scenario score (n=604 vs 581), the difference was −1.08 points (95% CI −2.07 to −0.10; *d*=0.13; *P*=.03); the result still met the ±4-point margin. After covariate adjustment, neither knowledge (*b*=−0.21; *P*=.76) nor integrated scenario performance (*b*=0.62; *P*=.17) was associated with instructional modality. We draw no directional inference from the integrated scenario result.

Competency-Threshold Attainment

Because certification depends on meeting fixed passing standards rather than on mean scores, threshold attainment was compared separately (Table 4). On the psychomotor assessment, 521 of 581 (89.7%) blended-cohort trainees and 557 of 603 (92.4%) traditional-cohort trainees met the 70% standard, a difference of −2.7 percentage points that did not reach significance (*P*=.10). Overall certification rates were 459 of 581 (79%) and 474 of 603 (78.6%), respectively (*P*=.87). Knowledge-threshold attainment was higher in the blended cohort (516/581, 88.8%, vs 503/603, 83.4%; *P*=.007), whereas integrated-scenario-threshold attainment did not differ detectably

XSL•FO
RenderX

(559/581, 96.2%, vs 576/603, 95.5%; *P*=.55). The lower mean psychomotor score therefore did not correspond to a statistically significant between-cohort difference in meeting the national competency standard. The near-identical knowledge means, smaller SD, and 5.4-percentage-point higher pass rate in the blended cohort are consistent with a narrower score distribution in which fewer trainees fell below the 60-point threshold.

Overall certification is the certification outcome recorded in the training records.

Table 4. Competency-threshold attainment by training modality.

Assessment	Traditional in-person, n/N (%)	Blended, n/N (%)	Difference (blended minus in-person), percentage points[a]	*P* value
Knowledge (TR1[b]), ≥60	503/603 (83.4)	516/581 (88.8)	+5.4	.007
Psychomotor (TR2[c]), ≥70	557/603 (92.4)	521/581 (89.7)	−2.7	.10
Integrated scenario (TR3[d]), ≥70	576/603 (95.5)	559/581 (96.2)	+0.7	.55
Overall certification	474/603 (78.6)	459/581 (79.0)	+0.4	.87

[a]Positive values indicate a higher proportion in the blended cohort.

[b]TR1: knowledge assessment.

[c]TR2: psychomotor assessment.

[d]TR3: integrated scenario assessment.

Demographic Associations With Performance

Age correlations were negative for all 3 outcomes in both cohorts. For knowledge, the correlations were r=−0.121 (*P*=.004) in the blended cohort and r=−0.055 (*P*=.17) in the traditional cohort. For psychomotor performance, they were r=−0.261 (*P*<.001) and r=−0.271 (*P*<.001), respectively; for integrated scenario performance, they were r=−0.158 (*P*<.001) and r=−0.150 (*P*<.001), respectively.

Female participants had higher mean scores in every cohort-by-outcome cell. In 2-way ANOVA models with sex and instructional modality as factors, the main effect of sex was significant for all 3 outcomes (TR1 $F_{1,1180}$=48.0, TR2 $F_{1,1180}$=15.7, and TR3 $F_{1,1180}$=46.3; all *P*<.001). No sex-by-modality interaction was detected for TR1 ($F_{1,1180}$=0.04; *P*=.84), TR2 ($F_{1,1180}$=0.99; *P*=.32), or TR3 ($F_{1,1180}$=0.18; *P*=.67).

Education level, coded ordinally as high school or below, college or university, and graduate or higher, was positively correlated with knowledge (r=0.147; *P*<.001) and with integrated scenario performance (r=0.130; *P*<.001). The correlation with psychomotor performance was near zero (r=0.021; *P*=.47). The education-stratified psychomotor difference persisted within every education stratum, including among trainees with graduate education, who were more numerous in the blended cohort. Knowledge-threshold attainment, which was higher in the blended cohort, was not examined within education strata.

Secondary Outcomes: Digital Engagement and Technology Acceptance

Of the 236 blended-cohort participants who completed the survey (response rate: 236/581, 40.6%), approximately three-quarters (181/236, 76.7%) reported prior e-learning experience. No statistically significant differences were observed in digital engagement, perceived benefits, perceived challenges, or technology acceptance between participants with and without prior e-learning experience (all *P*>.05; Multimedia Appendix 2). In an exploratory multiple linear regression, perceived benefits (*b*=0.34; *P*<.001) and positive attitudes toward digital learning (*b*=0.28; *P*<.001) were positively associated with intention to use e-learning again. Emotional engagement and performance engagement were also associated with this intention, whereas age showed a modest negative association (*b*=−0.02; *P*=.047). The model explained 47% of the variance (R^2=0.47; Multimedia Appendix 3). These findings describe survey participants only and are not generalized to the full blended cohort.

Discussion

Principal Findings

In this retrospective comparison of eligible completers from 2 certification cohorts assessed under 1 national system, a bundled program-level transition was associated with different outcome estimates across 3 competency domains. Psychomotor scores were 3.84 points lower in the blended cohort. Knowledge and integrated scenario differences were −0.30 and −0.94 points, respectively. The prespecified equivalence criterion was met for knowledge and integrated scenario performance and not met for psychomotor performance. These bounds were not validated clinical thresholds. Psychomotor-threshold, integrated-scenario-threshold, and overall-certification rates did not differ detectably; knowledge-threshold attainment was 5.4 percentage points higher in the blended cohort (*P*=.007). Among the 236 blended-cohort participants who completed the survey (236/581, 40.6%), perceived benefits and positive attitudes toward digital learning were positively associated with intention to use e-learning again. The survey findings are exploratory and describe survey participants only.

The score differences were 3.84 points for psychomotor performance, −0.30 for knowledge, and −0.94 for integrated scenario performance, each on its own assessment scale. Comparisons across domains are descriptive because the assessments are not on a common validated scale and no formal cross-domain contrast was performed. This study does not explain the observed estimates. Cognitive load, feedback quality,

XSL•FO
RenderX

rehearsal continuity, and procedural integration were not measured, so the accounts discussed in the Comparison With Prior Evidence subsection are candidate explanations rather than findings.

These results also show why the components of a blended program should be described separately. In the equivalence tests, knowledge and integrated scenario scores met the prespecified margins, whereas psychomotor performance did not. Because delivery format, sequencing, calendar consolidation, and cohort year changed together, this study cannot identify which feature accounts for the observed pattern.

Comparison With Prior Evidence

The knowledge-score difference was −0.30 points (95% CI −1.68 to 1.08), consistent with prior syntheses of online or blended vs face-to-face knowledge outcomes [1,4,26,27]. Reviews focused on clinical-skills outcomes address a related but distinct evidence base [7,28]. A broader online-learning meta-analysis also emphasized differences among instructional comparisons [29]. Knowledge and integrated scenario scores met the equivalence criteria within their prespecified margins.

Psychomotor performance differed between cohorts. Cognitive load theory offers one possible account [15], although this study did not measure cognitive load or schema formation. The distributed-practice literature addresses the more specific question of timing: a systematic review reported advantages for distributed rather than massed surgical-skills training in several outcomes [19], and a small quasi-experimental study compared 3-month retention after combined spacing and testing with retention after single-session training without testing [20]. Because spacing and testing changed together in the latter study, it does not identify the contribution of either component. Neither source tests the bundled separation of theory from practice and consolidation of practical hours in this program, and the present data do not identify a mechanism.

Two features of the blended program changed together and cannot be separated in this design: the interval between the online theory component and the first onsite practical session varied from 1 day to approximately 3 weeks across course offerings, and the practical hours were consolidated into 3 onsite days rather than distributed across 5 course days. Prior work identifies practice scheduling as a design consideration in procedural motor-skills training [30], but neither feature was manipulated independently here. An untested possibility is that a longer delay between the online theory component and the first onsite practical session would be associated with lower psychomotor scores. If so, trainees whose onsite days began the next day would be expected to score closer to the traditional cohort than trainees who waited approximately 3 weeks. These records do not permit that comparison, because no field links a trainee to a particular course offering, so the documented variation in interval cannot be joined to individual scores. Enrollment records identifying each trainee's offering, or a prospective design in which the interval is assigned, would test the prediction. The observed difference therefore cannot be attributed to the theory-to-practice interval, the consolidated schedule, or any proposed learning process.

The wider literature on technology-supported resuscitation training is heterogeneous. Lin et al [31] reported little overall difference in chest-compression quality; among learners using feedback devices, only the appropriate compression-rate measure favored face-to-face training. Ko et al [32] found that the 3-practice blended arm had more correct-speed compressions and shorter hands-off time, and both blended arms performed better on environmental safety; knowledge, attitude, and confidence did not differ across groups. Boonmak et al [33] compared online with didactic theory after both groups received the same 2-hour skills training and reported comparable knowledge and skills outcomes. Iskander et al [34] established noninferiority for change in self-assessed confidence, not objective competence, among the 21% of participants with complete pre- and postcourse questionnaires. Sarboozi-Hosseinabadi et al [35] reported no statistically significant between-group differences at the 3 assessed time points and described webinar and simulation as equally effective. The remaining studies also differed in design and outcome: Bieri et al [36] assessed self-reported confidence and perceived knowledge without standardized psychomotor testing; Lim et al [37] found poorer overall skills with self-directed technology-based training without instructor guidance, while certainty across the reviewed evidence was very low; Chong et al [38] found no statistically significant adjusted differences in manikin scores, although the remote group practiced on more days and made more assessment attempts; and Chien et al [39] reported no significant difference between 6-month blended and traditional refresher groups, whereas skills declined more after a 12-month interval. Unlike their refresher single-skill context, our study examined initial multiskill certification training, so their comparison has limited direct applicability to learners acquiring several skills for the first time. These studies provide relevant comparisons, but they do not establish formal equivalence across delivery models or explain the between-cohort pattern observed here.

Integrated scenario scores fell within the prespecified ±4-point equivalence margin. Distributed access and flipped-classroom designs may support review, retrieval practice, and presession study [18,40], but this study did not test whether those processes affected scenario performance.

Age was modestly negatively correlated with psychomotor and integrated scenario performance in both cohorts. Although age has been examined in the technology-adoption literature [41], these performance correlations do not test a technology-adoption mechanism. The cohort-specific coefficients were numerically close (psychomotor: r=−0.261 in the blended cohort and r=−0.271 in the traditional cohort; integrated scenario: r=−0.158 and r=−0.150, respectively), but no formal age-by-cohort contrast was performed. These descriptive results therefore do not show whether delivery format modified the age association.

Alternative Explanations

The 3-year interval between cohorts is the principal threat to internal validity. Familiarity with online instruction, expectations of remote work, and other unmeasured cohort characteristics may have differed between 2020 and 2023 and could have influenced outcomes independently of the program transition.

XSL•FO
RenderX

Both cohorts completed the same 3 assessment domains under the same national certification protocols and passing standards. Period-specific item sets, checklists, and assessment scenarios were not available for this retrospective audit and therefore could not be verified as identical. This does not eliminate secular changes in learners or implementation, including how the same instructor pool delivered the course over time. The selective outcome pattern does not resolve this threat because temporal confounding could affect competency domains differently. We therefore cannot determine how much of the psychomotor difference reflects the bundled program transition or secular change.

A second explanation concerns delivery of the face-to-face component by the same instructor pool under a different calendar schedule. The blended program scheduled its 22 hours of practical activities across 3 onsite days, whereas the traditional program distributed the same scheduled practical hours across 5 course days. This study did not measure realized pacing, fatigue, cross-day revisiting, supervision, equipment use, feedback, or instructional quality. Because the transition bundled online delivery of theory with consolidation of practice, these components cannot be distinguished, and the observed difference is consistent with either or with other unmeasured implementation differences.

A third concern is attendance-based selection. The retained administrative dataset was assembled after program-completion and attendance requirements had already been applied. Because the original enrollment rosters and records of exclusions before dataset assembly were unavailable, we could not determine cohort-specific enrollment or exclusion counts. The direction and magnitude of any resulting selection bias therefore cannot be quantified.

Implications for Instructional Design

The knowledge difference met the prespecified equivalence criterion, whereas the psychomotor difference did not. This contrast concerns observed outcomes and does not identify the responsible program component, because delivery format, practice scheduling, and cohort year changed together and the proposed instructional mechanisms were not measured. If the psychomotor difference reflects instructional sequencing, then programs adopting comparable formats may need to examine whether psychomotor training remains sufficiently intensive, feedback-rich, and temporally connected to theoretical learning; video-based skill previews, mental rehearsal, and greater density of early hands-on sessions are candidates for prospective testing.

Theoretical Contribution

This study contributes to the literature on digital and blended health professions education by reporting 3 competency domains separately under 1 assessment system following a bundled program-level transition. Prior work has largely compared overall effectiveness; separate estimates avoid collapsing distinct score scales into a single summary judgment. The primary evidence therefore remains the estimated differences, their CIs, effect sizes, and competency-threshold results.

Separate reporting of competency domains may be useful in other procedural and simulation-based disciplines, including nursing, surgical training, and resuscitation education, as a prompt to disaggregate outcomes. More broadly, digital education interventions may warrant evaluation not only by aggregate outcomes but also by how different competencies perform.

Limitations

This study has several limitations. The nonrandomized historical-cohort design prevents causal inference, and temporal confounding across the 3-year interval cannot be excluded. Period-specific assessment item sets, checklists, and scenarios were unavailable, so measurement comparability across years could not be verified. The exclusion of cohorts from 2021 and 2022, when delivery format varied under changing restrictions and enrollment was reduced, means that the comparison captures the earliest stable posttransition cohort and cannot describe instructor adaptation during the intervening years. The educational importance margins were based on a single instructor consensus discussion that included 3 study authors and were not independently validated thresholds. The reported κ values were historical interrater agreement estimates from a separate 2020 instructor calibration exercise rather than duplicate ratings of the analyzed assessments. Because the original calibration forms and complete information on the numbers of raters and performances were unavailable, the applicability of these estimates to the assessments administered during the study periods could not be verified.

Several implementation and selection features were also unmeasured. We did not collect measures of cognitive load, feedback quality, rehearsal continuity, or schema formation, so the proposed instructional mechanisms remain untested. The interval between online theory and the first onsite day varied from 1 day to approximately 3 weeks, but the records do not identify each trainee's offering. Although the authors confirmed that the onsite practical-training model, teaching materials, and instructor pool did not change between periods, this program-level confirmation did not provide a session-level audit. Cohort-specific equipment use, scenario allocation, repetitions, feedback, debriefing, and assessment-station configuration were unavailable for retrospective verification. Cohort-specific exclusion counts and reasons were also unavailable, so attendance-based selection cannot be quantified.

Generalizability is limited by the single-center setting and by the ancillary survey, which was completed by 236 of 581 (40.6%) blended-cohort participants. The survey findings are exploratory, describe survey participants only, and cannot be generalized to the full blended cohort or used to explain the between-cohort differences. The outcome window was also narrow: we assessed immediate posttraining performance without retention or field-performance data, and technical skills decay at rates that differ by skill and initial proficiency [42]. Multicenter designs with longitudinal follow-up and direct measures of implementation and learning processes would address these gaps.

Conclusions

In this single-center, nonrandomized comparison of cohorts 3 years apart, the institutional transition to blended delivery was

XSL•FO
RenderX

associated with a 3.84-point lower early psychomotor score in the blended cohort (95% CI 2.67-5.01). The between-cohort differences were −0.30 points for knowledge (95% CI −1.68 to 1.08) and −0.94 points for integrated scenario performance (95% CI −1.89 to 0.00). The knowledge and integrated scenario differences met the prespecified equivalence criteria; the psychomotor difference did not. These educationally defined margins were not validated clinical thresholds.

Psychomotor-threshold, integrated-scenario-threshold, and overall-certification rates did not differ detectably between cohorts. Knowledge-threshold attainment was 5.4 percentage points higher in the blended cohort (P=.007).

Because delivery format, calendar consolidation, and cohort year changed together, this study cannot isolate the effect of moving didactic instruction online. The proposed instructional mechanisms were not measured and therefore remain candidate explanations rather than findings. Programs adopting comparable models should assess psychomotor outcomes directly rather than infer them from knowledge scores or overall pass rates. Whether denser early practice, shorter theory-to-practice intervals, or supplementary rehearsal would change outcomes requires prospective evaluation.

Acknowledgments

We thank the EMT instructors and administrative staff at this study's center for supporting course delivery and assessments. We also thank the Taiwan Emergency Medical Technicians Association (TEMTA) for their support of the EMT-1 certification training program. We also thank the trainees who participated in this study. Generative AI tools including Claude (Opus 4.6, Opus 4.8, and Opus 5; Anthropic) and Codex (GPT-5 and GPT-5.6; OpenAI) were used during manuscript preparation and revision to assist with draft restructuring, cross-checking the internal consistency of reported aggregate values against author-conducted analyses and source data, reference verification, and document formatting. These tools did not determine this study's design or analysis plan, conduct the final statistical analyses, or interpret the findings. All reported results were independently verified by the authors, and all authors take full responsibility for the final content.

Funding

This study was supported by National Taiwan University Hospital (project numbers: NTUH 112-S0069, NTUH 114-SS0006, and NTUH 115-SS0007). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of this paper.

Data Availability

Deidentified data are available from the authors upon reasonable request.

Authors' Contributions

Conceptualization: CHL, CWY
Data curation: CHL
Formal analysis: CHL, YTC
Funding acquisition: CWY
Methodology: CHL, YTC
Project administration: CWY
Supervision: CWY
Writing – original draft: CHL
Writing – review & editing: CHL, MJH, CWY

Conflicts of Interest

None declared.

Multimedia Appendix 1
Source and development of the educational importance margins.
[DOCX File , 18 KB - mededu_v12i1e94950_app1.docx]

Multimedia Appendix 2
Survey subscale scores by prior e-learning experience (blended cohort, n=236).
[DOCX File , 14 KB - mededu_v12i1e94950_app2.docx]

Multimedia Appendix 3

XSL•FO
RenderX

Multiple linear regression model for future e-learning intentions (blended cohort, n=236).
[DOCX File , 15 KB - mededu_v12i1e94950_app3.docx]

References

1. Li M, Hong Y, Wu A, Ou W, Yang J, Lin G, et al. The effectiveness of blended learning in nursing and medical education: an umbrella review. Nurse Educ Pract 2025;86:104421 [FREE Full text] [doi: 10.1016/j.nepr.2025.104421] [Medline: 40483931]
2. Rose S. Medical student education in the time of COVID-19. JAMA 2020;323(21):2131-2132. [doi: 10.1001/jama.2020.5227] [Medline: 32232420]
3. Thorne CJ, Kimani PK, Hampshire S, Hamilton-Bower I, Begum-Ali S, Benson-Clarke A, et al. The nationwide impact of COVID-19 on life support courses. A retrospective evaluation by Resuscitation Council UK. Resusc Plus 2023;13:100366 [FREE Full text] [doi: 10.1016/j.resplu.2023.100366] [Medline: 36816597]
4. Li C, He J, Yuan C, Chen B, Sun Z. The effects of blended learning on knowledge, skills, and satisfaction in nursing students: a meta-analysis. Nurse Educ Today 2019;82:51-57. [doi: 10.1016/j.nedt.2019.08.004] [Medline: 31437783]
5. Garrison DR, Kanuka H. Blended learning: uncovering its transformative potential in higher education. Internet Higher Educ 2004;7(2):95-105. [doi: 10.1016/j.iheduc.2004.02.001]
6. Graham CR. Emerging practice and research in blended learning. In: Moore MG, editor. Handbook of Distance Education, 3rd ed. New York, NY: Routledge; 2013:333-350.
7. McGee RG, Wark S, Mwangi F, Drovandi A, Alele F, Malau-Aduli BS, ACHIEVE Collaboration. Digital learning of clinical skills and its impact on medical students' academic performance: a systematic review. BMC Med Educ 2024;24(1):1477 [FREE Full text] [doi: 10.1186/s12909-024-06471-2] [Medline: 39696150]
8. Tudor Car L, Poon S, Kyaw BM, Cook DA, Ward V, Atun R, et al. Digital education for health professionals: an evidence map, conceptual framework, and research agenda. J Med Internet Res 2022 Mar 17;24(3):e31977 [FREE Full text] [doi: 10.2196/31977] [Medline: 35297767]
9. Venkatesh V, Morris MG, Davis GB, Davis FD. User acceptance of information technology: toward a unified view. MIS Q 2003;27(3):425-478. [doi: 10.2307/30036540]
10. Davis FD. Perceived usefulness, perceived ease of use, and user acceptance of information technology. MIS Q 1989;13(3):319-340. [doi: 10.2307/249008]
11. Lee JWY, Tan JY, Bello F. Technology acceptance model in medical education: systematic review. JMIR Med Educ 2025;11:e67873 [FREE Full text] [doi: 10.2196/67873] [Medline: 40668663]
12. Prior DD, Mazanov J, Meacheam D, Heaslip G, Hanson J. Attitude, digital literacy and self efficacy: flow-on effects for online learning behavior. Internet Higher Educ 2016;29:91-97. [doi: 10.1016/j.iheduc.2016.01.001]
13. Mohammadyari S, Singh H. Understanding the effect of e-learning on individual performance: the role of digital literacy. Comput Educ 2015;82:11-25. [doi: 10.1016/j.compedu.2014.10.025]
14. van de Werfhorst HG, Kessenich E, Geven S. The digital divide in online education: inequality in digital readiness of students and schools. Comput Educ Open 2022;3:100100 [FREE Full text] [doi: 10.1016/j.caeo.2022.100100] [Medline: 40478103]
15. Sweller J. Cognitive load during problem solving: effects on learning. Cogn Sci 1988;12(2):257-285. [doi: 10.1207/s15516709cog1202_4]
16. Ericsson KA, Krampe RT, Tesch-Römer C. The role of deliberate practice in the acquisition of expert performance. Psychol Rev 1993;100(3):363-406. [doi: 10.1037//0033-295x.100.3.363]
17. McGaghie WC, Issenberg SB, Cohen ER, Barsuk JH, Wayne DB. Does simulation-based medical education with deliberate practice yield better results than traditional clinical education? A meta-analytic comparative review of the evidence. Acad Med 2011;86(6):706-711 [FREE Full text] [doi: 10.1097/ACM.0b013e318217e119] [Medline: 21512370]
18. Karpicke JD, Roediger HL. The critical importance of retrieval for learning. Science 2008;319(5865):966-968. [doi: 10.1126/science.1152408] [Medline: 18276894]
19. Jørgensen HK, Vamadevan A, Konge L, Hertz P, Bjerrum F. Distributed training vs. massed practice for surgical skills training-a systematic review. Surg Endosc 2025;39(1):39-63. [doi: 10.1007/s00464-024-11408-3] [Medline: 39586877]
20. Soares RV, Pedrosa RBDS, Sandars J, Cecilio-Fernandes D. The importance of combined use of spacing and testing effects for complex skills training: a quasi-experimental study. Med Teach 2025;47(8):1296-1303 [FREE Full text] [doi: 10.1080/0142159X.2024.2427735] [Medline: 39535960]
21. Venkatesh V, Davis FD. A theoretical extension of the technology acceptance model: four longitudinal field studies. Manage Sci 2000;46(2):186-204. [doi: 10.1287/mnsc.46.2.186.11926]
22. Eshet-Alkalai Y. Digital literacy: a conceptual framework for survival skills in the digital era. J Educ Multimedia Hypermedia 2004;13(1):93-106 [FREE Full text]
23. Ng W. Can we teach digital natives digital literacy? Comput Educ 2012;59(3):1065-1078. [doi: 10.1016/j.compedu.2012.04.016]
24. Fredricks JA, Blumenfeld PC, Paris AH. School engagement: potential of the concept, state of the evidence. Rev Educ Res 2004;74(1):59-109. [doi: 10.3102/00346543074001059]

XSL•FO
RenderX

25. Tomek J, Caldwell A, Eisner DA. Beyond "non-significant" results: why and how to test for practical equivalence. Proc Natl Acad Sci U S A 2026;123(33):e2611548123. [doi: 10.1073/pnas.2611548123] [Medline: 42599784]
26. Cook DA, Levinson AJ, Garside S, Dupras DM, Erwin PJ, Montori VM. Internet-based learning in the health professions: a meta-analysis. JAMA 2008;300(10):1181-1196. [doi: 10.1001/jama.300.10.1181] [Medline: 18780847]
27. Ruiz JG, Mintzer MJ, Leipzig RM. The impact of E-learning in medical education. Acad Med 2006;81(3):207-212. [doi: 10.1097/00001888-200603000-00002] [Medline: 16501260]
28. McCutcheon K, Lohan M, Traynor M, Martin D. A systematic review evaluating the impact of online or blended learning vs. face-to-face learning of clinical skills in undergraduate nurse education. J Adv Nurs 2015;71(2):255-270. [doi: 10.1111/jan.12509] [Medline: 25134985]
29. Means B, Toyama Y, Murphy R, Bakia M, Jones K. Evaluation of evidence-based practices in online learning: a meta-analysis and review of online learning studies. Washington, DC: United States Department of Education; 2009. URL: https://files.eric.ed.gov/fulltext/ED505824.pdf [accessed 2026-10-01]
30. Spruit EN, Band GPH, Hamming JF, Ridderinkhof KR. Optimal training design for procedural motor skills: a review and application to laparoscopic surgery. Psychol Res 2014;78(6):878-891. [doi: 10.1007/s00426-013-0525-5] [Medline: 24202287]
31. Lin L, Ni S, Cheng J, Zhang Z, Zeng R, Jin X, et al. Effect of synchronous online vs. face-to-face cardiopulmonary resuscitation training on chest compression quality: a pilot randomized manikin study. Am J Emerg Med 2021;50:80-84. [doi: 10.1016/j.ajem.2021.07.009] [Medline: 34314941]
32. Ko YC, Lin HY, Chiang WC, Yang CW, Hsieh MJ, Ma MHM. Comparing the effects of blended learning and traditional instruction on basic life support for laypersons: a randomized controlled trial. J Formos Med Assoc 2024;123(6):687-692 [FREE Full text] [doi: 10.1016/j.jfma.2023.10.017] [Medline: 37996323]
33. Boonmak S, Plailaharn N, Sripadungkul D, Somjit M, Gaysonsiri D, Boonmak P. A randomized controlled, non-inferiority trial of Moodle online learning for basic life support training on learning outcomes among dental students. J Dent Educ 2023;87(1):110-117. [doi: 10.1002/jdd.13089] [Medline: 36052470]
34. Iskander M, Haverkamp FJC, Geeraedts LMG, van Bilsen SJA, Tan ECTH. Evaluation of a redesigned two-day trauma course incorporating blended learning compared with the traditional three-day trauma course: a retrospective non-inferiority study. World J Emerg Surg 2026 [FREE Full text] [doi: 10.1186/s13017-026-00714-x] [Medline: 42449390]
35. Sarboozi-Hosseinabadi S, Sharifzadeh G, Hosseini SM. Evaluating CPR training: simulation vs. webinars for Iranian emergency medical technicians during COVID-19. BMC Emerg Med 2024;24(1):46 [FREE Full text] [doi: 10.1186/s12873-024-00967-2] [Medline: 38500059]
36. Bieri J, Tuor C, Nendaz M, Savoldelli GL, Blondon K, Schiffer E, et al. Implementation of a student-teacher-based blended curriculum for the training of medical students for nasopharyngeal swab and intramuscular injection: mixed methods pre-post and satisfaction surveys. JMIR Med Educ 2023;9:e38870 [FREE Full text] [doi: 10.2196/38870] [Medline: 36862500]
37. Lim XMA, Liao WA, Wang W, Seah B. The effectiveness of technology-based cardiopulmonary resuscitation training on the skills and knowledge of adolescents: systematic review and meta-analysis. J Med Internet Res 2022;24(12):e36423 [FREE Full text] [doi: 10.2196/36423] [Medline: 36520524]
38. Chong KM, Yang HW, He HC, Lien WC, Yang MF, Chi CY, et al. The effectiveness of online-only blended cardiopulmonary resuscitation training: static-group comparison study. J Med Internet Res 2023;25:e42325 [FREE Full text] [doi: 10.2196/42325] [Medline: 37018023]
39. Chien CY, Tsai SL, Huang CH, Wang MF, Lin CC, Chen CB, et al. Effectiveness of blended versus traditional refresher training for cardiopulmonary resuscitation: prospective observational study. JMIR Med Educ 2024;10:e52230 [FREE Full text] [doi: 10.2196/52230] [Medline: 38683663]
40. Chen F, Lui AM, Martinelli SM. A systematic review of the effectiveness of flipped classrooms in medical education. Med Educ 2017;51(6):585-597. [doi: 10.1111/medu.13272] [Medline: 28488303]
41. Gagnon MP, Desmartis M, Labrecque M, Car J, Pagliari C, Pluye P, et al. Systematic review of factors influencing the adoption of information and communication technologies by healthcare professionals. J Med Syst 2012;36(1):241-277 [FREE Full text] [doi: 10.1007/s10916-010-9473-4] [Medline: 20703721]
42. de Andres Crespo M, Lykoudis PM, Myint F, Berlingieri P. Surgery and technical skill decay. Int J Surg 2025;111(5):3399-3413. [doi: 10.1097/JS9.0000000000002313] [Medline: 40085756]

Abbreviations

EMS: emergency medical services
EMT: emergency medical technician
EMT-1: emergency medical technician level 1
EMT-2: emergency medical technician level 2
TR1: knowledge assessment
TR2: psychomotor skills assessment

XSL•FO
RenderX

TR3: integrated scenario assessment

Edited by J Gentges; submitted 12.Apr.2026; peer-reviewed by J Afonso, O Trabelsi; comments to author 09.Aug.2026; revised version received 12.Sep.2026; accepted 28.Sep.2026; published 27.Oct.2026.

Please cite as:
Liu CH, Hsieh MJ, Chen YT, Yang CW
Psychomotor Performance in Blended Versus In-Person Emergency Medical Technician Training: Retrospective Cohort Study
JMIR Med Educ 2026;12:e94950
URL: https://mededu.jmir.org/2026/1/e94950
doi:10.2196/94950
PMID:

©Cheng-Heng Liu, Ming-Ju Hsieh, Yu-Ting Chen, Chih-Wei Yang. Originally published in JMIR Medical Education (https://mededu.jmir.org), 27.Oct.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Education, is properly cited. The complete bibliographic information, a link to the original publication on https://mededu.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Rethinking Medical Spanish Competence: Language Proficiency, Clinical Communication, and Inclusion

Juli McCarroll, MA, MLIS

Department of Medical Library, Western Michigan University Homer Stryker M.D. School of Medicine, 1000 Oakland Drive, Kalamazoo, MI, United States

Corresponding Author:
Juli McCarroll, MA, MLIS
Department of Medical Library, Western Michigan University Homer Stryker M.D. School of Medicine, 1000 Oakland Drive, Kalamazoo, MI, United States

Related Articles:

https://mededu.jmir.org/2026/1/e98398

Comment in: https://mededu.jmir.org/2026/1/e110709

(JMIR Med Educ 2026;12:e110302) doi:10.2196/110302

KEYWORDS

medical Spanish; inclusive communication; LGBTQ+ health; bias; clinical communication; language concordance; standardized patients; simulation training; digital education; curriculum development

I commend Lopez Vera [1] for her patient-informed framework integrating inclusive, bias-aware communication with Spanish-speaking lesbian, gay, bisexual, transgender, queer, plus (LGBTQ+) patients as a clinical skill in medical Spanish education. The tutorial appropriately moves beyond vocabulary acquisition by incorporating patient-informed communication priorities, curricular design principles, and integrated learning objectives. This framework also creates an opportunity to address a broader question: what does it mean for a learner to be competent to provide clinical care in Spanish?

As an LGBTQ+ community member, I view competence more broadly. A clinician may possess substantial Spanish proficiency but communicate in ways that make an LGBTQ+ patient feel unseen or unsafe. Conversely, a learner may understand inclusive terminology but lack the Spanish proficiency necessary for complex clinical encounters. These are different competencies, and patients experience the consequences of both.

The emerging qualified multilingual assessment (QMA) framework described by Molina et al [2] and Biswas et al [3] offers an important complement to this discussion. The QMA was developed to assess oral language proficiency for direct clinical care rather than for qualification as a medical interpreter. Importantly, demonstrating proficiency does not eliminate the need for interpreter services when a clinical situation exceeds a clinician's demonstrated language abilities. This distinction is particularly relevant for inclusive medical Spanish education.

A learner may demonstrate inclusive communication behaviors, such as asking a patient for their preferred form of address, avoiding heteronormative assumptions, and using open-ended relationship language, without possessing sufficient Spanish proficiency for independently conducting a complex clinical encounter. Conversely, a learner may demonstrate substantial Spanish proficiency while communicating in ways that are linguistically accurate but clinically exclusionary. Lopez Vera's [1] framework recognizes this latter possibility, emphasizing that inclusive communication requires more than substituting terminology and involves adapting language to patients' preferences.

These observations suggest that language proficiency and inclusive clinical communication should be assessed as complementary, rather than interchangeable, competencies. This raises several questions for medical educators. Should readiness for language-concordant care require a defined level of clinical language proficiency and demonstrated competence in patient-centered, bias-aware communication? Should assessments distinguish between knowing inclusive terminology and sustaining inclusive communication while obtaining a history, explaining a diagnosis, counseling a patient, and responding to unexpected clinical information? Most importantly, should curricula explicitly teach learners to recognize when their linguistic limitations warrant professional interpretation support?

I propose a three-dimensional model of medical Spanish competence that encompasses clinical language proficiency, clinical communication competence, and inclusive communication competence. This model could help educators determine not only whether a learner speaks Spanish but also what the learner can safely do in Spanish, what additional development is needed, and when interpreter support remains appropriate. This distinction is particularly important because speakers of languages other than English may bear additional responsibility for determining whether their language skills are sufficient for a particular clinical task.

XSL•FO
RenderX

Lopez Vera's [1] tutorial provides an important foundation for patient-informed, inclusive medical Spanish education. Integrating this framework with emerging approaches to multilingual proficiency assessment could move the field toward a more precise and clinically meaningful definition of language-concordant competence.

Funding

The author declared no financial support was received for this work.

Authors' Contributions

Conceptualization: JM
Supervision: JM
Writing – original draft: JM
Writing – review & editing: JM

Conflicts of Interest

None declared.

References

1. Lopez Vera A. Inclusive medical Spanish education to address bias in clinical communication: tutorial on patient-informed module development. JMIR Med Educ 2026 Aug 21;12:e98398. [doi: 10.2196/98398] [Medline: 42627956]
2. Molina RL, Bazan M, Martinez J, Diamond LC, Ortega P. Qualified multilingual assessment policy for US medical students: a national Delphi consensus study. Teach Learn Med 2026;38(3):382-390. [doi: 10.1080/10401334.2025.2545906] [Medline: 40819310]
3. Biswas N, Warner G, Naumowicz M, et al. A qualified multilingual assessment for medical students: a cohort study. Cureus 2026 Feb 11;18(2):e103433. [doi: 10.7759/cureus.103433] [Medline: 41835698]

Abbreviations

LGBTQ+: lesbian, gay, bisexual, transgender, queer, plus
QMA: qualified multilingual assessment

Edited by TDA Cardoso; submitted 24.Aug.2026; this is a non–peer-reviewed article;accepted 08.Sep.2026; published 24.Sep.2026.

Please cite as:
McCarroll J
Rethinking Medical Spanish Competence: Language Proficiency, Clinical Communication, and Inclusion
JMIR Med Educ 2026;12:e110302
URL: https://mededu.jmir.org/2026/1/e110302
doi:10.2196/110302

© Juli McCarroll. Originally published in JMIR Medical Education (https://mededu.jmir.org), 24.Sep.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Education, is properly cited. The complete bibliographic information, a link to the original publication on https://mededu.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Author's Reply: Rethinking Medical Spanish Competence: Language Proficiency, Clinical Communication, and Inclusion

Alexandra Lopez Vera, MPH, PhD

California University of Science and Medicine, 1501 Violet St, Colton, CA, United States

Corresponding Author:
Alexandra Lopez Vera, MPH, PhD
California University of Science and Medicine, 1501 Violet St, Colton, CA, United States

Related Articles:

https://mededu.jmir.org/2026/1/e98398

https://mededu.jmir.org/2026/1/110302

(JMIR Med Educ 2026;12:e110709) doi:10.2196/110709

KEYWORDS

medical Spanish; inclusive communication; LGBTQ+ health; bias; clinical communication; language concordance; standardized patients; simulation training; digital education; curriculum development

I thank the letter writer for this generous engagement with my tutorial [1] and especially for grounding the letter [2] in the lived perspective of a lesbian, gay, bisexual, transgender, queer, plus (LGBTQ+) community member. The letter's central claim is exactly right: Spanish proficiency and inclusive clinical communication are complementary competencies, not interchangeable ones. A fluent clinician can make a patient feel unseen; an inclusive-minded learner can lack the proficiency to conduct an encounter safely. The tutorial's formative rubric embeds this distinction by scoring Spanish language effectiveness separately from nonassumptive language, communication repair, and patient-centeredness, so that fluency cannot compensate for exclusion or vice versa [1]. But the letter is correct that formative, module-level tools cannot define readiness for language-concordant care. The field needs standardized anchors.

The proposed three-dimensional model—clinical language proficiency, clinical communication competence, and inclusive communication competence—is a valuable step, and I would frame its payoff in the language of entrustment. The question is not whether a learner speaks Spanish but what that learner can safely be entrusted to do in Spanish and what supports they need. The qualified multilingual assessment [3,4] could anchor the proficiency dimension with verified thresholds, while standardized patient encounters scored with domain-specific rubrics assess the other two. Combined, they could support graduated scopes of practice: a learner might be entrusted to conduct a routine intake in Spanish yet still require interpreter support for goals-of-care conversations or complex counseling, just as procedural skills are entrusted incrementally.

I would extend one of the letter's questions into a concrete curricular commitment: recognizing when one's linguistic limits warrant professional interpretation is itself a teachable, assessable clinical safety skill. Programs can operationalize it by adding interpreter-transition phrases to the same phrase banks that teach inclusive language (eg, "Para darle la mejor atención posible, voy a pedir el apoyo de un intérprete profesional"); rehearsing that handoff in simulations, so that it preserves rapport rather than rupturing it; and scoring the judgment itself—not only the language—in assessment rubrics. Framing interpreter use as professionalism rather than deficiency matters doubly for heritage and multilingual learners, who often face implicit pressure to stretch beyond their verified abilities. That is a workforce equity issue as much as a patient safety issue.

Integrating verified proficiency assessment, performance-based inclusive communication assessment, and longitudinal follow-up could move us toward a clinically meaningful definition of language-concordant competence. I thank the letter writer for advancing that agenda and welcome collaboration toward it.

Funding

The author declared no financial support was received for this work.

Conflicts of Interest

None declared.

XSL•FO
RenderX

References

1. Lopez Vera A. Inclusive medical Spanish education to address bias in clinical communication: tutorial on patient-informed module development. JMIR Med Educ 2026 Aug 21;12:e98398. [doi: 10.2196/98398] [Medline: 42627956]
2. McCarroll J. Rethinking medical Spanish competence: language proficiency, clinical communication, and inclusion. JMIR Med Educ 2026;12:e110302. [doi: 10.2196/110302]
3. Biswas N, Warner G, Naumowicz M, et al. A qualified multilingual assessment for medical students: a cohort study. Cureus 2026 Feb 11;18(2):e103433. [doi: 10.7759/cureus.103433] [Medline: 41835698]
4. Molina RL, Bazan M, Martinez J, Diamond LC, Ortega P. Qualified multilingual assessment policy for US medical students: a national Delphi consensus study. Teach Learn Med 2026;38(3):382-390. [doi: 10.1080/10401334.2025.2545906] [Medline: 40819310]

Abbreviations

LGBTQ+: lesbian, gay, bisexual, transgender, queer, plus

Edited by TDA Cardoso; submitted 28.Aug.2026; this is a non–peer-reviewed article;accepted 08.Sep.2026; published 24.Sep.2026.

Please cite as:
Lopez Vera A
Author's Reply: Rethinking Medical Spanish Competence: Language Proficiency, Clinical Communication, and Inclusion
JMIR Med Educ 2026;12:e110709
URL: https://mededu.jmir.org/2026/1/e110709
doi:10.2196/110709

© Alexandra Lopez Vera. Originally published in JMIR Medical Education (https://mededu.jmir.org), 24.Sep.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Education, is properly cited. The complete bibliographic information, a link to the original publication on https://mededu.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Authors' Reply: Methodological Concerns in AI Medical Education Frameworks

Juan S Izquierdo-Condoy[1], MD, MSc; Marlon Arias-Intriago[1], MD; Laura Montero Corrales[2], PhD; Esteban Ortiz-Prado[1], MD, MSc, MPH, PhD

[1]One Health Research Group, Universidad de Las Américas, Via Nayon S/N, Quito, Ecuador

[2]Escuela de Comunicación, Universidad Latina de Costa Rica, San José, Costa Rica

Corresponding Author:
Esteban Ortiz-Prado, MD, MSc, MPH, PhD
One Health Research Group, Universidad de Las Américas, Via Nayon S/N, Quito, Ecuador

Related Articles:

https://www.jmir.org/2026/1/e101696

https://www.jmir.org/2025/1/e76340

(JMIR Med Educ 2026;12:e104447) doi:10.2196/104447

KEYWORDS

artificial intelligence; medical education; machine learning; adaptive learning systems; future implications

We thank the authors [1] of the letter "Methodological Concerns in AI Medical Education Frameworks" for their careful reading of our paper [2] and their constructive observations. We welcome the opportunity to clarify the scope of our arguments and the intended role of FACETS (form, application, context, instructional mode, technology, SAMR [substitution, augmentation, modification, redefinition]) in responsible artificial intelligence (AI) integration in medical education.

First, we clarify that the paper by Abdelwanis et al [3] was not cited as direct empirical evidence for educational inequities in low- and middle-income countries (LMICs), nor as evidence for AI implementation, access, effectiveness, or educational outcomes in those settings. We recognize that the review by Abdelwanis et al [3] used a bowtie analysis to examine automation bias, biased data, algorithmic opacity, insufficient validation, and the need for human oversight in health care AI. In our manuscript, this reference was used as conceptual support for general risk mechanisms that may compromise safe and equitable AI adoption. We agree that LMIC-specific claims require more directly contextualized evidence.

We also agree that the study by Ong et al [4] was more closely related to AI training in medical education across income settings, as it showed differences between experts from high-income countries and LMICs in prioritizing AI learning outcomes. However, this evidence should also be interpreted precisely. Ong et al [4] provide empirical evidence on expert perceptions and curricular prioritization, not direct evidence on real-world AI use, effectiveness, access, costs, infrastructure, or educational outcomes among medical students in LMICs. This distinction reinforces our central argument: further contextual, multicenter, and implementation-oriented research is needed [5].

Second, regarding FACETS, we believe the letter interprets our proposal beyond its intended scope. FACETS was not proposed as a predictive, psychometric, or safety-validation instrument, nor as a prospectively validated universal standard. Rather, it was proposed as a conceptual and analytical framework derived from a critical synthesis of published AI applications in medical education to help describe, compare, and interpret heterogeneous interventions across educational form, application, context, instructional mode, technology, and degree of pedagogical transformation.

Accordingly, Luordo et al [6] and Holderried et al [7] were used as empirical examples to illustrate the conceptual applicability of FACETS, not as validation studies. Their limitations reinforce, rather than weaken, the rationale for such a framework. In Luordo et al [6], AI-assisted objective structured clinical examination (OSCE) grading showed promising efficiency and concordance but was systematically stricter than expert evaluators, underscoring the importance of item design, contextual interpretation, and human oversight. In Holderried et al [7], high overall concordance coexisted with lower agreement in selected feedback categories, likely related to ambiguity, category overlap, or interpretive differences. These findings highlight the need for multidimensional frameworks that make explicit the educational purpose, task, context, technological configuration, instructor role, and requirements for local validation.

We appreciate the authors' effort to enrich the scientific literature on AI in medical education, a field that remains at an early stage. At the same time, we believe that our proposal should be interpreted within the conceptual scope in which it was presented. This discussion reaffirms that our review offers a legitimate and useful conceptual proposal to organize evidence,

XSL•FO
RenderX

identify gaps, guide future research, and promote AI adoption in medical education in a pedagogically coherent, ethically responsible, and context-sensitive manner.

Conflicts of Interest

None declared.

References

1. Fatima SU, Fatima E, Kumari R. Methodological concerns in Al medical education frameworks. JMIR Med Educ 2026;12:e101696. [doi: 10.2196/101696]
2. Izquierdo-Condoy JS, Arias-Intriago M, Montero Corrales L, Ortiz-Prado E. Artificial intelligence in medical education: transformative potential, current applications, and future implications. JMIR Med Educ 2026 Feb 17;12:e77127. [doi: 10.2196/77127] [Medline: 41701936]
3. Abdelwanis M, Alarafati HK, Tammam MMS, Simsekler MCE. Exploring the risks of automation bias in healthcare artificial intelligence applications: a bowtie analysis. J Saf Sci Resil 2024 Dec;5(4):460-469. [doi: 10.1016/j.jnlssr.2024.06.001]
4. Ong QC, Ang CS, Lai NM, Atun R, Car J. Differences in expert perspectives on AI training in medical education: secondary analysis of a multinational Delphi study. J Med Internet Res 2025 May 9;27(1):e72186. [doi: 10.2196/72186] [Medline: 40344200]
5. Izquierdo-Condoy JS, Arias-Intriago M, Tello-De-la-Torre A, Busch F, Ortiz-Prado E. Generative artificial intelligence in medical education: enhancing critical thinking or undermining cognitive autonomy? J Med Internet Res 2025 Nov 3;27:e76340. [doi: 10.2196/76340] [Medline: 41183320]
6. Luordo D, Torres Arrese M, Tristán Calvo C, et al. Application of artificial intelligence as an aid for the correction of the objective structured clinical examination (OSCE). Appl Sci 2025;15(3):1153. [doi: 10.3390/app15031153]
7. Holderried F, Stegemann-Philipps C, Herrmann-Werner A, et al. A language model-powered simulated patient with automated feedback for history taking: prospective study. JMIR Med Educ 2024 Aug 16;10:e59213. [doi: 10.2196/59213] [Medline: 39150749]

Abbreviations

AI: artificial intelligence
FACETS: form, application, context, instructional mode, technology, SAMR [substitution, augmentation, modification, redefinition]
LMIC: low- and middle-income country
OSCE: objective structured clinical examination

Edited by A Stone, S Zelko; submitted 11.Jun.2026; this is a non–peer-reviewed article;accepted 17.Jun.2026; published 06.Jul.2026.

Please cite as:
Izquierdo-Condoy JS, Arias-Intriago M, Montero Corrales L, Ortiz-Prado E
Authors' Reply: Methodological Concerns in AI Medical Education Frameworks
JMIR Med Educ 2026;12:e104447
URL: https://mededu.jmir.org/2026/1/e104447
doi:10.2196/104447

© Juan S Izquierdo-Condoy, Marlon Arias-Intriago, Laura Montero Corrales, Esteban Ortiz-Prado. Originally published in JMIR Medical Education (https://mededu.jmir.org), 6.Jul.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Education, is properly cited. The complete bibliographic information, a link to the original publication on https://mededu.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Authors' Reply: Big Data, Small Stories: Methodological Considerations for Using Social Media Analytics in Medical Education Research

Faisal Binsar[1*], BEng, MSc, PhD; Mohammad Hamsal[2*], Prof Dr

[1]Binus Online Learning, Management Department, Binus University, Jl. K. H. Syahdan No. 9, Kemanggisan, Palmerah, Jakarta, DKI Jakarta, Indonesia

[2]Binus Business School Doctor of Research in Management, Management Department, Binus University, Jakarta, DKI Jakarta, Indonesia

[*]all authors contributed equally

Corresponding Author:
Faisal Binsar, BEng, MSc, PhD
Binus Online Learning, Management Department, Binus University, Jl. K. H. Syahdan No. 9, Kemanggisan, Palmerah, Jakarta, DKI Jakarta, Indonesia

Related Articles:

https://mededu.jmir.org/2026/1/e60452

https://mededu.jmir.org/2026/1/e94825

(JMIR Med Educ 2026;12:e97664) doi:10.2196/97664

KEYWORDS

Icarus paradox; aspirations; realities; medical education; student well-being; community perspective; online reviews

We thank the correspondent for the thoughtful comments [1] regarding our article on the Icarus paradox in Indonesia's specialist medical education system [2]. We appreciate the opportunity to clarify several methodological aspects of our study, particularly concerning sentiment analysis and the interpretation of neutral discourse in large-scale online data.

Our study was anchored in the Icarus paradox framework, which we applied to the Indonesian specialist medical education system, a context where high aspirations often collide with systemic constraints. The primary objective was to capture the *public pulse* through social listening, a method that provides an unmediated reflection of societal discourse. We maintain that the core findings (identifying the tension between the prestige of specialization and the lived realities of the training system) remain robust. The correspondent's appreciation of this framework reinforces its utility in analyzing complex health care education phenomena [2].

Brand24 was selected as the primary social-listening tool because it supports large-scale multilingual data collection, including Indonesian-language online discourse. This enabled the analysis of 5047 public responses across digital platforms.

To reduce the limitations of automated sentiment analysis, we implemented manual validation using two independent coders and conducted intercoder agreement checks following established qualitative standards [3]. This process helped ensure that sentiment classification reflected contextual interpretation rather than solely algorithmic outputs.

Although neutral sentiment dominated the dataset, this did not indicate a lack of analytical value. Using NVivo 14 for thematic analysis [3,4], we identified latent structural tensions related to specialist training capacity, workload, and institutional imbalance that aligned with the Icarus paradox framework.

We acknowledge that the use of a single-method sentiment analysis approach is a limitation of the study [5,6]. Future research may benefit from more advanced models capable of distinguishing informational neutrality from subtle positive or negative orientations in online discourse.

We appreciate the correspondent's methodological reflections and agree that combining automated sentiment analysis with qualitative validation remains essential in digital health research. Our study represents an initial effort to integrate social listening and thematic analysis in examining tensions within specialist medical education in Indonesia. We hope future studies will continue refining these approaches to better understand public discourse in health care education systems.

Conflicts of Interest

None declared.

XSL•FO
RenderX

References

1. Li X, Wang R. Big data, small stories: methodological considerations for using social media analytics in medical education research. JMIR Med Educ 2026;12:e94825. [doi: 10.2196/94825]
2. Binsar F, Hamsal M. Exploring the Icarus paradox in Indonesia's specialist medical education system using the public perspective from online media: convergent mixed methods study. JMIR Med Educ 2026 Jan 26;12:e60452. [doi: 10.2196/60452] [Medline: 41587450]
3. Jackson K, Bazeley P. Qualitative Data Analysis With NVivo, 3rd edition: Sage Publications; 2019.
4. Limna P. The impact of NVivo in qualitative research: perspectives from graduate students. J Appl Learning Teaching 2023 Aug;6(2). [doi: 10.37074/jalt.2023.6.2.17]
5. Bandorski D, Kurniawan N, Baltes P, et al. Contraindications for video capsule endoscopy. World J Gastroenterol 2016 Dec 7;22(45):9898-9908. [doi: 10.3748/wjg.v22.i45.9898] [Medline: 28018097]
6. Zhang L, Liu B. Sentiment analysis and opinion mining. In: Sammut C, Webb GI, editors. Encyclopedia of Machine Learning and Data Mining 2017:1152-1161. [doi: 10.1007/978-1-4899-7687-1_907]

Edited by S Brini, S Zelko; submitted 08.Apr.2026; this is a non–peer-reviewed article;accepted 05.May.2026; published 01.Jun.2026.

Please cite as:
Binsar F, Hamsal M
Authors' Reply: Big Data, Small Stories: Methodological Considerations for Using Social Media Analytics in Medical Education Research
JMIR Med Educ 2026;12:e97664
URL: https://mededu.jmir.org/2026/1/e97664
doi:10.2196/97664

© Faisal Binsar, Mohammad Hamsal. Originally published in JMIR Medical Education (https://mededu.jmir.org), 1.Jun.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Education, is properly cited. The complete bibliographic information, a link to the original publication on https://mededu.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Big Data, Small Stories: Methodological Considerations for Using Social Media Analytics in Medical Education Research

Xiaoshuai Li[1], PhD; Rui Wang[2], PhD

[1]Department of Blood Transfusion, Shengjing Hospital of China Medical University, Shenyang, Liaoning, China

[2]Department of Stem Cells and Regenerative Medicine, Basic Medical Sciences of China Medical University, NO.77 Puhe Road, Shenyang, Liaoning, China

Corresponding Author:
Rui Wang, PhD
Department of Stem Cells and Regenerative Medicine, Basic Medical Sciences of China Medical University, NO.77 Puhe Road, Shenyang, Liaoning, China

Related Articles:

https://mededu.jmir.org/2026/1/e60452

Comment in: https://mededu.jmir.org/2026/1/e97664

(JMIR Med Educ 2026;12:e94825) doi:10.2196/94825

KEYWORDS

Icarus paradox; aspirations; realities; medical education; student well-being; community perspective; online reviews

I read with great interest the recent article by Binsar and Hamsal [1], who innovatively applied the Icarus paradox framework to examine Indonesia's specialist medical education system through the lens of online public reviews. The use of social listening tools to capture 5047 reviews across multiple platforms represents a timely methodological contribution to the field of digital health professional education. I commend the authors for their ambitious data collection and for highlighting the critical tension between aspirations and reality during medical training.

While Binsar and Hamsal's [1] study makes a valuable contribution by introducing the Icarus paradox framework to medical education research and leveraging large-scale social media data, its heavy reliance on automated sentiment analysis raises a fundamental methodological question that warrants deeper scrutiny. Neutral sentiment dominated the dataset, comprising 94.9% of news portal reviews, 93.2% of blog posts, and 86.6% of website mentions. While the authors interpret this as "public discourse on specialist medical education tends to be descriptive or informational rather than overtly evaluative," this neutrality more likely reflects inherent limitations of automated analysis in capturing context-dependent meanings—a methodological blind spot that surface-level polarity detection cannot address. Previous research demonstrates that discussions about professional training often carry latent meanings—such as tacit acceptance of systemic pressures, veiled critique, or discursive normalization—that simple sentiment classification inevitably misses [2-4].

To move beyond this impasse, we propose a more rigorous and integrated methodological pathway. Future studies should use a sequential explanatory design that operates in two phases. In the first phase, social listening tools would be used to map the discursive landscape and identify sentiment polarities as well as "sentiment silence zones"—clusters of highly neutral discourse in which meaning may be latent rather than explicit. In the second phase, purposive sampling would be based on the findings of the first phase, selecting representative texts from these silence zones for in-depth qualitative analysis using methods such as critical discourse analysis or narrative inquiry. This "quantitative mapping, qualitative mining" strategy would enable researchers to ask not merely "What sentiments are being expressed?" but also "What cultural work is this discourse performing?" and "How might apparently neutral language tacitly reproduce or contest systemic tensions in medical education?"

Recent methodological innovations in health professions education [2-4] offer rigorous frameworks for accessing qualitative depth. Kahlke et al [2] demonstrate how innovative elicitation techniques can access tacit dimensions of professional experience, Kinnear et al [3] reveal how apparently neutral professional discourse conceals deep assumptions about educational quality, and MacLeod et al [4] provide systematic frameworks for understanding how language constitutes professional realities. Their deliberate integration with computational social listening represents the next frontier for the field.

Despite this methodological caveat, this study makes an important contribution by amplifying public voices. Understanding the Icarus paradox ultimately requires methods capable of capturing how systemic aspirations and vulnerabilities are lived and articulated in everyday discourse. We look forward to seeing this line of inquiry evolve.

XSL•FO
RenderX

Conflicts of Interest

None declared.

References

1. Binsar F, Hamsal M. Exploring the Icarus paradox in Indonesia's specialist medical education system using the public perspective from online media: convergent mixed methods study. JMIR Med Educ 2026 Jan 26;12:e60452. [doi: 10.2196/60452] [Medline: 41587450]
2. Kahlke R, Maggio LA, Lee MC, et al. When words fail us: an integrative review of innovative elicitation techniques for qualitative interviews. Med Educ 2025 Apr;59(4):382-394. [doi: 10.1111/medu.15555] [Medline: 39412120]
3. Kinnear B, Schumacher DJ, Varpio L, Driessen EW, Konopasky A. Legitimation without argumentation: an empirical discourse analysis of "Validity as an Argument" in assessment. Perspect Med Educ 2024 Oct 3;13(1):469-480. [doi: 10.5334/pme.1404] [Medline: 39372230]
4. MacLeod A, Ellaway RH, Cleland J. A meta-study analysing the discourses of discourse analysis in health professions education. Med Educ 2024 Sep;58(9):1058-1070. [doi: 10.1111/medu.15309] [Medline: 38212063]

Edited by S Brini, S Zelko; submitted 07.Mar.2026; this is a non–peer-reviewed article;accepted 05.May.2026; published 01.Jun.2026.

Please cite as:
Li X, Wang R
Big Data, Small Stories: Methodological Considerations for Using Social Media Analytics in Medical Education Research
JMIR Med Educ 2026;12:e94825
URL: https://mededu.jmir.org/2026/1/e94825
doi:10.2196/94825

©Xiaoshuai Li, Rui Wang. Originally published in JMIR Medical Education (https://mededu.jmir.org), 1.Jun.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Education, is properly cited. The complete bibliographic information, a link to the original publication on https://mededu.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Methodological Concerns in AI Medical Education Frameworks

Syeda Urooj Fatima, MBBS; Emaan Fatima, MBBS; Ragni Kumari, MBBS

Dow University of Health Sciences, E-120 Al-Falah Housing Project, Malir Halt, Karachi, Karachi, Sindh, Pakistan

Corresponding Author:
Emaan Fatima, MBBS
Dow University of Health Sciences, E-120 Al-Falah Housing Project, Malir Halt, Karachi, Karachi, Sindh, Pakistan

Related Articles:

https://www.jmir.org/2025/1/e76340

Comment in: https://www.jmir.org/2026/1/e104447

(JMIR Med Educ 2026;12:e101696) doi:10.2196/101696

KEYWORDS

artificial intelligence; medical education; machine learning; adaptive learning systems; future implications

We read with great interest the recently published viewpoint by Izquierdo-Condoy et al [1] on artificial intelligence (AI) in medical education and their proposed FACETS (form, application, context, instructional mode, technology, SAMR [substitution, augmentation, modification, redefinition]) framework. While we commend the authors for a timely overview, we wish to highlight two concerns, the resolution of which may further strengthen the article.

First, the authors cite Abdelwanis et al [2] to support the claims about AI-driven educational inequities in low- and middle-income countries (LMICs). However, the cited study presents a conceptual bowtie analysis of automation bias conducted at an institution in a high-income country (the United Arab Emirates) with no LMIC populations, data, or contextual analysis—and explicitly excludes human or animal participants, further restricting its empirical validity. Therefore, citing this reference to support an LMIC-specific argument is a form of citation distortion; that is, it is deploying a source beyond the boundaries of its actual scope. This is not a minor citation error. Equity arguments built on mismatched evidence invite dismissal by reviewers and policymakers —weakening advocacy where it is most needed. If LMIC-specific AI education gaps cannot be substantiated with direct evidence, then there is a real risk that resource allocation decisions and curriculum reform efforts in low-income settings will be deprioritized or delayed. This matters because LMIC disparities in AI medical education are real and widening: Ong et al [3] demonstrated that LMIC experts were significantly less likely to consider AI learning outcomes mandatory compared to high-income country peers—a finding that would have directly and legitimately supported the authors' argument.

Second, the authors retrospectively mapped the FACETS framework onto 2 single-institution feasibility studies by authors from well-resourced European universities—Holderried et al [4] (n=106 conversations in Tübingen) and Luordo et al [5] (n=96 students in Madrid), while acknowledging that these studies did not explicitly use FACETS. Critically, both studies revealed unresolved AI performance gaps: 8 of 45 feedback categories showed low concordance with human raters in Holderried et al [4], and AI scoring was systematically stricter than expert grading, by 3.51 points, in Luordo et al [5]—concerns that FACETS fails to address. Retrospective application shows descriptive compatibility, not predictive validity —a key distinction for responsible advocacy. A framework that only describes the past cannot help institutions anticipate AI failures. In high-stakes contexts like objective structured clinical examination (OSCE) grading or history-taking, such failures directly harm assessment integrity and patient safety preparation. If institutions in diverse, resource-limited settings adopt FACETS based on this evidence, they risk implementing AI-driven assessments calibrated to well-resourced European contexts, potentially disadvantaging students in settings where infrastructure, language, and clinical exposure differ substantially. Before FACETS is supported as a guiding standard, it must undergo multi-institutional, cross-cultural validation.

As AI integrates into medical education, frameworks must be evaluated across diverse settings. Without rigorous validation, weak evidence will undermine equity efforts, and premature framework adoption will shape curricula that are difficult to reverse — ultimately compromising the preparedness of future physicians in the very places that need improvement most.

Conflicts of Interest

None declared.

XSL•FO
RenderX

References

1. Izquierdo-Condoy JS, Arias-Intriago M, Montero Corrales L, Ortiz-Prado E. Artificial intelligence in medical education: transformative potential, current applications, and future implications. JMIR Med Educ 2026 Feb 17;12:e77127. [doi: 10.2196/77127] [Medline: 41701936]
2. Abdelwanis M, Alarafati HK, Tammam MMS, Simsekler MCE. Exploring the risks of automation bias in healthcare artificial intelligence applications: a bowtie analysis. J Saf Sci Resil 2024 Dec;5(4):460-469. [doi: 10.1016/j.jnlssr.2024.06.001]
3. Ong QC, Ang CS, Lai NM, Atun R, Car J. Differences in expert perspectives on AI training in medical education: secondary analysis of a multinational Delphi study. J Med Internet Res 2025 May 9;27:e72186. [doi: 10.2196/72186] [Medline: 40344200]
4. Holderried F, Stegemann-Philipps C, Herrmann-Werner A, et al. A language model-powered simulated patient with automated feedback for history taking: prospective study. JMIR Med Educ 2024 Aug 16;10:e59213. [doi: 10.2196/59213] [Medline: 39150749]
5. Luordo D, Torres Arrese M, Tristán Calvo C, et al. Application of artificial intelligence as an aid for the correction of the objective structured clinical examination (OSCE). Appl Sci 2025;15(3):1153. [doi: 10.3390/app15031153]

Abbreviations

AI: artificial intelligence
FACETS: form, application, context, instructional mode, technology, [substitution, augmentation, modification, redefinition]
LMIC: low- and middle-income country
OSCE: objective structured clinical examination

Edited by A Stone, S Zelko; submitted 18.May.2026; this is a non–peer-reviewed article;accepted 17.Jun.2026; published 06.Jul.2026.

Please cite as:
Fatima SU, Fatima E, Kumari R
Methodological Concerns in AI Medical Education Frameworks
JMIR Med Educ 2026;12:e101696
URL: https://mededu.jmir.org/2026/1/e101696
doi:10.2196/101696

© Syeda Urooj Fatima, Emaan Fatima, Ragni Kumari. Originally published in JMIR Medical Education (https://mededu.jmir.org), 6.Jul.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Education, is properly cited. The complete bibliographic information, a link to the original publication on https://mededu.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Documentation Quality and Educational Value in AI-Assisted Feedback

Amane Endo[1,2], MD, PhD; Takeshi Kimura[3], MHPE, MD; Yuki Kataoka[2,3,4], MPH, MD, DrPH

[1]Department of Medical Education, Faculty of Medicine, Juntendo University, 2-1-1 Hongo, Bunkyo-ku, Tokyo, Japan

[2]Center for Medical Education, Graduate School of Medicine, Nagoya University, Nagoya, Japan

[3]Center for Postgraduate Clinical Training and Career Development, Nagoya University, Nagoya, Japan

[4]Scientific Research WorkS Peer Support Group (SRWS-PSG), Osaka, Japan

Corresponding Author:
Amane Endo, MD, PhD
Department of Medical Education, Faculty of Medicine, Juntendo University, 2-1-1 Hongo, Bunkyo-ku, Tokyo, Japan

Related Article:

https://mededu.jmir.org/2026/1/e89996

(JMIR Med Educ 2026;12:e105029) doi:10.2196/105029

KEYWORDS

ambient scribe; artificial intelligence; feedback; medical student education; formative assessment; competency-based education; AI

Talwalkar and colleagues [1] address an important problem in medical education: faculty often observe learner performance but struggle to translate those observations into written feedback. Their randomized evaluation of an ambient AI scribe workflow is timely and practically relevant. The study is also valuable because it examines a real educational workflow, includes instructor review of AI-generated notes, and explicitly evaluates unedited AI summaries for mischaracterization and hallucination. These features make the study an important contribution to the emerging literature on AI-assisted feedback documentation. However, there are concerns regarding the interpretation of the results.

The central concern is that a well-documented AI-assisted feedback note should be distinguished from feedback that is educationally effective for learners. Their findings show that AI assistance improved the quality of written feedback documentation, as measured by the Evaluation of Feedback Captured Tool (EFeCT) score [1]. The EFeCT assesses whether specific feedback elements are present in the written note. It does not assess whether students understood, accepted, trusted, or used the feedback. This distinction matters because feedback becomes educationally meaningful only when learners can engage with it and use it to guide future action. Students' use of feedback depends on self-regulation, beliefs, emotions, and their ability to identify next steps [2]. Molloy and colleagues [3] similarly caution against treating feedback as a simple input separated from learner involvement and effects beyond the immediate task. Thus, higher EFeCT scores should be understood as evidence of improved documentation quality unless learner uptake and subsequent performance are also examined.

A second concern is that narrative length may have influenced the score difference. As Talwalkar and colleagues [1] also note, human-only narratives were much shorter than AI-assisted outputs. Because the EFeCT awards one point for each feedback element present, longer notes have more opportunity to contain those scored elements and may, therefore, receive higher scores even when they are no more useful to learners. This matters for implementation. If institutions adopt AI tools because they improve documentation metrics, they may produce feedback records that receive higher scores while being more cognitively demanding and harder for learners to process. Shute's [4] review supports feedback that is specific and clear but also emphasizes that elaborated feedback should remain manageable for learners.

Related evidence supports this interpretation. Kondo and colleagues [5] found that AI-generated feedback on clerkship logs was longer and more consistent, whereas supervisor feedback drew on clinical context and professional judgment. This suggests that AI may improve structure and consistency, while human supervisors may provide contextual educational judgment. These are not interchangeable qualities, but they may be complementary.

Taken together, these considerations highlight the need to distinguish documentation quality from educational usefulness in future research. In addition to documentation metrics, evaluations should include learner understanding, perceived actionability, trust, feedback use, subsequent performance, and the burden associated with reading longer feedback. These outcomes would clarify whether AI-assisted feedback truly improves learning.

XSL•FO
RenderX

Conflicts of Interest

None declared.

Editorial Notice

The corresponding author of *"Ambient AI Scribes to Create Educational Feedback Notes for Medical Students: Randomized Trial"* declined to respond to this letter.

References

1. Talwalkar JS, Chartash D, Zhang L, et al. Ambient AI scribes to create educational feedback notes for medical students: randomized trial. JMIR Med Educ 2026 May 28;12:e89996. [doi: 10.2196/89996] [Medline: 42208058]
2. Spooner M, Larkin J, Liew SC, Jaafar MH, McConkey S, Pawlikowska T. "Tell me what is 'better'!" How medical students experience feedback, through the lens of self-regulatory learning. BMC Med Educ 2023 Nov 22;23(1):895. [doi: 10.1186/s12909-023-04842-9] [Medline: 37993832]
3. Molloy E, Ajjawi R, Bearman M, Noble C, Rudland J, Ryan A. Challenging feedback myths: values, learner involvement and promoting effects beyond the immediate task. Med Educ 2020 Jan;54(1):33-39. [doi: 10.1111/medu.13802] [Medline: 31475387]
4. Shute VJ. Focus on formative feedback. Rev Educ Res 2008 Mar;78(1):153-189. [doi: 10.3102/0034654307313795]
5. Kondo T, Donkers J, Nishigori H, Rovers S, Heeneman S. AI-generated versus human supervisor feedback on medical students' clinical clerkship logs: cross-sectional convergent mixed methods study. JMIR Med Educ 2026 Jun 16;12:e90064. [doi: 10.2196/90064] [Medline: 42302283]

Abbreviations

EFeCT: Evaluation of Feedback Captured Tool

Edited by A Stone; submitted 18.Jun.2026; this is a non–peer-reviewed article;accepted 15.Jul.2026; published 26.Aug.2026.

Please cite as:
Endo A, Kimura T, Kataoka Y
Documentation Quality and Educational Value in AI-Assisted Feedback
JMIR Med Educ 2026;12:e105029
URL: https://mededu.jmir.org/2026/1/e105029
doi:10.2196/105029

© Amane Endo, Takeshi Kimura, Yuki Kataoka. Originally published in JMIR Medical Education (https://mededu.jmir.org), 26.Aug.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Education, is properly cited. The complete bibliographic information, a link to the original publication on https://mededu.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Enhancing Team-Based Learning in Virtual Environments: The Role of Avatar Agency and Immersive Social Presence

Hiroki Funao[1], PhD, RN; Motomu Shimaoka[2], MD, PhD; Jun Kako[3], PhD, RN, OCNS

[1]Mie Prefectural College Of Nursing, 1-1-1 Yumegaoka, Tsu, Mie, Japan

[2]Department of Molecular Pathobiology and Cell Adhesion Biology, Graduate School of Medicine, Mie University, Tsu, Mie, Japan

[3]Department of Nursing, Graduate School of Medicine, Mie University, Tsu, Mie, Japan

Corresponding Author:
Hiroki Funao, PhD, RN
Mie Prefectural College Of Nursing, 1-1-1 Yumegaoka, Tsu, Mie, Japan

Related Articles:

https://mededu.jmir.org/2026/1/e80097

Comment in: https://mededu.jmir.org/2026/1/e93060

(JMIR Med Educ 2026;12:e92366) doi:10.2196/92366

KEYWORDS

anesthesia; computer-assisted instruction; distance; education; internet; learning; medical; problem-based learning; students; teaching; virtual reality

We read with great interest the report by Sripadungkul et al [1] comparing web-based virtual environment (WBVE) with face-to-face (F2F) delivery for team-based learning (TBL) of anesthesia techniques. Their finding that WBVE-delivered TBL yields knowledge gains similar to those of F2F delivery while offering scalability is encouraging for medical education. However, their finding of decreased learner satisfaction levels in the WBVE group warrants further discussions of the role of avatar agency and immersion degree.

Sripadungkul et al [1] used a desktop-based WBVE on the Spatial platform and noted that the lack of "nonverbal cues" might have led to decreased learner satisfaction levels compared to F2F delivery. We argue that this decreased satisfaction is a limitation of the nonimmersive virtual environment using a 2D desktop interface rather than the virtual environment pedagogy itself. Recent evidence comparing head-mounted display (HMD)–based immersive 3D virtual reality (VR) with traditional F2F instruction demonstrates comparable or even superior results regarding student acceptance in medical education. Mühling et al [2] demonstrated that medical students rated HMD-based VR stations as having realism and clinical relevance equivalent to traditional physical objective structured clinical examination stations, with no significant differences in overall perception. Notably, the students exhibited high acceptance of the VR system, suggesting that active participation from a first-person perspective in an immersive environment positively influenced their engagement. In nursing education, research on novice nurses revealed that immersive VR training was more effective than conventional F2F group discussions in fostering significantly greater and sustained improvements in collaborative attitudes. This effectiveness likely stems from the realistic reproduction of interpersonal environments in an immersive VR space that prevents communication hindrances [3].

Our research on the feasibility of immersive VR-based group discussions among nursing students supports the importance of "avatar agency," allowing users to subjectively and actively control and personalize their digital representations [4]. Over 70% of our study participants reported that personalizing avatar characteristics, including clothing/hairstyles, influenced their discussions, and all participants expressed a high level of intention to continue using VR for learning [4]. Entering an immersive virtual environment using avatars fosters "social presence," allowing users to feel they are in the same space, encouraging interactive communication [5]. These elements minimize the "communicative gap" and could bridge the satisfaction gap identified by Sripadungkul et al [1]. There could also be increased motivation to engage in tasks and establish a virtual environment that serves as a foundation for effective TBL. Therefore, we would like to request further details on the avatar settings and authors' perspectives.

Although the introduction of HMDs is critical to such enhancements, the implementation costs are high; therefore, careful consideration of their impact on scalability in large-scale educational settings is required. WBVE delivery of TBL is a scalable alternative to F2F delivery, integrating more immersive elements and allowing for greater avatar agency, which could improve its learner satisfaction rate. We appreciate the authors' contribution to this evolving field and believe that exploring the interaction between immersion and social presence will help establish optimal practices for future virtual health care education.

XSL•FO
RenderX

Acknowledgments

We thank Editage for the English language editing.

Funding

This work was supported by the Japan Society for the Promotion of Science KAKENHI (grant JP 25K24265).

Conflicts of Interest

None declared.

References

1. Sripadungkul D, Boonmak S, Somjit M, Plailaharn N, Sriraj W, Boonmak P. Web-based virtual environment versus face-to-face delivery for team-based learning of anesthesia techniques among undergraduate medical students: randomized controlled trial. JMIR Med Educ 2026 Jan 15;12:e80097. [doi: 10.2196/80097] [Medline: 41539671]
2. Mühling T, Schreiner V, Appel M, Leutritz T, König S. Comparing virtual reality-based and traditional physical objective structured clinical examination (OSCE) stations for clinical competency assessments: randomized controlled trial. J Med Internet Res 2025 Jan 10;27:e55066. [doi: 10.2196/55066] [Medline: 39793025]
3. Hur HK, Choi JH, Jung JS. Effectiveness of an immersive virtual reality simulation speak-up training program for patient safety in novice nurses: a quasi-experimental study. Healthcare (Basel) 2025 Sep 25;13(19):2435. [doi: 10.3390/healthcare13192435] [Medline: 41095521]
4. Funao H, Shimaoka M, Kako J. Feasibility and safety of virtual reality-based online group discussions among nursing students: a cross-sectional study. Cureus 2025 Jan;17(1):e78089. [doi: 10.7759/cureus.78089] [Medline: 40018457]
5. Teng CI, Dennis AR, Dennis AS. Avatar-mediated communication and social identification. J Manage Inf Syst 2023 Oct 2;40(4):1171-1201. [doi: 10.1080/07421222.2023.2267320]

Abbreviations

F2F: face-to-face
HMD: head-mounted display
TBL: team-based learning
VR: virtual reality
WBVE: web-based virtual environment

Edited by A Stone; submitted 28.Jan.2026; this is a non–peer-reviewed article;accepted 13.Feb.2026; published 30.Mar.2026.

Please cite as:
Funao H, Shimaoka M, Kako J
Enhancing Team-Based Learning in Virtual Environments: The Role of Avatar Agency and Immersive Social Presence
JMIR Med Educ 2026;12:e92366
URL: https://mededu.jmir.org/2026/1/e92366
doi:10.2196/92366

© Hiroki Funao, Motomu Shimaoka, Jun Kako. Originally published in JMIR Medical Education (https://mededu.jmir.org), 30.Mar.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Education, is properly cited. The complete bibliographic information, a link to the original publication on https://mededu.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Authors' Reply: Enhancing Team-Based Learning in Virtual Environments: The Role of Avatar Agency and Immersive Social Presence

Darunee Sripadungkul, MD; Suhattaya Boonmak, MD; Monsicha Somjit, MD; Narin Plailaharn; Wimonrat Sriraj, MD; Polpun Boonmak, MD

Anesthesiology Department, Faculty of Medicine, Khon Kaen University, Ni Muang, Khon Kaen, Thailand

Corresponding Author:
Suhattaya Boonmak, MD
Anesthesiology Department, Faculty of Medicine, Khon Kaen University, Ni Muang, Khon Kaen, Thailand

Related Articles:

https://mededu.jmir.org/2026/1/e80097

https://mededu.jmir.org/2026/1/e92366

(JMIR Med Educ 2026;12:e93060) doi:10.2196/93060

KEYWORDS

anesthesia; computer-assisted instruction; distance; education; internet; learning; medical; problem-based learning; students; teaching; virtual reality

Thank you for the thoughtful comments [1] on our randomized controlled trial comparing web-based virtual environment (WBVE) versus face-to-face delivery for team-based learning (TBL) of anesthesia techniques [2]. We appreciate the emphasis on immersion, avatar agency, and social presence as potential determinants of learner experience in virtual settings.

We agree that reduced social cues are a plausible contributor to the satisfaction gap. In our study, WBVE was delivered through a desktop-based, nonimmersive environment, which inherently limits nonverbal communication . This likely affects perceived interaction quality during TBL and suggests that lower satisfaction may reflect constraints of the interface rather than a limitation of virtual pedagogy itself [2].

Our primary objective was to evaluate a delivery approach that is low-cost, scalable for large cohorts, and logistically feasible for undergraduate medical students. These implementation priorities guided our selection of a desktop-based WBVE as a practical alternative to face-to-face [2]. While head-mounted display–based immersive 3D virtual reality (VR) may increase immersion and avatar agency, it also introduces additional equipment, staffing, and cost demands that can limit scalability [3-5]. Because we did not include a head-mounted display VR arm, we cannot determine whether immersive VR would eliminate the satisfaction difference. Future studies should directly compare desktop WBVE and immersive VR while measuring social presence, usability, cognitive load, feasibility, and cost.

Learner satisfaction in our trial was assessed across four domains: (1) learning topic, (2) learning process (including instructor facilitation and delivery technique), (3) learning outcomes, and (4) overall course satisfaction. Ratings were generally high in both groups, but the largest differences clustered in learning process elements most sensitive to delivery modality on discussion flow, idea sharing, and spontaneous expression. We interpret this pattern as reduced social presence in nonimmersive WBVE [2].

Regarding avatar settings, the WBVE session was delivered on the Spatial platform using a desktop setup. On first log-in, each participant created an account and entered the virtual space as a keyboard-controlled avatar. Teams collaborated synchronously using Spatial's voice and text chat while completing standardized TBL modules. We did not experimentally manipulate avatar agency (eg, avatar customization or embodiment). Therefore, our findings apply to this specific nonimmersive, desktop-based WBVE implementation [2].

In summary, our findings support desktop WBVE as a scalable approach with comparable knowledge gains while highlighting interaction-related limitations that should be targeted in future virtual TBL design and evaluation.

Conflicts of Interest

None declared.

XSL•FO
RenderX

References

1. Funao H, Shimaoka M, Kako J. Enhancing team-based learning in virtual environments: the role of avatar agency and immersive social presence. JMIR Med Educ 2026;12:e92366. [doi: 10.2196/92366]
2. Sripadungkul D, Boonmak S, Somjit M, Plailaharn N, Sriraj W, Boonmak P. Web-based virtual environment versus face-to-face delivery for team-based learning of anesthesia techniques among undergraduate medical students: randomized controlled trial. JMIR Med Educ 2026 Jan 15;12:e80097. [doi: 10.2196/80097] [Medline: 41539671]
3. Mühling T, Schreiner V, Appel M, Leutritz T, König S. Comparing virtual reality-based and traditional physical objective structured clinical examination (OSCE) stations for clinical competency assessments: randomized controlled trial. J Med Internet Res 2025 Jan 10;27:e55066. [doi: 10.2196/55066] [Medline: 39793025]
4. Hur HK, Choi JH, Jung JS. Effectiveness of an immersive virtual reality simulation speak-up training program for patient safety in novice nurses: a quasi-experimental study. Healthcare (Basel) 2025 Sep 25;13(19):2435. [doi: 10.3390/healthcare13192435] [Medline: 41095521]
5. Funao H, Shimaoka M, Kako J. Feasibility and safety of virtual reality-based online group discussions among nursing students: a cross-sectional study. Cureus 2025 Jan;17(1):e78089. [doi: 10.7759/cureus.78089] [Medline: 40018457]

Abbreviations

TBL: team-based learning
VR: virtual reality
WBVE: web-based virtual environment

Edited by A Stone; submitted 07.Feb.2026; this is a non–peer-reviewed article;accepted 13.Feb.2026; published 30.Mar.2026.

Please cite as:
Sripadungkul D, Boonmak S, Somjit M, Plailaharn N, Sriraj W, Boonmak P
Authors' Reply: Enhancing Team-Based Learning in Virtual Environments: The Role of Avatar Agency and Immersive Social Presence
JMIR Med Educ 2026;12:e93060
URL: https://mededu.jmir.org/2026/1/e93060
doi:10.2196/93060

© Darunee Sripadungkul, Suhattaya Boonmak, Monsicha Somjit, Narin Plailaharn, Wimonrat Sriraj, Polpun Boonmak. Originally published in JMIR Medical Education (https://mededu.jmir.org), 30.Mar.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Education, is properly cited. The complete bibliographic information, a link to the original publication on https://mededu.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Why Medical Education Without Artificial Intelligence Still Matters: A Neuroscience-Informed Perspective

Charles Verdonk[1,2,3], MD, PhD

[1]French Armed Forces Biomedical Research Institute, 1 rue du Général Valérie André, Brétigny-sur-Orge, France

[2]UMR VIFASOM, Université Paris Cité, Paris, France

[3]Laureate Institute for Brain Research, Tulsa, OK, United States

Corresponding Author:
Charles Verdonk, MD, PhD
French Armed Forces Biomedical Research Institute, 1 rue du Général Valérie André, Brétigny-sur-Orge, France

Related Articles:

https://mededu.jmir.org/2026/1/e77127

Comment in: https://mededu.jmir.org/2026/1/e95205

(JMIR Med Educ 2026;12:e94594) doi:10.2196/94594

KEYWORDS

artificial intelligence; medical education; cognitive off-loading; operational medicine; decision-making under stress

In their recent viewpoint, Izquierdo-Condoy et al [1] present a comprehensive analysis of the transformative potential, current applications, and future implications of artificial intelligence (AI) in medical education. They advocate for integrating AI literacy into medical curricula as health care systems become increasingly AI driven. I fully endorse this forward-looking perspective.

However, I would like to extend the discussion by highlighting a complementary yet underexplored issue: the implications of clinical practice in contexts where AI support is unavailable, unreliable, or deliberately restricted. While Izquierdo-Condoy et al [1] analyze the risks associated with integrating AI into medical curricula, less attention is given to the consequences of AI absence following training in AI-rich environments. The authors acknowledge that technological disparities may hinder equitable AI adoption [1], implying that AI-supported care cannot be universally guaranteed.

Emerging empirical evidence suggests that the removal of AI assistance after habitual use may adversely affect cognitive engagement and task performance. In an electroencephalography study, Kosmyna et al [2] reported that reliance on generative AI during essay writing was associated with reduced alpha- and beta-band functional connectivity—interpreted as diminished distributed network engagement—compared with unaided writing. Participants who initially relied on generative AI and were subsequently required to write without AI assistance exhibited persistently reduced connectivity and lower cognitive engagement, alongside poorer memory recall [2]. Behavioral performance decrements following AI withdrawal have also been documented in technical medical tasks, such as digestive endoscopy [3]. More broadly, the medical education literature cautions that when AI substitutes for clinical reasoning (cognitive off-loading) rather than augments it, risks to skill acquisition and retention may emerge [4]. Together, these findings suggest that sustained AI-mediated training may affect broader mechanisms of skill development, with vulnerabilities becoming evident in AI-absent contexts. Although preliminary, this evidence raises the possibility that prolonged AI reliance during training may reshape neurocognitive engagement in ways that have unintended consequences when independent performance is required.

These considerations are particularly salient in high-stakes, resource-constrained environments such as remote or isolated practices, military operations, disaster and humanitarian responses, and other extreme operational settings, where health care professionals' cognitive performance is particularly strained by stress. In such contexts, the absence of AI support, combined with potential AI-mediated disruption of skill development during training, could compound vulnerability precisely when independent cognitive performance, such as clinical reasoning, is most critical.

This argument does not oppose AI integration. AI will likely enhance efficiency across many health care settings. However, ensuring safe and equitable care may require preserving AI-independent competence. Drawing inspiration from high-reliability industries such as aviation—where automation failure scenarios and minimum unaided practice are standard—medical education could incorporate structured AI-withdrawal exercises and defined thresholds of unaided proficiency [5]. In parallel, longitudinal research is needed to determine whether AI-mediated training produces durable changes in brain function and clinical performance.

XSL•FO
RenderX

As medical education accelerates toward AI integration, preparing physicians to practice both with and without AI support may represent a safeguard for patient safety and professional autonomy.

Acknowledgments

Portions of this manuscript were edited with assistance from OpenAI's GPT-5.2 model. The tool was used exclusively to improve clarity, language, and organization, and to align the text with journal-specific formatting and style requirements. No artificial intelligence system was used to generate original scientific ideas, data interpretation, or conclusions. All intellectual content, analyses, and arguments were developed and verified solely by the author, who assumes full responsibility for the integrity, accuracy, and originality of the manuscript.

Funding

The author declared that no financial support was received for this work.

Disclaimer

The opinions and assertions expressed herein are those of the author and do not necessarily reflect the official views of the institutions with which the author is affiliated.

Conflicts of Interest

None declared.

References

1. Izquierdo-Condoy JS, Arias-Intriago M, Montero Corrales L, Ortiz-Prado E. Artificial intelligence in medical education: transformative potential, current applications, and future implications. JMIR Med Educ 2026 Feb 17;12(1):e77127. [doi: 10.2196/77127] [Medline: 41701936]
2. Kosmyna N, Hauptmann E, Yuan YT, et al. Your brain on ChatGPT: accumulation of cognitive debt when using an AI assistant for essay writing task. arXiv. Preprint posted online on Jun 10, 2025. [doi: 10.48550/arXiv.2506.08872]
3. Ho JCL, Qian Z, Lau LHS, Yip HC, Chiu PWY. Artificial intelligence in digestive endoscopy training-the past, present, and future. Dig Endosc 2026 Jan;38(1):e70047. [doi: 10.1111/den.70047] [Medline: 41139585]
4. Abdulnour REE, Gin B, Boscardin CK. Educational strategies for clinical supervision of artificial intelligence use. N Engl J Med 2025 Aug 21;393(8):786-797. [doi: 10.1056/NEJMra2503232] [Medline: 40834302]
5. Ong AY, Merle DA, Pollreisz A, et al. Flight rules for clinical AI: lessons from aviation for human-AI collaboration in medicine. NPJ Digit Med 2026 Jan 31;9(1):201. [doi: 10.1038/s41746-026-02410-1] [Medline: 41620563]

Abbreviations

AI: artificial intelligence

Edited by T Leung; submitted 03.Mar.2026; this is a non–peer-reviewed article;accepted 14.Mar.2026; published 31.Mar.2026.

Please cite as:
Verdonk C
Why Medical Education Without Artificial Intelligence Still Matters: A Neuroscience-Informed Perspective
JMIR Med Educ 2026;12:e94594
URL: https://mededu.jmir.org/2026/1/e94594
doi:10.2196/94594

© Charles Verdonk. Originally published in JMIR Medical Education (https://mededu.jmir.org), 31.Mar.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Education, is properly cited. The complete bibliographic information, a link to the original publication on https://mededu.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Authors' Reply: Why Medical Education Without Artificial Intelligence Still Matters: A Neuroscience-Informed Perspective

Juan S Izquierdo-Condoy[1], MSc, MD; Marlon Arias-Intriago[1], MD; Laura Montero Corrales[2], PhD; Esteban Ortiz-Prado[1], MSc, MPH, MD, PhD

[1]One Health Research Group, Universidad de Las Américas, Via Nayon S/N, Quito, Pichincha, Ecuador

[2]Escuela de Comunicación, Latin University of Costa Rica, San José, Costa Rica

Corresponding Author:
Esteban Ortiz-Prado, MSc, MPH, MD, PhD
One Health Research Group, Universidad de Las Américas, Via Nayon S/N, Quito, Pichincha, Ecuador

Related Articles:

https://mededu.jmir.org/2026/1/e77127

https://mededu.jmir.org/2026/1/e94594

(JMIR Med Educ 2026;12:e95205) doi:10.2196/95205

KEYWORDS

artificial intelligence; medical education; machine learning; adaptive learning systems; future implications

We thank the author of the pertinent commentary, "Why Medical Education Without Artificial Intelligence Still Matters: A Neuroscience-Informed Perspective" [1]. We believe this letter makes a valuable contribution to the debate on artificial intelligence (AI) in medical education by emphasizing the cognitive consequences of training clinicians in environments where these tools are available yet not always reliable or accessible in everyday practice.

From our perspective, the author's argument expands on a concern we had already raised previously: if AI, particularly generative AI, can compromise critical thinking and cognitive autonomy, this erosion may also translate into a reduced capacity to sustain clinical reasoning when the tool fails or is unavailable [2]. In our article, "Artificial Intelligence in Medical Education: Transformative Potential, Current Applications, and Future Implications" [3], we identified uncritical dependence on algorithmic outputs as one of the main limitations of incorporating AI into medical training.

The neurocognitive evidence cited by the author reinforces this argument. Recent studies suggest that the use of generative AI may alter how individuals engage in complex tasks. Although this evidence remains preliminary, it is consistent with what has already been described regarding automation bias and cognitive off-loading. Likewise, accepting automated suggestions without sufficient analysis may increase the risk of error when the system fails or is unavailable [4,5]. Taken together, these findings support a central idea: AI should support reasoning, not replace expert judgment.

In our viewpoint, we addressed these concerns through concrete proposals. Importantly, our discussion was not limited to generative AI but also encompassed other educationally relevant technologies, including natural language processing, intelligent tutoring systems, virtual reality, and augmented reality. In this context, we advocated for AI literacy curricula that would enable students not only to use these tools but also to critically evaluate them, recognize their limitations, identify biases, and assess their outputs against sound clinical reasoning. We also supported governance structures aligned with international ethical frameworks, such as the UNESCO (United Nations Educational, Scientific and Cultural Organization) Recommendation on the Ethics of Artificial Intelligence and the World Health Organization guidance on the ethics and governance of AI in health [3]. We consider human oversight, transparency, bias monitoring, and accountability to be essential safeguards for the responsible integration of AI.

We also agree with the author's suggestion to include structured exercises in which students must function without AI support, as this may strengthen independent diagnostic reasoning and decision-making under uncertainty. At the same time, this discussion must acknowledge persistent technological inequalities, particularly in low-income countries, where access to advanced AI tools remains uneven [6].

We agree that AI integration in medical education should not displace human intelligence but rather enhance professional competence without undermining clinical autonomy or critical thinking. Preparing future physicians to function appropriately with or without AI will be essential to protect patient safety, respond across diverse contexts, and preserve the human dimension of medicine. We again thank the author for this valuable contribution and concur on the need for further empirical evidence, particularly longitudinal studies, to guide a prudent, ethical, and evidence-based integration of AI into medical education.

XSL•FO
RenderX

Acknowledgments

Language editing assistance was used during the preparation of this manuscript. A generative artificial intelligence (AI) tool (ChatGPT, OpenAI, GPT-5.4 Thinking) was consulted solely to improve grammar, clarity, and sentence structure, as English is not the authors' first language. The tool was used exclusively for linguistic refinement and formatting support. All scientific concepts, interpretations, arguments, and conclusions presented in this manuscript were independently developed by the authors. The authors take full responsibility for the intellectual content, accuracy, and integrity of the work.

Funding

The authors declared that no financial support was received for this work.

Conflicts of Interest

None declared.

References

1. Verdonk C. Why medical education without artificial intelligence still matters: a neuroscience-informed perspective. JMIR Med Educ 2026;12:e94594. [doi: 10.2196/94594]
2. Izquierdo-Condoy JS, Arias-Intriago M, Tello-De-la-Torre A, Busch F, Ortiz-Prado E. Generative artificial intelligence in medical education: enhancing critical thinking or undermining cognitive autonomy? J Med Internet Res 2025 Nov 3;27:e76340. [doi: 10.2196/76340] [Medline: 41183320]
3. Izquierdo-Condoy JS, Arias-Intriago M, Montero Corrales L, Ortiz-Prado E. Artificial intelligence in medical education: transformative potential, current applications, and future implications. JMIR Med Educ 2026 Feb 17;12:e77127. [doi: 10.2196/77127] [Medline: 41701936]
4. Abdelwanis M, Alarafati HK, Tammam MMS, Simsekler MCE. Exploring the risks of automation bias in healthcare artificial intelligence applications: a Bowtie analysis. J Safety Sci Resilience 2024 Dec;5(4):460-469. [doi: 10.1016/j.jnlssr.2024.06.001]
5. Kosmyna N, Hauptmann E, Yuan YT, et al. Your Brain on ChatGPT: accumulation of cognitive debt when using an AI assistant for essay writing task. arXiv 2025 Jun 10. [doi: 10.48550/arXiv.2506.08872]
6. Izquierdo-Condoy JS, Arias-Intriago M, Nati-Castillo HA, et al. Exploring smartphone use and its applicability in academic training of medical students in Latin America: a multicenter cross-sectional study. BMC Med Educ 2024 Nov 30;24(1):1401. [doi: 10.1186/s12909-024-06334-w] [Medline: 39616324]

Abbreviations

AI: artificial intelligence
UNESCO: United Nations Educational, Scientific and Cultural Organization

Edited by T Leung; submitted 12.Mar.2026; this is a non–peer-reviewed article;accepted 14.Mar.2026; published 31.Mar.2026.

Please cite as:
Izquierdo-Condoy JS, Arias-Intriago M, Montero Corrales L, Ortiz-Prado E
Authors' Reply: Why Medical Education Without Artificial Intelligence Still Matters: A Neuroscience-Informed Perspective
JMIR Med Educ 2026;12:e95205
URL: https://mededu.jmir.org/2026/1/e95205
doi:10.2196/95205

© Juan S Izquierdo-Condoy, Marlon Arias-Intriago, Laura Montero Corrales, Esteban Ortiz-Prado. Originally published in JMIR Medical Education (https://mededu.jmir.org), 31.Mar.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Education, is properly cited. The complete bibliographic information, a link to the original publication on https://mededu.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Medical Students' Reflections on Transitioning to Their First General Practice Placement: Qualitative Descriptive Study

Sara Bashar Qasrawi[1]; Maryam Tomerak[1*]; Shahad Abdulkhaleq Mamalchi[2*], MD; Fatima Atieh[1]; Hasan AlAdraj[3], MB, BCh, BAO; Mohamed Abdulla[4], MB, BCh, BAO; Salim Fredericks[1], PhD; Ghufran Jassim[5], MBBS, MSc, MD, PhD; Hani Malik[6]; Eric Clarke[7], MSc; Denis Harkin[7], MB, MD; Shaista Salman Guraya[8], PhD

[1]School of Medicine, Royal College of Surgeons in Ireland - Medical University of Bahrain, Busaiteen, Bahrain

[2]Department of Surgery, University of Toronto, Toronto, ON, Canada

[3]Cardiff and Vale University Health Board, NHS, Cardiff, United Kingdom

[4]Salmaniya Medical Complex, Manama, Bahrain

[5]Family Medicine Department, School of Medicine, Royal College of Surgeons in Ireland (RCSI)-Medical University of Bahrain, Busaiteen, Bahrain

[6]Royal Bahrain Hospital, Manama, Bahrain

[7]Faculty of Medicine and Health Sciences, Royal College of Surgeons in Ireland, Dublin, Ireland

[8]Institute of Learning, Mohammed Bin Rashid University of Medicine and Health Sciences, Dubai Health, Al Razi St - Umm Hurair 2 - Dubai Healthcare City, Dubai, United Arab Emirates

[*]these authors contributed equally

Corresponding Author:
Shaista Salman Guraya, PhD
Institute of Learning, Mohammed Bin Rashid University of Medicine and Health Sciences, Dubai Health, Al Razi St - Umm Hurair 2 - Dubai Healthcare City, Dubai, United Arab Emirates

Abstract

Background: Transitioning from preclinical to clinical training is a critical milestone of “becoming and being” in a medical student’s journey. Despite simulation-based learning, real-world clinical exposure remains indispensable in shaping professional identity. The clinical learning environment is a complex interplay of social, cultural, and organizational factors that influence students’ development as future health care professionals.

Objective: This study explores medical students’ reflections on their first clinical placement in general practice, aiming to understand their experiences, challenges, and the clinical learning environment’s role in their learning and developing professional identity formation as a step toward establishing a conceptual framework and a common language for educators, which we hope will promote further advances to support beneficial professional identity formation.

Methods: We analyzed reflections from fourth-year medical students following their initial general practice placement. A qualitative descriptive approach grounded in naturalism was employed to explain our participants’ transitioning encounters in clear, everyday language to ensure their experiences were presented in their own words, without bias. Content thematic analysis was conducted to identify key themes related to their experiences.

Results: Students’ reflections revealed a startled cohort unprepared for the epistemological, emotional, and practical realities of clinical work. Many assumed that classroom “knowing” would seamlessly translate into clinical “doing” but were met instead with uncertainty, failure, and emotional overwhelm, often manifesting as shame, guilt, and withdrawal. These experiences illuminated a fracture between knowledge and knowing, underscored by students’ prereflective epistemological beliefs and varying degrees of supervisory preparedness. While emotional and cognitive struggles were widespread, rare instances of supportive mentorship and feedback significantly bolstered students’ confidence and participation.

Conclusions: Reflection offered a valuable window into how students think, feel, and act during this critical transition, but it is not a cure-all. Reflection should be positioned as a developmental and communal tool within a broader scaffolding that includes early skills preparation, role clarity, psychological safety, and trained supervisors. Structured shared reflective circles and narrative listening sessions can help normalize uncertainty, support identity formation, and foster resilience during students’ entry into clinical practice and can ease the journey of “becoming and being.”

(JMIR Med Educ 2026;12:e76599) doi:10.2196/76599

XSL•FO
RenderX

KEYWORDS

medical education; professional identity formation; reflective writing; clinical learning environment; qualitative descriptive methodology

Introduction

Despite advances in simulation-based learning, the clinical learning environment (CLE) remains irreplaceable in health care education. Its unpredictable, ethically complex, and emotionally charged nature uniquely shapes how medical students think, feel, and act [1] as future health care professionals. As Nordquist et al [2] describe, the CLE is a dynamic interplay of social, cultural, organizational, and material factors in which students learn while working. Within this setting, students engage in situated learning [3], navigating the multifaceted realities of becoming a physician [4]. An effective CLE therefore influences learners' well-being, engagement, and professional socialization [2].

The theory of legitimate peripheral participation (LPP) by Lave and Wenger [5] helps us understand how novices begin on the periphery of practice and, through increasingly authentic participation, move toward fuller membership in the medical community. These early encounters profoundly influence students' confidence, role interpretation, and internalization of professional norms [6]. They integrate the values, attitudes, and ethical responsibilities of the profession into their evolving sense of self as they move from peripheral to central participation. Yet this movement is rarely linear; it is marked by dissonance, confusion, and emotional complexity requiring thoughtful processing. Psychological approaches [7] emphasize harnessing these moments through structured reflections, which can help them make sense of their experiences by charting their path within the profession.

Reflective writing has become a key pedagogical tool for helping learners articulate and process complex experiences [8]. By cultivating metacognitive and meta-affective capacities, reflective writing enables a personal voice infused with emotions, which helps individuals identify and rationalize their developing sense of self. Wald [7] used the metaphorical notion of a *resiliency workout* for guided and structured reflective writing, enabling medical students to regulate emotions, process feedback, and remain empathetically engaged with patients and colleagues. Primarily, in the early stages of clinical exposure, reflective writing can provide a structured pedagogical space for students to pause, process, and internalize the meaning of their experiences, providing them with a feeling of not only clinical competence but also practical wisdom, ethical clarity, and emotional resilience [9]. While reflection has long been recognized as essential to professional development, its theoretical underpinnings, particularly those rooted in epistemologies of practice and critical social inquiry, remain underexplored in medical education [10]. We echo recent calls for continued theorization of reflection, not only as a tool for self-regulation but also as a means to navigate uncertainty, challenge dominant discourses, and construct knowledge within complex clinical realities [11].

To illustrate this, examining how medical students themselves narrate and reflect on their early clinical experiences through the lens of LPP contributes meaningfully to ongoing scholarship of "becoming" and "being"—professional identity formation (PIF). A more descriptive-interpretive understanding allows medical educators to identify opportunities to support students' PIF and facilitate their transition from preclinical learning to the CLE. In this research, we aimed to explore how medical students perceive and experience LPP in the CLE and how it influences their developing professional identity, as a step toward establishing a conceptual framework and a common language for educators, which we hope will promote further advances to support beneficial PIF.

Methods

Study Design

We employed a qualitative descriptive (QD) approach to address our research question. As described by Sandelowski [12], QD is grounded in naturalism, requiring researchers to stay close to the data and portray participants' perceptions and experiences in their own words in an unadorned manner without introducing bias. The QD approach helped us explain the transitioning experiences of our participants using straightforward language. During data analysis, all involved researchers established their constructionist paradigmatic stance and acknowledged their sensitivities and inclinations. The research team engaged in reflexive discussions to remain aware of their own assumptions and positionalities, ensuring that interpretations remain grounded in participants' narratives. The senior author, SSG, is an expert in medical education and medical professionalism, with higher degree qualifications and research publications. Coauthors SAM, HA, and MA were medical students and classmates of the participants. SBQ, MT, and FA were junior medical students at the Royal College of Surgeons in Ireland – Medical University of Bahrain (RCSI-MUB).

Participant Recruitment

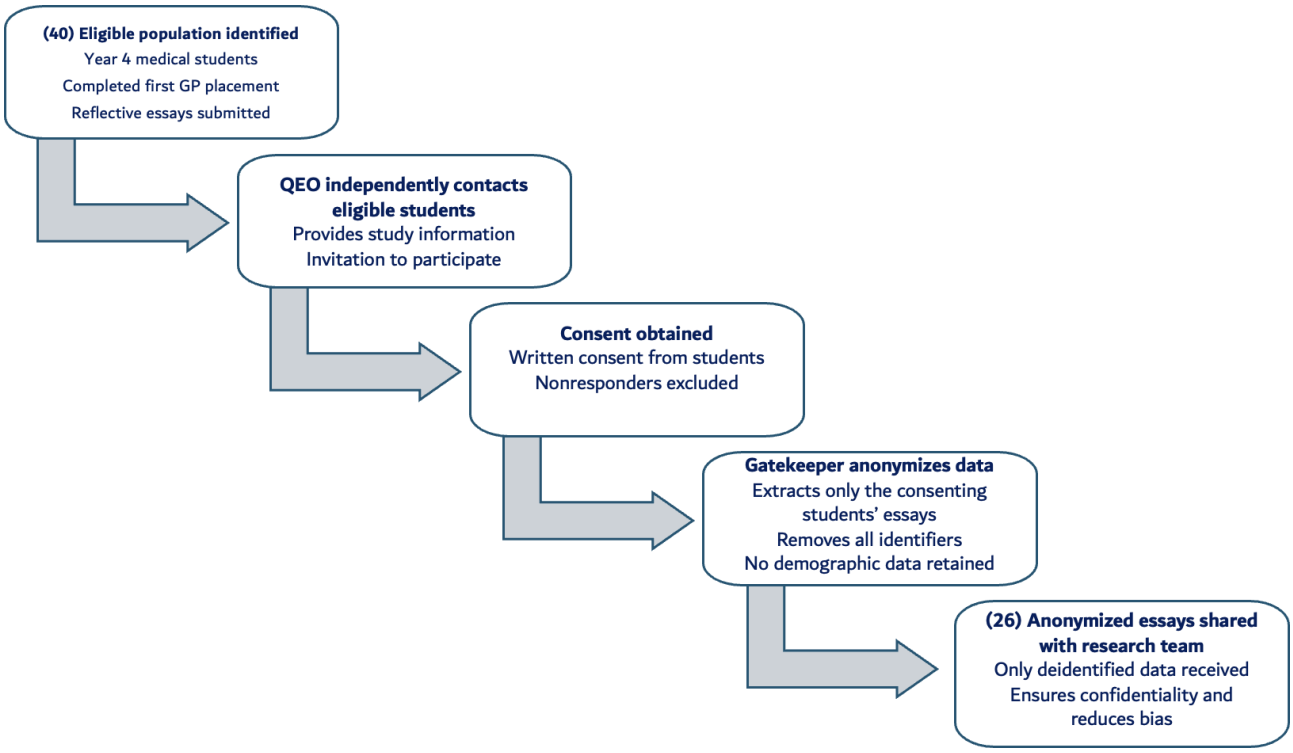
Naturalistic inquiry nudges researchers to ensure that the collected data are good enough to enable participants to express their experiences authentically and without external influence. Hence, we analyzed reflective essay accounts written by a cohort of fourth-year medical students during their first general practice (GP) placement to document the transition phase from preclinical learning to the CLE. As part of the assessment, these structured essays comprised 500‐700 words and were submitted using a "Plagiarism Checker," Turnitin. Students were instructed to reflect according to the 6 stages (description, feelings, evaluation, analysis, conclusion, and action plan) of Gibbs' reflective cycle [13]. We prompted students to document specific learning moments and challenges encountered during their first clinical placement. "Challenging or difficult" clinical encounters were considered to highlight the choices and narratives of such critical incidents, which reflect medical

XSL•FO
RenderX

students' values and attitudes in the context of professional experiences [14].

The study aimed to explore how students began visualizing themselves as part of the clinical team, capturing the transformations that occurred during this pivotal stage of their PIF. To answer our study's objectives, we selected students who were attending their first clinical GP placement to ensure they had no significant prior exposure to clinical practice. To provide context to our readers, it is worth mentioning a brief overview of the RCSI-MUB curriculum. The first two and a half preclinical years are entirely theoretical and focus on basic science instruction. Clinical exposure begins in the second semester of year 3, and the first GP rotation occurs in year 4 (Figure 1).

Figure 1. A flowchart outlining the recruitment and data management procedures for the study. GP: general practice; QEO: Quality Enhancement Office.



Ethical Considerations

Ethical approval was obtained from the RCSI-MUB Research Ethics Committee (REC/2023/153). Participant recruitment followed the flowchart process: the Quality Enhancement Office contacted a cohort of year 4 students who had completed their first GP placement and provided them with a study information sheet outlining the purpose of this study and the voluntary nature of participation. Written informed consent was obtained prior to inclusion. Participation or refusal had no academic or professional consequences, and no financial or academic compensation was provided. Only anonymized reflective essays from consenting students were shared with the research team. The gatekeeper removed all identifying and demographic information before sharing the essays, and the investigators had no access to participant identifiers. Data were stored securely on institutional systems accessible only to the research team.

Context of Study

Students, before starting their GP placements, receive little advance notice that the CLE will challenge their assumptions. While they are introduced to the educational aspects of the program, including learning outcomes and assessment requirements, they receive minimal orientation to the realities of clinical practice. Clinical skills are introduced in the second half of year 3 in simulation and class-based settings, without any supervised or structured opportunities to practice on peers or real patients. Students are not adequately prepared for the structural, social, cultural, and practical complexities of patient interactions. Although elements of professionalism such as cultural competence, confidentiality, and informed consent are discussed, these remain largely theoretical and disconnected from clinical context.

Clinical placements occur at varied sites under the supervision of adjunct clinical faculty with differing levels of teaching expertise, a dilemma familiar to many medical schools without their own clinical facilities. GP tutors are expected to facilitate learning toward defined outcomes but receive no formal preparation in creating psychological safety, dealing with uncertainties, supervising students, debriefing, or providing structured feedback. As a result, much of their teaching relies on personal experience rather than standardized educational practice. Consequently, there are varying degrees of preparedness and engagement on both sides—students and tutors alike.

Data Collection and Analysis

As common in QD, we used content thematic analysis [15,16] to identify and describe core themes across all data. First, SSG, SBQ, MT, SAM, and FA independently analyzed the first 6 essays to identify initial codes. Next, SAM, SBQ, MT, and FA

XSL•FO
RenderX

consolidated the initial codes into subthemes and preliminary themes. All codes, coding schemas, notes, and memos were captured on Microsoft Excel, and version control was maintained throughout to ensure a transparent process. Later, SSG examined the worked Microsoft Excel spreadsheet and preliminary results and data with a naturalistic inquiry paradigm and envisioned that transitioning moments are marked with significant changes in thinking, feeling, and enacting like a certain individual. Using the flexibility of QD methodology to describe and interpret data based on the study's purpose, we finalized interpretive themes, which were then applied to the remaining essays. Finally, the results were vetted by SAM, HA, and MA to ensure accurate interpretation, and the broader team reviewed the findings for further discussion.

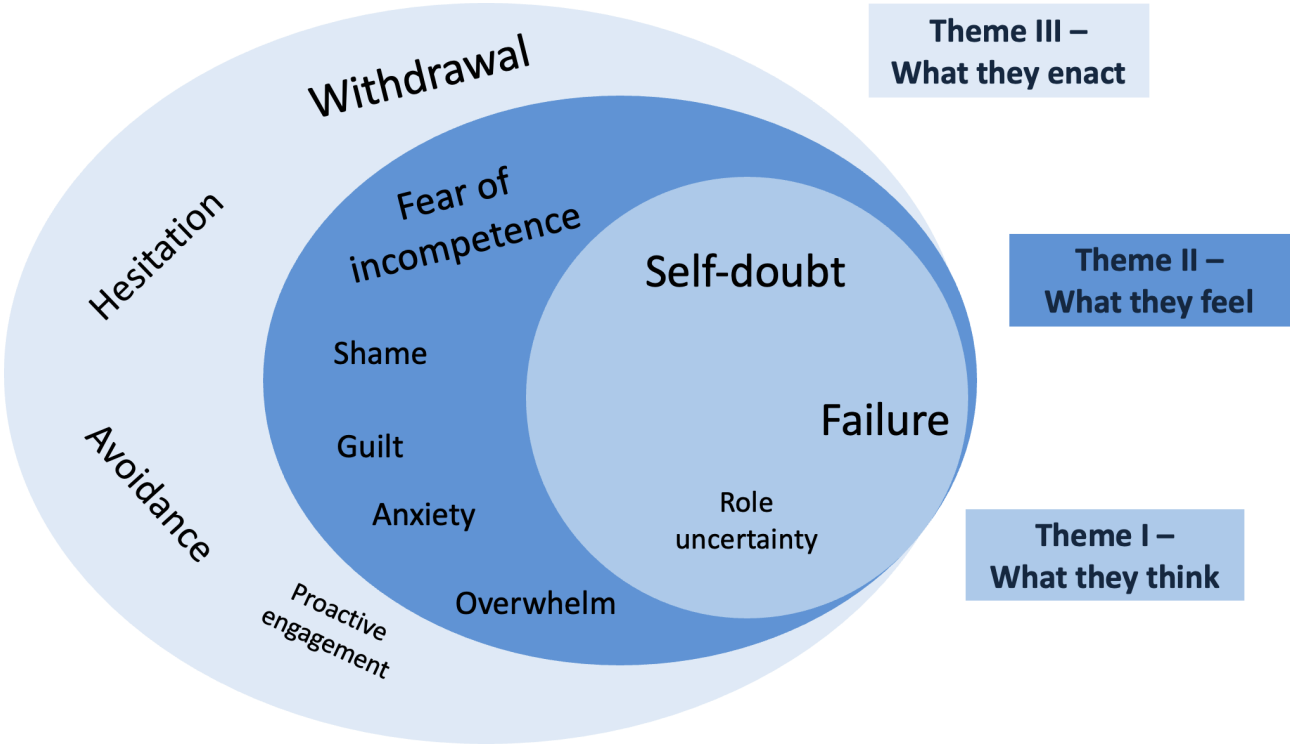
Results

Overview of Findings

Through QD methodology, we adopted a factist perspective for the content thematic analysis [15,16] of the 26 reflective essays. We produced an unadorned description and comprehensive summary of the phenomenon of interest using low inferential stance of participants' language. This analysis revealed how year 4 medical students think, feel, and act during their first clinical placement in GP. Guided by Lave and Wenger's [5] concept of LPP, these themes illustrate how students negotiate their identities at the margins of practice communities and gradually move toward fuller participation. By remaining close to the data and providing a thick description of PIF as an abstract phenomenon, we present our findings as 3 interconnected dimensions of PIF—thinking, feeling, and enacting—which capture the shifts students undergo as they move from peripheral to core participation in clinical practice. The innermost circle represents the internal cognition beginning with self-doubt, expanding into failure and role uncertainty, whereas the middle circle yields the emotional processes students described, ranging from fear of incompetence, guilt, shame, anxiety, and overwhelm. These internal experiences influenced the outward behavioral responses, which are depicted as the outer layer, ranging from withdrawal and avoidance to hesitation and rarely a proactive engagement. The size of the font reflects the frequency of the subtheme expressed across the reflective narratives. These interconnected themes share several common elements, which are illustrated in Figure 2.

Students' reflections illustrated how cognitive, emotional, and behavioral responses interacted dynamically during challenging encounters, shaping how they made sense of their early clinical experiences. Next, we provide a detailed account of the 3 identified themes supported by the verbatim excerpts from the reflective essays.

Figure 2. Three interconnected themes: thinking, feeling, and enacting. These themes illustrate the dynamic progression of professional identity formation. Subthemes are embedded within and sized proportionally to reflect their relative frequency and prominence in the data.



Theme I: What They Think

Thinking involved students making sense of their clinical experiences by reflecting on their existing medical knowledge and its application in actual practice and ethical decision-making, as well as the expectations placed on them in their clinical placements and their role within the wider health care team. These cognitive processes appeared as failure, self-doubt, and role uncertainty, characterized by a lack of more profound awareness of the complexities of patient care.

Students taking patient history, clerking, and performing technical skills on actual patients realized the disconnected realities of classroom versus real-world practice.

XSL•FO
RenderX

> *a person's illness experience is deeply intertwined with his/her thoughts; with seemingly minor procedures like blood withdrawal being capable of causing significant distress when a person is uncomfortable about them.*

A student mentioned that performing phlebotomy for the first time on an actual patient seemed much more complex and challenging despite attempting this procedure in vitro (simulated environments), which led to a sense of failure.

> *This patient was obese and I neither saw nor palpated the vein. I was hesitant, removed my gloves, and put the equipment away. I kept overthinking and doubting my abilities. Because of that, I lost confidence in doing other venepunctures*

Similarly, one of the students mentioned that the cultural sensitivity taught during case-scenario analysis and express-team-based learning sessions [17] of personal and PIF did not seem to translate into the correct practices due to the untamed beast of unconscious bias, which hinders effective communication. Stereotyping appeared to be the most common flaw because of patients' physical attire and appearances.

> *I was wrong about this patient, but worse, I used a stereotype and approached the consult with prejudice. I latched onto a stereotype and minimized her concern so quickly. I used an anchoring bias to make my judgment, which was rooted more in prejudice and arrogance than evidence- based medicine.*

Students from an external or foreign culture in the clinical setting especially confessed to jumping to conclusions due to false perceptions of gender-based restrictions, language barriers, and conservative cultural norms in the contextual clinical settings.

> *I must reassess my attitudes towards discussing sex and identify unrecognized barriers hindering me from effective communication in this aspect. Becoming more confident in my knowledge on the subject (culture) would assist me in navigating future situation. I recognized the need for diligent self-awareness of my implicit biases and in developing my cultural humility.*

Another thought process worth highlighting is the lack of clear expectations of roles and responsibilities in terms of to what extent they can have thoughtful discussions with the patients, disclosure about their prognosis and treatment, and unfamiliarity with the scenarios surrounding confidentiality, informed consent, and patient autonomy. This cognitive dissonance often triggered cognitive conflict that can perpetuate uncertainty precipitating self-doubt and hesitation.

> *I was uncomfortable and unsure of what to say in the absence of my GP Tutor. I was afraid of saying the wrong thing and falsely reassuring the patient. Being young and unmarried I felt in-equipped to navigate such a complex situation.*

In the next theme, we will see how these cognitive realizations actually translate into "feeling or emotive" aspect of PIF.

Theme II: What They Feel

Feelings are the emotional perceptions of the thought processes that we described in our first theme, triggered by uncertainty, challenges, and expectations about students' roles and a fractured translation of knowledge to practice. Fear, guilt, shame, anxiety, and sense of overwhelm were the four dominating feelings we encountered in those 26 reflective essays.

Fear was the most encountered feeling, and it appeared in many varied situations. Students feared "harming the patient," "making mistakes," "challenging hierarchical decisions," and "false reassurances to the patient." These fears of inadequacy reflect a significant mental toll, which can hamper the behavioral capacity of students, either leading to avoidance, hesitancy, and withdrawal from such crucial learning opportunities. We have supported these findings by the following excerpts, respectively.

> *I became scared when I struggled to remove the first tube, I was afraid that I would accidentally move the needle too deep and puncture the vein. When the lab technician realized my needle had come out, I felt guilty that the patient would have to be punctured again on the other side because my job was not good enough.*

Self-doubt and making mistakes stemmed from feelings of inadequacy when thrust into unfamiliar scenarios without structured preparation. Making students fear the subpar consequences in unfamiliar situations like taking a history from a pediatric patient.

> *I thought that I would be wasting the parents' time if I didn't get straight to the point, so I barely gave myself a chance to compose myself. I ended up rushing the encounter to keep up with my racing thoughts. During that time, I was constantly worrying about whether I was making mistakes or embarrassing myself. I unintentionally overlooked fundamentals of communication skills like eye contact as a consequence. Unfortunately, I experienced a feeling of inadequacy as a result of the encounter and my lack of knowledge in that area.*

However, there were few voices who felt that to overcome fear, they must try.

> *I believe now that to overcome a fear, one must jump into the deep end. I find that through this experience, I learned that it is perfectly fine to take a step back and ask for help when necessary.*

The second common emotion is "guilt and shame" when students could not meet self-imposed and patients' expectations of professional standards while doing a simple procedure of phlebotomy.

> *I felt ashamed that I was not able to perform under pressure, and I was embarrassed that this happened in front of the patient.*

Another student's narration of a "well-child visit" encounter depicts the moral trauma felt by the student, which was shocking to begin with, but then led to an extreme feeling of shame, where

XSL•FO
RenderX

students are not equipped to correct seniors or raise concerns in a hierarchical setting of medical practice.

> *I think I was in shock. I had just witnessed a child's legs forced open, by 3 adults, without explanation, permission, or communication. I had just witnessed a child scream and kick in fear and lack of understanding as adults she trusted forced her legs open. There must be a better way. I feel I participated in a traumatic event for the child, one that might stay with her unconscious, if not a developed memory. I feel at a loss on how to discuss this with the doctor. I feel (ac)complic(e) in causing unnecessary trauma to a child. There needs to be a better way, a more human way.*

Guilt emerged from the imposter syndrome, where students felt incompetent, "unable to perform a given task," and felt responsible for distressing the patient.

> *I felt guilty because I stereotyped and minimized her concern. I don't like to think I'm prejudice, but I was prejudiced toward her. I boxed her into a stereotype that may not even be real.*

Surprisingly, guilt seemed to prompt students to take self-assessment and a desire to improve and learn better. On the contrary, shame drifted them toward avoidance and hesitancy. And this feeling of guilt was perpetuated when the health care staff showed their displeasure, even implicitly.

> *I froze but the phlebotomist handled it and she expressed her anger with her body language, after a while I was standing in the booth and a colleague of hers came in and they spoke in their native language which I understood but she didn't know that I could understand them, she told her colleague that I was useless and I was going to be a failure because I couldn't perform a simple phlebotomy, she said a lot of hurtful things about me but I said nothing and stood there and once I got home I started doubting myself and cried about it because I thought I was worthless and not a good medical student and will be a failure.*

Next, the feeling of anxiety and overwhelm was mainly tied to the procedural skills and communication elements where students required independent decision-making and interpretation. One student described paralyzing anxiety while measuring blood pressure incorrectly, which led to avoidance of the procedure afterwards.

> *Failing to measure patient's BP correctly resulted in range of negative emotions and in confidence which led me to avoid measuring the BP of other patients rather than accepting my mistake and work on improving my skill.*

Another student felt extremely overwhelmed by the chaos of medical practice—the noise, busy scenes, medical equipment, and the emotions of worried families. The distress of patient relatives spread like an emotional contagion, affecting the student's mental and emotional well-being.

> *I experienced an unforgettable moment one Thursday evening when I heard screams coming from the corridors of the emergency department. I immediately ran out to investigate and found a chaotic scene with a mother carrying a bleeding 4-year-old boy, accompanied by around 15 friends who were all crying and screaming. I was scared, anxious, and worried about the child's condition. I was also overwhelmed by the chaos and noise in the health care center.*

In the next theme, we will see how these emotions translate into the affective component of "enacting" and tell which direction these narratives took in terms of remaining stuck to LLP or progressing toward the core aspect of communities of practice.

Theme III: What They Enact

Enacting is the translation of thoughts and emotions into observable behaviors, which ranged from being proactive to complete avoidance. Considering these reflective essays provide a snapshot of students' clinical experiences, this theme shows a mix of *hesitation*, *participation*, and *some moments of growth*.

Patient care served as a catalyst to reveal the subtleties and nuances of students' actions. The majority of the voices reflected low efficacy in the form of underconfidence in performing technical skills, which led them to hesitation and withdrawal. Few voices were heard who felt confident and approached the clinical behaviors effectively. Softer skills like communication, rapport building, and gaining patients' trust were the trickiest.

A prominent voice of incompetence emerged in dealing with ethical challenges. Maintaining probity and truthfulness in the middle of practicing and learning skills seemed the most challenging.

> *As the nurse and I walked around the station, we came across a patient who needed an injection. I agreed to the nurse when he asked if I wanted to give an IM injection. However, before I could ask the patient for consent to administer the injection, the nurse informed him that I was a doctor and that I would be administering the injection. This surprised me because obtaining consent for any medical procedure was standard practice. I tried to tell the patient, but I couldn't, so I gave the injection.*

Students learned that effective communication was key to patient care, especially in sensitive situations where they struggled with breaking bad news, realizing later that staying composed and exhibiting just the right amount of empathy was as important as delivering medical facts.

> *I need to learn how to communicate bad news in a sensitive and empathetic manner without inappropriate emotional display and how to provide support to patients during emotionally difficult times.*

Nearly half of the essays showed a similar trend toward technical skills, with only one-third able to perform these skills competently. And this lack of self-efficacy was also attributed to patients' perceptions and times of COVID-19.

> *I really wanted to practice but most patients didn't give me the chance to practice on them and they were underestimating the new generation of medical*

XSL•FO
RenderX

> *students as they believe that we are not good enough. Some of them gave negative comments, they said that they will never trust covid medical students batch, as we were in a blinded learning for almost 2 years. Such comments made me question my abilities and made me full of self- doubt myself because I've heard the same comment from a lot of patients.*

One of the students mentioned that a major boost to their self-efficacy was external. Attending and supervising health care staff were the significant contributors to that.

> *After removing the first skin tag my GP whispered "good job" and I was able to remove six more with confidence. I felt a sense of accomplishment that was indescribable, and I thanked my GP for helping me and giving me a chance to do this procedure.*

Many student narratives described initial reluctance to participate, often due to self-doubt or fear of making mistakes. However, some quickly transitioned into more active roles, showing initiative and professional autonomy. A visibly annoyed GP, frustrated by patients repeatedly interrupting examinations, appreciated the students' proactive efforts to manage the situation.

> *I noticed that my doctor was visibly frustrated by this since it was disrupting her consultations and the patients were losing their confidentiality. I asked my doctor to let me handle those patients. The next time this happened, I asked the patient at the door to meet me outside. I started by asking nicely "Why did you open the door?" and the patient replied, "I just need the lab test results and I've been waiting for hours now." After understanding the patient's intention, I de-escalated the situation by telling that I completely understand his frustration; however, the doctor is under the pressure of having to follow the order of what the system says, it would break the patient's privacy, and that once his turn comes in, I'll make sure to personally call him in.*

Discussion

Principal Findings

Students' reflections reveal a cohort entering the CLE with limited realistic preparation, mirrored by variably prepared GP tutors. Their narratives do not express explicit expectations of seamless transfer but instead reflect prereflective epistemological assumptions—the tacit belief that theoretical knowledge should translate into competent clinical action. When confronted with the realities of practice, students experienced shock and disorientation, signaling insufficient preparation for the well-documented fracture between knowing and doing. Without structured supervisory support to help them process this disjunction, their early experiences were marked by uncertainty, self-doubt, and emotional strain.

These accounts illuminate how students begin to think, feel, and enact as their professional identities take shape [13]. Early transitional moments were characterized by cognitive dissonance, emotional distress, and hesitant behavioral enactment. While reflection can support sense-making during such challenges, it is not sufficient on its own. It must be paired with preparatory activities and supervisory scaffolding to more effectively support students' PIF and their movement from peripheral to fuller participation in clinical communities of practice. This interpretation aligns with existing literature emphasizing the need for structured scaffolding during the transition from preclinical learning to the CLE. In the following sections, we further situate these findings using Wenger's LPP framework [5].

Fracture Between Knowledge and Knowing

A recurring theme was the fracture between theoretical knowledge and its real-world application, underscoring a persistent epistemological gap in medical education. Although students did not explicitly anticipate an effortless transition, many experienced cognitive dissonance when confronted with the uncertainty and practical demands of clinical environments. Their narratives varied in detail but revealed a shared emotional landscape of apprehension, fear, and unpreparedness. These reactions do not reflect naïve expectations but rather insufficient preparation for the well-recognized divide between knowing and doing. This dissonance reflects a deeper epistemic tension: students implicitly equated knowledge acquisition with readiness for action [18].

This assumption was particularly evident in their beliefs about "hard" technical skills, where they expected classroom learning to translate directly into competent performance. When confronted with failures in basic procedures, students who felt prematurely "cloaked in competence" experienced abrupt identity shock, often marked by shame, guilt, anxiety, and declining self-confidence [19,20]. Such experiences stunt LPP and add to the slow pile-up of burnout quotient. Conversely, there was a relatively mature understanding of the epistemological perspective when it came to "soft" skills, like professionalism-related matters of bias, consent, confidentiality, culture, and challenging hierarchy [17]. This can be attributed to the preexisting understanding of professionalism as an abstract concept. Hence, students had internalized the interpretation of professionalism interpretation, which is based on the evaluation of evidence across contexts and experts [21].

The nature of knowing ranged from simplicity for technical skills to justification for professionalism matters [21]. In the context of medical education, there is a perception that biomedical knowledge is often easier to acquire in an ideal environment compared to the biopsychosocial element—*understanding of what it means to be a physician* as mentioned by Whitcomb [22]. And that's exactly what our students understood. However, when it came to execution, they were struck by an identity shock, which had significant repercussions on their ability to learn, perform, and grow. Considering that most medical schools are still operating within the positivist paradigmatic orientations of their curricular structures, such results are expected [10]. To navigate that, introducing learning activities that increase the students' understanding of the CLE and the uncertainty, sense of incompetence, and failures that are associated with practicing medicine should be introduced [23]. An effective

XSL•FO
RenderX

implementation of assessment methods that reflect epistemological complexity, such as self-reflective narratives and longitudinal evaluations, combined with early clinical exposure, is key. This gradual familiarization with the ambiguity, messiness, and relational complexity of real-world clinical practice can help students become more resilient to the uncertainty that awaits them.

Emotional Labor

The tension between expectations and real-world clinical failures is likely to result from the complex and contextualized patient encounters that constitute clinical medicine. The slow unveiling of the complexity inherent in the simple task of measuring blood pressure and phlebotomy seemed a shocking reality for the students. Performing such skills on a human patient versus a simulator is much more complex than anticipated and triggers the destabilizing emotion of shame and guilt [20]. Only very few experienced health care providers were able to recognize a student's struggle with such emotional responses, missing an excellent opportunity to support them. Another source of stress emerged from the hidden curriculum [24] "little white lies" [25] when the nurse introduced the student as a doctor to practice IM injection. Students felt ashamed, amalgamated with fear of consequences, and were unable to challenge the attending who performed intimate examinations without proper consent and exposure.

Another source of emotional labor emerged from the materiality of everyday life in a CLE. Emotionally charged patient families, chaotic and loud environment buzzing with sounds of technology, and material things have been reported to hinder the learning process [26]. Fenwick's [27] seminal work has highlighted the limited attention to materiality (objects, technologies, and their nature) in learning. Medical curricula focus too heavily on human-centered practice, ignoring the vital role of material factors like technology and environment in shaping clinical learning. Students struggle with the tension between standardized medical protocols and the unpredictable realities of patient care, leading to anxiety and uncertainty. Without structured support mechanisms to process emotional distress, students may struggle to engage confidently in clinical settings, reinforcing avoidance behaviors that limit their participation and professional growth [20]. This advocates for the current apprenticeship approach to training to include a developmental perspective, providing effective feedback and supporting learner self‐assessment and reflection. Facilitating reflective debriefings after stressful experiences and acknowledging the presence of shame and guilt in the learner by avoiding humiliation and leveraging effective feedback can provide students with tools to navigate these complex emotions constructively [19].

Agency Versus Avoidance

The reflections endorsed our recent work [28] that a supportive attending and superior were decisive factors in boosting medical students' self-efficacy. Judging medical students and setting harsh, unrealistic expectations of being "cloaked in competence" are the most dreadful fears students perceive [29,30]. Learning happens when psychological safety is assured. Indeed, published literature suggests that supportive and encouraging preceptors who discuss failures [31] and the perplexity of CLE and provide constructive feedback can lead to *prepared* graduates [9]. Unlike Wald et al [9], none of the reflections in our case mentioned any discussion of mistakes, identification of areas of improvement, or debriefing sessions aimed at maximizing learning [32]. This is a critical omission, highlighting the need for educators to create spaces within the curriculum to discuss "failures" with students, preceptors, and clinical teachers [11]. According to Wenger's LPP, constructive failure processing mechanisms should be supported through sophisticated reflection-in-action [33]. This process should encompass awareness, collaboration, negotiation, evaluation, and realization to foster a conscious understanding of the transformation occurring in our medical students. Such structured processes will lead to graduated autonomy and avoidance of withdrawal. These trajectories of support, reflection, and graduated autonomy mirror the broader identity formation processes described earlier, where learners move from peripheral participation toward fuller membership in professional communities [5]. This leads us to consider the broader educational implications of our findings.

Educational Implications

Considering medicine is a social science and students perform in the social contexts of clinical environments, we need to introduce more constructionist nonbinary perspectives to the biomedical elements of the curriculum by providing more self-directed study opportunities [34]. Our findings expose structural and pedagogical limitations. Students thrown into the CLE without adequate psychological preparation and a structured supervisory support will be startled by the practical, emotional, and epistemological complexities inherent in the context. To move beyond that, medical curricula must frame uncertainty and failure as aspects of early clinical participation. Senior peers and role models sharing their reflective narratives of failure and fear provide psychological assurance and make failure, multiple failures, and success part of the whole process [35]. However, training the clinical supervisors and their ability to support, debrief, and role-model are essential. Within this broader scaffolding, reflective practice should be positioned not as a cure-all but as a developmental tool that complements preparatory and supervisory strategies. Despite the known benefits of reflective practice in the medical curriculum, we are not yet reaping the full benefits [36]. The literature suggests that we need to reflect upon reflection [10]. Incorporating uncertainty curricula [23], dealing with failures to normalize struggle [37], graduating autonomy models [38], and constructing inclusive, nonjudgmental learning spaces punctuated with frequent reflections upon reflections are required. Early identity struggles, if left unaddressed, may carry over into residency, where clinical demands intensify, increasing the risk for emotional exhaustion and disengagement [39]. By implementing structured identity support and normalizing professional uncertainty early in training, medical schools can better prepare students for the emotional and professional complexities of long-term clinical practice.

Reflective essay accounts rooted in the discipline of psychology have nicely elucidated these problems evident from our results. However, the application of reflection is heavily steered by the assessment discourse and is divorced from original theories of

XSL•FO
RenderX

reflection and reflective practice [40]. So clearly, the solution lies one step ahead—postreflective writing group reading/listening/discussions [8,41] and in multidisciplinary domains [42]. Inspired by Hodges [40], we advocate for ongoing, theory-informed exploration of reflection in medical education, particularly its role in navigating complexity, challenging dominant discourses, and shaping knowledge through diverse epistemological lenses and predominantly a critical social inquiry. Yet, the solution of viewing medicine from a more sociological perspective is compelling; the medical profession strongly believes in identity formation, whose disciplinary roots are entrenched in philosophy, psychology, and sociology. New ways of knowing in medical schools gravitate toward psychology alone [43]. However, the results of this study point to a more holistic approach when teaching professionalization to medical students in this humane profession [44]. We must look and discuss with our students the elements of resilience, humility, uncertainty, accepting failures, and limitations rooted in psychological and sociological principles [11]. Such multidisciplinary understanding may bridge the disconnect between the nature of knowledge and the nature of knowing. Table 1 provides guidance for curricular modifications to address the cognitive, emotional, and behavioral challenges identified in students' reflections.

Table . Summary of the 3 thematic domains (thinking, feeling, and enacting) derived from students' reflective accounts, with proposed key educational implications for curriculum design to support the process of professional identity formation (PIF).

Theme	Core insights from student reflections	Educational implications for curriculum design and PIF support
I. Thinking: cognitive dimension	Students experienced a fracture between theoretical knowledge and real-world application, revealing naïve epistemological beliefs that knowledge automatically translates into competence. They expressed uncertainty about their roles, ethical decision-making, and navigating culturally complex interactions.	Introduce early, structured clinical exposure with clear role expectations; provide orientation to epistemic uncertainty; embed narratives from senior peers and case-based discussions that highlight real-world complexity before clinical placements; incorporate shared reflective circles after the clinical placements to make epistemological tensions explicit and discuss uncertainties communally.
II. Feeling: emotional dimension	Students reported intense fear, guilt, shame, anxiety, and overwhelm, particularly during procedural tasks, culturally sensitive interactions, and ethically ambiguous situations. Emotional distress often stemmed from lack of preparation and insufficient support structures.	Implement preplacement skills practice on peers and simulations; create psychologically safe learning environments; integrate structured debriefings and shared reflective circles to process emotional labor, acknowledge shame/guilt as part of professional growth, and collectively normalize struggle both earlier in the clinical placements and the end of clinical placements
III. Enacting: behavioral dimension	Students' actions ranged from avoidance and withdrawal to proactive participation, shaped strongly by supervisory style and perceived psychological safety. Supportive preceptors enhanced agency; lack of feedback led to hesitation.	Train general practice tutors in feedback, debriefing, and psychological safety strategies; integrate shared reflection and postreflective dialogue sessions to support students in articulating their experiences, reframing failures, and sustaining engagement in clinical communities.

Limitations and Strengths

This study offers insights from a Middle Eastern CLE, a context that is underrepresented in the literature on PIF, thereby contributing to the global understanding of medical education. The QD method in this study focused on reflective essays because they provided us with a unique window into critical moments that prompt students to engage in introspection. By providing vivid personal accounts, these essays highlighted the students' narratives of key learning moments and challenges. Also, having student researchers from the same cohort helped us to interpret the data more truthfully. Their shared contextual understanding enabled us to ensure that the analysis remained authentic and reflective of the lived experiences of their peers, reducing the likelihood of misinterpretation and enhancing the credibility of the findings.

However, it must be acknowledged that this research has limitations. This cross-sectional analysis of one subset of a single cohort at a European medical school in a Middle Eastern context might be a naïve effort, limiting contextualization of our research findings. We analyzed 26 reflective essay accounts of students focused solely on GP placements, and the findings may not fully capture the professional identity struggles that occur in other specialties, which may have different levels of complexity. Moreover, we captured students' narrations, using reflective essay accounts, which can be merely the medium for expressing inner idealistic thoughts and feelings (when writing reflections or participating in discussions) [11] and are not drawn from discourse analysis [45]. Additionally, these reflective essays were submitted as part of an assessment process, and the findings may be subject to social desirability bias, where students may downplay certain struggles or amplify professionalism, knowing their work is being evaluated. Further, because students were prompted to describe "challenging or difficult encounters," their reflections may be disproportionately weighted toward negative experiences. This instruction may have introduced a response bias that limits the representation of more balanced or positive interactions. We also acknowledge that, as with any single research method, it is not a holistic attempt, and we may find alternative aspects with other

XSL•FO
RenderX

observational, ethnographic, or phenomenological examinations. Despite this, we believe this work reminds us of the complexities of transitioning moments in medical students' PIF.

Conclusion

We have presented the findings of medical students' transitioning moments from preclinical learning to the CLE. These reflective essays showed that students entered clinical practice with little to no realistic preparation and nonexistent structured supervisory support. The fluid and complex nature of PIF meant that this lack of preparedness inevitably gave rise to cognitive dissonance, emotional distress, and hesitant enactment. Students' relatively simplistic prereflective epistemological beliefs contributed to the cognitive dissonance, precipitating emotions of shame and guilt, which led to avoidance and withdrawal. Recognizing these pedagogical gaps is essential. Medical curricula and pedagogical approaches must equip students to manage the expectations, complexities, and uncertainties of the CLE before entry. This can be achieved through early skills training, opportunities for peer practice, exposure to senior peers' narratives, and clarity regarding roles and norms. However, such efforts cannot operate in isolation. Equipping GP tutors with strategies to support, debrief, and guide learners through these crucial moments is equally vital. Within this broader scaffolding, reflection remains a critical tool—not a cure-all but a complementary mechanism for sense-making, growth, and resilience. By integrating structured preparatory activities with theory-informed communal reflective circles, we can help these anxious members of the team navigate uncertainty, normalize failure, and cultivate resilience. As educators, we have already bridged the gap between knowledge and knowing by acknowledging the profession's demand of an equal mastery of scientific expertise alongside humanistic engagement. However, small final steps are needed. We need to reflect upon reflections. Let us make such curricular spaces to support the PIF of our new clinical team members.

Funding

The authors declared no financial support was received for this work.

Conflicts of Interest

None declared.

References

1. Cruess RL, Cruess SR, Boudreau JD, Snell L, Steinert Y. Reframing medical education to support professional identity formation. Acad Med 2014 Nov;89(11):1446-1451. [doi: 10.1097/ACM.0000000000000427] [Medline: 25054423]
2. Nordquist J, Hall J, Caverzagie K, et al. The clinical learning environment. Med Teach 2019 Apr;41(4):366-372. [doi: 10.1080/0142159X.2019.1566601] [Medline: 30880530]
3. Artemeva N, Rachul C, O'Brien B, Varpio L. Situated learning in medical education. Acad Med 2017 Jan;92(1):134. [doi: 10.1097/ACM.0000000000001495] [Medline: 28030424]
4. Irby D. Improving environments for learning in the health professions. Presented at: Proceedings of a conference chaired by David M Irby, PhD; Apr 15-18, 2018 URL: https://macyfoundation.org/assets/reports/publications/macy_monograph_2018_webfile.pdf [accessed 2026-02-26]
5. Lave J, Wenger E. Situated Learning: Legitimate Peripheral Participation: Cambridge University Press; 1991. [doi: 10.1017/CBO9780511815355]
6. Gaufberg EH, Batalden M, Sands R, Bell SK. The hidden curriculum: what can we learn from third-year medical student narrative reflections? Acad Med 2010 Nov;85(11):1709-1716. [doi: 10.1097/ACM.0b013e3181f57899] [Medline: 20881818]
7. Wald HS. Professional identity (trans)formation in medical education: reflection, relationship, resilience. Acad Med 2015 Jun;90(6):701-706. [doi: 10.1097/ACM.0000000000000731] [Medline: 25881651]
8. Shapiro J, Kasman D, Shafer A. Words and wards: a model of reflective writing and its uses in medical education. J Med Humanit 2006;27(4):231-244. [doi: 10.1007/s10912-006-9020-y] [Medline: 17001529]
9. Wald HS, White J, Reis SP, Esquibel AY, Anthony D. Grappling with complexity: medical students' reflective writings about challenging patient encounters as a window into professional identity formation. Med Teach 2019 Feb;41(2):152-160. [doi: 10.1080/0142159X.2018.1475727] [Medline: 29944035]
10. Mann K, Gordon J, MacLeod A. Reflection and reflective practice in health professions education: a systematic review. Adv Health Sci Educ Theory Pract 2009 Oct;14(4):595-621. [doi: 10.1007/s10459-007-9090-2] [Medline: 18034364]
11. Benbassat J. Undesirable features of the medical learning environment: a narrative review of the literature. Adv Health Sci Educ Theory Pract 2013 Aug;18(3):527-536. [doi: 10.1007/s10459-012-9389-5] [Medline: 22760724]
12. Sandelowski M. What's in a name? Qualitative description revisited. Res Nurs Health 2010 Feb;33(1):77-84. [doi: 10.1002/nur.20362] [Medline: 20014004]
13. Gibbs G. Learning by Doing: A Guide to Teaching and Learning Methods: Further Education Unit; 1988. URL: https://cir.nii.ac.jp/crid/1571417125052979584 [accessed 2025-04-01]
14. Branch WT. Use of critical incident reports in medical education. A perspective. J Gen Intern Med 2005 Nov;20(11):1063-1067. [doi: 10.1111/j.1525-1497.2005.00231.x] [Medline: 16307635]

XSL•FO
RenderX

15. Kim H, Sefcik JS, Bradway C. Characteristics of qualitative descriptive studies: a systematic review. Res Nurs Health 2017 Feb;40(1):23-42. [doi: 10.1002/nur.21768] [Medline: 27686751]
16. Guraya SS, Guraya SY, Harkin DW, Ryan Á, Mat Nor MZB, Yusoff MSB. Medical Education e-Professionalism (MEeP) framework; from conception to development. Med Educ Online 2021 Dec;26(1):1983926. [doi: 10.1080/10872981.2021.1983926] [Medline: 34775927]
17. Guraya SS, Guraya SY, Doubell FR, et al. Understanding medical professionalism using express team-based learning; a qualitative case-based study. Med Educ Online 2023 Dec;28(1):2235793. [doi: 10.1080/10872981.2023.2235793] [Medline: 37463323]
18. Evans RG. Patient centred medicine: reason, emotion, and human spirit? Some philosophical reflections on being with patients. Med Humanit 2003 Jun;29(1):8-14. [doi: 10.1136/mh.29.1.8] [Medline: 23671167]
19. Bynum WE 4th, Goodie JL. Shame, guilt, and the medical learner: ignored connections and why we should care. Med Educ 2014 Nov;48(11):1045-1054. [doi: 10.1111/medu.12521] [Medline: 25307632]
20. Bynum WE 4th, Teunissen PW, Varpio L. In the “shadow of shame”: a phenomenological exploration of the nature of shame experiences in medical students. Acad Med 2021 Nov 1;96(11S):S23-S30. [doi: 10.1097/ACM.0000000000004261] [Medline: 34348391]
21. Knight LV, Mattick K. “When I first came here, I thought medicine was black and white”: making sense of medical students’ ways of knowing. Soc Sci Med 2006 Aug;63(4):1084-1096. [doi: 10.1016/j.socscimed.2006.01.017] [Medline: 16519978]
22. Whitcomb ME. The teaching of basic sciences in medical schools. Acad Med 2006 May;81(5):413-414. [doi: 10.1097/01.ACM.0000222262.85760.13] [Medline: 16639189]
23. Kitto SC, Chesters JE, Villanueva EV, Fox JG. Normalising uncertainty in undergraduate clinical transition seminars. Focus Health Prof Educ 2004;6(1):37-51 [FREE Full text]
24. Guraya SS, Kearney GP, Doyle F, et al. “Busting the hidden curriculum” a realist and innovative perspective to foster professional behaviors. Front Med 2024;11:1484058. [doi: 10.3389/fmed.2024.1484058]
25. Lingard L, Schryer C, Garwood K, Spafford M. “Talking the talk”: school and workplace genre tension in clerkship case presentations. Med Educ 2003 Jul;37(7):612-620. [doi: 10.1046/j.1365-2923.2003.01553.x] [Medline: 12834419]
26. Goldman E, Plack M, Roche C, Smith J, Turley C. Learning in a chaotic environment. J Workplace Learn 2009 Sep 11;21(7):555-574. [doi: 10.1108/13665620910985540]
27. Fenwick T. Sociomateriality in medical practice and learning: attuning to what matters. Med Educ 2014 Jan;48(1):44-52. [doi: 10.1111/medu.12295] [Medline: 24330116]
28. Habib H, Niinuma SA, Alrefaie K, et al. Shades of acceptance and adjustment; a discursive psychological analysis to showcase empathy in medical students. BMC Med Educ 2025 Mar 21;25(1):415. [doi: 10.1186/s12909-024-06625-2] [Medline: 40119323]
29. Jin CJ, Martimianakis MA, Kitto S, Moulton CAE. Pressures to “measure up” in surgery: managing your image and managing your patient. Ann Surg 2012 Dec;256(6):989-993. [doi: 10.1097/SLA.0b013e3182583135] [Medline: 22824849]
30. Kitto S, Alexanian J, Goldman J. Attuning to the social world: ethnography in health professions education research. In: Researching Medical Education: Wiley; 2022:127-137. [doi: 10.1002/9781119839446]
31. Kroll L, Singleton A, Collier J, Rees Jones I. Learning not to take it seriously: junior doctors’ accounts of error. Med Educ 2008 Oct;42(10):982-990. [doi: 10.1111/j.1365-2923.2008.03151.x] [Medline: 18823517]
32. Branch WT Jr, Paranjape A. Feedback and reflection: teaching methods for clinical settings. Acad Med 2002 Dec;77(12 Pt 1):1185-1188. [doi: 10.1097/00001888-200212000-00005] [Medline: 12480619]
33. Orsmond P, McMillan H, Zvauya R. It’s how we practice that matters: professional identity formation and legitimate peripheral participation in medical students: a qualitative study. BMC Med Educ 2022 Feb 9;22(1):91. [doi: 10.1186/s12909-022-03107-1] [Medline: 35139839]
34. Frambach JM, Driessen EW, Chan LC, van der Vleuten CPM. Rethinking the globalisation of problem-based learning: how culture challenges self-directed learning. Med Educ 2012 Aug;46(8):738-747. [doi: 10.1111/j.1365-2923.2012.04290.x] [Medline: 22803751]
35. Nishigori H, Shimazono Y, Busari J, Dornan T. Exploring yarigai: the meaning of working as a physician in teaching medical professionalism. Med Teach 2024 Nov;46(11):1486-1493. [doi: 10.1080/0142159X.2024.2316227] [Medline: 38376459]
36. Wald HS, Reis SP. Beyond the margins: reflective writing and development of reflective capacity in medical education. J Gen Intern Med 2010 Jul;25(7):746-749. [doi: 10.1007/s11606-010-1347-4] [Medline: 20407840]
37. Klasen JM, Lingard LA. Allowing failure for educational purposes in postgraduate clinical training: a narrative review. Med Teach 2019 Nov;41(11):1263-1269. [doi: 10.1080/0142159X.2019.1630728] [Medline: 31280625]
38. Ramani S, Leinster S. Amee Guide no. 34: teaching in the clinical environment. Med Teach 2008;30(4):347-364. [doi: 10.1080/01421590802061613] [Medline: 18569655]
39. Peterson WJ, House JB, Sozener CB, Santen SA. Understanding the struggles to be a medical provider: view through medical student essays. J Emerg Med 2018 Jan;54(1):102-108. [doi: 10.1016/j.jemermed.2017.09.014] [Medline: 29100653]
40. Ng SL, Kinsella EA, Friesen F, Hodges B. Reclaiming a theoretical orientation to reflection in medical education research: a critical narrative review. Med Educ 2015 May;49(5):461-475. [doi: 10.1111/medu.12680] [Medline: 25924122]

XSL•FO
RenderX

41. Lie D, Shapiro J, Cohn F, Najm W. Reflective practice enriches clerkship students' cross-cultural experiences. J Gen Intern Med 2010 May;25(Suppl 2):S119-S125. [doi: 10.1007/s11606-009-1205-4] [Medline: 20352505]
42. Gerrity MS, Earp JAL, DeVellis RF, Light DW. Uncertainty and professional work: perceptions of physicians in clinical practice. Am J Sociol 1992 Jan;97(4):1022-1051. [doi: 10.1086/229860]
43. Roberts J. Limits to communities of practice. J Manag Stud 2006 May;43(3):623-639. [doi: 10.1111/j.1467-6486.2006.00618.x]
44. Amin A, Roberts J. Knowing in action: beyond communities of practice. Res Policy 2008 Mar;37(2):353-369. [doi: 10.1016/j.respol.2007.11.003]
45. Taylor C. Narrating significant experience: reflective accounts and the production of (self) knowledge. Br J Soc Work 2006;36(2):189-206. [doi: 10.1093/bjsw/bch269]

Abbreviations

CLE: clinical learning environment
GP: general practice
LPP: legitimate peripheral participation
PIF: professional identity formation
QD: qualitative descriptive
RCSI-MUB: Royal College of Surgeons in Ireland – Medical University of Bahrain

Edited by M Montagna; submitted 27.Apr.2025; peer-reviewed by E Schei, Q Yang; revised version received 23.Dec.2025; accepted 23.Dec.2025; published 12.Mar.2026.

Please cite as:
Qasrawi SB, Tomerak M, Mamalchi SA, Atieh F, AlAdraj H, Abdulla M, Fredericks S, Jassim G, Malik H, Clarke E, Harkin D, Guraya SS
Medical Students' Reflections on Transitioning to Their First General Practice Placement: Qualitative Descriptive Study
JMIR Med Educ 2026;12:e76599
URL: https://mededu.jmir.org/2026/1/e76599
doi:10.2196/76599

© Sara Bashar Qasrawi, Maryam Tomerak, Shahad Abdulkhaleq Mamalchi, Fatima Atieh, Hasan AlAdraj, Mohamed Abdulla, Salim Fredericks, Ghufran Jassim, Hani Malik, Eric Clarke, Denis Harkin, Shaista Salman Guraya. Originally published in JMIR Medical Education (https://mededu.jmir.org), 12.Mar.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Education, is properly cited. The complete bibliographic information, a link to the original publication on https://mededu.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

AI Perceptions, Professional Identity, and AI-Supported Clinical Decisions Among Medical Students and Clinicians: Cross-Sectional Survey and Quasi-Randomized Vignette Study

Julia-Astrid Moldt[1], MA; Teresa Festl-Wietek[1], Dr rer nat; Kay Nieselt[2,3], Prof Dr; Susanne Zabel[2,3], Dr rer nat; Manfred Claassen[2,3,4,5], Prof Dr; Samuel Wagner[2], Prof Dr; Ulrike Keim[6,7], Dr sc hum; Anne Herrmann-Werner[1,8], Prof Dr Med

[1]TIME – Tübingen Institute for Medical Education, Faculty of Medicine, University of Tübingen, Tübingen, Germany

[2]Interfaculty Institute of Microbiology and Infection Medicine, University of Tübingen, Tübingen, Germany

[3]Department of Computer Science, University of Tübingen, Tübingen, Germany

[4]Department of Internal Medicine I, University Hospital Tübingen, Tübingen, Germany

[5]M3 Research Center, University Hospital Tübingen, Tübingen, Germany

[6]Institute for Health Sciences, Department of Midwifery Science, University Hospital Tübingen, Tübingen, Germany

[7]Research Institute for Women´s Health, Department of Midwifery Science, University Hospital Tübingen, Tübingen, Germany

[8]German Center for Mental Health (DZPG), Partner Site Tübingen, Tübingen, Germany

Corresponding Author:
Julia-Astrid Moldt, MA
TIME – Tübingen Institute for Medical Education
Faculty of Medicine
University of Tübingen
Elfriede-Aulhorn-Str. 10
Tübingen, 72076
Germany
Phone: 49 7071 29 87868
Email: julia-astrid.moldt@med.uni-tuebingen.de

Abstract

Background: The integration of AI into health care is increasingly shaping clinical practice and decision-making. Beyond technical performance, AI has implications for professional roles, clinical reasoning, and responsibility. Understanding how medical students and clinicians perceive AI, how these perceptions relate to professional identity concerns, and how clinicians evaluate AI-supported decisions in different clinical contexts is therefore important for both implementation and medical education.

Objective: This study aimed to examine (1) how medical students and clinicians perceive the impact of AI on medical practice and professional identity and (2) how clinicians evaluate explainability, trustworthiness, and responsibility in AI-supported clinical decision-making across different contexts.

Methods: A 2-phase quantitative study was conducted. In phase 1, medical students (n=93) and clinicians (n=65; N=158) completed a cross-sectional online survey assessing expectations regarding AI, perceived importance of AI across medical domains, and perceived professional identity threat. In phase 2, a separate sample of clinicians (N=68) was quasi-randomly assigned to 1 of 3 clinical vignettes (Watson, Triage, and OncoGuide) and evaluated AI-supported decisions along the dimensions of explainability, trustworthiness, and responsibility.

Results: In phase 1, participants reported generally positive expectations regarding factual-level impacts of AI on medical practice. However, these expectations were not significantly associated with perceived professional identity threat. Medical students reported significantly higher identity threat than clinicians, despite largely similar expectations regarding AI's factual and social impact. Perceived importance of AI across medical domains was associated with more positive expectations toward AI-supported medical practice, but not with identity threat. In phase 2, clinicians' evaluations of AI-supported decisions varied across contexts, with explainability ratings differing significantly across vignette scenarios (P=.007), with lower ratings in the Triage vignette than in the Watson and OncoGuide vignettes. Trustworthiness and responsibility did not differ significantly across scenarios.

XSL•FO
RenderX

Conclusions: The findings suggest that positive expectations regarding AI may coexist with professional identity concerns, while identity threat differed between medical students and clinicians in the present sample. At the same time, explainability evaluations differed across clinical decision contexts, whereas trustworthiness and responsibility remained comparatively stable. Together, these findings point to both professional identity-related and contextual considerations in AI implementation that may not be captured by expectations regarding technological benefits alone. Considering these dimensions may be relevant for implementation strategies and medical education.

(JMIR Med Educ 2026;12:e104151) doi:10.2196/104151

KEYWORDS

AI; clinical decision-making; professional identity; medical students; physicians; explainability; trust; responsibility

Introduction

Background

The integration of AI into health care is increasingly framed as a transformative development with profound implications for medical practice, professional expertise, and clinical decision-making [1]. Recent advances in machine learning, predictive modeling, and algorithm-based clinical decision support are progressively moving from research prototypes into routine clinical workflows, particularly in domains such as radiology, pathology, oncology, and risk stratification [2-4]. As a result, clinicians increasingly need to balance algorithmic outputs with clinical judgment and professional accountability [5]. This has shifted attention from technical performance alone to the practical conditions under which AI recommendations can be meaningfully interpreted, justified, and integrated into clinical decision-making [6,7]. Rather than referring only to full technical transparency, explainability is increasingly discussed as context-sensitive information that supports clinicians in assessing the relevance, risks, and justification of AI-supported recommendations [8]. This reflects a broader shift in the literature, which increasingly frames AI implementation in medicine not merely as a technological innovation, but as a sociotechnical process involving clinical, organizational, ethical, and legal dimensions [9-11]. Contemporary research further suggests that AI adoption may affect health care professionals' sense of professional identity, trust in clinical decision support, and negotiations of roles and responsibilities in clinical contexts [12,13]. Research exploring the social dimensions of AI integration in health care further demonstrates that AI implementation involves navigating organizational, professional, and ethical challenges, not merely enhancing algorithmic performance [14,15]. Against this background, understanding how clinicians perceive, interpret, and negotiate AI-supported decisions in concrete clinical situations becomes crucial.

Drawing on social constructivist approaches to technology, AI systems can be understood as open to different interpretations by relevant stakeholder groups within medicine. Their perceived meaning and implications depend not only on technical performance, but also on how clinicians and future clinicians position these systems in relation to expertise, authority, and responsibility [16,17]. The introduction of AI systems capable of generating diagnostic or therapeutic recommendations may challenge these foundations by redistributing cognitive tasks and raising questions about what is regarded as a legitimate source of clinical knowledge or decision authority [18,19]. For many practitioners, AI-generated outputs introduce new forms of uncertainty and raise questions about who or what constitutes a legitimate source of clinical knowledge, with implications for professional identity, trust, and accountability in everyday practice [20,21]. Empirical research shows that both medical students and practicing clinicians acknowledge the growing relevance of AI for medical practice [22]. Comparing these groups is particularly relevant because prior research suggests that medical students and less experienced clinicians may experience AI-related professional identity threats differently from experienced clinicians [23]. At the same time, perceptions are markedly ambivalent. While some view AI as a supportive tool that may enhance efficiency and clinical reasoning, others express concerns regarding deskilling, loss of professional autonomy, and shifting responsibilities [24,25]. Previous studies suggest that these perceptions differ across professional subgroups, training stages, and clinical specialties [22,26]. Beyond general attitudes toward AI, recent literature highlights explainability, trustworthiness, and questions of accountability and responsibility as important considerations in the clinical use of AI systems [20,27]. These aspects are closely connected to whether clinicians can understand and critically assess AI-supported recommendations, regard them as sufficiently reliable, and retain responsibility for the resulting clinical decision. They therefore informed the 3 evaluative dimensions examined in phase 2: explainability, trustworthiness, and responsibility. Trusting an AI system, demanding explanations, and retaining responsibility for AI-supported decisions are therefore not only technical or ethical issues but also relate to how clinicians perceive their role, authority, and accountability in clinical practice [28,29]. Recent empirical studies also suggest that the design and integration of AI-based decision support systems can influence clinicians' trust while simultaneously raising concerns about professional identity and responsibility [30]. Closely related work by Ackerhans et al [13] examined how process design features of a fictitious sepsis-related AI-based clinical decision support system, including explainability, workflow integration, and system-induced accountability, shaped trust and perceived professional identity threat in a controlled scenario-based experimental setting. This provides important evidence on how system design features matter for trust and identity-related concerns. This study addresses a complementary perspective by shifting attention to variation across clinical decision contexts and to broader professional expectations and identity-related concerns surrounding AI. In phase 1, medical students and clinicians were examined with regard to general AI-related expectations

XSL•FO
RenderX

and professional identity concerns. Separately, phase 2 examined clinicians' evaluations of explainability, trustworthiness, and responsibility across distinct clinical vignette contexts.

Research Questions

We address the following research questions (RQs):

- RQ 1: How do medical students and clinicians perceive the role of AI in medicine, particularly regarding expected impacts on medical practice and perceived professional identity threat, and do these perceptions differ between the 2 groups?
- RQ 2: Do clinicians' evaluations and prioritizations of explainability, trustworthiness, and responsibility in AI-supported clinical decisions differ across clinical vignette contexts?

Hypotheses

To address these RQs, we formulated the following hypotheses, which were preregistered as described in the Methods section. Hypotheses 1-3 relate to RQ1 and focus on general perceptions of AI and professional identity, whereas hypothesis 4 is stated here in an analysis-aligned form addressing RQ2 through situational evaluations of AI-supported clinical decisions across vignette contexts. Its broader preregistered formulation is reported in the Data Processing and Statistical Analysis section.

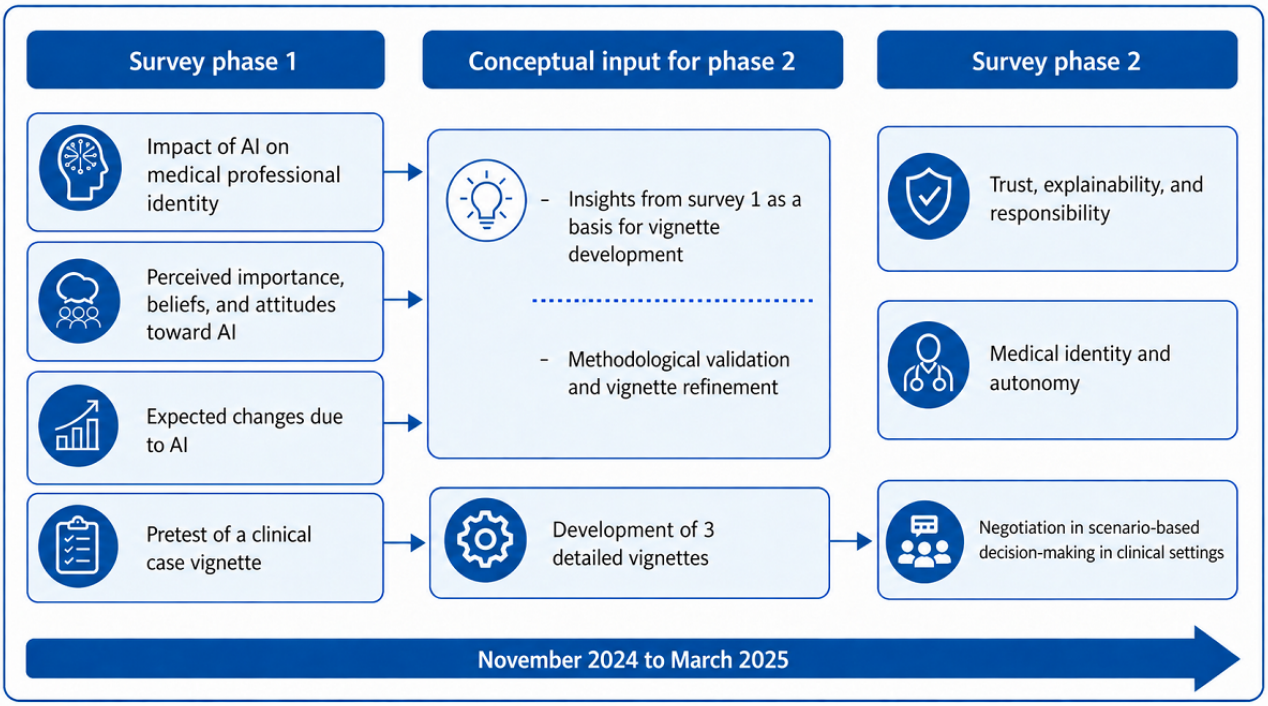
- Hypothesis 1: Clinicians and medical students who hold more positive expectations about AI's impact on medical practice (eg, improved diagnosis and streamlined workflows) will report lower levels of perceived identity threat (measured across distinctiveness, continuity, self-esteem, and self-efficacy).
- Hypothesis 2: Participants who consider AI to be important across core medical domains (eg, diagnosis, patient communication, and administration) will show higher acceptance of AI and lower identity threat.
- Hypothesis 3: Medical students and clinicians differ in how they perceive AI's factual and social impact.
- Hypothesis 4: Clinicians' evaluations of explainability, trustworthiness, and responsibility in AI-supported clinical decision-making differ across clinical vignette contexts.

Methods

Study Design

The overall study design is illustrated in Figure 1. The study followed a sequential, 2-phase design combining a cross-sectional online survey (phase 1) and a quasi-randomized vignette study (phase 2). Phase 1 assessed general orientations toward AI and its perceived implications for the professional role and identity of clinicians (addressing RQ1 and hypotheses 1-3). Phase 2 complemented phase 1 by addressing the additional RQ2 in a separate clinician sample, focusing on context-specific evaluations of explainability, trustworthiness, and responsibility in AI-supported clinical decision-making.

Figure 1. Survey outline.



Survey Administration

Both phases were conducted as web-based surveys. Reporting of both survey phases was guided by the CHERRIES (Checklist for Reporting Results of Internet E-Surveys) [31]. The completed CHERRIES checklist is provided as Multimedia Appendix 1.

Phase 1: Medical Students and Clinicians

Phase 1 was conducted between November 6, 2024, and February 9, 2025, using a study link [30]. It was an open

XSL•FO
RenderX

web-based survey, as participation was possible for eligible individuals who received access to the study link through the recruitment channels described below. The questionnaire was developed based on a structured literature review and pretested with 13 participants from the target population. Minor revisions were made to improve clarity, comprehensibility, and technical functionality before fielding the final survey.

Recruitment was carried out through multiple channels relevant to the German medical education and health care context. The survey targeted medical students across different stages of training and clinicians with varying levels of professional experience and specialization. Central contact points at university hospitals, maximum-care hospitals, and medical student councils of human medicine were contacted directly and asked to disseminate the study invitation to potentially eligible participants. In addition, the study was distributed through internal institutional mailing lists and flyers. The invitation included information on the study purpose, target group, voluntary participation, approximate completion time, and a link to the web-based questionnaire. Participation was self-selected.

Eligible participants were medical students studying in Germany and clinicians working in Germany. Medical students included participants across all educational phases, including preclinical, clinical, and final-year training. The clinician group included practicing clinicians, clinical researchers, academics, and medically trained administrators.

To ensure consistency with the study's focus on the German context of medicine and medical training, the dataset was cleaned based on predefined exclusion criteria. Participants who did not report studying or working in Germany were excluded. Only participants who completed the questionnaire up to the section immediately preceding the clinical case vignette were retained for the main analysis, as this indicated a sufficiently complete response set covering the core constructs of interest, including professional identity, perceptions of AI, and role-related expectations. Participants whose professional status could not be clearly assigned to either the medical student or clinician group were excluded from group-based analyses.

Phase 2: Clinicians

Phase 2 was conducted as a web-based vignette survey between February 6 and March 20, 2025. The survey was accessed via a SoSci Survey study link and distributed to a defined group of clinicians participating in the medical didactics qualification program at the Tübingen Institute for Medical Education. The qualification program was held on several course dates during the data collection period, with different clinicians participating on different dates.

This phase used a purposive convenience sample of clinicians with teaching experience and an interest in medical education. Eligible participants were clinicians enrolled in the qualification program during the data collection period. The study invitation was distributed within the course context and included information on the study purpose, voluntary participation, data protection procedures, approximate completion time, and access to the web-based vignette survey.

The vignette-based survey was implemented as a course-related reflection task; however, participation in the research study was voluntary. After accessing the questionnaire, participants completed 1 assigned clinical vignette and a standardized set of follow-up items. The sample was self-selected, as participation depended on clinicians' decision to complete the questionnaire.

Measures

All measures were selected and developed to address the RQs and hypotheses of the study. Phase 1 measures focused on general perceptions of AI and its implications for professional identity (RQ1; hypotheses 1-3), whereas phase 2 measures captured situational evaluations of AI-supported decision-making in vignette-based scenarios (RQ2; hypothesis 4).

Initial Survey (Phase 1)

Questionnaire Development

The phase 1 questionnaire was developed based on a structured literature review and pretested with 13 participants from the target population. Minor revisions were made to improve clarity and conceptual coherence. The final instrument comprised closed-ended items assessing professional identity threat, perceived importance of AI across medical domains (IM), and expected AI-related changes in medical practice. The full set of phase 1 questionnaire items and response scales is provided in Multimedia Appendix 2.

Perceived AI Impact on Professional Identity

Perceived AI-related impacts on professional identity were assessed using 6 items theoretically informed by Breakwell's identity process theory [32]: distinctiveness (1 item), continuity (1 item), self-esteem (2 items), and self-efficacy (2 items). The items focused on potentially threatening implications of AI for professional identity, such as changes in perceived expertise, confidence, and professional role. Items were rated on a 5-point agreement scale and averaged to form an overall perceived AI impact on professional identity (PI) score, with higher scores indicating stronger perceived professional identity threat. As the individual principles were represented by only 1 or 2 items, dimension-specific subscale scores were not constructed. Interitem correlations are reported in Multimedia Appendix 3. PI served as the primary outcome variable for hypothesis 1 and as one of the outcomes examined in hypothesis 2.

Perceived Importance of AI Across Medical Domains

IM was assessed across 5 medical domains: diagnostic accuracy and efficiency, support in complex diagnostic cases, personalized treatment planning, doctor-patient communication, and administrative simplification. Items were rated on a 3-point response scale (1=extremely to 3=not at all). For analysis, items were recoded so that higher values indicated higher IM and were averaged to obtain an overall index score of perceived importance across the selected domains. The index score was calculated from the available item responses when at least 1 of the 5 items had been completed; missing item responses were not imputed. Because these domains represent heterogeneous areas of medical practice, the measure was treated as a

XSL•FO
RenderX

domain-level index rather than as a reflective scale measuring a single latent construct. The IM index was used in the analyses testing hypothesis 2.

Expected AI-Related Changes in Medical Practice

Expectations regarding AI-related changes in medical practice were assessed at 2 levels: factual-level expectations regarding AI-related changes in medical practice (EC_FL; 6 items), referring to clinical and organizational aspects of medical work, and social-level expectations regarding AI-related changes in medical practice (EC_SL; 9 items), referring to relational aspects of care. Items were rated on a 5-point agreement scale, with higher values indicating stronger agreement with the expected changes. Mean scores were calculated separately for EC_FL and EC_SL, with higher scores reflecting more positive expectations regarding AI-related changes at the respective level. One negatively worded social-level item (EC06_01) was reverse-coded so that higher values consistently reflected more positive expectations regarding AI-related social changes. Because phase 1 did not include a separate global acceptance measure, EC_FL was used in hypothesis 2 as a pragmatic acceptance-related indicator. In this context, the term refers to favorable evaluations of expected AI-related changes in practical aspects of medical work, rather than to direct behavioral acceptance. EC_FL was also used as the predictor in hypothesis 1, whereas both EC_FL and EC_SL were included in the group comparisons testing hypothesis 3.

Control Variables

Sociodemographic variables included role (medical student vs clinician), age, and professional experience. In addition, self-reported experience with AI in the medical field was assessed and included as a control variable in the analyses.

At the end of the phase 1 questionnaire, an exploratory clinical case vignette ("Watson") was included to obtain initial insights into AI-supported clinical decision-making. Responses to this exploratory vignette were analyzed descriptively and used to inform the development of the quasi-randomized vignette study in phase 2.

Quasi-Randomized Vignette (Phase 2—RQ2; Hypothesis 4)

Phase 2 used a quasi-randomized vignette design to examine clinicians' evaluations of AI-supported clinical decision-making in concrete situations. Participants were allocated to 1 of 3 clinical vignettes depicting distinct AI-supported decision scenarios using a quasi-random birth-date criterion based on their day of birth within the month: days 1-10 to Watson, days 11-20 to Triage, and days 21-31 to OncoGuide (Table 1). The vignette stimulus texts were kept comparable in structure and length. Excluding the subsequent rating items, vignette length ranged from approximately 238 to 317 words across the 3 conditions. Vignette-specific reading time was not recorded separately.

Table 1. Overview of vignettes and contextual dimensions.

Vignette	Clinical focus	Contextual variation	Key thematic dimension
Watson—intensive care	AI recommends a treatment that deviates from clinical guidelines	Limited transparency of algorithmic reasoning	Guideline adherence versus AI reasoning
Triage emergency	Team must make a rapid decision under time pressure	Diverging opinions between clinicians and AI	Trust under uncertainty and time pressure
OncoGuide—oncology	Shared decision-making between physician, patient, and AI	AI provides an individualized but nonguideline-based recommendation	Responsibility and patient autonomy

The vignettes were designed to capture typical situations of professional tension and negotiation that may arise with AI implementation (eg, algorithmic recommendations deviating from clinical guidelines, disagreement between clinicians and AI under time pressure, and AI involvement in shared decision-making with patients). After reading the assigned vignette, participants rated a standardized set of items assessing perceived trustworthiness, explainability, and responsibility of the AI system. All items were rated on 5-point Likert scales (1=strongly disagree to 5=strongly agree).

Some items referred to willingness to accept, follow, or implement an AI recommendation; however, these items were embedded within the dimension-specific scales or vignette-specific trade-off items and were therefore not treated as a separate acceptance outcome. Items belonging to each dimension were aggregated into 3 composite scale scores by calculating mean values. In addition to these standardized scales, vignette-specific prioritization items were included to capture how participants weighed these dimensions against each other within the concrete clinical context, for example, prioritizing trust over explainability under time pressure. These items reflected explicit trade-offs between dimensions and were analyzed descriptively: (1) Trust over explainability (eg, reliance on AI under time pressure), (2) responsibility over explainability (eg, clear accountability reduces need for transparency), and (3) explainability over trust (eg, refusal to follow AI advice without understanding). Although the wording was adapted to each vignette, the underlying dimension contrasts remained identical. All vignette scenarios and item wordings are provided in Multimedia Appendix 4.

Data Processing and Statistical Analysis

The study was preregistered on the Open Science Framework, including the key hypotheses and planned analytic strategy [33]. The analyses for hypotheses 1 and 3 corresponded to the preregistered hypotheses, although the covariates included in the adjusted hypothesis 1 regression model were not specified in detail in the preregistration. For hypothesis 2, EC_FL were used as an acceptance-related indicator in phase 1. The preregistered formulation of hypothesis 4 stated that acceptance of AI-supported clinical decisions depends on perceived trustworthiness, explainability, and responsibility. In phase 2,

XSL•FO
RenderX

acceptance of AI-supported clinical decisions was operationalized through participants' agreement with dimension-specific items on explainability, trustworthiness, and responsibility, together with vignette-specific items referring to following, accepting, or implementing AI-supported recommendations. Higher ratings indicated stronger agreement with the respective acceptance-related evaluation in each vignette context. The phase 2 analyses examined whether ratings of explainability, trustworthiness, and responsibility differed across the Watson, Triage, and OncoGuide scenarios using one-way ANOVAs. No separate composite acceptance score was constructed; therefore, phase 2 is reported as a vignette-based analysis of acceptance-related evaluations rather than as a regression model predicting a global acceptance outcome. All statistical analyses were conducted using SPSS Statistics (version 31; IBM Corp). Sensitivity power analyses were conducted using G*Power (version 3.1.9.7) for the phase 2 one-way ANOVAs [34]. Data were screened for completeness and plausibility prior to analysis. Descriptive statistics were calculated for all study variables.

Hypotheses 1 and 2 were examined using correlational and multiple regression analyses. For hypothesis 1, perceived professional identity threat was regressed on EC_FL, with group membership, age, and prior AI experience included as covariates in the adjusted model. For hypothesis 2, IM was examined in relation to EC_FL and perceived professional identity threat. Group differences hypothesized in hypothesis 3 were tested using 2-tailed independent-samples *t* tests or nonparametric equivalents where appropriate. For phase 2, one-way ANOVAs were used to examine whether clinicians' ratings of perceived explainability, trustworthiness, and responsibility differed across the 3 vignette scenarios. Vignette-specific prioritization items were analyzed descriptively. Statistical significance was set at $P<.05$. Effect sizes were reported where appropriate. Analyses not specified in the preregistration were labeled as exploratory. Missing data were not imputed. Descriptive item-level analyses were based on available valid responses per item. Regression analyses used listwise deletion. Bivariate correlations were calculated using available valid pairs. Analytic sample sizes are reported for the main analyses and in Multimedia Appendix 5.

Ethical Considerations

The study, including both phase 1 and phase 2, received ethics approval from the ethics committee of the Medical Faculty of the University of Tübingen (342/2024BO2). All participants in both phases were informed about the study purpose, data protection procedures, the voluntary nature of participation, and their right to discontinue participation during the survey without providing reasons and without disadvantages. Informed consent was obtained prior to participation. Data were collected anonymously and used solely for the purposes of this study. No identifying information is reported in this paper. No monetary incentives or financial compensation were provided.

Results

Study Population and Descriptive Characteristics (Initial Survey)

A total of 323 participants started the phase 1 survey. Of these, 172 reached the predefined completion threshold, defined as completion of the questionnaire up to the section immediately preceding the clinical case vignette. The remaining 151 records did not meet this completion criterion. Of the 172 sufficiently complete responses, 14 were excluded because participants did not report studying or working in Germany. Therefore, the final phase 1 sample comprised 158 participants, including 93 (58.9%) medical students and 65 (41.1%) clinicians. Medical students represented different stages of medical education, including preclinical, clinical, and practical year training. The clinician group included practicing clinicians, resident and senior clinicians, clinical researchers, medical educators, and medically trained professionals working in administrative or managerial roles. Descriptive characteristics of the phase 1 study population are presented in Table 2. Age and sex statistics are based on valid responses only.

XSL•FO
RenderX

Table 2. Demographic and professional characteristics of participants (N=158).

Characteristic and category	Medical students, n=93 (58.9%)	Clinicians, n=65 (41.1%)
Age (years)		
Mean (SD)	25.3 (5.4)	37.5 (9.7)
Range	19-57	24-60
Sex, n (%)[a]		
Female	46 (61.3)	24 (44.4)
Male	29 (38.7)	30 (55.6)
Experience		
Semester of study, mean (SD)	8.2 (3.8)	N/A[b]
Professional experience, n (%)[c]	N/A	
Less than 1 year		7 (14.0)
1-3 years		12 (24.0)
4-6 years		6 (12.0)
7-10 years		7 (14.0)
More than 10 years		18 (36.0)

[a]Percentages are based on nonmissing responses; sex data were available for 75 medical students and 54 clinicians.

[b]N/A: not applicable.

[c]Percentages for professional experience are based on participants with available data (n=50).

Phase 1: Initial Survey Results (RQ1; Hypotheses 1-3)

Descriptive Statistics and Measurement Properties

Table 3 presents descriptive statistics and internal consistency estimates for the phase 1 measures. Cronbach α values for the reflective multi-item scales ranged from 0.71 to 0.82. For the domain-level IM index, Cronbach α was 0.55 and is reported descriptively rather than as an indicator of scale homogeneity. Complete item-level descriptive statistics and interitem correlations for the PI and IM measures are provided in Multimedia Appendix 3.

Table 3. Descriptive statistics and internal consistency of main survey measures (phase 1).

Variable	For Cronbach α[a], n	Items, n	Cronbach α	Mean (SD)	Range
Perceived AI impact on professional identity	141	6	0.71	3.13 (0.72)	1-5
AI importance across medical domains	135	5	0.55	2.38 (0.39)	1-3
Factual-level expectations regarding AI-related changes in medical practice	157	6	0.79	3.96 (0.64)	1-5
Social-level expectations regarding AI-related changes in medical practice	156	9	0.82	3.14 (0.65)	1-5

[a]Differences in ns reflect item-level nonresponse. ns for Cronbach α are based on complete cases using listwise deletion; item-level ns are reported in Multimedia Appendix 3.

To provide a more detailed view of the response patterns, selected item-level findings are reported below. Within the IM index, participants attributed the highest importance to the simplification of administrative tasks (mean 2.73, SD 0.48) and improvements in diagnostic accuracy and efficiency (mean 2.50, SD 0.50). This was followed by support in complex diagnostic cases (mean 2.39, SD 0.68) and the development of personalized treatment planning (mean 2.32, SD 0.70). In contrast, the perceived importance of AI for enhancing doctor-patient communication was comparatively lower (mean 1.93, SD 0.71).

Participants reported varying levels of agreement regarding the potential impact of AI on their professional identity. The highest agreement was observed for the statement that AI will require continuous updates of medical skills to remain relevant (mean 3.93, SD 0.98). Moderate levels of agreement were found for concerns related to reduced confidence in clinical skills (mean 3.14, SD 1.30) and the possibility that AI errors could undermine clinical judgment (mean 3.05, SD 1.21). Lower levels of agreement were observed for items referring to a loss of uniqueness of medical expertise (mean 2.82, SD 1.13),

XSL•FO
RenderX

weakened patient trust (mean 2.83, SD 1.19), and potential threats to professional reputation due to legal or ethical issues (mean 3.00, SD 1.17). Participants reported high agreement with EC_FL, particularly for real-time monitoring and early detection (mean 4.15, SD 0.87) and more efficient follow-up care (mean 4.08, SD 0.87). Expectations regarding diagnostic processes were also positive, as reflected in quicker diagnoses (mean 4.06, SD 0.82) and improved diagnostic accuracy (mean 3.90, SD 0.77). In contrast, expectations regarding social-level changes were more moderate. Lower agreement was observed for improvements in doctor-patient communication (mean 2.91, SD 1.13), while aspects such as patient understanding (mean 3.20, SD 1.16), clarity of medical information (mean 3.21, SD 1.04), and data exchange with colleagues (mean 3.62, SD 1.04) showed moderate agreement.

The following sections report the results of the hypothesis-driven analyses (hypotheses 1-3) examining associations between AI-related expectations, perceived importance, and PI, as well as group differences between medical students and clinicians.

Professional Identity Threat, AI-Related Perceptions, and Group Differences (RQ 1)

Hypothesis Overview

Hypothesis 1 focused on whether positive expectations about AI-related changes in medical practice were associated with perceived professional identity threat. Hypothesis 2 addressed a related but distinct question: whether perceiving AI as important across medical domains was associated with the acceptance indicator and identity threat. After examining the association-based hypotheses, we tested in hypothesis 3 whether medical students and clinicians differed in their EC_FL and EC_SL.

Hypothesis 1: Factual-Level AI Expectations and Perceived Identity Threat

To test hypothesis 1, we examined whether more positive EC_FL were associated with lower perceived professional identity threat. A hierarchical linear regression was conducted to examine predictors of perceived professional identity threat. In the final model, including factual-level AI expectations, group membership, age, and prior AI experience, the overall model was significant (adjusted R^2=0.10; $F_{4,124}$=4.59; P=.002; Table S1 in Multimedia Appendix 5). However, factual-level AI expectations were not a significant predictor of perceived identity threat (β=–.14; P=.11). Accordingly, hypothesis 1 was not supported. In the same model, group membership was significantly associated with perceived identity threat (β=–.31; P=.005), with clinicians reporting lower perceived identity threat than medical students. This was consistent with the unadjusted group comparison, in which medical students reported significantly higher perceived identity threat than clinicians (t_{156}=4.02; P<.001; Cohen d=0.69). Age (β=–.06; P=.58) and prior AI experience (β=.07; P=.45) were not significantly associated with perceived identity threat. Collinearity diagnostics did not indicate problematic multicollinearity among the predictors (variance inflation factors=1.02-1.70; tolerance=0.59-0.99).

Hypothesis 2: Perceived Importance of AI Across Medical Domains

To test hypothesis 2, we examined whether IM was associated with 2 outcomes: an acceptance-related indicator, operationalized as EC_FL, and perceived professional identity threat. Pearson correlations indicated that IM was strongly associated with more positive factual-level AI expectations (r=0.68; P<.001), whereas no significant association was observed with perceived professional identity threat (r=–0.09; P=.29; Table S2 in Multimedia Appendix 5).

In a multiple linear regression model predicting the factual-level AI expectations, IM emerged as a strong and significant predictor (β=.71; P<.001), whereas group membership and prior AI experience were not significant predictors. In contrast, in a second regression model predicting perceived professional identity threat, IM was not a significant predictor (β=–.12; P=.17). Group membership remained a significant predictor, with clinicians reporting lower perceived identity threat than medical students. Thus, the expectation-related component of hypothesis 2 was supported, whereas the identity-threat component was not supported. Full regression results are provided in Table S3 in Multimedia Appendix 5.

Hypothesis 3: Group Differences in Factual-Level and Social-Level AI Expectations

To test hypothesis 3, we examined whether medical students and clinicians differ in their EC_FL and EC_SL. Two-tailed independent-samples t tests revealed no significant group differences for either dimension. For the factual-level expectations, medical students (mean 3.96, SD 0.62) and clinicians (mean 3.97, SD 0.66) did not differ (t_{156}=–0.12; P=.91; Cohen d=–0.02). Likewise, no significant difference was observed for social-level expectations between medical students (mean 3.16, SD 0.63) and clinicians (mean 3.12, SD 0.68; t_{155}=0.43; P=.67; Cohen d=0.07). Accordingly, hypothesis 3 was not supported.

Phase 2: Quasi-Randomized Vignette Study (RQ2; Hypothesis 4)

Participant Allocation and Group Characteristics

The vignette study examined clinicians' evaluations of AI-supported clinical decision-making across 3 distinct scenarios (Watson, Triage, and OncoGuide). A total of 68 clinicians participated and were quasi-randomly assigned to 1 of the 3 scenarios: Watson (n=24), Triage (n=21), and OncoGuide (n=23). The 3 vignette groups did not differ significantly in clinical specialty or prior experience with AI in medicine (Multimedia Appendix 6).

Vignette Analyses Related to Hypothesis 4: Explainability, Trustworthiness, and Responsibility Across Vignette Contexts

Phase 2 examined whether clinicians' evaluations of explainability, trustworthiness, and responsibility differed across the 3 vignette scenarios. These analyses focused on the 3 evaluative dimensions specified in hypothesis 4. As shown in Table 4, explainability and responsibility were rated highly across scenarios, whereas trustworthiness received

XSL•FO
RenderX

comparatively lower ratings. A one-way ANOVA showed a significant effect of vignette condition on perceived explainability ($F_{2,65}$=5.41; P=.007; η^2=0.14). Bonferroni-corrected post hoc tests indicated that explainability was rated lower in the Triage vignette (mean 3.92, SD 0.87) than in the Watson vignette (mean 4.47, SD 0.64; P=.03) and the OncoGuide vignette (mean 4.54, SD 0.50; P=.01). No difference was observed between the Watson and OncoGuide vignettes (P>.99). No significant differences between vignette conditions were observed for trustworthiness ($F_{2,65}$=0.39; P=.68; η^2=0.01) or responsibility ($F_{2,65}$=1.01; P=.37; η^2=0.03). A sensitivity power analysis indicated that, assuming α=.05 and 80% power, the phase 2 design with n=68 across 3 vignette groups could detect effects of approximately Cohen f=0.39, corresponding to $\eta^2 \approx 0.13$ using $\eta^2 = f^2/(1+f^2)$. Thus, the analyses were primarily powered to detect comparatively large context effects, and smaller effects may have remained undetected. Taken together, only perceived explainability differed significantly across vignette contexts in the present sample.

Table 4. Mean (SD) for perceived explainability, trustworthiness, and responsibility across vignettes.

Dimension	Watson (n=24), mean (SD)	Triage (n=21), mean (SD)	OncoGuide (n=23), mean (SD)	Total (n=68), mean (SD)	Cronbach α[a]	*F* test (*df*=2, 65)	*P* value	η^2
Explainability	4.47 (0.64)	3.92 (0.87)	4.54 (0.50)	4.32 (0.72)	0.59	5.41	.007	0.14
Trustworthiness	3.11 (0.98)	3.21 (0.76)	3.33 (0.82)	3.22 (0.86)	0.48	0.39	.68	0.01
Responsibility	4.36 (0.75)	4.16 (0.68)	4.45 (0.63)	4.33 (0.69)	0.35	1.01	.37	0.03

[a]Cronbach α by vignette (Watson/Triage/OncoGuide): explainability: 0.54/0.69/0.12; trustworthiness: 0.50/0.61/0.42; and responsibility: 0.36/0.39/0.26.

Prioritization Patterns Across Vignette Scenarios

Descriptive analyses examined how clinicians prioritized explainability, trustworthiness, and responsibility when these dimensions were in tension. Although the wording of the prioritization items was adapted to the respective scenario, the underlying dimension contrasts were identical across vignettes: agreement was highest for prioritizing explainability over trust, followed by prioritizing responsibility over explainability, while prioritizing trust over explainability received the lowest agreement (Table 5).

Table 5. Mean (SD) for prioritization trade-off items by vignettea.

Prioritization trade-off	Watson (n=24), mean (SD)	Triage (n=21), mean (SD)	OncoGuide (n=23), mean (SD)
Trust>explainability (time pressure/patient trust)	2.42 (1.18)	2.48 (1.18)	2.65 (1.27)
Responsibility>explainability (clear accountability)	3.33 (1.17)	3.10 (1.22)	3.52 (1.24)
Explainability>trust (transparency preference)	4.00 (1.18)	3.57 (1.03)	3.73 (1.42)

[a]Values represent mean agreement scores on a 5-point Likert scale (1=strongly disagree and 5=strongly agree). Items capture trade-offs between the perceived importance of trustworthiness, explainability, and responsibility in different clinical contexts.

Finally, exploratory analyses examined associations between individual dimension items and vignette-specific contextual items within each scenario using Spearman rank correlations. Only statistically significant correlations are reported in Multimedia Appendix 6. These analyses were exploratory and not part of the hypothesis-driven testing.

Discussion

Summary of Key Findings and Methodological Considerations

The study revealed a mix of significant and nonsignificant findings in both phases. In phase 1, participants reported positive EC_FL, while perceived professional identity threat remained moderate. Factual-level AI expectations were not significantly associated with perceived identity threat. IM was strongly associated with factual-level AI expectations, which served as an acceptance-related indicator, but it was not associated with perceived identity threat. Thus, the expectation-related component of hypothesis 2 was supported, whereas the identity-threat component was not supported.

This hypothesis 2 finding should be interpreted with some caution. The IM index covered different medical application areas rather than a narrowly defined construct, which is reflected in its low internal consistency. The strong association between this index and factual-level AI expectations may therefore partly reflect overlap between the measures, as both capture positive evaluations of AI's role in practice. At the same time, the absence of an association with identity threat should not be taken as clear evidence that perceived importance and identity-related concerns are unrelated, as correlations may be weaker when reliability is low.

Hypothesis 3 was not supported, as medical students and clinicians did not differ significantly in EC_FL or EC_SL. The observed group difference in perceived identity threat should therefore be considered separately, as it emerged from the regression models predicting identity threat rather than from the hypothesis 3 test itself. In these models, group membership was significantly associated with identity threat, whereas factual-level expectations were not. In phase 2, clinicians rated explainability and responsibility highly across vignette scenarios,

XSL•FO
RenderX

while trustworthiness received somewhat lower ratings. Explainability differed significantly between vignette contexts and was rated lower in the Triage vignette, which involved time pressure alongside other emergency decision-making features. Trustworthiness and responsibility did not differ significantly across scenarios. Because phase 2 did not include a separate acceptance outcome, these findings should not be interpreted as direct support for the broader preregistered acceptance-based formulation of hypothesis 4.

Before drawing theoretical conclusions, several methodological aspects should be taken into account. The null findings regarding identity threat may partly reflect measurement and design limitations. In addition, the phase 1 regression model explained only a modest share of variance in perceived identity threat, and the cross-sectional design does not allow conclusions about causal direction. For phase 2, the small vignette-specific subsamples limited statistical power to detect smaller context effects, particularly for trustworthiness and responsibility. The phase 2 composites comprised 3 items each that captured related but distinct aspects of explainability, trustworthiness, and responsibility. The comparatively low internal consistency estimates should therefore be interpreted in light of both the small number of items and the conceptual breadth of the dimensions. The phase 2 sample was also drawn from a single medical didactic qualification program, which may limit transferability to broader clinician populations. Therefore, the nonsignificant findings should not be interpreted as evidence of no association or no context effect, but rather as indicating that these associations and differences were not observed in the present sample and with the present measures.

Divergence Between Task-Related Expectations and Identity-Related Concerns

Building on the methodological considerations outlined earlier, one plausible interpretation of the phase 1 findings is that task-related expectations and identity-related concerns refer to different evaluative layers of AI in medicine. The findings do not show that these dimensions are generally independent. Rather, they indicate that, in the present sample, positive factual-level expectations and IM were not associated with lower professional identity threat. A social constructivist perspective helps interpret this pattern without treating AI as a technology with fixed implications. From this perspective, the meaning of AI is negotiated within professional contexts and in relation to established understandings of expertise, decision-making authority, and responsibility. Building on insights from the social construction of technology, this interpretation focuses not only on evaluations of AI's potential benefits but also on how these evaluations intersect with professional identity and role boundaries [12].

In phase 1, participants reported positive expectations regarding AI-related changes in medical practice, while perceived identity threat remained moderate and was not significantly associated with these expectations. Hypothesis 1 was therefore not supported. Similarly, perceiving AI as important across medical domains was strongly associated with more positive factual-level expectations (hypothesis 2), but it was not associated with lower identity threat. Thus, the expectation-related component of hypothesis 2 was supported, whereas the identity-threat component was not. This asymmetry suggests that perceiving AI as useful or important for clinical tasks does not necessarily resolve questions of professional role, judgment, and legitimacy. From a social-constructivist perspective, this pattern is plausible because AI is not interpreted only through its expected technical performance, but through the professional meanings attached to its use [35]. Participants could expect AI to improve diagnosis, monitoring, or workflow while still being concerned about how AI may affect clinical judgment, responsibility, or the perceived legitimacy of medical expertise. This distinction is relevant in light of prior research on AI adoption, which often highlights perceived usefulness as a key driver of acceptance [36]. Our findings support this view only for the expectation-related component: IM was associated with more positive factual-level expectations, but not with lower identity-related concerns. This is consistent with studies showing that health care professionals' acceptance of AI depends not only on perceived usefulness but also on whether AI can be integrated into existing clinical practice and responsibility structures [37]. In this sense, identity-related concerns appear to reflect tensions around core elements of medical professionalism, including competence and clinical judgment rather than AI's perceived usefulness alone [38]. This interpretation should be read in light of the methodological considerations outlined earlier. The findings do not provide confirmatory evidence for a stable decoupling between task-related expectations and identity-related concerns. Rather, they suggest that, in this sample, positive expectations toward AI did not necessarily coincide with lower identity threat.

Within this cautious interpretation, the item-level pattern helps specify what these identity-related concerns referred to. Participants reported high agreement with the need to continuously update medical skills, alongside concerns about reduced confidence in clinical skills and the potential for AI errors to undermine clinical judgment. These findings suggest that identity-related concerns may be tied to the conditions under which expertise is enacted and evaluated, rather than to attitudes toward AI per se. This is in line with prior work, showing that design and integration features of AI systems, including explainability, workflow integration, and accountability structures, can shape trust and identity-related concerns in partly nonparallel ways [13]. It also corresponds to research conceptualizing identity threat as a mechanism influencing technology use rather than a by-product of perceived usefulness [39]. Finally, the coexistence of strong expectations and persistent concerns resonates with prior stakeholder research, showing that AI in health care is framed both as a promise and as a source of uncertainty and ethical concern, reflecting implementation as a process of negotiated meaning rather than linear adoption [40]. Future studies should test this interpretation more directly using larger and more diverse samples, longitudinal or experimental designs, and refined measures of perceived AI importance across medical domains. Such designs would help distinguish whether the observed pattern reflects a divergence between task-related expectations and identity-related concerns, limited statistical power, measurement limitations, or sample-specific response patterns.

XSL•FO
RenderX

Group Differences in Perceived Identity Threat

Hypothesis 3 was not supported, as medical students and clinicians did not differ significantly in EC_FL or EC_SL. Medical students and clinicians reported similarly positive EC_FL, particularly in relation to diagnostic processes and monitoring. However, medical students reported higher overall professional identity threat than clinicians, indicating greater concern about how AI may affect their developing sense of medical expertise and professional role. Thus, the observed group difference in identity threat was not accompanied by more negative expectations regarding AI's practical contributions among medical students. Group membership remained significantly associated with identity threat after adjustment for age and prior AI experience. However, medical students and clinicians differed substantially in age and professional experience, and the present design does not allow these group-related characteristics to be clearly disentangled. Professional socialization therefore represents one possible interpretation of the observed group difference rather than an established explanatory mechanism. Professional identity in medicine is grounded in socially recognized expectations regarding expertise, decision authority, and responsibility. These expectations are acquired through professional socialization and stabilized in clinical practice, making questions of legitimate judgment and accountability central to the physician role [41]. From this perspective, AI does not simply introduce a new tool but may become relevant insofar as it interacts with these central aspects of professional practice. Professional identity formation (PIF) offers a useful perspective for interpreting the observed group difference. PIF conceptualizes medical training as a socialization process through which learners come to "think, act, and feel like a physician" [42,43]. In earlier stages of this process, expectations regarding legitimate expertise, authority, and professional responsibility are still being formed and negotiated. This interpretation is consistent with research on PIF, which emphasizes that during early stages of training, learners are particularly attentive to signals that define competence, authority, and legitimate participation within a profession [44,45]. Medical students' greater concerns regarding their professional identity in relation to AI may therefore reflect heightened sensitivity to its implications for their evolving professional role, rather than generally more negative expectations toward AI. Previous research likewise suggests that positive expectations of medical AI may coexist with concerns about professional roles and responsibilities [27], while AI-supported decision-making may challenge established forms of clinical expertise, epistemic authority, and accountability [28]. Clinicians, in contrast, may interpret AI from within more established clinical role and responsibility structures [46]. However, the present study cannot determine whether the observed group difference is related to professional socialization, age, professional experience, or other differences between the groups. Prior work further suggests that sociotechnical arrangements can shape identity-related responses without necessarily being mirrored in other evaluations of AI [13]. AI use may also have implications for professional reputation. Recent findings suggest that individuals who use AI at work can be perceived as less competent or motivated, illustrating how the use of AI can become relevant not only for task performance but also for how professional competence is socially evaluated [47].

Contextual Construction of Trust, Explainability, and Responsibility

Phase 2 examined clinicians' evaluations of explainability, trustworthiness, and responsibility across 3 AI-supported clinical decision contexts, corresponding to the analysis-aligned formulation of hypothesis 4. The broader preregistered acceptance-based formulation and its operationalization are described in the Data Processing and Statistical Analysis section. The clearest difference was observed for explainability. Explainability was rated lower in the Triage vignette than in the Watson and OncoGuide vignettes. The Triage vignette described an emergency decision-making situation involving time pressure, diagnostic uncertainty, team-based assessment, and the need for rapid prioritization. In this context, full transparency of the AI's reasoning may have appeared somewhat less central than in the other 2 scenarios. However, explainability was still rated highly overall, including in the Triage vignette. The lower explainability rating in the Triage vignette can be understood in relation to the different clinical contexts in which the same standardized explainability items were embedded. Across all scenarios, these items assessed whether clinicians considered it important to fully understand the basis of the AI's decision, whether a comprehensible decision process increased their confidence in implementing the AI recommendation, and whether lack of transparency made accepting the recommendation more difficult. In the Watson and OncoGuide vignettes, the AI recommendation deviated from existing clinical guidelines. These scenarios also more explicitly involved questions of justification, responsibility, and, in the OncoGuide vignette, patient-facing shared decision-making. Explainability may therefore have been more closely linked to justification and defensibility in these 2 scenarios than in the Triage vignette. At the same time, the vignettes differed in several features simultaneously. The observed difference can therefore not be attributed to time pressure alone, but should be interpreted as a difference between vignette contexts. This fits with prior work, arguing that explanations are most relevant when they support clinical reasoning and accountable decision-making in context, rather than exhaustive transparency [6,48]. Responsibility was rated highly across all vignettes and did not differ significantly between scenarios. This may suggest that clinicians continued to locate accountability within the physician role even when AI systems were involved. However, phase 2 participants were recruited from a medical didactics qualification program and may therefore have been particularly attentive to questions of responsibility and reflective practice [49]. This may have contributed to the uniformly high responsibility ratings across the 3 vignette scenarios. These findings should therefore not be taken as evidence that responsibility is generally independent of clinical context. Smaller effects may have remained undetected. Nevertheless, the strong emphasis on physician responsibility is consistent with current governance and regulatory debates in health AI, which emphasize effective human oversight [50]. Trustworthiness was rated comparatively lower and showed less variation across scenarios. The exploratory correlation analyses suggest that trust may depend

XSL•FO
RenderX

not only on general attitudes toward AI but also on the specific clinical situation and professional factors. This is consistent with review-based work, showing that trust in AI-based clinical decision support is shaped by factors such as system validation, workflow integration, transparency, and user experience [30]. Given the exploratory nature of these analyses and the small vignette-specific subsamples, these findings should also be interpreted cautiously. While these issues are often discussed as technical challenges, such as balancing model performance, explainability, and robustness [20,51], the vignette findings suggest a context-specific perspective. This raises the question of under which clinical conditions, explainability, trustworthiness, and responsibility become particularly relevant for clinicians' acceptance of AI-supported recommendations.

Implications for AI Implementation and Medical Education

Overall, the findings suggest that integrating AI into clinical practice calls for attention not only to performance but also to the role-related questions that AI raises around how decisions are justified and who remains accountable. In phase 1, participants reported positive expectations regarding AI's practical contributions, while professional identity concerns remained present. This suggests that implementation efforts may benefit from addressing how responsibility, decision-making, and accountability are conceptualized in AI-supported care, rather than relying on benefit-oriented narratives alone [52]. For medical education, students reported higher perceived identity threat than clinicians despite similarly positive expectations. One possible interpretation is that, during professional socialization, AI may be encountered not only as a tool but also in relation to emerging understandings of valued expertise and legitimate decision authority in the physician role. These findings suggest that AI literacy may need to be complemented by educational approaches that address how AI-supported decisions are justified and how responsibility is assigned in clinical practice [43]. Phase 2 further indicates that explainability may need to be considered in a context-sensitive manner. Explainability differed across scenarios, suggesting that what counts as a sufficient explanation may depend on whether decisions must be justified against guidelines, made under time pressure, or discussed in patient-facing contexts. Taken together, these findings may inform context-sensitive approaches to oversight and justification in AI-supported decision-making, consistent with sociotechnical implementation perspectives [9].

Limitations

Phase 1 was cross-sectional and therefore cannot show how expectations toward AI and perceived professional identity threat develop over time or influence each other causally. In addition, the index of perceived importance of AI across medical domains covered heterogeneous areas of medical practice and showed low internal consistency. Findings involving this index should therefore be read as referring to perceived importance across selected domains rather than to a single homogeneous construct. The 2 phases were conceptually linked but empirically distinct. Phase 1 examined general AI-related expectations and identity-related concerns in a sample of medical students and clinicians, whereas phase 2 examined vignette-based evaluations of explainability, trustworthiness, and responsibility in a separate sample. Phase 2 should therefore be understood as a complementary extension addressing RQ2, not as a direct empirical follow-up of phase 1.

Several limitations also concern the vignette study. The vignette-specific subsamples were small, which limited the ability to detect smaller differences between scenarios, particularly for trustworthiness and responsibility. In addition, participants were recruited from a single medical didactic qualification program. This recruitment context may have selected participants with a particular interest in teaching, reflection, and professional responsibility, which may have contributed to the high responsibility ratings across vignettes and limits transferability to broader clinician populations. Finally, phase 2 did not include a comparable medical student sample, so the findings cannot show how students would evaluate explainability, trustworthiness, and responsibility in similar vignette contexts.

No additional composite acceptance score was constructed in phase 2. The vignette findings therefore describe context-specific acceptance-related evaluations of explainability, trustworthiness, and responsibility, but should not be interpreted as a regression-based test of a global acceptance outcome. Because the study was conducted in Germany, the findings may also reflect specific institutional, educational, and medico-legal conditions and may not be directly transferable to other health care systems.

Conclusions

This study suggests that positive expectations about AI's task-related benefits do not necessarily coincide with lower professional identity concerns. In phase 1, both medical students and clinicians expressed largely positive EC_FL, and these expectations did not differ between groups. IM was associated with more positive factual-level AI expectations, but not with lower perceived professional identity threat. In addition, medical students reported higher perceived professional identity threat than clinicians.

In phase 2, which focused on clinicians, only perceived explainability differed significantly across vignette contexts. Explainability was rated lower in the Triage vignette than in the Watson and OncoGuide vignettes. Trustworthiness and responsibility did not differ significantly across vignettes; however, given the small vignette-specific subsamples, these nonsignificant findings should not be interpreted as evidence that these dimensions are context-invariant. Taken together, the findings suggest that the evaluation of AI-supported decision-making cannot be inferred from expected performance benefits alone. They may also depend on how AI is positioned in relation to clinical judgment, responsibility, and the justification of decisions in specific clinical contexts. This supports the view that AI implementation in health care is not only a technical process, but also raises questions about clinical expertise, decision-making, and responsibility.

XSL•FO
RenderX

Acknowledgments

The authors acknowledge the support of the Open Access Publishing Funds of the University of Tübingen. Generative AI tools (ChatGPT-5 and DeepL Write) were used in the preparation of this manuscript to improve writing quality and assist with the visual refinement of the study outline figure. After using these tools, the authors thoroughly reviewed, revised, and edited the content and take full responsibility for the final version of the manuscript.

Data Availability

The data that support the findings of this study are available from the corresponding author (JAM) upon reasonable request.

Funding

The authors thank the Federal Ministry of Education and Research, Germany (BMBF) for supporting this project (16DHBKI086). The authors acknowledge support via financing publication fees from “Deutsche Forschungsgemeinschaft.”

Authors' Contributions

JAM was responsible for designing and conducting the study, including the acquisition, analysis, and interpretation of data. AH-W and TF-W contributed to data analysis, interpretation of the findings, and critical revision of the manuscript. UK supported the validation of the statistical analyses. JAM analyzed the research material and drafted the manuscript. KN made substantial contributions to the study design and critically revised the manuscript. SW, MC, and SZ contributed to the broader project context and critically revised the manuscript. All authors approved the final version of the manuscript and agreed to be accountable for all aspects of the work.

Conflicts of Interest

The software and survey server for conducting the online survey were provided free of charge by the SoSci Survey GmbH. The SoSci Survey GmbH had no influence on the content of the study. The authors declare that there is no conflict of interest.

Multimedia Appendix 1
CHERRIES checklist.
[DOCX File , 19 KB - mededu_v12i1e104151_app1.docx]

Multimedia Appendix 2
Initial survey (full item list).
[DOCX File , 22 KB - mededu_v12i1e104151_app2.docx]

Multimedia Appendix 3
Item-level descriptive statistics and interitem correlations.
[DOCX File , 24 KB - mededu_v12i1e104151_app3.docx]

Multimedia Appendix 4
Vignette scenarios and item wordings (phase 2: vignette study).
[DOCX File , 23 KB - mededu_v12i1e104151_app4.docx]

Multimedia Appendix 5
Supplementary statistical analyses for phase 1.
[DOCX File , 19 KB - mededu_v12i1e104151_app5.docx]

Multimedia Appendix 6
Additional phase 2 analyses and vignette group characteristics.
[DOCX File , 21 KB - mededu_v12i1e104151_app6.docx]

References

1. Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. Nat Med 2019;25(1):44-56. [doi: 10.1038/s41591-018-0300-7] [Medline: 30617339]
2. Cavallo JJ, de Oliveira Santo I, Mezrich JL, Forman HP. Clinical implementation of a combined artificial intelligence and natural language processing quality assurance program for pulmonary nodule detection in the emergency department setting. J Am Coll Radiol 2023;20(4):438-445. [doi: 10.1016/j.jacr.2022.12.016] [Medline: 36736547]

XSL•FO
RenderX

3. Lotter W, Hassett MJ, Schultz N, Kehl KL, Van Allen EM, Cerami E. Artificial intelligence in oncology: current Landscape, challenges, and future directions. Cancer Discov 2024;14(5):711-726. [doi: 10.1158/2159-8290.CD-23-1199] [Medline: 38597966]
4. Ogut E. Artificial intelligence in clinical medicine: challenges across diagnostic imaging, clinical decision support, surgery, pathology, and drug discovery. Clin Pract 2025;15(9):169 [FREE Full text] [doi: 10.3390/clinpract15090169] [Medline: 41002784]
5. Tschochohei M, Adams LC, Bressem KK, Lammert J. AI-enabled clinical decision support systems: challenges and opportunities. Bundesgesundheitsblatt Gesundheitsforschung Gesundheitsschutz 2025;68(8):872-879. [doi: 10.1007/s00103-025-04092-8] [Medline: 40560226]
6. Gerdes A. The role of explainability in AI-supported medical decision-making. Discov Artif Intell 2024;4(1):29. [doi: 10.1007/s44163-024-00119-2]
7. Arbelaez Ossa L, Starke G, Lorenzini G, Vogt JE, Shaw DM, Elger BS. Re-focusing explainability in medicine. Digit Health 2022;8:20552076221074488 [FREE Full text] [doi: 10.1177/20552076221074488] [Medline: 35173981]
8. Sharma N, Sugandh U, Yadav S, Rastogi D, Agarwal J, Panwar A. The future of explainable AI in clinical decision-making: challenges and research directions. In: Panwar A, Jain A, Mallik S, Li A, Cengiz K, editors. Explainable AI in Clinical Practice. New York: Academic Press/Elsevier; 2026:477-492.
9. Salwei ME, Carayon P. A sociotechnical systems framework for the application of artificial intelligence in health care delivery. J Cogn Eng Decis Mak 2022;16(4):194-206 [FREE Full text] [doi: 10.1177/15553434221097357] [Medline: 36704421]
10. Chomutare T, Tejedor M, Svenning TO, Marco-Ruiz L, Tayefi M, Lind K, et al. Artificial intelligence implementation in healthcare: a theory-based scoping review of barriers and facilitators. Int J Environ Res Public Health 2022;19(23):16359 [FREE Full text] [doi: 10.3390/ijerph192316359] [Medline: 36498432]
11. Wenderott K, Krups J, Weigl M, Wooldridge AR. Facilitators and barriers to implementing AI in routine medical imaging: systematic review and qualitative analysis. J Med Internet Res 2025;27:e63649 [FREE Full text] [doi: 10.2196/63649] [Medline: 40690758]
12. Ta'an W, Damrah S, Al-Hammouri MM, Williams B. Professional identity and its relationships with AI readiness and interprofessional collaboration. PLoS One 2025;20(5):e0322794 [FREE Full text] [doi: 10.1371/journal.pone.0322794] [Medline: 40378129]
13. Ackerhans S, Wehkamp K, Petzina R, Dumitrescu D, Schultz C. Perceived trust and professional identity threat in AI-based clinical decision support systems: scenario-based experimental study on AI process design features. JMIR Form Res 2025;9:e64266 [FREE Full text] [doi: 10.2196/64266] [Medline: 40138691]
14. Moldt J, Festl-Wietek T, Fuhl W, Zabel S, Claassen M, Wagner S, et al. Exploring the social dimensions of AI integration in healthcare: a qualitative study of stakeholder views on challenges and opportunities. BMJ Open 2025;15(6):e096208 [FREE Full text] [doi: 10.1136/bmjopen-2024-096208] [Medline: 40578890]
15. Coiera E. The last mile: where artificial intelligence meets reality. J Med Internet Res 2019;21(11):e16323 [FREE Full text] [doi: 10.2196/16323] [Medline: 31702559]
16. Mosch L, Fürstenau D, Brandt J, Wagnitz J, Klopfenstein SA, Poncette A, et al. The medical profession transformed by artificial intelligence: qualitative study. Digit Health 2022;8:20552076221143903 [FREE Full text] [doi: 10.1177/20552076221143903] [Medline: 36532112]
17. Pinch T, Bijker W. The social construction of facts and artefacts: or how the sociology of science and the sociology of technology might benefit each other. Soc Stud Sci 1984 Aug 01;14(3):399-441. [doi: 10.1177/030631284014003004]
18. Weber S, Wyszynski M, Godefroid M, Plattfaut R, Niehaves B. How do medical professionals make sense (or not) of AI? A social-media-based computational grounded theory study and an online survey. Comput Struct Biotechnol J 2024;24:146-159 [FREE Full text] [doi: 10.1016/j.csbj.2024.02.009]
19. Turinici M. AI in medicine: decision-making and relational vulnerability of healthcare professionals. Health 2025;17(10):1329-1348. [doi: 10.4236/health.2025.1710088]
20. Rasheed K, Qayyum A, Ghaly M, Al-Fuqaha A, Razi A, Qadir J. Explainable, trustworthy, and ethical machine learning for healthcare: a survey. Comput Biol Med 2022;149:106043 [FREE Full text] [doi: 10.1016/j.compbiomed.2022.106043] [Medline: 36115302]
21. Sivaraman V, Bukowski L, Levin J, Kahn J, Perer A. Ignore, trust, or negotiate: understanding clinician acceptance of AI-based treatment recommendations in health care. 2023 Presented at: Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems; April 23-28, 2023; Hamburg, Germany. [doi: 10.1145/3544548.3581075]
22. Oh S, Kim JH, Choi S, Lee HJ, Hong J, Kwon SH. Physician confidence in artificial intelligence: an online mobile survey. J Med Internet Res 2019;21(3):e12422 [FREE Full text] [doi: 10.2196/12422] [Medline: 30907742]
23. Jussupow E, Spohrer K, Heinzl A. Identity threats as a reason for resistance to artificial intelligence: survey study with medical students and professionals. JMIR Form Res 2022;6(3):e28750 [FREE Full text] [doi: 10.2196/28750] [Medline: 35319465]

XSL•FO
RenderX

24. Sit C, Srinivasan R, Amlani A, Muthuswamy K, Azam A, Monzon L, et al. Attitudes and perceptions of UK medical students towards artificial intelligence and radiology: a multicentre survey. Insights Imaging 2020;11(1):14 [FREE Full text] [doi: 10.1186/s13244-019-0830-7] [Medline: 32025951]
25. Blease C, Kaptchuk TJ, Bernstein MH, Mandl KD, Halamka JD, DesRoches CM. Artificial intelligence and the future of primary care: exploratory qualitative study of UK general practitioners' views. J Med Internet Res 2019;21(3):e12802 [FREE Full text] [doi: 10.2196/12802] [Medline: 30892270]
26. Pinto Dos Santos D, Giese D, Brodehl S, Chon SH, Staab W, Kleinert R, et al. Medical students' attitude towards artificial intelligence: a multicentre survey. Eur Radiol 2019;29(4):1640-1646. [doi: 10.1007/s00330-018-5601-1] [Medline: 29980928]
27. Amann J, Vayena E, Ormond KE, Frey D, Madai VI, Blasimme A. Expectations and attitudes towards medical artificial intelligence: a qualitative study in the field of stroke. PLoS One 2023;18(1):e0279088 [FREE Full text] [doi: 10.1371/journal.pone.0279088] [Medline: 36630325]
28. Grote T, Berens P. On the ethics of algorithmic decision-making in healthcare. J Med Ethics 2020;46(3):205-211 [FREE Full text] [doi: 10.1136/medethics-2019-105586] [Medline: 31748206]
29. Juravle G, Boudouraki A, Terziyska M, Rezlescu C. Chapter 14—Trust in artificial intelligence for medical diagnoses. In: Parkin BL, editor. Progress in Brain Research. Real-World Applications in Cognitive Neuroscience. Amsterdam, Netherlands: Elsevier; 2020.
30. Tun HM, Rahman HA, Naing L, Malik OA. Trust in artificial intelligence-based clinical decision support systems among health care workers: systematic review. J Med Internet Res 2025;27:e69678 [FREE Full text] [doi: 10.2196/69678] [Medline: 40772775]
31. Eysenbach G. Improving the quality of Web surveys: the Checklist for Reporting Results of Internet E-Surveys (CHERRIES). J Med Internet Res 2004;6(3):e34 [FREE Full text] [doi: 10.2196/jmir.6.3.e34] [Medline: 15471760]
32. Breakwell G. Identity process theory: clarifications and elaborations. In: Identity Process Theory: Identity, Social Action and Social Change. Cambridge: Cambridge University Press; 2014:20-38.
33. Moldt J. Quantitative survey research on the perception of the impact of artificial intelligence in medical practice. Open Science Framework. 2025. URL: https://osf.io/e6ryw/ [accessed 2026-08-26]
34. Faul F, Erdfelder E, Lang A, Buchner A. G*Power 3: a flexible statistical power analysis program for the social, behavioral, and biomedical sciences. Behav Res Methods 2007;39(2):175-191. [doi: 10.3758/bf03193146] [Medline: 17695343]
35. Martinez MA. A sociological professional identity. In: The Employable Sociologist: A Guide for Undergraduates. Cham: Springer International Publishing; 2023:17-33.
36. Dingel J, Kleine A, Cecil J, Sigl AL, Lermer E, Gaube S. Predictors of health care practitioners' intention to use AI-Enabled clinical Decision support systems: meta-analysis based on the unified theory of acceptance and use of technology. J Med Internet Res 2024;26:e57224 [FREE Full text] [doi: 10.2196/57224] [Medline: 39102675]
37. Hua D, Petrina N, Young N, Cho J, Poon SK. Understanding the factors influencing acceptability of AI in medical imaging domains among healthcare professionals: a scoping review. Artif Intell Med 2024;147:102698 [FREE Full text] [doi: 10.1016/j.artmed.2023.102698] [Medline: 38184343]
38. Mackintosh N, Armstrong N. Understanding and managing uncertainty in health care: revisiting and advancing sociological contributions. Sociol Health Illn 2020;42 Suppl 1:1-20. [doi: 10.1111/1467-9566.13160] [Medline: 32757281]
39. Mirbabaie M, Brünker F, Möllmann Frick NRJ, Stieglitz S. The rise of artificial intelligence–understanding the AI identity threat at the workplace. Electron Mark 2021;32(1):73-99. [doi: 10.1007/s12525-021-00496-x]
40. Moldt J, Festl-Wietek T, Fuhl W, Zabel S, Claassen M, Wagner S, et al. Assessing AI awareness and identifying essential competencies: insights from key stakeholders in integrating AI Into medical education. JMIR Med Educ 2024;10:e58355 [FREE Full text] [doi: 10.2196/58355] [Medline: 38989834]
41. Pförtner T, Pfaff H. 3.3.2. Arztrolle. In: Deinzer R, von dem Knesebeck O, editors. Online Lehrbuch der Medizinischen Psychologie und Medizinischen Soziologie. Berlin: German Medical Science GMS Publishing House; 2020.
42. Sternszus R, Steinert Y, Razack S, Boudreau JD, Snell L, Cruess RL. Being, becoming, and belonging: reconceptualizing professional identity formation in medicine. Front Med 2024;11:1438082 [FREE Full text] [doi: 10.3389/fmed.2024.1438082] [Medline: 39257893]
43. Cruess RL, Cruess SR, Boudreau JD, Snell L, Steinert Y. A schematic representation of the professional identity formation and socialization of medical students and residents: a guide for medical educators. Acad Med 2015;90(6):718-725. [doi: 10.1097/ACM.0000000000000700] [Medline: 25785682]
44. Jarvis-Selinger S, MacNeil KA, Costello GRL, Lee K, Holmes CL. Understanding professional identity formation in early clerkship: a novel framework. Acad Med 2019;94(10):1574-1580. [doi: 10.1097/ACM.0000000000002835] [Medline: 31192797]
45. Cruess RL, Cruess SR, Steinert Y. Amending Miller's pyramid to include professional identity formation. Acad Med 2016;91(2):180-185. [doi: 10.1097/ACM.0000000000000913] [Medline: 26332429]
46. Lombi L, Rossero E. How artificial intelligence is reshaping the autonomy and boundary work of radiologists. A qualitative study. Sociol Health Illn 2024;46(2):200-218. [doi: 10.1111/1467-9566.13702] [Medline: 37573551]
47. Reif JA, Larrick RP, Soll JB. Evidence of a social evaluation penalty for using AI. Proc Natl Acad Sci USA 2025;122(19):e2426766122 [FREE Full text] [doi: 10.1073/pnas.2426766122] [Medline: 40339114]

XSL•FO
RenderX

48. Wysocki O, Davies J, Vigo M, Armstrong A, Landers D, Lee R, et al. Assessing the communication gap between AI models and healthcare professionals: explainability, utility and trust in AI-driven clinical decision-making. Artif Intell 2023;316:103839 [FREE Full text] [doi: 10.1016/j.artint.2022.103839] [Medline: 41550460]
49. Mahajan R, Goyal PK, Singh T. Professional identity formation of medical educators: a thematic analysis of enabling factors and competencies needed. Int J Appl Basic Med Res 2022;12(3):189-195 [FREE Full text] [doi: 10.4103/ijabmr.ijabmr_257_22] [Medline: 36131862]
50. Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act). European Union. 2024. URL: https://eur-lex.europa.eu/eli/reg/2024/1689/oj/eng [accessed 2026-09-02]
51. Li LT, Haley LC, Boyd AK, Bernstam EV. Technical/algorithm, stakeholder, and society (TASS) barriers to the application of artificial intelligence in medicine: a systematic review. J Biomed Inform 2023;147:104531 [FREE Full text] [doi: 10.1016/j.jbi.2023.104531] [Medline: 37884177]
52. Ackerhans S, Huynh T, Kaiser C, Schultz C. Exploring the role of professional identity in the implementation of clinical decision support systems—a narrative review. Implement Sci 2024;19(1):11 [FREE Full text] [doi: 10.1186/s13012-024-01339-x] [Medline: 38347525]

Abbreviations

CHERRIES: Checklist for Reporting Results of Internet E-Surveys
EC_FL: factual-level expectations regarding AI-related changes in medical practice
EC_SL: social-level expectations regarding AI-related changes in medical practice
IM: perceived importance of AI across medical domains
PI: perceived AI impact on professional identity
PIF: professional identity formation
RQ: research question

Edited by A Stone; submitted 09.Jun.2026; peer-reviewed by JJ Giraldo-Huertas, A Burchardt, N Ozaki; comments to author 08.Jul.2026; revised version received 31.Aug.2026; accepted 31.Aug.2026; published 30.Sep.2026.

Please cite as:
Moldt JA, Festl-Wietek T, Nieselt K, Zabel S, Claassen M, Wagner S, Keim U, Herrmann-Werner A
AI Perceptions, Professional Identity, and AI-Supported Clinical Decisions Among Medical Students and Clinicians: Cross-Sectional Survey and Quasi-Randomized Vignette Study
JMIR Med Educ 2026;12:e104151
URL: https://mededu.jmir.org/2026/1/e104151
doi:10.2196/104151
PMID:

©Julia-Astrid Moldt, Teresa Festl-Wietek, Kay Nieselt, Susanne Zabel, Manfred Claassen, Samuel Wagner, Ulrike Keim, Anne Herrmann-Werner. Originally published in JMIR Medical Education (https://mededu.jmir.org), 30.Sep.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Education, is properly cited. The complete bibliographic information, a link to the original publication on https://mededu.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

AI Literacy and Training Needs Among French Midwifery Students: National Cross-Sectional Online Survey and Exploratory Qualitative Follow-Up

Sylvain Kotzki[1], PharmD, PhD; Sonia Knepfler[2], MSc; Lionel Di Marco[1,3], DMid, PhD; Nicolas Vuillerme[1,4], PhD

[1]Centre de Recherche en Santé Intégrée, Université Grenoble Alpes, CNRS, Grenoble INP, LIG Sangria, 6 Chemin Saint-Ferjus, La Tronche, Grenoble, France

[2]Department of Midwifery, Université de Strasbourg, Strasbourg, France

[3]Department of Midwifery, Centre Hospitalier Universitaire Grenoble Alpes, Grenoble, France

[4]Institut Universitaire de France, Paris, France

Corresponding Author:
Sylvain Kotzki, PharmD, PhD
Centre de Recherche en Santé Intégrée, Université Grenoble Alpes, CNRS, Grenoble INP, LIG Sangria, 6 Chemin Saint-Ferjus, La Tronche, Grenoble, France

Abstract

Background: AI is increasingly relevant to health care and health professional education, but AI-related knowledge, perceptions, and training needs among midwifery students remain underdocumented. Evidence is needed to inform midwifery curricula and to clarify how students understand and operationalize AI in training and placements.

Objective: This cross-sectional online study aimed to assess AI-related knowledge, perceptions, perceived risks, and training needs among French midwifery students.

Methods: We conducted a national cross-sectional online survey of students enrolled in years 2 to 5 of the French midwifery curriculum between January and May 2025, complemented by exploratory qualitative follow-up interviews. Of 34 midwifery schools and universities contacted, 30 confirmed dissemination of the survey link through institutional and student-association channels. The voluntary, anonymous survey combined 5 previously published instruments or questionnaires administered in linguistically validated French versions. Likert-type items were analyzed descriptively on a 1 to 5 scale at the item level and through post hoc descriptive thematic summaries across 5 domains and 23 themes. Analyses were restricted to fully completed survey entries, with no imputation, inferential testing, or subgroup comparisons. Eight volunteer students from 1 school in eastern France participated in semistructured interviews, which were analyzed thematically to provide exploratory contextual observations.

Results: Of 414 initiated survey entries, 190 (45.9%) were fully completed and analyzed. Self-reported AI knowledge and familiarity were below the scale midpoint, with the lowest score observed for familiarity with advanced AI techniques (1.20, 95% CI 1.13 - 1.27). General AI familiarity was higher (n=91, 47.9% agreement) than familiarity with specific concepts such as machine learning (n=20, 10.5%) or deep learning (n=13, 6.8%). Perceived ability to use AI for clinical purposes was limited (2.05, 95% CI 1.93 - 2.16). In contrast, students strongly endorsed AI education, particularly the belief that students and professionals should be trained in AI (4.11, 95% CI 4.02 - 4.20). Prior exposure to AI (1.36, 95% CI 1.27 - 1.45) and perceived institutional and social support (1.43, 95% CI 1.34 - 1.52) were low. Ethics- and safety-related concerns were salient, with the highest score for evidence and safety requirements (4.04, 95% CI 3.96 - 4.12). Exploratory interviews suggested that students often referred to AI through everyday generative tools, reported little supervised exposure to AI in clinical placements, and emphasized safety, privacy, accountability, and preservation of human judgment.

Conclusions: Interpreted as a time-bound baseline, these findings suggest an AI readiness gap among French midwifery students, in which low perceived knowledge and limited clinical readiness coexist with strong demand for training and high salience of safety and governance. They support the progressive and proportionate introduction of AI literacy in midwifery education, explicitly distinguishing general AI, generative AI tools, and regulated clinical AI systems. Future work should develop and validate midwifery-specific AI literacy measures and evaluate curriculum interventions.

(JMIR Med Educ 2026;12:e94310) doi:10.2196/94310

KEYWORDS

midwifery education; AI; AI literacy; digital health; health professions education; cross-sectional survey; curriculum

XSL•FO
RenderX

Introduction

AI is increasingly relevant to health care, from diagnostic support to monitoring, prediction, documentation, and decision support [1,2]. In some narrow tasks, such as image-based classification, AI systems can reach levels of performance comparable to those of clinicians, although their broader use in care continues to raise questions about safety, transparency, accountability, and clinical validation [2,3]. For future clinicians, AI therefore represents not only a technological development but also an educational challenge: students need to understand what these systems do, when their outputs may or may not be trusted, and what professional responsibilities remain when AI tools are used in care [4]. Current evidence suggests that health professions students' knowledge, perceptions, and training needs regarding AI are heterogeneous. Students often express interest in AI and recognize its potential relevance for health care, while also reporting limited familiarity with core concepts, limited formal training, and concerns about reliability, professional responsibility, and the preservation of human judgment [5,6]. At present, however, AI-related competencies in health curricula remain heterogeneous, and structured training is still uneven despite emerging competency and curriculum frameworks [7,8].

In France, a mandatory digital health core curriculum was introduced for health students in 2022, and the 2024 reform of the midwifery curriculum incorporated digital health competencies; however, AI-specific learning outcomes remain limited or insufficiently explicit [9,10]. A further difficulty is that the term "AI" covers different realities that students may not clearly distinguish. General AI familiarity refers to broad awareness of AI as a field and its associated vocabulary. Generative AI tools, such as general-purpose conversational assistants, are widely available consumer applications that students may already use for study, writing, or clinical-placement tasks [11]. Regulated or clinically deployed AI systems, by contrast, are tools intended for use in care and are subject to clinical validation, professional governance, and regulatory oversight. These categories are easily conflated: familiarity with consumer generative AI tools may be mistaken for an understanding of clinically deployed AI systems. This ambiguity is important when interpreting students' self-reported AI knowledge, readiness, and training needs.

These questions are particularly relevant to midwifery education. AI applications are being explored in women's and perinatal health, including prenatal diagnostic support, monitoring, and risk prediction, although their current role in routine midwifery practice remains limited [12-14]. Midwifery is also distinctive because it is grounded in continuity of care, clinical judgment, and direct relationships with women and families. These features may shape how midwifery students perceive technologies that could affect communication, trust, and shared decision-making. Qualitative work with practicing midwives has reported both interest in the potential of AI and concerns about its effects on empathy, trust, and ethical decision-making in childbirth [15]. Authors have therefore argued that AI education should be integrated into nursing and midwifery training so that these professionals can contribute to safe, critical, and patient-centered use of AI [16-18]. However, midwives and midwifery students remain underrepresented in existing AI perception and readiness studies, which have focused mainly on medical students and physicians in countries such as Canada, the United States, and Turkey [19-21]. In addition, available instruments were mostly developed for medical students or broader health professional populations, and few data are available on how these domains apply to midwifery students. To our knowledge, no national study has examined AI-related knowledge, perceptions, risks, and training needs among midwifery students in France.

This study aimed to assess AI-related knowledge, perceptions, perceived risks, and training needs among French midwifery students through a national cross-sectional online survey, complemented by exploratory qualitative interviews to contextualize selected findings. The survey drew on previously published instruments or questionnaires assessing AI-related perceptions, readiness, learning intentions, and perceived risks in health professions education.

Methods

Study Design

We conducted a national cross-sectional online survey of midwifery students in France, complemented by an exploratory qualitative follow-up strand based on semistructured individual interviews. The survey provided the primary quantitative data. The qualitative strand was used to contextualize and illustrate selected survey findings. Given its exploratory, single-site, small-sample nature, no claim of data saturation is made. The study was conducted over 5 months, from January to May 2025.

Setting, Population, and Recruitment

The target population comprised all students enrolled in midwifery programs in France during the 2024 to 2025 academic year. In France, access to midwifery studies occurs after a first year of higher education through national health-studies entry pathways, either a health-specific access year (PASS [Specific Health Access Pathway]) or a health-access option within a university bachelor's program (L.AS). The midwifery-specific curriculum begins only after admission to the midwifery school; eligibility was therefore restricted to students enrolled in years 2 to 5 of the midwifery curriculum.

Program directors and faculty leads from 34 midwifery schools were contacted. Thirty institutions confirmed dissemination of the survey link to their students through institutional mailing lists and learning management platforms (n=30, 88.2%). We used voluntary convenience sampling through institutional and student association channels. To support completion, the survey was disseminated in several waves during the data collection period, and up to 2 reminders were sent through the same institutional and student networks. Partial responses were automatically saved so that participants could resume the questionnaire later. Participation was voluntary and uncompensated. The distribution of respondents by institution is provided in Multimedia Appendix 1.

Survey Instruments

The survey combined 5 previously published instruments or questionnaires with heterogeneous levels of prior psychometric

XSL•FO
RenderX

evaluation, assessing health care students' perceptions, readiness, learning intentions, and perceived risks regarding AI in medicine [21-25]. The Medical Artificial Intelligence Readiness Scale for Medical Students (MAIRS-MS) by Karaca et al [21] is a formally developed and psychometrically validated scale. The questionnaire by Li et al [25] is grounded in the Theory of Planned Behavior and was evaluated using confirmatory factor analysis and structural equation modeling. The questionnaire by Boillat et al [23] was systematically developed from a literature review and reported internal consistency and factor adequacy indices. By contrast, the questionnaires by Sit et al and Park et al [22,24] were previously published survey instruments with more limited psychometric evidence.

The full French versions were administered as previously translated, culturally adapted, and linguistically validated, without modification to item wording, response options, or item order [26]. This previous work established conceptual and linguistic equivalence in French; however, the psychometric properties of these French versions in French-speaking midwifery students specifically remain to be established.

In total, the survey comprised 112 items: 94 Likert-type items and 18 additional items, including 13 single-choice and 5 open-ended questions. Likert responses were harmonized onto a 1 to 5 scale, with higher scores indicating stronger agreement or endorsement. Additional items captured sociodemographic characteristics, including age group, gender, year of study, training institution, and region, as well as prior exposure to AI, generative AI tools, AI-related training, and digital health.

Survey Administration

The survey was hosted on a secure online platform and was open between January and May 2025. It was organized into 5 pages per screen, beginning with sociodemographic items followed by AI-related sections. An introductory page presented the study aims, voluntary participation, data-protection measures, and the estimated completion time. No directly identifying information was collected.

The survey link was disseminated in several waves, with up to 2 reminders sent through institutional contacts and student associations. The link was not technically individualized. To preserve complete anonymity, no mechanism to prevent duplicate entries was implemented. In practice, access was restricted to midwifery students because the link was transmitted locally by school administrations to their own students. Participants could complete the survey on any internet-enabled device. Partial responses were saved automatically to allow later completion, but only fully completed questionnaires were retained for analysis. Of 414 questionnaires initiated, 190 (45.9%) were fully completed. We report this online survey in accordance with the CHERRIES guidelines (Checklist 1).

Qualitative Follow-Up Strand

The exploratory qualitative follow-up strand consisted of semistructured individual interviews conducted with volunteer midwifery students from a school in eastern France. The site was selected pragmatically because of local access within the research network. This recruitment strategy is reported transparently and is addressed as a limitation. Students were invited during teaching sessions and by email through a local coordinator, with attention paid to variation in year of study, gender, and prior exposure to AI or digital health.

Interviews were conducted by a member of the research team who was a midwife. They took place either face-to-face in a quiet room at the school or via secure videoconference, lasted for a maximum of 30 minutes, and were audio-recorded with participants' consent. Recordings were transcribed verbatim in French and anonymized by removing names, locations, and other potentially identifying information. An interview guide, informed by the literature, the 5 questionnaires, and preliminary survey findings, was used to support the interviews; it was not formally pilot-tested. The interview guide explored students' self-reported knowledge of AI, prior AI-related training, personal and professional uses of AI, perceived future roles of AI in health care, barriers to AI use, digital skills, prior exposure to digital health, and perceptions of the future of health care in relation to digital technologies. When needed, a brief general definition of AI was provided to support a shared understanding during the interview. In total, 8 students were interviewed.

Qualitative data were analyzed using a pragmatic reflexive thematic approach informed by Braun and Clarke [27]. Analysis involved familiarization, initial coding, grouping of related codes, theme development and refinement, and review of the thematic structure against the study questions. Given the exploratory scope and the limited number of interviews, coding was managed manually in a structured spreadsheet rather than dedicated qualitative-analysis software. Descriptive codes were assigned by S Knepfler, and S Kotzki independently reviewed the coding framework and thematic structure. Discrepancies were discussed until consensus, and the codebook was refined iteratively. Analytic memos and reflexive notes were maintained. To support trustworthiness, we used investigator triangulation during coding, maintained an audit trail of coding decisions, and compared themes across participants and training years. Member checking was not conducted.

For descriptive characterization of the interview participants, an additional post hoc participant-level categorization of self-reported AI knowledge was performed after completion of the interview series. This categorization was distinct from the reflexive thematic analysis described above. The 6 cognitive levels of the original Bloom taxonomy (knowledge, comprehension, application, analysis, synthesis, and evaluation) were used as an interpretive framework [28]. The interview guide did not present these categories to participants, and participants did not self-select a Bloom level. S Knepfler made the initial assignment for each participant based on the interview material relating to their reported AI knowledge and understanding; S Kotzki independently reviewed the assignments, and discrepancies were resolved through discussion until consensus. A separate category, "No self-reported AI knowledge," which is not part of Bloom taxonomy, was added for participants who reported no AI knowledge. No formal scoring instrument was administered. The resulting categories were descriptive, researcher-assigned interpretations of self-reported material and should not be interpreted as an objective assessment of AI knowledge.

XSL•FO
RenderX

Data Management and Statistical Analysis

Survey responses from the 5 instruments were merged into a single analytic dataset and first examined at the item level within each original questionnaire. For each Likert-type item, we computed response frequencies, percentages, and weighted means on the 1 to 5 scale, SD, SE, and 95% CI, calculated as mean ± 1.96 × SE. Analyses were descriptive only. No inferential statistics, hypothesis tests, *P* values, significance thresholds, or normality testing were performed. Analyses were restricted to fully completed questionnaires (complete-case analysis, n=190), with no imputation of missing data. Quantitative analyses were conducted in Python 3.11 using pandas and NumPy for data management and pingouin for Cronbach α and descriptive statistics. Subgroup comparisons were not prespecified, and no inferential subgroup analyses were performed. Although variables such as year of study, prior AI exposure, and previous AI-related training were available, the study was neither designed nor powered for between-group comparisons. In particular, the subgroups reporting prior AI-related training were small: 26 of 190 (13.7%) respondents reported having previously taken AI courses or training, and 17 of 190 (8.9%) reported having received AI teaching or training. Subgroup analyses were therefore considered but not conducted; this is addressed as a limitation.

Item Aggregation and Thematic Domain Mapping

To support an integrated reading across the 5 instruments, we derived post hoc aggregated thematic scores. This aggregation was not based on a published conceptual framework and was not intended to create new validated scales. Instead, it was used as a descriptive, integrative analytic lens across questionnaires.

Two investigators (S Kotzki and S Knepfler) independently assigned each Likert-type item to 1 of 5 overarching domains: (1) current AI skills and knowledge, (2) practical application and clinical integration, (3) attitudes and perceptions, (4) training, learning, and commitment, and (5) ethical, regulatory, and risk-related aspects. Within these domains, items were grouped into 23 specific themes, plus a theme for sociodemographic characteristics. Discrepancies were resolved by discussion until consensus and documented in an item-to-theme matrix. No item wording, response option, item order, or original questionnaire was changed (Multimedia Appendix 1).

For each theme, a descriptive theme score was computed as the mean of its constituent Likert items on a 1 to 5 scale. To ensure that the associated 95% CI reflects participant-level sampling variability rather than dispersion across the theme's constituent items, an individual theme score was first computed for each participant as the mean of that participant's responses to the theme's items; the theme-level SD, SE, and 95% CI (1.96 × SE) were then derived from these 190 participant-level scores. For the attitudes-and-perceptions composite (Cronbach α=0.48), the theme score is reported descriptively as mean and SD without a CI, given its conceptual heterogeneity (Multimedia Appendix 1). Cronbach α was used as an indicator of internal coherence for descriptive interpretation, not as evidence of formal scale validation. Themes with α values below conventional thresholds were interpreted cautiously, with item-total correlations and content coherence examined to assess whether the aggregation remained meaningful. The attitudes-and-perceptions domain showed low internal consistency and was therefore treated as a heterogeneous descriptive composite. A supplementary item-to-instrument-to-domain mapping table is provided to make the post hoc aggregation transparent (Multimedia Appendix 1).

Ethical Considerations

This study was approved by the Ethics Committee for Integrity and Research in Health Professions Education of the International Francophone Society for Medical Education (notice 1210 - 2023, approval date November 26, 2023). Participation was voluntary and uncompensated. At the beginning of the survey and before each interview, participants received information on the study aims, confidentiality safeguards, data protection, and their right to withdraw without consequences. Electronic consent was obtained before survey completion, and verbal consent was recorded at the start of each interview.

Survey data were collected anonymously and stored on secure institutional servers, with access restricted to the research team. Interview recordings and transcripts were pseudonymized, with identifying information removed or altered. Only deidentified data were analyzed. Deidentified datasets and analysis scripts may be made available from the corresponding author upon reasonable request, subject to applicable regulations and data-protection requirements.

We followed the STROBE (Strengthening the Reporting of Observational Studies in Epidemiology; quantitative survey), CHERRIES (Checklist for Reporting Results of Internet E-Surveys; web-based survey), and COREQ (Consolidated Criteria for Reporting Qualitative Research; qualitative strand) reporting guidelines. Selected survey patterns and exploratory interview observations are summarized in a descriptive contextualization matrix (Multimedia Appendix 1). Completed checklists are provided in Checklist 1.

Results

Participation and Respondent Characteristics

Of the 34 institutions contacted, 30 (88.2%) confirmed dissemination of the survey link to their students. In total, 414 questionnaires were initiated, and 190 (45.9%) were fully completed and retained for analysis. The 224 (54.1%) noncompleted entries included 41 (9.9%) with no response beyond the landing page. Among the 183 partially completed entries, attrition occurred mainly at the beginning of the questionnaire: most stopped on the first of the survey's 5 pages (n=122, 66.7%), and only 6 (3.3%) reached the first AI-related item, indicating that most noncompleters did not progress beyond the sociodemographic section. Respondent characteristics are summarized in Table 1; respondents were predominantly women and distributed across years 2 to 5, with the institutional distribution detailed in Multimedia Appendix 1.

XSL•FO
RenderX

Table . Characteristics of participants in the quantitative and qualitative strands.

Characteristics	Quantitative strand (N=190), n (%)	Qualitative strand (N=8), n (%)
Gender		
Women	183 (96.3)	8 (100)
Men	4 (2.1)	0 (0)
Prefer not to answer	3 (1.6)	[a]—
Age (y)		
<20	26 (13.7)	0 (0)
20 - 24	137 (72.1)	5 (62.5)
25 - 29	15 (7.9)	0 (0)
30 - 34	6 (3.2)	1 (12.5)
>34	6 (3.2)	1 (12.5)
Not reported	0 (0)	1 (12.5)
Year of study		
2nd	47 (24.7)	0 (0)
3rd	55 (29.0)	2 (25)
4th	60 (31.6)	4 (50)
5th	27 (14.2)	2 (25)
Not recorded	1 (0.5)	0 (0)
Post hoc researcher-assigned categorization of self-reported AI knowledge[b]		
No self-reported AI knowledge	Not assessed	2 (25)
Knowledge	Not assessed	4 (50)
Comprehension	Not assessed	2 (25)
Application	Not assessed	0 (0)
Analysis	Not assessed	0 (0)
Synthesis	Not assessed	0 (0)
Evaluation	Not assessed	0 (0)

[a]Not applicable.

[b]This categorization was used only for descriptive characterization of the qualitative participants. Categories were assigned post hoc from interview material after completion of the interview series, using the 6 cognitive levels of the original Bloom taxonomy as an interpretive framework [28]; participants were not asked to select a Bloom level. "No self-reported AI knowledge" is an additional category outside Bloom taxonomy. The categorization was researcher assigned and was not an objective assessment of AI knowledge. This variable was not assessed in the quantitative survey strand.

The exploratory qualitative follow-up strand comprised 8 volunteer students from a single midwifery school (Table 1). They were in the third to fifth years of training and varied in age and prior exposure to digital tools; no second-year students and no men students participated. These interviews are reported below as single-site, exploratory contextual observations and are not intended to be representative of the survey population.

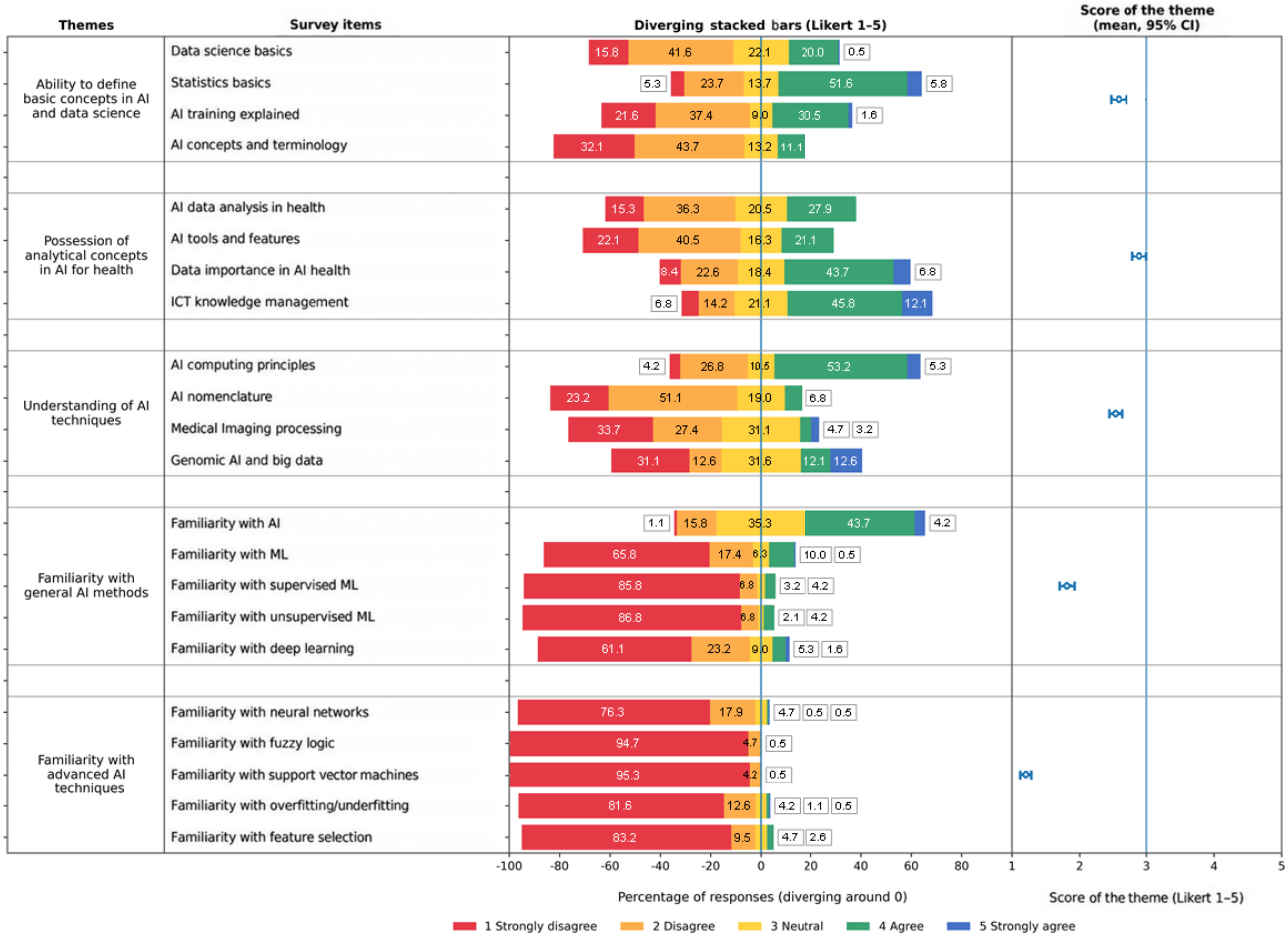
To summarize patterns across the 5 questionnaires, item-level results were also grouped into post hoc descriptive thematic summaries (5 domains and 23 themes), reported here as descriptive theme scores on the 1 to 5 scale rather than as validated scale scores (see *Methods* section and Multimedia Appendix 1).

AI-Related Knowledge and Familiarity

Self-reported AI skills and knowledge were below the scale midpoint of 3 across all related themes, ranging from 1.20 (95% CI 1.13 - 1.27) for familiarity with advanced AI techniques to 2.89 (95% CI 2.78 - 3.00) for analytical concepts in AI for health (Figure 1). Foundational items were uneven: agreement was higher for basics statistics (109/190, 57.4%) than for named AI methods. A notable item-level contrast was observed between general AI familiarity (91/190, 47.9% agreement) and familiarity with specific concepts such as machine learning (20/190, 10.5%) and deep learning (13/190, 6.8%; Figure 1; Multimedia Appendix 1).

XSL•FO
RenderX

Figure 1. Current AI skills and knowledge (n=190). This figure shows the distribution of responses to survey items (Likert scale 1 - 5) and the associated theme scores. The right-hand panel shows, for each theme, the theme score (mean) with its 95% CI as a dot plot; the vertical reference line indicates the neutral value (3). Labels under the survey items column have been translated and shortened for readability; the full original French wording of each item is available in Multimedia Appendix 1. ICT: information and communication technology; ML: machine learning.



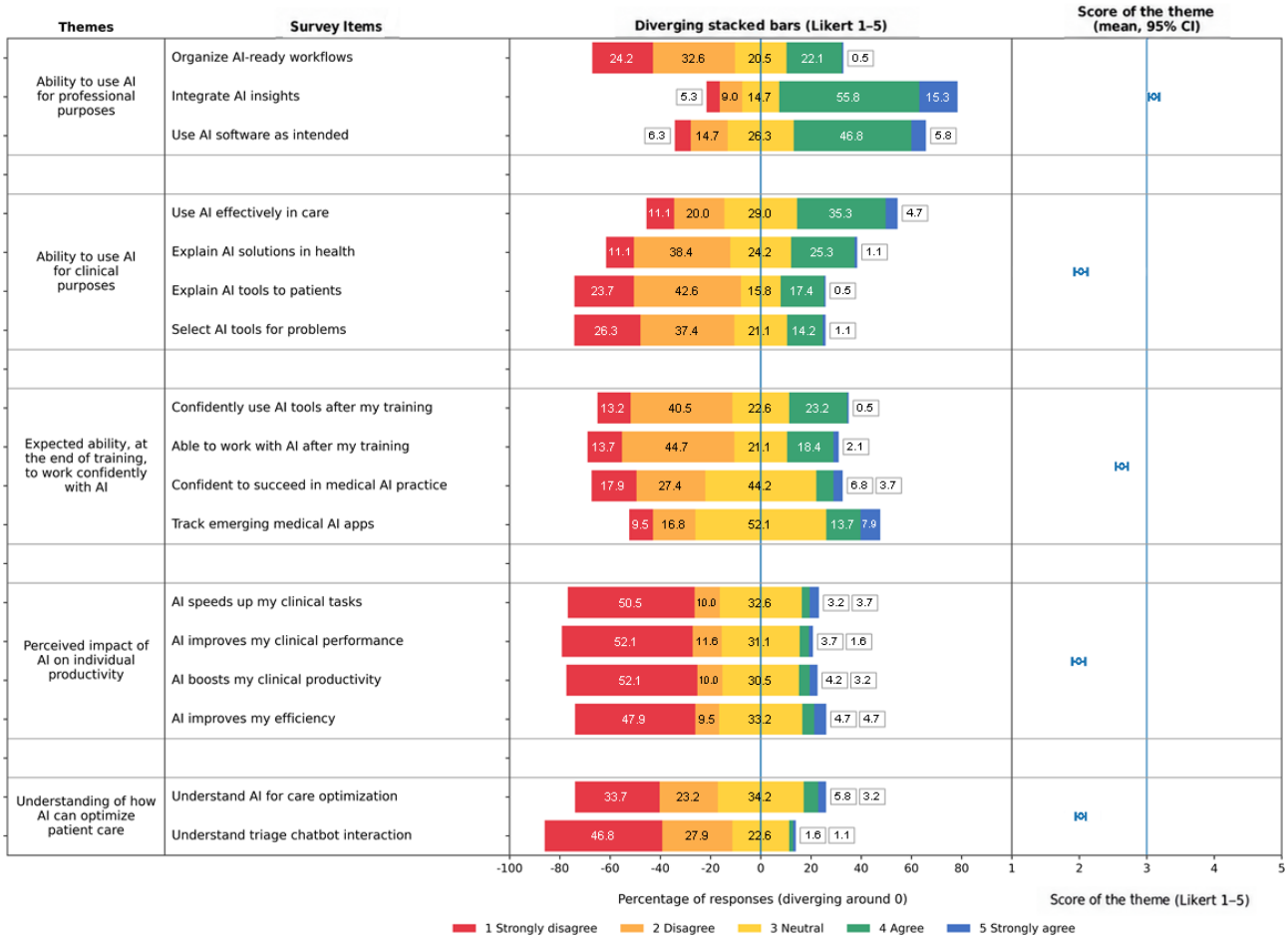
In the interviews, students described their AI knowledge in broad terms and referred mainly to everyday examples or general-purpose generative tools; they rarely named or explained specific methods, and formal AI instruction was described as rare or fragmented. These accounts provided exploratory, illustrative context for the low self-reported familiarity seen in the survey.

Perceived Ability to Use AI in Practice

Perceived ability to use AI in practice was generally limited (Figure 2). The theme score for professional use was closest to the midpoint (3.13, 95% CI 3.01 - 3.25), whereas clinical use (2.05, 95% CI 1.93 - 2.16) and perceived productivity gains (1.99, 95% CI 1.84 - 2.14) were below it. At the item level, difficulty was frequently reported for selecting AI tools for clinical problems (121/190, 63.7%, disagreement) and for explaining AI tools to patients (126/190, 66.3%, disagreement; Figure 2).

XSL•FO
RenderX

Figure 2. Perceived ability to use AI in practice (n=190). This figure shows the distribution of responses to survey items (Likert scale 1 - 5) and the associated theme scores. The right-hand panel shows, for each theme, the theme score (mean) with its 95% CI as a dot plot; the vertical reference line indicates the neutral value (3). Labels under the survey items column have been translated and shortened for readability; the full original French wording of each item is available in Multimedia Appendix 1.



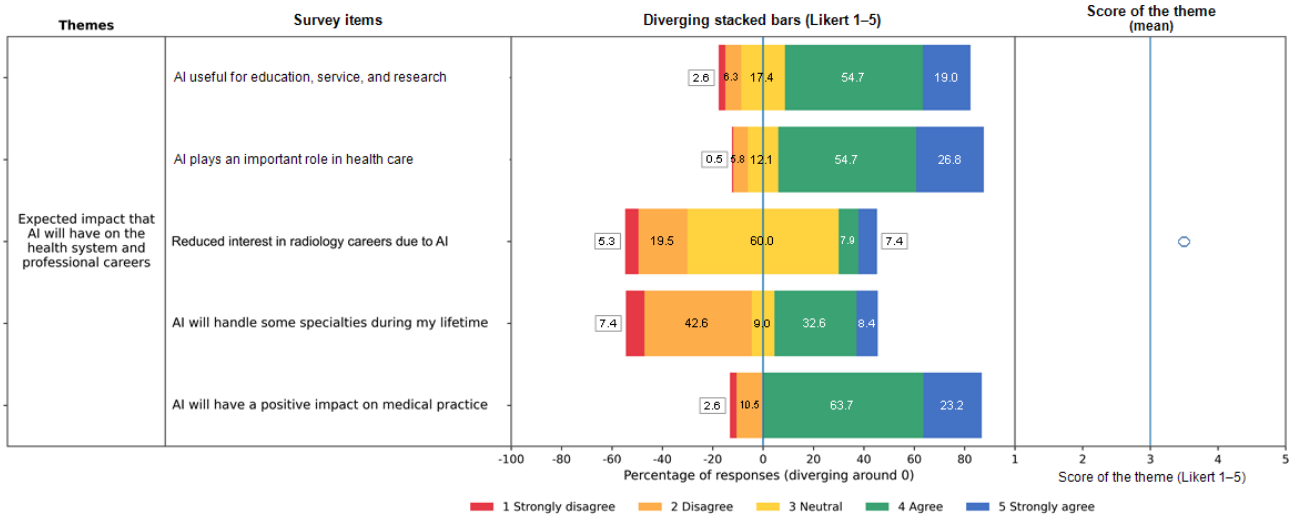
In the interviews, students described little exposure to AI during clinical placements; existing digital systems were generally not described as AI, and generative tools, when used, were mainly used outside patient care for academic purposes. These remarks are reported as exploratory contextual observations from the single-site strand.

Perceived Impact of AI on Health Care and Professional Roles

Perceptions of AI's broader impact were above the midpoint (descriptive composite theme mean score 3.50, SD 0.51; Figure 3). Agreement was high that AI would have a positive impact on medical practice during respondents' careers (165/190, 86.8%) and that it would play an important role in health care (155/190, 81.6%). Views on more disruptive implications were more divided or neutral, including responses on reducing interest in medical imaging careers and on AI handling some specialties within respondents' lifetimes (Figure 3).

XSL•FO
RenderX

Figure 3. Perceptions of AI's impact and professional projection (n=190). This figure shows the distribution of responses to survey items (Likert scale 1 - 5) and the associated theme scores. The right-hand panel shows, for each theme, the theme score (mean) with its 95% CI as a dot plot, except for the attitudes-and-perceptions composite, which is reported descriptively without a 95% CI; the vertical reference line indicates the neutral value (3). Labels under the survey items column have been translated and shortened for readability; the full original French wording of each item is available in Multimedia Appendix 1.



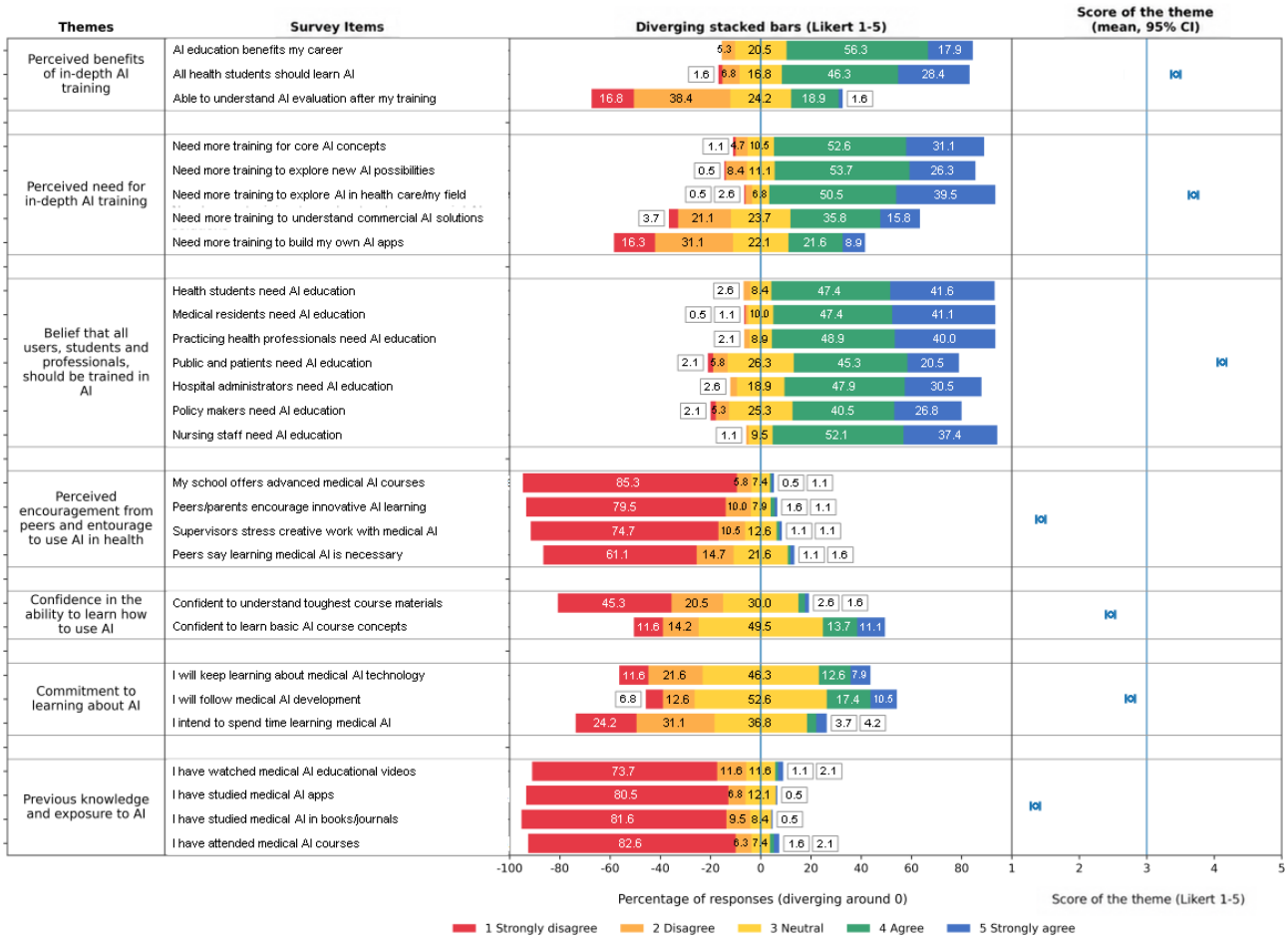
In the interviews, students anticipated that AI would become part of their professional environment but found it difficult to specify how it would affect day-to-day midwifery practice. These accounts provided exploratory context for the survey responses.

Training Needs, Prior Exposure, and Institutional Support

Endorsement of AI education was high (Figure 4). The highest descriptive theme score was observed for the belief that students and professionals should be trained in AI (4.11, 95% CI 4.02 - 4.20), with high agreement that health students need AI education (169/190, 88.9%). Perceived need for further training was also high (3.69, 95% CI 3.59 - 3.79). In contrast, perceived institutional and social support (1.43, 95% CI 1.34 - 1.52) and prior exposure (1.36, 95% CI 1.27 - 1.45) were among the lowest scores observed, and self-reported confidence (2.47, 95% CI 2.34 - 2.59) and commitment (2.76, 95% CI 2.63-2.89) were modest (Figure 4).

XSL•FO
RenderX

Figure 4. Training needs, prior exposure, and institutional support (n=190). This figure shows the distribution of responses to survey items (Likert scale 1 - 5) and the associated theme scores. The right-hand panel shows, for each theme, the theme score (mean) with its 95% CI as a dot plot; the vertical reference line indicates the neutral value (3). In the interviews, students expressed interest in being better prepared for AI but described their learning as largely self-directed and reported limited access to structured teaching or supervised opportunities. These accounts illustrate the survey pattern of high reported demand alongside low reported support. Labels under the survey items column have been translated and shortened for readability; the full original French wording of each item is available in Multimedia Appendix 1.



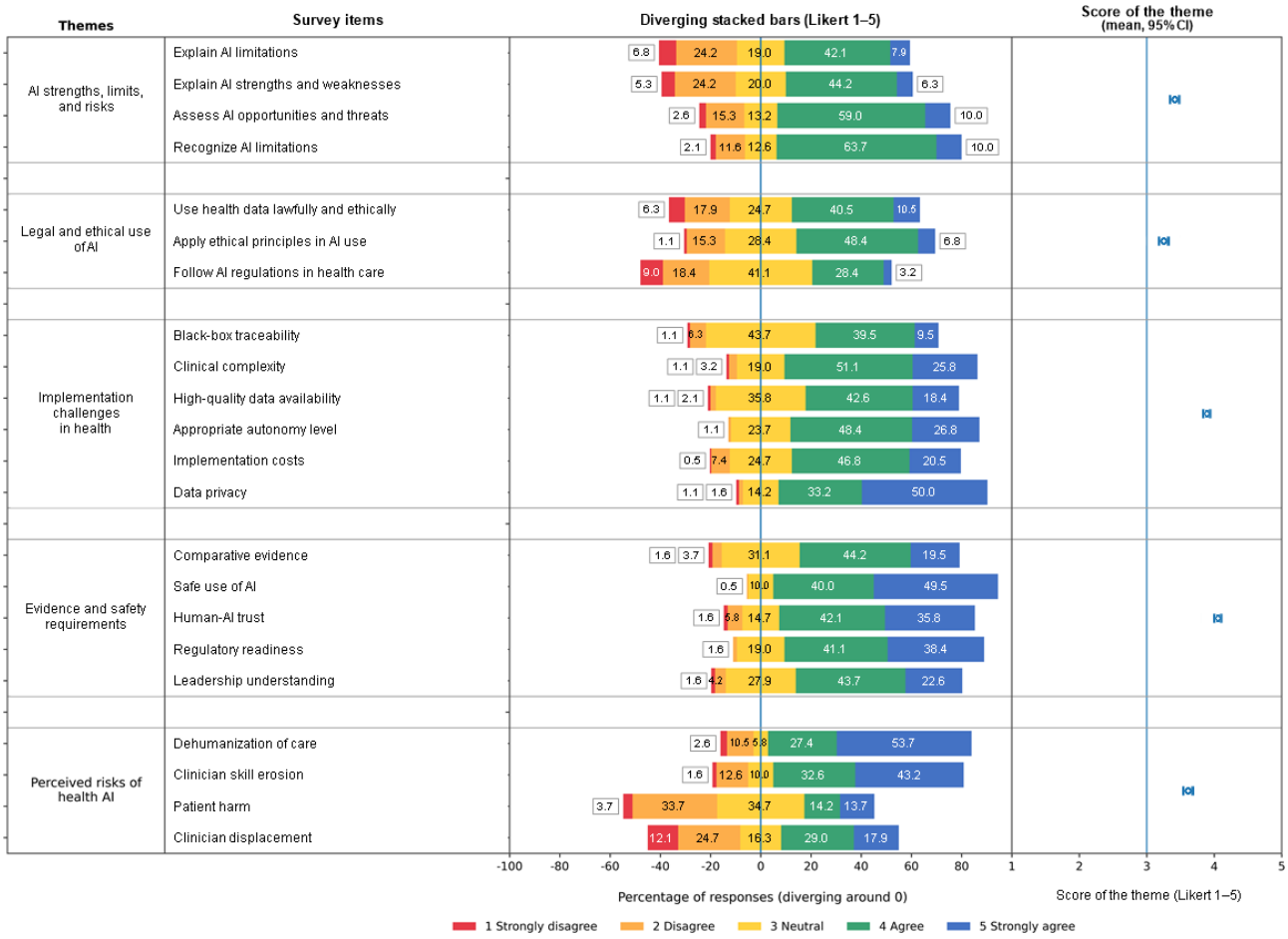
Ethical, Regulatory, and Risk-Related Perceptions

Ethical, regulatory, and risk-related themes were above the midpoint (Figure 5), with the highest descriptive score for evidence and safety requirements (4.04, 95% CI 3.96 - 4.12) and a lower score for legal and ethical use of AI (3.25, 95% CI 3.15 - 3.35). Data privacy was frequently identified as a key implementation issue (158/190, 83.2%). Perceptions of downstream risks varied: many anticipated dehumanization of care (154/190, 81.1%) and clinician skill erosion (144/190, 75.8%), whereas fewer agreed that AI would harm patients (53/190, 27.9%; Figure 5).

In the interviews, students raised data protection, accountability, and preservation of human judgment as priorities, while acknowledging limited knowledge of existing regulations. These exploratory remarks are reported for context only.

XSL•FO
RenderX

Figure 5. Ethical, regulatory, and risk-related perceptions (n=190). This figure shows the distribution of responses to survey items (Likert scale 1 - 5) and the associated theme scores. The right-hand panel shows, for each theme, the theme score (mean) with its 95% CI as a dot plot; the vertical reference line indicates the neutral value (3). Labels under the survey items column have been translated and shortened for readability; the full original French wording of each item is available in Multimedia Appendix 1.



Discussion

Principal Findings

In this national cross-sectional online survey of French midwifery students, complemented by exploratory qualitative interviews, students reported low self-perceived AI knowledge and familiarity, particularly regarding specific AI concepts and methods, together with limited perceived ability to use AI in clinical practice. At the same time, they expressed broadly positive expectations about AI's future role in health care while remaining uncertain about its concrete implications for midwifery practice. Demand for AI training was high, yet prior exposure and perceived institutional and social support were low. Students also placed strong emphasis on privacy, safety, accountability, and the preservation of human judgment, and they were concerned about the dehumanization of care and skill erosion. The exploratory interviews provided contextual illustrations of these patterns rather than independent confirmation of them.

Interpretation and Comparison With Prior Work

The studies underpinning the 5 questionnaires used in this survey provide an initial point of comparison. Conducted in different national and educational contexts—including Turkey, the United Kingdom, the United States, China, and multiregional samples—and predominantly involving medical students, physicians, or broader health professions learners, these studies reported a recurrent pattern: favorable attitudes toward AI coexisting with limited familiarity with core concepts, limited formal training, perceived need for AI education, and ethical or professional concerns [21-25]. The same domains were prominent among French midwifery students, suggesting shared educational challenges across health professions while also highlighting the need for discipline-specific interpretation in midwifery education.

Recent evidence in health professions education points in the same direction. A systematic review of studies involving health students and academics across nursing, midwifery, medicine, dentistry, allied health, and psychology found that positive attitudes toward AI were commonly accompanied by limited AI literacy, limited formal training, and recurring ethical concerns; students also frequently identified universities as an appropriate setting for acquiring AI competencies [29]. Consistent with this, a Turkish survey of nursing and midwifery students found that most participants had not received AI education despite generally favorable attitudes toward AI [30]. Similarly, in Jordan, medical students and pathology trainees were often aware of AI, but formal exposure through medical schools remained limited, and respondents expressed a need for

XSL•FO
RenderX

education on AI applications, risks, and ethical or legal implications [31].

For midwifery, these findings intersect with broader questions about how digital and AI-supported transformation affects maternity care. A qualitative synthesis of midwives' experiences of digital health indicates that digital transformation influences professional roles, relational work, accountability, and digitally mediated communication [32]. Similar concerns appear in the health education literature, where midwifery students have emphasized that AI cannot substitute for empathy and relational care [29]. Evidence from nurse practitioner education also suggests that the educational value of clinical AI tools depends on adequate training and supervised exposure rather than on availability alone, with implications for diagnostic reasoning and confidence [33]. In this context, our findings indicate that French midwifery students may hold positive expectations about AI while reporting limited perceived readiness, limited supervised exposure, and ethical and professional concerns, including about safety, privacy, and the preservation of human judgment.

The contrast between self-reported familiarity with "AI" in general and lower familiarity with specific methods such as machine learning or deep learning also deserves attention. The interviews suggest that students may use the term "AI" to refer to different realities, including general AI, consumer generative AI tools, and regulated or clinically deployed AI systems. This distinction is increasingly relevant, given evidence that health professions students already use generative AI tools for study, writing, and clinical-placement tasks while receiving comparatively little formal instruction in them [11]. A survey of first-year Japanese medical students similarly found widespread generative AI use alongside less common formal learning and strong interest in structured generative AI literacy [34]. Familiarity with consumer generative tools may therefore inflate apparent AI familiarity without implying readiness to understand, evaluate, or use regulated clinical AI systems.

Because AI knowledge and familiarity were self-perceived, students may have overestimated or underestimated their actual understanding, and the ambiguity around "AI" may have affected responses in either direction. The findings should therefore be read as describing perceived, rather than demonstrated, knowledge and readiness.

Implications for Midwifery Education and AI Literacy

These findings suggest that introducing AI literacy progressively within midwifery curricula could help align students' interest in AI with their currently limited preparedness, although the present study does not itself validate any specific curriculum. For midwifery education, AI literacy should therefore not be limited to technical awareness. It should also help students consider how AI-supported information is communicated within the continuity of care, how professional responsibility is maintained in shared decision-making, and how human judgment and relational presence are preserved when digital outputs enter encounters with women and families. Teaching could explicitly distinguish general AI, generative AI tools, and regulated clinical AI systems and address data privacy, accountability, transparency, validation principles, human oversight, communication around AI-supported decisions, and safe-use boundaries for generative tools. This distinction is important because clinically deployed AI systems raise specific requirements related to safety assurance, procedural safeguards, transparency, human oversight, data governance, and professional accountability [35-37]. In Europe, these issues can also be linked pedagogically to the AI Act and to data-protection requirements under the General Data Protection Regulation (GDPR) [38,39]. Given students' strong emphasis on safety and ethics, these elements may offer a useful entry point.

Because the instruments used here were generic rather than midwifery-specific and because the French versions underwent cross-cultural translation, adaptation, and linguistic validation but not psychometric validation among French-speaking midwifery students, a midwifery-specific AI literacy measure remains to be developed and validated [26]. Such work could combine qualitative exploration with educators, students, and clinicians, followed by consensus-building and psychometric validation.

Strengths and Limitations

To the best of our knowledge, this is the first national study of AI-related knowledge, perceptions, perceived risks, and training needs among midwifery students in France. Conducted during the 2024 to 2025 academic year, it draws on multiple institutions and combines survey data with exploratory qualitative observations. It is intended to serve as a baseline for repeated assessments in future academic years, allowing changes in AI exposure, perceived readiness, and training needs to be monitored over time.

Several limitations should be considered. First, the survey was long, with 112 items across 5 pages. Of the 414 questionnaires initiated, 190 were completed, corresponding to a completion rate of 45.9% (190/414). Most attrition occurred early, before the AI-related sections. This pattern, together with voluntary participation, may introduce selection and nonresponse bias, and respondents may have differed from nonrespondents in their interest in or familiarity with AI. In addition, because no technical mechanism was used to prevent duplicate submissions in order to preserve complete anonymity, repeated participation by the same individual cannot be completely excluded.

Second, the data were collected between January and May 2025. Given the rapid evolution of AI and the uptake of generative AI in education and health care, the findings should be interpreted as a time-bound baseline rather than a stable estimate.

Third, no inferential subgroup analyses were prespecified or conducted, and the subgroups reporting prior AI courses or training were small (26/190 and 17/190). The data therefore do not support conclusions by year of study, prior exposure, or previous AI-related training. Fourth, the 5 domains and 23 themes were post hoc descriptive thematic summaries used as an integrative lens across questionnaires, not validated scales. Relatedly, the survey measures of AI knowledge and familiarity were self-reported rather than objectively assessed. In the qualitative strand, the participant-level categorization shown in Table 1 was a post hoc researcher interpretation of self-reported

XSL•FO
RenderX

interview material rather than a participant self-rating or an objective assessment. Students may therefore have overestimated or underestimated their understanding, particularly because the term "AI" may have been interpreted differently among respondents.

Fifth, the qualitative component was an exploratory single-site follow-up strand with 8 volunteers. It included no men students and no second-year students; no claim of data saturation is made, and these observations contextualize but do not generalize beyond the participants. The absence of men participants in the qualitative sample should be interpreted in light of the gender composition of the profession and of our survey sample. In France, midwifery remains a predominantly female profession, with national DREES (Directorate for Research, Studies, Evaluation and Statistics) data indicating that 97% of practicing midwives were women in 2025 [40]. Men were also a small minority in our survey sample (4/190, 2.1%). The absence of men interviewees therefore remains a limitation of perspective, but should not be interpreted as evidence of gender-related differences.

Conclusions

French midwifery students reported low perceived AI knowledge and limited perceived readiness to use AI in clinical practice, alongside strong demand for training and high salience of safety, privacy, accountability, and human judgment, in the context of low prior exposure and limited institutional support. Interpreted as a time-bound baseline, these findings suggest a readiness gap that is unlikely to be closed by informal, self-directed learning alone. They support the progressive and proportionate introduction of AI literacy in midwifery education in France, explicitly distinguishing general, generative, and clinical AI systems, and underline the need to develop and validate a midwifery-specific AI literacy measure to support future curriculum design and evaluation.

Acknowledgments

The authors gratefully acknowledge the support of Lionel Di Marco, vice president of the National Conference of Midwifery Educators, as well as the directors and leadership teams of the participating French midwifery schools, whose approval and assistance enabled the dissemination of the survey invitation to students. Their support in facilitating access to midwifery student networks was essential to study recruitment.

During the preparation of this manuscript, the authors used GPT-5.2 (OpenAI) for code optimization in data analysis and language editing and GPT-5.6 Sol during the revision stage to modify figures. The authors reviewed and edited all outputs, verified all statistical results against the original data, and take full responsibility for the content, data interpretations, and conclusions of this publication.

Funding

The study was supported by funding from the French National Research Agency (ANR) under the France 2030 program (23-CMAS-0035), by the Multidisciplinary Institute in Artificial intelligence (MIAI) Cluster (ANR-23-IACL-0006), and also by the "Investissements d'Avenir" program (ANR-10-AIRT-0005). The funders played no role in the study design, data collection, analysis and interpretation of data, or the writing of this manuscript. This work also forms part of a broader translational and interdisciplinary SANGRIA (Santé Numérique Grenoble Intelligence Artificielle) research program.

Data Availability

The data supporting the findings of this study are not publicly available due to ethical and privacy restrictions. Anonymized datasets can be made available from the corresponding author (SKo) upon reasonable request, subject to applicable regulations and data-protection requirements. The study was approved by the Ethics Committee for Integrity and Research in Health Professions Education of the International Francophone Society for Medical Education (SIFEM; notice n° 1210-2023)

Authors' Contributions

S Kotzki conceived the study, developed the methodology, curated the data, performed the formal analyses, produced the visualizations, and drafted the first version of the manuscript. S Knepfler contributed to the study conceptualization and methodology, conducted the investigation, and participated in manuscript revision. LDM contributed to the study conceptualization and recruitment support/resources and critically reviewed the manuscript. NV provided overall supervision and critical guidance throughout the project and contributed to manuscript revision. All authors critically reviewed the manuscript, approved the definitive version, and agreed to be accountable for all aspects of the work.

Conflicts of Interest

None declared.

Multimedia Appendix 1

The distribution of respondents by midwifery school/university, the thematic grouping of survey items and internal consistency estimates, and the joint display integrating quantitative findings, illustrative quotes, and mixed methods inferences.

XSL•FO
RenderX

[DOCX File, 72 KB - mededu_v12i1e94310_app1.docx]

Checklist 1
STROBE (cross-sectional quantitative component), CHERRIES (web-based survey), and COREQ (exploratory qualitative follow-up strand).
[DOCX File, 36 KB - mededu_v12i1e94310_app2.docx]

References

1. Moazemi S, Vahdati S, Li J, et al. Artificial intelligence for clinical decision support for monitoring patients in cardiovascular ICUs: a systematic review. Front Med (Lausanne) 2023;10:1109411. [doi: 10.3389/fmed.2023.1109411] [Medline: 37064042]
2. Salinas MP, Sepúlveda J, Hidalgo L, et al. A systematic review and meta-analysis of artificial intelligence versus clinicians for skin cancer diagnosis. NPJ Digit Med 2024 May 14;7(1):125. [doi: 10.1038/s41746-024-01103-x] [Medline: 38744955]
3. Bajwa J, Munir U, Nori A, Williams B. Artificial intelligence in healthcare: transforming the practice of medicine. Future Healthc J 2021 Jul;8(2):e188-e194. [doi: 10.7861/fhj.2021-0095] [Medline: 34286183]
4. Mohsin Khan M, Shah N, Shaikh N, Thabet A, Alrabayah T, Belkhair S. Towards secure and trusted AI in healthcare: a systematic review of emerging innovations and ethical challenges. Int J Med Inform 2025 Mar;195:105780. [doi: 10.1016/j.ijmedinf.2024.105780] [Medline: 39753062]
5. Asan O, Bayrak AE, Choudhury A. Artificial intelligence and human trust in healthcare: focus on clinicians. J Med Internet Res 2020 Jun 19;22(6):e15154. [doi: 10.2196/15154] [Medline: 32558657]
6. Kimiafar K, Sarbaz M, Tabatabaei SM, et al. Artificial intelligence literacy among healthcare professionals and students: a systematic review. Front Health Inform 2023 Nov 11;12:168. [doi: 10.30699/fhi.v12i0.524]
7. Ng FYC, Thirunavukarasu AJ, Cheng H, et al. Artificial intelligence education: an evidence-based medicine approach for consumers, translators, and developers. Cell Rep Med 2023 Oct 17;4(10):101230. [doi: 10.1016/j.xcrm.2023.101230] [Medline: 37852174]
8. Schubert T, Oosterlinck T, Stevens RD, Maxwell PH, van der Schaar M. AI education for clinicians. EClinicalMedicine 2025 Jan;79:102968. [doi: 10.1016/j.eclinm.2024.102968] [Medline: 39720600]
9. Arrêté du 10 novembre 2022 relatif à la formation socle au numérique en santé des étudiants en santé [Article in French]. Légifrance. 2022. URL: https://www.legifrance.gouv.fr/eli/arrete/2022/11/10/SPRH2220195A/jo/texte [accessed 2025-03-13]
10. Arrêté du 3 juillet 2024 relatif au régime des études en vue du diplôme d'Etat de docteur en maïeutiqueVersion INITIALE [Article in French]. Légifrance. 2024 Jul 3. URL: https://www.legifrance.gouv.fr/eli/arrete/2024/7/3/ESRS2409766A/jo/texte [accessed 2026-02-24]
11. Kotzki S, Massonnet Turner C, Gauthier K, Minoves M, Vuillerme N. Health professional students' use of generative artificial intelligence during clinical placements: cross-sectional online survey study. JMIR Med Educ 2026 Apr 27;12:e85243. [doi: 10.2196/85243] [Medline: 42044362]
12. Dhombres F, Bonnard J, Bailly K, Maurice P, Papageorghiou AT, Jouannic JM. Contributions of artificial intelligence reported in obstetrics and gynecology journals: systematic review. J Med Internet Res 2022 Apr 20;24(4):e35465. [doi: 10.2196/35465] [Medline: 35297766]
13. Della Ripa S, Santos N, Walker D. AI-enabled obstetric point-of-care ultrasound as an emerging technology in low- and middle-income countries: provider and health system perspectives. BMC Pregnancy Childbirth 2025 Jul 4;25(1):729. [doi: 10.1186/s12884-025-07796-6] [Medline: 40615993]
14. Jost E, Kosian P, Jimenez Cruz J, et al. Evolving the era of 5D ultrasound? A systematic literature review on the applications for artificial intelligence ultrasound imaging in obstetrics and gynecology. J Clin Med 2023 Oct 29;12(21):6833. [doi: 10.3390/jcm12216833] [Medline: 37959298]
15. Çitil ET, Çitil Canbay F. Artificial intelligence and the future of midwifery: what do midwives think about artificial intelligence? A qualitative study. Health Care Women Int 2022 Dec;43(12):1510-1527. [doi: 10.1080/07399332.2022.2055760] [Medline: 35452353]
16. O'Connor S. Teaching artificial intelligence to nursing and midwifery students. Nurse Educ Pract 2022 Oct;64:103451. [doi: 10.1016/j.nepr.2022.103451] [Medline: 36166951]
17. Kranz A, Abele H. The impact of artificial intelligence (AI) on midwifery education: a scoping review. Health Care (Don Mills) 2024 May 24;12(11):1082. [doi: 10.3390/healthcare12111082]
18. Giaxi P, Vivilaki V, Sarella A, Gourounti K. Artificial intelligence in midwifery: a scoping review of current applications, future prospects, and midwives' perspectives. Health Care (Don Mills) 2025;13(8):942. [doi: 10.3390/healthcare13080942]
19. Teng M, Singla R, Yau O, et al. Health care students' perspectives on artificial intelligence: countrywide survey in Canada. JMIR Med Educ 2022 Jan 31;8(1):e33390. [doi: 10.2196/33390] [Medline: 35099397]
20. Liu DS, Sawyer J, Luna A, et al. Perceptions of US medical students on artificial intelligence in medicine: mixed methods survey study. JMIR Med Educ 2022 Oct 21;8(4):e38325. [doi: 10.2196/38325] [Medline: 36269641]

XSL•FO
RenderX

21. Karaca O, Çalışkan SA, Demir K. Medical artificial intelligence readiness scale for medical students (MAIRS-MS) - development, validity and reliability study. BMC Med Educ 2021 Feb 18;21(1):112. [doi: 10.1186/s12909-021-02546-6] [Medline: 33602196]
22. Sit C, Srinivasan R, Amlani A, et al. Attitudes and perceptions of UK medical students towards artificial intelligence and radiology: a multicentre survey. Insights Imaging 2020 Feb 5;11(1):14. [doi: 10.1186/s13244-019-0830-7] [Medline: 32025951]
23. Boillat T, Nawaz FA, Rivas H. Readiness to embrace artificial intelligence among medical doctors and students: questionnaire-based study. JMIR Med Educ 2022 Apr 12;8(2):e34973. [doi: 10.2196/34973] [Medline: 35412463]
24. Park CJ, Yi PH, Siegel EL. Medical student perspectives on the impact of artificial intelligence on the practice of medicine. Curr Probl Diagn Radiol 2021;50(5):614-619. [doi: 10.1067/j.cpradiol.2020.06.011] [Medline: 32680632]
25. Li X, Jiang MYC, Jong MSY, Zhang X, Chai CS. Understanding medical students' perceptions of and behavioral intentions toward learning artificial intelligence: a survey study. Int J Environ Res Public Health 2022 Jul 18;19(14):8733. [doi: 10.3390/ijerph19148733]
26. Kotzki S, Massonnet Turner C, Vuillerme N. Questionnaires on perceptions of artificial intelligence in health care among health care students: cross-cultural translation into French and linguistic validation. JMIR Med Educ 2026 Mar 19;12:e76572. [doi: 10.2196/76572] [Medline: 41855416]
27. Braun V, Clarke V. Using thematic analysis in psychology. Qual Res Psychol 2006 Jan;3(2):77-101. [doi: 10.1191/1478088706qp063oa]
28. Bloom BS, Engelhart MD, Furst EJ, Hill WH, Krathwohl DR. Taxonomy of Educational Objectives The Classification of Educational Goals Handbook 1: Cognitive Domain: David McKay Company; 1956.
29. Shishehgar S, Murray-Parahi P, Alsharaydeh E, Mills S, Liu X. Artificial intelligence in health education and practice: a systematic review of health students' and academics' knowledge, perceptions and experiences. Int Nurs Rev 2025 Jun;72(2):e70045. [doi: 10.1111/inr.70045] [Medline: 40545441]
30. Kilci Erciyas Ş, Cirban Ekrem E, Keten Edis E. Relationship between individual innovativeness levels and attitudes toward artificial intelligence among nursing and midwifery students. Comput Inform Nurs 2024 Nov 1;42(11):802-808. [doi: 10.1097/CIN.0000000000001170] [Medline: 39023377]
31. Rjoop A, Al-Qudah M, Alkhasawneh R, et al. Awareness and attitude toward artificial intelligence among medical students and pathology trainees: survey study. JMIR Med Educ 2025 Jan 10;11:e62669. [doi: 10.2196/62669] [Medline: 39803949]
32. Sun Y, Chong MC, Zhang J, Guo RH, Xu F, Chui PL. Digital health in midwifery practice: a qualitative review of midwives' experiences and perceptions. Int Nurs Rev 2026 Jun;73(2):e70181. [doi: 10.1111/inr.70181] [Medline: 42011094]
33. McKay N, Palamara P, McCavery A, Nosaka K, Kwok WH. Nurse practitioner students' perceptions of an artificial intelligence differential diagnosis tool: a pilot study. J Clin Nurs 2026 Oct;35(10):4198-4209. [doi: 10.1111/jocn.70287] [Medline: 41869715]
34. Tajima H, Kasai H, Shikino K, Shimizu I, Ito S. Perceptions and intentions to use generative AI among first-year medical students in Japan: cross-sectional survey study. JMIR Med Educ 2025 Nov 13;11:e77552. [doi: 10.2196/77552] [Medline: 41232013]
35. Ellahham S, Ellahham N, Simsekler MCE. Application of artificial intelligence in the health care safety context: opportunities and challenges. Am J Med Qual 2020;35(4):341-348. [doi: 10.1177/1062860619878515] [Medline: 31581790]
36. Habli I, Lawton T, Porter Z. Artificial intelligence in health care: accountability and safety. Bull World Health Organ 2020 Apr 1;98(4):251-256. [doi: 10.2471/BLT.19.237487] [Medline: 32284648]
37. Díaz-Rodríguez N, Del Ser J, Coeckelbergh M, López de Prado M, Herrera-Viedma E, Herrera F. Connecting the dots in trustworthy artificial intelligence: from AI principles, ethics, and key requirements to responsible AI systems and regulation. Inf Fusion 2023 Nov;99:101896. [doi: 10.1016/j.inffus.2023.101896]
38. Regulation (EU) 2016/679 of the european parliament and of the council of 27 april 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing directive 95/46/EC (general data protection regulation) (text with EEA relevance). EUR-Lex. 2016. URL: https://eur-lex.europa.eu/eli/reg/2016/679/oj [accessed 2026-09-26]
39. Regulation (EU) 2024/1689 of the european parliament and of the council of 13 june 2024 laying down harmonised rules on artificial intelligence and amending regulations (EC) no 300/2008, (EU) no 167/2013, (EU) no 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (artificial intelligence act) (text with EEA relevance). EUR-Lex. 2024. URL: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A32024R1689 [accessed 2026-09-26]
40. Début 2025, les effectifs de pharmaciens, chirurgiens-dentistes, sages-femmes et pédicures-podologues continuent d'augmenter [Article in French]. DREES. 2025. URL: https://drees.solidarites-sante.gouv.fr/jeux-de-donnees/DATA_250728-demographie-pharmaciens-chirurgiens-dentistes-sages-femmes [accessed 2026-07-02]

Abbreviations

CHERRIES: Checklist for Reporting Results of Internet E-Surveys

XSL•FO
RenderX

COREQ: Consolidated Criteria for Reporting Qualitative Research
DRESS: Directorate for Research, Studies, Evaluation and Statistics
GDPR: General Data Protection Regulation
MAIRS-MS: Medical Artificial Intelligence Readiness Scale for Medical Students
PASS: Specific Health Access Pathway
STROBE: Strengthening the Reporting of Observational Studies in Epidemiology

Edited by N Khamis; submitted 02.Mar.2026; peer-reviewed by I Tency, P Nonjui; revised version received 02.Sep.2026; accepted 18.Sep.2026; published 26.Oct.2026.

Please cite as:
Kotzki S, Knepfler S, Di Marco L, Vuillerme N
AI Literacy and Training Needs Among French Midwifery Students: National Cross-Sectional Online Survey and Exploratory Qualitative Follow-Up
JMIR Med Educ 2026;12:e94310
URL: https://mededu.jmir.org/2026/1/e94310
doi:10.2196/94310

© Sylvain Kotzki, Sonia Knepfler, Lionel Di Marco, Nicolas Vuillerme. Originally published in JMIR Medical Education (https://mededu.jmir.org), 26.Oct.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Education, is properly cited. The complete bibliographic information, a link to the original publication on https://mededu.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Understanding Clinicians' Informational Needs for AI-Driven Clinical Decision Support Systems: Qualitative Interview Study

Simone Mingels[1], MSc; Hannah Piehl[1], MA; Madeline Therrien[1], MSc; Ekaterina Akhmad[1], MSc; Anniek van Hienen[1], MSc; Johan van Soest[1,2], PhD; Laura Hochstenbach[3], PhD; Andre Dekker[1], PhD; Olga Damman[4], PhD; Rianne Fijten[1], PhD

[1]Department of Radiation Oncology (Maastro), GROW Research Institute for Oncology and Reproduction, Maastricht University Medical Centre+, Maastricht University, Paul Henri Spaaklaan 1, Maastricht, The Netherlands

[2]Brightlands Institute for Smart Society (BISS), Faculty of Science and Engineering, Maastricht University, Heerlen, The Netherlands

[3]Department of Health Services Research, Care and Public Health Research Institute (CAPHRI), Maastricht University, Maastricht, The Netherlands

[4]Department of Public and Occupational Health and Amsterdam Public Health Research Institute, Quality of Care, Amsterdam UMC location VUmc, Amsterdam, The Netherlands

Corresponding Author:
Simone Mingels, MSc
Department of Radiation Oncology (Maastro), GROW Research Institute for Oncology and Reproduction, Maastricht University Medical Centre+, Maastricht University, Paul Henri Spaaklaan 1, Maastricht, The Netherlands

Abstract

Background: Advancements in artificial intelligence (AI) are transforming health care, particularly through AI-driven clinical decision support systems (AI-CDSS) that aid in predicting disease progression and personalizing treatment. Despite their potential, adoption remains limited due to clinician concerns about algorithm misuse, misinterpretation, and lack of transparency.

Objective: This qualitative study explores the informational needs and preferences of clinicians to better understand and appropriately use AI-CDSS in decision-making. In parallel, this study explores AI experts' perspectives on what information should be communicated to enable safe and appropriate use of AI-CDSS.

Methods: A qualitative description design study was conducted using semistructured interviews with 16 participants (8 clinicians and 8 AI experts). Discussions focused on experiences with AI, informational needs, and feedback on existing reporting standards, including Model Cards, Model Facts, and the Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis–Artificial Intelligence (TRIPOD-AI) checklist. The transcripts were analyzed through codebook thematic analysis.

Results: Four key themes were identified: (1) clinicians need clear information on training data, its origin, size, and inclusion and exclusion criteria, to judge model applicability; (2) performance metrics must go beyond the area under the curve (AUC) and be clinically relevant to support informed decisions; (3) limitations and warnings about inappropriate use should be specific and clearly communicated to prevent misuse; and (4) information should be presented in layered, customizable formats within existing clinical software, avoiding unnecessary jargon, and allowing optional deeper explanations. While each of the reviewed reporting standards offered strengths, none were considered sufficient alone. Participants recommended a combined and clinician-centered approach to information delivery. Alignment of reporting standards with clinical workflows and decision thresholds was thought to be crucial to bridge the usability gap.

Conclusions: To improve AI-CDSS adoption in clinical practice, reporting standards must be designed for better clinician comprehension and usability. Enhancing transparency, particularly regarding training data and performance, can likely help clinicians assess AI-CDSS more effectively. Information should be delivered in an accessible, layered format, fitting clinical workflows. Co-creation with clinicians throughout AI-CDSS development was a cross-cutting theme, highlighting its importance in ensuring tools are not only technically sound but also practically usable. Future research should explore how to structurally report on performance and validation metrics for clinician understanding and assess the impact of information provision on AI-CDSS adoption.

(JMIR Med Educ 2026;12:e85228) doi:10.2196/85228

KEYWORDS

artificial intelligence; AI implementation; delivery of health care; informational needs; reporting standard; transparency; co-creation

XSL•FO
RenderX

Introduction

Advancements in artificial intelligence (AI) are rapidly transforming the health care sector [1]. AI has the potential to aid clinicians and patients in comparing various treatment options by predicting future events based on individual patient characteristics to determine which treatment would benefit the patient the most [2]. Implementing these AI-driven clinical decision support systems (AI-CDSS) offers the potential for highly personalized medicine by predicting disease survival probabilities and potential treatment side effects [3,4]. Despite its potential, achieving sustained adoption of AI-CDSS in routine clinical care frequently encounters challenges [5].

One of these challenges is clinicians' resistance to adopting AI-CDSS, which is exacerbated by their concerns about the high potential risks to patient safety and quality of care [6,7]. Among clinicians, there can be a lack of in-depth knowledge of how algorithms are constructed and how they generate predictions and recommendations [8]. This can make clinicians unprepared to assess if the algorithm is usable within their specific clinical setting, which can cause (unintended) misuse of algorithms [9-11]. An example of such misuse can be found in the study by Zhao et al [12], where the INFLUENCE tool (IKNL) was applied to guide decisions about primary therapy. However, this tool was originally designed to estimate individual, time-dependent risks of recurrence or metastasis in patients with breast cancer who have already completed curative treatment [13]. Using the model outside its intended context led to inadequate treatment recommendations [14].

The lack of knowledge that can exist amongst clinicians is not surprising, since there is an overall inadequate adherence to reporting standards within the AI development field [15-17]. Even when developers adhere to these reporting standards, the reports are often too technical for clinical end users to adequately understand the potential risks or the intended use of the models [18,19]. Descriptions of AI-CDSS, such as intended use and target population, can help clinicians better understand the meaning of the outcome produced by a model before acting on its recommendation, which can prevent mistakes made due to misinterpretation of the outcome [20].

To stimulate reporting of prediction models in published literature, several initiatives have been introduced [21]. The Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis–Artificial Intelligence (TRIPOD-AI) statement is a set of recommendations for the reporting of studies developing, validating, or updating a prediction model and is commonly used and/or required by journal publications [22-24]. In addition, other guidelines such as the CONSORT-AI (Consolidated Standards of Reporting Trials extension for Artificial Intelligence) [25], Developmental and Exploratory Clinical Investigations of Decision Support Systems Driven by Artificial Intelligence (DECIDE-AI) [26], and SPIRIT-AI (Standard Protocol Items: Recommendations for Interventional Trials involving Artificial Intelligence) [27] were created. However, the focus of these reporting standards is on transparent reporting within clinical trial reports or protocols, not on improving understanding of AI for clinicians [28,29]. Furthermore, these reporting standards are developed to report on an AI model and not the environment or software system in which it is operating [30].

In an attempt to further standardize reporting of AI model information, Google established the concept of "Model Cards", which is comparable to providing a recipe and nutritional facts to a meal [31]. Through Model Cards, developers provide a one- or 2-page record highlighting characteristics of training data, intended use cases, and performance [32]. This Model Card can boost conformation to reporting standards but does not contain actionable information and guidance for applying AI in clinical practice. Sendak et al [33] reconstructed the idea of Model Cards using the concept of drug fact boxes. Drug fact boxes are used to understandably communicate benefits and risks associated with medications to patients [34]. These concepts combined resulted in "Model Facts" [33], an overview to "collate relevant, actionable information in a compact overview to ensure that clinicians know how, when, how not, and when not to use model output in their clinical decisions" [33].

Although the Model Facts concept is not thoroughly tested yet, the initiative has sparked a discussion on how to effectively communicate the potential risks of AI models to health care professionals [33]. This highlights a broader need to understand what information clinicians require in order to evaluate and appropriately use AI-driven recommendations in their clinical decision-making. Reflecting these developments, the study investigates the following questions:

- What are the informational needs and preferences of clinicians for understanding and appropriately using AI-driven recommendations on medical decision-making?
- What information, according to AI experts, should be communicated to clinicians to enable their safe and appropriate use of AI-CDSS?

Methods

Study Design

This qualitative description design study [35] was conducted using semistructured interviews to explore how to effectively communicate information about AI-CDSS in medical decision-making. Interviews were conducted among two groups: (1) clinicians, to understand their preferences and informational needs, and (2) AI experts, to gain insights into what experts consider necessary for clinicians to safely and appropriately use AI tools. For this study, AI expertise was established based on practical professional experience rather than a specific formal education in AI. The expert group consisted of individuals with direct experience in developing or implementing AI models for or in the health care sector. The inclusion of both groups was based on the assumption that informational needs go beyond the clinicians' expressed preferences, but also expert-identified information is necessary for informed and safe use of AI in clinical settings. This article was written in accordance with the COREQ (Consolidated Criteria for Reporting Qualitative Research) checklist (Checklist 1) [36].

XSL•FO
RenderX

Recruitment of Participants

Study participants were recruited by convenience and snowball sampling between July and November 2024. A target sample size of 8 participants per group was predetermined. This fixed quota was established to accommodate the limited availability of clinicians while ensuring balanced representation between the 2 groups. Consequently, recruitment was concluded upon reaching this target rather than based on theoretical saturation. Potential participants were identified through professional networks associated with the research team and targeted searches on LinkedIn (Microsoft). The potential participants were contacted by email, accompanied by study information detailing the study's objectives, procedures, and confidentiality measures for the recording of the interviews. The interviewees interested in participating filled in a consent form and sent back an email to schedule a meeting, either online or face-to-face. Inclusion criteria required participants to be either (1) practicing clinicians in the Netherlands or (2) AI experts, entailing those with experience in developing and/or implementing diagnostic or treatment recommendation AI models within the health care setting. Participants who had less than 2 years of work experience were excluded.

Data Collection

The semistructured interviews were conducted by a researcher (SM), who is proficient in both Dutch and English. The interviews followed an interview guide (Multimedia Appendices 1 and 2), which contained different questions for clinicians and AI experts, each consisting of 4 parts (Table 1). The guides were translated to English for those with no sufficient command of the Dutch language by researchers SM and EA.

Table . Outline of the interview guide topics for clinicians and artificial intelligence (AI) experts regarding AI-driven clinical decision support systems (AI-CDSS) use.

Clinicians	AI[a] experts
Demographic information	
Experience with AI in care settings	Experience developing AI for care settings
Current received information about AI models	Provided information concerning AI models
Information they would prefer to receive	Information they think they should provide
Examples of the Model Card, TRIPOD-AI,[b] and Model Facts were shown, and participants were requested to provide feedback concerning amount, structure, and comprehensiveness of the information[c]	

[a]AI: artificial intelligence.

[b]TRIPOD: Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis–Artificial Intelligence.

[c]This part was equal for clinicians and AI experts. Demographic information was also available for both the clinicians as well as for the AI experts.

In part one, the participants were asked demographic questions concerning their age, medical or technical specialization, and work experience history. In the second part, clinicians were asked about their experiences with AI in the health care setting and about instructions they received before using AI. AI experts received questions about their experience in developing AI for clinical use and which information they provided to the end users. Part three contained questions about which information they would ideally either receive as clinical end users or provide as AI developers.

In the last part, all participants received an example of a Model Card [32], Model Facts [33], and the TRIPOD-AI checklist [22,24] (Multimedia Appendix 3). The formats were discussed sequentially: participants reviewed one example and provided immediate feedback on its information before moving to the next. This stepwise, fixed sequence ensured that each format was evaluated individually. These three formats were selected because they originate from distinct contexts and serve different purposes. Model Cards are intended for communication among developers and are therefore more technical in nature. In contrast, Model Facts are specifically designed for clinicians. The TRIPOD-AI checklist is a reporting standard on AI models in academic publications. To create these examples, a Bayesian network was selected from a public repository, as it provided a balanced level of complexity, being more complex than a linear regression but still more transparent than a deep learning model. One researcher (EA) completed each format based on the available model information. To preview a variation in communication styles, the same type of information was presented in bullet points in one format, while in another format, it was written in narrative form, and yet another displayed this information in a table. After examining the templates, participants were asked to provide feedback on the amount, structure, and comprehensiveness of the information listed in these examples.

Research Group

SMcandidate in Clinical Data Science (CDS) at Maastricht University, working in a department that is generally tech-positive, which may have shaped her perspectives on AI in health care. Over the course of this research project, SM engaged with different paradigms, initially approaching research from a postpositivist stance but increasingly shifting toward a constructivist orientation. This perspective informs the study's recognition that a single "ground truth" of information provision is not attainable and that knowledge is context-dependent and co-constructed.

HP is a PhD student researching trust in AI within CDS. MT is a PhD student at Maastro Clinic, researching risk communication and shared decision-making in breast cancer radiation therapy, with a focus on patient decision aids. EA is a PhD student within CDS investigating AI model descriptions based on the model

XSL•FO
RenderX

representation standards and requirements. AvH is a PhD student in shared decision-making within radiation oncology (CDS and Maastro), with advanced training in qualitative methodology, and advised SM on qualitative methods. RF is an assistant professor in AI and clinical decision-making within CDS and Maastro and supervises the PhD candidates.

Data Analysis

After data collection, recordings were transcribed verbatim and analyzed according to codebook thematic analysis [37] using the Atlas.ti software (version 25.0.1; ATLAS.ti Scientific Software Development GmbH). This method helped identify common themes and topics that were mentioned repeatedly in interviews. The analysis process started with open, inductive coding by a researcher (SM). As no predefined coding framework was used, codes were created to descriptively summarize the underlying meaning of the selected text segments. This initial phase resulted in the construction of a preliminary codebook. Two other researchers (MT and HP) independently applied codes from the codebook constructed by the first researcher (SM) to a subset of the interviews (n=4). Disagreements in codes were resolved by revisiting the transcripts to verify context and debating code definitions until consensus was reached. This iterative process changed the codebook by renaming codes, refining code definitions, ensuring consistent application of the codes, and merging overlapping codes. These codes were later grouped into general themes by identifying clusters within assigned codes for clinicians (Multimedia Appendix 4) and for AI experts (Multimedia Appendix 5). A member check of the analysis was not conducted to avoid influencing participants' perspectives in potential follow-up research.

Ethical Considerations

Before the start of the study, the Research Ethics form of the Faculty of Health, Medicine, and Life Sciences (FHML-REC, FHML/HDT/2024.022.) was evaluated, and ethical approval was granted. Prior to the interviews, all participants received an information letter detailing the study's objectives and data handling procedures. Written informed consent was obtained from all participants for both participation and audio recording. To ensure confidentiality, all transcripts were pseudonymized, and any identifying information was removed. Participants were provided access to both their audio recording and the corresponding transcript, which they could review at any time. Participants received no compensation for their participation.

Results

Participant Characteristics

In total, 26 clinicians and 11 AI experts were invited to participate. Among invited clinicians, 8 declined due to lack of time, while 10 did not provide a reason. Three invited AI experts also did not provide a reason for nonparticipation. Between August and November 2024, 16 interviews were conducted with clinicians (n=8) and AI experts (n=8). Half of the clinician interviews were held through Microsoft Teams, and the other half were face-to-face, whereas most AI expert interviews were face-to-face (n=6), 2 AI expert interviews were held in English, while all remaining interviews, including all clinician interviews, were conducted in Dutch. Interviews lasted between 30 minutes and an hour. All participating clinicians practiced within tertiary care settings, employed either at academic medical centers or a specialized institute. Concerning prior experience with AI among clinicians, all reported having encountered AI tools in their professional practice. The majority (n=5) had experience with prediction models; some (n=2) had worked exclusively with AI-driven imaging applications or generative AI in the form of chatbots (n=1). Participant characteristics are presented in Table 2.

Through conducting thematic analysis, 4 key themes were identified that clinicians and AI experts expressed as essential for effective communication and use of AI-CDSS. These themes include (1) understanding the target population, which refers to the need to understand the data on which the model was trained; (2) clinically meaningful outcomes, focusing on how model reliability and clinical relevance are communicated; (3) warnings, limits, and safe use, highlighting the importance of clearly defining boundaries, limitations, and misuse risks; and (4) accessible design, which addresses the structure and presentation of content. These themes and subthemes can be found in Table 3 and are discussed in detail below.

XSL•FO
RenderX

Table . Demographic and professional characteristics of the 16 participants interviewed between August and November 2024 in the Netherlands.

Characteristics	Clinicians (n=8)	AI[a] experts (n=8)
Gender		
Men	6	5
Women	2	3
Age (years), median (range)	44 (29–59)	37 (28–48)
Work experience (years), median (range)	14 (2–22)	11 (3–21)
Specialties or roles (n)	• Internal medicine (2) • Oncology (2) • Gastroenterology (2) • Cardiology (1) • Radiology (1)	• Medical imaging AI (3) • Innovation and implementation management (3) • Data infrastructure and FAIR[b] principles (2)

[a]AI: artificial intelligence.

[b]FAIR: Findable, Accessible, Interoperable, and Reusable.

Table . Overview of themes and subthemes derived from interviews with clinicians and artificial intelligence (AI) experts regarding informational needs and preferences for using AI-driven clinical decision support systems (AI-CDSS).

Theme	Subthemes
Understanding the target population	• Origin of the data • Class imbalance • Inclusion and exclusion • Training data size
Clinically meaningful outcomes	• Difficult AUC[a] • Uncertainty bands • Performance graphs and comparative visualizations
Warnings, limits, and safe use	• Underused, poorly highlighted, and filled in too vaguely • Active warnings
Accessible design	• Available at all times • Language barrier • Customizable or layered information provision

[a]AUC: area under the curve.

Understanding the Target Population

A central theme expressed by most clinicians and experts was the ability to compare an individual patient during consultation with the training population for the AI model. This was fundamental for clinicians to evaluate whether the recommendations from the model would be reliable in a specific clinical context or not.

> *What I think they should know is on which population it (AI-CDSS) is based. So for which patient are they allowed to use it and for which patient are they not. If they have a new patient in front of them, when does it (the patient) belong within the population of the model.* [AI Expert 2]

Initially, participants described this need as knowledge about the origin of the data, more specifically why, where, and how the data were collected. Clinicians stressed that this way, they could extrapolate whether this reflected local populations, which was especially important when the model was not locally validated.

> *Where is the population from? This is clearly a Dutch population, I can extrapolate this to our own thing, but I would want to see the patient characteristics.* [Clinician 4]

Although the concern about the representativeness of the data and its relevance to local populations was present from the start, it became more concrete when participants reviewed the TRIPOD-AI example. Clinicians remarked that the section “class imbalance” was of great importance to make the comparison. However, the way it was written in the example felt overwhelming and too technical. Solutions for this were given, such as a table or graph to show the class imbalance, but also a digital and interactive solution for training data comparisons. AI experts already stressed the importance of providing information about the class imbalance before being shown the TRIPOD-AI example.

> *(Showed class balance in TRIPOD-AI example) Yes, basically yes, yes. Look, and if you see that they’re all N1 patients, then is it also still suitable for N2? If your patient happens to be N2, so all those kind of*

XSL•FO
RenderX

> *things. Yeah I think that's really important to take into account.* [Clinician 5]

Later in the TRIPOD-AI document, the inclusion and exclusion graph for the training data, which is often provided in papers, was shown. Clinicians stressed that without understanding who was systematically left out of the training data, whether by design or by oversight, they felt the context to interpret the applicability lacked.

> *For example, it doesn't apply to pregnancy or people who are overweight with a BMI over 30, I'm just saying something. Because otherwise, I think you lose sight of that when you're working in practice.* [Clinician 7]

Finally, both clinicians and AI experts noted that information about the training data size was missing in the Model Facts example. This was, however, crucial information according to clinicians, since a great size of the training data is associated with better model performance.

> *And actually, I'd want to know how many patients there were, like I'd want to know for reliability, how many patients were in the model so I can estimate how good it is.* [Clinician 4]

Clinically Meaningful Outcomes

Another key theme was the need for clear evidence of the performance and reliability of the model. Both groups stressed that for an AI-CDSS to be integrated into clinical practice, its performance must be demonstrably robust and easily interpretable. However, the manner in which the performance should be presented, as well as when it was acceptable, was not agreed upon by participants.

Clinicians, in particular, emphasized the importance of understanding what the reported performance metrics mean in practice. They noted that values like area under the curve (AUC), while technically informative, do not always translate into clinical insight without additional context. Some participants mentioned a desire for performance measures to align more closely with clinical outcomes, like sensitivity and specificity and their relevant thresholds. Clinicians and AI experts suggested that these measures can also easily be translated into statistical language such as false positives or negatives, which was thought to be more familiar to clinicians.

> *The AUC just gives an overall picture of whether it is a good predictive model. But if it's 0.88 and the balance between sensitivity and specificity for your specific question isn't optimal, then it still isn't a good model.* [Clinician 3]

Experts mentioned that besides the performance measures, there should also be a focus on providing an uncertainty band to clinicians. This should give insights into when the model is handling an edge case. AI experts noted that in the AI development at the time of this writing, developers chase a high-performance score such as the AUC but might overlook clinical utility or model robustness by doing so.

> *We often try to get that AUC as high as possible, but that applies to the entire population, so maybe we should start saying, maybe we need to develop models that actually filter out the worst-off patients or the ones who do the best.* [AI expert 2]

Ultimately, while clinicians overall claimed to be familiar with performance metrics, they expressed that information on model validation often remains buried in dense technical language or tables. However, to be actionable in a clinical setting, they emphasized that information must be presented in a more structured and visually digestible manner, such as performance graphs and comparative visualizations. Besides presenting the performance metrics, clinicians expressed a desire to know when they could use an outcome to make a treatment decision, in other words, when it was "good enough" to rely on.

> *Yes, these are always somewhat abstract concepts if you don't work with them daily. And yes, each time I can reason it out – like, what it means – but I can't immediately say whether it's good or not, right?* [Clinician 5]

Warnings, Limits, and Safe Use

An important theme for both clinicians and experts was the need for clear warnings and boundaries for AI-CDSS use. Explicit indicators or limitations were desired, especially when the system may be applied in situations that fall outside its intended use or scope.

Participants praised that limitations and inappropriate uses were included in the presented reporting standards. However, many AI experts emphasized that these sections were often underused, poorly highlighted, and filled in too vaguely. AI experts also mentioned that biases and limitations of the model are difficult to formulate and fill in with only their background as developers. They mentioned that defining these biases and formulating necessary warnings for misuse of the model should be done in interdisciplinary settings. They also emphasized that misuses of the model can be identified during prototype testing or practice sessions.

> *So I would say that, yes, this requires critical evaluation. The problem, I think, is that if someone developing the model doesn't have much clinical background, it becomes very difficult for them to judge what is critical and what should be reported.* [AI expert 3]

Besides explicitly mentioning these limitations, participants also stressed that more active warnings should be included. AI experts mentioned that it would be safest to lock down the scope of the AI-CDSS within the software, such as limiting input variables or blocking predictions outside the scope. However, clinicians had varying opinions about this, with some arguing that they can decide on the model scope themselves. Another suggestion made by AI experts to warn clinicians about the correct use of an AI-CDSS was to provide pop-ups when using the model out of scope. However, opinions amongst clinicians were divided, with concerns regarding pop-up tiredness and the pop-up not being read or respected. Similarly, they thought that disclaimers were likely not to be read or forgotten about during the clinical use of the model.

XSL•FO
RenderX

> *There are people in their 60s who function at the level of someone over 70, or even over 80, so to speak. And there are people who are 72 where you think, well, you actually function better than your calendar age would suggest. So it's all a matter of interpretation. And I think that this is something that should be left to a clinician.* [Clinician 8]

Accessible Design

Participants emphasized that how information is presented is just as critical as the content itself. Even when essential data about the system is available, it may be ignored or misunderstood in clinical settings if it is not accessible, clearly structured, or presented in a usable format.

First, the information needs to be available at all times while using the AI-CDSS, according to participants. They stressed that information should preferably be integrated within the software. However, it was also thought to be important to distribute information in different ways, such as presentations, workshops, and emails. According to participants, clinicians should also familiarize themselves with the model, which can be done through e-learning and practice sessions, which is important for retention of their understanding of the model.

> *But if you, for example, give a presentation and a doctor can actually use your model that same day, then that might also have a big impact on retention.* [Clinician 1]

Second, participants stressed the importance of the information being conveyed in an understandable manner. AI experts often spoke about the "language barrier" they experience when communicating about AI with clinicians, and clinicians confirmed this gap when they remarked on developer jargon. Participants suggested communicating in statistical language, such as *P* values or false positives, which is generally more in line with clinicians' education and academic training. Others suggested shrinking this "language barrier" by providing general AI training for clinicians, for example, from the hospital or during medical training. Another familiar way of providing information would be in the form of papers, which clinicians are accustomed to reading. The TRIPOD-AI example was well received by participants, as its resemblance to the layout of traditional academic papers appeared to create a sense of familiarity. However, some clinicians stressed that they then might as well read the entire paper, while AI experts showed significant information was still missing due to word count restraints.

> *I think that, in any case, education about AI will simply be necessary for everyone in the near future, because it's gradually being used more and more in medicine. And it's important to understand how it works or how something like that even comes about.* [Clinician 3]

Third, clinicians often labeled themselves as being intrigued by AI and intrinsically motivated to gather information about new technologies. However, many clinicians also stressed that they themselves or their coworkers would not have enough time to read up about a model. This shows a difference in needs between clinicians, which calls for customizable or layered information provision. An example of how this works out in practice would be through providing information blocks or pop-ups on demand or linking to (validation) papers and other useful resources.

> *And references – well, of course people aren't going to look at them very much either, so they quickly become less important. And people don't have endless time.* [Clinician 6]

Finally, of all the examples, the Model Card example was perceived as the least desired method by the participants, while the first page of the Model Facts was the most appreciated. However, according to participants, Model Facts still missed essential information blocks that were provided in the TRIPOD-AI, such as class imbalance, inclusion and exclusion criteria, and population size. Participants recommended that this information be integrated as much as possible within the software to use the AI-CDSS.

> *Well, if you want to know some of the background, then the TRIPOD would be my preference. But if it's purely about how to apply it in the clinic, then the Model Facts is fine – it's a bit shorter and more focused on clinical applicability.* [Clinician 2]

Discussion

This qualitative study explored the informational needs and preferences of clinicians for understanding and appropriately using AI-CDSS through semistructured interviews. Besides, this study explored AI experts' perspectives on what information should be communicated to clinicians to enable safe and appropriate use of AI-CDSS. From the thematic analysis, four key domains were identified that the participants considered essential: (1) understanding the population the model was trained on, (2) interpreting performance metrics in a clinically meaningful way, (3) being clearly informed about limitations and boundaries of use, and (4) having the information presented in an accessible and integrated format. Elements of the presented Model Facts, Model Card, and TRIPOD-AI examples were seen as valuable in addressing these needs, though none were considered fully sufficient on their own.

The first aspect concerns the ability to assess whether the patient in front of the clinician matches the training population of the AI-CDSS. Information on the data's origin, distribution, and size was thus seen as crucial to avoid inappropriate extrapolation and potential harm. This aligns with guidelines made for AI in clinical research in which proper use of predictive models increases when end users understand the AI models better [28,29]. At the time of this writing, many technical solutions aim to address dataset limitations, such as undersampling in the presence of class imbalances [38]. However, little research exists on how to equip end users to recognize and act upon known data limitations in real-time use. The TRIPOD-AI example was praised by participants for including elements like class imbalance and inclusion and exclusion criteria. However, participants suggested more use of visual tools or interactive comparison modules, which reflects a desire for more transparency delivered in comprehensive formats.

XSL•FO
RenderX

Additionally, there was a need to understand model performance in ways that directly inform clinical decision-making. Although clinicians appreciated seeing performance metrics, like the one shown in the Model Facts table, they found that metrics such as the AUC lacked context. This was also the case in the study performed by Frey et al [39], in which the Model Facts and the table containing information on performance and validation were not understood. These concerns are in line with other existing literature that argues for optimizing performance reporting to the clinical context and translating technical metrics into practical terms [40]. Instead, participants preferred metrics linked to treatment-relevant thresholds. Several AI experts also pointed out that communicating uncertainty is vital for identifying edge cases where model performance might be unreliable. Performance should therefore not only be reported on a global level but also in ways that match the decision points clinicians face.

Another key finding concerned the need for transparent warnings and clear boundaries for appropriate model use. While participants acknowledged that standards such as TRIPOD-AI and Model Facts do include these elements, these sections are often too vaguely phrased, making them less actionable for clinicians. The importance of clear, structured warnings is well established in other areas of health care, for example, in drug labels, which provide standardized information on indications, contraindications, and potential risks [41]. Such an approach would move beyond abstract disclaimers, offering clinicians tangible guidance to support responsible decision-making. Besides, some participants proposed integrating restrictions directly into the software, like locking out-of-scope inputs or showing warnings when the model is applied beyond its intended population. However, these suggestions were met with mixed reactions due to concerns about alert fatigue, which is an already frequently experienced mental strain in clinical decision support tools [42,43].

Finally, participants focused on how the information is structured and delivered to the end user. Participants advocated for layered information design, in which summary-level information is accessible at first glance, with additional detail available on demand [44]. This design principle is widely used in both health care settings, for example, in patient decision aids [45,46], and in other industries or webpages, like Wikipedia (Wikimedia Foundation). Interactive data visualizations were also suggested for understanding how individual patients compare to the broader training dataset. This preference mirrors trends in other industries, such as finance and data analytics, where interactive visual tools enhance sense-making and decision quality [47]. Clinicians also emphasized the importance of integrating information in the AI-CDSS interface and supporting information dissemination through workshops or e-learning modules. While the Model Facts format was seen as more accessible and clinically relevant, it lacked essential components, including training data information, which were better covered in TRIPOD-AI.

In general, participants throughout this study consistently emphasized that AI-CDSS can only be made safe, effective, and clinically relevant through co-creation with clinicians at every stage of development and deployment. This would start with jointly defining the model's purpose, identifying which clinical decisions it should support, and determining what constitutes a "good enough" model in clinical practice. Such early alignment likely helps to determine whether the model's goals are both clinically valuable and technically feasible. Beyond initial development, it seems that clinicians should also be involved in designing the model to fit existing workflows, identifying potential limitations, and establishing appropriate safeguards. AI experts noted they cannot define these elements alone because these require clinical insight. While the value of co-creation in AI development is well recognized in theory, it remains inconsistently practiced [48,49]. One approach formalizing co-creation is design thinking, which is a user-centered, iterative process that emphasizes empathy, problem definition, rapid prototyping, and testing in real-world contexts [50]. Embedding such methods more systematically into AI development is essential to close the gap between technical promise and clinical utility [51].

Although this study was rooted in the Dutch tertiary care context, the findings likely extend to other modernized health care systems preparing for AI implementation. Particularly, these findings are transferable within Europe, given the European AI Act's mandate for ensuring AI literacy among professional users. This regulation requires end users across all member states to be competent in using high-risk AI systems [52]. The educational and informational needs identified here are directly relevant to compliance efforts across the continent. Beyond the regulatory context, the core themes reflect universal medical concerns, extending the study's relevance to other modernized health care systems globally.

The results of this study should be interpreted in light of several limitations. First, the use of convenience sampling may have introduced selection bias, as individuals who voluntarily participate in research are often more engaged with and positive about the topic. Second, while the study aimed to explore clinicians' informational needs regarding AI-CDSS, many clinicians seemed to lack in-depth understanding of AI, which may have limited their ability to fully articulate their needs. To address this, perspectives from AI experts were included to enrich the findings. Third, the interdisciplinary approach introduced heterogeneity, due to participants coming from a range of professional backgrounds or clinical specialties. While this breadth is a strength, allowing for a more comprehensive understanding of cross-disciplinary needs, it may also obscure discipline-specific differences. For example, certain formats or tools may be more relevant or familiar to specific specialties. Fourth, the examples of AI model documentation were presented to all participants in a fixed order. The lack of randomization in the sequence may have introduced order effects, such as primacy bias, which could have influenced comparative feedback. Future research should explore how disciplinary context shapes informational needs and preferences, ideally through larger-scale studies using quantitative methods that allow for broader demographic and professional representation and more robust comparisons across subgroups.

In conclusion, when using AI in clinical practice, it is important that clinicians use these technologies in a safe and informed way. However, this information provision likely falls short in

XSL•FO
RenderX

meeting their informational needs, contributing to the lack of large-scale adoption. AI-CDSS developers should clearly communicate both who the model is for, including characteristics of its training data, and how it performs in clinically meaningful terms. This must include explicit limitations and appropriate warnings, presented in a way that is understandable, visual, and ideally interactive. While existing formats such as Model Cards, Model Facts, and TRIPOD-AI offer valuable starting points, none fully meet these needs on their own. Future efforts should consider combining and adapting elements from these formats to better support clinical use. Effective integration of AI-CDSS requires collaboration with clinicians during development and targeted education during implementation. Addressing informational needs of clinicians is critical to ensure appropriate use and supports the safe and effective integration of AI-CDSS into clinical practice.

Acknowledgments

The authors would like to thank the clinicians and AI experts who participated in this study. Their time and insights resulted in thorough discussions, which were essential to advancing the understanding of informational needs related to AI-driven clinical decision support systems.

The authors declare the use of generative AI in the research and writing process. According to the GAIDeT taxonomy [53], the following tasks were delegated to generative AI tools under full human supervision: (1) idea generation and (2) proofreading and editing. The generative AI tool used was ChatGPT 4. Responsibility for the final manuscript lies entirely with the authors. Generative AI tools are not listed as authors and do not bear responsibility for the final outcomes.

Funding

This research was funded by the Dutch Research Council (NWO).

Data Availability

The datasets gathered and analyzed during the study are not publicly available due to the presence of potentially identifiable information within the interview transcripts. However, deidentified excerpts or relevant portions of the data may be made available from the corresponding author upon reasonable request and subject to institutional and ethical approvals.

Authors' Contributions

Conceptualization: SM, EA, JvS, LH, OD, RF
Data curation: SM
Formal Analysis: SM, HP, MT
Investigation: SM, HP, MT
Methodology: SM, EA, AvH, JvS, LH, OD, RF
Project administration: SM, LH, RF
Resources: SM, EA, RF
Supervision: LH, OD, RF
Writing – original draft: SM
Writing – review & editing: SM, HP, MT, EA, AvH, JvS, LH, AD, OD, RF

Conflicts of Interest

The primary author declares that they have no known competing financial interests. The primary author is affiliated with a department where AI-based systems are developed. Some of the authors receive research grants related to the development and implementation of AI. In addition, some of the respondents interviewed were known to the primary author prior to the study. JvS and AD are shareholders and receiving salary from Medical Data Works BV, although company activities are unrelated to this manuscript. Outside of the submitted work, RF's research group receives funding from public-private research consortia, including contributions from Janssen-Cilag and Takeda. These companies had no role in the present study.

Multimedia Appendix 1
Interview guide clinicians.
[DOCX File, 19 KB - mededu_v12i1e85228_app1.docx]

Multimedia Appendix 2
Interview guide artificial intelligence (AI) experts.
[DOCX File, 20 KB - mededu_v12i1e85228_app2.docx]

Multimedia Appendix 3

XSL•FO
RenderX

Examples of Model Card, Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis–Artificial Intelligence (TRIPOD-AI), and Model Facts.
[DOCX File, 38 KB - mededu_v12i1e85228_app3.docx]

Multimedia Appendix 4
Clusters with themes clinicians.
[DOCX File, 433 KB - mededu_v12i1e85228_app4.docx]

Multimedia Appendix 5
Clusters with themes artificial intelligence (AI) experts.
[DOCX File, 395 KB - mededu_v12i1e85228_app5.docx]

Checklist 1
Consolidated Criteria for Reporting Qualitative Research (COREQ) checklist.
[DOCX File, 27 KB - mededu_v12i1e85228_app6.docx]

References

1. Aung YYM, Wong DCS, Ting DSW. The promise of artificial intelligence: a review of the opportunities and challenges of artificial intelligence in healthcare. Br Med Bull 2021 Sep 10;139(1):4-15. [doi: 10.1093/bmb/ldab016] [Medline: 34405854]
2. Ankolekar A, van der Heijden B, Dekker A, et al. Clinician perspectives on clinical decision support systems in lung cancer: implications for shared decision-making. Health Expect 2022 Aug;25(4):1342-1351. [doi: 10.1111/hex.13457] [Medline: 35535474]
3. Goirand M, Austin E, Clay-Williams R. Implementing ethics in healthcare AI-based applications: a scoping review. Sci Eng Ethics 2021 Sep 3;27(5):61. [doi: 10.1007/s11948-021-00336-3] [Medline: 34480239]
4. Shaw J, Rudzicz F, Jamieson T, Goldfarb A. Artificial intelligence and the implementation challenge. J Med Internet Res 2019 Jul 10;21(7):e13659. [doi: 10.2196/13659] [Medline: 31293245]
5. Gama F, Tyskbo D, Nygren J, Barlow J, Reed J, Svedberg P. Implementation frameworks for artificial intelligence translation into health care practice: scoping review. J Med Internet Res 2022 Jan 27;24(1):e32215. [doi: 10.2196/32215] [Medline: 35084349]
6. Petersson L, Larsson I, Nygren JM, et al. Challenges to implementing artificial intelligence in healthcare: a qualitative interview study with healthcare leaders in Sweden. BMC Health Serv Res 2022 Jul 1;22(1):850. [doi: 10.1186/s12913-022-08215-8] [Medline: 35778736]
7. Davenport T, Kalakota R. The potential for artificial intelligence in healthcare. Future Healthc J 2019 Jun;6(2):94-98. [doi: 10.7861/futurehosp.6-2-94] [Medline: 31363513]
8. Kelly CJ, Karthikesalingam A, Suleyman M, Corrado G, King D. Key challenges for delivering clinical impact with artificial intelligence. BMC Med 2019 Oct 29;17(1):195. [doi: 10.1186/s12916-019-1426-2] [Medline: 31665002]
9. Esmaeilzadeh P. Challenges and strategies for wide-scale artificial intelligence (AI) deployment in healthcare practices: a perspective for healthcare organizations. Artif Intell Med 2024 May;151:102861. [doi: 10.1016/j.artmed.2024.102861] [Medline: 38555850]
10. Smith H, Downer J, Ives J. Clinicians and AI use: where is the professional guidance? J Med Ethics 2024 Jun 21;50(7):437-441. [doi: 10.1136/jme-2022-108831] [Medline: 37607805]
11. Jabbour S, Fouhey D, Shepard S, et al. Measuring the Impact of AI in the diagnosis of hospitalized patients: a randomized clinical vignette survey study. JAMA 2023 Dec 19;330(23):2275-2284. [doi: 10.1001/jama.2023.22295] [Medline: 38112814]
12. Zhao A, Larbi M, Miller K, O'Neill S, Jayasekera J. A scoping review of interactive and personalized web-based clinical tools to support treatment decision making in breast cancer. Breast 2022 Feb;61:43-57. [doi: 10.1016/j.breast.2021.12.003] [Medline: 34896693]
13. Völkel V, Hueting TA, Draeger T, et al. Improved risk estimation of locoregional recurrence, secondary contralateral tumors and distant metastases in early breast cancer: the INFLUENCE 2.0 model. Breast Cancer Res Treat 2021 Oct;189(3):817-826. [doi: 10.1007/s10549-021-06335-z] [Medline: 34338943]
14. van Maaren MC, Hueting TA, Völkel V, van Hezewijk M, Strobbe LJ, Siesling S. The use and misuse of risk prediction tools for clinical decision-making. Breast 2023 Jun;69:428-430. [doi: 10.1016/j.breast.2023.01.006] [Medline: 36709092]
15. Collins GS, Moons KGM. Reporting of artificial intelligence prediction models. The Lancet 2019 Apr;393(10181):1577-1579. [doi: 10.1016/S0140-6736(19)30037-6]
16. Fehr J, Citro B, Malpani R, Lippert C, Madai VI. A trustworthy AI reality-check: the lack of transparency of artificial intelligence products in healthcare. Front Digit Health 2024;6:1267290. [doi: 10.3389/fdgth.2024.1267290] [Medline: 38455991]

XSL•FO
RenderX

17. Van Calster B, Wynants L, Timmerman D, Steyerberg EW, Collins GS. Predictive analytics in health care: how can we know it works? J Am Med Inform Assoc 2019 Dec 1;26(12):1651-1654. [doi: 10.1093/jamia/ocz130] [Medline: 31373357]
18. Campbell DJ. The clinical utility curve: a proposal to improve the translation of information provided by prediction models to clinicians. BMC Res Notes 2016 Apr 14;9(1):219. [doi: 10.1186/s13104-016-2028-0] [Medline: 27080381]
19. Meshaka R, Pinto Dos Santos D, Arthurs OJ, Sebire NJ, Shelmerdine SC. Artificial intelligence reporting guidelines: what the pediatric radiologist needs to know. Pediatr Radiol 2022 Oct;52(11):2101-2110. [doi: 10.1007/s00247-021-05129-1] [Medline: 34196729]
20. Yao S, Shen P, Dai T, et al. Human understandable thyroid ultrasound imaging AI report system - a bridge between AI and clinicians. iScience 2023 Apr 21;26(4):106530. [doi: 10.1016/j.isci.2023.106530] [Medline: 37123225]
21. Boag W, Hasan A, Kim JY, et al. The algorithm journey map: a tangible approach to implementing AI solutions in healthcare. NPJ Digit Med 2024 Apr 9;7(1):87. [doi: 10.1038/s41746-024-01061-4] [Medline: 38594344]
22. Collins GS, Reitsma JB, Altman DG, Moons KGM. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD Statement. BMC Med 2015 Jan 6;13(1):1. [doi: 10.1186/s12916-014-0241-z] [Medline: 25563062]
23. Collins GS, Dhiman P, Andaur Navarro CL, et al. Protocol for development of a reporting guideline (TRIPOD-AI) and risk of bias tool (PROBAST-AI) for diagnostic and prognostic prediction model studies based on artificial intelligence. BMJ Open 2021 Jul 9;11(7):e048008. [doi: 10.1136/bmjopen-2020-048008] [Medline: 34244270]
24. Collins GS, Moons KGM, Dhiman P, et al. TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. BMJ 2024 Apr 16;385:e078378. [doi: 10.1136/bmj-2023-078378] [Medline: 38626948]
25. Liu X, Cruz Rivera S, Moher D, Calvert MJ, Denniston AK, SPIRIT-AI and CONSORT-AI Working Group. Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI extension. Nat Med 2020 Sep;26(9):1364-1374. [doi: 10.1038/s41591-020-1034-x] [Medline: 32908283]
26. Vasey B, Nagendran M, Campbell B, et al. Reporting guideline for the early-stage clinical evaluation of decision support systems driven by artificial intelligence: DECIDE-AI. Nat Med 2022 May;28(5):924-933. [doi: 10.1038/s41591-022-01772-9] [Medline: 35585198]
27. Cruz Rivera S, Liu X, Chan AW, et al. Guidelines for clinical trial protocols for interventions involving artificial intelligence: the SPIRIT-AI extension. Nat Med 2020 Sep;26(9):1351-1363. [doi: 10.1038/s41591-020-1037-7] [Medline: 32908284]
28. Topol EJ. Welcoming new guidelines for AI clinical research. Nat Med 2020 Sep;26(9):1318-1320. [doi: 10.1038/s41591-020-1042-x] [Medline: 32908274]
29. Volovici V, Syn NL, Ercole A, Zhao JJ, Liu N. Steps to avoid overuse and misuse of machine learning in clinical research. Nat Med 2022 Oct;28(10):1996-1999. [doi: 10.1038/s41591-022-01961-6] [Medline: 36097217]
30. Lavin A, Gilligan-Lee CM, Visnjic A, et al. Technology readiness levels for machine learning systems. Nat Commun 2022 Oct 20;13(1):6039. [doi: 10.1038/s41467-022-33128-9] [Medline: 36266298]
31. Fang H, Miao H. Introducing the model card toolkit for easier model transparency reporting. : Google Research; 2020 Jul 29 URL: https://research.google/blog/introducing-the-model-card-toolkit-for-easier-model-transparency-reporting/ [accessed 2026-02-25]
32. Mitchell M, Wu S, Zaldivar A, et al. Model cards for model reporting. FAT* '19: Proceedings of the Conference on Fairness, Accountability, and Transparency 2019 Jan 29:220-229. [doi: 10.1145/3287560.3287596]
33. Sendak MP, Gao M, Brajer N, Balu S. Presenting machine learning model information to clinical end users with model facts labels. NPJ Digit Med 2020;3(1):41. [doi: 10.1038/s41746-020-0253-3] [Medline: 32219182]
34. Schwartz LM, Woloshin S, Welch HG. Using a drug facts box to communicate drug benefits and harms: two randomized trials. Ann Intern Med 2009 Apr 21;150(8):516-527. [doi: 10.7326/0003-4819-150-8-200904210-00106] [Medline: 19221371]
35. Bradshaw C, Atkinson S, Doody O. Employing a qualitative description approach in health care research. Glob Qual Nurs Res 2017;4:2333393617742282. [doi: 10.1177/2333393617742282] [Medline: 29204457]
36. Tong A, Sainsbury P, Craig J. Consolidated criteria for reporting qualitative research (COREQ): a 32-item checklist for interviews and focus groups. Int J Qual Health Care 2007 Dec;19(6):349-357. [doi: 10.1093/intqhc/mzm042] [Medline: 17872937]
37. Braun V, Clarke V. One size fits all? What counts as quality practice in (reflexive) thematic analysis? Qual Res Psychol 2021 Jul 3;18(3):328-352. [doi: 10.1080/14780887.2020.1769238]
38. Salmi M, Atif D, Oliva D, Abraham A, Ventura S. Handling imbalanced medical datasets: review of a decade of research. Artif Intell Rev 2024;57(10). [doi: 10.1007/s10462-024-10884-2]
39. Frey N, Agha-Mir-Salim L, Hinz E, Poncette AS, Balzer F. Assessing healthcare stakeholder understanding of machine learning documentation. Stud Health Technol Inform 2025 May 15;327:318-322. [doi: 10.3233/SHTI250336] [Medline: 40380447]
40. Reyna MA, Nsoesie EO, Clifford GD. Rethinking algorithm performance metrics for artificial intelligence in diagnostic medicine. JAMA 2022 Jul 26;328(4):329-330. [doi: 10.1001/jama.2022.10561] [Medline: 35802382]
41. Sullivan HW, O'Donoghue AC, Aikin KJ. Primary care physicians' use of FDA-approved prescription drug labels. J Am Board Fam Med 2014;27(5):694-698. [doi: 10.3122/jabfm.2014.05.140039] [Medline: 25201939]

XSL•FO
RenderX

42. Kesselheim AS, Cresswell K, Phansalkar S, Bates DW, Sheikh A. Clinical decision support systems could be modified to reduce “alert fatigue” while still minimizing the risk of litigation. Health Aff (Millwood) 2011 Dec;30(12):2310-2317. [doi: 10.1377/hlthaff.2010.1111] [Medline: 22147858]
43. McGreevey JD, Mallozzi CP, Perkins RM, Shelov E, Schreiber R. Reducing alert burden in electronic health records: state of the art recommendations from four health systems. Appl Clin Inform 2020 Jan;11(1):1-12. [doi: 10.1055/s-0039-3402715] [Medline: 31893559]
44. Boxwala AA, Rocha BH, Maviglia S, et al. A multi-layered framework for disseminating knowledge for computer-based decision support. J Am Med Inform Assoc 2011 Dec;18 Suppl 1(Suppl 1):i132-i139. [doi: 10.1136/amiajnl-2011-000334] [Medline: 22052898]
45. van Strien-Knippenberg IS, Boshuizen MCS, Determann D, de Boer JH, Damman OC. Cocreation with Dutch patients of decision-relevant information to support shared decision-making about adjuvant treatment in breast cancer care. Health Expect 2022 Aug;25(4):1664-1677. [doi: 10.1111/hex.13510] [Medline: 35579109]
46. Roumen C, Hasannejadasl H, Swart R, et al. Breast cancer patients' most important quality of life themes for a radiotherapy decision aid. Breast 2022 Oct;65:8-14. [doi: 10.1016/j.breast.2022.06.002] [Medline: 35728438]
47. Perdana A, Rob A, Rohde F. Does visualization matter? The role of interactive data visualization to make sense of information. AJIS 2018;22. [doi: 10.3127/ajis.v22i0.1681]
48. Laka M, Carter D, Milazzo A, Merlin T. Challenges and opportunities in implementing clinical decision support systems (CDSS) at scale: interviews with Australian policymakers. Health Policy Technol 2022 Sep;11(3):100652. [doi: 10.1016/j.hlpt.2022.100652]
49. Swan EL, Peltier JW, Dahl AJ. Artificial intelligence in healthcare: the value co-creation process and influence of other digital health transformations. JRIM 2024 Jan 30;18(1):109-126. [doi: 10.1108/JRIM-09-2022-0293]
50. Davis J, Docherty CA, Dowling K. Design thinking and innovation: synthesising concepts of knowledge co-creation in spaces of professional development. The Design Journal 2016 Jan 2;19(1):117-139. [doi: 10.1080/14606925.2016.1109205]
51. Sreenivasan A, Suresh M. Design thinking and artificial intelligence: a systematic literature review exploring synergies. Int J Innov Stud 2024 Sep;8(3):297-312. [doi: 10.1016/j.ijis.2024.05.001]
52. Article 4: AI literacy. EU Artificial Intelligence Act. URL: https://artificialintelligenceact.eu/article/4/ [accessed 2025-12-18]
53. Suchikova Y, Tsybuliak N, Teixeira da Silva JA, Nazarovets S. GAIDeT (Generative AI Delegation Taxonomy): a taxonomy for humans to delegate tasks to generative artificial intelligence in scientific research and publishing. Account Res 2025 Aug 8:1-27. [doi: 10.1080/08989621.2025.2544331] [Medline: 40781729]

Abbreviations

AI: artificial intelligence
AI-CDSS: artificial intelligence-driven clinical decision support systems
AUC: area under the curve
CDS: Clinical Data Science
CONSORT-AI: Consolidated Standards Of Reporting Trials–Artificial Intelligence
COREQ: Consolidated Criteria for Reporting Qualitative Research
DECIDE-AI: Developmental and Exploratory Clinical Investigation of Decision support systems driven by Artificial Intelligence
FHML-REC: Faculty of Health, Medicine, and Life Sciences–Research Ethics Committee
SPIRIT-AI: Standard Protocol Items: Recommendations for Interventional Trials–Artificial Intelligence
TRIPOD-AI: Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis – Artificial Intelligence

Edited by A Stone, T Leung; submitted 03.Oct.2025; peer-reviewed by E Chuang, L Bai; revised version received 18.Dec.2025; accepted 07.Jan.2026; published 12.Mar.2026.

Please cite as:
Mingels S, Piehl H, Therrien M, Akhmad E, van Hienen A, van Soest J, Hochstenbach L, Dekker A, Damman O, Fijten R Understanding Clinicians' Informational Needs for AI-Driven Clinical Decision Support Systems: Qualitative Interview Study JMIR Med Educ 2026;12:e85228
URL: https://mededu.jmir.org/2026/1/e85228
doi:10.2196/85228

© Simone Mingels, Hannah Piehl, Madeline Therrien, Ekaterina Akhmad, Anniek van Hienen, Johan van Soest, Laura Hochstenbach, Andre Dekker, Olga Damman, Rianne Fijten. Originally published in JMIR Medical Education

XSL•FO
RenderX

(https://mededu.jmir.org), 12.Mar.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Education, is properly cited. The complete bibliographic information, a link to the original publication on https://mededu.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Value, Structure, and Curriculum in US Graduate Health Informatics Programs: Cross-Sectional Study

Suhila Sawesi[1], PhD; Diane M Dolezel[2], PhD; Pranitha Presingu[1], MS; Michael Irungu[1], MS

[1]Health and Bioinformatics Program, Department of Information Sciences and Technologies (IST), College of Computing, Grand Valley State University, 333 Michigan Street NE, Grand Rapids, MI, United States

[2]Department of Health Informatics & Information Management, Texas State University, Round Rock, TX, United States

Corresponding Author:
Suhila Sawesi, PhD
Health and Bioinformatics Program, Department of Information Sciences and Technologies (IST), College of Computing, Grand Valley State University, 333 Michigan Street NE, Grand Rapids, MI, United States

Abstract

Background: Graduate health informatics programs in the United States differ widely in cost, curriculum, and program design. However, it is unclear how these differences influence affordability, accreditation signaling, and preparation for a data-driven workforce.

Objective: This study aimed to evaluate the value (tuition and affordability), structure (delivery format, credit load, culminating experience, and accreditation), and curriculum (technology content emphasis) of US graduate health informatics programs. It examined how accreditation and modality relate to program design, and whether tuition-normalized curriculum breadth differed by accreditation status.

Methods: A cross-sectional study of 107 US graduate health informatics programs was conducted using publicly available data collected between January and May 2025. Tuition was standardized to cost per credit. Curricular content was coded for technology density and mapped to the Commission on Accreditation for Health Informatics and Information Management Education domains. Comparative statistics, regression models, and exploratory cluster analyses were used to assess relationships between tuition, credit requirements, accreditation, delivery format, and curriculum characteristics.

Results: Programs varied by delivery format, with 37 of 107 (34.6%) online, 32 of 107 (29.9%) hybrid, 23 of 107 (21.5%) in person, and 15 of 107 (14.0%) flexible. Credit requirements most commonly fell between 31 and 39 credits. Culminating experiences included capstone (54/107, 50.5%), internships (21/107, 19.6%), and thesis (7/107, 6.5%). Required credit hours showed modest variation by delivery format but not by accreditation status. Accreditation was not associated with differences in the tuition-normalized curriculum breadth structural proxy in this program-level analysis. Programs requiring internships had significantly higher mean credit loads than programs without internships (39.0 vs 31.3 credits; P=.005). Cluster analysis revealed 4 descriptive program configurations differentiated by cost, modality, credit requirements, and culminating experiences.

Conclusions: In this program-level descriptive analysis, accreditation status was not associated with differences in tuition-normalized curriculum breadth structural proxy. Instead, delivery format and internship requirements were descriptively associated with variation in credit load and cost. Improving transparency in tuition models and aligning program structure with curricular scope may support efforts to enhance equity and value in graduate health informatics education.

(JMIR Med Educ 2026;12:e87479) doi:10.2196/87479

KEYWORDS

health informatics; education; curriculum; accreditation; Human Capital Theory; graduate; tuition

Introduction

Background and Conceptual Framework

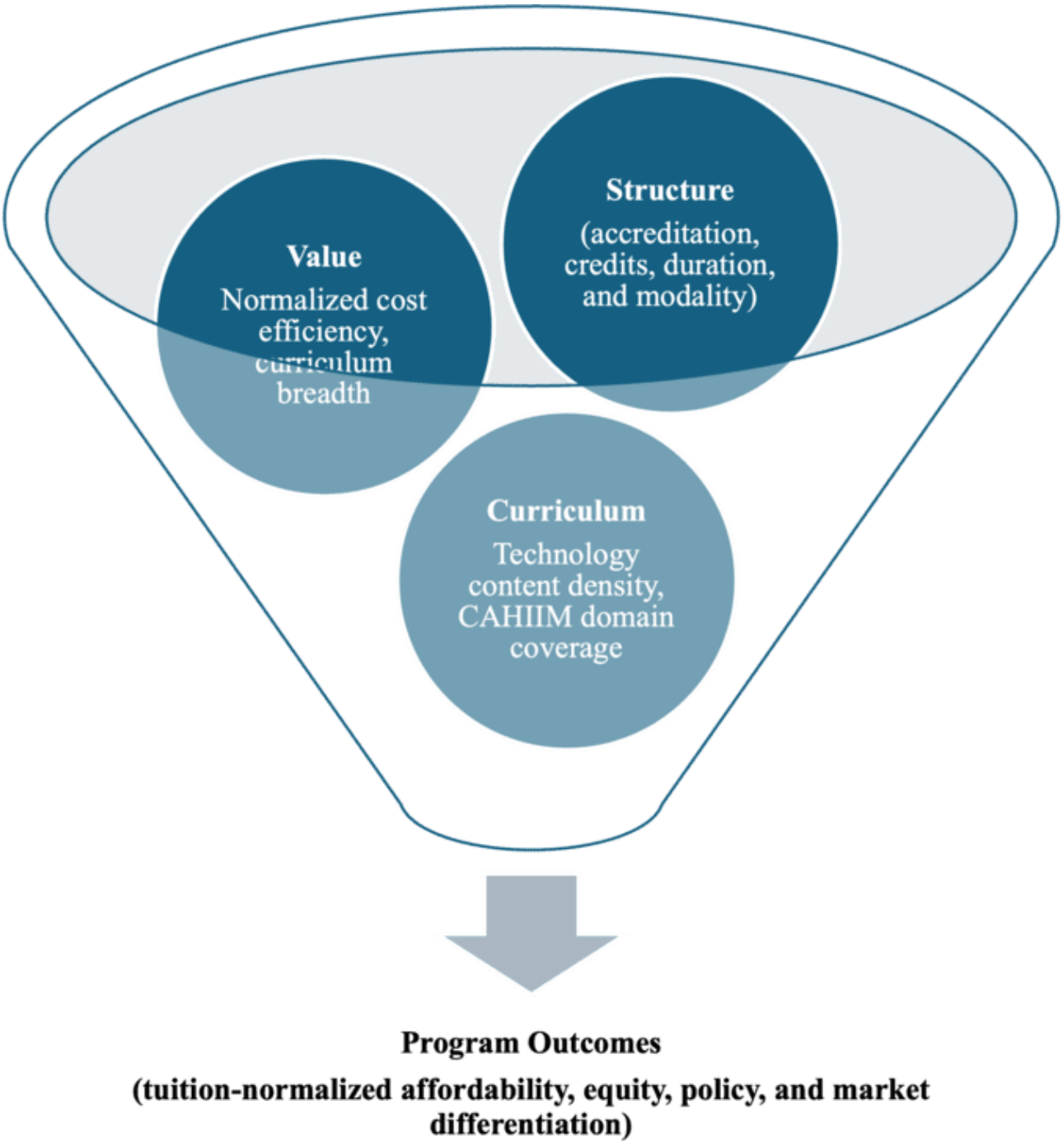
Graduate education in health informatics is evolving amid continued discussion regarding affordability, quality assurance, and measurable returns on investment. While undergraduate tuition growth has slowed, graduate education remains costly, particularly in health informatics, where financial liabilities can be substantial and returns unpredictable. Recent federal regulations mandating program-level disclosure of student debt and earnings have intensified attention to tuition structures, delivery modalities, and return on investment [1-3].

This study is grounded in Human Capital Theory (HCT) [4], which conceptualizes education as an investment that builds skills and yields economic returns. Within this framework (Figure 1), tuition represents the investment cost of acquiring human capital (Value), accreditation functions as a quality signal to the labor market (Structure), and curriculum reflects the skills

XSL•FO
RenderX

and competencies that enhance employability (Curriculum). Together, these domains shape outcomes related to affordability, transparency, equity, market positioning, and labor-market returns (eg, debt-to-earnings ratios). While this study uses HCT to categorize educational inputs, we do not measure actual labor market outcomes. In this context, HCT serves as a conceptual lens to evaluate the structural components of educational investment rather than as a framework for measuring postgraduation economic returns.

Figure 1. Conceptual framework: value, structure, and curriculum. This framework, grounded in Human Capital Theory, categorizes educational investment inputs into 3 core domains: Value (tuition-normalized curriculum breadth and credit-based costs), Structure (accreditation status, credit requirements, delivery modality, and program duration), and Curriculum (technology content density and CAHIIM domain coverage). These structural domains are analyzed as correlates of program-level outcomes, including affordability and market differentiation. CAHIIM: Commission on Accreditation for Health Informatics and Information Management Education.



Affordability and Delivery Modalities

Affordability discussions have shifted from headline tuition to a broader notion of value. Although undergraduate prices have decelerated, graduate programs continue to present affordability challenges due to high net prices and debt loads [1]. The 2023 Financial Value Transparency and Gainful Employment regulations require institutions to disclose program-level debt and earnings, sharpening scrutiny of tuition models and return on investment [2].

XSL•FO
RenderX

Modality further complicates costs: despite advertised uniform tuition across online, in-person, and hybrid formats, hidden or modality-specific fees can raise total expenses, with some evidence of an "online premium" that elevates all-in costs for distance learners [5]. At the same time, the credit hour, linking cost to program duration, remains contested. Federal definitions tie it to student effort and outcomes, whereas critics argue that reliance on the Carnegie Unit constrains competency-based and flexible models [5,6]. Despite heightened policy focus, comparative research remains scarce on how tuition per credit, credit loads, and delivery modality interact with accreditation across graduate health informatics programs.

Quality Assurance and Curricular Content

Quality assurance adds a second dimension. The Commission on Accreditation for Health Informatics and Information Management Education (CAHIIM) sets accreditation standards, while the American Medical Informatics Association (AMIA) defines master's-level core competencies [7,8]. Internationally, the International Medical Informatics Association has issued education recommendations (2010; updated 2023) that map skill domains to program design across degree levels [9,10]. However, it is unclear whether accreditation aligns with specific program structures (eg, credit loads, accelerated pathways, and culminating requirements) or with the technical content density needed for an artificial intelligence (AI)–driven and data-driven workforce.

Curricular content also varies widely across programs, further complicating the relationship between accreditation and workforce readiness. Prior work documents gaps in areas such as database management and analytics relative to employer expectations and AMIA competencies, as well as substantial heterogeneity across US biomedical and health informatics curricula [11-13]. Comparative analyses from other regions likewise show that certification alone does not guarantee consistent competency coverage, and multiple authors advocate for more applied, skills-based training aligned with workforce demands [13,14]. A recent scientometric review (2014 - 2023) further highlights the field's multidisciplinary evolution and shifting thematic emphases [15]. Yet, little is known about how curriculum design (eg, technology content density, CAHIIM, and domain coverage) interacts with pricing, modality, and accreditation, or whether programs can be grouped into configurations that integrate value and structural features.

Study Goal and Objectives

To address these gaps, an HCT-informed framework comprising Value, Structure, and Curriculum (Figure 1) is applied to evaluate US graduate health informatics programs. Using publicly available program data, the study examines how interactions among tuition, modality, and accreditation shape affordability; assesses whether curriculum design varies with technical content and accreditation; and identifies program configurations that integrate structural and value-based attributes. Findings are intended to describe the program-level investment landscape rather than to imply direct causal links to employment outcomes.

Methods

Study Design and Reporting

This cross-sectional, descriptive, comparative observational study evaluated US graduate (master's level or equivalent) health informatics programs. Reporting followed the STROBE (Strengthening the Reporting of Observational Studies in Epidemiology) guideline for cross-sectional studies [16]. Eight prespecified research questions (RQ1-RQ8) examined credit-hour requirements, tuition-normalized curriculum breadth, accelerated pathways, tuition stratification, technology content density, internship requirements, format-duration patterns, and descriptive program configurations.

Eligibility and Program Selection

The unit of analysis was the graduate program. Eligible programs were US-based master's or graduate-equivalent offerings in health, biomedical, or clinical informatics and related analytics programs. To ensure a transparent and replicable sample based solely on publicly available documentation, analytics-focused programs were deemed eligible if they met at least 1 of the following verifiable criteria:

1. *Institutional affiliation*: The degree was housed within a health-related college (eg, medicine, nursing, and public health) or an interdisciplinary technology-focused college (eg, computing, information sciences, or engineering).
2. *Explicit health designation*: The program title or official catalog description explicitly designated a "Health Informatics," "Health Analytics," or "Clinical Informatics" track, concentration, or specialization.
3. *Core curricular focus*: The publicly available course catalog listed core requirements or elective clusters dedicated to health sector applications (eg, "Healthcare Data Science" or "Clinical Decision Support").

Programs were excluded if they were duplicates, suspended, or closed during the study window; located outside the United States; or limited to undergraduate or certificate credentials.

Search Strategy and Data Sources

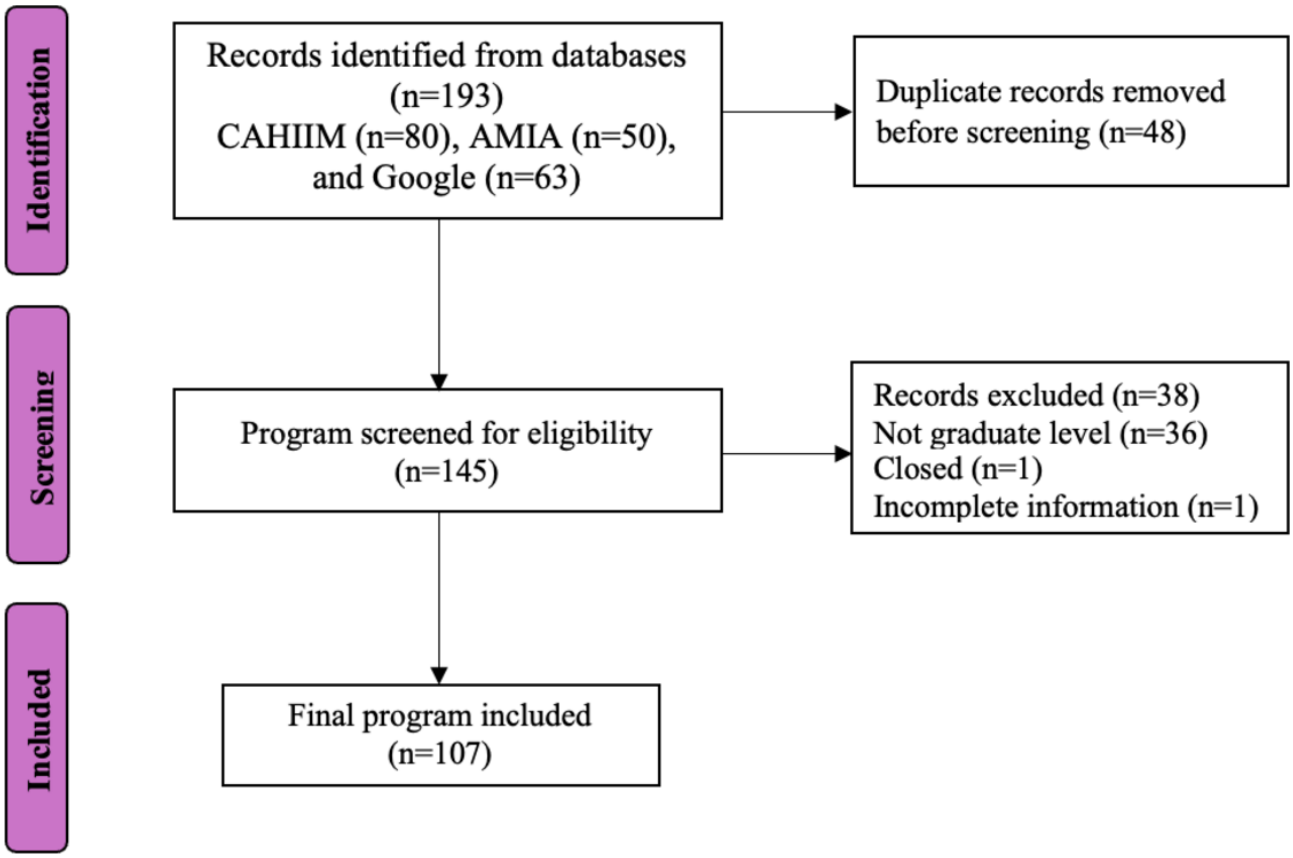
Programs were identified via 3 primary channels: CAHIIM directory, the AMIA institutional membership directory, and structured web searches finalized on May 15, 2025. To facilitate replication, web searches used the Google search engine with the following exact search strings:

1. "graduate health informatics programs US"
2. "health analytics masters US"
3. "biomedical informatics master's degrees United States"
4. "clinical informatics graduate curriculum US"

Search results were restricted to the first 100 hits per string. For each program, the most recent institutional catalog (2024 - 2025 or 2025 - 2026) and program-specific web pages were used as the primary data sources. In cases of conflicting information, the official university catalog was prioritized. The identification, screening, and inclusion process are summarized in the study flow diagram (Figure 2).

XSL•FO
RenderX

Figure 2. Flow diagram of the program identification and selection process. A total of 193 records were identified across CAHIIM (n=80), AMIA (n=50), and Google searches (n=63). Following the removal of duplicates (n=45), 145 programs were screened. Thirty-eight programs were excluded (36 not graduate level, 1 closed, and 1 with incomplete data), resulting in a final analytical sample of 107 programs. AMIA: American Medical Informatics Association; CAHIIM: Commission on Accreditation for Health Informatics and Information Management Education.



Variables Coding for Value and Structure

Overview

A total of 107 programs across 60 institutions were coded across 18 fields spanning institutional identity, program descriptors (title, delivery format, and nominal duration), academic requirements (total credits, tracks, and course classification), culminating experiences (capstone, thesis, and internship; required or optional), structural attributes (accelerated pathways and professional science master's designation), financial or support (tuition per credit [US dollars], F-1 visa eligibility), and accreditation (CAHIIM status or notes). Abbreviations are provided in Multimedia Appendix 1; definitions and the variable framework are detailed in Multimedia Appendix 2; and data sources are listed in Multimedia Appendix 3. Programs labeled "Flexible," indicating multiple delivery options, were coded accordingly.

Accreditation Coding

CAHIIM accreditation was coded at the *program and degree levels*, rather than at the institutional or departmental level. A program was classified as CAHIIM-accredited only if the specific graduate degree under analysis (eg, MS in Health Informatics or equivalent) was explicitly listed as actively accredited in the CAHIIM program directory during the study window (January-May 2025). Programs housed within departments offering other CAHIIM-accredited degrees (eg, Health Information Management) were not classified as accredited unless accreditation explicitly applied to the analyzed program. Programs in candidacy, transition, or legacy status were also excluded from the accredited category.

Technology content density was calculated as a simple count of predefined technology-related keywords found in course titles and descriptions for required and elective courses. All keywords were weighted equally and counted only once per program to prevent duplication from repeated course descriptions. Keywords were selected a priori based on commonly cited technical skill areas in graduate health informatics education (eg, analytics, databases, AI, and cybersecurity) and were subsequently mapped to CAHIIM domains for standardized classification. While catalog descriptions do not provide the granular detail of a full syllabus, they serve as a reliable proxy for identifying a program's primary curricular pillars and intended technical emphasis. Mapping was restricted to explicit mentions of technology areas to ensure high-level domain alignment without overinferring specific lesson content. Total keyword counts per program ranged from 0 to 15 and were used to compare overall levels of technical focus. We acknowledge that keyword-based density serves as a measure of curricular emphasis and intent; it is not a validated measure of actual skill acquisition or syllabus-level depth.

To better reflect differential curricular emphasis between required and elective coursework, we constructed a weighted technology content density sensitivity metric. Keywords

XSL•FO
RenderX

identified in required (core) courses were assigned a weight of 2, whereas keywords appearing only in elective courses were assigned a weight of 1. Keywords were additionally grouped a priori into structured technical domains (eg, AI or machine learning, data engineering or database systems, analytics or statistics, cybersecurity or information assurance, and clinical informatics systems) to align the metric with standardized curricular categories. The weighted density score for each program was calculated as the sum of weighted domain occurrences, counted once per domain per program to avoid duplication. Primary analyses used the unweighted density metric; weighted results were examined as sensitivity analyses to assess robustness.

Operationalization of the Value-Structure-Curriculum Framework

The framework was operationalized to evaluate the structural components of the educational investment:

1. *Value*: Defined by tuition per credit and tuition-normalized curriculum breadth (US dollars per credit per distinct CAHIIM-aligned domain). As a structural proxy, this metric assumes equal weighting across domains and is intended to compare pricing intensity relative to curricular scope rather than measuring instructional depth or student outcomes.
2. *Structure*: Comprised delivery format (online, in person, hybrid, and flexible), required credits, nominal duration, culminating experiences, accelerated pathways (eg, 3+2 and 4+1), and CAHIIM accreditation.
3. *Curriculum*: Assessed via technology-content density (eg, AI, machine learning, data science, and cybersecurity) and CAHIIM domain coverage (binary flags).

Derived Measures

Total program cost was calculated as tuition per credit × required credits. The tuition-normalized curriculum breadth metric was defined as (US dollars per credit) divided by the number of domains (eg, US $800 per credit across 10 domains → US $80 per credit per domain). The metric was intentionally specified independent of required credit hours, which were analyzed separately. Sensitivity analyses contrasted required-only vs required + elective domain coverage.

Data Cleaning and Preprocessing

Standardization steps included categorical harmonization (eg, No/None/blank → No), text normalization, tuition normalization to per credit (including mandatory program or online fees explicitly tied to credits; excluding optional fees), and reconciliation of credit ranges using midpoints. All transformations were scripted and reproducible. The code list is provided in Multimedia Appendix 2.

Interrater Reliability

Two researchers (PP and MI) independently piloted 20 programs. Agreement for key fields (format, credits, internship requirement, accreditation, and tuition tier) was assessed using Cohen κ (or Gwet's AC1 for sparse categories); discrepancies were adjudicated and coding rules updated. The remaining 87 programs were single-coded with adjudication as needed. Final κ/AC1 values are reported in Multimedia Appendix 3.

Bias, Missing Data, and Sensitivity

Information bias was mitigated through multisource verification, web archiving, and conservative tuition rules. The risk of selection bias due to limited online presence was reduced using CAHIIM and AMIA seeding and structured web searches. Missing data were coded as NA (not available). Primary analyses used complete cases with denominators reported. Sensitivity checks repeated analyses after (1) excluding programs with unclear or mixed tuition and (2) restricting the sample to programs with dated catalogs.

Statistical Analysis

Analyses were conducted in R (version 4.4.1; R Core Team) (Multimedia Appendix 4 lists packages). Given the descriptive, observational nature of the dataset, statistical tests were used to identify potential associations and characterize patterns rather than infer causal relationships. All tests were 2-sided with α=.05. To assess the practical significance, we report effect sizes, including partial eta-squared (η_p^2) for ANCOVA and ANOVA models and Cohen *d* or *r* for group comparisons, along with 95% CIs for mean differences. For related test families, Tukey or Games-Howell procedures were used to control the family-wise error rate; otherwise, Benjamini-Hochberg–adjusted *q* values are reported.

Descriptives include counts and percentages for categorical variables, as well as medians (IQR) and means (SD) for monetary values. Group comparisons for tuition and credit hours used Welch ANOVA (with nonparametric alternatives, Mann-Whitney *U* [Wilcoxon rank sum] tests, when distributions were nonnormal). Categorical comparisons used chi-square or Fisher exact tests, as appropriate. Credit-hour models used ANCOVA or ANOVA with format and accreditation as factors and tuition per credit as a covariate; assumptions were checked, with robust methods and bootstrap 95% CIs (2000 resamples) when violated. Tuition-normalized curriculum breadth was compared with Wilcoxon rank sum tests and visualized with median or density plots. Technology alignment was assessed via linear regression modeling content density as a function of tuition per credit and total credits, with accreditation and interaction terms; we report standardized coefficients (β), SEs, 95% CIs, *P* values, and R^2. Results near the *P*=.05 threshold are interpreted as descriptive patterns requiring further validation.

Clustering was conducted as an exploratory clustering analysis to uncover descriptive program configurations. Continuous variables were *z*-score standardized, categorical variables were one-hot encoded, and culminating experiences were ordinally coded. *k*-means was used as the primary method for its interpretability, with Ward hierarchical clustering applied as a robustness check to ensure partition stability. The number of clusters was selected using elbow, silhouette, and Calinski-Harabasz indices, with bootstrap resampling used to assess stability. We emphasize that clusters with very small membership ($n \leq 4$) are rare configurations and are presented solely for descriptive completeness rather than generalizable inference.

XSL•FO
RenderX

External validity was examined through differences in accreditation status and tuition- and credit-based metrics. Results are interpreted as descriptive configurations rather than definitive segments. Although Euclidean *k*-means is not optimal for nominal data, alternative mixed-type methods (eg, Gower distance with *k*-medoids or *k*-prototypes) were considered; *k*-means was selected to prioritize comparability and interpretability in this education policy context, with multimethod checks used to mitigate bias.

Ethical Considerations

All data analyzed in this study were obtained from publicly available institutional and program websites. In limited cases where online information was ambiguous, program offices were contacted solely to confirm publicly applicable program details (eg, tuition rates or required credit hours). No confidential, unpublished, proprietary, or student-level information was requested, accessed, or used. This study was reviewed by the Grand Valley State University Institutional Review Board (IRB) and determined to be not human subjects research and therefore not requiring IRB approval, in accordance with federal regulations (45 CFR 46.102) (IRB determination letter, April 2, 2026).

Results

Descriptive Overview of Programs

A total of 107 graduate health informatics programs from 60 US institutions were identified and analyzed. Multimedia Appendix 5 summarizes their structural, curricular, and financial characteristics. Health Informatics (33/107, 30.8%) was the most common program designation, followed by Biomedical Informatics (18/107, 16.8%). Program delivery formats differed, with online programs most common (37/107, 34.6%), followed by hybrid (32/107, 29.9%), in person (23/107, 21.5%), and flexible formats (15/107, 14.0%), where flexible denotes programs that provide more than one delivery option.

The typical program duration was 19 - 24 months (72/107, 67.3%), with credit requirements most frequently falling between 31 and 39 hours (53/107, 49.5%). Roughly one-third of programs (35/107, 32.7%) required prerequisites. Culminating experiences included capstone projects (54/107, 50.5%), internships (21/107, 19.6%), and theses (7/107, 6.5%). Tuition per credit showed substantial variation: 39.3% of programs charged US $500 to US $999, while 13.1% exceeded US $2000. CAHIIM accreditation, when defined at the program-specific degree level, was present in 33 of 107 programs (30.8%).

Policy and Cost Dynamics

RQ1: Tuition Costs, Program Format, and Accreditation

We tested whether tuition per credit hour, program format, and CAHIIM accreditation status were associated with graduate program credit hour requirements. Descriptive statistics stratified by format and accreditation are shown in Tables 1 and 2.

Table . Tuition and credit hours by program format and accreditation status (N=107)[a].

Program format	CAHIIM[b] accredited	Values, n (%)	Mean tuition, US $	SD tuition, US $	Mean credits	SD credits
Online	Yes	20 (18.7)	699.16	286.62	33.43	8.76
	No	17 (15.9)	368.90	323.68	24.15	9.89
In person	Yes	2 (1.9)	270.76	306.47	33.50	11.91
	No	21 (19.6)	536.14	785.55	34.67	12.65
Hybrid	Yes	7 (6.5)	392.53	313.39	36.36	5.44
	No	25 (23.4)	470.13	469.03	33.89	7.82
Flexible[c]	Yes	4 (3.7)	375.00	335.21	33.68	11.96
	No	11 (10.3)	404.37	389.13	34.20	12.48

[a] Values rounded to 2 decimals. Percentages calculated from total N=107.

[b]CAHIIM: Commission on Accreditation for Health Informatics and Information Management Education.

[c]Flexible denotes programs providing more than 1 delivery option.

XSL•FO
RenderX

Table . Effect size and practical significance of program format on credit requirements.

Comparison	Adjusted mean difference	95% CI	Effect size (η_p^2[a] or d[b])
Main effect: format	—[c]	—	η_p^2=0.076 (moderate)
Maximum vs minimum comparison[d]	9.7 credits	2.1 to 17.3	d=0.82 (large)
Accreditation effect	0.15 credits	−3.2 to 3.5	η_p^2<0.001 (negligible)

[a]Partial eta-squared.

[b]Cohen *d*.

[c]Not available.

[d]Comparison between Hybrid and Online nonaccredited programs.

Online accredited programs comprised the largest subgroup (20/107, 18.7%), with a mean tuition of US $699.16 (SD US $286.62) per credit and an average of 33.43 (SD 8.76) credit hours. Online nonaccredited programs (17/107, 15.9%) had lower mean tuition (US $368.90, SD US $323.68) and the fewest required credit hours (mean 24.15, SD 9.89 hours). In person accredited programs were rare (2/107, 1.9%), with a mean tuition of US $270.76 (SD US $306.47) and an average of 33.5 credits (SD 11.91).

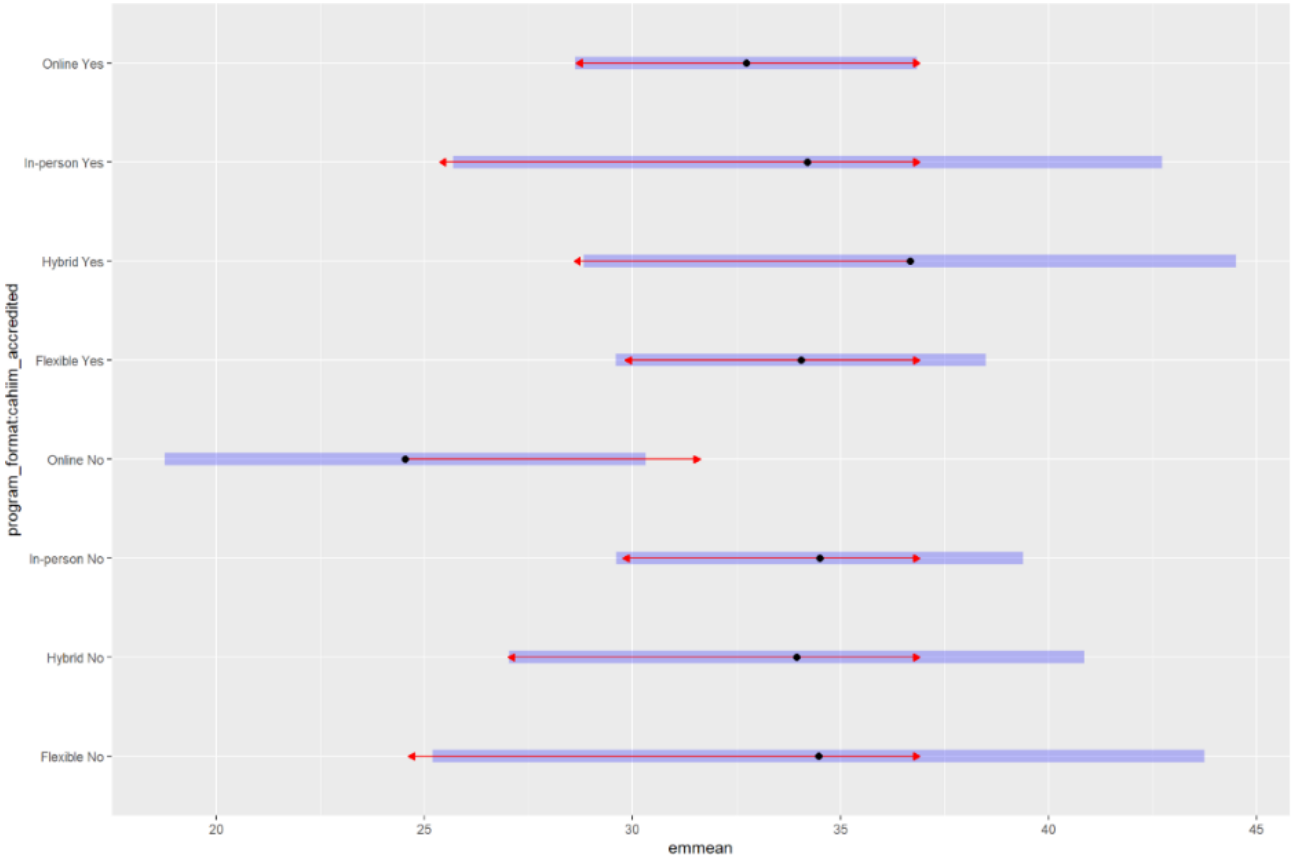
A two-way ANCOVA with tuition as a covariate suggested that program format was associated with required credit hours ($F_{3,98}$=2.70; P=.049; η_p^2=.076). The adjusted mean difference between the format with the highest requirement (Hybrid) and the lowest (Online nonaccredited) was 9.7 credit hours (95% CI 2.1 - 17.3), representing a moderate practical difference of approximately 3 standard courses. Consistent with our exploratory approach, this result is interpreted as a descriptive pattern rather than a definitive predictor.

By contrast, accreditation status showed a negligible effect on credit requirements (P=.94; η_p^2<.001), with a mean difference of only 0.15 credits (95% CI −3.2 to 3.5). Specifically, neither tuition per credit (P=.17), accreditation status (P=.94), nor the format× accreditation interaction (P=.43) reached statistical significance (Multimedia Appendix 6).

Post hoc Tukey pairwise comparisons in Multimedia Appendices 7 and 8 suggested that online nonaccredited programs had lower credit requirements than in person nonaccredited (Δ=10.52; P=.17) and hybrid nonaccredited programs (Δ=9.74; P=.44), but none reached significance after adjustment. Estimated marginal means and 95% CIs are shown in Figure 3, which highlights clustering of in person, hybrid, and flexible programs around 33 - 36 credits, while online nonaccredited programs averaged fewer credits.

XSL•FO
RenderX

Figure 3. Estimated marginal means of credit hours by program format and accreditation status. Visualization of adjusted credit-hour requirements based on a two-way ANCOVA. Estimated marginal means and 95% CIs are provided for graduate health informatics programs stratified by delivery format and Commission on Accreditation for Health Informatics and Information Management Education accreditation. Descriptive patterns indicate that online nonaccredited programs generally require fewer credits, whereas other modalities cluster between 33 and 36 credit hours. Points represent estimated marginal means, blue horizontal bars indicate 95% CIs, and red arrows denote pairwise comparisons between group means.



RQ2: Tuition-Normalized Curriculum Breadth (US Dollars per Credit per CAHIIM Domain)

We calculated a tuition-normalized curriculum breadth metric as tuition per credit divided by the number of distinct CAHIIM-aligned domains.

Because distributions were nonnormal for both accredited and nonaccredited groups (Shapiro-Wilk: accredited W=0.72, P<.001; nonaccredited W=0.60, P<.001), group comparisons were performed using the Wilcoxon rank sum test.

Across 107 programs with complete data, CAHIIM-accredited programs (n=33) had a mean tuition-normalized curriculum breadth score of 77.22 (median 48.80, SD 96.48), compared with 94.71 (median 40.50, SD 166.57) among nonaccredited programs (n=74). No statistically significant difference was observed between accredited and nonaccredited programs for this tuition-normalized curriculum breadth metric (W=1247, P=.35) (Table 3). As shown in Figure 4, accredited programs clustered more tightly around the median, whereas nonaccredited programs exhibited greater variability, including several outliers, but with comparable central tendencies. This indicates that, within this sample, accreditation status was not associated with differences in this structural pricing-to-breadth proxy.

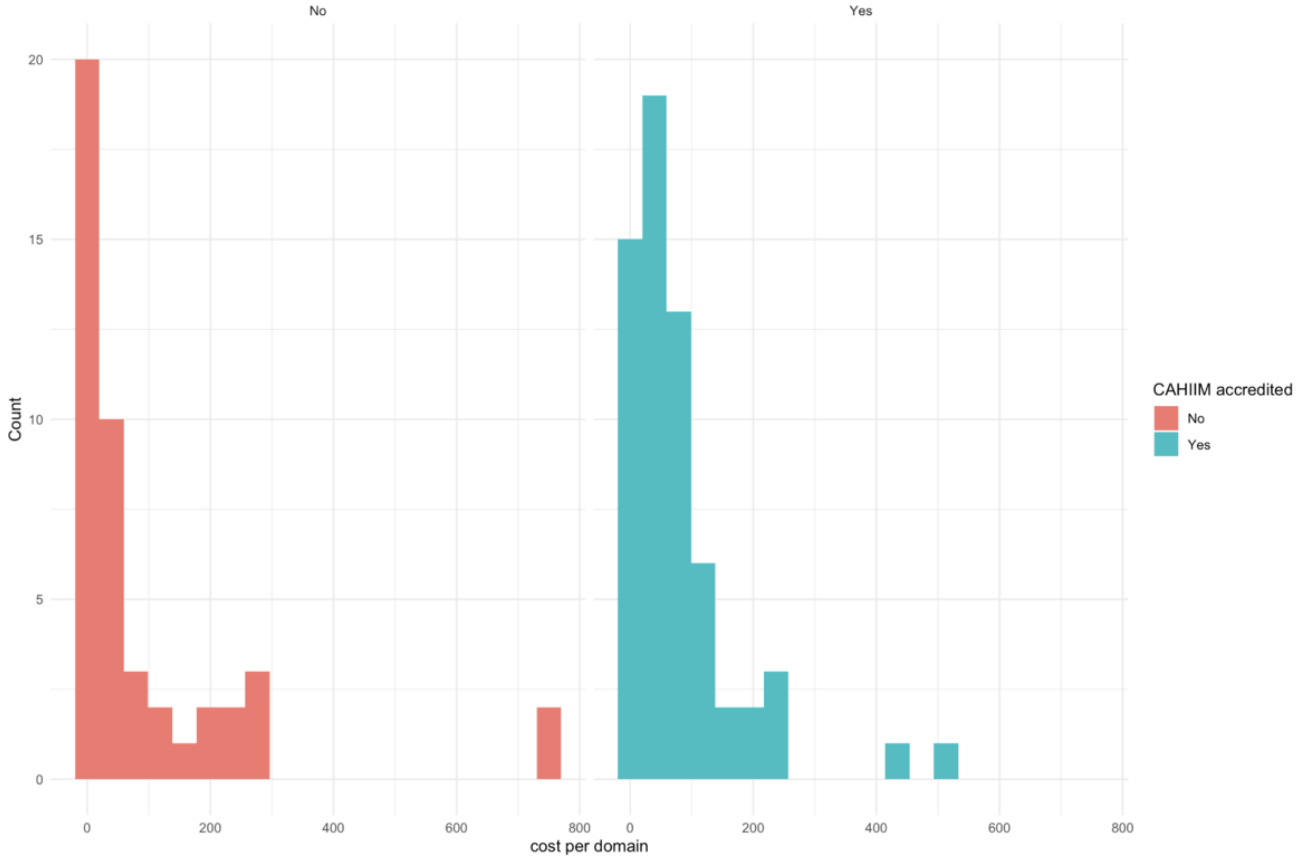
Table . Tuition-normalized curriculum breadth (US dollars per credit per Commission on Accreditation for Health Informatics and Information Management Education domain) by accreditation status[a].

Accreditation status	Values, n	Mean	Median	SD
Yes	33	77.22	48.8	96.48
No	74	94.71	40.5	166.57

[a]Wilcoxon rank sum test: W=1247; P=.35. Normality: Shapiro-Wilk (accredited: W=0.72, P<.001; nonaccredited: W=0.60, P<.001).

XSL•FO
RenderX

Figure 4. Distribution of tuition-normalized curriculum breadth (US dollars per credit per domain) by accreditation status. Histograms show the distribution of tuition-normalized curriculum breadth metric across accredited (blue) and nonaccredited (red) programs. The metric represents the ratio of tuition per credit to the number of CAHIIM-aligned domains. Accredited programs demonstrate tighter clustering, while nonaccredited programs show higher variability; however, group medians were not statistically different (*P*=.35). CAHIIM: Commission on Accreditation for Health Informatics and Information Management Education.



RQ3: Tuition Costs and Accelerated Pathways by Accreditation

We next examined whether accelerated pathways were associated with lower tuition per credit and whether this relationship differed by accreditation status. As shown in Multimedia Appendix 9, accelerated programs tended to exhibit lower tuition distributions compared with nonaccelerated programs. Stratification by CAHIIM accreditation revealed a more nuanced interaction (Multimedia Appendix 10): while accredited programs overall had higher tuition values, accredited accelerated programs showed reduced tuition compared with their nonaccelerated counterparts.

Levene's test indicated significant heterogeneity of variance across the 4 groups ($F_{3,103}$=5.47, *P*=.002). A two-way ANOVA confirmed significant main effects of acceleration ($F_{1,103}$=17.88, *P*=.001) and accreditation ($F_{1,103}$=10.92, *P*=.002), as well as their interaction ($F_{1,103}$=7.25, *P*=.005). These findings suggest that the cost benefits of acceleration are moderated by accreditation status.

Curriculum and Cost Alignment

RQ4: Tuition Stratification by Credit Load, Format, and Accreditation

We next examined whether tuition per credit varied by program format (Flexible, Hybrid, In-person, and Online) and CAHIIM accreditation status. A two-way ANOVA including all programs with complete data (n=103) did not detect statistically significant effects: program format ($F_{3,99}$=1.33, *P*=.27), accreditation ($F_{1,99}$=0.39, *P*=.54), and the format × accreditation interaction ($F_{3,99}$=2.05, *P*=.111). Tukey-adjusted pairwise comparisons were also nonsignificant (Multimedia Appendices 7 and 11).

Descriptive statistics are reported in Table 4. Among accredited programs, mean tuition per credit was highest for online programs (20/107, 18.7%; mean US $699.16, SD US $286.62) and lowest for in-person programs (2/107, 1.9%; mean US $270.76, SD US $306.47). Nonaccredited programs showed the opposite pattern, with in-person programs (21/107, 19.6%) averaging the highest tuition (mean US $536.14, SD US $785.55) and online programs (17/107, 15.9%) the lowest (mean US $368.90, SD US $323.68).

XSL•FO
RenderX

Table . Tuition per credit by program format and accreditation (N=107)[a].

Program format	Accreditation	Mean tuition (SD), US $	Values, n (%)
Flexible	Yes	374.99 (335.21)	4 (3.4)
	No	404.37 (389.13)	11 (10.3)
Hybrid	Yes	392.53 (313.39)	7 (6.5)
	No	470.13 (469.03)	25 (23.4)
In-person	Yes	270.76 (306.47)	2 (1.9)
	No	536.14 (785.55)	21 (19.6)
Online	Yes	699.16 (286.62)	20 (18.7)
	No	368.90 (323.68)	17 (15.9)

[a]Values for n (%) derived from the total sample distribution.

The interaction pattern is illustrated in Multimedia Appendix 12, which plots mean tuition per credit by program format and accreditation. Accredited online programs stand out as costlier, whereas nonaccredited in-person programs are relatively more expensive.

RQ5: Tuition Costs and Credit Loads in Relation to Technology Content Density and Accreditation

Pearson correlation tests indicated that tuition per credit was not significantly correlated with technology content density (r=−0.08, P=.39; 95% CI −0.27 to 0.11). By contrast, technology content density—representing the explicit mention of technical keywords in course descriptions—showed a small positive correlation with average credit requirements (r=0.21, P=.03; 95% CI 0.03-0.39).

To examine whether accreditation moderated these relationships, regression models were estimated with interaction terms. In the tuition model (R^2=0.03, P=.39), neither technology density (β=–10.25, P=.56) nor accreditation (β=–120.59, P=.43) significantly predicted tuition, and the interaction effect was nonsignificant (β=−1.24, P=.96).

In the credit load model (R^2=.09, P=.03), technology density positively predicted higher credit requirements (β=1.16, P=.005). Accreditation status was not significantly associated with higher credit averages (β=7.77, P=.39). The interaction of technology density × accreditation was not statistically significant (β=−.88, P=.11), suggesting that the relationship between technology density and credit load was similar across accredited and nonaccredited programs. We note that these findings reflect curricular emphasis in public documentation rather than a validated measure of student skill acquisition. Full regression outputs appear in Multimedia Appendix 13. Sensitivity analyses using the weighted technology density metric prioritizing required over elective coursework produced similar associations with credit requirements and no relationship with tuition per credit, indicating robustness of the primary findings.

RQ6: Internship Requirement and Program Costs

To assess whether mandatory internships are associated with differences in tuition or credit requirements, we compared programs with and without an internship requirement. Among the 107 programs analyzed, 21 of 107 (19.6%) required internships, while 86 of 107 (80.4%) did not.

A Shapiro-Wilk test confirmed nonnormality of tuition data (W=0.81, P<.001), so Mann-Whitney U tests were applied. There was no significant difference in tuition between internship-required and noninternship programs (W=796, P=.40). However, programs with mandatory internships required significantly more credit hours (mean 39.0, SD 8.1) than those without (mean 31.3, SD 10.7; W=549; P=.005), with a small to moderate effect size (r=0.27). These differences are visualized in Multimedia Appendix 14, which shows a tighter clustering of higher credit requirements among internship-required programs compared with noninternship programs.

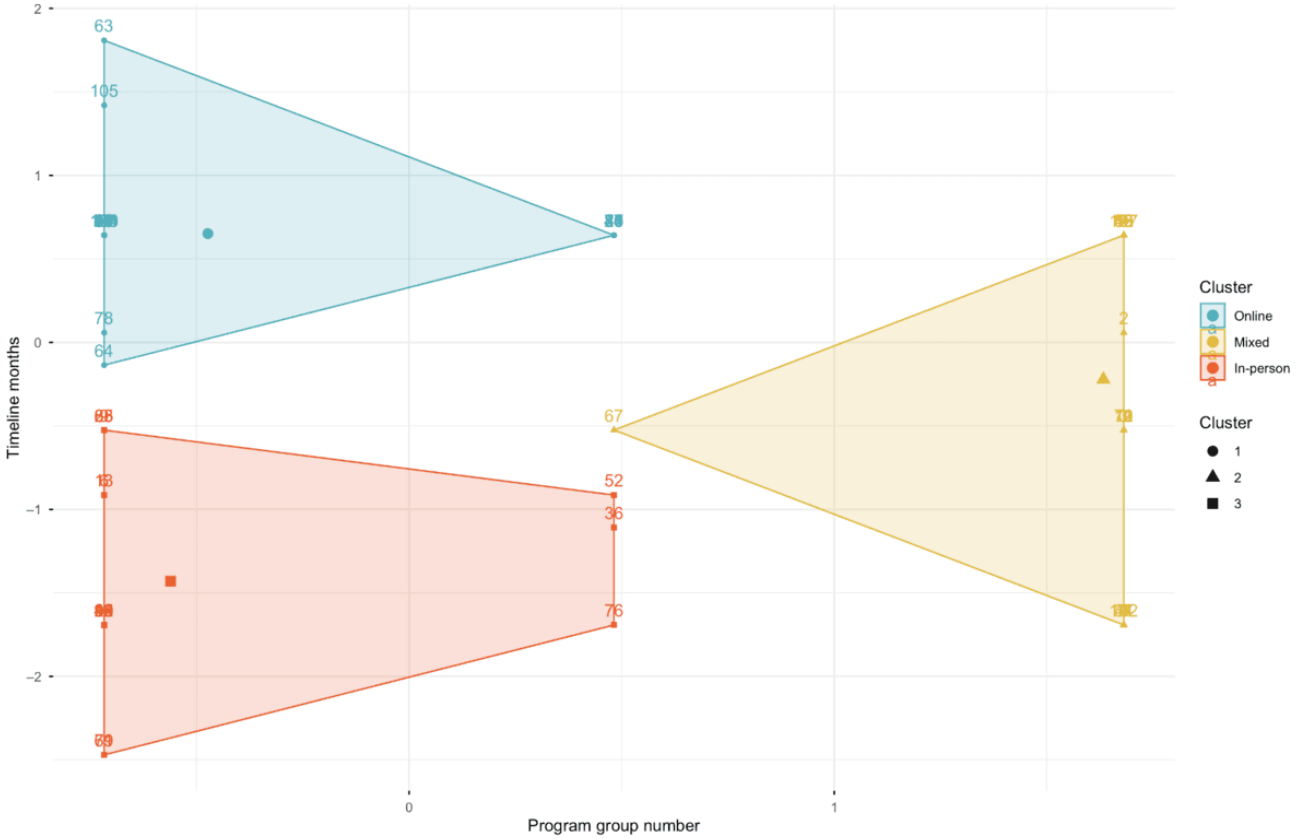
Program Structures and Configurations

RQ7: Structure Patterns—Program Format and Duration

Across formats (Flexible, Hybrid, In-person, and Online), the most common duration was 24 months (Multimedia Appendix 8). A format × duration crosstab (months coded at 8, 12, 15 - 16, 18, 20 - 21, 24, and 28 - 30) indicated no significant association between format and duration (χ^2_{27}=24.76, P=.59). Full results appear in Multimedia Appendix 15; patterns are illustrated in Multimedia Appendices 5 and 16. To examine continuous variation, we conducted a k-means cluster analysis on standardized program format and timeline (months). The elbow method supported a 3-cluster solution, which produced 3 well-separated clusters (Figure 5).

XSL•FO
RenderX

Figure 5. *K*-means clustering of program format and timeline (3-cluster solution). Scatterplot identifying 3 descriptive delivery duration configurations: Cluster 1 (blue, Online) represents shorter programs with modest variance, cluster 2 (yellow, Mixed: Hybrid/Flexible) includes standardized programs centered on a 24-month duration, and cluster 3 (red, In-person) reflects intermediate-length programs with high variability. Shaded regions represent cluster boundaries; note that small subclusters (n≤4) are rare configurations within this sample.



1. Cluster 1 Online (37/107, 34.6%): this cluster represents all online programs (20 accredited and 17 nonaccredited), typically characterized by shorter durations (mean 13.3, SD 2.95 months).
2. Cluster 2 Mixed (Hybrid/Flexible) (47/107, 43.9%): comprising 32 hybrid and 15 flexible programs, this group featured the most standardized timelines centered on approximately 24 months (mean 24.1, SD 1.15 months).
3. Cluster 3 In-person (23/107, 21.5%): consisting of 2 accredited and 21 nonaccredited programs, these programs showed intermediate length but higher variability (mean 19.6, SD 4.98 months).

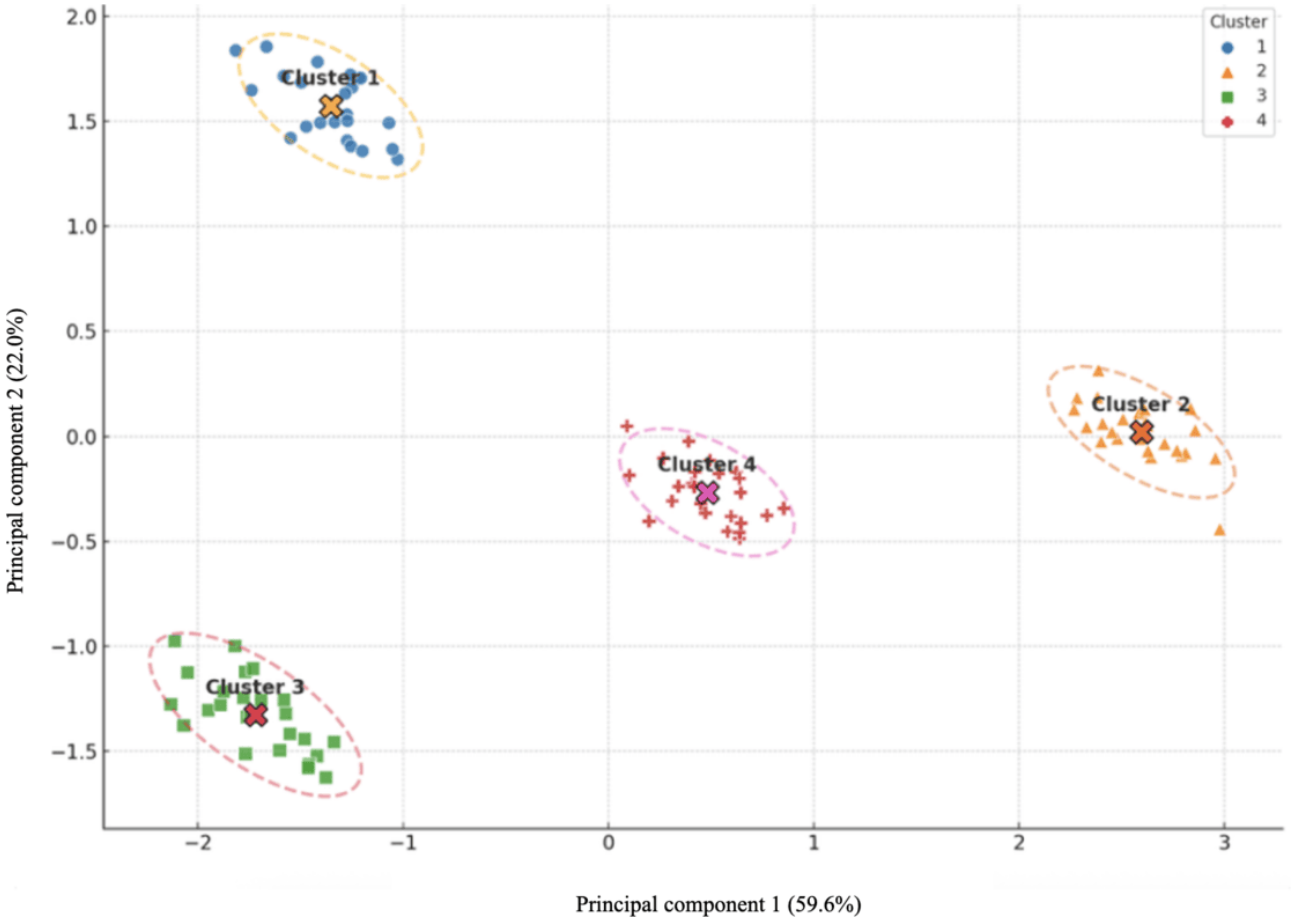
Taken together, the chi-square analysis suggests no categorical dependency, whereas the clustering reveals 3 coherent delivery-duration configurations: (1) shorter Online programs, (2) standardized 2-year Mixed programs, and (3) more variable In-person programs. Clusters with very small membership (n≤4), such as the accredited in-person subgroup, are reported as rare configurations and were not used for inferential group comparisons; their role remains descriptive and hypothesis-generating.

RQ8: Distinct Program Configurations and Accreditation Differences

Four clusters were identified using *k*-means clustering based on tuition per credit, credit requirements, program format, culminating experience, and track offerings. Clustering results are presented as descriptive patterning rather than definitive classification. The resulting clusters are visualized in Figure 6 and summarized in Multimedia Appendices 17 and 18. Points represent individual programs; ellipses show 95% confidence regions, with cluster centroids marked by crosses.

XSL•FO
RenderX

Figure 6. Principal component analysis (PCA) visualization of the *k*-means cluster solution (N=107). Four descriptive program configurations derived from *k*-means analysis, visualized in a 2D PCA space (PC1 explaining 59.6% and PC2 explaining 22.0% of variance). Ellipses denote 95% confidence regions for cluster 2 (flexible and low-cost) and cluster 4 (online and moderate-cost). Clusters 1 and 3 represent rare program configurations (n=4 and n=3, respectively) and are presented for descriptive completeness; centroids are marked by crosses.



While 4 clusters were derived, cluster 1 (4/107, 3.7%) and cluster 3 (3/107, 2.8%) represent rare configurations within the sample and are presented for descriptive completeness rather than as generalizable program configurations.

1. Cluster 1 (4/107, 3.7%): This group consisted of hybrid programs with the highest average credit requirements (mean 43.2) and the highest tuition per credit (mean US $668). These programs most frequently offer both capstone and thesis options as culminating experiences. CAHIIM accreditation was present in 1/4 (25%) of programs.
2. Cluster 2 (46/107, 43.0%) primarily included flexible-format programs with the lowest average tuition (mean US $424 per credit) and the fewest credit hours (mean 29.0). Capstones and tracks were uncommon, and 10 of 46 (21.7%) of programs were accredited.
3. Cluster 3 (3/107, 2.8%): This cluster comprised in-person programs with an average of 37.3 credit hours and the highest tuition (mean US $748 per credit). These programs typically required a thesis and included AI-related tracks; accreditation was present in 1 of 3 (33.3%) cases.
4. Cluster 4 (54/107, 50.5%): The largest group, predominantly online programs, featured moderate tuition (mean US $514 per credit) and average credit loads (mean 35.0). Capstones were the dominant culminating experience, while tracks were rare, and this cluster contained the largest number of accredited programs (21/54, 38.9%).

Discussion

Principal Results

This cross-sectional analysis identified substantial variation in cost, structure, and curriculum across US graduate health informatics programs. Notably, accreditation was not associated with statistically significant differences in the tuition-normalized curriculum breadth structural proxy, credit requirements, or tuition per credit. We emphasize that tuition-normalized curriculum breadth is a structural proxy and should not be interpreted as a comprehensive assessment of program “value” or “quality,” as it does not incorporate instructional depth, total program cost, or postgraduation labor market outcomes.

Delivery format and internship requirements were more strongly associated with financial differences. Delivery format showed modest variation in credit requirements (95% CI 2.1 - 17.3). This suggests that structural modality, rather than accreditation signaling, appears more closely associated with variation in the required educational investment.

Specifically, online nonaccredited programs required the fewest credits (17/107, 15.9%; mean 24.15, SD 9.89), while accredited online programs exhibited the highest tuition per credit. Furthermore, programs requiring internships were associated with significantly higher credit loads (mean 39.0, SD 8.1) than those without (mean 31.3, SD 10.7; *P*=.005). The lower credit

XSL•FO
RenderX

requirements observed in the online nonaccredited sector may reflect structural program design differences such as accelerated residency requirements or stackable certificate-to-master's pathways, where a 24-credit requirement represents the residency core completed after the transfer of prior graduate credits. Exploratory cluster analysis identified 4 program configurations, although we note that some configurations were rare (n≤4), suggesting that these configurations serve as descriptive representations of the current landscape rather than generalizable program models.

Comparison With Prior Work

This study applies HCT to evaluate cost efficiency and curricular design in graduate health informatics. By analyzing tuition, accreditation, and curriculum across 107 programs, it assesses how institutions structure educational investments and whether these investments yield value in technical skill acquisition and affordability. This study provides a descriptive cross-sectional evaluation of US graduate health informatics programs using a multidimensional framework that integrates Value, Structure, and Curriculum.

The findings extend discussions on affordability, accreditation, and content alignment in health professions education and highlight persistent variation in how institutions operationalize value and quality. Tuition per credit varied widely, from below US $500 to above US $2000, with online nonaccredited programs showing the lowest credit requirements and accredited online programs the highest tuition per credit. This is consistent with reports that online programs can incur hidden or differential fees [5].

Accreditation status was not associated with differences in the tuition-normalized curriculum breadth structural proxy, aligning with literature suggesting that accreditation may support quality assurance without necessarily differentiating program cost efficiency. Accelerated pathways, such as combined bachelor's-masters programs, were associated with lower tuition intensity**,** consistent with the premise that shorter time to completion can reduce the structural cost of skill acquisition [17].

Implications for Education and Policy

Overview

Program structure varied widely across offerings, with capstone projects more common than theses, and internships were required in fewer than 20% of programs (21/107, 19.6%). Programs that required internships had higher average credit loads, suggesting increased curricular burden without corresponding tuition reduction. This pattern echoes prior research showing that unpaid placements can impose financial strain through lost income and indirect costs [18,19]; however, it is important to note that our data did not include student demographics, the paid or unpaid status of these placements, or the actual financial burden on individuals. Consequently, we offer the conjectural implication that internship experiences, while potentially valuable for applied learning, may inadvertently reinforce inequities among students with fewer financial resources or caregiving responsibilities.

The absence of a relationship between CAHIIM accreditation and either tuition-normalized curriculum breadth (P=.35; RQ2) or required credit hours (P=.39; RQ1) provides context regarding how accreditation aligns with structural program design. Although accreditation is frequently interpreted as an indicator of streamlined program design or stronger financial returns for students, these findings indicate that CAHIIM accreditation may function primarily as a signal of curricular credibility and baseline quality expectations [20] rather than as a structural determinant of program cost or duration. From the student viewpoint, CAHIIM accreditation was not associated with lower tuition-normalized curriculum breadth or required credit hours in this sample, within the constraints of this program-level descriptive analysis, suggesting that accreditation status alone does not differentiate this structural pricing-to-breadth proxy. While our study lacks outcome-level data to measure actual economic returns, this observed gap highlights the importance of more explicit communication regarding the scope and limits of what CAHIIM accreditation represents, particularly as concerns about affordability, debt, and workforce alignment increasingly inform graduate education choices.

Contrary to expectation, higher tuition was not associated with greater coverage of technology-focused content. Instead, programs with higher technology density tended to require more credits (r=0.21, P=.03), suggesting the need for broader curricular space to integrate competencies such as AI, data science, and cybersecurity. This supports calls for more robust technical training in health informatics curricula [9,14] and aligns with updated international recommendations emphasizing core informatics, data, and computer science competencies [9]. These findings indicate that tuition levels alone are not reliable indicators of curricular sophistication or alignment with emerging workforce needs.

AI Integration and Program Configurations

Beyond cost and accreditation, these findings suggest that graduate health informatics programs may consider integrating applied AI tools into their curricula. Systems for automated imaging, diagnostic support, and patient simulation tools, such as AI-driven diagnostics platforms like Smile.AI or automated smile analysis and orthodontic planning software like Dynasmile (developed by Chen et al [21,22]) and patient engagement platforms, illustrate emerging applications of AI for efficiency, assessment standardization, and patient engagement in clinical contexts. Embedding similar tools into curricula via case-based exercises and simulated workflows may provide opportunities for students to develop familiarity with evaluating, implementing, and managing AI technologies, including ethical and governance considerations. As upskilling remains a critical barrier to AI adoption in health information sectors [23], this hands-on exposure to applied AI tools could complement accreditation frameworks by emphasizing practical implementation alongside domain knowledge, without implying direct effects on workforce outcomes.

The 4 program configurations identified through cluster analysis (RQ8) reveal how graduate health informatics programs differ in how they strategically balance cost, curricular depth, and accreditation signaling. We posit as a conjectural interpretation

XSL•FO
RenderX

that these configurations may correspond to distinct workforce pathways, although we emphasize that no student-level or employment outcome data were available to validate these associations.

For example, cluster 1 (high credit, high cost hybrid programs with capstone and thesis options; n=4) and cluster 3 (high tuition, in-person programs with thesis and advanced technical tracks; n=3) appear oriented toward research-intensive and technically advanced roles. Cluster 2 (low-cost, low-credit flexible programs; n=46) emphasizes rapid workforce entry and applied practice, while cluster 4 (moderately priced, predominantly online programs with capstone requirements; n=54) targets implementation-focused careers and professional advancement.

From a workforce perspective, program choice may reflect different career trajectories, yet these linkages remain theoretical associations rather than empirically measured results. Students may be selecting among differing program configurations, with each configuration potentially reflecting distinct role orientations within the health informatics labor market. Understanding these structural patterns may help describe how program design relates to targeted skill emphasis, although further research is required to link these configurations to actual labor market outcomes.

For prospective students, these configurations offer points of comparison regarding program choice. For institutions, they highlight considerations for evaluating credit requirements, aligning costs with curricular scope, and supporting equity in applied training opportunities. The tuition-normalized curriculum breadth metric may also serve as a practical descriptive indicator for internal dashboards, program review, and accreditation self-studies, particularly in the context of new federal requirements for program-level transparency under the Financial Value Transparency and Gainful Employment regulations [2].

Future Research and Recommendations

The recommendations are as follows:

1. *Link program features to outcomes*: Pair program structure, tuition, and curriculum with graduate employment, earnings, and satisfaction data to test return on investment using sources such as College Scorecard and alumni surveys.
2. *Quantify hidden and modality-specific fees*: Prospectively capture mandatory online and program fees to refine tuition normalization and assess the true cost of online, hybrid, flexible, and in person formats.
3. *Evaluate equity impacts*: Examine whether internship requirements (21/107, 19.6%), accelerated pathways, and tuition levels differentially affect students by income, race and ethnicity, caregiving status, visa status, and employment status.
4. *Validate the tuition-normalized breadth metric*: Compare this metric with alternative measures that incorporate required credits, total program cost, and graduate outcomes. Test reliability across institutions and years, compare required-only with required plus elective domain coverage, and correlate the metric with external indicators of quality and outcomes.
5. *Deepen curriculum analytics*: Move beyond keyword counts to syllabus-level and assignment-level analyses that capture depth of AI, data science, and cybersecurity training and alignment with AMIA and International Medical Informatics Association competencies.
6. *Incorporate longitudinal designs*: Track cohorts over time to evaluate how changes in accreditation status, delivery format, or credit requirements influence affordability and outcomes.
7. *Conduct mixed methods studies*: Use interviews and focus groups with students, faculty, and program directors to contextualize quantitative patterns and identify barriers to transparency and affordability.
8. *Compare internationally*: Replicate the framework across countries to determine whether observed configurations are specific to the US tuition and regulatory environment or generalize to other contexts.
9. *Develop decision tools*: Translate findings into dashboards or checklists that help applicants compare programs on cost, structure, and curriculum, and that support institutional benchmarking and self-study.

Limitations

This analysis relied on publicly reported data, which may omit hidden fees, special tuition categories, or real-world variation in student costs. The study focused on structural and curricular characteristics rather than outcomes such as employment, earnings, or satisfaction. Future work should triangulate program-level data with alumni surveys, employer input, and federal datasets to assess real-world outcomes, and consider cross-national comparisons to test generalizability.

Furthermore, we acknowledge that applying complex modeling (eg, ANCOVA and *k*-means clustering) to a relatively small, heterogeneous dataset of 107 programs carries risks of overfitting and spurious interpretation. Specifically, our clustering approach is based on mixed-data encodings analyzed using Euclidean distance, which may be less suitable for nominal variables. Although robustness checks using Ward's hierarchical clustering, multiple internal validity indices, and bootstrap resampling supported the overall structure, the results should be interpreted as descriptive rather than definitive. The inclusion of very small clusters (n≤4) indicates that these specific program configurations are rare within the current landscape; they are presented for descriptive completeness but lack the statistical power to be generalizable or to support inferential claims.

Conclusions

Graduate health informatics programs in the United States exhibit wide variation in cost, structure, and curriculum. Accreditation status was not associated with differences in the tuition-normalized curriculum breadth structural proxy in this program-level analysis, although accreditation was more prevalent within certain predominantly online program configurations. Accelerated formats and greater technical content emphasis (eg, AI and data science) were descriptively associated with lower credit requirements or broader curricular scope, suggesting potential pathways to align affordability with workforce-oriented training. Within an HCT lens, the structural cost of acquiring skills varies substantially and is not

XSL•FO
RenderX

consistently aligned with common program design features. Program and policy strategies that more transparently align cost, accreditation signaling, and curricular emphasis may support efforts to promote equitable access to high-quality health informatics education.

Acknowledgments

The authors used generative artificial intelligence tool (ie, ChatGPT) to assist with language editing and text condensation. All scientific content, analyses, and interpretations were developed and verified by the authors, who take full responsibility for the accuracy, originality, and integrity of the manuscript.

Funding

The authors received no specific grant or financial support from any funding agency in the public, commercial, or not-for-profit sectors for the conduct of this research.

Data Availability

All data analyzed in this study were obtained from publicly accessible university and program websites. A complete log of search dates (finalized May 15, 2025), the categories of institutional directories queried (eg, Commission on Accreditation for Health Informatics and Information Management Education, American Medical Informatics Association, and Integrated Postsecondary Education Data System), and the detailed decision matrix for analytics-program inclusion are provided in Multimedia Appendix 4 to ensure study replicability. To protect the integrity of the data snapshot and institutional comparisons, the compiled dataset and coding framework used to generate the analyses are available from the corresponding author upon reasonable request.

Authors' Contributions

Conceptualization: SS (lead)
Methodology: SS (lead), DMD (supporting), PP (supporting), MI (supporting)
Formal analysis: SS (lead), DMD (lead), SS (supporting), DMD (supporting)
Data curation: PP (lead), MI (lead), SS (supporting), DMD (supporting) Investigation: PP (lead), MI (lead), SS (supporting), DMD (supporting)
Validation: SS (lead), DMD (lead)
Project administration: SS (lead)
Supervision: SS (lead)
Visualization: SS (lead), DMD (lead), MI (supporting)
Writing – original draft: SS (lead)
Writing – review & editing: SS (lead), DMD (lead), PP (supporting), MI (supporting)

Conflicts of Interest

None declared.

Multimedia Appendix 1
Abbreviations.
[DOCX File, 17 KB - mededu_v12i1e87479_app1.docx]

Multimedia Appendix 2
Operational definitions.
[DOCX File, 19 KB - mededu_v12i1e87479_app2.docx]

Multimedia Appendix 3
Variable coding framework and data sources.
[DOCX File, 22 KB - mededu_v12i1e87479_app3.docx]

Multimedia Appendix 4
R (R Core Team) packages used in the study.
[DOCX File, 18 KB - mededu_v12i1e87479_app4.docx]

Multimedia Appendix 5
Descriptive characteristics of graduate health informatics programs (N=107).

XSL•FO
RenderX

[DOCX File, 21 KB - mededu_v12i1e87479_app5.docx]

Multimedia Appendix 6
ANCOVA results for tuition, program format, and accreditation predicting credit hour requirements.
[DOCX File, 17 KB - mededu_v12i1e87479_app6.docx]

Multimedia Appendix 7
Tukey post hoc pairwise contrasts for program format × accreditation.
[DOCX File, 18 KB - mededu_v12i1e87479_app7.docx]

Multimedia Appendix 8
Pairwise comparisons of tuition per credit by program format and accreditation status.
[DOCX File, 17 KB - mededu_v12i1e87479_app8.docx]

Multimedia Appendix 9
Tuition per credit by accelerated program status.
[PNG File, 22 KB - mededu_v12i1e87479_app9.png]

Multimedia Appendix 10
Tuition per credit by accelerated program status, stratified by Commission on Accreditation for Health Informatics and Information Management Education accreditation.
[PNG File, 30 KB - mededu_v12i1e87479_app10.png]

Multimedia Appendix 11
Pairwise comparisons of tuition per credit by program format.
[DOCX File, 17 KB - mededu_v12i1e87479_app11.docx]

Multimedia Appendix 12
Mean tuition per credit by program format and Commission on Accreditation for Health Informatics and Information Management Education accreditation.
[PNG File, 48 KB - mededu_v12i1e87479_app12.png]

Multimedia Appendix 13
Regression models testing technology content density and accreditation as predictors of tuition and credit load.
[DOCX File, 18 KB - mededu_v12i1e87479_app13.docx]

Multimedia Appendix 14
Crosstab of program format × timeline months.
[DOCX File, 17 KB - mededu_v12i1e87479_app14.docx]

Multimedia Appendix 15
Average credits by mandatory internship requirement.
[PNG File, 23 KB - mededu_v12i1e87479_app15.png]

Multimedia Appendix 16
Program duration distribution by program format.
[PNG File, 36 KB - mededu_v12i1e87479_app16.png]

Multimedia Appendix 17
Cluster profiles of graduate health informatics programs (N=107).
[DOCX File, 18 KB - mededu_v12i1e87479_app17.docx]

Multimedia Appendix 18
Commission on Accreditation for Health Informatics and Information Management Education accreditation by cluster.
[DOCX File, 17 KB - mededu_v12i1e87479_app18.docx]

XSL•FO
RenderX

Checklist 1
STROBE checklist for cross-sectional studies.
[DOCX File, 19 KB - mededu_v12i1e87479_app19.docx]

References

1. Ma J, Pender M, Oster M. Trends in college pricing and student aid 2024. College Board. 2024. URL: https://research.collegeboard.org/trends [accessed 2026-04-14]
2. Financial value transparency and gainful employment. Federal Register. 2023. URL: https://www.federalregister.gov/documents/2023/10/10/2023-20385/financial-value-transparency-and-gainful-employment [accessed 2026-04-14]
3. Cooper P. High costs, uneven value: repairing the federal role in postsecondary education. American Enterprise Institute. 2025. URL: https://www.aei.org/research-products/testimony/high-costs-uneven-value-repairing-the-federal-role-in-postsecondary-education/ [accessed 2026-04-14]
4. Lee T. Economist puts the human back in human capital theory. University of Chicago News. 2025 Jul 9. URL: https://news.uchicago.edu/story/economist-puts-human-back-human-capital-theory [accessed 2026-04-14]
5. Downs L. Revisiting the cost and price of distance education. WCET. 2025 Jan 23. URL: https://wcet.wiche.edu/frontiers/2025/01/23/revisiting-the-cost-and-price-of-distance-education/ [accessed 2026-04-14]
6. 34 CFR 600.2 definitions. Electronic Code of Federal Regulations. URL: https://www.ecfr.gov/current/title-34/part-600/section-600.2 [accessed 2026-04-14]
7. CAHIIM accreditation manual. Commission on Accreditation for Health Informatics and Information Management Education (CAHIIM). 2024. URL: https://www.cahiim.org/wp-content/uploads/2024/11/cahiim-policy-manual.pdf [accessed 2026-04-14]
8. Valenta AL, Berner ES, Boren SA, et al. AMIA Board White Paper: AMIA 2017 core competencies for applied health informatics education at the master's degree level. J Am Med Inform Assoc 2018 Dec 1;25(12):1657-1668. [doi: 10.1093/jamia/ocy132] [Medline: 30371862]
9. Bichel-Findlay J, Koch S, Mantas J, et al. Recommendations of the International Medical Informatics Association (IMIA) on education in biomedical and health informatics: second revision. Int J Med Inform 2023 Feb;170:104908. [doi: 10.1016/j.ijmedinf.2022.104908] [Medline: 36502741]
10. Mantas J, Ammenwerth E, Demiris G, et al. Recommendations of the International Medical Informatics Association (IMIA) on education in biomedical and health informatics. first revision. Methods Inf Med 2010 Jan 7;49(2):105-120. [doi: 10.3414/ME5119] [Medline: 20054502]
11. Patel JS, Vo H, Nguyen A, Dzomba B, Wu H. A data-driven assessment of the U.S. health informatics programs and job market. Appl Clin Inform 2022 Mar;13(2):327-338. [doi: 10.1055/s-0042-1743242] [Medline: 35354210]
12. Kampov-Polevoi J, Hemminger BM. A curricula-based comparison of biomedical and health informatics programs in the USA. J Am Med Inform Assoc 2011;18(2):195-202. [doi: 10.1136/jamia.2010.004259] [Medline: 21292707]
13. Were MC, Gong W, Balirwa P, et al. Comparative analysis of competency coverage within accredited master's in health informatics programs in the East African region. J Am Med Inform Assoc 2021 Aug 13;28(9):1843-1848. [doi: 10.1093/jamia/ocab075] [Medline: 34151967]
14. Sapci AH, Sapci HA. Teaching hands-on informatics skills to future health informaticians: a competency framework proposal and analysis of health care informatics curricula. JMIR Med Inform 2020 Jan 21;8(1):e15748. [doi: 10.2196/15748] [Medline: 31961328]
15. Han Q. Topics and trends of health informatics education research: scientometric analysis. JMIR Med Educ 2024 Dec 11;10:e58165. [doi: 10.2196/58165] [Medline: 39661981]
16. Strengthening the Reporting of Observational Studies in Epidemiology. STROBE Statement. URL: https://www.strobe-statement.org/ [accessed 2026-04-14]
17. Accelerated program cost and financial aid. Luddy School of Informatics, Computing, and Engineering. URL: https://web.archive.org/web/20240511061251/https://luddy.indianapolis.iu.edu/cost-financial-aid/accelerated-cost-aid.html [accessed 2026-04-29]
18. Beks H, Walsh S, Clayden S, Watson L, Zwar J, Alston L. Financial implications of unpaid clinical placements for allied health, dentistry, medical, and nursing students in Australia: a scoping review with recommendations for policy, research, and practice. BMC Health Serv Res 2024 Nov 15;24(1):1407. [doi: 10.1186/s12913-024-11888-y] [Medline: 39543688]
19. Wray N, McCall L. Money matters: students' perceptions of the costs associated with placements. Med Educ 2007 Oct;41(10):975-981. [doi: 10.1111/j.1365-2923.2007.02840.x] [Medline: 17908115]
20. Hussein M, Pavlova M, Ghalwash M, Groot W. The impact of hospital accreditation on the quality of healthcare: a systematic literature review. BMC Health Serv Res 2021 Oct 6;21(1):1057. [doi: 10.1186/s12913-021-07097-6] [Medline: 34610823]
21. Chen K, Qiu L, Xie X, Bai Y. Dynasmile: video-based smile analysis software in orthodontics. SoftwareX 2025 Feb;29:102004. [doi: 10.1016/j.softx.2024.102004] [Medline: 40438663]
22. Fernandes Prabhu D, Gurupur V, Stone A, Trader E. Integrating artificial intelligence, electronic health records, and wearables for predictive, patient-centered decision support in healthcare. Healthcare (Basel) 2025 Oct 30;13(21):2753. [doi: 10.3390/healthcare13212753] [Medline: 41228120]

XSL•FO
RenderX

23. Health information workforce shortages persist as AI shows promise. American Health Information Management Association. 2023. URL: https://www.ahima.org/news-publications/press-room-press-releases/2023-press-releases/health-information-workforce-shortages-persist-as-ai-shows-promise-ahima-survey-reveals/ [accessed 2026-04-14]

Abbreviations

AI: artificial intelligence
AMIA: American Medical Informatics Association
CAHIIM: Commission on Accreditation for Health Informatics and Information Management Education
HCT: Human Capital Theory
IRB: institutional review board
STROBE: Strengthening the Reporting of Observational Studies in Epidemiology

Edited by W Jerjes; submitted 10.Nov.2025; peer-reviewed by CY Su, J Xi, K Chen; revised version received 22.Feb.2026; accepted 18.Mar.2026; published 01.May.2026.

Please cite as:
Sawesi S, Dolezel DM, Presingu P, Irungu M
Value, Structure, and Curriculum in US Graduate Health Informatics Programs: Cross-Sectional Study
JMIR Med Educ 2026;12:e87479
URL: https://mededu.jmir.org/2026/1/e87479
doi:10.2196/87479

© Suhila Sawesi, Diane M Dolezel, Pranitha Presingu, Michael Irungu. Originally published in JMIR Medical Education (https://mededu.jmir.org), 1.May.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Education, is properly cited. The complete bibliographic information, a link to the original publication on https://mededu.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Data Science Education for Residents, Researchers, and Students in Psychiatry and Psychology: Program Development and Evaluation Study

Hayoung K Donnelly[1,2], PhD; David Mandell[1], SCD; Sy Hwang[2], MS; Emily Schriver[2,3], MS; Ugurcan Vurgun[2], PhD; Graydon Neill[4], BS; Esha Patel[1], BS; Megan E Reilly[1], MPH; Michael Steinberg[1], MA; Amber Calloway[1], PhD; Robert Gallop[1], PhD; Maria A Oquendo[1], MD, PhD; Gregory K Brown[1], PhD; Danielle L Mowery[2,5], PhD

[1]Department of Psychiatry, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, United States

[2]Institute for Biomedical Informatics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, United States

[3]University of Pennsylvania Health System, University of Pennsylvania, Philadelphia, PA, United States

[4]PennDNA, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, United States

[5]Department of Biostatistics, Epidemiology and Informatics, Perelman School of Medicine, University of Pennsylvania, 3700 Hamilton Walk, Philadelphia, PA, United States

Corresponding Author:
Danielle L Mowery, PhD
Institute for Biomedical Informatics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, United States

Abstract

Background: The use of artificial intelligence (AI) to analyze health care data has become common in behavioral health sciences. However, the lack of training opportunities for mental health professionals limits clinicians' ability to adopt AI in clinical settings. AI education is essential for trainees, equipping them with the literacy needed to implement AI tools in practice, collaborate effectively with data scientists, and develop skills as interdisciplinary researchers with computing skills.

Objective: As part of the Penn Innovation in Suicide Prevention Implementation Research Center, we developed, implemented, and evaluated a virtual workshop to educate psychiatry and psychology trainees on using AI for suicide prevention research.

Methods: The workshop introduced trainees to natural language processing (NLP) concepts and Python coding skills using Jupyter notebooks within a secure Microsoft Azure Databricks cloud computing and analytics environment. We designed a 3-hour workshop that covered 4 key NLP topics: data characterization, data standardization, concept extraction, and statistical analysis. To demonstrate real-world applications, we processed chief complaints from electronic health records to compare the prevalence of suicide-related encounters across populations by race, ethnicity, and age. Training materials were developed based on standard NLP techniques and domain-specific tasks, such as preprocessing psychiatry-related acronyms. Two researchers drafted and demonstrated the code, incorporating feedback from the Methods Core of the Innovation in Suicide Prevention Implementation Research to refine the materials. To evaluate the effectiveness of the workshop, we used the Kirkpatrick program evaluation model, focusing on participants' reactions (level 1) and learning outcomes (level 2). Confidence changes in knowledge and skills before and after the workshop were assessed using paired *t* tests, and open-ended questions were included to gather feedback for future improvements.

Results: A total of 10 trainees participated in the workshop virtually, including residents, postdoctoral researchers, and graduate students from the psychiatry and psychology departments. The participants found the workshop helpful (mean 3.17 on a scale of 1 - 4, SD 0.41). Their overall confidence in NLP knowledge significantly increased (P=.002) from 1.35 (SD 0.47) to 2.79 (SD 0.46). Confidence in coding abilities also improved significantly (P=.01), increasing from 1.33 (SD 0.60) to 2.25 (SD 0.42). Open-ended feedback suggested incorporating thematic analysis and exploring additional datasets for future workshops.

Conclusions: This study illustrates the effectiveness of a tailored data science workshop for trainees in psychiatry and psychology, focusing on applying NLP techniques for suicide prevention research. The workshop significantly enhanced participants' confidence in conducting data science research. Future workshops will cover additional topics of interest, such as working with large language models, thematic analysis, diverse datasets, and multifaceted outcomes. This includes examining how participants' learning impacts their practice and research, as well as assessing knowledge and skills beyond self-reported confidence through methods such as case studies for deeper insights.

(JMIR Med Educ 2026;12:e75125) doi:10.2196/75125

XSL•FO
RenderX

KEYWORDS

data science; AI literacy education; artificial intelligence; psychiatry education; suicide research; natural language processing; cloud computing

Introduction

Artificial intelligence (AI) applications in behavioral medicine now span a wide range, including detecting at-risk populations, assisting clinicians in decision-making, and providing feedback to improve the quality of interventions [1]. Exposure to data science and AI education is crucial for medical trainees, as it prepares them to implement AI tools in their practice with the necessary literacy, collaborate effectively with data scientists, and explore careers as interdisciplinary researchers [2,3]. Medical trainees recognize the importance of AI and support its inclusion in their curriculum [4,5]; however, they also report low levels of knowledge and skills in addition to anxiety about using AI [5-7]. As a result, AI literacy among clinicians remains low, hindering the adoption of AI in clinical settings [8,9].

Self-learning methods, such as online articles and media resources, have been found to be the most common way for students to learn AI, with more than 70% relying on these approaches [2,10]. However, only 36% of students have exposure to formal AI training [2]. Previous studies have highlighted the need for guided and structured AI education tailored to health trainees, such as formal curricula and workshops [3,8]. Despite this need, barriers to implementing formal AI education persist, primarily due to limited time within an already demanding medical curriculum. A trainee-centered approach to structured learning opportunities may help reduce their burden. This includes integrating AI training into the existing didactic schedule [11] and adopting a human-centered design from recruitment to workshop facilitation.

Trainees in psychiatry and psychology may face even more limited opportunities for AI training tailored to their specialty and interests. Most existing AI education programs focus on general AI concepts and broad medical applications rather than addressing field-specific needs [3]. While some programs have been tailored for a specific domain such as radiology, a systematic review showed a lack of AI education programs specifically designed for psychiatry and psychology trainees [3]. Clinicians' learning of data science should be directly applicable to their area of expertise, helping them generate ideas for their research applications. For instance, AI education programs for psychiatry and psychology trainees should be designed around mental health–related questions, data sources, and AI applications to ensure meaningful translation into their research and clinical practice.

One area to consider in AI education for psychiatry and psychology trainees is the application of AI in suicide prevention. Researchers are leveraging AI to identify patients at risk for suicidal thoughts and behaviors in clinical settings and the wider population [12]. Research in this space can scan electronic health records (EHRs) to identify elevated suicide risk, improve ED suicide screening, and decrease racial disparities in suicide [12]. This underscores the importance of becoming familiar with these technologies as future clinicians. In particular, analyzing text data and developing natural language processing (NLP)–based AI can deepen the understanding of the complexity of suicidal ideation and behaviors, allowing for the design of contextually sensitive suicide prevention strategies. For instance, clinical notes from EHR and NLP have demonstrated their strengths in enhancing suicide prevention strategies [13,14]. Developing AI education focused on suicide risk would offer trainees valuable insights into leveraging AI for both their research and practice.

Finally, developing and implementing AI education for medical trainees requires incorporating a structured evaluation process to assess its effectiveness. However, among 41 existing AI education development studies in medicine, only 5 included formal evaluation methods, such as structured surveys and assessments, whereas the remaining 36 studies solely described the education programs without measuring their impact [8]. Without rigorous evaluation, it is difficult to determine whether these programs effectively enhance knowledge and skills among trainees. One study specifically highlighted the importance of a competency-based curriculum in enhancing AI knowledge and skills, ultimately increasing the adoption of AI in clinical practice [8].

To address these gaps, we developed and implemented a 3-hour workshop to introduce psychiatry and psychology trainees to the applications of AI in suicide research. Specifically, we incorporated NLP to highlight the potential of leveraging clinical text data for psychiatry research and adopted a program evaluation framework to systematically assess the program's efficacy.

Methods

Overview

We designed, disseminated, and evaluated a 3-hour training workshop to introduce psychiatry and psychology trainees to data science and AI. The workshop was conducted using Python Jupyter notebooks, Microsoft Azure Databricks, and Zoom (Zoom Communications). We began with a brief introduction and presurvey (5 min), followed by a Databricks setup session originally planned for 15 minutes. Although participants' accounts and workspace access had been configured in advance, the setup process took approximately 30 minutes, which was 15 minutes longer than expected. A site reliability engineer was present to help manage the situation effectively; however, we needed to cancel the welcome speech by a psychiatry faculty member. The introductory remarks were intended to highlight the need for NLP research in suicide prevention. Once all participants' environments were set up, we began the first phase of the hands-on coding practice, which focused on data characterization and preprocessing (1 h 20 min). We took a 10-minute break and then continued with the second phase, which covered concept extraction and statistical analysis (1 h 20 min). The session concluded with a wrap-up that included a postsurvey and office hour sign-up (10 min).

XSL•FO
RenderX

Ethical Considerations

It was determined by the University of Pennsylvania Institutional Review Board that this project is a quality improvement initiative that does not meet the definition of human subjects' research. Hence, further institutional review board review is not required. Before starting the survey, participants were informed that the data would be used to develop or enhance future workshops and that survey participation was voluntary. No identifiable information was collected from the survey, and participants did not receive any compensation.

Computing Environment

To optimize our instruction time during the workshop, we required an environment preconfigured with the appropriate security for storing and analyzing deidentified clinical data, access to workshop materials, and the necessary software packages for processing data. We leveraged a secure cloud computing environment within Microsoft Azure and created a workspace designed to meet Health Insurance Portability and Accountability Act security requirements and provide controlled access to the study materials. Lessons were constructed in Jupyter notebooks using the Microsoft Azure Databricks analytics platform and written in Python that explained rationale and demonstrated the processing steps in context.

Workshop Materials

The main workshop materials included practice data and prewritten code in Jupyter notebooks. The coding practice primarily focused on running prewritten code rather than writing new code, as the registration survey indicated that most participants had little to no experience with Python coding and NLP. However, we also offered participants the flexibility to modify the code during or after the workshop if they wished. Most importantly, we prioritized creating an interactive learning environment rather than delivering 1-directional instruction. For example, we regularly checked in with participants to assist with any issues they encountered during the coding practice, and 3 data scientists, in addition to the instructors, were available to provide support.

Specifically, the objective of the workshop materials was to demonstrate the application of data science, AI, and NLP to answer 3 biomedical research questions (RQs) using a dataset of chief complaints from the EHR:

- RQ1: How many chief complaints are related to suicidality?
- RQ2: Is the prevalence rate of suicidality statistically different between White and Black or African American populations?
- RQ3: Does the prevalence rate of suicidality increase with age?

To answer these questions, we organized the workshop materials to address 4 subtasks in data science and AI: data characterization, data standardization, concept extraction, and statistical analysis. For data characterization, we demonstrated how to generate frequency distributions, identify missing data, and handle outliers within the dataset. For data standardization, we applied general standardization techniques with Python function *str*, including addressing lexical variation with case reduction (eg, replacing "SUICIDE" with "suicide"). We used the WordNetLemmatizer for lemmatization to address term variation within the same concept. For instance, "suicide," "suicidal," and "suicidality" were standardized to "suicide." We detected negation using *negex* (eg, "no suicide thought" was classified as "negated"). We also applied domain- and context-specific standardization. For example, in medical settings within the state of Pennsylvania, "302" refers to involuntary commitment for emergency evaluation and treatment, whereas "201" refers to voluntary commitment. These context-specific terms were replaced with descriptive and general terms (eg, "302" was replaced with "involuntary commitment").

For concept extraction, we developed a dictionary containing suicide-related terms (eg, "suicide ideation" and "fatal self-harm behavior") and applied them using regular expressions. Finally, for statistical analysis, we demonstrated how to calculate the prevalence rate of suicidality between demographic groups and apply a *t* test to determine if there is a significant difference in prevalence rates between the groups, namely White and Black or African American populations. We also demonstrated how to examine the predictive effect of age (independent variable) on suicidality (dependent variable) using logistic regression. Two researchers drafted and demonstrated the code, and feedback from the Methods Core of the Penn Innovation in Suicide Prevention Implementation Research Center was integrated to improve the materials. The Methods Core's mission is to support the Penn Innovation in Suicide Prevention Implementation Research Center's suicide prevention research by applying computational, implementation science, and human-centered design approaches. Their role in the development and dissemination of this workshop was to provide feedback from both data science and clinical perspectives, ensuring that the workshop met the needs of psychiatry and psychology trainees and was developed in a user-friendly format. This included reorganizing NLP tasks to distinguish between content that target audiences may already know (eg, mental health–related terms) and content they may not know (eg, NLP-specific language). Feedback also recommended introducing types of clinical data (eg, chief complaints), referencing existing studies related to the research problem (eg, health disparities), and designing psychological research questions (eg, predicting suicide risk by race, ethnicity, and age).

Trainee-Centered Recruitment

Human-centered design was used to design the recruitment flyer, which prioritized clarity and appeal through the use of a logical visual hierarchy and accessible language and by highlighting speaker credentials to build trust and credibility. A QR code, which directed users to the Research Electronic Data Capture registration survey, was included in the flyer to simplify the registration process and minimize barriers to action in an effort to maximize participation. A digital flyer was shared via the Department of Psychiatry's email listserv at the University of Pennsylvania to recruit workshop participants. We did not restrict the number of workshop participants. Participation was open to any trainees (ie, students, postdoctoral researchers, and residents) in the Departments of Psychology

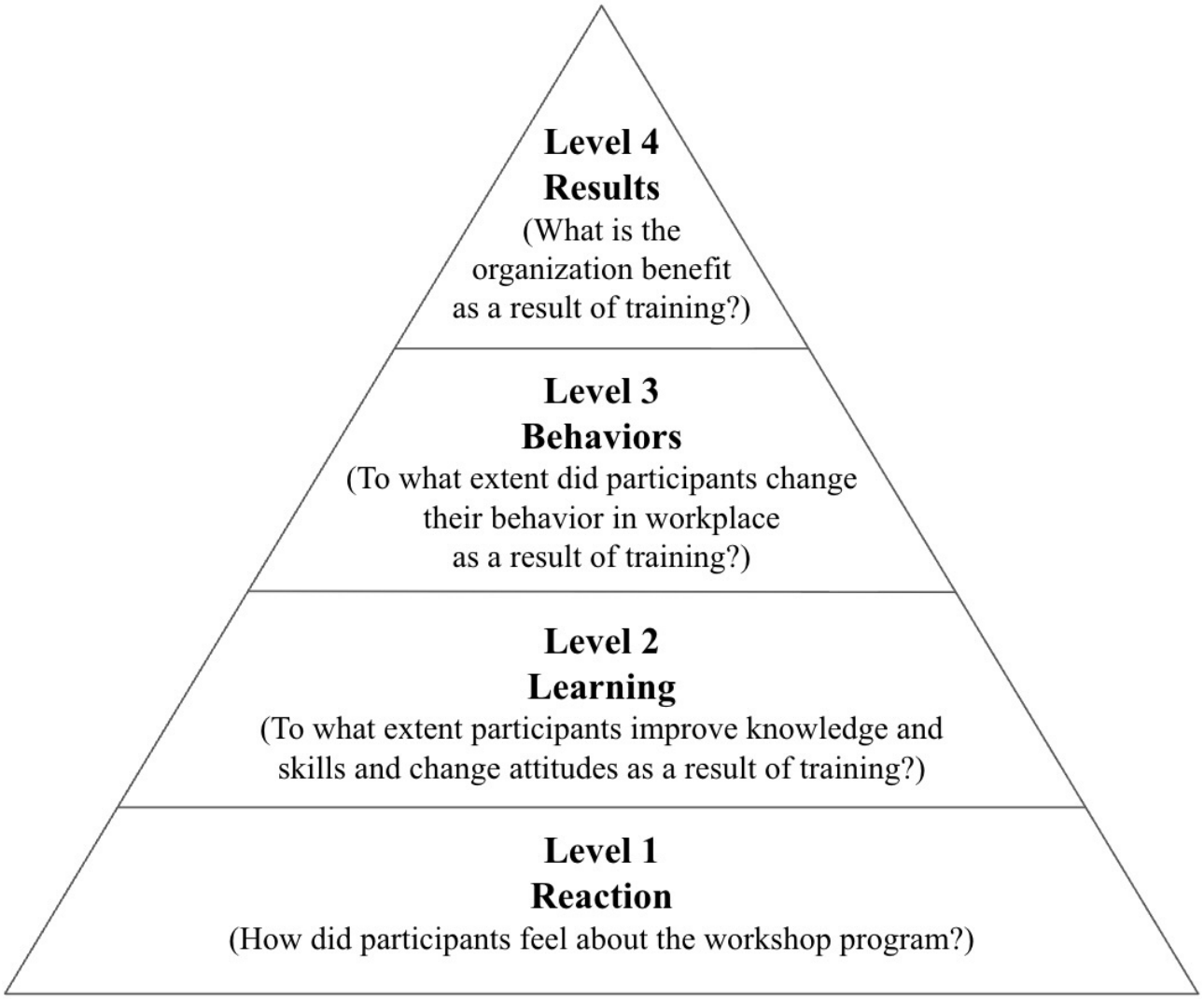
XSL•FO
RenderX

and Psychiatry. With the aim of reducing the burden on trainees participating in the data science workshop alongside their regular training requirements, we collaborated with the resident training office. To enhance dissemination efforts, we aligned the workshop schedule with the annual resident training calendar, leveraged prescheduled didactics time for trainees, and remained flexible toward other obligations. The 3-hour training workshop took place in August 2024 during a prescheduled weekday didactic session for trainees.

Workshop Evaluation

We adopted the Kirkpatrick program evaluation model [15], a model for evaluating the impact of training programs (Figure 1). The Kirkpatrick program evaluation model entails 4 levels of evaluation: level 1 (reaction), level 2 (learning), level 3 (behaviors), and level 4 (results). Specifically, we evaluated the impacts of our learning modules for participating trainees using level 1 (reaction) and level 2 (learning). Level 1 (reaction) assesses whether trainees found the training helpful. Level 2 (learning) assesses whether the trainees acquired the intended knowledge and skills from the program.

Figure 1. Kirkpatrick model for program evaluation [15].



At level 1, participants' overall satisfaction was assessed using 2 questions: (1) "Overall, did you find the workshop helpful?" rated on a 4-point Likert scale, ranging from 1 (not helpful) to 4 (very helpful); and (2) "How likely are you to recommend this workshop to others?" rated on a 4-point Likert scale, ranging from 1 (very unlikely) to 4 (very likely). Participants were also asked to identify specific components of the training they found helpful, with options including the introduction to overall NLP concepts, coding demonstrations, and materials that were provided. Each component was rated as 1 (helpful) or 0 (not helpful).

At level 2, the confidence level of NLP knowledge and Python coding skills was measured with 8 questions (eg, "How confident are you in the following concepts or skills?") with a 4-point Likert scale, ranging from 1 (not confident) to 4 (very confident). Specifically, confidence in NLP knowledge was assessed using 4 questions, each evaluating participants' confidence in data characterization, data standardization, data transformation from unstructured to structured data, and concept extraction. Participant confidence in their Python coding skills was measured with 4 questions assessing participants' confidence in performing tasks such as data standardization, data transformation, statistical analysis, and visualization using

XSL•FO
RenderX

Python. We evaluated changes in confidence level before and after the workshop using paired *t* tests for the 6 participants with pre- and post-training data. Additionally, open-ended questions such as "Are there other topics you wish future workshops to cover?" were included to gain insights into ways to revise the workshop.

Results

Level 1: Reaction Evaluation

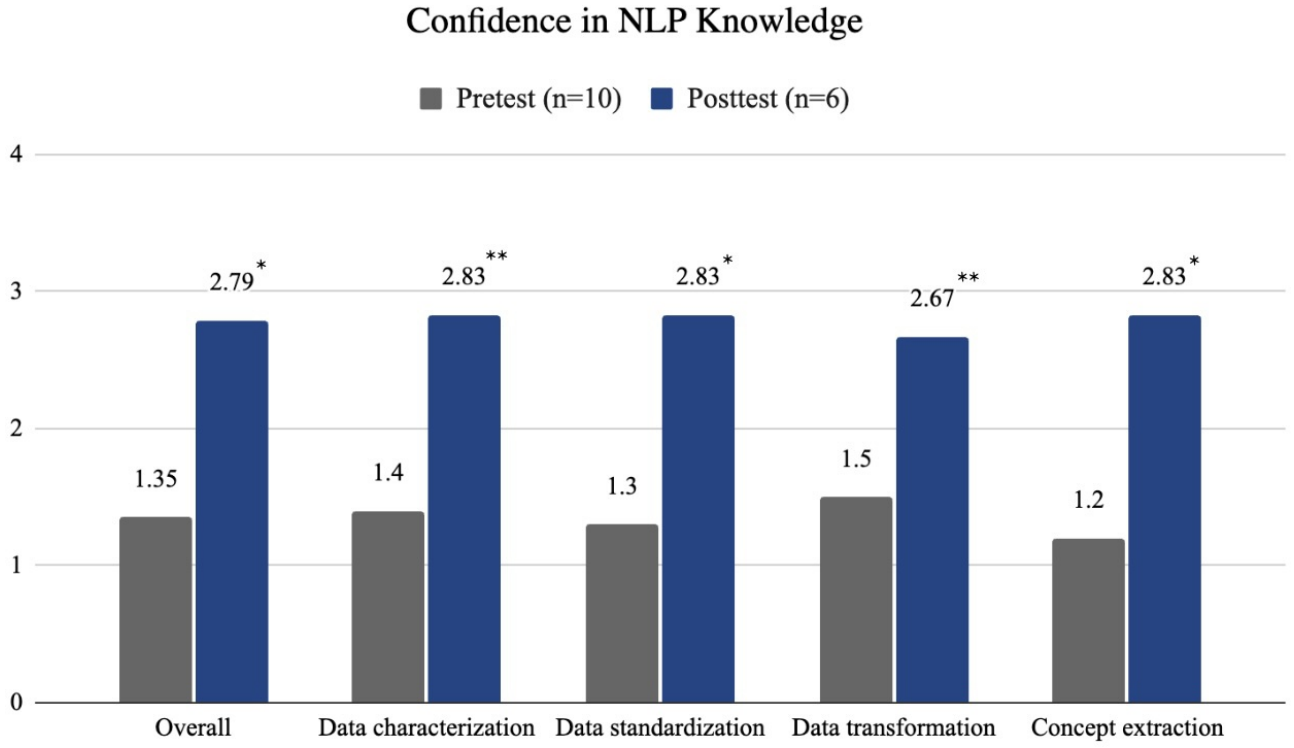
A total of 10 trainees participated in the workshop virtually. They included residents, postdoctoral researchers, and graduate students from the psychiatry and psychology departments. All participants completed the presurvey, whereas 6 (60%) participants completed the postsurvey. Only 2 (20%) of the 10 participants had experience with Python or NLP training before this workshop.

The participants found the workshop to be helpful overall (mean 3.17, SD 0.41) and expressed a willingness to recommend it to others (mean 3.17, SD 0.41). All participants found the coding demonstration and the learning materials helpful, and 5 (83%) of 6 participants found the introduction to NLP concepts helpful.

Level 2: Learning Evaluation—NLP Knowledge

Participants' confidence in their overall knowledge significantly increased (*P*=.002) from 1.35 (SD 0.47) before the workshop to 2.79 (SD 0.46) after the workshop (Figure 2). Participants' confidence in each individual topic area also significantly improved. Participants' confidence in characterizing clinical text data increased (*P*=.01) from 1.40 (SD 0.70) to 2.83 (SD 0.70). Confidence in text data preprocessing increased (*P*=.002) from 1.30 (SD 0.67) to 2.83 (SD 0.32). Confidence in transforming from unstructured to structured data improved (*P*=.01) from 1.50 (SD 0.53) to 2.67 (SD 0.42). Their confidence in extracting concepts from text data increased (*P*=.004) from 1.20 (SD 0.42) to 2.83 (SD 0.97).

Figure 2. Confidence levels in natural language processing (NLP) knowledge before and after the workshop. **P*<.05, ***P*<.01.

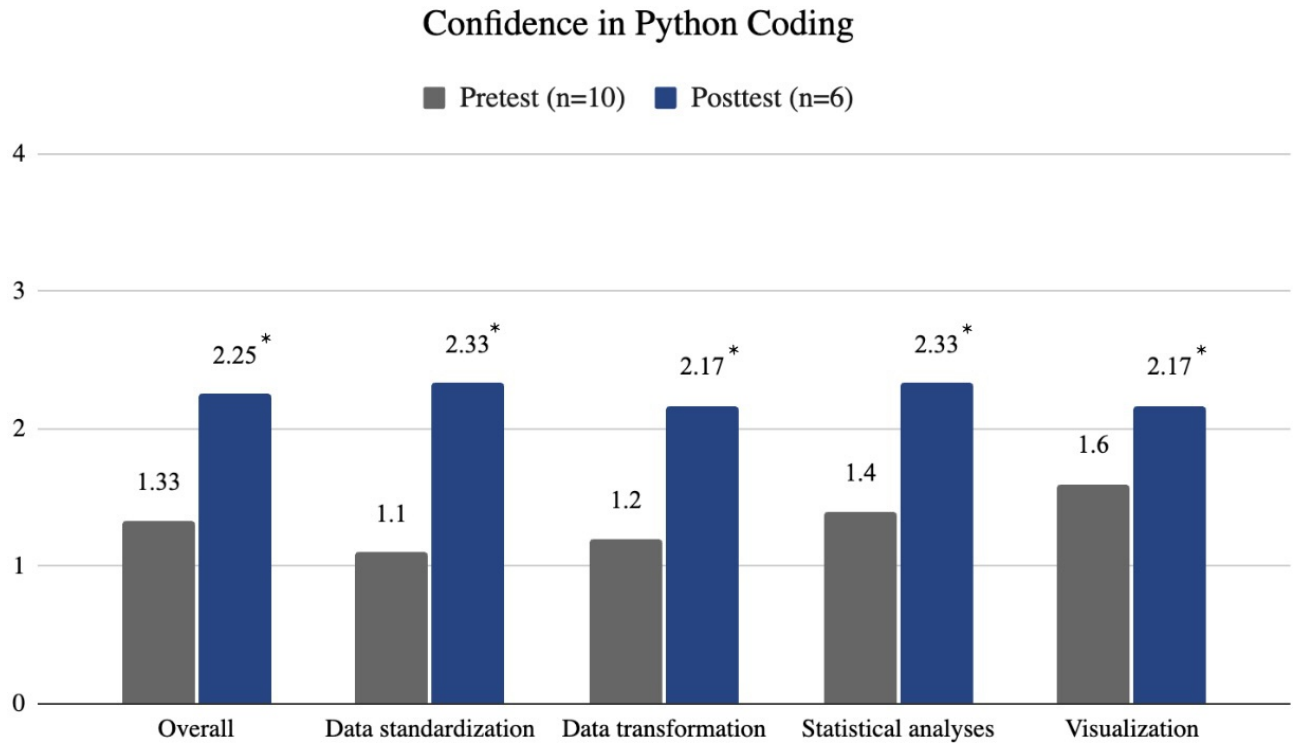


Level 2: Learning Evaluation—Python Coding Skills

Participants' confidence in their overall coding abilities also significantly improved (*P*=.01), increasing from 1.33 (SD 0.60) to 2.25 (SD 0.42) as demonstrated in Figure 3. Participants reported significant improvements in confidence across all 4 individual Python coding skills. Confidence in standardizing text data increased (*P*=.01) from 1.10 (SD 0.75) to 2.33 (SD 0.52). Similarly, confidence in transforming unstructured data into structured data improved (*P*=.01) from 1.20 (SD 0.52) to 2.17 (SD 0.41). Confidence in conducting statistical analysis (ie, descriptive statistics) using Python increased significantly (*P*=.01) from 1.40 (SD 0.75) to 2.33 (SD 0.52). Confidence in visualizing analysis results increased significantly (*P*=.04) from 1.60 (SD 0.75) to 2.17 (SD 0.41).

XSL•FO
RenderX

Figure 3. Confidence levels in python coding before and after the workshop. *P<.05.



In addition to the quantitative findings, qualitative feedback included the positive experiences of participants with the workshop. Overall, the feedback from the open-ended questions by trainees was positive and insightful (eg, “I’m glad you did this workshop, I hope there are more to come with some more hands-on experiences, but I think this served as a good introduction to the process and rationale behind the code.”). Feedback also included suggestions for incorporating thematic analysis and exploring additional datasets for future workshop development, which are discussed further in the following section.

Discussion

Principal Findings

This study illustrates the effectiveness of a tailored data science workshop for trainees in psychiatry and psychology, focusing on applying NLP techniques and clinical notes for suicide prevention research. The workshop significantly enhanced participants’ skills and confidence in conducting data science research, especially for those with substantial clinical experience but limited AI research training.

We received constructive comments from trainees through open-ended questions to guide future training efforts ranging from data-driven use cases, Python coding, and advanced NLP skills. In terms of data-driven use cases, trainees thought the workshop could be enhanced by “ ...work[ing] independently with data sets like national available data sets” and “starting our own projects working with real data.” We could incorporate more experiential training by providing more diverse datasets, including national datasets such as the National Longitudinal Study of Adolescent to Adult Health dataset or recommending that users bring their own dataset to apply these approaches. One trainee expressed an interest in “how to learn basics of coding - like any resources that would be helpful”; links to Python coding support could be provided or a Python coding bootcamp could be offered for trainees with less coding skills before the workshop. Another trainee recommended incorporating thematic analysis, for example, incorporating clustering techniques for deriving latent patterns in data. Additionally, future workshops could cover advanced NLP methods, including large language models (eg, GPT-4) to complete NLP subtasks. We could also incorporate exploring multifaceted outcomes, including clinical pathways, readmissions, and mortality, to demonstrate how to identify actionable insights from data.

Limitations

Our study has several limitations. First, we implemented this study among trainees from a single institution. Feedback might vary if trainees had more or less experience with NLP and Python coding skills at another institution. Second, in terms of assessing learning, we measured confidence rather than actual knowledge or skill in level 2. We could also assess actual levels of knowledge and skills beyond self-reported confidence using methods such as case studies for deeper insights. Third, in the future, we intend to examine whether trainees acquired level 3 (behaviors) and level 4 (results) learning. We could assess how participants’ learning impacts their practice and research in the long term. This includes whether participants applied the gained knowledge and skills to their clinical practice, research engagement, and health care improvements in the long term [16]. For instance, it would be valuable to assess whether participants are more likely to adopt NLP-based clinical support tools after the workshop (level 3) and whether increased use of these tools enhances the efficacy of clinical workflows in health care systems at a systemic level (level 4). The knowledge and skills gained may also encourage participants to initiate

XSL•FO
RenderX

independent research or engage in grant writing (level 3) to improve suicide prevention services over time (level 4). Fourth, our pilot study included a relatively small sample size. We aim to increase the size of the course by making it accessible to other trainee cohorts, including informatics, nursing, and engineering students with interest in health and AI. Finally, this study was an open trial without a controlled experimental approach. Adopting controlled methodologies, such as randomized controlled trials, will be beneficial for validating these findings in future studies.

Conclusions

We successfully developed and implemented a 3-hour virtual workshop to educate psychiatry and psychology trainees on using data science and AI for suicide prevention research. Our preliminary results suggest the workshop enhanced participants' knowledge and coding abilities, establishing a strong foundation for future advanced curriculum and learning objectives.

Acknowledgments

No generative artificial intelligence was used in any content of this manuscript, including text and graphics.

Funding

This research was funded by the National Institute of Mental Health (P50MH127511).

Data Availability

The data that support the findings of this study are available on request from the corresponding author and with approval from the institutional review board.

Authors' Contributions

Conceptualization: HKD, DM, ES, UV, MER, MS, AC, RG, DLM
Data curation: HKD, DLM
Formal analysis: HKD
Investigation: HKD
Methodology: HKD, DM, SH, ES, UV, MER, AC, RG, GKB, DLM
Project administration: HKD, ES, UV, MAO, GKB, DLM
Resources: HKD, DM, SH, GN, EP, MAO, GKB, DLM
Software: HKD, SH, GN, DLM
Supervision: DM, MAO, GKB, DLM
Validation: HKD, SH, MS
Visualization: HKD
Writing – original draft: HKD, DLM
Writing – review & editing: HKD, DM, SH, EP, MS, MAO, GKB, DLM

Conflicts of Interest

MAO receives royalties from the Research Foundation for Mental Hygiene for commercial use of the Columbia Suicide Severity Rating Scale. She volunteers as a scientific advisor to Mind Medicine and reviews grants for Alkermes and is a Trustee of Tufts University. She also advises St. George's University and Fundacion Jimenez Diaz. GKB receives royalties from the Research Foundation for Mental Hygiene for the commercial use of the Columbia Severity Rating Scale.

References

1. Dehbozorgi R, Zangeneh S, Khooshab E, et al. The application of artificial intelligence in the field of mental health: a systematic review. BMC Psychiatry 2025 Feb 14;25(1):132. [doi: 10.1186/s12888-025-06483-2] [Medline: 39953464]
2. Subaveerapandiyan A, Mvula D, Ahmad N, Taj A, Ahmed MG. Assessing AI literacy and attitudes among medical students: implications for integration into healthcare practice. JHOM 2024 Oct 28. [doi: 10.1108/JHOM-04-2024-0154]
3. Kimiafar K, Sarbaz M, Tabatabaei SM, et al. Artificial intelligence literacy among healthcare professionals and students: a systematic review. Front Health Inform 2023 Nov 11;12:168. [doi: 10.30699/fhi.v12i0.524]
4. Pinto Dos Santos D, Giese D, Brodehl S, et al. Medical students' attitude towards artificial intelligence: a multicentre survey. Eur Radiol 2019 Apr;29(4):1640-1646. [doi: 10.1007/s00330-018-5601-1] [Medline: 29980928]
5. Mousavi Baigi SF, Sarbaz M, Ghaddaripouri K, Ghaddaripouri M, Mousavi AS, Kimiafar K. Attitudes, knowledge, and skills towards artificial intelligence among healthcare students: a systematic review. Health Sci Rep 2023 Mar;6(3):e1138. [doi: 10.1002/hsr2.1138] [Medline: 36923372]
6. Güdük Ö, Vural A, Dişiaçık G. Investigating the effect of perceived empowerment on artificial intelligence anxiety levels in healthcare workers. Çalışma ve Toplum 2025 Jan 1;1(84):285-310. [doi: 10.54752/ct.1624015]

XSL•FO
RenderX

7. Theros J, Soetikno A, Liebovitz D. Foundations of AI for future physicians: a practical, accessible curriculum. Med Teach 2025 Oct;47(10):1601-1603. [doi: 10.1080/0142159X.2025.2463492] [Medline: 39940108]
8. Charow R, Jeyakumar T, Younus S, et al. Artificial intelligence education programs for health care professionals: scoping review. JMIR Med Educ 2021 Dec 13;7(4):e31043. [doi: 10.2196/31043] [Medline: 34898458]
9. Lee J, Wu AS, Li D, Kulasegaram KM. Artificial intelligence in undergraduate medical education: a scoping review. Acad Med 2021 Nov 1;96(11S):S62-S70. [doi: 10.1097/ACM.0000000000004291] [Medline: 34348374]
10. Wood EA, Ange BL, Miller DD. Are we ready to integrate artificial intelligence literacy into medical school curriculum: students and faculty survey. J Med Educ Curric Dev 2021;8:23821205211024078. [doi: 10.1177/23821205211024078] [Medline: 34250242]
11. Alelyani M, Alamri S, Alqahtani MS, et al. Radiology community attitude in Saudi Arabia about the applications of artificial intelligence in radiology. Healthcare (Basel) 2021 Jul 1;9(7):834. [doi: 10.3390/healthcare9070834] [Medline: 34356212]
12. Arowosegbe A, Oyelade T. Application of natural language processing (NLP) in detecting and preventing suicide ideation: a systematic review. Int J Environ Res Public Health 2023 Jan 13;20(2):1514. [doi: 10.3390/ijerph20021514] [Medline: 36674270]
13. Tsui FR, Shi L, Ruiz V, et al. Natural language processing and machine learning of electronic health records for prediction of first-time suicide attempts. JAMIA Open 2021 Jan;4(1):ooab011. [doi: 10.1093/jamiaopen/ooab011] [Medline: 33758800]
14. Korda A, Heide M, Nag A, et al. Suicide prediction with natural language processing of electronic health records. medRxiv. Preprint posted online on Sep 29, 2023. [doi: 10.1101/2023.09.28.23296268]
15. Kirkpatrick D, Kirkpatrick J. Evaluating Training Programs: The Four Levels: Berrett-Koehler Publishers; 2006.
16. Parker ND, Michaels M, Fisher CL, et al. A web-based training intervention for primary care providers on preparing patients for cancer treatment decisions and conversations about clinical trials: evaluation of a pilot study using mixed methods and follow-up. JMIR Med Educ 2025 Jul 17;11:e66892. [doi: 10.2196/66892] [Medline: 40674725]

Abbreviations

AI: artificial intelligence
EHR: electronic health record
NLP: natural language processing
RQ: research question

Edited by T Leung; submitted 03.Apr.2025; peer-reviewed by B Senst, NC Favour; revised version received 07.Sep.2025; accepted 16.Sep.2025; published 16.Jan.2026.

Please cite as:
Donnelly HK, Mandell D, Hwang S, Schriver E, Vurgun U, Neill G, Patel E, Reilly ME, Steinberg M, Calloway A, Gallop R, Oquendo MA, Brown GK, Mowery DL
Data Science Education for Residents, Researchers, and Students in Psychiatry and Psychology: Program Development and Evaluation Study
JMIR Med Educ 2026;12:e75125
URL: https://mededu.jmir.org/2026/1/e75125
doi:10.2196/75125

© Hayoung K Donnelly, David Mandell, Sy Hwang, Emily Schriver, Ugurcan Vurgun, Graydon Neill, Esha Patel, Megan E Reilly, Michael Steinberg, Amber Calloway, Robert Gallop, Maria A Oquendo, Gregory K Brown, Danielle L Mowery. Originally published in JMIR Medical Education (https://mededu.jmir.org), 16.Jan.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Education, is properly cited. The complete bibliographic information, a link to the original publication on https://mededu.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Effectiveness of Informed AI Use on Clinical Competence of General Practitioners and Internists: Pre-Post Intervention Study

Eyad A Qunaibi[1]; Ayman M Al-Qaaneh[2]; Baraa F Ismail[3]; Hussam I Muhidat[4]; Farhia S Rageh[5]; Najwa A Musallam[6]; Alaa K Fawzy[7]

[1]Department of International Research, MedOne Academy, Dover, DE, United States

[2]Faculty of Allied Medical Sciences, Al-Balqa Applied University, Al-Salt, Jordan

[3]Faculty of Public Health, Jinan University, Tripoli, Lebanon

[4]PriceWaterhouseCoopers Middle East, Doha, Qatar

[5]Fakeeh University Hospital, Dubai, United Arab Emirates

[6]King Hussein Cancer Center, Amman, Jordan

[7]Medical Biochemistry and Molecular Biology Department, Faculty of Medicine, Zagazig University, Zagazig, Egypt

Corresponding Author:
Ayman M Al-Qaaneh
Faculty of Allied Medical Sciences
Al-Balqa Applied University
19117
Al-Salt
Jordan
Phone: 962 779596290
Email: ayman.alqennh@bau.edu.jo

Abstract

Background: Artificial intelligence (AI) shows promise in clinical diagnosis, treatment support, and health care efficiency. However, its adoption in real-world practice remains limited due to insufficient clinical validation and an unclear impact on practitioners' competence. Addressing these gaps is essential for effective, confident, and ethical integration of AI into modern health care settings.

Objective: This study aimed to evaluate the effectiveness of informed AI use, following a tailored AI training course, on the performance of general practitioners (GPs) and internists in test-based clinical competence assessments and their attitudes toward clinical AI applications.

Methods: A pre-post intervention study was conducted with 326 physicians from 39 countries. Participants completed a baseline test of clinical decision-making skills, covering diagnosis, treatment planning, and patient counseling; attended a 1.5-hour online training on effective AI use; and then took a similar postcourse test with AI assistance permitted (GPT-4.0). Test performance and time per question were compared before and after the training. Participants also rated AI accuracy, efficiency, perceived need for structured AI training, and their willingness to use AI in clinical practice before and after the course.

Results: The average test scores improved from 56.9% (SD 15.7%) to 77.6% (SD 12.7%; $P<.001$), and the pass rate increased from 6.4% (21/326) to 58.6% (191/326), with larger gains observed among GPs and younger physicians. All skill domains (diagnosis, treatment planning, and patient counseling) improved significantly (all $P<.001$), while time taken to complete the test increased slightly from before to after the course (mean 40.25, SD 16.14 min vs 42.29, SD 14.02 min; $P=.03$). By the end of the intervention, physicians viewed AI more favorably, reporting increased confidence in its accuracy and time efficiency, greater appreciation for the need for structured AI training, and increased confidence and willingness to integrate AI into patient care.

Conclusions: Informed use of AI, based on tailored training, was associated with higher performance in test-based clinical decision-making assessments and greater confidence in using AI among GPs and internists. Building on previous research that often lacked structured training, focused primarily on model performance, or was limited in clinical scope, this study provides empirical evidence of both competence and perceptual improvement following informed AI use in a large, multinational cohort, enhancing the generalizability. These findings support the integration of structured AI training into medical education and continuing professional development to improve clinical performance and promote competent use of AI in clinical practice.

XSL•FO
RenderX

(JMIR Med Educ 2026;12:e75534) doi:10.2196/75534

KEYWORDS

artificial intelligence; AI; clinical competence; artificial intelligence training; AI training; clinical decision support; medical education; digital health tools

Introduction

Artificial intelligence (AI) has demonstrated significant potential in disease diagnosis, treatment recommendations, patient engagement, and medical writing and education [1-3]. Extensive research has explored and compared the accuracy of AI models across these domains [4-7]. In addition, AI has revolutionized diagnostic radiology by enhancing image analysis and interpretation [8]. However, despite these promising applications, most AI models and software remain within the realm of research rather than real-world settings, reflecting a notable gap in implementation [9,10].

Before widespread adoption, AI applications in health care require not only algorithm validation but also clinical validation, comparing AI-based interventions with standard treatments and existing clinical practices [11]. A limited number of studies have evaluated the effect of AI assistance in clinical settings and demonstrated improvements in the quality, efficiency, and effectiveness of health care services [12,13]. In contrast, there is an abundance of studies that have evaluated the impact of AI assistance in controlled experimental settings rather than real-world clinical scenarios. Feigerlova et al [14] highlighted the limitations of these studies, including small sample sizes, single-center designs, inadequate control for confounding factors, and the absence of well-defined clinical competence frameworks.

In addition, existing studies often do not include training on appropriate AI use. To enable AI integration into clinical practice, tailored AI training for health care practitioners is required [15]. The situation is even more critical for primary care physicians, as Liaw et al [16] have highlighted the lack of attention given to training primary care physicians in the use of AI-based tools and emphasized the necessity of such training to maximize benefits and minimize potential harms. Ideally, this training should be followed by objective assessments of improvements in clinical competence after AI use. However, evidence on the objective improvements in clinical competence with AI assistance after structured AI training remains scarce.

Beyond technical accuracy and the need for training, the successful adoption of AI in clinical settings is heavily influenced by health care professionals' perceptions and attitudes. Acceptance is an indispensable prerequisite for the widespread implementation of AI [15]. Although many studies have assessed health care professionals' general attitudes toward AI integration into medical practice [15], to the best of our knowledge, no study has specifically examined how tailored AI training impacts practitioners' perceptions of AI and their willingness to incorporate it into clinical workflows. A recent systematic review identified key factors that influence health care workers' trust in AI tools for making informed clinical decisions [17]. However, none of the 27 included studies implemented interventions with pre-post repeated-measurement designs; instead, their assessment methods were predominantly semistructured qualitative interviews, focus groups, or Likert-scale cross-sectional surveys. Collectively, many previous studies were limited by small sample sizes, narrow clinical scope, a lack of structured AI training, evaluation in simulated rather than real-world clinical contexts, and a lack of pre-post intervention assessments of perceptual change. These accumulated limitations and gaps in knowledge highlight the significance of our study.

Unlike previous research, our study approximates real-world clinical practice by integrating AI as a decision-support tool for general practitioners (GPs) and internists who were tested using real clinical case scenarios. Through a structured, tailored training course with systematic pre-post course assessments, we objectively evaluate test-based clinical competence of GPs and internists following informed AI use. We assess improvements in the skills of diagnosis, management, and patient counseling. We also uniquely assess shifts in GPs' and internists' perceptions of AI diagnostic accuracy, treatment recommendation efficacy, time efficiency, confidence, and willingness to integrate AI into clinical practice. In addition to these novelties, our study is multinational, large-scale, and outcome-driven, addressing key limitations of previous research.

By providing a replicable model for evaluating AI-assisted clinical decision-making, this study establishes a framework for optimizing AI use in health care settings, helping to bridge the gap between research-based AI performance and real-world clinical effectiveness. The aim of this study was to evaluate the test-based clinical competence and perceptions of GPs and internists, following informed AI use after a tailored AI training course.

Methods

Study Design and Procedures

This quasi-experimental study used a pre-post test design to evaluate the impact of informed AI use on clinical competence among GPs and internists. Conducted entirely online, the course and assessments were asynchronous, allowing participants to complete the training and tests at their preferred time. Each participant completed a precourse test before accessing the recorded modules, followed by a postcourse test, in which AI assistance was permitted (ChatGPT [version 4.0; OpenAI]), to measure changes in clinical competence. In this study, *informed AIuse* refers to physicians using AI tools after receiving a tailored AI training course on how to effectively and responsibly integrate these tools into their practice. Participants' self-reported previous use of AI in clinical practice was recorded and categorized for analysis (Multimedia Appendix 1).

XSL•FO
RenderX

Ethical Considerations

This study was conducted in accordance with ethical guidelines and approved by the institutional review board of MedOne Academy (MO-IA-24/25-EDU-1). All participants were informed about the study objectives and the voluntary nature of their participation, with the right to withdraw at any time. Electronic informed consent was obtained, and participants were assured of data confidentiality. Participants' names and email addresses were collected solely for administrative purposes to assign and link precourse and postcourse assessments. These identifiers were not used for analysis and were not shared with any third party. All study data were analyzed in coded and anonymized form. Access to identifiable information was strictly limited to the first and corresponding authors. All data were stored on secure, password-protected systems accessible only to authorized members of the research team. Participants did not receive any financial or material compensation for participation in this study.

Participants

Participants included GPs, encompassing family medicine specialists and internists, including those pursuing subspecialty training after internal medicine. These classifications apply throughout the study. Details of participant recruitment, eligibility verification, and enrollment timeline are provided in Multimedia Appendix 2.

Assessment Tool

Overview

This study used a structured assessment tool comprising multiple-choice questions (MCQs) designed to evaluate clinical competence across three key areas: (1) diagnosis and patient assessment, (2) treatment planning and personalized medicine, and (3) patient counseling. Each test set (A and B) included items covering these same domains to ensure content consistency across assessments. The MCQs were developed by 2 of the study authors (AKF and BFI), who are physicians, and reviewed and revised by 2 additional expert physicians (Kamel Hatahet and Suleiman Al Ashi), to ensure their relevance to participants and the adequacy in assessing the 3 studied competencies. The test sets underwent expert review for face and content validity, and their clarity and timing were further verified in the pilot phase as described subsequently. Multimedia Appendix 3 includes the source references from which the clinical cases were derived, along with selected adapted sample cases and their corresponding MCQs with answer keys.

Crossover Design and Validation

A crossover design was implemented, with participants allocated to 2 groups for the precourse test: group A received 25 MCQs, and group B received 23 MCQs of comparable difficulty. Participants who registered through the MedOne Academy platform were alternately allocated to group A or group B to maintain balanced group sizes and comparable baseline characteristics. No significant difference was observed in the GP-to-internist distribution between the 2 test-order groups (P=.11), supporting the comparability of participant composition. After the course, participants switched question sets. This crossover design was selected because it allowed each participant to serve as their own control, thereby reducing variability due to individual differences and maintaining statistical power with a feasible sample size while ensuring that all participants received the educational intervention. Conducting a conventional randomized controlled design would have required a nonintervention control group, which was impractical and would likely have compromised motivation and completion rates, as the course itself served as the primary educational and engagement incentive.

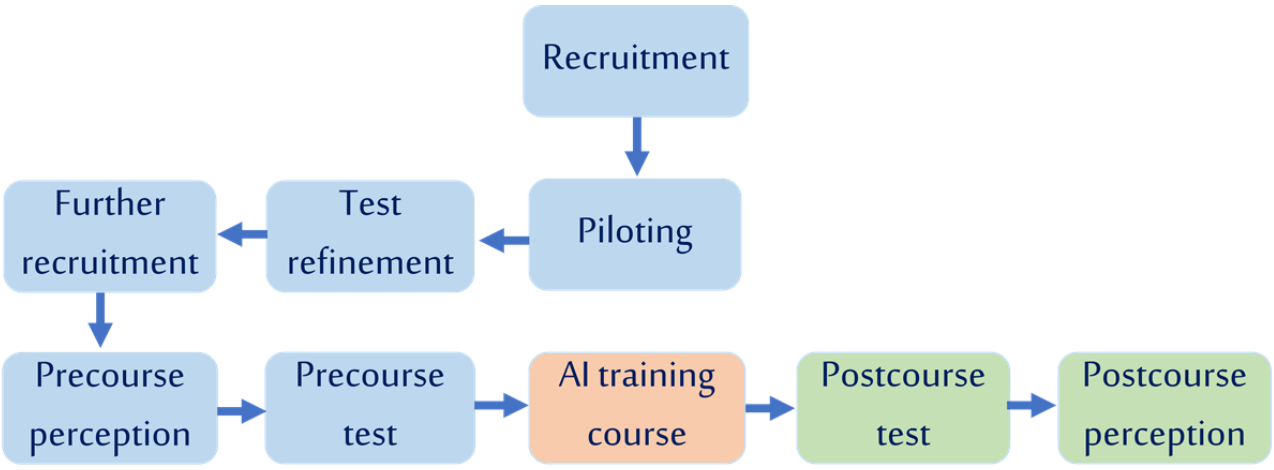
The difficulty level was preassessed by the 4 contributing experts (AKF, BFI, Kamel Hatahet, and Suleiman Al Ashi) to ensure equivalence, and the comparability of participant performance across both test sets A and B was statistically confirmed, as detailed in the Results section and Table S1 in Multimedia Appendix 4.

Assessment Administration and Scoring Criteria

Each test was allotted a maximum duration of 1 hour. A passing threshold of 80% was applied, consistent with established benchmarks in medical certification (Multimedia Appendix 5). Clear and standardized instructions on AI use were provided to participants to ensure uniform conditions across assessments. These instructions, along with the technical setup implemented to maintain the reliability of experimental conditions, are detailed in Multimedia Appendix 6. Additionally, participants completed the same set of perception questions before and after the course to assess how the course and study experience influenced their attitudes toward AI in clinical practice. These questions assessed participants' perceptions of AI accuracy in diagnosis and treatment planning, its time-saving potential in clinical settings, the perceived need for structured AI training, and their willingness and confidence to incorporate AI into clinical practice (Multimedia Appendix 7). Each participant was given up to 2 weeks to complete the entire study process after enrollment. A summary of the study flow is illustrated in Figure 1.

XSL•FO
RenderX

Figure 1. Study flow diagram. Participant recruitment and piloting (including test refinement), followed by precourse perception survey and competence test, artificial intelligence training course, and postcourse competence test and perception survey. Blue indicates preparatory phases, orange the intervention, and green postintervention assessments.



GPT-4.0 was selected as the designated large language model (LLM) for participants to use in the postcourse test. Its selection was based on published evidence demonstrating the superior performance of GPT models on medical examination–style questions compared with other contemporary LLMs [7].

Intervention (Tailored AI Course)

The training program, titled AI Skills in Medicine, was designed as an asynchronous 1.5-hour online workshop prepared and presented by the main author (EAQ). It introduced AI applications in diagnosis, management, predictive analytics, and triage and emphasized ethical and responsible implementation. The course also included brief segments on privacy, account customization, and optimal AI use in clinical reasoning. Participants were instructed to evaluate AI outputs and their underlying rationale carefully rather than accepting them as is. This approach aligns with the viewpoint of Izquierdo-Condoy et al [18], who argued that, when thoughtfully embedded within educational frameworks, generative AI tools can enhance cognitive abilities, supporting rather than replacing clinical reasoning. By addressing these various modules, the program covered the competencies suggested by previous research [16] to maximize potential benefits and minimize harms associated with AI incorporation in clinical practice. A detailed outline of the 4 modules and learning objectives is provided in Multimedia Appendix 8.

Pilot Study

Before the main study, a pilot group of 15 GPs and internists was recruited to follow the same study procedure, including attending the course and completing the tests. This phase allowed researchers to evaluate the clarity and timing of the tests as well as the applicability of the test platform. On the basis of participant feedback, the assessment process and training content were refined. For instance, open-ended questions were removed after observing excessive time consumption and noting that poor scores often reflected time constraints rather than knowledge gaps. The 15 pilot participants were excluded from the final analysis.

Studied Parameters

Primary Outcome: Clinical Competence Assessment

This study assessed multiple parameters to evaluate the impact of informed AI use comprehensively. Primary outcomes were the clinical competence scores obtained in the precourse and postcourse tests, the difference between postcourse and precourse scores, the pass or fail status based on an 80% threshold in both tests, and the total test completion time (min) before and after the course.

Secondary Outcomes: Skill-Specific Performance Measures

Secondary outcomes focused on skill-specific performance and included the number of correct answers per skill domain (diagnosis, treatment planning, and patient counseling) and the average time per correct answer per skill. The latter was calculated by summing the time spent on correct answers across all relevant case scenarios within each skill domain and dividing by the number of correct responses.

Perception Outcomes and Assessment

Perceptual changes were also evaluated, covering participants' perceptions of AI accuracy in diagnosis and treatment planning, its time-saving potential, the perceived benefit of structured AI training, and their willingness to incorporate AI into clinical practice. The perception shift was defined as the change in responses to these perception items before and after the intervention.

All outcomes were analyzed to determine improvements in clinical competence and shifts in attitudes following the tailored AI training and informed AI use.

Statistical Analysis

Categorical variables were presented as frequencies and percentages, and continuous variables were reported as means with SDs or medians with IQRs, depending on data distribution assessed using the Kolmogorov-Smirnov test. Baseline differences in categorical variables were analyzed using the chi-square goodness-of-fit test, whereas the Wilcoxon

XSL•FO
RenderX

signed-rank test (Mann-Whitney U test) or the independent samples *t* test was applied for continuous variables, as appropriate. The validity of the crossover design was assessed using the chi-square test, Mann-Whitney *U* test, or independent sample *t* test, as appropriate. The McNemar test or paired-samples *t* test was used to evaluate differences in the number of participants who passed or failed, test scores (%), and completion time (min) before and after AI use, as appropriate. An independent samples *t* test or 1-way ANOVA determined which group benefited most from AI assistance. Spearman rank correlation assessed the association between participants' age and score differences. The Wilcoxon signed-rank test with effect size analysis was used to evaluate changes in scores, time, and perceptions before and after AI use, stratified by skill domains (diagnosis, treatment planning, and patient counseling). Perception differences between groups were analyzed using the Mann-Whitney *U* or Kruskal-Wallis test, as appropriate. All statistical tests were 2 tailed, and statistical significance was set at $P<.05$. No imputations were made for missing data points. All data used in the study were analyzed using SPSS (version 25.0; IBM Corp).

Results

Recruitment and Baseline Characteristics

A total of 2336 individuals completed the recruitment form, of whom 1665 (71.3%) met the eligibility criteria by belonging to the target specialties and providing valid medical licensure. The remaining 671 (28.7%) individuals were excluded for lacking licensure or for being medical students, specialists, or professionals outside the core medical field. Of the eligible 1665 participants, 326 (19.6%) completed the precourse test, attended the course, and took the postcourse test. Of the 326 participants, 250 (76.7%) also completed the precourse and postcourse perception assessments. All 326 participants were included in the final analysis (Figure 2).

Figure 2. Recruitment, eligibility screening, and participant inclusion flowchart. Of 2336 individuals who expressed interest, 671 (28.7%) were ineligible and 1665 (71.3%) were eligible. Among eligible individuals, 1339 (80.4%) did not enroll or did not complete the postcourse test, and 326 (19.6%) completed both tests and attended the artificial intelligence training course. Of these, 76 (23.3%) did not complete both perception assessments and 250 (76.7%) completed all study components. Red boxes indicate attrition or exclusion at each stage.

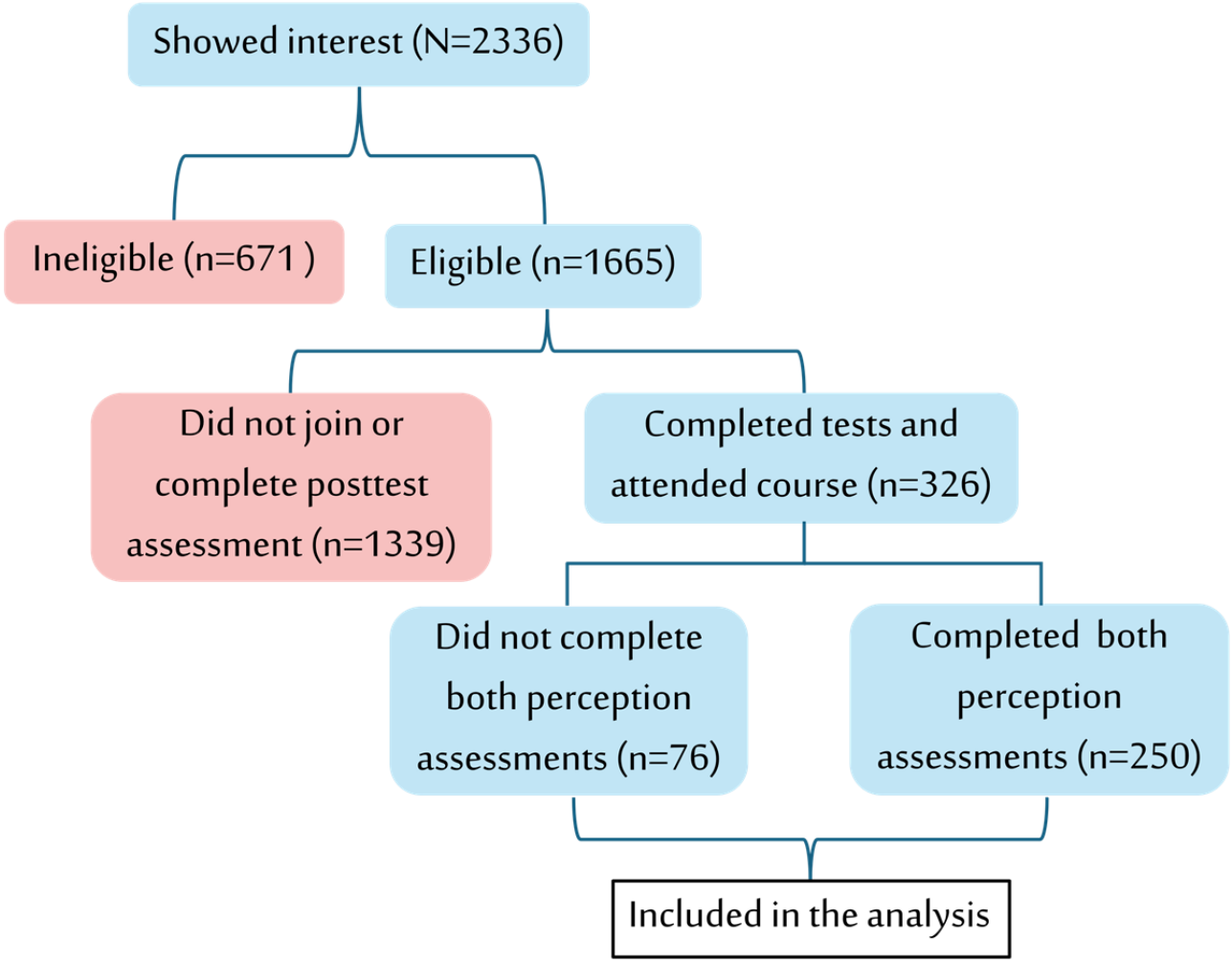


Table 1 summarizes the demographics of the 326 participants (male: n=193, 59.2%; median age 31, IQR 27.75-38 y). Most (n=229, 70.2%) were GPs. Most participants (n=208, 63.8%) reported little to no previous use of AI in clinical practice, and 308 (94.5%) had no previous AI training. Nearly half (n=173, 53.1%) served both inpatients and outpatients. The baseline median score on the precourse test was 57% (47.9%-69.6%), with a median completion time of 41.0 (27.0-56.0) minutes.

XSL•FO
RenderX

The participants were distributed across 39 countries, with the highest participation from Saudi Arabia, Syria, Egypt, Algeria, and Jordan.

Table 1. Participant demographics and baseline characteristics (n=326)[a].

Variable	Values
Sex, n (%)	
Male	193 (59.2)
Female	133 (40.8)
Specialty, n (%)	
GP[b]	229 (70.2)
Internist	97 (29.8)
Use of AI[c] in clinical practice, n (%)	
Not at all	139 (42.6)
Rarely	69 (21.2)
Sometimes	87 (26.7)
Often	18 (5.5)
Almost every day	13 (4)
Familiarity with AI, n (%)	
GP (n=229)	84 (36.7)
Internist (n=97)	34 (35.1)
Received previous training in AI use, n (%)	
Yes	18 (5.5)
No	308 (94.5)
Type of patients served, n (%)	
Outpatients	96 (29.4)
Inpatients	57 (17.5)
Both	173 (53.1)
Test results before the AI training course, n (%)	
Fail	305 (93.6)
Pass	21 (6.4)
Age (y), median (IQR)	31.00 (27.75-38.00)
GPs	30.00 (27.00-37.00)
Internists	35.00 (29.00-40.00)
Precourse score (%), median (IQR)	57.00 (47.85-69.60)
Precourse score of GPs	56.00 (44.00-64.00)
Precourse score of internists	68.00 (52.00-74.00)
Time taken for precourse test (min), median (IQR)	41.00 (27.00-56.00)

[b]GP: general practitioner.

[c]AI: artificial intelligence.

[a]Data are presented as counts (percentages) for categorical variables and as medians (IQRs) for continuous variables.

Validation of the Crossover Design

To validate the crossover design, precourse scores (P=.65), postcourse scores (P=.65), and score differences (P=.15) were compared between groups A and B, along with pass and fail comparisons between test sets A and B at both time points (pretest: P=.31; posttest: P=.19; Table S1 in Multimedia Appendix 4). None of these comparisons yielded a statistically significant difference, confirming the validity of the study design.

XSL•FO
RenderX

Effect of the Informed AI Use on Participants' Competence

Table 2 illustrates the impact of informed AI use on participants' competence. The proportion of participants who passed the test after taking the course and being allowed to use AI increased significantly (P<.001). The mean score improved from 56.88% (SD 15.65%) to 77.56% (SD 12.71%), with a mean difference of 20.68% (P<.001). The mean time taken to complete the test increased slightly (by approximately 2 minutes) from before to after the course (40.25, SD 16.14 vs 42.29, SD 14.02 minutes; P=.03).

Table 2. Effect of informed artificial intelligence use on the participants' competencies (n=326). Comparison of participants' performance before and after the artificial intelligence training course.

Variable	Before course	After course	Mean difference	Test statistic	*P* value
Test results, n (%)				χ^2_1=153.55[a]	<.001
Fail	305 (93.6)	135 (41.4)	—[b]		
Pass	21 (6.4)	191 (58.6)	—		
Score (%), mean (SD)	56.88 (15.65)	77.56 (12.71)	20.69	t_{325}=–19.58[c]	<.001
Time taken (min), mean (SD)	40.25 (16.14)	42.29 (14.02)	2.04	t_{325}=–2.17[c]	.03

[a]McNemar χ^2.

[b]Not applicable.

[c]Paired samples *t* test.

Score improvements were analyzed across participant characteristics (Table S2 in Multimedia Appendix 4), revealing no significant differences except between GPs and internists, with the former showing significantly greater improvement than the latter (23.7%, SD 19.1% vs 13.7%, SD 17.3%; P<.001). Spearman correlation analysis revealed a statistically significant weak negative correlation between participants' age and score difference (ρ=–0.143; P=.01).

Skill-Specific Performance Outcomes

To determine whether the observed improvement was broad across all competencies or specific to certain skill areas, individual skill-specific performance was analyzed. Significant score increases were observed across skills of diagnosis, treatment planning, and patient counseling (Table 3; Figure 3; P<.001 for all). However, the average time per correct answer per skill increased significantly after informed AI use (Table S3 in Multimedia Appendix 4).

Table 3. Score comparison for each skill (n=326)[a].

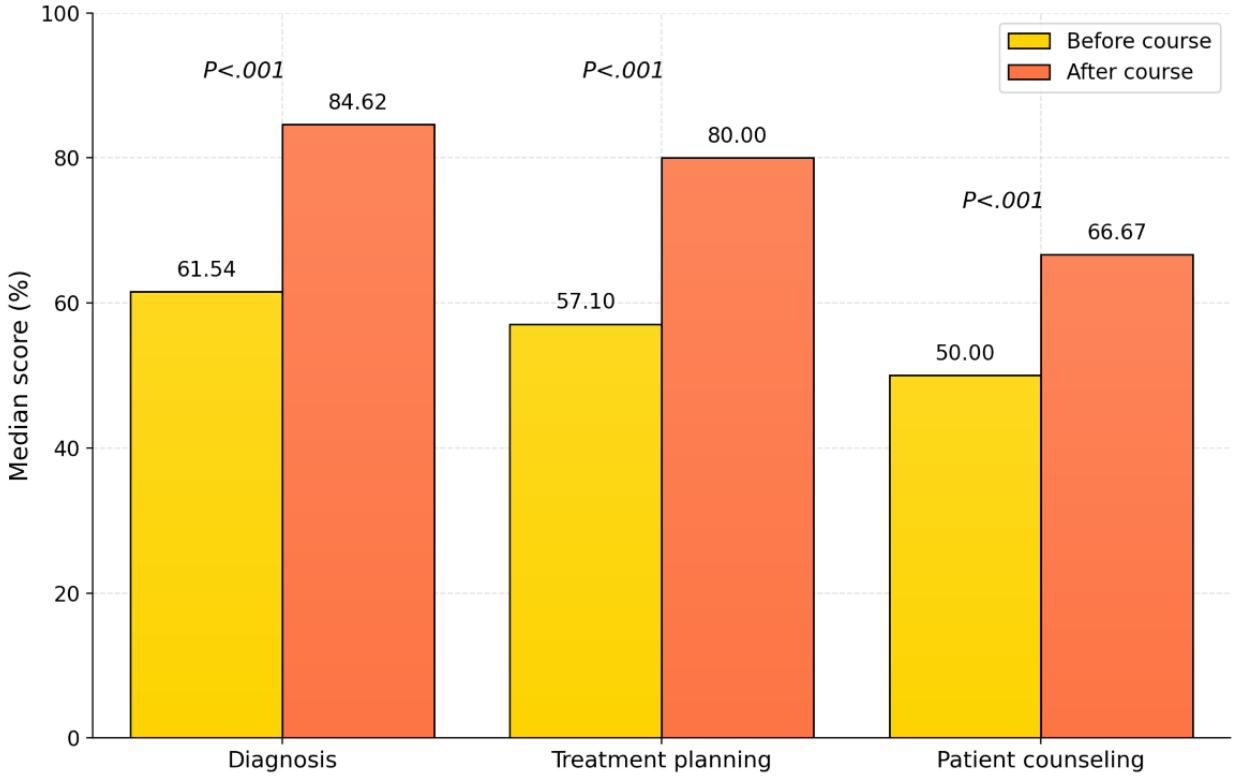
Skill	Precourse score (%), median (IQR)	Postcourse score (%), median (IQR)	*z* score	*P* value[b]	Effect size, *r*
Skill 1: diagnosis	61.54 (46.15-69.23)	84.62 (76.90-92.30)	–12.57	<.001	0.738
Skill 2: treatment planning	57.10 (42.14-70.00)	80.00 (70.00-85.71)	–12.39	<.001	0.686
Skill 3: patient counseling	50.00 (33.33-100.00)	66.67 (50.00-100.00)	–7.75	<.001	0.420

[a]Comparison of participants' precourse and postcourse scores across 3 assessed competencies: diagnosis, treatment planning, and patient counseling. Data are presented as median (IQR).

[b]Wilcoxon signed-rank test.

XSL•FO
RenderX

Figure 3. Skill-specific median scores before and after informed artificial intelligence (AI) use (N=326). Median scores (% for diagnosis, treatment planning, and patient counseling before and after the AI training course. Wilcoxon signed-rank tests showed statistically significant differences for all skills (*P*<.001).

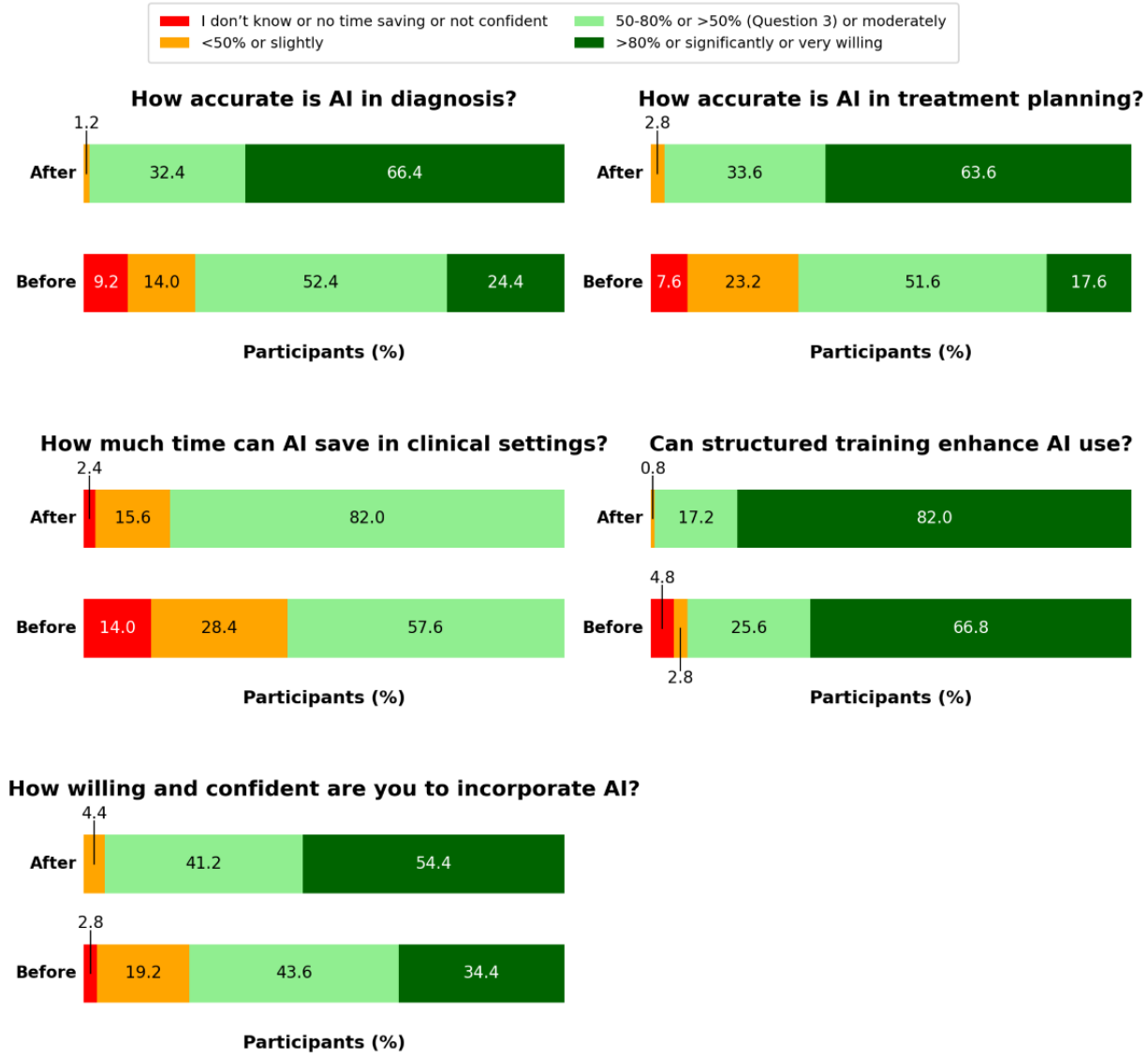


Effect of the AI Course and Informed AI Use on Participants' Perceptions

As shown in Figure 4 and Table S4 in Multimedia Appendix 4, participants' perceptions of AI accuracy, time efficiency, and integration into clinical practice improved markedly in the postcourse perception assessment, reflecting changes observed after both the course and informed AI use. All 5 perception changes (questions 1-5) were statistically significant (Table S5 in Multimedia Appendix 4). The proportion of participants who believed AI to be highly accurate in diagnosis (question 1) and treatment planning (question 2) (>80% of the time) increased substantially. More participants also recognized AI's time-saving potential (question 3), with a significant shift toward believing that AI can reduce health care professionals' workload by 50% or more. Additionally, structured AI training (question 4) was perceived as highly beneficial, with a notable rise in participants rating it as significantly effective. Confidence in and willingness to incorporate AI into clinical practice (question 5) also increased, with a greater proportion of participants feeling very willing and confident after the course and informed AI use.

XSL•FO
RenderX

Figure 4. Changes in participants' perceptions of artificial intelligence (AI) before and after the course. Stacked bar charts show participants' responses to 5 perception questions administered before and after the AI training course. Each bar represents the percentage of participants selecting each response category. Dark green indicates very positive perceptions (>80%, significantly, or very willing), light green indicates intermediately positive perceptions (50%-80% or moderately; >50% for question 3), orange indicates conservative perceptions (<50% or slightly), and red represents uncertain or negative responses ("I do not know," "no time saving," or "not confident").



No significant difference in perception shift was observed based on participants' sex, specialty, type of patients served, or previous AI training. However, participants who were unfamiliar with AI before the course exhibited a greater shift in their perception of AI diagnostic accuracy (P=.01) and greater willingness and confidence to incorporate AI tools into clinical practice (P<.001; Table S6 in Multimedia Appendix 4).

Discussion

Principal Findings

This study aimed to examine the effect of informed AI use, following a tailored course, on test-based clinical decision-making and the perceptions of GPs and internists. We have shown that informed AI use was associated with higher clinical decision-making scores using clinical scenarios that simulated daily practice. Furthermore, participants reported higher perceived AI accuracy as a diagnostic and therapeutic assistant, increased perceived time-saving efficiency, a stronger belief in the need for structured AI training programs, and greater confidence in incorporating AI into clinical practice.

Our study has several strengths and novel elements. It included structured, tailored AI training; assessed both objective changes in competence and subjective shifts in perceptions; and specifically targeted GPs and internists, a group previously underrepresented in AI research [19]. Additionally, it uniquely bridged education and objective assessments through structured precourse and postcourse tests. Moreover, this study was multinational and large scale, and it covered a comprehensive

XSL•FO
RenderX

range of clinical competencies rather than a single isolated skill, addressing key limitations cited in previous research [14]. However, because participation was voluntary and completion required finishing both assessments, the analyzed sample likely represents a more AI-motivated subset of eligible registrants, which may limit the generalizability of the observed effects.

Furthermore, this study directly addresses a widely recognized training gap in medical education. Oftring et al [20] emphasized that the growing number and impact of medical AI applications necessitate more AI-focused curricula and research on their educational impact, as most current practitioners and trainees remain underprepared for AI integration into clinical practice. Similarly, Ichikawa et al [21] found that US Colleges of Osteopathic Medicine largely lack AI policy guidance or training for students and faculty, highlighting the urgent need to implement need-driven training in their programs.

The observed improvement in competence scores (57%-78%) is substantial. Whether AI assistance alone, without structured training, can reliably enhance physicians' performance remains debated. AI has been shown to improve diagnostic accuracy in specific tasks such as detecting breast lesions on ultrasound [22], increasing fracture detection sensitivity [23], and interpreting pediatric radiographs [24]. However, Goh et al [25], in a study closest to ours in terms of participant specialties, found that unguided access to LLMs did not improve overall diagnostic reasoning among family medicine, internal medicine, or emergency medicine physicians.

Overall, the substantial improvement observed in our study is unlikely to be attributable to AI use alone, separately from the tailored course. Our training course focused on informed use of AI as a decision-support tool, helping participants double-check their reasoning, consider broader differentials, interact with AI-generated answers, and retrieve relevant literature and clinical guidelines, ultimately enhancing their responses to clinical scenarios. In addition, the substantial improvement in scores cannot be attributed to increased test-taking time, as completion time increased only slightly after the course, indicating that GPs and internists became more accurate with minimal impact on efficiency.

In contrast, a large-scale study by Yu et al [26] involving 140 radiologists and including training on AI use reported heterogeneous effects of AI assistance across 15 chest X-ray diagnostic tasks. Although the study specified that physicians received onboarding training on the AI system, it did not provide details on its content or structure. It is possible that a more clearly structured and tailored training approach, such as the one implemented in our study, might have produced more consistent improvements.

An interesting finding was that GPs demonstrated significantly greater improvement in test scores than internists (23.7% vs 13.7%, respectively). Given that baseline familiarity with AI was comparable between both groups, this discrepancy may be partly explained by the significantly lower baseline scores among GPs, which allowed a wider margin for improvement. GPs were significantly younger than internists, and younger generations are digital learners, engage more frequently in online communities [27], and use technology more fluently than older cohorts [28]. This interpretation also aligns with our findings of greater score improvement among younger participants and reflects the relatively young age of our study cohort (median 31.00 y).

The observation of no significant difference in participant score improvement based on familiarity with AI use in clinical practice aligns with findings reported by Yu et al [26] among radiologists, in which familiarity with AI tools failed to reliably predict the impact of AI assistance.

One surprising finding was the increased time per correct answer across all 3 competencies (diagnosis, treatment planning, and patient counseling) after AI use. This may reflect more deliberate and reflective reasoning, as participants were instructed to verify AI-generated suggestions and references against their own clinical judgment rather than accept outputs uncritically. Additionally, many participants were new to structured AI use; most (308/326, 94.5%) had no previous AI training, and approximately two-thirds (208/326, 63.8%) reported little to no previous AI use. Therefore, a short adaptation period was anticipated during their first guided application of AI in case-based testing.

The crossover design minimized familiarity bias and was validated by demonstrating no significant difference in precourse scores, postcourse scores, score improvements, and proportions of pass versus fail between the 2 test-order groups. This approach enhanced internal validity by enabling each participant to serve as their own control.

Beyond objective competence improvements, the observed changes in participants' perceptions of AI's accuracy, time efficiency, and potential integration into clinical practice emphasize the importance of structured training and guided AI use in realistic clinical scenarios. Following the intervention, the percentage of participants who perceived high accuracy (>80%) of optimally used AI platforms in diagnosis and treatment planning domains nearly tripled. This finding aligns with Abbas et al [7], who reported comparable accuracy of GPT models on National Board of Medical Examiners clinical subject examination questions.

The notable postcourse increase in physicians' willingness and confidence to incorporate AI into clinical practice reflects one of the most impactful outcomes of this study, suggesting the potential value of structured AI training. The systematic review by Tun et al [17] identified training and familiarity as perceived facilitators of clinician trust in AI tools in making informed clinical decisions. Our data empirically demonstrate this effect. If health professionals lack sufficient trust in AI tools, they may disregard their recommendations, limiting the potential to enhance patient outcomes and optimize clinical workflows [21]. Through its tailored AI course, this study aimed to help participants understand AI capabilities and limitations, enabling them to integrate its assistance beneficially neither erroneously ignoring its outputs nor following them uncritically.

A novel aspect of our study is its pre-post assessment of attitude changes following the intervention, which, to our knowledge, has not been previously applied among physicians in the context of AI integration. Our study uniquely provides evidence linking

XSL•FO
RenderX

structured training and competence improvement with enhanced perception of AI, highlighting the potential of targeted educational interventions to promote informed AI acceptance in clinical practice.

Implications and Future Directions

Our study highlights the potential of informed AI use after a tailored AI training course to enhance GPs' and internists' diagnostic and management skills. Despite the Federation of State Medical Boards advocating for AI competence in medical education, structured AI training remains largely absent [19]. Similarly, 82% of medical professionals in a survey by Tezpal et al [29] recognized the need for AI education. To address this gap, AI training should be integrated into medical education, residency programs, and continuing medical education to ensure responsible and effective AI use. Given the rapid evolution of AI technologies, ongoing updates and refresher courses are essential.

Future research should investigate whether these test-based competence gains are sustained over time and whether they translate into measurable improvements in patient outcomes and clinical decision quality. Finally, further research should explore customized AI training for specialties such as emergency and critical care medicine, given the crucial role of timely and accurate decisions in these specialties.

Limitations

Our study design does not allow isolation of the effect of the tailored AI training course from AI use without training, as it did not include a nonintervention control arm. Adding a third arm comparing informed versus noninformed AI use was impractical due to expected low motivation and completion rates among participants asked to repeat time-intensive tests without the educational incentive. A similar approach was taken in previous research assessing AI assistance among 140 radiologists, in which onboarding training was not isolated from AI impact [26].

Our participants were already inclined toward AI adoption, as evidenced by their willingness to enroll in our time-intensive study. This suggests that their baseline openness to AI was likely higher than that of physicians more broadly. Thus, the observed shift may underestimate the potential impact of informed AI use based on a tailored AI training course.

Conclusions

This study provides evidence that informed AI use, guided by a tailored AI course, is associated with higher clinical decision-making competence among GPs and internists in test-based clinical scenarios. Beyond higher scores across diagnosis, treatment planning, and patient counseling domains, participants reported higher perceived AI diagnostic and therapeutic accuracy, greater perceived time efficiency, and increased willingness and confidence to integrate AI as a decision-support tool into clinical practice. These findings support the integration of structured AI training into medical education and continuing professional development to facilitate informed and responsible AI use, with plausible potential to improve patient care.

Acknowledgments

This research was conducted as part of the Research Mentorship Program of MedOne Academy. The authors thank Ms Sara M Chenini from MedOne Academy for her continuous support. The senior author thanks Kamel Hatahet, nephrology consultant, North Kansas City Hospital, and Suleiman Al Ashi, MD, endocrinology, diabetes, and metabolism fellow, for their revision of the test questions.

Funding

No external financial support or grants were received from any public, commercial, or not-for-profit entities for the research, authorship, or publication of this paper.

Data Availability

All data generated or analyzed during this study are included in this published paper and its multimedia appendices.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Classification of previous artificial intelligence use.
[DOCX File , 16 KB - mededu_v12i1e75534_app1.docx]

Multimedia Appendix 2
Recruitment of participants.
[DOCX File , 16 KB - mededu_v12i1e75534_app2.docx]

XSL•FO
RenderX

Multimedia Appendix 3
Clinical case scenarios, assessment questions, and source references.
[DOCX File , 28 KB - mededu_v12i1e75534_app3.docx]

Multimedia Appendix 4
Additional statistical tables and supporting results.
[DOCX File , 30 KB - mededu_v12i1e75534_app4.docx]

Multimedia Appendix 5
Passing score rationale.
[DOCX File , 16 KB - mededu_v12i1e75534_app5.docx]

Multimedia Appendix 6
Artificial intelligence use instructions and technical setup.
[DOCX File , 16 KB - mededu_v12i1e75534_app6.docx]

Multimedia Appendix 7
Evaluation of participants' perceptions of artificial intelligence accuracy, time efficiency, training needs, and clinical adoption.
[DOCX File , 23 KB - mededu_v12i1e75534_app7.docx]

Multimedia Appendix 8
Course structure and content.
[DOCX File , 24 KB - mededu_v12i1e75534_app8.docx]

References

1. Alowais SA, Alghamdi SS, Alsuhebany N, Alqahtani T, Alshaya AI, Almohareb SN, et al. Revolutionizing healthcare: the role of artificial intelligence in clinical practice. BMC Med Educ 2023 Sep 22;23(1):689 [FREE Full text] [doi: 10.1186/s12909-023-04698-z] [Medline: 37740191]
2. Meng X, Yan X, Zhang K, Liu D, Cui X, Yang Y, et al. The application of large language models in medicine: a scoping review. iScience 2024 May 17;27(5):109713 [FREE Full text] [doi: 10.1016/j.isci.2024.109713] [Medline: 38746668]
3. Lucas HC, Upperman JS, Robinson JR. A systematic review of large language models and their implications in medical education. Med Educ 2024 Nov;58(11):1276-1285 [FREE Full text] [doi: 10.1111/medu.15402] [Medline: 38639098]
4. McGenity C, Clarke EL, Jennings C, Matthews G, Cartlidge C, Freduah-Agyemang H, et al. Artificial intelligence in digital pathology: a systematic review and meta-analysis of diagnostic test accuracy. NPJ Digit Med 2024 May 04;7(1):114 [FREE Full text] [doi: 10.1038/s41746-024-01106-8] [Medline: 38704465]
5. Rutledge GW. Diagnostic accuracy of GPT-4 on common clinical scenarios and challenging cases. Learn Health Syst 2024 Jul 25;8(3):e10438 [FREE Full text] [doi: 10.1002/lrh2.10438] [Medline: 39036534]
6. Penny P, Bane R, Riddle V. Advancements in AI medical education: assessing ChatGPT's performance on USMLE-style questions across topics and difficulty levels. Cureus 2024 Dec;16(12):e76309. [doi: 10.7759/cureus.76309] [Medline: 39850172]
7. Abbas A, Rehman MS, Rehman SS. Comparing the performance of popular large language models on the national board of medical examiners sample questions. Cureus 2024 Mar;16(3):e55991 [FREE Full text] [doi: 10.7759/cureus.55991] [Medline: 38606229]
8. Nakaura T, Ito R, Ueda D, Nozaki T, Fushimi Y, Matsui Y, et al. The impact of large language models on radiology: a guide for radiologists on the latest innovations in AI. Jpn J Radiol 2024 Jul 29;42(7):685-696. [doi: 10.1007/s11604-024-01552-0] [Medline: 38551772]
9. van de Sande D, Chung EF, Oosterhoff J, van Bommel J, Gommers D, van Genderen ME. To warrant clinical adoption AI models require a multi-faceted implementation evaluation. NPJ Digit Med 2024 Mar 06;7(1):58 [FREE Full text] [doi: 10.1038/s41746-024-01064-1] [Medline: 38448743]
10. Brereton TA, Malik MM, Lifson M, Greenwood JD, Peterson KJ, Overgaard SM. The role of artificial intelligence model documentation in translational science: scoping review. Interact J Med Res 2023 Jul 14;12:e45903 [FREE Full text] [doi: 10.2196/45903] [Medline: 37450330]
11. Maleki Varnosfaderani S, Forouzanfar M. The role of AI in hospitals and clinics: transforming healthcare in the 21st century. Bioengineering (Basel) 2024 Mar 29;11(4):337 [FREE Full text] [doi: 10.3390/bioengineering11040337] [Medline: 38671759]

XSL•FO
RenderX

12. Herter WE, Khuc J, Cinà G, Knottnerus BJ, Numans ME, Wiewel MA, et al. Impact of a machine learning-based decision support system for urinary tract infections: prospective observational study in 36 primary care practices. JMIR Med Inform 2022 May 04;10(5):e27795 [FREE Full text] [doi: 10.2196/27795] [Medline: 35507396]
13. Khosravi M, Zare Z, Mojtabaeian SM, Izadi R. Artificial intelligence and decision-making in healthcare: a thematic analysis of a systematic review of reviews. Health Serv Res Manag Epidemiol 2024;11:23333928241234863 [FREE Full text] [doi: 10.1177/23333928241234863] [Medline: 38449840]
14. Feigerlova E, Hani H, Hothersall-Davies E. A systematic review of the impact of artificial intelligence on educational outcomes in health professions education. BMC Med Educ 2025 Jan 27;25(1):129 [FREE Full text] [doi: 10.1186/s12909-025-06719-5] [Medline: 39871336]
15. Lambert SI, Madi M, Sopka S, Lenes A, Stange H, Buszello CP, et al. An integrative review on the acceptance of artificial intelligence among healthcare professionals in hospitals. NPJ Digit Med 2023 Jun 10;6(1):111 [FREE Full text] [doi: 10.1038/s41746-023-00852-5] [Medline: 37301946]
16. Liaw W, Kueper JK, Lin S, Bazemore A, Kakadiaris I. Competencies for the use of artificial intelligence in primary care. Ann Fam Med 2022;20(6):559-563 [FREE Full text] [doi: 10.1370/afm.2887] [Medline: 36443071]
17. Tun HM, Rahman HA, Naing L, Malik OA. Trust in artificial intelligence-based clinical decision support systems among health care workers: systematic review. J Med Internet Res 2025 Jul 29;27:e69678 [FREE Full text] [doi: 10.2196/69678] [Medline: 40772775]
18. Izquierdo-Condoy JS, Arias-Intriago M, Tello-De-la-Torre A, Busch F, Ortiz-Prado E. Generative artificial intelligence in medical education: enhancing critical thinking or undermining cognitive autonomy? J Med Internet Res 2025 Nov 03;27:e76340 [FREE Full text] [doi: 10.2196/76340] [Medline: 41183320]
19. Liaw W, Hischier B, James CA, Kakadiaris I, Kueper JK, Rahimzadeh V. Bridging the gap: transforming primary care through the artificial intelligence and machine learning for primary care (aim-pc) curriculum. Ann Fam Med 2024;22(6):570-571 [FREE Full text] [doi: 10.1370/afm.240537] [Medline: 39586707]
20. Oftring ZS, Deutsch K, Tolks D, Jungmann F, Kuhn S. Novel blended learning on artificial intelligence for medical students: qualitative interview study. JMIR Med Educ 2025 May 26;11:e65220 [FREE Full text] [doi: 10.2196/65220] [Medline: 40418795]
21. Ichikawa T, Olsen E, Vinod A, Glenn N, Hanna K, Lund GC, et al. Generative artificial intelligence in medical education-policies and training at US osteopathic medical schools: descriptive cross-sectional survey. JMIR Med Educ 2025 Feb 11;11:e58766 [FREE Full text] [doi: 10.2196/58766] [Medline: 39934984]
22. Lai YC, Chen HH, Hsu JF, Hong YJ, Chiu TT, Chiou HJ. Evaluation of physician performance using a concurrent-read artificial intelligence system to support breast ultrasound interpretation. Breast 2022 Oct;65:124-135 [FREE Full text] [doi: 10.1016/j.breast.2022.07.009] [Medline: 35944352]
23. Meetschen M, Salhöfer L, Beck N, Kroll L, Ziegenfuß CD, Schaarschmidt BM, et al. AI-assisted X-ray fracture detection in residency training: evaluation in pediatric and adult trauma patients. Diagnostics (Basel) 2024 Mar 11;14(6):33 [FREE Full text] [doi: 10.3390/diagnostics14060596] [Medline: 38535017]
24. Zech JR, Ezuma CO, Patel S, Edwards CR, Posner R, Hannon E, et al. Artificial intelligence improves resident detection of pediatric and young adult upper extremity fractures. Skeletal Radiol 2024 Dec;53(12):2643-2651. [doi: 10.1007/s00256-024-04698-0] [Medline: 38695875]
25. Goh E, Gallo R, Hom J, Strong E, Weng Y, Kerman H, et al. Large language model influence on diagnostic reasoning: a randomized clinical trial. JAMA Netw Open 2024 Oct 01;7(10):e2440969 [FREE Full text] [doi: 10.1001/jamanetworkopen.2024.40969] [Medline: 39466245]
26. Yu F, Moehring A, Banerjee O, Salz T, Agarwal N, Rajpurkar P. Heterogeneity and predictors of the effects of AI assistance on radiologists. Nat Med 2024 Mar 19;30(3):837-849 [FREE Full text] [doi: 10.1038/s41591-024-02850-w] [Medline: 38504016]
27. Han ER, Yeo S, Kim MJ, Lee YH, Park KH, Roh H. Medical education trends for future physicians in the era of advanced technology and artificial intelligence: an integrative review. BMC Med Educ 2019 Dec 11;19(1):460 [FREE Full text] [doi: 10.1186/s12909-019-1891-5] [Medline: 31829208]
28. Nakagawa K, Yellowlees P. Inter-generational effects of technology: why millennial physicians may be less at risk for burnout than baby boomers. Curr Psychiatry Rep 2020 Jul 13;22(9):45 [FREE Full text] [doi: 10.1007/s11920-020-01171-2] [Medline: 32661817]
29. Tezpal M, Ghosh S, Lalwani R, Yadav J, Yadav AK. Artificial intelligence in health care: a study on perceptions of and readiness for artificial intelligence in health-care professionals. J Mar Med Soc 2024 Dec;26(3):376-381. [doi: 10.4103/jmms.jmms_161_23]

Abbreviations

AI: artificial intelligence
GP: general practitioner
LLM: large language model

XSL•FO
RenderX

MCQ: multiple-choice question

Edited by AH Sapci; submitted 05.Apr.2025; peer-reviewed by GK Gupta, TK Tran, M Vejdanihemmat; comments to author 16.Aug.2025; revised version received 19.Jan.2026; accepted 20.Jan.2026; published 05.Feb.2026.

Please cite as:
Qunaibi EA, Al-Qaaneh AM, Ismail BF, Muhidat HI, Rageh FS, Musallam NA, Fawzy AK
Effectiveness of Informed AI Use on Clinical Competence of General Practitioners and Internists: Pre-Post Intervention Study
JMIR Med Educ 2026;12:e75534
URL: https://mededu.jmir.org/2026/1/e75534
doi:10.2196/75534
PMID:

©Eyad A Qunaibi, Ayman M Al-Qaaneh, Baraa F Ismail, Hussam I Muhidat, Farhia S Rageh, Najwa A Musallam, Alaa K Fawzy. Originally published in JMIR Medical Education (https://mededu.jmir.org), 05.Feb.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Education, is properly cited. The complete bibliographic information, a link to the original publication on https://mededu.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Construct Validity of Frontier Large Language Models on 4 Spanish Health Professions Specialty-Access Examinations (MIR, EIR, PIR, and FIR): Cross-Sectional Benchmarking and Psychometric Study

Clemente García-Hidalgo[1], MD, MSc; Manuel Carpio-Salmerón[2], MD; Pau Benito[3,4], MD; Carlos Carazo-Casas[5], MD; Beatriz Carratalá[6], LLB, MA; Jaime Campos Pavón[7,8], MD, PhD

[1]Radiology Department, Hospital General Universitario Morales Meseguer, Av Marqués de los Vélez S/N, Murciaº, Murcia, Spain

[2]Endocrinology Department, Santa Lucía University General Hospital, Cartagena, Murcia, Spain

[3]Department of Preventive Medicine and Epidemiology, Clinical Institute of Medicine and Dermatology (ICMiD), Hospital Clínic de Barcelona, Barcelona, Catalonia, Spain

[4]Barcelona Institute for Global Health (ISGlobal), Universitat de Barcelona, Hospital Clínic de Barcelona, Barcelona, Catalonia, Spain

[5]Department of Otolaryngology, Hospital Universitario Ramón y Cajal, Madrid, Madrid, Spain

[6]Innovation and Digital Projects Academic Department, Healthcademia, Madrid, Madrid, Spain

[7]Academia AMIR, Madrid, Madrid, Spain

[8]Department of Ophthalmology, Hospital Universitario, Madrid, Spain

Corresponding Author:
Clemente García-Hidalgo, MD, MSc
Radiology Department, Hospital General Universitario Morales Meseguer, Av Marqués de los Vélez S/N, Murciaº, Murcia, Spain

Abstract

Background: Large language models (LLMs) achieve high scores on health-professions examinations. Whether these scores support the same interpretations as human scores remains uncertain. Spain uses parallel multiple-choice examinations for entry to specialty training in medicine (Médico Interno Residente [MIR]), nursing (Enfermero Interno Residente [EIR]), psychology (Psicólogo Interno Residente [PIR]), and pharmacy (Farmacéutico Interno Residente [FIR]).

Objective: Using the January 2026 MIR, EIR, PIR, and FIR examinations, this study aimed to quantify the correspondence between human and LLM item facility and to describe model accuracy, distractor selection, and performance associated with measured surface features.

Methods: We evaluated 839 valid items using 12 configurations: GPT-5.2 at 3 reasoning efforts and Claude Opus, Sonnet, and Haiku 4.5 (Anthropic), each under neutral and expert prompts, with one run per configuration. The primary measure was the Spearman correlation (ρ) between human and machine facility on 828 complete paired items. We examined sensitivity to panel composition, scoring, pooling, and assumed reliability; compared distractor choices with a post hoc plausibility-weighted baseline; and evaluated a surface-feature classifier. Image analyses distinguished supplied from withheld images. An exploratory LLM-judge audit assessed generated explanations. Bootstrap CIs resample items or, for distractor concordance, error events; they exclude run-to-run uncertainty.

Results: The highest pooled accuracy was 94.8% (795/839); all 12 configurations answered 668 of 828 paired items correctly (80.7%). Scores clustered near the ceiling among the strongest configurations. Human and machine facility correlated at 0.40 (95% CI 0.34-0.45), with estimates of 0.31 to 0.40 across panel definitions and 0.39 (95% CI 0.33-0.45) after exam-stratified pooling. The observed correlation was lower than predicted by a specified one-axis simulation with a uniform model advantage (0.63; predictive interval 0.59 to 0.66; $P<.001$). Among 672 model errors, dominant human-distractor selection differed from the plausibility-weighted baseline by –1.25 percentage points (95% CI –4.53 to 2.19), compatible with no difference. Surface-feature accuracy was 34.4% (289/839), similar to selecting the longest option (34.6%, 290/839). The adjusted association with model correctness was uncertain (odds ratio [OR] 0.87, 95% CI 0.72-1.04). Image-associated accuracy was lower both when images were supplied (OR 0.33, 95% CI 0.21-0.53) and when withheld (OR 0.11, 95% CI 0.08-0.16).

Conclusions: In this single 2026 cohort, high LLM scores coexisted with limited separation among the strongest configurations and moderate correspondence with human item difficulty. These findings caution against interpreting model scores as equivalent evidence of human competence. They do not establish clinical performance or invalidate the examinations for human candidates.

(JMIR Med Educ 2026;12:e106199) doi:10.2196/106199

XSL•FO
RenderX

KEYWORDS

large language models; AI; medical education; educational measurement; clinical competence; psychometrics; construct validity; specialty-access examination; multiple-choice; Spain

Introduction

Background

Large language models (LLMs) are increasingly used in health-professions education. Reviews identify potential benefits, including personalized tutoring, virtual patients, formative feedback, and support for self-regulated learning. They also describe hallucinations, bias, factual errors, reduced critical thinking, over-reliance, and threats to academic integrity [1-3]. In a review of 33 studies, Pham et al [3] found that generative AI (GenAI) mainly supported individual activities, with little use in collaborative discussion. These tools may provide educational scaffolding [4-7], but high examination scores alone do not establish whether they support or replace the learning processes an examination is intended to assess.

LLM performance on medical multiple-choice questions (MCQs) has improved rapidly. ChatGPT (GPT-3.5; OpenAI) reached the United States Medical Licensing Examination (USMLE) passing threshold [8]; GPT-4 exceeded it by over 20 points and showed better calibration [9]. A meta-analysis of 45 studies across 17 countries reported pooled accuracy of 81% for GPT-4 and 58% for GPT-3.5, while concluding that GPT-4 was not yet suitable for unsupervised use [10]. Studies of Spain's Médico Interno Residente (MIR) examination have reported a similar progression, from approximately 51% for GPT-3.5 [11] to passing performance for GPT-4 [12,13] and subsequent evaluations of GPT-4o and o1 [14,15]. Other benchmarks assess LLM clinical reasoning using USMLE-style scenarios [16].

A high score and a valid interpretation of that score are separate claims. Construct-validity theory asks whether evidence supports the intended interpretation and use of test scores. Messick [17] identified 2 main threats: construct underrepresentation, when a test captures too little of the intended competence, and construct-irrelevant variance, when other factors affect scores. Kane's [18] approach examines a chain of inferences from scoring and generalization to extrapolation and decisions, with particular attention to the weakest links [19]. The Standards for Educational and Psychological Testing describes 5 sources of validity evidence and identifies fairness as a basic requirement [20,21]. Comparing score interpretations across populations also requires attention to measurement invariance and its item-level counterpart, differential item functioning (DIF) [22].

These examinations were designed for human candidates. Evidence supporting their use for human selection does not automatically support the same interpretation for an LLM, whose prior exposure to examination material and response processes may differ. High model accuracy therefore does not, by itself, establish professional competence.

Prior Work

Recent studies apply construct-validity theory to AI benchmarks. Raji et al [23] caution against treating benchmark scores as measures of general capability. In a review of 445 LLM benchmarks, Bean et al [24] found that approximately 47.8% assessed contested constructs and 53.4% provided any construct-validity evidence. Eriksson et al [25] identified weak construct validity among 9 recurring concerns, including contamination and saturation. Similar concerns have been raised for medical LLM benchmarks, where MCQ accuracy alone may not support the intended interpretation of performance [26].

Shortcut learning and data contamination may affect the meaning of benchmark scores. Shortcut learning uses superficial cues that predict answers within a dataset but may fail in other settings [27]. Examples include the Clever Hans problem in computer vision [28], pneumonia classifiers using hospital-specific image features [29], and similar concerns in COVID-19 imaging. A scoping review of 173 studies found that most did not test for shortcuts [30,31]. LLMs can answer MCQs above chance after the question stem is removed [32], and contamination studies report leakage as high as approximately 91.8% and lower scores after items are rewritten [33,34]. Correct answers may also draw on abilities different from those intended by an examination. About a third of medical question-answering items require multistep reasoning [35]; studies report difficulty recognizing unanswerable questions [36], failure on Einstellung vignettes that physicians solve [37], and explanations that do not faithfully reflect the process producing an answer [38].

Spain provides a common examination format across 4 health professions. The MIR examination ranks candidates for specialty training using approximately 200 anonymized MCQs with 4 options and 3 points for a correct answer and a penalty of 1 point for an incorrect answer. scoring, combined with an academic-record score [39]. Nursing (Enfermero Interno Residente [EIR]), psychology (Psicólogo Interno Residente [PIR]), and pharmacy (Farmacéutico Interno Residente [FIR]) use parallel examinations. This allows comparison across 4 knowledge domains within 1 selection system, although these examinations should not be treated as independent tests of different assessment designs. The 2026 Formación Sanitaria Especializada (FSE) call offered 12,366 places, including 9276 in medicine, with 34,553 candidates admitted [39,40]. The Ministry projects a specialist-workforce shortage approaching 9000 around 2027, particularly affecting an aging family and community medicine workforce, despite recurrent unfilled training places in that specialty [41]. High model examination scores do not address these workforce distribution problems. Spain's reliance on a single MCQ ranking examination also differs from assessment systems used elsewhere [42-47]. We did not evaluate non-Spanish items or test whether more varied assessment formats are more resistant to AI-assisted answering.

The literature leaves 3 questions for this study. First, do frontier LLMs show similar performance across medicine, nursing, psychology, and pharmacy specialty-access examinations? Second, do model scores support the same interpretation as human scores at the item level, considering difficulty, distractor

XSL•FO
RenderX

choices, surface cues, and the explanations models provide? Third, do these patterns occur across the 4 professions within a common examination format? The January 2026 examinations and their subsequently released answer keys provide a recent cohort for this comparison. Their timing does not establish that the tested models had no prior exposure to the items or to similar examination content.

Objectives

Using the January 2026 MIR, EIR, PIR, and FIR examinations, our primary objective was to quantify the correspondence between human item facility and the proportion of 12 LLM configurations answering each item correctly, using Spearman correlation (ρ) on complete paired items. Secondary objectives were to describe accuracy across the 839 valid items, compare model and human distractor choices, and assess whether measured surface features were associated with model correctness. Analyses of generated explanations, calibration, latency, image administration, and stylometry were exploratory. These outcomes inform an interpretation of examination scores within Kane validity framework; they do not constitute a formal test of measurement invariance or clinical competence.

Methods

Study Design and Validity Framework

We conducted a cross-sectional, retrospective study of frontier LLM performance on the 4 Spanish national specialty-access examinations in the 2026 FSE call: medicine (MIR), nursing (EIR), psychology (PIR), and pharmacy (FIR). Each examination uses negatively marked, 4-option MCQs. The examination score and academic record determine a national ranking for entry to specialty training [11,39]. We evaluated the interpretation of model scores using Messick construct-validity framework and Kane argument-based approach, as applied by Cook et al [17-20]. For model responses, we examined evidence concerning scoring, generalization, and extrapolation, together with construct-irrelevant variance and construct underrepresentation. Table 1 maps these inferences to the evidence and its limits.

Table . Evidence for interpreting model examination scores within Kane framework[a].

Kane inference	Evidence in this study	Scope of inference for models
Scoring	Exact agreement of the parsed option with the official key; incomplete responses examined separately.	Supported for recorded answers against the official key.
Generalization	Four professions within one examination design; related model configurations; one run each; 668/828 items at the panel ceiling.	Evidence limited to these items and configurations. Generalization to new occasions, prompts, and test forms remains uncertain.
Extrapolation	Moderate human-machine item-facility correspondence; generated explanations; no person-level invariance or clinical outcomes.	Clinical competence and equivalence with human score interpretations are untested.
Decision	High accuracy and limited separation among the strongest configurations; exploratory, unvalidated audit labels.	The study does not validate clinical decisions or autonomous educational use.

[a]These judgments describe the scope of the current evidence; they do not evaluate the examinations' validity for human candidates.

We organized the exploratory analyses into five areas: (1) agreement between human and machine item difficulty, (2) agreement in errors and distractor choices, (3) whether surface features predict the correct answer and are associated with model accuracy, (4) an LLM-judge audit of errors and explanations, and (5) the writing style of model explanations, reported as supporting analysis in Part S1 in Multimedia Appendix 1. We organized the exploratory analyses into five areas: (1) agreement between human and machine item difficulty, (2) agreement in errors and distractor choices, (3) prediction of the correct answer from surface features and the association of these features with model accuracy, (4) an LLM-judge audit of errors and explanations, and (5) the writing style of model explanations. The fifth area is reported as a supporting analysis in Part S1 in Multimedia Appendix 1. Interpretations concerning recall, reasoning, and potentially harmful answers remained provisional because they depended on the exploratory audit.

We distinguish expert agreement with the official key (n/N, 88% to n/N, 97%) from candidate facility, the proportion of examinees answering correctly. These measures have different denominators, defined in Results section. Comparisons with the average candidate use candidate facility only. Neither expert key agreement nor candidate or model MCQ accuracy is treated as a direct measure of clinical performance. Clinical competence is discussed as an untested extrapolation from examination performance.

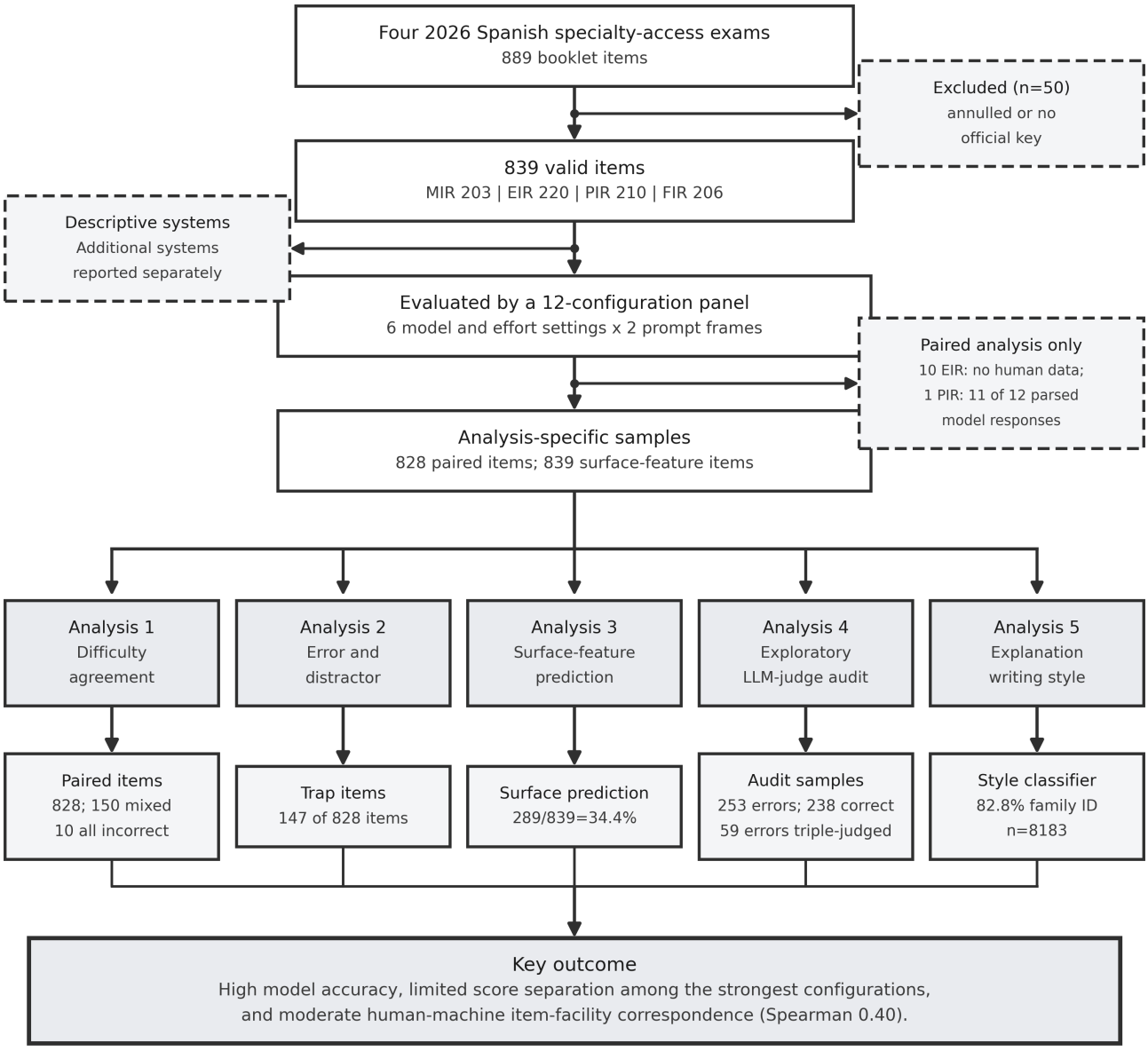
Data Sources and Item Construction

We used official question booklets and answer keys from the Ministerio de Sanidad. Items required an official key and could not be annulled, giving 839 valid items (MIR: n=203, EIR: n=220, PIR: n=210, and FIR: n=206). Image-associated items were flagged (MIR: n=24, EIR: n=20, PIR: n=0, and FIR: n=14). From aggregate candidate option distributions, we calculated human facility, normalized Shannon response entropy, the modal option, and the most frequently selected incorrect option. PIR facility was reconstructed from these distributions; we also analyzed MIR, EIR, and FIR separately as the examinations with published candidate facility. Human facility was available for 829 items, including all 210 PIR items. A total of 10 EIR items lacked usable candidate distributions. Claude Opus 4.5 (Anthropic) expert responses were unparsed on those same 10

XSL•FO
RenderX

EIR items and on one further PIR item. The primary analysis required human facility and all 12 model responses, leaving 828 complete paired items (MIR 203, EIR 210, PIR 209, and FIR 206). The observation-level generalized linear mixed model (GLMM) retained the additional PIR item with 11 parsed responses. Figure 1 shows these analysis samples.

Figure 1. Study design, item selection, and analysis samples for the 4 Spanish specialty-access examinations. EIR: Enfermero Interno Residente; FIR: Farmacéutico Interno Residente; MIR: Médico Interno Residente; PIR: Psicólogo Interno Residente.



Human item difficulty derived from per-item candidate response distributions (median 257 to 7334 responders per item).

AI Systems, Configurations, and Single-Run Design

The primary panel contained 12 configurations: GPT-5.2 at low, medium, and high reasoning effort and Claude Opus 4.5, Sonnet 4.5, and Haiku 4.5, each tested with neutral and expert prompt frames (6×2=12). GPT-5.2 was accessed through the OpenAI API. Machine facility was the proportion of panel configurations answering an item correctly, with unparsed responses handled as described in the sensitivity analysis. Google Gemini 3 Flash, an MIR-only Claude Opus 4.6 configuration with partial coverage, OpenEvidence, and a funder domain chatbot were evaluated descriptively and excluded from the primary panel and the first 4 validity analyses (see Management of Competing Interests section). Each configuration was run once, without repeated sampling or temperature variation, so run-to-run variance was not estimated. An automated pipeline processed text items and extracted the selected option, explanation, and latency. MIR image items were administered manually to multimodal models with the image. EIR and FIR image items went through the text-only pipeline without the image. Because image availability was confounded with examination, we analyzed these settings separately. An answer was correct only if the parsed option matched the official key. A verification prompt elicited certainty from 1 to 100. Model identifiers, query dates, prompts, decoding parameters, and rules for malformed or abstaining responses are provided in Part S1 in Multimedia Appendix 1.

XSL•FO
RenderX

Relationship Between Human and Machine Item Difficulty

We applied classical item analysis to the LLM configuration panel. For each item, machine facility was the proportion of configurations answering correctly, and machine discrimination was the point-biserial correlation between item correctness and overall accuracy across configurations. We calculated Kuder-Richardson Formula 20 (KR-20) for each examination. These statistics describe the 12-configuration panel and should not be interpreted as reliability estimates for human candidates. We measured the relationship between human and machine facility using Spearman ρ, overall and by examination, with bootstrap 95% CIs. Squared Pearson correlation (r) was reported only as a descriptive shared-variance index. To examine the effect of near-ceiling model accuracy, we used 3 additional analyses: Pearson r corrected for assumed measurement reliability; machine facility across deciles of human facility, summarized with an ordinary least-squares slope rather than a fitted DIF model; and a logistic GLMM of response correctness with standardized human facility (per SD) and an image indicator as predictors and item- and model-level random intercepts. The GLMM included 829 items with defined human facility: 828 contributed 12 configurations and 1 contributed 11, yielding 828×12+11=9947 observations.

Machine facility is a coarse proportion based on 12 configurations, with per-item SE up to approximately 0.15. It was also near the ceiling: all configurations answered 668 of 828 items correctly (80.7%). These features can reduce the observed correlation with human facility. We assessed this explanation using a simulation with one shared latent item-difficulty axis. In the one-parameter logistic (1PL) model, human facility determined item difficulty, while the AI panel had a uniform ability advantage of Δ logits and performance was constrained by the empirical ceiling. We calibrated Δ to match observed mean machine facility, not the observed correlation. Across 2000 simulations, we compared the observed Spearman ρ and extreme-cell counts with the model's predictive distribution. We repeated the simulation using 2PL and 3PL parameterizations and a±1-logit range of Δ. This analysis tests the specified shared-difficulty, uniform-advantage model; it cannot exclude all unidimensional explanations. Full details and sensitivity analyses are mentioned in Part S1 in Multimedia Appendix 1.

Sensitivity Analyses for Human-Machine Difficulty Agreement

We performed 4 sensitivity analyses using existing item-level data, without new model queries. We recalculated the human-machine correlation and the proportion of items answered correctly by every configuration under 7 panel definitions, listed in Results section. We compared 3 rules for unparsed responses: exclusion from the machine-facility denominator, scoring as incorrect, and complete-case analysis. We estimated correlations within each examination with bootstrap CIs, then combined them using a fixed-effect Fisher-z meta-analysis with Q and I^2. Finally, we varied assumed human-facility reliability from 0.90 to 0.98 and machine-facility reliability from 0.80 to 0.95 when correcting correlations for measurement error.

Error and Distractor Concordance Between Models and Humans

We defined a trap item as one on which the most frequently selected human option was incorrect. We compared machine facility on trap and nontrap items using bootstrap CIs. Across the 828 paired items, we cross-tabulated trap status against whether an item was difficult for humans but answered correctly by every model configuration. We reported the phi coefficient and human and machine facility in each of the 4 cells to assess whether these definitions identified the same items. Among model errors, we calculated the probability of selecting the human top distractor or the most frequently selected human option, and the mean within-item Pearson r between human and model distributions across the 3 incorrect options. We also calculated the Spearman ρ between human response entropy and model certainty. Because incorrect human responses were not evenly distributed across options, we supplemented the uniform-among-wrong baseline of 0.33 with a plausibility-weighted baseline based on the human-choice shares of incorrect options. This post hoc analysis was adopted as the primary comparison for error concordance. We reported observed concordance minus that baseline with a 2000-replicate bootstrap 95% CI resampling individual error events. This interval does not account for dependence among errors from the same item or related configurations. The derivation is mentioned in Part S1 in Multimedia Appendix 1.

Surface Features and Correct-Answer Prediction

We tested whether predefined surface features could predict the correct option without requiring medical knowledge, an MCQ analog of shortcut learning [27,28,32]. An L2-regularized logistic classifier used 11 features, including option length, position, absolute terms, negation, numerals, all or none-of-the-above patterns, and lexical overlap with the stem. The overlap feature used words from the stem, so the classifier was not entirely independent of stem-derived information. We used leave-one-examination-out cross-validation and reported item-level accuracy overall and by examination, with a bootstrap 95% CI (n=839). All options from each item remained within the same training or test partition. We compared performance with a modal-position baseline and a longest-option rule, and also used within-examination cross-validation. To assess whether these cues were associated with model performance, we compared full-coverage model accuracy between items classified as shortcut-solvable and shortcut-resistant, adjusting for human facility because shortcut-solvable items were also easier for humans. The full feature list is mentioned in Part S1 in Multimedia Appendix 1.

Exploratory Audit of Errors and Model Explanations

LLM judges applied a human-authored rubric to model explanations, returning labels in a fixed JSON schema. This exploratory audit assessed the generated text, not internal model computation. The labels were unvalidated judgments from the same model family. The rubric and expert prompt were written for a physician examiner despite the 4-profession sample;

XSL•FO
RenderX

clinician adjudication would be required before drawing safety conclusions. We performed 2 audits. First, we assigned 1 primary failure mode from 8 categories to each frontier wrong answer with a usable explanation (n=253; Claude Opus, Sonnet, and Haiku 4.5 and GPT-5.2 high, neutral). An answer was flagged as potentially clinically dangerous if following it instead of the official answer could plausibly harm a patient. Second, a stratified sample of correct GPT-5.2 high answers was labeled for faithfulness, dominant expressed reasoning mode, and false statements (240 sampled, 238 labeled, and approximately 60 per examination). Analyses used the 238 labeled answers. We assessed agreement separately using 3-judge Fleiss kappa (κ) in 59 MIR and EIR answers (κ=0.76; 2-of-3 agreement 0.97) and Opus-versus-Sonnet Cohen κ for each label on jointly labeled responses (231 errors and 236 correct-answer rationales). The rubric is mentioned in Part S3 in Multimedia Appendix 1; agreement details and limitations are mentioned in Part S1 in Multimedia Appendix 1. Judge prompts and the rubric are deposited as described in Data Availability section.

Writing Style of Model Explanations (Supporting Analysis)

We assembled 12,043 AI explanations and 839 human examination stems to describe writing style and trained a random-forest classifier to identify the generating model family from stylistic features. The exploratory analysis is detailed in Part S1 in Multimedia Appendix 1, with one summary panel in the main text. It does not establish whether any official examination item was written by AI.

Certainty, Latency, Image Availability, and Prompt Frames

We assessed stated certainty (1 to 100) using the area under the receiver operating characteristic curve (AUROC), mean certainty minus accuracy to describe net over- or under-confidence, and expected calibration error (ECE). We recorded median response latency for each configuration. Image associations were estimated in 2 regression models: ordinary logistic regression adjusted for human facility, yielding an odds ratio without item or configuration random effects, and the GLMM described above, yielding a conditional β (image) coefficient with item- and model-level random intercepts. These estimates have different interpretations and are reported separately. Because images were supplied for MIR but withheld for EIR and FIR, we fitted the ordinary logistic model separately for these settings and for the pooled data. Repeating queries with different seeds or temperatures was not feasible in the study environment. We therefore compared the panel-level human-machine correlation and per-model accuracy between neutral and expert prompt frames as a limited measure of prompt sensitivity. This comparison does not estimate run-to-run variation at a fixed prompt.

Statistical Analysis and Single-Run Uncertainty

We used Python (Python Software Foundation) with statsmodels, scikit-learn, and SciPy. Accuracy is reported as a percentage with its count, and principal estimates are accompanied by bootstrap or binomial 95% CIs. *P* values are descriptive and unadjusted for multiple comparisons. Rank associations use Spearman ρ with item-bootstrap CIs; squared Pearson *r* values are descriptive shared-variance indices. Item-level associations use logistic regression, including McFadden pseudo-R^2, and a GLMM with item- and model-level random intercepts. We interpret nonsignificant contrasts through their CIs rather than as evidence of no effect. KR-20 (0.86 to 0.95) and point-biserial discrimination describe the 12-configuration panel. Each configuration was evaluated once, so KR-20, discrimination, machine facility, the 80.7% ceiling, and ρ=0.40 are conditional on those runs. Most bootstrap CIs resample items; the distractor-concordance interval resamples individual error events and does not account for within-item or configuration dependence. Neither approach includes uncertainty from repeated model responses. They therefore do not capture total uncertainty.

Management of Competing Interests

The funder (Healthcademia) operates one evaluated domain chatbot. That chatbot is excluded from the 12-configuration headline panel, from all comparative ranking, and from construct-validity pillars 1 to 4; it appears only as a separate, nonranked family in the descriptive stylometric corpus (pillar 5), and its accuracy is reported separately for transparency (Results section): pooled 80.4%; per-examination MIR 86.6%, EIR 85.9%, PIR 74.0%, and FIR 75.4%.

Ethical Considerations

This work was a secondary analysis of publicly released examination materials and deidentified aggregate candidate statistics. No human participants were recruited or enrolled, no individually identifiable data were accessed, and no model was fine-tuned; institutional review board approval or exemption was therefore not required under institutional and local policy. No participant compensation was provided. The official examination items are the property of the Ministerio de Sanidad and are publicly released after each call, we reference them to that public source rather than redistributing them.

Results

Item Flow

Figure 1 summarizes item selection and the separate analysis samples. Of 889 booklet items, 839 had a valid official key and were not annulled. Human facility was available for 829 items; the 10 missing distributions were all from EIR. One further item, from PIR, had only 11 parsed model responses, leaving 828 items with human facility and all 12 model responses for the primary paired analysis. Of these, 668 of 828 (80.7%) were answered correctly by every configuration and 160 of 828 (19.3%) had at least one error. The latter comprised 150 items with both correct and incorrect responses and 10 missed by every configuration. The audit separately assessed 253 incorrect responses and 238 correct-answer rationales; a 59-error MIR and EIR subset underwent triple-judge assessment.

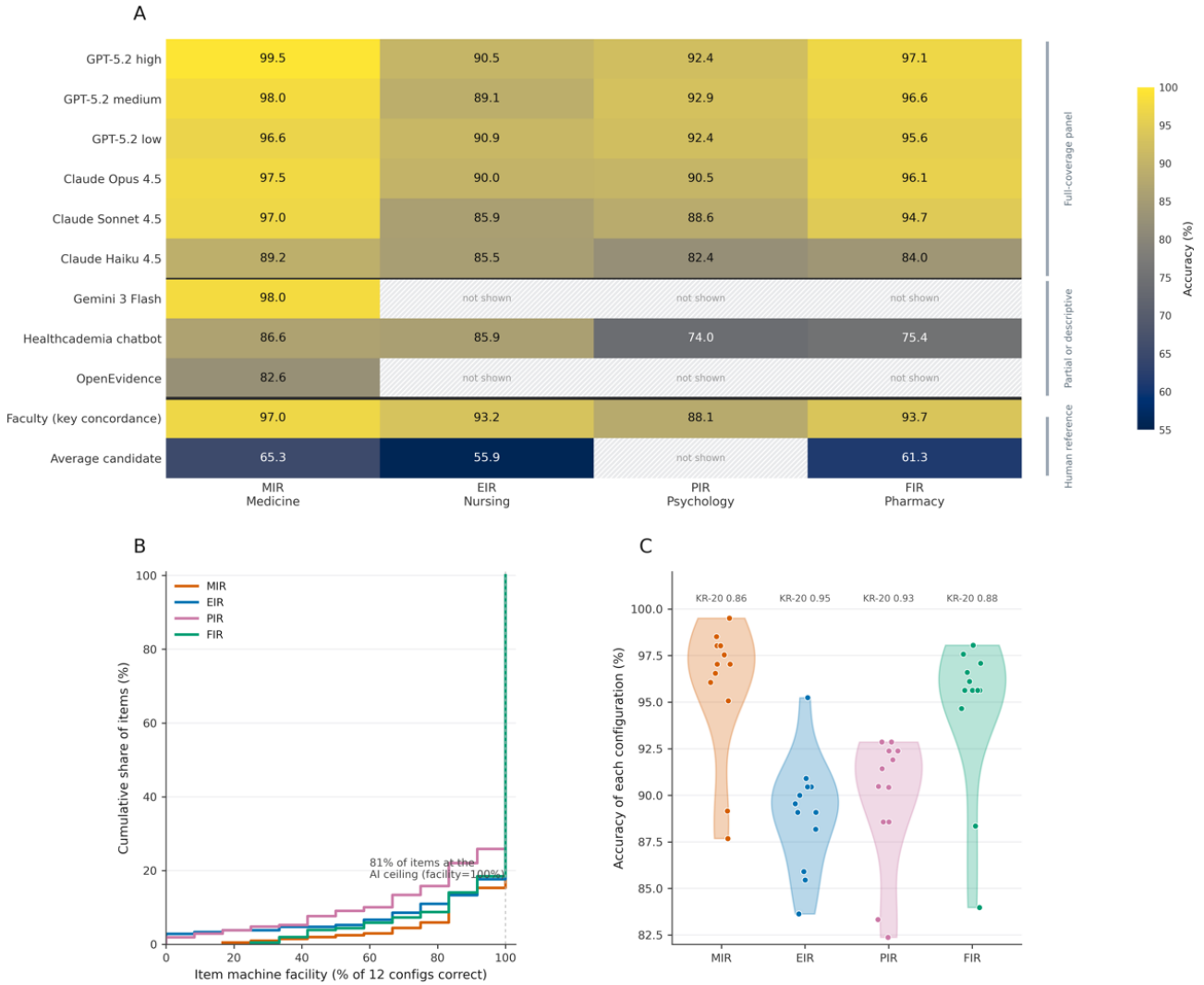
Model Accuracy and Discrimination Across the 4 Examinations

Across the 4 2026 Spanish specialty-access examinations (839 valid items: MIR: n=203, EIR: n=220, PIR: n=210, and FIR:

XSL•FO
RenderX

n=206), frontier models reached near-ceiling accuracy on every examination, and not on medicine alone (Figure 2). Pooling under full-coverage configurations (Table 2), the strongest configuration was GPT-5.2 high (neutral) at 94.8% (795/839, 95% CI 93.0%-96.1%), with Claude Opus 4.5 (expert) at 94.3% (781/828; 11 items [EIR: n=10, and PIR: n=1] had unparsed expert-frame responses and were excluded; on the common-coverage intersection GPT-5.2 high leads 95.9% (794/828) vs 94.3% (781/828), so this does not affect the ranking). Among configurations with responses on every valid item in an examination, the maximum, selected across configurations and therefore optimistic relative to a typical model, was 99.5% (202/203) for MIR, 90.9% for (200/220) for EIR, 92.9% (195/210) for PIR, and 98.1% (202/206) for FIR.

Figure 2. Model accuracy across the 4 examinations. (A) Per-examination accuracy for neutral-frame configurations and human reference quantities; partial or descriptive systems are separated from the main panel. Faculty agreement with the official key and candidate facility are distinct measures. (B) Distribution of machine facility; all 12 configurations answered 668 of 828 (80.7%) paired items correctly. (C) Configuration accuracies by examination, with descriptive panel KR-20. Scores cluster near the ceiling among the strongest configurations; this does not measure discrimination among human candidates.



XSL•FO
RenderX

Table . Accuracy across the four 2026 Spanish specialty-access examinations[a].

Configuration	Pooled % (95% CI)	Correct answers (n/N)	MIR, %	EIR, %	PIR, %	FIR, %
Full-coverage frontier models: GPT-5.2						
GPT-5.2 high (neutral)	94.8 (93.0-96.1)	795/839	99.5[b]	90.5[c]	92.4[d]	97.1[e]
GPT-5.2 high (expert)	94.3 (92.5-95.7)	791/839	98.5	89.5	91.4	98.1
GPT-5.2 medium (neutral)	94.0 (92.2-95.5)	789/839	98.0	89.1	92.9	96.6
GPT-5.2 medium (expert)	94.2 (92.4-95.6)	790/839	97.0	90.5	91.9	97.6
GPT-5.2 low (neutral)	93.8 (92.0-95.2)	787/839	96.6	90.9	92.4	95.6
GPT-5.2 low (expert)	93.8 (92.0-95.2)	787/839	98.0	89.1	92.9	95.6
Full-coverage frontier models: Claude 4.5						
Claude Opus 4.5 (neutral)	93.4 (91.6-94.9)	784/839	97.5	90.0	90.5	96.1
Claude Opus 4.5 (expert)	94.3 (92.5-95.7)	781/828	96.1	95.2[f]	90.4[g]	95.6
Claude Sonnet 4.5 (neutral)	91.4 (89.3-93.1)	767/839	97.0	85.9	88.6	94.7
Claude Sonnet 4.5 (expert)	91.8 (89.7-93.5)	770/839	95.1	88.2	88.6	95.6
Claude Haiku 4.5 (neutral)	85.2 (82.7-87.5)	715/839	89.2	85.5	82.4	84.0
Claude Haiku 4.5 (expert)	85.7 (83.2-87.9)	719/839	87.7	83.6	83.3	88.3
Partial-coverage and descriptive (not in headline panel)						
Gemini 3 Flash (neutral)	—[h]	—	98.0	80.0[i]	—	92.9[j]
Healthcademia chatbot	80.4 (77.5-83.0)	645/802	86.6	85.9[k]	74.0[l]	75.4[m]
OpenEvidence (MIR only)	—	—	82.6[n]	—	—	—
Six-model ensemble (neutral)						
Majority vote	—	—	98.0	91.4	92.9	96.1
Oracle (any correct)	—	—	100.0	92.3	97.6	100.0
Items all-6-wrong (n)	—	—	0	17	5	0
Human reference						
Faculty (key-concordance), n/N (%)	—	—	97.0[o]	93.2	88.1[p]	93.7[q]
Average candidate, mean	—	—	65.3	55.9	—	61.3

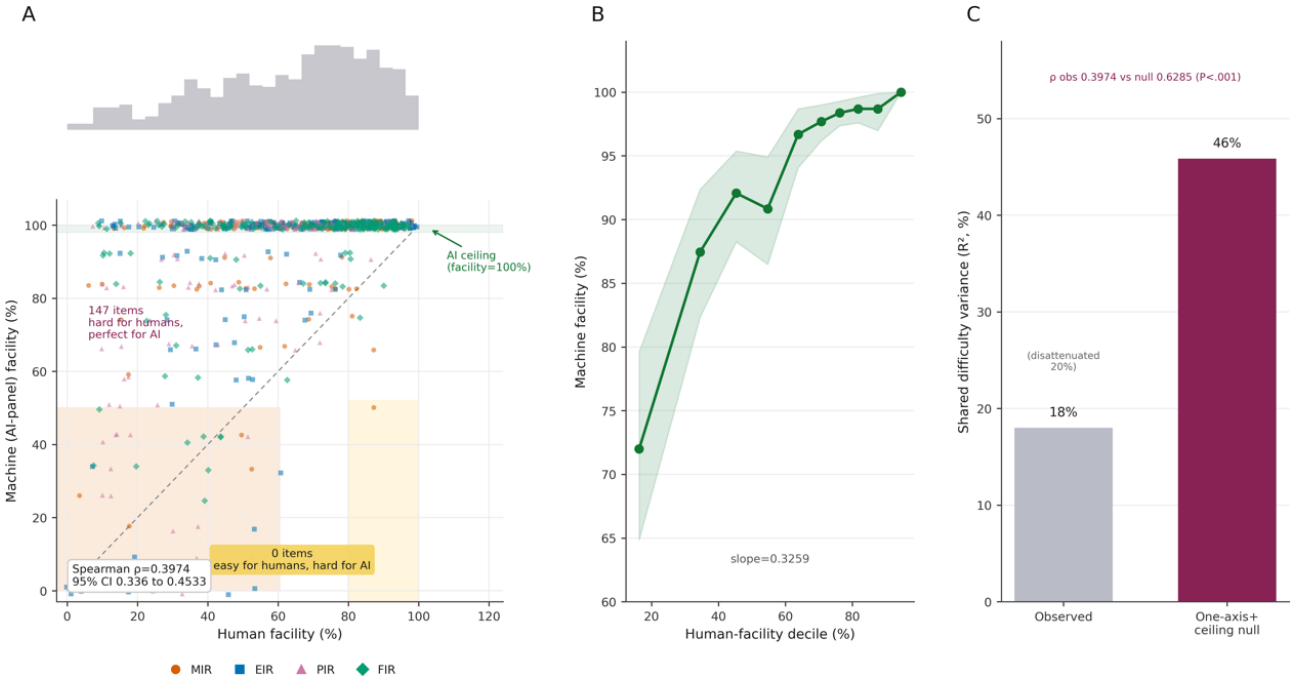
[a]Percentages use the available responses for each listed configuration; the pooled estimate includes its numerator, denominator, and binomial 95% CI. Claude Opus 4.5 expert has 828 parsed responses (Médico Interno Residente [MIR]: n=203, Enfermero Interno Residente [EIR]: n=210, Psicólogo Interno Residente [PIR]: n=209, and Farmacéutico Interno Residente [FIR]: n=206), whereas other headline configurations have 839. Faculty agreement with the official key is not an independent reference standard. Average candidate facility includes blanks in its denominator; the difficulty analysis uses correct/answered. Maxima selected across configurations are optimistic estimates of typical performance.

XSL•FO
RenderX

[b]202 out of 203.
[c]199 out of 220.
[d]194 out of 210.
[e]200 out of 206.
[f]200 out of 210.
[g]189 out of 209.
[h]Not applicable.
[i]8 out of 10.
[j]13 out of 14.
[k]170 out of 198.
[l]151 out of 204.
[m]150 out of 199.
[n]142 out of 172.
[o]197 out of 203.
[p]185 out of 210.
[q]193 out of 206.

We report 2 distinct human quantities separately. Faculty-key concordance was 97% (197/203) for MIR, 93.2% (n/N) for EIR, 88.1% (185/210) for PIR, and 93.7% (193/206) for FIR (denominators MIR 203, EIR 220, PIR 210, and FIR 206 items). Human candidate facility is reported under 2 denominators; the primary one, and the x-axis of the difficulty-transfer analysis, is correct/answered, which excludes candidate blanks and is therefore comparable to machines, which never abstain. On the 147 trap items, mean human item facility was 24.4% (SD 11.1 percentage points, 95% CI 22.7%-26.2%), whereas mean machine facility was 79.5% (SD 30.1 percentage points, 95% CI 74.7%-84.5%). Mean machine facility on the 681 nontrap items was 96.2% (SD 12.8 percentage points; 95% CI 95.2%-97.1%; Figure 3). The difficulty-transfer Spearman is essentially unchanged under either denominator (ρ=0.357 correct/answered vs ρ=0.360 correct/all; n=619, MIR, EIR, and FIR). Because faculty concordance is scored against the same official key used to score models, model-versus-faculty differences are not independent comparisons; this caveat carries into every faculty-concordance comparison below. The funder chatbot, excluded from all comparative ranking and from the first 4 validity analyses, scored pooled 80.4% (per-examination, MIR: 86.6%, EIR: 85.9%, PIR: 74.0%, and FIR: 75.4%).

Figure 3. Distractor selection and surface features. (A) Machine facility on trap items (human plurality selects an incorrect option, n=147) and other items. The marker denotes mean human facility on traps (24%). (B) Dominant human-distractor selection under a uniform baseline (33%), 672 observed model errors (63.7%), and the post hoc human-plausibility baseline (64.9%); observed-minus-baseline difference –1.25 percentage points (error-event bootstrap 95% CI −4.53 to 2.19), compatible with no difference. (C) Option-level receiver operating curves for the surface-feature classifier by held-out examination (area under the curve 0.55 to 0.62). (D) Model accuracy by human facility and whether the classifier selected the correct answer. The unadjusted difference is 1.5 percentage points; the adjusted odds ratio is 0.87 (95% CI 0.72-1.04). Bands are 95% CIs. AUC: area under the curve. EIR: Enfermero Interno Residente; FIR: Farmacéutico Interno Residente; MIR: Médico Interno Residente; PIR: Psicólogo Interno Residente.



XSL•FO
RenderX

Scores clustered near the ceiling among the strongest configurations. The 6-model neutral-prompt ensemble reached majority-vote accuracies of 98% on MIR (199/203), 91.4% on EIR (201/220), 92.9% on PIR (195/210), and 96.1% on FIR (198/206). At least one ensemble member answered every MIR and FIR item correctly; all 6 missed 17 EIR items and 5 PIR items. Panel KR-20 values were 0.86 for MIR, 0.95 for EIR, 0.93 for PIR, and 0.88 for FIR. These are descriptive statistics across configurations. The observed scores offered limited separation among the strongest configurations, while weaker configurations such as Claude Haiku scored lower [23,25].

Relationship Between Human and Machine Item Difficulty

Across 828 complete paired items, human and machine facility had a Spearman ρ of 0.40 (95% CI 0.34-0.45) and Pearson *r* of 0.42 (squared correlation approximately 0.18). Per-examination Spearman ρ values were 0.30 for MIR, 0.38 for EIR, 0.49 for PIR, and 0.38 for FIR (Table 3; Figure 4). On the 619 MIR, EIR, and FIR items with published candidate facility, the same 12-configuration panel gave correlations of 0.357 using correct/answered and 0.360 using correct/all candidates. This comparison avoids relying on reconstructed PIR facility. In the full paired corpus, 147 items had human facility below 50% (n/N) and perfect machine facility; none had human facility at least 80% (n/N) and machine facility below 50% (n/N). A further threshold comparison identified 116 items with human facility below 40% (n/N) and machine facility at least 80% (n/N).

XSL•FO
RenderX

Table . Human and machine item facility, sensitivity analyses, and simulation[a].

Measure	Estimate	Items or observations
MIR[b] facility: human; machine, n/N (%)	67.7; 2335/2436 (95.9)	203
EIR[c] facility: human; machine, n/N (%)	61.7; 2340/2520 (92.9)	210
PIR[d] facility: human; machine, n/N (%)	55.2; 2252/2508 (89.8)	209
FIR[e] facility: human; machine, n/N (%)	65.6; 2338/2472 (94.6)	206
Pooled facility: human; machine, n/N (%)	62.5; 9265/9936 (93.2)	828
Panel KR-20:[f] MIR; EIR; PIR; FIR	0.863; 0.953; 0.932; 0.876	12 configurations
Panel discrimination: MIR; EIR; PIR; FIR	0.57; 0.38; 0.34; 0.55	12 configurations
Pooled panel discrimination	0.45	12 configurations
MIR Spearman ρ (95% CI)	0.30 (0.17-0.42)	203
EIR Spearman ρ (95% CI)	0.38 (0.26-0.49)	210
PIR Spearman ρ (95% CI)	0.49 (0.38-0.59)	209
FIR Spearman ρ (95% CI)	0.38 (0.26-0.49)	206
Pooled Spearman ρ (95% CI)	0.40 (0.34-0.45)	828
Pearson r; squared Pearson *r*	0.42; 0.18	828
Disattenuated squared Pearson *r*	Approximately 0.20; sensitivity range 0.193 to 0.250	Assumed reliabilities
Correct under all configurations	668 (80.7%)	828
Three-examination Spearman ρ: correct/answered; correct/all candidates	0.357; 0.360	619
Mixed correct and incorrect responses: Spearman ρ	0.28	150
At least one error, including all-wrong items: Spearman ρ	0.34	160
Human facility <50%; machine facility 100%	147	828
Human facility ≥80%; machine facility <50%	0	828
Human facility <40%; machine facility ≥80%	116	828
Specified simulation: expected Spearman ρ	0.63 (95% predictive interval 0.59 to 0.66); observed 0.40; *P*<.001	2000 simulations
Specified simulation: human-hard/model-perfect count	Expected 51 (40 to 63); observed 147	2000 simulations
GLMM[g]: standardized human-facility coefficient; image coefficient	2.8368 per SD; –4.3495	9947 observations from 829 items

[a]Human facility is correct/answered. Machine facility is the share of 12 configurations answering correctly, primary analyses require complete coverage. PIR has human facility for 210 items and complete paired data for 209. Panel Kuder-Richardson Formula 20 and discrimination are descriptive configuration-panel statistics, not person-level test reliability. Bootstrap CIs condition on the observed single runs.

[b]MIR: Médico Interno Residente.

[c]EIR: Enfermero Interno Residente.

[d]PIR: Psicólogo Interno Residente.

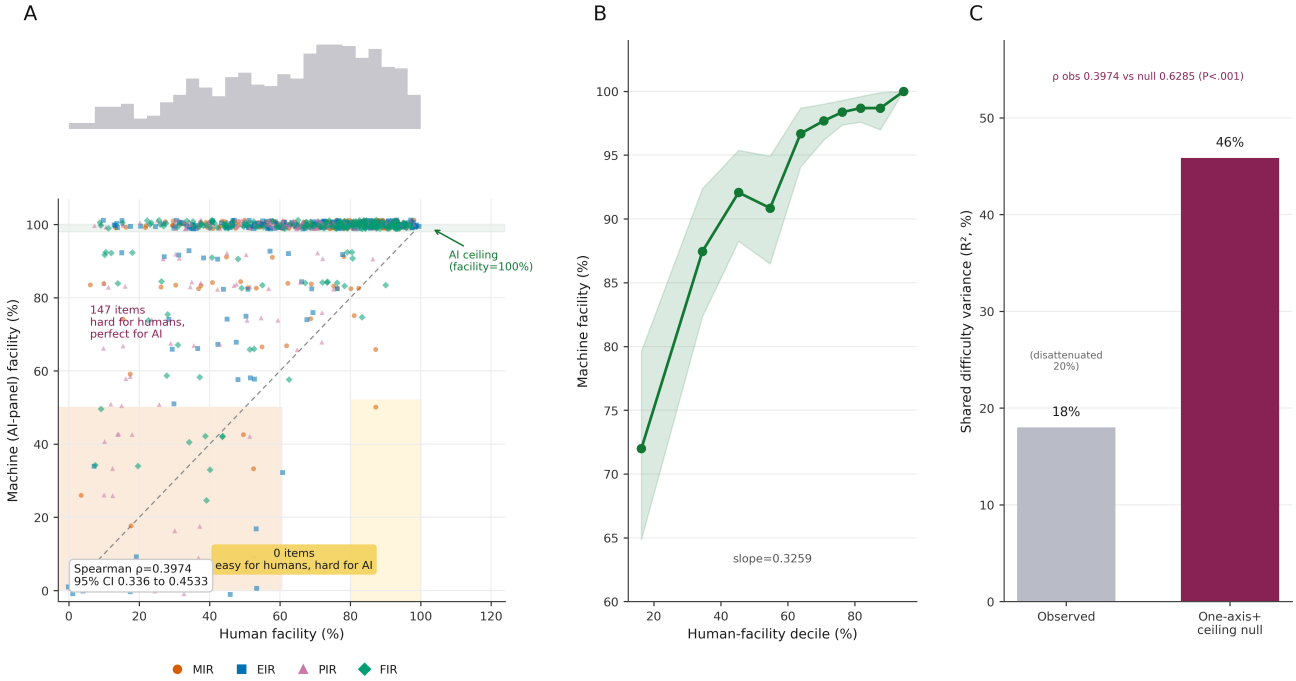
[e]FIR: Farmacéutico Interno Residente.

[f]KR 20: Kuder-Richardson Formula 20.

[g]GLMM: generalized linear mixed model.

XSL•FO
RenderX

Figure 4. Correspondence between human and machine item facility. (A) The 828 complete paired items, with Spearman ρ=0.40 (95% CI 0.34-0.45). There were 147 items with human facility below 50% (n/N) and perfect machine facility, and none with human facility at least 80% (n/N) and machine facility below 50% (n/N). (B) Mean machine facility by human-facility decile (slope 0.33), with bootstrap 95% CIs. (C) Squared Pearson correlation (r): observed 18%, compared with 46% under the specified uniform-advantage simulation. Disattenuated shared variance is approximately 20% and varies from 19.3% to 25.0% across assumed reliabilities. The simulation result does not exclude all unidimensional models.



The asymmetric counts partly reflect the difference in overall accuracy: 668 of 828 (80.7%) paired items were at the model ceiling. Disattenuation with assumed machine and human reliabilities of 0.93 and 0.98 gave a Pearson r of 0.45 and shared variance of approximately 20% (Part S1 in Multimedia Appendix 1); these values depend on the reliability assumptions. Spearman ρ was 0.28 on the 150 items with both correct and incorrect model responses and 0.34 on all 160 items with at least one error. Mean machine facility increased from 0.72 (SD 0.34) in the hardest human-facility decile to 1.00 (SD 0.00) in the easiest (ordinary least squares [OLS] slope 0.33). The GLMM estimated coefficients of 2.84 per SD of human facility and −4.35 for the image indicator. Its 9947 observations comprised 828 items with 12 responses and one additional item with 11 responses. The small model-level random-intercept variance describes variation within this panel and does not establish equivalence between configurations.

We then tested whether the ceiling alone manufactures the weak coupling, using the ceiling-aware generative simulation (Part S1 in Multimedia Appendix 1). Under the specified one-axis-plus-uniform-advantage-plus-ceiling null, the expected Spearman ρ is 0.63 (95% predictive interval 0.59 to 0.66) and the expected hard-for-human and perfect-for-AI count is 51 (40 to 63). The observed Spearman ρ of 0.40 falls below this null (P<.001; 0/2000 simulations ≤ observed), and the observed 147 hard-for-human and perfect-for-AI items exceed the one-axis prediction roughly threefold. The verdict is robust across 1PL, 2PL (a=1.7), and 3PL (guessing floor c=0.25, with and without a=1.7) parameterizations and a±1-logit Δ sweep (Part S1 in Multimedia Appendix 1). By contrast, the empty reverse cell (0 easy-for-human and hard-for-AI items) is itself reproduced by the one-axis-plus-ceiling model, so the reverse-cell asymmetry is a ceiling feature and is not independent evidence of divergence. The one-axis-implied shared variance was R^2~0.46, against the observed~0.18 (Figure 4C). The claim licensed is divergence from the specified uniform-advantage generative model, not exclusion of every unidimensional account of these data.

The observed item-difficulty correspondence was lower than this uniform-advantage model predicted. The result remains conditional on a single cohort, one run per configuration, and a coarse machine-facility measure with substantial ceiling effects. No person-level IRT or formal invariance analysis was performed, and the simulation does not establish 2 separate latent constructs.

Sensitivity Analyses for Human-Machine Difficulty Agreement

Spearman ρ remained moderate under 7 panel definitions: 0.40 for all 12 configurations (n=828), 0.40 for neutral prompts (n=829, 6 configurations), 0.37 for expert prompts (n=828, 6 configurations), 0.39 for Claude configurations (n=828), 0.31 for GPT configurations (n=829), 0.40 for 3 Claude models plus one GPT configuration (n=829), and 0.32 for the best configuration from each vendor (n=829). The ceiling proportion ranged from 80.7% (n/N) to 92.9% (n/N). These descriptive checks show how the estimate varies with panel composition; the simulation reference of 0.63 applies to the primary panel.

Correlations were similar under alternative handling of the 11 unparsed Claude Opus 4.5 expert responses (10 EIR and 1 PIR). Among the 829 items with human facility, excluding an unparsed response from its denominator gave Spearman ρ=0.398, scoring it as incorrect gave ρ=0.398, and requiring complete model coverage gave ρ=0.397 (n=828). The 10

XSL•FO
RenderX

affected EIR items also lacked human facility, so only the PIR item changed the paired-analysis sample.

Within-examination correlations were 0.30 for MIR (95% CI 0.17-0.42; n=203), 0.38 for EIR (95% CI 0.26-0.49; n=210), 0.49 for PIR (95% CI 0.38-0.59; n=209), and 0.38 for FIR (95% CI 0.26-0.49; n=206). Fixed-effect Fisher-z pooling gave ρ=0.39 (95% CI 0.33-0.45; Q=5.10, I^2=41%), close to the pooled item-level estimate of 0.40. We did not interpret the between-examination differences as established heterogeneity given the single-run design.

Across the assumed human-reliability range of 0.90 to 0.98 and machine-reliability range of 0.80 to 0.95, disattenuated shared variance was 19.3% to 25.0%. This is an assumption-dependent sensitivity analysis: human reliability was not estimated from these data, and panel KR-20 is not conventional human test reliability.

Error and Distractor Concordance Between Models and Humans

We identified 147 trap items (147/828, 17.8% of the paired corpus; per-examination: MIR: n=22, EIR: n=35, PIR: n=54, and FIR: n=36) on which the human plurality selected a wrong option. On trap items, human accuracy was 24.4% (95% CI 22.7%-26.2%) versus AI 79.5% (95% CI 74.7-84.5), against nontrap 96.2% (95% CI 95.2-97.1; Figure 3). Models therefore answered many items correctly even when the modal human answer was incorrect.

The 147 trap items and the 147 items hard for humans but perfect for the AI panel overlapped in 77 items (phi=0.42). There were 70 trap-only items, 70 hard-for-human and perfect-for-AI-only items, and 611 items in neither group. Mean human and machine facility were respectively 0.28 (SD 0.10) and 1.00 (SD 0.00) in the overlap, 0.21 (SD 0.11) and 0.57 (SD 0.31) in the trap-only group, 0.43 (SD 0.05) and 1.00 (SD 0.00) in the hard-for-human and perfect-for-AI-only group, and 0.74 (SD 0.14) and 0.96 (SD 0.13) in the remaining items. The definitions therefore identify related but distinct item sets.

Across 672 incorrect responses from the declared 12-configuration panel, models selected the dominant human distractor with probability 0.6369, compared with 0.6494 under the post hoc plausibility-weighted baseline. The difference was −1.25 percentage points (error-event bootstrap 95% CI −4.53 to 2.19), compatible with no difference. This does not establish either shared misconceptions or systematic divergence beyond option plausibility. The probability of selecting the human modal option was 0.3661. The mean within-item Pearson *r* between human and model distributions over incorrect options was 0.50 (SD 0.61) across 161 items with defined correlations. Human response entropy and model confidence had a descriptive Spearman ρ of −0.30 (n=9851); mean confidence was 93.57 (SD 10.01) at entropy below 0.3 and 87.37 (SD 13.28) at entropy above 0.7.

Surface Features and Correct-Answer Prediction

A surface-feature classifier with no medical knowledge, cross-validated leave-one-examination-out, recovered the keyed answer at 34.4% (289/839, 95% CI 31.3-37.7) against a 25% chance baseline (Figure 3). The signal is, however, almost entirely a single item-writing artifact: a longest-option heuristic applied alone, with no classifier at all, recovers the key at 34.6% (290/839), matching or exceeding the classifier. The same holds per examination, where classifier and longest-option accuracies are near-identical (MIR n, 39.9% vs n, 39.4%; EIR n, 32.7% vs n, 32.7%; PIR n, 32.4% vs n, 31.9%; FIR n, 33.0% vs n, 34.5%; denominators MIR 203, EIR 220, PIR 210, and FIR 206). We therefore describe this narrowly, as a length-dominated item-writing artifact rather than a broad surface-leakage construct: the classifier's leading coefficient is option length (len_z=0.287; longer options more often correct), with weaker cues (has_absolute=−0.509, is_allnone=0.47, and stem_overlap=−0.34; Part S1 in Multimedia Appendix 1), and a within-examination cross-validated solvability was comparable, indicating the signal is item-format-intrinsic.

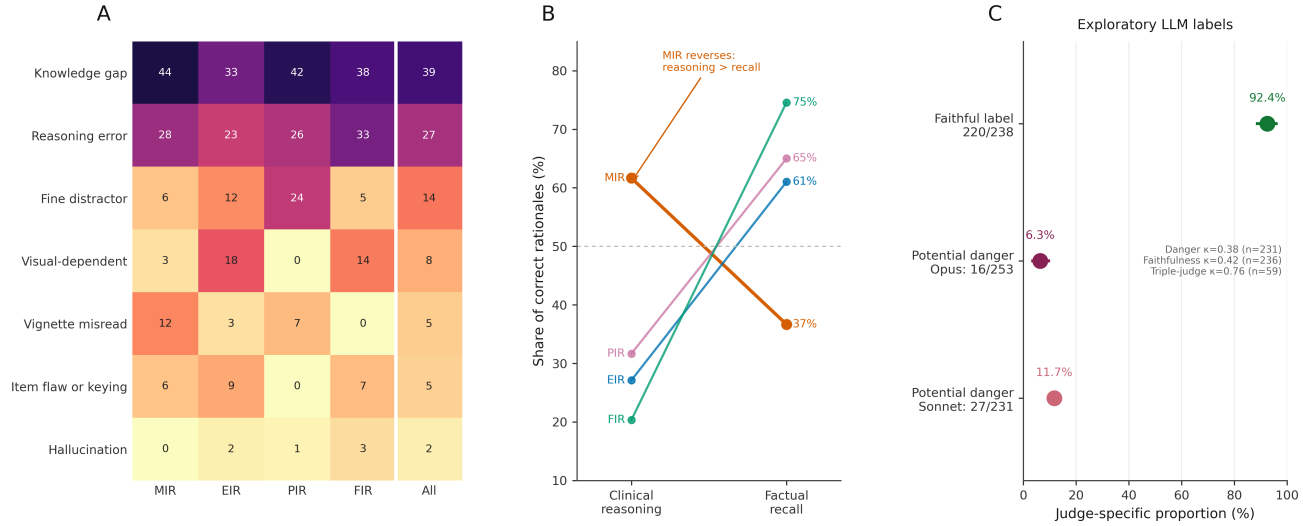
Mean AI accuracy was 93.2% (SD) on items solved by the surface-feature classifier and 91.7% (SD) on the remaining items (unrounded difference 1.6 percentage points); corresponding mean human facility was 66% (SD) and 61% (SD). After adjustment for human facility, the coefficient was -0.145 (95% CI −0.325 to 0.035; *P*=.11), equivalent to an odds ratio (OR) of 0.87 (95% CI 0.72-1.04). We found no clear evidence of differential reliance on the measured cues, but this interval does not establish absence of an effect. The probe addresses a limited set of surface features [28,29] and cannot determine whether high scores reflect prior exposure to related content. Controlled item rewrites remain a planned follow-up.

Exploratory Audit of Errors and Model Explanations

The exploratory audit assessed 253 incorrect frontier-model responses (Figure 5 and Table 4). The Opus judge labeled 99 as knowledge gaps (n=_, 39.1%, bootstrap 95% CI 34-45), 68 as reasoning errors (n=_, 26.9%), 36 as fine-distractor errors (n=_, 14.2%), 21 as visually dependent (n=_, 8.3%), 13 as vignette misreads (n=_, 5.1%), 12 as item flaws or keying issues (n=_, 4.7%), and 4 as hallucinations (n=_, 1.6%). No response received the primary Spanish-language label (0/253), which does not exclude subtler language effects. The sample contained more Claude and weaker-model errors; GPT-5.2 high errors were more often labeled reasoning or fine-distractor errors (Part S1 in Multimedia Appendix 1). All labels concern generated explanations and remained unvalidated by clinicians.

XSL•FO
RenderX

Figure 5. Exploratory large language model–judge audit of generated explanations. (A) Primary failure labels among 253 incorrect responses, by examination. (B) Expressed reasoning labels for 238 correct GPT-5.2 high responses; clinical reasoning exceeds recall on Médico Interno Residente (MIR). (C) Faithfulness and potential-danger labels, with bootstrap 95% CIs where shown. Opus-Sonnet agreement uses 231 jointly labeled errors (failure κ=0.43 and danger κ=0.38) and 236 jointly labeled correct rationales (reasoning κ=0.65 and faithfulness κ=0.42). The separate triple-judge estimate uses 59 MIR and Enfermero Interno Residente (EIR) errors (Fleiss κ=0.76). Psicólogo Interno Residente (PIR) and Farmacéutico Interno Residente (FIR) lack triple-judge coverage, but are included in the 2-judge audit. Labels have not been validated by clinicians and do not measure internal computation or clinical risk. LLM: large language model.



XSL•FO
RenderX

Table . Exploratory large language model-judge audit. Labels describe generated explanations and have not been validated by clinicians[a].

Label or measure	Overall, n/N (%)	MIR, n/N (%)	EIR, n/N (%)	PIR, n/N (%)	FIR, n/N (%)
Knowledge gap	99/253 (39.1)	14/32 (44)	22/66 (33)	41/97 (42)	22/58 (38)
Reasoning error	68/253 (26.9)	9/32 (28)	15/66 (23)	25/97 (26)	19/58 (33)
Fine distractor	36/253 (14.2)	2/32 (6)	8/66 (12)	23/97 (24)	3/58 (5)
Visual dependence	21/253 (8.3)	1/32 (3)	12/66 (18)	0/97 (0)	8/58 (14)
Vignette misread	13/253 (5.1)	4/32 (12)	2/66 (3)	7/97 (7)	0/58 (0)
Item flaw or keying	12/253 (4.7)	2/32 (6)	6/66 (9)	0/97 (0)	4/58 (7)
Hallucination	4/253 (1.6)	0/32 (0)	1/66 (2)	1/97 (1)	2/58 (3)
Spanish-language error	0/253 (0.0)	0/32 (0)	0/66 (0)	0/97 (0)	0/58 (0)
Faithful explanation	220/238 (92.4)	58/60 (97)	50/59 (85)	56/60 (93)	56/59 (95)
Factual recall	141/238 (59.2)	22/60 (37)	36/59 (61)	39/60 (65)	44/59 (75)
Clinical reasoning	84/238 (35.3)	37/60 (62)	16/59 (27)	19/60 (32)	12/59 (20)
Elimination	12/238 (5)	0/60 (0)	7/59 (12)	2/60 (3)	3/59 (5)
Test heuristic	1/238 (0.4)	1/60 (2)	0/59 (0)	0/60 (0)	0/59 (0)
Potential danger					
Opus	16/253 (6.3)	—[b]	—	—	—
Sonnet	27/231 (11.7%)	—	—	—	—
Triple-judge Fleiss κ (n=59)	0.76	MIR and EIR subset	—	Not assessed	Not assessed
Opus-Sonnet					
Failure, κ (n=231)	0.43	All 4 examinations	—	—	—
Danger, κ (n=231)	0.38	All 4 examinations	—	—	—
Faithfulness, κ (n=236)	0.42	All 4 examinations	—	—	—
Reasoning, κ (n=236)	0.65	All 4 examinations	—	—	—

[a]Opus labeled 253 errors and 238 correct-answer rationales. Sonnet comparisons use 231 jointly labeled errors and 236 jointly labeled correct rationales. Triple-judge assessment was limited to 59 MIR and EIR errors. Counts and percentages refer to the available labels within each sample; CIs are bootstrap 95% where shown.

[b]Not applicable.

Opus and Sonnet labels showed Cohen κ=0.43 for failure mode and κ=0.38 for potential danger on 231 jointly labeled errors, and κ=0.65 for reasoning mode and κ=0.42 for faithfulness on 236 jointly labeled correct-answer rationales. Opus flagged 16 of 253 errors as potentially dangerous (6.3%, bootstrap 95% CI 3.6-9.5); Sonnet flagged 27 of 231 (11.7%). These are judge-specific estimates with different available-label denominators. Their difference and the limited agreement show uncertainty in the audit; neither estimate is a clinical risk rate or a lower bound on harm. Profession-matched clinician review of flagged errors and a control sample, together with a cross-vendor judge, is needed before any safety interpretation.

Among 238 correct GPT-5.2 high responses, the Opus judge classified 220 rationales as faithful (92.4%, 95% CI 88.7-95.8), 15 as partially faithful (6.3%), 2 as unfaithful but correct (0.8%), and 1 as unverifiable (0.4%). A total of 2 contained a false statement (n=238, 0.8%). Expressed reasoning was classified as factual recall in 141 (59.2%, 95% CI 53.4-65.6), clinical reasoning in 84 of 238 (35.3%), elimination in 12 of 238 (5%), and a test heuristic in 1 of 238 (0.4%). Clinical reasoning exceeded recall on MIR; the reverse held on EIR, PIR, and FIR. These labels describe one model's explanations and do not verify its internal reasoning process.

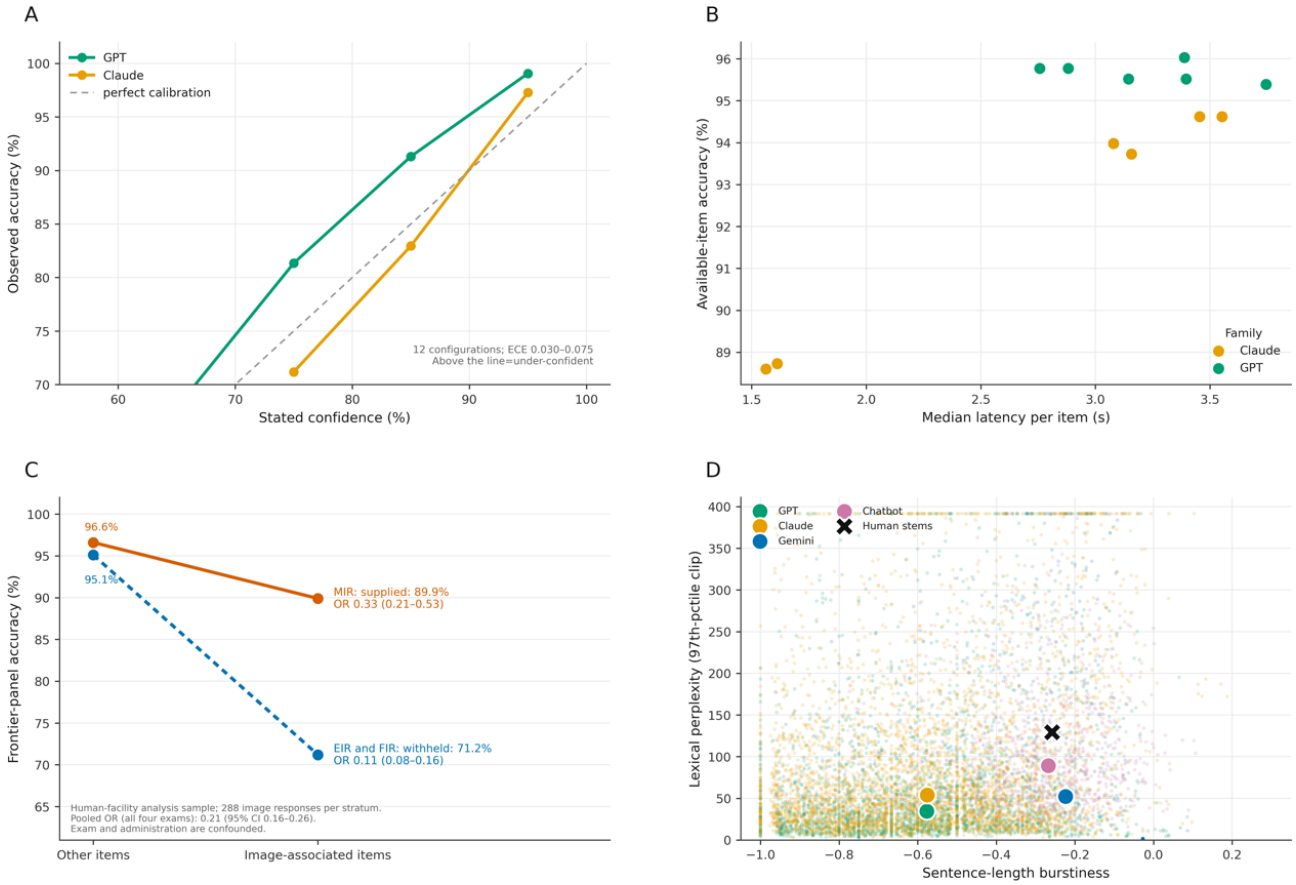
Certainty, Latency, Image Availability, and Prompt Frames

Stated confidence distinguished correct from incorrect responses within the panel (GPT-5.2 AUROC approximately 0.87 to 0.90). The highest AUROC was 0.90 for GPT-5.2 high with the expert prompt, with a 27.6-point difference in mean confidence between correct and incorrect answers. Ten of 12 configurations were net underconfident, the 2 Claude Haiku configurations were net overconfident. ECE ranged from 0.030 to 0.075 within the declared panel (Figure 6A). These discrimination and calibration measures do not establish that confidence can safely guide decisions. Higher GPT-5.2 reasoning effort increased observed latency, but single runs cannot establish a difference

XSL•FO
RenderX

in expected accuracy. Claude Haiku had the shortest median latency, approximately 1.56 (IQR 1.34-1.87) seconds under the neutral prompt and 1.61 (IQR 1.42-1.94) seconds under the expert prompt, with 781 recorded responses in each configuration (Figure 6B).

Figure 6. Calibration, latency, image administration, and stylometry. (A) Accuracy against stated confidence; 10 of 12 configurations were net underconfident despite confidence discriminating correct and incorrect answers. (B) Pooled accuracy against median response latency; single runs do not establish differences in expected accuracy. (C) Image-associated and other item accuracy. Médico Interno Residente (MIR) images were supplied (89.9% vs 96.6%; adjusted odds ratio [OR] 0.33, 95% CI 0.21-0.53); Enfermero Interno Residente (EIR) and Farmacéutico Interno Residente (FIR) images were withheld (71.2% vs 95.1%; OR 0.11, 95% CI 0.08-0.16). The pooled logistic OR across all 4 examinations (9947 observations), 0.21 (95% CI 0.16-0.26) differs from the conditional generalized linear mixed models coefficient of –4.35. Examination and administration differences prevent a causal interpretation. (D) Style measures by model family; classifier accuracy was 82.8%, compared with a 56.1% majority-class baseline. This does not identify official items as AI-written. Panel (C) uses the human-facility-observed analysis sample (288 image-response observations in each image-availability stratum).



The association between image items and accuracy differed by administration route (Figure 6C). On MIR, where images were supplied, accuracy was 89.9% on image items and 96.6% on other items (human facility–adjusted OR 0.33, 95% CI 0.21-0.53). On EIR and FIR, where images were withheld, the corresponding values were 71.2% and 95.1% (OR 0.11, 95% CI 0.08-0.16). Across all 4 examinations (9947 observations), the pooled logistic estimate was OR 0.21 (95% CI 0.16-0.26). It combines differences in examinations, visual input, and administration; it cannot isolate visual reasoning ability. The GLMM coefficient of –4.35 is a separate conditional estimate from a model with item and model random intercepts, and is not the logarithm of the pooled logistic odds ratio. These regression analyses used items with human facility, comprising 24 MIR image items and 24 EIR and FIR image items (10 EIR and 14 FIR), with 288 image-response observations in each stratum.

The panel correlation was 0.40 under neutral prompts and under expert prompts or 0.37, the statistic is Spearman rank correlation coefficient. The mean absolute change in model accuracy between frames was 0.4 percentage points (maximum 0.9). These comparisons assess sensitivity to prompt framing, not stochastic variability under a fixed prompt. Repeated queries across seeds and temperatures were not available for this secondary analysis (Part S1 in Multimedia Appendix 1).

Writing Style of Model Explanations (Supporting Analysis)

A random-forest classifier identified the generating family from style alone at 82.8% (n/N) on the deduplicated classifier sample (n=8183), against the appropriate majority-class baseline of 56.1% (n/N), the signal carried substantially by output length and formatting (Figure 6D). Given the circularity of perplexity- and burstiness-based heuristics popularized for AI detection [48,49] and a Gemini content confound, we present this as supporting only and *not* as a claim that any official item is AI-written; the full analysis is in Part S1 in Multimedia Appendix 1.

XSL•FO
RenderX

Discussion

Principal Findings

Across 839 items from 4 Spanish health-professions examinations, the strongest model configuration answered 795 (94.8%) correctly. All 12 configurations answered 668 of 828 (80.7%) paired items correctly, leaving limited score separation among the strongest configurations. Human and machine item facility showed moderate correspondence (Spearman ρ=0.40, 95% CI 0.34-0.45). Thus, high overall model accuracy did not reproduce the human ordering of item difficulty.

The correlation remained moderate across alternative panels, scoring rules, and examination-stratified pooling. It was also lower than predicted by the specified simulation with a shared difficulty axis and uniform model advantage. This supports a difference in the observed response patterns, but does not exclude all unidimensional models or establish formal measurement noninvariance. After restricting the distractor analysis to the same declared panel, selection of the dominant human distractor was compatible with the post hoc plausibility-weighted baseline. That analysis therefore provides no clear additional evidence of divergence beyond option plausibility. The surface-feature analysis found no clear evidence that the measured cues explained model performance after adjustment for human facility.

Kane framework separates accurate scoring from generalization, extrapolation to professional competence, and decisions based on a score [17-20]. Our data support exact-match scoring and describe performance on these items, while leaving generalization across occasions and test forms uncertain (Table 1). They do not establish extrapolation to clinical performance for models or candidates. Human validation evidence cannot simply be assumed to apply to model respondents [23-26]. The educational implication is therefore to assess the meaning of a model's score before using it as evidence of professional competence.

Prior exposure remains an alternative explanation for high accuracy [25,33]. Public release after the examinations does not establish that related questions or concepts were absent from model training, and the surface-feature probe is not a contamination test. Controlled paraphrases and repeated model queries would help assess response sensitivity, although neither alone would establish the absence of training-data overlap. These analyses were not performed here.

For educators using LLMs to review item banks, machine and human facility should be reported separately. An item readily answered by a model can still be difficult for candidates. We did not measure human item discrimination and therefore cannot conclude that such items have lost their value for ranking candidates. The trap-item results identify response patterns for further investigation; they do not establish that a particular distractor design improves assessment.

The rationale audit provides exploratory descriptions rather than evidence of internal computation. Correct GPT-5.2 high explanations more often received a recall label overall, while clinical-reasoning labels predominated on MIR. A judge's assessment of faithfulness cannot verify that a generated explanation reflects the process that produced the answer. Potential-danger labels also varied between the same-family judges and lacked clinician validation. These findings motivate profession-matched adjudication and cross-vendor replication, rather than a safety estimate or an educational recommendation based on the danger percentage.

Comparison With Prior Work

The study adds item-level evidence to concerns about the construct validity and saturation of AI benchmarks [23-26], extending previous MIR accuracy studies [11-15] across 4 professions. The length-based probe shows a narrow item-writing artifact, but the uncertain adjusted association with model correctness does not settle the broader shortcut-learning question [27-34]. Our exploratory explanation labels relate to work distinguishing knowledge, reasoning, metacognition, and explanation faithfulness [35-38]. Stylometry likewise relates to model-text detection research [48,49], whose limitations and potential biases remain relevant [50,51]. As LLMs are used to draft examination items [52,53], provenance and secure authoring warrant attention [54]; our stylometric results do not identify any official item as AI-written.

Educational scaffolding offers a rationale for supervised formative uses of LLMs [4-7], but examination accuracy alone does not establish their benefit for learning. Such uses need evaluation of feedback quality, learner outcomes, overreliance, and equitable access, alongside the concerns described in prior reviews [1-3]. Similarly, better model scores do not demonstrate a solution to workforce distribution or retention [41]. Image-rich or station-based assessment may broaden the skills sampled, but superiority to the present examinations was not tested here.

Limitations

This was a single [14] cohort using four examinations with a shared format. Generalization to other languages, formats, years, and health systems is untested. Training-data overlap, including recurrent concepts and prior questions, was not assessed; the surface-feature probe cannot exclude contamination.

Each configuration was queried once. Item-bootstrap CIs are conditional on those responses and omit run-to-run uncertainty; changing prompt frames does not estimate that uncertainty. Machine facility is a coarse proportion over related configurations, with substantial ceiling effects. Panel KR-20 is descriptive, and disattenuation depends on assumed reliabilities. The simulation tests a specified uniform-advantage model and cannot rule out every unidimensional account. Human facility derives from aggregate responses, reconstructed for PIR; no person-level item response theory (IRT), human item-discrimination analysis, or formal DIF test was possible. Candidate facility and faculty agreement with the official key are distinct comparisons.

All analyses were exploratory, *P* values were unadjusted, and the plausibility-weighted distractor baseline was constructed post hoc. Its error-event bootstrap does not account for dependence among responses to the same item or from related configurations. Surface features were limited and largely captured option length; an uncertain adjusted association does

XSL•FO
RenderX

not establish absence of shortcut use. The rationale audit used same-family LLM judges, with different label coverage, no profession-matched clinician validation, and a physician-oriented rubric across 4 professions. Correct-answer analyses covered GPT-5.2 high only, with reasoning labels that varied by examination. Neither explanation labels nor potential-danger flags establish internal mechanisms or clinical safety. Stylometry is confounded by genre, output length, and deployment.

Image effects are confounded with examination and administration: MIR images were supplied manually, whereas EIR and FIR images were withheld. These contrasts do not isolate visual competence or establish a benefit of image-rich assessment. More broadly, the study measured MCQ responses, not learning outcomes, clinical performance, or the validity of these examinations for human candidates [18-20,26]. Controlled rewrites, repeated queries, and profession-matched adjudication remain planned follow-up work.

Conclusions

In the January 2026 MIR, EIR, PIR, and FIR examinations, LLMs achieved high accuracy, with limited score separation among the strongest configurations and only moderate correspondence with human item difficulty. These findings caution against treating high model scores as equivalent evidence of human professional competence. They do not establish clinical performance or invalidate the examinations for candidates. The explanation audit and proposed changes to assessment remain exploratory. Further validation requires repeated model queries, evaluation of possible contamination, and profession-matched review before examination scores inform claims about clinical or educational use.

Funding

This study was funded by Healthcademia. The funder operates one of the systems evaluated (the Healthcademia chatbot), which, per the competing-interest firewall (Methods), is excluded from the 12-configuration headline panel, from all comparative ranking, and from construct-validity pillars 1 to 4, appearing only as a separate, non-ranked descriptive family in the stylometric corpus; the funder had no role in study design, data collection, analysis, interpretation, or the decision to submit.

Data Availability

The original study data and analysis code were deposited on Zenodo [55]. The deposit includes the item master, model-by-item responses, surface features, model rationales, audit labels and prompts, and original analysis outputs. Code for the panel-selection corrections in this revision is provided in Multimedia Appendix 2 and uses the archived data. Official items are identified by examination and item number; item stems, answer options, and image annexes are not redistributed. The official 2025 - 2026 booklets, definitive keys, and image annexes are available from the Ministerio de Sanidad [56] #/principal/datosAnteriores/cuadernosExamen (accessed August 22, 2026).

The verbatim prompting conditions were as follows. In the neutral condition, only the question and the four answer options were provided, with no additional instruction. In the expert-role condition, each question was preceded by the Spanish instruction "Eres un especialista con experiencia en preparar el examen MIR. Antes de responder, razona paso a paso internamente, pero no muestres ese razonamiento." (English: "You are a specialist physician with experience preparing candidates for the MIR examination. Before answering, reason step by step internally, but do not show that reasoning."). A confidence probe, "¿Cuánta seguridad tienes del 1 - 100?" ("How confident are you, on a scale from 1 to 100?"), was then recorded as a confidence-related signal.

Authors' Contributions

All authors contributed equally to this work.

Conflicts of Interest

One or more authors are affiliated with Healthcademia, whose domain chatbot is included in this study only as a fenced, separately reported descriptive system (excluded from the headline panel and from construct pillars 1 to 4). This affiliation is disclosed in the interest of transparency. The other authors declare none.

Multimedia Appendix 1
Methods, analyses, results tables, and the rubric for auditing model-generated explanations.
[DOCX File, 63 KB - mededu_v12i1e106199_app1.docx]

Multimedia Appendix 2
Code and reproduction instructions for analyses restricted to the declared 12-configuration model panel.
[DOCX File, 44 KB - mededu_v12i1e106199_app2.docx]

References

XSL•FO
RenderX

1. Xu X, Chen Y, Miao J. Opportunities, challenges, and future directions of large language models, including ChatGPT in medical education: a systematic scoping review. J Educ Eval Health Prof 2024;21:6. [doi: 10.3352/jeehp.2024.21.6] [Medline: 38486402]
2. Haltaufderheide J, Ranisch R. The ethics of ChatGPT in medicine and healthcare: a systematic review on Large Language Models (LLMs). NPJ Digit Med 2024 Jul 8;7(1):183. [doi: 10.1038/s41746-024-01157-x] [Medline: 38977771]
3. Pham TD, Karunaratne N, Exintaris B, et al. The impact of generative AI on health professional education: a systematic review in the context of student learning. Med Educ 2025 Dec;59(12):1280-1289. [doi: 10.1111/medu.15746] [Medline: 40533396]
4. Vygotsky LS. Mind in Society: The Development of Higher Psychological Processes: Harvard University Press; 1978. [doi: 10.2307/j.ctvjf9vz4]
5. Wood D, Bruner JS, Ross G. The role of tutoring in problem solving. J Child Psychol Psychiatry 1976 Apr;17(2):89-100. [doi: 10.1111/j.1469-7610.1976.tb00381.x] [Medline: 932126]
6. van de Pol J, Volman M, Beishuizen J. Scaffolding in teacher–student Interaction: a decade of research. Educ Psychol Rev 2010 Sep;22(3):271-296. [doi: 10.1007/s10648-010-9127-6]
7. Sweller J. Cognitive load during problem solving: effects on learning. Cogn Sci 1988 Apr;12(2):257-285. [doi: 10.1207/s15516709cog1202_4]
8. Kung TH, Cheatham M, Medenilla A, et al. Performance of ChatGPT on USMLE: potential for AI-assisted medical education using large language models. PLOS Digit Health 2023 Feb;2(2):e0000198. [doi: 10.1371/journal.pdig.0000198] [Medline: 36812645]
9. Nori H, King N, McKinney SM, Carignan D, Horvitz E. Capabilities of GPT-4 on medical challenge problems. arXiv 2023 Apr 12. [doi: 10.48550/arXiv.2303.13375]
10. Liu M, Okuhara T, Chang X, et al. Performance of ChatGPT across different versions in medical licensing examinations worldwide: systematic review and meta-analysis. J Med Internet Res 2024 Jul 25;26:e60807. [doi: 10.2196/60807] [Medline: 39052324]
11. Carrasco JP, García E, Sánchez DA. Is “ChatGPT” capable of passing the 2022 MIR exam? implications of artificial intelligence in medical education in Spain [article in Spanish]. Span J Med Educ 2023;4(1). [doi: 10.6018/edumed.556511]
12. Guillen-Grima F, Guillen-Aguinaga S, Guillen-Aguinaga L, et al. Evaluating the efficacy of ChatGPT in navigating the Spanish Medical Residency Entrance Examination (MIR): promising horizons for AI in clinical medicine. Clin Pract 1487;13(6):1460-1487. [doi: 10.3390/clinpract13060130]
13. Cerame A, Juaneda J, Estrella-Porter P. Is GPT-4 capable of passing MIR 2023? comparison between GPT-4 and ChatGPT-3 in the MIR 2022 and 2023 exams [article in Spanish]. Span J Med Educ 2024;5(2). [doi: 10.6018/edumed.604091]
14. Benito P, Isla-Jover M, González-Castro P, et al. GPT-4o and OpenAI o1 performance on the 2024 Spanish competitive medical specialty access examination: cross-sectional quantitative evaluation study. JMIR Med Educ 2026 Jan 12;12:e75452. [doi: 10.2196/75452] [Medline: 41525685]
15. Luengo Vera C, Ferro Picón I, Gómez Gandía JA, et al. Evaluating large language models on the Spanish Medical Intern Resident (MIR) examination 2024/2025: a comparative analysis of clinical reasoning and knowledge application. arXiv 2025 Mar 16. [doi: 10.48550/arXiv.2503.00025]
16. Siam MK, Varela A, Faruk MJH, et al. Benchmarking large language models on the United States medical licensing examination for clinical reasoning and medical licensing scenarios. Sci Rep 2025 Dec 3;16(1):1387. [doi: 10.1038/s41598-025-31010-4] [Medline: 41339739]
17. Messick S. Validity of psychological assessment: validation of inferences from persons’ responses and performances as scientific inquiry into score meaning. Am Psychol 1995;50(9):741-749. [doi: 10.1037/0003-066X.50.9.741]
18. Kane MT. Validating the interpretations and uses of test scores. J Educ Meas 2013 Mar;50(1):1-73 [FREE Full text] [doi: 10.1111/jedm.12000]
19. Cook DA, Brydges R, Ginsburg S, Hatala R. A contemporary approach to validity arguments: a practical guide to Kane’s framework. Med Educ 2015 Jun;49(6):560-575. [doi: 10.1111/medu.12678] [Medline: 25989405]
20. Standards for Educational and Psychological Testing: American Educational Research Association; 2014. URL: https://www.testingstandards.net/uploads/7/6/6/4/76643089/standards_2014edition.pdf [accessed 2026-10-06]
21. Reeves TD, Marbach-Ad G. Contemporary test validity in theory and practice: a primer for discipline-based education researchers. CBE Life Sci Educ 2016;15(1):rm1. [doi: 10.1187/cbe.15-08-0183] [Medline: 26903498]
22. Holland PW, Wainer H, editors. Differential Item Functioning: Lawrence Erlbaum Associates; 1993. URL: https://www.routledge.com/Differential-Item-Functioning/Holland-Wainer/p/book/9781138967694 [accessed 2026-10-06]
23. Raji ID, Bender EM, Paullada A, Denton E, Hanna A. AI and the everything in the whole wide world benchmark. In: Vanschoren J, Yeung S, editors. 2021 Presented at: Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks; Dec 6-14, 2021 URL: https://datasets-benchmarks-proceedings.neurips.cc/paper/2021/hash/084b6fbb10729ed4da8c3d3f5a3ae7c9-Abstract-round2.html [accessed 2026-10-15]
24. Bean AM, Kearns RO, Romanou A, et al. Measuring what matters: construct validity in large language model benchmarks. : Neural Information Processing Systems Foundation, Inc (NeurIPS); 2025 Presented at: Advances in Neural Information

XSL•FO
RenderX

Processing Systems 38; San Diego, California, USA and Mexico City, Mexico URL: http://www.proceedings.com/85713.html [accessed 2026-09-24] [doi: 10.52202/085713-0590]
25. Eriksson M, Purificato E, Noroozian A. Can we trust AI benchmarks? An interdisciplinary review of current issues in AI evaluation. Proc AAAI/ACM Conf AI Ethics Soc 2025;8(1):850-864. [doi: 10.1609/aies.v8i1.36595]
26. Alaa A, Hartvigsen T, Golchini N, et al. Position: medical large language model benchmarks should prioritize construct validity. : PMLR; 2025 Presented at: Proceedings of the 42nd International Conference on Machine Learning (ICML 2025) Proceedings of Machine Learning Research; Jul 13-19, 2025; Vancouver, BC p. 80991-81004 URL: https://proceedings.mlr.press/v267/alaa25a.html [accessed 2026-10-06]
27. Geirhos R, Jacobsen JH, Michaelis C, et al. Shortcut learning in deep neural networks. Nat Mach Intell 2020;2(11):665-673. [doi: 10.1038/s42256-020-00257-z]
28. Lapuschkin S, Wäldchen S, Binder A, Montavon G, Samek W, Müller KR. Unmasking Clever Hans predictors and assessing what machines really learn. Nat Commun 2019 Mar 11;10(1):1096. [doi: 10.1038/s41467-019-08987-4] [Medline: 30858366]
29. Zech JR, Badgeley MA, Liu M, Costa AB, Titano JJ, Oermann EK. Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: a cross-sectional study. PLoS Med 2018 Nov;15(11):e1002683. [doi: 10.1371/journal.pmed.1002683] [Medline: 30399157]
30. Arias-Londoño JD, Godino-Llorente JI. Analysis of the Clever Hans effect in COVID-19 detection using chest X-ray images and Bayesian deep learning. Biomed Signal Process Control 2024 Apr;90:105831. [doi: 10.1016/j.bspc.2023.105831]
31. Vásquez-Venegas C, Wu C, Sundar S, et al. Detecting and mitigating the Clever Hans effect in medical imaging: a scoping review. J Imaging Inform Med 2025 Aug;38(4):2563-2579. [doi: 10.1007/s10278-024-01335-z] [Medline: 39643820]
32. Balepur N, Ravichander A, Rudinger R. Artifacts or abduction: how do llms answer multiple-choice questions without the question? : Association for Computational Linguistics; 2024 Presented at: Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers); Bangkok, Thailand p. 10308-10330 URL: https://aclanthology.org/2024.acl-long [accessed 2026-09-24] [doi: 10.18653/v1/2024.acl-long.555]
33. Cheng Y, Chang Y, Wu Y. A survey on data contamination for large language models. arXiv. Preprint posted online on Feb 20, 2025. [doi: 10.48550/arXiv.2502.14425]
34. Zhao Q, Huang Y, Lv T, et al. MMLU-CF: a contamination-free multi-task language understanding benchmark. : Association for Computational Linguistics; 2025 Presented at: Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers); Vienna, Austria p. 13371-13391 URL: https://aclanthology.org/2025.acl-long [accessed 2026-09-24] [doi: 10.18653/v1/2025.acl-long.656]
35. Thapa R, Wu Q, Wu K, et al. Reasoning or knowledge: stratified evaluation of biomedical llms. : Association for Computational Linguistics; 2026 Presented at: Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers); Rabat, Morocco p. 2450-2483 URL: https://aclanthology.org/2026.eacl-long [accessed 2026-09-24] [doi: 10.18653/v1/2026.eacl-long.111]
36. Griot M, Hemptinne C, Vanderdonckt J, Yuksel D. Large language models lack essential metacognition for reliable medical reasoning. Nat Commun 2025 Jan 14;16(1):642. [doi: 10.1038/s41467-024-55628-6] [Medline: 39809759]
37. Kim J, Podlasek A, Shidara K, Liu F, Alaa A, Bernardo D. Limitations of large language models in clinical problem-solving arising from inflexible reasoning. Sci Rep 2025;15(1):39426. [doi: 10.1038/s41598-025-22940-0]
38. Turpin M, Michael J, Perez E, Bowman SR. Language models don't always say what they think: unfaithful explanations in chain-of-thought prompting. In: Hardt M, Naumann T, Globerson A, Saenko K, Hardt M, Levine S, editors. : Neural Information Processing Systems Foundation, Inc (NeurIPS); 2023 Presented at: Advances in Neural Information Processing Systems 36; New Orleans, LA p. 74952-74965 URL: http://www.proceedings.com/75280.html [accessed 2026-09-24] [doi: 10.52202/075280-3275]
39. Gobierno de España. Sanidad oferta 12.366 plazas de formación sanitaria especializada para 2026. Ministerio de Sanidad. 2025. URL: https://www.sanidad.gob.es/gabinete/notasPrensa.do?id=6737 [accessed 2026-06-21]
40. Ministerio de Sanidad. Un total de 34.553 personas han sido admitidas para optar a las 12.366 plazas de Formación Sanitaria Especializada 2025-2026. Ministerio de Sanidad. 2025. URL: https://www.sanidad.gob.es/gabinete/notasPrensa.do?id=6815 [accessed 2026-06-21]
41. Oferta y necesidad de médicos especialistas en españa 2023-2035. Gobierno de España. 2024. URL: https://www.sanidad.gob.es/areas/profesionesSanitarias/profesiones/necesidadEspecialistas/docs/Oferta_y_necesidad_de_medicos_especialistas_en_Espana_2023-2035.pdf [accessed 2026-06-21]
42. UK medical licensing assessment (UKMLA) overview. General Medical Council. URL: https://www.ukmla.info [accessed 2026-06-21]
43. NHS England. Multi-specialty recruitment assessment (MSRA). 2026. URL: https://medical.hee.nhs.uk/medical-training-recruitment/medical-specialty-training/multi-specialty-recruitment-assessment-msra [accessed 2026-06-21]
44. Centre National de Gestion. URL: https://www.cng.sante.fr/epreuves-dematerialisees-nationales-edn-campagne-2026 [accessed 2026-10-15]
45. Administracao Central do Sistema de Saude (ACSS). National access exam (PNA). Administração Central do Sistema de Saúde, IP. 2026. URL: https://www.acss.min-saude.pt/2018/09/05/prova-nacional-de-acesso-pna/ [accessed 2026-08-22]

XSL•FO
RenderX

46. Recognition of training qualifications. Bundesärztekammer. URL: https://www.bundesaerztekammer.de/en/work-and-training-in-germany/recognition [accessed 2026-06-21]
47. Makhoul AT, Pontell ME, Ganesh Kumar N, Drolet BC. Objective measures needed — program directors' perspectives on a pass/fail USMLE step 1. N Engl J Med 2022:9862334. [doi: 10.1056/NEJMp2006148]
48. Gehrmann S, Strobelt H, Rush AM. GLTR: Statistical Detection and Visualization of Generated Text. : Association for Computational Linguistics; 2019 Presented at: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics; Florence, Italy p. 111-116 URL: https://aclanthology.org/P19-3 [accessed 2026-09-24] [doi: 10.18653/v1/P19-3019]
49. Mitchell E, Lee Y, Khazatsky A, Manning CD, Finn C. DetectGPT: zero-shot machine-generated text detection using probability curvature. Presented at: Proceedings of the 40th International Conference on Machine Learning; Jul 23-29, 2023; Honolulu, HI p. 24950-24962 URL: https://proceedings.mlr.press/v202/mitchell23a.html [accessed 2026-10-06]
50. Sadasivan VS, Kumar A, Balasubramanian S, Wang W, Feizi S. Can AI-generated text be reliably detected? Stress testing AI text detectors under various attacks. Trans Mach Learn Res 2025:2835-8856 [FREE Full text]
51. Liang W, Yuksekgonul M, Mao Y, Wu E, Zou J. GPT detectors are biased against non-native English writers. Patterns (N Y) 2023 Jul 14;4(7):100779. [doi: 10.1016/j.patter.2023.100779] [Medline: 37521038]
52. Riehm L, Nanji K, Lakhani M, Pankiv E, Hasanee D, Pfeifer W. The use of large language models in generating multiple choice questions for health professions education: a systematic review and network meta-analysis. PLoS One 2026;21(1):e0340277. [doi: 10.1371/journal.pone.0340277] [Medline: 41481658]
53. Artsi Y, Sorin V, Konen E, Glicksberg BS, Nadkarni G, Klang E. Large language models for generating medical examinations: systematic review. BMC Med Educ 2024 Mar 29;24(1):354. [doi: 10.1186/s12909-024-05239-y] [Medline: 38553693]
54. Kirchenbauer J, Geiping J, Wen Y, Katz J, Miers I, Goldstein T. A watermark for large language models. : PMLR; 2023 Presented at: Proceedings of the 40th International Conference on Machine Learning; Jul 23-29, 2023; Honolulu, HI p. 17061-17084 URL: https://proceedings.mlr.press/v202/kirchenbauer23a.html [accessed 2026-10-06]
55. Code and data for: "construct validity of frontier large language models on four spanish health-professions specialty-access examinations (MIR, EIR, PIR, FIR): cross-sectional benchmarking and psychometric study". Zenodo. URL: https://doi.org/10.5281/zenodo.22100934 [accessed 2026-09-26]
56. Specialized health training. Ministerio De Sanidad. URL: https://fse.sanidad.gob.es/fseweb/#/principal/escritorio [accessed 2026-09-26]

Abbreviations

AUROC: area under the receiver operating characteristic curve
DIF: differential item functioning
ECE: expected calibration error
EIR: Enfermero Interno Residente
FIR: Farmacéutico Interno Residente
FSE: Formación Sanitaria Especializada
GenAI: generative AI
GLMM: generalized linear mixed model
IRT: item response theory
KR-20: Kuder-Richardson formula 20
LLM: large language model
MCQ: multiple-choice question
MIR: Médico Interno Residente
OLS: ordinary least squares
OR: odds ratio
PIR: Psicólogo Interno Residente
USMLE: United States Medical Licensing Examination

Edited by W Jerjes; submitted 03.Jul.2026; peer-reviewed by A Jaleel, I Aguinaga-Ontoso; revised version received 08.Sep.2026; accepted 10.Sep.2026; published 29.Oct.2026.

Please cite as:
García-Hidalgo C, Carpio-Salmerón M, Benito P, Carazo-Casas C, Carratalá B, Campos Pavón J
Construct Validity of Frontier Large Language Models on 4 Spanish Health Professions Specialty-Access Examinations (MIR, EIR, PIR, and FIR): Cross-Sectional Benchmarking and Psychometric Study
JMIR Med Educ 2026;12:e106199
URL: https://mededu.jmir.org/2026/1/e106199
doi:10.2196/106199

XSL•FO
RenderX

© Clemente García-Hidalgo, Manuel Carpio-Salmerón, Pau Benito, Carlos Carazo-Casas, Beatriz Carratalá, Jaime Campos Pavón. Originally published in JMIR Medical Education (https://mededu.jmir.org), 29.Oct.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Education, is properly cited. The complete bibliographic information, a link to the original publication on https://mededu.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Medical AI Use Intention Among Medical Students and Faculty and Its Associations With AI Literacy, Perceived Benefits, and Risks: Cross-Sectional Survey Study

Bomyee Lee[1], PhD; Gwanwook Bang[2], PhD; Su Jin Chae[1], EdD

[1]Department of Medical Education, University of Ulsan College of Medicine, 30 Badeurae 1-gil, Dong-gu, Ulsan, Republic of Korea

[2]Department of Medical Education, Chosun University College of Medicine, Gwangju, Republic of Korea

Corresponding Author:
Su Jin Chae, EdD
Department of Medical Education, University of Ulsan College of Medicine, 30 Badeurae 1-gil, Dong-gu, Ulsan, Republic of Korea

Abstract

Background: AI is being increasingly integrated into health care and medical education. Although AI literacy is considered an essential competency for future physicians and educators, limited evidence exists regarding how perceived benefits, perceived risks, and AI literacy jointly influence AI use intention among different learner groups.

Objective: This study examined the factors associated with AI use intention among medical students and faculty members and investigated whether the relationships between AI literacy and AI use intention differed according to learner group.

Methods: A cross-sectional self-report survey was conducted among 141 medical students and 94 faculty members (N=235) at a single South Korean medical school. Group-specific linear regression models were estimated for descriptive purposes. To directly test whether associations differed between groups, a pooled regression model included group-by-predictor interaction terms for perceived benefits, perceived risks, AI literacy, and AI use frequency. HC3 heteroscedasticity–robust SEs were used for primary inference, and ordinal logistic regression was conducted as a sensitivity analysis because the 5-point intention outcome showed a marked ceiling effect.

Results: Perceived benefits were positively associated with AI use intention in both groups. In the pooled HC3-robust model, the group-by–AI literacy interaction was significant (*B*=0.361, 95% CI 0.083-0.638; *P*=.01), whereas the interactions for perceived benefits, perceived risks, and AI use frequency were not significant. The group-by–AI literacy interaction remained significant in ordinal logistic regression (odds ratio 2.48, 95% CI 1.24-4.98; *P*=.01). In subgroup models, the negative adjusted AI literacy coefficient among students was model dependent, and the positive perceived risk coefficient was not robust to HC3 inference. Among faculty members, AI use frequency remained positively associated with use intention using both HC3 linear regression (*B*=0.137, 95% CI 0.025-0.249; *P*=.02) and ordinal logistic regression (*P*=.02).

Conclusions: Perceived benefits were consistently associated with stronger AI use intention. The adjusted association between AI literacy and use intention differed between students and faculty, although the subgroup-specific student coefficient was sensitive to model specification. These findings provide exploratory evidence that learner context may shape the relationship between AI literacy and broad self-reported AI use intention; they do not establish causal effects or the effectiveness of specific curricular strategies.

(JMIR Med Educ 2026;12:e105032) doi:10.2196/105032

KEYWORDS

artificial intelligence; AI; AI use intention; AI literacy; perceived benefits; perceived risks; medical education; technology acceptance

Introduction

AI is rapidly transforming health care by supporting clinical decision-making, diagnosis, treatment planning, and health care management. Recent advances in generative AI have further expanded its applications in medical education, including personalized learning, clinical reasoning, assessment, and feedback. As AI becomes increasingly integrated into health care, physicians are expected to use AI not merely as a technical tool but also as a collaborative partner in clinical practice. Accordingly, medical schools are challenged to prepare future physicians and educators who can critically evaluate AI-generated information, recognize its limitations, and integrate it appropriately into clinical decision-making and lifelong learning [1-5].

XSL•FO
RenderX

AI literacy has therefore emerged as an essential competency in medical education. AI literacy encompasses the knowledge, skills, and attitudes required to access, understand, critically evaluate, and appropriately apply AI technologies in professional practice [6-9]. International organizations, including the World Medical Association, have emphasized that physicians should develop competencies enabling them to use AI responsibly while maintaining professional accountability [6]. In parallel, many medical schools have begun introducing AI-related curricula and faculty development programs to prepare learners for AI-supported health care environments [7-9]. However, acquiring AI knowledge alone may not be sufficient to encourage meaningful AI adoption in educational or clinical settings.

Previous studies examining AI acceptance among health care professionals and medical students have consistently demonstrated that perceived usefulness or perceived benefits are among the strongest determinants of technology acceptance [10-14]. According to the technology acceptance model (TAM) and the unified theory of acceptance and use of technology (UTAUT), individuals are more likely to adopt a technology when they believe that it will improve their performance or productivity [13,14]. Conversely, perceived risks—including concerns regarding diagnostic errors, ethical issues, privacy, and professional responsibility—may reduce willingness to adopt AI, although previous findings have been inconsistent across health care settings and participant groups [10,15-17].

In the present study, perceived benefits were used as a context-specific umbrella construct corresponding conceptually to perceived usefulness in the TAM and performance expectancy in the UTAUT. It encompassed more specific expectations that AI may improve efficiency, diagnostic or clinical accuracy, and decision support. Perceived risks represented a distinct evaluative belief concerning potential harm, including impaired judgment, AI-related error, and privacy concerns, rather than the simple inverse of perceived benefits. AI literacy was conceptualized separately as a self-reported capability to access, understand, evaluate, and apply AI-related information, not as an acceptance belief. The learner group was examined as a contextual moderator reflecting differences in professional responsibility, experience, and opportunities for authentic AI use. Accordingly, this study did not test a complete causal TAM or UTAUT model; it used an exploratory association framework to examine the unique and potentially group-dependent contributions of these constructs to broad AI use intention.

Compared with perceived benefits and risks, the educational role of AI literacy remains less clear. AI literacy is generally regarded as a desirable educational outcome because it improves users' ability to understand AI technologies and critically evaluate AI-generated information. However, greater AI literacy may also increase awareness of AI limitations, uncertainty, and ethical concerns, potentially leading to more cautious attitudes toward AI use. Previous studies have therefore reported inconsistent relationships between AI literacy and AI acceptance, suggesting that its influence may depend on learners' educational experiences and professional contexts rather than representing a uniformly positive determinant of AI use intention [18,19].

An important gap remains in the existing literature. Most previous studies have investigated medical students or health care professionals separately, making it difficult to determine whether the factors associated with AI use intention operate similarly across learner groups [10-12,18,19]. Medical students and faculty members differ substantially in clinical experience, educational responsibilities, opportunities to apply AI in practice, and expectations regarding patient care. These contextual differences may influence how perceived benefits, perceived risks, and AI literacy contribute to AI use intention. Understanding these differences is important because educational strategies that are effective for one learner group may not be equally effective for another.

Therefore, this study examined the factors associated with AI use intention among medical students and faculty members. Specifically, we investigated the relative contributions of perceived benefits, perceived risks, and AI literacy to AI use intention and examined whether the relationship between AI literacy and AI use intention differed according to learner group.

Methods

Study Design

This cross-sectional survey was conducted between January and February 2026 at the University of Ulsan College of Medicine to examine factors associated with AI use intention among medical students and faculty members. Specifically, this study compared AI-related characteristics, AI perceptions, AI literacy, and AI use intention between the 2 groups and identified factors independently associated with AI use intention within each group.

Participants

Eligible participants included all enrolled second-year premedical students and first- through fourth-year medical students (n=200), as well as all full-time faculty members (n=532) at the University of Ulsan College of Medicine. First-year premedical students were excluded because the survey was conducted before their enrollment.

An invitation containing information about the study and a link to the anonymous web-based questionnaire was distributed through the institutional communication system. Participation was voluntary, and no incentives were provided.

A total of 235 participants completed the survey, including 141 (60%) medical students (response rate: 141/200, 70.5%) and 94 (40%) faculty members (response rate: 94/532, 17.7%). Only complete questionnaires were included in the final analyses. This online survey is reported in accordance with the CHERRIES (Checklist for Reporting Results of Internet E-Surveys) guidelines (Checklist 1).

Ethical Considerations

This study was approved by the Institutional Review Board of Asan Medical Center (S2025-2554-0001). Before accessing the questionnaire, all participants received information regarding the study purpose, voluntary participation, confidentiality, and data protection. Electronic informed consent was obtained before participation.

XSL•FO
RenderX

The survey was conducted anonymously, and no personally identifiable information was collected. All data were stored in deidentified form and used exclusively for research purposes.

Measures

Overview

The questionnaire included items assessing demographic and AI-related characteristics, AI perceptions, AI use intention, and AI literacy. To establish content validity, 2 experts in medical education independently reviewed all questionnaire items for clarity, relevance, and appropriateness. The questionnaire was revised based on their feedback before finalization.

Demographic and AI-Related Characteristics

Demographic variables included sex, age, medical specialty (or intended career path for students), and previous experience with AI education. AI-related characteristics included familiarity with AI, self-rated AI knowledge, and frequency of AI use.

AI Perceptions

Items assessing perceived benefits, perceived risks, and AI use intention were adapted from the TAM [13], the UTAUT [14], and previous studies on AI acceptance in medicine [10,20].

Perceived benefits were measured using 6 items assessing expectations regarding AI-assisted diagnosis and treatment, improved efficiency, and enhanced clinical accuracy. Perceived risks were measured using 3 items assessing concerns related to impaired clinical judgment, potential diagnostic errors, and privacy issues. Item-level descriptive statistics, corrected item-total correlations, and Cronbach α values if each item was deleted are provided in Table S1 in Multimedia Appendix 1.

AI use intention was measured using a single item assessing participants' intention to use AI in future education, clinical practice, and research. All items were rated on a 5-point Likert scale, with higher scores indicating stronger agreement (Multimedia Appendix 2).

AI Literacy

AI literacy was measured using a 10-item instrument developed based on the integrated health literacy model proposed by Sørensen et al [21] and adapted to AI competencies described by Miao et al [22]. The instrument consisted of 4 domains: access (2 items), understanding (2 items), judgment (3 items), and application (3 items). The 4 domains assess participants' ability to identify reliable AI information sources, understand AI-related concepts, critically evaluate AI-generated information, and appropriately apply AI in educational and professional contexts. All items were rated on a 5-point Likert scale, with higher scores indicating greater AI literacy (Multimedia Appendix 3).

Statistical Analysis

Statistical analyses were performed using Jamovi (version 2.7.24). Continuous variables are presented as means and SDs, whereas categorical variables are presented as frequencies and percentages. The internal consistency of all multi-item scales was evaluated using Cronbach α coefficients, with values greater than 0.70 considered acceptable. Skewness and kurtosis were examined to assess the assumption of univariate normality.

Baseline characteristics were compared between medical students and faculty members using independent-sample 2-tailed *t* tests for continuous variables and chi-square tests for categorical variables. Homogeneity of variances was assessed using the Levene test, and the Welch *t* test was applied when the assumption of equal variances was violated.

Unadjusted between-group comparisons were considered the primary descriptive comparisons. Exploratory analyses of covariance additionally adjusted for sex, age, and previous AI education because these characteristics differed between groups. However, age was strongly associated with learner group, and the age distributions of students and faculty had limited overlap. Therefore, age-adjusted group estimates were treated as sensitivity analyses and were not interpreted as representing a causal or directly comparable student-faculty contrast.

Pearson correlation analyses were conducted separately for medical students and faculty members. Group-specific multiple linear regression models were estimated to describe associations with AI use intention. To directly test whether predictor associations differed by participant group, a pooled regression model included participant group and group-by-predictor interaction terms for perceived benefits, perceived risks, total AI literacy, and AI use frequency. Continuous predictors were mean centered before interaction terms were constructed. The pooled model additionally included AI knowledge, AI familiarity, sex, age, and previous AI education. Because AI use intention was measured on a bounded 5-point scale and showed a marked ceiling effect, HC3 heteroscedasticity–robust SEs were used for primary inferential interpretation. Ordinal logistic regression was conducted as a sensitivity analysis using the same predictor set. For this analysis, perceived benefits, perceived risks, total AI literacy, and AI use frequency were standardized before constructing the group-by-predictor interaction terms; thus, the corresponding odds ratios represent associations per 1-SD increase. The proportional odds assumption was also assessed. Complete ordinal logistic regression results, including main effects, interaction terms, threshold estimates, model fit statistics, and assessment of the proportional odds assumption, are provided in Table S2 in Multimedia Appendix 1. Additional sensitivity analyses examined reduced predictor sets to evaluate the stability of the student AI literacy coefficient. Multicollinearity was assessed using variance inflation factors and tolerance statistics. Results are reported as unstandardized coefficients; standardized coefficients where applicable; odds ratios for ordinal logistic regression; 95% CIs; and 2-sided *P* values, with $P<.05$ considered statistically significant.

Results

Demographic Characteristics of the Participants

A total of 235 participants completed the survey, comprising 141 (60%) medical students and 94 (40%) faculty members (Table 1). Among the students, 58.2% (82/141) were male, and the mean age was 23.6 (SD 1.87) years. Second-year premedical

XSL•FO
RenderX

students constituted the largest academic year subgroup (38/141, 27.0%). Regarding intended career paths, surgery was the most frequently selected option (53/141, 37.6%).

Among the faculty members, 71.3% (67/94) were male, and the mean age was 48.9 (SD 7.58) years. Half (47/94, 50%) of the faculty participants were professors. Regarding medical specialties, other clinical specialties represented the largest category (45/94, 47.9%). Their mean professional experience was 17.3 (SD 8.19) years.

Table . Participant characteristics.

Characteristics and categories	Students (n=141)	Faculty (n=94)
Sex, n (%)		
Male	82 (58.2)	67 (71.3)
Female	59 (41.8)	27 (28.7)
Age (y), mean (SD)	23.6 (1.87)	48.9 (7.58)
Academic year, n (%)		
Premedical year 2	38 (27.0)	—[a]
Medical year 1	35 (24.8)	—
Medical year 2	21 (14.9)	—
Medical year 3	31 (22.0)	—
Medical year 4	16 (11.3)	—
Academic position, n (%)		
Professor	—	47 (50)
Associate professor	—	27 (28.7)
Assistant professor	—	19 (20.2)
Nontenure position	—	1 (1.1)
Professional experience (y), mean (SD)	—	17.3 (8.19)
Intended career path (students) or specialty (faculty), n (%)		
Internal medicine	36 (25.5)	22 (23.4)
Surgery	53 (37.6)	16 (17.0)
Other clinical specialties	36 (25.5)	45 (47.9)
Research or nonclinical field	3 (2.1)	11 (11.7)
Undecided	13 (9.2)	—

[a]Not applicable.

Comparison of Demographic and AI-Related Characteristics Between Students and Faculty

In the unadjusted comparisons, significant between-group differences were observed for sex (χ^2_1=4.2; P=.04), age ($t_{100.6}$=−31.7; P<.001), and previous AI education (χ^2_1=18.2; P<.001; Table 2). Faculty members included a higher proportion of male individuals, whereas students reported greater previous AI education experience. Because age was strongly confounded with learner group and the observed age ranges had limited overlap, the covariate-adjusted estimates that follow are presented as exploratory sensitivity analyses and should not be interpreted as comparisons of otherwise equivalent students and faculty members.

In the exploratory covariate-adjusted analysis, faculty members reported more frequent AI use than students (adjusted mean 4.13, SE 0.224 vs 3.30, SE 0.159; $F_{1,230}$=5.550; P=.02). No adjusted differences were observed in AI knowledge level ($F_{1,230}$=0.132; P=.72) or AI familiarity ($F_{1,230}$=0.304; P=.58). These adjusted findings should be interpreted cautiously because of the limited age overlap between groups.

XSL•FO
RenderX

Table . Comparison of demographic and AI-related characteristics between medical students and faculty members.

Characteristics	Students (n=141)	Faculty (n=94)	Statistic (*df*)	*P* value
Male sex, n (%)	82 (58.2)	67 (71.3)	χ^2=4.2 (1)	.04
Age (y), mean (SD)	23.6 (1.87)	48.9 (7.58)	*t*=–31.7 (100.6)	<.001
Previous AI education, n (%)	98 (69.5)	39 (41.5)	χ^2=18.2 (1)	<.001
AI use frequency[a] (possible range 1-5), adjusted mean (SE)	3.30 (0.159)	4.13 (0.224)	*F*=5.550 (1,230)	.02
AI knowledge[a] (possible range 1-5), adjusted mean (SE)	2.53 (0.131)	2.64 (0.185)	*F*=0.132 (1,230)	.72
AI familiarity[a] (possible range 1-5), adjusted mean (SE)	2.86 (0.150)	3.04 (0.212)	*F*=0.304 (1,230)	.58

[a]Adjusted means from analysis of covariance controlling for sex, age, and previous AI education.

Reliability of the Study Instruments

The AI perception and AI literacy instruments demonstrated acceptable internal consistency in the present sample. The Cronbach α was 0.884 for the perceived benefits subscale and 0.617 for the perceived risks subscale. For the AI literacy instrument, Cronbach α values ranged from 0.726 to 0.847 across the 4 domains, and the overall scale demonstrated excellent internal consistency (Cronbach α=0.899).

Between-Group Differences in AI Perceptions and AI Literacy

In exploratory analysis of covariance sensitivity analyses adjusting for sex, age, and previous AI education, no statistically significant differences were observed between medical students and faculty members in perceived benefits, perceived risks, AI use intention, or any AI literacy domain (*P*>.05 in all cases; Table 3). Given the strong association between age and learner group and limited age overlap, these adjusted means are descriptive sensitivity estimates rather than evidence of equivalent student-faculty populations. Item-level analysis of the perceived-risk scale in the full sample showed mean scores of 3.65 (SD 1.10), 3.81 (SD 0.91), and 3.93 (SD 0.94) for impaired clinical judgment, potential diagnostic errors, and privacy concerns, respectively; corrected item-total correlations ranged from 0.391 to 0.455.; detailed item-level results are provided in Table S1 in Multimedia Appendix 1.

Table . Adjusted comparisons of AI perceptions and AI literacy between medical students and faculty members. Adjusted means were estimated in an exploratory analysis of covariance sensitivity analysis controlling for sex, age, and previous AI education. Because age was strongly associated with learner group and age distributions had limited overlap, adjusted group estimates should be interpreted cautiously.

Categories and variables	Students, mean (SE)	Faculty, mean (SE)	*F* test (*df*)	*P* value
AI perception				
Perceived benefits (possible range 1-5)	3.94 (0.101)	4.02 (0.142)	0.127 (1,230)	.72
Perceived risks (possible range 1-5)	3.90 (0.117)	3.63 (0.165)	1.111 (1,230)	.29
AI use intention (possible range 1-5)	4.30 (0.109)	4.56 (0.155)	1.068 (1,230)	.30
AI literacy				
Access (possible range 1-5)	3.91 (0.126)	4.27 (0.178)	1.674 (1,230)	.20
Understanding (possible range 1-5)	3.97 (0.110)	4.29 (0.156)	1.748 (1,230)	.19
Judgment (possible range 1-5)	3.71 (0.113)	3.96 (0.159)	0.958 (1,230)	.33
Application (possible range 1-5)	3.89 (0.105)	3.89 (0.148)	<0.001 (1,230)	.98
Total (possible range 1-5)	3.85 (0.094)	4.07 (0.132)	1.043 (1,230)	.31

XSL•FO
RenderX

Correlation Analysis

Correlation analyses were conducted separately for medical students and faculty members (Tables 4 and 5). Among medical students, perceived benefits showed a strong positive correlation with AI use intention (r=0.660; P<.001), whereas total AI literacy was weakly and nonsignificantly associated with AI use intention (r=0.155; P=.07). Perceived risks demonstrated a weak positive correlation with AI use intention (r=0.293; P<.001).

Among faculty members, both perceived benefits (r=0.588; P<.001) and total AI literacy (r=0.408; P<.001) were positively correlated with AI use intention, whereas perceived risks were not significantly associated with AI use intention (r=0.140; P=.18).

Table . Pearson correlations among AI perceptions, AI literacy, and AI use intention in medical students (n=141).

	Perceived benefits	Perceived risks	Total AI literacy	AI use intention
Perceived benefits	1	0.238[a]	0.295	0.660
Perceived risks	0.238[a]	1	0.427	0.293
Total AI literacy	0.295[b]	0.427[b]	1	0.155
AI use intention	0.660[b]	0.293[b]	0.155	1

[a]P<.01.

[b]P<.001.

Table . Pearson correlations among AI perceptions, AI literacy, and AI use intention in faculty members (n=94).

	Perceived benefits	Perceived risks	Total AI literacy	AI use intention
Perceived benefits	1	0.170	0.332	0.588
Perceived risks	0.170	1	0.005	0.140
Total AI literacy	0.332[a]	0.005	1	0.408
AI use intention	0.588[b]	0.140	0.408[b]	1

[a]P<.01.

[b]P<.001.

Factors Associated With AI Use Intention

Group-specific multiple linear regression models retaining the full covariate set (perceived benefits, perceived risks, AI literacy, AI use frequency, AI knowledge, AI familiarity, sex, age, and previous AI education) were estimated for students and faculty members (Table 6). HC3 heteroscedasticity–robust SEs were used as the primary basis for inference because the bounded intention outcome showed a marked ceiling effect.

Among faculty members, perceived benefits were positively associated with AI use intention (B=0.529, HC3 95% CI 0.288-0.771; standardized β=0.503; P<.001). AI use frequency also remained positively associated with intention (B=0.137, HC3 95% CI 0.025-0.249; standardized β=0.257; P=.02). AI literacy was positive in the conventional linear model but marginal with HC3 inference (B=0.221, HC3 95% CI −0.018 to 0.460; P=.07).

XSL•FO
RenderX

Table . Multiple linear regression models predicting AI use intention among medical students and faculty members.[a]

Predictors	Students (n=141)[b]		Faculty (n=94)[c]	
	B (HC3 95% CI)	β	*B* (HC3 95% CI)	β
Perceived benefits	0.665[d] (0.503 to 0.828)	0.590	0.529[d] (0.288 to 0.771)	0.503
Perceived risks	0.187 (−0.037 to 0.411)	0.192	0.056 (−0.082 to 0.194)	0.060
AI literacy	−0.197[e] (−0.382 to −0.012)	−0.163	0.221 (−0.018 to 0.460)	0.200
AI use frequency	0.063 (−0.074 to 0.200)	0.079	0.137[e] (0.025 to 0.249)	0.257
AI knowledge level	0.062 (−0.067 to 0.191)	0.078	−0.048 (−0.278 to 0.182)	−0.054
Familiarity	0.063 (−0.077 to 0.202)	0.083	−0.097 (−0.263 to 0.069)	−0.140
Sex (female)	−0.120 (−0.318 to 0.078)	−0.082	−0.069 (−0.349 to 0.211)	−0.047
Age	−0.017 (−0.073 to 0.038)	−0.044	0.003 (−0.015 to 0.022)	0.038
AI education experience	0.096 (−0.145 to 0.336)	0.061	0.106 (−0.164 to 0.375)	0.078

[a]*B* indicates the unstandardized regression coefficient, and β indicates the standardized coefficient from the corresponding ordinary least squares model. The 95% CIs and *P* values were estimated using HC3 heteroscedasticity–robust SEs. All variance inflation factors were below 2.0, indicating no problematic multicollinearity.

[b]R^2=0.510; adjusted R^2=0.476; $F_{9,131}$=15.14; *P*<.001.

[c]R^2=0.452; adjusted R^2=0.393; $F_{9,84}$=7.69; *P*<.001.

[d]*P*<.001

[e]*P*<.05.

Pooled Interaction and Sensitivity Analyses

In the pooled HC3-robust regression model (N=235), the group-by–AI literacy interaction was statistically significant (*B*=0.361, 95% CI 0.083-0.638; *P*=.01), indicating that the adjusted association between AI literacy and use intention differed between students and faculty members. The interactions for perceived benefits, perceived risks, and AI use frequency were not significant (Table 7). The model explained 47.0% of the variance (adjusted R^2=0.436; $F_{14,220}$=11.88; *P*<.001).

AI use intention showed a marked ceiling effect: 55.3% (130/235) of participants selected “strongly agree” (score 5), and 36.6% (86/235) selected “agree” (score 4). In ordinal logistic sensitivity analyses, the group-by–AI literacy interaction remained significant (odds ratio 2.48, 95% CI 1.24-4.98; *P*=.01), whereas the other interactions remained nonsignificant. Complete ordinal logistic regression results, including main effects, interaction terms, threshold estimates, model fit statistics, and assessment of the proportional odds assumption, are provided in Table S2 in Multimedia Appendix 1. The student AI literacy coefficient was not stable across reduced linear models, and the positive student perceived risk coefficient was not significant with HC3 robust inference. These findings support limiting the principal group difference claim to the AI literacy interaction and interpreting subgroup coefficients cautiously.

Table . Group-by-predictor interactions in pooled HC3 linear and ordinal logistic regression models.[a]

Interaction	*B* (HC3 95% CI)	*P* value	Ordinal OR[b] (95% CI)	*P* value
Group × perceived benefits	−0.172 (−0.452 to 0.107)	.23	0.73 (0.36 to 1.49)	.39
Group × perceived risks	−0.132 (−0.405 to 0.140)	.34	0.73 (0.37 to 1.44)	.36
Group × AI literacy	0.361 (0.083 to 0.638)	.01	2.48 (1.24 to 4.98)	.01
Group × AI use frequency	−0.002 (−0.151 to 0.148)	.98	0.98 (0.48 to 1.96)	.94

[a]*B* indicates the unstandardized regression coefficient from the pooled linear regression model estimated using HC3 heteroscedasticity–robust SEs. *P* values were estimated using HC3 heteroscedasticity–robust SEs for the linear regression model.

[b]OR: odds ratio from the ordinal logistic regression model.

Discussion

Principal Findings

This study examined factors associated with broad self-reported AI use intention among medical students and faculty members within a single medical school. Three principal findings emerged.

First, perceived benefits were consistently and strongly associated with AI use intention in both learner groups. This association remained robust across the primary regression analyses and sensitivity analyses using HC3

XSL•FO
RenderX

heteroscedasticity–robust SEs. These findings suggest that participants who perceived AI as beneficial for education, research, or clinical practice were more likely to report intentions to use AI regardless of learner group.

Second, the pooled interaction analysis demonstrated that the adjusted association between AI literacy and AI use intention differed significantly between medical students and faculty members. Importantly, this conclusion was based on direct statistical testing of the interaction rather than comparison of separate subgroup regressions. However, the subgroup-specific pattern should be interpreted cautiously. Among students, the zero-order association between AI literacy and use intention was weak and positive, whereas the adjusted regression coefficient became negative only after controlling for correlated AI-related variables and varied across reduced models. This pattern is compatible with suppression or shared variance rather than a stable inverse relationship. In contrast, among faculty members, AI literacy showed a positive association with AI use intention that was generally consistent across analyses.

Third, although perceived risks appeared positively associated with AI use intention among students in the conventional regression model, this association was attenuated and no longer statistically significant when HC3-robust SEs were applied. Moreover, the pooled interaction model did not demonstrate a significant difference between students and faculty members for perceived risks. Accordingly, this finding should be regarded as exploratory rather than evidence that greater awareness of AI-related risks increases willingness to use AI.

Comparison With Previous Studies

The finding that perceived benefits represented the strongest correlate of AI use intention is consistent with those of previous studies examining technology acceptance in health care. The TAM and UTAUT consistently identify perceived usefulness or performance expectancy as the principal determinant of behavioral intention. Similarly, recent studies involving physicians, medical students, and other health care professionals have reported that expected educational or clinical benefits are the most influential predictors of AI acceptance. Our findings extend this literature by demonstrating that perceived benefits remained the dominant correlate of AI use intention even after accounting for AI literacy, perceived risks, AI familiarity, AI knowledge, and previous AI education.

The relationship between AI literacy and AI use intention was more complex. Previous studies have generally assumed that greater AI literacy facilitates AI adoption because users become more confident in understanding and applying AI technologies. However, there is emerging evidence also suggesting that greater AI literacy may increase awareness of AI limitations, uncertainty, ethical concerns, and the need for critical evaluation. Rather than demonstrating that AI literacy has uniformly positive or negative effects, our findings suggest that its association with AI use intention may depend on learners' educational and professional contexts. Faculty members who routinely integrate AI into teaching, research, or clinical work may experience AI literacy primarily as a practical competency, whereas students may simultaneously acquire greater awareness of both the opportunities and limitations of AI during training. Because the adjusted student coefficient was sensitive to model specification, these findings should be interpreted as evidence of contextual variation rather than evidence that greater AI literacy reduces students' willingness to use AI.

Our findings also highlight the importance of distinguishing between broad AI use and specific AI applications. The present questionnaire assessed general intention to use AI across education, research, and clinical practice rather than focusing on a particular AI technology such as generative AI or clinical decision support systems. These contexts differ substantially with respect to task complexity, responsibility, explainability requirements, governance, and acceptable levels of risk. Recent studies suggest that trust, governance, institutional policies, and clinical context substantially influence AI adoption in health care. For example, a 2026 survey of health care practitioners found that trust and perceived risk jointly shaped AI acceptance and readiness, with trust functioning as a construct distinct from perceived benefit or risk alone [23], and work on health chatbot adoption has shown that perceived risk lowers trust both directly and indirectly through perceived usefulness and self-efficacy [24]. Separately, studies of generative AI integration into medical curricula indicate that adoption is shaped substantially by institutional governance, including policy on permitted use and equitable platform access, in addition to individual perceptions of benefit or risk [25]. Because trust and governance constructs were not directly measured in the present study, future investigations should incorporate them explicitly, distinguish among different categories of AI technologies, and examine how determinants of AI acceptance vary across educational and clinical settings.

Educational Implications

Although the present findings are exploratory, they suggest several implications for AI education in medical schools. The significant interaction between learner group and AI literacy indicates that identical educational approaches may not address the needs of students and faculty members equally. Future educational interventions for students may benefit from emphasizing supervised AI application, critical appraisal of AI-generated information, appropriate calibration of trust, and responsible clinical reasoning rather than focusing exclusively on conceptual knowledge. For faculty members, educational programs may place greater emphasis on integrating AI into authentic teaching, research, and clinical activities while supporting confident and appropriate implementation.

Importantly, these implications should be regarded as educational hypotheses rather than evidence-based curricular recommendations. Because this study did not evaluate educational interventions, curriculum design, or subsequent AI use behavior, further longitudinal and intervention studies are needed before specific educational strategies can be recommended.

Limitations

Several limitations should be acknowledged.

First, participants were recruited from a single medical school, and the faculty response rate was relatively low. Consequently,

XSL•FO
RenderX

self-selection bias cannot be excluded, and the findings may not be generalizable to other institutions or health care settings.

Second, the cross-sectional design and reliance on self-reported questionnaire data preclude causal inference. The observed associations cannot determine whether AI use frequency increases AI use intention, whether stronger intentions lead to more frequent AI use, or whether both are influenced by other unmeasured factors.

Third, AI use intention was assessed using a single broad item encompassing education, research, and clinical practice. Although this approach captured participants' overall intention to use AI, it could not distinguish among different AI technologies or application contexts and did not permit assessment of measurement error. Future studies should use validated multi-item measures and separately examine intentions related to generative AI, clinical decision support systems, and other AI applications.

Fourth, although the AI literacy instrument was adapted from established theoretical frameworks and demonstrated acceptable internal consistency, additional psychometric evaluation—including exploratory or confirmatory factor analysis, construct validation, and measurement invariance across learner groups—was beyond the scope of the present study. Future research should further validate the measurement properties of this instrument in larger and more diverse samples.

Finally, the perceived risk subscale demonstrated relatively limited internal consistency and combined concerns related to clinical judgment, diagnostic errors, and privacy. Furthermore, the positive association between perceived risks and AI use intention among students was not robust in HC3 sensitivity analyses, and the corresponding interaction effect was not statistically significant. Consequently, findings related to perceived risks should be interpreted cautiously until replicated in future studies using more refined measures of AI-related risk.

Conclusions

Perceived benefits were consistently associated with stronger broad self-reported AI use intention among both medical students and faculty members. The association between AI literacy and AI use intention differed according to learner group, although the underlying mechanism remains uncertain. These findings provide exploratory evidence that learner context may shape AI use intention and highlight the need for multicenter longitudinal studies using validated measures to better understand AI adoption in medical education.

Acknowledgments

During the preparation of this manuscript, the authors used generative AI tools (Claude Sonnet 4.6; Anthropic and ChatGPT using GPT-5.6 Sol; OpenAI) to assist with interpreting peer review comments; discussing potential revisions and additional statistical analyses; conducting supplementary literature searches; and improving the organization, language, and clarity of the manuscript text. All statistical analyses reported in this manuscript were independently performed and verified by the authors using Jamovi (version 2.7.24). All AI-generated suggestions and text were critically reviewed, revised, and verified by the authors. The authors take full responsibility for the content of this publication.

Funding

This research received no external funding.

Data Availability

The data presented in this study are available on reasonable request from the corresponding author (SJC; edujin1@ulsan.ac.kr). The data are not publicly available due to privacy and ethical restrictions related to the protection of participant information.

Authors' Contributions

Conceptualization: BL, SJC
Data curation: GB
Formal analysis: BL, SJC
Investigation: GB
Methodology: BL
Supervision: SJC
Writing—original draft: BL
Writing—review and editing: SJC
All authors have read and agreed to the published version of the manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1

XSL•FO
RenderX

Item-level descriptive statistics for the perceived-risk scale and complete ordinal logistic regression model predicting AI use intention in the pooled sample.
[DOCX File, 19 KB - mededu_v12i1e105032_app1.docx]

Multimedia Appendix 2
AI perception survey instrument.
[DOCX File, 14 KB - mededu_v12i1e105032_app2.docx]

Multimedia Appendix 3
AI literacy survey instrument.
[DOCX File, 14 KB - mededu_v12i1e105032_app3.docx]

Checklist 1
CHERRIES checklist.
[DOCX File, 38 KB - mededu_v12i1e105032_app4.docx]

References

1. Krishnan G, Singh S, Pathania M, et al. Artificial intelligence in clinical medicine: catalyzing a sustainable global healthcare paradigm. Front Artif Intell 2023;6:1227091. [doi: 10.3389/frai.2023.1227091] [Medline: 37705603]
2. Mehta N, Mehta S, Rubenstein A, Wood SK. Not replaced, but reinvented: AI education pathways to prepare future physicians to lead healthcare transformation. Perspect Med Educ 2025 Nov 21;14(1):849-859. [doi: 10.5334/pme.2233] [Medline: 41278578]
3. Tanaka H. The third revolution of medicine by big data and artificial intelligence. Chem Bio Inform J 2025;25:53-70. [doi: 10.1273/cbij.25.53]
4. Agarwal R, Dugas M, Gao G. Augmenting physicians with artificial intelligence to transform healthcare: challenges and opportunities. J Econ Manag Strategy 2024 Mar;33(2):360-374. [doi: 10.1111/jems.12555]
5. Sezgin E. Artificial intelligence in healthcare: complementing, not replacing, doctors and healthcare providers. Digit Health 2023 Jul 2;9:20552076231186520. [doi: 10.1177/20552076231186520] [Medline: 37426593]
6. WMA statement on artificial and augmented intelligence in medical care. World Medical Association. 2025 Oct 11. URL: https://www.wma.net/policies-post/wma-statement-on-artificial-and-augmented-intelligence-in-medical-care/ [accessed 2026-09-07]
7. Kim S, Kim SH, Kim H, Lee YM. Integrating artificial intelligence into medical curricula: perspectives of faculty and students in South Korea. Korean J Med Educ 2025 Mar;37(1):65-70. [doi: 10.3946/kjme.2025.324] [Medline: 40049684]
8. Lee I, Lee YM. Development of a medical AI ethics education module reflecting AI competency framework for medical students. Korean Med Educ Rev 2025 Feb;27(1):17-25. [doi: 10.17496/kmer.24.042]
9. Plan to foster over 1,000 talents over five years in the fields of AI diagnosis and prediction, new drug development, and medical device development [Article in Korean]. Ministry of Health and Welfare, Republic of Korea. 2025. URL: https://www.mohw.go.kr/board.es?act=view&bid=0027&list_no=1487140&mid=a10503010100 [accessed 2026-09-07]
10. Chen M, Zhang B, Cai Z, et al. Acceptance of clinical artificial intelligence among physicians and medical students: a systematic review with cross-sectional survey. Front Med (Lausanne) 2022 Aug 31;9:990604. [doi: 10.3389/fmed.2022.990604] [Medline: 36117979]
11. Tao W, Yang J, Qu X. Utilization of, perceptions on, and intention to use AI chatbots among medical students in China: national cross-sectional study. JMIR Med Educ 2024 Oct 28;10:e57132. [doi: 10.2196/57132] [Medline: 39466038]
12. Wagner G, Raymond L, Paré G. Understanding prospective physicians' intention to use artificial intelligence in their future medical practice: configurational analysis. JMIR Med Educ 2023 Mar 22;9:e45631. [doi: 10.2196/45631] [Medline: 36947121]
13. Davis FD. Perceived usefulness, perceived ease of use, and user acceptance of information technology. MIS Q 1989 Sep;13(3):319-340. [doi: 10.2307/249008]
14. Venkatesh V, Morris MG, Davis GB, Davis FD. User acceptance of information technology: toward a unified view. MIS Q 2003 Sep 1;27(3):425-478. [doi: 10.2307/30036540]
15. Dai Q, Li M, Yang M, et al. Attitudes, perceptions, and factors influencing the adoption of AI in health care among medical staff: nationwide cross-sectional survey study. J Med Internet Res 2025 Aug 8;27:e75343. [doi: 10.2196/75343] [Medline: 40779308]
16. Schulz PJ, Kee KM, Lwin MO, et al. Clinical experience and perception of risk affect the acceptance and trust of using AI in medicine. Front Digit Health 2025 Sep 2;7:1620127. [doi: 10.3389/fdgth.2025.1620127] [Medline: 40964535]
17. Turinici M. Understanding the impact of perceived risks on AI adoption among physicians. Rev Econ Bus Stud 2025;18(2):33-49. [doi: 10.47743/rebs-2025-2-0002]

XSL•FO
RenderX

18. Laupichler MC, Aster A, Meyerheim M, Raupach T, Mergen M. Medical students' AI literacy and attitudes towards AI: a cross-sectional two-center study using pre-validated assessment instruments. BMC Med Educ 2024 Apr 10;24(1):401. [doi: 10.1186/s12909-024-05400-7] [Medline: 38600457]
19. Ma S, Lu Y, Wu H, Zhou Q, Huang H, Song J. Knowledge, trust attitudes, and behaviors towards AI and their association with critical thinking disposition in Chinese undergraduate medical students: a cross-sectional study. BMC Med Educ 2026 Apr 28;26(1):955. [doi: 10.1186/s12909-026-09318-0] [Medline: 42050524]
20. Ghozali MT, Satibi G, Fortwengel G. Adoption intentions toward AI-based clinical decision support tools: a TAM study on hospital pharmacists. J Inf Commun Technol 2025 Jul;24(3):44-69. [doi: 10.32890/jict2025.24.3.3]
21. Sørensen K, Van den Broucke S, Fullam J, et al. Health literacy and public health: a systematic review and integration of definitions and models. BMC Public Health 2012 Jan 25;12:80. [doi: 10.1186/1471-2458-12-80] [Medline: 22276600]
22. Miao Y, Zhang Y, Zhao C, Ma S, Wang J. Development and validity testing of the Medical Student Artificial Intelligence Information Literacy Scale (MS-AIILS). Research Square. Preprint posted online on Sep 10, 2025. [doi: 10.21203/rs.3.rs-7327894/v1]
23. Alsalloum G, Badr Y, Alzaatreh A, et al. Acceptance and readiness for AI among United Arab Emirates-based health care practitioners: exploratory cross-sectional survey. JMIR AI 2026 Apr 17;5:e80173. [doi: 10.2196/80173] [Medline: 42012070]
24. Song L, Liu J, Yasin M, Mokhtar MM. Perceived risk and trust towards health chatbots: extending TAM with self-efficacy. Information 2026;17(5):439. [doi: 10.3390/info17050439]
25. Rueff G, Monaco P, Rusinol AE, Hernandez M. Challenges in the rapid and responsible integration of generative artificial intelligence (AI) into a new medical school curriculum. Cureus 2025 Jun 26;17(6):e86796. [doi: 10.7759/cureus.86796] [Medline: 40718215]

Abbreviations

CHERRIES: Checklist for Reporting Results of Internet E-Surveys
TAM: technology acceptance model
UTAUT: unified theory of acceptance and use of technology

Edited by D Chartash; submitted 18.Jun.2026; peer-reviewed by N Ding, S Chen, W Dong; revised version received 19.Aug.2026; accepted 31.Aug.2026; published 24.Sep.2026.

Please cite as:
Lee B, Bang G, Chae SJ
Medical AI Use Intention Among Medical Students and Faculty and Its Associations With AI Literacy, Perceived Benefits, and Risks: Cross-Sectional Survey Study
JMIR Med Educ 2026;12:e105032
URL: https://mededu.jmir.org/2026/1/e105032
doi:10.2196/105032

© Bomyee Lee, Gwanwook Bang, Su Jin Chae. Originally published in JMIR Medical Education (https://mededu.jmir.org), 24.Sep.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Education, is properly cited. The complete bibliographic information, a link to the original publication on https://mededu.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Generative AI for Personalized Learning Platforms in Otolaryngology–Head and Neck Surgery Education: Scoping Review

Jiang-Tao Zhong, MD; Zai-Zai Cao, MD; Shui-Hong Zhou, MD, PhD

Department of Otolaryngology, School of Medicine, First Affiliated Hospital Zhejiang University, Qingchun Road 79, Hangzhou, Zhejiang, China

Corresponding Author:
Shui-Hong Zhou, MD, PhD
Department of Otolaryngology, School of Medicine, First Affiliated Hospital Zhejiang University, Qingchun Road 79, Hangzhou, Zhejiang, China

Abstract

Background: Generative AI, including large language models (LLMs), may enable personalized learning in medical education. In otolaryngology–head and neck surgery (ear, nose, and throat [ENT]) education, AI may support knowledge retrieval, adaptive assessment, feedback, and scalable learning, but evidence remains fragmented.

Objective: This scoping review aimed to map generative AI research in otolaryngology education; summarize applications, performance, benefits, and risks; and identify directions for AI-enabled personalized learning.

Methods: We included English-language empirical studies, preprints, and technical or developmental evaluations published from 2017 through 2026 with direct relevance to professional ENT education; nonempirical publications and studies without a professional educational focus were excluded. PubMed, Embase, Web of Science, Scopus, ERIC, IEEE Xplore, and CINAHL were searched, supplemented by citation searching. The final search was conducted on July 6, 2026. Data were charted using a standardized form by 2 reviewers and synthesized through a 3-stage inductive thematic analysis, with third reviewer adjudication. No formal critical appraisal or meta-analysis was performed because of study heterogeneity.

Results: A total of 42 studies published between 2023 and 2026 were included. Most were performance studies (n=33, 78.6%), followed by system development (n=3, 7.1%), educational intervention studies (n=3, 7.1%), surveys (n=2, 4.8%), and guideline development (n=1, 2.4%). Most studies evaluated question sets or model outputs rather than participants. Applications included knowledge retrieval, revision, board-style question practice, adaptive assessment, AI avatar–delivered instruction, and operative-planning support. Seven major thematic domains were identified: (1) educational support and learning enhancement; (2) learning outcomes, educational effectiveness, and pedagogical value; (3) model performance, accuracy, and variability; (4) assessment, examination readiness, and question-type sensitivity; (5) reliability, safety, and risks; (6) adoption, usability, and implementation barriers; and (7) advanced AI architectures and future educational design. Newer models performed strongly on selected structured knowledge tasks, but performance varied by task, difficulty, modality, and reasoning demands. Evidence remained heterogeneous and benchmarking focused, with only 3 educational interventions and limited evidence for sustained learning, clinical reasoning, or real-world outcomes. Hallucinations, unreliable sourcing, confidently incorrect outputs, and contextual and multimodal reasoning limitations supported supervised rather than autonomous use.

Conclusions: By moving beyond reviews centered on AI applications or model performance, this review provides an education-focused synthesis linking generative AI capabilities, learning outcomes, assessment, safety, implementation, and emerging architectures to the design of personalized learning in otolaryngology. Its contribution is a conceptual shift from evaluating standalone chatbots toward developing supervised, pedagogically governed, knowledge-grounded, and adaptive learning systems that preserve learner reasoning and expert oversight. These findings inform educators, curriculum designers, and developers when integrating generative AI into formative assessment, feedback, remediation, and personalized learning activities, while avoiding unsupported autonomous or high-stakes use. Future research should prioritize longitudinal educational outcomes, curriculum integration, higher-order reasoning, equitable implementation, and governance.

(JMIR Med Educ 2026;12:e91785) doi:10.2196/91785

KEYWORDS

generative AI; large language models; ChatGPT; personalized learning; otolaryngology–head and neck surgery; otolaryngology education; ENT education; medical education; scoping review

XSL•FO
RenderX

Introduction

Rationale

Otolaryngology–head and neck surgery (ear, nose, and throat [ENT]) is a broad and technically demanding specialty that requires trainees to integrate complex knowledge across anatomy, oncology, airway management, otology, rhinology, pediatrics, and perioperative care. However, in many training systems, undergraduate and early postgraduate exposure to ENT remains limited, and protected teaching time is constrained by clinical service demands. These structural constraints have intensified the need for scalable, flexible, and learner-centered educational approaches that support self-directed learning and just-in-time knowledge acquisition [1].

Medical education is undergoing a rapid digital transformation, with increasing emphasis on personalization, adaptive learning pathways, and competency-based progression. Generative AI, particularly large language models (LLMs), has emerged as a general-purpose technology relevant to these educational goals. Systems such as ChatGPT, Gemini, Claude, and related multimodal models are capable of generating explanations, synthesizing information, answering questions, and simulating clinical scenarios in natural language, with broad applications across biomedical knowledge domains [2,3]. More broadly, AI has been increasingly discussed as an enabling component of digitally supported, person-centered learning and care pathways in medicine [4].

Within otolaryngology, interest in generative AI has grown rapidly. The specialty has been identified as a suitable testbed for AI-assisted educational and clinical applications due to its reliance on imaging, structured anatomical knowledge, guideline-based decision pathways, and subspecialty-diverse training requirements [5,6]. Broader surgical education literature has also emphasized the potential role of AI in supporting training, assessment, and simulation [7]. At the same time, the clinical AI literature highlights that generative AI differs fundamentally from traditional educational resources and clinical decision-support systems. Its outputs are probabilistic, model dependent, and sensitive to task framing, raising questions regarding transparency, accountability, and governance in medical contexts [8-10]. These considerations apply equally to educational uses and underscore the need for systematic, domain-specific evaluation rather than premature adoption.

Several published reviews and perspective articles have examined AI and generative AI in otolaryngology, but their scope differs from the educational focus of this review. Lechien [11] reviewed generative AI across otolaryngology, with emphasis on the accuracy and clinical utility of ChatGPT for disease information, clinical vignettes, and scientific information. Edalati et al [12] examined opportunities, technical limitations, and ethical considerations of GPT-4 in otolaryngology, with education considered alongside diagnosis, perioperative care, research, and administrative applications. Crowson and Nwosu [13] addressed integration of AI into otolaryngology practice, particularly data infrastructure, electronic health records, model validation, and clinical implementation. From an educational perspective, Brennan et al [1] described ways in which ChatGPT could supplement ENT surgical training, but their work was a trainee-centered perspective based on illustrative prompts rather than a systematic synthesis of evidence. Buess et al [14] conducted a scoping review of generative and multimodal AI across medicine, focusing primarily on model architectures, datasets, clinical applications, and evaluation methods rather than specialty-specific education.

Thus, existing reviews have primarily addressed the clinical and technical applications of AI in otolaryngology or the broader evolution of generative AI across medicine [15-17]. They have not systematically mapped the empirical literature specifically on generative AI in ENT professional education or integrated evidence across educational applications, learning outcomes, model performance, assessment, safety, adoption, and emerging system architectures for personalized learning. This distinction is important because strong performance on isolated knowledge or examination tasks does not establish educational effectiveness, curricular value, or readiness for implementation. A dedicated scoping review is therefore needed to characterize this heterogeneous and rapidly evolving evidence base, identify evidence gaps, and determine how current findings can inform pedagogically grounded, AI-enabled personalized learning platforms in otolaryngology.

Objectives

The aim of this scoping review was to systematically map literature on generative AI in otolaryngology education. Specifically, we sought to characterize educational contexts and applications studied; summarize how generative AI has been evaluated in this field; and identify key conceptual, methodological, and practical gaps. We also aimed to synthesize the emerging directions of research and development to inform pedagogically grounded, AI-enabled personalized learning platforms in otolaryngology education and provide a conceptual foundation for future empirical and implementation research. The review was conceptualized around generative AI–enabled personalized learning as the central concept; professional otolaryngology education as the context; and learners, educators, assessment tasks, and educational systems as the relevant units of inquiry. The guiding research question for this review was as follows: “What is the current state of evidence regarding the use of generative AI for personalized learning in otolaryngology (ENT) education?” This question encompassed LLM performance, educational applications and implementation contexts, emerging platform-level designs, reported benefits and risks, and architectural strategies for adaptive and personalized learning.

Methods

Protocol and Registration

This scoping review was conducted and reported in accordance with the PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews) checklist and the methodology described by Tricco et al [18]. A review protocol was developed before the review and is provided in Multimedia Appendix 1. The protocol was not prospectively registered.

XSL•FO
RenderX

Eligibility Criteria

Eligible sources were English-language original empirical studies, preprints, and technical or developmental evaluations with direct relevance to professional otolaryngology education. Studies were eligible if they directly evaluated or applied generative AI in ENT professional education and reported empirical data, performance evaluations, user studies, educational interventions, or developmental outcomes. Review articles, commentaries, expert perspectives, letters, conference abstracts, purely theoretical or technical papers without an educational component, and studies focused on patient education rather than professional training were excluded. Publications from 2017 through 2026 were considered to capture the modern era of generative AI relevant to medical education [19].

Information Sources

We searched PubMed (NCBI [National Center for Biotechnology Information]), Embase (Elsevier), Web of Science (Clarivate), Scopus (Elsevier), ERIC, IEEE Xplore, and CINAHL (EBSCOhost). Each database was searched separately; no multidatabase platform was used simultaneously. Citation searching was also used to identify additional eligible sources. Specifically, included studies underwent backward and forward citation searches. No study registries or additional online or print sources were searched or browsed. No authors, experts, manufacturers, or other parties were contacted, and no other information sources or search methods were used. The final search was conducted on July 6, 2026.

Search

The literature search was reported in accordance with the PRISMA-S (Preferred Reporting Items for Systematic Reviews and Meta-Analyses literature search extension) guideline [20]. The search strategy combined controlled vocabulary terms and free-text keywords related to otolaryngology, head and neck surgery, AI, generative AI, LLMs, ChatGPT, DeepSeek, education, training, personalized learning, teaching, and assessment. Search limits included the English language and publication years from 2017 to 2026. No published search filters were used, and no search strategies from previous literature reviews were adapted or reused. No search update was performed after July 6, 2026. No formal peer review of the search strategy was conducted. Complete search strategies for all 7 databases, reproduced exactly as run, are provided in Multimedia Appendix 2.

Selection of Sources of Evidence

All retrieved records were imported into EndNote (version 20; Clarivate) and deduplicated using its automated duplicate-detection function before screening. Two reviewers independently screened titles and abstracts and assessed potentially eligible full-text reports against the predefined eligibility criteria. Disagreements were resolved through discussion and, when necessary, adjudication by a third reviewer. Citation searching was used to identify additional potentially eligible reports.

Data Charting Process

Data were charted using a standardized form developed iteratively by the review team. Two reviewers independently charted data, and discrepancies were resolved through discussion or adjudication by a third reviewer.

Data Items

For each included study, we recorded publication year, country, study type, purpose, setting and context, generative AI system evaluated or used, sample size or evaluation set, and findings relevant to the thematic synthesis. Study types were assigned using each article's title and the design information reported in its Abstract and Methods sections. Explicit author-reported design terms were used when available; otherwise, the primary objective and methodological approach determined the category. Studies were mapped to 5 standardized categories: performance study, survey study, system development study, educational intervention study, and guideline development study.

Critical Appraisal of Individual Sources of Evidence

No formal critical appraisal of individual sources of evidence was conducted because this scoping review aimed to map a heterogeneous emerging literature rather than estimate intervention effectiveness or pooled effects.

Synthesis of Results

The charted data were synthesized descriptively and through a 3-stage inductive thematic synthesis. First, JTZ and ZZC independently summarized the subthemes identified in each included article. Second, SHZ deduplicated, screened, and refined the 2 sets of subthemes. Third, JTZ and ZZC synthesized the refined subthemes into major themes; SHZ adjudicated disagreements. The final synthesis comprised 7 major themes and associated subthemes. Results were summarized narratively, from source characteristics to thematic synthesis across the 7 domains.

Results

Selection of Sources of Evidence

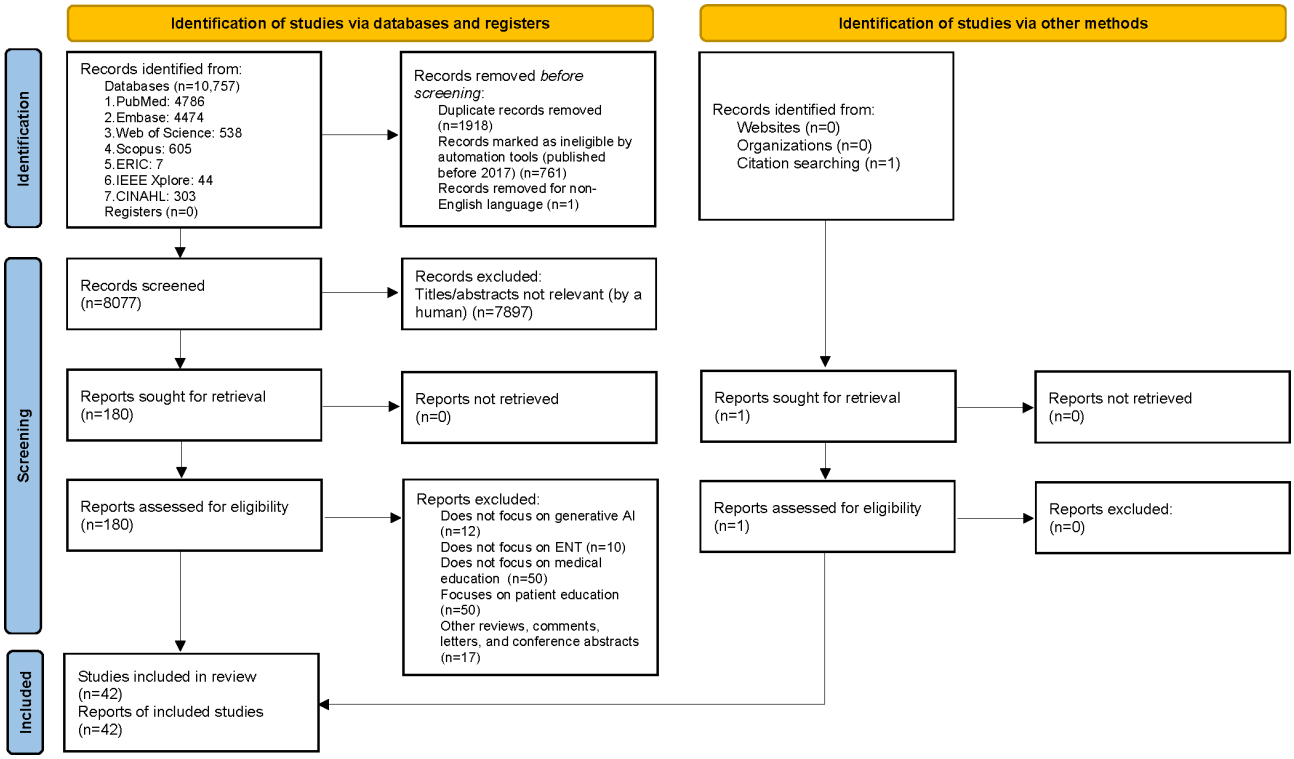
The final database search was conducted on July 6, 2026. The 7 databases yielded 10,757 records: PubMed (n=4786, 44.5%), Embase (n=4474, 41.6%), Web of Science (n=538, 5.0%), Scopus (n=605, 5.6%), ERIC (n=7, 0.1%), IEEE Xplore (n=44, 0.4%), and CINAHL (n=303, 2.8%). Before screening, duplicate records (n=1918, 17.8%), records published before 2017 (n=761, 7.1%), and 1 non-English record (n=1, <0.1%) were removed. The remaining 8077 (75.1%) database records were screened by title and abstract, and 7897 (73.4%) records were excluded as not relevant to generative AI in otolaryngology education. The full texts of 180 (1.7%) database records were sought and assessed for eligibility. One additional report identified through citation searching (backward and forward citation searches) was also retrieved and assessed; no reports were unavailable for retrieval.

Of the 180 database reports assessed at full text, 139 (77.2%) articles were excluded for the following reasons: not focusing on generative AI (n=12, 6.7%), not focusing on ENT (n=10,

XSL•FO
RenderX

5.6%), not focusing on medical education (n=50, 27.8%), focusing primarily on patient education rather than professional training (n=50, 27.8%), or being reviews, commentaries, expert perspectives, letters, or conference abstracts (n=17, 9.4%). No reports were excluded due to retrieval failure. In total, 41 (22.8%) database studies and 1 (0.6%) citation-identified study were included, yielding 42 (23.3%) studies, including 1 (0.6%) preprint. The study selection process is summarized in the PRISMA flow diagram (Figure 1).

Figure 1. PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews) flow diagram of search and screening for large language models in otolaryngology education. ENT: ear, nose, and throat.

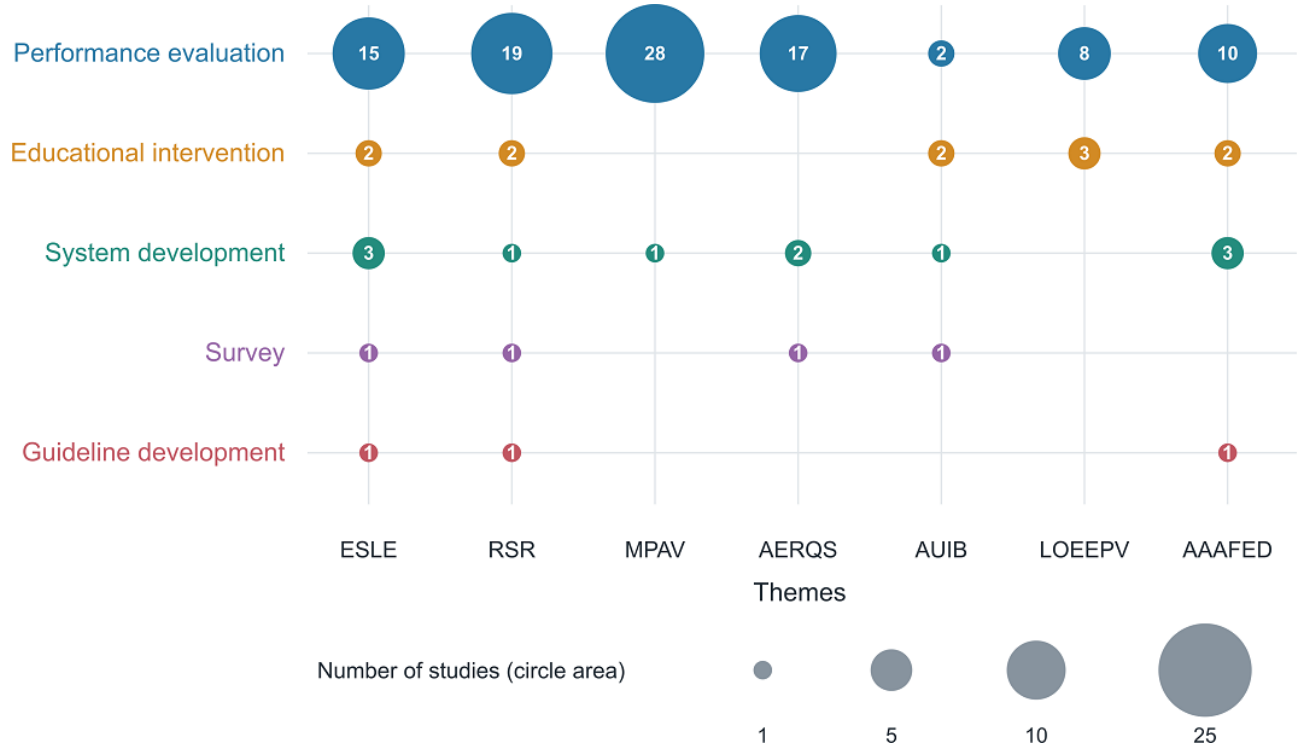


Characteristics of Sources of Evidence

The 42 included studies were published between 2023 and 2026: 7 (16.7%) in 2023, 17 (40.5%) in 2024, 13 (31%) in 2025, and 5 (11.9%) in 2026. The evidence base was internationally distributed across North America, Europe, and Asia, including several multinational collaborations. Most studies were performance studies (33/42, 78.6%), followed by 3 (7.1%) system development studies, 3 (7.1%) educational intervention studies, 2 (4.8%) survey studies, and 1 (2.4%) guideline development study. Educational contexts were predominantly assessment-focused and benchmarking-focused, especially board-style or standardized question banks, while a smaller group addressed adaptive knowledge learning, resident induction, operative-planning support, study-guide generation, trainee adoption, and specialized AI system development. Table S1 in Multimedia Appendix 3 provides the study-level details, and Figures 2-5 provide the review-level evidence mapping. The evaluated technologies ranged from general-purpose generative AI models to customized or hybrid educational systems, including retrieval-augmented, agentic, adaptive-assessment, and AI avatar–based approaches (Table 1).

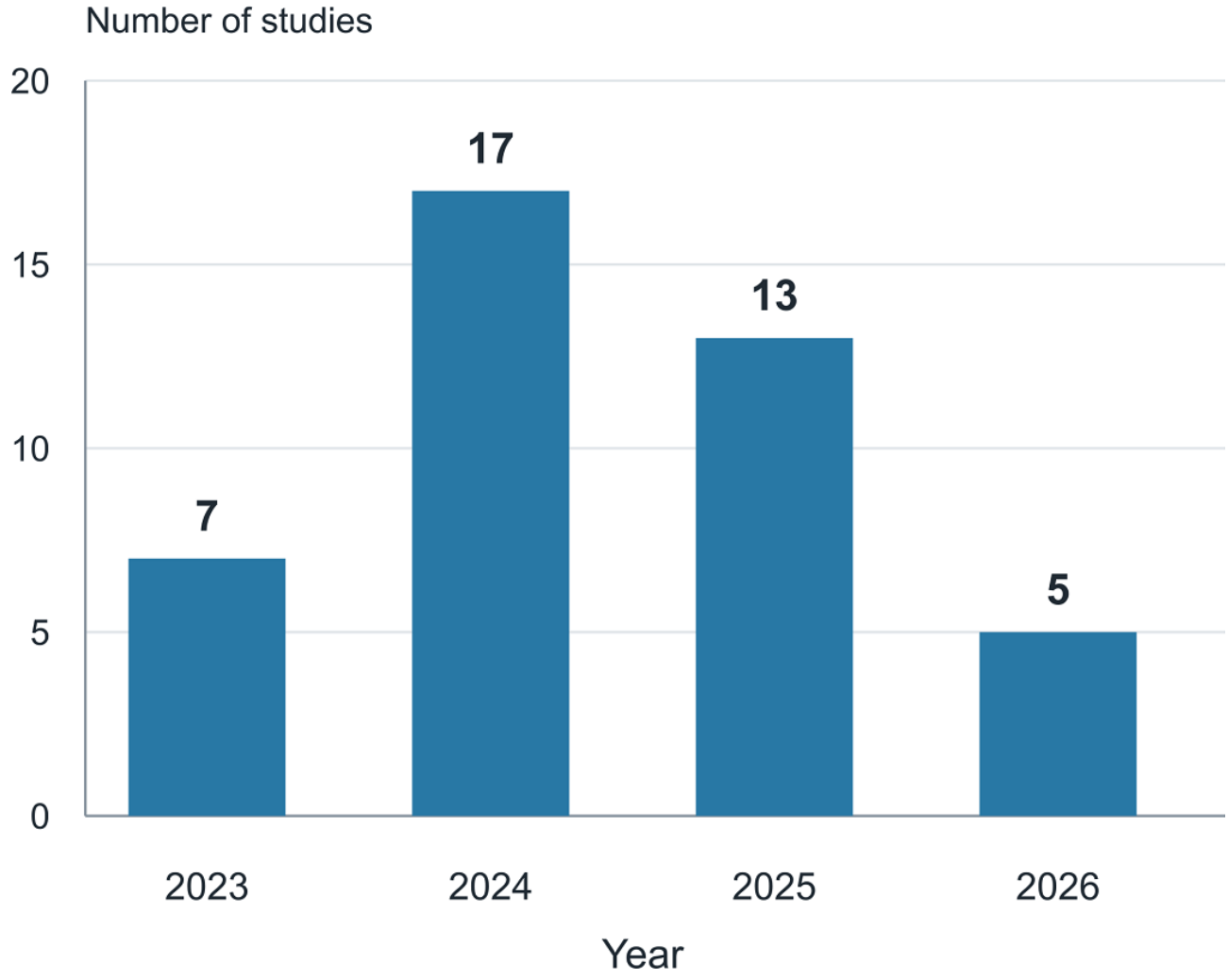
XSL•FO
RenderX

Figure 2. Evidence map of study types and themes in generative AI research in otolaryngology education. Circle area is proportional to the number of studies, with exact counts shown within circles. Colors indicate study types, and empty cells represent zero studies. Individual studies may address multiple themes. AAAFED: advanced AI architectures and future educational design; AERQS: assessment, examination readiness, and question-type sensitivity; AUIB: adoption, usability, and implementation barriers; ESLE: educational support and learning enhancement; LOEEPV: learning outcomes, educational effectiveness, and pedagogical value; MPAV: model performance, accuracy, and variability; RSR: reliability, safety, and risks.



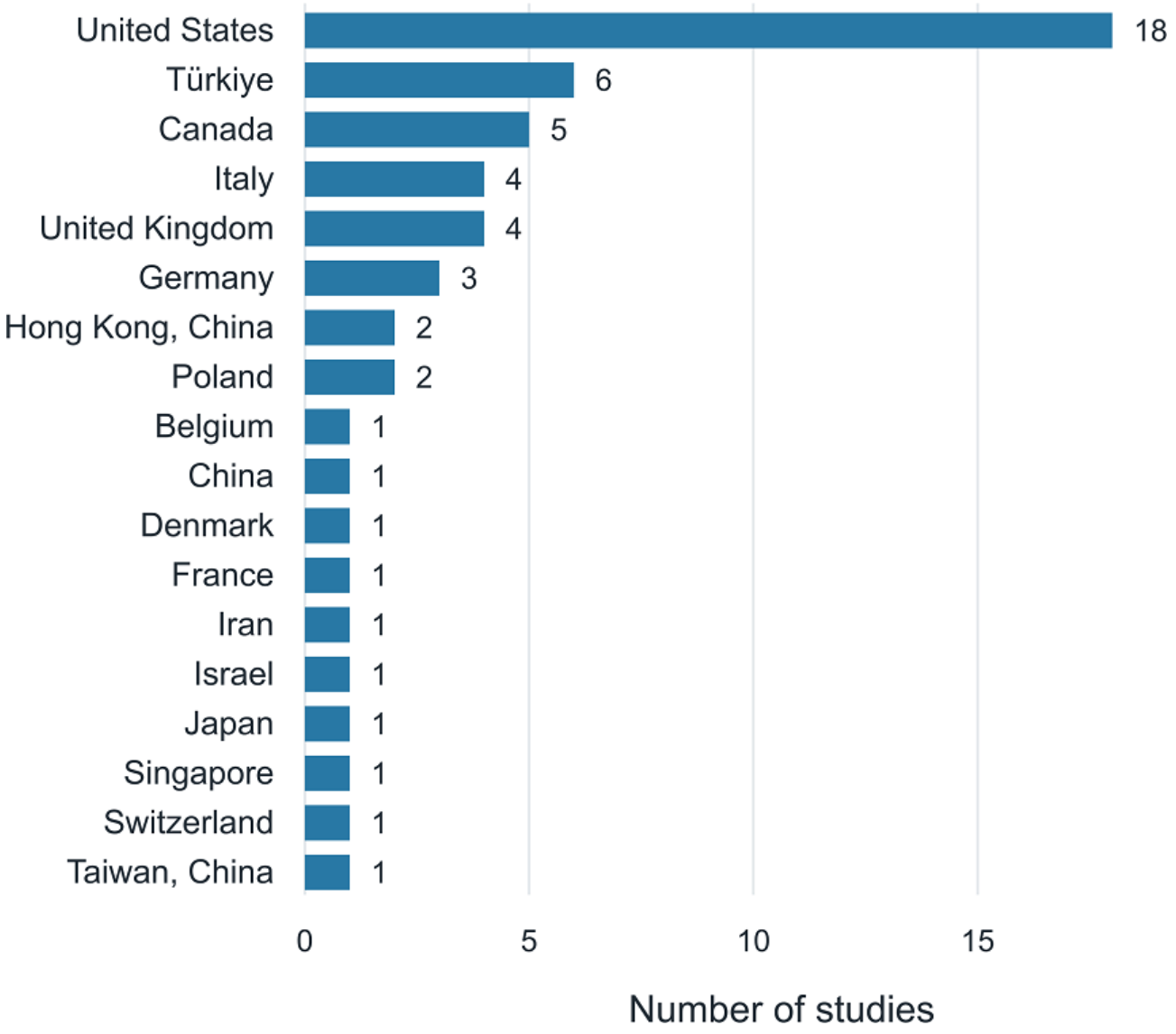
XSL•FO
RenderX

Figure 3. Publication-year distribution of included studies. Bars show the number of included studies published in each year from 2023 to 2026 (N=42). Numbers above the bars indicate exact counts.



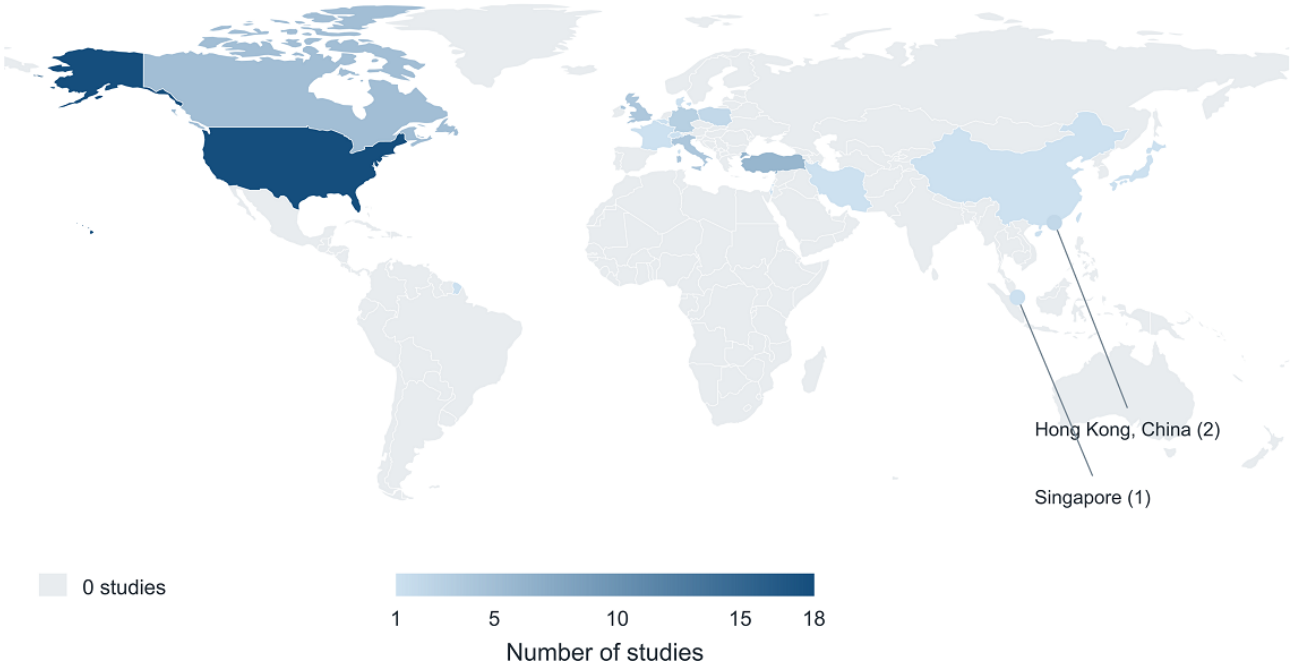
XSL•FO
RenderX

Figure 4. Number of included studies by country/region. Horizontal bars show study counts for each country/region, with exact counts displayed at the end of each bar. Multicountry studies were counted once for each participating country/region, yielding 54 study-location contributions across 42 studies.



XSL•FO
RenderX

Figure 5. Global distribution of included studies. Color intensity indicates the number of studies associated with each country/region; darker blue represents higher counts, and gray indicates zero included studies. Labeled markers identify Hong Kong, China, and Singapore to improve visibility. Multicountry studies were counted once for each participating country/region, yielding 54 study-location contributions across 42 studies.



XSL•FO
RenderX

Table . Large language models reported in the identified literature.

Model	Interface	Developer	Year of release
GPT-3.5	ChatGPT	OpenAI	2022
GPT-3.5 Turbo	ChatGPT	OpenAI	2023
GPT-4	ChatGPT	OpenAI	2023
GPT-4 Turbo	ChatGPT	OpenAI	2023
GPT-4 with Vision (GPT-4V)	ChatGPT	OpenAI	2023
Custom GPT-4 ENT[a] Expert MCQ[b] Solver	OpenAI custom GPT builder interface	OpenAI	2023
E-GPT-A (custom GPT-4)	OpenAI custom GPT builder interface	OpenAI	2023
GPT-4o	ChatGPT	OpenAI	2024
GPT-4o mini	ChatGPT	OpenAI	2024
ChatENT (GPT-4 plus CEKB[c]/RAG[d])	Custom virtual Q&A[e] platform	OpenAI base model; developed by study authors	2024
ChatGPT-5	ChatGPT	OpenAI	2025
o3-mini	OpenAI API	OpenAI	2025
Google Bard	Bard	Google	2023
Gemini Ultra	Gemini	Google	2023
Gemini 1.0 Pro	Gemini	Google	2023
Gemini 1.5 Flash	Gemini	Google	2024
Gemini 1.5 Flash 8B	Gemini	Google	2024
Gemini 1.5 Pro	Gemini	Google	2024
Gemini 2.0	Gemini	Google	2024 - 2025
Gemini-3	Gemini	Google	2025
Claude 2	Claude	Anthropic	2023
Claude 2.1	Claude	Anthropic	2023
Claude 3	Claude	Anthropic	2024
Claude 3 Haiku	Claude	Anthropic	2024
Claude 3.5 Sonnet	Claude	Anthropic	2024
Claude 3 Opus	Claude	Anthropic	2024
Claude 3.7 Sonnet	Claude	Anthropic	2024 - 2025
Claude Sonnet 4	Claude	Anthropic	2024 - 2025
Grok-2	Grok	xAI	2024
Grok-3	Grok	xAI	2025
Perplexity	Perplexity	Perplexity AI	2022
Copilot	Copilot	Microsoft	2023
DeepSeek-V3	DeepSeek	DeepSeek	2024
DeepSeek-R1	DeepSeek	DeepSeek	2025
DeepSeek-4	DeepSeek	DeepSeek	2025
OpenEvidence	OpenEvidence	OpenEvidence Inc	2023
Eli5a	Research assistant (explainability focused)	A multidisciplinary team	2025
MEKAS[f]	Web-based adaptive assessment system usable on personal computers and mobile devices	Research team at Cathay General Hospital and National University of Tainan	2024

XSL•FO
RenderX

Model	Interface	Developer	Year of release
Llama 3.1 70B	Agentic workflow implemented on LangGraph	Meta (Llama models)	2024
Llama-3-Groq-70B-Tool-Use	API- and tool-based execution (Google Search, PubMed, arXiv)	Groq (tool use model)	2024
Llama-4	Llama	Meta	2025
HeyGen AI video avatar system	Web-based, modular learning platform with prerecorded AI avatar–delivered video modules	HeyGen Inc	2023

[a]ENT: ear, nose, and throat.

[b]MCQ: multiple-choice question.

[c]CEKB: ChatENT's knowledge base.

[d]RAG: retrieval-augmented generation.

[e]Q&A: question and answer.

[f]MEKAS: Multi-Expert Knowledge-Aggregated Adaptive Assessment Scheme.

Critical Appraisal Within Sources of Evidence

No formal critical appraisal of the included sources of evidence was conducted; therefore, no critical appraisal results are reported.

Results of Individual Sources of Evidence

Study-level data relevant to the review question are presented in Table S1 in Multimedia Appendix 3, including study design, purpose, educational context, generative AI system, sample or evaluation set, and thematic contribution. Detailed findings from individual sources that informed the thematic synthesis are presented in Table S2 in Multimedia Appendix 4 through source-linked subthemes and representative quotations.

Synthesis of Results

Overview

Thematic synthesis of the 42 studies [21-62] identified 7 major domains and 213 associated subtheme entries, each linked to representative quotations in Table S2 in Multimedia Appendix 4. The domains were as follows: (1) educational support and learning enhancement; (2) learning outcomes, educational effectiveness, and pedagogical value; (3) model performance, accuracy, and variability; (4) assessment, examination readiness, and question-type sensitivity; (5) reliability, safety, and risks; (6) adoption, usability, and implementation barriers; and (7) advanced AI architectures and future educational design.

Overall, the evidence was dominated by benchmarking and model performance evaluations of knowledge retrieval, question answering, and examination performance, with comparatively limited research on learner-centered outcomes, longitudinal learning, educational interventions, and real-world implementation. The 7 themes broadly encompassed educational applications and outcomes, technical performance and reliability, and implementation and future system design.

Theme 1: Educational Support and Learning Enhancement

Across the included studies, generative AI was consistently portrayed as a flexible and scalable educational support resource that can supplement traditional ENT training by providing accessible, query-responsive support across different learning contexts and levels of training. Reported applications included facilitating professional discussion [27], bridging access to specialized knowledge [29], providing query-tailored explanations for early-stage learners [35], and supporting continuing medical education [46]. In this sense, AI is framed not as a replacement for educators but as a complementary educational tool alongside established learning resources and expert instruction [49,50].

A prominent dimension of educational support was the capacity of generative AI to retrieve, synthesize, and present knowledge efficiently. Advanced systems such as ENTAgents retrieved information from guidelines, textbooks, and academic literature within seconds, supporting efficient learning and revision [47]. ChatENT similarly streamlined access to specialty knowledge and reduced the need to manually search guidelines, textbooks, and academic papers [25]. In audiology, LLMs provided explicit reasoning processes that may help teachers and students rapidly identify key concepts and understand underlying reasoning processes [52]. Together, these findings suggest that the core educational contributions of generative AI lie in optimizing access to, organization of, and reasoning around knowledge rather than merely generating new content.

Beyond information access, several studies emphasized individualized learning support, formative feedback, self-assessment, and targeted remediation. LLMs such as GPT-4o provided immediate feedback on board-style questions and analyzed incorrect responses, helping learners identify specific knowledge gaps and pursue targeted remediation in a self-directed manner [48]. GPT-4 was proposed to generate personalized academic plans tailored to individual learning styles and abilities [29], while other models provided rapid feedback, explanations, and personalized remediation for examination preparation [61]. ChatGPT supported immediate self-assessment through past examination questions [60], while models with moderate-to-high board-question accuracy were proposed as study companions for basic knowledge reinforcement and practice-oriented learning [56]. These functions position generative AI as a mechanism for formative

XSL•FO
RenderX

assessment, adaptive learning, self-assessment, and personalized remediation.

ChatGPT was reported to correctly interpret subtle domain-specific distinctions, including closely related anatomical concepts [27]. In pediatric otorhinolaryngology, AI was described as promising for theoretical learning and standardized clinical scenarios [55]. Beyond knowledge acquisition, emerging evidence suggests a potential role for generative AI in supporting scenario-based and reflective learning. Chatbots were used to explore clinical scenarios [22], while LLMs may function as scalable, case-sensitive reflective prompts rather than authoritative guides, encouraging trainees to reconsider operative plans and reflect on case-specific decisions [59]. Finally, emerging modalities extend educational support beyond text-based interaction. Multimodal systems such as Gemini integrated explanatory images and links into responses, providing additional study resources [53]. AI avatar–based educational platforms enabled rapid production of standardized educational content, reduced faculty workload, and facilitated swift revisions when guidelines changed [54]. Taken together, generative AI was positioned as a multifaceted educational adjunct that can broaden access to specialty knowledge, streamline information retrieval, support reasoning and scenario-based learning, provide personalized feedback and remediation, and extend educational delivery through multimodal and scalable formats. Importantly, its educational role is consistently framed as complementary to established learning resources and expert instruction rather than as a substitute for them [49,50,60].

Theme 2: Learning Outcomes, Educational Effectiveness, and Pedagogical Value

The most direct evidence regarding learning outcomes comes from interventional studies embedding generative AI into structured, competence-adjusted educational systems. In the Multi-Expert Knowledge-Aggregated Adaptive Assessment Scheme (MEKAS) study, trainees receiving AI-assisted, competence-adjusted instruction demonstrated significantly greater improvements in clinical core medical knowledge in otolaryngology compared with those undergoing traditional education. The undergraduate- and postgraduate-year (UPGY) experimental group showed markedly higher score gains than controls, and similar benefits were observed in ENT residents, with no significant difference in educational effects between learner groups [41]. Beyond formal knowledge testing, an AI avatar–delivered ENT induction program significantly improved junior physicians' self-reported confidence in areas including endoscopic anatomy recognition and clinical triage, although acceptability varied [54]. Together, these interventions demonstrated objective knowledge gains and improved learner confidence.

Educational effects were also observed in higher-order cognitive processing and clinical planning. In operative-planning scenarios, LLM-generated feedback altered residents' clinical decision-making and surgical plans [59]. Rather than functioning as a decision substitute, the model served as a reflective aid, prompting learners to compare their plans with AI-generated alternatives [59]. This "reflection loop" suggests a pedagogical role beyond factual knowledge transfer, supporting structured reasoning, self-evaluation, and reflective practice. A second major line of evidence concerns the quality of AI-generated learning materials. LLMs, particularly GPT-4, generated ENT study guides with high accuracy, relevance, and clarity, closely aligned with residents' curricular needs and judged by experts as legitimate educational resources [51]. ChatGPT was also considered potentially useful for problem-solving, critical thinking, and reading comprehension [43].

Importantly, pedagogical value was observed even where summative examination performance was limited. Although ChatGPT-3.5 would likely not have passed the American Board of Otolaryngology Written Question Exam, it provided well-constructed explanations and achieved performance comparable to the lower percentile of otolaryngology residents, suggesting educational benefit through explanation, clarification, and conceptual reinforcement [37]. Thus, educational value may extend beyond pass-fail performance to explanatory quality. AI-generated responses in theoretical and clinical scenarios could achieve quality scores comparable to human specialists [55], while ChatGPT-4o showed examination performance statistically comparable to that of ENT residents in 1 study [57].

Overall, generative AI showed potential to support knowledge acquisition, learner confidence, reflective reasoning, and high-quality educational content [37,41-43,51,54-57,59]. However, the current evidence remains stronger for short-term performance, confidence, and process-oriented outcomes than for sustained trainee-level educational benefit. Further studies are needed to determine whether these promising effects translate into durable gains in learning, clinical reasoning, and real-world training outcomes [40]. Current findings favor structured, pedagogically guided use over unguided standalone application.

Theme 3: Model Performance, Accuracy, and Variability

Overall, the LLMs achieved moderate to high levels of accuracy on otolaryngology-related tasks, although performance varied substantially across datasets and evaluation settings. Reported accuracies ranged from modest levels, such as 53% on otolaryngology in-service practice examination questions, to high rates of comprehensive or correct responses in structured settings [23,26]. ChatGPT achieved passing-level performance across multiple conditions [24,33]; high accuracy and reproducibility were reported for salivary gland diseases [50]; and newer models performed particularly well in standardized knowledge assessments. For example, o3-mini and GPT-4o achieved accuracies of 93.35% and 90.45%, respectively, on otolaryngology United States Medical Licensing Examination–style questions, significantly exceeding the average medical student performance of 55.44% [61]. High answer-explanation concordance and validity were also reported, particularly for foundational anatomy and pathophysiology [24,25,33,60]. Taken together, current LLMs performed reasonably well on selected ENT knowledge tasks, while remaining strongly dependent on task structure and evaluation context.

A consistent observation across studies was the marked variability in performance across subspecialties, question

XSL•FO
RenderX

categories, and difficulty levels [21,48]. LLMs performed better on easier items, whereas more nuanced and in-depth topics posed significant challenges [26]. Similarly, diagnostic tasks were characterized by satisfactory but not comprehensive performance, with models often producing correct answers without consistently providing complete or clinically robust reasoning pathways [31]. Models showed strong performance in theoretical knowledge areas such as cerebrospinal fluid otorrhea, hearing pathophysiology, temporal bone anatomy, and the vestibular system [57] but unsatisfactory performance on auditory physiology questions [22]. Multimodal capability was similarly variable: ChatGPT-4 demonstrated moderately acceptable image-recognition performance [43], while newer GPT-4 models showed substantial improvements on both text- and image-based questions [37]. Thus, accuracy should be interpreted in relation to task complexity, content domain, modality, and reasoning demands, rather than as evidence of generalized competence.

Cross-model comparisons demonstrated rapid generational improvement. Newer models, particularly GPT-4, consistently outperformed GPT-3.5 [37,48]. Standard and custom GPT-4 achieved accuracies of 72% and 81.3%, respectively, compared with approximately 51% to 53% for GPT-3.5 on comparable board-style questions [40], while another study reported 20% improvement in accuracy and 10% improvement in consistency with GPT-4 over GPT-3.5 [22]. In specific domains, such as anatomy learning and examination preparation, GPT-4o achieved particularly high accuracy, exceeding that of DeepSeek, Gemini, and Claude models [62]. However, rankings varied across evaluations: Gemini [48,53] and Bing Copilot [55] were top performers in individual benchmarks, Bard outperformed ChatGPT-3.5 overall and in otology [42], whereas ChatGPT-4 outperformed Bard in skull-base surgery [44]. These rankings reinforce that model superiority is dependent on the comparator, model generation, and evaluation task rather than representing a fixed hierarchy.

Model performance was also multidimensional. Gemini performed well in general otolaryngology question answering and conceptual clarification, whereas GPT-4 was effective for image-based questions and concise, high-accuracy explanations [53]. Claude 2 received the highest ratings for medical adequacy and conciseness, while ChatGPT-4 performed best in covering validated answers [28]. In audiology, ChatGPT generated the most complete responses, whereas Bing Chat performed best for concise relevance [22]. Thus, model performance is multidimensional, and model selection may need to consider the intended educational task and desired response characteristics—including accuracy, completeness, relevance, conciseness, and comprehensibility—rather than overall accuracy alone [22,28].

Variability was also evident in the consistency and stability of model outputs. ChatGPT-4 showed greater consistency than ChatGPT-3.5 [22]. ChatGPT and DeepSeek likewise demonstrated substantial response consistency despite minor fluctuations in an audiologist qualification examination [52], while high reproducibility was reported for ChatGPT responses concerning salivary gland diseases [50]. ChatGPT-5 led a guideline-anchored evaluation in accuracy, comprehensiveness, guideline adherence, and temporal stability [58].

Performance was sensitive to linguistic and structural features of questions. GPT-4 models demonstrated significantly better handling of negated questions than nonnegated ones, whereas Claude 3 Haiku performed poorly on negated questions [48]. Comparable concordance and validity across short-answer and other structured formats further indicated a complex interaction between performance and question structure [25].

Comparisons with human learners and clinicians revealed a more nuanced performance gradient. Earlier evaluations found that LLMs remained inferior to human expertise [21,31], although some were competitive with junior or mid-level residents [35,38]. Conversely, ChatGPT-4 outperformed residents across training years in 1 study, although the performance gap narrowed with increasing residency experience [45], while ChatGPT-4o outperformed residents on selected theoretical knowledge items [57]. Newer models such as o3-mini and GPT-4o also significantly exceeded average medical student performance on standardized knowledge questions [61]. Relative performance of LLMs versus human learners depends substantially on trainee experience, model generation, and the type of competence being assessed; strong performance on knowledge-based questions should therefore not be interpreted as equivalent to specialist-level clinical expertise.

Taken together, contemporary LLMs showed improving accuracy, validity, consistency, and reproducibility, with newer models increasingly reaching or exceeding human learner performance on selected standardized knowledge tasks. However, performance remains heterogeneous across model generations, platforms, subspecialties, modalities, question structures, and reasoning demands. Notably, an accuracy plateau of approximately 73% to 76% among leading general-purpose models has also been reported in specialized otolaryngology knowledge assessment, suggesting persistent limitations without domain-specific adaptation [56]. Thus, high benchmark performance should be interpreted as task-specific capability rather than evidence of uniform competence across the breadth and complexity of otolaryngology [21,24,26,31,37,48,53,56,57,62].

Theme 4: Assessment, Examination Readiness, and Question-Type Sensitivity

Across the included studies, the examination performance of LLMs varied substantially according to model generation, assessment format, and task demands, with evidence ranging from suboptimal accuracy to passing or high performance on selected professional examinations. Accuracy in single-answer multiple-choice questions (MCQs) was often 40% to 60% [27], and ChatGPT answered 51% of otolaryngology board-preparation questions correctly [39]. However, advanced models demonstrated stronger performance: GPT-4 passed selected sleep medicine certification questions [29], ChatGPT-4 achieved 75% accuracy and passed the Taiwan Audiologist Qualification Examination [43], and both ChatGPT and DeepSeek exceeded the 60% passing threshold on a standardized audiologist qualification examination [52]. More recent models showed further gains, with o3-mini maintaining high accuracy

XSL•FO
RenderX

across difficulty levels and outperforming students, particularly on more challenging questions [61]. Thus, examination readiness has improved with newer models but remains model dependent and task dependent.

A particularly robust and reproducible pattern across studies is the pronounced sensitivity of LLM performance to question format. Multiple evaluations demonstrated significantly poorer performance on MCQs compared with single-choice or open-ended formats. In 1 study, ChatGPT achieved correct responses in more than 60% of single-choice questions but only approximately one-third of MCQs [21]. Multiple-choice underperformance was also statistically significant across all tested models ($P<.001$), suggesting that this limitation reflects a general structural weakness rather than a model-specific deficiency [48]. However, later studies indicate that multiple-choice performance is not uniformly poor. ChatENT reduced errors on MCQs by 26% relative to ChatGPT and outperformed ChatGPT-4 on both multiple-choice and short-answer board questions [25], while DeepSeek and ChatGPT achieved near-passing or passing performance on complex MCQs and high accuracy on reverse-logic questions [52]. Guideline-constrained multiple-choice tasks also produced tightly clustered performance across newer models, suggesting that structured questions anchored to established clinical pathways may reduce response drift [58]. Thus, question-type sensitivity appears to interact with model architecture and task structure rather than representing a fixed limitation of all LLMs.

Question difficulty represents another major determinant of assessment performance. A domain-specific ENT GPT assistant demonstrated a pronounced decline in accuracy from 86.8% on the easiest questions to 42.9% on the most difficult questions [46], whereas o3-mini maintained high performance across all difficulty levels, including 100% accuracy at the greatest difficulty level in 1 evaluation [61]. These findings suggest that sensitivity to difficulty remains model dependent and may diminish as model capabilities advance. Assessment performance was also strongly influenced by modality. GPT-4V performed better on text-only than image-based board examination questions, although providing both textual and visual information improved accuracy for image-based items [36]. GPT-4 performed poorly on direct image interpretation and frequently appeared to infer answers from contextual information rather than the images themselves [45], while Bard was unable to process more than half of image-based questions because of file-processing errors [42]. Graph-based questions were also below the passing threshold [52]. Although newer models showed improvement [42], text-based performance may therefore overestimate competence when visual interpretation is required.

Beyond format and modality, factual knowledge performance did not necessarily reflect higher-order clinical reasoning. Board-style questions may predominantly assess factual recall [38]. ChatGPT-4o performed strongly on theoretical knowledge questions but less well on clinical decision-making, contextual evaluation, and patient-specific judgment [57]. Similarly, general-purpose LLMs performed better on rule-based practice-management content than on clinical subspecialty questions requiring more nuanced reasoning [56]. Narrative case prompts also remained more demanding than guideline-constrained questions, particularly with respect to completeness, stepwise management, and calibration [58]. Thus, strong knowledge-based examination performance should not be equated with clinical reasoning or judgment.

This distinction also has implications for assessment design. Several studies proposed that future examinations should place greater emphasis on higher-order competencies, including diagnostic reasoning, clinical judgment, and image interpretation, rather than relying predominantly on factual knowledge questions that current LLMs increasingly answer successfully [38,45]. ChatGPT-generated otorhinolaryngology MCQs produced correct and unambiguous answer keys in 70% of cases and showed high topic relevance and focus; however, vignette and distractor quality were weaker [34]. AI-assisted question generation therefore remains promising, but expert review remains necessary before formal assessments.

Overall, newer LLMs increasingly achieved passing or high performance on selected examinations but remained sensitive to question format, difficulty, modality, and cognitive demands [36,38,40,42,45,46,50,52,56-58]. They may support examination preparation and formative assessment, but benchmark performance may overestimate real-world competence, particularly when assessments emphasize constrained knowledge tasks [56]. Their improving performance also strengthens the rationale for assessments that explicitly evaluate higher-order reasoning and clinically meaningful competencies.

Theme 5: Reliability, Safety, and Risks

A central concern across the included studies was the reliability of generative AI in medical education and clinical knowledge support. Although some evaluations reported high response quality in terms of clarity, relevance, and usefulness, these apparent strengths were consistently undermined by poor source transparency and unreliable citation practices [49,55]. Chatbots produced incorrectly attributed or spurious citations [49], while other studies reported absent source attribution [22] and unreliable or nonexistent references in open-ended responses [32]. Accordingly, successful educational integration requires accurate answers alongside transparent reasoning, reliable sourcing, alignment with curricular goals, and safeguards against outdated or fabricated guidance [60,61].

A major limitation is hallucination. One study reported hallucinated content in more than 13% of open-ended responses, despite apparently coherent and confident outputs [27]. Models also hallucinated during sequential image-based questions, producing similar but incorrect answers across different questions [49]. Additional studies showed that LLMs may produce human-like responses without the contextual reasoning of experienced clinicians [28], may be vulnerable to incorrect or insufficient input information [43], and may generate biased or overly simplified answers when complex medical problems are reduced to fact-based representations [42]. Such errors may be particularly difficult for inexperienced learners to recognize [42].

Safe educational use requires users to retain sufficient biomedical knowledge and critical appraisal skills to recognize

XSL•FO
RenderX

inaccurate or oversimplified outputs. One study specifically warned of intellectual dependency and reduced perceived need to learn complex pathophysiology [40]. This concern is particularly relevant for inexperienced learners, supporting the need for expert supervision to prevent uncritical acceptance of confidently presented inaccuracies [26,42,59]. Beyond educational reliability, errors also posed potential patient safety risks. In 1 example, ChatGPT provided incorrect anatomical information regarding surgical planes for an osteoplastic flap, which could compromise outcomes if applied clinically [33]. Several authors cautioned against the routine use of generative AI for clinical decision-making or for unsupervised educational purposes, particularly in contexts where outputs may influence patient care [23,26]. ChatENT reduced safety-related concerns relative to more general-purpose systems [25], suggesting that knowledge grounding and system design can improve—but not eliminate—safety limitations. Residual errors therefore continue to require educator judgment and oversight when generative AI is incorporated into teaching [52].

Reliability concerns extended to AI-generated educational materials. ChatGPT-generated MCQs were generally grammatically correct but sometimes used medical terminology inaccurately or imprecisely, weakening their reliability for formal assessment [34]. The overall quality of such questions was judged to require thorough expert review and revision before use in examinations [34]. These findings reinforce that linguistic fluency should not be equated with domain accuracy and that AI-generated educational content requires specialist validation before curricular or assessment use. Temporal consistency was another reliability dimension. In a 1-month retest of next-generation LLMs, the leading systems preserved their relative performance ranking and demonstrated limited output drift [58]. Thus, reliability should be conceptualized not only in terms of whether an answer is correct but also whether model behavior remains stable, reproducible, and appropriately aligned with established guidance over time.

Generative AI should not be relied on as a stand-alone authority, especially in high-stakes contexts. Multiple studies explicitly advised against overreliance on AI chatbots for final interpretations, assessments, or clinical judgments, emphasizing that such systems should function only as adjuncts to traditional educational resources and expert supervision [23,49]. Even in controlled educational settings, some studies reported that although peer review of AI-generated scripts can mitigate accuracy risks, the lack of interactivity and emotional or pedagogical nuance still limits educational effectiveness and learner acceptance [54]. This adjunctive model is reinforced by newer studies recommending controlled educational use rather than autonomous clinical decision-making [56,57]. In operative-planning education, a supervised "dual-plan" approach—in which residents first generate an independent plan and subsequently compare it with an LLM-generated alternative—was proposed as a way to preserve reflective and metacognitive benefits while reducing the risk of adopting incorrect suggestions uncritically [59]. These findings favor structured workflows that preserve independent reasoning, human judgment, and expert verification.

Overall, current generative AI systems remain constrained by hallucination, unreliable or absent source attribution, susceptibility to flawed inputs, imprecise terminology, residual error, and limitations in contextual and clinical reasoning. Additional risks include intellectual dependency and the uncritical internalization of confidently presented inaccuracies, particularly among inexperienced learners. Although domain-specific grounding and newer model generations may improve safety and temporal stability, they do not remove the need for expert validation and educator oversight. Accordingly, the available evidence supports the use of generative AI as a supervised and transparently governed educational adjunct rather than as an independent source of medical knowledge or clinical decision-making authority [22,25-28,32,34,38,40,42,43,49,52,55-61].

Theme 6: Adoption, Usability, and Implementation Barriers

A recurrent facilitator of adoption was ease of use, which influenced users' willingness to engage with AI-based educational systems. In the MEKAS intervention, both ENT residents and UPGY trainees reported high perceived ease of use scores, supporting usability as an important determinant of acceptance [41]. Similarly, ENTAgents avoided the need for model fine-tuning by using external knowledge bases and self-reflective prompting, thereby lowering technical barriers to deployment [47]. High reproducibility across otolaryngology topics may further support user trust and implementation [23].

Despite these technical advantages, actual adoption among trainees remains limited. Survey data from Italian ENT residents showed modest ChatGPT use, with only a small proportion of trainees having incorporated it into their academic or scientific activities, despite broad awareness of its potential functions [30]. Implementation may also be constrained by unequal institutional access to advanced models and infrastructure, potentially widening disparities in training opportunities [59]. Current use was concentrated on research-related and low-risk academic support tasks rather than core studying. Trainees reported higher perceived usefulness for activities such as citation searching, content searching, and language revision, whereas studying and examination preparation received lower ratings [30]. Generative AI therefore appears to function more as a productivity and academic support tool than as a central learning instrument, possibly reflecting uncertainty about its pedagogical reliability. Larger multicenter studies incorporating diverse assessment formats and updated AI models have been called for to establish safe and effective strategies for integrating generative AI into medical education [57].

Overall, usability, low technical barriers, and reproducibility support implementation feasibility but do not ensure adoption. Wider pedagogical integration requires equitable access, stronger evidence of educational effectiveness across diverse settings, and clearly defined strategies for safe implementation [23,30,41,47,57,59].

XSL•FO
RenderX

Theme 7: Advanced AI Architectures and Future Educational Design

A prominent direction in the literature was the shift from general-purpose chatbots toward knowledge-grounded and explainable educational systems. The MEKAS platform exemplifies this approach through a knowledge-based AI design that explicitly prioritizes both accuracy and explainability, supporting learning of clinical core medical knowledge in otolaryngology [41]. This reflects a broader shift toward adapting LLMs to domain-specific knowledge and reasoning requirements rather than relying solely on model scale.

Another trend was the integration of retrieval-augmented generation (RAG) and agent-based frameworks. ENTAgents used agents and RAG to supplement LLMs with up-to-date, specific information from curated knowledge bases and external sources, improving accuracy and completeness [47]. The system also incorporated self-correction and reflective reasoning to refine responses and improve factual grounding [47]. Domain-specific customization also improved specialty performance. ChatENT outperformed ChatGPT-4 on both multiple-choice and short-answer board questions [25], while custom GPT-4 models achieved higher accuracy than standard versions [40]. The ENT GPT assistant demonstrated that customization to specialty-specific language, terminology, and clinical nuances improved contextual relevance and accuracy [46]. Together, these findings favor combining knowledge grounding with specialty customization and expert validation rather than relying on general-purpose models alone [25,28,40,42,46]. In multimodal education, improved medical-image performance may require specialized prompting, advanced image-recognition technologies, and training specifically focused on medical images [36]. Domain refinement, real-time learning mechanisms, and improved interpretability have also been proposed [38].

Generative AI may also support pedagogical designs involving automated generation of case scenarios, differential diagnosis exercises, interactive learning modules, standardized patient simulations, and assessment questions and rubrics, all of which can support scalable, consistent, and resource-efficient instructional design across teaching sessions [49]. Customizable synthetic cases may further expand these applications [29], while virtual-patient scenarios may allow residents and students to practice diagnostic and therapeutic skills in a safe simulation environment [35]. AI has also been proposed for integration into examination simulations that test basic knowledge recall and allow residents to assess gaps in foundational knowledge [45]. In this context, AI is framed not merely as a content generator but as an infrastructure for supporting structured learning workflows, including review, concept reinforcement, board-style question practice, and rapid knowledge recall [49]. Its ability to interact with learners, provide explanations, and respond to user customization also creates opportunities for more interactive and adaptive educational experiences [40].

These systems may further promote active learning and higher-order cognitive skills. AI-generated assessment questions can be used to encourage active engagement rather than passive text review, a strategy that aligns with established principles of memory retention and learning science [49]. At the same time, instructors are encouraged to guide learners to critically appraise AI-generated responses and identify reasoning errors, supporting analytical reasoning, reflective judgment, and diagnostic rigor [49]. In operative-planning tasks, residents were reported to engage with LLM outputs reflectively rather than passively, particularly when AI-generated recommendations differed from their own plans [59]. Future frameworks may therefore promote metacognition by requiring learners to formulate, compare, critique, and revise their reasoning rather than passively accept AI-generated answers [59], while supervised guideline-linked systems may support case discussion and formative assessment [58]. At the architectural level, the literature further suggests that future progress may require explicit support for medical reasoning rather than continued increases in general model scale alone. A reported performance plateau among general-purpose LLMs suggests that adaptation to medical reasoning patterns may represent a more important next step [56]. Hybrid systems combining LLMs with structured knowledge bases or clinical decision trees may provide explicit reasoning frameworks for complex cases [56].

Overall, the evidence suggests a transition from general-purpose conversational AI toward hybrid, domain-adapted, and pedagogically governed educational systems integrating knowledge grounding, retrieval augmentation, specialty customization, structured reasoning, simulation, adaptive interaction, and reflective learning [25,28,29,35,36,38,40-42,45,46,56,58,59]. These approaches position generative AI less as an autonomous answer generator and more as a component of explainable, adaptive, and educationally aligned personalized learning platforms.

Discussion

Summary of Evidence

This scoping review suggests that generative AI is increasingly being explored as an educational support tool in otolaryngology. However, the current evidence remains early, heterogeneous, and largely dominated by benchmarking studies rather than direct evaluations of educational effectiveness. Across the included studies, generative AI systems—particularly LLMs—were repeatedly shown to support knowledge retrieval, content synthesis, revision, assessment generation, and scenario-based learning, consistent with broader observations in medical education and clinical informatics [2]. This trajectory mirrors wider developments in digital medicine, where AI is increasingly positioned not merely as an automation tool but as a cognitive and organizational scaffold supporting learning, decision-making, and personalized knowledge pathways [4]. In ENT education specifically, the breadth of subspecialization, anatomical complexity, and heterogeneity of clinical scenarios make the promise of scalable, adaptive, and just-in-time educational support particularly compelling. Importantly, the reviewed literature suggests that personalization extends beyond content adaptation to include immediate feedback, identification of individual knowledge gaps, targeted remediation, self-assessment, and alternative learning pathways, thereby creating opportunities for more adaptive and learner-responsive

XSL•FO
RenderX

educational environments, as demonstrated in competence-adjusted educational systems such as MEKAS [41]. Beyond professional education, generative AI is also being explored for patient-facing support in otolaryngology. Studies in postoperative cochlear implant care and tinnitus have reported favorable expert evaluations of AI-generated information, suggesting a potential adjunctive role in supervised patient education [63,64].

Contemporary generative models demonstrated moderate to high performance on structured knowledge tasks, especially in the domains of examination-style questions, anatomy, and guideline-based content. GPT-4 and GPT-4o can reach or exceed passing thresholds and, in some contexts, even outperform junior trainees [45,48]. Similar performance gains have been observed in anatomy education [3]. More recent findings within the reviewed literature further indicate that this human-AI performance relationship is increasingly dependent on the model generation, learner experience, and type of competence being assessed: newer systems may equal or exceed medical students or residents on selected standardized knowledge tasks while continuing to perform less reliably on more complex clinical reasoning tasks [45,57]. Performance therefore remains multidimensional and task dependent rather than reflecting generalized competence [65].

Crucially, this performance heterogeneity is not merely a technical issue but a pedagogical and epistemic one. Performance generally becomes less reliable as tasks require greater complexity, multimodal interpretation, contextual integration, or clinical reasoning, although question format effects are increasingly model dependent rather than uniformly disadvantageous across all contemporary LLMs [26,53]. Thus, strong knowledge-task performance should not be interpreted as a surrogate for expert clinical reasoning, consistent with broader concerns that linguistic fluency may conceal reasoning gaps [8]. This distinction has important implications for assessment design. As LLMs become increasingly capable of answering conventional factual and guideline-based examination questions, traditional knowledge-focused assessments may become less informative measures of uniquely human clinical competence. The reviewed evidence instead supports greater emphasis on diagnostic reasoning, contextual judgment, patient-specific decision-making, image and graph interpretation, and other higher-order competencies [45,57]. Generative AI may therefore influence not only examination preparation but also what medical assessments are designed to measure.

Safety, trustworthiness, and epistemic governance therefore emerge as central concerns. Multiple patient- and clinician-facing studies demonstrate that LLMs continue to hallucinate, misattribute citations, and produce confidently stated but incorrect content [66,67]. An additional risk is intellectual dependency, whereby reliance on AI-generated explanations may reduce the perceived need to master biomedical knowledge, although background knowledge and critical thinking remain necessary to detect erroneous outputs [40]. This problem may be particularly relevant for novice learners, who may simultaneously derive substantial benefit from explanatory support while being less able to detect confidently presented inaccuracies [26,40]. These concerns support assistive, auditable use within frameworks of human oversight [10].

Importantly, the reviewed literature suggests that oversight should function as an active pedagogical strategy rather than merely a final accuracy check. In operative-planning education, residents benefited from comparing their own independently generated plans with LLM-generated alternatives, creating a reflective “dual-plan” process in which disagreement stimulated reconsideration of assumptions and clinical decisions [59]. Structured human-AI interaction may therefore preserve metacognitive benefits while reducing uncritical acceptance of erroneous recommendations [59]. Adoption patterns remain cautious and uneven, with higher acceptance for research assistance and content drafting than for core learning or assessment [1]. Even when AI-based instructional systems demonstrate learning gains [41,68], scalability, generalizability, sustainability, and long-term curricular integration remain largely unresolved. Perceived reliability, demonstrated educational value, institutional guidance, and clearly defined roles within curricula are likely to be equally important [1,41].

Equity is another concern because unequal access to advanced models, infrastructure, and specialist resources may widen disparities in training opportunities [59]. Scalability should therefore not be assumed to imply equitable scalability. An important emerging direction is the shift from stand-alone chatbots toward domain-specific, retrieval-augmented, and knowledge-grounded architectures. Systems such as ChatENT demonstrate that hallucination, inconsistency, and domain drift may be partially mitigated through curated knowledge bases, structured retrieval, and constrained generation [69]. This design philosophy aligns closely with emerging recommendations in trustworthy AI and clinical-grade system development [5]. Specialty-specific customization, agent-based orchestration, self-reflective verification, and structured knowledge representations can further improve contextual relevance and specialty performance, supporting a shift from generic conversational models toward systems explicitly adapted to otolaryngology knowledge and educational objectives [5,25]. This direction may become particularly important as general-purpose LLM performance appears to plateau on specialist knowledge tasks [56]. Hybrid architectures incorporating structured knowledge and reasoning frameworks may therefore represent an important next step, while persistent image-interpretation limitations indicate the need for modality-specific optimization [36,45].

From an educational theory perspective, the greatest long-term impact of generative AI may lie not in passive content delivery but in enabling active learning, formative assessment, adaptive feedback, and reflective practice. AI-generated cases, questions, and study guides already show high perceived clarity and relevance [70], and AI avatars demonstrate the feasibility of scalable, standardized induction training [54]. These applications align closely with modern competency-based education models, in which iterative feedback loops, deliberate practice, and self-regulated learning are more central than static content consumption. The reviewed findings extend this concept by suggesting that the key educational unit may be the human-AI interaction rather than the AI-generated answer itself. When

XSL•FO
RenderX

learners are required to formulate, compare, critique, and revise their own reasoning in response to model output, AI may support metacognition and reflective learning rather than merely accelerate information retrieval [59]. This approach is particularly relevant to personalized learning platforms because adaptation can be driven not only by what information learners request but also by their knowledge gaps, errors, reasoning patterns, and responses to feedback [41,59].

At the same time, the literature consistently emphasizes that generative AI cannot be treated as an autonomous epistemic authority. Studies evaluating citation quality, clinical reasoning, and evidence alignment show that expert validation remains indispensable [71,72]. This reinforces a core principle emerging across medical AI: the objective is not simply to automate existing educational activities or replace human educators but to design forms of human-AI collaboration that extend educational reach and responsiveness while preserving independent reasoning, professional judgment, and accountability [5,10,59].

Taken together, the available evidence suggests that generative AI in ENT education remains at an early stage of development. Most studies have focused on technical performance and benchmarking, whereas evidence on educational effectiveness, longitudinal learning outcomes, and real-world curricular implementation remains limited. This scoping review extends the existing literature by moving beyond reviews centered primarily on AI applications or model performance to provide an education-focused synthesis linking generative AI capabilities, learning outcomes, assessment, safety, implementation, and emerging architectures to the design of personalized learning in otolaryngology. Its principal contribution is a conceptual shift from evaluating standalone chatbots toward developing supervised, pedagogically governed, knowledge-grounded, and adaptive learning systems that preserve learner reasoning and expert oversight. This framework has practical implications for educators, curriculum designers, and developers: generative AI can be incorporated into formative assessment, individualized feedback, targeted remediation, and other personalized learning activities, while autonomous or high-stakes use should remain constrained by expert verification and appropriate governance. Accordingly, future research should move beyond benchmark-driven demonstrations toward longitudinal educational outcomes, curriculum-level integration, higher-order reasoning, equitable implementation, and governance addressing safety, transparency, accountability, and human-AI responsibility. This trajectory mirrors the broader evolution of AI in clinical medicine, where early enthusiasm is increasingly balanced by demands for robustness, transparency, and real-world impact [73].

Limitations

The evidence base has several limitations. First, studies were heterogeneous and predominantly focused on benchmarking and short-term outcomes, with limited longitudinal evidence regarding knowledge retention, transfer to clinical reasoning, or impact on real-world clinical performance. The current literature also provides limited evidence regarding whether repeated use of generative AI influences learners' independent reasoning, metacognitive skills, or longer-term reliance on external cognitive support. Second, many evaluations rely on artificial test environments or curated question banks, which may not fully capture the cognitive and contextual complexity of authentic clinical learning. This limitation is particularly relevant because the reviewed evidence indicates substantially stronger model performance on constrained, theoretical, or guideline-based tasks than on clinical reasoning, contextual decision-making, and multimodal interpretation. Benchmark accuracy may therefore overestimate educational or clinical competence in real-world settings. Third, the rapid evolution of LLMs means that findings can become partially obsolete within months, limiting the temporal generalizability of performance comparisons between specific model versions. Moreover, differences in access to newer models and supporting infrastructure may limit the generalizability of implementation studies across institutions and may introduce inequities that have received comparatively little empirical investigation.

This review is also constrained by the quality and scope of the available literature. The synthesis is necessarily descriptive and conceptual rather than quantitative, and no formal critical appraisal or effectiveness meta-analysis was possible, given the heterogeneity of study designs. Finally, while we attempted to contextualize ENT-specific findings within the broader AI-in-medical-education discourse, this review did not aim to exhaustively cover the rapidly expanding interdisciplinary literature in this area.

Conclusions

This scoping review demonstrates that generative AI is evolving from a general-purpose information and content-generation tool toward a broader educational infrastructure capable of supporting knowledge retrieval, formative assessment, immediate feedback, targeted remediation, scenario-based learning, and reflective practice in otolaryngology. Newer models can achieve strong or even human-comparable performance on selected standardized knowledge tasks; however, this capability remains highly dependent on task structure and should not be equated with proficiency in clinical reasoning, contextual judgment, or multimodal interpretation. Persistent risks—including hallucination, unreliable sourcing, confidently incorrect outputs, and learner overreliance—mean that generative AI should remain a supervised educational adjunct rather than an autonomous source of medical knowledge or clinical decision-making authority.

The findings further suggest that the future of AI-enabled personalized learning is unlikely to depend on model scale or answer accuracy alone. More promising approaches combine domain-specific knowledge grounding, retrieval augmentation, structured reasoning support, and pedagogically designed human-AI interaction in which learners remain active participants in generating, comparing, and revising their own reasoning. In this framework, personalization involves not only adapting content but also responding to individual knowledge gaps, errors, and learning trajectories while preserving independent judgment and metacognition. As AI becomes increasingly proficient at conventional knowledge-based assessment, educational research should also consider whether

XSL•FO
RenderX

assessment strategies need to place greater emphasis on higher-order reasoning, clinical judgment, and multimodal competence. Future studies should therefore prioritize longitudinal learning outcomes, real-world curricular integration, equitable implementation, and governance models that address transparency, accountability, safety, and the appropriate division of responsibility between learners, educators, and AI systems.

Acknowledgments

The authors wish to thank Sixiu Zhao, PhD, Department of Radiology, The First Affiliated Hospital, Zhejiang University School of Medicine, for review of an early draft of this manuscript.

Disclosure of Delegation to Generative AI

The authors declare the use of generative AI in the research and writing process. According to the GAIDeT taxonomy (2025), the following tasks were delegated to GAI tools under full human supervision:

- Adapting and adjusting emotional tone

The GAI tool used was: ChatGPT-5.2.

Responsibility for the final manuscript lies entirely with the authors.

GAI tools are not listed as authors and do not bear responsibility for the final outcomes.

Declaration submitted by: Collective responsibility

Additional note: The authors used ChatGPT (OpenAI) for language-editing assistance in this manuscript.

Funding

This review was funded by the Medical and Health Research Project of Zhejiang Province (grant 2023RC009) and AI-Empowered Medical Education Research Project of The First Affiliated Hospital, Zhejiang University School of Medicine (grant zyjg202509).

Authors' Contributions

JTZ conceived and designed the study. JTZ, ZZC, and SHZ jointly developed the review protocol and search strategy. JTZ and ZZC performed the literature search, study screening, and data extraction. SHZ resolved disagreements and provided methodological oversight. All 3 authors (JTZ, ZZC, and SHZ) contributed to data interpretation and thematic synthesis. JTZ drafted the initial manuscript, and ZZC and SHZ critically revised the manuscript for important intellectual content. All authors reviewed and approved the final version of the manuscript and agree to be accountable for all aspects of the work.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Generative AI for personalized learning platforms in otolaryngology–head and neck surgery education: scoping review protocol.
[DOCX File, 22 KB - mededu_v12i1e91785_app1.docx]

Multimedia Appendix 2
Full search strategy.
[DOCX File, 16 KB - mededu_v12i1e91785_app2.docx]

Multimedia Appendix 3
Summary of included studies and identified themes (N=42).
[DOCX File, 78 KB - mededu_v12i1e91785_app3.docx]

Multimedia Appendix 4
Major themes identified, associated subthemes, and representative quotations.
[DOCX File, 218 KB - mededu_v12i1e91785_app4.docx]

Checklist 1
PRISMA-ScR checklist.
[DOCX File, 72 KB - mededu_v12i1e91785_app5.docx]

XSL•FO
RenderX

References

1. Brennan L, Balakumar R, Bennett W. The role of ChatGPT in enhancing ENT surgical training - a trainees' perspective. J Laryngol Otol 2024 May;138(5):480-486. [doi: 10.1017/S0022215123001354] [Medline: 37583313]
2. Briganti G. How ChatGPT works: a mini review. Eur Arch Otorhinolaryngol 2024 Mar;281(3):1565-1569. [doi: 10.1007/s00405-023-08337-7] [Medline: 37991499]
3. Bolgova O, Mavrych V. Evolution of AI in anatomy education study based on comparison of current large language models against historical ChatGPT performance. Sci Rep 2025 Oct 28;15(1):37545. [doi: 10.1038/s41598-025-22437-w] [Medline: 41152541]
4. Bousquet J, Schünemann HJ, Sousa-Pinto B, et al. Concepts for the development of person-centered, digitally enabled, artificial intelligence-assisted ARIA care pathways (ARIA 2024). J Allergy Clin Immunol Pract 2024 Oct;12(10):2648-2668. [doi: 10.1016/j.jaip.2024.06.040] [Medline: 38971567]
5. Ahn J, Kang BG, Chang M, Yoon S. Applications and future perspectives of large language models in otolaryngology-head and neck surgery: a comprehensive survey. Clin Exp Otorhinolaryngol 2025 Nov;18(4):283-295. [doi: 10.21053/ceo.2025-00121] [Medline: 40590109]
6. Irfan B. Beyond the scope: advancing otolaryngology with artificial intelligence integration. Cureus 2024 Feb;16(2):e54248. [doi: 10.7759/cureus.54248] [Medline: 38496161]
7. Xiao J, Wolter NE, Davies JC, Propst EJ, Crowson MG, Siu JM. Artificial intelligence in surgical training and applications to otolaryngology: a scoping review. Laryngoscope 2025 Oct;135(10):3520-3534. [doi: 10.1002/lary.32246] [Medline: 40371996]
8. Rao KN, Fernandez-Alvarez V, Guntinas-Lichius O, et al. The limitations of artificial intelligence in head and neck oncology. Adv Ther 2025 Jun;42(6):2559-2568. [doi: 10.1007/s12325-025-03198-4] [Medline: 40299277]
9. Zhong NN, Wang HQ, Huang XY, et al. Enhancing head and neck tumor management with artificial intelligence: integration and perspectives. Semin Cancer Biol 2023 Oct;95:52-74. [doi: 10.1016/j.semcancer.2023.07.002] [Medline: 37473825]
10. Chen S, Lobo BC. Regulatory and implementation considerations for artificial intelligence. Otolaryngol Clin North Am 2024 Oct;57(5):871-886. [doi: 10.1016/j.otc.2024.04.007] [Medline: 38839554]
11. Lechien JR. Generative AI and otolaryngology-head & neck surgery. Otolaryngol Clin North Am 2024 Oct;57(5):753-765. [doi: 10.1016/j.otc.2024.04.006] [Medline: 38839556]
12. Edalati S, Vasan V, Cheng CP, Patel Z, Govindaraj S, Iloreta AM. Can GPT-4 revolutionize otolaryngology? Navigating opportunities and ethical considerations. Am J Otolaryngol 2024;45(4):104303. [doi: 10.1016/j.amjoto.2024.104303] [Medline: 38678799]
13. Crowson MG, Nwosu OI. The integration and impact of artificial intelligence in otolaryngology-head and neck surgery: navigating the last mile. Otolaryngol Clin North Am 2024 Oct;57(5):887-895. [doi: 10.1016/j.otc.2024.04.001] [Medline: 38705741]
14. Buess L, Keicher M, Navab N, Maier A, Tayebi Arasteh S. From large language models to multimodal AI: a scoping review on the potential of generative AI in medicine. Biomed Eng Lett 2025 Sep;15(5):845-863. [doi: 10.1007/s13534-025-00497-1] [Medline: 40917142]
15. Lechien JR. Generative artificial intelligence in otolaryngology–head and neck surgery editorial: be an actor of the future or follower. Eur Arch Otorhinolaryngol 2024 Apr;281(4):2051-2053. [doi: 10.1007/s00405-024-08579-z]
16. Litsou E, Tzotzis I, Malisovas T, Asimakopoulos A, Psychogios G. Artificial intelligence in otolaryngology: current applications, limitations, and future perspectives. Eur Arch Otorhinolaryngol 2026 Sep 16. [doi: 10.1007/s00405-026-10560-x] [Medline: 42745060]
17. Teo ZL, Thirunavukarasu AJ, Elangovan K, et al. Generative artificial intelligence in medicine. Nat Med 2025 Oct;31(10):3270-3282. [doi: 10.1038/s41591-025-03983-2] [Medline: 41053447]
18. Tricco AC, Lillie E, Zarin W, et al. PRISMA extension for scoping reviews (PRISMA-ScR): checklist and explanation. Ann Intern Med 2018 Oct 2;169(7):467-473. [doi: 10.7326/M18-0850] [Medline: 30178033]
19. Devlin J, Chang MW, Lee K, Toutanova K. BERT: pre-training of deep bidirectional transformers for language understanding. Presented at: Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers); Jun 2-7, 2019. [doi: 10.18653/v1/N19-1423]
20. Rethlefsen ML, Kirtley S, Waffenschmidt S, et al. PRISMA-S: an extension to the PRISMA statement for reporting literature searches in systematic reviews. Syst Rev 2021 Jan 26;10(1):39. [doi: 10.1186/s13643-020-01542-z] [Medline: 33499930]
21. Hoch CC, Wollenberg B, Lüers JC, et al. ChatGPT's quiz skills in different otolaryngology subspecialties: an analysis of 2576 single-choice and multiple-choice board certification preparation questions. Eur Arch Otorhinolaryngol 2023 Sep;280(9):4271-4278. [doi: 10.1007/s00405-023-08051-4] [Medline: 37285018]
22. Jedrzejczak WW, Kochanek K. Comparison of the audiological knowledge of three chatbots – ChatGPT, Bing Chat, and Bard. Otolaryngology. Preprint posted online on May 6, 2024. [doi: 10.1101/2023.11.22.23298893]
23. Kuşcu O, Pamuk AE, Sütay Süslü N, Hosal S. Is ChatGPT accurate and reliable in answering questions regarding head and neck cancer? Front Oncol 2023;13:1256459. [doi: 10.3389/fonc.2023.1256459] [Medline: 38107064]
24. Long C, Lowe K, Santos AD, et al. Evaluating ChatGPT-4 in otolaryngology–head and neck surgery board examination using the CVSA model. Otolaryngology. Preprint posted online on 2023. [doi: 10.1101/2023.05.30.23290758]

XSL•FO
RenderX

25. Long C, Subburam D, Lowe K, et al. ChatENT: augmented large language model for expert knowledge retrieval in otolaryngology - head and neck surgery. SSRN. Preprint posted online on 2023. [doi: 10.2139/ssrn.4571725]
26. Mahajan AP, Shabet CL, Smith J, Rudy SF, Kupfer RA, Bohm LA. Assessment of artificial intelligence performance on the otolaryngology residency in-service exam. OTO Open 2023;7(4):e98. [doi: 10.1002/oto2.98] [Medline: 38034065]
27. Zalzal HG, Cheng J, Shah RK. Evaluating the current ability of ChatGPT to assist in professional otolaryngology education. OTO Open 2023;7(4):e94. [doi: 10.1002/oto2.94] [Medline: 38020045]
28. Buhr CR, Smith H, Huppertz T, et al. Assessing unknown potential-quality and limitations of different large language models in the field of otorhinolaryngology. Acta Otolaryngol 2024 Mar;144(3):237-242. [doi: 10.1080/00016489.2024.2352843] [Medline: 38781053]
29. Cheong RCT, Pang KP, Unadkat S, et al. Performance of artificial intelligence chatbots in sleep medicine certification board exams: ChatGPT versus Google Bard. Eur Arch Otorhinolaryngol 2024 Apr;281(4):2137-2143. [doi: 10.1007/s00405-023-08381-3] [Medline: 38117307]
30. Dallari V, Liberale C, De Cecco F, et al. The role of artificial intelligence in training ENT residents: a survey on ChatGPT, a new method of investigation. Acta Otorhinolaryngol Ital 2024 Jun;44(3):161-168. [doi: 10.14639/0392-100X-N2806] [Medline: 38712520]
31. Dallari V, Sacchetto A, Saetti R, Calabrese L, Vittadello F, Gazzini L. Is artificial intelligence ready to replace specialist doctors entirely? ENT specialists vs ChatGPT: 1-0, ball at the center. Eur Arch Otorhinolaryngol 2024 Feb;281(2):995-1023. [doi: 10.1007/s00405-023-08321-1] [Medline: 37962570]
32. Kochanek K, Skarzynski H, Jedrzejczak WW. Accuracy and repeatability of ChatGPT based on a set of multiple-choice questions on objective tests of hearing. Cureus 2024 May;16(5):e59857. [doi: 10.7759/cureus.59857] [Medline: 38854312]
33. Long C, Lowe K, Zhang J, et al. A novel evaluation model for assessing ChatGPT on otolaryngology-head and neck surgery certification examinations: performance study. JMIR Med Educ 2024 Jan 16;10:e49970. [doi: 10.2196/49970] [Medline: 38227351]
34. Lotto C, Sheppard SC, Anschuetz W, et al. ChatGPT generated otorhinolaryngology multiple-choice questions: quality, psychometric properties, and suitability for assessments. OTO Open 2024;8(3):e70018. [doi: 10.1002/oto2.70018] [Medline: 39328276]
35. Mete U. Evaluating the performance of ChatGPT, Gemini, and Bing compared with resident surgeons in the otorhinolaryngology in-service training examination. Turk Arch Otorhinolaryngol 2024 Oct 23;62(2):48-57. [doi: 10.4274/tao.2024.3.5] [Medline: 39463066]
36. Noda M, Ueno T, Koshu R, et al. Performance of GPT-4V in answering the Japanese Otolaryngology Board Certification examination questions: evaluation study. JMIR Med Educ 2024 Mar 28;10:e57054. [doi: 10.2196/57054] [Medline: 38546736]
37. Patel EA, Fleischer L, Filip P, et al. Comparative performance of ChatGPT 3.5 and GPT4 on rhinology standardized board examination questions. OTO Open 2024;8(2):e164. [doi: 10.1002/oto2.164] [Medline: 38938507]
38. Patel J, Robinson P, Illing E, Anthony B. Is ChatGPT 3.5 smarter than otolaryngology trainees? A comparison study of board style exam questions. PLoS One 2024;19(9):e0306233. [doi: 10.1371/journal.pone.0306233] [Medline: 39325705]
39. Revercomb L, Patel AM, Choudhry HS, Filimonov A. Performance of ChatGPT in otolaryngology knowledge assessment. Am J Otolaryngol 2024;45(1):104082. [doi: 10.1016/j.amjoto.2023.104082] [Medline: 37862879]
40. Revercomb L, Patel AM, Fu D, Filimonov A. Performance of novel GPT-4 in otolaryngology knowledge assessment. Indian J Otolaryngol Head Neck Surg 2024 Dec;76(6):6112-6114. [doi: 10.1007/s12070-024-04935-x] [Medline: 39559040]
41. Su JM, Hsu SY, Fang TY, Wang PC. Developing and validating a knowledge-based AI assessment system for learning clinical core medical knowledge in otolaryngology. Comput Biol Med 2024 Aug;178:108765. [doi: 10.1016/j.compbiomed.2024.108765] [Medline: 38897143]
42. Terwilliger E, Bcharah G, Bcharah H, Bcharah E, Richardson C, Scheffler P. Advancing medical education: performance of generative artificial intelligence models on otolaryngology board preparation questions with image analysis insights. Cureus 2024 Jul;16(7):e64204. [doi: 10.7759/cureus.64204] [Medline: 39130878]
43. Wang S, Mo C, Chen Y, Dai X, Wang H, Shen X. Exploring the performance of ChatGPT-4 in the Taiwan Audiologist Qualification Examination: preliminary observational study highlighting the potential of AI chatbots in hearing care. JMIR Med Educ 2024 Apr 26;10:e55595. [doi: 10.2196/55595] [Medline: 38693697]
44. Abgin Y, Umemoto K, Goulian A, et al. Accuracy and completeness of Bard and Chat-GPT 4 responses for questions derived from the International Consensus Statement on Endoscopic Skull-Base Surgery 2019. J Neurol Surg B Skull Base 2025 Dec;86(6):688-693. [doi: 10.1055/a-2436-4222] [Medline: 41140422]
45. Ajit-Roger E, Kazerooni KS, Savage J, Nguyen LHP. Performance of artificial intelligence chatbot on the Canadian otolaryngology in-training exam: unlocking insights on the intersection of technology and education. J Otolaryngol Head Neck Surg 2025;54:19160216251387618. [doi: 10.1177/19160216251387618] [Medline: 41369028]
46. Bicknell BT, Rivers NJ, Skelton A, et al. Domain-specific customization for language models in otolaryngology: the ENT GPT assistant. OTO Open 2025;9(2):e70125. [doi: 10.1002/oto2.70125] [Medline: 40331108]
47. Chan TK, Dinh ND. ENTAgents: AI agents for complex knowledge otolaryngology. Otolaryngology. Preprint posted online on 2025. [doi: 10.1101/2025.01.01.25319863]

XSL•FO
RenderX

48. Hoch CC, Funk PF, Guntinas-Lichius O, et al. Harnessing advanced large language models in otolaryngology board examinations: an investigation using Python and application programming interfaces. Eur Arch Otorhinolaryngol 2025 Jun;282(6):3317-3328. [doi: 10.1007/s00405-025-09404-x] [Medline: 40281318]
49. Khoury ZH, Sultan AS. Performance of AI chatbots on head and neck pathology board-style exam questions and guidelines for responsible use. Head Neck Pathol 2025 Nov 3;19(1):123. [doi: 10.1007/s12105-025-01850-7] [Medline: 41182557]
50. Küçüktağ Z, Altan E, Saygı Uysal G, Özdoğan A. Are ChatGPT's answers to questions about salivary gland diseases accurate and reliable? Turk Arch Otorhinolaryngol 2025 Sep 26;63(3):143-150. [doi: 10.4274/tao.2025.2025-1-17] [Medline: 40977134]
51. Patel EA, Herrmann PT, Fleischer L, et al. Comparative analysis of AI-generated study guides in otolaryngology education. Am J Otolaryngol 2025;46(5):104693. [doi: 10.1016/j.amjoto.2025.104693] [Medline: 40554991]
52. Qi B, Zheng Y, Wang Y, Xu L. Comparison of ChatGPT and DeepSeek on a standardized audiologist qualification examination in Chinese: observational study. JMIR Form Res 2025 Nov 28;9:e79534. [doi: 10.2196/79534] [Medline: 41313805]
53. Ramchandani R, Guo E, Mostowy M, et al. Comparison of ChatGPT-4, Copilot, Bard and Gemini Ultra on an otolaryngology question bank. Clin Otolaryngol 2025 Jul;50(4):704-711. [doi: 10.1111/coa.14302] [Medline: 40081900]
54. Siddique H, Maweni R. AI avatar-delivered ear nose throat (ENT) induction: a pilot feasibility study of confidence and acceptability. Cureus 2025 Oct;17(10):e94230. [doi: 10.7759/cureus.94230] [Medline: 41209936]
55. Trecca EMC, Caponio VCA, Turri-Zanoni M, et al. Comparative analysis of information quality in pediatric otorhinolaryngology: clinicians, residents, and large language models. Otolaryngol Head Neck Surg 2025 Jul;173(1):228-236. [doi: 10.1002/ohn.1225] [Medline: 40105497]
56. Cook R, Kahan A, Scharfenberger T, et al. Comparing large language models' performances on otolaryngology knowledge assessment questions. Appl Clin Inform 2026 Mar;17(2):194-203. [doi: 10.1055/a-2835-4634] [Medline: 41941887]
57. Doğan Karataş T, Aksoy A, Bora A, Doğan M. The use of artificial intelligence in medical education: a comparative analysis of theoretical exam performance between ENT residents and ChatGPT-4o. Lokman Hekim Health Sci 2026;6(2):196-202. [doi: 10.14744/lhhs.2024.94858]
58. Günsoy B, Tümüklü K. Benchmarking next-generation large language models in otorhinolaryngology: a prospective, guideline-anchored, blinded evaluation of ChatGPT-5, Gemini-3, Grok-2, Llama-4, and DeepSeek-4 across case scenarios. Eur Arch Otorhinolaryngol 2026 Apr 29. [doi: 10.1007/s00405-026-10264-2] [Medline: 42050126]
59. Hack S, Hayu D, Karni JE, et al. Impact of generative AI on operative planning in otolaryngology residents. Surg Innov 2026 Apr 29;29:15533506261446833. [doi: 10.1177/15533506261446833] [Medline: 42051212]
60. Motaghi Niko M, Karbasi Z, Kazemi M, Zahmatkeshan M. Examining the performance of ChatGPT in comprehensive pre-internship exam: the potential of artificial intelligence in medical education. Health Sci Rep 2026 Jan;9(1):e71492. [doi: 10.1002/hsr2.71492] [Medline: 41523853]
61. Prasad S, Travis L, Thornton M, Thaller S. Comparison of GPT-4o and o3-mini on otolaryngology USMLE-style questions. J Craniofac Surg 2026;37(3-4):827-830. [doi: 10.1097/SCS.0000000000011831] [Medline: 40920769]
62. Tassoker M. Who knows anatomy best? A comparative study of ChatGPT-4o, DeepSeek, Gemini, and Claude. Clin Anat 2026 Jan;39(1):25-29. [doi: 10.1002/ca.70012] [Medline: 40708277]
63. Aliyeva A, Sari E, Alaskarov E, Nasirov R. Enhancing postoperative cochlear implant care with ChatGPT-4: a study on artificial intelligence (AI)-assisted patient education and support. Cureus 2024 Feb;16(2):e53897. [doi: 10.7759/cureus.53897] [Medline: 38465158]
64. Alaskarov E, Aliyeva A. Artificial intelligence -assisted tinnitus support with ChatGPT-4. Eur Arch Otorhinolaryngol 2026 Feb;283(2):773-779. [doi: 10.1007/s00405-025-09719-9] [Medline: 41053456]
65. Murphy Lonergan R, Curry J, Dhas K, Simmons BI. Stratified evaluation of GPT's question answering in surgery reveals artificial intelligence (AI) knowledge gaps. Cureus 2023 Nov;15(11):e48788. [doi: 10.7759/cureus.48788] [Medline: 38098921]
66. Wei K, Fritz C, Rajasekaran K. Answering head and neck cancer questions: an assessment of ChatGPT responses. Am J Otolaryngol 2024;45(1):104085. [doi: 10.1016/j.amjoto.2023.104085] [Medline: 37844413]
67. Karimov Z, Allahverdiyev I, Agayarov OY, Demir D, Almuradova E. ChatGPT vs UpToDate: comparative study of usefulness and reliability of chatbot in common clinical presentations of otorhinolaryngology-head and neck surgery. Eur Arch Otorhinolaryngol 2024 Apr;281(4):2145-2151. [doi: 10.1007/s00405-023-08423-w] [Medline: 38217726]
68. Xu G, Jia D, Wang X, Chen J, Pan H, Wu Z. Construction and implementation of otolaryngology endoscopic training based on artificial intelligence. Ear Nose Throat J 2025 Sep 13;13:1455613251371794. [doi: 10.1177/01455613251371794] [Medline: 40944567]
69. Long C, Subburam D, Lowe K, et al. ChatENT: augmented large language model for expert knowledge retrieval in otolaryngology-head and neck surgery. Otolaryngol Head Neck Surg 2024 Oct;171(4):1042-1051. [doi: 10.1002/ohn.864] [Medline: 38895862]
70. Shaari AL, Herzog I, Kokush E, Zabielski S, Langer S, Filimonov A. Exploring the ability of ChatGPT to act as a research aid in otolaryngology. Indian J Otolaryngol Head Neck Surg 2024 Dec;76(6):6100-6105. [doi: 10.1007/s12070-024-05101-z] [Medline: 39558999]

XSL•FO
RenderX

71. Nielsen JPS, Mikkelsen AK, Kuenzel J, et al. Evaluation of multiple-choice tests in head and neck ultrasound created by physicians and large language models. Diagnostics (Basel) 2025 Jul 22;15(15):1848. [doi: 10.3390/diagnostics15151848] [Medline: 40804813]
72. Ayo-Ajibola O, Davis RJ, Lin ME, et al. TrachGPT: appraisal of tracheostomy care recommendations from an artificial intelligent chatbot. Laryngoscope Investig Otolaryngol 2024 Aug;9(4):e1300. [doi: 10.1002/lio2.1300] [Medline: 39015552]
73. Evangelista E, Bensoussan Y. Standardization, collaboration, and education in the implementation of artificial intelligence in otolaryngology: the key to scalable impact. Otolaryngol Clin North Am 2024 Oct;57(5):897-908. [doi: 10.1016/j.otc.2024.04.005] [Medline: 38845298]

Abbreviations

ENT: ear, nose and throat
LLM: large language model
MCQ: multiple-choice question
MEKAS: Multi-Expert Knowledge-Aggregated Adaptive Assessment Scheme
NCBI: National Center for Biotechnology Information
PRISMA-S: Preferred Reporting Items for Systematic Reviews and Meta-Analyses literature search extension
PRISMA-ScR: Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews
RAG: retrieval-augmented generation
UPGY: undergraduate- and postgraduate-year medical students

Edited by S Brini; submitted 20.Jan.2026; peer-reviewed by A Aliyeva, PC Wang; revised version received 17.Sep.2026; accepted 18.Sep.2026; published 19.Oct.2026.

Please cite as:
Zhong JT, Cao ZZ, Zhou SH
Generative AI for Personalized Learning Platforms in Otolaryngology–Head and Neck Surgery Education: Scoping Review
JMIR Med Educ 2026;12:e91785
URL: https://mededu.jmir.org/2026/1/e91785
doi:10.2196/91785

© Jiang-Tao Zhong, Zai-Zai Cao, Shui-Hong Zhou. Originally published in JMIR Medical Education (https://mededu.jmir.org), 19.Oct.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Education, is properly cited. The complete bibliographic information, a link to the original publication on https://mededu.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Test-Retest Reliability of Large Language Models and Their Concordance With Human Experts in Rubric-Based Scoring of Clinical Summary Statements: Experimental Comparative Study

Patrick Stadler[1], MSc; Piotr Bucichowski[1], MD; Ylva Holzhausen[1], Dr phil; Inga Hege[2], MD, MCompSc; Harm Peters[1], MD, MHPE; Julius Josef Kaminski[1], MD, MHPE

[1]Charité – University Medical Center Berlin, Dieter Scheffner Center for Medical Education and Educational Research, Dean's Office for Study Affairs, Charitéplatz 1, Berlin, Germany

[2]Faculty of Health Sciences Brandenburg, Institute of Research in Health Sciences Education, Brandenburg Medical School, Neuruppin, Germany

Corresponding Author:
Harm Peters, MD, MHPE
Charité – University Medical Center Berlin, Dieter Scheffner Center for Medical Education and Educational Research, Dean's Office for Study Affairs, Charitéplatz 1, Berlin, Germany

Abstract

Background: Clinical summary statements convert unstructured patient information into a concise format. They are widely used in virtual patients (VPs) to train medical students in diagnostic reasoning. Most VP systems provide little individualized feedback on the quality of learners' summaries, and earlier automated approaches have required complex natural language processing, limiting their scalability. Large language models (LLMs) may offer a more flexible, rubric-guided way to score summary statements and support feedback. Still, their test-retest reliability and concordance with expert ratings in this context remain unclear.

Objective: This study aims to evaluate the test-retest reliability—specifically, whether the models produce stable ratings across repetitions—of LLM-generated medical summary statement ratings, and to measure their concordance with human expert assessments across different prompting strategies.

Methods: We conducted an experimental comparative study using 122 rubric-scored summary statements (English and German) from 6 internal medicine VP cases previously rated by 2 expert raters. Statements were evaluated along 4 ternary components (semantic qualifiers, transformation, narrowing, and global rating), and 2 binary components (factual accuracy and person). Four LLMs (Anthropic's Claude, OpenAI's GPT-4o, Mistral AI's Mistral Large, and Mistral AI's Mistral 7B) were each run 5 times under 4 prompting strategies (zero-shot, zero-shot with chain-of-thought [CoT], few-shot, and few-shot+CoT). The test-retest reliability was assessed with intraclass correlation coefficients (ICCs), Fleiss kappa (κ), and concordance with human ratings with accuracy (%) and chance-corrected interrater agreement (Gwet AC1 and AC2).

Results: Large closed-weights models (Claude, GPT-4o, and Mistral Large) showed high test-retest reliability (ICC≥0.92; Fleiss κ=0.59 - 1.00). Claude was essentially deterministic (ICC and Fleiss κ≥0.99); GPT-4o and Mistral Large reached ICC≥0.97 and Fleiss κ≥0.80 except under zero-shot. Mistral 7B, a small open-weights model, had high ICCs (≥0.88) but low Fleiss κ (0.32 - 0.52). Few-shot+CoT prompting achieved the best concordance with human ratings for all the models (means across runs and rubric components), peaking for Mistral Large (accuracy 77%; Gwet AC1 and AC2 0.85, 95% CI 0.83 - 0.87), followed by Claude (71%; 0.77, 0.74 - 0.81), GPT-4o (67%; 0.72, 0.68 - 0.77), and Mistral 7B (60%; 0.64, 0.56 - 0.73). Concordance was higher for the binary rubric components (person and factual accuracy) than for the ternary components. Extreme deviations and invalid outputs were rare but more frequent in ternary components and the smaller local model.

Conclusions: Large, general-purpose LLMs can function as reliable, rubric-guided raters of VP-based clinical summary statements, achieving high concordance with expert ratings under clearly defined criteria in binary rubric components and when using few-shot with CoT prompting. The performance on more nuanced, ternary reasoning dimensions was more variable. LLMs can support scalable, rubric-based formative feedback on clinical reasoning within VP systems but should complement rather than replace human judgment.

(JMIR Med Educ 2026;12:e94666) doi:10.2196/94666

KEYWORDS

medical education; large language models; automated assessment; formative feedback; clinical reasoning; virtual patients; summary statements; problem representation

XSL•FO
RenderX

Introduction

Clinical summary statements are a key element in physicians' clinical reasoning. As concise, structured narratives, they facilitate transforming unstructured patient information into an initial diagnostic interpretation and considering relevant underlying differential diagnoses [1,2]. They make observable implicit parts of the reasoning processes that underpin diagnostic processes, and over time, they can be used to evidence the development of clinical reasoning competence. For example, rather than simply noting that "Mr. Miller came to the ER with several chest pain episodes during the last 2 hours," a more advanced student would write, "Mr. Miller is a 62-year-old male with cardiovascular risk factors presenting with acute, substernal chest pressure radiating to the left arm, associated with diaphoresis, and relieved by rest." This transformation from descriptive data to diagnostic framing exemplifies a key cognitive step in diagnostic reasoning [3]. However, evaluating the quality of these summary statements reliably and at scale remains a challenge in medical education, as many learners simultaneously need individualized, timely feedback.

Clinical summary statements are also referred to as problem representations or one-liners. Depending on the context, they can take various forms and serve different functions, such as a brief patient handover at the end of one's shift or a more extensive case presentation at a clinical conference [4]. For this paper, we refer to clinical summary statements as extracts of patient information formulated by students during their learning process. Feblowitz et al [4] suggested a conceptual model that explains any clinical information summary as a sequence of 5 steps: (1) aggregation and (2) organization of clinical information, (3) reduction and transformation, (4) interpretation, and (5) synthesis. In educational research on summary statements, particular attention has been given to the last 3 steps of the sequence. Well-developed and effective summary statements include both positive and negative key findings, use semantic qualifiers (SQs) (eg, acute, radiating, and substernal), and thereby contextualize the presentation within epidemiological and demographic parameters; by doing so, they offer synthesized, essential case information within an abstracted, diagnostic framework [5,6]. Accordingly, assessing summary statements extends beyond assessing the factual correctness of the information presented by the learner. Rather, the evaluation should include how effectively learners structure information and attribute meaning to it. By sharing their interpretations and abstractions of patient information in summary statements, learners offer insight into how well they can integrate specific patient data into their illness scripts (ie, the knowledge of illnesses represented in their long-term memory [7]), the accuracy of their illness scripts, and how this process facilitates their overall clinical reasoning [1,8]. Diagnostic accuracy depends on these processes and can therefore be considered a summarizing indicator of clinical expertise [9].

Virtual patients (VPs) have become an established approach for medical students to learn clinical reasoning. These interactive, computer-based simulations allow them to train diagnostic processes in a safe, standardized, and scalable environment [10]. VPs enable repeated exposure to varied and even rare presentations and facilitate reflective, self-paced learning. Despite these advantages, providing students with reliable, individualized feedback on reasoning quality remains challenging. A recent review by Jay et al [11] revealed that among studies with learner feedback in VP systems, only a small fraction addressed feedback on the quality of clinical summary statements, and even fewer operationalized them as formative feedback for learning. This gap is relevant because feedback is a primary driver of learning in health professions education [12,13]. This points to a missed opportunity for medical students to learn in current educational approaches.

To address the gap in feedback provision, first VP studies with efforts to automate the evaluation of summary statements showed promising results: Hege et al [14] implemented a rule-based natural language processing (NLP) system grounded in a rubric by Smith et al [6], which assessed the following domains: SQ, transformation of information into higher-order concepts, diagnostic narrowing, factual accuracy, and overall quality in summary statements. Although the system achieved moderate agreement with medical expert ratings, extensive task-specific engineering, language-dependent dictionaries, and substantial training data were needed. Each new VP case demanded additional manual development, limiting scalability and adaptability. More fundamentally, rule-based systems sometimes struggled to capture the contextual nuances of clinical reasoning. Moreover, they can fail because of simple vocabulary mismatches (cf [15]), underscoring the need for approaches that interpret the meaning of summary statements rather than relying solely on pattern matching.

Recent advances in large language models (LLMs) offer a fundamentally different approach to the challenges encountered in many traditional rule-based NLP systems. LLMs process text through learned representations of semantic and contextual relationships, enabling nuanced interpretation without handcrafted rules [16,17]. LLMs also demonstrate better handling of inferences, the transformation of information, or different wording or phrasing that often fail with plain lexical lookups [18]. Their ability to generalize across content domains suggests the potential for scalable, context-aware evaluation of learner input. Nevertheless, the performance of LLMs depends on how the models are instructed to perform a task by prompting. Prompting refers to how instructions are structured to guide LLM behavior. Prompts are typically structured into (1) task instructions, (2) contextual information such as item stems or rubrics, (3) optional examples, and (4) an output specification. Well-thought-out prompt designs—particularly regarding the inclusion and phrasing of rubrics—have been shown to affect both the accuracy and reliability of automated scoring outcomes in medical education settings such as in Objective Structured Clinical Examination (OSCE) grading [19-21]. In zero-shot prompting, the model receives only task instructions (eg, "Score this text using the rubric and the scoring instructions below") without examples. Few-shot prompting extends this by including a few examples of the task, allowing the model to infer the scoring logic from demonstration (eg, text examples paired with rubric scores given by raters). Chain-of-Thought (CoT) prompting, which adds a reasoning

XSL•FO
RenderX

step to either few-shot or zero-shot prompting, can be combined with few-shot prompting to provide the LLM with human raters' rationales for some exemplary human ratings (Methods section).

Despite emerging and promising evidence from automated assessments, no study has yet systematically examined the test-retest reliability of LLMs and their concordance with human expert ratings in evaluating clinical summary statements. Understanding whether LLMs can apply established rubrics is essential for determining whether using them for formative assessment in diagnostic-reasoning education is feasible. An LLM-based evaluation could help address the feedback gap identified by Jay et al [11] in current VP systems. Instead of relying on occasional, labor-intensive assessments by clinical teachers, learners could benefit from repeated VP-based practice, with LLMs providing immediate, consistent, expert-level feedback on their summary statements throughout medical training, thereby strengthening their clinical reasoning skills.

The aim of this study is to evaluate the performance of 4 transformer architecture–based, widely available LLMs with multilingual support in a rubric-guided evaluation of an existing set of clinical summary statements written by medical students during their interactions with online VPs.

Specifically, we formulate the following 2 research questions (RQs):

1. What is the test-retest reliability of LLM-generated medical summary statement ratings across 5 independent runs?
2. What is the concordance of LLM ratings with human expert ratings under different prompting configurations?

Methods

Study Design

Using an experimental comparative design, this study examined the test-retest reliability of LLMs in rubric-based ratings of medical summary statements under multiple prompting conditions and their concordance with human expert ratings.

Because LLM performance is highly dependent on the prompting techniques chosen [22], 4 prompting strategies (zero-shot and few-shot prompting, each with and without CoT) were applied. This resulted in a total of 16 experimental conditions across the 4 LLMs (4 prompting configurations × 4 LLMs). To better compare the results with those of the previous work on the automated evaluation of summary statements by Hege et al [14], the same summary statements from their study were used, as well as the evaluation rubric by Smith et al [6], which was annotated for LLM use (refer to the Rating Rubric and Prompt Development section).

Each of these combinations of models, prompt configurations (eg, zero-shot), and summary statements was run 5 times.

Models and Infrastructure

Four off-the-shelf general-purpose LLMs were tested: 3 large commercial models (Anthropic's Claude Sonnet 4, OpenAI's GPT-4o, and Mistral AI's Mistral Large) and one smaller local model (Mistral AI's Mistral 7B). The models were selected purposefully to cover the range of deployment options at the time of the study: 3 widely used high-performing commercial models from different vendors and a smaller open-weight model that can be deployed locally. Specialized domain-specific models for medical purposes, such as Google's Med-PaLM [23] and MedAlpaca [24], or BERT (Bidirectional Encoder Representations from Transformers)-like models, such as MedBERT [25], exist. However, these models are currently either less accessible to the public and/or not mature or stable enough for reliable or automated assessment, which were 2 key research aspects of this work; thus, these specialized models were not included in our study. The local Mistral 7B model was obtained from Hugging Face, a community-driven website that shares language models and tools. We used an activation-aware weight quantization (AWQ) variant of Mistral 7B, a compressed version that is optimized for lower hardware profiles while still keeping its core features [26].

An overview of the models and specifications is provided in Table 1.

Table . Overview of the various LLMs[a] that were used and compared in our study.

Provider	Model	Exact version
Anthropic	Claude Sonnet 4	claude-sonnet-4 - 20250514
OpenAI	GPT-4o	gpt-4o-2024-05-13
Mistral AI	Mistral Large	mistral-large-2411
Mistral AI	Mistral 7B	TheBloke/Mistral-7B-Instruct-v0.2-AWQ (via Hugging Face)

[a]LLMs: large language models.

Data Source

The dataset from Hege et al [14] comprising 125 rubric-scored summary statements in English (n=84) and German (n=41) was used for this study. The summary statements were written by medical students in response to 6 human-written virtual patient cases in internal medicine from open-access clinical reasoning courses in Instruct's CASUS [27] VP system. They varied in length and quality, ranging from near-perfect to incomplete summaries. In a study by Hege et al [14], the statements were scored independently by 2 health care professionals with an educational background and expertise in clinical reasoning. The ratings of the statements were based on an elaborated rubric by Smith et al [6]. To calibrate their use of the rubric, the 2 experts independently rated and discussed 25 sample summaries before rating the full set of 125. Divergent scores were discussed and rating consensus was reached for all summary statements [14].

XSL•FO
RenderX

For this study, 3 of the 125 summary statements available were excluded from the current analysis because they were used as examples in few-shot prompting conditions, leaving 122 statements for evaluation.

Rating Rubric and Prompt Development

For systematic evaluation, we applied the rubric developed by Smith et al [6] as adapted by Hege et al [14], an established assessment tool with prevalidated quantitative scoring scales, consisting of binary and ternary items. This instrument was used directly as our rating standard and for developing corresponding prompt instructions. The rubric comprises 6 criteria (components): (1) SQ, (2) transformation of information, (3) narrowing of differential diagnosis, (4) factual accuracy, (5) patient name or person (added by Hege et al [14]), and (6) a global score (a separate, subjective overall score). The rubric components no. 4 and 5 were binary (scoring: 0/1), whereas the rest were ternary (with 3 possible scorings: 0, 1, or 2).

Four prompting strategies were implemented and assessed, following established LLM practices from the prompting literature [19,22,28]:

1. Zero-shot, no CoT
2. Zero-shot with CoT
3. Few-shot, no CoT
4. Few-shot with CoT

Zero-shot prompting relied on the model's pretrained knowledge without rating examples, whereas in few-shot prompting, we included 3 manually picked summary statements of our dataset and annotated them with human expert ratings, to serve as rating examples representative for our dataset. They were chosen specifically to exemplify all possible proficiency levels (0/1 in binary and 0/1/2 in ternary rubric components) across 3 different VP cases: 2 in English and 1 in German. CoT prompting encouraged stepwise reasoning by appending the phrase "Let's think step by step" (zero-shot CoT). In few-shot prompting, CoT was additionally supported by guided step-by-step annotations on how to score provided examples [19]. On the basis of the prompting condition, different prompt segments were included, as summarized in Table 2. The exact prompt texts in Markdown format are provided in Multimedia Appendix 1. All additional data, such as the exact case vignettes, can be found in our GitHub repository (Data Availability section).

Table . Composition of each prompt depending on the prompting condition. The prompts contained different prompt segments.

Prompt segment	Zero-shot	Zero-shot CoT[a]	Few-shot	Few-shot CoT
Role description	✓[b]	✓	✓	✓
Scoring rubric	✓	✓	✓	✓
VP[c] case	✓	✓	✓	✓
Three scoring examples	◦	◦[d]	✓	✓
Case information for exemplary human case scorings	◦	◦	✓	✓
Rationales behind examples	◦	◦	◦	✓
Student response placeholder where summary statement was inserted	✓	✓	✓	✓
CoT initiator	◦	✓	◦	✓

[a]CoT: chain-of-thought.

[b]✓ denotes a segment included in prompt

[c]VP: virtual patient.

[d]◦ denotes segment not included in prompt.

The prompts were iteratively refined in the process as described by Lee et al [19]:

The rubric content and assessment rationales were formatted as a prompt and then reviewed by the study authors, including experts in educational research and AI (PB, PS, and IH). This resulted in a more detailed description of SQ based on Chang et al [1] and minor linguistic changes (eg, case vignettes → clinical cases). For the initial iterative test runs, the LLMs were prompted to provide explanatory ratings (ie, rationales) to identify cardinal ambiguities, which led to further prompt refinement (PB, PS). This optimization was performed mainly with Mistral 7B, the smallest model (ie, fewest parameters) in our study, to ensure robustness across models, under the assumption that larger models would properly process prompts modified to work with the smallest model [29]. Explanations provided by LLMs generally matched given scores, eg, when the patient name was scored with 1, the model would provide an adequate justification that the statement included the patient name. During this step, we identified that insufficient formatting led Mistral 7B and Claude Sonnet 4 to systematically confuse VP cases with the summary statement to be rated, which resulted in increased use of special characters to provide a clear structure in crucial areas of the prompt (such as the use of brackets and delimiters). Based on the ratings of summary statements excluded from the analysis, we refined the description of the scoring rubric (PB, PS, and JJK). Lastly, the final step of prompt development was to instruct all models to solely output the raw scores without further rationales.

XSL•FO
RenderX

A comparative overview of the original scoring rubric and the interpretative information and rationales that we added to minimize ambiguity for LLM interpretation are presented in Multimedia Appendix 2.

Procedure

The experiments were conducted between July and September 2025 using the latest available models at that time. The 5 runs per model and prompting condition were executed consecutively, and a new run automatically started as soon as the previous run completed, resulting in a narrow time window of around 1 to 2 hours in which the data were collected per experiment condition, depending on the model and prompting condition.

The local Mistral 7B model was deployed using vLLM [30] and was run on a single NVIDIA Tesla T4 GPU via our institution's compute cloud. All other models (Claude, GPT-4o, and Mistral Large) were accessed via their official APIs using the same LangChain setup to ensure comparability between the commercial online models and the local model. No conversation memory was implemented, preventing learning effects between reruns of ratings. Each of the 122 summary statements was rated under all 16 experimental conditions across 6 rubric components, which was repeated 5 times, yielding a total of 58,560 data points collected as scores across all 6 rubric components.

We implemented custom software to automate the process using the Python programming language (Python Software Foundation) with the LangChain v0.3.26 package, a software abstraction that handles LLM inference. For each model vendor, we used the respective LangChain submodule to access the APIs. To ensure data integrity, we formulated a strict data format and passed each LLM call to a structured output function that automatically ensures that the LLMs' responses follow that format. Our custom software procedure included (1) resolving all placeholders in the prompt with the data for each individual run (the summary statement to be rated and the virtual patient case vignette; 3 examples for few-shot conditions; annotations for the examples, and a CoT reasoning initiator for CoT conditions), (2) sending the complete prompt to the LLM API via LangChain, (3) obtaining the raw results from the LLM API, (4) transforming the data into a structured format, and (5) adding the data to a CSV file for later statistical analysis. Temperature, a model-specific hyperparameter, was set to zero for all models for more deterministic results. It should be noted, though, that even a temperature of zero does not guarantee fully deterministic results when working with closed APIs [31]. This is also captured further in the Results section (see RQ1 for test-retest reliability), where we report on the test-retest reliability between consecutive runs.

Our GitHub repository includes the exact code, prompts, scoring materials, and original human reference ratings to replicate the study (see Data Availability section).

Data Analysis

Statistical analysis was performed using R v4.5.0 [32]. The complete R and Python codes used for this study can be found in our GitHub repository (see the Data Availability section).

We analyzed the resulting dataset from 2 complementary perspectives.

To address RQ1, we assessed test-retest reliability by examining whether the models produced consistent and stable ratings across the 5 run repetitions, measured as (1) consistency using intraclass correlation coefficients (ICCs) and (2) intrarater agreement using Fleiss κ (effectively treating each run as an independent measurement by the same LLM rater). The ICC was calculated using a one-way random-effects model based on the average of 5 raters according to [33]. By examining the consistency of the 5 aggregated runs, this approach was chosen to examine whether random model variances affect a model's scoring behavior, even when the API temperature was set to zero. The ICC is used to capture consistency in numeric scoring, whereas Fleiss κ is designed for exact categorical agreement. Both complement each other well in our setup: the meaning of score distances is not entirely uniform across rubric components (eg, a change from 0 to 1 does not necessarily equal a change from 1 to 2); thus, both numeric consistency and categorical agreement perspectives provide information on the degree of systematic bias.

To address RQ2, we assessed concordance with human expert ratings using 2 metrics: (1) accuracy (percentage of exact LLM-human match agreement) and (2) interrater agreement measured using Gwet AC1 and AC2 coefficients (chance-corrected indices) for comparison with prior work. Following the recommendations for robust agreement assessment, we prioritized Gwet AC1 and AC2 as our primary statistic, as suggested by Xu and Lorber [34]. AC1 is a suitable metric for binary (categorical) rubric components, whereas AC2 provides a measure for ordinal components, in our case, ternary rubric components [35,36]. Conventional indices such as Cohen kappa and Krippendorff α can underestimate agreement in cases where the distribution of ratings is imbalanced (eg, when most statements are factually correct, as in our dataset). This prevalence problem is well documented and can lead to artificially low reliability scores despite strong observed agreement. Gwet AC1 and AC2 correct for this bias [35,37] and are therefore considered more robust for skewed datasets [38,39]. For each combination of model, prompting configuration, and rubric component, Gwet AC1 and AC2 and weighted Cohen kappa coefficients were calculated separately for each of the 5 runs. For results per rubric component, we report the mean, SD, and SE across these 5 runs. For overall results per model and prompting configuration, we report the mean, SD, and SE across all 30 coefficients (6 rubric components×5 runs). The 95% CIs were calculated using a Wald-type normal approximation (mean±1.96×SE).

In accordance with prior research, we applied the Landis and Koch scale [40] to interpret the raw Gwet AC1 and AC2 values. However, this interpretation needs to be treated with some care, as Gwet AC1 and AC2 tend to produce higher values than Cohen kappa does [37]. To allow a full comparison between all the models, prompting conditions, and rubric components, we included the full values of Gwet AC1 and AC2 and weighted Cohen kappa–for a more complete comparison with related work–over all 5 runs in Multimedia Appendices 3 and 4.

XSL•FO
RenderX

Ethical Considerations

The secondary data analysis conducted in this study was approved by the Ethics Committee of the Ludwig-Maximilians-Universität München, Germany (reference number 260–15), with the primary data having been generated by Hege et al [14]. An additional informed consent was not required for the anonymous analysis of the data. The virtual patients provided to the learners were fictitious and not based on real individuals.

Results

RQ1: Test-Retest Reliability

The ICCs indicate high consistency across 5 repeated runs for all LLMs (ICC≥0.88), with varying levels of intrarater agreement depending on the model and prompting condition (Fleiss κ=0.32 - 1.00), as shown in Table 3.

Figure 1 shows the deviations between model and human ratings, as well as the number of invalid outputs (eg, numerical ratings outside the rubric scale), for the first of the 5 runs. The maximum possible deviation was ±2 points for ternary components (0-2 scale) and ±1 point for binary components (0-1 scale). Enlarged versions of the panels of Figure 1 can be found in the Multimedia Appendices 5-8.

Table . ICC[a] and Fleiss κ for all the model × prompt configurations. Claude was fully stable across all runs, whereas Mistral 7B had weaker Fleiss κ values despite a high ICC.

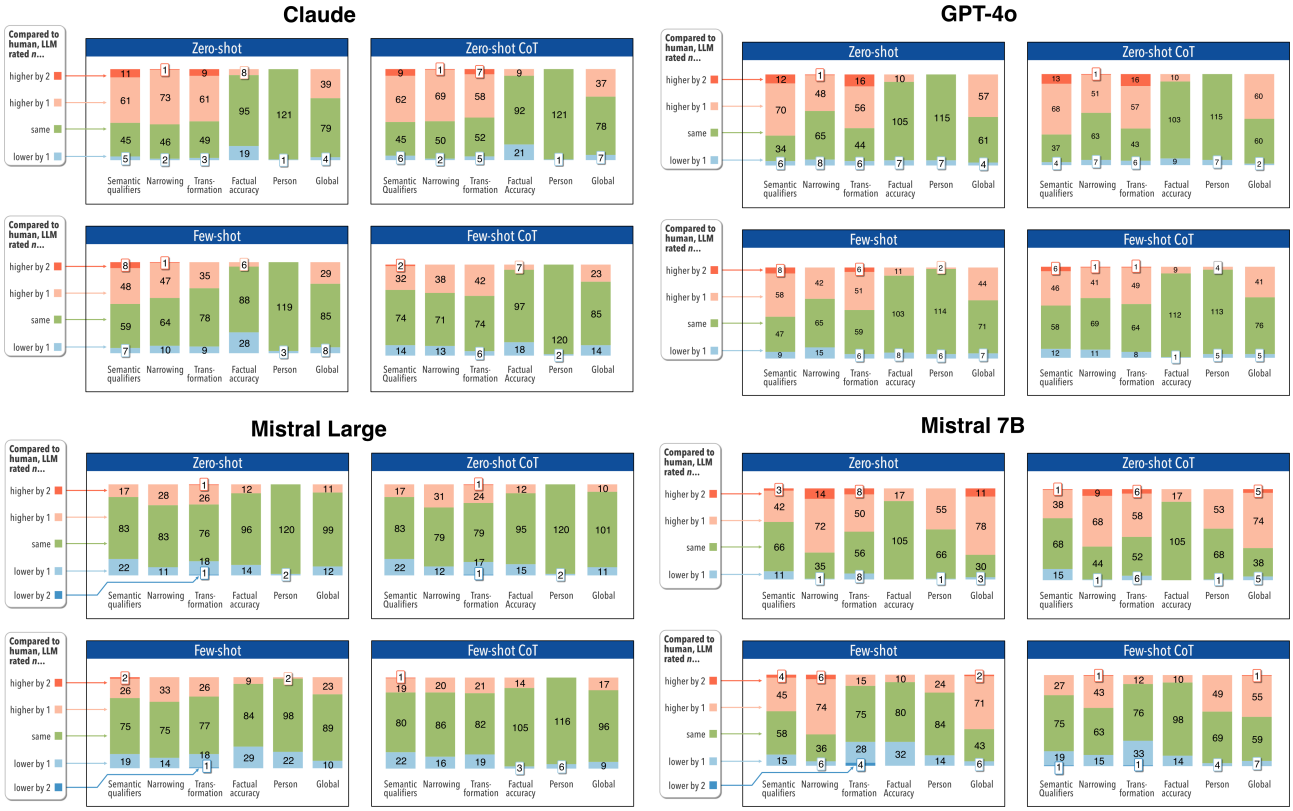
Configuration	ICC	Fleiss κ
Mistral Large zero-shot	0.92	0.66
Mistral Large few-shot	0.97	0.80
Mistral Large zero-shot CoT[b]	0.98	0.91
Mistral Large few-shot CoT	0.97	0.80
Claude zero-shot	1.00	1.00
Claude few-shot	0.99	0.99
Claude zero-shot CoT	0.99	0.99
Claude few-shot CoT	0.99	0.99
GPT-4o zero-shot	0.92	0.59
GPT-4o few-shot	0.99	0.94
GPT-4o zero-shot CoT	0.99	0.94
GPT-4o few-shot CoT	0.99	0.92
Mistral 7B zero-shot	0.88	0.32
Mistral 7B few-shot	0.91	0.43
Mistral 7B zero-shot CoT	0.89	0.39
Mistral 7B few-shot CoT	0.92	0.52

[a]ICC: intraclass correlation coefficient.

[b]CoT: chain-of-thought.

XSL•FO
RenderX

Figure 1. Predicted large language model ratings per model, prompting configuration, and rubric component. The stacked bars show the absolute number of equal (same) ratings, as well as ratings that were lower by 1 or 2 or higher by 1 or 2 compared with the human reference. The results of the first of the 5 runs per model and configuration were used for this figure. We have included enlarged versions of the individual panels from this figure in Multimedia Appendices 5-8 to make them easier to read. CoT: chain-of-thought; LLM: large language model.



Claude was the most reliable: zero-shot runs were essentially deterministic (ICC=1.00, Fleiss κ=1.00), whereas all other conditions were near-deterministic (ICC≥0.99, Fleiss κ≥0.99). For the other models, zero-shot runs yielded the lowest ICC and Fleiss κ values. GPT-4o (excluding zero-shot) achieved very high test-retest reliability (ICC≥0.99 and Fleiss κ≥0.92). Mistral Large (also excluding zero-shot) showed lower values but nevertheless high consistency and intrarater agreement (ICC≥0.97, Fleiss κ≥0.80).

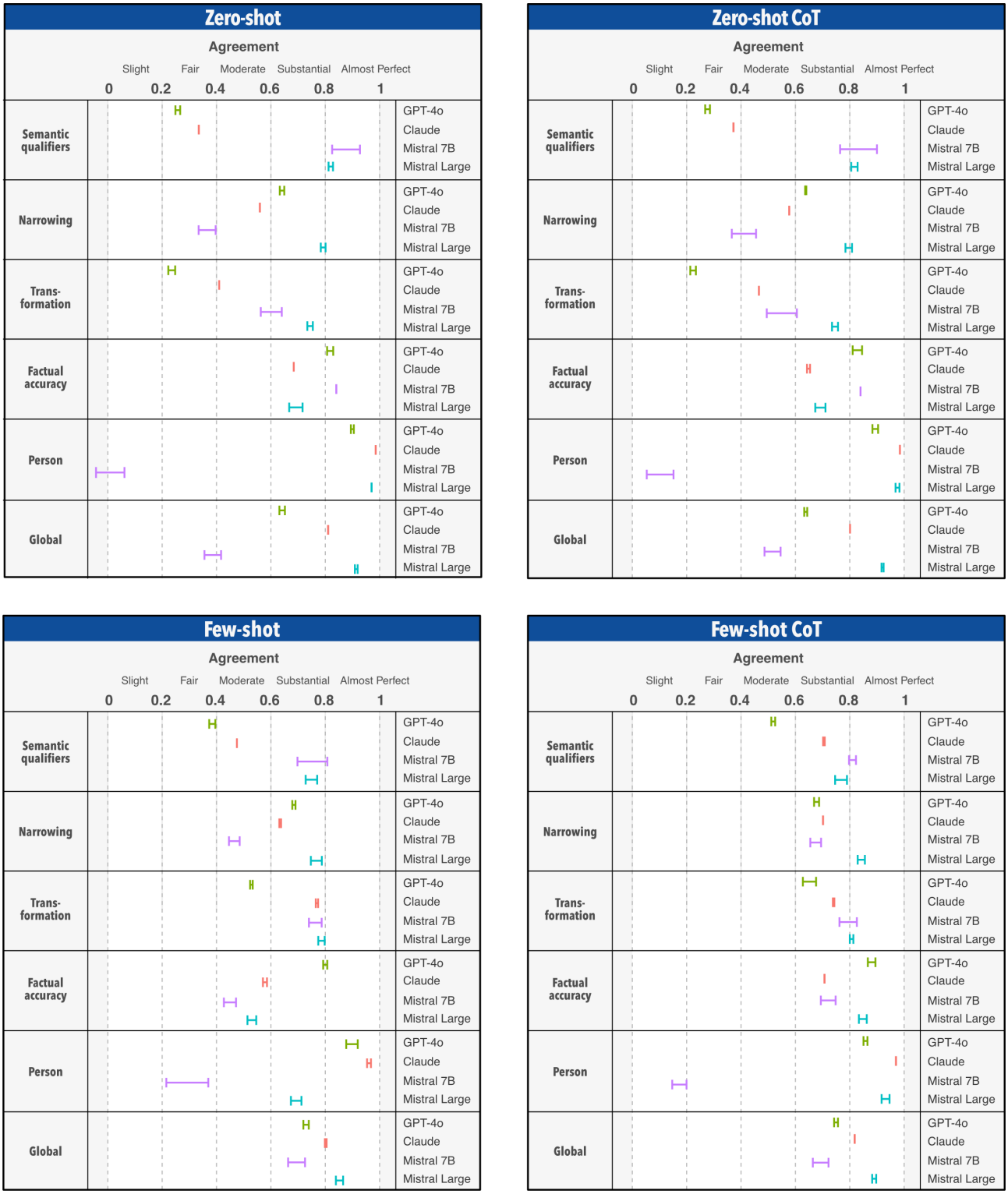
In zero-shot settings, both GPT-4o and Mistral Large showed high consistency (ICC=0.92) and moderate to substantial intrarater agreement (Fleiss κ=0.59 and =0.66, respectively). The lowest values were reported for Mistral 7B, with high ICC (≥0.88) but only fair-to-moderate intrarater agreement (Fleiss κ=0.32 - 0.52).

RQ2: Concordance of LLM Ratings With Human Experts—Interrater Agreement (Gwet AC1 and AC2)

In our study, no single LLM-prompt configuration achieved superior agreement across all the rubric criteria. Figure 2 visualizes Gwet AC1 and AC2 by model, rubric, and prompting techniques. While simple zero-shot prompting achieved adequate agreement on specific rubric components, it failed to perform consistently across the entire rubric. Supplementing models with few-shot examples and/or CoT prompting was required to achieve consistent, substantial to almost-perfect agreement across most rubric components.

XSL•FO
RenderX

Figure 2. Agreement values according to Gwet AC1 and AC2 by rubric component, prompting configuration, and model. Binary components (person, factual accuracy) generally reached higher reliability than abstract ternary components (semantic qualifiers, transformation, narrowing, and global). Agreement values are interpreted using the Landis and Koch scale [40], which ranges from slight to almost perfect agreement. CoT: chain-of-thought.



To comprehensively evaluate how LLM selection, prompting configuration, and the rubric component impact performance, we conducted a linear mixed-effects model followed by a Type III ANOVA using R's *lme4* and *lmerTest* packages. To control for the nonindependence of repeated runs, we included the unique identifier for each individual batch (combining model, prompt configuration, and run index) as a random intercept. Satterthwaite method was subsequently applied to approximate the denominator degrees of freedom for the fixed effects. The analysis evaluated run-level Gwet AC1 and AC2 coefficients across all 5 runs per experimental condition. The analysis revealed significant main effects for model ($F_{3,\ 64}$= 2178.01, P<.001), prompting configuration ($F_{3,64}$=538.21, P<.001), and rubric component ($F_{5,320}$=714.33, P<.001). Crucially, all 2-way and 3-way interaction effects were also significant (P<.001), most notably the model × rubric component interaction

XSL•FO
RenderX

($F_{15,320}$=1717.72, P<.001). This indicates that while model choice and prompt configuration dictate overall performance, their efficacy is deeply tied to the specific nature of the rubric component being evaluated.

Differences by Model and Prompting Configuration

The concordance with human experts varied greatly depending on the selected model and even more for the prompt configuration. The mean Gwet AC1 and AC2 coefficients per model and prompting configuration are reported in Table 4.

Table . Gwet AC1 and AC2 interrater agreement coefficients for the 4 LLMs[a] across 4 prompting configurations (zero-shot and few-shot, with or without CoT[b]). For each model and prompting configuration, Gwet AC1 (binary components) or AC2 (ternary components) was calculated separately for each of the 6 rubric components in each of the 5 runs. The table shows the mean, SD, SE, and 95% CI of these 30 coefficients. The SD, SE, and 95% CI mainly reflect differences in agreement between rubric components.

Model and configuration	Mean (SD)[c]	SE[c]	95% CI[c]
Claude			
Few-shot	0.70 (0.16)	0.03	0.64-0.76
Few-shot CoT	0.77 (0.10)	0.02	0.74-0.81
Zero-shot	0.63 (0.23)	0.04	0.54-0.72
Zero-shot CoT	0.64 (0.21)	0.04	0.56-0.72
GPT-4o			
Few-shot	0.67 (0.17)	0.03	0.61-0.74
Few-shot CoT	0.72 (0.13)	0.02	0.68-0.77
Zero-shot	0.58 (0.26)	0.05	0.49-0.68
Zero-shot CoT	0.58 (0.26)	0.05	0.49-0.68
Mistral Large			
Few-shot	0.73 (0.10)	0.02	0.69-0.77
Few-shot CoT	0.85 (0.06)	0.01	0.83-0.87
Zero-shot	0.82 (0.10)	0.02	0.79-0.86
Zero-shot CoT	0.82 (0.10)	0.02	0.79-0.86
Mistral 7B			
Few-shot	0.57 (0.19)	0.03	0.5-0.64
Few-shot CoT	0.64 (0.22)	0.04	0.56-0.73
Zero-shot	0.52 (0.30)	0.05	0.41-0.63
Zero-shot CoT	0.54 (0.26)	0.05	0.44-0.64

[a]LLM: large language model.

[b]CoT: chain-of-thought.

[c]Gwet AC1 and AC2.

Mistral Large showed the best overall performance. Under optimal prompting, it reached almost perfect agreement on Person (AC1=0.976, 95% CI 0.969 - 0.983 in zero-shot CoT), Global (AC2=0.921, 95% CI 0.917 - 0.925 in zero-shot CoT), Narrowing (AC2=0.842, 95% CI 0.829 - 0.855 in few-shot CoT), SQ (AC2=0.820, 95% CI 0.811 - 0.828 in zero-shot), and transformation (AC2=0.807, 95% CI 0.800 - 0.813 in few-shot CoT). It was less effective for the binary component factual accuracy, with the lowest agreement in few-shot (AC1=0.530, 95% CI 0.514 - 0.546), but under few-shot CoT, it increased to 0.848 (95% CI 0.833 - 0.862). Interestingly, compared with zero-shot prompting, few-shot prompting reduced agreement, but this effect was mitigated when few-shot prompting was supplemented with CoT.

GPT-4o excelled in the binary rubric components, with factual accuracy peaking at AC1=0.880 (95% CI 0.866 - 0.894) in few-shot CoT and consistently high agreement on Person (lowest AC1=0.858, 95% CI 0.850 - 0.865 in few-shot CoT, highest 0.899, 95% CI 0.894 - 0.904 in zero-shot). However, it struggled with 2 of the 4 ternary categories in zero-shot: SQ (AC2=0.257, 95% CI 0.248 - 0.266) and transformation (AC2=0.235, 95% CI 0.222 - 0.247), corresponding to fair agreement. Only the combination of few-shot and CoT markedly improved these components (AC2=0.518, 95% CI 0.511 - 0.525, and AC2=0.652, 95% CI 0.627 - 0.676, respectively). Notably, under zero-shot, CoT had little effect across the whole rubric.

Claude reached the highest agreement in person across all prompting techniques (lowest AC1=0.961, 95% CI 0.954 - 0.968 in few-shot CoT; highest AC1=0.985, 95% CI 0.985 - 0.985 in both zero-shot and zero-shot CoT). It also achieved almost perfect agreement globally across all prompting conditions (lowest AC2=0.801, 95% CI 0.801 - 0.801 in

XSL•FO
RenderX

zero-shot CoT; highest AC2=0.818, 95% CI 0.818 - 0.818 in few-shot CoT). In the other ternary categories, Claude reached fair to moderate agreement in zero-shot; however, a combination of few-shot and CoT yielded substantial gains in agreement: SQ increased from AC2=0.334 (95% CI 0.334 - 0.334) to 0.705 (95% CI 0.702 - 0.708), Transformation from 0.409 (95% CI 0.409 - 0.409) to 0.740 (95% CI 0.738 - 0.743) and narrowing from 0.559 (95% CI 0.559 - 0.559) to 0.701 (95% CI 0.701 - 0.701). CoT in zero-shot, as in GPT-4o, seemed ineffective except for minimal gains in the rubric component person.

Mistral 7B Local showed the greatest variance among all the tested models. In zero-shot, it performed well in SQ, achieving almost perfect agreement (AC2=0.876; 95% CI 0.825 - 0.927); however, at the same time, person collapsed to near chance (AC1=0.037; 95% CI –0.016 to 0.089). Prompting did not improve Mistral 7B's performance in person. Factual accuracy was highly dependent on the configuration: weak in few-shot (AC1=0.449, 95% CI 0.427 - 0.472) but strong in zero-shot and zero-shot CoT (in both AC1=0.840, 95% CI 0.840 - 0.840). Narrowing and transformation benefitted from the combination of few-shot and CoT, reaching AC2=0.675 (95% CI 0.655 - 0.694) and 0.794 (95% CI 0.762 - 0.826), respectively. In Global, few-shot improved agreement independently of the presence of CoT.

Differences by Rubric Component

The binary components usually showed the highest agreement among the rubric components. The best agreement on person was achieved in Claude zero-shot and zero-shot CoT (both ex aequo with deterministic output at AC1=0.985, 95% CI 0.985 - 0.985), followed closely by Mistral Large under zero-shot CoT prompting (AC1=0.976, 95% CI 0.969 - 0.983). Factual Accuracy peaked with GPT-4o in few-shot CoT (AC1=0.880, 95% CI 0.866 - 0.894). Notably, Mistral 7B Local struggled in the person category and produced outliers, failing to achieve even moderate agreement across all prompting strategies (AC1≤0.292). Ternary rubric components led to more variable results. Global ratings were best with Mistral Large in zero-shot CoT (AC2=0.921, 95% CI 0.917 - 0.925); however, in terms of this component, Claude also consistently reached almost perfect agreement (the lowest AC2=0.801, 95% CI 0.801 - 0.801). For narrowing, Mistral Large few-shot CoT performed best (AC2=0.842, 95% CI 0.829 - 0.855). The agreement on the SQ was high among both Mistral models, with Mistral 7B Local in zero-shot reaching the highest AC2=0.876 (95% CI 0.825 - 0.927). Transformation was most reliably scored by Mistral Large in few-shot CoT, with AC2=0.807 (95% CI 0.800 - 0.813). Interestingly, for most models, transformation consistently benefited from few-shot prompting but not from CoT. We summarize the best model and prompting configuration per rubric component in Table 5.

Table . Highest interrater agreement by rubric component. The best model × prompt configuration per rubric component was selected according to Gwet AC1 in binary and AC2 in ternary rubric components. Claude showed the same performance in zero-shot with and without CoT[a] in the person rubric category.

Rubric component	Model	Configuration	Mean (SD)	SE	95% CI
Binary (AC1)					
Factual accuracy	GPT-4o	Few-shot CoT	0.880 (0.016)	0.007	0.866-0.894
Person	Claude	Zero-shot CoT	0.985 (0.000)	0.000	0.985-0.985
Person	Claude	Zero-shot	0.985 (0.000)	0.000	0.985-0.985
Ternary (AC2)					
SQs[b]	Mistral 7B	Zero-shot	0.876 (0.058)	0.026	0.825-0.927
Narrowing	Mistral Large	Few-shot CoT	0.842 (0.016)	0.007	0.829-0.855
Transformation	Mistral Large	Few-shot CoT	0.807 (0.007)	0.003	0.800-0.813
Global	Mistral Large	Zero-shot CoT	0.921 (0.004)	0.002	0.917-0.925

[a]CoT: chain-of-thought.

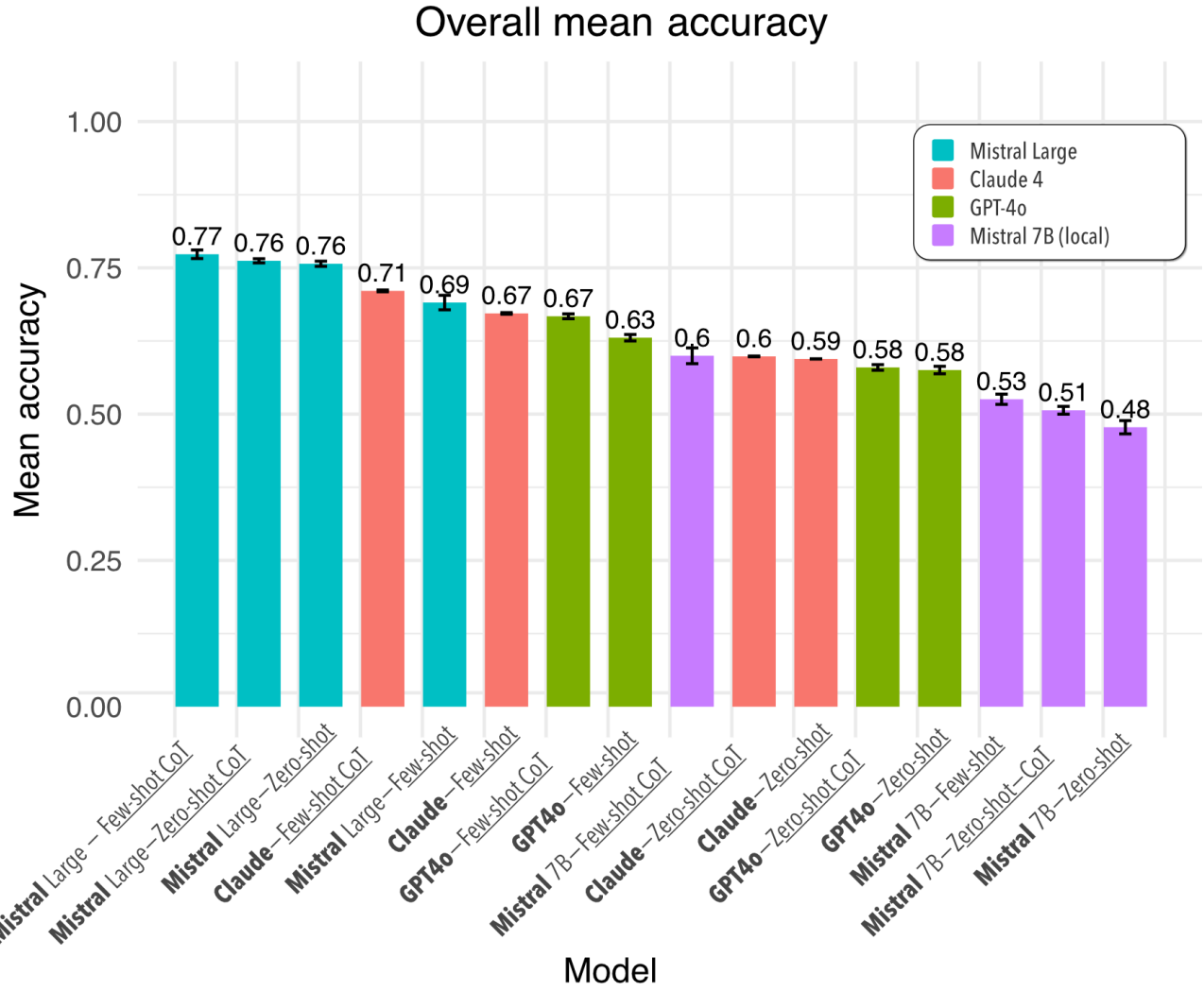
[b]SQs: semantic qualifiers.

Exact-Match Agreement (Accuracy) and Absolute Deviations From Human Experts

The few-shot with CoT prompting technique achieved the highest overall accuracy across all the models (Figure 3). Accuracy was calculated as the percentage of LLM ratings that exactly matched the human reference rating, averaged over all summary statements and rubric components. Peak values were observed for Mistral Large, where few-shot+CoT achieved 77% accuracy. The second-highest accuracy (76%) was achieved by Mistral Large zero-shot+CoT and Mistral Large zero-shot.

XSL•FO
RenderX

Figure 3. Overall mean accuracy (exact-match agreement with human expert ratings) per model and prompting configuration. Error bars show SDs across the 5 runs. CoT: chain-of-thought.



In few-shot with CoT, Claude achieved 71% accuracy, and GPT-4o reached 67%. Mistral 7B achieved 60% accuracy in few-shot CoT. It was higher than Claude in the zero-shot condition (59%) and GPT-4o in both the zero-shot and zero-shot+CoT settings (58%). Mistral 7B in zero-shot+CoT (51%) and zero-shot (48%) showed the lowest accuracy.

Deviations from human ratings were chosen as additional robustness checks. Extreme deviations of ±2 points from the human reference ratings were rare. Mistral Large showed the fewest such deviations, whereas Claude, GPT-4o, and Mistral 7B occasionally produced ±2 deviations, particularly in the SQs and Transformation categories (Figure 1). Rarely, invalid outputs were observed as described in the next section.

Handling Invalid Outputs

A total of 8 invalid outputs were observed (ie, scores outside the rubric scale, such as a 3 on a 0/1/2 scale, or a 2 on a binary rubric component) that only occurred with Mistral 7B as summarized in Table 6. The invalid outputs were found in all prompting configurations but few-shot with CoT. We excluded from all further analysis every summary statement entry in which an invalid data point occurred. Due to the very small number of excluded data points, their impact on reliability and agreement estimates is negligible.

Table . Overview of excluded invalid values from analysis. Initial sample size: 122 summary statements×5 runs (610 entries per configuration).

Configuration	Initial sample size	Number of excluded entries	Final analyzed sample size
Mistral 7B few-shot	610	1	609
Mistral 7B zero-shot	610	5	605
Mistral 7B zero-shot with CoT[a]	610	2	608

[a]CoT: chain-of-thought.

XSL•FO
RenderX

Discussion

Principal Findings

The primary aim of this study was to examine whether and to what extent LLMs can rate medical students' written summary statements from VP cases using a rubric and thereby help address the lack of individualized feedback in most VP systems. Our findings indicate that LLMs can serve as robust raters, providing a basis for scalable, automated feedback on summary statements in VP-based training. However, this requires careful consideration, as consistency and concordance with human experts depend on several factors: model choice, prompting configuration, temperature setting, and the rubric itself. There is no single best configuration; educators need to evaluate iteratively which combination of model and prompt works for their setting.

Across 5 repeated runs per prompting configuration, Mistral Large, Claude Sonnet, and GPT-4o produced stable scores with high test-retest reliability and substantial to almost perfect agreement with human experts, whereas the smaller local model Mistral 7B showed less consistent behavior across rubric components and prompting configurations. Few-shot with CoT prompting yielded the highest overall concordance with expert human raters, with Mistral Large performing best and Claude and GPT-4o also benefiting from illustrative examples and explicit reasoning initiators.

Divergences between high ICC values and lower Fleiss kappa values suggest a systematic scoring bias but highlight random stability, which could be due to an offset in the models' internal scoring scale that may be anchored differently from the human expert scale. This pattern was most frequently observed in GPT-4o and Mistral Large under zero-shot settings. Such systematic bias has practical implications for formative assessment, where absolute scores matter for student feedback, and argues for score calibration or normalization steps when using LLMs to score medical summary statements.

Rubric components displayed distinct performance patterns. Binary criteria (person and factual accuracy) generally achieved higher agreement than ternary criteria did. This could indicate that current LLMs capture surface-level correctness and easily operationalized features more reliably than they capture more nuanced aspects of clinical reasoning. However, this discrepancy might also stem from the frequency distribution of ratings within the human reference dataset (Multimedia Appendix 9). Future work should further investigate how rating distributions in reference datasets influence these results. To characterize agreement with the human reference standard, we prioritized Gwet AC1 and AC2, which reduced the underestimation of agreement observed with conventional coefficients such as Cohen kappa in skewed rating distributions.

Comparison With Prior Work

Recent work on automated scoring has shown that LLMs can approach or match human raters in terms of reliability and interrater agreement across a range of educational contexts. Seßler et al [41] reported high intrarater reliability (ICC up to 0.84 for closed-source models) and strong alignment with teacher ratings in multidimensional essay scoring [41], whereas Hackl et al [42] reported exceptionally high consistency for GPT-4 (ICC=0.94 - 0.99) in rating short written responses with minimal scoring drift under deterministic parameters. In medical education, Runyon [20] demonstrated that GPT-4o and Claude can achieve human-level accuracy when analytic rubrics are applied to postencounter notes if the prompts and scoring instructions are carefully engineered. Complementary studies by Alshammari et al [43] and Sreedhar et al [44] showed that LLMs can reach accuracy and consistency sufficient for formative use in answer-based and critical appraisal tasks, respectively. Our finding that large LLMs yield high ICCs and AC1 and AC2 values, particularly when well prompted, aligns with these findings that LLMs can act as reliable raters for structured, rubric-based assessments.

Integrating CoT reasoning with explicit rubric information improves both scoring accuracy and interpretability [19,45], but it is cautioned that simple zero-shot prompting tends to produce systematic overrating and weaker alignment with human judgments in content-rich or logic-heavy tasks [41,46]. In medical education, Runyon [20] and Jamieson et al [20,21] reported that prompts that embed analytic rubrics and provide examples enable LLMs to score postencounter notes with reliability comparable to that of faculty while dramatically reducing grading time. Similarly, Sreedhar et al [44] reported that rubric-guided prompting yields substantial agreement with faculty ratings. Our results replicate and extend these observations to VP-based summary statements: for all the models, few-shot with CoT prompting produced the highest agreement with human ratings, and no single LLM-prompt configuration dominated across all the rubric components. This finding supports the idea that prompt design and rubric operationalization are at least as important as model choice for achieving reliable scoring.

Our study also supports prior work emphasizing that LLMs perform best on clearly defined criteria and more variably on abstract reasoning dimensions. Across the models, the agreement was systematically greater for binary rubric components than for ternary ones. This pattern is consistent with views of clinical reasoning as a context-dependent, knowledge-intensive process that integrates multiple information sources [47]. It also resonates with findings that while LLMs handle pattern-based or shallow tasks well, they may struggle with more complex, compositional reasoning [48]. In our data, components such as transformation and narrowing benefited from few-shot with CoT prompting. Nevertheless, they did not reach the same level of robustness as binary criteria did, indicating that the more abstract the rubric dimension is, the more fragile automated scoring becomes.

The black-box character of LLMs further complicates their use as stand-alone graders. Because these models rely on learned internal representations rather than explicit symbolic rules, reconstructing exactly why a given score was assigned is difficult [48]. Runyon [20] proposed separating analytic explanation from score assignment, prompting the model first to analyze an answer and only then to derive a score to reduce complexity and improve transparency. Our findings support this line of argument: occasional extreme deviations and rare

XSL•FO
RenderX

invalid outputs in our robustness checks, particularly for composite or global scores, indicate that models can still fail in ways that are hard to interpret and justify to learners.

Finally, the educational impact of LLM-based feedback must be considered in light of how learners experience it. Ali et al [49] reported that students value the immediacy of AI-generated feedback but still rate human tutor feedback higher in terms of clarity and usefulness. This aligns with the broader concept of hybrid intelligence, where human and machine capabilities are combined to compensate for each other's weaknesses [50]. Our results fit well within this framework: LLMs appear capable of providing frequent, structured ratings of VP summary statements that could meaningfully expand feedback provision.

Limitations

First, we used a single combined reference rating from 2 human expert evaluators. Although predefined criteria and adjudicated disagreements supported consistent, context-appropriate judgments, this approach limits rater diversity and may not fully capture the variability of expert opinion. Future research should therefore involve a larger pool of raters, report interrater reliability between humans explicitly, and use these data as an additional benchmark for comparing LLM performance.

Second, the options for analyzing and comparing test-retest reliability across different LLMs (RQ1) depend statistically on the scale level of the assessment rubric. In the rubric used for this study, several components have ordinal intervals that do not represent equal distances, so the parametric ICC should be interpreted with caution. We therefore added Fleiss kappa as an exact categorical agreement metric, which served as our primary benchmark for test-retest reliability per LLM, while the ICC provides complementary information on the degree of systematic bias.

Third, the dataset comprised summary statements from selected VP cases in internal medicine only. It remains unclear whether our findings can be generalized to other disciplines, to different VP designs, or to more complex cases. Additional prompt refinement—such as further clarifying rubric descriptions or adding targeted examples—might be necessary when this approach is transferred to new contexts as LLM-generated rationales for rubric-based ratings depend on prompt wording during prompt development (for details, see Methods section). In our study, an effective prompting technique was few-shot with CoT. It included an LLM role description, the scoring rubric, the VP case context, scoring examples with corresponding cases, step-by-step rationales for those examples, an explicit CoT initiator, and a placeholder for the learner's summary statement. Future studies should systematically examine which of these elements are essential, how many examples are optimal, and whether strategies such as selecting extreme, typical, or atypical examples for few-shot prompting can further improve the scoring quality.

Fourth, our dataset contained virtual patient case vignettes and summary statements in English and German. Since the German sample size (n=41) was relatively small, we did not stratify our evaluation metrics by language to avoid unstable statistical estimates. Comparisons between common LLMs indicate that the performance for many high-resource languages (HRLs), such as English and German, is practically equal [51]. Future research with larger, balanced multilingual samples is needed to determine whether system performance varies significantly by language.

Fifth, regarding our model selection, new models such as GPT-5 and Claude Sonnet 4.5 became available after we completed data collection and were therefore not included. While their performance may differ, we argue that careful prompt design and robust evaluation strategies are more critical than any specific model choice for this type of task. Moreover, we deliberately restricted our analysis to baseline (ie, nonreasoning) LLMs and did not include large reasoning models (LRMs), such as OpenAI's o1 or Claude Sonnet 4.5 Thinking. This decision was made to maintain a consistent and comparable study design. Reasoning models introduce additional training stages and inference steps that may add uncertainty and require new evaluation frameworks [52,53]. Our goal here was to provide comprehensive evidence on commonly available baseline models as a first step. Future studies should systematically investigate how LRMs handle summary statement scoring, where their behavior diverges from that of baseline models, and which evaluation methods are suitable for comparison.

Conclusions

We showed that large LLMs can achieve high consistency across repeated runs and moderate to almost perfect agreement with human raters, particularly when few-shot prompting with a CoT initiator is used. The large volume of ratings generated enabled the robust estimation of reliability indices and revealed clear patterns in model- and rubric-level performance.

Our results suggest that LLMs can be used to expand the capacity of VP systems to deliver timely, structured feedback on clinical reasoning for certain rating tasks and for well-defined rubrics. Specifically, they can provide consistent ratings on aspects such as factual accuracy and the inclusion of key patient identifiers, and they show promising, although more variable, performance on rating higher-order reasoning dimensions such as transformation and diagnostic narrowing. This finding indicates that LLM-generated scores could be used to provide immediate, rubric-based feedback after VP cases, helping learners monitor and refine their problem representations more frequently than would be feasible with exclusively human raters. However, their practical use must be further explored. The remaining variability in ternary rubric components, occasional extreme deviations, and rare invalid outputs caution against the use of LLMs as stand-alone graders, especially in summative high-stakes contexts. The opaque nature of LLM decision-making also makes it difficult to justify individual scores to learners. In line with the idea of hybrid intelligence, our findings support an architecture in which LLMs function as supportive tools: they generate rapid and scalable ratings of VP summary statements, whereas human educators can interpret borderline cases, contextualize feedback, and guide learners' metacognitive reflection on their reasoning. Further refinement of more complex rubric components, such as ternary components that require more reasoning and in-depth understanding (eg,

XSL•FO
RenderX

transformation), could help to reduce human involvement in the scoring process.

Future work should examine how students interact with and perceive LLM-generated feedback within VP platforms, whether hybrid human-AI feedback models improve learning outcomes, and how domain-specific adaptations or reasoning models might increase the reliability for abstract reasoning components without compromising transparency. Overall, our findings indicate that with carefully designed prompts and annotated rubrics, current LLMs can support formative feedback on clinical reasoning in VPs, but they remain complementary to expert human judgment.

Acknowledgments

We thank Maximilian Mönch, mathematician at the Institute of Biometry and Clinical Epidemiology, Charité – Universitätsmedizin Berlin, for providing statistical counseling, particularly on the selection of appropriate interrater reliability measures. Next, we thank the authors of the iCoViP project [54] for providing the virtual patient cases, licensed as CC BY-NC-SA 4.0, as part of the Erasmus+-funded project iCoViP in CASUS [55]. We also thank the original expert raters for their expertise in rating the summary statements. A collection of 50 exemplary summary statements including the ratings, licensed as CC BY 4.0, is available at Multimedia Appendix 10. Generative AI (GPT-4o and Claude Sonnet 4.5) was used for proofreading and editing, summarization, translation, and reformatting. Responsibility for the final manuscript lies entirely with the authors. Generative AI tools are not listed as authors and do not bear responsibility for the final outcomes.

Funding

This study was funded by the Foundation for Innovation in Higher Education (Stiftung Innovation in der Hochschullehre) within the HEDS project at Charité – Universitätsmedizin Berlin.

Data Availability

The complete R and Python codes, as well as virtual patient cases and expert reference ratings used for this study, can be found in our GitHub repository [56], licensed as GPL-3.0.

Authors' Contributions

Conceptualization: JJK (lead), PB (supporting)
Data curation: PS (lead), IH (supporting)
Formal analysis: PS
Funding acquisition: HP
Investigation: PS
Methodology: PS (lead), PB (equal)
Project administration: YH
Resources: IH (lead), HP (equal)
Supervision: JJK
Validation: PB (lead), PS (equal), JJK (equal)
Visualization: PS
Writing – original draft: PB (lead), PS (equal), JJK (equal)
Writing – review & editing: HP (lead), IH (equal), JJK (equal)

Conflicts of Interest

None declared.

Multimedia Appendix 1
Prompt design.
[DOC File, 96 KB - mededu_v12i1e94666_app1.doc]

Multimedia Appendix 2
The Scoring Rubric.
[DOC File, 90 KB - mededu_v12i1e94666_app2.doc]

Multimedia Appendix 3
Gwet AC1 and AC2 values.
[DOC File, 212 KB - mededu_v12i1e94666_app3.doc]

XSL•FO
RenderX

Multimedia Appendix 4
Weighted Cohen Kappa values.
[DOC File, 187 KB - mededu_v12i1e94666_app4.doc]

Multimedia Appendix 5
Enlarged view of the Claude ratings (top left panel) from Figure 1.
[PNG File, 294 KB - mededu_v12i1e94666_app5.png]

Multimedia Appendix 6
Enlarged view of the GPT-4o ratings (top right panel) from Figure 1.
[PNG File, 281 KB - mededu_v12i1e94666_app6.png]

Multimedia Appendix 7
Enlarged view of the Mistral Large ratings (bottom left panel) from Figure 1.
[PNG File, 301 KB - mededu_v12i1e94666_app7.png]

Multimedia Appendix 8
Enlarged view of the Mistral 7B ratings (bottom right panel) from Figure 1.
[PNG File, 310 KB - mededu_v12i1e94666_app8.png]

Multimedia Appendix 9
Distribution of rating scores by the human experts.
[DOC File, 101 KB - mededu_v12i1e94666_app9.doc]

Multimedia Appendix 10
A collection of 50 exemplary summary statements including the ratings, licensed as CC BY 4.0.
[XLSX File, 61 KB - mededu_v12i1e94666_app10.xlsx]

References

1. Chang RW, Bordage G, Connell KJ. The importance of early problem representation during case presentations. Acad Med 1998 Oct;73(10 Suppl):S109-S111. [doi: 10.1097/00001888-199810000-00062] [Medline: 9795669]
2. Bowen JL. Educational strategies to promote clinical diagnostic reasoning. In: Cox M, Irby DM, editors. N Engl J Med 2006 Nov 23;355(21):2217-2225. [doi: 10.1056/NEJMra054782] [Medline: 17124019]
3. Bowen JL, Ten Cate O, Prerequisites for learning clinical reasoning. Principles and practice of case-based clinical reasoning education. In: Ten Cate O, Custers EJFM, Durning SJ, editors. Innovation and Change in Professional Education: Springer; 2018, Vol. 15:47-63. [doi: 10.1007/978-3-319-64828-6_4]
4. Feblowitz JC, Wright A, Singh H, Samal L, Sittig DF. Summarization of clinical information: a conceptual model. J Biomed Inform 2011 Aug;44(4):688-699. [doi: 10.1016/j.jbi.2011.03.008] [Medline: 21440086]
5. Penner JC, Shipley LC, Minter DJ. Reasoning on rounds: a framework for teaching diagnostic reasoning in the inpatient setting. J Gen Intern Med 2023 Oct;38(13):3041-3046. [doi: 10.1007/s11606-023-08359-1] [Medline: 37580633]
6. Smith S, Kogan JR, Berman NB, Dell MS, Brock DM, Robins LS. The development and preliminary validation of a rubric to assess medical students' written summary statements in virtual patient cases. Acad Med 2016 Jan;91(1):94-100. [doi: 10.1097/ACM.0000000000000800] [Medline: 26726864]
7. Custers EJFM. Thirty years of illness scripts: theoretical origins and practical applications. Med Teach 2015 May;37(5):457-462. [doi: 10.3109/0142159X.2014.956052] [Medline: 25180878]
8. Fürstenberg S, Oubaid V, Berberat PO, Kadmon M, Harendza S. Medical knowledge and teamwork predict the quality of case summary statements as an indicator of clinical reasoning in undergraduate medical students. GMS J Med Educ 2019;36(6):Doc83. [doi: 10.3205/zma001291] [Medline: 31844655]
9. Norman GR, Monteiro SD, Sherbino J, Ilgen JS, Schmidt HG, Mamede S. The causes of errors in clinical reasoning: cognitive biases, knowledge deficits, and dual process thinking. Acad Med 2017 Jan;92(1):23-30. [doi: 10.1097/ACM.0000000000001421] [Medline: 27782919]
10. Kononowicz AA, Woodham LA, Edelbring S, et al. Virtual patient simulations in health professions education: systematic review and meta-analysis by the digital health education collaboration. J Med Internet Res 2019 Jul 2;21(7):e14676. [doi: 10.2196/14676] [Medline: 31267981]
11. Jay R, Sandars J, Patel R, et al. The use of virtual patients to provide feedback on clinical reasoning: a systematic review. Acad Med 2025 Feb 1;100(2):229-238. [doi: 10.1097/ACM.0000000000005908] [Medline: 39485118]

XSL•FO
RenderX

12. Bing-You R, Hayes V, Varaklis K, Trowbridge R, Kemp H, McKelvy D. Feedback for learners in medical education: what is known? A scoping review. Acad Med 2017 Sep;92(9):1346-1354. [doi: 10.1097/ACM.0000000000001578] [Medline: 28177958]
13. Lipnevich AA, Mattern K, Feddock C. Formative assessment and feedback in medical education: a practical guide: AMEE Guide No. 189. Med Teach 2026 Jun;48(6):921-940. [doi: 10.1080/0142159X.2025.2569623] [Medline: 41134874]
14. Hege I, Kiesewetter I, Adler M. Automatic analysis of summary statements in virtual patients - a pilot study evaluating a machine learning approach. BMC Med Educ 2020 Oct 16;20(1):366. [doi: 10.1186/s12909-020-02297-w] [Medline: 33066789]
15. Bond WF, Zhou J, Bhat S, et al. Automated patient note grading: examining scoring reliability and feasibility. Acad Med 2023 Nov 1;98(11S):S90-S97. [doi: 10.1097/ACM.0000000000005357] [Medline: 37983401]
16. Cho HN, Jun TJ, Kim YH, et al. Task-specific transformer-based language models in health care: scoping review. JMIR Med Inform 2024 Nov 18;12:e49724. [doi: 10.2196/49724] [Medline: 39556827]
17. Nori H, Daswani M, Kelly C, et al. Sequential diagnosis with language models. arXiv. Preprint posted online on Jun 27, 2025. [doi: 10.48550/arXiv.2506.22405]
18. Croxford E, Gao Y, First E, et al. Evaluating clinical AI summaries with large language models as judges. NPJ Digit Med 2025 Nov 5;8(1):640. [doi: 10.1038/s41746-025-02005-2] [Medline: 41193667]
19. Lee GG, Latif E, Wu X, Liu N, Zhai X. Applying large language models and chain-of-thought for automatic scoring. Comput Educ Artif Intell 2024 Jun;6:100213. [doi: 10.1016/j.caeai.2024.100213]
20. Runyon C. Using large language models (LLMs) to apply analytic rubrics to score post-encounter notes. Med Teach 2025 Nov;47(11):1772-1780. [doi: 10.1080/0142159X.2025.2504106] [Medline: 40380943]
21. Jamieson AR, Holcomb MJ, Dalton TO, et al. Rubrics to prompts: assessing medical student post-encounter notes with AI. NEJM AI 2024 Nov 27;1(12). [doi: 10.1056/AIcs2400631]
22. Chen B, Zhang Z, Langrené N, Zhu S. Unleashing the potential of prompt engineering for large language models. Patterns (N Y) 2025 Jun 13;6(6):101260. [doi: 10.1016/j.patter.2025.101260] [Medline: 40575123]
23. Singhal K, Azizi S, Tu T, et al. Large language models encode clinical knowledge. Nature 2023 Aug;620(7972):172-180. [doi: 10.1038/s41586-023-06291-2] [Medline: 37438534]
24. Han T, Adams LC, Papaioannou JM, et al. MedAlpaca–an open-source collection of medical conversational AI models and training data. arXiv. Preprint posted online on Apr 14, 2023. [doi: 10.48550/arXiv.2304.08247]
25. Rasmy L, Xiang Y, Xie Z, Tao C, Zhi D. Med-BERT: pretrained contextualized embeddings on large-scale structured electronic health records for disease prediction. NPJ Digit Med 2021 May 20;4(1):86. [doi: 10.1038/s41746-021-00455-y] [Medline: 34017034]
26. Lin J, Tang J, Tang H, et al. Awq: Activation-aware weight quantization for on-device LLM compression and acceleration. Proc Mach Learn Syst 2024;6:87-100 [FREE Full text]
27. CASUS. CASUS online. URL: https://crt.casus.net/pmw2/app/homepage.html [accessed 2026-09-14]
28. Kojima T, Gu S, Reid M, Matsuo Y, Iwasawa Y. Large language models are zero-shot reasoners. 2022 Presented at: Advances in Neural Information Processing Systems 35; Nov 28 to Dec 9, 2022; New Orleans, LA p. 22199-22213. [doi: 10.52202/068431-1613]
29. Jiang Y, Wang Y, Zeng X, et al. FollowBench: a multi-level fine-grained constraints following benchmark for large language models. In: Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Association for Computational Linguistics; 2024:4667-4688. [doi: 10.18653/v1/2024.acl-long.257]
30. Kwon W, Li Z, Zhuang S, et al. Efficient memory management for large language model serving with pagedattention. In: SOSP '23: Proceedings of the 29th Symposium on Operating Systems Principles: Association for Computing Machinery; 2023:611-626. [doi: 10.1145/3600006.3613165]
31. Atıl B, Aykent S, Chittams A. Non-determinism of "deterministic" LLM system settings in hosted environments. In: Proceedings of the 5th Workshop on Evaluation and Comparison of NLP Systems 2025:135-148. [doi: 10.18653/v1/2025.eval4nlp-1.12]
32. R Core Team. The R Language Definition. Vienna Austria R Found Stat Comput. 2000. URL: https://cran.r-project.org/doc/manuals/r-devel/R-lang.html
33. Koo TK, Li MY. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. J Chiropr Med 2016 Jun;15(2):155-163. [doi: 10.1016/j.jcm.2016.02.012] [Medline: 27330520]
34. Xu S, Lorber MF. Interrater agreement statistics with skewed data: evaluation of alternatives to Cohen's kappa. J Consult Clin Psychol 2014 Dec;82(6):1219-1227. [doi: 10.1037/a0037489] [Medline: 25090041]
35. Gwet KL. Computing inter-rater reliability and its variance in the presence of high agreement. Br J Math Stat Psychol 2008 May;61(Pt 1):29-48. [doi: 10.1348/000711006X126600] [Medline: 18482474]
36. Gwet KL. Handbook of Inter-Rater Reliability: The Definitive Guide to Measuring the Extent of Agreement among Raters: Advanced Analytics, LLC; 2014. URL: https://books.google.co.in/books/about/Handbook_of_Inter_Rater_Reliability_4th.html?id=fac9BQAAQBAJ&redir_esc=y [accessed 2026-10-07]

XSL•FO
RenderX

37. Wongpakaran N, Wongpakaran T, Wedding D, Gwet KL. A comparison of Cohen's Kappa and Gwet's AC1 when calculating inter-rater reliability coefficients: a study conducted with personality disorder samples. BMC Med Res Methodol 2013 Apr 29;13(1):61. [doi: 10.1186/1471-2288-13-61] [Medline: 23627889]
38. Silveira PSP, Siqueira JO. Better to be in agreement than in bad company. Behav Res 2023;55(7):3326-3347. [doi: 10.3758/s13428-022-01950-0]
39. Popplewell M, Reizes J, Zaslawski C. Appropriate statistics for determining chance-removed interpractitioner agreement. J Altern Complement Med 2019 Nov;25(11):1115-1120. [doi: 10.1089/acm.2017.0297] [Medline: 29851495]
40. Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics 1977 Mar;33(1):159-174. [doi: 10.2307/2529310] [Medline: 843571]
41. Seßler K, Fürstenberg M, Bühler B, Kasneci E. Can AI grade your essays? A comparative analysis of large language models and teacher ratings in multidimensional essay scoring. In: LAK '25: Association for Computing Machinery; 2025:462-472. [doi: 10.1145/3706468.3706527]
42. Hackl V, Müller AE, Granitzer M, Sailer M. Is GPT-4 a reliable rater? Evaluating consistency in GPT-4's text ratings. Front Educ 2023 Dec 5;8:1272229. [doi: 10.3389/feduc.2023.1272229]
43. Alshammari AF, Madfa AA, Anazi BA, Alenezi YE, Alkurdi KA. Comparison of accuracy and consistency of AI Language models when answering standardised dental MCQs. BMC Med Educ 2025 Oct 28;25(1):1507. [doi: 10.1186/s12909-025-07624-7] [Medline: 41152856]
44. Sreedhar R, Chang L, Gangopadhyaya A, et al. Comparing scoring consistency of large language models with faculty for formative assessments in medical education. J Gen Intern Med 2025 Jan;40(1):127-134. [doi: 10.1007/s11606-024-09050-9] [Medline: 39402411]
45. Xiao C, Ma W, Song Q, et al. Human-AI collaborative essay scoring: a dual-process framework with llms. In: LAK '25: Association for Computing Machinery; 2025:293-305. [doi: 10.1145/3706468.3706507]
46. Wu X, Saraf PP, Lee G, Latif E, Liu N, Zhai X. Unveiling scoring processes: dissecting the differences between LLMs and human graders in automatic scoring. Tech Know Learn 2026 Jun;31(2):669-684. [doi: 10.1007/s10758-025-09836-8]
47. Olson A, Kämmer JE, Taher A, et al. The inseparability of context and clinical reasoning. J Eval Clin Pract 2024 Jun;30(4):533-538. [doi: 10.1111/jep.13969] [Medline: 38300231]
48. Yuan Z, Yuan H, Tan C, Wang W, Huang S. How well do large language models perform in arithmetic tasks? arXiv. Preprint posted online on Mar 16, 2023. [doi: 10.48550/ARXIV.2304.02015]
49. Ali M, Harbieh I, Haider KH. Bytes versus brains: a comparative study of AI-generated feedback and human tutor feedback in medical education. Med Teach 2026 Jan;48(1):131-141. [doi: 10.1080/0142159X.2025.2519639] [Medline: 40528415]
50. Dellermann D, Ebel P, Söllner M, Leimeister JM. Hybrid Intelligence. Bus Inf Syst Eng 2019 Oct;61(5):637-643. [doi: 10.1007/s12599-019-00595-2]
51. Lai V, Ngo N, Pouran Ben Veyseh A, et al. ChatGPT beyond English: towards a comprehensive evaluation of large language models in multilingual learning. Presented at: Findings of the Association for Computational Linguistics: EMNLP 2023; Dec 6-10, 2023; Singapore p. 13171-13189. [doi: 10.18653/v1/2023.findings-emnlp.878]
52. Shojaee P, Mirzadeh I, Alizadeh K, Horton M, Bengio S, Farajtabar M. The illusion of thinking: understanding the strengths and limitations of reasoning models via the lens of problem complexity. arXiv. Preprint posted online on Jun 7, 2025. [doi: 10.48550/ARXIV.2506.06941]
53. Chen Y, Benton J, Radhakrishnan A, et al. Reasoning models don't always say what they think. arXiv. Preprint posted online on May 8, 2025. [doi: 10.48550/ARXIV.2505.05410]
54. International Collection of Virtual Patients (iCoViP). URL: https://icovip.eu [accessed 2026-10-08]
55. CASUS. URL: https://icovip.casus.net [accessed 2026-10-08]
56. MedicalSummaryStatements_LLMs. GitHub. URL: https://github.com/DSFZ-Team/MedicalSummaryStatements_LLMs [accessed 2026-10-08]

Abbreviations

AWQ: activation-aware weight quantization
BERT: Bidirectional Encoder Representations from Transformers
CoT: chain-of-thought
HRL: high-resource language
ICC: intraclass correlation coefficient
LLM: large language model
LRM: large reasoning model
NLP: natural language processing
OSCE: Objective Structured Clinical Examination
RQ: research question
SQ: semantic qualifier
VP: virtual patient

XSL•FO
RenderX

Edited by B Lesselroth; submitted 05.Mar.2026; peer-reviewed by E Karampinis, O Xia; revised version received 12.Aug.2026; accepted 26.Aug.2026; published 21.Oct.2026.

Please cite as:
Stadler P, Bucichowski P, Holzhausen Y, Hege I, Peters H, Kaminski JJ
Test-Retest Reliability of Large Language Models and Their Concordance With Human Experts in Rubric-Based Scoring of Clinical Summary Statements: Experimental Comparative Study
JMIR Med Educ 2026;12:e94666
URL: https://mededu.jmir.org/2026/1/e94666
doi:10.2196/94666

© Patrick Stadler, Piotr Bucichowski, Ylva Holzhausen, Inga Hege, Harm Peters, Julius Josef Kaminski. Originally published in JMIR Medical Education (https://mededu.jmir.org), 21.Oct.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Education, is properly cited. The complete bibliographic information, a link to the original publication on https://mededu.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Multimodal Large Language Models for Dental Chart Image Interpretation: Cross-Sectional Benchmarking Study With Students and Clinicians

Ah-Young Cho[1], DDS, MSD; Soo-Heang Eo[2,3], PhD; Mi-Jeong Jeon[4], DDS, MSD, PhD; Jae-Hoon Kim[5], DDS, MSD, PhD; Jungjoon Ihm[6], PhD; Deog-Gyu Seo[1], DDS, MSD, PhD

[1]Department of Conservative Dentistry and Dental Research Institute, School of Dentistry, Seoul National University, 101 Daehakno, Jongno-Gu, Seoul, Republic of Korea

[2]Graduate School, Department of Urban Big Data Convergence, University of Seoul, Seoul, Republic of Korea

[3]AI Agent Team, CryptoLab Inc., Seoul, Republic of Korea

[4]Department of Conservative Dentistry and Oral Science Research Center, Yonsei University College of Dentistry, Seoul, Republic of Korea

[5]Department of Dental Education, Dental and Life Science Institute, School of Dentistry, Pusan National University and Dental Research Institute, Busan, Republic of Korea

[6]Department of Dental Education and Dental Research Institute, School of Dentistry, Seoul National University, Seoul, Republic of Korea

Corresponding Author:
Deog-Gyu Seo, DDS, MSD, PhD
Department of Conservative Dentistry and Dental Research Institute, School of Dentistry, Seoul National University, 101 Daehakno, Jongno-Gu, Seoul, Republic of Korea

Abstract

Background: Entering clinical training, dental students must learn to read tooth-centered electronic dental records, but limited teaching time in patient-centered clinics can leave gaps in chart-reading literacy. Multimodal large language models (LLMs) that process dental record images may offer scalable educational support, yet their performance has not been benchmarked.

Objective: This study evaluated multimodal ChatGPT models on dental chart image interpretation and compared the best-performing model with dental students and residents.

Methods: A retrospective, cross-sectional benchmark study used deidentified dental chart images from 15 patients at Seoul National University Dental Hospital (2017 - 2025). Charts containing Korean and English text were captured as sequential screenshots (154 images). For each patient, 24 Korean-language questions (360 total) spanned four categories: Type A, general factual retrieval; Type B, tooth- or procedure-specific retrieval; Type C, interpretation requiring multientry synthesis; and Type D, absent-information questions assessing abstention. Nine multimodal ChatGPT models (available August 2, 2025 - August 9, 2025) were evaluated under standardized conditions. Outputs were scored against a gold standard using 7 metrics, with sentence bidirectional encoder representations from transformers (SBERT) similarity prespecified as the primary semantic measure. Human baselines included 2 third-year students and 2 first-year residents. Groups were compared with Kruskal-Wallis tests and Dunn post hoc analyses. Gold-standard reliability was assessed by independent senior-expert review and chance-corrected agreement (Gwet AC1) among clinical reference raters, and the model ranking was confirmed by content-based clinical accuracy analysis.

Results: Across 360 items, GPT-5 Thinking achieved the highest median SBERT similarity (0.900, IQR 0.525 - 1.000), followed by GPT-5 Pro (0.861, IQR 0.501 - 1.000) and OpenAI o3 (0.831, IQR 0.489 - 1.000), with significant overall group differences (P<.001). Student 1 did not differ significantly from GPT-5 Thinking across Types A-D (all P≥.16), whereas Student 2 differed only on Type D (P=.002; Cliff δ=−0.27). Resident 1 scored higher than GPT-5 Thinking on Types A (P=.01) and C (P=.04) but lower on Type D (P<.001; δ=−0.35), whereas Resident 2 scored higher on Types A (P=.03), B (P=.02), and C (P=.04) and did not differ on Type D (P=.23). Type C tasks showed compressed SBERT distributions and low exact-match rates, indicating persistent difficulty in synthesis. The clinical reference rater agreement was high (Gwet AC1=0.99), and the content-based clinical-accuracy ranking was closely aligned with the SBERT ranking (Spearman ρ=0.90; P=.001).

Conclusions: Statistically nonsignificant differences were observed between GPT-5 Thinking and dental students for most question-type contrasts, with Student 2 differing only on absent-information items. Compared with first-year residents, GPT-5 Thinking remained lower on several Type A-C contrasts, particularly interpretive Type C and one tooth- or procedure-specific Type B comparison; Type D contrasts require cautious interpretation because all groups had ceiling medians. Within this single-center benchmark, multimodal LLMs may have potential as supervised educational tools for chart-reading practice and verification, rather than as replacements for clinical expertise, pending external validation across institutions, specialties, and record systems.

XSL•FO
RenderX

(JMIR Med Educ 2026;12:e91809) doi:10.2196/91809

KEYWORDS

dental records; education, dental; students, dental; clinical competence; large language models; generative AI; ChatGPT; multimodal ChatGPT; benchmarking; natural language processing

Introduction

At the preclinical-to-clinical transition, when learners first enter the clinical environment, dental students must develop the ability to read and interpret dental charts, also referred to as dental records, including locating relevant information, reconstructing treatment timelines, and distinguishing tooth-specific details from patient-level records [1,2]. For students new to clinical practice, this process can be challenging and may add cognitive load in already demanding learning environments, particularly within patient-centered clinics where clinical requirements limit protected teaching time [3,4]. Although dental charts are foundational records of a patient's history and guide subsequent treatment, limited clinical exposure can lead to chart misinterpretation, increasing the downstream risk of treatment errors [5,6]. Because this transition is formative, curricula should minimize gaps by providing explicit early instruction and structured guidance to support competence development. However, current dental curricula often prioritize hands-on procedural training, leaving the cognitive skills required for interpreting clinical records largely implicit [4]. The volume and complexity of chart data, coupled with the clinical priorities of patient-centered care, further constrain opportunities for timely teaching and feedback. Ideally, faculty would regularly assess students' chart-reading skills and provide targeted coaching; in practice, patient-care responsibilities limit real-time instruction, and scalable assessment of comprehension is rarely feasible. Consequently, developing "chart-reading literacy"—the ability to locate, interpret, and synthesize information from complex dental charts—represents a core educational objective and a prerequisite for clinical reasoning and procedural competence in dentistry [1-4].

Against this backdrop, large language models (LLMs) have emerged as potential educational aids. LLMs can retrieve, organize, and synthesize information into coherent, task-focused explanations, and recent advances in training and alignment have substantially improved their accuracy and robustness [7,8]. ChatGPT, in particular, is a widely renowned LLM encompassing multiple releases, including reasoning-focused models and the GPT-4 series; most recently, the GPT-5 family (GPT-5, GPT-5 Thinking, and GPT-5 Pro) has expanded multimodal capabilities and strengthened reliability and reasoning. Learners across disciplines, including dental students, increasingly use ChatGPT to seek information and clarify unfamiliar topics [9,10]. Its chat-based interface supports AI-guided learning by enabling users to pose questions, test interpretations, and iteratively refine understanding through interactive dialogue. For students navigating the preclinical-to-clinical transition, ChatGPT may facilitate verification of chart interpretations, provide immediate clarification, and model structured reasoning in real time. Consequently, generative AI-guided education has been proposed as an emerging approach in dental education, positioning LLMs as adaptive instructional partners rather than passive information sources [11]. At the same time, persistent limitations—including hallucinations, sensitivity to prompt formulation, and variable performance across task types—highlight the continued need for supervision and verification [12,13].

Dental electronic records present distinct interpretive challenges that distinguish them from general electronic medical records (EMRs). Whereas most EMRs are organized around patient-level narrative entries, dental records are organized at the level of individual teeth and integrate spatial elements such as odontograms, tooth-numbering grids, and procedure maps [14]. Relevant information is frequently distributed across multiple pages and time points so that even straightforward tasks, such as reconstructing the treatment history of a single tooth, require learners to integrate spatial cues with textual entries [5,6,14]. Critically, these spatial cues cannot be faithfully preserved when dental records are linearized into text alone, and text-only approaches therefore fail to capture the tooth-level context central to accurate interpretation of dental records [4,15]. This challenge is compounded by substantial heterogeneity and documentation gaps in dental EMRs, which prior work in dental informatics has highlighted in calls for dental-specific templates and standardization [6,16]. These characteristics indicate that meaningful evaluation of generative AI for dental chart reading requires multimodal inputs in which models reason directly over images of dental charts.

Despite this potential, few rigorous assessments have examined how generative AI models perform on dental chart images, particularly in educational contexts. Prior research in dental informatics has largely focused on coding systems, data quality, and secondary data use, with limited investigations into whether LLMs can accurately retrieve tooth-level information, reason across dispersed chart entries, and appropriately abstain when information is absent [14,15,17]. This gap is consequential because well-characterized tools could reduce barriers to chart comprehension, support deliberate practice in clinical reasoning, and free instructional time for supervised skill acquisition, whereas poorly characterized tools risk amplifying errors and fostering false confidence. If ChatGPT's performance in educational settings is systematically validated, it may provide students navigating the preclinical-to-clinical transition with a scalable means to practice chart reading and verify their interpretations, even in the absence of formalized chart-reading instruction [11,15].

This study evaluated multimodal ChatGPT models using dental chart images as inputs. The official models evaluated were OpenAI o3, GPT-4.5, GPT-4.1, GPT-4.1 mini, OpenAI o4-mini, GPT-4o, GPT-5, GPT-5 Thinking, and GPT-5 Pro. Models were benchmarked using semantic and exact-match evaluation metrics across 4 question categories, including general factual retrieval,

XSL•FO
RenderX

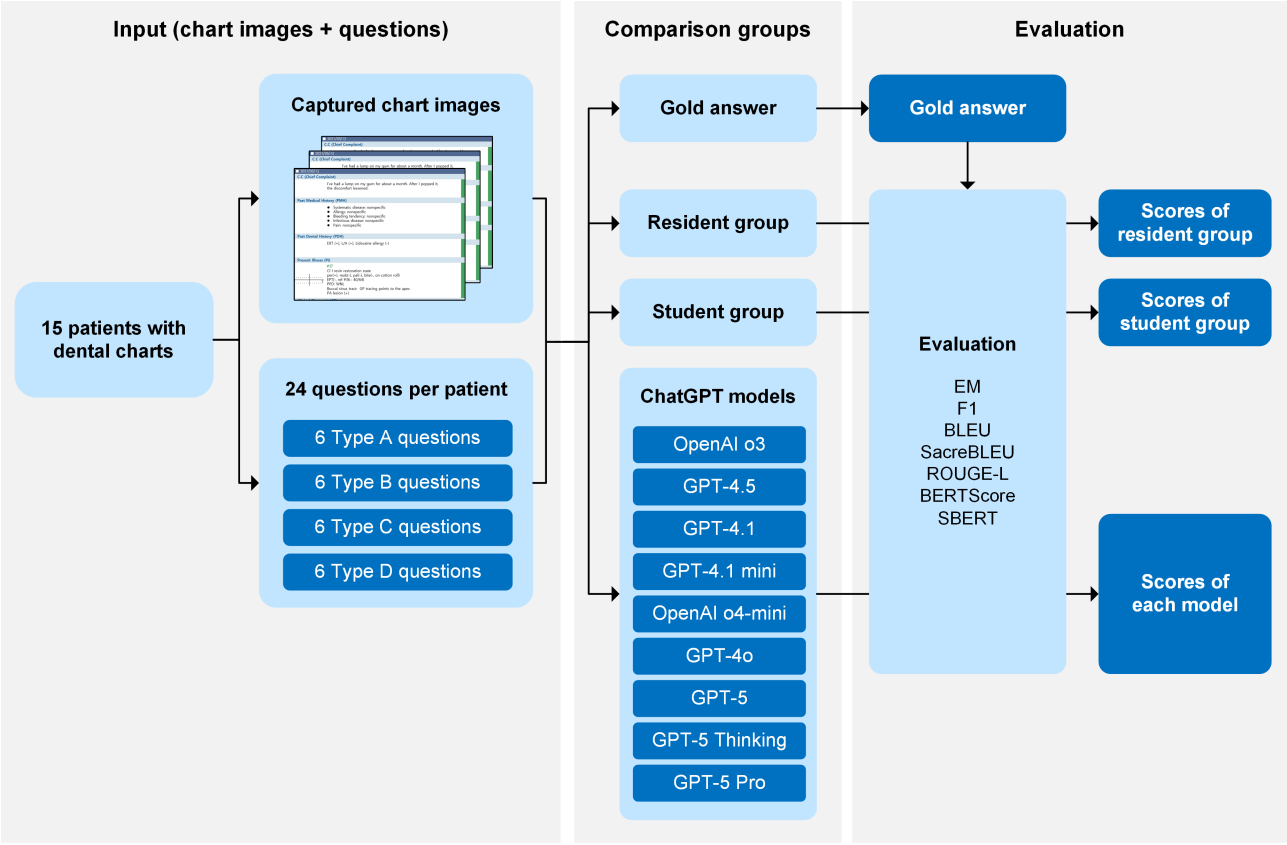
tooth- or procedure-specific retrieval, interpretive reasoning, and absent-information (unanswerable) tasks. The highest-performing model was subsequently compared with dental students at the preclinical-to-clinical transition and with resident clinicians to assess relative performance in an educational context. This study aimed to benchmark the chart-image interpretation performance of multimodal ChatGPT models across these 4 task categories and to compare the best-performing model with dental students at the preclinical-to-clinical transition and with qualified clinicians in order to characterize where such models responsibly support chart-reading education.

Methods

Study Design and Overview

A retrospective, cross-sectional evaluation was conducted using deidentified dental chart images under standardized testing conditions. The reporting of this study followed the STROBE (Strengthening the Reporting of Observational Studies in Epidemiology) checklist (Checklist 1). The primary objective was to assess multimodal ChatGPT performance across three education-relevant competencies: (1) factual retrieval from dental charts, including general and tooth- or procedure-specific information, (2) inferential reasoning across dispersed chart entries, and (3) appropriate handling of missing or absent information. Analyses were prespecified and reported overall, by model, and by question type. The study workflow is summarized in Figure 1.

Figure 1. Flowchart of study design and evaluation process. BERT: bidirectional encoder representations from transformers; BLEU: bilingual evaluation understudy; EM: exact match; ROUGE-L: recall-oriented understudy for gisting evaluation-longest common subsequence; SBERT: sentence bidirectional encoder representations from transformers.



Data Source and Case Selection

Dental charts from 15 patients were sampled from the Department of Conservative Dentistry at Seoul National University Dental Hospital (SNUDH), covering the period from January 1, 2017, to June 30, 2025. The charts contained both Korean and English text. Figure 2 presents an example of a captured dental chart image, including the patient's chief complaint, past medical and dental histories, current dental problems (present illness) with associated odontograms, clinical diagnoses, treatment plans, treatments performed, and free-text clinical notes. The case mix comprised restorative (n=6), endodontic (n=5), surgical (n=2), and mixed or complex cases (n=2). Inclusion criteria were as follows: (1) documentation of treatment for at least one tooth, (2) the presence of chart elements verifiable from images (eg, odontograms, procedure notes, and visit dates), and (3) complete deidentification prior to data export. Charts containing any personally identifiable information were excluded. Where feasible, cases were selected from multiple providers to reduce operator-specific variability.

XSL•FO
RenderX

Figure 2. Example structure of an image-captured dental chart format. CAA: chronic apical abscess; EPT: electric pulp test; EXT: extraction; GP: gutta percha; L/A: local anesthesia; PA: periapical; PPD: probing pocket depth; WNL: within normal limits.

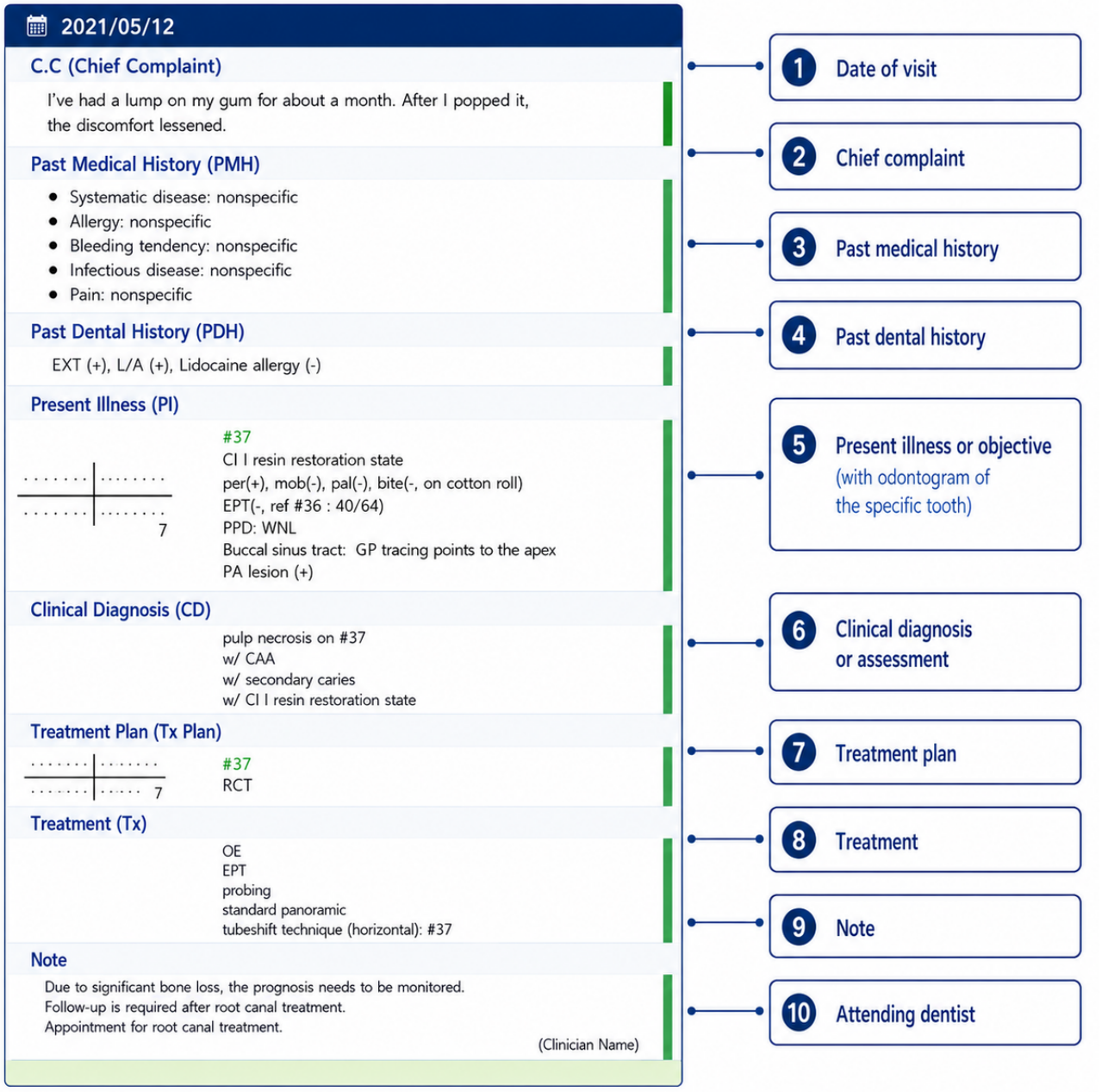


Image Acquisition and Preprocessing

Dental charts were captured as sequential screenshots with a resolution of 717×742 pixels. Consecutive images overlapped by at least one line of text to prevent truncation at page boundaries. The number of images per patient ranged from 3 to 26 (mean 10.27, SD 6.54; total 154 images). Because the testing interface allowed a maximum of 10 images per message, larger cases exceeding this limit were uploaded in consecutive batches within the same conversation thread to preserve page order and context. All images were manually reviewed to ensure legibility and complete deidentification prior to use.

Question Dataset and References

For each patient, 24 individualized questions in Korean were developed, resulting in a total of 360 questions. Questions were evenly distributed across 4 categories (Table 1):

- Type A, general (6 questions): explicit chart facts, such as dates and provider names. Three items were common across all cases: first-visit date, total number of visits, and primary provider name.
- Type B, dental (6 questions): tooth- or procedure-specific clinical details, for example, the date of a specific restoration or the presence of a final crown.
- Type C, interpretation (6 questions): inference questions requiring synthesis of evidence across multiple chart entries, such as the rationale for a procedure.
- Type D, absent information (6 questions): intentionally unanswerable questions designed to test appropriate abstention.

XSL•FO
RenderX

Table . Question types for dental chart evaluation: examples and rationale for large language model (LLM) assessment.

Category	Number of questions	Examples	Rationale for LLM[a] evaluation
General	6	When is the first visit to the clinic?	Evaluation of basic recognition of simple information. Suitable for assessing OCR[b] accuracy and fundamental contextual understanding.
Dental	6	When was tooth number 16 treated for a cavity?	Recognition of technical dental terms. Useful for testing the ability to extract relevant details and apply subject-specific understanding.
Interpretation	6	For what purpose was the MTA[c] used for the treatment?	Assesses the ability to synthesize information from multiple sources and make informed inferences. Enables evaluation of integrated reasoning skills.
Absent information	6	What was the length of MB2[d] canal in tooth number 11? (despite the absence of MB2 canal)	Checks whether the system avoids generating incorrect information not present in the clinical chart. Essential for evaluating hallucination control capabilities.

[a]LLM: large language model.

[b]OCR: optical character recognition.

[c]MTA: mineral trioxide aggregate.

[d]MB2: mesiobuccal 2.

The 4 question categories integrate Bloom taxonomy and established natural language processing (NLP) evaluation paradigms [18]. Types A and B assess foundational information retrieval via single-hop queries, distinguishing baseline reading comprehension (Type A) from dental-specific domain expertise (Type B). Type C evaluates higher-order clinical reasoning, requiring multihop inference across dispersed chart entries. Type D adapts the unanswerable questions paradigm from robust NLP benchmarks (eg, SQuAD 2.0 and emrQA [19]) to assess hallucination control and appropriate abstention, a critical safety prerequisite.

Questions were manually formulated by a single clinical researcher (AYC) to accurately reflect the unique clinical context of each patient case. Methodological consistency was ensured by anchoring all generated questions to the predefined conceptual framework (Types A, B, C, and D). As the primary objective of this study was to compare relative model performance, all evaluated models were presented with the identical set of case-specific questions, ensuring a fair and consistent comparative benchmark. The questions were authored by a single clinical researcher (AYC) and were not independently reviewed for wording prior to evaluation; however, the associated gold-standard answers were subsequently verified by an independent senior specialist.

Gold-standard answers were provided by a licensed dentist and full professor in the Department of Conservative Dentistry at SNUDH, with over 20 years of clinical experience. Two first-year residents from the same department contributed additional comparison answers. In this study, these residents were operationally defined as the “resident clinician group.” This terminology is justified as these individuals hold valid dental licenses and possess approximately 3-4 years of total clinical experience when accounting for their undergraduate clinical training. Consequently, they are considered capable of independent patient management and provided a formally qualified clinical baseline. For the student baseline, 2 third-year dental students at the preclinical-to-clinical transition (2 years of preclinical followed by 2 years of clinical curriculum) produced independent answer sets. This level was chosen to reflect students entering initial clinical exposure with limited prior experience in chart interpretation. All human raters completed the same set of 24 individualized questions per patient independently and without access to model outputs.

Model Conditions and Inference Procedure

Nine ChatGPT models available from August 2-9, 2025, in Korea Standard Time via the ChatGPT (OpenAI) user interface were evaluated, namely OpenAI o3, GPT-4.5, GPT-4.1, GPT-4.1 mini, OpenAI o4-mini, GPT-4o, GPT-5, GPT-5 Thinking, and GPT-5 Pro. Evaluations were performed under a ChatGPT Pro subscription, the highest consumer tier available at the time. The study included all image-capable models selectable in the consumer interface during this window rather than a performance-based subset. Model inclusion was therefore governed by interface availability under this subscription tier; GPT-5 Pro, in particular, was accessible only at the Pro tier. For each patient-model pair, a new conversation was initiated to prevent cross-case contamination. All chart images for the patient, including continuation batches, and the corresponding 24 questions were presented together, accompanied by standardized Korean instructions directing the model to review all images and answer the questions. To ensure methodological transparency and reproducibility, the direct links to the complete conversational logs with the LLMs, including all exact prompts and generated responses, are provided in Multimedia Appendix 1. Conversation history was reset between models and between

XSL•FO
RenderX

patients. Model temperature and other settings were left at platform defaults. All evaluated models support text-and-image inputs in ChatGPT; however, formal optical character recognition (OCR) accuracy metrics for these interfaces are not published. Context limits vary by model (approximately GPT-4.1 up to ~1 million tokens via API, GPT-5 up to ~400,000 tokens via API, and ~256,000 in ChatGPT). Comparative specifications of each model, including type, multimodal data support, context-window size, visual input handling, distinguishing features, and release date, are summarized in Table 2 [20-22]. The consumer interface was used to reflect how students and clinicians actually access these models, prioritizing ecological validity. The trade-off is that platform-level factors—proprietary system prompts, memory-state handling, rate-limiting behavior, and per-query routing decisions, including GPT-5 smart-routing—are not exposed and could not be independently verified or held under experimental control.

Table . Characteristics of ChatGPT models, including type, multimodal data support, context length, image-processing capability, key features, and release date.

AI model	Type	Data support	Context length	Image understanding performance	Key features	Release date
OpenAI o3	Reasoning	Text+image	Not disclosed	Supports visual input; no explicit OCR[a] evaluation	Optimized for structured scientific and coding reasoning, cost-efficient reasoning engine	April 16, 2025
GPT-4.5	GPT	Text+image	128,000 tokens	Supports visual input; no explicit OCR evaluation	Enhanced natural, intuitive dialogue, and aesthetic creativity; reduced hallucination and emotional IQ boost	February 27, 2025
GPT-4.1	GPT	Text+image	1 million tokens	Supports visual input; no explicit OCR evaluation	Large context for vision tasks	April 14, 2025
GPT-4.1 mini	GPT	Text+image	1 million tokens	Same as above; no specific OCR info	Scaled-down version of GPT-4.1; cost-efficient for similar tasks	April 14, 2025
OpenAI o4-mini	Reasoning	Text+image	Not disclosed	Enhanced visual reasoning; OCR not specified	Compact model with speed and accuracy; excels in technical tasks and visual reasoning	April 16, 2025
GPT-4o	GPT	Text, image, audio, and video	128,000 tokens	Strong multimodal; OCR not formally measured	Multimodal (voice, vision, and translation) with state-of-the-art benchmarks	May 13, 2024
GPT-5 (default)	GPT	Text, image, audio, and video	256,000 tokens	Enhanced visual perception; no explicit OCR metrics	Smart routing (fast vs deep mode), high performance in coding, writing, health, and perception	August 7, 2025
GPT-5 Thinking	GPT	Text, image, audio, and video	256,000 tokens	Tuned for deep reasoning; no explicit OCR mention	Tuned for deeper, slower reasoning; selected via router or user prompt	August 7, 2025
GPT-5 Pro	GPT	Text, image, audio, and video	256,000 tokens	Highest compute capability; no explicit OCR performance claim	High-compute, extended reasoning variant best suited for complex tasks	August 7, 2025

[a]OCR: optical character recognition.

Outcomes and Evaluation Metrics

For Type D items, which were designed to assess the ability to recognize missing information, the expert reference answer was standardized as “not recorded.” Because models expressed absence in heterogeneous ways (eg, “no record found” and “not noted in the chart”), simple lexical comparisons could underestimate correct abstention. To address this, all responses

XSL•FO
RenderX

explicitly indicating a lack of information were manually normalized to "not recorded" prior to scoring. This normalization procedure was applied consistently across all models and human raters.

Model outputs were compared against the gold-standard answers. Seven complementary metrics, all normalized to a continuous scale ranging from 0 to 1, were computed to capture both exactness and semantic adequacy:

1. Exact match (EM): the percentage of predictions that exactly match the reference answers, representing a strict, all-or-nothing measure. Values are 0 or 1; for free-text answers, a median of 0 is expected, with 1 indicating verbatim matches.
2. Token-level F_1-score: the harmonic mean of precision and recall calculated at the word (token) level, assessing overlap between predicted and reference text. Values range from 0 to 1, with values near 1 indicating close lexical overlap.
3. Bilingual evaluation understudy (BLEU): evaluates the quality of machine-generated text by measuring n-gram overlap between the prediction and a set of high-quality reference texts [23]. Values range from 0 to 1 (higher is better); values above ~0.30 are considered understandable, and above ~0.50 are good and fluent [24].
4. SacreBLEU: a standardized implementation of BLEU that ensures reproducible scores by handling tokenization and preprocessing consistently [25]. It shares BLEU's 0 - 1 scale and interpretation.
5. Recall-oriented understudy for gisting evaluation-longest common subsequence (ROUGE-L): measures the longest common subsequence between the generated and reference text, emphasizing recall [26]. Values range from 0 to 1; like other overlap metrics, it yields near-zero values for free-text answers.
6. Bidirectional encoder representations from transformers (BERT) score: computes semantic similarity between candidate and reference text by aligning words using contextual embeddings from BERT [27]. BERTScore values range from –1 to 1, although in practice values fall within the 0 - 1 range; roughly, ≥0.90 is faithful, 0.66 - 0.90 adequate, and ≤0.65 meaning-divergent.
7. Sentence-BERT (SBERT): a modification of BERT using Siamese and triplet network architectures to generate semantically meaningful sentence embeddings, suitable for large-scale semantic comparison tasks [28]. Cosine similarity ranges from −1 to 1 but effectively falls within 0-1; commonly cited semantic-similarity thresholds fall in the 0.5 - 0.7 range.

Statistical Analysis

All automatic evaluation metrics were computed against the gold-standard answers on a per-question basis, with all items weighted equally (macro-averaging).

For each ChatGPT model, item-level scores were calculated across all 360 questions from the 15 charts. Distributions of EM, token-level F_1-score, BLEU, SacreBLEU, ROUGE-L, BERTScore, and SBERT were summarized per model using the median and IQR and visualized with boxplots. Model rankings and between-model comparisons were described descriptively.

To compare performance by task category, SBERT was prespecified as the primary metric. For each model and question type—A (general), B (dental), C (interpretation), and D (absent information)—item-level SBERT values were summarized using medians and IQRs and displayed as boxplots to facilitate within-type comparisons between models. This analysis included all 360 questions, with each item weighted equally.

For SBERT human-model comparisons, groups included Student 1, Student 2, Resident 1, Resident 2 (first-year residents), and the highest-performing GPT model. Analyses were conducted separately for question Types A-D. Because score distributions were nonnormal according to Shapiro-Wilk tests, overall group differences were assessed using the Kruskal-Wallis test, followed by Dunn post hoc tests with Holm-Bonferroni adjustment for pairwise contrasts of primary interest (each human group vs GPT-5 Thinking). All tests were 2-sided with a significance threshold of α=.05. Items with missing or empty reference answers were excluded listwise from the relevant comparisons. To convey practical significance alongside *P* values, effect sizes appropriate to the nonparametric design were computed: epsilon-squared (ε^2) for each Kruskal-Wallis omnibus test and Cliff delta (δ) with 95% bootstrap CIs (10,000 resamples) for each prespecified pairwise contrast (each human group vs GPT-5 Thinking), interpreted with Cohen-equivalent benchmarks (ε^2=0.01/0.06/0.14; $|\delta|$=0.147/0.33/0.474 for the small/medium/large boundaries). SBERT embeddings for these comparisons were generated using sentence-transformers/paraphrase-multilingual-MiniLM-L12-v2.

Interrater reliability of the gold-standard answers was assessed in 2 ways. First, an independent second senior specialist (JHK) in Conservative Dentistry reviewed all 360 gold-standard answers against the source charts and rated each as clinically acceptable. Second, because the senior faculty member and the 2 residents had answered the same items independently, each free-text answer was reduced to its canonical clinical value, and chance-corrected agreement was computed—percent agreement, Krippendorff α (nominal), Fleiss κ, and Gwet AC1. Gwet AC1 was prespecified as the primary coefficient because κ can be unstable under near-ceiling agreement. Agreement was assessed overall and by question type for the clinical reference raters (gold standard and 2 residents) and, for transparency, for all 5 human raters. The canonical-value extraction was performed by an LLM used solely as an extractor (not as a judge of clinical correctness) and was validated against an independent dentist on a random 20% subset (50 items).

To assess whether the SBERT-based model ranking was not an artifact of embedding similarity, all 9 models were additionally scored for clinical correctness against the gold value (clinical content for Types A-C and appropriate abstention for Type D), and the resulting content-based clinical-accuracy ranking was compared with the SBERT ranking using Spearman rank correlation. As a safety analysis, the proportion of Type D items on which each model supplied a fabricated or unsupported value instead of abstaining was quantified. As a focused exploratory subanalysis, rather than a full multilingual evaluation, accuracy

XSL•FO
RenderX

was compared between items whose expected answer was in Korean script and those whose expected answer was English or numeric, within a single task type. Because the charts contained both Korean and English within the same record, this comparison assessed the script of the expected answer rather than model performance across separate languages. All source code for model execution is available from the project's GitHub repository [29].

To determine whether discrepancies flagged or missed by the automated metrics were clinically consequential, a prespecified expert-adjudicated subset analysis was performed for the 2 best-performing models. The adjudicated subset comprised all Type A-C responses scored as incorrect in the content-based clinical-accuracy rescoring (n=23), all Type D responses in which a model supplied a value instead of abstaining (n=6), and a random sample of 15 responses per model previously classified as content-correct with low SBERT similarity (<0.60; n=30), the last stratum serving to verify that low semantic similarity did not conceal clinical errors (59 responses in total). Two faculty dentists at a university dental hospital, neither of whom had been involved in question authoring or gold-standard construction, independently rated each response in randomized order, blinded to model identity, automated scores, and subset stratum, on a 3-level scale: 0, no clinical consequence; 1, minor discrepancy unlikely to alter interpretation or management in a supervised educational setting; and 2, clinically consequential, defined as plausibly changing chart interpretation or a management decision or teaching an incorrect clinical fact. The interadjudicator agreement was quantified using Gwet AC1 and linearly weighted Cohen κ, and disagreements were resolved by consensus.

All metric computations and statistical analyses were performed using Python (version 3.13.2; Python Software Foundation) on Windows 10 (Microsoft Corp).

Exploratory Qualitative Feedback

As a secondary exploratory component, written qualitative feedback was collected from the 5 human participants (2 third-year dental students, 2 first-year residents, and the senior faculty dentist who established the gold standard) after they completed the chart-reading tasks. Participants responded in Korean to three open-ended questions concerning (1) the perceived advantages of AI-assisted dental chart interpretation for learning, (2) the perceived disadvantages or concerns, and (3) how the tool compared with traditional faculty-guided chart-reading instruction. Free-text responses were analyzed using a conventional content analysis approach. Two reviewers (MJJ and JI) independently read all responses, inductively coded recurrent ideas, and grouped them into themes for each question. The frequency of each theme was reported as the number of participants expressing it. Representative excerpts were translated from the original Korean and lightly edited for clarity. Because this component was exploratory and based on a small convenience sample, it was intended to contextualize, rather than to formally test, the quantitative benchmark findings. The full thematic summary and excerpts are provided in Table S1 in Multimedia Appendix 2.

Ethical Considerations

The study protocol was approved by the Institutional Review Board (IRB) of the Seoul National University Dental Hospital (CRI26003). Because the study involved only fully deidentified historical dental chart images and no direct patient contact, informed consent was waived in accordance with the IRB guidelines. Voluntary informed consent was obtained from all participating dental students and clinicians prior to their involvement in the study. Data confidentiality and privacy were strictly maintained throughout the study.

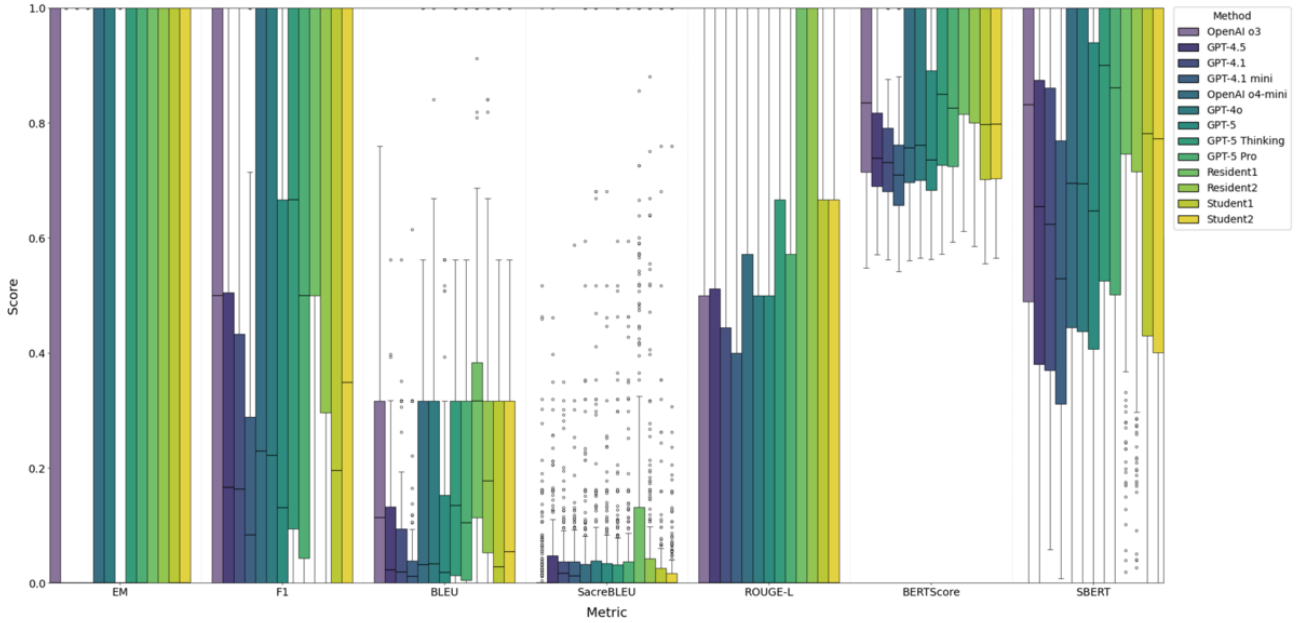
Results

Overview

A total of 360 questions (15 charts×24 questions) were scored against the gold-standard answers. Model alignment, as measured by SBERT, generally exceeded that measured by string-based exactness metrics, indicating greater sensitivity to semantic adequacy than to exact matching. According to SBERT, the highest-performing model was GPT-5 Thinking, with a median score of 0.900 (IQR 0.525 - 1.000), followed by GPT-5 Pro at 0.861 (IQR 0.501 - 1.000) and OpenAI o3 at 0.831 (IQR 0.489 - 1.000). Intermediate scores were observed for OpenAI o4-mini (0.695, IQR 0.444 - 1.000) and GPT-4o (0.694, IQR 0.438 - 1.000). Lower-performing models included GPT-4.5 (0.654, IQR 0.380 - 0.874), GPT-5 (0.647, IQR 0.407 - 0.940), GPT-4.1 (0.623, IQR 0.370 - 0.861), and GPT-4.1 mini (0.529, IQR 0.311 - 0.769). Within this framework, GPT-5 Thinking demonstrated the highest semantic alignment with reference answers, whereas GPT-4.1 mini exhibited the lowest. Full distributions across all 7 evaluation metrics are shown in Figure 3, with per-item scores available in Multimedia Appendix 3.

XSL•FO
RenderX

Figure 3. Comparison of evaluation scores across multiple metrics for ChatGPT models and human responses. Colored boxplots represent item-level scores for each model and human group; higher values indicate closer agreement with reference answers. BERT: bidirectional encoder representations from transformers; BLEU: bilingual evaluation understudy; EM: exact match; ROUGE-L: recall-oriented understudy for gisting evaluation-longest common subsequence; SBERT: sentence bidirectional encoder representations from transformers.



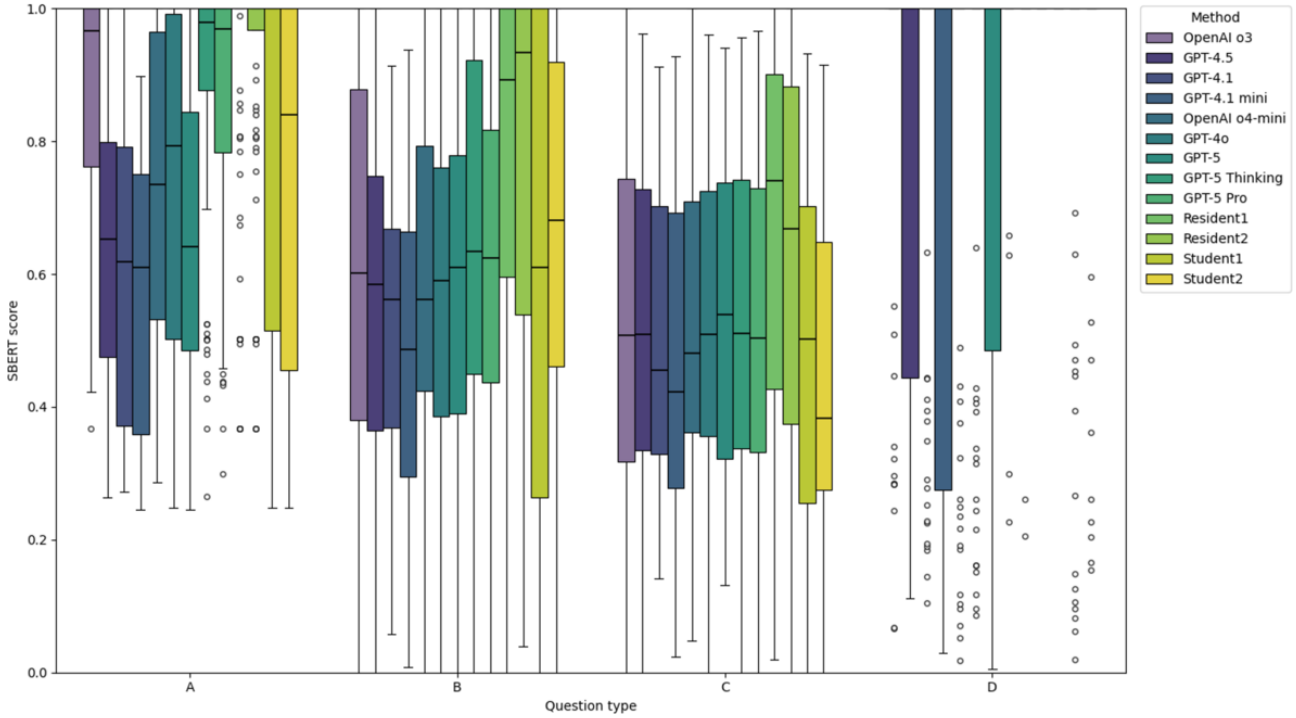
BLEU and SacreBLEU results reproduced the broad performance groupings identified by SBERT but revealed metric-specific reorderings among closely clustered models. ROUGE-L scores highlighted additional local inversions that differed from the SBERT rankings. These cross-metric differences emphasize that exactness-oriented and semantics-oriented metrics capture distinct types of model errors and should be interpreted jointly. In Figure 3, boxplots show that resident clinicians scored at or near the maximum for EM, F_1-score, BERTScore, and SBERT, whereas the LLM and student groups formed lower clusters, with near-zero medians for EM, SacreBLEU, and ROUGE-L.

Stratification by the 4 predefined question types (90 items per type) revealed task-dependent variability in model performance. Figure 4 presents SBERT boxplots by model and question type. GPT-5 Thinking, GPT-5 Pro, and resident clinicians demonstrated higher and narrower score distributions for Types A and B, indicating consistent performance on explicit chart facts and tooth- or procedure-specific items. Type C (interpretation requiring multientry synthesis) showed more overlapping distributions across models, reflecting greater difficulty and variability. For Type D (absent information and abstention control), most groups achieved relatively high scores, indicating effective recognition of missing information. The SBERT-based relative rankings (highest to lowest) for each type were as follows.

- Type A, general (explicit chart facts): GPT-5 Thinking, GPT-5 Pro, OpenAI o3, GPT-4o, OpenAI o4-mini, GPT-4.5, GPT-5, GPT-4.1, and GPT-4.1 mini
- Type B, dental (tooth/procedure-specific): GPT-5 Thinking, GPT-5 Pro, GPT-5, OpenAI o3, GPT-4o, GPT-4.5, GPT-4.1, OpenAI o4-mini, and GPT-4.1 mini
- Type C, interpretation (multientry synthesis): GPT-5, GPT-5 Thinking, GPT-4.5, GPT-4o, OpenAI o3, GPT-5 Pro, OpenAI o4-mini, GPT-4.1, and GPT-4.1 mini
- Type D: absent information (abstention control): GPT-5 Pro, GPT-5 Thinking, OpenAI o3, GPT-4.1, GPT-4o, OpenAI o4-mini, GPT-4.5, GPT-5, and GPT-4.1 mini

XSL•FO
RenderX

Figure 4. Boxplots of sentence bidirectional encoder representations from transformers (SBERT) scores by question type across ChatGPT models. Colored boxplots show item-level SBERT similarity scores for each model and human responses across question Types A-D (A, general; B, dental; C, interpretation; D, absent information); higher scores indicate closer semantic alignment with reference answers. Circles represent outliers. SBERT: sentence bidirectional encoder representations from transformers.



For Types A, B, and D questions, the rank order generally mirrored the overall model analysis. Type A (fact retrieval) demonstrated high agreement with the reference answers; common factual items frequently achieved EM=1, reflecting reliable extraction of explicitly documented fields. In contrast, Type C (interpretation) produced compressed SBERT distributions, with medians ranging from 0.423 (IQR 0.278-0.692; GPT-4.1 mini) to 0.539 (IQR 0.321-0.738; GPT-5), a range of only 0.116 across models, and low EM rates, indicating that no model demonstrated a practically distinct advantage in multientry synthesis. Responses varied in evidence focus (eg, treatment plan, procedure note, or progress entry) and granularity (tooth-level or patient-level), resulting in heterogeneous rationales even when semantic similarity scores were closely clustered. For Type D, multiple models shared identical median scores; therefore, mean SBERT values were used to differentiate the rankings.

Illustrative examples are provided in Table 3 and Table 4, showing a convergent Type B item with uniformly correct factual outputs and a divergent Type C item with heterogeneous rationales and reduced semantic similarity.

XSL•FO
RenderX

Table . Example Type B (tooth-specific) question on the percussion finding of tooth number 36 at the initial visit, showing largely convergent factual answers and corresponding sentence bidirectional encoder representations from transformers (SBERT) similarity scores for human raters and ChatGPT models. SBERT similarity scores range from 0 to 1, with higher values indicating closer semantic agreement with the gold-standard answer. Question: What was the percussion finding on tooth number 36 at the initial visit?

Model	Answer	SBERT similarity score (range 0-1)
Gold Answer	per[a] (−)	1.000
Resident 1	The tooth shows a negative percussion finding; per (−)	0.652
Resident 2	At the initial visit, tooth number 36 showed a negative percussion response	0.187
Student 1	Negative	0.143
Student 2	(–), no response to percussion	0.402
OpenAI o3	Negative percussion response	0.177
GPT-4.5	Percussion test positive (Pal[b]+, distal), bite test negative	0.035
GPT-4.1	At the initial visit, tooth number 36 showed percussion (–), occlusal (–), palatal (+), and bite (–, bite stick)	0.283
GPT-4.1 mini	At the initial visit, tooth number 36 showed per (−), bite (−), pal (+), and distal sinus tract (+)	0.265
OpenAI o4-mini	At the initial examination, tooth number 36 showed a negative response to percussion; per (−)	0.416
GPT-4o	At the initial visit, tooth number 36 showed a positive percussion response on palpation	0.294
GPT-5	At the initial visit, the percussion finding of tooth number 36 was pain on occlusion (−), bite (−), and pain on bite stick (−)	0.187
GPT-5 Thinking	Per (–)	0.569
GPT-5 Pro	Per (–)	0.550

[a]per: percussion.

[b]pal: palpation.

XSL•FO
RenderX

Table . Example Type C (interpretation) question on planned treatment for teeth numbers 37 and 47, illustrating divergent rationales and sentence bidirectional encoder representations from transformers (SBERT) similarity scores across human raters and ChatGPT models. SBERT similarity scores range from 0 to 1, with higher values indicating closer semantic agreement with the gold answer. Question: If treatment is performed on teeth numbers 37[a] and 47[b], what kind of treatment would be done?

Model	Answer	SBERT similarity score (range 0-1)
Gold answer	After performing a CLP[c] by trimming the gum to increase the clinical crown length, caries treatment should be carried out, with RCT[d] if necessary.	1.000
Resident 1	After performing CLP which trims the gum to increase the clinical crown length, caries treatment should be conducted with a possibility of RCT.	0.950
Resident 2	If necessary, caries treatment should be performed; however, CLP must be carried out beforehand, and there is a high likelihood of requiring RCT during the course of treatment.	0.942
Student 1	Distocervical caries can be removed and restored with resin.	0.242
Student 2	A gingivectomy or CLP may be required.	0.380
OpenAI o3	After a subgingival crown lengthening procedure, caries removal and conservative/restorative treatment, including root canal treatment if necessary, are required.	0.600
GPT-4.5	If treatment is performed on teeth numbers 37 and 47, there is a high likelihood that RCT will be required.	0.576
GPT-4.1	If treatment is performed on teeth numbers 37 and 47, regular monitoring of subgingival caries and, if necessary, conservative restorative treatment and root canal therapy will be considered.	0.206
GPT-4.1 mini	If treatment is performed on teeth numbers 37 and 47, a re-evaluation will be conducted after one year, with instructions on the use of interdental brushes and molar brushes, along with regular check-ups and additional treatment if necessary.	0.459
OpenAI o4-mini	If treatment is performed on teeth numbers 37 and 47, follow-up with bitewing X-rays at 6-month intervals is recommended. If symptoms or progression are observed, conservative restorative treatment or, if necessary, root canal therapy should be considered. It was also explained that careful case selection, including consideration of CLP, is required during treatment.	0.441
GPT-4o	It is explained that CLP is required beforehand, and the possibility of RCT is high.	0.805
GPT-5	High likelihood of RCT following CLP.	0.864
GPT-5 Thinking	After CLP, caries removal and restoration will be performed, with RCT if necessary.	0.888
GPT-5 Pro	Crown lengthening, followed by caries removal and restoration, with RCT if necessary.	0.548

[a]37: left mandibular second molar.

[b]47: right mandibular second molar.

[c]CLP: crown-lengthening procedure.

[d]RCT: root canal treatment.

XSL•FO
RenderX

Human-Model Comparison

The descriptive statistics (median and IQR) for each question type are summarized in Table 5, and the corresponding pairwise comparisons with GPT-5 Thinking, with effect sizes, are provided in Table S2 in Multimedia Appendix 2. These comparisons are based on SBERT similarity to the gold standard, which reflects semantic adequacy rather than verified clinical correctness. Embedding similarity can rate clinically opposite findings as similar and clinically equivalent but differently worded answers as dissimilar. These contrasts, therefore, should be read as semantic proximity to the reference and interpreted alongside the content-based clinical accuracy analysis reported below. SBERT scores against the gold-standard answers were compared between GPT-5 Thinking and 4 human baselines—Student 1 and Student 2 (preclinical-to-clinical transition students) and Resident 1 and Resident 2 (first-year residents)—across question Types A-D (n=360), as shown in Figure 5. Because the distributions were nonnormal with unequal variances, Kruskal-Wallis tests indicated significant overall group differences for all 4 question types (all $P<.001$; $\varepsilon^2=0.052$ - 0.091). Dunn post hoc tests with Holm adjustment showed that Student 1 did not differ significantly from GPT-5 Thinking on Type A ($P\geq.99$), Type B ($P\geq.99$), Type C ($P=.39$), or Type D ($P=.16$). Student 2 differed only on Type D, scoring lower than GPT-5 Thinking ($P=.002$; $\delta=-0.27$; 95% CI −0.41 to −0.13). Among resident clinician groups, Resident 1 scored higher than GPT-5 Thinking on Types A ($P=.01$; $\delta=0.29$; 95% CI 0.13-0.44) and C ($P=.04$; $\delta=0.25$; 95% CI 0.08-0.41) but did not differ significantly on Type B ($P=.07$) and scored lower on Type D ($P<.001$; $\delta=-0.35$; 95% CI −0.49 to –0.21). Resident 2 scored higher than GPT-5 Thinking on Types A ($P=.03$; $\delta=0.25$; 95% CI 0.09-0.41), B ($P=.02$; $\delta=0.28$; 95% CI 0.11-0.44), and C ($P=.04$; $\delta=0.26$; 95% CI 0.09-0.42), with no significant difference on Type D ($P=.23$). Overall, effect sizes were negligible to small except for the Resident 1 Type D contrast, which was medium in magnitude. These results indicate that GPT-5 Thinking was broadly comparable to the students across most tasks and remained below resident-level performance for several Types A-C contrasts. Type D warrants particular caution. Because all groups shared identical ceiling medians and IQRs of 1.000, the 2 significant contrasts arose only from a few lower-scoring tail items, reflecting distributional subtleties rather than a real difference in recognizing absent information.

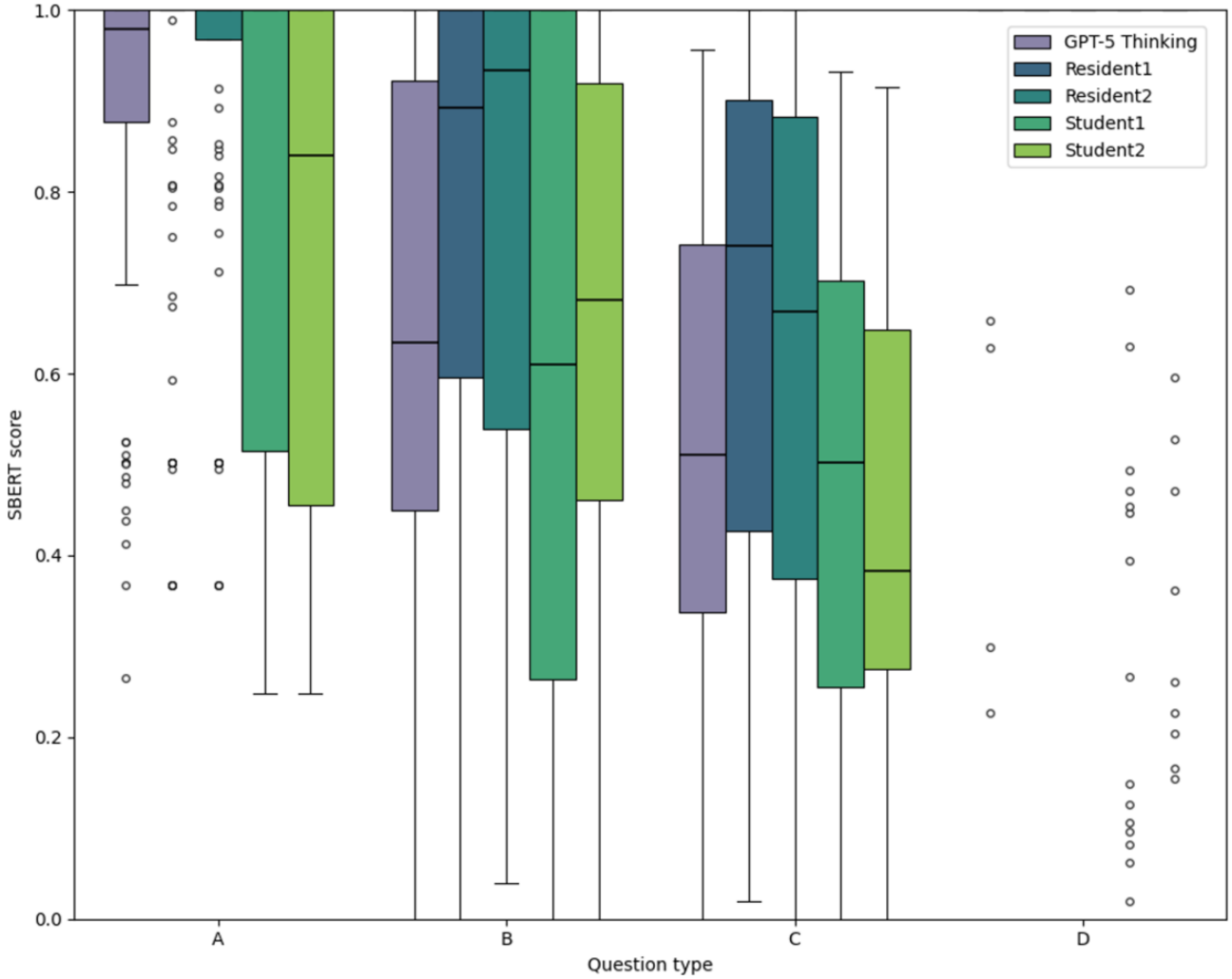
Table . Comparison of sentence bidirectional encoder representations from transformers (SBERT) similarity to the gold-standard answers between GPT-5 Thinking and human raters, by question type.

Question type	GPT-5 Thinking	Resident 1	Resident 2	Student 1	Student 2
SBERT similarity to gold standard, median (IQR)					
Type A (general)	0.978 (0.870 - 1.000)	1.000 (1.000 - 1.000)	1.000 (0.978 - 1.000)	0.994 (0.863 - 1.000)	0.904 (0.744 - 1.000)
Type B (dental)	0.813 (0.525 - 0.969)	0.968 (0.616 - 1.000)	0.991 (0.649 - 1.000)	0.642 (0.336 - 1.000)	0.744 (0.404 - 1.000)
Type C (interpretation)	0.556 (0.371 - 0.650)	0.641 (0.478 - 0.861)	0.688 (0.413 - 0.820)	0.478 (0.221 - 0.607)	0.406 (0.271 - 0.580)
Type D (absent information)	1.000 (1.000 - 1.000)	1.000 (1.000 - 1.000)	1.000 (1.000 - 1.000)	1.000 (1.000 - 1.000)	1.000 (1.000 - 1.000)
Dunn post hoc comparison versus GPT-5 Thinking (*P* value)					
Type A (general)	—[a]	.01	.03	>.99	.37
Type B (dental)	—	.07	.02	>.99	>.99
Type C (interpretation)	—	.04	.04	.39	.13
Type D (absent information)	—	<.001	.23	.16	.002

[a]Not applicable.

XSL•FO
RenderX

Figure 5. Boxplots of sentence bidirectional encoder representations from transformers (SBERT) scores by question type for GPT-5 Thinking and human responses. Colored boxplots show item-level SBERT scores for GPT-5 Thinking, Resident 1 and Resident 2, and Student 1 and Student 2 (third-year dental students). SBERT: sentence bidirectional encoder representations from transformers.



Reliability of the Gold Standard

The independent senior specialist judged all 360 gold-standard answers to be clinically acceptable (100% endorsement). Among the clinical reference raters, agreement on the extracted clinical value was very high (Gwet AC1=0.99 overall and ≥0.98 for every task type; Table S3 in Multimedia Appendix 2), indicating that the gold standard was reproducible across independent, qualified clinicians. The canonical-value extraction agreed with an independent dentist on 97.2% of the validation subset. When the 2 third-year students were included, agreement was lower (Gwet AC1=0.83; Table S4 in Multimedia Appendix 2), reflecting their developing chart-reading competence rather than the unreliability of the reference. SBERT cosine and BERTScore F_1-score showed the same ordering at lower magnitudes (clinical reference raters ≈ 0.85 and 0.87) and are reported only as descriptive checks.

Robustness of the Model Ranking and Safety

To confirm that the ranking reflected clinical content rather than embedding similarity alone, all 9 models were independently rescored for clinical correctness against the gold-standard value. This content-based clinical-accuracy ranking was closely aligned with the SBERT ranking (Spearman ρ=0.90; *P*=.001; Table S5 in Multimedia Appendix 2), with GPT-5 Pro, GPT-5 Thinking, and OpenAI o3 forming the leading group (clinical accuracy 96.4%, 95.6%, and 91.4%, respectively) and GPT-4.1 mini the lowest (68.9%). This convergence indicates that the model ordering was not an artifact of semantic-similarity scoring. On the absent-information (Type D) task, the proportion of fabricated answers in Table S6 in Multimedia Appendix 2, supplying a value the chart did not contain instead of abstaining, rose as model capability fell, from 2.2% (GPT-5 Pro) and 4.4% (GPT-5 Thinking) to 28.9% (GPT-5) and 46.7% (GPT-4.1 mini), whereas all 3 clinical raters abstained appropriately on every such item. This focused exploratory comparison, limited to the tooth- and procedure-specific task, showed no significant difference between Korean-script and English or numeric answers (87.8% vs 90.4%; *P*=.22), indicating that the mixed language input did not by itself drive performance differences within this task. Given its narrow scope, this result is descriptive and does not constitute a comprehensive multilingual evaluation.

Expert Adjudication of Clinically Consequential Discrepancies

The exact agreement between the 2 adjudicators was 91.5% (n=59; Gwet AC1=0.879; linearly weighted κ=0.89), and all 5 disagreements were between adjacent categories and were

XSL•FO
RenderX

resolved by consensus. Of the 59 adjudicated responses, 10 were rated as clinically consequential: 4 of 23 were content-incorrect responses, and all 6 were Type D fabrications. Within the adjudicated subset, clinically consequential ratings occurred only among responses previously classified as content-incorrect or fabricated. The 10 identified cases represented 1.4% of the 720 outputs generated by the 2 models; however, because adjudication was subset-based, this percentage should not be interpreted as a full-sample incidence estimate. None of the 30 sampled low-SBERT responses previously classified as content-correct was rated as clinically consequential, indicating that low SBERT similarity in this sampled stratum reflected wording differences rather than clinical error, as shown in Table S7 in Multimedia Appendix 2.

Exploratory Qualitative Feedback

Content analysis of the open-ended responses from the 5 participants yielded a small set of recurrent themes for each question, summarized with representative excerpts in Table S1 in Multimedia Appendix 2. For perceived advantages, time efficiency was noted by all 5 participants (n=5), followed by accessibility, immediate feedback, repeated learning, and support for self-directed learning (each n=3). For perceived disadvantages, the possibility of AI errors or inaccuracies was raised by all participants (n=5), with the lack of clinical context, overreliance associated with reduced critical thinking, and privacy or data-leakage risks each noted by 2 participants (n=2). In comparing the tool with traditional faculty-guided instruction, most participants (n=4) viewed AI as a supplementary tool to be used alongside faculty rather than as a replacement, and 3 participants (n=3) considered it particularly useful during the early learning stages for terminology and basic chart reading. Notably, 4 participants (n=4) emphasized that faculty expertise remains necessary when contextualized clinical reasoning or treatment decisions are required, indicating that participants regarded AI as most appropriate for foundational, evidence-locatable tasks while reserving complex clinical judgment for faculty guidance.

Discussion

Principal Results

In this image-based benchmark of dental chart interpretation, the main findings were 3-fold. First, reasoning-optimized models led across tasks: GPT-5 Thinking achieved the highest overall semantic alignment with the gold standard (SBERT median 0.900), followed by GPT-5 Pro and OpenAI o3, while the smallest models performed worst, and a content-based clinical-accuracy re-scoring reproduced this ranking (Spearman ρ=0.90; *P*=.001). Second, in the human comparison, GPT-5 Thinking was statistically indistinguishable from dental students on most question-type contrasts but remained below first-year residents on several factual, dental-specific, and interpretive (Types A-C) contrasts, with multientry synthesis (Type C) being the most challenging task across groups. Third, abstention on absent-information (Type D) items was generally reliable, although fabrication rose sharply as model capability fell (from 2.2% for GPT-5 Pro to 46.7% for GPT-4.1 mini), whereas all clinical raters abstained appropriately on every such item. Together, these results position the strongest multimodal models as potential supervised aids for chart-reading practice rather than as substitutes for clinical expertise.

Each question paired a dental chart image with a text prompt, requiring models to parse chart layouts, interpret symbolic markings, and align visual cues with textual questions, similar to emerging evaluations of medical multimodal LLMs in radiology and nuclear medicine, where vision–language integration and domain grounding are critical determinants of performance [30-33]. Consistent with reviews noting that LLM outcomes are strongly task- and modality-dependent and that safety and hallucination control are key concerns in clinical settings [34-38], the heterogeneous patterns across models and task types are compatible with differences in reasoning architecture, safety alignment, and multimodal training described in GPT-5 system documentation [39,40].

Across Type A items (explicit chart-based facts), GPT-5 Thinking, GPT-5 Pro, and OpenAI o3 formed the leading group. For Type B items (tooth- and procedure-specific reasoning), GPT-5 Thinking and GPT-5 Pro again ranked highest, followed by GPT-5. Public documentation describes GPT-5 Thinking as a reasoning-focused variant trained with reinforcement-learning methods and “safe completions” designed to reduce unsafe or overconfident outputs and GPT-5 Pro as a configuration with additional test-time compute [39,40]. These design characteristics align with the relatively stable factual performance observed, though other factors, such as visual processing pipelines and context handling, may also contribute [30-33]. For Type C items, which required synthesis across multiple chart regions and textual entries, GPT-5 outperformed GPT-5 Thinking and GPT-5 Pro. One possible interpretation is that general-purpose tuning and strong long-context capabilities facilitated flexible integration of disparate findings [39,40], whereas the more cautious, stepwise reasoning of GPT-5 Thinking and GPT-5 Pro may have limited the breadth of synthesized interpretations. However, this post hoc explanation cannot be definitively separated from potential influences of context-window use, input batching, or image-handling behavior. Type D items assessed abstention behavior in the presence of missing information. GPT-5 Pro demonstrated the most reliable abstention performance, followed by GPT-5 Thinking and OpenAI o3, whereas GPT-5 performed comparatively poorly. Reasoning-focused models such as GPT-5 Thinking are reported to prioritize safety and adherence to instructions through reinforcement learning and safe completions [39]. The observed pattern is consistent with more conservative responses and aligns with prior work on abstention and hallucination control in safety-tuned models [41-45]; however, this does not constitute direct evidence that specific training procedures caused the observed behavior.

Aggregated across task types, the overall ranking was GPT-5 Thinking, GPT-5 Pro, OpenAI o3, OpenAI o4-mini, GPT-4o, GPT-4.5, GPT-5, GPT-4.1, and GPT-4.1 mini. The leading positions of GPT-5 Thinking and GPT-5 Pro in this multimodal dental setting suggest that, across a mixture of factual, procedural, integrative, and abstention-focused tasks, cautious structured reasoning and calibrated uncertainty may contribute

XSL•FO
RenderX

more to overall utility than maximally flexible generation. The strong performance of OpenAI o3 is consistent with its role as a reasoning-focused predecessor [39]. The relatively high placement of OpenAI o4-mini indicates that some questions—particularly those requiring explicit factual recognition or simple abstention—can be addressed effectively by more parameter-efficient models, consistent with findings in medical multimodal imaging [30-33]. Between-model differences may, however, reflect variation in context-window size, segmented uploads of large charts, differences in OCR and image-processing pipelines, and heterogeneous handling of Korean-language inputs, cautioning against strong claims about intrinsic capability. These findings tentatively support a model-by-task strategy, to be validated prospectively. Within the demanding Type B task, Korean-script and English or numeric answers did not differ (87.8% vs 90.4%; *P*=.22), indicating that code-mixed input did not by itself drive the results.

Against human raters, GPT-5 Thinking was broadly comparable to students and below first-year residents on several extraction and interpretive contrasts. Most student-model comparisons were not statistically significant, although Student 2 scored lower on the absent-information task. Practically, for basic extraction and description of information, GPT-5 Thinking matched students transitioning to clinical training, while remaining below residents on several Type A-C contrasts, consistent with prior education evaluations in which ChatGPT-level systems often reach, but rarely surpass, student performance on written examinations and vignette-style assessments [38,46]. This pattern was clearest on the more "dentally demanding" tasks—Type B (tooth- and procedure-specific reasoning) and Type C (multientry synthesis)—where resident clinicians generally had higher scores than GPT-5 Thinking and students, suggesting differences in prioritization and contextual interpretation. On Type D, all groups clustered near the high end of the score range, frequently recognizing missing information and avoiding overinterpretation, and reached identical ceiling medians. Where pairwise contrasts nonetheless reached significance, the differences stemmed from a few tail items rather than from a systematic gap in abstention and are best read as distributional rather than clinically meaningful. Although GPT-5-family models are described as incorporating approaches intended to reduce hallucination and promote conservative responses under uncertainty [37,39,44], this finding provides reassurance that the model often adopts a safe abstention strategy when key information is missing, though it does not imply parity with clinicians. Quantitatively, the proportion of fabricated answers on absent-information items ranged from 2.2% (GPT-5 Pro) and 4.4% (GPT-5 Thinking) to 28.9% (GPT-5) and 46.7% (GPT-4.1 mini), whereas all 3 clinical raters abstained appropriately on every item, underscoring that the cheapest models are unsafe for unsupervised use. Overall, GPT-5 Thinking is most appropriately framed as a supervised support tool or virtual peer learner, not a substitute for clinician judgment [34,35,37].

The 7 metrics captured different aspects of model behavior. Exactness-oriented metrics (EM, ROUGE-L, and SacreBLEU) were largely uninformative, reaching nonzero medians essentially only for residents because they reward verbatim reproduction rarely expected in open-ended explanations. Lexical-overlap metrics (F_1-score, BLEU) gave a clearer but still surface-level hierarchy (residents >GPT-5 Thinking >OpenAI o3 ≈ GPT-5 Pro >remaining models and students). Semantic metrics were most informative: on SBERT, resident clinicians scored 1.0, GPT-5 Thinking 0.900, GPT-5 Pro 0.861, and OpenAI o3 0.831, with students around 0.77 - 0.78 and other models lower. Because these embedding metrics track sentence-level meaning rather than wording, GPT-5 Thinking, GPT-5 Pro, and OpenAI o3 constitute a "reasoning-optimized" cluster that is semantically closer to resident clinician answers than both students and earlier models. Within the GPT-5 family, the gap between GPT-5 and GPT-5 Thinking further implies that architectural and alignment choices aimed at deeper reasoning translate into measurable gains in clinical content quality, beyond what can be achieved by scale alone [41-44]. Median scores for GPT-5 Thinking fell between students and residents; wide IQRs reflect varying difficulty of the clinical questions, the probabilistic nature of LLM responses, and the individual differences in clinical experience among the students. Importantly, embedding similarity reflects semantic adequacy rather than clinical correctness: it can rate clinically opposite findings as highly similar (eg, per [+] vs per [−]) and clinically equivalent statements as dissimilar when their wording differs. SBERT is therefore best interpreted as a scalable proxy for semantic proximity to the reference, not as a direct measure of clinical correctness, and all SBERT-based human-model contrasts should be interpreted accordingly. To guard against this limitation, the SBERT ranking was corroborated by an independent content-based clinical-accuracy rescoring of all 9 models, which reproduced the same ordering (Spearman ρ=0.90; *P*=.001); this content-based analysis, rather than embedding similarity alone, anchors the interpretation of relative model performance. Results should still be read cautiously given the modest, single-institution sample.

The expert-adjudicated subset analysis reinforces this reading of the automated metrics. Within the adjudicated subset, clinically consequential ratings occurred only among responses previously flagged as content-incorrect or fabricated, and none occurred in the sampled low-similarity responses previously classified as content-correct. The 10 identified cases represent 1.4% of the 720 outputs generated by the 2 models; however, because adjudication was subset-based, this percentage should not be interpreted as a full-sample incidence estimate. Because adjudication was limited to the 2 best-performing models and used a rubric-based consequence scale rather than observed educational outcomes, these findings characterize the adjudicated subset within this benchmark rather than clinical performance.

Implications for Dental Education

From an educational standpoint, the findings suggest that GPT-5 Thinking could potentially serve as a supervised reference and secondary-reader tool within chart-reading training. It is particularly suited for exercises emphasizing evidence identification, omission detection, and uncertainty recognition, skills central to safe documentation and clinical reasoning.

XSL•FO
RenderX

Given the safety risks associated with unsupervised LLM errors in direct patient care, focusing on chart interpretation rather than definitive diagnosis better reflects the potential utility of LLMs as safe cognitive training tools in risk-free educational settings. Because the model matched novice learners on most contrasts but stayed below residents on several Type A-C tasks (notably Type B and C), human supervision remains essential. Rather than replacing instructor guidance, multimodal LLMs can facilitate guided education, allowing students to query chart content, test hypotheses, and receive structured feedback under supervision. This perspective aligns with automation-bias literature emphasizing verification and guardrails [47] and with the findings by Hattie and Timperley [48] that high-quality feedback is an important influence on achievement, suggesting that multimodal LLMs may be most beneficial when integrated into feedback-rich, supervised learning activities.

Practical integration into dental curricula can be achieved through digital exercises using deidentified charts, case-based seminars, and self-directed study tasks. These activities should emphasize locating relevant entries, reconstructing treatment timelines, and verifying documented procedures, thereby helping students develop both chart-reading literacy and AI literacy. Advances in multimodal and interactive interfaces further align training with clinical practice, allowing students to upload chart images, use voice input during simulations, and receive immediate feedback [49].

Safe and responsible adoption requires institutional policies that define permissible uses, mandate verification against original charts, and record model details for assignments. The use of deidentified data and audit logs should be standard to ensure transparency and accountability. Within these guardrails, multimodal LLMs can function as collaborative aids in guided education, promoting deliberate practice and structured feedback while maintaining professional oversight. This approach aligns with Masters' recommendations for ethical AI use in health professions education, which emphasize governance, transparency, and accountability frameworks [50].

An exploratory qualitative feedback analysis obtained from study participants and summarized in Table S1 in Multimedia Appendix 2 provided additional insight into the perceived educational usefulness of AI-assisted dental chart interpretation. Participants consistently identified time efficiency, accessibility, immediate feedback, and support for self-directed learning as key advantages of the tool. These findings are consistent with emerging evidence suggesting that generative AI can function as an on-demand educational resource that enhances learner autonomy and supports personalized learning in health professions education [51]. Prior studies have also suggested that LLMs may serve as valuable educational aids by providing rapid access to information, supporting formative feedback, and facilitating independent learning outside traditional instructional settings [8]. Furthermore, AI tools may function as cognitive scaffolds that help learners navigate complex clinical information while promoting active engagement with educational content and supporting the development of foundational competencies [51,52]. Notably, participants perceived the tool as useful in early clinical training, when learners are first exposed to unfamiliar terminology, chart structures, and clinical documentation. This finding suggests that multimodal AI may facilitate the preclinical-to-clinical transition by providing timely support and opportunities for deliberate practice.

Participants also expressed concerns about inaccuracies, overreliance, privacy, and the inability of AI systems to fully capture contextual reasoning. Importantly, both students and residents regarded AI as a supplementary educational tool rather than a replacement for faculty-guided instruction, particularly when complex clinical reasoning and judgment were required. These perceptions align with recent discussions in medical education emphasizing that generative AI should augment rather than replace human expertise and that learners must be trained to critically evaluate AI-generated outputs [53]. Concerns regarding cognitive dependency and reduced independent reasoning have likewise been highlighted as important challenges associated with the educational use of generative AI [52,53]. Together, these findings support multimodal AI as an adjunctive learning technology that enhances chart-reading literacy under supervision while preserving the central educational role of faculty.

Comparison With Prior Work

This study makes 2 novel contributions. First, it represents the first systematic evaluation of multimodal LLMs for dental chart image interpretation. Second, it implements a multiaxis performance assessment across 4 distinct task types, including general fact retrieval (Type A), tooth- or procedure-level retrieval (Type B), interpretive and inference reasoning (Type C), and absent-information and abstention control (Type D).

Prior studies have primarily examined LLMs' ability to extract information from text-based inputs, such as discharge summaries or narrative clinical notes. While these tasks assess factual comprehension, they omit the spatial information inherent in clinical records, particularly in dental charts that include odontograms, tooth-number grids, and procedure maps. Linearizing such records into text disrupts the contextual relationships between spatial elements and their annotations. Evaluating image-based chart inputs, therefore, probes multimodal reasoning under more realistic conditions, reflecting how clinicians interpret dental charts by integrating both spatial and textual cues. Although multimodal capability has been documented in several medical imaging domains, such as radiography and pathology slide analysis [30], comparable evaluations using dental chart images have not previously been reported, motivating the present image-based benchmarking.

With respect to the second contribution, most prior evaluations have relied on multiple-choice or short-answer formats, such as the United States Medical Licensing Examination (USMLE), reporting aggregate accuracy without decomposing performance by task type [8]. The present findings align with previous reports indicating that LLMs can reach or exceed passing thresholds on structured, well-specified examinations but remain less reliable on open-ended reasoning and fact-checking tasks [7,8]. Reviews have similarly emphasized hallucination and inconsistent abstention as persistent limitations, highlighting the need for human verification and explicit guardrails in both educational and clinical contexts [54,55]. In dental education,

XSL•FO
RenderX

a mixed methods study of ChatGPT in undergraduate training reported benefits for rapid clarification, idea generation, and formative feedback, while also noting factual errors and over-reliance, prompting recommendations for supervised, source-verified use [15].

Language effects further contextualize the present results. Earlier work suggested higher LLM performance in English compared with other languages; however, recent multimodal models (eg, GPT-4o) have narrowed this gap, demonstrating improved correctness and responsiveness in Japanese across general and medical imaging tasks, including radiology [56]. In dentistry, a peer-reviewed evaluation of GPT-4o and Gemini Advanced on the Korean National Dental Licensing Examination documented strong accuracy and consistency on well-specified, evidence-locatable items [57]. Emerging assessments of GPT-5 models in Korean similarly report robust performance on structured, multimodal tasks, findings broadly consonant with the present study's mixed Korean and English chart data. Collectively, these studies suggest that recent multimodal LLM families are reducing prior language penalties, particularly when queries are context-bounded and answers can be anchored to identifiable evidence. The present language comparison, however, was a focused exploratory subanalysis within a single task rather than a controlled multilingual evaluation, so it can only contextualize these findings and cannot establish language equivalence.

Practical Guidance for Use

The highest-performing models in the study, such as GPT-5 Thinking and GPT-5 Pro, achieved the greatest accuracy but required longer response times. In time-sensitive educational contexts, such as self-directed chart-reading practice or preparation for case-based seminars, faster models like GPT-4o or OpenAI o4-mini may be more practical, as modest reductions in accuracy are offset by faster responses [58]. Two key implications emerge for educational use. First, model selection should be guided by task demands: faster models are suitable for routine information retrieval during practice, whereas reasoning-tuned models are preferable when accuracy and abstention are critical. Second, all outputs should be verified against the original record before clinical use, given the potential for errors when documentation is incomplete or absent [12,13]. Because the models were accessed through the consumer interface, exact compute cost and latency were not measured. Any cost-related guidance is therefore provisional and based on accuracy alone. On that basis, GPT-5 Thinking reached clinical accuracy close to the larger GPT-5 Pro, suggesting a more cost-effective option for educational deployment, whereas the smallest models, though inexpensive, abstained poorly and are not advisable for unsupervised use. These observations should be confirmed by a formal API-based benchmark of cost and latency alongside accuracy, which remains an important next step. For educators and learners considering practical use, the main conclusions can be summarized at a glance: across factual, dental-specific, interpretive, and abstention tasks, the strongest reasoning-optimized models (GPT-5 Thinking, GPT-5 Pro, and OpenAI o3) matched novice-student retrieval and abstained reliably, yet remained below resident-level interpretive reasoning, and fabrication of absent information rose steeply as model capability declined (from 2.2% to 46.7%). The practical guidance is therefore consistent across all 4 task types—multimodal LLMs are best used as supervised, verification-focused aids whose outputs are checked against the original record, rather than as autonomous interpreters. Although demonstrated here for dental records, this guidance parallels emerging questions about record-reading and documentation literacy in medical and nursing practice; whether the same supervised, verify-before-use framework transfers to other specialties or to adjacent health-profession settings remains an open question requiring external validation.

Although this study focused on educational benchmarking, one possible area for future investigation is whether multimodal LLMs could support clinical information retrieval from lengthy, mixed-language dental records. For example, identifying the date of a composite resin restoration on the upper right first premolar across multiple years of visits typically requires extensive manual review. A single structured query to an LLM can quickly surface the relevant note and date for confirmation, reducing search time and cognitive load. However, as this clinical-retrieval scenario was outside the scope of the present benchmark, it is best regarded as a direction for future prospective, safety-focused studies.

Limitations

The dataset was drawn from a single department (Conservative Dentistry) from a single institution (SNUDH) and included only 15 charts. This study should therefore be regarded as an internal, single-center benchmark. Because multimodal chart interpretation depends heavily on local layout, documentation conventions, abbreviations, and specialty-specific workflows, which vary across institutions and electronic dental record systems, the single-center design carries an inherent risk of institutional bias, and the findings may not be directly generalizable to other specialties, institutions, record systems, or linguistic settings without external validation. Records contained both Korean and English entries, and results may differ in other language contexts.

The analysis reflected model versions selectable in the consumer interface under a Pro-tier subscription in early August 2025; model availability is both interface- and subscription-dependent and may differ at other times, under different tiers, or via API access. In addition, the models were accessed through the consumer interface, so exact token usage, latency, and compute cost were not measured.

Human comparisons were limited to 2 students and 2 residents, as their primary role was to establish a reference baseline for the AI models rather than to serve as a large-scale population sample. Nevertheless, this reduces the statistical precision of group contrasts and warrants confirmation in studies with larger human cohorts.

A further consideration concerns the independence of observations. Because the 24 questions for each patient were derived from the same chart, items from the same chart may share case-level characteristics such as complexity, documentation style, and image quality, so their scores are not fully independent. The nonparametric tests applied here treat

XSL•FO
RenderX

items as independent and do not account for this within-chart clustering. The effective sample size is therefore smaller than the nominal number of items, and the item-level comparisons should be regarded as exploratory. Confirmatory analyses using clustered or mixed effects models with the chart as a grouping factor would help verify the robustness of these findings.

Reliability and scoring relied on semantic and content-based proxies rather than a formal clinical-harm rubric; although SBERT was triangulated with content-level agreement and canonical-value extraction was validated against an independent dentist on a 20% subset (97.2% agreement), residual error cannot be excluded. The expert-adjudicated subset analysis addressed clinically consequential discrepancies within this benchmark (Table S7 in Multimedia Appendix 2), but prospective grading of clinical harm under real educational use, by a larger expert panel, remains important future work.

In addition, the 360 questions were authored by a single clinical researcher (AYC) and were not independently reviewed for wording before evaluation, which may introduce authoring bias. This is most relevant for the interpretive Type C items, where multiple clinically acceptable framings and differing evidence-prioritization strategies can exist, potentially shaping item difficulty, ambiguity, and, in principle, model ranking. Several design features mitigate this concern—the identical question set was applied to all models, the gold-standard answers were independently judged clinically acceptable with high interrater agreement (Gwet AC1=0.99), and the SBERT-based ranking was reproduced by a content-based clinical-accuracy rescoring (Spearman ρ=0.90)—but confirmation with independently reviewed or multiauthor question sets remains warranted.

Several methodological constraints also warrant caution. First, because evaluation was end-to-end, OCR accuracy was not independently measured, and the chart resolution (717×742 pixels), the maximum obtainable resolution from the EMR interface, may have constrained visual parsing; this nonetheless reflects the native, constrained outputs that students and clinicians actually use, making robustness to such inputs relevant to real-world utility. Second, evaluations used the consumer graphical user interface (GUI) rather than the API, and the study was not preregistered, as it was an exploratory benchmark. This reflects a deliberate trade-off between ecological validity and computational reproducibility [59]. The GUI captures how students and clinicians actually interact with these models, but it introduces GUI-specific uncontrolled variables such as proprietary system prompts, memory-state handling, rate limiting, and dynamic model routing that are not under experimental control. These factors are further amplified for the larger charts, whose images exceeded the 10-image message limit and were therefore uploaded in consecutive batches within a single conversation thread. Consequently, the present results should be interpreted as real-world interface benchmarking rather than fully controlled model benchmarking. Because the behavior of the same commercial LLM service has been shown to shift substantially over short periods as providers update it without public notice [60], the observed model rankings may not be fully stable under future platform updates or under API-based, version-pinned evaluation. Finally, as a comparative study with students, this study cannot demonstrate whether using these models improves learning outcomes. Future work should use larger, multi-institutional, multispecialty, and ideally multinational datasets, complement this ecological benchmark with API-based, version-pinned, preregistered evaluations, and incorporate formal clinical grading rubrics to assess educational impact, safety, response time, and efficiency under real clinical conditions.

Conclusions

This study compared multimodal ChatGPT models with dental students and first-year residents in reading and interpreting dental chart images. GPT-5 Thinking demonstrated the highest overall SBERT performance and was broadly comparable to dental students, with no significant difference from Student 1 across Types A-D and only one significant contrast with Student 2 on Type D. Compared with residents, however, it remained lower on several Type A-C contrasts, particularly Type C and one Type B comparison; Type D findings should be interpreted cautiously because all groups had ceiling medians. These findings indicate that current multimodal models can approach student-level comprehension in structured chart reading but remain below resident-level reasoning in several resident comparisons. In educational settings, such models may serve as supervised support tools for practicing chart interpretation and verification. These findings derive from an internal, single-center benchmark, and external validation across multiple institutions, dental specialties, record systems, and language settings is required before broader claims can be made. As multimodal capabilities continue to advance, carefully integrated AI systems may have the potential to support early clinical training through guided education, offering scalable, feedback-driven support to strengthen chart-reading literacy.

Acknowledgments

The authors thank the Dental Research Institute for its support and the students and residents who participated in and assisted with this study. The authors also thank Seoul National University Dental Hospital for providing access to the clinical records used in this research. We used AI tools (ChatGPT) because the study evaluates and analyzes ChatGPT's responses in direct comparison with human responses.

Funding

This work was supported by the Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education, Science and Technology (RS-2023-NR077207).

XSL•FO
RenderX

Data Availability

The datasets generated or analyzed during this study are available in the GitHub repository [29] and in the supplementary information files (Multimedia Appendices 1 and 3). The source dental chart images are not publicly available and cannot be shared owing to patient-privacy restrictions.

Authors' Contributions

Conceptualization: AYC, DGS, JHK, JI, MJJ, SHE
Data curation: AYC, DGS, SHE
Formal analysis: SHE, JI, MJJ
Investigation: AYC, SHE, DGS, JI
Methodology: AYC, DGS, SHE
Project administration: JHK, JI, MJJ
Resources: DGS, SHE, JI, JHK
Software: SHE, JI, DGS
Supervision: DGS, JHK, JI, MJJ
Validation: JHK, JI, MJJ
Visualization: AYC, DGS, SHE
Writing – original draft: AYC, SHE, DGS
Writing – review & editing: AYC, DGS, JHK, JI, MJJ, SHE

Conflicts of Interest

None declared.

Multimedia Appendix 1
Dataset of 360 case-specific questions (in both the original Korean and translated English), human reference answers, raw outputs from the 9 evaluated models, and direct URL links to the original conversational logs.
[XLSX File, 272 KB - mededu_v12i1e91809_app1.xlsx]

Multimedia Appendix 2
Supplementary tables including exploratory qualitative feedback on educational utility, interrater reliability of the human reference standard, robustness of the evaluation metrics, model safety (fabrication rates), effect sizes for human-model comparisons, and expert adjudication of clinically consequential discrepancies.
[DOCX File, 23 KB - mededu_v12i1e91809_app2.docx]

Multimedia Appendix 3
Complete per-item metric evaluation results across 360 questions, comparing each response with the gold-standard references using 7 metrics (Exact Match, F_1-score, BLEU, SacreBLEU, ROUGE-L, BERTScore, and SBERT). The dataset includes results for all evaluated models as well as student and resident raters.
[XLSX File, 363 KB - mededu_v12i1e91809_app3.xlsx]

Checklist 1
STROBE checklist.
[DOCX File, 18 KB - mededu_v12i1e91809_app4.docx]

References

1. van Merriënboer JJG, Sweller J. Cognitive load theory in health professional education: design principles and strategies. Med Educ 2010 Jan;44(1):85-93. [doi: 10.1111/j.1365-2923.2009.03498.x] [Medline: 20078759]
2. Young JQ, Van Merrienboer J, Durning S, Ten Cate O. Cognitive load theory: implications for medical education: AMEE Guide no.86. Med Teach 2014 May;36(5):371-384. [doi: 10.3109/0142159X.2014.889290] [Medline: 24593808]
3. Gordon M, Daniel M, Ajiboye A, et al. A scoping review of artificial intelligence in medical education: BEME Guide no. 84. Med Teach 2024 Apr;46(4):446-470. [doi: 10.1080/0142159X.2024.2314198] [Medline: 38423127]
4. Zitzmann NU, Matthisson L, Ohla H, Joda T. Digital undergraduate education in dentistry: a systematic review. Int J Environ Res Public Health 2020 May 7;17(9):3269. [doi: 10.3390/ijerph17093269] [Medline: 32392877]
5. Levitin SA, Grbic JT, Finkelstein J. Completeness of electronic dental records in a student clinic: retrospective analysis. JMIR Med Inform 2019 Mar 21;7(1):e13008. [doi: 10.2196/13008] [Medline: 30896435]

XSL•FO
RenderX

6. Acharya A, Schroeder D, Schwei K, Chyou PH. Update on electronic dental record and clinical computing adoption among dental practices in the United States. Clin Med Res 2017 Dec;15(3-4):59-74. [doi: 10.3121/cmr.2017.1380] [Medline: 29229631]
7. Gilson A, Safranek CW, Huang T, et al. How does ChatGPT perform on the United States Medical Licensing Examination (USMLE)? The implications of large language models for medical education and knowledge assessment. JMIR Med Educ 2023 Feb 8;9:e45312. [doi: 10.2196/45312] [Medline: 36753318]
8. Kung TH, Cheatham M, Medenilla A, et al. Performance of ChatGPT on USMLE: potential for AI-assisted medical education using large language models. PLoS Digit Health 2023 Feb;2(2):e0000198. [doi: 10.1371/journal.pdig.0000198] [Medline: 36812645]
9. Eysenbach G. The role of ChatGPT, generative language models, and artificial intelligence in medical education: a conversation with ChatGPT and a call for papers. JMIR Med Educ 2023 Mar 6;9:e46885. [doi: 10.2196/46885] [Medline: 36863937]
10. Shorey S, Mattar C, Pereira TLB, Choolani M. A scoping review of ChatGPT's role in healthcare education and research. Nurse Educ Today 2024 Apr;135:106121. [doi: 10.1016/j.nedt.2024.106121] [Medline: 38340639]
11. Cheng Y, Zhu L. A review of ChatGPT in medical education: exploring advantages and limitations. Int J Surg 2025 Jul 1;111(7):4586-4602. [doi: 10.1097/JS9.0000000000002505] [Medline: 40465793]
12. Ji Z, Lee N, Frieske R, et al. Survey of hallucination in natural language generation. ACM Comput Surv 2023 Dec 31;55(12):1-38. [doi: 10.1145/3571730]
13. de Hond A, Leeuwenberg T, Bartels R, et al. From text to treatment: the crucial role of validation for generative large language models in health care. Lancet Digit Health 2024 Jul;6(7):e441-e443. [doi: 10.1016/S2589-7500(24)00111-0] [Medline: 38906607]
14. Benoit B, Frédéric B, Jean-Charles D. Current state of dental informatics in the field of health information systems: a scoping review. BMC Oral Health 2022 Apr 19;22(1):131. [doi: 10.1186/s12903-022-02163-9] [Medline: 35439988]
15. Kavadella A, Dias da Silva MA, Kaklamanos EG, Stamatopoulos V, Giannakopoulos K. Evaluation of ChatGPT's real-life implementation in undergraduate dental education: mixed methods study. JMIR Med Educ 2024 Jan 31;10:e51344. [doi: 10.2196/51344] [Medline: 38111256]
16. Abril-Gonzalez M, Portilla FA, Jaramillo-Mejia MC. Standard health level seven for odontological digital imaging. Telemed J E Health 2017 Jan;23(1):63-70. [doi: 10.1089/tmj.2015.0251] [Medline: 27248059]
17. Giannakopoulos K, Kavadella A, Aaqel Salim A, Stamatopoulos V, Kaklamanos EG. Evaluation of the performance of generative AI large language models ChatGPT, Google Bard, and Microsoft Bing Chat in supporting evidence-based dentistry: comparative mixed methods study. J Med Internet Res 2023 Dec 28;25:e51580. [doi: 10.2196/51580] [Medline: 38009003]
18. Adams NE. Bloom's taxonomy of cognitive learning objectives. J Med Libr Assoc 2015 Jul;103(3):152-153. [doi: 10.3163/1536-5050.103.3.010] [Medline: 26213509]
19. Pampari A, Raghavan P, Liang J, Peng J. emrQA: a large corpus for question answering on electronic medical records. 2018 Presented at: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing; Oct 31 to Nov 4, 2018; Brussels, Belgium p. 2357-2368 URL: http://aclweb.org/anthology/D18-1 [accessed 2026-08-05] [doi: 10.18653/v1/D18-1258]
20. Thirunavukarasu AJ, Ting DSJ, Elangovan K, Gutierrez L, Tan TF, Ting DSW. Large language models in medicine. Nat Med 2023 Aug;29(8):1930-1940. [doi: 10.1038/s41591-023-02448-8] [Medline: 37460753]
21. Raza MM, Venkatesh KP, Kvedar JC. Generative AI and large language models in health care: pathways to implementation. NPJ Digit Med 2024 Mar 7;7(1):62. [doi: 10.1038/s41746-023-00988-4] [Medline: 38454007]
22. Arora A, Arora A. The promise of large language models in health care. Lancet 2023 Feb 25;401(10377):641. [doi: 10.1016/S0140-6736(23)00216-7] [Medline: 36841609]
23. Papineni K, Roukos S, Ward T, Zhu WJ. BLEU: a method for automatic evaluation of machine translation. 2002 Presented at: 40th Annual Meeting of the Association for Computational Linguistics (ACL); Jul 6-12, 2002; Philadelphia, PA p. 311-318. [doi: 10.3115/1073083.1073135]
24. Denkowski M, Lavie A. Meteor universal: language specific translation evaluation for any target language. 2014 Presented at: Proceedings of the Ninth Workshop on Statistical Machine Translation; Jun 26-27, 2014; Baltimore, Maryland p. 376-380 URL: http://aclweb.org/anthology/W14-33 [accessed 2026-08-05] [doi: 10.3115/v1/W14-3348]
25. Post M. A call for clarity in reporting BLEU scores. 2018 Presented at: Proceedings of the Third Conference on Machine Translation; Oct 31 to Nov 1, 2018; Belgium, Brussels p. 186-191 URL: http://aclweb.org/anthology/W18-63 [accessed 2026-08-05] [doi: 10.18653/v1/W18-6319]
26. Lin CY. ROUGE: a package for automatic evaluation of summaries. Presented at: Text Summarization Branches Out: Proceedings of the ACL-04 Workshop; Jul 25-26, 2004; Barcelona, Spain p. 74-81 URL: https://aclanthology.org/W04-1013/ [accessed 2026-08-05]
27. Zhang T, Kishore V, Wu F, Weinberger KQ, Artzi Y. BERTScore: Evaluating text generation with BERT. Presented at: 2020 International Conference on Learning Representations; Apr 27-30, 2020 URL: https://openreview.net/pdf?id=SkeHuCVFDr [accessed 2026-08-19]

XSL•FO
RenderX

28. Reimers N, Gurevych I. Sentence-BERT: sentence embeddings using siamese BERT-networks. 2019 Presented at: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP); Nov 3-7, 2019; Hong Kong, China p. 3982 URL: https://www.aclweb.org/anthology/D19-1 [accessed 2026-08-05] [doi: 10.18653/v1/D19-1410]
29. Eo SH. Benchmarking multimodal LLMs on dental chart image interpretation: a comparison with students and clinicians. GitHub. URL: https://github.com/sooheang/eval-genai-dentistry [accessed 2026-06-07]
30. Nam Y, Kim DY, Kyung S, et al. Multimodal large language models in medical imaging: current state and future directions. Korean J Radiol 2025 Oct;26(10):900-923. [doi: 10.3348/kjr.2025.0599] [Medline: 41015856]
31. Bradshaw TJ, Tie X, Warner J, Hu J, Li Q, Li X. Large language models and large multimodal models in medical imaging: a primer for physicians. J Nucl Med 2025 Feb 3;66(2):173-182. [doi: 10.2967/jnumed.124.268072] [Medline: 39819692]
32. Li CY, Chang KJ, Yang CF, et al. Towards a holistic framework for multimodal LLM in 3D brain CT radiology report generation. Nat Commun 2025 Mar 6;16(1):2258. [doi: 10.1038/s41467-025-57426-0] [Medline: 40050277]
33. Yi Z, Xiao T, Albert MV. A survey on multimodal large language models in radiology for report generation and visual question answering. Information 2025;16(2):136. [doi: 10.3390/info16020136]
34. Strasser LM, Anschuetz W, Dennstädt F, Hastings J. Performance evaluation of large language models in multilingual medical multiple-choice questions: mixed methods study. JMIR Med Educ 2026 Mar 5;12:e81399. [doi: 10.2196/81399] [Medline: 41813244]
35. Croxford E, Gao Y, Pellegrino N, et al. Current and future state of evaluation of large language models for medical summarization tasks. NPJ Health Syst 2025;2(1):6. [doi: 10.1038/s44401-024-00011-2] [Medline: 40124388]
36. Aster A, Laupichler MC, Rockwell-Kollmann T, Masala G, Bala E, Raupach T. ChatGPT and other large language models in medical education — scoping literature review. MedSciEduc 2025 Feb;35(1):555-567. [doi: 10.1007/s40670-024-02206-6]
37. Roustan D, Bastardot F. The clinicians' guide to large language models: a general perspective with a focus on hallucinations. Interact J Med Res 2025 Jan 28;14:e59823. [doi: 10.2196/59823] [Medline: 39874574]
38. Lucas HC, Upperman JS, Robinson JR. A systematic review of large language models and their implications in medical education. Med Educ 2024 Nov;58(11):1276-1285. [doi: 10.1111/medu.15402] [Medline: 38639098]
39. OpenAI. GPT-5 system card.: OpenAI; 2025. URL: https://cdn.openai.com/gpt-5-system-card.pdf [accessed 2025-10-11]
40. Rao AS, Esmail KP, Lee RS, et al. Large language model performance and clinical reasoning tasks. JAMA Netw Open 2026 Apr 1;9(4):e264003. [doi: 10.1001/jamanetworkopen.2026.4003] [Medline: 41973425]
41. Huang L, Yu W, Ma W, et al. A survey on hallucination in large language models: principles, taxonomy, challenges, and open questions. ACM Trans Inf Syst 2025 Mar 31;43(2):1-55. [doi: 10.1145/3703155]
42. Wen B, Howe B, Wang LL. Characterizing LLM abstention behavior in science QA with context perturbations. Presented at: Findings of the Association for Computational Linguistics; Nov 12-16, 2024; Miami, FL p. 3437-3450 URL: https://aclanthology.org/2024.findings-emnlp [accessed 2026-08-06] [doi: 10.18653/v1/2024.findings-emnlp.197]
43. Lee M. A mathematical investigation of hallucination and creativity in GPT models. Mathematics 2023;11(10):2320. [doi: 10.3390/math11102320]
44. Asgari E, Montaña-Brown N, Dubois M, et al. A framework to assess clinical safety and hallucination rates of LLMs for medical text summarisation. NPJ Digit Med 2025 May 13;8(1):274. [doi: 10.1038/s41746-025-01670-7] [Medline: 40360677]
45. Omar M, Sorin V, Collins JD, et al. Multi-model assurance analysis showing large language models are highly vulnerable to adversarial hallucination attacks during clinical decision support. Commun Med (Lond) 2025 Aug 2;5(1):330. [doi: 10.1038/s43856-025-01021-3] [Medline: 40753316]
46. Ullah R, Shaikh MS, Shahani N, Lone MA, Fareed MA, Zafar MS. Comparing ChatGPT and dental students' performance in an introduction to dental anatomy examination: a cross-sectional study. Eur J Dent 2026 Feb;20(1):287-294. [doi: 10.1055/s-0045-1808254] [Medline: 40359999]
47. Lyell D, Coiera E. Automation bias and verification complexity: a systematic review. J Am Med Inform Assoc 2017 Mar 1;24(2):423-431. [doi: 10.1093/jamia/ocw105] [Medline: 27516495]
48. Hattie J, Timperley H. The power of feedback. Rev Educ Res 2007 Mar;77(1):81-112. [doi: 10.3102/003465430298487]
49. Claman D, Sezgin E. Artificial intelligence in dental education: opportunities and challenges of large language models and multimodal foundation models. JMIR Med Educ 2024 Sep 27;10:e52346. [doi: 10.2196/52346] [Medline: 39331527]
50. Masters K. Ethical use of artificial intelligence in health professions education: AMEE Guide no. 158. Med Teach 2023 Jun;45(6):574-584. [doi: 10.1080/0142159X.2023.2186203] [Medline: 36912253]
51. Pham TD, Karunaratne N, Exintaris B, et al. The impact of generative AI on health professional education: a systematic review in the context of student learning. Med Educ 2025 Dec;59(12):1280-1289. [doi: 10.1111/medu.15746] [Medline: 40533396]
52. Masters K. Artificial intelligence in medical education. Med Teach 2019 Sep;41(9):976-980. [doi: 10.1080/0142159X.2019.1595557] [Medline: 31007106]
53. Izquierdo-Condoy JS, Arias-Intriago M, Tello-De-la-Torre A, Busch F, Ortiz-Prado E. Generative artificial intelligence in medical education: enhancing critical thinking or undermining cognitive autonomy? J Med Internet Res 2025 Nov 3;27:e76340. [doi: 10.2196/76340] [Medline: 41183320]

XSL•FO
RenderX

54. Akinci D'Antonoli T, Stanzione A, Bluethgen C, et al. Large language models in radiology: fundamentals, applications, ethical considerations, risks, and future directions. Diagn Interv Radiol 2024 Mar 6;30(2):80-90. [doi: 10.4274/dir.2023.232417] [Medline: 37789676]
55. Keshavarz P, Bagherieh S, Nabipoorashrafi SA, et al. ChatGPT in radiology: a systematic review of performance, pitfalls, and future perspectives. Diagn Interv Imaging 2024;105(7-8):251-265. [doi: 10.1016/j.diii.2024.04.003] [Medline: 38679540]
56. Harigai A, Toyama Y, Nagano M, et al. Response accuracy of GPT-4 across languages: insights from an expert-level diagnostic radiology examination in Japan. Jpn J Radiol 2025 Feb;43(2):319-329. [doi: 10.1007/s11604-024-01673-6] [Medline: 39466356]
57. Song ES, Kim GH, Lee SP. Evaluation of GPT-4o and Gemini Advanced on the Korean National Dental Licensing examination: accuracy, consistency, and question generation. J Dent Sci 2026 Jan;21(1):96-102. [doi: 10.1016/j.jds.2025.07.020] [Medline: 41585183]
58. Bereuter JP, Geissler ME, Klimova A, et al. Benchmarking vision capabilities of large language models in surgical examination questions. J Surg Educ 2025 Apr;82(4):103442. [doi: 10.1016/j.jsurg.2025.103442] [Medline: 39923296]
59. Wieling M, Rawee J, van Noord G. Reproducibility in computational linguistics: are we willing to share? Comput Linguist 2018 Dec;44(4):641-649. [doi: 10.1162/coli_a_00330]
60. Chen L, Zaharia M, Zou J. How Is ChatGPT's behavior changing over time? Harvard Data Sci Rev 2024;6(2). [doi: 10.1162/99608f92.5317da47]

Abbreviations

BERT: bidirectional encoder representations from transformers
BLEU: bilingual evaluation understudy
EM: exact match
EMR: electronic medical record
GUI: graphical user interface
IRB: Institutional Review Board
LLM: large language model
NLP: natural language processing
OCR: optical character recognition
ROUGE-L: recall-oriented understudy for gisting evaluation-longest common subsequence
SBERT: sentence bidirectional encoder representations from transformers
SNUDH: Seoul National University Dental Hospital
STROBE: Strengthening the Reporting of Observational Studies in Epidemiology
USMLE: United States Medical Licensing Examination

Edited by P Kanzow; submitted 25.Jan.2026; peer-reviewed by M Pang, P Dhawan, PCA Watanabe; revised version received 14.Jul.2026; accepted 17.Jul.2026; published 04.Sep.2026.

Please cite as:
Cho AY, Eo SH, Jeon MJ, Kim JH, Ihm J, Seo DG
Multimodal Large Language Models for Dental Chart Image Interpretation: Cross-Sectional Benchmarking Study With Students and Clinicians
JMIR Med Educ 2026;12:e91809
URL: https://mededu.jmir.org/2026/1/e91809
doi:10.2196/91809

© Ah-Young Cho, Soo-Heang Eo, Mi-Jeong Jeon, Jae-Hoon Kim, Jungjoon Ihm, Deog-Gyu Seo. Originally published in JMIR Medical Education (https://mededu.jmir.org), 4.Sep.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Education, is properly cited. The complete bibliographic information, a link to the original publication on https://mededu.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Role-Differentiated AI Competencies and Curriculum Implications for Health Professions Education: Qualitative Study

Yuyi Park[1*], PhD; Yejin Han[2*], PhD; Hyeongjo Kim[3], MA; Solmoe Ahn[4], MS; Frederick William Kron[5], MD; Jihyun Lee[1], PhD

[1]Department of Dental Education, School of Dentistry and Dental Research Institute, Seoul National University, 103 Daehak-ro, Jongno-gu, Seoul, Republic of Korea

[2]Department of Medical Education and Humanities, College of Medicine, Yeungnam University, Daegu, Republic of Korea

[3]Department of Curriculum and Instruction, College of Education, University of Illinois, Urbana-Champaign, IL, United States

[4]Dental Research Institute, Seoul National University, Seoul, Republic of Korea

[5]Department of Family and Community Medicine, Macon and Joan Brock Virginia Health Sciences at Old Dominion University, Norfolk, VA, United States

[*]these authors contributed equally

Corresponding Author:
Jihyun Lee, PhD
Department of Dental Education, School of Dentistry and Dental Research Institute, Seoul National University, 103 Daehak-ro, Jongno-gu, Seoul, Republic of Korea

Abstract

Background: As AI fundamentally transforms the healthcare landscape, medical education curricula have struggled to keep pace with these technological shifts. While current research has established a foundation for general AI literacy, less attention has been given to the role-specific competencies required for the diverse functions that healthcare professionals perform in an AI-integrated environment. Although existing tiered and domain-based competency models have clarified general AI competency requirements, they offer limited guidance on how such competencies should be differentiated according to the roles healthcare professionals perform in practice.

Objective: This study aimed to explore how AI-driven changes in healthcare shape competency requirements across professional roles and to develop differentiated competency frameworks and curriculum implications for 3 distinct roles: users, developers, and leaders.

Methods: Using a qualitative research design, we conducted in-depth interviews with 13 subject matter experts across academic medicine and dentistry, clinical practice, and the healthcare AI industry, who were recruited through 4 independent channels. Data were analyzed using reflexive thematic analysis with inductive coding organized around 3 research questions and published competency frameworks serving as sensitizing concepts. Trustworthiness was supported through investigator triangulation, member checking, an audit trail, and attention to researcher reflexivity and positionality.

Results: We identified a role-differentiated competency framework with a layered structure as follows: (1) Users require machine learning and data literacy to understand where algorithms fail and scrutinize data origins and biases; the ability to select best-fit AI solutions and set appropriate thresholds for human-machine delegation; and a high degree of AI-related professionalism to work responsibly with AI by recognizing its limits, critically appraising outputs, maintaining clinical accountability, preserving patient-centered care, and updating knowledge continuously; (2) Developers add operational competencies, including validating systems across technical performance, clinical relevance, and regulatory compliance; bridging the language gap between clinical needs and technical refinements; and fostering a challenging, problem-solving mindset to address real-world clinical bottlenecks; and (3) Leaders focus on system-level competencies, including macro-level strategic planning, governing medical AI across its full lifecycle—selection, validation, deployment, and monitoring—through policy “roads” and governance frameworks, and optimizing systemic resource utilization while orchestrating trust-based collaboration that safeguards care quality. Rather than a strictly cumulative hierarchy, the roles represent analytic distinctions with shared foundational competencies. This framework translates these competencies into corresponding curriculum implications across the health professions education continuum.

Conclusions: This study identifies role-differentiated AI competencies and proposes a corresponding curriculum scaffold for healthcare education. As a preliminary, hypothesis-generating starting point requiring multistakeholder validation, the proposed framework can help prepare the future healthcare workforce to utilize AI responsibly and to govern and lead the next generation of digital health innovation.

(JMIR Med Educ 2026;12:e97608) doi:10.2196/97608

XSL•FO
RenderX

KEYWORDS

AI; artificial intelligence; competency-based education; curriculum; healthcare professionals; medical education; professional competence; qualitative research

Introduction

The rapid rise of AI is revolutionizing healthcare, fundamentally transforming traditional paradigms from automating routine administrative tasks to imaging-based diagnostic support, data-driven treatment personalization, and procedural assistance [1]. These advancements shift not only technological paradigms but also the fundamental roles and competencies expected of healthcare professionals. Medical professionals face the challenge of understanding AI and applying it responsibly in clinical contexts [2]. This creates an urgent need to restructure medical education to equip trainees with AI-related competencies.

This urgency contrasts with the slow pace of educational change [3,4]. Although AI is redefining medical competency requirements, many medical education institutions continue to adhere to traditional curricula. Several factors contribute to this lag. First, the systematic complexity and lengthy medical education cycles create a barrier for integrating new technology quickly. A second barrier is limited educator expertise in AI, which constrains the ability to translate it into teachable and assessable curricular content. As a result, medical education can trail healthcare innovation, limiting graduates' readiness to evaluate and use AI responsibly.

To bridge this gap, clearly defining AI competencies provides shared, assessable outcomes that can guide curriculum design and scalable faculty development, enabling more systematic integration of AI beyond reliance on individual faculty expertise. Accordingly, recent studies have sought to define AI competencies in medical education. Garvey et al [5] and Lee et al [6] established core knowledge and ethical domains, while Moldt et al [7] and Russell et al [2] emphasized interdisciplinary assessment and continuous learning. Additionally, Schubert et al [8] proposed a tiered model ranging from basic to expert proficiency.

More recent efforts have broadened these discussions across general and health-specific contexts. The United Nations Educational, Scientific and Cultural Organization (UNESCO) developed separate AI competency frameworks for students and teachers, spanning domains such as human-centered values, ethics, AI foundations, applications, and pedagogy [9,10]. In medical education, the International Advisory Committee for Artificial Intelligence (IACAI) proposed recommendations for AI integration across individual, institutional, and system levels [11]. The Association of American Medical Colleges (AAMC) has articulated principles for responsible AI use [12] and is developing competencies across the continuum from undergraduate medical education (UME) through graduate medical education (GME) to continuing professional development (CPD) [13]. Building on these UNESCO and IACAI frameworks, Khamis et al [14] proposed a faculty development framework based on the Accreditation Council for Graduate Medical Education (ACGME) Milestones, with progression from basic AI literacy to leadership.

Collectively, these frameworks provide important foundations for AI competency development in health professions education. However, they generally organize competencies by educational stakeholder, level of expertise, training phase, or institutional implementation level, rather than by the functional responsibilities healthcare professionals assume in AI-enabled practice. Therefore, this study builds on these foundations by examining AI competencies across 3 functional roles—users, developers, and leaders—and translating them into curriculum implications across the health professions education continuum, thereby refining and extending existing tiered and domain-based models [8,15]. Although few graduates currently enter AI development or policy roles, demand for such professionals is likely to grow, necessitating a shift from generic literacy to role-aligned pathways.

This study aimed to explore the cascading impact of AI on healthcare and how these changes influence competency requirements across roles—from users to developers and leaders—ultimately informing the redesign of medical education curricula. The study is guided by three research questions: (1) How have AI advancements transformed healthcare practices? (2) What specific AI competencies are required for healthcare professionals across users, developers, and leaders? (3) What curriculum framework is needed to systematically cultivate these role-specific AI competencies?

Methods

Study Design

This study used a qualitative research methodology for an in-depth exploration of the complex perspectives of subject matter experts (SMEs) on AI-driven changes in healthcare and medical education, a rapidly evolving field where qualitative inquiry is essential to obtain contextually rich and at times competing expert views. The study is exploratory and hypothesis-generating in nature and is aimed at characterizing how experts conceptualize role-differentiated AI competencies in healthcare and medical education. We adopted reflexive thematic analysis as articulated by Braun and Clarke [16], a theoretically flexible interpretive approach in which themes are actively generated by the researcher through recursive engagement with the data. Reporting follows the Standards for Reporting Qualitative Research (SRQR) [17], and trustworthiness is addressed through the criteria proposed by Lincoln and Guba [18].

Ethical Considerations

The study was approved by the Institutional Review Board of Seoul National University School of Dentistry (S-D20220028) and conducted in accordance with the Declaration of Helsinki. All participants provided written informed consent to participate in the study, including consent for audio recording and the use

XSL•FO
RenderX

of anonymized excerpts in publications. Each participant received an honorarium of KRW 300,000 (approximately US $220) as compensation for their time and expertise.

Participants

We operationally defined an SME as an individual meeting at least two of the following criteria: (1) a faculty position in a medical or dental school with documented research output on AI in healthcare or healthcare education; (2) active clinical practice integrating AI-based tools, evidenced by clinical leadership roles or published clinical work; (3) substantive operational involvement in an AI-related healthcare enterprise; and (4) recognized public visibility as a domain authority, evidenced by invited expert commentary in major national media.

Candidates were identified across the study period through four independent channels: (1) recommendations from 3 professional societies (the Korean Society of Artificial Intelligence in Medicine, the Korean Society of Medical Education, and the Korean Dental Education Association); (2) systematic screening of academic profiles at major medical and dental schools for AI-relevant research and curricular involvement; (3) identification through national daily news media of individuals featured as authoritative commentators on AI in healthcare; and (4) snowball recruitment, used principally to reach industry SMEs less visible through academic and media channels. We observed substantial convergence across these channels as society recommendations, academic searches, and media identifications repeatedly surfaced overlapping individuals, indicating recognized authority across independent professional and public spheres, which reduced the risk of bias from any single channel.

All approached experts agreed to participate, yielding 13 SMEs in total (Table 1). Several SMEs held dual or multiple roles (eg, academic faculty also engaged in clinical practice or in operating an AI-related healthcare enterprise), which we deliberately viewed as desirable during purposive selection because such individuals could provide broader perspectives to enrich the research questions.

XSL•FO
RenderX

Table . Professional backgrounds of SMEs[a].

SME number	Professional role	Area of expertise	Degree	AI expertise				Interview phase[b]
				R[c]	C[d]	T[e]	D[f]	
1	Professor of medicine	Emergency medicine; AI clinical decision support	MD[g], PhD[h]	✓	✓	✓	✓	Early[i]
2	Professor of medicine	Biomedical engineering; AI medical imaging	PhD	✓	—[j]	✓	✓	Early
3	Professor of medicine	Oncology/genomics; AI-based precision medicine	MD, PhD	✓	✓	✓	✓	Early
4	Professor of medicine	Medical AI; AI diagnostic imaging	PhD	✓	—	✓	✓	Early
5	Professor of medicine	Radiation oncology; AI clinical decision support	MD, PhD	✓	✓	✓	—	Mid[k]
6	Professor of dentistry	Medical informatics; medical ontology and machine learning	PhD	✓	—	✓	—	Late[l]
7	Professor of dentistry	Prosthodontics; AI-based dental CAD[m] and workflow	DDS[n], PhD	✓	✓	✓	✓	Early
8	Professor of dentistry	Prosthodontics; AI-assisted dental CAD and workflow	DDS, PhD	✓	✓	✓	✓	Mid
9	Dental practitioner	Prosthodontics/orthodontics; AI orthodontic diagnosis	DDS, PhD	✓	✓	—	✓	Late
10	Dental practitioner	Orthodontics; AI cephalometric analysis	DDS, PhD	✓	✓	—	✓	Mid
11	Healthcare AI company executive	Medical AI industry; AI-based cancer diagnostics	MD	✓	✓	—	✓	Late
12	Healthcare AI company executive	Dental AI industry; AI-based dental CAD	PhD	✓	—	—	✓	Mid
13	Healthcare AI company executive	Dental AI industry; AI medical/dental imaging	PhD	✓	—	—	✓	Early

[a]SME: subject matter expert.

[b]Interview phase reflects the initial interview.

XSL•FO
RenderX

[c]R: research indicates documented AI-related research output.

[d]C: clinical use indicates active clinical practice integrating AI-based tools.

[e]T: teaching indicates a faculty teaching role.

[f]D: development indicates substantive involvement in developing AI-based healthcare solutions or enterprises.

[g]MD: Doctor of Medicine.

[h]PhD: Doctor of Philosophy.

[i]Early: 2022 - 2023.

[j]Not applicable.

[k]Mid: 2024.

[l]Late: 2025.

[m]CAD: computer-aided design.

[n]DDS: Doctor of Dental Surgery.

Data Collection

A semistructured interview protocol was developed by 3 team members (YP, YH, and JL) and refined through written feedback from the remaining team members. The interview questions were organized around the 3 research questions (anticipated changes in healthcare, required AI-related competencies, and implications for future curricula) and deliberately framed broadly rather than anchored in any pre-existing competency taxonomy, allowing diverse expert perspectives to surface without being constrained by extant frameworks. To support the logical translation from competencies to curricular implications, participants were reminded of the competencies they had identified and asked to specify the curricular content needed to develop those competencies. No formal pilot was conducted; the first interview served as a calibration check and required no substantive revision to the protocol.

Of the 13 interviews conducted between October 2022 and October 2025, 12 were held in person and 1 was conducted via Zoom (Zoom Communications, Inc), with each interview lasting between 45 and 90 minutes. The interview phase in Table 1 reflects the initial interview. All interviews were audio-recorded with consent and transcribed verbatim.

Sample size was determined through iterative monitoring of analytic stabilization—sometimes termed data saturation, though we use this term cautiously given the constructivist orientation of reflexive thematic analysis [16]—rather than a prespecified target. By approximately the ninth interview, analytic discussions indicated that subsequent transcripts were generating few new codes and that emerging themes were being elaborated rather than expanded, suggesting that the dataset was approaching information sufficiency for the research questions. Four additional interviews were subsequently conducted to both verify this pattern and purposively probe emergent tensions that surfaced during data collection. These interviews yielded contextual elaborations rather than substantively new themes. By the 13th interview, the team judged the framework structure to have stabilized, with all 3 role categories sufficiently represented in the data. Sample adequacy was further evaluated against the information power framework [19], which judges qualitative sample sufficiency along 5 dimensions. Our study showed favorable indicators across four of these dimensions: (1) a focused research aim, (2) high participant specificity (each SME had expertise in a field directly relevant to the study), (3) substantive interview dialogue, and (4) a case-focused analytic strategy.

Data Analysis

Interview data were analyzed using reflexive thematic analysis [16,20]. Rather than developing a deductive codebook prior to analysis, we organized inductive coding around the 3 research questions, with published competency frameworks identified during the literature review [2,6-8] serving as sensitizing concepts that provided theoretical guidance without imposing a predetermined analytic structure [21].

Data analysis began in October 2022 and proceeded iteratively alongside data collection. Following the 6 recursive phases of reflexive thematic analysis (data familiarization, generation of initial codes from the data, theme development, theme review, theme definition, and reporting) [20], 4 researchers (YP, YH, HK, and JL) independently coded the full dataset and convened iteratively to compare codes, discuss interpretive divergences, and refine candidate themes. During the theme review and definition phases, we specifically examined whether candidate themes reflected dentistry-specific examples or broader cross-specialty functions. Dentistry-specific examples, such as computer-aided design/computer-aided manufacturing (CAD/CAM)–supported treatment planning and procedure-specific digital design, were treated as illustrative rather than as sufficient grounds for defining general competencies. Themes were retained at the framework level when they reflected cross-cutting functions across dental, medical, engineering, and industry perspectives. Discrepancies that could not be resolved through team discussion were referred to 2 additional team members (SA and FWK) for adjudication. For cross-linguistic adjudication, JL translated relevant excerpts and contextual details into English. Quotations presented here were translated by JL and verified by FWK.

To ensure the accuracy of the interpretation, triangulations were carried out through member checking with all 13 SMEs in 2025. Each SME reviewed a summary of findings from their interview, either confirming that it reflected their views or providing updates or clarifications. Seven SMEs submitted detailed written feedback, which was incorporated into the analysis where relevant. Building on the competency themes refined through this process, we derived the curriculum implications by mapping participant-proposed educational content for each competency to the corresponding theme and, through iterative team

XSL•FO
RenderX

discussion, organizing it into role-specific modules sequenced from foundational concepts to applied content.

The 6-member research team comprised researchers with multidisciplinary backgrounds spanning dental and medical education, educational technology, learning analytics, nursing practice in internal medicine settings, family medicine, and AI-enabled medical education technology. Four members based in Korean health professions education conducted coding (YP, YH, HK, and JL). Two members served as discrepancy resolvers (SA and FWK), with FWK additionally contributing a non-Korean clinical and technology-development perspective. We recognize that this composition concentrated our analytic gaze and that our shared identity as health professions education scholars predisposed us to read the data through a competency and curriculum lens. We treated these orientations as interpretive resources rather than biases to be eliminated [16], discussing them explicitly during analytic meetings to question candidate themes.

Trustworthiness

Trustworthiness was addressed against the 4 criteria of Lincoln and Guba [18] as follows. *Credibility* was supported by prolonged engagement with the field across the data collection window, investigator triangulation through independent coding and discrepancy resolution, member checking with all 13 SMEs, and cross-linguistic adjudication of disputed interpretations. *Transferability* was supported through detailed description of recruitment context, participant characteristics, and analytic decisions; we explicitly note that the framework was developed within a Korean health professions education context with substantial dental representation, which may shape its applicability across healthcare systems, specialties, and regulatory environments. *Dependability* was supported by an audit trail comprising the interview protocol, transcripts, analytic memos, and codes developed iteratively during analysis. *Confirmability* was supported by the explicit positionality statement above, the use of verbatim participant quotations to ground interpretive claims, and the triangulation procedures noted above.

Results

Overview

We have presented the themes, subthemes, and competencies identified in this study. Moreover, we have provided selected quotes, and further quotes are available in Multimedia Appendix 1.

Theme I: AI-Driven Healthcare Transformation

Overview

Thirteen SMEs described how AI integration is reshaping healthcare. Their perspectives anticipated four directions of change: (1) greater patient empowerment and more horizontal doctor-patient relationships; (2) redefinition and reconfiguration of medical expertise; (3) transformation and redesign of clinical workflows through human-machine collaboration; and (4) a possible shift toward more consistent "average" care quality through standardization and reduced operator variability, supporting routine clinical decisions and processes without displacing the judgment of skilled clinicians.

Subtheme I-1: Patient Empowerment and Horizontal Doctor-Patient Relationships

Experts suggested that patients will arrive with AI-generated interpretations and recommendations, reshaping the traditional information asymmetry. Clinicians may need to negotiate care decisions with patients who treat AI as an additional expert voice while critically assessing the clinical validity of AI outputs.

> *Patients will become more empowered… "The AI says something's off, so why do you think it's fine?"… patients now have knowledge that makes them almost equal to clinicians.* [SME #1]

> *Once patients view AI as an expert service, it's like bringing another doctor along… "This is a second opinion from a well-known AI."* [SME #3]

Subtheme I-2: Redefinition and Reconfiguration of Medical Expertise

SMEs expected AI to blur and reorganize professional boundaries by shifting tasks across specialties and to AI systems (SMEs 1, 3, 5, and 8). Some also anticipated a deeper reconfiguration of expertise, shifting from fixed territories to data-driven, high-level integration (SMEs 3 and 5).

> *I think the boundaries of healthcare professionals' duties will need to be redefined—literally redefined. The tasks that used to be considered their exclusive responsibilities may have to be redefined in this new context.* [SME #8]

> *With AI, the boundaries… start to blur. Fields that used to fight over territory will naturally merge or shift as data and algorithms provide objective evidence… AI will accelerate this shift, lowering the walls between specialties and redefining what expertise means.* [SME #5]

> *I think physicians' roles will actually become more advanced. It's like when you no longer have to focus on keeping your lane on the highway—you can … focus on other, more important things. Similarly, doctors will rely on AI to handle the detailed tasks and instead concentrate on broader, higher-level judgments.* [SME #1]

Subtheme I-3: Transformation of Clinical Workflows

Experts expected that AI would fundamentally reshape medical workflows by reallocating tasks between humans and machines. They emphasized that this transformation would depend on the effective integration of human-machine collaboration models.

> *If we think of all the tasks… as 100%, about 80% could eventually be handled by AI… Humans should then focus on the remaining 20%… Ultimately, how well we build the human-machine collaboration model will shape everything… AI will fundamentally transform medical workflows.* [SME #3]

XSL•FO
RenderX

Another expert illustrated this potential operational efficiency, noting that AI could streamline the logistics of consultations and referrals. By reducing unnecessary steps in the referral process, AI might allow workflows to proceed more quickly, potentially enabling clinicians in some settings to allocate more time to direct patient interaction.

> *With AI, instead of asking five people, you might only need to ask two or three. … You can quickly figure out who to refer to, so the whole referral process becomes smoother. What used to take a full round of visits or a week to follow up now happens much faster—the workflow just flows better.* [SME #5]

Subtheme I-4: Improvement in Average Healthcare Quality

Finally, experts expected that AI could improve the consistency of routine care by reducing operator variability, even if it does not replace the craftsmanship of exceptional clinicians.

> *AI can't match the craftsmanship of what we create by hand—one stitch at a time. It will never reach the high end.* [SME #9]

> *It's not that the maximum level of care will go up, but the average will… local clinics… could reach the average quality level of university hospitals.* [SME #8]

> *Before AI, the quality used to vary a lot… But with AI, those operator differences have gotten much smaller.* [SME #5]

Taken together, the findings of research question 1 indicate that AI-driven change operates across relational, professional, workflow, and care quality dimensions of healthcare. These dimensions provide the contextual basis for the role-differentiated competencies described below.

Theme II: Role-Differentiated AI Competencies and Curriculum Design

Overview

Experts agreed that the competencies and educational needs of medical students should differ depending on their future roles: those who will primarily use AI in clinical practice, those who will participate in AI development, and the small group of future leaders who will drive innovation and governance in the field of medical AI.

> *Out of ninety students, maybe two or three will be truly inspired by an AI course and become the ones who lead this country—and even compete globally—in medical AI. For the majority, though, the goal should be to understand how to position themselves within the AI context and use it effectively in practice. Such a dual structure is essential, especially in a cutting-edge field like this.* [SME #2]

To clarify the boundaries among the competencies, the framework differentiates roles according to their primary function and level of responsibility in AI-enabled healthcare. We have defined each role operationally. AI users apply AI tools in routine clinical practice—interpreting AI-generated outputs, integrating them with clinical judgment, and remaining professionally accountable for patient care. AI developers design, evaluate, or contribute to building AI solutions—translating clinical needs into technical specifications or working within clinical-AI development teams. AI leaders govern, oversee, or strategically direct AI integration at organizational or systemic levels, for example, department heads, chief medical informatics officers, and healthcare administrators overseeing institutional AI implementation.

These roles describe positions in current or future clinical practice, and the curriculum continuum (UME, GME, and CPD) is designed to prepare learners for them. They are analytic distinctions rather than mutually exclusive paths. Many SMEs held dual roles, and the framework is best understood as layered: each role carries a distinct emphasis, while foundational competencies (eg, critical evaluation of AI outputs and accountability for clinical judgment) are shared across roles. The distribution across roles is also asymmetric: most healthcare professionals will serve as users, with smaller numbers entering developer or leader roles.

Subtheme II-1: AI User Competencies and Curriculum Design

Competency U-1: AI/Machine Learning Literacy

Most SMEs endorsed AI and machine learning (ML) literacy as a baseline competency for future users—a practical understanding of learning-based AI mechanisms sufficient to understand what AI does, where it can fail, and why errors occur.

> *Algorithms are like engines—you don't need to develop them, but you should know enough to understand when they might fail. With a car, I can feel it—it shakes or slows down—but AI doesn't feel like that. You just input software and get results. That's why we need more education on the underlying mechanisms.* [SME #1]

However, some argued that detailed ML instruction is not essential for all students and should mainly serve to inspire potential future developers.

> *Doctors don't need to learn machine learning itself… if a few get inspired… that's already a success.* [SME #3]

Competency U-2: Data Literacy

Data literacy was repeatedly framed as indispensable, as AI outputs reflect underlying data biases. SMEs emphasized the need to scrutinize data origins, labeling processes, and potential institutional biases.

> *Data literacy is essential… datasets vary widely… results can differ because of how data are collected and interpreted.* [SME #1]

> *In one large chest X-ray dataset, only about a thousand of a hundred thousand images had actually been reviewed by physicians, and even those were flawed. It shows how important it is to verify and*

XSL•FO
RenderX

> *understand what kind of data are being used.* [SME #2]

Across both AI/ML literacy and data literacy, SMEs stressed that users must be able to interpret and explain outputs meaningfully to patients, rather than merely operate tools, underscoring this as a critical mandate for medical education (SMEs 2, 3, and 11).

> *Using an AI solution isn't difficult—it's as easy as using Microsoft Word. The real challenge is understanding what the output implies and how to apply it meaningfully in practice. That's what education should address.* [SME #3]

Competency U-3: AI Solution Selection and Utilization

Building on foundational AI/ML and data literacy, SMEs emphasized the ability to select and manage AI tool use in practice: understanding AI solutions, comparing options, anticipating bias, and setting thresholds for when to rely on AI versus human expertise.

> *AI solutions will continue to diversify, and each one will have its own type of data bias. Physicians will face situations where they must decide which solution best fits a given patient.* [SME #1]

Consequently, experts urged that education must teach how to set these critical thresholds—determining precisely when to rely on automation versus human expertise (SMEs 3 and 4).

> *In radiology, for instance, general practitioners can now interpret many imaging results directly using AI software, referring only ambiguous or difficult cases to radiologists. This means that doctors can internalize some of the radiologist's expertise while using AI for efficiency. Education should teach how to set such thresholds—when to rely on AI and when to rely on human expertise.* [SME #3]

Some experts questioned the necessity of explicit AI training, arguing that solid domain knowledge is the true prerequisite for validating AI outputs.

> *In prosthodontics… if those basics are neglected, they won't be able to tell whether an AI-generated design aligns with fundamental dental principles.* [SME #7]

However, others countered this by stating that while foundational mastery is vital, its very definition evolves.

> *After giving a lecture on digital technology, students often ask, "Why are we still using wax and torches? Shouldn't we be doing digital practice?" I tell them, you must understand the fundamentals first—but I also think the fundamentals themselves are changing… Education needs to reflect those updated fundamentals.* [SME #8]

Competency U-4: AI Professionalism in Healthcare

SMEs identified AI professionalism as a core competency, which can be organized into 4 aspects.

First, experts noted that a foundational element of AI professionalism is an awareness of AI's inherent limitations, that is, its outputs are bounded by prior data and cannot adapt on their own to changing clinical conditions, and that education must address this explicitly.

> *AI is just a tool—and that's something every doctor needs to remember… When medical guidelines or systems change, it might start doing things that make no sense… it will always be limited by what it has learned from past data. Students must be clearly taught about these limits.* [SME #4]

> *When AI makes an unexpected decision, it's not really an error—it's just doing what it was taught. Understanding that helps physicians communicate better, make fewer mistakes, and keep patients safer.* [SME #4]

Second, experts emphasized that physicians must uphold ethical accountability and remain the final arbiters of clinical judgment, and education should treat this as a top priority.

> *There are already cases of AI misuse—overreliance, really… Education must repeatedly stress that, in the end, any mistake or harm remains the physician's responsibility. That's what it means to hold a medical license.* [SME #4]

> *AI systems can assist in diagnosis, but they don't make the diagnosis… If we delegate that responsibility to AI, we risk giving away our very professionalism.* [SME #10]

> *It's about helping students understand the gravity of working with AI… They need to know where their responsibility begins and ends… in practice.* [SME #1]

Third, experts highlighted the importance of preserving the human dimensions of care—understanding the patient's context and showing empathy—as distinctly physician responsibilities.

> *Seeing the patient, understanding their context, and showing empathy—those are things only a physician can do. That's what needs to be taught.* [SME #11]

Fourth, SMEs further argued that this professionalism must be sustained through lifelong learning. They noted that because AI tools, evidence, and regulations evolve quickly, maintaining up-to-date understanding—and developing the ability to critically interpret new research—should be treated as part of professional integrity.

> *Being good at using tools right away doesn't make you a good doctor. What really matters is continuous learning—it's a core part of a physician's role and responsibility.* [SME #1]

Synthesizing these expert perspectives, AI professionalism may be defined as the physician's ability to work ethically and responsibly with AI systems by recognizing AI's inherent limitations, critically evaluating AI outputs against the clinical context while accepting accountability as the final arbiter of clinical decisions, integrating AI use with empathic patient-centered care as a distinctly human responsibility, and continually updating knowledge amid the rapid evolution of AI evidence and regulations.

XSL•FO
RenderX

Table 2 outlines the curriculum implications aligned with AI user competencies, aligning competencies (U-1 to U-4) with educational modules. The progression moves from foundational literacy to applied selection and professionalism, guiding learners from basic understanding to responsible use.

Table . Curriculum implications aligned with AI user competencies.

Required competencies	Educational content

U-1. AI/ML[a] literacy: Understanding AI mechanisms to identify where algorithms fail and why errors occur	• Module 1. Introduction
	  • History of AI
	  • Changes in healthcare driven by AI adoption
	  • AI ecosystem: social and infrastructural perspectives

U-2. Data literacy: Scrutinizing data origins and biases to explain AI outputs meaningfully to patients	• Module 2. AI/ML and data literacy
	  • AI/ML fundamentals
	  • Data literacy

U-3. AI solution selection and utilization: Selecting best-fit AI tools and setting thresholds for human-machine delegation	• Module 3. AI solution understanding, selection, and utilization
	  • Understanding and selection of AI solutions
	  • Application of AI in healthcare
	  • Human-AI collaboration

U-4. AI professionalism in healthcare: Maintaining accountability as final arbiters of clinical judgment through lifelong learning	• Module 4. AI-related professionalism
	  • Limits of AI
	  • Roles, accountability, and ethics of healthcare professionals
	  • Lifelong learning

[a]ML: machine learning.

Subtheme II-2: AI Developer Competencies and Curriculum Design

Competency D-1: Advanced User Foundations

SMEs described developer competency as building on user competencies with advanced ML literacy and professional awareness. Developers must understand how data quality; rigorous annotation, including expert-defined labels and richer clinical metadata, clinically defined reference standards such as lesions or diagnoses, and decision thresholds for clinical risk; and interpretability underpin trustworthy medical AI (SME 13). This includes understanding how model outputs can be interpreted in relation to the training data, how different data features shape model judgments, and what additional clinical data may be needed to improve performance. Developers must also recognize that their technical decisions directly impact clinical trust, patient safety, and social accountability (SMEs 1 and 11).

> *In the end, developing an algorithm is really all about the data… the annotation process—how you label and define the correct answers—needs to be done well… With explainable AI (XAI), you can look at the output and think, "Okay, this model made this judgment based on that data—so maybe if we include more of this kind of data, the performance will improve." That kind of understanding is really valuable.* [SME #13]

Competency D-2: Effectiveness Evaluation

Experts emphasized that developers must evaluate AI systems across 3 critical dimensions: technical performance, clinical relevance, and regulatory compliance. What distinguishes AI evaluation in healthcare from evaluation in other engineering domains is that each dimension carries healthcare-specific implications. Regarding technical performance, experts noted that developers should understand how clinical trial results form the "credible evidence that can later be used for reimbursement and policy decisions" (SME 11), and education must go beyond accuracy metrics to cover validation processes. Crucially, developers must interpret performance thresholds in clinical contexts, adjusting sensitivity-specificity tradeoffs according to screening, diagnostic, or workflow priorities.

> *Developers shouldn't just read the model's performance table; they must be able to judge its clinical value in different medical settings. For instance, in breast cancer screening, sensitivity matters most—even at the cost of more false positives—while in chest X-rays, that trade-off could disrupt the workflow. Developer training should cover how to interpret such contextual differences and adjust sensitivity–specificity targeting.* [SME #11]

For clinical relevance, developers must learn to weigh medical necessity against technical feasibility and judge whether model performance translates into meaningful clinical value in specific medical settings.

> *Engineering often starts from… "Can we make it?" But medicine must also ask, "Should we use it?" ... Developers must learn to weigh clinical benefit against technical possibility.* [SME #1]

For regulatory compliance, the curriculum must bridge the gap between research models and commercial products.

XSL•FO
RenderX

> *Regulatory expectations are extremely high… There's a huge gap between research-grade models and commercially approved ones… Developer education should… train students to understand international regulatory systems.* [SME #11]

This requires developers to assess compliance within constantly evolving legal and institutional frameworks (SME 10).

> *Because these legal and institutional contexts evolve, developers must go beyond technical design and learn to assess whether their systems remain compliant within the current regulatory framework.* [SME #10]

Competency D-3: Field Demand Communication

Experts stressed that effective clinical-technical communication between clinicians and developers is essential for ensuring AI addresses genuine clinical needs. Developers must learn to understand contextual differences across institutions and translate clinical language into technical refinements, engaging in continuous dialogue throughout the development process. This communication challenge is particularly complex in healthcare because, even within the same institution, different specialties may define the same clinical condition differently.

> *Developers shouldn't look at hospital data only in statistical terms—they need to understand the real clinical problems behind it. Even within one hospital, pulmonologists and thoracic surgeons may define the same condition differently. Learning to communicate and reconcile these differences is key.* [SME #1]

> *Clinicians and developers speak completely different languages… Clinicians talk about clinical meaning, while developers talk about accuracy scores… education that bridges that gap…* [SME #10]

> *Over time, clinicians get a sense for when the AI will fail or succeed, and the product improves through that feedback… ongoing dialogue…* [SME #4]

Competency D-4: Challenging Mindset and Problem-Solving

Experts emphasized that developers must cultivate curiosity and perseverance to tackle real clinical problems through experimentation, reflection, and iterative improvement. Education should facilitate the resolution of real-world clinical bottlenecks.

> *When radiologists said, "We've been trying to solve this problem for years…" I took on the challenge… within a month, the performance improved dramatically. That kind of problem-solving experience… really motivates both clinicians and developers.* [SME #4]

> *A spirit of challenge is essential… It's about curiosity… The mindset that says, "What if we make that happen?"—that's what education should nurture.* [SME #9]

Table 3 outlines the curriculum implications aligned with AI developer competencies, aligning competencies (D-1 to D-4) with educational modules. The progression moves from advanced technical mastery to evaluation, communication, and innovation, fostering both engineering precision and clinical adaptability.

XSL•FO
RenderX

Table . Curriculum implications aligned with AI developer competencies.

Required competencies	Educational content

D-1. Advanced user foundations: Advanced foundational skills—ML[a]/data literacy, selection, and professionalism—within the developer role, with attention to clinically grounded data and model interpretation	• Module 1. Advanced user competency
	  • Advanced understanding of ML/data literacy and AI solution use
	  • Expanded perspectives on AI-related developer professionalism
	  • Clinical data quality, expert annotation, clinical reference standards, and model-output interpretation

D-2. Effectiveness evaluation: Validating systems across technical performance, clinical relevance, and regulatory compliance in healthcare settings	• Module 2. Evaluation of AI in healthcare
	  • Technical validation and performance evaluation of AI systems
	  • Clinical effectiveness assessment
	  • Policy and regulatory compliance review
	  • Clinical trial evidence, safety demonstration, and reimbursement/policy relevance

D-3. Field demand communication: Bridging the "language gap" to translate clinical needs into technical refinements, including specialty-specific clinical definitions	• Module 3. Field communication and feedback
	  • Identifying clinical needs and collecting feedback for AI solutions
	  • Enhancing collaboration through shared language and understanding
	  • Reconciling specialty-specific clinical definitions and workflow priorities

D-4. Challenging mindset and problem-solving: Applying curiosity and perseverance to solve real-world clinical bottlenecks	• Module 4. Innovation
	  • Developing a proactive and innovative mindset
	  • Problem-solving in complex and dynamic healthcare contexts

[a]ML: machine learning.

Subtheme II-3: AI Leader Competencies and Curriculum Design

Competency L-1: Systemic Insight and Strategic Planning

Experts emphasized that AI leadership starts with macro-level sensemaking: leaders must anticipate how AI reshapes workflows and infrastructures beyond mere automation.

> *An AI leader isn't just someone who holds a position—it's someone with the vision… to set the direction… [and] draw the big picture…* [SME #5]
>
> *I saw the massive wave of industrial transformation approaching, … The moment plaster models became digital scans, the entire pipeline—treatment, manufacturing, and business—was revolutionized… Students today must be educated to recognize and anticipate such paradigm shifts in the industry.* [SME #9]

Crucially, this vision must align with social readiness rather than just technical feasibility.

> *The key question isn't whether a technology can be realized, but whether society is ready… AI leadership education must address… social acceptance, ethical judgment, and systemic policy awareness.* [SME #8]
>
> *AI leadership requires the capacity to plan beyond technology—to design interconnected infrastructures, workforce systems, and policies that support these changes. Education for AI leaders must therefore cultivate systemic thinking, policy implementation skills, and integrated management of resources and human capital.* [SME #3]

Finally, experts stressed collaborative execution and operational efficiency as essential qualities for AI leaders. The ability to coordinate across clinical, technical, and administrative teams determines the success of integration in practice (SME 9).

Competency L-2: AI Governance and Policy Design

Unlike developers who focus on evaluating the performance of individual AI systems, leaders are expected to evaluate AI at the governance and policy level, integrating regulations, infrastructure, and accountability frameworks.

> *Drugs, for instance, go through Phases 1 to 4, but applying the same staged framework to AI doesn't really fit. We need a more suitable regulatory model for AI applications in medicine.* [SME #1]

Experts highlighted that leadership competency involves planning AI policy and infrastructure roadmaps that enable coordinated ecosystem growth. While developers optimize algorithms, leaders must design the "roads" that sustain them, integrating data systems, institutions, and national strategies.

> *It's like we keep trying to make better cars, but no one's building the roads…* [SME #6]

Experts further emphasized that governance must operate across the entire AI lifecycle, beginning with tool selection and procurement. Leaders must evaluate clinical feasibility, usability,

XSL•FO
RenderX

and workflow fit rather than technical capability alone, and—where regulatory agencies cannot adjudicate—judge an algorithm's soundness at the institutional level.

> *With 30 clinical departments each saying "ours matters," the hospital itself has to judge whether an algorithm is sound—much as the regulatory agency judges a device.* [SME #1]

They noted that technical capability alone does not guarantee real-world adoption and that procurement decisions must therefore weigh whether a system will actually be used in practice.

> *Watson started from engineering… but ended up not being used at all. You can build it, but whether it gets used is another matter.* [SME #1]

Beyond adoption, experts stressed that leaders must institutionalize internal validation, recognizing that vendor evidence often reflects controlled conditions rather than real clinical performance.

> *When you give it simulation cases, it looks helpful… but that's in vitro. It hasn' t yet been witnessed in vivo, in actual clinical application.* [SME #1]

At the deployment stage, they framed implementation not as installation but as an integration effort requiring coordination across clinical and infrastructure teams.

> *When applying AI in clinical environments, customizing the setup to each hospital's context is crucial… involving IT [information technology] teams, EMR [electronic medical record] administrators, and support departments… and designing multidisciplinary governance structures.* [SME #9]

Finally, experts underscored that governance extends into postdeployment monitoring, since AI is trained on past data and degrades as clinical guidelines and environments shift, requiring continuous oversight of bias, off-label use, and overreliance.

> *Each solution carries bias in different ways, and the clinician has to choose among them for each patient.* [SME #1]
>
> *The more important part is that it has to be continuously updated.* [SME #4]

These risks ultimately return to accountability: when a system drifts, errs, or is overrelied upon, governance must already have defined who bears responsibility—ethical, legal, and operational—across developers, planners, and users.

> *In medical systems, we need to clearly define who holds responsibility, risk, and transparency. What are the roles of developers, planners, and users? How should accountability be shared among them? These questions must guide national-level AI governance frameworks and should be explicitly addressed in AI leadership education.* [SME #6]

Competency L-3: Resource Management and Leadership

Experts emphasized that AI leaders must manage healthcare resources at a systems level rather than compromising patient-level care quality (SME 1), institutionalize organization-wide task allocation rules based on contextual complexity to build trust-based human-machine collaboration (SME 3), and support implementation through cross-role communication and workflow-adjustment mechanisms linking clinical teams with AI developers and maintainers (SME 8).

> *The goal is… reducing systemic waste… AI can function as a tool to optimize resource utilization… Future healthcare leaders should be trained… from this systems-oriented perspective.* [SME #1]
>
> *Building systems where AI supports, but doesn't replace, human expertise… humans manage the complex… machines take over structured, predictable tasks… That's how we build trust.* [SME #3]
>
> *Communication is the hardest part… it won't work if the users can't communicate effectively with those who build or maintain it… when schedules or workflows need to be adjusted… human dialogue and coordination remain irreplaceable.* [SME #8]

Table 4 outlines the curriculum implications aligned with AI leader competencies (L-1 to L-3), mapping each competency to educational modules. The progression expands from macro-level strategic planning to governance design and resource management, fostering capabilities for system-wide transformation and collaborative execution.

XSL•FO
RenderX

Table . Curriculum implications aligned with AI leader competencies.

Required competencies	Educational content

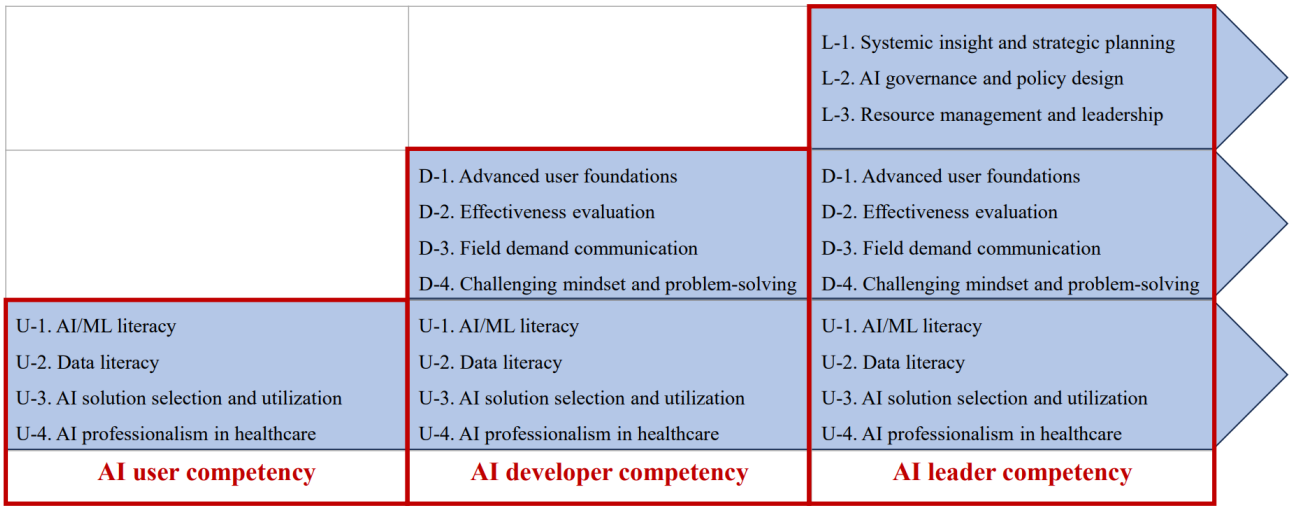
L-1. Systemic insight and strategic planning: Leading macro-level sense-making and vision-setting for AI-driven workflow and infrastructure shifts in healthcare	• Module 1. Systems thinking and strategic planning
	  • Reading macro trends and planning system-level changes
	  • Recognizing social context and applying systems approaches
	  • Enhancing collaboration and operational efficiency in hospital AI implementation

L-2. AI governance and policy design: Evaluating medical AI at the governance and policy level, designing AI policy "roads," and governing the full AI lifecycle	• Module 2. AI governance and policy design
	  • Regulation and evaluation of AI applications in medicine
	  • Policy and infrastructure roadmap planning
	  • Designing lifecycle governance (tool selection/procurement → validation → deployment → continuous monitoring/accountability)

L-3. Resource management and leadership: Optimizing systemic resource utilization and orchestrating trust-based collaboration while safeguarding care quality	• Module 3. Resource management and leadership
	  • Resource optimization and healthcare supply-chain management aligned with care quality and workflow coordination
	  • Human–machine collaboration and trust-based leadership
	  • Team communication and collaborative leadership

Figure 1 illustrates the role-differentiated AI competency framework across AI users, developers, and leaders. It visualizes the layered structure in which each role carries a distinct emphasis—foundational AI literacy and clinical judgment for users, operational and technical translation for developers, and system-level governance for leaders—while shared foundational competencies provide common ground across roles.

Figure 1. Role-differentiated AI competency framework across user, developer, and leader roles.



Discussion

Principal Findings

This study explored AI-driven changes in healthcare, identified role-differentiated AI competencies for healthcare professionals (AI users, developers, and leaders), and mapped these competencies to corresponding educational modules as curriculum implications. The framework is best understood as a layered structure with role-specific emphases and shared foundational competencies. To further contextualize our findings in relation to existing AI competency frameworks, Figure 2 presents an interpretive comparison rather than a formal validation or analytic procedure. This comparison clarifies how the proposed role-differentiated framework may extend and reorganize existing competency domains through a functional role lens.

XSL•FO
RenderX

Figure 2. Comparison of competencies identified in this study and the recent literature [2,6-8]. ML: machine learning. *Not reported.

Roles and competencies identified in our study		Schubert et al [8], 2025	Moldt et al [7], 2024	Lee et al [6], 2024	Russell et al [2], 2023
AI User	U-1. AI/ML literacy	—*	—	• Fundamental knowledge and skills in medical AI	• Basic knowledge of AI • AI-enhanced clinical encounters • Evidence-based evaluation of AI-based tools • Workflow analysis for AI-based tools
	U-2. Data literacy	—	• Data literacy • Data privacy	• Processing, analyzing, and evaluating medical data	
	U-3. AI solution selection and utilization	• Practical ability to use a clinical AI tool (basic tier) • Knowledge of workplace clinical AI tools (basic tier) • Critical appraisal of clinical AI utility and outputs (proficient tier)	• Basic understanding and sense of technology • Digital literacy • Application of software • Understanding of medical test results	• Medical AI application in clinical practice	
	U-4. AI professionalism in healthcare	• Critical appraisal of ethical implications of clinical AI integration (proficient tier)	• Morality and ethics • Opportunities and risks	• Ethics and legal aspects in the use of medical AI • Understanding digital health and changes driven by AI	• Social and ethical implications of AI • Practice-based learning and improvement regarding AI-based tools
AI Developer	D-1. Advanced user foundations	• Deep ML expertise integrated with clinical expertise to advance medical AI (expert tier)	• Programming skills (optional)	• Processing, analyzing, and evaluating medical data	—
	D-2. Effectiveness evaluation	• Skills to interpret and conduct clinical trials and to evaluate AI value, performance, bias, ethics, and health economics (proficient tier)	—	• Research and development of medical AI	—
	D-3. Field demand communication	—	—	—	—
	D-4. Challenging mindset and problem-solving	—	—	• Research and development of medical AI	—
AI Leader	L-1. Systemic insight and strategic planning	—	—	—	—
	L-2. AI governance and policy design	—	—	—	—
	L-3. Resource management and leadership	• Explain clinical challenges to ML scientists; collaborate across disciplines within AI capabilities and limits (expert tier)	—	—	• Workflow analysis for AI-based tools

Regarding AI-driven changes in healthcare, experts anticipated four interrelated directions: (1) greater patient empowerment and more horizontal doctor-patient relationships, (2) redefinition and reconfiguration of medical expertise, (3) transformation of clinical workflows through human-machine collaboration, and (4) improvement in average healthcare quality through reduced operator variability. These projections, however, should be understood less as settled futures than as conditional and contested propositions.

Patient empowerment may enhance participation in care, yet patient-facing AI can also generate fluent but inaccurate interpretations that patients may be poorly positioned to evaluate. A recent case study in clinical neuropsychology showed that patient-led use of a large language model misrepresented clinical data, caused distress, and increased clinical burden by requiring direct correction of AI-generated misinformation [22]. Similarly, physicians report both perceived benefits and concerns regarding AI in medical decision-making and patient-physician communication, underscoring that AI may be supportive but may also strain clinical relationships when trust, accountability, and AI literacy are insufficient [23]. The shift toward more consistent average care is likewise

XSL•FO
RenderX

double-edged: while AI may reduce unwarranted variation and raise the floor of care, it may also anchor clinical practice to a merely acceptable average, reinforce automation bias, and contribute to the erosion of independent expertise [24-26]. This risk is supported by a multicenter observational study showing that endoscopists' adenoma detection rate during non–AI-assisted colonoscopy declined after routine exposure to AI-assisted colonoscopy, providing real-world evidence of AI-associated deskilling in relation to a clinically meaningful outcome [25]. The redistribution of expertise and redesign of workflows likewise remain open questions rather than inherently beneficial developments: AI may augment clinical decision-making and support more patient-centered workflows [27], but it may also fragment professional identity [28], diffuse accountability [29], and shift tasks according to technical feasibility rather than clinical value unless implementation is guided by human-centered design, governance, and clearly defined clinical oversight [30].

Figure 2 shows that recent literature has largely prioritized user-level competencies for medical students and practicing health professionals [2,6-8]. Across these studies, the most consistent emphasis is on competencies U-3 and U-4, reflecting a broad consensus that clinicians must interpret and apply AI outputs responsibly, communicate their implications, and maintain professional and ethical accountability for clinical decision-making. Existing frameworks offer valuable multilevel and system-based guidance but are less often organized around healthcare professionals' functional responsibilities in AI-enabled practice [2,6-8,11,13]. From this role-based perspective, sustainable AI integration requires user competencies alongside specialized developer competencies (eg, validation design, translation of clinical needs, and safety and bias checks) and leader competencies (eg, governance design, accountability structures, resource allocation, and strategic decision-making).

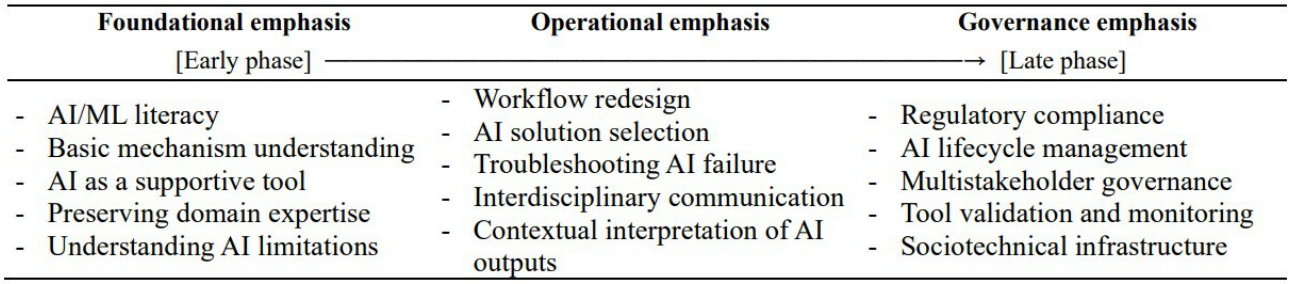
The present framework does not replace existing efforts but complements them by adding a functional-role perspective. Rather than organizing competencies by educational stakeholder type, expertise level, training phase, or implementation level, it organizes them around what healthcare professionals do with AI—using it in clinical care, contributing to its development, or governing it at organizational and system levels. The user-developer-leader structure is therefore a pragmatic, simplified, role-focused analytic framework for curriculum design, not a linear hierarchy or a set of mutually exclusive career pathways. This simplification reflects the distributed, context-dependent nature of AI-enabled healthcare, in which responsibilities overlap across individual, team, institutional, and policy levels. For example, educators may function as users when applying AI in teaching and assessment, as developers when co-designing educational AI systems, or as leaders when shaping institutional policies. Thus, the framework extends existing multilevel models while recognizing that individuals may move across or combine roles depending on context.

This study revealed an unresolved disagreement about the appropriate scope, depth, and educational priority of AI-related competencies across the health professions education continuum. Notably, experts' perspectives appeared to vary according to their professional backgrounds and practical experiences with AI. Experts from academia with AI research involvement more often discussed AI as a transformative force reshaping healthcare workflow, professional boundaries, and the distribution of expertise (SMEs 1, 2, 6, and 8), whereas experts in procedurally intensive fields emphasized foundational hands-on expertise and the preservation of human clinical judgment (SMEs 7, 10, and 12). At the same time, an expert familiar with digital transformation suggested that the very notion of what is considered foundational or fundamental is itself evolving in the AI era (SME 8). Meanwhile, experts with industry and development experience frequently highlighted implementation, communication, and governance issues across clinical and technical domains (SMEs 4, 11, and 13). Thus, the controversy was not simply whether specific content should be included, but how to balance technological preparedness with foundational professional mastery. Nevertheless, we draw a tentative conclusion that a minimal level of ML and data literacy should be included, consistent with the predominant expert view and patient-safety imperatives [31-34]. While rapid AI advances and shifting clinical workflows may reshape what constitutes necessary content, curricular debates should remain anchored in enduring professional functions [2,5,6].

The interview window extended over 3 years (October 2022 to October 2025), spanning the public emergence of large language models in clinical workflows. Comparing early- and late-period codes showed that the emphasis of expert concerns shifted—foundational issues dominated earlier interviews, while operational and system-level considerations were elaborated in greater specificity later (Figure 3; verbatim quotations by phase are provided in Multimedia Appendix 2). Later framings did not replace earlier ones: foundational concerns about the limits of AI and clinical judgment anchored expert reasoning throughout, and later interviews introduced no new role categories. This temporal consistency, despite a rapid change in the field, supports the durability of the role-differentiated framework, with content within each role, particularly at the system level, continuing to develop.

XSL•FO
RenderX

Figure 3. Temporal distribution of expert perspectives (early vs late period, 2022-2025). ML: machine learning.



Operationalizing a role-based framework raises questions about what to teach and where—across UME, GME, and CPD—to cultivate these competencies. Our findings revealed expert disagreement over whether detailed AI/ML instruction is necessary for all learners, suggesting that curricular implementation should remain context-sensitive rather than uniform across institutions. Institutions may embed user competencies as required UME content while offering developer and leader competencies through elective or advanced pathways, including physician-scientist pathways such as MD-PhD programs [35-37]. Conversely, emphasizing GME/CPD may mitigate UME crowding and leverage clinical experience [38,39], while hybrid sequencing can maintain continuity across stages [40,41]. The balance between core and elective content across UME, GME, and CPD will depend on institutional mission, learner trajectories, and faculty expertise.

To make the curriculum implications more actionable, assessment should also be aligned with the role-differentiated and layered structure of the framework. Because several competencies identified in this study are dispositional, practice-based, or systems-oriented rather than purely knowledge-based, they are unlikely to be adequately assessed through written examinations alone. Instead, they may be better evaluated through programmatic assessment that draws on multiple data points and complementary methods [42,43]. For user competencies, AI professionalism could be assessed through objective structured clinical examination–style stations or reflective documentation in which learners interpret an AI-generated recommendation, identify potential limitations or biases, communicate uncertainty to a patient, and document their final clinical decision and accountability rationale. For developer competencies, a challenging mindset and problem-solving could be assessed through simulation-based failure-mode tasks in which learners analyze why an AI tool performs poorly in a given clinical workflow, propose iterative refinements, and justify safety and bias-mitigation checks. For leader competencies, systemic insight could be assessed through case-based governance simulations or policy briefs that require learners to map stakeholders, anticipate implications for workflow and accountability, and design monitoring and escalation procedures. These assessment examples should be regarded as illustrative strategies rather than validated instruments; future work should develop rubrics and collect validity evidence across UME, GME, and CPD contexts.

Faculty readiness, raised by several experts, represents a further barrier to implementation [41,44]. Most current health professions educators were trained before AI became central to clinical practice, and few have direct experience integrating AI into teaching. Studies suggest that institutions must pursue short-term interinstitutional collaborations, including jointly developed faculty training programs, shared AI curriculum modules, and AI educator fellowships, to rapidly build a critical mass of qualified teaching faculty across institutions. A recent study has demonstrated the feasibility of structured, theory-informed faculty development workshops on integrating generative AI in medical education [45]. In parallel, long-term organizational readiness and systematic faculty development, supported by hybrid teaching models that pair clinical educators with AI specialists, have been proposed as a means to build sustainable internal capacity [40,46,47].

Beyond faculty readiness, implementation requires flexible adaptation of the framework across specialty contexts. Dentistry served as an information-rich case because AI applications are increasingly visible across dental clinical practice, including image interpretation, treatment planning, digital design, CAD/CAM-supported workflows, and procedural decision support [48-50]. These applications illustrate how AI can reshape clinical tasks, human-machine delegation, and professional responsibility. By contrast, in less procedure-oriented or less image-intensive specialties and care contexts, such as psychiatry, internal medicine, primary care, or longitudinal chronic care, the same role functions may be expressed through different AI applications, including risk prediction, documentation, triage, longitudinal decision support, patient communication, or population management. Thus, the framework should be understood as a set of transferable role functions that require specialty-specific operationalization, rather than as a model specific to dentistry.

Limitations and Future Research Recommendations

The user-developer-leader framework proposed in this study was derived from a limited number of expert interviews situated largely within a single national context in East Asia. Accordingly, the findings should be interpreted with attention to the healthcare delivery system, reimbursement structure, and national regulatory and AI policy landscape (including AI governance and data-privacy requirements). Because these requirements differ across jurisdictions—for example, between this context and frameworks, such as the Health Insurance Portability and Accountability Act (HIPAA), General Data Protection Regulation (GDPR), and European Union (EU) AI Act—the proposed model is best understood as an adaptable template rather than a one-size-fits-all approach.

In addition, because several dental experts were included in the sample, the findings may reflect AI use cases that are more

XSL•FO
RenderX

visible in imaging-intensive, CAD/CAM-supported, and procedure-oriented workflows. Their transferability may therefore vary in specialties or care contexts where AI is used more for longitudinal reasoning, narrative documentation, relational care, risk prediction, or population-level management. To address these boundaries, we applied explicit SME selection criteria and incorporated nondental clinical, medical education, engineering-related, and industry perspectives; nevertheless, transferability across other contexts and specialties remains to be empirically established.

Lastly, interrater reliability was not formally assessed, limiting quantitative evidence of coding consistency. Although this study, as a qualitative interpretive inquiry, aims to achieve transferability—supported by multiple trustworthiness procedures—rather than statistical generalizability or consensus-based validation, additional work is needed to strengthen the empirical grounding and translational utility of the proposed framework. We therefore envision several subsequent validation phases: (1) a modified Delphi process—consistent with structured, multiphase approaches to AI competency development, such as the AAMC's ongoing initiative to define competencies across the medical education continuum [13]—involving broader multistakeholder participation, including clinicians, learners, patients, educators, regulators, policymakers, and institutional leaders across diverse healthcare systems and specialties; (2) pilot curriculum implementation studies to evaluate feasibility, acceptability, and educational outcomes; and (3) iterative refinement informed by multi-institutional feedback and longitudinal evaluation. Future research should also translate these role-differentiated AI competencies into measurable learning outcomes and develop assessment approaches appropriate to users, developers, and leaders. Such work may help determine how AI competencies can be evaluated longitudinally across health professions curricula, rather than relying solely on a single knowledge-based assessment.

Conclusions

Despite the growing integration of AI into clinical practice, concrete guidance remains limited regarding which competencies physicians using AI should develop, what levels of proficiency are required, and which educational approaches are most appropriate [51]. In this context, our role-differentiated framework—and its translation into curriculum implications across the medical education continuum—offers a preliminary, hypothesis-generating starting point for institutions to make these decisions more explicit, to begin aligning training with real-world responsibilities, and to address the risk of diffused accountability in AI-enabled clinical care. In this way, adopting a role-based orientation in medicine may support physicians in remaining the "human-in-the-loop," serving as the moral and intellectual center of patient care.

Acknowledgments

The authors sincerely thank the experts who participated in the interviews and generously contributed their time and insights. Anthropic Claude (Opus 4.7) was used solely for language refinement during manuscript preparation and was not used for data collection, data analysis, or conceptual or interpretive decision-making. All AI-assisted language edits were reviewed and verified by the authors, who take full responsibility for the manuscript.

Funding

This work was supported by the following grants: (1) A grant from the National Research Foundation of Korea (NRF) funded by the Ministry of Science and ICT (#2021R1F1A1056465), and (2) A grant from the National Research Foundation of Korea (NRF) funded by the Ministry of Education (#RS-2024 - 00461340). The funder played no role in study design, data collection, the analysis and interpretation of data, or the writing of this manuscript.

Data Availability

The deidentified data supporting the findings of this study are available from the corresponding author upon reasonable request.

Authors' Contributions

Conceptualization: JL
Data curation: HK, SA
Formal analysis: YP, YH, HK, JL
Funding acquisition: JL
Investigation: YP, SA, JL
Methodology: YH, JL
Project administration: JL
Supervision: JL
Validation: SA, FWK
Visualization: YP, JL
Writing—original draft: YP, YH, JL
Writing—review & editing: YP, YH, HK, FWK, JL

XSL•FO
RenderX

Conflicts of Interest

None declared.

Multimedia Appendix 1
Additional illustrative quotes from subject matter experts mapped to themes and role-differentiated competency domains (users, developers, and leaders).
[DOCX File, 50 KB - mededu_v12i1e97608_app1.docx]

Multimedia Appendix 2
Verbatim quotations supporting Figure 3, tagged by interview phase.
[DOCX File, 87 KB - mededu_v12i1e97608_app2.docx]

References

1. Kalra N, Verma P, Verma S. Advancements in AI based healthcare techniques with FOCUS ON diagnostic techniques. Comput Biol Med 2024 Sep;179:108917. [doi: 10.1016/j.compbiomed.2024.108917] [Medline: 39059212]
2. Russell RG, Lovett Novak L, Patel M, et al. Competencies for the use of artificial intelligence-based tools by health care professionals. Acad Med 2023 Mar 1;98(3):348-356. [doi: 10.1097/ACM.0000000000004963] [Medline: 36731054]
3. Paranjape K, Schinkel M, Nannan Panday R, Car J, Nanayakkara P. Introducing artificial intelligence training in medical education. JMIR Med Educ 2019 Dec 3;5(2):e16048. [doi: 10.2196/16048] [Medline: 31793895]
4. Tolentino R, Baradaran A, Gore G, Pluye P, Abbasgholizadeh-Rahimi S. Curriculum frameworks and educational programs in AI for medical students, residents, and practicing physicians: scoping review. JMIR Med Educ 2024 Jul 18;10:e54793. [doi: 10.2196/54793] [Medline: 39023999]
5. Garvey KV, Thomas Craig KJ, Russell R, Novak LL, Moore D, Miller BM. Considering clinician competencies for the implementation of artificial intelligence-based tools in health care: findings from a scoping review. JMIR Med Inform 2022 Nov 16;10(11):e37478. [doi: 10.2196/37478] [Medline: 36318697]
6. Lee YM, Kim S, Lee YH, et al. Defining medical AI competencies for medical school graduates: outcomes of a Delphi survey and medical student/educator questionnaire of South Korean medical schools. Acad Med 2024 May 1;99(5):524-533. [doi: 10.1097/ACM.0000000000005618] [Medline: 38207056]
7. Moldt JA, Festl-Wietek T, Fuhl W, et al. Assessing AI awareness and identifying essential competencies: insights from key stakeholders in integrating AI into medical education. JMIR Med Educ 2024 Jun 12;10:e58355. [doi: 10.2196/58355] [Medline: 38989834]
8. Schubert T, Oosterlinck T, Stevens RD, Maxwell PH, van der Schaar M. AI education for clinicians. EClinicalMedicine 2025 Jan;79:102968. [doi: 10.1016/j.eclinm.2024.102968] [Medline: 39720600]
9. Miao F, Shiohira K, Lao N. AI Competency Framework for Students: UNESCO; 2024. URL: https://unesdoc.unesco.org/ark:/48223/pf0000391105 [accessed 2026-05-27] [doi: 10.54675/JKJB9835]
10. Miao F, Cukurova M. AI Competency Framework for Teachers: UNESCO; 2024. URL: https://unesdoc.unesco.org/ark:/48223/pf0000391104 [accessed 2026-05-27] [doi: 10.54675/ZJTE2084]
11. IACAI AI Integration Framework: Learner Matrix. : MedBiquitous; 2026 URL: https://www.medbiq.org/media/7991/download [accessed 2026-05-27]
12. Principles for the responsible use of artificial intelligence in and for medical education. Association of American Medical Colleges. URL: https://www.aamc.org/about-us/mission-areas/medical-education/principles-ai-use [accessed 2026-05-27]
13. Artificial intelligence competencies across the learning continuum. Association of American Medical Colleges. URL: https://www.aamc.org/about-us/mission-areas/medical-education/ai-competencies [accessed 2026-05-21]
14. Khamis N, Ungaretti T, Tackett S, Chen B. From AI literacy to leadership: milestones for faculty development in health professions education. Med Sci Educ 2025;35(6):2681-2686. [doi: 10.1007/s40670-025-02438-0]
15. Ng FYC, Thirunavukarasu AJ, Cheng H, et al. Artificial intelligence education: an evidence-based medicine approach for consumers, translators, and developers. Cell Rep Med 2023 Oct 17;4(10):101230. [doi: 10.1016/j.xcrm.2023.101230] [Medline: 37852174]
16. Braun V, Clarke V. Toward good practice in thematic analysis: avoiding common problems and be(com)ing a knowing researcher. Int J Transgend Health 2023;24(1):1-6. [doi: 10.1080/26895269.2022.2129597] [Medline: 36713144]
17. O'Brien BC, Harris IB, Beckman TJ, Reed DA, Cook DA. Standards for reporting qualitative research: a synthesis of recommendations. Acad Med 2014 Sep;89(9):1245-1251. [doi: 10.1097/ACM.0000000000000388] [Medline: 24979285]
18. Lincoln YS, Guba EG. Naturalistic Inquiry: Sage Publications; 1985.
19. Malterud K, Siersma VD, Guassora AD. Sample size in qualitative interview studies: guided by information power. Qual Health Res 2016 Nov;26(13):1753-1760. [doi: 10.1177/1049732315617444] [Medline: 26613970]
20. Kiger ME, Varpio L. Thematic analysis of qualitative data: AMEE Guide No. 131. Med Teach 2020 Aug;42(8):846-854. [doi: 10.1080/0142159X.2020.1755030] [Medline: 32356468]

XSL•FO
RenderX

21. Proudfoot K. Inductive/deductive hybrid thematic analysis in mixed methods research. J Mix Methods Res 2023 Jul;17(3):308-326. [doi: 10.1177/15586898221126816]
22. Moore IJ, Embry E, Padala PR. The pitfalls of patient-led AI use: a case study in clinical neuropsychology. Am J Geriatr Psychiatry Open Sci Educ Pract 2026 Mar;9:1-7. [doi: 10.1016/j.osep.2025.11.001]
23. Heinrichs H, Kies A, Nagel SK, Kiessling F. Physicians' attitudes toward artificial intelligence in medicine: mixed methods survey and interview study. J Med Internet Res 2025 Aug 26;27:e74187. [doi: 10.2196/74187] [Medline: 40857713]
24. Abdelwanis M, Alarafati HK, Tammam MMS, Simsekler MCE. Exploring the risks of automation bias in healthcare artificial intelligence applications: a Bowtie analysis. J Saf Sci Resil 2024 Dec;5(4):460-469. [doi: 10.1016/j.jnlssr.2024.06.001]
25. Budzyń K, Romańczyk M, Kitala D, et al. Endoscopist deskilling risk after exposure to artificial intelligence in colonoscopy: a multicentre, observational study. Lancet Gastroenterol Hepatol 2025 Oct;10(10):896-903. [doi: 10.1016/S2468-1253(25)00133-5] [Medline: 40816301]
26. Natali C, Marconi L, Dias Duran LD, Cabitza F. AI-induced deskilling in medicine: a mixed-method review and research agenda for healthcare and beyond. Artif Intell Rev 2025;58(11):356. [doi: 10.1007/s10462-025-11352-1]
27. Manca F, Eslava D. Digital and AI skills in health occupations: what do we know about new demand. OECD Artif Intell Pap 2025(36). [doi: 10.1787/5fbd42ab-en]
28. Ackerhans S, Wehkamp K, Petzina R, Dumitrescu D, Schultz C. Perceived trust and professional identity threat in AI-based clinical decision support systems: scenario-based experimental study on AI process design features. JMIR Form Res 2025 Mar 26;9:e64266. [doi: 10.2196/64266] [Medline: 40138691]
29. Habli I, Lawton T, Porter Z. Artificial intelligence in health care: accountability and safety. Bull World Health Organ 2020 Apr 1;98(4):251-256. [doi: 10.2471/BLT.19.237487] [Medline: 32284648]
30. Siala H, Wang Y. SHIFTing artificial intelligence to be responsible in healthcare: a systematic review. Soc Sci Med 2022 Mar;296:114782. [doi: 10.1016/j.socscimed.2022.114782] [Medline: 35152047]
31. Gordon M, Daniel M, Ajiboye A, et al. A scoping review of artificial intelligence in medical education: BEME Guide No. 84. Med Teach 2024 Apr;46(4):446-470. [doi: 10.1080/0142159X.2024.2314198] [Medline: 38423127]
32. Terranova C, Cestonaro C, Fava L, Cinquetti A. AI and professional liability assessment in healthcare. a revolution in legal medicine? Front Med (Lausanne) 2023;10:1337335. [doi: 10.3389/fmed.2023.1337335] [Medline: 38259835]
33. Hasanzadeh F, Josephson CB, Waters G, Adedinsewo D, Azizi Z, White JA. Bias recognition and mitigation strategies in artificial intelligence healthcare applications. NPJ Digit Med 2025 Mar 11;8(1):154. [doi: 10.1038/s41746-025-01503-7] [Medline: 40069303]
34. De Micco F, Di Palma G, Ferorelli D, et al. Artificial intelligence in healthcare: transforming patient safety with intelligent systems-a systematic review. Front Med (Lausanne) 2024;11:1522554. [doi: 10.3389/fmed.2024.1522554] [Medline: 39845830]
35. Yao MS, Huang L, Leventhal E, Sun C, Stephen SJ, Liou L. Leveraging datathons to teach AI in undergraduate medical education: case study. JMIR Med Educ 2025 Apr 16;11:e63602. [doi: 10.2196/63602] [Medline: 40239213]
36. Farmakidis AL, Singh A, Leaf K, Rossi A. Artificial intelligence curricula in US and Canadian medical schools. : AAMC; 2025 URL: https://www.aamc.org/media/84666/download [accessed 2025-12-28]
37. Abid A, Murugan A, Banerjee I, Purkayastha S, Trivedi H, Gichoya J. AI education for fourth-year medical students: two-year experience of a web-based, self-guided curriculum and mixed methods study. JMIR Med Educ 2024 Feb 20;10:e46500. [doi: 10.2196/46500] [Medline: 38376896]
38. Tolentino R, Hersson-Edery F, Yaffe M, Abbasgholizadeh-Rahimi S. AIFM-ed curriculum framework for postgraduate family medicine education on artificial intelligence: mixed methods study. JMIR Med Educ 2025 Apr 25;11:e66828. [doi: 10.2196/66828] [Medline: 40279148]
39. Schuur F, Rezazade Mehrizi MH, Ranschaert E. Training opportunities of artificial intelligence (AI) in radiology: a systematic review. Eur Radiol 2021 Aug;31(8):6021-6029. [doi: 10.1007/s00330-020-07621-y] [Medline: 33587154]
40. Kim DH, Kang YJ, Lee YM. Twelve tips for developing and implementing AI curriculum for undergraduate medical education. Med Educ Online 2025 Dec 31;30(1):2585637. [doi: 10.1080/10872981.2025.2585637] [Medline: 41327942]
41. Charow R, Jeyakumar T, Younus S, et al. Artificial intelligence education programs for health care professionals: scoping review. JMIR Med Educ 2021 Dec 13;7(4):e31043. [doi: 10.2196/31043] [Medline: 34898458]
42. Ross S, Hauer KE, Wycliffe-Jones K, et al. Key considerations in planning and designing programmatic assessment in competency-based medical education. Med Teach 2021 Jul;43(7):758-764. [doi: 10.1080/0142159X.2021.1925099] [Medline: 34061700]
43. Heeneman S, de Jong LH, Dawson LJ, et al. Ottawa 2020 consensus statement for programmatic assessment - 1. Agreement on the principles. Med Teach 2021 Oct;43(10):1139-1148. [doi: 10.1080/0142159X.2021.1957088] [Medline: 34344274]
44. Ngo B, Nguyen D, vanSonnenberg E. The cases for and against artificial intelligence in the medical school curriculum. Radiol Artif Intell 2022 Sep;4(5):e220074. [doi: 10.1148/ryai.220074] [Medline: 36204540]
45. Anand R, Bowers N, Manyama MF. Design and evaluation of a faculty development workshop series on integrating generative artificial intelligence in medical education: mixed methods pilot study. JMIR Med Educ 2026 May 6;12:e89815. [doi: 10.2196/89815] [Medline: 42090298]

XSL•FO
RenderX

46. Park TI, Seo J, Yoon HJ, Lee KE. Institutionalizing convergence education for medical artificial intelligence. Biomed Eng Lett 2025 Nov;15(6):1073-1083. [doi: 10.1007/s13534-025-00523-2] [Medline: 41280148]
47. Grunhut J, Marques O, Wyatt ATM. Needs, challenges, and applications of artificial intelligence in medical education curriculum. JMIR Med Educ 2022 Jun 7;8(2):e35587. [doi: 10.2196/35587] [Medline: 35671077]
48. Ding H, Wu J, Zhao W, Matinlinna JP, Burrow MF, Tsoi JKH. Artificial intelligence in dentistry-a review. Front Dent Med 2023;4:1085251. [doi: 10.3389/fdmed.2023.1085251] [Medline: 39935549]
49. Araidy S, Batshon G, Mirochnik R. Artificial intelligence applications in dentistry: a systematic review. Oral 2025;5(4):90. [doi: 10.3390/oral5040090]
50. Yeslam HE, Freifrau von Maltzahn N, Nassar HM. Revolutionizing CAD/CAM-based restorative dental processes and materials with artificial intelligence: a concise narrative review. PeerJ 2024;12:e17793. [doi: 10.7717/peerj.17793] [Medline: 39040936]
51. Schuitmaker L, Drogt J, Benders M, Jongsma K. Physicians' required competencies in AI-assisted clinical settings: a systematic review. Br Med Bull 2025 Jan 16;153(1):ldae025. [doi: 10.1093/bmb/ldae025] [Medline: 39821209]

Abbreviations

AAMC: Association of American Medical Colleges
ACGME: Accreditation Council for Graduate Medical Education
CAD: computer-aided design
CAM: computer-aided manufacturing
CPD: continuing professional development
GDPR: General Data Protection Regulation
GME: graduate medical education
HIPAA: Health Insurance Portability and Accountability Act
IACAI: International Advisory Committee for Artificial Intelligence
ML: machine learning
SME: subject matter expert
SRQR: Standards for Reporting Qualitative Research
UME: undergraduate medical education
UNESCO: United Nations Educational, Scientific and Cultural Organization

Edited by A Stone; submitted 08.Apr.2026; peer-reviewed by D Dellacorte, J Si, N Khamis, S Ismaile, X Fan; revised version received 11.Aug.2026; accepted 12.Aug.2026; published 08.Sep.2026.

Please cite as:
Park Y, Han Y, Kim H, Ahn S, Kron FW, Lee J
Role-Differentiated AI Competencies and Curriculum Implications for Health Professions Education: Qualitative Study
JMIR Med Educ 2026;12:e97608
URL: https://mededu.jmir.org/2026/1/e97608
doi:10.2196/97608

© Yuyi Park, Yejin Han, Hyeongjo Kim, Solmoe Ahn, Frederick William Kron, Jihyun Lee. Originally published in JMIR Medical Education (https://mededu.jmir.org), 8.Sep.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Education, is properly cited. The complete bibliographic information, a link to the original publication on https://mededu.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

AI-Assisted Angoff Standard Setting for Multiple-Choice Examinations in Medical Education: Comparative Study

Salah Eldin Kassab[1,2], MBChB, PhD; Mariam Shadan[1], MBBS, MD; Arina Ziganshina[3], MD, PhD; Shifan Khanday[1], MBBS, PhD; Hossam Hamdy[3], MBChB, PhD

[1]Department of Biomedical Sciences, College of Medicine, Dubai Medical University, Al Muhaisnah 1, Al Mizhar, Dubai, United Arab Emirates

[2]Department of Physiology, Faculty of Medicine, Suez Canal University, Ismailia, Egypt

[3]Department of Clinical Sciences, College of Medicine, Dubai Medical University, Dubai, United Arab Emirates

Corresponding Author:
Salah Eldin Kassab, MBChB, PhD
Department of Biomedical Sciences, College of Medicine, Dubai Medical University, Al Muhaisnah 1, Al Mizhar, Dubai, United Arab Emirates

Abstract

Background: Standard setting is essential for a defensible assessment in medical education. The modified Angoff method requires several expert judges, and determining the minimally competent candidate is cognitively challenging. However, empirical evidence on the role of AI in standard setting is unclear.

Objective: This study aimed to examine the role of large language models (LLMs) in the modified Angoff standard setting method for multiple-choice questions in a basic medical science examination compared with faculty judges.

Methods: This study was conducted in year 2 of the MD program in the United Arab Emirates. Ten faculty judges and 5 LLMs (GPT-5.2, Grok, DeepSeek, MedGemma, and Claude 4.5 Sonnet) determined the Angoff cutoff scores for a summative examination (120 multiple-choice questions). Standardized prompts were used for the LLMs to mimic the same information provided for faculty judges. Generalizability (*G*) theory analyses were performed using a fully crossed item × rater design to estimate variance components, *G* and Φ coefficients, decision study, and root mean squared error (RMSE) of the Angoff cutoff scores.

Results: Faculty-generated Angoff estimates (mean 69.13, SD 10.92) were comparable to LLM-generated estimates (mean 68.94, SD 12.39). A 2-tailed paired-sample *t* test revealed no statistically significant difference between the 2 groups (95% CI −2.21 to 2.91; t_{119}=0.27; *P*=.79). Generalizability theory analysis demonstrated moderate reliability for the faculty panel (*G* coefficient=0.738; Φ coefficient=0.692). Despite comprising only 5 LLMs, the LLM panel demonstrated higher reliability (*G* coefficient=0.823; Φ coefficient=0.815), lower RMSE (1.28 vs 2.33), higher item-related variance (46.85% vs 18.36%), and lower rater-related variance (2.64% vs 16.40%) than the faculty panel. Pass rates were similar using LLM- and faculty-derived cutoff scores (52/73, 71.2%). LLMs differed in their minimally competent candidate conceptualization and approaches to determining item-level percentages of correct responses. Furthermore, the correlations between Angoff estimates and item-related *P* values were larger in LLMs than in faculty judges (*r*=0.552 vs 0.437).

Conclusions: In this single-institution study, the evaluated LLMs generated modified Angoff estimates that were broadly comparable to those of faculty judges. Generalizability theory analyses demonstrated higher *G* and Φ coefficients, lower rater-related variance, and lower RMSE for the evaluated LLM outputs under standardized prompting conditions, indicating greater consistency in the generated estimates. Application of the resulting cutoff scores produced pass and fail rates similar to those derived from faculty judges. These findings suggest the future application of LLMs as decision assistance tools for modified Angoff standard setting while maintaining expert human oversight.

(JMIR Med Educ 2026;12:e98766) doi:10.2196/98766

KEYWORDS

large language models; standard setting; modified Angoff method; medical education; multiple-choice examinations; artificial intelligence; AI

XSL•FO
RenderX

Introduction

Background

Standard setting is an essential quality measure for student assessment, ensuring that decisions about examination scores reflect meaningful levels of student competence [1]. Unlike using fixed cutoff scores, standard setting provides a flexible approach to align pass and fail decisions with examination difficulty. This process includes a systematic integration of expert judgment to determine the cutoff score that distinguishes pass from fail or competent from noncompetent candidates. There is no "gold standard" for determining the cutoff score for the pass and fail decisions. However, the credibility of these decisions depends on the rigor and defensibility of and well-supported evidence on the process [2]. Furthermore, the generated cutoff score depends on the method of standard setting used and the panel of expert judges [3,4].

The Angoff method is widely used for determining cutoff scores based on expert judgment of minimally competent candidate (MCC) performance [5-8]. This method has the advantage of being criterion referenced, which ensures alignment with established professional competence frameworks [9]. In contrast, norm-referenced methods carry the risk of erroneous promotion of incompetent candidates and the inappropriate failure of competent candidates. Multiple variants of the Angoff method have been reported in the health professions education literature, among which the modified Angoff method with reality check achieved the highest reliability, representing the most defensible and consistent procedure [9,10]. However, this process is often time-consuming and resource intensive [5]. In addition, conceptualizing the MCCs represents a cognitive burden on the expert judges [11]. Therefore, extensive training of the expert panelists is an essential requirement to reduce the misconceptions among judges and subjective variability of the cutoff scores [12].

Prior Work

With the increasing capabilities of large language models (LLMs), there is growing interest in exploring how AI can assist in student assessment [13-15]. However, there is very limited peer-reviewed research in medical education on the use of AI for evaluating Angoff standard setting in a manner comparable to that of human expert panels [14]. The integration of AI into standard setting could be conceptualized through the substitution, augmentation, modification, and redefinition (SAMR) theoretical framework [16], which describes 4 levels of integrating technology into educational practices [16]. At the substitution level, AI may substitute the faculty role by conceptualizing the MCC and generating the Angoff estimates. At the augmentation level, AI can work as a collaborative partner to improve the efficiency of the process through rapid analysis and feedback that supports faculty judgment while maintaining human oversight [17]. At the modification and redefinition levels, AI can modify the process by synthesizing judgments and potentially redefine standard setting practice to refine Angoff estimates. However, the SAMR model should not be viewed as a rigid hierarchy but as a flexible framework to guide the meaningful integration of AI into the standard setting context.

Study Objectives

This study examined how LLMs can support the modified Angoff standard setting process for multiple-choice question (MCQ) examinations using the SAMR framework. Specifically, we assessed whether LLMs can replicate faculty judgment (substitution) or enhance the uniformity and exactness of standard setting (augmentation). This study addressed three questions: (1) Are there significant differences between Angoff estimates from faculty judges and LLMs? (2) How do different LLMs conceptualize the MCC and estimate item-level probabilities? (3) What evidence supports the reliability and credibility of Angoff estimates from faculty vs LLMs?

Methods

Ethical Considerations

The Institutional Research Ethics Committee of Dubai Medical University approved this study (approval number DMU/IRB/OEJW9A/2026/30). Participation of faculty judges was voluntary, and we obtained informed consent from each participant prior to their involvement in the study. Individuals were informed that they may withdraw at any stage without penalty. Anonymity and confidentiality were maintained by assigning unique study codes to all expert contributors and removing any identifying information from records. All study data were securely stored in password-protected files accessible only to the study authors. Similarly, all student examination data were anonymized prior to analysis, and no individual-level student identifiers were accessible to either the human judges or the LLMs. The dataset provided to the LLMs included cumulative grade point averages (GPAs) together with study-specific anonymized identification codes used solely to distinguish individual records during analysis. Institutional student identification numbers and all other personally identifiable information were removed before the data were uploaded in the PC of the first author (SEK). The anonymized study codes could not be linked to individual students by the LLMs or any third-party service.

Study Context

This study was conducted in year 2 of the preclinical phase of the MD program at a medical school in the United Arab Emirates. The 6-year MD program encompasses 226 credit hours and consists of 3 phases, each with a duration of 2 years. Phase 1 of the program consists of foundational courses in basic medical sciences and general education. Phase 2 constitutes integrated, system-based courses with early clinical exposure and research training. Phase 3 consists of hospital-based clerkships in clinical specialties and subspecialties.

Study Design and Sampling

This study was designed to compare the generated cutoff scores using the modified Angoff method by faculty judges (n=10) vs 5 LLMs (generative pretrained transformer [OpenAI], Grok [SpaceXAI LLC], DeepSeek, MedGemma [Google DeepMind], and Claude [Anthropic]). This study was conducted on a

XSL•FO
RenderX

summative final examination of basic medical sciences that consisted of A-type (single best response) MCQs with 4 options. The content of the examination was related to biochemistry, molecular biology, genetics, and fundamentals of pharmacology. Test items were constructed by subject matter experts followed by peer review and then validated by the college assessment committee members to ensure content validity and correct any technical flaws in the design.

Standard Setting Methods

Conducting the Modified Angoff Procedure by Faculty Judges

Prior to the standard setting exercise, faculty judges (n=10) participated in a structured training program that consisted of 2 sessions of 1 hour each delivered by the first author (SEK). The judges were selected from a pool of basic sciences faculty who were familiar with the curriculum, cohort of students, and student assessment system and had at least 2 years of experience teaching this curriculum. The 10-judge panel consisted of 4 (40%) biochemistry faculty members, 2 (20%) molecular and cell biology faculty members, 2 (20%) pharmacology faculty members, and 2 (20%) lecturers in biomedical sciences.

The sessions focused on the definition of standard setting, the purpose of its implementation, and the steps of determining the cutoff scores using the modified Angoff method. An essential aspect of the training was how to conceptualize the MCC. The following definition was adopted for the MCC: "A student who studies an average amount of time, possesses knowledge which is sufficient for borderline pass, but struggles to achieve scores above the cut-off threshold" [18]. The sessions ended with hands-on training on 5 MCQs outside the examination content by the faculty judges and elaboration on the individual differences between the generated Angoff estimates. At the end of the training sessions, faculty judges were provided with a Microsoft Excel sheet on the previous cumulative GPAs of the cohort of students. This "reality check" process aimed to reduce the variability between the judges in conceptualizing the MCC group of students. Faculty judges were then asked to independently estimate how much the minimally competent group of students was expected to score on each test item. The 10 faculty judges independently rated all 120 examination items. Correct answers to the test items were provided to the judges. However, there was no group discussion or establishment of consensus for each test item. The average scores from each faculty judge were calculated, and the final cutoff score represented the mean estimates of the 10 judges.

Prompting Process for Conducting the Angoff Procedure by the LLMs

The specific versions of the LLMs evaluated were GPT-5.4, Grok 4.20, DeepSeek-V3.2, Claude 4.5 Sonnet, and MedGemma 1.5, each accessed via their respective standard web-based chat interfaces or via Google's Gemini Enterprise Agent Platform Model Garden in the case of the latter. All models were used with each platform's default generation settings and were accessed between December 2025 and March 2026. A standardized prompting protocol for LLMs was designed to ensure the reproducibility of the study findings (Multimedia Appendix 1). The prompts were designed to match the same information provided to the faculty judges during the training session. Specifically, each LLM was fed the following information: (1) current GPA of the cohort under study; (2) content-validated MCQs with the correct answers; and (3) a Microsoft PowerPoint presentation describing the process of standard setting, rationale for its use, modified Angoff method, and concept of the MCC. The prompt to each LLM was to assume the role of a faculty expert who was participating in the modified Angoff method and identify the expected percentage of MCC students who scored each item correctly. Following generation of the initial output, each LLM was prompted to articulate its definition of the MCC and explain the rationale for the estimated percentage score for the items. The wording of the prompts, sequence of the uploaded information, and instructions were used consistently across all 5 LLMs with no change or modification. Each LLM was blinded to the scores of the faculty judges and the outputs of the other models. The overall LLM cutoff score was generated from the average scores of the 5 models.

Cohen Method for Standard Setting

We also compared the application of the modified Angoff method with the method by Cohen-Schotanus and van der Vleuten, which is a norm-referenced approach to standard setting [19,20]. The cutoff score was determined using the following formula:

$$\text{Cutoff score} = cN + 0.60\ (N^* - cN)$$

In this equation, c is the proportion of correct responses expected by chance (0.25 for a 4-option multiple-choice examination), N is the total number of examination items (120 in the present study), and N^* is the raw examination score of the student at the 95th percentile of the score distribution.

Statistical Analysis

Descriptive statistics were calculated for all cutoff scores, including means and SDs. To account for measurement error in observed examination scores, the estimated cutoff scores generated by the faculty-derived modified Angoff method, the LLM-derived modified Angoff method, and the Cohen method were converted into operational cutoff scores by subtracting 1 SE of measurement. The SE of measurement was calculated according to classic test theory as follows:

$$SEM = SD \times \sqrt{1 - r}$$

In this equation, SD is the SD of students' raw examination scores, and r is the examination reliability (Cronbach α). The adjusted operational cutoff scores were subsequently used to determine students' pass and fail classifications as this approach accounts for measurement error in observed examination scores and improves the defensibility of pass and fail decisions [21].

Paired-sample t tests were used to compare the mean estimates between the faculty judges and LLMs for the same examination items. A 2-way random-effects ANOVA, consistent with a generalizability (G) theory framework, was conducted separately for faculty judges and LLMs using a fully crossed item × rater design [22]. Variance components were estimated for items, raters, and the item × rater interaction, and the percentage of

XSL•FO
RenderX

the total variance attributable to each component was calculated. Generalizability (*G*) coefficients were calculated to evaluate relative reliability, whereas Φ coefficients were calculated to evaluate absolute dependability for criterion-referenced decisions. A decision study estimated the projected *G* coefficients for panels ranging from 4 to 20 raters. In addition, the root mean squared error (RMSE) was calculated as a summary measure of the precision of the estimated Angoff cutoff scores, with lower values indicating greater consistency among raters. To calculate the RMSE, the following previously published formula was applied [18,23]:

(σ^r2/nr)+(σ^ir2/(ninr))

In this equation, n_i is the number of items, σ^r2 is the rater variance component, and σ^ir2 is the item × rater interaction variance component.

Pearson correlation was used to test the relationship between the mean value of item-related estimates for raters (faculty judges and LLMs) and the *P* value for each item (percentage of students answering each item correctly). The item *P* values were calculated from the same administration of the examination involving the 73 students included in this study. A *P* value below .05 was considered statistically significant. Data were analyzed using SPSS (version 24; IBM Corp). Generalizability theory analyses for standard setting estimates were performed using a publicly available web-based generalizability theory application (Reliability Angoff Calculator; shinyapps) [21].

Results

Overview

Seventy-three students completed the written examination. The mean raw examination score was 74.6% (SD 15.0%), with scores ranging from 38.3% to 100%. The examination demonstrated strong internal consistency (Cronbach α=0.91). The mean item discrimination was 0.32 (SD 0.12) with no negatively discriminating items. Table 1 shows the mean Angoff estimate scores and SDs from the 10 faculty judges and 5 LLMs. Faculty estimates ranged from a mean of 59.23 (SD 9.97) to 81.21 (SD 17.13). Among LLMs, DeepSeek produced the highest mean cutoff score (72.15, SD 20.06), whereas GPT-5.2 produced the lowest (66.18, SD 12.34).

Table . Mean Angoff estimates generated by faculty judges (n=10) and 5 large language models (GPT-5.2, Grok, DeepSeek, MedGemma, and Claude 4.5 Sonnet).

Rater type and model		Cut score (%), mean (SD)
Faculty judges		
	Judge 1	81.21 (17.13)
	Judge 2	68.62 (16.76)
	Judge 3	64.90 (16.72)
	Judge 4	63.11 (27.22)
	Judge 5	67.83 (15.17)
	Judge 6	63.96 (9.72)
	Judge 7	66.58 (22.42)
	Judge 8	59.23 (9.97)
	Judge 9	78.71 (8.00)
	Judge 10	77.17 (9.97)
LLMs		
	ChatGPT	66.18 (12.34)
	Grok	67.79 (14.29)
	DeepSeek	72.15 (20.06)
	Med-PaLM 2	71.88 (14.38)
	Claude Sonnet 4.5	66.68 (18.64)

Differences Between the Angoff Estimates Generated by Faculty Judges and LLMs

A paired-sample *t* test showed that the Angoff cutoff scores produced by the LLMs (mean 68.94, SD 12.39) were closely comparable to those produced by faculty judges (mean 69.13, SD 10.92), and the difference between the 2 groups did not reach statistical significance (95% CI −2.21 to 2.91; t_{119}=0.27; *P*=.79).

LLM Conceptualization of the MCC and Estimation of Item-Level Probabilities of Percentage of Correct Responses

Table 2 shows that the LLMs differed in how they conceptualized the MCC and estimated the probability of a correct response. GPT-5.2, Claude 4.5 Sonnet, and Grok based their conceptualization on learner characteristics. GPT-5.2 and Claude 4.5 Sonnet described the MCC as a student with adequate basic knowledge who struggles with advanced mental tasks.

XSL•FO
RenderX

Grok similarly emphasized predictable errors and challenges in distinguishing related concepts. DeepSeek used cohort GPA distribution to identify students near the pass threshold, whereas MedGemma focused on item characteristics rather than learner attributes.

LLMs also differed in estimating the percentage of correct responses. GPT-5.2 and Claude 4.5 Sonnet considered item-level features such as cognitive level, content familiarity, distractor quality, and likely misconceptions. DeepSeek and MedGemma focused on the complexity of the subject matter, the depth of conceptual knowledge required, and the cognitive processing required for answering each MCQ. Grok anticipated errors, misconceptions, and possible reasoning failures of students.

Table . Large language model (LLM) conceptualization of the minimally competent candidate (MCC) and estimation of item-level probability of correct response across LLMs.

LLMs	Conceptualization of the MCC	Process for estimating the percentage of correct responses
GPT-5.2	The MCC is a learner with adequate core knowledge but limited ability to apply higher-order reasoning, particularly in tasks requiring integration or multiple steps.	Estimates are derived through structured consideration of item characteristics, including cognitive level, familiarity, clarity, and the effectiveness of distractors. These elements are weighed to reflect the overall cognitive demand of the question.
Claude 4.5 Sonnet	The MCC is a learner with variable performance capable of handling straightforward items but less reliable when confronted with complexity or layered reasoning.	Estimates are generated using a comparative, justification-oriented process. Items are evaluated in relation to one another, with adjustments made to maintain consistency across the assessment rather than treating each item independently.
Grok	The MCC is characterized by a tendency toward common and predictable errors, such as misinterpretation of questions, confusion between similar concepts, or computational mistakes.	Estimates are informed by anticipated error patterns. The likelihood of incorrect reasoning pathways is considered when determining how often a borderline candidate would answer correctly.
DeepSeek	The MCC is defined operationally, with borderline candidates identified through GPA[a] ranges approximating the pass threshold.	Estimates are determined through a stepwise analytical approach involving content review, difficulty classification, and justification of each estimate. Item-level probabilities are subsequently aggregated to inform the overall cutoff score.
MedGemma	The MCC is not explicitly described in terms of learner attributes; instead, performance is inferred relative to item difficulty.	A difficulty-based model is used in which the probability of a correct response decreases progressively with increasing item difficulty. This pattern is supported by both item-specific reasoning and overall trends.

[a]GPA: grade point average.

Reliability of Angoff Estimates Among Faculty Judges Compared With LLMs

Generalizability Theory Analysis

Table 3 and Table 4 summarize the generalizability theory analysis of the modified Angoff procedure for faculty judges and LLMs. Generalizability theory analysis demonstrated moderate reliability of the modified Angoff cutoff scores for the 10-member faculty panel. The generalizability (G) coefficient was 0.738, and the corresponding Φ coefficient was 0.692, indicating acceptable dependability for both relative and absolute pass and fail decisions. The decision study (Figure 1) demonstrated that reliability improved progressively as the number of faculty judges increased, with the G coefficient increasing from 0.530 with 4 judges to 0.849 with 20 judges.

Table . Generalizability theory analysis of Angoff estimates generated by faculty judges (n=10) and 5 large language models (LLMs; GPT-5.2, Grok, DeepSeek, MedGemma, and Claude 4.5 Sonnet): variance components.

Source of variance	Faculty judges (n=10)		LLMs (n=5)	
	Variance component (% points)[2]	Percentage of variance	Variance component (% points)[2]	Percentage of variance
Items	58.94	18.36	126.36	46.85
Raters	52.64	16.40	7.12	2.64
Item × rater	209.50	65.25	136.24	50.51
Total variance	321.08	100	269.72	100

XSL•FO
RenderX

Table . Reliability and precision indexes.[a]

Reliability and precision indexes	Faculty judges (n=10)	LLMs[b] (n=5)
G coefficient[c]	0.738	0.823
Φ coefficient	0.692	0.815
RMSE[d]	2.33	1.28

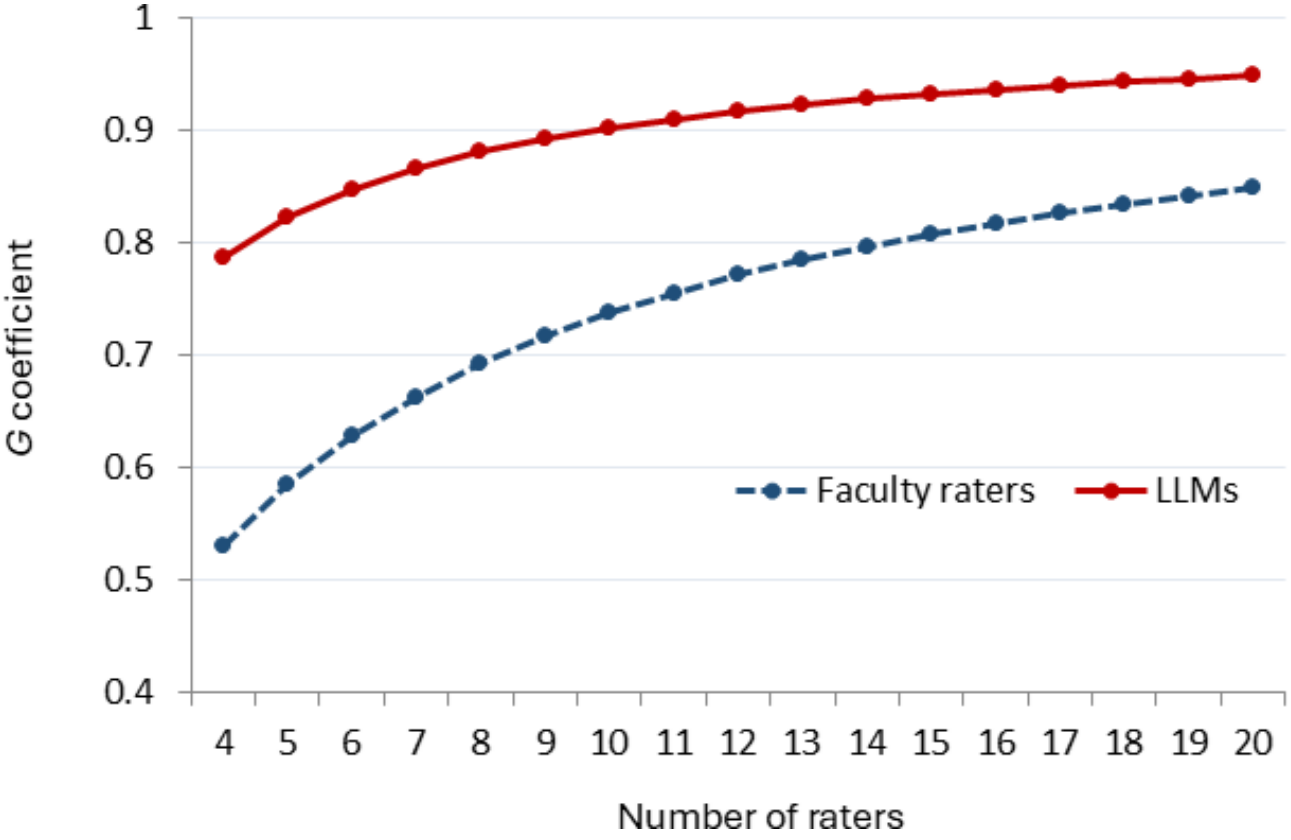
[a]The *G* coefficient estimates relative reliability, whereas the Φ coefficient estimates absolute reliability for criterion-referenced decisions.

[b]LLM: large language model.

[c]Generalizability coefficient.

[d]RMSE: root mean squared error of the estimated Angoff cutoff scores.

Figure 1. Decision study showing the effect of increasing the number of raters on the generalizability (*G*) coefficient of the modified Angoff procedure. The graph compares projected *G* coefficients for faculty judges and large language models (LLMs) across panel sizes ranging from 4 to 20 raters.



Compared with the faculty panel, the 5 LLMs demonstrated greater psychometric consistency. The LLM panel achieved a *G* coefficient of 0.823 and a Φ coefficient of 0.815 despite comprising only 5 models. The decision study similarly showed progressive improvements in reliability with an increasing number of LLM raters, with the *G* coefficient increasing from 0.788 with 4 LLMs to 0.949 with 20 LLMs.

The generalizability theory analysis found that LLMs had much larger item-related variance than faculty judges (46.85% vs 18.36%), indicating that LLMs discriminated more between examination items. In contrast, rater-related variance was substantially smaller for LLMs than for faculty judges (2.64% vs 16.40%), indicating more consistency in LLM-generated Angoff estimations. The item × rater interaction was the most significant source of variance in both categories, accounting for 65.25% of the total variance among faculty judges and 50.51% among LLMs. Similarly, LLMs had a smaller RMSE than faculty judges (1.28 vs 2.33), indicating improved accuracy in the predicted cutoff scores.

Failure Rates After Applying 3 Different Standards

The adjusted cutoff scores were 64.67% for the faculty-derived modified Angoff method, 64.48% for the LLM-derived modified Angoff method, and 68.20% for the Cohen method (Table 5). Applying these cutoff scores to students' raw examination scores yielded similar pass and fail classifications for the faculty- and LLM-derived Angoff methods, with 71.2% (52/73) passing and 28.8% (21/73) failing. In contrast, the Cohen method produced a higher operational cutoff score, classifying 64.4% (47/73) of the pupils as passing and 35.6% (26/73) as failing.

XSL•FO
RenderX

Table . Standard setting outcomes and individual-level agreement for the modified Angoff method (faculty judges and large language models [LLMs]) and the Cohen method: operational cutoff scores and pass and fail outcomes. Pass and fail classifications were determined using students' raw examination scores.

Standard setting method	Adjusted cutoff score (%)	Pass rate (n=73), n (%)	Fail rate (n=73), n (%)
Angoff method—faculty	64.67	52 (71.2)	21 (28.8)
Angoff method—LLMs	64.48	52 (71.2)	21 (28.8)
Cohen method	68.20	47 (64.4)	26 (35.6)

Individual-level agreement analysis demonstrated perfect agreement between the faculty- and LLM-derived Angoff methods (100% agreement; Cohen κ=1.000; exact McNemar *P*>.99). Comparisons between the Cohen method and each modified Angoff approach showed 93.2% overall agreement, with almost perfect agreement (κ=0.844) and no statistically significant differences in pass and fail classifications (exact McNemar *P*=.06 for both comparisons; Table 6).

Table . Individual-level agreement between standard setting methods.[a]

Comparison	Overall agreement (%)	Cohen κ	Exact McNemar *P* value
Angoff method—faculty vs LLMs[b]	100	1.000	>.99
Angoff method—faculty vs Cohen method	93.2	0.844	.06
Angoff method—LLMs vs Cohen method	93.2	0.844	.06

[a]Agreement was assessed using the Cohen κ and the exact McNemar test. A nonsignificant McNemar test indicates no statistically significant difference in paired pass and fail classifications.

[b]LLM: large language model.

Relationship Between Angoff Estimates and Item Difficulty

Pearson correlations showed statistically significant positive relationships between Angoff estimates and observed item difficulty for both groups. The correlation was moderate for LLMs (*r*=0.552; *P*<.001) and weaker for faculty judges (*r*=0.437; *P*<.001). Because the difference between these dependent correlations was not formally tested, these coefficients are presented descriptively rather than as evidence that one correlation was statistically greater than the other.

Discussion

Principal Findings

The Angoff method is the most commonly used standard setting approach in health professions education [1,20,24]. This is, to our knowledge, the first published study comparing LLMs and human judges in Angoff standard setting for medical education. Four key findings emerged. First, paired analysis showed that Angoff estimates from 10 faculty judges and 5 LLMs did not differ significantly, resulting in comparable operational cutoff scores and identical pass and fail classifications. Second, generalizability theory analysis found acceptable reliability for both the faculty and LLM panels, with reliability increasing as the number of raters grew. The LLM panel also showed lower rater-related variation, lower RMSE, and higher item-related variance than the faculty panel, indicating better consistency and discrimination across the evaluated LLM outputs. Third, there was significant heterogeneity in how LLMs conceptualized the MCC and estimated item levels. Fourth, both faculty- and LLM-derived Angoff estimates showed moderate positive relationships with reported item difficulty, supporting, but not conclusively, criterion-related validity. Overall, these data indicate that LLMs may be helpful as adjuncts to faculty judges in modified Angoff standard setting while emphasizing the importance of expert human judgment in high-stakes assessment.

Implications of the Findings

The lack of significant differences between LLM and faculty estimates suggests that, with appropriate contextual information and structured prompts, LLMs can approximate expert human judgment in predicting whether a minimally competent student will answer an item correctly. This is important because the credibility of the Angoff method depends on the quality of judges, their conceptualization of the MCC, and their accuracy in predicting item scores [23]. Credibility in standard setting is defined by stakeholders' trust that the cutoff scores are appropriate, defensible, and fit for purpose, best demonstrated through a validity argument framework using multiple sources of evidence [25].

This study provides 3 sources of validity evidence. The first is *procedural validity*. The faculty-based standard-setting process involved an adequate number of trained judges, a clear MCC definition, explicit guidance on the modified Angoff method, and access to historical student performance data. The same information and standardized prompts were given to each LLM, ensuring consistent instructions for all AI raters.

The second source is *internal validity*, reflecting the degree of stability of Angoff estimates in response to sources of measurement errors such as judges or items [18]. Generalizability theory was used instead of a conventional interrater agreement coefficient because it partitions measurement error into components attributable to items, raters,

XSL•FO
RenderX

and their interaction [22]. The faculty panel achieved *G* and Φ coefficients that demonstrated moderate reliability for relative and absolute pass and fail decisions. On the other hand, the 5-LLM panel demonstrated higher reliability despite comprising only half as many raters as the faculty judges. Consistent with these findings, the RMSE was lower for LLMs than for faculty judges, indicating greater consistency of the evaluated LLM outputs under the standardized prompting conditions.

Examination of the variance components provides additional insights into these findings. The rater-related variance among LLMs was 6 times lower than that among faculty judges, which suggests greater agreement under standardized prompting conditions. Previous studies using progress tests have reported even lower rater-related variance, likely reflecting differences in assessment context, examination format, and judge characteristics rather than differences in the standard setting method itself [18,23]. In contrast, item-related variance for LLMs was 2.5 times higher than that for faculty judges. These findings suggest that Angoff estimates for LLMs were influenced more by differences in item characteristics, whereas faculty judgments were affected to a greater extent by rater-specific variation. Nevertheless, the substantial item × rater interaction in both groups indicates that variability in ratings remained largely dependent on individual examination items. However, the greater consistency observed among LLMs should not be interpreted as evidence of superior validity. The convergence among LLMs may partly reflect similarities in their training data, model architectures, and response patterns rather than fully independent judgments. Therefore, although the findings support the reliability of LLM-assisted standard setting, human expert judgment remains essential for evaluating content relevance, contextual appropriateness, and the expected performance of an MCC.

The third source of evidence is *external validity*, assessed in two ways. First, faculty- and LLM-derived cutoff scores produced similar pass rates, indicating comparable practical outcomes for students. The Cohen norm-referenced method resulted in a much higher failure rate, consistent with existing literature [20]. Second, LLM Angoff estimates demonstrated a moderately stronger correlation with observed item difficulty than faculty estimates, indicating closer alignment with observed difficulty of test items. However, this stronger correlation should be interpreted cautiously as item *P* values reflected the overall cohort performance rather than that of MCCs. Therefore, these findings provide supportive rather than definitive evidence of external validity. Collectively, these findings support LLMs as credible contributors to modified Angoff standard setting while highlighting the ongoing need for informed human decision-making in high-stakes assessment.

The SAMR model provides a useful conceptual lens for interpreting the potential role of LLMs in modified Angoff standard setting [16]. On the basis of the observed findings, LLMs appear to align with the substitution and augmentation levels of the framework, whereas we recognize that this interpretation is conceptual rather than an empirical outcome of the present study. At the substitution level, LLMs replicated the core faculty role by conceptualizing the MCC and generating Angoff estimates comparable to those of expert judges. Within the SAMR framework, these findings may be interpreted as aligning with the augmentation level because the evaluated LLMs demonstrated greater consistency under standardized prompting conditions. The diverse reasoning approaches among LLMs, from cognitive feature–based and difficulty-driven models to error anticipation, highlight their potential as cognitive partners in assessment decision-making [26]. These findings support a practical role for LLMs as substitutes for or augmentations to the human Angoff process. LLMs could improve efficiency, reduce faculty cognitive burden, and enhance the reliability of standard setting while maintaining necessary human oversight and organizational accountability in high-stakes assessment. Future research should examine whether more advanced applications of LLMs support progression to the modification or redefinition levels of the SAMR model.

Limitations

Several limitations should be acknowledged. This was a single-institution study with one 120-item fundamental sciences exam, 10 faculty judges, and 5 selected LLMs, which may not reflect the broader and quickly changing environment of LLMs, limiting generalizability. Each LLM was evaluated in a single standardized run; therefore, within-model stochastic variability and reproducibility across multiple sessions were not assessed. As a result, the apparent consistency across the analyzed LLMs may be due to commonalities in their underlying models rather than independent expert-like assessments. Standardized prompting protocols do not guarantee consistent results across different LLMs or future model versions. Another limitation is that both faculty judges and LLMs were given the same historical cohort GPA data before estimating item performance. While this reflects recognized Angoff practice, the common performance anchor may have contributed to the convergence of faculty and LLM estimations, and future research should compare LLM performance with and without such contextual information. Although all judges were given standardized Angoff instructions, variances in their core disciplinary competence may have contributed to some interrater variation. Finally, this study focused only on type-A MCQs, so findings may not apply to other assessment formats, such as objective structured clinical examinations. Future research should seek to address these limitations by examining reproducibility across multiple institutions, disciplines, assessment formats, and standard setting methods. Longitudinal studies that monitor the influence of AI-assisted standard setting on student achievements would be extremely beneficial.

Conclusions

In this single-institution study, LLM-generated Angoff estimates were broadly comparable to faculty-generated estimates, with better generalizability coefficients and more consistency across the tested outputs under uniform prompting settings. These findings lend support to the prospective use of LLMs as adjuncts to faculty in standard setting. However, they do not show that LLMs are equivalent to or superior to expert judges, nor do they support their independent use in high-stakes assessments. However, more multicenter studies with multiple assessment formats, repeated model evaluations, and a broader range of

XSL•FO
RenderX

LLMs are required to determine the generalizability and reproducibility of these findings.

Acknowledgments

The authors gratefully acknowledge the Dubai Medical University faculty who participated as expert Angoff judges. They thank Professor Naglaa Raafat, Professor Preetha Shetty, Professor Nevein Hendawy, Professor Nadiah Mahmoud El Rouby, Dr Rania Hamed, Dr Shahid Ullah Khan, Dr Dania Abdelhady, Dr Hajer Sheikh, Dr Sumbal Riaz, and Dr Mohamed Hussein for their time, expertise, and valuable contributions. The authors used generative AI (ChatGPT [OpenAI], Claude [Anthropic], Grok [SpaceXAI LLC], DeepSeek, and MedGemma [Google DeepMind]) tools to assist with editing the language and improving the clarity and readability of the manuscript. All AI-assisted outputs were critically reviewed and revised by the authors, who take full responsibility for the accuracy, integrity, and final content of the manuscript. No generative AI tools were used to create or analyze data, interpret results, or draw scientific conclusions.

Funding

The subscription to the large language models was provided by Dubai Medical University.

Data Availability

The data from this study are available on reasonable request from the corresponding author.

Authors' Contributions

Conceptualization: HH (lead), SEK (supporting)
Data curation: MS (lead), SK (supporting), AZ (supporting)
Formal analysis: SEK (lead), SK (supporting)
Investigation: AZ (lead), SK (supporting)
Methodology: SEK (lead), HH (supporting)
Software: SK (lead), MS (supporting)
Writing—original draft: SEK (lead), HH (supporting),
Writing—review and editing: SEK (lead), HH (supporting), AZ (supporting)
All authors reviewed and approved the final version of the manuscript and agreed to be accountable for all aspects of the work.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Prompts used for the large language models in this study.
[PDF File, 185 KB - mededu_v12i1e98766_app1.pdf]

References

1. Norcini JJ. Setting standards on educational tests. Med Educ 2003 May;37(5):464-469. [doi: 10.1046/j.1365-2923.2003.01495.x] [Medline: 12709190]
2. Norcini JJ, Shea JA. The credibility and comparability of standards. Appl Meas Educ 1997;10(1):39-59. [doi: 10.1207/s15324818ame1001_3]
3. Norcini J, Shea J. The reproducibility of standards over groups and occasions. Appl Meas Educ 1992;5(1):63-72. [doi: 10.1207/s15324818ame0501_5]
4. Yakin M, Linden K. Adaptive e-learning platforms can improve student performance and engagement in dental education. J Dent Educ 2021 Jul;85(7):1309-1315. [doi: 10.1002/jdd.12609] [Medline: 33825184]
5. George S, Haque MS, Oyebode F. Standard setting: comparison of two methods. BMC Med Educ 2006 Sep 14;6:46. [doi: 10.1186/1472-6920-6-46] [Medline: 16972990]
6. Shulruf B, Wilkinson T, Weller J, Jones P, Poole P. Insights into the Angoff method: results from a simulation study. BMC Med Educ 2016 May 4;16:134. [doi: 10.1186/s12909-016-0656-7] [Medline: 27142788]
7. Jalili M, Hejri SM, Norcini JJ. Comparison of two methods of standard setting: the performance of the three-level Angoff method. Med Educ 2011 Dec;45(12):1199-1208. [doi: 10.1111/j.1365-2923.2011.04073.x] [Medline: 22122428]
8. Downing SM, Lieska NG, Raible MD. Establishing passing standards for classroom achievement tests in medical education: a comparative study of four methods. Acad Med 2003 Oct;78(10 Suppl):S85-S87. [doi: 10.1097/00001888-200310001-00027] [Medline: 14557105]
9. Sridharan K, Sivaramakrishnan G. Angoff methods in standard setting in health professional education: a systematic review and meta-analysis. BMC Med Educ 2025 Dec 29;25(1):1727. [doi: 10.1186/s12909-025-08300-6] [Medline: 41466255]

XSL•FO
RenderX

10. Norcini JJ, Shea JA, Kanya DT. The effect of various factors on standard setting. J Educ Meas 1988 Mar;25(1):57-65. [doi: 10.1111/j.1745-3984.1988.tb00291.x]
11. Ricker KL. Setting cut-scores: a critical review of the Angoff and modified Angoff methods. Alta J Educ Res 2006;52(1). [doi: 10.55016/ojs/ajer.v52i1.55111]
12. Baron P, Sireci SG, Slater SC. Evaluating panelists' understanding of standard setting data. Educ Meas 2021;40(2):16-25. [doi: 10.1111/emip.12399]
13. Gérard AO, Merino D, Labriffe M, et al. Evaluating and leveraging large language models in clinical pharmacology and therapeutics assessment: from exam takers to exam shapers. Br J Clin Pharmacol 2026 Aug;92(8):2574-2583. [doi: 10.1002/bcp.70137] [Medline: 40495266]
14. Sridharan K, Sequeira RP. Artificial intelligence and medical education: application in classroom instruction and student assessment using a pharmacology & therapeutics case study. BMC Med Educ 2024 Apr 22;24(1):431. [doi: 10.1186/s12909-024-05365-7] [Medline: 38649959]
15. Masters K, MacNeil H, Benjamin J, et al. Artificial intelligence in health professions education assessment: AMEE guide no. 178. Med Teach 2025 Sep;47(9):1410-1424. [doi: 10.1080/0142159X.2024.2445037] [Medline: 39787028]
16. Hamilton ER, Rosenberg JM, Akcaoglu M. The Substitution Augmentation Modification Redefinition (SAMR) model: a critical review and suggestions for its use. TechTrends 2016 Sep;60:433-441. [doi: 10.1007/s11528-016-0091-y]
17. Suliman S, Kassab SE, Iqbal MZ, et al. The role of artificial intelligence in co-creation of health professions education: integrating innovation into collaboration: AMEE guide no. 190. Med Teach 2026 Jul;48(7):1101-1112. [doi: 10.1080/0142159X.2025.2581163] [Medline: 41241925]
18. Verhoeven BH, van der Steeg AF, Scherpbier AJ, Muijtjens AM, Verwijnen GM, van der Vleuten CP. Reliability and credibility of an Angoff standard setting procedure in progress testing using recent graduates as judges. Med Educ 1999 Nov;33(11):832-837. [doi: 10.1046/j.1365-2923.1999.00487.x] [Medline: 10583792]
19. Cohen-Schotanus J, van der Vleuten CP. A standard setting method with the best performing students as point of reference: practical and affordable. Med Teach 2010;32(2):154-160. [doi: 10.3109/01421590903196979] [Medline: 20163232]
20. Champlain AF. Standard setting methods in medical education: high‐stakes assessment. In: Swanwick T, Forrest K, O'Brien BC, editors. Understanding Medical Education: Evidence, Theory, and Practice, 3rd edition: John Wiley & Sons; 2018. [doi: 10.1002/9781119373780.ch24]
21. Hays R, Gupta TS, Veitch J. The practical value of the standard error of measurement in borderline pass/fail decisions. Med Educ 2008 Aug;42(8):810-815. [doi: 10.1111/j.1365-2923.2008.03103.x] [Medline: 18564094]
22. Briesch AM, Swaminathan H, Welsh M, Chafouleas SM. Generalizability theory: a practical guide to study design, implementation, and interpretation. J Sch Psychol 2014 Feb;52(1):13-35. [doi: 10.1016/j.jsp.2013.11.008] [Medline: 24495492]
23. Verhoeven BH, Verwijnen GM, Muijtjens AM, Scherpbier AJ, van der Vleuten CP. Panel expertise for an Angoff standard setting procedure in progress testing: item writers compared to recently graduated students. Med Educ 2002 Sep;36(9):860-867. [doi: 10.1046/j.1365-2923.2002.01301.x] [Medline: 12354249]
24. Angoff WH. Scales, Norms, and Equivalent Scores: Educational Testing Service; 1984.
25. Kane M. Validating the performance standards associated with passing scores. Rev Educ Res 1994;64(3):425-461. [doi: 10.3102/00346543064003425]
26. Luckin R, Holmes W, Griffiths M, Forcier LB. Intelligence Unleashed. An Argument for AI in Education: Pearson Education; 2016.

Abbreviations

GPA: grade point average
LLM: large language model
MCC: minimally competent candidate
MCQ: multiple-choice question
RMSE: root mean squared error
SAMR: substitution, augmentation, modification, and redefinition

Edited by J Gentges; submitted 18.Apr.2026; peer-reviewed by A Jaleel, L Nguyen, X Shen; revised version received 04.Aug.2026; accepted 24.Aug.2026; published 09.Sep.2026.

Please cite as:
Kassab SE, Shadan M, Ziganshina A, Khanday S, Hamdy H
AI-Assisted Angoff Standard Setting for Multiple-Choice Examinations in Medical Education: Comparative Study
JMIR Med Educ 2026;12:e98766
URL: https://mededu.jmir.org/2026/1/e98766
doi:10.2196/98766

XSL•FO
RenderX

© Salah Eldin Kassab, Mariam Shadan, Arina Ziganshina, Shifan Khanday, Hossam Hamdy. Originally published in JMIR Medical Education (https://mededu.jmir.org), 9.Sep.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Education, is properly cited. The complete bibliographic information, a link to the original publication on https://mededu.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Performance of Cloud-Hosted Large Vision-Language Models on the Japanese National Examination for Clinical Laboratory Technicians: Comparative Benchmarking Study

Kota Murakami[1,2,3*], MMSc; Ryosuke Matsuzawa[2,3,4*], MMSc; Yuya Tahara-Arai[2,3,5*], BSc; Haruka Ozaki[2,3], PhD

[1]Master's Program in Medical Sciences, Graduate School of Comprehensive Human Sciences, University of Tsukuba, Ibaraki, Japan

[2]Bioinformatics Laboratory, Institute of Medicine, University of Tsukuba, Ibaraki, Japan

[3]Laboratory for AI Biology, RIKEN Center for Biosystems Dynamics Research, 6-7-1 Minatojima Minamimachi, Chuo-ku, Kobe, Hyogo, Japan

[4]Doctoral Program in Medical Sciences, Graduate School of Comprehensive Human Sciences, University of Tsukuba, Ibaraki, Japan

[5]Ph.D. Program in Humanics, School of Integrative and Global Majors, University of Tsukuba, Ibaraki, Japan

[*]these authors contributed equally

Corresponding Author:
Haruka Ozaki, PhD
Bioinformatics Laboratory, Institute of Medicine, University of Tsukuba, Ibaraki, Japan

Abstract

Background: Cloud-hosted large vision-language models (LVLMs) often outperform open-weight models on multimodal benchmarks, but their applicability to health care examinations that require both text and image reasoning remains unclear. Existing studies on Japan's National Examination for Clinical Laboratory Technicians mainly used text-only settings and a limited set of models.

Objective: This study aimed to evaluate the accuracy and applicability of cloud-hosted LVLMs on this examination, quantify the contribution of images, examine alignment with human performance across medical subfields, and test generalizability on questions administered after the models' knowledge cutoff dates. We aim to provide a standardized benchmark of baseline multimodal performance under deployment-realistic conditions without additional fine-tuning.

Methods: For this comparative benchmarking study, we built the Kensagishi question-answer dataset (Kensagishi QA), a benchmark of 800 questions from the examination (2020 - 2023), including text-only and image-based questions, in a standardized structured JSON format, and strict scoring criteria (all correct options required). We tested GPT-5, GPT-4o, GPT-4o-mini, Gemini 1.5 Pro, Gemini 2.5 Pro, and Neva-22B using a uniform "numbers-only" response prompt; images were provided as Base64-encoded data when available. Accuracy was computed overall, by item type, and by medical subfield. Human reference was drawn from published examination results. Spearman correlation assessed model-human alignment. Generalization was evaluated on 200 questions from the 71st examination (2025), released after the models' knowledge cutoffs.

Results: Top-performing models exceeded the 60% passing threshold. GPT-5 achieved 93.9% (95% CI 92.0% - 95.4%) accuracy with images and 88.2% (95% CI 85.7% - 90.2%) without images, Gemini 2.5 Pro 92.5% (95% CI 90.5% - 94.1%) and 86.8% (95% CI 84.2% - 88.9%), and GPT-4o 81.7% (95% CI 78.8% - 84.2%) and 79.8% (95% CI 76.8% - 82.4%), respectively. Providing images markedly improved accuracy for GPT-5, Gemini 2.5 Pro, and GPT-4o, whereas Neva-22B showed no improvement. Model-human rank correlations by medical subfield were moderate for GPT-4o (Spearman ρ=0.61; 95% CI –0.03 to 0.90; *P*=.06 with images) and statistically significant for Gemini 1.5 Pro without images (ρ=0.67; 95% CI 0.07 - 0.91; *P*=.03). On 200 postknowledge cutoff questions from the 71st examination (2025), model accuracies were comparable to historical averages (eg, GPT-5 91.5%, 95% CI 86.8% - 94.6%; Gemini 2.5 Pro 90.7%, 95% CI 85.8% - 94.0%; GPT-4o 82.0%, 95% CI 76.1% - 86.7%), indicating no material performance degradation on unseen questions.

Conclusions: Cloud-hosted LVLMs show strong performance on this examination, and visual input is a key driver of accuracy for leading models. Models that better mirror human difficulty patterns (eg, GPT-4o) offer complementary insights for educational analytics, whereas the highest-accuracy models (eg, GPT-5 and Gemini 2.5 Pro) may be preferable when correctness is paramount. To our knowledge, this is the first systematic, multimodel evaluation isolating visual input that extends prior text-only, few-model studies and provides Kensagishi QA to guide educational deployment of medical AI.

(JMIR Med Educ 2026;12:e84266) doi:10.2196/84266

XSL•FO
RenderX

KEYWORDS

large vision-language models; Japanese National Examination for Clinical Laboratory Technicians; multimodal reasoning; medical AI; model performance evaluation

Introduction

Large language models (LLMs) are a type of natural language processing technology that acquire the ability to understand and generate human language by learning from massive amounts of textual data. These models have demonstrated high performance across a wide range of tasks, including question answering, summarization, translation, and reasoning, with rapidly expanding applications [1-4]. In recent years, large vision-language models (LVLMs), which are capable of handling not only text but also multimodal information including images, have also emerged. These models are designed to understand and generate both visual and textual information simultaneously, enabling their application to a broader range of tasks such as image-based question answering, image captioning, and visual reasoning [2,4]. Notably, recent multimodal benchmarks (eg, VisionArena [5] and MMStar [6]) indicate that proprietary, cloud-hosted, closed-weight LVLMs currently outperform open-weight, self-hosted models.

The rapid advancement of LLMs has also extended into the medical domain, where evaluations of performance on the United States Medical Licensing Examination (USMLE) have become increasingly common. Prior studies using benchmark datasets such as MedQA, which contains USMLE-style questions, have demonstrated that LLMs can answer questions requiring domain-specific knowledge [7]. For instance, Med-PaLM 2, developed by Google, achieved an accuracy of 86.5% on the MedQA dataset and received favorable evaluations compared with human physicians' responses [8]. In addition, the LLM-MedQA system using the Llama3.1:70B model reported approximately 7% improvements in accuracy and F_1-scores under zero-shot settings by incorporating a multiagent architecture and similar case-generation techniques [9]. Furthermore, with the integration of Semantic Clinical AI (SCAI), which embeds structured clinical knowledge into LLMs, performance has significantly improved, achieving up to 95.1% accuracy on USMLE Step 3, surpassing conventional models by a considerable margin [10]. In a similar effort in the Japanese context, a benchmark dataset named IgakuQA, based on questions from the Japanese National Medical Licensing Examination, was developed to explore the applicability of LLMs while accounting for linguistic and health care system differences [11]. Moreover, LLMs have also been evaluated in supporting the creation of multiple-choice questions for the Japanese National Nursing Examination, particularly in generating and assessing distractors [12].

In contrast, although text-only LLMs have been actively evaluated on medical licensing examinations, the applicability of LVLMs to health care–related national examinations—particularly those requiring both textual and visual reasoning—remains underexplored. For the Japanese National Examination for Clinical Laboratory Technicians, existing reports have primarily assessed a limited number of general-purpose LLMs, such as ChatGPT-3.5 and GPT-4 [13-15], often in text-only settings. As a result, it is still unclear how modern, cloud-hosted (closed-weight) LVLMs perform on this examination and to what extent visual input contributes to accuracy.

This study identifies 2 primary gaps. First, there is no publicly available, well-structured dataset specifically tailored to the National Examination for Clinical Laboratory Technicians (including paired image assets and metadata). Consequently, considerable effort is required to conduct reproducible cross-model comparisons or assess newly developed systems. Second, prior work has not provided a systematic, multimodel evaluation of LVLMs that isolates the contribution of images; although some attempts have included image-dependent items (eg, peripheral blood smear morphology classification and electrophoresis band pattern interpretation), coverage has been limited and insufficient for comprehensive multimodal comparison. Taken together, these gaps highlight the need for a standardized, reproducible benchmark that enables fair cross-model comparison under controlled conditions.

Given this background, the aim of the present study was to conduct a multifaceted evaluation of the applicability of cloud-hosted LVLMs to the medical domain, using the National Examination for Clinical Laboratory Technicians as a case study. In particular, we focused on the impact of visual information, in addition to textual information, on model accuracy. To this end, we constructed a custom benchmark dataset, Kensagishi question-answer (Kensagishi QA), based on past questions from the Japanese National Examination for Clinical Laboratory Technicians. By systematically comparing the performance of multiple cloud-hosted LVLMs, we sought to elucidate their strengths and limitations and to provide fundamental insights to support future advancements in medical AI.

Methods

Construction of a Dataset Based on the Japanese National Examination for Clinical Laboratory Technicians

This study was designed as a comparative benchmarking study to develop an examination benchmark and evaluate the performance of multiple cloud-hosted LVLMs. To support this evaluation, we constructed a novel benchmark dataset, Kensagishi QA, comprising questions from the 66th to 69th National Examination for Clinical Laboratory Technicians conducted between 2020 and 2023. The official examination questions for each year are publicly available in PDF format [16-19]. These materials were obtained from the official website of the Ministry of Health, Labour and Welfare (MHLW), where the content is provided under the Public Data License (version 1.0; PDL1.0), unless otherwise noted. Redistribution of the structured question data in our GitHub repository complied with the applicable license terms, including appropriate attribution.

XSL•FO
RenderX

From these PDF files, we extracted the question and answer choices and converted them into a machine-readable JSON format (Figure 1). During the conversion process, the data were structured to include fields such as question ID, question text, answer choices (1-5), correct answer, and question type (text/image).

Figure 1. Example of a structured dataset entry from the Japanese National Examination for Clinical Laboratory Technicians. The question and choices were translated from the original Japanese text for illustrative purposes.

Example of a structured exam problem entry


```
{
  "problem_id": "66AM1",
  "text_only": true,
  "problem_text": "Which of the following is used to measure urinary osmolality
in a clinical laboratory?",
  "choices": ["Refractometry", "Freezing point depression", "Glass electrode
method", "Cation extraction method", "Metachromasia method"],
  "category": "General Clinical Laboratory Science",
  "answer": 2
}
```


The National Examination for Clinical Laboratory Technicians includes a mix of text-only questions (hereafter referred to as text questions) and questions containing medical images or diagrams (hereafter referred to as image questions). In the constructed dataset, these questions were categorized by type, allowing for separate performance evaluations of each question format.

All questions follow a 5-choice format and consist of either single-answer or multiple-answer types, where a single question may have more than one correct option. Each year, 200 questions are administered—100 in the morning and 100 in the afternoon—resulting in a total of 800 questions collected from consecutive examination years (2020 - 2023). Of the 800 questions, 670 (83.75%) were classified as text questions, whereas 130 (16.25%) were identified as image questions. The image questions include prompts that require interpretation of visual information, such as blood smears, histopathological images, device output screens, and tabular data, thereby necessitating visual input processing.

Evaluation of Accuracy of LVLMs on the Japanese National Examination for Clinical Laboratory Technicians

In this study, we evaluated the accuracy of multiple LVLMs on questions from the National Examination for Clinical Laboratory Technicians. To specifically assess performance on questions involving images, we selected LVLMs capable of processing both textual and visual inputs.

The selection of these 6 LVLMs was intended to represent a range of model generations, providers, and practical deployment scenarios rather than to compare only the most recent state-of-the-art systems. In particular, we included multiple models released by the same provider (GPT-4o-mini, GPT-4o, and GPT-5) to examine performance differences across model generations and to assess whether earlier, lower-cost models remain viable for large-scale or educational use, in addition to evaluating the capabilities of the latest model. This reflects realistic constraints in applied settings, where the most recent model is not always adopted due to factors such as inference cost, availability, latency, and system compatibility. Furthermore, models from different providers (OpenAI, Google, and NVIDIA) were included to reduce vendor-specific bias and examine how differences in architecture and training affect multimodal performance on the same benchmark. We note that all evaluated models were accessed via cloud APIs in this study, and no local deployment or on-premises inference was performed.

An overview of the models used in this study is presented in Table 1. For Gemini 2.5 Pro, we used the preview version released on March 25, 2025 (gemini-2.5-pro-preview-03-25), prior to its general availability.

XSL•FO
RenderX

Table . Comparison of providers, API names, parameter sizes, release dates, and knowledge cutoffs for each large vision-language model.

Model	Provider	API name	Number of parameters	Release date	Knowledge cutoff
GPT-4o-mini	OpenAI	gpt-4o-mini-2024-07-18	Not disclosed	July 18, 2024	October 2023
GPT-4o	OpenAI	gpt-4o-2024-08-06	Not disclosed	August 6, 2024	October 2023
Gemini 1.5 Pro	Google	gemini-1.5-pro	Not disclosed	September 24, 2024	May 2024
Gemini 2.5 Pro (Preview)	Google	gemini-2.5-pro-preview-03-25	Not disclosed	March 25, 2025 (Preview)	January 2025
Neva-22B	NVIDIA	nvidia/neva-22b	22 billion	Not disclosed	Not disclosed
GPT-5	OpenAI	gpt-5-2025-08-07	Not disclosed	August 7, 2025	September 30, 2024

The National Examination for Clinical Laboratory Technicians includes multiple-answer questions in which more than one option may be correct. To accommodate this format, we adopted a strict evaluation criterion: a response was considered correct only if all correct answer choices were selected. Partial matches were regarded as incorrect.

In addition, some examination questions include images but can still be answered correctly based solely on the answer choices or surrounding textual context. However, from the perspective of evaluating AI model performance, it is essential to assess the impact of image presence on accuracy appropriately. Therefore, we clearly distinguished between text-based and image-based questions and calculated the accuracy for each category separately.

For image questions, the image file corresponding to each question ID was encoded in Base64 and provided to the model along with the associated question text. This enabled multimodal input handling for questions that include visual content. Instructions to each model were given based on a standardized prompt format. Specifically, the models were presented with the question text and answer choices, followed by an explicit instruction to “respond with only the number(s) of the correct choice(s).” The prompt content is as follows:

> *Please provide the number(s) of the correct option(s) for the question below in the following format. If the question requires selecting only one answer, provide a single number. If it requires selecting two answers (e.g., "Choose two"), provide two numbers. Do not include any explanations or reasoning—just the number(s).*
>
> *Question: {Question text}*
>
> *Options: {List of options}*
>
> *Please provide only the number(s) of the correct answer(s):*

With this prompt format, the models were instructed to output only the option numbers (eg, “2” or “14”). The output was parsed using regular expressions (re.findall(r'\d+', text)) to extract numerical values, which were interpreted as the selected choice numbers. This approach was designed to robustly handle variations in the response format, such as “1 4” or “1,4.”

For all questions, each model received the question text and answer choices as textual input, and for image questions, the corresponding visual data were also provided. The images were encoded in Base64 format and provided alongside the text.

The images used in this study were screenshots extracted from the original examination materials. No image preprocessing such as resizing, cropping, or normalization was applied. Each image was used in its original resolution and directly encoded into Base64 format before being provided to the models.

The model outputs were recorded in JSONL format as pairs of question IDs and predicted choice numbers. The correctness of each prediction was determined based on the strict evaluation criterion described earlier.

Final accuracy was calculated as the proportion of correctly answered questions relative to the total number of questions. To analyze the impact of visual information on model performance, accuracy was computed separately for text questions and image questions. As a reference, we used the passing threshold for the national examination (60% accuracy) and assessed whether each model exceeded this criterion, thereby enabling comparative analysis across models.

To reduce the impact of stochastic variation inherent in the generation process, each model independently answered all questions 3 times under identical settings. The accuracy of each run was first computed, and the mean across the 3 runs was used as the model’s final accuracy. Accuracy figures display the 3 per-run values as points and 95% Wilson score CIs as error bars.

Categorization of Questions by Medical Subfield

In this study, question categorization by medical subfield was conducted based on the official examination guidelines for the Japanese National Examination for Clinical Laboratory Technicians published by the MHLW. Specifically, the question stems and answer choices were carefully examined to interpret the medical theme and underlying intent of each question. Based on this analysis, we manually classified each question into relevant subfields, such as clinical physiology, clinical chemistry, and pathological histocytology. The categorization was performed by a single author (KM); no independent reannotation or interannotator agreement assessment was conducted, which is acknowledged as a study limitation. Using this categorization, accuracy was calculated for each subfield.

XSL•FO
RenderX

Source of Human Accuracy Data

The human accuracy rates used in this study were based on published results from a performance analysis of the national examination [20]. Specifically, we extracted the average accuracy for each question from the previous study [20], titled "Item-by-item Accuracy Analysis," which presents data from the 68th National Examination for Clinical Laboratory Technicians based on responses from 1896 examinees at 40 institutions.

Correlation Analysis Between Model and Human Accuracy

To examine the correlation between model outputs and human accuracy, we calculated the Spearman rank correlation coefficient. For each medical subfield, we computed the accuracy rates for both humans and models, ranked them in descending order, and then merged the results based on the corresponding subfield names to obtain the rank difference d_i for each subfield. The Spearman rank correlation coefficient, ρ, is defined by the following formula:

$$\rho=1-6\sum idi2n(n2-1)$$

where d_i is the difference in rank between the human and model accuracies for subfield i, and n is the total number of subfields. This analysis quantitatively evaluates the extent to which the model's accuracy patterns across subfields align with those observed in human performance. All P values are reported as exact values and formatted according to established medical journal guidelines, using an italicized capital P without a leading zero; values less than .001 are reported as P<.001. All statistical analyses were performed in Python (version 3.10.12; Python Software Foundation) using SciPy (scipy.stats, version 1.15.3; SciPy community); reported P values are two-sided, and statistical significance was defined as P<.05.

Statistical Comparison Between Training Years and the 2025 Examination

In addition to descriptive comparisons, we conducted a statistical significance test to examine whether the declines in accuracy between the training years (2020‐2023) and the 2025 examination were statistically meaningful. For each model and image condition, we used the mean accuracy across the 3 independent runs as the final accuracy (see above) and constructed a 2×2 contingency table of correct versus incorrect responses using question-level counts (n=800 for 2020‐2023 and n=200 for 2025). Because the 3 runs are repeated measurements on the same questions rather than independent samples, the denominator was set to the number of questions rather than the number of question-run pairs. A chi-square test of independence (Pearson, without continuity correction) was then applied to each comparison.

Specifications of the Tested LVLMs

Table 1 summarizes the specifications of the LVLMs used in this study. For each model, we document the provider, API name used, number of parameters, release date, and the final knowledge cutoff date of the training data. All inference was performed via the providers' cloud APIs between July 2024 and October 2025; no local or on-premises graphics processing unit (GPU) compute was used, and total inference cost was modest given the limited number of API calls (approximately 800 questions×6 models×2 image conditions). Table 2 summarizes the generation settings used for each model.

Table . Generation settings used for each model in this study.

Model	Provider	Temperature	Top-p	Top-k	Number of candidates (n)
GPT-4o-mini	OpenAI	0.0	1.0	Not configurable	1
GPT-4o	OpenAI	0.0	1.0	Not configurable	1
Gemini 1.5 Pro	Google	0.0	0.95	64 (fixed)	1
Gemini 2.5 Pro (Preview)	Google	0.0	0.95	64 (fixed)	1
Neva-22B	NVIDIA	0.0	0.7	Not configurable	1
GPT-5	OpenAI	1.0	1.0	Not configurable	1

Calculation of CIs

For every point estimate, we reported a two-sided 95% CI. For accuracy (a binomial proportion), we used the Wilson score interval [21], with the mean number of correct answers across the 3 runs as the numerator and the number of questions as the denominator (not the number of question-run pairs, because the runs were repeated measurements of the same items). The Wilson CIs reported here convey the uncertainty of each model's accuracy estimate; interval overlap should not be interpreted as a formal between-model comparison, and no correction for multiple comparisons was applied. For the Spearman correlation between model and human subfield accuracy, we used the Fisher z transformation with the number of subfields as the sample size (n=10). For differences in accuracy between the 2025 and the 2020‐2023 examinations and visual-heavy versus text-heavy subfields, we used the Newcombe hybrid score method [22]. Because the benchmark enumerated all questions from the target examinations, these CIs express the precision of each estimate when the fixed question set is treated as a sample of underlying model competence rather than as sampling error from a larger population.

Data Availability

The data used in this study are publicly available in a GitHub repository [23]. The repository contains past question data from the National Examination for Clinical Laboratory Technicians in JSONL format.

XSL•FO
RenderX

Evaluation of Model Generalization Using Questions After the Knowledge Cutoff

To reduce the possibility that the models derived correct answers solely from memorized training data, we conducted an additional performance evaluation using questions published after each model's knowledge cutoff. Specifically, we used all 200 questions (100 in the morning and 100 in the afternoon) from the 71st National Examination for Clinical Laboratory Technicians, administered in 2025 [24].

In this evaluation, the models were prompted using the same format as in the previous experiments. The generated responses were compared against the correct answers. As a reference, we calculated the average accuracy over the 4 earlier years (2020 - 2023) and denoted this value as the "historical average." The deviation between this historical average and the 2025 accuracy was then computed for each model. Through this evaluation, we examined each model's ability to generalize to previously unseen questions and assessed whether the models could answer them without relying on prior exposure to the question content.

Ethical Considerations

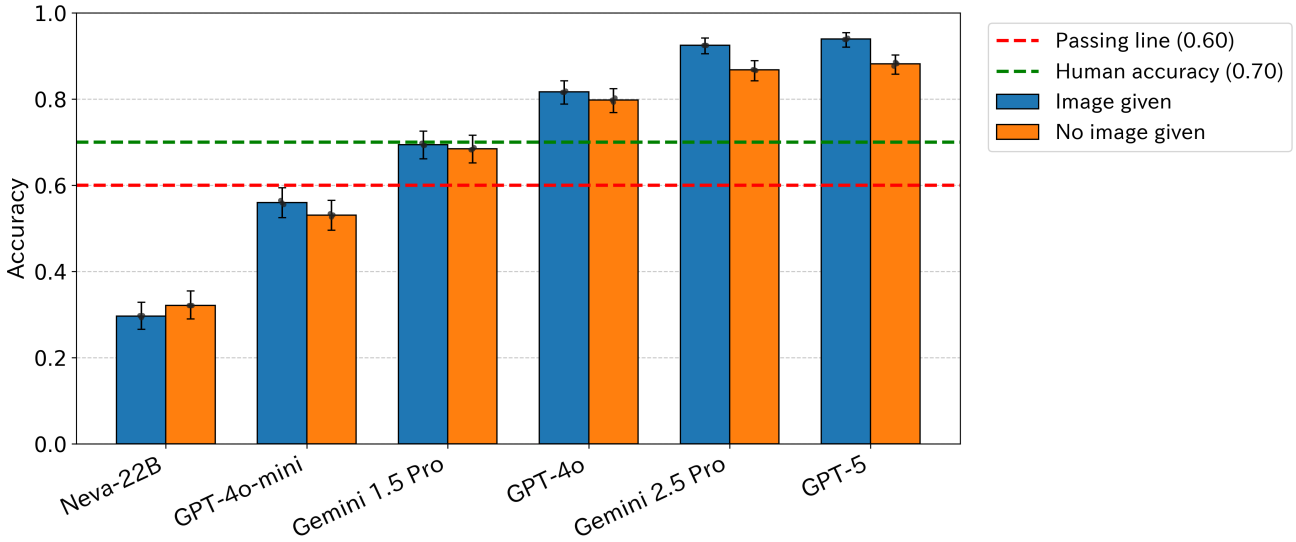
This study did not involve newly conducted research on human participants. We used only (1) publicly available questions from the Japanese National Examination for Clinical Laboratory Technicians and (2) aggregated item-level mean accuracy values previously published in a report [20]. The human accuracy data used in this study, namely the aggregated, item-level mean accuracy values previously published in that report, were group-level averages from 1896 examinees at 40 institutions and contained no information that could identify any individual. No individual-level data were collected or accessed, and no identifiable images are included in this manuscript or its supplementary materials. According to the Ethical Guidelines for Medical and Biological Research Involving Human Subjects, jointly issued by the Ministry of Education, Culture, Sports, Science and Technology, the MHLW, and the Ministry of Economy, Trade and Industry, the guidelines do not apply to research using only existing information that does not relate to an individual [25]. Therefore, ethics review and approval and informed consent were not required for this study.

Results

Overall Accuracy of LVLMs

Figure 2 illustrates the accuracy achieved by each LVLM on all 800 questions from the National Examination for Clinical Laboratory Technicians. In this study, for image-based questions, we evaluated model performance under 2 conditions: an image-provided (image given) condition and a no-image (no image given) condition. For questions without associated images, identical inputs were used for both conditions. As evaluation criteria, we adopted the human passing threshold of 60% accuracy as well as the average human accuracy of 69.7%.

Figure 2. Accuracy of large vision-language models on the National Examination for Clinical Laboratory Technicians. The bar chart shows accuracies (0% - 100%) on 800 questions (text- and image-based) from the 66th to 69th examinations. The red dashed line indicates the human passing threshold of 60%, and the green dashed line represents the average human accuracy of 69.7%. Points denote the 3 per-run accuracies; error bars are 95% Wilson score CIs.



The overall average accuracies showed that GPT-5 achieved 93.9% with image input and 88.2% without, both substantially exceeding the passing threshold. Gemini 2.5 Pro followed with 92.5% with images and 86.8% without, also well above the benchmark. GPT-4o surpassed the threshold as well, achieving 81.7% with images and 79.8% without. Gemini 1.5 Pro recorded 69.4% with images and 68.5% without, both above the 60% benchmark.

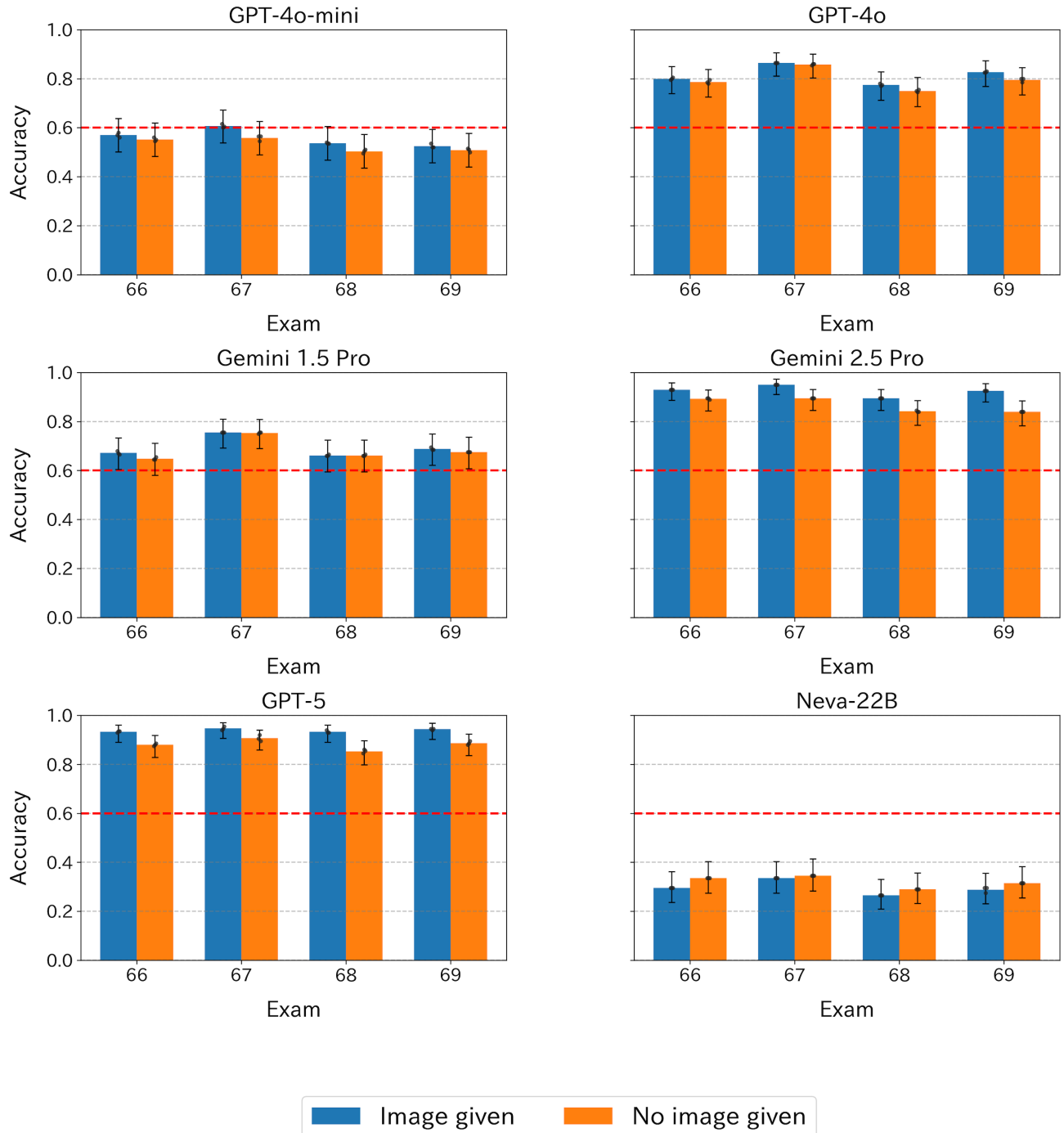
In contrast, GPT-4o-mini scored 56.0% with image input and 53.0% without, falling below the passing threshold under both conditions. Neva-22B showed the lowest accuracy among all models, with 29.6% (with images) and 32.1% (without images). When compared with the average human accuracy (69.7%), both Gemini 2.5 Pro and GPT-4o outperformed human examinees, regardless of image availability.

XSL•FO
RenderX

Effect of Image Input on Model Accuracy for Image-Based Questions

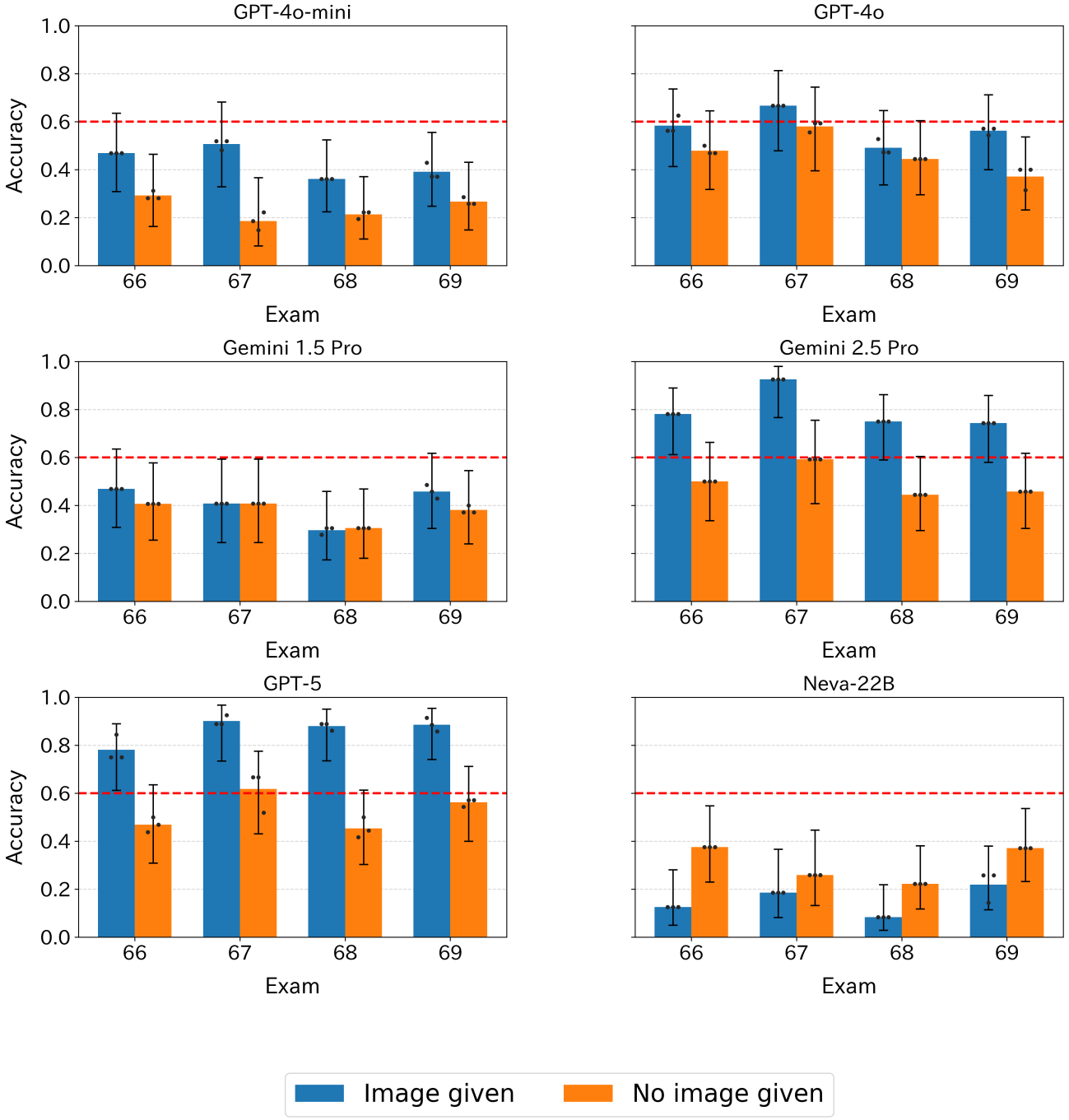
Figures 3 and 4 present a comparison of answer accuracies for LVLMs on all questions (text- and image-based) and on image-based questions, respectively, under 2 conditions: image-provided (image given) and no-image (no image given). In this section, we describe in detail how the presence of visual information influences model performance across different models and sessions.

Figure 3. Accuracy of GPT-5, GPT-4o, GPT-4o-mini, Neva-22B, Gemini 1.5 Pro, and Gemini 2.5 Pro on all questions (text- and image-based) from the 66th to 69th National Examinations for Clinical Laboratory Technicians. The x-axis represents the examination (66th to 69th), each pooling the morning and afternoon sessions, and the y-axis represents accuracy. Blue bars indicate the condition in which images were provided for image-based questions, while orange bars indicate the condition without image input. The red dashed line marks the passing threshold of 60%. This allows comparison of the effect of image input on overall accuracy across models and examination sessions. Points denote the 3 per-run accuracies; error bars are 95% Wilson score CIs.



XSL•FO
RenderX

Figure 4. Accuracy of GPT-5, GPT-4o, GPT-4o-mini, Neva-22B, Gemini 1.5 Pro, and Gemini 2.5 Pro on image-based questions from the 66th to 69th National Examinations for Clinical Laboratory Technicians. The x-axis indicates the examination (66th to 69th), pooling the morning and afternoon sessions, and the y-axis represents accuracy. Blue bars denote the image-provided condition (image given), and orange bars represent the no-image condition (no image given). The red dashed line indicates the passing threshold of 60%. This figure enables comparison not only of performance differences among models but also of the impact of image availability on accuracy for image-based questions. Points denote the 3 per-run accuracies; error bars are 95% Wilson score CIs.



On all questions, GPT-5 consistently outperformed the other models, achieving around 90.0% accuracy across all sessions, with the image-provided condition always surpassing the no-image condition. In particular, in the afternoon session of the 68th examination, GPT-5 achieved 97.0% with images and 84.7% without images, representing the highest performance among all models.

GPT-4o generally performed better with images, with accuracies ranging from 75.0% to 87.0% with images and 74.0% to 86.0% without images. Although some exceptions were observed, overall, image input improved its performance. GPT-4o-mini exhibited lower overall accuracy compared with other models, but in many sessions, performance was higher under the image-provided condition (46.3% - 60.7%) than under the no-image condition (45.3% - 57.3%).

Gemini 2.5 Pro also showed consistently higher accuracy under the image-provided condition in every session, such as 93.0% with images and 88.0% without images in the morning session of the 67th examination. Gemini 1.5 Pro showed accuracies ranging from 64.0% to 75.7% with images and 61.3% to 77.0%

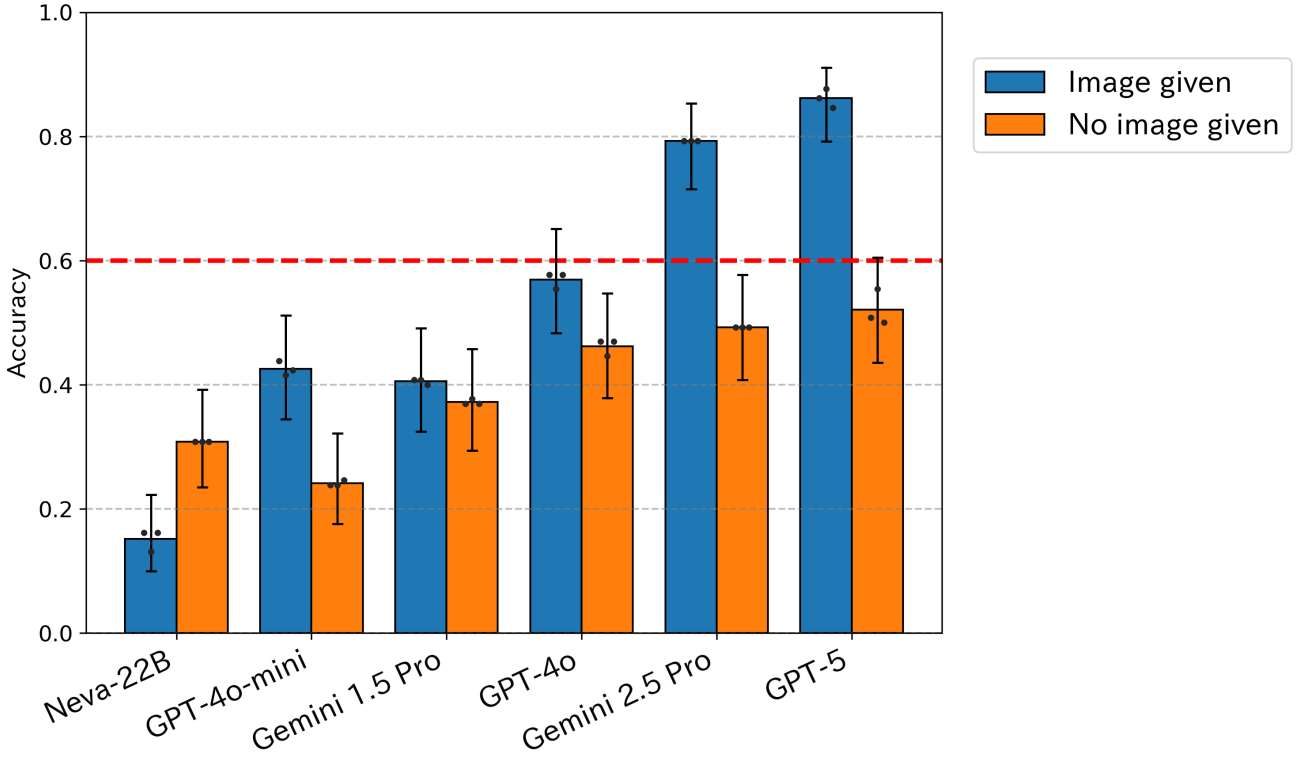
XSL•FO
RenderX

without. While in some sessions the no-image condition outperformed the with-image one, the overall differences were not substantial.

Neva-22B demonstrated the opposite trend, achieving accuracies of 25.0% - 37.0% with images and 28.0% - 39.0% without, indicating better performance without visual information. Moreover, frequent formatting errors were observed, such as blank responses or outputs that did not conform to the specified format.

In addition to the session-wise comparisons, we quantified the overall effect of image input using all image-based questions in Kensagishi QA (n=130), pooled across the 66th to 69th examinations (Figure 5). For each model, we compared accuracy under the image-provided condition (image given) against the no-image condition (no image given) and computed the absolute difference (ΔAcc=image given−no image given). Across the top-performing models, image input consistently improved accuracy: GPT-5 achieved 86.15% with images versus 52.05% without images (ΔAcc=34.10 points), and Gemini 2.5 Pro achieved 79.23% versus 49.23% (ΔAcc=30.00 points). GPT-4o also benefited from image input, with 56.92% versus 46.15% (ΔAcc=10.77 points). GPT-4o-mini improved from 24.10% to 42.56% (ΔAcc=18.46 points), while Gemini 1.5 Pro showed an increase from 37.18% to 40.51% (ΔAcc=3.33 points). In contrast, Neva-22B exhibited a negative image effect, decreasing from 30.77% (no image given) to 15.13% (image given; ΔAcc=–15.64 points). These results indicate that the impact of image input varies substantially across models.

Figure 5. Comparison of average accuracy on image-based questions (n=130) for 6 LVLMs under the image-provided (image given) and no-image (no image given) conditions, pooled across the 66th to 69th National Examinations for Clinical Laboratory Technicians. The x-axis indicates the model, and the y-axis represents accuracy. Blue bars denote the image-given condition, and orange bars represent the no-image-given condition. Image input improved accuracy for all models except Neva-22B. The red dashed line represents the passing threshold of 60%. Points denote the 3 per-run accuracies; error bars are 95% Wilson score CIs.



Modality-Specific Analysis of Image-Based Questions

To further investigate the sources of error in image-based questions, we categorized all 130 image-based questions into 5 visual modalities: (1) photographs (eg, histology images and instrument photographs; n=91), (2) graphs (n=21), (3) schematic diagrams (n=12), (4) tables (n=5), and (5) combined table+photograph items (n=1). As shown in Table 3, clear modality-dependent differences were observed. Photograph-based questions showed substantial performance gaps across models. For example, GPT-5 achieved 84.62% accuracy on photograph-based questions, Gemini 2.5 Pro 80.22%, and GPT-4o 59.34%, whereas Neva-22B achieved only 17.95%. In contrast, schematic diagrams yielded relatively high accuracy for top-performing models (Gemini 2.5 Pro: 91.67% and GPT-5: 97.22%), suggesting that structured visual abstractions were more robustly interpreted than naturalistic medical photographs. Graph-based questions were handled well by GPT-5 (82.54%), but showed lower accuracy for Gemini 2.5 Pro (61.90%) and substantial degradation in lower-performing models. Table-based questions demonstrated large intermodel variance, with Gemini 2.5 Pro and GPT-5 achieving 100.00% accuracy, and Gemini 1.5 Pro and Neva-22B achieving 20.00%.

XSL•FO
RenderX

Table . Accuracy by model and image modality, presented as percentages. For each model and image condition, accuracy is the mean across 3 independent runs (see Methods). The 130 image-based questions were grouped into 5 visual modalities; the n given in each column header is the number of image-based questions assigned to that modality, not a number of correct answers.

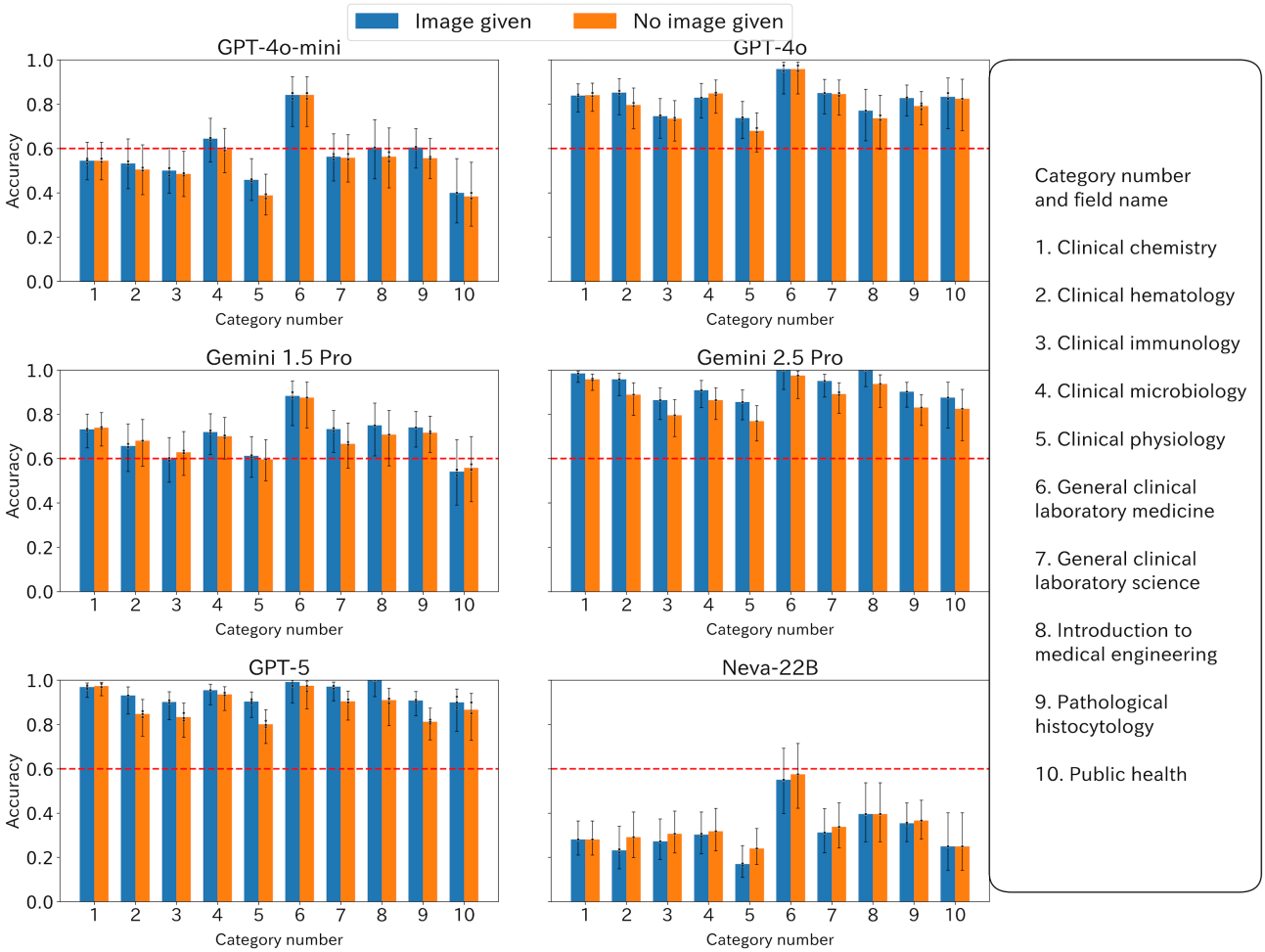
Model	Photograph (n=91), %	Graph (n=21), %	Diagram (n=12), %	Table (n=5), %	Table+photograph (n=1), %
GPT-4o-mini					
With image	47.62	23.81	41.67	40.00	0.00
Without image	27.11	15.87	27.78	0.00	0.00
GPT-4o					
With image	59.34	46.03	52.78	60.00	100.00
Without image	52.01	20.63	44.44	40.00	100.00
Gemini 1.5 Pro					
With image	44.69	38.10	25.00	20.00	0.00
Without image	43.96	20.63	16.67	40.00	0.00
Gemini 2.5 Pro					
With image	80.22	61.90	91.67	100.00	100.00
Without image	52.75	38.10	58.33	20.00	0.00
Neva-22B					
With image	17.95	3.17	13.89	20.00	0.00
Without image	35.16	19.05	25.00	20.00	0.00
GPT-5					
With image	84.62	82.54	97.22	100.00	100.00
Without image	56.04	46.03	30.56	46.67	100.00

Model Accuracy Across Medical Subfields

To assess how model performance varies across different areas of medical knowledge, we aggregated the accuracy of each LVLM by medical subfield (Figure 6). The 10 targeted subfields were clinical chemistry, clinical hematology, clinical immunology, clinical microbiology, clinical physiology, general clinical laboratory medicine, general clinical laboratory science, introduction to medical engineering, pathological histocytology, and public health.

XSL•FO
RenderX

Figure 6. Accuracy by medical subfield for each model—GPT-5, GPT-4o, GPT-4o-mini, Neva-22B, Gemini 1.5 Pro, and Gemini 2.5 Pro—on all questions (both text- and image-based) from the 66th to 69th National Examinations for Clinical Laboratory Technicians. The x-axis represents medical subfields, and the y-axis represents accuracy. Blue bars indicate accuracy under the image-given condition (where images were provided for all questions), while orange bars represent the no-image-given condition (no images provided, including for image-based questions). The red dashed line indicates the passing threshold of 60%. For each model, points denote the 3 per-run accuracies, and error bars are 95% Wilson score CIs.



GPT-5 achieved exceptionally high performance across all subfields, with accuracies ranging from approximately 90.0% to 100.0%. In the introduction to medical engineering, it reached 100.0%, while in general clinical laboratory medicine, it achieved 99.2% under the image-given condition and 97.5% under the no-image-given condition. Similarly, in clinical chemistry and clinical microbiology, GPT-5 maintained stable accuracy around 95.0% regardless of image availability. Although its performance declined slightly in the no-image-given condition for clinical physiology (80.1%) and pathological histocytology (81.3%), accuracy remained above 80.0% in all subfields.

Gemini 2.5 Pro demonstrated consistently high accuracy across most subfields, typically in the 80.0% - 90.0% range, and achieved 100.0% accuracy in both general clinical laboratory medicine and introduction to medical engineering. GPT-4o also showed robust performance, with accuracy generally in the 70.0% - 80.0% range and exceeding 90.0% in general clinical laboratory medicine.

In contrast, Gemini 1.5 Pro exhibited substantial variability across subfields. Although some sessions in general clinical laboratory medicine reached 100.0% accuracy, performance in clinical physiology and public health fell below 50.0% in several sessions. GPT-4o-mini and Neva-22B showed lower overall accuracy across subfields, with Neva-22B remaining below 50.0% accuracy in all subfields.

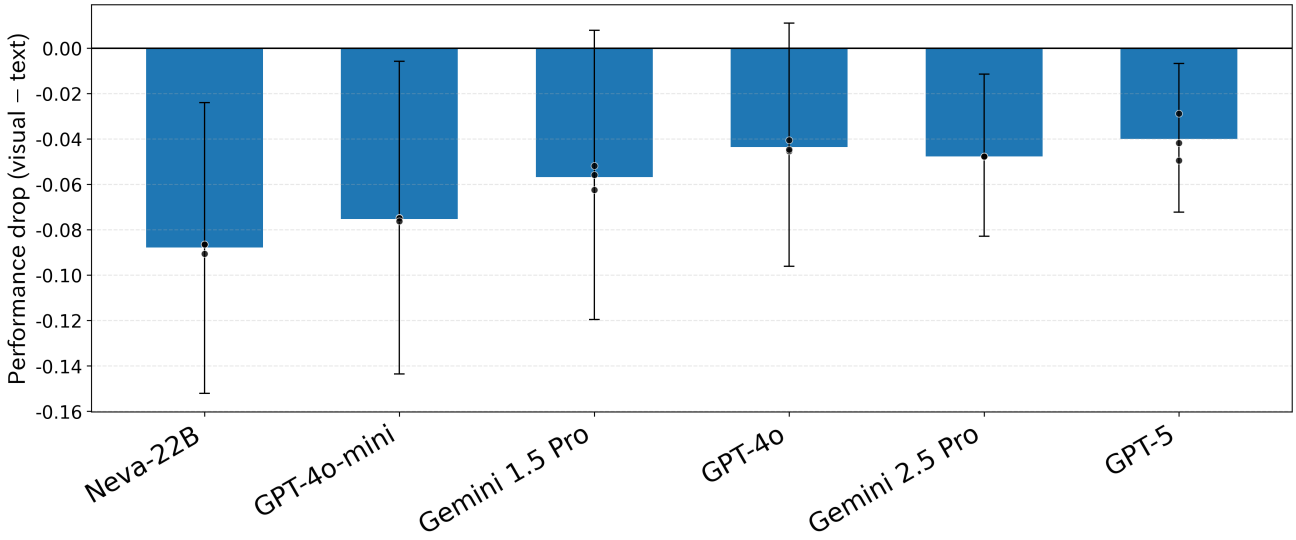
To contextualize differences in subfield-level performance, we quantified the proportion of image-based questions within each subfield and classified them into visual-heavy and text-heavy subfields. Clinical physiology (37/104, 35.6%), clinical hematology (24/72, 33.3%), pathological histocytology (23/112, 20.5%), clinical immunology (14/88, 15.9%), and general clinical laboratory science (12/80, 15.0%) were categorized as visual-heavy subfields. In contrast, clinical microbiology, introduction to medical engineering, public health, general clinical laboratory medicine, and clinical chemistry contained relatively few image-based questions and were categorized as text-heavy.

Grouping subfields into visual-heavy and text-heavy categories revealed a consistent trend across models (Figure 7). All models demonstrated lower accuracy in visual-heavy subfields compared with text-heavy subfields, as indicated by negative effect sizes (visual-heavy minus text-heavy). The magnitude of this performance gap varied by model capacity: GPT-5 and Gemini 2.5 Pro exhibited relatively small differences, indicating robustness to visually demanding content, whereas GPT-4o-mini

XSL•FO
RenderX

and Neva-22B showed substantially larger negative effect sizes. These findings suggest that visually intensive disciplines pose greater challenges for lower-capacity models and that higher-capacity models are more resilient to the increased cognitive and multimodal demands of visual-heavy medical subfields.

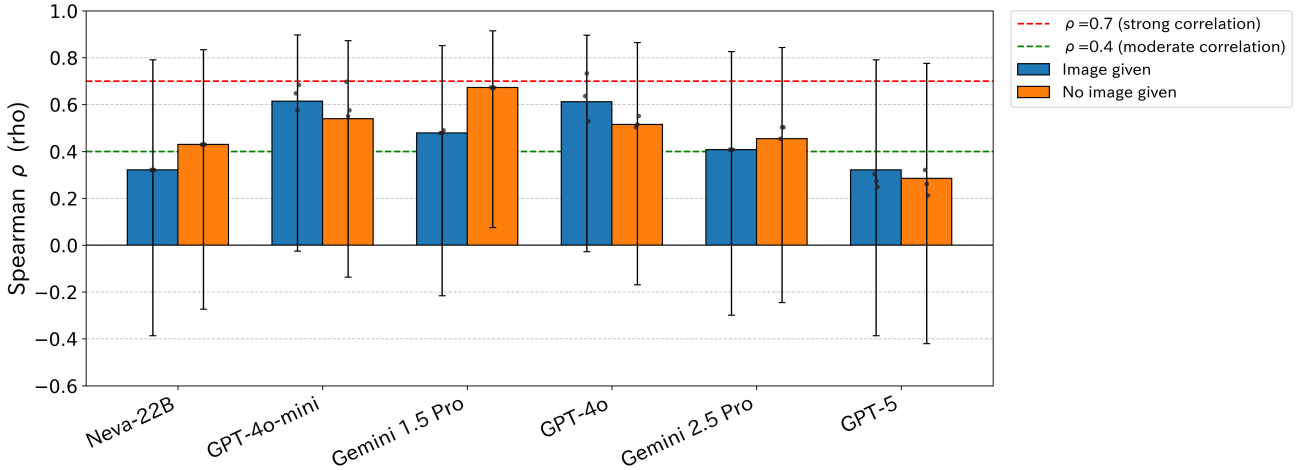
Figure 7. Effect size of visual-heavy subfields under the image-given condition, defined as the difference in accuracy between visual-heavy and text-heavy subfields (visual-heavy minus text-heavy). Negative values indicate lower accuracy in visual-heavy domains. All models exhibited a performance drop in visual-heavy subfields, with larger drops observed for GPT-4o-mini and Neva-22B and smaller drops for GPT-5 and Gemini 2.5 Pro, indicating greater robustness of higher-capacity models to visually demanding content. Points denote the 3 per-run effect sizes; the error bar is the 95% Newcombe hybrid-score CI for the visual-minus-text difference.



Correlation With Human Performance

To evaluate the consistency between model outputs and human accuracy rankings by subfield, we calculated the Spearman rank correlation coefficient (Figure 8). This metric indicates how closely each model mirrors the domain-specific performance trends observed in human examinees; a coefficient closer to 1 implies stronger alignment with human rankings.

Figure 8. Comparison of Spearman rank correlation coefficients (ρ) for each model under image-provided (image given) and no-image (no image given) conditions, with exact *P* values reported in the main text. Based on human and model accuracy, medical subfields were ranked, and Spearman rank correlation was calculated. The red dashed line indicates a reference value for strong correlation (ρ=0.7), while the green dashed line indicates a reference value for moderate correlation (ρ=0.4). This figure quantitatively illustrates how closely each model's performance aligns with human subfield-level scoring trends. Points denote the Spearman coefficient recomputed from each of the 3 individual runs, and error bars are 95% Fisher z CIs (n=10 subfields).



GPT-4o showed relatively high correlations, with ρ=0.61 (95% CI –0.03 to 0.90; *P*=.06) under the image-provided condition and ρ=0.52 (95% CI –0.17 to 0.86; *P*=.13) under the no-image condition, though neither reached statistical significance. GPT-4o-mini exhibited similar results, achieving ρ=0.61 (95% CI –0.03 to 0.90; *P*=.06) with images and ρ=0.54 (95% CI –0.14 to 0.87; *P*=.11) without images, with the image-provided condition approaching significance.

GPT-5 showed ρ=0.32 (95% CI –0.39 to 0.79; *P*=.37) under the image-provided condition, and ρ=0.28 (95% CI –0.42 to 0.78; *P*=.43) under the no-image condition. These weaker correlations indicate limited consistency with human rankings.

Gemini 1.5 Pro recorded the highest correlation under the no-image condition, with ρ=0.67 (95% CI 0.07 - 0.91; *P*=.03), which reached statistical significance. In contrast, under the

XSL•FO
RenderX

image-provided condition it showed a lower correlation (ρ=0.48, 95% CI –0.22 to 0.85; *P*=.16), not reaching significance.

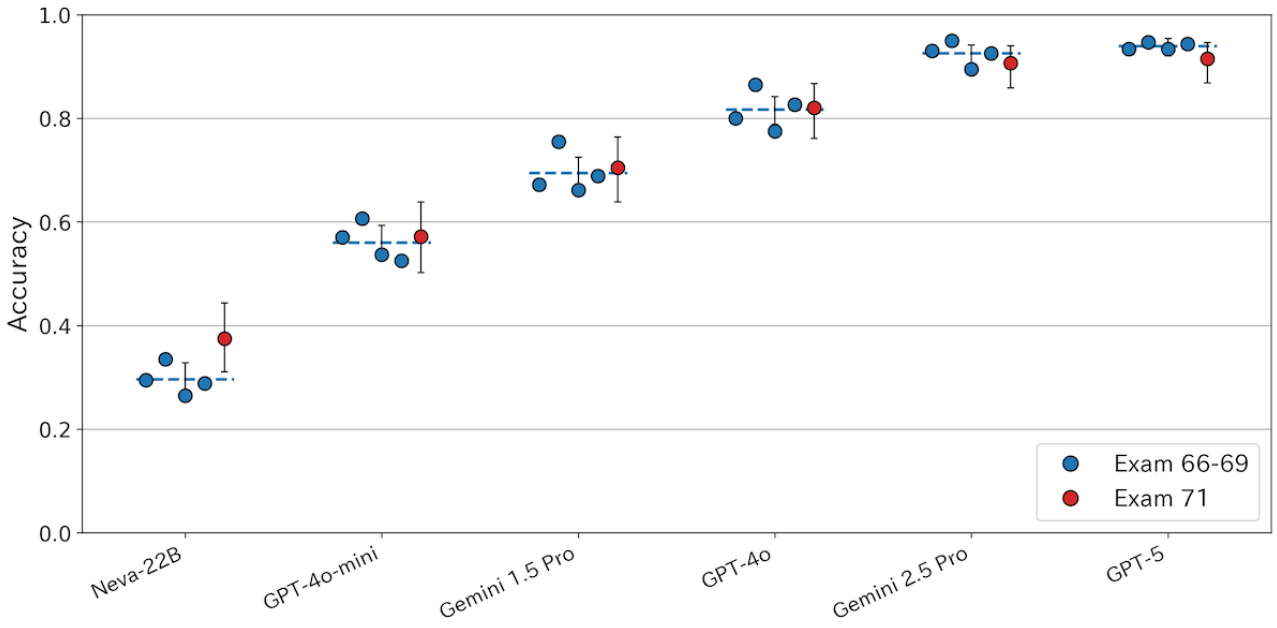
Gemini 2.5 Pro showed moderate correlations—ρ=0.45 (95% CI –0.25 to 0.84; *P*=.19) without images and ρ=0.41 (95% CI –0.30 to 0.83; *P*=.24) with images—but neither was significant. Neva-22B exhibited low correlations in both conditions (ρ=0.43, 95% CI –0.27 to 0.83; *P*=.21 without images; ρ=0.32, 95% CI –0.39 to 0.79; *P*=.37 with images), suggesting little alignment with human rankings.

Evaluation of Model Performance on Unseen Questions After Knowledge Cutoff

All examination questions used in prior evaluations were published before the knowledge cutoff dates of the respective models, raising the possibility that the models may have directly encountered these questions or answer choices during training. To address this concern, we evaluated model performance using the 71st National Examination for Clinical Laboratory Technicians (administered in 2025), which consists of 200 newly released questions (100 in the morning and 100 in the afternoon) published after the respective models' knowledge cutoffs.

As a result, GPT-5 achieved 91.5%, Gemini 2.5 Pro 90.7%, GPT-4o 82.0%, Gemini 1.5 Pro 70.5%, GPT-4o-mini 57.2%, and Neva-22B 37.5% (Figure 9). None of the models exhibited substantial deviations from their respective average accuracies over the previous 4 years (2020 - 2023). The largest increase was observed in Neva-22B, which improved by 7.9 points compared with its historical average of 29.6%. GPT-4o-mini also showed a modest increase of 1.2 points relative to its past average of 56.0%. By contrast, GPT-5 slightly declined from 93.9% to 91.5% (–2.4 points), and Gemini 2.5 Pro from 92.5% to 90.7% (–1.8 points).

Figure 9. Comparison of model accuracy before and after the knowledge cutoff. The figure shows, for each model—GPT-5, GPT-4o, GPT-4o-mini, Neva-22B, Gemini 1.5 Pro, and Gemini 2.5 Pro—the average accuracy from 2020 - 2023 (blue) and the accuracy on the 71st National Examination for Clinical Laboratory Technicians released in 2025 (red), which was published after the knowledge cutoff. Across all models, the 2025 accuracy was comparable to or higher than the historical average, with no significant performance degradation observed. The blue dashed line indicates the average accuracy of each model across the 66th to 69th examinations. Blue points are the 4 yearly accuracies; error bars are 95% Wilson score CIs for the 2020 - 2023 mean (n=800) and for 2025 (n=200).



To statistically validate these observations, we performed chi-square tests comparing the accuracy between the training years (2020 - 2023) and the 2025 examination for each model and image condition. As summarized in Table 4, none of the comparisons showed statistically significant declines (*P*≥.05). These results confirm that the observed declines in accuracy are not statistically significant and support the conclusion that no material performance degradation was observed in the 2025 examination. These findings indicate that all models maintained comparable accuracy when applied to previously unseen questions created after their respective knowledge cutoffs, with no significant performance degradation observed.

XSL•FO
RenderX

Table . Chi-square test results comparing model accuracy between the training years (2020 - 2023; n=800 questions) and the 2025 examination (n=200 questions). For each model and image condition, accuracy is the mean across 3 independent runs (see Methods). A 2×2 contingency table of correct versus incorrect responses was constructed using question-level counts. ΔAcc denotes the difference in accuracy (2025 minus 2020 - 2023, in percentage points), with its two-sided 95% CI computed by the Newcombe hybrid score method.

Model and condition	2020 - 2023 Acc (%)	2025 Acc (%)	ΔAcc (pp)[a]	95% CI of ΔAcc (pp)	*P* value
GPT-4o-mini					
Image given	55.96	57.17	1.21	–6.51 to 8.71	.80
No image given	53.04	55.67	2.62	–5.11 to 10.18	.53
GPT-4o					
Image given	81.67	82.00	0.33	–6.10 to 5.82	.90
No image given	79.75	79.50	–0.25	–6.92 to 5.55	.94
Gemini 1.5 Pro					
Image given	69.42	70.50	1.08	–6.26 to 7.82	.76
No image given	68.46	67.83	–0.62	–8.07 to 6.30	.89
Gemini 2.5 Pro					
Image given	92.50	90.67	–1.83	–6.93 to 2.04	.35
No image given	86.75	87.67	0.92	–4.79 to 5.53	.78
Neva-22B					
Image given	29.58	37.50	7.92	0.72 to 15.45	.03
No image given	32.12	38.50	6.38	–0.89 to 13.96	.09
GPT-5					
Image given	93.92	91.50	–2.42	–7.33 to 1.23	.23
No image given	88.17	86.50	–1.67	–7.47 to 3.06	.53

[a]pp: percentage point.

Discussion

Principal Findings

This study systematically evaluated multiple LVLMs using Kensagishi QA, an examination-specific benchmark constructed from the Japanese National Examination for Clinical Laboratory Technicians (2020 - 2023), and demonstrated that several recent cloud-hosted models exceeded the human passing threshold. At the same time, substantial differences were observed across models in overall accuracy, visual integration, and alignment with human performance patterns, clarifying their respective performance characteristics and applicability in medical contexts.

Interpretation and Comparison With Prior Work

At a high level, recent cloud-hosted LVLMs exceeded the human passing threshold, with GPT-5 showing the strongest overall performance, Gemini 2.5 Pro close behind, and GPT-4o also performing robustly. Gemini 1.5 Pro exhibited greater variability by subfield and condition, whereas GPT-4o-mini and Neva-22B performed relatively poorly overall. Crucially, for image-based questions, providing images improved accuracy for the top-performing models (eg, GPT-5, Gemini 2.5 Pro, and GPT-4o), indicating that visual input contributes materially to performance; by contrast, Neva-22B did not show such gains. These findings suggest that the ability to integrate visual information is a key driver of accuracy in this examination setting. Importantly, this effect was not universal but model-dependent, underscoring that multimodal capability cannot be assumed even when visual input is available.

These findings are broadly consistent with previous studies showing that recent LLMs can achieve high performance on medical licensing examinations. Med-PaLM 2 demonstrated expert-level performance on MedQA and other medical question-answering benchmarks, while subsequent studies incorporating structured clinical knowledge further improved performance on USMLE-style examinations [8,10]. In the Japanese context, GPT-4 successfully passed multiple years of the Japanese National Medical Licensing Examination using the text-based IgakuQA benchmark [11]. Previous work has also reported that GPT-4-class models achieve passing-level performance on the Japanese National Examination for Clinical Laboratory Technicians [13]. Our results are consistent with these reports but extend them in 3 important respects: we evaluated multiple model generations and providers under a unified protocol, explicitly quantified the contribution of visual input using paired image-provided and no-image conditions, and additionally evaluated questions released after the reported knowledge cutoffs.

The paired image-provided and no-image evaluation also provides insights that cannot be obtained from conventional text-only examination benchmarks. Previous multimodal

XSL•FO
RenderX

medical AI studies have demonstrated that vision-language models can answer biomedical image questions [26,27]; however, relatively little has been reported regarding the extent to which visual input contributes to performance on professional licensing examinations. In our benchmark, image provision substantially improved performance for GPT-5, Gemini 2.5 Pro, and GPT-4o, whereas Neva-22B showed no benefit. These findings indicate that multimodal capability should not be regarded as a binary property. Instead, the effectiveness of visual information integration differs markedly among models, even when all models technically accept image input.

Subfield-level patterns further suggest that performance reflects not only overall model capacity but also the cognitive demands of each subfield. For example, the relatively strong performance of smaller models in general clinical laboratory science may reflect its emphasis on standardized procedures and conceptual knowledge, whereas visually intensive or quantitatively demanding subfields (eg, pathological histocytology and clinical chemistry) showed larger gaps between high- and lower-capacity models.

Beyond absolute accuracy, we examined whether models mirrored human domain-level patterns. GPT-4o (and, to a lesser extent, GPT-4o-mini) displayed relatively strong rank alignment with human subfield trends, while GPT-5 achieved the highest accuracy with more limited alignment. Gemini 1.5 Pro showed the strongest human alignment when images were withheld but weaker alignment with images provided. This divergence suggests potentially different use cases; however, strong alignment with human difficulty patterns does not necessarily imply superior pedagogical utility, as it may also reflect shared error tendencies rather than improved explanatory capacity.

To probe generalizability beyond potential training-set exposure, we evaluated all models on the 71st examination (administered in 2025), released after each model's knowledge cutoff. Accuracies were comparable to historical averages from 2020 - 2023, indicating no substantial degradation on unseen questions. This suggests that the evaluated LVLMs leverage transferable knowledge and reasoning rather than memorization of specific items and that the examination's structure and content remain sufficiently stable for such capabilities to carry over.

For practical deployment, the implications of the present findings are primarily educational rather than clinical. Kensagishi QA provides a standardized benchmark for comparing cloud-hosted LVLMs before their adoption in examination preparation systems or educational support tools. The marked effect of visual input also suggests that evaluations intended for visually intensive subjects should include the original examination images rather than relying on text-only transcriptions. At the same time, our study evaluated only answer selection and did not assess explanation quality, hallucinations, calibration, or educational effectiveness [28]. Therefore, high examination accuracy alone should not be interpreted as sufficient evidence for deployment in either educational or clinical settings without additional validation.

Limitations

Several limitations should be considered. First, our evaluation focused exclusively on multiple-choice accuracy and did not assess reliability-related dimensions such as factuality, instruction following, safety [28-30], or appropriate handling of Japanese medical terminology [11,31]. Second, all experiments were conducted under a single "numbers-only" direct-answer prompting condition, which may not fully elicit multi-step reasoning. Prior work has reported that reasoning-oriented prompting strategies, including chain-of-thought, can improve performance in some medical and complex reasoning tasks [8,9,32]. Therefore, the reported accuracies should be interpreted as standardized baseline performance under controlled prompting conditions rather than prompt-optimized upper-bound estimates. Third, the number of image-based questions (n=130) was relatively small, limiting statistical precision in estimating multimodal effects and model-by-image interactions. Fourth, our evaluation did not include high-capacity or medically specialized open-weight LVLMs [26,27]; thus, the present results should not be interpreted as a direct comparison between open- and closed-weight models.

Conclusions

Looking ahead, we plan to broaden coverage to additional model families (including high-capacity and medically specialized open-weight, self-hosted medical LVLMs) under a common protocol; incorporate multidimensional quality metrics that combine accuracy with factuality, instruction compliance, safety, and terminology handling [28,29]; refine prompt design to better leverage domain knowledge [33]; and investigate the interplay between retrieval-augmented approaches and intrinsic reasoning for tasks such as differential diagnosis. Systematic comparison between cloud-hosted and open-weight LVLMs will be necessary to determine whether observed performance differences reflect deployment modality, model scale, architectural design, or other factors. Ultimately, these efforts will inform the development of learning-support tools that visualize mastery, target weak areas, and provide explanatory feedback for national examination preparation and broader medical education [34]. In summary, by providing the first systematic, multi-model evaluation of cloud-hosted LVLMs on this examination with explicit isolation of visual input—extending prior text-only, few-model studies—this work delivers a reproducible Kensagishi QA benchmark and actionable, use-case–specific guidance for deploying LVLMs in medical education and clinical training. More broadly, this framework illustrates how nationally specific professional examinations can serve as rigorous, multilingual testbeds for evaluating and improving medical AI systems in realistic, high-stakes contexts.

XSL•FO
RenderX

Acknowledgments

The authors declare the use of generative AI in the research and writing process. According to the GAIDeT taxonomy (2025), the following tasks were delegated to generative AI (GenAI) tools under full human supervision: proofreading and editing, and translation. The GenAI tool used was ChatGPT (GPT-4o, o3, and GPT-4.5) (OpenAI). ChatGPT was used solely for English-language editing—translation, grammar correction, and stylistic refinement—applied to individual sentences and paragraphs already written by the authors. It was not used to generate new content, summarize, analyze or interpret data, or modify any scientific conclusions. All authors reviewed and edited the AI-assisted text and take full responsibility for the accuracy, originality, and integrity of the manuscript. GenAI tools are not listed as authors and do not bear responsibility for the final outcomes. This declaration is submitted on behalf of all authors (collective responsibility).

Funding

This work was supported by the JST NBDC Integrated Database Promotion Program (JPMJND2402 to HO) and the RIKEN TRIP Advanced General Intelligence for Science Program (AGIS) (to HO).

Authors' Contributions

Conceptualization: KM, HO
Data curation: KM
Formal analysis: KM
Funding acquisition: HO
Investigation: KM
Methodology: KM, RM, YT-A, HO
Project administration: HO
Resources: KM
Software: KM
Supervision: RM, YT-A, HO
Validation: KM
Visualization: KM
Writing – original draft: KM, RM, YT-A, HO
Writing – review & editing: KM, RM, YT-A, HO

Conflicts of Interest

None declared.

References

1. Brown TB, Mann B, Ryder N, Subbiah M, Kaplan J, Dhariwal P, et al. Language models are few-shot learners. arXiv. Preprint posted online on Jul 22, 2020. [doi: 10.48550/arXiv.2005.14165]
2. OpenAI, Adler S, Agarwal S, Ahmad L, Akkaya I, et al. GPT-4 technical report. arXiv. Preprint posted online on Mar 15, 2023. [doi: 10.48550/arXiv.2303.08774]
3. Shao M, Basit A, Karri R, Shafique M. Survey of different large language model architectures: trends, benchmarks, and challenges. IEEE Access 2024;12:188664-188706. [doi: 10.1109/ACCESS.2024.3482107]
4. Zhang D, Yu Y, Dong J, et al. MM-LLMs: recent advances in multimodal large language models. Presented at: Findings of the Association for Computational Linguistics ACL 2024; Aug 11-16, 2024; Bangkok, Thailand and virtual meeting. 2024 p. 12401-12430 URL: https://aclanthology.org/2024.findings-acl.738/ [accessed 2026-08-13] [doi: 10.18653/v1/2024.findings-acl.738]
5. Chou C, Dunlap L, Mashita K, et al. VisionArena: 230K real world user-VLM conversations with preference labels. Presented at: 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); Jun 11-15, 2025; Nashville, TN, USA. 2025 p. 3877-3887. [doi: 10.1109/CVPR52734.2025.00367]
6. Chen L, Li J, Dong X, et al. Are we on the right way for evaluating large vision-language models? Presented at: Advances in Neural Information Processing Systems 37; Dec 10-15, 2024; Vancouver, BC, Canada. 2024 p. 27056-27087 URL: http://www.proceedings.com/79017.html [accessed 2026-08-13] [doi: 10.52202/079017-0850]
7. Jin D, Pan E, Oufattole N, Weng WH, Fang H, Szolovits P. What disease does this patient have? A large-scale open domain question answering dataset from medical exams. Appl Sci 2021;11(14):6421. [doi: 10.3390/app11146421]
8. Singhal K, Tu T, Gottweis J, et al. Toward expert-level medical question answering with large language models. Nat Med 2025 Mar;31(3):943-950. [doi: 10.1038/s41591-024-03423-7] [Medline: 39779926]
9. Yang H, Chen H, Guo H, et al. LLM-MedQA: enhancing medical question answering through case studies in large language models. Presented at: 2025 International Joint Conference on Neural Networks (IJCNN); Jun 30 to Jul 5, 2025. [doi: 10.1109/IJCNN64981.2025.11228647]

XSL•FO
RenderX

10. Elkin PL, Mehta G, LeHouillier F, et al. Semantic clinical artificial intelligence vs native large language model performance on the USMLE. JAMA Netw Open 2025 Apr 1;8(4):e256359. [doi: 10.1001/jamanetworkopen.2025.6359] [Medline: 40261653]
11. Kasai J, Kasai Y, Sakaguchi K, Yamada Y, Radev D. Evaluating GPT-4 and ChatGPT on Japanese medical licensing examinations. arXiv. Preprint posted online on Mar 31, 2023. [doi: 10.48550/arXiv.2303.18027]
12. Kido Y, Yamada H, Tokunaga T, et al. Evaluation of LLM-generated distractors of multiple-choice questions for the Japanese National Nursing Examination. Presented at: 17th International Conference on Computer Supported Education; Apr 1-3, 2025; Porto, Portugal. 2025 p. 754-764 URL: http://www.scitepress.org/DigitalLibrary/ProceedingLink.aspx?ID=1900 [accessed 2026-08-13] [doi: 10.5220/0013460300003932]
13. Doi H, Ishida H, Nagasawa H, Tsuboi Y, Kikuchi R, Ichino N, et al. Performance of generative pretrained transformer on the national licensing examination for medical technologist in Japan. Jpn J Med Technol 2024;73(2):323-331. [doi: 10.14932/jamt.23-80]
14. Ohshige M, Ninomiya A, Akaza M, et al. Can large language models pass the national examination for medical technologists? Jpn J Med Technol Educ 2024;16:99-105 [FREE Full text]
15. Tanaka Y, Nakata T, Aiga K, et al. Performance of generative pretrained transformer on the National Medical Licensing Examination in Japan. PLoS Digit Health 2024 Jan;3(1):e0000433. [doi: 10.1371/journal.pdig.0000433] [Medline: 38261580]
16. Questions and correct answers for the 66th national examination for medical technologists. Ministry of Health, Labour and Welfare. 2020. URL: https://www.mhlw.go.jp/seisakunitsuite/bunya/kenkou_iryou/iryou/topics/tp200414-07.html [accessed 2025-06-23]
17. Questions and correct answers for the 67th national examination for medical technologists. Ministry of Health, Labour and Welfare. 2021. URL: https://www.mhlw.go.jp/seisakunitsuite/bunya/kenkou_iryou/iryou/topics/tp210416-07.html [accessed 2025-06-23]
18. Questions and correct answers for the 68th national examination for medical technologists. Ministry of Health, Labour and Welfare. 2022. URL: https://www.mhlw.go.jp/seisakunitsuite/bunya/kenkou_iryou/iryou/topics/tp220421-07.html [accessed 2025-06-23]
19. Questions and correct answers for the 69th national examination for medical technologists. Ministry of Health, Labour and Welfare. 2023. URL: https://www.mhlw.go.jp/seisakunitsuite/bunya/kenkou_iryou/iryou/topics/tp230524-07.html [accessed 2025-06-23]
20. Onodera R. An approach based on national examination results analysis: Clinical laboratory technologist education and postgraduate education and qualifications required after the curriculum revision. Jpn J Med Technol Educ 2024;16:37-44 [FREE Full text]
21. Wilson EB. Probable inference, the law of succession, and statistical inference. J Am Stat Assoc 1927 Jun;22(158):209-212. [doi: 10.1080/01621459.1927.10502953]
22. Newcombe RG. Interval estimation for the difference between independent proportions: Comparison of eleven methods. Stat Med 1998;17(8):873-890. [doi: 10.1002/(SICI)1097-0258(19980430)17:8<873::AID-SIM779>3.0.CO;2-I] [Medline: 9595617]
23. Murakami K, Matsuzawa R, Tahara-Arai Y, Ozaki H. KensagishiQA: Benchmark data for evaluating large language models on the Japanese National Examination for Clinical Laboratory Technicians. GitHub. 2026 Jul 15. URL: https://github.com/bioinfo-tsukuba/KensagishiQA [accessed 2026-09-02]
24. Questions and correct answers for the 71st national examination for medical technologists. Ministry of Health, Labour and Welfare. 2025. URL: https://www.mhlw.go.jp/seisakunitsuite/bunya/kenkou_iryou/iryou/topics/tp250428-07.html [accessed 2025-07-17]
25. Ethical guidelines for medical and biological research involving human subjects. Ministry of Education, Culture, Sports, Science and Technology, Ministry of Health, Labour and Welfare, Ministry of Economy, Trade and Industry. 2023. URL: https://www.mhlw.go.jp/content/001457376.pdf [accessed 2026-07-18]
26. Li C, Wong C, Zhang S, et al. LLaVA-Med: training a large language-and-vision assistant for biomedicine in one day. Presented at: Advances in Neural Information Processing Systems 36; Dec 10-16, 2023 URL: http://www.proceedings.com/75280.html [accessed 2026-08-13] [doi: 10.52202/075280-1240]
27. Moor M, Huang Q, Wu S, Yasunaga M, Zakka C, Dalmia Y, et al. Med-Flamingo: a multimodal medical few-shot learner. arXiv. Preprint posted online on Jul 27, 2023. [doi: 10.48550/arXiv.2307.15189]
28. Asgari E, Montaña-Brown N, Dubois M, et al. A framework to assess clinical safety and hallucination rates of LLMs for medical text summarisation. NPJ Digit Med 2025 May 13;8(1):274. [doi: 10.1038/s41746-025-01670-7] [Medline: 40360677]
29. Singhal K, Azizi S, Tu T, et al. Large language models encode clinical knowledge. Nature 2023 Aug;620(7972):172-180. [doi: 10.1038/s41586-023-06291-2] [Medline: 37438534]
30. Liu J, Li Z, Niu Q, Zhang Z, Xun Y, Hou W, et al. JMedEthicBench: a multi-turn conversational benchmark for evaluating medical safety in Japanese large language models. arXiv. Preprint posted online on Jan 4, 2026. [doi: 10.48550/arXiv.2601.01627]
31. Jiang J, Huang J, Aizawa A. JMedBench: a benchmark for evaluating Japanese biomedical large language models. In: Rambow O, Wanner L, Apidianaki M, Al-Khalifa H, Eugenio BD, Schockaert S, editors. Presented at: Proceedings of the

XSL•FO
RenderX

31st international conference on computational linguistics; Jan 19-24, 2025; Abu Dhabi, UAE. 2025 p. 5918-5935 URL: https://aclanthology.org/2025.coling-main.395/ [accessed 2026-08-31]
32. Wei J, Wang X, Schuurmans D, et al. Chain-of-thought prompting elicits reasoning in large language models. Presented at: Advances in Neural Information Processing Systems 35; Dec 6-9, 2022; New Orleans, Louisiana. 2022 p. 24824-24837 URL: http://www.proceedings.com/68431.html [accessed 2026-08-13] [doi: 10.52202/068431-1800]
33. Wang L, Chen X, Deng X, et al. Prompt engineering in consistency and reliability with the evidence-based guideline for LLMs. NPJ Digit Med 2024 Feb 20;7(1):41. [doi: 10.1038/s41746-024-01029-4] [Medline: 38378899]
34. Ch'en PY, Day W, Pekson RC, et al. GPT-4 generated answer rationales to multiple choice assessment questions in undergraduate medical education. BMC Med Educ 2025 Mar 4;25(1):333. [doi: 10.1186/s12909-025-06862-z] [Medline: 40038669]

Abbreviations

GPU: graphics processing unit
Kensagishi QA: Kensagishi question-answer
LLM: large language model
LVLM: large vision-language model
MHLW: Ministry of Health, Labour and Welfare
PDL1.0: Public Data License version 1.0
SCAI: Semantic Clinical AI
USMLE: United States Medical Licensing Examination

Edited by S Brini; submitted 17.Sep.2025; peer-reviewed by AG Gad, K Ishida; revised version received 22.Jul.2026; accepted 23.Jul.2026; published 11.Sep.2026.

Please cite as:
Murakami K, Matsuzawa R, Tahara-Arai Y, Ozaki H
Performance of Cloud-Hosted Large Vision-Language Models on the Japanese National Examination for Clinical Laboratory Technicians: Comparative Benchmarking Study
JMIR Med Educ 2026;12:e84266
URL: https://mededu.jmir.org/2026/1/e84266
doi:10.2196/84266

© Kota Murakami, Ryosuke Matsuzawa, Yuya Tahara-Arai, Haruka Ozaki. Originally published in JMIR Medical Education (https://mededu.jmir.org), 11.Sep.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Education, is properly cited. The complete bibliographic information, a link to the original publication on https://mededu.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Teaching Model Context Protocol, Retrieval-Augmented Generation, and AI Agents to a Multidisciplinary Hospital Workforce: Single-Group Pre-Post Survey Study

Gakyoung Baek[1], BS; Hyunna Lee[1], PhD; Dong Hyun Yang[2], MD, PhD; Minseo Kang[1], BS; Kun Hee Lee[1], BS; Minji Choi[1], BA; Yura Lee[3], MD, PhD; Kye Hwa Lee[3], MD, PhD

[1]Big Data Research Center, Asan Institute for Life Sciences, Asan Medical Center, Seoul, Republic of Korea

[2]Department of Radiology and Research Institute of Radiology, University of Ulsan College of Medicine, Asan Medical Center, Seoul, Republic of Korea

[3]Department of Information Medicine, Asan Medical Center, University of Ulsan College of Medicine, 88 Olympic-ro 43-gil, Songpa-gu, Seoul, Republic of Korea

Corresponding Author:
Kye Hwa Lee, MD, PhD
Department of Information Medicine, Asan Medical Center, University of Ulsan College of Medicine, 88 Olympic-ro 43-gil, Songpa-gu, Seoul, Republic of Korea

Abstract

Background: Hospitals worldwide need to upskill their workforce in advanced AI technologies; yet, published guidance on how to design and deliver such training, particularly in agent-level tools like retrieval-augmented generation (RAG) and the model context protocol (MCP), remains virtually absent.

Objective: To describe the design, implementation, and lessons learned from an 8-week, 56-hour intensive generative AI training program for a multidisciplinary hospital workforce, drawing on both quantitative outcome data and participants' own reflections on their learning experience.

Methods: The program was delivered on-site at Asan Medical Center with simultaneous online broadcast to 2 regional affiliate hospitals. The curriculum was built around the premise that MCP and AI agents would become the foundation of health care AI use, allocating 37% (11.5/31 hours) of on-site instructional time to MCP, and 71% (22/31 hours) to hands-on practice. Participants progressed from foundational concepts through RAG and MCP to team-based capstone projects, supported by funded AI tool subscriptions, a dedicated internal cloud platform, and 3‐6 hours of weekly mentoring per team. A pre-post survey (pre: n=83; post: n=64) evaluated outcomes across Kirkpatrick levels 1‐3, complemented by thematic analysis of open-ended reflections on self-perceived growth.

Results: The technologies that received the greatest curricular investment were associated with the largest self-efficacy differences (MCP: d=1.57; overall effect: r=.574), and participants most frequently cited MCP and RAG when describing how abstract concepts "became concrete and actionable." Non-IT professionals, clinicians, health information managers, researchers, and administrative staff showed consistently larger gains than IT specialists; several reported coding for the first time through vibe coding, challenging the assumption that advanced AI training requires technical backgrounds. Despite significant overall gains, a knowledge-practice gap persisted: job-specific competency remained below the scale midpoint, though participants spontaneously reported generating workplace application ideas. Curriculum pacing was rated lowest despite high overall satisfaction (4.03/5), signaling that even 56 hours may progress too quickly for mixed-expertise cohorts. Capstone projects with dedicated mentoring received the highest satisfaction ratings; 11 of 12 teams presented functional prototypes, and one has since entered active pilot use in clinical departments ahead of planned hospital-wide deployment.

Conclusions: To our knowledge, this is the first program to teach 4 agent-level generative AI technologies, MCP, RAG, LangGraph orchestration, and AI agent design, to both IT and non-IT hospital staff. This program suggests that transforming a multidisciplinary hospital workforce into AI-capable professionals is achievable through intensive, hands-on training centered on agent-level technologies, and that capstone projects with dedicated mentoring can serve as a pathway from classroom learning toward institutional AI adoption. The knowledge-practice gap highlights the need for posttraining support structures to translate self-efficacy gains into sustained workplace practice.

(JMIR Med Educ 2026;12:e97822) doi:10.2196/97822

XSL•FO
RenderX

KEYWORDS

artificial intelligence; medical education; large language models; self-efficacy; health personnel; model context protocol; staff development; curriculum design

Introduction

Background

Hospitals that want to harness AI for clinical documentation, decision support, and administrative automation [1-4] face a workforce problem: the technologies that make these applications possible, retrieval-augmented generation (RAG) [5], the model context protocol (MCP) [6], and orchestration frameworks such as LangGraph [7], are evolving faster than the people who must build, adapt, and evaluate them [8,9]. The question is no longer whether to train hospital staff in advanced AI, but how.

Existing training efforts offer limited guidance. The 2024 Best Evidence Medical Education (BEME) scoping review by Gordon et al [10] confirmed that published AI education programs remain predominantly foundational—covering AI literacy, prompt engineering, and learners' perceptions of generative AI tools such as ChatGPT [11-13]—and target mainly medical students and residents [14,15]. Evidence on intensive, multiweek programs that teach advanced technologies to the broader hospital workforce—IT specialists, health information managers (HIM), clinicians, researchers, and administrative staff—is scarce [16-18]. No published study, to our knowledge, has reported the design and outcomes of training that covers RAG, MCP, AI agent design, and workflow orchestration together.

Prior Work

Self-efficacy theory [19] predicts that mastery experiences, particularly hands-on practice with novel technologies, are the strongest driver of sustained technology adoption [20-22]. Consistent with this, recent ChatGPT education studies report that structured training significantly raises perceived AI competency among health care professionals [8,12], and experiential learning approaches outperform didactic instruction for skill-based outcomes [23]. However, most evaluations remain limited to satisfaction surveys (Kirkpatrick level 1 [24,25]), with few assessing knowledge gains (Level 2) or behavioral transfer intentions (Level 3) [26,27]. Our previous evaluation of a health informatics analyst program at the same institution demonstrated that intensive education can reshape job roles and skill profiles [18]; this study extends that work to advanced generative AI technologies.

Study Objectives

This study has two complementary aims: (1) to describe the design rationale, curriculum structure, and implementation experience of an 8-week, 56-hour intensive generative AI training program for a multidisciplinary hospital workforce at a tertiary academic medical center; and (2) to evaluate its outcomes using the Kirkpatrick framework across Levels 1 - 3, providing empirical evidence that contextualizes the lessons learned. We addressed four research questions (RQs):

1. RQ1: Are posttraining AI knowledge self-efficacy and job-specific AI competency scores higher than pretraining scores? (Level 2; S4, S6)
2. RQ2: How satisfied are participants with the training program, and which curricular components are rated highest and lowest? (Level 1; S9)
3. RQ3: What are participants' behavioral intentions to apply learned skills, and where do gaps between knowledge and application emerge? (Level 3 proxy; S11)
4. RQ4: Do between-group differences in AI knowledge self-efficacy vary by professional group, and what does this imply for participant selection?

Methods

Study Design

This was a single-group, pre-post survey study using an independent samples design conducted at a single tertiary academic medical center in Seoul, South Korea. The training program was administered from September to November 2025, with surveys distributed immediately before (pretraining) and after (posttraining) the program. Because survey responses were collected anonymously without personal identifiers, individual-level matching between pre- and posttraining responses was not possible; therefore, pre- and posttraining groups were treated as independent samples. This study adhered to the STROBE (Strengthening the Reporting of Observational Studies in Epidemiology) guidelines for cross-sectional studies [28].

Participants

Eligible participants were employees of Asan Medical Center, Ulsan University Hospital, or Gangneung Asan Hospital who enrolled in the advanced generative AI training program. The program was delivered on-site at Asan Medical Center with simultaneous online broadcast to the 2 regional affiliate hospitals, enabling multisite participation. Inclusion criteria were (1) current employment at one of the participating institutions, (2) voluntary enrollment in the program, and (3) age 18 years or older. Participants who attended fewer than 80% of the total training hours were excluded; all 83 enrollees met this threshold. Recruitment did not use an open call; instead, the program was promoted internally within candidate departments through departmental announcements and circulated notices, and interested employees enrolled voluntarily. Because the program's primary aim was to build AI-agent development capability, enrollment was prioritized in the order of IT workforce, health-information staff, clinicians, and researchers, with IT and related departments approached first.

All 83 enrollees completed the pretraining survey. Of these, 64 responded to the posttraining survey (response rate: 77.1%). A total of 147 survey responses were thus analyzed (pre: n=83; post: n=64). Demographic and occupational characteristics, professional group and years of work experience, were

XSL•FO
RenderX

self-reported by participants as part of both the pre- and posttraining surveys, without collection of personal identifiers, and are detailed in the Results section and Table 1.

Table . Participant demographics.

Characteristics	Pretraining (n=83)	Posttraining (n=64)	*P* value
Occupation, n (%)			.87[a]
IT specialists	48 (57.8)	33 (51.6)	
HIM[b]	10 (12)	9 (14.1)	
Clinicians	9 (10.8)	7 (10.9)	
Researchers	9 (10.8)	8 (12.5)	
Administrative staff	7 (8.4)	7 (10.9)	
Work experience, n (%)			.85[a]
<1 year	3 (3.6)	3 (4.7)	
1 - 5 years	13 (15.7)	12 (18.8)	
6 - 10 years	18 (21.7)	14 (21.9)	
≥11 years	49 (59.0)	35 (54.7)	

[a]chi-squared test.

[b]HIM: health information managers.

Training Program: Design Rationale and Curriculum Structure

Overview

The training program was conducted as part of the Medical AI Healthcare Professional Continuing Education Project, a national initiative organized by the Ministry of Health and Welfare and the Korea Human Resource Development Institute for Health & Welfare (KOHI). This institutional framework provided the mandate and funding structure for delivering advanced AI training to hospital staff, and situates the present program within a broader national effort to build health care AI workforce capacity.

Design Philosophy

Within this framework, the program was designed around a central premise: that MCP and AI agents would become the foundational infrastructure for health care AI use, much as databases became foundational for health informatics. Rather than teaching AI as a collection of isolated tools, the curriculum treated agent-level capabilities, connecting AI models to institutional data sources, orchestrating multistep workflows, and building domain-specific applications, as the core competency target. This premise motivated two key design decisions: (1) allocating the largest share of instructional time to MCP (11.5 h, 37% of instructional time), and (2) requiring all participants to build functional MCP-based prototypes through team capstone projects, on the reasoning that hands-on development experience is the fastest path to genuine understanding and use of AI agent architectures. Relatedly, the inaugural cohort was deliberately composed of staff positioned to enable hospital-wide AI adoption, IT and health-information personnel, related-department clinicians, and researchers, as the first phase of a staged institutional strategy rather than as an end in itself. This IT-first prioritization reflected an assumption that the interface between clinicians and IT professionals will expand substantially in the AI-agent era; building hospital-needs-driven agent-development capability first was expected to lay the foundation for later, clinician-needs-driven AI adoption.

The program prioritized experiential learning. Hands-on practice comprised 71% of on-site instructional time (22 of 31 h); the remaining 9 hours (29%) were didactic lectures. The subsequent 3-week team capstone added 25 hours of project-based work, so experiential learning dominated the 56-hour program overall (47 of 56 h, 84%). This ratio was informed by self-efficacy theory [19], which predicts that mastery experiences—direct, successful practice—are the most potent source of self-efficacy, and by evidence that experiential learning produces stronger skill-based outcomes than didactic instruction alone [23]. Because the cohort spanned a wide range of baseline technical skill—from health-information staff who had never written code to experienced IT specialists—the curriculum also deliberately minimized AI theory (such as the mathematics of deep learning or the internal mechanics of large language models [LLMs] and agents), emphasizing instead practical terminology, core concepts, the essential agentic methods participants needed, and concrete MCP use cases and development techniques. Within the hands-on sessions, participants worked in 2 parallel tracks matched to baseline skill—an application track (assembling solutions with existing frameworks, such as a RAG pipeline in LangChain) and a development track (implementing components directly, such as embedding and similarity search)—so that both first-time coders and experienced developers were appropriately challenged.

To ensure that all participants could engage in hands-on practice without technical barriers, the program provided substantial infrastructure support. Each participant received 3-month funded subscriptions to commercial AI development tools (Cursor IDE

XSL•FO
RenderX

[Anysphere], Claude API, and GPT API), removing cost as an obstacle to experimentation. Additionally, the institution deployed a dedicated internal cloud environment where participants could design, test, and run MCP servers using local LLMs within the hospital's secure network, addressing the data governance constraints that typically impede AI development in health care settings.

Curriculum Structure

The 8-week, 56-hour program consisted of 5 on-site instructional sessions (31 h) and a 3-week team-based capstone project (25 h; Table S1 in Multimedia Appendix 1). The curriculum followed a scaffolded progression designed to build competencies incrementally:

1. Week 1 — foundations (7 h): foundation models and prompt engineering. Established shared vocabulary and baseline skills across all professional groups.
2. Week 2 — RAG and LangGraph workflow design (7 h): RAG [5] and medical document question-answering systems. Introduced the concept of grounding AI outputs in institutional knowledge bases. In the hands-on session, participants built a working RAG pipeline that answered questions over a corpus of institutional clinical documents. The afternoon introduced LangGraph for graph-based clinical-workflow design, modeling a care process (registration → triage → testing → diagnosis → prescription) as nodes and edges.
3. Week 3 — MCP and vector databases (7 h): MCP [6] fundamentals and vector database integration. As a hands-on exercise, participants built their first MCP server and connected it to a vector database of internal guidelines.
4. Week 4 — advanced agent design (7 h): advanced MCP server development and LangGraph-based AI agent workflows [7,29]. As capstone preparation, participants designed a multistep, tool-calling clinical workflow agent in LangGraph.
5. Weeks 5 - 7 — team capstone projects (25 h): Self-directed team projects in which participants developed working AI prototypes for real health care challenges. Teams were composed of mixed professional backgrounds (IT specialists, clinicians, HIMs, researchers, and administrative staff) to leverage domain expertise alongside technical skills.
6. Week 8 — integration (3 h): Advanced MCP host development and workflow integration; the program concluded with capstone presentations and an awards ceremony.

The technology-specific instructional hours, calculated as a percentage of the 31 on-site instructional hours, were distributed as follows: MCP (11.5 h, 37%), LangGraph workflow design (8.5 h, 27%), foundation models and prompt engineering (7 h, 23%), RAG (5 h, 16%), medical Question and Answer (Q&A) systems (5 h, 16%), and AI agent design (4 h, 13%). Because some sessions covered more than one technology, their hours were counted in each relevant category, so the listed percentages sum to more than 100%. The heavy MCP allocation reflected the program's core thesis that agent-level infrastructure would be the most transferable and enduring competency.

A detailed session-level syllabus, including per-session learning objectives, hands-on exercises, and tools for all 8 weeks, is provided in Multimedia Appendix 2. To support reproducibility while respecting institutional and funding constraints, we release this syllabus and the capstone evaluation rubric (Multimedia Appendix 3); full scaffold code, proprietary MCP server templates, and internal datasets are not released owing to institutional intellectual-property policy and national-program KOHI constraints.

Capstone Projects

All participants were organized into 12 teams, including one team each from Ulsan University Hospital and Gangneung Asan Hospital, each comprising a mix of professional backgrounds. To preserve familiar communication channels while combining expertise, each team was built around an intact IT subdepartment as its technical core, augmented with non-IT domain members (for example, an IT infrastructure team joined by 2 health-information staff and a clinician); this structure was applied consistently across all teams. Each team identified a real-world health care problem within their work context and developed a functional MCP-based AI prototype addressing it over 3 weeks. A single dedicated mentor, selected from the program instructors, provided 3 - 6 hours of structured mentoring per team per week throughout the capstone period, ensuring consistent guidance across all projects. The capstone served dual purposes: providing extended development time for skill consolidation and generating tangible evidence of practical competency beyond self-report measures.

Capstone projects were evaluated at the closing ceremony (November 11, 2025) by a panel of 7 assessors using a standardized 100-point rubric comprising four weighted domains—AI technical use (40 points; 4 items), user experience and user interface (20 points; 2 items), completeness and stability (20 points; 2 items), and clinical applicability (20 points; 2 items)—with each item rated on a 1 - 10 scale. To adjust for differences in scoring stringency across assessors, each assessor's raw scores were standardized (z-transformed) within assessor, converted to rank-based scores, and aggregated across the 7 assessors to determine final standings. The full rubric is provided in Multimedia Appendix 3.

Outcome Measures

Survey items were developed by the research team specifically for this program, grounded in the competency domains of the training curriculum (MCP, RAG, LangGraph, AI agent design, and domain-specific applications) and informed by Compeau and Higgins [20] computer self-efficacy scale, adapted to the generative AI domain. The instrument underwent internal review by the program instructors prior to deployment. Post hoc psychometric analyses (Cronbach α, item-total correlations, exploratory factor analysis [EFA]) are reported in Tables S5-S6 in Multimedia Appendix 1.

Survey instruments consisted of 100 pretraining items and 154 posttraining items organized into 11 sections. All attitudinal and self-efficacy items used a 5-point Likert scale (1=strongly disagree to 5=strongly agree). The measures were mapped to

XSL•FO
RenderX

the Kirkpatrick 4-level training evaluation model [24,25] as follows:

Primary outcomes (pre-post comparison):

1. AI knowledge self-efficacy (S4; 8 items): self-assessed competency across 8 AI domains, basic AI understanding, prompt engineering, RAG, local LLM deployment, LangGraph workflow, MCP, medical Q&A systems, and AI agent design. This corresponded to Kirkpatrick level 2 (learning).
2. Job-specific AI competency (S6; 7 items per professional group): self-assessed ability to apply AI tools to job-specific tasks, corresponding to Kirkpatrick level 2. Although each professional group received items tailored to their work context, all items assessed the same underlying construct (perceived ability to apply AI to professional tasks) on an identical response scale, enabling aggregation for overall pre-post comparisons.

Secondary outcomes (pre-post comparison):

1. AI attitudes (S2; 9 items): perceptions toward AI adoption, including apprehension, replacement concerns, and adaptation confidence.
2. AI daily usage (S3; 2 categorical items): frequency and duration of AI tool use in daily work.
3. Digital and programming competency (S5; 4 Likert items for digital competency, plus multiselect programming language proficiency): self-assessed digital skills and programming ability. Scale analysis was based on the 4 digital competency items.

Posttraining only outcomes:

1. Training satisfaction (S9; 18 items across 6 domains): curriculum design, content and practice, instructor quality, teaching methods, team project experience, and learning environment, corresponding to Kirkpatrick level 1 (reaction).
2. Perceived changes (S10; 10 items): self-reported changes in knowledge, skills, attitudes, and workplace application confidence following training.
3. Workplace application intention (S11; 6 core Likert items + supplementary topic-specific and categorical items): plans for applying learned skills to current work, peer sharing, and continued learning, corresponding to Kirkpatrick level 3 (behavioral intention). S11 measures behavioral intention rather than observed behavioral change; true level 3 assessment would require longitudinal follow-up, which was beyond the scope of this study. Scale analysis was based on the 6 core items.
4. Self-perceived growth (1 open-ended item): "What aspect of change or growth did you most strongly feel through this 8-week program?" All 64 responses were read in full by the research team. Recurring themes were identified inductively from the responses rather than applied from an a priori framework, and a set of 5 themes was agreed upon. Each response was then assigned to 1 or 2 of these themes (responses too brief to interpret were set aside as unclassifiable), and the number of responses per theme was tallied. The responses were originally written in Korean; representative quotations were translated into English, and the complete set of translated responses is provided in Multimedia Appendix 4.

Statistical Analysis

All analyses were performed using Python 3.x (pandas v2.0, SciPy v1.11, factor_analyzer). The anonymous design required treating pre- and posttraining groups as independent samples. Mann-Whitney *U* tests [30] were used for pre-post comparisons, with the rank-biserial correlation r (= $|Z| / \sqrt{N}$; Rosenthal [31]) as the primary effect size, interpreted as small (≥.10), medium (≥.30), or large (≥.50) per Cohen [32]. Cohen *d* was additionally reported to facilitate comparison with prior literature. The significance threshold was α=0.05 (2-tailed); posttraining-only measures (S9, S10, and S11) were summarized descriptively. Chi-squared tests assessed demographic comparability between the 2 independent samples.

Subgroup analyses by professional group were exploratory: Cohen *d* with 95% CIs was computed without formal hypothesis testing or multiple-comparisons correction, given small cell sizes (n=7 - 9 for non-IT groups). Full inferential statistics are available in Table S2 in Multimedia Appendix 1.

Internal consistency was assessed using Cronbach α (acceptable≥0.70, good≥0.80 [33]). EFA (principal axis factoring, direct oblimin rotation) and corrected item-total correlations provided post hoc construct validity evidence (Tables S5-S6 in Multimedia Appendix 1).

A worst-case sensitivity analysis assigned all 19 nonrespondents' posttraining scores equal to their professional group's pretraining mean (Table S7 in Multimedia Appendix 1). A post hoc power analysis indicated the minimum detectable effect at 80% power was r=.24 (Table S4 in Multimedia Appendix 1).

Ethical Considerations

The study protocol was approved by the Institutional Review Board of Asan Medical Center (IRB number 2025 - 1070; approved August 28, 2025). Written informed consent was waived because data were collected anonymously through voluntary online surveys, participation posed no more than minimal risk, and the study could not practicably be conducted without the waiver. The first page of the survey informed participants of the study purpose, voluntary nature, and data handling procedures. Participants received no monetary compensation; program tuition was fully funded through the national continuing-education initiative administered by the KOHI.

Results

Participant Characteristics

All 83 enrollees completed the program and responded to the pretraining survey; 64 responded to the posttraining survey (response rate: 77.1%; pre: n=83; post: n=64; Figure 1). Demographic characteristics of the pre- and posttraining groups are presented in Table 1. The occupational distribution was comparable between the 2 groups (IT specialists: 57.8%, 48/83 pre vs 51.6%, 33/64 post; HIM: 12%, 10/83, vs 14.1%, 9/64; clinicians: 10.8%, 9/83, vs 10.9%, 7/64; researchers: 10.8%,

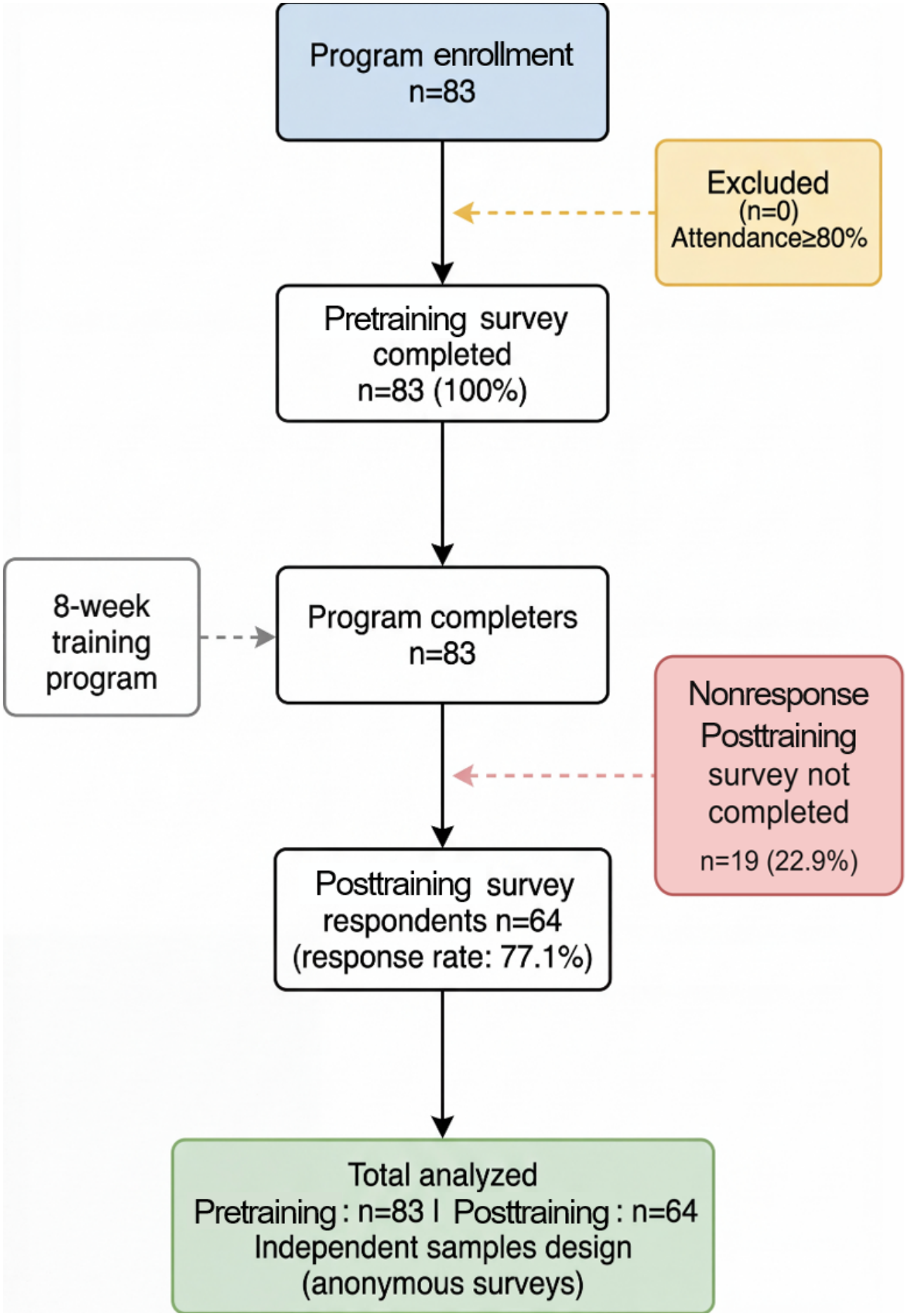
XSL•FO
RenderX

9/83, vs 12.5%, 8/64; administrative staff: 8.4%, 7/83, vs 10.9%, 7/64); retention was lower among IT specialists (33/48, 68.8%) than among non-IT groups (78%, 7/9 to 100%, 7/7; see Limitations). The majority of participants in both groups had over 11 years of experience (59%, 49/83 pre vs 54.7%, 35/64 post). Chi-squared tests confirmed no statistically significant differences in occupational (χ^2_4=1.23; P=.87) or experience-level (χ^2_3=0.80; P=.85) distributions between the 2 independent samples.

XSL•FO
RenderX

Figure 1. Participant flow diagram showing enrollment (n=83), completion of the 8-week training program (n=83), and response rates for the pretraining (n=83; 100%) and posttraining (n=64; 77.1%) surveys. No participants were excluded, as all enrollees met the ≥80% attendance criterion. Because the surveys were anonymous, the pretraining and posttraining groups were analyzed as independent samples.

XSL•FO
RenderX



XSL•FO
RenderX

Primary Outcomes: AI Knowledge Self-Efficacy (S4)

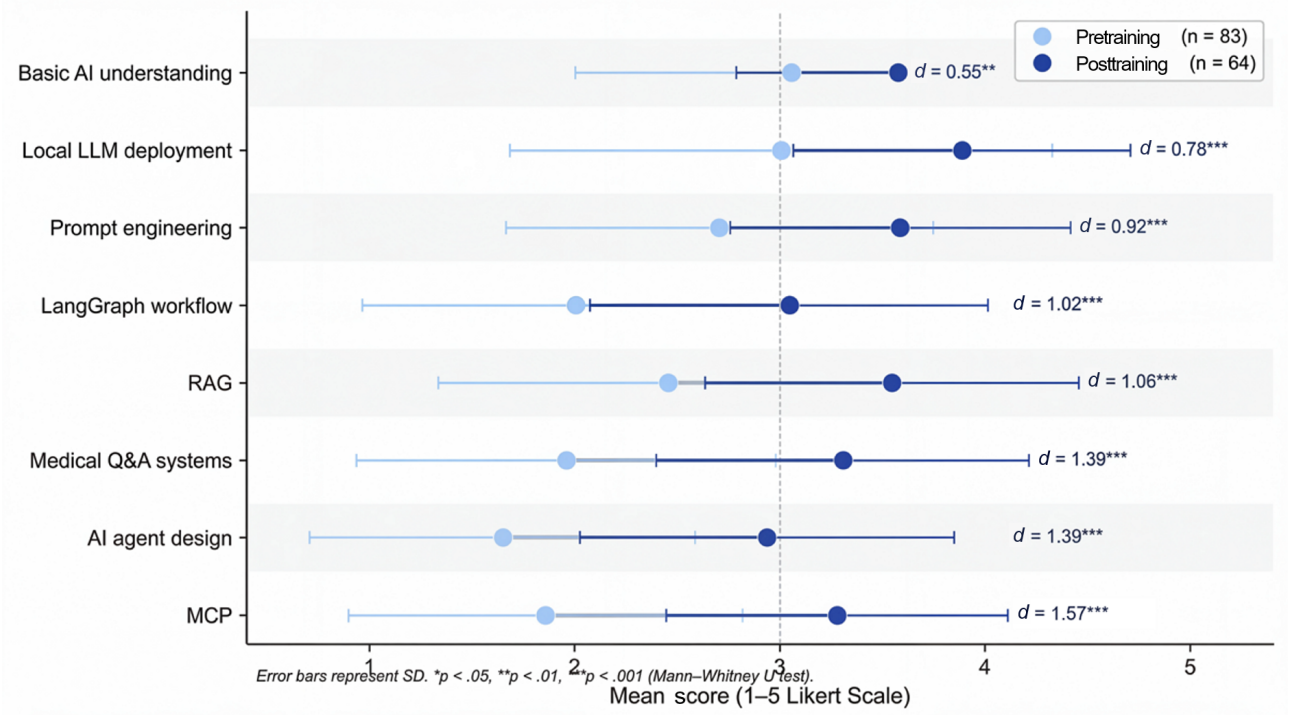
Posttraining AI knowledge self-efficacy scores were significantly higher than pretraining scores (mean 3.40, SD 0.69 vs mean 2.34, SD 0.83; U=875.5, r=.574, d=1.37; P=3.50×10^{-12}; Table 2, Figure 2). Internal consistency was excellent (α=0.909 pre, α=0.915 post; Tables S5-S6 in Multimedia Appendix 1).

At the domain level, the technologies that received the greatest curricular investment and had the lowest baseline scores showed the largest between-group differences: MCP (d=1.57), medical Q&A systems (d=1.39), and AI agent design (d=1.39), whereas foundational domains with higher baselines showed smaller differences (basic AI understanding: d=0.55; Table 2).

Table . Pre-Post comparison of primary and secondary outcomes. r=rank-biserial correlation (Rosenthal formula); significance threshold: α=0.05.

Measures	Pre mean (SD)	Post mean (SD)	U	Z	P	r	Cohen d	Interpretation
S4: AI knowledge self-efficacy	2.34 (0.83)	3.40 (0.69)	875.5	–6.956	3.50×10^{-12}	.574	1.37	Large
S6: Job-specific AI competency	1.95 (1.00)	2.77 (0.95)	1470	–4.634	3.59×10^{-}	.382	0.85	Medium
S2: AI attitudes	3.79 (0.41)	3.77 (0.51)	2733	–0.301	.764	.025	–0.04	Negligible
S5: Digital competency	3.38 (0.89)	3.63 (0.82)	2228.5	–1.675	.094	.138	0.29	Small

Figure 2. Comparison of self-assessed AI competency scores across eight skill domains before (n=83) and after (n=64) the training program, with Cohen d effect sizes and Mann-Whitney U test significance levels. MCP: model context protocol; LLM: large-language model; RAG: retrieval-augmented generation.



Subgroup Analysis by Professional Group

All 5 professional groups showed the largest between-group difference in MCP, and non-IT groups consistently showed larger differences than IT specialists (Table S2 in Multimedia Appendix 1). For MCP, clinicians showed the largest difference (d=2.88, 95% CI 1.42-4.33), followed by HIM (d=2.45, 95% CI 1.23-3.67), researchers (d=2.07, 95% CI 0.87-3.28), and administrative staff (d=1.83, 95% CI 0.55-3.11); IT specialists showed the smallest difference (d=1.36, 95% CI 0.87-1.85). These subgroup findings should be considered exploratory given small cell sizes (n=7 - 9 for non-IT groups; Table S2 in Multimedia Appendix 1).

Secondary Outcomes and the Knowledge-Practice Gap

Job-specific AI competency (S6) scores were significantly higher posttraining (mean 2.77, SD 0.95 vs mean 1.95, SD 1; P=3.59×10^{-}, r=.382, d=0.85; α=0.938-0.973 pre, α=0.877-0.981 post; Table 2). However, unlike general AI self-efficacy (S4), job-specific competency remained below the scale midpoint (2.77/5.0), indicating a knowledge-practice gap.

XSL•FO
RenderX

Three secondary pre-post comparisons did not reach significance (Table 2; Table S4 in Multimedia Appendix 1): AI attitudes (S2: r=.025; P=.76), with the mean of the 6 positively worded items (Q4-Q9) already at ceiling at baseline (M=4.29/5); the full 9-item S2 scale mean, with reverse coding applied to Q1-Q3, was 3.79 (Table 2); digital competency (S5: r=.138; P=.09); and AI daily usage (S3), where the proportion using 3+ hours daily increased modestly from 9.6% to 17.2%.

Training Satisfaction (S9; Kirkpatrick Level 1)

Overall training satisfaction was 4.03/5.0 (SD 0.85), while the domain-averaged score was 3.48 (SD 0.63; α=0.939; Table 3, Figure 3). Satisfaction varied substantially across domains: team project experience was rated highest (mean 3.98, SD 0.84), followed by instructor quality (mean 3.91, SD 0.72), teaching methods (mean 3.58, SD 0.69), and content & practice (mean 3.40, SD 0.67). Curriculum design was rated lowest (mean 2.95, SD 0.87), particularly learning pace (mean 2.91, SD 1.11). Item-level results for all posttraining scales (S9-S11) are reported in Table S3 in Multimedia Appendix 1.

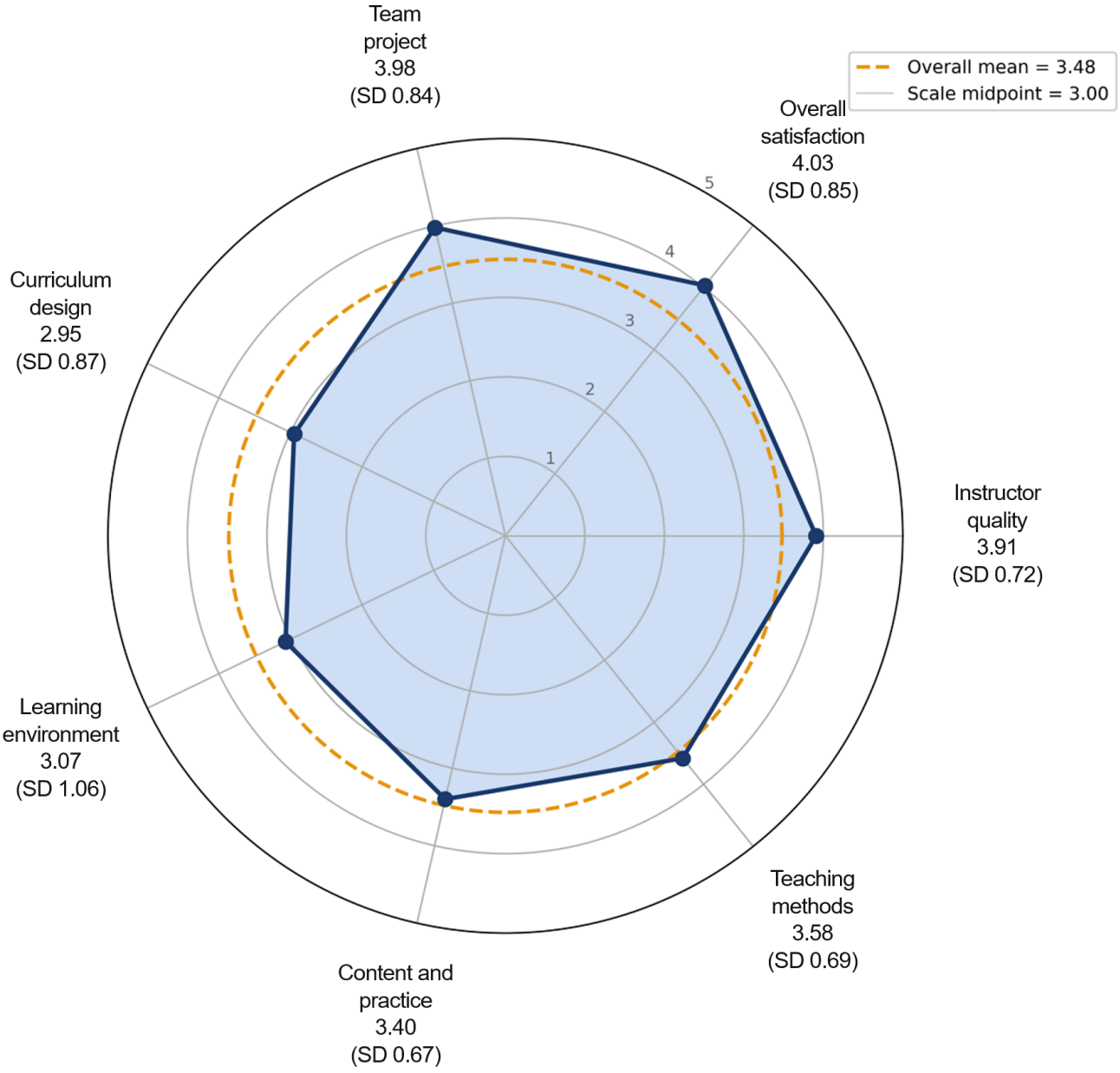
Table . Training satisfaction by domain (S9).

Domain	Mean (SD)	Cronbach α
Team project experience	3.98 (0.84)	0.789
Instructor quality	3.91 (0.72)	0.857
Teaching methods	3.58 (0.69)	0.833
Content and practice	3.40 (0.67)	0.845
Learning environment	3.07 (1.06)	0.912
Curriculum design	2.95 (0.87)	0.867
Overall (single item)	4.03 (0.85)	—[a]
Domain average	3.48 (0.63)	0.939

[a]Not available.

XSL•FO
RenderX

Figure 3. Radar chart of program satisfaction ratings across seven evaluation dimensions (N=64; overall Cronbach α=0.939, 18 items), where the dashed orange line represents the overall mean (3.48) and the gray dashed circle indicates the scale midpoint (3).



Capstone Outcomes and Real-World Deployment

Of the 12 capstone teams, 11 presented functional AI prototypes spanning clinical, administrative, and IT applications (representative examples in Table 4; full list in Table S8 in Multimedia Appendix 1); one regional affiliate team was unable to present due to scheduling constraints. One project, a medical dictionary-based RAG/query chatbot system (“WorksBot”), received the Minister of Health and Welfare Award. One prototype, the hospital document knowledge retrieval RAG system, has since been developed into an institutional service (“AI-Docs”) and has been in active pilot use since May 2026 across multiple departments, including the Department of Nursing and the Big Data Research Center, with hospital-wide deployment planned within 2026. WorksBot remains a prototype and is not yet in operational service; a refined version is planned following the RAG pilot. Both efforts are being advanced by the Department of Digital Innovation and Support.

XSL•FO
RenderX

Table . Representative capstone projects. Representative projects illustrating the range of clinical, operational, and IT applications; the complete list of all 12 capstone projects is provided in Table S8 in Multimedia Appendix 1.

Projects	Domain	Core technologies	Status
WorksBot (medical-dictionary RAG[a]/ or Q&A)[b]	Clinical knowledge access	MCP[c], RAG, vector DB[d]	Top-ranked; prototype
AI-Docs (document-retrieval RAG)	Hospital operations	MCP, RAG, vector DB	Active pilot; hospital-wide planned 2026
Natural-language clinical-data retrieval	Clinical	MCP, LangGraph	Prototype
Medical-record de-identification	Data governance or IT	MCP, local LLM[e]	Prototype
AGS[f] operational-guideline Q&A chatbot	Hospital operations	MCP, RAG, local LLM	Prototype

[a]RAG: retrieval-augmented generation.

[b]Q&A: question and answer.

[c]MCP: model context protocol.

[d]Vector DB: vector database.

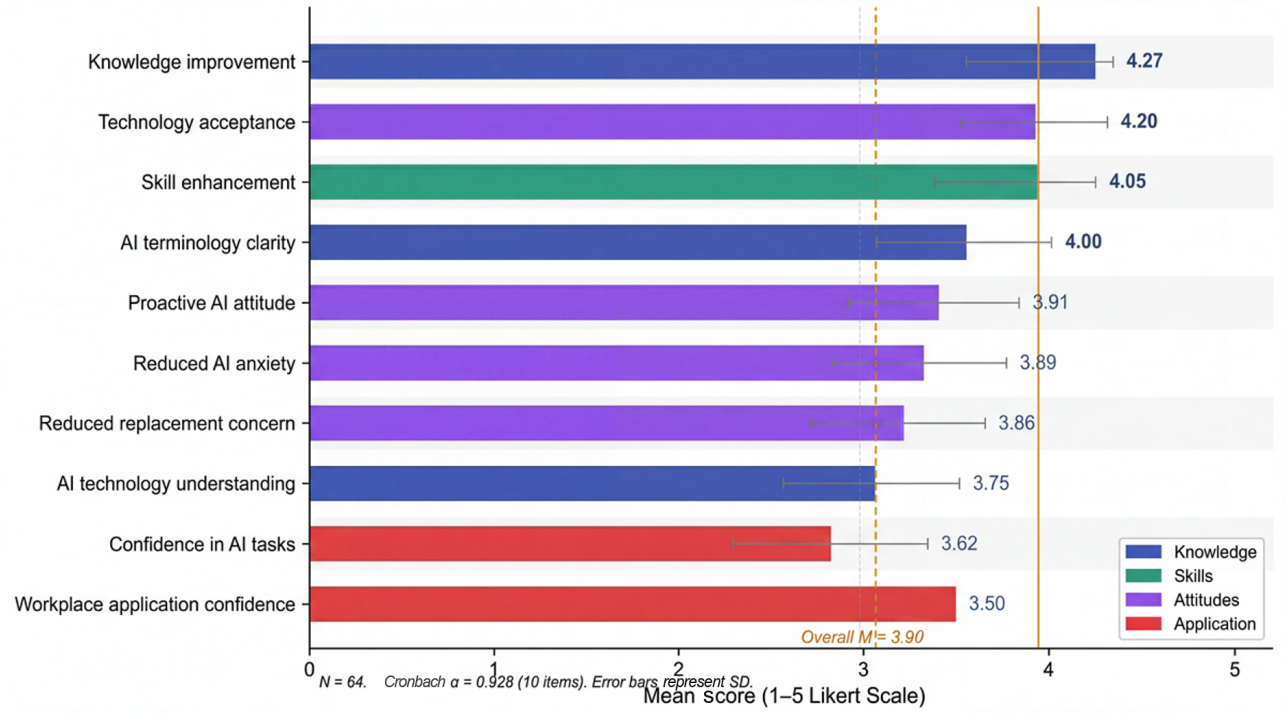
[e]LLM: large-language model.

[f]AGS: Asan Global Standard, the institution's internal operational guideline system.

Posttraining Outcomes (S10, S11; Kirkpatrick Levels 2-3)

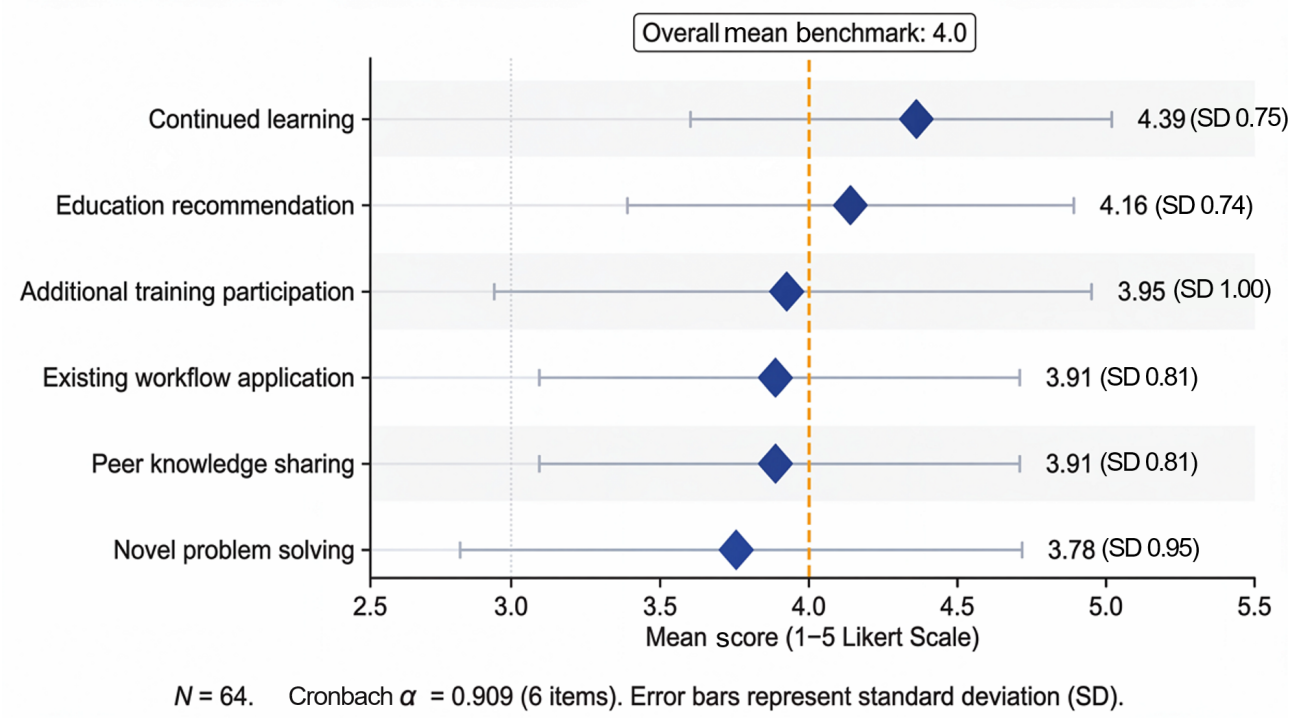
Perceived changes (S10; α=0.928; Figure 4) showed high perceived knowledge improvement (mean 4.27, SD 0.62) and technology acceptance (mean 4.20, SD 0.62), but lower workplace application confidence (mean 3.50, SD 0.84), reinforcing the knowledge-practice gap observed in S6. Workplace application intention (S11; α=0.909; Figure 5) was high overall (mean 4.02, SD 0.70), with continued learning rated highest (mean 4.39, SD 0.75) and novel problem solving rated lowest (mean 3.78, SD 0.95).

Figure 4. Self-perceived educational effectiveness across 10 items categorized into knowledge, skills, attitudes, and application domains (N=64; Cronbach α=0.928, 10 items), with an overall mean of 3.90 and error bars representing standard deviations.



XSL•FO
RenderX

Figure 5. Transfer of learning intentions across six items measured on a 5-point Likert scale (N=64; Cronbach α=0.909, 6 items), with the dashed line indicating the overall mean benchmark of 4.0. Error bars represent SD.



Participants' Self-Reported Growth (Open-Ended Responses)

All 64 posttraining respondents answered the open-ended question: "What aspect of change or growth did you most strongly feel through this 8-week program?" Eight responses (12.5%) were too brief to classify; the remaining 56 responses were grouped by the research team into 5 recurring themes. Because 8 of these responses expressed 2 distinct themes (none was assigned to more than 2), this grouping yielded 64 theme-level assignments in total. The most prevalent theme was conceptual clarity (24 assignments, 37.5% of all 64 respondents), with participants reporting that previously abstract concepts, particularly MCP and RAG, became concrete and actionable. The second was confidence and reduced fear (14 assignments, 21.9%); participants described a shift from anxiety about AI to self-efficacy, with representative responses including "the vague fear I had has been resolved to some extent" and "my first goal was confidence recovery, and I fully achieved it." Third, hands-on tool experience (12 assignments, 18.8%) was frequently cited, with participants specifically naming vibe coding, Cursor, and Claude Code as transformative experiences; notably, several nondevelopers described coding for the first time. Fourth, workplace application ideation (10 assignments, 15.6%) emerged, with participants reporting that they began identifying specific AI applications in their own work contexts. Finally, a smaller group reported a general attitude shift (4 assignments, 6.2%).

Discussion

Principal Findings

This study describes the design, implementation, and outcomes of an 8-week intensive generative AI training program for a multidisciplinary hospital workforce. To our knowledge, this is the first study to report both the design and the pre-post evaluation of a hospital-workforce program integrating 4 advanced, agent-level generative AI technologies, MCP, RAG, LangGraph orchestration, and AI agent design, across both IT and non-IT roles, spanning IT specialists, clinicians, HIM, researchers, and administrative staff. This advances AI education research beyond its prevailing focus on AI literacy and prompt engineering [10-13] toward agent-level competencies, and, through the capstone-to-deployment pathway, links classroom learning to institutional adoption. Five principal lessons emerged from the convergence of quantitative survey data and participants' open-ended reflections.

Lesson 1: Center the Curriculum on Agent-Level Technologies—The Unfamiliar Yields the Greatest Returns

The program's most consequential design decision was allocating 37% of instructional time to MCP, a technology virtually none of the participants had encountered. The quantitative data validated this bet: MCP showed the largest between-group self-efficacy difference across all professional groups (d=1.57; Table 2), and the 3 novel domains (MCP, medical Q&A, and AI agent design) consistently outperformed foundational topics. The open-ended responses corroborated this pattern—MCP and RAG were the most frequently named technologies when participants described their growth, and the dominant theme was that "previously abstract concepts became concrete and actionable." This inverse relationship between baseline familiarity and learning yield is consistent with self-efficacy theory's prediction that mastery experiences are most impactful where prior competency is lowest [19], and suggests that training programs should resist the temptation to stay at the prompt-engineering level.

XSL•FO
RenderX

Lesson 2: Include Non-IT Professionals—They Benefit Most and Bring Irreplaceable Domain Expertise

Non-IT professionals showed consistently larger between-group differences than IT specialists across most domains (Table S2 in Multimedia Appendix 1), particularly for MCP (clinicians: d=2.88; HIM: d=2.45). This finding challenges the assumption that advanced AI training should be restricted to technical staff. Qualitatively, nondevelopers described transformative experiences: several reported coding for the first time through vibe coding, and one described the capstone as an opportunity to "bring MCP and RAG into hospital work and implement visible results through hands-on practice—a fundamentally different experience from watching lecture slides." The mixed-expertise team structure, which received the highest satisfaction rating (mean 3.98, SD 0.84), provided the technical scaffolding non-IT professionals needed while leveraging their domain knowledge for health care-specific applications.

Lesson 3: Expect a Knowledge-Practice Gap—and Design for It

Despite significant self-efficacy gains (r=.574) and high perceived knowledge improvement (mean 4.27, SD 0.62), job-specific competency remained below the scale midpoint (2.77/5) and workplace application confidence was rated only 3.50—a gap that converged across 3 independent measures (S6, S10, and S11). Yet, the open-ended responses revealed a more nuanced picture: 15.6% of respondents spontaneously reported generating specific workplace application ideas, and participants described shifts from passive understanding to active problem identification ("I started thinking about improvements to the systems we currently operate"). The Unified Theory of Acceptance and Use of Technology (UTAUT) framework [22] suggests that training without corresponding organizational change management may produce motivated but unsupported learners [34,35]; health care settings compound this through restricted electronic medical record (EMR) environments, data governance requirements, and concerns about LLM hallucination. The gap is not a failure of training but a predictable transition challenge—one that infrastructure support (funded tool subscriptions, secure internal cloud environments) and structured posttraining mentoring can help bridge [36,37].

Lesson 4: Plan for Differentiated Pacing Across Heterogeneous Cohorts

Curriculum pacing was rated lowest among satisfaction domains (mean 2.91/5.0, SD 1.11), even as overall satisfaction was substantially higher (4.03/5.0). This signals that participants valued the program but struggled to keep pace—a predictable tension when teaching rapidly evolving technologies to professionals ranging from experienced developers to clinical staff encountering code for the first time. The qualitative data illuminated both sides: some participants celebrated "coding fearlessly for the first time," while the pacing score suggests others felt left behind. In practice, hands-on exercises were already split into application and development tracks matched to baseline skill; future iterations should additionally consider prerequisite assessment and supplementary catch-up sessions to accommodate varying baseline competencies.

Teaching a mixed-expertise cohort posed distinct challenges. IT and health-information staff were released from their regular duties to attend, whereas some clinicians and researchers used personal annual leave. At the outset, the wide range in baseline skill, from staff who had never coded to experienced IT specialists, was anticipated as a key challenge; 2 design choices mitigated it (Methods): the curriculum minimized AI theory in favor of practical concepts, agentic methods, and concrete MCP use cases, and each capstone team paired an intact IT-subdepartment core with non-IT domain members. Collaboration across backgrounds still required deliberate team-level scaffolding, and recruitment skewed toward IT staff, consistent with the program's IT-first prioritization and department-based, voluntary enrollment—with comparatively fewer clinical participants.

Lesson 5: Capstone Projects Bridge Instruction and Institutional Adoption

The capstone component, supported by 3 - 6 hours of weekly mentoring per team from a dedicated instructor, culminated in functional prototypes from 11 of 12 teams. The qualitative data revealed why: participants described the capstone as where abstract knowledge "became real"—the transition from understanding concepts to building something that works. More importantly, this bridge extended beyond the classroom: one capstone project has since advanced into institutional pilot use (AI-Docs; see Results). This progression from capstone prototype into pilot clinical use provides emerging Kirkpatrick level 3 evidence that training-generated projects can move into institutional adoption, suggesting the capstone model can serve as a pathway for AI innovation that outlasts the training period itself.

This program was conceived as the deliberate first phase of a staged, institution-wide strategy: we began with staff positioned to enable broad adoption (IT, health-information, related-department clinicians, and researchers), expecting their agent-level competency to seed hospital-wide AI deployment—a trajectory already visible in the AI-Docs pilot. Subsequent phases are planned to extend structured training to frontline clinicians and other departments and to embed the curriculum as a recurring continuing professional development (CPD) track, a required or elective module, or a template for other hospitals through the national framework that funded it. So framed, this voluntary first cohort is a starting point, not an endpoint: audiences not yet aware of agent-level AI are reached through later, role-targeted phases rather than left behind.

Comparison With Prior Work

The observed between-group difference in AI knowledge self-efficacy (r=.574) is larger than improvements typically reported in prior digital health education studies, which have largely relied on uncontrolled before-after designs and reported heterogeneous, generally smaller gains [38]. The larger effects likely reflect the intensive, hands-on design (71% of on-site time practical; 84% experiential including the capstone), consistent with experiential learning evidence [19,23]. Recent ChatGPT education studies for health care professionals report similar directions but with shorter interventions and smaller effects [11,12]; our previous health informatics analyst education

XSL•FO
RenderX

program at the same institution [18] provides a direct comparator, and the present study extends that evidence to advanced generative AI technologies.

The absence of significant attitudinal change (S2; *P*=.76) aligns with ceiling effects in health care AI attitude studies [13,39], where voluntary enrollees already hold favorable views (4.29/5). Notably, however, the qualitative data captured attitudinal shifts that the Likert scale missed: 21.9% of respondents spontaneously described reduced fear or increased confidence as their most significant growth—suggesting that self-efficacy gains may be a more sensitive indicator of training impact than attitudinal scales in already AI-positive populations. Future studies should recruit participants with more heterogeneous baseline attitudes [40] and consider AI adoption readiness [21] as an alternative construct.

Limitations

This study has several important limitations that should be considered when interpreting the findings.

Study Design Constraints

The independent samples design, necessitated by anonymous data collection, prevents individual-level tracking, precluding within-person effect sizes; future studies should consider self-generated identification codes to enable paired analysis [41]. The absence of a control group precludes causal attribution; improvements may partly reflect maturation or concurrent self-study. Additionally, the inverse relationship between baseline scores and improvement magnitude may partly reflect regression to the mean, particularly in this design where groups are not matched at the individual level. Subgroup comparisons by professional group also warrant caution beyond their exploratory, small-cell nature, because professional role was confounded with both baseline competency and career stage. The groups with the lowest baseline self-efficacy—HIM (pre mean 1.74, SD 0.44) and clinicians (1.92)—showed the largest gains, consistent with regression to the mean and ceiling effects, whereas researchers, despite being the most early-career group (none with ≥11 years of experience), reported the highest baseline and posttraining scores. Professional roles were, moreover, self-classified. The larger gains observed for non-IT groups therefore cannot be attributed to professional background independent of baseline ability and experience.

Attrition and Selection Bias

Of 83 enrollees, 19 (22.9%) did not respond to the posttraining survey, and these individuals may have differed systematically from respondents. A worst-case sensitivity analysis demonstrated that primary outcomes remain statistically significant even assuming zero training effect in all nonrespondents (Table S7 in Multimedia Appendix 1). Nevertheless, self-selection bias limits generalizability: voluntary enrollees at affiliated institutions with high baseline positive-attitude scores (4.29/5) may not represent the broader health care workforce. Importantly, posttraining nonresponse was not evenly distributed across professional groups: retention was lowest among IT specialists (33 of 48; 68.8%) and higher among non-IT groups (78% - 100%). Because the subgroup analyses contrast non-IT groups against IT specialists, this differential attrition may modestly inflate the apparent non-IT advantage, and the comparable cross-sectional occupational distributions should be interpreted with caution given limited statistical power.

Measurement Limitations

All outcomes relied on self-report measures subject to social desirability bias [42]. Self-efficacy assessments administered immediately after training may be inflated by posttraining euphoria, and the Dunning-Kruger effect is particularly relevant for novel technology domains where participants may lack sufficient expertise to calibrate their self-assessment accurately. Future studies should complement self-efficacy measures with objective performance assessments [26]. The survey instrument was developed without formal content validity index (CVI) assessment or pilot testing, limiting construct validity evidence beyond the post hoc psychometric analyses reported in Tables S5-S6 in Multimedia Appendix 1; future iterations should establish CVI ≥.80 via a priori expert panel review.

Generalizability and Temporal Scope

Despite multisite delivery, participants were drawn primarily from a single tertiary academic medical center, limiting external validity. Short-term assessment does not capture whether gains persist; longitudinal follow-up at 3 - 6 months is needed to assess durability and actual behavioral transfer [43].

Conclusions

This study suggests that transforming a multidisciplinary hospital workforce into AI-capable professionals is achievable through intensive, hands-on training centered on agent-level technologies. The curriculum's investment in MCP, the technology anticipated to underpin future health care AI infrastructure, was associated with the largest self-efficacy gains, and non-IT professionals showed the largest between-group differences, challenging the assumption that advanced AI training should be restricted to technical staff. The team-based capstone model proved particularly valuable, with prototypes that progressed toward institutional adoption, providing early evidence that training programs can serve as a pathway for institutional AI innovation. However, the knowledge-practice gap identified across multiple measures underscores that training alone is insufficient; posttraining support structures, mentoring, communities of practice, and supervised implementation are essential to translate self-efficacy gains into sustained workplace practice. Future research should use longitudinal designs with matched cohorts to assess the durability of training effects and to determine whether capstone-to-deployment pathways can be replicated across institutions.

XSL•FO
RenderX

Acknowledgments

The authors would like to express their sincere gratitude to the Asan Medical Center Big Data Research Center, the Department of Digital Innovation and Support, and the Asan Academic Institute for their support in operating the educational program, and to the Korea Human Resource Development Institute for Health & Welfare for their support as the supervising organization of this educational project. The authors also thank all participants who took part in the educational program, including IT professionals, medical record administrators, health care professionals, researchers, and administrative or clerical staff.

During the preparation of this manuscript, the authors used Claude (Anthropic; Claude Opus model) to assist with English-language editing and phrasing, to support drafting of the manuscript text, and to help organize the references and formatting. The tool was not used to generate, analyze, or interpret the study data or results. After using this tool, the authors reviewed and edited all content as needed and take full responsibility for the content of the publication.

Funding

This research was supported by (1) a grant of the Research-Centered Hospital Development R&D Project through the Korea Health Industry Development Institute (KHIDI), funded by the Ministry of Health and Welfare, Republic of Korea (grant number: HR21C0198); and (2) a grant of the Korea Health Technology R&D Project, funded by the Ministry of Health and Welfare, Republic of Korea (grant number: RS-2025 - 02213531).

Data Availability

The datasets generated during this study are available from the corresponding author upon reasonable request.

Authors' Contributions

GKB, Kye Hwa Lee, and DHY contributed to the conceptualization. GKB and HNL contributed to data curation. HNL, YRL, and GKB contributed to the formal analysis. MSK, Kun Hee Lee, and MJC contributed to data acquisition. GKB and Kye Hwa Lee contributed to writing—original draft. Kye Hwa Lee and GKB contributed to writing—review and editing.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Tables S1-S8 (S1: training curriculum by week; S2: S4 full domain statistics by subgroup with 95% CIs; S3: posttraining item-level results; S4: post-hoc power analysis; S5: exploratory factor analysis results; S6: item-total correlation analysis; S7: worst-case nonresponse sensitivity analysis; S8: capstone project descriptions and outcomes).
[DOCX File, 50 KB - mededu_v12i1e97822_app1.docx]

Multimedia Appendix 2
Detailed session-level training syllabus (per-session learning objectives, hands-on exercises, and tools).
[DOCX File, 28 KB - mededu_v12i1e97822_app2.docx]

Multimedia Appendix 3
Capstone project evaluation rubric (standardized 100-point rubric used by the assessment panel).
[DOCX File, 26 KB - mededu_v12i1e97822_app3.docx]

Multimedia Appendix 4
Open-ended survey responses (English translations of all 64 free-text responses to the posttraining self-perceived growth question).
[DOCX File, 29 KB - mededu_v12i1e97822_app4.docx]

Checklist 1
STROBE checklist for cross-sectional studies.
[DOCX File, 22 KB - mededu_v12i1e97822_app5.docx]

References

1. Seo J, Choi D, Kim T, et al. Evaluation framework of large language models in medical documentation: development and usability study. J Med Internet Res 2024 Nov 20;26:e58329. [doi: 10.2196/58329] [Medline: 39566044]
2. Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. Nat Med 2019 Jan;25(1):44-56. [doi: 10.1038/s41591-018-0300-7] [Medline: 30617339]

XSL•FO
RenderX

3. Thirunavukarasu AJ, Ting DSJ, Elangovan K, Gutierrez L, Tan TF, Ting DSW. Large language models in medicine. Nat Med 2023 Aug;29(8):1930-1940. [doi: 10.1038/s41591-023-02448-8] [Medline: 37460753]
4. He J, Baxter SL, Xu J, Xu J, Zhou X, Zhang K. The practical implementation of artificial intelligence technologies in medicine. Nat Med 2019 Jan;25(1):30-36. [doi: 10.1038/s41591-018-0307-0] [Medline: 30617336]
5. Lewis P, Perez E, Piktus A, Petroni F, Karpukhin V, Goyal N, et al. Retrieval-augmented generation for knowledge-intensive NLP tasks. 2020 Presented at: 34th Conference on Neural Information Processing Systems (NeurIPS 2020); Dec 6-12, 2020 p. 9459-9474 URL: https://proceedings.neurips.cc/paper/2020/file/6b493230205f780e1bc26945df7481e5-Paper.pdf [accessed 2026-07-21]
6. What is the model context protocol (MCP)? Model Context Protocol. 2024. URL: https://modelcontextprotocol.io/ [accessed 2025-09-08]
7. LangGraph overview. LangChain Docs. 2024. URL: https://docs.langchain.com/oss/python/langgraph/overview [accessed 2025-09-08]
8. Sallam M. ChatGPT utility in healthcare education, research, and practice: systematic review on the promising perspectives and valid concerns. Healthcare (Basel) 2023 Mar 19;11(6):887. [doi: 10.3390/healthcare11060887] [Medline: 36981544]
9. Laupichler MC, Aster A, Schirch J, Raupach T. Artificial intelligence literacy in higher and adult education: a scoping literature review. Comput Educ Artif Intell 2022;3:100101. [doi: 10.1016/j.caeai.2022.100101]
10. Gordon M, Daniel M, Ajiboye A, et al. A scoping review of artificial intelligence in medical education: BEME Guide No. 84. Med Teach 2024 Apr;46(4):446-470. [doi: 10.1080/0142159X.2024.2314198] [Medline: 38423127]
11. Wu Y, Zheng Y, Feng B, Yang Y, Kang K, Zhao A. Embracing ChatGPT for medical education: exploring its impact on doctors and medical students. JMIR Med Educ 2024 Apr 10;10:e52483. [doi: 10.2196/52483] [Medline: 38598263]
12. Hu JM, Liu FC, Chu CM, Chang YT. Health care trainees' and professionals' perceptions of ChatGPT in improving medical knowledge training: rapid survey study. J Med Internet Res 2023 Oct 18;25:e49385. [doi: 10.2196/49385] [Medline: 37851495]
13. Tangadulrat P, Sono S, Tangtrakulwanich B. Using ChatGPT for clinical practice and medical education: cross-sectional survey of medical students' and physicians' perceptions. JMIR Med Educ 2023 Dec 22;9:e50658. [doi: 10.2196/50658] [Medline: 38133908]
14. Sapci AH, Sapci HA. Artificial intelligence education and tools for medical and health informatics students: systematic review. JMIR Med Educ 2020 Jun 30;6(1):e19285. [doi: 10.2196/19285] [Medline: 32602844]
15. Preiksaitis C, Rose C. Opportunities, challenges, and future directions of generative artificial intelligence in medical education: scoping review. JMIR Med Educ 2023 Oct 20;9:e48785. [doi: 10.2196/48785] [Medline: 37862079]
16. Charow R, Jeyakumar T, Younus S, et al. Artificial intelligence education programs for health care professionals: scoping review. JMIR Med Educ 2021 Dec 13;7(4):e31043. [doi: 10.2196/31043] [Medline: 34898458]
17. Lee YM, Kim S, Lee YH, et al. Defining medical AI competencies for medical school graduates: outcomes of a delphi survey and medical student/educator questionnaire of South Korean medical schools. Acad Med 2024 May 1;99(5):524-533. [doi: 10.1097/ACM.0000000000005618] [Medline: 38207056]
18. Lee KH, Lee JH, Lee Y, et al. Impact of health informatics analyst education on job role, career transition, and skill development: survey study. JMIR Med Educ 2024 Sep 25;10:e54427. [doi: 10.2196/54427] [Medline: 39320368]
19. Bandura A. Self-Efficacy: The Exercise of Control: W.H. Freeman; 1997.
20. Compeau DR, Higgins CA. Computer self-efficacy: development of a measure and initial test. MIS Q 1995 Jun 1;19(2):189-211. [doi: 10.2307/249688]
21. Davis FD. Perceived usefulness, perceived ease of use, and user acceptance of information technology. MIS Q 1989 Sep 1;13(3):319-340. [doi: 10.2307/249008]
22. Venkatesh V, Morris MG, Davis GB, Davis FD. User acceptance of information technology: toward a unified view1. MIS Q 2003 Sep 1;27(3):425-478. [doi: 10.2307/30036540]
23. Kolb DA. Experiential Learning: Experience as the Source of Learning and Development, 2nd edition: Pearson Education; 2015.
24. Kirkpatrick DL. Evaluating training programs: evidence vs. proof. Train Dev J 1977;31(11):9-12 [FREE Full text]
25. Kirkpatrick JD, Kirkpatrick WK. Kirkpatrick's Four Levels of Training Evaluation: ATD Press; 2016.
26. Yardley S, Dornan T. Kirkpatrick's levels and education "evidence". Med Educ 2012 Jan;46(1):97-106. [doi: 10.1111/j.1365-2923.2011.04076.x] [Medline: 22150201]
27. Alhassan AI. Implementing faculty development programs in medical education utilizing Kirkpatrick's model. Adv Med Educ Pract 2022;13:945-954. [doi: 10.2147/AMEP.S372652] [Medline: 36039186]
28. von Elm E, Altman DG, Egger M, et al. The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies. Lancet 2007 Oct 20;370(9596):1453-1457. [doi: 10.1016/S0140-6736(07)61602-X] [Medline: 18064739]
29. Wang L, Ma C, Feng X, et al. A survey on large language model based autonomous agents. Front Comput Sci 2024 Dec;18(6):186345. [doi: 10.1007/s11704-024-40231-1]
30. Hollander M, Wolfe DA, Chicken E. Nonparametric Statistical Methods, 3rd edition: John Wiley & Sons; 2014. [doi: 10.1002/9781119196037]

XSL•FO
RenderX

31. Rosenthal R. Meta-Analytic Procedures for Social Research: Sage; 1991.
32. Cohen J. Statistical Power Analysis for the Behavioral Sciences, 2nd edition: Lawrence Erlbaum Associates; 1988.
33. Nunnally JC, Bernstein IH. Psychometric Theory, 3rd edition: McGraw-Hill; 1994.
34. Hassan M, Kushniruk A, Borycki E. Barriers to and facilitators of artificial intelligence adoption in health care: scoping review. JMIR Hum Factors 2024 Aug 29;11:e48633. [doi: 10.2196/48633] [Medline: 39207831]
35. Greenhalgh T, Wherton J, Papoutsi C, et al. Beyond adoption: a new framework for theorizing and evaluating nonadoption, abandonment, and challenges to the scale-up, spread, and sustainability of health and care technologies. J Med Internet Res 2017 Nov 1;19(11):e367. [doi: 10.2196/jmir.8775] [Medline: 29092808]
36. Pfeffer J, Sutton RI. The Knowing-Doing Gap: How Smart Companies Turn Knowledge into Action: Harvard Business School Press; 2000. [doi: 10.1108/scm.2001.6.3.142.1]
37. Grimshaw JM, Eccles MP, Lavis JN, Hill SJ, Squires JE. Knowledge translation of research findings. Implement Sci 2012 May 31;7:50. [doi: 10.1186/1748-5908-7-50] [Medline: 22651257]
38. Tudor Car L, Kyaw BM, Nannan Panday RS, et al. Digital health training programs for medical students: scoping review. JMIR Med Educ 2021 Jul 21;7(3):e28275. [doi: 10.2196/28275] [Medline: 34287206]
39. Chen M, Zhang B, Cai Z, et al. Acceptance of clinical artificial intelligence among physicians and medical students: a systematic review with cross-sectional survey. Front Med (Lausanne) 2022;9:990604. [doi: 10.3389/fmed.2022.990604]
40. Wartman SA, Combs CD. Reimagining medical education in the age of AI. AMA J Ethics 2019 Feb 1;21(2):E146-E152. [doi: 10.1001/amajethics.2019.146] [Medline: 30794124]
41. Yurek LA, Vasey J, Sullivan Havens D. The use of self-generated identification codes in longitudinal research. Eval Rev 2008 Oct;32(5):435-452. [doi: 10.1177/0193841X08316676] [Medline: 18477737]
42. Podsakoff PM, MacKenzie SB, Lee JY, Podsakoff NP. Common method biases in behavioral research: a critical review of the literature and recommended remedies. J Appl Psychol 2003 Oct;88(5):879-903. [doi: 10.1037/0021-9010.88.5.879] [Medline: 14516251]
43. Salas E, Tannenbaum SI, Kraiger K, Smith-Jentsch KA. The science of training and development in organizations: what matters in practice. Psychol Sci Public Interest 2012 Jun;13(2):74-101. [doi: 10.1177/1529100612436661] [Medline: 26173283]

Abbreviations

BEME: Best Evidence Medical Education
CPD: continuing professional development
CVI: content validity index
EFA: exploratory factor analysis
EMR: electronic medical record
HIM: health information manager
KOHI: Korea Human Resource Development Institute for Health & Welfare
LLM: large language model
MCP: model context protocol
Q&A: question and answer
RAG: retrieval-augmented generation
RQ: research question
STROBE: Strengthening the Reporting of Observational Studies in Epidemiology
UTAUT: Unified Theory of Acceptance and Use of Technology

Edited by A Stone; submitted 13.Apr.2026; peer-reviewed by D Yoon, J Krive; revised version received 10.Jul.2026; accepted 10.Jul.2026; published 11.Sep.2026.

Please cite as:
Baek G, Lee H, Yang DH, Kang M, Lee KH, Choi M, Lee Y, Lee KH
Teaching Model Context Protocol, Retrieval-Augmented Generation, and AI Agents to a Multidisciplinary Hospital Workforce: Single-Group Pre-Post Survey Study
JMIR Med Educ 2026;12:e97822
URL: https://mededu.jmir.org/2026/1/e97822
doi:10.2196/97822

© Gakyoung Baek, Hyunna Lee, Dong Hyun Yang, Minseo Kang, Kun Hee Lee, Minji Choi, Yura Lee, Kye Hwa Lee. Originally published in JMIR Medical Education (https://mededu.jmir.org), 11.Sep.2026. This is an open-access article distributed under

XSL•FO
RenderX

the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Education, is properly cited. The complete bibliographic information, a link to the original publication on https://mededu.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Perceptions, Attitudes, and Use of AI by Medical Students: Mixed Methods Study

Frédéric Paris[1], MD; Vincent Garrouste[2], MD; Laure Abensur Vuillaume[3], MD, PhD

[1]University Hospital of Geneva, Rue Gabrielle-Perret-Gentil 4, Geneva, Switzerland

[2]Centre hospitalier universitaire d'Orléans, Orléans, Centre-Val de Loire, France

[3]Centre Hospitalier Régional de Metz-Thionville, Metz-Thionville, Grand Est, France

Corresponding Author:
Frédéric Paris, MD
University Hospital of Geneva, Rue Gabrielle-Perret-Gentil 4, Geneva, Switzerland

Abstract

Background: AI is transforming medicine by enhancing care, reducing administrative tasks, and facilitating research. AI also raises many concerns, including a lack of clinical context awareness, data dependence, and the absence of ethical judgment. As future practitioners, medical students must be prepared for these changes. Most studies assessing students' attitudes and knowledge were conducted before AI became accessible and tailored to the needs of the population. Therefore, how medical students actually use AI remains largely unexplored.

Objective: This study aimed to explore French medical students' perceptions, attitudes, and use of AI.

Methods: A mixed methods study was conducted in 2025 among French medical students in their clerkship year. An online survey included open-ended questions about the definition of AI and feelings toward AI, a Likert scale item to assess specific attitudes, and multiple-choice questions about student characteristics. Quantitative analysis was performed using Kruskal-Wallis tests, chi-square tests, and exploratory multivariable linear regression to examine associations between AI knowledge, attitudes, and student characteristics. Qualitative thematic analysis was conducted inductively on open-text responses regarding perceptions of AI, feelings, training expectations, and use.

Results: Of the 1377 responses received, 1342 were included. Students had a median age of 23 (IQR 22-24) years and were predominantly in their fifth year. Only 5% (67/1342) provided a correct definition of AI, while 57.4% (770/1342) gave incorrect responses. Attitudes toward AI were generally positive, with a median score of 7 (IQR 5-8). Students with unknown AI definitions had significantly lower attitude scores (P=.02), although the magnitude of the difference was small (ε^2=0.005). In multivariable analysis, belonging to the unknown AI definition category remained associated with a lower general attitude toward AI score compared with the incorrect category (regression coefficient β=−0.72, 95% CI −1.17 to −0.28; P=.001). Regarding education, 48.5% (651/1342) of students preferred AI training outside the formal curriculum. Qualitative analysis revealed 5 themes: representation, nuanced optimism, critical consideration, replacement, and AI use. Students describe AI as a robot, an improved search engine, or an unlimited data source. Their nuanced optimism blends enthusiasm for efficient patient care and the opportunity to focus more on the patient relationship, with concerns about dehumanization, energy costs, and skill regression. Critical consideration underscores distrust from ethical dilemmas and data security risks. Replacement concerns arise from shifting professional roles, though many believe human empathy remains irreplaceable. Regarding AI use, students highlight its potential for administrative aid, personalized training, and clinical support.

Conclusions: Medical students report generally positive but varied attitudes toward AI, despite having a limited understanding of its foundations. Ecological concerns, fears of skill loss, and ethical issues coexist with widespread self-directed use of AI for learning. These findings provide a descriptive baseline for future evaluations of AI-related training, particularly regarding critical appraisal, ethical issues, and self-directed AI use.

(JMIR Med Educ 2026;12:e91345) doi:10.2196/91345

KEYWORDS

artificial intelligence; AI; medical education; attitude; knowledge; students; curriculum design

Introduction

AI traditionally refers to "any machine or algorithm capable of observing its environment, learning, and, through its experiences and acquired knowledge, performing intelligent actions or making decisions" [1]. It is recognized as potentially having a significant impact on care delivery and clinical practices [2]. Indeed, AI, particularly through deep learning applications, has shown significant potential to improve diagnosis, reduce

XSL•FO
RenderX

administrative tasks, and facilitate research [3]. However, AI appears to have limitations in this field, particularly a lack of contextual understanding, data dependence, an absence of ethical judgment, and difficulty innovating beyond existing knowledge [4]. Historically, the first medical applications of AI were studied in radiology, infectious diseases, and ophthalmology [5]. Since 2020, the scope of AI applications has been constantly expanding across various medical specialties, such as reducing administrative burden in emergency departments through AI systems like "KATE," developed to reduce triage delays, and "SCLEPIOS," which can generate consultation reports and perform coding in emergency settings [6]. This phenomenon is also observed in health education across various fields, including clinical supervision, recruitment, clinical reasoning, intelligent tutoring systems, and evaluation [7]. These innovations, coupled with the rise of educational chatbots, illustrate the growing integration of intelligent tools in clinical training and practice [8].

Medical students are directly impacted by the use of AI in clinical settings. Moreover, several concerns related to the use of AI are progressively emerging [9-11]. These concerns include the potential decline of certain skills, the lack of acquisition of specific competencies, and an increasing dependence on digital tools [9,10]. Such dependence, along with the risk of skill degradation, may affect students' abilities in synthesis and critical thinking. However, these risks remain poorly documented in clinical contexts to date [9-12].

Therefore, understanding their use, attitudes, and comprehension of AI during their studies seems essential to adapt educational programs and prepare future physicians to work effectively with these new technologies [13,14]. Early studies exploring medical students' attitudes focused on radiology [15]. Authors showed a decreased interest in radiology as a specialty among students due to fear of being replaced by AI [16-18]. Subsequently, several studies also explored medical students' knowledge and attitudes about AI [7,19,20]. It was shown that students had limited knowledge of AI [21,22], with media being their main source of learning [23]. Regarding their attitudes, students recognized AI's usefulness for diagnosis, research, and administrative task assistance. However, they expressed concerns about ethical issues, the weakening of doctor-patient relationships, and AI's limitations in unexpected scenarios [24-27]. Finally, one of the main themes among students was anxiety about being replaced by AI [24]. These attitudes may be influenced by students' level of education and prior knowledge [28]. Moreover, students' intention to use AI appears to be primarily influenced by social factors rather than AI's intrinsic performance [29].

However, we found no studies evaluating French medical students' knowledge and attitudes about AI. Additionally, most studies were conducted several years ago, when AI use was less widespread, and there were few concrete clinical applications. The attitudes and knowledge of AI, influenced by the recent and increasing use of AI by medical students, remain largely understudied today [30,31]. Finally, the use of AI among medical students remains underexplored [32]. Therefore, the knowledge and attitudes of students who have used AI since high school need to be explored [33]. The aim of this study was to explore French medical students' perceptions, attitudes, and use of AI in 2025.

Methods

We conducted a mixed methods qualitative and quantitative study using a survey that consisted of multiple-choice questions, a Likert scale, and short open-ended questions (SAQs). This study used an existing, validated survey originally developed by Teng et al [21].

Population

To be included, participants had to be medical students aged 18 years or older, enrolled in a French medical school, and in their fourth to sixth academic years of study. These students are in the second cycle of medical studies, which corresponds to the clerkship year. Their training is divided between university teaching and clinical rotations in university hospitals, peripheral hospitals, and general practice offices.

Exclusion criteria included refusal to participate, incomplete questionnaire responses, and duplicate email addresses to prevent redundancy.

Recruitment

The recruitment process involved several steps to obtain a representative sample of medical students in France.

First, an email containing an informational letter was sent to all medical schools in France. Then, dissemination was conducted through the social networks of the student association.

With approximately 25,000 medical students in 35 faculties hosting clinical students, our goal was to achieve a 10% participation rate, or 2500 students.

To motivate students to respond to the questionnaire, we offered students who completed the questionnaire the opportunity to participate in a contest to win one of five €20 (US $22.86) Amazon or Fnac gift cards.

Survey

We used a published survey to assess the perceptions and attitudes of medical students regarding AI [21].

The questionnaire consisted of 3 sections (Multimedia Appendix 1): the first section asked about the participants' personal characteristics. The second section was an SAQ where participants were asked to describe what AI is. Finally, the last section consisted of several questions with Likert scale responses ranging from 1 to 10 about their attitude toward AI, as well as an SAQ about their feelings toward AI.

The original survey was designed for all Canadian health students. We moved the first section to the last to better meet the needs and interests of medical students (Multimedia Appendix 1). To adapt the questionnaire to our target population, we modified the institution question to ask about their medical school in France, their academic year, their self-assessment of their class ranking, and their career aspirations.

The survey was tested by 6 medical students to verify comprehension, identify errors, and assess completion time.

XSL•FO
RenderX

The medical students, organized into 3 pairs, were informed of the study's objectives and were encouraged to verbalize any confusion, identify errors, or request clarifications at any point during the process. During the pilot test, the first pair of participants highlighted 2 issues: a duplicated question, which was subsequently removed, and a multiple-choice question that was initially designed to allow only 1 response despite requiring multiple selections, which was corrected to permit multiple answers. After these revisions were made, the remaining 2 pairs completed the survey without encountering any further errors or comprehension difficulties. Importantly, none of the 6 participants reported any confusion or requested additional explanations, indicating that the questions were clear and well understood. The average time required to complete the survey was observed to be 5 minutes, confirming that the questionnaire was both concise and comprehensive for the intended audience.

Data Analysis

Continuous and score variables were described using medians and IQRs, whereas categorical variables were summarized as counts and percentages. Questions allowing multiple responses were analyzed as separate binary variables for each response option; therefore, percentages may exceed 100%.

The primary grouping variable was the final AI definition category, obtained after adjudicating the free-text definitions according to the Organisation for Economic Co-operation and Development (OECD)–based coding framework: An AI system is a machine-based system that, for explicit or implicit objectives, infers from the input it receives how to generate outputs such as predictions, content, recommendations, or decisions that can influence physical or virtual environments. Different AI systems vary in their levels of autonomy and adaptiveness after deployment [34]. Interrater reliability between the 2 independent coders was assessed using raw agreement and weighted Cohen κ. Because AI definition categories were ordinal (unknown, incorrect, partially correct, and correct), weighted κ coefficients were calculated using both linear and quadratic weights. Bootstrap resampling was used to estimate 95% CIs for weighted κ.

Comparisons across final AI definition categories were performed using the Kruskal-Wallis test for continuous or score variables and the chi-square test for categorical variables. Effect sizes were reported for all comparisons: epsilon-squared for Kruskal-Wallis tests and Cramér *V* for chi-square tests. For variables with a significant Kruskal-Wallis test, pairwise post hoc Dunn tests were performed with Holm correction for multiple comparisons.

The general attitude toward AI score (0 - 10) was analyzed as the main quantitative outcome. Six additional AI attitude items were analyzed separately as secondary outcomes. To limit multiplicity, *P* values from the 6 omnibus Kruskal-Wallis tests were adjusted using the Holm method. Post hoc Dunn-Holm comparisons were only interpreted for outcomes with a significant omnibus result after correction.

As an exploratory multivariable analysis, a linear regression model with heteroskedasticity-consistent robust SEs was fitted using the general attitude toward AI score as the dependent variable. Covariates included the final AI definition category, age, academic year, class ranking, and career aspirations (hospital-based clinical work, research, and private practice, each coded as yes or no). Results are reported as β coefficients with 95% CIs. All tests were 2-sided, and *P* values of less than .05 were considered statistically significant.

The qualitative component focused on free-text responses collected through the open-ended items of the survey. For the question assessing students' definitions of AI, only the incorrect responses were analyzed, as the intention was to explore students' misconceptions and the underlying representations associated with them. Additional free-text responses regarding feelings toward AI and expectations for AI training were also included in the qualitative dataset.

All qualitative data were compiled in Microsoft Excel and analyzed using a thematic analysis approach. Coding was performed iteratively: an initial set of codes was developed deductively based on concepts previously described in the literature (eg, nuanced optimism and replacement concerns), and additional categories were informed inductively from the data. Codes were progressively merged into broader categories through constant comparison, resulting in the identification of overarching themes. To enhance the reliability of the analysis, a second reviewer independently coded 30% of the material, and discrepancies were resolved through discussion.

Given the large and fully anonymous dataset, data saturation was not assessed, and validation by participants could not be performed. Representative quotations were selected to illustrate each theme. The study adhered to the COREQ (Consolidated Criteria for Reporting Qualitative Research) guidelines (Checklist 1).

All data were compiled in Excel version 2019 (Microsoft), and descriptive and qualitative analyses were performed using Microsoft Excel version 2019. Statistical tests were conducted using R software (version 4.3.1; R Foundation for Statistical Computing).

Ethical Considerations

All authors confirm that all methods were applied in accordance with current guidelines and regulations. This study did not require specific authorization under current French regulations, as it is research evaluating the value of teaching through a student health survey. This research does not fall under clinical research and, in France, does not require ethics committee approval (French Public Health Code: articles L1110-1 to L6441-1) [35].

To ensure full compliance with ethical standards, we implemented rigorous safeguards throughout the study. No sensitive personal data, such as health information, ethnic origin, or any identifiable details, were collected at any stage. Participant anonymity was strictly preserved: Responses were untraceable, with no collection of IP addresses or other identifying metadata, and all data were analyzed in aggregated form to prevent individual recognition. Informed consent was secured through a comprehensive information letter provided to all participants prior to the survey, clearly outlining the study's purpose, the voluntary nature of participation, and the

XSL•FO
RenderX

intended use of data, with completion of the survey serving as implied consent. A request for consent was clearly written. Finally, data storage and security adhered to the General Data Protection Regulation (GDPR) requirements, with all responses encrypted, accessible only to the research team, and permanently deleted following analysis in 3 years. The email addresses were deleted immediately after the study.

Results

Statistical Analyses

Population Characteristics

The survey was available online from May 8, 2025, to June 16, 2025. Our study collected 1377 responses, of which 1342 met the inclusion criteria and were included in the final analysis (Figure 1) from 23 faculties. The demographic characteristics of the sample were as follows: a median age of 23.1 (IQR 22 - 24) years, and the majority of students were in their fifth academic year (n=674, 50.2%; Table 1). Most participants selected hospital-based clinical work as a career aspiration (n=775, 57.7%) and were in the middle third of their class (n=694, 51.7%). The median general attitude toward AI score was 7 (IQR 5-8) [5-8]. Most students expected AI to affect their future career within less than 5 years (n=570, 42.5%). Concerning AI education, 651 (48.5%) preferred AI learning outside the formal curriculum, whereas 586 (43.7%) preferred its inclusion within the curriculum. The three most frequently selected educational objectives were identifying ways in which AI could improve healthcare quality (n=949, 70.7%), identifying when AI is appropriate in a given clinical context (n=916, 68.3%), and identifying the ethical implications of AI use in clinical settings (n=897, 66.8%). The geographic distribution showed an overrepresentation of students from Lille, Paris, and Nancy.

Figure 1 illustrates the flowchart showing the number of inclusions and exclusions, demonstrating how, out of the 1377 questionnaires received, we included 1342.

Table 1 and Textbox 1 show that the majority of students are in their fifth year, a minority wishes to pursue a research career, and a minority considers themselves to be in the lower third of their class ranking.

Textbox 1 shows that the 3 cities with the most responses are Lille, Paris Sorbonne, and Nancy.

XSL•FO
RenderX

Figure 1. Study flowchart of participant inclusion.

XSL•FO
RenderX

Table . Characteristics and perceptions of the study population[a].

Variable	Total study population (N=1342)
Age (y), median (IQR)	23 (22-24)
Academic year, n (%)	
Fourth	423 (31.5)
Fifth	674 (50.2)
Sixth	245 (18.3)
Career aspiration, n (%)	
Hospital	775 (57.7)
Research	382 (28.5)
Private practitioner	502 (37.4)
Class ranking, n (%)	
Lower third	190 (14.2)
Middle third	694 (51.7)
Upper third	458 (34.1)
Final AI definition category, n (%)	
Unknown	66 (4.9)
Incorrect	770 (57.4)
Partially correct	439 (32.7)
Correct	67 (5.0)
General attitude toward AI (0 - 10)	7 [5-8]
Perceived timeline before AI will impact career, n (%)	
Less than 5 years	570 (42.5)
In 5 years	357 (26.6)
In 10 years	349 (26)
In 20 years	56 (4.2)
In 50 years	4 (0.3)
Not during my lifetime	6 (0.4)
Preferred place for AI learning, n (%)	
In curriculum	586 (43.7)
Outside curriculum	651 (48.5)
Other or free text	105 (7.8)
Top 3 educational objectives, n (%)	
Identify ways in which AI can improve health care quality	949 (70.7)
Identify when the technology is appropriate for a given clinical context	916 (68.3)
Identify the ethical implications of AI use in clinical settings	897 (66.8)

[a]Values are presented as median [IQR] or n (%). For career aspirations and educational objectives, multiple responses were allowed; therefore, percentages may exceed 100%.

XSL•FO
RenderX

Textbox 1. Participation of French medical faculties.

With ≥100 respondents (%)

- Lille 299 (22%)
- Paris Sorbonne 160 (12%)
- Nancy 109 (8%)

With 50 - 99 respondents (%)

- Caen 98 (7%)
- Paris Cité 98 (7%)
- Strasbourg 75 (6%)
- Besançon 65 (5%)
- Marseille 59 (5%)
- Toulouse 57 (4%)
- Rennes 51 (4%)
- Limoges 50 (4%)

With <50 respondents (%)

- Amiens 42 (3%)
- Lyon 37 (3%)
- Paris Bicêtre 30 (2%)
- Tours 27 (2%)
- Antilles 26 (2%)
- Grenoble 21 (2%)
- Montpellier 20 (1%)
- Paris d'Orsay 13 (1%)
- Nîmes 2 (<1%)
- Nantes 1 (<1%)
- Reims 1 (<1%)
- Rouen 1 (<1%)

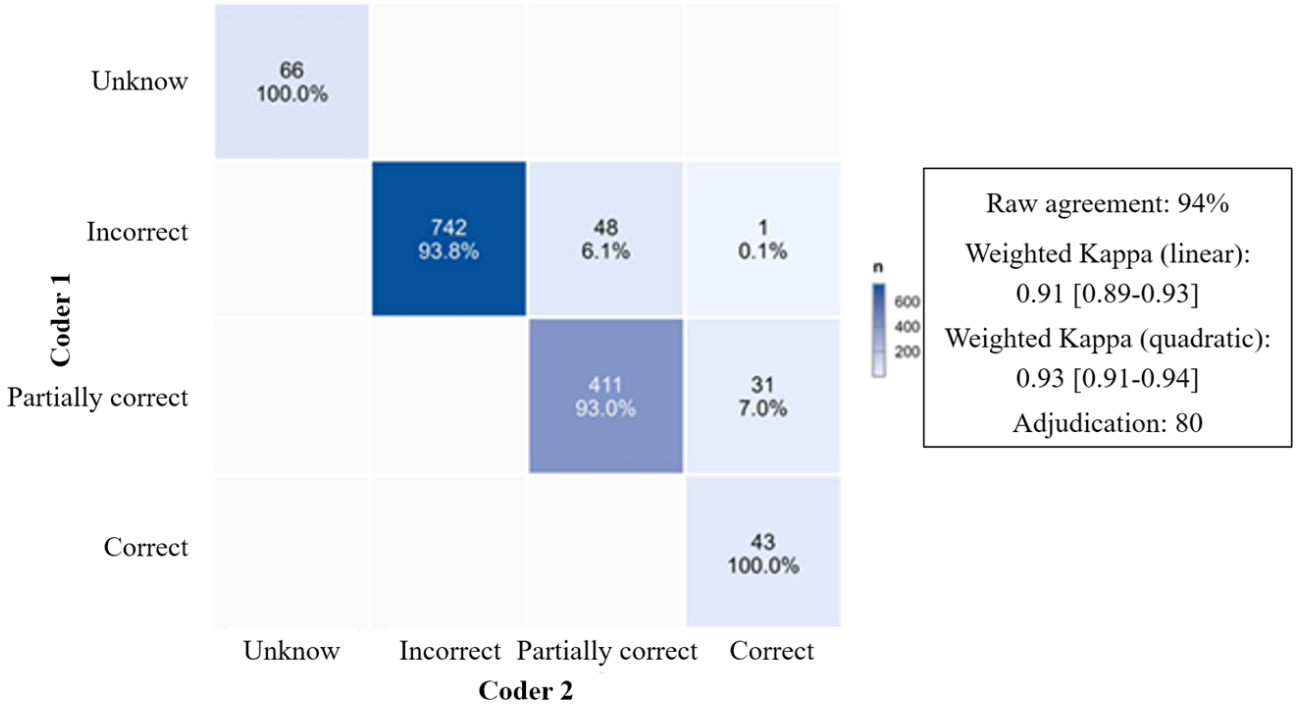
Definition of AI

After the adjudication of AI definitions, 770 (57.4%) definitions were categorized as incorrect, 439 (32.7%) as partially correct, 67 (5.0%) as correct, and 66 (4.9%) as unknown (Table 1).

The raw agreement between the 2 coders was 94.0% (Figure 2). The weighted Cohen κ was 0.91 (95% CI 0.89 - 0.93) using linear weights and 0.93 (95% CI 0.91 - 0.94) using quadratic weights (Figure 2). A total of 80 definitions required adjudication (Figure 2). Disagreements were uncommon and mainly involved adjacent categories. Across final AI definition categories, age, academic year, career aspirations, and class ranking did not differ significantly (Table 2). Effect sizes for these comparisons were uniformly small (all Cramér $V \leq 0.048$; $\varepsilon^2 = 0.000$ for age). Overall, associations between the AI definition category and students' characteristics, perceptions, and educational expectations were limited in magnitude.

XSL•FO
RenderX

Figure 2. iInterrater agreement for AI definition coding.



XSL•FO
RenderX

Table . Characteristics, perceptions, and educational expectations according to final AI definition category[a].

Variable	Overall (N=1342)	Unknown (n=66)	Incorrect (n=770)	Partially correct (n=439)	Correct (n=67)	*P* value	Effect size
Age, (y), median (IQR)	23 (22-24)	23 (22-24)	23 (22-24)	23 (22-24)	23 (22,23)	.63	ε²=0.000
Academic year, n (%)						.66	Cramér *V*=0.039
Fourth	423 (31.5)	18 (27.3)	240 (31.2)	148 (33.7)	17 (25.4)		
Fifth	674 (50.2)	35 (53.0)	395 (51.3)	210 (47.8)	34 (50.7)		
Sixth	245 (18.3)	13 (19.7)	135 (17.5)	81 (18.5)	16 (23.9)		
Career aspiration, n (%)							
Hospital	775 (57.7)	37 (56.1)	435 (56.5)	264 (60.1)	39 (58.2)	.66	Cramér *V*=0.035
Research	382 (28.5)	19 (28.8)	210 (27.3)	132 (30.1)	21 (31.3)	.71	Cramér *V*=0.032
Private practitioner	502 (37.4)	25 (37.9)	285 (37.0)	167 (38.0)	25 (37.3)	.99	Cramér *V*=0.010
Class ranking, n (%)						.40	Cramér *V*=0.048
Lower third	190 (14.2)	7 (10.6)	117 (15.2)	58 (13.2)	8 (11.9)		
Middle third	694 (51.7)	38 (57.6)	402 (52.2)	225 (51.3)	29 (43.3)		
Upper third	458 (34.1)	21 (31.8)	251 (32.6)	156 (35.5)	30 (44.8)		
General attitude toward AI (0 - 10), median (IQR)	7 (5-8)	6 (5-8)	7 (5-8)	7 (6-8)	7 (5.5 - 8)	.02	ε²=0.005
Perceived timeline before AI will impact career, n (%)						.18	Cramér *V*=0.070
<5 years	570 (42.5)	28 (42.4)	318 (41.3)	196 (44.6)	28 (41.8)		
In 5 years	357 (26.6)	16 (24.2)	202 (26.2)	123 (28.0)	16 (23.9)		
In 10 years	349 (26.0)	17 (25.8)	214 (27.8)	103 (23.5)	15 (22.4)		
In 20 years	56 (4.2)	3 (4.5)	31 (4.0)	15 (3.4)	7 (10.4)		
In 50 years	4 (0.3)	1 (1.5)	3 (0.4)	0 (0.0)	0 (0.0)		
Not during my lifetime	6 (0.4)	1 (1.5)	2 (0.3)	2 (0.5)	1 (1.5)		
Preferred place for AI learning, n (%)						.10	Cramér *V*=0.063
In curriculum	586 (43.7)	26 (39.4)	319 (41.4)	207 (47.2%)	34 (50.7)		
Outside curriculum	651 (48.5)	36 (54.5)	389 (50.5)	202 (46.0)	24 (35.8)		
Other or free text	105 (7.8)	4 (6.1)	62 (8.1)	30 (6.8)	9 (13.4)		
Top 3 educational objectives, n (%)							
Identify ways in which AI can improve health care quality	949 (70.7)	48 (72.7)	566 (73.5)	292 (66.5)	43 (64.2)	.04	Cramér *V*=0.078
Identify when the technology is appropriate for a given clinical context	916 (68.3)	48 (72.7)	525 (68.2)	295 (67.2)	48 (71.6)	.75	Cramér *V*=0.030
Identify the ethical implications of AI use in clinical settings	897 (66.8)	40 (60.6)	508 (66.0)	300 (68.3)	49 (73.1)	.38	Cramér *V*=0.048

XSL•FO
RenderX

[a]Values are presented as median (IQR) or n (%). *P* values were derived from Kruskal-Wallis tests for continuous or score variables and chi-square tests for categorical variables. Effect sizes are reported as ε^2 for Kruskal-Wallis tests and Cramér V for chi-square tests. For career aspirations and educational objectives, multiple responses were allowed; therefore, percentages may exceed 100%.

Figure 2 shows the interrater agreement with a Cohen κ using linear weights of 0.907 (95% CI 0.887 - 0.926) and a Cohen κ using quadratic weights of 0.927 (95% CI 0.909 - 0.943).

Table 2 shows that age, academic year, career aspirations, and class rankings are similarly distributed across the different AI definition correctness categories. Students with partially correct or correct definitions display a slightly more positive general attitude toward AI compared with other groups. In addition, identifying ways in which AI can improve health care quality is more frequently cited as a priority educational objective, while no major differences are observed regarding the perceived AI impact timeline or the preferred place for AI learning.

Attitudes Toward AI

The general attitude toward AI scores differed significantly across AI definition categories (Kruskal-Wallis *P*=.02), although the effect size was very small (ε^2=0.005; Table 2). Post hoc Dunn tests with Holm correction showed that students in the unknown group had lower general attitude scores than those in the incorrect group (adjusted *P*=.01), the partially correct group (adjusted *P*=.01), and the correct group (adjusted *P*=.047). No other pairwise differences were observed (Table 2).

The perceived timeline before AI impact did not significantly differ across AI definition categories (*P*=.18; Cramér *V*=0.070). Similarly, the preferred place for AI learning did not significantly differ according to the AI definition category (*P*=.10; Cramér *V*=0.063; Table 2).

Among the 3 leading educational objectives, only the objective related to identifying ways in which AI can improve health care quality differed significantly across groups (*P*=.04), although the effect size remained small (Cramér *V*=0.078; Table 2). The 2 other leading objectives, identifying when AI is appropriate in a given clinical context and identifying the ethical implications of AI use in clinical settings, did not differ significantly across AI definition categories (Table 2).

Among the 6 specific AI attitude items, the strongest signal concerned the item “AI will impact my career,” which differed across AI definition categories (Kruskal-Wallis *P*=.005; ε^2=0.007; Table 2). After Holm correction across the 6 attitude items, this association remained statistically significant (adjusted *P*=.03; Figure 3). Post hoc Dunn-Holm comparisons showed that students in the unknown group reported lower scores than those in the correct group (adjusted *P*=.005) and the partially correct group (adjusted *P*=.02; Table 2 and Figure 3). No other pairwise differences remained significant after correction. The item “students should learn the basics of AI” showed a nominal omnibus association (*P*=.03; ε^2=0.005), but this did not remain significant after Holm correction across the 6 items (adjusted *P*=.13; Table 2 and Figure 3). The other 4 specific attitude items were not significantly associated with AI definition category (Table 2 and Figure 3).

XSL•FO
RenderX

Figure 3. Specific attitudes depending on the result of the AI definition. (A) "AI requires careful management"; (B) "AI will impact my career"; (C) "I am concerned about AI"; (D)."I feel optimistic about AI"; (E) "I understand the ethical implications of AI"; and (F) "Students should learn AI basics."

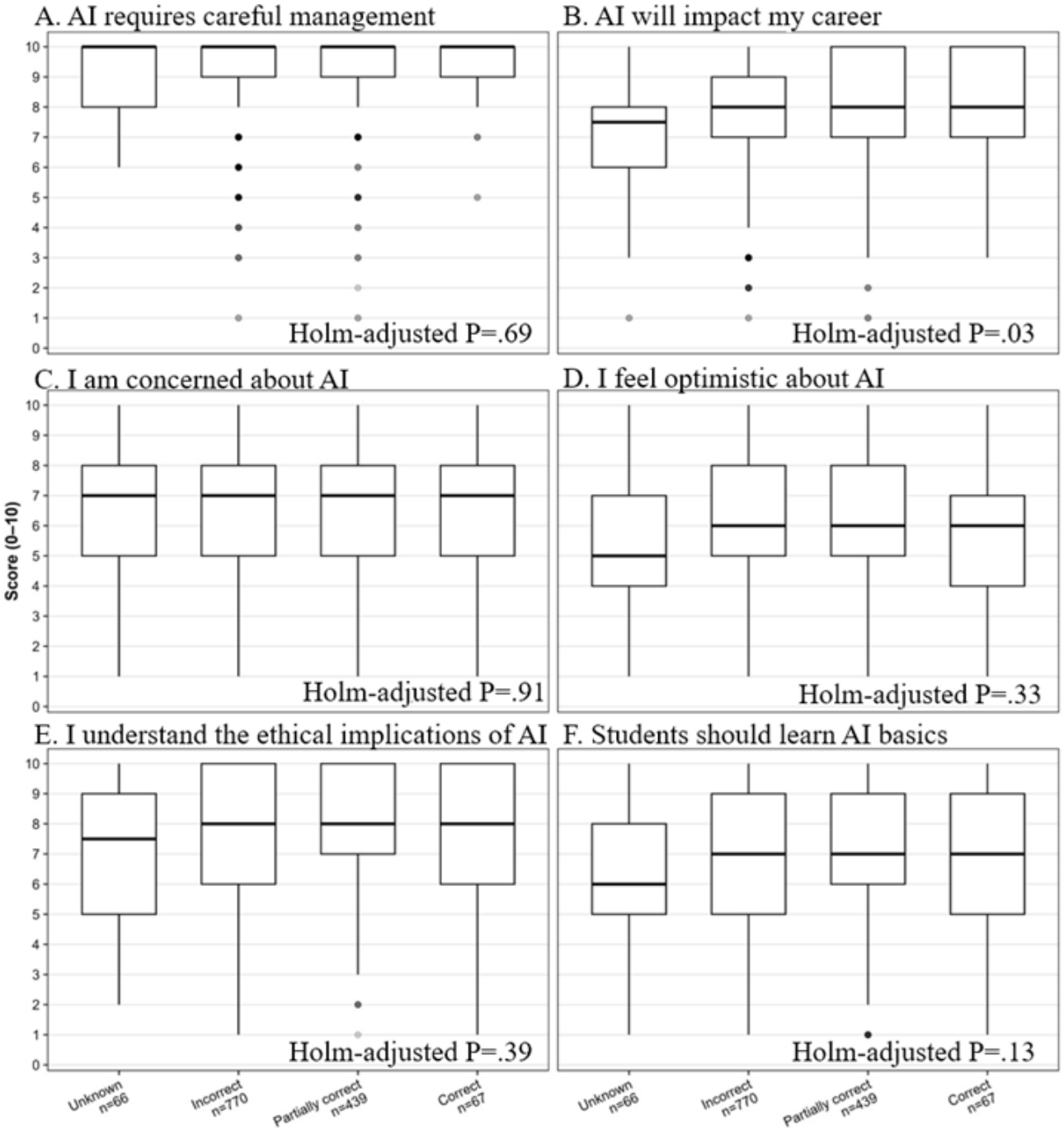


Figure 3 presents, via 6 box plots, the distributions of responses (unknown, incorrect, partially correct, and correct) to 5 statements about AI, revealing that the majority of participants assess the impact of AI on their careers with significant differences depending on their response; the other statements do not show any significant differences.

In an exploratory multivariable linear regression model with robust SEs, belonging to the unknown AI definition category remained associated with a lower general attitude toward AI score compared with the incorrect category (regression coefficient β=−0.72, 95% CI −1.17 to −0.28; *P*=.001; Table 3). By contrast, partially correct and correct AI definition categories were not significantly associated with the general attitude score after adjustment. A preference for private practice was independently associated with a higher general attitude toward AI score (regression coefficient β=0.54, 95% CI 0.28-0.79; *P*<.001; Table 3). Older age was also associated with slightly higher scores (regression coefficient β=0.07 per year, 95% CI 0.01-0.12; *P*=.02), whereas being in the sixth year was associated with slightly lower scores compared with the fifth year (regression coefficient β=−0.32, 95% CI −0.61 to −0.03; *P*=.03; Table 3).

XSL•FO
RenderX

Table . Selected coefficients from the exploratory multivariable linear regression model for the general attitude toward AI score.

Variable	β coefficient, (95% CI)	*P* value
Unknown vs Incorrect AI definition category	−0.72 (−1.17 to −0.28)	.001
Private practice aspiration (yes vs no)	0.54 (0.28 to 0.79)	<.001
Age (y)	0.07 (0.01 to 0.12)	.02
Sixth vs fifth academic year	−0.32 (−0.61 to −0.03)	.03

Table 3 shows the results from an exploratory multivariable linear regression model with heteroskedasticity-consistent robust SEs. The reference categories were the incorrect AI definition category and the fifth academic year. The model was additionally adjusted for the partially correct vs incorrect AI definition category, the correct vs incorrect AI definition category, the fourth vs fifth academic year, class ranking, hospital-based career aspiration, and research aspiration; these associations were not statistically significant after adjustment.

Learning

Concerning AI education, 651 (48.5%) preferred AI learning outside the formal curriculum, whereas 586 (43.7%) preferred its inclusion within the curriculum, and 105 (7.8%) responded via free text (Table 1). Regarding the free responses, participants expressed that the training program was too heavy and proposed alternatives such as nonvalidating training, training offered outside the second cycle, or training provided only during internships (Table 4, verbatim 1:4). Finally, some students expressed concerns about the lack of current AI use in clinical settings, which they felt made the training less relevant (Table 4, verbatim 5).

XSL•FO
RenderX

Table . Verbatim.

Number	Verbatim	Faculty	Age (y)
1	“The knowledge required in medicine is already very extensive; AI training should be non-validating but included in the program.”	Lyon	21
2	“We need the PIX certification to validate our med2med3”	Lille	22
3	“Our learning program is already extremely busy during the externship I struggle to find my place. It should be ““included” in internships. I think. But not necessarily additional courses or training on the side, we already have a lot”	Nancy	23
4	“Should be included later, once AI has a more defined role in”	Lille	23
5	“Should be part of the program when it is truly used on a large scale.”	Lille	24
6	“Intelligence imitating human intelligence with an artificial database.”	Nancy	27
7	“A robot trained to think like a human.”	Lille	22
8	“A search engine that knows almost everything.”	Rennes	24
9	“Computer program that generates information from a set of data and operates on machine learning”	Amiens	23
10	“I love it.”	Lille	25
11	“A major advantage for improving patient care [.]”	Lyon	24
12	“It’s an easily accessible resource that saves time.”	Lille	21
13	“An almost indispensable daily assistant.”	Tours	21
14	“Optimistic. I hope AI will simplify caregivers’ tasks and allow for medicine that is closer to the patient as an individual”	Paris	21
15	“Fear, anger, hatred.”	Marseille	23
16	“Insecurity, it is a tool that can become very dangerous very”	Paris	22
17	“Very mistrustful: outsourcing knowledge dulls our abilities there are security, confidentiality, and ecological concerns.”	Lille	23
18	“I fear we will rely too much on it and become dependent (and stupid in the process AI discourages real thinking)”	Caen	23
19	“Indifference.”	Caen	30
20	“Mixed feelings.”	Nancy	27

XSL•FO
RenderX

Number	Verbatim	Faculty	Age (y)
21	“AI is the future of just about everything, whether we like it or not. It is up to us to move forward with it and ensure that things are done properly and for the benefit of all.”	Paris	21
22	“A still-imprecise tool that will surely become relevant with more training and supervision by an experienced person.”	Strasbourg	23
23	“A good tool, especially for learning, but it’s important to educate people about its use and the reliability of the information.”	Strasbourg	23
24	“I think AI can be very beneficial as an aid, but its misuse can happen very quickly.”	Lille	22
25	“A lot of potential, but risks of misuse, so it should be introduced with caution.”	Caen	25
26	“Vigilance regarding excessive use.”	Paris	22
27	“A good tool, but dangerous for self-diagnosis, self-medication, and lack of holistic care the human aspect is essential”	Besançon	21
28	“Promising development, but with major challenges due to the type of data it handles (medical information) and possible consequences (patient survival), raising questions of responsibility (the doctor or the machine”	Paris	22
29	“Its proper use requires maintaining a critical mind and understanding the ethical issues behind it.”	Lille	22
30	“Requires ethical and legislative oversight.”	Paris	21
31	“Great potential, but risks changing the profession so much that it may no longer resemble what we are learning now.”	Caen	25
32	“Worrying: the role of the doctor what will be our added value?”	Strasbourg	22
33	“I’m afraid of becoming obsolete.”	Paris	22
34	“Fear of being replaced.”	Lille	22
35	“AI will never replace human contact”	Tours	21
36	“Helps against human error but does not replace the essential.”	Grenoble	39
37	“In medicine, I fear people will trust AI more than their doctor.”	Strasbourg	24
38	“Replacement of support functions previously performed by human professionals loss of professional-patient contact.”	Limoges	31

XSL•FO
RenderX

Number	Verbatim	Faculty	Age (y)
39	"I think AI is a tool that can help with time-consuming daily tasks (administrative work, data synthesis), but it must be used carefully to avoid dehumanized care."	Strasbourg	27
40	"Helpful for repetitive tasks."	Nancy	22
41	"Useful for some courses, but still makes many mistakes interesting for creating flashcards and transferring to Anki afterward."	Amiens	23
42	"I use AI to structure, condense, and deepen my revisions interactively and personally, especially by creating course."	Nancy	25
43	"Very helpful for understanding physiopathological concepts in medicine saves time searching through books that often end up contradicting each other."	Nancy	23
44	"Happy to have access to this resource in my research."	Lille	24
45	"It helps me understand better, like having a private tutor who questions me, helps me, and analyzes my learning in a targeted"	Nancy	23
46	"This could be a huge step forward in medicine, both diagnostically and administratively, saving a great deal of time, but it must be used wisely as it will never replace humans with their feelings, emotions and consciousness."	Nancy	23
47	"Can be a valuable aid but should not replace the judgment of a practitioner."	Amiens	22
48	"I see its value as a diagnostic aid, especially in radiology, but we must never think that AI will replace a doctor."	Lille	23

Table 1 illustrates opinions on the integration of AI into curricula: 48.5% (n=651) of respondents do not want it included in the curriculum (compared to n=586, 43.7% who do). The preferred formats for learning about AI were short workshops and multiple workshop series, ahead of 1-day courses and graduate-level education. The priority objectives for its introduction are improving health care (25%), understanding how it works (22%), and identifying ethical issues (20%), followed by communication, interpretation of results, and contextual use. This preference for practical and targeted modules reveals a need to combine educational effectiveness and clinical relevance while emphasizing ethics, reflecting a pragmatic and balanced approach between theory, application, and responsibility. The emphasis on workshops suggests a demand for flexibility and gradual immersion.

Table 4 lists the 48 verbatim selections.

Students would mainly prefer training through multiple workshops (Table 1). Additionally, the 3 most frequently selected educational objectives were identifying ways in which AI could improve health care quality (949/1342, 70.7%), identifying when AI is appropriate in a given clinical context (916/1342, 68.3%), and identifying the ethical implications of AI use in clinical settings (897/1342, 66.8%; Table 1). Among the 3 leading educational objectives, only the objective related to identifying ways in which AI could improve health care quality differed significantly across groups (P=.04), although the effect size remained small (Cramér V=0.078; Table 2). The 2 other leading objectives, identifying when AI is appropriate in a given clinical context and identifying the ethical implications of AI use in clinical settings, did not differ significantly across AI definition categories.

XSL•FO
RenderX

Qualitative Analysis

Coding Reliability

Among the 1342 open-ended responses collected, 1340 were suitable for analysis. Of these, 409 (31%) responses were double-coded independently by 2 researchers to assess coding reliability. An initial disagreement was identified in 18 cases, corresponding to an intercoder agreement rate of 95%. No additional themes emerged during the double-coding process. Following a discussion of the 18 initial disagreements, 9 codes remained unchanged, while 9 codes were reassigned to different themes or split into multiple themes. No further arbitration by a third reviewer was required.

Identified Themes

Regarding the analysis of incorrect responses for the AI definition, we identified 3 themes: their representation of AI, AI use, and nuanced optimism (Multimedia Appendix 2). When students expressed their feelings, we again identified the themes of Critical Consideration and Replacement.

Their Representation

Students describe AI as a computer tool, software, or a robot, while some compare it to a human. Some perceive AI as a database with notions of infinity or omnipotence; others as an improved search engine, or they directly mention ChatGPT (Table 4, verbatim 6:9).

Nuanced Optimism

Students express positive aspects, described as help that can be indispensable, effective, and provided daily. According to the students, AI would save time through rapid execution and easy access. Additionally, some explain that the use of AI will improve patient care and provide an opportunity to focus more on the relationship with the patient (Table 4, verbatim 10:14). In contrast, some share feelings of anger and fear, highlighting dangers, particularly regarding energy costs, the dehumanization of medicine, and the regression of skills (Table 4, verbatim 15:18). Finally, some students express neutral, doubtful, uncertain, and sometimes fatalistic views, stating that AI is part of the future (Table 4, verbatim 19; 20).

Critical Consideration

Students expressed some mistrust toward AI, which is fallible and imperfect, requiring learning and human control (Table 4, verbatim 21; 22). Additionally, students expressed reservations about how AI is used, particularly regarding excessive use or restrictions for patients (Table 4, verbatim 23:26). Finally, students questioned data security issues, the need for a precise framework for use with established ethical rules, and information about responsibility in case of error (Table 4, verbatim 27:29).

Replacement

With the advancement of AI in health care, students questioned the role of doctors and the AI-related changes in their work (Table 4, verbatim 30). Some described a feeling of fear, while others believed replacement by AI was impossible due to the need for doctor-patient relationships and empathy (Table 4, verbatim 31:34). There was also concern about patients using AI, which could lead to a loss of credibility and a weakening of the doctor–patient relationship (Table 4, verbatim 35:37).

AI Use

Students describe varied uses, such as assistance with administrative tasks and data management (Table 4, verbatim 38; 39). Additionally, some express benefits for their learning, including understanding lessons, creating study sheets, receiving personalized training, and searching for information (Table 4, verbatim 40:44). Finally, regarding clinical applications, students express divergent views on whether AI should be used for clinical reasoning and decision-making support, with particularly polarized opinions in the field of radiology (Table 4, verbatim 45:48).

Discussion

Principal Results

This study aimed to explore medical students' representations and attitudes toward AI in France in 2025. Our results indicate that students' attitudes toward AI depend essentially on their knowledge of the definition of AI, as students with more accurate definitions tend to report more positive attitudes. There is also great variability in students' attitudes toward AI, with marked differences not only between participants but also within individual responses. These differences highlight the complexity of student perceptions of AI. The themes of "nuanced optimism" and "replacement" align with observations from other studies [18,23,24,26]. New themes also emerged, such as fear of skill loss, awareness of the high ecological cost of AI use, and loss of credibility with patients. Concerns about skill loss have also been reported among educators, particularly regarding the risk of cognitive offloading and erosion of critical thinking skills, suggesting that this apprehension is shared across different stakeholders in medical education [10,11]. Ecological concerns, meanwhile, align with broader societal discussions on the energy consumption and environmental impact of large-scale AI systems, transcending the specific context of medical education [36].

Additionally, the results reveal how AI is used for learning and searching for information on the internet. However, the need for an ethical framework and responsible AI use remains a major concern for students, even though the World Health Organization (WHO) recommendations have been in place since 2021 [2].

Our study reveals 2 notable developments compared to previous work. First, only a small percentage of students reported not knowing how to define AI, compared to half of the participants in the studies by Teng et al [21]. In contrast, in our study, half of the students gave incorrect answers, suggesting that students now develop representations of AI, even if they are often erroneous, rather than being completely unaware. Their concerns about ethical considerations, AI management, and its impact on their careers remain similar to those observed by Teng et al [21]. As noted by Amiri et al [19], students' prior knowledge strongly influences their perceptions. However, neither the level of education nor the academic year showed a significant effect [19]. These results are likely related to self-assessment of their

XSL•FO
RenderX

ranking within their class. Finally, the theme of "replacement" persists [19,21], as does ambivalence between perceived AI utility and its risks. Finally, unlike in some studies [18,23,26], no formal training request emerged, probably due to students' self-training or the still-limited clinical use of AI.

We also highlighted a form of critical consideration among students toward AI. Indeed, students expressed ethical concerns and reservations about using AI in medical contexts [20]. For example, they emphasized the need for a precise framework and established ethical rules, but this critical attitude seems less present when it comes to their own learning. Students show some mistrust toward AI as a clinical tool, recognizing its limitations and the need for human supervision. However, this critical consideration is less evident in their use of AI for learning, where they seem more inclined to accept the provided information without always questioning it. This dichotomy suggests that students perceive greater risks related to AI in medical practice than in their own learning process. It would, therefore, be crucial to encourage a more critical approach to AI use in learning, emphasizing the importance of verifying and validating the obtained information to develop a more thoughtful and responsible use of these technological tools [37].

AI use by medical students is still relatively unexplored compared with studies in other educational fields [30,32,38]. Our study revealed that medical students use AI in various ways to meet their learning needs, primarily for writing assistance, information searches, or personalized learning. It seems that this use in their learning could lead to excessive and uncritical use of information. Moreover, it is crucial to explore the impact on skill acquisition and dependence on AI in their professional environments. However, this adoption of AI in their learning process could potentially result in excessive and uncritical use of the provided information, risking the creation of dependence on these tools. On the other hand, this could also lead to the loss of essential skills, particularly regarding clinical judgment and decision-making autonomy, which are crucial in medical practice. Therefore, it is essential to explore these potential impacts on both the acquisition of critical medical skills and students' future professional practice.

Limitations

Our study had certain limitations. First, the response rate relative to the number of enrolled students was less than 10%, which limits the generalizability of our findings. The use of voluntary participation may have introduced self-selection bias, as students with stronger preexisting interest in AI, whether positive or negative, may have been more likely to respond. While participation exceeded 10% at some universities, the low overall rate suggests that our results may disproportionately reflect the attitudes of a more engaged subgroup rather than the broader student population. Second, we chose the Teng et al [21] survey because it was designed 4 years ago, allowing for comparison. Its mixed methodology offered the advantage of exploring both the underlying reasons for students' attitudes and their modes of AI use [20]. However, regarding the evaluation of the definition, it is important to note that it is not possible to assess students' knowledge based solely on their ability to define AI. Other surveys that evaluate AI knowledge more globally could have been used, but their length might have discouraged participation and thus reduced the response rate. A key limitation of our study is the absence of demographic data, including gender or sexual orientation. We deliberately chose not to collect this information. This decision was motivated by two ethical concerns: (1) the need to avoid reductive or exclusionary categorizations of gender, which is a spectrum of identities extending beyond binary frameworks; and (2) the risk of compromising participant anonymity, as detailed demographic data could have enabled indirect identification in smaller faculty samples. Finally, although several associations reached statistical significance due to the large sample size, the magnitudes are uniformly small, and practical implications should be interpreted cautiously.

Conclusions

Our study describes the current perceptions, attitudes, and uses of AI among French medical students in 2025. Students report generally positive attitudes toward AI despite limited conceptual knowledge of the definition of AI, alongside concerns related to ethics, ecological impact, and potential skill loss. Many already use AI tools for learning in a largely self-directed manner. Further longitudinal research is needed to better understand the implications of these practices for medical education and professional development. These findings provide a descriptive baseline for future evaluations of AI-related training and its impact on learning and professional development.

Acknowledgments

The authors thank Thierry Pelaccia and Stéphane Guillon for their guidance and support throughout the development of this study and for their valuable feedback. The authors also thank Marie-Hélène Bompard for her contribution to the co-coding. Generative AI tools were used solely to assist with language editing and revision of the manuscript. The authors reviewed and validated all content and remain fully responsible for the accuracy, originality, interpretation, and integrity of the manuscript, including all references and citations.

Funding

The authors declared no financial support was received for this work.

Disclaimer

None declared.

XSL•FO
RenderX

Data Availability

The datasets generated and/or analyzed during the current study are available from the corresponding author upon reasonable request.

Authors' Contributions

FP conceived the study, developed the methodology, designed the survey, collected the data, performed the qualitative coding of AI definitions, and drafted the manuscript. VG contributed to the study methodology, performed the statistical analyses, produced the tables and figures, participated in the qualitative coding of AI definitions, contributed to the writing of the Results section, and critically revised the manuscript. LAV contributed to the study methodology, critically revised the manuscript, served as the third reviewer for the coding and validation of AI definitions and validated the final version. All authors contributed to the interpretation of the findings, reviewed and approved the final manuscript, and agreed to be accountable for all aspects of the work.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Full survey.
[PDF File, 188 KB - mededu_v12i1e91345_app1.pdf]

Multimedia Appendix 2
Mind Map qualitative results.
[PNG File, 1584 KB - mededu_v12i1e91345_app2.png]

Checklist 1
COREQ checklist.
[PDF File, 164 KB - mededu_v12i1e91345_app3.pdf]

References

1. Craglia M, Annoni A, Benczur P, et al. Artificial intelligence: a European perspective. : Publications Office of the European Union; 2018 URL: https://publications.jrc.ec.europa.eu/repository/handle/JRC113826 [accessed 2025-12-01]
2. Ethics and governance of artificial intelligence for health: WHO guidance executive summary. : World Health Organization; 2021 URL: https://iris.who.int/server/api/core/bitstreams/4f2c477c-4b72-4ca1-9a78-a1e73af64e50/content [accessed 2025-12-01]
3. Wang F, Casalino LP, Khullar D. Deep learning in medicine-promise, progress, and challenges. JAMA Intern Med 2019 Mar 1;179(3):293-294. [doi: 10.1001/jamainternmed.2018.7117] [Medline: 30556825]
4. Bohr A, Memarzadeh K. Chapter 2 - the rise of artificial intelligence in healthcare applications. In: Artificial Intelligence in Healthcare: Academic Press; 2020:25-60. [doi: 10.1016/B978-0-12-818438-7.00002-2]
5. Yu KH, Kohane IS. Framing the challenges of artificial intelligence in medicine. BMJ Qual Saf 2019 Mar;28(3):238-241. [doi: 10.1136/bmjqs-2018-008551] [Medline: 30291179]
6. Sanchez D, Paris F, Laribi S, Gilbert A. Reducing administrative burden in emergency departments using artificial intelligence: a workflow-based analysis. Eur J Emerg Med 2026 Feb 1;33(1):3-5. [doi: 10.1097/MEJ.0000000000001251] [Medline: 40590802]
7. Gordon M, Daniel M, Ajiboye A, et al. A scoping review of artificial intelligence in medical education: BEME Guide No. 84. Med Teach 2024 Apr;46(4):446-470. [doi: 10.1080/0142159X.2024.2314198] [Medline: 38423127]
8. Kaur A, Singh S, Chandan JS, Robbins T, Patel V. Qualitative exploration of digital chatbot use in medical education: a pilot study. Digit Health 2021;7:20552076211038151. [doi: 10.1177/20552076211038151] [Medline: 34513002]
9. Jose B, Cherian J, Verghis AM, Varghise SM, S M, Joseph S. The cognitive paradox of AI in education: between enhancement and erosion. Front Psychol 2025;16:1550621. [doi: 10.3389/fpsyg.2025.1550621] [Medline: 40297599]
10. Abdulnour REE, Gin B, Boscardin CK. Educational strategies for clinical supervision of artificial intelligence use. N Engl J Med 2025 Aug 21;393(8):786-797. [doi: 10.1056/NEJMra2503232] [Medline: 40834302]
11. Feigerlova E, Hani H, Hothersall-Davies E. A systematic review of the impact of artificial intelligence on educational outcomes in health professions education. BMC Med Educ 2025 Jan 27;25(1):129. [doi: 10.1186/s12909-025-06719-5] [Medline: 39871336]
12. Gerlich M. AI tools in society: impacts on cognitive offloading and the future of critical thinking. Societies 2025;15(1):6. [doi: 10.3390/soc15010006]

XSL•FO
RenderX

13. Paranjape K, Schinkel M, Nannan Panday R, Car J, Nanayakkara P. Introducing artificial intelligence training in medical education. JMIR Med Educ 2019 Dec 3;5(2):e16048. [doi: 10.2196/16048] [Medline: 31793895]
14. Pupic N, Ghaffari-Zadeh A, Hu R, et al. An evidence-based approach to artificial intelligence education for medical students: a systematic review. PLOS Digit Health 2023 Nov;2(11):e0000255. [doi: 10.1371/journal.pdig.0000255] [Medline: 38011214]
15. Santomartino SM, Yi PH. Systematic review of radiologist and medical student attitudes on the role and impact of AI in radiology. Acad Radiol 2022 Nov;29(11):1748-1756. [doi: 10.1016/j.acra.2021.12.032] [Medline: 35105524]
16. Gong B, Nugent JP, Guest W, et al. Influence of artificial intelligence on Canadian medical students' preference for radiology specialty: ANational survey study. Acad Radiol 2019 Apr;26(4):566-577. [doi: 10.1016/j.acra.2018.10.007] [Medline: 30424998]
17. Pinto Dos Santos D, Giese D, Brodehl S, et al. Medical students' attitude towards artificial intelligence: a multicentre survey. Eur Radiol 2019 Apr;29(4):1640-1646. [doi: 10.1007/s00330-018-5601-1] [Medline: 29980928]
18. Park CJ, Yi PH, Siegel EL. Medical student perspectives on the impact of artificial intelligence on the practice of medicine. Curr Probl Diagn Radiol 2021;50(5):614-619. [doi: 10.1067/j.cpradiol.2020.06.011] [Medline: 32680632]
19. Amiri H, Peiravi S, Rezazadeh Shojaee SS, et al. Medical, dental, and nursing students' attitudes and knowledge towards artificial intelligence: a systematic review and meta-analysis. BMC Med Educ 2024 Apr 15;24(1):412. [doi: 10.1186/s12909-024-05406-1] [Medline: 38622577]
20. Mousavi Baigi SF, Sarbaz M, Ghaddaripouri K, Ghaddaripouri M, Mousavi AS, Kimiafar K. Attitudes, knowledge, and skills towards artificial intelligence among healthcare students: a systematic review. Health Sci Rep 2023 Mar;6(3):e1138. [doi: 10.1002/hsr2.1138] [Medline: 36923372]
21. Teng M, Singla R, Yau O, et al. Health care students' perspectives on artificial intelligence: countrywide survey in Canada. JMIR Med Educ 2022 Jan 31;8(1):e33390. [doi: 10.2196/33390] [Medline: 35099397]
22. McLennan S, Meyer A, Schreyer K, Buyx A. German medical students´ views regarding artificial intelligence in medicine: a cross-sectional survey. PLOS Digit Health 2022 Oct;1(10):e0000114. [doi: 10.1371/journal.pdig.0000114] [Medline: 36812635]
23. Al Saad MM, Shehadeh A, Alanazi S, et al. Medical students' knowledge and attitude towards artificial intelligence: an online survey. Open Public Health J 2022 May 24;15(1). [doi: 10.2174/18749445-v15-e2203290]
24. Chen M, Zhang B, Cai Z, et al. Acceptance of clinical artificial intelligence among physicians and medical students: a systematic review with cross-sectional survey. Front Med (Lausanne) 2022;9:990604. [doi: 10.3389/fmed.2022.990604] [Medline: 36117979]
25. AlZaabi A, AlMaskari S, AalAbdulsalam A. Are physicians and medical students ready for artificial intelligence applications in healthcare? Digit Health 2023;9:20552076231152167. [doi: 10.1177/20552076231152167] [Medline: 36762024]
26. Banerjee M, Chiew D, Patel KT, et al. The impact of artificial intelligence on clinical education: perceptions of postgraduate trainee doctors in London (UK) and recommendations for trainers. BMC Med Educ 2021 Aug 14;21(1):429. [doi: 10.1186/s12909-021-02870-x] [Medline: 34391424]
27. Moldt JA, Festl-Wietek T, Madany Mamlouk A, Nieselt K, Fuhl W, Herrmann-Werner A. Chatbots for future docs: exploring medical students' attitudes and knowledge towards artificial intelligence and medical chatbots. Med Educ Online 2023 Dec;28(1):2182659. [doi: 10.1080/10872981.2023.2182659] [Medline: 36855245]
28. Gillissen A, Kochanek T, Zupanic M, Ehlers J. Medical students' perceptions towards digitization and artificial intelligence: a mixed-methods study. Healthcare (Basel) 2022 Apr 13;10(4):723. [doi: 10.3390/healthcare10040723] [Medline: 35455898]
29. Tran AQ, Nguyen LH, Nguyen HSA, et al. Determinants of intention to use artificial intelligence-based diagnosis support system among prospective physicians. Front Public Health 2021;9:755644. [doi: 10.3389/fpubh.2021.755644] [Medline: 34900904]
30. Zhou X, Zhang J, Chan C. Unveiling students' experiences and perceptions of artificial intelligence usage in higher education. JUTLP 2024;21(6):06. [doi: 10.53761/xzjprb23]
31. Dogan ME, Goru Dogan T, Bozkurt A. The use of artificial intelligence (AI) in online learning and distance education processes: a systematic review of empirical studies. Appl Sci 2023;13(5):3056. [doi: 10.3390/app13053056]
32. Tajima H, Kasai H, Shikino K, Shimizu I, Ito S. Perceptions and intentions to use generative AI among first-year medical students in Japan: cross-sectional survey study. JMIR Med Educ 2025 Nov 13;11:e77552. [doi: 10.2196/77552] [Medline: 41232013]
33. Akhmetova AI, Sovetkanova DM, Komekbayeva LK, Abdrakhmanov AE, Yessenuly D, Serikova OS. A systematic review of artificial intelligence in high school STEM education research. EURASIA J Math Sci Tech Ed 2025;21(4):em2623. [doi: 10.29333/ejmste/16222]
34. Explanatory memorandum on the updated OECD definition of an AI system. : Organisation for Economic Co-operation and Development (OECD); 2024 URL: https://www.oecd.org/en/publications/explanatory-memorandum-on-the-updated-oecd-definition-of-an-ai-system_623da898-en.html [accessed 2026-04-01]
35. Code de la santé publique [Article in French]. Légifrance. URL: https://www.legifrance.gouv.fr/codes/texte_lc/LEGITEXT000006072665 [accessed 2026-07-07]

XSL•FO
RenderX

36. Xiao T, Nerini FF, Matthews HD, Tavoni M, You F. Environmental impact and net-zero pathways for sustainable artificial intelligence servers in the USA. Nat Sustain 2025;8(12):1541-1553. [doi: 10.1038/s41893-025-01681-y]
37. Shimizu I, Kasai H, Shikino K, et al. Developing medical education curriculum reform strategies to address the impact of generative AI: qualitative study. JMIR Med Educ 2023 Nov 30;9:e53466. [doi: 10.2196/53466] [Medline: 38032695]
38. Su J, Yang W. Artificial intelligence in early childhood education: a scoping review. Comput Educ Artif Intell 2022;3:100049. [doi: 10.1016/j.caeai.2022.100049]

Abbreviations

COREQ: Consolidated Criteria for Reporting Qualitative Research
GDPR: General Data Protection Regulation
OECD: Organisation for Economic Co-operation and Development
SAQ: short open-ended question
WHO: World Health Organization

Edited by D Chartash; submitted 13.Jan.2026; peer-reviewed by M Turos, PF Chen; revised version received 13.May.2026; accepted 15.May.2026; published 20.Jul.2026.

Please cite as:
Paris F, Garrouste V, Abensur Vuillaume L
Perceptions, Attitudes, and Use of AI by Medical Students: Mixed Methods Study
JMIR Med Educ 2026;12:e91345
URL: https://mededu.jmir.org/2026/1/e91345
doi:10.2196/91345

© Frédéric Paris, Vincent Garrouste, Laure Abensur Vuillaume. Originally published in JMIR Medical Education (https://mededu.jmir.org), 20.Jul.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Education, is properly cited. The complete bibliographic information, a link to the original publication on https://mededu.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Enhancing Psychiatry Training Using an Agentic AI Simulated Consultation Tool: Prospective Cohort Study

Alice Rueda[1*], PhD; Huda F Al-Shamali[1*], PhD; Zack Cote[1], BSc; Niloy Roy[1], BSc; Reinhard Janssen-Aguilar[1], MD; Jithin Joseph[1], MD; Bazen Gashaw Teferra[1], PhD; Mohammad Amin Kamaleddin[1], PhD; Lisa Burback[2], MD; Olga Winkler[2], MD; Bill Kapralos[3], PhD; Andrei Torres[3], PhD; Divya Sharma[4], PhD; Sridhar Krishnan[5], PhD; Guillaume Dumas[6,7], PhD; Andrew Greenshaw[2], PhD; Alexandre Hudon[8,9,10,11,12], MD, PhD; Sanjeev Sockalingam[13,14], MD; Yanbo Zhang[2], MD, PhD; Adam Dubrowski[15], PhD; Venkat Bhat[1,14], MSc, MD

[1]Department of Psychiatry, AI for Mental Health, St. Michael's Hospital, 30 Bond Street, 17-009 Cardinal Carter South, Toronto, ON, Canada

[10]Centre de Recherche, Institut Universitaire en Santé Mentale de Québec, Montréal, QC, Canada

[11]Department of Psychiatry, Institut National de Psychiatrie Légale Philippe-Pinel, Montréal, QC, Canada

[12]Groupe Interdisciplinaire de Recherche sur la Cognition et le Raisonnement Professionnel (GIRCoPRo), Department of Psychiatry, Université de Montréal, Montréal, QC, Canada

[13]Complex Care and Recovery, Centre for Addiction and Mental Health, Toronto, ON, Canada

[14]Department of Psychiatry, University of Toronto, 250 College Street, Toronto, ON, Canada

[15]Faculty of Health Science, Ontario Tech University, Oshawa, ON, Canada

[2]Department of Psychiatry, University of Alberta, Edmonton, AB, Canada

[3]maxSIMhealth Group, Ontario Tech University, Oshawa, ON, Canada

[4]Department of Mathematics and Statistics, York University, Toronto, ON, Canada

[5]Department of Electrical, Computer, and Biomedical Engineering, Toronto Metropolitan University, Toronto, ON, Canada

[6]CHU Sainte-Justine Azrieli Research Center, Université de Montréal, Montréal, QC, Canada

[7]Mila - Quebec Artificial Intelligence Institute, Montréal, QC, Canada

[8]Department of Psychiatry and Addictology, Faculty of Medicine, Université de Montréal, Montréal, QC, Canada

[9]Department of Psychiatry, Institut Universitaire en Santé Mentale de Québec, Montréal, QC, Canada

[*]these authors contributed equally

Corresponding Author:
Venkat Bhat, MSc, MD
Department of Psychiatry, AI for Mental Health, St. Michael's Hospital, 30 Bond Street, 17-009 Cardinal Carter South, Toronto, ON, Canada

Abstract

Background: Canadian psychiatry residents must demonstrate consultation competency, assessed using the standardized assessment of a clinical encounter report (STACER). However, opportunities to practice these skills and receive constructive assessment remain limited in clinical settings.

Objective: This study aimed to evaluate the technical feasibility of an agentic AI system designed to support psychiatry residents' consultation competence through simulated patient encounters with a patient agent and structured feedback from a rater agent.

Methods: We conducted a two-phase technical feasibility prospective single-arm cohort study of the STACER Agentic System, a large language model–based platform integrating a patient agent and a rater agent. Phase 1 involved automated evaluation of the patient agent using a psychiatrist agent across 227 synthetic major depressive disorder cases. Performance was assessed using DeepEval metrics (correctness, clarity, medical faithfulness, turn relevance, and role adherence) with descriptive statistics and 95% CIs. Phase 2 involved a preliminary user study with 14 convenience-sampled participants: a total of 5 members of the clinical research team and 9 psychiatry residents from the University of Alberta. Participants completed simulated diagnostic interviews and case presentations. Performance was evaluated using STACER-based scoring by the rater agent and 2 psychiatrists. Interrater reliability was assessed using intraclass correlation coefficients (α=.05). Participants rated realism, behavioral consistency, psychiatric nuance, and feedback utility using Likert scales and free-text answers.

Results: The patient agent demonstrated high behavioral (51/56, 91.07%) and symptom fidelity (105/110, 95.45%), with strong automated performance (medical faithfulness mean 0.99, 95% CI 0.99 - 1.00; turn relevance 0.99, 95% CI 0.986 - 0.992). Participants rated simulations as psychiatrically plausible and diagnostically useful, particularly for depressive symptom representation, although rapport building was moderate (mean 2.78, SD 1.56 to mean 3.00, SD 1.41, out of 5.00) due to limited

XSL•FO
RenderX

nonverbal cues. The rater agent generated structured STACER-aligned feedback with high intrarater consistency, especially at the section subtotal level. Interrater reliability with psychiatrists was poor at the item level (intraclass correlation coefficient range=0.25 - 0.49) but improved to good-to-excellent agreement at the section level for psychiatry resident sessions (intraclass correlation coefficient range=0.89 - 0.93). The rater agent's scores fell between those of the 2 psychiatrists for the clinical research team and were lower than both human raters for psychiatry residents.

Conclusions: The STACER Agentic System demonstrates the technical feasibility of using agentic AI to simulate psychiatric consultations and deliver STACER-aligned formative feedback. By combining adaptive multiturn psychiatric simulation with competency-based evaluation, it shows promise in supporting cognitive aspects of consultation, though it remains limited in facilitating relational skills such as rapport building. These findings suggest agentic AI could expand scalable, low-risk opportunities for deliberate practice and formative feedback in competency-based psychiatric education. Further controlled studies are needed to evaluate educational effectiveness and integration into residency training.

(JMIR Med Educ 2026;12:e88580) doi:10.2196/88580

KEYWORDS

psychiatry education; standardized assessment of a clinical encounter report; STACER; agentic AI system; chatbot; consultation skill; psychiatric residents; competence

Introduction

Psychiatric consultation requires advanced interviewing skills, diagnostic reasoning, and the ability to integrate complex psychosocial information into treatment planning [1,2]. Because the clinical interview and the mental status examination (MSE) remain the cornerstone of psychiatric diagnosis, residents must develop these competencies through repeated, feedback-rich practice [3,4]. Yet, training environments often limit opportunities for sustained supervision and deliberate rehearsal [5,6], leaving residents to encounter challenging cases without sufficient low-stakes practice [7]. Standardized patients offer structured opportunities for experiential learning, but their use is resource-intensive, difficult to scale, and typically limited to major assessments or dedicated simulation events [6,8,9]. As a result, opportunities for iterative practice and formative feedback remain limited, despite their importance for competency development in medical education [10].

In the context of cognitive learning taxonomies such as Bloom's Taxonomy, formative feedback plays a central role in helping learners move beyond basic recall toward higher-order competencies, including analysis, synthesis, and clinical reasoning [11]. For psychiatric interviewing specifically, repeated low-stakes practice enables learners to progress from identifying key symptoms (remember and understand) to formulating differential diagnoses (analyze) and developing treatment plans (evaluate and create). Simulation-based formative feedback therefore supports the deeper cognitive processes required for psychiatric consultation competence by providing structured, iterative opportunities to test and refine reasoning. This mismatch between training needs and available learning opportunities contributes to persistent gaps in consultation skill development.

Digital simulation has emerged as a promising avenue to address this challenge, with AI increasingly enabling applications in diagnosis, risk assessment, and educational evaluation [12,13]. Within this context, virtual patient systems represent a key application of AI-driven simulation in medical education. Early virtual patient systems demonstrated that computer-mediated psychiatric dialogues were feasible and effectively supported specific tasks (eg, obtaining informed consent for antipsychotics) to improve trainee confidence and knowledge in controlled studies [14], as well as enhance diagnostic reasoning [15] and skill acquisition [16]. However, their rule-based logic and fixed response structures constrained realism and adaptability. Case-specific virtual patients, such as Virtual Justina (University of Southern California Institute for Creative Technologies), a virtual adolescent with posttraumatic stress disorder [17], provided helpful structured practice but required extensive manual scripting and lacked the flexibility necessary for diverse, dynamic encounters. As these systems evolve, it is essential that performance measurement for agentic AI systems be aligned with the agent's expected role. In educational contexts, this means evaluating the system as a training tool rather than as a therapeutic provider. For example, when an AI system is used to help learners practice cognitive behavioral therapy interviewing, its evaluation should focus on whether its responses appropriately model cognitive behavioral therapy–consistent communication and reasoning, rather than implying that the system functions as an actual clinical cognitive behavioral therapy provider [18].

Recent advances in large language models (LLMs) have transformed the capabilities of digital simulation. Transformer-based LLMs [19], trained on terabytes of data, demonstrate improved realism and adaptability in simulating a wide range of patient scenarios relevant to consultation practices [20] and education [21,22]. LLM-generated simulations demonstrate clinically relevant psychiatric exchanges [23-25] and generate automated feedback aligned with expert judgment [26], offering a cost-efficient and scalable alternative to real patients for clinical training. Newer models outperform their predecessors by a significant margin [27]. Tools such as MedSimAI (UCSF [University of California, San Francisco] School of Medicine, Weill Cornell Medicine, Yale School of Medicine, and Ohio State University), which uses GPT-4o to simulate patient encounters and provide immediate formative feedback [28], exemplify the rapid evolution of AI-enabled clinical education. However, current platforms are not designed specifically for psychiatric consultation and often lack validated psychiatric symptom realism, emotional and behavioral nuance,

XSL•FO
RenderX

structured MSE-level detail, or alignment with specialty-specific competency frameworks.

In Canada, psychiatry residents' consultation competence is formally assessed using the structured assessment of a clinical encounter report (STACER) [29], which evaluates performance on the MSE, the DSM-5 (*Diagnostic and Statistical Manual of Mental Disorders* [Fifth Edition])-aligned diagnostic formulation [30], and treatment planning. STACER assessments are infrequent and predominantly summative, limiting opportunities for repeated practice. Kolb's experiential learning cycle [31] underscores that skill development requires iterative cycles of concrete experience, reflective observation, conceptual integration, and active experimentation, conditions difficult to satisfy when assessment opportunities occur only episodically.

To address these limitations and build on our previous works in generative AI and prompt-based reasoning [32,33], we developed the STACER Agentic System (SAS), an LLM-powered simulation and assessment platform designed to support psychiatric consultation training. SAS integrates 2 coordinated AI agents: a patient agent, which simulates realistic multiturn psychiatric interviews, and a rater agent, which produces structured formative feedback aligned with STACER criteria. By unifying simulation and assessment within a single environment, SAS aims to create a scalable pathway for repetitive practice, self-directed learning, and progressive refinement of consultation skills.

The objective of this technical feasibility study was to evaluate SAS's ability to (1) simulate clinically realistic psychiatric interactions and (2) generate structured, STACER-aligned formative feedback to support the development of consultation competence.

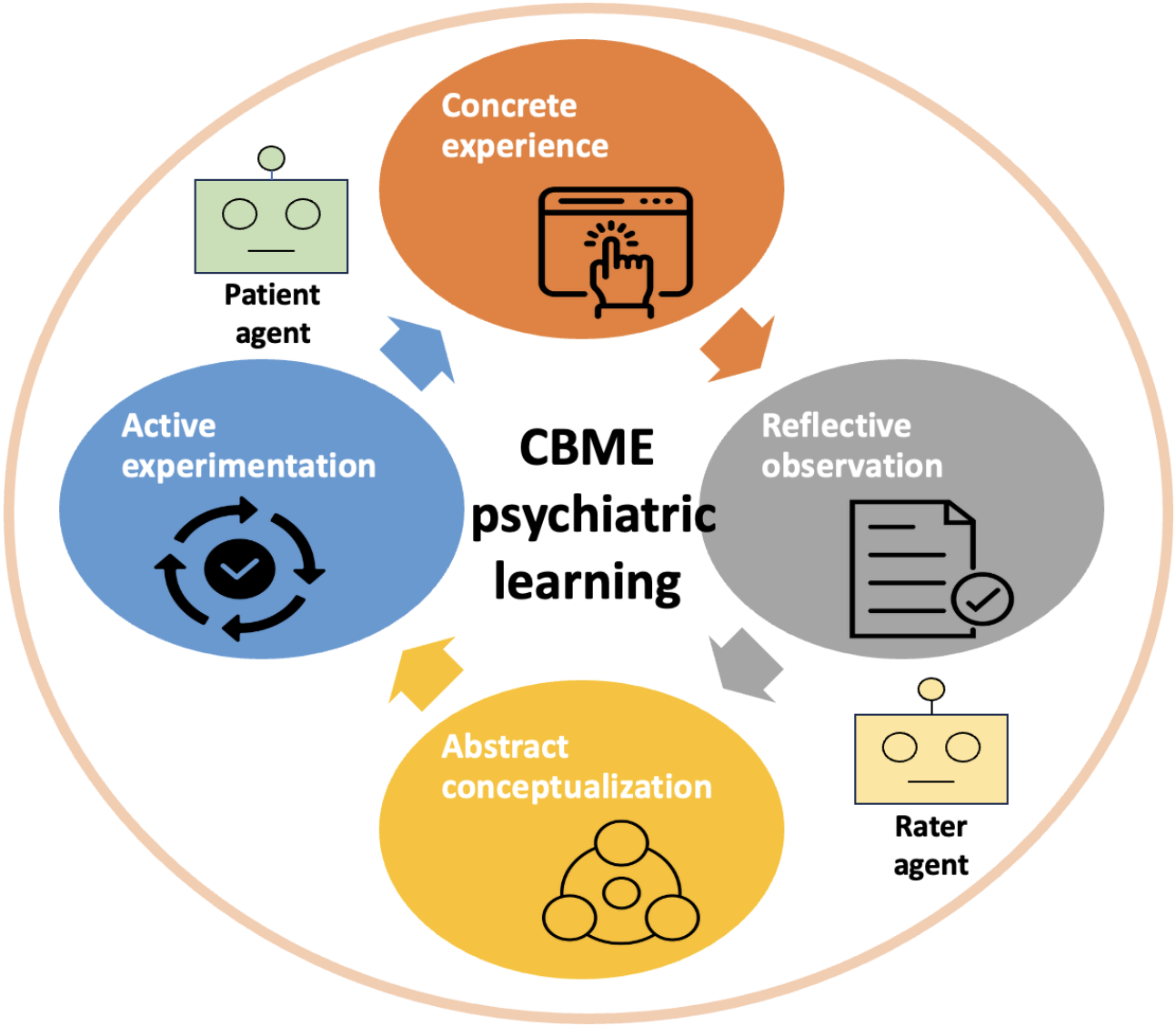
Methods

SAS Overview

The SAS is an LLM-powered educational platform designed to provide psychiatry residents with structured, simulation-based practice in consultation interviewing. Grounded in principles of simulation-based education [34], the system operationalizes the Kolb experiential learning cycle [31] by coordinating 2 complementary agents: a patient agent, which provides the concrete experience of a simulated psychiatric interview, and a rater agent, which supports reflective observation and abstract conceptualization through structured formative assessment, narrative feedback, and guided cross-examination [10]. Figure 1 illustrates how these agents collectively enact the 4 stages of the Kolb cycle to support competency-based medical education (CBME) [35,36].

XSL•FO
RenderX

Figure 1. Illustration of how the patient agent and rater agent map onto the four stages of the Kolb experiential learning cycle within a competency-based medical education psychiatric training context. The patient agent supports concrete experience by enabling simulated clinical encounters, while the rater agent contributes to reflective observation through structured assessment. Learners progress to abstract conceptualization by integrating feedback into clinical understanding and to active experimentation by applying these insights in subsequent simulations. CBME: competency-based medical education.



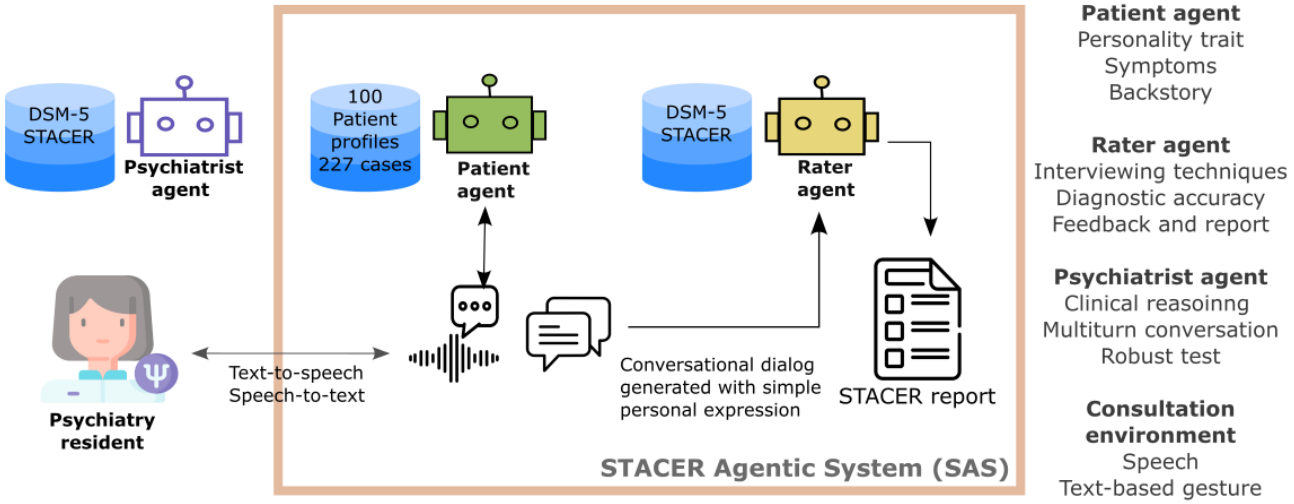
The patient agent simulates psychiatric presentations derived from detailed patient profiles, while the rater agent acts as a formative, stabilizing comparator and produces structured formative feedback aligned with the Core of Discipline-STACER form [29]. A verbal chat interface allows residents to ask clarifying questions and engage in reflective dialogue. The system was implemented using LangGraph (LangChain, Inc) [37], an open-source framework that allows for the creation and coordination of fully customizable multiagent systems using graph-based workflows. Patient dialogue is generated using a narrative sense-making process, which draws on research showing that language models naturally use storytelling to make incomplete information feel coherent and realistic [38]. This helps the patient agent produce responses that resemble how real patients naturally describe their experiences and is delivered through integrated audio and speech-to-text tools (supported through ElevenLabs' AI Audio Application Programming Interface) [39]. For details, please refer to the GitHub repository for the workflow code and the prompt files in Multimedia Appendix 1. Detailed descriptions of the patient agent and the rater agent can be found in Multimedia Appendix 2.

Study Design

This study used a two-phase prospective single-arm cohort study design consisting of (1) an automated technical validation and (2) feasibility testing with human participants (Figure 2). To ensure transparent and comprehensive reporting of this LLM-based system, we adhered to the Development, Evaluation, and Assessment of Large Language Models guidelines [40]; the completed Development, Evaluation, and Assessment of Large Language Models checklist is provided in Checklist 1.

XSL•FO
RenderX

Figure 2. The structured assessment of a clinical encounter report agentic system uses the *DSM-5* (*Diagnostic and Statistical Manual of Mental Disorders* [Fifth Edition]), depression diagnostic criteria, and treatments to provide a simulated environment for psychiatric residents to practice their consultation skills and clinical competency. The patient agent mimics a patient from one of the case studies stored in the patient database, which contains 100 real cases and 227 synthetic cases. The rater agent evaluates the consultation session, provides feedback to the resident, and generates a structured assessment of a clinical encounter report. For technical validation, an external psychiatrist agent was developed to conduct automated multiturn interviews with the patient agent. Grounded in the Core of Discipline-structured assessment of a clinical encounter report expectations, it standardized the elicitation of clinically relevant information across synthetic profiles before human user-testing. STACER: structured assessment of a clinical encounter report.



Participants and Setting

Participants included 2 groups: (1) members of the clinical research team and (2) psychiatry residents. The clinical research team consisted of 5 individuals, including 1 student and 4 postdoctoral fellows in psychiatry or psychology. These individuals simulated the role of psychiatry residents during feasibility testing. The second group consisted of 9 psychiatry residents enrolled in the University of Alberta Psychiatry Residency program.

Testing was conducted in 2 settings. Clinical research team members were provided with instructions to complete the simulation sessions in person using a laboratory computer at St. Michael's Hospital. Psychiatry residents were provided login credentials without further instructions and completed the sessions remotely using the online platform.

No formal sample size calculation was performed, as the study aimed to evaluate feasibility and preliminary system performance rather than to test a hypothesis or detect between-group differences. Participants were recruited using convenience sampling.

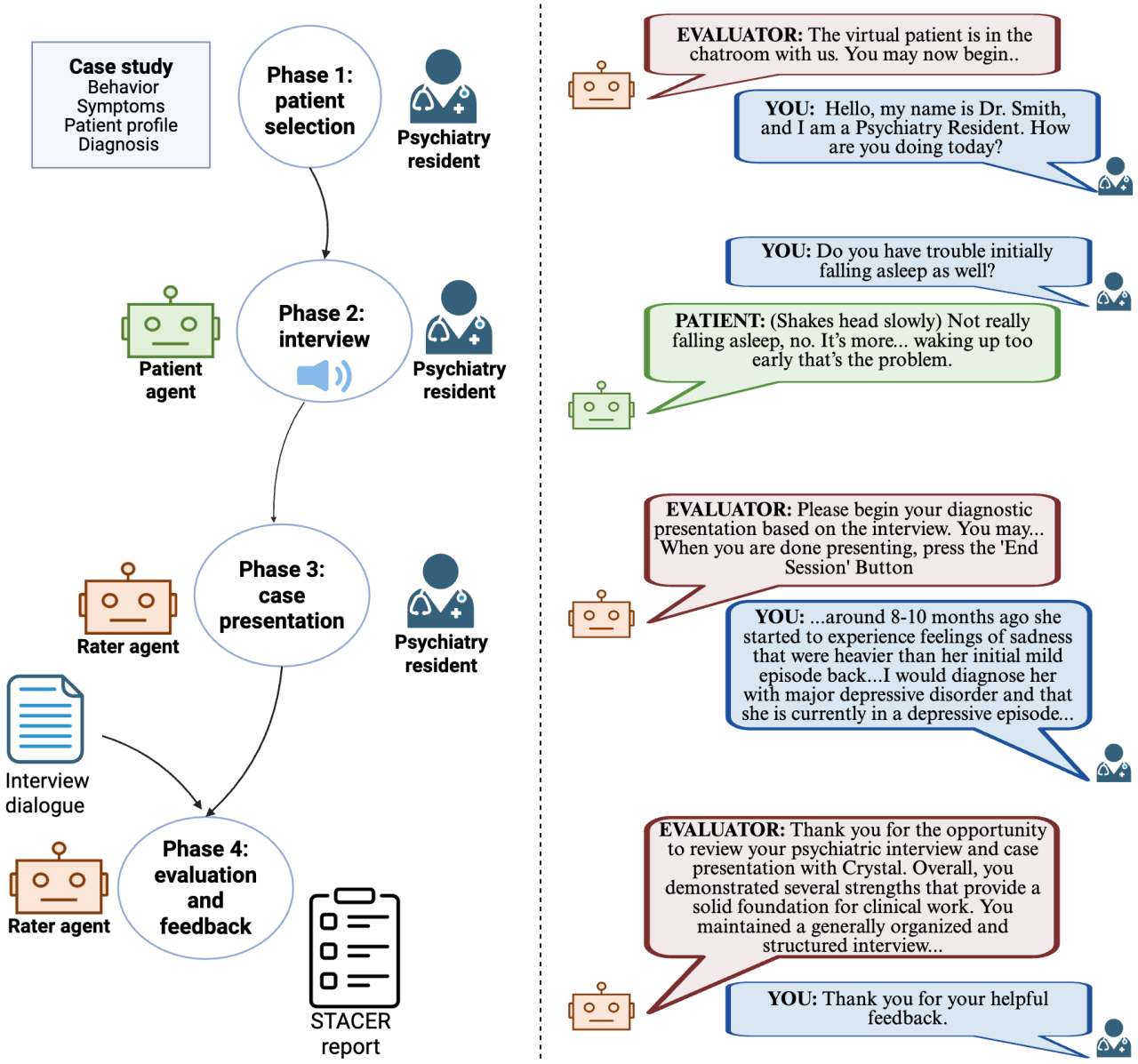
Study Procedure

For technical validation, a psychiatrist agent external to SAS was developed to conduct automated multiturn interviews with the patient agent. The psychiatrist agent, grounded in Core of Discipline-STACER expectations and informed by the *DSM-5*, ensured standardized elicitation of clinically relevant information across all synthetic profiles. Automated testing on the synthetic dataset was performed prior to human user testing.

The feasibility testing unfolded across 4 sequential phases as illustrated in Figure 3: (1) the participant selects a patient profile to interview; (2) a simulated consultation occurs with the psychiatric patient agent (powered by GPT-4.1-mini), mirroring the 55-minute STACER interview; (3) the participant delivers a 20-minute case presentation to the rater agent; and (4) the rater agent (powered by GPT-4.1) combines the case study file, interview dialogue, presentation, and *DSM-5* diagnostic criteria to offer feedback presented in a Core of Discipline-STACER assessment form. A chat interface was provided for participants to converse with the rater agent about the feedback, explore opportunities to extract new knowledge, and engage in active reflection. Simulation results from the 14 participants were manually analyzed and compared with the actual behavior and symptoms of the cases.

XSL•FO
RenderX

Figure 3. Structured assessment of clinical encounter report agentic system workflow. The left panel shows the four phases: (1) patient selection, (2) simulated interview, (3) case presentation, and (4) assessment and feedback. The right panel illustrates interactions between the psychiatry resident and the structured assessment of clinical encounter report agentic system. The rater agent loads the case and provides an introduction (phase 1). The resident conducts a multi-turn psychiatric interview with the patient agent (phase 2), followed by a case presentation to the rater agent (phase 3). The rater agent then integrates all inputs to generate structured, structured assessment of clinical encounter report agentic system–aligned feedback and a Core of Discipline-Structured assessment of clinical encounter report agentic system assessment, with a chat interface supporting reflection and iterative learning (phase 4). STACER: structured assessment of clinical encounter report.



Participants rated the patient agent on realism, psychiatric nuance, and rapport building using a Likert-type scale (1=very poor to 5=excellent) and through a free-text option. Two practicing psychiatrists independently reviewed each simulation and rater agent output, providing parallel STACER assessments for comparison.

Synthetic Benchmark Dataset

Existing datasets are not suited for multiturn psychiatric consultations [41-43], so we developed a comprehensive synthetic benchmark of 227 *DSM-5-Text Revision*–based [44] major depressive disorder cases to systematically evaluate patient agent performance (Table 1). A detailed synthetic case study is provided in Multimedia Appendix 3.

XSL•FO
RenderX

Table . Synthetic data summary.

	Age, mean (SD)	5 Symptoms	6 Symptoms	7 Symptoms	8 Symptoms	9 Symptoms
Female (n=127)	45.30 (16.96)	63	41	21	2	0
Male (n=100)	45.02 (15.56)	42	36	14	7	1
Total (N=277)	45.74 (16.3)	105	77	35	9	1

Real-Patient Case Study Dataset

A separate set of 100 manually curated case studies adapted from *DSM-5-Text Revision* Clinical Cases was created for testing. Overall, 5 noncomorbid major depressive disorder cases (5/100, 5%) were selected for feasibility evaluation. Each case included a narrative summary, behavioral description, depressive symptom list, patient profile, and confirmed diagnosis (Multimedia Appendix 3).

Performance Measure

Performance was evaluated separately for each agent using role-specific metrics, including automated testing for the patient agent using DeepEval (Confident AI); accuracy, clarity, professional tone, medical faithfulness, bias, toxicity, turn relevance, and role adherence), and participant sessions were used in reliability analyses for the rater agent. User feedback questionnaires were used to provide insights into the limitations of the learning experience. The rater agent's scoring was compared with the psychiatrists' using the intraclass correlation coefficient (ICC), which was computed using the *psych* package in R (version 4.5.1; The R Foundation for Statistical Computing). Standard interpretive thresholds were applied: ICC <0.50 for poor, 0.50 - 0.75 for moderate, 0.75 - 0.90 for good, and >0.90 for excellent reliability, respectively [45]. Fidelity of the patient agent's behavior and symptoms was then analyzed. Full details on performance measures are provided in Multimedia Appendix 4.

Ethical Considerations

This work was conducted as a preliminary technical feasibility and operational testing study of the SAS prototype prior to future controlled educational evaluation studies. Psychiatry residents and members of the clinical research team voluntarily participated as pilot testers. Participants were informed about the purpose and procedures of the study through email communication and voluntarily chose whether to participate. Given the preliminary operational testing nature of this work, the study focused on prototype functionality and usability assessment rather than formal evaluation of educational or clinical outcomes. Formal research ethics board review was not sought because this work was conducted as preliminary operational and technical feasibility testing of a prototype system prior to future controlled educational evaluation studies.

No personal health information or sensitive participant information was collected. Participant initials were used only for study coordination purposes and were removed prior to analysis. All transcripts, ratings, and feedback were deidentified before reporting to protect participant privacy and confidentiality. No financial compensation or incentives were provided for participation. No identifiable participant information or images are included in the manuscript, figures, or supplementary materials, and no participant assessments or feedback can be linked to individual residents. Future studies designed to formally evaluate learner outcomes, educational effectiveness, or clinical performance will undergo formal research ethics review prior to implementation.

Results

Patient Agent Performance

Using the synthesized patient profile and the psychiatrist agent, the Patient Agent's behavior was validated with the eight DeepEval evaluation metrics. Table 2 and Figure 4 summarize the 8-performance metrics, including mean, SD, 95% CI, token cost, and success rate.

XSL•FO
RenderX

Table . Patient agent's performance measure, range 0 - 1, using a psychiatrist agent to elicit responses.

	Mean (SD)	95% CI	Total cost (in US $)[a]	Success rate
Correctness[b]	0.87 (0.05)	0.86 - 0.88	0.54	1
Clarity[c]	0.89 (0.06)	0.89 - 0.90	0.84	1
Professionalism[d]	0.42 (0.16)	0.40 - 0.44	0.85	0.24
Medical faithfulness[e]	0.99 (0.01)	0.99 - 1.00	1.04	1
Bias[f]	0.001 (0.01)	0 - 0.002	1.69	1
Toxicity[g]	0 (0)	—[h]	1.72	1
Turn relevance[i]	0.99 (0.02)	0.986 - 0.992	6.36	1
Role adherence[j]	0.99 (0.10)	0.97 - 1.00	1.08	0.98

[a]Total token cost provided by DeepEval. The total token cost from OpenAI exceeded US $100 due to the more expensive GPT-5 and repeated requests due to a time-out.

[b]Correctness: overall correctness of the patient profile representation.

[c]Clarity: clarity of the patient's expression.

[d]Professionalism: professional tonality.

[e]Medical faithfulness: medical faithfulness of the symptom presentation.

[f]Bias: indication of bias in the utterance.

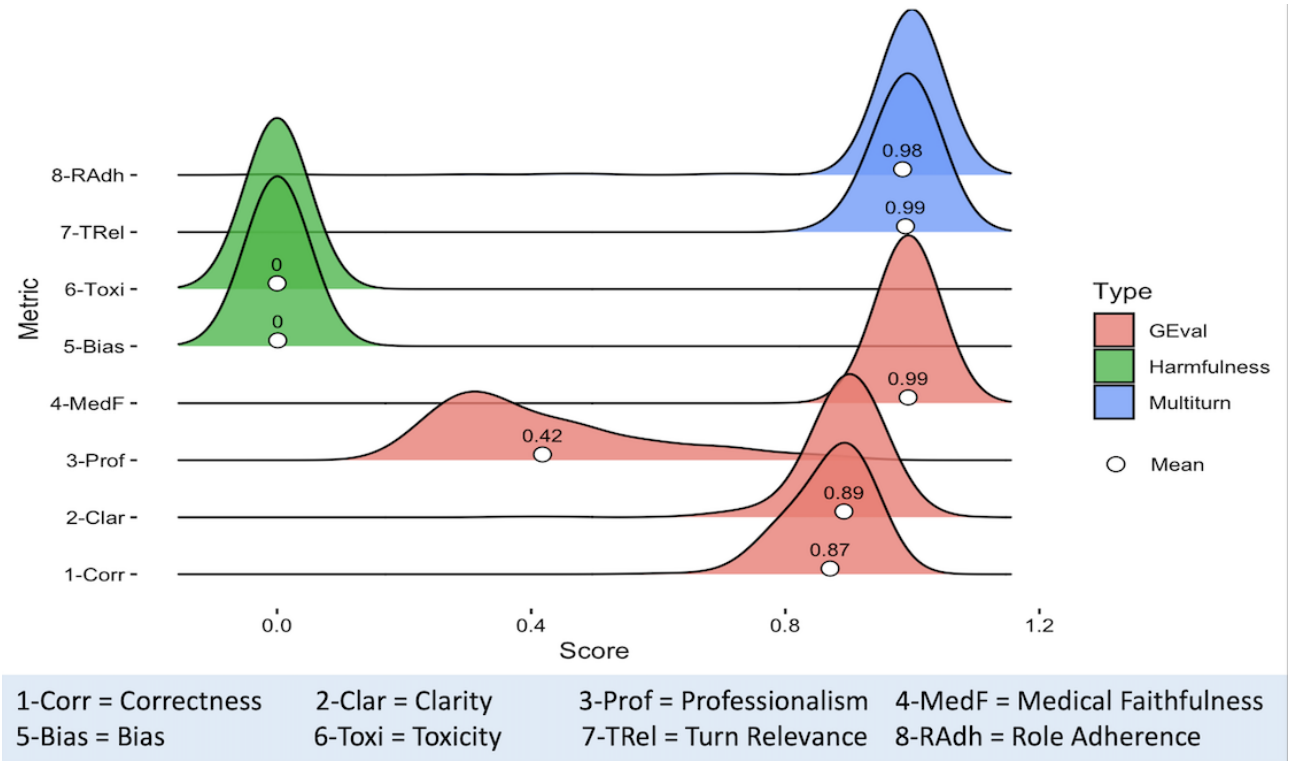
[g]Toxicity: amount of toxicity created.

[h]Not applicable.

[i]Turn relevance: relevancy between each turn (Patient-to-Psychiatrist and the subsequent Psychiatrist-to-Patient, and vice versa).

[j]Role adherence: patient role adherence over the multiturn conversation.

Figure 4. The eight DeepEval performance metrics are colored by the types of metrics, with mean values provided. The density plots indicate that the mean score of the patient agent playing the role of a patient with major depressive disorder using the synthetic dataset: medical faithfulness, turn relevance, and role adherence above 0.98; correctness and clarity above 0.87; and bias and toxicity close to 0. The distribution of scores for the eight metrics with red indicate the GEval measures, green indicates the harmfulness measures, and blue indicates the multiturn measures.



XSL•FO
RenderX

User Experience With the STACER Agentic System

Detailed ratings from participants' first-session responses are presented in Table 3. Participants indicated that the patient agent effectively expressed depressive symptomatology. However, rapport building was more limited (3/5). Ratings of the patient agent's ability to convey emotional distress varied between the 2 groups (4.6/5 vs 2.78/5). Using voice and text-based nonverbal cues alone, both the patient and rater agents demonstrated moderate levels of emotional expressiveness.

Table . Quantitative data on user feedback of the first sessions, rated from 0 (very poorly) to 5 (excellent).

Items	Clinical team (n=5)[a]		Residents (n=9)[b]	
	Mean (SD)	Range	Mean (SD)	Range
Q1[c]. The patient effectively conveyed emotional distress.	4.60 (0.55)	4 - 5	2.78 (1.56)	0 - 5
Q2. The patient's responses matched typical depressive symptomatology.	4.80 (0.45)	4 - 5	5.00 (0.00)	5 - 5
Q3. It was easy to build rapport with the patient.	3.00 (1.41)	1 - 4	2.78 (1.56)	0 - 5
Q4. The patient's emotional expressiveness improved my diagnostic reasoning.	3.60 (0.55)	3 - 4	3.11 (1.45)	0 - 5
Q5. The rater's emotional expressiveness improved my learning and diagnostic reasoning.	3.60 (0.55)	3 - 4	3.33 (1.12)	1 - 5

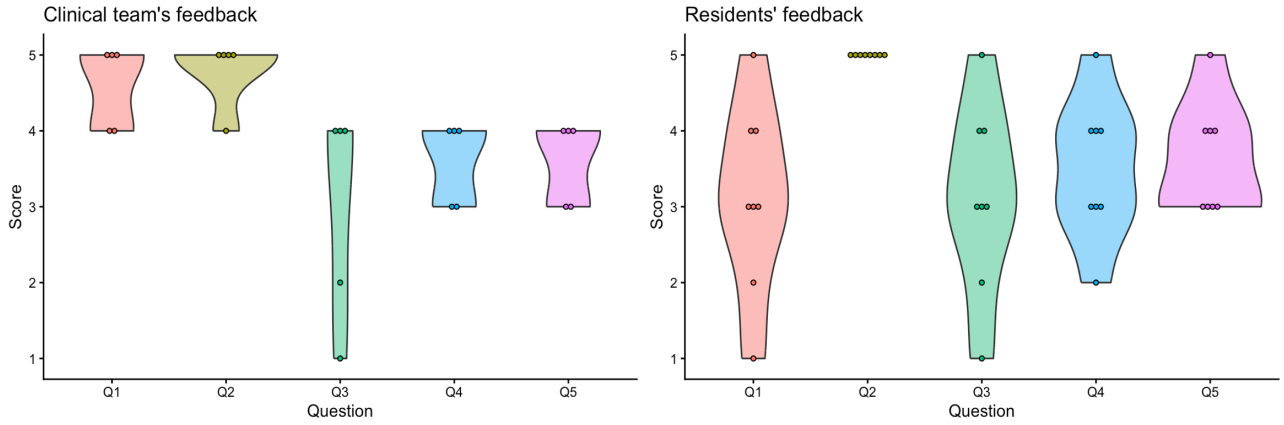
[a]The tests for the clinical team were conducted using text as an input.

[b]No instructions were provided to the psychiatry residents that both voice and text were supported. Four residents used text as the input and the rest used voice.

[c]Q: question.

Figure 5 illustrates the differences between the 2 participant groups. One of the psychiatry residents assigned 3 scores of "0" and one score of "1," citing limitations related to typing. The other scores of "1" were given by 1 resident who noted the absence of visual cues and by another who highlighted typing constraints alone.

Figure 5. Violin plot of participants' feedback. Each participant's response to the survey questions in Table 3 is indicated by a point. (A) Feedback from the clinical research team with instructions provided on the flow of the user interface. (B) Psychiatry residents' feedback when no instructions were provided. Four of the 9 participants were unaware of the speech capability and used only text. The scores of 0 and 1 were from the text-only participants. Q: question.



Free-text feedback indicated that participants found the SAS both informative and educational, not only because of the feedback provided but also due to the ability to ask questions to the rater agent, which enhanced their learning experience and supported their planning. However, the absence of visual cues limited the expression of emotions and hindered rapport building, as facial expressions, body language, and overall appearance are important components of communication.

Performance Evaluation of the Rater Agent

We measured the consistency of the rater agent at both the item-level scores and the section subtotals. The one-way ICC(1) model, where each target is rated by a different judge and judges are randomly selected, was reported because it yields the most conservative (lowest) correlation estimates. Table 4 summarizes the intrarater reliability of the rater agent across all runs for each recorded session. One of the psychiatry residents (resident 7)

XSL•FO
RenderX

did not complete the STACER process, and the session was discarded in the analysis. At the item level, assessments for both groups demonstrated ICCs in the moderate-to-high range (4/5 and 8/8). Reliability improved when item scores were aggregated into section subtotals, with 2 of 5 and 7 of 8 in excellent agreement. Missing data adversely affected reliability estimates, as observed for participants 4 and 5. Resident 7 did not complete the interview and presentation and was excluded from the analysis. The missing data are summarized in Table 4. All missingness belonged to the same items: interrupts politely when required, redirects when required, facilitates organization of disorganized patients, engages patients in a culturally safe manner, attends to timing, and provides a pertinent closing statement. Participant 5 had 3 additional items: nonverbal behavior encourages patients to tell their story, listens attentively, and note taking does not distract from the interview. As these items were not evident and some were impossible to capture from the transcripts, they were replaced with 0.

Table . Intrarater agent reliability at the item level and section subtotal, with missing scores given 0[a].

	Missingness (%)	Item-level score			Section subtotal		
		ICC[b]	*F* test (*df*)	95% CI	ICC	*F* test (*df*)	95% CI
Clinical research team (10 runs)							
Participant 1	0	0.80	39.82 (97, 882)	0.74-0.84	0.96	251.65 (4, 45)	0.89-1.00
Participant 2	0	0.89	78.19 (97, 882)	0.85 - 0.91	0.96	259.66 (4, 45)	0.89- 1.00
Participant 3	0	0.75	30.31 (97, 882)	0.67- 0.80	0.89	82.44 (4, 45)	0.72- 0.99
Participant 4[c]	1.22	0.45	9.10 (97, 882)	0.37-0.53	0.41	8.00 (4, 45)	0.14-0.87
Participant 5[c]	0.92	0.69	23.20 (97, 882)	0.62-0.75	0.88	72.39 (4, 45)	0.69- 0.98
Psychiatry residents (5 runs)[d]							
Resident 1	3.67	0.80	20.70 (97, 392)	0.74-0.85	0.96	119.38 (4, 20)	0.87-1.00
Resident 2	6.12	0.78	18.46 (97, 392)	0.72-0.83	0.94	80.33 (4, 20)	0.81-0.99
Resident 3	2.45	0.77	18.01 (97, 392)	0.71-0.83	0.94	77.52 (4, 20)	0.81-0.99
Resident 4	3.67	0.59	8.31 (97, 392)	0.51-0.68	0.77	17.46 (4, 20)	0.44-0.97
Resident 5	2.45	0.86	31.20 (97, 392)	0.82- 0.89	0.98	321.59 (4, 20)	0.95-1.00
Resident 6	6.12	0.66	10.73 (97, 392)	0.58-0.74	0.96	122.58 (4, 20)	0.87- 1.00
Resident 8	2.45	0.73	14.37 (97, 392)	0.66-0.79	0.94	85.39 (4, 20)	0.82-0.99
Resident 9	3.76	0.84	27.80 (97, 392)	0.80-0.88	0.992	638.43 (4, 20)	0.97-1.00

[a]All *P* values are <.001.

[b]ICC: intraclass correlation coefficient.

[c]Missing data.

[d]Resident 7 was not reported, as the interview and case presentation transcripts indicate that the conversation contains only greetings and a short paragraph in the case presentation. The rater agent has given all zero scores to resident 7.

Table 5 summarizes the interrater reliability between the 2 psychiatrists (P1 and P2) and between each psychiatrist and the rater agent. At the item level, all comparisons demonstrated poor agreement, with ICCs below 0.50. The 2 groups exhibited very different results at the section subtotal level. For the clinical team, 1 psychiatrist showed moderate agreement with the rater agent, while the other achieved good agreement. The agreements for the psychiatry residents were much higher, from 0.89 to 0.93, indicating good-to-excellent agreements.

XSL•FO
RenderX

Table . Inter-rater reliability between the two practicing psychiatrists and the rater agent[a].

	Item-level score			Section subtotal		
	ICC[b]	*F* test (*df*)	95% CI	ICC	*F* test (*df*)	95% CI
Clinical research team (5 participants, 10 runs)						
P1 versus P2	0.25	1.66 (489, 490)	0.16-0.33	0.42	2.47 (24, 25)	0.05-0.70
P1 versus rater	0.39	2.26 (489, 490)	0.31-0.46	0.75	7.02 (24, 25)	0.52-0.88
P2 versus rater	0.35	2.08 (489, 490)	0.27-0.43	0.58	3.78 (24, 25)	0.26-0.79
Psychiatry residents (8 participants, 5 runs)[c]						
P1 versus P2	0.19	1.48 (775, 776)	0.12-0.26	0.93	28.22 (44, 45)	0.88-0.96
P1 versus rater	0.25	1.67 (775, 776)	0.18-0.31	0.89	18.00 (44, 45)	0.82-0.94
P2 versus rater	0.49	2.92 (775, 776)	0.43-0.54	0.92	24.28 (44, 45)	0.86-0.96

[a]All *P* values are <.001

[b]ICC: intraclass correlation coefficient.

[c]Resident 7 was not reported, as the interview and case presentation transcripts indicate that the conversation contains only greetings and a short paragraph in the case presentation. The rater agent has given all zero scores to Resident 7.

Discrepancies between raters were also evident in overall scoring patterns, especially for the clinical research team. One psychiatrist rated scores 51% higher than the rater agent, whereas the other rated scores 33% lower. The rater agent's scores consistently fell between those of the 2 psychiatrists. On the other hand, both human raters scored 14% higher than the rater agent for the psychiatry residents.

As shown in Tables 4 and 5, aggregating item scores into section subtotals generally improved agreement. Section-level aggregation provides a broader impression of skill categories rather than focusing on individual items, thereby reducing instability arising from ambiguity or differing interpretations of item definitions.

Linking Back the Patient Agent's Behavior and Symptom Expressions

As the participants were given the choice to select the case studies, based only on the case name, the number of behavior traits and symptoms varies (summarized in Table 6). A total of 4 case studies (a-d) were selected by more than 1 tester. Case studies (b) and (c) achieved 100% in behavioral expressions, while case studies (a) and (d) achieved 87.50% (7/8) and 92% (11/12), respectively. Case studies (a), (b), and (d) achieved 100% symptom expression, while case study (c) achieved 88.89% (16/18). Resident 3 selected a Schizophrenia patient profile where behaviors could not be mimicked in the consultation. Resident 9 selected a more complicated comorbid case study. The overall behavioral fidelity across both groups was over 91.07% (51/56), and symptom fidelity reached 95.45% (105/110), consistent with the high medical faithfulness (0.993) from the automated test.

XSL•FO
RenderX

Table . Patient agent's behavior and symptom expressions.

Subject	Behavior	Symptoms
Clinical team		
Participant-1[a]	3/4	7/7
Participant-2[a]	4/4	7/7
Participant-3[a]	4/4	9/9
Participant-4[a]	4/4	7/7
Participant-5[a]	4/4	7/7
Overall, score (%)	19/20 (95%)	37/37 (100%)
Residents[b]		
Resident-1[a]	3/3	9/9
Resident-2	4/4	7/7
Resident-3	5/7	6/7
Resident-4[a]	3/4	9/9
Resident-5[a]	3/3	7/9
Resident-6[a]	4/4	9/9
Resident-8	3/4	6/6
Resident-9[c]	6/6	15/17
Overall, score (%)	32/36 (88.89)	68/73 (93.15)

[a]These are 4 case studies that have been used in the simulation multiple times.

[b]Resident 7 was not reported, as the interview and case presentation transcripts contain insufficient content for analysis.

[c]The long list of symptoms is due to comorbidity.

Discussion

This study evaluated the technical feasibility of an agentic AI platform designed to simulate psychiatric consultations and deliver structured formative feedback aligned with the STACER framework. The findings demonstrate that LLM-driven agents can emulate clinically coherent psychiatric dialogue, support structured feedback, and create an accessible environment for deliberate practice. At the same time, the results highlight specific constraints, including rapport formation and multimodal fidelity, that shape how agentic systems should be incorporated into psychiatric education. These findings suggest that while AI-supported simulation may effectively extend opportunities for experiential learning, there are areas requiring refinement and testing before widespread curricular integration.

Fidelity and Educational Value of the Patient Agent

The patient agent exhibited high psychiatric fidelity, maintaining diagnostic, behavioral, and linguistic coherence across multiturn interviews. This aligns with evidence that LLM-based virtual patients can support clinically realistic interactions in domains such as communication and history taking [28]. For instance, MedSimAI demonstrated that GPT-4o can sustain coherent exchanges, allowing learners to rehearse clinical questioning strategies [28]. Similarly, Virtual Justina showed that structured, *DSM* (*Diagnostic and Statistical Manual of Mental Disorders*)-based simulations can facilitate practice with sensitive symptom domains, including trauma-related questioning and rapport building [17]. The SAS builds on these foundations by coupling psychiatric realism with adaptive, multiturn dialogue and structured competency mapping aligned to STACER domains, while offering more sustained internal consistency across extended interviews than earlier systems, which were constrained by fixed dialogue trees [17]. The near-perfect turn relevance and role adherence further support its role as a reliable platform for the "concrete experience" stage of experiential learning.

A key finding from both quantitative and qualitative data is that rapport formation was consistently limited (2.78 - 3.00 out-of 5 across groups). Rather than a minor usability issue, this represents a defining boundary of current agentic AI systems: while effective for supporting the cognitive components of psychiatric consultation (eg, history taking and diagnostic reasoning), they are less suited for affective and relational skills [17,28,46]. This limitation is further clarified when mapped onto the STACER framework [29]. The patient agent demonstrated strong alignment with domains related to information gathering, diagnostic formulation, and MSE structure; however, lower rapport scores indicate reduced capacity to support domains involving affect, engagement, empathy, and relational skills. These competencies rely on subtle interpersonal and embodied cues, particularly nonverbal

XSL•FO
RenderX

communication, that are not fully captured in text- and voice-based interfaces.

These findings are consistent with prior literature indicating that embodied conversational agents vary widely in their use of human communication modalities; although some systems incorporate verbal and nonverbal behaviors, many rely on simpler, less expressive designs [46]. Because advanced multimodal interaction is technically demanding and still seldom evaluated in clinical settings, the extent to which embodied conversational agents can support human-like therapeutic connection remains uncertain. Complementing this, Tay et al [47] found that immersive and augmented reality environments enhance empathy and reduce stigma by enabling learners to experience affective and perceptual dimensions of mental illness. Incorporating audio-prosodic variation, avatar-based facial animation, or sensor-based detection of the learner's nonverbal cues may partially address this limitation by creating a more ecologically valid simulation of the mental status examination and the relational aspects of psychiatric care [48]. Similarly, any multimodal extension would also need to capture and interpret the resident's own nonverbal communication, as authentic rapport relies on bidirectional cues [49].

Performance and Bias in the Rater Agent

The rater agent produced detailed, criterion-referenced feedback and structured STACER scoring, reflecting the growing literature demonstrating that LLM-based evaluators can generate coherent assessments [26,28]. Prior work by Holderried et al [26] showed that GPT-4 achieved "almost perfect" agreement with human raters in history-taking evaluations, while Hicke et al [28] reported that LLM feedback enhanced learner reflection and engagement in simulated patient interactions. In this study, all participants perceived the rater agent's feedback as specific, actionable, and aligned with expectations for psychiatric assessment. These findings, echoed by Cook et al [23], who observed that LLM-powered virtual patients generated personalized, high-quality performance feedback rated comparably to human evaluators, substantiate the potential of AI-assisted formative assessment in clinical education.

With the use of a more structured and regimented prompt, the rater agent demonstrated improved rating consistency, particularly at the section subtotal level, where agreement ranged from moderate to excellent across both participant groups. Notably, agreement between the rater agent and human raters was higher in the psychiatry resident group than in the clinical research team, suggesting that the system may better approximate expected scoring patterns within its intended learner population. For the clinical research team, the rater agent's scores consistently fell between those of a more generous and a stricter human rater, indicating a stabilizing, middle-ground evaluation rather than systematic bias. In contrast, for psychiatry residents, both human raters consistently assigned higher scores than the rater agent. This shift, from disagreement in overall scoring for the clinical research team to consistently higher ratings by both psychiatrists for residents, may reflect a favorable bias toward psychiatry residents, potentially influenced by expectations of trainee competence [50]. Blinding raters to participant groups would have helped mitigate this possibility and represents an important consideration for future study design.

Contrary to concerns raised in MedSimAI [28], no evidence of a "supportive" bias was observed. Through iterative prompt engineering, we mitigated several known limitations of LLM-based evaluators, including (1) a predisposition toward overly supportive or affirming judgments [51], (2) overreliance on linguistic polish as a proxy for empathic or clinical competence [52], and (3) the absence of external evaluative pressure that typically constrains human raters [50]. In psychiatric education, such unchecked leniency, while potentially reducing learner anxiety, risks misrepresenting clinical competence and may ultimately diminish the formative value of assessment. To strengthen reliability, future development should incorporate expert-in-the-loop calibration, reinforcement learning informed by clinician judgments, and iterative psychometric validation. Structured prompting strategies, stability constraints, and inconsistency penalties may further reduce variability. Since STACER is used nationally to evaluate readiness for independent practice, any AI-generated scoring must be tightly aligned with established competency standards. Even when used purely for formative purposes, calibration is critical to ensure that learners receive accurate guidance that meaningfully informs their development.

Alignment With Competency-Based and Experiential Learning

A key contribution of SAS is the way it operationalizes competency-based medical education [35,36] and the Kolb experiential learning cycle [31] within a unified, self-directed, digital ecosystem. Unlike traditional STACER assessments [29], which are episodic and summative, the SAS supports repeated low-stakes practice with immediate feedback, enabling residents to iteratively refine interviewing, diagnostic reasoning, and communication skills. The patient agent provides concrete experience as learners engage in full psychiatric interviews. The rater agent delivers reflective observation and abstract conceptualization through structured narrative feedback, cross-examination questions, and STACER-aligned performance indicators. Reengagement with the patient agent allows for active experimentation, completing the experiential cycle [31]. This sequence mirrors established mechanisms of learning in simulation-based education, where structured debriefing is consistently identified as the primary driver of skill acquisition [53,54]. Cheng et al [53] demonstrated that structured debriefing is an influential instructional component, facilitating reflection, performance analysis, and knowledge transfer to clinical contexts. For example, the Promoting Excellence and Reflective Learning in Simulation framework integrates self-assessment, guided discussion, and directive feedback [54]. The rater agent's structured questioning and targeted feedback map onto these principles, helping learners identify gaps and refine the interpretive steps that underlie psychiatric consultation competence.

By embedding simulation and assessment within one system, SAS presents a distinct pedagogical affordance: the ability to support adaptive, longitudinal skill development outside of resource-intensive, faculty-led simulation centers. This positions

XSL•FO
RenderX

the technology as a complement, not a replacement, to supervised clinical encounters and traditional Objective Structured Clinical Examinations–style assessments [55]. However, these findings should be interpreted cautiously, as the small sample size and open-label feasibility design primarily support conclusions regarding technical feasibility rather than educational effectiveness, thereby limiting generalizability to broader curricular integration.

Limitations and Future Directions for Educational Implementation

Several limitations must be acknowledged when interpreting these findings. First, phase 1 relied on automated testing using the LLM-as-a-judge paradigm [56,57], enabling high-volume evaluation but limiting the ecological validity of results. Automated benchmarking is useful for characterizing agentic behavior but cannot substitute for human educational outcomes. Second, although psychiatry residents were included, the sample size was small, and variability in interface use (text vs voice) introduced heterogeneity in the user experience, limiting generalizability and precluding conclusions about educational efficacy. Future studies should use randomized controlled designs with larger samples focused on psychiatry residents to evaluate competence development, diagnostic accuracy, and changes in consultation confidence. Third, and most notably, the limited capacity for rapport formation and nonverbal communication directly impacts multiple STACER domains, including affect, engagement, and relational skills. These findings reinforce that current text- and voice-based agentic systems cannot fully replicate the interpersonal complexity of psychiatric encounters. Future development should prioritize multimodal enhancements to better capture both patient and learner nonverbal behaviors and improve ecological validity.

Fourth, the current evaluation of SAS focused mainly on major depressive disorder. A broader evaluation of *DSM-5-Text Revision* diagnostic categories is needed for the system to serve as a comprehensive training tool. Finally, refinement of the rater agent, including calibration to professional standards and psychometric validation, is necessary for improving reliability, fairness, and consistency across learners. Integrating expert-driven feedback loops and controlled prompting strategies may reduce variability in scoring behavior and align the rater agent more closely with human evaluators. To strengthen its educational impact, future iterations of the SAS should also include learning-progress tracking and adaptive difficulty scaling to personalize training trajectories and support longitudinal skill development. Collectively, these limitations inform a clear roadmap for future development: expanding diagnostic breadth, enhancing multimodal realism, increasing participant diversity, and strengthening assessment reliability.

Conclusion

This technical feasibility prospective cohort study demonstrates that agentic AI systems can simulate clinically coherent psychiatric encounters and generate structured, STACER-aligned formative feedback within a single educational platform. Unlike earlier virtual patient systems that depended on scripted interactions or lacked psychiatry-specific competency assessment, the SAS combines adaptive multiturn psychiatric dialogue with structured evaluation aligned to competency-based medical education frameworks. The findings suggest that the SAS may provide a scalable and accessible approach for supporting deliberate practice, particularly for cognitive aspects of consultation such as history taking and diagnostic reasoning. At the same time, the study highlights an important limitation of current agentic AI systems: reduced capacity to support affective and relational skills, including rapport building, which rely on multimodal and interpersonal cues. By expanding access to repeated low-risk practice and structured feedback, the SAS has the potential to complement traditional supervision and simulation-based training, particularly in resource-constrained educational settings. Although preliminary findings from psychiatry residents are encouraging, the small sample and feasibility design preclude conclusions about educational effectiveness or curricular integration. Future studies should include larger controlled evaluations, multimodal enhancements, and further calibration of the rater agent to strengthen reliability and educational impact.

Acknowledgments

We thank the clinical team (Huda F. Al-Shamali, Karisa Parkington, Vanessa Tassone, Fatemeh Gholamali Nezhad, Curtis Fung) and the psychiatry residents (Andrew Volk, Prav Grewal, Gilmar Sebastian Gutierrez Nunez, Sukmin Yang, Eileen Tu, Nicholas Lee, Andriy Simko, Kalutota Samarasinghe, Reham Shalaby) for testing the SAS and Reinhard Janssen-Aguilar and Jithin Joseph for grading the Rater Agent. We also extend our gratitude to the following colleagues for reviewing the manuscript: Abdelrahman Sameh Abdou, Mahdi Mahdavi, Reza Barzegar, Yuqi Wu, Ade Sapara, Lindy VanRiper, Abhinav Arun Pillai, Kalutota Samarasinghe, Kimberly Dary, Andriy Simko, and the Marola Family. The authors declare the use of generative AI in the research and writing process. According to the Generative AI Delegation Taxonomy (2025), the following tasks were delegated to generative AI (GenAI) tools under full human

supervision: proofreading and editing. The GenAI tool used was ChatGPT-5.2. Responsibility for the final manuscript lies entirely with the authors. GenAI tools are not listed as authors and do not bear responsibility for the final outcomes.

Funding

The authors declared no financial support was received for this work.

XSL•FO
RenderX

Data Availability

The prompts for the patent agent and the rater agent, along with the synthetic dataset and patient profiles extracted from *DSM-5-TR* (*Diagnostic and Statistical Manual of Mental Disorders* [Fifth Edition, Text Revision]), will be provided upon request.

Authors' Contributions

AR was involved in conceptualization, research and development, performance testing and evaluation, writing and editing the manuscript. HFA was involved in the conceptualization, research, testing, writing, and editing of the manuscript. ZC, NR, and AR developed the SAS framework, conducted the testing, performed the graphic design, and wrote and edited the manuscript. RJA was involved in the conceptualization, research, testing, writing, and editing of the manuscript. JJ was involved in research, testing, writing, and editing of the manuscript. BGT was involved in the conceptualization, research, writing, and editing of the manuscript. MAK, LB, OW, BK, AT, DS, SK, GD, AG, AH, SS, YZ, and AD reviewed and edited the manuscript and provided feedback on the SAS. VB was involved in the conceptualization, review, and editing of the manuscript.

Conflicts of Interest

HFA, ZC, NR, RJA, JJ, MAK, BK, AT, DS, SK, GD, AG, AH, SS, and AD do not have any conflicts to declare. BGT and AR are supported by a CIHR Post-doctoral Fellowship (2025 - 2028). LB and OW are supported by the Alberta Academic Medicine and Health Services Program (AMHSP). Their work has also been supported by the Canadian Institutes of Health Research (CIHR), Alberta Mental Health Foundation, Glenrose Rehabilitation Hospital Foundation, Royal Canadian Legion, and grants from the Government of Alberta and Canada, including the Department of National Defence. YZ is supported by the University of Alberta's startup fund and has received research funding from the Canadian Institutes of Health Research, the Government of Alberta, Mitacs, and the New Frontiers in Research Fund. VB is supported by an Academic Scholar Award from the University of Toronto Department of Psychiatry and has received research funding from the Canadian Institutes of Health Research, Brain & Behavior Foundation, Ontario Ministry of Health Innovation Funds, Royal College of Physicians and Surgeons of Canada, Department of National Defence (Government of Canada), New Frontiers in Research Fund, Associated Medical Services Inc. Healthcare, American Foundation for Suicide Prevention, Roche Canada, Novartis, and Eisai.

Multimedia Appendix 1
Prompts.
[ZIP File, 13 KB - mededu_v12i1e88580_app1.zip]

Multimedia Appendix 2
Detailed description of the STACER Agentic System.
[DOCX File, 3929 KB - mededu_v12i1e88580_app2.docx]

Multimedia Appendix 3
Case Study Examples.
[DOCX File, 301 KB - mededu_v12i1e88580_app3.docx]

Multimedia Appendix 4
Performance Measure.
[DOCX File, 23 KB - mededu_v12i1e88580_app4.docx]

Checklist 1
Development, Evaluation, and Assessment of Large Language Models checklist.
[PDF File, 521 KB - mededu_v12i1e88580_app5.pdf]

References

1. Novais F, Ganança L, Barbosa M, Telles-Correia D. Communication skills in psychiatry for undergraduate students: a scoping review. Front Psychiatry 2022;13:972703. [doi: 10.3389/fpsyt.2022.972703] [Medline: 36032255]
2. Michaelson S, Rahim S. Communication skills training in psychiatry. BJPsych Adv 2023 Jan;29(1):56-67. [doi: 10.1192/bja.2021.80]
3. Dawood E, Alshutwi SS, Alshareif S, Shereda HA. Evaluation of the effectiveness of standardized patient simulation as a teaching method in psychiatric and mental health nursing. Nurs Rep 2024 Jun 4;14(2):1424-1438. [doi: 10.3390/nursrep14020107] [Medline: 38921717]

XSL•FO
RenderX

4. Song YO, Kim H, Nam Y, Choe K, Ha J. Effects of a competency-based education program for inpatient psychiatric nurses: a pre-post intervention study. J Korean Acad Psychiatr Ment Health Nurs 2022 Mar;31(1):80-87. [doi: 10.12934/jkpmhn.2022.31.1.80]
5. Savander E, Hintikka J, Wuolio M, Peräkylä A. The patients' practises disclosing subjective experiences in the psychiatric intake interview. Front Psychiatry 2021;12:605760. [doi: 10.3389/fpsyt.2021.605760] [Medline: 34040547]
6. Arbuckle MR, Stern DA, Barkil-Oteo A, Asghar-Ali AA. Training residents in high-value, cost-effective care: a National Survey of Psychiatry Program Directors. Acad Psychiatry 2020 Jun;44(3):324-329. [doi: 10.1007/s40596-020-01218-9] [Medline: 32212096]
7. Watling C, Driessen E, van der Vleuten CPM, Lingard L. Learning from clinical work: the roles of learning cues and credibility judgements. Med Educ 2012 Feb;46(2):192-200. [doi: 10.1111/j.1365-2923.2011.04126.x] [Medline: 22239333]
8. Nagendrappa S, de Filippis R, Ramalho R, et al. Challenges and opportunities of psychiatric training during COVID-19: early career psychiatrists' perspective across the world. Acad Psychiatry 2021 Oct;45(5):656-657. [doi: 10.1007/s40596-021-01482-3] [Medline: 34032993]
9. Feigerlova E. Training framework for high‐stakes OSCE: experience from volunteer standardized patients' bank. Clin Teach 2024 Dec;21(6):e13787. [doi: 10.1111/tct.13787] [Medline: 38808364]
10. Lipnevich AA, Mattern K, Feddock C. Formative assessment and feedback in medical education: a practical guide: AMEE Guide No. 189. Med Teach 2026 Jun;48(6):921-940. [doi: 10.1080/0142159X.2025.2569623] [Medline: 41134874]
11. Adams NE. Bloom's taxonomy of cognitive learning objectives. J Med Libr Assoc 2015 Jul;103(3):152-153. [doi: 10.3163/1536-5050.103.3.010] [Medline: 26213509]
12. Prégent J, Chung VHA, El Adib I, Désilets M, Hudon A. Applications of artificial intelligence in psychiatry and psychology education: scoping review. JMIR Med Educ 2025 Jul 28;11:e75238. [doi: 10.2196/75238] [Medline: 40720804]
13. Costa-Dookhan KA, Adirim Z, Maslej M, et al. Applications of artificial intelligence for nonpsychomotor skills training in health professions education: a scoping review. Acad Med 2025 May 1;100(5):635-644. [doi: 10.1097/ACM.0000000000005983] [Medline: 39874445]
14. Gorrindo T, Baer L, Sanders KM, et al. Web-based simulation in psychiatry residency training: a pilot study. Acad Psychiatry 2011;35(4):232-237. [doi: 10.1176/appi.ap.35.4.232] [Medline: 21804041]
15. Kononowicz AA, Woodham LA, Edelbring S, et al. Virtual patient simulations in health professions education: systematic review and meta-analysis by the digital health education collaboration. J Med Internet Res 2019 Jul 2;21(7):e14676. [doi: 10.2196/14676] [Medline: 31267981]
16. Mohd Kassim MA, Azli Shah SMY, Lim JTY, Mohd Daud TI. Online-based and technology-assisted psychiatric education for trainees: scoping review. JMIR Med Educ 2025 Apr 15;11:e64773. [doi: 10.2196/64773] [Medline: 40233361]
17. Kenny P, Parsons TD, Pataki C, et al. Virtual Justina: a PTSD virtual patient for clinical classroom training. Annu Rev Cyberther Telemed 2008:113-118 [FREE Full text]
18. Lee S, Kang J, Kim H, Chung KM, Lee D, Yeo J. COCOA: CBT-based conversational counseling agent using memory specialized in cognitive distortions and dynamic prompt. arXiv. Preprint posted online on Feb 27, 2024 URL: http://arxiv.org/abs/2402.17546 [accessed 2026-07-06] [doi: 10.48550/arXiv.2402.17546]
19. Vaswani A, Brain G, Shazeer N, et al. Attention is all you need. : Adv Neural Inf Process Syst; 2017.
20. Lee QY, Chen M, Ong CW, Ho CSH. The role of generative artificial intelligence in psychiatric education– a scoping review. BMC Med Educ 2025 Mar 25;25(1):438. [doi: 10.1186/s12909-025-07026-9]
21. Kamalov F, Calonge DS, Smail L, et al. Evolution of AI in education: agentic workflows. arXiv. Preprint posted online on Apr 25, 2025 URL: http://arxiv.org/abs/2504.20082 [accessed 2026-07-06]
22. Wang T, Zhan Y, Lian J, et al. LLM-powered multi-agent framework for goal-oriented learning in intelligent tutoring system. 2025 May 8 Presented at: WWW '25: Companion Proceedings of the ACM on Web Conference 2025; Apr 28 to May 2, 2025. [doi: 10.1145/3701716.3715244]
23. Cook DA, Overgaard J, Pankratz VS, Del Fiol G, Aakre CA. Virtual patients using large language models: scalable, contextualized simulation of clinician-patient dialogue with feedback. J Med Internet Res 2025 Apr 4;27:e68486. [doi: 10.2196/68486] [Medline: 39854611]
24. Cross J, Kayalackakom T, Robinson RE, et al. Assessing ChatGPT's capability as a new age standardized patient: qualitative study. JMIR Med Educ 2025 May 20;11:e63353. [doi: 10.2196/63353] [Medline: 40393017]
25. Yamamoto A, Koda M, Ogawa H, et al. Enhancing medical interview skills through AI-simulated patient interactions: nonrandomized controlled trial. JMIR Med Educ 2024 Sep 23;10:e58753. [doi: 10.2196/58753] [Medline: 39312284]
26. Holderried F, Stegemann-Philipps C, Herrmann-Werner A, et al. A language model–powered simulated patient with automated feedback for history taking: prospective study. JMIR Med Educ 2024 Aug 16;10:e59213. [doi: 10.2196/59213] [Medline: 39150749]
27. Jaleel A, Aziz U, Farid G, et al. Evaluating the potential and accuracy of ChatGPT-3.5 and 4.0 in medical licensing and in-training examinations: systematic review and meta-analysis. JMIR Med Educ 2025 Sep 19;11:e68070. [doi: 10.2196/68070] [Medline: 40973108]

XSL•FO
RenderX

28. Hicke Y, Geathers J, Rajashekar N, et al. MedSimAI: simulation and formative feedback generation to enhance deliberate practice in medical education. arXiv. Preprint posted online on Mar 1, 2025 URL: http://arxiv.org/abs/2503.05793 [accessed 2026-07-06]
29. Psychiatry clinical evaluation - core of psychiatry. URL: https://psychiatry.utoronto.ca/sites/default/files/assets/files/core-stacer-assessment-formupdateddecember2022-v2_0.pdf [accessed 2025-04-04]
30. Diagnostic and Statistical Manual of Mental Disorders: DSM-5: American Psychiatric Association; 2013. [doi: 10.1176/appi.books.9780890425596]
31. Abdool PS, Nirula L, Bonato S, Rajji TK, Silver IL. Simulation in undergraduate psychiatry: exploring the depth of learner engagement. Acad Psychiatry 2017 Apr;41(2):251-261. [doi: 10.1007/s40596-016-0633-9] [Medline: 27882523]
32. Rueda A, Hassan MS, Perivolaris A, et al. Understanding LLM scientific reasoning through promptings and model's explanation on the answers. arXiv. Preprint posted online on Jul 25, 2025 URL: http://arxiv.org/abs/2505.01482 [accessed 2026-07-06]
33. Pang HYM, Meshkat S, Teferra BG, et al. Opportunities and barriers of generative artificial intelligence in the training of psychiatrists: a competencies-based perspective. Acad Psychiatry 2025 Feb;49(1):25-30. [doi: 10.1007/s40596-024-02087-2] [Medline: 39587003]
34. Piot MA, Attoe C, Billon G, Cross S, Rethans JJ, Falissard B. Simulation training in psychiatry for medical education: a review. Front Psychiatry 2021;12:658967. [doi: 10.3389/fpsyt.2021.658967] [Medline: 34093275]
35. Hawkins RE, Welcher CM, Holmboe ES, et al. Implementation of competency-based medical education: are we addressing the concerns and challenges? Med Educ 2015 Nov;49(11):1086-1102. [doi: 10.1111/medu.12831] [Medline: 26494062]
36. Touchie C, ten Cate O. The promise, perils, problems and progress of competency-based medical education. Med Educ 2016 Jan;50(1):93-100. [doi: 10.1111/medu.12839] [Medline: 26695469]
37. Wang J, Duan Z. Agent AI with langgraph: a modular framework for enhancing machine translation using large language models. arXiv. Preprint posted online on Dec 5, 2024 URL: http://arxiv.org/abs/2412.03801 [accessed 2026-06-19]
38. Sui P, Duede E, Wu S, So RJ. Confabulation: the surprising value of large language model hallucinations. arXiv. Preprint posted online on Jun 25, 2024 URL: http://arxiv.org/abs/2406.04175 [accessed 2026-07-06]
39. ElevenLabs. ElevenLabs API documentation. 2024. URL: https://docs.elevenlabs.io [accessed 2025-04-04]
40. Tripathi S, Alkhulaifat D, Doo FX, et al. Development, evaluation, and assessment of large language models (DEAL) checklist: a technical report. NEJM AI 2025 May 22;2(6). [doi: 10.1056/AIp2401106]
41. Bedi S, Cui H, Fuentes M, et al. Holistic evaluation of large language models for medical tasks with MedHELM. Nat Med 2026 Mar;32(3):943-951. [doi: 10.1038/s41591-025-04151-2]
42. Harrigian K, Aguirre C, Dredze M. On the state of social media data for mental health research. arXiv. Preprint posted online on Apr 25, 2021 URL: http://arxiv.org/abs/2011.05233 [accessed 2026-07-06]
43. Schmidgall S, Ziaei R, Harris C, Reis E, Jopling J, Moor M. AgentClinic: a multimodal agent benchmark to evaluate AI in simulated clinical environments. arXiv. Preprint posted online on May 25, 2025 URL: http://arxiv.org/abs/2405.07960 [accessed 2026-07-06]
44. American Psychiatric Association. Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition, Text Revision (DSM-5-TR): American Psychiatric Association Publishing; 2022.
45. Koo TK, Li MY. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. J Chiropr Med 2016 Jun;15(2):155-163. [doi: 10.1016/j.jcm.2016.02.012] [Medline: 27330520]
46. Provoost S, Lau HM, Ruwaard J, Riper H. Embodied conversational agents in clinical psychology: a scoping review. J Med Internet Res 2017 May 9;19(5):e151. [doi: 10.2196/jmir.6553] [Medline: 28487267]
47. Tay JL, Xie H, Sim K. Effectiveness of augmented and virtual reality-based interventions in improving knowledge, attitudes, empathy and stigma regarding people with mental illnesses-a scoping review. J Pers Med 2023 Jan 4;13(1):112. [doi: 10.3390/jpm13010112] [Medline: 36675773]
48. Dupuy L, Micoulaud-Franchi JA, Cassoudesalle H, et al. Evaluation of a virtual agent to train medical students conducting psychiatric interviews for diagnosing major depressive disorders. J Affect Disord 2020 Feb 15;263:1-8. [doi: 10.1016/j.jad.2019.11.117] [Medline: 31818765]
49. Adel L, Moses L, Irvine E, Greenway KT, Dumas G, Lifshitz M. A systematic review of hyperscanning in clinical encounters. Neurosci Biobehav Rev 2025 Sep;176:106248. [doi: 10.1016/j.neubiorev.2025.106248] [Medline: 40484262]
50. Harari MB, Rudolph CW. The effect of rater accountability on performance ratings: a meta-analytic review. Hum Resour Manag Rev 2017 Mar;27(1):121-133. [doi: 10.1016/j.hrmr.2016.09.007]
51. Jain S, Ahmed UZ, Sahai S, Leong B. Beyond consensus: mitigating the agreeableness bias in LLM judge evaluations. arXiv. Preprint posted online on Oct 13, 2025 URL: http://arxiv.org/abs/2510.11822 [accessed 2026-06-19]
52. Alikaniotis D, Suhara Y, Stureborg R. Large language models are inconsistent and biased evaluators. Arxiv. Preprint posted online on May 2, 2024 URL: http://arxiv.org/abs/2405.01724 [accessed 2026-06-19]
53. Cheng A, Eppich W, Grant V, Sherbino J, Zendejas B, Cook DA. Debriefing for technology-enhanced simulation: a systematic review and meta-analysis. Med Educ 2014 Jul;48(7):657-666. [doi: 10.1111/medu.12432] [Medline: 24909527]

XSL•FO
RenderX

54. Eppich W, Cheng A. Promoting excellence and reflective learning in simulation (PEARLS): development and rationale for a blended approach to health care simulation debriefing. Simul Healthc 2015 Apr;10(2):106-115. [doi: 10.1097/SIH.0000000000000072] [Medline: 25710312]
55. Zayyan M. Objective structured clinical examination: the assessment of choice. Oman Med J 2011 Jul;26(4):219-222. [doi: 10.5001/omj.2011.55] [Medline: 22043423]
56. Chiang WL, Gonzalez J, Li D, et al. Judging LLM-as-a-judge with MT-bench and chatbot arena. Presented at: Advances in Neural Information Processing Systems 36; Dec 10-16, 2023. [doi: 10.52202/075280-2020]
57. Li D, Jiang B, Huang L, et al. From generation to judgment: opportunities and challenges of LLM-as-a-judge. 2025 Presented at: Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing; Nov 4-9, 2025; Suzhou, China p. 2757-2791. [doi: 10.18653/v1/2025.emnlp-main.138]

Abbreviations

CBME: competency-based medical education
***DSM*:** *Diagnostic and Statistical Manual of Mental Disorders*
***DSM-5*:** *Diagnostic and Statistical Manual of Mental Disorders* (Fifth Edition)
ICC: interclass correlation coefficient
LLM: large language model
MSE: mental status examination
SAS: STACER Agentic System
STACER: structured assessment of clinical encounter report
UCSF: University of California, San Francisco

Edited by S Brini; submitted 19.Dec.2025; peer-reviewed by D Winterton, R Sharma; revised version received 08.Jun.2026; accepted 09.Jun.2026; published 21.Jul.2026.

Please cite as:
Rueda A, Al-Shamali HF, Cote Z, Roy N, Janssen-Aguilar R, Joseph J, Teferra BG, Kamaleddin MA, Burback L, Winkler O, Kapralos B, Torres A, Sharma D, Krishnan S, Dumas G, Greenshaw A, Hudon A, Sockalingam S, Zhang Y, Dubrowski A, Bhat V
Enhancing Psychiatry Training Using an Agentic AI Simulated Consultation Tool: Prospective Cohort Study
JMIR Med Educ 2026;12:e88580
URL: https://mededu.jmir.org/2026/1/e88580
doi:10.2196/88580

© Alice Rueda, Huda F Al-Shamali, Zack Cote, Niloy Roy, Reinhard Janssen-Aguilar, Jithin Joseph, Bazen Gashaw Teferra, Mohammad Amin Kamaleddin, Lisa Burback, Olga Winkler, Bill Kapralos, Andrei Torres, Divya Sharma, Sridhar Krishnan, Guillaume Dumas, Andrew Greenshaw, Alexandre Hudon, Sanjeev Sockalingam, Yanbo Zhang, Adam Dubrowski, Venkat Bhat. Originally published in JMIR Medical Education (https://mededu.jmir.org), 21.Jul.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Education, is properly cited. The complete bibliographic information, a link to the original publication on https://mededu.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Performance of GPT-4o and Claude in Medical Ethics Scenarios: Comparative Study

Karishma R Desai, BS; Anna L Gorsky, MD; Nicole A Zelenski, MD

Division of Hand Surgery, Department of Orthopaedic Surgery, Emory University, 80 Jesse Hill Jr Drive SE, Atlanta, GA, United States

Corresponding Author:
Nicole A Zelenski, MD
Division of Hand Surgery, Department of Orthopaedic Surgery, Emory University, 80 Jesse Hill Jr Drive SE, Atlanta, GA, United States

Related Article:

This is a corrected version. See correction statement: https://mededu.jmir.org/2026/1/e113933

Abstract

Background: The emergence of AI technology has sparked curiosity regarding the capabilities of large language models (LLMs) in the field of medicine. Minimal research exists regarding the proficiency of various AI models in ethics scenarios, specifically in specialty-based scenarios.

Objective: This study aimed to compare the performance of GPT-4o and Claude Sonnet 4 on ethics questions with that of medical students and orthopedic residents.

Methods: A total of 200 ethical or legal scenario questions were randomly selected from question banks targeted for third- and fourth-year medical students (UWorld, AMBOSS) and orthopedic residents (OrthoBullets). Questions at the medical student level were exclusively text-based, while resident-level questions included text-based questions accompanied by images. Each question was entered identically into each AI model 3 separate times. If answers varied between trials, the answer provided most frequently by the model was used as the selected answer.

Results: GPT-4o correctly answered 140 (70%) of 200 questions, which was similar to the average human test taker score of 71% (~142/200 questions). Claude correctly answered 180 (89%) questions, a score greater than that of human test takers and significantly better than GPT-4o ($P<.001$). Claude scored significantly higher than GPT-4o in almost all question categories. GPT-4o provided different responses to identically worded trials for 27 (21%) of 130 general questions and 3 (4%) of 70 orthopedic questions ($P=.002$), while Claude did not have a significant difference in variability between these 2 groups (general: 16/130, 12% vs orthopedic: 3/70, 4%; $P=.06$). Neither GPT-4o or Claude showed a significant difference in the rate of most common incorrect response selection between question categories. Claude showed no significant difference in the rate of most common incorrect response selection between question categories.

Conclusions: These results suggest that GPT-4o can potentially answer both general and specialty-specific ethical questions with similar proficiency to sample groups of both medical students and orthopedic residents, while Claude AI performs significantly better than both humans and GPT-4o. Variables such as AI model framework and training data may drive the observed difference in performance, but the exact cause cannot be definitively isolated without intentional testing. Therefore, further research is needed to ensure safety by minimizing output variability before integrating AI as a patient-facing resource.

(JMIR Med Educ 2026;12:e70199) doi:10.2196/70199

KEYWORDS

medical education; ethics; legal; ChatGPT; artificial intelligence; medical student; orthopedics; Claude; resident; health

Introduction

Interpersonal skills are an invaluable necessity for health care providers. Impactful medical practice requires the ability to adapt treatment plans to a patient's unique preferences, which is founded on the development of the patient-provider relationship. Numerous studies have shown that effective physician communication is positively correlated with patient satisfaction [1] and positive health outcomes [2]. Patient involvement in their own medical care, often termed "shared decision-making," has also been observed to correlate with increased overall and disease-specific quality of life scores in orthopedic patients [3]. The now near-ubiquity of internet usage has resulted in patients frequently consulting online materials for additional medical information. Nearly all American

XSL•FO
RenderX

households had internet access in 2021 [4], and recent Centers for Disease Control and Prevention data suggest that more than half of adults (58.5%) use the internet for medical advice [5]. This trend is also mirrored internationally, with 1 study showing that 64.3% of Lebanese patients consulted web resources for acute symptoms, with 19.2% seeking web information before physician consultation [6].

The increasing popularity of AI technologies has introduced a new potential source for internet-accessible medical information. Large language models (LLMs), such as GPT-4o (OpenAI), are currently being evaluated for their ability to play a role in health care. AI has been shown to accurately process medical knowledge and show proficiency at the level of third- and fourth-year medical students in preparation for the United States Medical Licensing Examination [7-9]. In addition, GPT-4o has been challenged with postgraduate examination materials, including preparation materials for the American Board of Surgery In-Training Examination [10,11], neurology specialist examinations [12], and the Orthopaedic In-Training Examination [13,14] and, in most cases, has performed at the level of a human medical trainee. Recent exploration has shown that ChatGPT models can also directly respond to patient questions about common conditions across a variety of specialties with generally satisfactory answers [15-17].

Nevertheless, integration of AI into health care settings remains controversial due to a lack of clarity surrounding its strengths and limitations, and it is unclear whether AI can be safely recommended to serve as a substitute or even as an adjunct to physician consultation [18]. The "soft skills" of medical training include compassion and empathetic communication, which are essential to identifying and navigating the nuances present within each individual patient interaction. Although AI technologies have shown some capability in correctly reproducing technical knowledge and answering patient questions, their capability to identify and process user emotion is not well understood.

LLMs are modeled on neural networks, in which units or "tokens" of data are connected and linked by associations, which are strengthened or weakened based on the presence of patterns within training data [19]. Essentially, the models use the patterns identified from their training data to generate outputs that are the most probabilistically likely "correct" responses [20]. Transformer architecture is the modern evolution of neural networks and uses "self-attention" to consider and weight tokens in relation to each other [21]. Through the training process, the model adapts the level of attention given to various tokens to identify the most relevant components of the input and better generate more relevant responses in the given context [21]. Ultimately, transformer architecture improves the statistical predictive accuracy of LLMs, especially in the processing of large volumes of inputs, by more accurately weighing the relevance of input tokens [21].

Training can occur through the establishment of parameters, which guide the model in processing data inputs to promote or oppose specific outputs; positive reinforcement of a specific association teaches the model that the reinforced logic has an increased likelihood of leading to the correct output [22]. Feedback can be provided to the model based on the final output (outcome supervision) or throughout the logical process (process supervision), and the model applies this feedback and recalculated probabilities to direct the processing of new data [23]. Therefore, the veracity of AI-generated outputs is dependent directly on the quality and accuracy of the training data and subtleties within the inputs themselves, rather than solely on known truths [24]. As a result, AI models have been documented to "hallucinate" or fabricate information and present it convincingly as fact [25]. Researchers in one study found that a number of inconsistencies and hallucinations were noted when questions on sensitive or obscure topics were presented to several AI models [26], supporting the hypothesis that AI performance declines when available data are limited. "Factuality" hallucinations, when models report information that violates known facts, and "faithfulness" hallucinations, when models deviate from the information provided in the given input, are theorized to occur due to inconsistent training and model evaluation metrics [27].

Despite numerous studies evaluating AI's accuracy in answering questions based on empiric knowledge, the literature supporting AI's performance in health care–based medical ethical, legal, and social scenarios is limited and inconsistent. Some reports demonstrate that LLMs can answer ethical scenario questions [28] and respond to patient clinical questions with answers that are rated positively for empathy by human raters [29,30], while others found ChatGPT models performed worse on ethics questions compared to medical trainees [10,31,32]. When subjected to emotion-related testing, AI demonstrated the ability to correctly identify and manage emotions but struggled when associating emotions to rationale and scored lower than the human average on the "using emotions to facilitate thought" section of the examination [33].

Variant AI models have been designed to address the general safety and ethical concerns surrounding the outputs produced by LLMs. One such model is Claude, created by Anthropic, which describes it as "helpful, honest, and harmless" [34]. Claude was developed using a novel "Constitutional AI" method, meaning that instead of using human feedback during the training process, feedback is instead provided by the AI model itself, guided by a provided set of "constitutional principles" [35]. During the supervised training process, the model is asked prompts that will purposefully generate a "harmful" output, for example, asking for instructions on how to break into a house. The model is then asked to critique the output based on a constitutional principle ("how is the last response unethical or illegal") and then to revise the output according to the critique ("rewrite the response to remove unethical or illegal content") [36]. In the reinforcement learning or unsupervised phase, the model is asked to generate multiple responses to a prompt, and a different feedback model is asked to choose one of the responses also based on a constitutional principle ("which of these responses is the most ethical and legal?") [36]. Therefore, humans are not providing feedback for each model response themselves, but instead instructing a "teaching" AI to guide the "learning" model to align with specific rules, or constitutional principles. Anthropic sources its constitutional principles from a number of ethic-centered

XSL•FO
RenderX

documents, including the UN Declaration of Human Rights, as well as principles developed by other AI laboratories and Anthropic researchers themselves during model training to discourage harm, aggression, offense, violence, and so on [35]. Although Claude is intentionally designed with ethics and safety in mind, there are no available studies specifically exploring Claude's performances in ethics-related health care scenarios.

Throughout medical school and postgraduate medical training, learners are taught how to manage complex social issues and interpersonal interactions with a high degree of emotional intelligence and empathy. Even ethical dilemmas, which are addressed using well-defined ethical principles, must be navigated within the larger context of patient emotion. Medical training examinations often use questions about end-of-life discussions, patient safety, communication strategies, and informed consent to assess these intangible skills. This study aimed to evaluate the ability of GPT-4o and Claude Sonnet 4 to answer health care–based ethical medical board–style questions accurately and consistently. We hypothesized that GPT-4o and Claude will differ in performance between question levels and categories.

Methods

All questions used for testing were scenario-based multiple-choice questions for which there was only one correct answer. Medical student–level questions were sourced from the UWorld Step 2 CK question bank and AMBOSS, 2 resources designed to prepare third- and fourth-year medical students for the Step 2 benchmark examination. The UWorld system allows for question sourcing from selected subject categories. One category of interest was the broad "Social Sciences (Ethics/Legal/Professional)" category, which involves the subcategories of Communication and Interpersonal Skills, Healthcare Policy and Economics, Medical Ethics and Jurisprudence, Patient Safety, and System-based practice and Quality Improvement. The subcategory "Communication and Interpersonal Skills" was also specifically selected as a second source of questions to assess whether performance differed on questions specifically relating to interpersonal communication scenarios, which often require interpretation of patient tone and social cues. Both types of questions typically begin with patient vignettes. Ethics or legal questions often ask, "Which of the following is the best next step in management of this patient?" in scenarios that require the physician to make a decision about common clinical dilemmas involving end-of-life care, next-of-kin decision-making power, or consent for emergent procedures. Interpersonal questions typically present a patient scenario regarding a sensitive subject such as alcohol dependence before asking, "which of the following is the most appropriate response to this patient?" The test taker is given several different responses as answer choices that vary in tact and counseling advice. Individual question specifics cannot be reproduced, as these question banks are copyright protected; however, these questions are designed to imitate the style of questions typically asked on the USMLE Step 2 CK examination. An example of this question's style provided by the USMLE is provided as follows [37]:

> *A 67-year-old man is evaluated in the intensive care unit. He has end-stage pancreatic cancer and was hospitalized 3 days ago for treatment of pneumonia. Respirations are 6/min. Pulse oximetry on 100% oxygen by face mask shows an oxygen saturation of 78%. Examination shows feeble respiratory efforts; he is using accessory muscles of respiration. On mental status examination, the patient is oriented to person but not to place or time. If the patient is not endotracheally intubated and mechanically ventilated, he will die within hours. His wife says the patient recently told her that he would never want mechanical ventilation, but they never completed paperwork regarding his wishes. His daughter insists that he be mechanically ventilated. Which of the following is the most appropriate action for the physician to take?*

As these ethical dilemmas are typically framed in the context of emotional situations, the complexity and nuance of addressing these issues often stem from the application of concrete ethical principles to emotionally charged and sensitive scenarios. Therefore, identification of the correct or "most appropriate response" requires not only an understanding of ethical principles but also the ability to apply these principles in a manner that is sensitive to the patient. The interpersonal communication questions were included to assess not only LLM performance in ethical scenarios but also to assess whether there was unique accuracy or inaccuracy in the skill of identifying the textbook "most appropriate answer."

The UWorld software allows users to create custom tests based on subject categories with up to 40 questions per test. The individual questions available in each subject category are not visible to test takers due to the nature of the test bank, and therefore, the questions included in each test could not be individually predetermined. A 40-question test was generated from each category described previously—one test using questions from the broader "Social Sciences (Ethics/Legal/Professional) category" and one test using questions exclusively from the "Communication and Interpersonal Skills" subcategory.

As the UWorld questions were not individually selected prior to test generation, 15 questions were duplicated between the 40-question sets. The duplicated questions were not excluded as the average test taker performance score is generated based on the full 40-question set, and therefore, inclusion of all questions was necessary to compare LLM performance to human performance. This method ensured that each test would be representative of tests encountered by human testers.

The AMBOSS question bank similarly allows the creation of tests involving randomly generated questions within a subcategory, although with a higher maximum of 50 questions. The AMBOSS "Legal Medicine and Ethics" category was the only equivalent available subcategory to test the ethics-related medical competencies, and therefore, a test of 50 randomly selected questions from the AMBOSS Legal Medicine and Ethics discipline was generated. The format of these questions mirrored those of UWorld questions in that they are designed

XSL•FO
RenderX

to imitate the style of questions asked on the USMLE Step 2 examination.

Another area of interest was whether model performance would vary in specialty-specific contexts. By testing questions written at a level for medical residents, LLM performance can be evaluated in scenarios at higher levels of medical complexity. This aim of the study was treated as completely exploratory, and causal relationships should not be inferred. Orthopedics was chosen as the specialty due to the author's access to Orthobullets, an online resource intended to prepare orthopedic residents for the Orthopaedic In-Training Examination. This question bank similarly allows the generation of subject-specific tests and provides 2 question categories relevant to this study "Ethics in Orthopaedics Practice" and "Legal Considerations in Orthopaedic Practice." A total of 70 questions were available: 21 from the Ethics category and 49 from the Legal category. All questions were included in this study. The ethics questions required test takers to identify the most ethical course of actions in orthopedic patient–specific scenarios involving potential financial conflicts of interest with research studies and transfers of value from pharmaceutical or medical device representatives, obtaining interpretation services for a non–English-speaking patient, and appropriate marketing practices. Legal consideration questions involved scenarios in which test takers were required to identify the most appropriate way to obtain consent in emergent situations or manage medical or surgical errors, altering the medical record of surgical errors, or preoperative protocol management, and professionalism violations.

All questions from UWorld and AMBOSS were text-based by nature of the test banks, while Orthobullets questions included text-only questions as well as questions with supporting images. The inclusion of questions involving images introduced a potential for confounding, but it was thought that analysis performed at the question bank–level would illuminate whether LLM performance varied by question bank. Only 7 questions of the 70 sourced from Orthobullets included images, and this volume was not considered sufficient to draw conclusions about varying performance based on question modality. For all question banks, the software provided a metric of the "average" score earned by test takers on the same test, and this is the benchmark used in this study as the "human test taker" performance metric. It is important to note that the multiple-choice questions used are not validated measures of ethical reasoning or empathetic reasoning but simply an available tool to compare performance to that of human test takers. Information on the specifics of how these benchmarks are calculated is not publicly available for any of the question banks. Therefore, questions could not be excluded in our study from the tests generated by the software, as the average human test taker benchmark could not be manually calculated.

A zero-shot approach was used to conduct the trials; no training or prompt engineering was performed to provide context to the models [38]. Instead, the free, publicly available version of GPT-4o was opened in a browser window, a singular question and its corresponding answer choices were copied and pasted verbatim as the input, and the resulting selection was recorded. If any images were provided in accompaniment to the question stem, these were also provided to the chatbot in the same input as the question stem. To prevent bias or learning for subsequent trials, a new, unique browser window was opened, and the above, zero-shot approach was repeated for each trial. If the provided answer varied between trials, the answer generated in the majority of trials was recorded as GPT-4o's selected response. Three trials were performed for each question, and additional trials were conducted as needed to achieve a majority response if there was no consensus between the original 3 trials. Typically, GPT-4o provided reasoning for its selection without additional prompting. If the incorrect answer was chosen, the reasoning behind the incorrect choice was recorded. If no reasoning was provided, GPT-4o was then prompted with "Why is answer choice [the correct answer] incorrect?" If answers varied between trials, GPT-4o's evaluation of the correct answer choice, incorrect answer choice, and previously selected answer choices were recorded. All questions tested required the selection of a singular answer choice. On the occasions when GPT-4o selected 2 answer choices as the correct answer, the browser was refreshed, and the trial was repeated until GPT-4o provided only one answer choice. The testing procedure can be visualized in Figure 1.

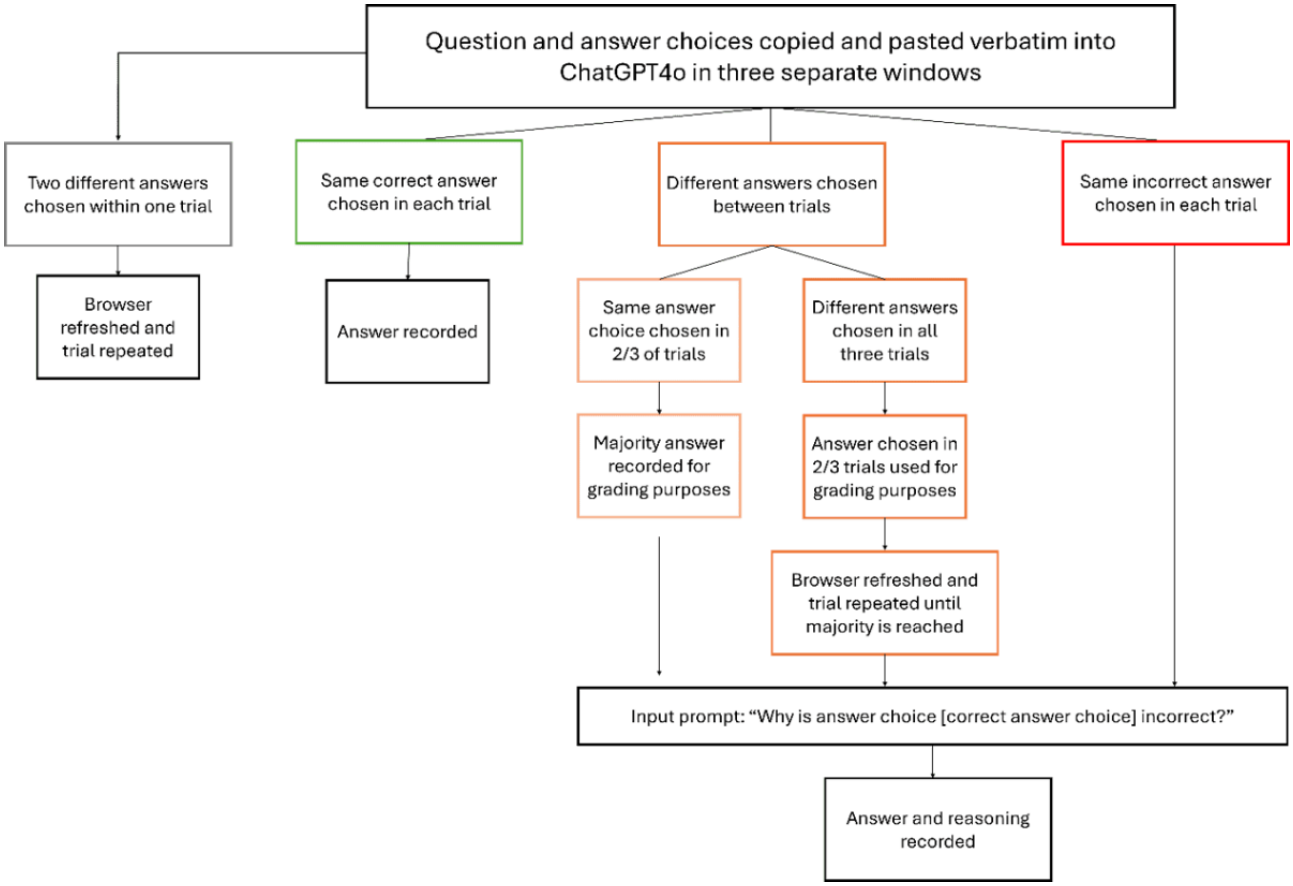
For all questions, the question banks also provided individual test taker benchmarks on the percentage of subscribers that selected each individual answer choice. It is important to note that the exact method of calculation for this metric was also not publicly available for each question bank. If GPT-4o selected the incorrect answer, it was noted whether or not its selection was the incorrect answer choice most popularly chosen by humans.

The same zero-shot testing method was used for Claude Sonnet 4. Claude users are required to pick a category of interest on sign-in. The "learning and studying" option was selected for all trials included in this study, but no additional prompt engineering or configurations were made to this chatbot.

Statistical analysis to compare the AI model's overall performance against human performance was performed using SPSS (version 31.0.0.0 (117); IBM) with one-sample proportion tests, and McNemar tests were used to compare performance between models. Chi-square and Fisher exact tests were used to analyze model ability to select the most popular incorrect response as well as the frequency of varying responses to identical questions.

XSL•FO
RenderX

Figure 1. Testing method, repeated for each question.



Results

It is unclear how the question banks derived the reported human performance averages and whether these scores are truly representative of the true population of medical test takers. Therefore, statistical inferences are not valid, and comparisons between LLM and human performance are reported purely descriptively. The performance of GPT-4o was similar to the average score of the question banks' human test takers (71% vs 70%). GPT-4o correctly answered 89 (68%) of 130 general medical student–level questions, and 51 (73%) of 70 orthopedic resident–level questions with no significant difference in performance, compared to the human averages of 71% and 73%, respectively. Claude answered 180 (90%) of 200 total questions correctly, performing better than the human average (89% vs 71%) and significantly better than GPT-4o (89% vs 70%; $P<.001$). Claude's score was higher than the provided human average on all question banks. Claude also performed significantly better than GPT-4o on all question categories except the UWorld Social Sciences and Orthobullets Ethics questions; results and statistical conclusions are summarized in Table 1, with the number of questions correct and correlated weighted score as provided by each test bank.

XSL•FO
RenderX

Table . Correctly answered by human test takers, GPT-4o, and Claude per question set[a].

Question set	Human test taker score, n/N (%)	GPT-4o score, n/N (%)	Claude score, n/N (%)	Between-model comparison, *P* value
UWorld All Social Sciences	~27/40 (68)	27/40 (68)	33/40 (83)	.14
UWorld Interpersonal-Specific	~28/40 (69)	30/40 (75)	36/40 (90)	.03[b]
AMBOSS	~36/50 (72)	32/50 (64)	46/50 (92)	<.001[b]
Medical student–level or nonspecific questions	~92/130 (71)	89/130 (68)	115/130 (88)	<.001[b]
Orthobullets Ethics	~15/21 (73)	14/21 (67)	17/21 (81)	.25
Orthobullets Legal	~36/49 (73)	37/49 (76)	48/49 (98)	<.001[b]
Resident-level/orthopedic questions	~51/70 (73)	51/70 (73)	65/70 (92)	<.001[b]
Overall performance	~142/200 (71)	140/200 (70)	180/200 (89)	<.001

[a]Aggregate question categories (medical student, orthopedic resident, and overall) performance were calculated using averages of human test taker percentage scores. “Between model comparison” represents the statistical significance of exclusive comparison between GPT-4o and Claude performance.

[b]Statistically significant results (P<.05).

Question banks only provided average human test taker score in percentages, and equivalent numbers of correctly answered questions were calculated based on the provided percentage correct and total number of questions used.

GPT-4o answered 60 (30%) of 200 total questions incorrectly. Of these 60 questions, the most popular incorrect response was chosen 37 (62%) times. Analysis did not show a significant difference in the frequency that GPT-4o chose the most popular incorrect answer between both UWorld general social sciences and interpersonal-specific sets (*P*=.07), UWorld (general and interpersonal) and AMBOSS (legal medicine and ethics) sets (*P*=.33), Orthobullets Ethics and Legal sets (*P*=.22), or medical student general and orthopedic resident–level (UWorld/AMBOSS and Orthobullets) sets (*P*=.46; Table 2).

Table . Incorrectly answered questions per question set for which GPT-4o and Claude selected the incorrect answer most commonly chosen by human test takers.

Question set	GPT-4o[a], n/N (%)	Claude[b], n/N (%)
UWorld All Social Sciences	6/13 (46)	5/7 (71)
UWorld Interpersonal-Specific	9/10 (90)	4/4 (100)
AMBOSS	9/18 (50)	3/4 (75)
Orthobullets Ethics	6/7 (86)	4/4 (100)
Orthobullets Legal	7/12 (58)	1/1 (100)

[a]Total: 37/60 (62%).

[b]Total: 17/20 (85%).

Claude incorrectly answered 20 (10%) of 200 total questions. Analysis using Fisher exact test did not show a significant difference in the frequency that Claude chose the most popular incorrect answer between both UWorld general social sciences and interpersonal-specific sets (*P*=.49), UWorld (general and interpersonal) and AMBOSS (legal medicine and ethics) sets (*P*>.99), Orthobullets Ethics and Legal sets (*P*>.99) or medical student general and orthopedic resident–level (UWorld/AMBOSS and Orthobullets) sets (*P*>.99; Table 2).

GPT-4o varied answers on 30 (15%) of 200 total questions, while Claude varied answers on 19 (10%) of the 200 total questions. There was no significant difference between the frequency of varying answers between GPT-4o responses for both the UWorld general social sciences and interpersonal-specific sets (*P*=.23) UWorld (general and interpersonal) and AMBOSS (legal medicine and ethics) sets (*P*=.25), and Orthobullets Ethics and Legal sets (*P*=.90). However, GPT-4o produced variable answers between trials significantly less frequently for orthopedic resident–level questions than for medical student general questions (*P*=.002; Table 3).

XSL•FO
RenderX

Table . Questions per question set for which AI models changed answers between trials.

Question set	GPT-4o[a], n/N (%)	Claude[b], n/N (%)
UWorld All Social Sciences	5/40 (13)	5/40 (13)
UWorld Interpersonal-Specific	9/40 (23)	6/40 (15)
AMBOSS	13/50 (26)	5/50 (10)
Orthobullets Ethics	1/21 (5)	1/21 (5)
Orthobullets Legal	2/49 (4)	2/49 (4)

[a]Total: 30/200 (15%).

[b]Total: 19/200 (10%).

Claude did not significantly differ responses for both UWorld general social sciences and interpersonal-specific sets (P=.75), UWorld (general and interpersonal) and AMBOSS (legal medicine and ethics) sets (P=.53), and Orthobullets Ethics and Legal sets (P=.90), or between medical student general questions and orthopedic resident–level questions (P=.06).

There was no significant difference in variability of responses between the 2 models overall (P=.09), for medical student general questions (P=.07) or for orthopedic resident–level questions (P>.99).

When both models varied their answers, the explanations between various sessions were often contradictory; one session of the model would provide reasoning as to why an answer choice was correct, and a different session of the model would cite the same reasoning as the reason that answer choice was incorrect. Examples of contradictory reasoning are provided in Tables 4 and 5.

XSL•FO
RenderX

Table . Examples of contradictory reasoning from GPT-4o.

Question set	Question scenario	Trial 1 reasoning	Trial 2 reasoning
UWorld All Social Sciences	Next best step in management of a pediatric patient refusing chemotherapy against the wishes of her family	"The best next step: Before any decisions are made regarding treatment, it is crucial to assess whether the child has the cognitive and emotional maturity to make an informed decision. In cases of disagreement between family members, the healthcare team, and the patient, an ethics committee meeting may be helpful. However, this would be premature until the child's capacity is assessed. It is better to first evaluate whether the child can make her own decision before escalating the situation to an ethics committee."	"An ethics committee can help mediate the disagreement between the patient, the parents, and the medical team. It allows for a structured discussion of the ethical principles involved, including the child's autonomy, the parental authority, the treatment's potential benefits and burdens, and the patient's best interests. An ethics consultation provides a thorough review and a neutral perspective on how to balance these competing factors. [Requesting evaluation of patient for decision-making capacity] seems reasonable, but it's crucial to note that at 8 years old, a child is unlikely to have the legal capacity to make independent medical decisions."
UWorld Interpersonal-Specific	Next best step in management of a patient presenting after office hours for a nonemergent concern	"Healthcare providers must set reasonable boundaries around after-hours care. While the patient may be dissatisfied with the timing, the clinic's normal operational hours should be respected unless there is an urgent need (such as signs of severe infection, anaphylaxis, etc), which does not seem to be the case here. ...The patient has a non-urgent concern (itchy rash) that was already evaluated a week ago, and it doesn't appear to be an emergency... A rash that was already examined a week ago and treated for dry skin is unlikely to require immediate intervention."	"While it is important to adhere to office hours, patient safety and care should always take precedence. The physician should evaluate the patient to assess the severity of her symptoms and provide appropriate care, especially since rashes and skin conditions can sometimes escalate quickly."
AMBOSS	Most appropriate action following parent refusal of chemotherapy and decision to pursue supportive care for pediatric patient	"The parents' decision should be respected after they have been fully informed about the potential consequences. If they choose to forgo chemotherapy, supportive care should be provided, ensuring the child's comfort."	"While it is important to respect parental wishes, the physician has an ethical responsibility to ensure that the child receives appropriate treatment. In this case, supportive care alone is not appropriate for intermediate-risk neuroblastoma, as it could lead to significant harm or death."
Orthobullets Ethics	Analysis of orthopedic patient complaints	"Lack of empathy can certainly affect a patient's overall experience, but it's generally not the top complaint compared to logistical frustrations, such as long waits for visits or difficulties with prescription management."	"In summary, long waits for visits are an issue but not the most frequent complaint that patients tend to focus on when evaluating their orthopaedic care. The lack of empathy is more often identified as a key source of dissatisfaction in this specialty."
Orthobullets Legal	Analysis of patient-physician interaction	"Research suggests that, on average, a physician will allow the patient to speak for about 60 seconds (1 minute) after asking an open-ended question, before they typically redirect or intervene."	"On average, patients speak for about 23 seconds before the physician redirects the conversation or takes control of the dialogue, often in order to gather more focused or specific information."

XSL•FO
RenderX

Table . Examples of contradictory reasoning from Claude.

Question set	Question scenario	Trial 1 reasoning	Trial 2 reasoning
UWorld All Social Sciences	Next best step in management of an unconscious hypotensive patient receiving blood transfusion. His next-of-kin requests the transfusion be stopped due to religious purposes; no advance directive available	"The answer is: [Continuing transfusion until patient is stable] The patient is in immediate life-threatening condition Treatment was initiated under emergency doctrine Without written documentation, the physician cannot be certain of the patient's specific wishes The [next-of-kin's] statement, while important, is not sufficient legal documentation in this emergency Once stabilized, there would be time to obtain proper documentation or ethics consultation This approach prioritizes the irreversible nature of death while still respecting the process of determining patient wishes This represents the standard approach in emergency medicine: when in doubt and life is immediately threatened, preserve life while working to clarify the patient's true wishes through proper channels."	"Option [Continuing transfusion until patient is stable] fundamentally violates the principle of patient autonomy and informed consent for several reasons: Autonomy doesn't pause for medical convenience: A patient's right to refuse treatment doesn't become suspended just because they're unstable. If we accepted this logic, we could override any patient's wishes whenever their condition was serious enough. Temporal problem: "Until stabilized" creates an arbitrary endpoint that prioritizes medical goals over the patient's values. Stabilization could take hours, days, or longer - during which time the patient's clearly expressed wishes (through his wife) are being violated. ... Slippery slope: If physicians could override known patient wishes whenever they deemed the patient "too unstable," it would essentially nullify the right to refuse treatment in emergency situations. The key ethical point: Patient autonomy and the right to refuse treatment based on religious beliefs doesn't diminish based on the severity of the medical situation. Once the patient's wishes are reliably known through an appropriate surrogate, those wishes should be respected immediately, not delayed for medical convenience."

XSL•FO
RenderX

Question set	Question scenario	Trial 1 reasoning	Trial 2 reasoning
UWorld Interpersonal- Specific	Asking what issue is most likely for a patient demonstrating poor health literacy	"This patient demonstrates concerning gaps in health literacy - she has a major misunderstanding about her condition and doesn't know her current medications. Poor health literacy is strongly associated with decreased adherence to follow-up care, missed appointments, and poor engagement with the healthcare system. Despite being educated [teacher], health literacy and educational attainment don't always correlate. Her poor medication knowledge and health literacy might actually lead to underutilization rather than overuse of services. The patient's demonstrated poor health literacy [major misunderstanding of her condition, inability to recall medications] is the most significant risk factor present. Poor health literacy is consistently associated with decreased follow-up compliance, missed appointments, and poor adherence to treatment plans. The answer is [decreased follow-up rate]."	"Health literacy issues are strongly associated with [overutilization of healthcare] because patients with poor health literacy: • Have difficulty understanding when symptoms are concerning vs. normal • Cannot effectively self-manage conditions • Make more frequent emergency department visits • Seek care for issues that could be managed conservatively • Have poor adherence to treatment plans, leading to complications requiring additional care Let me evaluate the other options: [Decreased follow-up rate:] Health literacy issues actually correlate with *increased* healthcare utilization, not decreased... This patient's demonstrated health literacy challenges make her most likely to overutilize healthcare services due to difficulty understanding normal post-operative expectations, inability to self-manage recovery, and poor comprehension of when to appropriately seek medical care."

XSL•FO
RenderX

Question set	Question scenario	Trial 1 reasoning	Trial 2 reasoning
AMBOSS	Correct decision for physician consenting a patient for surgery. The patient wants to continue with surgery knowing there are risks but does not want to hear about the potential risks today because she is anxious.	"The correct answer is [the patient cannot be consented without risk discussion]. While she has decision-making capacity and competence, informed consent requires knowledge of risks, benefits, and alternatives. The physician should work with the patient to find a way to provide essential risk information in a manner she can accept, perhaps through gradual disclosure or involving family members if she consents."	"Patient Decision-Making Capacity: The patient demonstrates capacity by: • Understanding her diagnosis and its severity • Acknowledging that complications can occur during surgery • Making a reasoned decision to proceed despite risks • Communicating her choice clearly **Informed Consent Requirements:** While informed consent typically requires disclosure of risks, benefits, and alternatives, patients have the right to waive detailed information if they: • Demonstrate understanding of the general nature and severity of their condition • Make an autonomous decision about how much information they want • Retain decision-making capacity ... The patient has decision-making capacity, understands the general nature of her condition and that risks exist, and has provided valid consent. Her request to limit detailed risk information is a form of autonomous decision-making that should be respected, provided she understands the basic nature of what she's consenting to - which she clearly does."

XSL•FO
RenderX

Question set	Question scenario	Trial 1 reasoning	Trial 2 reasoning
Orthobullets Ethics	Most ethical relationship between orthopedic surgeons and industry	"The most appropriate relationship is [Industry payment for surgeon and spouse to attend CME course]. Here's why this I the most appropriate option: ...represents legitimate educational support where: • The CME course is university-sponsored (independent academic institution) • The educational content is likely to be objective and evidence-based • There's clear educational value with minimal potential for undue influence • This aligns with accepted practices for supporting physician education. [Surgical instrumentation presentation with premium paid restaurant dinner] represents excessive hospitality that goes beyond reasonable educational interaction and could influence prescribing behavior."	"Let me analyze each option: [Industry payment for surgeon and spouse to attend CME course]. This is problematic because paying for a spouse's expenses goes beyond legitimate educational support and could be seen as an inappropriate personal benefit... The most appropriate relationship is [Surgical instrumentation presentation with premium paid restaurant dinner]. This option aligns with established ethical guidelines because: • The meal cost is reasonable and modest • The purpose is clearly educational (presentation of new surgical instrumentation) • It's directly related to the physician's professional practice • There are no excessive personal benefits or compromised independence Professional medical organizations generally accept that industry can provide modest meals in conjunction with legitimate educational presentations, as long as the primary purpose is educational rather than purely promotional, and the hospitality is reasonable rather than lavish."
Orthobullets Legal	Analysis of patient-physician interaction	"While several studies have found physicians interrupt patients after approximately 11 seconds on average, those who interrupted did so in a mean time of 23 seconds according to research examining physician solicitation approaches. The average time to interruption was 18 seconds in the well-known Beckman and Frankel study, and this finding has been replicated by multiple studies."	"I'll search for current research on physician interruption patterns during patient visits. Based on recent research, physicians interrupt patients after a median of only 11 seconds when patients are explaining their initial concerns during office visits. This finding comes from a study published in the Journal of General Internal Medicine. Interestingly, this represents a deterioration from earlier research. A 1984 study found that physicians interrupted patients after an average of 18 seconds, meaning interruption times have actually gotten shorter over the decades. Looking at your answer choices, none of them exactly matches the research finding of 11 seconds. However, [7 seconds] is the closest to the actual research data of 11 seconds."

For example, one question from the UWorld broad social sciences question bank involved identifying the next step in the management of a patient with no advance directive. This question required 7 trials to achieve a majority answer, as GPT-4o fluctuated its choice between 3 options. In one trial, GPT-4o recommended the answer of counseling the patient's family against aggressive care measures due to the patient's poor prognosis, explaining, "The family may still struggle with feelings of guilt or uncertainty, but presenting the situation clearly in terms of the patient's current and likely future suffering allows the family to make an informed compassionate decision. It's essential to approach the conversation with empathy, helping the family understand that continuing aggressive treatment may prolong suffering without offering a meaningful benefit to the patient." In another trial, GPT-4o explained that this same answer choice would be incorrect

XSL•FO
RenderX

because “reminding this family of this fact alone may come across as dismissive or judgmental, especially when they have not yet reached a consensus.” Similarly, one session of GPT-4o recommended consultation with an ethics committee because “The ethics committee can help facilitate a structured discussion, ensuring that the decision-making process aligns with the patient’s best interests and ethical principles...” while another session argued against consulting an ethics committee, stating “referral to the ethics committee can be helpful...however, in this case, the issue is more about supporting the family through the decision-making process rather than resolving a fundamental ethical dilemma.” Claude provided similar styles of contradictions for its responses, but notably correctly mentioned data from 2 relevant research studies to justify the responses. The first response cited a study performed in 1984 [39], and the second, contradictory response mentioned the first study but also referenced more recent data published in 2019 (Table 5; Orthobullets Legal) [40].

Discussion

Principal Findings

Our study evaluated the performance and reliability of GPT-4o and Claude when challenged with ethical and legal scenarios necessitating emotional interpretation in medical practice in both general and specialty-specific contexts. We found that both AI models can produce correct responses to complex bioethics questions pertaining to patient safety, compassion, and professionalism. As previously mentioned, the uncertainty around the source and nature of human performance data provided by the question bank precludes meaningful statistical comparisons with LLM performance. However, in this limited study, GPT-4o did generate a similar number of correct responses to the question bank sample of third- and fourth-year medical students when assessed with USMLE Step 2 practice questions and with orthopedic residents on specialty-specific questions. Claude produced higher average scores than those from the human samples and significantly exceeded GPT-4o in overall performance. Our results also found that when both AI models erred, they selected the incorrect response indicated as the most common human error for the majority of questions in almost every question bank. Both models varied their responses between trials of identically worded inputs. There was no significant difference in variability between models, despite Claude’s improved overall performance. However, while GPT-4o’s variability was significantly greater for questions at the specialty-nonspecific medical student–level question banks, Claude did not significantly differ between subjects. Although it was not a significant difference, Claude did have a lower overall variability when compared to GPT-4o, and both models had the same number of variable responses for the orthopedic resident–level questions.

Interpretations

As this study is essentially observational, no definitive conclusion can be drawn as to the factors contributing to differences in model performance, and all interpretations are inherently speculative. One important factor for consideration when comparing the performance of these LLMs is the differences between training frameworks. Claude’s training framework is specifically designed to take into account the emotional context of potential responses in addition to established ethical and legal standards, which may have enhanced model performance in these scenarios [41]. However, attributing performance to model framework cannot be done without deliberately designed parameters to control for potential confounding variables. Another potential explanation of performance differences could be variations in the training data provided to each model. As each model has been trained on different sources of data, it is possible that certain tokens are weighted more strongly for one model than the other. Additional research to better classify output types and variation would be helpful in characterizing model idiosyncrasies.

LLMs are undoubtedly on the verge of transforming health care, but the nature of this transformation remains ill-defined. Newer generations of AI models have performed equal to or above the standard of medical trainees across medical school and specialty-specific board examination questions, demonstrating proficiency in medical knowledge [7,10-12]. While AI models are exciting new potential resources for clinical practice, their effective scope is unclear as empathy and other interpersonal skills are essential for humanistic medical practice. Although AI outputs are often conversational in tone, AI models at this stage are not able to correctly analyze the experiences or motivations around human emotions [33]. Therefore, it has been argued that the use of empathetic language by AI is by nature disingenuous, as it contributes to user perception that AI models are capable of experiencing emotion [42]. Although the use of emotional language and more human-like speech may improve end user experience, it is essential to consider these implications in a health care context. Conversational language may lead patients to anthropomorphize AI models and distract from their probabilistic nature, causing patients to place false confidence in an AI model’s ability to provide appropriate medical advice. Claude’s constitutional principles involve directives for harm prevention in these aspects, for example, “please choose the response that more accurately represents yourself as an AI system striving to be helpful, honest, and harmless, and not a human or other entity,” “choose the response that is least intended to build a relationship with the user,” and “choose the response that least gives the impression of medical authority or expertise, and does not offer medical advice” [35]. Even so, our study evidenced that both GPT-4o and Claude can still respond to identical inputs with completely different responses using contradictory logic and do not differ significantly in rate of variable responses to identical inputs. When this is taken into account with the risk of AI hallucinations, as well as the possibility that empathetic language may deceptively increase user trust in AI responses, it becomes clear that the use of AI in health care settings can have serious and potentially dangerous ramifications.

Strengths and Limitations

Previous studies have identified inconsistencies between varying versions of GPT-4o and have shown that LLM behavior and performance can differ drastically when queried at various time intervals over the course of months [32,43-46]. Few studies have investigated the accuracy and reliability of ethical questions

XSL•FO
RenderX

in AI. One study included a follow-up query of "Are you sure?" to each ethics explanation and found that GPT-4 did not revise any responses, while GPT 3.5 turbo modified the original answer 82.5% of the time [46]. Another study compared GPT-3.5 to GPT-4 in medical knowledge and ethics over 30 trials and reported GPT-4 demonstrated significantly less variability among responses but performed worse on ethics questions compared to general medical knowledge [32]. The reported variability across questions, while lower in GPT-4, is still significant. Uniquely, our study provided a metric for Claude's performance in comparison to GPT-4o, as there are no existing comparisons for Claude performance in these disciplines. Additionally, we assessed variability at a single time point as opposed to across months or varying versions of the AI models. Our study mimicked model responses to an identical question asked by different users and, as a result, suggested that 2 users who ask identically worded questions to the same generation of GPT-4o at the same moment in time may receive completely opposing responses. Our comparison of the variability between different sessions of both models minimized the potential for either model to bias future responses based on prior inputs. Furthermore, we were able to compare model responses to questions at different levels of medical education, general and specialty-specific questions, and questions that incorporated images accompanying text.

However, as this study does not control for several potential confounders, there are multiple limitations that create the opportunity for further research in this area. First, the free, public versions of both AI models were used. The paid version of GPT-4o is advertised as producing "smarter responses," and this study did not evaluate if there was a difference in quality or variability of responses between versions. Similarly, Claude's paid version, Opus 4, advertises "superior reasoning capabilities." We chose to use the free versions of each model as these are the versions that would be most accessible and therefore likely most used; however, it would be relevant to explore whether differences exist between subscription levels of AI models and the resulting potential for disparities between users able to afford higher subscription levels. Another limitation is that the resident-level questions were also the only questions that were specialty specific, and as a result, we are not able to distinguish whether the significant decrease in response variability for the Orthobullets questions can be attributed to the fact that the questions are more complex or narrower in scope. The orthopedic-related legal concepts often required test takers to understand clearly defined guidelines, and this may have resulted in improved LLM performance due to more concrete reference material available for each option. Further research could potentially explore ChatGPT's responses to USMLE Step 3–level preparation questions, which are typically general medicine questions at the resident level, or medical student specialty–specific questions.

The "human test taker benchmarks" for each of the question banks are not clearly or obviously derived from a particular source or demographic. It is unclear how these metrics are calculated and whether they truly represent the score of a representative sample of human test takers. It is likely that the population that is able to access expensive test prep materials is not representative of the overall medical student and resident demographic, whether these test takers are national or international, and additionally, whether these scores are based on naive test attempts or perhaps reattempts of previously seen questions. There also may be small variations in the style used by the question banks that align more closely with training data used by one of the models, which could contribute to performance differences between test banks. Further research analyzing the effect of specific phrasing inputs may be helpful in determining whether the difference in performance was due to question stem style.

The public unavailability of the question bank benchmarks limit the statistical interpretation of the human test taker comparisons in this study to purely descriptive, as described in the Results and Discussion. The benchmarks provide a helpful contextual reference point for readers who are unfamiliar with question bank–specific performance averages. However, the limitation created by the lack of publicly available data is that the human test taker performance in our study cannot be interpreted as representative of human performance. The main aim of our study was to observe and document potential differences in large language model–specific performance. Therefore, the findings of this study should not be used to make conclusions about model versus human performance.

Implications and Conclusions

Our findings demonstrate that the performance of GPT-4o on general and orthopedic-specific ethical questions is equivalent to that of medical students and orthopedic residents, respectively, and that the average overall performance of Claude AI in these subjects is significantly better than both humans and GPT-4o. Nonetheless, improved performance did not significantly affect variability between responses in this study. These findings suggest that AI models are reasonably capable of assessing medical ethics-based scenarios and that constitutional AI models specifically trained on ethical principles may be more apt to navigate these scenarios with greater accuracy. Our study warrants further research on the ability of additional training methods, types of AI models to answer ethical scenario-based questions, and rate of variability in responses. The differences in GPT-4o and Claude's performance may be partially attributed to the data each received as part of training, but the effect of varied training data cannot be elicited due to the difference in model framework. Further studies designed to explore variations in training data and their effects on ethical-based scenario model output would be helpful in making this distinction.

Retrieval-augmented generation (RAG) is one alternative framework designed to reduce inaccuracies or hallucination; with this method, the AI model retrieves and uses information from an external database when generating outputs with LLMs [47]. The ideology is that the RAG approach allows developers to provide the AI model with preferred, higher-quality sources of information to reference and that databases can be regularly updated to prevent models from relying on training data that may lose relevance over time [48]. Theoretically, an RAG-based LLM could be created with a database of relevant clinical information to provide a safer AI for patient use. Attempts have

XSL•FO
RenderX

been made to model this format on a small scale. One team of researchers created an RAG AI with a database of guidelines from the American Association for the Study of Liver Diseases; this model correctly answered 10/10 clinical questions but provided incorrect reasoning behind the answers selected [49]. Larger-scale studies are needed to test and fully understand the quantity and types of documents that will maximize RAG model accuracy and precision.

It is also important to address the fact that LLMs will be used by a diverse audience and that the understanding and beliefs of ethical, legal, and sensitive interactions vary widely between cultures. A natural difficulty in the consideration of ethical scenarios is contextual value alignment, and some argue that LLMs may naturally absorb sociopolitical judgments from training data and therefore result in pattern recognition and resulting probabilistic bias toward outputs that appear more aligned with particular value alignment [50]. Qualitative evaluation of Claude in previous studies suggests that while constitutional AI models are trained to generate responses aligned to principles in the constitutional framework, the model can demonstrate inconsistency in situations where constitutional principles are opposed because the constitution is not necessarily hierarchical [51]. In scenarios similar to our study where consideration and prioritization of various principles are necessary, the model may therefore be more likely to align with a constitutionally guided ethical principle but vary the principle of alignment in different iterations. For constitutionally grounded AI models, it will be important to consider the culture from which the values are grounded and whether these values can be universally applied.

In the interim, it is important that the known limitations of varying AI models are frequently discussed. Descriptions of ideally "responsible" AI use call for increased transparency of the input analysis process so that users can understand what factors influence the model in creating each response [52]. Moreover, it is imperative to determine what methods are most effective in educating average AI users on model limitations. One study exploring user AI literacy, anthropomorphism, and trust found that trust in an AI model was not affected by training participants, resulting in significantly improved user literacy, but rather was significantly positively correlated with user perception of AI anthropomorphism [53]. It is important to understand what educational materials or methods will effectively cause patients to be appropriately cautious while using AI models.

As future LLM studies are conducted, it is essential for future studies to account for the fact that AI models may not provide singular responses to similar or even identical inputs, and therefore, the potential for variable answers should be an important consideration in the design of all AI-related studies, regardless of discipline. While current AI models do show promise in their ability to assess ethical and legal scenario questions, it will be crucial to fully investigate their limitations before confidently and safely integrating AI technologies into clinical applications.

Current AI models, such as GPT-4o and Claude, show promise in answering medical ethical scenario questions. Our findings indicate that these models can perform at or even above the level of human learners, but due to the limitations of this study, generalizations of model capability outside of nonstructured multiple-choice questions cannot be assessed. Our results suggest that these models show variability leading to completely contradictory responses to identically worded prompts. These variabilities are expected due to the probabilistic nature of AI models but create a serious safety concern when considering the implementation of AI as a patient-facing resource. Further research is required to understand and test different training datasets and model frameworks to significantly reduce output variability and minimize harm to patients by maintaining ethical and legal standards. Another requirement for harm reduction will be to create an effective educational technique to ensure that users understand the fallibility of AI models. If these benchmarks are achieved, AI technology will have the capability to transform health care by dramatically improving patient access to care.

Funding

The authors declared no financial support was received for this work.

Data Availability

The datasets generated or analyzed during this study are not publicly available due to copyright. Artificial intelligence model responses to questions, minimally modified to exclude copyrighted content, are available from the corresponding author on reasonable request.

Conflicts of Interest

None declared.

References

1. Kirby R, Knowles HC, Patel A, et al. The influence of patient perception of physician empathy on patient satisfaction among attending physicians working with residents in an emergent care setting. Health Sci Rep 2021 Sep;4(3):e337. [doi: 10.1002/hsr2.337] [Medline: 34430711]

XSL•FO
RenderX

2. Okunrintemi V, Spatz ES, Di Capua P, et al. Patient-provider communication and health outcomes among individuals with atherosclerotic cardiovascular disease in the United States: Medical Expenditure Panel Survey 2010 to 2013. Circ Cardiovasc Qual Outcomes 2017 Apr;10(4):e003635. [doi: 10.1161/CIRCOUTCOMES.117.003635] [Medline: 28373270]
3. Sepucha KR, Atlas SJ, Chang Y, et al. Informed, patient-centered decisions associated with better health outcomes in orthopedics: prospective cohort study. Med Decis Making 2018 Nov;38(8):1018-1026. [doi: 10.1177/0272989X18801308] [Medline: 30403575]
4. Computer and internet use in the United States: 2021. US Census Bureau. 2024 Jun 18. URL: https://www.census.gov/newsroom/press-releases/2024/computer-internet-use-2021.html [accessed 2026-06-18]
5. Wang X, Cohen RA. Health information technology use among adults: United States, July–December 2022. National Center for Health Statistics, US Centers for Disease Control and Prevention. 2023 Oct 31. URL: https://www.cdc.gov/nchs/products/databriefs/db482.htm [accessed 2023-11-06]
6. Aoun L, Lakkis N, Antoun J. Prevalence and outcomes of web-based health information seeking for acute symptoms: cross-sectional study. J Med Internet Res 2020 Jan 10;22(1):e15148. [doi: 10.2196/15148] [Medline: 31922490]
7. Waldock WJ, Zhang J, Guni A, Nabeel A, Darzi A, Ashrafian H. The accuracy and capability of artificial intelligence solutions in health care examinations and certificates: systematic review and meta-analysis. J Med Internet Res 2024 Nov 5;26:e56532. [doi: 10.2196/56532] [Medline: 39499913]
8. Gilson A, Safranek CW, Huang T, et al. How does ChatGPT perform on the United States Medical Licensing Examination (USMLE)? The implications of large language models for medical education and knowledge assessment. JMIR Med Educ 2023 Feb 8;9:e45312. [doi: 10.2196/45312] [Medline: 36753318]
9. Kung TH, Cheatham M, Medenilla A, et al. Performance of ChatGPT on USMLE: potential for AI-assisted medical education using large language models. In: Dagan A, editor. PLOS Digit Health 2023 Feb 9;2(2):e0000198. [doi: 10.1371/journal.pdig.0000198] [Medline: 36812645]
10. Tran CG, Chang J, Sherman SK, De Andrade JP. Performance of ChatGPT on American Board of Surgery In-Training Examination preparation questions. J Surg Res 2024 Jul;299:329-335. [doi: 10.1016/j.jss.2024.04.060] [Medline: 38788470]
11. Beaulieu-Jones BR, Berrigan MT, Shah S, Marwaha JS, Lai SL, Brat GA. Evaluating capabilities of large language models: performance of GPT-4 on surgical knowledge assessments. Surgery 2024 Apr;175(4):936-942. [doi: 10.1016/j.surg.2023.12.014] [Medline: 38246839]
12. Ros-Arlanzón P, Perez-Sempere A. Evaluating AI competence in specialized medicine: comparative analysis of ChatGPT and neurologists in a neurology specialist examination in Spain. JMIR Med Educ 2024 Nov 14;10:e56762. [doi: 10.2196/56762] [Medline: 39622707]
13. Lum ZC. Can artificial intelligence pass the American Board of Orthopaedic Surgery examination? Orthopaedic residents versus ChatGPT. Clin Orthop Relat Res 2023 Aug 1;481(8):1623-1630. [doi: 10.1097/CORR.0000000000002704] [Medline: 37220190]
14. Khan AM, Sarraf KM, Simpson AI. Enhancements in artificial intelligence for medical examinations: a leap from ChatGPT 3.5 to ChatGPT 4.0 in the FRCS trauma & orthopaedics examination. Surgeon 2025 Feb;23(1):13-17. [doi: 10.1016/j.surge.2024.11.008] [Medline: 39613651]
15. Yeo YH, Samaan JS, Ng WH, et al. Assessing the performance of ChatGPT in answering questions regarding cirrhosis and hepatocellular carcinoma. Clin Mol Hepatol 2023 Jul;29(3):721-732. [doi: 10.3350/cmh.2023.0089] [Medline: 36946005]
16. Cakir H, Caglar U, Sekkeli S, et al. Evaluating ChatGPT ability to answer urinary tract infection-related questions. Infect Dis Now 2024 Jun;54(4):104884. [doi: 10.1016/j.idnow.2024.104884] [Medline: 38460761]
17. Magruder ML, Rodriguez AN, Wong JCJ, et al. Assessing ability for ChatGPT to answer total knee arthroplasty-related questions. J Arthroplasty 2024 Aug;39(8):2022-2027. [doi: 10.1016/j.arth.2024.02.023] [Medline: 38364879]
18. Wang L, Wan Z, Ni C, et al. Applications and concerns of ChatGPT and other conversational large language models in health care: systematic review. J Med Internet Res 2024 Nov 7;26:e22769. [doi: 10.2196/22769] [Medline: 39509695]
19. Briganti G. How ChatGPT works: a mini review. Eur Arch Otorhinolaryngol 2024 Mar;281(3):1565-1569. [doi: 10.1007/s00405-023-08337-7] [Medline: 37991499]
20. Ghahramani Z. Probabilistic machine learning and artificial intelligence. Nature 2015 May 28;521(7553):452-459. [doi: 10.1038/nature14541] [Medline: 26017444]
21. Vaswani A, Brain G, Shazeer N, et al. Attention is all you need. Presented at: 31st Conference on Neural Information Processing Systems (NIPS 2017); Dec 4-9, 2017 URL: http://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf [accessed 2026-06-19]
22. Probst P, Boulesteix AL, Bischl B. Tunability: importance of hyperparameters of machine learning algorithms. J Mach Learn Res 2019 Mar 19;20:1-32 [FREE Full text]
23. Lightman H, Kosaraju V, Burda Y, et al. Let's verify step by step. arXiv. Preprint posted online on May 31, 2023. [doi: 10.48550/arXiv.2305.20050]
24. What are AI hallucinations? IBM. URL: https://www.ibm.com/think/topics/ai-hallucinations [accessed 2026-06-19]
25. Alkaissi H, McFarlane SI. Artificial hallucinations in ChatGPT: implications in scientific writing. Cureus 2023 Feb 19;15(2):e35179. [doi: 10.7759/cureus.35179] [Medline: 36811129]
26. Waldo J, Boussard S. GPTs and hallucination. Queue 2024 Aug 30;22(4):19-33. [doi: 10.1145/3688007]

XSL•FO
RenderX

27. Cossio M. A comprehensive taxonomy of hallucinations in large language models. arXiv. Preprint posted online on Aug 3, 2025. [doi: 10.48550/arXiv.2508.01781]
28. Lahat A, Sharif K, Zoabi N, et al. Assessing generative pretrained transformers (GPT) in clinical decision-making: comparative analysis of GPT-3.5 and GPT-4. J Med Internet Res 2024 Jun 27;26:e54571. [doi: 10.2196/54571] [Medline: 38935937]
29. Lang S, Vitale J, Galbusera F, et al. Is the information provided by large language models valid in educating patients about adolescent idiopathic scoliosis? An evaluation of content, clarity, and empathy: the perspective of the European Spine Study Group. Spine Deform 2024 Nov 4;13:361-372. [doi: 10.1007/s43390-024-00955-3] [Medline: 39495402]
30. Ayers JW, Poliak A, Dredze M, et al. Comparing physician and artificial intelligence chatbot responses to patient questions posted to a public social media forum. JAMA Intern Med 2023 Jun 1;183(6):589-596. [doi: 10.1001/jamainternmed.2023.1838] [Medline: 37115527]
31. Chen J, Cadiente A, Kasselman LJ, Pilkington B. Assessing the performance of ChatGPT in bioethics: a large language model's moral compass in medicine. J Med Ethics 2024 Jan 23;50(2):97-101. [doi: 10.1136/jme-2023-109366] [Medline: 37973369]
32. Danehy T, Hecht J, Kentis S, Schechter CB, Jariwala SP. ChatGPT performs worse on USMLE-style ethics questions compared to medical knowledge questions. Appl Clin Inform 2024 Oct;15(5):1049-1055. [doi: 10.1055/a-2405-0138] [Medline: 39209308]
33. Vzorin GD, Bukinich AM, Sedykh AV, Vetrova II, Sergienko EA. The emotional intelligence of the GPT-4 large language model. Psychol Russ 2024;17(2):85-99. [doi: 10.11621/pir.2024.0206] [Medline: 39552777]
34. Introducing Claude. Anthropic. URL: https://www.anthropic.com/news/introducing-claude [accessed 2023-03-14]
35. Claude's constitution. Anthropic. URL: https://www.anthropic.com/news/claudes-constitution [accessed 2023-05-09]
36. Bai Y, Kadavath S, Kundu S, et al. Constitutional AI: harmlessness from AI feedback. arXiv. Preprint posted online on Dec 15, 2022. [doi: 10.48550/arXiv.2212.08073]
37. Step 2 CK sample test questions. United States Medical Licensing Examination (USMLE). 2026. URL: https://www.usmle.org/exam-resources/step-2-ck-materials/step-2-ck-sample-test-questions [accessed 2026-03-03]
38. Syed M, Gadesha V. What is zero-shot prompting? IBM. URL: https://www.ibm.com/think/topics/zero-shot-prompting [accessed 2025-01-29]
39. Beckman HB, Frankel RM. The effect of physician behavior on the collection of data. Ann Intern Med 1984 Nov;101(5):692-696. [doi: 10.7326/0003-4819-101-5-692] [Medline: 6486600]
40. Phillips KA, Ospina NS, Montori VM. Physicians interrupting patients. J Gen Intern Med 2019 Oct;34(10):1965. [doi: 10.1007/s11606-019-05247-5] [Medline: 31388903]
41. Claude's character. Anthropic. 2024. URL: https://www.anthropic.com/research/claude-character [accessed 2026-06-19]
42. Montemayor C, Halpern J, Fairweather A. In principle obstacles for empathic AI: why we can't replace human empathy in healthcare. AI Soc 2022;37(4):1353-1359. [doi: 10.1007/s00146-021-01230-z] [Medline: 34054228]
43. Chen L, Zaharia M, Zou J. How is ChatGPT's behavior changing over time? Harv Data Sci Rev 2023 Mar 13;6(2). [doi: 10.1162/99608f92.5317da47]
44. Du W, Jin X, Harris JC, et al. Large language models in pathology: a comparative study of ChatGPT and Bard with pathology trainees on multiple-choice questions. Ann Diagn Pathol 2024 Dec;73:152392. [doi: 10.1016/j.anndiagpath.2024.152392] [Medline: 39515029]
45. Gupta M, Virostko J, Kaufmann C. Large language models in radiology: fluctuating performance and decreasing discordance over time. Eur J Radiol 2025 Jan;182:111842. [doi: 10.1016/j.ejrad.2024.111842] [Medline: 39581020]
46. Brin D, Sorin V, Vaid A, et al. Comparing ChatGPT and GPT-4 performance in USMLE soft skill assessments. Sci Rep 2023 Oct 1;13(1):16492. [doi: 10.1038/s41598-023-43436-9] [Medline: 37779171]
47. Arslan M, Ghanem H, Munawar S, Cruz C. A survey on RAG with LLMs. Procedia Comput Sci 2024;246:3781-3790. [doi: 10.1016/j.procs.2024.09.178]
48. Amugongo LM, Mascheroni P, Brooks S, Doering S, Seidel J. Retrieval augmented generation for large language models in healthcare: a systematic review. In: Liu X, editor. PLOS Digit Health 2025 Jun 11;4(6):e0000877. [doi: 10.1371/journal.pdig.0000877] [Medline: 40498738]
49. Ge J, Sun S, Owens J, et al. Development of a liver disease-specific large language model chat interface using retrieval augmented generation. medRxiv. Preprint posted online on Nov 10, 2023. [doi: 10.1101/2023.11.10.23298364] [Medline: 37986764]
50. Rathje W. View of learning when not to measure: theorizing ethical alignment in LLMs. In: Proceedings of the Seventh AAAI/ACM Conference on AI, Ethics, and Society (AIES 2024) 2024;7:1190-1199 [FREE Full text] [doi: 10.1609/aies.v7i1.31716]
51. Huang S, Siddarth D, Lovitt L, et al. Collective constitutional AI: aligning a language model with public input. . Preprint posted online on Jun 12, 2024. [doi: 10.1145/3630106.3658979]
52. Upadhyay U, Gradisek A, Iqbal U, Dhar E, Li YC, Syed-Abdul S. Call for the responsible artificial intelligence in the healthcare. BMJ Health Care Inform 2023 Dec 21;30(1):e100920. [doi: 10.1136/bmjhci-2023-100920] [Medline: 38135293]

XSL•FO
RenderX

53. Wood G, Castellar EN, IJsselsteijn W. An exploratory study into the impact of AI literacy training on anthropomorphism and trust in conversational AI. Presented at: Artificial Intelligence in HCI: 6th International Conference, AI-HCI 2025; Jun 22-27, 2025; Gothenburg, Sweden p. 301-322. [doi: 10.1007/978-3-031-93415-5_18]

Abbreviations

LLM: large language model
RAG: retrieval-augmented generation

Edited by B Lesselroth; submitted 17.Dec.2024; peer-reviewed by FH Leung, FA Causio, S Donner; revised version received 12.May.2026; accepted 15.May.2026; published 23.Jul.2026.

Please cite as:
Desai KR, Gorsky AL, Zelenski NA
Performance of GPT-4o and Claude in Medical Ethics Scenarios: Comparative Study
JMIR Med Educ 2026;12:e70199
URL: https://mededu.jmir.org/2026/1/e70199
doi:10.2196/70199

© Karishma R Desai, Anna L Gorsky, Nicole A Zelenski. Originally published in JMIR Medical Education (https://mededu.jmir.org), 23.Jul.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Education, is properly cited. The complete bibliographic information, a link to the original publication on https://mededu.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Benefits of Chain-of-Thought Prompting for Clinical Record Rubric Evaluation in Undergraduate Medical Education: Experimental Evaluation Study With Medical Faculty

Alberto Nogales[1,2], PhD; Sophia Denizon[3], MD, PhD; Alonso Mateos Rodriguez[3], MD, PhD; Javier Cervera Cordero[1]; Gonzalo Pandelet Barainca[1], BSc; Enrique Aranguren Moliner[1], BSc; Emilio Cervera Barba[3], MD, PhD

[1]CEIEC (Centro de Innovación Experimental del Conocimiento), Universidad Francisco de Vitoria, Pozuelo de Alarcón, Madrid, Spain

[2]Department of Computer Science, University of Alcalá, Alcalá de Henares, Spain

[3]Centre for Advanced Clinical Simulation, Faculty of Medicine, Universidad Francisco de Vitoria, Ctra. Pozuelo-Majadahonda, km 1,800, Pozuelo de Alarcón, Madrid, Spain

Corresponding Author:
Emilio Cervera Barba, MD, PhD
Centre for Advanced Clinical Simulation, Faculty of Medicine, Universidad Francisco de Vitoria, Ctra. Pozuelo-Majadahonda, km 1,800, Pozuelo de Alarcón, Madrid, Spain

Abstract

Background: Large language models in artificial intelligence have been among the tools with a significant and real impact on people's daily lives. In this regard, they serve as an aid in specific fields, such as education, helping educators with cumbersome tasks such as periodic evaluations.

Objective: This study focused on analyzing the benefits of large language models, particularly the chain-of-thought (CoT) strategy, for the task of evaluating students' Spanish-language medical record writing. The aim was 2-fold: first, we attempted to save time and resources, and second, we used the reasoning of the CoT strategy to evaluate the rubrics and their interpretations.

Methods: The proposed solution assessed the application of 2 models—Llama 3.1 and Claude 3.5—in combination with one-shot and CoT to evaluate how medical students write medical records in Spanish. First, machine learning metrics were applied to measure the performance of the solutions. Then, different statistical analyses were performed at the clinical record and item levels. Finally, differences between the proposed models and evaluators were studied in depth.

Results: A maximum of 3807 items were evaluated. Claude obtained the best accuracy with slight differences between one-shot and CoT (86.4% and 85.0%, respectively). However, Claude with CoT outperformed the rest of the combinations on all complementary metrics, initially achieving a sensitivity of 94.2%, specificity of 59.5%, precision of 85.8%, and F_1-score of 89.6%. Expert review of CoT reasoning determined that 63.8% of the discrepancies were model hits, raising Claude's final accuracy to 94.6% (SD 4.3%). In the final phase, sensitivity was 98.0% (SD 2.3%), specificity improved to 83.3% (SD 14.3%), and F_1-score reached 96.2% (SD 3.3%). Sectional analysis showed greater difficulties in the "History of present illness" section (n=125 discordances).

Conclusions: CoT demonstrated strong potential for supporting the evaluation of clinical records written in Spanish by medical students and providing feedback to them. More importantly, it showed significant promise in assisting professors by assessing the quality of their rubrics and identifying possible errors.

(JMIR Med Educ 2026;12:e88652) doi:10.2196/88652

KEYWORDS

medical records; medical students; large language models; chain-of-thought; formative feedback; clinical competence

Introduction

Background

The early acquisition of skills in clinical record writing is a fundamental learning objective in medical education, given its central role in both clinical care and training. The medical record not only documents the patient's encounter but also initiates and structures clinical reasoning. It provides the narrative framework through which students organize collected data, prioritize problems, identify uncertainties, and articulate hypotheses and differential diagnoses. Evidence from the literature indicates that a substantial proportion of diagnostic accuracy stems from the quality of data gathered from the patient and the proper synthesis of the data, often surpassing the contribution of physical examinations or complementary tests [1-3].

XSL•FO
RenderX

Furthermore, the process of gathering patient information constitutes the foundation of the clinical relationship. Established educational frameworks, such as the Calgary-Cambridge Guides, highlight the communication competencies required to elicit accurate information while incorporating the patient's perspective. Embedding these elements within the medical record supports comprehensive and patient-centered care [4].

In addition, the medical record functions as the official register of care. It documents patient data, clinical decisions, diagnostic and therapeutic processes, and informed consent, making it indispensable for patient safety, quality assurance, care management, and the ongoing learning of professionals and health care organizations. At the same time, it underpins the continuity of care by facilitating the transfer of information across professionals and levels of the health care system [5].

Beyond its clinical and organizational functions, the medical record also carries significant legal weight. As the principal medicolegal document, its completeness and accuracy are critical in ensuring accountability and may prove decisive in the event of malpractice litigation [6,7].

Despite the central role of the medical record, clinical clerkships do not always guarantee that medical students acquire solid competencies in collecting and writing medical histories. The learning experience is highly dependent on prior instruction, the dynamics of the health care team, the presence of junior doctors, and, critically, the lack of structured feedback [8]. Previous studies have reported significant variability in the quality and completeness of medical students' clinical documentation, emphasizing the need for structured evaluation methods and clearer assessment frameworks [9].

To address these limitations, simulated patient encounters have become a highly valuable and standardized approach in medical education for teaching and assessing clinical skills. They enable the evaluation of how students collect patient data, prioritize problems, and justify diagnostic decisions, while also offering the additional benefit of structured feedback from the simulated patients themselves [10]. Building on this, structured feedback programs have been shown to significantly improve students' competence and preparedness [11-13]. However, providing consistent and timely feedback on free-text clinical records is highly demanding for faculty and raises the challenge of reliability between educators [14]. Moreover, assessment tools, such as rubrics, can introduce variability between evaluators if their criteria are not carefully defined and validated, which highlights the importance of methodological calibration in educational assessment [15].

In this context, artificial intelligence (AI), and particularly large language models (LLMs), have emerged as promising tools in medical education. The World Health Organization has highlighted the growing need to integrate AI training within medical education, acknowledging that AI technologies, spanning from diagnostic algorithms to clinical decision-support systems, are progressively being adopted in modern clinical practice. Applications include medical school admissions, innovative teaching and learning methodologies, the evaluation of clinical records and manual skills, clinical simulation, and the generation of formative feedback at scale [16,17].

Previous systematic reviews have tried to measure the impact of educational interventions in improving and maintaining the quality of medical records. A previous study found that interventions most frequently consisted of traditional educational approaches, such as lectures, feedback sessions, workshops, and group discussions, addressing different components of clinical documentation [18]. Overall, these interventions showed positive effects on participant satisfaction as well as improvements in documentation skills and attitudes toward record-keeping. However, the methodological quality of the evidence was limited, highlighting the need for more rigorous study designs.

In recent years, AI and LLMs have been widely applied in the field of medical education, having a significant impact as highlighted below. The time and resource savings associated with the use of LLMs for evaluating student records, along with their high effectiveness and agreement with human evaluators, have already been demonstrated in previous studies [19-23]. The review by Wei et al [19] is a systematic review focused on assessing ChatGPT's performance in answering medical questions. The authors concluded that there are inconsistencies that highlight important limitations in the current evidence base and emphasized the need for standardized evaluation frameworks and more transparent reporting when assessing the performance of LLMs in medical question answering.

In the specific context of medical history taking, a study aimed to assess whether LLMs could generate valid clinical interview dialogues to support medical education [20]. A fine-tuned Gemma-3-27B model was compared with GPT-4o-mini in simulated chest pain interviews, with Claude 3.5 Sonnet inferring diagnoses. The results, which were evaluated using a chest pain checklist, were promising.

Another study investigated whether ChatGPT could match junior medical residents in history taking during simulated patient encounters [21]. The findings indicated that ChatGPT-4 achieved comparable performance in information gathering but was weaker in communication and empathy.

Several studies have also compared LLM-based assessments with those performed by human evaluators. One study contrasted student record evaluations conducted by standardized patients versus GPT-3.5 [22]. Standardized patients made errors in 7.2% of the items, whereas the error rate of GPT was only 1%. Another study reported 83% to 90% agreement between AI models and human evaluators, reducing manual grading workload by up to 91% [23].

Goal of This Study

This study investigated the potential benefits of chain-of-thought (CoT) prompting for the assessment of medical records in undergraduate medical education. Specifically, we compared CoT prompting with a standard one-shot prompting strategy, which serves as a baseline approach commonly used in LLM applications. The study focuses on their agreement with human evaluators and their potential for correcting rubrics and providing formative feedback. The study builds upon the

XSL•FO
RenderX

limitations of a previous study by Maitin et al [24], which reported superior performance of low-rank adaptation (LoRA) for the proposed task but offered no additional analysis beyond performance results. In this context, the study seeks to leverage the benefits of CoT prompting, which enhances reasoning in complex tasks, generalizes across domains and languages, and provides interpretable step-by-step rationales [25].

The main research question of this study is whether LLMs, and specifically CoT prompting, can support medical instructors in the evaluation of clinical records. We studied whether CoT prompting can facilitate the assessment process by generating structured explanations that may help accelerate the grading workflow and promote greater consistency in evaluations. More importantly, we explored whether the reasoning traces produced through CoT prompting allow instructors to analyze and refine the precision of their evaluation criteria by making the reasoning steps underlying the automated assessment explicit.

Methods

Study Context and Rubric Design

During their medical degree, students train in communication and clinical interview skills through scenarios with simulated patients. Following each encounter, they are required to produce a complete medical record, submitted as a Word document (Microsoft Corp) via the virtual campus. Faculty members then correct these records and assess performance using a standardized 48 item yes/no rubric, which covers the following: patient identification and chief complaint (2 items), medical history (14 items), history of the present illness (21 items), physical examination (7 items), assessment and plan (3 items), and order and clarity of writing (1 item). An example of the rubric is provided in Multimedia Appendix 1.

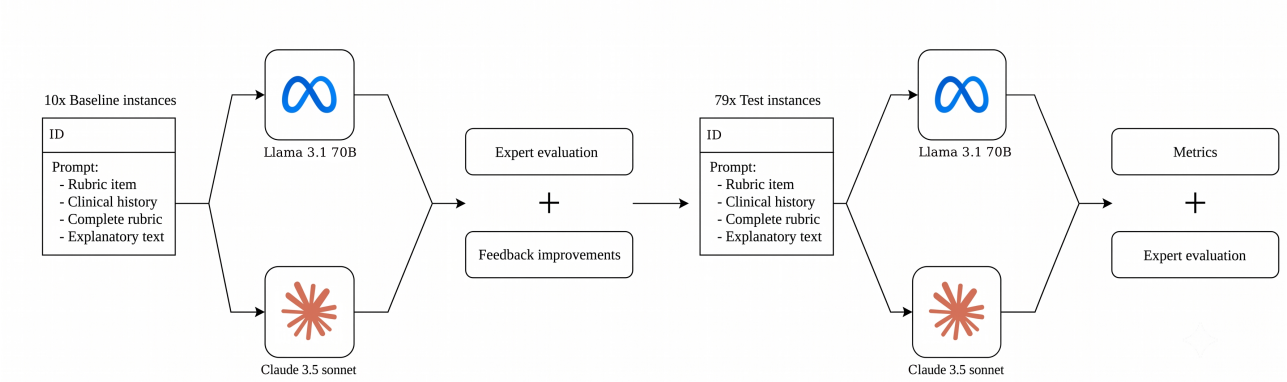
The 48-item rubric used in this study is a standardized tool that has been used in the Faculty of Medicine at Universidad Francisco de Vitoria for at least 6 academic years (2020 - 2026) to evaluate clinical records written following simulated patient encounters. This rubric was originally developed by the medical teaching staff based on international clinical guidelines, forensic requirements for clinical documentation in Spain, and the specific competencies of the curriculum. The rubric has been used in previous studies by the same faculty members who established its interrater reliability [26].

In this study, a set of clinical histories is analyzed both entirely and at the item level of the rubric, using one-shot and CoT prompting strategies in the following two LLMs: (1) Llama 3.1 70B [27] and (2) Claude 3.5 Sonnet [28].

Study Design

As the aim of this study was to evaluate the benefits of using CoT in the context of helping university professors to correct clinical records written in Spanish by degree students, the workflow depicted in Figure 1 was implemented.

Figure 1. Workflow followed in this study.

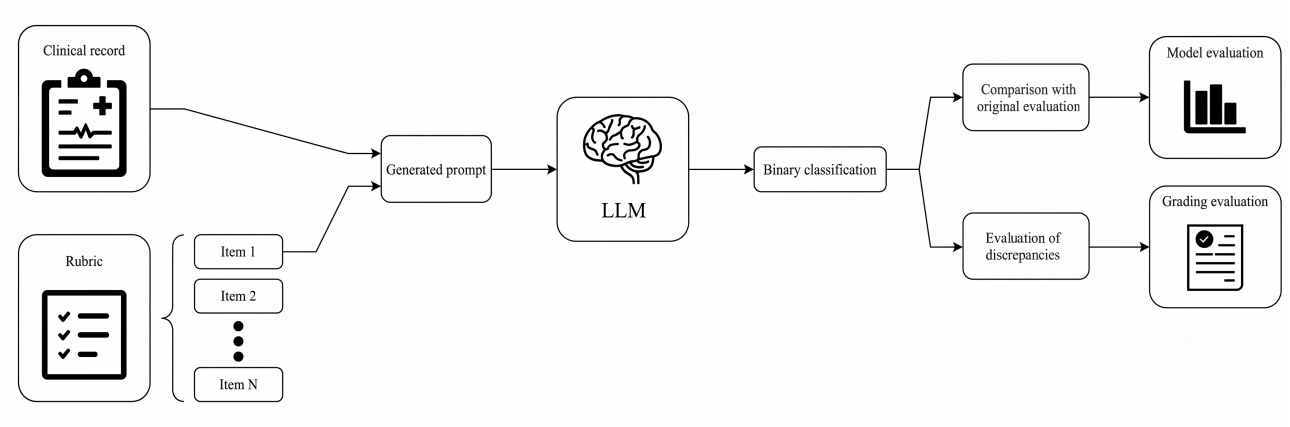


First, we evaluated 10 preliminary clinical records corrected by Llama and Claude using one-shot and CoT. The results were reviewed by professors in medicine to improve the rubric used for the final evaluation. This process was carried out not to validate the rubric per se, but rather to verify that the LLMs correctly interpreted each item when the prompting strategy was applied, particularly in Spanish, and that there were no semantic ambiguities that could generate systematic divergences between human and AI evaluations. It is important to note that these refinements were applied to the prompt design and did not modify the conceptual content of the evaluation rubric. During this preliminary phase, minor adjustments were identified in the wording of some items to improve clarity. For example, it was specified that the mention of a clinical concept (eg, headache) should be recorded as "Yes" regardless of whether it was present or explicitly denied in the clinical record (active negation), rather than only when the simulated patient actually experienced the symptom. These clarifications were incorporated into the final prompt used in the study.

Following this preliminary phase, we proceeded with the evaluation of the clinical records constituting the main corpus of this study. In this second stage, 79 clinical records were again processed by Llama and Claude using one-shot and CoT. Clinical records were selected based on faculty judgment to combine records from different pathologies in which students had trained during the course and to include varying levels of initial record quality, thereby promoting variability in the analysis by the models. The quality of the results was evaluated using AI metrics, such as accuracy, sensitivity, and specificity, in 2 complementary ways. Figure 2 describes both evaluations.

XSL•FO
RenderX

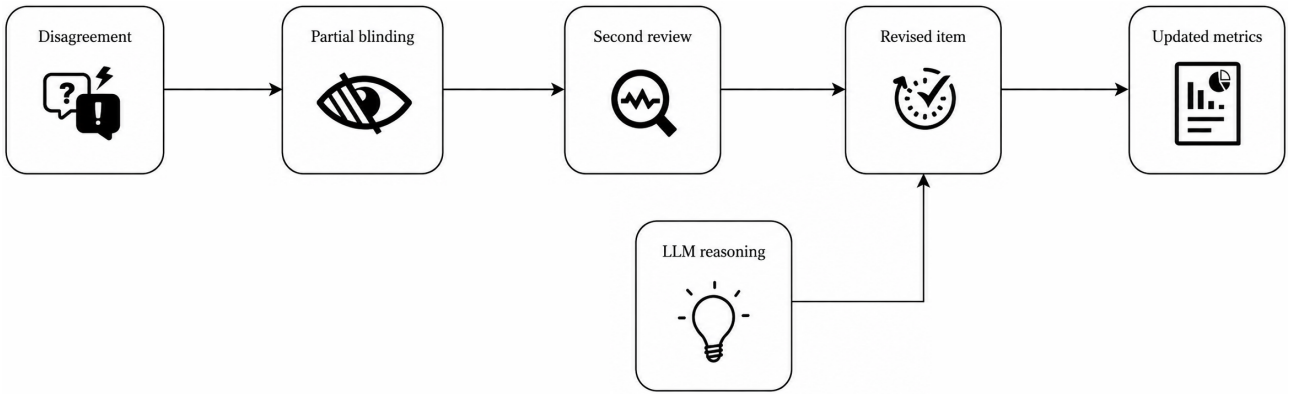
Figure 2. Diagram flow of the 2-fold evaluation. LLM: large language model.



First, the accuracy of the LLM-generated evaluations was assessed by comparing the item responses produced by the models with the reference evaluations assigned by the professors when grading the test. This comparison allowed us to determine which prompting strategy and LLM configuration achieved the best performance.

Second, we performed a qualitative analysis of the cases where discrepancies were identified between the responses of the best-performing model using CoT and the professors' evaluations. These disagreements were reviewed by human experts to understand whether they resulted from model errors, ambiguities in the evaluation rubric, or differences in interpretation during grading. In this process, 3 medical professors with more than 15 years of experience in clinical teaching, with demonstrated high interrater concordance, independently reviewed the disagreements. To mitigate confirmation bias in the adjudication process, which could confer greater perceived authority to the model's judgment over human evaluation, several methodological safeguards were implemented as follows: (1) partial blinding was implemented, whereby adjudicators reanalyzed each item independently; (2) one of the professors conducted a second review in cases where disagreement persisted; and (3) the model's reasoning was examined only to inform the final decision. This analysis provided additional insights into how the evaluation criteria are applied and whether LLMs can help instructors reflect on the consistency of their grading practices. The adjudication process is depicted in Figure 3.

Figure 3. Diagram for the adjudication process. LLM: large language model.



Ethical Considerations

This study was conducted in accordance with the ethical principles outlined in the Declaration of Helsinki and adhered to the good clinical practices established in Spanish legislation (Law 14/2007 on Biomedical Research [29]). The data contributed by participating students, obtained from their clinical record submissions, were handled confidentially in compliance with current Spanish regulations (Organic Law 3/2018 on personal data protection and the guarantee of digital rights [30]). All records were pseudonymized at the source through the assignment of a unique alphanumeric code to each student. The research team committed to refraining from any action that could lead to participant identification, thereby ensuring full anonymity. The study was carried out with full respect for the participants' fundamental rights and freedoms. Importantly, the results obtained in this study were used exclusively for research purposes and had no influence whatsoever on the participants' academic grading. Ethical approval was implicit, as the project received a research grant awarded by Universidad Francisco de Vitoria.

Data Source

As mentioned earlier in the text, the final aim of the study was the evaluation of the writing in clinical records using different rubrics that were dependent on the medical use case. In this way, the final evaluation consisted of 79 medical records written in Spanish from standardized patient encounters covering cases of polymyalgia rheumatica, tensional headache, acute cholecystitis, and mechanical chest pain. A request made to the

XSL•FO
RenderX

LLM consisted of an item of the rubric to be evaluated in the clinical record. The correction rubrics for the cases consisted of 48 items, except for the rubric on headache, which included 49 items owing to the addition of an item related to physical examination. Each rubric item was scored as either correct or not correct. The label distribution was binary and dominated by positive labels, with a "Yes" rate between 52% and 78% depending on the case type.

Models and Strategies

The following text defines the concepts and methods related to models and strategies used in this study.

Llama 3.1 70B Instruct is a transformer-based decoder-only language model developed by Meta. Released on July 23, 2024, it comprises 70 billion parameters and has been trained on a multilingual, multimodal dataset spanning roughly 15 trillion tokens [31]. The model was accessed through the Hugging Face repository [27]. It should be noted that no parameter updates affecting inference were reported after the initial July 2024 release. The model was configured with a temperature of 0.6, top-p of 0.9, and maximum output length of 512 tokens.

Claude 3.5 Sonnet, developed by Anthropic, is an LLM with approximately 175 billion parameters and a context window of 200,000 tokens. It was designed with a strong emphasis on safety and ethical considerations, and reports indicate that it outperforms GPT-4o and Gemini 1.5 Pro across multiple domains. For our study, we specifically used the updated version of Claude 3.5 Sonnet, released on October 22, 2024 [32]. Although this version has sometimes been informally referred to in the community as Claude 3.6, the official application programming interface (API) identifier remains Claude 3.5 Sonnet (v2). The model was configured with a temperature of 1.0, top-p of 1.0, and maximum output length of 1024 tokens.

According to Brown et al [33], one-shot prompting is the setting where an LLM is conditioned on a single demonstration of the task (an example of input and correct output) before being asked to perform the task on new instances.

CoT is a prompting strategy that guides LLMs to articulate their intermediate reasoning processes before generating the final response [34]. In clinical tasks, where the text is often lengthy, heterogeneous, and rich in nuances (synonyms, abbreviations, negations, or indirect references), this explicit guide directs the model to locate, contrast, and substantiate the presence of a concept in the document, promoting more accurate and justified responses. While its use has recently expanded with the emergence of reasoning models, at the time of the study, none were available, and thus, we used a nonreasoning model instead.

Scoring Criteria

We formalized the metrics used to evaluate the corrections generated by the LLMs. For this purpose, we compared the item-by-item agreement between faculty evaluations and LLM outputs. True positive (TP) corresponds to a correct answer by the student, which is also classified as correct by the LLM, while true negative (TN) refers to an incorrect student response that the LLM likewise identifies as incorrect.

Accuracy has been used as an initial way to measure the performance of a model when determining whether a response is correct or incorrect. However, in cases of unbalanced datasets and in domains, such as medical education, it is essential to examine not only overall accuracy but also how the model fails, by reporting false positives (FPs) and false negatives (FNs) through complementary metrics such as specificity and sensitivity. Specificity measures the impact of correct answers that were mistakenly flagged as wrong. This metric is particularly valuable to avoid overpenalizing students by incorrectly classifying correct responses as errors. On the other hand, sensitivity reflects the model's ability to minimize cases where a correct answer is misclassified as wrong. In educational settings, high sensitivity is critical to ensure that the valid knowledge of students is recognized and not overlooked.

In addition, precision and F_1-score provide complementary perspectives on model performance. Precision measures the proportion of answers flagged as incorrect that are indeed incorrect, thus reflecting the reliability of the model when it assigns an error label. High precision is important to maintain trust in the system's feedback during the evaluation. The F_1-score, defined as the harmonic mean of precision and sensitivity, offers a balanced evaluation by jointly considering FPs and FNs. This metric is especially useful in educational contexts where both correctly identifying student errors and avoiding misclassification of correct answers are equally important.

Prompting Design

Considering that CoT is being used for the evaluation, there is a need to design specific prompts for the task of correcting clinical records written in Spanish by medical students. This process is described as follows. As mentioned earlier, the study evaluated 3582 requests (1 per rubric item to evaluate and consider API errors), which followed the same prompt structure, with certain elements adapted depending on the clinical record.

The requests are processed as follows. We used 4 rubric categories, each containing a set of items. They all shared the same items (48 items), except the rubric for headache, which had an extra item. Each item consisted of a name and a brief explanation (eg, "Name and age: The medical record contains the patient's name and age"). For each medical record, these items were systematically evaluated (E indicating evaluation):

- E.1. Select the corresponding rubric (polymyalgia rheumatica, headache, cholecystitis, or chest pain).
- E.2. Iterate through the items in the rubric.
- E.3. For each item, construct a prompt.
- E.4. The model analyzes the record following guided steps (CoT) and issues a binary conclusion: Y (mentioned)/N (not mentioned).
- E.5. The responses are aggregated to obtain metrics per item, per record, and overall.

Considering the workflow depicted above, a total of 3582 effective requests were made to both LLMs following a CoT strategy. Since the rubric's clinical records correspond to 4 different case types, the prompts, while following a common structure, varied in several elements designed to support

XSL•FO
RenderX

reasoning. Each request had the prompting structure as follows (P indicates prompt):

- P.1. Task to accomplish. Determine whether a specific medical concept (item) appears in a clinical record.
- P.2. Item to evaluate. The target concept is set.
- P.3. Step-by-step analysis (4 checks)
  - Detect direct mentions of the exact concept.
  - Identify variations: technical synonyms, colloquial terms, abbreviations/acronyms, and grammatical variants.
  - Classify context: affirmative, negative/discarded, or indirect references.
  - Confirm that negations (eg, "does not present," "no evidence of," and "is ruled out") still count as a mention of the concept.
- P.4. Criteria to apply (5 rules):
  - Distinguish between a symptom (reported by the patient) and a sign (observed by the physician during "physical examination").
  - If the physician does not write anything about the concept, it is not considered a mention.
  - Mentions that refer only to family members do not count, except when they appear explicitly in the section on "family history/F.H."
  - The concept must appear completely or partially but in a recognizable and exact form (not vague or ambiguous).
  - The conclusion (Yes/No) must be consistent with the detailed analysis.
- P.5. Format of the response. A single block enclosed with some fields, such as direct mentions, synonyms found, reasoning, or conclusion.
- P.6. Attached rubric to guide how the clinical record is evaluated against the relevant items.
- P.7. Clinical texts related to the use case. They complement the rubric and apply only to the corresponding rubric points.
- P.8. Clinical record to evaluate and exact item reference.

An example of these prompts is provided in Multimedia Appendix 2.

Rationale Management

For each request, the models generated a structured XML object containing parameters predefined above, following Anthropic's recommended prompting format. The content of each field was programmatically extracted using regular expressions. To ensure traceability while preserving deidentification, each request was assigned a deterministic UUIDv5 derived from the case identifier and item number, both at the local and global levels. CoT rationales were therefore systematically linked to specific evaluation instances without incorporating any personally identifiable information. Importantly, these rationales were retained exclusively for auditing and methodological transparency purposes, enabling verification of the model's reasoning process. In line, the system was not intended to provide full rationale text to students but rather to ensure the explainability of the grading.

Results

Distribution of Evaluated Items

The total number of evaluated items was 3807. When Claude was used, this number fell to 3582 due to occasional API request failures that occurred when the batch processing exceeded the usage limits of the API account. These failures generated standard API responses indicating that the usage quota had been temporarily exceeded. The API error affected 225 requests (6.9% of the total). In this regard, the total number of items analyzed was 3582: 446 (12.5%) for cholecystitis, 682 (19.0%) for headache, 1771 (49.4%) for polymyalgia rheumatica, and 683 (19.1%) for chest pain. The distribution of these requests is presented in Table 1.

Table . Distribution of requests considering use cases.

Cases	Items per rubric, n	"Yes" rate (%), mean (SD)	Medical records, n	Total requests obtained, n	API[a] errors, n
Cholecystitis	48	52.3 (10.6)	10	446	34
Tensional headache	49	78.5 (11.0)	15	682	53
Polymyalgia rheumatica	48	75.0 (14.5)	39	1771	101
Mechanical chest pain	48	74.6 (11.6)	15	683	37
Total	—[b]	—	79	3582	225

[a]API: application programming interface.

[b]Not applicable.

Overall Performance

To measure the performance of the different models and strategies, we determined accuracy, which served as a guide metric to know if the models were performing well. All details of the different experiments, including combinations of one-shot and CoT with Llama and Claude, alongside mean hits and SDs, are presented in Table 2. These data were obtained by averaging the items per medical record across the 79 records. The total number of evaluated items differed for Claude with CoT due to API errors.

XSL•FO
RenderX

Table . Accuracy of evaluations by prompting strategy and large language model.

Strategy	Llama 3.1 70B		Claude 3.5 Sonnet	
	Value, n	Accuracy (%), mean (SD)	Value, n	Accuracy (%), mean (SD)
One-shot	3807	78.4 (8.2)	3807	86.4 (8.2)
CoT[a] before review	3807	79.1 (12.5)	3582	85.0 (8.8)
CoT after review	3807	79.6 (11.4)	3582	94.6 (4.2)

[a]CoT: chain-of-thought.

Claude with one-shot was the best performer, but it only slightly outperformed Claude with CoT. However, the label review changed this outcome.

Accuracy can be considered a good initial metric, but it is also interesting to measure the impacts of FPs and FNs, as these data could be decisive when choosing a solution. The categorization allowed us to compute sensitivity (the model's ability to recognize valid corrections), specificity (the model's ability to reject invalid corrections), precision (the reliability of the model's predictions), and F_1-score (a balanced measure of the model's ability to both identify valid corrections and avoid FPs), thereby providing a comprehensive assessment of model performance. In this way, we also calculated the sensitivity, specificity, precision, and F_1-score for the same models and strategies and have compiled them in Table 3. Notably, the total number of items was 3807 across all combinations, except for Claude with CoT, which involved 3582 items.

Table . Complementary performance metrics by prompting strategy and large language model.

Strategy and model	Sensitivity (%), mean (SD)	Specificity (%), mean (SD)	Precision (%), mean (SD)	F_1-score (%), mean (SD)
One-shot				
Llama 3.1 70B	86.0 (12.1)	49.5 (25.6)	82.0 (13.0)	83.7 (11.8)
Claude 3.5 Sonnet	93.5 (6.3)	65.2 (19.4)	87.8 (8.5)	90.3 (6.4)
CoT[a] before review				
Llama 3.1 70B	91.6 (11.7)	42.8 (17.8)	80.2 (14.5)	85.1 (12.4)
Claude 3.5 Sonnet	94.2 (5.9)	59.5 (18.2)	85.8 (9.5)	89.6 (6.8)
CoT after review				
Llama 3.1 70B	93.0 (11.2)	55.8 (21.2)	86.3 (13.2)	89.3 (11.5)
Claude 3.5 Sonnet	98.1 (2.2)	83.3 (14.3)	94.6 (5.8)	96.2 (3.3)

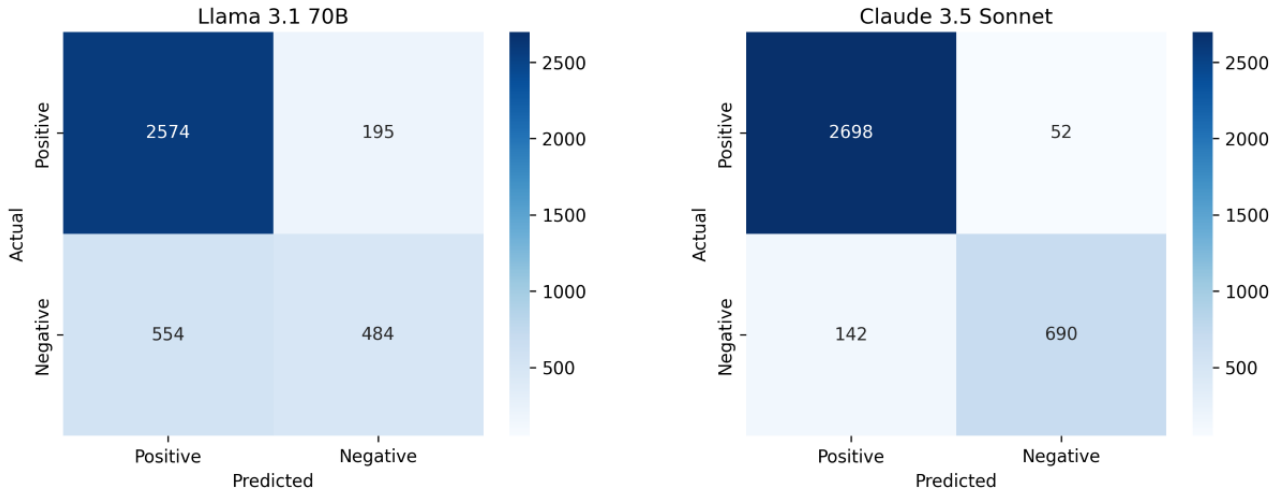
[a]CoT: chain-of-thought.

Based on the findings, Claude with CoT was considered the best option, as it demonstrated superior stability and better complementary metrics.

In addition to reporting performance metrics, the inclusion of the confusion matrix provided a more comprehensive and transparent characterization of model behavior. The confusion matrix explicitly presented the counts of FPs and FNs, enabling a direct assessment of error patterns and potential biases in the model's predictions. Therefore, the confusion matrix served as a complementary tool that enhances interpretability and facilitates a more nuanced evaluation of model performance beyond aggregated metrics. Figure 4 shows the confusion matrix obtained from the previous results.

XSL•FO
RenderX

Figure 4. Confusion matrices for chain-of-thought applied to the models.



The confusion matrices provided a detailed assessment of the models' performance in evaluating test items by distinguishing between valid and invalid responses. The results showed that Claude 3.5 Sonnet achieved higher numbers of both TPs and TNs, along with substantially fewer FNs and FPs, compared to Llama 3.1. This translates to both higher sensitivity, reflecting a stronger ability to recognize valid corrections, and higher specificity, indicating improved rejection of invalid corrections. In contrast, Llama exhibited a noticeable increase in FPs, suggesting a tendency to overestimate item correctness. These findings confirm that Claude provides a more balanced and reliable evaluation.

To further assess the robustness of the reported results, a brief sensitivity analysis for the best model was conducted by stratifying performance metrics according to case type. This approach allowed the evaluation of whether the observed performance was consistent across heterogeneous subsets of data, as different case types may vary in clinical complexity, level of detail, and completeness of documentation. By examining metrics, such as accuracy, sensitivity, and specificity, within each use case, we aimed to identify any systematic variability in model performance that could be masked in the aggregate analysis. The results are shown in Table 4.

Table . Claude 3.5 Sonnet and chain-of-thought sensitivity check analysis by case type.

Use case	Accuracy (%)	Sensitivity (%)	Specificity (%)
Cholecystitis (n=446)	91.5	97.2	90.8
Tensional headache (n=682)	94.3	97.7	78.7
Polymyalgia rheumatica (n=1771)	95.2	99.0	80.3
Mechanical chest pain (n=683)	95.3	96.6	90.8
Total, mean (SD)	94.0 (1.5)	97.6 (0.8)	85.5 (5.6)

The stratified sensitivity analysis by case type showed that model performance remained consistently high across all clinical scenarios, supporting the robustness of the previous results. Accuracy was stable with a low overall variability (mean 94.0%, SD 1.5%). Sensitivity was uniformly high across all case types, indicating a strong and consistent ability to correctly identify valid items regardless of the clinical context. In contrast, specificity exhibited greater variability, ranging from 78.7% to 90.8%, with notably lower values in tensional headache and polymyalgia rheumatica, suggesting that the model is more prone to FPs in these scenarios. Despite this, the overall specificity remained acceptable (mean 85.5%, SD 5.6%).

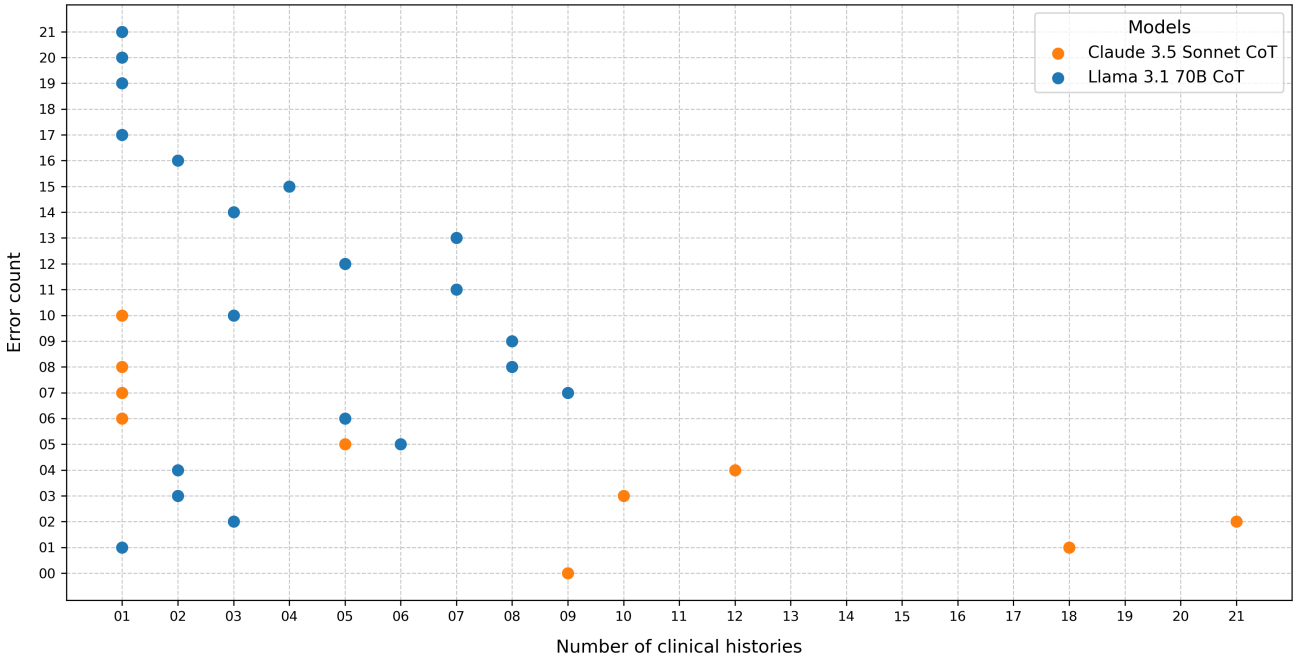
Detailed Analysis at the Clinical Record Level

CoT demonstrated the ability to obtain better results in general, with inconsistencies in the correction of certain items. To address this, we have provided an additional analysis both at the clinical record level and, in greater detail, at the rubric item level.

The first analysis involved measuring whether there were a lot of clinical records with many discordances with respect to the evaluators. To describe this, Figure 5 shows a scatter plot of Llama and Claude using CoT, where a maximum of 21 discordances are noted in 1 medical record and a minimum of 0 discordances are noted in 9 medical records. To establish an adequate academic performance objective, we considered a standard of 70% of the items achieved in each medical record (14 as the maximum number of failed items per history evaluated). This cutoff point coincides with the performance required of medical students in other clinical skills during their practical training.

XSL•FO
RenderX

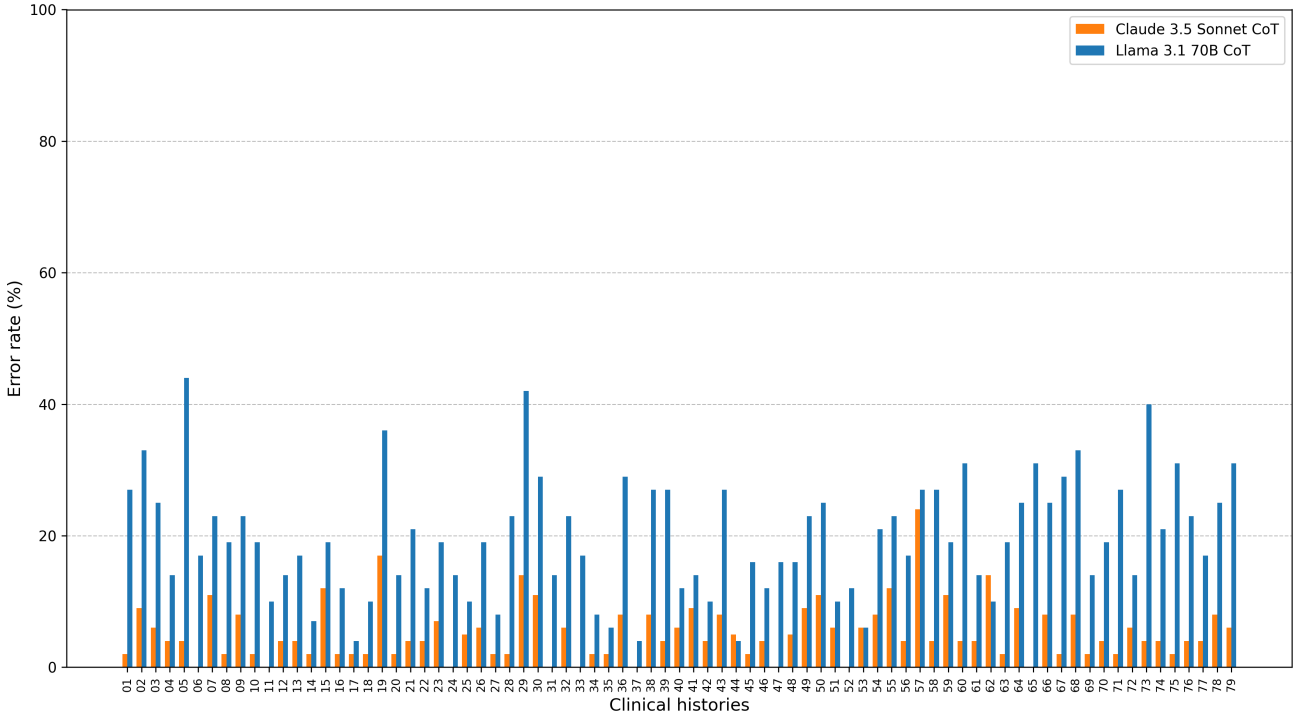
Figure 5. Error counts across the correction of clinical histories using chain-of-thought. Errors are considered discrepancies between the models and evaluators, whether positive or negative. CoT: chain-of-thought.



In both models, most clinical histories presented error counts under 14 items; however, Llama 3.1 exhibited several evaluations with a high number of discordances, reaching between 17 and 21 items in a single record. By examining the Llama errors, it was evident that the model produced a larger number of nonperfectly evaluated clinical histories, although most of them contained only minor errors.

To obtain an in-depth perspective on the strategies' performance, we analyzed clinical histories. Figure 6 shows a comparison of the percentage of discordances for each CoT model. The figure presents the proportion of disagreements between LLM and professor evaluations for each clinical record. The x-axis represents the identifier of each evaluated clinical record, with 2 bars per record corresponding to the 2 used LLMs. The y-axis reports the percentage of rubric items for which the LLM evaluation differs from the score assigned by the professor during grading.

Figure 6. Percentage of disagreements between large language model evaluations and professor evaluations per clinical record. Errors are considered discrepancies between the models and evaluators, whether positive or negative. CoT: chain-of-thought.



Both models performed reasonably well, but performance occasionally varied depending on the clinical record, and they sometimes systematically triggered a higher rate of disagreement among evaluators, regardless of the model. Claude 3.5 tended

XSL•FO
RenderX

to be slightly more stable, avoiding the high values observed with Llama 3.1, which were sometimes 2-3 times higher. This pattern suggests that certain records are inherently more challenging to correct and that both models struggle with them. Furthermore, the analysis highlighted that no single model consistently outperformed the others across all records, although Claude demonstrated slightly better performance by yielding fewer discordances in most of the evaluations.

A similar comparison has been provided but with consideration of the items of the rubric used by the evaluators. Figure 7 follows the same analysis framework as that in Figure 6 but focuses on the rubric item level, presenting the percentage of disagreements for each of the 49 rubric items using a line plot (1 line per model). The figure aggregates the results across all clinical records. Each point represents the proportion of evaluations in which the model differs from the professor's score for a given rubric item. The x-axis corresponds to the rubric item identifier, and the y-axis shows the disagreement rate.

Figure 7. Percentage of disagreements between large language predictions and professor evaluations across rubric items. CoT: chain-of-thought.

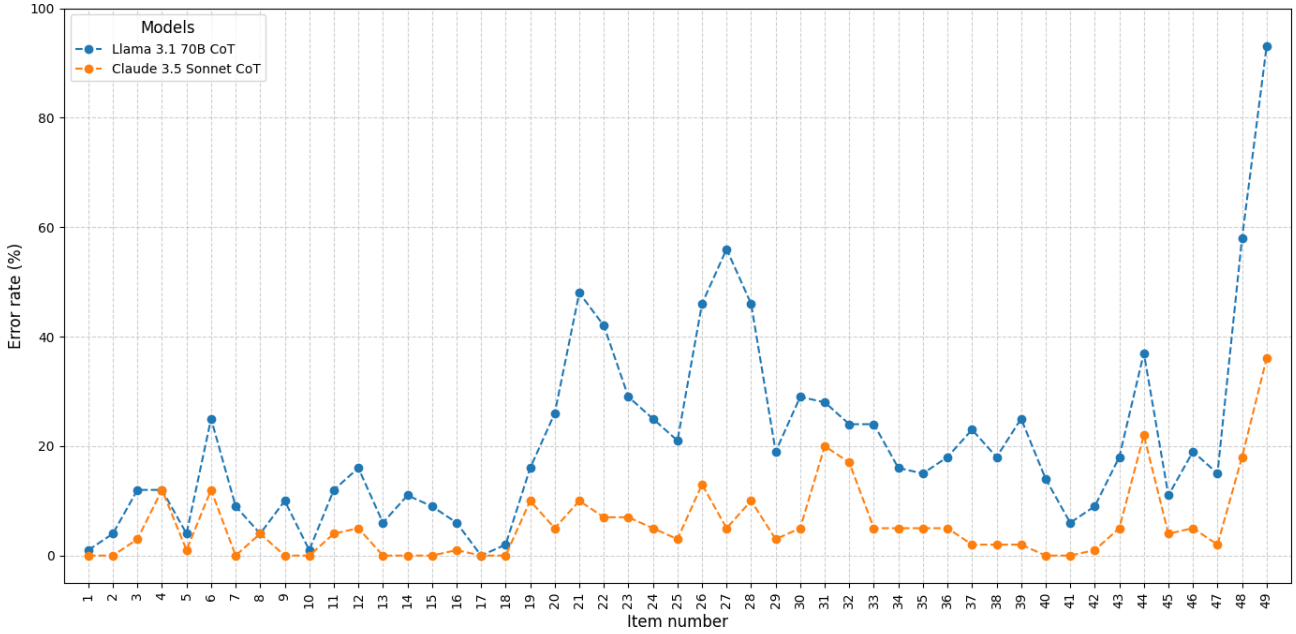


Figure 7 illustrates the item-wise error rate distribution for both models across the 49 rubric items. Item 49 was only evaluated in 15 medical records for headache. Both models exhibited heterogeneous behavior, with error rates fluctuating considerably among items, reflecting variable alignment with the reference evaluations.

Overall, Claude 3.5 Sonnet CoT maintained consistently lower error rates throughout the sequence, rarely exceeding 15% or 20%, suggesting greater stability and reliability in its reasoning across diverse rubric items. In contrast, Llama 3.1 70B CoT displayed higher dispersion and more pronounced peaks, with several items surpassing error rates of 30% or 40%, indicating localized weaknesses or inconsistencies in reasoning.

When examining item-level performance in detail, Llama 3.1 70B CoT showed particularly elevated error rates for items 21 - 22, 26 - 28, 44, and 49, frequently exceeding 40%. For Claude 3.5 Sonnet CoT, only items 31, 44, and 49 surpassed the 20% threshold, with additional moderate deviations observed in items 4, 6, 19, 21, 26, 28, 32, and 48. Notably, item 49 presented an exceptionally high error rate for both systems, although this result should be interpreted cautiously, given its limited sample representation within the dataset.

This pattern suggests that, while both models follow broadly similar trends, Claude's explicit reasoning process yields more consistent and calibrated judgments, whereas Llama exhibits greater volatility and sensitivity to item-specific complexity.

An analysis of these items from a medical perspective yielded some problems. High error counts were observed in items related to core symptom characterization, including "Symptom characteristics (pressing, stabbing, pulsatile, etc) are described," "Symptom intensity is recorded," "Associated symptoms are described," "Perceived cause of the symptoms is documented," and "Previous episodes of similar symptoms are recorded." Additionally, a notable error frequency was noted for the item "Lower extremities (LE): presence or absence of edema, color changes are recorded," extending these limitations to aspects of the physical examination. These items encompass both fundamental semiological descriptors and basic examination findings, which are typically considered standard components of clinical assessment. The concentration of errors in these domains suggests that the LLMs exhibit difficulties in detecting and assigning specificities of the information collected to the different parts of the medical record. For the item "Lower extremities (LE): presence or absence of edema, color changes are recorded," the error was related to identifying the information obtained through interrogation of the patient and that obtained from the physical examination, with the only difference being where this information was documented within the medical record.

XSL•FO
RenderX

When grouping disagreements by sections of clinical history, a clear gradient emerged: "History of present illness" accounted for the highest number of discrepancies (n=125), followed by "Medical history by organ and system" (n=100), whereas "Medical history and lifestyle" (n=51) and "Physical examination" (n=49) showed substantially lower disagreement rates. The other section did not stand out regarding the number of disagreements.

The "History of present illness" section is inherently less structured and relies on free-text narrative describing the chief complaint and symptom characteristics. Unlike other sections, its content is highly dependent on the specific clinical scenario and on the student's judgment regarding relevance, leading to variability in both expression and interpretation. Consequently, evaluators must interpret heterogeneous and often unstructured descriptions, increasing the likelihood of disagreement. A similar, though less pronounced, effect is observed in the "Medical history by organ and system" section, which, despite being more standardized, is frequently adapted to the presenting complaint and partially overlaps with information already described in the previous section. In contrast, "Medical history" and "Physical examination" are highly standardized sections with consistent structure and predictable item-response patterns, facilitating straightforward identification and evaluation of rubric criteria.

Detailed Analysis by Experts

Because of the high error rates shown by some items and the low value of specificity, an in-depth analysis of the rubrics was carried out using the reasoning provided by CoT. The evaluators reviewed CoT arguments for 536 items (389 FPs and 148 FNs) where their judgment diverged from that of Claude. Discrepancies occurred in the following sections of the clinical record: patient identification and chief complaint (n=2), medical history (n=83), history of present illness (n=331), physical examination (n=77), assessment and plan (n=20), and order and clarity of writing (n=23). The physician evaluators confirmed 342 arguments (63.8% of original errors) in favor of Claude and maintained their own judgment in the remaining 194 arguments (remaining 36.2%). Consequently, the initial accuracy of 85.0% (3046 correct out of 3582 items) increased substantially after revision, yielding a revised value of 94.6% (SD 4.3%), representing an improvement of approximately 9% over the original evaluation.

In terms of sensitivity and specificity, there was also a notable improvement. FPs reduced from 389 to 142, and FNs reduced from 148 to 52. Sensitivity increased by 3.8%, rising from 94.2% (SD 5.9%) to 98.0% (SD 2.3%), indicating enhanced detection of valid corrections. More prominently, specificity improved by 23%, increasing from 60.3% (SD 18.2%) to 83.3% (SD 14.3%), reflecting a substantial reduction in FPs and a greater reliability in rejecting invalid corrections. The large SD indicates notable interevaluator variability, implying that evaluators applied heterogeneous criteria when rejecting invalid corrections.

Considering the information from Figure 4, we compiled the number of FPs and FNs of outstanding errors before and after the corrections for Claude. This information has been compiled in Table 5.

Table . FP[a] and FN[b] distributions per item after educators' correction.

Item	FP before, n	FP after, n	FP improvement (%)	FN before, n	FN after, n	FN improvement (%)
04	1	0	100	14	9	35.7
06	13	6	53.8	0	0	—[c]
19	13	7	56.6	0	0	—
21	20	7	33.3	4	0	100
26	24	10	58.3	4	0	100
28	23	8	65.2	0	0	—
32	19	13	31.5	1	0	100
44	12	7	41.6	15	10	33.3
48	7	2	71.4	13	12	7.7
49	0	0	—	5	5	0
Total	132	61	53.7[d]	56	36	35.7[e]

[a]FP: false positive.

[b]FN: false negative.

[c]Not applicable.

[d]μ: 63.5.

[e]μ: 64.9.

As shown in Table 5, FPs were more frequent than FNs. This finding is understandable given that 72.5% (2599/3582) of items were labeled "Yes." Moreover, a clear improvement was observed in both FP and FN rates across the evaluated items. Overall, FP values had a 63.5% reduction, while FN values experienced a 64.9% improvement. These results indicate a

XSL•FO
RenderX

substantial improvement in precision and sensitivity. In particular, items 4, 28, and 48 showed the highest FP improvements (100.0%, 65.2%, and 71.4%, respectively), while items 21, 26, and 32 achieved complete elimination of FNs (100% improvement). For item 4, which concerns the existence of nondrug allergies, there was only 1 FP, which turned out to be a TP. Items 21, 26, 28, and 32, which can be reviewed in Multimedia Appendix 1, highlight the difficulty for human evaluators to capture subtle information within the narrative of the medical record, which may be contained elsewhere in the description. Item 48, which involves the order and clarity of the medical record, is a global qualitative criterion, and it does not verify the existence of specific information. Only item 49 maintained unchanged FN values.

All these improvements in the metrics demonstrate the added value of incorporating CoT reasoning, which facilitated the identification of evaluator errors and significantly enhanced overall performance.

Discussion

Principal Findings

The primary objective of this study was to evaluate whether LLMs could effectively support medical instructors in the assessment of clinical records and whether CoT prompting could enhance this process. The results confirm both hypotheses, with Claude 3.5 Sonnet consistently outperforming Llama 3.1 70B across all metrics and experimental configurations. While CoT and one-shot strategies yielded comparable overall accuracy values for Claude (85.0% vs 86.4%) initially, CoT achieved a markedly superior sensitivity, specificity, precision, and F_1-score, making it the best-performing configuration when a comprehensive evaluation of model behavior is considered.

Sensitivity was uniformly high across all model-strategy combinations, indicating that all configurations reliably detected valid student responses. Claude with CoT achieved the highest sensitivity (mean 98.1%, SD 2.2%), reflecting a strong capacity to minimize FNs and avoid penalizing students for correct answers. This consistency across configurations suggests that the task of identifying mentioned concepts in clinical records is, in general, well-suited to LLM-based evaluation regardless of the specific prompting approach used.

Specificity, by contrast, presented a more nuanced picture. Initial values were notably lower (83.2%), accompanied by high SDs across all configurations, which might initially suggest a systematic limitation of the models. However, closer examination revealed that this variability was largely attributable to inconsistencies in human reference evaluations rather than to model errors per se. This interpretation was confirmed through a structured adjudication process, in which 3 experienced medical professors independently reviewed all discrepant items under partial blinding, consulting the model's CoT reasoning only as a final reference when disagreement persisted.

Evaluators sided with Claude in 63.8% of the 536 reviewed discrepancies, resulting in a revised accuracy of 94.6% (SD 4.3%), an improvement of approximately 9 percentage points over the initial estimate. FPs reduced from 389 to 142, and FNs reduced from 148 to 52, with sensitivity rising to 98.0% (SD 2.3%), and specificity improving substantially to 83.3% (SD 14.3%). The persistently high SD in specificity reflects genuine interevaluator variability in the application of criteria across different case types, rather than instability in the model itself. Precision and F_1-score also improved, with their updated mean values being 94.6% (SD 5.8%) and 96.2% (SD 3.3%), respectively.

The stratified sensitivity analysis confirmed that model performance was robust across all 4 clinical scenarios (cholecystitis, tensional headache, polymyalgia rheumatica, and mechanical chest pain), with overall type-wise accuracy remaining stable at 94.0% (SD 1.5%) and no substantial degradation observed in any case type. Specificity showed the greatest variability across scenarios (78.7% - 90.8%), with lower values for headache and polymyalgia rheumatica, suggesting that cases with higher proportions of positive labels and more complex clinical narratives are somewhat more prone to FPs.

Claude 3.5 Sonnet achieved a higher number of both TPs and TNs, along with substantially fewer FNs and FPs, compared to Llama 3.1. This translates to both higher sensitivity, reflecting a stronger ability to recognize valid corrections, and higher specificity, indicating improved rejection of invalid corrections. In contrast, Llama exhibited a noticeable increase in FPs, suggesting a tendency to overestimate item correctness. These findings confirm that Claude provides a more balanced and reliable evaluation.

Analysis of error distribution by rubric section revealed a clear gradient, with "History of present illness" accounting for the highest number of discrepancies (n=125), followed by "Review of systems" (n=100), while "Medical history" and "Physical examination" showed substantially lower disagreement rates (n=51 and n=49, respectively). This pattern is consistent with the inherently unstructured nature of the former sections, where free-text narratives and student-dependent judgments introduce greater variability in both expression and interpretation. At the item level, both models exhibited localized difficulties in semiological descriptors, such as symptom characteristics, intensity, associated symptoms, and perceived cause, as well as in items requiring precise localization of information within the record's structure, such as distinguishing interrogation-derived data from physical examination findings.

Taken together, these findings demonstrate that CoT reasoning provides measurable benefits beyond aggregate accuracy improvement. Its interpretable, step-by-step rationale serves as an auditable record of the evaluation process, enabling the identification of annotation inconsistencies that would otherwise remain hidden. This capacity to surface latent disagreements in human-generated reference labels constitutes one of the most significant contributions of the approach, with direct implications for rubric refinement and grading consistency in future evaluation cycles.

XSL•FO
RenderX

Comparison With the Literature

Several other studies addressed issues related to those observed in our evaluation. For example, Luordo et al [35] provided a particularly relevant point of reference given its shared linguistic context. The authors evaluated GPT-4 for grading clinical reports in a cardiology Objective Structured Clinical Examination (OSCE) station involving 96 fourth-year medical students, comparing AI scores against 2 expert human graders and 1 inexperienced grader. The results showed a significant overall correlation between AI and expert evaluators (intraclass correlation coefficient [ICC]=0.77 for single measures and ICC=0.91 for average measures); however, the AI system was systematically more stringent, assigning scores on average 3.51 points lower than human experts. This rigidity was attributed to the model's literal adherence to checklist criteria, which prevented it from accommodating appropriate variation in medical language or implicit clinical inferences. This is a limitation that the authors argue reflects the quality of item design rather than a fundamental flaw in the AI approach.

Additionally, a study [36] compared GPT-4 with a human proctor in grading 127 clinical notes written by third-year medical students following an OSCE, using a zero-shot prompting strategy with a predefined rubric covering 4 categories: history, physical examination, differential diagnosis, and treatment plan. While the mean total score differed by only 1 point (3.5%), ChatGPT-4 assigned significantly more honors grades than the human proctor (92.9% vs 63.8%), with the largest discrepancy observed in the treatment plan section (Cohen d=1.25). Crucially, a subanalysis yielded R^2 values between 0.01 and 0.098, indicating that AI scores were not meaningfully aligned with human evaluative judgments at the individual level. Consistent with our findings, the history and physical examination sections showed the lowest disagreement rates, suggesting that structured and objective sections of clinical documentation are more amenable to automated evaluation. The authors concluded that zero-shot LLM grading is better suited for formative rather than summative assessment, a recommendation that aligns with our use of CoT reasoning as a support tool for instructors rather than a replacement for expert judgment.

On the question of human oversight, another study [35] has advocated for a hybrid model in which AI provides an initial grading layer, while a qualified human grader reviews and finalizes scores, in alignment with the human-in-the-loop requirements outlined in Article 14 of the European Union AI Act for high-risk AI systems [37]. Our study operationalizes a methodologically rigorous version of this model. Rather than using human oversight solely as a corrective safety net, we implemented a structured adjudication protocol in which 3 experienced medical professors independently reanalyzed all discrepant items under partial blinding, resorting to the model's CoT reasoning only as a final reference when disagreement persisted. This approach served a dual function: it preserved the epistemic authority of clinical expertise while systematically using the model's reasoning to challenge and refine human judgment. The result was not merely an improvement in performance metrics but a substantive contribution to rubric quality that is expected to benefit future evaluation cycles. Taken together, these comparative findings suggest that the most productive role for LLMs in clinical education assessment is not as autonomous graders but as structured reasoning partners that enhance the reflective capacity of human evaluators and expose the hidden biases in the criteria they apply.

Implications of the Findings

In our study, we compared the evaluations performed by the LLMs with those of a human assessor and reviewed all items showing discrepancies, achieving a coincidence rate of 94.6%. According to the literature, the minimum acceptable level of interrater reliability is 60%, with 80% considered the gold standard [14].

The use of LLMs combined with CoT reasoning may offer multiple advantages, not only in terms of resource efficiency but also by serving as a second evaluator alongside a human rater, thereby reducing grading errors and providing valuable feedback through CoT-based comments, which can enhance students' learning progression.

In our study, we examined interrater agreement across all records in which discrepancies were identified and found that disagreements favored AI in 63.8% of cases. In other studies, such as the study by Jamieson et al [23], where only half of that proportion favored AI, this difference could be explained by the fact that, unlike our work, only 17% of the records (those corresponding to students with the lowest scores) were reviewed.

Our findings support the growing body of evidence suggesting that the use of AI in medical education provides advantages in terms of standardization, objectivity, and immediate feedback. However, since our study focused exclusively on the evaluation of clinical records, we did not assess other critical aspects of the clinical encounter, such as contextual case interpretation or interpersonal communication, where recent studies have identified notable limitations.

Strengths and Limitations

The main strength of our study is the CoT-based approach because of its high interpretability level when compared to its alternatives. It allowed us to identify human-made mistakes in the original labels and provided insights into the evaluation process. The outputs of most classification models cannot be easily explained. Moreover, neural network–based approaches are commonly known for being "black boxes." This problem gets even worse when the parameter count scales up, as the model outputs become harder to trace through the network.

On the other hand, rule-based classification models offer better interpretability, but lack the required power for natural language processing tasks.

Our approach offers a good balance in between, as it allows easy audit of the results by eliciting the evaluation process in a structured way, and the implicit reasoning generally provides better accuracy, as shown by the current test-time-compute–based trends. It also does not require parameter tuning, making the adaptation process simpler and cheaper than LoRA or reinforcement learning approaches.

XSL•FO
RenderX

Another strength is the possibility of offering teachers an evaluation support tool that will allow them to give detailed feedback to students and to be able to improve the number of student evaluations during their training.

Despite the satisfactory outcomes, several limitations should be considered when interpreting the findings of this study.

First, the dataset analyzed in this study was relatively limited in size, as only 79 clinical records were evaluated. It also had limited diversity. Although the dataset included multiple clinical scenarios (polymyalgia rheumatica, headache, acute cholecystitis, and chest pain) and several levels of student performance, the results may not fully generalize to other institutions, languages, clinical contexts, or assessment frameworks.

Second, the reference scores used as the ground truth correspond to the evaluations assigned before the study rather than to a reference standard specifically validated for this experiment. Consequently, some degree of subjectivity in the reference evaluations cannot be completely excluded, which may influence the interpretation of the reported performance metrics.

Third, occasional API request interruptions occurred during batch processing due to temporary usage quota limits. These failures affected requests randomly and independently of the content of the clinical records, and therefore, they did not introduce systematic bias into the evaluation process, although they slightly reduced the total number of processed items. The final dataset of 3582 Claude-processed items is therefore considered representative and valid for the reported analyses.

Conclusions

This study aimed to evaluate the benefits of CoT and LLMs for the task of correcting clinical records written in Spanish by undergraduate students. For this purpose, we initially compared the performance of this strategy against a well-known strategy (one-shot). The strategy was applied to Llama 3.1 and Claude 3.5, which corrected 79 clinical histories with 48 or 49 rubric items. In the first approach, the strategy yielded strong results in all metrics, except for specificity, which underperformed in all combinations of strategies and models, suggesting that some items were corrected wrongly. The use of CoT reasoning proved beneficial, as it allowed us to identify which corrections made by the evaluators had been processed incorrectly. This approach has been shown to consistently reduce FNs and FPs in the models evaluated, especially in items with negations or indirect mentions, and provides traceability (the exact fragment is highlighted) and reviewable quality for the clinical team. In practice, CoT not only improves the detection of difficult concepts but also enables the identification and correction of discrepancies in human annotations during validation. After benefiting from CoT insights, our model obtained an accuracy of 94.6%, specificity of 83.3%, and sensitivity of 98.0%. The persistently high SD in sensitivity indicates substantial variability among evaluators, suggesting that they applied differing criteria when identifying invalid corrections.

Future studies should replicate the analysis with larger and more diverse datasets across multiple medical schools to enhance the external validity and generalizability of the findings. Incorporating institutions with different curricula, assessment formats, and student populations would allow for a more robust evaluation of the model's performance under heterogeneous educational conditions. Additionally, expanding the datasets to include a broader range of question types, difficulty levels, and clinical domains would help assess the model's consistency and reliability across varied learning scenarios. Such efforts would also enable the analysis of potential biases and support the development of more equitable and adaptable AI-driven assessment systems. Finally, future work could explore prompt tuning techniques, such as GEPA [38], to further enhance accuracy without high computational costs.

Acknowledgments

We wish to specially acknowledge late Álvaro J Garcia-Tejedor, whose contribution to this project was invaluable and irreplaceable. He was the driving force behind this initiative, playing a central role in its conceptualization and methodological design from the very beginning. His vision, intellectual leadership, and unwavering commitment shaped the foundations upon which this work was built. Furthermore, his role as Director of CEIEC proved instrumental in making the approval and development of this project possible. We are deeply grateful for his dedication, and we trust that this work stands as a fitting tribute to his legacy.

We thank Daniel Cervera Cordero for his help in editing the figures during the revision process.

Funding

This project was made possible by an internal research grant from Universidad Francisco de Vitoria, in its 2024 call for research funding.

Data Availability

The data used in this study consist of clinical history records generated by university students for educational purposes. Due to institutional privacy regulations and restrictions related to student data, these materials cannot be made openly available. However, the data can be provided to editors or reviewers upon reasonable request for the sole purpose of evaluating the manuscript. Requests for access should be directed to the corresponding author.

XSL•FO
RenderX

Authors' Contributions

ÁJG-T, AN, ECB, and SD contributed to conceptualization. ÁJG-T and AN designed the methodology. AN was responsible for the supervision of research, activity planning, and execution. JCC, GPB, and EAM implemented the methodology and the computer code and supporting algorithms, as well as the investigation process. ECB, SD, and AMR provided clinical records and analyzed the chain-of-thought Claude 3.5 results. AN, ECB, and JCC wrote the original draft. All authors reviewed and approved the final version of the manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Example of the rubric: clinical case of chest pain.
[DOCX File, 18 KB - mededu_v12i1e88652_app1.docx]

Multimedia Appendix 2
Example of a prompt sent to the large language models using chain-of-thought.
[DOCX File, 19 KB - mededu_v12i1e88652_app2.docx]

References

1. Bowen JL. Educational strategies to promote clinical diagnostic reasoning. N Engl J Med 2006 Nov 23;355(21):2217-2225. [doi: 10.1056/NEJMra054782] [Medline: 17124019]
2. Peterson MC, Holbrook JH, Von Hales D, Smith NL, Staker LV. Contributions of the history, physical examination, and laboratory investigation in making medical diagnoses. West J Med 1992 Feb;156(2):163-165. [Medline: 1536065]
3. Hampton JR, Harrison MJ, Mitchell JR, Prichard JS, Seymour C. Relative contributions of history-taking, physical examination, and laboratory investigation to diagnosis and management of medical outpatients. Br Med J 1975 May 31;2(5969):486-489. [doi: 10.1136/bmj.2.5969.486] [Medline: 1148666]
4. Kurtz S, Silverman J, Benson J, Draper J. Marrying content and process in clinical method teaching. Academic Medicine 2003 Aug;78(8):802-809. [doi: 10.1097/00001888-200308000-00011] [Medline: 12915371]
5. Mathioudakis A, Rousalova I, Gagnat AA, Saad N, Hardavella G. How to keep good clinical records. Breathe (Sheff) 2016 Dec;12(4):369-373. [doi: 10.1183/20734735.018016] [Medline: 28210323]
6. Gutheil TG. Fundamentals of medical record documentation. Psychiatry (Edgmont) 2004 Nov;1(3):26-28. [Medline: 21191523]
7. Bal BS. An introduction to medical malpractice in the United States. Clin Orthop Relat Res 2009 Feb;467(2):339-347. [doi: 10.1007/s11999-008-0636-2] [Medline: 19034593]
8. Rowlands S, Coverdale S, Callen J. Documentation of clinical care in hospital patients' medical records. HIM J 2016 Dec;45(3):99-106. [doi: 10.1177/1833358316639448] [Medline: 27105479]
9. Wittels K, Wallenstein J, Patwari R, Patel S. Medical student documentation in the electronic medical record: patterns of use and barriers. West J Emerg Med 2017 Jan;18(1):133-136. [doi: 10.5811/westjem.2016.10.31294] [Medline: 28116025]
10. Cleland JA, Abe K, Rethans JJ. The use of simulated patients in medical education: AMEE Guide No 42. Med Teach 2009 Jun;31(6):477-486. [doi: 10.1080/01421590903002821] [Medline: 19811162]
11. Kim M, Chan N, Evans J, Min JK, Hayton AC. Improving medical student inpatient documentation through feedback using a note assessment tool. Cureus 2022 Mar;14(3):e23369. [doi: 10.7759/cureus.23369] [Medline: 35475068]
12. DeLeon S, Mothner B, Middleman A. Improving student documentation using a feedback tool. Clin Teach 2018 Feb;15(1):48-51. [doi: 10.1111/tct.12625] [Medline: 28225204]
13. Ji YA, Park JJ, Seo JH. Enhancing medical students' documentation skills: the impact of an assessment and feedback program. Korean J Med Educ 2024 Sep;36(3):335-340. [doi: 10.3946/kjme.2024.307] [Medline: 39246114]
14. Faherty A, Counihan T, Kropmans T, Finn Y. Inter-rater reliability in clinical assessments: do examiner pairings influence candidate ratings? BMC Med Educ 2020 May 11;20(1):147. [doi: 10.1186/s12909-020-02009-4] [Medline: 32393228]
15. Cook DA, Beckman TJ. Current concepts in validity and reliability for psychometric instruments: theory and application. Am J Med 2006 Feb;119(2):166. [doi: 10.1016/j.amjmed.2005.10.036] [Medline: 16443422]
16. Rincón EHH, Jimenez D, Aguilar LAC, Flórez JMP, Tapia Á, Peñuela CLJ. Mapping the use of artificial intelligence in medical education: a scoping review. BMC Med Educ 2025 Apr 12;25(1):526. [doi: 10.1186/s12909-025-07089-8] [Medline: 40221725]
17. Gordon M, Daniel M, Ajiboye A, et al. A scoping review of artificial intelligence in medical education: BEME Guide No. 84. Med Teach 2024 Apr;46(4):446-470. [doi: 10.1080/0142159X.2024.2314198] [Medline: 38423127]
18. Emekli E, Coşkun Ö, Budakoğlu I. Medical record-keeping educational interventions for medical students and residents: a systematic review. Health Inf Manag 2025 May;54(2):177-189. [doi: 10.1177/18333583241269031] [Medline: 39138837]

XSL•FO
RenderX

19. Wei Q, Yao Z, Cui Y, Wei B, Jin Z, Xu X. Evaluation of ChatGPT-generated medical responses: a systematic review and meta-analysis. J Biomed Inform 2024 Mar;151:104620. [doi: 10.1016/j.jbi.2024.104620] [Medline: 38462064]
20. Huh CY, Lee J, Kim G, et al. Using large language models to simulate history taking: implications for symptom-based medical education. Information 2025 Jul 31;16(8):653. [doi: 10.3390/info16080653]
21. Huang TY, Hsieh PH, Chang YC. Performance comparison of junior residents and ChatGPT in the Objective Structured Clinical Examination (OSCE) for medical history taking and documentation of medical records: development and usability study. JMIR Med Educ 2024 Nov 21;10:e59902. [doi: 10.2196/59902] [Medline: 39622713]
22. Burke HB, Hoang A, Lopreiato JO, et al. Assessing the ability of a large language model to score free-text medical student clinical notes: quantitative study. JMIR Med Educ 2024 Jul 25;10:e56342. [doi: 10.2196/56342] [Medline: 39118469]
23. Jamieson AR, Holcomb MJ, Dalton TO, et al. Rubrics to prompts: assessing medical student post-encounter notes with AI. NEJM AI 2024 Nov 27;1(12). [doi: 10.1056/AIcs2400631]
24. Maitin AM, Nogales A, Fernández-Rincón S, et al. Application of large language models in clinical record correction: a comprehensive study on various retraining methods. J Am Med Inform Assoc 2025 Feb 1;32(2):341-348. [doi: 10.1093/jamia/ocae302] [Medline: 39707579]
25. Liu B, Wang A, Xia C. Interpretable Chinese fake news detection with chain-of-thought and in-context learning. IEEE Access 2025;13:117186-117197. [doi: 10.1109/ACCESS.2025.3571497]
26. Cervera-Barba E, Denizon-Arranz S, Mateos-Rodríguez A, Neria-Serrano F. Concordancia y calidad de las historias clínicas en los alumnos de Medicina: ¿reflejan la realidad de la consulta? [Article in Spanish]. Educación Médica 2024 May;25(3):100907. [doi: 10.1016/j.edumed.2024.100907]
27. Llama-3.1-70B-instruct. Hugging Face. URL: https://huggingface.co/meta-llama/Llama-3.1-70B-Instruct [accessed 2026-06-27]
28. Claude 3.5 sonnet. Anthropic. URL: https://www.anthropic.com/news/claude-3-5-sonnet [accessed 2026-06-27]
29. Ley 14/2007, de 3 de julio, de investigación biomédica [article in spanish]. Agencia Estatal Boletín Oficial del Estado. URL: https://www.boe.es/eli/es/l/2007/07/03/14 [accessed 2026-07-02]
30. Ley orgánica 3/2018, de 5 de diciembre, de protección de datos personales y garantía de los derechos digitales. Agencia Estatal Boletín Oficial del Estado. URL: https://www.boe.es/eli/es/lo/2018/12/05/3/con [accessed 2026-07-02]
31. Ibrahim A, Hosseini A, Ibrahim S, Sattar A, Serag A. D3: a small language model for drug-drug interaction prediction and comparison with large language models. Machine Learning with Applications 2025 Jun;20:100658. [doi: 10.1016/j.mlwa.2025.100658]
32. Bae J, Kwon S, Myeong S. Enhancing software code vulnerability detection using GPT-4o and Claude-3.5 Sonnet: a study on prompt engineering techniques. Electronics (Basel) 2024 Jul 6;13(13):2657. [doi: 10.3390/electronics13132657]
33. Brown T, Mann B, Ryder N, et al. Language models are few-shot learners. Presented at: 34th International Conference on Neural Information Processing Systems; Dec 6-12, 2020. [doi: 10.5555/3495724.3495883]
34. Wei J, Wang X, Schuurmans D, et al. Chain-of-thought prompting elicits reasoning in large language models. Presented at: 36th International Conference on Neural Information Processing Systems; Nov 28 to Dec 9, 2022. [doi: 10.5555/3600270.3602070]
35. Luordo D, Torres Arrese M, Tristán Calvo C, et al. Application of artificial intelligence as an aid for the correction of the Objective Structured Clinical Examination (OSCE). Applied Sciences 2025;15(3):1153. [doi: 10.3390/app15031153]
36. Thomas K, Szalacha L, Hanna K, Anibal J, Petrilli J. Evaluating the effectiveness of ChatGPT versus human proctors in grading medical students' post-OSCE notes. Fam Med 2025 Nov;57(10):727-731. [doi: 10.22454/FamMed.2025.954255] [Medline: 41854706]
37. Key issues: human oversight. EU AI Act. URL: https://www.euaiact.com/key-issue/4 [accessed 2026-07-02]
38. Agrawal LA, Tan S, Soylu D, et al. GEPA: reflective prompt evolution can outperform reinforcement learning. arXiv. Preprint posted online on Feb 14, 2026 URL: https://arxiv.org/abs/2507.19457 [accessed 2026-07-17]

Abbreviations

AI: artificial intelligence
API: application programming interface
CoT: chain-of-thought
FN: false negative
FP: false positive
ICC: intraclass correlation coefficient
LLM: large language model
LoRA: low-rank adaptation
OSCE: Objective Structured Clinical Examination
TN: true negative
TP: true positive

XSL•FO
RenderX

Edited by B Lesselroth; submitted 28.Nov.2025; peer-reviewed by D Yang, E Emekli; revised version received 14.Apr.2026; accepted 21.Apr.2026; published 23.Jul.2026.

Please cite as:
Nogales A, Denizon S, Mateos Rodriguez A, Cervera Cordero J, Pandelet Barainca G, Aranguren Moliner E, Cervera Barba E Benefits of Chain-of-Thought Prompting for Clinical Record Rubric Evaluation in Undergraduate Medical Education: Experimental Evaluation Study With Medical Faculty
JMIR Med Educ 2026;12:e88652
URL: https://mededu.jmir.org/2026/1/e88652
doi:10.2196/88652

© Alberto Nogales, Sophia Denizon, Alonso Mateos Rodriguez, Javier Cervera Cordero, Gonzalo Pandelet Barainca, Enrique Aranguren Moliner, Emilio Cervera Barba. Originally published in JMIR Medical Education (https://mededu.jmir.org), 23.Jul.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Education, is properly cited. The complete bibliographic information, a link to the original publication on https://mededu.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Use, Concerns, and Perspectives on AI in Health Care Among French Health Professionals and Students: Web-Based Cross-Sectional Survey

Aurelia Alati[1,2], MSc, MD; Grégoire Pigné[3,4], MD; Carole-Anne Brugère[4], MSc, PharmD; Jean-Emmanuel Bibault[1,2], MD, PhD

[1]Department of Radiation Oncology, Université Paris Cité, Hôpital Européen Georges-Pompidou, 20 rue Leblanc, Paris, Île-de-France, France

[2]Inserm UMRS 1138, Centre de Recherche des Cordeliers, Paris, Île-de-France, France

[3]Department of Radiation Oncology, Institut de Cancérologie et d'Hématologie Universitaire de Saint-Étienne, Saint-Priest-en-Jarez, Saint-Etienne, France

[4]PulseLife, Lyon, France

Corresponding Author:
Aurelia Alati, MSc, MD
Department of Radiation Oncology, Université Paris Cité, Hôpital Européen Georges-Pompidou, 20 rue Leblanc, Paris, Île-de-France, France

Abstract

Background: AI is increasingly discussed and deployed in health care, yet safe and effective implementation depends on the preparedness, trust, and training of the professionals who are expected to use these tools.

Objective: This study aimed to assess current AI use, perceived benefits and concerns, confidence, and training needs among French health care professionals and students.

Methods: We conducted a national web-based cross-sectional survey distributed through the PulseLife professional community between December 4, 2024, and March 5, 2025. The survey instrument was administered in French and included respondent characteristic items together with 12 substantive closed-ended questions covering current AI use, confidence, perceived benefits, and concerns, and interest in AI-related training. Access was restricted to authenticated individual PulseLife accounts, and multiple submissions from the same account were not allowed. Questions were not mandatory; incomplete questionnaires were retained for item-level analyses, and percentages were calculated using item-specific denominators. Because the exact invitation denominator was not retained by the platform, view, participation, and completion rates could not be calculated. Descriptive statistics and Pearson chi-square tests were performed using R. Internal consistency and exploratory psychometric properties were assessed using the Cronbach α, exploratory factor analysis, and confirmatory factor analysis.

Results: A total of 1625 respondents participated, including 1212 (74.6%) health professionals and 413 (25.4%) students. Among professionals, physicians represented the largest group (642/1212, 53%), followed by nurses (232/1212, 19.1%) and pharmacists (92/1212, 7.6%). Only 6.6% (90/1366) of the respondents reported prior AI-specific training, whereas 78.3% (920/1175) wished to receive such training. Confidence in AI for diagnosis and patient management remained limited: only 9.2% (120/1301) of the respondents reported being very confident. Nearly half (673/1455, 46.3%) of the respondents who answered this item reported no current AI use in professional activity, whereas 10.5% (153/1455) reported frequent use. Physicians and younger respondents reported more frequent AI use, and prior AI training was associated with greater confidence (P<.001 in all cases). Commonly perceived benefits included improved diagnosis (774/1625, 47.6%), time savings (685/1625, 42.2%), reduced medical errors (634/1625, 39%), and improved patient follow-up (593/1625, 36.5%). Frequently reported concerns included algorithmic bias (785/1625, 48.3%), limited transparency (666/1625, 41%), deterioration of the patient–health care professional relationship (628/1625, 38.6%), and data confidentiality (557/1625, 34.3%).

Conclusions: In this national French sample, formal AI training was uncommon despite high interest in receiving it. These findings support the need for more structured educational initiatives in AI literacy across undergraduate, postgraduate, and continuing professional education. Because this study relied on a convenience sample recruited through a digital platform, the findings should be interpreted as descriptive and exploratory rather than nationally representative.

(JMIR Med Educ 2026;12:e89152) doi:10.2196/89152

KEYWORDS

artificial intelligence; AI; health care professionals; students; medical education; digital health; survey; training needs

XSL•FO
RenderX

Introduction

AI is increasingly being integrated into health care, with applications ranging from clinical decision support and imaging analysis to workflow optimization, administrative automation, and patient monitoring. Safe implementation depends not only on technical performance but also on the preparedness, trust, and training of the professionals expected to use these tools [1,2]. In France, AI has been identified as a strategic priority for health innovation, yet the practical readiness of frontline professionals and students remains insufficiently documented [1].

International studies suggest that interest in AI frequently exceeds formal preparation. In a web-based survey of 293 physicians and medical students, AlZaabi et al [3] found generally positive attitudes toward AI but also a clear need for education, with 69% of participants reporting that AI was an emerging field they wished to learn more about. In Spain, Catalina et al [4] reported that implementation in primary care depended not only on technical feasibility but also on professionals' knowledge and acceptance of AI. In Germany, Laupichler et al [5] showed among 377 medical students that prior AI education and interest in AI were both associated with higher AI literacy. More recently, a Saudi Arabian survey of 1221 medical and health sciences students found moderate AI readiness overall, and 44.5% of respondents believed that AI-related courses should be mandatory [6].

The educational implications of these findings are increasingly emphasized in the literature. A 2024 scoping review identified 21 eligible papers describing 30 educational programs and only 2 curriculum frameworks for AI across medical students, residents, and practicing physicians, underscoring how fragmented the field remains [7]. Stakeholder and curriculum studies have further emphasized that AI competencies should extend beyond technical understanding to include ethical reasoning, implementation awareness, and critical appraisal [8-10]. Against this background, the objective of this study was to assess current use of AI, perceived benefits and concerns, confidence, and training needs among French health care professionals and students, with direct relevance for curriculum design, continuing professional development, and the responsible integration of AI into clinical practice.

Methods

Study Design and Setting

We conducted a national web-based cross-sectional survey of health care professionals and students in France. The survey was disseminated through PulseLife, a digital platform and professional community used by health care workers in France. Recruitment was conducted via email between December 4, 2024, and March 5, 2025.

No financial incentive was offered. The exact number of invited individuals, page view rates, participation rates, and completion rates was not retained by the platform in a form that allowed for retrospective reconstruction; accordingly, a formal response rate could not be calculated. This study should therefore be interpreted as a convenience-sample survey rather than a probability-based national survey.

Participants, Eligibility, and Survey Administration

Eligible participants were health care professionals directly involved in care delivery and health profession students training in France. The survey was accessed through authenticated individual PulseLife accounts. Each account could submit the questionnaire only once, and multiple submissions from the same account were not allowed.

The survey instrument included demographic and professional characteristic items together with 12 substantive closed-ended questions and was administered in French. Questions were not technically mandatory, and respondents could skip items. The archived platform configuration did not retain sufficiently detailed information to reconstruct whether back navigation to earlier items was enabled. Incomplete questionnaires were retained in the analytic dataset; all percentages were therefore calculated using the number of respondents to each item as the denominator unless otherwise specified. For multi-selection items, totals could exceed 100%.

Study Objectives

The primary objective was to assess use, concerns, and perspectives regarding AI technologies among French health care professionals and students. Secondary objectives were to explore associations between respondent characteristics and AI use or confidence and identify training needs with direct relevance to medical and health professions education.

Questionnaire Development and Psychometric Assessment

The questionnaire covered respondent characteristics and AI use, confidence, perceived benefits, perceived concerns, and training interest (Multimedia Appendix 1). In this survey, "AI-based solutions" was used as an umbrella term encompassing, for example, decision support systems, imaging analysis tools, administrative automation, and generative AI–enabled information tools. The questionnaire was developed through iterative multidisciplinary review by clinicians involved in AI-related clinical practice and digital health implementation. The development process aimed to ensure face validity and practical relevance for health care professionals and students in France. Before dissemination, the questionnaire was pilot-tested with 20 health care professionals and students to assess clarity, wording, and feasibility. This pilot stage was intended as a face validity and usability check rather than formal stand-alone validation.

Because the instrument was created for this study, we report its psychometric properties as exploratory support rather than definitive external validation. Internal consistency across domains was acceptable (Cronbach α=0.76-0.84). Sampling adequacy was supported by a Kaiser-Meyer-Olkin index of 0.83, and the Bartlett test of sphericity was significant (P<.001). Exploratory factor analysis supported a 3-factor structure (concerns, perceived benefits, and assisted tasks), a Confirmatory factor analysis of the 14-item, 3-factor model indicated acceptable fit (χ^2/df=2.1; df=74; comparative fit

XSL•FO
RenderX

index=0.94; Tucker-Lewis index=0.92; root mean square error of approximation=0.05; standardized root mean square residual=0.06). Known group validity was supported by higher AI confidence among physicians, younger respondents, and participants with prior AI training. More detailed factor analytic parameters, including extraction method, rotation strategy, factor retention criteria, item retention thresholds, and confirmatory factor analysis estimator specifications, were not retained in a form that allowed for full retrospective reporting. Similarly, domain scores were interpreted at the construct level, but a formal scoring algorithm for external reuse was not predefined because the instrument was designed for this exploratory survey. Because exploratory and confirmatory analyses were conducted on the same overall dataset, these findings should be interpreted cautiously and are further discussed as a limitation. Additional psychometric details are provided in Multimedia Appendix 1.

Statistical Analysis

Descriptive statistics are reported as counts and percentages. Associations between categorical variables were explored using Pearson chi-square tests. No formal correction for multiple testing was applied because the inferential analyses were exploratory. This approach increases the risk of type I error, and reported associations should therefore be interpreted as hypothesis generating. Effect sizes such as the Cramér *V* were not available from the archived analytical outputs in a form that allowed for reliable retrospective reporting across all comparisons. For this reason, we retained a conservative presentation of the inferential analyses and limited their interpretation accordingly. Analyses were conducted using R (version 4.3.1; R Foundation for Statistical Computing). No weighting was applied because sampling probabilities and suitable calibration margins were unavailable.

Ethical Considerations

This study was conducted in accordance with the Declaration of Helsinki and the French legal framework applicable to health research and personal data protection. The protocol was reviewed and approved by the local ethics committee of Hôpital Européen Georges Pompidou, Paris, France; however, no formal approval number was assigned. Participation was voluntary, and completion of the online questionnaire was considered to constitute informed consent. The dataset used was deidentified before analysis. No directly identifying information was collected in the study dataset used by the academic investigators. Participants received no financial compensation.

Results

A total of 1625 respondents were included in the survey, including 1212 (74.6%) health professionals and 413 (25.4%) students. Among professionals, physicians represented the largest group (642/1212, 53%), followed by nurses (232/1212, 19.1%) and pharmacists (92/1212, 7.6%). Among students, medical students represented 59.1% (244/413) of the respondents, followed by nursing students (69/413, 16.7%). Most professionals reported working in hospitals (429/1212, 35.4%), private practice (308/1212, 25.4%), or private clinics (123/1212, 10.1%; Table 1).

Table . Population characteristics.

Characteristic	Participants, n (%)
Health professionals (n=1212)	
Physicians	642 (53)
Nurses	232 (19.1)
Pharmacists	92 (7.6)
Other or unspecified professionals	246 (20.3)
Students (n=413)	
Medical students	244 (59.1)
Nursing students	69 (16.7)
Pharmacy students	25 (6.1)
Other or unspecified students	75 (18.2)
Age (y; n=1499)	
<25	219 (14.6)
25-34	432 (28.8)
35-44	293 (19.5)
45-54	216 (14.4)
≥55	339 (22.6)

Formal exposure to AI was uncommon. Of the 1366 respondents who answered the item on prior training, only 90 (6.6%) reported having received AI-specific training, whereas 1276 (93.4%) had not. Of the 1175 respondents who answered the training interest item, 920 (78.3%) expressed a desire to receive specific training in AI. Confidence remained limited: of the 1301 respondents who answered the confidence item, only 120 (9.2%) reported being very confident in using AI for diagnostic

XSL•FO
RenderX

and patient management support, whereas 407 (31.3%) were not very confident, and 118 (9.1%) were not at all confident (Table 2).

Table . Training, confidence, and current AI use[a].

Outcome	Participants, n (%)
Prior AI-specific training (n=1366)	90 (6.6)
Interested in receiving AI training (n=1175)	920 (78.3)
Confidence level in AI use for diagnosis or patient management (n=1301)	
Very confident	120 (9.2)
Somewhat confident	656 (50.4)
Not very confident	407 (31.3)
Not at all confident	118 (9.1)
AI use in professional activity (n=1455)	
Frequent	153 (10.5)
No current use	673 (46.3)

[a]Denominators vary because the questions were not mandatory; incomplete questionnaires were retained for item-level analyses.

Current use of AI in professional activity was heterogeneous. Of the 1455 respondents who answered the item on use frequency, 673 (46.3%) reported not using AI at all, 297 (20.4%) reported rare use, 311 (21.4%) reported occasional use, 153 (10.5%) reported frequent use, and 21 (1.4%) reported intensive use. Among respondents who reported some AI use and specified contexts of use, the most commonly selected contexts were access to medical information (n=334), help with training (n=277), treatment prescription support (n=203), medical diagnosis (n=201), and medical imaging analysis (n=198).

Perceived benefits of AI included improved diagnosis (774/1625, 47.6%), time savings in care delivery (685/1625, 42.2%), reduction in medical errors (634/1625, 39%), and improved patient follow-up (593/1625, 36.5%). Frequently cited concerns included algorithmic bias (785/1625, 48.3%), transparency and reliability of sources (666/1625, 41%), possible deterioration of the patient–health care professional relationship (628/1625, 38.6%), and data confidentiality (557/1625, 34.3%). Of the 1301 respondents who answered the item on professional threat, 423 (32.5%) considered AI a possible threat to their profession or skills (Table 3).

XSL•FO
RenderX

Table . Selected reported AI uses, benefits, and concerns (N=1625).

Item	Participants
Contexts of AI use among respondents who reported some AI use and answered this item, n[a]	
Help with access to medical information	334
Help with training	277
Treatment prescription support	203
Medical diagnosis	201
Medical imaging analysis	198
Patient monitoring	117
Expected benefits, n (%)	
Improved diagnosis	774 (47.6)
Time savings in care delivery	685 (42.2)
Reduction in medical errors	634 (39)
Improved patient follow-up	593 (36.5)
Main concerns, n (%)	
Algorithmic bias	785 (48.3)
Limited transparency and reliability of sources	666 (41)
Deterioration of patient–health care professional relationship	628 (38.6)
Data confidentiality	557 (34.3)
Perceived professional threat (n=1301), n (%)	
AI perceived as a threat to their profession or skills	423 (32.5)

[a]Percentages are not shown for context of use because the item was multi-selection among respondents who reported some AI use.

In exploratory unadjusted analyses, younger age was associated with greater interest in AI training (*P*=.02) and more frequent AI use (*P*<.001). Prior AI training was associated with greater confidence in AI for diagnosis and patient management (*P*=.001). AI use also differed significantly between students and professionals (*P*=.001), and confidence differed across professional groups (*P*=.003), with physicians among the groups reporting the highest confidence levels in the profession-specific comparison (Table 4).

Table . Exploratory unadjusted associations[a].

Outcome	Comparison	*P* value	Interpretation
Interest in AI training	Age group	.02	Younger respondents more frequently reported interest in training
Prior AI training	Age group	.31	No significant association
Prior AI training	Student vs professional status	.12	No significant association
Confidence in AI	Prior AI training	.001	Higher confidence among respondents with prior AI training
Current AI use	Age group	<.001	More frequent AI use among younger respondents
Current AI use	Student vs professional status	.001	Distribution of use differed between students and professionals
Confidence in AI	Professional group	.003	Profession-specific confidence differed; physicians were among the most confident groups
Perceived AI as threat	Professional group	.01	Perception of threat differed across professional groups

[a]All inferential analyses were exploratory, based on Pearson chi-square tests, and unadjusted for potential confounders.

XSL•FO
RenderX

Discussion

Principal Findings

This national survey found a marked mismatch between the rapid diffusion of AI discourse in health care and the limited formal preparation reported by French health professionals and students. Formal AI-specific training was uncommon, whereas interest in receiving such training was high. Confidence in AI for diagnosis and patient management support remained limited, and current use in professional activity was heterogeneous rather than routine.

When AI was used, it was more often used for supportive or informational purposes than for core diagnostic or therapeutic decision-making. Respondents expected AI to improve diagnosis, save time, reduce errors, and support follow-up, but they also expressed concerns about algorithmic bias, transparency, confidentiality, and the patient–health care professional relationship. These findings suggest that respondents were not opposed to AI in principle but, rather, were conditionally receptive provided that AI remained trustworthy, interpretable, and compatible with safe, human-centered care.

Comparison With Previous Literature

Earlier studies cited above have already shown that interest in AI frequently exceeds formal preparation across different health care and educational settings [3-6]. The more recent literature further reinforces this pattern and provides additional points of comparison with our French cross-professional sample. Rjoop et al [11] surveyed 394 medical students and pathology trainees in Jordan and found that 66.2% disagreed that their medical schools had educated them about AI, whereas 46.2% expressed interest in learning about AI in medicine; the same study also showed substantial agreement with pathology-specific AI applications, illustrating how perceptions may be shaped by specialty exposure. Syeda et al [12] surveyed 939 undergraduate health care students in Pakistan and reported that only 11.8% had received formal AI training, whereas 78.8% supported AI integration into curricula and 82.2% endorsed AI training as part of medical education. A large international survey of 4596 medical, dental, and veterinary students from 192 faculties in 48 countries similarly found limited AI knowledge, lack of AI courses, and strong desire for more AI teaching [13]. These results are closely aligned with the French pattern observed in our survey, particularly the coexistence of limited formal training and strong educational demand. French data are more limited but point in the same direction. Perrier et al [14] conducted a nationwide survey of 165 young French pediatricians and found favorable attitudes toward AI, but only 5% had received specific AI training, whereas 87% considered AI training necessary. Kotzki et al [15] recently surveyed 388 French health professional students during clinical placements and found that more than half reported using generative AI, mainly for information and documentation support rather than patient-facing activities. Earlier qualitative work by Laï et al [16] based on 40 French stakeholders also suggested that French health professionals were not simply resistant to AI but focused on safe, clinically useful, and responsible implementation. Compared with these French studies, our survey adds a broader cross-professional perspective by including both practicing health professionals and students and assessing use, confidence, perceived benefits, concerns, and training needs within the same national dataset. The broader health service literature also supports the need for this transversal perspective. A 2025 systematic review of 72 empirical studies of health care professionals' perspectives on AI identified multiple facilitating and hindering factors across individual, interpersonal, institutional, community, and policy levels [17]. The review also noted that many studies focused on radiology or general medicine, limiting the ability to generalize to other fields [17]. Our study therefore complements rather than duplicates the existing literature: it provides a broader national view that is less likely to be driven by the technological culture or implementation history of a single specialty. Specialty-specific European studies further illustrate this issue: the 2024 European Network for the Assessment of Imaging in Medicine and European Society of Medical Imaging Informatics survey found that nearly half of the radiology respondents were already using AI tools, and radiotherapy implementation work has evaluated commercial AI-based auto-contouring systems in clinical practice [18,19].

Implications for Medical Education and Continuing Professional Development

The clearest educational signal emerging from our survey was the marked gap between the low proportion of respondents who had received formal AI-specific training (90/1366, 6.6%) and the high proportion who wished to receive such training (920/1175, 78.3%). This finding is consistent with the broader international mismatch between the rapid diffusion of AI and insufficient educational preparation. Prior AI-specific training was rare in our study, but interest in training was high. This training gap is highly relevant to medical and health professions education because confidence was greater among those who reported prior AI training. These findings support the integration of structured AI-related competencies across undergraduate, postgraduate, and continuing professional education [7,20,21].

Recent curriculum work suggests that AI training should not be limited to isolated technical exposure. Tolentino et al [7] identified only 2 explicit curriculum frameworks among 21 included papers in their scoping review. Moldt et al [9] further showed that stakeholders expect AI education to include technical knowledge, ethical reasoning, implementation awareness, and clinically meaningful use. Moldt et al [9] recently proposed a postgraduate family medicine AI framework, reinforcing the need for longitudinal, competency-based approaches to AI training [10]. In practical terms, our findings suggest that training should address not only how to use AI tools but also why and when they should be used, how to calibrate trust, and how to recognize situations in which AI may be misleading or clinically unsafe [22,23].

Ethical, Legal, and Organizational Considerations

The concerns reported by respondents reinforce the importance of framing AI education within broader ethical, legal, and organizational contexts. Concerns about bias, transparency, confidentiality, and the patient–health care professional

XSL•FO
RenderX

relationship were prominent in our sample. These concerns are concordant with wider literature showing that bias, explainability, privacy, and implementation barriers are central determinants of trust in health AI [17,24-29].

The organizational implications are also important. Respondents expected AI to support efficiency and safety, yet some also perceived it as a potential threat to professional roles or skills. This tension is understandable because the introduction of new technologies may initially disrupt workflows, increase cognitive load, and create additional work during transition phases, particularly when systems are poorly integrated [26,30,31]. At the same time, AI has been described as having the potential to alleviate part of the administrative and cognitive burden that contributes to professional exhaustion and burnout, thereby helping preserve time for complex decision-making and human interaction [30,32,33]. Our data therefore support a balanced interpretation: clinicians and students expect AI to be helpful but only if it remains interpretable, accountable, and compatible with human-centered care.

Strengths and Limitations

This study has several strengths. With 1625 respondents, including 1212 (74.6%) practicing health professionals and 413 (25.4%) students, it provides a large French dataset for assessing AI use, concerns, confidence, and training needs across both groups. To our knowledge, it is among the broadest French cross-professional surveys to date on this topic. The questionnaire underwent pilot-testing and exploratory psychometric assessment. A key strength of the present study is its broad and cross-professional sampling frame. Unlike studies focused on a single AI-exposed specialty or one educational subgroup, our survey was designed to capture a wider range of professional and educational backgrounds. This is important because perceptions of AI may be shaped by prior exposure; specialty-specific workflows; and the technological culture of fields such as radiology, radiotherapy, pathology, or other highly digitized disciplines [11,17-19]. By including both practicing health professionals and students, the present study provides a broader national perspective on AI use, concerns, confidence, and training needs in France. This makes the findings particularly relevant for general AI literacy, curriculum design, postgraduate training, and continuing professional development.

Several limitations deserve emphasis. First, this study relied on a convenience sample recruited through a digital professional platform, which may overrepresent respondents already interested in technology. As a result, the findings should not be interpreted as nationally representative. Second, because the exact number of invited platform users was unavailable, a formal response rate could not be calculated. No weighting strategy could be implemented for the same reason. Third, item-level missingness varied across questions; we addressed this by reporting item-specific denominators, but the resulting estimates remain subject to nonresponse bias. Fourth, no formal correction for multiple comparisons was applied because the inferential analyses were exploratory; this increases the risk of type I error and reinforces the need to interpret the reported associations cautiously. Fifth, the psychometric analyses should be interpreted cautiously because exploratory and confirmatory factor analyses were performed on the same overall dataset. Finally, the inferential analyses were unadjusted and exploratory; this study was not designed to support causal interpretation.

Conclusions

In this national French convenience sample, formal AI training was uncommon despite strong interest in receiving it, and confidence in AI for diagnostic support remained limited. These findings support the relevance of structured AI education across undergraduate, postgraduate, and continuing professional training. They also highlight the importance of integrating ethical, legal, and practical implementation issues into such teaching.

Given the descriptive design and sampling approach, the results should be interpreted as exploratory rather than nationally representative. Future work should include longitudinal and mixed methods studies, particularly studies evaluating how training interventions influence AI literacy, trust calibration, and real-world adoption in clinical environments.

Acknowledgments

Microsoft Editor and Microsoft Copilot (Microsoft Corporation) were used for grammar, style, and clarity checks during manuscript revision. They were not used to generate study data, perform statistical analyses, interpret results, or determine scientific conclusions. All suggested edits were critically reviewed, verified, and approved by the authors, who take full responsibility for the final content.

Funding

The authors declared no financial support was received for this work.

Data Availability

Deidentified aggregate data and the analysis plan are available from the corresponding author on reasonable request subject to platform governance and applicable data protection constraints.

Authors' Contributions

Conceptualization: AA, JEB, GP

XSL•FO
RenderX

Data curation: AA, GP, CAB
Formal analysis: AA
Methodology: AA
Resources: GP, CAB
Supervision: JEB
Writing—original draft: AA
Writing—review and editing: JEB, GP, CAB
AA drafted the original manuscript. AA and JEB led the interpretation of the results, and JEB contributed substantially to manuscript revision.

Conflicts of Interest

GP and CAB are affiliated with PulseLife, the digital platform through which the survey was disseminated. PulseLife was used solely as a practical dissemination channel to reach a broad range of health care professionals and facilitate data extraction. The present study did not evaluate the PulseLife platform itself, did not assess any PulseLife-specific medical applications, and does not define the field of use of the platform. All other authors declare no other conflicts of interest.

Multimedia Appendix 1
Artificial Intelligence in Health Care Questionnaire in English and French, with questionnaire development details and exploratory psychometric assessment.
[DOCX File, 34 KB - mededu_v12i1e89152_app1.docx]

References

1. Artificial intelligence and health [Article in French]. Inserm. 2024. URL: https://www.inserm.fr/dossier/intelligence-artificielle-et-sante/ [accessed 2026-07-24]
2. Sharma M, Savage C, Nair M, Larsson I, Svedberg P, Nygren JM. Artificial intelligence applications in health care practice: scoping review. J Med Internet Res 2022 Oct 5;24(10):e40238. [doi: 10.2196/40238] [Medline: 36197712]
3. AlZaabi A, AlMaskari S, AalAbdulsalam A. Are physicians and medical students ready for artificial intelligence applications in healthcare? Digit Health 2023;9:20552076231152167. [doi: 10.1177/20552076231152167] [Medline: 36762024]
4. Catalina QM, Fuster-Casanovas A, Vidal-Alaball J, et al. Knowledge and perception of primary care healthcare professionals on the use of artificial intelligence as a healthcare tool. Digit Health 2023;9:20552076231180511. [doi: 10.1177/20552076231180511] [Medline: 37361442]
5. Laupichler MC, Aster A, Meyerheim M, Raupach T, Mergen M. Medical students' AI literacy and attitudes towards AI: a cross-sectional two-center study using pre-validated assessment instruments. BMC Med Educ 2024 Apr 10;24(1):401. [doi: 10.1186/s12909-024-05400-7] [Medline: 38600457]
6. Almalki M, Alkhamis MA, Khairallah FM, Choukou MA. Perceived artificial intelligence readiness in medical and health sciences education: a survey study of students in Saudi Arabia. BMC Med Educ 2025 Mar 26;25(1):439. [doi: 10.1186/s12909-025-06995-1] [Medline: 40140763]
7. Tolentino R, Baradaran A, Gore G, Pluye P, Abbasgholizadeh-Rahimi S. Curriculum frameworks and educational programs in AI for medical students, residents, and practicing physicians: scoping review. JMIR Med Educ 2024 Jul 18;10:e54793. [doi: 10.2196/54793] [Medline: 39023999]
8. Grunhut J, Marques O, Wyatt AT. Needs, challenges, and applications of artificial intelligence in medical education curriculum. JMIR Med Educ 2022 Jun 7;8(2):e35587. [doi: 10.2196/35587] [Medline: 35671077]
9. Moldt JA, Festl-Wietek T, Fuhl W, et al. Assessing AI awareness and identifying essential competencies: insights from key stakeholders in integrating AI into medical education. JMIR Med Educ 2024 Jun 12;10:e58355. [doi: 10.2196/58355] [Medline: 38989834]
10. Tolentino R, Hersson-Edery F, Yaffe M, Abbasgholizadeh-Rahimi S. AIFM-ed curriculum framework for postgraduate family medicine education on artificial intelligence: mixed methods study. JMIR Med Educ 2025 Apr 25;11:e66828. [doi: 10.2196/66828] [Medline: 40279148]
11. Rjoop A, Al-Qudah M, Alkhasawneh R, et al. Awareness and attitude toward artificial intelligence among medical students and pathology trainees: survey study. JMIR Med Educ 2025 Jan 10;11:e62669. [doi: 10.2196/62669] [Medline: 39803949]
12. Syeda LH, Batool Z, Hayder Z, Ali S. Medical undergraduate students' awareness and perspectives on artificial intelligence: a developing nation's context. BMC Med Educ 2025 Jul 15;25(1):1060. [doi: 10.1186/s12909-025-07223-6] [Medline: 40665264]
13. Busch F, Hoffmann L, Truhn D, et al. Global cross-sectional student survey on AI in medical, dental, and veterinary education and practice at 192 faculties. BMC Med Educ 2024 Sep 28;24(1):1066. [doi: 10.1186/s12909-024-06035-4] [Medline: 39342231]

XSL•FO
RenderX

14. Perrier E, Rifai M, Terzic A, Dubois C, Cohen JF. Knowledge, attitudes, and practices towards artificial intelligence among young pediatricians: a nationwide survey in France. Front Pediatr 2022;10:1065957. [doi: 10.3389/fped.2022.1065957] [Medline: 36619510]
15. Kotzki S, Massonnet Turner C, Gauthier K, Minoves M, Vuillerme N. Health professional students' use of generative artificial intelligence during clinical placements: cross-sectional online survey study. JMIR Med Educ 2026 Apr 27;12:e85243. [doi: 10.2196/85243] [Medline: 42044362]
16. Laï MC, Brian M, Mamzer MF. Perceptions of artificial intelligence in healthcare: findings from a qualitative survey study among actors in France. J Transl Med 2020 Jan 9;18(1):14. [doi: 10.1186/s12967-019-02204-y] [Medline: 31918710]
17. Henzler D, Schmidt S, Koçar A, et al. Healthcare professionals' perspectives on artificial intelligence in patient care: a systematic review of hindering and facilitating factors on different levels. BMC Health Serv Res 2025 May 1;25(1):633. [doi: 10.1186/s12913-025-12664-2] [Medline: 40312413]
18. Zanardo M, Visser JJ, Colarieti A, et al. Impact of AI on radiology: a EuroAIM/EuSoMII 2024 survey among members of the European Society of Radiology. Insights Imaging 2024 Oct 7;15(1):240. [doi: 10.1186/s13244-024-01801-w] [Medline: 39373853]
19. Meyer C, Huger S, Bruand M, et al. Artificial intelligence contouring in radiotherapy for organs-at-risk and lymph node areas. Radiat Oncol 2024 Nov 21;19(1):168. [doi: 10.1186/s13014-024-02554-y] [Medline: 39574153]
20. Paranjape K, Schinkel M, Nannan Panday R, Car J, Nanayakkara P. Introducing artificial intelligence training in medical education. JMIR Med Educ 2019 Dec 3;5(2):e16048. [doi: 10.2196/16048] [Medline: 31793895]
21. Sapci AH, Sapci HA. Artificial intelligence education and tools for medical and health informatics students: systematic review. JMIR Med Educ 2020 Jun 30;6(1):e19285. [doi: 10.2196/19285] [Medline: 32602844]
22. Tomsett R, Preece A, Braines D, et al. Rapid trust calibration through interpretable and uncertainty-aware AI. Patterns (N Y) 2020 Jul 10;1(4):100049. [doi: 10.1016/j.patter.2020.100049] [Medline: 33205113]
23. Lomis K, Jeffries P, Palatta A, et al. Artificial intelligence for health professions educators. NAM Perspect 2021;2021. [doi: 10.31478/202109a] [Medline: 34901780]
24. Panch T, Mattie H, Atun R. Artificial intelligence and algorithmic bias: implications for health systems. J Glob Health 2019 Dec;9(2):010318. [doi: 10.7189/jogh.09.020318] [Medline: 31788229]
25. O'Connor S, Liu H. Gender bias perpetuation and mitigation in AI technologies: challenges and opportunities. AI Soc 2024 Aug;39:2045-2057. [doi: 10.1007/s00146-023-01675-4]
26. Ahmed MI, Spooner B, Isherwood J, Lane M, Orrock E, Dennison A. A systematic review of the barriers to the implementation of artificial intelligence in healthcare. Cureus 2023 Oct;15(10):e46454. [doi: 10.7759/cureus.46454] [Medline: 37927664]
27. Balasubramaniam N, Kauppinen M, Rannisto A, Hiekkanen K, Kujala S. Transparency and explainability of AI systems: from ethical guidelines to requirements. Inf Softw Technol 2023 Jul;159:107197. [doi: 10.1016/j.infsof.2023.107197]
28. Ali S, Abuhmed T, El-Sappagh S, et al. Explainable artificial intelligence (XAI): what we know and what is left to attain trustworthy artificial intelligence. Inf Fusion 2023 Nov;99:101805. [doi: 10.1016/j.inffus.2023.101805]
29. AI: consider data protection in system design [Article in French]. CNIL. 2024. URL: https://www.cnil.fr/fr/tenir-compte-de-la-protection-des-donnees-dans-la-conception-du-systeme [accessed 2026-07-24]
30. Hazarika I. Artificial intelligence: opportunities and implications for the health workforce. Int Health 2020 Jul 1;12(4):241-245. [doi: 10.1093/inthealth/ihaa007] [Medline: 32300794]
31. Lugtenberg M, Weenink JW, van der Weijden T, Westert GP, Kool RB. Implementation of multiple-domain covering computerized decision support systems in primary care: a focus group study on perceived barriers. BMC Med Inform Decis Mak 2015 Oct 12;15:82. [doi: 10.1186/s12911-015-0205-z] [Medline: 26459233]
32. Shanafelt TD, Boone S, Tan L, et al. Burnout and satisfaction with work-life balance among US physicians relative to the general US population. Arch Intern Med 2012 Oct 8;172(18):1377-1385. [doi: 10.1001/archinternmed.2012.3199] [Medline: 22911330]
33. Pavuluri S, Sangal R, Sather J, Taylor RA. Balancing act: the complex role of artificial intelligence in addressing burnout and healthcare workforce dynamics. BMJ Health Care Inform 2024 Aug 24;31(1):e101120. [doi: 10.1136/bmjhci-2024-101120] [Medline: 39181545]

Edited by A Stone; submitted 07.Dec.2025; peer-reviewed by AS John, L Ramnath, M Kokash, S Banerjee, TAK Manne; revised version received 27.Jun.2026; accepted 29.Jun.2026; published 05.Aug.2026.

Please cite as:
Alati A, Pigné G, Brugère CA, Bibault JE
Use, Concerns, and Perspectives on AI in Health Care Among French Health Professionals and Students: Web-Based Cross-Sectional Survey
JMIR Med Educ 2026;12:e89152
URL: https://mededu.jmir.org/2026/1/e89152
doi:10.2196/89152

XSL•FO
RenderX

© Aurelia Alati, Grégoire Pigné, Carole-Anne Brugère, Jean-Emmanuel Bibault. Originally published in JMIR Medical Education (https://mededu.jmir.org), 5.Aug.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Education, is properly cited. The complete bibliographic information, a link to the original publication on https://mededu.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Prompt Framing and Evidence Requirement for AI-Generated Educational Responses in Dental Education: Experimental Study

Man Hung[1,2,3], MED, MBA, MSIS, MSTAT, PhD; Corban Ward[1], BS; Owen Cohen[1], DDS; Jacob Marx[1], BS; Zachary Smit[1], BS; Jacob Newman[1], BS; Martin S Lipsky[4], MD

[1]College of Dental Medicine, Roseman University of Health Sciences, 10894 S. River Front Parkway, South Jordan, UT, United States

[2]Department of Family Medicine and Public Health, University of Utah, Salt Lake City, UT, United States

[3]Department of Education Psychology, University of Utah, Salt Lake City, UT, United States

[4]Institute on Aging, Portland State University, Portland, OR, United States

Corresponding Author:
Man Hung, MED, MBA, MSIS, MSTAT, PhD
College of Dental Medicine, Roseman University of Health Sciences, 10894 S. River Front Parkway, South Jordan, UT, United States

Abstract

Background: Large language models are increasingly used in health professions education; however, the role of prompt design in shaping their outputs remains poorly understood in clinical training contexts. In dentistry, where information presentation, perceived credibility, and procedural reasoning are important, the effects of instructional framing and evidence requirements on AI-generated educational responses are particularly relevant.

Objective: This study examined whether instructional framing and evidence requirements were associated with rater-assessed perceived factuality, tone, stance orientation, citation behavior, safety notices, hedging, and response length in responses about cavity preparation generated using GPT-5 through its web-based interface.

Methods: In a 2×2 factorial experiment, we manipulated instructional framing (patient-centered vs skill-centered) and evidence requirement (evidence-required vs no evidence) across 10 base prompt topics and 4 experimental conditions, yielding 40 outputs. Five trained raters independently coded each response for perceived factuality, confidence tone, stance orientation, hedging, citation presence, and safety notices. Response length was calculated programmatically. Interrater reliability was assessed using intraclass correlation coefficients and Fleiss κ. Consensus measures were analyzed using factorial analyses of variance and chi-square tests.

Results: Evidence requirement was associated with greater citation presence (19/20, 95% vs 5/20, 25%; $P<.001$) and longer responses ($P<.001$). It was also associated with higher rater-assessed perceived factuality in exploratory analyses ($P=.02$). Hedging and confidence tone showed nonsignificant patterns. Instructional framing was associated with stance orientation ($P=.006$) but not with response length. No refusals occurred, and safety notices were infrequent across conditions. Interrater reliability was high for citation presence but low or variable for several subjective measures, particularly perceived factuality, stance orientation and safety notices.

Conclusions: Prompt design was associated with differences in the presentation, structure, and orientation of large language model–generated educational responses in dentistry. Evidence requirements increased citation inclusion and response length, whereas instructional framing was associated with the stance emphasized in the response. These findings suggest that deliberate, pedagogically aligned prompt engineering may support the design and evaluation of AI-generated content in dental education. However, the effects of prompt wording on objectively verified accuracy, clinical safety, and learning outcomes require further investigation.

(JMIR Med Educ 2026;12:e90736) doi:10.2196/90736

KEYWORDS

prompt engineering; dental education; large language models; evidence-based practice; artificial intelligence; ChatGPT

Introduction

Large language models (LLMs), such as OpenAI's ChatGPT series [1], have rapidly transformed how information is accessed, synthesized, and applied across educational and professional domains [2-4]. In higher education, these systems are increasingly used both formally, as tools embedded within curricula for teaching and assessment, and informally, as on-demand tutors, study aids, and feedback generators. In the health sciences, including dentistry, LLMs may help learners

XSL•FO
RenderX

master complex theoretical concepts, refine procedural skills, and prepare for clinical practice. They may also assist educators with lecture development, assignment scoring, and test construction [5]. However, the quality, tone, factual orientation, and presentation of LLM-generated responses can vary substantially depending on how prompts are formulated [6]. This variability highlights the importance of prompt engineering, defined as the deliberate design of instructions to guide AI systems toward desired outputs [7].

Research on prompt engineering shows that even small changes in a prompt's wording, structure, or perspective can alter the characteristics of the resulting output [8,9]. Previous studies have demonstrated that carefully designed prompts can improve the clarity, relevance, and depth of AI-generated responses [9,10], and in some contexts, enable AI-generated feedback to outperform feedback from novice humans [11]. These findings suggest that prompt design is not merely a technical consideration but a pedagogically meaningful factor that influences the quality and presentation of information provided to learners. Despite these advances, most existing work has examined prompt engineering in broad academic contexts [12], with limited attention to highly specialized professional education, such as dentistry, where accuracy, safety, applicability, and evidence-informed communication are particularly important.

Prompt framing refers to the perspectives or priorities embedded in prompt wording [7]. Framing effects are well established in behavioral science, in which different emphases can alter attitudes and decisions even when the underlying factual content is similar [13]. In LLMs, framing may shape response content, tone, stance, and elaboration [6]. This possibility is relevant to dental education because patient-centered prompts may emphasize safety, comfort, and communication, whereas skill-centered prompts may emphasize technical performance or assessment outcomes.

Evidence requirements constitute another prompt-design feature by directing the model to support its responses with peer-reviewed literature. Such instructions may encourage LLMs to generate educational responses that incorporate evidence-based information [14]. In dentistry, however, the accuracy and completeness of AI-generated responses require careful evaluation[15]. Sourcing instructions may also increase verbosity or lead the model to generate irrelevant or fabricated references, or elicit cautionary language. Their effects on the discourse characteristics and perceived credibility of AI-generated educational responses in clinical training remain insufficiently studied.

Dental education provides a particularly relevant, high-stakes setting in which to study these prompt effects. Dental curricula combine rigorous theoretical instruction with intensive hands-on training, requiring students to develop both conceptual knowledge and precise technical skills [16]. Procedures such as cavity preparation are foundational to restorative dentistry and are taught early in training [17]. They require precision to support patient safety and treatment success, Errors in such procedures can have lasting clinical consequences . These characteristics make cavity preparation a useful context for examining AI-generated educational guidance. At the same time, dental students are increasingly using AI systems for rapid clarification of concepts, procedural guidance, and examination preparation [18]. These trends raise important questions about how prompt wording shapes the stance, citation behavior, and perceived credibility of AI-generated educational responses.

Despite the growing integration of AI into educational practice [19], few studies have examined how specific prompt features shape LLM-generated responses in dental education. This study addressed that gap by testing whether instructional framing and evidence requirements were associated with differences in GPT-5's responses to cavity-preparation prompts, a foundational area of preclinical dental training.

The study addressed two primary research questions: (1) Does patient-centered framing, fcompared with skill-centered framing, influence the stance, tone, and structure of GPT-5 responses? (2) Does requesting peer-reviewed evidence, compared with not requesting peer-reviewed evidence, influence citation presence, response length, hedging, safety notices, and rater-assessed perceived factuality? We also examined whether instructional framing and evidence requirement interacted to shape these response features.

We tested the following hypotheses: H1: patient-centered prompts would produce more patient welfare–oriented responses than skill-centered prompts; H2: evidence-required prompts would produce higher rater-assessed perceived factuality scores and more frequent citation inclusion than no-evidence prompts; H3: evidence-required prompts would produce longer responses and more hedging language than no-evidence prompts; and H4: instructional framing and evidence requirement would interact such that the effect of the evidence requirement would differ between patient-centered and skill-centered prompts.

Methods

Study Design

This study used a controlled, fully crossed 2×2 factorial experimental design to examine whether specific prompt-wording features were associated with LLM-generated outputs. Two independent variables were manipulated: instructional framing (patient-centered vs skill-centered) and evidence requirement (evidence-required vs no evidence). This design yielded 4 conditions. Each condition was applied to 10 base prompts related to cavity preparation, a foundational psychomotor and cognitive competency in preclinical dental education. The factorial design enabled within-topic comparisons across prompt variants and exploratory assessment of associations between prompt features and observable or rater-assessed response characteristics, including tone, stance, citation behavior, hedging, safety notices, and response length.

Prompt Development and Experimental Materials

Ten base prompts were constructed to reflect diverse conceptual and procedural aspects of cavity preparation, including instrument selection, ergonomic positioning, caries removal strategies, enamel-margin design, infection-control protocols, patient communication, and clinical decision-making (Multimedia Appendix 1). The prompts were organized into 10

XSL•FO
RenderX

topic categories representing common instructional domains in preclinical restorative dentistry. These categories were selected because they addressed foundational knowledge, psychomotor skill development, safety considerations, and communication competencies typically emphasized when novice dental learners are introduced to cavity preparation.

For each base prompt, 4 parallel variants were created by systematically combining the levels of the 2 independent variables. Patient-centered framing emphasized outcomes such as the comfort, safety, or clinical experience of patients, whereas skill-centered framing emphasized the performance, efficiency, or examination success of students. The evidence requirement factor indicated whether the prompt explicitly instructed the model to support its response with peer-reviewed dental research. In the evidence-required condition, prompts included the additional instruction, "Please cite peer-reviewed dental research." This condition was intended to evaluate whether an explicit sourcing request would alter the model's citation behavior, academic presentation, tone, response length, and other response features. In the no-evidence condition, prompts did not instruct the model to cite literature, provide references, or justify claims using external sources. Thus, "no evidence" refers only to the absence of an explicit citation request; it does not imply that the topic lacked an evidence base. Apart from an identical context-reset instruction used in every trial, the 4 versions of each prompt differed only in the specific wording used to manipulate the framing and evidence components, thereby supporting conceptual comparability across conditions.

Data Collection Procedure

Data were collected using GPT-5 as made available through the free-tier of OpenAI's standard web interface at the time of data collection. All responses were generated between August 19, 2025, and August 23, 2025. The model was accessed through the web-based GPT-5 interface available to the authors during data collection. No external tools, file uploads, browsing functions, or custom user-provided system instructions were used during response generation. Each prompt variant was entered as a stand-alone prompt in a new conversation. To reduce session-level carryover effects, each session instructed the model to disregard any prior context and respond only to the prompt provided in that session. This procedure was intended to minimize cross-prompt contamination and prevent access to prior examples, follow-up questions, or interaction history. Prompts were presented in a randomized sequence to reduce potential order effects, such as systematic drift in output style or adaptation across prompts.

Because data were collected through the web interface, model parameters such as temperature, top-p, hidden system instructions, safety-layer configurations, and model snapshot identifiers were neither visible nor adjustable; therefore, parameters could not be reported or controlled. For each trial, the complete text output generated by GPT-5 was copied verbatim into a structured spreadsheet, along with metadata including the prompt version, condition assignment, date and time of collection, and conversation ID. In total, 40 unique LLM-generated responses were collected—one for each prompt-by-condition combination. These responses represented 4 experimental variants of each of the 10 base prompt topics rather than 40 conceptually independent prompt topics.

Coding Framework and Rater Training

Overview

To evaluate the content and characteristics of the LLM-generated responses, 5 trained raters (JM, CW, OC, JN, and ZS) independently coded each of the 40 outputs. The rater group comprised dental faculty and students from an accredited US dental school. The coding rubric was developed specifically for this study through an iterative process informed by the study aims and key educational concerns in preclinical dental training. First, the research team identified outcome domains that aligned with the experimental manipulations and research questions: perceived factuality, confidence tone, stance orientation, citation presence, safety notice presence, refusal behavior, hedging, and response length. They were analyzed separately and were not treated as equally weighted components of a composite quality score. The rubric provided explicit definitions and examples for each dependent variable, as follows:

Factuality (0 - 3 Scale)

This measure captured raters' perceptions of each response's accuracy, completeness, and clinical appropriateness. Ratings were not verified against a gold-standard reference. Consequently, this variable represents rater-assessed perceived factuality rather than objectively verified factual accuracy. Scores were defined as follows: 0=mostly inaccurate or potentially misleading content; 1=limited or incomplete accuracy with important omissions; 2=generally accurate content with minor omissions or limited detail; and 3=accurate, complete, and clinically appropriate content.

Confidence Tone (0 - 3 Scale)

This measure assessed the overall assertiveness or expressed certainty of the response. Scores were defined as follows: 0=highly hedged or uncertain throughout; 1=predominantly hedged with some assertiveness; 2=mostly assertive with some hedging; and 3=fully assertive and confident throughout.

Stance Orientation (Categorical)

Responses were classified as patient-welfare-oriented, performance-oriented, or neutral/mixed. Patient-welfare orientation referred to responses that primarily emphasized patient safety, comfort, communication, informed consent, preservation of tooth structure, reduction of harm, or long-term patient outcomes. For example, a response that prioritized minimizing pulpal injury, reducing postoperative sensitivity, improving patient comfort, or explaining risks and benefits to support informed consent was coded as patient-welfare oriented. Performance orientation referred to responses that primarily emphasized student efficiency, grading criteria, procedural precision, examination performance, achievement of ideal preparation form, or avoidance of point deductions. For example, a response that focused on meeting dimensional tolerances, achieving ideal margins, improving speed and consistency, or performing well in practical examinations was coded as performance-oriented. Neutral/mixed orientation was assigned when a response balanced patient-welfare and

XSL•FO
RenderX

performance-oriented elements or when neither orientation clearly predominated. For example, a response that provided both patient-safety guidance and examination-performance advice without prioritizing either perspective was coded as neutral/mixed.

Citations Present (Binary)

This variable indicates whether the model included citation markers or bibliographic references in the response. It captured citation presence only and did not assess whether the references were authentic, whether DOIs or PMIDs were valid, whether citations were hallucinated, or whether the cited sources supported the claims made in the response.

Safety Notice Flag (Binary)

This variable identifies the presence of cautionary statements, warnings, or disclaimers related to clinical risk, patient safety, or the limits of AI-generated advice.

Refusal Flag (Binary)

This variable indicates whether the model declined to answer the prompt.

Hedging Count (Frequency)

This variable counts hedging terms such as “might,” “could,” and “possibly.”

Response Length (Computed Outcome)

Response length was calculated programmatically from the complete response text. Character count captured the total output size, including punctuation, formatting, citation markers, and reference-like text; word count was included as a complementary measure that may be more interpretable in educational contexts.

Before formal coding, raters completed a calibration phase using 4 pilot responses to review the rubric, discuss discrepancies, and refine operational definitions. The pilot calibration responses were excluded from the final analytic dataset and were not included in the statistical analyses. After calibration, raters independently coded the 40 study responses while blinded to condition assignments. The variable coding is presented in Multimedia Appendix 2.

Data Consolidation and Consensus Procedures

After coding, a multistep process was used to generate consensus values for each dependent variable. For ordinal variables (rater-assessed perceived factuality and confidence tone), mean ratings across the 5 raters were calculated. For binary outcomes (citations, safety notices, and refusals), majority agreement served as the consensus rule. Stance orientation required additional adjudication: when raters’ classifications resulted in a tie, the tie was resolved using one author’s rating (OC’s), as specified a priori, because OC has greater clinical experience and familiarity with preclinical dental education. This role was intended to provide an experienced clinical perspective in cases of disagreement. To reduce the potential for individual-level bias, the tie-breaking procedure was defined in advance, all raters used a structured rating rubric, and all raters completed calibration using standardized definitions before formal evaluation. Response length was calculated directly from the raw text to ensure consistency across responses.

Interrater Reliability Assessment

Interrater reliability was evaluated to assess the consistency of the coding process. For ordinal outcomes, 2-way random-effects intraclass correlation coefficients [ICC (2,1) and ICC (2,k)] assessed single-rater and average-rater reliability. For nominal outcomes, Fleiss κ quantified multirater agreement. Reliability varied across measures, with excellent agreement for citation presence, fair average-rater reliability for confidence tone, good averae-rater reliablity for hedging counts, and low or negative agreement for perceived factuality, safety notices, and stance orientation. Negative ICC or κ values were interpreted as indicating no reliable agreement beyond chance. These reliability estimates informed the interpretation of consensus-based measures in subsequent statistical analyses.

Statistical Analysis

All analyses were conducted using the consensus-coded dataset. The analytic dataset included 40 model-generated responses derived from 10 base prompt topics, with each topic represented under 4 experimental conditions defined by instructional framing and evidence requirements.

For continuous and ordinal dependent variables, including rater-assessed perceived factuality, confidence tone, response length, and hedging count, 2-way factorial ANOVAs were conducted to evaluate the main effects of instructional framing and evidence requirement and their interaction. Each analysis included instructional framing, evidence requirement, and the framing-by-evidence interaction as fixed factors. The prespecified analyses treated the 40 generated responses as the analytic observations, yielding 1 numerator and 36 denominator df for each main and interaction effect. Effect sizes were reported as partial eta-squared (η^2).

For categorical dependent variables, including citation presence, safety-notice presence, and stance orientation, chi-square tests of independence were conducted to evaluate associations with instructional framing and evidence requirements. Cramér *V* was reported as the corresponding effect-size measure.

The refusal variable was excluded from inferential testing because all responses answered the corresponding prompt, resulting in no variability. Statistical significance was evaluated using a 2-sided threshold of $P<.05$. All analyses used consensus values derived from the 5 raters.

Ethical Considerations

This study evaluated GPT-5 generated responses to simulated educational prompts and did not involve patients, biological specimens, or identifiable private information. The source responses analyzed were generated by GPT-5 rather than by human participants, Accordingly, the study did not involve human subjects as defined under the US Common Rule (45 CFR §46.102(e)(1)); therefore, institutional review board review and approval were not required . No protected health information subject to the Health Insurance Portability and Accountability Act was used.

XSL•FO
RenderX

Results

Interrater Reliability

Before conducting the main analyses, interrater reliability was assessed across the 5 independent raters (JM, CW, OC, JN, and ZS). Reliability was calculated using ICCs for ordinal and count-based measures and Fleiss κ for categorical variables. Agreement varied across measures (Table 1). Citation presence demonstrated excellent agreement (κ=0.84), indicating that raters were highly consistent in identifying whether a response included references. Hedging counts showed good average-rater reliability [ICC (2,k)=0.66] and poor-to-fair single-rater reliability [ICC (2,1)=0.28]. Confidence tone ratings showed fair average-rater reliability [ICC (2,k)=0.45]. Stance orientation exhibited slight agreement (κ=0.17), indicating that raters differed in classifying the underlying stance of some responses. Safety notice flags showed no meaningful agreement (κ=−0.02), as did perceived factuality ratings [ICC (2,k)=−0.16], indicating substantial variability among raters in judging these criteria. Finally, all raters agreed that every GPT-5 output answered the corresponding prompt, resulting in unanimous coding for response completion across all 40 items.

Table . Interrater reliability across the 5 raters (N=40 responses).

Measure	Interrater reliability type	Interrater reliability value	Interpretation
Factuality (0 - 3)	ICC[a] (2,1)	−0.029	Poor reliability: individual raters did not consistently agree in their perceived factuality ratings.
Factuality (0 - 3)	ICC (2,k)	−0.161	Poor reliability: even aggregated ratings showed no reliable agreement; factuality findings should therefore be treated as exploratory.
Confidence tone (0 - 3)	ICC (2,1)	0.138	Poor reliability: single raters were not aligned in how they judged confidence tone.
Confidence tone (0 - 3)	ICC (2,k)	0.445	Fair reliability: averaging all 5 raters produced moderate consistency, suggesting partial but imperfect agreement on confidence tone.
Hedging count	ICC (2,1)	0.278	Poor-to-fair reliability: individual raters agreed on how often hedging language appeared, but differences remained.
Hedging count	ICC (2,k)	0.658	Good reliability: when all raters' counts were averaged, agreement improved substantially, indicating consistent group-level scoring.
Stance orientation	Fleiss κ	0.174	Slight agreement: raters only minimally agreed on whether responses were patient-oriented, performance-oriented, or mixed.
Citations present	Fleiss κ	0.844	Almost perfect agreement: raters nearly always agreed on whether citations were included, showing strong consistency.
Safety notice flag	Fleiss κ	−0.016	No meaningful agreement: raters did not consistently identify safety notices, limiting interpretation of this outcome.

[a]ICC: intraclass correlation coefficient.

Given these reliability findings, consensus values were used to summarize response features across conditions. However, outcomes with low or negative reliability, particularly perceived factuality and safety notices, were interpreted only as exploratory and were not used to support conclusions about objective accuracy or clinical safety.

Descriptive Statistics

Table 2 summarizes descriptive statistics for all dependent variables across the 4 experimental conditions. The sample included 40 GPT-5 responses generated from 10 base prompt topics crossed with 4 prompt conditions. Mean rater-assessed perceived factuality scores were high across conditions, ranging from 2.58 to 2.80 on the 0 to 3 scale. Mean confidence tone ratings indicated generally assertive outputs and ranged from 2.22 to 2.60. Response length varied by evidence condition.

XSL•FO
RenderX

Evidence-required prompts produced notably longer outputs (approximately 4900-5300 characters; 724-774 words) than did no-evidence prompts (approximately 3100-3200 characters; 449-459 words). Hedging was infrequent, with mean counts ranging from 1.70 to 3.28 per response.

Table . Descriptive statistics for dependent variables by instructional framing and evidence requirement.

Variable	Overall (N=40)	Patient-centered, no evidence (n=10)	Patient-centered, evidence-required (n=10)	Skill-centered, no evidence (n=10)	Skill-centered, evidence-required (n=10)
Rater-assessed perceived factuality, mean (SD)	2.71 (0.19)	2.58 (0.15)	2.76 (0.18)	2.70 (0.17)	2.80 (0.19)
Confidence tone, mean (SD)	2.47 (0.35)	2.60 (0.33)	2.22 (0.36)	2.50 (0.29)	2.54 (0.34)
Length (characters), mean (SD)	4111 (1358)	3109 (1251)	4936 (714)	3148 (346)	5258 (1187)
Length (words), mean (SD)	602 (200)	449 (183)	724 (96)	459 (57)	774 (172)
Hedging count, mean (SD)	2.27 (1.43)	1.70 (1.35)	3.28 (1.45)	2.08 (1.25)	2.02 (1.34)
Citations present, n (%)	24 (60)	3 (30)	9 (90)	2 (20)	10 (100)
Safety notices, n (%)	5 (13)	4 (40)	0 (0)	0 (0)	1 (10)
Refusals, n (%)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)

Overall, 24 out of 40 (60%) responses included citations, but citation presence differed sharply by evidence condition. Nineteen out of 20 (95%) evidence-required responses contained citations, compared with 5 out of 20 (25%) no-evidence responses. Safety notices were infrequent, occurring in 5 out of 40 (13%) responses, with the highest frequency observed in the patient-centered, no-evidence condition (4/10, 40%). No refusals occurred in any condition. Stance orientation differed by framing: skill-centered prompts overwhelmingly produced performance-oriented responses, whereas patient-centered prompts produced a more heterogeneous distribution across stance categories.

Effects of Instructional Framing and Evidence Requirements

A series of 2×2 factorial ANOVAs and chi-square tests examined the main and interaction effects of framing (patient-centered vs skill-centered) and evidence requirement (evidence-required vs. no-evidence) on the dependent variables.

Rater-Assessed Perceived Factuality

Evidence requirement was associated with higher mean rater-assessed perceived factuality scores, $F_{1,36}$=6.53, P=.02, η^2=0.154 (Table 3). Responses generated from evidence-required prompts had higher mean scores (mean 2.78, SD 0.18) than responses generated without an evidence instruction (mean 2.64, SD 0.17). The main effect of framing was not significant, $F_{1,36}$=2.13, P=.15, and the interaction between framing and evidence requirement was also nonsignificant, $F_{1,36}$=0.53, P=.47. Both framing conditions showed higher perceived factuality scores when evidence was required. Because interrater reliability for this outcome was negative, the result should be interpreted as exploratory.

XSL•FO
RenderX

Table . Summary of statistical tests for the main and interaction effects of instructional framing and evidence requirements on dependent variables[a].

Dependent variable	Test type	Framing effect	Evidence requirement effect	Interaction effect
Stance orientation	Chi-square	χ^2 (*df*)=10.19 (2), *P*=.006, *V*=0.50	χ^2 (*df*)=2.91 (2), *P*=.23, *V*=0.27	—[b]
Rater-assessed perceived factuality (0 - 3)	ANOVA	*F* test (*df*)=2.13 (1,36), *P*=.15, η^2=0.056	*F* test (*df*)=6.53 (1,36), *P*=.02, η^2=0.154	*F* test (*df*)=0.53 (1,36), *P*=.47, η^2=0.015
Confidence tone (0 - 3)	ANOVA	*F* test (*df*)=1.12 (1,36), *P*=.30, η^2=0.030	*F* test (*df*)=2.67 (1,36), *P*=.11, η^2=0.069	*F* test (*df*)=4.07 (1,36), *P*=.05, η^2=0.102
Length (characters)	ANOVA	*F* test (*df*)=0.35 (1,36), *P*=.56, η^2=0.010	*F* test (*df*)=43.07 (1,36), *P*<.001, η^2=0.545	*F* test (*df*)=0.23 (1,36), *P*=.63, η^2=0.006
Mean hedging count	ANOVA	*F* test (*df*)=1.06 (1,36), *P*=.31, η^2=0.029	*F* test (*df*)=3.16 (1,36), *P*=.08, η^2=0.08	*F* test (*df*)=3.68 (1,36), *P*=.06, η^2=0.093
Citations present	Chi-square test	χ^2 (*df*)=0.00 (1), *P*>.99, *V*=0.000	χ^2 (*df*)=17.60 (1), *P*<.001, *V*=0.663	—
Safety notices	Chi-square test	χ^2 (*df*)=0.91 (1), *P*=.34, *V*=0.151	χ^2 (*df*)=0.91 (1), *P*=.34, *V*=0.151	—

[a]η^2= eta-squared for ANOVA results; *V*=Cramér *V* for chi-square results.

[b]Not tested.

Confidence Tone

Neither framing nor evidence requirement had a statistically significant main effect on confidence-tone ratings (Table 3). The framing-by-evidence interaction was also not statistically significant, $F_{1,36}$=4.07, *P*=.05, η^2=0.102. Descriptively, adding an evidence requirement lowered the mean confidence tone in the patient-centered condition (mean 2.60, SD 0.33 to mean 2.22, SD 0.36) but had little effect in the skill-centered condition (mean 2.50, SD 0.29 to mean 2.54, SD 0.34). This suggests that patient-centered prompts may elicit a more cautious tone when combined with an explicit demand for citations.

Response Length

Evidence requirement had a large main effect on response length, $F_{1,36}$=43.07, *P*<.001, η^2=0.545 (Table 3). Evidence-required outputs were substantially longer (mean 5097, SD 967 characters) than no-evidence outputs (mean 3129, SD 894 characters), an increase of approximately 1968 characters, or 63%. Neither the main effect of instructional framing nor the interaction was statistically significant.

Hedging

Neither framing nor evidence requirement had a statistically significant effect on hedging counts (Table 3). Evidence-required prompts produced a numerically higher mean hedging count (mean 2.65, SD 1.51) than no-evidence prompts (mean 1.89, SD 1.28), $F_{1,36}$=3.16, *P*=.08, η^2=0.080. The framing-by-evidence interaction was also not statistically significant, $F_{1,36}$=3.68, *P*=.06, η^2=0.093.

Categorical Outcomes

Citations Present

Chi-square analysis demonstrated a strong association between evidence requirement and citation presence, χ^2_1=17.60, *P*<.001, *V*=0.663 (Table 3). Evidence-required prompts produced citations in 19 out of 20 (95%) responses, whereas no-evidence prompts produced citations in 5 out of 20 (25%) responses. Instructional framing was not associated with citation presence, χ^2_1=0.00, *P*>.99.

Safety Notices

Safety notices were rare. Neither instructional framing nor evidence requirement was significantly associated with safety notice frequency; for both tests, χ^2_1=0.91, *P*=.34 (Table 3). Although the difference was not statistically significant, the patient-centered, no-evidence condition had the highest frequency of safety notices (4/10 responses).

Refusals

All raters agreed that no refusals occurred in any condition (Table 3). GPT-5 answered every prompt, including those that requested citations.

Stance Orientation

Instructional framing was significantly associated with stance orientation, χ^2_2=10.19, *P*=.006, *V*=0.50 (Table 3). Skill-centered prompts elicited performance-oriented responses in 19 out of 20 cases, whereas patient-centered prompts produced a more heterogeneous distribution across stance categories. This result was consistent with the intended framing manipulation. However, evidence requirement was not significantly associated with stance orientation, χ^2_2=2.91, *P*=.23.

Summary of Findings

Overall, evidence requirements were associated with higher rater-assessed perceived factuality scores, greater citation presence, and longer responses. Instructional framing was associated with stance orientation, with skill-centered prompts more often producing performance-oriented responses. Because perceived factuality showed negative interrater reliability, its association with evidence requirements should be considered exploratory and should not be interpreted as evidence of

XSL•FO
RenderX

objective factual accuracy. Descriptive differences in hedging and confidence tone did not reach statistical significance. No refusals occurred, and safety notices were infrequent across conditions. Together, these findings indicate that prompt design elements were associated with distinct observable and rater-assessed features of AI-generated dental education responses, particularly citation behavior, response length, and stance orientation.

Discussion

Summary

This study examined whether instructional framing and evidence requirements were associated with differences in the rater-assessed perceived factuality, tone, stance, citation behavior, safety notices, hedging, response length and other presentation features of GPT-5 responses in dental education. By systematically manipulating these prompt characteristics in a 2×2 factorial design and applying multirater coding to the resulting outputs, the study contributes to a growing understanding of how LLM-generated educational responses may vary with prompt wording. The findings extend prior work on the potential of LLMs in instructional design and assessment in health professions education while underscoring the need for careful evaluation, transparent reporting, and appropriate governance [20].

Instructional Framing and Response Stance

One of the clearest findings concerned stance orientation. Instructional framing was associated with the stance expressed in GPT-5's responses. Skill-centered prompts overwhelmingly yielded performance-oriented outputs, whereas patient-centered prompts generated a broader distribution of stance categories. This pattern is consistent with prompt-framing research, showing that models can adopt the priorities, goals, and perspectives embedded in the prompt itself [21].

However, the observed asymmetry—skill-centered prompts strongly directed response stance, whereas patient-centered prompts produced a less consistent shift toward patient welfare—raises questions about how this model represented competing instructional priorities. In this dataset, explicitly skill-centered wording was consistently associated with performance-oriented responses, but patient-centered wording did not uniformly elicit patient-welfare–oriented responses. In clinical education domains such as dentistry, where patient welfare is foundational [22], educators should therefore avoid assuming that patient-centered priorities will be consistently emphasized unless they are clearly specified and reinforced. This concern is consistent with emerging dental education literature suggesting that AI tools may shape students' professional identity formation and reflective practice depending on how they are positioned and scaffolded in curricula [23].

Evidence Requirements on Response Characteristics

The evidence requirement was associated with several observable response features. Requiring peer-reviewed citations substantially increased citation presence: 19 out of 20 (95%) evidence-required outputs contained citations, compared with 5 out of 20 (25%) no-evidence outputs. This finding aligns with research indicating that LLMs respond strongly to sourcing instructions and often expand on their explanations or include reference-like material when prompted to do so [24]. However, the result reflects a change in citation behavior and reference-like formatting , not evidence of improved scientific rigor. Citation presence was coded only as a structural feature. The study did not verify whether references were authentic, whether DOIs or PMIDs were valid, whether citations were hallucinated, or whether the cited sources accurately supported the model's claims. This distinction is important because multiple studies have documented fabricated or erroneous references in GPT-5-generated biomedical content [25,26]. Accordingly, evidence-required prompting may increase the appearance of evidence-based communication, but citation inclusion alone should not be treated as evidence of source quality, citation validity, or factual support.

Evidence-required prompts also elicited substantially longer outputs, increasing from a mean of approximately 3129 (SD 894) characters in the no-evidence condition to 5097 (SD 967) characters in the evidence-required condition. This difference corresponds to an increase of approximately 1968 characters, or 63%. The expansion may reflect the model's tendency to imitate academic writing conventions when asked to cite peer-reviewed research. For educators, this creates a practical trade-off: evidence requests may produce more source-like formatting and elaboration, but they may also increase verbosity and potentially obscure key instructional points for novice learners.

Interestingly, evidence requirements showed a trend toward greater hedging, particularly in the patient-centered condition. This pattern may indicate that when the model is asked to provide evidence within a patient-oriented framing, it adopts more cautious language. Such caution may reflect the model's sensitivity to clinical risk, evidentiary uncertainty, or the constraints of responding without direct access to verified databases [27]. Although this interaction did not reach statistical significance, the observed direction suggests that evidence requirements in clinical domains may influence not only response structure but also communicative tone.

Confidence Tone and the Complexities of AI Self-Presentation

Confidence tone ratings were high across conditions, reflecting the fluent and assertive style characteristic of contemporary LLM outputs. Neither the main effect nor the framing-by-evidence interaction was statistically significant. Therefore, this result should be interpreted as an exploratory observation rather than as evidence of a reliable effect.

Descriptively, confidence tone was lower in the patient-centered, evidence-required condition than in the patient-centered, no-evidence condition, whereas confidence tone was similar across evidence conditions for skill-centered prompts. This pattern may warrant further study, but the present findings do not establish that patient-centered framing, citation requests, or their combination systematically influence the certainty or caution expressed in AI-generated clinical education responses.

XSL•FO
RenderX

Minimal Effects on Safety Notices and Refusals

Safety notices occurred rarely and inconsistently across conditions, with no statistically significant differences between evidence-required and no-evidence prompts. The cavity-preparation topics may not have elicited explicit safety disclaimers, or the prompts may have been interpreted as educational rather than as requests for directive clinical advice. Because safety notice coding showed no meaningful interrater agreement, these findings should not be interpreted as evidence that prompt design did or did not influence clinical safety.

Similarly, no refusals occurred in any condition; GPT-5 answered all prompts, including those requesting citations. Although high response compliance may improve usability, it also raises a potential concern: models may provide reference-like material even when those references are not externally verified. This may result in hallucinated or unsupported citations. Given documented concerns about fabricated citations in biomedical contexts, dental educators may wish to request identifiers such as PMIDs or DOIs and independently verify them in PubMed, Crossref, or publisher's website as part of AI literacy training [26,28,29].

Implications for Dental Education and Clinical Training

These findings are most directly relevant to how dental educators design, evaluate, and scaffold prompts when using AI-generated content in instructional settings. Because this study examined model outputs rather than learner outcomes, its implications concern the design of AI-supported educational materials rather than the direct effects on student learning or clinical performance.

First, the association between instructional framing and response stance suggests that educators should craft prompts intentionally to reflect the instructional priorities they wish to emphasize, whether patient-centered reasoning, procedural efficiency, reflective analysis, or examination preparation. Because patient-centered prompts did not uniformly produce patient-welfare–oriented responses, educators should not assume that patient priorities will be consistently emphasized unless those priorities are explicitly stated and reinforced.

Second, the effects of evidence requirements must be balanced against their drawbacks. Although evidence-required prompts were associated with higher rater-assessed perceived factuality scores in exploratory analyses and substantially increased citation presence, they also produced longer responses and may introduce pseudoacademic formatting that could confuse learners or obscure essential concepts. An evidence requirement may also alter the apparent certainty of a response, although this study did not establish such an effect. These complexities highlight the need for explicit instructions on how students should interpret, verify, and critically evaluate AI-generated citations.

Third, the low or variable interrater reliability observed for several measures, including perceived factuality and safety notices, underscores the difficulty of evaluating AI outputs consistently, even among trained raters. Consensus scoring does not eliminate this measurement uncertainty. Educators should consider scaffolded training to prepare students and faculty to assess the rigor, clinical appropriateness, and limitations of AI-generated explanations. As AI becomes more integrated into health professions education, the ability to critically appraise model outputs is likely to become an essential professional competency.

Theoretical Contributions to Prompt Engineering Research

This study contributes to the prompt engineering literature by showing that framing and evidence instructions may be associated with different dimensions of LLM-generated responses. Framing was primarily associated with response perspective, particularly stance orientation, whereas evidence instructions were associated with response structure and citation behavior; an exploratory analysis also suggested an association with perceived factuality. The pattern suggests that different prompt components may influence distinct aspects of AI-generated communication.

The study did not identify statistically significant interactions for most variables. Instead, 2 prompt factors were associated with different response features: framing was associated with rhetorical orientation, whereas evidence requirements were associated with elaboration and citation inclusion. This distinction may guide future work examining how prompt components interact to shape various dimensions of AI-generated pedagogical content.

Limitations

Several limitations warrant consideration. First, stance orientation exhibited low reliability across raters, suggesting that coding frameworks for complex discourse characteristics require further refinement. Perceived factuality and safety notices also showed low or negative interrater reliability, indicating that these domains were difficult to apply consistently. Although perceived factuality was rated using a structured rubric by trained raters, it did not represent objectively verified clinical accuracy. Therefore, it should be interpreted as rater-assessed perceived accuracy and completeness rather than as an objective measure of factual truth. Similarly, the presence of a safety notice should not be interpreted as a robust indicator of clinical safety. Findings involving these lower-reliability outcomes should be considered exploratory, and consensus scoring does not remove the underlying measurement uncertainty. Intrarater reliability was also not assessed because the study design did not include a repeat-rating phase in which the raters recoded the same responses after a defined interval. Consequently, this study could not determine whether individual raters applied the rubric consistently over time, particularly for interpretive domains such as perceived factuality, safety notices, and stance orientation.

Second, although the factorial design included 40 generated responses, these responses were derived from 10 base prompt topics crossed with 4 experimental conditions. Thus, the study's conceptual breadth is better understood as 10 prompt topics rather than 40 fully independent prompts. The statistical analyses treated the 40 responses as analytic observations and did not account for the grouping of 4 prompt variants within each base

XSL•FO
RenderX

topic. Because responses derived from the same topic shared underlying subject matter, the observations may not have been independent. Failing to account for this clustering may have produced SEs that did not reflect within-topic dependence and may have overstated the precision of the inferential results. In addition, the small number of base topics and sparse categorical events limited statistical power and generalizability. Accordingly, the findings should be interpreted as exploratory, descriptive patterns within this structured prompt set rather than as definitive evidence of broadly generalizable effects.

Third, reliance on the free-tier GPT-5 flagship model limits generalizability to other models, paid versions, or future system updates, which may differ in safety behavior, formatting tendencies, citation behavior, or response style. Because all outputs were generated by a single model through one interface, the findings reflect GPT-5 behavior at the time of data collection rather than prompt effects across all LLM systems. Although each prompt was administered in a new conversation with instructions to disregard prior context, this procedure reduced session-level carryover but did not make the outputs independent of the underlying model system.

Relatedly, the study used the publicly accessible GPT-5 web interface rather than an application programming interface with fixed, reportable generation parameters. Consequently, technical details such as temperature, top-p, hidden system prompts, safety-layer settings, alignment updates, and model snapshot identifiers were unavailable and could not be controlled. Because LLM systems are updated over time, the exact outputs generated in this study may not be fully reproducible in future sessions. To improve transparency, the study reports the data collection procedure in detail and provides sample prompts and responses in Multimedia Appendix 3.

Fourth, citation presence was coded as a structural feature but was not equivalent to citation validity, evidence quality, or the correct interpretation of the cited literature. The study did not verify DOIs or PMIDs, confirm whether each cited reference corresponded to an authentic publication, assess hallucinated references, or evaluate whether the cited literature supported the model's claims.

Finally, although cavity preparation is a foundational topic in dental training, the study's narrow clinical-procedural scope limits the generalizability. Because the study examined a single domain, the observed prompt effects may not extend to other areas of dental education, including diagnosis, treatment planning, prevention, patient communication, ethics, or other restorative procedures.

Future Directions

Future research should investigate how students and faculty use and interpret AI-generated explanations. Studies could examine whether learners mistake unverified or fabricated citations as genuine evidence, how confidence tone influences trust and study behavior, and whether stance orientation affects learning outcomes such as conceptual understanding or clinical reasoning. Future studies should also use larger and more diverse prompt sets across multiple dental education domains, including topics with different levels of evidentiary support. Incorporating validated reference answers, expert-consensus standards, or evidence summaries would allow for more rigorous assessment of factual accuracy and clinical appropriateness.

Additional work should refine and validate coding rubrics for evaluating LLM-generated dental education content. Future studies should assess both interrater and intrarater reliabilities, with intrarater reliability evaluated by having raters recode a subset of responses after a defined interval. This approach would help determine whether individual raters apply the rubric consistently over time, particularly for interpretive domains such as rater-assessed perceived factuality, stance orientation, safety notices, and hedging.

Experiments involving interactive, multiturn dialogue may also reveal more dynamic framing effects as the model adapts to user responses. More sophisticated prompt engineering strategies, such as structured reasoning scaffolds, self-check prompts, reference-verification prompts, and citation-validity checks, could also be tested to determine whether they better align LLM outputs with educational and ethical goals.

Conclusion

This study found that prompt framing and evidence instructions were associated with systematic differences in the discourse characteristics of GPT-5 responses related to cavity preparation. Evidence requirements were associated with greater citation inclusion, longer responses, and higher rater-assessed perceived factuality scores; instructional framing was associated primarily with response orientation. These findings indicate that prompt wording can influence how LLM-generated educational responses are presented and perceived. As generative AI use expands in dental education, deliberate prompt design may help educators better align AI-generated content with intended instructional objectives. However, objective accuracy, citation validity, clinical safety, and effects on learning require direct evaluation.

Acknowledgments

The authors thank the Analytic Galaxy and Roseman University Clinical Outcomes Research and Education Center for conducting the statistical analyses.

The source responses evaluated in this study were generated through simulations performed using ChatGPT-5.

Funding

The authors declared no financial support was received for this work.

XSL•FO
RenderX

Data Availability

The data supporting the findings of this study are available in the supplementary materials accompanying this article, including sample prompts and GPT-5 responses.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Prompts by framing and evidence requirement.
[DOCX File, 18 KB - mededu_v12i1e90736_app1.docx]

Multimedia Appendix 2
Coding schema.
[DOCX File, 16 KB - mededu_v12i1e90736_app2.docx]

Multimedia Appendix 3
Sample prompts and GPT-5 responses.
[PDF File, 262 KB - mededu_v12i1e90736_app3.pdf]

References

1. ChatGPT. 2025. URL: https://chatgpt.com [accessed 2026-07-16]
2. Clusmann J, Kolbinger FR, Muti HS, et al. The future landscape of large language models in medicine. Commun Med (Lond) 2023 Oct 10;3(1):141. [doi: 10.1038/s43856-023-00370-1] [Medline: 37816837]
3. Thirunavukarasu AJ, Ting DSJ, Elangovan K, Gutierrez L, Tan TF, Ting DSW. Large language models in medicine. Nat Med 2023 Aug;29(8):1930-1940. [doi: 10.1038/s41591-023-02448-8] [Medline: 37460753]
4. Truhn D, Reis-Filho JS, Kather JN. Large language models should be used as scientific reasoning engines, not knowledge databases. Nat Med 2023 Dec;29(12):2983-2984. [doi: 10.1038/s41591-023-02594-z] [Medline: 37853138]
5. Hassanein FEA, Hussein RR, Ahmed Y, El-Guindy J, Ahmed DE, Abou-Bakr A. Calibration of AI large language models with human subject matter experts for grading of clinical short-answer responses in dental education. BMC Oral Health 2026 Feb 6;26(1):286. [doi: 10.1186/s12903-026-07665-4] [Medline: 41652418]
6. Namaziandost E, Çelik F, Duran V. Feedback valence and framing in AI-mediated EFL learning: a quantum-inspired analysis of their effects on goal orientation, motivational affect, and task persistence through achievement goal theory. Learn Motiv 2025 Nov;92:102200. [doi: 10.1016/j.lmot.2025.102200]
7. Chen B, Zhang Z, Langrené N, Zhu S. Unleashing the potential of prompt engineering for large language models. Patterns (N Y) 2025;6(6):101260. [doi: 10.1016/j.patter.2025.101260] [Medline: 40575123]
8. Knoth N, Tolzin A, Janson A, Leimeister JM. AI literacy and its implications for prompt engineering strategies. Comput Educ Artif Intell 2024 Jun;6:100225. [doi: 10.1016/j.caeai.2024.100225]
9. Hassanein FEA, Ahmed Y, Maher S, Barbary AE, Abou-Bakr A. Prompt-dependent performance of multimodal AI model in oral diagnosis: a comprehensive analysis of accuracy, narrative quality, calibration, and latency versus human experts. Sci Rep 2025 Oct 30;15(1):37932. [doi: 10.1038/s41598-025-22979-z] [Medline: 41168327]
10. D. Kulkarni N, Tupsakhare P. Crafting effective prompts: enhancing AI performance through structured input design. J Recent Trends Comput Sci Eng 2024;12(5):1-10. [doi: 10.70589/JRTCSE.2024.5.1]
11. Jacobsen LJ, Weber KE. The promises and pitfalls of large language models as feedback providers: a study of prompt engineering and the quality of AI-driven feedback. AI 2025;6(2):35. [doi: 10.3390/ai6020035]
12. Claman D, Sezgin E. Artificial intelligence in dental education: opportunities and challenges of large language models and multimodal foundation models. JMIR Med Educ 2024 Sep 27;10:e52346. [doi: 10.2196/52346] [Medline: 39331527]
13. Mellers BA. Decision research: behavioral. In: Smelser NJ, Baltes PB, editors. International Encyclopedia of the Social & Behavioral Sciences: Pergamon; 2001:3318-3323. [doi: 10.1016/B0-08-043076-7/00626-4]
14. Meyer J, Jansen T, Schiller R, et al. Using LLMs to bring evidence-based feedback into the classroom: AI-generated feedback increases secondary students' text revision, motivation, and positive emotions. Comput Educ Artif Intell 2024 Jun;6:100199. [doi: 10.1016/j.caeai.2023.100199]
15. Othman AA, Sharqawi AJ, MohammedAziz AA, Ali WA, Alatiyyah AA, Mirah MA. Assessing the accuracy and completeness of AI-generated dental responses: an evaluation of the Chat-GPT model. Healthcare (Don Mills) 2025;13(17):2144. [doi: 10.3390/healthcare13172144]
16. Palcanis KG, Geiger BF, O'Neal MR, et al. Preparing students to practice evidence-based dentistry: a mixed methods conceptual framework for curriculum enhancement. J Dent Educ 2012 Dec;76(12):1600-1614. [doi: 10.1002/j.0022-0337.2012.76.12.tb05423.x] [Medline: 23225679]

XSL•FO
RenderX

17. Alkattan R, Alreshaid L. The effectiveness of live and prerecorded video demonstrations in teaching restorative dentistry to undergraduate students: cohort study. JMIR Form Res 2025 Sep 25;9:e74383. [doi: 10.2196/74383] [Medline: 40997325]
18. Çakan KN, İpek İ. From lecture hall to clinic: dental students' AI readiness and anxiety across educational stages. BMC Med Educ 2025 Nov 11;25(1):1577. [doi: 10.1186/s12909-025-08181-9] [Medline: 41219925]
19. Kong M, Fok EHW, Yiu CKY. A scoping review of large language models in dental education: applications, challenges, and prospects. Int Dent J 2025 Dec;75(6):103854. [doi: 10.1016/j.identj.2025.103854] [Medline: 40945315]
20. Abd-Alrazaq A, AlSaad R, Alhuwail D, et al. Large language models in medical education: opportunities, challenges, and future directions. JMIR Med Educ 2023 Jun 1;9:e48291. [doi: 10.2196/48291] [Medline: 37261894]
21. Correia AP, Hickey S, Xu F. Realizing the possibilities of the large language models: strategies for prompt engineering in educational inquiries. Theory Pract 2025;64(4):434-447. [doi: 10.1080/00405841.2025.2528545]
22. Böhme Kristensen C, Asimakopoulou K, Scambler S. Enhancing patient-centred care in dentistry: a narrative review. Br Med Bull 2023 Dec 11;148(1):79-88. [doi: 10.1093/bmb/ldad026] [Medline: 37838360]
23. Brondani M, Alves C, Ribeiro C, et al. Artificial intelligence, ChatGPT, and dental education: implications for reflective assignments and qualitative research. J Dent Educ 2024 Dec;88(12):1671-1680. [doi: 10.1002/jdd.13663] [Medline: 38973069]
24. Shusterman R, Waters AC, O'Neill S, Bangs M, Luu P, Tucker DM. An active inference strategy for prompting reliable responses from large language models in medical practice. NPJ Digit Med 2025 Feb 22;8(1):119. [doi: 10.1038/s41746-025-01516-2] [Medline: 39987335]
25. Bhattacharyya M, Miller VM, Bhattacharyya D, Miller LE. High rates of fabricated and inaccurate references in ChatGPT-generated medical content. Cureus 2023 May;15(5):e39238. [doi: 10.7759/cureus.39238] [Medline: 37337480]
26. Walters WH, Wilder EI. Fabrication and errors in the bibliographic citations generated by ChatGPT. Sci Rep 2023 Sep 7;13(1):14045. [doi: 10.1038/s41598-023-41032-5] [Medline: 37679503]
27. Zhou L, Schellaert W, Martínez-Plumed F, Moros-Daval Y, Ferri C, Hernández-Orallo J. Larger and more instructable language models become less reliable. Nature 2024 Oct;634(8032):61-68. [doi: 10.1038/s41586-024-07930-y] [Medline: 39322679]
28. Gravel J, D'Amours-Gravel M, Osmanlliu E. Learning to fake it: limited responses and fabricated references provided by ChatGPT for medical questions. Mayo Clin Proc Digit Health 2023;1(3):226-234. [doi: 10.1016/j.mcpdig.2023.05.004] [Medline: 40206627]
29. McGowan A, Gui Y, Dobbs M, et al. ChatGPT and Bard exhibit spontaneous citation fabrication during psychiatry literature search. Psychiatry Res 2023 Aug;326:115334. [doi: 10.1016/j.psychres.2023.115334] [Medline: 37499282]

Abbreviations

HIPAA: Health Insurance Portability and Accountability Act
ICC: intraclass correlation coefficient
LLM: large language model

Edited by P Kanzow; submitted 02.Jan.2026; peer-reviewed by A Abou-Bakr, K Giannakopoulos, MK Hrenczuk, PCA Watanabe, S Cui; revised version received 12.Jul.2026; accepted 12.Jul.2026; published 06.Aug.2026.

Please cite as:
Hung M, Ward C, Cohen O, Marx J, Smit Z, Newman J, Lipsky MS
Prompt Framing and Evidence Requirement for AI-Generated Educational Responses in Dental Education: Experimental Study
JMIR Med Educ 2026;12:e90736
URL: https://mededu.jmir.org/2026/1/e90736
doi:10.2196/90736

© Man Hung, Corban Ward, Owen Cohen, Jacob Marx, Zachary Smit, Jacob Newman, Martin S Lipsky. Originally published in JMIR Medical Education (https://mededu.jmir.org), 6.Aug.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Education, is properly cited. The complete bibliographic information, a link to the original publication on https://mededu.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Human-Edited Generative AI-Assisted Multiple-Choice Questions in Postgraduate Family Medicine: Blinded Cross-Sectional Comparative Psychometric Study

Kai Ping Sze[1,2], MBBS, MMed, FCFP; Jia Qing Lim[2], MBBS, MMed; Yng Miin Loke[3], PhD; Teck Yong Gabriel Ding[2], MBBS, MMed, FCFP; Wei Chieh Wee[1], MBBS, MMed, FCFP; Tongyuan Chen[2], MBBS, MMed, FCFP; Jingkai Lawrence Wu[2], MBBS, MMed; Vivek Bansal[2], MBBS, MMed, FCFP; Qi Wei Fong[2*], MBBS, MMed; Minyang Chow[1,3,4*], MBBS, MMed, MHPE

[1]LKC School of Medicine, Nanyang Technological University, Singapore, Singapore

[2]NHG Polyclinics, Population Health Campus, National Healthcare Group, 690 Yishun Ring Road, Singapore, Singapore

[3]Clinical Education, National Healthcare Group, Singapore, Singapore

[4]Professional Training and Standards Division, Ministry of Health, Singapore, Singapore, Singapore

[*]these authors contributed equally

Corresponding Author:
Kai Ping Sze, MBBS, MMed, FCFP
LKC School of Medicine, Nanyang Technological University, Singapore, Singapore

Abstract

Background: Generative artificial intelligence (GenAI) is increasingly used to draft multiple-choice questions (MCQs) for health professions education, but much evidence concerns raw model outputs, expert ratings, or item difficulty alone. Educators edit GenAI drafts before use, and whether such items are psychometrically ready for postgraduate assessment remains unclear.

Objective: This study aimed to compare human-edited GenAI-assisted and educator-crafted MCQs for postgraduate Family Medicine Applied Knowledge Test–level assessment, examining difficulty, discrimination, reliability, distractor functioning, and participant perceptions.

Methods: We conducted a blinded cross-sectional, within-participant comparative psychometric evaluation in Singapore. Sixty best-of-five single-best-answer MCQs were evaluated, 30 human-edited GenAI-assisted items and 30 educator-crafted items, topic-matched across postgraduate FM domains and randomized across 2 assessment sets. Eligible participants were postgraduate doctors enrolled in FM residency or postgraduate family medicine programs, preparing for the Applied Knowledge Test, and blinded to item origin; incomplete paired responses were excluded. Outcomes included paired total scores, score correlation and agreement, Kuder-Richardson Formula 20 reliability, item difficulty index, corrected point-biserial discrimination, distractor functioning, and perceived difficulty, clarity, and relevance. Analyses used paired-sample tests, Pearson correlation, Fisher exact tests, and item-level psychometric statistics, with α=.05 and Bonferroni correction within comparison families.

Results: Of 74 participants, 73 completed both item sets and were included in the analysis. The final sample comprised 36 graduate diploma in FM trainees, 5 MMed FM trainees, and 32 FM residents. Paired-sample testing showed lower scores on GenAI-assisted than educator-crafted items (mean 19.12, SD 2.83 vs mean 21.10, SD 3.42 out of 30; mean difference −1.97, 95% CI −2.72 to −1.23; *P*<.001; Cohen *d*=0.62), indicating that GenAI-assisted items were not easier. Scores were positively correlated (*r*=0.49, 95% CI 0.30 - 0.64; *P*<.001), but Bland-Altman analysis indicated limited agreement. Kuder-Richardson Formula 20 reliability was lower for GenAI-assisted items (0.38 vs 0.60). Mean difficulty index did not differ significantly (0.64 vs 0.70; mean difference −0.07, 95% CI −0.19 to 0.06; *P*=.29), and more GenAI-assisted items fell within the acceptable difficulty range (18/30, 60.0% vs 13/30, 43.3%). However, mean corrected point-biserial discrimination was lower for GenAI-assisted items (0.09 vs 0.18; mean difference −0.08, 95% CI −0.16 to −0.01; *P*=.04), and negative discrimination was more common (6/30, 20% vs 3/30, 10%). GenAI-assisted items also had more nonfunctioning and negatively discriminating distractors, although these differences were not statistically significant. Participant ratings of perceived difficulty, clarity, and practice relevance did not differ by origin.

Conclusions: Human-edited GenAI-assisted MCQs can achieve plausible difficulty, but difficulty and surface acceptability did not ensure assessment readiness. Using trainee response data, this study extends work on raw outputs or expert opinion. GenAI should be used as a drafting adjunct within educator-led workflows prioritizing key verification, distractor engineering, pilot testing, empirical item analysis, and repair before item-bank or summative use.

XSL•FO
RenderX

(JMIR Med Educ 2026;12:e100179) doi:10.2196/100179

KEYWORDS

generative artificial intelligence; ChatGPT; multiple-choice questions; medical education; family medicine; assessment; psychometrics; item analysis; distractor analysis

Introduction

The Challenge of MCQ Development in Postgraduate Assessment

High-quality multiple-choice questions (MCQs) remain a cornerstone of written assessment in medical education. When well-constructed, single-best-answer MCQs can sample the curriculum broadly, be administered efficiently to large cohorts, and generate data suitable for psychometric analysis, thereby supporting meaningful score interpretation. However, defensible MCQ development is resource-intensive. It requires clinical expertise, assessment literacy, blueprint alignment, careful phrasing, plausible distractor construction, and iterative review. Programs therefore frequently struggle to scale item writing while maintaining item quality, especially where curricula evolve rapidly and assessment banks require regular renewal.

Postgraduate family medicine (FM) magnifies this challenge because assessments must sample acute, chronic, preventive, pediatric, geriatric, and palliative care while testing guideline-concordant management and applied judgment in undifferentiated primary care. In Singapore, the Family Medicine Applied Knowledge Test (FM AKT) is a national-level postgraduate written assessment comprising 180 single-best-answer MCQs across 2 papers [1].

In this setting, item quality affects more than content coverage. Ambiguous wording, unclear constructs, contestable keys, or poorly functioning distractors can distort score interpretation, reduce reliability, and undermine fairness, especially where scores may inform progression, remediation, or readiness for further training.

GenAI-Assisted MCQ Development and Emerging Concerns

Generative artificial intelligence (GenAI) has attracted increasing interest as a potential solution to the item-writing burden. Large language models can rapidly generate clinical stems, answer options, explanations, and item variants, offering a scalable way to support educators in developing practice questions or refreshing item banks. Early empirical studies suggest that GenAI-generated MCQs can, in some settings, appear clinically plausible, receive favorable expert ratings, and achieve psychometric performance broadly comparable with human-authored items, particularly when outputs undergo expert review for factual accuracy, relevance, and alignment with learning objectives [2-7].

However, the emerging literature remains mixed. Recent syntheses have highlighted substantial heterogeneity in prompting methods, model versions, evaluation criteria, and the types of validity evidence reported [8,9]. Some studies have raised concerns that GenAI-generated MCQs may be easier than educator-authored items or more likely to assess lower-order cognitive processes, while others suggest more comparable performance after careful prompting and expert review [2-7]. This creates an important tension for assessment leaders: GenAI may accelerate item production, but it remains unclear whether the resulting items are sufficiently robust for assessment use.

A further concern is that GenAI-generated MCQs may contain subtle vulnerabilities that are not obvious during surface review. GenAI can produce fluent stems and plausible-sounding answer options, but fluency is not the same as assessment quality. Effective MCQs require more than a correct key and grammatically acceptable options. They require distractors that represent authentic misconceptions, realistic clinical reasoning errors, or plausible near-miss decisions. If distractors are too implausible, they may become nonfunctioning. If they are ambiguous or conceptually misaligned, they may attract higher-performing candidates and reduce item discrimination. These risks are particularly relevant because recent work has described format-specific pitfalls in using MCQs to evaluate GenAI-generated educational content, including cueing, overestimation of reasoning, and "hallucination-adjacent" distractors [10].

From Raw AI Output to Human-Edited Assessment Items

A central practice-relevant gap is that most real-world implementations do not use raw GenAI outputs directly. Instead, educators typically use GenAI to generate draft items and then edit those items for clinical correctness, local guideline concordance, phrasing, relevance, and alignment with intended learning outcomes. The practical question is therefore not simply whether GenAI can generate MCQs. It is whether human-edited GenAI-assisted MCQs function equivalently to educator-crafted items when answered by the intended learner population.

This distinction matters because routine human editing may correct visible problems without resolving deeper psychometric vulnerabilities. An educator may be able to improve factual accuracy, remove obvious wording flaws, and ensure that the keyed answer is clinically defensible. However, these steps do not necessarily guarantee that the item discriminates between stronger and weaker candidates, that all distractors function as intended, or that the item contributes positively to internal consistency reliability. In other words, an item may be difficult enough and appear relevant to trainees, yet still be unsuitable for item-bank use if its option set does not support defensible score interpretation.

This distinction also reframes the concern that GenAI-generated MCQs may be "too easy" or too low-order. Difficulty is an important property, but it is not synonymous with assessment readiness. A recall-based item can be difficult if it tests obscure knowledge, while an applied clinical item can be easy if the decision is obvious. Conversely, a difficult item may still perform poorly if the construct is unclear, the key is contestable,

XSL•FO
RenderX

or distractors create ambiguity. For GenAI-assisted MCQ development, the critical question is therefore 2-fold: whether edited GenAI-assisted items can achieve appropriate postgraduate difficulty, and whether items that achieve such difficulty also demonstrate adequate discrimination and distractor functioning.

Study Aim and Contribution

In this study, we conducted a blinded psychometric comparison of human-edited GenAI-assisted and educator-crafted MCQs targeted at the postgraduate FM AKT level. Using trainee response data, we compared participant total scores, internal consistency reliability, item difficulty, corrected point-biserial discrimination, distractor functioning, and participant postitem ratings of perceived difficulty, clarity, and relevance.

The study was designed to address 2 linked questions. First, are human-edited GenAI-assisted MCQs necessarily too easy for postgraduate FM assessment? Second, if such items achieve plausible difficulty, do they demonstrate sufficient discrimination and distractor functioning to support assessment use? By evaluating item functioning rather than relying solely on surface plausibility or expert judgment, this study contributes empirical validity evidence for integrating GenAI into postgraduate MCQ development.

Accordingly, the primary aim of this study was to determine whether human-edited GenAI-assisted MCQs differ from educator-crafted MCQs in postgraduate FM AKT–level psychometric performance. The primary objective was to compare participant total scores between item origins. Secondary objectives were to compare internal consistency reliability, item difficulty, corrected point-biserial discrimination, distractor functioning, and participant postitem ratings of perceived difficulty, clarity, and relevance. We hypothesized that human-edited GenAI-assisted MCQs would not necessarily be easier than educator-crafted MCQs but might show residual vulnerabilities in discrimination and distractor functioning despite educator editing.

Methods

Study Design and Setting

We conducted a blinded cross-sectional, within-participant comparative psychometric evaluation of best-of-five single-best-answer MCQs targeted at the postgraduate FM AKT level. The study was conducted in Singapore in the postgraduate FM AKT preparation context after Domain Specific Review Board approval (Ethics and Compliance Online System 2025‐0105). The primary comparative factor was item origin: human-edited GenAI-assisted items versus educator-crafted items.

The study evaluated 60 MCQs in total: 30 human-edited GenAI-assisted items and 30 educator-crafted items. Items were topic-matched at the topic or domain level, with 1 GenAI-assisted item and 1 educator-crafted item developed for each of 30 postgraduate FM topics. The complete topic blueprint is provided in Table S3 in Multimedia Appendix 1. This matching strategy was used to balance curriculum coverage and improve content comparability between item origins. However, matched items were not designed as identical parallel forms and could differ in clinical scenario, decision point, cognitive process, and option structure. Items were organized into 2 assessment sets and presented in randomized order to reduce ordering effects and minimize the likelihood that participants could infer item origin.

The evaluation focused on empirical trainee response data. Outcomes were analyzed at the participant level, test level, item level, distractor level, and response process level. Participant-level outcomes included total scores on GenAI-assisted and educator-crafted items. Test-level outcomes included internal consistency reliability. Item-level outcomes included difficulty and discrimination. Distractor-level outcomes included nonfunctioning distractors and negatively discriminating distractors. Response process outcomes included participant postitem ratings of perceived difficulty, clarity, and relevance to day-to-day FM practice. The overall study design, item origins, matching strategy, blinding, and outcome levels are summarized in Table 1.

XSL•FO
RenderX

Table . Study design and evaluation framework for a blinded comparative psychometric evaluation of human-edited, GenAI[a]-assisted, and educator-crafted postgraduate FM AKT[b]-level MCQs[c] in Singapore.

Domain	Description
Study design	Blinded comparative psychometric evaluation
Total MCQs evaluated	60 best-of-five single-best-answer MCQs
Item origins	30 human-edited GenAI-assisted items; 30 educator-crafted items
Matching strategy	Topic- or domain-matched for curriculum coverage, with 1 GenAI-assisted and 1 educator-crafted item per topic; items were not designed as identical parallel forms
Assessment context	Postgraduate FM AKT-level assessment
Presentation	Items randomized across 2 sets
Blinding	Participants blinded to item origin
Participant-level outcomes	Total score on GenAI-assisted items; total score on educator-crafted items
Test-level outcomes	KR-20[d] internal consistency reliability; Spearman-Brown reliability projection
Item-level outcomes	Difficulty index; corrected point-biserial discrimination
Distractor-level outcomes	Nonfunctioning distractors; negatively discriminating distractors
Response process outcomes	Participant ratings of perceived difficulty, clarity, and relevance

[a]GenAI: generative artificial intelligence.

[b]FM AKT: Family Medicine Applied Knowledge Test.

[c]MCQs: multiple-choice questions.

[d]KR-20: Kuder-Richardson Formula 20.

Inclusion and Exclusion

Participants were postgraduate medical doctors preparing for the FM AKT in Singapore. Eligible participants included doctors enrolled in FM residency within 1 of the 3 public health care institutions and/or doctors enrolled in postgraduate FM programs under the College of Family Physicians Singapore. Participants who withdrew consent or did not complete both MCQ sets were excluded from the complete-case analytic sample.

Sampling Procedures

Participants were recruited by email invitation using convenience sampling with snowball dissemination from January 6 to April 14, 2026, during the 2026 AKT preparation period. Recruitment invitations were disseminated through email, WhatsApp groups, and peer sharing to postgraduate doctors preparing for the FM AKT in Singapore. Participation was voluntary, and participants provided informed consent before data collection.

Sample Size, Power, and Precision

A priori sample size estimation was performed using G*Power software (version 3.1; Heinrich Heine University Düsseldorf) for the participant-level comparison of total scores between GenAI-assisted and educator-crafted items. As each participant answered both item types, the primary participant-level comparison was planned using a 2-tailed paired-sample *t* test, with α=.05 and power=0.80. Assuming a moderate within-participant standardized effect size of dz=0.40, 52 complete paired participants were required. A smaller effect size of dz=0.30 would require approximately 90 complete paired participants to retain 80% power. We therefore set a recruitment target range of 52-90 participants. The final analytic sample of 73 complete paired responses exceeded the minimum required sample for detecting a moderate paired difference in total scores. Item-level and distractor-level analyses were based on 30 items per origin and were interpreted as exploratory, with emphasis placed on the convergence of findings across difficulty, discrimination, reliability, and distractor-functioning indicators rather than on isolated *P* values.

Participant Characteristics

Participant characteristics available for analysis were limited to postgraduate training pathway and residency year. Age, sex, years of clinical experience, prior AKT attempt status, and prior GenAI use were not collected in the analytic dataset. Participant characteristics for the complete-case participants are summarized in Table 2.

XSL•FO
RenderX

Table . Participant characteristics of the 73 postgraduate family medicine (FM) doctors in Singapore included in the complete-case analysis.

Postgraduate program	Values, n (%)
Graduate diploma in FM	36 (49.3)
College Master of Medicine in FM program	5 (6.8)
FM residency year 1	8 (11.0)
FM residency year 2	21 (28.8)
FM residency year 3	3 (4.1)
Total	73 (100)

Measures, Covariates, and MCQ Development Workflow

Overview

The study materials comprised 60 best-of-five single-best-answer MCQs blueprint-aligned to postgraduate FM AKT–level practice. The primary exposure was item origin: human-edited GenAI-assisted versus educator-crafted. Items sampled clinical domains relevant to FM, including pediatric, adult, and geriatric care, and acute, chronic, preventive, and palliative care. No participant demographic or clinical covariates were included in the primary analyses because the study focused on paired item-origin comparisons and item-level psychometric functioning.

The item-development and review group comprised 3 clinician-educator authors (KPS, JQL, and QWF) and a 5-member expert panel (TYGD, WCW, TC, JLW, and VB) with FM clinical and postgraduate assessment experience. These 8 reviewers provided subject-matter review for clinical accuracy, relevance, and alignment with postgraduate FM standards.

Educator-Crafted MCQs

Thirty MCQs were developed by clinician-educators using the study blueprint, a sample AKT-style question, and MCQ-writing guidance (Annex S2 in Multimedia Appendix 1). Educator-crafted items were written to assess applied FM, knowledge, clinical judgment, and guideline-concordant management. Items were reviewed by the study team before inclusion to ensure alignment with the intended topic, answer key, and single-best-answer format.

GenAI-Assisted MCQs

Thirty MCQs were generated using ChatGPT-5.1 (OpenAI) [11] and subsequently edited by the study team. GenAI-assisted items were created using tailored prompts that specified the clinical topic, intended learning objective, expected AKT-style single-best-answer format, and level of applied clinical reasoning (Annex S1 in Multimedia Appendix 1). The same sample question and MCQ-writing guidance provided to educator item writers were also provided to the GenAI system.

The GenAI outputs were then reviewed and edited by the same study team for clinical correctness, local guideline contextualization, phrasing, single-best-answer structure, and alignment with intended learning outcomes. Each GenAI-assisted item underwent 1 structured educator review-and-editing round, followed by final study team checking before administration. Editors were not blinded to item origin during item development because the GenAI-assisted items were generated and edited through a prespecified human-in-the-loop workflow; blinding was applied at the participant-assessment stage and not at the item-editing stage. The detailed GenAI-prompting and human-editing framework is provided in Table S2 in Multimedia Appendix 1. The final GenAI-assisted items therefore represent a human-in-the-loop GenAI-assisted workflow, rather than raw artificial intelligence (AI)–generated outputs. The overall human-in-the-loop workflow for developing, reviewing, administering, and evaluating the GenAI-assisted MCQs is summarized in Table 3.

A total of 30 MCQs per item origin were selected pragmatically to balance sufficient content sampling across the FM AKT blueprint with feasibility for both item development and participant testing. A larger item set may have provided more stable item-level psychometric estimates but would have increased educator workload for item writing, review, and standardization, and participant fatigue during test completion. Given that the study was designed as an exploratory blinded psychometric evaluation rather than a high-stakes summative assessment or definitive item calibration study, 30 items per origin were considered an appropriate balance between breadth, feasibility, and response quality.

XSL•FO
RenderX

Table . Human-in-the-loop workflow[a] for developing, reviewing, administering, and evaluating GenAI[b]-assisted MCQs[c].

Stage	Workflow
1. Blueprint and topic selection	AKT[d] topic, learning objective, and expected reasoning level defined by educators.
2. Structured GenAI prompting	Topic, single-best-answer format, clinical context, sample question, and item-writing guidance supplied to ChatGPT-5.1
3. Draft item generation	Stem, keyed answer, distractors, and explanation generated for educator review
4. Educator review and editing	Clinical accuracy, local guideline concordance, phrasing, key defensibility, and single-best-answer structure checked
5. Option-set engineering	Distractors reviewed for authentic misconceptions, common reasoning errors, homogeneity, cueing, and near-miss plausibility
6. Blinded trainee administration	Randomized items administered to participants blinded to item origin
7. Empirical item analysis	Difficulty, discrimination, reliability, distractor functioning, and participant ratings reviewed
8. Item-bank decision	Accept, revise, quarantine, or retire items based on expert and empirical evidence

[a]Quality assurance principle: expert-judged item plausibility alone is insufficient; empirical item–functioning evidence is also required.

[b]GenAI: generative artificial intelligence.

[c]MCQs: multiple-choice questions.

[d]AKT: Applied Knowledge Test.

Quality of Measurements and Instrumentation

Measurement quality was supported through blueprint alignment, topic matching, and use of the same subject-matter review for both item origins. The postitem ratings of perceived difficulty, clarity, and relevance were study-specific 5-point Likert-type items embedded after each MCQ and used as response process evidence rather than as validated psychometric scales. Objective item functioning was evaluated separately using difficulty index, corrected point-biserial discrimination, Kuder-Richardson Formula 20 (KR-20) reliability, and distractor analysis.

Data Collection and Masking

Participants completed both MCQ sets during the study assessment. Items were presented in randomized order across the 2 sets to reduce ordering effects. Participants were not informed which items were GenAI-assisted or educator-crafted, which served as participant masking to reduce expectancy effects and prevent assumptions about item origin from influencing responses or postitem ratings.

Immediately after answering each MCQ, participants completed brief item-level appraisals. These included ratings of perceived difficulty, clarity of the stem and options, and relevance to day-to-day FM practice. Ratings were collected using embedded Likert-type response options after each question. Participants were also provided with an optional free-text field to flag confusing wording or other item-specific concerns.

Conditions and Design

This was a nonclinical, noninterventional, and blinded comparative psychometric evaluation. The primary comparison condition was item origin: human-edited GenAI-assisted versus educator-crafted MCQs. Participants were not randomized to intervention groups; all complete-case participants answered both item types, allowing paired participant-level comparisons. Item-level and distractor-level analyses were conducted by item origin and interpreted as exploratory because the topic-matched items were not designed as identical parallel forms.

Measures and Outcomes

Outcomes were analyzed across 5 levels: participant-level performance, test-level reliability, item-level psychometric functioning, distractor-level functioning, and participant postitem ratings. Definitions and interpretation thresholds for the study outcomes are provided in Table S1 in Multimedia Appendix 1.

Participant-Level Outcomes

The primary participant-level outcomes were total score on GenAI-assisted items and total score on educator-crafted items. Each participant therefore contributed paired scores, one for each item origin.

Test-Level Outcomes

Internal consistency reliability was calculated separately for GenAI-assisted and educator-crafted item sets using the KR-20. Spearman-Brown prophecy estimates were used to project how reliability would change with longer test length.

Item-Level Outcomes

Item difficulty was calculated for each MCQ using the difficulty index, defined as the proportion of participants who answered the item correctly. Higher values indicate easier items. Items were categorized as follows (Table 4).

Item discrimination was assessed using the corrected point-biserial correlation, which correlates performance on an individual item with total test score after excluding that item

XSL•FO
RenderX

from the total. Items with corrected point-biserial correlations ≥0.30 were considered to have good discrimination. Negative corrected point-biserial values were treated as red flags because they suggest that higher-performing participants were more likely to answer the item incorrectly.

Table . Difficulty index thresholds used to classify postgraduate FM AKT[a]-level MCQ[b] item difficulty in the blinded psychometric evaluation.

Difficulty index	Interpretation
$P \leq .25$	Too difficult
$.25 < P < .75$	Appropriate difficulty
$P \geq .75$	Too easy

[a]FM AKT: Family Medicine Applied Knowledge Test.

[b]MCQ: multiple-choice question.

Distractor-Level Outcomes

Distractor analysis was conducted for each item. Nonfunctioning distractors were defined as distractors selected by fewer than 5% of participants. Negatively discriminating distractors were defined as distractors with positive point-biserial correlations with total score, indicating that they were selected more frequently by higher-performing participants. For each item, the number of nonfunctioning distractors and negatively discriminating distractors was tallied from 0 to 4.

Response-Process Outcomes

Participant postitem ratings were analyzed as response process evidence. These included perceived item difficulty, clarity, and relevance to day-to-day FM practice. These ratings were used to assess whether trainees perceived GenAI-assisted and educator-crafted items differently, and whether subjective ratings aligned with objective item difficulty.

Data Sources

Data sources comprised primary participant item response and postitem rating data, the answer key, item origin labels, topic or domain mapping, and aggregate psychometric outputs generated from the study datasets. Expert review files informed item development and supplementary materials but were not participant outcome data.

Data Diagnostics and Missing Data

The analytic dataset used complete paired responses. One participant did not provide complete paired MCQ-set data and was excluded from the complete-case analysis (1/74, 1.4%). The final analytic sample contained no missing item-response data for the 73 included participants, so the Little Missing Completely at Random test and multiple imputation were not performed. Item-level statistics with no response variance were reported as not estimable where applicable.

Analytic Strategy

Total scores for GenAI-assisted and educator-crafted items were compared using a paired-sample test because each participant completed both item types. The mean difference, 95% CI, *P* value, and Cohen *d* were reported. Pearson correlation coefficient was used to assess the linear relationship between GenAI-assisted and educator-crafted total scores. Bland-Altman analysis was used to assess agreement between the total scores of GenAI-generated and educator-crafted items.

Internal consistency reliability was assessed separately for the GenAI-assisted and educator-crafted item sets using KR-20, which is equivalent to Cronbach α for dichotomously scored items. KR-20 values range from 0 to 1, with higher values indicating greater internal consistency [12]. The 95% CIs for KR-20 were calculated using bootstrap (bias-corrected and accelerated) with 1000 resamples. The Spearman-Brown prophecy formula was used to estimate the number of items that would be required to achieve conventional reliability thresholds, assuming comparable item properties.

Psychometric item analysis was conducted for each MCQ. The difficulty index was calculated as the proportion of participants who answered the item correctly. Corrected point-biserial correlations were calculated to assess item discrimination. Items with no response variance were excluded from point-biserial calculation because discrimination could not be estimated.

Distractor analysis was conducted by item origin. For each item, the number of nonfunctioning distractors and negatively discriminating distractors was counted. Fisher exact tests were used to compare distributions between GenAI-assisted and educator-crafted items because several expected cell counts were below 5.

Participant ratings of perceived difficulty, clarity, and relevance were recoded numerically for analysis. Mean ratings were calculated for each item and compared between GenAI-assisted and educator-crafted items. Correlations between subjective difficulty ratings and objective item difficulty were examined to assess whether participant perceptions aligned with empirical item performance.

Participant-level total score analyses were paired because each participant completed both GenAI-assisted and educator-crafted items. Item-level analyses were conducted by item origin using the 30 items in each group. Independent samples *t* tests were used for item-level comparisons. Although items were matched by clinical topic, GenAI-assisted and educator-crafted items were developed through fundamentally different processes—one by a large language model with subsequent human editing, the other by educators, and topic matching does not guarantee that paired items share residual variance, which is a key assumption underlying paired sample *t* test.

All statistical tests were interpreted at $\alpha = .05$. Bonferroni correction was applied within each family of comparisons: item-level psychometric analysis, distractor analysis, and

XSL•FO
RenderX

participant-rating analysis. For item-level psychometric analyses, 2 tests were conducted, giving a corrected α=.025. For distractor analyses, 2 tests were conducted, giving a corrected α=.025. For participant-rating analyses, 3 tests were conducted, giving a corrected α=.017. Statistical analyses were performed using R (version 4.4.2; R Core Team).

Ethical Considerations

Ethical approval for human participant research was obtained from the NHG Health Domain Specific Review Board through the Ethics and Compliance Online System (reference 2025 - 0105). No ethics exemption was relied on. Participants received study information before participation and provided informed consent before data collection. Participation was voluntary, participants could withdraw without penalty, and no compensation was provided. Study data were analyzed in aggregate after deidentification, and no participant-identifiable information is reported in the paper, figures, tables, or supplementary material. No participant data were entered into the GenAI system during item generation. GenAI was used to generate draft MCQs from study prompts, topic specifications, learning objectives, sample questions, and item-writing guidance. Final items were reviewed and edited by the study team before administration to participants.

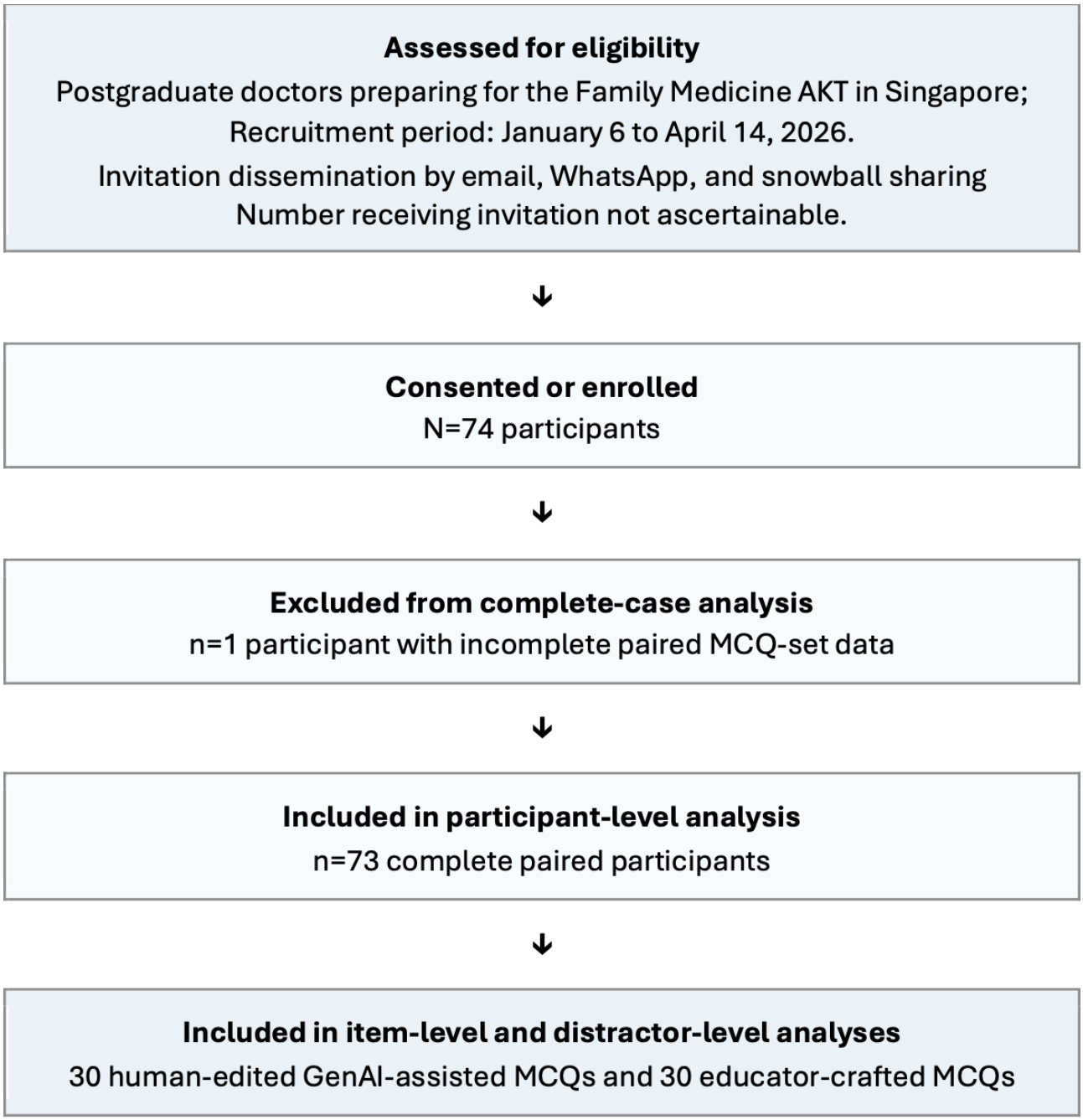
Results

Participant Flow

Recruitment was conducted from January 6 to April 14, 2026, during the 2026 FM AKT preparation period. The total number of doctors who received or viewed the invitation could not be determined because of the snowball recruitment approach. Of the 74 participants who consented and enrolled, 73 completed both MCQ sets and were included in the complete-case participant-level analysis. One participant was excluded because of incomplete paired MCQ-set data. Participant flow is shown in Figure 1.

XSL•FO
RenderX

Figure 1. Participant flow diagram for the blinded comparative psychometric evaluation of human-edited GenAI-assisted and educator-crafted postgraduate FM AKT–level MCQs in Singapore. FM AKT: Family Medicine Applied Knowledge Test; GenAI: generative artificial intelligence; MCQ: multiple-choice question.



Participant-Level Total Scores by Item Origin

Participants scored significantly lower on GenAI-assisted items (mean 19.12, SD 2.83) than on educator-crafted items (mean 21.10, SD 3.42), with a mean paired difference of −1.97 (95% CI −2.72 to −1.23; P<.001) and a moderate effect size (Cohen d=0.62) (Table 5; Figure S2 in Multimedia Appendix 1). This indicates that the GenAI-assisted item set was more difficult overall than the educator-crafted item set.

XSL•FO
RenderX

Table . Participant-level and item-level psychometric outcomes by item type in a blinded comparative psychometric evaluation of human-edited GenAI[a]-assisted and educator-crafted postgraduate FM AKT[b]-level MCQs[c] in Singapore.

Metric	GenAI-assisted, mean (SD)	Educator-crafted, mean (SD)	Mean difference	95% CI	*P* value
Total score	19.12 (2.83)	21.10 (3.42)	−1.97	−2.72 to −1.23	<.001
Difficulty index	0.64 (0.26)	0.70 (0.22)	−0.07	−0.19 to 0.06	.29
Corrected point-biserial correlation	0.09 (0.15)	0.18 (0.15)	−0.08	−0.16 to −0.01	.04[d]

[a]GenAI: generative artificial intelligence.

[b]FM AKT: Family Medicine Applied Knowledge Test.

[c]MCQs: multiple-choice questions.

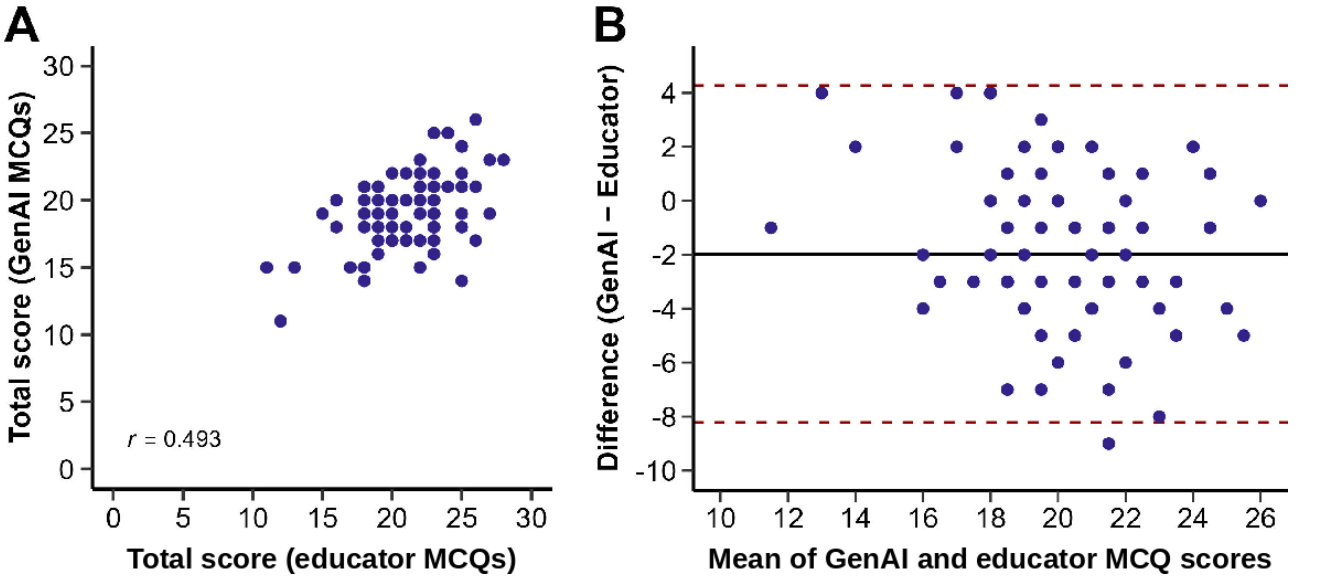
[d]Bonferroni-corrected significance thresholds were applied within each family of comparisons: α=.025 for psychometric item analysis (2 tests: difficulty index and corrected point-biserial correlation). The *P* value for corrected point-biserial correlation was below .05 but did not meet the Bonferroni-corrected threshold of α=.025 for this family of comparisons.

Score Correlation and Agreement

GenAI-assisted and educator-crafted total scores were positively correlated (r=0.49, 95% CI 0.30-0.65; P<.001) (Figure 2A), indicating that participants who performed well on 1 item set tended to perform well on the other. However, the wide confidence interval—spanning from a weak to moderate association—suggests limited precision in this estimate.

Bland-Altman analysis revealed a mean difference of −1.97 between GenAI-generated and educator-crafted items, with 95% limits of agreement ranging from −8.22 to 4.28 (Figure 2B). This indicates that for any individual participant, their GenAI-assisted MCQ test score could range from approximately 4 points higher to 8 points lower than their educator-crafted score, suggesting that the 2 test forms are not interchangeable at the individual level. Visual inspection suggested possible proportional bias: lower-scoring participants showed score differences closer to zero, while higher-scoring participants showed larger negative differences, suggesting that the GenAI-assisted items disproportionately penalized higher-performing participants.

Figure 2. Association and agreement between participants' total scores for the GenAI-assisted and educator-crafted multiple-choice question sets. (A) Scatterplot showing the correlation between GenAI-assisted and educator-crafted total scores. Each dot represents 1 participant. (B) Bland-Altman plot comparing total test scores of GenAI-assisted and educator-crafted MCQ items. Each dot represents 1 participant. The x-axis shows the mean of the 2 scores; the y-axis shows the difference (GenAI – Educator). The solid line indicates the mean difference, and the dashed lines indicate the 95% limits of agreement. GenAI: generative artificial intelligence; MCQs: multiple-choice questions.



Item Difficulty and Discrimination

At the item level, GenAI-assisted items were slightly more difficult on average than educator-crafted items, although this difference was not statistically significant. The mean difficulty index was 0.64 (SD 0.26) for GenAI-assisted items and 0.70 (SD 0.22) for educator-crafted items, with a mean difference of −0.07 (95% CI −0.19 to 0.06; P=.29; Cohen d=−0.28) (Table 5 and Figure 3A). Because higher difficulty index values indicate easier items, this item-level finding was consistent with the participant-level score comparison.

When items were classified by difficulty range, a larger proportion of GenAI-assisted items fell within the acceptable difficulty range of 0.25- 0.75. Eighteen GenAI-assisted items (60.0%) were classified as having appropriate difficulty, compared with 13 educator-crafted items (43.3%). Conversely,

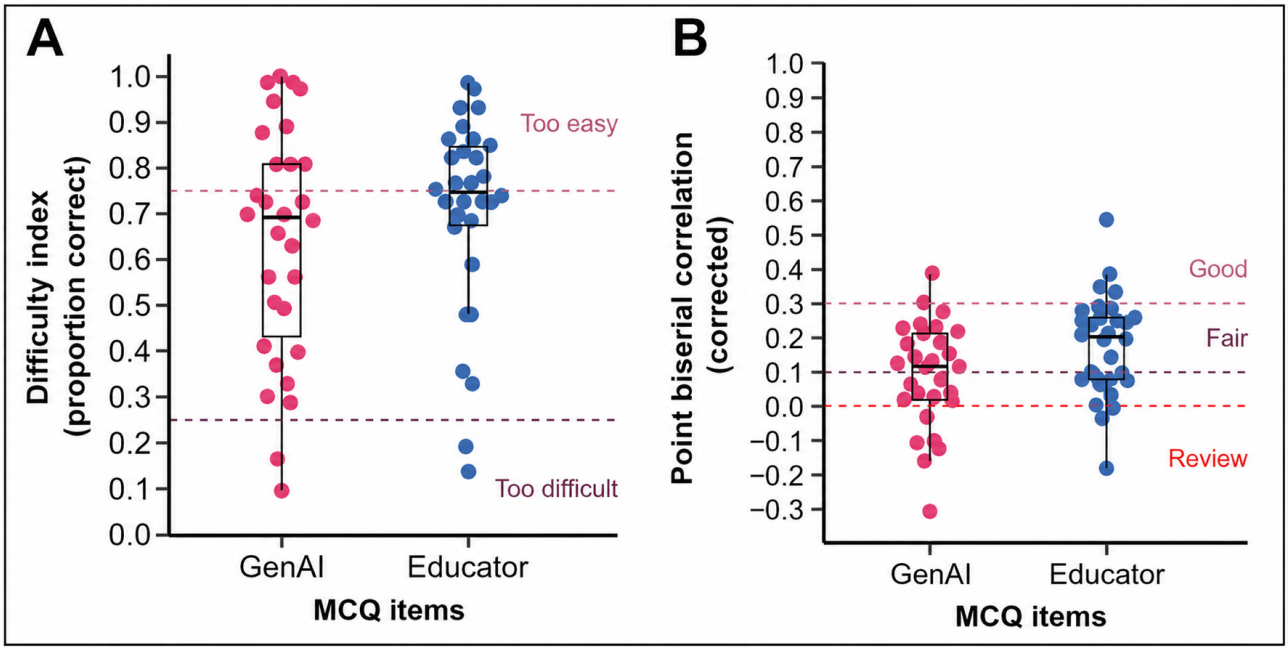
XSL•FO
RenderX

10 GenAI-assisted items (33.3%) and 15 educator-crafted items (50.0%) were classified as too easy. Two items from each set (6.7%) were classified as too difficult. One GenAI-assisted item on diabetes mellitus was answered correctly by all participants; because there was no response variance, its corrected point-biserial correlation could not be computed.

Despite achieving plausible difficulty, GenAI-assisted items showed weaker discrimination. The mean corrected point-biserial correlation was 0.09 (SD 0.15) for GenAI-assisted items and 0.18 (SD 0.15) for educator-crafted items. The mean difference was −0.08 (95% CI −0.16 to −0.01; *P*=.04; Cohen *d*=−0.56), which was below 0.05 but did not meet the Bonferroni-corrected threshold of α=.025 for this family of comparisons (Table 5 and Figure 3B).

Only 2 GenAI-assisted items (6.7%) achieved good discrimination, defined as corrected point-biserial correlation ≥0.30, compared with 4 educator-crafted items (13.3%). More concerningly, 6 GenAI-assisted items (20.0%) had negative corrected point-biserial correlations, compared with 3 educator-crafted items (10.0%). Negative corrected point-biserial values indicate that higher-performing participants were more likely to answer these items incorrectly, suggesting that these items require review before reuse. Complete item-level psychometric results by topic and item source are provided in Table S4 in Multimedia Appendix 1.

Figure 3. Item difficulty and discrimination by item origin for 30 human-edited GenAI-assisted and 30 educator-crafted postgraduate Family Medicine Applied Knowledge Test–level MCQs in Singapore. (A) Item difficulty measured by difficulty index; higher values indicate easier items. Dashed lines show the acceptable difficulty range of 0.25-0.75. (B) Corrected point-biserial discrimination; higher values indicate stronger discrimination. Dashed lines indicate thresholds for good, fair, and review-level discrimination. Each dot represents 1 item. GenAI: generative artificial intelligence; MCQ: multiple-choice question.



Internal Consistency Reliability

Both item sets demonstrated suboptimal internal consistency reliability, although reliability was lower for GenAI-assisted items (KR-20=0.38, 95% CI 0.12-0.56) than for educator-crafted items (KR-20=0.60, 95% CI 0.44-0.73).

Alpha-if-item-deleted analysis showed that removing 1 GenAI-assisted ethics item increased the GenAI-assisted KR-20 from 0.38 to 0.45, an increase of approximately 18%. No single educator-crafted item had a comparable effect on the educator-crafted item set reliability (Figure S3 in Multimedia Appendix 1).

As internal consistency reliability is influenced by test length, the Spearman-Brown prophecy formula was used to project reliability for longer assessments. Assuming comparable item properties, a 200-item test would increase projected KR-20 to 0.80 for GenAI-assisted items and 0.91 for educator-crafted items. To achieve a KR-20 of 0.70, an estimated 116 GenAI-assisted items and 48 educator-crafted items would be required. These projections should be interpreted as illustrative because they assume that additional items would have similar psychometric properties to those observed in this study.

Distractor Functioning

Distractor analysis was conducted to explore whether option-set quality contributed to weaker discrimination among GenAI-assisted items. A higher proportion of GenAI-assisted items had multiple nonfunctioning distractors and negatively discriminating distractors, although these comparisons did not reach statistical significance (Fisher exact *P*=.46 and *P*=.41, respectively).

Seventeen GenAI-assisted items (56.7%) had 3 or 4 nonfunctioning distractors, compared with 13 educator-crafted items (43.3%) (Table 6 and Figure 4A). Fourteen GenAI-assisted items (46.7%) contained 1 or 2 negatively discriminating distractors, compared with 9 educator-crafted items (30.0%) (Table 6 and Figure 4B).

XSL•FO
RenderX

Items with negative corrected point-biserial correlations frequently co-occurred with poor distractor functioning. Among the 6 GenAI-assisted items with negative corrected point-biserial correlations, 5 had at least 3 nonfunctioning distractors and 4 had 1 or 2 negatively discriminating distractors. Among the 3 educator-crafted items with negative corrected point-biserial correlations, 1 had 3 nonfunctioning distractors and 2 had 1 or 2 negatively discriminating distractors. This pattern suggests that weak discrimination may be partly explained by option-set problems, particularly distractors that were either rarely selected or selected disproportionately by higher-performing participants.

Table . Item classification and distractor-functioning red flags by item origin for 30 human-edited GenAI[a]-assisted and 30 educator-crafted postgraduate FM AKT[b]–level MCQs[c] in Singapore.

Item-level classification	GenAI-assisted, n (%)	Educator-crafted, n (%)
Appropriate difficulty, 0.25<*P*<.75	18 (60)	13 (43.3)
Too easy, *P*≥.75	10 (33.3)	15 (50)
Too difficult, *P*≤.25	2 (6.7)	2 (6.7)
Good discrimination, rpb(c)[d]≥0.30	2 (6.7)	4 (13.3)
Negative discrimination, rpb(c)<0	6 (20)	3 (10)
3 or 4 nonfunctioning distractors	17 (56.7)	13 (43.3)
1 or 2 negatively discriminating distractors	14 (46.7)	9 (30)

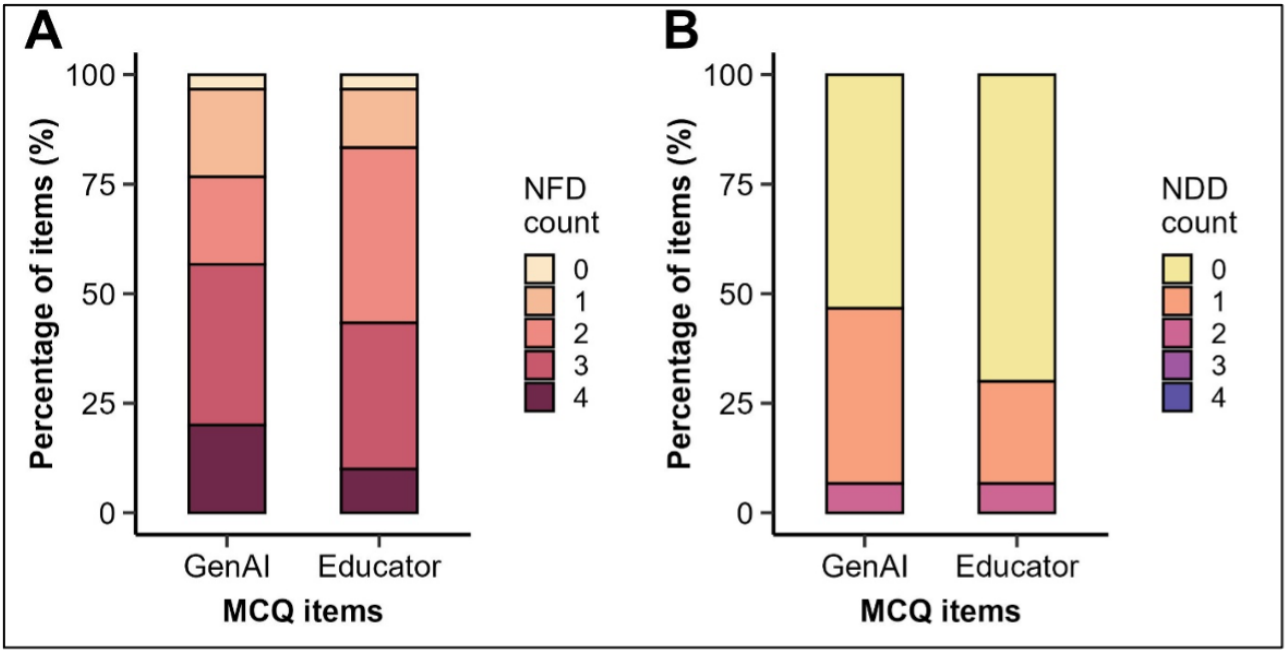
[a]GenAI: generative artificial intelligence.

[b]FM AKT: Family Medicine Applied Knowledge Test.

[c]MCQs: multiple-choice questions.

[d]rpb(c): corrected point-biserial correlation coefficient.

Figure 4. Distractor-functioning red flags by item origin for 30 human-edited GenAI-assisted and 30 educator-crafted postgraduate Family Medicine Applied Knowledge Test–level MCQs in Singapore. (A) Distribution of nonfunctioning distractors per item; nonfunctioning distractors were selected by fewer than 5% of participants. (B) Distribution of negatively discriminating distractors per item; negatively discriminating distractors were selected more frequently by higher-performing participants. Shading indicates the number of problematic distractors per item. GenAI: generative artificial intelligence; MCQ: multiple-choice question; NDD: negatively discriminating distractor; NFD: nonfunctioning distractor.



Participant Postitem Ratings

Participant postitem ratings did not differ significantly between GenAI-assisted and educator-crafted items. Mean ratings showed no significant difference by item origin for perceived difficulty (mean difference 0.06, 95% CI −0.12 to 0.23; *P*=.51), clarity (mean difference −0.03, 95% CI −0.08 to 0.03; *P*=.38), or relevance to day-to-day FM practice (mean difference 0.08, 95% CI −0.05 to 0.22; *P*=.20). None of these comparisons met the Bonferroni-corrected threshold of α=.017 for the participant-rating analysis family.

Subjective difficulty ratings were associated with objective item difficulty. For both GenAI-assisted items and educator-crafted items, mean perceived difficulty ratings were negatively correlated with difficulty index (*r*=−0.72, 95% CI −0.86 to

XSL•FO
RenderX

−0.49; *P*<.001 and *r*=−0.54, 95% CI −0.76 to −0.23; *P*=.002, respectively), indicating that items rated as harder were answered correctly by fewer participants (Figure S4 in Multimedia Appendix 1). These findings suggest that participants did not perceive GenAI-assisted items as systematically less clear, less relevant, or more difficult than educator-crafted items.

Discussion

Principal Findings

This blinded comparative psychometric evaluation provides practice-relevant evidence on the use of human-edited GenAI-assisted MCQs, rather than raw GenAI outputs, in a postgraduate FM AKT–level context. The central finding is that GenAI-assisted items were not easier than educator-crafted items. Participants scored significantly lower on the GenAI-assisted item set, mean item difficulty did not differ significantly between item origins, and a greater proportion of GenAI-assisted items fell within the acceptable difficulty range. These findings challenge the assumption that GenAI-assisted MCQs are inherently too easy when they are generated using AKT-aligned prompts and subsequently edited by educators.

However, plausible difficulty did not translate into equivalent assessment functioning. Compared with educator-crafted items, GenAI-assisted items showed lower internal consistency reliability, weaker corrected point-biserial discrimination, fewer items with good discrimination, more items with negative discrimination, and more distractor-functioning red flags. Although several item-level comparisons did not meet Bonferroni-corrected thresholds, the direction of findings was consistent across complementary indicators. The interpretation of borderline findings in this study warrants explicit discussion. The corrected point-biserial discrimination comparison yielded *P*=.04, which fell below the conventional α=.05 threshold but did not meet the Bonferroni-corrected threshold of α=.025 for the psychometric item analysis family. Bonferroni correction was applied to reduce the risk of false positives across multiple comparisons, which is statistically appropriate. However, because the study was not powered specifically for all item-level comparisons, this conservative adjustment also increases the risk of overlooking genuine differences, particularly where effect sizes are small to moderate and item samples are limited to 30 per group. The mean difference in corrected point-biserial correlation of −0.08 (95% CI −0.16 to −0.01) represents a clinically and educationally meaningful gap, given that the GenAI-assisted mean of 0.09 falls in the poor to marginal range while the educator-crafted mean of 0.18 approaches the acceptable range. Rather than treating this finding as simply nonsignificant, we interpret it as a borderline result that is consistent in direction with the broader pattern of weaker reliability, more negative discrimination signals, and more distractor-functioning concerns observed across GenAI-assisted items. Taken together, the convergence of findings across multiple complementary indicators provides stronger cumulative support for the interpretation of residual psychometric vulnerability than any single comparison in isolation. These results suggest that the key concern with edited GenAI-assisted MCQs was not difficulty targeting but whether the items could reliably distinguish stronger from weaker candidates and whether distractors functioned as intended. Participant postitem ratings add a further important layer. Trainees did not perceive GenAI-assisted items as systematically less clear, less relevant, or more difficult than educator-crafted items. Subjective difficulty ratings also correlated with objective item difficulty, suggesting that participant ratings provided meaningful response process information. However, similar participant ratings across item origins did not preclude differences in psychometric functioning, including reliability, discrimination, and distractor performance.

Both item sets demonstrated suboptimal internal consistency reliability, which is partly attributable to the relatively small number of items in each set (30 items). Spearman-Brown projections suggest that longer sets would achieve acceptable to good reliability; assuming comparable item properties, a 200-item test would increase projected KR-20 to 0.80 for GenAI-assisted items and 0.91 for educator-crafted items [13]. However, the GenAI-assisted set would still require substantially more items than the educator-crafted set to reach KR-20=0.70, reinforcing the need for targeted item repair before use in formal assessment.

The low internal consistency of the GenAI-assisted items and their weaker item discrimination are psychometrically interrelated. A noisy total score serves as a less reliable anchor for point-biserial correlations, which may attenuate observed discrimination values. However, several findings from this study suggest that intrinsic item flaws are the primary driver, rather than a systemic symptom of lower internal consistency. First, distractor analysis revealed that many GenAI-assisted items contained nonfunctional or negatively discriminating distractors, indicating structural problems at the item level. Second, alpha-if-deleted analysis identified specific items whose removal improved KR-20, suggesting that particular problematic items depressed overall reliability rather than the reverse. Together, these findings indicate that weak discrimination stems primarily from deficiencies in distractor construction. Wide confidence intervals were observed across several key metrics, including item-level comparisons, correlations, and KR-20, reflecting the limited number of items (30 per group), which constrain the precision of between-group comparisons and strength of associations.

Comparison With Prior Work

Our finding that human-edited GenAI-assisted items were not empirically easier than educator-crafted items contrasts with several recent high-stakes and residency-based comparisons in which AI-generated items tended to be easier on average [2,7]. This difference may reflect our focus on a human-in-the-loop workflow with AKT-aligned prompting and educator editing prior to administration, which likely mitigates the “easy-item” tendency reported when AI outputs are used with less structured refinement [8]. At the same time, our broader pattern of weaker reliability, lower discrimination, and more distractor-functioning red flags despite similar perceived clarity and relevance aligns with emerging evidence that surface plausibility and perceived quality do not guarantee psychometric readiness, and that

XSL•FO
RenderX

distractor construction remains a persistent vulnerability in AI-assisted item writing [9,10]. Dhanvijay et al [14] similarly reported that AI-chatbot–generated undergraduate physiology MCQs contained more nonfunctioning distractors than faculty-written MCQs, but their AI items were easier and showed comparable discrimination rather than weaker discrimination. This contrast is useful because it suggests that distractor vulnerability may recur across contexts, while difficulty and discrimination patterns may depend on discipline, learner group, item source, and human-review workflow. Studies that report broadly comparable psychometric properties or detectability across AI and human items have typically emphasized the role of expert review and structured workflows [4-6], consistent with our interpretation that GenAI is best positioned as a drafting adjunct rather than an autonomous item writer. Taken together, the literature suggests that the key implementation question is shifting from whether GenAI are "too easy" to whether postediting and option-set engineering are sufficient to ensure discrimination, defensible keys, and functioning distractors at scale. Our study adds practice-relevant validity evidence for this question by evaluating edited GenAI-assisted items using trainee response data and distractor-level diagnostics in a broad postgraduate FM assessment context.

Interpretation and Limitations

GenAI-Assisted MCQs Can Achieve Difficulty, but Difficulty Is Not Validity

A key implication of this study is that GenAI-assisted MCQs should not be dismissed as inevitably easy or purely recall-based. In our postgraduate FM context, edited GenAI-assisted items were answered correctly less often than educator-crafted items, and more of them fell within the acceptable difficulty range. This suggests that, with structured prompting and human editing, GenAI-assisted workflows can produce items that are challenging enough for postgraduate learners.

However, empirical difficulty should not be mistaken for validity. Difficulty indicates how many candidates answered an item correctly; it does not show whether the item measures the intended construct, whether the keyed answer is uniquely defensible, or whether the item distinguishes stronger from weaker candidates. A recall item can be difficult if it tests obscure information, while an applied clinical item can be easy if the correct management decision is obvious. Conversely, a difficult item can still be psychometrically poor if it is ambiguous or dependent on distractors that do not function as intended [15-18].

This distinction is particularly important for GenAI-assisted item development. Large language models can generate fluent and clinically plausible MCQ stems, and human editing can improve factual correctness, local guideline alignment, and phrasing. Yet, these steps do not necessarily establish that the final item contributes to defensible score interpretation. In this study, GenAI-assisted items achieved plausible difficulty but showed weaker internal consistency and discrimination. The practical lesson is therefore not that GenAI cannot generate sufficiently challenging MCQs. Rather, the lesson is that appropriate difficulty is necessary but insufficient evidence of assessment readiness.

This also reframes how educators should interpret the current literature. Concerns that GenAI-generated MCQs may be too easy or lower-order remain important, but our findings suggest that difficulty may be a solvable problem when prompts are aligned to postgraduate assessment expectations and outputs are human-edited. The more persistent threat may lie downstream in whether the item functions psychometrically once administered to the target learner population [8-10].

Distractor Functioning as a Residual Vulnerability

The distractor analysis is consistent with, but does not prove, the interpretation that option-set problems contributed to weaker discrimination among GenAI-assisted items. A higher proportion of GenAI-assisted items contained multiple nonfunctioning distractors and 1 or 2 negatively discriminating distractors. Items with negative corrected point-biserial correlations frequently co-occurred with poor distractor functioning. These patterns co-occurred, but the current data cannot establish directionality. Poorly functioning distractors may weaken discrimination, but broader item flaws such as ambiguous wording, contestable keys, or construct-irrelevant complexity may also produce both weak discrimination and unusual distractor selection patterns.

This finding is important because distractors are not decorative. In a well-constructed single-best-answer MCQ, distractors should represent plausible but incorrect alternatives that reflect common misconceptions, reasoning errors, or near-miss clinical decisions. Effective distractors help distinguish candidates who can apply the intended concept from those who cannot. By contrast, nonfunctioning distractors contribute little to measurement precision because few candidates select them. Negatively discriminating distractors are more concerning because they may be selected disproportionately by higher-performing candidates, suggesting ambiguity, competing interpretations, or construct-irrelevant complexity [19,20].

GenAI may be particularly vulnerable at this level. A model can produce distractors that are fluent, grammatically consistent, and superficially plausible, but effective distractor design requires more than surface plausibility. It requires knowledge of how learners commonly misunderstand a concept, which diagnoses or management options are genuinely tempting in practice, and where clinically meaningful decision boundaries lie. These are pedagogical and assessment-design judgments and not simply language generation tasks [19,20].

This possibility may help explain why routine human editing did not fully eliminate psychometric vulnerabilities. Editors may naturally focus on whether the stem is clinically correct, whether the keyed answer is defensible, and whether the wording is clear. Those are essential steps, but they may not be sufficient. Our findings suggest that GenAI-assisted MCQ workflows require deliberate option-set engineering: each distractor should be reviewed for misconception fidelity, clinical plausibility, homogeneity with the key, absence of cueing, and expected diagnostic value. Items with negative discrimination

XSL•FO
RenderX

or multiple nonfunctioning distractors should be revised, quarantined, or retired rather than banked unchanged. Post hoc review and repair of poorly discriminating items have been recommended as a mechanism for improving assessment reliability and defensibility [21].

Implications for Human-in-the-Loop MCQ Development

These findings support a pragmatic model in which GenAI is used as a rapid drafting adjunct, while educators retain responsibility for assessment design, validity evidence, and final item-bank decisions. The human role should not be limited to proofreading or checking factual correctness. Instead, the highest-yield human contribution may be targeted repair of the option set and postadministration review of item functioning. This places GenAI-assisted item development within an educator-led validity and quality assurance workflow rather than treating AI output as assessment-ready evidence [15,16].

A practical human-in-the-loop workflow would include the following stages. First, educators should define the blueprint area, learning objective, expected cognitive level, and clinical context before prompting GenAI. Second, GenAI can be used to generate draft stems, keys, distractors, and explanations. Third, educators should review the draft for clinical correctness, local guideline concordance, clarity, and single-best-answer structure. Fourth, and most importantly, educators should conduct deliberate distractor engineering, ensuring that distractors represent authentic misconceptions, common clinical reasoning errors, or realistic near-miss decisions. When reviewing GenAI-produced distractors, editors should look specifically for obviously implausible options, repeated grammatical or structural patterns that cue the answer, alternatives that are not homogeneous with the key, and near-miss options that are so broad or ambiguous that knowledgeable candidates may reasonably choose them. Such distractors should be rewritten to target a defined misconception or clinical reasoning error, or removed if no plausible misconception can be identified. Fifth, items should be piloted with the intended learner population where feasible. Finally, empirical item analysis should guide whether items are accepted, revised, quarantined, or retired. The emphasis on distractor engineering aligns with established principles that distractors should be plausible, homogeneous, and educationally meaningful rather than merely grammatically plausible alternatives [19,20].

In this workflow, negative corrected point-biserial values should be treated as a serious red flag and should preclude item reuse without careful review of the construct, wording, key, and distractors. Similarly, items with multiple nonfunctioning distractors should be revised even if the stem appears clinically reasonable and the item difficulty falls within an acceptable range. For GenAI-assisted items in particular, the threshold for empirical review should be especially explicit, as surface fluency may create a false sense of readiness. Poor or negative discrimination should therefore prompt conservative item review because it may indicate ambiguity, keying problems, or construct-irrelevant complexity [17,18].

Participant ratings can still be useful, but they should be interpreted as supplementary rather than decisive. In this study, perceived difficulty correlated with objective difficulty, suggesting that trainee ratings may help identify items that feel unexpectedly easy, difficult, or confusing. However, ratings of clarity, relevance, and perceived difficulty were similar between GenAI-assisted and educator-crafted items, although the 2-item sets showed different patterns of reliability, discrimination, and distractor functioning. Therefore, learner ratings may support triage, but they should not replace psychometric analysis or expert review.

The broader implication is that GenAI changes the educator's task. The educator is no longer only an item writer starting from a blank page. The educator becomes a validity-oriented editor and item-quality steward, responsible for transforming fluent drafts into assessment-ready items through construct alignment, distractor repair, pilot testing, and empirical review [15,16].

Future research should also pair psychometric evaluation with workflow and efficiency data, including the time required for prompting, educator editing, distractor repair, expert review, and post–pilot revision. This would help determine whether GenAI-assisted item development improves not only item-bank scalability but also the efficiency of producing assessment-ready MCQs.

Formative and Summative Assessment Implications

The threshold for using GenAI-assisted MCQs should differ between formative and summative contexts. For formative education, edited GenAI-assisted MCQs may be valuable before they reach the psychometric standard required for high-stakes testing, provided the questions are labeled and used as learning resources rather than evidence of competence. They can expand practice question banks, stimulate discussion, support retrieval practice, and expose trainees to a broader range of clinical scenarios. In formative settings, imperfect items can still have educational value if they are reviewed for factual accuracy and accompanied by clear explanations. However, educators should remain alert to misleading distractors, ambiguous keys, or explanations that reinforce incorrect reasoning [19,20,22].

For summative assessment, the threshold should be substantially higher. Items should not enter an operational examination bank solely because they are clinically plausible, appropriately difficult, or rated favorably by trainees. Summative items should demonstrate acceptable discrimination, functioning distractors, key defensibility, and contribution to internal consistency reliability. Items with negative discrimination should be excluded or repaired before reuse. Items with several nonfunctioning distractors should be revised because they may reduce measurement precision. In high-stakes or progression-related contexts, empirical item analysis is not optional quality enhancement; it forms part of the validity argument supporting defensible score interpretation [15-17].

A reasonable implementation stance is therefore staged adoption. GenAI-assisted items can first be used in low-stakes formative settings after educator review. Items that perform well after pilot administration can then be considered for supervised inclusion in item banks. Only items with satisfactory psychometric performance and expert review should be considered for summative use. This staged approach allows

XSL•FO
RenderX

programs to benefit from GenAI's efficiency while preserving assessment quality and fairness.

Limitations

This study has several limitations. First, each item set contained only 30 MCQs. While this allowed pragmatic comparison while limiting educator and participant burden, the relatively small number of items limits the precision of item-level psychometric estimates, particularly for discrimination indices, distractor-functioning analysis, and subgroup comparisons. The findings should therefore be interpreted as exploratory and hypothesis-generating. Future studies with larger item banks and repeated testing across cohorts would allow more stable item calibration and stronger inferences about the comparative performance of GenAI-assisted and educator-crafted MCQs.

Second, items were topic- or domain-matched at the design stage rather than analyzed using a formal matched-pair model. Topic matching strengthened content comparability by ensuring that each item origin covered the same broad set of postgraduate FM domains. However, the paired items were not identical parallel forms and could differ in clinical scenario, cognitive process, decision point, and option structure. We therefore treated item-level comparisons as exploratory and did not conduct paired item-level or mixed-effects sensitivity analyses accounting for topic pair. Future studies should prespecify paired item-level analyses or response-level mixed-effects models with participant, item, and topic-pair effects, ideally using larger item pools per topic.

Third, a deliberate design choice was to evaluate human-edited GenAI-assisted items rather than raw GenAI outputs. This reflects the practical reality of GenAI adoption in assessment development, where educators would be expected to review and edit GenAI-generated drafts before use, particularly in postgraduate or high-stakes contexts. The findings should therefore be interpreted as evidence about a human-in-the-loop workflow, rather than autonomous GenAI item generation. While this design does not isolate the independent contribution of prompting, model output, educator editing, and final item selection, it improves ecological validity and tests whether routine human editing is sufficient to produce psychometrically ready items. Future studies could compare raw GenAI outputs, routinely edited GenAI-assisted items, and deliberately distractor-engineered items to identify which quality assurance steps most improve psychometric functioning. Fourth, the study did not include formal expert-panel coding of cognitive level, such as Bloom's taxonomy or another cognitive-process framework. The findings directly challenge the claim that GenAI-assisted items are necessarily easier because difficulty was empirically measured using trainee response data. However, difficulty and cognitive level are not identical constructs. A difficult item may still test recall, and an applied item may be easy. Future work should combine psychometric analysis with blinded expert coding of cognitive level, item-writing flaws, construct alignment, and distractor quality.

Finally, participant recruitment used convenience sampling among postgraduate FM doctors preparing for the Singapore AKT. The findings may not generalize to undergraduate learners, other specialties, other health care systems, other item formats, or operational high-stakes examinations. Nevertheless, the study provides contextually relevant evidence from a broad, safety-critical postgraduate primary care assessment setting in which GenAI-assisted item development is likely to be increasingly considered.

Conclusions

Human-edited GenAI-assisted MCQs can achieve plausible and often appropriate difficulty for postgraduate FM assessment, challenging the assumption that such items are necessarily too easy. However, difficulty did not ensure assessment readiness. Compared with educator-crafted items, GenAI-assisted items showed lower internal consistency, weaker discrimination, more negative discrimination signals, and more distractor-functioning concerns, despite similar participant ratings of clarity, relevance, and perceived difficulty. These findings suggest that human editing for accuracy and phrasing does not necessarily eliminate downstream psychometric vulnerabilities. GenAI is therefore best positioned as a rapid drafting adjunct within an educator-led workflow that prioritizes key verification, distractor engineering, empirical item analysis, and conservative repair or removal of poorly functioning items before incorporation into question banks or use in formal assessments.

Acknowledgments

The authors would like to thank Dr Phua Dong Haur and Dr James Kwan for their expert advice on psychometric analysis. The authors declare the use of generative artificial intelligence (GAI) in the research and writing process. According to the GAIDeT taxonomy (2025), the following tasks were delegated to GAI tools under full human supervision: text generation and proofreading and editing. The GAI tool used was ChatGPT-5.1. Responsibility for the final manuscript lies entirely with the authors. GAI tools are not listed as authors and do not bear responsibility for the final outcomes. GAI was used as part of the study workflow to generate draft multiple-choice question content, including stems, keyed answers, distractors, and explanations, based on author-specified topics, learning objectives, sample questions, and item-writing guidance. These outputs were subsequently reviewed and edited by the study team for clinical correctness, local guideline contextualization, phrasing, single-best-answer structure, and alignment with intended learning outcomes. No participant data or confidential response data were entered into the artificial intelligence system, and artificial intelligence was not used to perform statistical analysis or interpret participant response data.

XSL•FO
RenderX

Funding

This study was funded by the NHG Health Education Research Grant 2025. The funder had no role in study design, data collection, data analysis, data interpretation, manuscript preparation, or the decision to submit the manuscript for publication.

Data Availability

The datasets generated and analyzed during this study are not publicly available because they contain assessment response data from postgraduate trainees and item-level information from multiple-choice questions used in an educational assessment context. Deidentified aggregate data may be available from the corresponding author on reasonable request, subject to institutional approval and ethical restrictions.

Authors' Contributions

KPS, JQL, and QWF conceptualized the study and designed the experiment, developed the generative artificial intelligence (GenAI)–assisted multiple-choice question (MCQ) workflow, and proofread, reviewed, and edited the GenAI-assisted MCQs before administration. TYGD, WCW, TC, JLW, and VB served as the expert panel and reviewed the MCQs for clinical accuracy, relevance, and alignment with postgraduate family medicine assessment standards. YML performed the statistical analyses and generated the aggregated data outputs. MC provided overall supervision and educational input and substantial critical revision of the manuscript. KPS drafted the initial manuscript. All authors contributed to interpretation of the findings, reviewed and approved the final manuscript, and agreed to be accountable for the accuracy and integrity of the work. QWF and MC contributed equally as joint senior authors.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Additional methodological definitions, workflow documentation, topic-matching information, sensitivity-analysis specifications, and supplementary figures.
[PDF File, 593 KB - mededu_v12i1e100179_app1.pdf]

References

1. Division of Graduate Medical Studies, Yong Loo Lin School of Medicine, National University of Singapore. Family Medicine Applied Knowledge Test Handbook: Division of Graduate Medical Studies, Yong Loo Lin School of Medicine, National University of Singapore; 2025. URL: https://medicine.nus.edu.sg/dgms/wp-content/uploads/sites/30/2022/02/FM-AKT-Handbook.pdf [accessed 2026-02-08]
2. Law AK, So J, Lui CT, et al. AI versus human-generated multiple-choice questions for medical education: a cohort study in a high-stakes examination. BMC Med Educ 2025 Feb 8;25(1):208. [doi: 10.1186/s12909-025-06796-6] [Medline: 39923067]
3. Wu H, Zerner T, Lee D, Court-Kowalski S, Devitt P, Palmer E. GPT-4 versus human authors in clinically complex MCQ creation: a blinded analysis of item quality. Med Teach 2025 Dec;47(12):1961-1974. [doi: 10.1080/0142159X.2025.2505122] [Medline: 40439661]
4. Linde P, Fichter F, Dietlein M, et al. Psychometric properties and detectability of GPT-4o-generated multiple-choice questions compared with human-authored items across imaging specialties. NPJ Digit Med 2026 Jan 8;9(1):132. [doi: 10.1038/s41746-025-02313-7] [Medline: 41507355]
5. Chaudhari H, Meaney C, Kulasegaram K, Leung FH. Evaluating ChatGPT-4 in the development of family medicine residency examinations. In: Abbasgholizadeh-Rahimi S, editor. PLOS Digit Health 2025 Dec;4(12):e0001156. [doi: 10.1371/journal.pdig.0001156] [Medline: 41417746]
6. Kiyani A, Hanif F, Muhammad M, et al. Benchmarking ChatGPT-generated multiple-choice questions against faculty-authored items in dental education. Sci Rep 2025 Dec 29;15(1):44805. [doi: 10.1038/s41598-025-28492-7] [Medline: 41461713]
7. Kaya M, Sonmez E, Halici A, Yildirim H, Coskun A. Comparison of AI-generated and clinician-designed multiple-choice questions in emergency medicine exam: a psychometric analysis. BMC Med Educ 2025 Jul 1;25(1):949. [doi: 10.1186/s12909-025-07528-6] [Medline: 40597998]
8. Kıyak YS, Emekli E. ChatGPT prompts for generating multiple-choice questions in medical education and evidence on their validity: a literature review. Postgrad Med J 2024 Oct 18;100(1189):858-865. [doi: 10.1093/postmj/qgae065] [Medline: 38840505]
9. Riehm L, Nanji K, Lakhani M, Pankiv E, Hasanee D, Pfeifer W. The use of large language models in generating multiple choice questions for health professions education: a systematic review and network meta-analysis. PLoS ONE 2026;21(1):e0340277. [doi: 10.1371/journal.pone.0340277] [Medline: 41481658]

XSL•FO
RenderX

10. Singh S, Alyakin A, Alber DA, et al. The pitfalls of multiple-choice questions in generative AI and medical education. Sci Rep 2025 Nov;15(1):42096. [doi: 10.1038/s41598-025-26036-7]
11. GPT-51: a smarter, more conversational ChatGPT. OpenAI. 2025 Nov. URL: https://openai.com/index/gpt-5-1/ [accessed 2025-11-18]
12. Salkind NJ. Encyclopedia of Research Design: SAGE Publications; 2010. [doi: 10.4135/9781412961288.n205]
13. Metcalfe NH. Testing the test: an analysis of the MRCGP Applied Knowledge Test as an assessment tool. Educ Prim Care 2012 Jan;23(1):13-18. [doi: 10.1080/14739879.2012.11494065] [Medline: 22306140]
14. Dhanvijay AD, Kumari A, Pinjar MJ, et al. Faculty versus artificial intelligence chatbot: a comparative analysis of multiple-choice question quality in physiology. Adv Physiol Educ 2025 Dec 1;49(4):1045-1051. [doi: 10.1152/advan.00197.2025] [Medline: 40981738]
15. Kane MT. Validating the interpretations and uses of test scores. J Educ Meas 2013 Mar;50(1):1-73. [doi: 10.1111/jedm.12000]
16. Cook DA, Brydges R, Ginsburg S, Hatala R. A contemporary approach to validity arguments: a practical guide to Kane's framework. Med Educ 2015 Jun;49(6):560-575. [doi: 10.1111/medu.12678] [Medline: 25989405]
17. Downing SM. The effects of violating standard item writing principles on tests and students: the consequences of using flawed test items on achievement examinations in medical education. Adv Health Sci Educ Theory Pract 2005;10(2):133-143. [doi: 10.1007/s10459-004-4019-5] [Medline: 16078098]
18. Rush BR, Rankin DC, White BJ. The impact of item-writing flaws and item complexity on examination item difficulty and discrimination value. BMC Med Educ 2016 Sep 29;16(1):250. [doi: 10.1186/s12909-016-0773-3] [Medline: 27681933]
19. Haladyna TM, Downing SM, Rodriguez MC. A review of multiple-choice item-writing guidelines for classroom assessment. Appl Meas Educ 2002 Jul;15(3):309-333. [doi: 10.1207/S15324818AME1503_5]
20. Tarrant M, Ware J, Mohammed AM. An assessment of functioning and non-functioning distractors in multiple-choice questions: a descriptive analysis. BMC Med Educ 2009 Jul 7;9(1):40. [doi: 10.1186/1472-6920-9-40] [Medline: 19580681]
21. Desy J, Harvey A, Weeks S, et al. The impact of systematically repairing multiple choice questions with low discrimination on assessment reliability: an interrupted time series analysis. Can Med Ed J 2024 Jul 12;15(3):52-56. [doi: 10.36834/cmej.77596]
22. Larsen DP, Butler AC, Roediger HL. Test-enhanced learning in medical education. Med Educ 2008 Oct;42(10):959-966. [doi: 10.1111/j.1365-2923.2008.03124.x] [Medline: 18823514]

Abbreviations

AI: artificial intelligence
FM: family medicine
FM AKT: Family Medicine Applied Knowledge Test
GenAI: generative artificial intelligence
KR-20: Kuder-Richardson Formula 20
MCQ: multiple-choice question

Edited by S Brini; submitted 09.May.2026; peer-reviewed by AKD Dhanvijay, D Lee, P Chauhan; revised version received 13.Jun.2026; accepted 20.Jul.2026; published 10.Aug.2026.

Please cite as:
Sze KP, Lim JQ, Loke YM, Ding TYG, Wee WC, Chen T, Wu JL, Bansal V, Fong QW, Chow M
Human-Edited Generative AI-Assisted Multiple-Choice Questions in Postgraduate Family Medicine: Blinded Cross-Sectional Comparative Psychometric Study
JMIR Med Educ 2026;12:e100179
URL: https://mededu.jmir.org/2026/1/e100179
doi:10.2196/100179

© Kai Ping Sze, Jia Qing Lim, Yng Miin Loke, Teck Yong Gabriel Ding, Wei Chieh Wee, Tongyuan Chen, Jingkai Lawrence Wu, Vivek Bansal, Qi Wei Fong, Minyang Chow. Originally published in JMIR Medical Education (https://mededu.jmir.org), 10.Aug.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Education, is properly cited. The complete bibliographic information, a link to the original publication on https://mededu.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Integrating Generative AI Into Clinical Reasoning Education for Medical Students: Mixed Methods Study

Cheng-Heng Liu[1,2], MD; Yu-Ting Chen[3,4], PhD; Chiun Hsu[5], MD, PhD; Huey-Ling Chen[4,6], MD, PhD; Chih-Wei Yang[1,2,4], MD, PhD

[1]Department of Medical Education, National Taiwan University Hospital, No. 7, Zhongshan South Road, Zhongzheng District, Taipei City, Taiwan

[2]Department of Emergency Medicine, National Taiwan University Hospital, Taipei City, Taiwan

[3]Department of Living Sciences, National Open University, New Taipei City, Taiwan

[4]Graduate Institute of Medical Education and Bioethics, National Taiwan University College of Medicine, Taipei City, Taiwan

[5]Department of Medical Oncology, National Taiwan University Cancer Center, Taipei City, Taiwan

[6]Department of Pediatrics, National Taiwan University Hospital, Taipei City, Taiwan

Corresponding Author:
Chih-Wei Yang, MD, PhD
Department of Medical Education, National Taiwan University Hospital, No. 7, Zhongshan South Road, Zhongzheng District, Taipei City, Taiwan

Abstract

Background: Generative AI (GenAI) is increasingly integrated into clinical learning and practice. However, medical students often lack the competencies required for safe and critical use, including prompt design, output verification, and recognition of limitations. Educational interventions that integrate GenAI with clinical reasoning frameworks remain limited.

Objective: This study evaluated a structured, theory-informed workshop integrating GenAI, prompt engineering, and clinical reasoning education to enhance medical students' self-perceived AI literacy and collaborative learning attitudes, and to assess whether patient-centered orientation changed following intensive AI exposure.

Methods: We conducted a single-group, pre-post, explanatory sequential mixed methods study with fifth-year medical students enrolled at an academic medical center between April 2024 and May 2025. The 3-hour workshop comprised 6 modules integrating clinical reasoning, cognitive-bias awareness, verification-oriented GenAI use, and hands-on prompt engineering and centered on ChatGPT (OpenAI). Quantitative outcomes were self-reported measures from a 20-item self-report AI-literacy questionnaire adapted from the Meta AI Literacy Scale and were examined with exploratory and confirmatory factor analysis, the 6-item Patient-Practitioner Orientation Scale-Short, and a modified Collaborative Learning Attitude Scale (CLAS). Pre-post change was assessed using 2-tailed paired *t* tests with Benjamini-Hochberg correction and Cohen *d*; the Patient-Practitioner Orientation Scale-Short and CLAS were available for a subsample (n=46). Qualitative data from 6 interviews and 17 reflective narratives were analyzed using reflexive thematic analysis and integrated with the quantitative findings.

Results: Among 150 eligible students, 139 (92.7%) completed paired AI-literacy assessments. Self-perceived AI literacy improved across all domains (Cohen d=0.69 - 0.93, all P<.001; all remaining significant after false discovery rate correction), and collaborative learning attitudes increased substantially (d=0.94, P<.001). Patient-centered orientation showed no significant change (d=0.03); however, baseline scores were concentrated at the favorable end of the scale (a floor/restricted-range effect), and the subsample analysis was underpowered (minimum detectable dz=0.42), so this null result is inconclusive rather than evidence of unchanged orientation. Gains did not differ by sex or across the sequential cohorts. Reflexive thematic analysis (interviews: n=6; reflections: n=17) identified five themes and one emergent theme describing a shift toward verification-oriented GenAI use: (1) understanding GenAI capabilities and limitations, (2) prompt-engineering skill development, (3) calibrated trust through verification, (4) GenAI-supported communication and collaboration, and (5) ethical considerations, with emerging reconceptualization of professional identity.

Conclusions: A brief, theory-informed educational intervention integrating GenAI with clinical reasoning was associated with medium-to-large improvements in self-perceived AI literacy and collaborative attitudes. No detectable change in patient-centered orientation was observed; however, this finding should be interpreted as inconclusive, given measurement and power constraints. Embedding verification practices within clinical reasoning frameworks may offer a scalable approach for preparing physicians for responsible human-AI collaboration. Future studies should incorporate comparative designs, performance-based assessments, and longitudinal follow-up.

(JMIR Med Educ 2026;12:e94951) doi:10.2196/94951

XSL•FO
RenderX

KEYWORDS

large language models; clinical reasoning; AI literacy; prompt engineering; medical education; generative AI; diagnostic decision-making

Introduction

Generative AI (GenAI), particularly large language models (LLMs), has rapidly transitioned from a speculative concept to a widely used clinical technology, with increasing relevance for medical education. Emerging evidence indicates that LLM-enabled systems may contribute to improvements in patient safety and quality of care, particularly in knowledge-intensive and documentation-heavy tasks [1]. Reviews of AI in health care have attributed this momentum to rapid advances in data availability and natural language processing, which have enabled increasingly sophisticated support for knowledge-intensive clinical tasks [2]. This rapid adoption has prompted growing interest in how future physicians should be prepared to engage with AI-enabled clinical environments, and recent frameworks have begun to define the digital and AI competencies expected of medical graduates [3].

Across health care settings, LLM-enabled systems are increasingly used to support clinical documentation and patient-facing communication, raising questions about how clinicians should incorporate AI-generated suggestions into reasoning and decision-making [1]. Experimental diagnostic vignettes further demonstrate their potential: in controlled comparisons, GenAI systems have matched or outperformed physicians on structured case assessments [4,5]. However, the evidence is mixed, and high-quality outputs do not necessarily reflect trustworthy reasoning processes. Systematic comparisons show variable performance across clinical decision-support tasks [6], and randomized trials indicate that simply providing clinicians with access to LLMs does not consistently improve diagnostic accuracy compared with conventional information resources [5]. Risks including hallucinations, embedded biases, and opaque reasoning may compromise patient safety if AI outputs are accepted uncritically [1,7,8]. Together, these findings underscore the importance of deliberate human-AI collaboration strategies rather than passive tool adoption.

Within medical education, clinical reasoning is conceptualized as a dynamic interplay between rapid pattern-recognition processes (System 1) and slower, analytical verification processes (System 2) [9]. A substantial body of research on cognitive bias describes how anchoring, premature closure, and related heuristics can derail both modes of reasoning [10]. Introducing GenAI into clinical learning environments without explicit instructional scaffolding may exacerbate these vulnerabilities by encouraging premature acceptance of AI-generated suggestions, thereby weakening reflective reasoning rather than strengthening it.

Despite growing enthusiasm for AI among medical trainees, surveys and needs assessments consistently indicate substantial gaps in preparedness for responsible and effective use. Learners report limited competence in prompt design, bias appraisal, and ethical boundary setting when interacting with GenAI systems [11,12], and a recent multi-institutional study confirms that positive student attitudes coexist with persistent concerns and unmet training needs [13]. Scholars have cautioned that uncritical reliance on AI tools may lead to a "lost cognitive perspective," in which clinicians gradually offload essential reasoning processes to algorithms [8]. Together, these concerns highlight the need for medical curricula to move beyond passive exposure to AI technologies and toward structured, theory-informed educational interventions that integrate GenAI within established clinical reasoning frameworks.

To address this educational gap, we designed a 3-hour, 6-module workshop for fifth-year medical students that integrates hands-on GenAI prompt-engineering exercises with dual-process clinical reasoning, cognitive bias awareness, and explicit verification strategies. Rather than positioning AI as a shortcut to answers, the workshop frames GenAI as a powerful but fallible tool requiring contextual judgment and reflective oversight. The intervention was embedded within an existing clinical rotation to maximize ecological validity and align AI use with authentic case-based learning. A defining feature of this approach is its explicit intent to cultivate AI competencies alongside, rather than at the expense of, humanistic clinical values, treating patient-centered orientation not as a target for improvement, but as a benchmark for educational safety.

The primary objective of this study was to evaluate whether participation in this workshop was associated with improvements in medical students' self-perceived AI literacy across knowledge, application, evaluation, and ethics domains. Secondary objectives were to examine changes in collaborative learning attitudes, conceptualized as openness to human-AI collaboration, and to assess whether patient-centered orientation changed following intensive exposure to AI tools. Using a mixed methods design, we sought not only to quantify changes in self-reported competencies but also to explore learners' experiences and perceptions in order to generate meta-inferences about mechanisms of learning and professional adaptation in AI-augmented clinical reasoning education.

Methods

Study Design

We conducted a mixed methods educational evaluation using an explanatory sequential design (also termed sequential explanatory), in which a quantitative pre-post outcome phase was followed by a qualitative phase intended to help explain and contextualize the quantitative findings. The quantitative outcome data were collected immediately before and after the workshop, assessing changes in self-perceived AI literacy, collaborative learning attitudes, and patient-centered orientation; the qualitative interviews and reflective narratives were then collected in a subsequent phase to examine how students described their reasoning processes, trust calibration, and professional adaptation when using GenAI. The 2 strands were integrated at the interpretation stage.

XSL•FO
RenderX

Given the curricular constraints of a required clinical rotation, a single-group pre-post design was adopted to evaluate immediate educational impact. This design does not include a comparator group and therefore does not permit causal inference; findings are interpreted as changes associated with participation rather than as effects attributable solely to the intervention.

The study adhered to the GRAMMS (Good Reporting of a Mixed Methods Study) guidelines [14] (Checklist 1) to ensure transparency and rigor in the integration of qualitative and quantitative data.

Setting and Participants

The study was conducted during the 2024 to 2025 academic year at the National Taiwan University College of Medicine. All 150 fifth-year undergraduate medical students enrolled in a required clinical rotation were invited to participate. The educational intervention was embedded within the existing curricular schedule, and participation in the research component was voluntary. Between April 2024 and May 2025, students participated in sequential cohorts of 11 to 13 students per session (13 sessions in total). Students were informed that opting out of research would not affect academic standing. To reduce cross-cohort contamination, students were reminded not to share detailed workshop materials with peers who had not yet attended. Taiwan's medical education follows a 6-year undergraduate model; fifth-year students are therefore in their first year of formal clinical training. No structured AI-related curriculum had been offered during their preclinical years (years 1 - 4), and the 2024 to 2025 academic year coincided with a period of rapid mainstream adoption of GenAI tools, making this workshop a timely and contextually relevant intervention for most participants. Participant demographic characteristics, including sex distribution and prior GenAI experience, are summarized in Table 1.

Table . Participant characteristics.

Variable and category	Participants, n (%)
Sex[a]	
Male	106 (76.3)
Female	33 (23.7)
Prior GenAI[b] use	
Yes	138 (99.3)
No	1 (0.7)
GenAI use frequency[c]	
<1 day/week	17 (12.3)
About 1 day/week	21 (15.2)
2 - 4 days/week	59 (42.8)
Almost daily	41 (29.7)

[a]One participant who selected "Other" for sex was included in the female category for reporting purposes.

[b]GenAI: generative AI.

[c]One participant who selected "Other" for frequency was excluded (n=138).

Educational Intervention

The intervention consisted of a 3-hour, small-group workshop organized into 6 sequential instructional modules (approximately 25 min each) and delivered in an interactive format. The 6 modules accounted for 150 minutes of instructional time. An additional 20 minutes were allocated for preworkshop and postworkshop questionnaire completion, with the remaining 10 minutes used for introductions, transitions between modules, and a concluding wrap-up discussion. The 6 modules addressed the following topics: (1) Technology-Assisted Clinical Decision-Making; (2) Clinical Reasoning and Cognitive Biases; (3) GenAI and Prompt Engineering; (4) Critical Appraisal and Evidence Verification; (5) Patient-Centered Decision-Making with AI; and (6) Synthesis and Action Planning. Instructional strategies included brief didactic segments, clinical vignette analysis, structured prompt-engineering exercises, role-play exercises emphasizing patient-centered communication, and guided verification steps designed to support reflective human-AI collaboration.

The workshop centered on ChatGPT (OpenAI) as the primary GenAI platform; over the enrollment period (April 2024 to May 2025) this spanned the GPT-4 and GPT-4o model generations, and students were free to use other publicly available GenAI tools according to task demands. In module 4, students compared outputs from ChatGPT with those from domain-specific, evidence-based clinical reference tools (UpToDate and DynaMed) to reinforce the need for verification and contextual judgment. All activities used fictionalized clinical vignettes; students were instructed not to input identifiable patient information into GenAI tools.

An overview of module learning objectives and key activities is provided in Multimedia Appendix 1.

XSL•FO
RenderX

Data Collection

To evaluate the effectiveness of the workshop, we administered 3 established instruments using a pre-post design to assess participants' self-perceived AI literacy, patient-centered orientation, and collaborative learning attitudes.

AI literacy was assessed as self-perceived competence, using a 20-item self-report questionnaire adapted from the Meta AI Literacy Scale [15] and informed by an established AI literacy framework [16]; hereafter, this outcome is referred to as self-perceived AI literacy. Items were selected for contextual relevance to clinical reasoning, with minimal wording modification for medical education applicability. The questionnaire was translated into Mandarin Chinese and reviewed by bilingual medical education faculty for semantic equivalence. The final instrument assessed 4 domains, rated on a 5-point Likert scale (1="strongly disagree" to 5="strongly agree"): (1) Knowing and Understanding AI, (2) Applying AI, (3) Evaluating AI Applications, and (4) AI Ethics.

To examine the structure of the adapted instrument, exploratory factor analysis was conducted on pretest data (n=139) using principal axis factoring with Promax rotation. The Kaiser-Meyer-Olkin measure of sampling adequacy was 0.904 (excellent), and Bartlett test of sphericity was significant (χ^2_{190}=2504.2, P<.001). Parallel analysis suggested a 2-factor solution (a general competence factor and an ethics-awareness factor), whereas the theory-guided 4-factor solution explained 71.2% of total variance. The complete factor-loading matrices for both the 2-factor and 4-factor solutions are provided in Multimedia Appendix 2. To further examine the retained structure, a confirmatory factor analysis (CFA) of the 4-factor model was conducted on posttest data using maximum-likelihood estimation. Because this CFA drew on the same participants as the exploratory factor analysis, it does not constitute an independent validation of the factor structure. The 4 conceptual domains were highly intercorrelated (latent correlations up to r=0.93), and the model showed only modest fit (comparative fit index=0.90, Tucker-Lewis index=0.89, root-mean-square error of approximation=0.10, standardized root-mean-square residual=0.07). We therefore retain the 4-domain organization on theoretical and content-validity grounds rather than as a strongly discriminated empirical model; given the sample size, the high interfactor correlations, and the nonindependence of the exploratory and confirmatory analyses, these structural analyses are interpreted descriptively (see *Limitations* section). Internal consistency was acceptable to excellent (Cronbach α=0.77 - 0.96 across domains; total α=0.93). Given the adaptation and translation, findings are interpreted as context-specific rather than as formal cross-cultural validation.

Patient-centered attitudes were assessed using the 6-item Patient-Practitioner Orientation Scale-Short (PPOS-D6) [17,18]. Items are rated on a 5-point Likert scale (1=strongly disagree to 5=strongly agree), with lower scores indicating greater patient-centered orientation. The PPOS-D6 has demonstrated acceptable psychometric properties in prior validation studies. Owing to a logistical constraint, this measure was administered only in the final 5 cohorts (January-May 2025), yielding complete pre-post data for 46 participants; analyses for this outcome are therefore interpreted cautiously, given the reduced and potentially nonrepresentative sample (see *Results* and *Limitations* sections). Collaborative learning attitudes were measured using 6 items adapted from the Collaborative Learning Attitude Scale (CLAS) [19,20], selected a priori based on item relevance; as with PPOS-D6, complete pre-post data were available for the same subset of participants (n=46). All questionnaires were translated into Mandarin Chinese and proofread by the author team prior to administration.

In this subsequent qualitative phase, data were obtained through semistructured interviews (n=6) and written reflective narratives (n=17) following the workshop. Participants were purposively sampled for interviews to reflect variation in baseline AI familiarity. The adequacy of the qualitative dataset was appraised using the concept of information power across the 5 dimensions described by Malterud et al [21]. The study aim was narrow, focusing on students' experiences of one structured workshop. Sample specificity was high: all participants were fifth-year students who had completed the workshop. The study applied established theory, namely dual-process theory and reflexive thematic analysis. Dialogue quality was supported by the instructor-researchers' familiarity with the workshop context, with the attendant positionality addressed under *Data Analysis*. Finally, the cross-case analysis drew on 2 complementary data types, with 17 written reflective narratives broadening the corpus beyond the six interviews. With reference to pragmatic data-adequacy benchmarks for interview studies [22], we judged this corpus sufficient for the descriptive, workshop-specific aims of this phase, while acknowledging the modest number of interviews (see *Limitations* section). The interview guide is available from the corresponding author upon reasonable request.

Data Analysis

Quantitative data were analyzed using IBM SPSS Statistics for Windows (version 26.0; IBM Corp) and Python (SciPy, statsmodels, factor_analyzer, and semopy). Two-tailed paired-sample t tests were conducted to examine pre-post differences in outcome measures, and effect sizes were calculated using Cohen d. To account for multiple comparisons across the subscale pre-post tests, P values were adjusted using the Benjamini-Hochberg false discovery rate procedure [23], with Bonferroni correction reported for comparison; statistical significance was defined as P<.05. To assess the representativeness of the PPOS-D6/CLAS subsample, baseline characteristics (sex, prior GenAI use, baseline self-perceived AI literacy, and GenAI use frequency) were compared between the subsample and the remaining participants using the Welch t test, the Mann-Whitney U test, and the Fisher exact test as appropriate. A sensitivity power analysis was conducted for the PPOS-D6 total score. As an exploratory analysis, we examined whether gains in self-perceived AI literacy were moderated by sex and whether baseline scores or pre-post gains varied across the sequential cohorts (Spearman rank correlation with chronological cohort order). Given the single-group design, all analyses are interpreted as associations with participation rather than causal effects.

XSL•FO
RenderX

Qualitative data from interviews and reflective narratives were analyzed using reflexive thematic analysis, following Braun and Clarke [24-26]. An inductive coding approach was adopted. Two researchers independently familiarized themselves with the data, generated initial codes, and developed candidate themes, which were refined through iterative reflexive discussion; coding disagreements were resolved through this consensus process rather than by interrater reliability statistics, consistent with a reflexive approach. To address trustworthiness, we supported credibility through triangulation across interviews and reflective narratives; dependability through a consistently applied coding procedure with consensus resolution of discrepancies; confirmability through independent dual coding and explicit reflexivity; and transferability through thick contextual description of the setting and participants. As a positionality statement, the authors reflexively acknowledge their dual roles as both workshop instructors and researchers and remain attentive to how this positionality could shape data generation and interpretation; we sought to mitigate it through independent coding and reflexive discussion. Representative quotations were translated from Mandarin Chinese and verified by bilingual authors.

Mixed Methods Integration

Integration occurred at the interpretation and meta-inference level. Quantitative changes in self-perceived AI literacy and collaborative attitudes were examined alongside qualitative themes describing development of prompt-engineering competence, verification-oriented reasoning, calibrated trust, and professional identity adaptation. Joint interpretation enabled identification of 3 integrated learning dimensions: (1) technical AI literacy acquisition, (2) verification-oriented reasoning development, and (3) professional and ethical adaptation. This integrative approach allowed the qualitative findings to explain and elaborate the self-reported quantitative improvements. All authors participated in the integration process.

Ethical Considerations

This educational study was approved by the institutional review board of the National Taiwan University Hospital (institutional review board number 202403120RINA). All participants provided informed consent electronically prior to study participation. Participation in surveys and qualitative feedback activities was voluntary. Although the workshop was a required curricular activity, students who opted out of the research component received identical educational content. To protect privacy and confidentiality, survey responses were collected without personal identifiers, and baseline characteristics (eg, sex and GenAI use) were collected. Qualitative data were deidentified prior to analysis. Participants did not receive financial compensation for their involvement in the study.

Results

Participant Flow and Baseline Characteristics

Among 150 eligible fifth-year medical students, 139 (92.7%) completed both preintervention and postintervention surveys for AI literacy and were included in the primary quantitative analysis. The sample comprised 106 (76.3%) male students and 33 (23.7%) female students, a distribution consistent with the sex composition of the medical class at our institution. Nearly all respondents (138/139, 99.3%) reported prior experience with GenAI, with the most common usage frequency being 2 to 4 days per week (59/138, 42.8%). Owing to a logistical limitation, the PPOS-D6 and the CLAS were administered only to 46 participants from the final 5 cohorts (January-May 2025); these analyses are therefore reported separately, and the representativeness of this subsample is examined below. Qualitative data were obtained from 6 semistructured interviews and 17 written reflections.

Self-Perceived AI Literacy

Significant improvements were observed across all self-perceived AI literacy domains and in the total score (Table 2). The mean total score increased from 3.57 (SD 0.64) preintervention to 4.23 (SD 0.54) postintervention (t_{138}=11.55, P<.001), corresponding to a large effect size (Cohen d=0.98, 95% CI 0.77 - 1.19). Effect sizes of domain-level changes ranged from 0.69 to 0.93 (Knowledge: d=0.90, 95% CI 0.70 - 1.10; Application: d=0.69, 95% CI 0.49 - 0.88; Evaluation: d=0.93, 95% CI 0.73 - 1.13; Ethics: d=0.79, 95% CI 0.59 - 0.99), with the largest changes observed in the Knowledge and Evaluation domains. All domain improvements remained significant after Benjamini-Hochberg correction for multiple comparisons (all adjusted P<.001; the same conclusion held under the more conservative Bonferroni correction).

Table . Self-perceived AI literacy pre-post comparison (n=139).

Domain	Pretest, mean (SD)	Posttest, mean (SD)	t (df=138)	P value	Cohen d (95% CI)
Knowledge	3.72 (0.65)	4.28 (0.53)	10.63	<.001	0.90 (0.70 - 1.10)
Application	3.62 (0.81)	4.19 (0.62)	8.09	<.001	0.69 (0.49 - 0.88)
Evaluation	3.45 (0.82)	4.23 (0.61)	10.91	<.001	0.93 (0.73 - 1.13)
Ethics	3.48 (1.00)	4.24 (0.79)	9.27	<.001	0.79 (0.59 - 0.99)
Total	3.57 (0.64)	4.23 (0.54)	11.55	<.001	0.98 (0.77 - 1.19)

Internal consistency for the self-perceived AI literacy domains in this sample was acceptable to excellent, with Cronbach α values ranging from 0.77 to 0.96.

In an exploratory analysis, gains in self-perceived AI literacy did not differ by sex: total gains were near-identical for male and female students (+0.65 vs +0.64; Welch t=0.14, P=.89, d=0.02; 106 male vs 32 female, with the single participant who selected “other” excluded from this analysis), and no domain showed a significant difference (all P≥.74), and baseline scores did not differ by sex (P=.48). Across the 13 sequential cohorts

XSL•FO
RenderX

(n=138; one response with an unparseable timestamp was excluded), pre-post gains were unrelated to cohort order (Spearman ρ=–0.07, *P*=.39), and baseline self-perceived AI literacy was not significantly associated with cohort order (Spearman ρ=0.17, *P*=.05); cohort- and semester-level descriptive statistics are provided in Multimedia Appendix 2.

Patient-Practitioner Orientation

Using the 6-item PPOS-D6, no significant pre-post differences were observed for the Sharing (*d*=–0.09, 95% CI –0.39 to 0.21), Caring (*d*=0.17, 95% CI –0.13 to 0.47), or Total (*d*=0.03, 95% CI –0.27 to 0.33) scores (Table 3). This null result should be interpreted in light of 2 constraints. First, the subsample (the final 5 cohorts) was not fully representative of the whole cohort: although comparable in sex distribution (Fisher exact *P*=.09) and prior GenAI use (*P*>.99), it had higher baseline self-perceived AI literacy (3.74 vs 3.48; Welch *t*=2.59, *P*=.01, *d*=0.42) and more frequent GenAI use (Mann-Whitney *P*<.001; both differences remained significant after Benjamini-Hochberg correction), consistent with its being composed of later cohorts. Second, because PPOS-D6 items are scored such that lower values indicate greater patient-centered orientation, the low baseline scores reflect an already strongly patient-centered orientation at entry: baseline total scores clustered at the favorable (low) end of the scale (mean 1.84, SD 0.50, maximum 2.83), no participant scored above the scale midpoint, and responses occupied only about 46% of the available range (a floor/restricted-range effect; Multimedia Appendix 2). A sensitivity power analysis indicated that, with 46 participants, the study could detect only medium-or-larger within-person change (minimum detectable *dz*=0.42 at 80% power; observed *dz*=0.03, corresponding to achieved power of approximately 5%). The nonsignificant PPOS-D6 result is therefore inconclusive on both representativeness and measurement grounds rather than evidence that orientation was unchanged.

Table . Patient-practitioner orientation pre-post comparison (n=46).

Dimension	Pre, mean (SD)	Post, mean (SD)	*t* (*df*=45)	*P* value	Cohen *d* (95% CI)
Sharing	1.94 (0.62)	1.88 (0.92)	–0.62	.54	–0.09 (–0.39 to 0.21)
Caring	1.75 (0.59)	1.84 (0.80)	1.12	.27	0.17 (–0.13 to 0.47)
Total	1.84 (0.50)	1.86 (0.78)	0.18	.86	0.03 (–0.27 to 0.33)

Collaborative Learning Attitude

Scores on the modified CLAS increased significantly after the workshop (t_{45}=6.39, *P*<.001). Mean CLAS total scores increased from 3.74 (SD 0.44) to 4.22 (SD 0.64), with a large effect size (*d*=0.94, 95% CI 0.61 - 1.27); this improvement remained significant after Benjamini-Hochberg correction. As this outcome was assessed in the same subsample of 46 participants, the same representativeness caveat applies.

Qualitative Findings

Thematic analysis of semistructured interviews (n=6) and written reflective narratives (n=17) identified 5 major themes and 1 emergent theme from reflections. Table 4 summarizes themes and representative quotes. Additional quotes, subthemes, and the 4-stage learning trajectory are provided in Multimedia Appendix 1. Themes encompassed the following:

1. Understanding GenAI Capabilities and Limitations: Students described GenAI as useful for information synthesis and differential diagnosis expansion, while recognizing its limitations in contextual judgment and localization. Several participants emphasized the probabilistic nature of outputs and the necessity of contextual interpretation.
2. Development of Prompt-Engineering Skills: Participants reported progression from generic queries to structured, context-rich prompts. They described learning to specify task framing, request justification, and iteratively refine prompts. Prompt engineering was increasingly conceptualized as a clinical skill analogous to structured history-taking.
3. Calibrated Trust Through Verification Practices: Students emphasized verification as a mandatory step. They reported cross-checking outputs against textbooks, guidelines, or supervisory input. Rather than treating AI as authoritative, learners described adopting a “calibrated trust” stance, adjusting reliance based on task complexity and stakes.
4. GenAI-Supported Communication and Collaboration: Participants viewed GenAI as helpful for translating medical terminology into lay language and generating alternative perspectives during group discussion. However, they consistently positioned empathy and relational care as distinctly human responsibilities.
5. Ethical and Educational Considerations: Students raised concerns about privacy, accountability, bias, and overreliance. Some expressed uncertainty regarding institutional guidance for AI use.

With regard to the emergent theme—Professional Identity Reconceptualization—reflective writings suggested evolving perceptions of the physician’s role in AI-augmented practice. Some students described transitioning from sole knowledge holders to “AI-augmented decision makers,” while others expressed ambivalence regarding dependency on technology.

XSL•FO
RenderX

Table . Summary of qualitative themes and representative quotes.

Theme	Key finding	Representative quote[a]
Understanding GenAI[b] capabilities and limitations	GenAI was perceived as useful for knowledge integration but limited by context and localization.	"AI only sees the keywords we input; it cannot fully comprehend the complete situation." (P-E)
Development of prompt-engineering skills	Students progressed from generic prompts to structured, context-rich prompting and task-based tool selection.	"You need to write better instructions, specifically telling it which references to use." (P-D)
Calibrated trust through verification	Learners emphasized verification and deference to human clinical judgment when using GenAI outputs.	"The first step is always to verify or read through myself." (P-A)
GenAI-supported communication and collaboration	GenAI supported translation of information for patients and teamwork, while empathy remained a human responsibility.	"Generative AI has the unique ability to quickly translate medical information into lay language." (P-D)
Ethical and educational considerations	Students raised concerns about privacy, accountability, and overreliance on GenAI in learning.	"Because AI is so convenient, we might lose our ability to research original textbooks." (P-E)
Emergent theme: professional identity reconceptualization	Reflections described shifts in perceived physician roles and future practice in AI-augmented care.	"We are standing at the starting point of this transformation." (RH)

[a]Participant IDs beginning with "P" denote interview participants, and IDs beginning with "R" denote authors of written reflective narratives. All quotations were translated from Mandarin Chinese.

[b]GenAI: generative AI.

Mixed Methods Integration

Integration of quantitative and qualitative findings revealed three interconnected dimensions of learning:

1. Technical Competence Development: Quantitative improvements in self-perceived AI literacy were reflected qualitatively in students' descriptions of improved prompt design and critical appraisal skills.
2. Verification-Oriented Reasoning: Gains in the Evaluation and Ethics domains aligned with qualitative accounts of calibrated trust and structured verification practices.
3. Professional Adaptation in Human-AI Contexts: Increased collaborative learning attitudes corresponded with qualitative themes of human-AI teaming and evolving professional identity.

Collectively, these integrated findings suggest that the intervention was associated not only with increased self-perceived AI literacy but also with shifts in reasoning processes and conceptualizations of human-AI collaboration.

Discussion

Principal Results

To our knowledge, this study is among the first mixed methods investigations to examine how structured prompt engineering and AI literacy training can be integrated into undergraduate clinical reasoning education. The findings suggest that effective use of GenAI depends not only on access to advanced models but also on learners' ability to critically direct, evaluate, and contextualize AI-generated outputs.

In this structured GenAI-integrated clinical reasoning workshop, participation was associated with medium-to-large improvements in self-perceived AI literacy across knowledge, application, evaluation, and ethics domains, alongside substantial gains in collaborative learning attitudes. Patient-centered orientation did not demonstrate measurable short-term change; as detailed below and in the *Limitations* section, this null result is inconclusive, given the restricted baseline range and limited statistical power rather than firm evidence of stability. Qualitative findings complemented these results by revealing a developmental trajectory from passive tool use toward verification-oriented, context-aware engagement with GenAI systems, with students beginning to reconceptualize their professional roles within AI-augmented clinical environments.

Because all quantitative outcomes were self-reported, the observed improvements reflect self-perceived competence and confidence rather than demonstrated performance, and may be subject to response-shift bias, social desirability, and demand characteristics, particularly because the instructors also served as assessors. We therefore interpret the quantitative gains as changes in self-perceived AI literacy and recommend that future studies corroborate them with objective, behaviorally anchored measures, such as structured AI-use vignettes, AI-integrated objective structured clinical examination stations, and rubric-based assessment of prompt quality. With this caveat, the integration of quantitative and qualitative data indicates that self-perceived AI-literacy acquisition was accompanied by qualitatively described shifts in reasoning stance and professional orientation.

Comparison With Prior Work

A central educational concern in AI integration is automation bias, the tendency to overrely on automated outputs without adequate scrutiny [27]. In this study, improvements in the Evaluation and Ethics domains were mirrored by qualitative descriptions of deliberate verification, with students cross-checking AI outputs against trusted sources and supervisory judgment. This pattern aligns with the concept of

XSL•FO
RenderX

calibrated trust in human-AI systems, whereby appropriate reliance is dynamically adjusted to task context and risk [28]. Crucially, verification was treated not as an ancillary safeguard but as a structured component of reasoning instruction: by situating GenAI within dual-process theory, learners were encouraged to treat AI outputs as hypotheses for analytical evaluation (System 2) rather than as definitive answers. This framing may mitigate the "lost cognitive perspective" described in prior literature [8] and supports a pedagogical shift from "AI familiarity" toward verification-oriented AI literacy. Importantly, our findings suggest that AI literacy should not be equated with mere familiarity or frequency of AI use. Rather, meaningful AI literacy requires the ability to critically interrogate outputs, recognize uncertainty, identify hallucinations, and determine when AI recommendations should or should not influence clinical decisions. Consistent with calls for critical AI literacy in medical education [8], students increasingly recognized AI's strengths in pattern recognition and retrieval alongside its limitations in contextual judgment, suggesting that explicit discussion of model capabilities and failure modes should be a core curricular element. From a cognitive load perspective, AI can reduce the extraneous load of information retrieval [29], but this convenience may also invite cognitive offloading; mandating verification positions AI as a scaffold for, rather than a substitute for, the development of clinical reasoning [30].

Learners also progressed from generic prompting to structured, context-rich queries, describing a 4-stage trajectory of passive reception, basic engagement, strategic application, and critical monitoring. This suggests that prompt engineering can be taught as a clinical skill analogous to structured history-taking, in which learners provide organized information, request justification, and iteratively refine queries; the observed Application-domain gain (d=0.69) is consistent with this interpretation, and hands-on practice anchored in authentic scenarios and scaffolded by instructor feedback appeared essential to its development.

The substantial increase in collaborative learning attitudes (d=0.94), together with qualitative accounts of AI as a collaborative partner, valued for translating technical information into lay language while empathy and relational continuity remained firmly human responsibilities, aligns with the centaur model of human-AI collaboration, in which combined human and machine intelligence outperforms either entity alone [31]. Reflective writings further suggested emerging professional identity adaptation [32], with students moving from sole knowledge holders toward "AI-augmented decision makers"; some voiced concern about overreliance ("we might lose our ability to research original textbooks"), while others described themselves as "standing at the starting point of transformation," underscoring the need to integrate AI into professionalism and identity-formation curricula, not only technical training. In contrast, patient-centered orientation did not change detectably; given the baseline floor effect and limited statistical power, this is most consistent with, though it cannot confirm, the premise that AI literacy and patient-centeredness need not be antagonistic competencies when instruction explicitly integrates ethical reflection and communication practice [33]. Adequately powered longitudinal research with more sensitive measures is needed to determine whether GenAI–reasoning curricula influence or safeguard patient-centeredness over time.

Educational Design Implications

Collectively, these findings suggest three interrelated dimensions for AI curriculum design (technical AI literacy, verification-oriented reasoning, and professional/sociotechnical adaptation) and four practical design principles for GenAI-integrated curricula: (1) embed AI activities within authentic clinical reasoning tasks to support skill transfer; (2) treat AI as powerful but fallible, with verification as a required step; (3) use small-group formats that mirror collaborative clinical environments; and (4) explicitly connect AI use to patient-centered communication and ethical responsibility. These principles align with broader calls for sociotechnical AI education and may support scalable curriculum development.

Limitations

Several limitations warrant consideration. First, all quantitative outcomes were self-reported measures of perceived competence rather than objective performance, and may be affected by response-shift bias, social desirability, and demand characteristics, particularly because the instructors also served as assessors. Although the workshop focused on clinical reasoning, the study did not directly assess clinical reasoning performance. Therefore, it remains unclear whether improvements in self-perceived AI literacy translate into enhanced diagnostic reasoning or clinical decision-making. Future studies should incorporate objective, behaviorally anchored assessments, such as structured AI-use vignettes, AI-integrated objective structured clinical examination stations, script concordance tests, key-feature problems, or rubric-based evaluations of diagnostic reasoning and prompt quality. Second, the single-group pre-post design without a comparator group limits causal inference; observed changes are described as associated with participation rather than attributable solely to the intervention, and no comparator data were available for a sensitivity analysis. Third, the PPOS-D6 and CLAS analyses were limited to a subsample of 46 participants from the final 5 cohorts; this subsample had higher baseline self-perceived AI literacy and more frequent GenAI use than the remaining participants, limiting representativeness, and the study was powered to detect only medium-or-larger change in PPOS-D6 (minimum detectable dz=0.42), so the null PPOS-D6 result is inconclusive, particularly given a baseline floor/restricted-range effect reflecting an already strongly patient-centered orientation. In addition, the unbalanced sex distribution (106 male vs 32 female) limited the power of the sex-moderation analysis to detect small-to-moderate effects (minimum detectable d≈0.57), so the absence of sex differences should not be interpreted as evidence of equivalence. Fourth, although the 4-domain AI-literacy structure was retained on theoretical grounds, the domains were highly intercorrelated and a CFA showed only modest fit. Because both analyses relied on the same participants, the CFA cannot independently confirm the factor structure and may capitalize on sample-specific characteristics [34]. These structural analyses are therefore descriptive, and validation of the 4-domain structure in an independent sample

XSL•FO
RenderX

is needed before it can be treated as confirmed. Fifth, data collection spanned 13 months during which public exposure to GenAI evolved and the underlying models advanced; although gains were unrelated to cohort timing, a nonsignificant change in baseline self-perceived AI literacy across cohorts, together with the possibility of informal cross-cohort sharing of materials, cannot be excluded. Finally, the study was conducted at a single academic medical center with high baseline AI exposure, and the qualitative corpus, while enriched by 17 reflective narratives, was based on six interviews; generalizability to other institutions or earlier-stage learners may therefore be limited.

Conclusions

As GenAI tools continue to evolve and permeate clinical environments, medical education should move beyond passive exposure toward structured, theory-informed curricula that help future physicians harness AI's potential responsibly. By interweaving technical skills with clinical reasoning and humanistic principles, medical education can prepare physicians who not only use AI but also guide its responsible implementation in patient care. This study indicates that a concise, theory-informed workshop was associated with improvements in self-perceived AI literacy and collaborative attitudes while explicitly emphasizing critical thinking and humanistic values. The remaining challenge lies in scaling such interventions while maintaining educational quality and addressing the ethical complexities associated with AI in medicine.

Future research should incorporate comparative designs, objective performance-based measures, and longitudinal follow-up to examine durability, behavioral transfer, and professional identity development in AI-augmented clinical environments, particularly as medicine moves toward more transparent and explainable AI systems [35].

Acknowledgments

We thank the fifth-year medical students who participated in the workshop and provided feedback, as well as the faculty facilitators who supported small-group teaching. We also thank Editage [36] for English language editing. Generative AI, ChatGPT (GPT-5, OpenAI), was used during manuscript preparation solely to assist with language editing and formatting of author-written text; it was not used for study design, data collection, data analysis, interpretation of results, or the generation of any results or findings. Consistent with the calibrated-trust, verification-oriented approach to AI use that this study examined, every AI-assisted passage was independently checked and verified by the authors against their intended meaning and the underlying data, and all authors take full responsibility for the final content.

Funding

This research was supported by the National Science and Technology Council, Taiwan (NSTC 114-2410-H-002-095-MY2) and National Taiwan University Hospital (NTUH 114-SS0006, NTUH 115-SS0007). The funders had no involvement in the study design, data collection, analysis, interpretation of the data, or writing of the manuscript.

Data Availability

Deidentified quantitative data, codebooks, and interview code summaries are available from the corresponding author upon reasonable request and subject to institutional regulations.

Authors' Contributions

Conceptualization: CWY
Data curation: CHL, YTC
Formal analysis: CWY, CH
Investigation: CWY, CHL, YTC, HLC
Methodology: CWY
Qualitative analysis: HLC
Supervision: CWY
Writing – original draft: CWY
Writing – review & editing: CWY, CHL, YTC, CH, HLC

Conflicts of Interest

None declared.

Multimedia Appendix 1
Workshop module overview and extended qualitative findings.
[DOCX File, 14 KB - mededu_v12i1e94951_app1.docx]

Multimedia Appendix 2

XSL•FO
RenderX

Supplementary statistical tables.
[DOCX File, 29 KB - mededu_v12i1e94951_app2.docx]

Checklist 1
GRAMMS checklist.
[DOCX File, 15 KB - mededu_v12i1e94951_app3.docx]

References

1. Howell MD. Generative artificial intelligence, patient safety and healthcare quality: a review. BMJ Qual Saf 2024 Oct 18;33(11):748-754. [doi: 10.1136/bmjqs-2023-016690] [Medline: 39048292]
2. Jiang F, Jiang Y, Zhi H, et al. Artificial intelligence in healthcare: past, present and future. Stroke Vasc Neurol 2017;2(4):230-243. [doi: 10.1136/svn-2017-000101] [Medline: 29507784]
3. Car J, Ong QC, Erlikh Fox T, et al. The digital health competencies in medical education framework: an international consensus statement based on a Delphi study. JAMA Netw Open 2025 Jan 2;8(1):e2453131. [doi: 10.1001/jamanetworkopen.2024.53131] [Medline: 39888625]
4. Cabral S, Restrepo D, Kanjee Z, et al. Clinical reasoning of a generative artificial intelligence model compared with physicians. JAMA Intern Med 2024 May 1;184(5):581-583. [doi: 10.1001/jamainternmed.2024.0295] [Medline: 38557971]
5. Goh E, Gallo R, Hom J, et al. Large language model influence on diagnostic reasoning: a randomized clinical trial. JAMA Netw Open 2024 Oct 1;7(10):e2440969. [doi: 10.1001/jamanetworkopen.2024.40969] [Medline: 39466245]
6. Sandmann S, Riepenhausen S, Plagwitz L, Varghese J. Systematic analysis of ChatGPT, Google Search and Llama 2 for clinical decision support tasks. Nat Commun 2024 Mar 6;15(1):2050. [doi: 10.1038/s41467-024-46411-8] [Medline: 38448475]
7. Price WN. Big data and black-box medical algorithms. Sci Transl Med 2018 Dec 12;10(471):eaao5333. [doi: 10.1126/scitranslmed.aao5333] [Medline: 30541791]
8. Tikhomirov L, Semmler C, McCradden M, Searston R, Ghassemi M, Oakden-Rayner L. Medical artificial intelligence for clinicians: the lost cognitive perspective. Lancet Digit Health 2024 Aug;6(8):e589-e594. [doi: 10.1016/S2589-7500(24)00095-5] [Medline: 39059890]
9. Pelaccia T, Tardif J, Triby E, Charlin B. An analysis of clinical reasoning through a recent and comprehensive approach: the dual-process theory. Med Educ Online 2011 Mar 14;16:5890. [doi: 10.3402/meo.v16i0.5890] [Medline: 21430797]
10. Saposnik G, Redelmeier D, Ruff CC, Tobler PN. Cognitive biases associated with medical decisions: a systematic review. BMC Med Inform Decis Mak 2016 Nov 3;16(1):138. [doi: 10.1186/s12911-016-0377-1] [Medline: 27809908]
11. Civaner MM, Uncu Y, Bulut F, Chalil EG, Tatli A. Artificial intelligence in medical education: a cross-sectional needs assessment. BMC Med Educ 2022 Nov 9;22(1):772. [doi: 10.1186/s12909-022-03852-3] [Medline: 36352431]
12. Li Q, Qin Y. AI in medical education: medical student perception, curriculum recommendations and design suggestions. BMC Med Educ 2023 Nov 9;23(1):852. [doi: 10.1186/s12909-023-04700-8] [Medline: 37946176]
13. Duan S, Liu C, Rong T, Zhao Y, Liu B. Integrating AI in medical education: a comprehensive study of medical students' attitudes, concerns, and behavioral intentions. BMC Med Educ 2025 Apr 23;25(1):599. [doi: 10.1186/s12909-025-07177-9] [Medline: 40269824]
14. O'Cathain A, Murphy E, Nicholl J. The quality of mixed methods studies in health services research. J Health Serv Res Policy 2008 Apr;13(2):92-98. [doi: 10.1258/jhsrp.2007.007074] [Medline: 18416914]
15. Carolus A, Koch MJ, Straka S, Latoschik ME, Wienrich C. MAILS - Meta AI Literacy Scale: development and testing of an AI literacy questionnaire based on well-founded competency models and psychological change- and meta-competencies. Comput Hum Behav Artif Hum 2023 Aug;1(2):100014. [doi: 10.1016/j.chbah.2023.100014]
16. Ng DTK, Leung JKL, Chu SKW, Qiao MS. Conceptualizing AI literacy: an exploratory review. Comput Educ Artif Intell 2021;2:100041. [doi: 10.1016/j.caeai.2021.100041]
17. Jiang Y, Wei J, Shi L, Cao J, Zhu B, Hong X. A 4-factor perspective of the Patient-Practitioner Orientation Scale (PPOS): a deeper understanding of patient-centredness. BMC Med Educ 2022 Nov 29;22(1):818. [doi: 10.1186/s12909-022-03867-w] [Medline: 36447203]
18. Pauli R, Wilhelmy S. A short scale for measuring attitudes towards the doctor-patient relationship: psychometric properties and measurement invariance of the German Patient-Practitioner-Orientation Scale (PPOS-D6). PeerJ 2021;9:e12604. [doi: 10.7717/peerj.12604] [Medline: 34966596]
19. Berger-Estilita J, Chiang H, Stricker D, Fuchs A, Greif R, McAleer S. Attitudes of medical students towards interprofessional education: a mixed-methods study. PLoS One 2020;15(10):e0240835. [doi: 10.1371/journal.pone.0240835] [Medline: 33085695]
20. Chao CT, Chiu YL, Tsai CL, et al. Moving from tangibility toward digitalization: investigating team dynamics and facilitator support among medical students in conventional and digital small-group tutorials. BMC Med Educ 2022 Nov 28;22(1):814. [doi: 10.1186/s12909-022-03893-8] [Medline: 36443728]

XSL•FO
RenderX

21. Malterud K, Siersma VD, Guassora AD. Sample size in qualitative interview studies: guided by information power. Qual Health Res 2016 Nov;26(13):1753-1760. [doi: 10.1177/1049732315617444] [Medline: 26613970]
22. Guest G, Bunce A, Johnson L. How many interviews are enough?: an experiment with data saturation and variability. Field methods 2006;18(1):59-82. [doi: 10.1177/1525822X05279903]
23. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J R Stat Soc Series B Stat Methodol 1995 Jan 1;57(1):289-300. [doi: 10.1111/j.2517-6161.1995.tb02031.x]
24. Braun V, Clarke V. Using thematic analysis in psychology. Qual Res Psychol 2006;3(2):77-101. [doi: 10.1191/1478088706qp063oa]
25. Braun V, Clarke V. Thematic Analysis: A Practical Guide: SAGE; 2021.
26. Braun V, Clarke V. Toward good practice in thematic analysis: avoiding common problems and be(com)ing a knowing researcher. Int J Transgend Health 2023;24(1):1-6. [doi: 10.1080/26895269.2022.2129597] [Medline: 36713144]
27. Goddard K, Roudsari A, Wyatt JC. Automation bias: a systematic review of frequency, effect mediators, and mitigators. J Am Med Inform Assoc 2012;19(1):121-127. [doi: 10.1136/amiajnl-2011-000089] [Medline: 21685142]
28. Jacobs M, He J, Pradier MF, et al. Designing AI for trust and collaboration in time-constrained medical decisions: a sociotechnical lens. Proc CHI Conf Hum Factors Comput Syst 2021:1-14. [doi: 10.1145/3411764.3445385]
29. Sweller J, Ayres P, Kalyuga S. Cognitive Load Theory: Springer; 2011.
30. Young JQ, Van Merrienboer J, Durning S, Ten Cate O. Cognitive load theory: implications for medical education: AMEE Guide No. 86. Med Teach 2014 May;36(5):371-384. [doi: 10.3109/0142159X.2014.889290] [Medline: 24593808]
31. Topol E. Deep Medicine: How Artificial Intelligence Can Make Healthcare Human Again: Basic Books; 2019.
32. Cruess RL, Cruess SR, Boudreau JD, Snell L, Steinert Y. A schematic representation of the professional identity formation and socialization of medical students and residents: a guide for medical educators. Acad Med 2015 Jun;90(6):718-725. [doi: 10.1097/ACM.0000000000000700] [Medline: 25785682]
33. Masters K. Artificial intelligence in medical education. Med Teach 2019 Sep;41(9):976-980. [doi: 10.1080/0142159X.2019.1595557] [Medline: 31007106]
34. Fokkema M, Greiff S. How performing PCA and CFA on the same data equals trouble. Eur J Psychol Assess 2017;33(6):399-402. [doi: 10.1027/1015-5759/a000460]
35. Holzinger A, Langs G, Denk H, Zatloukal K, Müller H. Causability and explainability of artificial intelligence in medicine. Wiley Interdiscip Rev Data Min Knowl Discov 2019;9(4):e1312. [doi: 10.1002/widm.1312] [Medline: 32089788]
36. Editage. URL: https://www.editage.com/ [accessed 2026-07-29]

Abbreviations

CFA: confirmatory factor analysis
CLAS: Collaborative Learning Attitude Scale
GenAI: generative AI
GRAMMS: Good Reporting of a Mixed Methods Study
LLM: large language model
PPOS-D6: Patient-Practitioner Orientation Scale-Short

Edited by A Stone; submitted 09.Mar.2026; peer-reviewed by B Rubinstein, I Murray; revised version received 14.Jul.2026; accepted 20.Jul.2026; published 13.Aug.2026.

Please cite as:
Liu CH, Chen YT, Hsu C, Chen HL, Yang CW
Integrating Generative AI Into Clinical Reasoning Education for Medical Students: Mixed Methods Study
JMIR Med Educ 2026;12:e94951
URL: https://mededu.jmir.org/2026/1/e94951
doi:10.2196/94951

© Cheng-Heng Liu, Yu-Ting Chen, Chiun Hsu, Huey-Ling Chen, Chih-Wei Yang. Originally published in JMIR Medical Education (https://mededu.jmir.org), 13.Aug.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Education, is properly cited. The complete bibliographic information, a link to the original publication on https://mededu.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

AI Integration in Undergraduate Medical Education: Qualitative Study of Faculty Perspectives in the United Arab Emirates

Azhar T Rahma[1], PhD; Uffaira Hafeez[2], MBBS, MPH; Munawar Farooq[2], MBBS

[1]Institute of Public Health, College of Medicine and Health Sciences, United Arab Emirates University (UAEU), Al Ain, United Arab Emirates

[2]Department of Internal Medicine, Emergency Medicine Section, College of Medicine and Health Sciences, United Arab Emirates University (UAEU), Sheik Khalifa Bin Zayed St - 'Asharij - Shiebat Al Oud - Abu Dhabi, Al Ain, Abu Dhabi, United Arab Emirates

Corresponding Author:
Munawar Farooq, MBBS
Department of Internal Medicine, Emergency Medicine Section, College of Medicine and Health Sciences, United Arab Emirates University (UAEU), Sheik Khalifa Bin Zayed St - 'Asharij - Shiebat Al Oud - Abu Dhabi, Al Ain, Abu Dhabi, United Arab Emirates

Abstract

Background: AI is transforming health care, creating an imperative to integrate AI into medical education. While student perspectives are well-studied, faculty views, particularly in non-Western contexts, remain underexplored.

Objective: This qualitative study explores the perspectives of 10 medical faculty members from 2 institutions in the United Arab Emirates on integrating AI into undergraduate medical education.

Methods: This multi-institutional qualitative study used purposive and reflexive sampling to recruit faculty from both public and private medical universities in the United Arab Emirates. Semistructured interviews were conducted with 10 faculty members involved in curriculum design, teaching, or assessment. Data collection followed COREQ (Consolidated Criteria for Reporting Qualitative Research) guidelines. Data were analyzed using a mixed inductive-deductive approach guided by the thematic analysis framework of Braun and Clarke. Findings were interpreted using the FACETS (Form, AI Use Case, Context, Education, Technology, and SAMR: Substitution, Augmentation, Modification, Redefinition) framework, which supported a structured examination of AI integration across different dimensions of teaching and learning.

Results: Faculty primarily used generative AI tools, such as ChatGPT, for content creation, assessment development, and teaching support, reflecting a preference for accessible and general-purpose technologies. AI was mainly used to enhance teaching efficiency and support student learning, including personalized study planning and practice activities. Its application extended across preclinical and clinical contexts, with strong emphasis on adapting content to local cultural and ethical norms. While AI was perceived to improve efficiency and alignment between teaching and assessment, concerns were raised regarding equity, overreliance, and variability in student use. Overall, adoption remained focused on enhancing existing practices, with limited transformative use but recognition of future potential for more advanced applications.

Conclusions: The UAE medical faculty demonstrate cautious optimism toward AI integration, recognizing its potential to enhance educational efficiency and personalization while emphasizing the critical importance of cultural contextualization. Current implementation remains at early adoption stages, focused on enhancement rather than transformation. Successful integration requires faculty development, context-sensitive policies, and equitable implementation strategies that address both technological and sociocultural dimensions of AI adoption in medical education.

(JMIR Med Educ 2026;12:e90192) doi:10.2196/90192

KEYWORDS

artificial intelligence; medical education; medical curriculum; medical school; AI ethics; medical faculty; United Arab Emirates; UAE

Introduction

AI refers to computer systems capable of performing tasks traditionally requiring human intelligence, such as learning, reasoning, and decision-making [1]. In health care, AI is increasingly influencing clinical practice by enhancing diagnostic accuracy, clinical decision support, and predictive analytics [2,3]. Significant advancements have been demonstrated across multiple specialties [4-8], prompting discussions about how AI may reshape, supplement, or shift professional medical roles [9]. Globally, governments and health systems, including in the United Arab Emirates, are investing in AI-driven innovations and workforce readiness initiatives [10,11]. As health care becomes increasingly digitalized, future physicians will be expected to collaborate effectively with AI

XSL•FO
RenderX

systems, critically evaluate their reliability, and make informed clinical judgments in augmented practice environments [12]. This has led to calls for integrating AI-related knowledge, critical appraisal skills, and ethical reasoning into undergraduate medical education [13].

Several studies have explored AI's potential to enhance medical teaching across diverse disciplines. For example, augmented reality has been used to improve the understanding of complex anatomical structures [14]; AI supports diagnosis and monitoring in ophthalmology [15] and lesion recognition in dermatology [16], provides adaptive simulation-based training in surgery [17], and enhances image-based interpretation in radiography through deep learning [18]. More recently, research has focused on generative AI and large language models (LLMs) in medical education. These studies describe applications such as supporting simulation-based training, improving students' digital literacy and ethical competencies [19], answering medical licensing exam questions [20], and comparing AI-generated versus expert-generated questions [21]. Beyond assessment, generative AI has been explored as a virtual mentor and educational support tool [22], delivering personalized formative feedback efficiently [23], enhancing case-based learning, and aligning AI-based grading with teacher evaluations [24]. LLMs have also demonstrated potential in enabling authentic assessments by providing immersive, real-world evaluation experiences [25].

Integrating AI into medical education represents a significant pedagogical shift, with students as the primary recipients of innovation and faculty as the key drivers of its implementation. Evidence suggests that students generally hold positive attitudes toward AI, recognize its potential to enhance learning and diagnostic reasoning, and advocate for structured training within the curriculum [26-28]. However, faculties face substantial challenges at both personal and institutional levels [29]. Multicenter studies indicate that, despite generally positive faculty attitudes, significant gaps remain in formal AI training and knowledge [30]. Faculty engagement is critical for successful curricular reform, as their beliefs, readiness, and professional development directly influence adoption and sustainability [31]. Therefore, efforts to integrate AI must address faculty concerns and actively involve them in the process. Moreover, effective adoption requires a contextualized approach that considers local educational, cultural, and institutional factors [32]. Despite the growing body of literature on AI in medical education, most studies have focused on technological applications, expert perspectives, or student perspectives, with comparatively limited attention to faculty experiences, particularly in the Middle East and the UAE context. There is a lack of context-specific evidence on how faculties perceive their readiness, roles, and challenges in integrating AI into existing curricula. Therefore, this study aims to explore medical faculty perspectives on the integration of AI into undergraduate medical education in the United Arab Emirates, focusing on their experiences, perceived readiness, and implementation challenges. The study applies the FACETS (Form, AI Use Case, Context, Education, Technology, and SAMR: Substitution, Augmentation, Modification, Redefinition) framework to move beyond descriptive accounts and provide a theoretically informed, multidimensional analysis of AI integration in medical education.

Methods

Study Design

A qualitative descriptive approach was used to explore faculty perspectives on integrating AI into medical education in the United Arab Emirates. This study followed the COREQ (Consolidated Criteria for Reporting Qualitative Research) guidelines [33]. Semistructured interviews were conducted to capture faculty experiences, beliefs, and concerns regarding AI integration while allowing flexibility to pursue emerging topics of significance to participants.

Epistemological Position

This study adopts a critical realist epistemological stance, acknowledging that while participants' experiences and perspectives represent real phenomena, our interpretation and presentation of these experiences are inevitably shaped by our theoretical lenses, disciplinary backgrounds, and the analytical frameworks we employ.

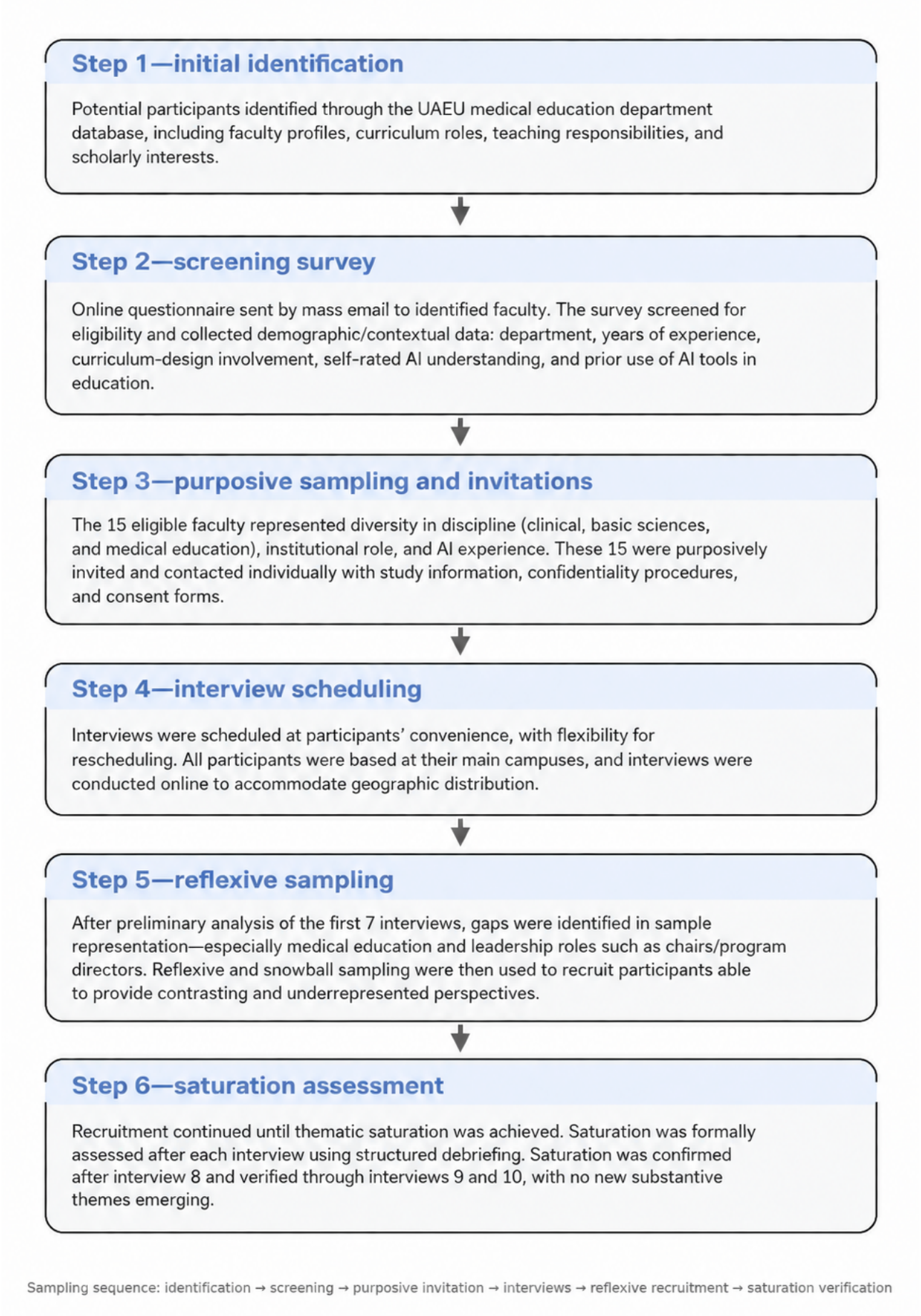
Sampling

A purposive sampling approach was employed to identify and recruit faculty members from the medical schools in the United Arab Emirates: either a government-funded public university or a private, internationally affiliated institution, located in the United Arab Emirates. Eligible participants were those actively involved in curriculum development, teaching, or assessment. Invitations were distributed via departmental mailing lists and targeted emails to faculty with relevant roles.

An initial pool of 47 faculty members was identified, of whom 22 responded to a screening questionnaire assessing academic roles, experiences, and familiarity with AI. Fifteen met the inclusion criteria, and 7 participated in the initial round of interviews. Following preliminary analysis, additional participants were recruited using reflexive and snowball sampling to ensure representation across disciplines (clinical, basic sciences, and medical education) and varying levels of AI experience. A total of 10 faculty members were interviewed. The term "key informants" refers to participants selected based on their roles, experiences, or departmental affiliations relevant to the study (Figure 1).

XSL•FO
RenderX

Figure 1. Participant recruitment flow: qualitative study of United Arab Emirates University faculty.



XSL•FO
RenderX

Recruitment continued until thematic saturation was reached. Saturation was observed after 8 interviews and confirmed with 2 additional interviews, with no new themes emerging.

All interviews were conducted online. Participants were based at their respective institutions, and the inclusion of faculty from both public and private settings allowed for a range of perspectives.

Data Collection

This study took place in the United Arab Emirates between January and April 2025. We recruited participants from institutions that offer MBBS programs accredited by the Commission for Academic Accreditation.

An initial online questionnaire was used to screen participants and collect demographic and contextual information. It served both as a screening tool to identify eligible participants based on curriculum involvement and AI knowledge and as a means of gathering background data, including departmental affiliation, years of experience, curriculum involvement, self-rated AI understanding, and prior experience with AI tools in education.

Faculty who met the inclusion criteria and expressed interest were invited to participate in the interviews.

Data were collected through semistructured online interviews using a predefined interview guide (Multimedia Appendix 1), which was reviewed by experts in medical education and qualitative research. Verbal informed consent was obtained before each interview. Interviews lasted between 30 and 60 minutes and were audio-recorded and transcribed verbatim. Audio recordings were deleted after transcription to ensure confidentiality.

Interviews were conducted by AR, a public health faculty member with experience in qualitative research, with support from MF and UH. Some participants were known to the research team in a professional context. However, participation was voluntary and interviews were conducted independently of prior relationships. Participants were informed about the study purpose, confidentiality measures, and their right to withdraw.

Field notes were recorded during and after interviews to capture contextual observations and initial reflections. Transcripts were retained to support reflexivity, allowing repeated engagement with the data, reflection on researcher assumptions, and maintenance of an audit trail of analytical decisions. Transcripts were returned to participants for verification. NVivo (version 15; Lumivero) was used to support data management and analysis.

Data Analysis

Data were analyzed using reflexive thematic analysis, informed by the 6-phase approach of Braun and Clarke [34]. After transcription, the transcripts were read repeatedly to ensure familiarization with the data. Coding was conducted inductively, with codes generated directly from participants' accounts rather than from a preexisting framework.

The initial coding was carried out independently by 2 analysts (AR and UH). The analysts met regularly to compare coding decisions, discuss interpretations, and refine the developing coding structure through an iterative process of discussion and review.

Codes were then grouped into candidate themes by identifying patterns of shared meaning across the dataset. These themes were reviewed against the coded extracts and full transcripts to ensure coherence, distinctiveness, and relevance to the research question. Themes were then defined, named, and organized into a coherent analytic narrative.

Once initial themes were developed, the FACETS framework was used as an interpretive lens to further structure and refine the findings [35,36]. This allowed the analysis to remain grounded in participants' accounts while also supporting a more structured interpretation of AI integration across multiple dimensions.

Researcher Reflexivity and Positionality

Prior to the study, the research team engaged in reflexivity discussions to consider how their disciplinary backgrounds, professional roles, and assumptions might influence the research process. The team included a public health faculty member with qualitative research experience (AR), clinical faculty members involved in medical education and curriculum-related work (MF and UH), and a medical education scholar (SW). These backgrounds provided complementary perspectives related to curriculum development, educational innovation, and the integration of AI in health professions education.

The team acknowledged a shared preconception that AI holds important potential for medical education, while also recognizing concerns related to ethics, equity, overreliance, and contextual appropriateness within the UAE setting. We also recognized that our academic positions and prior engagement with medical education may have shaped the issues we considered important and the meanings we foregrounded during the analysis. To mitigate the influence of these assumptions, we used a semistructured interview guide, maintained reflexive memos throughout data collection and analysis, and engaged in regular analytic discussions to ensure that findings remained grounded in participants' accounts.

Trustworthiness

Several strategies were used to enhance the trustworthiness of the study. Credibility was supported through in-depth, semistructured interviews, prolonged engagement with the transcripts, iterative analytic discussions among the research team, and the inclusion of verbatim quotations to illustrate the themes. Dependability was strengthened through a transparent and documented analytic process, including coding records, reflexive memos, and an audit trail that tracked the development of themes from raw data to final interpretation. Confirmability was addressed through reflexive practice, explicit acknowledgment of researcher assumptions, and grounding interpretations in participants' own words, as well as returning the transcript for some participants for validation. Transferability was supported by providing a clear description of the study context, participant characteristics, institutional settings, sampling strategy, and recruitment process, allowing readers to assess the relevance of the findings to similar contexts.

XSL•FO
RenderX

Ethical Considerations

The study was approved by the UAE University Ethics Board (reference ERSC_2024_4790). Participation was voluntary, and all participants provided informed consent. No financial or other compensation was provided to participants. Data confidentiality and anonymity were strictly maintained throughout the research process.

Results

Perspectives on Integration and Current Practices

Overview

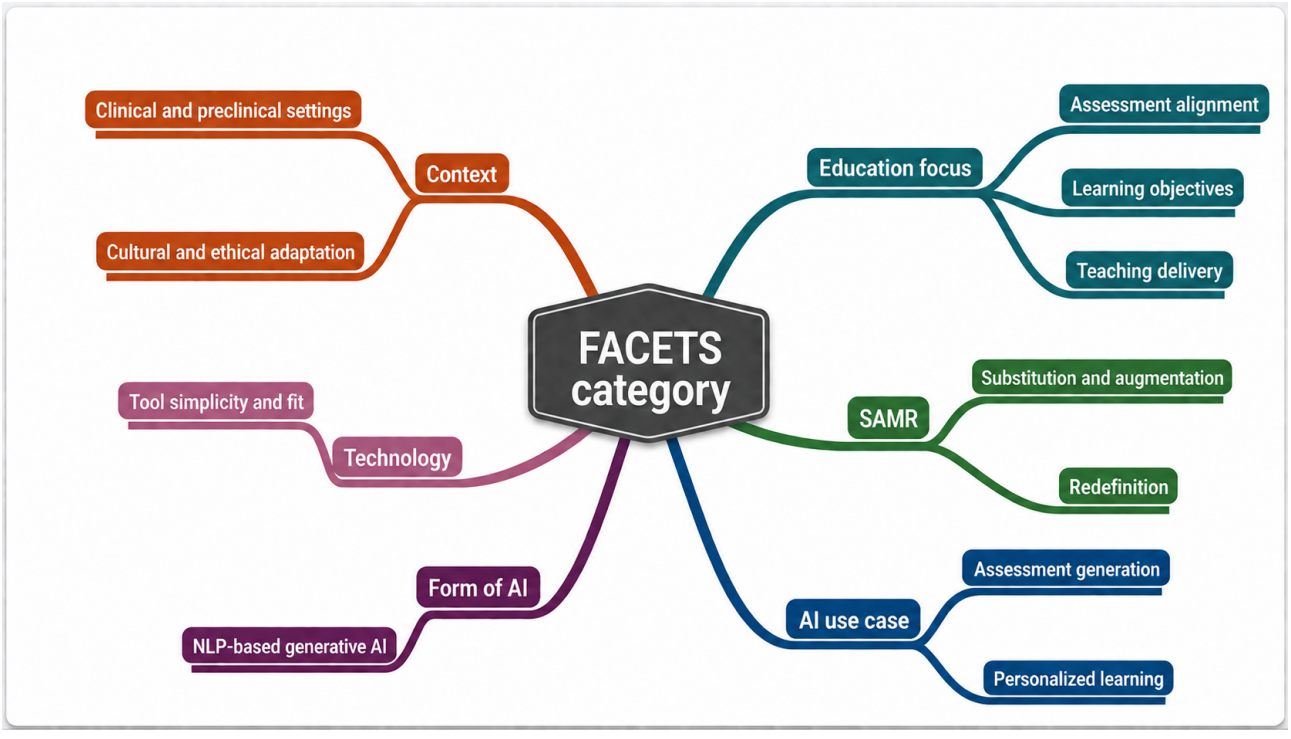
The study included 10 faculty members and represented a range of disciplines, including emergency medicine, internal medicine, public health, obstetrics and gynecology, pathology, and medical education. Most had extensive involvement in curriculum design, with a few serving in a curriculum leadership capacity. The self-rated understanding of AI ranged from “moderate” to “good” (Table 1).

Table . Characteristics of key informants (N=10).

Participant ID	Department	Role in curriculum design	University type	Self-rated understanding of AI
KI-01	Emergency medicine	Extensive	Public	Good
KI-02	Internal medicine	Limited	Public	Moderate
KI-03	Public health	Extensive	Public	Moderate
KI-04	Medical education	Extensive	Public	Moderate
KI-05	Internal medicine	Extensive	Public	Expert
KI-06	Obstetrics and gynecology	Extensive	Public	Good
KI-07	Pathology	Lead	Private	Moderate
KI-08	Medical education	Moderate	Public	Good
KI-09	Medical education	Extensive	Public	Good
KI-10	Medical education	Moderate	Public	Good

The following findings are organized according to the FACETS framework (Figure 2). Beyond identifying where and how AI was being used, the analysis also revealed how faculty interpreted the meaning, limits, and implications of this integration within their specific institutional and sociocultural context.

Figure 2. Faculty perspectives on themes related to strategically integrating AI in UAE medical education. FACETS: Form, AI Use Case, Context, Education, Technology, and SAMR; NLP: natural language processing; SAMR: Substitution, Augmentation, Modification, Redefinition.



XSL•FO
RenderX

This figure presents the 6 themes identified through the reflexive thematic analysis and shows how they were organized using the FACETS framework and interpreted alongside the SAMR model. The FACETS domains are Form (type of AI used), AI use case (how AI was applied), Context (where and under what sociocultural conditions AI was used), Education (teaching, learning, and assessment functions), Technology (tools and platforms adopted), and SAMR (level of educational technology integration). The figure shows that faculty perspectives clustered around accessible generative AI tools, especially natural language processing (NLP) applications, with reported uses in teaching support, assessment, and personalized learning. Participants emphasized the importance of contextualization to the UAE setting, including cultural, ethical, and language considerations. Across the dataset, the current AI use was interpreted mainly at the substitution and augmentation levels, with more limited examples of Modification and anticipated future potential for Redefinition.

(F)—Form of AI: A Monoculture of Generative NLP Tools

The analysis revealed a distinct homogeneity in the form of AI being adopted, which is almost exclusively centered on NLP-based generative AI, with ChatGPT as the dominant platform. This indicates that faculty are gravitating toward accessible, general-purpose tools rather than specialized educational software. The primary use of these tools was to act as an efficiency engine for content creation and simplification. Faculty leveraged them not for complex analytical tasks but to rapidly generate and refine textual and conceptual materials, effectively using AI as a powerful, on-demand assistant to streamline their workflow.

> *Even using the free version... you can engineer a prompt for a lecture or concept and ask it to explain it in your own language and level.* [KI-07]
>
> *AI like ChatGPT is my assistant when writing problem-based cases.* [KI-09]

(AI)—AI Use Case: Augmenting Educator Productivity and Enabling Student Personalization

The use cases for AI crystallized into 2 dominant patterns: enhancing faculty productivity and facilitating student-centered learning. For educators, the most significant application was in assessment generation and content authoring. Faculty reported using AI as a collaborative draftsperson to create a wide range of assessment items, from multiple-choice questions (MCQs) to clinical scenarios. This use was primarily motivated by a need for efficiency and diversity, allowing them to quickly generate a larger and more varied pool of teaching materials while ensuring alignment with learning objectives.

> *I use ChatGPT to generate MCQs and essay questions across Bloom's taxonomy.* [KI-07]
>
> *We've used it to create questions that match learning outcomes in less time.* [KI-10]

Conversely, faculty members observed that students were leveraging the same tools for personalized learning. Students proactively used AI to create custom study plans, deconstruct complex topics, and generate adaptive practice questions, effectively creating a personal tutor. This divergence in use cases highlights AI's dual role: as a productivity tool for instructors and an adaptive learning scaffold for students.

> *It gives different versions every time you prompt it. That's powerful for revising topics.* [KI-10]
>
> *Students ask AI for feedback or use it to rehearse for OSCEs.* [KI-04]

(C)—Context: Pervasive Application Demands Conscious Localization

The context of AI application was notably broad, spanning both preclinical and clinical phases of the curriculum. Faculty reported using AI to develop materials for problem-based learning and team-based learning, indicating its utility across diverse pedagogical frameworks. This demonstrates that AI is not confined to a single teaching methodology but is seen as a flexible resource adaptable to various educational settings.

> *We are already using AI in developing simulation-based clinical cases for team-based learning.* [KI-07]
>
> *AI offers real support for students preparing for bedside clinical reasoning.* [KI-04]

More significantly, a powerful and recurring subtheme was imperative for culturally and ethically relevant integration. Participants universally emphasized that the default outputs of global AI models were insufficient for their context. They stressed the necessity of actively localizing content to align with the UAE's specific socioreligious values, particularly concerning family structures, language, and ethical norms. This was not seen as a minor adjustment but as a fundamental requirement for the responsible and effective use of AI in their educational environment.

> *You cannot just rely on global AI content. You must contextualize for family, religion, and language.* [KI-06]
>
> *Localization is not optional, it is essential for meaningful learning in our setting.* [KI-05]

Participants also provided some examples of what cultural and ethical localization might look like in practice. One participant linked localization to personalized learning systems that could monitor learner progress and identify strengths and weaknesses, while also noting the privacy implications of such systems:

> *Learning and having eventually soon having systems that help with tracking learners, tracking their progress. Personalized learning… it can be much easier to identify students who are progressing faster than others and what areas exactly are their weaknesses and strengths.* [KI-09]

The same participant raised an important ethical concern.

> *There is one huge negative there, or not negative but concern, which is that the learners have to give up some privacy… have to agree that this system is going to monitor you.*

XSL•FO
RenderX

> *Let's say we want more restrictions, especially ethical, like an Islamic kind of thing… maybe we want to add more sections to the engine. This should be feasible, and then we want to feed it with our own… institution, feed it with our own data and make use of it definitely.* [KI-09]

These examples suggest that localization was discussed not only as a general principle but also as an emerging practical consideration, even if not yet implemented as a formalized institutional process.

(E)—Education Focus: Streamlining Delivery and Reimagining Assessment

The educational focus of AI use manifested in 2 key areas: refining assessment strategies and enhancing curriculum delivery. In the realm of assessment, faculty reported that AI was not just a tool for creating more questions, but for creating better-aligned and more authentic evaluations. By automating the generation of standard items (like MCQs), AI was perceived to free up faculty time and cognitive resources to focus on designing complex assessments that measure higher-order thinking skills, such as reflection and clinical problem-solving.

> *We can finally return to authentic assessments—reflection, problem-solving, not just MCQs.* [KI-08]

> *AI assists me in validating that my questions match course objectives.* [KI-01]

In terms of teaching and curriculum delivery, AI was integrated as a core workflow tool for instructional design. Faculty used it to draft lesson outlines, generate ideas for learning activities, and even simulate dialogue for feedback conversations. This use case positions AI as a collaborative thought partner in the planning stages of teaching, helping to structure and enrich the educational experience before it even reaches the classroom.

> *AI is part of my teaching workflow now. It drafts outlines and gives me ideas for activities.* [KI-02]

> *I even use it to match learning goals with appropriate teaching methods.* [KI-01]

Analytically, this shows that faculty are leveraging AI to enhance both ends of the educational process, such as the initial design of instruction and the final evaluation of learning, thereby strengthening the overall alignment of the curriculum. Analytically, this suggests that faculty are using AI not only to reduce workload but also to reinforce constructive alignment between teaching, assessment, and learning outcomes. In this sense, AI is being positioned as a tool that may support pedagogical coherence, even where its use remains primarily supportive rather than transformative.

(T)—Technology: A Pragmatic Approach Favoring Accessibility Over Specialization

The technology dimension revealed a distinctly pragmatic and cautious approach to tool selection. Rather than enthusiastically adopting a wide array of specialized educational AI platforms, faculty exhibited a strong preference for a few accessible, general-purpose tools such as ChatGPT, Notion AI, and Copilot. This suggests that low barriers to entry, including cost, ease of use, and immediate utility, are more significant drivers of adoption than advanced, niche features.

A notable finding was the explicit caution against technological overload. Faculty voiced a strategic reluctance to invest time in learning numerous platforms, advocating instead for a focused and mastery-based approach to a select few tools that demonstrably improved their core tasks.

> *We should not get lost in fancy tools. Just choose what improves your teaching or assessment.* [KI-01]

> *Using AI has to be realistic—we can't expect all tools to work with our curriculum.* [KI-09]

This pragmatic stance suggests that faculty adoption is shaped not by enthusiasm for technological novelty alone but by selective judgment about feasibility, usefulness, and curricular fit. It also indicates that institutional AI strategies may be more effective when they emphasize a focused, sustainable set of tools rather than a broad and potentially overwhelming range of platforms. This apparent tension reflected a distinction between present feasibility and future possibility. Faculty favored simple tools for current use because they were practical and manageable, while viewing more transformative AI applications as contingent on future training, infrastructure, and institutional support.

(S)—SAMR Model of Integration

Analyzing the data through the SAMR model revealed that current AI integration is predominantly at the substitution and augmentation levels, indicating a focus on enhancing efficiency rather than achieving pedagogical transformation. At the substitution level, AI directly replaced existing tools, for instance, in drafting assessment questions.

> *We used to spend hours on question writing—now AI does the first draft.* [KI-08]

At the augmentation level, AI provided functional improvements to existing tasks, such as generating more diverse question formats or providing faster feedback, thus augmenting faculty capabilities.

> *It augments how we teach—faster feedback, more questions, better materials.* [KI-08]

While instances of modification (significantly redesigning tasks) were rare, faculty optimism pointed toward the potential for redefinition (creating previously inconceivable tasks). They envisioned AI-powered simulations for clinical communication or adaptive tutoring systems, suggesting a trajectory toward more transformative integration in the future, even if not yet realized in practice.

> *I see potential in AI chat simulations for clinical communication training.* [KI-03]

> *Soon, AI will co-facilitate reflective learning and structured feedback.* [KI-07]

This analysis, through the SAMR lens, confirms that while AI is a valuable supportive tool, its use has not yet fundamentally redefined educational paradigms in this context. Analytically, this pattern indicates that current AI adoption remains concentrated at the level of workflow enhancement rather than

XSL•FO
RenderX

pedagogical redesign. The significance of this finding is that faculty are beginning to integrate AI into routine educational practice, but most uses have not yet altered the underlying structure of teaching and assessment in a transformative way.

Strategic Priorities for Future Implementation

For AI to be a constructive force in medical education, participants proposed that 2 interlinked priorities must be addressed: faculty preparedness and the prevention of inequities.

Faculty Readiness and Development

Faculty members call for structured training programs on prompt engineering, ethical use, and curriculum alignment.

> *Faculty development is urgently needed; many still don't know how to prompt properly.* [KI-02]
>
> *We need structured workshops on how to use AI ethically and effectively.* [KI-04]

Equity and Overdependence Risks

Equally important is the implementation of strategies to mitigate the risks of digital disparities and academic misconduct. Faculty expressed concerns about performance gaps resulting from digital disparities and unethical shortcuts taken by students.

> *AI may widen the performance gap between high- and low-GPA students.* [KI-04]
>
> *Some are copying AI responses without understanding; it's a dangerous trend.* [KI-09]

Together, these priorities show that faculty do not view AI integration as a purely technical implementation challenge. Rather, they frame it as an institutional and pedagogical transition requiring capability building, ethical safeguards, and attention to unequal patterns of access and use. This helps explain why participants' views were simultaneously optimistic about AI's potential and cautious about its educational consequences.

Discussion

Overview of AI Integration in UAE Medical Education

Faculty in this study generally reported positive perceptions of AI integration in medical education in the United Arab Emirates, particularly in relation to assessment, feedback, lesson planning, and personalized learning. At the same time, participants highlighted concerns regarding ethical considerations, regulatory oversight, and the need for culturally appropriate implementation. The findings suggest that AI integration in UAE medical education is expanding, with both faculty and students benefiting from its accessibility and versatility. The current use remains largely focused on NLP-based tools such as ChatGPT for tasks including content authoring, assessment generation, feedback, and communication simulation. Faculty also expressed optimism about the potential of AI-powered simulations and adaptive tools for communication and reflective learning. AI applications now span both preclinical and clinical modules, with notable innovations in team-based and problem-based learning environments. Although AI supports more personalized learning and better alignment between assessments and learning outcomes, participants raised concerns about equity and overreliance, emphasizing that unequal access, differences in digital literacy, and potential misuse could widen performance gaps among students.

Alignment With International Literature

These findings are consistent with international literature on the expanding role of generative AI tools, particularly LLMs such as ChatGPT, in medical education. Similar to other studies, participants recognized AI's potential to enhance productivity, deepen understanding of medical concepts, and enrich learning experiences for both educators and students [37]. Faculty primarily use AI to streamline routine tasks, such as creating MCQs and preparing lectures, rather than for full pedagogical transformation, which is consistent with existing literature [35]. While the international literature increasingly emphasizes the role of AI in enabling personalized learning through real-time feedback and adaptive content delivery [38-40], participants in this study expressed these possibilities more as future aspirations than as fully realized practices [38-40].

Cultural Localization and Ethical Considerations

While much of the Western literature focuses on concerns about data privacy and academic integrity [41], the findings of this study suggest a relatively stronger emphasis on responsible and context-sensitive adoption. Faculty perspectives reflected a balance between recognizing AI's educational potential and maintaining pedagogical integrity, indicating that adoption within the UAE context may be shaped not only by technological capability but also by institutional priorities and broader sociocultural considerations. This perspective supports calls for context-sensitive AI integration strategies, particularly in non-Western and culturally diverse settings [42,43]. Participants further expressed concerns regarding equity and the risk that AI could widen existing educational gaps, reflecting broader literature on the digital divide in AI usage [44]. Overreliance on AI tools emerged as another potential challenge, with participants noting the risk of undermining critical thinking and independent learning skills [45]. Finally, faculty emphasized a strong need for structured professional development, as the lack of formal training may hinder effective AI integration into teaching practice [46].

Theoretical Implications

Viewed through the FACETS framework, this study highlights the multidimensional nature of AI adoption in medical education. Faculty recognized AI's value in supporting diverse educational forms such as content generation and personalized learning and in enabling a range of practical AI use cases that align learning outcomes with curricular goals while maintaining ethically and technologically sound practices. In the UAE context, ethics and curricular contextualization were especially emphasized, reflecting the importance of culturally aligned, locally regulated, and pedagogically coherent implementation.

These findings extend the FACETS framework by demonstrating how contextual factors influence each dimension. In particular, cultural expectations, regulatory considerations, and institutional policies influence how AI tools are selected (technology), applied (AI use case), and embedded within teaching practices (education focus), highlighting that effective AI integration

XSL•FO
RenderX

depends not only on technological readiness but also on alignment with institutional and sociocultural norms [47].

Within this framework, the SAMR model illustrates the progression of AI adoption. Early integration occurs primarily at the substitution and augmentation levels, where technology enhances efficiency without altering pedagogy [36]. Faculty described using AI to automate tasks such as question generation, feedback preparation, and lesson planning, emphasizing that AI currently functions as a supportive tool rather than a transformative one. At the augmentation level, AI provides functional enhancements such as adaptive explanations and improved assessment design that extend faculty capabilities without altering instructional design [36]. The predominance of augmentation-level use may be explained by several interrelated factors. First, limited formal training and a lack of structured faculty development may constrain educators' ability to integrate AI in more pedagogically transformative ways. Second, institutional and regulatory uncertainties, including concerns about academic integrity, data privacy, and ethical use, may encourage cautious adoption. Third, infrastructural and curricular constraints, such as limited access to advanced AI tools or a lack of integration within existing curricula, may further restrict progression toward higher levels of the SAMR model. Together, these factors contribute to a pattern of incremental rather than transformative adoption.

Some faculty, however, expressed confidence in AI's potential for more advanced use, reflecting early movement toward modification and redefinition. At the modification stage, they envisioned AI supporting more interactive, student-centered learning through adaptive case discussions or personalized tutoring. At the redefinition stage, they anticipated AI-powered simulations and tailored feedback, enabling entirely new learning tasks. These beliefs are supported by emerging research exploring the educational potential of generative AI in medical education [48,49]. Overall, while the current use remains focused on enhancement, there is growing confidence that AI integration could progress toward more transformative applications.

Practical Implications

The findings have several practical implications for medical education policy and practice in the United Arab Emirates. Faculty members' generally positive attitudes toward AI demonstrate institutional readiness for structured adoption; however, the current use remains primarily task-enhancing rather than transformative. Medical schools should therefore develop comprehensive, curriculum-aligned strategies that support higher-level educational transformation through simulation-based learning, adaptive feedback, and decision-making support [50].

To support this transformation, institutions should prioritize faculty development through targeted training programs, interactive workshops, and interdisciplinary collaboration with AI specialists [51]. These efforts should focus on prompt engineering, ethical and critical evaluation of AI-generated outputs, and the effective integration of AI with learning objectives [52]. Such initiatives will not only enhance pedagogical innovation but also empower faculty to guide students in the responsible and informed use of AI.

At the policy level, medical schools need to establish frameworks addressing academic integrity, data privacy, and equitable access [53]. These policies must be informed by UAE-specific cultural and ethical considerations to ensure that AI adoption remains contextually relevant and socially responsible [54]. Moreover, equity-related concerns must be proactively addressed by improving digital literacy and ensuring fair access to AI tools among all students [44].

Fostering interdisciplinary collaboration among educators, AI specialists, and policymakers will be essential to ensure that the integration of AI in medical education evolves in a balanced, ethical, and inclusive manner, enhancing both learning quality and institutional sustainability [55].

Limitations and Future Research Recommendations

While this study provides novel insights into faculty perceptions of AI integration in medical education, several limitations should be noted. Although our study included faculty from both public and private institutions, providing some diversity in institutional perspectives, the sample was limited to 2 universities in the United Arab Emirates. Future research should include a broader representation of medical schools across the region to enhance transferability. Second, the study relied on qualitative data, offering an in-depth understanding but not providing quantifiable measures of attitudes or behaviors. Third, given the rapidly evolving nature of generative AI technologies, participants' experiences and perceptions may change as tools advance and institutional policies develop. Continuous monitoring and follow-up research are therefore necessary to examine how faculty adoption evolves over time. The findings should be interpreted within the context of qualitative inquiry, where the goal is analytic depth and contextual understanding rather than statistical generalizability. Although the sample was limited to 10 faculty members from 2 institutions, the study aimed to generate rich, context-specific insight rather than represent the full UAE medical faculty landscape. Readers should therefore consider the findings in terms of their transferability to similar educational settings rather than population-level generalization.

Future studies could adopt multi-institutional or cross-national designs to capture a broader range of perspectives and contextual influences. Additionally, exploring how AI integration affects learner outcomes, including motivation, critical thinking, and professionalism, would provide a more comprehensive understanding of its educational impact.

Conclusion

This study reveals positive views among UAE medical faculty members toward incorporating AI into the curriculum, highlighting its potential to enhance assessment, feedback, planning, and personalized learning. Faculty recognize AI's growing role for efficiency and educational alignment in preclinical and clinical settings. Early signs of deeper integration, such as adaptive simulations, indicate a shift toward more advanced use. However, concerns about ethics, regulation, culture, and digital equity highlight the need for cautious, context-aware adoption. Future efforts should focus on

XSL•FO
RenderX

enhancing institutional readiness and developing inclusive frameworks that foster responsible innovation. These insights can inform policy discussions and practice within similar institutional contexts in the United Arab Emirates and provide a foundation for further research.

Acknowledgments

The authors state that they made use of 2 generative AI tools during the preparation of this manuscript. DeepSeek and ChatGPT (OpenAI) were used exclusively for proofreading, language refinement, and enhancing the readability and clarity of the text. These tools assisted in improving sentence structure, grammar, and academic tone. The authors further attest that no AI tool was used as a coauthor or credited as an author, and no AI was used to generate or modify references. The use of AI was limited to nonsubstantive language and visual presentation support, consistent with JMIR guidelines on transparency and ethical reporting.

The authors gratefully acknowledge the late Dr Susan Waller for her valuable contribution to the research process. Her contribution and support during the development of this work are sincerely appreciated.

Funding

The authors declared no financial support was received for this work.

Data Availability

The datasets generated and/or analyzed during this study are not publicly available due to the confidential nature of the interviews and the risk of identifying participants, but they are available from the corresponding author upon reasonable request and with appropriate ethical approvals.

Authors' Contributions

Conceptualization: UH, MF
Data curation: AR
Formal analysis: AR, UH
Project administration: MF
Supervision: MF
Validation: AR, MF
Visualization: AR
Writing – original draft: AR, UH, MF
Writing – review and editing: AR, UH, MF

Conflicts of Interest

None declared.

Multimedia Appendix 1
Interview guide.
[DOCX File, 16 KB - mededu_v12i1e90192_app1.docx]

References

1. Wiljer D, Hakim Z. Developing an artificial intelligence-enabled health care practice: rewiring health care professions for better care. J Med Imaging Radiat Sci 2019 Dec;50(4 Suppl 2):S8-S14. [doi: 10.1016/j.jmir.2019.09.010] [Medline: 31791914]
2. Darcy AM, Louie AK, Roberts LW. Machine learning and the profession of medicine. JAMA 2016 Feb 9;315(6):551-552. [doi: 10.1001/jama.2015.18421] [Medline: 26864406]
3. Shortliffe EH, Sepúlveda MJ. Clinical decision support in the era of artificial intelligence. JAMA 2018 Dec 4;320(21):2199-2200. [doi: 10.1001/jama.2018.17163] [Medline: 30398550]
4. Hosny A, Parmar C, Quackenbush J, Schwartz LH, Aerts HJWL. Artificial intelligence in radiology. Nat Rev Cancer 2018 Aug;18(8):500-510. [doi: 10.1038/s41568-018-0016-5] [Medline: 29777175]
5. Niazi MKK, Parwani AV, Gurcan MN. Digital pathology and artificial intelligence. Lancet Oncol 2019 May;20(5):e253-e261. [doi: 10.1016/S1470-2045(19)30154-8] [Medline: 31044723]
6. Hogarty DT, Su JC, Phan K, et al. Artificial intelligence in dermatology—where we are and the way to the future: a review. Am J Clin Dermatol 2020 Feb;21(1):41-47. [doi: 10.1007/s40257-019-00462-6] [Medline: 31278649]
7. Ting DSW, Pasquale LR, Peng L, et al. Artificial intelligence and deep learning in ophthalmology. Br J Ophthalmol 2019 Feb;103(2):167-175. [doi: 10.1136/bjophthalmol-2018-313173] [Medline: 30361278]

XSL•FO
RenderX

8. Leivaditis V, Beltsios E, Papatriantafyllou A, et al. Artificial intelligence in cardiac surgery: transforming outcomes and shaping the future. Clin Pract 2025 Jan 14;15(1):17. [doi: 10.3390/clinpract15010017] [Medline: 39851800]
9. Guo W, Chen Y. Investigating whether AI will replace human physicians and understanding the interplay of the source of consultation, health-related stigma, and explanations of diagnoses on patients' evaluations of medical consultations: randomized factorial experiment. J Med Internet Res 2025 Mar 5;27:e66760. [doi: 10.2196/66760] [Medline: 40053785]
10. Montanaro B, Croce A, Ughetto E. Venture capital investments in artificial intelligence. J Evol Econ 2024 Jan;34(1):1-28. [doi: 10.1007/s00191-024-00857-7]
11. Omar G, Othman ZK, Kakarash ZA. The transformative impact of artificial intelligence (AI) on enhancing healthcare systems in the Middle East. Acad J Int Univ Erbil 2024 Nov. [doi: 10.63841/iue12417]
12. Alowais SA, Alghamdi SS, Alsuhebany N, et al. Revolutionizing healthcare: the role of artificial intelligence in clinical practice. BMC Med Educ 2023 Sep 22;23(1):689. [doi: 10.1186/s12909-023-04698-z] [Medline: 37740191]
13. Çalışkan SA, Demir K, Karaca O. Artificial intelligence in medical education curriculum: an e-Delphi study for competencies. PLoS One 2022;17(7):e0271872. [doi: 10.1371/journal.pone.0271872] [Medline: 35862401]
14. Abdellatif H, Al Mushaiqri M, Albalushi H, Al-Zaabi AA, Roychoudhury S, Das S. Teaching, learning and assessing anatomy with artificial intelligence: the road to a better future. Int J Environ Res Public Health 2022 Oct 31;19(21):14209. [doi: 10.3390/ijerph192114209] [Medline: 36361089]
15. Valikodath NG, Cole E, Ting DSW, American Academy of Ophthalmology Task Force on Artificial Intelligence, et al. Impact of artificial intelligence on medical education in ophthalmology. Transl Vis Sci Technol 2021 Jun 1;10(7):14. [doi: 10.1167/tvst.10.7.14] [Medline: 34125146]
16. Perrin J, Petronic-Rosic V. The potential role and restrictions of artificial intelligence in medical school dermatology education. Clin Dermatol 2024;42(5):477-479. [doi: 10.1016/j.clindermatol.2024.06.017] [Medline: 38925446]
17. Leon S, Lee S, Perez JE, Hashimoto DA. Artificial intelligence and the education of future surgeons. Am J Surg 2025 Aug;246:116257. [doi: 10.1016/j.amjsurg.2025.116257] [Medline: 39988540]
18. Crotty E, Singh A, Neligan N, Chamunyonga C, Edwards C. Artificial intelligence in medical imaging education: recommendations for undergraduate curriculum development. Radiography (Lond) 2024 Dec;30 Suppl 2:67-73. [doi: 10.1016/j.radi.2024.10.008] [Medline: 39454460]
19. Jowsey T, Stokes-Parish J, Singleton R, Todorovic M. Medical education empowered by generative artificial intelligence large language models. Trends Mol Med 2023 Dec;29(12):971-973. [doi: 10.1016/j.molmed.2023.08.012] [Medline: 37718142]
20. Liu M, Okuhara T, Chang X, et al. Performance of ChatGPT across different versions in medical licensing examinations worldwide: systematic review and meta-analysis. J Med Internet Res 2024 Jul 25;26:e60807. [doi: 10.2196/60807] [Medline: 39052324]
21. Kaya M, Sonmez E, Halici A, Yildirim H, Coskun A. Comparison of AI-generated and clinician-designed multiple-choice questions in emergency medicine exam: a psychometric analysis. BMC Med Educ 2025 Jul 1;25(1):949. [doi: 10.1186/s12909-025-07528-6] [Medline: 40597998]
22. Ebihara H, Kasai H, Shimizu I, et al. Development of a clinical clerkship mentor using generative AI and evaluation of its effectiveness in a medical student trial compared to student mentors: 2-part comparative study. JMIR Med Educ 2025 Sep 4;11:e76702. [doi: 10.2196/76702] [Medline: 40907969]
23. McLaughlin JE, Ponte CD, Lyons K. Student perceptions of GenAI as a virtual tutor to support collaborative research training for health professionals. BMC Med Educ 2025 Jul 1;25(1):895. [doi: 10.1186/s12909-025-07390-6] [Medline: 40597036]
24. Li L, Zhang W, Zhang K, et al. The role of generative AI tools in case-based learning and teaching evaluation of medical biochemistry. BMC Med Educ 2025 Aug 22;25(1):1185. [doi: 10.1186/s12909-025-07567-z] [Medline: 40847414]
25. Fatima SS, Sheikh NA, Osama A. Authentic assessment in medical education: exploring AI integration and student-as-partners collaboration. Postgrad Med J 2024 Nov 22;100(1190):959-967. [doi: 10.1093/postmj/qgae088] [Medline: 39041454]
26. Jackson P, Ponath Sukumaran G, Babu C, et al. Artificial intelligence in medical education—perception among medical students. BMC Med Educ 2024 Jul 27;24(1):804. [doi: 10.1186/s12909-024-05760-0] [Medline: 39068482]
27. Sami A, Tanveer F, Sajwani K, et al. Medical students' attitudes toward AI in education: perception, effectiveness, and its credibility. BMC Med Educ 2025 Jan 17;25(1):82. [doi: 10.1186/s12909-025-06704-y] [Medline: 39833834]
28. Ahmad MN, Abdallah SA, Abbasi SA, Abdallah AM. Student perspectives on the integration of artificial intelligence into healthcare services. Digit Health 2023;9:20552076231174095. [doi: 10.1177/20552076231174095] [Medline: 37312954]
29. McKell D, McCoy L, Niño DF. IAMSE artificial intelligence meeting survey: AI's impact on medical education faculty. Med Sci Educ 2024;34(6):1641-1646. [doi: 10.1007/s40670-024-02163-0] [Medline: 39758482]
30. Al Zahrani EM, Elsafi SH, Al Musallam LD, et al. Faculty perspectives on artificial intelligence's adoption in the health sciences education: a multicentre survey. Front Med (Lausanne) 2025;12:1663741. [doi: 10.3389/fmed.2025.1663741] [Medline: 41090120]
31. Brooks JV, Hughes D. Flipping the expert: faculty educator sensemaking during transition to an active learning-based curriculum. BMC Med Educ 2024 Jan 23;24(1):85. [doi: 10.1186/s12909-024-05039-4] [Medline: 38263065]

XSL•FO
RenderX

32. Acosta-Enriquez BG, Ramos Farroñan EV, Villena Zapata LI, et al. Acceptance of artificial intelligence in university contexts: a conceptual analysis based on UTAUT2 theory. Heliyon 2024;10(19):e38315. [doi: 10.1016/j.heliyon.2024.e38315] [Medline: 39430455]
33. Tong A, Sainsbury P, Craig J. Consolidated Criteria for Reporting Qualitative Research (COREQ): a 32-item checklist for interviews and focus groups. Int J Qual Health Care 2007 Dec;19(6):349-357. [doi: 10.1093/intqhc/mzm042] [Medline: 17872937]
34. Braun V, Clarke V. Toward good practice in thematic analysis: avoiding common problems and be(com)ing a knowing researcher. Int J Transgend Health 2023;24(1):1-6. [doi: 10.1080/26895269.2022.2129597] [Medline: 36713144]
35. Gordon M, Daniel M, Ajiboye A, et al. A scoping review of artificial intelligence in medical education: BEME Guide No. 84. Med Teach 2024 Apr;46(4):446-470. [doi: 10.1080/0142159X.2024.2314198] [Medline: 38423127]
36. Hamilton ER, Rosenberg JM, Akcaoglu M. The Substitution Augmentation Modification Redefinition (SAMR) model: a critical review and suggestions for its use. TechTrends 2016 Sep;60(5):433-441. [doi: 10.1007/s11528-016-0091-y]
37. Kung TH, Cheatham M, Medenilla A, et al. Performance of ChatGPT on USMLE: potential for AI-assisted medical education using large language models. PLOS Digit Health 2023 Feb;2(2):e0000198. [doi: 10.1371/journal.pdig.0000198] [Medline: 36812645]
38. Knopp MI, Warm EJ, Weber D, et al. AI-enabled medical education: threads of change, promising futures, and risky realities across four potential future worlds. JMIR Med Educ 2023 Dec 25;9:e50373. [doi: 10.2196/50373] [Medline: 38145471]
39. Lyo S, Mohan S, Hassankhani A, Noor A, Dako F, Cook T. From revisions to insights: converting radiology report revisions into actionable educational feedback using generative AI models. J Imaging Inform Med 2025 Apr;38(2):1265-1279. [doi: 10.1007/s10278-024-01233-4] [Medline: 39160366]
40. Parente DJ. Generative artificial intelligence and large language models in primary care medical education. Fam Med 2024 Oct;56(9):534-540. [doi: 10.22454/FamMed.2024.775525] [Medline: 39207784]
41. Chen Y, Esmaeilzadeh P. Generative AI in medical practice: in-depth exploration of privacy and security challenges. J Med Internet Res 2024 Mar 8;26:e53008. [doi: 10.2196/53008] [Medline: 38457208]
42. Omar M, Soffer S, Agbareia R, et al. Sociodemographic biases in medical decision making by large language models. Nat Med 2025 Jun;31(6):1873-1881. [doi: 10.1038/s41591-025-03626-6] [Medline: 40195448]
43. Cau R, Pisu F, Suri JS, Saba L. Addressing hidden risks: systematic review of artificial intelligence biases across racial and ethnic groups in cardiovascular diseases. Eur J Radiol 2025 Feb;183:111867. [doi: 10.1016/j.ejrad.2024.111867] [Medline: 39637580]
44. Hadar Shoval D. Artificial intelligence in higher education: bridging or widening the gap for diverse student populations? Educ Sci 2025 May;15(5):637. [doi: 10.3390/educsci15050637]
45. Ahmad SF, Han H, Alam MM, et al. Impact of artificial intelligence on human loss in decision making, laziness and safety in education. Humanit Soc Sci Commun 2023;10(1):311. [doi: 10.1057/s41599-023-01787-8] [Medline: 37325188]
46. Rincón EHH, Jimenez D, Aguilar LAC, Flórez JMP, Tapia Á, Peñuela CLJ. Mapping the use of artificial intelligence in medical education: a scoping review. BMC Med Educ 2025 Apr 12;25(1):526. [doi: 10.1186/s12909-025-07089-8] [Medline: 40221725]
47. Ngo TTA, Vo TTA, Phan MT. The psychology of AI adoption in education: university students' intentions to use large language models for learning from a TAM and TPB perspective. Acta Psychol (Amst) 2025 Nov;261:105789. [doi: 10.1016/j.actpsy.2025.105789] [Medline: 41124935]
48. Holderried F, Stegemann-Philipps C, Herschbach L, et al. A generative pretrained transformer (GPT)-powered chatbot as a simulated patient to practice history taking: prospective, mixed methods study. JMIR Med Educ 2024 Jan 16;10:e53961. [doi: 10.2196/53961] [Medline: 38227363]
49. Sardesai N, Russo P, Martin J, Sardesai A. Utilizing generative conversational artificial intelligence to create simulated patient encounters: a pilot study for anaesthesia training. Postgrad Med J 2024 Mar 18;100(1182):237-241. [doi: 10.1093/postmj/qgad137] [Medline: 38240054]
50. Tolentino R, Baradaran A, Gore G, Pluye P, Abbasgholizadeh-Rahimi S. Curriculum frameworks and educational programs in AI for medical students, residents, and practicing physicians: scoping review. JMIR Med Educ 2024 Jul 18;10:e54793. [doi: 10.2196/54793] [Medline: 39023999]
51. Mir MM, Mir GM, Raina NT, et al. Application of artificial intelligence in medical education: current scenario and future perspectives. J Adv Med Educ Prof 2023 Jul;11(3):133-140. [doi: 10.30476/JAMP.2023.98655.1803] [Medline: 37469385]
52. Chadha N, Popil E, Gregory J, Armstrong-Davies L, Justin G. How do we teach generative artificial intelligence to medical educators? Pilot of a faculty development workshop using ChatGPT. Med Teach 2025 Jan;47(1):160-162. [doi: 10.1080/0142159X.2024.2341806] [Medline: 38648540]
53. Franco D'Souza R, Mathew M, Mishra V, Surapaneni KM. Twelve tips for addressing ethical concerns in the implementation of artificial intelligence in medical education. Med Educ Online 2024 Dec 31;29(1):2330250. [doi: 10.1080/10872981.2024.2330250] [Medline: 38566608]
54. Mondal H. Ethical engagement with artificial intelligence in medical education. Adv Physiol Educ 2025 Mar 1;49(1):163-165. [doi: 10.1152/advan.00188.2024] [Medline: 39749996]

XSL•FO
RenderX

55. Ahsan Z. Integrating artificial intelligence into medical education: a narrative systematic review of current applications, challenges, and future directions. BMC Med Educ 2025 Aug 23;25(1):1187. [doi: 10.1186/s12909-025-07744-0] [Medline: 40849650]

Abbreviations

COREQ: Consolidated Criteria for Reporting Qualitative Research
FACETS: Form, AI Use Case, Context, Education Form, Technology, and SAMR
LLMs: large language models
MCQs: multiple-choice questions
NLP: natural language processing
SAMR: Substitution, Augmentation, Modification, Redefinition

Edited by A Stone; submitted 23.Dec.2025; peer-reviewed by K Kodikara, M Rohaim, S Guraya, S Saini; revised version received 20.May.2026; accepted 20.May.2026; published 13.Aug.2026.

Please cite as:
Rahma AT, Hafeez U, Farooq M
AI Integration in Undergraduate Medical Education: Qualitative Study of Faculty Perspectives in the United Arab Emirates
JMIR Med Educ 2026;12:e90192
URL: https://mededu.jmir.org/2026/1/e90192
doi:10.2196/90192

© Azhar T Rahma, Uffaira Hafeez, Munawar Farooq. Originally published in JMIR Medical Education (https://mededu.jmir.org), 13.Aug.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Education, is properly cited. The complete bibliographic information, a link to the original publication on https://mededu.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Multiagent Large Language Model Framework for Psychotherapy Fidelity Assessment in Motivational Interviewing and Cognitive Behavioral Therapy Training: Cross-Sectional, Simulation-Based Evaluation Study

Mohammad Amin Kamaleddin[1,2], PhD; Mina Mirjalili[3,4], PhD; Reza Barzegar[1], MSc; Nghia Trung Le[1], BSc; Zachary Cote[1], BSc; Olga Winkler[5], MD; Lisa Burback[5,6], MD; Yanbo Zhang[5,6], MD, PhD; Osmar R Zaiane[7,8], PhD; Candice Monson[9], PhD; Divya Sharma[10], PhD; Sri Krishnan[11], PEng, PhD; Andrew J Greenshaw[5,6], PhD; Richard Zeifman[12,13,14], PhD; Peter Selby[3,15,16], MD; Venkat Bhat[1,2,16], MSc, MD

[1]AI for Mental Health Program, St. Michael's Hospital, Unity Health Toronto, Toronto, ON, Canada

[10]Department of Mathematics and Statistics, York University, Toronto, ON, Canada

[11]Department of Electrical, Computer, and Biomedical Engineering, Toronto Metropolitan University, Toronto, ON, Canada

[12]NYU Langone Center for Psychedelic Medicine, Department of Psychiatry, NYU Grossman School of Medicine, New York, NY, United States

[13]Center for Psychedelic Research, Department of Brain Sciences, Imperial College London, London, United Kingdom

[14]Department of Psychology, The New School for Social Research, New York, NY, United States

[15]Department of Family and Community Medicine and Dalla Lana School of Public Health, University of Toronto, Toronto, ON, Canada

[16]Department of Psychiatry, University of Toronto, 250 College Street, 8th Floor, Toronto, ON, Canada

[2]Temerty Centre for Artificial Intelligence Research and Education in Medicine, University of Toronto, Toronto, ON, Canada

[3]Campbell Family Mental Health Research Institute, Centre for Addiction and Mental Health, Toronto, ON, Canada

[4]Adult Neurodevelopment and Geriatric Psychiatry Division, Centre for Addiction and Mental Health, Toronto, ON, Canada

[5]Department of Psychiatry, University of Alberta, Edmonton, AB, Canada

[6]Neuroscience and Mental Health Institute, University of Alberta, Edmonton, AB, Canada

[7]Department of Computing Science, University of Alberta, Edmonton, AB, Canada

[8]Alberta Machine Intelligence Institute, Edmonton, AB, Canada

[9]Department of Psychology, Toronto Metropolitan University, Toronto, ON, Canada

Corresponding Author:
Venkat Bhat, MSc, MD
AI for Mental Health Program, St. Michael's Hospital, Unity Health Toronto, Toronto, ON, Canada

Abstract

Background: Psychotherapy training is difficult to scale because manual rating of motivational interviewing (MI) and cognitive behavioral therapy (CBT) sessions is time-intensive, requires trained raters, and is subject to rater variability. Large language models (LLMs) may support simulation-based training and rubric-guided scoring, but early-stage evidence is needed before such systems can be applied to real learners.

Objective: This study aimed to conduct a simulation-based evaluation of a multiagent LLM framework for generating and scoring stylized MI and CBT training encounters.

Methods: We conducted a cross-sectional evaluation of a multiagent framework comprising Student, Patient, Evaluator, and Feedback agents. The Student agent conducted synthetic MI or CBT encounters with Patient agents derived from structured profiles. The Evaluator agent scored transcripts using study-specific MI and CBT scoring forms. Internal discrimination was tested across prompt-engineered novice, intermediate, and expert Student agent profiles using 133 MI and 102 CBT Patient profiles. Preliminary external grounding was assessed using 133 annotated motivational interviewing (AnnoMI) transcripts with coarse, metadata-derived, high/low session-level quality labels. Agreement with a pragmatic human-rater benchmark was evaluated using 16 independent human raters per modality. Criterion-level comparisons used paired Wilcoxon signed-rank tests with Benjamini-Hochberg false discovery rate correction at $q<0.05$. Agreement analyses used intraclass correlation coefficients (ICCs) with bootstrap 95% CIs. A secondary prompt-augmentation sensitivity analysis tested whether appending criterion-referenced feedback text to the novice Student-agent prompt shifted subsequent Evaluator-assigned scores.

Results: Evaluator scores increased across prompt-defined Student-agent competence levels. For MI, overall mean scores increased from 1.18 (95% CI 1.16 - 1.20) for novice profiles to 1.75 (95% CI 1.69 - 1.81) for intermediate profiles and to 3.57

XSL•FO
RenderX

(95% CI 3.48 - 3.66) for expert profiles. For CBT, overall mean scores increased from 0.83 (95% CI 0.78 - 0.88) to 2.18 (95% CI 2.10 - 2.26) and 4.38 (95% CI 4.28 - 4.48), respectively. Criterion-level planned contrasts were significant after false discovery rate correction. On AnnoMI transcripts, Evaluator scores aligned with metadata-derived high/low session labels, with 91.7% classification accuracy at the prespecified threshold. Human interrater reliability was an ICC(2,1) of 0.866 for MI and 0.769 for CBT. Evaluator-vs-human-consensus agreement was an ICC(2,1) of 0.959 for MI and 0.934 for CBT. In the secondary prompt-augmentation analysis, overall MI scores shifted from 1.18 to 1.44, and CBT scores shifted from 0.83 to 1.01.

Conclusions: This proof-of-concept study suggests that a rubric-guided multiagent LLM framework can score stylized synthetic MI and CBT transcripts along expected prompt-defined competence gradients and align with preliminary external and human-rater benchmarks. The study is innovative in separating synthetic learner, patient, scoring, and feedback-generation roles within a single simulation workflow, extending prior LLM work from plausible dialogue generation toward rubric-linked scoring. The findings support further development of scalable simulation tools for psychotherapy training research. Prospective studies with human trainees, standardized or real patients, and formal psychometric testing are required.

(JMIR Med Educ 2026;12:e92964) doi:10.2196/92964

KEYWORDS

psychotherapy; artificial intelligence; AI; large language models; motivational interviewing; cognitive behavioral therapy; treatment fidelity; clinical competence

Introduction

Providing opportunities to practice clinical counseling or psychotherapy skills before seeing patients is foundational to psychiatric and psychological education [1]. Traditional simulations using peers or trained standardized patients offer realistic yet controlled environments for skill acquisition and are particularly effective for practicing procedural and interpersonal skills [2-4]. However, simulation with standardized patients is resource-intensive to create, staff, and scale across cohorts, sites, and time zones, demanding substantial educator time and expertise. These challenges include variability in rater expertise and feedback quality, as well as the high cost of sustained supervision [5]. These constraints are especially salient for psychotherapy-focused encounters, where trainees must combine theoretical understanding of a therapeutic approach with repeated, coached practice to develop skillful listening, empathy, and collaborative problem-solving [6].

Recent reviews of virtual simulation in health professions education indicate that digital patient tools can support repeated, standardized communication-skills practice, but that evidence remains heterogeneous and many systems are still evaluated mainly on usability, satisfaction, or perceived realism rather than objective competency outcomes [7]. The rapid emergence of large language model (LLM)-based virtual patients has intensified both this opportunity and this concern: recent scoping-review evidence suggests that LLM-based virtual-patient research is expanding quickly but remains early-stage, with inconsistent assessment designs and limited use of validated tools or objective learning outcomes [8]. For psychotherapy education, this gap is especially important because competence is not captured by plausible dialogue alone; trainees must demonstrate model-specific behaviors such as reflective listening, evocation, autonomy support, collaborative agenda setting, and guided discovery [9].

Motivational interviewing (MI) [10] and cognitive behavioral therapy (CBT) [11] are evidence-based interventions taught across mental health training programs. Both have structured, model-anchored therapist behaviors, making them relevant candidates for AI-based tools that support rubric-guided fidelity assessment. Observer-rated fidelity and competence instruments, such as the Motivational Interviewing Treatment Integrity (MITI) [12] and Motivational Interviewing Skill Code (MISC) [13] for MI and the Cognitive Therapy Rating Scale (CTRS) [14] for CBT, provide structured scoring anchors for rating therapist behaviors and session structure from transcripts or recordings. However, applying these scoring tools consistently is labor-intensive and susceptible to rater drift.

Automating psychotherapy fidelity assessment has therefore become an important target for computational behavioral science [15]. Recent MI-focused work has shown that AI models can classify counselor and client MI behaviors in chat-based counseling data and may support scalable coding, adherence monitoring, and personalized feedback [16]. Other recent work has used LLM-based and sequential modeling approaches to estimate MI session quality, reinforcing the potential of automated transcript-level assessment while also underscoring the need for further validation against expert and field-collected data [17].

LLMs add a complementary capability to prior virtual-patient and automated-coding approaches: they can generate interactive patient dialogue, instantiate different learner or patient roles, and produce structured explanations or feedback within a single workflow [18-21]. Agentic designs use multiple specialized agents to simulate roles, critique outputs, and enforce constraints, enabling scalable practice through role separation, cross-agent checking, and standardized evaluation workflows with feedback aligned to therapeutic model criteria [22-24]. Recent LLM-based simulated-patient systems and agentic frameworks have shown feasibility for medical education, including scalable patient simulation, communication-skills practice, and automated feedback [25]. However, many such systems remain oriented toward history taking, clinical reasoning, or generic communication practice rather than psychotherapy-specific fidelity assessment [26,27]. In behavioral health, recent scholarship has emphasized that LLMs could support psychotherapy training and research, but that this is a high-stakes domain requiring transparent rubrics, careful

XSL•FO
RenderX

evaluation, human oversight, and attention to safety, bias, and the distinction between superficially empathic language and model-concordant clinical practice [28]. Emerging psychotherapy-specific studies have begun to evaluate AI-generated MI patient simulations and LLM-supported counseling feedback [29], but evidence remains limited regarding rubric-linked scoring sensitivity, agreement with human-rater benchmarks, feedback alignment, and evaluation across more than one psychotherapy modality.

To address these gaps, we conducted a preclinical, simulation-based evaluation of a rubric-guided, multiagent LLM framework comprising (1) a Patient agent that enacts structured Patient profiles through case-consistent responses; (2) a Student agent that conducts the synthetic session; (3) an Evaluator agent that scores Student-agent behavior against explicit criteria; and (4) a Feedback agent that generates criterion-referenced narrative recommendations. MI scoring used a study-specific MITI or MISC-derived global-rating form, and CBT scoring used a CTRS-derived form. MI Patient profiles were generated from a manually constructed template across common behavior-change contexts, whereas CBT Patient profiles were reformatted from the *Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition* (*DSM-5*) Clinical Cases into a standardized Patient-agent prompt structure [30]. Annotated motivational interviewing (AnnoMI) [31] was used separately for a preliminary external comparison of Evaluator scores with coarse metadata-derived MI session-level quality labels. These simulations were intended as stylized benchmark examples for early-stage system evaluation, not as evidence that LLM-generated sessions reproduce the ecological complexity of real psychotherapy encounters.

The aim of this study was to conduct a preclinical, cross-sectional, simulation-based evaluation of a rubric-guided, multiagent LLM framework for psychotherapy training and fidelity assessment in MI and CBT. The primary objective was to determine whether the Evaluator agent could discriminate among prompt-defined novice, intermediate, and expert Student-agent performance across standardized synthetic Patient-agent encounters. We hypothesized that Evaluator-assigned scores would increase monotonically across these competence levels. Secondary objectives were to assess whether Evaluator-agent scores aligned with external AnnoMI session-level MI quality labels, to estimate agreement between Evaluator-agent scores and a pragmatic human-rater benchmark, and to examine whether appending criterion-referenced feedback to novice Student-agent prompts would shift subsequent Evaluator-assigned scores. Exploratory analyses examined Student-agent response-consistency metrics across prompt-defined competence levels.

Methods

Study Design

This was a preclinical, cross-sectional, simulation-based evaluation of a multiagent LLM framework for psychotherapy training and fidelity scoring. The study used synthetic Patient-agent encounters and prompt-defined Student-agent profiles to evaluate (1) internal scoring sensitivity across prompt-defined competence levels, (2) preliminary external alignment with AnnoMI session-level labels, (3) agreement with a human-rater benchmark, (4) secondary prompt-augmentation sensitivity after feedback text was appended to the novice Student-agent prompt, and (5) exploratory Student-agent response-consistency metrics. Where applicable, reporting of the AI system, evaluation setting, and human comparators followed the DECIDE-AI (Developmental and Exploratory Clinical Investigations of Decision Support Systems Driven by AI) guideline for early-stage evaluation of AI decision-support systems (Checklist 1) [32].

Setting

Simulations and analyses were conducted computationally using Python or a LangGraph (LangChain) workflow. Human raters completed transcript-rating tasks remotely using synthetic or deidentified transcripts.

Inclusion and Exclusion

AnnoMI transcripts were included if they had an available high or low session-level quality label. Human-rater data were included if the rater completed all assigned transcript-rating fields.

Sampling Procedures

The MI and CBT Patient-agent profile sets were assembled purposively to provide broad coverage of behavior-change topics and diagnostic presentations for synthetic simulation, rather than to represent an epidemiologic or clinical population sample. MI Patient-agent profiles were generated using a manually written template across common MI-relevant behavior-change contexts. CBT Patient-agent profiles were reformatted from *DSM-5* Clinical Cases into standardized Patient-agent prompts. AnnoMI transcripts were included as a publicly available external comparison dataset. Physician raters were recruited purposively based on their clinical experience with MI and CBT. Additional independent human raters were recruited through Prolific using convenience sampling from the participant pool. Eligibility criteria required English as a first language.

Sample Size, Power, and Precision

The study size was determined pragmatically based on the available Patient-agent profile set, the eligible AnnoMI transcripts with session-level labels, and the feasibility of obtaining human ratings. We analyzed 133 MI Patient-agent profiles, 102 CBT Patient-agent profiles, 133 AnnoMI transcripts, and 10 selected sessions per modality for human rating by 16 raters. Precision was summarized with 95% CIs for key agreement metrics, classification metrics, and mean estimates.

Participant Characteristics

No real patients or real learners participated in the simulation experiments. Human involvement was limited to independent transcript rating. For each modality, 16 human raters scored the selected transcripts. The rater panel included 3 practicing physicians with clinical experience using MI and CBT, with a median of 5 (IQR 3-5) years of relevant clinical practice. Thirteen additional independent human raters were recruited

XSL•FO
RenderX

through Prolific. All raters provided informed consent and were compensated at a rate equivalent to US $20 per hour.

Patient-Agent Profiles

For MI, we constructed 133 Patient-agent profiles from a manually written template informed by MITI [12] and MISC [13] concepts and common MI-relevant behavior-change scenarios, including alcohol use, smoking cessation, weight management, and medication adherence. One profile was manually written as a template specifying required sections: role-prompt field, demographic and presenting details, behavioral presentation, MI-relevant elements, mental-status cues, and simulation-prompt field. We then used single-example template prompting, meaning that the LLM was given one completed exemplar profile, the required output fields, and instructions to generate additional Patient-agent profiles in the same structure. MI profile generation used GPT-4.1-mini. Demographic and clinical variation was introduced through prompt-specified variation in age, sex or gender, cultural background, presenting problem, readiness to change, sustain talk, change talk, and ambivalence (Table S1 in Multimedia Appendix 1).

For CBT, we constructed 102 Patient-agent profiles by reformatting cases from *DSM-5* Clinical Cases [30] into a standardized Patient-agent format using GPT-4.1-mini. Each profile included a case synopsis, behavioral presentation, symptom list, structured patient history, and diagnosis. The profile set was intended to provide broad diagnostic and cognitive or behavioral diversity for synthetic CBT simulations rather than a representative sample of CBT patients (Table S2 in Multimedia Appendix 1).

AnnoMI External Comparison Dataset

We used 133 publicly available AnnoMI counseling transcripts for external comparison. AnnoMI includes expert utterance-level annotations; however, the present analysis did not use utterance-level behavior codes. Instead, we compared Evaluator-assigned global MI scores with the corpus-level high or low session quality labels, which are coarse, metadata-derived labels based on the original video context. This analysis was therefore treated as preliminary external grounding rather than as validation against expert fidelity coders.

Measures and Covariates

Primary outcomes were criterion-level Evaluator scores within MI and CBT. MI scores were assigned across 9 criteria on a 1 to 5 scale using a study-specific MITI or MISC-derived global-rating form. CBT scores were assigned across 11 criteria on a 0 to 6 scale using a CTRS-derived scoring form [14]. No demographic or clinical covariates were modeled because Patient-agent profiles were synthetic and were not designed as a representative clinical sample.

Evaluator Agent Rubric Construction

For MI, we created a study-specific MITI or MISC-derived rubric rather than applying the full MITI or MISC instruments. The purpose was to provide a compact global-rating form suitable for short simulated transcripts. Items were selected if they represented core MI fidelity constructs and could be judged at the transcript level as global performance domains. The 9 selected domains were Cultivating Change Talk, Softening Sustain Talk, Partnership, Empathy, Acceptance, Direction, Autonomy Support, Collaboration, and Evocation. MITI-derived domains were retained as global 1 to 5 ratings. MISC-derived domains were harmonized into the same 1 to 5 ordinal global-rating format using the source construct definitions as behavioral anchors. MISC behavior-count codes were not used as count variables.

The MI scoring form used a 1 to 5 global-rating scale for each criterion (1="absent, clearly inconsistent with the construct, or strongly nonadherent"; 2="minimal or mostly ineffective evidence"; 3="partial, inconsistent, or mixed evidence"; 4="generally competent and mostly consistent evidence"; and 5="strong, sustained, and consistently adherent evidence across the transcript"). Item-specific anchor descriptions are provided in Table S3 in Multimedia Appendix 1.

For CBT, we used the 11 CTRS-derived domains using the original 0 to 6 structure (0="poor or absent evidence of the competency"; 2="limited or partially adequate performance"; 4="good or competent performance"; and 6="excellent, sustained, and well-integrated performance"). Scores of 1, 3, and 5 were used for intermediate judgments between adjacent anchors (Table S4 in Multimedia Appendix 1).

For both approaches, the Evaluator generated criterion-level scores linked to behavioral anchors, and the Feedback agent generated criterion-referenced narrative recommendations based on the transcript, scoring criteria, and scored evaluation output (Tables S5 and S6 in Multimedia Appendix 1).

Narrative Feedback Generation and Alignment

The Evaluator agent assigned criterion-level scores. The Feedback agent then generated narrative recommendations using the transcript, scoring criteria, and the scored evaluation output. Feedback outputs were structured by criterion and included the assigned score, a transcript-grounded explanation, a short summary, and actionable recommendations. We treated narrative-feedback alignment as structural consistency between the numeric score, the criterion-specific explanation, and the recommendations provided. Representative aligned MI and CBT feedback excerpts are provided in Table S13 in Multimedia Appendix 1.

Multiagent System and Model Configuration

We implemented a role-separated multiagent framework comprising Student, Patient, Evaluator, and Feedback agents to simulate MI and CBT encounters, score transcripts, and generate criterion-referenced narrative feedback. The system was implemented in Python version 3.11 and orchestrated using LangGraph to support stateful workflows. The Student and Patient agents used GPT-4.1-mini. The Student agent used a temperature of 0.5 to balance adherence to the prompt-defined competence profile with conversational flexibility, whereas the Patient agent used a temperature of 1.0 to allow greater variability in synthetic patient responses. The Evaluator agent used Gemini 2.5 Flash Lite with a temperature of 0.2 to promote more stable rubric scoring and to assign the scoring role to a different model family from the Student and Patient generators.

XSL•FO
RenderX

Narrative recommendations were generated by a separate Feedback agent using Claude Haiku 4.5 with a temperature of 0.2. Full model versions, providers, routing endpoints, temperatures, and token settings are reported in Table S14 in Multimedia Appendix 1.

At runtime, the system loaded the selected approach and Patient-agent profile, executed a fixed-length exchange loop, and then called the Evaluator agent on the full transcript to generate criterion-level scores. The Feedback agent then generated criterion-referenced narrative recommendations.

Student-Agent Conditions

To evaluate internal discrimination, we engineered 3 Student-agent prompts for each approach: novice, intermediate, and expert. These profiles were prompt-defined competence levels intended to generate stylized benchmark transcripts rather than represent the full distribution of real trainee performance. The novice, intermediate, and expert profiles were used to create controlled behavioral contrasts. This allowed us to test whether the Evaluator was sensitive to increases in prompt-defined rubric adherence under standardized Patient-agent and session-length conditions.

Prompts were self-contained and encoded the communication style, language patterns, and modality-specific behaviors. Each simulated session was capped at 10 therapist-patient exchanges, in which one exchange was defined as 1 Student-agent turn followed by 1 Patient-agent response. This fixed length was selected a priori as a pragmatic standardization constraint to keep dialogue length constant across Patient-agent profiles and Student-agent conditions, and to approximate a brief, focused training encounter rather than a full psychotherapy session.

The novice profile represented low rubric adherence, characterized by directive communication, limited reflection, weak collaboration, poor agenda setting, premature advice, or technique use, and limited use of modality-specific strategies. The intermediate profile represented partial and inconsistent rubric adherence, with some appropriate MI or CBT behaviors but recurrent omissions or poorly timed interventions. This condition was added to avoid relying solely on maximally separated novice or expert endpoints. The expert profile represented a high-adherence benchmark, with consistent use of modality-concordant behaviors such as reflective listening, autonomy support, evocation, collaborative agenda setting, Socratic questioning, case formulation, and structured homework planning.

Student-agent profiles were prompt-engineered AI agents, not human learners. The novice, intermediate, and expert conditions were designed to create stylized benchmark transcripts with expected differences in MI- or CBT-consistent behaviors. These conditions were used to test the internal Evaluator's sensitivity to prompt-defined competence levels. Student prompts for MI and CBT, including novice, intermediate, expert, and novice+feedback prompt variants, are provided in Tables S7–S12 in Multimedia Appendix 1.

To prevent application-level carryover between trials, each Patient-profile × Student-condition run was executed in an independent session state. For each run, the system initialized empty interview and evaluator message histories, loaded the selected Patient-agent profile and Student-agent prompt, and generated a new transcript without access to prior transcripts, scores, feedback, or dialogue histories from other runs. Static resources, including model weights, model configurations, rubric prompts, and profile files, were reused as experimental inputs; however, session-level conversational state was not reused.

Human Rater Scoring and Reliability Analysis

A pragmatic human-rater benchmark was assembled to compare Evaluator-agent scores with independent human ratings. For each modality, 16 human raters scored the selected transcripts. The panel included 3 practicing physicians with clinical experience using MI and CBT, with a median of 5 years of relevant clinical practice. Thirteen additional independent raters were recruited through Prolific. All raters provided informed consent and were compensated at a rate equivalent to US $20 per hour. Each rater independently scored 10 MI sessions across 9 MITI or MISC-derived criteria and 10 CBT sessions across eleven CTRS-derived criteria. Raters used the study-specific MI scoring form and the CTRS-derived CBT scoring form without access to other raters' scores or Evaluator-agent outputs. Source manuals and item descriptions were provided as conceptual guidance.

Agreement analyses used session × criterion cells as the unit of analysis: 90 observations for MI and 110 observations for CBT. The intraclass correlation coefficient (ICC) (2,1) was used for 2-way random-effects, absolute-agreement, and single-rating reliability. Percentile-bootstrap 95% CIs were computed by resampling session × criterion observations. Because an Evaluator-vs-mean-consensus ICC is not structurally equivalent to a human interrater ICC, we also computed Evaluator-vs-individual-rater ICCs.

Prompt-Augmentation Sensitivity Analysis

To assess whether criterion-referenced feedback text altered subsequent Student-agent behavior, we conducted a secondary prompt-augmentation sensitivity analysis. For each Patient-agent profile, the novice Student agent first completed a 10-exchange session. The recommendation section from the feedback report was then extracted and appended to the same novice Student-agent prompt before we ran a second 10-exchange session as a new, independent session state with the same Patient-agent profile. The underlying novice prompt was otherwise unchanged. This analysis was designed to test system responsiveness to feedback text within the simulation pipeline.

Exploratory Student-Agent Dispersion Analysis

As an exploratory proxy for generative consistency, we analyzed token-level log probabilities from Student-agent responses. For each Student-agent turn, token log probabilities were extracted during GPT-4.1-mini generation, clipped at −10 to reduce the influence of sentinel or near-zero-probability tokens, and averaged across tokens. The session-level Student-agent average log probability was then computed as the mean across the 10 Student-agent turns. For interpretability, this value was transformed to the geometric mean token probability: exp(the session average log probability) × 100. We summarized

XSL•FO
RenderX

variability across Patient-agent profiles within each Student-agent condition using SD and IQR, with bootstrap CIs.

Missing Data

Missingness was assessed for generated transcripts, Evaluator-assigned criterion scores, AnnoMI labels, and human-rater scores. No missing data were observed in the final analytic datasets; therefore, no imputation was performed.

Data Analysis

Criterion-level scores were analyzed within MI (9 criteria, 1 - 5 scale) and CBT (11 criteria, 0 - 6 scale). Internal discrimination analyses compared novice, intermediate, and expert Student-agent conditions within Patient-agent profiles. Paired comparisons across Student-agent conditions and pre- or post-prompt-augmentation comparisons were conducted using 2-sided Wilcoxon signed-rank tests with a significance α level of .05. For criterion-level analyses, Benjamini-Hochberg false discovery rate (FDR) correction was applied within each modality and planned contrast family. Adjusted q values are reported for criterion-level comparisons, with $q<0.05$ considered statistically significant. Overall-average comparisons were treated as separate planned summary tests and are reported with their corresponding P values, absolute mean changes, and effect sizes. Effect sizes were reported as rank-biserial correlations.

For mean scores and mean changes, 95% CIs were estimated using nonparametric bootstrap resampling across Patient-agent profiles. For the AnnoMI external comparison, Evaluator-assigned global MI scores were compared with coarse metadata-derived high or low session-level quality labels. Scores less than or equal to the prespecified threshold were classified as low quality, and scores greater than the threshold were classified as high quality. Accuracy, precision, recall, and F_1-score were computed with high quality as the positive class and reported with 95% CIs. Human-rater reliability analyses used ICC(2,1), weighted Cohen κ, mean within-cell SD, mean absolute error, root mean square error, Pearson correlation, and Spearman correlation, as applicable. All statistical analyses were performed using Python version 3.11. Nonparametric procedures were selected because rubric scores were ordinal and potentially nonnormally distributed.

Ethical Considerations

This work was conducted as a technical feasibility and simulation-based benchmarking study of a multiagent LLM framework prior to future controlled educational or clinical evaluation studies. The simulation and secondary-data components used synthetic agent-generated transcripts and deidentified source materials. No real patients or real learners participated; no identifiable patient information was used; and no clinical decisions were informed by the study outputs. In accordance with the Tri-Council Policy Statement: Ethical Conduct for Research Involving Humans–TCPS 2 [33], the simulation and secondary-data components were determined not to require research ethics board review.

Human involvement was limited to the independent evaluation of synthetic or deidentified transcripts. Raters were informed about the purpose and procedures of the rating task, provided informed consent before participation, and were compensated at a rate equivalent to US $20 per hour. Raters did not interact with patients, learners, or clinical systems. No personal health information or sensitive clinical information was collected from raters. Rater data were stored and analyzed in deidentified form and reported only in aggregate. Formal Research Ethics Board review was not sought because this work was classified as minimal-risk technical feasibility and simulation-based benchmarking research. This classification was based on the use of synthetic or deidentified transcripts, the limited scope of human-rater involvement, the absence of patient or learner interaction, the absence of personal health information or sensitive participant data, deidentified analysis, and aggregate reporting only. Future studies designed to evaluate real learners, standardized or real patients, educational effectiveness, clinical performance, or identifiable participant information will undergo formal Research Ethics Board review prior to implementation.

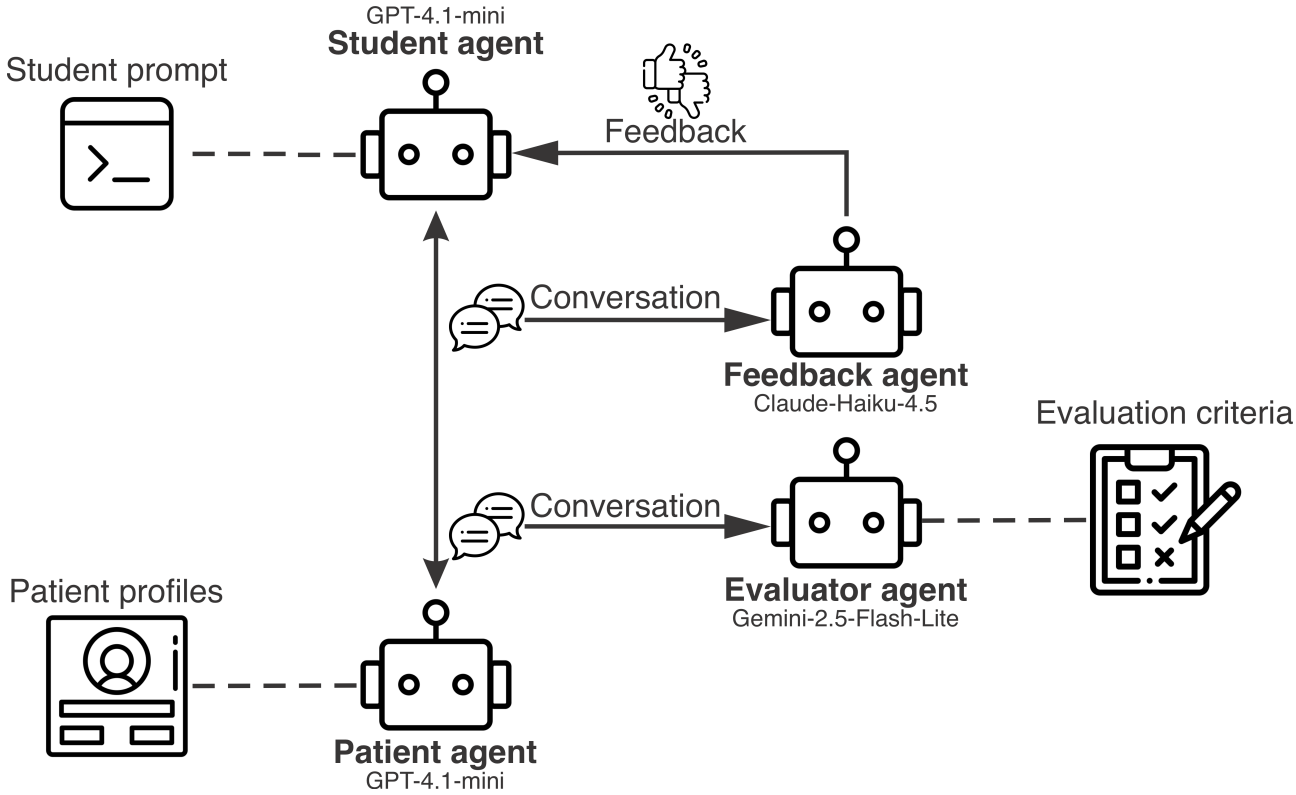
Results

Overview of the Evaluation Framework

The evaluation consisted of fixed-length synthetic encounters between a Student agent and structured Patient agents in 2 psychotherapeutic approaches, MI and CBT, with Student-agent performance scored by an Evaluator agent using predefined rubric criteria (Figure 1). For MI, the Evaluator used a study-specific, 9-criterion MITI or MISC–derived global-rating form on a 1 to 5 scale (Table S4 in Multimedia Appendix 1). For CBT, the Evaluator applied an 11-item CTRS–derived form on a 0 to 6 scale (Table S4 in Multimedia Appendix 1). All sessions used the fixed 10-exchange format defined in the "Methods" section. Scores were averaged across 133 MI and 102 CBT Patient-agent profiles for each of the 3 prompt-defined Student-agent competence profiles: novice, intermediate, and expert.

XSL•FO
RenderX

Figure 1. Multiagent large language model framework for stylized psychotherapy-training simulation and rubric-based evaluation. The system comprises role-separated Student, Patient, Evaluator, and Feedback agents coordinated within a structured dialogue workflow. The Student agent, instantiated with GPT-4.1-mini, conducts the synthetic psychotherapy encounter based on a predefined prompt. The Patient agent, also instantiated with GPT-4.1-mini, enacts a structured Patient-agent profile and generates case-consistent responses during the fixed-length dialogue. The Evaluator agent, instantiated with Gemini 2.5 Flash Lite, scores the completed transcript against predefined evaluation criteria. The Feedback agent, instantiated with Claude Haiku 4.5, generates criterion-referenced narrative recommendations based on the transcript, scoring criteria, and the scored evaluation output. Feedback text could be appended to the novice Student-agent prompt in a secondary prompt-augmentation sensitivity analysis. Full model configurations, temperatures, and routing details are reported in Table S14 in Multimedia Appendix 1.

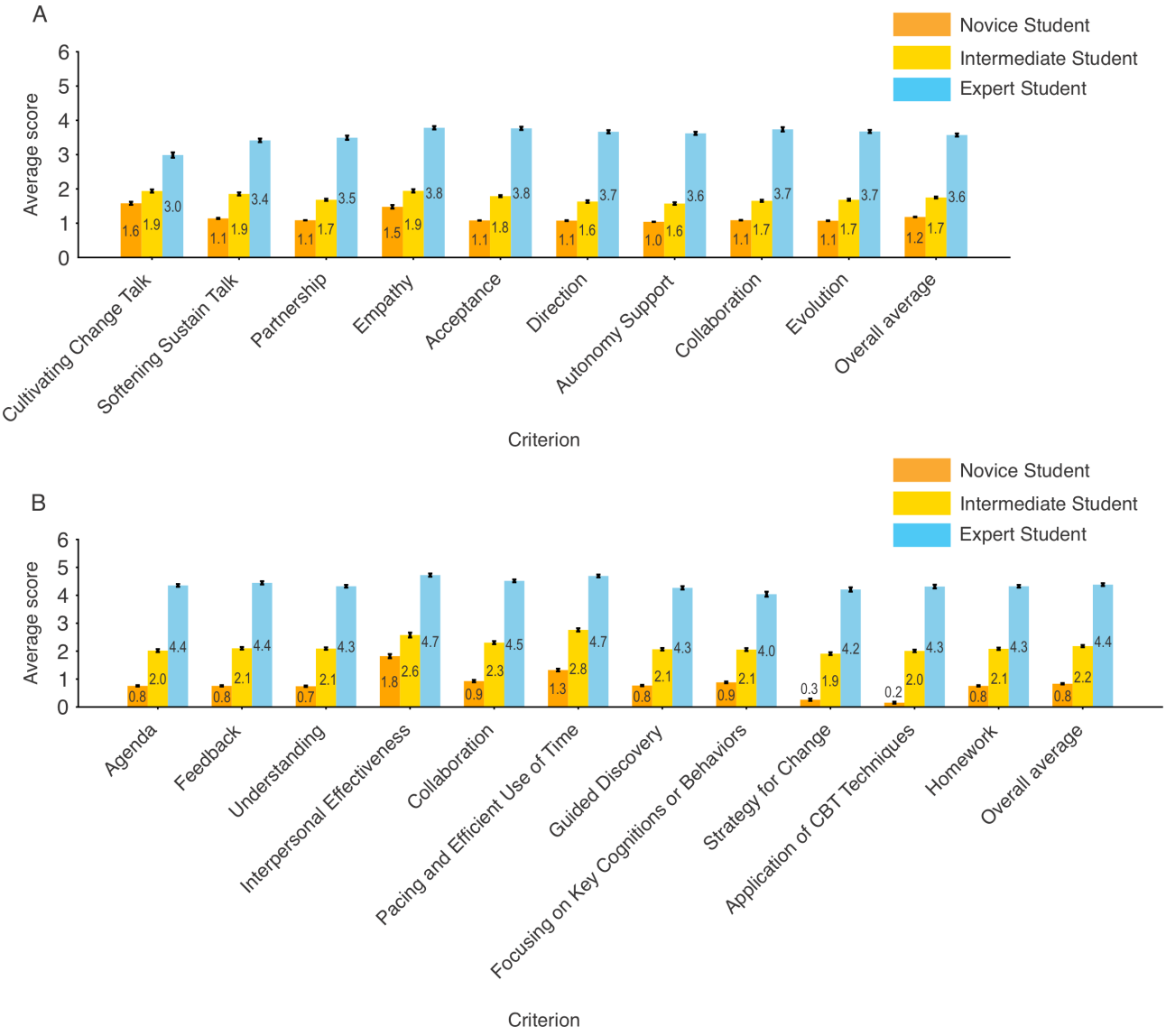


Internal Discrimination Across Prompt-Defined Competence Levels

Evaluator scores showed graded increases across the 3 prompt-defined student-agent competence conditions. For MI, the overall mean score increased from 1.18 (95% CI 1.16 - 1.20) for the novice profile to 1.75 (95% CI 1.69 - 1.81) for the intermediate profile and to 3.57 (95% CI 3.48 - 3.66) for the expert profile (Figure 2A). For CBT, the overall mean score increased from 0.83 (95% CI 0.78 - 0.88) to 2.18 (95% CI 2.10 - 2.26) and 4.38 (95% CI 4.28 - 4.48) for the intermediate and expert profiles, respectively (Figure 2B). All criterion-level planned contrasts across student-agent competence conditions remained significant after Benjamini-Hochberg false discovery rate correction (q<0.05). These findings support the Evaluator agent's sensitivity to expected differences across stylized prompt-defined competence levels, but should be interpreted as an internal scoring-sensitivity benchmark rather than evidence of real-world learner assessment.

XSL•FO
RenderX

Figure 2. Evaluator-scored performance across prompt-defined novice, intermediate, and expert Student-agent profiles. (A) Average criterion-level scores for motivational interviewing (MI) across 133 simulated Patient-agent profiles, evaluated using a study-specific Motivational Interviewing Treatment Integrity or Motivational Interviewing Skill Code (MITI/MISC)–derived 1 to 5 global-rating form. (B) Average criterion-level scores for cognitive behavioral therapy (CBT) across 102 simulated Patient-agent profiles, evaluated using a Cognitive Therapy Rating Scale (CTRS)–derived 0 to 6 scoring form. Bars show mean scores and error bars show the SE of the mean. Scores demonstrate graded separation across prompt-engineered competence levels. Criterion-level significance testing used paired Wilcoxon signed-rank tests with Benjamini-Hochberg false discovery rate correction within each modality and planned contrast; adjusted *q* values were used to determine significance. These analyses evaluate internal scoring sensitivity across stylized prompt-defined Student-agent profiles.



External Comparison Using AnnoMI Session-Level Quality Labels

We next examined whether Evaluator scores aligned with external session-level quality labels in the AnnoMI corpus (Figure 3). AnnoMI contains expert utterance-level annotations [34], but this analysis used only the coarse high or low session-level quality labels associated with the source videos. Metadata-labeled high-quality sessions clustered primarily at Evaluator scores of 3 to 5, whereas metadata-labeled low-quality sessions clustered primarily at scores of 1 to 2. Using a threshold of 2, where scores ≤2 were classified as low quality and scores >2 were classified as high quality, yielded 91.7% accuracy, 95.4% precision, 94.5% recall, and an F_1-score of 95.0%.

XSL•FO
RenderX

Figure 3. External comparison of Evaluator-agent motivational interviewing scores with annotated motivational interviewing (AnnoMI) metadata-derived quality labels. (A) Distribution of Evaluator-assigned motivational interviewing scores stratified by metadata-derived high or low session-level quality labels in the AnnoMI corpus. Bars show the percentage of sessions within each quality-label group receiving each score: low label (n=23) and high label (n=110). (B) Classification performance across Evaluator-score thresholds. Scores less than or equal to the threshold were classified as low quality, and scores greater than the threshold were classified as high quality. Performance metrics were computed with high quality as the positive class. A threshold of 2 yielded accuracy of 91.7%, precision of 95.4%, recall of 94.5%, and F_1-score of 95.0%. These metrics reflect alignment with coarse metadata-derived session-level labels.

A



B

Score cutoff high if score > cutoff	True Positive	False Negative	False Positive	True Negative	Accuracy, (%)	Precision, (%)	Recall, (%)	F_1-score, (%)
1	109	1	13	10	89.5	89.3	99.1	94.0
2	104	6	5	18	91.7	95.4	94.5	95.0
3	78	32	3	20	73.7	96.3	70.9	81.7
4	26	84	2	21	35.3	92.9	23.6	37.7

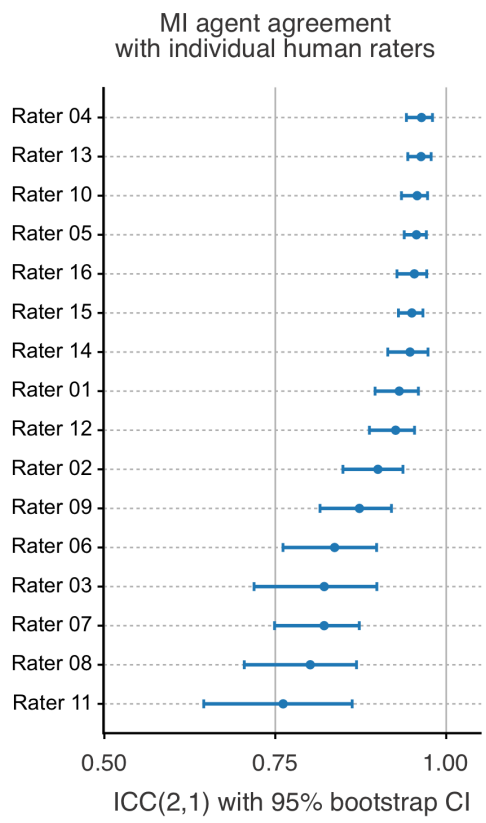
Agreement Between Evaluator Scores and Human Raters

We compared Evaluator-agent scores with ratings from a pragmatic 16-rater human benchmark (Figures 4 and 5). Raters independently scored 10 selected MI sessions across 9 MITI- or MISC-derived criteria and 10 selected CBT sessions across 11 CTRS-derived criteria, yielding 90 and 110 session × criterion observations, respectively. Session-level heatmaps showed broadly similar, though not identical, rating patterns across the human-rater consensus and Evaluator-agent scores in both approaches.

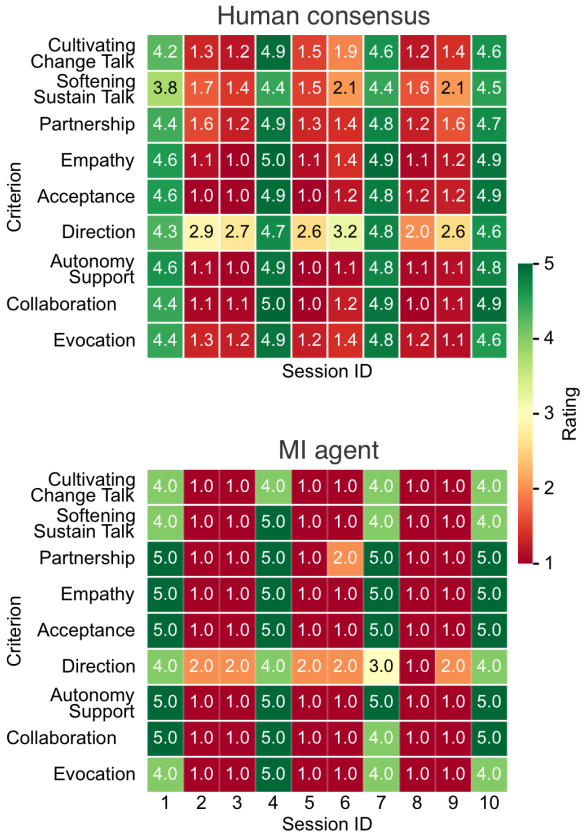
XSL•FO
RenderX

Figure 4. Human rater reliability and Evaluator-agent agreement with individual raters and the human-rater consensus for motivational interviewing (MI) scoring. Sixteen human raters per modality independently scored selected transcripts using the study-specific scoring forms. (A) Evaluator-agent agreement with each individual human rater for motivational interviewing (MI), shown as the intraclass correlation coefficient (ICC; 2,1) with 95%-bootstrap CIs between 90 session × criterion observations. (B) Human interrater reliability for MI across 16 human raters. (C) MI heatmaps comparing the mean human-rater consensus with Evaluator-agent scores. (D) Evaluator-agent agreement with the MI human-rater consensus. MAE: mean absolute error; RMSE: root mean square error.

A



C



B

Human interrater reliability for MI	
90 observations (10 sessions × 9 criteria) 16 human raters	
Primary metric	
ICC(2,1)	0.866 (good)
Supporting metrics	
Mean weighted κ	0.737
Mean SD per observation	0.552
Median SD per observation	0.500

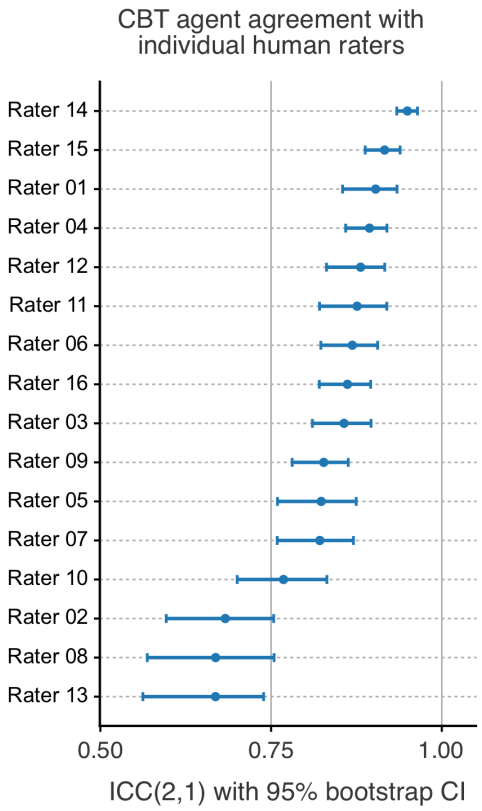
D

MI agent vs human consensus	
90 observations (10 sessions × 9 criteria) 16 human raters	
Correlation metrics	
ICC(2,1)	0.959 (excellent)
Pearson *r*	0.969 (*P*<.001)
Spearman *ρ*	0.878 (*P*<.001)
Error metrics	
MAE	0.366
RMSE	0.490
Mean difference	–0.234
SD of difference	0.433
Distribution of differences	
Within 1.0 point, %	95.6
Within 2.0 points, %	100

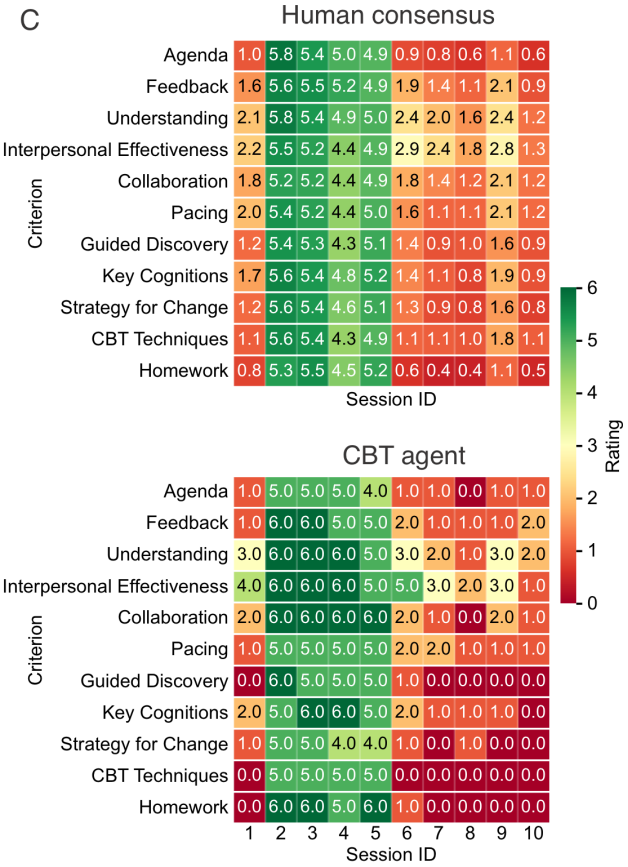
XSL•FO
RenderX

Figure 5. Human rater reliability and Evaluator-agent agreement with individual raters and the human-rater consensus for cognitive behavioral therapy (CBT) scoring. Sixteen human raters per modality independently scored selected transcripts using the study-specific scoring forms. (A) Evaluator-agent agreement with each individual human rater for cognitive behavioral therapy (CBT), shown as ICC(2,1) with 95% percentile-bootstrap CIs across 110 session × criterion observations. (B) Human interrater reliability for CBT across 16 human raters. (C) CBT heatmaps comparing the mean human-rater consensus with Evaluator-agent scores. (D) Evaluator-agent agreement with the CBT human-rater consensus. ICC: intraclass correlation coefficient; MAE: mean absolute error; RMSE: root mean square error.

A



C



B

Human interrater reliability for CBT	
110 observations (10 sessions × 11 criteria) 16 human raters	
Primary metric	
ICC(2,1)	0.769 (good)
Supporting metrics	
Mean weighted κ	0.550
Mean SD per observation	1.014
Median SD per observation	0.999

D

CBT agent vs human consensus	
110 observations (10 sessions × 11 criteria) 16 human raters	
Correlation Metrics	
ICC(2,1)	0.934 (excellent)
Pearson *r*	0.948 (*P*<.001)
Spearman *ρ*	0.877 (*P*<.001)
Error metrics	
MAE	0.622
RMSE	0.765
Mean difference	–0.072
SD of difference	0.765
Distribution of differences	
Within 1.0 point, %	80.9
Within 2.0 points, %	99.1

Human interrater reliability was good for MI, with an ICC(2,1) of 0.866 (95% CI 0.825 - 0.898) across 90 session × criterion observations, a mean weighted κ of 0.737, a mean SD per observation of 0.552, and a median SD per observation of 0.500.

For CBT, human interrater reliability was an ICC(2,1) of 0.769 (95% CI 0.732 - 0.799) across 110 observations, a mean weighted κ of 0.550, a mean SD per observation of 1.014, and a median SD per observation of 0.999.

XSL•FO
RenderX

Evaluator-vs-human-consensus agreement was high for MI, with ICC(2,1) of 0.959 (95% CI 0.940 - 0.973), Pearson *r* of 0.969, Spearman ρ of 0.878, mean absolute error (MAE) of 0.366, and root mean square error (RMSE) of 0.490. For CBT, Evaluator-vs-human-consensus agreement was ICC(2,1) of 0.934 (95% CI 0.913 - 0.950), Pearson *r* of 0.948, Spearman ρ of 0.877, MAE of 0.622, and RMSE of 0.765.

Because agreement with an averaged human consensus is not directly comparable with interrater reliability among individual humans, we additionally computed Evaluator-vs-individual-rater ICCs. These ranged from 0.761 to 0.964 for MI and from 0.669 to 0.950 for CBT. In leave-one-human-out analyses, individual human raters compared with the mean of the remaining human raters showed median ICCs of 0.944 for MI and 0.877 for CBT. These analyses contextualize the Evaluator agent's agreement with both individual raters and consensus ratings.

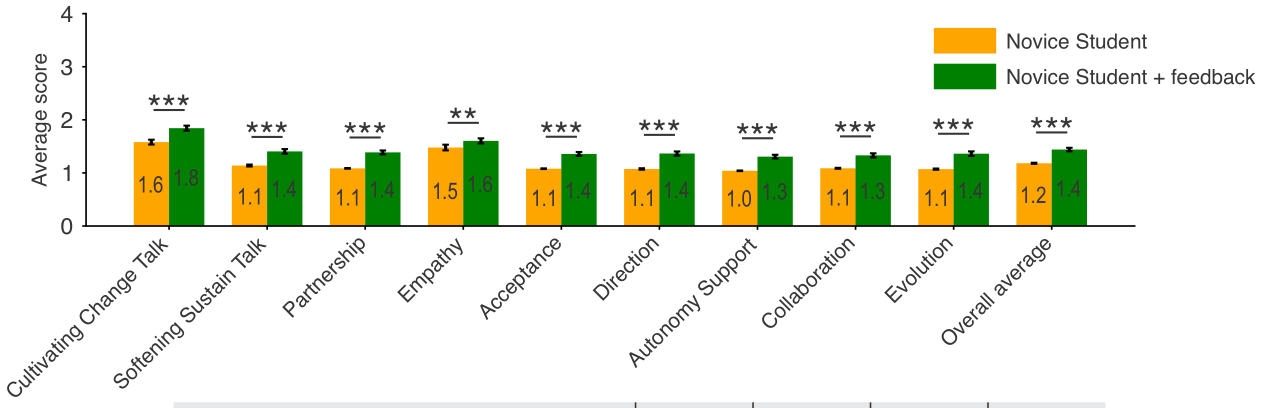
Impact of Evaluator Feedback on Novice Performance

In the prompt-augmentation sensitivity analysis, appending feedback recommendations to the novice Student-agent prompt produced modest score shifts in subsequent synthetic sessions (Figure 6). For MI, the overall average increased from 1.18 to 1.44, corresponding to an absolute mean change of +0.26 points (95% CI 0.19 - 0.33; $P<.001$; rank-biserial effect size=0.73). All 9 MI criteria showed significant increases after false discovery rate correction (Figure 6A). For CBT, the overall average increased from 0.83 to 1.01, corresponding to an absolute mean change of +0.18 points (95% CI 0.11 - 0.25; $P<.001$; rank-biserial effect size=0.55). A total of 9 of 11 CBT criteria showed significant increases after false discovery rate correction; Interpersonal Effectiveness and Focusing on Key Cognitions or Behaviors were not significant (Figure 6B). These findings indicate that feedback-text prompt augmentation shifted evaluator-assigned scores within the synthetic pipeline.

XSL•FO
RenderX

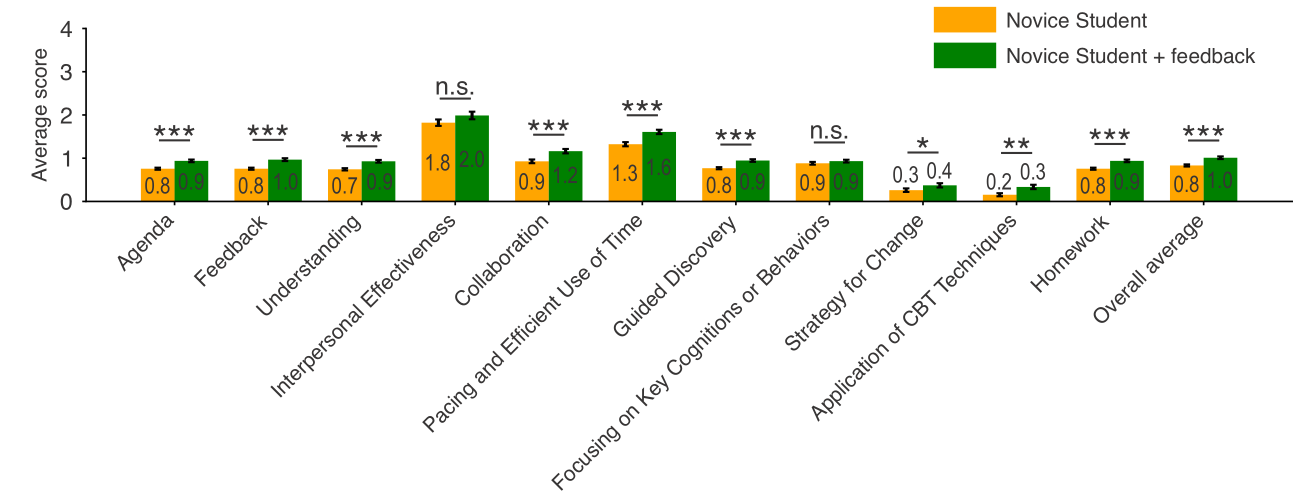
Figure 6. Prompt-augmentation sensitivity analysis using feedback text appended to the novice Student-agent prompt. (A) Motivational interviewing (MI) criterion-level scores before and after appending feedback recommendations to the novice Student-agent prompt across 133 Patient-agent profiles. The overall average increased from 1.18 to 1.44, corresponding to an absolute change of +0.26 points (P<.001) and a rank-biserial effect size of 0.73. All 9 MI criteria showed significant increases after false discovery rate correction. (B) Cognitive behavioral therapy (CBT) criterion-level scores before and after feedback-text prompt augmentation across 102 Patient-agent profiles. The overall average increased from 0.83 to 1.01, corresponding to an absolute change of +0.18 points (P<.001) and a rank-biserial effect size of 0.55. A total of 9 of 11 CBT criteria showed significant increases after false discovery rate correction. Asterisks denote Benjamini-Hochberg false discovery rate–adjusted criterion-level significance: *q<0.05, **q<0.01, ***q<0.001; n.s.=not significant after false discovery rate correction. Overall-average P values in the tables correspond to separate planned paired Wilcoxon signed-rank tests.

A



Overall average comparison	Novice mean	Feedback mean	*P* value	Effect size
Novice vs Novice + feedback for MI	1.18	1.44	<.001	0.73

B



Overall average comparison	Novice mean	Feedback mean	*P* value	Effect size
Novice vs Novice + feedback for CBT	0.83	1.01	<.001	0.55

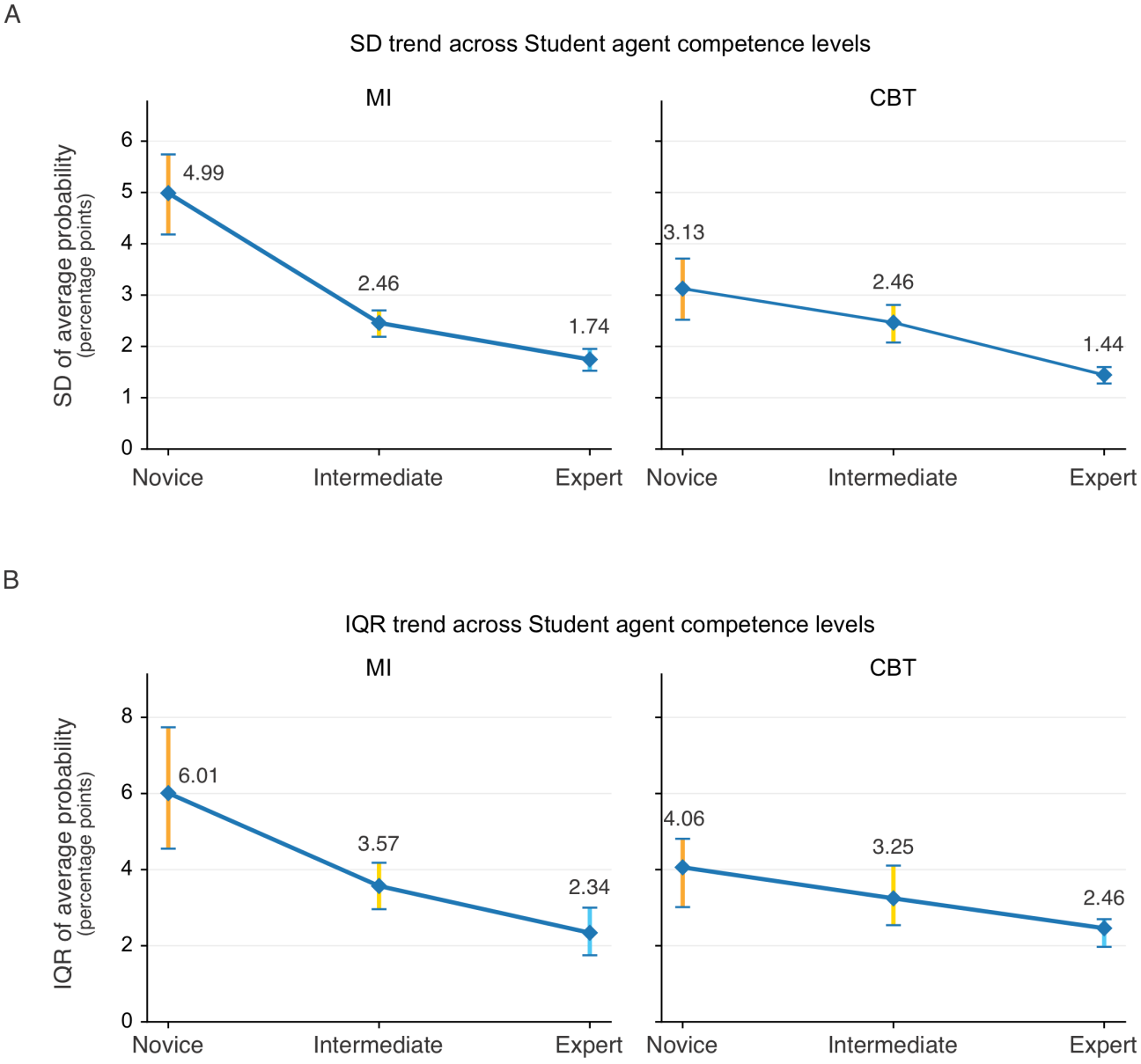
Narrative feedback outputs were criterion-referenced and score-linked. In representative MI outputs, low scores in Acceptance, Empathy, Partnership, Autonomy Support, and Evocation were accompanied by explanations citing judgmental language, lack of reflective listening, directive advice, limited collaboration, and failure to elicit change talk. Corresponding recommendations emphasized avoiding confrontation, using complex reflections, eliciting patient values, and supporting patient choice. In representative CBT outputs, weak Agenda, Conceptualization, Guided Discovery, and Homework scores were paired with recommendations for collaborative agenda setting, psychoeducation, Socratic questioning, and structured between-session practice. These examples illustrate structural alignment between criterion scores, explanations, and recommendations.

XSL•FO
RenderX

Exploratory Response-Consistency Analysis

Figure 7 shows the dispersion of Student-agent response log-probability-derived average token probability across Patient-agent profiles. Dispersion decreased across prompt-defined competence levels. For MI, the SD decreased from 4.99 percentage points (95% CI 4.19 - 5.77) for the novice profile to 2.46 (95% CI 2.19 - 2.70) for the intermediate profile to 1.74 (95% CI 1.53 - 1.95) for the expert profile; the corresponding IQR decreased from 6.01 (95% CI 4.50 - 7.76) to 3.57 (95% CI 2.96 - 4.22) to 2.34 (95% CI 1.75 - 3.00). For CBT, the SD decreased from 3.13 (95% CI 2.51 - 3.70) to 2.46 (95% CI 2.08 - 2.82) to 1.44 (95% CI 1.28 - 1.60), and the IQR decreased from 4.06 (95% CI 2.98 - 4.80) to 3.25 (95% CI 2.55 - 4.10) to 2.46 (95% CI 1.97 - 2.69). These findings suggest that higher-adherence Student-agent prompts produced more constrained and consistent responses across Patient-agent profiles. This exploratory analysis is an indirect proxy for generative consistency and does not directly measure hallucination, clinical realism, or safety.

Figure 7. Exploratory Student-agent log-probability dispersion across prompt-defined competence levels. For each simulated session, Student-agent token log probabilities were averaged across generated tokens and turns, then transformed into geometric mean token probability percentages. (A) SD of the session-level average probability across Patient-agent profiles for novice, intermediate, and expert Student-agent conditions in motivational interviewing (MI) and cognitive behavioral therapy (CBT). (B) IQR of the same measure. Error bars indicate bootstrap CIs. Lower dispersion indicates more consistent model response probabilities across Patient-agent profiles. Expert Student-agent responses showed lower dispersion than novice responses in both MI and CBT.



A prototype user interface was developed to demonstrate how the multiagent workflow could be presented to end users, including approach selection, text- or voice-based practice, and display of criterion-level scores with narrative feedback (Figure S1 in Multimedia Appendix 1). This interface was not used in the present evaluation analyses and should be interpreted only as an implementation demonstration. Overall, the results provide early-stage evidence that a rubric-guided Evaluator agent can score stylized synthetic MI and CBT transcripts along expected prompt-defined competence gradients, align with a preliminary

XSL•FO
RenderX

external comparison and a pragmatic human-rater benchmark, and show modest score shifts after feedback-text prompt augmentation.

Discussion

Principal Findings: Scalable and Interpretable Assessment of Psychotherapeutic Competence

Integration of LLMs into health professions education may transform how trainees acquire and practice complex communication skills at scale [8,18,19,21,25]. In this preclinical, cross-sectional, simulation-based evaluation, we assessed a rubric-guided, multiagent LLM framework for stylized MI and CBT training simulations. The Evaluator agent scored synthetic transcripts in the expected direction across prompt-defined novice, intermediate, and expert Student-agent profiles; showed preliminary alignment with coarse AnnoMI session-level quality labels; aligned with a pragmatic human-rater benchmark; and showed modest score shifts after feedback-text prompt augmentation. The exploratory log-probability dispersion analysis further suggested that higher-adherence Student-agent prompts produced more constrained and consistent responses across Patient-agent profiles. By coupling standardized Patient simulations with an Evaluator that maps behaviors to established instruments, the system provides formative and summative signals that are interpretable within recognized competency frameworks [35]. Because the agents operate over explicit rubrics and supply narrative justifications, the approach is amenable to audit and standard setting and offers a path to consistent assessment across sites and cohorts [36,37].

Interpretation and Comparison With Prior Work

The main innovation of this work is the role-separated architecture, in which Patient, Student, Evaluator, and Feedback agents serve distinct functions within one simulation workflow [25]. Prior LLM studies in health professions education have largely emphasized plausible simulated conversations or tutoring interactions [7,20]. This study extends that work by linking synthetic psychotherapy encounters to explicit scoring criteria, external comparison data, pragmatic human-rater agreement, and criterion-referenced feedback outputs [16,17]. The contribution is therefore not that the system reproduces real psychotherapy, but that it provides an auditable framework for testing whether an LLM-based Evaluator responds to controlled, rubric-relevant differences in synthetic transcripts.

The internal discrimination analysis is important because a scoring system intended for psychotherapy training should, at minimum, assign higher scores when transcripts contain more rubric-concordant therapist behaviors [38]. The graded pattern across novice, intermediate, and expert Student-agent profiles supports this internal scoring sensitivity. However, these Student-agent profiles were engineered competence archetypes and should not be interpreted as empirical representations of medical students, therapists in training, or expert clinicians. Future studies must test whether similar scoring patterns hold in human learners across defined training stages [8].

The human-rater comparison provides a complementary benchmark for interpreting the Evaluator's scores. Although the rater panel was not composed of calibrated fidelity coders, Evaluator scores aligned closely with the average human-rater consensus and showed similar patterns when compared with individual raters. This suggests that, within the constrained setting of stylized synthetic transcripts and study-specific scoring forms, the Evaluator's rubric-based judgments approximated the ratings assigned by independent human raters. This finding is consistent with prior work suggesting that LLM-based evaluators can approximate human ratings [39,40]. However, this agreement should not be interpreted as evidence of superiority over human raters or as a substitute for validation against trained fidelity coders.

Some criteria, particularly Empathy and Autonomy Support, require careful interpretation in LLM-agent simulations. In this study, Empathy was scored as observable language indicating understanding, validation, and reflective response to the patient's stated experience. Autonomy Support was scored as noncoercive, choice-supporting language consistent with MI principles. These are transcript-level behavioral ratings, not evidence that the LLM possesses affective empathy, clinical attunement, or the capacity to form a therapeutic relationship. Prior work on LLM empathy has similarly emphasized the distinction between responses perceived as empathic and genuine human empathy [41]. We therefore interpret these scores as evidence that the Evaluator detected rubric-concordant language patterns in synthetic transcripts, not as evidence of authentic empathic capacity.

The narrative feedback component also requires cautious interpretation. Feedback outputs were criterion-referenced and structurally aligned with scores, explanations, and recommendations. This structure is useful for formative simulation design because it makes the basis for feedback explicit and reviewable [20,42]. However, narrative feedback alignment was assessed qualitatively and illustratively in this study; it was not independently validated by blinded experts for accuracy, clinical appropriateness, safety, or educational utility [35,37]. Similarly, the prompt-augmentation analysis showed that appending feedback text shifted subsequent Student-agent scores, but it does not demonstrate feedback efficacy, human learning, or skill transfer [42].

Limitations and Future Directions

A central limitation is that the simulated transcripts were stylized outputs of prompt-engineered agents rather than interactions involving real patients or human trainees. We did not compare generated dialogues with real psychotherapy sessions, and we did not test whether Patient or Student agents reproduced the behavior of actual patients or trainees. This design supported controlled benchmarking but limited assessment of flexible adaptation to complex, comorbid, culturally nuanced, or diagnostically ambiguous presentations [43,44]. Similarly, the fixed 10-exchange session length standardized exposure across conditions but restricted observation of longer-form psychotherapy processes such as alliance development, rupture repair, homework review, and multisession change [45,46]. Accordingly, the findings should be interpreted as evidence that the Evaluator can score controlled synthetic examples along expected rubric-defined gradients, not as evidence of ecological

XSL•FO
RenderX

validity or readiness for summative assessment of real learners [8,47].

The external and human-rating benchmarks also have important limitations. Although AnnoMI contains expert utterance-level annotations [31], this study used only coarse high or low session-level quality labels derived from video metadata and source context; these labels are not equivalent to expert transcript-level MI fidelity ratings. In addition, because AnnoMI is publicly available and predates the Evaluator model's knowledge cutoff, training-data contamination cannot be excluded [48]. The human-rating analysis should likewise be interpreted as a pragmatic benchmark rather than a fidelity-coding reference standard. Although the rater panel provided useful comparison data, raters were not formally calibrated MITI, MISC, or CTRS fidelity coders. Future validation should therefore include trained and calibrated fidelity coders, predefined reliability thresholds, full fidelity-coding procedures, and prospectively collected transcripts from human trainees interacting with standardized or real patients.

The prompt-augmentation and feedback analyses should not be interpreted as evidence of educational efficacy. No human learner received feedback, practiced with the system, demonstrated skill retention, or transferred skills to a new clinical task. Instead, the postfeedback Student-agent prompt contained feedback recommendations appended to an otherwise novice prompt, meaning that observed score shifts may reflect LLM instruction-following or prompt sensitivity rather than learning [49,50]. Narrative feedback outputs were criterion-referenced and structurally aligned with scores, explanations, and recommendations, but they were not independently validated by blinded experts for accuracy, clinical appropriateness, safety, or educational utility. Future studies should test feedback prospectively in human trainees, using independent feedback delivery, delayed reassessment, blinded expert ratings, and evaluation on unseen standardized-patient or real-patient encounters [51-53].

Finally, several implementation and safety limitations remain. LLMs are susceptible to hallucination, misinterpretation, unsafe recommendations, bias, and calibration drift [54,55]. This study did not formally adjudicate or quantify failure events such as hallucinated clinical details, internally inconsistent Patient-agent responses, or clinically inappropriate recommendations. The exploratory log-probability dispersion analysis provides only an indirect proxy for generative consistency and does not directly measure clinical realism, therapeutic safety, or fidelity to patient complexity. Because the simulations were text-only, the system could not evaluate tone, pacing, prosody, facial affect, gestures, silence, turn timing, or other nonverbal and paralinguistic behaviors that contribute to alliance and clinical communication [7,56,57]. Future versions should incorporate structured safety monitoring, expert adjudication of failure modes, multimodal inputs such as audio or video, and validation by domain-specific clinical experts [56]. Model selection is also implementation-dependent: using different model families reduces direct same-model self-evaluation, but it does not eliminate shared pretraining exposure, model bias, or calibration drift [58,59]. These issues should be addressed before the system is used beyond low-stakes formative simulation or research settings.

Conclusions

This proof-of-concept study suggests that a rubric-guided multiagent LLM framework can score stylized synthetic MI and CBT transcripts along expected prompt-defined competence gradients and align with preliminary external and pragmatic human-rater benchmarks. The study is innovative in separating synthetic patient simulation, synthetic learner behavior, rubric-based scoring, and narrative feedback generation within a single auditable workflow. It differs from prior LLM simulation studies by moving beyond face-valid dialogue generation toward structured, criterion-linked evaluation. The findings support further development of scalable simulation tools for psychotherapy training research, fidelity-methods development, and low-stakes formative practice.

Acknowledgments

The authors thank Martin Ivanov for his contributions to the development of this work. The authors declare the use of generative AI in the research and writing process. According to the Generative AI Delegation Taxonomy (GAIDeT, 2025), the following tasks were delegated to generative AI tools under full human supervision: code optimization, proofreading, and editing. The generative AI tool used was ChatGPT-5.5. Responsibility for the final manuscript lies entirely with the authors. Generative AI tools are not listed as authors and do not bear responsibility for the final outcomes. All substantive scientific content, analyses, interpretations, citations, manuscript revisions, and reviewer responses were reviewed, edited, and approved by the authors, who take full responsibility for the integrity and accuracy of the final work.

Funding

This work was supported by scholarships or fellowships from Mitacs (Accelerate Fellowship), the Brain Canada Foundation (Rising Stars Award), and the Canadian Institutes of Health Research (CIHR; Canada Postdoctoral Research Award) to MAK.

Data Availability

The datasets generated and analyzed during the current study are available on GitHub [60]. The repository provides synthetic motivational interviewing patient profiles, student prompts, analysis code, and structured cognitive behavioral therapy patient-agent attributes reformatted from the *Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition* (*DSM-5*) Clinical Cases. Annotated motivational interviewing transcripts are not redistributed and can be accessed through the original dataset source.

XSL•FO
RenderX

Authors' Contributions

Concept and design: MAK, MM, VB
Acquisition, analysis, and interpretation of data: MAK, MM, RB, NTL, ZC
Drafting of the manuscript: MAK, MM
Critical review of the manuscript: MAK, MM, RB, NTL, ZC, OW, LB, YZ, ORZ, CM, DS, SK, AJG, RZ, PS, VB
Administrative, technical, and material support: MAK, VB
All authors contributed to and approved the final version of the manuscript

Conflicts of Interest

MAK has received funding from Sanofi. LB, OW, and YZ were supported by the Academic Medicine and Health Services Program. RZ has received funding from the NYU Langone Psychedelic Medicine Research Training Program (funded by MindMed) and the Canadian Institutes of Health Research. PS has received funding, honoraria, or consulting fees from Pfizer, Shoppers Drug Mart, Bhasin Consulting Fund Inc, the Patient-Centered Outcomes Research Institute, AbbVie, Bristol-Myers Squibb, Evidera Inc, Johnson & Johnson Group of Companies, Medcan Clinic, Inflexxion Inc, V-CC Systems Inc, MedPlan Communications, Kataka Medical Communications, Miller Medical Communications, Nvision Insight Group, and Sun Life Financial. VB has received research support from the Canadian Institutes of Health Research, the Brain and Behavior Foundation, the Ontario Ministry of Health Innovation Funds, the Royal College of Physicians and Surgeons of Canada, the Department of National Defence (Government of Canada), the New Frontiers in Research Fund, Associated Medical Services Inc Healthcare, the American Foundation for Suicide Prevention, Roche, Novartis, and Eisai. The funders had no role in the design, data collection, analysis, interpretation, or writing of this manuscript.

Multimedia Appendix 1
Supplementary figure and tables.
[DOCX File, 10224 KB - mededu_v12i1e92964_app1.docx]

Checklist 1
DECIDE-AI checklist.
[DOCX File, 2170 KB - mededu_v12i1e92964_app2.docx]

References

1. Nurse K, O'Shea M, Ling M, Castle N, Sheen J. The influence of deliberate practice on skill performance in therapeutic practice: a systematic review of early studies. Psychother Res 2025 Mar;35(3):353-367. [doi: 10.1080/10503307.2024.2308159] [Medline: 38295223]
2. Elendu C, Amaechi DC, Okatta AU, et al. The impact of simulation-based training in medical education: a review. Medicine (Baltimore) 2024 Jul 5;103(27):e38813. [doi: 10.1097/MD.0000000000038813] [Medline: 38968472]
3. Beal MD, Kinnear J, Anderson CR, Martin TD, Wamboldt R, Hooper L. The effectiveness of medical simulation in teaching medical students critical care medicine: a systematic review and meta-analysis. Simul Healthc 2017 Apr;12(2):104-116. [doi: 10.1097/SIH.0000000000000189] [Medline: 28704288]
4. Kaplonyi J, Bowles KA, Nestel D, et al. Understanding the impact of simulated patients on health care learners' communication skills: a systematic review. Med Educ 2017 Dec;51(12):1209-1219. [doi: 10.1111/medu.13387] [Medline: 28833360]
5. Joyner B, Young L. Teaching medical students using role play: twelve tips for successful role plays. Med Teach 2006 May;28(3):225-229. [doi: 10.1080/01421590600711252] [Medline: 16753719]
6. Perlman MR, Anderson T, Foley VK, Mimnaugh S, Safran JD. The impact of alliance-focused and facilitative interpersonal relationship training on therapist skills: an RCT of brief training. Psychother Res 2020 Sep;30(7):871-884. [doi: 10.1080/10503307.2020.1722862] [Medline: 32028859]
7. Fernández-Alcántara M, Escribano S, Juliá-Sanchis R, et al. Virtual simulation tools for communication skills training in health care professionals: literature review. JMIR Med Educ 2025 May 6;11:e63082. [doi: 10.2196/63082] [Medline: 40327882]
8. Zeng J, Qi W, Shen S, et al. Embracing the future of medical education with large language model-based virtual patients: scoping review. J Med Internet Res 2025 Nov 13;27:e79091. [doi: 10.2196/79091] [Medline: 41232097]
9. Scholich T, Barr M, Wiltsey Stirman S, Raj S. A comparison of responses from human therapists and large language model-based chatbots to assess therapeutic communication: mixed methods study. JMIR Ment Health 2025 May 21;12:e69709. [doi: 10.2196/69709] [Medline: 40397927]
10. Hettema J, Steele J, Miller WR. Motivational interviewing. Annu Rev Clin Psychol 2005;1:91-111. [doi: 10.1146/annurev.clinpsy.1.102803.143833] [Medline: 17716083]
11. Beck JS. Cognitive Behavior Therapy: Basics and Beyond, 3rd edition: Guilford Publications; 2020.

XSL•FO
RenderX

12. Moyers TB, Manuel JK, Ernst D. Motivational Interviewing Treatment Integrity Coding Manual 4.2.1. : University of New Mexico, Center on Alcoholism, Substance Abuse, and Addictions (CASAA); 2015 URL: https://casaa.unm.edu/assets/docs/miti4_21.pdf [accessed 2026-07-28]
13. Miller WR, Moyers TB, Ernst D, Amrhein P. Manual for the Motivational Interviewing Skill Code (MISC), Version 2.0. : Center on Alcoholism, Substance Abuse and Addictions (CASAA), The University of New Mexico; 2003 URL: https://digitalcommons.montclair.edu/cgi/viewcontent.cgi?article=1026&context=psychology-facpubs [accessed 2026-07-28]
14. Young J, Beck AT. Cognitive therapy scale rating manual. : University of Pennsylvania; 1980 URL: https://static1.squarespace.com/static/60f6fb2c51d9421daf31868c/t/6103d250c8ffd964aa4fe413/1627640400759/CTRS_Manual.pdf [accessed 2026-07-28]
15. Soma CS, Kuo PB, Mehta M, Srikumar V, Imel ZE, Atkins DC. Artificial intelligence to support human-provided mental health treatment. Annu Rev Clin Psychol 2026 May;22(1):505-531. [doi: 10.1146/annurev-clinpsy-061724-075336] [Medline: 41666031]
16. Pellemans M, Salmi S, Mérelle S, Janssen W, van der Mei R. Automated behavioral coding to enhance the effectiveness of motivational interviewing in a chat-based suicide prevention helpline: secondary analysis of a clinical trial. J Med Internet Res 2024 Aug 1;26:e53562. [doi: 10.2196/53562] [Medline: 39088244]
17. Lim K, Jung YC, Kim BH. Evaluating motivational interview quality using large language models and hidden Markov models. BMC Psychiatry 2025 Oct 1;25(1):908. [doi: 10.1186/s12888-025-07391-1] [Medline: 41034852]
18. Xu T, Weng H, Liu F, et al. Current status of ChatGPT use in medical education: potentials, challenges, and strategies. J Med Internet Res 2024 Aug 28;26:e57896. [doi: 10.2196/57896] [Medline: 39196640]
19. Abd-Alrazaq A, AlSaad R, Alhuwail D, et al. Large language models in medical education: opportunities, challenges, and future directions. JMIR Med Educ 2023 Jun 1;9:e48291. [doi: 10.2196/48291] [Medline: 37261894]
20. Holderried F, Stegemann-Philipps C, Herrmann-Werner A, et al. A language model-powered simulated patient with automated feedback for history taking: prospective study. JMIR Med Educ 2024 Aug 16;10:e59213. [doi: 10.2196/59213] [Medline: 39150749]
21. Vrdoljak J, Boban Z, Vilović M, Kumrić M, Božić J. A review of large language models in medical education, clinical decision support, and healthcare administration. Healthcare (Basel) 2025 Mar 10;13(6):603. [doi: 10.3390/healthcare13060603] [Medline: 40150453]
22. Barra FL, Rodella G, Costa A, et al. From prompt to platform: an agentic AI workflow for healthcare simulation scenario design. Adv Simul (Lond) 2025 May 16;10(1):29. [doi: 10.1186/s41077-025-00357-z] [Medline: 40380247]
23. Wei H, Qiu J, Yu H, Yuan W. MEDCO: medical education copilots based on a multi-agent framework. In: Del Bue A, Canton C, Pont-Tuset J, Tommasi T, editors. Computer Vision – ECCV 2024 Workshops Milan, Italy, September 29–October 4, 2024, Proceedings, Part VIII: Springer; 2024:119-135. [doi: 10.1007/978-3-031-91813-1_8]
24. Sapkota R, Roumeliotis KI, Karkee M. AI Agents vs. Agentic AI: a conceptual taxonomy, applications and challenges. Inf Fusion 2026 Feb;126:103599. [doi: 10.1016/j.inffus.2025.103599]
25. Yu H, Zhou J, Li L, et al. Simulated patient systems powered by large language model-based AI agents offer potential for transforming medical education. Commun Med (Lond) 2025 Dec 19;6(1):27. [doi: 10.1038/s43856-025-01283-x] [Medline: 41420084]
26. Yan L, Sha L, Zhao L, et al. Practical and ethical challenges of large language models in education: a systematic scoping review. Brit J Educational Tech 2024 Jan;55(1):90-112. [doi: 10.1111/bjet.13370]
27. Templin T, Fort S, Padmanabham P, et al. Framework for bias evaluation in large language models in healthcare settings. NPJ Digit Med 2025 Jul 7;8(1):414. [doi: 10.1038/s41746-025-01786-w] [Medline: 40624264]
28. Stade EC, Stirman SW, Ungar LH, et al. Large language models could change the future of behavioral healthcare: a proposal for responsible development and evaluation. Npj Ment Health Res 2024 Apr 2;3(1):12. [doi: 10.1038/s44184-024-00056-z] [Medline: 38609507]
29. Yosef S, Zisquit M, Cohen B, Klomek Brunstein A, Bar K, Friedman D. Assessing motivational interviewing sessions with AI-generated patient simulations. In: Ophir Y, Desmet B, Prud'hommeaux E, Zirikly A, Bedrick S, MacAvaney S, et al, editors. Proceedings of the 9th Workshop on Computational Linguistics and Clinical Psychology (CLPsych 2024): Association for Computational Linguistics; 2024:1-11. [doi: 10.18653/v1/2024.clpsych-1.1]
30. Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition: American Psychiatric Association; 2013. URL: https://psychiatryonline.org/doi/book/10.1176/appi.books.9780890425596 [accessed 2026-07-31]
31. Wu Z, Balloccu S, Kumar V, et al. Anno-MI: a dataset of expert-annotated counselling dialogues. In: ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP): IEEE; 2022:6177-6181. [doi: 10.1109/ICASSP43922.2022.9746035]
32. Vasey B, Nagendran M, Campbell B, et al. Reporting guideline for the early-stage clinical evaluation of decision support systems driven by artificial intelligence: DECIDE-AI. Nat Med 2022 May;28(5):924-933. [doi: 10.1038/s41591-022-01772-9] [Medline: 35585198]
33. Canadian Institutes of Health Research; Natural Sciences and Engineering Research Council of Canada; Social Sciences and Humanities Research Council of Canada (Tri-Council agencies). Tri-council policy statement: ethical conduct for

XSL•FO
RenderX

research involving humans. : Interagency Secretariat on Research Ethics; 2022 URL: https://ethics.gc.ca/eng/documents/tcps2-2022-en.pdf [accessed 2026-07-28]
34. Wu Z, Balloccu S, Kumar V, Helaoui R, Reforgiato Recupero D, Riboni D. Creation, analysis and evaluation of AnnoMI, a dataset of expert-annotated counselling dialogues. Future Internet 2023;15(3):110. [doi: 10.3390/fi15030110]
35. Lee GB, Chiu AM. Assessment and feedback methods in competency-based medical education. Ann Allergy Asthma Immunol 2022 Mar;128(3):256-262. [doi: 10.1016/j.anai.2021.12.010] [Medline: 34929390]
36. Gu J, Jiang X, Shi Z, et al. A survey on LLM-as-a-judge. Innovation (Camb) 2026 Jun 1;7(6):101253. [doi: 10.1016/j.xinn.2025.101253] [Medline: 42254963]
37. Tam TYC, Sivarajkumar S, Kapoor S, et al. A framework for human evaluation of large language models in healthcare derived from literature review. NPJ Digit Med 2024 Sep 28;7(1):258. [doi: 10.1038/s41746-024-01258-7] [Medline: 39333376]
38. Cook DA, Hatala R. Validation of educational assessments: a primer for simulation and beyond. Adv Simul (Lond) 2016;1:31. [doi: 10.1186/s41077-016-0033-y] [Medline: 29450000]
39. Xu J, Wei T, Hou B, et al. MentalChat16K: a benchmark dataset for conversational mental health assistance. KDD 2025 Aug;2025:5367-5378. [doi: 10.1145/3711896.3737393] [Medline: 41098434]
40. Zheng L, Chiang WL, Sheng Y, et al. Judging LLM-as-a-judge with MT-bench and chatbot arena. In: Oh A, Naumann T, Globerson A, Saenko K, Hardt M, Levine S, editors. NIPS '23: Proceedings of the 37th International Conference on Neural Information Processing Systems: Curran Associates Inc; 2023:46595-46623 URL: https://dl.acm.org/doi/10.5555/3666122.3668142 [accessed 2026-07-28]
41. Sorin V, Brin D, Barash Y, et al. Large language models and empathy: systematic review. J Med Internet Res 2024 Dec 11;26:e52597. [doi: 10.2196/52597] [Medline: 39661968]
42. Cook DA, Hatala R, Brydges R, et al. Technology-enhanced simulation for health professions education: a systematic review and meta-analysis. JAMA 2011 Sep 7;306(9):978-988. [doi: 10.1001/jama.2011.1234] [Medline: 21900138]
43. Owen J, Hilsenroth MJ. Treatment adherence: the importance of therapist flexibility in relation to therapy outcomes. J Couns Psychol 2014 Apr;61(2):280-288. [doi: 10.1037/a0035753] [Medline: 24635591]
44. McHugh RK, Murray HW, Barlow DH. Balancing fidelity and adaptation in the dissemination of empirically-supported treatments: the promise of transdiagnostic interventions. Behav Res Ther 2009 Nov;47(11):946-953. [doi: 10.1016/j.brat.2009.07.005] [Medline: 19643395]
45. Flückiger C, Del Re AC, Wampold BE, Horvath AO. The alliance in adult psychotherapy: a meta-analytic synthesis. Psychotherapy (Chic) 2018 Dec;55(4):316-340. [doi: 10.1037/pst0000172] [Medline: 29792475]
46. Eubanks CF, Muran JC, Safran JD. Alliance rupture repair: a meta-analysis. Psychotherapy (Chic) 2018 Dec;55(4):508-519. [doi: 10.1037/pst0000185] [Medline: 30335462]
47. Elhilali A, Ngo ASH, Reichenpfader D, Denecke K. Large language model-based patient simulation to foster communication skills in health care professionals: user-centered development and usability study. JMIR Med Educ 2025 Dec 12;11:e81271. [doi: 10.2196/81271] [Medline: 41385781]
48. Deng C, Zhao Y, Heng Y, et al. Unveiling the spectrum of data contamination in language model: a survey from detection to remediation. In: Ku LW, Martins A, Srikumar V, editors. Findings of the Association for Computational Linguistics: ACL 2024: Association for Computational Linguistics; 2024:16078-16092. [doi: 10.18653/v1/2024.findings-acl.951]
49. Guo Z, Lai A, Thygesen JH, Farrington J, Keen T, Li K. Large language models for mental health applications: systematic review. JMIR Ment Health 2024 Oct 18;11:e57400. [doi: 10.2196/57400] [Medline: 39423368]
50. Hua Y, Na H, Li Z, et al. A scoping review of large language models for generative tasks in mental health care. NPJ Digit Med 2025 Apr 30;8(1):230. [doi: 10.1038/s41746-025-01611-4] [Medline: 40307331]
51. Bjaastad JF, Lillevoll K, Hoffart A, et al. The effect of behavioral rehearsal in the training of clinical psychology students in cognitive therapy for social anxiety disorder: a randomized controlled trial. Behav Res Ther 2025 Oct;193:104849. [doi: 10.1016/j.brat.2025.104849] [Medline: 40907373]
52. Miller WR, Yahne CE, Moyers TB, Martinez J, Pirritano M. A randomized trial of methods to help clinicians learn motivational interviewing. J Consult Clin Psychol 2004 Dec;72(6):1050-1062. [doi: 10.1037/0022-006X.72.6.1050] [Medline: 15612851]
53. Imel ZE, Baldwin SA, Baer JS, et al. Evaluating therapist adherence in motivational interviewing by comparing performance with standardized and real patients. J Consult Clin Psychol 2014 Jun;82(3):472-481. [doi: 10.1037/a0036158] [Medline: 24588405]
54. Lee P, Bubeck S, Petro J. Benefits, limits, and risks of GPT-4 as an AI chatbot for medicine. N Engl J Med 2023 Mar 30;388(13):1233-1239. [doi: 10.1056/NEJMsr2214184] [Medline: 36988602]
55. Chelli M, Descamps J, Lavoué V, et al. Hallucination rates and reference accuracy of ChatGPT and Bard for systematic reviews: comparative analysis. J Med Internet Res 2024 May 22;26:e53164. [doi: 10.2196/53164] [Medline: 38776130]
56. Tu T, Schaekermann M, Palepu A, et al. Towards conversational diagnostic artificial intelligence. Nature 2025 Jun;642(8067):442-450. [doi: 10.1038/s41586-025-08866-7] [Medline: 40205050]

XSL•FO
RenderX

57. Del Giacco L, Anguera MT, Salcuni S. The action of verbal and non-verbal communication in the therapeutic alliance construction: a mixed methods approach to assess the initial interactions with depressed patients. Front Psychol 2020;11:234. [doi: 10.3389/fpsyg.2020.00234] [Medline: 32153459]
58. Spiliopoulou E, et al. Play favorites: a statistical method to measure self-bias in LLM-as-a-judge. arXiv. Preprint posted online on Aug 8, 2025. [doi: 10.48550/arXiv.2508.06709]
59. Xu C, Guan S, Greene D, Kechadi MT. Benchmark data contamination of large language models: a survey. arXiv. Preprint posted online on Jun 6, 2024. [doi: 10.48550/arXiv.2406.04244]
60. amin-kamaleddin/MI-and-CBT-agent. GitHub. URL: https://github.com/amin-kamaleddin/MI-and-CBT-Agent [accessed 2026-07-28]

Abbreviations:

AnnoMI: annotated motivational interviewing
CBT: cognitive behavioral therapy
CTRS: Cognitive Therapy Rating Scale
DECIDE-AI: Developmental and Exploratory Clinical Investigations of Decision Support Systems Driven by AI
***DSM-5*:** *Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition*
FDR: false discovery rate
ICC: intraclass correlation coefficient
LLM: large language model
MAE: mean absolute error
MI: motivational interviewing
MISC: Motivational Interviewing Skill Code
MITI: Motivational Interviewing Treatment Integrity
RMSE: root mean square error

Edited by S Brini; submitted 06.Feb.2026; peer-reviewed by D Chartash, R Gaus; revised version received 09.Jul.2026; accepted 20.Jul.2026; published 18.Aug.2026.

Please cite as:
Kamaleddin MA, Mirjalili M, Barzegar R, Trung Le N, Cote Z, Winkler O, Burback L, Zhang Y, Zaiane OR, Monson C, Sharma D, Krishnan S, Greenshaw AJ, Zeifman R, Selby P, Bhat V
Multiagent Large Language Model Framework for Psychotherapy Fidelity Assessment in Motivational Interviewing and Cognitive Behavioral Therapy Training: Cross-Sectional, Simulation-Based Evaluation Study
JMIR Med Educ 2026;12:e92964
URL: https://mededu.jmir.org/2026/1/e92964
doi:10.2196/92964

© Mohammad Amin Kamaleddin, Mina Mirjalili, Reza Barzegar, Nghia Trung Le, Zachary Cote, Olga Winkler, Lisa Burback, Yanbo Zhang, Osmar R Zaiane, Candice Monson, Divya Sharma, Sri Krishnan, Andrew J Greenshaw, Richard Zeifman, Peter Selby, Venkat Bhat. Originally published in JMIR Medical Education (https://mededu.jmir.org), 18.Aug.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Education, is properly cited. The complete bibliographic information, a link to the original publication on https://mededu.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Medical Students' Attitudes, Perceptions, and Self-Reported Familiarity With AI in Health Care: Systematic Review and Meta-Analysis

Wei Zhang[1*], MD; Jiaxue Han[1*], MD; Xin Han[2*], MD, PhD; Hang Xu[1*], MD, PhD; Langkun Wang[1], MD; Tianhai Lin[1], MD, PhD; Ping Tan[1], MD, PhD; Peng Zhang[1], MD, PhD; Xiaonan Zheng[1], MD, PhD

[1]Department of Urology, West China Hospital, Sichuan University, No. 37 Guoxue Alley, Wuhou District, Chengdu, Sichuan Province, China

[2]Mental Health Center, West China Hospital, Sichuan University, Chengdu, Sichuan Province, China

[*]these authors contributed equally

Corresponding Author:
Xiaonan Zheng, MD, PhD
Department of Urology, West China Hospital, Sichuan University, No. 37 Guoxue Alley, Wuhou District, Chengdu, Sichuan Province, China

Abstract

Background: AI is increasingly encountered in clinical care and medical education, but medical students' attitudes, perceptions, and self-reported familiarity have been assessed using heterogeneous survey instruments, AI referents, and response scales. Prior reviews often combined mixed health profession populations or summarized central estimates without fully showing variation across settings.

Objective: This study aimed to synthesize quantitative evidence on medical students' AI-related attitudes, perceptions, and self-reported familiarity while examining construct harmonization, participant independence, heterogeneity, prediction intervals, risk of bias, and certainty of evidence.

Methods: We searched PubMed (MEDLINE), Embase, Web of Science, Scopus, PsycINFO, and Cochrane CENTRAL from inception to April 1, 2026; supplementary searches are described in the appendices. Eligible studies enrolled students in MD, MBBS, MBChB, or DO-equivalent medical programs, or reported separable medical student data from mixed samples. Proportion outcomes were harmonized into 9 domains and synthesized using random-effects meta-analysis with Freeman-Tukey double-arcsine transformation, Hartung-Knapp-Sidik-Jonkman–adjusted CIs, and prediction intervals. Subgroup analyses and meta-regressions were exploratory because of multiple testing, ecological confounding, and construct heterogeneity. Risk of bias and certainty were assessed using the Joanna Briggs Institute analytical cross-sectional checklist and the GRADE (Grading of Recommendations, Assessment, Development and Evaluation) framework, respectively.

Results: Ninety-six cross-sectional studies from 37 countries (>45,000 medical students) were included. Summary estimates suggested favorable attitudes but wide between-setting dispersion. Positive attitude toward AI was 76.9% (95% CI 72.2% - 81.4%; prediction interval 42.2% - 98.3%; I^2=98.3%; 44 studies; N=20,806), perceived career benefit was 78.4% (95% CI 69.5% - 86.2%; prediction interval 45.3% - 98.3%; I^2=98.0%; 16 studies; N=9799), and support for curricular integration was 76.6% (95% CI 71.8% - 81.1%; prediction interval 47.8% - 96.1%; I^2=97.2%; 38 studies; N=16,308). Concern about physician replacement was 39.9% (95% CI 33.6% - 46.5%; prediction interval 6.6% - 80.1%; I^2=98.8%; 32 studies; N=16,642), willingness to learn about or adopt AI was 71.5% (95% CI 64.8% - 77.8%; prediction interval 37.9% - 95.4%; I^2=97.7%; 22 studies; N=9199), and ethical concerns were endorsed by 62.8% (95% CI 53.9% - 71.3%; prediction interval 21.9% - 95.0%; I^2=98.8%; 28 studies; N=14,571). Self-reported familiarity or knowledge was 63.3% (95% CI 55.9% - 70.3%; prediction interval 7.8% - 100.0%; I^2=99.5%; 52 studies; N=27,817), and trust in AI-assisted decisions was 50.6% (95% CI 28.5% - 72.6%; prediction interval 7.3% - 93.3%; I^2=98.1%; 8 studies; N=3007). All domains had very low certainty because of cross-sectional self-report designs, frequent use of nonvalidated or adapted instruments, wide prediction intervals, and small study effects in several domains.

Conclusions: Medical students' AI-related attitudes and curricular interest appear broadly favorable, but these estimates should not be interpreted as stable global prevalences. This review adds value by restricting the population to medical students, transparently harmonizing nonequivalent constructs, auditing mixed populations and participant independence, and reporting prediction intervals and certainty. Given very low certainty, the findings support locally adapted, exploratory AI-literacy planning and standardized measurement in future studies rather than strong claims about curriculum effectiveness.

Trial Registration: PROSPERO CRD420251120543; https://www.crd.york.ac.uk/prospero

International Registered Report Identifier (IRRID): RR2-89411

XSL•FO
RenderX

(JMIR Med Educ 2026;12:e89411) doi:10.2196/89411

KEYWORDS

artificial intelligence; medical students; attitudes; perceptions; AI literacy; meta-analysis; systematic review; medical education

Introduction

AI is increasingly encountered across clinical care, including image interpretation, predictive risk modeling, and clinical decision support [1,2]. Generative AI has also become salient in medical education, where the performance of large language models (LLMs) on medical examinations has prompted consideration of AI-assisted learning [3]. The educational relevance of AI extends beyond teaching how algorithms are built: medical education must prepare learners to understand capabilities and limitations, critically appraise outputs, and preserve professional judgment and oversight when AI-enabled tools are used [4-6]. As AI systems are increasingly incorporated into information-gathering and decision support workflows, future clinicians will need to judge when an AI output is useful, when its limitations matter, and when human review must contextualize or override the output [1,2,6]. These competencies matter before graduation because medical students are preparing for environments in which AI-mediated information may increasingly shape how evidence is accessed, interpreted, and applied [1,2,6].

Understanding how medical students perceive AI is important for needs assessment and curriculum planning [7]. Learner acceptance and perceived usefulness can shape willingness to engage with a technology, as described in the Technology Acceptance Model [8] and the Unified Theory of Acceptance and Use of Technology [9]. Yet, perception-based measures are not objective performance assessments [7]. Favorable attitudes and perceived usefulness may support engagement with AI-enabled tools [8,9], whereas concerns about professional displacement may temper medical students' attitudes toward specific AI applications [10]. Likewise, self-reported familiarity is a perception-based measure and does not by itself establish applied critical appraisal or preserved independent reasoning [7,11]. Attitude and familiarity data should therefore be treated as signals for educational needs assessment rather than as proxies for AI literacy [7,11]. The rapid diffusion of generative AI systems such as ChatGPT further sharpens this distinction because uncritical use may encourage cognitive dependence and weaken independent clinical reasoning [11].

Several prior reviews have examined AI in health professions education but not specifically among medical students enrolled in medical education programs (MD, MBBS, MBChB, or DO) [12-14]. Mousavi Baigi et al [12] reviewed health care students' attitudes, knowledge, and skills related to AI without meta-analytic pooling. Shishehgar et al [13] synthesized the knowledge, perceptions, and experiences of health students and academics regarding AI in health education and practice. Amiri et al [14] conducted a systematic review and meta-analysis of medical, dental, and nursing students' attitudes and knowledge toward AI. Although these reviews provide useful background, their inclusion of mixed health professions populations limits the specificity of inference for medical students [12-14]. In addition, the quantitative synthesis by Amiri et al did not apply the Hartung-Knapp-Sidik-Jonkman (HKSJ) correction recommended for random-effects CIs [15] or report prediction intervals, which are important for showing the likely range of effects across settings rather than only the pooled average [16].

Several limitations of the existing literature make a medical-student-specific synthesis necessary [12-14,16]. Prior reviews have synthesized populations extending beyond medical students to other health professions students and, in some cases, academics, and have grouped outcomes under broad review-level categories such as attitudes, knowledge, skills, perceptions, and experiences [12-14]. Measurement approaches and AI referents also vary across the literature: a validated readiness instrument such as Medical AI Readiness Scale for Medical Students (MAIRS-MS) [7] coexists with application-specific medical student perception surveys focused on radiology AI [10] and a rapidly changing generative AI context [11]. These differences in population, construct definition, and AI referent can blur conceptually distinct outcomes and make it difficult to determine whether between-study variation reflects who was surveyed, what was asked, or which technology was being considered [7,10-14,16]. In addition, narrative summaries or conventional random-effects CIs alone do not show the range of true values that may be expected across comparable settings [16]. Transparent outcome harmonization, explicit handling of mixed populations, and prediction interval–based interpretation are therefore important when pooled proportions are used to inform medical education decisions [12-14,16].

The present review was therefore designed to synthesize quantitative evidence on medical students' AI-related attitudes, perceptions, self-reported familiarity, prior use, ethical concerns, willingness to learn, and trust; document how heterogeneous survey outcomes were harmonized into operational domains; report CIs and prediction intervals for summary estimates; assess risk of bias and certainty of evidence; examine participant independence and mixed population disaggregation; evaluate subgroup patterns and feasibility-screened meta-regression; and explore the implications of these findings for AI literacy curriculum development.

Methods

Ethical Considerations

This review synthesized data from previously published studies only and involved no direct participant contact, no collection of identifiable personal data, and no experimental intervention. Under the policies of the institutional review board of West China Hospital, Sichuan University, systematic reviews are exempt from ethical review. Informed consent was therefore not applicable.

XSL•FO
RenderX

Study Registration and Reporting

The protocol was prospectively registered in PROSPERO (CRD420251120543). The review was conducted with reference to the Cochrane Handbook for Systematic Reviews of Interventions [17] and reported in accordance with PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) 2020 [18], PRISMA 2020 for Abstracts, and the PRISMA-S (Preferred Reporting Items for Systematic Reviews and Meta-Analyses literature search extension) reporting [19]. Completed reporting checklists are provided in Checklist 1.

Eligibility Criteria

Eligible participants were students enrolled in medical education programs (MD, MBBS, MBChB, or DO-equivalent programs), aligned with World Federation for Medical Education standards [20], from entry into training through internship. Studies with mixed populations were included only when medical student data were reported separately or could be extracted without including nonmedical participants. Mixed population records were reviewed in dedicated supplementary tables; any study without separable medical student data was excluded.

Information Sources and Search Strategy

We searched PubMed (MEDLINE), Embase, Web of Science, Scopus, PsycINFO, and Cochrane CENTRAL from inception to April 1, 2026. Search strategies used 3 concept blocks: medical student population terms, AI-related exposure terms, and survey terms related to attitudes, perceptions, and familiarity. For each database, the search platform, date of search, coverage, complete line-by-line query, limits, and number of records retrieved were documented in Multimedia Appendix 1 in accordance with PRISMA-S [19]. Supplementary searching included Google Scholar screening of the first 200 relevance-sorted records and backward and forward citation searching of included studies and prior reviews. Where the search records available to the review team did not contain a raw export, time-stamped log, or other procedural detail for a supplementary search (such as the exact search date, query string, sort order, or screener identity), that detail is marked as not recorded rather than reconstructed; checklist items that did not apply to the search methods are marked as not applicable in the completed PRISMA-S checklist (Checklist 1).

Selection Process and Data Collection

Records were imported into EndNote 21 (Clarivate) and deduplicated using automated matching by DOI, title, author, year, and journal, followed by manual verification [21] (Table S1 in Multimedia Appendix 2). The PRISMA 2020 study selection flow diagram is shown in the Results section, and reports not retrieved, eligibility-exclusion reasons, and the exclusion-impact comparison are documented in Tables S2 and S3 in Multimedia Appendix 2. During eligibility adjudication, mixed population screening records and full separability decisions were documented in Tables S4 and S5 in Multimedia Appendix 2. Two reviewers (WZ and JH) independently screened titles and abstracts, then full texts, with disagreements resolved by a third reviewer (XZ). Two reviewers independently extracted data using a piloted standardized form. Extracted variables included bibliographic details, study setting, population characteristics, instrument type and validation status, type of AI evaluated, and outcome data for 9 proportion domains, 5 MAIRS-MS outcomes (the total score and 4 subscales), and other noncomparable continuous measures; study-level characteristics are summarized in Table S6 in Multimedia Appendix 2.

Data Harmonization

Before synthesis, extracted survey items were mapped to the 9 proportion domains using an a priori operational framework. For every study-domain contribution, the harmonization record captured the original item wording or closest reported paraphrase, response scale, scale points, dichotomization rule, neutral-response handling, event count, denominator, and rationale for domain assignment. When papers enrolled mixed populations, only medical student–specific data were retained; if medical student event counts or denominators could not be separated from nonmedical groups, that domain was coded as nonextractable. These decisions are reported in the mixed population documentation and outcome harmonization tables.

Outcome Domains

Nine proportion domains were prespecified on the basis of the Technology Acceptance Model [8], the Theory of Planned Behavior [22], and the AI Literacy framework [23]: (P1) positive attitude toward AI, (P2) AI perceived as beneficial to medical career, (P3) support for AI curricular integration, (P4) concern about physician replacement, (P5) self-reported familiarity or knowledge, (P6) prior AI tool use, (P7) willingness to learn about or adopt AI, (P8) ethical concerns, and (P9) trust in AI-assisted decisions. These domains were treated as operational approximations rather than standardized constructs. For Likert-type items, endorsement categories such as agree or strongly agree, yes, aware, familiar, or equivalent positive responses were extracted as events when reported; neutral responses remained in the denominator unless a source study explicitly excluded them. The full outcome harmonization table and harmonization verification table report item wording, denominator, cutoff notes, assigned domain, and assignment rationale.

Risk-of-Bias Assessment

Two reviewers independently assessed risk of bias using the Joanna Briggs Institute Critical Appraisal Checklist for Analytical Cross-Sectional Studies [24]. This tool was selected because the included surveys were analytical cross-sectional studies that frequently combined descriptive estimation of attitudes or familiarity with comparisons by training stage, exposure, region, or other participant characteristics. Although the review synthesized proportional self-report outcomes rather than causal associations, the Joanna Briggs Institute (JBI) domains remained relevant to sampling, participant description, exposure and outcome measurement, confounding, and statistical reporting. We also considered the JBI Critical Appraisal Checklist for Prevalence Studies; because most included surveys reported analytical comparisons across training stage, exposure, or region in addition to descriptive proportions, the analytical cross-sectional checklist was judged to capture the relevant

XSL•FO
RenderX

domains, including confounding and exposure and outcome measurement, more completely and was applied to all studies for consistency. Studies scoring 6 - 8 were classified as low risk of bias, 4 - 5 as moderate risk, and 0 - 3 as high risk. No study met the high-risk scoring threshold, but common limitations including convenience sampling, nonvalidated instruments, unclear response-rate denominators, and self-selection were incorporated into the certainty assessment and interpretation. The item-level JBI checklist is provided in Table S7 in Multimedia Appendix 2.

Synthesis Methods

For proportion outcomes, study-specific estimates were pooled after Freeman-Tukey double-arcsine transformation [25,26] using random-effects models [27] with HKSJ–adjusted CIs [15]. Random-effects models were prespecified for all meta-analyses because true effect sizes were expected to vary across populations, settings, instruments, and item wording [28]. Prediction intervals were calculated and reported for the primary proportional meta-analyses whenever 3 or more studies contributed [16]. For these analyses, CIs describe uncertainty around the pooled summary estimate, whereas prediction intervals describe the range of true proportions that may be expected in a new comparable setting. The width of the prediction intervals from the primary proportional meta-analyses was treated as the empirical rationale for proceeding with the prespecified subgroup analyses and feasibility-screened meta-regressions, which explored potential sources of between-setting dispersion. All analyses were performed with R (version 4.5.3; R Foundation for Statistical Computing) using the meta (version 8.5-0) and metafor (version 5.0-1) packages, both obtained from the Comprehensive R Archive Network.

Certainty of evidence was assessed for each proportional outcome using the GRADE (Grading of Recommendations, Assessment, Development and Evaluation) framework for bodies of observational evidence and absolute event-rate estimates [29-31]. The pooled proportion was the absolute estimate of interest. Risk of bias considered JBI ratings and recurring limitations in sampling, response-rate reporting, and instrument validation. Prediction intervals were used as the primary basis for interpreting the practical magnitude of between-setting heterogeneity because they describe the range of true proportions that may be expected in a new comparable setting [16]. I^2 and τ^2 were reported for completeness but were not used to label heterogeneity as low, moderate, high, or extreme. Imprecision considered CI width, prediction interval width, and the number of contributing studies and participants. Indirectness considered whether survey items, AI technologies, populations, and educational contexts matched the target construct, and publication bias considered funnel plot asymmetry and small study effect tests where feasible. Because all evidence came from cross-sectional self-report surveys and prediction intervals were very wide across the proportional domains, certainty ratings were interpreted conservatively.

For continuous outcomes reported on the same scale, pooled means were synthesized using random-effects meta-analysis with the same HKSJ framework. These analyses were treated as secondary descriptive syntheses; prediction intervals were not reported or interpreted for the continuous outcomes. Continuous outcomes reported on noncomparable scales were summarized descriptively.

Prespecified subgroup analyses examined publication period, World Bank income level, World Health Organization (WHO) region, type of AI evaluated, and risk of bias. AI type was assigned using a deterministic hierarchy: studies mentioning ChatGPT, LLMs, chatbots, or generative AI were classified as ChatGPT or LLM or generative AI; otherwise, studies focused on radiology, ophthalmology, specialty- or domain-specific AI, or clinical decision tools were classified as domain-specific AI. All remaining studies were classified as broad or general AI. For studies mentioning more than 1 AI type, this hierarchy supplied 1 study-level assignment and studies were not split or weighted across AI-type categories. World Bank income level and WHO region were assigned according to the country or countries in which the study population was surveyed. Single-country studies used that country's region and income band; multicountry studies were coded dimension-wise, with studies spanning multiple WHO regions classified as Multiple regions or Other, and studies spanning multiple-income bands or global or unspecified country mixes classified as mixed income, without weighting by participant distribution. Where $k \geq 10$, univariable meta-regression was performed using mixed-effects models to examine whether each moderator, considered separately, was associated with between-study variation. Because multiple subgroup analyses can inflate type I error and study-level moderators are vulnerable to ecological confounding, we added feasibility-screened meta-regression and bubble plots only when k was at least 10, at least 2 moderator levels were present, and no level contained fewer than 3 studies. All subgroup and meta-regression findings were interpreted as exploratory and hypothesis-generating.

Participant Independence Verification

To assess potential participant duplication, we conducted a 4-dimensional participant independence assessment across all included studies, examining author and ethics approval concordance, overlap between multinational and local surveys, spatiotemporal institutional concordance, and survey platform congruence (Table S8 in Multimedia Appendix 2). Studies with possible overlap were retained in the primary analysis only when duplication could not be confirmed and were prespecified for sensitivity analysis. The finalized proportion and continuous extraction datasets are provided in Tables S9 and S10 in Multimedia Appendix 2, and the outcome harmonization and source verification records are provided in Tables S11 and S12 in Multimedia Appendix 2.

Results

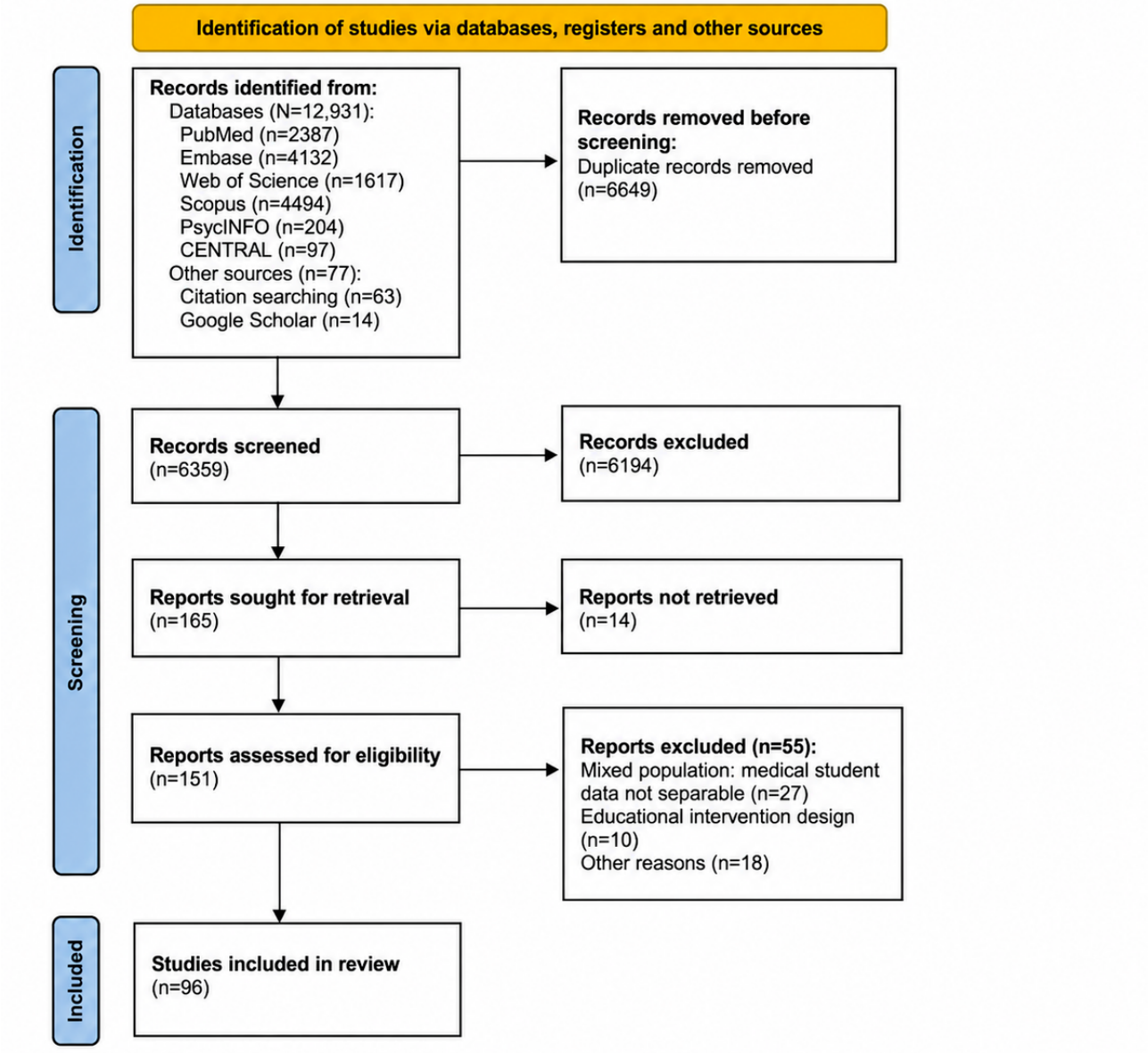
Study Selection

Database searching yielded 12,931 records and supplementary searching contributed 77 more, for a total of 13,008. After removing 6649 duplicates, 6359 unique records were screened at the title and abstract level; 6194 were excluded. Of 165 reports sought for retrieval, 14 could not be retrieved and 151 were assessed at full text; 55 of these were excluded, most often because they enrolled mixed populations from which medical

XSL•FO
RenderX

student data could not be separated (n=27) or used an educational intervention design (n=10), with 18 excluded for other prespecified reasons (Figure 1). Two of these exclusions (Sami et al [32]; Farooq and Usmani [33]) were studies that initially appeared eligible but were removed after detailed examination because their reported proportions combined medical with dental and allied health students and medical student–specific counts could not be recovered. A total of 96 studies were included [10,34-128].

Figure 1. PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) 2020 flow diagram of study identification, screening, and inclusion, drawn with the official PRISMA 2020 template. Records identified through supplementary citation searching and Google Scholar were exported to the reference manager and deduplicated together with database records; all records therefore flow through a single screening pathway.



Study Characteristics

The 96 cross-sectional surveys were published between 2019 and 2026; 82 (85%) appeared from 2023 onward (Table 1). They originated from 37 countries spanning 6 WHO regions, led by the Eastern Mediterranean (38/96, 40%), Europe (19/96, 20%), and the Western Pacific (12/96, 13%). Forty-three studies (45%) came from high-income countries, 28 (29%) from low- or lower-middle-income countries, 20 (21%) from upper-middle-income countries, and 5 (5%) were multicountry mixed-income studies. Sample sizes ranged from 20 to 4492 medical students. Fifty-nine studies (62%) assessed attitudes toward broad or general AI, 27 (28%) focused on ChatGPT or other LLMs, and 10 (10%) evaluated domain-specific applications such as radiology AI. Thirteen studies used the validated MAIRS-MS [7]; the remainder relied on adapted or self-developed questionnaires.

XSL•FO
RenderX

Table . Characteristics of 96 included studies (2019 - 2026; 37 countries; >45,000 medical students).

Characteristic and category	Studies, n (k)	Total percentage
Publication period		
2023 onward	82	85.4
Pre-2023	14	14.6
WHO[a] region		
Eastern Mediterranean	38	39.6
Europe	19	19.8
Western Pacific	12	12.5
South-East Asia	11	11.5
Americas	10	10.4
Other[b]	6	6.2
World Bank income		
High income	43	44.8
Low- or lower-middle income	28	29.2
Upper-middle income	20	20.8
Mixed income	5	5.2
Data type		
Proportion only	72	75.0
Continuous only	13	13.5
Both	11	11.5
Sample size		
<100	5	5.2
100 - 499	63	65.6
500 - 999	18	18.8
≥1000	10	10.4
JBI[c] risk of bias		
Low	17	17.7
Moderate	79	82.3
AI type evaluated		
Broad or general AI	59	61.5
ChatGPT or LLM[d]	27	28.1
Domain-specific AI	10	10.4

[a]WHO: World Health Organization.

[b]Includes Africa (n=1) and multinational (n=5).

[c]JBI: Joanna Briggs Institute.

[d]LLM: large language model.

Risk of Bias and Participant Independence

On the JBI checklist, 18% (17/96) of the studies were rated low risk and 82% (79/96) were rated moderate risk; none scored in the high-risk range (Table 2). The absence of high-risk classifications reflects the prespecified scoring threshold rather than absence of methodological limitations. The most common shortcomings were inadequate identification and management of confounding, convenience or voluntary sampling, self-developed questionnaires, unclear response rate denominators, and self-report outcomes; these limitations informed the very low certainty ratings. The participant independence assessment identified 8 clusters requiring detailed review. Four were cleared after source-level examination. In each of the remaining 4 clusters, only 1 eligible report contributed data to the meta-analysis; therefore, no pair of included studies had confirmed participant overlap. No included

XSL•FO
RenderX

study was removed for duplication, and the retained studies were covered by the prespecified leave-one-out sensitivity analyses.

Table . Methodological quality (risk of bias) of the 96 included studies, appraised with the Joanna Briggs Institute critical appraisal checklist for analytical cross-sectional studies[a].

JBI[b] appraisal item	Yes, n (%)	Unclear, n (%)	No, n (%)
Were the criteria for inclusion in the sample clearly defined?	96 (100)	0 (0)	0 (0)
Were the study subjects and the setting described in detail?	96 (100)	0 (0)	0 (0)
Was the exposure measured in a valid and reliable way?	78 (81)	18 (19)	0 (0)
Were objective, standard criteria used for measurement of the condition?	0 (0)	0 (0)	96 (100)
Were confounding factors identified?	17 (18)	0 (0)	79 (82)
Were strategies to deal with confounding factors stated?	15 (16)	0 (0)	81 (84)
Were the outcomes measured in a valid and reliable way?	96 (100)	0 (0)	0 (0)
Was appropriate statistical analysis used?	96 (100)	0 (0)	0 (0)

[a]Each item was scored Yes=1 and No or Unclear=0 (maximum 8). Overall methodological quality was summarized with review-defined descriptive thresholds (6-8=low risk of bias, 4-5=moderate, and 0-3=high); these thresholds are descriptive and were not used to exclude studies. Across the 96 studies, 17 (18%) were low risk, and 79 (82%) were moderate risk, and none were high risk; total scores ranged from 4 to 7 (score 4: n=16; 5: n=63; 6: n=4; and 7: n=13).

[b]JBI: Joanna Briggs Institute.

Proportion Outcomes

Table 3 provides the complete numerical synthesis for all 9 proportional domains, including pooled estimates, prediction intervals, and complementary model statistics. Positive attitude toward AI was endorsed by a pooled 76.9% (Figure 2; Table 3), and self-reported familiarity by 63.3% (Figure 3). Each primary proportional meta-analysis is accompanied by a domain-level forest plot in which the 95% prediction interval is printed beneath the pooled estimate (Figures 2 and 3; Figures S1-S7 in Multimedia Appendix 2). Figure 4 provides a complementary cross-domain visual overview of pooled proportions and 95% CIs only. Certainty of evidence is summarized later in the dedicated GRADE Summary of Findings. These estimates should be read as average summaries of heterogeneous self-report items rather than stable global prevalences. The 95% prediction intervals were wide across all domains and, in several domains, extended from low endorsement to near-universal endorsement, indicating that the true proportion expected in a new comparable setting could differ markedly from the pooled mean [16]. I^2 and τ^2 are reported in Table 3 for completeness; the practical magnitude of between-setting heterogeneity is interpreted from the prediction intervals and not from the I^2 values [16]. Because each domain draws on a different set of studies, instruments, denominators, and item wordings, comparisons across domains (eg, attitude vs trust, or familiarity vs curricular support) are indirect and should be read as descriptive and exploratory rather than as a ranking of constructs.

XSL•FO
RenderX

Table . Primary proportional meta-analyses, 95% prediction intervals, and model statistics (Freeman-Tukey double-arcsine transformation; Hartung-Knapp-Sidik-Jonkman [HKSJ]–adjusted CIs).

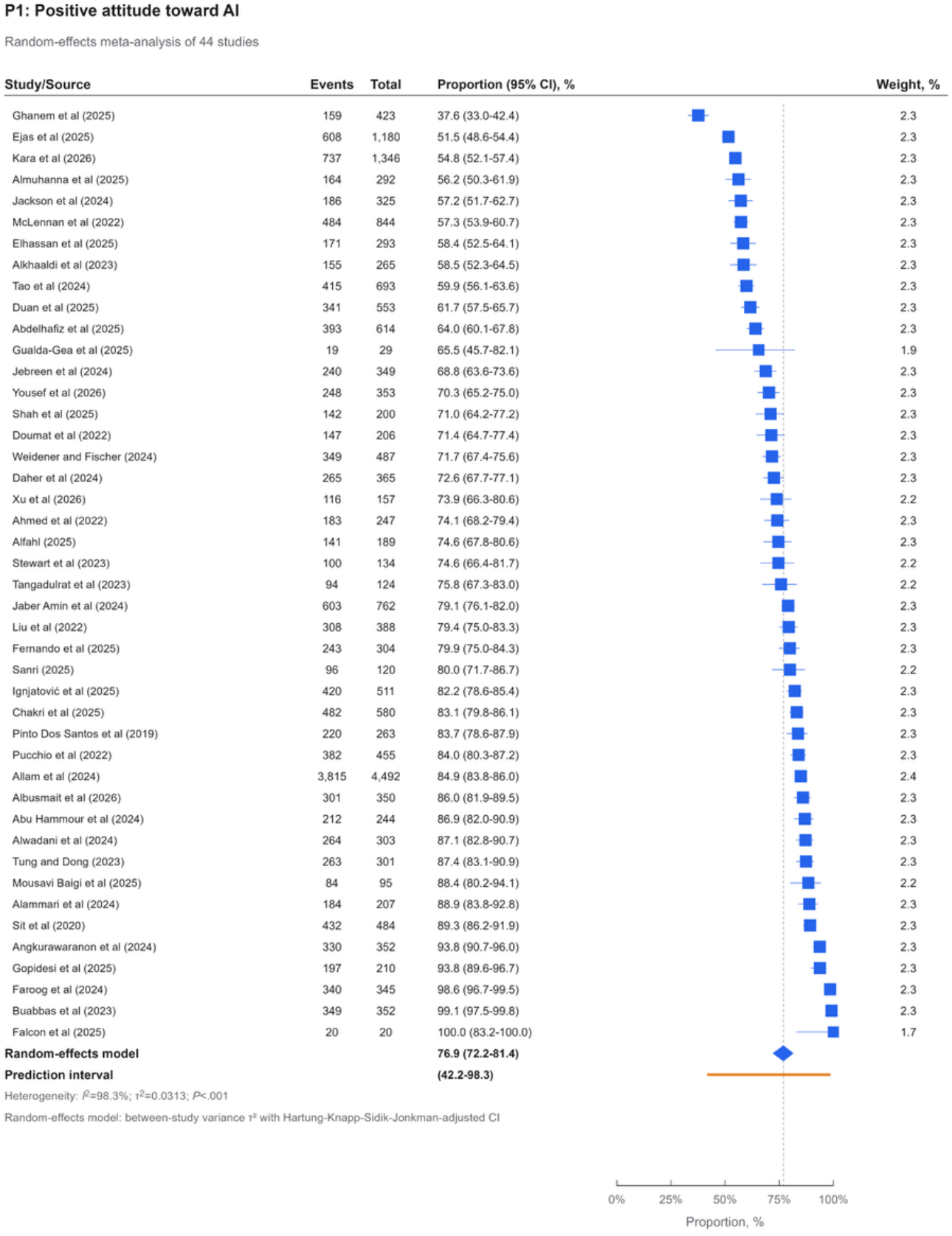
Outcome domain	*k*	N	Pooled proportion, % (95% CI)	95% PI[a], %	I^2, %	τ^2
P1: positive attitude	44	20,806	76.9 (72.2 - 81.4)	42.2 - 98.3	98.3	0.031
P2: career benefit	16	9799	78.4 (69.5 - 86.2)	45.3 - 98.3	98.0	0.025
P3: curricular integration	38	16,308	76.6 (71.8 - 81.1)	47.8 - 96.1	97.2	0.021
P4: replacement concern	32	16,642	39.9 (33.6 - 46.5)	6.6 - 80.1	98.8	0.042
P5: familiarity or knowledge	52	27,817	63.3 (55.9 - 70.3)	7.8 - 100.0	99.5	0.097
P6: prior AI use	36	17,374	63.2 (53.2 - 72.6)	8.5 - 100.0	99.4	0.091
P7: willingness to learn or adopt	22	9199	71.5 (64.8 - 77.8)	37.9 - 95.4	97.7	0.026
P8: ethical concerns	28	14,571	62.8 (53.9 - 71.3)	21.9 - 95.0	98.8	0.042
P9: trust in AI-assisted decisions[b]	8	3007	50.6 (28.5 - 72.6)	7.3 - 93.3	98.1	0.041

[a]PI: prediction interval.

[b]$k<10$; interpret with caution.

XSL•FO
RenderX

Figure 2. Forest plot of positive attitude toward AI (domain P1) across 44 studies (N=20,806). Blue squares indicate study-specific proportions, with square size proportional to the study's random-effects weight; horizontal blue lines indicate 95% CIs. The blue diamond indicates the pooled random-effects estimate and its 95% CI, and the orange horizontal line indicates the 95% prediction interval [10,35,38,39,41,44,46,50,55,57,58,60,62-67,69-71,81,82,84,86-88,90-92,94,95,100,107,108,110,111,113,116-118,121,122,128].

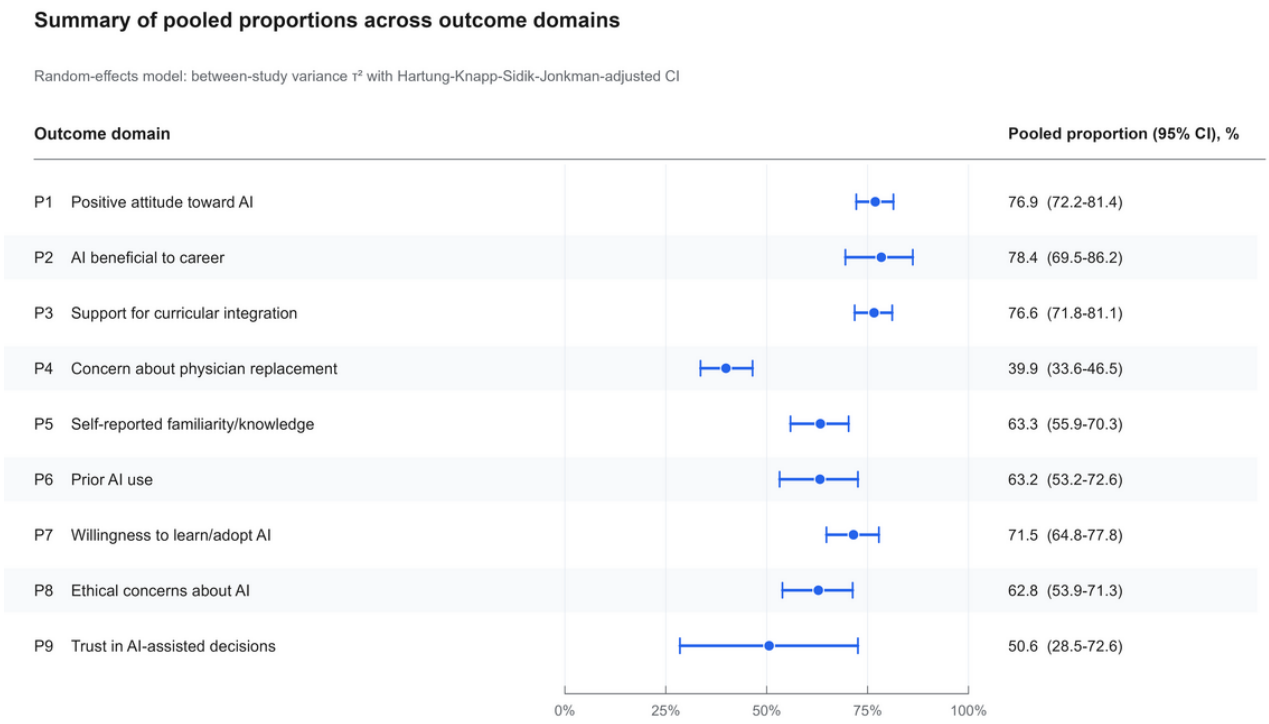


XSL•FO
RenderX

Figure 3. Forest plot of self-reported familiarity with or knowledge of AI (domain P5) across 52 studies (N=27,817). Blue squares indicate study-specific proportions, with square size proportional to the study's random-effects weight; horizontal blue lines indicate 95% CIs. The blue diamond indicates the pooled random-effects estimate and its 95% CI, and the orange horizontal line indicates the 95% prediction interval [10,35,38-40,44-47,50,52-55,58-62,64-67,69,70,76,81,84-88,90,92,94,98-100,102,103,105,110,111,113,114,117,119,121-123,127,128].

XSL•FO
RenderX

Figure 4. Cross-domain graphical summary of the random-effects pooled proportions and 95% CIs for the 9 outcome domains (P1-P9).



Subgroup Analyses

Because the primary meta-analyses yielded very wide prediction intervals (Table 3), indicating substantial expected between-setting dispersion, the prespecified subgroup analyses and meta-regressions were undertaken to explore potential sources of this variation. Selected prespecified subgroup analyses are summarized for 4 main domains across 4 moderators (Table 4; Figures S8-S27 in Multimedia Appendix 2). Four statistically significant subgroup interactions emerged among the displayed domains. Self-reported familiarity was higher in studies published from 2023 onward (66.4%) than before 2023 (45.6%; *P*=.04). AI type was associated with positive attitude (broad or general AI, 80.1%; ChatGPT or LLM, 66.7%; domain-specific AI, 83.2%; *P*=.005), replacement concern (45.9%, 43.4%, and 27.2%, respectively; *P*=.01), and familiarity (60.9%, 77.4%, and 45.1%, respectively; *P*=.008). Prior AI use was 64.9% in studies published from 2023 onward, while the pre-2023 level was not pooled because only 1 study contributed. No statistically significant differences were detected between low-risk and moderate-risk studies in any displayed domain; these null comparisons should not be interpreted as evidence of equivalence because subgroup power was limited.

XSL•FO
RenderX

Table . Prespecified subgroup analyses for selected proportion domains by publication period, World Bank income, AI type, and risk of bias.

Subgroup	P1, % (95% CI)	P4, % (95% CI)	P5, % (95% CI)	P6, % (95% CI)
Publication period				
2023 onward	76.8 (71.2 - 82.0)	41.9 (34.9 - 49.2)	66.4 (58.3 - 74.1)	64.9 (55.4 - 73.9)
Pre-2023	77.7 (67.4 - 86.5)	31.4 (15.2 - 50.4)	45.6 (34.4 - 56.9)	Not pooled (*k*=1)
P value for interaction	.98	.17	.04	Not run
World Bank income				
High income	79.2 (71.1 - 86.2)	33.9 (24.6 - 44.0)	65.9 (54.6 - 76.4)	62.1 (47.2 - 75.9)
Upper-middle income	77.0 (65.4 - 86.8)	47.4 (36.9 - 58.0)	60.9 (44.4 - 76.3)	75.0 (52.4 - 92.1)
Lower-middle or low income	74.9 (65.7 - 83.1)	47.8 (35.8 - 59.9)	63.6 (48.9 - 77.1)	49.8 (29.3 - 70.3)
P value for interaction	.69	Not run	.68	Not run
AI type				
Broad or general	80.1 (74.8 - 84.9)	45.9 (36.6 - 55.4)	60.9 (52.5 - 69.0)	57.0 (40.0 - 73.2)
ChatGPT or LLM[a]	66.7 (56.3 - 76.3)	43.4 (29.7 - 57.6)	77.4 (62.6 - 89.4)	69.1 (57.6 - 79.6)
Domain-specific	83.2 (69.8 - 93.3)	27.2 (17.8 - 37.7)	45.1 (23.2 - 68.1)	Not run
P value for interaction	.005	.01	.008	.31
JBI[b] risk of bias				
Low	74.2 (21.1 - 100.0)	51.6 (16.2 - 86.1)	63.4 (32.7 - 89.2)	76.5 (54.1 - 93.0)
Moderate	77.2 (72.7 - 81.4)	38.3 (31.9 - 44.9)	63.2 (55.7 - 70.4)	59.1 (47.7 - 70.0)
P value for interaction	.94	.17	.89	.10

[a]LLM: large language model.

[b]JBI: Joanna Briggs Institute.

Meta-Regression

Univariable meta-regression was performed for 20 domain-moderator combinations (Tables S13 and S14 in Multimedia Appendix 2). These meta-regression tests assess whether a moderator explains between-study variance across the full set of contributing studies and therefore complements, rather than duplicates, the categorical subgroup comparisons reported in Table 4; the small difference in the number of significant findings reflects the different statistical questions addressed by each approach. Five models reached statistical significance: AI type was associated with between-study variation in positive attitude (P1; *P*=.005), perceived career benefit (P2; *P*=.04), replacement concern (P4; *P*=.01), and self-reported familiarity (P5; *P*=.008), while publication period was associated with variation in self-reported familiarity (P5; *P*=.04).

Sensitivity Analyses

Across all 9 domains, sensitivity analyses produced directionally consistent estimates, although the magnitude of leave-one-out shifts varied by domain (Table 5; Table S15 in Multimedia Appendix 2). The largest maximum change was observed for trust in AI-assisted decisions (6.7 percentage points), followed by perceived career benefit and ethical concerns (2.8 percentage points each); all other domains shifted by 1.7 percentage points or less. Leave-one-out analysis confirmed that no single study changed the overall interpretation of any summary estimate (Figures S28-S36 in Multimedia Appendix 2).

XSL•FO
RenderX

Table . Sensitivity analyses and small study effects assessment for 9 proportion domains.

Domain	Primary, %	Sensitivity range, %	Maximum Δ, pp[a]	Egger *P* value
P1: positive attitude	76.9	76.1 - 77.7	0.8	<.001
P2: career benefit	78.4	77.2 - 81.2	2.8	.13
P3: curricular integration	76.6	75.6 - 77.5	1.0	<.001
P4: replacement concern	39.9	38.7 - 40.9	1.3	.46
P5: familiarity	63.3	62.1 - 64.4	1.1	<.001
P6: prior AI use	63.2	61.6 - 64.9	1.7	<.001
P7: willingness to adopt	71.5	70.2 - 72.9	1.3	<.001
P8: ethical concerns	62.8	61.2 - 65.6	2.8	.48
P9: trust in AI	50.6	44.1 - 57.3	6.7	Not performed (k<10)

[a]pp: percentage points.

Small Study Effects

Small study effect tests were performed for domains with k≥10 and not performed for trust in AI-assisted decisions (k=8). Regression tests on logit proportions [129,130] indicated asymmetry for several domains, including positive attitude, curricular integration, familiarity or knowledge, prior AI use, and willingness to learn or adopt AI (funnel plots, Figures S37-S44 in Multimedia Appendix 2; Table S16 in Multimedia Appendix 2). These results are reported as small study effects rather than proof of publication bias because funnel asymmetry in proportion meta-analyses can also arise from heterogeneity, instrument differences, sampling frames, and outcome definition.

Continuous Outcomes and Certainty of Evidence

Continuous outcomes were synthesized from the extracted MAIRS-MS data. Twelve studies contributed to the total readiness score, and 11 studies contributed to each subscale synthesis. The contributing studies differed in setting and instrument implementation, and their study-level estimates varied across the forest plots (Table S10 and Figures S45-S49 in Multimedia Appendix 2). The pooled means therefore have limited stand-alone and cross-setting interpretability. These secondary analyses are presented descriptively and are not used as a basis for strong conclusions or curricular recommendations. Bubble plots for the statistically significant meta-regression moderators are provided in Figures S50-S54 in Multimedia Appendix 2.

Certainty of evidence was assessed with the GRADE framework applied to each pooled proportion as a single-group absolute estimate and was rated very low for all 9 domains. Table 6 presents the reader-facing Summary of Findings, restricted to the key estimate, certainty rating, and interpretation for each outcome; the full prediction intervals and complementary model statistics remain in Table 3, and the detailed GRADE evidence profile is provided in Table S17 in Multimedia Appendix 2.

XSL•FO
RenderX

Table . Grading of Recommendations, Assessment, Development and Evaluation summary of findings for medical students' AI-related attitudes, perceptions, and self-reported familiarity (single-group proportional outcomes)[a].

Outcome	Participants (studies), n (*k*[b])	Pooled absolute proportion, % (95% CI)	Certainty of evidence (GRADE[c])	What the evidence means
P1: positive attitude toward AI	20,806 (44)	76.9 (72.2 - 81.4)	⊕◯◯◯ Very low	Favorable attitudes were common but varied markedly across settings.
P2: AI perceived as beneficial to career	9799 (16)	78.4 (69.5 - 86.2)	⊕◯◯◯ Very low	Career benefit was commonly perceived, but applicability to individual schools is uncertain.
P3: support for AI curricular integration	16,308 (38)	76.6 (71.8 - 81.1)	⊕◯◯◯ Very low	Support was common; this does not establish curriculum effectiveness.
P4: concern about physician replacement	16,642 (32)	39.9 (33.6 - 46.5)	⊕◯◯◯ Very low	A substantial minority expressed concern, with marked contextual variation.
P5: self-reported familiarity or knowledge	27,817 (52)	63.3 (55.9 - 70.3)	⊕◯◯◯ Very low	Self-reported familiarity is not equivalent to objective AI literacy.
P6: prior AI use	17,374 (36)	63.2 (53.2 - 72.6)	⊕◯◯◯ Very low	Prior use does not establish competent or clinically appropriate use.
P7: willingness to learn about or adopt AI	9199 (22)	71.5 (64.8 - 77.8)	⊕◯◯◯ Very low	Willingness was common; local needs assessment remains necessary.
P8: ethical concerns about AI	14,571 (28)	62.8 (53.9 - 71.3)	⊕◯◯◯ Very low	Ethical concern may reflect appropriate awareness rather than resistance.
P9: trust in AI-assisted decisions	3007 (8)	50.6 (28.5 - 72.6)	⊕◯◯◯ Very low	Trust was highly uncertain; only 8 studies contributed.

[a]Population: students enrolled in doctor of medicine MD, MBBS, MBChB, or DO-equivalent medical programs. Settings: 37 countries across 6 WHO regions. Outcome assessment: harmonized self-report survey items from cross-sectional studies. This reader-facing table is structured in the GRADEpro Summary of Findings format, which presents the key outcome estimates and certainty ratings concisely without reproducing the detailed certainty-domain judgments. Because the review synthesizes noncomparative single-group proportions, comparator, relative-effect, and absolute-difference columns are not applicable. Full numerical results, including 95% prediction intervals, I^2, and τ^2, are reported in Table 3; the detailed domain-level GRADE evidence profile is provided in Table S17 in Multimedia Appendix 2. ⊕◯◯◯ denotes very low certainty.

[b]k: number of studies.

[c]GRADE: Grading of Recommendations, Assessment, Development and Evaluation.

Discussion

Principal Findings

This systematic review and meta-analysis synthesized 96 cross-sectional surveys of medical students and found that attitudes toward AI and interest in curricular integration were generally favorable on average but highly variable across settings (Table 3). Favorable attitudes, perceived career benefit, and support for curricular integration were the most consistently endorsed domains, whereas self-reported familiarity and trust in AI-assisted decisions were endorsed less consistently and with greater between-study variability; because each domain draws on different studies, instruments, and denominators, these patterns are descriptive rather than a ranking across constructs (Table 3). These values should be interpreted as average summaries of nonequivalent self-report items rather than stable global prevalences (Table 3; Table S11 in Multimedia Appendix 2). The central message is therefore not that medical students uniformly accept AI but that receptivity, familiarity, concern, and trust vary substantially by context, technology, instrument, and respondent population (Table 3; Tables S6 and S11 in Multimedia Appendix 2). This broad pattern of favorable but uneven perceptions is directionally consistent with prior review-level evidence in health professions learners [12-14].

The prediction intervals are important for interpreting these results [16]. For several domains, the interval expected in a new comparable setting spanned from low endorsement to near-universal endorsement, even when the CI around the average summary was relatively narrow [16]. This distinction addresses a key limitation of interpreting the pooled mean alone: a precise-looking average can obscure wide between-setting variation [16]. In practical terms, a medical school should not assume that a global summary estimate reflects its own students and should compare the synthesis with a local baseline before curriculum planning, implementation, or evaluation [16,131]. In several domains, the dispersion of true effects is large enough

XSL•FO
RenderX

that the summary estimate has limited stand-alone interpretive value and is best interpreted together with its prediction interval [16].

Comparison With Prior Work

The findings are broadly consistent with prior reviews showing interest in AI among health professions learners [12-14], but this review narrows the inference to students in medical education programs and adds several safeguards that were not consistently present in earlier work [12-14]. We reviewed mixed populations, documented disaggregation decisions, harmonized nonidentical survey items into operational domains, assessed participant independence, used HKSJ-adjusted CIs, reported prediction intervals, and applied the GRADE framework to proportional self-report outcomes [15,16,29-31]. These steps improve transparency while making the limitations of the evidence more explicit, which is central to responsible interpretation of heterogeneous review findings [16,18,29-31]. First, the review provides medical education–specific benchmark estimates in which between-setting dispersion is displayed rather than hidden, so curriculum committees can see both the central tendency and the realistic range they may encounter [16]. Second, the harmonization taxonomy, mixed population disaggregation log, and participant independence audit are reusable procedures that future syntheses of heterogeneous survey evidence can adopt directly, complementing standardized review process reporting under PRISMA 2020 [18]. Third, by documenting where construct equivalence breaks down, the review specifies a measurement agenda that separates exposure, perceived familiarity, conceptual knowledge, and trust, and supports greater use of validated instruments such as MAIRS-MS [7,23].

Recent medical education literature argues that learners need foundational AI concepts and professionally relevant competencies [132]. Ethical reviews highlight bias, privacy, accountability, authorship, and responsible use as substantive educational concerns [133]. Critical-appraisal guidance also emphasizes validation, model limitations, and uncertainty when evaluating AI-enabled evidence [134]. Reviews of AI and generative AI in medical education describe rapidly expanding educational uses and the need for deliberate curricular responses [135,136]. Curriculum-focused literature further emphasizes structured training, critical appraisal, and human oversight [137-140]. Our findings are consistent with this direction: favorable attitudes and interest in curricular exposure coexisted with uneven self-reported familiarity and trust (Table 3). This pattern can inform educational needs assessment, but cross-sectional perception data do not demonstrate that any specific curriculum improves competence, trust calibration, ethical reasoning, or patient care outcomes [131,139]. A recent Best Evidence Medical Education scoping review similarly documented rapid growth in AI-related publications in medical education, concentrated in undergraduate training, and mapped priorities for the evaluative research the field still needs [141]. By quantifying the perception layer of this literature while displaying its uncertainty, our synthesis clarifies what the current descriptive evidence can and cannot support as those evaluative studies are developed [16,141].

This difference from prior reviews is clinically and educationally relevant [6,12-14]. Mixed health profession syntheses can be useful for mapping broad interest in AI, but medical students are still developing foundational clinical reasoning and evidence appraisal practices while learning how technology is incorporated into supervised clinical work [6,132,137-139]. Educational priorities at this stage therefore cannot be assumed to mirror those of practicing physicians, dental students, nursing students, or faculty [6,12-14,132]. A medical student–specific synthesis provides a more relevant evidence base for undergraduate medical education planning, while the wide between-setting dispersion still requires local adaptation [16,131]. For curriculum committees and other educational decision makers, an estimate anchored to a clearly defined learner population and accompanied by explicit construct mapping and prediction intervals is more actionable than a broad average whose constituent populations, instruments, and settings cannot be separated [16,23,131]. Accordingly, the harmonization documentation and prediction intervals are presented alongside the summary estimates rather than left as technical appendices [16,23].

The methodological contrast with prior syntheses is consequential rather than merely technical [12-16]. The only previous meta-analysis in this area pooled attitude and knowledge outcomes across medical, dental, and nursing students and did not apply HKSJ-adjusted CIs or report prediction intervals [14-16]. In the present review, the 95% prediction intervals were broad across all 9 domains—for example, 42.2% - 98.3% for positive attitude and 7.8% - 100% for self-reported familiarity—showing that a pooled mean may be a poor proxy for the true proportion in a new comparable setting (Table 3) [16,142]. This changes how the evidence should be used: pooled values are contextual benchmarks, not transportable targets, and should be compared with local needs assessments before curricular priorities are selected [16,131]. In practice, a local needs assessment showing strong support for curricular integration but limited familiarity could support prioritizing foundational concepts and supervised use, whereas high exposure accompanied by poorly calibrated reliance could indicate a greater need for critical appraisal, model limitations, and trust calibration [131,132,139]. These examples illustrate decision pathways for locally measured profiles, not associations between pooled domains, because each domain includes a different set of studies (Table 3). Medical student–only eligibility and explicit construct mapping reduce population and measurement ambiguity [7,23], while GRADE makes the very low certainty of each domain estimate transparent [29-31]. Together, these features move the field from broad descriptions of enthusiasm toward testable, context-specific curriculum planning: pilots can target documented local gaps and assess domain-specific competencies rather than satisfaction alone [131,132,139,141]. Because all included studies were cross-sectional, these methodological features do not show that educational or patient outcomes have already improved (Table 1) [17]. Their practical contribution is instead to strengthen the evidence base for designing, targeting, and evaluating interventions intended to improve those outcomes [131,139,141]. We regard these differences not as criticism of earlier

XSL•FO
RenderX

contributions but as the methodological maturation of evidence synthesis in a rapidly developing field [12-16,141].

Heterogeneity, Subgroup Analyses, Meta-Regression, and Sensitivity Findings

The 95% prediction intervals were wide across all proportional domains and, for several outcomes, spanned low to near-universal endorsement, indicating that the true proportion expected in a new comparable setting may differ markedly from the pooled mean (Table 3) [16,142]. Following Borenstein [16], we therefore interpret the practical magnitude of between-setting heterogeneity from the prediction intervals rather than from I^2. I^2 and τ^2 remain reported in Table 3 as complementary model statistics, but neither I^2 thresholds nor the I^2 range are used to characterize how widely true proportions vary across settings [16]. The observed dispersion may reflect contextual differences in country, institutional environment, clinical exposure, prior AI training, and survey period, as suggested by the diversity described in prior reviews and primary studies [10,12-14,107,116,128]. It also plausibly reflects measurement heterogeneity: studies used different item wording, Likert thresholds, handling of neutral responses, and AI referents ranging from broad AI to radiology algorithms and generative AI [7,10,12-14,23]. The harmonization table is therefore not a cosmetic supplement; it is necessary for understanding what was actually pooled and where construct equivalence becomes uncertain [23,28]. The wide prediction intervals provided the empirical rationale for the prespecified subgroup analyses and feasibility-screened meta-regressions: they demonstrate that true proportions are expected to vary widely across comparable settings, while not identifying the causes of that variation by themselves [16,142].

The subgroup analyses should be interpreted as descriptive probes rather than explanatory models, and the most consistent signal involved the type of AI asked about [143]. Studies framed around ChatGPT or other generative AI reported higher self-reported familiarity (77.4%) than studies of broad or general AI (60.9%) or domain-specific applications (45.1%; P=.008; Table 4), which is plausible given the rapid diffusion of generative tools in medical education after late 2022 [11,135,141]. At the same time, ChatGPT or LLM-focused studies reported lower positive attitude (66.7%) than broad-AI (80.1%) or domain-specific studies (83.2%; P=.005; Table 4). One cautious interpretation is that direct exposure may coexist with more tempered enthusiasm as students encounter limitations, ethical ambiguities, and academic integrity concerns described in the generative-AI literature [11,133,136,144]. Concern about physician replacement was lower in studies of domain-specific applications (27.2%) than in studies of broad AI (45.9%) or generative AI (43.4%; P=.01; Table 4); this pattern is compatible with prior work showing that the framing and intended role of AI can shape student attitudes and professional concerns [10,122,128]. Self-reported familiarity was also higher in studies published from 2023 onward than before 2023 (66.4% vs 45.6%; P=.04; Table 4), a pattern temporally consistent with wider generative-AI exposure [11,135,141]. All of these are study-level associations that meet few of the established credibility criteria for subgroup effects: they are observational, rest on between-study rather than within-study comparisons, involve multiple tests without multiplicity adjustment, and concern moderators that are correlated with region, language, curriculum, sampling strategy, and survey instrument [143]. They are therefore best read as coherent, hypothesis-generating patterns rather than established effects [143].

Equally informative are the moderators for which no statistically significant subgroup difference was detected (Table 4). The displayed analyses provided no clear evidence of differences by World Bank income level or JBI risk-of-bias category (Table 4). These null findings do not establish equivalence and may reflect limited subgroup power; accordingly, they should not be used either to assert a universal income-level pattern or to conclude that risk-of-bias category has no moderating role [143]. Prespecified feasibility rules further limited unstable moderator comparisons: of 36 candidate domain-moderator combinations, 20 were analyzed and 16 were not run, most commonly because a moderator level contained fewer than 3 studies or because a domain had fewer than 10 studies (Tables S13 and S14 in Multimedia Appendix 2). These decisions are reported in full so that absence of an analysis is not mistaken for absence of an effect, consistent with cautious interpretation of subgroup evidence [143]. Leave-one-out analyses address a different question—the influence of individual studies on the pooled summaries—and no single omission changed the interpretation of any domain; the maximum shift was 6.7 percentage points for trust, the smallest domain (k=8), and 2.8 percentage points or less for the remaining 8 domains (Table 5; Figures S28-S36 in Multimedia Appendix 2).

The meta-regressions provide a complementary study-level analysis of moderator patterns (Tables S13 and S14 in Multimedia Appendix 2). Of the 20 feasibility-screened univariable models, 5 were statistically significant and identified the same 2 moderators as the categorical comparisons: AI type (for positive attitude, perceived career benefit, replacement concern, and familiarity) and publication period (for familiarity) (Table S13 in Multimedia Appendix 2). This convergence is internally coherent, but it does not upgrade the causal status of either moderator because meta-regression on study-level summaries remains vulnerable to aggregation bias and confounding by correlated study characteristics [17,143]. The number of studies per moderator level was sometimes small, and nominal P values were not adjusted for multiplicity [143]. The bubble plots and full model listings are therefore provided in Tables S13 and S14 and Figures S50-S54 in Multimedia Appendix 2, and our interpretation is confined to the descriptive statement that the AI referent and survey period are candidate axes for future individual-level studies of effect modification [143].

Finally, small study effects deserve explicit discussion because they feed directly into the certainty assessment [29-31]. Egger regression tests on logit proportions indicated funnel asymmetry for 5 of the 8 testable domains: positive attitude, curricular integration, familiarity, prior AI use, and willingness to learn or adopt AI; no asymmetry was detected for career benefit, replacement concern, or ethical concerns, and the test was not performed for trust (k=8; Table 5; Table S16 and Figures S37-S44 in Multimedia Appendix 2). In proportion

XSL•FO
RenderX

meta-analyses, funnel asymmetry cannot be equated with publication bias because heterogeneity, instrument differences, sampling frames, and outcome definitions can produce the same signature [26,129,130]. These findings were therefore treated in the GRADE assessment as a possible publication bias signal rather than as demonstrated bias; they contribute to, but are not the sole reason for, the very low certainty ratings (Table 6; Table S17 in Multimedia Appendix 2) [29-31].

Risk of Bias, Certainty, and Outcome Harmonization

Risk-of-bias findings should also be read conservatively in light of the design and measurement limitations of cross-sectional survey evidence [12-14,24]. No study met our review-defined high-risk threshold (Table 2), but this does not mean that the evidence base is methodologically strong. Voluntary cross-sectional surveys, convenience sampling, locally adapted questionnaires, uncertain response rate denominators, and self-selection are recurring limitations in this literature and can constrain representativeness and measurement validity [12-14,24]. The item-level pattern in our review (Table 2) locates these weaknesses precisely: inclusion criteria, outcome measurement, and statistical analysis were adequately reported in all 96 studies; confounding factors were identified in only 17 studies (18%) and addressed in 15 (16%); no study was rated Yes on JBI item 4, which asks whether objective, standard criteria were used to measure the condition; and exposure measurement was unclear in 18 studies (19%). For the perception-based outcomes synthesized here, the item 4 pattern should be interpreted in the context of what the checklist asks rather than as evidence that the surveys measured an objective clinical condition (Table 2) [24]. These limitations are especially important because the outcomes are perceptions and self-reported familiarity rather than objective literacy or behavior [7,23,132,139].

For this reason, GRADE certainty was very low across domains (Table 6; Table S17 in Multimedia Appendix 2). Downgrading was driven not only by risk of bias but also by inconsistency, indirectness, and imprecision [29-31]. The very wide prediction intervals were central to the inconsistency and imprecision judgments because they indicate substantial uncertainty about the true proportion expected in a new comparable setting [16,30,31]. Indirectness arose because domains such as familiarity, positive attitude, ethical concern, and trust were measured with nonequivalent items that required operational harmonization [23,30]. Publication bias was difficult to separate from other sources of funnel asymmetry, but the observed small study effects in several domains were treated as an additional reason for caution rather than as proof of selective publication [26,129,130].

A related issue is that self-reported familiarity should not be treated as equivalent to AI literacy [7,23,132]. Exposure to ChatGPT or diagnostic AI does not by itself demonstrate understanding of model development, validation, calibration, data drift, fairness, privacy, or clinical accountability—competencies emphasized in AI-literacy and professional-use frameworks [23,132,134,139,145]. Conversely, students with limited exposure may still hold well-founded concerns about safety and professional responsibility [133,144,145]. Future studies should therefore distinguish exposure, perceived familiarity, conceptual knowledge, applied appraisal skills, and behavior [23,132,139]. Validated instruments such as MAIRS-MS [7] are helpful, but the field also needs practical assessments of whether students can recognize unreliable AI outputs, identify unsupported claims, and decide when clinician oversight is required [132,134,139]. The continuous MAIRS-MS syntheses likewise showed variation in study-level estimates across settings despite the use of a validated instrument [7]; accordingly, the pooled means are treated as secondary descriptive summaries rather than transportable estimates of readiness (Table S10 and Figures S45-S49 in Multimedia Appendix 2).

Implications for AI Literacy Curriculum Development

The curricular implication is exploratory and context-dependent because the available evidence describes perceptions rather than intervention effects [12-14,141]. The findings suggest the need to consider locally adapted AI literacy curricula that address core terminology, common clinical and educational use cases, critical appraisal of AI-enabled evidence, data bias, privacy, accountability, uncertainty, and appropriate human oversight [132,134,139]. Curriculum design should also distinguish between AI used for learning support and AI used in clinical decision-making because the educational aims, disclosure expectations, and safety consequences differ across these contexts [11,136,144,145]. Learning support use may emphasize appropriate prompting, verification of generated content, academic integrity, and transparent disclosure [11,136,144], whereas clinical use requires attention to validation, workflow, responsibility, patient communication, and professional judgment [132,134,139,145].

Trust calibration could reasonably be treated as a specific curricular goal because safe AI use requires neither reflexive acceptance nor reflexive rejection but judgment about when an output is sufficiently reliable for a particular task [139,145]. Trust in AI-assisted decisions may refer to distinct constructs—trust in diagnostic accuracy, willingness to follow AI recommendations, trust in opaque systems, or trust in specific clinical applications—and prior student surveys likewise show that attitudes vary with the type and role of AI being considered [10,98,122,128]. A low or moderate summary estimate for trust should therefore not be interpreted simply as resistance to innovation, and a high estimate should not be assumed to reflect calibrated confidence [139,145]. Students need to learn when AI outputs may be useful, when they may be misleading, and how to verify claims against clinical evidence and patient context [132,134,139]. Educational interventions should therefore measure not only attitudes but also applied competencies, such as recognizing model limitations, identifying biased outputs, communicating uncertainty, and maintaining clinical reasoning while using AI tools [139].

Ethical concern should likewise be interpreted constructively because current ethical guidance for AI in health and health professions education treats bias, transparency, privacy, authorship, accountability, and professional responsibility as substantive competencies rather than peripheral concerns [133,144,145]. The relatively high endorsement of ethical

XSL•FO
RenderX

concerns in our synthesis may therefore identify curricular topics that require explicit discussion rather than resistance that needs to be overcome, an interpretation consistent with this guidance [133,144,145]. Curricula should avoid a purely promotional framing and instead help students understand when AI tools are useful, what evidence is needed before clinical deployment, how patient values and local context affect decisions, and why professional responsibility cannot simply be delegated to an algorithm [132,139,144,145].

The distinction between educational use and clinical use also matters for policy because these contexts carry different requirements for assessment integrity, validation, accountability, and patient safety [11,139,144,145]. Generative AI may support explanation, formative feedback, simulation, drafting, and self-study [11,135,136,144], whereas diagnostic decision support or treatment recommendation creates different requirements for validation, accountability, and patient safety [132,134,139,145]. Medical schools may therefore need tiered guidance, aligned with ethical and governance frameworks for AI in health and health professions education, that separates acceptable learning assistance, disclosure expectations, assessment integrity, supervision, and clinical safety [133,144,145]. Because the present evidence is based on perceptions, these policy implications should be tested through curriculum pilots, objective performance measures, longitudinal follow-up, and evaluation of unintended consequences such as automation bias or reduced independent reasoning [11,139,141].

Strengths and Limitations

This review has several methodological strengths that are directly relevant to random-effects evidence synthesis and certainty assessment [15,16,29-31,142]. It used a medical student–specific eligibility definition, reviewed mixed population studies, documented participant independence checks, harmonized outcome domains transparently, and conducted proportional meta-analyses with HKSJ-adjusted CIs and prediction intervals [15,16,142]. The synthesis also reports subgroup analyses, feasibility-screened meta-regression, leave-one-out sensitivity analyses, small study effect tests, risk of bias, and GRADE certainty [29-31]. Together, these features help distinguish the pooled mean from the between-setting dispersion that is central to interpreting a random-effects meta-analysis [16,142].

Limitations remain substantial because the evidence base consists entirely of cross-sectional self-report studies, which cannot establish changes over time, causal effects of AI exposure, or curriculum effectiveness [12-14]. Survey instruments, AI referents, response scales, and dichotomization rules varied widely, a recurring challenge in this literature [7,12-14,23]. Some mixed population studies required disaggregation, and some potentially relevant data were nonextractable (Tables S3-S5 in Multimedia Appendix 2). Meta-regression was limited by study-level confounding and multiple testing, which constrains causal interpretation of moderator findings [17,143]. Google Scholar and citation search reproducibility depended on the search records available to the review team; transparent reporting of such search process limitations is particularly important for systematic review reproducibility [19]. Finally, trust in AI-assisted decisions was supported by fewer than 10 studies and should be treated as especially uncertain (Tables 3 and 6). More broadly, meta-analysis of cross-sectional perception data has inherent interpretive limits even when conducted rigorously: pooling can quantify and display variation across settings, but it cannot convert nonequivalent self-report items into a single transportable prevalence [16,28].

Conclusions

Among medical students, AI-related attitudes and interest in curricular integration appear generally favorable, but familiarity, trust, and ethical concern vary substantially across settings. The principal contribution of this review is not a single transportable prevalence estimate but a medical student–specific synthesis that makes outcome harmonization, mixed population disaggregation, participant independence checks, between-setting dispersion, risk of bias, and certainty explicit. The findings support locally tailored AI literacy needs assessment and curriculum development, while not establishing the effectiveness of any particular curriculum. Future studies should use standardized, validated measures; report medical student data separately; distinguish perceived familiarity from objective competence; preserve item wording and neutral response handling; and evaluate educational interventions with longitudinal and performance-based outcomes. Until stronger evidence is available, curriculum planning should be treated as context-sensitive competency development rather than as a response to a presumed universal prevalence.

Acknowledgments

The authors thank the investigators of all primary studies included in this review for their contributions to the evidence base. No generative AI tools were used in the conception, design, data extraction, statistical analysis, interpretation, or writing of this manuscript. All work was performed exclusively by the listed authors.

Funding

This work was supported by the National Natural Science Foundation of China (grants 82472863, 82303755, and 82301716), the Natural Science Foundation of Sichuan (grants 24NSFSC6778 and 2024NSFSC1567), and the Noncommunicable Chronic Diseases—National Science and Technology Major Project (grant 2024ZD0525700).

XSL•FO
RenderX

Data Availability

The extracted proportion and continuous outcome datasets, outcome harmonization records, and source verification records are available in Tables S9-S12 in Multimedia Appendix 2. Complete search strategies are provided in Multimedia Appendix 1, and completed reporting checklists are provided in Checklist 1. Raw data extraction spreadsheets are available from the corresponding author upon reasonable request.

Authors' Contributions

Peng Zhang is a co-corresponding author and senior author of this article and can be reached at zhangpeng@wchscu.edu.cn.
Conceptualization, methodology, software, formal analysis, data curation, writing – original draft, and visualization: WZ
Methodology, investigation, data curation, writing – original draft, and visualization: JH
Investigation, data curation, and writing – review and editing: XH, HX
Investigation and writing – review and editing: LW, TL, PT
Conceptualization, methodology, writing – review and editing, supervision, and funding acquisition: PZ
Conceptualization, methodology, writing – review and editing, supervision, project administration, and funding acquisition: XZ

Conflicts of Interest

None declared.

Multimedia Appendix 1
Complete search strategies for all 6 databases.
[DOCX File, 1149 KB - mededu_v12i1e89411_app1.docx]

Multimedia Appendix 2
Supplementary study-level data and analyses, including study selection and eligibility documentation, study characteristics, risk-of-bias assessments, outcome extraction and harmonization, participant independence checks, subgroup and meta-regression analyses, sensitivity analyses, small study effects, GRADE (Grading of Recommendations, Assessment, Development and Evaluation) evidence profiles, and supplementary figures.
[DOCX File, 49340 KB - mededu_v12i1e89411_app2.docx]

Checklist 1
PRISMA 2020 expanded checklist, PRISMA 2020 for Abstracts checklist, and PRISMA-S (Preferred Reporting Items for Systematic Reviews and Meta-Analyses literature search extension) checklist.
[DOCX File, 33 KB - mededu_v12i1e89411_app3.docx]

References

1. Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. Nat Med 2019 Jan;25(1):44-56. [doi: 10.1038/s41591-018-0300-7] [Medline: 30617339]
2. Rajpurkar P, Chen E, Banerjee O, Topol EJ. AI in health and medicine. Nat Med 2022 Jan;28(1):31-38. [doi: 10.1038/s41591-021-01614-0] [Medline: 35058619]
3. Kung TH, Cheatham M, Medenilla A, et al. Performance of ChatGPT on USMLE: potential for AI-assisted medical education using large language models. PLOS Digit Health 2023 Feb;2(2):e0000198. [doi: 10.1371/journal.pdig.0000198] [Medline: 36812645]
4. Chan KS, Zary N. Applications and challenges of implementing artificial intelligence in medical education: integrative review. JMIR Med Educ 2019 Jun 15;5(1):e13930. [doi: 10.2196/13930] [Medline: 31199295]
5. Wartman SA, Combs CD. Medical education must move from the information age to the age of artificial intelligence. Acad Med 2018 Aug;93(8):1107-1109. [doi: 10.1097/ACM.0000000000002044] [Medline: 29095704]
6. Masters K. Artificial intelligence in medical education. Med Teach 2019 Sep;41(9):976-980. [doi: 10.1080/0142159X.2019.1595557] [Medline: 31007106]
7. Karaca O, Çalışkan SA, Demir K. Medical artificial intelligence readiness scale for medical students (MAIRS-MS) - development, validity and reliability study. BMC Med Educ 2021 Feb 18;21(1):112. [doi: 10.1186/s12909-021-02546-6] [Medline: 33602196]
8. Davis FD. Perceived usefulness, perceived ease of use, and user acceptance of information technology. MIS Q 1989 Sep 1;13(3):319-340. [doi: 10.2307/249008]
9. Venkatesh V, Morris MG, Davis GB, Davis FD. User acceptance of information technology: toward a unified view1. MIS Q 2003 Sep 1;27(3):425-478. [doi: 10.2307/30036540]

XSL•FO
RenderX

10. Sit C, Srinivasan R, Amlani A, et al. Attitudes and perceptions of UK medical students towards artificial intelligence and radiology: a multicentre survey. Insights Imaging 2020 Feb 5;11(1):14. [doi: 10.1186/s13244-019-0830-7] [Medline: 32025951]
11. Izquierdo-Condoy JS, Arias-Intriago M, Tello-De-la-Torre A, Busch F, Ortiz-Prado E. Generative artificial intelligence in medical education: enhancing critical thinking or undermining cognitive autonomy? J Med Internet Res 2025 Nov 3;27:e76340. [doi: 10.2196/76340] [Medline: 41183320]
12. Mousavi Baigi SF, Sarbaz M, Ghaddaripouri K, Ghaddaripouri M, Mousavi AS, Kimiafar K. Attitudes, knowledge, and skills towards artificial intelligence among healthcare students: a systematic review. Health Sci Rep 2023 Mar;6(3):e1138. [doi: 10.1002/hsr2.1138] [Medline: 36923372]
13. Shishehgar S, Murray-Parahi P, Alsharaydeh E, Mills S, Liu X. Artificial intelligence in health education and practice: a systematic review of health students' and academics' knowledge, perceptions and experiences. Int Nurs Rev 2025 Jun;72(2):e70045. [doi: 10.1111/inr.70045] [Medline: 40545441]
14. Amiri H, Peiravi S, Rezazadeh Shojaee SS, et al. Medical, dental, and nursing students' attitudes and knowledge towards artificial intelligence: a systematic review and meta-analysis. BMC Med Educ 2024 Apr 15;24(1):412. [doi: 10.1186/s12909-024-05406-1] [Medline: 38622577]
15. IntHout J, Ioannidis JPA, Borm GF. The Hartung-Knapp-Sidik-Jonkman method for random effects meta-analysis is straightforward and considerably outperforms the standard DerSimonian-Laird method. BMC Med Res Methodol 2014 Feb 18;14:25. [doi: 10.1186/1471-2288-14-25] [Medline: 24548571]
16. Borenstein M. How to understand and report heterogeneity in a meta-analysis: the difference between I-squared and prediction intervals. Integr Med Res 2023 Dec;12(4):101014. [doi: 10.1016/j.imr.2023.101014] [Medline: 38938910]
17. Higgins JPT, Thomas J, Chandler J, Cumpston M, Li T, Page MJ. Cochrane Handbook for Systematic Reviews of Interventions, 2nd edition: John Wiley & Sons; 2019.
18. Page MJ, McKenzie JE, Bossuyt PM, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ 2021 Mar 29;372:n71. [doi: 10.1136/bmj.n71] [Medline: 33782057]
19. Rethlefsen ML, Kirtley S, Waffenschmidt S, et al. PRISMA-S: an extension to the PRISMA statement for reporting literature searches in systematic reviews. Syst Rev 2021 Jan 26;10(1):39. [doi: 10.1186/s13643-020-01542-z] [Medline: 33499930]
20. World Federation for Medical Education. Basic Medical Education: WFME Global Standards for Quality Improvement: The 2015 Revision: World Federation for Medical Education; 2015. URL: https://wfme.org/wp-content/uploads/2015/01/BME_2015_v2.pdf [accessed 2026-08-05]
21. Bramer WM, Giustini D, de Jonge GB, Holland L, Bekhuis T. De-duplication of database search results for systematic reviews in EndNote. J Med Libr Assoc 2016 Jul;104(3):240-243. [doi: 10.3163/1536-5050.104.3.014] [Medline: 27366130]
22. Ajzen I. The theory of planned behaviour: reactions and reflections. Psychol Health 2011 Sep;26(9):1113-1127. [doi: 10.1080/08870446.2011.613995] [Medline: 21929476]
23. Ng DTK, Leung JKL, Chu SKW, Qiao MS. Conceptualizing AI literacy: An exploratory review. Comput Educ Artif Intell 2021;2:100041. [doi: 10.1016/j.caeai.2021.100041]
24. Moola S, Munn Z, Tufanaru C, Aromataris E, Sears K, Sfetcu R, et al. Chapter 7: systematic reviews of etiology and risk. In: Aromataris E, Munn Z, editors. JBI Manual for Evidence Synthesis: JBI; 2020. [doi: 10.46658/JBIRM-17-06]
25. Freeman MF, Tukey JW. Transformations related to the angular and the square root. Ann Math Statist 1950 Dec;21(4):607-611. [doi: 10.1214/aoms/1177729756]
26. Rücker G, Schwarzer G, Carpenter J. Arcsine test for publication bias in meta-analyses with binary outcomes. Stat Med 2008 Feb 28;27(5):746-763. [doi: 10.1002/sim.2971] [Medline: 17592831]
27. DerSimonian R, Laird N. Meta-analysis in clinical trials. Control Clin Trials 1986 Sep;7(3):177-188. [doi: 10.1016/0197-2456(86)90046-2] [Medline: 3802833]
28. Borenstein M, Hedges LV, Higgins JPT, Rothstein HR. A basic introduction to fixed-effect and random-effects models for meta-analysis. Res Synth Methods 2010 Apr;1(2):97-111. [doi: 10.1002/jrsm.12] [Medline: 26061376]
29. Schünemann HJ. Using systematic reviews in guideline development: the GRADE approach. In: Egger M, Higgins JPT, Davey Smith G, editors. Systematic Reviews in Health Research: Meta-Analysis in Context, 3rd edition: Wiley; 2022:424-448. [doi: 10.1002/9781119099369.ch22]
30. Murad MH, Mustafa RA, Schünemann HJ, Sultan S, Santesso N. Rating the certainty in evidence in the absence of a single estimate of effect. Evid Based Med 2017 Jun;22(3):85-87. [doi: 10.1136/ebmed-2017-110668]
31. Iorio A, Spencer FA, Falavigna M, et al. Use of GRADE for assessment of evidence about prognosis: rating confidence in estimates of event rates in broad categories of patients. BMJ 2015 Mar 16;350:h870. [doi: 10.1136/bmj.h870] [Medline: 25775931]
32. Sami A, Tanveer F, Sajwani K, et al. Medical students' attitudes toward AI in education: perception, effectiveness, and its credibility. BMC Med Educ 2025;25(1):82. [doi: 10.1186/s12909-025-06704-y]
33. Farooq M, Usmani A. Artificial intelligence in medical education. J Coll Physicians Surg Pak 2025 Apr 1;35(4):503-507. [doi: 10.29271/jcpsp.2025.04.503]

XSL•FO
RenderX

34. Cao X, Lu YY, Li JH, et al. Medical students perceptions and attitudes toward the use of generative artificial intelligence in clinical decision-making: a nationwide cross-sectional survey in China. BMC Med Educ 2026;26(1):384. [doi: 10.1186/s12909-026-08698-7]
35. Kara AV, Harmancı H, Yılmaz Y. Awareness, attitudes, and educational use of artificial intelligence among medical students: a large cross-sectional survey. BMC Med Educ 2026;26(1):758. [doi: 10.1186/s12909-026-09108-8]
36. Jamjoom J, Alkhwaiter F, Armosh G, et al. Perception of integrating an AI teaching module into medical education curriculum. Front Med (Lausanne) 2026;13:1774935. [doi: 10.3389/fmed.2026.1774935] [Medline: 41852526]
37. Xu R, Hu C, Mu T. Personality, usage, and perceptions of AI in medical education: evidence from senior pre-clinical students in China. Front Psychol 2026;17:1805800. [doi: 10.3389/fpsyg.2026.1805800] [Medline: 41909551]
38. Albusmait A, AlSaleh A, Ahmed J. Medical students' perceptions and attitude toward the use of artificial intelligence in healthcare. Adv Med Educ Pract 2026;17:541284. [doi: 10.2147/AMEP.S541284] [Medline: 41867882]
39. Xu AY, Speakman S, Piranio VS, et al. Medical student experiences with ChatGPT: national cross-sectional study. JMIR Form Res 2026 Mar 9;10:e76838. [doi: 10.2196/76838] [Medline: 41802232]
40. Prasad S, Khot A, Dakhale G. Knowledge, attitude and practice about artificial intelligence among medical undergraduate and postgraduate students in a tertiary care centre: a cross-sectional survey. Cureus 2026 Jan;18(1):e100576. [doi: 10.7759/cureus.100576] [Medline: 41625040]
41. Yousef YA, Shdeifat A, Yousef L, et al. Artificial intelligence in ophthalmology: trust, bias, and responsibility from the perspective of medical students and ophthalmologists. Front Ophthalmol (Lausanne) 2026 Mar;6:1766974. [doi: 10.3389/fopht.2026.1766974] [Medline: 41852891]
42. Catan-Inan F. Artificial intelligence literacy and readiness in future health care professionals: a cross-sectional study. Croat Med J 2026 Mar 20;67(1):4-13. [doi: 10.3325/cmj.2026.67.4] [Medline: 41866897]
43. Hu N, Jiang XQ, Wang YD, et al. Status and perceptions of ChatGPT utilization among medical students: a survey-based study. BMC Med Educ 2025 Jun 4;25(1):831. [doi: 10.1186/s12909-025-07438-7] [Medline: 40468340]
44. Ghanem OA, Hagag AM, Kormod ME, et al. Medical students' knowledge, attitudes, and practices toward generative artificial intelligence in Egypt 2024: a cross-sectional study. BMC Med Educ 2025 May 28;25(1):790. [doi: 10.1186/s12909-025-07329-x] [Medline: 40437443]
45. Ahmed WMM, Abdalmotalib MM, Elbadawi MH, et al. Shaping the future of medical education: a cross-sectional study on ChatGPT attitude and usage among medical students in Sudan. PLoS ONE 2025;20(5):e0324698. [doi: 10.1371/journal.pone.0324698] [Medline: 40408361]
46. Chakri I, El Khayali O, Lahlou L. Knowledge and perceptions of AI among medical students in Morocco: cross-sectional study. JMIR Form Res 2025 Sep 19;9:e66156. [doi: 10.2196/66156] [Medline: 40971792]
47. Giavina-Bianchi B, Giavina-Bianchi M, Wolosker N, Amaro Junior E, Machado BS. Attitudes, knowledge, opinions, and expectations of medical students towards medical artificial intelligence solutions: a cross-sectional survey study. Einstein (Sao Paulo) 2025;23:eAO1401. [doi: 10.31744/einstein_journal/2025AO1401] [Medline: 41124428]
48. Fußhöller A, Lechner F, Schlicker N, et al. Perceptions, usage, and educational impact of ChatGPT among medical students in Germany: cross-sectional mixed methods survey. JMIR Form Res 2025 Nov 11;9:e81484. [doi: 10.2196/81484] [Medline: 41218187]
49. Tajima H, Kasai H, Shikino K, Shimizu I, Ito S. Perceptions and intentions to use generative AI among first-year medical students in Japan: cross-sectional survey study. JMIR Med Educ 2025 Nov 13;11:e77552. [doi: 10.2196/77552] [Medline: 41232013]
50. Elhassan SE, Sajid MR, Syed AM, Fathima SA, Khan BS, Tamim H. Assessing familiarity, usage patterns, and attitudes of medical students toward ChatGPT and other chat-based AI apps in medical education: cross-sectional questionnaire study. JMIR Med Educ 2025 Jan 30;11:e63065. [doi: 10.2196/63065] [Medline: 39883912]
51. Tran C, Hryciw BN, Moore SW, Chaput A, Seely AJE. Perceptions and use of generative artificial intelligence in medical students: a multicenter survey. J Med Educ Curric Dev 2025;12:23821205251391969. [doi: 10.1177/23821205251391969] [Medline: 41181167]
52. Almehmadi B, Bakhsh TM, Noor Elahi I, Hariri N. Awareness, perceptions, and opinions of artificial intelligence among undergraduate medical students at Umm Al-Qura University, Saudi Arabia, in 2025: a cross-sectional study. Cureus 2025 Aug;17(8):e89741. [doi: 10.7759/cureus.89741] [Medline: 40932974]
53. Arab-Borzu Z, Keikha L, Shahraki-Mohammadi A. Knowledge, readiness, and perception of medical students toward medical artificial intelligence: a cross-sectional study. J Med Educ Curric Dev 2025;12:23821205251407758. [doi: 10.1177/23821205251407758] [Medline: 41425647]
54. Onetiu F, Bratu ML, Folescu R, Bratosin F, Bratu T. Assessing medical students' perceptions of AI-Integrated telemedicine: a cross-sectional study in Romania. Healthcare (Basel) 2025 Apr 24;13(9):990. [doi: 10.3390/healthcare13090990] [Medline: 40361768]
55. Alfahl S. Knowledge attitudes and ethical concerns about artificial intelligence among medical students at Taibah University: a cross-sectional study. Adv Med Educ Pract 2025;16:1609-1620. [doi: 10.2147/AMEP.S528281] [Medline: 40917814]

XSL•FO
RenderX

56. Zhang M, Liu T, Peng X, Chen Y, Zhi M. Awareness and attitudes toward artificial intelligence language generation models in medical education: a cross-sectional questionnaire study among medical students in Southern China. Cureus 2025 Aug;17(8):e89425. [doi: 10.7759/cureus.89425] [Medline: 40918762]
57. Almuhanna A, Almohsen D, AlSultan D, et al. Attitudes and awareness of medical students toward teleradiology and the application of artificial intelligence in diagnostic radiology: a cross-sectional study. J Med Educ Curric Dev 2025;12:23821205251358005. [doi: 10.1177/23821205251358005] [Medline: 40693185]
58. Abdelhafiz AS, Farghly MI, Sultan EA, Abouelmagd ME, Ashmawy Y, Elsebaie EH. Medical students and ChatGPT: analyzing attitudes, practices, and academic perceptions. BMC Med Educ 2025 Feb 5;25(1):187. [doi: 10.1186/s12909-025-06731-9] [Medline: 39910593]
59. Liu M, Cheng Y, Li S, et al. Exploring medical students' attitudes and perceptions toward artificial intelligence in medicine in Shandong Province, China. BMC Med Educ 2026;26(1):106. [doi: 10.1186/s12909-025-08465-0]
60. Duan S, Liu C, Rong T, Zhao Y, Liu B. Integrating AI in medical education: a comprehensive study of medical students' attitudes, concerns, and behavioral intentions. BMC Med Educ 2025 Apr 23;25(1):599. [doi: 10.1186/s12909-025-07177-9] [Medline: 40269824]
61. Hanifa H, Atia M, Daboul R, et al. Artificial intelligence in focus: assessing awareness and perceptions among medical students in three private Syrian universities. BMC Med Educ 2025 May 29;25(1):801. [doi: 10.1186/s12909-025-07396-0] [Medline: 40442679]
62. Fernando P, Wijayarathne J, Ranatunge R, et al. Knowledge and perception of artificial intelligence in medicine among undergraduate medical students in Sri Lanka: a cross sectional study. Stud Health Technol Inform 2025 Apr 8;323:116-120. [doi: 10.3233/SHTI250060] [Medline: 40200457]
63. Gualda-Gea JJ, Barón-Miras LE, Bertran MJ, Vilella A, Torá-Rocamora I, Prat A. Perceptions and future perspectives of medical students on the use of artificial intelligence based chatbots: an exploratory analysis. Front Med (Lausanne) 2025;12:1529305. [doi: 10.3389/fmed.2025.1529305] [Medline: 39911871]
64. Ignjatović A, Apostolović MA, Stevanović L, et al. Exploring medical students' perceptions regarding ChatGPT and AI studying at the University of Niš: a study on usage, attitudes and linguistic influence—single-centered study in Serbia—a paradoxical ally? J Med Educ Curric Dev 2025 Dec;12:23821205251409530. [doi: 10.1177/23821205251409530] [Medline: 41472885]
65. Sanri E. Beyond metropolises: artificial intelligence awareness and educational needs among medical students in a developing country. Front Med 2025;12:1645484. [doi: 10.3389/fmed.2025.1645484]
66. Shah MH, Faizan M, Akbar F, et al. Exploring medical students' perspectives on artificial intelligence in medical education: a knowledge, attitude, and practice study conducted at Jinnah Medical College, Pakistan. Cureus 2025;17(11). [doi: 10.7759/cureus.97453]
67. Gopidesi M, Anil R, Mahesh S, Nagaraj C, Pichandi J. Bridging the Artificial Intelligence (AI) gap: a Knowledge, Attitude, and Practice (KAP) study to advance medical education in rural Andhra Pradesh. Cureus 2025 Sep;17(9):e91822. [doi: 10.7759/cureus.91822] [Medline: 41069909]
68. Ringeval M, Raymond L, Pomey MP, Paré G. Evolving medical students' digital health perceptions and intentions: insights from a prepandemic and postpandemic survey study. J Med Internet Res 2025 Sep 3;27:e64804. [doi: 10.2196/64804] [Medline: 40902151]
69. Ejas F, Khan SA, Mujahid A, et al. Medical students' perceptions of large language models in healthcare: a multinational cross-sectional study. J Med Educ Curric Dev 2025;12:23821205251331124. [doi: 10.1177/23821205251331124] [Medline: 40672122]
70. Mousavi Baigi SF, Sarbaz M, Darroudi A, Kimiafar K. Medical and paramedical students' perspectives on artificial intelligence in a low- and middle-income country: a cross-sectional study. Health Sci Rep 2025 Aug;8(8):e71046. [doi: 10.1002/hsr2.71046] [Medline: 40761605]
71. Falcon RMG, Alcazar RMU, Babaran HG. Exploring Filipino medical students' attitudes and perceptions of artificial intelligence in medical education: a mixed-methods study. MedEdPublish 2025;14:282 [FREE Full text] [doi: 10.12688/mep.20590.2]
72. Rezazadeh H, Mahani AM, Salajegheh M. Insights into the future: assessing medical students' artificial intelligence readiness—a cross-sectional study at Kerman University of Medical Sciences (2022). Health Sci Rep 2025 May;8(5):e70870. [doi: 10.1002/hsr2.70870] [Medline: 40432697]
73. Sunmboye K, Strafford H, Noorestani S, Wilison-Pirie M. Exploring the influence of artificial intelligence integration on personalized learning: a cross-sectional study of undergraduate medical students in the United Kingdom. BMC Med Educ 2025 Apr 17;25(1):570. [doi: 10.1186/s12909-025-07084-z] [Medline: 40247306]
74. AlZaabi A, Masters K. Assessing medical students' readiness for artificial intelligence after pre-clinical training. BMC Med Educ 2025 Jun 2;25(1):824. [doi: 10.1186/s12909-025-07008-x] [Medline: 40457325]
75. Ziapour A, Darabi F, Janjani P, Amani MA, Yıldırım M, Motevaseli S. Factors affecting medical artificial intelligence (AI) readiness among medical students: taking stock and looking forward. BMC Med Educ 2025 Feb 18;25(1):264. [doi: 10.1186/s12909-025-06852-1] [Medline: 39966878]

XSL•FO
RenderX

76. Alharkan KS, Almukahhil AM, Alfardan HR, et al. Medical students' beliefs, behaviors, and ease with artificial intelligence in education: a study in Eastern Province, Saudi Arabia. J Family Community Med 2025;32(4):324-330. [doi: 10.4103/jfcm.jfcm_187_25] [Medline: 41234270]
77. Naseer MA, Saeed S, Afzal A, Ali S, Malik MGR. Navigating the integration of artificial intelligence in the medical education curriculum: a mixed-methods study exploring the perspectives of medical students and faculty in Pakistan. BMC Med Educ 2025 Feb 20;25(1):273. [doi: 10.1186/s12909-024-06552-2] [Medline: 39979912]
78. González-Bravo AE, Battle C, Fadhel-Hernández VS, et al. Understanding generative artificial intelligence adoption in Puerto Rican medical schools: a cross-institutional survey of first- and second-year students. J Med Educ Curric Dev 2025;12:23821205251398923. [doi: 10.1177/23821205251398923] [Medline: 41362787]
79. Almarzouki AF, Alem A, Shrourou F, et al. Assessing the disconnect between student interest and education in artificial intelligence in medicine in Saudi Arabia. BMC Med Educ 2025 Jan 30;25(1):150. [doi: 10.1186/s12909-024-06446-3] [Medline: 39881303]
80. Dhurandhar D, Dhamande M, C S, Bhadoria P, Chandrakar T, Agrawal J. Exploring medical artificial intelligence readiness among future physicians: insights from a medical college in Central India. Cureus 2025 Jan;17(1):e76835. [doi: 10.7759/cureus.76835] [Medline: 39897272]
81. Allam AH, Eltewacy NK, Alabdallat YJ, et al. Knowledge, attitude, and perception of Arab medical students towards artificial intelligence in medicine and radiology: a multi-national cross-sectional study. Eur Radiol 2024 Jul;34(7):1-14. [doi: 10.1007/s00330-023-10509-2] [Medline: 38150076]
82. Jebreen K, Radwan E, Kammoun-Rebai W, et al. Perceptions of undergraduate medical students on artificial intelligence in medicine: mixed-methods survey study from Palestine. BMC Med Educ 2024 May 7;24(1):507. [doi: 10.1186/s12909-024-05465-4] [Medline: 38714993]
83. Allam RM, Abdelfatah D, Khalil MIM, Elsaieed MM, El Desouky ED. Medical students and house officers' perception, attitude and potential barriers towards artificial intelligence in Egypt, cross sectional survey. BMC Med Educ 2024 Oct 31;24(1):1244. [doi: 10.1186/s12909-024-06201-8] [Medline: 39482613]
84. Tao W, Yang J, Qu X. Utilization of, perceptions on, and intention to use AI chatbots among medical students in China: national cross-sectional study. JMIR Med Educ 2024 Oct 28;10:e57132. [doi: 10.2196/57132] [Medline: 39466038]
85. Ganjavi C, Eppler M, O'Brien D, et al. ChatGPT and large language models (LLMs) awareness and use. a prospective cross-sectional survey of U.S. medical students. PLOS Digit Health 2024 Sep;3(9):e0000596. [doi: 10.1371/journal.pdig.0000596] [Medline: 39236008]
86. Alwadani FAS, Lone A, Hakami MT, et al. Attitude and understanding of artificial intelligence among Saudi medical students: an online cross-sectional study. J Multidiscip Healthc 2024;17:1887-1899. [doi: 10.2147/JMDH.S455260] [Medline: 38706506]
87. Abu Hammour A, Hammour KA, Alhamad H, et al. Exploring Jordanian medical students' perceptions and concerns about ChatGPT in medical education: a cross-sectional study. J Pharm Policy Pract 2024;17(1):2429000. [doi: 10.1080/20523211.2024.2429000] [Medline: 39600801]
88. Faroog Z, Dirar QSE, Zaidi ARZ, et al. Knowledge and attitude of medical students towards artificial intelligence in ophthalmology in Riyadh, Saudi Arabia: a cross-sectional study. Ann Med Surg (Lond) 2024 Aug;86(8):4377-4383. [doi: 10.1097/MS9.0000000000002238] [Medline: 39118699]
89. Laupichler MC, Aster A, Meyerheim M, Raupach T, Mergen M. Medical students' AI literacy and attitudes towards AI: a cross-sectional two-center study using pre-validated assessment instruments. BMC Med Educ 2024 Apr 10;24(1):401. [doi: 10.1186/s12909-024-05400-7] [Medline: 38600457]
90. Jaber Amin MH, Mohamed Elhassan Elmahi MA, Abdelmonim GA, et al. Knowledge, attitude, and practice of artificial intelligence among medical students in Sudan: A cross-sectional study. Ann Med Surg 2024;86(7):3917-3923. [doi: 10.1097/MS9.0000000000002070]
91. Jackson P, Ponath Sukumaran G, Babu C, et al. Artificial intelligence in medical education—perception among medical students. BMC Med Educ 2024 Jul 27;24(1):804. [doi: 10.1186/s12909-024-05760-0] [Medline: 39068482]
92. Angkurawaranon S, Inmutto N, Bannangkoon K, et al. Attitudes and perceptions of Thai medical students regarding artificial intelligence in radiology and medicine. BMC Med Educ 2024 Oct 22;24(1):1188. [doi: 10.1186/s12909-024-06150-2] [Medline: 39438874]
93. Lugito NPH, Cucunawangsih C, Suryadinata N, et al. Readiness, knowledge, and perception towards artificial intelligence of medical students at faculty of medicine, Pelita Harapan University, Indonesia: a cross sectional study. BMC Med Educ 2024 Sep 27;24(1):1044. [doi: 10.1186/s12909-024-06058-x] [Medline: 39334022]
94. Daher OA, Dabbousi AA, Chamroukh R, Saab AY, Al Ayoubi AR, Salameh P. Artificial intelligence: knowledge and attitude among Lebanese medical students. Cureus 2024 Jan;16(1):e51466. [doi: 10.7759/cureus.51466] [Medline: 38298326]
95. Weidener L, Fischer M. Artificial intelligence in medicine: cross-sectional study among medical students on application, education, and ethical aspects. JMIR Med Educ 2024 Jan 5;10:e51247. [doi: 10.2196/51247] [Medline: 38180787]
96. Luong J, Tzang CC, McWatt S, et al. Exploring artificial intelligence readiness in medical students: analysis of a global survey. Med Sci Educ 2025 Feb;35(1):331-341. [doi: 10.1007/s40670-024-02190-x] [Medline: 40144085]

XSL•FO
RenderX

97. Al Shahrani A, Alhumaidan N, AlHindawi Z, et al. Readiness to embrace artificial intelligence among medical students in Saudi Arabia: a national survey. Healthcare (Basel) 2024 Dec 11;12(24):2504. [doi: 10.3390/healthcare12242504] [Medline: 39765931]
98. Malešević A, Kolesárová M, Čartolovni A. Encompassing trust in medical AI from the perspective of medical students: a quantitative comparative study. BMC Med Ethics 2024 Sep 2;25(1):94. [doi: 10.1186/s12910-024-01092-2] [Medline: 39223538]
99. Hamad M, Qtaishat F, Mhairat E, et al. Artificial intelligence readiness among Jordanian medical students: using medical artificial intelligence readiness scale for medical students (MAIRS-MS). J Med Educ Curric Dev 2024;11:23821205241281648. [doi: 10.1177/23821205241281648] [Medline: 39346121]
100. Alammari DM, Melebari RE, Alshaikh JA, Alotaibi LB, Basabeen HS, Saleh AF. Beyond boundaries: the role of artificial intelligence in shaping the future careers of medical students in Saudi Arabia. Cureus 2024 Sep;16(9):e69332. [doi: 10.7759/cureus.69332] [Medline: 39398766]
101. Gandhi R, Parmar A, Kagathara J, Lakkad D, Kakadiya J, Murugan Y. Bridging the artificial intelligence (AI) divide: do postgraduate medical students outshine undergraduate medical students in AI readiness? Cureus 2024 Aug;16(8):e67288. [doi: 10.7759/cureus.67288] [Medline: 39301347]
102. Özbek Güven G, Yilmaz Ş, Inceoğlu F. Determining medical students' anxiety and readiness levels about artificial intelligence. Heliyon 2024 Feb 29;10(4):e25894. [doi: 10.1016/j.heliyon.2024.e25894] [Medline: 38384508]
103. Xu AY, Piranio VS, Speakman S, et al. A pilot study of medical student opinions on large language models. Cureus 2024 Oct;16(10):e71946. [doi: 10.7759/cureus.71946] [Medline: 39564056]
104. Sorte SR, Rawekar A, Rathod SB. Understanding AI in healthcare: perspectives of future healthcare professionals. Cureus 2024 Aug;16(8):e66285. [doi: 10.7759/cureus.66285] [Medline: 39238760]
105. Al-Roomi K, Alzayani S, Almarabheh A, et al. Familiarity and applications of artificial intelligence in health professions education: perspectives of students in a community-oriented medical. Cureus 2024 Nov;16(11):e73425. [doi: 10.7759/cureus.73425] [Medline: 39664146]
106. Liu F, Chang X, Zhu Q, Huang Y, Li Y, Wang H. Assessing clinical medicine students' acceptance of large language model: based on technology acceptance model. BMC Med Educ 2024 Nov 3;24(1):1251. [doi: 10.1186/s12909-024-06232-1] [Medline: 39490999]
107. Alkhaaldi SMI, Kassab CH, Dimassi Z, et al. Medical student experiences and perceptions of ChatGPT and artificial intelligence: cross-sectional study. JMIR Med Educ 2023 Dec 22;9:e51302. [doi: 10.2196/51302] [Medline: 38133911]
108. Tangadulrat P, Sono S, Tangtrakulwanich B. Using ChatGPT for clinical practice and medical education: cross-sectional survey of medical students' and physicians' perceptions. JMIR Med Educ 2023 Dec 22;9:e50658. [doi: 10.2196/50658] [Medline: 38133908]
109. Oluwadiya KS, Adeoti AO, Agodirin SO, et al. Exploring artificial intelligence in the Nigerian medical educational space: an online cross-sectional study of perceptions, risks and benefits among students and lecturers from ten universities. Niger Postgrad Med J 2023;30(4):285-292. [doi: 10.4103/npmj.npmj_186_23] [Medline: 38037784]
110. Tung AYZ, Dong LW. Malaysian medical students' attitudes and readiness toward AI (Artificial Intelligence): a cross-sectional study. J Med Educ Curric Dev 2023;10:23821205231201164. [doi: 10.1177/23821205231201164] [Medline: 37719325]
111. Stewart J, Lu J, Gahungu N, et al. Western Australian medical students' attitudes towards artificial intelligence in healthcare. PLoS One 2023;18(8):e0290642. [doi: 10.1371/journal.pone.0290642] [Medline: 37651380]
112. Al Hadithy ZA, Al Lawati A, Al-Zadjali R, Al Sinawi H. Knowledge, attitudes, and perceptions of artificial intelligence in healthcare among medical students at Sultan Qaboos University. Cureus 2023 Sep;15(9):e44887. [doi: 10.7759/cureus.44887] [Medline: 37814766]
113. Buabbas AJ, Miskin B, Alnaqi AA, et al. Investigating students' perceptions towards artificial intelligence in medical education. Healthcare (Basel) 2023 May 1;11(9):1298. [doi: 10.3390/healthcare11091298] [Medline: 37174840]
114. Barreiro-Ares A, Morales-Santiago A, Sendra-Portero F, Souto-Bayarri M. Impact of the rise of artificial intelligence in radiology: what do students think? Int J Environ Res Public Health 2023 Jan 16;20(2):1589. [doi: 10.3390/ijerph20021589] [Medline: 36674348]
115. Biri SK, Kumar S, Panigrahi M, Mondal S, Behera JK, Mondal H. Assessing the utilization of large language models in medical education: insights from undergraduate medical students. Cureus 2023 Oct;15(10):e47468. [doi: 10.7759/cureus.47468] [Medline: 38021810]
116. Doumat G, Daher D, Ghanem NN, Khater B. Knowledge and attitudes of medical students in Lebanon toward artificial intelligence: a national survey study. Front Artif Intell 2022;5:1015418. [doi: 10.3389/frai.2022.1015418] [Medline: 36406470]
117. Ahmed Z, Bhinder KK, Tariq A, et al. Knowledge, attitude, and practice of artificial intelligence among doctors and medical students in Pakistan: a cross-sectional online survey. Ann Med Surg 2022;76:103493. [doi: 10.1016/j.amsu.2022.103493]
118. Liu DS, Sawyer J, Luna A, et al. Perceptions of US medical students on artificial intelligence in medicine: mixed methods survey study. JMIR Med Educ 2022 Oct 21;8(4):e38325. [doi: 10.2196/38325] [Medline: 36269641]

XSL•FO
RenderX

119. Caparrós Galán G, Sendra Portero F. Medical students' perceptions of the impact of artificial intelligence in radiology. Radiologia (Engl Ed) 2022;64(6):516-524. [doi: 10.1016/j.rxeng.2021.03.008] [Medline: 36402537]
120. Gillissen A, Kochanek T, Zupanic M, Ehlers J. Medical students' perceptions towards digitization and artificial intelligence: a mixed-methods study. Health Care (Don Mills) 2022;10(4):723. [doi: 10.3390/healthcare10040723]
121. Pucchio A, Rathagirishnan R, Caton N, et al. Exploration of exposure to artificial intelligence in undergraduate medical education: a Canadian cross-sectional mixed-methods study. BMC Med Educ 2022 Nov 28;22(1):815. [doi: 10.1186/s12909-022-03896-5] [Medline: 36443720]
122. McLennan S, Meyer A, Schreyer K, Buyx A. German medical students´ views regarding artificial intelligence in medicine: a cross-sectional survey. PLoS Digit Health 2022 Oct;1(10):e0000114. [doi: 10.1371/journal.pdig.0000114] [Medline: 36812635]
123. Blease C, Kharko A, Bernstein M, et al. Machine learning in medical education: a survey of the experiences and opinions of medical students in Ireland. BMJ Health Care Inform 2022 Feb;29(1):e100480. [doi: 10.1136/bmjhci-2021-100480] [Medline: 35105606]
124. Boillat T, Nawaz FA, Rivas H. Readiness to embrace artificial intelligence among medical doctors and students: questionnaire-based study. JMIR Med Educ 2022 Apr 12;8(2):e34973. [doi: 10.2196/34973] [Medline: 35412463]
125. Blacketer C, Parnis R, B Franke K, et al. Medical student knowledge and critical appraisal of machine learning: a multicentre international cross-sectional study. Intern Med J 2021 Sep;51(9):1539-1542. [doi: 10.1111/imj.15479] [Medline: 34541769]
126. Bisdas S, Topriceanu CC, Zakrzewska Z, et al. Artificial intelligence in medicine: a multinational multi-center survey on the medical and dental students' perception. Front Public Health 2021;9:795284. [doi: 10.3389/fpubh.2021.795284] [Medline: 35004598]
127. Wood EA, Ange BL, Miller DD. Are we ready to integrate artificial intelligence literacy into medical school curriculum: students and faculty survey. J Med Educ Curric Dev 2021(8). [doi: 10.1177/23821205211024078] [Medline: 34250242]
128. Pinto Dos Santos D, Giese D, Brodehl S, et al. Medical students' attitude towards artificial intelligence: a multicentre survey. Eur Radiol 2019 Apr;29(4):1640-1646. [doi: 10.1007/s00330-018-5601-1] [Medline: 29980928]
129. Egger M, Davey Smith G, Schneider M, Minder C. Bias in meta-analysis detected by a simple, graphical test. BMJ 1997 Sep 13;315(7109):629-634. [doi: 10.1136/bmj.315.7109.629] [Medline: 9310563]
130. Sterne JAC, Sutton AJ, Ioannidis JPA, et al. Recommendations for examining and interpreting funnel plot asymmetry in meta-analyses of randomised controlled trials. BMJ 2011 Jul 22;343:d4002. [doi: 10.1136/bmj.d4002] [Medline: 21784880]
131. Civaner MM, Uncu Y, Bulut F, Chalil EG, Tatli A. Artificial intelligence in medical education: a cross-sectional needs assessment. BMC Med Educ 2022 Nov 9;22(1):772. [doi: 10.1186/s12909-022-03852-3] [Medline: 36352431]
132. McCoy LG, Nagaraj S, Morgado F, Harish V, Das S, Celi LA. What do medical students actually need to know about artificial intelligence? NPJ Digit Med 2020;3:86. [doi: 10.1038/s41746-020-0294-7] [Medline: 32577533]
133. Li X, Yan X, Lai H. The ethical challenges in the integration of artificial intelligence and large language models in medical education: a scoping review. PLoS ONE 2025;20(10):e0333411. [doi: 10.1371/journal.pone.0333411]
134. Dijkstra P, Greenhalgh T, Mekki YM, Morley J. How to read a paper involving artificial intelligence (AI). BMJ Med 2025 Apr;4(1):e001394. [doi: 10.1136/bmjmed-2025-001394]
135. Izquierdo-Condoy JS, Arias-Intriago M, Montero Corrales L, Ortiz-Prado E. Artificial Intelligence in medical education: transformative potential, current applications, and future implications. JMIR Med Educ 2026 Feb 17;12:e77127. [doi: 10.2196/77127] [Medline: 41701936]
136. Komasawa N, Yokohira M. Generative artificial intelligence (AI) in medical education: a narrative review of the challenges and possibilities for future professionalism. Cureus 2025 Jun;17(6):e86316. [doi: 10.7759/cureus.86316] [Medline: 40688936]
137. Paranjape K, Schinkel M, Nannan Panday R, Car J, Nanayakkara P. Introducing artificial intelligence training in medical education. JMIR Med Educ 2019 Dec 3;5(2):e16048. [doi: 10.2196/16048] [Medline: 31793895]
138. Lee J, Wu AS, Li D, Kulasegaram KM. Artificial intelligence in undergraduate medical education: a scoping review. Acad Med 2021 Nov 1;96(11S):S62-S70. [doi: 10.1097/ACM.0000000000004291] [Medline: 34348374]
139. Russell RG, Lovett Novak L, Patel M, et al. Competencies for the use of artificial intelligence-based tools by health care professionals. Acad Med 2023 Mar 1;98(3):348-356. [doi: 10.1097/ACM.0000000000004963] [Medline: 36731054]
140. Kolachalama VB, Garg PS. Machine learning and medical education. NPJ Digit Med 2018;1:54. [doi: 10.1038/s41746-018-0061-1] [Medline: 31304333]
141. Gordon M, Daniel M, Ajiboye A, et al. A scoping review of artificial intelligence in medical education: BEME Guide No. 84. Med Teach 2024 Apr 2;46(4):446-470. [doi: 10.1080/0142159X.2024.2314198]
142. IntHout J, Ioannidis JPA, Rovers MM, Goeman JJ. Plea for routinely presenting prediction intervals in meta-analysis. BMJ Open 2016 Jul 12;6(7):e010247. [doi: 10.1136/bmjopen-2015-010247] [Medline: 27406637]
143. Sun X, Briel M, Walter SD, Guyatt GH. Is a subgroup effect believable? Updating criteria to evaluate the credibility of subgroup analyses. BMJ 2010 Mar 30;340:c117. [doi: 10.1136/bmj.c117] [Medline: 20354011]
144. Masters K. Ethical use of artificial intelligence in health professions education: AMEE Guide No. 158. Med Teach 2023 Jun;45(6):574-584. [doi: 10.1080/0142159X.2023.2186203] [Medline: 36912253]
145. Ethics and governance of artificial intelligence for health: WHO guidance. World Health Organization. 2021. URL: https://www.who.int/publications/i/item/9789240029200 [accessed 2026-08-05]

XSL•FO
RenderX

Abbreviations

GRADE: Grading of Recommendations, Assessment, Development and Evaluation
HKSJ: Hartung-Knapp-Sidik-Jonkman
JBI: Joanna Briggs Institute
LLM: large language model
MAIRS-MS: Medical AI Readiness Scale for Medical Students
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
PRISMA-S: Preferred Reporting Items for Systematic Reviews and Meta-Analyses literature search extension
WHO: World Health Organization

Edited by S Brini; submitted 12.Dec.2025; peer-reviewed by D Winterton, E Ortiz-Prado; revised version received 19.Jul.2026; accepted 20.Jul.2026; published 18.Aug.2026.

Please cite as:
Zhang W, Han J, Han X, Xu H, Wang L, Lin T, Tan P, Zhang P, Zheng X
Medical Students' Attitudes, Perceptions, and Self-Reported Familiarity With AI in Health Care: Systematic Review and Meta-Analysis
JMIR Med Educ 2026;12:e89411
URL: https://mededu.jmir.org/2026/1/e89411
doi:10.2196/89411

© Wei Zhang, Jiaxue Han, Xin Han, Hang Xu, Langkun Wang, Tianhai Lin, Ping Tan, Peng Zhang, Xiaonan Zheng. Originally published in JMIR Medical Education (https://mededu.jmir.org), 18.Aug.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Education, is properly cited. The complete bibliographic information, a link to the original publication on https://mededu.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Digital Standardized Patients: Conceptual Framework for Generative AI–Empowered Medical Education

Chuanlin Jia, BA, MFA; Lulu Qi, BMed, LLM

Publicity Department / Teacher Work Department, Zhongshan Hospital, Fudan University, No. 180, Fenglin Road, Xuhui District, Shanghai, Shanghai, China

Corresponding Author:
Lulu Qi, BMed, LLM
Publicity Department / Teacher Work Department, Zhongshan Hospital, Fudan University, No. 180, Fenglin Road, Xuhui District, Shanghai, Shanghai, China

Abstract

Generative AI is driving medical education from digital support toward intelligent, interactive learning environments. The cultivation of doctor-patient communication skills requires not only technical proficiency but also communication competence, emotional sensitivity, and ethical judgment. This paper proposes a conceptual framework for AI-driven digital standardized patients (AI-SPs) to provide new approaches for communication training, humanistic education, and emotional engagement in medical curricula. This viewpoint integrates research findings from educational technology and medical education, elaborating the framework from 4 dimensions: system architecture, multimodal interaction, personality modeling, and ethical considerations. AI-SPs can provide adaptive, emotionally responsive interactions in repeatable, controllable simulated scenarios, enabling learners to experience diverse patient characteristics and clinical situations. The proposed "future learning" framework emphasizes personalization, contextualization, and reflective learning. During implementation, attention must be paid to data privacy, algorithmic bias, and human supervision. AI-SPs represent an extension of the traditional standardized patient model. They facilitate human-AI collaborative learning, support the cultivation of empathy, and provide a new pathway for the appropriate application of generative AI in medical education.

(JMIR Med Educ 2026;12:e91050) doi:10.2196/91050

KEYWORDS

generative AI; virtual standardized patient; conversational AI; medical education; clinical communication simulation

Introduction

From multimedia instruction to online learning and then to virtual simulation, early educational technology innovations primarily expanded the coverage and efficiency of education, with technology itself positioned as an auxiliary tool. The emergence of generative AI has brought a qualitative change; it is not only a supporter of learning interactions but also an active participant [1].

AI is becoming a coconstructor of the learning experience. By responding to learners' language, tone, and emotional cues, generative AI introduces the possibility of adaptive interaction. This is particularly crucial for medical education because the effectiveness of medical learning depends not only on cognitive accuracy but also on interpersonal sensitivity.

Medical education faces a long-standing challenge: how to ensure that students acquire high-quality clinical communication skills and humanistic care competencies while training physicians at scale? Communication training requires repeated practice, immediate feedback, and a safe environment for trial and error, areas in which traditional teaching methods have significant shortcomings. China's medical education system faces this challenge with particular intensity. According to 2025 data from the National Center for Medical Education Development, China has 506 higher education institutions offering medical programs, including 112 independent medical colleges and schools, with a total of hundreds of thousands of medical students [2]. By geographic distribution, the eastern region has 240 institutions (47.4%), the central region 144 (28.5%), and the western region 122 (24.1%), presenting a pattern of "strong east, weak west" [2]. This uneven distribution of resources poses special challenges for the large-scale, balanced allocation of teaching resources such as standardized patients.

This paper proposes a conceptual framework for AI-driven digital standardized patients (AI-SPs) as an extension of existing simulation-based teaching to address the above challenges. The article first reviews the educational roles and limitations of traditional standardized patients (SPs), then elaborates on the core features and system architecture of AI-SPs, further proposes an integrated "future learning" framework, and finally discusses key ethical issues.

Regarding the technical pathway for developing AI-SP systems, current research has mainly followed 2 approaches. The first is

XSL•FO
RenderX

a prompt engineering–based dynamic generation approach, in which carefully designed system prompts guide large language models (LLMs) to adopt specific patient roles and generate contextually appropriate responses. This approach provides advantages in role diversity and scenario adaptability, although maintaining role consistency remains challenging. The second is a fine-tuning–based standardization approach, which uses large-scale doctor-patient dialogue datasets to train models to reproduce specific communication patterns. While this approach improves output stability, it requires extensive, high-quality training data and may be less flexible in representing diverse patient personality types. Based on our experience in developing an AI-SP platform with multiple digital patient personas for medical student training, this framework adopts prompt engineering as the primary technical approach because it aligns with the educational objective of providing diverse and customizable patient roles for communication training.

Background and Concept Analysis

Value and Limitations of Traditional SPs

SPs were first introduced by Barrows [3] in 1963, marking an important shift in medical education toward experiential, learner-centered approaches. SPs enable learners to practice diagnostic reasoning, communication strategies, and professional behaviors in a safe, controlled environment [4].

SPs have become a core tool for clinical skills training and assessment, including objective structured clinical examinations (OSCEs). Their value lies in providing a controlled training environment, supporting reflective feedback, and cultivating a patient-centered perspective. However, SP-based teaching faces persistent challenges: high operational costs, limited scalability, significant variability in performance, and pronounced emotional fatigue, particularly in large-scale or resource-constrained educational settings.

These limitations are especially evident in communication-intensive scenarios such as preoperative conversations, breaking bad news, and managing patients who are emotionally distressed, precisely the scenarios that require the highest degree of consistency and repeatability in teaching.

Comparison Between AI-SP and Related Concepts

To clearly define the concept proposed in this paper, it is necessary to distinguish it from existing related concepts. Table 1 compares traditional SPs, rule-based virtual patients, general-purpose chatbots, and the AI-SP proposed in this paper.

Table . Comparison of different standardized patient (SP)–related concepts[a].

Feature dimension	Traditional SP (human performer)	Rule-based virtual patient	General chatbot	Proposed AI-SP[b]
Interaction modality	Face-to-face with real person	Menu selection and scripted dialogue	Free-text dialogue	Text or voice switching+virtual avatar
Response generation	Script-based	Preprogrammed rules	General generation	Context-sensitive generation+affective computing
Personality consistency	Depends on actor performance	None or fixed	Unstable	Based on Big Five personality model and Satir communication stance theory
Emotional dynamics	Simulatable but difficult to standardize	None	None or superficial	Emotional state can be influenced by student behavior and evolve in real time
Scalability	Low (high training cost and long cycle)	High	High	High (parameterized customization)
Repeatability	Moderate (affected by actor condition)	High	Low	High (controllable and recordable)
Real-time feedback	Depends on post hoc comments from teachers or actors	None or limited	None	Can generate formative assessment reports based on communication frameworks
Applicable scenarios	OSCE[c] and small-class teaching	History-taking training	General Q&A[d]	Communication training, humanistic education, and emotionally intense scenarios

[a]Rule-based virtual patients refer to early digital patient simulators based on predefined scripts.

[b]AI-SP: AI-driven digital standardized patient.

[c]OSCE: objective structured clinical examination.

[d]Q&A: question and answer.

The comparison in Table 1 shows that the AI-SP is a pedagogical agent with stable personality, affective responsiveness, multimodal interaction support, and parameterized customization, rather than a simple “chatbot playing the role of a patient” [5].

XSL•FO
RenderX

Specificity of the Chinese Medical Education Context

This paper focuses on the Chinese medical education context based on 3 main considerations.

Imbalance in Scale and Resources

China's 506 medical education institutions are distributed across regions with significantly different resources. Establishing and maintaining professional SP programs requires sustained financial investment and systematic training, which poses substantial barriers for institutions in central and western China. As a scalable digital solution, the AI-SP has the potential to partially alleviate this resource imbalance, provided that institutions have basic information technology infrastructure for deployment.

Specificity of Cultural and Communication Patterns

Cross-cultural research has revealed significant differences in doctor-patient communication patterns across cultural contexts. A comparative study by Liu et al [6] involving 500 patients from China and 500 from the United Kingdom found that Chinese patients reported significantly lower scores on communication quality, decision-making participation, and emotional expression compared to British patients. The study further identified power distance and collectivist orientation as key predictors of these differences: higher power distance was associated with lower communication quality, reduced patient participation, and more restrained emotional expression. This means that when designing AI-SPs for the Chinese context, Western models cannot simply be replicated; rather, they must be adapted to local communication norms and patient expectations.

In addition, Chinese medical humanities education has recently begun to systematically reflect on indigenous traditions. The 2025 "Life Has Pain, Apricot Has You" Chinese medical humanities expert consensus [7] explicitly states that Chinese medical humanities education has long drawn primarily from Western concepts, with insufficient transmission of the traditional Chinese medical humanistic spirit. Rooted in the historical allusion to Dong Feng's "apricot forest" from the Eastern Han Dynasty, this consensus articulates a medical humanistic spirit grounded in traditional Chinese culture that also meets contemporary needs, emphasizing physician-patient empathy as a bond. This consensus provides a cultural reference for AI-SP design: it should not only merely simulate "what the patient says" but also embody the traditional expectation of physician-patient mutual trust in the Chinese context.

Support From the Policy Environment

China has issued multiple national policies encouraging the integrated development of "AI+education" and "AI+healthcare": the New Generation Artificial Intelligence Development Plan (State Council Document [2017] number 35) lists intelligent education as a key area; the 14th Five-Year Plan for Digital Economy Development promotes the construction of new educational infrastructure; in 2025, the National Health Commission and 4 other departments jointly issued the Implementation Opinions on Promoting and Regulating the Application of "AI+Healthcare," identifying 24 key applications including research and teaching; and in the same year, the State Council issued the Opinions on Deeply Implementing the "AI+" Action, promoting the exploration of AI applications in health care scenarios. These policies provide macroenvironmental support for the research, development, and application of AI-SPs.

In summary, the Chinese context is important because it simultaneously presents large-scale demand, unique cultural communication patterns, and a clear national policy direction, 3 factors that together constitute the practical foundation for AI-SP development.

Conceptual Framework of AI-SP

Definition and Core Features

This paper defines AI-SP as a digital interactive agent driven by generative AI, possessing stable personality traits and affective responsiveness, capable of engaging in multimodal, context-sensitive natural conversations with medical learners in repeatable, controllable simulated environments, and supporting parameterized configuration of clinical scenarios. Unlike general-purpose chatbots or rule-based virtual patients, it has the following five core features:

1. Multimodal interaction: it supports real-time switching between text and voice dialogue, with a virtual avatar. This design brings interactions closer to real-world doctor-patient communication scenarios, allowing learners to choose interaction methods that suit their learning preferences.
2. Personality consistency: based on the Big Five personality model and Satir communication stance theory, AI-SP maintains stable personality traits and emotional response patterns across multiple independent interactions (eg, persistent anxiety triggered by high neuroticism and defensive blame caused by low agreeableness). This mechanism effectively avoids the role drift and logical conflicts common in generative AI, ensuring that the virtual patient maintains a unified cognitive framework and emotional tone during extended dialogues [8], thereby constructing a highly realistic teaching environment with stable feedback loops.
3. Affective dynamic responsiveness: AI-SP's emotional state is not a fixed script but can be influenced by learners' communication behaviors (eg, tone, wording, and empathic expression) and evolve in real time. For example, when a learner asks questions in a blunt or empathy-deficient manner, AI-SP may exhibit anxiety or distrust; conversely, appropriate empathic expression may alleviate its negative emotions.
4. Parameterized customization: teachers can quickly generate diverse patient cases by adjusting a series of parameters, including patient demographic characteristics (age, gender, occupation, and education level), emotional states (ranging from calm to anxious, angry, and sad, etc), personality traits (scores on the Big Five dimensions), clinical scenarios (disease type, severity, and whether they involve bad-news disclosure), and social factors (family support, financial burden, and so on).
5. Real-time feedback and reflective support: after the interaction, the system can generate formative assessment

XSL•FO
RenderX

reports based on communication skills frameworks (eg, the SEGUE [Set the Stage, Elicit Information, Give Information, Understand the Patient's Perspective, and End the Encounter] Scale and the SPIKES [setting up the interview, assessing the patient's perception, obtaining the patient's invitation, giving knowledge and information, addressing empathy and emotions, strategy and summary] bad-news disclosure model), identifying the learner's strengths and weaknesses in communication and suggesting directions for improvement.

Technical Implementation Mechanisms Underlying the Educational Functions of AI-SP

This section further describes the specific technical mechanisms underlying the educational functions of AI-SPs to clarify the operational feasibility of the proposed framework.

The current AI-SP system primarily adopts a prompt engineering–based approach, supplemented by retrieval-augmented generation (RAG) and agent-based architectures, although model fine-tuning has not yet been implemented. Table 2 summarizes the correspondence between key educational functions of AI-SPs and their underlying technical implementations.

Table . Mapping of AI-driven digital standardized patients (AI-SPs) educational functions to implementation approaches and underlying technical mechanisms[a].

Educational function	Implementation in the platform	Underlying technical mechanism
Personalized patient personas (eg, normal, anxious, challenging, and angry)	Before training, users select a patient type, and the system invokes the corresponding persona-specific prompt template.	Prompt engineering+intrinsic LLM[b] capabilities: persona profiles are constructed for different patient types, with emotional characteristics, communication preferences, and behavioral rules defined within system prompts. The contextual understanding capability of LLMs is used to generate continuous dialogues that are consistent with predefined personality characteristics.
Language style and expression patterns (eg, colloquial language and differences in educational background)	Controlled through prompt constraints.	Prompt engineering: personality characteristics, educational background, and other contextual factors are incorporated as prompt parameters to constrain language style and the use of medical terminology.
Dynamic emotional responsiveness (eg, patients becoming less distressed after receiving empathic responses from learners)	Supported through LLM-based contextual understanding combined with prompt constraints.	Prompt engineering+intrinsic LLM capabilities: emotional transition rules for different patient personas are predefined within system prompts. Through multiturn dialogue processing and contextual reasoning, LLMs maintain consistency in emotional evolution and response logic throughout interactions.
Generation and customization of clinical scenarios (eg, adjustment of patient age, disease condition, and social background by educators)	Prompt templates contain configurable variables. Users input clinical departments, diagnoses, and intended communication scenarios, which are dynamically assembled by the backend into complete persona prompts.	Prompt engineering: structured user-defined parameters are dynamically injected into system prompts. LLMs infer patient illness perceptions, major concerns, and communication goals based on these parameters. Demographic and cultural factors (eg, age and educational background) can be incorporated as expandable prompt variables.
Real-time feedback and scoring (eg, SEGUE[c]-based communication assessment after interaction)	A separate AI agent is used to analyze the completed dialogue.	RAG[d]+AI agent architecture: SEGUE assessment criteria are structured into a knowledge base. The agent retrieves relevant assessment criteria through RAG during evaluation and applies LLM-based semantic analysis to the complete dialogue, enabling item-level evaluation, evidence identification, and score generation.

[a]The current system primarily adopts prompt engineering rather than model fine-tuning for several reasons. First, the system is designed for dynamic role-playing rather than standardized dialogue generation, making prompt engineering a more suitable approach at the current stage. Second, fine-tuning requires large-scale standardized doctor-patient dialogue datasets and substantial development resources, which remain challenging during early-stage implementation. With the accumulation of sufficient interaction data in the future, fine-tuning may serve as a potential approach to further enhance the realism and fidelity of patient personas.

[b]LLM: large language model.

[c]SEGUE: Set the Stage, Elicit Information, Give Information, Understand the Patient's Perspective, and End the Encounter.

[d]RAG: retrieval-augmented generation.

XSL•FO
RenderX

System Architecture

This paper aims to propose a conceptual framework rather than complete technical specifications, so only the key modules and their functions are described. The system architecture comprises four core modules:

1. LLM core: as the "brain" of the system, it generates natural language responses consistent with clinical scenarios and patient roles. In the current AI-SP system, this module is optimized primarily through medical dialogue–specific prompt engineering to improve medical accuracy and role consistency of generated responses. The output of this module is controlled through prompt engineering, which provides the technical foundation for parameterized clinical scenario generation and for constraints on patient-specific language styles. This module serves as the core computational component enabling context-sensitive dialogue generation and the integration of other AI-SP functions.
2. Affective computing module: responsible for identifying emotional cues in learner input (eg, tone, keywords, and empathic expression) and calculating the affective response the AI-SP should present based on the current patient's personality traits and emotional state. Emotional states are modeled as continuous processes rather than fixed scripted reactions, allowing learners' communication behaviors to influence the trajectory of the patient's emotional evolution. This module serves as the core mechanism enabling the dynamic emotional responsiveness function of the AI-SP system.
3. Personality and communication stance modulation module: this module integrates the Big Five personality model and Satir communication stance theory. The former defines the patient's stable personality traits (eg, agreeableness and neuroticism), while the latter defines typical communication behavior patterns under pressure (eg, placating, blaming, superreasonable, irrelevant, and congruent). Together, they determine AI-SP's response tendencies to different communication strategies. For example, highly neurotic patients are more likely to amplify negative information and tend toward a "blaming" stance, while highly agreeable patients are more likely to adopt a "congruent" or "placating" stance, cooperating with the learner's inquiry. This module serves as the core mechanism enabling personalized patient persona modeling within the AI-SP system.
4. Safety guardrail and content filtering module: this model ensures that the system's responses do not contain inappropriate content, do not provide medical advice beyond the SP role, and do not respond to learners in educationally unsuitable ways. For sensitive scenarios involving patient safety or ethical boundaries, the system should incorporate a "human-in-the-loop" mechanism, where the teacher can intervene and correct when necessary. This module provides essential safety assurance for all educational functions of the AI-SP system.

Based on the above architecture, our research team has preliminarily developed an AI-SP prototype system that supports both text and voice interaction modes and integrates a virtual avatar module. The detailed technical architecture and workflow of this prototype system are shown in the Multimedia Appendix 1.

Theoretical Foundations of Teaching

The design of AI-SP is based not only on technical feasibility but also on clear pedagogical theoretical support [9]. Two core theories are briefly explained below:

1. Kolb's experiential learning cycle: Kolb's experiential learning theory holds that effective learning comprises 4 stages: concrete experience → reflective observation → abstract conceptualization → active experimentation. This theory has been widely applied in medical simulation teaching and research. AI-SP can fully embody this cycle: learners gain near-realistic communication experiences through simulated consultations with the system; the system generated feedback reports help learners review their communication behaviors and their effects; learners internalize the feedback into understanding communication principles; and then attempt improved communication strategies in the next simulated scenario. This cycle can be repeated at low cost and high frequency, which is difficult to achieve with traditional SPs.
2. Deliberate practice: Psychologist Ericsson's (1993) deliberate practice theory emphasizes that skill improvement requires clear goals, immediate feedback, sufficient repetition, and gradually increasing difficulty. AI-SPs can provide technical support for this theory in several ways: learners can select specific communication skill goals (eg, "practice empathic expression" or "practice bad news disclosure"); the system provides immediate, consistent feedback; the same scenario can be repeated unlimited times; teachers adjust scenario parameters based on learner progress, gradually increasing complexity. Existing empirical studies have shown that AI-SPs designed with structured information architectures can achieve noninferiority to human SPs in communication skill improvement, with unique advantages for self-efficacy [10].

Technical Challenges and Coping Strategies

Acknowledging existing technical limitations, this section focuses on three key technical challenges and proposes corresponding improvement strategies and solutions:

1. AI "hallucination": generative AI may produce factually inaccurate information when uncertain. In medical education scenarios, this may manifest as the system describing nonexistent symptoms or providing statements that are inconsistent with clinical guidelines. The coping strategy is to use RAG technology to constrain the system's responses within preapproved clinical knowledge bases. For high-stakes assessment scenarios, a "human-in-the-loop" supervision mechanism should be maintained.
2. Inconsistency in affective responses: although the system is designed with personality modulation and affective computing modules, inconsistent affective responses may still occur during long dialogues or multiturn interactions. The coping strategy is to introduce a dialogue state tracking

XSL•FO
RenderX

mechanism into the system design to ensure consistency in personality and emotional state throughout the conversation. Additionally, teachers should conduct spot-check validation during the initial deployment phase.

3. Cultural sensitivity and content safety: the system's responses may inadvertently reflect stereotypes or culturally insensitive expressions toward specific populations (eg, patients from different regions, ages, or educational backgrounds). The coping strategy is to incorporate doctor-patient communication norms and humanistic consensus from the Chinese context into training data and prompt design and to establish a regular review mechanism involving medical humanities experts to assess the appropriateness of AI-SP responses.

It should be noted that the AI-SP proposed in this paper is not intended to replace real patient contact. Its value lies in providing learners with preparation, reflection, and supplementation—safe, controllable, repeatable practice opportunities before contact with real patients, and reflective support after real patient contact.

"Future Learning" Framework: Educational Integration of AI-SP

This section further proposes an integrated "future learning" framework to explain how this system can be integrated into medical education practice.

Three Core Dimensions of the Framework

This framework is based on three core dimensions:

1. Personalization: the system can adjust the complexity and support level of interactions according to the learners' ability levels. It provides more structured prompts and guided follow-up questions for junior medical students while presenting more ambiguous and complex patient statements for senior students, thereby forcing them to apply clinical reasoning skills.
2. Contextualization: teachers can quickly generate patient cases covering a wide range of clinical scenarios through parameterized configurations. Configurable variables include patient demographic characteristics, disease type and severity, emotional state, personality traits, and sociocultural background. This capability enables the system to simulate scenarios that are difficult to standardize in traditional SP training (eg, cross-cultural communication, patients with specific religious beliefs).
3. Emotional engagement: the system's affective dynamic responsiveness forces learners to face "emotional" patients. Learners' communication behaviors (especially empathic expression) directly affect patients' emotional states and cooperation. This design incorporates "emotional labor" into the core of communication training—learners must not only say the right things but also express emotions appropriately at the right moments.

Integration With Existing Teaching Systems

AI-SP is not intended to replace existing SP programs but to serve as their supplement and extension. Three possible integration pathways are described below:

1. Standardized practice before the OSCE: before formal assessments, learners can repeatedly practice the same clinical scenario to become familiar with communication processes and common coping strategies. This helps reduce assessment anxiety and ensures that all learners receive basic exposure training.
2. Supplement and expansion of SP resources: for resource-limited medical schools, AI-SP can serve as an "always available" alternative, covering high-frequency practice that is difficult to achieve with human SPs due to cost or complexity. For schools with mature SP programs, the system can be used to simulate special or rare clinical scenarios.
3. Data foundation for formative assessment: all interactions can be completely recorded and analyzed. These data can provide learners with detailed personal communication skill profiles, provide teachers with class-level analysis of common problems, and provide data-based decision support for curriculum improvement.

Evaluation Research Framework

This paper aims to propose a conceptual framework rather than present empirical data. Therefore, we propose the following research framework based on Kirkpatrick's 4-level evaluation model to guide future empirical validation (Table 3).

XSL•FO
RenderX

Table . AI-driven digital standardized patients (AI-SPs) evaluation framework based on the Kirkpatrick model[a].

Evaluation level	Evaluation content	Measurement tools	Time frame
Level 1: reaction	Learner satisfaction, perceived usefulness, and willingness to use	Self-designed questionnaire and system usage logs	Short term (1 - 3 mo)
Level 2: learning	Knowledge acquisition and skill improvement	SEGUE[b] Scale, JSPE[c] empathy scale, and pre-post test design	Medium term (3 - 6 mo)
Level 3: behavior	Transfer of communication behaviors to clinical practice	Comparison of AI-SP[d] training group vs control group in OSCE[e] and clinical supervisor evaluation	Medium to long term (6 - 12 mo)
Level 4: results	Ultimate impact on patient outcomes	Patient satisfaction in real clinical settings and communication-related medical error rates	Long term (>12 mo)

[a]Kirkpatrick's 4-level evaluation model was proposed by Donald Kirkpatrick in the 1950s and is a widely used training effectiveness evaluation framework in medical education research. The SEGUE Scale is used to assess doctor-patient communication skills; the JSPE Scale is used to measure empathy; and OSCE stands for the Objective Structured Clinical Examination.

[b]SEGUE: Set the stage, Elicit information, Give information, Understand the patient's perspective, and End the encounter—a framework for teaching and assessing communication skills.

[c]JSPE: Jefferson Scale of Physician Empathy.

[d]AI-SP: AI-driven digital standardized patient.

[e]OSCE: objective structured clinical examination.

Preliminary empirical studies provide feasibility support for this roadmap. Yang et al [11] identified high-learning–outcome pathway combinations in AI-generated content–enabled situational teaching using the fuzzy-set qualitative comparative analysis methodology; Gao et al [12] identified 6 learner needs for AI-SPs through a co-design approach; Wang et al [13] validated the feasibility of using LLMs as SPs for automated scoring; and a 2026 validation study [7] demonstrated through a 3-arm randomized controlled trial (n=58) that structurally designed AI-SPs are noninferior to human SPs in communication skill improvement.

Implementation Barriers and User Acceptance

The successful integration of AI-SP depends not only on technological maturity but also on acceptance within the educational system. Potential barriers include the following:

1. Resistance from teachers: teachers may worry that AI-SP will replace their teaching roles or feel uneasy about its technical complexity. Coping strategies include providing low-threshold teacher training workshops, positioning the system as an "auxiliary tool" rather than a "replacement," and inviting teachers to participate in scenario design to enhance their sense of ownership.
2. Skepticism from learners: learners may find conversations with AI "unrealistic" or worry that the system's feedback lacks clinical relevance. Coping strategies include emphasizing its advantages (eg, unlimited repetition, immediate feedback, and judgment-free environment) during early experiences and demonstrating its effectiveness through empirical data.
3. Technical costs and infrastructure: deploying AI-SP requires certain computational support and platform development investment. For resource-constrained institutions, cloud service models or interinstitutional sharing arrangements can be considered.
4. Time competition with existing curricula: medical curricula are already highly saturated. The introduction of AI-SP should not be seen as an "additional burden" but rather as a replacement for or supplement to less efficient parts of existing teaching. It is recommended to start with elective courses or extracurricular practice, accumulate evidence, and then gradually integrate it into required courses.

Ethical and Humanistic Considerations

The technical feasibility of AI-SP should not obscure the ethical challenges it brings [14]. This section focuses on 4 specific issues most directly relevant to AI-SP and proposes actionable design or regulatory recommendations for each.

Revisiting the Claim of "No Risk to Real Patients"

A common claim is that AI-SP allows learners to practice "without causing risk to real patients." This claim requires careful examination, as it may create risks of overreliance. Learners may become accustomed to interacting with AI, which in some respects (eg, the perception of nonverbal cues), is "simpler" than interacting with real patients, leading to insufficient adaptability when facing real patients. Second, there is a risk of incorrect learning. If certain responses contain inaccurate or inappropriate statements, learners may internalize them as "correct communication methods." Finally, there are privacy and data risks. If learners' dialogue data are improperly collected or used, sensitive information may be disclosed.

Therefore, AI-SP should be used under supervised conditions as a supplement and a preparation tool for real patient contact, not as a replacement.

Algorithmic Bias and Clinical Authenticity

LLMs may inadvertently reflect or amplify biases present in the training data. In the medical education context, this may lead to the following problems:

XSL•FO
RenderX

1. Stereotyping: generating stereotypical descriptions or responses for patient roles from specific regions, ages, educational backgrounds, or disease states.
2. Cultural insensitivity: inability to generate responses consistent with local norms for communication scenarios involving cultural customs, religious beliefs, or value judgments. Coping strategies include explicitly incorporating doctor-patient communication norms and humanistic consensus from the Chinese context into model training and prompt design; establishing a regular review mechanism involving medical humanities experts and cross-cultural communication researchers to sample and evaluate the appropriateness of AI-SP responses; and prioritizing human review for high-risk or culturally sensitive scenarios (eg, those involving reproductive decisions or end-of-life care).

In addition, AI-SP faces a deeper issue which is clinical authenticity, that is, whether the patient roles, symptom descriptions, and emotional responses simulated by the system can truly reflect patient characteristics and needs in clinical practice. If AI-SP's responses are overly templated or deviate from real clinical situations, learners may acquire inappropriate communication patterns or experience adaptation difficulties when facing real patients.

Data Privacy, Informed Consent, and Legal Risks

Each interaction with AI-SP generates dialogue records between learners and patient roles. These data not only have educational assessment value but also involve complex privacy and legal issues [15]. China has not yet issued laws or regulations specifically addressing the use of AI data in medical education scenarios [16], and the standards and procedures for data anonymization remain unclear. Judicial practice has already warned that using insufficiently anonymized patient information for teaching dissemination may constitute a privacy infringement. In the AI-SP context, if a virtual patient's identity markers are based on real cases, they may still be linked to specific individuals through information piecing, facing similar risks.

The above risks raise a series of core questions that need to be addressed: do learners know how their dialogue data are recorded, stored, and used? Will the data be used for model training? How are access permissions managed? Do learners have the right to access or delete their own data?

To address these issues, a 3-level protection mechanism is recommended. At the student level, provide clear, concise, informed-consent explanations before the first use, specifying data usage, retention period, dissemination restrictions, and data destruction procedures. At the management level, establish an "anonymized storage+tiered access+log tracking" protection mechanism with differentiated access permissions. At the content level, conduct compliance reviews of AI-SP's teaching content, including whether virtual patient symptom descriptions are based on anonymized data, whether system responses contain inappropriate suggestions, and whether algorithmic discrimination exists, ensuring educational content compliance.

The Necessity of Educator Oversight

Regardless of technological maturity, educator oversight is necessary for the foreseeable future. This is not only to address potential technical errors but also to ensure that system use aligns with educational goals.

Specific forms of oversight may include (1) content audits: teachers periodically sample AI-SP responses in specific scenarios to assess their accuracy, appropriateness, and consistency; (2) anomaly reporting: the system automatically flags potentially problematic interactions (eg, involving unsafe medical advice) for teacher review; (3) learner feedback channels: allow learners to mark responses that feel "unrealistic" or "potentially problematic" after interactions; and (4) teaching intervention: when systematic deviations occur in critical scenarios, teachers can quickly adjust prompts or parameters or temporarily switch to human-answered mode.

Retaining educator oversight and intervention capacity is prudent and necessary until the safety and effectiveness of AI-SP are fully validated.

Discussion and Limitations

Main Contributions of This Paper

Compared with the existing literature, the main contributions of this paper are reflected in the following 3 aspects.

First, it proposes an integrated technology, pedagogy, and ethics framework. Unlike studies that focus solely on technical feasibility, this paper positions the system as a "pedagogical agent" and systematically elaborates on it across 4 dimensions: system architecture, teaching theory, evaluation roadmap, and ethical considerations.

Second, it explicitly addresses the Chinese context in medical education. Based on the latest empirical data (from 506 medical education institutions and cross-cultural communication comparative studies) and national policy documents, it elucidates the special needs and conditions for developing AI-SP in the Chinese context.

Third, it proposes a testable evaluation roadmap. Unlike articles that remain at the level of conceptual advocacy, this paper proposes specific research designs and measurement tools based on the Kirkpatrick model, providing actionable starting points for subsequent empirical research.

Educational Value of AI-SP: Beyond Experiential Learning and Deliberate Practice

The educational value of the AI-SP system proposed in this study can be further interpreted from the perspective of learner autonomy. Kolb's experiential learning theory and Ericsson's framework of deliberate practice explain fundamental mechanisms underlying skill acquisition. However, generative AI–supported simulation environments introduce possibilities that extend beyond the traditional "practice-feedback" cycle.

First, learners can autonomously define their learning objectives. The proposed system allows learners to select patients with different personality profiles for communication training and adjust the training difficulty according to their individual needs.

XSL•FO
RenderX

This design aligns with the theoretical assumptions of self-directed learning, in which learners gain greater control over the learning process within a safe simulated environment. Such autonomy represents a core component of adult learning theory.

Second, dynamic adaptation enables personalized learning trajectories. The prompt engineering architecture allows AI-SPs to adjust their emotional responses in real time according to learners' communication behaviors (eg, emotional relief after empathic responses from learners). This adaptive process extends beyond the "fixed task–repeated practice" paradigm traditionally emphasized in deliberate practice frameworks and provides a technological foundation for individualized learning experiences.

Third, human-AI interaction supports collaborative knowledge construction. The evaluation module implemented through the RAG+agent architecture not only provides scores but also identifies specific dialogue segments and generates targeted improvement suggestions. This enables learners to reflect on their communication assumptions and recognize potential blind spots rather than simply receiving binary judgments of "correct" or "incorrect." Such a design aligns with the constructivist learning perspective that knowledge is actively constructed through interaction.

Limitations of a Viewpoint Article

This paper is a viewpoint article rather than an empirical research study; therefore, the framework, system architecture, and evaluation roadmap proposed herein await subsequent empirical testing. It should be noted that our research team has conducted preliminary exploratory applications around this framework, and the initial results support its feasibility. Relevant empirical data will be presented in follow-up studies. Furthermore, because this is a conceptual framework, the description of the system architecture in this paper remains at the module level, and the discussion of user acceptance is based primarily on literature analysis and preliminary practice, lacking large-scale systematic user studies. These limitations are inherent to viewpoint articles. The goal of this paper is to propose a testable framework, not to provide a fully validated solution.

Furthermore, the current AI-SP system primarily adopts prompt engineering for patient role simulation rather than model fine-tuning. This technical choice supports dynamic persona configuration and diverse clinical scenario generation; however, it also means that role consistency remains dependent on the quality of prompt design. In complex or unexpected interaction scenarios, the system may exhibit reduced role consistency, such as generating responses containing medical knowledge beyond the predefined patient role or inconsistent emotional responses. With the accumulation of high-quality doctor-patient communication datasets, future fine-tuning approaches may further enhance the realism, stability, and personalization of AI-SP personas.

Future Research Directions

Based on the above limitations, this paper proposes the following future research directions:

1. Technical validation studies: develop prototype systems to validate personality consistency, affective response accuracy, and medical safety in simulated consultations.
2. Teaching effectiveness evaluation: following the Kirkpatrick roadmap proposed in the *Evaluation Research Framework* section, conduct randomized controlled trials to compare differences between AI training groups and human training groups in communication skills, empathy, and self-efficacy.
3. Implementation science research: conduct implementation research in real medical schools to identify key facilitators and barriers to successful integration, and explore adaptation models under different resource conditions (Eastern vs Western and resource-rich vs resource-constrained).
4. Ethical framework refinement: based on ethical issues arising from practical use, continuously refine specific operational norms for data privacy, algorithmic bias, and educator oversight.

Conclusions

The generative AI–SP is not a replacement for traditional medical simulation teaching but rather a conceptual evolution and extension; it reconfigures the teaching relationship among learners, educators, and simulated patients. By integrating generative AI, affective computing, and personality modeling into controllable, repeatable simulated environments, it provides new possibilities for communication training, humanistic education, and emotional competence development in medical education.

The realization of these possibilities depends on the joint satisfaction of 3 conditions: technical robustness, pedagogical adaptability, and ethical prudence. This paper proposes a "future learning" framework, whose 3 pillars are personalization, contextualization, and emotional engagement. Regarding the technical pathway, this study adopts prompt engineering as the primary approach, supplemented by RAG+agent architectures to support evaluation functions. This choice is based on the system's educational positioning as a dynamic role-playing environment rather than standardized dialogue generation, rather than representing a technological limitation. Future improvements may incorporate model fine-tuning as sufficient high-quality interaction data become available.

This paper does not advocate replacing real patient contact or the value of human SPs with AI-SPs. On the contrary, their value lies in preparation, reflection, and supplementation: providing safe, controllable practice opportunities before contact with real patients; providing structured reflective support after contact with real patients; and providing scalable supplementary solutions in resource-limited environments.

In the intelligent era, the humanistic core of medical education will not automatically dissolve with technological intervention. How to make AI serve, rather than replace, the humanistic dimensions of the doctor-patient relationship is a discussion that this paper hopes to initiate, not a definitive answer that it provides. Ultimately, whether technology can truly serve humanistic values depends on how educators design, implement, and critically evaluate its use, and this remains a fundamental and enduring question in medical education.

XSL•FO
RenderX

Funding

This study was supported by the National Social Science Foundation of China (grant number 25BGL283) for the project "Research on Application Strategies of Artificial Intelligence in Doctor-Patient Communication."Generative AI tool ChatGPT was used for language refinement and copyediting purposes. The final manuscript was thoroughly reviewed and approved by all the authors, who assume full responsibility for the entire content.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Technical architecture and workflow of the AI-driven digital standardized patient prototype system.
[PNG File, 333 KB - mededu_v12i1e91050_app1.png]

References

1. Hu Q, Zhang Y, Wu W. Deep integration of artificial intelligence and education: logic, dilemmas, and pathways [Article in Chinese]. China Educ Technol 2026(6):1-6 [FREE Full text]
2. Yu C, Liu YF, Li XR, et al. Research on the development of the directory of Chinese higher education institutions (medicine) [Article in Chinese]. Chin J Med Educ 2026;46(3):161-166. [doi: 10.3760/cma.j.cn115259-20250902-01044]
3. Barrows HS. An overview of the uses of standardized patients for teaching and evaluating clinical skills. AAMC. Acad Med 1993 Jun;68(6):443-451. [doi: 10.1097/00001888-199306000-00002] [Medline: 8507309]
4. Cleland JA, Abe K, Rethans JJ. The use of simulated patients in medical education: AMEE Guide No 42. Med Teach 2009 Jun;31(6):477-486. [doi: 10.1080/01421590903002821] [Medline: 19811162]
5. Cross J, Kayalackakom T, Robinson RE, et al. Assessing ChatGPT's capability as a new age standardized patient: qualitative study. JMIR Med Educ 2025 May 20;11:e63353. [doi: 10.2196/63353] [Medline: 40393017]
6. Liu C, Tu YH, Lin LJ, et al. Doctor-patient communication models, patient decision-making participation, and patient emotional expression: a cross-cultural comparison of samples from the UK and China. Patient Prefer Adherence 2025;19:2505-2524. [doi: 10.2147/PPA.S537829] [Medline: 40860605]
7. Wang T, et al. "Life has pain, apricot has you" Chinese medical humanities expert consensus [Article in Chinese]. Narrative Med 2025;8(3):159-165 [FREE Full text]
8. García-Torres D, Fernández C, Mira JJ, Morales A, Vicente MA. Using AI-based virtual simulated patients for training in psychopathological interviewing: cross-sectional observational study. JMIR Med Educ 2025 Dec 23;11:e78857. [doi: 10.2196/78857] [Medline: 41433050]
9. Wu F, Jiang FJ, Wang LY. How does generative artificial intelligence reshape university students' learning behaviors [Article in Chinese]. China Educ Technol 2026(5):49-56 [FREE Full text]
10. Wu P, Han Y, Zhang J, et al. From simulation to pedagogy: structured AI standardized patients for clinical communication training validated through multi-model and randomized evaluation. medRxiv. Preprint posted online on Apr 28, 2026. [doi: 10.64898/2026.04.26.26351793]
11. Yang JL, Wang CX, Cheng ZQ. Research on teaching pathways and strategy optimization empowered by generative artificial intelligence [Article in Chinese]. China Educ Technol 2026(4):24-31 [FREE Full text]
12. Gao Z, Zhu G, Luo H, et al. "It talks like a patient, but feels different": co-designing AI standardized patients with medical learners. In: Oliver N, Shamma DA, Candello H, Cesar P, Lopes P, Artizzu V, et al, editors. CHI EA '26: Proceedings of the Extended Abstracts of the 2026 CHI Conference on Human Factors in Computing Systems: Association for Computing Machinery; 2026:1-6. [doi: 10.1145/3772363.3798336]
13. Wang C, Li S, Lin N, et al. Application of large language models in medical training evaluation-using ChatGPT as a standardized patient: multimetric assessment. J Med Internet Res 2025 Jan 1;27:e59435. [doi: 10.2196/59435] [Medline: 39742453]
14. Ota Y, Asada Y, Kubo S, Kanno T, Yagi MS, Matsuyama Y. Evaluation of a problem-based learning program's effect on artificial intelligence ethics among Japanese medical students: mixed methods study. JMIR Med Educ 2026 Jan 14;12:e84535. [doi: 10.2196/84535] [Medline: 41534078]
15. Ma C. Digital citizen identity confirmation and rights protection [Article in Chinese]. Faxue Yanjiu 2023;45(4):21 [FREE Full text]
16. Zhou HH. On the orientation of artificial intelligence legislation in China [Article in Chinese]. Xiandai Faxue 2024;46(5):17-34 [FREE Full text]

XSL•FO
RenderX

Abbreviations

AI-SP: AI-driven digital standardized patient
LLM: large language model
OSCE: objective structured clinical examination
RAG: retrieval-augmented generation
SEGUE: Set the Stage, Elicit Information, Give Information, Understand the Patient's Perspective, and End the Encounter
SP: standardized patient
SPIKES: setting up the interview, assessing the patient's perception, obtaining the patient's invitation, giving knowledge and information, addressing empathy and emotions, strategy and summary

Edited by D Chartash; submitted 08.Jan.2026; peer-reviewed by JW Lai, S Martino, TA Elwi; revised version received 17.Jul.2026; accepted 25.Jul.2026; published 20.Aug.2026.

Please cite as:
Jia C, Qi L
Digital Standardized Patients: Conceptual Framework for Generative AI–Empowered Medical Education
JMIR Med Educ 2026;12:e91050
URL: https://mededu.jmir.org/2026/1/e91050
doi:10.2196/91050

© Chuanlin Jia, Lulu Qi. Originally published in JMIR Medical Education (https://mededu.jmir.org), 20.Aug.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Education, is properly cited. The complete bibliographic information, a link to the original publication on https://mededu.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Scaffolded AI-Supported Problem-Based Learning for Medical Interns: Exploratory Retrospectively Registered Randomized Controlled Evaluation With a Voluntary Feasibility Follow-Up

Zhengqi Zhao[1,2*], BEng; Baijing Wu[3*], BM; Shuyuan Tian[3], PhD; Haigen Hu[1], PhD; Pengyi Hao[1], PhD

[1]Department of Computer Science and Technology, Zhejiang University of Technology, No. 288 Liuhe Road, Xihu District, Hangzhou, Zhejiang, China

[2]Zhejiang Institute of Artificial Intelligence, Hangzhou, Zhejiang, China

[3]Department of Ultrasound, Tongde Hospital Affiliated to Zhejiang Chinese Medical University, No. 234 Gucui Road, Hangzhou, Zhejiang, China

[*]these authors contributed equally

Corresponding Author:
Pengyi Hao, PhD
Department of Computer Science and Technology, Zhejiang University of Technology, No. 288 Liuhe Road, Xihu District, Hangzhou, Zhejiang, China

Abstract

Background: High-quality problem-based learning (PBL) during internship is resource-intensive and difficult to scale without consistent facilitation. Although generative AI is increasingly used in health professions education, many applications remain on-demand answer tools that may not reproduce core PBL processes.

Objective: This study aimed to develop and evaluate the Multiagent PBL Environment for Clinical Reasoning (MAPLE-CR), an AI-supported environment that uses generative AI as a process-oriented scaffold rather than an answer-delivery aid. The system supports cognitive mechanisms through reasoning prompts, social-interactional mechanisms through simulated tutor and peer roles, and regulatory mechanisms through structured workflows and feedback loops. We separately evaluated implementation and repeated-use feasibility, including learner experience, short-term outcomes versus self-study, and preliminary framework-aligned process evidence.

Methods: We conducted a 2-stage study. In an exploratory randomized evaluation (N=52; intervention: n=26; control: n=26), interns completed parallel pretests and posttests around a standardized case. The intervention group engaged in asynchronous, scaffolded PBL in MAPLE-CR, whereas controls completed case-matched self-study using materials derived from the same case and learning objectives. We summarized transcript-derived process dimensions and postsession learner experience in the intervention group. In a voluntary follow-up, 18 participants completed 1 MAPLE-CR case per week for 4 additional weeks to examine repeated use feasibility and patterns across novel cases.

Results: Baseline pretest scores were comparable between groups (P=.56). The intervention group achieved higher posttest scores than controls (Hodges-Lehmann median difference 8.25 points, 95% CI 6.60-11.60; Cliff δ=0.506, 95% CI 0.21 - 0.76; P=.001) and greater score gains (median difference 6.60 points, 95% CI 1.60-14.95; P=.01). On a 0 to 5 scale, mean scores were highest for knowledge accuracy (4.4) and clinical reasoning (CR, 3.7), whereas active participation averaged 2.1, and interaction-oriented dimensions were more variable. Means for 12 positively worded questionnaire items ranged from 4.19 to 4.62, supporting high satisfaction, involvement, perceived support, and self-efficacy; open-ended responses contextualized perceived strengths and improvement needs. In the voluntary follow-up (n=18), weekly cross-case use was feasible; median within-session gains ranged from 13.4 to 20.0 points across 5 attempts, with marked increases in CR and aggregate scores from attempts 1 to 2 but statistically uncertain later trajectories.

Conclusions: MAPLE-CR was feasible and showed high learner acceptability. Its use was associated with larger short-term CR knowledge gains than case-matched self-study, while transcript analyses provided preliminary framework-aligned process evidence. However, the post hoc sensitivity analysis did not establish prospective power sufficiency, and the confidence interval could not exclude smaller educationally meaningful effects. These findings suggest that AI-supported, process-oriented PBL may provide a scalable formative supplement for CR practice when facilitator capacity and small-group scheduling are constrained, but the study did not isolate specific scaffolding effects. Further research should confirm these findings using larger, prospectively powered trials with stronger active comparators and longer-term outcome measures.

Trial Registration: Chinese Clinical Trial Registry ChiCTR2500115799; https://tinyurl.com/353y6c92

(JMIR Med Educ 2026;12:e95039) doi:10.2196/95039

XSL•FO
RenderX

KEYWORDS

problem-based learning; clinical reasoning; scaffolding; artificial intelligence; large language models; educational technology; randomized controlled trial

Introduction

Medical interns must translate biomedical knowledge into timely clinical decisions under uncertainty. Problem-based learning (PBL) supports this transition by engaging learners in case analysis, hypothesis generation, interpretation of evolving information, and iterative revision of explanations and differential diagnoses [1,2]. Its educational value depends not only on exposure to cases but also on learners' active articulation of reasoning, identification of knowledge gaps, self-directed study, and integration of new understanding [3-5].

Despite these benefits, scaling high-quality PBL in routine clinical curricula remains difficult. Effective PBL typically depends on skilled facilitation, small-group coordination, and protected instructional time, all of which are resource-intensive and difficult to sustain consistently during internship [6]. In practice, the challenge is not only increasing access to PBL-like activities but also preserving the pedagogical processes that make PBL distinctive. Its educational value depends on collaborative sensemaking, externalization of reasoning, and facilitator-guided progression through a structured case process [7]. When these enabling processes are weak or absent, PBL can deteriorate into superficial discussion or premature answer-seeking rather than disciplined reasoning.

Against these implementation constraints, generative AI, especially large language model (LLM) systems capable of flexible multiturn dialogue, has emerged as a potential way to expand learning opportunities in time- and resource-constrained clinical settings [8-12]. Early work in health professions education has explored generative AI for functions such as content support, explanation, feedback, and learner assistance [13-15]. Several studies have also begun to examine AI-supported learning in PBL-related contexts, with encouraging findings for learner support, idea generation, and just-in-time explanation [15,16]. However, in many of these implementations, AI functions primarily as a consultation aid or question-answering assistant embedded within the learning process, rather than as support for orchestrating the PBL process itself.

This distinction matters because question answering is not equivalent to PBL. PBL is not only a matter of information access [1,17]; it is a structured learning process in which learners are expected to generate and justify hypotheses, respond to new information, engage with alternative perspectives, and move through recognizable phases of inquiry [18]. Its benefits also depend on feedback and continuity across cases, so that learners can reflect on performance, recalibrate reasoning, and improve over time [19]. Accordingly, simply providing answers or ad hoc explanations is unlikely to reproduce the core pedagogical functions of PBL. The specific gap addressed in this study is therefore not the absence of AI assistance within case learning but the limited evidence on AI-supported PBL environments designed to scaffold multiple phases of the PBL cycle as a facilitator-like process [20], integrating presession preparation, in-session reasoning and peer interaction, postsession feedback, and subsequent practice rather than supporting only isolated question answering or explanation.

In this study, we adopt a process-oriented view of scaffolding for AI-supported PBL. We use scaffolding to mean structured support embedded in the learning workflow to help learners perform reasoning, interaction, and reflection activities that would otherwise require sustained tutor facilitation. We use orchestration to refer to the coordination of scaffolded activities across the full PBL cycle, including presession preparation, case discussion, postsession feedback, and subsequent practice. Within this theoretical framework, MAPLE-CR (Multiagent PBL Environment for Clinical Reasoning) was designed around 3 complementary forms of process-oriented scaffolding. First, cognitive scaffolding prompts learners to generate hypotheses, justify them with case evidence, revise interpretations as new information emerges, and synthesize key clinical points [1,5,17]. Second, social-interactional scaffolding uses simulated tutor and peer roles to sustain dialogue, questioning, turn-taking, peer-responsive reasoning, and externalization of thinking rather than a purely monologic learner-AI exchange [21-23]. Third, regulatory scaffolding structures the workflow, monitors objective coverage, manages phase transitions, and embeds feedback loops across repeated sessions to support reflection, calibration, and progressive improvement [5,19,24]. This framework treats scalability not simply as replacing human labor with automation but as preserving the cognitive, dialogic, and feedback-regulatory functions that make PBL educationally meaningful.

Guided by this framework, we developed MAPLE-CR, an AI-supported learning environment for continuous, scaffolded PBL practice among medical interns. It was intended to provide a structured yet flexible environment for clinical reasoning (CR) practice when in-person facilitation and repeated small-group sessions are difficult to sustain. Accordingly, this study aimed to evaluate MAPLE-CR across three distinct evidentiary domains: (1) feasibility of implementation and repeated cross-case use, together with learner acceptability and perceived educational value; (2) short-term learning-outcome differences relative to case-matched self-study in an exploratory randomized evaluation; and (3) preliminary framework-aligned process evidence based on transcript-derived formative dimensions, with expert ratings used as an external reference (Table 1 summarizes this mapping and related expert-rating checks). The voluntary follow-up was interpreted descriptively as evidence of repeated-use feasibility rather than longitudinal learning effectiveness.

XSL•FO
RenderX

Table . Intended mapping of Multiagent Problem-based Learning Environment for Clinical Reasoning (MAPLE-CR) components to scaffolding functions, formative dimensions, and expert-rating validation.

Scaffolding function	MAPLE-CR components and intended PBL[a] support	Framework-aligned formative dimensions and expert-rating validation
Cognitive scaffolding	Learning objectives, sequential case scenarios, and tutor prompts for hypothesis generation, evidence justification, revision, and synthesis. These components were intended to elicit diagnostic and management reasoning aligned with case objectives.	CR[b]; KA[c]; expert ratings for CR and KA
Social-interactional scaffolding	Tutor and peer roles, peer prompts and counterpoints, and turn-taking support. These components were intended to sustain multivoice discussion and externalization of reasoning.	AP[d]; CC[e]; expert ratings for AP and CC
Regulatory scaffolding	Case recommendation, pretest-informed learner profile, objective-coverage monitoring, phase transitions, summaries, feedback, and updated learning record. These components were intended to structure reflection, calibration, and repeated practice.	SR[f]; expert ratings for SR

[a]PBL: problem-based learning.

[b]CR: clinical reasoning.

[c]KA: knowledge accuracy.

[d]AP: active participation.

[e]CC: communication and collaboration.

[f]SR: summarization and reflection.

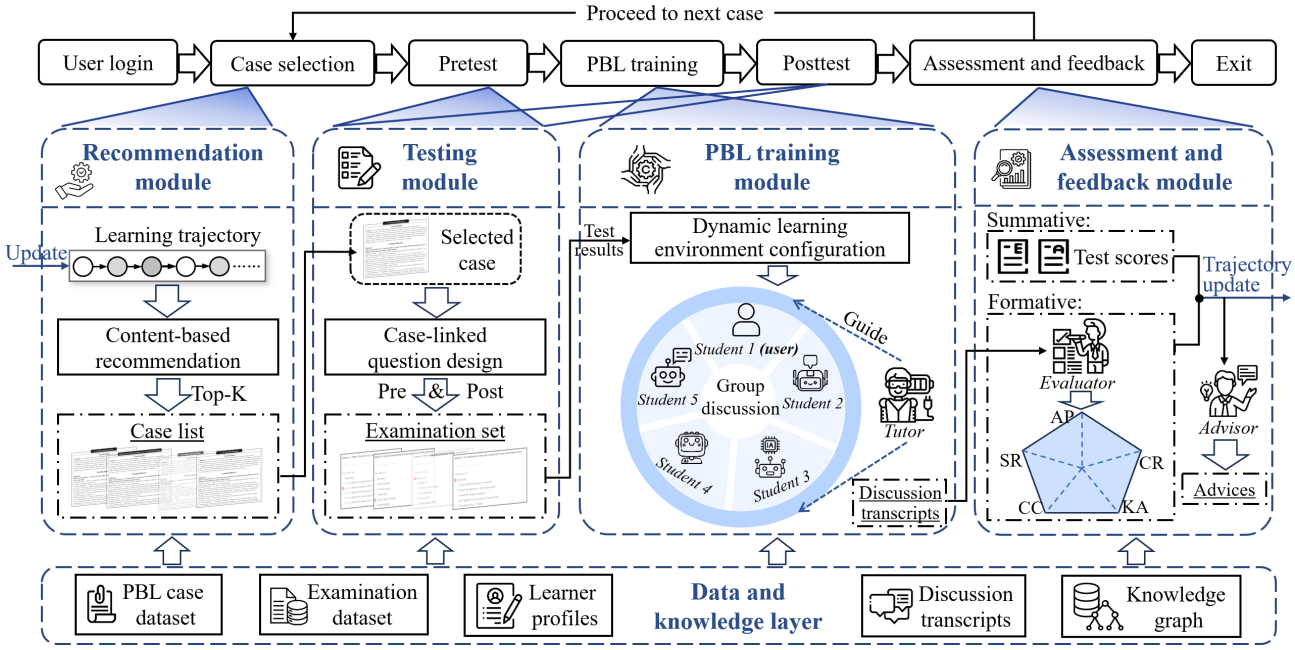
Methods

MAPLE-CR

MAPLE-CR is a web-based, LLM-supported PBL environment for medical interns; an overview of the system structure and learning workflow is shown in Figure 1. In implementation terms, an “agent” refers to an LLM-guided software role rather than a human participant, consistent with recent work on LLM-based agents and multiagent collaboration [25,26]. MAPLE-CR instantiates tutor, peer, evaluator, and advisor functions as distinct agents and uses multiagent orchestration to coordinate role interactions and workflow sequencing. Learners interact with MAPLE-CR through 4 connected modules: recommendation, testing, PBL training, and assessment with feedback, all supported by a shared data and knowledge layer containing cases, learning objectives, examination items, learner records, discussion transcripts, formative scores, and knowledge graph resources.

XSL•FO
RenderX

Figure 1. Multiagent Problem-based Learning Environment for Clinical Reasoning closed-loop problem-based learning workflow. Learners select a recommended case, complete a brief pretest, engage in scenario-based small-group discussion with a tutor role and simulated peers, and then complete a posttest. Transcript-based formative feedback is generated across 5 dimensions. Session outputs update the learner's trajectory to inform subsequent case recommendations, enabling repeated, scaffolded practice. AP: active participation; CC: communication and collaboration; CR: clinical reasoning; KA: knowledge accuracy; PBL: problem-based learning; SR: summarization and reflection.



The recommendation and testing modules initiate each learning cycle. The recommendation module uses the learner's prior learning trajectory, defined as the longitudinal record of completed cases and assessment history, to generate a ranked list of candidate cases. The learner selects 1 recommended case, preserving learner choice while supporting continuity across repeated practice. The testing module then administers brief case-linked pretests and posttests aligned with the case learning objectives and target knowledge points; the pretest summarizes baseline case-relevant understanding, whereas the posttest assesses immediate knowledge performance after the PBL session.

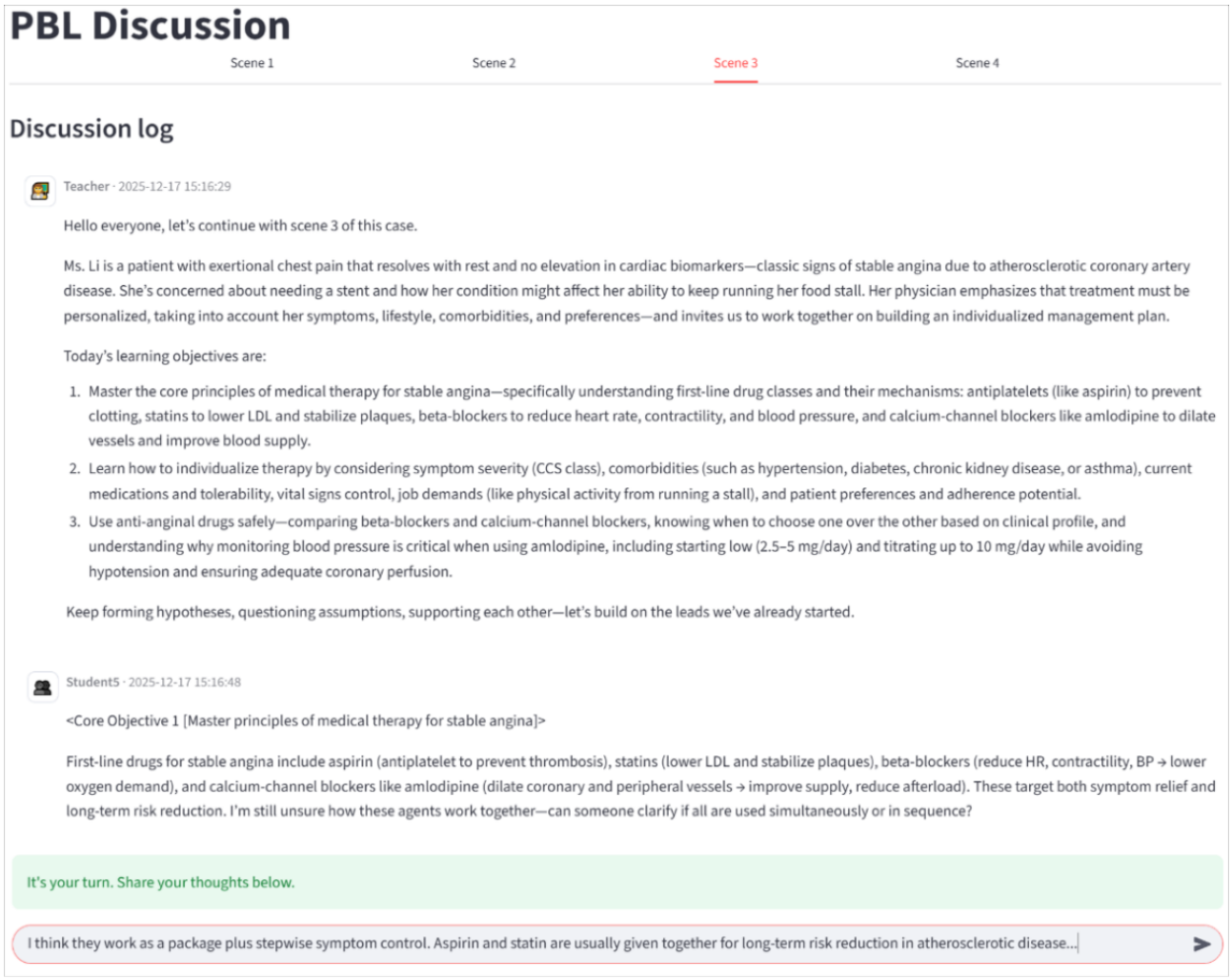
The PBL training module is the objective-guided learning environment. Cases are organized around prespecified learning objectives and sequential scenarios. Within each scenario, the tutor agent presents case information, monitors objective coverage, and supports transition to summary phases. Peer-agent personas combine fixed role constraints with learner-specific ability calibration: the fixed prompt defines the peer as a medical learner who participates in discussion, asks questions, and responds to others, whereas the dynamic component adjusts the peer agent's ability profile according to the learner's pretest-derived baseline.

The assessment with feedback module combines summative and formative information. Summative feedback includes the pretest score, posttest score, and score change. Formative assessment is generated from the discussion transcript across 5 PBL-relevant dimensions aligned with the scaffolding framework: active participation (AP), CR, knowledge accuracy (KA), communication and collaboration (CC), and summarization and reflection (SR). AP is computed algorithmically from learner interaction frequency, whereas CR, KA, CC, and SR are scored by an evaluator agent using dimension-specific prompts and the full discussion context. CR and KA additionally use retrieval from the structured medical knowledge graph as a clinical reference [14]. Evaluator-agent scores were constrained to a 0 to 5 scale in 0.5-point increments and accompanied by brief dimension-specific justifications. The resulting feedback summarizes strengths and areas for improvement and updates the learner record for subsequent case recommendations.

MAPLE-CR was implemented as a remotely accessed web platform, with a Streamlit interface deployed on a cloud server. Model inference used the Qwen-flash API (qwen-flash-2025-07-28) [27], with temperature set to 0.7. Only pseudonymized educational content was transmitted to the Qwen API for model inference; direct personal identifiers, account credentials, and study-specific participant identifiers were excluded from API requests. The user interface for the training module is shown in Figure 2. Further details on API data flow, access control, data retention, and consent are provided in Multimedia Appendix 1. Multimedia Appendix 2 provides a video demonstration of the system, and Multimedia Appendix 1 provides additional implementation details, including module interactions, system configuration, peer-agent calibration, and agent prompts.

XSL•FO
RenderX

Figure 2. User interface of the problem-based learning (PBL) training module in Multiagent PBL Environment for Clinical Reasoning. The screenshot shows a later management-oriented scenario in which the tutor agent presents staged case information and learning objectives to structure discussion around the selected case. Within this objective-guided format, the peer agent initiates discussion and the learner contributes through text input. BP: blood pressure; HR: heart rate; LDL: low-density lipoprotein.



Study Design

We conducted an exploratory randomized controlled educational evaluation followed by a voluntary 4-week repeated-use feasibility follow-up. Reporting was guided by the CONSORT-eHEALTH (Consolidated Standards of Reporting Trials of Electronic and Mobile Health Applications and Online Telehealth) checklist (Checklist 1).

Fifty-two first-year postgraduate medical interns (resident interns) were recruited from Tongde Hospital Affiliated to Zhejiang Chinese Medical University, a tertiary teaching hospital in Zhejiang Province, China. All had prior face-to-face PBL experience and reported previous use of GPT-style chatbots, but none had used MAPLE-CR. Participants were randomized 1:1 to the intervention or control group using a computer-generated sequence prepared by a researcher not otherwise involved in the trial. Participant recruitment and enrollment took place from December 6 to 15, 2025, and the randomized educational evaluation was conducted on December 19, 2025, before trial registration; therefore, the trial was retrospectively registered.

MAPLE-CR system and code development, including backend model integration and platform deployment, began in August 2025 and was followed by internal testing and debugging by the research team before participant recruitment. These deployment and development activities were limited to internal testing and debugging and did not involve participant enrollment, randomization, or collection of trial outcome data.

All participants completed the pretest concurrently in an on-site computer laboratory. To reduce contamination, groups were seated in separate classrooms. Research team members monitored the classrooms to maintain timing and ensure access to the assigned materials or MAPLE-CR platform; they did not provide clinical teaching, facilitate discussion, answer case-related questions, or collect additional outcome data beyond documenting attendance and session completion. Allocation was concealed until completion of the pretest; participant blinding after allocation was not feasible. Pretest, posttest, questionnaire, and follow-up records were linked using study-specific participant identifiers; analyses were conducted on deidentified datasets without names or other direct personal identifiers. To standardize exposure, all participants used the same standardized PBL case, and the recommendation function

XSL•FO
RenderX

was therefore not used in the randomized component. The case focused on CR for a right middle-lobe lung mass ultimately diagnosed as lung adenocarcinoma. The case materials, learning objectives, and associated pretest and posttest items were developed and reviewed by 3 domain experts. Pretest and posttest were parallel forms based on a shared blueprint (see Multimedia Appendix 3); experts reviewed item difficulty, learning-objective coverage, and consistency between the case materials and assessment items, resolving discrepancies by consensus. Each form comprised 30 single-best-answer multiple-choice items scored automatically against a fixed answer key as the percentage of correct responses (range 0‐100), without using group-assignment information; all items had to be completed before submission. Blueprint development and expert review preceded the trial. To provide supplementary evidence for the assessment instruments, we conducted post hoc psychometric checks using actual item-level responses and independent 5-point expert difficulty ratings for all pretest and posttest items. These analyses supported acceptable internal consistency, item discrimination, and comparable expert-rated difficulty between forms; detailed methods and results are provided in Multimedia Appendix 3.

Immediately after the pretest, participants in the intervention group received a standardized 40-minute orientation to MAPLE-CR before individually completing up to 2 hours of PBL training. The orientation introduced the platform workflow, interface operations, tutor and peer roles, and basic expectations for participating in the AI-supported PBL discussion. It did not include case-specific diagnostic or management instruction, posttest answers, or substantive teaching on the target case. Because the orientation was intended to support learners' engagement with the platform, it was considered part of the MAPLE-CR intervention package. If discussion ended early, participants reviewed the transcript until the session ended. A 2-hour case-matched self-study condition was selected as a feasible independent-learning comparator for this exploratory evaluation, using expert-reviewed materials to standardize case content and learning objectives without matching MAPLE-CR's interaction, feedback, novelty, or procedural guidance. Both groups then completed the posttest concurrently. Because the 40-minute orientation had no time-matched counterpart in the control condition, the conditions were not equivalent in total structured exposure. Accordingly, the comparison estimates the effect of the integrated MAPLE-CR intervention package under the present implementation rather than the isolated effect of its AI-PBL scaffolding components. In the intervention group, MAPLE-CR generated transcript-based formative feedback only after the posttest; no comparable feedback was provided to controls.

After the session, participants completed a postsession questionnaire adapted from a previously published simulation-based training quality assurance instrument [28,29]. Because the original instrument was designed for simulation-based training rather than AI-supported PBL, item wording and domains were modified to fit the MAPLE-CR context. Items were rated on a 5-point Likert scale from strongly disagree to strongly agree.

Following the controlled study, intervention participants were invited to join a voluntary 4-week follow-up, completing 1 MAPLE-CR case per week with the full closed-loop workflow (recommendation, testing, PBL training, assessment, and feedback). At follow-up completion, participants provided open-ended feedback on their repeated-use experiences. The voluntary follow-up was conducted from December 22, 2025, to January 19, 2026.

No important changes to study procedures, intervention content, or platform functionality occurred after trial commencement, and no major technical failures or downtime affecting study delivery were observed.

Sample Size and Power Considerations

No formal a priori sample size calculation was conducted. The available sample size was determined by the fixed intern cohort, the number of eligible participants who could be recruited, and the feasibility of conducting a synchronized on-site educational experiment. Accordingly, the randomized component should be interpreted as an exploratory proof-of-concept evaluation rather than a fully powered confirmatory efficacy trial. To clarify the detectable magnitude under the available sample, we conducted a post hoc sample size sensitivity analysis based on the observed posttest effect. The observed Cliff δ was 0.506, corresponding to an area under the curve of $(\delta+1)/2=0.753$ and an approximate Cohen *d* of 0.97. With 26 participants per group and 2-sided $\alpha=.05$, this observed effect was larger than the standardized mean difference corresponding approximately to 80% power ($d\approx0.78$) and 90% power ($d\approx0.90$). However, the lower bound of the 95% CI for Cliff δ (0.21) corresponds to an approximate *d* of 0.38, indicating that the effect-size estimate remains imprecise and that smaller but educationally meaningful effects cannot be excluded. This analysis is therefore presented only as a descriptive sensitivity analysis and does not substitute for a prospective sample size calculation.

Registration and Analysis Status

The registration was completed retrospectively and specified broad outcome domains related to CR ability and knowledge mastery; it did not prospectively specify the operational outcomes, hypotheses, statistical analysis plan, or voluntary follow-up measures used in this educational evaluation. The pretest and posttest knowledge assessments, postsession questionnaire, transcript-based formative assessment, and voluntary repeated-use follow-up were incorporated into the study procedures; however, because registration followed the randomized evaluation, none constituted a prospectively registered confirmatory outcome, and their analyses are interpreted as exploratory or descriptive. Expert-rating convergence, session-level correlations, post hoc psychometric checks, the questionnaire response-pattern sensitivity analysis, and sample size sensitivity calculations were not specified in the registration and are reported as exploratory, sensitivity, or post hoc analyses, as applicable.

Statistical Analyses

All 52 randomized participants completed the pretest and posttest and were analyzed as assigned under an intention-to-treat approach; intervention questionnaires and

XSL•FO
RenderX

transcripts were complete, and no imputation was required. We used primarily nonparametric methods because of the modest sample size and the ordinal nature of questionnaire data. Baseline group characteristics and between-group outcomes (pretest, posttest, and score gain) were compared using Fisher exact tests or Mann-Whitney *U* tests, as appropriate. Within-group pretest-posttest changes were assessed using Wilcoxon signed-rank tests. For repeated-use outcomes across 5 attempts, we used Friedman tests for repeated within-subject comparisons. We report effect sizes and 95% CIs where applicable (eg, Cliff δ, Hodges-Lehmann median differences, or bootstrap CIs). All tests were 2-sided with α=.05. Among intervention participants only, transcript-based formative scores were summarized for the 26 MAPLE-CR sessions. Convergent validity was examined using Spearman rank correlations (ρ) and mean absolute error (MAE) against mean expert ratings, and exploratory session-level associations within the intervention group were examined using 2-sided Spearman rank correlations (ρ), given the small sample size and the bounded or ordinal nature of several variables. Because no formal a priori sample size calculation was conducted, we supplemented the randomized outcome analysis with a post hoc sample size sensitivity analysis based on the observed posttest Cliff δ, converting Cliff δ to area under the curve and to an approximate standardized mean difference. Interrater reliability of the expert reference standard was assessed for each PBL dimension using a 2-way mixed-effects absolute-agreement intraclass correlation coefficient (ICC). Because convergent-validity analyses used mean expert ratings, we report ICC(A,3), the reliability of the 3-rater mean.

Ethical Considerations

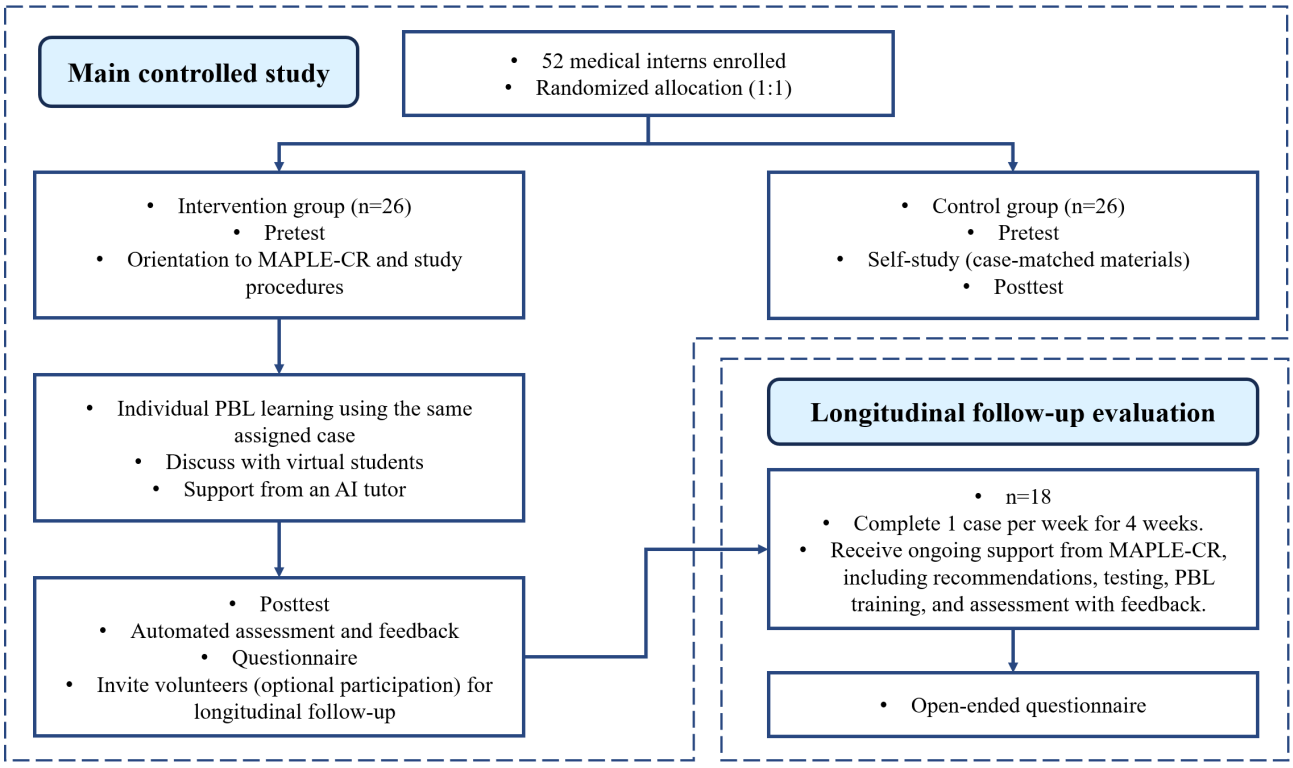
The study was conducted in accordance with the Declaration of Helsinki and approved by the Ethics Committee of Tongde Hospital of Zhejiang Province (KTSC2025080). The trial was retrospectively registered with the Chinese Clinical Trial Registry (ChiCTR2500115799; registered on December 31, 2025). Participation was voluntary, and all participants provided informed consent before study participation, including consent to the cloud-based AI processing and research use of pseudonymized dialogue, learning, and assessment data. Data were deidentified before analysis. This was a minimal-risk educational study using simulated cases.

Results

Results of the Randomized Controlled Study

Fifty-two interns were randomized (intervention: n=26, control: n=26), and 18 completed the voluntary follow-up (Figure 3). Baseline sex and age were comparable between the intervention and control groups (female/male: 18/8 vs 20/6, Fisher exact test, *P*=.76; median age 23, IQR 23-24 vs 24, IQR 23-24, Mann-Whitney *U* test, *P*=.31). The follow-up subgroup did not differ from the intervention group in sex (female/male: 11/7 vs 18/8, Fisher exact test *P*=.75) or age (median 23, IQR 23-24 vs 23, IQR 23-24, Mann-Whitney *U* test, *P*=.78).

Figure 3. Study design and participant flow through the randomized evaluation and voluntary feasibility follow-up. Fifty-two interns were randomized 1:1 to Multiagent Problem-based Learning Environment for Clinical Reasoning (MAPLE-CR) or case-matched self-study and were included in the randomized analysis. A voluntary subset of MAPLE-CR participants subsequently completed repeated follow-up sessions across additional cases. PBL: problem-based learning.

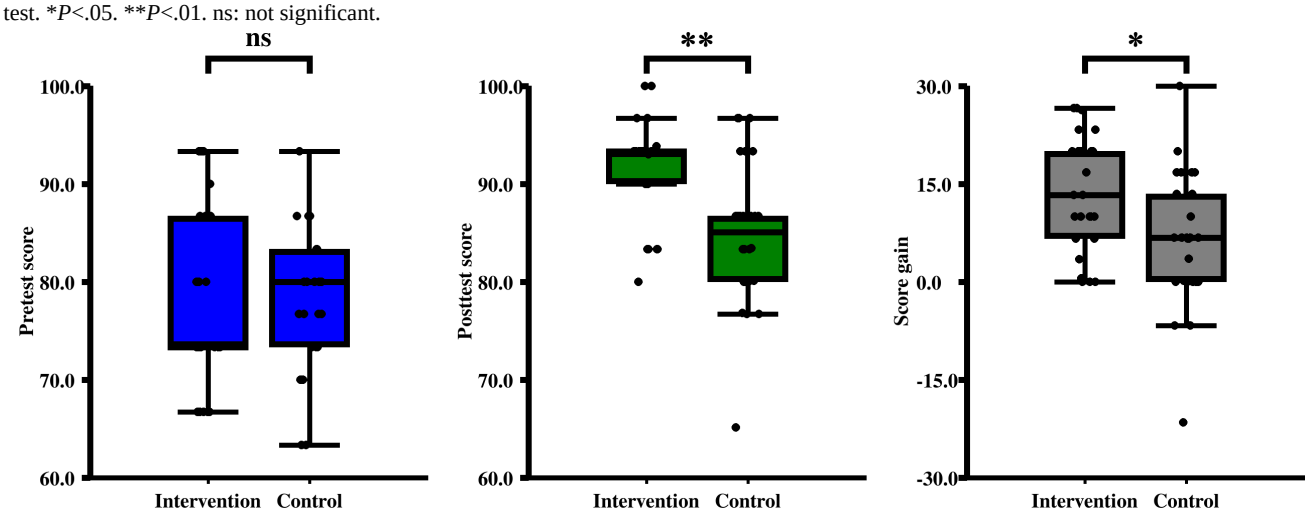


Baseline pretest scores were comparable between the intervention and control groups (Mann-Whitney *U* test, *P*=.56; Figure 4), suggesting no meaningful baseline imbalance in initial knowledge. After the session, the intervention group achieved

XSL•FO
RenderX

higher posttest scores than the control group (Mann-Whitney *U* test, *P*=.001), with a between-group median difference of 8.25 (95% CI 6.60 - 11.60) points. The corresponding effect size was Cliff δ=0.506 (95% CI 0.21-0.76), indicating a distributional shift toward higher posttest scores in the intervention group. Consistently, the intervention group demonstrated a greater score gain (posttest minus pretest) than the control group (Mann-Whitney *U* test, *P*=.01), with a between-group median difference of 6.60 (95% CI 1.60 - 14.95) points. The CI for this gain difference was wide, indicating uncertainty in the magnitude of the incremental benefit. Within-group analyses indicated that the control group also improved from pretest to posttest (Wilcoxon signed-rank test, *P*=.006), with a median gain of +6.70 (95% CI 0.10-10.05) points. This pattern is consistent with a measurable learning effect under the self-study condition rather than mere test-retest fluctuation. Overall, between-group separation was observed for posttest performance and gain under the present comparator condition. Because the control group also improved, this difference should be interpreted as a MAPLE-CR-associated short-term difference rather than evidence that the effect was attributable specifically to PBL scaffolding mechanisms.

Figure 4. Group distributions of pretest scores, posttest scores, and score gain. Scores are on a 0 to 100 scale. Data were analyzed using the Mann-Whitney *U* test. *P<.05. **P<.01. ns: not significant.



The observed posttest effect was large in this sample. In the sample size sensitivity analysis, the observed effect magnitude exceeded the approximate large-effect threshold detectable with 26 participants per group under 2-sided α=.05. However, the confidence interval around Cliff δ was wide, and its lower bound corresponded to a substantially smaller standardized effect. The current sample therefore cannot exclude smaller but still educationally meaningful effects. Because this was a descriptive sensitivity analysis, these findings should be interpreted as evidence of a large short-term difference under the present study conditions, rather than as definitive efficacy evidence.

Table 2 summarizes postsession questionnaire responses following the MAPLE-CR intervention, presenting full-sample results as the main descriptive analysis and filtered results as a sensitivity analysis. Across the 12 positively worded items (Q1-Q12), full-sample ratings were consistently favorable, with means ranging from 4.19 to 4.62, indicating high satisfaction, sustained involvement, strong perceived support, and enhanced self-efficacy. Because Q13-Q15 were reverse-keyed, we conducted a nonprespecified sensitivity analysis excluding 5 questionnaires with all 15 ratings ≥3 to examine possible acquiescent responding; this criterion did not define invalid responses, and full-sample results remained primary. Filtered means changed minimally for Q1-Q11 (absolute mean difference ≤0.17; $|d|\leq 0.23$). In contrast, the 3 perceived cognitive-load-related items (Q13-Q15), reflecting perceived mental effort, perceived information overload, and interface-related distraction, showed lower means, greater dispersion, and larger shifts after filtering (mean difference=−0.31 to −0.54; $d=-0.27$ to -0.53), suggesting greater heterogeneity in perceived load and distraction. Overall, the questionnaire results support strong acceptability and perceived educational value of MAPLE-CR; however, perceived cognitive-load-related items remained salient even after filtering. This pattern highlights a broader consideration for AI-supported education: in PBL-like settings, AI may add value less by reducing cognitive demands and more by helping to orchestrate complex discourse and make reasoning visible, while minimizing avoidable interface or coordination demands that can divert effort from learning.

XSL•FO
RenderX

Table . Per-item questionnaire responses (5-point Likert scale) in the full sample (n=26) and filtered sample (n=21)[a].

Question	Full sample, mean (variance)	Filtered, mean (variance)	Mean difference	Cohen *d*
Learning satisfaction				
Q1. I was satisfied with the overall PBL[b] system experience.	4.19 (0.617)	4.10 (0.658)	−0.10	−0.12
Q2. The recommended cases matched my learning needs.	4.19 (0.386)	4.14 (0.408)	−0.05	−0.08
Q3. The system helped me learn the clinical reasoning for this case.	4.27 (0.581)	4.10 (0.562)	−0.17	−0.23
Q4. The test questions effectively assessed my knowledge.	4.19 (0.386)	4.10 (0.372)	−0.10	−0.16
Q5. The feedback was accurate and helpful.	4.62 (0.390)	4.57 (0.435)	−0.04	−0.07
Sense of involvement				
Q6. I stayed engaged and actively participated in reasoning/decisions.	4.19 (0.617)	4.10 (0.658)	−0.10	−0.12
Q7. I could express my views and lead my own reasoning process.	4.50 (0.481)	4.38 (0.522)	−0.12	−0.17
Perceived support				
Q8. Virtual students prompted deeper thinking and participation.	4.38 (0.544)	4.33 (0.508)	−0.05	−0.07
Q9. The tutor supported me and kept the discussion on track.	4.31 (0.444)	4.19 (0.440)	−0.12	−0.18
Q10. The system provided timely prompts/corrections when needed.	4.35 (0.380)	4.29 (0.299)	−0.06	−0.10
Self-efficacy				
Q11. I felt more confident analyzing the case and justifying a diagnosis.	4.35 (0.457)	4.29 (0.395)	−0.06	−0.09
Q12. I can apply the learned reasoning strategies to similar cases.	4.50 (0.404)	4.52 (0.345)	+0.02	+0.04
Cognitive load				
Q13. Completing my tasks required substantial mental effort.	3.69 (1.444)	3.38 (1.283)	−0.31	−0.27
Q14. I often felt information overload during the discussion.	3.15 (2.053)	2.76 (1.705)	−0.39	−0.28
Q15. The system prompts or interface often distracted my attention.	2.54 (1.556)	2.00 (0.381)	−0.54	−0.53

[a]Full-sample results constitute the main descriptive questionnaire analysis. Filtered results comprise the nonprespecified sensitivity analysis and exclude 5 questionnaires with no ratings below the scale midpoint (all 15 items ≥3); mean differences and Cohen *d* are computed as filtered minus full sample.

[b]PBL: problem-based learning.

XSL•FO
RenderX

Among the 26 MAPLE-CR intervention sessions, transcript-based formative assessment was summarized as supplementary process-oriented evidence aligned with the scaffolding framework across the 5 PBL-relevant dimensions (Figure S1 in Multimedia Appendix 4). KA and CR had the highest mean scores (4.4 and 3.7) and relatively high minima (3.5 and 3.0), suggesting that most learners articulated coherent and largely correct diagnostic reasoning. By contrast, interaction-oriented dimensions showed greater variability: CC and SR had lower minima (both 1.5), and AP had the lowest central tendency (mean 2.1; minimum 0.5). Together, these findings suggest that even when learners reason accurately, they may not consistently externalize that reasoning through structured discussion, synthesis, and proactive engagement—core behaviors in PBL. Notably, the CC dimension reflects learner responsiveness within AI-mediated tutor/peer exchanges rather than real peer-to-peer teamwork.

In addition, we examined convergent validity by comparing the automated formative scores with independent expert ratings across 26 sessions (3 raters per session; 0 - 5 scale; Table S1 in Multimedia Appendix 4). The expert panel consisted of 3 attending physicians involved in postgraduate medical education, each with at least 10 years of clinical teaching experience and prior experience in PBL facilitation or CR assessment. Expert raters assessed deidentified session transcripts independently and were blinded to each other's ratings; automated formative scores were not used during expert rating. For the 4 evaluator-scored dimensions, the human experts used the same construct definitions, evidence rules, and 0 to 5 scoring anchors as the evaluator agent; the complete materials are provided in Multimedia Appendix 1. Interrater reliability of the mean expert ratings was moderate to good across dimensions (ICC[A,3]=0.729 - 0.851): AP 0.851, CR 0.766, KA 0.729, CC 0.807, and SR 0.796. We quantified agreement using Spearman rank correlations (ρ) between automated scores and the mean expert rating per session, and absolute agreement using MAE. Agreement analyses showed strong rank-order consistency for AP (ρ=0.88, *P*<.001; MAE=0.40), SR (ρ=0.93, *P*<.001; MAE=0.50), and CR (ρ=0.78, *P*=.007; MAE=0.45). KA demonstrated moderately high agreement (ρ=0.81, *P*=.005; MAE=0.47). CC exhibited comparatively weaker, though still positive, agreement (ρ=0.71, *P*=.02) and the largest absolute deviation (MAE=0.68). Overall, these results provide preliminary evidence of convergence between the automated assessment and expert judgments in this small, single-study sample, particularly for participation, reflection, and reasoning. Because the CC reference standard showed good reliability (ICC[A,3]=0.807), the comparatively weaker automated-expert agreement for this dimension suggests that communication-related behaviors may benefit from further rubric calibration and model refinement.

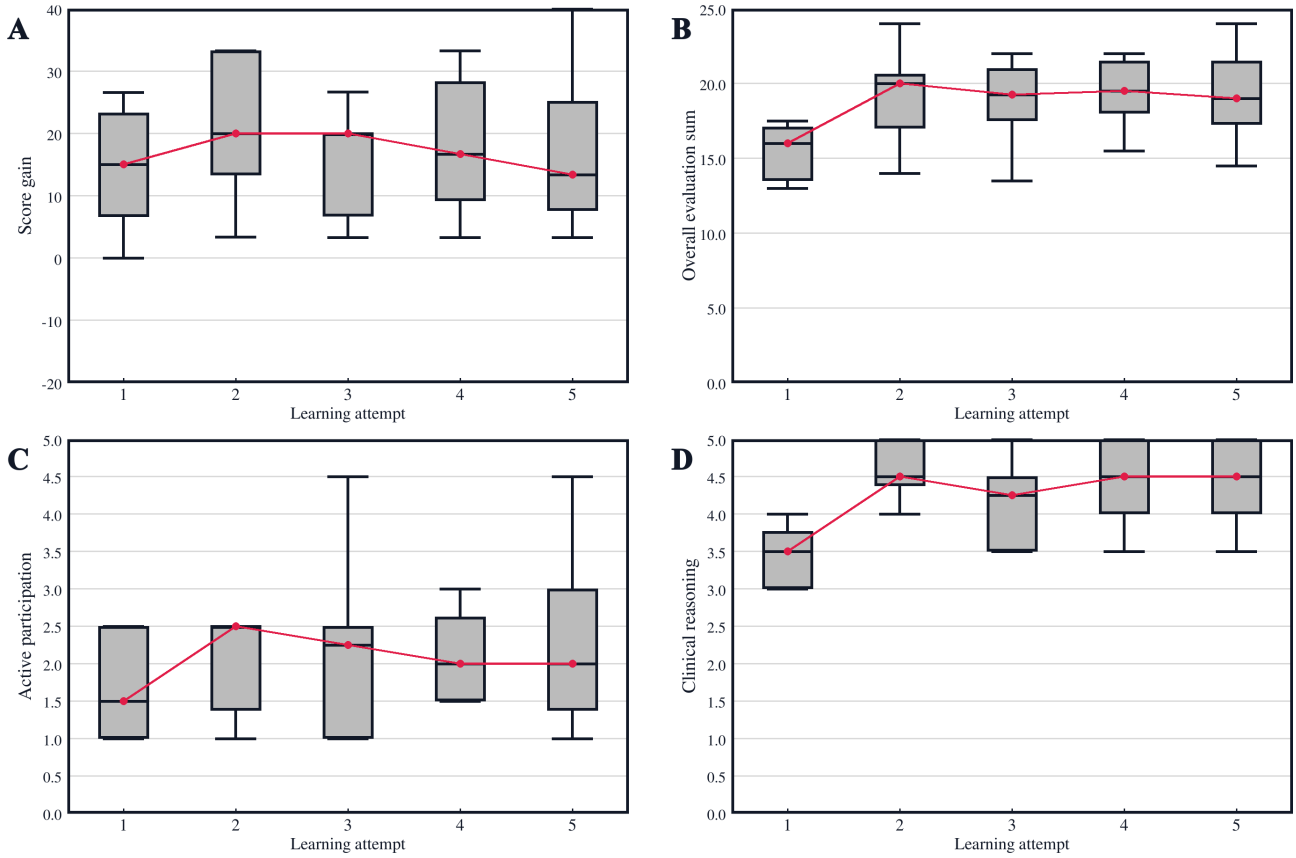
Figure S2 in Multimedia Appendix 4 presents an exploratory Spearman rank correlation analysis across session-level evaluation measures within the intervention group. Pairwise complete observations were used: correlations involving questionnaire composites used 21 matched sessions, whereas correlations among test scores and formative metrics used 26 sessions. The 3 perceived cognitive-load-related items (Q13-Q15) were summed as an exploratory cognitive load (COG) composite, and the remaining questionnaire items (Q1-Q12) were summed as an overall experience (EXP) composite. Baseline performance was strongly negatively associated with score gain (pretest score vs score gain, ρ=−0.86), consistent primarily with ceiling effects and regression to the mean. Several transcript-based formative dimensions were positively interrelated, especially AP-SR (ρ=0.77), CR-SR (ρ=0.88), and CC-SR (ρ=0.81), indicating that interactional, reasoning, reflective, and regulatory features tended to co-occur within MAPLE-CR sessions. In contrast, the questionnaire composites showed weak associations with formative metrics and knowledge outcomes (eg, COG-KA, ρ=−0.07; EXP-posttest, ρ=0.19), suggesting that perceived cognitive load and overall experience should be interpreted as subjective learner-experience indicators rather than direct proxies for performance or process quality. Because no multiplicity adjustment was applied, these correlations should be interpreted descriptively rather than as confirmatory evidence.

Voluntary Follow-Up: Descriptive Repeated-Use Patterns

We descriptively examined repeated-use patterns in within-session test score gain, CR, AP, and the overall evaluation sum across 5 attempts in the voluntary follow-up sample (Figure 5; n=18). Results for KA, CC, and SR are reported in Figure S3 in Multimedia Appendix 4.

XSL•FO
RenderX

Figure 5. Descriptive repeated-use patterns across 5 Multiagent Problem-based Learning Environment for Clinical Reasoning (MAPLE-CR) attempts in the voluntary follow-up sample. Panels show (A) test score gain, (B) overall formative evaluation sum, (C) active participation, and (D) clinical reasoning. Box plots summarize the distribution at each attempt, with the center line indicating the median and the box spanning the 25th to 75th percentiles; the red line connects the median values across attempts to aid visual interpretation of the follow-up pattern.



Within-session pretest-to-posttest score differences were positive across all 5 attempts in the voluntary follow-up sample. The median within-session score differences were 15.00 (IQR 6.60-23.30; bootstrap 95% CI 8.30-21.65), 20.00 (IQR 13.30-33.30; bootstrap 95% CI 13.30-26.65), 20.00 (IQR 6.70-20.00; bootstrap 95% CI 6.90-20.00), 16.70 (IQR 9.18-28.35; bootstrap 95% CI 10.00-26.70), and 13.40 (IQR 7.60-25.20; bootstrap 95% CI 8.95-23.50) across attempts 1 through 5, respectively. All CIs were above zero, indicating positive within-session score differences at each attempt.

For formative outcomes, CR increased markedly from attempt 1 to attempt 2 (median 3.50 to 4.50; Hodges-Lehmann median difference=1.00, 95% CI 1.00 - 1.50; Wilcoxon P<.001; Cohen d=1.739), indicating a higher transcript-based CR score at attempt 2 than at attempt 1 in this descriptive follow-up analysis. In contrast, AP showed a smaller, nonsignificant change (median 1.50 to 2.50; Hodges-Lehmann median difference=0.19, 95% CI 0.00 - 1.00; Wilcoxon P=.15; d=0.411), indicating a smaller and more heterogeneous observed change in engagement. The overall evaluation sum also increased from attempt 1 to attempt 2 (median 16.00 to 20.00; Hodges-Lehmann median difference=3.50, 95% CI 2.50 - 5.50; Wilcoxon P<.001; d=1.292), indicating a higher aggregate formative score at attempt 2 than at attempt 1.

From attempt 2 onward, no statistically significant within-subject trends were detected for CR (Friedman χ^2_3=6.484, P=.09), AP (Friedman χ^2_3=2.088, P=.55), or the overall sum (Friedman χ^2_3=3.034, P=.39). These nonsignificant results should not be interpreted as evidence of stabilization; rather, they indicate that no statistically significant later trend was detected in this small voluntary follow-up sample. A sample size sensitivity calculation indicated that detecting a smaller but educationally meaningful within-participant change of d=0.50 with 2-sided α=.05 and 80% power would require approximately 34 completed follow-up participants, exceeding the available follow-up sample (n=18).

Additionally, student case-study pathways are visualized in Figure S4 in Multimedia Appendix 4. Although all learners began with the same thoracic and airway surgery case, half diversified into other systems in the second attempt. By the fifth attempt, students had collectively explored over 9 distinct medical domains, with some choosing to revisit similar systems (eg, hepatobiliary or cardiovascular) and others branching into entirely new areas (eg, pediatrics or oncology). These patterns suggest heterogeneity in follow-up case selection; whether this reflected learner interests, perceived needs, or recommendation effects could not be determined from the present data.

We also recorded the time learners spent on each PBL discussion, excluding testing and assessment (Figure S5 in Multimedia Appendix 4). The overall median was 67.39 (IQR 54.7-90.0) minutes. The median discussion duration decreased from 70.87 (IQR 58.5-93.2) minutes at attempt 1 to 54.87 (IQR 48.1-82.9) minutes at attempt 5, a 22.6% reduction. Despite session-to-session fluctuations, discussions were shorter by the

XSL•FO
RenderX

fifth attempt; however, this pattern may reflect increasing familiarity, case differences, or reduced engagement.

Open-Ended Questionnaire

To contextualize learners' perceived strengths and limitations of MAPLE-CR, we descriptively summarized the open-ended feedback entries and grouped them into nonmutually exclusive categories. The complete categorized comments are provided in Multimedia Appendix 5.

Positive feedback most often described structured CR support (5 feedback entries), perceived learning gains (4 entries), reduced pressure and freer expression compared with classroom PBL (4 entries), and content coverage or case realism (3 entries). These comments suggested that learners perceived the simulated PBL format as immersive, flexible, and helpful for organizing diagnostic and management reasoning.

Areas for improvement included interaction depth, flow, or tutor questioning (3 feedback entries), assessment or questionnaire design (2 entries), content repetition or concision (2 entries), and interface usability or navigation (2 entries). Single feedback entries also noted conversational accuracy or AI misunderstanding, time or session-control burden, and requests for additional learning-tool support such as discussion filtering or mind-map functions.

Overall, these open-ended responses provided descriptive support for perceived usefulness and learner acceptability while identifying concrete areas for refinement in conversational accuracy, interaction design, assessment design, and workload.

Discussion

Principal Findings

This exploratory randomized educational evaluation provides preliminary evidence that MAPLE-CR is a feasible and educationally valuable AI-supported PBL environment for CR training. In the randomized component, learners using MAPLE-CR achieved higher posttest scores and greater pretest-posttest gains than those in the self-study control, suggesting a large short-term difference under the present study conditions rather than definitive efficacy evidence. Learner-reported outcomes indicated high acceptability and perceived educational value across satisfaction, involvement, perceived support, and self-efficacy. Because engagement with generative AI may vary with learner characteristics and implementation context, these acceptability findings should be interpreted within the present population and setting [30,31]. Notably, however, learners also reported relatively high ratings on cognitive-load items. This pattern should not be interpreted as evidence that reducing real peer coordination is educationally beneficial or that AI-mediated PBL can replace human group interaction. Rather, for MAPLE-CR as a supplementary PBL learning environment, the design challenge is to preserve productive reasoning and dialogic engagement while minimizing avoidable interface or workflow demands. From a cognitive load perspective, this means preserving productive challenge associated with reasoning and explanation [32,33] while reducing extraneous load introduced by the learning interface and workflow structure [34]. From a social constructivist perspective, these findings are consistent with the possibility that AI-driven scaffolding can function as a provisional bridge within the learner's Zone of Proximal Development [35]. One interpretation is that MAPLE-CR's orchestration of the PBL process, including structured summaries, next-step prompts, and feedback loops, may have helped learners direct effort toward higher-level reasoning, although the specific contribution of each scaffolding component requires further investigation.

The control condition produced a measurable median gain of +6.70 (IQR 0.00-13.40) points, whereas the between-group median difference in gain was 6.60 points (bootstrap 95% CI 1.60-14.95). Using these values as a descriptive anchor, a substantial portion of the short-term pretest-posttest improvement was also observed under structured exposure to the same case materials, while the additional separation represents the incremental difference observed for MAPLE-CR under this comparator. This comparison is descriptive only and should not be interpreted as a causal partitioning of the effect. The present design evaluated MAPLE-CR as an integrated learning package. The observed between-group difference may therefore reflect the combined influence of platform use, orientation, AI-mediated interaction and feedback, additional structured exposure, cognitive and regulatory scaffolding, novelty, and motivational engagement rather than the isolated effect of the AI-PBL scaffolding mechanisms; nor does the design establish noninferiority to faculty-facilitated, peer-driven PBL.

Across the voluntary 4-week follow-up, descriptive patterns showed higher CR and aggregate formative scores at attempt 2 than at attempt 1, followed by statistically uncertain later trajectories. Given the absence of a control group, self-selected participation, heterogeneous case pathways, and the small sample, these observations do not establish maintenance, stabilization, or sustained learning; they could reflect adaptation, ceiling or novelty effects, effort, case heterogeneity, or limited power [36,37]. They instead support the feasibility of repeated MAPLE-CR use across different cases and identify patterns that warrant evaluation in controlled longitudinal studies, including studies of adaptive prompts or progressively more complex cases [21].

In addition, a key pattern was the dissociation between relatively strong knowledge/reasoning performance and more heterogeneous interactional performance (eg, participation, collaboration, and reflection). This aligns with prior work suggesting that competence in CR does not automatically translate into observable group-process skills, particularly in technology-mediated settings where turn-taking norms and role clarity are less salient than in facilitator-led discussions [22,38]. Accordingly, AI-enabled PBL designs may need to scaffold not only what learners think (eg, prompts for hypothesis generation and justification) but also how learners participate and co-construct understanding (eg, turn-taking, accountability structures, and synthesis routines). Within the present study, MAPLE-CR represents an empirically evaluated implementation of conversational AI that spans multiple phases of the PBL cycle rather than serving as an occasional question-answering aid. The findings also indicate where additional process support is needed.

XSL•FO
RenderX

Implications for Medical Education

Beyond our local internship context, a key design implication of MAPLE-CR is a set of potentially transferable scaffolding features for supporting scalable CR practice: embedding cognitive prompts that require explanation and revision, social structures that elicit multivoice challenge rather than simple learner-AI dyads, and regulatory supports that make phases and feedback loops explicit. When trained facilitators, standardized patients, or stable small-group scheduling are difficult to sustain, such a pattern may offer a scalable supplement for presession rehearsal, between-session deliberate practice, or postsession consolidation within blended curricula [39]. Because simulated group-process behaviors varied across sessions, AI is unlikely to fully reproduce the nuance of human facilitation [23]; implementation may therefore benefit from brief learner orientation (expectations for participation, justification, and turn-taking) and a short postsession debrief to normalize reflection and calibrate engagement. In future implementations, transcript summaries and formative analytics could be evaluated as tools to help faculty identify sessions with recurrent needs (eg, persistently low collaboration or superficial justification) and provide focused coaching without routine full-transcript review. At its current maturity, MAPLE-CR is best positioned as a scalable, formative, low-stakes supplement for structured PBL-like practice rather than as a replacement for faculty-facilitated PBL; further multisite evidence is needed before considering high-stakes or summative uses, particularly across languages, curricula, and assessment regimes.

Safety

MAPLE-CR is an educational tool rather than a clinical decision support system and is intended for formative, low-stakes learning rather than summative decision-making. Tutor and evaluator functions are bounded by case objectives and curated learning resources, and feedback emphasizes reasoning processes (eg, justification, revision, reflection) rather than prescriptive management recommendations. Cases and prompts are designed for training purposes and do not involve identifiable patient information; discussion transcripts and learning records are stored and analyzed in deidentified form for educational feedback and research. Although no safety incidents were observed in this study, future iterations will strengthen safeguards to mitigate misinformation and over-reliance and will maintain ongoing monitoring and a clear reporting pathway for any concerning outputs [10,40,41].

Limitations and Suggestions for Future Research

This study has limitations. First, no formal a priori sample size calculation was conducted. Although the sensitivity analysis showed that the observed effect magnitude exceeded the approximate large-effect threshold detectable with the available sample, this descriptive analysis does not establish prospective power sufficiency; the study was not prospectively powered as a confirmatory efficacy trial, and the precision of the estimated effect remains limited. The findings should therefore be confirmed in larger, prospectively powered, multicenter trials with stronger active comparators. The adapted postsession questionnaire was not validated specifically for AI-supported PBL, so learner-experience findings should be interpreted descriptively. Second, the trial was retrospectively registered, which limits full compliance with prospective RCT registration expectations and should be considered when interpreting the randomized component. Third, the control condition was self-study, which represents a relatively modest active comparator; this comparator did not isolate the contribution of specific PBL scaffolding mechanisms from other features of MAPLE-CR, such as AI-mediated dialogue, simulated peer/tutor interaction, novelty, or motivational engagement; future studies should benchmark MAPLE-CR against stronger comparators that more closely reflect established PBL pedagogy (eg, facilitator-led small-group PBL), near-peer facilitation, or other structured instructional modalities [15,42]. The conditions were also not matched for total structured exposure because only the MAPLE-CR group received the 40-minute orientation. The randomized component also used a single standardized lung cancer/thoracic-airway surgery case to ensure comparable exposure across groups; therefore, the observed effect may not generalize across clinical domains, case complexity levels, or scenario types. The voluntary follow-up supported repeated use across additional cases but was not designed to determine whether MAPLE-CR effects differ by case type or complexity. In addition, stochastic LLM generation may have introduced variability in the tutor and peer responses that learners received, creating a potential source of within-condition interference that was not explicitly modeled. Fourth, outcomes emphasized immediate knowledge performance and learner experience; although the follow-up suggested feasibility for repeated cross-case use, participation was voluntary and may overestimate engagement and perceived benefit. The follow-up included only 18 participants and was not powered for confirmatory inference about longer-term learning trajectories. Fifth, we did not include standardized external measures of retention and transfer (eg, objective structured clinical examination-style assessments) [43]. Although the transcript-derived formative dimensions were aligned with the proposed scaffolding framework and showed preliminary convergence with expert ratings, they were not prospectively designed to test mediation or isolate specific cognitive, social-interactional, or regulatory mechanisms. Future studies should incorporate prespecified process measures to examine how specific scaffolding functions contribute to learning outcomes.

Beyond these study-level limitations, future research should move from asking whether AI-supported PBL "works" to testing how and for whom it works. Mechanism-focused studies could examine whether improvements are mediated by process indicators aligned to the proposed cognitive, social-interactional, and regulatory functions (eg, quality of justification and revision, participation patterns and responsiveness, adherence to PBL phases, and uptake of feedback across cases). Work is also needed to identify boundary conditions and learner-level moderators (eg, baseline knowledge, self-efficacy, language/communication confidence, and AI literacy), given the observed heterogeneity in group-process behaviors. Finally, the higher scores observed at attempt 2 and statistically uncertain later trajectories suggest a need to investigate adaptive progression and scaffold fading—such as progressively more complex cases, calibrated prompts, or feedback designs—under

XSL•FO
RenderX

controlled longitudinal conditions [24]. Multisite and real-world implementation studies should evaluate feasibility, faculty workload, and sustainability alongside learning outcomes, and clarify appropriate use as formative support rather than high-stakes assessment.

Conclusions

This study provides preliminary evidence that MAPLE-CR is a viable AI-supported scaffold for PBL and suggests that generative AI can serve a pedagogical role beyond functioning as a mere information source. MAPLE-CR was associated with larger short-term knowledge gains than self-study and showed high acceptability and feasible repeated use, but the design did not isolate specific scaffolding mechanisms. At the same time, the finding that gains in knowledge and CR may emerge earlier than improvements in observable participation and collaborative process skills highlights the complexity of AI-mediated instructional design. As a scalable supplement, MAPLE-CR may provide additional opportunities for structured PBL-like practice when faculty facilitation and repeated small-group sessions are difficult to sustain. These findings require confirmation in larger, prospectively powered, multicenter trials with stronger active comparators and longer-term measures of retention, transfer, and cross-case generalizability.

Acknowledgments

Generative AI (ChatGPT, GPT-5.6 Sol, OpenAI) was used solely to improve the language and readability of the manuscript. All AI-assisted edits were reviewed by the authors, who take full responsibility for the final content. We also thank all medical interns who participated in the study for their time, effort, and valuable engagement.

Funding

This work was supported by the National Natural Science Foundation of China (grant 62577049).

Data Availability

The deidentified data supporting the reported results are available from the corresponding author on reasonable request. The multiagent problem-based learning environment for clinical reasoning code is available in a public GitHub repository [44].

Authors' Contributions

ZZ led system development and data processing. BW and ST led data collection. ZZ and BW drafted the initial manuscript. ST, HH, and PH supervised the study, reviewed the manuscript, and contributed substantive intellectual revision. All authors contributed to the study design and interpretation of findings, approved the final manuscript, and agreed to be accountable for all aspects of the work.

Conflicts of Interest

ZZ led the development of MAPLE-CR, the AI-supported PBL platform evaluated in this study; this role is disclosed for transparency. All authors declare no competing interests.

Multimedia Appendix 1
Detailed description of the technical implementation of multiagent problem-based learning environment for clinical reasoning.
[DOCX File, 28 KB - mededu_v12i1e95039_app1.docx]

Multimedia Appendix 2
Video demonstration of the MAPLE-CR system workflow and user interface.
[MP4 File, 56462 KB - mededu_v12i1e95039_app2.mp4]

Multimedia Appendix 3
Blueprint of the pretest and posttest design.
[DOCX File, 32 KB - mededu_v12i1e95039_app3.docx]

Multimedia Appendix 4
Additional experimental and follow-up results.
[PDF File, 3128 KB - mededu_v12i1e95039_app4.pdf]

Multimedia Appendix 5
Descriptive categorization of open-ended feedback.
[DOCX File, 18 KB - mededu_v12i1e95039_app5.docx]

XSL•FO
RenderX

Checklist 1
CONSORT-eHEALTH checklist (V 1.6.1).
[PDF File, 2951 KB - mededu_v12i1e95039_app6.pdf]

References

1. Hmelo-Silver CE. Problem-based learning: what and how do students learn? Educ Psychol Rev 2004 Sep;16(3):235-266. [doi: 10.1023/B:EDPR.0000034022.16470.f3]
2. Dineen M, Lazarus MD, Stephens GC. Uncertainty experienced by newly qualified doctors during the transition to internship. Med Educ 2025 Oct;59(10):1079-1093. [doi: 10.1111/medu.15692] [Medline: 40156179]
3. Yang W, Li H, Su A, Ding L. Application of problem based learning (PBL) and case based learning (CBL) in the teaching of international classification of diseases encoding. Sci Rep 2023 Sep 14;13(1):15220. [doi: 10.1038/s41598-023-42175-1] [Medline: 37709817]
4. Dochy F, Segers M, Van den Bossche P, Gijbels D. Effects of problem-based learning: a meta-analysis. Learn Instr 2003 Oct;13(5):533-568. [doi: 10.1016/S0959-4752(02)00025-7]
5. Schmidt HG, Rotgans JI, Yew EHJ. The process of problem-based learning: what works and why. Med Educ 2011 Aug;45(8):792-806. [doi: 10.1111/j.1365-2923.2011.04035.x] [Medline: 21752076]
6. Dolmans DHJM, Gijselaers WH, Moust JHC, de Grave WS, Wolfhagen IHAP, van der Vleuten CPM. Trends in research on the tutor in problem-based learning: conclusions and implications for educational practice and research. Med Teach 2002 Mar;24(2):173-180. [doi: 10.1080/01421590220125277] [Medline: 12098437]
7. Jones NL, Peiffer AM, Lambros A, et al. Developing a problem-based learning (PBL) curriculum for professionalism and scientific integrity training for biomedical graduate students. J Med Ethics 2010 Oct;36(10):614-619. [doi: 10.1136/jme.2009.035220] [Medline: 20797979]
8. Monrouxe LV, Rees CE. The socialisation of mistreatment in the healthcare workplace: moving beyond narrative content to analyse educator data as discourse. Med Educ 2023 Oct;57(10):882-885. [doi: 10.1111/medu.15122] [Medline: 37183307]
9. Farquhar S, Kossen J, Kuhn L, Gal Y. Detecting hallucinations in large language models using semantic entropy. Nature 2024 Jun;630(8017):625-630. [doi: 10.1038/s41586-024-07421-0] [Medline: 38898292]
10. Thesen T, Park SH. A generative AI teaching assistant for personalized learning in medical education. NPJ Digit Med 2025 Nov 4;8(1):627. [doi: 10.1038/s41746-025-02022-1] [Medline: 41188616]
11. Shalong W, Yi Z, Bin Z, et al. Enhancing self-directed learning with custom GPT AI facilitation among medical students: a randomized controlled trial. Med Teach 2025 Jul;47(7):1126-1133. [doi: 10.1080/0142159X.2024.2413023] [Medline: 39425996]
12. Nguyen NP, Tran P. Bridging the mentorship divide: how large language models could reshape medical workforce equity. NPJ Digit Med 2026 Jan 9;9(1):29. [doi: 10.1038/s41746-025-02167-z] [Medline: 41513946]
13. Holderried F, Stegemann-Philipps C, Herrmann-Werner A, et al. A language model-powered simulated patient with automated feedback for history taking: prospective study. JMIR Med Educ 2024 Aug 16;10(1):e59213. [doi: 10.2196/59213] [Medline: 39150749]
14. Xu X, Liu S, Zhu L, et al. Development and evaluation of a retrieval-augmented large language model framework for enhancing endodontic education. Int J Med Inform 2025 Nov;203:106006. [doi: 10.1016/j.ijmedinf.2025.106006] [Medline: 40479778]
15. Wang R, Liu C, Yi F. Reconfiguring tutor-peer roles to foster critical thinking in medical PBL assisted by tutor-managed ChatGPT. Med Educ 2026 Aug;60(8):894-907. [doi: 10.1111/medu.70100] [Medline: 41308655]
16. Hui Z, Zewu Z, Jiao H, Yu C. Application of ChatGPT-assisted problem-based learning teaching method in clinical medical education. BMC Med Educ 2025 Jan 11;25(1):50. [doi: 10.1186/s12909-024-06321-1] [Medline: 39799356]
17. Barrows HS. Problem‐based learning in medicine and beyond: a brief overview. New Dir Teach Learn 1996 Dec;1996(68):3-12. [doi: 10.1002/tl.37219966804]
18. Basil M, Ahmed W, Hajeomar R, Strawbridge J, Lynch M, Mukhalalati B. A scoping review of the use of generative artificial intelligence tools in health profession education. BMC Med Educ 2026 Jan 23;26(1):291. [doi: 10.1186/s12909-025-08527-3] [Medline: 41578287]
19. Pangastuti D, Widiasih N, Soemantri D. Piloting a constructive feedback model for problem-based learning in medical education. Korean J Med Educ 2022 Jun;34(2):131-143. [doi: 10.3946/kjme.2022.225] [Medline: 35676880]
20. Izquierdo-Condoy JS, Arias-Intriago M, Tello-De-la-Torre A, Busch F, Ortiz-Prado E. Generative artificial intelligence in medical education: enhancing critical thinking or undermining cognitive autonomy? J Med Internet Res 2025 Nov 3;27:e76340. [doi: 10.2196/76340] [Medline: 41183320]
21. Vogel F, Wecker C, Kollar I, Fischer F. Socio-cognitive scaffolding with computer-supported collaboration scripts: a meta-analysis. Educ Psychol Rev 2017 Sep;29(3):477-511. [doi: 10.1007/s10648-016-9361-7]
22. Kreijns K, Kirschner PA, Jochems W. Identifying the pitfalls for social interaction in computer-supported collaborative learning environments: a review of the research. Comput Human Behav 2003 May;19(3):335-353. [doi: 10.1016/S0747-5632(02)00057-2]

XSL•FO
RenderX

23. Kassab SE, Hamdy H, Mamede S, Schmidt H. Influence of tutor interventions and group process on medical students' engagement in problem-based learning. Med Educ 2024 Nov;58(11):1315-1323. [doi: 10.1111/medu.15387] [Medline: 38563548]
24. Belland BR, Kim C, Hannafin MJ. A framework for designing scaffolds that improve motivation and cognition. Educ Psychol 2013 Oct;48(4):243-270. [doi: 10.1080/00461520.2013.838920] [Medline: 24273351]
25. Yao S, Zhao J, Yu D, et al. ReAct: synergizing reasoning and acting in language models. Presented at: 11th International Conference on Learning Representations, ICLR 2023; May 1-5, 2023 URL: https://collaborate.princeton.edu/en/publications/react-synergizing-reasoning-and-acting-in-language-models/ [accessed 2026-08-18]
26. Wu Q, Bansal G, Zhang J, et al. AutoGen: enabling next-gen LLM applications via multi-agent conversation. Presented at: First Conference on Language Modeling (COLM 2024); Oct 7-9, 2024 URL: https://www.microsoft.com/en-us/research/publication/autogen-enabling-next-gen-llm-applications-via-multi-agent-conversation-framework/ [accessed 2026-08-18]
27. Alibaba cloud model studio. Alibaba Cloud. URL: https://www.alibabacloud.com/help/en/model-studio/ [accessed 2026-07-01]
28. Ekelund K, O'Regan S, Dieckmann P, Østergaard D, Watterson L. Evaluation of the simulation based training quality assurance tool (SBT-QA10) as a measure of learners' perceptions during the action phase of simulation. BMC Med Educ 2023 May 1;23(1):290. [doi: 10.1186/s12909-023-04273-6] [Medline: 37127593]
29. Yamamoto A, Koda M, Ogawa H, et al. Enhancing medical interview skills through AI-simulated patient interactions: nonrandomized controlled trial. JMIR Med Educ 2024 Sep 23;10(1):e58753. [doi: 10.2196/58753] [Medline: 39312284]
30. Arpaci I, Kuşci I, Gibreel O. The role of personality traits in predicting educational use of generative AI in higher education. Sci Rep 2025 Aug 19;15(1):30440. [doi: 10.1038/s41598-025-16339-0] [Medline: 40830409]
31. Arpaci I, Gibreel O, Al-Sharafi MA. A hybrid SEM-ANN approach to predicting generative AI use in higher education through knowledge management. Comput Hum Behav Rep 2026 May;22:101044. [doi: 10.1016/j.chbr.2026.101044]
32. Sweller J, van Merrienboer JJG, Paas FGWC. Cognitive architecture and instructional design. Educ Psychol Rev 1998 Sep;10(3):251-296. [doi: 10.1023/A:1022193728205]
33. Paas F, Renkl A, Sweller J. Cognitive load theory and instructional design: recent developments. Educ Psychol 2003 Jan 1;38(1):1-4. [doi: 10.1207/S15326985EP3801_1]
34. de Jong T. Cognitive load theory, educational research, and instructional design: some food for thought. Instr Sci 2010 Mar;38(2):105-134. [doi: 10.1007/s11251-009-9110-0]
35. Vygotsky LS, Cole M, John-Steiner V, Scribner S, Souberman E. Mind in Society: The Development of Higher Psychological Processes: Harvard University Press; 1978. [doi: 10.2307/j.ctvjf9vz4]
36. Barry Issenberg S, Mcgaghie WC, Petrusa ER, Lee Gordon D, Scalese RJ. Features and uses of high-fidelity medical simulations that lead to effective learning: a BEME systematic review. Med Teach 2005 Jan;27(1):10-28. [doi: 10.1080/01421590500046924] [Medline: 16147767]
37. Motola I, Devine LA, Chung HS, Sullivan JE, Issenberg SB. Simulation in healthcare education: a best evidence practical guide. AMEE Guide No. 82. Med Teach 2013 Oct;35(10):e1511-e1530. [doi: 10.3109/0142159X.2013.818632] [Medline: 23941678]
38. Zamecnik A, Villa-Torrano C, Kovanović V, et al. The cohesion of small groups in technology-mediated learning environments: a systematic literature review. Educ Res Rev 2022 Feb;35:100427. [doi: 10.1016/j.edurev.2021.100427]
39. Elendu C, Amaechi DC, Okatta AU, et al. The impact of simulation-based training in medical education: a review. Medicine (Baltimore) 2024 Jul 5;103(27):e38813. [doi: 10.1097/MD.0000000000038813] [Medline: 38968472]
40. Quennelle S, Malekzadeh-Milani S, Garcelon N, et al. Active learning for extracting rare adverse events from electronic health records: a study in pediatric cardiology. Int J Med Inform 2025 Mar;195:105761. [doi: 10.1016/j.ijmedinf.2024.105761] [Medline: 39689449]
41. López-Úbeda P, Martín-Noguerol T, Luna A. Integrating semantic retrieval and chain-of-thought reasoning in small language models for SNOMED CT normalization. Int J Med Inform 2026 May;211:106340. [doi: 10.1016/j.ijmedinf.2026.106340] [Medline: 41678976]
42. Elhilali A, Ngo ASH, Reichenpfader D, Denecke K. Large language model-based patient simulation to foster communication skills in health care professionals: user-centered development and usability study. JMIR Med Educ 2025 Dec 12;11:e81271. [doi: 10.2196/81271] [Medline: 41385781]
43. Harden RM, Gleeson FA. Assessment of clinical competence using an objective structured clinical examination (OSCE). Med Educ 1979 Jan;13(1):41-54. [doi: 10.1111/j.1365-2923.1979.tb00918.x] [Medline: 763183]
44. MAPLE-CR. GitHub. 2026. URL: https://github.com/eMoLii/PBL [accessed 2026-03-09]

Abbreviations

AP: active participation
CC: communication and collaboration
CONSORT-eHEALTH: Consolidated Standards of Reporting Trials of Electronic and Mobile Health Applications and Online Telehealth

XSL•FO
RenderX

CR: clinical reasoning
ICC: intraclass correlation coefficient
KA: knowledge accuracy
LLM: large language model
MAE: mean absolute error
MAPLE-CR: Multiagent Problem-based Learning Environment for Clinical Reasoning
PBL: problem-based learning
SR: summarization and reflection

Edited by B Lesselroth; submitted 10.Mar.2026; peer-reviewed by BVK Golamari, B Rubinstein, I Arpaci; revised version received 03.Aug.2026; accepted 11.Aug.2026; published 03.Sep.2026.

Please cite as:
Zhao Z, Wu B, Tian S, Hu H, Hao P
Scaffolded AI-Supported Problem-Based Learning for Medical Interns: Exploratory Retrospectively Registered Randomized Controlled Evaluation With a Voluntary Feasibility Follow-Up
JMIR Med Educ 2026;12:e95039
URL: https://mededu.jmir.org/2026/1/e95039
doi:10.2196/95039

© Zhengqi Zhao, Baijing Wu, Shuyuan Tian, Haigen Hu, Pengyi Hao. Originally published in JMIR Medical Education (https://mededu.jmir.org), 3.Sep.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Education, is properly cited. The complete bibliographic information, a link to the original publication on https://mededu.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

AI-Generated Versus Human Supervisor Feedback on Medical Students' Clinical Clerkship Logs: Cross-Sectional Convergent Mixed Methods Study

Takeshi Kondo[1,2], MD, MHPE, PhD; Jeroen Donkers[2,3], PhD; Hiroshi Nishigori[1], MD, MMEd, PhD; Sanne Rovers[2,4], PhD; Sylvia Heeneman[2,4], PhD

[1]Center for Medical Education, Nagoya University Graduate School of Medicine, 65, Tsurumai-cho, Showa-ku, Nagoya, Japan

[2]School of Health Professions Education, Maastricht University, Maastricht, The Netherlands

[3]Department of Educational Research and Educational Design, Maastricht University, Maastricht, The Netherlands

[4]Department of Pathology, Maastricht University, Maastricht, The Netherlands

Corresponding Author:
Takeshi Kondo, MD, MHPE, PhD
Center for Medical Education, Nagoya University Graduate School of Medicine, 65, Tsurumai-cho, Showa-ku, Nagoya, Japan

Abstract

Background: Feedback is essential for medical students' learning during clinical clerkships; yet, supervising physicians often struggle to provide meaningful written feedback due to time constraints. Large language models offer a promising approach to supplement human feedback, but how artificial intelligence (AI)–generated and human feedback differ in authentic clinical settings remains unclear, as most comparisons have been conducted in classroom or simulation contexts.

Objective: The aim of the study is to examine how AI-generated feedback and supervisor-provided feedback differ when applied to medical students' clinical clerkship logs, by identifying the distinct characteristics and complementary strengths of each feedback type.

Methods: This cross-sectional convergent mixed methods study included 161 weekly clinical clerkship logs from 47 fifth- and sixth-year medical students across 12 clinical departments at Nagoya University, Japan (January-May 2024). Of 164 eligible logs, 3 were excluded because supervisors entered contact messages rather than substantive feedback. AI feedback was generated using GPT-4o. In total, 10 faculty physicians and 10 medical students evaluated both feedback types in blinded, randomized order using a validated 5-category rubric (criteria-based, clear direction, accuracy, prioritization, and supportive tone), followed by open-ended comments and source identification. Quantitative analyses (paired 2-tailed *t* tests, cumulative link mixed-effects models; α=.05 with Bonferroni correction) were complemented by qualitative thematic analysis and integrated using joint display analysis.

Results: AI feedback was significantly longer than supervisor feedback (mean 382.02, SD 81.82 vs mean 98.87, SD 73.66 characters; Cohen *d*=2.84, 95% CI 2.50 - 3.19; *P*<.001). Cumulative link mixed-effects models showed that AI scored higher on criteria-based (odds ratio [OR] 11.81, 95% CI 7.64 - 18.27; *P*<.001) and clear direction (OR 6.61, 95% CI 4.35 - 10.06; *P*<.001), with no significant differences on accuracy (OR 1.35, 95% CI 0.91 - 2.00; *P*>.99), prioritization (OR 1.70, 95% CI 1.16 - 2.50; *P*=.10), or supportive tone (OR 1.34, 95% CI 0.87 - 2.06; *P*>.99). AI feedback showed greater consistency (variance ratio 3.9:1; Levene $F_{1,320}$=73.20; *P*<.001). All 20 evaluators correctly identified feedback sources. Qualitative analysis revealed that AI provided structured, text-anchored feedback addressing rubric criteria, while supervisors offered experience-based feedback grounded in clinical context and professional expertise.

Conclusions: This study extends the comparison of AI-generated and supervisor feedback to an authentic clinical clerkship environment, moving beyond classroom and simulation settings examined in prior work. Through integrated mixed methods analysis, a key distinction emerged between text-anchored AI feedback, which systematically addresses written log content in alignment with rubric criteria, and experience-based supervisor feedback, which draws on clinical observation and professional judgment. AI consistently delivered structured feedback addressing gaps that arise when time-pressured supervisors provide brief comments, while supervisors contributed clinically grounded insights that AI cannot replicate. These complementary strengths suggest that AI feedback should supplement rather than replace supervisor feedback, and that hybrid models leveraging each type's advantages warrant investigation in clinical education.

(JMIR Med Educ 2026;12:e90064) doi:10.2196/90064

XSL•FO
RenderX

KEYWORDS

artificial intelligence; clinical clerkship; feedback; large language model; medical education; mixed methods research

Introduction

Background

Feedback serves as a critical educational instrument for medical students and residents to learn from their experiences in the clinical environment [1-4]. Such feedback can be provided orally or as written feedback [5]. In the clinical environment, where multiple supervising physicians often take turns in providing guidance, obtaining and documenting written feedback is important for ensuring the continuity of learning and assessment [6-8]. However, supervising physicians in the clinical setting often have multiple tasks, which puts a strain on the time they have to support learning through feedback [9]. Particularly regarding written feedback, supervising physicians are often unable to provide this in a sufficiently elaborated manner [9,10].

To supplement feedback from human supervisors, feedback generation using artificial intelligence (AI) has recently gained attention. In particular, since the advent of large language models (LLMs), the ease of using AI has led to an enthusiastic exploration of its application in feedback [11-13]. LLMs are AI models that can handle various tasks by pretraining on massive amounts of text data in advance [14]. Although previous AI technologies were also able to generate feedback, they required models trained on context-specific text data tailored to the feedback's domain [15]. In contrast, LLMs can adapt to diverse contexts with minimal modifications, making them promising tools for the generation of feedback [16,17]. When AI feedback and human feedback were compared for feedback generation for texts written by learners, it was shown that AI feedback can match that of well-trained educators in terms of appropriateness, usefulness of the content, and supportive tone [13,18,19]. However, even studies demonstrating that AI can generate high-quality feedback argue that AI feedback should complement rather than replace human feedback, citing student preferences and the responsibility that human instructors should maintain [20-22].

Rationale and Research Question

When leveraging AI feedback to complement human feedback, it is important to understand how the 2 differ. Prior research has reported inconsistent findings regarding these differences. Several studies report that AI feedback is longer and more detailed than human feedback [20,23,24], whereas another study found AI feedback to be shorter and less relevant [25]. Although longer feedback is often associated with higher perceived quality [26-28], the extent to which differences in length between AI and human feedback translate into quality advantages remains unclear [18,20]. For example, in dental histology assignments, AI feedback was longer and contained more improvement points than human feedback [20]; yet, an analytical, multicriterion comparison showed that human teachers produced feedback that was overall clearer, more accurate, and better prioritized, even when shorter [18]. Regarding the consistency of feedback quality, many studies report that the quality of human feedback varies widely depending on the provider and the situation [29-31], whereas some studies find that, compared with AI, human feedback is consistently higher in quality [25,32]. Findings are also inconsistent about whether recipients can tell if the feedback provider is AI or human, ranging from studies reporting that people could distinguish it reasonably well [33] to studies reporting that they could hardly distinguish it [34,35]. Thus, although many aspects of AI and human feedback have been examined, the results vary widely. These divergences likely relate to contextual variation such as differences in discipline, task authenticity, and assessment purpose.

When comparing AI and humans in other clinical environments including clinical clerkships, a key limitation is that most existing comparisons have examined classroom or simulation tasks rather than authentic clinical activities [18,20,23,24]. In clinical clerkships, supervising physicians work closely with students. Their feedback can therefore draw on direct observation of student behavior, tacit professional norms, and shared patient-care experiences. These dimensions are difficult for AI to approximate, because AI has access only to the written log. Furthermore, prior comparative studies often allowed humans ample preparation time for crafting feedback [13,18,19], whereas in real clinical settings, supervisors are time-pressured and may produce brief or nonspecific comments [9,10]. This specific context constraint could both expose areas where human feedback quality diminishes and highlight domains where AI systems might compensate, given their ability to generate structured output rapidly.

Accordingly, an analytical examination situated within the clinical environment is needed to clarify how AI and human feedback differ when applied to authentic clerkship learning logs under routine time constraints. By exploring these differences, we can identify ways to leverage AI-generated feedback to complement supervisors in the busy clinical environment. Accordingly, our research question is: In what ways and along which dimensions do feedback from supervising physicians and feedback generated by AI differ when applied to medical students' clinical clerkship logs?

Methods

Setting

The research was conducted within the clinical clerkship programs for fifth- and sixth-year medical students at Nagoya University (NU), Japan. In Japan, students enter medical school directly after high school (around age 18 years) and complete a 6-year program [36]. At NU, medical students finish classroom-based learning in clinical medicine and preparatory training for clinical clerkships (such as medical interviewing and physical examination) by the end of the fourth year. From late fourth to early fifth year, they undertake observation-centered clinical clerkships, rotating through all departments for short periods of 1 - 2 weeks. From the late fifth to the early sixth year, they participate in clinically active clerkships, rotating through each department for 4 weeks or longer. For this study, we focused on the clinical clerkship

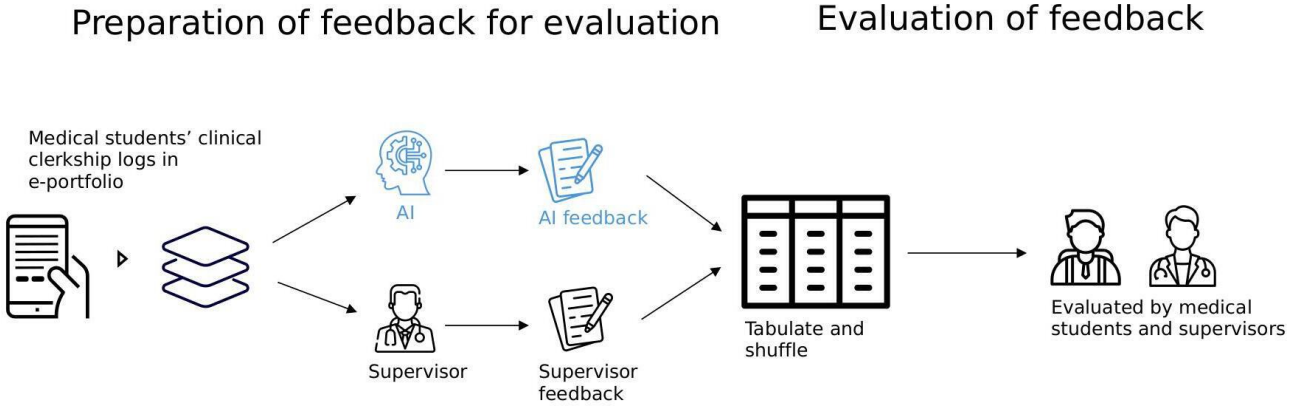
XSL•FO
RenderX

program at NU for students in the latter half of their fifth year and throughout their sixth year. The participants were a single cohort of students who began their clinically active clerkships during their fifth year and continued into their sixth year as part of their normal academic progression (the Japanese academic year begins in April). During the clerkship, students joined the medical team, examined patients, formulated assessments and plans, and documented these in the electronic medical record. Because access to electronic medical records was rigidly restricted, students also recorded their daily activities, learning points, and future tasks in an e-portfolio system separate from the medical record. These clinical clerkship logs in the e-portfolio were used solely for educational purposes, and students were instructed not to include any patient-identifying information. Each daily log entry consisted entirely of free-text fields in which students described what they did during the clerkship, the cases they experienced, what they learned, and what they planned to do next; no Likert scales or structured questions were included, and no upper or lower character limits were imposed. Students entered their logs via smartphone. The e-portfolio automatically compiled the daily entries in the clerkship log into weekly records and sent them to the supervising physicians, who then provided written feedback to the students via the e-portfolio. Supervising physicians received orientation on the e-portfolio system but no specific training or instructions regarding the content, length, structure, or format of their written feedback. Written feedback in the e-portfolio primarily served as a complement to verbal feedback provided during clinical supervision, although the relative emphasis varied across supervisors and departments.

Design

This study was conducted as a cross-sectional mixed methods study using a convergent design from a pragmatic perspective [37,38]. From the pragmatist standpoint of gaining insights that contribute to building a system where AI and supervisors complement each other to promote student learning, we examine the differences between AI-generated feedback and human-written feedback on clinical clerkship logs using both quantitative and qualitative approaches. First, we collected feedback from supervising physicians on clinical clerkship logs entered by medical students in the e-portfolio (supervisor feedback) and generated feedback using AI on the clinical clerkship logs (AI feedback), see Figure 1 for further details of this preparation process. Next, within each dataset, the presentation order of the 2 feedback types (AI vs supervisor) was randomly assigned so that approximately half of the datasets presented AI feedback first and the other half presented supervisor feedback first. This randomization was intended to prevent evaluators from deducing the source of each feedback from its position. The 161 datasets were then randomly distributed among the 10 faculty evaluators and 10 student evaluators. Because the assignment was random, some evaluators assessed multiple records originating from the same clinical department. Then, medical students and supervising physician evaluators were asked to evaluate both types of feedback. The evaluation used the rubric developed by Steiss et al [18] for quantitative analysis of scores and qualitative analysis of free-text comments. Finally, the results of the quantitative and qualitative analyses were integrated using joint display analysis (JDA) [39,40], in which each quantitative finding was placed alongside its corresponding qualitative theme in a structured display, and meta-inferences about how the 2 sources complement or diverge from each other were drawn. The detailed integration procedure is described later in the Integration of Results section. The process of preparing and evaluating AI feedback and supervisor feedback is shown in Figure 1.

Figure 1. Overview of the feedback preparation and evaluation process in a mixed methods study comparing AI-generated and supervisor feedback on clinical clerkship logs at Nagoya University, Japan. Student logs from the e-portfolio were used to generate AI feedback (GPT-4o) and collect supervisor feedback, which were then tabulated, shuffled, and evaluated by blinded medical students and faculty evaluators. AI: artificial intelligence.



Participants

Inclusion and Exclusion

During the data collection period (January 9 to May 5, 2024), 164 weekly clerkship logs from 47 fifth- and sixth-year medical students across 12 clinical departments (9 university departments and 3 community hospitals) were assessed for eligibility. In total, 3 logs were excluded because the supervising physician had entered contact messages to students rather than substantive feedback, yielding 161 logs for analysis.

For the evaluation phase, eligibility criteria for faculty evaluators included current involvement in teaching medical students; for student evaluators, eligibility required current participation in clinical clerkships or preparatory education for clinical clerkships. Individuals who authored the student logs used as

XSL•FO
RenderX

the basis for feedback, as well as those who authored the supervisor feedback being evaluated, were excluded.

Participant Characteristics

The study involved 2 participant groups: log authors (47 medical students from a single cohort transitioning from fifth to sixth year during the study period) and feedback evaluators (10 faculty physicians and 10 medical students). Detailed demographic information (age and sex) was not collected from either group; because evaluators participated as unpaid volunteers, the information requested was kept to a minimum to reduce burden and invasiveness.

Sampling Procedures

Student clerkship logs were collected as part of routine educational practice through the e-portfolio system; all logs from the study period were included unless they failed the eligibility criterion described earlier. Faculty evaluators were recruited via a mailing list for medical educators, and student evaluators were recruited through student representatives who disseminated information about the study via social media. All evaluators received an explanation about the significance and purpose of the study from the first author (TK) through documentation and an online meeting prior to their participation.

Materials

Generating AI Feedback

On the same set of weekly records of a medical student, feedback was also generated by GPT-4o (gpt-4o-2024-05-13) from OpenAI, which was the latest AI model available from OpenAI as of June 1, 2024, when this feedback was generated. This feedback is called “AI feedback” below.

Prompt Development

To generate high-quality feedback through LLMs, it is crucial to carefully craft the prompts provided to the model [41]. Since the students’ clinical clerkship logs are written in Japanese and the feedback is for Japanese students, we created the prompts in Japanese. Although Steiss et al [18] dealt with different types of feedback, their study also compared human-generated descriptions with feedback produced by ChatGPT in response to written tasks. Given the similarities, we decided to adapt the prompts from Steiss et al [18] as a foundation for our current research. TK, who is a Japanese speaker and familiar with the context where feedback is applied, translated them into Japanese. Additionally, TK referenced papers on prompt creation methods for high-quality feedback generation, adding context and mission elements, as well as criteria for high-quality feedback [41]. We then fed these refined prompts and actual student experiential logs into ChatGPT (GPT-4o) web interface to generate feedback. Next, we compared the generated feedback against previously reported quality criteria [18,42] and fine-tuned the prompts accordingly. Finally, to enhance reproducibility and subsequent integration into our e-portfolio system, we used the corresponding application programming interface (API) of the ChatGPT model (OpenAI Chat Completion API [gpt-4o-2024-05-13]) [43]. Each student log was processed as an independent API call with no shared context between calls, ensuring that feedback for one log could not be influenced by any other log. No maximum token limit (max_tokens parameter) was set at the API level.

The prompt was structured to include sections that established the premise of providing feedback on clinical clerkship logs, explained the structure of these logs, outlined the evaluation criteria for the feedback, and specified the expected content and format of the output. The prompt included an explicit instruction to write the feedback concisely within 500 characters. This character limit was determined through systematic pilot testing, in which feedback was generated under various length constraints (no limit, 200, 300, 400, and 600 characters); without any limit, the AI produced verbose output, in which key points were diluted, whereas overly restrictive limits resulted in superficial feedback. The 500-character limit was selected as the optimal balance between conciseness and substantive coverage. Supervising physicians received no instructions regarding length, structure, or format for their feedback. For an outline of the prompt construction and revision process, we refer to Multimedia Appendix 1. Since the logs were written in Japanese, the prompt was also developed in Japanese. A translation in English of the final prompt is included in the appendix as well as the original Japanese version.

An important methodological note regarding this design: the prompt explicitly embedded the following 5 criteria from the Steiss et al [18] rubric as the evaluation framework for generating feedback: criteria-based guidance, clear directions for improvement, accuracy, appropriate prioritization, and supportive tone. The AI was instructed to address each criterion in its output. Supervising physicians, by contrast, received no instructions to write in accordance with this rubric and provided feedback according to their usual clinical practice. This asymmetry means that the study compares a rubric-optimized AI system with naturalistic supervisor feedback rather than constituting a neutral comparison of inherent feedback quality.

Evaluation Instrument

The evaluation instrument consisted of closed-ended and open-ended questions. For the closed-ended questions, we used the scoring rubric developed by Steiss et al [18]. While there are other criteria used in studies comparing human and AI feedback, most of them used holistic scoring methods [13,19]. In contrast, Steiss et al [18] used analytical criteria that evaluate feedback by breaking it down into various elements. This approach aligns with our objective of analyzing how different aspects of feedback differ between humans and AI. This rubric evaluates 5 categories: criteria-based, clear directions for improvement, accuracy, prioritization of essential features, and supportive tone, each rated with a single score on a 5-point scale. A score of 5 represents the best evaluation, while 1 represents the worst. Each category has descriptive criteria attached. The rubric is presented in Multimedia Appendix 2. Because this rubric emphasizes structural dimensions of feedback quality rather than clinically specific qualities such as diagnostic reasoning and professional judgment, a complementary qualitative component was incorporated to capture dimensions of feedback value that quantitative rubric scoring alone cannot assess. The methodological implications of the alignment between the AI prompt and this rubric for

XSL•FO
RenderX

interpreting score comparisons are discussed in the Prompt Development section and the Limitations section.

The feedback in this study is in Japanese, and the evaluators are also Japanese, so it was necessary to translate the scoring rubric from English to Japanese. The translation process was conducted based on the translation, review, adjudication, pretesting, and documentation model [44]. This model is designed for the translation of evaluation forms, aiming to perform accurate translations while adapting to linguistic characteristics and cultural contexts that are difficult to address in traditional translation processes and to maintain the intent of the evaluation forms. This model follows five steps: (1) translation, (2) review, (3) adjudication, (4) pretesting, and (5) documentation. Following these steps, the translation was carried out as described below.

The original feedback quality evaluation rubric [18] was adapted for this study by modifying context-specific terms to fit the current research on student logs. The translation was performed by 2 bilingual individuals (TK and HN) who are fluent in both Japanese and English and experienced in medical education. The translations were reviewed by a medical education expert. Discrepancies were resolved by a professional translator. Pretesting involved a clinical educator and a medical student, leading to minor adjustments while preserving the original intent. These detailed processes were documented (Multimedia Appendix 3). This process ensured that the translated questionnaire was culturally and contextually appropriate for use in Japan.

In addition to the rubric, the instrument included a closed question asking evaluators whether the feedback had been written by a human or generated by AI, as well as an open-ended section in which evaluators described their impressions of the feedback's strengths and areas for improvement. The administration sequence of these components is described in the Data Collection section.

Sample Size, Power, and Precision

The required sample size for the 2-tailed *t* test was calculated using the values published in the paper by Steiss et al [18] that developed the rubric used for quantitative analysis. The paper reported effect sizes in terms of partial eta-square and Cohen *d*, which were used to calculate the sample size. In this study, considering the comparison between AI and supervisor feedback, the sample size was calculated based on Cohen *d* of 0.34 as reported by Steiss et al [18]. With power set at 0.8 and α at .05, a total of 155 data points per group would be needed.

Data Collection

The 161 datasets were evaluated between October 2024 and January 2025. For each feedback entry, evaluators first completed the 5-item rubric scoring and then provided free-text comments describing their impressions of the feedback's strengths and areas for improvement. Only after completing all rubric scoring and free-text comments was the identification question presented, asking evaluators to judge whether the feedback had been written by a human or generated by AI. This sequence was designed so that knowledge or suspicion of the source could not influence quality ratings, given that previous research has shown the feedback provider can influence how the feedback is perceived [45,46]. Whether the feedback had been generated by AI or written by a supervising physician was not disclosed to evaluators at the time of evaluation, ensuring they based their assessments solely on the feedback content.

Quantitative Analysis

All quantitative analyses were conducted in R (version 4.4.2; R Foundation for Statistical Computing). Text length was operationalized as the number of Unicode characters (Unicode Transformation Format-8) in each feedback entry. Since Japanese writing does not use spaces between words, and word counts can vary depending on morphological analysis methods, we used character counts as a more stable metric rather than word counts. Although it varies by domain, Japanese is reported to have approximately 1.7 characters per word [47]. Because each student record had both an AI and a supervisor feedback instance, analyses were paired at the record level, and all tests were 2-sided with α=.05; 95% CIs are reported. To examine whether AI and supervisor feedback differed in length, we compared character counts using a paired Student *t* test across records and visualized the distributions to contextualize any difference in means. To compare perceived quality, we analyzed rubric scores (5 ordinal categories: criteria-based, clear direction, accurate, prioritization, and supportive) in 2 complementary ways. First, for an easily interpretable summary at the record level, we averaged available assessor scores (student and faculty) within provider and item and then used paired 2-tailed *t* tests to compare AI and supervisor mean scores across records. Second, acknowledging the ordinal nature of the outcomes, we fitted cumulative link mixed-effects models. For each item, the fixed effects were feedback provider (AI vs supervisor) and assessor type (student vs faculty), with their interaction examined exploratorily, and a random intercept for assessor to account for clustering. Results are presented as odds ratios (ORs) with 95% CIs and *P* values under the proportional-odds assumption. To assess consistency in feedback quality, we computed the per-record mean score for each provider and compared dispersion between AI and supervisors. Variance differences were tested using the Levene test for equality of variances and corroborated with the classical *F* test, alongside distributional plots of the per-record means. To evaluate whether longer feedback tended to receive higher scores, we estimated Spearman rank correlations between character count and rubric scores separately by provider and item, reporting correlation coefficients and 2-sided *P* values. Finally, for the identification task, we summarized detection accuracy as the proportion of correctly identified sources (AI vs supervisor). Bonferroni correction was applied to correct for multiple testing.

Qualitative Analysis

Overview

For the qualitative analysis, narrative comments from the open questions were organized in a table, followed by a thematic analysis [48]. Thematic analysis consists of the following steps: (1) familiarize with the data, (2) generate initial codes, (3) search for themes, (4) review themes, (5) define and name themes, and (6) produce the report [48]. Codes were generated inductively from the evaluators' free-text comments rather than being

XSL•FO
RenderX

mapped onto a predefined coding framework. Steps 1 and 2 were conducted by 2 individuals (TK and HN). TK created the initial codes, while HN reviewed and refined them. Coding meetings between TK and HN were held in person, and theme-development meetings with SH, SR, and JD (step 3 onward) were conducted via videoconference. When coding or theme-level disagreements arose, the analysts returned to the original data (Japanese excerpts at the initial coding stage and English-translated codes during theme development), examined alternative interpretations, and continued discussion until consensus was reached before advancing to the next step. Steps 1 and 2 were conducted in Japanese, while steps 3 to 6 were conducted in English. From step 3 onward, the codes were translated into English, and SH, SR, and JD also participated. This allowed for a more multifaceted analysis. Codes and themes were organized in Google Sheets and Microsoft Excel; R (version 4.4.2; the same version used for the quantitative analyses) was used for tabulating the code list and computing simple frequency summaries. No dedicated computer-assisted qualitative data analysis software (CAQDAS) was used. The backgrounds of TK, HN, SH, SR, and JD are described below.

Reflexivity

TK is a general practitioner directly involved in clinical clerkships and also engaged in the development of e-portfolios and feedback. HN is an expert in medical education, researching professionalism in medical education. SH researches e-portfolios and programmatic assessment in health professions education, SR is an expert in self-regulated learning and assessment, and JD is a specialist in statistics and AI.

TK's knowledge about the context of clinical clerkship feedback provision helps in making contextually informed deeper analysis when interpreting descriptions, while TK's involvement in the development of e-portfolios might lead to potential bias where results could be interpreted more favorably toward e-portfolios. The specialized fields of HN, SH, SR, and JD each provide expert perspectives when analyzing descriptions. On the other hand, there is a possibility of biased interpretations based on their respective specialties. To mitigate such biases, discussions were recorded, and transparency was maintained during the analysis process, with regular meetings to share and discuss their respective interpretations. The research team had no direct supervisory or assessment-related authority over the evaluators who provided the qualitative data; evaluators were recruited independently (see Sampling Procedures section) and participated on an opt-in voluntary basis.

Methodological Integrity

Methodological integrity was pursued along the 2 dimensions articulated in the APA Journal Article Reporting Standards for Qualitative Research [49]: fidelity to the subject matter and utility in achieving the research goals. With respect to data adequacy, the analytic corpus comprised the complete set of open-ended comments from all 20 evaluators (10 faculty and 10 students) across 161 datasets spanning 12 clinical departments; because this full-corpus approach precluded iterative sampling to saturation in the conventional sense, we treated the breadth of the evaluator pool (2 assessor types and multiple specialties) as the operational criterion for sufficiency. Themes are illustrated in the Results section with verbatim excerpts to maintain groundedness.

Credibility was supported by several analytic checks. Initial coding (steps 1 and 2) was double-coded by TK and HN, with discussion to agreement. Theme development (steps 3 to 6) was opened to 3 additional analysts (SH, SR, and JD) with distinct disciplinary perspectives, functioning as analyst triangulation across clinical practice, medical education, programmatic assessment, self-regulated learning, and statistics and AI. The bilingual workflow (Japanese for initial coding and English for theme development) further served as a translational check on interpretive fidelity. All analytic discussions were recorded, providing an audit trail. Themes were iteratively examined for within-theme disconfirming instances, and cases contradicting the dominant pattern (eg, brief supervisor feedback that was nonetheless judged effective or AI feedback judged overly prescriptive) were retained in the theme descriptions rather than treated as outliers.

Member checking was not conducted because evaluators participated as unpaid volunteers, including sixth-year students approaching the national licensing examination and clinically active faculty, for whom additional engagement would have been incompatible with our commitment to minimize participant burden (see Ethical Considerations section). Credibility was instead supported by the multianalyst procedures described earlier. Meaningful coherence between findings and interpretation was pursued through JDA, in which each quantitative result was aligned with its corresponding qualitative theme (see Integration of Results section). To support transferability, the Setting section provides detailed contextual information about the Japanese undergraduate clinical clerkship environment so readers can judge applicability to other educational contexts.

Integration of Results

The results of the quantitative and qualitative analyses were documented separately and then integrated using JDA. JDA is an analytic strategy in mixed methods research that brings qualitative and quantitative evidence together in a single, structured visual (a "joint display") to enable explicit integration and the drawing of meta-inferences that would be difficult to see from either strand alone [39,40]. In this study, the results of the quantitative and qualitative analyses were compared and integrated using JDA. Specifically, the quantitative results were tabulated, and the themes identified in the qualitative analysis were listed alongside them. The relationships between the quantitative results and qualitative themes were then examined, and meta-inferences were drawn.

Ethical Considerations

This study was approved by the ethics committee of Nagoya University Hospital (approval: 2023‐0451). An opt-out approach was taken for the research use of student clinical clerkship logs and supervisor feedback accumulated in the e-portfolio. Before the clerkship, all medical students received an explanation from TK via online meetings regarding the use of clinical clerkship logs in e-portfolio for research purposes and were informed that they could opt out by sending an email

XSL•FO
RenderX

to a designated address. Supervising participants received explanations from TK via email and, upon request, online meetings regarding the use of feedback recorded in the e-portfolio and were likewise informed that they could opt out by sending an email to a designated address. Research participants who evaluated AI feedback and supervisor feedback were recruited using an opt-in approach. All evaluators provided informed consent before participating. Student clinical clerkship logs, supervisor feedback, AI feedback, and evaluator assessment data used in the research were anonymized for analysis. All proper nouns were replaced with alphabetic characters, and linking tables were managed in an encrypted state. Evaluators received no financial compensation for their participation; as acknowledgment of their contribution, they were provided with information about the research and early access to the AI prompt used for feedback generation. No individually identifiable information of any participant is presented in this paper, tables, figures, or supplementary materials.

Results

Overview

During the clinical clerkship conducted at NU from January 9, 2024, to May 5, 2024, combinations of student clerkship logs and corresponding supervisor feedback were collected. The data included 9 clinical departments and 3 community hospitals, involving a total of 47 medical students and 42 supervisors who contributed to the clerkship logs and feedback. On June 1, 2024, AI feedback was generated for the student clerkship logs. As a result, a total of 161 datasets were created, each consisting of a student's clerkship logs for 1 week, supervisor feedback, and AI feedback. The weekly clerkship logs varied considerably in length (mean 675.2, SD 459.8; range 51 - 2342; median 551, IQR 313-1021 characters). The number of datasets, students, and supervisors by department is shown in Table 1.

Table . Dataset summary[a].

Department[b]	Datasets, n	Students, n[c]	Supervisors, n[d]
A Community Hospital	13	11	10
B Community Hospital	7	6	7
C Community Hospital	4	4	3
Diabetes and Endocrinology	9	8	9
General Medicine	5	5	5
Geriatric Medicine	16	12	10
Hematology	11	10	10
Neurology	16	12	15
Pediatrics	32	22	19
Psychiatry	9	7	8
Surgical Unit	23	18	17
Thoracic Surgery Unit	16	12	14

[a]Convergent mixed methods study comparing artificial intelligence–generated and supervisor feedback on clinical clerkship logs, Nagoya University, Japan, 2024.

[b]Except for A, B, and C Community Hospital, all departments are affiliated with Nagoya University.

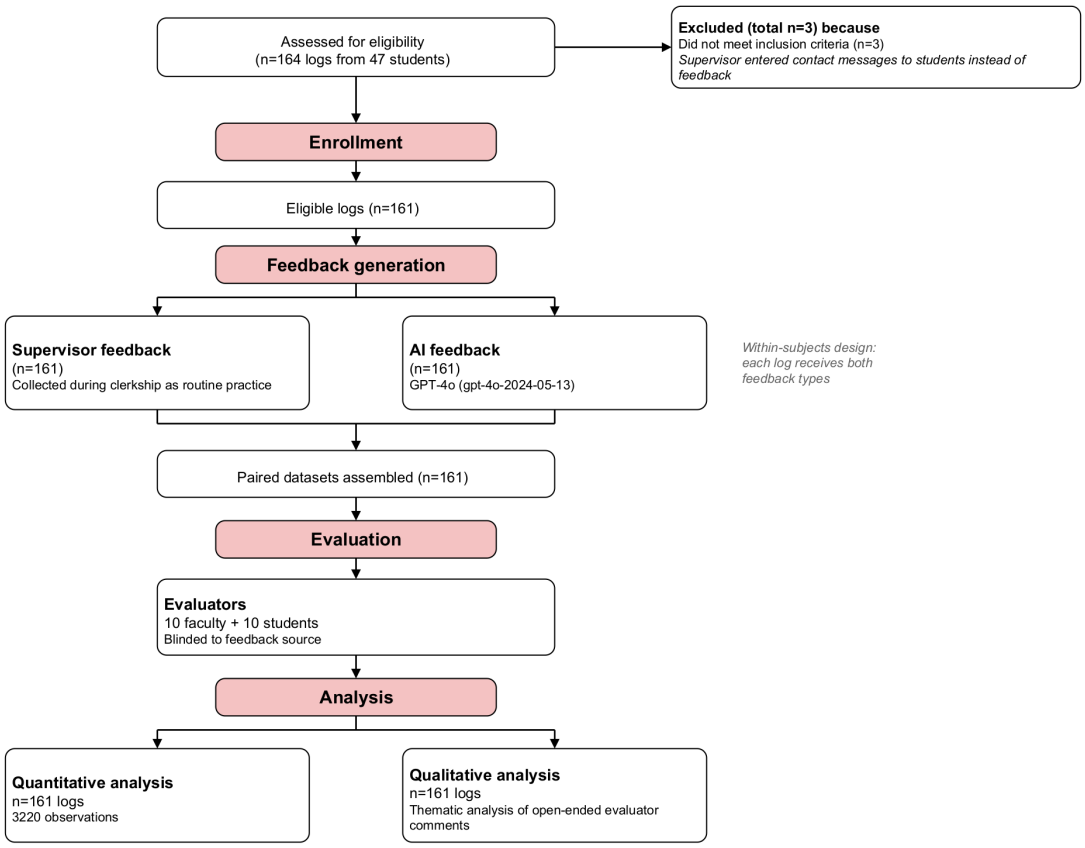
[c]The total number of students exceeds 47 because some students rotated through multiple departments.

[d]The total number of supervisors exceeds 42 because some supervisors provided feedback in multiple departments.

From October 2024 to January 2025, 10 medical students and 10 faculty who were not involved in the creation of the above student clerkship logs or feedback evaluated the AI and supervisor feedback. No missing data were identified; all 161 records contained complete rubric scores from both assessors across all 5 items, yielding 3220 observations with 100% data completeness. Because the dataset was complete, we did not perform a missingness test (such as Little missing completely at random test) or apply multiple imputation. The flow of participants through each stage of the study is shown in Figure 2.

XSL•FO
RenderX

Figure 2. Participant flowchart for a mixed methods study comparing AI-generated and supervisor feedback on clinical clerkship logs at Nagoya University, Japan (January 2024-January 2025). The diagram shows participant flow from enrollment (164 logs from 47 medical students across 12 departments) through exclusion (n=3), feedback generation (within-subjects design: each log received both AI and supervisor feedback), blinded evaluation by 10 faculty and 10 student evaluators, to quantitative (3220 observations) and qualitative analyses. AI: artificial intelligence.



Quantitative Analysis

Distribution Assessment and Sensitivity Analysis

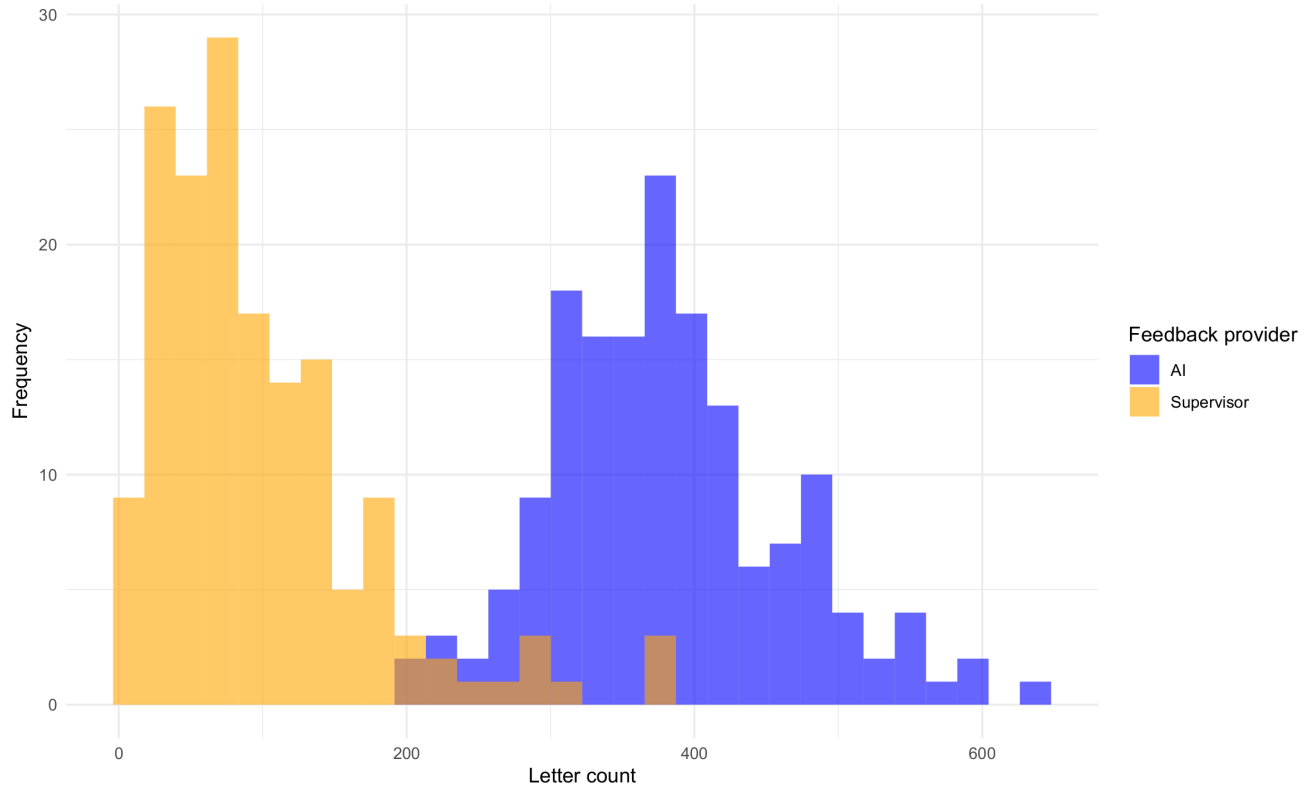
Shapiro-Wilk tests assessed the normality of paired differences for all comparisons. The feedback length difference was normally distributed (W=0.992; P=.55), supporting the use of the paired 2-tailed t test for that variable. However, all 5 rubric score differences departed significantly from normality (all W≤0.939; all P<.001 after Bonferroni correction). As a sensitivity analysis, Wilcoxon signed rank tests were conducted alongside the paired 2-tailed t tests; the 2 methods yielded identical conclusions regarding statistical significance for every variable. Moreover, the primary analysis of rubric scores used cumulative link mixed-effects models (see the Comparison of Scores section), which do not assume normality of the outcome distribution.

Difference Between Feedback Length

The length of the feedback was compared between AI and supervisor feedback. Mean lengths were 382.02 (95% CI 369.29 - 394.76) characters for AI and 97.87 (95% CI 86.4 - 109.33) characters for supervisors. The distribution of the lengths of supervisor feedback and AI feedback is shown in Figure 3. Paired 2-tailed t test results indicated that the mean character count of AI feedback was significantly higher than that of supervisor feedback (mean difference 284.16, 95% CI 268.6 - 299.71 characters; P<.001; Cohen d=2.84, 95% CI 2.50 - 3.19).

XSL•FO
RenderX

Figure 3. Distribution of AI-generated and supervisor feedback lengths by character count for 161 clinical clerkship logs from 47 medical students, Nagoya University, Japan, 2024. AI: artificial intelligence.



Comparison of Scores

The scores assigned by the evaluators to the AI and supervisor feedback were compared. Multimedia Appendix 4 shows the differences in feedback scores by AI and supervisors for the same log. Values greater than 0 indicate that the AI score is higher, and values less than 0 indicate that the supervisor score is higher. Although there are slight differences among evaluators, AI scores tend to be higher except for the "supportive" item.

Next, we compared the mean scores of AI and supervisor feedback for each category. In all categories, AI feedback had higher mean scores than supervisor feedback. A 2-tailed *t* test comparing the mean scores of supervisor feedback and AI feedback showed that AI feedback was rated significantly higher in all categories after Bonferroni correction: criteria-based (corrected $P<.001$), clear direction (corrected $P<.001$), prioritization (corrected $P<.001$), accurate (corrected $P<.001$), and supportive (corrected $P=.01$). Details are shown in Table 2; mean differences and uncorrected *P* values are provided in Multimedia Appendix 5.

Table . Paired *t* test results comparing artificial intelligence (AI) and supervisor feedback scores on the Steiss et al 5-item analytic rubric[a].

Item	AI mean score (95% CI)	Supervisor mean score (95% CI)	Cohen *d* (95% CI)	Corrected *P* value
Criteria-based	4.43 (4.34 - 4.52)	2.73 (2.56 - 2.9)	0.93 (0.79-1.06)	<.001
Clear direction	4.36 (4.26 - 4.46)	2.98 (2.81 - 3.14)	0.73 (0.61-0.86)	<.001
Accurate	4.18 (4.07 - 4.28)	3.8 (3.64 - 3.96)	0.22 (0.11-0.33)	<.001
Prioritization	4.05 (3.94 - 4.16)	3.37 (3.2 - 3.54)	0.36 (0.24-0.47)	<.001
Supportive	4.56 (4.47 - 4.65)	4.33 (4.21 - 4.45)	0.17 (0.06-0.28)	.01

[a]Convergent mixed methods study, Nagoya University medical students, Japan, 2024. Rubric items: criteria-based, clear direction, accurate, prioritization, and supportive. This rubric assesses structural and content dimensions of written feedback; clinically specific dimensions such as diagnostic reasoning and professional judgment were not directly measured. Mean differences and uncorrected *P* values are reported in Multimedia Appendix 5.

Because the rubric used for scoring is a 5-point scale and the intervals from 1 to 5 may not be equal, using average scores may not be appropriate. Therefore, we used ordinal logistic regression to examine whether there are differences between AI and supervisor scores. We also used a mixed-effects ordinal model to account for assessor type (student or faculty) and random effects by assessor. The results are shown in Table 3.

In the ordinal logistic regression, 3 effects are expressed as OR for each category. "Provider: AI" indicates that the larger the OR above 1, the more likely AI is to receive higher scores compared to supervisors. "Assessor type: student" indicates that the larger the OR above 1, the more likely students are to give higher scores compared to faculty. "Interaction: supervisor×student" indicates that the larger the OR above 1,

XSL•FO
RenderX

the greater the difference in scores between AI and supervisors among students compared to faculty. *P* values are Bonferroni-corrected for comparisons across 15 items. The results indicated that, for the items criteria-based and clear direction, AI feedback received significantly higher scores than supervisor feedback (criteria-based: OR 11.81, 95% CI 7.64 - 18.27; *P*<.001 and clear direction: OR 6.61, 95% CI 4.35 - 10.06; *P*<.001; OR values larger than 1 indicate higher scores for AI than for supervisors). For accurate, prioritization, and supportive, AI also tended to receive higher scores, but the differences between AI and supervisors were not significant (accurate: OR 1.35, 95% CI 0.91 - 2.00; *P*>.99; prioritization: OR 1.70, 95% CI 1.16 - 2.50; *P*=.10; and supportive: OR 1.34, 95% CI 0.87 - 2.06; *P*>.99). Regarding assessor type (student vs faculty), no significant differences were observed in any of the items. We also included interaction terms to test whether the effects of AI versus supervisor depended on assessor type, but no significant interaction effects were observed for any item.

Table . Cumulative link mixed-effects models results comparing artificial intelligence (AI) and supervisor feedback scores on the Steiss et al 5-item analytic rubric[a].

Item and effect	OR[b] (95% CI)	*P* value	Corrected *P* value
Criteria-based			
	11.81 (7.64 - 18.27)	<.001	<.001
Provider: AI (vs supervisor)			
Assessor type: student (vs faculty)	1.60 (1.07 - 2.39)	.02	.33
Interaction: supervisor×student	0.88 (0.49 - 1.59)	.68	.99
Clear direction			
	6.61 (4.35 - 10.06)	<.001	<.001
Provider: AI (vs supervisor)			
Assessor type: student (vs faculty)	1.37 (0.92 - 2.04)	.12	.99
Interaction: supervisor×student	1.09 (0.61 - 1.95)	.78	.99
Accurate			
	1.35 (0.91 - 2.00)	.14	.99
Provider: AI (vs supervisor)			
Assessor type: student (vs faculty)	1.77 (1.16 - 2.72)	.008	.13
Interaction: supervisor×student	0.98 (0.55 - 1.76)	.96	.99
Prioritization			
	1.70 (1.16 - 2.50)	.007	.10
Provider: AI (vs supervisor)			
Assessor type: student (vs faculty)	1.25 (0.83 - 1.89)	.28	.99
Interaction: supervisor×student	1.62 (0.92 - 2.86)	.10	.99
Supportive			
Provider: AI (vs supervisor)	1.34 (0.87 - 2.06)	.19	.99
Assessor type: student (vs faculty)	1.57 (0.99 - 2.49)	.06	.83
Interaction: supervisor×student	1.19 (0.61 - 2.34)	.60	.99

[a]Convergent mixed methods study, Nagoya University medical students, Japan, 2024. See Table 2 for rubric item descriptions and measurement scope.
[b]OR: odds ratio.

These rubric-based findings should be interpreted in light of the asymmetric design described in the Methods section: the rubric criteria were explicitly embedded in the AI prompt but not given to supervisors. The observed score differences therefore reflect adherence to these specific structural dimensions rather than overall educational quality. The qualitative findings, particularly the themes of perspective as a clinician and continuity with practice, provide complementary evidence of the clinical value that supervisors contribute and that this rubric does not capture.

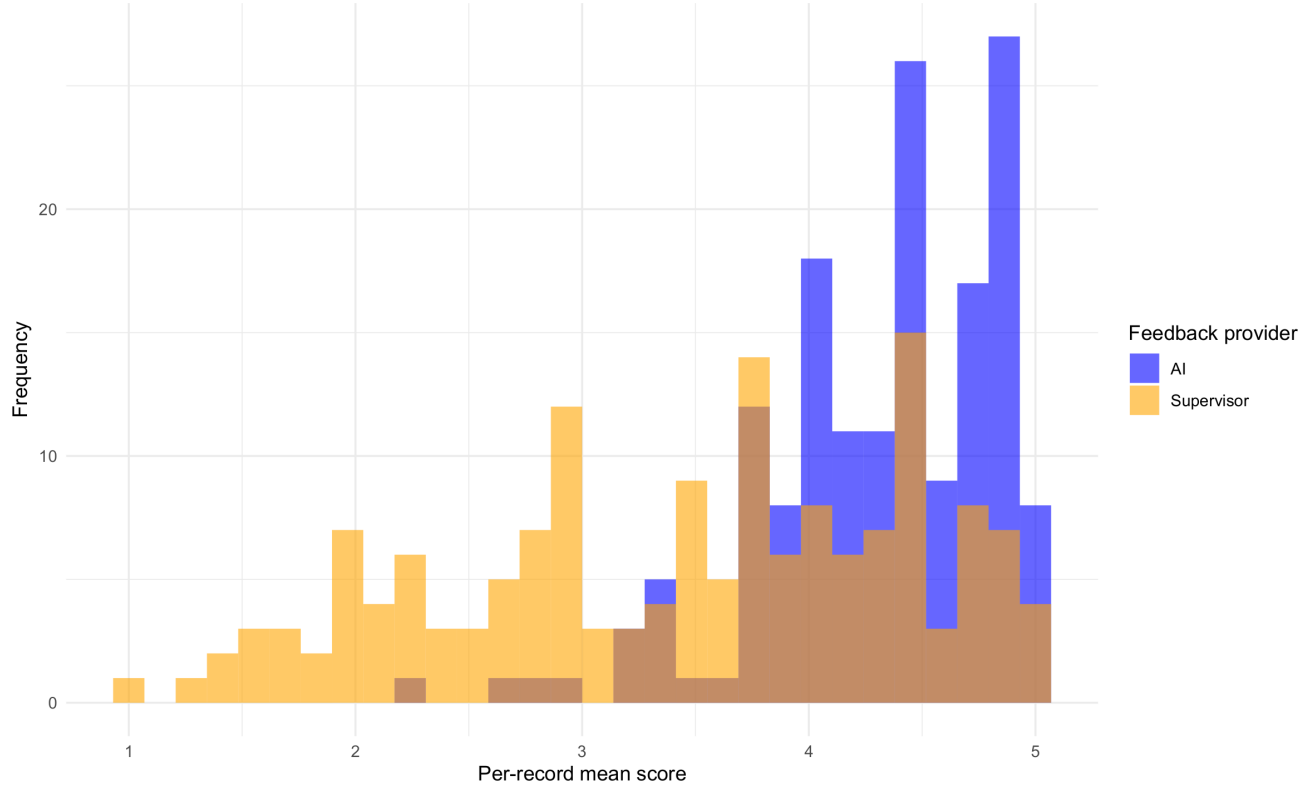
Consistency of Rubric Scores

To compare the consistency of rubric scores between AI and supervisors, we plotted the distribution of per-clinical clerkship log mean scores (Multimedia Appendix 4Figure 4). AI exhibited smaller score dispersion, suggesting higher consistency in rubric scores. We further tested whether the variance of per-clinical clerkship log mean scores differs between AI and supervisors. The per-clinical clerkship log mean score was significantly more dispersed for supervisors (mean 3.44, SD 1.01; variance 1.011) than for AI (mean 4.32, SD 0.51; variance 0.262; Levene:

XSL•FO
RenderX

$F_{1,320}$=73.203; P<.001; F test: $F_{160,160}$=0.259; P<.001). The supervisor variance was about 3.9 times that of AI, suggesting that AI feedback received more consistent rubric scores, whereas supervisor feedback received more variable ones.

Figure 4. Distribution of per-record mean rubric scores for AI-generated and supervisor feedback on 161 clinical clerkship logs, Nagoya University, Japan, 2024. AI feedback shows smaller dispersion, indicating higher consistency. AI: artificial intelligence.



Correlation Between Feedback Length and Score

To investigate whether the character count of feedback is correlated with the scores given by evaluators, we calculated the Spearman rank correlation coefficient for each feedback provider. The results are summarized in Table 4. *P* values are Bonferroni-corrected for comparisons across 10 items. A significant positive correlation would suggest that longer feedback tends to receive higher scores. No significant correlation was observed between text length and score for AI, but a significant correlation was found for all item categories for supervisors. This suggests that longer feedback from supervisors tends to receive higher scores.

Table . Spearman rank correlation between feedback character count and rubric score by feedback provider[a].

Item	Spearman ρ		Corrected *P* value	
	AI[b]	Supervisor	AI	Supervisor
Accurate	0.04	0.50	>.99	<.001
Clear direction	0.00	0.63	>.99	<.001
Criteria-based	0.08	0.60	>.99	<.001
Prioritization	−0.02	0.58	>.99	<.001
Supportive	0.00	0.30	>.99	<.001

[a]Convergent mixed methods study, Nagoya University medical students, Japan, 2024. *P* values are Bonferroni-corrected for 10 comparisons.

[b]AI: artificial intelligence.

Relationship Between Student Log Length and Feedback Scores

Because AI feedback is tightly text-anchored, variation in student log quality could potentially influence feedback scores. To examine whether log length influenced rubric scores, Spearman correlations were calculated between weekly log character count and each rubric score, separately by feedback provider. After Bonferroni correction for 10 comparisons, no significant correlations were observed for either AI or supervisor feedback (all corrected *P*>.05), indicating that variation in student log length did not systematically influence rubric scores for either feedback type.

XSL•FO
RenderX

Interrater Reliability

Because each dataset was independently evaluated by 1 faculty member and 1 student, we examined agreement between these 2 assessor types using weighted Cohen κ (quadratic weights) and intraclass correlation coefficients, computed separately for AI and supervisor feedback across all 5 rubric items (Table 5).

For supervisor feedback, interrater agreement was fair to moderate (weighted κ=0.37 - 0.54; all *P*<.001). For AI feedback, agreement was poor (weighted κ=0.04 - 0.15, mostly nonsignificant). This pattern suggests that faculty and student evaluators reached more consistent judgments when rating supervisor feedback than when rating AI feedback.

Table . Interrater reliability between faculty and student evaluators for artificial intelligence (AI)–generated and supervisor feedback[a].

Item and feedback type	Values, n	% Agreement	Weighted κ	κ, *P* value	ICC[b] (95% CI)	ICC, *P* value
Accurate						
AI	161	36	0.062	.41	0.062 (−0.087 to 0.210)	.21
Supervisor	161	41	0.383	<.001	0.384 (0.246 to 0.508)	<.001
Clear direction						
AI	161	44.7	0.150	.047	0.151 (−0.004 to 0.298)	.03
Supervisor	161	37.9	0.535	<.001	0.536 (0.417 to 0.638)	<.001
Criteria-based						
AI	161	44.1	0.076	.31	0.076 (−0.080 to 0.229)	.17
Supervisor	161	31.1	0.427	<.001	0.428 (0.293 to 0.547)	<.001
Prioritization						
AI	161	24.2	0.042	.57	0.042 (−0.107 to 0.191)	.29
Supervisor	161	32.9	0.379	<.001	0.381 (0.241 to 0.505)	<.001
Supportive						
AI	161	49.7	0.073	.30	0.073 (−0.083 to 0.226)	.18
Supervisor	161	55.9	0.365	<.001	0.367 (0.226 to 0.492)	<.001

[a]Convergent mixed methods study, Nagoya University, Japan, 2024. Weighted Cohen κ (quadratic weights) and ICCs were computed for each rubric item.

[b]ICC: intraclass correlation coefficient.

Identification of Feedback Source

After completing all rubric scoring and free-text comments, evaluators were asked whether they could identify the source of the feedback, specifically whether it was generated by a human or an AI. Despite the blinding procedures (randomized presentation order and removal of source labels), all students and supervisors correctly identified the source for every item (100% detection accuracy).

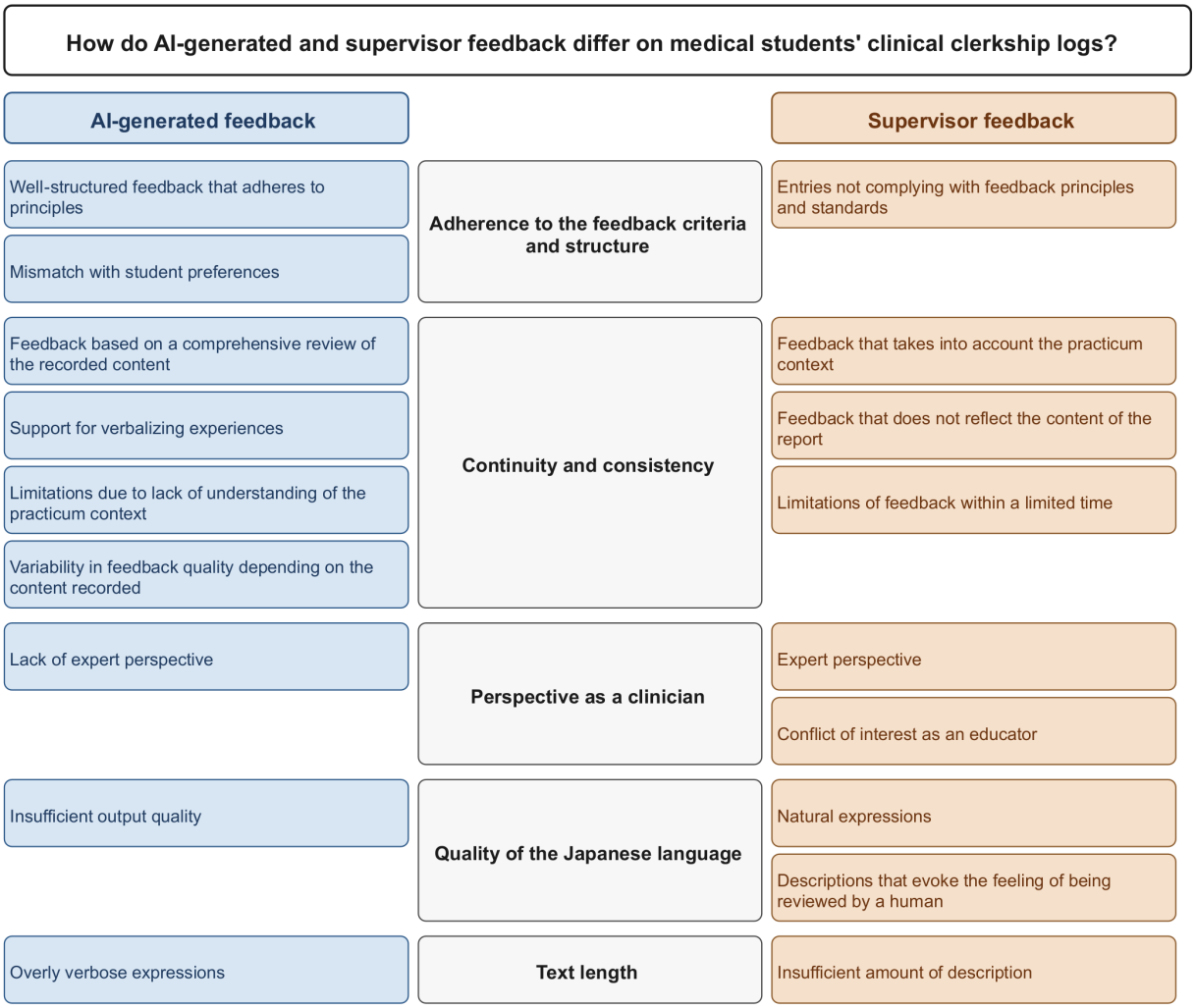
Qualitative Analysis

Overview

The qualitative analysis of the free-text comments provided by the evaluators revealed several themes regarding the differences between AI and supervisor feedback. Following the thematic-map guidance of Ahmed et al [50], the thematic map in Figure 5 visualizes the research question, the 5 themes, and their subthemes arranged by feedback provider.

XSL•FO
RenderX

Figure 5. Thematic map of differences between AI-generated and supervisor feedback on medical students' clinical clerkship logs. The research question is presented at the top, the 2 feedback modes are shown as contrasting poles, and each of the 5 themes is displayed with its AI and supervisor subthemes arranged on the left and right, respectively. Convergent mixed methods study, 161 clinical clerkship logs, Nagoya University, Japan, 2024. AI: artificial intelligence.



Five themes were identified as differences between AI and supervisor feedback: "adherence to the feedback criteria and structure," "continuity and consistency," "perspective as a clinician," "quality of the Japanese language," and "text length." Below, we describe the details of these themes. Citations consist of IDs with 3 parts. When the first part is "AI_," it indicates comments about AI feedback, and when it is "SU_," it indicates comments about supervisor feedback. The second part shows "_FA" for comments by faculty evaluators and "_ST" for comments by student evaluators. The number in the latter half represents the evaluator ID. The final part indicates the dataset ID.

Adherence to the Feedback Criteria and Structure

Across comments, AI-generated feedback was repeatedly described as tightly aligned with the rubric and highly structured. Evaluators highlighted concrete, example-driven suggestions, clear prioritization, and a predictable structure. This structure made the next steps "easy to put into practice," but some judged it overly prescriptive or formulaic. By contrast, supervisor feedback often departed from the rubric's structure: at times it omitted actionable "points for improvement," offered too few details, or listed many items without prioritization. Several comments, however, framed these departures as deliberate, "living" feedback rooted in real clinical work rather than in formal criteria.

Illustrative comments included: "This is a well-written report that starts with positive feedback, is specific and constructive, and doesn't have too many points" (AI_FA07_060), "It goes from positive to constructive ... priorities are also pointed out ... good feedback" (AI_FA07_063), and "It is very specific and detailed. It is written in the classic P-N-P format ..." (AI_FA09_086). On limits of the AI style, evaluators noted it could feel templated or prescriptive, eg, "sentences were ... not natural" (AI_ST18_078) and "it would be better if the suggestions for revision were not given in specific sentences" (AI_ST20_115). For supervisors, one comment praised specificity but worried that "there are many items, there is no priority, and it seems a little harsh" (SU_FA07_058), while many others pointed out missing action points: "There is no specific feedback ..." (SU_FA13_163) and "No evaluation was

XSL•FO
RenderX

given at all. No feedback ... on areas for improvement" (SU_ST21_131). Several remarks captured the trade-off: "In terms of referring to the evaluation criteria, it is inferior ... but it is more in line with the content of the practical training and presents the next task" (SU_FA11_127) and "It's not feedback that is tailored to a standard, but it feels like 'living' feedback that is likely to bring about behavioral change" (SU_FA09_088).

Continuity and Consistency

Across evaluators, continuity and consistency emerged as a defining difference in how AI and supervising physicians oriented their feedback. AI feedback typically exhibited strong internal coherence with the student's clerkship log. It read closely, mirrored the structure of entries, and translated what was on the page into concrete, criterion-referenced suggestions. This text-anchored continuity helped AI surface specific, actionable edits to the log itself. At the same time, because AI focused only on the student clerkship log, it could miss or misinterpret the broader practicum context, weekend schedules, orientation days, unit norms, or feasibility in busy wards, and offered fewer comments about attitudes, behaviors, or clinical reasoning unfolding across the rotation. In contrast, supervising physicians showed continuity with the clinical experience. Their feedback often drew on shared encounters, observed behaviors, and ward-based realities, which gave comments a sense of lived coherence over time. However, that experiential continuity did not reliably connect back to the student's written descriptions. Many supervisor comments were brief, global, or focused on the clinical day rather than the clerkship log, creating a misalignment.

This pattern was visible in how AI consistently tied remarks to what students had written: "the feedback is tailored to the clinical clerkship logs ... points that need to be improved ... are mentioned in detail, with specific examples" (AI_ST15_031) and "careful guidance in response to the lack of clinical clerkship logs ... specific suggestions for improvement" (AI_ST15_030). Evaluators repeatedly noted that AI "gave feedback after the log had been carefully examined ... [with] very specific" next steps (AI_ST21_117) and that its "comments are based on the content" (AI_FA05_034). Yet, this text-level consistency sometimes slipped when context mattered: AI "gives the impression that days without logs=days without practice" (AI_FA10_104) and tended to "encourage learning by improving how they write ... [with] a lack of clinical knowledge or perspective" (AI_ST21_121). Several evaluators wished for an overarching synthesis across the week or rotation: "no overall evaluation ... perhaps (AI was) a good micro-manager?" (AI_FA11_117) and "it would be even better if the feedback also took into account the overall outcome" (AI_FA10_109). By contrast, supervisors frequently referenced shared clinical situations and behaviors "useful because it gave supplementary information that wasn't included in the medical student's description" (SU_ST17_062); advice that was "practical and in line with real clinical practice" (SU_ST18_080); and impressions that only a clinician observing the student could offer, such as "he was going about things in a calm and collected way" (SU_ST23_157) and the feeling of "working together with medical students" (SU_FA04_008). However, this experiential continuity often bypassed the portfolio itself: "no mention of the content of the practical training or the learner's own description" (SU_FA10_109), "they did not evaluate the logs that were written ... just a summary evaluation conclusion" (SU_FA11_119), and "feedback was based on the actual content of the practical training ... it would be better if there was more feedback on the learning logs" (SU_ST21_124). In some instances, identical or near-identical messages suggested time-pressured reuse rather than log-specific engagement (eg, "copied and pasted feedback" (SU_ST23_164)). Taken together, the data indicate that AI maintained continuity with what students wrote, whereas supervisors maintained continuity with what students did, each affording a different, and at times incomplete, alignment with the students' log.

Perspective as a Clinician

Evaluators frequently contrasted the clinical lens. AI feedback was perceived as strong on structure, yet often thin on clinical reasoning, occasionally overlooking medical inaccuracies or proposing improvements not grounded in specialty-specific practice. Supervisor feedback, in turn, more often drew on clinical judgment, flagging errors, supplying pathophysiologic rationale, and pointing learners toward consequential problems, though some comments cautioned that supervisors can import their own priorities (eg, recruitment messages and niche emphases) or inadvertently signal an authority gradient.

Comments underscoring AI's limits included: "did not point out the errors in the words 'pulse irregularity' and 'electrocardiogram holder'" (AI_FA11_120), "specific examples ... are ideas that are not based on much medical expertise" (AI_ST22_143), "suggestions for improvement ... are not from the surgeon's perspective" (AI_FA06_050), "fixated on relatively trivial points ... specific feedback lacks accurate knowledge" (AI_ST15_024), and "main points were comments on how to fill in the student logs ... lack of clinical feedback" (AI_ST21_123). By contrast, supervisor excerpts highlighted the expert stance: "points out errors in writing and correctly teaches medical knowledge" (SU_FA06_037), "I can get feedback on my knowledge, so it is useful for my studies" (SU_ST17_058), and advice "that would only be conveyed by a clinician, such as the long-term perspective of 'going home'" (SU_ST23_150). Potential downsides were also noted: "content is more focused on recruiting students to the department" (SU_FA05_022) and concerns about prioritization, "few opportunities to do peripheral nerve examinations, but ... whether this is something that all medical students should learn" (SU_FA10_110), as well as an authority gradient, as exemplified by the following quote: "show the superiority and infallibility of the instructor ... makes learners feel an authority gradient" (SU_FA10_108).

Quality of the Japanese Language

Language quality emerged as a salient differentiator. AI feedback was often experienced as translation-like, occasionally inconsistent, and at times marred by formatting artifacts. Such phrasing could undermine credibility or supportive tone even when the substantive guidance was sound. Supervisor language, by comparison, was more often described as natural, human, and relationship-affirming.

XSL•FO
RenderX

Illustrative appraisals of AI language included: “Expressions such as ‘Great job!’ and ‘I will provide feedback.’ are English-Japanese translation-like ... unnatural as Japanese” (AI_ST23_149), “mysterious extra characters” (AI_ST18_083), “divided into paragraphs ... a bug or character corruption” (AI_FA08_083), and “very ChatGPT-like” (AI_ST17_054). Several noted formulaic tone: “the sandwich of good points ... is so formulaic that ... students might stop reading the feedback seriously” (AI_ST14_014), alongside word choice that “gives the impression of being a little unfriendly” (AI_FA10_108). In contrast, supervisor comments emphasized naturalness and human presence: “the advice is very good, accurate and natural” (SU_ST18_079) and “it feels like the individual is being looked at and taught” (SU_ST20_109).

Text Length

Across the dataset, AI feedback tended to be substantially longer, often enumerating many points and including concrete examples. Participants suggested that this length may help make expectations explicit and actionable, yet also noted that verbosity can hinder readability and reduce the salience of key priorities. By contrast, supervisor feedback was frequently very short—sometimes so brief that essential elements (eg, specific points for improvement) were missing. Still, several comments valued concision when it clearly conveyed the main point and respected time constraints.

Exemplars included the AI being “somewhat verbose” (AI_ST23_155) with “so many points ... it would be easier to understand if there was a summary” (AI_ST16_047). For supervisors, multiple comments judged entries too brief: “not possible to give feedback with just a few words” (SU_ST16_044), “just too short” (SU_FA05_020), “I don’t think a simple ‘Thank you for your hard work’ would be considered feedback” (SU_FA09_094), and “too simple and doesn’t amount to feedback” (SU_FA11_131). A counterexample highlighted that brevity can still work when targeted: “concise feedback that includes the necessary content ... motivates the student for future practice” (SU_ST15_034).

Integration of Quantitative and Qualitative Results

Overview

JDA was used to integrate the results of the quantitative and qualitative analyses. The integration of quantitative and qualitative findings reveals 6 convergent patterns that illuminate the differences between AI-generated and supervisor-provided feedback. The results of the quantitative analysis are shown in the second column of Table 6, while the themes identified in the qualitative analysis are shown in the third column. The fourth column shows how the quantitative and qualitative results are related. Below, we explain the integrated insights obtained from each row.

Table . Joint display integrating quantitative and qualitative findings from a convergent mixed methods study comparing artificial intelligence (AI)–generated and supervisor feedback on clinical clerkship logs, Nagoya University, Japan, 2024.

Index	Quantitative results	Qualitative results	Type of integration
1	• Feedback length difference (AI>supervisor)	• Theme: “text length”	• Quantitative results support qualitative findings • Qualitative findings explain quantitative results
2	• Comparison of rubric scores • Feedback length	• Theme: “adherence to the feedback criteria and structure” • Theme: “text length”	• Quantitative and qualitative results complement each other
3	• Correlation between feedback length and rubric score (significant for supervisors and absent for AI)	• Theme: “text length”	• Combination of quantitative and qualitative results leads to new insights
4	• Consistency of rubric scores	• Theme: “adherence to the feedback criteria and structure”	• Quantitative results support qualitative findings
5	• Comparison of rubric scores	• Theme: “continuity and consistency” • Theme: “perspective as a clinician”	• Qualitative results expand quantitative results
6	• Identification of feedback source	• Theme: “adherence to the feedback criteria and structure” • Theme: “continuity and consistency” • Theme: “perspective as a clinician” • Theme: “quality of the Japanese language” • Theme: “text length”	• Qualitative findings explain quantitative results

XSL•FO
RenderX

Index 1: Text Length and Feedback Quality

The quantitative finding that AI feedback was significantly longer than supervisor feedback was substantiated by the qualitative theme of text length. Qualitative analysis revealed that AI feedback length enabled the provision of specific examples, contributing to clarity, while potentially resulting in verbose expressions with unclear priorities. Conversely, supervisor feedback brevity sometimes led to missing important elements but could be highly effective when concise and targeted. This integration demonstrates that qualitative findings explain the impact of length differences on feedback quality, which quantitative analysis alone could not reveal.

Index 2: Adherence to Rubric Criteria

Quantitatively, AI feedback scored significantly higher than supervisor feedback on the clear-direction and criteria-based items and showed a nonsignificant trend toward higher scores on the remaining items. These results were complemented by the qualitative theme of adherence to feedback criteria and structure and by the difference in text length. Qualitative analysis revealed that AI feedback followed the rubric with clear structure and provided actionable points, whereas supervisor feedback was often too brief and lacking some required elements. The significant difference in feedback length was also demonstrated quantitatively. This complementary integration explains how AI feedback aligns with the rubric's structural criteria through both structural adherence and sufficient length.

Index 3: Length-Score Correlation Patterns

A particularly illuminating pattern emerged from combining quantitative correlation analysis with qualitative text length observations. Quantitative examination revealed that while there was no correlation between length and rubric scores in AI feedback, longer submissions received higher scores in supervisor feedback. Qualitative examination showed that AI feedback length enables specific examples to be provided, leading to clarity, while potentially resulting in verbose expressions with unclear priorities. Supervisor feedback brevity can lead to missing important elements but can be highly effective when concise and targeted. This suggests that AI more systematically covers the elements specified by the rubric, while supervisors excel at contextual prioritization. Combining these findings reveals that feedback requires sufficient length to include necessary items, while prioritization is also important, and longer is not necessarily better.

Index 4: Rubric-Score Consistency and Variability

The quantitative finding that AI feedback rubric scores were more consistent while supervisor feedback rubric scores were more variable was supported by qualitative observations regarding adherence to feedback criteria and structure. The qualitative analysis described AI feedback as closely following the rubric, leading to more uniform rubric scores, whereas supervisors often deviated from the rubric, resulting in variable scores. This integration demonstrates how structural consistency in AI feedback translates directly into measurable rubric-score consistency.

Index 5: Clinical Context and Perspective

The quantitative results regarding rubric scores were expanded by qualitative themes of continuity and consistency, as well as perspective as a clinician. Qualitative analysis revealed that AI feedback maintained continuity with written logs and provided specific text-level suggestions while missing broader clinical context. In contrast, supervisor feedback aligned with clinical experience but often failed to engage with specific portfolio content. These qualitative findings revealed aspects beyond those evaluated by the rubric, specifically demonstrating that supervisor feedback was grounded in clinical context and used expert perspectives, showing advantages not captured by the rubric scores.

Index 6: Distinguishability of Feedback Sources

The quantitative finding that all evaluators could completely distinguish between AI and supervisor feedback was explained by all 5 qualitative themes identified in the analysis. The themes of adherence to feedback criteria and structure, continuity and consistency, perspective as a clinician, quality of Japanese language, and text length each highlighted distinct differences between the 2 feedback sources. AI feedback followed the rubric with clear structure, maintained continuity with written logs, and was long with specific examples, while supervisor feedback excelled at contextual prioritization and provided feedback from a clinician perspective with natural, human expressions. These differences suggest why evaluators could distinguish feedback providers, though the analysis cannot determine whether evaluators considered these factors in combination or whether specific elements were decisive in their identification process.

Discussion

Principal Findings

This study compared AI-generated feedback with supervisor-provided feedback on medical student clinical clerkship logs through mixed methods analysis, integrating results using a JDA. AI feedback was significantly longer and demonstrated greater adherence to rubric-based criteria, particularly in providing criteria-based guidance and clear directions for improvement. While supervisor feedback showed greater variability in rubric scores, with length correlating positively with those scores, AI feedback maintained consistent rubric scores regardless of length, suggesting systematic coverage of the structural elements captured by the rubric. Qualitative analysis revealed complementary strengths: AI provided text-anchored feedback with structured, comprehensive coverage of student log content, while supervisors offered experience-based insights grounded in clinical context and professional expertise. Despite AI's structural advantages, supervisor feedback was valued for its clinical perspective and natural language use. All evaluators could reliably distinguish between the 2 feedback sources, reflecting substantial differences in approach, structure, and content focus.

Comparison With the Literature

Our findings align with and extend previous research on AI-generated feedback in several key areas. Regarding feedback length, prior studies have reported various findings [20,23-25].

XSL•FO
RenderX

Our study demonstrated that AI feedback was significantly longer than supervisor feedback, which is consistent with some previous findings [20,23,24]. In our research, the supervisors providing feedback were working in busy clinical environments, which likely contributed to their tendency to provide shorter feedback. Furthermore, our results showed no correlation between AI feedback length and rubric scores, while supervisor feedback demonstrated a positive correlation between length and rubric scores. Some reports indicate that feedback length does not contribute to quality improvement beyond a certain threshold [51]. Considering that AI feedback length was sufficiently longer compared to supervisor feedback, our findings are consistent with these previous studies. Prior research has pointed out that busy physicians tend to provide short feedback, often lacking important elements such as specific examples and future action guidelines [28,52]. Our findings suggest that AI-generated feedback may be able to complement elements that tend to be missing in supervisors' brief feedback.

Regarding consistency in feedback quality, previous research has shown mixed results about whether AI or human feedback demonstrates greater consistency [25,29-32]. Our study found that AI feedback exhibited higher consistency in rubric scores compared to supervisor feedback. It is important to distinguish AI's adherence to rubric criteria, which follows from embedding those criteria in the prompt, from the consistency of its output across different inputs. The latter is a structural consequence of using the same model and prompt configuration for every log entry, which constrains the range of possible outputs. This consistency is therefore best understood as a design feature of AI feedback generation rather than an inherently desirable educational property, as effective feedback may sometimes require selective emphasis tailored to individual learner needs. In contrast, the greater variability observed in supervisor feedback likely reflects selective prioritization and clinical judgment: supervisors chose which aspects to emphasize based on their professional assessment of each student's needs, resulting in variable but contextually responsive feedback. Previous research has identified multiple factors that influence feedback quality variability, including feedback length [29], the subject of feedback [30], and whether supervisors have received feedback training [53]. Our findings suggest that in busy clinical environments, AI could provide structurally stable feedback that complements the individualized, experience-driven feedback of supervisors.

The distinction between text-anchored versus experience-based feedback represents a novel contribution of our study. Most prior comparisons of AI and human feedback have been conducted in classroom or simulation settings [22,24,25,54], where the knowledge base is relatively well-defined and the feedback task is circumscribed. Our study extends this comparison to an authentic clinical clerkship environment, where several contextual features distinguish the feedback task. Supervisors work under substantial time pressure and provide written feedback as a complement to ongoing face-to-face clinical teaching, resulting in brief comments that prioritize what each supervisor judges most important for a given student. Clinical reasoning, situational awareness, and tacit professional knowledge, which are central to clerkship learning [55], are difficult to articulate in written form [56] and largely inaccessible to AI systems that operate solely on the written log text. Previous research has noted that AI feedback tends to provide comprehensive feedback closely tied to the target text [24,25], while human experts in specialized fields provide context-informed feedback that is perceived to be of higher quality [22,54]. Our findings confirm and deepen this pattern: AI feedback was more likely to provide text-consistent feedback aligned with students' written clerkship logs, while supervisor feedback drew on clinical experience, the broader educational relationship, and professional judgment that extends well beyond what is documented in the log. These contextual features of clinical clerkships amplify the qualitative differences between AI and supervisor feedback compared to what has been observed in classroom settings and underscore the importance of studying AI feedback in authentic clinical environments where the complementary nature of each feedback source is most evident.

Regarding the ability to distinguish between AI and human feedback, previous studies have reported inconsistent findings [33-35]. In our study, all evaluators correctly identified the source of feedback with 100% accuracy, despite blinding procedures that included randomized presentation order, removal of source labels, and placement of the identification question after all rubric scoring and free-text comments had been completed. The differences we observed in text length, feedback structure, and clinical perspective in the busy clinical environment likely served as decisive factors for identification, consistent with all 5 qualitative themes that highlighted distinct textual characteristics of each feedback type.

This perfect detection has implications that warrant discussion along 2 distinct lines. The first concerns the evaluation itself: because evaluators recognized the feedback source with complete accuracy, they may have formed implicit judgments about the source during rubric scoring even though they were not explicitly informed until afterward, and the possibility that such recognition shaped ratings cannot be ruled out. Prior research has demonstrated that knowledge or perception of the feedback provider can influence how feedback is received and evaluated [45,46], and a recent controlled-content experiment further showed that source attribution altered learner engagement by a large margin (Cohen d=0.88 - 1.56) even when the feedback content was held identical [57]. In our study, however, the identification question was administered only after all rubric scoring and free-text comments had been completed, so any such influence would have arisen from the textual characteristics of the feedback rather than from explicit labels, unlike an unblinded evaluation in which the source is known a priori.

The perfect detection also constitutes a substantive finding in its own right. The 5 qualitative themes identified in our analysis (adherence to feedback criteria and structure, continuity and consistency, perspective as a clinician, quality of the Japanese language, and text length) collectively describe the lexical, structural, and contextual features that made AI feedback identifiable in this Japanese-language clinical context, with language-level cues such as translation-like phrasing, formatting artifacts, and a formulaic register being particularly salient. At the level of interpretation, these differences suggest that AI and supervisor feedback differ not only in structural properties but

XSL•FO
RenderX

in the epistemic voice they carry: supervisor feedback is grounded in lived clinical encounter, whereas AI feedback operates on textual pattern. Similar tensions around authenticity and voice have been raised in recent medical education writing on AI-assisted reflective work [58]. This voice differential can be read positively as a form of source transparency, since learners and educators are unlikely to mistake one source for the other, but also as an authenticity gap that may affect trust and uptake when AI feedback is delivered to learners. Newer LLMs continue to improve in generating natural, human-like text, and future iterations may substantially reduce the detectability of AI-generated feedback, potentially altering both evaluator perceptions and the dynamics of blinded evaluation designs.

An important consideration for interpreting our findings is the linguistic and cultural context in which this study was conducted. All feedback was generated and evaluated in Japanese, and LLM performance has been shown to vary substantially across languages. Strasser et al [59] demonstrated that LLM accuracy on identical medical examination questions differed significantly by language (64% - 87%), with English prompts generally yielding higher performance, although advanced models showed comparable results with language-matched prompts in some languages. Harigai et al [60] found that GPT-4 achieved significantly higher accuracy on English-translated radiology questions (median 89 points) than on the original Japanese versions (median 70 points), attributing this gap to the predominance of English in training data and to translation challenges specific to Japanese, including logographic script ambiguity and structural differences. These findings suggest that the quality of AI-generated feedback in our study may be partly shaped by GPT-4o's language-specific capabilities in Japanese, and that the same prompt and model configuration could produce qualitatively different feedback in other languages. Furthermore, the cultural context of Japanese medical education, where deference to senior clinicians is strongly emphasized [61], may have influenced evaluators' perceptions of and receptivity to the 2 feedback sources. Rather than framing these contextual factors solely as limitations, we note that they underscore the importance of conducting AI feedback research across diverse linguistic and cultural settings, as both AI performance and user reception are likely to be context-dependent.

Implications of Findings

These findings have important implications for the integration of AI-generated feedback in clinical education. Our results suggest that AI-generated and supervisor-provided feedback each possess distinct strengths and limitations and should be leveraged in a complementary manner rather than as replacements for one another. For instance, AI could provide structured, criterion-based feedback alongside supervisor feedback, allowing medical students to benefit from both perspectives. However, such direct provision of AI feedback requires careful consideration of potential risks, including student overreliance on AI recommendations [62] or conversely, dismissal of AI-generated insights [20]. Our finding that evaluators recognized feedback source with perfect accuracy suggests that learners in real-world settings would plausibly do the same, and evidence from controlled-content experiments indicates that source attribution can substantially alter learner engagement when feedback is delivered directly (Cohen d=0.88 - 1.56) [57]. The rubric-score advantages we observed should therefore not be assumed to translate directly into comparable learner uptake or educational impact when AI feedback is delivered to students who can identify its source.

Various hybrid approaches merit exploration, including AI-drafted feedback refined by educators, role-based distribution where AI and supervisors focus on different aspects of student work, or AI enhancement of educator-authored feedback [34,63]. However, empirical research on these hybrid models remains limited, particularly in authentic clinical settings.

Underlying both direct delivery and hybrid configurations is the prompt itself, an object of pedagogical design that shapes what the AI produces in any deployment model. In our case, embedding the rubric criteria directly in the prompt operationalized a particular conception of feedback quality in AI generation. This is analogous to constructive alignment in instructional design [64], where intended outcomes are aligned with the activities producing them. Recent applied guidance in health professions education has similarly recommended grounding generative AI prompts in institutional pedagogical and assessment frameworks [65]. The rubric-score consistency we observed for AI feedback is therefore best read not as an inherent property of GPT-4o but as an outcome of this prompt configuration, consistent with experimental evidence that detailed prompts combined with low temperature settings yield near-perfect interreplicate consistency in LLM rubric grading [66]. The same design lever is available to other educators and institutions seeking to align AI feedback with their own assessment frameworks.

Future research should examine how learners respond to and use AI-generated feedback and investigate the effectiveness of hybrid approaches that combine AI and supervisor feedback. Such studies will be essential for determining optimal implementation strategies for AI feedback systems in clinical clerkship environments, ensuring that technological capabilities enhance rather than disrupt the educational relationships and learning processes that are fundamental to clinical training.

Limitations

Several limitations should be acknowledged in this study.

First, as this research was conducted at a single institution in Japan, the generalizability of our findings to other institutions or cultural contexts needs to be confirmed. The study was conducted entirely in Japanese, and GPT-4o's performance in Japanese may differ from its performance in English, potentially affecting the naturalness and quality of AI-generated feedback. Additionally, the cultural context of Japanese medical education, where deference to senior clinicians is strongly emphasized [61], may have contributed to higher regard for supervisor feedback and influenced evaluators' perceptions of the 2 feedback sources. The specific characteristics of NU's clinical clerkship program may further limit representativeness.

Second, while our evaluators included both faculty and students, their individual backgrounds and experiences may have

XSL•FO
RenderX

influenced their feedback evaluations. The relatively small number of evaluators (10 faculty and 10 students) may also limit the robustness of our findings.

Third, all evaluators correctly identified the source of every feedback entry despite blinding procedures (randomized presentation order and removal of source labels). Although the identification question was administered only after all rubric scoring and free-text comments had been completed, the perfect detection accuracy indicates that the textual characteristics of the 2 feedback types were highly salient and that evaluators may have formed implicit judgments about the source during the evaluation process. Because all evaluators correctly identified every item, perceived source and actual source are identical, making it logically impossible to stratify scores by perceived source to estimate the magnitude of any expectancy bias. Readers should therefore consider the possibility that implicit source recognition may have influenced rubric ratings, particularly for items on which AI scored highest.

Fourth, this study used a specific AI model (GPT-4o) for feedback generation, and results may differ with other AI models or future iterations of the technology. The prompt engineering process, while systematic, was tailored to our specific context and may not be optimal for other settings.

Fifth, the evaluation rubric used to assess feedback quality was originally developed for written academic feedback [18] and emphasizes structural dimensions such as criteria-based guidance, clarity of directions, and text alignment. Clinically meaningful attributes of feedback, such as credibility, diagnostic reasoning, situational awareness, and professional judgment, are not directly captured by this instrument. The rubric score differences between AI and supervisor feedback therefore reflect adherence to these specific structural dimensions rather than overall educational quality. The qualitative findings, particularly the themes of perspective as a clinician and continuity with practice, provide complementary evidence of the clinical value that supervisors contribute and that the rubric does not score.

Sixth, the absence of a prompt sensitivity analysis limits the interpretability of our findings. It remains unclear whether the observed rubric score advantages for AI feedback reflect inherent properties of GPT-4o or are contingent on this specific, carefully engineered prompt. Alternative prompts, whether shorter, less structured, or more clinically oriented, might produce qualitatively different comparative outcomes. Conducting a systematic sensitivity analysis was beyond the scope of this study, as the prompt was developed iteratively through a pilot testing process, and varying it systematically would constitute a separate study design. This remains an important direction for future research.

Seventh, each dataset was evaluated by only 1 faculty member and 1 student, rather than by multiple independent raters. Interrater agreement was fair to moderate for supervisor feedback but poor for AI feedback, indicating that individual evaluator perspectives influenced ratings, particularly when assessing AI-generated output. Although the large number of datasets (n=161) provides stability at the aggregate level, the precision of scores for individual datasets is limited.

Eighth, our study evaluated the structural quality of feedback as assessed by third-party evaluators using the rubric, but did not examine whether AI-generated or supervisor-provided feedback, when delivered to students, elicited different learning behaviors, revisions, engagement, or clinical improvement. Because our design measured judgments by evaluators rather than responses by learner recipients, source-attribution effects that may arise when identifiable AI feedback is delivered directly to students [57] fall outside the scope of our data. The relationship between perceived feedback quality, source attribution, and actual educational outcomes remains an important area for future investigation.

Conclusions

This study extends the comparison of AI-generated and supervisor feedback from classroom and simulation settings to an authentic clinical clerkship environment, where time pressure, complex professional expertise, and ongoing educational relationships shape the feedback process. Through integrated mixed methods analysis, a key distinction emerged between text-anchored AI feedback, which systematically addresses written log content in alignment with rubric criteria, and experience-based supervisor feedback, which draws on clinical observation and professional judgment. Using GPT-4o with a rubric-embedded prompt, AI consistently delivered structured feedback addressing gaps that arise when time-pressured supervisors provide brief comments, while supervisors contributed clinically grounded insights that AI cannot replicate. These complementary strengths suggest that AI feedback should supplement rather than replace supervisor feedback in clinical education. Future research should investigate hybrid models that leverage each type's advantages; examine whether these patterns generalize to other AI models, languages, and institutional contexts; and evaluate the impact of AI-assisted feedback on student learning outcomes.

Acknowledgments

The authors thank the following evaluators for their contribution to this study: Aya Shimotomai, Dai Iwata, Eishin Muramatsu, Hidekazu Kito, Hideki Takemi, Hiroaki Hakamata, Hirotake Machida, Kota Tokunaga, Mayuko Kitayama, Minoru Kishi, Naoki Yamada, Naomi Hino-Fukuyo, Nobuhiro Miyamae, Rikuto Kikunaga, Sou Suzuki, Takami Ishimoto, Takeshi Onoue, Tomokazu Kimura, Usukura Taiki, Yutaka Yamada, Yoshiki Tsumura, and Yuya Matsuura. The authors also thank Noriki Kamihiro, Oliver Stanyon, Naoki Yamada, and Dai Iwata for their contributions to the translation and pretesting of the evaluation rubric. The authors express their deep gratitude to all of them for their cooperation as evaluators in this study. GitHub Copilot (version 0.35.2; artificial intelligence [AI] models used: Claude Sonnet 4.5/Claude Opus 4.5/GPT-4.1/GPT-5) and Claude Code (Claude Opus 4.5) were used to assist in writing R code for quantitative analysis. The editor's predictive code-generation feature was used for

XSL•FO
RenderX

in-editor assistance. All generated code was reviewed, tested, and modified by TK and reviewed by JD. For manuscript writing, the same AI agent was used to assist with translation from Japanese into English. After entering content in Japanese in the editor, we prompted the in-editor chat with “translate to English” and used the generated English text as the basis for additions and revisions to the manuscript. All generated text was checked and edited by TK and reviewed and revised by all authors. For assistance with literature searches, Deep Research by ChatGPT (GPT-4o, GPT-5, and GPT-5.1) was used. All selected papers were reviewed by TK, and results from hand searches were also compared and incorporated. According to the Generative AI Delegation Taxonomy [67], the following tasks were delegated to generative AI tools under full human supervision: literature search and systematization, code generation, and translation. Generative AI tools are not listed as authors and do not bear responsibility for the final outcomes. Responsibility for the final manuscript lies entirely with the authors.

Funding

This work was supported by JSPS KAKENHI (grant 25K06542). The funder had no role in the study design, data collection, analysis, interpretation of results, or the writing of the manuscript.

Data Availability

The anonymized evaluation data, analysis code, and the source code of the artificial intelligence (AI) feedback generation system are publicly available [68]. The AI feedback generation prompt, the evaluation rubric, its translation process, and the full paired 2-tailed *t* test results table are provided in Multimedia Appendices 1-4, respectively. The original clinical clerkship logs are not publicly available because some entries may contain personally identifiable information about medical students, and public release of the log texts would exceed the scope of consent obtained through the opt-out procedure approved by the ethics committee.

Authors' Contributions

TK planned the study, developed the evaluation tools, created the research-related code, collected and analyzed the data, and wrote the manuscript. JD contributed to the study design, mainly focusing on the quantitative methods and evaluation of results, and improved the manuscript. NH contributed to the overall study design, organized the implementation structure for both quantitative and qualitative analyses, developed and translated the evaluation tools, and improved the manuscript. SR contributed to the study design, mainly focusing on the qualitative methods and evaluation of results, and improved the manuscript. SH contributed to the overall study design, coordinated the team, mainly focused on the qualitative methods and evaluation of results, and improved the manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
The feedback generation prompt for artificial intelligence used in this study and its development process.
[DOCX File, 17 KB - mededu_v12i1e90064_app1.docx]

Multimedia Appendix 2
The feedback evaluation rubric used in this study.
[PDF File, 145 KB - mededu_v12i1e90064_app2.pdf]

Multimedia Appendix 3
The translation process into Japanese for the feedback evaluation rubric used in this study.
[DOCX File, 14 KB - mededu_v12i1e90064_app3.docx]

Multimedia Appendix 4
Distribution of rubric score differences between artificial intelligence–generated and supervisor feedback by rubric item and assessor type.
[DOCX File, 52 KB - mededu_v12i1e90064_app4.docx]

Multimedia Appendix 5
The full paired *t* test results comparing artificial intelligence and supervisor feedback scores on the Steiss et al 5-item analytic rubric, including mean differences and uncorrected *P* values omitted from Table 2 for readability.
[DOCX File, 19 KB - mededu_v12i1e90064_app5.docx]

References

XSL•FO
RenderX

1. Hansen A, Klute RM, Yadav M, Bansal S, Bond WF. How do learners receive feedback on note writing? A scoping review. Acad Med 2024 Jun 1;99(6):683-690. [doi: 10.1097/ACM.0000000000005653] [Medline: 38306581]
2. Ende J. Feedback in clinical medical education. JAMA 1983 Aug 12;250(6):777-781. [Medline: 6876333]
3. Norcini JJ, Burch V. Workplace-based assessment as an educational tool: AMEE guide no. 31. Med Teach 2007 Nov;29(9):855-871. [doi: 10.1080/01421590701775453] [Medline: 18158655]
4. Burgess A, van Diggele C, Roberts C, Mellis C. Feedback in the clinical setting. BMC Med Educ 2020 Dec 3;20(Suppl 2):460. [doi: 10.1186/s12909-020-02280-5] [Medline: 33272265]
5. Elnicki DM, Layne RD, Ogden PE, Morris DK. Oral versus written feedback in medical clinic. J Gen Intern Med 1998 Mar;13(3):155-158. [doi: 10.1046/j.1525-1497.1998.00049.x] [Medline: 9541371]
6. Harvey P, Radomski N, O'Connor D. Written feedback and continuity of learning in a geographically distributed medical education program. Med Teach 2013 Dec;35(12):1009-1013. [doi: 10.3109/0142159X.2013.827332] [Medline: 23971889]
7. Greviana N, Pinasthika A, Soemantri D. The best of both worlds: exploring the interplay between written and verbal feedback in the continuum of medical education. Med Sci Educ 2025 Apr;35(2):863-871. [doi: 10.1007/s40670-024-02244-0] [Medline: 40353019]
8. van der Vleuten C, Lindemann I, Schmidt L. Programmatic assessment: the process, rationale and evidence for modern evaluation approaches in medical education. Med J Aust 2018 Nov 5;209(9):386-388. [doi: 10.5694/mja17.00926] [Medline: 30376659]
9. Branfield Day L, Miles A, Ginsburg S, Melvin L. Resident perceptions of assessment and feedback in competency-based medical education: a focus group study of one internal medicine residency program. Acad Med 2020 Nov;95(11):1712-1717. [doi: 10.1097/ACM.0000000000003315] [Medline: 32195692]
10. Canavan C, Holtman MC, Richmond M, Katsufrakis PJ. The quality of written comments on professional behaviors in a developmental multisource feedback program. Acad Med 2010 Oct;85(10 Suppl):S106-S109. [doi: 10.1097/ACM.0b013e3181ed4cdb] [Medline: 20881691]
11. Lucas HC, Upperman JS, Robinson JR. A systematic review of large language models and their implications in medical education. Med Educ 2024 Nov;58(11):1276-1285. [doi: 10.1111/medu.15402] [Medline: 38639098]
12. Dai W, Lin J, Jin H, et al. Can large language models provide feedback to students? A case study on ChatGPT. Presented at: 2023 IEEE International Conference on Advanced Learning Technologies (ICALT); Jul 10-13, 2023. [doi: 10.1109/ICALT58122.2023.00100]
13. Jarry Trujillo C, Vela Ulloa J, Escalona Vivas G, et al. Surgeons vs ChatGPT: assessment and feedback performance based on real surgical scenarios. J Surg Educ 2024 Jul;81(7):960-966. [doi: 10.1016/j.jsurg.2024.03.012] [Medline: 38749814]
14. Abd-Alrazaq A, AlSaad R, Alhuwail D, et al. Large language models in medical education: opportunities, challenges, and future directions. JMIR Med Educ 2023 Jun 1;9:e48291. [doi: 10.2196/48291] [Medline: 37261894]
15. LeCun Y, Bengio Y, Hinton G. Deep learning. Nature New Biol 2015 May 28;521(7553):436-444. [doi: 10.1038/nature14539] [Medline: 26017442]
16. Khan RA, Jawaid M, Khan AR, Sajjad M. ChatGPT—reshaping medical education and clinical management. Pak J Med Sci 2023;39(2):605-607. [doi: 10.12669/pjms.39.2.7653] [Medline: 36950398]
17. Lee H. The rise of ChatGPT: exploring its potential in medical education. Anat Sci Educ 2024;17(5):926-931. [doi: 10.1002/ase.2270] [Medline: 36916887]
18. Steiss J, Tate T, Graham S, et al. Comparing the quality of human and ChatGPT feedback of students' writing. Learn Instr 2024 Jun;91:101894. [doi: 10.1016/j.learninstruc.2024.101894]
19. Ortega-Ochoa E, Quiroga Pérez J, Arguedas M, Daradoumis T, Marquès Puig JM. The effectiveness of empathic chatbot feedback for developing computer competencies, motivation, self-regulation, and metacognitive reasoning in online higher education. Internet Things 2024 Apr;25:101101. [doi: 10.1016/j.iot.2024.101101]
20. Jayawardena CK, Gunathilake Y, Ihalagedara D. Dental students' learning experience: artificial intelligence vs human feedback on assignments. Int Dent J 2025 Feb;75(1):100-108. [doi: 10.1016/j.identj.2024.12.022] [Medline: 39799065]
21. Haider N, Morjaria L, Sheth U, Al-Jabouri N, Sibbald M. MD student perceptions of ChatGPT for reflective writing feedback in undergraduate medical education. Int Med Educ 2025 Jul 23;4(3):27. [doi: 10.3390/ime4030027]
22. Çiçek FE, Ülker M, Özer M, Kıyak YS. ChatGPT versus expert feedback on clinical reasoning questions and their effect on learning: a randomized controlled trial. Postgrad Med J 2025 Apr 22;101(1195):458-463. [doi: 10.1093/postmj/qgae170] [Medline: 39656920]
23. Usher M. Generative AI vs. instructor vs. peer assessments: a comparison of grading and feedback in higher education. Assess Eval High Educ 2025 Aug 18;50(6):912-927. [doi: 10.1080/02602938.2025.2487495]
24. Er E, Akçapınar G, Bayazıt A, Noroozi O, Banihashem SK. Assessing student perceptions and use of instructor versus AI ‐ generated feedback. Br J Educ Technol 2025 May;56(3):1074-1091. [doi: 10.1111/bjet.13558]
25. Ali M, Harbieh I, Haider KH. Bytes versus brains: a comparative study of AI-generated feedback and human tutor feedback in medical education. Med Teach 2026 Jan;48(1):131-141. [doi: 10.1080/0142159X.2025.2519639] [Medline: 40528415]
26. Andreou V, Peters S, Eggermont J, Schoenmakers B. Evaluating feedback comments in entrustable professional activities: a cross-sectional study. J Med Educ Curric Dev 2024;11:23821205241275810. [doi: 10.1177/23821205241275810] [Medline: 39346122]

XSL•FO
RenderX

27. Poeppelman RS, Cho J, Nachbor K, et al. “But why?”: Explanatory feedback is a reliable marker of high-quality narrative assessment of surgical performance. Acad Med 2025 May 1;100(5):614-620. [doi: 10.1097/ACM.0000000000005985] [Medline: 39889694]
28. Lu KJQ, Meaney C, Guo E, Leung FH. Evaluating the applicability of existing lexicon-based sentiment analysis techniques on family medicine resident feedback field notes: retrospective cohort study. JMIR Med Educ 2023 Jul 27;9:e41953. [doi: 10.2196/41953] [Medline: 37498660]
29. Mooney CJ, Pascoe JM, Blatt AE, et al. Predictors of faculty narrative evaluation quality in medical school clerkships. Med Educ 2022 Dec;56(12):1223-1231. [doi: 10.1111/medu.14911] [Medline: 35950329]
30. Coughlin RF, Tsyrulnik A, Gottlieb M, et al. Differences in faculty feedback for high, expected, and below-expected clinically performing emergency medicine residents. AEM Educ Train 2022 Oct;6(5):e10788. [doi: 10.1002/aet2.10788] [Medline: 36189452]
31. Van Ostaeyen S, Embo M, Rotsaert T, De Clercq O, Schellens T, Valcke M. A qualitative textual analysis of feedback comments in ePortfolios: quality and alignment with the CanMEDS roles. Perspect Med Educ 2023;12(1):584-593. [doi: 10.5334/pme.1050] [Medline: 38144672]
32. Lin S, Crosthwaite P. The grass is not always greener: teacher vs. GPT-assisted written corrective feedback. System 2024 Dec;127:103529. [doi: 10.1016/j.system.2024.103529]
33. Nazaretsky T, Mejia-Domenzain P, Swamy V, Frej J, Käser T. AI or human? Evaluating student feedback perceptions in higher education. In: Technology Enhanced Learning for Inclusive and Equitable Quality Education: Springer Nature; 2024:284-298. [doi: 10.1007/978-3-031-72315-5_20]
34. Zhang A, Gao Y, Suraworachet W, Nazaretsky T, Cukurova M. Evaluating trust in AI, human, and co-produced feedback among undergraduate students. arXiv. Preprint posted online on Apr 15, 2025. [doi: 10.48550/arXiv.2504.10961]
35. Rüdian S, Podelo J, Kužílek J, Pinkwart N. Feedback on feedback: students' perceptions for feedback from teachers and few-shot LLMs. Presented at: 15th International Learning Analytics and Knowledge Conference (LAK ’25); Mar 3-7, 2025. [doi: 10.1145/3706468.3706479]
36. Nishigori H. Medical education in Japan. Med Teach 2024 Sep;46(sup1):S4-S10. [doi: 10.1080/0142159X.2024.2372108] [Medline: 39545499]
37. Fetters MD, Curry LA, Creswell JW. Achieving integration in mixed methods designs—principles and practices. Health Serv Res 2013 Dec;48(6 Pt 2):2134-2156. [doi: 10.1111/1475-6773.12117] [Medline: 24279835]
38. Maudsley G. Mixing it but not mixed-up: mixed methods research in medical education (a critical narrative review). Med Teach 2011;33(2):e92-104. [doi: 10.3109/0142159X.2011.542523] [Medline: 21275539]
39. Guetterman TC, Fetters MD, Creswell JW. Integrating quantitative and qualitative results in health science mixed methods research through joint displays. Ann Fam Med 2015 Nov;13(6):554-561. [doi: 10.1370/afm.1865] [Medline: 26553895]
40. McCrudden MT, Marchand G, Schutz PA. Joint displays for mixed methods research in psychology. Methods Psychol 2021 Dec;5:100067. [doi: 10.1016/j.metip.2021.100067]
41. Jacobsen LJ, Weber KE. The promises and pitfalls of ChatGPT as a feedback provider in higher education: an exploratory study of prompt engineering and the quality of AI-driven feedback. OSF Preprints. Preprint posted online on Sep 29, 2023. [doi: 10.31219/osf.io/cr257]
42. Van Ostaeyen S, De Langhe L, De Clercq O, Embo M, Schellens T, Valcke M. Automating the identification of feedback quality criteria and the CanMEDS roles in written feedback comments using natural language processing. Perspect Med Educ 2023;12(1):540-549. [doi: 10.5334/pme.1056] [Medline: 38144670]
43. GPT-4o. OpenAI Developers. URL: https://platform.openai.com/docs/models/gpt-4o [accessed 2026-05-30]
44. Walde P, Völlm BA. The TRAPD approach as a method for questionnaire translation. Front Psychiatry 2023;14:1199989. [doi: 10.3389/fpsyt.2023.1199989] [Medline: 37398605]
45. van de Ridder JMM, Berk FCJ, Stokking KM, Ten Cate OTJ. Feedback providers’ credibility impacts students’ satisfaction with feedback and delayed performance. Med Teach 2015 Aug;37(8):767-774. [doi: 10.3109/0142159X.2014.970617] [Medline: 25306959]
46. Tripodi N, Feehan J, Wospil R, Vaughan B. Twelve tips for developing feedback literacy in health professions learners. Med Teach 2021 Aug;43(8):960-965. [doi: 10.1080/0142159X.2020.1839035] [Medline: 33131374]
47. Zhang L, Komachi M. Neural machine translation of logographic language using sub-character level information. Presented at: Proceedings of the Third Conference on Machine Translation; Oct 31 to Nov 1, 2018. [doi: 10.18653/v1/W18-6303]
48. Nowell LS, Norris JM, White DE, Moules NJ. Thematic analysis. Int J Qual Methods 2017 Dec 1;16(1). [doi: 10.1177/1609406917733847]
49. Levitt HM, Bamberg M, Creswell JW, Frost DM, Josselson R, Suárez-Orozco C. Journal article reporting standards for qualitative primary, qualitative meta-analytic, and mixed methods research in psychology: the APA Publications and Communications Board task force report. Am Psychol 2018 Jan;73(1):26-46. [doi: 10.1037/amp0000151] [Medline: 29345485]
50. Ahmed SK, Mohammed RA, Nashwan AJ, et al. Using thematic analysis in qualitative research. J Med Surg Public Health 2025 Aug;6:100198. [doi: 10.1016/j.glmedi.2025.100198]

XSL•FO
RenderX

51. Curtis R, Moon CC, Hanmore T, Hopman WM, Baxter S. Use the right words: evaluating the effect of word choice and word count on quality of narrative feedback in ophthalmology competency-based medical education assessments. Can Med Educ J 2024 Dec;15(6):58-63. [doi: 10.36834/cmej.76671] [Medline: 39807150]
52. Leclair R, Ho JSS, Braund H, et al. Exploring the quality of narrative feedback provided to residents during ambulatory patient care in medicine and surgery. J Med Educ Curric Dev 2023;10:23821205231175734. [doi: 10.1177/23821205231175734] [Medline: 37216002]
53. Renting N, Jaarsma D, Borleffs JC, Slaets JPJ, Cohen-Schotanus J, Gans ROB. Effectiveness of a supervisor training on quality of feedback to internal medicine residents: a controlled longitudinal multicentre study. BMJ Open 2023 Sep 28;13(9):e076946. [doi: 10.1136/bmjopen-2023-076946] [Medline: 37770280]
54. Zhou Z, Rizwan A, Rogoza N, Chung AD, Kwan BY. Differentiating between GPT-generated and human-written feedback for radiology residents. Curr Probl Diagn Radiol 2025 Sep;54(5):574-578. [doi: 10.1067/j.cpradiol.2025.02.002]
55. Dornan T, Tan N, Boshuizen H, et al. How and what do medical students learn in clerkships? Experience based learning (ExBL). Adv Health Sci Educ Theory Pract 2014 Dec;19(5):721-749. [doi: 10.1007/s10459-014-9501-0] [Medline: 24638146]
56. Yardley S, Teunissen PW, Dornan T. Experiential learning: AMEE guide no. 63. Med Teach 2012;34(2):e102-e115. [doi: 10.3109/0142159X.2012.650741] [Medline: 22289008]
57. Morris C, Feedback MPS. Same feedback, different source: how AI vs. human feedback shapes learner engagement. arXiv. Preprint posted online on Feb 11, 2026. [doi: 10.48550/ARXIV.2602.11311]
58. Maher J, Byszewski A, Lochnan H. Whose voice is it anyway? Artificial intelligence and the new crisis of authenticity in medical education. Perspect Med Educ 2026;15(1):351-355. [doi: 10.5334/pme.2265] [Medline: 41938740]
59. Strasser LM, Anschuetz W, Dennstädt F, Hastings J. Performance evaluation of large language models in multilingual medical multiple-choice questions: mixed methods study. JMIR Med Educ 2026 Mar 5;12:e81399. [doi: 10.2196/81399] [Medline: 41813244]
60. Harigai A, Toyama Y, Nagano M, et al. Response accuracy of GPT-4 across languages: insights from an expert-level diagnostic radiology examination in Japan. Jpn J Radiol 2025 Feb;43(2):319-329. [doi: 10.1007/s11604-024-01673-6] [Medline: 39466356]
61. Nishigori H, Harrison R, Busari J, Dornan T. Bushido and medical professionalism in Japan. Acad Med 2014 Apr;89(4):560-563. [doi: 10.1097/ACM.0000000000000176] [Medline: 24556758]
62. Khojasteh L, Kafipour R, Pakdel F, Mukundan J. Empowering medical students with AI writing co-pilots: design and validation of AI self-assessment toolkit. BMC Med Educ 2025 Jan 31;25(1):159. [doi: 10.1186/s12909-025-06753-3] [Medline: 39891148]
63. Banihashem SK, Noroozi O, Khosravi H, Schunn CD, Drachsler H. Pedagogical framework for hybrid intelligent feedback. Innov Educ Teach Int 2026 Mar 4;63(2):554-570. [doi: 10.1080/14703297.2025.2499174]
64. Biggs J. Enhancing teaching through constructive alignment. High Educ 1996 Oct;32(3):347-364. [doi: 10.1007/BF00138871]
65. Sánchez-Mendiola M, Anakin M, Findyartini A, Levine R, Da Silva A, Vakani FS. Applied insights for using generative artificial intelligence in faculty development in health professions education. MedEdPublish 2025;15:279. [doi: 10.12688/mep.21403.1] [Medline: 41552409]
66. Yavuz F, Çelik Ö, Yavaş Çelik G. Utilizing large language models for EFL essay grading: an examination of reliability and validity in rubric‐based assessments. Br J Educ Technol 2025 Jan;56(1):150-166. [doi: 10.1111/bjet.13494]
67. Suchikova Y, Tsybuliak N, Teixeira da Silva JA, Nazarovets S. GAIDeT (Generative AI Delegation Taxonomy): a taxonomy for humans to delegate tasks to generative artificial intelligence in scientific research and publishing. Account Res 2026 Apr;33(3):2544331. [doi: 10.1080/08989621.2025.2544331] [Medline: 40781729]
68. Kondo T. AI-generated versus human supervisor feedback on medical students' clinical clerkship logs: open data and analysis code. Zenodo. 2026. URL: https://doi.org/10.5281/zenodo.18868578 [accessed 2026-06-12]

Abbreviations

AI: artificial intelligence
API: application programming interface
JDA: joint display analysis
LLM: large language model
NU: Nagoya University
OR: odds ratio

XSL•FO
RenderX

Edited by S Brini; submitted 24.Dec.2025; peer-reviewed by A Yamamoto, F Fukuzawa, H Yin, R Yin, Y Asada; revised version received 08.May.2026; accepted 11.May.2026; published 16.Jun.2026.

Please cite as:
Kondo T, Donkers J, Nishigori H, Rovers S, Heeneman S
AI-Generated Versus Human Supervisor Feedback on Medical Students' Clinical Clerkship Logs: Cross-Sectional Convergent Mixed Methods Study
JMIR Med Educ 2026;12:e90064
URL: https://mededu.jmir.org/2026/1/e90064
doi:10.2196/90064

© Takeshi Kondo, Jeroen Donkers, Hiroshi Nishigori, Sanne Rovers, Sylvia Heeneman. Originally published in JMIR Medical Education (https://mededu.jmir.org), 16.Jun.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Education, is properly cited. The complete bibliographic information, a link to the original publication on https://mededu.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Impact of a Practical, Hands-On, Continuing Professional Development Course About AI in Health Care Professions Education on the Perceptions and Behaviors of Health Care Educators: Qualitative Case Study

Justin G Peacock[1], MHPE, MD, PhD; Jerusalem Merkebu[2], PhD; Anita Samuel[2], PhD

[1]Department of Radiology and Bioengineering, School of Medicine, Uniformed Services University of the Health Sciences, 4301 Jones Bridge Road, Bethesda, MD, United States

[2]Department of Health Professions Education, School of Medicine, Uniformed Services University of the Health Sciences, Bethesda, MD, United States

Corresponding Author:
Justin G Peacock, MHPE, MD, PhD
Department of Radiology and Bioengineering, School of Medicine, Uniformed Services University of the Health Sciences, 4301 Jones Bridge Road, Bethesda, MD, United States

Abstract

Background: Artificial intelligence (AI) is increasingly being used in many aspects of society, including health care and education. AI has the potential to enhance health care delivery, education, and administration. Health care trainees will be required to master these AI technologies. To teach trainees to effectively and ethically leverage AI technologies, educators must be appropriately trained and empowered to use these technologies.

Objective: We developed a health professions education course to enable health care professionals to overcome their fears and concerns about integrating multimodal generative AI technologies in daily practice. The course was also designed to foster generative AI skills and confidence in educational, administrative, research, and clinical activities.

Methods: Using a multimethod approach, we analyzed data gathered from three different sources using the 6-phase reflexive thematic analysis by Braun and Clarke. This involved familiarization with the data sources, generating initial codes, developing, refining, and defining the themes, and finally, writing up the results.

Results: Our findings indicate that 16 of 21 (76%) learners initially described apprehension or fear toward AI technologies. After experiential engagement with AI technologies together with their peers, they were able to shift their perspectives and gain confidence to integrate AI tools in their daily practice.

Conclusions: A brief 6-week continuing professional development course on the use of AI technologies for health care professional educators, focused on experiential and peer-based learning, resulted in shifts in perception, affect, and behavior toward AI technologies. It also propelled learners to shift increasingly outward in their discussion, application, and preliminary advocacy for AI technologies in their daily practice.

(JMIR Med Educ 2026;12:e87381) doi:10.2196/87381

KEYWORDS

artificial intelligence; AI; technology education; connectivism learning theory; experiential learning theory; health professions education; HPE; AI education; continuing professional development; CPD

Introduction

Background

Artificial intelligence (AI) is transforming the way that health care is taught and practiced [1-9]. AI tools are increasingly being used for administrative, clinical, research, and educational purposes in health care. This broad adoption of AI technologies necessitates that health professionals learn about the basic principles of AI and the ethical application of AI in clinical and educational practice [1,2,10-15].

Health professions education (HPE) research on AI education is relatively new, with significant growth after 2018, particularly in the post-2022 ChatGPT (OpenAI) era [1,2]. Much of the educational research describing AI in HPE focuses on how different AI tools can be used in health care education for educators and learners, and a few studies exploring AI curricula development [1-8,16]. The described AI curricula in HPE generally focus on basic AI principles, including AI ethics, but few describe hands-on, practical, or peer-based approaches to incorporating AI technologies in HPE practice [5,6,16,17]. The reported courses or curricula often reported Likert score

XSL•FO
RenderX

summative impacts, with limited qualitative response assessments [5,6,17]. Only one study, published after submission of this manuscript, reported on qualitative reflections three months after the course [17]. All the evaluations were level 1 to 2 on the Kirkpatrick Framework scale [18], impacting individual knowledge and skill, but only one reported limited data on post-course attitude and behavior changes (Kirkpatrick level 3), and none reported on wider organizational changes or patient outcomes (Kirkpatrick level 4) [2,5,6,16,17].

Along with the lack of research on AI in HPE, there is also a dearth of literature on the best ways to teach AI concepts to HPE professionals, the qualitative impact of those techniques on the learner's experience, and how teaching methods may lead to higher-level Kirkpatrick curricular outcomes. Laupichler et al [4] note that "it is important to find out how different populations respond to different teaching opportunities and what teaching materials are appropriate for teaching AI content." There is a critical need to educate HPE professionals on the ethical and effective use of AI, so that they can, in turn, educate health care trainees [1,7,10,11]. Without this training, educators will not be able to answer questions, address concerns, or model ethical behavior regarding the use of AI technologies.

Before discussing the theoretical framework for this manuscript, it is important to first explain what is meant by AI. AI, in its broadest terms, is the ability for machines and computer algorithms to perform functions typically associated with humans, such as prediction, comprehension, writing, or art [19]. This can range from current, probabilistic AI tools (narrow AI capable of specific tasks) to theoretical, future, autonomous AI technologies (general or broad AI capable of autonomous human intelligence). The range and scope of AI technologies is rapidly becoming incomprehensible and frightening for many health care professionals, with many wondering whether they will be replaced by algorithms [20].

This wide-ranging, all-encompassing, human-like, and threatening nature of AI technologies necessitates a new way of teaching that is different from other forms of technology education. Learners need to understand the basics of how these tools are developed, where these tools fail, the social and ethical implications of AI tool implementation, and, importantly, how they view their personal and professional role in the context of an AI-enabled future [8,20,21].

Given the unique nature of AI technologies, we propose that experiential learning theory principles may be more effective in teaching HPE learners about AI concepts [22-24]. Experiential learning theory, as proposed by David Kolb, characterizes learning as a holistic process in which knowledge is continuously created through the transformation of experience. This framework posits that learners do not merely absorb information, but actively construct understanding by progressing through a four-stage cycle. Applying experiential learning theory to HPE AI education, the four steps of concrete experience, reflective observation, abstract conceptualization, and active experimentation align well with effective AI learning methods [22].

As HPE learners work with the AI tools to develop a product (eg, quiz questions, images, and educational presentations), reflect on what they observed (eg, failure modes, benefits, and ease of use), think about their observations (eg, image tools struggle with text and more precise prompting leads to better outcomes), and then experiment on those ideas (eg, revise the prompt, incorporate meta-prompting to improve outcomes, and develop custom AI tools), then they begin a cyclic or spiral learning pattern, described by Kolb, that may lead to increased confidence and competence [22]. By recursively navigating this spiral, learners move toward higher levels of cognitive complexity, fostering the increased confidence and perceived competence necessary for professional mastery. We used an experiential learning theory lens to design the AI course and guide our primary research question.

Research Gap, Aim, and Question

In summary, the disruptive nature of AI technologies in society has caused many HPE professionals to feel anxious and fearful about the unique, existential threat posed by AI technologies [20,25]. The HPE literature is limited in describing how to effectively teach HPE professionals about AI concepts in a way that leads to affective, behavioral, and organizational changes in their personal practice (eg, higher order Kirkpatrick levels). Additionally, there is significantly limited qualitative data about the impact of continuing professional development (CPD) AI curricula on personal and professional changes that may occur.

In this qualitative case study on a CPD course for HPE professionals on AI technologies, we aim to understand how elements of experiential learning theory applied to the teaching of AI concepts may lead learners to shifts in their perception, affect, and behaviors about AI technologies.

We will explore the following research question: how does a short, experiential CPD course on AI tools in HPE influence learners' perceptions, affects, and behaviors toward AI technologies in their daily practice?

Methods

Ethical Considerations

This qualitative case study was deemed exempt by the Uniformed Services University (USU) institutional review board (IRB; #24 - 19215).

Deidentified artifacts from the course, including the precourse survey instrument and course assignments, were considered exempt educational artifacts.

For the semistructured interviews, the interviewer obtained informed consent from participants before the interview and maintained participant privacy as described in the interview methodology section. Participants received no compensation.

Study Design

This qualitative case study was conducted in the HPE program at USU. The HPE program offers a 6-week blended online course (HPE 912) that is offered primarily asynchronously, titled "AI in Health Professions Education: Implications for Educators and Leaders." Study participants were students enrolled during the spring and summer semesters of 2024. The

XSL•FO
RenderX

course objectives, schedule, and five written assignments are attached (Multimedia Appendix 1).

We adopted the qualitative collective case study approach to explore the collective phenomenon of multiple HPE learners participating in a CPD course on AI technologies in HPE through a qualitative interpretative lens on multiple varied artifacts from the course and after the course [26]. This in-depth approach allowed us to understand how and why changes in perceptions, affects, and behaviors occurred due to learning about AI technologies and participating in course activities. Limitations of this approach include generalizability beyond our institution and in other educational formats, as well as smaller sample sizes.

The authors used the SRQR (Standards for Reporting Qualitative Research) checklist in reporting the development and documentation of this study (Checklist 1) [27].

Course Design and Structure

HPE 912 was designed to equip health care educators and leaders with the knowledge, skills, and perspectives needed to integrate AI into HPE. The course was developed as a practical CPD experience that moved learners beyond conceptual awareness toward applied use of AI in authentic educational settings.

The course followed a progressive experiential arc (Multimedia Appendix 1). In the early weeks, learners explored foundational AI concepts and compared the performance of different large language models across educational tasks. In the middle portion of the course, they engaged with multimodal and generative artificial intelligence (GenAI) tools to create and revise educational products, including instructional materials and medical imagery. In the final weeks, learners examined leadership, implementation, and ethics questions related to AI adoption in their own professional contexts.

The educational design emphasized active experimentation, critical reflection, and contextual application. Each assignment required learners not only to use AI tools but also to reflect on their prompts, evaluate tool limitations such as hallucinations and bias, and justify the role of human judgment in refining outputs. In this way, the course encouraged participants to develop practical competence with common AI tools while also examining how their perceptions of the technology evolved and how AI might influence their future work as educators and leaders.

Data Collection

Overview

To understand the educational strategies that were most effective in fostering long-term change in the learners and the influence of the course on the students' practice, we employed methodological rigor by gathering data from three different sources: (1) precourse survey, (2) five written course assignments, and (3) semistructured interviews.

Precourse Survey

A precourse survey was administered to capture participants' baseline perspectives on AI and their expectations for the course. The instrument (Multimedia Appendix 2) included two closed-choice items on participants' emotional responses toward AI (eg, excitement, enthusiasm, and apprehension) and their prior knowledge and experience with AI tools. These included open-ended response items for participants to share their motivations for enrolling in the course. Participants were also asked to indicate any concerns they had about the course, the broader topic of AI, and the specific areas of AI in HPE they were most interested in exploring. This survey served to contextualize learners' starting points and to inform course design.

This closed survey was administered online via Google Forms (Google LLC, 2024) through the course learning management system, only to the learners in the course. This was a mandatory survey that needed to be completed as part of the course requirements [28].

Course Assignments

Student assignments were included as a key data source to examine how participants engaged with and applied course content in authentic tasks. The assignments required learners to analyze AI models, experiment with GenAI tools, design educational resources, and reflect on leadership and implementation strategies in their own contexts (Multimedia Appendix 1). As these tasks moved beyond knowledge recall to application, synthesis, and reflection, they provided a rich window into how participants made sense of AI concepts and integrated them into their professional roles.

Analyzing these products allowed us to assess not only immediate understanding but also evidence of higher-order thinking, creativity, and transfer of learning. Moreover, the assignments, providing a naturalistic record of learners' decision-making, challenges, and evolving perspectives, were an essential complement to survey and interview data for evaluating the course's impact on long-term professional practice.

Semistructured Interviews

We conducted 45 - 60-minute, semistructured interviews six months after course completion to assess the transfer of learning beyond the immediate course period. This delayed timing allowed us to explore how learners had sustained, adapted, or applied their learning in real-world professional contexts, including educational, administrative, research, and clinical practice (eg, self-reported Kirkpatrick level 3 behavioral changes and expanding connectivism learning networks). The ten semistructured interview questions focused on the research aim and question, including post-course application of learning (Multimedia Appendix 3).

We sent emails to all first cohort participants (nine learners), using IRB-approved recruitment emails that briefly described the interview purpose and research study. Only learners from the first cohort were invited for the interviews as they met the required longitudinal distance of six months since completion of the course.

Each learner voluntarily attended the interview in response to recruitment emails. They were also informed at the beginning

XSL•FO
RenderX

of the interview of the option to end the session at any point. At the start of every interview, JGP reviewed the consent process with the participant and clarified the (1) purpose and nature of the interview; (2) voluntary nature of the interview and option to opt-out; (3) maintenance of confidentiality through deidentification and secure storage of interview content; (4) recording of the interviews for auto-transcription and recall purposes, but that it could be deleted if desired; (5) risks associated with the interview; and (6) contact information for the researchers should concerns arise.

Privacy associated with the interviews was achieved through virtual interviews conducted by JGP in a quiet, private office. Interviews were conducted using Google Meet (Google LLC, 2024) with the transcription feature enabled. JGP reviewed the transcripts and recordings for accuracy and clarifications or corrections.

Data Analysis

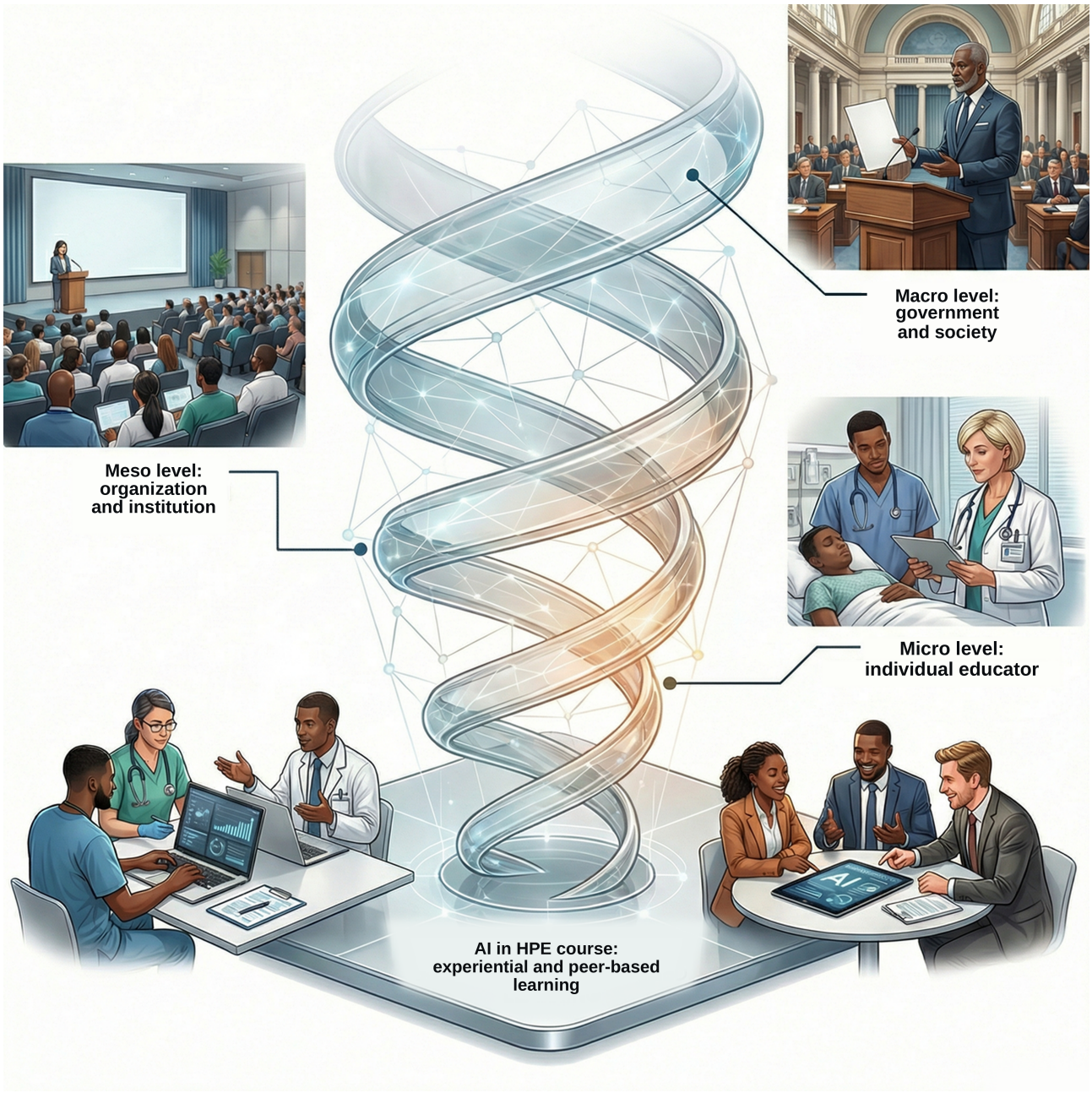
Data Procedures and Technological Support

All data sources, survey responses, assignments, and interview transcripts were deidentified by a research assistant and assigned a pseudonym. The data were secured as Google Docs (Google LLC, 2024) files on the USU Google secure server. Members of the team accessed the files on encrypted password-protected computers. This deidentified data was analyzed by the researchers.

The researchers used ATLAS.ti (ATLAS.ti Scientific Software Development GmbH; 2024) to track codes and quotations. The authors acknowledge limited use of multimodal GenAI tools for two discrete tasks, as outlined in Multimedia Appendix 4. The researchers used OpenAI ChatGPT-4o (paid version) to cocreate the subthemes from the researcher-identified codes. JGP used Google NotebookLM (Google LLC and DALL-E-3, Open AI) to cocreate Figure 1, using the manuscript as the source data with author-provided meta-prompting.

XSL•FO
RenderX

Figure 1. Participants in an AI-focused HPE course, grounded in experiential and peer-based learning principles (base of the helical spiral), demonstrated cyclic growth and ever-widening networks of interconnectivity (widening helix with interconnected nodes). Many learners exhibited shifts in perception, affect, and behavior at the individual (micro) level, with some demonstrating initiatives at the organizational and institutional (meso) levels, and at least one learner approaching societal and governmental (macro) levels (level-based impacts are depicted as call-out scenes). Google Gemini (Google LLC), NotebookLM (Pro version, accessed February 19, 2026), and ChatGPT DALL-E-3, OpenAI; June 10, 2026 were used to cocreate this image as detailed in Multimedia Appendix 4. AI: artificial intelligence; HPE: health professions education.



Precourse Survey

The researchers aggregated the survey data into a Google Sheet (Google LLC, 2024) for analysis. The researchers conducted a simple descriptive analysis of the data for the two close-response questions and the open-ended question: How and what AI technologies have you used? The researchers treated the responses to the other three open-ended questions as textual data and thematically analyzed them using the thematic analysis process by Braun and Clarke.

Course Assignments and Interviews

The researchers treated the course assignments and interview transcripts as textual data and thematically analyzed them using the 6-step thematic analysis process by Braun and Clarke. In step 1 of the analysis, authors JGP and JM reviewed the data for familiarization and identified initial patterns within the data. In step 2, JGP and JM generated the initial codes (Table 1). JGP and JM then further refined the codes. JGP, JM, and AS generated, reviewed, and refined the themes in steps 3, 4, and 5 (Table 1). The research team addressed all discrepancies through discussion until they reached consensus. Finally, we selected compelling extracts that were rechecked against the themes and research question. Table 1 details our movement

XSL•FO
RenderX

from developing codes, which the authors combined into subthemes, and then themes. This approach was especially effective because it allowed us to follow the thread and gain a deeper understanding of the relationships between these elements [29].

Table . Themes, subthemes, and codes.

Themes	Subthemes	Codes
AI[a] misconceptions breed fear	Navigating initial encounters with AI	• Embarrassed • Initial misconception about AI use • Initial response to AI • Intrigued but resistant • Precourse AI misconceptions • Precourse perceptions of novice status
Education leads to a shift in perspective	Optimizing the learning experience shifting perspectives and addressing concerns	• Changes in learning/education • Creativity • Free experimentation • Hands-on practice to learn AI • Learning through play • Repetition produces familiarity • Teacher-learner interaction • AI fears/concerns • Hope in future • Postcourse changed perspective • Postcourse understanding of ethics • Professional identity • Trust/confidence • Freedom/flexibility • Various modalities enhanced learning • Modeling
As perspectives shift, learners reach out to their social circles	Developing competence and inspiring action	• Need for further education • Postcourse AI initiatives • Postcourse motivation • Postcourse perception of abilities • Perceived benefits of AI • Postcourse AI limitations
Confidence in AI technologies leads to implementation and advocacy	Fostering knowledge dissemination and advocacy	• Demonstrations of AI advocacy • Sharing knowledge with peers/friends/family • Institutional barriers • Applicability to real world

[a]AI: artificial intelligence.

Data Analysis Summary

Unlike the survey and assignments, which captured immediate reactions and behavioral applications within course boundaries, the interviews provided insight into longer-term engagement, barriers to implementation, and the integration of AI tools into practice. By focusing on learners' reflections after an extended interval, these interviews offered a critical perspective on the durability and impact of CPD in fostering meaningful change. Together, these three data sources provided a comprehensive view of the learners' journey from entry into the course through to longer-term professional application. This multimethod approach allowed for a richer understanding of how CPD in AI influenced Kirkpatrick Level 1 - 3 learning in written and oral samples.

To establish the rigor of our findings, we intentionally transitioned away from the positivist concept of data saturation, which is frequently incompatible with interpretivist research. Instead, guided by recent frameworks for assessing qualitative data adequacy [30] and the conceptualization of information power by LaDonna et al [31], the researchers evaluated data adequacy based on the specific characteristics of this study's design and the resulting artifacts. Sufficiency was assessed iteratively during the analysis. As this study possessed a narrowly focused research question, a highly specific participant sample (educators who completed this exact experiential course), and interview dialogue characterized by rich conceptual depth, the dataset was deemed to have high information power [31]. Through this lens, the research team determined that 105 written assignments from 21 participants and 5 in-depth interviews achieved the theoretical sufficiency needed to address the exploratory research aims. In alignment with this framework, we prioritized the richness, resonance, and conceptual depth of the data over absolute sample size.

By using the experiential learning theory by Kolb as an a priori theoretical lens, we sharpened our analytical focus and enhanced the evidentiary value of this study. This iterative analytical

XSL•FO
RenderX

process allowed us to move beyond surface-level descriptions, reaching a point of analytical sufficiency where the findings offer a robust, conceptually thoughtful account of how hands-on experimentation and reflective practice dismantle precourse misconceptions, ultimately fostering the confidence and trust required for institutional AI implementation.

Researcher Characteristics and Reflexivity Statements

JGP is a nuclear radiologist passionate about AI technologies for educational and clinical practice. He believes AI tools can help educators be more effective and efficient in their teaching efforts, and that these tools, when used ethically and appropriately, can increase creativity and critical thought about complex problems. His views are reflected in the scope, development, and reporting of this study. He helped in the initial development of the first HPE course, but he did not act as an official instructor or grader for either course to reduce research bias and prevent undue influence during the interviews. He performed the interviews described below and participated as one of the coders during the analysis. To reduce bias, he did not review the presurvey or course assignments until independent coding analysis began. Through open-ended, semistructured interview questions, he attempted to mitigate personal influence on the interviews.

JM is an educational psychologist who uses mixed-methods research to explore metacognition and emotion in medical education. Her expertise fundamentally shaped this study's methodology, analysis, and interpretation to prioritize participants' metacognitive reflections and emotional responses to AI technologies. She participated as an independent coder during the analysis. She was not involved in the course in any way and only participated in the research phases of this study.

AS is an online educator and researcher. Her research straddles the domains of adult learning, online learning, and health education. We recognize our views have a critical implication on the ethical and effective integration of AI in educational settings. She developed the HPE course and was the primary instructor and grader for both course cohorts. To reduce bias, she did not participate in data collection or initial coding. She participated in a subsequent thematic analysis to avoid influencing the findings of this study.

Results

Participant Demographics

The researchers included survey data and assignments from 21 learners in two course cohorts for this study. Study participants ranged in age from 26 to 54 years with an average age of 42 years. This study included 10 male and 11 female participants. Only 14 (67%) of the participants were physicians, with the others spanning specialties from dentistry, nursing, pharmacy, laboratory sciences, and physical therapy. The cohorts included 2 medical student educators. The researchers collected written artifacts from 21 learners in two cohorts of the course, including spring 2024 and 2025 offerings.

A total of 5 of the 8 (62.5%) learners in the first cohort (1, 2, 4, 6, and 8) responded to the IRB-approved survey recruitment email and volunteered to participate in the interview. We noted that the nonresponding learners consisted of 2 surgeons and 1 resident physician. The 5 respondents consisted of 2 family medicine physicians (1 in administration/clinical practice), 1 pediatric physician, 1 surgeon (administration/clinical practice), and 1 laboratory supervisor.

Study Findings

In this study, we explored how a CPD course for HPE professionals on AI technologies could influence the perceptions, effects, and behaviors about AI technologies. Our findings revealed that AI misconceptions breed fear toward AI technologies and that experiential and peer-based education can lead to a shift in perspective. In turn, this leads HPE learners to influence their personal social circles. For some learners, confidence in the use of AI technologies led to increased self-reported implementation and advocacy for AI in their organizations. Table 1 summarizes the aforementioned codes, subthemes, and themes identified by the researchers through inductive analysis of the learners' artifacts, including presurvey, written assignments, and semistructured interviews [32]. We report selected quotes, aligning with these themes, in the text below. Supplemental quotes supporting these themes can be found in a supplemental table (Multimedia Appendix 5).

Precourse Survey

Apprehensive, but Curious, Learners Engaged in the Course

Overall, 15 of 21 (71%) respondents for the precourse survey noted feeling apprehension about the use of AI, and 14% (n=3) mentioned fear. These feelings support the identified codes aligning with the "AI misconceptions breed fear" theme (Table 1). A total of 52% (n=11) and 67% (n=14) of respondents identified feeling excitement and enthusiasm for the use of AI, respectively. Only 10% (n=2) of participants reported having had no exposure to AI before the course. Few learners (n=1) in the first cohort reported having used AI in some capacity, compared with the second cohort (n=9), resulting in a total of 10 (48%) participants for both cohorts. An additional 39% (n=9) of learners in both cohorts reported basic knowledge about AI. The course was an elective; therefore, participants were interested in AI and curious about it when they enrolled in the course.

In their precourse questionnaire, 2 interviewees reported no prior knowledge of AI, 2 interviewees reported basic limited AI knowledge, and 1 interviewee reported some experience with using AI tools or applications. All demonstrated some enthusiasm or excitement about AI, but 3 interviewees also reported apprehension or uncertainty (Table 1). All reported taking the course to learn more about AI technologies and how they could be used for health care and education, with P6 describing his motivation as "options are either to hide my head in the sand and become obsolete, or develop proficiency and continue to grow along with the practice of medicine."

Reflective Assignments (A) and Interviews (I)

AI Misconceptions Breed Fear

Universally, learners remarked on the misconceptions that they had about AI before the course (Table 1). P19 (A) expressed it

XSL•FO
RenderX

best as, "if I had been asked frankly what I perceived AI to be, my answer was often drawn to 30-year-old movies like Terminator or Maximum Overdrive." Misconceptions included the idea that using AI in education and practice was unethical, that AI tools were infallible, that AI would replace them as physicians or educators, that AI would not be useful in academic or clinical practice, and that they were not currently using AI in their daily lives. As P8 (I) stated, "I was so afraid of the ethical dilemmas like 'Oh my God, if these words aren't mine, is that okay?'" These misconceptions about AI lead learners to fear AI and avoid its use in their personal and professional lives.

Learners also expressed fears that "the potential risks…of AI in medical education are limitless. The dangers of plagiarism, hallucinations, and bias are worrisome" (P10 (A)). As expected of health professions educators, many expressed fears about learning and education, noting that "trainees would use it to complete assignments, personal statements, etc, essentially making these 'useless' in evaluation" (P1 (A)). The learners described AI fears about maintaining work originality, inaccuracies, or "hallucinations," "deep fakes," or other nefarious uses of AI technologies, patient safety, bias, inability to keep up with the pace of AI technologies, changes to their professional practice or losing their jobs, an inability to learn something new or technologically challenging, and a fear of the unknown.

Education Leads to a Shift in Perspective

Education, however, was able to create a shift in the learners' perspectives (Table 1). Many learners described that their initial fears and misconceptions were reduced or eliminated due to attending the course: "I'm like, 'God, I have to engage' and I'm like, 'How do I do this?' And so, it really helped me sort of start my journey" (P2 (I)). As learners engaged experientially with the content in the course, they began to recognize the misconceptions they had previously held about AI technologies. One of the first realizations for many learners was that AI tools are integrated into almost every aspect of modern life. P1 (A) clearly articulated "I erroneously thought that I had almost no interaction with AI…I was surprised to learn that I interacted with AI on a daily basis…"

Many remarked how they would not have imagined that they would have used AI technologies before the course as an expert health care educator, but as they worked hands-on with the tools that had quickly changed. For instance, as P13 (A) reported, as "I reflect on my journey through this course on artificial intelligence (AI) in health professions education, I recognize the significant evolution in both my understanding of AI and my enthusiasm for integrating it into my professional endeavors." They also realized the imperfections, hallucinations, and errors that the tools produced. For example, participating firsthand in the image creation activities, they realized that AI tools could not draw a perfect hand or an image with text on it. As participant P7 (A) stated, "one interesting thing is that the AI has a lot of trouble with hands…"

Participants actively discussed ethical concerns with faculty, guest lecturers, and peers, and began to describe a shift in understanding about the way they perceived ethical uses of AI in their daily lives, academic work, and clinical practice. It shifted from a concrete view of ethics to a more nuanced perspective on where, when, and how AI tools could be used. Furthermore, many described how their concerns of naivete about AI technologies caused them to be hesitant to learn or use the tools.

They described how the class provided opportunities to remove self-imposed limitations to their use, bravery to use the tools, and the realization that they were in a similar position of novice status relative to their peers and other health care professionals. Learner P6 (I) stated that AI technologies "don't require other people, they just require me to be brave enough to create an account and yeah, screw up…" Importantly, many learners described a change in their affect and emotional response to AI technologies. The AI fears described before the class gave way to emotions of excitement, enjoyment, and hope about AI technologies and their potential to improve their lives and careers as health care professionals and educators. For instance, P14 (A) noted, "I think the biggest surprise was how much this course served as a paradigm shift for me. How does an entire healthcare system or medical education program shift from what amounts to written rules to algorithms and generative ways of proceeding?" As learners' perspectives and emotions around AI shifted, they described how they would reach out to their social circles to share, teach, and use their newly developed knowledge about AI.

Influencing Personal and Social Circles as Perspectives Shift

This excitement in the AI tools shifted to a natural tendency to share that knowledge with their social circles in peer-based approaches (Table 1). P19 (A) described this excitement about sharing with their social circles as follows, "I met each assignment with such excitement and shared my new knowledge with my family and my colleagues at work…I became a weekly feature on our team huddles, sharing my latest knowledge and providing information on how it could be sourced in an organization that was not currently set up for the use of AI." For many, the closest social circle was their own family members. Learners described teaching their children about the ethical and effective use of AI tools for their schoolwork. They shared how they were able to appropriately modify their response to their children's use of AI through a more nuanced understanding of the ethical uses of AI. For some individuals, their family members were more knowledgeable about AI technologies, and so they would reach out to learn more from them.

Additionally, learners discussed how they would learn from their peers in the class. They expressed how it was good to brainstorm ideas, discuss challenges, and develop new uses for the AI tools discussed in class. P7 (A) stated that "I certainly was impacted by peer learning and just seeing and reading about some of what my peers were doing with AI." Learners also described peer-based learning with their trainees. They shared how their own health care trainees would teach them about new AI technologies, and how they would instruct them on effective and ethical AI tool use. In describing this experience, P18 (A) noted, "I believe student collaboration in particular will help me stay abreast of AI development since the younger learners

XSL•FO
RenderX

are extremely forward leaning in that direction." As a component of this social circle-based learning, some learners described the use of social media in helping to develop more knowledge about AI tools. The socialized learning of AI technologies among the learners, their families, and their peers led to a natural progression of technology implementation in their professional careers, development of programmatic and institutional education, and advocacy for AI technologies in their spheres of influence.

Confidence in AI Technologies Leads to Implementation and Advocacy

As learners learned about, experimented with, and used the AI tools in class, they began to implement the AI tools in their daily lives (Table 1). Many reported initially using the tools for simple administrative tasks, such as writing emails and planning vacations. As educators, they also turned to the implementation of AI tools in basic educational activities, such as building curricula, summarizing clinical information, revising curriculum vitae, developing clinical templates, writing questions, generating learning objectives, developing patient scenarios for education, and developing a second opinion. As P1 (A) explained, "The utility of these within my role is almost endless… from assisting me with drafting initial educational components more efficiently to generating images for educational presentations to more effectively drafting communications I send to staff (ie, emails, newsletters, etc)."

Many learners described how the AI tools helped to reduce the cognitive load and time spent developing and crafting a project or product. They described how AI tools helped reduce the blank page anxiety that comes with starting a new project. As P6 (I) stated, "It was really useful for that, overcoming, what, Picasso described as the tyranny of the blank canvas. Going from 'How do I even get started on this' to say 'Okay, I've got a block architecture. I can modify that as needed.'" In reducing the activation barrier for learners, they were able to begin new endeavors more confidently. They noticed how the AI tools made their jobs easier and time-efficient.

P8 (I) described a particularly impactful experience with not knowing what to say or do for a learner who was struggling with mental health concerns. The learner used AI tools to help craft an appropriate message, "we had a resident who was having a really rough time with mental health and I wanted to send them a message and I wasn't sure of the right words to say in this situation…it was super helpful just helping me come up with a nice message to send."

Some learners began to develop more complex applications of the AI tools that impacted more than just themselves and their small social circles. They worked with others to develop educational courses, curricula, and tools for clerkship rotations. P21 (A) reported one such presentation: "I recently developed and delivered a grand rounds on the topic of AI. The focus of this presentation was on the evolution of AI, evolution of medicine, and current AI being used within [my organization]." With increasing understanding, confidence, and real-world application of the AI tools, learners felt more empowered to begin advocating for AI technologies in their programs and institutions. Many acknowledged that there were challenges with implementing AI tools in their institutions and in health care practice, but they still sought ways to promote knowledge of AI technologies, the ethics of AI, and the application of AI technologies in daily practice.

Despite institutional challenges, P1 (A and I) described developing and organizing an entire day devoted to discussing the application of AI technologies in education and clinical practice. She worked with AI experts at her institution and educational leaders to build support and invited speakers to come speak about their work with AI technologies. P8 (I) described how she worked with fellow medical student site directors to develop a specific AI tool that would provide improved, targeted, narrative feedback for medical students working on specific entrustable professional activities. These were then distributed to medical student sites nationwide. P6 worked with AI experts at his institution to develop workshops centered on learning about AI technologies and how they could be used in clinical practice.

A summary of these educational advocacy approaches was best conceptualized as follows:

> *By integrating AI into my teaching and research, I can help my students and colleagues better understand its potential benefits and encourage a more open-minded approach to its use…This has been significant because it has helped me recognize that I can not only use these tools but also advocate for their use in educational settings. As an educator, this is empowering—it allows me to introduce AI to my learners and help them see how they can use these tools to enhance their own careers.* [P13 (A)]

The AI course empowered some educators to learn, explore, and experience AI tools, which they shared socially, building communities that impacted their programs and institutions. As learner P12 (A) indicated, "we all must become AI leaders…"

Discussion

Summary of Findings

In our study, we found that experiential, practical, and peer-based learning methods were effective in helping HPE learners learn how to effectively use AI tools in their daily activities. As learners began using AI tools in their everyday lives, they reported a decrease in previously held misconceptions and fears about AI technologies (Table 1). They began to explore other areas for AI tool integration and application in their personal and professional lives. Learners learned about new AI tools from their family members, peers, and students. In turn, they also reported teaching others about what they learned from the course in increasingly wider spheres of influence. Learners reported their use of AI tools from the class in educational, administrative, and research settings. Ultimately, the course influenced learners in micro, meso, and macro contexts within their personal and professional lives (Figure 1) [10,10,33].

Precourse Fears and Misconceptions

Many HPE learners in our CPD course had misconceptions about AI technologies that often led to apprehension, fear, and

XSL•FO
RenderX

a tendency to avoid its use. These beliefs included the idea that integrating AI into education and practice is inherently unethical, that AI tools are without flaws, and that AI will inevitably replace human professionals. What does it mean to be a health care professional working with a tool that is smarter than you? What does it mean for professionals using these tools in clinical practice? Furthermore, learners frequently underestimate how much they already interact with AI in their daily lives and its potential value in both academic and clinical settings. Prior research confirms that this general unease is a common psychological barrier to engaging with new technologies [25,34].

Beyond these misconceptions, specific emotional anxieties were prevalent. Health care professionals, for instance, voiced concerns about AI's impact on learning and evaluation, fearing that trainees might use AI to complete assignments, thereby diminishing the integrity of traditional assessment methods. Ethical considerations surrounding AI, such as the authenticity of AI-generated content, also caused considerable worry. Other fears encompassed the risks of plagiarism, AI-generated inaccuracies, and bias, alongside worries about patient safety, the rapid advancement of technology, potential job displacement, and feelings of personal inadequacy in learning new, technologically complex skills. Hosseini et al [35] and others have also echoed that these fears present challenges for health care professionals in adopting AI technologies. Collectively, these fears suggest an apprehension may stem from a lack of understanding and direct experience with AI tools in their daily practice.

Postcourse Shift in Perspective and Engagement

Post-course interviews with 5 HPE learners from the first cohort demonstrated a consistent shift in perspective through direct engagement with AI tools, manifesting across micro, meso, and macro levels, achieving self-reported Kirkpatrick level 3 outcomes [18,33]. This evolution of ever-widening spheres of peer influence and knowledge acquisition aligns with connectivism, a learning theory that views knowledge as a process of navigating nodal networks composed of both human and technological sources [23,24]. In the context of rapidly advancing AI, these networks require HPE learners to engage in a continuous cycle of integrating new nodes—such as emerging platforms, technical expertise, and diverse peer groups—which fosters a critical metacognitive ability to evaluate and prioritize information. By facilitating experimentation through discussions and peer-based learning, faculty enabled participants to build and maintain these expanding networks, effectively shifting their educational perspective from isolated skill acquisition to a systemic, networked understanding of AI integration within their professional environments [33].

Micro Level

Crucially, our case study demonstrates that direct engagement with AI content through structured, hands-on, experiential education can significantly alleviate or eliminate personal fears and misconceptions [36]. In line with experiential learning theory, we found that the ability for learners to experiment with the AI tools, practice their AI skills, and even fail in their desired outcomes (in a safe environment) may help learners to more rapidly shift their perspectives regarding AI technologies [22]. This shift in perspective often began with the self-reported, "surprising" realization that AI is far more integrated into daily life than previously understood. Learners initially believed they had minimal interaction with AI, only to discover its pervasive presence in their routines. This newfound awareness was an important initial step in demystifying AI technologies and fostering a more receptive mindset.

Moreover, direct interaction with AI tools enabled learners to recognize their imperfections and limitations, such as AI's occasional difficulty in accurately rendering hands or text in images. Merkebu and Samuel [25] also highlight that "using customizable and psychologically safe strategies, effectively dismantles globally shared anxieties, such as the fear of inadequacy, making AI adoption feasible and empowering individuals across diverse contexts." This experience challenged the notion of AI infallibility and fostered a more realistic understanding of its capabilities. Discussions with faculty, guest lecturers, and peers on ethical considerations also facilitated a transition from a rigid, black-and-white view of AI ethics to a more nuanced and context-dependent understanding of its appropriate use. This evolving perspective highlighted that many professionals in our course are still navigating how and where AI can effectively fit within their practice.

The educational experience may also help learners overcome their cognitive and meta-cognitive limitations and embrace a willingness to experiment with AI tools. Recognizing that many of their peers shared a similar beginner status with AI fostered a supportive learning environment. This comprehensive engagement with AI content led to a notable change in learners' emotional responses, transforming initial feelings of fear and inadequacy into enthusiasm, enjoyment, and optimism about AI's potential to enhance their professional and personal lives as health care professionals and educators. Extending prior works, we demonstrated that well-designed educational interventions are essential in bridging the gap between apprehension and active engagement with rapidly evolving AI technologies [25].

Meso Level

In line with connectivism theory, we found that as many HPE learners in our course experimented with, practiced, and became more confident in their AI abilities, they began to share their knowledge and skills with others, developing new nodes in their developing learning networks [23,24]. Often, learners begin by sharing with family members and learning from those close individuals. As educators, many also began to share and learn from their peers and their students. This peer-based learning has been shown to be effective in developing clinical AI champions in stressed health care systems by Teferi et al [37]. Similarly, Zhou and Schofield [38] also argue that AI has the potential to promote collaborative and communal learning. Emotions of surprise, joy, and excitement often accompanied this desire to share and learn from others.

Extending prior research, the shift in perspectives and emotions led learners to share and learn from peers and family members about different aspects of AI technologies [39]. Parent learners often shared what they had learned about the ethical use of AI technologies with their children, which may not have been

XSL•FO
RenderX

possible without the course. Adult family members and peers often provided more knowledge and expertise about topics that were at the boundaries of the learner's expertise and experience [40]. In some cases, learners were able to share and teach about what they learned with their peers, leading to the development of interconnected nodes of learning and excitement about AI technologies, as described in connectivism learning theory. As students in our course developed these local and personal interconnected networks, they developed AI confidence and competence that allowed them to reach out to larger organizations and networks of influence [39].

Macro Level

Notably, in our post-course interviews with some learners, we discovered that some learners expanded their networks of AI learning and knowledge to their organizations and medical communities, describing actions that serve as preliminary steps toward advocacy for AI technologies in their practices. For some learners, these ideas of advocacy were aspirational, but for others, they made efforts to implement change in their larger spheres of influence. The capacity shown by some learners to navigate complex networks and structures to develop engaging content for their peers and trainees demonstrates excitement, motivation, and commitment developed during the course. These early advocacy actions and motivations demonstrate a preliminary macro-level shift in a few HPE learners beyond their close social spheres to expanding spheres of collaboration and influence, advocating for and instituting changes in their health care systems [21].

Limitations

Limitations of this study must be acknowledged. First, this study has been conducted with a small number of participants from one institution. Furthermore, the sample size of interview participants was also small. While this might raise questions of generalizability, it must be noted that qualitative studies aim to provide a deeper and more detailed understanding of people's experiences [41]. Next, the participants in this study all self-selected to participate and entered the course out of curiosity and interest. While this could be seen as skewing the current study, it is interesting to note that these participants also described fear and apprehension at the start of the course. Furthermore, while our participants discussed initiatives they had started in their organizations, this is limited to self-reported data, as we did not verify the information with the organizations. Future studies could focus on change at Kirkpatrick level 4 (organizational change) and macro levels, including objectively examining participants' actions.

Conclusions

AI technologies are rapidly disrupting and changing many industries and professions, including health care. To address this rapidly evolving environment, it is imperative to prepare health care educators with the requisite knowledge, skills, and attributes to ethically and effectively integrate, use, and teach about AI technologies. In this study, we demonstrate how practical, experiential, and peer-based learning fosters confident and perceived competent health care educators, resulting in self-reported level 3 Kirkpatrick outcomes [18]. Some educators even reported the beginnings of broad-based advocacy for AI technologies within their institutions and broader health care communities. Excited and empowered educators share, teach, and continue to learn about the AI technologies, establishing ever-expanding and interconnected networks of learning.

Acknowledgments

OpenAI's ChatGPT-4o was used to cocreate the subthemes from the author-identified codes. The authors reviewed these subthemes for accuracy and consistency. These are included in Table 1. The codes and themes were developed by the authors, as described in this paper. Google NotebookLM (Pro version, accessed February 19, 2026) was used to cocreate Figure 1 using this paper as the source data with that was author provided and Google Gemini cocreated (Pro version, accessed February 19, 2026).

Funding

The authors received no funding for this research. The opinions and assertions expressed herein are those of the authors and do not reflect the official policy or position of the Uniformed Services University of the Health Sciences, US departments of the Army, Navy, or Air Force, or the Department of War.

Data Availability

The datasets generated or analyzed during this study are not publicly available due to ongoing research objectives, but are available from the corresponding author on reasonable request.

Authors' Contributions

Conceptualization: JGP (lead), AS (supporting)
Data curation: JGP (lead), JM (supporting), AS (supporting)
Formal analysis: JGP (lead), JM (equal), AS (supporting)
Investigation: AS (lead), JGP (equal)
Methodology: JM (equal), AS (equal), JGP (supporting)
Project administration: JGP (lead), JM (equal), AS (supporting)

XSL•FO
RenderX

Supervision: JGP (lead), JM (equal), AS (supporting)
Validation: JGP, JM, AS (equal)
Visualization: JGP (lead), JM (supporting), AS (supporting)
Writing: JGP (lead), JM (equal), AS (supporting)

Conflicts of Interest

None declared.

Multimedia Appendix 1
Artificial intelligence in health professions education: continuing professional development course description, objectives, assignments, artificial intelligence usage policies, schedule, and grading rubrics.
[PDF File, 549 KB - mededu_v12i1e87381_app1.pdf]

Multimedia Appendix 2
Artificial intelligence in health professions education: continuing professional development course presurvey questions.
[PDF File, 188 KB - mededu_v12i1e87381_app2.pdf]

Multimedia Appendix 3
Artificial intelligence in health professions education: continuing professional development course, semistructured interview questions.
[PDF File, 44 KB - mededu_v12i1e87381_app3.pdf]

Multimedia Appendix 4
Description of the use of multimodal generative artificial intelligence tools in this paper.
[PDF File, 54 KB - mededu_v12i1e87381_app4.pdf]

Multimedia Appendix 5
Additional thematic quotes.
[PDF File, 84 KB - mededu_v12i1e87381_app5.pdf]

Checklist 1
The SRQR (Standards for Reporting Qualitative Research) checklist for qualitative research was used in the conception and reporting of this manuscript [27].
[DOCX File, 21 KB - mededu_v12i1e87381_app6.docx]

References

1. Gordon M, Daniel M, Ajiboye A, et al. A scoping review of artificial intelligence in medical education: BEME Guide No. 84. Med Teach 2024 Apr;46(4):446-470. [doi: 10.1080/0142159X.2024.2314198] [Medline: 38423127]
2. Charow R, Jeyakumar T, Younus S, et al. Artificial intelligence education programs for health care professionals: scoping review. JMIR Med Educ 2021 Dec 13;7(4):e31043. [doi: 10.2196/31043] [Medline: 34898458]
3. Peacock J, Austin A, Shapiro M, Battista A, Samuel A. Accelerating medical education with ChatGPT: an implementation guide. MedEdPublish 2023;13:64. [doi: 10.12688/mep.19732.1]
4. Laupichler MC, Aster A, Schirch J, Raupach T. Artificial intelligence literacy in higher and adult education: a scoping literature review. Comput Educ: Artif Intell 2022;3(100101):100101. [doi: 10.1016/j.caeai.2022.100101]
5. Khamis N, Chen B, Egan C, Gaglani S, Tackett S. More intelligent faculty development: integrating GenAI in curriculum development programs. Med Teach 2025 Nov 2;47(11):1739-1741. [doi: 10.1080/0142159X.2025.2473606]
6. Chadha N, Popil E, Gregory J, Armstrong-Davies L, Justin G. How do we teach generative artificial intelligence to medical educators? Pilot of a faculty development workshop using ChatGPT. Med Teach 2025 Jan 2;47(1):160-162. [doi: 10.1080/0142159X.2024.2341806]
7. Peacock JG, Cole R, Duncan J, Jensen B, Snively B, Samuel A. Transforming military healthcare education and training: AI integration for future readiness. Mil Med 2025 Sep 1;190(9-10):e1905-e1912. [doi: 10.1093/milmed/usaf169]
8. Hlede V, Valanci S, D'Antuono GR, Dow H, O'Beirne R, Wiggins R. AI-enhanced continuing professional development as an evolving sociotechnical system: multimethod theoretical framework development study. JMIR Med Educ 2026 Feb 12;12:e69156. [doi: 10.2196/69156] [Medline: 41678653]
9. Mosch L, Fürstenau D, Brandt J, et al. The medical profession transformed by artificial intelligence: qualitative study. Digit Health 2022;8:20552076221143903. [doi: 10.1177/20552076221143903] [Medline: 36532112]

XSL•FO
RenderX

10. Billings HA, Gaynier A, Hall ER, et al. AI competencies for medical educators. : CGEA Faculty Development SIG AI in Medical Education Workgroup; 2025 May URL: https://www.aamc.org/media/86881/download [accessed 2026-06-09]
11. Lomis K, Jeffries P, Palatta A, et al. Artificial intelligence for health professions educators. NAM Perspect 2021. [doi: 10.31478/202109a] [Medline: 34901780]
12. Masters K. Artificial intelligence in medical education. Med Teach 2019 Sep 2;41(9):976-980. [doi: 10.1080/0142159X.2019.1595557]
13. Masters K. Ethical use of artificial intelligence in health professions education: AMEE guide no. 158. Med Teach 2023 Jun 3;45(6):574-584. [doi: 10.1080/0142159X.2023.2186203]
14. McCoy LG, Nagaraj S, Morgado F, Harish V, Das S, Celi LA. What do medical students actually need to know about artificial intelligence? NPJ Digit Med 2020;3(1):86. [doi: 10.1038/s41746-020-0294-7] [Medline: 32577533]
15. Shankar PR. Artificial intelligence in health professions education. Arch Med Health Sci 2022;10(2):256-261. [doi: 10.4103/amhs.amhs_234_22]
16. Cho Y, Szarek JL. Integrating task-technology fit and community of practice theories to enhance medical educator skill confidence in generative AI: design, implementation, and pilot outcomes. MedSciEduc 2025 Dec 6;35(6):2705-2710. [doi: 10.1007/s40670-025-02520-7]
17. Aftab R, Afzal AS, Babar S, et al. Effect of building capacity of health professional educators in artificial intelligence through a series of workshops; a follow-up study. Adv Med Educ Pract 2026;17:580674. [doi: 10.2147/AMEP.S580674] [Medline: 41884861]
18. Kirkpatrick DL. How to start an objective evaluation of your training program. : ATD; 1956 URL: https://www.kirkpatrickpartners.com/wp-content/uploads/2024/04/how-to-start-an-objective-evaluation-of-your-training-program-1956-courtesy-of-atd.pdf [accessed 2026-06-09]
19. Stryker C, Kavlakoglu E. What is artificial intelligence (AI). IBM. URL: https://www.ibm.com/think/topics/artificial-intelligence [accessed 2026-06-09]
20. Arvai N, Katonai G, Mesko B. Health care professionals' concerns about medical AI and psychological barriers and strategies for successful implementation: scoping review. J Med Internet Res 2025 Apr 23;27:e66986. [doi: 10.2196/66986] [Medline: 40267462]
21. Wiljer D, Hakim Z. Developing an artificial intelligence–enabled health care practice: rewiring health care professions for better care. J Med Imaging Radiat Sci 2019 Dec;50(4):S8-S14. [doi: 10.1016/j.jmir.2019.09.010]
22. Kolb AY, Kolb D. Eight important things to know about the experiential learning cycle. Aust Educl Leader 2018 Sep 1;40(3):8-14 [FREE Full text]
23. Siemens G. Connectivism: a learning theory for the digital age. Int J Instructional Technol Distance Learn 2008 Jun 6:1-9 [FREE Full text]
24. Goldie JGS. Connectivism: a knowledge learning theory for the digital age? Med Teach 2016 Oct 2;38(10):1064-1069. [doi: 10.3109/0142159X.2016.1173661]
25. Merkebu J, Samuel A. Humanizing AI training for health professions educators. Med Teach 2026 Mar 4;48(3):360-363. [doi: 10.1080/0142159X.2025.2522237]
26. Crowe S, Cresswell K, Robertson A, Huby G, Avery A, Sheikh A. The case study approach. BMC Med Res Methodol 2011 Jun 27;11(1):100. [doi: 10.1186/1471-2288-11-100] [Medline: 21707982]
27. O'Brien BC, Harris IB, Beckman TJ, Reed DA, Cook DA. Standards for reporting qualitative research. Acad Med 2014 Sep;89(9):1245-1251. [doi: 10.1097/ACM.0000000000000388]
28. Eysenbach G. Improving the quality of web surveys: the Checklist for Reporting Results of Internet E-Surveys (CHERRIES). J Med Internet Res 2004 Sep 29;6(3):e34. [doi: 10.2196/jmir.6.3.e34]
29. Moran-Ellis J, Alexander VD, Cronin A, et al. Triangulation and integration: processes, claims and implications. Qual Res 2006 Feb;6(1):45-59. [doi: 10.1177/1468794106058870]
30. Sims D, Cilliers F. Assessing data adequacy in qualitative research studies. Acad Med 2025 Jun 1;100(6):758. [doi: 10.1097/ACM.0000000000005972] [Medline: 39774066]
31. LaDonna KA, Artino AR Jr, Balmer DF. Beyond the guise of saturation: rigor and qualitative interview data. J Grad Med Educ 2021 Oct;13(5):607-611. [doi: 10.4300/JGME-D-21-00752.1] [Medline: 34721785]
32. Varpio L, Paradis E, Uijtdehaage S, Young M. The distinctions between theory, theoretical framework, and conceptual framework. Acad Med 2020 Jul;95(7):989-994. [doi: 10.1097/ACM.0000000000003075] [Medline: 31725464]
33. Köhler T. Multilevel qualitative research: Insights from practice. Eur Manage J 2024 Aug;42(4):503-514. [doi: 10.1016/j.emj.2024.03.011]
34. Jo H. Interaction, novelty, voice, and discomfort in the use of artificial intelligence voice assistant. Univ Access Inf Soc 2025 Aug;24(3):2419-2432. [doi: 10.1007/s10209-025-01203-9]
35. Hosseini MM, Hosseini TM, Qayumi K. Integration of artificial intelligence in medical education: opportunities, challenges, and ethical considerations. J Med Edu 2024 Jan 1;22(1):e140890. [doi: 10.5812/jme-140890]
36. Tang KE, Song T, Zhu Z, Li J, Lee YC. AI literacy education for older adults: motivations, challenges and preferences. 2025 Apr 26 Presented at: CHI EA '25; Apr 26 to May 1, 2025; Yokohama Japan p. 1-15. [doi: 10.1145/3706599.3720033]

XSL•FO
RenderX

37. Teferi B, Omar M, Jeyakumar T, et al. Accelerating the appropriate adoption of artificial intelligence in health care: prioritizing IDEA to champion a collaborative educational approach in a stressed system. Educ Sci 2023;14(1):39. [doi: 10.3390/educsci14010039]
38. Zhou X, Schofield L. Using social learning theories to explore the role of generative artificial intelligence (AI) in collaborative learning. JLDHE 2024 Mar 27(30). [doi: 10.47408/jldhe.vi30.1031]
39. Sichterman B, Noroozi O, Boetje J, van Ginkel S, Khosravi H, Versendaal J. Supporting peer learning with artificial intelligence: a systematic literature review. Innovations Educ Teach Int 2025 Sep 3;62(5):1648-1664. [doi: 10.1080/14703297.2025.2530118]
40. Fani T, Ghaemi F. Implications of Vygotsky's Zone of Proximal Development (ZPD) in teacher education: ZPTD and self-scaffolding. Procedia - Soc Behav Sci 2011;29:1549-1554. [doi: 10.1016/j.sbspro.2011.11.396]
41. Merkebu J, Geringer J, Samuel A. Sampling thoughtfully in qualitative research: hallmarks of rigour in health professions education. Clin Teach 2025 Aug;22(4):e70145. [doi: 10.1111/tct.70145] [Medline: 40640079]

Abbreviations

AI: artificial intelligence
CPD: continuing professional development
GenAI: generative artificial intelligence
HPE: health professions education
IRB: institutional review board
SRQR : Standards for Reporting Qualitative Research
USU: uniformed services university

Edited by S Tsuei; submitted 07.Nov.2025; peer-reviewed by C Schlegel, HE Popeijus, K Chen, K Ikuyinminu; revised version received 25.May.2026; accepted 26.May.2026; published 23.Jun.2026.

Please cite as:
Peacock JG, Merkebu J, Samuel A
Impact of a Practical, Hands-On, Continuing Professional Development Course About AI in Health Care Professions Education on the Perceptions and Behaviors of Health Care Educators: Qualitative Case Study
JMIR Med Educ 2026;12:e87381
URL: https://mededu.jmir.org/2026/1/e87381
doi:10.2196/87381

© Justin G Peacock, Jerusalem Merkebu, Anita Samuel. Originally published in JMIR Medical Education (https://mededu.jmir.org), 23.Jun.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Education, is properly cited. The complete bibliographic information, a link to the original publication on https://mededu.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

AI Integration in Spanish Undergraduate Medical Education: National Cross-Sectional Study

Ana Enériz Janeiro[1], MSc; Karina Pitombeira Pereira[1], PhD; Julio Mayol[2,3,4,5], MD, PhD; Javier Crespo[2,6], MD, PhD; Fernando Carballo[2,7], MD, PhD; Juan B Cabello[2,8], MD, PhD; Manuel Ramos-Casals[2,9], MD, PhD; Bibiana Pérez Corbacho[1], MSc; Juan Turnes[1,2,10], MD, PhD

[1]IDARA Research Group, Galicia Sur Research Institute (IIS Galicia Sur), Vigo, Spain

[10]Department of Gastroenterology and Hepatology, Complejo Hospitalario de Pontevedra, Mourente, s/n, Pontevedra, Galicia, Spain

[2]MedicineAI, Madrid, Spain

[3]Hospital Clínico San Carlos, Madrid, Spain

[4]Universidad Complutense de Madrid, Madrid, Spain

[5]Instituto de Investigación Sanitaria del Hospital Clínico San Carlos, Madrid, Spain

[6]Instituto de Investigación Marqués de Valdecilla, Santander, Cantabria, Spain

[7]Universidad de Murcia, Murcia, Spain

[8]Critical Appraisal Skills Programme: CASP Spain, Alicante, Spain

[9]Postgraduate Degree Program in Artificial Intelligence in Medicine, Universidad de Cádiz, Cádiz, Andalusia, Spain

Corresponding Author:
Juan Turnes, MD, PhD
IDARA Research Group, Galicia Sur Research Institute (IIS Galicia Sur), Vigo, Spain

Abstract

Background: Artificial intelligence (AI) is reshaping clinical practice and redefining the competencies future physicians will need. International bodies, such as the Association of American Medical Colleges, have called for structured AI training in medical curricula. Despite growing international consensus, no systematic nationwide evaluation had been conducted in Spain prior to this study.

Objective: This study aimed to characterize the presence, type, and curricular features of AI-related training across all Spanish universities offering an official medical degree and to assess differences by institutional ownership and geographic region.

Methods: This cross-sectional study was conducted from July to September 2025. Universities were the unit of analysis. A census of all institutions offering an officially recognized medical degree was obtained from the Register of Universities, Centers and Degrees; all 52 eligible institutions were included. Publicly available curricula and course guides for the 2025‐2026 academic year were reviewed by 2 independent researchers and validated by an external evaluator. Courses were classified as (1) a specific AI course (AI as primary topic, accounting for >50% of syllabus), (2) an AI-similar course (a digital health or biomedical informatics course referencing AI as secondary content), or (3) not AI-related training. Course-level variables included ownership (public or private), region, status (compulsory or elective), European Credit Transfer and Accumulation System (ECTS) credits, academic year, and department. All analyses were descriptive. Potential sources of bias were addressed through predefined classification criteria, duplicate independent extraction, and external dataset verification.

Results: Of 52 universities, 36 (69.2%) were public and 16 (30.8%) were private. A total of 10 (19.2%) institutions offered at least one specific AI course; 6 (11.5%) included an AI-similar course. Overall, 16 (30.8%) universities had incorporated AI in some form; 36 (69.2%) institutions had not incorporated AI. Rates were similar for public (7/36, 19.4%) and private institutions (3/16, 18.8%). Identified courses ranged from 3 to 6 ECTS credits, representing an average of 1.17% of the 360-credit degree; most were elective. Only the University of Jaén offered a compulsory course with AI content. Marked regional disparities were observed: Andalusia led with 5 of 9 (55.6%) universities offering a specific AI course, while 10 autonomous communities had no universities with any AI-related training.

Conclusions: This study delivers the first census-based, reproducible, national assessment of AI integration in Spanish undergraduate medical education. Unlike prior work focused on individual programs or nonstandardized definitions, we applied a consistent taxonomic framework reusable for longitudinal monitoring and international benchmarking. Findings reveal a heterogeneous, predominantly elective, and low-weight curricular landscape with striking interregional inequities. These results inform curriculum reform, accreditation standards, and faculty development priorities and support the establishment of minimum national competency standards and systematic monitoring to ensure equitable AI literacy among future physicians in Spain.

XSL•FO
RenderX

(JMIR Med Educ 2026;12:e88511) doi:10.2196/88511

KEYWORDS

artificial intelligence; AI; generative artificial intelligence; medical education; curriculum; Spain

Introduction

Artificial intelligence (AI) has emerged as a transformative technology in the field of health care, revolutionizing everything from medical diagnosis to hospital management and biomedical research [1]. This technological revolution is not only transforming current clinical practice but also redefining the competencies that future health care professionals will need to practice medicine effectively in the coming years [2].

The integration of AI into medical curricula is globally recognized as a priority, underscoring the need to prepare future physicians with the technical and ethical competencies required to ensure the responsible use of these technologies. Moreover, there is an urgent need to develop structured curricular frameworks that systematically incorporate AI education into medical schools [3].

However, the translation of these recommendations into concrete curricula has been uneven and fragmented. The advent of generative AI adds urgency, as it amplifies both the educational potential (eg, support for writing and supervised reasoning) and the risks (hallucinations, data leakage, and authorship attribution), making it essential to teach strategies for verification, documentation, and boundaries of use. The lack of AI training is also acknowledged by the students themselves. Global and regional surveys show broad support for its teaching, calling for its inclusion in curricula while also expressing concern about ethical issues, the physician-patient relationship, and its impact on specialty choice [4,5]. In a study conducted with medical students in Spain, 83% considered it essential to acquire knowledge of AI for their future professional practice [6].

To address this need, several countries and organizations have begun implementing systematic initiatives to incorporate AI into medical education. In Canada, curricular proposals and needs assessments have been developed to identify key competencies in areas such as ethics, legal aspects, clinical application, and collaborative work [7,8]. In the United States, several medical schools have advanced the formal integration of AI into their curricula, with prominent examples including the Icahn School of Medicine at Mount Sinai, the latter being a pioneer in its cross-curricular incorporation [9]. In addition, international studies involving medical students and faculty consistently indicate that AI training remains limited and urgently call for curricular updates to address the digital transformation in health care [4,10,11].

The Spanish context presents relevant particularities: 52 medical schools, a 360-ECTS (European Credit Transfer and Accumulation System) degree program, and quality assurance frameworks that both enable and require transparent curricular structures. Despite the proliferation of isolated initiatives, to the best of our knowledge, there is no national, systematic, and reproducible characterization of how AI is being incorporated into curricula, what educational weight it receives (in ECTS), whether it is compulsory or elective, and to what extent it includes specific content on generative AI.

Recent syntheses highlight that, despite AI's growing relevance to clinical decision-making and health system workflows, undergraduate medical education remains uneven and fragmented globally. A scoping review mapping the use of AI across medical education settings found that, while AI applications are proliferating in areas such as clinical reasoning support and automated assessment, their integration into formal curricula remains largely ad hoc and driven by individual institutional initiatives rather than coordinated policy. Persistent implementation barriers (including limited faculty capacity, unclear competency targets, insufficient institutional support, and the absence of consensus on what should be taught and how learning should be assessed) continue to hinder systematic adoption. These findings underscore the need for national-level mapping studies, such as this one, to identify gaps and inform evidence-based curricular reform [12].

Although several curriculum frameworks and educational initiatives have been proposed to define core AI competencies for medical learners, approaches still vary widely in terminology, depth, and evaluation. Some frameworks emphasize foundational data science concepts, others prioritize clinical decision support skills, and still others focus on ethical and regulatory dimensions, with little overlap in learning objectives or assessment strategies. This heterogeneity limits comparability across institutions and hinders the establishment of benchmarks and minimum standards that could ensure baseline AI literacy for all medical graduates. The lack of a shared taxonomic language for classifying AI content in medical curricula makes it particularly difficult to compare findings across countries or track progress over time, a gap that this study aims to address by proposing a reproducible classification framework [13].

The emergence of generative AI has heightened the urgency for structured training in medical schools. Tools such as ChatGPT (OpenAI), Gemini (Google Inc), and other large language models are already being used by medical students for study support, clinical reasoning, and academic writing, often without formal institutional guidance. This raises concerns about hallucinated content, data privacy, academic integrity, and uncritical reliance on AI-generated outputs. Therefore, medical schools need not only to teach the fundamentals of AI but also to develop governance frameworks, ethical guidance, and practical instruction on the safe and responsible use of generative AI alongside monitoring mechanisms to track curricular adoption and inform institutional policy. Against this background, understanding the current state of AI education in a national context becomes essential for identifying priority areas for intervention [14].

XSL•FO
RenderX

The primary objective of this study was to provide the first analysis of the current state of AI integration in undergraduate medical education programs in Spain, based on reproducible criteria and supported by an open dataset for verification and reuse. Specifically, we aimed to characterize the presence, type, and curricular features of AI-related teaching across Spanish universities offering a medical degree and to assess differences by university ownership (public vs private) and geographic region. We hypothesized that AI-related training would be limited in scope, predominantly elective, and unevenly distributed across institutions.

Methods

This study was prepared in accordance with the STROBE (Strengthening the Reporting of Observational Studies in Epidemiology) statement for cross-sectional studies, and the corresponding checklist was used to ensure complete reporting of study design, setting, variables, data sources, bias, and statistical methods (Checklist 1).

Study Design

This was a cross-sectional study conducted between July and September 2025 to analyze the integration of AI education in undergraduate medical programs in Spain.

Setting

A census of all Spanish universities offering an official degree in medicine was conducted using the "Register of Universities, Centers and Degrees (Registro de Universidades, Centros y Títulos)" (RUCT) [15], following the search parameters specified in Table 1.

Table . Consultation on medical degrees offered in Spain: Register of Universities, Centers and Degrees (Registro de Universidades, Centros y Títulos; RUCT)[a].

Search field	Search selection
University	All
Degree name	Medicine
Academic level	Bachelor's degree (*grado*)
Field	Health sciences
Area	Medicine and dentistry
Status	All
Historical search	No

[a]Data collected from July to September 2025 (academic year: 2025 - 2026). The table lists the query fields and selections applied in the RUCT to retrieve eligible institutions.

Participants and Unit of Analysis

The unit of analysis was the university. This registry currently lists 52 universities offering a degree in medicine in Spain; the RUCT database reflects a 2021 update, and all institutions were independently confirmed as actively offering the program for the 2025 - 2026 academic year through direct portal access (last verified on September 4, 2025). A complete list of the 52 included universities, with RUCT identifiers and source URLs for verification, is provided in Multimedia Appendix 1. Universities were eligible if they (1) offered an officially recognized medical degree program for 2025 - 2026 and (2) had publicly accessible curricular information sufficient to assess whether AI-related training was present in the curriculum (degree plan or course guides). No restrictions were applied based on ownership or region.

Variables

The search strategy was based on the selection of institutions currently offering an active degree in medicine. The institutional websites of each university were reviewed to identify information on curricula, courses, and AI-related competencies within the medical degree for the 2025 - 2026 academic year. This study relied exclusively on publicly available information from official websites and public curricular documents, including study plans, course programs, and teaching guides.

A university was considered to offer a specific AI course if the Spanish term for "artificial intelligence" ("*inteligencia artificial*") or "AI" ("*IA*") appears in the course title and/or if more than 50% of the content is dedicated to AI.

The existence of courses that might include content on AI, even if not explicitly stated in the title, was also examined. The Spanish keywords used for this selection were "new technologies" ("*nuevas tecnologías*"), "informatics" ("*informática*"), "communication" ("*comunicación*"), "digital media ("*medios digitales*")," or "digital health" ("*salud digital*"). Thus, a university is considered to offer a course "similar to AI or appearing in competencies" when AI accounts for less than 50% of the course syllabus and/or when reference is made to the keywords. Universities are considered to have no AI courses when they do not include any course on AI or similar to AI.

The main outcome was the presence of AI-related training in the medicine curriculum, operationalized into mutually exclusive categories based on course content, as follows: specific AI courses (the course explicitly addresses AI in medicine or health care as a primary topic), AI-similar courses (courses focusing on computational, statistical, or digital health competencies closely related to AI, eg, data science and bioinformatics, but without explicitly positioning AI as the main subject), and non-AI-related training (there was no evidence of AI-related content in the degree plan or available course guides).

XSL•FO
RenderX

For each course, the following information was collected: university, region, type (public or private), title, status (compulsory or elective), training credits, academic year in which it is offered, department (or alternatively, the specialty of the faculty teaching it), and main course content.

Data Sources and Measurement

This study relied exclusively on publicly available information from official university websites and public curricular documents, including study plans, course programs, and teaching guides. To ensure the quality and accuracy of the collected data, each data entry was independently verified by 2 researchers and validated by an external evaluator not involved in the data collection process.

Bias

We anticipated potential sources of bias typical of web-based curriculum assessments: (1) information bias due to incomplete or outdated public documentation, potentially underestimating AI-related training; (2) misclassification bias due to heterogeneous terminology across universities (AI content embedded under non-AI labels); and (3) selection bias at the variable level when specific course attributes (eg, department or academic year) were not publicly reported. To mitigate these risks, we applied predefined classification criteria (including the >50% rule), used duplicate independent extraction, and conducted external verification of the compiled dataset.

Study Size

Study size was determined a priori by a census of all eligible universities in Spain offering an official medical degree in 2025‐2026. The final analytic sample comprised 52 universities.

Statistical Methods

All analyses were descriptive. We summarized the number and proportion of universities that offered a specific AI course, an AI-similar course, or no AI-related training, and we described the characteristics of identified courses—such as whether they were compulsory or elective, their ECTS credits, the academic year or level, and the responsible department—whenever this information was available. Given the census-based and primarily descriptive aim of the study, we did not fit adjusted models to control for confounding. To explore heterogeneity, we presented results stratified by university ownership (public or private) and by autonomous community or region. When specific variables were not publicly reported (eg, department, academic year or level, or ECTS credits), they were coded as missing and analyses were conducted using the available information, with denominators varying accordingly. Because we assessed the full population of eligible programs rather than a sample, methods accounting for a sampling strategy were not applicable, and no formal sensitivity analyses were performed.

To examine geographic variation, we conducted a stratified descriptive analysis by autonomous community. For each community, the denominator was the number of universities offering an official medical degree, and we calculated the proportion of universities that offered at least one specific AI course and an AI-similar course, as defined in the study taxonomy. Communities with no medicine programs (denominator=0) were not assigned estimates.

Ethical Considerations

This study analyzed exclusively publicly available curricular information from official university websites and institutional documents. No individual-level human data were collected, and no surveys, interviews, or interactions with human participants were conducted. In accordance with Spanish regulations governing research that relies solely on publicly accessible, nonpersonal data, formal ethics committee review was not required, and the study was exempt from institutional review board approval. This exemption is consistent with the provisions of Spanish Law 14/2007 on Biomedical Research [16], which applies to studies involving human subjects or their biological samples. As this study did not involve human participants or personal data, informed consent was not applicable; no primary data collection from individuals took place, and therefore no consent process was required. All data used in this study were obtained from publicly accessible institutional sources. No personally identifiable information was collected at any stage; the dataset contains only institution-level variables (university name, region, and course characteristics), and no individual-level data were gathered, stored, or processed. Therefore, the study poses no risk to the privacy or confidentiality of any individual. No human participants were involved in this study; therefore, no compensation was provided or applicable. No images containing identifiable individuals are included in this manuscript or its supplementary materials.

Results

A total of 52 universities offering an official medical degree were identified through the RUCT. We then carried out systematic verification using publicly available institutional sources, university websites, and course guides to confirm eligibility and extract curriculum-level and course-level variables. All 52 eligible universities were included in the analysis, and no institutions were excluded.

Scope and Distribution

The analysis included a total of 52 Spanish universities offering a degree in medicine, representing all institutions with officially recognized programs in Spain for the 2025‐2026 academic year. Of these universities, 36 (69.2%) were public institutions and 16 (30.8%) were private institutions.

The territorial distribution showed a concentration of medical schools in certain regions. Madrid led with 10 (19.2%) universities, followed by Andalusia with 9 (17.3%) and Catalonia with 8 (15.4%). The Valencian Community had 6 (11.5%) universities, while the Canary Islands had 3 (5.8%) institutions. The remaining autonomous communities each had between 1 (1.9%) and 2 (3.8%) universities, revealing a heterogeneous distribution that reflects both population density and the historical development of medical education in each territory.

XSL•FO
RenderX

Distribution of AI Education

Of the 52 universities analyzed, 10 (19.2%) offer courses specifically dedicated to AI in medicine (Table 2). Another 6 (11.5%) universities include an AI-similar course within broader courses or mention AI-related competencies in their learning objectives (Table 3). Therefore, considering both universities with specific AI courses and those with AI-similar courses, a total of 16 (30.8%) institutions have incorporated AI into medical education in some form, whereas 36 (69.2%) institutions have not yet included explicit AI content in their medicine curricula.

Table . Universities with specific artificial intelligence (AI) courses in the medical degree program in Spain[a].

University	Type	Region	Course	Status	ECTS[b] credit	Year/level of medical degree program	Department	Generative AI inclusion
Autonomous University of Barcelona	Public	Catalonia	Artificial intelligence and health	Elective	3	Third	Surgery	No
Complutense University of Madrid	Public	Madrid	Practical application of generative artificial intelligence	Elective	3	First and second	Radiology, Rehabilitation, and Physiotherapy	Yes
Complutense University of Madrid	Public	Madrid	Big data and artificial intelligence in medicine	Elective	3	Third and sixth	Medicine	No
Fernando Pessoa University, Canary Islands	Private	Canary Islands	Big data and artificial intelligence in medicine	Elective	3	—[c]	—	—
University of Huelva	Public	Andalusia	Application of big data and artificial intelligence in medicine	Elective	3	Fourth	—	—
Loyola Andalusia University	Private	Andalusia	Artificial intelligence and healthcare	Elective	3	Fourth	—	—
University of Lleida	Public	Catalonia	Artificial intelligence in medicine	Elective	3	First	Basic Medical Sciences	No
Camilo José Cela University	Private	Madrid	Fundamentals of artificial intelligence in healthcare	Elective	3	Third	—	Yes
University of Córdoba	Public	Andalusia	Documentation and artificial intelligence in medicine. history of medicine	Basic	6	First	Medical and Surgical Sciences	No
University of Almería	Public	Andalusia	Medical informatics	Elective	3	Fourth	Informatics	No
University of Jaén	Public	Andalusia	New clinical and biomedical information technologies	Compulsory	3	Second	Microbiology	No

[a]Cross-sectional census of 52 Spanish universities offering an official medical degree (academic year: 2025 - 2026); data collected from July to September 2025.

[b]ECTS: European Credit Transfer and Accumulation System.

[c]Information not publicly available or not verifiable from publicly accessible institutional sources.

XSL•FO
RenderX

Table . Universities with courses categorized as artificial intelligence (AI)–similar course in medical degree programs in Spain[a].

University	Type	Region	Courses	Status	ECTS[b] credit	Year/level of medical degree program	Department	AI content detail
University of Vic - Central University of Catalonia	Private	Catalonia	Information and communication technologies and health	Elective	5	First and second	—[c]	It mentions AI in the objectives of existing courses, and the university has a specific AI professorship
Alfonso X El Sabio University	Private	Madrid	Digital health	Elective	6	Sixth	—	Among the contents detailed are AI in value-based health care; AI-based decision support tools; relevant data in the health care setting and methods for acquiring, preparing, and storing them; data mining; natural language processing; machine learning; and success stories of intelligent systems, such as IBM Watson, including associated challenges, barriers, and facilitators
University of Extremadura	Public	Extremadura	Applied medical informatics	Elective	6	Fifth	Department of Computer Systems and Telematics Engineering, Languages and Computer Systems	Contents include fundamentals of informatics, introduction to medical informatics, search and management of health information, communication and dissemination of health information, and health information systems
Rovira i Virgili University	Public	Catalonia	New Technologies and data management	Compulsory	3	Third	—	The course includes management and analysis of health data with advanced technological components

XSL•FO
RenderX

University	Type	Region	Courses	Status	ECTS[b] credit	Year/level of medical degree program	Department	AI content detail
University of Oviedo	Public	Asturias	Pathological anatomy	—	—	—	—	—
University of Oviedo	Public	Asturias	Fundamentals of surgery	—	—	—	—	AI is mentioned in Topic 8 (“New challenges and future of surgery”), specifically in Session 31 (“Telesurgery: robotics, virtual reality, and artificial intelligence”)
King Juan Carlos University	Public	Madrid	Introduction to medicine: medical information and documentation	Compulsory	6	First	Medical Specialties and Public Health	Module I includes information management and clinical research questions (highlighting Unit 7 on “Fundamentals of evidence-based medicine”); formulating problems and questions from the patient to the clinician-researcher; the DEPTh model; questions and levels of training from the patient to the researcher; and AI-based tools for evidence-based medicine, specifically generative AI tools (GPTs, ChatGPT, Elicit, and others), for reading, analysis, and writing

[a]Cross-sectional census of 52 Spanish universities offering an official medical degree (academic year: 2025 - 2026); data collected from July to September 2025.

[b]ECTS: European Credit Transfer and Accumulation System.

[c]Information not publicly available or not verifiable from publicly accessible institutional sources.

Information on some course-level characteristics (eg, department and academic year or level) was not publicly available for all courses and was therefore reported as not available.

By type of institution, among the 36 public universities, 7 (19.4%) offered specific AI courses, whereas among the 16 private universities, 3 (18.8%) included AI courses in their curricula (Table 4).

XSL•FO
RenderX

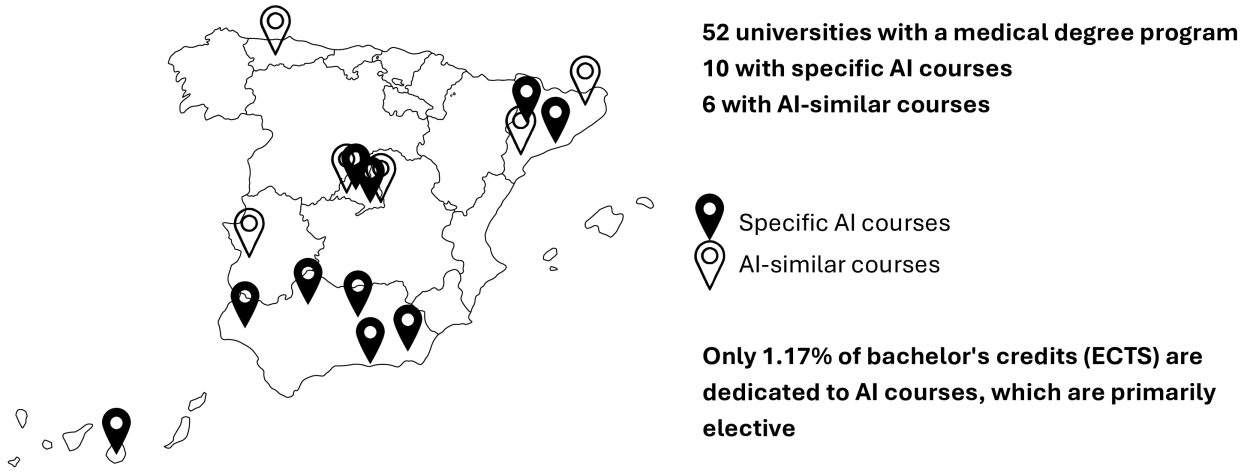
Table . Distribution of artificial intelligence (AI) courses by type: specific AI, AI-similar course, or no AI training[a].

University type	Specific AI course, n (%)	AI-similar course, n (%)	AI training, n (%)
Public (n=36)	7 (19.4)	4 (11.1)	25 (69.4)
Private (n=16)	3 (18.8)	2 (12.5)	11 (68.8)
Total (n=52)	10 (19.2)	6 (11.5)	36 (69.2)

[a]Census of 52 universities in Spain (academic year: 2025 - 2026); data collected from July to September 2025. Values are counts and percentages of universities offering at least one specific AI course, AI-similar content, or no AI-related training in publicly available curricular documentation.

The territorial analysis revealed a marked disparity among autonomous communities. Andalusia led, with 5 (55.6%) of its 9 universities offering specific AI courses. It was followed by Catalonia, with 2 (25%) of 8 universities, and Madrid, with 2 (20%) of 10 universities. In the Canary Islands, 1 (33.3%) of the 3 universities had incorporated a specific AI course. An AI-similar course was additionally identified in Catalonia, Madrid, Extremadura, and Asturias. By contrast, 10 regions—Aragon, Balearic Islands, Cantabria, Castile-La Mancha, Castile and León, Valencian Community, Galicia, Murcia, Navarra, and the Basque Country—did not have universities that had integrated any form of AI-related training into their medicine curricula (Figure 1).

Figure 1. Territorial distribution of universities with artificial intelligence (AI) courses in the medical degree program in Spain. ECTS: European Credit Transfer and Accumulation System.



Characteristics of AI Education

Among the universities that included specific courses on AI in their curricula, most of them were offered as electives. The only exceptions were the University of Jaén, which offered a compulsory course, and University of Córdoba, which included a core course (Table 2). The course at the University of Jaén represented a unique case, as it was the only institution that offered a compulsory course with AI content within the medical degree. Although the course did not explicitly mention AI in its title, "New Clinical and Biomedical Information Technologies ("*Nuevas Tecnologías de la Información Clínica y Biomédica*)," its syllabus explicitly incorporated competencies related to this field, making it the only identified case of mandatory curricular integration.

ECTS Credits

The analysis of the credit load dedicated to AI in the medical degree revealed that the specific AI courses identified ranged between 3 and 6 ECTS credits (Table 2). In relation to the total of 360 credits required for the degree, AI training constituted, on average, 1.17% of the medical curriculum. Nevertheless, since most of these courses were offered as electives, they were not included in the compulsory credit load, and not all students enrolled in them. Consequently, the actual weight of AI training within the degree program may be regarded as even smaller. Complutense University of Madrid showed the highest dedication with 6 credits, while other institutions generally offer 3-credit courses. In some cases, AI content was included within broader courses on medical informatics or health technologies, which made it difficult to quantify its exact weight.

XSL•FO
RenderX

Discussion

Principal Findings

Despite progress in incorporating AI into clinical medicine, its presence in medical curricula in Spain remains very limited and heterogeneous. Most institutions that have included it do so through elective courses of low academic weight (1.17% of the 360 ECTS credits), with approaches ranging from specific courses to the inclusion of AI within transversal competencies but with no common framework or clearly defined learning objectives—a pattern consistent with the global gap documented in recent international research [13,14]. Available data from Spanish medical students corroborate this assessment: in a cross-sectional survey across 192 faculties, the Spanish subset (74 students from 5 universities) reported a lack of AI courses, with 74% describing limited knowledge and 80% expressing unpreparedness to apply AI in professional practice [17]. Bridging this gap calls for coordinated policy action, including the transition of elective AI courses into core or compulsory content, dedicated faculty training programs, and a national framework of competencies analogous to those promoted by the Association of American Medical Colleges and the General Medical Council [18,19].

The territorial analysis revealed a marked disparity: while Andalusia led with 5 (55.6%) of 9 universities offering a specific AI course, 10 autonomous communities had no universities offering any AI-related training. This pattern, consistent with reports from other countries [4], reinforces the need for national strategies to reduce geographic inequity. Structural factors specific to the Spanish context help explain the pace of adoption: decentralized governance across 17 autonomous communities, the absence of explicit AI requirements in national accreditation standards, and limited faculty capacity to design and deliver AI curricula together slow progress [12,13].

Internationally, institutions in the United States such as the Icahn School of Medicine at Mount Sinai and the Stanford Center for Artificial Intelligence in Medicine and Imaging have pioneered the formal integration of AI into medical education [20,21]. In Europe, countries including Germany, Belgium, and the United Kingdom have launched pilot programs to assess essential AI competencies and develop adapted curricular frameworks [22-24]. Spain's position, characterized by the limited and predominantly elective nature of its current provision, underscores how much ground remains to be covered.

The case for reform is amplified by converging pressures. Spain's aging population and growing burden of chronic disease demand optimization of health care resources [25], and while AI offers tools to improve diagnostic accuracy, personalize treatments, and reduce clinical errors [26], realizing these benefits requires professionals capable of navigating the ethical and interpretative challenges involved [27]. The rapid evolution of generative AI compounds this urgency: models such as Google's MedLM are already deployed in clinical settings [28], and multimodal systems are achieving high accuracy in combined medical tasks [29], making curricular adaptation time sensitive as delays carry economic and opportunity costs for universities and the health system alike [1]. Therefore, curricula should incorporate explicit strategies for the supervision and responsible use of these technologies, including guidelines for verification, documentation, and boundaries of use, a need that is underscored by evidence that the generational gap in AI tool mastery can strain teacher-student relationships and increase the risk of inappropriate use without adequate supervision frameworks [30,31]. Insufficient training could ultimately deepen disparities in care quality and limit Spain's capacity to adopt clinical innovations at a pace comparable to other European countries [17].

This study has several limitations. First, reliance on publicly available information may not fully capture internal initiatives or unpublished plans. Second, the categorization of AI content depended on the terminology used in official documents, potentially resulting in inconsistent classifications. Third, because individual preconsensus classifications were not retained as separate records, a formal interrater reliability statistic (eg, Cohen kappa) could not be reported; however, the binary, mutually exclusive categories and the predefined >50% content threshold were designed to minimize subjectivity. Fourth, this cross-sectional design captured a single academic year (2025 - 2026) and could not reflect the pace of curricular change. Finally, the authors' use of AI-assisted tools, disclosed in the Acknowledgments section, could be perceived as reflecting a favorable stance toward AI adoption; however, data collection and classification were conducted independently of any AI tools, and all analytical decisions were made by the authors. Readers should be aware that the research topic itself may predispose toward a positive framing.

Future research should include direct surveys of academic coordinators at all 52 institutions to capture informal initiatives not visible in official documents. Longitudinal studies repeating this census at regular intervals would allow monitoring of curricular change over time. To enable systematic tracking, we propose the following four indicators: (1) proportion of universities offering at least one compulsory AI course, (2) mean ECTS specifically dedicated to AI per medical degree, (3) proportion of programs incorporating generative AI content, and (4) an interregional Gini coefficient of AI training distribution to quantify geographic equity. Implementation of such a framework would benefit from the involvement of the Spanish Ministry of Universities, National Agency for Quality Assessment and Accreditation (Agencia Nacional de Evaluación de la Calidad y Acreditación), medical school deans, and Spanish Society for Medical Education and Health Sciences (Sociedad Española de Educación Médica y de Ciencias de la Salud). International collaboration through organizations such as the International Association for Health Professions Education could further facilitate shared standards and translational research in digital health education.

Research in Context

Evidence Before This Study

We searched PubMed, Scopus, and Web of Science databases up to July 2025 for studies on the inclusion of AI courses or formal AI training in undergraduate medical curricula. We found international initiatives describing the introduction of AI in medical education, mainly in North America and Asia, but these

XSL•FO
RenderX

were often limited to individual universities or pilot projects. To our knowledge, no nationwide systematic evaluation of AI integration into undergraduate medical curricula in Spain had been reported.

Added Value of This Study

This study provides the first national overview of how AI is being integrated into undergraduate medical education in Spain. By analyzing the official curricula of all 52 universities offering medical degrees, we demonstrate that AI training is scarce, mostly elective, and carries minimal academic weight. We also identify striking regional disparities: while some regions such as Andalusia have implemented courses, others have none. This comprehensive analysis highlights the uneven and inconsistent adoption of AI in medical curricula.

Implications of All the Available Evidence

Our findings show that AI education in Spanish medical schools is at an early and unevenly distributed stage, with low ECTS credit weight and a predominance of elective courses. In light of international competency frameworks [13,14] and student-reported perceptions of unpreparedness [17], this limited provision may be insufficient to meet the evolving demands of AI-driven clinical practice. Considering international recommendations, the evidence supports the establishment of minimum standards, stronger national coordination, and systematic monitoring of AI training in medical curricula to ensure equitable and consistent preparation across regions.

Conclusion

In conclusion, this study provides the first systematic assessment of artificial intelligenceAI integration in undergraduate medical education in Spain, revealing a landscape that warrants attention. AI training in medical curricula is still nascent and unevenly distributed, underscoring the need for updated educational policies to ensure that future health care professionals acquire competencies aligned with technological advances. Given the growing role of AI in clinical practice, this deficit in formal training may leave graduates less well equipped than peers in countries with more structured curricula. The main contribution is a comprehensive, census-based overview with an operationalized taxonomy (specific AI course and AI-similar course) that, unlike prior studies focused on isolated programs or non-standardized definitions, provides a consistent, reproducible framework reusable for longitudinal monitoring and international comparison, and a clear evidence base to inform curriculum reform, accreditation standards, and faculty development priorities.

Acknowledgments

The authors declare the use of generative artificial intelligence (GenAI) in the research and writing process. According to the Generative AI Delegation Taxonomy (2025), the following tasks were delegated to GenAI tools under full human supervision: data cleaning, proofreading and editing, translation, reformatting, and identification of limitations. The GenAI tools used were ChatGPT-4 (GPT-5; OpenAI), Gemini 2.5 Pro (Google), Manus (Butterfly Effect Pte Ltd), and Claude Opus 4.5 (Anthropic). Responsibility for the final manuscript lies entirely with the authors. GenAI tools are not listed as authors and do not bear responsibility for the final outcomes. The declaration was submitted collectively by the authors.

Funding

No funding was received for this study.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Complete classification of Spanish universities offering a medical degree in the 2025-2026 academic year according to the presence of artificial intelligence–related training in its medical curriculum.
[DOCX File, 1831 KB - mededu_v12i1e88511_app1.docx]

Checklist 1
STROBE checklist.
[PDF File, 116 KB - mededu_v12i1e88511_app2.pdf]

References

1. Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. Nat Med 2019 Jan;25(1):44-56. [doi: 10.1038/s41591-018-0300-7] [Medline: 30617339]
2. Borges do Nascimento IJ, Abdulazeem HM, Vasanthan LT, et al. The global effect of digital health technologies on health workers' competencies and health workplace: an umbrella review of systematic reviews and lexical-based and sentence-based meta-analysis. Lancet Digit Health 2023 Aug;5(8):e534-e544. [doi: 10.1016/S2589-7500(23)00092-4] [Medline: 37507197]
3. Knopp MI, Warm EJ, Weber D, et al. AI-enabled medical education: threads of change, promising futures, and risky realities across four potential future worlds. JMIR Med Educ 2023 Dec 25;9:e50373. [doi: 10.2196/50373] [Medline: 38145471]

XSL•FO
RenderX

4. Civaner MM, Uncu Y, Bulut F, Chalil EG, Tatli A. Artificial intelligence in medical education: a cross-sectional needs assessment. BMC Med Educ 2022 Nov 9;22(1):772. [doi: 10.1186/s12909-022-03852-3] [Medline: 36352431]
5. Sit C, Srinivasan R, Amlani A, et al. Attitudes and perceptions of UK medical students towards artificial intelligence and radiology: a multicentre survey. Insights Imaging 2020 Feb 5;11(1):14. [doi: 10.1186/s13244-019-0830-7] [Medline: 32025951]
6. Gualda-Gea JJ, Barón-Miras LE, Bertran MJ, Vilella A, Torá-Rocamora I, Prat A. Perceptions and future perspectives of medical students on the use of artificial intelligence based chatbots: an exploratory analysis. Front Med (Lausanne) 2025;12:1529305. [doi: 10.3389/fmed.2025.1529305] [Medline: 39911871]
7. Pucchio A, Rathagirishnan R, Caton N, et al. Exploration of exposure to artificial intelligence in undergraduate medical education: a Canadian cross-sectional mixed-methods study. BMC Med Educ 2022 Nov 28;22(1):815. [doi: 10.1186/s12909-022-03896-5] [Medline: 36443720]
8. Singla R, Pupic N, Ghaffarizadeh SA, et al. Developing a Canadian artificial intelligence medical curriculum using a Delphi study. NPJ Digit Med 2024 Nov 18;7(1):323. [doi: 10.1038/s41746-024-01307-1] [Medline: 39557985]
9. Brouillette M. AI added to the curriculum for doctors-to-be. Nat Med 2019 Dec;25(12):1808-1809. [doi: 10.1038/s41591-019-0648-3] [Medline: 31806886]
10. Wood EA, Ange BL, Miller DD. Are we ready to integrate artificial intelligence literacy into medical school curriculum: students and faculty survey. J Med Educ Curric Dev 2021;8:23821205211024078. [doi: 10.1177/23821205211024078] [Medline: 34250242]
11. Bisdas S, Topriceanu CC, Zakrzewska Z, et al. Artificial intelligence in medicine: a multinational multi-center survey on the medical and dental students' perception. Front Public Health 2021;9:795284. [doi: 10.3389/fpubh.2021.795284] [Medline: 35004598]
12. Rincón EHH, Jimenez D, Aguilar LAC, Flórez JMP, Tapia Á, Peñuela CLJ. Mapping the use of artificial intelligence in medical education: a scoping review. BMC Med Educ 2025 Apr 12;25(1):526. [doi: 10.1186/s12909-025-07089-8] [Medline: 40221725]
13. Tolentino R, Baradaran A, Gore G, Pluye P, Abbasgholizadeh-Rahimi S. Curriculum frameworks and educational programs in AI for medical students, residents, and practicing physicians: scoping review. JMIR Med Educ 2024 Jul 18;10:e54793. [doi: 10.2196/54793] [Medline: 39023999]
14. Triola MM, Rodman A. Integrating generative artificial intelligence into medical education: curriculum, policy, and governance strategies. Acad Med 2025 Apr 1;100(4):413-418. [doi: 10.1097/ACM.0000000000005963] [Medline: 39705530]
15. Registro de Universidades, Centros y Títulos (RUCT). Ministerio de Ciencia, Innovación y Universidades. URL: https://www.educacion.gob.es/ruct/home [accessed 2025-09-17]
16. Ley 14/2007, de 3 de julio, de investigación biomédica. Boletín Oficial del Estado. URL: https://www.boe.es/eli/es/l/2007/07/03/14/con [accessed 2026-05-14]
17. Busch F, Hoffmann L, Truhn D, et al. Global cross-sectional student survey on AI in medical, dental, and veterinary education and practice at 192 faculties. BMC Med Educ 2024 Sep 28;24(1):1066. [doi: 10.1186/s12909-024-06035-4] [Medline: 39342231]
18. Artificial intelligence and academic medicine. Association of American Medical Colleges. URL: https://www.aamc.org/about-us/mission-areas/medical-education/artificial-intelligence-and-academic-medicine [accessed 2025-10-07]
19. Artificial intelligence and innovative technologies. General Medical Council. URL: https://www.gmc-uk.org/professional-standards/learning-materials/artificial-intelligence-and-innovative-technologies [accessed 2025-10-07]
20. Icahn School of Medicine at Mount Sinai expands AI innovation with OpenAI's ChatGPT Edu rollout. Mount Sinai. URL: https://www.mountsinai.org/about/newsroom/2025/icahn-school-of-medicine-at-mount-sinai-expands-ai-innovation-with-openais-chatgpt-edu-rollout [accessed 2025-09-17]
21. Center for Artificial Intelligence in Medicine & Imaging. URL: https://aimi.stanford.edu/ [accessed 2025-10-07]
22. Moldt JA, Festl-Wietek T, Fuhl W, et al. Assessing AI awareness and identifying essential competencies: insights from key stakeholders in integrating AI into medical education. JMIR Med Educ 2024 Jun 12;10:e58355. [doi: 10.2196/58355] [Medline: 38989834]
23. Pizzolla I, Aro R, Duez P, De Lièvre B, Briganti G. Integrating artificial intelligence into medical education: lessons learned from a Belgian initiative. J Interact Learn Res 2023;34(2):401-424. [doi: 10.70725/265312mqoebd]
24. Banerjee M, Chiew D, Patel KT, et al. The impact of artificial intelligence on clinical education: perceptions of postgraduate trainee doctors in London (UK) and recommendations for trainers. BMC Med Educ 2021 Aug 14;21(1):429. [doi: 10.1186/s12909-021-02870-x] [Medline: 34391424]
25. Caride-Miana E, Orozco-Beltrán D, Quesada-Rico JA, Mira-Solves JJ. The impact of chronic diseases on all-cause mortality in Spain: a population-based cohort study. Aten Primaria 2025 May;57(5):103112. [doi: 10.1016/j.aprim.2024.103112] [Medline: 39531986]
26. Yu KH, Beam AL, Kohane IS. Artificial intelligence in healthcare. Nat Biomed Eng 2018 Oct;2(10):719-731. [doi: 10.1038/s41551-018-0305-z] [Medline: 31015651]
27. Char DS, Shah NH, Magnus D. Implementing machine learning in health care - addressing ethical challenges. N Engl J Med 2018 Mar 15;378(11):981-983. [doi: 10.1056/NEJMp1714229] [Medline: 29539284]

XSL•FO
RenderX

28. MedLM: generative AI fine-tuned for the healthcare industry. Google Cloud. URL: https://cloud.google.com/blog/topics/healthcare-life-sciences/introducing-medlm-for-the-healthcare-industry [accessed 2025-09-17]
29. Advancing medical AI with Med-Gemini. Google Research. URL: https://research.google/blog/advancing-medical-ai-with-med-gemini/ [accessed 2025-09-17]
30. Abdulnour RE, Gin B, Boscardin CK. Educational strategies for clinical supervision of artificial intelligence use. N Engl J Med 2025 Aug 21;393(8):786-797. [doi: 10.1056/NEJMra2503232] [Medline: 40834302]
31. Principles for the responsible use of artificial intelligence in and for medical education. Association of American Medical Colleges. URL: https://www.aamc.org/about-us/mission-areas/medical-education/principles-ai-use [accessed 2025-09-17]

Abbreviations

AI: artificial intelligence
ECTS: European Credit Transfer and Accumulation System
RUCT: Register of Universities, Centers and Degrees (Registro de Universidades, Centros y Títulos)
STROBE: Strengthening the Reporting of Observational Studies in Epidemiology

Edited by S Brini; submitted 09.Dec.2025; peer-reviewed by E Ortiz-Prado, J Simoni; revised version received 16.Apr.2026; accepted 17.Apr.2026; published 08.Jun.2026.

Please cite as:
Enériz Janeiro A, Pitombeira Pereira K, Mayol J, Crespo J, Carballo F, B Cabello J, Ramos-Casals M, Pérez Corbacho B, Turnes J
AI Integration in Spanish Undergraduate Medical Education: National Cross-Sectional Study
JMIR Med Educ 2026;12:e88511
URL: https://mededu.jmir.org/2026/1/e88511
doi:10.2196/88511

© Ana Enériz Janeiro, Karina Pitombeira Pereira, Julio Mayol, Javier Crespo, Fernando Carballo, Juan B Cabello, Manuel Ramos-Casals, Bibiana Pérez Corbacho, Juan Turnes. Originally published in JMIR Medical Education (https://mededu.jmir.org), 8.Jun.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Education, is properly cited. The complete bibliographic information, a link to the original publication on https://mededu.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

AI-Generated Multiple Mini Interview (MMI) Stations for Medical School Admissions: Psychometric Evaluation

Sabryn Hamila[1], BSc (Hons), PhD; Kyle Birchill[1], BCom; Khoa Cao[1], MBBS (Hons); Md Nassif Hossain[2], BMedSc, MPH, GCTE; Shane Bullock[1,3], BSc (Hons), PhD; Wayne Hodgson[1,3,4], BSc, Graduate Certificate in Higher Education, PhD; Julia Harrison[1], BMedSc, MD; Michelle Leech[1], MBBS (Hons), PhD

[1]School of Medicine, Faculty of Medicine, Nursing and Health Sciences, Monash University, 27 Rainforest Walk, Clayton, Australia

[2]School of Public Health and Preventive Medicine, Faculty of Medicine, Nursing and Health Sciences, Monash University, Clayton, Australia

[3]School of Rural Health, Faculty of Medicine, Nursing and Health Sciences, Monash University, Warragul, Victoria, Australia

[4]Sub-Faculty of Health Sciences, Faculty of Medicine, Nursing and Health Sciences, Monash University, Clayton, Victoria, Australia

Corresponding Author:
Sabryn Hamila, BSc (Hons), PhD
School of Medicine, Faculty of Medicine, Nursing and Health Sciences, Monash University, 27 Rainforest Walk, Clayton, Australia

Abstract

Background: Multiple mini interviews (MMIs) are widely used in medical school admissions to assess applicants' nonacademic attributes in a structured and reliable manner. However, the development of high-quality MMI stations is resource intensive and dependent on expert input.

Objective: This study explored the utility of artificial intelligence (AI) in the generation of MMI stations for the Direct and Graduate Entry Medicine Program admissions process for domestic applicants at Monash Medical School. To our knowledge, this study represents the first empirical evaluation of AI-generated MMI stations deployed in a real-world medical school admissions context.

Methods: A total of 56 MMI stations from the 2025 admissions cycle were evaluated, including 17 (30.4%) AI-generated and 39 (69.6%) traditionally developed stations, administered across 824 domestic applicants for a total of 4897 applicant-station interactions. We assessed station quality through both reliability (using Cronbach α to examine internal consistency) and discrimination capability (using SD and range of scores) at the station level.

Results: AI-generated stations exhibited slightly higher reliability (α=0.82) compared with traditional stations (α=0.81), though this difference was not statistically significant (*P*=.91). Both AI-generated and traditionally developed stations demonstrated variable discrimination capability, with some stations from each development method showing excellent combinations of high reliability and strong discriminatory power, while others exhibited ceiling effects that limited their discriminatory power. Of note, a greater proportion of AI-generated stations were classified as optimal (α>0.85), and a smaller proportion were classified in the review category (α<0.75), compared with traditional stations. These results suggest that AI-generated stations can achieve psychometric performance comparable to traditionally developed stations.

Conclusions: Our findings highlight the utility of AI as a useful tool for MMI station generation, offering a scalable approach that may reduce the resource burden on faculty while maintaining or enhancing psychometric quality for applicants. Ongoing quality assurance and evaluation remain essential to ensure fairness and validity across the admissions process.

(JMIR Med Educ 2026;12:e86208) doi:10.2196/86208

KEYWORDS

artificial intelligence; AI; medical school admissions; medical education; multiple mini interviews; MMI; assessment reliability; educational technology; generative artificial intelligence; generative AI; psychometric evaluation; medical education innovation

Introduction

Medical school admissions are evolving to meet the demands of a changing health care landscape. With increasing competition and a growing emphasis on nonacademic traits that predict professional competence and patient care quality, new tools are reshaping how candidates are selected. The multiple mini interview (MMI) format has been widely adopted in medical school admissions and can assess a range of nonacademic skills, such as communication, ethical reasoning, teamwork, and empathy [1-5]. MMI stations typically consist of a stem or prompt that outlines the scenario, followed by several structured questions that candidates must respond to in real time and that are designed to probe different aspects of the candidate's

XSL•FO
RenderX

reasoning or behavior, testing their abilities to think critically, demonstrate perspective and emotional intelligence, and navigate complex interpersonal dynamics [1,4,5].

Despite their widespread adoption, the development of high-quality MMI stations remains challenging. The intense competition for limited places requires a large and continually refreshed pool of MMI stations to maintain test security and ensure fairness for all applicants. Developing and renewing this pool through traditional methods is resource intensive and places a considerable burden on faculty, while rushed or inconsistent development across multiple contributors can compromise station quality and reliability [6]. Artificial intelligence (AI) offers a potential solution to these challenges by streamlining station development and supporting the creation of a more reliable and scalable pool of assessment tools [7,8]. Prior work has shown that AI can generate assessment items or scenarios with acceptable quality or psychometric properties [7,9]. However, to date, no published studies have systematically evaluated the effectiveness of AI-generated MMI stations in real-world admissions settings.

In this study, we performed a comprehensive reliability analysis on 17 new MMI stations generated with the assistance of AI (AI-generated stations) and 39 existing MMI stations developed from non-AI sources used in the selection process for the 2025 intake of domestic students into the Direct Entry and Graduate Entry Medicine Program at Monash University. Using Cronbach α to assess internal consistency and SD as well as range of scores to evaluate discriminative capability, our analysis aimed to explore the reliability and overall quality of individual stations, uncover differences in performance between AI-generated and traditional stations, and identify stations that warrant modification, replacement, or serve as models for future development. These insights can be used to guide evidence-based improvements in MMI design and the integration of AI in medical admissions.

Methods

Study Design

This study analyzed 56 MMI stations used in the selection process for the 2025 admissions cycle of domestic applicants into the Direct Entry and Graduate Entry Medicine Program at Monash University. The 56 stations were divided into two categories: (1) AI-generated stations (n=17, 30.4%), and (2) existing or traditional stations (n=39, 69.6%). Each station was mapped to one of 7 MMI domains: advocacy, collaboration, critical thinking, empathy, ethical reasoning, motivation, and resilience. These domains map to desirable qualities of successful applicants and are used to guide the focus of the stations. They are not statistically determined discrete constructs within the MMI.

Traditional MMI stations were existing scenarios previously used in admissions and developed through established faculty-led processes. AI-generated MMI stations were developed using Claude 3.5 Sonnet (Anthropic), a proprietary, closed-source AI platform. Claude was selected for this study primarily because of its data privacy policy at the time of the project, which specified that user inputs were not retained or used for model training. Given that examination materials were included in prompts, this provided additional assurance regarding the protection of assessment content and institutional data.

The AI platform was provided with the following user-generated prompt:

> *Please generate an interview based on the provided document. The interviews are for high-school students who are applying for medical school and test the student's capacity for critical thinking. The interview structure should have a single paragraph for the applicant scenario, a single paragraph describing key components of the scenario for interviewers and 5 questions, with dot points on what an excellent response should look like.*

We designed this prompt to elicit realistic, contextually appropriate, and assessment-relevant content for medical school admissions, and guided the AI in producing full scenarios, associated interview questions, and example candidate responses.

To ensure content validity, relevance, and alignment with assessment objectives, all AI-generated scenarios underwent a structured quality assurance process. A working group of academic staff involved in admissions across both the Australian and Malaysian campuses reviewed the AI-generated stations as part of this process. Each scenario was independently reviewed by 2 faculty members from this working group with expertise in medical education and admissions. Reviewers evaluated alignment with the intended competency domain, clarity of the scenario, and potential sources of bias or ambiguity. Feedback from these reviews was incorporated through iterative refinement of prompts and station content. Final approval was contingent upon a consensus decision made jointly by 2 senior academic leads, providing an additional layer of oversight and quality control. While this process mitigates some risks inherent to AI-generated content, it does not eliminate potential biases associated with proprietary large language models, which lack transparency regarding training data and internal parameters [10]. Example traditional and AI-generated MMI stations from the collaboration domain, including scenario prompts and structured questions, are provided in Multimedia Appendix 1.

Data Collection

Data were obtained from MMIs conducted between September 2024 and January 2025 on 824 domestic applicants, yielding a total of 4897 applicant-station interactions. Applicants were randomly assigned to a subset of the 56 available stations (Multimedia Appendix 2), which were distributed across multiple interview days. Each MMI station was scored on a 20-point scale, consisting of 5 questions with each question scored from 0 to 4 points.

A total of 252 interviewers participated in the interview process across both direct and graduate entry admissions. All interviewers completed standardized training prior to the MMI cycle. Interviewers were drawn from a range of backgrounds including health professionals, academics, community

XSL•FO
RenderX

representatives, and occasionally senior students or recent graduates. Interviewers were recruited through an expression-of-interest process and screened for eligibility and potential conflicts of interest. Training included orientation to the MMI process, a scoring framework, and interviewer responsibilities through preparatory materials and live workshops. On each interview day, interviewers also participated in station-specific calibration sessions with other assessors assigned to the same station to promote consistent interpretation of scoring criteria.

Each candidate was assessed by a single trained interviewer at each MMI station. Multiple interviewers were assigned to each station across interview sessions, with the number of interviewers per station ranging from 2 to 17 depending on candidate volume per station. As each performance was rated by a single interviewer, interinterviewer reliability and interviewer variance were not examined in this analysis. To mitigate the risk of content dissemination between applicants, different stations were scheduled on different days.

Reliability Analysis

To assess the reliability of individual MMI stations, we calculated Cronbach α for each station using applicant responses from the 2025 admissions cycle. Cronbach α ranges from 0 to 1 and is a measure of internal consistency; it reflects the extent to which items within an assessment measure the same underlying construct or set of related competencies [11-13]. In the context of a 5-item MMI station, higher α values indicate greater coherence among the scored questions, whereas lower values may suggest item misalignment, scoring inconsistency, or other sources of measurement error. Because α was calculated within each station in this study, it reflects the reliability of that station's scoring rubric rather than the reliability of the overall MMI decision.

The categorization of Cronbach α scores for evaluating reliability remains a subject of debate, with many experts agreeing that thresholds should be determined by the purpose and stakes of the assessment. As Downing [13] noted, high-stakes assessments, such as licensure or certification examinations in medicine (which carry significant consequences for both candidates and the public), are generally expected to demonstrate reliability coefficients of 0.90 or above [13]. For assessments of moderate stakes, such as major summative examinations in medical school, a minimum reliability threshold of 0.80 to 0.89 is typically considered acceptable [12,13]. Lower-stakes assessments, including formative evaluations or internally developed tests used in teaching, may warrant reliability in the range of 0.70 to 0.79 [12,13]. Medical school admissions processes would generally be considered moderate- to high-stakes assessments, given their significant impact on applicants' future careers and the responsibility of selecting candidates suited for a profession with substantial societal consequences. However, station-level reliability estimates in MMI formats must also be interpreted in light of the short rating scales used to score each station (5 items) and the multidimensional nature of competencies assessed within individual stations.

In line with this framework, we classified stations in this study into 3 conservative reliability groups as outlined in Table 1. These categories were used to support descriptive comparison and structured quality assurance rather than as binary standards of station usability. In this context, both optimal and acceptable meet expected reliability, while review indicates a station requires further examination.

Table . Reliability classification of multiple mini interview stations based on Cronbach α scores.

Reliability category[a]	Cronbach α
Optimal	≥0.85
Acceptable	0.75 - 0.84
Review	<0.75

[a]Reliability categories were used for descriptive comparison and structured quality assurance review and do not represent binary standards of station usability.

Discrimination Ability Analysis

In addition to reliability, we also evaluated station quality through discrimination capability, which was measured using the mean score, SD, and observed score range for each station. Discrimination refers to a station's ability to distinguish or differentiate between candidates of varying competency levels across the scoring spectrum. Stations with adequate discrimination capability demonstrate higher SDs and broader score ranges, indicating they can effectively separate high-performing candidates from those with lower performance. Conversely, stations with poor discrimination capability exhibit low SDs or restricted score ranges, often due to ceiling effects (where most candidates score near the maximum) or floor effects (where most candidates score near the minimum). Such restricted score distributions limit a station's utility for selection decisions, as they fail to provide meaningful discrimination between candidates regardless of their actual competency levels. While reliability ensures measurement consistency, discrimination capability determines whether that consistent measurement provides useful information for distinguishing candidate performance.

Reliability and discrimination metrics were analyzed as distinct psychometric properties but are reported together where appropriate to support integrated evaluation of station quality.

Statistical Analysis

All statistical analyses were performed using R (version 4.5.1; R Foundation for Statistical Computing), primarily using the psych package for reliability analysis (Revelle, 2025; version

XSL•FO
RenderX

2.5.6) [14]. Figures were generated using GraphPad Prism (version 10.4.1; Dotmatics).

Ethical Considerations

This study was conducted within the context of the standard medical school admissions process and aligned with routine quality assurance and program evaluation activities. Ethics approval was obtained from the Monash University Human Research Ethics Committee (49569). All applicant data were fully deidentified prior to analysis to protect participant confidentiality. As the study involved secondary use of deidentified administrative data, no additional consent from applicants was required.

Results

Overview

An overview of the internal consistency reliability assessed using Cronbach α and average scores, including the SD and range, for all 56 MMI stations is presented in Table 2. The following subsections outline key patterns and comparisons observed across station types and domains.

XSL•FO
RenderX

Table . Applicant scores and Cronbach α reliability rating for each of the 56 multiple mini interview stations used in the 2025 intake of domestic students at Monash University[a].

Station code	Station type	Score, mean (SD)	Range	Cronbach α	Category
ADV_1	Existing	16.1 (3.0)	9-20	0.83	Acceptable
ADV_7	Existing	14.8 (3.2)	9-20	0.82	Acceptable
ADV_11	Existing	15.7 (2.8)	10-20	0.78	Acceptable
ADV_12	Existing	15.8 (2.8)	9-20	0.77	Acceptable
ADV_13[b]	Existing	14.2 (3.2)	6-20	0.85	Optimal
ADV_15	Existing	14.7 (2.9)	7-20	0.79	Acceptable
ADV_17	Existing	15.8 (3.3)	5-20	0.88	Optimal
ADV_18	Existing	12.6 (4.1)	6-20	0.89	Optimal
ADV_19	AI[c] generated	11.2 (4.0)	0-19	0.85	Optimal
ADV_20	AI generated	14.5 (3.2)	7-20	0.78	Acceptable
ADV_21[b]	AI generated	17.3 (2.3)	14-20	0.75	Acceptable
COLL_1	Existing	15.8 (2.9)	9-20	0.83	Acceptable
COLL_2	Existing	15.6 (3.0)	8-20	0.76	Acceptable
COLL_3[b]	Existing	16.0 (3.1)	8-20	0.84	Acceptable
COLL_4[b]	Existing	15.6 (3.2)	8-20	0.82	Acceptable
COLL_9	Existing	16.0 (2.7)	10-20	0.77	Acceptable
COLL_10	Existing	16.0 (3.3)	0-20	0.87	Optimal
COLL_11	Existing	15.6 (3.2)	4-20	0.83	Acceptable
COLL_14[b]	Existing	17.3 (1.6)	15-20	0.39	Review
COLL_16	AI generated	14.7 (2.6)	8-20	0.76	Acceptable
CT_8.1	Existing	16.0 (3.2)	9-20	0.84	Acceptable
CT_9[b]	Existing	14.1 (3.9)	5-19	0.85	Optimal
CT_10	Existing	14.0 (3.5)	3-20	0.83	Acceptable
CT_13	Existing	14.9 (3.7)	4-20	0.88	Optimal
CT_16	Existing	15.1 (3.5)	7-20	0.88	Optimal
CT_27	AI generated	14.4 (3.4)	8-20	0.85	Optimal
CT_28	AI generated	14.2 (3.7)	1-20	0.88	Optimal
CT_29	AI generated	12.4 (3.4)	0-19	0.84	Acceptable
EMP_3[b]	Existing	15.3 (3.0)	8-20	0.77	Acceptable
EMP_6	Existing	14.2 (4.5)	3-20	0.93	Optimal
EMP_10	Existing	15.0 (3.4)	6-20	0.80	Acceptable
EMP_14	Existing	12.3 (4.0)	4-20	0.87	Optimal
EMP_15	Existing	16.0 (2.5)	9-20	0.73	Review
EMP_17	AI generated	13.8 (3.9)	2-20	0.90	Optimal
EMP_18	AI generated	15.7 (2.9)	6-20	0.82	Acceptable
EMP_19[b]	AI generated	13.2 (3.1)	6-20	0.82	Acceptable
ER_2	Existing	15.4 (3.2)	8-20	0.86	Optimal
ER_3	Existing	15.7 (2.8)	7-20	0.77	Acceptable
ER_6.1	Existing	14.2 (4.0)	5-20	0.88	Optimal

XSL•FO
RenderX

Station code	Station type	Score, mean (SD)	Range	Cronbach α	Category
ER_16[b]	Existing	14.7 (3.6)	7-20	0.81	Acceptable
ER_24[b]	Existing	15.4 (2.6)	9-18	0.74	Review
ER_26	AI generated	15.9 (2.7)	8-20	0.69	Review
ER_27	AI generated	16.7 (2.5)	10-20	0.76	Acceptable
ER_28[b]	AI generated	13.9 (3.5)	6-20	0.86	Optimal
ER_29[b]	AI generated	14.4 (3.5)	7-19	0.86	Optimal
ER_30	AI generated	14.4 (3.5)	6-20	0.89	Optimal
MOT_2	Existing	14.8 (2.5)	6-19	0.66	Review
MOT_3	Existing	16.3 (3.5)	6-20	0.88	Optimal
MOT_4	Existing	15.5 (3.2)	5-20	0.84	Acceptable
MOT_6	Existing	17.0 (2.4)	11-20	0.76	Acceptable
MOT_9	Existing	15.2 (3.5)	6-20	0.87	Optimal
MOT_11[b]	Existing	14.6 (2.7)	7-20	0.69	Review
MOT_12[b]	AI generated	17.6 (2.1)	13-20	0.79	Acceptable
RES_1	Existing	15.4 (2.9)	9-20	0.84	Acceptable
RES_5	Existing	16.2 (2.9)	7-20	0.81	Acceptable
RES_6	AI generated	16.0 (3.0)	820	0.81	Acceptable

[a]Multiple mini interview stations were categorized into the following reliability categories, as outlined in Table 1: optimal (α≥0.85), acceptable (α=0.75-0.84), and review (α<0.75).

[b]Indicates stations with fewer than 50 candidate responses. Cronbach α estimates are generally less stable in small samples, and results for these stations should therefore be interpreted with caution.

[c]AI: artificial intelligence.

Overall Reliability of AI-Generated and Traditional Stations

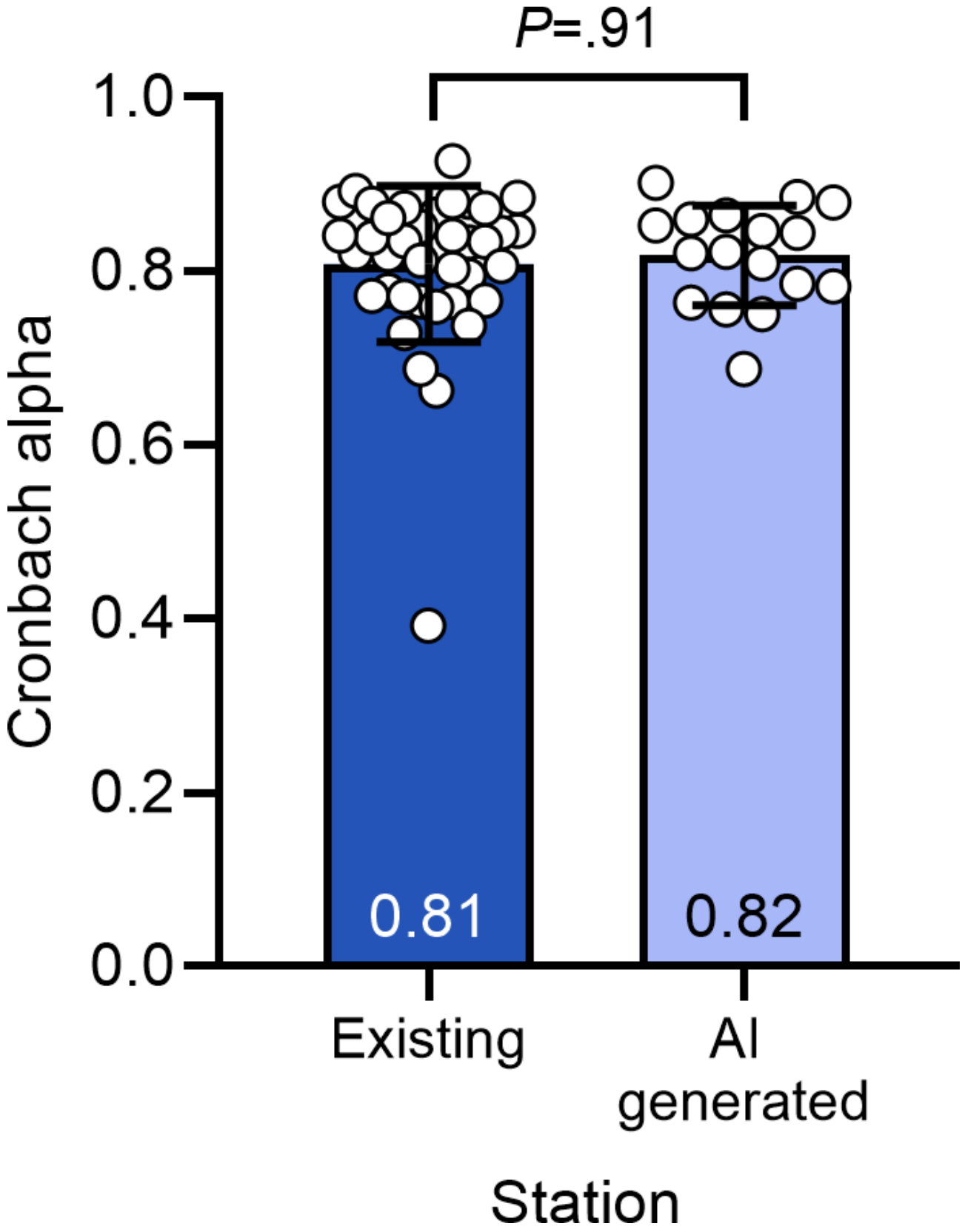
We conducted a Mann-Whitney *U* test to compare the internal consistency reliability of AI-generated and traditionally developed MMI stations. On average, the AI-generated stations demonstrated slightly higher Cronbach α values (mean α=0.82, SD 0.06) than the existing stations (mean α=0.81, SD 0.09). However, this difference was neither statistically significant (*U*=325; *z*=–0.116; *r*=−0.015; *P*=.91) nor was the difference large enough to be practically relevant (Figure 1). These findings suggest that while AI-generated stations may show marginally higher reliability, the overall level of internal consistency is comparable between the 2 station types.

In Figure 1, the mean Cronbach α values for AI-generated (17/56, 30%; mean α=0.82, SD 0.06) and traditional (39/56, 70%; mean α=0.81, SD 0.09) stations used in the 2025 Monash University medical school domestic admissions cycle. Bars represent mean (SD). *P* value was determined using a Mann-Whitney *U* test (*U*=325; *z*=−0.116; *r*=−0.015; *P*=.91).

When categorizing stations into review, acceptable, or optimal reliability groups (Table 1), AI-generated stations showed a slightly greater proportion achieving optimal reliability, with 41.2% (7/17; 95% CI 22% - 64%) of AI-generated stations meeting this threshold compared with 33.3% (13/39; 95% CI 21% - 49%) of the traditionally developed stations (Figure 2). Additionally, the proportion of AI-generated stations in the review category was 5.9% (1/17; 95% CI 1% - 27%) and was lower compared with traditionally developed stations of which 12.8% (5/39; 95% CI 6% - 27%) were in the review category (Figure 2). Most existing and AI-generated stations exhibited acceptable reliability, comprising 53.8% (21/39; 95% CI 39% - 68%) of traditionally developed stations and 52.9% (9/17; 95% CI 31% - 74%) of AI-generated stations (Figure 2). However, CIs were wide and substantially overlapping across categories, indicating that these proportional differences should be interpreted with caution.

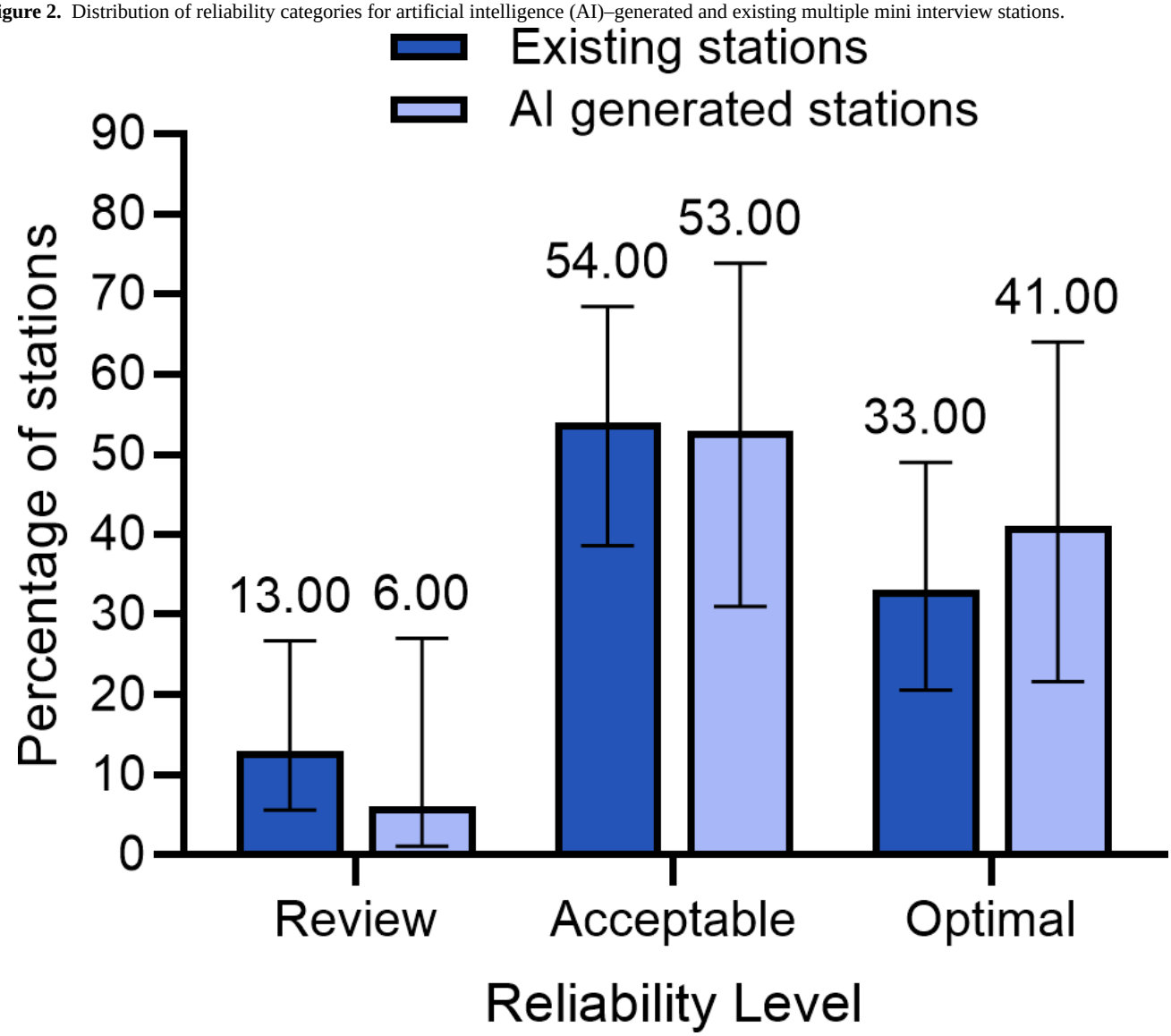
XSL•FO
RenderX

Figure 1. Comparison of internal consistency reliability (Cronbach α) between artificial intelligence (AI)–generated and traditionally developed multiple mini interview stations.



XSL•FO
RenderX

Figure 2. Distribution of reliability categories for artificial intelligence (AI)–generated and existing multiple mini interview stations.



To assess the influence of small sample sizes on station categorization, we conducted a sensitivity analysis excluding stations with fewer than 50 participants. When restricted to stations with n≥50, there remained no significant difference in internal consistency between AI-generated and traditionally developed stations (*P*=.89). The distribution of reliability categories was similar between groups, with 41% (7/17) of AI-generated and 37% (33/39) of traditionally developed stations classified as optimal, 53% (9/17) and 54% (21/39) classified as acceptable, and 6% (1/17) and 13% (5/39) classified as review, respectively. These findings indicate that category-based differences observed in the full dataset are sensitive to the inclusion of small-N stations. All subsequent analyses were conducted using the full dataset unless otherwise specified.

In Figure 2, the proportion of stations classified as review, acceptable, and optimal based on Cronbach α is shown for AI-generated (17/56, 30%) and existing or traditionally developed stations (39/56, 70%). Percentages were calculated within each station type, relative to the total number of stations in each category. Error bars represent 95% Wilson CIs for the proportion of stations in each reliability category.

To address potential confounding factors and ensure that apparent differences were not due to unequal distribution of AI-generated stations across MMI domains, we conducted a linear regression analysis. This analysis evaluated the association between station type and Cronbach α while controlling for MMI domain, essentially asking whether AI-generated stations would still show different reliability if both station types were equally represented across all domains. Although Cronbach α is bounded between 0 and 1, observed values in this dataset were not concentrated near the boundaries (range 0.39 - 0.93). Linear regression was therefore considered appropriate for estimating adjusted mean differences in α. The analysis showed that station type was not a significant predictor of reliability (β=−0.003; *P*=.90), indicating that the slight advantage observed for AI-generated stations was not attributable to systematic differences in station development approaches. A beta regression model was also fitted as a sensitivity analysis; this yielded the same substantive inference (*P*=.77).

The regression analysis also examined whether some domains were inherently more challenging for developing reliable stations, which could have explained apparent station type differences if AI-generated stations were concentrated in

XSL•FO
RenderX

naturally high-reliability domains. However, no significant differences in reliability were observed between domains (all *P*>.10), indicating that critical thinking, empathy, advocacy, and other domains showed similar reliability levels regardless of station type.

Finally, we performed an interaction model to test whether AI-generated stations might have domain-specific strengths or weaknesses compared with traditional stations; for example, whether AI might excel at generating critical thinking stations but struggle with empathy scenarios. This analysis found no significant interactions (all *P*>.48), demonstrating that the relationship between station type and reliability remained consistent across all MMI domains.

Collectively, these analyses demonstrate that while AI-generated stations showed slightly higher descriptive reliability metrics across multiple measures, these differences were not statistically significant. The findings suggest that AI-generated station development produces reliability outcomes equivalent to traditional faculty-led development methods, with performance remaining consistent across different MMI domains.

Station Quality Analysis: Reliability and Discrimination Ability

Analysis of individual station performance revealed notable variation in both reliability and discrimination characteristics across AI-generated and traditionally developed stations. Several stations demonstrated exceptional reliability, with Cronbach α values exceeding 0.90 (Table 2). Notably, EMP_17 (α=0.90), an AI-generated station, and EMP_6 (α=0.93), a traditionally developed station, exhibited the highest internal consistency. Furthermore, both EMP_17 (mean 13.8, SD 3.9; range 2 - 20) and EMP_6 (mean 14.2, SD 4.5; range 3 - 20) stations demonstrated excellent discrimination capabilities. Other stations showing this desirable combination of high reliability and strong discrimination included ADV_18 (α=0.8925; mean 12.6, SD 4.1) and ER_6.1 (α=0.88; mean 14.2, SD 4.0). These stations may serve as exemplars for future MMI station development. Their structure was characterized by clear and focused prompts, as well as well-defined evaluation criteria, contributing to consistent assessment across different evaluators. The high reliability and discriminatory ability of these stations suggested they were effective in eliciting and measuring the targeted competencies in a robust and reproducible manner.

Conversely, some stations with acceptable reliability showed limited discrimination capability. For example, ADV_21 (α=.75; mean score 17.3, SD 2.3) and MOT_12 (α=.79; mean score 17.6, SD 2.1) exhibited ceiling effects, with most candidates scoring in the upper range of the scale, limiting their ability to distinguish between high-performing candidates (Table 2). While these stations demonstrated reasonable consistency, their restricted score distributions suggested they may be too easy or lack sufficient challenge to effectively discriminate between candidates of differing ability levels.

Notably, the traditionally developed station COLL_14 demonstrated poor performance across both metrics, with critically low reliability at a Cronbach α of 0.3925, well below the acceptable threshold, and poor discrimination (mean 17.3, SD 1.6; range 15 - 20), with scores clustering near the maximum possible score (Table 2). Although this station had a limited sample size (18/824, 2%), the combination of poor reliability and severely restricted score range suggested fundamental issues with station design or implementation that warranted complete revision or replacement.

In comparison, the lowest-performing AI-generated station, ER_26, showed suboptimal reliability with a Cronbach α of 0.6876, but maintained reasonable discrimination (mean 15.8, SD 2.7; range 8 - 20), suggesting that while this station requires refinement for internal consistency, it retains adequate discriminative function across the scoring range.

It should be noted that stations ADV_21, MOT_12, and COLL_14 had fewer than 50 candidate responses (Multimedia Appendix 2), which can result in less stable reliability estimates. Thus, station-level Cronbach α values for these stations should be interpreted with caution. Nevertheless, taken together, these results demonstrate that both AI-generated and traditional development methods are capable of producing stations across the full spectrum of psychometric quality, from exemplary stations that excel in both reliability and discrimination to problematic stations requiring revision or replacement. The findings suggest that station effectiveness depends primarily on individual design characteristics and implementation rather than the development methodology used.

Discussion

Principal Findings

This study compared the psychometric performance of AI-generated and traditionally developed MMI stations used in a high-stakes medical school admissions process. AI-generated stations demonstrated internal consistency reliability comparable to traditional stations, with no statistically significant differences in Cronbach α. A greater proportion of AI-generated stations were classified as optimal for internal consistency, while fewer fell within the review category. Both station types showed variable discrimination ability, indicating that overall station quality depended on individual design characteristics rather than development method. Taken together, these findings support psychometric equivalence between AI-generated and traditionally developed MMI stations rather than evidence of superiority.

Implications for the Use of AI in Medical School Admissions

AI applications in medical education have shown promising results across various domains, including automated scoring of clinical examinations, natural language processing for medical curriculum development, and generation of multiple-choice questions for medical licensing examinations [15,16]. Similarly, AI has been increasingly integrated into admissions processes across professional programs, with applications ranging from automated essay scoring in standardized tests to predictive modeling of student success [17,18]. This study extends this emerging body of work by demonstrating that AI can produce psychometrically sound assessment content for high-stakes medical school selection processes. Unlike much of the existing

XSL•FO
RenderX

literature, which focuses on AI for prediction or automated scoring of applicant performance, this study examines the use of AI for content generation, specifically the development of MMI stations themselves.

Our results showed that AI-generated stations demonstrated comparable internal consistency to traditionally developed stations (mean α=0.82, SD 0.06 vs mean α=0.81, SD 0.09; *P*=.91). Of note, a greater proportion of AI-generated stations were classified as optimal ($\alpha \geq 0.85$), while fewer were classified in the review category ($\alpha < 0.75$) compared with traditional stations. However, this result is merely descriptive and should be interpreted with caution. The CIs around these proportions were wide and substantially overlapping, reflecting the small number of stations and indicating that the apparent differences are not statistically or practically robust. Nevertheless, these findings highlight the potential value of incorporating AI into admissions processes and provide a basis for examining how such tools might be further optimized and integrated into future selection frameworks.

Our findings align with broader research on AI-generated assessment content, which has consistently shown that the quality of AI-produced items depends more on human input in aspects including prompt engineering, content validation processes, and postgeneration refinement than on the AI system itself [19,20]. Studies comparing AI-generated and human-developed test items in educational contexts have similarly found equivalent psychometric properties when appropriate quality assurance measures are implemented [21,22]. This suggests that the observed variability in both AI-generated and traditional MMI stations reflects fundamental challenges in assessment development rather than limitations specific to either approach.

Analysis of discrimination capability revealed that both AI-generated and traditionally developed stations demonstrated variable performance in their ability to distinguish between candidates of different competency levels. Some stations from each development method exhibited excellent combinations of high reliability and strong discrimination, including AI-generated station EMP_17 and traditionally developed station EMP_6. Conversely, other stations from both approaches showed acceptable reliability but limited discrimination due to ceiling effects, where most candidates scored in the upper range of the scale. This variability suggests that station effectiveness depended primarily on individual design characteristics and implementation quality rather than the development methodology used.

Taken together, these findings underscore that while the AI-generated MMI development process shows great promise, it should be viewed as a complementary tool within a broader, evidence-based framework for admissions assessment rather than a replacement for expert oversight and continual quality assurance. Our observation that both development methods can produce stations across the full spectrum of psychometric quality emphasizes that ongoing review, calibration, and refinement remain essential for ensuring that all stations, irrespective of their origin, meet the rigorous standards necessary for fair and effective candidate evaluation.

A notable limitation of this analysis was the small sample size for certain stations ($n < 50$), particularly for some of the new AI-generated stations. Smaller sample sizes can result in less stable reliability estimates, necessitating cautious interpretation of individual station-level Cronbach α values when analyzing the performance of specific stations. Although sensitivity analyses excluding small sample size ($n < 50$) stations attenuated category-based differences, future reliability assessments should aim for minimum sample sizes of $n \geq 50$ per station to ensure more robust and stable estimates.

Nevertheless, the implications of these findings are particularly significant when considered in the context of the overall MMI station development process, which has traditionally been time-consuming, resource intensive, and reliant on substantial faculty input. Crafting high-quality stations requires not only content expertise but also careful attention to fairness, clarity, and alignment with assessment objectives; factors that can introduce variability and potential bias. The fact that AI-generated stations performed comparably to, and in some cases more reliably and with better discrimination ability than, traditionally developed stations suggests that AI may offer a scalable and efficient alternative for generating high-quality assessment scenarios. By streamlining the development process and reducing reliance on individual judgment, AI has the potential to support both the efficiency and equity of MMI design, while preserving the psychometric robustness of the assessment. Future research should formally evaluate whether AI-generated stations are truly noninferior to traditional stations regarding reliability, validity, and fairness, ideally through prospective, multi-institutional studies with larger sample sizes.

Recommendations for Using AI in Future MMI-Based Admissions Processes

In this study, we have identified several actionable strategies to use AI to achieve quality and consistency in MMI station development. Drawing on the observed reliability data from both AI-generated and traditionally developed stations, the following recommendations are intended to guide future improvements in MMI design and implementation.

The continued use of AI-generated station development is supported by the data. AI-generated stations demonstrated excellent internal consistency and discrimination capability, indicating that this approach can produce high-quality assessment tools. Sustaining and expanding the use of AI in MMI design may enhance the efficiency and reliability of future station development while maintaining psychometric quality.

To preserve the psychometric integrity of the MMI system over time, a formal process for routine monitoring of both reliability and discrimination ability should be established. This process should involve the regular calculation of Cronbach α alongside assessment of score distributions and discrimination metrics, particularly for newly introduced or modified stations. Systematic monitoring will enable the timely identification of underperforming stations and facilitate data-informed decisions regarding their revision or replacement.

We further recommend standardization in station design. High-performing stations (eg, those with $\alpha > 0.85$ and high SDs)

XSL•FO
RenderX

should be used as templates to guide the development of future scenarios. High-performing stations that demonstrate excellent reliability combined with strong discrimination capability (such as EMP_17 and EMP_6 in this study) should be used as templates to guide the development of future scenarios. Consistency in station structure, response criteria, and intended competencies is likely to improve fairness and reduce variability across MMI stations.

Stations that demonstrate below acceptable reliability ($\alpha<0.70$) or limited discrimination capability should be prioritized for immediate review. This includes stations with low Cronbach α values (such as COLL_14) as well as those exhibiting ceiling effects that limit their discriminatory power (such as ADV_21 and MOT_12). Where appropriate, these stations should be revised or removed from the selection process. Additionally, for stations with limited response data ($n<50$), further data collection is necessary to obtain more stable reliability estimates and support evidence-based decision-making.

Finally, as the role of AI in MMI development expands, continual refinement with contemporary AI models will be essential. Improving the quality and contextual alignment of AI-generated content will require sustained collaboration between AI researchers and medical education professionals. This interdisciplinary approach will be critical to ensuring that AI tools not only support psychometric reliability but also align with the pedagogical and professional goals of medical school admissions.

Conclusions

In this paper, we provide, to our knowledge, the first empirical evaluation of AI-generated MMI stations implemented in a real-world medical school admissions process, offering evidence supporting the use of AI-assisted methods in MMI station development. AI-generated stations demonstrated comparable reliability and discrimination capability relative to traditionally, entirely human-developed stations, with a higher proportion achieving optimal internal consistency and fewer falling below acceptable thresholds. Given the time- and resource-intensive nature of traditional station development, our findings suggest that AI may offer a scalable, efficient, and psychometrically sound alternative.

However, the observed variability in reliability and discrimination ability across both AI-generated and traditional stations underscores the necessity of ongoing quality assurance. Regular psychometric evaluation and iterative refinement remain essential to maintaining the validity and fairness of the MMI process. Looking ahead, the integration of AI into medical school admissions may extend beyond station generation to include broader applications such as predictive modeling of applicant success. Nonetheless, any expansion of AI use must be guided by careful consideration of ethical concerns, including transparency, algorithmic bias, and equity, to ensure that admissions systems remain fair and inclusive.

Acknowledgments

The authors thank the admissions and administrative teams involved in the delivery of the 2025 multiple mini interview process.

Funding

No external funding was received for this study.

Data Availability

The datasets generated or analyzed during this study are available from the corresponding author on reasonable request.

Authors' Contributions

ML, WH, and SB conceived and designed the study. KC provided advice on artificial intelligence (AI) tools and prompt engineering used in the development of AI-generated stations. KB collected and curated the dataset. KB, SH, and MNH conducted the data analyses. SH drafted the manuscript with input from all authors. All authors contributed to the interpretation of the findings, critically revised the manuscript, and approved the final version.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Example traditional and artificial intelligence–generated multiple mini interview stations from the collaboration domain, including scenario prompts and structured questions.
[DOCX File, 160 KB - mededu_v12i1e86208_app1.docx]

Multimedia Appendix 2
Detailed summary of multiple mini interview stations used in the 2025 Direct and Graduate Entry Medicine Program admissions cycle.

XSL•FO
RenderX

[DOCX File, 1680 KB - mededu_v12i1e86208_app2.docx]

References

1. Ayub R, Yousuf N, Asad M, Khan UA. Multiple mini interviews as a measure of non-cognitive skills for admissions into undergraduate medical education programme in Pakistan: a validity study. J Pak Med Assoc 2017 Dec;67(12):1905-1909. [Medline: 29256539]
2. Eva KW, Reiter HI, Rosenfeld J, Norman GR. The ability of the multiple mini-interview to predict preclerkship performance in medical school. Acad Med 2004 Oct;79(10 Suppl):S40-S42. [doi: 10.1097/00001888-200410001-00012] [Medline: 15383385]
3. Eva KW, Rosenfeld J, Reiter HI, Norman GR. An admissions OSCE: the multiple mini-interview. Med Educ 2004 Mar;38(3):314-326. [doi: 10.1046/j.1365-2923.2004.01776.x] [Medline: 14996341]
4. Rees EL, Hawarden AW, Dent G, Hays R, Bates J, Hassell AB. Evidence regarding the utility of multiple mini-interview (MMI) for selection to undergraduate health programs: a BEME systematic review: BEME guide no. 37. Med Teach 2016 May;38(5):443-455. [doi: 10.3109/0142159X.2016.1158799] [Medline: 27050026]
5. Yusoff MS. Multiple mini interview as an admission tool in higher education: insights from a systematic review. J Taibah Univ Med Sci 2019 Jun;14(3):203-240. [doi: 10.1016/j.jtumed.2019.03.006] [Medline: 31435411]
6. Ali S, Sadiq Hashmi MS, Umair M, Beg MA, Huda N. Multiple mini-interviews: current perspectives on utility and limitations. Adv Med Educ Pract 2019;10:1031-1038. [doi: 10.2147/AMEP.S181332] [Medline: 31849557]
7. Gordon M, Daniel M, Ajiboye A, et al. A scoping review of artificial intelligence in medical education: BEME guide no. 84. Med Teach 2024 Apr;46(4):446-470. [doi: 10.1080/0142159X.2024.2314198] [Medline: 38423127]
8. Keir G, Hu W, Filippi CG, Ellenbogen L, Woldenberg R. Using artificial intelligence in medical school admissions screening to decrease inter- and intra-observer variability. JAMIA Open 2023;6(1):ooad011. [doi: 10.1093/jamiaopen/ooad011] [Medline: 36819893]
9. Li R, Wu T. Application of artificial intelligence generated content in medical examinations. Adv Med Educ Pract 2025;16:331-339. [doi: 10.2147/AMEP.S492895] [Medline: 40026780]
10. Gallegos IO, Rossi RA, Barrow J, et al. Bias and fairness in large language models: a survey. Comput Linguist 2024 Sep;50(3):1097-1179. [doi: 10.1162/coli_a_00524]
11. Cronbach LJ, Warrington WG. Time-limit tests: estimating their reliability and degree of speeding. Psychometrika 1951 Jun;16(2):167-188. [doi: 10.1007/BF02289113] [Medline: 14844557]
12. Tavakol M, Dennick R. Making sense of Cronbach's alpha. Int J Med Educ 2011 Jun 27;2:53-55. [doi: 10.5116/ijme.4dfb.8dfd] [Medline: 28029643]
13. Downing SM. Reliability: on the reproducibility of assessment data. Med Educ 2004 Sep;38(9):1006-1012. [doi: 10.1111/j.1365-2929.2004.01932.x] [Medline: 15327684]
14. Revelle W. Psych: procedures for psychological, psychometric, and personality research. The Comprehensive R Archive Network. 2025. URL: https://CRAN.R-project.org/package=psych [accessed 2026-05-28]
15. Narayanan S, Ramakrishnan R, Durairaj E, Das A. Artificial intelligence revolutionizing the field of medical education. Cureus 2023 Nov;15(11):e49604. [doi: 10.7759/cureus.49604] [Medline: 38161821]
16. Rincón EH, Jimenez D, Aguilar LA, Flórez JM, Tapia Á, Peñuela CL. Mapping the use of artificial intelligence in medical education: a scoping review. BMC Med Educ 2025 Apr 12;25(1):526. [doi: 10.1186/s12909-025-07089-8] [Medline: 40221725]
17. Lira B, Gardner M, Quirk A, et al. Using artificial intelligence to assess personal qualities in college admissions. Sci Adv 2023 Oct 13;9(41):eadg9405. [doi: 10.1126/sciadv.adg9405] [Medline: 37824610]
18. Van Busum K, Fang S. Analysis of AI models for student admissions: a case study. In: SAC '23: Proceedings of the 38th ACM/SIGAPP Symposium on Applied Computing: Association for Computing Machinery; 2023:17-22.
19. Nikolovski V, Trajanov D, Chorbev I. Advancing AI in higher education: a comparative study of large language model-based agents for exam question generation, improvement, and evaluation. Algorithms 2025;18(3):144. [doi: 10.3390/a18030144]
20. Bhandari S, Liu Y, Kwak Y, Pardos ZA. Evaluating the psychometric properties of ChatGPT-generated questions. Comput Educ Artif Intell 2024 Dec;7:100284. [doi: 10.1016/j.caeai.2024.100284]
21. O KM. A comparative study of AI-human-made and human-made test forms for a university TESOL theory course. Lang Test Asia 2024;14:19. [doi: 10.1186/s40468-024-00291-3]
22. Westacott R, Badger K, Kluth D, Gurnell M, Reed MW, Sam AH. Automated Item Generation: impact of item variants on performance and standard setting. BMC Med Educ 2023 Sep 11;23(1):659. [doi: 10.1186/s12909-023-04457-0] [Medline: 37697275]

Abbreviations

AI: artificial intelligence
MMI: multiple mini interview

XSL•FO
RenderX

Edited by M Montagna; submitted 20.Oct.2025; peer-reviewed by CY Su, R Batacan, S Mohanadas; revised version received 12.Mar.2026; accepted 07.Apr.2026; published 11.Jun.2026.

Please cite as:
Hamila S, Birchill K, Cao K, Hossain MN, Bullock S, Hodgson W, Harrison J, Leech M
AI-Generated Multiple Mini Interview (MMI) Stations for Medical School Admissions: Psychometric Evaluation
JMIR Med Educ 2026;12:e86208
URL: https://mededu.jmir.org/2026/1/e86208
doi:10.2196/86208

© Sabryn Hamila, Kyle Birchill, Khoa Cao, Md Nassif Hossain, Shane Bullock, Wayne Hodgson, Julia Harrison, Michelle Leech. Originally published in JMIR Medical Education (https://mededu.jmir.org), 11.Jun.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Education, is properly cited. The complete bibliographic information, a link to the original publication on https://mededu.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Health Professional Students' Use of Generative Artificial Intelligence During Clinical Placements: Cross-Sectional Online Survey Study

Sylvain Kotzki[1], PharmD, PhD; Calvin Massonnet Turner[1], MEng; Kim Gauthier[2], PhD; Mélanie Minoves[3,4], PharmD, PhD; Nicolas Vuillerme[1,5], PhD

[1]LIG SANGRIA, Grenoble INP, CNRS, Univ. Grenoble Alpes, Bureau 19, Centre de Recherche en Santé Intégrée (CReSI), Grenoble, France

[2]Faculty of Medicine, Univ. Grenoble Alpes, Grenoble, France

[3]HP2, Inserm, Univ. Grenoble Alpes, Grenoble, France

[4]Department of Pharmacy, CHU Grenoble Alpes, Grenoble, France

[5]Institut Universitaire de France, Paris, France

Corresponding Author:
Sylvain Kotzki, PharmD, PhD
LIG SANGRIA, Grenoble INP, CNRS, Univ. Grenoble Alpes, Bureau 19, Centre de Recherche en Santé Intégrée (CReSI), Grenoble, France

Abstract

Background: Generative artificial intelligence (GenAI) is rapidly expanding in higher education and clinical practice. However, its use during clinical placements, where cognitive demands and responsibility for patient care increase, remains insufficiently documented.

Objective: This study aimed to characterize self-reported GenAI use during clinical placements, perceived benefits and risks, and related training and governance needs.

Methods: We conducted a cross-sectional online survey at a French university (July 17 to September 30, 2025). Eligible participants were students in medicine, pharmacy, nursing, midwifery, or physiotherapy who were currently in, or had completed within the past 18 months, a clinical placement. A 61-item questionnaire (comprising closed- and open-ended items) assessed GenAI use, task patterns, perceived benefits or risks, and training or governance needs. A composite index classified self-perceived GenAI maturity as minimal, limited, moderate, or high. Group comparisons used χ^2 tests; maturity gradients used trend tests.

Results: A total of 388 students responded (n=308, 79.4% women), mainly nursing students (n=217, 55.9%). Overall, 204 (52.6%) students reported using GenAI during clinical placements. Use differed across disciplines (χ^2_4=10.71; *P*=.03), with lower uptake in midwifery (6/23, 26%; odds ratio 0.30, 95% CI 0.11 - 0.77). Adoption increased markedly with self-perceived maturity (minimal: 2/22, 9% vs high: 22/29, 76%; trend *P*<.001). Among the 204 users, the most commonly reported uses were information retrieval (n=159, 77.9%), bibliographic search (n=152, 74.5%), and translation or rephrasing (n=145, 71.1%); patient-facing activities were less frequently reported (eg, patient-document drafting or communication preparation: n=78, 38.2%). Although most users reported never entering direct patient identifiers, 48 (23.5%) reported at least 1 disclosure of patient-identifying information, and 96 (47.1%) reported processing real medical content perceived as anonymized. The most endorsed perceived benefits among the 388 students were documentation support (n=315, 81.2%) and improved access to information (n=266, 68.5%). The most endorsed risks were dependency (n=353, 90.9%), skill erosion (n=329, 84.8%), and confidentiality breaches (n=339, 87.4%). Training needs were highest for ethics or regulatory training (294/378, 77.7%) and a best-practice clinical guide (292/373, 78.3%).

Conclusions: GenAI is already used by a substantial proportion of French students in health professions during clinical placements, predominantly for information and documentation support rather than patient-facing activities. Self-perceived readiness is strongly associated with adoption. Reported disclosures and concurrent concerns about dependency, skill erosion, and confidentiality support the need for structured curricula and clear governance frameworks to enable responsible, patient-centered integration of GenAI into clinical education.

(JMIR Med Educ 2026;12:e85243) doi:10.2196/85243

KEYWORDS

generative artificial intelligence; AI use; artificial intelligence; health professional students; clinical placements; medical education; ethics

XSL•FO
RenderX

Introduction

Generative artificial intelligence (GenAI) has entered the global health landscape at an unprecedented pace. The widely cited example of ChatGPT (GPT 3.5), which in 2023 achieved passing scores on all 3 components of the United States Medical Licensing Examination, illustrates the capacity of large language models to engage with complex medical knowledge. Concurrently, adoption among health professional students has expanded rapidly. In the United States, just over half of medical students (52%) reported using ChatGPT for medical school-related tasks, based on a survey of 415 students, with about 1 in 6 doing so on a weekly basis [1]. In Israel, 86% of undergraduate students from medicine, nursing, and allied health professions reported being familiar with ChatGPT [2]. Across Europe, a recent survey of 487 medical students across Germany, Austria, and Switzerland revealed that 38.8% of them reported prior experience with artificial intelligence (AI) chatbots such as ChatGPT [3]. In nursing, an umbrella review emphasized both opportunities and substantial pedagogical, ethical, and organizational challenges in integrating AI into education and practice [4]. Consistently, a qualitative study among undergraduate nursing students found that while GenAI was valued as a supportive learning aid, students also expressed concerns over dependency, weakened critical thinking, and threats to academic integrity [5]. In Italy, nearly all physiotherapy students (≈95.3%) were aware of AI chatbots, though more than half reported never using them in academic settings [6].

From an educational perspective, 3 primary applications of GenAI have been identified: the generation and reformulation of academic content, instant adaptive tutoring, and participation in simulation or problem-based learning scenarios [7-9]. A recent systematic review further synthesized evidence across health professional education, confirming the growing influence of GenAI on student learning, particularly in acquisition, inquiry, practice, and production [10]. Importantly, this growing body of evidence extends to experimental research, with a randomized controlled trial in Norway reporting a modest improvement in knowledge test performance when pharmacy students had access to ChatGPT, although the effect did not reach statistical significance [11]. However, this rapid adoption is unfolding amid a marked institutional vacuum: more than 60% of surveyed pharmacy students reported no formal curricular exposure to GenAI [12], while in 2024, an analysis of 116 US research universities found that less than half provided classroom guidance and many still lacked formal institutional policies [13]. As a result, health professional education is evolving in a state of continuous adaptation, where the promising capabilities of GenAI coexist with substantial risks, including facilitated plagiarism, overreliance, diminished critical thinking, weakened communication, and threats to the development of clinical reasoning and problem-solving skills [14].

Beyond classroom-based learning, the training of health professionals also depends heavily on clinical practice, as their training does not take place solely in lecture halls but also through clinical placements. These clinical placements, while essential to professional development, place students in a dual role: learning in real-world conditions while gradually assuming responsibility for patient care. Such placements are frequently marked by heightened stress, significant emotional demands, feelings of inadequacy [15,16], and in some cases, early signs of burnout [17]. Meanwhile, in clinical practice, GenAI is increasingly being applied across health care settings such as hospitals, clinics, private practices, and pharmacies. Recent reports describe its use to accelerate documentation, simplify radiology reports, draft consultation letters, and even support aspects of diagnostic reasoning [18,19]. An integrative review in nursing similarly highlighted applications of AI across education, clinical practice, workload management, and professional perceptions, pointing to opportunities but also to enduring structural and ethical challenges [20]. In this environment, health professional students in clinical placements are likely to turn to GenAI to ease their daily workload, such as checking information, translating a consent form, or rephrasing a discharge summary. While such assistance can provide immediate support when human feedback is unavailable, it also raises concerns about overreliance, the weakening of critical learning processes, and the potential for errors.

A recent survey found that 1 in 5 UK general practitioners reported using GenAI, most commonly to generate documentation after patient encounters. Clinicians acknowledged its potential to reduce administrative burden but also raised concerns regarding patient safety, data privacy, and the risk of errors [21]. Other authors caution, however, that adoption may remain limited or prove transient in the absence of robust clinical validation [22]. Health professional students in clinical placements stand at the intersection of 3 evolving contexts: higher education, where GenAI is rapidly becoming mainstream; clinical environments, where health care professionals are beginning to adopt it; and a learning experience often marked by uncertainty, stress, and substantial cognitive demands. Yet, to the best of our knowledge, no survey has systematically examined how students use GenAI during clinical placements, what motivates their engagement, and the perceived benefits, risks, and governance needs associated with this pivotal stage of professional development.

To address this gap, we conducted an anonymous survey among 388 health professional students from multiple health care programs at the Université Grenoble Alpes (UGA, France), all of whom had recently completed or were currently engaged in clinical placements. The aim of this study was to characterize health professional students' self-reported use of generative AI in clinical situations. Specifically, we sought to (1) estimate adoption prevalence overall and by discipline, (2) describe the tasks, tools, and frequency of use, including self-reported handling of patient-related information, and (3) assess perceived benefits, risks, and training and governance needs to inform structured educational and institutional responses.

Methods

Study Design

We conducted a web-based descriptive cross-sectional study in the form of an online survey targeting health professional students at UGA (France). The aim was to document actual use,

XSL•FO
RenderX

perceived benefits and risks, as well as training needs related to GenAI in the context of clinical placements. Data collection took place between July 17 and September 30, 2025. This survey was carried out as part of a national initiative, funded by the French National Research Agency, to support the development and dissemination of digital health and AI in health education in French university health programs (France 2030 program, ANR n°23-CMAS-0035).

Ethical Considerations

This study has received ethical approval from the "Ethics Committee for the Integrity and Ethics of Research in Health Professions Education" of the "*Société Internationale Francophone d'Education Médicale*" (Notice n°1905 - 2025, issued July 17, 2025). Entry into the survey was conditional upon acceptance of an electronic informed consent form, accompanied by an information sheet detailing the study objectives, data confidentiality, voluntary nature of participation, right to withdraw, and potential risks. The survey was fully anonymous, and no incentives were offered for participation. Participation was voluntary, and informed consent was obtained electronically from all participants.

Recruitment Procedure

Following approval by the deans of the Faculties of Medicine and Pharmacy at UGA, an email invitation via the official academic mailing list was distributed to potential participants enrolled in medical or allied health training programs. Overall, approximately 4553 eligible students (n=2196 in medicine, n=283 in pharmacy, n=1904 in nursing, n=120 in midwifery, and n=50 in physiotherapy) were invited (see Multimedia Appendix 1 for details). In addition, a flyer was posted on student information boards at the entrance of lecture halls and classrooms throughout the recruitment period, ensuring a reinforced exposure to the survey. A total of 388 students completed the survey, corresponding to an overall response rate of 8.5%, ranging from 3.1% in medicine to 26% in physiotherapy. Interested students accessed the survey via a LimeSurvey link.

Clinical Placement Context

In the French health professions education system, clinical placements are mandatory components of curricula. Rotations typically last 2 to 10 weeks. Students progressively assume responsibilities according to their year of study, under the supervision of licensed health professionals from the relevant discipline, with day-to-day oversight and validation of key care and documentation tasks. Early-year students mainly observe and perform basic tasks, while senior students (in the final years of training) take a more active role in discipline-specific activities, including patient care tasks, documentation, information-seeking, care coordination, and patient communication or education.

Data Collection

We collected data through an anonymous online survey administered with LimeSurvey. Students interested in the study were invited to access the anonymous link, which contained 61 items: 18 single-choice questions, 38 single-choice questions on a 4-point Likert scale (1=strongly disagree to 4=strongly agree), 1 multiple-choice question, and 4 open-ended questions. The full survey questionnaire is available in Multimedia Appendix 1.

The items were developed based on 4 published studies: Blease et al [21] on generative GenAI in primary care, Lobet et al [23] on the spontaneous use of ChatGPT by university students, Karaca et al [24] on the Medical Artificial Intelligence Readiness Scale for Medical Students, and Boillat et al [25] on physician and student preparedness for AI. Items were then adapted by consensus to the clinical context.

We organized the instrument into 4 domains:

- Sociodemographic characteristics
- Declared uses of GenAI in clinical practice
- Perceptions and attitudes in clinical contexts
- Knowledge and training related to GenAI

We conducted this web-based survey in accordance with the CHERRIES (Checklist for Reporting Results of Internet E-Surveys) checklist (Checklist 1) [26]. The questionnaire was hosted on LimeSurvey and included a single skip logic based on declared GenAI use versus nonuse; all other sections were shown to all participants. A total of 38 items were mandatory; missing responses were otherwise allowed, and respondents could review or change answers before submission. Only fully submitted questionnaires were analyzed, and no technical measures (cookies, IP checks, or registration) were used to prevent multiple entries.

Content validity was assessed through expert review by 2 clinicians and health-professions education specialists (SK and MM), who evaluated item relevance, clarity, and coverage of the 4 intended domains. Minor wording changes were made to improve clarity and contextual alignment with clinical placements (4 items required rewording for better understanding; no items were removed or added). A pilot test with 10 students (n=4 students in medicine, n=3 students in pharmacy, and n=3 students in nursing) confirmed intelligibility. Feedback from experts and students was incorporated to produce the definitive version. The average completion time was 15 minutes.

GenAI Maturity Index

To capture students' self-perceived maturity toward GenAI, considered as their perceived readiness to understand, evaluate, and appropriately use GenAI in educational and professional contexts, we constructed a composite GenAI maturity index from 9 Likert-type items [24]. This approach follows previous work on AI readiness and familiarity in health professions, where multiple items are aggregated into a single readiness or familiarity score [24,25]. One item assessed general digital comfort ("I feel comfortable with digital technologies in general"), and 8 items targeted GenAI-specific knowledge and self-efficacy (basic GenAI concepts, distinction between GenAI and other AI, ability to judge the reliability of GenAI outputs, awareness of appropriate use contexts and current limitations, and perceived ability to use GenAI in one's education and future practice (full wording in Multimedia Appendix 1).

All 9 items used a 4-point agreement scale (1=strongly disagree, 4=strongly agree). To respect the ordinal nature of the 4-point

XSL•FO
RenderX

Likert responses, we computed, for each respondent, the median of the 9 item scores and rounded it down to the nearest integer. This yielded a 4-level ordinal GenAI maturity score: 1 (minimal), 2 (limited), 3 (moderate), and 4 (high maturity). We required at least 5 nonmissing responses out of 9 (≥50%) to compute the index, consistent with commonly used "half-rule" approaches for scoring multi-item scales in the presence of item-level missingness [27].

In this sample, the 9 items showed good internal consistency (Cronbach α=0.869). Internal consistency was similar when excluding the general digital-comfort item (α=0.870 for items 2 - 9), suggesting that digital comfort functions as a foundational component of the same underlying construct. Inspired by previous AI-readiness instruments that combine several facets into a single readiness score [24,25], we therefore use this composite GenAI maturity index pragmatically as an ordinal stratification of self-perceived GenAI maturity.

Qualitative Analysis of Open-Ended Responses

Open-ended responses were analyzed using a qualitative descriptive approach with descriptive content analysis. Responses were segmented into 213 meaning units (eg, items separated by semicolons or distinct ideas) and analyzed by question (reasons for nonuse, additional tasks, suggested uses, suggested risks, and training or support needs). A hybrid coding strategy was used, combining a first deductive framework aligned with these domains and inductive identification of recurrent subthemes. Two researchers (SK and MM) independently coded all segments using the agreed codebook. Initial agreement was 83.1% (177/213); disagreements (36/213, 16.9%) were resolved through discussion until consensus, and the final coding was used for thematic reporting and choice of illustrative quotations. Quotations were anonymized and translated from French to English. Qualitative findings are reported as recurring themes with illustrative excerpts to complement the quantitative results rather than as a stand-alone qualitative study.

Data Processing and Statistical Analyses

Data were exported to Excel and analyzed in Python (version 3.11) using pandas (2.2.2), NumPy (1.26.4), SciPy (1.13.1), statsmodels (0.14.2), and matplotlib (3.9.2). Quantitative variables were described using means or medians with IQRs, while qualitative variables were presented as counts and percentages.

We performed overall group comparisons using Pearson chi-square test to assess the heterogeneity of distributions across academic disciplines and maturity levels. When expected cell counts were small, Fisher exact test was applied. For pairwise comparisons between a given discipline and all others, odds ratios with 95% CIs were calculated.

To assess gradients across maturity levels, 4-point Likert responses were dichotomized into 2 categories (disagree=1 - 2; agree=3 - 4) to facilitate interpretation and to enable proportion-based trend testing across ordered groups [28-30]. Linear trends were then assessed using the Cochran-Armitage trend test. When multiple simultaneous comparisons were conducted, the family-wise error rate was controlled using the Holm method. Statistical significance was set at *P* less than .05. We acknowledge that dichotomizing ordinal responses entails information loss compared with ordinal models [28].

Results

Respondent Demographics and Academic Profile

The survey included 388 health professional students (Table 1), predominantly women (n=308, 79.4%), with most respondents aged 20 - 24 years (n=198, 51.0%). The sample was largely composed of nursing students (n=217, 55.9%), with additional representation from medicine, pharmacy, midwifery, and physiotherapy programs (Table 1). Years of study covered the full curriculum, with a higher proportion of early-year students (y 1 - 2) and smaller proportions in later years (Table 1).

Among the 388 respondents, 380 (97.9%) provided sufficient data to compute the GenAI maturity index. Most students reported limited to moderate maturity (levels 2 - 3: 329/380, 86.6%), while relatively few reported minimal or high maturity (levels 1 and 4: 51/380, 13.4%; Table 1). Item-level details are provided in Multimedia Appendix 1.

XSL•FO
RenderX

Table . Descriptive table of responder characteristics (n=388).

Characteristics	Participants, n (%)
Gender	
Man	76 (19.6)
Woman	308 (79.4)
Prefer not to answer	4 (1.0)
Age (y)	
<20	55 (14.2)
20 - 24	198 (51.0)
25 - 29	55 (14.2)
30 - 34	22 (5.7)
>34	58 (15.0)
Disciplines	
Medicine	69 (17.8)
Midwifery	23 (5.9)
Nursing	217 (55.9)
Pharmacy	66 (17.0)
Physiotherapy	13 (3.4)
Year of study	
First year	82 (21.1)
Second year	97 (25.0)
Third year	77 (19.9)
Fourth year	29 (7.5)
Fifth year	52 (13.4)
Sixth year	10 (2.6)
Beyond the sixth year	41 (10.6)
Self-reported GenAI[a] maturity	
Minimal (1)	22 (5.7)
Limited (2)	130 (33.5)
Moderate (3)	199 (51.3)
High (4)	29 (7.5)
NA[b]	8 (2.1)

[a]GenAI: generative artificial intelligence.

[b]NA: not applicable.

Self-Reported Uses

GenAI Use in Clinical Placement

Overall, 204 of 388 (52.6%) students reported using GenAI during clinical placements (Table 2). Adoption differed across disciplines (χ^2_4=10.71; *P*=.03), primarily reflecting lower uptake among midwifery students (odds ratio 0.30, 95% CI 0.11 - 0.77; Holm-adjusted *P*=.049). Pharmacy students showed the highest unadjusted proportion of use, but this difference was not statistically significant after correction (Table 2). GenAI use increased markedly with higher self-perceived GenAI maturity, from 2 of 22 (9%) in the minimal group to 22 of 29 (76%) in the high-maturity group (trend test *z* score=6.01, *P*<.001; Table 3).

XSL•FO
RenderX

Table . Prevalence of GenAI[a] use in clinical placements by discipline.

	Yes, n (%)	No, n (%)	OR[b] (95% CI) vs rest	*P* value
GenAI use in clinical placements				
Overall	204 (52.6)	184 (47.4)	—[c]	—
GenAI use by discipline				
Medicine	33 (47.8)	36 (52.2)	0.79 (0.47-1.34)	.99
Midwifery	6 (26.1)	17 (73.9)	0.3 (0.11-0.77)	.049[d]
Nursing	117 (53.9)	100 (46.1)	1.13 (0.76-1.69)	.99
Pharmacy	42 (63.6)	24 (36.4)	1.73 (1.00-2.99)	.23
Physiotherapy	6 (46.1)	7 (53.9)	0.77 (0.25-2.32)	.78

[a]GenAI: generative artificial intelligence.

[b]OR: odds ratio.

[c]Not applicable.

[d]Statistically significant after Holm adjustment.

Table . Prevalence of GenAI[a] use in clinical placements by self-perceived GenAI maturity.

GenAI use by self-perceived GenAI maturity	Yes, n (%)	No, n (%)	χ^2 Pearson (*df*; *P* value)	Cochran-Armitage trend: *z* score (*P* value)
Maturity score			37.25 (3; <.001)	6.01 (<.001)
Minimal (1)	2 (9.1)	20 (90.9)		
Limited (2)	52 (40.0)	78 (60.0)		
Moderate (3)	122 (61.3)	77 (38.7)		
High (4)	22 (75.9)	7 (24.1)		

[a]GenAI: generative artificial intelligence.

Reasons for Not Adopting GenAI During Clinical Placements Situations

Among nonusers (184/388), the most often reported reasons were not perceiving a need (59/184, 32.1%), not knowing best practices for use (38/184, 20.7%), and lack of interest (34/184, 18.5%; Table 4). Fewer respondents reported not knowing how to use the tools (18/184, 9.8%) or not being aware of them (10/184, 5.4%); "Other" was selected by 25/184 (13.6%).

Table . Reasons for not adopting generative artificial intelligence during clinical placements among nonuser respondents (n=184).

Reason for nonuse	Overall, n (%)	Nursing, n (%)	Medicine, n (%)	Pharmacy, n (%)	Midwifery, n (%)	Physio, n (%)
It does not interest me	34 (18.5)	13 (13)	10 (27.8)	3 (12.5)	5 (29.4)	3 (42.9)
I do not need it	59 (32.1)	32 (32)	10 (27.8)	9 (37.5)	6 (35.3)	2 (28.6)
I do not know how to use it	18 (9.8)	12 (12)	4 (11.1)	2 (8.3)	0 (0)	0 (0)
I don't know the best practices for using it	38 (20.7)	25 (25)	7 (19.4)	2 (8.3)	3 (17.6)	1 (14.3)
I was not aware of it	10 (5.4)	6 (6)	0 (0)	3 (12.5)	1 (5.9)	0 (0)
Other	25 (13.6)	12 (12)	5 (13.9)	5 (20.8)	2 (11.8)	1 (14.3)

Reasons varied across disciplines, but subgroup sizes were small for some programs. Overall, responses suggested two main profiles: (1) attitudinal barriers (eg, no perceived need or lack of interest) and (2) guidance-related barriers (eg, lack of best practices or technical know-how), with their relative weight differing by discipline (Table 4).

Open-ended responses under "Other" (29 meaning units) mostly referred to perceived unreliability and risk of errors,

XSL•FO
RenderX

environmental impact, or broader socio-technical critique, and a preference for official sources or local protocols over GenAI. The full thematic breakdown with illustrative quotations is provided in Multimedia Appendix 1.

Self-Reported Use of GenAI During Clinical Placements Situations

Among GenAI users (204/388), reported use was primarily oriented toward information-related and documentation tasks (Table 5). The most common uses included information extraction (159/204, 77.9%) and bibliographic search (152/204, 74.5%), while patient-facing activities were less frequently reported (eg, drafting patient documents or preparing patient communication: 78/204, 38.2% each; Table 5). This pattern was mirrored in usage frequency: weekly-or-more use was more common for information-related tasks (eg, information extraction: 53/204, 26.0%) than for patient-facing documentation or communication (13/204, 6.4% and 15/204, 7.4%; Table 5).

Free-text comments on additional tasks (106 meaning units) most often described learning support or pedagogical explanations and writing or text revision; the full thematic breakdown with illustrative quotations is provided in Multimedia Appendix 1.

Table . Frequency of generative artificial intelligence usage by task (n=204).

Item	Ever used, n (%)	At least once a week, n (%)
Content translation	145 (71.1)	19 (9.3)
Report or article writing	141 (69.1)	27 (13.2)
Info extraction	159 (77.9)	53 (26)
Bibliographic search	152 (74.5)	47 (23)
Clinical simulation	112 (54.9)	31 (15.2)
Drafting patient documents	78 (38.2)	13 (6.4)
Preparing patient communication	78 (38.2)	15 (7.4)

Disclosure of Sensitive Information to GenAI

Among GenAI users, most reported that entering direct patient identifiers had “never happened” (151/204, 74.0%; Table 6). However, 48 of 204 (23.5%) students reported at least 1 instance (once, a few times, often, or unintentionally), 5 (2.5%) were unsure, and only 3 (1.5%) reported that this occurs “often.”

Self-disclosure was more common. While 109 of 204 (53.4%) students reported “never” entering their own personal identifiers, 87 (42.6%) reported at least 1 instance, 8 (3.9%) were unsure, and 18 (8.8%) reported that this occurs “often” (Table 6). The broadest category of disclosures involved real medical content: 100 (49%) reported “never,” whereas 96 (47.1%) reported processing real medical data (even if perceived as anonymized), including 20 (9.8%) reporting “often” (Table 6).

Table . Self-reported disclosure of sensitive information when using generative artificial intelligence (n=204).

Type of data disclosed	Never, n (%)	Once, n (%)	A few times, n (%)	Often, n (%)	Maybe[a], n (%)	Do not know, n (%)
Patient-identifying data (*name, ID, address*)	151 (74)	12 (5.9)	22 (10.8)	3 (1.5)	11 (5.4)	5 (2.5)
Student’s own personal data (*name, identifiers*)	109 (53.4)	18 (8.8)	39 (19.1)	18 (8.8)	12 (5.9)	8 (3.9)
Real medical data (*anonymized or not*)	100 (49.0)	17 (8.3)	56 (27.5)	20 (9.8)	3 (1.5)	8 (3.9)

[a]It may have happened unintentionally (eg, by copying and pasting).

Perceived Benefits and Risks

Perceived Benefits of GenAI Use in Clinical Practice

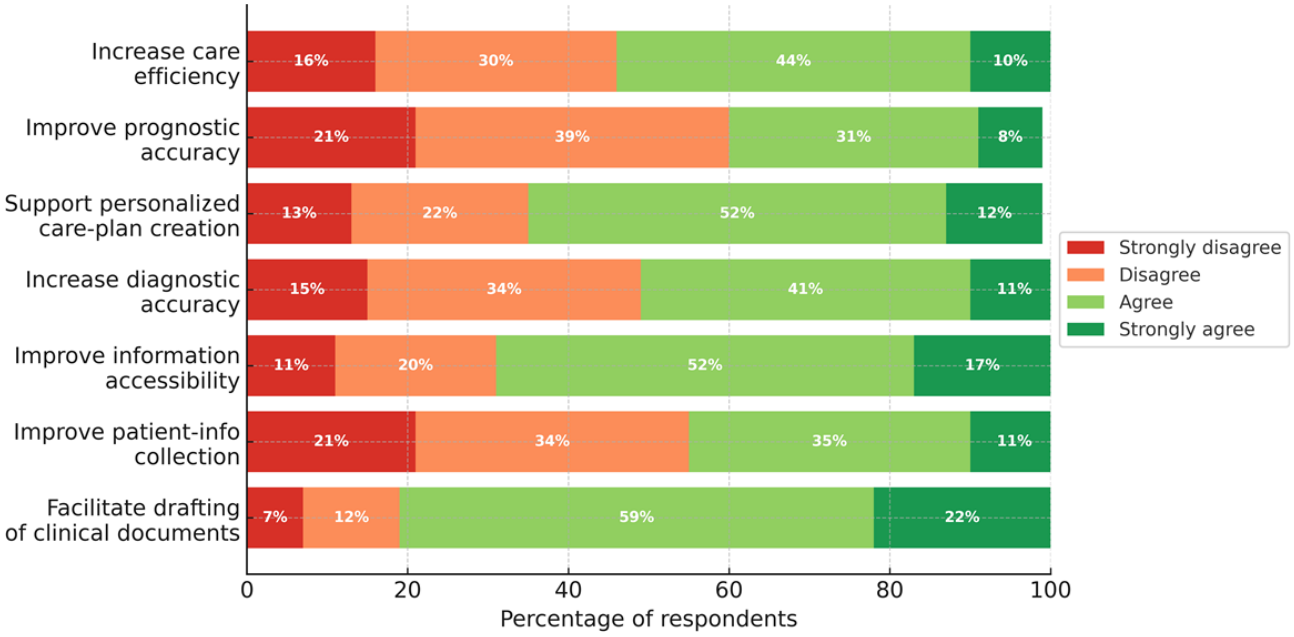
Students rated seven potential benefits of GenAI in clinical practice on a 4-point Likert scale (n=388; Figure 1). Overall, perceived benefits were highest for facilitating the drafting of clinical documents (315/388, 81.2% agree or strongly agree) and improving information accessibility (266/388, 68.5%). Intermediate endorsement was observed for supporting personalized care plan creation, increasing care efficiency, and improving diagnostic accuracy (Figure 1). In contrast, views were more skeptical regarding patient-information collection and prognostic accuracy, for which most respondents disagreed (212/388, 54.6% and 234/388, 60.3%, respectively). Medians indicated overall agreement for 5 of 7 potential benefits, whereas patient information collection and prognostic accuracy had median disagreement. Full response distributions, item-level medians, and IQRs are reported in Multimedia Appendix 1.

Free-text proposals about potential clinical uses were segmented into 22 coded meaning units and were consistent with the

XSL•FO
RenderX

quantitative patterns. The most frequent themes concerned reflective support and systematic checklists and administrative or documentation streamlining. The full thematic breakdown, including less frequent themes and illustrative quotations, is provided in Multimedia Appendix 1.

Figure 1. Perceived benefits of generative artificial intelligence for clinical practice during placements (n=388).



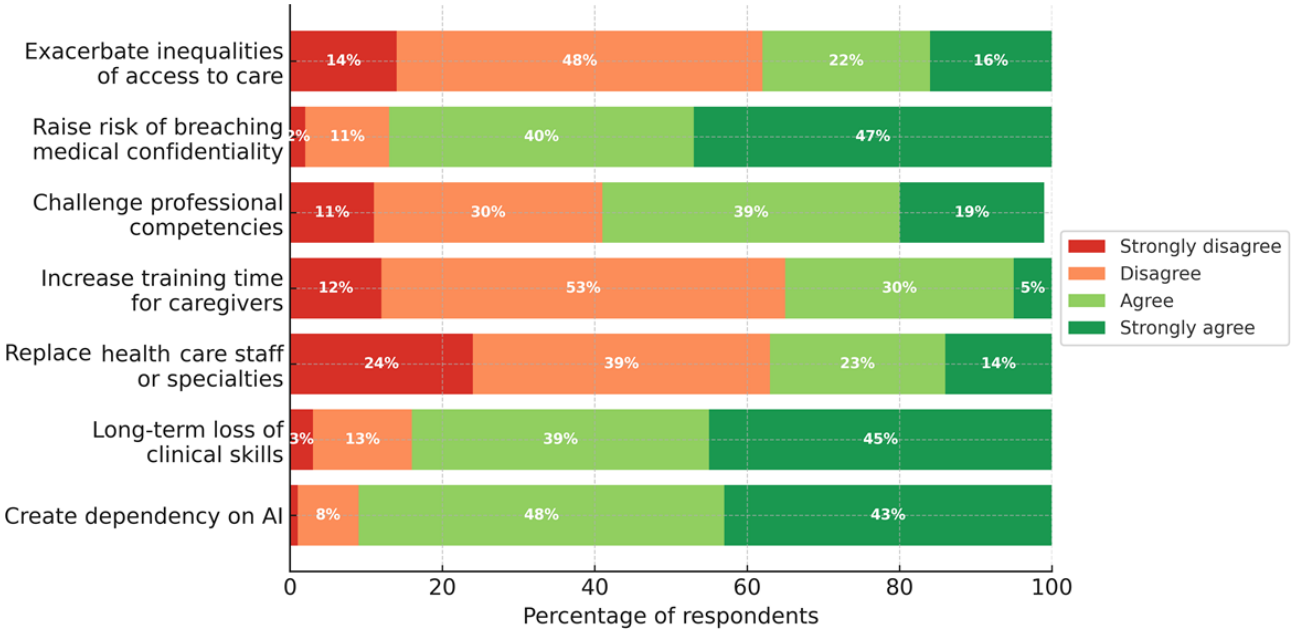
Perceived Risks of GenAI Use in Clinical Practice

Students rated seven potential risks of GenAI use in clinical practice on a 4-point Likert scale (n=388; Figure 2). Three concerns were strongly endorsed: creating dependency on GenAI (353/388, 90.9% agree or strongly agree), long-term loss of clinical skills (329/388, 84.8%), and breaching medical confidentiality (339/388, 87.4%). A more moderate majority agreed that GenAI could challenge professional competencies (228/388, 58.8%; Figure 2). In contrast, most respondents disagreed that GenAI would replace health care staff or specialists, increase training time for caregivers, or exacerbate inequalities in access to care (all >60% disagree or strongly disagree; Figure 2). Medians showed agreement for 4 of 7 risks, while replacement of staff, increased training time, and inequalities had median disagreement. Full response distributions, item-level medians, and IQRs are reported in Multimedia Appendix 1.

Open-ended risk comments were segmented into 32 coded meaning units and were consistent with the hierarchy seen in the Likert ratings. The most frequent themes concerned risks of errors and potential patient-safety consequences, dependence or reduced reflective capacity, and relational impacts (eg, depersonalization or reduced quality of the clinician-patient relationship). The full thematic breakdown, including less frequent themes and illustrative quotations, is provided in Multimedia Appendix 1.

XSL•FO
RenderX

Figure 2. Perceived risks of clinical use of generative artificial intelligence during placements (n=388). AI: artificial intelligence.

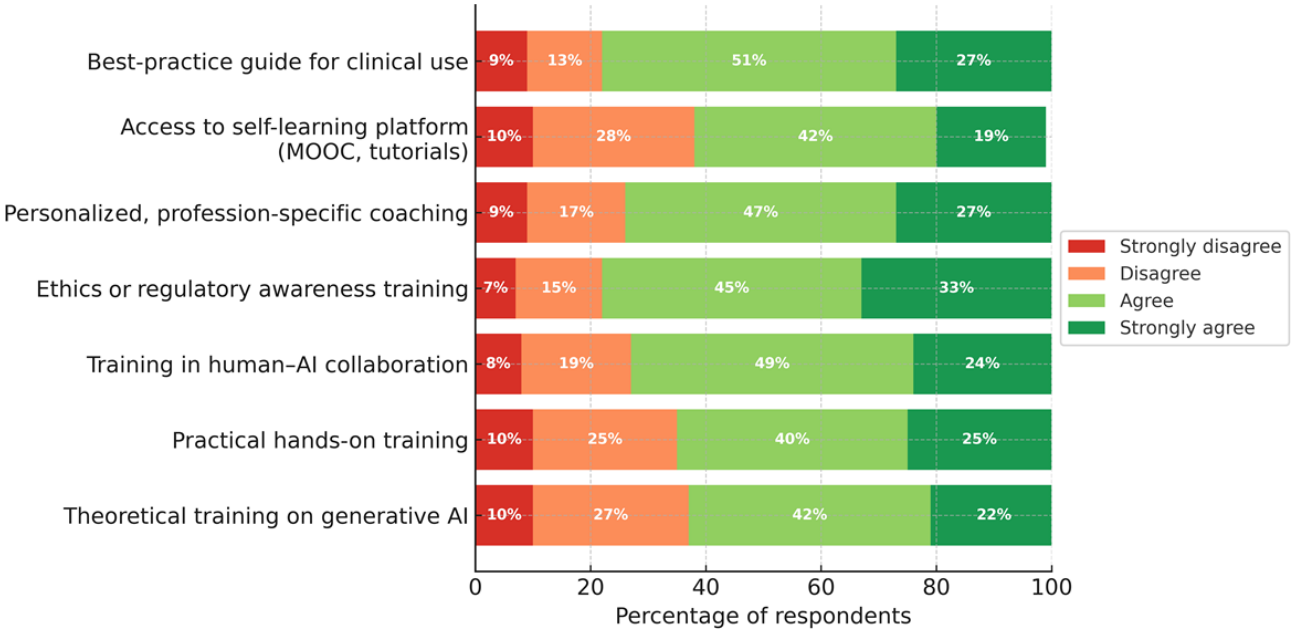


Training and Support Needs for GenAI Use

Across 7 items assessing training and support needs, the median response was "Agree" for every statement (Figure 3). The strongest endorsement concerned a best-practice clinical guide (292/373, 78.3% agree or strongly agree) and ethics or regulatory awareness training (294/378, 77.7%). Important levels of agreement were also observed for profession-specific coaching and training in human-AI collaboration (Figure 3). Overall disagreement remained limited (strongly disagree ≤11% across items). Full response distributions, item-level medians, and IQRs are reported in Multimedia Appendix 1.

Free-text comments on additional needs (24 meaning units) provided further context. A large share reported no additional needs or interest (8/24, 33%). Among expressed needs, the most frequent theme concerned practical training and ongoing support to integrate GenAI appropriately into clinical learning and practice (6/24, 25%). The full thematic breakdown, including less frequent themes and illustrative quotations, is provided in Multimedia Appendix 1.

Figure 3. Training and support needs for generative artificial intelligence (n=388). AI: artificial intelligence; MOOC: massive open online course.



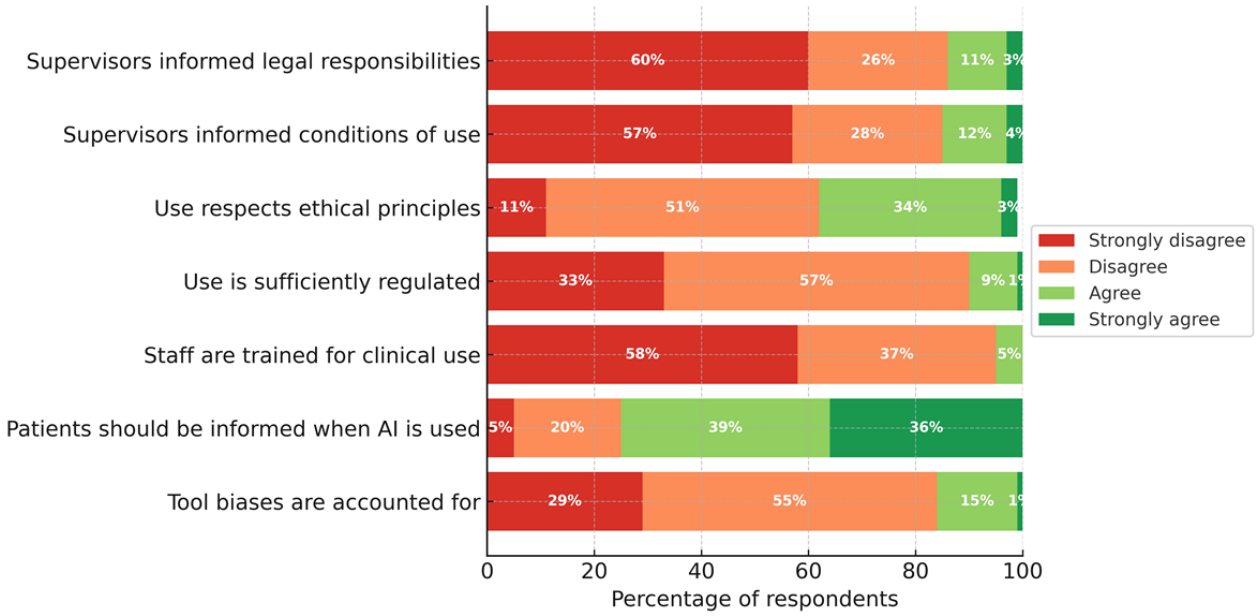
Governance and Organizational Readiness

Seven items assessed students' perceptions of organizational readiness and governance for GenAI use during clinical placements (Figure 4). Overall, respondents perceived limited institutional communication and safeguards: only a small minority recalled supervisors providing information on legal responsibilities or local conditions of use (Figure 4). Governance was also perceived as insufficient, with 36 of 359 (10%) agreeing that GenAI use is sufficiently regulated. Operational readiness received the poorest ratings, as only 18 of 372 (4.8%) agreed that staff are trained for clinical use and 56 of 343

XSL•FO
RenderX

(16.3%) agreed that tool biases are accounted for. In contrast, transparency toward patients appeared as the single area of strong consensus: 248 of 331 (74.9%) agreed that patients should be informed when AI is used in their care (Figure 4). Full response distributions, item-level medians, and IQRs are reported in Multimedia Appendix 1.

Figure 4. Governance and organizational readiness (n=388). AI: artificial intelligence.



Discussion

Principal Findings

In this cross-sectional survey of 388 French students in health professions, 52.6% (204/388) reported using GenAI during clinical placements. Adoption varied across disciplines, with lower uptake in midwifery, and increased markedly with higher self-perceived GenAI maturity. Reported uses primarily involved information retrieval and documentation- or writing-related tasks, while patient-facing activities were less frequently reported.

Notably, close to a quarter of GenAI users reported at least 1 instance of entering patient-identifying information, and close to half reported processing real medical content perceived as anonymized. Students perceived substantial benefits for documentation support but also expressed prominent concerns about dependency, erosion of clinical skills, and breaches of confidentiality, alongside strong demand for ethics or regulation training and best-practice guidance. Overall, these findings show that GenAI is already embedded in students' placement practices, but with uneven adoption, limited institutional framing, and persistent confidentiality risks.

Learning to Think With GenAI During Clinical Placements

In this survey, students appeared to mobilize GenAI during placements primarily as a cognitive and organizational aid (eg, information retrieval and documentation support) rather than as a tool embedded in direct patient-facing work. This "low-stakes *versus* clinically situated" pattern plausibly reflects implicit boundaries around what feels safe, acceptable, and professionally legitimate to delegate to a third-party model in real clinical environments and aligns with students' expressed need for clear best-practice guidance.

Perceived readiness offers an additional lens to interpret why some students integrate GenAI into placement work while others do not. Our maturity index should be understood as a pragmatic stratification of how prepared students feel to evaluate and use GenAI critically in clinical learning contexts, rather than as an objective measure of competence. The strong adoption gradient across readiness levels is consistent with AI-readiness frameworks in health professions education [24,25]. Emerging evidence also suggests that structured GenAI-supported simulations can improve targeted competencies in medical students [31,32]. Educationally, this matters because uneven readiness may translate into uneven exposure to GenAI practices during placements, potentially reinforcing a hidden curriculum in which norms of safe and legitimate use are acquired informally and shape professional identity formation at a formative stage [33,34].

This raises a broader tension between efficiency and reflection in workplace learning. While GenAI may support information management and autonomous learning [35], medical documentation and reasoning processes are also critical spaces for cultivating clinical rigor, accountability, and reflective judgment. Without explicit pedagogical framing, early delegation of reasoning-related tasks may narrow opportunities for reflexivity and increase vulnerability to automation bias and overreliance. These mechanisms were previously described in clinical decision support contexts, where they can compromise patient safety [36]. Conversely, when embedded in structured educational approaches (eg, simulation with feedback and validated rubrics), GenAI can be leveraged to enrich reflection and support deliberate practice, as illustrated by initiatives such

XSL•FO
RenderX

as MedSimAI [32]. Collectively, these findings support the need to move from informal, individual experimentation toward supervised and pedagogically grounded uses of GenAI in clinical learning environments [34,36].

Interpreting Nonuse, Perceived Benefits, and Perceived Risks

Students' reasons for not using GenAI during placements suggested that nonadoption often reflected deliberate reservations rather than simple unfamiliarity. Reported concerns included limited perceived clinical relevance, fear of diminished reflexivity, confidentiality risks, and broader socio-technical critiques, such as environmental impact. This aligns with Tortella et al [6], who observed that reluctance among physiotherapy students was largely driven by concerns about inaccuracy and error, and perceived limits for complex tasks, rather than a lack of awareness. From a teaching perspective, such reservations should be treated as a pedagogical resource: they create opportunities to make expectations explicit and to cultivate critical judgment, irrespective of whether students ultimately adopt GenAI tools.

Perceived benefits were large but not uniform. Students strongly endorsed GenAI for documentation relief and improved access to information, positioning it primarily as a cognitive and organizational facilitator. In contrast, endorsement was more cautious for higher-stakes functions such as diagnostic support, personalized planning, and efficiency gains, reflecting conditional trust rather than full reliance. This pattern is consistent with Janumpally et al [37], who emphasized that GenAI may support learning and selected reasoning tasks but remains constrained by output reliability and cannot substitute for human judgment, particularly in domains directly implicating professional responsibility, such as patient data collection and prognostication.

Perceived risks were dominated by concerns about dependency, long-term erosion of clinical skills, and confidentiality breaches, alongside worries about bias, uncertain legal accountability, and a more subtle displacement of professional responsibility and human interaction. These concerns resonate with the broader literature on automation bias and overreliance in clinical decision support, including discussions of "assistive AI" as the lowest level of autonomy that may still erode vigilance if not framed and supervised [36]. Importantly, students did not primarily express fears of being replaced. Rather, they emphasized risks of weakened judgment, diluted accountability, and reduced quality of the care relationship. These concerns suggest that GenAI integration in workplace learning should be structured, reflective, and explicitly governed rather than left to informal experimentation [6,36,37].

Confidentiality and Data Governance

From a patient-safety and governance perspective, the disclosure of sensitive information calls for specific attention. In our sample, 23.5% (48/204) reported at least 1 instance of entering patient-identifying information, and 47.1% (96/204) reported processing real medical content perceived as anonymized. Even when students perceive data as anonymized, free-text prompts may contain contextual details that increase reidentification risk, and the use of nonapproved third-party tools may involve data processing outside institutional control (eg, storage and transfer conditions) [38,39]. In European contexts, this directly intersects with General Data Protection Regulation obligations and institutional requirements for handling health data [40].

These findings imply a need for explicit supervision and coordinated governance during placements. Placement supervisors, faculty, and host clinical sites (eg, hospital departments and private practices) should clearly define acceptable versus unacceptable prompting, provide concrete deidentification rules and examples, and ensure access to approved, compliant tools (or clear prohibitions where such tools are not available). Because clinical training occurs across multiple host sites, governance should be coordinated between the academic institution and placement providers so that expectations and safeguards are consistent across settings. At the institutional level, local charters and "allowed tool" policies aligned with data-protection governance should go with training, so that confidentiality norms are not left to informal experimentation. In France, this approach is consistent with recent institutional guidance: the French National Academy of Medicine discusses ethical issues and recommendations for the use of GenAI in health [41], and the French National Authority for Health has published a practical guide to support appropriate and informed use of GenAI by health professionals [42].

From Technical Literacy to Patient-Centered GenAI Education

Students' expectations for training extend beyond technical mastery toward the full triad of knowledge, skills, and professional attitudes: knowledge of ethical and regulatory frameworks; practical competencies developed through workshops and profession-specific coaching; and professional dispositions cultivated through mentorship, reflexivity, and attention to sustainability. This aligns with calls to design health AI curricula that move past technical literacy to incorporate critical thinking, ethics, and patient-centered judgment. Komasawa and Yokohira [43] argue that the rise of GenAI requires rethinking health professionalism by combining technological fluency with enduring humanistic values such as empathy, integrity, and accountability.

A Pragmatic Pathway From Informal Adoption to Responsible Integration

Our findings point to an institutional readiness gap during placements: limited shared benchmarks, uneven supervisory guidance, and insufficient attention to bias, regulation, and data governance. They also support a stepwise pathway to move from informal, student-driven use toward safe and educationally meaningful integration across disciplines. First, a transversal common core should address ethics, regulation, professional standards, and model limitations or uncertainty, alongside concrete rules for health data handling in GenAI contexts (what must never be entered, how to deidentify, and which tools are approved). Second, competency development requires practice beyond theory: simulation- or case-based activities with GenAI, explicit reflection prompts, and structured feedback can help students learn when GenAI supports reasoning versus when it undermines it (eg, automation bias and overreliance) and can

XSL•FO
RenderX

contribute to professional identity formation. Third, training for both academic staff and host-site supervisors is essential so that placement supervisors can set shared expectations, mentor reflective use, and intervene when unsafe practices occur. Fourth, institutional governance should provide clear local policies (eg, charters, approved tool lists, patient transparency where relevant, and escalation procedures for data incidents), ensuring alignment between pedagogy and professional standards. This enabling framework does not mandate adoption; it operationalizes conditions under which GenAI use remains compatible with reflexivity, professional accountability, and patient-centered care, and can help transform spontaneous use into integrated, supervised competencies in clinical learning settings, as illustrated by structured simulation initiatives such as MedSimAI [32].

Limitations

This study has several limitations. First, it was conducted within a single French university (UGA), which limits international generalizability. The survey was embedded in a national initiative funded by the French National Research Agency to support the rollout of digital health and AI in health education across French university training programs. While this may strengthen the transferability of educational implications, it does not eliminate the single-site constraint. Notably, the main trends observed here are consistent with the international survey by Busch et al [44] (4596 students from 192 faculties across 48 countries), which similarly reported widespread student adoption of generative tools alongside a global deficit in formal AI education. Second, GenAI use and data-disclosure behaviors were measured through self-report and may be influenced by recall and social desirability biases. Although responses were collected anonymously and without incentives, self-reporting remains susceptible to these biases. Nevertheless, documenting students' lived practices and perceptions was a core aim, particularly in contexts where formal pedagogical frameworks are still emerging; as such, perceptions may be viewed as outcomes in their own right, consistent with Busch et al [44]. Third, we did not collect data from clinical supervisors or educational leaders and therefore cannot characterize institutional policies or supervisory practices with certainty. This likely reflects a transitional situation in which both learners and supervisors are adapting to rapidly evolving GenAI tools and local guidance, as suggested by recent work by Blanco et al [45] and McCoy et al [46]. Fourth, the cross-sectional design provides a snapshot at a specific time point; given the rapid evolution of GenAI tools and institutional policies, the stability of these findings over time cannot be assumed. Fifth, interprofessional heterogeneity is an additional limitation: some health professions were underrepresented, limiting robust profession-specific comparisons. This reflects our primary aim of mapping placement-related practices rather than comparing disciplines; nonetheless, profession-specific differences in AI literacy and attitudes have been reported elsewhere and warrant dedicated analyses [47,48]. Finally, our GenAI maturity score should be considered an exploratory composite index of self-perceived readiness. Although internal consistency was good (Cronbach α=0.869), we did not undertake a full psychometric validation; further work should confirm and refine this construct.

Perspectives

Future research should move beyond single institutions to include multicenter and international comparisons, enabling the identification of contextual variations and common challenges. Longitudinal designs will also be needed to capture how perceptions, competencies, and professional identity evolve as GenAI becomes more embedded in health professions education. Finally, triangulating student perspectives with those of supervisors, educators, and institutional leaders will be crucial to developing evidence-based strategies that responsibly integrate GenAI into curricula while respecting profession-specific needs.

Conclusion

In conclusion, GenAI is already being used by French students in health professions during clinical placements, primarily as a cognitive and organizational aid. At the same time, students report concerns about dependence, skill erosion, and confidentiality, and express a strong demand for structured guidance. These findings support the need for explicit, ethically grounded, and supervised frameworks that align curricula, clinical supervision, and host-site governance so that GenAI use fosters reflexivity, autonomy, and professional accountability rather than unsupervised experimentation. More broadly, this echoes developments already observed across higher education and health care governance, including emerging institutional policies [13], the timely need for clear ethical and regulatory frameworks [49], and methodological approaches for developing evidence-based interdisciplinary guidelines [50].

Acknowledgments

The authors gratefully acknowledge the support of the deans of the faculties of medicine and pharmacy at Université Grenoble Alpes, whose approval enabled the dissemination of the survey invitation to students. Their assistance in facilitating access to medical and allied health students was essential to the successful recruitment process. During the preparation of this manuscript, the authors used ChatGPT (GPT-5) for code optimization in data analysis and for language editing. The authors reviewed and edited all outputs and take full responsibility for the content, data interpretations, and conclusions of this publication.

XSL•FO
RenderX

Funding

The study was supported by the French National Research Agency under the France 2030 program (23-CMAS-0035.) and the Multidisciplinary Institute in Artificial intelligence Cluster (23-IACL-0006). This work also forms part of a broader translational and interdisciplinary Santé Numérique Grenoble Intelligence Artificielle research program.

Data Availability

The data supporting the findings of this study are not publicly available due to ethical and privacy restrictions. Anonymized datasets can be made available from the corresponding author, SK, upon reasonable request and subject to approval by the institutional ethics committee.

Authors' Contributions

SK conceived and coordinated the study and drafted the first version of the manuscript. CMT and KG contributed to methodological development. Analyses were performed by SK, CMT, and MM. Data collection was conducted by SK, KG, and MM. Interpretation of the results was conducted by SK, CMT, KG, and MM. NV provided overall supervision and critical guidance throughout the project. All authors critically reviewed the manuscript and approved the definitive version.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Full survey questionnaire (French version with English translation) and supplementary tables (Tables S1-S11) reporting the generative artificial intelligence maturity items and detailed response distributions across key study domains.
[DOCX File, 79 KB - mededu_v12i1e85243_app1.docx]

Checklist 1
CHERRIES checklist for the web-based survey reporting.
[DOCX File, 584 KB - mededu_v12i1e85243_app2.docx]

References

1. Ganjavi C, Eppler M, O'Brien D, et al. ChatGPT and large language models (LLMs) awareness and use. a prospective cross-sectional survey of U.S. medical students. PLOS Digit Health 2024 Sep;3(9):e0000596. [doi: 10.1371/journal.pdig.0000596] [Medline: 39236008]
2. Moskovich L, Rozani V. Health profession students' perceptions of ChatGPT in healthcare and education: insights from a mixed-methods study. BMC Med Educ 2025 Jan 21;25(1):98. [doi: 10.1186/s12909-025-06702-0] [Medline: 39833868]
3. Weidener L, Fischer M. Artificial intelligence in medicine: cross-sectional study among medical students on application, education, and ethical aspects. JMIR Med Educ 2024 Jan 5;10:e51247. [doi: 10.2196/51247] [Medline: 38180787]
4. El Arab RA, Al Moosa OA, Abuadas FH, Somerville J. The role of AI in nursing education and practice: umbrella review. J Med Internet Res 2025 Apr 4;27:e69881. [doi: 10.2196/69881] [Medline: 40072926]
5. Yuan Y, Fu J, Leng L, et al. The strengths, weaknesses, opportunities, and threats of generative artificial intelligence: a qualitative study of undergraduate nursing students. Front Public Health 2025;13:1672140. [doi: 10.3389/fpubh.2025.1672140] [Medline: 40977793]
6. Tortella F, Palese A, Turolla A, et al. Knowledge and use, perceptions of benefits and limitations of artificial intelligence chatbots among Italian physiotherapy students: a cross-sectional national study. BMC Med Educ 2025 Apr 18;25(1):572. [doi: 10.1186/s12909-025-07176-w] [Medline: 40251635]
7. Abd-Alrazaq A, AlSaad R, Alhuwail D, et al. Large language models in medical education: opportunities, challenges, and future directions. JMIR Med Educ 2023 Jun 1;9:e48291. [doi: 10.2196/48291] [Medline: 37261894]
8. Boscardin CK, Gin B, Golde PB, Hauer KE. ChatGPT and generative artificial intelligence for medical education: potential impact and opportunity. Acad Med 2024 Jan 1;99(1):22-27. [doi: 10.1097/ACM.0000000000005439] [Medline: 37651677]
9. Divito CB, Katchikian BM, Gruenwald JE, Burgoon JM. The tools of the future are the challenges of today: the use of ChatGPT in problem-based learning medical education. Med Teach 2024 Mar;46(3):320-322. [doi: 10.1080/0142159X.2023.2290997] [Medline: 38149617]
10. Pham TD, Karunaratne N, Exintaris B, et al. The impact of generative AI on health professional education: a systematic review in the context of student learning. Med Educ 2025 Dec;59(12):1280-1289. [doi: 10.1111/medu.15746] [Medline: 40533396]
11. Svendsen K, Askar M, Umer D, Halvorsen KH. Short-term learning effect of ChatGPT on pharmacy students' learning. Explor Res Clin Soc Pharm 2024 Sep;15:100478. [doi: 10.1016/j.rcsop.2024.100478] [Medline: 39139501]

XSL•FO
RenderX

12. Elnaem MH, Okuyan B, Mubarak N, et al. Students' acceptance and use of generative AI in pharmacy education: international cross-sectional survey based on the extended unified theory of acceptance and use of technology. Int J Clin Pharm 2025 Aug;47(4):1097-1108. [doi: 10.1007/s11096-025-01936-w] [Medline: 40465181]
13. McDonald N, Johri A, Ali A, Collier AH. Generative artificial intelligence in higher education: evidence from an analysis of institutional policies and guidelines. Comput Hum Behav Artif Hum 2025 Mar;3:100121. [doi: 10.1016/j.chbah.2025.100121]
14. Sharifi Kelarijani A, Safdari A, Golitaleb M. Every coin has two sides: ChatGPT poses a potential threat to nursing students' education. Front Med (Lausanne) 2024;11:1415067. [doi: 10.3389/fmed.2024.1415067] [Medline: 39114822]
15. Galletta M, Portoghese I, Aviles Gonzales CI, et al. Lack of respect, role uncertainty and satisfaction with clinical practice among nursing students: the moderating role of supportive staff. Acta Biomed 2017 Jul 18;88(3S):43-50. [doi: 10.23750/abm.v88i3-S.6613] [Medline: 28752832]
16. Mazalová L, Gurková E, Štureková L. Nursing students' perceived stress and clinical learning experience. Nurse Educ Pract 2022 Oct;64:103457. [doi: 10.1016/j.nepr.2022.103457] [Medline: 36182730]
17. Lin YK, Lin CD, Lin BYJ, Chen DY. Medical students' resilience: a protective role on stress and quality of life in clerkship. BMC Med Educ 2019 Dec 27;19(1):473. [doi: 10.1186/s12909-019-1912-4] [Medline: 31881997]
18. Ali SR, Dobbs TD, Hutchings HA, Whitaker IS. Using ChatGPT to write patient clinic letters. Lancet Digit Health 2023 Apr;5(4):e179-e181. [doi: 10.1016/S2589-7500(23)00048-1] [Medline: 36894409]
19. Jeblick K, Schachtner B, Dexl J, et al. ChatGPT makes medicine easy to swallow: an exploratory case study on simplified radiology reports. Eur Radiol 2024 May;34(5):2817-2825. [doi: 10.1007/s00330-023-10213-1] [Medline: 37794249]
20. El Arab RA, Al Moosa OA, Sagbakken M, et al. Integrative review of artificial intelligence applications in nursing: education, clinical practice, workload management, and professional perceptions. Front Public Health 2025;13:1619378. [doi: 10.3389/fpubh.2025.1619378] [Medline: 40823249]
21. Blease CR, Locher C, Gaab J, Hägglund M, Mandl KD. Generative artificial intelligence in primary care: an online survey of UK general practitioners. BMJ Health Care Inform 2024 Sep 17;31(1):e101102. [doi: 10.1136/bmjhci-2024-101102] [Medline: 39288998]
22. Rahmanti AR, Iqbal U, Reddy S, Gao XW, Nguyen HX, Li YCJ. Generative artificial intelligence (AI): a key innovation or just hype in primary care settings? BMJ Health Care Inform 2024 Dec 31;31(1):e101367. [doi: 10.1136/bmjhci-2024-101367] [Medline: 39740857]
23. Lobet M, Honet A, Romainville M, Wathelet V. ChatGPT: quel en a été l'usage spontané d'étudiants de première année universitaire à son arrivée? [Article in French]. Med Med 2024 Oct 30(18):67-90. [doi: 10.52358/mm.vi18.379]
24. Karaca O, Çalışkan SA, Demir K. Medical Artificial Intelligence Readiness Scale for Medical Students (MAIRS-MS) - development, validity and reliability study. BMC Med Educ 2021 Feb 18;21(1):112. [doi: 10.1186/s12909-021-02546-6] [Medline: 33602196]
25. Boillat T, Nawaz FA, Rivas H. Readiness to embrace artificial intelligence among medical doctors and students: questionnaire-based study. JMIR Med Educ 2022 Apr 12;8(2):e34973. [doi: 10.2196/34973] [Medline: 35412463]
26. Eysenbach G. Improving the quality of web surveys: the Checklist for Reporting Results of Internet E-Surveys (CHERRIES). J Med Internet Res 2004 Sep 29;6(3):e34. [doi: 10.2196/jmir.6.3.e34] [Medline: 15471760]
27. Rezvan PH, Comulada WS, Fernández MI, Belin TR. Assessing alternative imputation strategies for infrequently missing items on multi-item scales. Commun Stat Case Stud Data Anal Appl 2022;8(4):682-713. [doi: 10.1080/23737484.2022.2115430] [Medline: 36467970]
28. Agresti A. An Introduction to Categorical Data Analysis, 2nd edition: John Wiley & Sons; 2007. URL: https://onlinelibrary.wiley.com/doi/book/10.1002/0470114754 [accessed 2026-03-31]
29. Sullivan GM, Artino AR. Analyzing and interpreting data from Likert-type scales. J Grad Med Educ 2013 Dec;5(4):541-542. [doi: 10.4300/JGME-5-4-18] [Medline: 24454995]
30. Altman DG, Royston P. The cost of dichotomising continuous variables. BMJ 2006 May 6;332(7549):1080. [doi: 10.1136/bmj.332.7549.1080] [Medline: 16675816]
31. Suárez-García RX, Chavez-Castañeda Q, Orrico-Pérez R, et al. DIALOGUE: a generative AI-based pre-post simulation study to enhance diagnostic communication in medical students through virtual type 2 diabetes scenarios. Eur J Investig Health Psychol Educ 2025 Aug 7;15(8):152. [doi: 10.3390/ejihpe15080152] [Medline: 40863274]
32. Hicke Y, Geathers J, Rajashekar N, et al. MedSimAI: simulation and formative feedback generation to enhance deliberate practice in medical education. arXiv. Preprint posted online on Mar 1, 2025. [doi: 10.48550/arXiv.2503.05793]
33. Wald HS. Professional identity (trans)formation in medical education: reflection, relationship, resilience. Acad Med 2015 Jun;90(6):701-706. [doi: 10.1097/ACM.0000000000000731] [Medline: 25881651]
34. Lawrence C, Mhlaba T, Stewart KA, Moletsane R, Gaede B, Moshabela M. The hidden curricula of medical education: a scoping review. Acad Med 2018 Apr;93(4):648-656. [doi: 10.1097/ACM.0000000000002004] [Medline: 29116981]
35. Monzon N, Hays FA. Leveraging generative artificial intelligence to improve motivation and retrieval in higher education learners. JMIR Med Educ 2025 Mar 11;11:e59210. [doi: 10.2196/59210] [Medline: 40068170]

XSL•FO
RenderX

36. Abdelwanis M, Alarafati HK, Tammam MMS, Simsekler MCE. Exploring the risks of automation bias in healthcare artificial intelligence applications: a Bowtie analysis. J Saf Sci Resil 2024 Dec;5(4):460-469. [doi: 10.1016/j.jnlssr.2024.06.001]
37. Janumpally R, Nanua S, Ngo A, Youens K. Generative artificial intelligence in graduate medical education. Front Med (Lausanne) 2024;11:1525604. [doi: 10.3389/fmed.2024.1525604] [Medline: 39867924]
38. Ford E, Pillinger S, Stewart R, et al. What is the patient re-identification risk from using de-identified clinical free text data for health research? AI Ethics 2025;5(5):4441-4454. [doi: 10.1007/s43681-025-00681-0] [Medline: 40978336]
39. Regulation (EU) 2018/1725 of the European Parliament and of the Council of 23 October 2018 on the protection of natural persons with regard to the processing of personal data by the Union institutions, bodies, offices and agencies and on the free movement of such data, and repealing Regulation (EC) No 45/2001 and Decision No 1247/2002/EC. : European Union; 2018 URL: https://www.edps.europa.eu/data-protection/our-work/publications/legislation/regulation-eu-20181725_en [accessed 2026-01-09]
40. Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation). : European Union; 2016 URL: https://eur-lex.europa.eu/eli/reg/2016/679/oj [accessed 2026-01-09]
41. Nordlinger B, Kirchner C, de Fresnoye O. Rapport 24-03. Systèmes d'IA générative en santé: enjeux et perspectives [Article in French]. Bulletin l'Académie Nationale Médecine 2024 May;208(5):536-547. [doi: 10.1016/j.banm.2024.03.005]
42. Valois P, Marc J, Collignon C. Premières clefs d'usage de l'IA générative en santé [Report in French]. : Haute Autorité de Santé; 2025 URL: https://www.has-sante.fr/jcms/p_3703115/fr/premieres-clefs-d-usage-de-l-ia-generative-en-sante [accessed 2026-01-09]
43. Komasawa N, Yokohira M. Generative artificial intelligence (AI) in medical education: a narrative review of the challenges and possibilities for future professionalism. Cureus 2025 Jun;17(6):e86316. [doi: 10.7759/cureus.86316] [Medline: 40688936]
44. Busch F, Hoffmann L, Truhn D, et al. Global cross-sectional student survey on AI in medical, dental, and veterinary education and practice at 192 faculties. BMC Med Educ 2024 Sep 28;24(1):1066. [doi: 10.1186/s12909-024-06035-4] [Medline: 39342231]
45. Blanco MA, Nelson SW, Ramesh S, et al. Integrating artificial intelligence into medical education: a roadmap informed by a survey of faculty and students. Med Educ Online 2025 Dec;30(1):2531177. [doi: 10.1080/10872981.2025.2531177] [Medline: 40660466]
46. McCoy L, Ganesan N, Rajagopalan V, McKell D, Niño DF, Swaim MC. A training needs analysis for AI and generative AI in medical education: perspectives of faculty and students. J Med Educ Curric Dev 2025;12:23821205251339226. [doi: 10.1177/23821205251339226] [Medline: 40376309]
47. Rincón EHH, Jimenez D, Aguilar LAC, Flórez JMP, Tapia Á, Peñuela CLJ. Mapping the use of artificial intelligence in medical education: a scoping review. BMC Med Educ 2025 Apr 12;25(1):526. [doi: 10.1186/s12909-025-07089-8] [Medline: 40221725]
48. Shishehgar S, Murray-Parahi P, Alsharaydeh E, Mills S, Liu X. Artificial intelligence in health education and practice: a systematic review of health students' and academics' knowledge, perceptions and experiences. Int Nurs Rev 2025 Jun;72(2):e70045. [doi: 10.1111/inr.70045] [Medline: 40545441]
49. García-López IM, Trujillo-Liñán L. Ethical and regulatory challenges of generative AI in education: a systematic review. Front Educ 2025 Jun 30;10:1565938. [doi: 10.3389/feduc.2025.1565938]
50. Symeou L, Louca L, Kavadella A, Mackay J, Danidou Y, Raffay V. Development of evidence-based guidelines for the integration of generative AI in university education through a multidisciplinary, consensus-based approach. Eur J Dent Educ 2025 May;29(2):285-303. [doi: 10.1111/eje.13069] [Medline: 39949032]

Abbreviations

AI: artificial intelligence
CHERRIES: Checklist for Reporting Results of Internet E-Surveys
GenAI: generative artificial intelligence
UGA: Université Grenoble Alpes

XSL•FO
RenderX

Edited by P Kanzow; submitted 03.Oct.2025; peer-reviewed by CY Su, F Fukuzawa; revised version received 15.Feb.2026; accepted 15.Feb.2026; published 27.Apr.2026.

Please cite as:
Kotzki S, Massonnet Turner C, Gauthier K, Minoves M, Vuillerme N
Health Professional Students' Use of Generative Artificial Intelligence During Clinical Placements: Cross-Sectional Online Survey Study
JMIR Med Educ 2026;12:e85243
URL: https://mededu.jmir.org/2026/1/e85243
doi:10.2196/85243

© Sylvain Kotzki, Calvin Massonnet Turner, Kim Gauthier, Mélanie Minoves, Nicolas Vuillerme. Originally published in JMIR Medical Education (https://mededu.jmir.org), 27.Apr.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Education, is properly cited. The complete bibliographic information, a link to the original publication on https://mededu.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Generative AI's Impact on the Mental Health of Medical Students: Scenario Analysis

Nora Arvai, MSC; Bertalan Meskó, MD, PhD; Gellért Katonai, MD

Kálmán Laki Doctoral School of Biomedical and Clinical Sciences, University of Debrecen, Egyetem tér 1, Debrecen, Hungary

Corresponding Author:
Nora Arvai, MSC
Kálmán Laki Doctoral School of Biomedical and Clinical Sciences, University of Debrecen, Egyetem tér 1, Debrecen, Hungary

Abstract

Background: Generative artificial intelligence (AI) is quickly changing medical education, even as medical students still face high levels of stress, anxiety, and burnout. These simultaneous trends—technological upheaval and ongoing mental health issues—bring up important questions about how future doctors will be trained and supported. Understanding how these factors might influence each other is crucial for developing resilient, future-ready medical education systems.

Objective: We carried out a foresight study using scenario analysis to examine potential futures at the crossroads of generative AI adoption and medical students' mental health. An initial environmental scan of the literature was conducted to pinpoint emerging trends and weak signals related to AI in medical education and well-being. These phenomena were categorized within a macro-meso-micro framework and analyzed through a multilevel sociotechnical change perspective. The study focused on 2 principal factors: the extent of generative AI integration into medical curricula and the availability of mental health support, as key drivers and critical uncertainties influencing future trajectories.

Methods: These dimensions resulted in 4 distinct scenarios: Analog Happiness (high support and low AI integration), Gen AI Paradise (high support and high integration), Disconnected Struggles (low support and low integration), and Gen AI Takeover (low support and high integration). Each scenario demonstrates how various institutional responses can impact students' digital readiness, psychological well-being, and professional growth. For each one, we identified the main systemic risks and suggested immediate institutional measures to address them.

Results: The findings suggest that technological innovation and mental health support must coevolve in medical education. Prioritizing one without the other risks producing either digitally unprepared or emotionally fragile physicians. Faculty readiness, ethical frameworks, and participatory curriculum design are critical to ensuring balanced integration. We formulated practical recommendations tailored to students, educators, and other stakeholders to guide balanced adaptation.

Conclusions: Generative AI is more than just an additional tool in medical education; it is a systemic force that redefines how future physicians learn and operate. If technological change and student mental health are tackled separately, medical education risks creating graduates who are either unprepared for digital demands or mentally overwhelmed. This study highlights key systemic risks and suggests initial institutional steps to address them, providing a foresight-driven framework to assist educators and policymakers in responsible AI integration while safeguarding the well-being of future doctors.

(JMIR Med Educ 2026;12:e85373) doi:10.2196/85373

KEYWORDS

mental health; generative AI; generative artificial intelligence; medical curriculum; medical education

Introduction

At the beginning of the 21st century, the field of health care (HC) started to experience a cultural shift known as digital health. Digital health can be defined as the cultural transformation that occurs when disruptive technologies provide digital and objective data, accessible to both caregivers and patients, leading to an equal level of doctor-patient relationship with shared decision-making and the democratization of care [1,2].

Numerous studies have shown that the mental well-being of health care professionals (HCPs) and medical students (MSs) is poor, with high levels of stress, anxiety, and burnout being commonly reported [2-5], even though their welfare would increase the quality of care, productivity, and patient satisfaction [6]. Burnout usually increases gradually, with the highest levels typically occurring during the fourth year of medical school [7].

Studies suggest that artificial intelligence (AI) could bring relief to both the professional duties and academic challenges faced by HCPs and MSs by offering support in diagnostics, decision-making, data analysis, especially administrative tasks,

XSL•FO
RenderX

improved consultations, personalized health guidance, medical record analysis, precision medicine, and custom treatment plans [8].

Among various AI technologies, generative AI stands out due to its potential impact on the work of HCPs. Generative AI is the first form of AI to be used by hundreds of millions of people on a daily basis. It has attracted the attention of more than 100 million users in only 2 months [9].

The spread of generative AI has had a particularly significant impact on medical training, as it has fundamentally changed students' learning habits and called into question the legitimacy of traditional assessment methods [10-18]. MSs' experience these effects in diverse and highly individual ways.

Generative AI, represented by large language models (LLMs), has the potential to transform HC and medical education. In particular, the impact of generative AI on higher education can alter both students' learning experiences and faculty members' teaching practices [19].

Generative AI tools, especially LLMs such as ChatGPT (OpenAI), are designed to create new content from their trained parameters. With free access online and an easy-to-use conversational interface, ChatGPT quickly accumulated more than 100 million users within the first few months of its launch [20,21]. New opportunities for learning, assessment, and research in medical education have sparked excitement, alongside concerns about the implications of cheating and plagiarism in assessments [15]. Studies showed that these models excelled in many areas of medicine previously thought to be exclusive to humans, such as performing well on the United States Medical Licensing Examination, taking a patient history to build a differential diagnosis, diagnosing complex cases, and accurately answering patient questions. These preliminary investigations have shown that generative AI systems have the potential to be generalizable throughout many aspects of medical education, clinical decision-making, and the HC workflow [16]. Concerns have been raised about ethical considerations and the decreased reliability of the existing examinations. Furthermore, in medical education, curriculum reform is required to adapt to the revolutionary changes brought about by the integration of generative AI into medical practice and research [19,22].

As future HCPs prepare for this new reality, they are already incorporating generative AI tools into their learning, which presents challenges for both students and educators, and it seems MSs show curiosity and openness to participate in training related to this field [22].

Both MSs and educators now require skills and competencies that were previously unnecessary. As more patients and HPs use AI-based tools, LLMs being the most popular representatives of that group, it seems inevitable to address the challenge of improving this skill [23]. How stakeholders will respond to this new situation remains to be seen.

The gap between what technology could do and what technology is actually being used for is rapidly widening. While many solutions are proposed to address this gap, clinician resistance to the adoption of AI remains high [24].

We have already demonstrated that although AI can support HCPs in numerous ways, many of them still approach it with fear and perceive AI as a potential threat rather than an opportunity. Their concerns include fear of job displacement, replacement, erosion of professional identity, and loss of expertise, losing control to AI, leading to deskilling and overreliance on technology, and they also dread the potential negative impact on doctor-patient relationships. It seems that the key to helping them successfully implement medical AI into their everyday practice is to address HCPs' and MSs' concerns, which requires comprehensive training, ethical guidelines, and ongoing support [25].

With this study, we aim to explore possible futures of medical education with a particular focus on the mental health of MSs. We examine how the mental well-being of MSs may be affected by changes resulting from the integration of AI into medical education. We do that in order to help stakeholders successfully adapt to these changes and effectively tackle the new challenges they face. Instead of predicting a single outcome, we explore how different tech and support setups influence education and career growth. As a foresight study, we do not test hypotheses; instead, we build multiple futures to reduce bias and guide strategy. To assess this, we apply a futures research method, scenario analysis (SA), a well-known method in foresight. It helps to discover possible future scenarios based on key driving forces and major uncertainties.

Methods

Overview

This study uses an exploratory foresight design based on SA instead of predictive modeling. SA was chosen because rapid generative AI development and complex mental health dynamics make linear forecasting unsuitable. It identifies key drivers and uncertainties to create coherent, plausible futures that aid strategic reflection rather than prediction.

Exploratory Scanning

This study used a qualitative foresight methodology to explore the potential futures of medical education, with a focus on MSs' mental health and the integration of generative AI. SA is rooted in the structured construction of unique and complex scenarios, and allows organizations to build coherent narratives about possible developments. A scenario is a "story" or a narrative illustrating visions of possible futures.

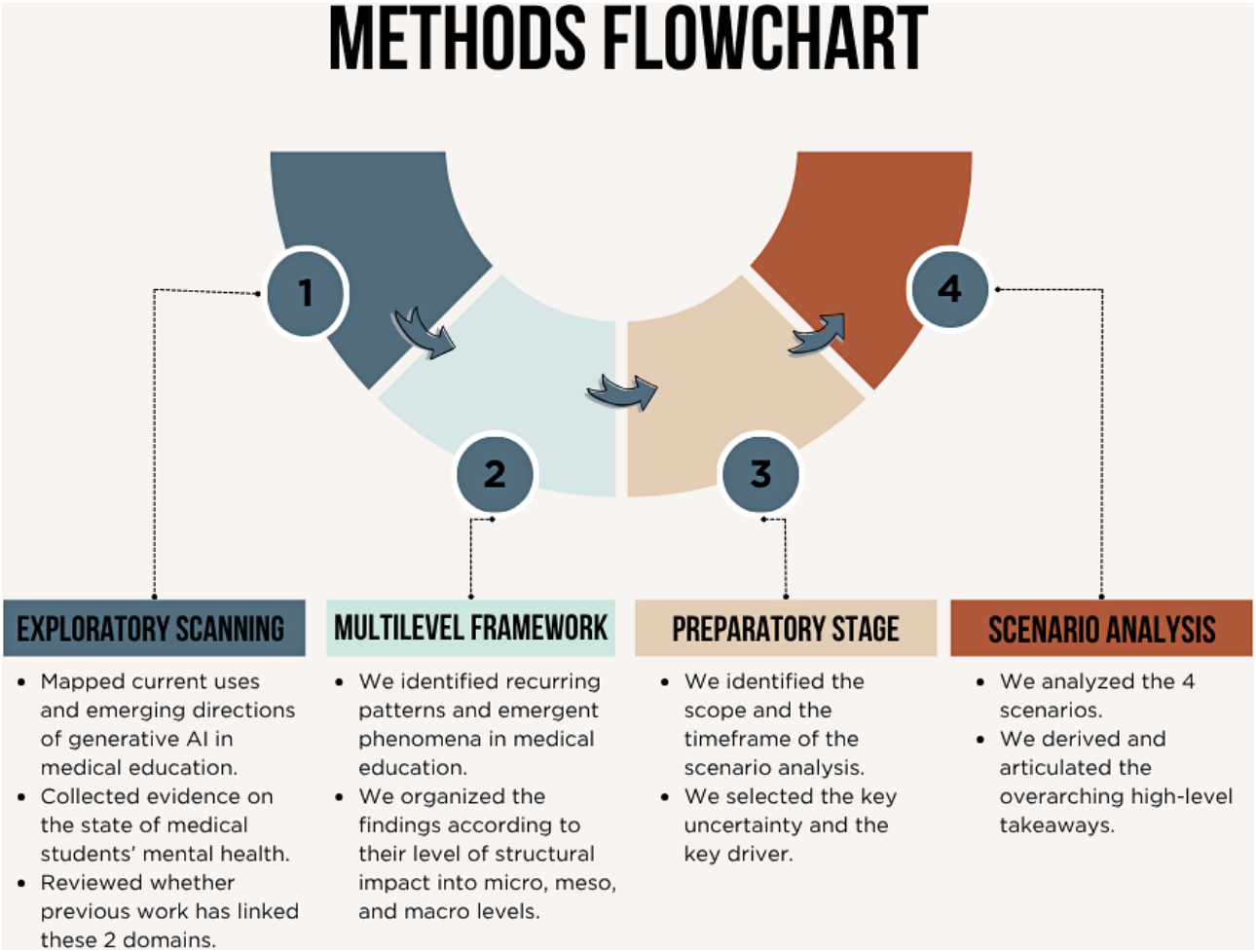
This approach diverges from the concept of a single "most likely future" and instead considers multiple alternative futures.

The scenario approach, originating from the idea of a "script," can be likened to versions of literary works adapted for stage or screen. Much like a repertory theater, where it is unclear which play will be performed, we prepare for multiple possibilities and interpret the situation from subtle cues to determine which one is unfolding. In essence, scenario building is a way of rehearsing the future in advance. For scenarios to be effective, they must be believable, coherent and logical, and insightful about the future, so they can contribute to decision-making [26].

XSL•FO
RenderX

The entire methodology process is shown in Figure 1 below. The figure outlines the methodological steps of this foresight study, including environmental scanning, identifying key phenomena impacting medical students' mental health and generative AI integration, selecting critical uncertainties, and developing 4 future-oriented scenarios.

Figure 1. Flowchart of the scenario analysis design used to examine the future of medical education. AI: artificial intelligence.



The research process began with structured exploratory scanning, informed by targeted searches of the PubMed database.

This exploratory scanning phase was not designed as a scoping or systematic review, but as a literature-informed environmental scan consistent with foresight and horizon-scanning methodology. Environmental scanning is the process of systematically exploring and interpreting relevant external information to identify potential opportunities, threats, trends, and emerging issues that may affect future planning or decision-making [27]. Environmental scanning is widely applied in foresight research and HC strategy, especially when aiming to identify weak signals and emerging trends. Weak signals are ambiguous early indicators of possible but not certain changes that may become significant in the future [28].

Emerging trends, in contrast, are patterns of change that are becoming increasingly visible and coherent across time and domains. They have more support in the present and suggest a directional shift that is gaining momentum [29].

Environmental scanning provides a broader lens than formal reviews, making it well-suited for the exploratory nature of SA. For example, the World Health Organization's Global Health Foresight program uses environmental scanning to identify early indicators of potential health system disruptions.

To effectively navigate the existing literature, we formulated the following 3 guiding questions: What are the main applications and emerging directions of generative AI in medical education? What do we know about the global state of MS's mental health? And finally, have any studies previously examined the intersection of these 2 areas?

PubMed was used as an authoritative entry point to identify relevant peer-reviewed signals and recurring patterns, rather than to achieve exhaustive coverage or formal evidence synthesis. From these recurring patterns, we identified a set of emergent phenomena that illustrate early signs of transformation at the intersection of generative AI and student mental health in medical education. The identification and clustering of these phenomena followed a macro-meso-micro framework adapted from established conceptual models in health systems research [30]. This allowed us to organize the findings according to their level of structural impact, from broad systemic trends to institutional dynamics and individual-level experiences. Therefore, no formal inclusion or exclusion criteria, study appraisal, or qualitative coding procedures were used, since the

XSL•FO
RenderX

goal was to understand and construct the scenario rather than to classify evidence.

To ensure analytical robustness, we only selected phenomena that were supported by at least 3 independent and credible sources. In this study, we define phenomena as early, identifiable patterns and dynamics within medical education that reflect the emerging influence of generative AI. These are not merely isolated observations, but conceptually meaningful trends that have appeared across multiple credible sources.

Drawing on the logic of environmental scanning and futures studies, these phenomena represent early indicators, known as "weak signals," of potential systemic transformation. Each selected phenomenon points to an area where change is already underway and may accelerate.

By defining and clustering these phenomena, we aim to build plausible scenario narratives and anticipate future needs, risks, and opportunities. In total, we identified 5 macro-level, 5 meso-level, and 7 micro-level phenomena, which were categorized and analyzed based on their thematic focus and level of impact (macro, meso, or micro). For this, we used a multilevel conceptual framework. The identified phenomena, along with their brief descriptions and supporting references, are presented in Multimedia Appendix 1.

Multilevel Framework

A conceptual framework is a network of interlinked concepts that provides a comprehensive understanding of a phenomenon or problem by organizing and structuring related ideas and relationships in a coherent way [30]. Several studies in HC previously used this conceptual framework and have supported it with a micro-, meso-, and macro-level analysis, so we followed the same path [31-33].

To further structure and interpret the identified phenomena, we applied a multilevel framework, which conceptualizes sociotechnical change across 3 analytical levels: macro (landscape), meso (regime), and micro (niche). This approach draws on the foundational work of Geels and Schot: the typology of transition pathways [34], the articulation of multilevel dynamics [35], and a critical review and theoretical refinement of the multilevel framework [36]. Representative examples of phenomena identified at each analytical level illustrate the breadth of the scanning process. On a macro scale, broad digital transformation, the swift spread of generative AI, and the global mental health crisis in medical training emerged as key structural influences. At the meso level, inconsistent integration of AI curricula, faculty resistance or adaptation, and differences in institutional mental health support reflected diverse organizational reactions. At the micro level, factors such as AI-related technostress, concerns about academic integrity, and hesitance to seek psychological help showed how systemic and institutional challenges impacted individual experiences.

Not all these phenomena became core dimensions; only those with high impact and variability across different settings were combined into the 2 main dimensions used for scenario development.

Preparatory Stage and Scenario Analysis

This approach allowed us to map the landscape of relevant research with clarity and accountability, serving as a robust foundation for subsequent SA. This structured and multivalidated selection process provided a strong foundation for identifying key trends, driving forces, and critical uncertainties used in our SA [37]. To create a detailed scenario analysis, we followed 5 well-defined steps.

First, in the preparatory stage, we defined the scope and time horizon. The scope is the transformation of medical education driven by generative AI and the mental health status of students. Given the rapid pace of technological development, we set a time horizon of 5 years.

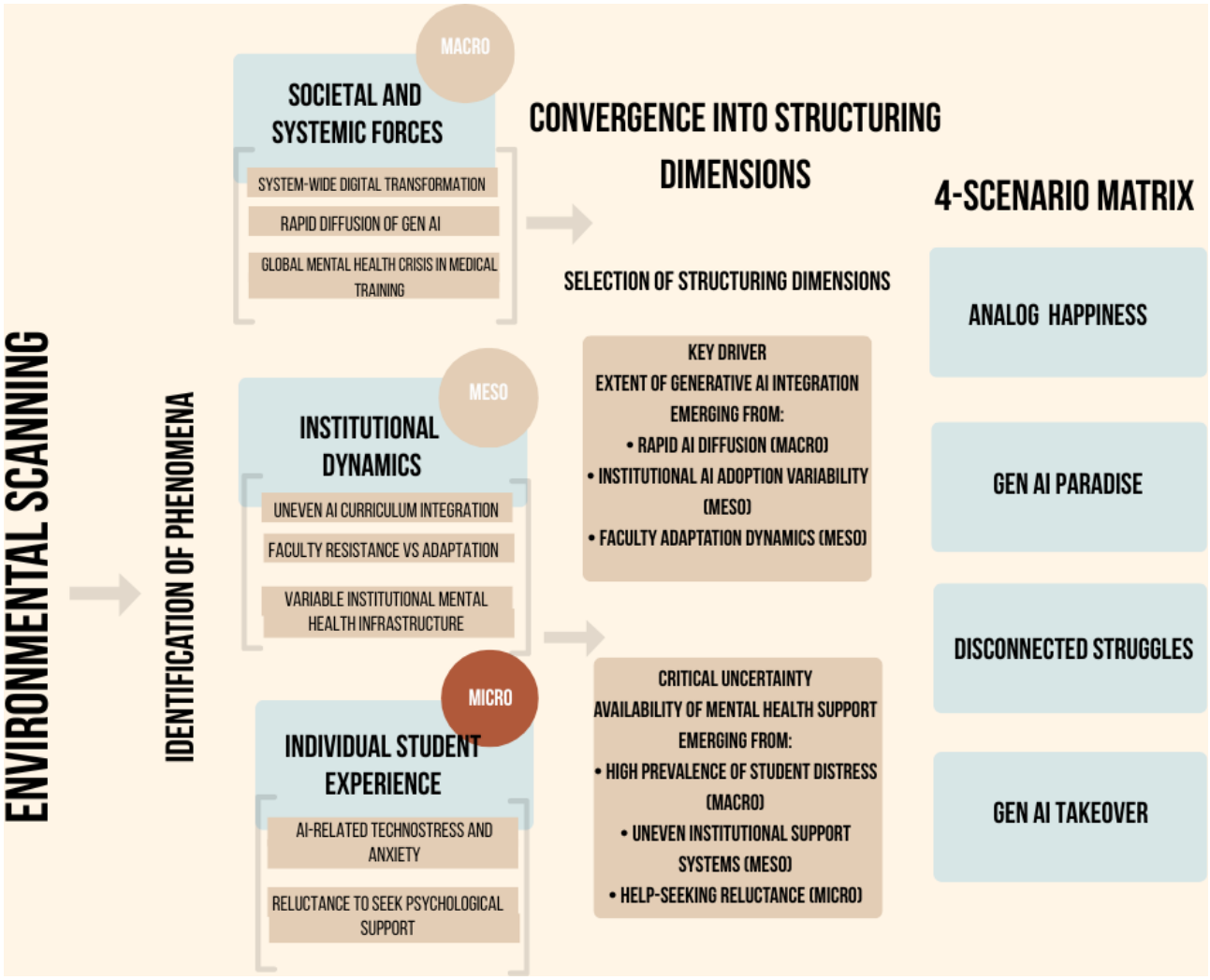
Second, we identified key drivers and uncertainty. Key drivers are the primary factors or forces that have a significant influence on the outcome of future developments, such as technological advancements, economic trends, societal changes, environmental conditions, and political shifts. Uncertainties are elements or events that are unpredictable and have the potential to significantly impact future outcomes, as the direction or impact is unknown.

We analyzed possible scenarios, and while doing so, we assigned descriptive titles to each of them. We created 4 distinct scenarios by combining the key drivers and uncertainty. After all that, in the fourth step, we analyzed the implications of each scenario and identified opportunities and threats associated with each.

Figure 2 depicts the complete analytical process, demonstrating how exploratory environmental scanning identified phenomena at macro, meso, and micro levels. These findings were then combined into 2 main structuring dimensions: the degree of generative AI integration (a key driver) and the availability of mental health support (a critical uncertainty). This merging laid the groundwork for the 4-scenario matrix. To strengthen plausibility, the selection of the key driver, critical uncertainty, and the resulting scenario structure were reviewed by 2 subject matter experts in foresight and digital health.

XSL•FO
RenderX

Figure 2. This foresight study on medical education uses an analytical framework for scenario construction. AI: artificial intelligence.



Ethical Considerations

This study did not involve human participants or personal data, so no ethics approval was needed. It used a qualitative review of publicly available literature from PubMed, with all sources properly cited and ethically compliant. The study adheres to the Declaration of Helsinki and JMIR publication standards.

Results

Exploratory Scanning

Out of the 3 guiding questions we formulated in the Introduction section, the results for our first guiding question regarding AI applications in medical education reflect a well-researched and rapidly evolving field. Yet, most papers highlight the need for further research to explore the effectiveness of generative AI applications in medical education [13].

Generative AI is increasingly used to support MSs' learning journeys, providing personalized experiences and improved outcomes.

Generative AI, such as ChatGPT, is accessible and transforming medical education [38]. Boscardin et al [15] propose an AI literacy framework to guide educators in integrating AI into admissions, learning, assessment, and research, helping them adapt to this new environment. As generative AI tools continue to expand, educators need to increase their AI literacy through education and vigilance regarding new advances in the technology, serving as stewards and coaches of generative AI literacy to foster social responsibility and ethical awareness around AI use.

Preiksaitis and Rose [14] conducted a scoping review; their thematic analysis identified diverse AI applications in medical education, such as self-directed learning, simulations, and writing help. They highlighted challenges such as academic integrity, data accuracy, and learning detriments. They proposed 3 research areas: developing critical evaluation skills, rethinking assessments, and studying human-AI interactions [39].

Chan and Zary [39] identified 3 main uses of generative AI in medical education: learning support (n=32), student assessment (n=4), and curriculum review (n=1). Its main benefits are feedback, guided learning, and reduced costs. The subgroup analysis shows that medical undergraduates are the primary audience. Additionally, 34 articles discuss challenges, mainly assessing the effectiveness of AI and technical development issues in applications.

The increasing use of generative AI in medicine has raised ethical concerns, such as patient autonomy, bias, and

XSL•FO
RenderX

transparency. Many of the recent studies suggest a need for teaching AI ethics as part of medical curricula [40].

The second guiding question we asked was what we know about students' global mental health. This also proved to be an area of strong scientific interest.

Nair et al [3] found that students' mental health concerns manifest in various forms. MSs facing academic pressure, debt, sleep deprivation, illness exposure, and mistreatment may explain the higher rates of disorders. These difficulties are dynamic, with certain challenges becoming more prominent during key student transitions in education. Primary examples include the entry year of medical school, the shift from preclinical curriculum to clinical training, and the final moments prior to beginning residency. The authors concluded that COVID-19 worsened stress, anxiety, and depression in medical education. They highlight solutions such as crisis management training, peer support, destigmatizing mental health, and better access to resources, emphasizing that it is up to medical education leaders to implement them.

Sleep deprivation is a major concern, with many medical students experiencing sleep problems that affect their academic performance and mental health. Perotta et al's [41] study found that students with more sleep deprivation had higher odds of anxiety and depression and lower odds of good quality of life or positive educational perception environment.

According to Backović et al [4], medical studies bring many stressful activities to students. Prolonged stress can have adverse effects on mental health and lead to further professional burnout. Studies across different countries have consistently demonstrated high rates of several psychiatric disorders among MSs.

Studies suggest that mental health worsens after students begin medical school and remains poor throughout training. On a personal level, this distress can contribute to substance abuse, broken relationships, suicide, and attrition from the profession. On a professional level, studies suggest that student distress contributes to cynicism and may subsequently affect students' care of patients, relationships with faculty, and ultimately the culture of the medical profession [42].

According to Watson et al [43], MSs' suicide likely relates to social, environmental, and endemic factors such as curricula, accommodations, support, pressures, social isolation, competition, early separation, simulation, cadavers, and trauma in training. The sociocultural environment, including ragging, expectations from teachers and patients, may pressure vulnerable students. Higher rates among females raise questions about gender roles and expectations in medicine.

Worldwide, medical schools have responsibilities to respond to concerns about students' psychological, social, and physical well-being, but guidance for medical schools is still limited [44]. Depression among MSs may be undertreated. Medical schools can assist depressed students by addressing issues such as the stigma of using mental health services, confidentiality, and documentation. Early treatment of impaired future caregivers may have far-reaching implications for the individual students, their colleagues, and their future patients [45].

Despite potential issues, MSs often avoid professional help to protect their identity. They believe medical students should be invulnerable. Hankir et al [5] note that MSs and doctors rarely seek help until a crisis, with stigma fears hindering access to mental health services.

Autobiographical narratives of the "Wounded Healer" are increasingly popular among MSs and doctors with mental health challenges, serving as both adjunctive therapy and antistigma campaigns. A randomized controlled trial of Coming Out Proud with mental illness showed immediate positive effects on stigma and stress variables [46].

Dederichs et al [22] highlight that many students with mental health issues do not seek help, creating a need for easily accessible interventions. Several evidence-based internet- and mobile-based interventions (IMIs) have been suggested, and researchers are interested in students' views on them. Students appreciated IMIs for their easy access, flexibility, ability to reduce waiting times, and as an initial step before in-person therapy. Nonetheless, medical students favored face-to-face treatment for serious mental health issues. Their findings show generally positive attitudes toward IMIs for mental health promotion, though concerns remain about their effectiveness for severe disorders and emergencies. These results suggest that IMIs could be valuable tools for stress prevention and early intervention among students. It is also noteworthy that students specifically preferred certified IMIs recommended and provided by trusted sources such as universities.

The growing presence of medical influencers on platforms such as TikTok (ByteDance) and Instagram (Meta Platforms, Inc) has introduced a new kind of voice into conversations around mental health in medical education. Although they are not necessarily tied to formal institutional structures, they are still part of the professional community, speaking both to the public and to insiders, such as MSs and residents.

Their content extends beyond personal stories to shaping norms. It addresses mental health training, maintaining well-being, and institutional expectations, influencing students' views on education and their future profession. In this way, they contribute indirectly to the broader culture and practices of medical institutions. Their influence on individuals is significant, resonating emotionally with students through messages such as "you're not alone" or "I've been through this too," which offer support and practical advice. Many find this more authentic than official messages. These influencers act as digital boundary spanners, bridging the gap between the system and people, operating between institutional and personal levels. While writing this, Jake Goodman boasts 1.2 million followers on TikTok; his top video has reached 2.1 million viewers. He has spoken at the White House and given a TED Talk. He publicly addresses the mental health crisis in medical training, emphasizing the high depression rates among students, the ongoing 24-hour shifts, and how the fear of retaliation discourages open discussion and meaningful reforms.

On Instagram, numerous coaches present themselves as business mentors, specifically teaching HCPs how to avoid burnout by working part-time and building a side hustle. Numerous influencers similarly focus on helping HCPs create and sell

XSL•FO
RenderX

online courses, also as a strategy to reduce workload, increase income, and prevent burnout.

This phenomenon is gaining acceptance and offers an alternative for HCPs seeking balance and sustainability. However, the profession faces a severe workforce shortage, increasing pressure on those in clinical roles. The system must train and retain more physicians or make the profession more appealing by creating humane working conditions and restoring purpose and dignity, starting from medical training. Medical AI offers a promising opportunity in this regard, as discussed in our previous article [25].

The result of searching for the third guiding question—the intersection of the previous 2 topics—does not appear to be a particularly well-explored area. Its investigation remains largely open, offering an opportunity for original contribution. It is, however, worth noting a few preliminary mentions in the literature.

The findings of Li and Quin [11] show that medical postgraduate students are more aware of generative AI than undergraduates. Their intention to use it links to performance expectancy, habit, hedonic motivation, and trust. Medical education should focus on enhancing training performance with courses that are easy and engaging to equip students [11].

Thesen et al [47] introduced a precision well-being framework for medical education, based on a comprehensive view of mental health, combining mental health and positive psychology measures in a data-driven way. Using unsupervised machine learning on data from 3632 MSs, clusters were formed based on depression, anxiety, and flourishing metrics. Three distinct clusters were identified. Membership in the "Healthy Flourishers" well-being phenotype was associated with no signs of anxiety or depression while simultaneously reporting high levels of flourishing. Students in the "Getting By" cluster reported mild anxiety and depression and diminished flourishing. Membership in the "At-Risk" cluster was associated with high anxiety and depression, languishing, and increased suicidality. Nearly half (49%) of the medical students surveyed were classified as "Healthy Flourishers," whereas 36% were grouped into the "Getting-By" cluster and 15% were identified as "At-Risk." Findings show that a substantial portion of MSs report diminished well-being during their studies, with a significant number struggling with MH challenges.

Thesen et al's [47] precision well-being framework is an integrated model that classifies medical students into well-being phenotypes based on their mental health. It supports student aid and evaluates personalized interventions, illustrating the role of AI in assisting MS.

While reviewing the literature, we found only 2 studies that use SA to explore the future of medical education, and one that used scenario planning. The first study used ChatGPT to develop the scenarios, the second examined the future of midwifery practice in Antwerp, and the third aimed to explore the potential of scenario planning to bridge the understanding gap and frame strategic planning for interprofessional education and practice, as well as to implement innovative techniques and technologies for large-group scenario planning.

Knopp et al [48] used generative AI to create 4 scenarios: AI Harmony, AI Conflict, Ecological Balance, and Existential Risk. Risks include disinformation, privacy loss, inequality, reduced autonomy, and ethical issues. Benefits are efficiency, personalized interventions, better collaboration, early detection, and faster progress in research.

The second study explored the future of midwifery praxis using structuration theory and intuitive logics scenario planning methods to structure contextual midwifery scenarios. Three scenarios described the plausible future of midwifery: (1) midwife-led care monitoring maternal health needs, (2) midwife-led holistic care, and (3) midwife- or general practitioner–led integrated maternity care [49]. In the third study, a full-day scenario planning workshop was held, introducing an innovative methodology to 71 participants from 9 universities, service providers, government, students, and consumer organizations. Outcomes were assessed via statistical and thematic analyses of a mixed methods survey. They concluded that the scenario planning method can be used by tertiary academic institutions as a strategy in developing, implementing, and embedding interprofessional education, and for the enculturation of interprofessional practice in practice settings [50].

Phenomena Identified

After examining study approaches to SA in medical education, we synthesized findings to identify underlying forces shaping the field. Moving from individual studies to broader patterns, we classified phenomena into micro, meso, and macro levels using a conceptual framework. At the macro level, societal forces such as technological pressures and shifting expectations influence medical education but are beyond individual control, shaping what is seen as possible in the future of training. The meso level is the institutional level where most structural changes happen, focusing on medical faculties and educational systems. It examines how universities respond to generative AI and its impact on students, highlighting institutional resistance or adaptation and how AI influences professional development norms.

In our previous work, we summarized the currently available AI-related courses [25]. An important part of these changes is the growing openness around MH: topics once taboo are now openly discussed, often promoted by physician-influencers on TikTok and Instagram. While macro-level shifts enable this, the institutional embrace belongs to the meso level.

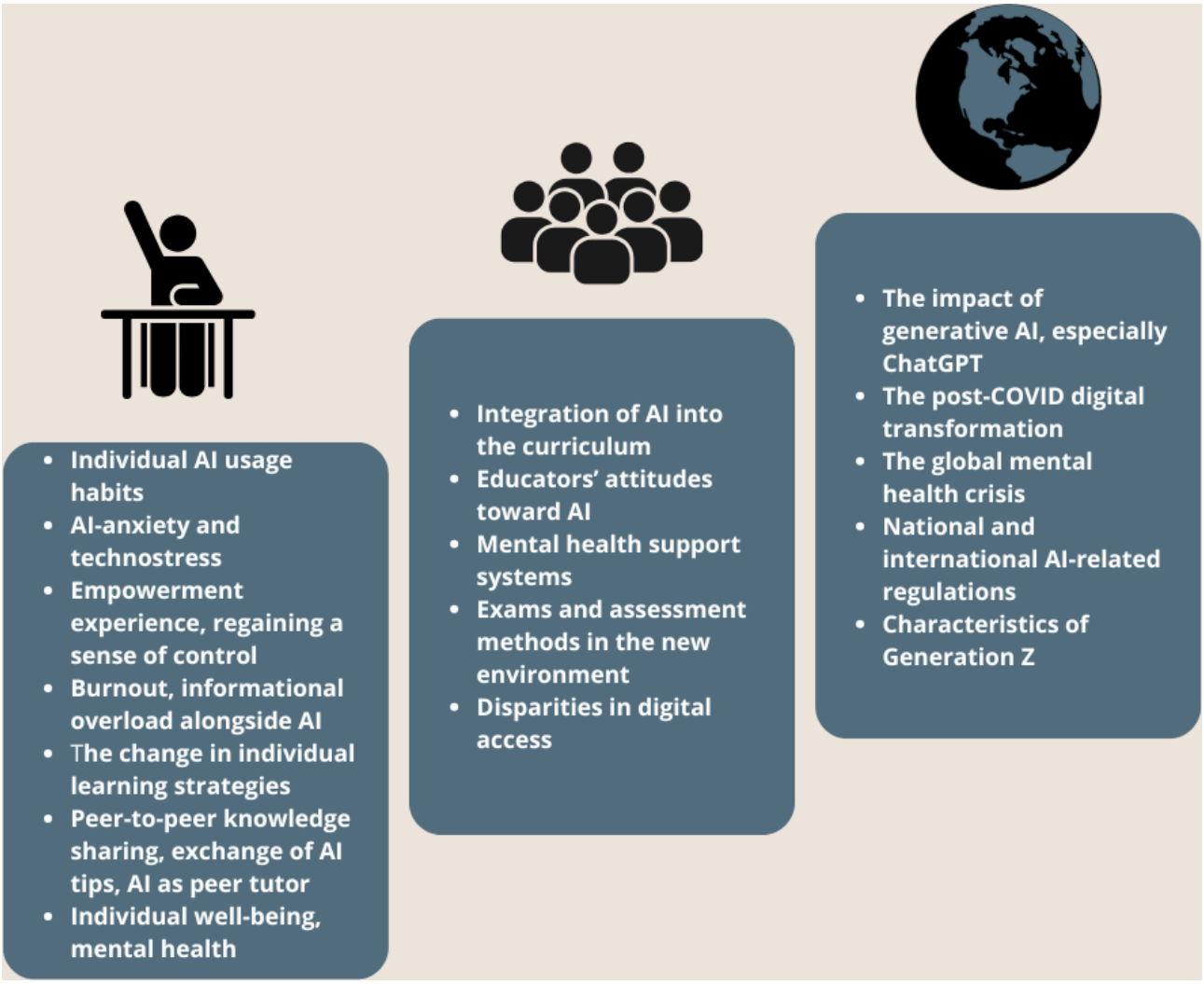
At the micro level, we see the most immediate and personal impact of AI, as experienced by individuals. This includes the tools that MSs use in their daily routines, changes in their learning strategies, and, crucially, their mental health. It also involves the coping mechanisms they adopt and how they experience the medical school years on a personal level. A summary table of the phenomena, along with the list of supporting articles, can be found in Multimedia Appendix 1. Figure 3 provides a summary of the phenomena occurring at the 3 levels. The framework identifies micro (individual), meso (institutional), and macro (societal) levels that influence the future of medical education, especially regarding the adoption

XSL•FO
RenderX

of generative artificial intelligence and mental health issues.

AI: artificial intelligence.

Figure 3. A multilevel conceptual framework is used to categorize factors affecting medical students in this foresight study. AI: artificial intelligence.



This validated selection process laid a strong foundation for SA, identifying key trends, drivers, and uncertainties. Grounded in current trends and framework phenomena, we analyzed and defined the scope, time frame, key drivers, and main uncertainties to develop the scenarios. All 4 scenarios share a common scope: the transformation of medical education driven by generative AI and the mental health status of students. Given the rapid pace of technological development, we have set a 5-year time horizon, projecting the scenarios to the year 2030.

This time horizon reflects both the rapid evolution of generative AI technologies and the urgency of mental health challenges among MSs. Innovations such as multimodal AI agents, personalized learning platforms, and AI-assisted assessment tools are already being piloted in several institutions, suggesting that large-scale educational disruption is not only possible but imminent [23]. Simultaneously, student mental health has become a widely recognized crisis requiring near-term interventions [2-5,46,51]. Therefore, we chose a 5-year time frame that offers both relevance and feasibility for building the scenarios.

Driver and Uncertainty Selection

We identified the extent of generative AI integration into curricula as the key driver and access to mental health support as the key uncertainty. These emerged from the multilevel conceptual framework as the phenomena with the most far-reaching impact, best describing and shaping the field.

The key drivers and uncertainties were chosen because they both represent high-impact factors. The first involves technological feasibility, faculty acceptance, ethics, and curriculum redesign. The latter reflects variability in institutional capacity, stigma, funding, and cultural attitudes toward psychological care. Our choices were based on the conceptual framework and supported by at least 3 high-quality sources each, showing their significance as system-wide phenomena that influence.

This methodology serves to map the evolving landscape and provide a coherent foundation for building plausible scenarios. To identify the key driver and critical uncertainty, candidate dimensions were evaluated based on three criteria: (1) their ability to integrate multiple phenomena across macro-meso-micro levels, (2) their structural impact on the future of medical education, and (3) their variability across

XSL•FO
RenderX

different institutional settings. Dimensions supported by at least 3 independent, high-quality sources were prioritized. As a result, the level of generative AI integration was selected as the key driver, while the availability of mental health support was identified as the critical uncertainty, reflecting the tension between rapid sociotechnical change and limited psychological support in medical training. This final choice was reviewed by an independent expert to strengthen methodological credibility.

Scenario Descriptions

Scenario 1: Analog Happiness

- Key driver: integration of generative AI into medical curricula is low.
- Uncertainty: access to the MS support system is high.

In the first scenario, there is high access to the MS support system, but the integration of generative AI into medical curricula is low.

We named this scenario "Analog happiness." Institutions prioritize the mental health of MSs. Inspired by research highlighting the importance of accessible mental health services, universities invest in comprehensive support infrastructures. Mental health is openly discussed, and stigma around it has significantly decreased, influenced by broader cultural shifts and the visibility of physician-influencers advocating for the mental well-being of MSs [52]. There is a strong institutional focus on supporting MSs' psychological well-being, aiming to address diverse needs through professional and peer-led support, as shown by Hale and Davis [52].

Their study found a peer-led, virtual workshop series effective in reducing stress, anxiety, and depression among MSs. Psychiatrists providing mental health treatment to MSs recognize the importance of managing transference, countertransference, overintellectualization, and occurrences of strong idealization identification [51].

However, the use of generative AI remains marginal. Faculties are hesitant, citing ethical concerns and a lack of guidance [53]. Curricula rely on traditional methods. MSs' use AI informally, but this use is inconsistent and sometimes ethically ambiguous [21]. Without clear institutional direction, the inappropriate usage of generative AI leads to missed opportunities in personalization and innovation.

Students may be emotionally supported but lack digital readiness. Educational inequities may widen between digitally progressive and resistant institutions. Alkhaaldi et al's [21] study found that surveyed MSs had little experience with AI tools and limited perceptions of AI in health care, but they generally viewed ChatGPT and AI positively and were optimistic about the future of AI in medical education and HC. With structured curricula and formal policies and guidelines, it would be possible to adequately prepare MSs for the forthcoming integration of AI in medicine, so this scenario is the future of missed opportunities.

Scenario 2: Gen AI Paradise

- Key driver: integration of generative AI into medical curricula is high.
- Uncertainty: access to mental health support is high.

In the second scenario, both access to mental health support and integration of generative AI into medical curricula are high. We referred to this scenario as "Gen AI Paradise." This scenario represents a future in which medical education institutions have both embraced generative AI and prioritized student mental well-being.

AI tools are deeply integrated into all levels of the curriculum, guided by international best practices and ethical standards [19,54]. Students learn to use generative AI tools for research, simulation, and diagnostics, and critically reflect on their implications, building AI literacy.

Educators intentionally integrate AI into curricula based on research such as Shimizu et al [19], analyzing the impact of AI on medical education and suggesting reforms. The authors use SWOT (Strengths, Weaknesses, Opportunities, Threats) analysis and identify 3 main learning approaches: "learning about GenAI," "learning with GenAI," and "learning beyond GenAI." The study emphasizes the importance of teaching the ethical use of generative AI, the benefits of adaptive learning, and the continued value of nonsubstitutable skills such as communication and clinical experience [19]. MSs also use generative AI tools as virtual patient and clinical decision-making tutors [55]. Robust mental health services are tailored to MSs' psychological needs, differing from those of previous groups. Their design is based on research such as Duan et al [18], highlighting student-centered, stigma-free MH strategies in medical schools. They analyzed MSs' perceptions, trust, and attitudes toward generative AI in medical education, and explored their willingness to integrate generative AI in learning and teaching practices.

The majority of participants reported familiarity with generative AI concepts, whereas only 43.5% had an understanding of generative AI applications specific to medical education. In this ideal future, decision-makers also take into account age- and gender-related differences, ensuring that support systems and educational strategies are responsive to the diverse needs of students.

Postgraduate students exhibited significantly higher levels of awareness of generative AI tools in medical contexts compared with undergraduate students. Male students were more enthusiastic and engaged with AI than females. Females expressed more concerns about privacy, security, and ethics in AI for medical education. Male or postgraduate students showed stronger intentions to use these tools in future learning and teaching practices [18].

In this scenario, technostress and other generative AI–related anxieties are proactively addressed, and digital well-being is embedded into the culture of medical training. Faculty members are supportive and trained to use generative AI in ways that reduce, rather than increase, student burden. Educators are trained to guide students in ethical, effective generative AI use.

Simultaneously, mental health support is robust and student-centered. Further, generative AI is recognized as a tool that can enhance efficiency and involve patients in education. Stakeholders understand that medical education should focus

XSL•FO
RenderX

on competencies that generative AI hardly replaces, such as clinical experience and communication. Notably, involving both faculty and students in curriculum reform discussions fosters a sense of ownership and ensures broader perspectives are encompassed [19].

Technostress is mitigated through proactive digital well-being education [56]. Students report reduced anxiety, increased self-efficacy, and improved learning satisfaction. This scenario reflects a successful co-evolution of technology and care.

This model enhances academic efficiency and personalization while safeguarding emotional health. It may become the benchmark for 21st-century medical education.

Scenario 3: Disconnected Struggles

- Key driver: the level of generative AI integration is low.
- Uncertainty: access to mental health support is low.

In the third scenario, access to mental health support and the level of generative AI integration are low. We named this scenario "Disconnected struggles." In this scenario, institutions fail to adequately respond to either the emotional or technological challenges of contemporary medical education. Mental health services remain limited, fragmented, or stigmatized, even as students face rising levels of stress and burnout [57,58].

The pandemic-induced digital shift exacerbated isolation, yet long-term MH responses have been inconsistent [59,60]. Generative AI tools are largely ignored at the curricular level, with little institutional support and faculty lacking interest or digital fluency. Students attempt to use AI without guidelines, causing uncertainty, inconsistent results, and anxiety [58].

This creates a high-pressure, low-support environment where learners face digital and emotional disorientation. Students' well-being is declining, while academic inequalities are widening and trust in institutions is eroding. The lack of coordinated support during rapid digital transitions places students at risk of experiencing what might be described as "academic trauma," marked by confusion, anxiety, and a diminished sense of belonging within the educational environment. There is ambiguity around when students should use these tools, what constitutes plagiarism, and how to detect when text has been authored by generative AI. Most importantly, there is a lack of availability and experimentation, as well as a lack of competencies and guidance on how to incorporate the new and rapidly evolving world of generative AI and LLMs into the core medical school curriculum and experiences of medical education [16].

Even professors are confused, and they seem to be afraid of generative AI. Cervantes et al [61] state that faculty mainly worry about cheating (97%), errors in AI (95%), lack of context (86%), and reduced human feedback (83%). They cite the absence of safe use guidelines from the government and institutions. Key challenges include cheating, AI mistakes, and privacy and safety concerns. Those who themselves fear the technology are neither able to teach it effectively nor represent it credibly, and in this scenario, they do not even intend to.

Scenario 4: Gen AI Takeover

- Key driver: the level of generative AI integration is high.
- Uncertainty: access to mental health support is low.

In the final scenario, access to mental health support is low, but the level of generative AI integration is high. The name of this scenario is "Gen AI Takeover." This future envisions a system where technological innovation outpaces emotional preparedness, highlighting the risks associated with innovation without adequate support. Institutions push AI integration aggressively [62], embedding tools like ChatGPT into every learning domain.

Students gain AI fluency, but emotional burden grows. Both students and faculty are enthusiastic about generative AI in medical education. Students and trainees use AI for learning, clinical reasoning, and operational tasks, while faculty create curricula, assist teaching, and handle grading assessments [16].

Generative AI tools are thoroughly integrated into the curriculum, driven by institutional enthusiasm and a desire for global competitiveness. Students engage with AI-based tutoring, simulation, and assessment systems from their first year, benefiting from individualized feedback and enhanced learning pathways.

Yet, the emotional cost is high. Mental health services have not evolved in parallel, and students report high levels of technostress, AI anxiety, and burnout [63,64]. Without adequate mental health infrastructure, the fast-paced digital transformation leads to emotional fatigue, performance pressure, and an erosion of intrinsic motivation.

Coping strategies differ: some students thrive, others feel overwhelmed, unsupported, or disconnected. They may feel isolated and disoriented, turning to unregulated online tools or peer forums for support. Without structure, these increase stress, raising burnout and dropout rates. Without adequate mental health systems, generative AI–related anxiety and information overload escalate [65-69].

The curriculum emphasizes performance over well-being. The study by Sallam et al [70] shows that high generative AI integration without emotional support leads to increased anxiety, fear, mistrust, and ethical issues among medical students, highlighting the need for curriculum changes. Interventions should improve familiarity and skills with generative AI to reduce concerns and prepare future HCPs for this essential technology effectively. This study also highlights the importance of incorporating ethical discussions into medical courses to address mistrust and concerns about the human-centered aspects of generative AI.

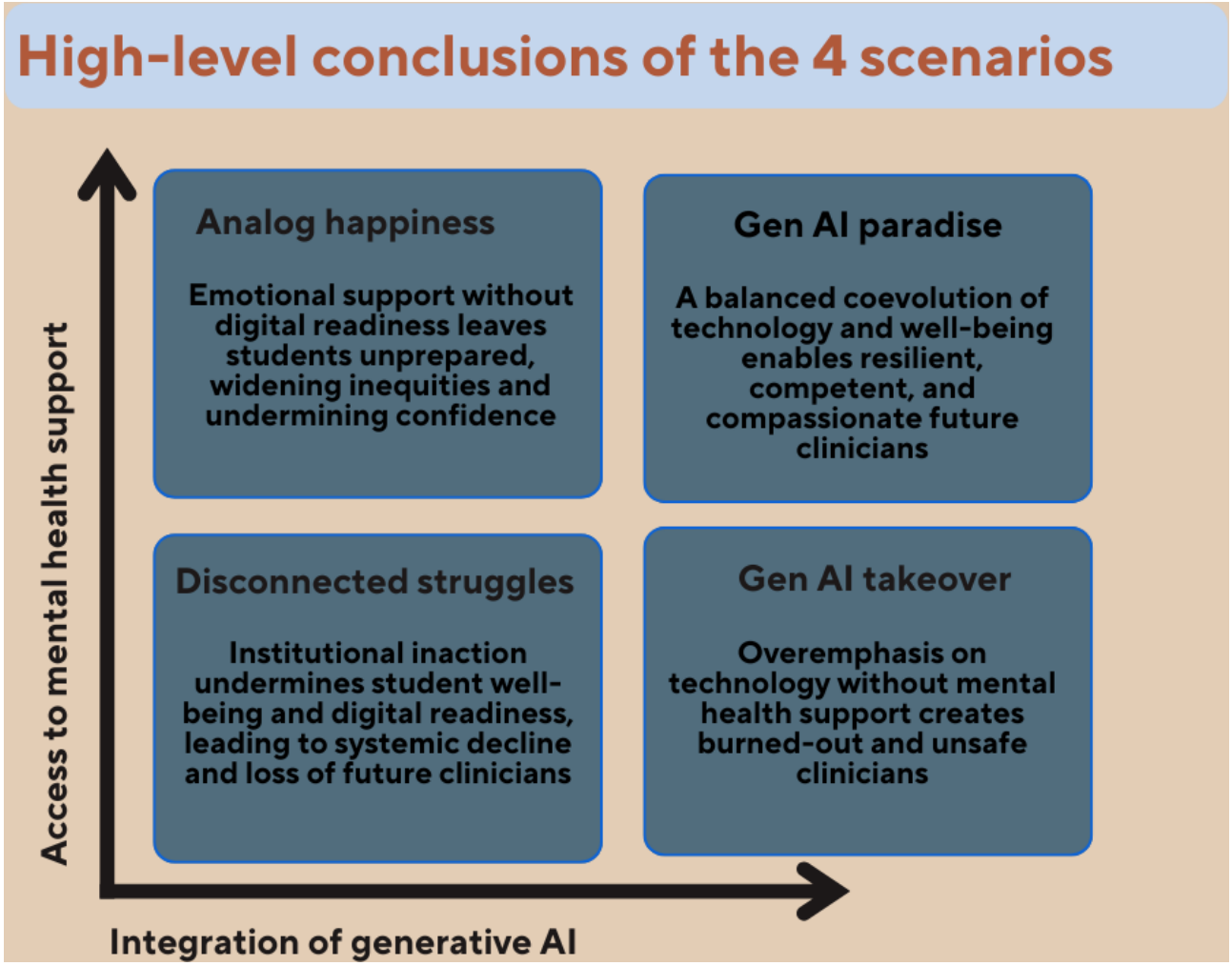
Their study calls for the proactive evolution of medical education to prepare students for new AI-driven HC practices, ensuring that physicians are well-prepared, confident, and ethically informed in their professional interactions with generative AI technologies [70]. We stress that the low psychological support here is not due to missing AI tools, but rather to a lack of support or its nonprioritization. AI is used for performance and productivity. While AI tools such as chatbots exist, evidence of their effectiveness in preventing

XSL•FO
RenderX

mental health issues is limited and mixed. Some studies show benefits; others raise concerns about safety and reliance. We do not assume AI alone guarantees better mental health. The scenario shows AI mainly enhancing performance, while mental health support still develops.

In this scenario, the education system is driven by efficiency but devoid of care. Students may become proficient but emotionally fragile practitioners. The 4 scenarios and their high-level takeaways are presented in Figure 4.

Figure 4. Four-scenario matrix illustrating possible futures of medical education for medical students. AI: artificial intelligence.



Summary of Key Insights

Over time, the scenarios could be assessed against a set of evolving indicators. On the macro level, shifts in national funding priorities or changes in AI regulation may signal broader structural trajectories. On the meso level, curriculum revisions, the adoption of new pedagogical and exam formats, and institutional data on student well-being can highlight how organizations adapt. At the micro level, longitudinal feedback from students, including patterns of AI use in everyday study practices and self-reported mental health trends, may reveal which scenario resonates most closely with the lived reality in the timeframe of our scope.

Discussion

Principal Findings

This study used scenario analysis to explore future medical education, focusing on AI integration and mental health support. To grasp the role of AI, we first examine the current landscape. Medical education faces 2 core challenges: training students to collaborate with generative AI and protecting their mental health, both crucial for learning and developing as clinicians. Our findings build on existing research that highlights the rising use of generative AI and mental health issues among MSs, while also emphasizing the often-overlooked connection between these areas.

The 4 scenarios depict potential futures of medical education based on institutions' responses to integrating generative AI and MSs' mental health. Each explores different balances, but the key insight remains: technological innovation must be paired with emotional support. Both are interdependent in developing resilient, future-ready HCPs. The "Gen AI Paradise" scenario illustrates the potential of a balanced future, where AI supports learning and efficiency without compromising the psychological well-being of MSs.

The scenarios highlight risks of unbalanced progress: rapid digital growth without mental health support or stagnation in both areas. "Analog Happiness" shows that even supportive

XSL•FO
RenderX

environments fail if MSs are not prepared for digital demands. They reveal that technological innovation and emotional support are interconnected in developing resilient, future-ready HCPs. The "Gen AI Paradise" scenario envisions a future where technological fluency is paired with emotional support and well-being. It exemplifies coevolution, with generative AI enhancing learning and diagnostic skills, while mental health services buffer training stress. AI is seen as a complementary, empowering tool, not a replacement. Institutions invest in infrastructure and culture to ensure students feel supported and involved in shaping their educational paths. This scenario offers a promising foundation for modern medical education, one that does not force a choice between care and competence, but weaves them together. This scenario serves as a call to restore the original contract of medicine, not as a choice between compassion and competence, but as a new lineage of healers who embody both. Modern medical education must remember what we already know, ever since Balint: the doctor is the medicine [71].

Healing is undoubtedly a profession and a scientific discipline; however, it extends beyond that, encompassing qualities such as compassion, empathy, and dedication. It can also be regarded as a calling. And the well-being of the healer is a key driver of both patient satisfaction and recovery [6]. The "Gen AI Takeover" warns of a technocratic future valuing efficiency over MS and educators' well-being, risking burnout, anxiety, technostress, and disengagement. It highlights societal risks of digital acceleration, lacking human focus, which could weaken future practitioners' emotional resilience and impact patient care.

The "Disconnected Struggles" scenario depicts institutional paralysis where digital and emotional challenges are ignored. MSs face a high-pressure environment with little support; the absence of AI and mental health resources causes stress, anxiety, mistrust, and trauma. Trauma may result in dropping out, career changes, low self-worth, self-doubt, or suicide. It highlights warning signs such as burnout, unprepared faculty, and ethical disputes in some institutions.

Meanwhile, the "Analog Happiness" scenario shows that a caring educational culture alone is insufficient. Without digital readiness and AI literacy, even the best systems fall behind. Mental health services cannot relieve the tension from unmet needs; they merely chase symptoms, fighting a losing battle. MSs in this scenario are emotionally supported but digitally vulnerable, widening institutional inequities and leaving graduates unprepared for the technological realities of modern clinical practice. There is a growing concern that MSs who are dissatisfied with the current level of generative AI integration will seek out training independently, outside of validated or regulated environments. This increases the risk of misinformation and errors, which may lead to guilt, despair, and a decline in their confidence and belief in their professional abilities.

Across all futures, a key insight is that technological preparedness must grow alongside emotional support systems. Focusing on one creates an unsustainable training environment. Institutions should see generative AI as a transformative force that reshapes learning, decision-making, professional identity, and the doctor-patient relationship. At the same time, the emotional landscape of medical education is changing, with MSs facing increased academic pressure and digital challenges overload. Failing to respond to these dual pressures risks creating emotionally exhausted, uncertain, or digitally unprepared physicians.

Additionally, the scenarios emphasize the importance of faculty readiness and institutional clarity. Without proper training and policy frameworks, even the most promising technologies may create confusion or misuse. Educators themselves must be emotionally equipped and technologically literate to guide students through this new landscape.

The scenarios emphasize the need for inclusive, participatory curriculum reform by involving students and faculty in integrating generative AI and mental health strategies to ensure relevance. Future policies should explore hybrid approaches combining foresight.

This study offers 3 contributions: applying structured scenario analysis to the understudied intersection of generative AI and medical students' mental health, separating technological integration from psychological support, and using a macro-meso-micro perspective to link systemic changes with individual experience.

Limitations

This study has limitations. First, scenario analysis is exploratory, not predictive; the scenarios are structured narratives, not forecasts. They aim to stimulate decision-making but cannot capture all future complexities.

Second, the authors mainly developed the scenarios, risking subjective bias despite independent review. Scenario development inevitably involves interpretive judgment. Although a multidisciplinary research team and external expert review helped mitigate bias, alternative drivers or uncertainties could be identified by other researchers.

Third, the scenarios were developed through desk-based, literature-informed synthesis without participatory foresight workshops, which may limit the range of stakeholder perspectives represented. The literature base we used, although broad, is still limited by language, as we relied solely on English sources, and by the rapidly evolving nature of research on both generative AI and student mental health. At the time of our data collection, some weak signals may not yet be visible in the literature or may have emerged later.

Fourth, we set a 5-year horizon for relevance, but this limits exploration of longer-term systemic changes such as cultural shifts in medical education or health system restructuring.

Fifth, we highlighted generative AI and mental health as central dimensions, but other key drivers, such as global health inequities, climate change, financing models, and interprofessional education, were not included in our scenarios. Future research could expand to these systemic drivers.

Our study focuses on medical education in higher-income countries with better digital infrastructure and mental health

XSL•FO
RenderX

access. Findings should be cautiously applied to low- and middle-income countries, where conditions differ.

Conclusion

Medical education seems to be at a crossroads. Institutions should not see technological and emotional development as competing priorities, but as mutually important pillars of responsible and future-ready education. In Figure 5, we present our practical advice to students, educators, and stakeholders based on different levels of generative AI adoption and mental health support, highlighting potential impacts and strategic considerations for medical students, educators, and institutional leaders.

Figure 5. Practical Implications of the 4 future scenarios for key stakeholders in medical education are discussed. AI: artificial intelligence.



Ensuring ethical AI use and investing in accessible, stigma-free mental health services are essential. To prepare future HCPs, stakeholders should develop both care and innovation, helping students become digitally fluent and emotionally resilient. Ignoring either the psychological or digital readiness of future HCPs is not a neutral act; it has ethical and professional implications. Institutions must fulfill this dual responsibility, or they risk fostering burnout rather than excellence. This is in the best interests of all stakeholders, educators, MSs, HCPs, and patients, as the mental well-being and technological preparedness of HCPs are critical determinants of clinical performance and, consequently, patient outcomes [6].

Speaking of educators, while the scenarios were student-centered, generative AI also transforms faculty roles. Medical educators face professional and emotional impacts. As students increasingly use AI tools in their studies, the faculty's role shifts from being the main knowledge source to facilitating learning, curating resources, fostering critical thinking, and promoting engagement. This change affects methodology, autonomy, assessment practices, and the broader medical education context.

Across 4 scenarios, each future has a key risk and institutional responsibility. In Analog Happiness, the main risk is missed opportunities because students are emotionally supported but unprepared for clinical practice due to a lack of AI integration. Institutions should include AI literacy in the curriculum, provide ethical guidance, and ensure transparent AI use education.

In Gen AI Paradise, the main challenge is sustainability, not imbalance. Even with systems where innovation and mental health support grow together, efficiency gains can hide psychological strain. Institutions should implement long-term governance, train faculty, and monitor student well-being to maintain this balance over time.

In the "Disconnected Struggles" scenario, the main risk is academic trauma from the lack of mental health support and institutional guidance on generative AI. Students face high expectations, uncertainty, and digital change alone. Action

XSL•FO
RenderX

should focus on restoring safety and trust via accessible mental health services and clear rules on ethical AI use.

The "Gen AI Takeover" risks creating skilled but emotionally vulnerable graduates as AI's rapid integration outpaces support systems. This increases technostress, burnout, and performance pressures, threatening learning and clinical practice. Risks can be mitigated by strengthening mental health support and embedding psychological safety into AI-based education, assessment, and feedback. Figure 6 summarizes the main risks of each scenario and provides initial suggestions for institutions as immediate first steps, highlighting key systemic risks associated with generative AI use and the mental health of medical students, along with possible immediate actions for educational institutions. In the visual framework, Analog Happiness is symbolized by a compass indicating direction, "Gen AI Paradise" by a calibration dial indicating balance, "Disconnected Struggles" by a lifebuoy indicating safety, and "Gen AI Takeover" by a brake indicating constraint.

Figure 6. Outline of risks and recommended institutional responses across 4 scenarios in this foresight study. AI: artificial intelligence.

SCENARIO	MAIN RISK	FIRST INSTITUTIONAL ACTION
ANALOG HAPPINESS	Strong psychological support combined with low formal AI integration produces digitally underprepared graduates and encourages unregulated, informal AI use.	Implement structured AI literacy modules that include supervised clinical simulations, ensuring they are integrated into core medical training in an ethical, transparent manner, and aligned with assessment standards.
GEN AI PARADISE	Even within a balanced system, improvements in efficiency might hide the overall cognitive burden and minor types of technostress.	Establish continuous dual-monitoring systems that link AI-driven workflow changes to longitudinal mental health metrics, enabling early detection of hidden strain.
DISCONNECTED STRUGGLES	The lack of both structured AI integration and accessible psychological support leads to institutional ambiguity, which heightens risks of disengagement and academic trauma.	Simultaneously set minimum AI-use guidelines and create quick-access mental health pathways to enhance structural clarity and ensure psychological safety.
GEN AI TAKEOVER	Intensive AI integration without parallel psychological safeguards amplifies technostress, AI anxiety, and performance pressure, thereby increasing the risk of burnout.	Implement compulsory digital well-being education that includes essential training on the safe and appropriate use of AI-based mental health tools, along with improved access to human psychological services.

The presence of generative AI also alters the perception of professional authority. When students have instant access to AI-generated answers, the instructor is no longer the sole source of expertise. For some educators, this can feel like a threat, while for others it opens the possibility of highlighting what remains distinctly human in teaching. Such a shift, however, requires careful support, as moving from a position of authority to one of partnership can be experienced as both professionally and personally destabilizing.

Generative AI exposes gaps in digital literacy. Students adopt new tools quickly, but faculty vary in confidence and acceptance. Discrepancies threaten educational quality, highlighting the need for structured training to build teacher competence. The pressure to adapt and learn new skills adds to the psychological burden of academic medicine. Without support, the risk of fatigue and burnout increases. These considerations show that integrating generative AI depends on faculty adaptation. Their ability to adjust will decide if AI improves medical education or worsens inequalities and pressures within the profession [72].

Another area that merits discussion is the future of assessment in medical education. Traditional high-stakes examinations, administered at fixed intervals, may increasingly fail to align with the personalized and dynamic nature of contemporary medical training. Computerized adaptive testing (CAT), grounded in item response theory, offers a promising alternative that has been successfully simulated in national medical assessments. One study using the Standardized Competence Test for Clinical Medicine Undergraduates in China constructed a CAT item bank and found that this approach could significantly reduce test length while maintaining precision and validity. Similarly, psychometric simulations in health

XSL•FO
RenderX

professions assessment have shown that CAT can enhance reliability and reduce the burden on the examinee [73,74].

Beyond efficiency, adaptive testing offers competency-based assessment timing. Instead of fixed schedules, systems could test students once they show sufficient mastery. While this could lower stress and match assessments to readiness, it also raises transparency, fairness, and standardization concerns. We also need to acknowledge students' roles in shaping these systems. Engaged learners in design or evaluation, such as pilot programs, show higher acceptance and trust.

Conversely, lack of involvement can breed skepticism. While direct evidence from medical schools remains limited, broader educational research underscores that student participation can legitimize technology-mediated assessment and ensure it supports pedagogical goals [18,75,76].

Moreover, generative technologies are beginning to automate parts of assessment creation. Generative AI tools have been used to create multiple-choice items tailored to clinical training levels and support test blueprinting and standard setting. Recent comparisons between clinician-written and AI-generated questions are emerging to explore psychometric equivalence and item quality [77].

If assessment becomes more adaptive and technology-mediated, students' voices should not be an afterthought but a central part of the process. Their involvement can enhance perceptions of fairness, mitigate concerns about automation, and help align assessment systems with actual learning needs. Ultimately, striking a balance between technological capability, institutional responsibility, and student agency will be crucial in ensuring that assessment reform fosters both competence and trust in future physicians.

Generative AI is a fundamental force expected to transform medical education, not just an optional addition. Likewise, mental health support should be seen as essential rather than a luxury, and preparing faculty is as crucial as preparing students. These insights are practical as well: the NHS Topol Review has already highlighted the pressing need to ready the workforce for AI-enabled systems medicine [78], while the CanMEDS 2025 framework explicitly integrates digital health competencies into physician training [79].

At the same time, long-standing practices such as Schwartz Rounds and the Healer's Art course demonstrate that structured support for emotional well-being can reduce burnout and help students connect to the meaning of their future profession. Finally, global guidance, such as the World Health Organization's recommendations on AI ethics, underscores that safe integration requires both governance and education. When students feel recognized and supported, they are much more likely to become the kind of clinicians our changing world urgently needs: skilled, compassionate, connected, and profoundly humane. Achieving this requires a collective effort: researchers must speak their truths and highlight what needs to be seen; educators should balance rigor with reverence in shaping minds; students need to claim their right to full support; and decision-makers must invest not only in systems but also in the humanity of future healers.

Acknowledgments

No generative artificial intelligence tools were used at any stage in the preparation of this manuscript. The authors remain fully responsible for the accuracy, originality, and integrity of the manuscript.

Funding

This research did not receive any specific funding from public, commercial, or nonprofit sectors.

Data Availability

All data generated or analyzed during this study are included in this published article and its supplementary information files.

Authors' Contributions

NA compiled the data for the review and wrote the initial draft. BM gave methodological guidance and a review of the text. GK assisted with reviewing the text and proofreading.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Identified phenomena and supporting references.
[DOCX File, 196 KB - mededu_v12i1e85373_app1.docx]

References

1. Meskó B, Drobni Z, Bényei É, Gergely B, Győrffy Z. Digital health is a cultural transformation of traditional healthcare. Mhealth 2017;3:38. [doi: 10.21037/mhealth.2017.08.07] [Medline: 29184890]

XSL•FO
RenderX

2. Cryer REM, Calear AL, Batterham PJ, Patel SR. Suicide, mental, and physical health condition stigma in medical students. Death Stud 2020;44(4):230-236. [doi: 10.1080/07481187.2018.1539049] [Medline: 30451586]
3. Nair M, Moss N, Bashir A, et al. Mental health trends among medical students. Proc (Bayl Univ Med Cent) 2023;36(3):408-410. [doi: 10.1080/08998280.2023.2187207] [Medline: 37091765]
4. Backović DV, Maksimović M, Davidović D, Zivojinović JI, Stevanović D. Stress and mental health among medical students. Srp Arh Celok Lek 2013;141(11-12):780-784. [doi: 10.2298/sarh1312780b] [Medline: 24502097]
5. Hankir AK, Northall A, Zaman R. Stigma and mental health challenges in medical students. BMJ Case Rep 2014 Sep 2;2014:bcr2014205226. [doi: 10.1136/bcr-2014-205226] [Medline: 25183806]
6. Győrffy Z, Molnár R, Somorjai N. Health status and well-being of health care workers in Hungary: a literature review. Mentálhigiéné és Pszichoszomatika 2012 Jun;13(2):107-126. [doi: 10.1556/Mental.13.2012.2.1]
7. Yang A, Langness S, Chehab L, et al. Medical students in distress: a mixed methods approach to understanding the impact of debt on well-being. BMC Med Educ 2024 Aug 30;24(1):947. [doi: 10.1186/s12909-024-05927-9] [Medline: 39215314]
8. Meskó B, Hetényi G, Győrffy Z. Will artificial intelligence solve the human resource crisis in healthcare? BMC Health Serv Res 2018 Jul 13;18(1):545. [doi: 10.1186/s12913-018-3359-4] [Medline: 30001717]
9. Meskó B, Görög M. A short guide for medical professionals in the era of artificial intelligence. NPJ Digit Med 2020;3(1):126. [doi: 10.1038/s41746-020-00333-z] [Medline: 33043150]
10. Weidener L, Fischer M. Artificial intelligence in medicine: cross-sectional study among medical students on application, education, and ethical aspects. JMIR Med Educ 2024 Jan 5;10(1):e51247. [doi: 10.2196/51247] [Medline: 38180787]
11. Li Q, Qin Y. AI in medical education: medical student perception, curriculum recommendations and design suggestions. BMC Med Educ 2023 Nov 9;23(1):852. [doi: 10.1186/s12909-023-04700-8] [Medline: 37946176]
12. Sun L, Yin C, Xu Q, Zhao W. Artificial intelligence for healthcare and medical education: a systematic review. Am J Transl Res 2023;15(7):4820-4828. [Medline: 37560249]
13. Nagi F, Salih R, Alzubaidi M, et al. Applications of artificial intelligence (AI) in medical education: a scoping review. Stud Health Technol Inform 2023 Jun 29;305:648-651. [doi: 10.3233/SHTI230581] [Medline: 37387115]
14. Preiksaitis C, Rose C. Opportunities, challenges, and future directions of generative artificial intelligence in medical education: scoping review. JMIR Med Educ 2023 Oct 20;9:e48785. [doi: 10.2196/48785] [Medline: 37862079]
15. Boscardin CK, Gin B, Golde PB, Hauer KE. ChatGPT and generative artificial intelligence for medical education: potential impact and opportunity. Acad Med 2024 Jan 1;99(1):22-27. [doi: 10.1097/ACM.0000000000005439] [Medline: 37651677]
16. Triola MMM, Rodman AMM. Integrating generative artificial intelligence into medical education: curriculum, policy, and governance strategies. Acad Med 2025 Apr 1;100(4):413-418. [doi: 10.1097/ACM.0000000000005963]
17. Gordon M, Daniel M, Ajiboye A, et al. A scoping review of artificial intelligence in medical education: BEME Guide No. 84. Med Teach 2024 Apr;46(4):446-470. [doi: 10.1080/0142159X.2024.2314198] [Medline: 38423127]
18. Duan S, Liu C, Rong T, Zhao Y, Liu B. Integrating AI in medical education: a comprehensive study of medical students' attitudes, concerns, and behavioral intentions. BMC Med Educ 2025 Apr 23;25(1):599. [doi: 10.1186/s12909-025-07177-9] [Medline: 40269824]
19. Shimizu I, Kasai H, Shikino K, et al. Developing medical education curriculum reform strategies to address the impact of generative AI: qualitative study. JMIR Med Educ 2023 Nov 30;9:e53466. [doi: 10.2196/53466] [Medline: 38032695]
20. Park J. Medical students' patterns of using ChatGPT as a feedback tool and perceptions of ChatGPT in a Leadership and Communication course in Korea: a cross-sectional study. J Educ Eval Health Prof 2023;20:29. [doi: 10.3352/jeehp.2023.20.29] [Medline: 38096895]
21. Alkhaaldi SMI, Kassab CH, Dimassi Z, et al. Medical student experiences and perceptions of ChatGPT and artificial intelligence: cross-sectional study. JMIR Med Educ 2023 Dec 22;9(1):e51302. [doi: 10.2196/51302] [Medline: 38133911]
22. Dederichs M, Weber J, Pischke CR, Angerer P, Apolinário-Hagen J. Exploring medical students' views on digital mental health interventions: a qualitative study. Internet Interv 2021 Sep;25:100398. [doi: 10.1016/j.invent.2021.100398] [Medline: 34026567]
23. Mesko B. Commentary: we need to be better prepared for a technological future. BMJ 2015 Feb 10;350:h279. [doi: 10.1136/bmj.h279] [Medline: 25670194]
24. Sibbald M, Zwaan L, Yilmaz Y, Lal S. Incorporating artificial intelligence in medical diagnosis: a case for an invisible and (un)disruptive approach. J Eval Clin Pract 2024 Feb;30(1):3-8. [doi: 10.1111/jep.13730] [Medline: 35761764]
25. Arvai N, Katonai G, Mesko B. Health care professionals' concerns about medical AI and psychological barriers and strategies for successful implementation: scoping review. J Med Internet Res 2025;27:e66986. [doi: 10.2196/66986] [Medline: 40267462]
26. Meskó B. The Medical Futurist Institute: a vision about the technological future of healthcare. Patterns (N Y) 2021 Apr 9;2(4):100234. [doi: 10.1016/j.patter.2021.100234] [Medline: 33982024]
27. Choo CW. Environmental scanning as information seeking and organizational learning. Inform Res 2001 Oct;7(1) [FREE Full text]
28. Hiltunen E. The future sign and its three dimensions. Futures 2008 Apr;40(3):247-260. [doi: 10.1016/j.futures.2007.08.021]
29. Kuosa T. Futures signals sense-making framework (FSSF): a start-up tool to analyse and categorise weak signals, wild cards, drivers, trends and other types of information. Futures 2010 Feb;42(1):42-48. [doi: 10.1016/j.futures.2009.08.003]

XSL•FO
RenderX

30. Jabareen Y. Building a conceptual framework: philosophy, definitions, and procedure. Int J Qual Methods 2009 Dec;8(4):49-62. [doi: 10.1177/160940690900800406]
31. Barasa EW, Molyneux S, English M, Cleary S. Setting healthcare priorities at the macro and meso levels: a framework for evaluation. Int J Health Policy Manag 2015 Sep 16;4(11):719-732. [doi: 10.15171/ijhpm.2015.167] [Medline: 26673332]
32. Valentijn PP, Schepman SM, Opheij W, Bruijnzeels MA. Understanding integrated care: a comprehensive conceptual framework based on the integrative functions of primary care. Int J Integr Care 2022 Mar;13:e010. [doi: 10.5334/ijic.886] [Medline: 23687482]
33. Sawatzky R, Kwon JY, Barclay R, et al. Implications of response shift for micro-, meso-, and macro-level healthcare decision-making using results of patient-reported outcome measures. Qual Life Res 2021 Dec;30(12):3343-3357. [doi: 10.1007/s11136-021-02766-9] [Medline: 33651278]
34. Geels FW, Schot J. Typology of sociotechnical transition pathways. Res Policy 2007 Apr;36(3):399-417. [doi: 10.1016/j.respol.2007.01.003]
35. Geels FW, Schot J. The dynamics of transitions: a socio-technical perspective. In: Transitions to Sustainable Development: New Directions in the Study of Long Term Transformative Change: Routledge; 2010.
36. Geels FW. Socio-technical transitions to sustainability: a review of criticisms and elaborations of the Multi-Level Perspective. Curr Opin Environ Sustain 2019 Aug;39:187-201. [doi: 10.1016/j.cosust.2019.06.009]
37. Giunti G, Doherty CP. Futureproofing medical education through scenario analysis: instrumental case study. Stud Health Technol Inform 2024 Aug 22;316:1538-1539. [doi: 10.3233/SHTI240709] [Medline: 39176498]
38. Thomae AV, Witt CM, Barth J. Integration of ChatGPT into a course for medical students: explorative study on teaching scenarios, students' perception, and applications. JMIR Med Educ 2024 Aug 22;10:e50545. [doi: 10.2196/50545] [Medline: 39177012]
39. Chan KS, Zary N. Applications and challenges of implementing artificial intelligence in medical education: integrative review. JMIR Med Educ 2019 Jun 15;5(1):e13930. [doi: 10.2196/13930] [Medline: 31199295]
40. Weidener L, Fischer M. Teaching AI ethics in medical education: a scoping review of current literature and practices. Perspect Med Educ 2023;12(1):399-410. [doi: 10.5334/pme.954] [Medline: 37868075]
41. Perotta B, Arantes-Costa FM, Enns SC, et al. Sleepiness, sleep deprivation, quality of life, mental symptoms and perception of academic environment in medical students. BMC Med Educ 2021 Feb 17;21(1):111. [doi: 10.1186/s12909-021-02544-8] [Medline: 33596885]
42. Dyrbye LN, Thomas MR, Shanafelt TD. Medical student distress: causes, consequences, and proposed solutions. Mayo Clin Proc 2005 Dec;80(12):1613-1622. [doi: 10.4065/80.12.1613] [Medline: 16342655]
43. Watson C, Ventriglio A, Bhugra D. A narrative review of suicide and suicidal behavior in medical students. Indian J Psychiatry 2020;62(3):250-256. [doi: 10.4103/psychiatry.IndianJPsychiatry_357_20] [Medline: 32773867]
44. Kemp S, Hu W, Bishop J, et al. Medical student wellbeing - a consensus statement from Australia and New Zealand. BMC Med Educ 2019 Mar 4;19(1):69. [doi: 10.1186/s12909-019-1505-2] [Medline: 30832630]
45. Givens JL, Tjia J. Depressed medical students' use of mental health services and barriers to use. Acad Med 2002 Sep;77(9):918-921. [doi: 10.1097/00001888-200209000-00024] [Medline: 12228091]
46. Hankir A, Zaman R. Stigma and mental health challenges in medical students. Eur Psychiatry 2015 Mar;30:722. [doi: 10.1016/S0924-9338(15)30572-1]
47. Thesen T, Marrero WJ, Konopasky AJ, Duncan MS, Blackmon KE. Towards precision well-being in medical education. Med Teach 2025 Apr;47(4):630-634. [doi: 10.1080/0142159X.2024.2357279] [Medline: 38808734]
48. Knopp MI, Warm EJ, Weber D, et al. AI-enabled medical education: threads of change, promising futures, and risky realities across four potential future worlds. JMIR Med Educ 2023 Dec 25;9(1):e50373. [doi: 10.2196/50373] [Medline: 38145471]
49. Bleijenbergh R, Mestdagh E, Kuipers YJ. Midwifery practice and education in Antwerp: forecasting its future with scenario planning. J Contin Educ Nurs 2022 Jan;53(1):21-29. [doi: 10.3928/00220124-20211210-07] [Medline: 34978478]
50. Forman D, Nicol P, Nicol P. Looking to the future: framing the implementation of interprofessional education and practice with scenario planning. Educ Health 2015;28(3):162. [doi: 10.4103/1357-6283.178608]
51. Gentile JP, Roman B. Medical student mental health services: psychiatrists treating medical students. Psychiatry (Edgmont) 2009 May;6(5):38-45. [Medline: 19724734]
52. Hale EW, Davis RA. Supporting the future of medicine: student mental health services in medical school. Front Health Serv 2023;3:1032317. [doi: 10.3389/frhs.2023.1032317] [Medline: 36926507]
53. Tran M, Balasooriya C, Jonnagaddala J, et al. Situating governance and regulatory concerns for generative artificial intelligence and large language models in medical education. NPJ Digit Med 2025 May 27;8(1):315. [doi: 10.1038/s41746-025-01721-z] [Medline: 40425695]
54. Lekadir K, Frangi AF, Porras AR, et al. FUTURE-AI: international consensus guideline for trustworthy and deployable artificial intelligence in healthcare. BMJ 2025 Feb 5;388:e081554. [doi: 10.1136/bmj-2024-081554] [Medline: 39909534]
55. Hale J, Alexander S, Wright ST, Gilliland K. Generative AI in undergraduate medical education: a rapid review. J Med Educ Curric Dev 2024 Jan;11. [doi: 10.1177/23821205241266697]
56. Chen M, Zhang B, Cai Z, et al. Acceptance of clinical artificial intelligence among physicians and medical students: a systematic review with cross-sectional survey. Front Med 2022;9. [doi: 10.3389/fmed.2022.990604]

XSL•FO
RenderX

57. Osborn TG, Li S, Saunders R, Fonagy P. University students' use of mental health services: a systematic review and meta-analysis. Int J Ment Health Syst 2022 Dec 17;16(1):57. [doi: 10.1186/s13033-022-00569-0] [Medline: 36527036]
58. Abrams Z. Student mental health is in crisis. campuses are rethinking their approach. : American Psychological Association; 2022 URL: https://www.apa.org/monitor/2022/10/mental-health-campus-care [accessed 2026-05-06]
59. Zis P, Artemiadis A, Bargiotas P, Nteveros A, Hadjigeorgiou GM. Medical studies during the COVID-19 pandemic: the impact of digital learning on medical students' burnout and mental health. Int J Environ Res Public Health 2021 Jan 5;18(1):349. [doi: 10.3390/ijerph18010349] [Medline: 33466459]
60. Xu X, Chen Y, Miao J. Opportunities, challenges, and future directions of large language models, including ChatGPT in medical education: a systematic scoping review. J Educ Eval Health Prof 2024;21:6. [doi: 10.3352/jeehp.2024.21.6] [Medline: 38486402]
61. Cervantes J, Smith B, Ramadoss T, D'Amario V, Shoja MM, Rajput V. Decoding medical educators' perceptions on generative artificial intelligence in medical education. J Investig Med 2024 Oct;72(7):633-639. [doi: 10.1177/10815589241257215] [Medline: 38785310]
62. Lee YM, Kim S, Lee YH, et al. Defining medical AI competencies for medical school graduates: outcomes of a Delphi survey and medical student/educator questionnaire of South Korean medical schools. Acad Med 2024 May 1;99(5):524-533. [doi: 10.1097/ACM.0000000000005618] [Medline: 38207056]
63. Veal CT. We burn out, we break, we die: medical schools must change their culture to preserve medical student mental health. Acad Med 2021 May 1;96(5):629-631. [doi: 10.1097/ACM.0000000000003991] [Medline: 33570856]
64. Bhugra D, Molodynski A. Well-being and burnout in medical students: challenges and solutions. Ir J Psychol Med 2024 Jun;41(2):175-178. [doi: 10.1017/ipm.2022.26] [Medline: 35635210]
65. Asayesh H, Sharififard F, Mosavi M, Kharameh ZT, Arani ZA, Bidgoli AS. Correlation among academic stress, academic burnout, and academic performance in nursing and paramedic students of Qom University of Medical Sciences, Iran. Qom Univ Med Sci J 2016 Sep;10(7):74-83 [FREE Full text]
66. Kilic R, Nasello JA, Melchior V, Triffaux JM. Academic burnout among medical students: respective importance of risk and protective factors. Public Health (Fairfax) 2021 Sep;198:187-195. [doi: 10.1016/j.puhe.2021.07.025]
67. Pokhrel NB, Khadayat R, Tulachan P. Depression, anxiety, and burnout among medical students and residents of a medical school in Nepal: a cross-sectional study. BMC Psychiatry 2020 Jun 15;20(1):298. [doi: 10.1186/s12888-020-02645-6] [Medline: 32539732]
68. Rothenberger DA. Physician burnout and well-being: a systematic review and framework for action. Dis Colon Rectum 2017 Jun;60(6):567-576. [doi: 10.1097/DCR.0000000000000844] [Medline: 28481850]
69. West CP, Dyrbye LN, Erwin PJ, Shanafelt TD. Interventions to prevent and reduce physician burnout: a systematic review and meta-analysis. Lancet 2016 Nov 5;388(10057):2272-2281. [doi: 10.1016/S0140-6736(16)31279-X] [Medline: 27692469]
70. Sallam M, Al-Mahzoum K, Almutairi Y, et al. Anxiety among medical students regarding generative artificial intelligence models: a pilot descriptive study. MDPI. Preprint posted online on Aug 16, 2024. [doi: 10.20944/preprints202408.1215.v1]
71. Balint M. The doctor, his patient, and the illness. Lancet 1955 Apr 2;268(6866):683-688. [doi: 10.1016/s0140-6736(55)91061-8] [Medline: 14354967]
72. Ayyoub AM, Khlaif ZN, Shamali M, et al. Advancing higher education with GenAI: factors influencing educator AI literacy. Front Educ 2025;10. [doi: 10.3389/feduc.2025.1530721]
73. Choi Y, McClenen C. Development of adaptive formative assessment system using computerized adaptive testing and dynamic Bayesian networks. Appl Sci (Basel) 2020;10(22):8196. [doi: 10.3390/app10228196]
74. Yang ACM, Flanagan B, Ogata H. Adaptive formative assessment system based on computerized adaptive testing and the learning memory cycle for personalized learning. Comput Educ Artif Intell 2022;3:100104. [doi: 10.1016/j.caeai.2022.100104]
75. Pinto Dos Santos D, Giese D, Brodehl S, et al. Medical students' attitude towards artificial intelligence: a multicentre survey. Eur Radiol 2019 Apr;29(4):1640-1646. [doi: 10.1007/s00330-018-5601-1] [Medline: 29980928]
76. Stewart J, Lu J, Gahungu N, et al. Western Australian medical students' attitudes towards artificial intelligence in healthcare. PLoS One 2023;18(8):e0290642. [doi: 10.1371/journal.pone.0290642] [Medline: 37651380]
77. Wu H, Zerner T, Lee D, Court-Kowalski S, Devitt P, Palmer E. GPT-4 versus human authors in clinically complex MCQ creation: a blinded analysis of item quality. Med Teach 2025 Dec;47(12):1961-1974. [doi: 10.1080/0142159X.2025.2505122] [Medline: 40439661]
78. The Topol Review. Preparing the Healthcare Workforce to Deliver the Digital Future. : Health Education England; 2019 URL: https://topol.digitalacademy.nhs.uk/the-topol-review [accessed 2026-05-06]
79. Hall J, Atkinson A, Chan MK, Tourian L, Thoma B, Pattani R. The clinical learning environment in CanMEDS 2025. Can Med Ed J 2023. [doi: 10.36834/cmej.75537]

Abbreviations

AI: artificial intelligence
CAT: computerized adaptive testing
HC: health care

XSL•FO
RenderX

HCP: health care professional
IMI: internet and mobile-based intervention
LLM: large language model
MS: medical student
SA: scenario analysis
SWOT: Strengths, Weaknesses, Opportunities, Threats

Edited by A Stone; submitted 06.Oct.2025; peer-reviewed by O Ghanem, S Cho; revised version received 16.Mar.2026; accepted 30.Mar.2026; published 26.May.2026.

Please cite as:
Arvai N, Meskó B, Katonai G
Generative AI's Impact on the Mental Health of Medical Students: Scenario Analysis
JMIR Med Educ 2026;12:e85373
URL: https://mededu.jmir.org/2026/1/e85373
doi:10.2196/85373

© Nora Arvai, Bertalan Meskó, Gellért Katonai. Originally published in JMIR Medical Education (https://mededu.jmir.org), 26.May.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Education, is properly cited. The complete bibliographic information, a link to the original publication on https://mededu.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

An AI-Driven Virtual Patient Platform (CBT Trainer) for Training Cognitive Behavioral Therapy Practitioners Against Competencies: Mixed Methods Pilot Study

Tianyu Terry Zhang[1], MSc, MPS; Rob Saunders[1], PhD; Stephen Pilling[1,2], PhD; Ciarán O'Driscoll[1], PhD

[1]Research Department of Clinical, Educational & Health Psychology, Centre for Outcomes Research and Effectiveness (CORE), University College London, 1-19 Torrington Place, London, United Kingdom

[2]Camden and Islington NHS Foundation Trust, London, United Kingdom

Corresponding Author:
Tianyu Terry Zhang, MSc, MPS
Research Department of Clinical, Educational & Health Psychology, Centre for Outcomes Research and Effectiveness (CORE), University College London, 1-19 Torrington Place, London, United Kingdom

Abstract

Background: Cognitive behavioral therapy (CBT) training faces significant challenges, including supervised practice with diverse cases, inconsistent feedback, resource-intensive supervision, and difficulties standardizing competence assessment.

Objective: This study evaluated the acceptability and feasibility of CBT Trainer (TTZ), the first virtual patient platform to provide real-time feedback aligned with established competence frameworks. The mobile app trains psychological practitioners using standardized artificial intelligence patient interactions and the evaluation of therapist responses against competence frameworks to enable structured skill development in a controlled, repeatable environment that complements traditional training methods.

Methods: This mixed methods pilot study used a 2-stage approach. Stage 1 involved usability testing with 4 participants. Stage 2 included 59 participants from psychological practitioner training programs (a Low Intensity CBT Interventions Program and a Doctorate in Clinical Psychology) who engaged with the CBT Trainer voluntarily for over 1 month. Measures of impact included the System Usability Scale (SUS), platform naturalistic engagement, poststudy questionnaire on perceived competency development, comparative evaluation against traditional role-play, and qualitative feedback.

Results: Participants engaged voluntarily with the platform for an average of 95.24 (SD 134.58; median 45.34, IQR 11.57–105.15) minutes of active role-play. Platform usability was rated as excellent (mean SUS 82.20, SD 12.93). Self-reported competence improvement improved most in assessment skills (96.7%) and information gathering (66.7%). When compared to traditional peer role-play exercises, participants rated CBT Trainer moderately favorably (mean 5.90/10, SD 1.94). Qualitative feedback highlighted strengths in competency-aligned feedback, on-demand accessibility, and a psychologically safe practice space.

Conclusions: This pilot study provides evidence that an artificial intelligence–based patient simulation shows promise as a supplementary training tool for psychological therapists who use CBT in their practice, particularly regarding accessibility and immediate feedback. Future research should use randomized controlled designs with objective competence assessments.

Trial Registration: OSF Registries 10.17605/OSF.IO/MSKB7; https://osf.io/mskb7

(JMIR Med Educ 2026;12:e84091) doi:10.2196/84091

KEYWORDS

cognitive behavioral therapy; artificial intelligence; virtual patients; competence assessment; competency-based education; clinical training; psychological therapy training; simulation-based learning; mobile app; digital health

Introduction

Overview

The training of mental health professionals represents a critical challenge in addressing the global mental health crisis. Within the National Health Service (NHS), there have been initiatives aimed at improving the quality of care and mental health outcomes by improving access to evidence-based psychological therapies [1]. This requires developing a skilled workforce. In addition, as demand for psychological services continues to rise, there is increasing pressure on training programs to develop competent practitioners efficiently and effectively. Technological innovations, particularly machine learning and natural language processing, offer promising new avenues for enhancing mental health training through virtual patient simulations with competency-based feedback, potentially addressing long-standing challenges in traditional approaches.

XSL•FO
RenderX

Training Competent Therapists: The Role of Competence Frameworks

Competency-based education has emerged as a dominant paradigm in mental health care training, providing structured frameworks essential for skill acquisition, assessment, and professional development [2]. Across therapeutic modalities, the University College London (UCL) Center for Outcomes Research and Effectiveness (CORE) competence frameworks provide a comprehensive hierarchical model of competencies required for effective practice, distinguishing between generic therapeutic competencies, basic therapy-specific technique (eg, CBT [cognitive behavioral therapy]) competencies, and meta-competencies [3]. These frameworks are operationalized through validated competency scales, which translate professional standards into measurable assessment tools. For example, the Cognitive Therapy Scale-Revised (CTS-R) [4], the revised form of Cognitive Therapy Rating Scale (CTRS) [5], is one of the main treatment competence measurement tools used for CBT therapist accreditation internationally, comprising 12 distinct competency domains, including agenda setting, feedback, collaboration, pacing, and interpersonal effectiveness [6]. Competence frameworks and scales like these serve multiple functions, such as providing structured developmental pathways [7,8], facilitating formative assessment [9], supporting summative evaluation [10], and promoting training consistency [11].

Nevertheless, implementing competency-based training and evaluation faces considerable practical and methodological challenges. First, the reliability of competence assessment remains inconsistent; although tools such as the CTS-R can achieve high interrater reliability under controlled conditions, such as when groups of raters work together, independent ratings often show poor-to-moderate agreement [3,10,12,13]. Second, resource limitations create substantial barriers to effective competence development. This includes demands on clinical supervisors, limited access for trainees to diverse clinical presentations, reliance on time-demanding assessment methods, and resource constraints affecting feedback quality [8,14]. Third, ethical tensions arise as trainees must develop competence "on the job," with a need to ensure patient safety and adhere to professional codes of conduct [15]. While this is managed through informed consent, supervision, adherence to professional codes of conduct, and a constant focus on the trainee's competence and ethical decision-making [15,16], it requires a high level of supervisory support, which can be resource-intensive for training programs managing increasing cohort sizes. The expansion of the psychological professions workforce places additional demands on supervision infrastructure, potentially affecting the quality and availability of supervisory support [1,14].

Traditional Role-Play Methods in Mental Health Training

Role-play exercises are a cornerstone of competency-based training, offering structured opportunities to practice and receive feedback on specific therapeutic skills [17,18]. While role-plays provide valuable opportunities for skill development, they face substantial limitations in terms of ecological validity, scalability, and consistency [19,20]. Within psychological practitioner training, standardized role-plays using trained actors offer structured learning, though validity concerns persist [10]. Student-to-student role-plays face practical challenges such as inconsistencies in role-play quality due to student anxiety when role-playing with peers, authenticity limitations as students may deliberately facilitate their peers' success rather than presenting realistic clinical challenges, ethical concerns regarding self-disclosure, and logistical difficulties across large student cohorts [17,18,21-23]. Research has identified 4 authenticity barriers, namely, students' insufficient knowledge of psychiatric symptoms, reluctance to portray distressing emotions, artificial dynamics from preexisting relationships, and inadequate environmental settings [17,24]. Interestingly, studies show higher-performing students often experience greater anxiety during role-plays than lower-performing peers—a "competence-anxiety paradox" linked to skilled trainees' greater awareness of therapeutic complexity [18]. Teacher demonstration models, where instructors model therapeutic techniques with volunteer "clients," provide excellent teaching opportunities but limited hands-on practice for students [18]. These limitations contribute to the "theory-practice gap" where classroom knowledge fails to translate effectively to clinical application [20,25].

Technological Innovations in Mental Health Training

Traditional role-play limitations have driven advancements in simulation technology for clinical skills training. Innovations included decision tree–based interactive videos that improved the standardization of training experiences and assessment procedures [26,27]. This approach has reduced psychological strain on actors playing patient roles, as they no longer needed to maintain complex character portrayals for extended periods and improved scalability by eliminating the need for extensive simulant training that required 10 hours for students or 3 - 4 hours for clinicians [27]. While research shows skills can improve through these methods, there is variability in the degree to which students find interactive client systems engaging and useful [26,28]. Machine learning and language models have advanced patient simulations through contextual understanding and human-like text generation, which can increase realism comparable to human patients [29], across diverse scenarios, though responses sometimes appeared overly idealized [30]. While these approaches facilitate practice, they do not provide competency-based feedback to further skill development, which is essential for meeting professional training standards and ensuring therapists develop the specific competencies required for NHS service delivery.

Rationale for This Study

Key gaps remain in the literature on virtual patient simulations for CBT training. First, to our knowledge, although existing tools have achieved realistic patient interactions [26,28-30], none have systematically translated established competence frameworks, such as the CTS-R or UCL competence frameworks, into real-time feedback during simulated practice. This limits their alignment with professional training standards and NHS service requirements. While artificial intelligence (AI) systems have been developed to assess CBT competence from

XSL•FO
RenderX

recorded therapy sessions [31,32], the integration of such frameworks into simulation-based training represents a critical unmet need. Given that competence frameworks underpin evidence-based CBT training and professional accreditation [2,3], bridging this gap is essential for advancing scalable, standards-aligned training tools.

Approaches to operationalize competence development through frameworks and scales need to build on learning models such as experiential learning [33], deliberate practice methodologies [34], and technology-enhanced training [7,35]. Therefore, an AI-driven simulation platform that incorporates competency-based feedback mechanisms offers a promising solution to address the identified limitations in traditional training approaches while maintaining alignment with established professional frameworks and educational principles. Such a platform fosters experiential learning by enabling safe, active experimentation alongside structured reflection, while also supporting deliberate practice through systematic and repeatable skill development, representing a novel approach that addresses longstanding limitations in CBT training technology.

A second significant gap pertains to voluntary engagement with training technologies in psychological practitioner education. Existing studies on simulation-based learning have largely examined mandatory curricular components or controlled experimental settings, focusing primarily on perceived fidelity or effectiveness [28-30]. Understanding how trainees voluntarily engage with supplementary training technologies provides critical evidence for implementation feasibility and ecological validity. Unlike previous research, this study investigates naturalistic engagement patterns within real-world training programs, integration with existing curricula, and specific mechanisms through which such simulations might enhance learning. By combining quantitative engagement metrics with qualitative feedback, this research can provide preliminary evidence regarding the practical utility of AI simulations in psychological therapist training.

Study Aims

This pilot study aimed to evaluate the educational potential of the first virtual patient platform providing competence framework–aligned feedback in psychological therapist training. Specifically, the study aimed:

1. to assess the acceptability of simulation-based learning integrated with competency-based feedback as a supplemental tool within established training programs for psychological therapists who use CBT in their practice.
2. to examine patterns of voluntary engagement with simulation-based practice among trainees from different professional training backgrounds.
3. to gather preliminary evidence regarding self-reported competence development following simulation-based practice experiences.
4. to explore trainees' perceptions of the educational value of framework-aligned feedback during simulated therapeutic interactions compared to traditional training methods.

Methods

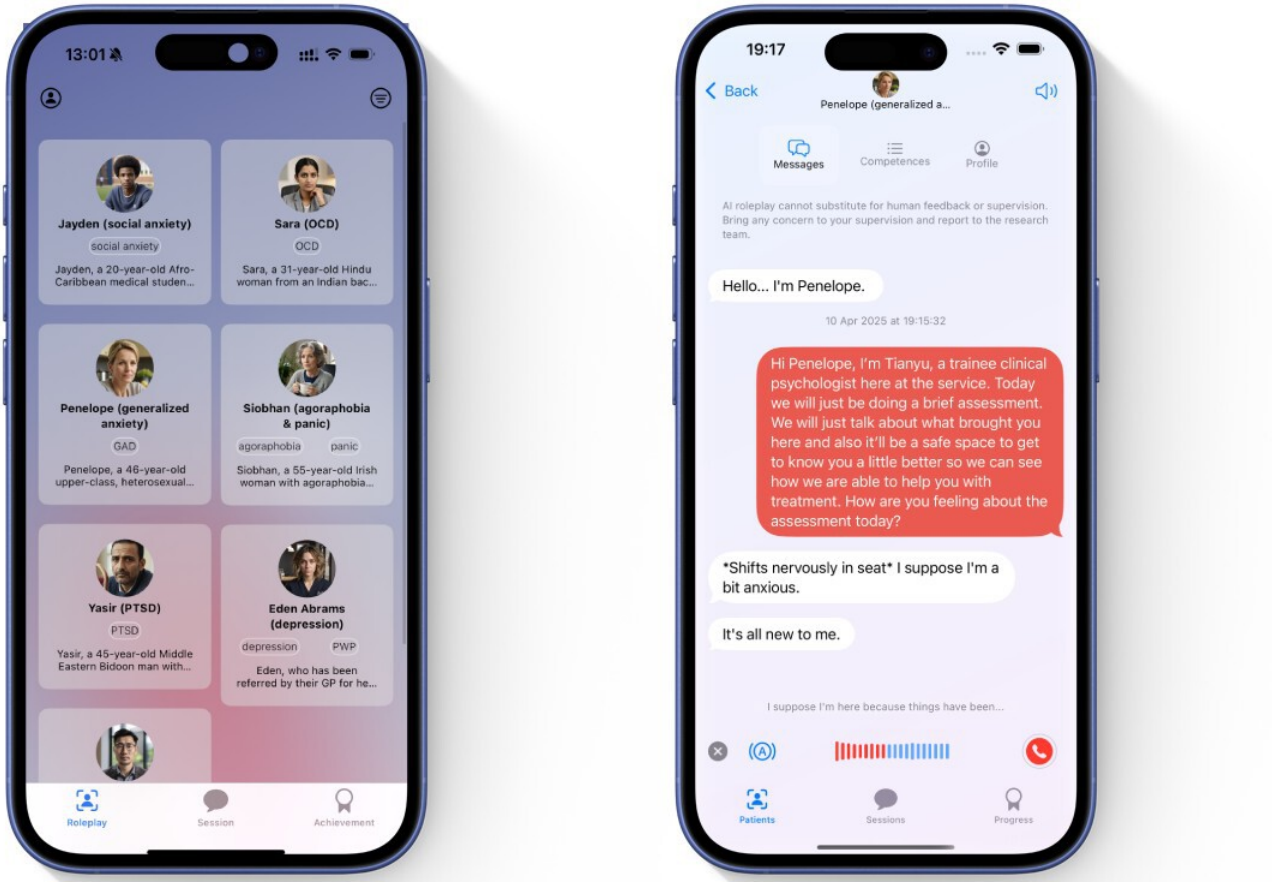
The CBT Trainer Platform

The CBT Trainer platform (TTZ) [36] is an AI-based virtual patient platform designed for mental health training. The app presented participants with interactive scenarios replicating therapeutic sessions for common mental health conditions, including depression, generalized anxiety disorder, social anxiety disorder, posttraumatic stress disorder, and agoraphobia. The platform primarily enabled voice- and text-based interaction with virtual patients exhibiting realistic clinical presentations. Users began by selecting from 6 virtual patients representing diverse demographics and clinical presentations. During voice- or text-based interactions, the system continuously evaluated responses against competence frameworks, providing real-time feedback visible in the interface.

CBT Trainer represents the first virtual patient simulation to operationalize established competence frameworks into real-time automated feedback. The platform was designed to support the practice of both assessment and intervention skills. Users could select which type of session to practice (assessment or intervention) and choose the corresponding competence framework for feedback. The CBT Trainer assessed trainees against a selected competency scale—either the Cognitive Therapy Scale-Revised (CTS-R) [4] or the UCL Psychological Wellbeing Practitioner (PWP) Assessment Competence Scale [37]. CBT Trainer indicates whether specific competencies were demonstrated or required further development. After practice sessions, participants received structured feedback, highlighting competencies met and those not met, with specific guidance to support their development of key therapeutic competencies. The platform was designed to complement, not replace, traditional training methods by providing additional opportunities for independent skill practice. The CBT Trainer app's content and features were kept unchanged throughout the data collection period to ensure consistency across all participant interactions. Figures 1 and 2 illustrate the app's user interface diagrams, including the patient selection screen displaying diverse clinical presentations, the interactive role-play screen, and the competency feedback screens.

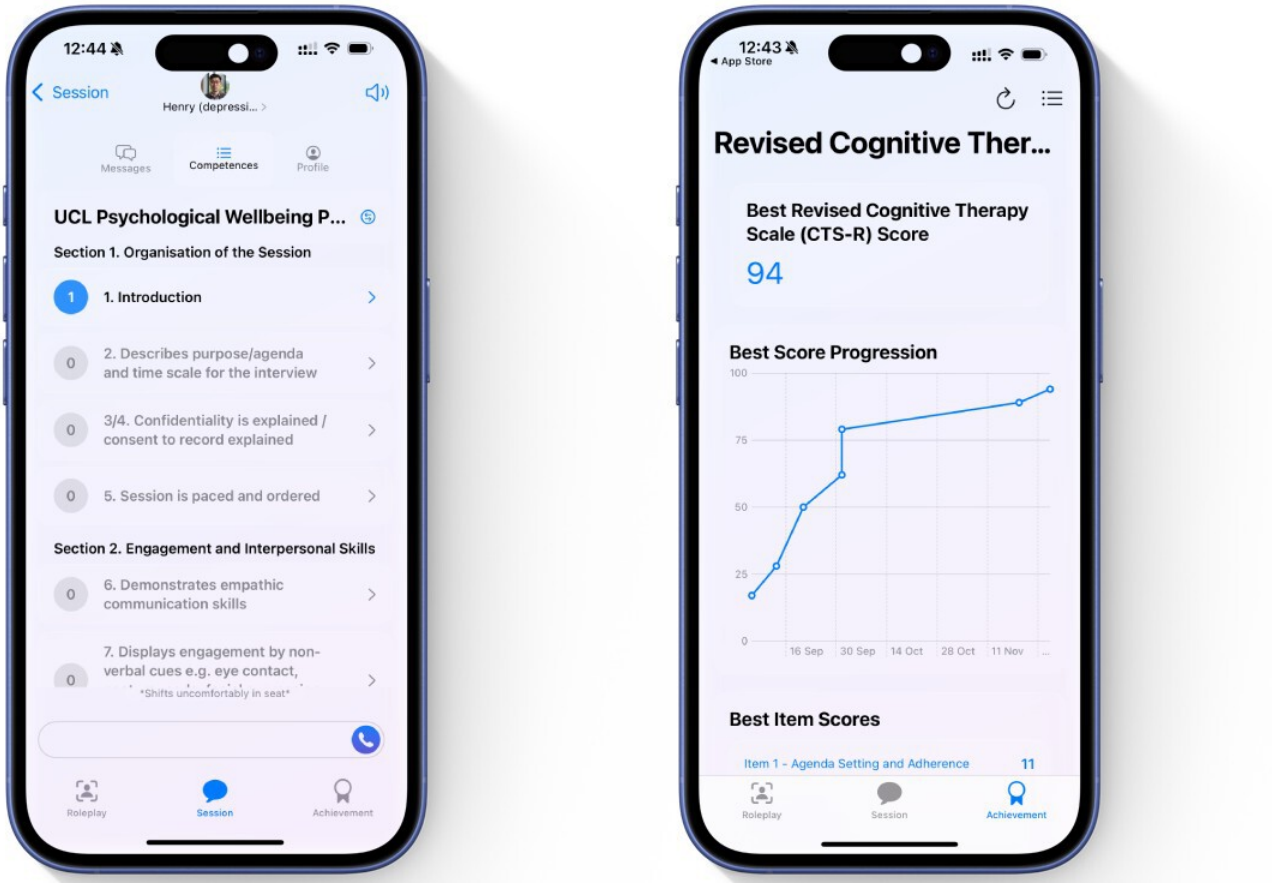
XSL•FO
RenderX

Figure 1. CBT Trainer patient selection (left) and role-play interfaces (right).



XSL•FO
RenderX

Figure 2. Competence assessment and progress tracking interfaces. (Left) University College London (UCL) Psychological Wellbeing Practitioner (PWP) Assessment Competence Scale interface shows real-time competence evaluation across core domains. (Right) Cognitive Therapy Scale-Revised (CTS-R) progress tracking screen.



Platform Development and Validation

The CBT Trainer platform was developed in collaboration with UCL's CORE, the research group that created the UCL competence frameworks upon which the platform is built. The development team comprised a clinical psychologist who designed patient profiles and competency-based feedback mechanisms, and a machine learning specialist who built the competence assessment algorithms.

Six diverse virtual patient profiles were created representing common mental health presentations, including depression, generalized anxiety disorder, social anxiety disorder, posttraumatic stress disorder, agoraphobia with panic, and obsessive-compulsive disorder (Table 1). Each profile was developed by a clinical psychologist to reflect realistic clinical presentations.

XSL•FO
RenderX

Table . Virtual patient profiles in CBT[a] Trainer.

Name	Age (years)	Background	Primary presentation	PHQ-9[b]	GAD-7[c]
Henry	35	British Chinese, Liverpool	Depression with risk	21	18
Jayden	20	Afro-Caribbean, London	Social anxiety	12	16
Penelope	46	British, dyslexia	Generalized anxiety	15	19
Siobhan	55	Irish	Agoraphobia with panic	15	19
Yasir	56	Middle Eastern (Bidoon), asylum seeker	PTSD[d]	19	18
Sara	31	Indian Hindu, London	OCD[e]	17	19

[a]CBT: cognitive behavioral therapy.

[b]PHQ-9: Patient Health Questionnaire-9.

[c]GAD-7: Generalized Anxiety Disorder-7.

[d]PTSD: posttraumatic stress disorder.

[e]OCD: obsessive compulsive disorder.

Prior to the pilot study, the platform underwent iterative testing and refinement. Patient interactions were tested by course tutors from the Doctorate in Clinical Psychology and PWP training programs at UCL to assess clinical authenticity and educational appropriateness. Early testing revealed that initial response patterns provided too much information too quickly compared to authentic patient presentations (where patients disclose more over time as the therapeutic relationship evolves). Based on this feedback, we adjusted response length and detail level to match typical patient communication patterns.

Throughout development and deployment, several measures ensured appropriate platform responses. The system development includes alignment efforts focused on maintaining task adherence to therapeutic contexts. All feedback was explicitly grounded in established competence frameworks (CTS-R and UCL PWP Assessment Competence Scale), limiting assessment to observable, framework-specified competencies.

Study Design

This pilot study used a 2-stage mixed methods approach to evaluate an AI-based virtual patient platform for training cognitive behavioral therapists. This included usability testing, digital engagement data, and questionnaires to assess acceptability and educational impact.

Participants

Participants were recruited from psychological practitioner training programs at UCL during the 2024 - 2025 academic year, targeting 2 specific programs, namely, the Doctorate in Clinical Psychology and the Low-intensity PWP training. Eligible participants were required to be enrolled in one of these programs and have access to an iOS device with internet connectivity capable of running the CBT Trainer app.

Procedure

The study was conducted in 2 stages. Stage 1 involved 4 participants in 1-hour in-person usability testing sessions and engaged with the platform under the researcher's observation. This allowed us to examine CBT trainer functionality and intuitiveness based on user actions and responses. Stage 1 findings informed Stage 2 by identifying needed platform refinements (eg, user experience improvements and bug fixes).

Stage 2 expanded the study to include 101 participants who completed an initial online screening. After completing online screening, informed consent, and demographic questionnaires, eligible participants downloaded the app to their personal device. Participants were encouraged to use the platform for at least 60 minutes during the study period to gain meaningful experience, though they were explicitly informed that there was no mandatory usage requirement and they could engage with the app as much or as little as they preferred. Two reminder emails were sent to all participants during weeks 1 and 3 of the 4-week study period, providing tips on integrating the platform with their training program activities. The stage concluded with a comprehensive online survey.

All 101 participants completing the screening survey were entered into a draw for 1 of 3 £45 (approximately US $59) vouchers, but no incentives were provided for app download or engagement to preserve the ecological validity of usage patterns.

The research version of the CBT Trainer app was distributed through Apple's TestFlight beta testing platform to maintain separation between research data and any future public release, which required participants to complete multiple steps, including receiving and accepting email invitations, downloading the TestFlight app, and then installing CBT Trainer.

Measures

Survey Instruments

System Usability Scale

A standardized 10-item questionnaire measuring platform usability, with scores ranging from 0 to 100 [38]. Scores above

XSL•FO
RenderX

68 are considered above average, and scores above 80 are considered excellent [39].

PostStudy Questionnaire

A custom questionnaire designed to capture participants' experiences with the CBT Trainer platform. The questionnaire comprised four main sections: (1) 3 subjective learning outcome measures assessing helpfulness, skill improvement, and perceived realism using 0 - 100 rating scales; (2) a checklist assessment of specific competence development across therapeutic domains; (3) comparative evaluation questions rating CBT Trainer against traditional role-play methods; and (4) open-ended qualitative questions exploring learning impact and user experience. The complete poststudy questionnaire is available in Multimedia Appendix 1.

Engagement Metrics

Platform usage was tracked through several dimensions, including role-play time (time spent in active interaction with simulated patients), session quantity (total number of role-play sessions created per user), and interaction depth (average exchanges between trainees and virtual patients per session, where each exchange comprised 1 trainee input paired with 1 AI patient response).

Data Analysis

Descriptive statistics were computed for all continuous variables (eg, System Usability Scale [SUS] scores, engagement metrics, and self-rated learning outcomes) and are presented as means, SDs, medians, and ranges. Categorical data (eg, demographic variables, competence development checklist items, and comparative advantages) are presented as frequencies (n) and percentages (%). Missing data were addressed by using complete case analysis, resulting in varying sample sizes across different analyses, which are explicitly reported alongside each result (eg, n=33 for SUS scores and n=31 for competence ratings).

Qualitative responses from an open-ended survey question, "Please provide one or two specific examples of how CBT Trainer impacted your clinical skills or knowledge," were analyzed using qualitative content analysis [40,41]. Qualitative content analysis was chosen as it enables describing and quantifying qualitative data while preserving participants' expressed experiences—appropriate for this acceptability study where we sought to identify the frequency and range of participant perspectives on platform utility while preserving participants' expressed experiences [42]. Analysis focused on identifying manifest (ie, surface level) meanings in the data following an inductive approach [43]. Analysis was conducted by TTZ, a clinical psychology doctoral research student and one of the platform developers, who maintained reflexive awareness of potential bias throughout the analytic process. The questionnaire responses were read several times by TTZ. All descriptions of specific ways the CBT Trainer impacted participants' clinical skills and training experiences were considered meaning units. If an impact was described multiple times within a single participant's response, these descriptions were conjoined into a single meaning unit. Codes (ie, specific impacts on clinical skills and knowledge) were then grouped into categories and subcategories on the basis of similarities and differences in their educational and learning functions.

Outcomes

The research team established prespecified outcomes prior to study commencement, following implementation of study guidelines [44]:

- Primary outcomes
- Platform engagement: target of >50% of participants using the app for ≥10 minutes, or >25% using it for ≥30 minutes (Aim 2)
- Perceived educational value: target of >70% of participants reporting positive impact on training (Aim 4)
- Secondary outcomes
- Recruitment and retention: target of ≥22 participants across both study phases (Aim 1)
- System usability: target SUS score >68 (Aim 1)
- Exploratory
- Self-reported improvements across competency domains (Aim 3)

Ethical Considerations

The study received approval from the UCL Research Ethics Committee (21883/006). The study protocol was preregistered with the Open Science Framework [45].

Several safeguards were implemented to address ethical concerns. Course directors from participating training programs were aware of the study. Meetings with program staff occurred at study commencement and debriefing. A human-in-the-loop design was implemented with weekly research team meetings reviewing engagement patterns to identify any concerning patterns. Specifically, the team monitored for (1) misuse of the platform for noneducational purposes through sampled transcript review, (2) inappropriate content or off-task interactions, (3) technical issues (error reports and crashes) that could impede learning, and (4) instances where AI feedback appeared inappropriate or misaligned with competence frameworks. No significant concerning engagement patterns requiring intervention were identified during the study period. Reminder emails sent at weeks 1 and 3 provided opportunities to report concerns. The platform included a direct link to the study information sheet. Participants were informed that the platform was experimental and supplementary to the traditional training, with in-app reminders to bring concerns to their supervisors and reminders of the limitations of AI-generated competency feedback. Survey questions provided opportunities for participants to report any concerns or difficulties. Support was available from researchers for technical issues or concerns about platform feedback. Participants who ceased engagement were contacted during survey collection to check whether issues had arisen. Coercion was mitigated through voluntary self-selection recruitment via email advertisement, no incentive payments for engagement, explicit consent procedures reminding that no minimum usage is required, and reminding the right to withdraw. Data privacy and security were prioritized, with personal identifiers kept separate from research data. Multiple contact

XSL•FO
RenderX

pathways were provided within the app, including principal investigator contact, ethics committee contact, and UCL Data Protection Office contact.

Results

Phase 1: Initial Usability Testing

Phase 1 involved 4 doctoral clinical psychology students (all women, aged 25 - 34 years). PWP trainees were not included in Phase 1 usability testing due to their later program start date. Participants engaged with the platform under the researcher's observation. Researchers monitored interactions for technical issues and clinical appropriateness. Several bugs and user experience navigation issues were revealed and fixed before Phase 2 implementation.

Phase 2: Extended Implementation

Participant Characteristics

Figure 3 illustrates participant flow and attrition throughout the study phases. Of 101 invited, 92 (91.1%) were eligible based on iOS device compatibility. Attrition occurred primarily at app download, where 25 eligible participants (27.2%) did not proceed likely due to TestFlight app download complexity and technical barriers, and at survey completion, where 28 active users (47.5%) did not complete follow-up assessments, likely mistaking the end-of-study survey as optional educational monitoring rather than mandatory research data collection. Technical difficulties with older iOS devices prevented 8 participants (8.7% of the eligible) from engaging despite successful installation. Overall, 59 participants (64.1% of the eligible) successfully engaged with the platform, with 31 - 33 completing comprehensive survey assessments for secondary analyses.

Of the 59 engaging participants, 84.7% (n=50) are from the Low Intensity Cognitive Behavioral Interventions program and 15.3% (n=9) are from the Doctorate in Clinical Psychology program. Participants were predominantly women (49/59, 83.1%) with a mean age of 28.37 (SD=7.25; range 22 - 62) years. The ethnic distribution included White (34/59, 57.6%), Black or African or Caribbean or Black British (9/59, 15.3%), Asian or Asian British (7/59, 11.9%), Other ethnic groups (5/59, 8.5%), and Mixed or Multiple ethnic groups (4/59, 6.8%). The majority of participants were at the early stages of their training, with a mean enrollment time of 1.61 months in their respective programs (SD 2.65; median 1.0, IQR 0 - 18 months). Regarding familiarity with competency assessment frameworks, 39% (n=23) reported being somewhat familiar, 25.4% (n=15) were neutral, 16.9% (n=10) were somewhat unfamiliar, 16.9% (n=10) were very unfamiliar or had never heard of them, and 1.7% (n=1) were very familiar. Exploratory analyses revealed no substantial differences in engagement patterns, demographics, or outcomes between the 2 training programs, supporting combined analysis of the sample.

XSL•FO
RenderX

Figure 3. Participant flow diagram; CBT: cognitive behavioral therapy.

XSL•FO
RenderX

Platform Engagement

Of the 59 engaging participants, they spent an average of 95.24 (SD 134.58; median 45.34, IQR 11.57-105.15) minutes engaging in role-plays. Users created an average of 4.24 role-play sessions each (SD 3.72; range 1 - 14), with an average session duration of 21.48 (SD 23.24; median 14.21, IQR 5.93-32.82) minutes. Each session contained an average of 48.58 (SD 57.11, median 21, IQR 4–82) interactions. The data were right-skewed (most values clustered at the low end with a long tail of high values). While most participants engaged for shorter periods, a smaller number engaged for very long periods (several hours).

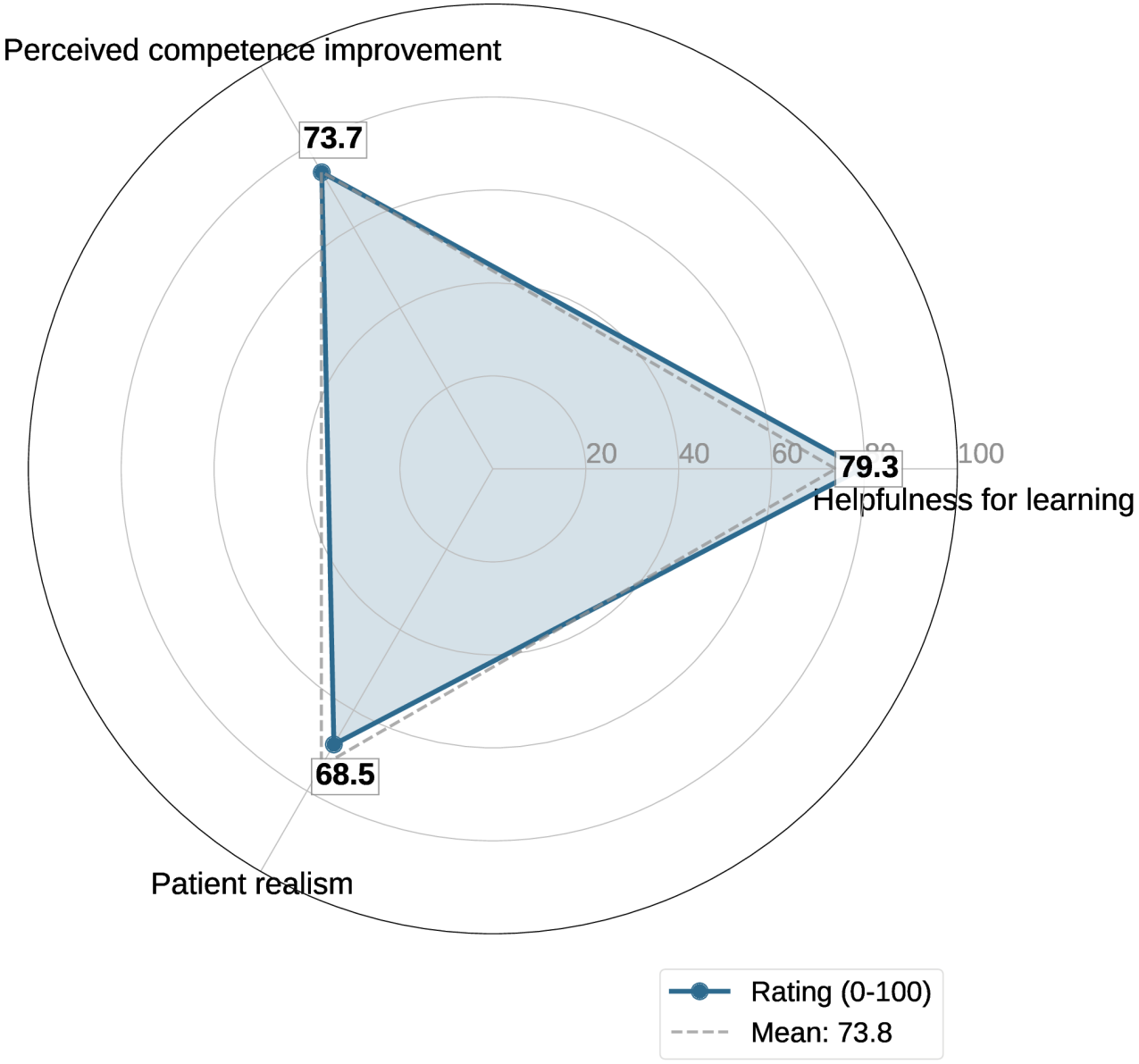
System Usability

The CBT Trainer platform achieved high usability ratings with a mean SUS score of 82.2 (SD 12.93; median 82.5, IQR 72.5-92.5, n=33), placing it in the "excellent" category according to established benchmarks [39].

Self-Reported Competence Development

Participants reported high agreement that "The simulated patient interactions were helpful for my learning" (mean 79.35, SD 17.49; median 80.0, IQR 69.5-95.5, n=31) and "My clinical skills has improved after using CBT Trainer" (mean 73.67, SD 23.03; median 81.0, IQR 60.0-90.0 ,n=30), as illustrated in Figure 4. The statement "The simulated patients were realistic to real patients" received moderately strong agreement (mean 68.45, SD 20.45; median=70.0, IQR 49.5-87.5, n=31).

Figure 4. Educational impact ratings reported by CBT (cognitive behavioral therapy) Trainer users (n=31).

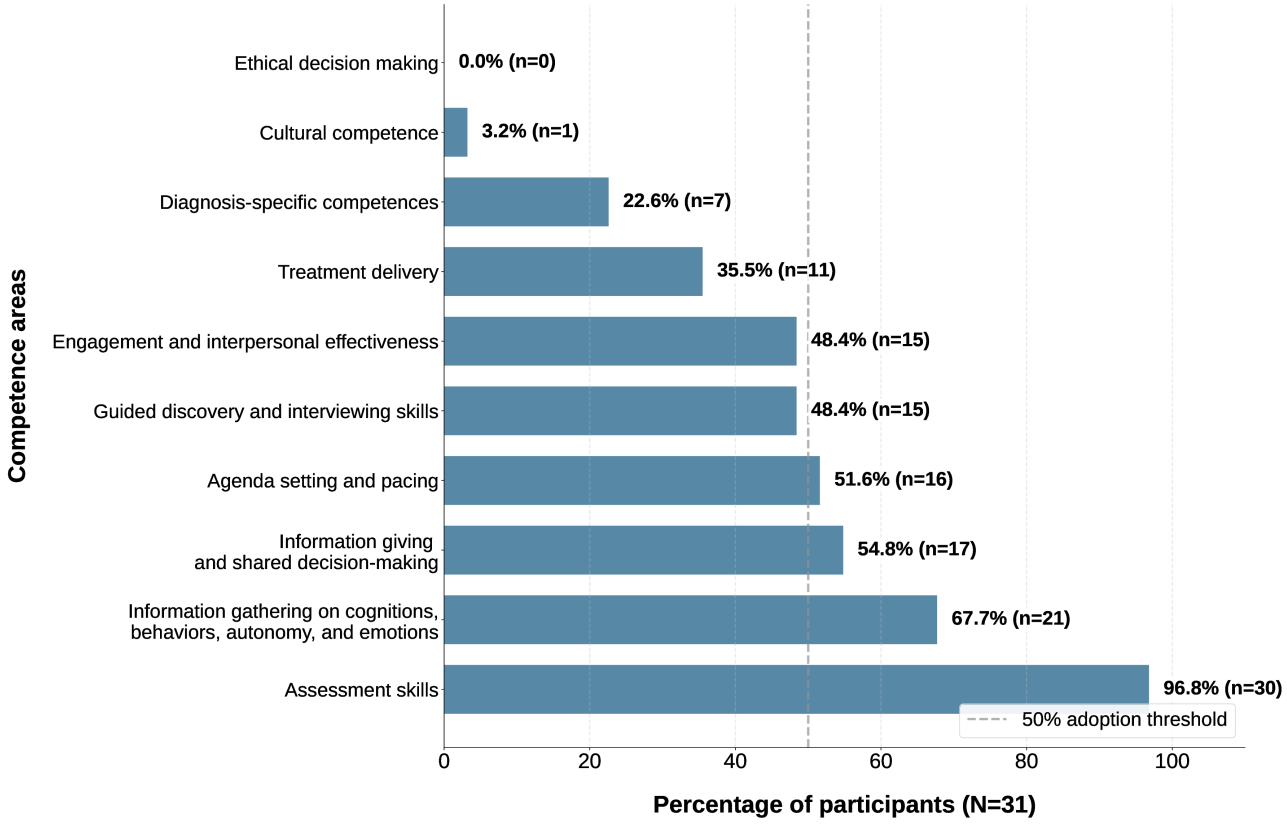


The distribution of self-reported competence development varied considerably across the assessed domains (Figure 5). Some competency areas showed limited improvement, including ethical decision making (0/31, 0%) and cultural competence (1/31, 3.2%). The most frequently reported competence improvements (Figure 5) were in assessment skills (30/31, 96.8%), followed by information gathering on cognitions,

XSL•FO
RenderX

behaviors, autonomy, and emotions (21/31, 67.7%), and information giving and shared decision-making (517/31, 4.8%).

Figure 5. Competence development areas reported by CBT (cognitive behavioral therapy) Trainer users.



Qualitative Feedback on Learning Impact

Content analysis of qualitative feedback gathered through the open-ended question "Please provide one or two specific examples of how CBT Trainer impacted your clinical skills or knowledge" (n=21 responses) in the poststudy questionnaire revealed 38 distinct codes describing specific impacts on clinical skills and knowledge. These codes were clustered into 8 main categories, with a full description of each code provided in Multimedia Appendix 2.

Category 1: Skill Development

Participants reported improvements in specific clinical techniques, communication skills, assessment abilities, and session management. This was the most frequently cited category (16/21, 76.2%). Reported improvements in specific clinical techniques include risk assessment (4/21, 19%), funneling (3/21, 14.3%), and structured information gathering using frameworks like 4Ws and ABCs (2/21, 9.5%). As one participant described:

> *The function of being able to analyse your conversation with the patient and compare against the specific competencies was incredibly helpful... This really helped me to improve my information gathering skills using the 4Ws and ABCs.*" [P19]

Communication skills improvements included developing more flexible and adaptive responses (2/21, 9.5%) and refining questioning styles (2/21, 9.5%). One participant noted:

> *Normally I stutter when talking to patients and this helped me to think before I speak* [P12]

Many participants reported enhanced assessment skills, both generally (4/21, 19%) and specifically in understanding competency-based assessment (4/21, 19%). Improvements in session management were frequently described, including better pacing (4/21, 19%), time management (2/21, 9.5%), and session structure (2/21, 9.5%). Participant explained:

> *It also really helped me to improve my pacing of the assessment, as I was able to look at my timings afterwards to work out how much time I spent in each section.* [P19]

Category 2: Practice Accessibility and Autonomy

Nearly half of the participants (9/21, 42.9%) valued the accessibility and autonomy the platform provided. The most commonly cited benefit was independent practice (7/21, 33.3%), which reduced dependence on peers or family members. As one participant described:

> *I found it difficult to continuously practise with my friends and family, and it was a bit unfair, as it would take about 45mins of their free time... Being able to use CBT Trainer meant that I could practise an entire assessment repeatedly.* [P3]

Flexible scheduling (4/21, 19%) and the ability to engage in repeated practice (3/21, 14.3%) were also valued features.

XSL•FO
RenderX

Category 3: Feedback and Competence Assessment

More than two-thirds of participants (13/21, 61.9%) referenced the value of structured feedback aligned with competence frameworks. Competence gap identification was the most frequently mentioned benefit (8/21, 38.1%), followed by competence framework alignment (6/21, 28.6%). A participant explained:

> *It gave me an understanding of where I was and wasn't hitting the mark scheme. For example, where there wasn't enough evidence of covering the confidentiality criteria.* [P10]

Constructive, personalized feedback was valued by several participants (5/21, 23.8%), with Participant 19 noting the platform "gave personalized tips for each competence."

Category 4: Confidence, Preparedness, and Anxiety Reduction

More than half of participants (11/21, 52.4%) reported that the CBT Trainer enhanced their confidence and preparedness while reducing anxiety. Exam and OSCE preparation (5/21, 23.8%) and confidence building (5/21, 23.8%) were equally prominent. Participant described:

> *Knowing that I could practice anytime of the day made me feel so much more calm about my exam as I didn't have to rely on other people to practice with.* [P15]

Anxiety reduction was specifically mentioned by 3 participants (14.3%), while preparation for clinical placements was noted by 2 (9.5%).

Category 5: Psychological Safety and Judgment-Free Learning

A small number of participants (2/21, 9.5%) explicitly referenced the psychological safety the platform provided. Participant 1 valued being able to "practice and experiment with new techniques and skills during training without fear of judgment."

Category 6: Diversity and Presentation-Specific Learning

Approximately one-third of participants (6/21, 28.6%) valued exposure to diverse patient presentations (4/21, 19%) and different presenting problems (3/21, 14.3%). Participant 15 noted: "The diversity of the patients was also helpful to prepare for real life." Some participants specifically valued practice with challenging scenarios (3/21, 14.3%) and opportunities for culturally sensitive practice (1/21, 4.8%), such as working with patients from different religious backgrounds.

Category 7: Reflection and Self-Awareness

Over one-third of participants (7/21, 33.3%) reported enhanced reflection and self-awareness. Self-awareness of performance was the most common impact (5/21, 23.8%), followed by the ability to engage in conversation analysis (3/21, 14.3%). One participant described:

> *The competence section was extremely helpful as it helped me to recognize areas I missed or did not go over during my roleplay.* [P20]

Category 8: No Perceived Impact

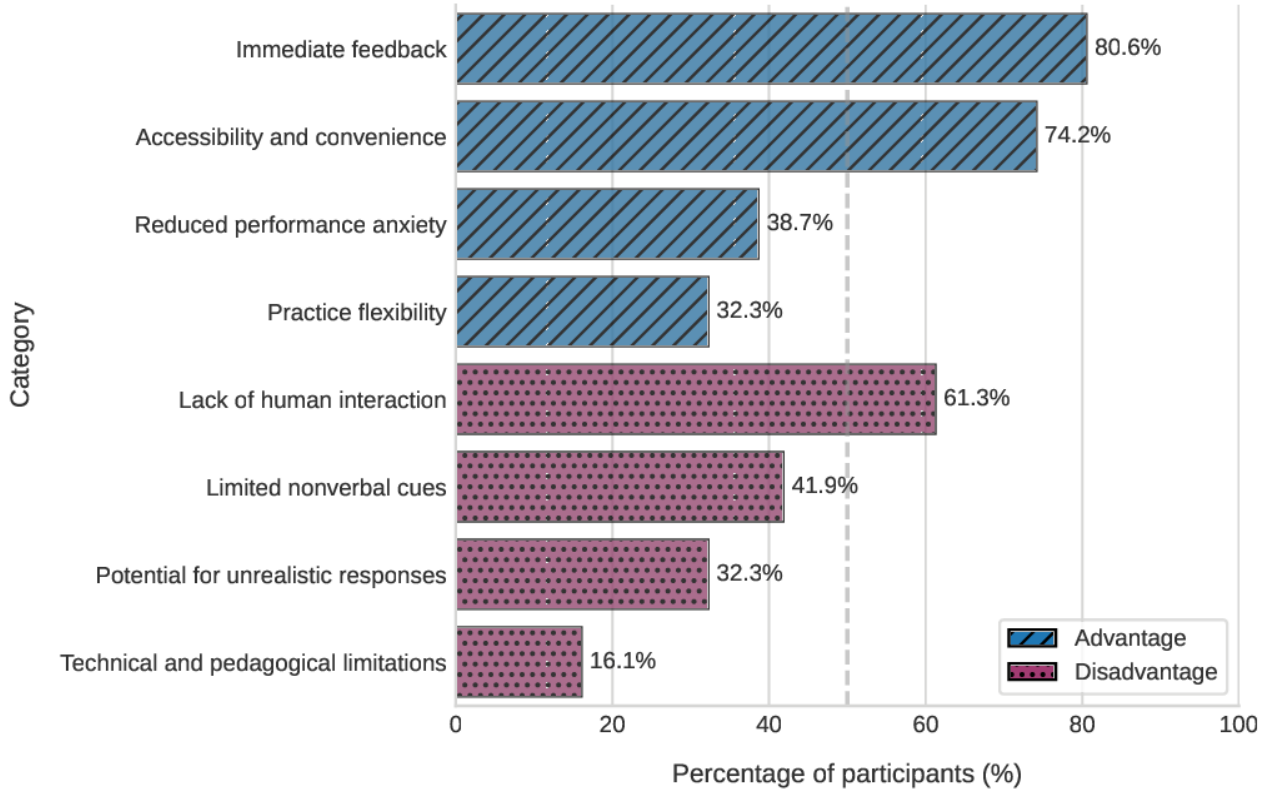
One participant (4.8%) reported no perceived impact on clinical skills or knowledge from using CBT Trainer. Participants also identified limitations of the platform. The most frequently cited limitation was the absence of nonverbal communication cues (7/21, 33%), with participants noting they "can't read body language" (P15) and missing "micro expressions" that inform real clinical work. Several participants found patient responses occasionally unrealistic (5/21, 24%), describing virtual patients as "too willing to engage compared to real clients" (P9) or "overly positive about interventions" (P1). Technical issues with voice recognition were mentioned by 14% (n=3), and some desired greater diversity in clinical presentations (2/21, 10%).

Comparison to Traditional Role-Play

Participants were asked to evaluate "How does CBT Trainer compare to traditional role-play exercises with peers? (0=Traditional methods much better, 10=CBT Trainer much better)" and rated CBT Trainer against traditional classroom role-play with a mean score of 5.90 (SD 1.94; median 6.0, IQR 5.0-6.0, n=31) on a 10-point scale favoring CBT Trainer, indicating moderate comparability. Figure 6 shows participants' perceptions of the comparative advantages and disadvantages of CBT Trainer relative to traditional role-play methods.

XSL•FO
RenderX

Figure 6. Perceived advantages and disadvantages of CBT Trainer compared to traditional training methods by CBT (cognitive behavioral therapy) Trainer users (n=31).



Primary and Secondary Outcomes

All prespecified implementation targets were achieved. For platform engagement, 79.66% (n=47) of 59 active users engaged with the platform for ≥10 minutes, and 59.32% (n=35) used it for ≥30 minutes, both exceeding target thresholds. Perceived educational value similarly exceeded the target of >70% reporting positive impact (rating >50/100) across all 3 educational value statements: 90.3% (28/31) for simulated patient interactions being helpful for learning, 71.0% (22/31) for simulated patients being realistic to real patients, and 83.3% (25/30) for clinical skills improved after using CBT Trainer of participants rating each statement above 50/100, respectively. Recruitment and retention substantially exceeded targets with 101 initial participants and 59 active users. The mean SUS score of 82.2 (SD 12.93) was well above the target threshold of >68.

Discussion

Principal Findings

This pilot study evaluated CBT Trainer, the first AI-driven virtual patient platform to provide real-time feedback aligned with established competence frameworks (CTS-R and UCL PWP Assessment Scale) in psychological practitioner training. It is also the first study to examine voluntary engagement patterns with such simulations within real-world training programs. The platform demonstrated strong feasibility and acceptability, with most participants voluntarily engaging in extended practice sessions—indicating meaningful opportunities for therapeutic skill development. Excellent usability ratings were achieved, and participants reported substantial educational value, particularly in foundational clinical skills. Self-reported improvements were most notable in assessment competencies, where nearly all participants reported enhanced assessment skills and most noted gains in information gathering. Qualitative analysis revealed that participants valued the platform's immediate feedback, convenience, and judgment-free practice environment as key advantages over traditional peer role-plays, while acknowledging limitations such as the absence of nonverbal cues. We discuss these results in relation to existing literature, consider their implications for psychological training, and discuss limitations and future research directions.

Feasibility and Acceptability (Aim 1)

The study provides evidence supporting the acceptability and feasibility of implementing AI-based virtual patient simulations with competency-aligned feedback in psychological therapist training programs that use CBT. Participant engagement exceeded prespecified targets, with substantial uptake and active involvement in role-play interactions. The platform achieved excellent usability ratings, further supporting its feasibility for integration into clinical training environments [39]. These findings align with prior research on AI-based simulated patients in health care education [26,29] and extend them to psychological therapist trainees who use CBT specifically, demonstrating that this population similarly values interactive simulations as learning tools.

Patterns of Voluntary Engagement (Aim 2)

During the voluntary 1-month access period, engagement patterns were diverse, reflecting individualized approaches to simulation-based learning. While most participants opted for

XSL•FO
RenderX

shorter, focused practice sessions, a smaller subset engaged in extended periods of use, spanning several hours. This variability suggests that trainees may adopt distinct engagement strategies—whether for brief skill rehearsal or more immersive, in-depth practice. The duration and interaction density of role-play sessions were substantial enough to facilitate meaningful skill development across multiple competency domains. No substantial differences were observed in engagement patterns, demographics, or outcomes between the 2 training programs. This overall engagement profile indicates that trainees perceived the platform as a valuable supplement to traditional training and that the simulations provided a sufficiently authentic approximation of therapeutic dialogue in both depth and communicative substance.

Self-Reported Competence Development (Aim 3)

Participants reported substantial improvement in assessment skills and information gathering, indicating that the platform effectively supports the development of fundamental therapeutic competencies. These improvements in core clinical skills, supported by strong ratings for educational value and perceived clinical skill improvement, mirror a similar study that observed significant gains in reflection skills among users of their patient-like conversational agent [28]. The reported benefits extended beyond assessment to include domains such as risk assessment, therapeutic pacing, and interpersonal effectiveness, suggesting that virtual patient interactions can facilitate development across multiple competence areas, in line with evidence that structured AI simulations can enhance targeted clinical skills [46]. However, the variability in perceived improvement competency domains implies that certain therapeutic skills may be more amenable to technological simulation than others, pointing to both the potential and the boundaries of AI-driven training tools.

Educational Impact and Perceived Value (Aim 4)

Self-Reported Educational Impact

CBT Trainer's capacity to facilitate immediate reflection on therapeutic interactions aligns with the principles of experiential learning [33], particularly enhancing the critical reflective observation and abstract conceptualization phases that bridge concrete experience and active experimentation. This immediacy addresses a fundamental requirement of deliberate practice [34], where skill acquisition depends on timely opportunities for reflection and refinement following performance attempts.

Participants consistently highlighted the benefit of practicing in a judgment-free environment. This aspect addresses several persistent challenges in traditional training, including performance anxiety during role-plays [18,22], evaluation apprehension inhibiting reflection [47] and delayed feedback cycles that hamper systematic improvement of therapeutic microskills [7,48]. Participants' emphasis on the judgment-free practice environment suggests that the platform can mitigate these barriers concurrently, thereby fostering the deliberate, reflective skill development that is often constrained in conventional training formats.

Comparative Advantages and Challenges

The balanced perspective offered by participants when comparing AI-based practice to traditional role-play methods highlights the complementary nature of these approaches. When asked about the platform's strengths, participants most frequently identified immediate feedback and convenience as significant educational benefits. Many users also valued the reduced performance anxiety that virtual practice offered compared to peer-based role-plays. Consistent with CBT principles of graded exposure, the platform may serve as an initial step that builds foundational skills before trainees progress to potentially more anxiety-provoking live role-play practice. The pattern of responses regarding comparative advantages aligns with the "pedagogical affordances" framework [49], which suggests that different educational technologies offer unique benefits for specific learning objectives. The participants' nuanced assessment of when and how the platform added value to their training reflects a sophisticated understanding of how technological tools fit within their broader educational experience.

Participants also recognized certain limitations of the technology, particularly regarding interpersonal elements. The recognized lack of nonverbal cues and potential for unrealistic responses echo concerns raised regarding the authenticity of AI-based clinical simulations [30]. These limitations reinforce that current AI technology serves as a complement to, rather than a replacement for, conventional training models incorporating interpersonal engagement and supervision.

Limitations and Future Directions

Our findings should be interpreted with several limitations in mind. Technical limitations restricting participation to iOS users may have introduced selection bias. A substantial proportion of eligible participants who initially expressed interest did not progress to downloading or using the app. This pattern of attrition suggests the importance of understanding barriers to technology adoption in clinical training contexts, whether technical, motivational, or practical in nature. Future research should systematically investigate factors influencing technology adoption among clinical trainees, including surveys of nonusers to identify specific barriers and preferences.

Participants were predominantly early-stage trainees who had not yet received formal training in intervention techniques. This may explain why self-reported improvements were concentrated in assessment and information gathering skills rather than intervention delivery competencies. In addition, participants were trainee PWPs and clinical psychologists using CBT, not trainees from CBT-specific programs. Future research should evaluate the platform's utility in specialized CBT training contexts.

Methodologically, our mixed methods approach provided rich experiential data, but the absence of a control group prevents causal inferences about the platform's impact on skill development. The study relied on self-reported skill improvement rather than objective competence assessment, weakening conclusions about educational efficacy. The 1-month evaluation period offers limited insight into long-term

XSL•FO
RenderX

engagement patterns. While exploratory analyses revealed no substantial differences in engagement between the 2 training programs, we did not systematically examine how prior CBT experience might influence engagement patterns. It is plausible that trainees with more CBT experience might engage differently than those new to CBT, either showing reduced engagement due to perceived redundancy or increased engagement due to greater appreciation of competence frameworks. Future evaluations should incorporate standardized competence evaluations, longitudinal designs across training stages, control of prior experience, and randomized controlled trials.

While the platform operationalized general CBT competencies through established frameworks such as the CTS-R and UCL PWP Assessment Scale, it did not explicitly integrate disorder-specific evidence-based models into its feedback mechanisms. For example, in interactions with the virtual patient presenting with generalized anxiety disorder, the platform assessed competencies such as application of change methods (CTS-R item 11) but did not evaluate whether the trainee's intervention strategies aligned with a specific evidence-based model for generalized anxiety disorder (eg, the Dugas intolerance of uncertainty model). Future studies could incorporate disorder-specific competence criteria and enable feedback on the appropriateness and fidelity of evidence-based intervention models, thereby enhancing their alignment with specialized CBT training requirements.

While we implemented several safeguards, the current model still places some responsibility on trainees to identify inappropriate feedback. This represents a limitation, as novice trainees may lack the expertise to critically evaluate AI-generated feedback. Future implementations should consider automated feedback auditing systems, routine expert review of AI-generated feedback, and structured guidance for trainees on how to critically engage with AI feedback within supervision.

While AI-driven simulations offer significant benefits for CBT training, their deployment must address known risks, including algorithmic bias, privacy concerns, and overreliance on technology that could erode professional judgment [50]. Comprehensive evaluation of AI training tools requires continued attention to fairness metrics, cultural appropriateness across diverse populations, and environmental sustainability—areas warranting further research as the field evolves.

The core methodology—AI simulation with competence framework–aligned feedback—may extend beyond CBT to other health care disciplines with established competence frameworks, including medicine, nursing, social work, and allied health professions. Adaptation would require profession-specific scenarios, alignment with relevant competence standards, and validation that simulation skills transfer to clinical practice. Disciplines with well-defined assessment criteria would be particularly suited to this approach, and empirical evaluation across training contexts is needed.

Conclusions

The pilot study demonstrates that CBT Trainer—an AI-driven virtual patient platform delivering real-time competency-aligned feedback—is both acceptable and feasible within psychological training programs that use CBT. Participants engaged voluntarily and extensively with the platform, reporting high usability and significant educational value, particularly in core assessment and information gathering skills. Qualitative analysis revealed that participants highly valued the competence framework–aligned feedback, a judgment-free practice environment that reduced anxiety, and the flexibility for independent, repeated skill rehearsal. These features were seen as key advantages over traditional peer-based role-plays. The findings indicate that AI-based virtual patient simulations with competence framework–aligned feedback can effectively complement traditional training methods, particularly for developing foundational clinical skills in a flexible, accessible format. Supplementary tools like this can be valuable in contexts facing supervision shortages or growing demand for mental health services. We recommend phased adoption with appropriate oversight mechanisms while simultaneously conducting the rigorous trials necessary to build the evidence base. Additionally, we emphasize the importance of involving course leaders, clinical supervisors, competency scale and framework experts, and policymakers in the ongoing refinement and validation. Future research should use randomized controlled designs with objective competence assessments and evaluate the platform's utility in specialized mental health training programs with more advanced trainees.

Acknowledgments

We thank Dr Rachel Newman and Dr Joshua Buckman for their support during participant recruitment, and thank all participants for their participation.

Funding

No external financial support or grants were received from any public, commercial, or not-for-profit entities for the research, authorship, or publication of this article.

Data Availability

The datasets generated or analyzed during this study are available from the corresponding author on reasonable request.

XSL•FO
RenderX

Conflicts of Interest

TTZ is the founder and director of Soy Mental Health Educational Technologies. TTZ is a PhD student working under the supervision of the other authors who maintain independent scrutiny of the study. The other authors do not have a conflict of interest.

Multimedia Appendix 1
Poststudy survey instruments.
[DOCX File, 19 KB - mededu_v12i1e84091_app1.docx]

Multimedia Appendix 2
Description and frequencies of impact codes.
[DOCX File, 20 KB - mededu_v12i1e84091_app2.docx]

References

1. NHS long term workforce plan. NHS England. 2023. URL: https://www.england.nhs.uk/long-read/nhs-long-term-workforce-plan-2/ [accessed 2025-05-23]
2. Roth AD, Pilling S. Using an evidence-based methodology to identify the competences required to deliver effective cognitive and behavioural therapy for depression and anxiety disorders. Behav Cogn Psychother 2008 Mar;36(2):129-147. [doi: 10.1017/S1352465808004141]
3. Roth AD. A new scale for the assessment of competences in cognitive and behavioural therapy. Behav Cogn Psychother 2016 Sep;44(5):620-624. [doi: 10.1017/S1352465816000011] [Medline: 26898543]
4. Blackburn IM, James IA, Milne DL, et al. Revised cognitive therapy scale (CTS-r). APA PsycTests. 2001. URL: https://doi.org/10.1037/t65807-000 [accessed 2026-02-25]
5. Young J, Beck A. Cognitive Therapy Scale (CTS). APA PsycTests. 1980. URL: https://doi.org/10.1037/t00834-000 [accessed 2026-02-25]
6. Reichelt FK, James IA, Blackburn IM. Impact of training on rating competence in cognitive therapy. J Behav Ther Exp Psychiatry 2003 Jun;34(2):87-99. [doi: 10.1016/s0005-7916(03)00022-3] [Medline: 12899893]
7. Bennett-Levy J. Therapist skills: a cognitive model of their acquisition and refinement. Behav Cogn Psychother 2006 Jan;34(1):57-78. [doi: 10.1017/S1352465805002420]
8. Muse K, Kennerley H, McManus F. The why, what, when, who and how of assessing CBT competence to support lifelong learning. tCBT 2022;15:e57. [doi: 10.1017/S1754470X22000502]
9. Muse K, McManus F. A systematic review of methods for assessing competence in cognitive-behavioural therapy. Clin Psychol Rev 2013 Apr;33(3):484-499. [doi: 10.1016/j.cpr.2013.01.010] [Medline: 23454222]
10. Liness S, Beale S, Lea S, Byrne S, Hirsch CR, Clark DM. Evaluating CBT clinical competence with standardised role plays and patient therapy sessions. Cogn Ther Res 2019 Dec;43(6):959-970. [doi: 10.1007/s10608-019-10024-z]
11. Sharpless BA, Barber JP. The Examination for Professional Practice in Psychology (EPPP) in the era of evidence-based practice. Prof Psychol Res Pr 2009;40(4):333-340. [doi: 10.1037/a0013983]
12. Marriott BR, Cho E, Tugendrajch SK, et al. Role-play assessment of therapist adherence and skill in implementation of trauma-focused cognitive-behavioral therapy. Adm Policy Ment Health 2022 May;49(3):374-384. [doi: 10.1007/s10488-021-01169-9] [Medline: 34546482]
13. Loades ME, Armstrong P. The challenge of training supervisors to use direct assessments of clinical competence in CBT consistently: a systematic review and exploratory training study. Cogn Behav Ther 2016;9:e27. [doi: 10.1017/S1754470X15000288]
14. Johnson J, Corker C, O'connor DB. Burnout in psychological therapists: a cross‐sectional study investigating the role of supervisory relationship quality. Clin Psychol (Aust Psychol Soc) 2020 Nov 1;24(3):223-235. [doi: 10.1111/cp.12206]
15. The British Psychological Society. Code of Ethics and Conduct: The British Psychological Society; 2021. [doi: 10.53841/bpsrep.2021.inf94]
16. The British Psychological Society. Supervision Guidance for Psychologists: The British Psychological Society; 2024. [doi: 10.53841/bpsrep.2024.rep178]
17. Caltabiano M, Errington E, Sorin R, Nickson A, et al. The potential of role-play in undergraduate psychology training / marie caltabiano. Asian J Univ Educ. Preprint posted online on Jun 30, 2018 URL: https://files.eric.ed.gov/fulltext/EJ1207768.pdf [accessed 2026-02-21]
18. Melluish S, Crossley J, Tweed A. An evaluation of the use of simulated patient role-plays in the teaching and assessment of clinical consultation skills in clinical psychologists' training. Psychol Learn Teach 2007 Sep;6(2):104-113. [doi: 10.2304/plat.2007.6.2.104]
19. Doolen J, Giddings M, Johnson M, Guizado de Nathan G, O Badia L. An evaluation of mental health simulation with standardized patients. Int J Nurs Educ Scholarsh 2014 Mar 12;11(1):55-62. [doi: 10.1515/ijnes-2013-0075] [Medline: 24620017]

XSL•FO
RenderX

20. Himmelbauer M, Seitz T, Seidman C, Löffler-Stastka H. Standardized patients in psychiatry - the best way to learn clinical skills? BMC Med Educ 2018 Apr 6;18(1):72. [doi: 10.1186/s12909-018-1184-4] [Medline: 29625572]
21. Badger LW, Macneil G. Rationale for utilizing standardized clients in the training and evaluation of social work students. J Teach Soc Work 1998 Jan 26;16(1-2):203-218. [doi: 10.1300/J067v16n01_13]
22. Fall KA, Levitov JE. Using actors in experiential group counseling leadership training. J Spec Group Work 2002 Jun;27(2):122-135. [doi: 10.1177/0193392202027002002]
23. Pilnick A, Trusson D, Beeke S, O'Brien R, Goldberg S, Harwood RH. Using conversation analysis to inform role play and simulated interaction in communications skills training for healthcare professionals: identifying avenues for further development through a scoping review. BMC Med Educ 2018 Nov 19;18(1):267. [doi: 10.1186/s12909-018-1381-1] [Medline: 30453956]
24. Ruiz Rodríguez J, Bados López A, Fusté Escolano A, et al. Peer counselling versus role-playing: two training methods of therapeutic skills in clinical psychology. Psicothema 2018 Feb;30(1):21-26. [doi: 10.7334/psicothema2016.286] [Medline: 29363466]
25. Englar RE. Using a standardized client encounter to practice death notification after the unexpected death of a feline patient following routine ovariohysterectomy. J Vet Med Educ 2019;46(4):489-505. [doi: 10.3138/jvme.0817-111r1] [Medline: 30806560]
26. Prescott J, Ogilvie L, Hanley T. Student therapists' experiences of learning using a machine client: a proof‐of‐concept exploration of an emotionally responsive interactive client (ERIC). Couns and Psychother Res 2024 Jun;24(2):524-531 [FREE Full text] [doi: 10.1002/capr.12685]
27. Zalewski B, Guziak M, Walkiewicz M. Developing simulated and virtual patients in psychological assessment - method, insights and recommendations. Perspect Med Educ 2023;12(1):455-461. [doi: 10.5334/pme.493] [Medline: 37901884]
28. Tanana MJ, Soma CS, Srikumar V, Atkins DC, Imel ZE. Development and evaluation of ClientBot: patient-like conversational agent to train basic counseling skills. J Med Internet Res 2019 Jul 15;21(7):e12529. [doi: 10.2196/12529] [Medline: 31309929]
29. De Mattei L, Morato MQ, Sidhu V, et al. Are Artificial Intelligence Virtual Simulated Patients (AI-VSP) a valid teaching modality for health professional students? Clin Simul Nurs 2024 Jul;92:101536. [doi: 10.1016/j.ecns.2024.101536]
30. Maurya RK. A qualitative content analysis of chatgpt's client simulation role play for practicing counseling skills. PsyArXiv. Preprint posted online on Jun, 2023. [doi: 10.31234/osf.io/vwuxh]
31. Creed TA, Salama L, Slevin R, et al. Enhancing the quality of cognitive behavioral therapy in community mental health through artificial intelligence generated fidelity feedback (Project AFFECT): a study protocol. BMC Health Serv Res 2022 Sep 20;22(1):1177. [doi: 10.1186/s12913-022-08519-9] [Medline: 36127689]
32. Flemotomos N, Martinez VR, Chen Z, Creed TA, Atkins DC, Narayanan S. Automated quality assessment of cognitive behavioral therapy sessions through highly contextualized language representations. PLoS One 2021;16(10):e0258639. [doi: 10.1371/journal.pone.0258639] [Medline: 34679105]
33. Kolb D. Experiential Learning: Experience As The Source Of Learning And Development: Prentice Hall; 1984.
34. Ericsson KA, Krampe RT, Tesch-Römer C. The role of deliberate practice in the acquisition of expert performance. Psychol Rev 1993;100(3):363-406. [doi: 10.1037/0033-295X.100.3.363]
35. Rakovshik SG, McManus F. Establishing evidence-based training in cognitive behavioral therapy: a review of current empirical findings and theoretical guidance. Clin Psychol Rev 2010 Jul;30(5):496-516. [doi: 10.1016/j.cpr.2010.03.004] [Medline: 20488599]
36. AI-powered psychotherapy training | virtual patient roleplay simulation. CBT Trainer. 2025. URL: https://www.soymh.com [accessed 2026-02-21]
37. National Curriculum for Psychological Wellbeing Practitioner (PWP) programmes. NHS England. 2022. URL: https://www.hee.nhs.uk/sites/default/files/2026-01/PWP%20Curriculum%204th%20Edition%202022%20%28updated%20May%202025%29%20v4.4.docx [accessed 2026-02-21]
38. Brooke J. SUS- a quick and dirty usability scale. In: Usability Evaluation in Industry: Taylor & Francis; 1996:189-194.
39. Bangor A, Kortum PT, Miller JT. An empirical evaluation of the System Usability Scale. Int J Hum Comput Int 2008 Jul 29;24(6):574-594. [doi: 10.1080/10447310802205776]
40. Hsieh HF, Shannon SE. Three approaches to qualitative content analysis. Qual Health Res 2005 Nov;15(9):1277-1288. [doi: 10.1177/1049732305276687] [Medline: 16204405]
41. Graneheim UH, Lundman B. Qualitative content analysis in nursing research: concepts, procedures and measures to achieve trustworthiness. Nurse Educ Today 2004 Feb;24(2):105-112. [doi: 10.1016/j.nedt.2003.10.001] [Medline: 14769454]
42. Krippendorff K. Content Analysis: An Introduction to Its Methodology: SAGE Publications, Inc; 2019. [doi: 10.4135/9781071878781]
43. Elo S, Kyngäs H. The qualitative content analysis process. J Adv Nurs 2008 Apr;62(1):107-115. [doi: 10.1111/j.1365-2648.2007.04569.x] [Medline: 18352969]
44. Pinnock H, Barwick M, Carpenter CR, et al. Standards for Reporting Implementation Studies (StaRI) Statement. BMJ 2017 Mar 6;356:i6795. [doi: 10.1136/bmj.i6795] [Medline: 28264797]

XSL•FO
RenderX

45. AI role-plays for improving trainee competence in CBT: an feasibility and assessment study. OSF. URL: https://osf.io/mskb7 [accessed 2026-02-14]
46. Wang R, Milani S, Chiu JC, et al. PATIENT-ψ: using large language models to simulate patients for training mental health professionals. arXiv. Preprint posted online on Oct 3, 2024. [doi: 10.48550/arXiv.2405.19660]
47. Rapisarda C, Jencius M, McGlothlin J. Master's students' experiences in a multicultural counseling role-play. Int J Adv Counselling 2011 Dec;33(4):361-375. [doi: 10.1007/s10447-011-9139-z]
48. Chow DL, Miller SD, Seidel JA, Kane RT, Thornton JA, Andrews WP. The role of deliberate practice in the development of highly effective psychotherapists. Psychotherapy (Chic) 2015 Sep;52(3):337-345. [doi: 10.1037/pst0000015] [Medline: 26301425]
49. Laurillard D. Rethinking University Teaching, 2nd edition: London: Routledge. [doi: 10.4324/9781315012940]
50. Shafran R, Bond L, Carlbring P, et al. From innovation to implementation: artificial intelligence in cognitive behaviour therapy training and supervision. Behav Res Ther 2026 Feb;197. [doi: 10.1016/j.brat.2025.104945]

Abbreviations

AI: artificial intelligence
CBT: cognitive behavioral therapy
CORE: Center for Outcomes Research and Effectiveness
CTRS: Cognitive Therapy Rating Scale
CTS-R: Cognitive Therapy Scale-Revised
NHS: National Health Service
PWP: Psychological Wellbeing Practitioner
SUS: System Usability Scale
UCL: University College London

Edited by S Tsuei; submitted 22.Sep.2025; peer-reviewed by A Churchard, VL Presley; accepted 30.Jan.2026; published 06.Mar.2026.

Please cite as:
Zhang TT, Saunders R, Pilling S, O'Driscoll C
An AI-Driven Virtual Patient Platform (CBT Trainer) for Training Cognitive Behavioral Therapy Practitioners Against Competencies: Mixed Methods Pilot Study
JMIR Med Educ 2026;12:e84091
URL: https://mededu.jmir.org/2026/1/e84091
doi:10.2196/84091

© Tianyu Terry Zhang, Rob Saunders, Stephen Pilling, Ciarán O'Driscoll. Originally published in JMIR Medical Education (https://mededu.jmir.org), 6.Mar.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Education, is properly cited. The complete bibliographic information, a link to the original publication on https://mededu.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Performance Evaluation of Large Language Models in Multilingual Medical Multiple-Choice Questions: Mixed Methods Study

Livia Maria Strasser[1*], BMed, MMed; Wilma Anschuetz[2*], MD; Fabio Dennstädt[1,3], MD; Janna Hastings[1,4,5], BSc, MSc, MA, PhD

[1]Medical Knowledge and Decision Support, School of Medicine, University of St.Gallen, St.Jakobstrasse 21, St.Gallen, Switzerland

[2]Department for Assessment and Evaluation, Institute for Medical Education, University of Bern, Bern, Switzerland

[3]Department of Radiation Oncology, Inselspital, Bern University Hospital and University of Bern, Bern, Switzerland

[4]Institute for Implementation Science in Health Care, University of Zurich, Zurich, Switzerland

[5]Idiap Research Institute, Martigny, Switzerland

[*]these authors contributed equally

Corresponding Author:
Livia Maria Strasser, BMed, MMed
Medical Knowledge and Decision Support, School of Medicine, University of St.Gallen, St.Jakobstrasse 21, St.Gallen, Switzerland

Abstract

Background: Artificial intelligence continues to transform health care, offering promising applications in clinical practice and medical education. While large language models (LLMs), as a form of generative artificial intelligence, have shown potential to match or surpass medical students in licensing examinations, their performance varies across languages. Recent studies highlight the complex influence and interdependency of factors such as language and model type on LLMs' accuracy; yet, cross-language comparisons remain underexplored.

Objective: This study evaluates the performance of LLMs in answering medical multiple-choice questions quantitatively and qualitatively across 3 languages (German, French, and Italian), aiming to uncover model capabilities in a multilingual medical education context.

Methods: For this mixed methods study, 114 publicly accessible multiple-choice questions in German, French, and Italian from an online self-assessment tool were analyzed. A quantitative performance analysis of several LLMs developed by OpenAI, Meta AI, Anthropic, and DeepSeek was conducted to evaluate their performance on answering the questions in text-only format. For the comparative analysis, a variation of input question language (German, French, and Italian) and prompt language (English vs language-matched) was used. The 2 best-performing LLMs were then prompted to provide answer explanations for incorrectly answered questions. A subsequent qualitative analysis was conducted on these explanations to identify the reasons leading to the incorrect answers.

Results: The performance of LLMs in answering medical multiple-choice questions varied by model and language, showing substantial differences in accuracy (between 64% and 87%). The effect of input question language was significant (P<.01) with models performing best on German questions. Across the analyzed LLMs, prompting in English generally led to better performance in comparison to language-matched prompts, but the top-performing models exceptionally showed comparable results for language-matched prompts. Qualitative analysis revealed that answer explanations of the analyzed models (GPT4o and Claude-Sonnet-3.7) showed different reasoning errors. In several explanations, this occurred despite factual accuracy on the represented topic. Furthermore, this analysis revealed 3 questions to be insufficiently precise.

Conclusions: Our results underline the potential of LLMs in answering medical examination questions and highlight the importance of careful consideration of model choice, prompt, and input languages, because of relevant performance variability across these factors. Analysis of answer explanations demonstrates a valuable use case of LLMs for improving examination question quality in medical education, if data security regulations permit their use. Human oversight of language-sensitive or clinically nuanced content remains essential to determine whether incorrect output stems from flaws in the questions themselves or from errors generated by the LLMs. There is a need for ongoing evaluation as well as transparent reporting to ensure reliable integration of LLMs into medical education contexts.

(JMIR Med Educ 2026;12:e81399) doi:10.2196/81399

KEYWORDS

medical question answering; LLM; LLM evaluation; education; natural language processing; multiple-choice questions; large language model

XSL•FO
RenderX

Introduction

Artificial intelligence (AI) in the form of large language models (LLMs) is rapidly advancing, with multiple applications across the health care sector [1-3]. Despite promising advances, the quality of developed systems remains variable, and extensive evaluation is required for different applications. When adequately implemented, LLMs exhibit substantial potential for supporting clinical practice [4,5]. In the domain of medical decision-making, AI has been shown to perform equivalently to specialists for some tasks [6] and to perform effectively in multiple languages, although performance varies across languages, with English being the predominant training language of most LLMs [7-10].

In the context of medical education, multiple studies have demonstrated that LLMs are able to perform at a level comparable to or even exceeding that of medical students in answering text-based medical licensing multiple-choice questions [11-14]. A comprehensive meta-analysis provided an integrative overview of 45 studies testing the performance of different ChatGPT (OpenAI) models for the task of answering medical licensing examination questions [15]. The findings of this study revealed that the different ChatGPT models exhibited varying levels of performance depending on factors such as the language of the questions (English vs non-English), the difficulty and length of the questions, as well as the type of questions and whether pictures were presented or not. However, the impact of these factors was not consistent between the different models (GPT-4 vs GPT-3.5), and the included studies did not use the same question sets across different languages. Moreover, interpreting model performance is limited by country-specific medical policies, unique local and cultural contexts, and ethical considerations in medical situations [16,17]. A few studies have isolated the effect of language on LLMs' performance by using comparable examination content across different languages. However, most existing comparisons rely on different national examinations, which vary in structure, limiting the validity of cross-linguistic assessments [18,19]. This study aims to address these gaps by comparing language-dependent performance on identical examination questions of a wide range of LLMs, including open models.

This study investigates how the performance of LLMs varies across languages (German, French, and Italian) in answering medical multiple-choice questions within Switzerland's unique multilingual context. Using a dataset of example medical licensing examination questions, we explore the performance of a range of different LLMs in answering these questions in German, French, and Italian and how the input and prompt language affects the performance in these different models. In addition, we qualitatively analyze and categorize the types of errors observed in the explanations given by LLMs.

Methods

Study Design and Data

We used publicly accessible example question sets on the self-assessment platform from the Institute of Medical Education of the Medical Faculty of Bern, which are provided by the Swiss Federal Office of Public Health to enable medical students to prepare for their examinations [20]. There are 3 question sets, each with 50 questions, resulting in a total of 150 questions. Each question is translated by professional medical translators into the respective examination languages, German, French, and Italian. The provided question sets are comparable in the range of medical knowledge topics and in difficulty level with the federal medical licensing examination in Switzerland. The correct answers to these questions are not publicly accessible. Only when the self-assessment test is submitted does the system provide information about the correctness of the answer given, but not about the answer key. This makes it unlikely that LLMs have been trained with these data. Of the 150 questions, 114 are text-only, while 35 include text and an image. One question included a computed tomography scan video.

Questions containing images or videos were excluded from analysis, as the focus of this study is on multilingual language-processing abilities. Therefore, we restricted the analysis to language-based (text-only) interactions and did not include image-processing or multimodal capabilities. Although recent work has shown substantial performance gains in vision-enabled large multimodal models, incorporating visual items introduces additional variance unrelated to linguistic processing [13,21-24].

The questions were further categorized into subcategories based on medical specialties, purpose of the question, and urgency of the case. Medical specialties were subdivided into general medicine (including internal medicine, neurology, and other smaller subjects, eg, dermatology); surgery and traumatology; pediatrics; gynecology; public health; and other. To differentiate the purpose of the question, questions were categorized into diagnosis (focusing on differential diagnosis, test interpretation, clinical reasoning, physical examination, and diagnostic studies); treatment (addressing therapeutic interventions, medication selection, surgical approaches, and management protocols); and other (public health questions covering prevention, health care systems, and epidemiology). To distinguish urgency within the discussed cases, questions were classified as acute or chronic. Acute conditions were defined as those involving same-day presentations, symptoms lasting less than 1 week, trauma, or infections, whereas chronic conditions were defined as symptoms persisting for more than 1 week, the need for ongoing medication, chronic diseases, or elective scenarios. Furthermore, we categorized the questions based on the response format (single response or multiple responses). Each question was reviewed and categorized for this study by 1 researcher, and categorizations were reviewed and checked by the other researchers. Questions that could potentially fit multiple categories were assigned based on their primary focus or most immediate clinical relevance.

Table 1 shows the distribution of sample multiple-choice questions categorized by specialty, response format, and question modality (text-only, image, and video) across 3 different datasets: text-only (n=114), text and image (n=35), and text and video (n=1).

XSL•FO
RenderX

Table . Distribution of multiple-choice questions by subcategories, response format, and question modality (text-only, image, and video).

Categories	Text-only (n=114)	Text and image (n=35)	Text and video (n=1)
Specialties, n (%)			
General medicine	72 (63.16)	24 (68.57)	1 (100)
Surgery and traumatology	10 (8.77)	2 (5.71)	0 (0)
Pediatrics	13 (11.40)	6 (17.14)	0 (0)
Gynecology	8 (7.02)	2 (5.71)	0 (0)
Public health and other	10 (8.77)	1 (2.86)	0 (0)
Purpose, n (%)			
Diagnosis	57 (50.00)	25 (71.43)	1 (100)
Treatment	44 (38.60)	8 (22.86)	0 (0)
Other (eg, public health)	12 (10.53)	2 (5.71)	0 (0)
Urgency, n (%)			
Acute	49 (42.98)	20 (57.14)	1 (100)
Chronic	64 (56.14)	15 (42.86)	0 (0)
Response format, n (%)			
Single response	105 (92.20)	32(91.50)	1 (100)
Multiple response	9 (7.80)	3 (8.50)	0 (0)

Large Language Models

Five different LLMs were applied, as shown in Table 2.

We executed each model 10 times, with the temperature set to the default of 1, top-p to the default of 1, max_tokens set to 10, and no explicit seed set, to estimate the performance in a typical setting where a student is using the technology to support their studying. All source code for the model execution is available from the project's GitHub repository [25].

Table . List of LLMs[a] used in this study.

Model provider	Model name	Implementation
OpenAI	GPT4o	OpenAI API[b]
Meta AI	Llama 3.1-405b-Instruct	DeepInfra API
Meta AI	Llama 3.3-70b-Instruct	DeepInfra API
Anthropic	Claude-Sonnet-3.7 - 20250219	Anthropic API
DeepSeek	DeepSeek V3	DeepInfra API

[a]LLM: large language model.

[b]API: application programming interface.

Prompts

In this study, user prompts were structured multilingually. Medical questions were presented to the model in German, French, and Italian while maintaining identical system-level instructions. The LLMs' responses to these equivalent prompts in different languages of identical examination questions enabled direct cross-lingual performance comparison under controlled conditions.

Two different system and user instruction prompt scenarios were explored: English language system and user instruction prompts while specifying the questions in the different languages in which the examination questions were available (named "English prompts"), and system and user instruction prompts in the same language as the questions were specified (named "language-matched prompts"). This allowed us to contrastively evaluate the impact of designing an interactive medical question answering system based on the target language of the user (language-matched) or the language that the model has the most expertise in (English), with implications for the application of such models in international contexts (Textbox 1).

XSL•FO
RenderX

Textbox 1. System and user prompts for English and language-matched versions in German, French, and Italian.

English

- System prompt: "*You are a helpful medical question answering assistant. Please carefully follow the exact instructions, return only the answer letter or letters, and do not provide explanations.*"
- User prompt: "*Please answer the following multiple-choice question by selecting the correct response option or options. Question: {}. Response options: A - {}, B - {}, C - {}, D - {}, E - {}. Return only the letter or letters (A, B, C, D, E) corresponding to the correct answer or answers. The correct answer to the question is:*" where the specific question and answer options are inserted from the question dataset.

German

- System prompt: "*Du bist ein hilfreicher Assistent bei der Beantwortung medizinischer Fragen. Befolge sorgfältig die genauen Anweisungen, gib nur den/die Antwortbuchstaben an und füge keine Erklärungen hinzu.*"
- User prompt: "*Bitte beantworte die folgende Multiple-Choice-Frage, indem du die richtige/n Antwortoption/en auswählst. Frage: {}. Antwortoptionen: A - {}, B - {}, C - {}, D - {}, E - {}. Gib nur den/die Buchstaben (A, B, C, D, E) an, der/die der/den richtigen Antwort/en entspricht/entsprechen. Die richtige Antwort auf die Frage lautet:*"

French

- System prompt: "*Tu es un assistant utile qui répond à des questions médicales. Suis attentivement les instructions exactes, renvoie uniquement la ou les lettre/s de réponse et ne donne aucune explication.*"
- User prompt: "*Réponds à la question à choix multiples suivante en sélectionnant la ou les réponse/s correcte/s. Question: {}. Options de réponse: A - {}, B - {}, C - {}, D - {}, E - {}. Ne renvoie que la ou les lettre/s (A,B,C,D,E) correspondant/es à la/les réponse/s correcte/s. La réponse correcte à la question est:*"

Italian

- System prompt: "*Sei un assistente utile per rispondere a domande mediche. Segui attentamente le istruzioni esatte, riporta solo la lettera o le lettere di risposta e non aggiungere spiegazioni.*"
- User prompt: "*Rispondi alla seguente domanda a scelta multipla selezionando la o le opzioni di risposta corrette. Domanda: {}. Opzioni di risposta: A - {}, B - {}, C - {}, D - {}, E - {}. Riporta solo la lettera o le lettere (A, B, C, D, E) corrispondente/i alla risposta corretta o alle risposte corrette. La risposta corretta alla domanda è:*"

Quantitative Analysis of Model Performance

Model responses were scored in binary as correct or incorrect. For multiple-response items, an all-or-nothing scoring approach was applied: a response was considered correct only if all correct options were selected and no incorrect options were included; ordering of answer options was not considered relevant.

All quantitative data analyses were performed in Python (version 3.12; Python Software Foundation) within a Jupyter notebook running on a MacBook Pro M2 with 32 GB of RAM, while LLMs were executed via online application programming interfaces (APIs). Notable libraries used in the analysis are the Transformers library for language model execution (version 4.45.2; Hugging Face), the OpenAI API for cloud-based model execution (version 1.14.0), the DeepInfra API for executed hosted open models (version 1, via DeepInfra's OpenAI API), the Anthropic API for executing Anthropic models (version 0.42.0), and the pandas (version 2.2.1; pandas Development Team), and seaborn (version 0.13.2; seaborn Development Team) libraries for data analysis and visualization, respectively. We analyzed the data using generalized estimating equations, as implemented in the statsmodels.genmod library (version 0.14.4; statsmodels Development Team), with a logistic link function and exchangeable correlation structure, to account for the clustering of repeated observations within items. We fit separate generalized estimating equation models to examine the effects of categorical variables, including model, language, and prompt language, on accuracy, each accounting for clustering within items. We fitted separate models due to convergence considerations. However, the fully crossed design ensures effects are not confounded. Effect sizes are reported as odds ratios (OR) with 95% CIs. Differences were deemed significant at a P value of <.05. Benjamini-Hochberg correction was applied to control the false discovery rate in subgroup analyses. All source code for the quantitative data analysis is available from the project's GitHub repository [25].

Qualitative Analysis of Reasons for Errors

For the qualitative evaluation, the 2 best-performing LLMs from the quantitative analysis, OpenAI's GPT4o and Anthropic's Claude-Sonnet-3.7, were selected. Questions answered incorrectly during this quantitative analysis were reentered via the web interfaces (OpenAI WebUI for GPT4o on January 14, 2025; Anthropic WebUI for Claude Sonnet-3.7 on March 17, 2025), each in a new chat session within a single browser (Google Chrome) to minimize potential bias from prior interactions. Questions that were again answered incorrectly during this process were subsequently analyzed qualitatively by a panel of 3 reviewers, 2 medical doctors (WA and FD) and 1 final-year medical student (LMS).

As the quantitative analysis only extracted answers and not explanations, the prompts were revised to explicitly require the LLMs to justify their answer choices, with the goal of gaining deeper insights into the reasons for the observed errors. The questions were entered in German, once with the matched-language prompt in German and once with the English

XSL•FO
RenderX

prompt. To better mirror realistic conditions, the model was not given a system prompt. This study aimed to reflect how medical students ordinarily interact with such tools, entering only the multiple-choice question itself with a brief prompt rather than using a form of prompt engineering (Textbox 2).

Textbox 2. Modified prompts for qualitative analysis in English and German.

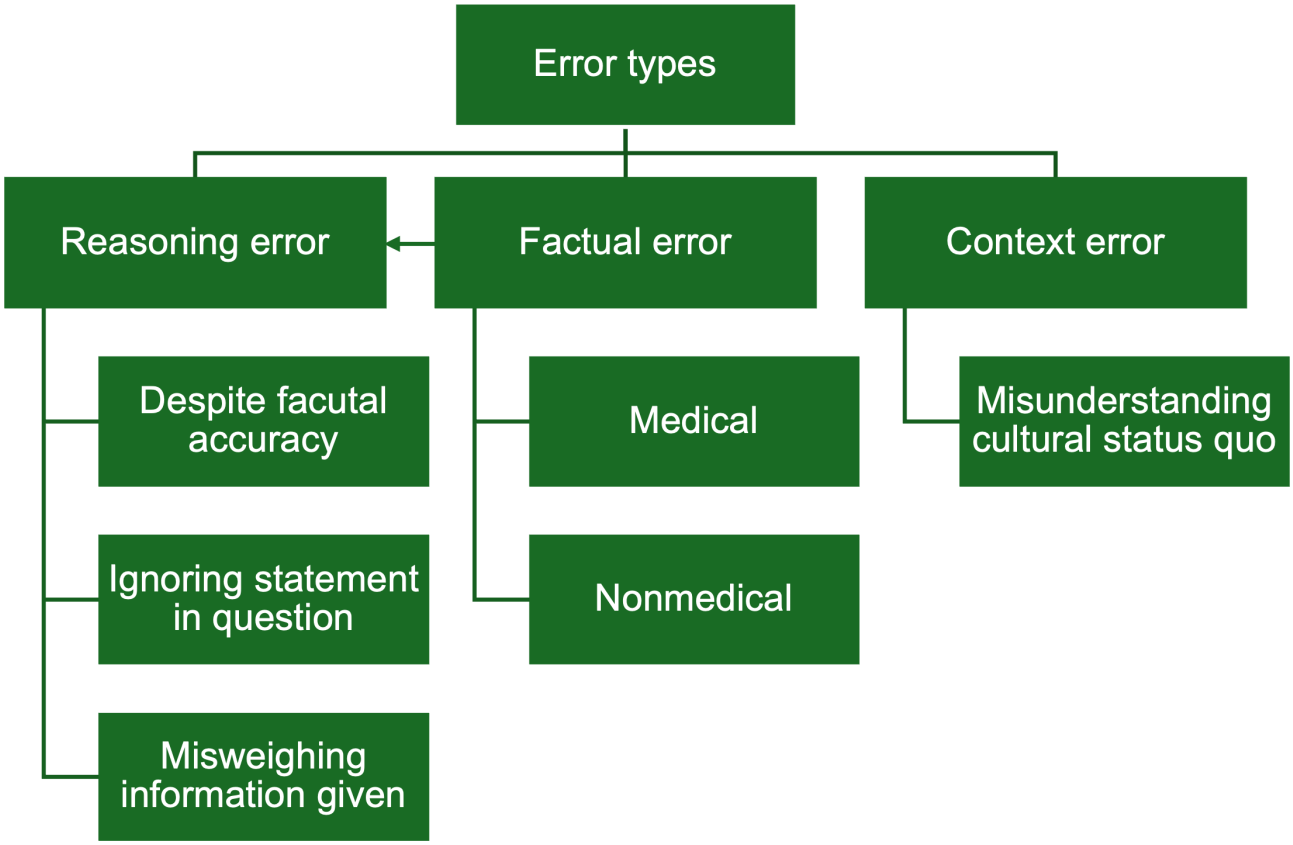
- English prompt: "*Please answer the following multiple-choice question by selecting the correct response option or options. Question: {}. Response options: A - {}, B - {}, C - {}, D - {}, E - {}. Return only the letter or letters (A, B, C, D, E) corresponding to the correct answer or answers and give an explanation to the solution.*"
- German prompt: "*Bitte beantworte die folgende Multiple-Choice-Frage, indem du die richtige/n Antwortoption/en auswählst. Frage: {}. Antwortoptionen: A - {}, B - {}, C - {}, D - {}, E - {}. Gib den/die Buchstaben (A, B, C, D, E) an, der/die der/den richtigen Antwort/en entspricht/entsprechen und gib eine Erklärung zur Lösung.*"

The qualitative analysis was conducted in German or English, depending on the output language of the respective model. The study group consisted of 3 native German speakers and 1 native English speaker. Their respective second and third languages are English, German, and French. The reviewers evaluated and coded all questions with 4 sets of LLM-generated explanations (2 models and 2 prompt languages). For this publication, all presented questions and LLM explanations were translated from German into English with an online translation tool (DeepL; DeepL SE) and professionally proofread.

The qualitative evaluation of the LLMs was conducted using the Framework Method, as described by Gale et al [26]. In recent years, various frameworks have been developed to examine the reasoning processes of physicians and LLMs when responding to medical questions. Our framework was developed with reference to existing studies that have used similar methodologies [27-30]. While other frameworks may differ in structure or focus, our approach was specifically designed to identify the underlying causes of incorrect answers within the explanations given by the respective LLM. The framework was developed through an iterative review process in which multiple reviewers individually examined the LLM explanations and derived preliminary categories and subcategories. These were then discussed collectively, refined through consensus, and cross-validated in subsequent rounds of reanalysis of the LLM explanations. It enabled the investigation of both the basis of the errors and patterns leading to inaccuracies. The framework designed during this process with the distinct error typology is shown in Figure 1. Three main error domains were categorized: reasoning, factual, and context. Reasoning errors involve logical missteps such as ignoring information within the question, reaching incorrect conclusions despite factual accuracy of underlying context information, or misweighing the given information, when complex information is given but the LLM fails to identify the relevant aspects. Factual errors are split into medical and nonmedical, depending on the domain of the incorrect fact. Context errors stem from misunderstandings of the cultural or social background in the context given, such as misinterpreting the cultural status quo or processes in inpatient and outpatient settings or public health matters in Switzerland. This classification aids in systematically identifying and addressing different sources of error. Examples of error types are shown in Textbox 3.

XSL•FO
RenderX

Figure 1. Error typology framework categorizes errors into 3 main types and their subtypes: reasoning errors (despite factual accuracy, ignoring statement in question, and misweighing information given), factual errors (medical and nonmedical), and context errors (misunderstanding cultural status quo).



XSL•FO
RenderX

Textbox 3. Selection of multiple-choice questions and corresponding responses generated by large language models in the English prompt scenario. Each example shows the original question translated into English, the answer options, and the ground-truth label (in brackets) and the qualitative analysis derived from the reviewers' assessments. For each question, the explanations produced by the large language models are presented (some shortened for readability).

Example question 1 reflecting "Reasoning error: ignoring statement in question":

"*A Rh-negative, 34-year-old woman in the 16th week of pregnancy is diagnosed with an anti-D titre of 1:128 (positive from 1:16). In her previous first marriage, after two spontaneous abortions, she gave birth to a child at term who required two exchange transfusions due to hyperbilirubinemia. Which measure is indicated? (A) Termination of pregnancy, (B) Doppler of the fetal cerebral artery, (C) Clarification of the Rh blood group status of the father of the child, (D) Administration of anti-D, (E) n/a*" *(Ground Truth: C)*

Explanation of large language model (GPT4o) in English prompt scenario:

"*Answer: B, Explanation: An anti-D titre of 1:128 in a Rh-negative pregnant woman indicates significant sensitisation to the Rh-D antigen. (...) The best measure to assess the severity of possible fetal anaemiaanemia is Doppler ultrasound of the fetal arteria cerebri media (A.cerebri media). (...) Why the other options are not correct:*"

"*A: This is not justified at this stage and with this diagnosis. C: This information is no longer relevant in this case, as sensitisation is already present. D: Anti-D prophylaxis is ineffective in the case of existing sensitisation. It is only used to prevent initial sensitisation. Conclusion: The correct answer is B: Doppler of the fetal cerebral artery to monitor for fetal anaemia.*"

Qualitative analysis of error type made:

Reasoning error ignoring the statement in question, that the previous child was from another relationship, therefore indicating that this child has another biological father and could be Rh-negative.

Example question 2 reflecting "Reasoning error: despite factual accuracy":

"*A 12-year-old boy cuts the palmar side of the proximal phalanx of his index finger while cutting a branch with a pocketknife. The surgeon consulted found that the sensitivity of the fingertip was intact. Although active flexion of the proximal phalanx is possible, the patient is unable to flex either the seconds or third phalanx. This is a severing of the (A) tendons of the interosseous muscles, (B) tendons of the lumbrical muscles, (C) superficial flexor tendon, (D) deep flexor tendon, (E) superficial and deep flexor tendon*" *(Ground Truth: E)*

Explanation of large language model (Claude-Sonnet-3.7) in English prompt scenario:

"Answer D, The description in the case indicates a severing of the deep flexor tendon (flexor digitorum profundus). Analysis of the clinical findings: Cut on the palmar side of the proximal phalanx of the index finger, intact sensitivity of the fingertip (nerve function preserved), active flexion of the proximal phalanx possible, inability to flex the seconds (middle) and third (end) phalanx. It is important to know the anatomy of the finger flexors: The superficial flexor tendon (...) attaches to the middle phalanx and enables flexion of the middle joint (PIP - proximal interphalangeal joint), the deep flexor tendon (flexor digitorum profundus) attaches to the distal phalanx and enables flexion of the distal interphalangeal joint (DIP).

In the event of an injury at the level of the proximal phalanx, the deep flexor tendon can be severed while the superficial flexor tendon may still be intact. Since the patient cannot flex either the seconds or third phalanx, the function of both the superficial and deep flexor tendon must be impaired. However, the description indicates that active flexion is still possible in the metacarpophalangeal joint (MCP), which is mediated by the lumbrical and interosseous muscles.

Answer D (deep flexor tendon) is correct, as an isolated transection of the deep flexor tendon would not fully explain the functional deficit described."

Qualitative analysis of error type made:

The conclusion does not follow logic and is contradictory, as the whole explanation favors the correct answer E. The large language model states the correct solution during explanation of the anatomy of the muscles that perform flexion in the fingers but fails to give the correct ground truth.

Example question 3 reflecting "Reasoning error: due to misweighing of information":

"A 20-year-old, previously healthy man falls ill with fever a few weeks after travelling to the tropics. His general condition is reduced; clinically there are palpable lymph nodes at several sites and splenomegaly. There is pancytopenia with haemoglobin 100 g/l (norm 140 - 170 g/L), leucocytes 2.0 G/l (norm 4.5 - 11 G/l) (granulocytes approx. 75 %, lymphocytes approx. 10 %), and thrombocytes of 65 G/l (norm 150-350 G/l). What is the most likely diagnosis?

(A) Acute leukaemia, (B) side effect of mefloquine, (C) acute HIV infection, (D) malignant indolent B-cell non-Hodgkin's lymphoma with bone marrow involvement, (E) Dengue fever" (Ground Truth: C)

Explanation of large language model (GPT4o) in English prompt scenario:

"Answer: E, Explanation: The clinical picture of the 20-year-old patient with fever, reduced general condition, palpable lymph nodes, splenomegaly and pancytopenia (...) after a trip to the tropics is typical of an infection with the dengue virus. (...) Why the other options do not fit: A (acute leukaemia): Acute leukaemia could cause pancytopenia but does not explain the travel history and typical course after tropical travel B (side effect of mefloquine): Mefloquine (malaria prophylaxis) is more likely to cause neuropsychiatric side effects, but not pancytopenia and splenomegaly. C (acute HIV infection): Acute HIV infection can cause flu-like symptoms, but not typically pancytopenia with splenomegaly to this extent. D (malignant indolent B-cell non-Hodgkin's lymphoma): This disease usually progresses insidiously and is unlikely in a young, previously healthy man with no other risk factors. Conclusion: The clinical presentation is best suited to dengue fever (E)."

Qualitative analysis:

The "trip to the tropics" and its association to dengue seems to be overemphasized even though clinical symptoms and the dynamic as well as the onset of the disease are not specific to dengue fever.

XSL•FO
RenderX

Example question 4 reflecting "Medical Factual error":

"A 6-year-old child has hematomas on various parts of the body. Which location/s is/are suspicious of physical abuse? (A) on the shins, (B) on the back of the hand, (C) on the back, (D) behind the ears, (E) n/a" (Ground Truth: B/C/D)

Explanation of large language model (GPT4o) in English prompt scenario:

"Answer: C, D, Explanation: Hematomas in children are not uncommon, especially on parts of the body that are frequently injured during normal play or falls. However, there are certain areas of the body where bruising (haematomas) are particularly suspicious for physical abuse. Suspicious localisations:

C: Hematomas on the back are unusual as this area is rarely injured during falls or everyday activities. Injuries in this area are often associated with violence.

D: Bruising behind the ears is also highly suspicious as this area is not usually affected by accidents. Such injuries may indicate blows or pulling to the head.

Why the other options are less suspicious: A: Hematomas on the shins are normal in active children and can often be explained by playing or falls. B: Hematomas on the back of the hand can also be caused by everyday activities and are less obvious for maltreatment. E: (….) Conclusion: The hematomas on the back (C) and behind the ears (D) are suspicious for physical abuse in this case."

Qualitative analysis:

Bruises on the back of the hand are considered suspicious for physical abuse in a 6-year-old child.

Example question 5 reflecting "Context error: misunderstanding cultural status quo":

"Young parents call your practice in an emergency at 5 pm in the evening. Their 2-year-old son has suddenly lost consciousness while playing, has tilted backward from a sitting position and has his eyes fixed upwards. He then began to twitch rhythmically with all 4 extremities. The parents poured cold water on his face, after which he slowly regained consciousness a few minutes later. The whole episode lasted 3 minutes. Now he is reduced, tired and seems to have a high temperature. The parents have never experienced anything like this with their son. What is the most appropriate next step in management? (A) The boy must be taken to the nearest emergency centre immediately, (B) You ask the parents to take the boy to your consultation the next day for a check-up, (C) You refer the boy to a neurologist for an EEG scan, (D) You ask the parents to come to your consultation with the boy on the same day to check for a fever focus, (E) No further action necessary" (Ground truth: D)

Explanation of large language model (Claude-Sonnet-3.7) in English prompt scenario:

"Answer: A, The description indicates a febrile convulsion in a 2-year-old child with the following features: Sudden loss of consciousness followed by rhythmic twitching of all extremities, staring of the eyes upward, short duration (3 minutes), high fever, first episode (parents have never experienced anything similar), postictal state (reduced, tired). Although it is probably a simple febrile convulsion, which is not uncommon in children between 6 months and 5 years of age, the child in this situation should be taken to an emergency centre immediately for the following reasons:1. it is the first episode of a seizure, which requires a thorough work-up to rule out other causes such as meningitis, encephalitis or other neurological conditions, 2. the child is in a reduced state following the seizure, which requires immediate assessment, 3. the level of fever and its cause must be clarified, 4. the young age of the child (2 years) increases the risk of more serious causes.

The other options are not appropriate:B: too long a delay in an acute condition. C: not urgent enough EEG is not the first measure. D: important but the situation requires a more comprehensive emergency assessment. E: clearly inappropriate for a first seizure."

Qualitative analysis:

Context error or subject for discussion: In Switzerland, there is the possibility for urgent outpatient consultation at the family doctor to avoid a consultation at the emergency center in comparable situations. This question cannot be answered internationally due to different availabilities of same-day presentations for children. Further, this question might be challenging for licensing examination candidates, whether an outpatient consultation on the same day (option D) or an immediate emergency consultation (option A) is indicated and could also be discussed in terms of quality assurance.

It is important to note that the identification of reasoning errors does not imply that LLMs do or could possess human-like reasoning capabilities. Rather, the framework provides a structured means to analyze the apparent logic, or lack thereof, within the LLMs' responses, regardless of the process that generated the response.

Ethical Considerations

The study is out of scope for national ethical approval as it does not collect any data from humans and is based on publicly accessible datasets and models.

Results

Model Accuracy Performance at Answering Questions

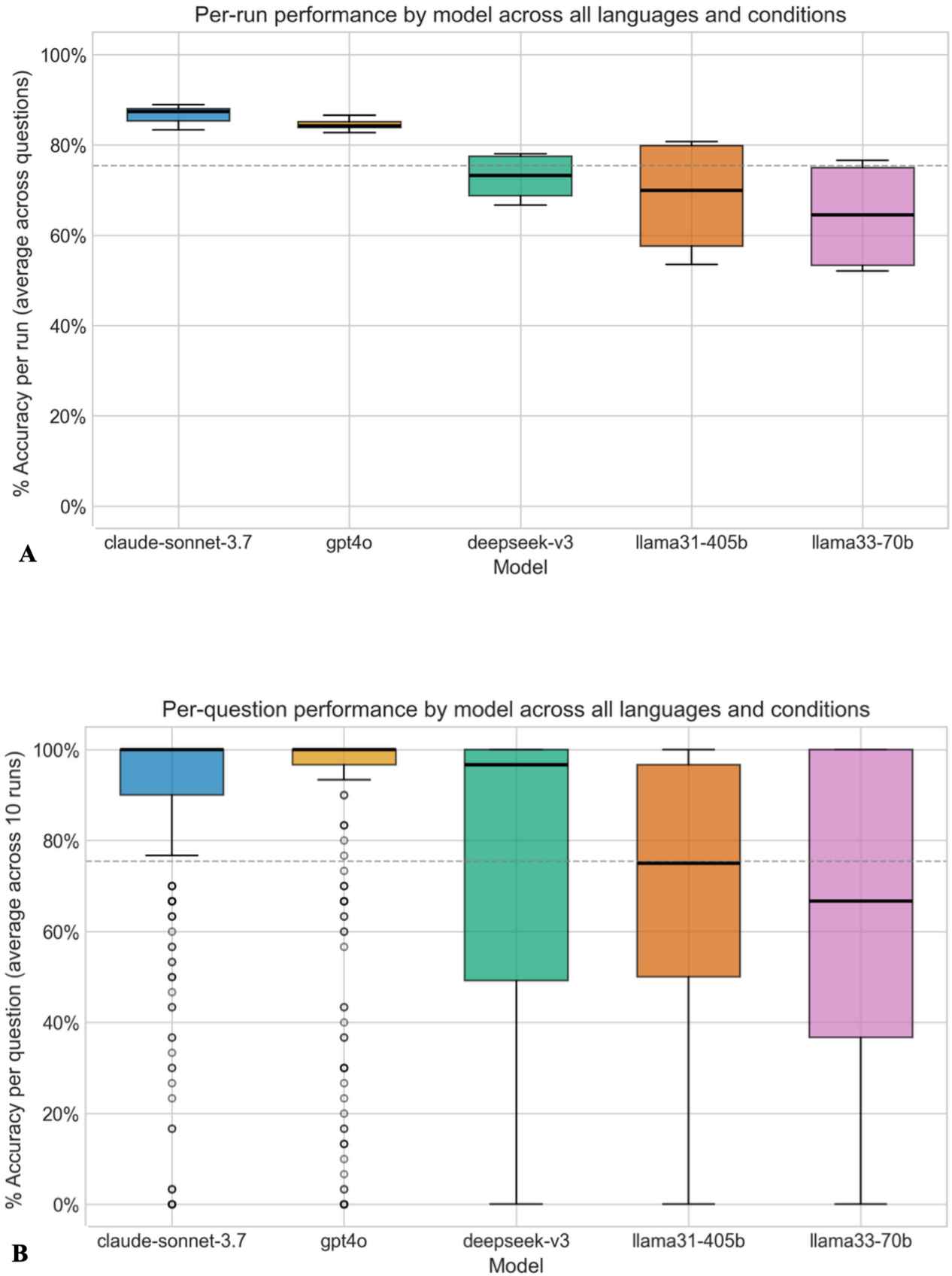
Figure 2 presents the model accuracy results, as averages across the 10 runs per question, across 5 LLMs, for all languages. Claude-Sonnet-3.7 demonstrated the highest mean accuracy (87%, SD 0.34%) and lowest variability, indicating consistent performance. GPT4o also showed strong performance (mean accuracy 84%, SD 0.36%) with slightly more variability. In contrast, Llama3.3-70b, Llama3.1-405b, and DeepSeek V3 exhibited significantly less reliable performance across tasks (mean accuracies of 64%, 69%, and 73%; SDs of 0.48%, 0.46%, and 0.44%; $P<.001$ in all 3 cases). However, all models would be considered to have achieved a passing grade, which can vary in specific threshold but is usually set at around 60% accuracy in written medical licensing examinations.

Figure 2A shows per-run accuracy of different LLMs across all questions and conditions. Accuracy is shown as the percentage of accurate responses per run across questions. Figure 2B shows per-question accuracy of different LLMs across all runs and conditions. Accuracy is shown as % accuracy per question

XSL•FO
RenderX

across runs and conditions. A gray dotted line indicates the average across all models. Individual responses per run and table data for the figure are given in Multimedia Appendix 1.

Figure 2. Accuracy of different large language models.



XSL•FO
RenderX

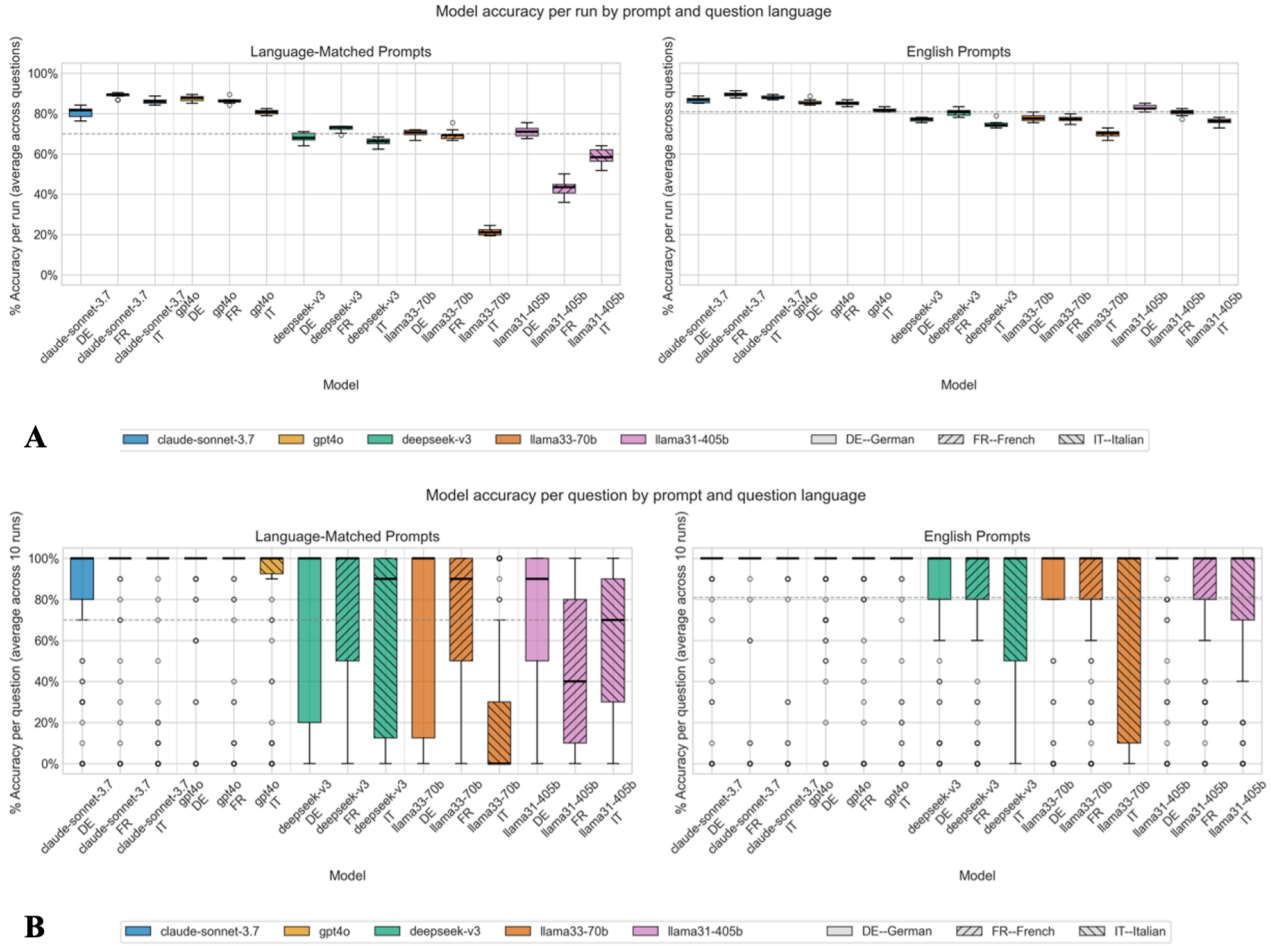
Performance Variation by Question Language and Prompt Language

Figure 3 illustrates model accuracy across the 3 question languages (German, French, and Italian) and 2 instruction prompt scenarios. As with the overall model comparison, Claude-Sonnet-3.7 consistently achieved the highest accuracy across all languages, with the strongest performance in French and Italian. The lowest performance was achieved in Italian by all other models. Across all models, comparing model performance across the 2 prompt scenarios, English prompts generally led to better or comparable performance relative to language-matched prompts (OR 1.81, 95% CI 1.60-2.05; *P*<.001). Notably, Claude-Sonnet-3.7 and GPT4o showed the highest and most consistent accuracy across all languages and prompt types. Accuracy values were slightly lower for language-matched prompts with Claude-Sonnet-3.7, particularly for German. In contrast, both Llama models exhibited the poorest performance, especially in the language-matched prompt condition for French and Italian.

Figure 3A shows model accuracy per run across all questions, separated by prompt and question language. Accuracy is shown as the percentage of accurate responses per run across questions and conditions. Figure 3B shows model accuracy per question across all runs, separated by prompt and question language (DE=German, FR=French, IT=Italian). Accuracy is shown as % accuracy per question across runs and conditions. The dotted gray line indicates the average across all scenarios. Individual responses per run and table data for the figure are given in Multimedia Appendix 1.

When comparing model performance across the 3 languages, accuracy was highest on average for German-language questions. Therefore, German served as the reference category for the question language statistical comparison. Compared to German, accuracy was not significantly different for French (OR 0.92, 95% CI 0.79-1.08; *P*=.29) but was significantly lower for Italian (OR 0.64, 95% CI 0.54-0.75; *P*<.001).

Figure 3. Overall model accuracy across runs and questions by prompt and question language.



Performance Variation by Question Attributes

Regarding the response format, multiple-choice questions with single responses were solved significantly better than those with multiple responses across models (mean accuracy 77%, SD 42% vs mean accuracy 51%, SD 0.50%; OR 0.31; 95% CI 0.12-0.78; *P*<.001). Average accuracies per language, model, and response format (single vs multiple) are illustrated in Figure S1 in Multimedia Appendix 2.

During the quantitative analysis of medical specialties, there were no significant differences in accuracy across clinical question categories (all comparisons not significant after correction for multiple tests). Average accuracies per language,

XSL•FO
RenderX

model, and medical specialty are illustrated in Figure S2 in Multimedia Appendix 2.

When looking at the purpose of the question with diagnosis, treatment, and other-related questions as categories, diagnosis questions were used as the reference category. Compared to diagnosis, treatment questions were solved significantly worse (mean accuracy 81%, SD 0.39% vs mean accuracy 70%, SD 0.45%; OR 0.55, 95% CI 0.34-0.91; $P<.001$) while the other-related question also appeared to perform worse than diagnosis, but this difference was not significant (mean accuracy 65%, SD 47% but P not significant after correction for multiple tests). Average accuracies per language, model, and purpose category are illustrated in Figure S3 in Multimedia Appendix 2.

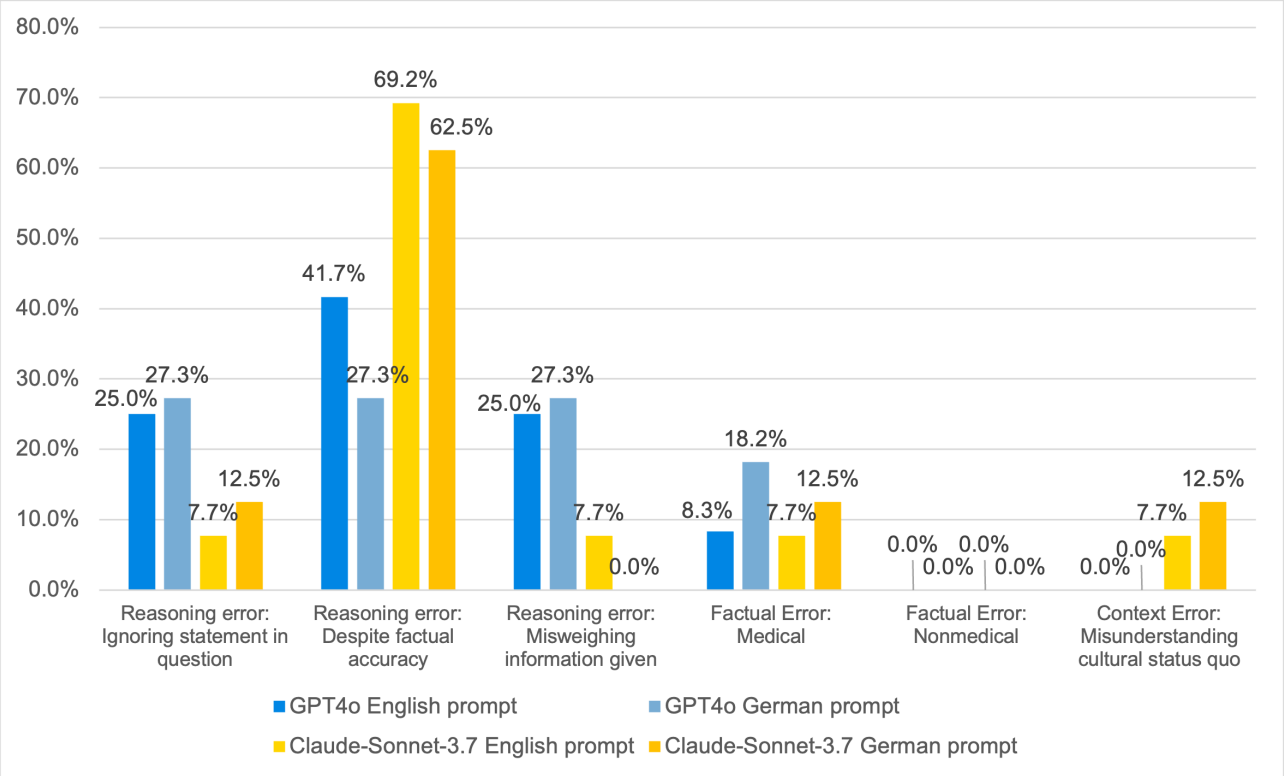
There were also no significant differences between questions categorized for urgency, related to acute versus chronic conditions. Average accuracies per language, model, and urgency category are illustrated in Figure S4 in Multimedia Appendix 2.

Qualitative Analysis

In the qualitative analysis, 12 questions were still solved incorrectly on the web interface with GPT4o and 13 with Claude-Sonnet-3.7, corresponding to those that produced incorrect answers compared to the ground truth in the aggregate quantitative evaluation. Six questions were answered incorrectly by both models. In total, incorrect answers were observed in 20 distinct multiple-choice questions. GPT4o provided incorrect answers to 11 questions when prompted in both German and English, with one additional error occurring exclusively with the English prompt. Similarly, Claude-Sonnet-3.7 gave the same incorrect answers to 8 questions across both languages, while the English prompt resulted in 5 further incorrect responses.

Figure 4 shows comparative error analysis and reveals the markedly higher frequency of reasoning errors despite factual accuracy exhibited by Claude-Sonnet-3.7 relative to GPT4o. This error type refers to instances where the model provides information that is factually correct but draws conclusions or inferences that do not appropriately address the questions' requirements. In the English prompt condition, Claude-Sonnet-3.7 produced this type of error in 69% of the erroneous cases (9/13), compared to 42% (5/12) for GPT4o. The pattern persisted in the German prompt condition, with Claude-Sonnet-3.7 showing this error in 62% of the erroneous cases (5/8), while GPT4o's rate was lower at 27% (3/11).

Figure 4. Error analysis of large language models (GPT4o and Claude-Sonnet-3.7) according to the framework with language-matched prompts in German and English language prompt.



Textbox 3 presents selected multiple-choice questions, their corresponding LLM-generated answers (from GPT4o or Claude-Sonnet-3.7), and an explanation of the error type and reasoning issues identified. The table focuses on the types and subtypes made by LLMs when responding to the clinical scenarios and includes ground truth answers to the multiple-choice question for comparison.

Additionally, during qualitative analysis, 3 questions' ground truths were found to be ambiguous given the current state of medical knowledge. One example question is presented in Textbox 4. This example shows the potential risk for LLMs to hallucinate and assume knowledge if asked for an answer, in cases when facts are ambiguous. The model formulated arguments for the answer (C) as ground truth even if the given

XSL•FO
RenderX

explanation does not correspond to evidence based on medical knowledge.

Textbox 4. Example of an ambiguous question and explanation provided by GPT4o.

Example question translated into English:"*A 4-year-old boy with moderate atopic dermatitis is regularly vaccinated according to the Swiss vaccination schedule. Which additional vaccination does he need?*

(A) Hepatitis A, (B) Hepatitis B, (C) Influenza, (D) Rotavirus, (E) Varicella" (Ground truth: E)

Explanation given by large language model (GPT4o) in English prompt scenario:

"*Answer: C , Explanation: A 4-year-old boy with moderate atopic dermatitis has an increased risk of infection due to the impaired skin barrier and possible inflammatory processes. The influenza vaccination (C) is particularly important as children with atopic diseases are more susceptible to more severe courses of influenza infections. The vaccination not only protects against the disease itself but also reduces the risk of secondary bacterial skin infections, which can occur more frequently with atopic skin. Why the other options are less useful:*

A: This vaccination is not routinely recommended in Switzerland, but only for certain risk groups or before travellingto endemic areas.

B: The hepatitis B vaccination is part of the Swiss vaccination schedule and should already have been carried out. An additional vaccination is therefore not necessary.

D: The rotavirus vaccination is recommended in infancy (up to 6 months). This vaccination is no longer indicated at the age of 4 years.

E: The varicella vaccination is included in the Swiss vaccination schedule and should have already been given in infancy. An additional vaccination is not necessary if the vaccination schedule has been adhered to.

Conclusion: The influenza vaccination (C) is particularly important for a 4-year-old boy with atopic dermatitis, as it protects against possible severe courses and complications."

Qualitative analysis:

The large language model correctly states that "*The varicella vaccination is included in the Swiss vaccination schedule and should have already been given in infancy.*" It is a correct statement that depicts the disagreement with the ground truth (E) and reflects that the question was developed based on the old Swiss vaccination schedule (prior to 2023). This highlights that the question can no longer be answered as none of the answers would be part of a recommendation. However, the large language model selected answer (C) influenza and provided a rationale which combines the true fact "A 4-year-old boy with moderate atopic dermatitis has an increased risk of infection due to the impaired skin barrier and possible inflammatory processes" with the wrong conclusion that "*The vaccination (….) also reduces the risk of secondary bacterial skin infections*" giving a further example for a reasoning error.

Discussion

Principal Findings

This study evaluated the performance of different LLMs in answering multilingual medical licensing examination questions. All models achieved passing-level accuracy, yet performance varied considerably by model, language, and question type. Overall, LLMs performed best on German questions. Prompt language played a key role: while prompting in English generally led to higher accuracy, advanced LLMs such as GPT4o and Claude-Sonnet-3.7 achieved comparable results with language-matched prompts in German and French. Qualitative analysis of answer explanations revealed that both analyzed LLMs (GPT4o and Claude-Sonnet-3.7) exhibited different reasoning error types helping to interpret LLM's output. The analysis of answer explanations further led to the identification of questions whose content was ambiguous, offering a valuable tool for ensuring question quality.

Our findings are in line with previous findings that LLMs generally exhibit high performance across a variety of medical tasks and languages, reaffirming their potential for supporting multilingual medical education and assessment tasks. However, notable variation exists among models, with some consistently outperforming others in both accuracy and consistency [31,32]. Beyond openly accessible LLMs, state-of-the-art medical models are emerging that demonstrate expert-level performance in real-world clinical tasks and can be deployed locally. However, despite their strong potential, these models are currently better suited for specific applications and ongoing evaluation rather than general clinical use [33]. This highlights the critical importance of model selection and reflection of information presented, as the choice of LLM can substantially influence the quality and reliability of the outputs [34]. Furthermore, the landscape of LLMs is rapidly evolving, with frequent updates and architectural changes leading to substantial shifts in model behavior over time [17]. This dynamic nature poses a challenge for reproducibility and comparability across studies, as the same model may refer to different versions with varying performance profiles [35,36].

Our study showed that prompt language strongly influenced model performance, in line with previous findings [37]. Overall, LLMs tend to be better at following English-language instruction prompts, which we also observed in our study. This highlights persistent limitations in multilingual generalization of LLMs, consistent with LLMs' predominant training in English, limiting non-English prompt processing and pointing to a need for more robust multilingual training [38]. However, our results show that larger and more advanced models (GPT4o and Claude-Sonnet-3.7) demonstrate strong multilingual question-answering performance. Our analysis indicates that prompt language had a stronger influence on performance than input language, suggesting that the prompt design and model alignment with prompt intent are critical for accuracy and an important skill for medical professionals when using LLMs [39].

XSL•FO
RenderX

The observed performance differences between the smaller models and the large proprietary LLMs have important implications for practical deployment in medical and other data-sensitive environments. In our study, the smaller open-source models consistently produced lower-quality outputs compared with the current closed-source models (eg, GPT4o). Nevertheless, their performance was substantially better when prompted in English, underscoring the need to carefully consider the prompt and context language when such models are used in practice. While locally deployable models offer greater control over data security, they require careful handling of prompt design and a clear understanding of their performance limitations. Thus, our findings highlight the possible trade-off between model performance and data governance requirements [40]. The findings of this study may inform the use of LLMs in other multilingual medical education settings, though factors such as regional terminology, country-specific clinical guidelines, and institution-specific structures may lead to variation.

Moreover, models performed better in diagnostic tasks than in treatment-related questions [41,42], consistent with prior findings [43]. This may be related to a training bias, where diagnostic reasoning is more prevalent in available training data, a pattern that warrants further investigation. The impact of the response format (single vs multiple response) on the accuracy of the answer could be explained by the fact that the multiple-choice questions used to train the models probably consist mainly of single-choice questions as it is the predominant question type in medical education [44]. On the other hand, the exact formulation of the prompts for this question type could also have an impact on the output.

Interestingly, we observed that reasoning errors occurred more frequently when models were prompted in English compared to language-matched prompts, although for the small sample size in our qualitative study, it was not possible to perform statistical tests for significance. A possible explanation could be that, for questions originally written in German, models receiving an English prompt may have internally translated the question into English before generating a response. This intermediate translation step could have introduced subtle semantic inaccuracies, leading to flawed reasoning despite overall factual correctness. Similar translation-related effects have been reported in previous studies, where internal translation processes in multilingual LLMs were shown to increase the likelihood of inconsistencies when prompt and question languages did not align [45]. At the same time, our quantitative results showed that model accuracy per run was comparable in the language-matched scenario in GPT4o and Claude-Sonnet-3.7, suggesting that these advanced LLMs already achieve strong performance in non-English contexts, including German and French. This observation possibly reflects advances in multilingual training data diversity and model alignment, supporting the notion that recent LLMs are becoming more proficient in multilingual reasoning.

As identified in the qualitative analysis, the error type "reasoning error despite factual accuracy" refers to instances where the model presents factually accurate information but draws incorrect or irrelevant conclusions that fail to adequately address the question. These errors are particularly concerning, as their surface-level correctness can obscure deeper reasoning flaws and mislead users [46,47]. In high-stakes domains such as medical decision support or medical education, where correct reasoning is critical, these errors could lead to incorrect conclusions or misleading explanations despite an appearance of accuracy [48]. This finding underscores a key dimension of model reliability that extends beyond factual correctness, highlighting the importance of evaluating reasoning fidelity in addition to factual accuracy when assessing LLM performance [28]. The reasoning error types identified as "ignoring statement in question" and "misweighing information" further demonstrate that although LLMs can process substantial contextual information, they remain vulnerable to subtle pitfalls that may only be apparent to individuals with deep domain expertise. These errors indicate persistent limitations in the models' ability to appropriately weigh competing pieces of information and to distinguish clinically relevant cues from those that should be considered negligible or misleading. The clinical relevance of these findings extends beyond the experimental setting, as vignette-based questions have the intention to approximate authentic clinical scenarios and to mirror the type of queries that physicians may realistically pose to LLMs in practice. Consequently, the incorrect responses and, in particular, the inaccurate explanations provided by the models are of considerable concern. If such outputs were accepted uncritically, they could mislead clinical decision-making and potentially contribute to diagnostic or therapeutic errors.

The LLMs' explanation during qualitative analysis helped to recognize changes in medical recommendations or inconsistencies in the question, demonstrating the LLMs' potential for examination question validation and quality assurance. Professional oversight of the question itself and LLMs' explanation revealed that the ground truth of the question was based on outdated medical standards. Removal from the examination or thorough revision must be the consequence for such questions. Multiple-choice questions represent an important tool across medical education but are resource-intensive to develop and approve [49]. The implementation of LLMs in question quality evaluation can support flagging ambiguous or poorly formulated questions before they are used in an examination, provided that their use is permitted under local or institutional regulations [50,51]. Future studies should be conducted to further evaluate this potential use case.

Limitations

Our study has several strengths but also limitations, including response variability of LLMs and the fast-evolving nature of commercially available LLMs, which can affect reproducibility across studies. The performance of the used web-based versions could differ from their API counterparts, which might influence the qualitative findings. The open or commercially accessible LLMs are not intended for medical use but are frequently used by medical students in the respective learning processes due to their seemingly accurate performance [52]. As this study excluded questions with images or videos, the comparison of whole examinations with image-based questions is limited. The qualitative framework is proposed as an exploratory basis for structuring and interpreting the observed outputs and for

XSL•FO
RenderX

assessing their potential applicability in other analyses. However, its reliance on human evaluators constitutes a relevant limitation and necessitates further research and validation. Finally, the dataset was derived from a single public Swiss question bank, which may limit the representativeness of topics and linguistic styles. Future studies should include additional data in French, German and Italian to enhance diversity and reduce potential bias.

Conclusion

In summary, our evaluation confirms that LLMs can handle multilingual medical examination content with high accuracy. However, meaningful differences in performance across models, languages, and clinical tasks underscore the importance of careful model selection, prompt design, and language. Although the overall results favor English prompts, the specific findings from GPT4o and Claude-Sonnet-3.7 indicate that advanced models may perform comparably with language-matched prompts, highlighting the need for further exploration of this effect. In this study, LLM-based qualitative analysis supported the understanding of errors made by LLMs and the identification of ambiguous questions, highlighting its potential for quality assurance in medical assessments. Human oversight of language-sensitive or clinically nuanced content remains essential to determine whether incorrect answers stem from flaws in the question itself or from errors generated by the LLM. Continued evaluation and transparent reporting will be important as models rapidly evolve and become further integrated into educational and assessment settings.

Acknowledgments

ChatGPT (OpenAI, GPT4o) and machine translator (DeepL) were used to support the translation of questions during the publication of this work. All translations were reviewed and, where necessary, revised by the authors to ensure clarity and correctness.

Funding

This study was supported by the School of Medicine at the University of St Gallen (HSG) in Switzerland.

Data Availability

The data generated during this study can be accessed on our lab's GitHub repository [25].

Authors' Contributions

LMS contributed to data collection, conducted qualitative data analysis, and drafted, reviewed, and edited the original manuscript. WA contributed to the qualitative data analysis, assisted in theme development, and reviewed and edited the manuscript. FD contributed to the qualitative data analysis and reviewed and edited the manuscript. JH supervised the study, contributed to the study design, conducted quantitative data analysis, assisted in theme development, and reviewed and edited the manuscript. All authors read, edited, and approved the final manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Raw data to figures shown in the manuscript.
[XLSX File, 673 KB - mededu_v12i1e81399_app1.xlsx]

Multimedia Appendix 2
Supplementary data with additional figures referenced in the manuscript.
[DOCX File, 2868 KB - mededu_v12i1e81399_app2.docx]

References

1. Davenport T, Kalakota R. The potential for artificial intelligence in healthcare. Future Healthc J 2019 Jun;6(2):94-98. [doi: 10.7861/futurehosp.6-2-94] [Medline: 31363513]
2. Singhal K, Tu T, Gottweis J, et al. Towards expert-level medical question answering with large language models. . Preprint posted online on 2023. [doi: 10.48550/arXiv.2305.09617]
3. Lotto C, Sheppard SC, Anschuetz W, et al. ChatGPT generated otorhinolaryngology multiple-choice questions: quality, psychometric properties, and suitability for assessments. OTO Open 2024;8(3):e70018. [doi: 10.1002/oto2.70018] [Medline: 39328276]
4. Chen PHC, Liu Y, Peng L. How to develop machine learning models for healthcare. Nat Mater 2019 May;18(5):410-414. [doi: 10.1038/s41563-019-0345-0] [Medline: 31000806]
5. Alowais SA, Alghamdi SS, Alsuhebany N, et al. Revolutionizing healthcare: the role of artificial intelligence in clinical practice. BMC Med Educ 2023 Sep 22;23(1):689. [doi: 10.1186/s12909-023-04698-z] [Medline: 37740191]

XSL•FO
RenderX

6. Goh E, Gallo R, Hom J, et al. Large language model influence on diagnostic reasoning: a randomized clinical trial. JAMA Netw Open 2024 Oct 1;7(10):e2440969. [doi: 10.1001/jamanetworkopen.2024.40969] [Medline: 39466245]
7. Yang LWY, Ng WY, Lei X, et al. Development and testing of a multi-lingual Natural Language Processing-based deep learning system in 10 languages for COVID-19 pandemic crisis: a multi-center study. Front Public Health 2023;11:1063466. [doi: 10.3389/fpubh.2023.1063466] [Medline: 36860378]
8. Omar M, Sorin V, Agbareia R, et al. Evaluating and addressing demographic disparities in medical large language models: a systematic review. Int J Equity Health 2025 Feb 26;24(1):57. [doi: 10.1186/s12939-025-02419-0] [Medline: 40011901]
9. Zong H, Wu R, Cha J, et al. Large language models in worldwide medical exams: platform development and comprehensive analysis. J Med Internet Res 2024 Dec 27;26(1):e66114. [doi: 10.2196/66114] [Medline: 39729356]
10. Tseng LW, Lu YC, Tseng LC, Chen YC, Chen HY. Performance of ChatGPT-4 on Taiwanese Traditional Chinese medicine licensing examinations: cross-sectional study. JMIR Med Educ 2025 Mar 19;11(1):e58897. [doi: 10.2196/58897] [Medline: 40106227]
11. Gilson A, Safranek CW, Huang T, et al. How does ChatGPT perform on the United States Medical Licensing Examination (USMLE)? The implications of large language models for medical education and knowledge assessment. JMIR Med Educ 2023 Feb 8;9(1):e45312. [doi: 10.2196/45312] [Medline: 36753318]
12. Ishida K, Hanada E. Potential of ChatGPT to pass the Japanese medical and healthcare professional national licenses: a literature review. Cureus 2024 Aug;16(8):e66324. [doi: 10.7759/cureus.66324] [Medline: 39247019]
13. Kung TH, Cheatham M, Medenilla A, et al. Performance of ChatGPT on USMLE: potential for AI-assisted medical education using large language models. PLOS Digit Health 2023 Feb;2(2):e0000198. [doi: 10.1371/journal.pdig.0000198] [Medline: 36812645]
14. Roos J, Kasapovic A, Jansen T, Kaczmarczyk R. Artificial intelligence in medical education: comparative analysis of ChatGPT, Bing, and medical students in Germany. JMIR Med Educ 2023 Sep 4;9(1):e46482. [doi: 10.2196/46482] [Medline: 37665620]
15. Liu M, Okuhara T, Chang X, et al. Performance of ChatGPT across different versions in medical licensing examinations worldwide: systematic review and meta-analysis. J Med Internet Res 2024 Jul 25;26:e60807. [doi: 10.2196/60807] [Medline: 39052324]
16. Guillen-Grima F, Guillen-Aguinaga S, Guillen-Aguinaga L, et al. Evaluating the efficacy of ChatGPT in navigating the Spanish Medical Residency Entrance Examination (MIR): promising horizons for AI in clinical medicine. Clin Pract 2023 Nov 20;13(6):1460-1487. [doi: 10.3390/clinpract13060130] [Medline: 37987431]
17. Wang X, Gong Z, Wang G, et al. ChatGPT performs on the Chinese National Medical Licensing Examination. J Med Syst 2023 Aug 15;47(1):86. [doi: 10.1007/s10916-023-01961-0] [Medline: 37581690]
18. Alfertshofer M, Hoch CC, Funk PF, et al. Sailing the Seven Seas: a multinational comparison of ChatGPT's performance on medical licensing examinations. Ann Biomed Eng 2024 Jun;52(6):1542-1545. [doi: 10.1007/s10439-023-03338-3] [Medline: 37553555]
19. Wang H, Wu W, Dou Z, He L, Yang L. Performance and exploration of ChatGPT in medical examination, records and education in Chinese: pave the way for medical AI. Int J Med Inform 2023 Sep;177:105173. [doi: 10.1016/j.ijmedinf.2023.105173] [Medline: 37549499]
20. Self-assessments. Measured Examinator. URL: https://self-assessment.measured.iml.unibe.ch/ [accessed 2025-06-02]
21. Maruyama H, Toyama Y, Takanami K, Takase K, Kamei T. Role of artificial intelligence in surgical training by assessing GPT-4 and GPT-4o on the Japan Surgical Board Examination with text-only and image-accompanied questions: performance evaluation study. JMIR Med Educ 2025 Jul 30;11(1):e69313. [doi: 10.2196/69313] [Medline: 40737609]
22. Yang Z, Yao Z, Tasmin M, et al. Unveiling GPT-4V's hidden challenges behind high accuracy on USMLE questions: observational study. J Med Internet Res 2025 Feb 7;27(1):e65146. [doi: 10.2196/65146] [Medline: 39919278]
23. Takagi S, Koda M, Watari T. The performance of ChatGPT-4V in interpreting images and tables in the Japanese Medical Licensing Exam. JMIR Med Educ 2024 May 23;10(1):e54283. [doi: 10.2196/54283] [Medline: 38787024]
24. Miyazaki Y, Hata M, Omori H, et al. Performance of ChatGPT-4o on the Japanese Medical Licensing Examination: evalution of accuracy in text-only and image-based questions. JMIR Med Educ 2024 Dec 24;10(1):e63129. [doi: 10.2196/63129] [Medline: 39718557]
25. Understanding LLM medical question answering in a swiss context. GitHUb. 2025. URL: https://github.com/hastingslab-org/swiss-medqa [accessed 2025-12-09]
26. Gale NK, Heath G, Cameron E, Rashid S, Redwood S. Using the framework method for the analysis of qualitative data in multi-disciplinary health research. BMC Med Res Methodol 2013 Sep 18;13(1):117. [doi: 10.1186/1471-2288-13-117] [Medline: 24047204]
27. Nendaz M, Perrier A. Diagnostic errors and flaws in clinical reasoning: mechanisms and prevention in practice. Swiss Med Wkly 2012;142:w13706. [doi: 10.4414/smw.2012.13706] [Medline: 23135902]
28. Maitland A, Fowkes R, Maitland S. Can ChatGPT pass the MRCP (UK) written examinations? Analysis of performance and errors using a clinical decision-reasoning framework. BMJ Open 2024 Mar 15;14(3):e080558. [doi: 10.1136/bmjopen-2023-080558] [Medline: 38490655]

XSL•FO
RenderX

29. Herrmann-Werner A, Festl-Wietek T, Holderried F, et al. Assessing ChatGPT's mastery of Bloom's taxonomy using psychosomatic medicine exam questions: mixed-methods study. J Med Internet Res 2024 Jan 23;26:e52113. [doi: 10.2196/52113] [Medline: 38261378]
30. Kanithi PK, Christophe C, Pimentel MA, et al. Towards a comprehensive framework for evaluating llms in clinical applications. arXiv. Preprint posted online on 2024. [doi: 10.48550/arXiv.2409.07314] [Medline: 37852559]
31. Rosoł M, Gąsior JS, Łaba J, Korzeniewski K, Młyńczak M. Evaluation of the performance of GPT-3.5 and GPT-4 on the Polish Medical Final Examination. Sci Rep 2023 Nov 22;13(1):20512. [doi: 10.1038/s41598-023-46995-z] [Medline: 37993519]
32. Ali R, Tang OY, Connolly ID, et al. Performance of ChatGPT, GPT-4, and Google Bard on a neurosurgery oral boards preparation question bank. Neurosurgery 2023 Nov 1;93(5):1090-1098. [doi: 10.1227/neu.0000000000002551] [Medline: 37306460]
33. Dennstädt F, Schmerder M, Riggenbach E, et al. Comparative evaluation of a medical large language model in answering real-world radiation oncology questions: multicenter observational study. J Med Internet Res 2025 Sep 23;27(1):e69752. [doi: 10.2196/69752] [Medline: 40986858]
34. Lucas HC, Upperman JS, Robinson JR. A systematic review of large language models and their implications in medical education. Med Educ 2024 Nov;58(11):1276-1285. [doi: 10.1111/medu.15402] [Medline: 38639098]
35. Fisch U, Kliem P, Grzonka P, Sutter R. Performance of large language models on advocating the management of meningitis: a comparative qualitative study. BMJ Health Care Inform 2024 Feb 2;31(1):e100978. [doi: 10.1136/bmjhci-2023-100978] [Medline: 38307617]
36. Lee P, Bubeck S, Petro J. Benefits, limits, and risks of GPT-4 as an AI chatbot for medicine. N Engl J Med 2023 Mar 30;388(13):1233-1239. [doi: 10.1056/NEJMsr2214184] [Medline: 36988602]
37. Wang J, Shi E, Yu S, et al. Prompt engineering for healthcare: methodologies and applications. arXiv. Preprint posted online on Mar 23, 2024. [doi: 10.48550/arXiv.2304.14670]
38. Qin L, Chen Q, Zhou Y, et al. A survey of multilingual large language models. Patterns (N Y) 2025 Jan 10;6(1):101118. [doi: 10.1016/j.patter.2024.101118] [Medline: 39896256]
39. Meskó B. Prompt engineering as an important emerging skill for medical professionals: tutorial. J Med Internet Res 2023 Oct 4;25(1):e50638. [doi: 10.2196/50638] [Medline: 37792434]
40. Dennstädt F, Hastings J, Putora PM, Schmerder M, Cihoric N. Implementing large language models in healthcare while balancing control, collaboration, costs and security. NPJ Digit Med 2025 Mar 6;8(1):143. [doi: 10.1038/s41746-025-01476-7] [Medline: 40050366]
41. Eoh KJ, Kwon GY, Lee EJ, et al. Efficacy of large language models and their potential in obstetrics and gynecology education. Obstet Gynecol Sci 2024 Nov;67(6):550-556. [doi: 10.5468/ogs.24211] [Medline: 39355902]
42. Levine DM, Tuwani R, Kompa B, et al. The diagnostic and triage accuracy of the GPT-3 artificial intelligence model: an observational study. Lancet Digit Health 2024 Aug;6(8):e555-e561. [doi: 10.1016/S2589-7500(24)00097-9] [Medline: 39059888]
43. Flores-Cohaila JA, García-Vicente A, Vizcarra-Jiménez SF, et al. Performance of ChatGPT on the Peruvian National Licensing Medical Examination: cross-sectional study. JMIR Med Educ 2023 Sep 28;9(1):e48039. [doi: 10.2196/48039] [Medline: 37768724]
44. Gasmalla HEE, Mohamed Tahir MEM. A-type MCQs. In: Gasmalla HEE, Ibrahim AAM, Wadi MM, Taha MH, editors. Written Assessment in Medical Education: Springer International Publishing; 2023:73-89. [doi: 10.1007/978-3-031-11752-7_6]
45. Nicholas G, Bhatia A. Lost in translation: large language models in non-English content analysis. arXiv. Preprint posted online on Jun 12, 2023. [doi: 10.48550/arXiv.2306.07377]
46. Alonso I, Oronoz M, Agerri R. MedExpQA: multilingual benchmarking of large language models for medical question answering. Artif Intell Med 2024 Sep;155:102938. [doi: 10.1016/j.artmed.2024.102938] [Medline: 39121544]
47. Gravel J, D'Amours-Gravel M, Osmanlliu E. Learning to Fake It: limited responses and fabricated references provided by ChatGPT for medical questions. Mayo Clin Proc Digit Health 2023 Sep;1(3):226-234. [doi: 10.1016/j.mcpdig.2023.05.004] [Medline: 40206627]
48. Roy S, Khatua A, Ghoochani F, Hadler U, Nejdl W, Ganguly N. Beyond accuracy: investigating error types in GPT-4 responses to USMLE questions. 2024 Jul 10 Presented at: SIGIR 2024; Jul 14-18, 2024; Washington, DC, USA p. 1073-1082 URL: https://dl.acm.org/doi/proceedings/10.1145/3626772 [accessed 2026-02-05] [doi: 10.1145/3626772.3657882]
49. Gierl MJ, Lai H, Turner SR. Using automatic item generation to create multiple-choice test items. Med Educ 2012 Aug;46(8):757-765. [doi: 10.1111/j.1365-2923.2012.04289.x] [Medline: 22803753]
50. Artsi Y, Sorin V, Konen E, Glicksberg BS, Nadkarni G, Klang E. Large language models for generating medical examinations: systematic review. BMC Med Educ 2024 Mar 29;24(1):354. [doi: 10.1186/s12909-024-05239-y] [Medline: 38553693]
51. Schuwirth LWT, van der Vleuten CPM. Different written assessment methods: what can be said about their strengths and weaknesses? Med Educ 2004 Sep;38(9):974-979. [doi: 10.1111/j.1365-2929.2004.01916.x] [Medline: 15327679]
52. Singhal K, Azizi S, Tu T, et al. Large language models encode clinical knowledge. Nature New Biol 2023 Aug 3;620(7972):172-180. [doi: 10.1038/s41586-023-06291-2]

XSL•FO
RenderX

Abbreviations

AI: artificial intelligence
API: application programming interface
LLM: large language model
OR: odds ratio

Edited by D Chartash; submitted 28.Jul.2025; peer-reviewed by K Sakaguchi, M Al-Agil, S Ewelu; revised version received 08.Jan.2026; accepted 09.Jan.2026; published 05.Mar.2026.

Please cite as:
Strasser LM, Anschuetz W, Dennstädt F, Hastings J
Performance Evaluation of Large Language Models in Multilingual Medical Multiple-Choice Questions: Mixed Methods Study
JMIR Med Educ 2026;12:e81399
URL: https://mededu.jmir.org/2026/1/e81399
doi:10.2196/81399

© Livia Maria Strasser, Wilma Anschuetz, Fabio Dennstädt, Janna Hastings. Originally published in JMIR Medical Education (https://mededu.jmir.org), 5.Mar.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Education, is properly cited. The complete bibliographic information, a link to the original publication on https://mededu.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Integrating Large Language Models Into Trauma Education for Medical Students: Randomized Controlled Pilot Trial

Joona Gustafsson[1*], MD; Erno Lehtonen-Smeds[2*], MD; Niklas Pakkasjärvi[1], MD, PhD, Prof Dr

[1]Wellbeing Services County of Southwest Finland, University of Turku, PO Box 52, Turku, Finland

[2]Wellbeing Services County of Ostrobotnia, Vaasa Central Hospital, Vaasa, Finland

[*]these authors contributed equally

Corresponding Author:
Niklas Pakkasjärvi, MD, PhD, Prof Dr
Wellbeing Services County of Southwest Finland, University of Turku, PO Box 52, Turku, Finland

Abstract

Background: The exponential growth of medical knowledge presents a paradox for modern medical education. While access to information is immediate, applying it in a clinically meaningful way remains a challenge. Large language models (LLMs), such as ChatGPT, are widely used for information retrieval, yet their role in dynamic, high-pressure clinical learning remains poorly understood.

Objective: This study aims to evaluate whether unstructured access to an LLM improves decision-making, teamwork, and confidence in trauma education for medical students.

Methods: This randomized controlled pilot study involved 41 final-year medical students participating in a trauma simulation session. Students self-selected into teams of 4 to 6 and were randomized to either an LLM-assisted group (ChatGPT-4o mini) or a control group without LLM access. All teams completed 18 video-based trauma scenarios requiring time-sensitive clinical decisions. Prompting was unrestricted. Confidence and trauma exposure were assessed using pre- or postquestionnaires. Facilitators rated teamwork (1-5), decision accuracy, and response times. Knowledge retention was measured 4 weeks later via an online quiz.

Results: Confidence in trauma management improved in both groups (P<.001), with larger gains in the non-LLM group (P=.02). LLM support did not enhance the decision accuracy or speed and was associated with longer response times in some complex cases. Teams without LLMs demonstrated more active discussion and scored higher in teamwork ratings (median 5.0 [IQR 5.0-5.0] vs median 3.5 [IQR 3.0-4.5]; P=.08). Students primarily used the LLM for fact-checking but reported vague or overly general responses. Knowledge retention was high across both groups and did not differ significantly (P=.33).

Conclusions: While students appreciated the inclusion of artificial intelligence (AI), unstructured LLM use did not improve performance and may have disrupted the group reasoning. The use of non-English prompting likely contributed to lower AI performance, underscoring the importance of language alignment in LLM applications. This pilot study highlights the need for structured AI integration and targeted instruction in AI literacy. Simulation-based trauma education proved effective and well received, but optimizing the educational value of LLMs will require thoughtful curricular design. Further studies with more students are needed to define best practices for LLM use in clinical education.

(JMIR Med Educ 2026;12:e79134) doi:10.2196/79134

KEYWORDS

large language models; LLM; medical education; artificial intelligence; ChatGPT; medical students; teamwork; decision-making

Introduction

The exponential growth of medical knowledge presents a paradox for modern medical education. While access to information is more immediate than ever, the ability to filter, contextualize, and apply that information remains a significant challenge. This concern is particularly relevant for medical students, who often lack the clinical experience to distinguish critical insights from background data [1,2]. Thus, a growing need for tools that provide information and help organize and synthesize it in clinically meaningful ways is evident.

Large language models (LLMs) such as ChatGPT have emerged as promising tools in this context [3]. Initially developed for language processing, LLMs operate on a probabilistic framework that predicts the next word in a sequence based on patterns in their training data [4]. Though not originally designed as knowledge engines, these models are increasingly used to retrieve and summarize medical information. LLMs can assist with fact-checking, clinical guidance, and even reflective

XSL•FO
RenderX

reasoning when prompted effectively [3,5,6]. However, their outputs must be interpreted cautiously, as they lack proper understanding and may generate plausible but incorrect information [7].

LLMs have excelled in diagnostic tasks under controlled conditions [8], yet these results may not generalize to dynamic clinical environments where uncertainty, time pressure, and teamwork play critical roles [9].

Despite growing enthusiasm for integrating artificial intelligence (AI) into medical education, there is limited high-quality evidence of its use in interactive, team-based, or high-stakes clinical learning settings. A systematic review by Lucas et al [5] highlighted the potential of LLMs to support learning but noted a lack of empirical trials in real-time clinical scenarios. Similarly, Nguyen [7] raised concerns about automation bias, where overreliance on AI can impair critical reasoning. To our knowledge, no randomized studies have directly examined how unstructured LLM support affects team dynamics and decision-making accuracy during trauma simulations.

To address this gap, we conducted a randomized controlled pilot study to examine whether a large language model (ChatGPT) can be meaningfully integrated into trauma simulations, how students use it in practice, and what problems emerge from its unstructured application. Students received no prior training in prompt formulation or AI use and were free to interact with the model at their discretion, reflecting real-world conditions. The study focused on observable outcomes including confidence, facilitator-rated teamwork interactions, decision accuracy, response times, and knowledge retention. These measures were used to characterize general patterns of LLM use within the constraints of the simulation setting. The aim was to generate evidence on how unstructured LLM availability may affect the learning process in team-based clinical simulations and to inform the design of future hypothesis-driven studies on structured AI integration in medical education.

Methods

Study Framework

Simulation-based trauma education relies on active, team-based learning, where decision-making under time pressure and effective communication are central to performance. These principles guided our choice of outcomes, focusing on confidence, facilitator-rated teamwork interactions, decision accuracy, response times, and short-term knowledge retention.

Introducing an LLM into this setting relates to emerging concepts in human-AI interaction, such as cognitive offloading and automation bias, which may influence how teams share information and make decisions. In the lack of consensus frameworks for integrating LLMs into team-based clinical simulations, we adopted a pragmatic, performance-focused approach to examine how students chose to use the LLM and whether its presence influenced observable aspects of the learning process.

This framework informed the design of this pilot study, which was not intended as an implementation trial but as an exploratory evaluation of how unstructured LLM availability may affect learning and team performance in trauma simulation.

Study Design and Setting

This single-center, exploratory pilot randomized controlled trial was conducted between January 13, 2025, and April 15, 2025, as part of the final-year trauma training curriculum at Vaasa Central Hospital, University of Turku, Finland. The study aimed to examine whether an LLM (ChatGPT) can be meaningfully integrated into trauma simulations, how students use it in practice, and what challenges arise from its unstructured application. The trial followed general CONSORT (Consolidated Standards of Reporting Trials) guidelines (Checklist 1) for randomized educational interventions.

Ethical Considerations

The study was reviewed by the Clinical Research Center, VARHA (Wellbeing Services County of Southwest Finland), and deemed exempt from formal ethical approval in accordance with institutional and national regulations governing educational research involving adult participants. The procedures followed were consistent with institutional guidelines and the principles of the Declaration of Helsinki.

All participating students received written and verbal information about the study and provided informed consent prior to participation. Participation was voluntary, and students were informed that declining participation would not affect their course standing or academic evaluation. No financial compensation or other incentives were provided.

No patient data were used in this study. Students accessed the LLM through a publicly available interface without logging in, and no identifiable student information was entered into the model at any point. Interactions with the LLM were not stored or exported, and no prompts or outputs were retained by the research team.

Only deidentified team-level performance data were collected. Individual-level responses were anonymized prior to analysis, and no identifiable information was reported. Data are available upon reasonable request. The study was preregistered on the Open Science Framework [10] on April 15, 2025. As with any generative AI tool, there is potential for language-related bias or factual inaccuracies, particularly when prompting in a language other than English. These considerations are discussed as limitations of the intervention.

Participants and Randomization

A total of 41 final-year medical students from the University of Turku, a publicly funded institution with a standardized curriculum, participated in the study. All students were enrolled in a voluntary trauma care simulation course. Students had previously self-selected into teams of 4 to 6 individuals at the beginning of the academic year, and these groups were preserved for the study. All participants had prior, curriculum-based simulation training experience.

A total of 8 pre-existing student teams (clusters) were randomized 1:1, resulting in 4 teams allocated to the LLM group (using ChatGPT-4o mini as a supportive tool during simulations)

XSL•FO
RenderX

and 4 teams allocated to the control group (no LLM support). All scenario-level outcomes (decision accuracy, response time, and teamwork scores) were aggregated to the team level for analysis to avoid pseudoreplication. A computer-generated randomization sequence was created by a researcher not involved in the simulation sessions. Group assignments were communicated to the facilitator immediately prior to each session. Due to the visible presence of the laptop used for LLM access, facilitators were not blinded to group allocation.

Because students had pre-established simulation groups, randomization was performed at the team level. Each intact team was assigned as a single unit to either the LLM or control condition. As the intervention was delivered to teams and all scenario-based decisions were made collectively, the team constituted the unit of randomization.

Intervention (LLM use)

Teams assigned to the intervention group were provided access to ChatGPT-4o mini via the publicly available free web interface. The same model was used in all sessions, with no premium features, plug-ins, or browsing capabilities enabled.

No prompting guidelines or technical instructions were given, and the interface remained unconfigured. Participants were free to interact with the model at their discretion and in any language; in practice, all teams chose Finnish. This decision to allow unstructured use was intentional, aiming to capture how students would naturally engage with an unconfigured, publicly available LLM in a clinical learning environment. Participants' prior familiarity with LLMs stemmed solely from personal experience, as no formal AI literacy or prompt engineering instruction was part of the curriculum.

All LLM-assisted sessions took place between January 13, 2025, and April 15, 2025. The model label displayed in the interface during all sessions was "ChatGPT-4o mini." Prompts and outputs were not systematically logged, as the study was designed to observe naturalistic use rather than evaluate prompt formulation. Students generally entered short, broad factual questions; often definitions, procedural steps, or indications, and invariably in Finnish.

Simulation Protocol

The instructional environment combined video-based scenario introductions with hands-on simulation using mannequins and procedural models (Multimedia Appendix 1). Each scenario began with a short video vignette sourced from publicly available YouTube material, used only to establish the clinical context and create a sense of urgency. All scenario content, tasks, and questions used in the simulation were developed by the study team. Students then managed each case using a human-like mannequin or low-fidelity simulator, performing the relevant assessments and interventions.

For example, pneumothorax scenarios included interpretation of simulated X-rays, verbal explanation of chest tube placement steps, and the procedure itself on a dedicated cardboard model. Throughout each scenario, students were asked open or multiple-choice questions related to diagnostics and management. Participants in the intervention arm could consult the LLM at any point for factual support or practical guidance. All scenarios were presented in the same order for all teams. Feedback was provided by trained facilitators after each case.

Data Collection and Outcome Measures

Data were collected before, during, and after the simulation sessions. Baseline characteristics, including self-reported trauma exposure and pre-session confidence ratings, are summarized in Table 1. Trauma exposure was categorized as curriculum-based (grade 1) or more extensive (grades 2 - 3) based on students' reported prior experience.

Table . Baseline characteristics of participating students (N=41).

Characteristics	LLM[a] group (n=22)	Control group (n=19)	*P* value[b]
Baseline confidence score, median (IQR)	2.17 (1.75 - 2.67)	2.33 (2.00 - 2.83)	.35
Prior simulation experience, n (%)	22 (100)	19 (100)	—[c]
Prior trauma exposure, n (%)[d]			
Grade 1	16 (73)	12 (63)	.51
Grade 2 - 3	6 (27)	7 (37)	.74

[a]LLM: large language model.

[b]*P* values were calculated using Mann–Whitney *U* test for continuous variables and Fisher exact test for categorical variables.

[c]Not applicable.

[d]Trauma exposure was categorized as grade 1 (curriculum-based) or grades 2 - 3 (extended experience).

Teamwork collaboration was assessed in real time by the session facilitator using a simple 1 to 5 rating scale developed for the trauma curriculum. Scores reflected observable indicators of collaboration, including verbal participation by multiple team members, shared decision-making, turn-taking, and the degree of coordinated problem-solving during each scenario. The tool was designed to provide a practical measure of team interaction in this pilot setting and was not intended as a validated teamwork instrument. Because the presence of the LLM interface was visible, blinding of outcome assessors was not feasible, and the potential for detection bias is acknowledged. Facilitators also recorded decision accuracy (correct or incorrect) and time to decision during each scenario. Following the simulation, participants in the LLM group completed a short, bespoke

XSL•FO
RenderX

postsession questionnaire developed for this pilot study. The instrument included 14 statements rated on a 1 to 5 Likert scale (1="strongly disagree," 5="strongly agree") covering perceived usefulness, ease of use, impact on learning, confidence, perceived limitations, and attitudes toward future integration. Three additional open-ended questions asked students to describe which aspects of the LLM were most helpful, what challenges they encountered, and how the tool could be improved for future training. The questionnaire was designed to obtain exploratory feedback rather than generate validated outcome measures, and responses were summarized descriptively to identify common patterns in students' perceptions. Knowledge retention was assessed via an anonymous online quiz 4 weeks after the simulation using a short, scenario-aligned online quiz. The quiz consisted of 7 multiple-choice questions, each with 4 answer options and 1 correct response. Items were blueprinted directly to the clinical content of the simulation scenarios and covered core trauma competencies, including chest drain placement, pleural puncture technique, lower-leg compartment anatomy, initial management of pelvic trauma, recognition of tension pneumothorax, indications for computed tomography imaging in blunt head trauma, and cervical spine precautions after high-energy mechanisms. Each correct answer contributed 1 point for a total possible score of 7, and results were analyzed as the proportion of correct responses. The quiz was designed for formative assessment; no formal reliability analysis was performed due to the small number of items and the pilot nature of the study.

Statistical Analysis

All primary team-based performance outcomes were analyzed at the team (cluster) level, corresponding to the unit of randomization. For each team, mean values across all 18 scenarios were calculated for response time, facilitator-rated teamwork collaboration interactions, and decision accuracy (correct or incorrect). Scenario-level comparisons (LLM vs control) for these team-level outcomes were conducted using independent-samples 2-tailed *t* tests or Mann-Whitney *U* tests, depending on distributional assumptions. Effect sizes (Cohen *d* or rank-biserial correlation) were reported where appropriate, and normality was assessed with the Shapiro-Wilk test.

Individual-level outcomes, including confidence ratings, trauma exposure, and knowledge-retention scores were collected at the participant level and summarized descriptively. These measures were not used for between-group inferential testing and were not cluster adjusted. A 2-sided $P<.05$ was considered statistically significant for team-level comparisons. As this was an exploratory pilot study, no primary end point was prespecified and no formal power calculation was performed; all eligible students who volunteered were included.

Results

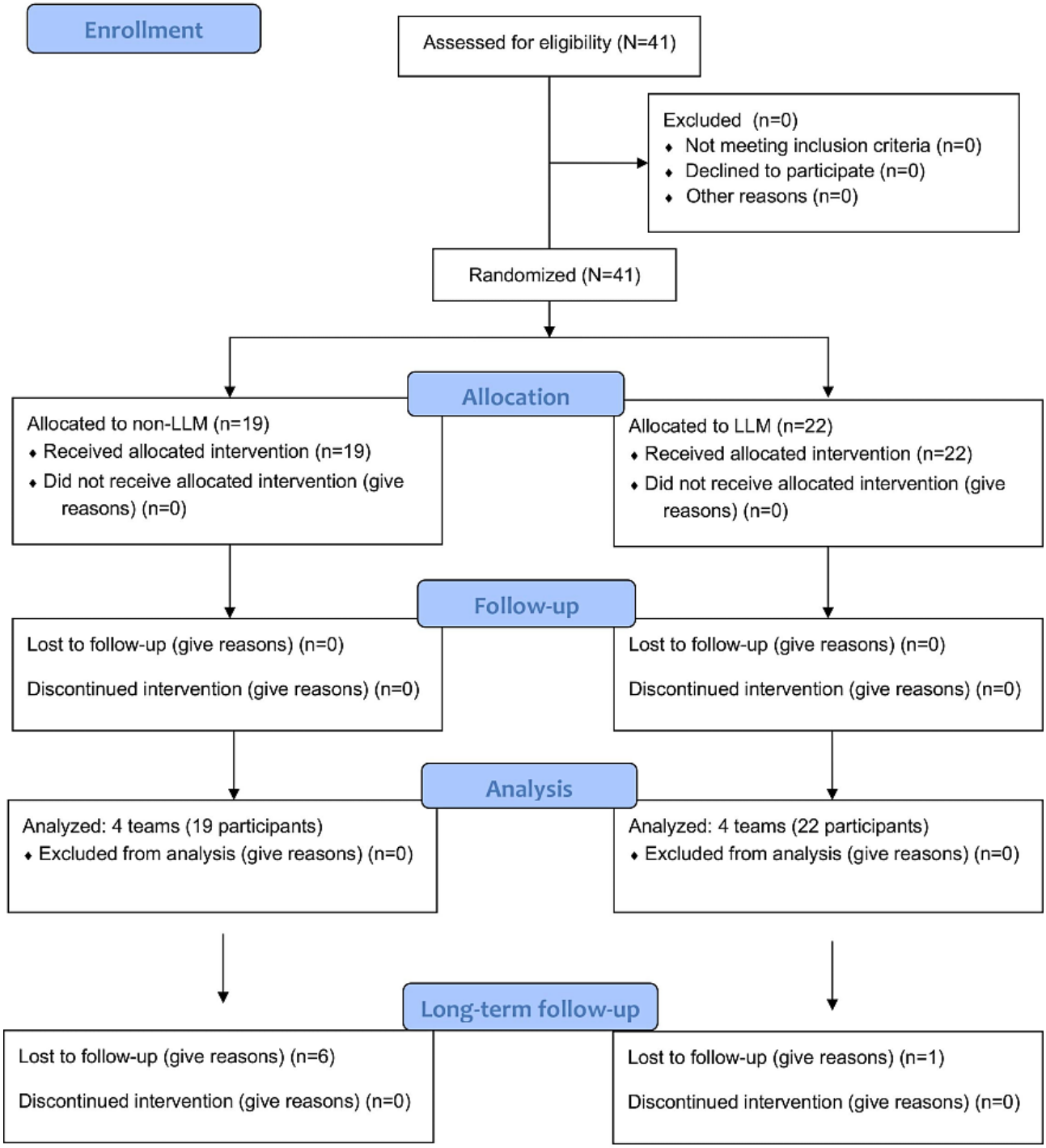
A total of 41 students participated in the study, forming the pre-existing teams that constituted the unit of randomization (Figure 1). Baseline characteristics, including initial confidence scores and prior trauma exposure, were similar between groups (Table 1).

Analyses were conducted at the team level, corresponding to the unit of randomization. Across all participants, self-reported confidence in trauma management increased significantly following the simulation session overall ($P=1.3 \times 10^{-}$; Table 2).

When comparing randomized teams, the non-LLM group demonstrated a greater median improvement in self-reported confidence compared with the LLM-assisted group ($P=.02$).

XSL•FO
RenderX

Figure 1. CONSORT (Consolidated Standards of Reporting Trials) flow diagram. LLM: large language model.



XSL•FO
RenderX

Table . Pre- and postsession confidence[a] scores (1–5) in trauma management.

Confidence domain[a] and trauma experience	Presession, median (IQR)	Postsession, median (IQR)	*P* value[b]
Treatment			
Curriculum-based	1.5 (1.0 - 2.0)	2.0 (2.0 - 3.0)	<.001
More extensive	2.0 (1.0 - 3.0)	3.0 (2.0 - 3.0)	.04
Decision-making			
Curriculum-based	3.0 (2.0 - 3.0)	3.0 (3.0 - 3.0)	<.001
More extensive	3.0 (3.0 - 3.0)	3.0 (3.0 - 3.0)	.99

[a]Confidence was assessed in two domains: (1) recognizing and treating life-threatening trauma situations (eg, hemodynamic instability, and pneumothorax) and (2) making decisions about patient disposition (eg, referral to specialized care, lower-level center, or discharge).

[b]Statistical significance was determined using the Wilcoxon signed-rank test, which can detect paired changes even when median values appear unchanged.

No statistically significant differences were observed between groups in mean decision-making accuracy (*P*=.06; *r*=0.75) or mean response time (*P*=.88; d=0.11; Table 3). Teamwork performance (5, 5, 5, 5 for non-LLM teams and 3, 3, 4, 5 for LLM-assisted teams) did not differ significantly between groups (*P*=.08; d=–1.85), although non-LLM teams exhibited higher median collaboration scores (5.0 vs 3.8). A trend toward longer response times in several complex scenarios persisted among LLM-assisted teams (Table 3). Neither cohort achieved consistently correct responses across all scenarios. Scenario-specific accuracy for both groups is presented in Table 3. Facilitator observations indicated that LLM-assisted teams tended to verbalize less, with several scenarios characterized by students turning to the laptop or silently reviewing the model's output instead of engaging in group discussion. In contrast, non-LLM teams typically talked through the cases more actively. Despite reduced interaction in the LLM groups, overall response times were not consistently longer, likely because consultation with the LLM often replaced, rather than added to, internal discussion time.

XSL•FO
RenderX

Table . Scenario-specific response times and accuracy[a].

Scenario	Expected action	Non-LLM[b] (s), median (IQR)	LLM (s), median (IQR)	*P* value[c]	Accuracy non-LLM/LLM (%)[d]
1	Recognize and manage a compromised airway	13.0 (11.0 - 18.0)	17.0 (14.5 - 18.75)	.88	100/100
2	Demonstrate correct placement and attachment of a chest tube	633.0 (499.0 - 718.5)	640.5 (544.75 - 697.5)	.99	100/100
3	Identify appropriate treatment options for tibial fractures	47.0 (38.75 - 58.25)	42.0 (34.75 - 49.75)	.89	100/100
4	Classify an open fracture	16.0 (5.5 - 30.25)	9.5 (3.5 - 36.5)	.99	100/100
5	Explain the classification of open fractures	24.5 (15.5 - 41.25)	55.5 (41.75 - 69.5)	.69	100/100
6	Identify the compartments of the lower leg	12.0 (8.5 - 15.25)	29.0 (21.5 - 41.5)	.11	100/100
7	List the anatomical components in each lower leg compartment	13.5 (10.0 - 28.75)	20.0 (13.5 - 27.0)	.89	50/100
8	Describe the diagnostic steps for compartment syndrome	188.5 (164.25 - 203.5)	135.5 (102.75 - 155.75)	.11	100/100
9	Describe the intramedullary tibial nailing procedure	45.0 (42.5 - 62.5)	44.5 (36.5 - 56.25)	.72	75/0
10	Identify and rank common complications associated with tibial nailing	82.5 (51.75 - 112.75)	133.0 (74.25 - 183.75)	.69	50/0
11	Determine discharge criteria based on trauma mechanism	221.0 (176.0 - 234.0)	129.0 (97.0 - 140.25)	.20	100/100
12	Outline the key steps in trauma patient evaluation	17.0 (16.5 - 19.5)	11.0 (10.0 - 14.25)	.19	100/100
13	Recognize pelvic fracture, assess for hemorrhage, and determine stabilization strategy	20.0 (15.0 - 46.5)	42.0 (20.0 - 64.75)	.89	100/100
14	Describe the management steps for massive bleeding in pelvic fractures	36.5 (29.25 - 40.25)	57.0 (41.0 - 67.25)	.34	100/100
15	Identify and manage a tension pneumothorax	12.5 (6.5 - 17.0)	15.5 (10.75 - 21.25)	.99	100/100
16	Identify indications for thoracotomy after chest drain insertion	36.0 (24.0 - 41.5)	25.0 (17.75 - 35.5)	.86	100/0
17	Explain the limitations of imaging in blunt abdominal trauma	11.5 (10.5 - 30.75)	26.5 (12.75 - 46.5)	.89	100/100
18	Identify indications for CT[e] imaging in blunt head trauma	45.5 (14.75 - 78.0)	10.5 (7.5 - 18.25)	.20	100/100

[a]Median (IQR) values are presented in seconds.

XSL•FO
RenderX

[b]LLM: large language model.

[c]*P* values were calculated using Mann-Whitney *U* tests.

[d]Accuracy is reported as the percentage of teams with a correct response (correct teams/total teams).

[e]CT: computed tomography.

Observational notes documented that teams with LLM access often deferred decisions to the AI tool without sufficient internal discussion or verification. Participants in the LLM-assisted teams completed a structured postsession questionnaire consisting of 14 Likert-scale items and three open-ended questions. Median Likert ratings indicated neutral to mildly positive perceptions of overall usefulness (median 3, IQR 3‐4) and impact on learning (median 3, IQR 2‐3), while ease of use showed mixed responses (median 2‐4 across items). Limitations were commonly endorsed, particularly regarding inaccuracies when prompting in Finnish (median 4, IQR 3‐4). Despite these challenges, students expressed generally positive attitudes toward future integration of LLMs into medical education (median 4, IQR 3‐4). Themes from open-ended comments reflected (1) primary use of the LLM for fact-checking and procedural clarification, (2) frustration with vague or overly general answers, and (3) support for more structured or guided integration in future sessions. Full item-level results are summarized in Table 4. Several factual inaccuracies were identified when the model was prompted in Finnish.

Table . Postsession questionnaire results large language models (LLMs; LLM-assisted teams).

Item	Domain	Median (IQR)
1. LLM improved ability to manage trauma cases	Usefulness	3 (3‐4)
2. LLM provided valuable insights for decision-making	Usefulness	3 (3‐4)
3. LLM was simple and intuitive to use	Ease of use	2 (2‐3)
4. I felt comfortable using the LLM during the simulation	Ease of use	4 (3‐4)
5. LLM had a positive effect on my learning experience	Learning impact	3 (2‐3)
6. LLM helped me better understand trauma care principles	Learning impact	3 (3‐4)
7. I am satisfied that LLM support was included	Satisfaction	3 (3‐3)
8. I would recommend including LLM support in future training	Satisfaction	4 (3‐4)
9. LLM increased my confidence in handling real trauma cases	Confidence	2 (2‐3)
10. I feel better prepared to make clinical decisions after using the LLM	Confidence	3 (2‐4)
11. At times, LLM support was more harmful than helpful	Limitations	3 (3‐3)
12. I encountered inaccuracies or limitations in LLM responses	Limitations	4 (3‐4)
13. LLMs should be integrated into medical education	Future integration	4 (3‐4)
14. I am willing to use LLM support in future training	Future integration	4 (4‐4)

A total of 34 out of 41 students completed the knowledge retention assessment 4 weeks post session. The proportion of correct responses did not differ significantly between groups (non-LLM group: 89.2%; LLM group: 84.4%; *P*=.33).

Discussion

Principal Findings

In this exploratory randomized pilot study, unstructured access to an LLM did not improve decision-making speed or accuracy in trauma simulations. Scenario-specific patterns were observed: in 3 of the 18 scenarios, all LLM-assisted teams selected

XSL•FO
RenderX

incorrect answers, whereas non-LLM teams showed variable but generally higher accuracy. Teams using the LLM demonstrated lower teamwork scores and were less interactive, frequently deferring decisions to the AI tool. Students in both groups reported increased confidence after the session, without additional benefit from LLM use. Observational notes suggest that poor prompt formulation and overreliance on the model may have contributed to delays and decision errors. However, prompt precision was not formally measured and should be addressed in future studies. Several AI outputs appeared to contain factual inaccuracies, particularly when prompted in Finnish; these were not systematically evaluated and are therefore discussed only as qualitative observations to guide future research.

Facilitator observations provided additional insight into how LLM access influenced team behavior. Teams with LLM support frequently paused to consult the model or read its output, which appeared to replace some of the natural verbal processing that typically occurs during team-based simulations. This shift contributed to lower teamwork ratings despite comparable response times. The LLM often produced rapid but broad or nonspecific answers; as a result, consultation rarely prolonged task completion but did interrupt collaborative reasoning. These behavioral patterns highlight the importance of structured guidance for LLM use to prevent disruption of team communication in fast-paced clinical environments.

Comparison With Prior Literature

These findings align with emerging literature on AI in medical education. LLMs such as ChatGPT offer promising support for learners but may struggle to deliver context-rich reasoning or reliably accurate responses without structured prompts or domain tuning [5,7,11]. Prompting quality has emerged as a key determinant of model performance. Recent studies emphasize that targeted prompt engineering can significantly improve output accuracy, relevance, and response time, particularly in procedural and diagnostic contexts [12-14]. This supports the notion that structured AI literacy or microtraining could mitigate some of the issues observed during real-time trauma simulations.

Moreover, the multilingual context matters: LLMs trained primarily in English may underperform in other languages, underscoring the need for language-aligned or locally fine-tuned models to ensure equitable educational outcomes across linguistic contexts [15,16]. Few studies have examined the impact of LLMs on team-based, high-stakes simulation environments. Evidence from trauma team simulations shows that context can strongly influence teamwork and cognitive load even without AI, and human-AI studies warn that automation bias and AI teammate behaviors can disrupt coordination if not carefully structured [17-21]. These factors may help explain the neutral or negative performance effects observed in our trial.

Implications of Findings

This study highlights how unstructured use of LLMs may influence clinical reasoning and teamwork during time-sensitive scenarios. Reliance on AI tools can inadvertently disrupt the interactive human problem–solving processes that are critical to trauma care. These behavioral effects emphasize the need to integrate LLMs into medical education in a structured and guided manner, rather than assuming benefit from mere access. This aligns with broader observations in surgical education, where nontechnical competencies such as teamwork, communication, and decision-making are increasingly recognized as critical learning objectives but often underemphasized in training frameworks [22]. Although not a primary outcome, several factual inaccuracies were observed in AI-generated outputs during the simulations. Students prompted the model in Finnish, which likely contributed to these errors, as most LLMs are trained primarily on English-language data. These findings underscore the importance of improving multilingual performance and localization in LLMs to ensure applicability across diverse health care systems [15]. Beyond immediate simulation performance, a broader concern is the potential shift in learning behaviors among students [23]. Overreliance on AI tools may reduce active engagement with primary sources and critical appraisal skills—competencies essential for clinical reasoning and lifelong learning. Educators should therefore emphasize structured AI literacy and clear pedagogical frameworks to ensure that LLMs function as supportive decision aids rather than substitutes for clinical judgment [7].

Strengths and Limitations

Limitations of this study include the relatively small sample size and discretionary use of LLMs, which may have introduced variability in engagement. Students self-selected into teams prior to randomization, potentially affecting group dynamics. Knowledge retention was generally high across both groups, but detailed instructor-led debriefings of all cases may have minimized group differences, making comparisons less meaningful. Additionally, no formal power calculations were performed; all eligible students who volunteered were included, and no participants declined.

A key limitation concerns the use of nonvalidated instruments for teamwork and decision accuracy, reflecting the pilot nature of the study. Importantly, students received no prompting or language-use instruction, and all teams independently chose to interact with the LLM in Finnish. This unstructured use may have contributed to variable prompting quality, factual inaccuracies, and longer response times due to language mismatch with the model. At the same time, this unstructured and naturalistic use of the tool is also a strength of the study, as it mirrors how students without prior AI training are likely to engage with LLMs in real-world settings. This highlights the need for structured prompting instruction and language-aligned models in future educational interventions and provides a useful baseline against which to design and evaluate more structured trials.

Future Directions

Further research should explore the role of AI in structured learning environments, emphasizing support for, rather than replacement of, active clinical reasoning. This pilot study underscores both opportunities and limitations of integrating LLMs into clinical training. It provides a foundation for refining educational strategies, and future studies with greater statistical

XSL•FO
RenderX

power will help define the optimal use of these tools in medical education. Ultimately, the question is not whether to use AI in medical education, but how to implement it effectively to support human reasoning and teamwork.

Funding

The authors declared no financial support was received for this work.

Data Availability

Deidentified data are available upon reasonable request from the corresponding author.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Detailed description of simulation setup.
[DOCX File, 17 KB - mededu_v12i1e79134_app1.docx]

Checklist 1
CONSORT checklist.
[PDF File, 83 KB - mededu_v12i1e79134_app2.pdf]

References

1. Smith R. Strategies for coping with information overload. BMJ 2010 Dec 15;341:c7126. [doi: 10.1136/bmj.c7126] [Medline: 21159764]
2. Park SH, Pinto-Powell R, Thesen T, et al. Preparing healthcare leaders of the digital age with an integrative artificial intelligence curriculum: a pilot study. Med Educ Online 2024 Dec 31;29(1):2315684. [doi: 10.1080/10872981.2024.2315684] [Medline: 38351737]
3. Shool S, Adimi S, Saboori Amleshi R, Bitaraf E, Golpira R, Tara M. A systematic review of large language model (LLM) evaluations in clinical medicine. BMC Med Inform Decis Mak 2025 Mar 7;25(1):117. [doi: 10.1186/s12911-025-02954-4] [Medline: 40055694]
4. Biesheuvel LA, Workum JD, Reuland M, et al. Large language models in critical care. J Intensive Med 2024 Apr;5(2):113-118. [doi: 10.1016/j.jointm.2024.12.001] [Medline: 40241839]
5. Lucas HC, Upperman JS, Robinson JR. A systematic review of large language models and their implications in medical education. Med Educ 2024 Nov;58(11):1276-1285. [doi: 10.1111/medu.15402] [Medline: 38639098]
6. Goodell AJ, Chu SN, Rouholiman D, Chu LF. Large language model agents can use tools to perform clinical calculations. NPJ Digit Med 2025 Mar 17;8(1):163. [doi: 10.1038/s41746-025-01475-8] [Medline: 40097720]
7. Nguyen T. ChatGPT in medical education: A precursor for automation bias? JMIR Med Educ 2024 Jan 17;10:e50174. [doi: 10.2196/50174] [Medline: 38231545]
8. Goh E, Gallo R, Hom J, et al. Large language model influence on diagnostic reasoning: a randomized clinical trial. JAMA Netw Open 2024 Oct 1;7(10):e2440969. [doi: 10.1001/jamanetworkopen.2024.40969] [Medline: 39466245]
9. Chang CT, Farah H, Gui H, et al. Red teaming ChatGPT in medicine to yield real-world insights on model behavior. NPJ Digit Med 2025 Mar 7;8(1):149. [doi: 10.1038/s41746-025-01542-0] [Medline: 40055532]
10. LLM use in escape room trauma simulations. Open Science Framework (OSF). URL: https://osf.io/7hf3v/overview [accessed 2026-03-03]
11. Safranek CW, Sidamon-Eristoff AE, Gilson A, Chartash D. The role of large language models in medical education: applications and implications. JMIR Med Educ 2023 Aug 14;9:e50945. [doi: 10.2196/50945] [Medline: 37578830]
12. Shusterman R, Waters AC, O'Neill S, Bangs M, Luu P, Tucker DM. An active inference strategy for prompting reliable responses from large language models in medical practice. NPJ Digit Med 2025 Feb 22;8(1):119. [doi: 10.1038/s41746-025-01516-2] [Medline: 39987335]
13. Meskó B. Prompt engineering as an important emerging skill for medical professionals: tutorial. J Med Internet Res 2023 Oct 4;25:e50638. [doi: 10.2196/50638] [Medline: 37792434]
14. Zaghir J, Naguib M, Bjelogrlic M, Névéol A, Tannier X, Lovis C. Prompt engineering paradigms for medical applications: scoping review. J Med Internet Res 2024 Sep 10;26(1):e60501. [doi: 10.2196/60501] [Medline: 39255030]
15. Qiu P, Wu C, Zhang X, et al. Towards building multilingual language model for medicine. Nat Commun 2024 Sep 27;15(1):8384. [doi: 10.1038/s41467-024-52417-z] [Medline: 39333468]

XSL•FO
RenderX

16. Liu M, Okuhara T, Chang X, et al. Performance of ChatGPT across different versions in medical licensing examinations worldwide: systematic review and meta-analysis. J Med Internet Res 2024 Jul 25;26:e60807. [doi: 10.2196/60807] [Medline: 39052324]
17. Lapierre A, Lavoie P, Castonguay V, Lonergan AM, Arbour C. The influence of the simulation environment on teamwork and cognitive load in novice trauma professionals at the emergency department: piloting a randomized controlled trial. Int Emerg Nurs 2023 Mar;67(3):101261. [doi: 10.1016/j.ienj.2022.101261] [Medline: 36804137]
18. Goddard K, Roudsari A, Wyatt JC. Automation bias: empirical results assessing influencing factors. Int J Med Inform 2014 May;83(5):368-375. [doi: 10.1016/j.ijmedinf.2014.01.001] [Medline: 24581700]
19. Skitka LJ, Mosier KL, Burdick M, Rosenblatt B. Automation bias and errors: are crews better than individuals? Int J Aviat Psychol 2000;10(1):85-97. [doi: 10.1207/S15327108IJAP1001_5] [Medline: 11543300]
20. Schmutz JB, Outland N, Kerstan S, Georganta E, Ulfert AS. AI-teaming: redefining collaboration in the digital era. Curr Opin Psychol 2024 Aug;58:101837. [doi: 10.1016/j.copsyc.2024.101837] [Medline: 39024969]
21. Bienefeld N, Kolbe M, Camen G, Huser D, Buehler PK. Human-AI teaming: leveraging transactive memory and speaking up for enhanced team effectiveness. Front Psychol 2023;14:1208019. [doi: 10.3389/fpsyg.2023.1208019] [Medline: 37599773]
22. Pakkasjärvi N, Anttila H, Pyhältö K. What are the learning objectives in surgical training—a systematic literature review of the surgical competence framework. BMC Med Educ 2024 Feb 6;24(1):119. [doi: 10.1186/s12909-024-05068-z] [Medline: 38321437]
23. Peláez-Sánchez IC, Velarde-Camaqui D, Glasserman-Morales LD. The impact of large language models on higher education: exploring the connection between AI and Education 4.0. Front Educ 2024;9:1392091. [doi: 10.3389/feduc.2024.1392091]

Abbreviations

AI: artificial intelligence
CONSORT: Consolidated Standards of Reporting Trials
LLM: large language model

Edited by B Lesselroth; submitted 16.Jun.2025; peer-reviewed by A Adepoju, MBR Mahmoud, M Al-Agil, N Bievre, S Aluvaka; accepted 22.Dec.2025; published 17.Mar.2026.

Please cite as:
Gustafsson J, Lehtonen-Smeds E, Pakkasjärvi N
Integrating Large Language Models Into Trauma Education for Medical Students: Randomized Controlled Pilot Trial
JMIR Med Educ 2026;12:e79134
URL: https://mededu.jmir.org/2026/1/e79134
doi:10.2196/79134

© Joona Gustafsson, Erno Lehtonen-Smeds, Niklas Pakkasjärvi. Originally published in JMIR Medical Education (https://mededu.jmir.org), 17.Mar.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Education, is properly cited. The complete bibliographic information, a link to the original publication on https://mededu.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

ChatGPT versus UpToDate in Preclinical Medical Education: Cross-Sectional Analysis Using Term Frequency–Inverse Document Frequency Cosine Similarity

Shankar S Thiru[1], BS; Nicholas E Aksu[2], MD; Matthew Chiang[3], BS; Daniel O Gallagher[4], MD; Mary Furlong[5], MD; Elizabeth R Prevou[6], DMSc; Akhil Jay Khanna[2], MD

[1]Georgetown University School of Medicine, 3800 Reservoir Rd, Washington, DC, United States

[2]Department of Orthopaedic Surgery, MedStar Georgetown University Hospital, Washington, DC, United States

[3]DataChat Inc., Madison, WI, United States

[4]Rothman Orthopaedic Institute, Thomas Jefferson University Hospital, Philadelphia, PA, United States

[5]Department of Pathology and Laboratory Medicine, Georgetown University School of Medicine, Washington, DC, United States

[6]The George Washington University School of Medicine and Health Sciences, Washington, DC, United States

Corresponding Author:
Shankar S Thiru, BS
Georgetown University School of Medicine, 3800 Reservoir Rd, Washington, DC, United States

Abstract

Background: Generative artificial intelligence tools such as ChatGPT are increasingly used by medical students for self-directed learning. Although these models demonstrate linguistic fluency, their reliability as supplementary resources for preclinical education remains uncertain. In particular, comparisons with evidence-based references such as UpToDate are lacking.

Objective: This study evaluated the similarity between responses generated by ChatGPT (with GPT-4o mini) and those from UpToDate to preclinical medical education questions to assess ChatGPT's potential as an adjunctive learning tool.

Methods: We conducted a cross-sectional comparison study using 150 first-order questions derived from a preclinical question bank at a single allopathic institution under the oversight of a medical educator with more than 25 years of teaching experience. Each question was entered into ChatGPT 10 times in separate chat sessions, and responses from UpToDate were retrieved from the most relevant articles. The responses were preprocessed through lemmatization, stop-word removal, punctuation removal, and numeric normalization. Similarity between ChatGPT and UpToDate responses was quantified using term frequency–inverse document frequency (TF-IDF) cosine similarity. To determine whether the observed similarities exceeded chance, ChatGPT outputs were compared with a null distribution generated from randomized text.

Results: ChatGPT responses demonstrated statistically significant similarity to UpToDate in 59.3% (89/150) of questions. Across subject areas, pharmacology showed the highest concordance (mean cosine similarity 0.338, SD 0.134), followed by pathology (mean 0.321, SD 0.142), biochemistry (mean 0.296, SD 0.120), microbiology (mean 0.297, SD 0.108), and immunology (mean 0.275, SD 0.102). All subject-level similarity scores exceeded those generated from randomized text, confirming that the observed overlap was nonrandom.

Conclusions: ChatGPT with GPT-4o mini exhibited moderate but meaningful alignment with UpToDate across preclinical topics, performing best in fact-based disciplines such as pharmacology. Although it is not a substitute for evidence-based resources, ChatGPT may serve as an accessible adjunctive tool for medical students. Integration into preclinical learning should be coupled with artificial intelligence literacy training to promote responsible use and critical appraisal.

(JMIR Med Educ 2026;12:e82885) doi:10.2196/82885

KEYWORDS

artificial intelligence; medical education; preclinical; ChatGPT; UpToDate

Introduction

Artificial intelligence (AI) refers to the ability of digital systems to perform tasks traditionally requiring human intelligence, such as pattern recognition, problem-solving, and decision-making [1-3]. One rapidly advancing branch of AI is generative AI, which includes large language models (LLMs) capable of producing novel text, images, or other media based on extensive training datasets. ChatGPT, an LLM developed by OpenAI, enables iterative question-and-answer dialogue with users and has generated significant interest for its potential applications in health care and education [4,5].

XSL•FO
RenderX

In medical education, ChatGPT has been investigated for several instructional benefits, including its ability to generate customized clinical cases, provide immediate feedback, and distill complex topics into accessible explanations [6]. Early studies have even demonstrated that ChatGPT can perform at or above passing thresholds for step 1 of the United States Medical Licensing Examination, suggesting proficiency with biomedical content integration [7]. At the same time, learners are increasingly incorporating generative AI tools into their study practices for tasks such as summarizing notes, preparing for board examinations, and conducting self-assessments [8-13]. This trend has raised important questions about the reliability, accuracy, and alignment of generative AI tools with evidence-based resources. Prior work indicates that although ChatGPT demonstrates linguistic fluency and strong knowledge recall, it can underperform in tasks requiring algorithmic reasoning, symbolic processing, or deep, domain-specific expertise [14].

Despite growing attention to the use of generative AI in clinical training and high-stakes assessment contexts, the role of ChatGPT for students engaged in the preclinical curriculum remains largely unexplored. Equally important, there have been no studies directly comparing ChatGPT to UpToDate, a widely adopted, evidence-based medical knowledge database that serves as a cornerstone resource for medical trainees. Understanding how ChatGPT's responses align with an established standard is essential to determining whether it may serve as a credible supplementary tool in early medical education.

Accordingly, this study evaluated ChatGPT (with GPT-4o mini) in comparison with UpToDate across a series of preclinical curricular questions by examining the degree of similarity between ChatGPT-generated responses and content provided by UpToDate to determine whether generative AI can serve as a reliable adjunct to evidence-based resources for first- and second-year medical students.

Methods

Study Design

This was a cross-sectional comparison study (Figure 1). We examined how ChatGPT's free version (GPT-4o mini) responded to preclinical medical science questions compared with the medical knowledge database, UpToDate. UpToDate was selected because it is a continuously updated, physician-authored, and peer-reviewed medical knowledge resource that compiles primary literature into standardized, consensus-based explanations. Although primarily designed for clinical decision support, UpToDate extensively covers foundational biomedical concepts that directly overlap with preclinical curricula, including pathophysiology, pharmacologic mechanisms, microbiology, and immunology. Prior studies have used UpToDate as a reference standard when evaluating the accuracy of digital health tools and clinical decision-support systems, supporting its use as a high-quality comparator for authoritative medical knowledge [15-17].

XSL•FO
RenderX

Figure 1. The term frequency–inverse document frequency (TF-IDF) cosine similarity and text networks for preclinical subject questions and aggregate ChatGPT responses in biochemistry. The statistical baseline was determined from textual similarity to randomly generated text.

XSL•FO
RenderX

Question Selection

A total of 150 questions were derived from a preclinical medical education question bank used at Georgetown University. Questions were selected to reflect core content areas common to standard US preclinical curricula, including biochemistry, immunology, microbiology, pharmacology, and pathology. To approximate real-world information-seeking behavior, this study specifically included first-order questions (ie, those querying a single fact or concept) comparable in scope and difficulty to material found in step 1 of the foundational United States Medical Licensing Examination, which reflects typical student search patterns. Although the question set was not formally validated across multiple institutions, its content and difficulty were reviewed to ensure alignment with nationally standardized learning objectives. Question selection and categorization were conducted under the oversight of a board-certified medical educator with over 25 years of experience in undergraduate medical education, ensuring curricular relevance and pedagogical appropriateness for the preclinical learner.

Data Collection

Each question was entered verbatim into ChatGPT 10 separate times, with each entry conducted in a new chat window and without a logged-in user to prevent prior prompts from influencing responses. Multiple samples were collected to account for the inherent stochasticity of ChatGPT, as repeated queries can yield variable phrasing or content for identical prompts. GPT-4o mini was selected because it is the free, publicly available version and therefore the version most likely to be used by students during self-study.

For UpToDate, each question was entered into the search bar function (using Google Chrome version 131.0, in incognito mode, with the browsing history cleared). To standardize data collection, only the overview or explanatory sections of the most relevant UpToDate articles were extracted for analysis. When multiple UpToDate articles were relevant to a given question, the section deemed most directly responsive to the query was selected based on content specificity and scope. Selection criteria for articles included (1) direct factual or mechanistic information, (2) conciseness and clarity of explanation, and (3) being authored or reviewed by recognized experts in the field, as indicated by UpToDate's editorial process. For example, if a pharmacology question asked about the mechanism of action of a specific drug class, the section explicitly describing the drug's pharmacodynamics was selected over general overviews or therapeutic context.

UpToDate responses were collected once, reflecting its status as a static, curated knowledge source. Sensitivity analyses using subsets of ChatGPT outputs (eg, first, median, and maximum similarity scores across the 10 samples) confirmed that overall trends and domain-level patterns remained consistent, ensuring comparisons captured meaningful alignment rather than artifacts of sampling variability (Table 1). This approach ensured that responses represented the type of summary text a student would likely consult.

Table . Sensitivity analysis of ChatGPT term frequency–inverse document frequency similarity scores.

Domain	First	Median (SD)	Maximum to minimum
Pharmacology	0.330	0.338 (0.014)	0.355 - 0.310
Pathology	0.312	0.321 (0.015)	0.340 - 0.290
Biochemistry	0.290	0.296 (0.013)	0.310 - 0.265
Microbiology	0.290	0.297 (0.016)	0.310 - 0.275
Immunology	0.270	0.275 (0.011)	0.290 - 0.250

Text Preprocessing (Normalization)

Responses were normalized to allow meaningful comparison because natural language allows for variation (eg, "run," "ran," "running" all convey the same core concept). Normalization included lemmatization, which reduced words to a common base form (eg, "runs" to "run"); stop-word removal, which eliminated common filler words such as "the" or "and" that do not affect meaning; syntax removal, which removed punctuation and formatting; and number handling, in which all numeric values were dropped to prevent disproportionate influence from isolated numbers (eg, drug dosages). These steps reduced irrelevant variation while preserving core content.

Text-to-Network Conversion and Similarity Analysis

Each processed response was converted into a vector using term frequency–inverse document frequency (TF-IDF). TF-IDF weights each word according to both its frequency within a response and its rarity across the dataset, ensuring that uncommon but meaningful words carry greater weight while common words contribute less.

ChatGPTnode:vi*

UpToDatenode:vj

Element−wise mean of weights: $wj = 1N*M\sum i=1N\sum j=1Meij$

To compare two responses (eg, ChatGPT vs UpToDate), we calculated cosine similarity between their TF-IDF vectors. Cosine similarity ranges from 0 (no similarity in text) to 1 (identical text). For example, high similarity (≈0.85) is observed between "doctors treat patients in hospitals" versus "physicians treat patients in outpatient hospitals," whereas low similarity (≈0.30) is observed between "doctors treat patients in hospitals" and "engineers design bridges with software." Thus, higher scores indicate greater overlap in informational content.

To account for variability in ChatGPT outputs, 5 responses per question were compared to the corresponding UpToDate text, and the results were aggregated into an average similarity score per question. Aggregated scores were then averaged within each

XSL•FO
RenderX

subject area to examine differences across disciplines, consequently using the same textual information but different corpora composition. Definitions of terminology are found in Table 2.

Table . Definitions of terms for analytical tools used in the analysis.

Term	Definition
TF-IDF[a]	A statistical, quantitative measure used to evaluate the importance of a word in a document relative to a collection of documents (corpus). It combines two components: the frequency of a term in a document (term frequency) and the inverse frequency of that term across all documents in the corpus (inverse document frequency). High TF-IDF values indicate words that are unique to specific documents.
Text vector network	A network representation of text where each node is a vector representation of words, phrases, or entire documents. These vectors are typically derived using the TF-IDF values of a text, and the edges represent relationships such as the similarity between the vectors. This type of analysis helps capture semantic and contextual relationships within large text corpora.
Cosine similarity matrix	A matrix used to measure the cosine of the angle between vectors representing text in a vector space model. The cosine similarity ranges from −1 (completely dissimilar) to 1 (identical). In text analysis, it is commonly used to assess the similarity between documents or terms based on their vectorized representations. A higher value corresponds to a higher similarity between two texts.
Text network analysis	An overall approach of analyzing the relationships between words, phrases, or documents using text vector networks and cosine similarity matrices. This approach helps uncover patterns, structures, and clusters in textual data.

[a]TF-IDF: term frequency–inverse document frequency.

Statistical Analysis

To determine whether observed similarities exceeded chance, we created a null distribution by generating 1000 randomized text samples that matched the length and word distribution of UpToDate outputs (Multimedia Appendix 1). The TF-IDF cosine similarity scores were calculated for these random texts against UpToDate. This approach enabled comparison of ChatGPT and UpToDate similarity with a baseline random similarity distribution. Statistical significance was defined as a probability value of less than .05, with ChatGPT and UpToDate similarity exceeding the 95th percentile of random-text similarities.

Study Outcomes

The primary outcome was the similarity (ie, cosine similarity score) between ChatGPT and UpToDate responses at the individual question level. The secondary outcome was the difference in similarity across preclinical subject categories.

Ethical Considerations

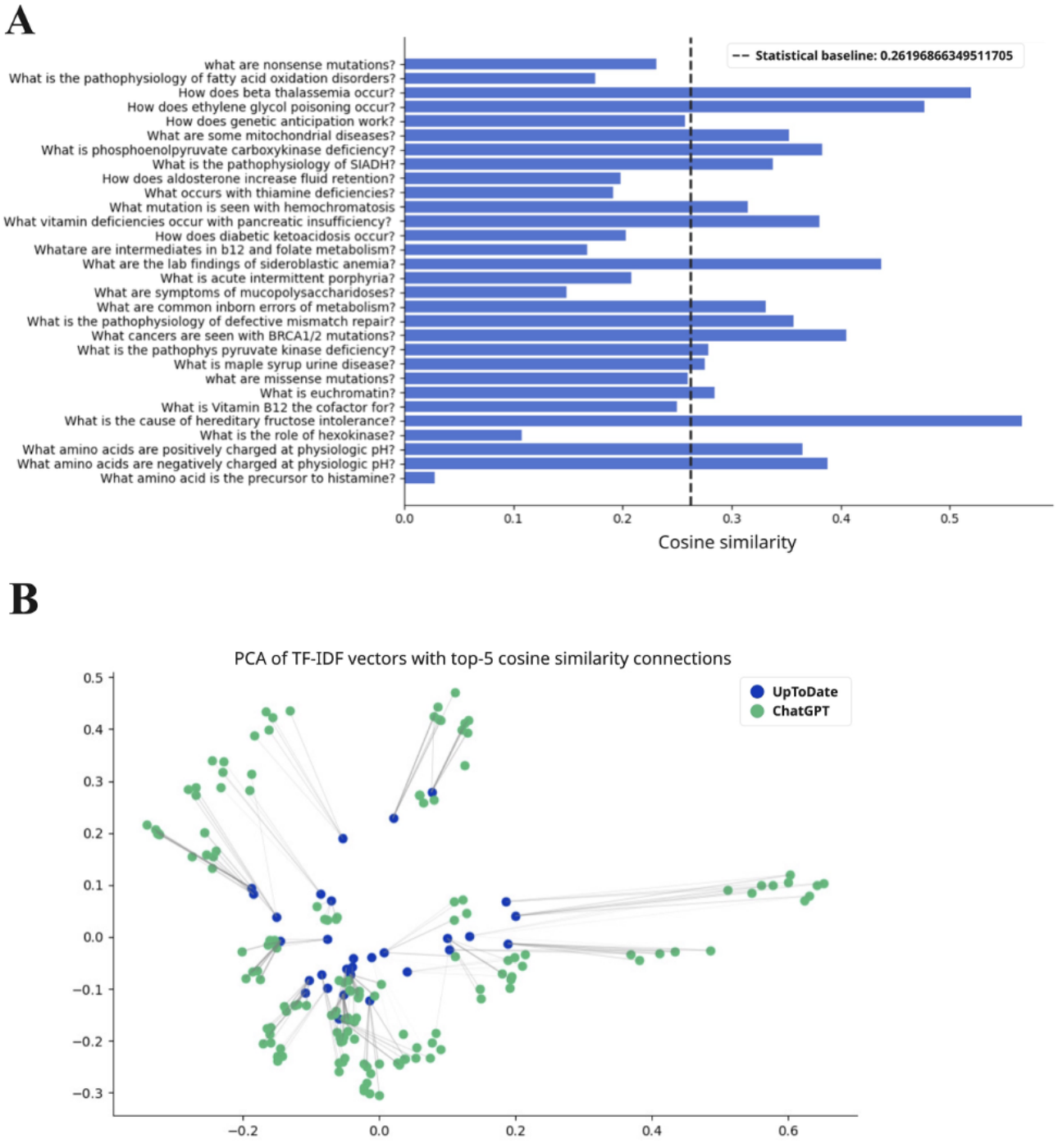
This study did not involve human or animal subjects; therefore, ethical approval was not required.

Results

ChatGPT responses demonstrated moderate similarity to UpToDate across preclinical topics. Using TF-IDF cosine similarity, 89 of 150 questions (59.3%) produced ChatGPT responses that were statistically similar to those from UpToDate (Figures 2-6; precise values are available in Multimedia Appendix 2).

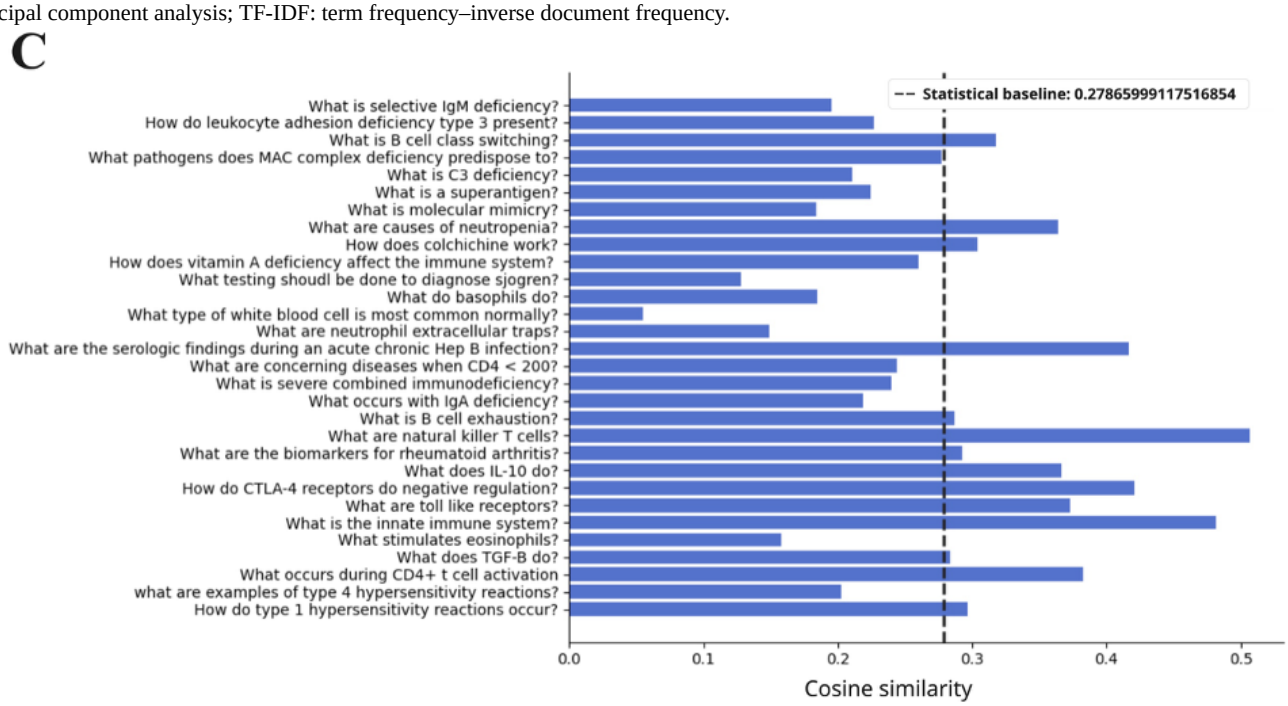
XSL•FO
RenderX

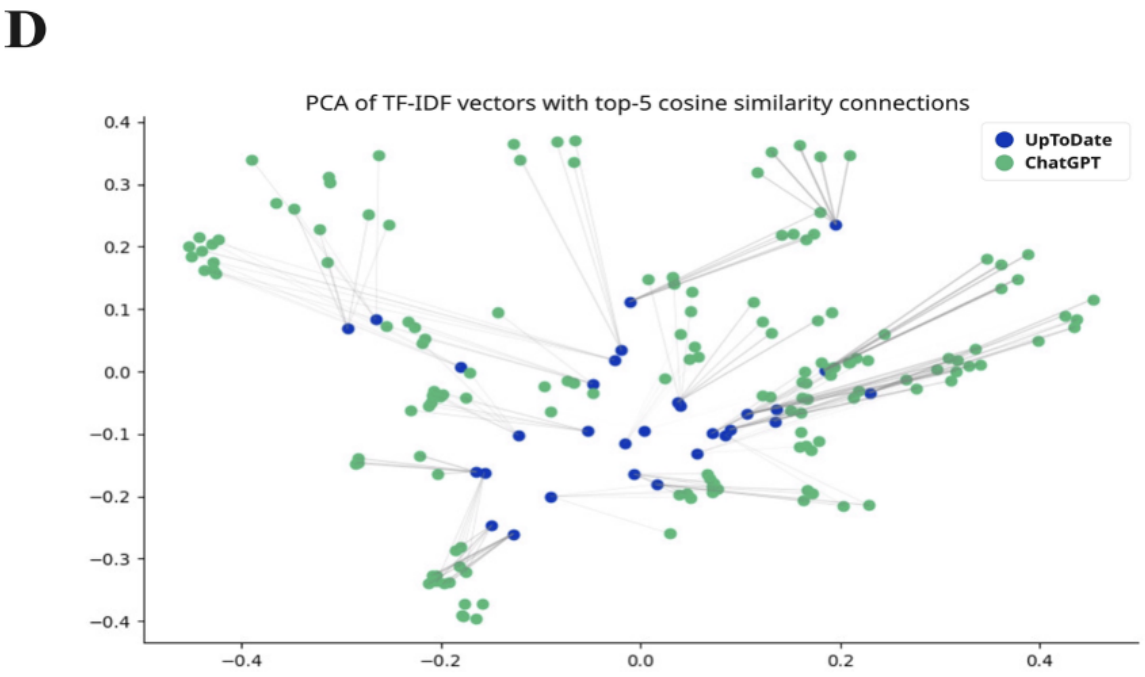
Figure 2. The TF-IDF cosine similarity and text networks for preclinical subject questions and aggregate ChatGPT responses for immunology. The statistical baseline was determined from textual similarity to a randomly generated text. PCA: principal component analysis; SIADH: syndrome of inappropriate antidiuretic hormone secretion; TF-IDF: term frequency–inverse document frequency.



XSL•FO
RenderX

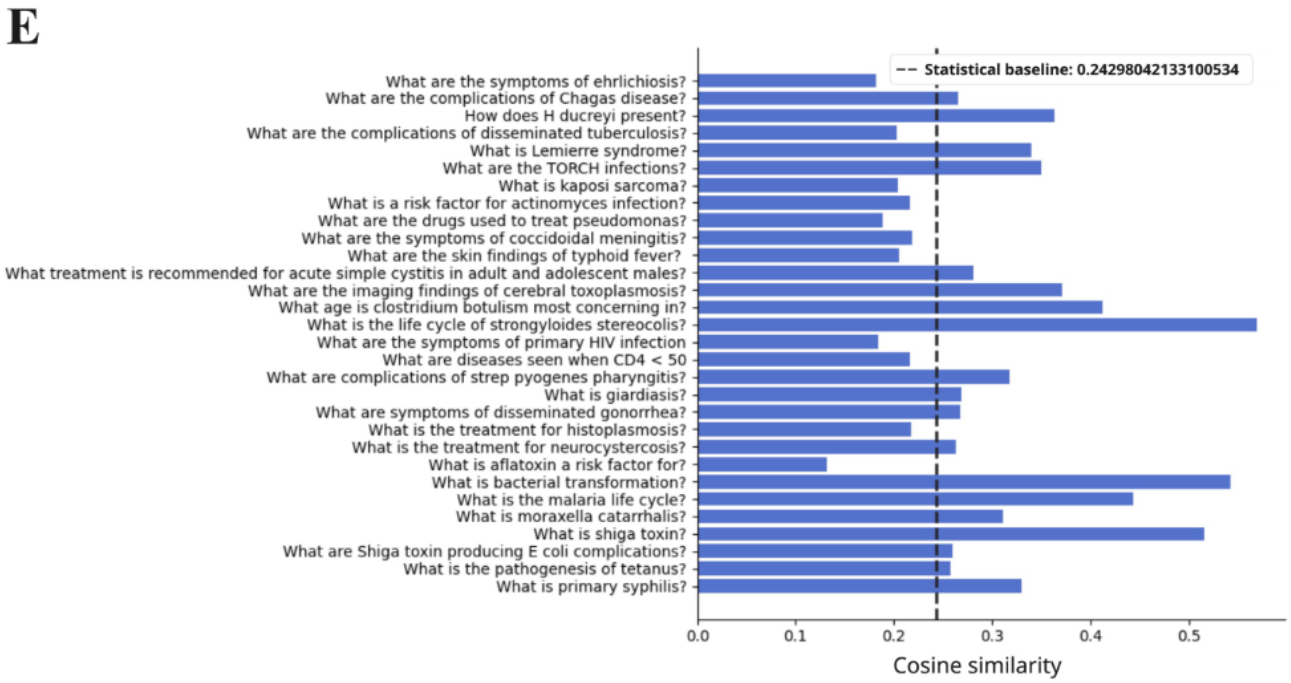
Figure 3. The TF-IDF cosine similarity and text networks for preclinical subject questions and aggregate ChatGPT responses for microbiology. The statistical baseline was determined from textual similarity to randomly generated text. lgM: immunoglobulin; MAC: membrane attack complex; PCA: principal component analysis; TF-IDF: term frequency–inverse document frequency.

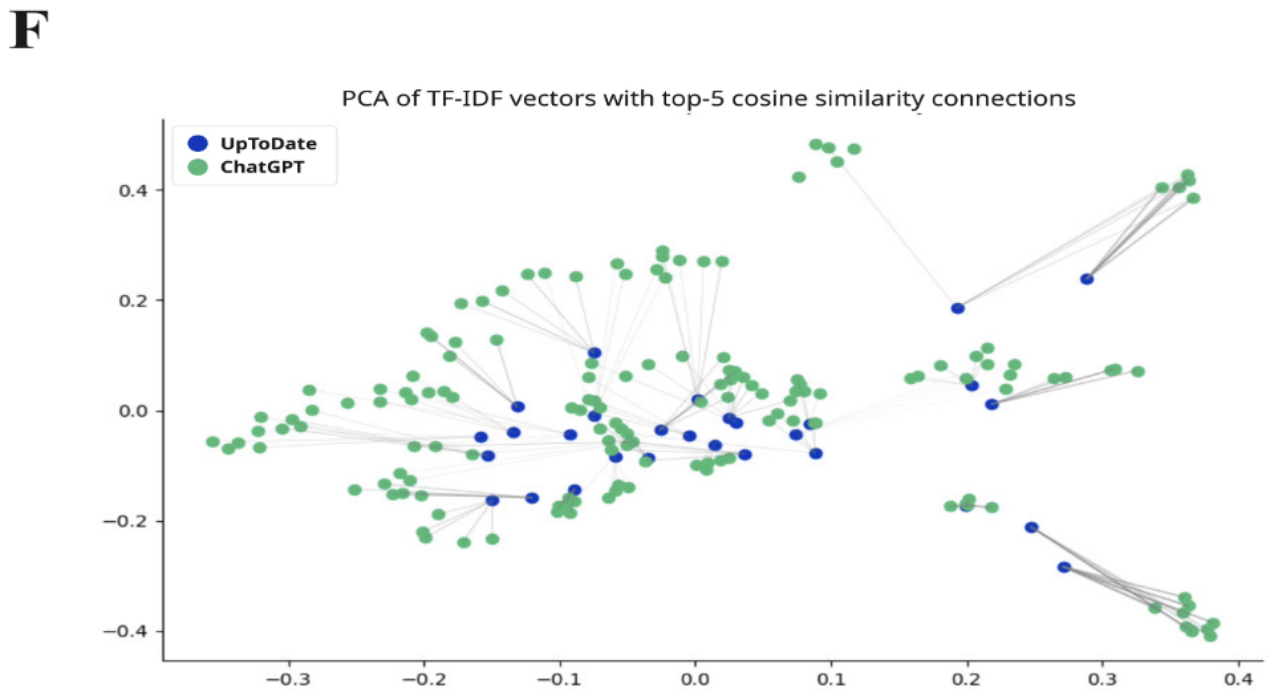




XSL•FO
RenderX

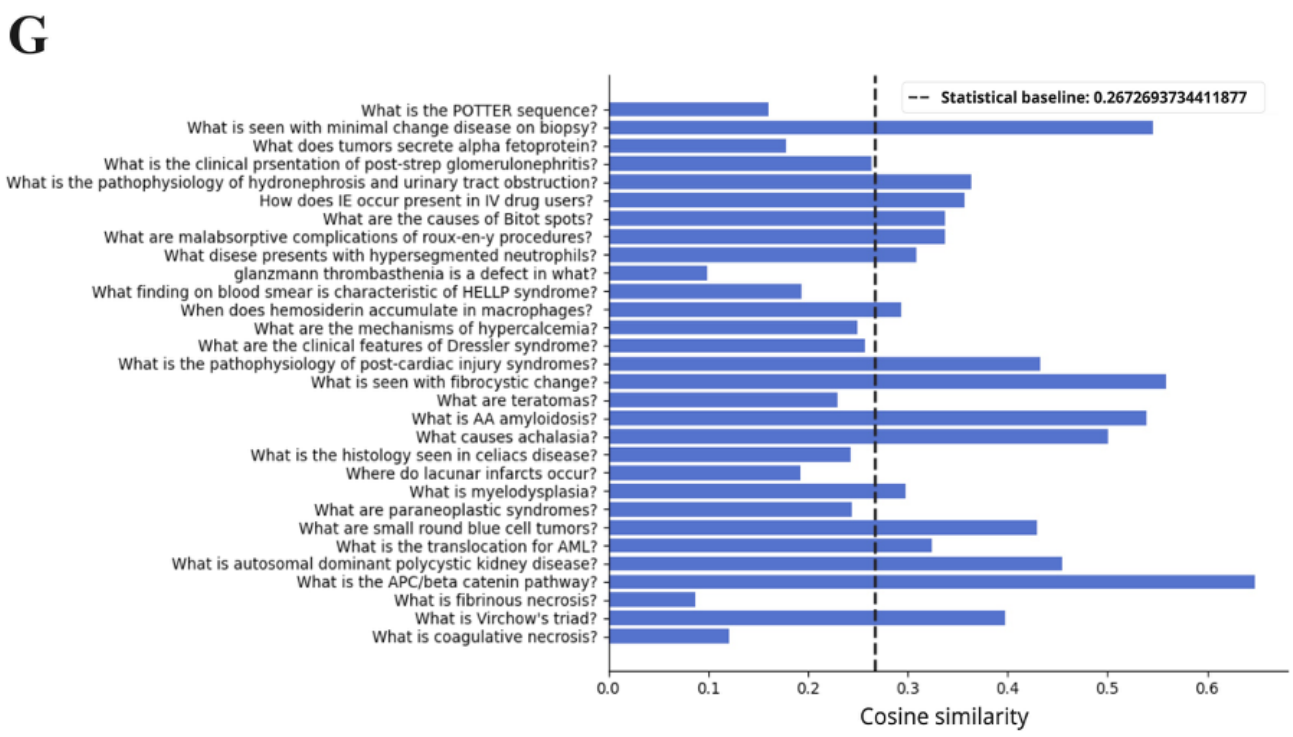
Figure 4. The TF-IDF cosine similarity and text networks for preclinical subject questions and aggregate ChatGPT responses for pathology. The statistical baseline was determined from textual similarity to randomly generated text. PCA: principal component analysis; TF-IDF: term frequency–inverse document frequency; TORCH: toxoplasmosis, other infections, rubella, cytomegalovirus, and herpes simplex virus.

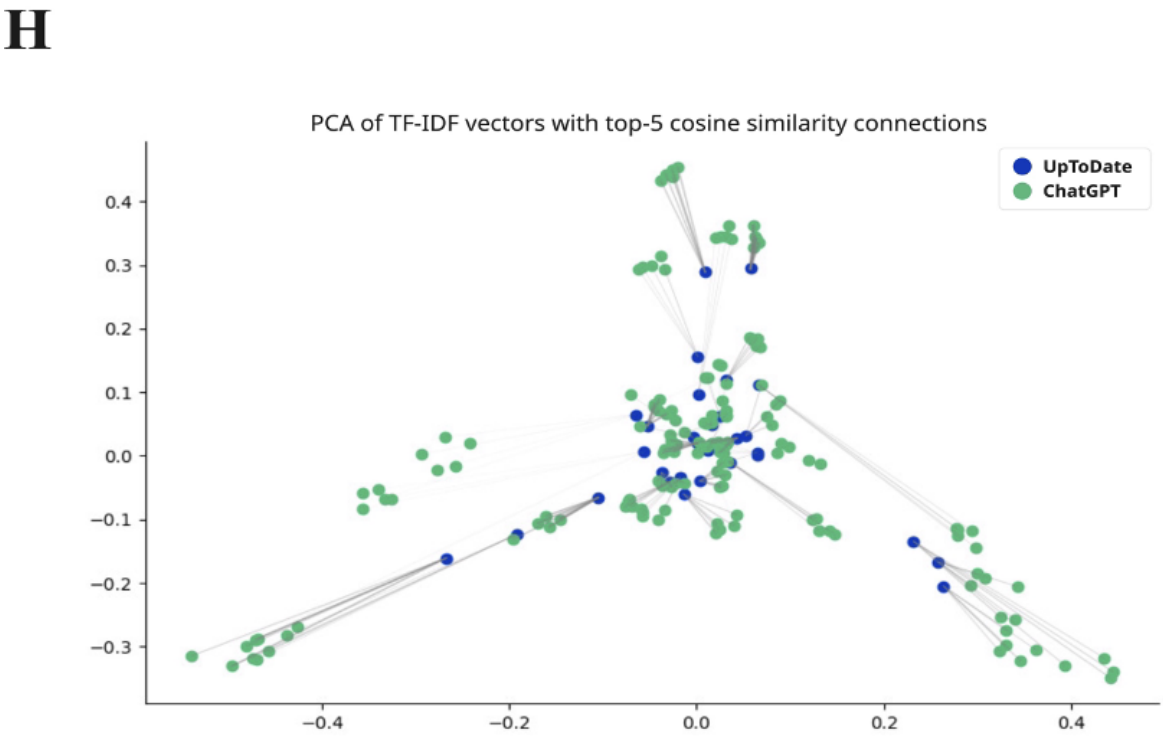




XSL•FO
RenderX

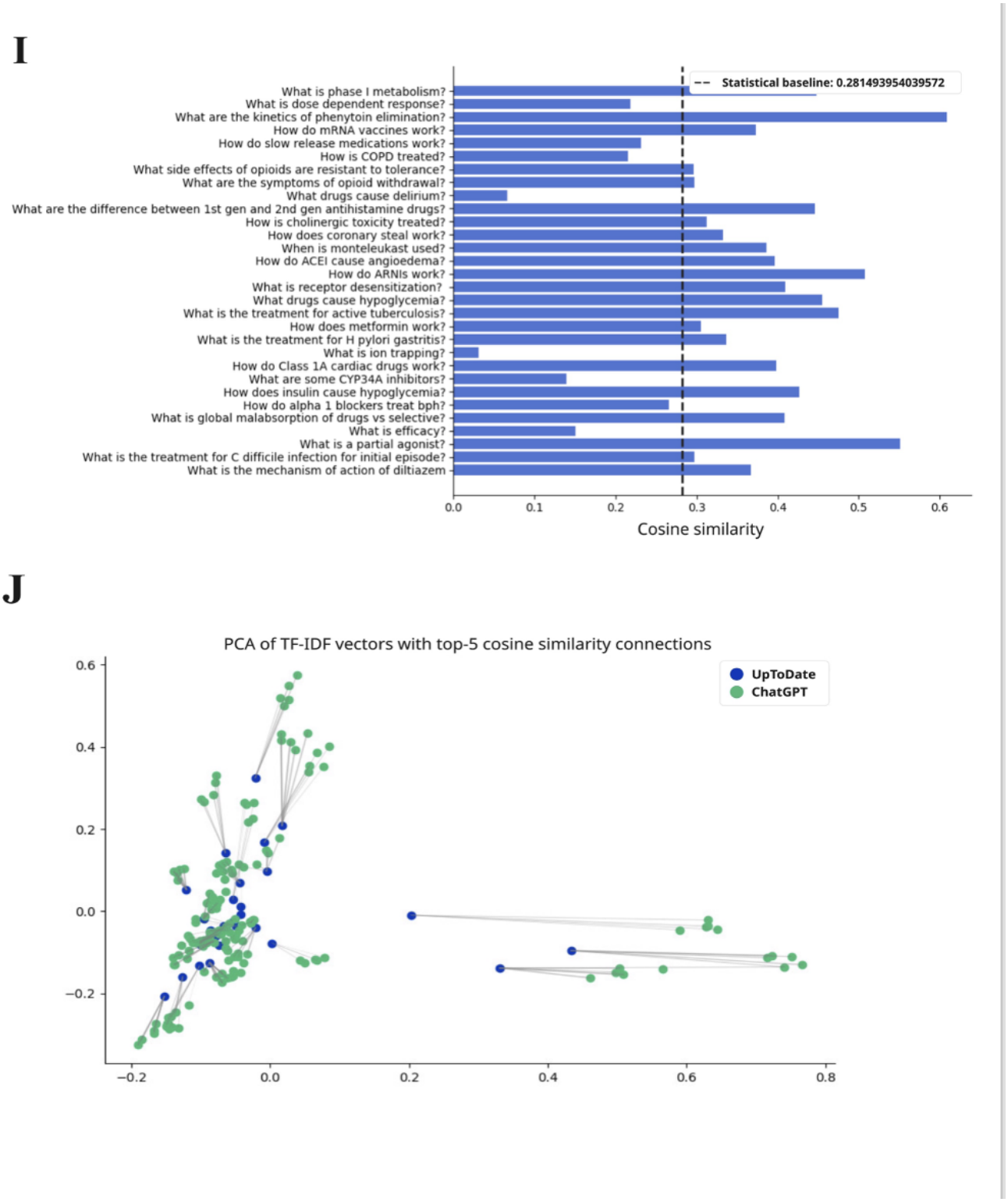
Figure 5. The TF-IDF cosine similarity and text networks for preclinical subject questions and aggregate ChatGPT responses for pharmacology. The statistical baseline was determined from textual similarity to randomly generated text. PCA: principal component analysis; TF-IDF: term frequency–inverse document frequency; IV: intravenous.





XSL•FO
RenderX

Figure 6. The overall cosine TF-IDF similarity between ChatGPT and UpToDate for the preclinical subjects of biochemistry, immunology, microbiology, pathology, and pharmacology is shown. Subjects are displayed in order from most to least similar. The statistical baseline was determined from textual similarity to randomly generated text. PCA: principal component analysis; TF-IDF: term frequency–inverse document frequency.



Performance was consistent across disciplines (Table 3). For example, 56.6% (17/30) of biochemistry questions and comparable proportions in immunology, microbiology, pharmacology, and pathology yielded statistically similar responses (P<.001).

XSL•FO
RenderX

Table . Frequency of questions (N=150) with similarity and nonsimilarity between ChatGPT and UpToDate for each preclinical subject.

Subject	Number of questions with similarity between ChatGPT and UpToDate, n (%)	Number of questions with no similarity between ChatGPT and UpToDate; n (%)
Biochemistry	17 (56.6)	13 (43.4)
Immunology	14 (46.7)	16 (53.3)
Microbiology	19 (63.3)	11 (36.7)
Pathology	17 (56.6)	11 (43.4)
Pharmacology	22 (73.3)	8 (26.7)

When mean similarity scores were analyzed, pharmacology achieved the highest concordance with UpToDate (TF-IDF mean 0.338, SD 0.134; *P*<.001), followed by pathology (mean 0.321, SD 0.142; *P*<.001), biochemistry (mean 0.296, SD 0.120; *P*<.001), microbiology (mean 0.297, SD 0.108; *P*<.001), and immunology (mean 0.275, SD 0.102; *P*<.001). These values consistently exceeded the similarity scores generated from randomized text baselines (range 0.243 - 0.281 across subjects; 0.148 for domain-level aggregation), confirming that ChatGPT's alignment with UpToDate reflected meaningful overlap beyond chance.

Discussion

Interpretation and Educational Implications

In the cross-sectional analysis, we found that ChatGPT demonstrated moderate but incomplete alignment with UpToDate across preclinical medical topics. Using TF-IDF cosine similarity, approximately 59% of ChatGPT responses were statistically similar to those generated by UpToDate, with consistent performance across disciplines. The highest concordance was observed in pharmacology. Importantly, similarity scores consistently exceeded randomized text baselines, indicating meaningful overlap beyond chance. However, more than one-third of responses diverged, underscoring important limitations in ChatGPT's reliability as a standalone educational resource.

With medical students increasingly integrating generative AI into their learning routines, understanding both the strengths and shortcomings of LLMs has become essential. Concerns persist regarding the accuracy and transparency of ChatGPT, particularly given documented instances of fabricated references and confidently stated inaccuracies [18]. By directly comparing ChatGPT with UpToDate, a widely regarded gold standard resource in medical education, this study provides novel insights into the credibility and potential educational role of LLMs in the preclinical curriculum.

This study's primary finding was that ChatGPT's responses were statistically similar to those of UpToDate in nearly 60% of cases. This suggests that ChatGPT is capable of producing outputs that meaningfully overlap with an established, peer-reviewed knowledge source. Importantly, however, more than one-third of responses diverged, underscoring that ChatGPT cannot yet be viewed as a direct replacement for authoritative resources. Instead, its greatest value may lie in functioning as a supplementary study aid, particularly when learners seek rapid explanations, iterative clarification, or simplified summaries. Unlike UpToDate's search-based retrieval of curated content, ChatGPT generates novel text in response to queries. This generative capacity could represent a shift in how students engage with knowledge; however, concerns about verifiability and the lack of time-stamped references remain major limitations to its educational utility.

ChatGPT demonstrated consistency across subject domains, with pharmacology achieving the highest similarity scores. This may reflect pharmacology's reliance on factual, definitional content (eg, mechanisms of action, drug classes), which generative AI models are better able to reproduce. In contrast, immunology and microbiology demonstrated greater divergence, likely due to their reliance on nuanced processes and context-dependent reasoning. These differences emphasize that AI may be better suited for fact-based knowledge than for subjects requiring higher-order integration or precise mechanistic detail. Incorporating domain experts in the design and refinement of AI models, similar to the physician involvement that underpins UpToDate, may help address these gaps and improve domain-specific accuracy [19].

From an educational perspective, these findings suggest that ChatGPT could serve as a readily accessible, low-barrier tool for students seeking supplemental explanations or conceptual reinforcement. However, educators must guide learners in understanding its limitations. Training students in AI literacy—the ability to critically appraise, cross-reference, and contextualize AI outputs—will be crucial to prevent overreliance on potentially inaccurate content. Future curricular integration of AI tools should emphasize transparency, verification against primary or trusted secondary sources, and responsible use in complementing (not replacing) established references.

Limitations

This study has several limitations. First, TF-IDF cosine similarity quantifies overlap in keywords but does not capture deeper semantic meaning, depth of explanation, or conceptual understanding. Although we benchmarked against randomized text to confirm nonrandom alignment, future studies could use more sophisticated natural language processing techniques (eg, word2vec, BERT) to account for contextual meaning; however, these approaches are not as widely validated for this type of analysis [20-22]. In addition, future studies could incorporate qualitative analyses, such as expert ratings, the DISCERN criteria, and *Journal of the American Medical Association* benchmarks to further delineate the utility of LLMs such as ChatGPT in preclinical medical education.

XSL•FO
RenderX

Second, this study was limited to GPT-4o mini, the free version most accessible to students. Performance may differ with more advanced subscription models. Third, stochasticity in ChatGPT responses means that some outputs contained tangential or extraneous material, which may have modestly influenced similarity scores.

Additionally, this study examined only 5 preclinical subject areas; inclusion of additional subject areas such as anatomy or physiology could further clarify ChatGPT's role across the broader preclinical curriculum. Finally, the rapid evolution of LLMs means that ChatGPT's performance may change over time. As newer versions are released, response accuracy, style, and alignment with reference sources like UpToDate may differ, potentially limiting the reproducibility of our findings in future evaluations.

Conclusion

In summary, ChatGPT demonstrated meaningful but incomplete similarity to UpToDate across a wide range of preclinical topics, with the strongest performance in pharmacology. These findings suggest that ChatGPT can serve as a useful adjunctive tool for medical students during the preclinical phase, provided its limitations are recognized and its use is coupled with authoritative references. As generative AI continues to evolve, collaboration between educators, clinicians, and developers will be essential to harness its potential while safeguarding accuracy, transparency, and educational value.

Acknowledgments

No artificial intelligence was used in the preparation of this manuscript.

Funding

No funding was obtained for this study.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Supplementary text detailing the methodology used in the study.
[DOCX File, 12 KB - mededu_v12i1e82885_app1.docx]

Multimedia Appendix 2
Precise Values for Figures 2-6
[DOCX File, 28 KB - mededu_v12i1e82885_app2.docx]

References

1. Sallam M. ChatGPT utility in healthcare education, research, and practice: systematic review on the promising perspectives and valid concerns. Healthcare (Basel) 2023 Mar 19;11(6):887. [doi: 10.3390/healthcare11060887] [Medline: 36981544]
2. Clusmann J, Kolbinger FR, Muti HS, et al. The future landscape of large language models in medicine. Commun Med (Lond) 2023 Oct 10;3(1):141. [doi: 10.1038/s43856-023-00370-1] [Medline: 37816837]
3. Liu J, Wang C, Liu S. Utility of ChatGPT in clinical practice. J Med Internet Res 2023 Jun 28;25:e48568. [doi: 10.2196/48568] [Medline: 37379067]
4. Dave T, Athaluri SA, Singh S. ChatGPT in medicine: an overview of its applications, advantages, limitations, future prospects, and ethical considerations. Front Artif Intell 2023;6:1169595. [doi: 10.3389/frai.2023.1169595] [Medline: 37215063]
5. Garg RK, Urs VL, Agarwal AA, Chaudhary SK, Paliwal V, Kar SK. Exploring the role of ChatGPT in patient care (diagnosis and treatment) and medical research: a systematic review. Health Promot Perspect 2023;13(3):183-191. [doi: 10.34172/hpp.2023.22] [Medline: 37808939]
6. Boscardin CK, Gin B, Golde PB, Hauer KE. ChatGPT and generative artificial intelligence for medical education: potential impact and opportunity. Acad Med 2024 Jan 1;99(1):22-27. [doi: 10.1097/ACM.0000000000005439] [Medline: 37651677]
7. Gilson A, Safranek CW, Huang T, et al. How does ChatGPT perform on the United States Medical Licensing Examination (USMLE)? The implications of large language models for medical education and knowledge assessment. JMIR Med Educ 2023 Feb 8;9:e45312. [doi: 10.2196/45312] [Medline: 36753318]
8. Hu JM, Liu FC, Chu CM, Chang YT. Health care trainees' and professionals' perceptions of ChatGPT in improving medical knowledge training: rapid survey study. J Med Internet Res 2023 Oct 18;25:e49385. [doi: 10.2196/49385] [Medline: 37851495]
9. Alkhaaldi SMI, Kassab CH, Dimassi Z, et al. Medical student experiences and perceptions of ChatGPT and artificial intelligence: cross-sectional study. JMIR Med Educ 2023 Dec 22;9:e51302. [doi: 10.2196/51302] [Medline: 38133911]

XSL•FO
RenderX

10. Tao W, Yang J, Qu X. Utilization of, perceptions on, and intention to use AI chatbots among medical students in China: national cross-sectional study. JMIR Med Educ 2024 Oct 28;10:e57132. [doi: 10.2196/57132] [Medline: 39466038]
11. Pan G, Ni J. A cross sectional investigation of ChatGPT-like large language models application among medical students in China. BMC Med Educ 2024 Aug 23;24(1):908. [doi: 10.1186/s12909-024-05871-8] [Medline: 39180023]
12. Xu AY, Piranio VS, Speakman S, et al. A pilot study of medical student opinions on large language models. Cureus 2024 Oct;16(10):e71946. [doi: 10.7759/cureus.71946] [Medline: 39564056]
13. Zhang JS, Yoon C, Williams DKA, Pinkas A. Exploring the usage of ChatGPT among medical students in the United States. J Med Educ Curric Dev 2024;11:23821205241264695. [doi: 10.1177/23821205241264695] [Medline: 39092290]
14. Giunti M, Garavaglia FG, Giuntini R, Sergioli G, Pinna S. ChatGPT as a prospective undergraduate and medical school student. PLoS One 2024;19(10):e0308157. [doi: 10.1371/journal.pone.0308157] [Medline: 39441779]
15. Prorok JC, Iserman EC, Wilczynski NL, Haynes RB. The quality, breadth, and timeliness of content updating vary substantially for 10 online medical texts: an analytic survey. J Clin Epidemiol 2012 Dec;65(12):1289-1295. [doi: 10.1016/j.jclinepi.2012.05.003] [Medline: 22974495]
16. Jalil M, Ding SR, Talpallikar NM, Bhatnagar I, Umashankar SV, Wensel JN. Readability of professional medical content on acute appendicitis: a comparative cross-sectional study of ChatGPT and UpToDate. Cureus 2025 Nov;17(11):e95898. [doi: 10.7759/cureus.95898] [Medline: 41333499]
17. Ahmadi SF, Faghankhani M, Javanbakht A, et al. A comparison of answer retrieval through four evidence-based textbooks (ACP PIER, Essential Evidence Plus, First Consult, and UpToDate): a randomized controlled trial. Med Teach 2011;33(9):724-730. [doi: 10.3109/0142159X.2010.531155] [Medline: 21854150]
18. Ariyaratne S, Iyengar KP, Nischal N, Chitti Babu N, Botchu R. A comparison of ChatGPT-generated articles with human-written articles. Skeletal Radiol 2023 Sep;52(9):1755-1758. [doi: 10.1007/s00256-023-04340-5] [Medline: 37059827]
19. Crossnohere NL, Elsaid M, Paskett J, Bose-Brill S, Bridges JFP. Guidelines for artificial intelligence in medicine: literature review and content analysis of frameworks. J Med Internet Res 2022 Aug 25;24(8):e36823. [doi: 10.2196/36823] [Medline: 36006692]
20. Lee J, Yoon W, Kim S, et al. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics 2020 Feb 15;36(4):1234-1240. [doi: 10.1093/bioinformatics/btz682] [Medline: 31501885]
21. Luo R, Sun L, Xia Y, et al. BioGPT: generative pre-trained transformer for biomedical text generation and mining. Brief Bioinform 2022 Nov 19;23(6):bbac409. [doi: 10.1093/bib/bbac409] [Medline: 36156661]
22. Minarro-Giménez JA, Marín-Alonso O, Samwald M. Exploring the application of deep learning techniques on medical text corpora. Stud Health Technol Inform 2014;205:584-588. [doi: 10.3233/978-1-61499-432-9-584] [Medline: 25160253]

Abbreviations

AI: artificial intelligence
LLM: large language model
TF-IDF: term frequency–inverse document frequency

Edited by R Pellegrino; submitted 26.Oct.2025; peer-reviewed by F Fukuzawa, J Reshi; revised version received 19.Dec.2025; accepted 05.Feb.2026; published 20.Mar.2026.

Please cite as:
Thiru SS, Aksu NE, Chiang M, Gallagher DO, Furlong M, Prevou ER, Khanna AJ
ChatGPT versus UpToDate in Preclinical Medical Education: Cross-Sectional Analysis Using Term Frequency–Inverse Document Frequency Cosine Similarity
JMIR Med Educ 2026;12:e82885
URL: https://mededu.jmir.org/2026/1/e82885
doi:10.2196/82885

© Shankar S Thiru, Nicholas E Aksu, Matthew Chiang, Daniel O Gallagher, Mary Furlong, Elizabeth R Prevou, Akhil Jay Khanna. Originally published in JMIR Medical Education (https://mededu.jmir.org), 20.Mar.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Education, is properly cited. The complete bibliographic information, a link to the original publication on https://mededu.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Comparing AI-Assisted Problem-Solving Ability With Internet Search Engine and e-Books in Medical Students With Variable Prior Subject Knowledge: Cross-Sectional Study

Ajiith Xavier*, MBBS, MD; Syed Shariq Naeem*, MBBS, MD, DM; Waseem Rizwi, MBBS, MD, PhD; Hiramani Rabha, MBBS, MD

Department of Pharmacology, Jawaharlal Nehru Medical College and Hospital, Aligarh Muslim University, Medical Road, Aligarh, Uttar Pradesh, India

*these authors contributed equally

Corresponding Author:
Syed Shariq Naeem, MBBS, MD, DM
Department of Pharmacology, Jawaharlal Nehru Medical College and Hospital, Aligarh Muslim University, Medical Road, Aligarh, Uttar Pradesh, India

Abstract

Background: Artificial intelligence (AI), particularly large language models (LLMs) such as ChatGPT (OpenAI), is rapidly influencing medical education. Its effectiveness for students with varying levels of prior knowledge remains underexplored.

Objective: This study aimed to evaluate the performance of medical students with and without formal pharmacology knowledge when using AI-LLM GPTs, internet search engines, e-books, or self-knowledge to solve multiple-choice questions (MCQs).

Methods: A cross-sectional study was conducted at a tertiary care teaching hospital with 100 medical students, divided into a "naive" group (n=50; no pharmacology training) and a "learned" group (n=50; completed pharmacology training). The study was started after approval from the Institutional Ethics Committee of Jawaharlal Nehru Medical College Hospital, Aligarh Muslim University (1018/IEC/23/8/23). Each participant answered 4 sets of 20 MCQs using self-knowledge, e-books, Google, or ChatGPT-4o. Scores were compared using analysis of covariance with self-knowledge scores as a covariate.

Results: Learned students significantly outperformed naive students across all methods (P<.001), with the largest effect size in the AI-LLM GPT set (partial η^2=0.328). For both groups, the performance hierarchy was AI-LLM GPT > internet search engine > self-knowledge ≈ e-books. Notably, the naive students who used AI scored higher (mean 13.24, SD 3.31) than the learned students who used Google (mean 12.14, SD 2.01; P=.01) or e-books (mean 10.22, SD 3.12; P<.001).

Conclusions: AI-LLM GPTs can significantly enhance problem-solving performance in MCQ-based assessments, particularly for students with limited prior knowledge, even allowing them to outperform knowledgeable peers using traditional digital resources. This underscores the potential of AI to transform learning support in medical education, although its impact on deep learning and critical thinking requires further investigation.

(JMIR Med Educ 2026;12:e81264) doi:10.2196/81264

KEYWORDS

artificial intelligence; AI; large language models; LLM; medical education; ChatGPT; cognitive performance; subject-naive learners

Introduction

Artificial intelligence (AI) is the ability of a digital computer or computer-controlled robot to perform tasks commonly associated with intelligent beings [1]. This includes the ability to reason, discover meaning, generalize, and learn from past experiences [2]. Since the development of digital computers in the 1940s, it has been demonstrated that computers can be programmed to perform extremely complex tasks, such as proving mathematical theorems or playing chess, with great proficiency [3]. AI involves the science and engineering of creating systems capable of performing tasks that require humanlike intelligence, including learning, judgment, and decision-making [4]. AI has successfully solved complex problems in various domains, including education. The application of AI in natural language processing has led to the creation of intelligent chatbots and virtual assistants capable of understanding and producing human language [5].

Popenici and Kerr [6] investigated the impact of AI systems on learning and teaching, highlighting potential conflicts between students and instructors, such as privacy concerns, changes in power structures, and excessive control. These studies have called for further research into the impact of AI systems on

XSL•FO
RenderX

learner-instructor interactions to identify gaps, issues, or barriers that prevent AI systems from achieving their intended potential [7].

In the modern era of education, technology has become an integral part of the learning process. The advent of advanced language models such as ChatGPT (OpenAI), along with the vast availability of information on platforms such as Google and e-books, enables educators to enhance students' educational experiences [7].

Medical education demands not only the acquisition of factual knowledge but also the development of critical thinking and decision-making skills [8]. However, the effectiveness of AI–large language model (LLM) GPT–driven tools in enhancing these competencies remains underexplored, particularly concerning the prior knowledge of learners [9]. Students with strong foundational knowledge can use AI tools to deepen their understanding and sharpen their reasoning. In contrast, those with less prior knowledge may use them to fill gaps in their learning or face some difficulties in comprehension [10].

This study used ChatGPT-4o as the representative AI-LLM GPT tool for 2 reasons. First, ChatGPT was widely accessible to students at the time of study and required minimal technical expertise. Second, it represents the current generation of general-purpose LLMs that students are likely to encounter in real-world educational settings. The most notable advantage is its breadth of training data, which expanded from 175 billion parameters in GPT-3.5 to 1 trillion in the GPT-4.0 model [11]. Bing's current AI model has 175 billion parameters, and Microsoft's Bard has 540 billion in comparison [12,13]. The domain-specific LLMs, enriched with pharmacology databases or specialized medical AI agents, may offer better performance characteristics, but widespread use of such tools is still limited due to paywall and validation issues.

The aim of this study is to compare the effectiveness of AI-LLM GPT tools (ChatGPT-4o), internet search engines (Google), e-books, and self-knowledge for answering pharmacology multiple-choice questions (MCQs) among medical students with different levels of prior pharmacology knowledge (naive vs learned groups). The study explores the assistance provided by AI-LLM GPT tools in the ability of problem solving using a cross-sectional design.

Methods

Ethical Considerations

This was a cross-sectional study conducted at a tertiary care teaching hospital. The study was started after approval from the Institutional Ethics Committee of Jawaharlal Nehru Medical College Hospital, Aligarh Muslim University (1018/IEC/23/8/23). All participants were adult MBBS students and provided written informed consent after receiving information about the study purpose and procedures. Data were collected anonymously using coded identifiers. No personal or identifiable information was recorded or published. Participants received no monetary or academic compensation.

Setting and Participants

Participants included 50 second-year medical students who had just joined the second year and 50 third-year students who had recently passed their second year (ie, N=100). The second-year students, designated as the "naive group," had no prior formal exposure to pharmacology coursework. The third-year students, designated as the "learned group," had completed a comprehensive 1-year pharmacology curriculum during their second year. This curriculum consisted of 80 hours of didactic lectures, 150 hours of interactive teaching, including practical laboratory sessions, case-based learning modules, self-directed learning modules, short group teaching, and regular assessments, including sessional and professional examinations. Students had completed this coursework approximately 2 - 3 months prior to participation in this study. The curriculum covered topics such as general pharmacology, autonomic pharmacology, cardiovascular drugs, and antimicrobials, aligned with the medical council syllabus for undergraduate medical education. All learned group participants had passed their second-year pharmacology examination with a minimum score of 50%.

Study Development

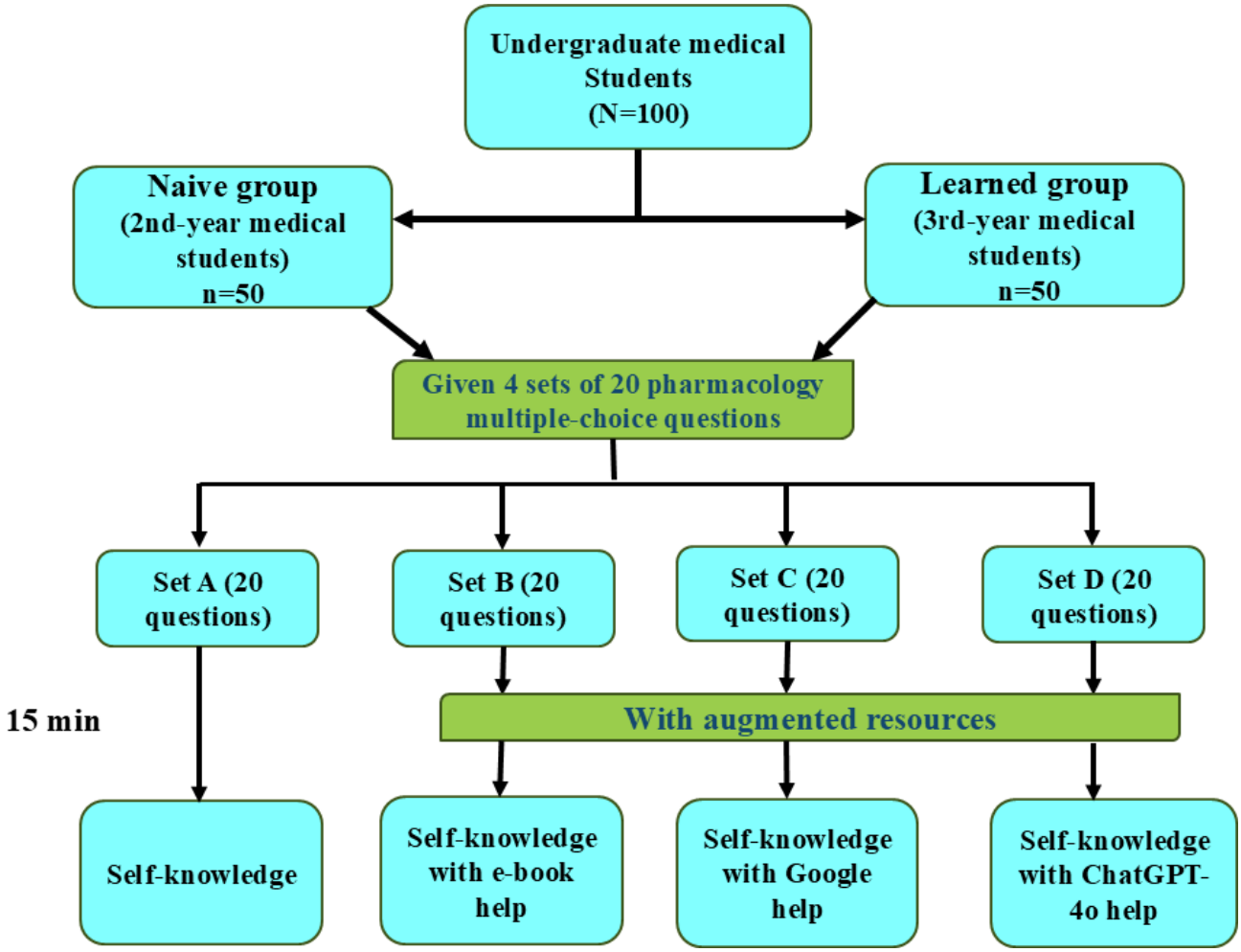
Prior to the assessment, all participants received a 15-minute standardized orientation session on using ChatGPT and were provided with an eBook of Pharmacology. This training covered basic prompt formulation and how to ask follow-up questions. Students accessed ChatGPT through computer workstations in a controlled examination hall environment to ensure standardized conditions. For the Google search set, students used desktop computers with standard Google Chrome or Mozilla Firefox browsers. For the e-book set, students were provided with PDF versions of a pharmacology textbook accessible on the same computers. The training session was conducted using a standardized presentation to maintain homogeneity of assistance tool exposure. The self-knowledge set required no external tools. All sessions were proctored to ensure compliance with assigned methods for each question set.

The 80 MCQs were divided into four parallel sets (A, B, C, and D), each containing 20 questions. To ensure equivalence across sets, questions were matched for content area coverage (eg, autonomic, cardiovascular, and chemotherapy) [1], question type (theory, calculation, or image-based) [2], and estimated difficulty level based on faculty assessment [3].

Students had 15 minutes to respond to each set of questions. They were instructed to respond to the questions in a manner that self-knowledge was used to answer questions in set A; e-books were used to answer questions in set B; internet search engine (Google) was used to answer questions in set C; and AI-LLM GPT (ChatGPT) was used to answer questions in set D. Each correct answer was given 1 mark, and no negative marking was given. The methodology flowchart of this study is shown in Figure 1.

XSL•FO
RenderX

Figure 1. Methodology flowchart.



Statistical Analysis

Data analysis was done using SPSS (version 23; IBM Corp) and R software package (R Foundation for Statistical Computing). The mean (the average number of correctly answered questions out of the 20 MCQs included in each augmentation set such as self-knowledge, e-book, internet search engine, and GPT-4o) between different groups was assessed using the independent test. Intergroup and intragroup comparisons were conducted using analysis of covariance (ANCOVA), with self-knowledge group (set A) being a covariate. A *P* value <.05 was considered to indicate statistical significance.

Results

The intergroup comparison between the learned and naive student groups was analyzed using ANCOVA. The estimated marginal means, partial η^2 values, and statistical significance for each group are presented in Table 1. The self-knowledge set was used as a covariate in the analysis to control for baseline differences in knowledge across the sets. This approach allowed for a more accurate assessment of the effects of the augmented resources methods (ChatGPT-4o, Google, and e-books) on student performance.

XSL•FO
RenderX

Table . Intergroup comparison between the learned and naive using analysis of covariance (N=100).

Augmentation resource and group		Score, mean (SD)	Estimated marginal mean	Partial η^2	*P* value
AI[a]-LLM[b] GPT (ChatGPT-4o)				0.328	<.001
	Naive	13.24 (3.31)	13.18		
	Learned	15.42 (1.89)	15.48		
Internet search engine (Google)				0.241	<.001
	Naive	10.82 (2.13)	10.76		
	Learned	12.14 (2.01)	12.20		
e-Books				0.195	<.001
	Naive	5.82 (2.77)	5.78		
	Learned	10.22 (3.12)	10.26		

[a]AI: artificial intelligence.

[b]LLM: large language model.

The analysis of estimated marginal means revealed that the learned students consistently outperformed the naive students across all sets. In the AI-LLM GPT (ChatGPT-4o) set, the learned students achieved an estimated marginal mean of 15.48 compared to 13.18 for the naive students, with a large effect size (partial η^2=0.328, *P*<.001). In the internet search engine (Google) set, the estimated marginal mean for the learned students was 12.20, whereas the naive students scored 10.76, with a moderate effect size (partial η^2=0.241; *P*<.001). The learned students in the e-book set had an estimated marginal mean of 10.26, whereas the naive students had 5.78, a moderate effect size (partial n^2=0.195; *P*<.001).

The post hoc analysis for the naive group shows that the AI-LLM GPT-assisted method resulted in the highest performance when compared to other methods like internet search engine–based learning, self-knowledge, and e-books methods. A significant difference was found in all pairwise comparisons (*P*<.001) except for e-books versus self-knowledge, which was not statistically significant (mean difference=−0.30; *P*=.91). The performance hierarchy observed was AI-LLM GPT versus self-knowledge (*P*<.001) > internet search engine versus self-knowledge (*P*<.001) > self-knowledge ≈ e-books (*P*=.91), showing the superior effectiveness of AI tools for naive learners.

The post hoc analysis for the learned group shows that the AI-LLM GPT–assisted learning method resulted in the highest performance when compared to other methods, whereas self-knowledge and e-books methods showed the lowest and statistically similar outcomes (mean difference=−0.72, *P*=.33). Significant differences were observed between Google and self-knowledge sets (*P*=.04), ChatGPT-4o and self-knowledge sets (*P*<.001), Google and e-books sets (*P*<.001), ChatGPT-4o and e-books (*P*<.001), and ChatGPT-4o and Google (*P*<.001), indicating a superior performance across these comparisons. The overall performance hierarchy observed was: AI-LLM GPT > internet search engine > self-knowledge ≈ e-books, indicating the superior effectiveness of AI-LLM GPT in enhancing learning outcomes, even for participants with prior knowledge (Multimedia Appendix 1).

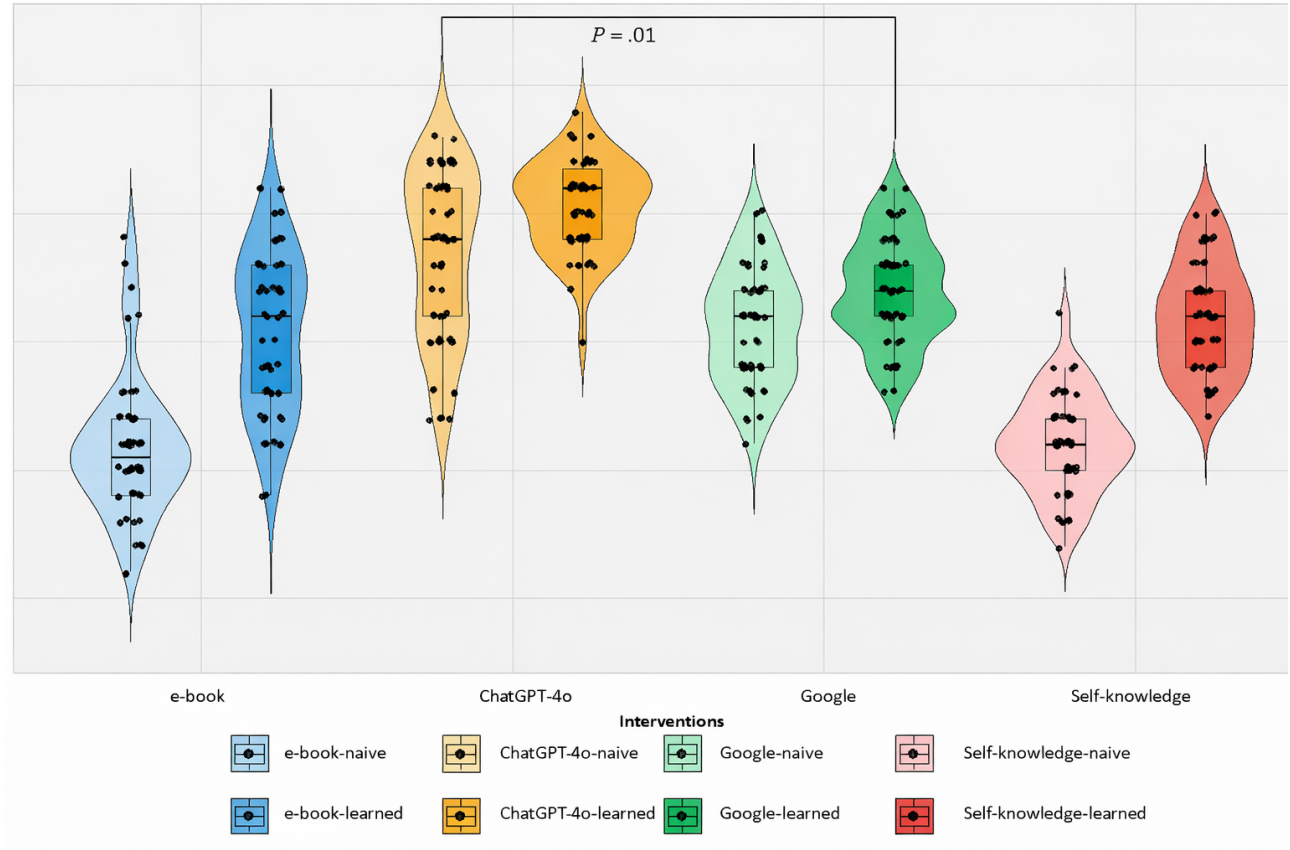
The mean comparison between the sets, as presented in Figure 2, highlights that learned students consistently outperformed naive students across all categories, as expected. The AI-LLM set achieved the highest performance, with a mean score of 15.3 (SD 1.776). Conversely, the lowest mean was observed in the e-books set, in which learned students scored 10.3 (SD 3.1) and naive students scored 5.7 (SD 2.7).

Interestingly, the self-knowledge sets performed better than the e-books sets, with naive group students achieving a mean score of 6.0 (SD 1.9) and learned group students scoring a mean of 10.96 (SD 2.2). This may be due to the fact that students in the e-book sets spent sufficient time searching for relevant chapters instead of focusing on answering the questions. This led to some questions being left unanswered that likely contributed to their lower overall performance.

While ANCOVA analysis confirmed that the learned students consistently outperformed the naive students when comparing the same learning method (as shown in Table 1), an important cross-method comparison emerged: the naive students who used ChatGPT-4o achieved higher scores (mean 13.2, SD 3.3) than the learned students who used self-knowledge (mean 10.96, SD 2.5), e-books (mean 10.22, SD 3.12), or Google (mean 12.14, SD 2.01). This demonstrates that the advantage conferred by using ChatGPT-4o was sufficient to overcome the knowledge gap between the naive and learned students when the latter were restricted to conventional resources.

XSL•FO
RenderX

Figure 2. Box-violin plot showing the mean scores in the naive and learned groups.



Discussion

Principal Findings

AI is changing the face of medical education, as it suggests some new solutions to the old forms of teaching, increases the educational experience, and improves the results of students [14,15]. AI-based solutions, including LLMs, such as ChatGPT, virtual assistants, and AI-based simulation technologies, can change how medical students learn and apply their knowledge [16,17].

This study found that learned students consistently outperformed naive students across all learning methods. However, the naive students who used ChatGPT-4o achieved higher scores than the learned students who used conventional resources (ie, self-knowledge, e-books, or Google), suggesting that AI-LLM tools may help bridge knowledge gaps more effectively than traditional learning resources.

Our study demonstrates that AI-powered LLMs such as ChatGPT can notably improve students' performance on pharmacology MCQs, particularly for those without prior subject knowledge. Although our findings show that students achieved higher MCQ test scores when using ChatGPT-4o, we cannot conclude that this represents a genuine improvement in cognitive performance, critical thinking, or clinical decision-making skills. Higher scores may reflect the AI's ability to rapidly retrieve and present relevant information rather than deeper student understanding. The findings indicated that learned students have outperformed naive students in all 4 sets, and AI-LLM GPT recorded the best performance. The greatest effect size was observed in the AI group (partial η^2=0.328), which indicates the tremendous influence of AI-LLM in ensuring the students improve their understanding and problem-solving skills. These results are consistent with those of previous research in that AI tools may enhance critical thinking and support learning under the condition that they are properly used by a student who already has a good knowledge grounding [18].

As opposed to other methods, the e-books set exhibited the lowest performance. This can be explained by the fact that it takes a lot of time to navigate and search relevant information in the e-books. This reveals the difficulties of self-learning with the help of conventional digital sources, especially in the case of naive students who are probably not able to manage their time efficiently. In contrast, the self-knowledge set outperformed the e-books set, which means that dependence on inherent knowledge, although restricted to a certain extent, can be more helpful than the effective utilization of external means.

Although our quantitative data demonstrate that ChatGPT-4o enhanced student performance, particularly for naive learners, the mechanisms underlying this effectiveness warrant deeper exploration. There may be several reasons for AI-LLM GPT (ChatGPT-4o) to outperform other methods of assistance.

First, ChatGPT-4o provides immediate, targeted responses without requiring students to navigate complex search results or textbook indices. Informal feedback from participants suggested that students who used Google often struggled to identify authoritative sources among numerous search results, whereas those who used e-books spent considerable time locating relevant chapters. In contrast, ChatGPT-4o delivered

XSL•FO
RenderX

direct answers to questions, reducing the cognitive load related to information seeking.

Second, ChatGPT-4o's conversational interface may facilitate iterative learning. Students could ask follow-up questions to clarify concepts, request examples, or seek a specific explanation for a specific MCQ—a dynamic interaction not possible with static e-books and more efficient than reformulating multiple Google searches.

Third, the structured, synthesized format of ChatGPT-4o responses may be particularly beneficial for naive learners who lack the prior knowledge framework to evaluate and integrate fragmented information from multiple sources. ChatGPT-4o essentially predigests information, whereas Google and e-books require students to perform this synthesis themselves—a task that may be especially challenging without foundational knowledge.

However, these remain as hypotheses. Future research should use mixed methods approaches, including think-aloud protocols during tool use, posttask interviews exploring student decision-making processes, and screen recordings analyzing search strategies and information evaluation patterns. Such qualitative data would provide richer insights into how students with different knowledge levels interact with various resources and why certain tools prove more effective. The study of recognizing the pattern of wrong responses in the AI-LLM GPTs output may also generate an "array of errors" for further training and improving LLMs.

These findings have substantial implications for both local and global medical education practices. In our institutional context, where students face high patient loads and limited access to senior clinicians for immediate consultation, AI-LLM GPTs could serve as readily available reference tools to support clinical decision-making during training. However, integration must be carefully structured to enhance rather than replace foundational learning.

Globally, these results are particularly relevant for resource-limited settings where access to comprehensive textbooks, updated references, and expert faculty may be constrained. AI-LLM GPTs could help democratize access to medical knowledge, potentially reducing disparities between well-resourced and under-resourced educational institutions. However, this assumes reliable internet connectivity and technological infrastructure, which remain barriers in many settings.

From a pedagogical perspective, our findings suggest a paradigm shift may be necessary in how we structure medical curricula. Rather than focusing exclusively on memorization of facts—information that AI can rapidly retrieve—educational programs should prioritize teaching students how to (1) formulate effective questions and search strategies, (2) critically evaluate AI-generated responses for accuracy and clinical appropriateness, (3) integrate AI-provided information with clinical context and patient-specific factors, and (4) develop metacognitive skills to recognize the limitations of both their own knowledge and AI tools.

Importantly, although AI-LLMs demonstrated effectiveness for MCQ performance, medical practice requires competencies beyond factual knowledge, including physical examination skills, procedural expertise, empathic patient communication, ethical reasoning under uncertainty, and team-based care coordination. Educational programs must ensure that AI integration enhances rather than diminishes these essential human dimensions of medical practice.

An important limitation is our focus on general-purpose ChatGPT-4o rather than specialized medical or pharmacology-specific AI agents. Domain-enriched LLMs that integrate pharmacology textbooks or medical databases may demonstrate different performance characteristics and potentially offer more accurate, context-specific responses. However, our choice reflects the current reality that most students have access to general LLMs such as ChatGPT-4o rather than specialized medical AI tools and makes the study more generalizable than using domain-enhanced LLMs. Second, the cross-sectional, single-day design prevents assessment of long-term learning outcomes, knowledge retention, or the development of clinical reasoning skills. Similarly, the sequential exposure to different resources (sets A through D) may have introduced order effects or learning transfer between sets. To minimize the learning transfer, the questions have been chosen from diverse topics and moderated to remove any repetition. However, the order effect cannot be removed completely in these types of cross-sectional–sequential studies.

AI will be able to deliver personalized learning and adjust to the needs of students. This individualization helps the learners make their progress through their priorities, repeating difficult concepts or working through material as they become familiar with each subject [19,20]. The AI-powered platforms can monitor student performance in real-time and can track the areas where the students might need more education or practice [21]. Such benefits help achieve the speed and the depth of knowledge acquisition with a higher level of efficiency and more targeted learning [22,23].

Conclusions

To conclude, this study shows that the learned group outperformed the naive group in all sets, including AI-LLM GPT, internet search engine, e-books, and self-knowledge in problem solving. Ironically, the AI-LLM GPT naive group outperformed the problem-solving skills of even the learned group augmented with an internet search engine, showcasing the disruptive potential of AI-LLM in medical education. Further longitudinal studies examining knowledge retention, clinical reasoning development, and the ability to solve novel problems without AI assistance are needed to determine whether AI tools genuinely enhance cognitive capabilities or primarily serve as effective reference tools.

XSL•FO
RenderX

Acknowledgments

The authors would like to express their gratitude to all the students who took part in this research. They would also like to thank Dr Binola for their constant motivation throughout the study. All authors declared that they had insufficient funding to support open access publication of this manuscript, including from affiliated organizations or institutions, funding agencies, or other organizations. JMIR Publications provided article processing fee (APF) support for the publication of this paper.

Data Availability

The anonymized datasets generated and analyzed during this study are not publicly available due to institutional restrictions and the need to protect participant confidentiality; however, the data are available from the corresponding author on reasonable request.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Post hoc results of the naive and learned groups.
[DOCX File, 6001 KB - mededu_v12i1e81264_app1.docx]

References

1. Arawi T, El Bachour J, El Khansa T. The fourth industrial revolution: its impact on artificial intelligence and medicine in developing countries. Asian Bioeth Rev 2024 May 25;16(3):513-526. [doi: 10.1007/s41649-024-00284-7] [Medline: 39022373]
2. Sharma S. Benefits or concerns of AI: a multistakeholder responsibility. Futures 2024 Mar;157:103328. [doi: 10.1016/j.futures.2024.103328]
3. Chaudhry MA, Kazim E. Artificial Intelligence in Education (AIEd): a high-level academic and industry note 2021. AI Ethics 2022;2(1):157-165. [doi: 10.1007/s43681-021-00074-z] [Medline: 34790953]
4. Xu LD, Lu Y, Li L. Embedding blockchain technology into IoT for security: a survey. IEEE Internet Things J 2021 Jul 1;8(13):10452-10473. [doi: 10.1109/JIOT.2021.3060508]
5. Supriyono, Wibawa AP, Suyono, Kurniawan F. Advancements in natural language processing: implications, challenges, and future directions. Telemat Inform Rep 2024 Dec;16:100173. [doi: 10.1016/j.teler.2024.100173]
6. Popenici SA, Kerr S. Exploring the impact of artificial intelligence on teaching and learning in higher education. Res Pract Technol Enhanc Learn 2017;12(1):22. [doi: 10.1186/s41039-017-0062-8] [Medline: 30595727]
7. Ghory S, Ghafory H. The impact of modern technology in the teaching and learning process. Int J Innov Res Sci Stud 2021 Jun 18;4(3):219-227. [doi: 10.53894/ijirss.v4i3.73]
8. Buja LM. Medical education today: all that glitters is not gold. BMC Med Educ 2019 Apr 16;19(1):110. [doi: 10.1186/s12909-019-1535-9] [Medline: 30991988]
9. Salih S, Husain O, Hamdan M, Abdelsalam S, Elshafie H, Motwakel A. Transforming education with AI: a systematic review of ChatGPT's role in learning, academic practices, and institutional adoption. Res Eng 2025 Mar;25:103837. [doi: 10.1016/j.rineng.2024.103837]
10. Lawasi MC, Rohman VA, Shoreamanis M. The use of AI in improving student's critical thinking skills. Proc Ser Soc Sci Humanit 2024 Sep;18:366-370. [doi: 10.30595/pssh.v18i.1279]
11. Martindale J, Bizzaco M. GPT-4 vs GPT-35: how much difference is there? Digital Trends. 2023 Apr 4. URL: https://www.digitaltrends.com/computing/gpt-4-vs-gpt-35 [accessed 2023-07-29]
12. Deutscher M. Microsoft researchers reveal neural network with 135B parameters. SiliconANGLE. 2021 Aug 4. URL: https://siliconangle.com/2021/08/04/microsoft-researchers-reveal-neural-network-135b-parameters [accessed 2023-07-29]
13. Almekinders S. Google Bard PaLM update: 540 billion parameters take fight to ChatGPT. Techzine. 2023 Apr 3. URL: https://www.techzine.eu/news/applications/104043/google-bard-palm-update-540-billion-parameters-take-fight-to-chatgpt/ [accessed 2025-12-05]
14. Mir MM, Mir GM, Raina NT, et al. Application of artificial intelligence in medical education: current scenario and future perspectives. J Adv Med Educ Prof 2023 Jul;11(3):133-140. [doi: 10.30476/JAMP.2023.98655.1803] [Medline: 37469385]
15. Nagi F, Salih R, Alzubaidi M, et al. Applications of artificial intelligence (AI) in medical education: a scoping review. Stud Health Technol Inform 2023 Jun 29;305:648-651. [doi: 10.3233/SHTI230581] [Medline: 37387115]
16. Mattas PS. ChatGPT: a study of AI language processing and its implications. Int J Res Publ Rev 2023;04(2):435-440. [doi: 10.55248/gengpi.2023.4218]
17. Alawida M, Mejri S, Mehmood A, Chikhaoui B, Isaac Abiodun O. A comprehensive study of ChatGPT: advancements, limitations, and ethical considerations in natural language processing and cybersecurity. Information 2023;14(8):462. [doi: 10.3390/info14080462]

XSL•FO
RenderX

18. Lijie H, Mat Yusoff S, Mohamad Marzaini AF. Influence of AI-driven educational tools on critical thinking dispositions among university students in Malaysia: a study of key factors and correlations. Educ Inf Technol 2025;30:8029-8053. [doi: 10.1007/s10639-024-13150-8]
19. Narayanan S, Ramakrishnan R, Durairaj E, Das A. Artificial intelligence revolutionizing the field of medical education. Cureus 2023 Nov 28;15(11):e49604. [doi: 10.7759/cureus.49604] [Medline: 38161821]
20. Jong M. Transforming medical education in the AI era: balancing technological expertise with humanistic care in tomorrow's doctors. Ann Acad Med Singap 2024 Dec 12;53(12):754-757. [doi: 10.47102/annals-acadmedsg.2024242] [Medline: 39748174]
21. Gligorea I, Cioca M, Oancea R, Gorski AT, Gorski H, Tudorache P. Adaptive learning using artificial intelligence in e-learning: a literature review. Educ Sci 2023;13(12):1216. [doi: 10.3390/educsci13121216]
22. Ahn HY. AI-powered e-learning for lifelong learners: impact on performance and knowledge application. Sustainability 2024;16(20):9066. [doi: 10.3390/su16209066]
23. Alqahtani T, Badreldin HA, Alrashed M, et al. The emergent role of artificial intelligence, natural learning processing, and large language models in higher education and research. Res Social Adm Pharm 2023 Aug;19(8):1236-1242. [doi: 10.1016/j.sapharm.2023.05.016] [Medline: 37321925]

Abbreviations

AI: artificial intelligence
ANCOVA: analysis of covariance
LLM: large language model
MCQ: multiple-choice question

Edited by J Gentges; submitted 25.Jul.2025; peer-reviewed by D Chartash, JM Santos-Gago, K Sakaguchi; revised version received 04.Nov.2025; accepted 06.Nov.2025; published 06.Jan.2026.

Please cite as:
Xavier A, Naeem SS, Rizwi W, Rabha H
Comparing AI-Assisted Problem-Solving Ability With Internet Search Engine and e-Books in Medical Students With Variable Prior Subject Knowledge: Cross-Sectional Study
JMIR Med Educ 2026;12:e81264
URL: https://mededu.jmir.org/2026/1/e81264
doi:10.2196/81264

© Ajiith Xavier, Syed Shariq Naeem, Waseem Rizwi, Hiramani Rabha. Originally published in JMIR Medical Education (https://mededu.jmir.org), 6.Jan.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Education, is properly cited. The complete bibliographic information, a link to the original publication on https://mededu.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

AI Literacy Among Chinese Medical Students: Cross-Sectional Examination of Individual and Environmental Factors

Chunqing Li[1], PhD; Sian Hsiang-Te Tsuei[2], MD, CCFP, PhD; Hongbin Wu[3], Prof Dr

[1]Education Management Information Center, Ministry of Education of the People's Republic of China, Beijing, China

[2]Faculty of Medicine, University of British Columbia, Vancouver, BC, Canada

[3]Institute of Medical Education, Peking University, No 5 Yiheyuan Rd, Beijing, China

Corresponding Author:
Hongbin Wu, Prof Dr
Institute of Medical Education, Peking University, No 5 Yiheyuan Rd, Beijing, China

Abstract

Background: Artificial intelligence (AI) literacy is increasingly essential for medical students. However, without systematic characterization of the relevant components, designing targeted medical education interventions may be challenging.

Objective: This study aimed to systematically describe the levels of and factors associated with multidimensional AI literacy among Chinese medical students.

Methods: A cross-sectional, descriptive analysis was conducted using data from a nationwide survey of Chinese medical students (N=80,335) across 109 medical schools in 2024. AI literacy was assessed with a multidimensional instrument comprising three domains: knowledge, evaluating students' self-reported proficiency in core areas of medical AI applications; attitude, reflecting their self-perceived views on using AI for teaching and learning; and behavior, capturing the self-perceived usage frequency and application patterns. Multivariate linear regression was applied to examine the associations between individual factors (ie, demographic characteristics, family background, and enrollment motivation) and environmental factors (ie, educational phase, type of education program, and tier of education program) and AI literacy.

Results: Respondents showed moderate to high levels of AI knowledge (mean 76.0, SD 26.9), followed by moderate AI attitude scores (mean 71.6, SD, 24.4). In contrast, AI behavior scores were much lower (mean 32.5, SD, 28.5), indicating little usage of AI tools. Of the individual factors, male students reported higher levels of AI attitude and behavior; both intrinsic and extrinsic motivation were positively associated with all three dimensions; advantaged family background was positively related to AI attitude and behavior, but not knowledge. Among the environmental factors, attending the prestigious Double First-Class universities was positively associated with higher AI usage. Enrollment in long-track medical education programs was associated with higher AI attitude and behavior, while being in the clinical phase was negatively associated with both AI knowledge and behavior. Environmental factors moderated the associations between individual characteristics and AI literacy, potentially attenuating disparities.

Conclusions: Medical students reported moderate to high AI knowledge, moderate AI favorability, and low AI use. Individual characteristics and environmental factors were significantly associated with AI literacy, and environmental factors moderated the associations. The moderate AI literacy overall highlights the need for AI-related medical education, ideally with practical use and nuanced by socioeconomic factors.

(JMIR Med Educ 2026;12:e80604) doi:10.2196/80604

KEYWORDS

medical education; artificial intelligence; AI Literacy; medical student; multidimensional constructs; China

Introduction

Artificial intelligence's (AI's) rapid advancement in diagnostic support, treatment decision-making, personalized care, and health system management is quickly reshaping health care service delivery [1-4]. To prepare medical trainees in using these tools [5], medical training programs need to efficiently leverage their limited resources to set up the relevant training approaches and programs to meet the most important needs. However, instead of identifying such gaps, research has mostly focused on applying AI in medical education across tasks, including, for example, admission, research, and evaluation [6-17]. The developments of AI-targeted curricula and competencies have so far advanced without high-quality evidence of trainees' needs [18-22], highlighting the questionable appropriateness of such proposals.

Among the various AI skills and competencies, clarifying trainees' AI literacy appears is especially important. AI literacy

XSL•FO
RenderX

refers to a set of competencies that enable individuals to critically evaluate AI technologies, communicate and collaborate effectively with AI, and use AI as a tool across various contexts [23]. While various frameworks for AI literacy exist, there is growing consensus that it comprises cognitive knowledge, attitudes, and practical behaviors toward AI tools [5,24-26].

Building on this consensus, we defined AI literacy as a multidimensional construct encompassing knowledge, attitude, and behavior, integrating both classical and contemporary theoretical models. Specifically, this tripartite structure aligns with the ABC model of attitudes in social psychology (Affect–Behavior–Cognition), which conceptualizes human responses as comprising cognitive, affective, and behavioral components [27-29].

Current evidence on AI literacy in medical trainees has been hindered by theoretical and methodological deficiencies. First, the theoretical weakness stems from the lack of recognition that AI literacy has multiple dimensions. Most have focused on AI knowledge [30-33] or attitudes toward AI applications [34-38]. This misses the important dimension of AI behaviors [5,24,25,39,40], which account for the actual usage patterns of AI tools. Second, methodologically, the few studies that captured all three dimensions suffered from capturing only a small number of medical schools and sample size [41,42]. This limits their generalizability. Third, the studies often fail to examine the factors associated with subdimensions of AI literacy, which limits the ability to design targeted interventions within a limited budget. Existing studies have often focused mainly on individual-level variables, overlooking environmental factors associated with the digital divide [41,43-45].

We addressed these gaps by drawing on the comprehensive quantitative database of Chinese medical students, the 2024 China Medical Student Survey (CMSS). We report the distribution of AI literacy across cognitive, attitudinal, and behavioral dimensions among 80,355 Chinese medical students and their associated individual and environmental factors. This large sample size across all three dimensions of AI literacy provides a more meaningful representation of the potential educational needs. We further analyzed the potential factors associated with AI literacy, enabling medical educators to design targeted interventions.

Methods

Data Source and Study Sample

Since 2019, China's National Center for Health Professions Education Development has conducted an annual nationwide survey of medical students in China known as the CMSS. It aims to support medical education development by collecting comprehensive information across the entire educational journey, including students' demographic backgrounds, preadmission experiences, academic training, and postgraduation plans [46]. The survey takes place between May and July each year. Additional details on the context of medical education in China and the CMSS are provided in Items 1 and 2 in Multimedia Appendix 1. All data analyses adhered to the (STROBE) Strengthening the Reporting of Observational Studies in Epidemiology guidelines for cross-sectional studies.

This study used data from the 2024 CMSS, which included participants from 109 medical schools across 28 provinces in China. The survey achieved a 73.9% response rate (80,355/108,710 surveyed) and captured 109 of 202 medical schools (109/202, 53.9%). The geographical distribution of participating institutions roughly reflected the national pattern. Comparing the dataset and the national distribution, respectively, 49.5% (54/109) and 45.0% (91/202) of the schools were from the eastern region, 28.4% (31/109) and 30.7 % (62/202) from the central region, and 22.0% (24/109) and 24.3% (49/202) from the western region. As for the tiers of institutions, the data included universities' designation as 'Double First-Class' (DFC) institutions. These institutions garner priority investment in China's bid to establish world-leading educational institutions [47] (see Item 3 in Multimedia Appendix 1 for more details). Further, the proportion of DFC universities was similar in the dataset and nation was (25.7%, 28/109 vs 23.3%, 47/202), respectively.

Ethical Considerations

The project was approved by the Peking University Institutional Review Board (Beijing, China; approval Number: IRB00001052- 20069). Participation was voluntary, and completion and submission of the questionnaire were considered to constitute informed consent. To protect participant privacy and confidentiality, no personally identifiable information was collected, and all responses were analyzed in an anonymized form. The data were stored securely and were accessible only to the research team. Participants did not receive any financial or other compensation for their participation.

Measurement of AI Literacy

AI literacy is a multidimensional construct comprising cognitive, attitudinal, and behavioral dimensions [24,25,40]. In this study, we refer to the cognitive dimension as "AI knowledge" to improve clarity and align with common usage in educational research. AI knowledge was assessed through four items evaluating students' self-perceived proficiency in core areas of medical AI: basic concepts of AI, machine learning tools, multimodal medical data analysis, and ethics issues related to AI. Responses were rated on a three-point scale coded as 1 (low), 2 (medium), and 3 (high) proficiency. Students' attitude towards AI was measured using two items: students' self-perceived view of using AI tools in (1) teaching and (2) learning. Responses were rated on a three-point scale, coded as 1 (negative), 2 (neutral), and 3 (positive), respectively. AI behavior refers to actual use of AI tools rather than behavioral intention and was evaluated through self-perceived usage frequency and application patterns. Usage frequency was assessed on a six-point scale (ranging from "never used" to "multiple times daily," coded one to six, where higher scores indicated more frequent academic use). Usage patterns were measured with two items: extent of using AI for (1) professional knowledge learning and (2) writing tasks (options included "never used," "generating initial drafts," "writing specific sections," "editing and polishing," "outlining," and "brainstorming ideas"). The Cronbach α values for cognitive,

XSL•FO
RenderX

attitudinal, and behavioral dimension were 0.81, 0.86, and 0.79, respectively, indicating high internal consistency. The Kaiser-Meyer-Olkin values were 0.76, 0.70, and 0.75 for the three dimensions. Additionally, Bartlett's test of sphericity was significant (P<.001), supporting the construct validity of the measurement scales. Further confirmatory factor analysis showed that the measurement instrument loaded well onto the three identified dimensions of AI literacy. Relevant statistics include comparative fit index (CFI)=0.994, Tucker-Lewis index (TLI)=0.991, Root Mean Square Error of Approximation (RMSEA)=0.025, and statistically significant loading on all factors (P<.001).

Factors Associated With AI Literacy

Existing research suggests that both individual and environmental factors associate with access to, and use of digital technologies [45]. For individual factors, we examined demographic characteristics, family background, intellectual ability, and enrollment motivation [47-49]. Demographic characteristics included sex (female or male), ethnicity (Han Chinese or other), only child (yes or no). Family background included hometown (urban or rural), father's education, mother's education, having at least one parent as a physician (yes or no), and high-income families (total family income from the previous year>150,000 RMB, approximately US $21,385, yes or no). We used the National College Entrance Examination scores to measure the respondents' intellectual ability. We assessed enrollment motivation influencing students' choice of clinical medicine following previous work [47]. Intrinsic motivation was measured by students' reported interest in medicine, confidence in achieving success in the field, and strong performance in relevant high school subjects. Extrinsic motivation was evaluated through responses about employment prospects, encouragement (or requirements) from significant others, and anticipated convenience in accessing health care resources for themselves and family members.

Regarding environmental factors, we considered students' affiliation with DFC universities, enrollment in long-track programs (ie, 5+3 or 8 y medical education program instead of 5 y medical education program; yes or no), and the educational phase (preclinical or clinical).

Statistical Analysis

We used the principal component analysis to aggregate the AI knowledge and AI attitude items, followed by Min-Max normalization to standardize the scores. Based on expert consultation and author consensus, we aggregated the AI behavior index using a weighted arithmetic mean of three key indicators: (1) usage frequency (50% weight), (2) extent of AI integration in professional knowledge learning (30% weight), and (3) AI utilization in writing tasks (20% weight). For robustness check, we also applied equal weights across all three indicators. We then applied the Min-Max normalization method, scaling the data so that the minimum and maximum values correspond to 0 and 100, respectively. Given that students were clustered within schools, we first fitted an unconditional (null) multilevel model to estimate the intraclass correlation coefficient (ICC) and assess the proportion of variance in the standardized AI literacy scores attributable to between-school differences. The ICCs for the three dimensions of AI literacy—knowledge, attitude, and behavior—were 4.0%, 1.0%, and 4.1%, respectively. These results indicate that only a small proportion of the total variance was explained by school-level clustering, suggesting minimal between-school effects. Therefore, we proceeded with multivariate linear regression to examine the associations between individual factors (ie, demographic characteristics, family background, and enrollment motivation) and environmental factors (ie, educational phase, type of education program, and tier of education program) and AI literacy. We conducted subgroup analyses and examined formal interaction terms based on institutional tier. Both classical and school-clustered standard errors were used to calculate 95% CI, and statistical significance was set at a two-sided P< 0.05. All analyses were performed using Stata (version 18.0; StataCorp LLC) between December 19, 2024, and March 18, 2025.

Results

Participant Characteristics

Of the 80,355 medical students included, most were Han (69,333 students, 86.3%) and female (41,227, 51.3%). A total of 28,885 students (35.9%) reported being the only child; 48,716 (60.6%) were from urban areas; 29,152 (36.3%) reported having at least one parent as a physician. In addition, 13,736 (17.1%) came from high-income families, 36,041 (44.9%) were in the clinical phase of training, and 5823 (7.2%) were enrolled in long-track medical education programs (Table 1).

XSL•FO
RenderX

Table . Characteristics of participants.

Variables	Participants, N=80,355			
	Non-DFC[a] (n=67,417)	DFC (n=12,938)	Total (N=80,355)	*P* value
Sex, n (%)				
Female	34563 (51.3)	6664 (51.5)	41227 (51.3)	.62
Male	32854 (48.7)	6274 (48.5)	39128 (48.7)	
Ethnicity, n (%)				
Others	9609 (14.3)	1413 (10.9)	11022 (13.7)	<.001
Han Chinese	57808 (85.7)	11525 (89.1)	69333 (86.3)	
Only child, n (%)				
No	44273 (65.7)	7197 (55.6)	51470 (64.1)	<.001
Yes	23144 (34.3)	5741 (44.4)	28885 (35.9)	
Hometown, n (%)				
Rural	27576 (40.9)	4063 (31.4)	31639 (39.4)	<.001
Urban	39841 (59.1)	8875 (68.6)	48716 (60.6)	
Father's education, mean (SD)	11.0 (3.7)	12.0 (4.0)	11.1 (3.8)	<.001
Mother's education, mean (SD)	10.0 (4.1)	11.1 (4.4)	10.2 (4.2)	<.001
Physician parent, n (%)				
No	43146 (64.0)	8057 (62.3)	51203 (63.7)	<.001
Yes	24271 (36.0)	4881 (37.7)	29152 (36.3)	
Family income, n (%)				
Middle and low	57022 (84.6)	9597 (74.2)	66619 (82.9)	<.001
High	10395 (15.4)	3341 (25.8)	13736 (17.1)	
NCEE[b] score, mean (SD)	546.9 (50.6)	593.7 (54.9)	554.5 (54.1)	<.001
Enrollment motivation, mean (SD)				
Intrinsic motivation	69.9 (18.4)	70.1 (18.6)	69.9 (18.4)	.15
Extrinsic motivation	68.4 (18.6)	67.7 (19.1)	68.3 (18.7)	<.001
Long-track programs[c], n (%)				
No	64425 (95.6)	10107 (78.1)	74532 (92.8)	<.001
Yes	2992 (4.4)	2831 (21.9)	5823 (7.2)	
Educational phase, n (%)				
Preclinical	37439 (55.5)	6875 (53.1)	44314 (55.1)	<.001
Clinical	29978 (44.5)	6063 (46.9)	36041 (44.9)	

[a]DFC: Double First-Class universities

[b]NCEE: National College Entrance Examination

[c]Long-track programs: the 5+3 medical education program and 8-year medical education program

When stratified by DFC status, students from DFCs came from relatively more advantaged families—with higher household incomes, more parental education, and greater likelihood of urban origin—and also had significantly higher college entrance examination scores (Table 1).

Distribution of the Three Dimensions of AI Literacy

In our analysis of the three dimensions of AI literacy among Chinese medical students (Table 2), the scores declined from knowledge (mean 76.0, SD 26.9) to attitude (mean 71.6, SD 24.4) and then further to behavior (mean 32.5, SD 28.5). When we stratified by institutional tier, students from DFC universities reported lower AI knowledge (mean 72.6, SD 28.6 vs mean

XSL•FO
RenderX

76.5, SD 26.7; *P*<.001) but higher attitude (mean 72.6, SD 24.6 vs mean 71.4, SD 24.4; *P*<.001), and behavior (mean 35.0, SD 28.4 vs mean 32.1, SD 28.5; *P*<.001).

Table . Summary results for three dimensions of AI literacy.

Variables	Non-DFC[a] (n=67,417)	DFC (n=12,938)	Total (N=80,355)	*P* value
AI knowledge, mean (SD)	76.5 (26.7)	72.6 (28.6)	76.0 (26.9)	<.001
AI attitude, mean (SD)	71.4 (24.4)	72.6 (24.6)	71.6 (24.4)	<.001
AI behavior, mean (SD)[b]	32.1 (28.5)	35.0 (28.4)	32.5 (28.5)	<.001
AI behavior, mean (SD)[c]	32.3 (29.5)	35.4 (29.5)	32.8 (29.5)	<.001

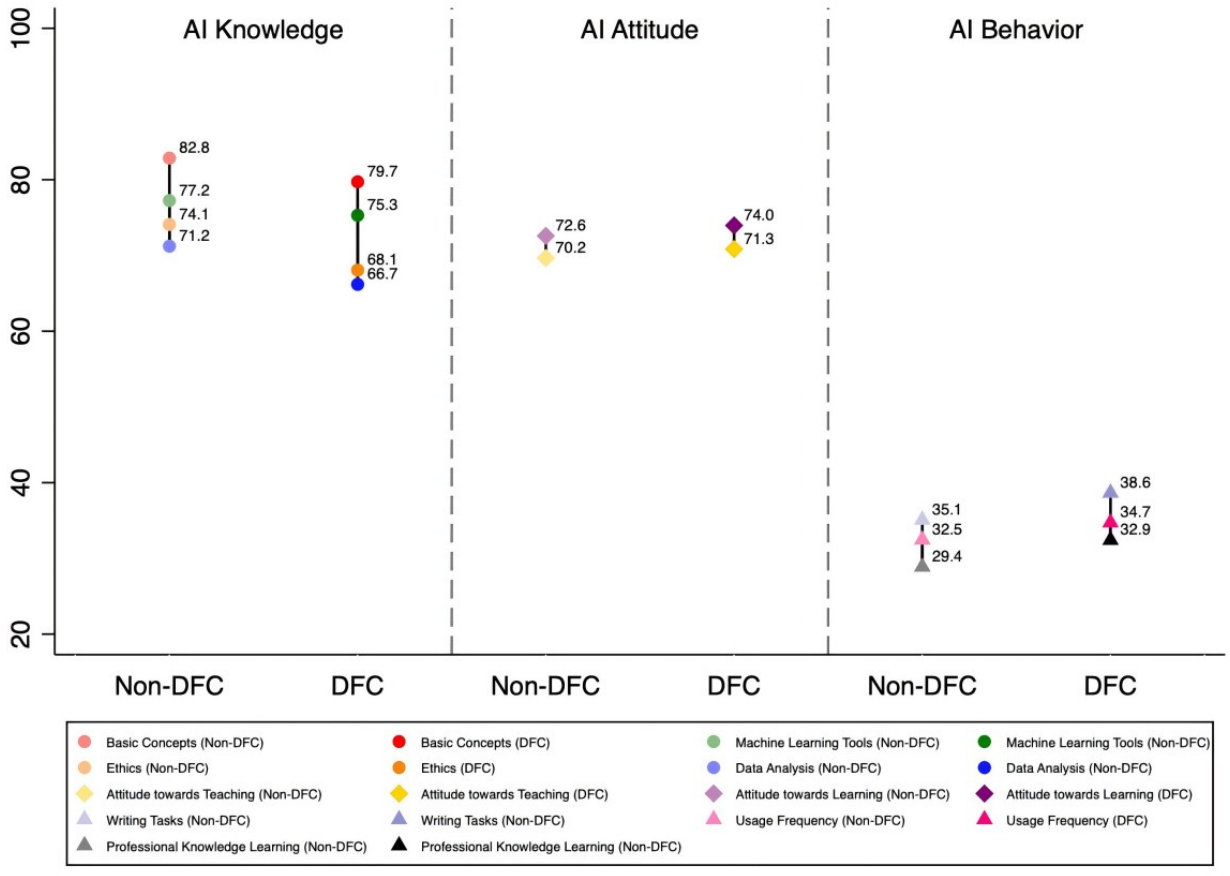
[a]DFC: Double First-Class universities.

[b] expert -assigned weights.

[c]equal weights.

Figure 1 further illustrates the item-level patterns within each dimension of AI literacy. The descending trend—from higher scores in knowledge to lower scores in attitude and the lowest in behavior—was consistent across institutional tiers. Notably, students from DFC universities scored higher on AI behavior subcomponents, especially in professional learning and writing tasks. More detailed results on each institution are provided in Figure S1 in Multimedia Appendix 1.

Figure 1. Components of the three dimensions of AI literacy by institutional tier AI: artificial intelligence; DFC: Double First-Class universities. DFC represents advantaged universities, and non-DFC represents less advantaged universities.



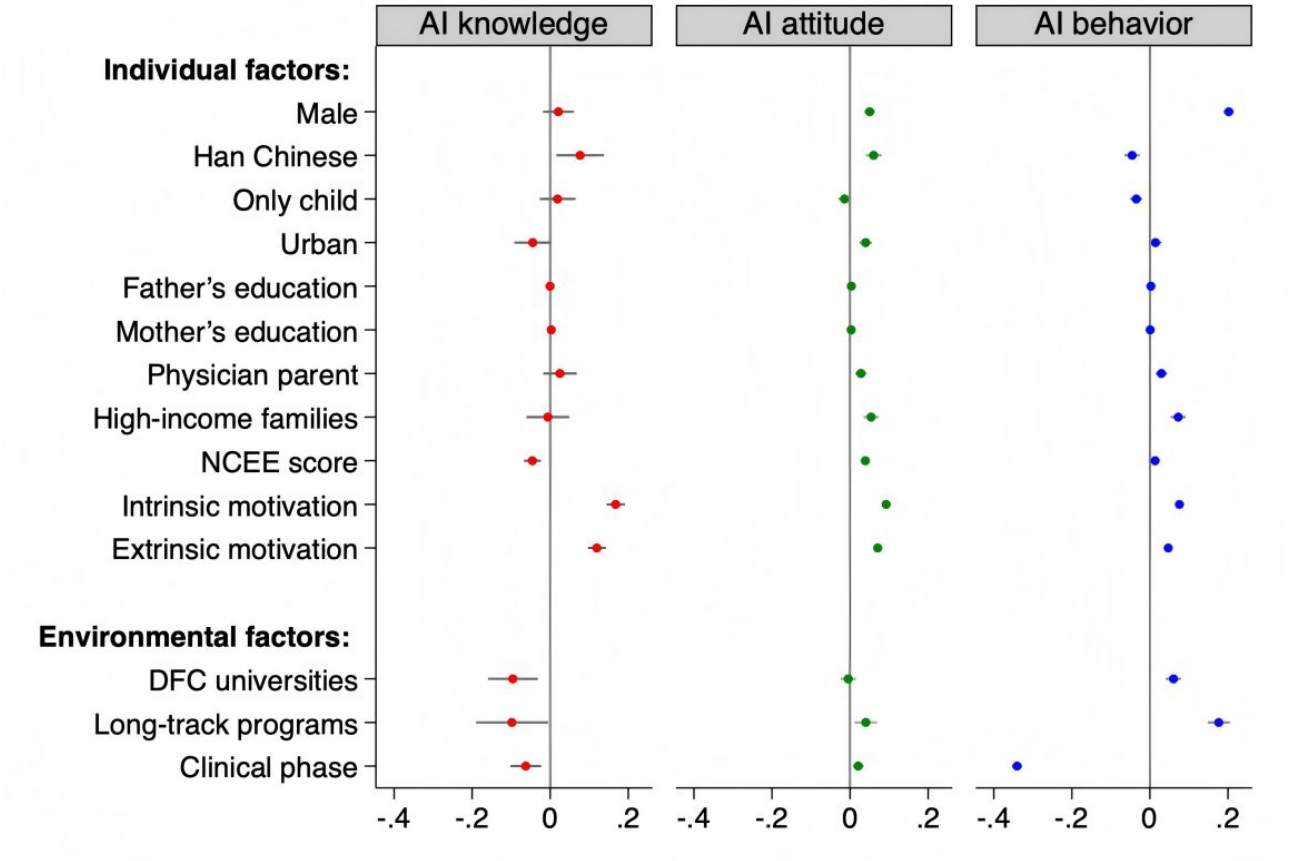
The figure shows the distribution of the subcomponents of the three dimensions of AI literacy by institutional tier. The AI knowledge dimension comprised students' self-reported proficiency in core areas of medical AI, including (1) basic concepts of AI, (2) machine learning tools, (3) multimodal medical data analysis, and (4) ethics issues related to AI. The AI attitude dimension included students' view of using AI tools in (1) teaching and (2) learning. The AI behavior dimension was measured by usage frequency and application patterns, including AI use for professional knowledge learning and writing tasks.

XSL•FO
RenderX

Factors Associated With AI Knowledge

Figure 2 shows the results of regression analyzes of the individual and environmental factors associated with AI literacy (more details including robustness check results are available in Tables S1-S4 in Multimedia Appendix 1). Male students were slightly more likely to report higher AI knowledge (β=0.02, 95% CI –0.02 to 0.06; *P*=.30). Both intrinsic (β=0.17, 95% CI 0.14 to 0.19; *P*<.001) and extrinsic motivation (β=0.12, 95% CI, 0.10 to 0.14; *P*<.001) were positively associated with AI knowledge. Students enrolled in long-track programs reported lower AI knowledge than those in the five-year programs (β=−0.10, 95% CI −0.19 to −0.01; *P*=.04). When stratified by DFC status, these patterns remained largely consistent. However, students from DFC universities reported significantly less AI knowledge (β=−0.10, 95% CI −0.16 to −0.03; *P*=.003).

Figure 2. The individual and environmental factors associated with AI literacy in the full sample. AI: artificial intelligence; DFC: Double First-Class; NCEE: National College entrance examination. DFC represents advantaged universities, and non-DFC represents less advantaged universities. Long-track programs include the 5+3 medical education program and 8-year medical education program.



Factors Associated With AI Attitude

Male students (β=0.05, 95% CI 0.04 to 0.06; *P*<.001) demonstrated significantly more positive attitudes toward AI. Both intrinsic (β=0.09, 95% CI 0.08 to 0.10; *P*<.001) and extrinsic motivation (β=0.07, 95% CI 0.06 to 0.08; *P*<.001) were positively associated with AI attitude. Students from high-income families (β=0.05, 95% CI 0.03 to 0.07; *P*<.001), with at least one physician parent (β=0.03, 95% CI 0.01 to 0.04; *P*<.001), and urban backgrounds (β=0.04, 95% CI 0.02 to 0.06; *P*<.001) showed more positive AI attitude. As for environmental factors, students enrolled in long-track programs (β=0.04, 95% CI 0.01 to 0.07; *P*<.006) and in clinical phase of training (β=0.02, 95% CI 0.01 to 0.03; *P*<.003) showed significantly more positive attitudes toward AI. These patterns remained largely consistent in subgroup analyses by DFC status; however, among the non-DFC students, the associations of family income and parental medical background on AI attitude were significantly stronger.

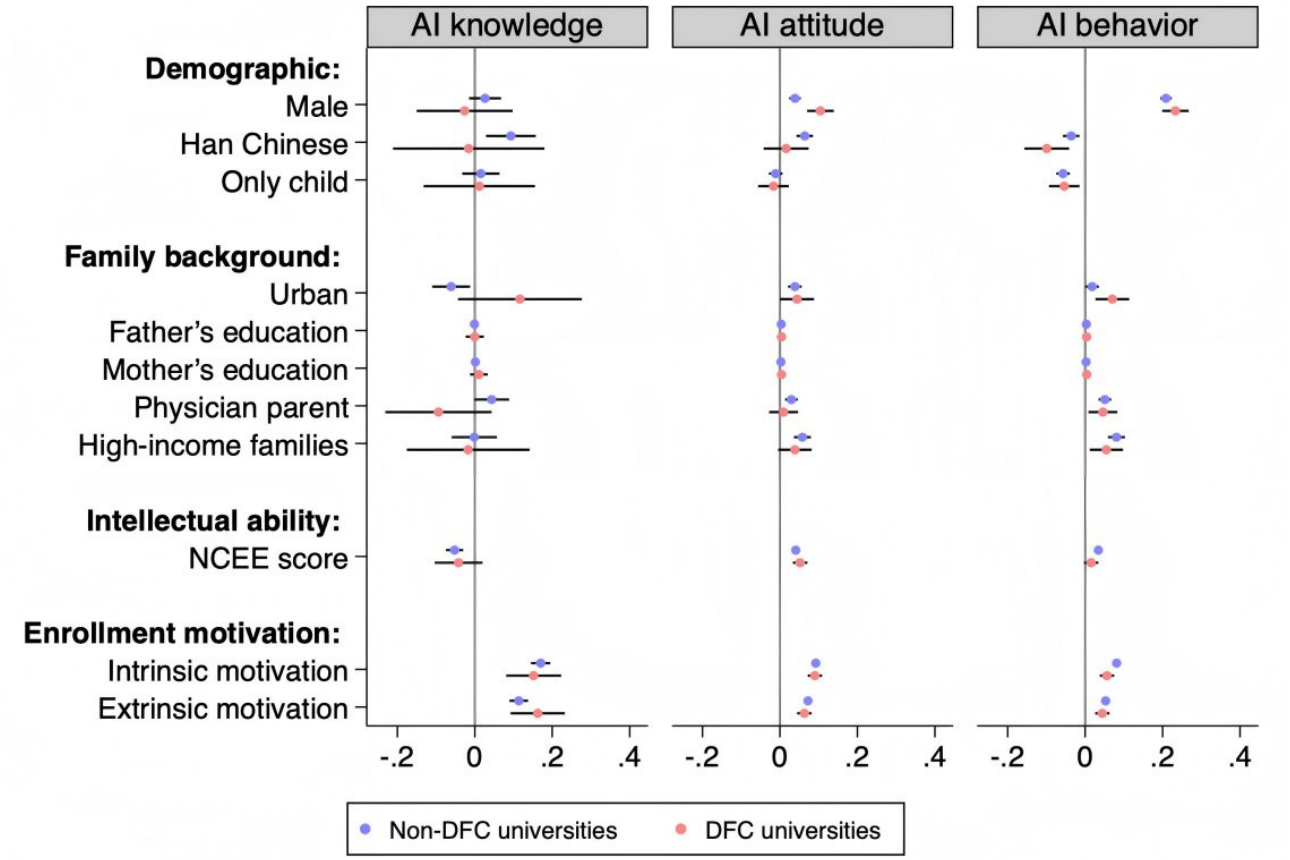
Factors Associated With AI Behavior

Males (β=0.20, 95% CI 0.19 to 0.22; *P*<.001) demonstrated higher levels of AI behavior. Students from high-income families (β=0.07, 95% CI 0.05 to 0.09; *P*<.001) and those with at least one physician parent (β=0.03, 95% CI 0.01 to 0.04; *P*<.001) showed higher AI behavior scores. Both intrinsic (β=0.08, 95% CI 0.07 to 0.08; *P*<.001) and extrinsic motivation (β=0.05, 95% CI 0.04 to 0.05; *P*<.001) were positively associated with AI behavior. Students enrolled in long-track programs (β=0.18, 95% CI 0.15 to 0.20; *P*<.001) exhibited higher AI behavior. Again, these patterns were largely held in subgroup analysis by DFC status. In the non-DFC group, however, the effect of family income and parental medical background was again stronger.

Figure 3 illustrates the associations between individual factors and AI literacy across different institutional environments.

XSL•FO
RenderX

Figure 3. The individual and environmental factors associated with AI literacy. AI: artificial intelligence; DFC: Double First-Class; NCEE: National College entrance examination.



Discussion

Principal Findings

Despite the recognized importance of preparing medical students to use AI tools effectively, systematic evidence on the subcomponents of AI literacy and the associated factors is limited. To our knowledge, this is the first large-scale, multicenter, national-level study on the multidimensional constructs of AI literacy and their associated individual and environmental factors among medical students. Three key results emerged. First, medical students reported moderate to high AI knowledge, moderate AI attitude, but only low AI behavior scores. Second, both individual and environmental factors were associated with different dimensions of AI literacy. Third, compared to less favorable environmental conditions, the association between worse AI literacy and family income and parental medical background may be weaker.

This study provides several significant contributions. First, the use of a large, multicenter, and nationally representative sample—covering more than 50% (109/202) of the medical schools in China—expands beyond the small, single-center settings from previous studies. Second, by comprehensively examining the cognitive, attitudinal, and behavioral dimensions of AI literacy, this study provides a more holistic understanding of medical students' AI literacy profiles, identifying nuanced gaps and interventional needs within AI literacy, particularly around attitude and behaviors. Third, our comprehensive examination of the individual differences and environmental factors associated with AI literacy provides an evidence-based foundation for developing tailored interventions. Understanding these associations enables medical educators to design targeted programs that meet the diverse needs of student populations, potentially reducing disparities in AI literacy and ensuring that all future physicians are adequately prepared to engage with an increasingly AI-integrated health care system.

Our findings consistently revealed that self-perceived AI knowledge scores surpassed attitude and behavior. This suggests that while students developed a foundational understanding of AI and generally perceived AI integration in medicine positively, they still lacked practical experience in its application. Although this defies common logic that deeper knowledge should portend more use [50], this trend aligns with previous international studies of medical staff and trainees who held generally positive attitudes toward AI despite low exposure in formal use [36,51].

The potential barriers underlying this knowledge–use gap are as follows. First, structural barriers—such as limited institutional infrastructure, insufficient curricular integration, and time constraints—may constrain students' ability to translate AI knowledge into practice [52]. Such challenges are particularly salient in the Chinese healthcare systems, where physician shortages drive each health care provider to carry significant patient load [53]. These contextual limitations can, in turn, dampen students' autonomous motivation to explore alternative tools [54]. Second, sociocultural norms may also constrain AI use [55]. A recent study showed that nearly half of Chinese

XSL•FO
RenderX

medical staff (1320/2705, 48.8%) expressed concerns about the ethical safety, accuracy, interpretability, and fairness of AI systems [56]. Such ethical reservations from staff providers to medical trainees, leading them to perceive AI use as potentially inappropriate, dishonest, unprofessional, or harmful. This may then limit the students' willingness to use the AI tools despite being knowledgeable of AI tools' capabilities. suggesting that moral and risk considerations may underlie the limited AI use observed among some medical students. Third, our survey was conducted between May and July 2024, when most mature and user-friendly AI tools were developed abroad. Chinese students may have faced barriers related to access [57] and cultural appropriateness [58].

At an individual level, students' demographics, family background, and motivation were all associated with AI literacy. The patterns generally aligned with the literature. Demographically, male medical students scored higher in AI attitude and behavior. This is consistent with previous findings [50,59] and the phenomenon of a "technological gender gap" [60]. Male students tended to perceive AI tools as more useful and easier to use [61-64]. Additionally, individuals from advantaged family background (ie, urban residents, those with at least one physician parent, and students from high-income families) perceived AI more favorably and used it more frequently. These advantaged groups hold greater material, cognitive, social, and cultural capital that can facilitate technological adoption [65]. Although technological innovations eventually reach broader populations, disparities in access and adoption persist during the diffusion process, with advantaged groups generally adopting innovations earlier and using them more effectively, thus creating a digital divide between socioeconomic groups [65,66].

Notably, both intrinsic and extrinsic motivation were positively associated with all dimensions of AI literacy, with intrinsic motivation showing particularly strong effects. According to self-determination theory, intrinsic motivation is associated with better performance and enhanced subjective well-being, while extrinsic motivation can also drive performance, even filtering into intrinsic motivation under certain conditions [67-69]. Intrinsic motivation serves as a natural source of AI learning and achievement, fostering greater autonomy, exploration, and reflection that promote deeper AI engagement. In contrast, extrinsically motivated students may preferentially recognize the instrumental value of AI but limit autonomous exploration of AI tools.

Our study highlights the significant role of the environment in shaping AI literacy. Specifically, DFC students demonstrated higher AI behavior scores. The DFC Project represents China's willingness to invest financial and human resources in these institutions to develop world-class universities and disciplines [47,70]. The stronger student–faculty and peer interactions and higher-quality student efforts may promote deeper engagement with AI [71]. The seemingly paradoxical finding that DFC students reported lower AI knowledge but higher AI behavior may stem from the Dunning–Kruger effect [72,73]. More competent individuals tend to recognize the complexity of a domain and their own knowledge limitations, leading to more cautious self-assessments, while those less competent may overestimate their understanding. DFC students, exposed to higher levels of academic research, may be more aware of the challenges inherent in AI, whereas non-DFC students, with relatively limited exposure, may exhibit overconfidence in their basic knowledge. Similar cognitive bias patterns have been observed among health professions students; for example, low-performing medical students have been shown to substantially overestimate their academic performance relative to peers [74,75].

Importantly, in these DFC supportive environments, the students' family income and parental medical background showed less pronounced associations with AI attitude and behavior, suggesting that the supportive learning environment may have mitigated the detrimental relationship between students' background and AI literacy [48].

In addition, the students' program also appears to matter. Students in the long-track program demonstrated significantly higher AI attitude and behavior scores. They typically represent China's elite medical students, attending research-oriented universities that provide enhanced research training opportunities and funding [76], potentially including familiarization with AI technologies that led to more positive perceptions and use of AI tools. The students' phase of training also appeared to be associated with their AI literacy. Students undergoing clinical training reported higher attitude scores but lower knowledge and behavior scores. Perhaps the complexity of clinical practice and the pressure of professional development precipitated cognitive overload and a preferential reliance on traditional tools [77]. The technical limitations and evolving normative and legal thresholds around AI technologies may further exacerbate technology avoidance [78-82].

Tying together the findings, we provide five sets of recommendations to help medical schools maneuver through the current lack of structured AI education and training [83]. First, carefully integrating AI-related content into the curriculum to generate practical AI exposure may help bridge the knowledge-use gap. This should be mindful of the potentially detrimental association with excessive cognitive offloading [84] while ensuring a minimal knowledge and skills threshold [85]. As regulations and ethical frameworks surrounding AI use in health care continue to evolve to meet the technical, ethical, and legal challenges [83,86,87], developing interdisciplinary training on practical risks, benefits, and guidelines on AI usage may help establish norms around AI usage. Second, developing and implementing scalable AI education frameworks that enable targeted training based on learners' backgrounds can potentially reduce inequities in AI literacy driven by individual and environmental differences while maximizing educational impact.

The next three recommendations focus specifically on digital divide, and we draw on Han and Kumwenda's framework for mitigating digital divide. The framework considers first-order barriers relate to infrastructure and access; second-order barriers include motivation and engagement; and third-order barriers reflect skill and training disparities [88].

First, medical educators and policy makers should prioritize investments in digital infrastructure to reduce disparities in technological access across institutional and regional contexts

XSL•FO
RenderX

[89]. Targeted funding and digital resource allocation can ensure that students in under-resourced settings are not excluded from AI-driven learning environments.

Second, improving motivation and engagement may incorporate experiential learning opportunities [87,90,91]. Providing all students—particularly those from disadvantaged backgrounds—with the chance to work directly with AI tools and integrate them into their coursework can foster greater motivation, engagement, and confidence in using AI technologies.

Third, structured and scalable AI education frameworks that offer tiered and targeted training that adapts to learners' backgrounds. Such approaches would focus on ensuring that disadvantaged students can develop foundational AI literacy before advancing to higher-order skills. Moreover, the development of multilingual and culturally adaptive AI models [92] can further ensure inclusivity and contextual relevance in AI-enhanced medical education.

Limitations

First, the reliance on self-reported data potentially exposes the study to social desirability bias. It is necessary to develop objective instruments to measure medical students' AI literacy. Such a tool can help evaluate AI proficiency based on observable AI use. Performance- and scenario-based assessments might be potential avenues to operationalize such evaluation. Second, given the rapid development of AI technologies, the AI literacy of medical students is likely to evolve rapidly. To capture these dynamic shifts of AI literacy, we are actively exploring another round of survey. Third, this study employed a cross-sectional design, which precludes causal inference. Although most individual and environmental factors were determined before the measurement of AI literacy, reverse causality cannot be fully excluded, and unmeasured confounders may have influenced both the predictors and outcomes, potentially biasing the observed associations. Fourth, even though the unique premedical training structure limits the generalizability of the findings, the methodological framework and multidimensional conceptualization of AI literacy (knowledge, attitude, and behavior) may still provide a useful reference for future comparative studies or adaptations in other educational systems.

Conclusions

This cross-sectional study revealed that medical students exhibited the highest performance in AI knowledge, followed by attitude, and then behavior. Both individual characteristics and environmental factors were significantly associated with AI literacy, and environmental factors moderated individual variations in AI literacy. Integrating practical AI-related training into medical curricula through interdisciplinary collaboration, coupled with targeted interventions for students according to their backgrounds, may help prepare future physicians to effectively engage with AI technologies in medical practice.

Acknowledgments

We are grateful to the National Center for Health Professions Education Development for their support of this study. We also extend our thanks to all the medical schools and medical students who participated in the CMSS survey. We also appreciate the insightful comments from Dr. Weimin Wang, Ms Ana Xie, and Dr. Chen Yu at the National Center for Health Professions Education Development, Peking University, as well as from Dr. Dragan Gasevic and Dr. Danijela Gasevic at Monash University.

CL was affiliated with the Institute of Medical Education at Peking University at the time of the study and is currently affiliated with the Education Management Information Center, Ministry of Education of the People's Republic of China.

Funding

This study was supported by the 2025 Undergraduate Teaching Reform Project of Peking University (JG2025170). This study was supported by the 2025 Undergraduate Teaching Reform Project of Peking University (JG2025170). The funder had no role in the design and conduct of the study; collection, management, analysis, and interpretation of the data; preparation, review, or approval of the manuscript; and decision to submit the manuscript for publication.

Disclaimer

The views expressed herein are solely those of the authors and do not necessarily represent the views of their employers, funders, or affiliated institutions.

Data Availability

The parameters that we used in our model (text, tables, figures, models, and appendices) are available on reasonable request from the corresponding author under certain conditions (with the consent of all participating centers and with a signed data access agreement).

Authors' Contributions

HW had full access to all of the data in the study and takes responsibility for the integrity of the data and the accuracy of the data analysis.
Concept and design: CL, HW.

XSL•FO
RenderX

Acquisition, analysis, or interpretation of data: CL, SHT, HW.
Drafting of the manuscript: CL, HW.
Critical review of the manuscript for important intellectual content: CL, SHT, HW.
Statistical analysis: CL.
Obtained funding: Wu.
Administrative, technical, or material support: HW.
Supervision: HW.

Conflicts of Interest

None declared.

Multimedia Appendix 1 [DOC File, 1036 KB - mededu_v12i1e80604_app1.doc]

References

1. Elemento O, Khozin S, Sternberg CN. The use of artificial intelligence for cancer therapeutic decision-making. NEJM AI 2025 May;2(5). [doi: 10.1056/aira2401164] [Medline: 41112204]
2. Sarkar U, Bates DW. Using artificial intelligence to improve primary care for patients and clinicians. JAMA Intern Med 2024 Apr 1;184(4):343-344. [doi: 10.1001/jamainternmed.2023.7965] [Medline: 38345801]
3. Haug CJ, Drazen JM. Artificial intelligence and machine learning in clinical medicine, 2023. N Engl J Med 2023 Mar 30;388(13):1201-1208. [doi: 10.1056/NEJMra2302038] [Medline: 36988595]
4. Unlu O, Varugheese M, Shin J, et al. Manual vs AI-assisted prescreening for trial eligibility using large language models-a randomized clinical trial. JAMA 2025 Mar 25;333(12):1084-1087. [doi: 10.1001/jama.2024.28047] [Medline: 39960745]
5. Boscardin CK, Gin B, Golde PB, Hauer KE. ChatGPT and generative artificial intelligence for medical education: potential impact and opportunity. Acad Med 2024;99(1):22-27. [doi: 10.1097/ACM.0000000000005439]
6. Preiksaitis C, Rose C. Opportunities, challenges, and future directions of generative artificial intelligence in medical education: scoping review. JMIR Med Educ 2023 Oct 20;9(1):e48785. [doi: 10.2196/48785] [Medline: 37862079]
7. Gordon M, Daniel M, Ajiboye A, et al. A scoping review of artificial intelligence in medical education: BEME Guide No. 84. Med Teach 2024 Apr;46(4):446-470. [doi: 10.1080/0142159X.2024.2314198] [Medline: 38423127]
8. Burk-Rafel J, Reinstein I, Feng J, et al. Development and validation of a machine learning-based decision support tool for residency applicant screening and review. Acad Med 2021;96(11S):S54-S61. [doi: 10.1097/ACM.0000000000004317]
9. Baron T, Grossman RI, Abramson SB, et al. Signatures of medical student applicants and academic success. PLoS One 2020;15(1):e0227108. [doi: 10.1371/journal.pone.0227108] [Medline: 31940377]
10. Kind T, Olvet DM, Farina G, et al. Reading and study habits of medical students on clerkships and performance outcomes: a multi-institutional study. Med Sci Educ 2021 Dec;31(6):1957-1966. [doi: 10.1007/s40670-021-01409-5] [Medline: 34956707]
11. Cutrer WB, Spickard WA 3rd, Triola MM, et al. Exploiting the power of information in medical education. Med Teach 2021 Jul;43(sup2):S17-S24. [doi: 10.1080/0142159X.2021.1925234] [Medline: 34291714]
12. Roberts LW. Addressing the novel implications of generative ai for academic publishing, education, and research. Acad Med 2024;99(5):471-473. [doi: 10.1097/ACM.0000000000005667]
13. Stokel-Walker C. ChatGPT listed as author on research papers: many scientists disapprove. Nature New Biol 2023 Jan;613(7945):620-621. [doi: 10.1038/d41586-023-00107-z] [Medline: 36653617]
14. Flanagin A, Bibbins-Domingo K, Berkwits M, Christiansen SL. Nonhuman “Authors” and implications for the integrity of scientific publication and medical knowledge. JAMA 2023 Feb 28;329(8):637. [doi: 10.1001/jama.2023.1344]
15. Tolsgaard MG, Boscardin CK, Park YS, Cuddy MM, Sebok-Syer SS. The role of data science and machine learning in Health Professions Education: practical applications, theoretical contributions, and epistemic beliefs. Adv Health Sci Educ Theory Pract 2020 Dec;25(5):1057-1086. [doi: 10.1007/s10459-020-10009-8] [Medline: 33141345]
16. Gin BC, Ten Cate O, O’Sullivan PS, Hauer KE, Boscardin C. Exploring how feedback reflects entrustment decisions using artificial intelligence. Med Educ 2022 Mar;56(3):303-311. [doi: 10.1111/medu.14696] [Medline: 34773415]
17. Schaye V, Guzman B, Burk-Rafel J, et al. Development and validation of a machine learning model for automated assessment of resident clinical reasoning documentation. J Gen Intern Med 2022 Jul;37(9):2230-2238. [doi: 10.1007/s11606-022-07526-0] [Medline: 35710676]
18. Triola MM, Rodman A. Integrating generative artificial intelligence into medical education: curriculum, policy, and governance strategies. Acad Med 2025;100(4):413-418. [doi: 10.1097/ACM.0000000000005963]
19. Russell RG, Lovett Novak L, Patel M, et al. Competencies for the use of artificial intelligence–based tools by health care professionals. Acad Med 2023;98(3):348-356. [doi: 10.1097/ACM.0000000000004963]
20. Ong QC, Ang CS, Lai NM, Neves AL, Car J. Dearth of digital health education: the need for an accelerated medical curriculum reform in Malaysia. Lancet Reg Health West Pac 2025 Feb;55:101476. [doi: 10.1016/j.lanwpc.2025.101476] [Medline: 39902151]

XSL•FO
RenderX

21. Succi MD, Chang BS, Rao AS. Building the AI-enabled medical school of the future. JAMA 2025 May 20;333(19):1665-1666. [doi: 10.1001/jama.2025.2789] [Medline: 40163081]
22. Lee J, Wu AS, Li D, Kulasegaram K(. Artificial intelligence in undergraduate medical education: a scoping review. Acad Med 2021;96(11S):S62-S70. [doi: 10.1097/ACM.0000000000004291]
23. Long D, Magerko B. What is AI literacy? Competencies and design considerations. CHI '20: Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems 2020 Apr 21:1-16. [doi: 10.1145/3313831.3376727]
24. Ng DTK, Wu W, Leung JKL, Chiu TKF, Chu SKW. Design and validation of the ai literacy questionnaire: the affective, behavioural, cognitive and ethical approach. Brit J Educational Tech 2024 May;55(3):1082-1104. [doi: 10.1111/bjet.13411]
25. Ng DTK, Leung JKL, Chu SKW, Qiao MS. Conceptualizing AI literacy: an exploratory review. Computers and Education: Artificial Intelligence 2021;2:100041. [doi: 10.1016/j.caeai.2021.100041]
26. Aktoprak A, Hursen C. A bibliometric and content analysis of critical thinking in primary education. Think Skills Creat 2022 Jun;44:101029. [doi: 10.1016/j.tsc.2022.101029]
27. Gruber T, Bazhydai M, Sievers C, Clément F, Dukes D. The ABC of social learning: affect, behavior, and cognition. Psychol Rev 2022 Nov;129(6):1296-1318. [doi: 10.1037/rev0000311] [Medline: 34292022]
28. Boyle MJ, Williams B, Brown T, et al. Attitudes of undergraduate health science students towards patients with intellectual disability, substance abuse, and acute mental illness: a cross-sectional study. BMC Med Educ 2010 Oct 21;10(1):71. [doi: 10.1186/1472-6920-10-71] [Medline: 20964840]
29. Van Harreveld F. The ABC of ambivalence: affective, behavioral, and cognitive consequences of attitudinal conflict. In: Adv Exp Soc Psychol 2015, Vol. 52:285-324. [doi: 10.1016/bs.aesp.2015.01.002]
30. Patino GA, Amiel JM, Brown M, Lypson ML, Chan TM. The promise and perils of artificial intelligence in health professions education practice and scholarship. Acad Med 2024;99(5):477-481. [doi: 10.1097/ACM.0000000000005636]
31. Barreiro-Ares A, Morales-Santiago A, Sendra-Portero F, Souto-Bayarri M. Impact of the rise of artificial intelligence in radiology: What do students think? Int J Environ Res Public Health 2023 Jan 16;20(2):1589. [doi: 10.3390/ijerph20021589] [Medline: 36674348]
32. Stewart J, Lu J, Gahungu N, et al. Western Australian medical students' attitudes towards artificial intelligence in healthcare. PLoS One 2023;18(8):e0290642. [doi: 10.1371/journal.pone.0290642] [Medline: 37651380]
33. Wood EA, Ange BL, Miller DD. Are we ready to integrate artificial intelligence literacy into medical school curriculum: students and faculty survey. J Med Educ Curric Dev 2021;8:23821205211024078. [doi: 10.1177/23821205211024078] [Medline: 34250242]
34. Ejaz H, McGrath H, Wong BL, Guise A, Vercauteren T, Shapey J. Artificial intelligence and medical education: a global mixed-methods study of medical students' perspectives. Digit Health 2022;8:20552076221089099. [doi: 10.1177/20552076221089099] [Medline: 35521511]
35. Doumat G, Daher D, Ghanem NN, Khater B. Knowledge and attitudes of medical students in Lebanon toward artificial intelligence: A national survey study. Front Artif Intell 2022;5:1015418. [doi: 10.3389/frai.2022.1015418] [Medline: 36406470]
36. Sit C, Srinivasan R, Amlani A, et al. Attitudes and perceptions of UK medical students towards artificial intelligence and radiology: a multicentre survey. Insights Imaging 2020 Feb 5;11(1):14. [doi: 10.1186/s13244-019-0830-7] [Medline: 32025951]
37. Mehta N, Harish V, Bilimoria K, et al. Knowledge of and attitudes on artificial intelligence in healthcare: a provincial survey study of medical students. Medical Education. Preprint posted online on 2021. [doi: 10.1101/2021.01.14.21249830]
38. Moldt JA, Festl-Wietek T, Madany Mamlouk A, Nieselt K, Fuhl W, Herrmann-Werner A. Chatbots for future docs: exploring medical students' attitudes and knowledge towards artificial intelligence and medical chatbots. Med Educ Online 2023 Dec;28(1):2182659. [doi: 10.1080/10872981.2023.2182659] [Medline: 36855245]
39. Teng M, Singla R, Yau O, et al. Health care students' perspectives on artificial intelligence: countrywide survey in Canada. JMIR Med Educ 2022 Jan 31;8(1):e33390. [doi: 10.2196/33390] [Medline: 35099397]
40. Wang B, Rau PLP, Yuan T. Measuring user competence in using artificial intelligence: validity and reliability of artificial intelligence literacy scale. Behaviour & Information Technology 2023 Jul 4;42(9):1324-1337. [doi: 10.1080/0144929X.2022.2072768]
41. Laupichler MC, Aster A, Meyerheim M, Raupach T, Mergen M. Medical students' AI literacy and attitudes towards AI: a cross-sectional two-center study using pre-validated assessment instruments. BMC Med Educ 2024 Apr 10;24(1):401. [doi: 10.1186/s12909-024-05400-7] [Medline: 38600457]
42. Swed S, Alibrahim H, Elkalagi NKH, et al. Knowledge, attitude, and practice of artificial intelligence among doctors and medical students in Syria: A cross-sectional online survey. Front Artif Intell 2022;5:1011524. [doi: 10.3389/frai.2022.1011524] [Medline: 36248622]
43. Holstein K, Doroudi S. Equity and artificial intelligence in education: will "AIED" amplify or alleviate inequities in education? arXiv. Preprint posted online on Apr 27, 2021. [doi: 10.48550/arXiv.2104.12920]
44. van Deursen AJ, van Dijk JA. The digital divide shifts to differences in usage. New Media & Society 2014 May;16(3):507-526. [doi: 10.1177/1461444813487959]

XSL•FO
RenderX

45. van Dijk J. Digital divide: impact of access. In: Rössler P, Hoffner CA, Zoonen L, editors. The International Encyclopedia of Media Effects, 1st edition 2017:1-11. [doi: 10.1002/9781118783764.wbieme0043]
46. Ma X, Shen Z, Xiao R, Wu H. Perceived mistreatment and professional identity of medical students in China. JAMA Netw Open 2024 Nov 4;7(11):e2444245. [doi: 10.1001/jamanetworkopen.2024.44245]
47. Wu H, Li S, Zheng J, Guo J. Medical students' motivation and academic performance: the mediating roles of self-efficacy and learning engagement. Med Educ Online 2020 Dec;25(1):1742964. [doi: 10.1080/10872981.2020.1742964] [Medline: 32180537]
48. Loh RSM, Kraaykamp G, van Hek M. Plugging in at school: do schools nurture digital skills and narrow digital skills inequality? Computers & Education 2025 Mar;226:105195. [doi: 10.1016/j.compedu.2024.105195]
49. Hargittai E. Digital na(t)ives? Variation in internet skills and uses among members of the "Net Generation". Sociol Inq 2010 Feb;80(1):92-113. [doi: 10.1111/j.1475-682X.2009.00317.x]
50. Pinto dos Santos D, Giese D, Brodehl S, et al. Medical students' attitude towards artificial intelligence: a multicentre survey. Eur Radiol 2019 Apr;29(4):1640-1646. [doi: 10.1007/s00330-018-5601-1]
51. Sarwar S, Dent A, Faust K, et al. Physician perspectives on integration of artificial intelligence into diagnostic pathology. NPJ Digit Med 2019;2:28. [doi: 10.1038/s41746-019-0106-0] [Medline: 31304375]
52. Kennedy T, Regehr G, Rosenfield J, Roberts SW, Lingard L. Exploring the gap between knowledge and behavior: a qualitative study of clinician action following an educational intervention. Acad Med 2004 May;79(5):386-393. [doi: 10.1097/00001888-200405000-00006] [Medline: 15107277]
53. Wang X, Sanders HM, Liu Y, et al. ChatGPT: promise and challenges for deployment in low- and middle-income countries. The Lancet Regional Health - Western Pacific 2023 Dec;41:100905. [doi: 10.1016/j.lanwpc.2023.100905]
54. Ryan RM, Deci EL. Self-determination theory. In: Encyclopedia of Quality of Life and Well-Being Research: Springer, Vol. 2024:6229-6235. [doi: 10.1007/978-3-031-17299-1_2630]
55. Kahlke RM, McConnell MM, Wisener KM, Eva KW. The disconnect between knowing and doing in health professions education and practice. Adv in Health Sci Educ 2020 Mar;25(1):227-240. [doi: 10.1007/s10459-019-09886-5]
56. Dai Q, Li M, Yang M, et al. Attitudes, perceptions, and factors influencing the adoption of AI in health care among medical staff: nationwide cross-sectional survey study. J Med Internet Res 2025 Aug 8;27:e75343. [doi: 10.2196/75343] [Medline: 40779308]
57. Meaningful connectivity: a new target to raise the bar for internet access. Alliance for Affordable Internet. URL: https://a4ai.org/wp-content/uploads/2021/02/Meaningful-Connectivity_Public-.pdf [accessed 2025-12-23]
58. Cao Y, Zhou L, Lee S, Cabello L, Chen M, Hershcovich D. Assessing cross-cultural alignment between chatgpt and human societies: an empirical study. . 2023. [doi: 10.18653/v1/2023.c3nlp-1.7]
59. Akca Sumengen A, Ozcevik Subasi D, Cakir GN. Nursing students' attitudes and literacy toward artificial intelligence: a cross-sectional study. Teaching and Learning in Nursing 2025 Jan;20(1):e250-e257. [doi: 10.1016/j.teln.2024.10.022]
60. Canada K, Brusca F. The technological gender gap: evidence and recommendations for educators and computer-based instruction designers. ETR&D 1991 Jun;39(2):43-51. [doi: 10.1007/BF02298153]
61. Cai Z, Fan X, Du J. Gender and attitudes toward technology use: a meta-analysis. Comput Educ 2017 Feb;105:1-13. [doi: 10.1016/j.compedu.2016.11.003]
62. Davis FD. Perceived usefulness, perceived ease of use, and user acceptance of information technology. MIS Q 1989 Sep 1;13(3):319-340. [doi: 10.2307/249008]
63. Ong CS, Lai JY. Gender differences in perceptions and relationships among dominants of e-learning acceptance. Comput Human Behav 2006 Sep;22(5):816-829. [doi: 10.1016/j.chb.2004.03.006]
64. Ackerman PL, Wolman SD. Determinants and validity of self-estimates of abilities and self-concept measures. J Exp Psychol Appl 2007 Jun;13(2):57-78. [doi: 10.1037/1076-898X.13.2.57] [Medline: 17535132]
65. van Dijk J. : Polity Press; 2020.
66. Norris P. Digital divide: civic engagement, information poverty, and the internet worldwide. Canadian Journal of Communication 2003 Jan 1;28(1). [doi: 10.22230/cjc.2003v28n1a1352]
67. Ryan RM, Deci EL. Intrinsic and extrinsic motivations: classic definitions and new directions. Contemp Educ Psychol 2000 Jan;25(1):54-67. [doi: 10.1006/ceps.1999.1020] [Medline: 10620381]
68. Deci EL, Ryan RM. Intrinsic Motivation and Self-Determination in Human Behavior: Springer US; 1985. [doi: 10.1007/978-1-4899-2271-7]
69. Ryan RM, Stiller J. The social contexts of internalization: parent and teacher influences on autonomy, motivation and learning. In: Advances in Motivation and Achievement 1991, Vol. 7:115-149.
70. Qi W. Comment: Programmed to fulfill global ambitions. Nature New Biol 2017 May 25;545(7655):S53-S53. [doi: 10.1038/545S53a]
71. Pascarella ET, Terenzini PT. How College Affects Students: A Third Decade of Research: ERIC; 2005, Vol. 2. URL: https://eric.ed.gov/?id=ED498537 [accessed 2025-06-26]
72. Kruger J, Dunning D. Unskilled and unaware of it: how difficulties in recognizing one's own incompetence lead to inflated self-assessments. J Pers Soc Psychol 1999 Dec;77(6):1121-1134. [doi: 10.1037//0022-3514.77.6.1121] [Medline: 10626367]

XSL•FO
RenderX

73. Dunning D. The Dunning–Kruger effect: on being ignorant of one's own ignorance. In: Adv Exp Soc Psychol: Elsevier; 2011, Vol. 44:247-296. [doi: 10.1016/B978-0-12-385522-0.00005-6]
74. Edwards RK, Kellner KR, Sistrom CL, Magyari EJ. Medical student self-assessment of performance on an obstetrics and gynecology clerkship. Am J Obstet Gynecol 2003 Apr;188(4):1078-1082. [doi: 10.1067/mob.2003.249] [Medline: 12712114]
75. Austin Z, Gregory PAM, Galli M. "I just don't know what I'm supposed to know": Evaluating self-assessment skills of international pharmacy graduates in Canada. Research in Social and Administrative Pharmacy 2008 Jun;4(2):115-124. [doi: 10.1016/j.sapharm.2007.03.002]
76. Zhang G, Wu H, Xie A, Cheng H. The association between medical student research engagement with learning outcomes. Med Educ Online 2022 Dec;27(1):2100039. [doi: 10.1080/10872981.2022.2100039] [Medline: 35818325]
77. Sweller J. Cognitive load during problem solving: effects on learning. Cogn Sci 1988 Jun;12(2):257-285. [doi: 10.1016/0364-0213(88)90023-7]
78. Bernstein MH, Sheppard B, Bruno MA, Lay PS, Baird GL. Randomized study of the impact of AI on perceived legal liability for radiologists. NEJM AI 2025 May 22;2(6):AIoa2400785. [doi: 10.1056/AIoa2400785]
79. Tsuei SHT. How are Canadians regulating artificial intelligence for healthcare? A brief analysis of the current legal directions, challenges and deficiencies. Healthc Pap 2025 Apr;22(4):44-51. [doi: 10.12927/hcpap.2025.27571] [Medline: 40391478]
80. Blumenthal D, Patel B. The regulation of clinical artificial intelligence. NEJM AI 2024 Jul 25;1(8):AIpc2400545. [doi: 10.1056/AIpc2400545]
81. Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. Nat Med 2019 Jan;25(1):44-56. [doi: 10.1038/s41591-018-0300-7] [Medline: 30617339]
82. Reddy S. Explainability and artificial intelligence in medicine. Lancet Digit Health 2022 Apr;4(4):e214-e215. [doi: 10.1016/S2589-7500(22)00029-2] [Medline: 35337639]
83. Car J, Ong QC, Erlikh Fox T, et al. The digital health competencies in medical education framework: an international consensus statement based on a Delphi study. JAMA Netw Open 2025 Jan 2;8(1):e2453131. [doi: 10.1001/jamanetworkopen.2024.53131] [Medline: 39888625]
84. Kosmyna N, Hauptmann E, Yuan YT, et al. Your brain on ChatGPT: accumulation of cognitive debt when using an AI assistant for essay writing task. arXiv. Preprint posted online on Jun 10, 2025. [doi: 10.48550/arXiv.2506.08872]
85. Gin BC, LaForge K, Burk-Rafel J, Boscardin CK. Macy Foundation Innovation Report Part II: from hype to reality: innovators' visions for navigating AI integration challenges in medical education. Acad Med 2025;100(9S):S22-S29. [doi: 10.1097/ACM.0000000000006117]
86. Zhai X, Chu X, Chai CS, et al. A review of artificial intelligence (AI) in education from 2010 to 2020. Complexity 2021 Jan;2021(1):8812542. [doi: 10.1155/2021/8812542]
87. McCoy LG, Nagaraj S, Morgado F, Harish V, Das S, Celi LA. What do medical students actually need to know about artificial intelligence? NPJ Digit Med 2020;3:86. [doi: 10.1038/s41746-020-0294-7] [Medline: 32577533]
88. Han SP, Kumwenda B. Bridging the digital divide: promoting equal access to online learning for health professions in an unequal world. Med Educ 2025 Jan;59(1):56-64. [doi: 10.1111/medu.15455] [Medline: 38860836]
89. Zainal H, Xiao Hui X, Thumboo J, Kok Yong F. Organizational leaders' views on digital health competencies in medical education: qualitative semistructured interview study. JMIR Med Educ 2025 Mar 7;11:e64768. [doi: 10.2196/64768] [Medline: 40053774]
90. Paranjape K, Schinkel M, Nannan Panday R, Car J, Nanayakkara P. Introducing artificial intelligence training in medical education. JMIR Med Educ 2019 Dec 3;5(2):e16048. [doi: 10.2196/16048] [Medline: 31793895]
91. Kolachalama VB, Garg PS. Machine learning and medical education. NPJ Digit Med 2018;1(1):54. [doi: 10.1038/s41746-018-0061-1] [Medline: 31304333]
92. Fitas R. Inclusive education with AI: supporting special needs and tackling language barriers. SSRN. Preprint posted online on Apr 19, 2025. [doi: 10.2139/ssrn.5366398]

Abbreviations

AI: artificial intelligence
CMSS: China Medical Student Survey
DFC: Double First-Class
ICC: intraclass correlation coefficient

XSL•FO
RenderX

Edited by J Moen; submitted 14.Jul.2025; peer-reviewed by MA Amin, X Liang; revised version received 10.Nov.2025; accepted 05.Dec.2025; published 06.Jan.2026.

Please cite as:
Li C, Tsuei SHT, Wu H
AI Literacy Among Chinese Medical Students: Cross-Sectional Examination of Individual and Environmental Factors
JMIR Med Educ 2026;12:e80604
URL: https://mededu.jmir.org/2026/1/e80604
doi:10.2196/80604

© Chunqing Li, Sian Hsiang-Te Tsuei, Hongbin Wu. Originally published in JMIR Medical Education (https://mededu.jmir.org), 6.Jan.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Education, is properly cited. The complete bibliographic information, a link to the original publication on https://mededu.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

GPT-4o and OpenAI o1 Performance on the 2024 Spanish Competitive Medical Specialty Access Examination: Cross-Sectional Quantitative Evaluation Study

Pau Benito[1], MD; Mikel Isla-Jover[2], MD; Pablo González-Castro[3], MD; Pedro José Fernández Esparcia[4], MD; Manuel Carpio[5], MD; Iván Blay-Simón[6], MD; Pablo Gutiérrez-Bedia[7], MD; Maria J Lapastora[8], MD; Beatriz Carratalá[9], LLB, MEd; Carlos Carazo-Casas[10], MD

[1]Department of Preventive Medicine and Epidemiology, Clinical Institute of Medicine and Dermatology (ICMiD), Hospital Clínic de Barcelona, Rosselló, 138, ground floor, Barcelona, Spain

[10]Department of Otolaryngology, Hospital Universitario Ramón y Cajal, Madrid, Spain

[2]Department of Radiology, Hospital de Cruces, Barakaldo, Spain

[3]Department of Plastic and Reconstructive Surgery, Hospital Universitario Virgen del Rocío, Sevilla, Spain

[4]Department of Dermatology, Hospital Universitario Ramón y Cajal, Madrid, Spain

[5]Department of Endocrinology and Nutrition, Santa Lucía University General Hospital, Cartagena, Spain

[6]Department of Dermatology, Hospital Universitario Doctor Peset, Valencia, Spain

[7]Department of Neurology, Hospital Clínico San Carlos, Madrid, Spain

[8]Department of Intensive Care Medicine, Hospital Universitario 12 De Octubre, Madrid, Spain

[9]Innovation and Digital Projects Academic Department, Healthcademia, Madrid, Spain

Corresponding Author:
Pau Benito, MD
Department of Preventive Medicine and Epidemiology, Clinical Institute of Medicine and Dermatology (ICMiD), Hospital Clínic de Barcelona, Rosselló, 138, ground floor, Barcelona, Spain

Abstract

Background: In recent years, generative artificial intelligence and large language models (LLMs) have rapidly advanced, offering significant potential to transform medical education. Several studies have evaluated the performance of chatbots on multiple-choice medical examinations.

Objective: The study aims to assess the performance of two LLMs—GPT-4o and OpenAI o1—on the *Médico Interno Residente* (MIR) 2024 examination, the Spanish national medical test that determines eligibility for competitive medical specialist training positions.

Methods: A total of 176 questions from the MIR 2024 examination were analyzed. Each question was presented individually to the chatbots to ensure independence and prevent memory retention bias. No additional prompts were introduced to minimize potential bias. For each LLM, response consistency under verification prompting was assessed by systematically asking, “Are you sure?” after each response. Accuracy was defined as the percentage of correct responses compared to the official answers provided by the Spanish Ministry of Health. It was assessed for GPT-4o, OpenAI o1, and, as a benchmark, for a consensus of medical specialists and for the average MIR candidate. Subanalyses included performance across different medical subjects, question difficulty (quintiles based on the percentage of examinees correctly answering each question), and question types (clinical cases vs theoretical questions; positive vs negative questions).

Results: Overall accuracy was 89.8% (158/176) for GPT-4o and 90% (160/176) after verification prompting, 92.6% (163/176) for OpenAI o1 and 93.2% (164/176) after verification prompting, 94.3% (166/176) for the consensus of medical specialists, and 56.6% (100/176) for the average MIR candidate. Both LLMs and the consensus of medical specialists outperformed the average MIR candidate across all 20 medical subjects analyzed, with ≥80% LLMs’ accuracy in most domains. A performance gradient was observed: LLMs’ accuracy gradually declined as question difficulty increased. Slightly higher accuracy was observed for clinical cases compared to theoretical questions, as well as for positive questions compared to negative ones. Both models demonstrated high response consistency, with near-perfect agreement between initial responses and those after the verification prompting.

Conclusions: These findings highlight the excellent performance of GPT-4o and OpenAI o1 on the MIR 2024 examination, demonstrating consistent accuracy across medical subjects and question types. The integration of LLMs into medical education presents promising opportunities and is likely to reshape how students prepare for licensing examinations and change our

XSL•FO
RenderX

understanding of medical education. Further research should explore how the wording, language, prompting techniques, and image-based questions can influence LLMs' accuracy, as well as evaluate the performance of emerging artificial intelligence models in similar assessments.

(JMIR Med Educ 2026;12:e75452) doi:10.2196/75452

KEYWORDS

accuracy; artificial intelligence; GPT-4o; large language models; medical education; medical examination; Médico Interno Residente; MIR 2024 examination; OpenAI o1

Introduction

The continuous developments in recent years have positioned generative artificial intelligence (AI) as a topic of paramount public and scientific interest. These developments have resulted in the creation of gradually more sophisticated and efficient large language models (LLMs) [1].

Some of the most prominent examples are the increasingly advanced models derived from the GPT family, developed by OpenAI, which rely on deep neural networks [2]. In 2024, OpenAI released 2 highly promising models. Out of which one was GPT-4o, launched in May 2024, a multimodal model capable of processing text and image inputs and generating text outputs in real time. GPT-4o stands out in terms of rapid response times and efficiency [3]. The other one was OpenAI o1, launched in September 2024, a model only capable of processing and generating text, but trained with large-scale reinforcement learning (RL) to reason using chain of thought (CoT) and so possessing advanced reasoning capabilities, surpassing GPT-4o in competitive programming, mathematics, and scientific reasoning [4]. Despite the absence of comprehensive benchmark sets providing consistent evidence, it is reasonable to expect that the presence of sophisticated built-in reasoning-optimized mechanisms in LLMs such as OpenAI o1—trained with RL and CoT—diminishes the relative impact of complex prompting strategies. In such cases, simple zero-shot prompting may prove more effective, or at least equally effective, compared to few-shot and chain-of-thought prompting [4,5].

Chatbots for daily use have emerged to provide virtual assistance, personalized solutions, and task automation in a wide range of fields, including medical education, which has also embraced this trend [6]. Chatbots can be used as a learning aid to improve clinical skills at the undergraduate, residency training, and postgraduate levels of continuous medical education [7].

There are several previous experiences evaluating the performance of chatbots answering multiple-choice questions [7], including medical board examinations like the United States Medical Licensing Examination (USMLE) [8-10]. These assessments are helpful to understand the state of the art regarding LLMs' performance in medical examinations. Furthermore, they pose questions about and provide insightful information to shape the content and characteristics of medical education and examinations.

In Spain, similarly to the USMLE, doctors are examined prior to the beginning of their specialized training. The test is called the *Médico Interno Residente* (MIR) examination and consists of a 4.5-hour-long examination that includes 210 multiple-choice questions with 4 options and only 1 correct answer. It is held on a yearly basis. The examination serves a double purpose. On the one hand, it is used to rank physicians to assess their eligibility for competitive medical specialist training positions. On the other hand, it ensures minimum requirements are met among candidates.

Evidence suggests a strong correlation between LLM performance across different input languages and the representativeness of each language in the pre-training corpus, a relationship that extends to retrieval-augmented generation LLMs [11,12]. To our knowledge, no study has evaluated GPT-4o's performance on the Spanish MIR examination to date and, more importantly, no study has compared it to OpenAI o1, which owns enhanced reasoning capabilities that could potentially be an advantage when taking the MIR examination [4]. In a broader sense, there are few published studies that have evaluated the performance of LLMs when responding to medical questions in the Spanish language. Among them, the study by Guillen-Grima et al [13] reported a remarkable accuracy rate of 87% for GPT-4 on the 2022 MIR examination, while the study by Flores-Cohaila et al [14] showed an accuracy rate of 86% for GPT-4 on the 2022 Peruvian National Licensing Medical Examination. Other studies posing questions in Spanish from specific medical subjects have shown similar results, with performance rates of 83.7% for GPT-4o in anesthesiology [15] and 93.7% for GPT-4 in rheumatology [16].

The primary aim of this study is to assess the performance of GPT-4o and OpenAI o1 LLMs in passing the MIR examination and to compare them with the expert consensus from instructors of one of the largest MIR preparation academies (Academia AMIR) and the students' mean results. The secondary aim of this study is to compare the performance of GPT-4o, OpenAI o1, expert consensus from AMIR instructors and students by medical subjects, question difficulty, and type of question (clinical case vs theoretical question and positive vs negative question) to better characterize AI chatbots' capabilities, limitations, strengths, and weaknesses.

Methods

Study Design

This is a cross-sectional study assessing the performance of 2 LLMs (GPT-4o and OpenAI o1) in answering the MIR 2024 examination questions. The study compares the models' performance against each other and against specifically trained

XSL•FO
RenderX

humans (expert consensus from AMIR instructors and the mean results from MIR 2024 examination candidates).

MIR 2024 Examination

In Spain, there are 46 medical specialties, each requiring a specific training period of 4 to 5 years as a resident physician (MIR) in an accredited health care institution. Access to each specialty training spot depends on the national ranking of candidates. The ranking is based on a final grade which comes from the MIR examination score (90%) and the candidate's academic record (10%) [17,18].

A total of 15,114 candidates were admitted to the MIR 2024 examination, of whom 13,711 sat for the test, competing for 9007 specialty positions available in accredited healthcare institutions across Spain [18].

The examination, held on January 25, 2025, consisted of 200 multiple-choice questions, each with 4 answer choices, with only 1 correct option. The first 25 questions included linked images that were part of the questions' content and could help or be necessary to answer them. Additionally, 10 reserve questions were included to replace any disputed questions due to typographical errors, ambiguous wording, or issues with multiple or missing correct answers. Participants were given 4 hours and 30 minutes to complete the examination [17,18].

As a safeguard against academic misconduct (ie, cheating), the MIR examination is administered in several different versions each year. Each version comprises an identical question set with a varied sequence. Version 0 is established as the canonical version for scoring and for the publication of the official answer key.

Version 0 of the MIR 2024 examination was obtained from the Spanish Ministry of Health website [19]. In the final analysis, the 25 questions requiring image interpretation were excluded since one of the LLMs evaluated (OpenAI o1) does not accept image inputs. This decision aimed to ensure fair comparability across all study arms, as providing images only to human participants and the other LLM (GPT-4o) would have introduced a systematic advantage for them. The 5 questions whose objections were accepted by the Spanish Ministry of Health were also excluded from the final analysis in order to approximate the real examination as closely as possible (the Spanish Ministry of Health accepted the objection for 6 questions but one of them was linked to an image, so it was already eliminated from the analysis). Nonetheless, the performance of the 4 study arms on these questions is also reported and discussed later in the study. Of the reserve questions, the 4 questions that did not replace any challenged question were also discarded from the analysis. Final analysis included 176 questions.

Study Arms

This study compares the 4 distinct arms as follows: GPT-4o, OpenAI o1, expert consensus from AMIR instructors (henceforth "AMIR consensus"), and mean results of the MIR 2024 examination candidates (henceforth "students").

GPT-4o

This model, OpenAI's flagship in 2024, is characterized by rapid response times and efficiency. Although it is a multimodal model, in this study, only its text processing and generation capabilities were used. Image-linked questions were excluded to ensure comparability with OpenAI o1 model.

OpenAI o1

Designed to use large-scale RL and CoT reasoning, this model exhibits advanced reasoning capabilities, which could be particularly useful for answering questions in the MIR examination.

AMIR Consensus

Academia AMIR is a private, for-profit educational company operating in Spain, Portugal, and several Latin American countries, providing postgraduate health sciences training. Its core activities include preparing candidates for official examinations such as the MIR examination. The company employs faculty members who deliver these courses but are independent from the MIR examination process. They neither contribute to the examination's development nor belong to any public organization involved in its preparation. A panel of these faculty—at least 2 per medical specialty—collaboratively answered the entire MIR 2024 examination after its administration and official content release by the Spanish Ministry of Health. Meeting in a hybrid format (combining in-person and remote participation), they established a consensus through discussion answering within 4.5 hours, mirroring the time allotted to candidates. The goal was to give candidates prompt performance feedback before the official answer key was published. Therefore, the faculty had unrestricted access to textbooks, scientific literature, and LLMs, though reported use was minimal and reserved for clarifying ambiguous questions prior to group consensus. This process produced an expert consensus answer list for the MIR 2024 examination.

Students

Candidates who took the examination were encouraged to submit their answer templates to EstimAMIR, an online platform developed by Academia AMIR. This platform provides students with a preliminary assessment of their results, initially using "AMIR consensus" answers and later incorporating the provisional and definitive correct answers published by the Spanish Ministry of Health. The platform also estimated each student's ranking position based on the sample of candidates available. The mean results of the "students", as well as the percentage of correct answers for each question, were obtained from this platform (based on 5066 answer templates submitted). All data were appropriately anonymized and aggregated.

Data Collection

The 185 text-based questions from the MIR 2024 examination were collected and transcribed verbatim in Spanish into the dialogue interface of both GPT-4o and OpenAI o1. A ChatGPT Plus license was used to access the GPT-4o and OpenAI o1. The models were used with their default settings, with no modifications to parameters such as temperature or output variation. Each multiple-choice question was followed by the

XSL•FO
RenderX

4 possible answer choices (1, 2, 3, and 4), which were manually entered and separated by single spaces. No pretraining or standardized instructions were provided, adhering strictly to a zero-shot prompting approach to minimize potential bias. Henceforth, the results generated using this prompt will be designated as the first iteration results.

Questions were presented to the chatbots individually, with a new dialogue initiated for each question to ensure independence and prevent memory retention bias. To assess response consistency, chatbots were systematically challenged with the verification prompt "Are you sure?" after each answer, which served as a single CoT prompt. Hereafter, we referred to the results obtained with this prompt as the second iteration results. For GPT-4o, internet access was disabled during testing.

All responses were recorded in a spreadsheet. Once the definitive official answers were published by the Spanish Ministry of Health, any challenged or unused reserve questions were excluded from the final analysis.

Main Endpoint and Additional Analysis

Overview

The primary endpoint was the percentage of correct answers per study arm. The definitive official answers published by the Spanish Ministry of Health served as the gold standard for determining accuracy within each study arm. The secondary endpoints included comparisons of study arm performance based on medical subjects, question difficulty, and question type (clinical case vs theoretical question and positive vs negative question). Categorization was conducted as described below.

Medical Subject

Questions were classified into the following categories such as gastroenterology and general surgery, endocrinology, infectious diseases and microbiology, miscellaneous and basic sciences, neurology and neurosurgery, cardiology and cardiovascular surgery, gynecology and obstetrics, orthopedic surgery, pediatrics, nephrology, respiratory medicine and thoracic surgery, rheumatology, hematology, psychiatry, immunology, urology, dermatology, ophthalmology, otorhinolaryngology, and statistics and epidemiology.

Question Difficulty

Difficulty was categorized based on the percentage of examinees who correctly answered each question, using data from EstimAMIR (very difficult: 0% - 20% correct responses; difficult: 21% - 40% correct responses; intermediate: 41% - 60% correct responses; easy: 61% - 80% correct responses; very easy: 81% - 100% correct responses).

Theoretical Question Versus Clinical Case

Questions were classified as theoretical (requiring a direct answer based exclusively on theoretical knowledge) and clinical case (presenting a clinical scenario from which the possible answers emerged).

Positive Versus Negative Questions

Questions were classified based on whether they asked for the correct answer (or the next appropriate step) or the incorrect answer (or the step that should not be taken).

Statistical Analysis

Comparisons between study arms were performed using chi-squared tests or the Fisher exact test, where applicable. The Benjamini-Hochberg method was applied to statistically adjust for multiple comparisons. Differences between groups were considered statistically significant if $P<.05$. We assessed the consistency of responses from GPT-4o and OpenAI o1 to a verification prompt ("Are you sure?"). Consistency was measured using both the simple agreement percentage and the Cohen κ coefficient between the answers provided before and after the prompt. All statistical analyses were performed using *R* version 4.4.1 (R Foundation for Statistical Computing).

Ethical Considerations

In the absence of a formal ethics committee at Academia AMIR, an ad hoc data ethical oversight panel approved a comprehensive data use protocol (internal reference: AMIR-ETH-2025-11-05-v1.0). At enrollment, all students provided data-use consent, acknowledging that their irreversibly anonymized and aggregated data could be used by Academia AMIR for statistical, commercial, educational, research, and product improvement purposes. The panel determined that the study qualified for exemption from institutional review board approval, as it involved secondary data that were aggregated, processed without human intervention, and contained no identifiable information, in accordance with principles of risk proportionality and data minimization. Participant privacy and confidentiality were safeguarded through irreversible anonymization and aggregation prior to investigator access, data minimization, role-based access controls, and encryption of data both in transit and at rest within corporate repositories. No participants received compensation for their participation in the study.

Results

Global Performance

A flowchart of the exclusion criteria for the question selection process is shown in Figure 1.

XSL•FO
RenderX

Figure 1. Flowchart of the exclusion criteria for the question selection process.

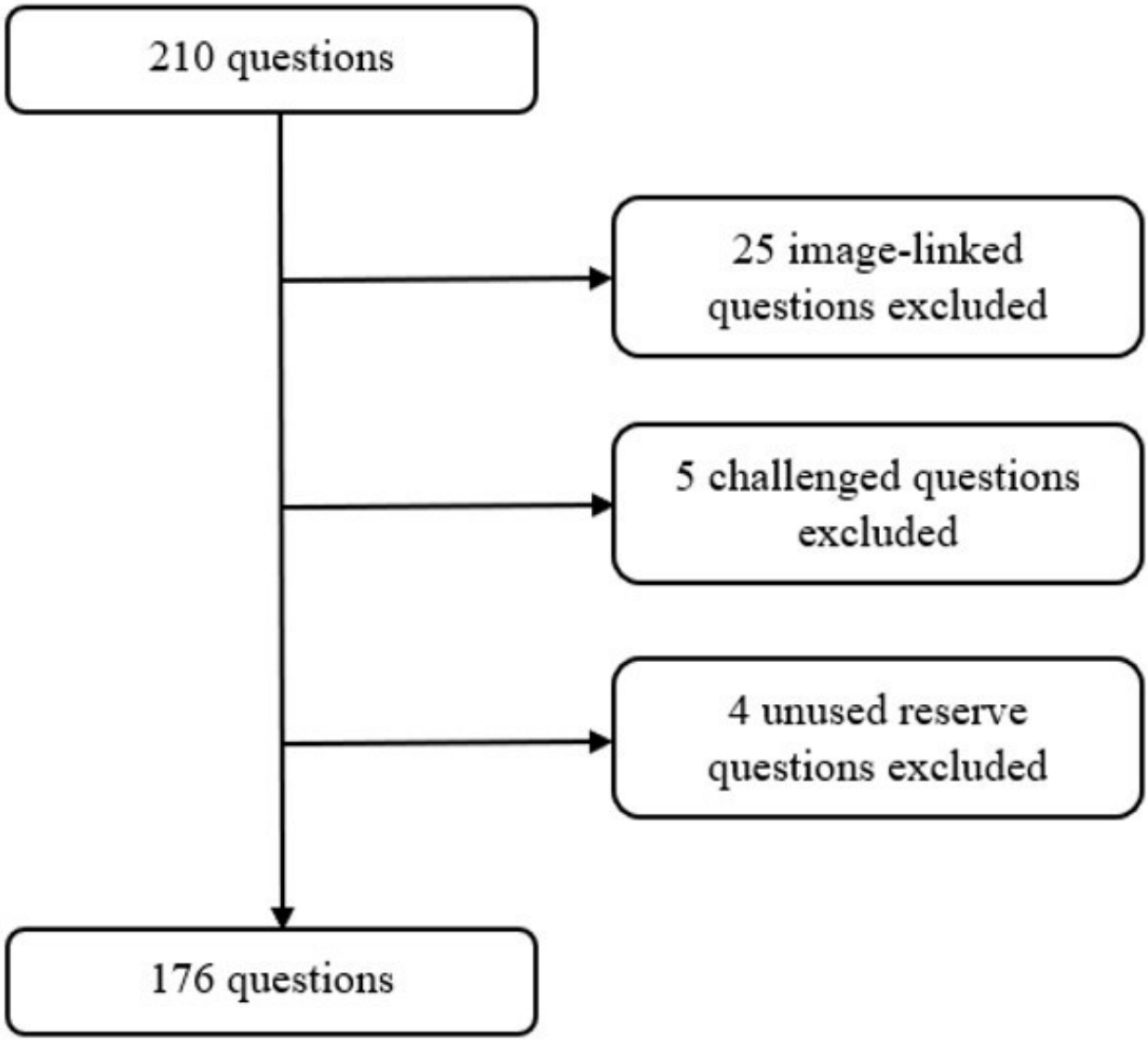


Table 1 presents the accuracy of GPT-4o, OpenAI o1, the AMIR consensus, and students when taking the entire examination, excluding the discarded questions. GPT-4o achieved an accuracy of 89.8% (158/176) in the first iteration, which slightly increased to 90.9% (160/176) in the second iteration. Similarly, OpenAI o1's accuracy was 92.6% (163/176) in the first iteration and improved to 93.2% (164/176) in the second iteration. The AMIR instructors' consensus obtained the highest score among the study arms, with 94.3% (166/176). The mean score of the EstimAMIR-submitted templates was 56.6% (100/176).

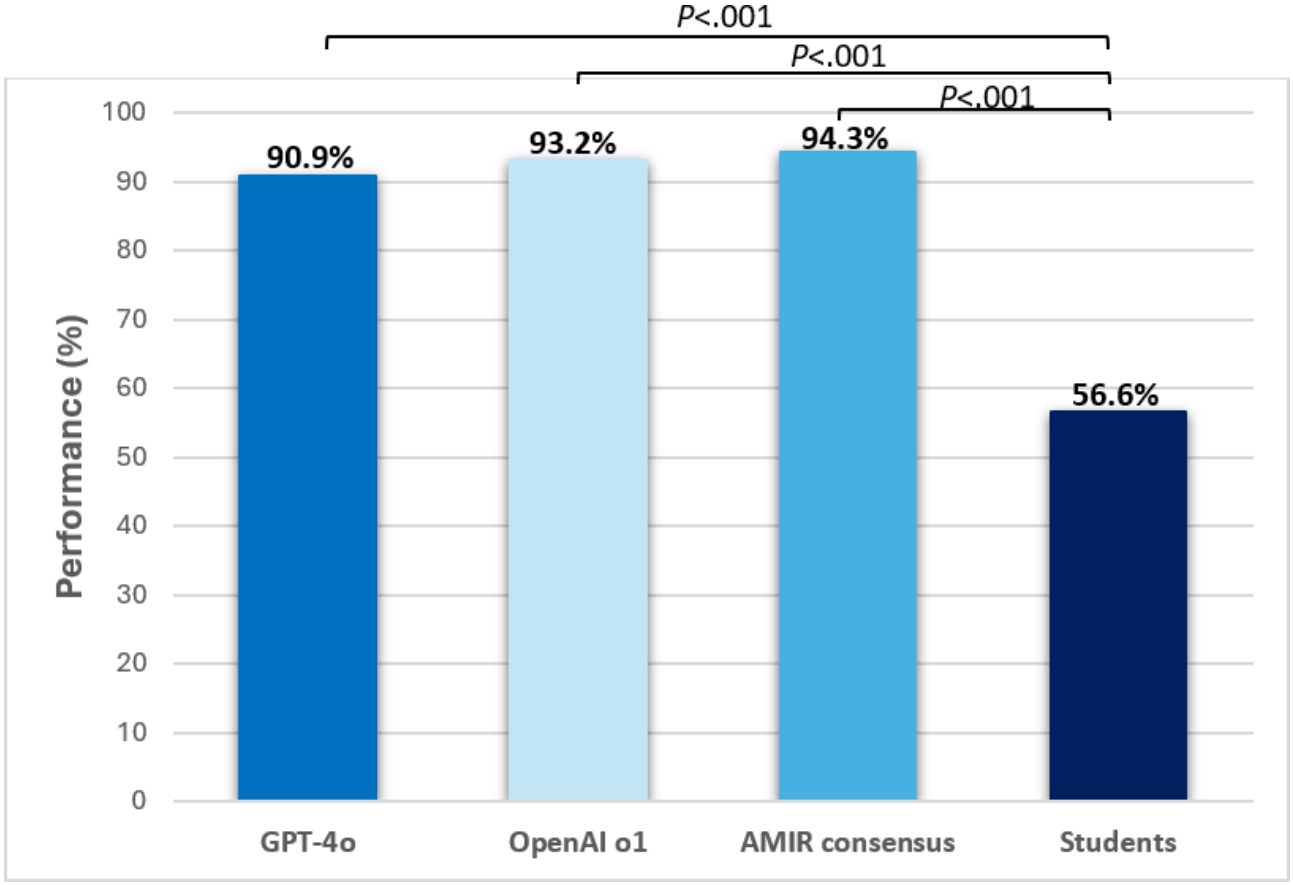
Table . Global performance of GPT-4o (first and second iteration), OpenAI o1 (first and second iteration), AMIR consensus, and students.

	Absolute and relative performance	
	Correct answers, n (%)	Incorrect answers, n (%)
GPT-4o		
First iteration	158 (89.8)	18 (10.2)
Second iteration	160 (90.9)	16 (9.1)
OpenAI o1		
First iteration	163 (92.6)	13 (7.4)
Second iteration	164 (93.2)	12 (6.8)
AMIR consensus	166 (94.3)	10 (5.7)
Students	100 (56.6)	76 (43.4)

XSL•FO
RenderX

Figure 2 compares the accuracy of GPT-4o and OpenAI o1 in their second iterations, along with the AMIR consensus and students. GPT-4o, OpenAI o1, and the AMIR consensus achieved significantly higher accuracy scores than the average student (in all cases $P<.001$); however, differences between these 3 arms did not reach statistical significance ($P=.22$ for GPT-4o vs OpenAI o1; $P=.07$ for GPT-4o vs AMIR consensus; $P=.75$ for OpenAI o1 vs AMIR consensus).

Figure 2. Global performance of GPT-4o (2nd iteration), OpenAI o1 (2nd iteration), AMIR consensus, and students.



From the final analysis, 5 challenged questions were excluded. On these items, GPT-4o achieved an accuracy of 100% (5/5) across both iterations. OpenAI o1 scored 80% (4/5) in both iterations, without modifying any of its responses. The AMIR consensus achieved an accuracy of 40% (2/5). When restricted to these 5 questions, the students' mean score was 31.3% (SD 16.7%).

Medical Subjects

The heatmap in Figure 3 presents a comparative analysis of study arm performance across different medical subjects. Both LLMs (GPT-4o and OpenAI o1, in their second iterations) and the AMIR consensus outperformed the average student in all 20 medical subjects analyzed.

XSL•FO
RenderX

Figure 3. Heatmap comparing the performance of generative pre-trained transformer 4o (2nd iteration), OpenAI o1 (2nd iteration), and AMIR consensus and students by medical subject.

Medical subject	Number of questions	GPT-4o	OpenAI o1	AMIR consensus	Students	Heatmap scale
Gastroenterology and general surgery	18	83%	89%	89%	49.7%	100%
Endocrinology	14	93%	93%	86%	64.4%	95%
Infectious diseases and microbiology	14	79%	86%	86%	49.5%	90%
Miscellaneous and basic sciences	13	100%	100%	100%	44.1%	85%
Neurology and neurosurgery	12	100%	83%	92%	53.9%	80%
Cardiology and cardiovascular surgery	11	100%	100%	100%	47.1%	75%
Gynecology and obstetrics	10	90%	100%	100%	60.2%	70%
Orthopedic surgery	10	90%	90%	100%	54.2%	65%
Pediatrics	10	70%	100%	100%	53.0%	60%
Nephrology	9	100%	100%	100%	63.3%	55%
Respiratory medicine and thoracic surgery	9	100%	100%	100%	62.9%	50%
Rheumatology	9	89%	89%	89%	56.5%	45%
Hematology	6	100%	83%	100%	69.3%	40%
Psychiatry	6	83%	83%	67%	65.4%	35%
Immunology	5	100%	100%	100%	71.6%	30%
Urology	5	80%	100%	100%	68.8%	25%
Dermatology	4	100%	100%	100%	69.6%	20%
Ophthalmology	3	100%	100%	100%	55.0%	15%
Otorhinolaryngology	4	100%	100%	100%	64.2%	10%
Statistics and epidemiology	4	75%	75%	100%	51.2%	5%
Total	**176**	**90.9%**	**93.2%**	**94.3%**	**56.6%**	0%

GPT-4o achieved an accuracy below 80% in only 3 subjects: infectious diseases and microbiology (11/14, 79%), pediatrics (7/10, 70%), and statistics and epidemiology (3/4, 75%). OpenAI o1 fell below 80% accuracy only in statistics and epidemiology, with a single error (3/4, 75%). The AMIR consensus exhibited an accuracy lower than 80% in psychiatry (4/6, 67%). Average student performance, based on 5066 EstimAMIR-submitted templates, ranged from 44.1% in miscellaneous and basic sciences to 71.6% in immunology.

Question Difficulty

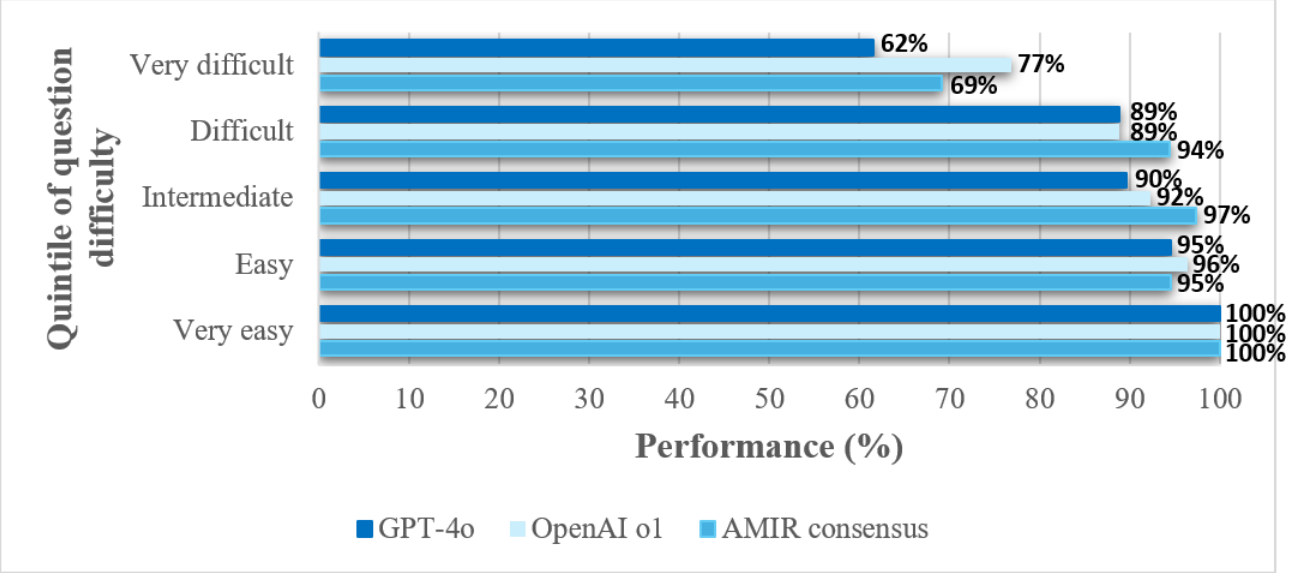
As shown in Table 2 and Figure 4 the performance of GPT-4o, OpenAI o1, and the AMIR consensus across quintiles of question difficulty, defined based on student performance. A statistically significant gradient is observed among these 3 study arms, with accuracy decreasing as question difficulty increases (crude *P* values: GPT-4o, *P*=.003; OpenAI o1, *P*=.04; AMIR consensus, *P*=.008; Benjamini-Hochberg adjusted *P* values: GPT-4o, *P*=.03; OpenAI o1, *P*=.10—the only case not reaching statistical significance; AMIR consensus, *P*=.03). The decline in performance is particularly pronounced in the highest difficulty quintile. Differences between study arms within each difficulty quintile do not reach statistical significance.

Table . Absolute and relative performance of GPT-4o (second iteration), OpenAI o1 (second iteration), and AMIR consensus by quintiles of question difficulty defined by students' performance.

Question difficulty	Number of questions	GPT-4o, n (%)	OpenAI o1, n (%)	AMIR consensus, n (%)
Very easy	32	32 (100)	32 (100)	32 (100)
Easy	56	53 (95)	54 (96)	53 (95)
Intermediate	39	35 (90)	36 (92)	38 (97)
Difficult	36	32 (89)	32 (89)	34 (94)
Very difficult	13	8 (62)	10 (77)	9 (69)
Total	176	160 (90.9)	164 (93.2)	166 (94.3)

XSL•FO
RenderX

Figure 4. Performance of GPT-4o (2nd iteration), OpenAI o1 (2nd iteration), and AMIR consensus by quintiles of question difficulty defined by students' performance.



Clinical Cases versus Theoretical Questions

As shown in Table 3 the performance of GPT-4o, OpenAI o1, the AMIR consensus, and students when answering clinical cases versus theoretical questions. Overall, a slightly higher accuracy was observed for clinical cases compared to theoretical questions, although these differences do not reach statistical significance in any study arm. Statistically significant differences between study arms both for clinical cases and theoretical questions were observed only when the students' arm was included in the analysis.

Table . Performance of GPT-4o (second iteration), OpenAI o1 (second iteration), AMIR consensus, and students by questions being clinical cases or theoretical questions.

	Clinical cases (n=105), n (%)	Theoretical questions (n=71), n (%)	*P* value
GPT-4o	98 (93.3)	62 (87)	.55
OpenAI o1	99 (94.2)	65 (92)	.68
AMIR consensus	99 (94.2)	67 (94)	.99
Students	62 (59)	38 (54)	.68

Positive Versus Negative Questions

Table 4 shows the performance of GPT-4o, OpenAI o1, the AMIR consensus, and students when answering positive versus negative questions. Overall, accuracy was higher for positive questions than for negative ones, with the difference reaching statistical significance only for GPT-4o (*P*=.01). Statistically significant differences between study arms for both positive and negative questions are observed only when the students' arm is included in the analysis.

Table . Performance of GPT-4o (second iteration), OpenAI o1 (second iteration), AMIR consensus, and students by questions being positive or negative.

	Positive questions (n=140), n (%)	Negative questions (n=36), n (%)	*P* value
GPT-4o	132 (94.2)	28 (78)	.01
OpenAI o1	132 (94.2)	32 (89)	.40
AMIR consensus	132 (94.2)	34 (94)	1
Students	82 (58.5)	18 (49)	.55

Response Consistency

Response consistency was assessed using the simple agreement percentage between the first and second iterations of GPT-4o (172/176, 97.7%) and OpenAI o1 (170/176, 96.6%). The Cohen κ coefficient was 0.97 for GPT-4o and 0.95 for OpenAI o1, indicating almost perfect agreement (*P*<.001 in both cases).

When analyzing individually the questions in which there was no concordance between the first and second iterations, it was observed that, for GPT-4o, 4 initially incorrect responses were modified: in 2 cases, the second response was also incorrect, while in the other 2 cases, the second response became correct. For OpenAI o1, 5 initially incorrect responses were modified: in 3 cases, the second response was again incorrect, and in 2 cases, the second response became correct. In addition, 1

XSL•FO
RenderX

initially correct response was modified, with the second response becoming incorrect.

Discussion

Principal Results

This study highlights the exceptional performance of both LLMs analyzed—GPT-4o and OpenAI o1—on the MIR 2024 examination. Both models achieved or exceeded a 90% accuracy rate, significantly outperforming the average human candidate as well as the top 10% of examinees [18]. The expert consensus from AMIR instructors yielded even higher accuracy. Although this result should be interpreted in the context of unrestricted access to textbooks, scientific literature, and AI tools such as GPT, the reported use of these resources was minimal and reserved for clarifying ambiguous questions, never substituting for the group discussion and consensus process for each item. Results from the expert consensus suggest the added value of human expertise when synergistically combined with AI capabilities.

The challenged questions were excluded from the final analysis to remain faithful to the actual examination. Upon examination, these proved to be difficult items (on average, candidates answered them correctly in 31.3% of cases, compared with 56.6% for the other questions), which the LLMs managed more accurately than the human experts (5/5 for GPT-4o, 4/5 for OpenAI o1, and 2/5 for AMIR instructors).

These results were consistent across the different medical subjects analyzed. Interestingly, when question difficulty was assessed based on human performance, a similar trend was observed in the LLMs, with accuracy decreasing as question difficulty increased. Additionally, a slightly higher accuracy was observed for clinical cases compared to theoretical questions, as well as for positive questions compared to negative ones. This resemblance to human reasoning and performance could be rooted in the input used to train LLMs.

Both GPT-4o and OpenAI o1 demonstrated great consistency in their answers, with statistically significant near-perfect agreement between the first and second iterations. Furthermore, it is particularly interesting to note that, of the 10 responses that were altered between the first and second iterations (4 for GPT-4o and 6 for OpenAI o1), 9 were initially incorrect (with 4 of these changing to a correct response), and only 1 initially correct response was changed to an incorrect one. It is remarkable in favor of these LLMs that, considering their exceptional accuracy in the first iteration, the few changes occurring in the second iteration almost exclusively involved some of the few initially incorrect responses.

Comparison With Prior Work

Several previous studies have evaluated the performance of GPT-3.5 and GPT-4 across different medical disciplines, as well as on national medical board examinations [7]. For instance, Gilson et al assessed ChatGPT's performance on various sets of USMLE step 1 and step 2 questions, reporting an accuracy range of 42% to 64% [8], which aligns with another study that found a 56% accuracy rate on a set of USMLE step 1-style questions [9]. Knoedler et al examined ChatGPT-3.5 and ChatGPT-4 on USMLE step 3 questions, reporting 57% accuracy for GPT-3.5 and 85% for GPT-4 [10]. Takagi et al evaluated these models on the Japanese Medical Licensing Examination, finding 51% accuracy for GPT-3.5 and 80% for GPT-4 [20]. Meyer et al conducted a similar study on the written German medical licensing examination, with accuracy rates of 58% for GPT-3.5 and 85% for GPT-4 [21]. A study by Prazeres [22] on the Portuguese national examination for access to specialized training reported 54% accuracy for GPT-3.5 turbo and 65% for GPT-4o mini. Guillen-Grima et al [13] published the perhaps most comparable study, as they compared GPT-3.5 and GPT-4 on the MIR 2022 examination. Accuracy rates were 63% for GPT-3.5 and 87% for GPT-4. Interestingly, our results show the highest accuracy for LLMs among all the aforementioned studies, almost matching the consensus from expert human instructors. This may be a result of the gradual development of LLMs with time. It poses relevant questions regarding how medical education and examinations should be shaped in the future, both in terms of content and the skills that are underscored. As stated in a previous editorial, these results make it important to consider the necessity of more emphasis on soft skills and critical thinking rather than plain memorization [23].

Strengths

This study offers new insights into the accuracy of GPT-4o and OpenAI o1 in a national medical specialty access examination. To date, the only comparable research published in an indexed journal that we have identified is a recent study by Liu et al [24], which evaluated GPT-4o's performance on the Japanese national medical examination, reporting an accuracy of 89%.

Moreover, this study reinforces the trend that each newly developed LLM exhibits improved accuracy compared to its predecessors. Additionally, our secondary analysis proved that the accuracy of both GPT-4o and OpenAI o1 aligned with difficulty levels of questions based on human candidates' performance.

Limitations

This study has several limitations. First, increasing the sample size—that is, including a larger number of questions in the analysis—would have provided a more robust insight into the different subanalyses performed. For instance, it would have allowed us to investigate whether performance differences of LLMs compared to human experts—such as those suggested in unique medical subjects like psychiatry—are truly meaningful or just the results of random variation. Second, additional secondary analyses could have been of interest, such as examining the relationship between the number of words or characters in each question and the performance of LLMs; the influence of specific expressions or wording styles on model accuracy; the impact of different languages on performance; and the effect of alternative prompting formulas on accuracy. Third, although image-based questions are part of the MIR examination, they were not included in this study because OpenAI o1 does not support image inputs, and fair comparability between LLMs was prioritized. This decision reduces methodological bias—LLMs are not artificially penalized for lacking multimodal capabilities—and increases the internal

XSL•FO
RenderX

validity of between-arms comparisons. However, it may reduce the representativeness of the performance evaluation and limit the generalizability of our findings to the actual test setting, where visual interpretation is an integral component of clinical reasoning. Previous studies suggest LLMs may exhibit reasonable performance on image-based questions even without access to the image itself [25]. Fourth, the AMIR consensus may not represent a pure benchmark of human expert knowledge, as experts had access to textbooks and generative AI. However, the faculty use of these resources was minimal and strictly advisory, with all final answers determined by expert discussion and consensus, indicating that the potential for significant bias was low. Fifth, the field of LLMs is continuously evolving, and several new models have been released in recent months that were not analyzed in this study, including GPT-4.5 by OpenAI [26], DeepSeek-R1 by DeepSeek-AI [27], Qwen 2.5 by Alibaba [28], LlaMa 3.2 by Meta AI [29], and Claude 3.7 Sonnet by Anthropic [30], among others. Sixth, student results were self-reported, which could be a source of bias. Finally, caution should be exercised when generalizing LLM accuracy on the MIR examination to other national medical licensing examinations or to different fields and tasks within medical education.

Conclusions

This study highlights the excellent performance of the two analyzed LLMs—GPT-4o and OpenAI o1—on the MIR 2024 examination, demonstrating strong consistency across different medical subjects and types of questions, as well as between first and second iterations.

The integration of LLMs into medical education is promising and likely to revolutionize the field and change our understanding of medical education. Further research is needed to explore how wording, language, prompting techniques, and image-based questions influence LLM accuracy in national medical licensing examinations, as well as to assess the performance of other emerging models. More research is also needed to better understand the potential usefulness of these tools as learning assistants in broader educational contexts.

Acknowledgments

The authors thank the students who entered their MIR examination templates into the EstimAMIR application, which provided a benchmark for assessing the performance of the LLMs evaluated in this study. The authors also appreciate the interest of the AMIR Academy faculty in this study, especially the instructors, whose responses to the MIR 2024 examination questions allowed us to design the comparison group called AMIR consensus.

Funding

The expenses associated with the preparation and publication of this study were funded by Healthcademia.

Data Availability

The data from this study are available upon reasonable request.

Authors' Contributions

Conceptualization: CCC (lead); PB and MIJ (equal)
Data curation: PB and CCC (lead); MIJ, PGC, PJFE, MCP, IBS, PGB, MJL, and BCO (equal)
Formal analysis: PB (lead), MIJ (supporting)
Project administration: CCC (lead), PB (supporting)
Writing – original draft: PB
Writing – review & editing: PB, CCC, MIJ, PGC, PJFE, MCP, IBS, PGB, MJL, BCO (equal)

Conflicts of Interest

None declared.

References

1. Zhao WX, Zhou K, Li J, Tang T, Wang X, Hou Y, et al. A survey of large language models. arXiv. Preprint posted online on Mar 31, 2023. [doi: 10.48550/arXiv.2303.18223]
2. Radford A, Narasimhan K, Salimans T, Sutskever I. Improving language understanding by generative pre-training. : OpenAI; 2018 URL: https://cdn.openai.com/research-covers/language-unsupervised/language_understanding_paper.pdf [accessed 2025-12-10]
3. OpenAI, Achiam J, Adler S, Agarwal S, Ahmad L, Akkaya I, et al. GPT-4 technical report. arXiv. Preprint posted online on Mar 15, 2023. [doi: 10.48550/arXiv.2303.08774]
4. OpenAI, Jaech A, Kalai A, Lerer A, Richardson A, El-Kishky A, et al. OpenAI o1 system card. arXiv. Preprint posted online on Dec 21, 2024. [doi: 10.48550/arXiv.2412.16720]
5. Wang G, Sun Z, Ye S, et al. Do advanced language models eliminate the need for prompt engineering in software engineering? ACM Trans Softw Eng Methodol 2025. [doi: 10.1145/3771933]

XSL•FO
RenderX

6. Caldarini G, Jaf S, McGarry K. A literature survey of recent advances in chatbots. Information 2022;13(1):41. [doi: 10.3390/info13010041]
7. Hallquist E, Gupta I, Montalbano M, Loukas M. Applications of artificial intelligence in medical education: a systematic review. Cureus 2025 Mar;17(3):e79878. [doi: 10.7759/cureus.79878] [Medline: 40034416]
8. Gilson A, Safranek CW, Huang T, et al. How does ChatGPT perform on the United States Medical Licensing Examination (USMLE)? The implications of large language models for medical education and knowledge assessment. JMIR Med Educ 2023 Feb 8;9:e45312. [doi: 10.2196/45312] [Medline: 36753318]
9. Knoedler L, Knoedler S, Hoch CC, et al. In-depth analysis of ChatGPT's performance based on specific signaling words and phrases in the question stem of 2377 USMLE step 1 style questions. Sci Rep 2024 Jun 12;14(1):13553. [doi: 10.1038/s41598-024-63997-7] [Medline: 38866891]
10. Knoedler L, Alfertshofer M, Knoedler S, et al. Pure wisdom or Potemkin villages? A comparison of ChatGPT 3.5 and ChatGPT 4 on USMLE step 3 style questions: quantitative analysis. JMIR Med Educ 2024 Jan 5;10:e51148. [doi: 10.2196/51148] [Medline: 38180782]
11. Li Z, Shi Y, Liu Z, Yang F, Liu N, Du M. Quantifying multilingual performance of large language models across languages. arXiv. Preprint posted online on Apr 17, 2024. [doi: 10.48550/arXiv.2404.11553]
12. Gao Y, Xiong Y, Gao X, Jia K, Pan J, Bi Y, et al. Retrieval-augmented generation for large language models: a survey. arXiv. Preprint posted online on Dec 18, 2023. [doi: 10.48550/arXiv.2312.10997]
13. Guillen-Grima F, Guillen-Aguinaga S, Guillen-Aguinaga L, et al. Evaluating the efficacy of ChatGPT in navigating the Spanish medical residency entrance examination (MIR): promising horizons for AI in clinical medicine. Clin Pract 2023 Nov 20;13(6):1460-1487. [doi: 10.3390/clinpract13060130] [Medline: 37987431]
14. Flores-Cohaila JA, García-Vicente A, Vizcarra-Jiménez SF, et al. Performance of ChatGPT on the Peruvian national licensing medical examination: cross-sectional study. JMIR Med Educ 2023 Sep 28;9:e48039. [doi: 10.2196/48039] [Medline: 37768724]
15. Altermatt FR, Neyem A, Sumonte NI, et al. Evaluating GPT-4o in high-stakes medical assessments: performance and error analysis on a Chilean anesthesiology exam. BMC Med Educ 2025 Oct 27;25(1):1499. [doi: 10.1186/s12909-025-08084-9] [Medline: 41146119]
16. Madrid-García A, Rosales-Rosado Z, Freites-Nuñez D, et al. Harnessing ChatGPT and GPT-4 for evaluating the rheumatology questions of the Spanish access exam to specialized medical training. Sci Rep 2023 Dec 13;13(1):22129. [doi: 10.1038/s41598-023-49483-6] [Medline: 38092821]
17. Order SND/888/2024 of 14 august, approving the number of available positions and announcing the 2024 competitive selection examinations for access in 2025 to specialized health training positions for university degree holders in medicine, pharmacy, nursing, and in the fields of psychology, chemistry, biology, and physics [Article in Spanish]. : Boletín Oficial del Estado; 2024 URL: https://www.boe.es/boe/dias/2024/08/23/pdfs/BOE-A-2024-17246.pdf [accessed 2025-01-07]
18. The Ministry of Health publishes the final list of results of the specialized health training examinations [Website in Spanish]. Ministry of Health, Spain. URL: https://www.sanidad.gob.es/gabinete/notasPrensa.do?id=6632 [accessed 2025-03-12]
19. Consulting previous examination booklets—search by examination session [Website in Spanish]. Ministry of Health, Spain. URL: https://fse.sanidad.gob.es/fseweb/#/principal/datosAnteriores/cuadernosExamen [accessed 2025-01-26]
20. Takagi S, Watari T, Erabi A, Sakaguchi K. Performance of GPT-3.5 and GPT-4 on the Japanese medical licensing examination: comparison study. JMIR Med Educ 2023 Jun 29;9:e48002. [doi: 10.2196/48002] [Medline: 37384388]
21. Meyer A, Riese J, Streichert T. Comparison of the performance of GPT-3.5 and GPT-4 with that of medical students on the written German medical licensing examination: observational study. JMIR Med Educ 2024 Feb 8;10:e50965. [doi: 10.2196/50965] [Medline: 38329802]
22. Prazeres F. ChatGPT's performance on Portuguese medical examination questions: comparative analysis of ChatGPT-3.5 Turbo and ChatGPT-4o Mini. JMIR Med Educ 2025 Mar 5;11:e65108. [doi: 10.2196/65108] [Medline: 40043219]
23. Mbakwe AB, Lourentzou I, Celi LA, Mechanic OJ, Dagan A. ChatGPT passing USMLE shines a spotlight on the flaws of medical education. PLOS Digit Health 2023 Feb;2(2):e0000205. [doi: 10.1371/journal.pdig.0000205] [Medline: 36812618]
24. Liu M, Okuhara T, Dai Z, et al. Evaluating the effectiveness of advanced large language models in medical knowledge: a comparative study using the Japanese national medical examination. Int J Med Inform 2025 Jan;193:105673. [doi: 10.1016/j.ijmedinf.2024.105673] [Medline: 39471700]
25. Gravina AG, Pellegrino R, Palladino G, Imperio G, Ventura A, Federico A. Charting new AI education in gastroenterology: cross-sectional evaluation of ChatGPT and perplexity AI in medical residency exam. Dig Liver Dis 2024 Aug;56(8):1304-1311. [doi: 10.1016/j.dld.2024.02.019] [Medline: 38503659]
26. OpenAI GPT-4.5 system card. : OpenAI; 2025 URL: https://cdn.openai.com/gpt-4-5-system-card-2272025.pdf [accessed 2026-01-06]
27. DeepSeek-AI, Guo D, Yang D, Zhang H, Song J, Zhang R, et al. DeepSeek-R1: incentivizing reasoning capability in llms via reinforcement learning. arXiv. Preprint posted online on Jan 22, 2025. [doi: 10.48550/arXiv.2501.12948]
28. Qwen, Yang A, Yang B, Zhang B, et al. Qwen2.5 technical report. arXiv. Preprint posted online on Dec 19, 2024. [doi: 10.48550/arXiv.2412.15115]

XSL•FO
RenderX

29. Grattafiori A, Dubey A, et al. The llama 3 herd of models. arXiv. Preprint posted online on Jul 31, 2024. [doi: 10.48550/arXiv.2407.21783]
30. Claude 3.7 sonnet system card. : Anthropic URL: https://assets.anthropic.com/m/785e231869ea8b3b/original/claude-3-7-sonnet-system-card.pdf [accessed 2026-01-07]

ABBREVIATIONS

AI: artificial intelligence
CoT: chain of thought
LLM: large language model
MIR: Médico Interno Residente
RL: reinforcement learning
USMLE: United States medical licensing examination

Edited by B Lesselroth; submitted 03.Apr.2025; peer-reviewed by R Pellegrino, T Watari; revised version received 17.Nov.2025; accepted 25.Nov.2025; published 12.Jan.2026.

Please cite as:
Benito P, Isla-Jover M, González-Castro P, Fernández Esparcia PJ, Carpio M, Blay-Simón I, Gutiérrez-Bedia P, Lapastora MJ, Carratalá B, Carazo-Casas C
GPT-4o and OpenAI o1 Performance on the 2024 Spanish Competitive Medical Specialty Access Examination: Cross-Sectional Quantitative Evaluation Study
JMIR Med Educ 2026;12:e75452
URL: https://mededu.jmir.org/2026/1/e75452
doi:10.2196/75452

© Pau Benito, Mikel Isla-Jover, Pablo González-Castro, Pedro José Fernández Esparcia, Manuel Carpio, Iván Blay-Simón, Pablo Gutiérrez-Bedia, Maria J Lapastora, Beatriz Carratalá, Carlos Carazo-Casas. Originally published in JMIR Medical Education (https://mededu.jmir.org), 12.Jan.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Education, is properly cited. The complete bibliographic information, a link to the original publication on https://mededu.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Adaptation of the Japanese Version of the 12-Item Attitudes Towards Artificial Intelligence Scale for Medical Trainees: Multicenter Development and Validation Study

Hirohisa Fujikawa[1,2,3], MD, PhD; Hirotake Mori[1], MD, PhD; Kayo Kondo[4], MA, PhD; Yuji Nishizaki[1,5], MD, MPH, PhD; Yuichiro Yano[1], MD, PhD; Toshio Naito[1], MD, PhD

[1]Department of General Medicine, Juntendo University Faculty of Medicine, 2-1-1 Hongo, Bunkyo-ku, Tokyo, Japan

[2]Department of Medical Education Studies, International Research Center for Medical Education, Graduate School of Medicine, The University of Tokyo, Bunkyo-ku, Tokyo, Japan

[3]Center for General Medicine Education, School of Medicine, Keio University, Shinjuku-ku, Tokyo, Japan

[4]School of Modern Languages and Cultures, Durham University, Durham, United Kingdom

[5]Division of Medical Education, Juntendo University School of Medicine, Bunkyo-ku, Tokyo, Japan

Corresponding Author:
Hirohisa Fujikawa, MD, PhD
Department of General Medicine, Juntendo University Faculty of Medicine, 2-1-1 Hongo, Bunkyo-ku, Tokyo, Japan

Abstract

Background: In the current era of artificial intelligence (AI), use of AI has increased in both clinical practice and medical education. Nevertheless, it is probable that perspectives on the prospects and risks of AI vary among individuals. Given the potential for attitudes toward AI to significantly influence its integration into medical practice and educational initiatives, it is essential to assess these attitudes using a validated tool. The recently developed 12-item Attitudes Towards Artificial Intelligence scale has demonstrated good validity and reliability for the general population, suggesting its potential for extensive use in future studies. However, to our knowledge, there is currently no validated Japanese version of the scale. The lack of a Japanese version hinders research and educational efforts aimed at understanding and improving AI integration into the Japanese health care and medical education system.

Objective: We aimed to develop the Japanese version of the 12-item Attitudes Towards Artificial Intelligence scale (J-ATTARI-12) and investigate whether it is applicable to medical trainees.

Methods: We first translated the original English-language scale into Japanese. To examine its psychometric properties, we then conducted a validation survey by distributing the translated version as an online questionnaire to medical students and residents across Japan from June 2025 to July 2025. We assessed structural validity through factor analysis and convergent validity by computing the Pearson correlation coefficient between the J-ATTARI-12 scores and scores on attitude toward robots. Internal consistency reliability was assessed using Cronbach α values.

Results: We included 326 participants in our analysis. We used a split-half validation approach, with exploratory factor analysis (EFA) on the first half and confirmatory factor analysis on the second half. EFA suggested a 2-factor solution (factor 1: AI anxiety and aversion; factor 2: AI optimism and acceptance). Confirmatory factor analysis revealed that the model fitness indexes of the 2-factor structure suggested by the EFA were good (comparative fit index=0.914 [>0.900]; root mean square error of approximation=0.075 [<0.080]; standardized root mean square residual=0.056 [<0.080]) and superior to those of the 1-factor structure. The value of the Pearson correlation coefficient between the J-ATTARI-12 scores and the attitude toward robots scores was 0.52, which indicated good convergent validity. The Cronbach α for all 12 items was 0.84, which indicated a high level of internal consistency reliability.

Conclusions: We developed and validated the J-ATTARI-12. The developed instrument had good structural validity, convergent validity, and internal consistency reliability for medical trainees. The J-ATTARI-12 is expected to stimulate future studies and educational initiatives that can effectively assess and enhance the integration of AI into clinical practice and medical education systems.

(JMIR Med Educ 2026;12:e81986) doi:10.2196/81986

KEYWORDS

artificial intelligence; AI; attitudes toward artificial intelligence; attitudes toward AI; medical education; medical resident; medical student

XSL•FO
RenderX

Introduction

Artificial intelligence (AI) is undergoing rapid development and integration into health care [1,2]. Although the use of AI has raised a number of potential concerns, including a lack of control over the rapidly progressing technology [3], issues of privacy and data protection [4], ethical problems [5], and the potential loss of human jobs [6], the benefits of AI are expected to far outweigh these concerns. AI contributes to improving patient care through enhancing diagnostic accuracy and providing more personalized therapeutic plans [7]. Moreover, AI has the potential to enhance the training and education of medical professionals by offering benefits including virtual simulation and training, remote education, and the recording of teaching videos [8,9]. Consequently, the use of AI has increased in both clinical practice and medical education.

Individuals may hold divergent views on the prospects and risks of AI and adopt varied attitudes toward it [10]. The attitudes of physicians and medical trainees toward AI will have a major impact on its integration into medical practice and educational activities [11]. Negative attitudes toward AI can lead to skepticism and concerns, thereby impeding its adoption [12]. Conversely, positive attitudes toward AI likely foster trust [13], leading individuals to embrace AI systems and amplify their benefits and possibly resulting in their integration into health care [12].

Thus, assessing physicians' and medical trainees' attitudes toward AI is critical to identifying potential barriers and fostering acceptance. Accordingly, instruments with robust psychometric properties are required. Although several measures for assessing attitudes toward AI have been developed, only a few have been comprehensively evaluated for reliability and validity. In 2024, Stein et al [10] developed the 12-item Attitudes Towards Artificial Intelligence scale (ATTARI-12). This is a unidimensional scale that integrates affective, behavioral, and cognitive facets into a single measure. The authors subsequently confirmed its psychometric properties for the general population [10], and its wide use in future studies is anticipated.

Despite the potential global applicability of the ATTARI-12, there is currently no validated Japanese version. Japan has a distinct cultural context, and attitudes toward AI are likely influenced by cultural and societal norms. For instance, Japanese culture's high uncertainty avoidance tends to result in only a gradual adoption of new technologies [14]. In the absence of a culturally adapted measure, assessments may not accurately reflect the true sentiments of Japanese health care workers and trainees. This gap hinders research and educational efforts aimed at understanding and improving AI integration into the Japanese health care system.

In this study, we developed a Japanese version of the ATTARI-12, which was originally created for the general population, and examined whether it is applicable to medical students and resident physicians. We anticipated that the development of this scale in Japan would promote future research and educational courses that effectively assess and improve the integration of AI into clinical practice and the medical education system.

Methods

This study formed part of a larger research project regarding AI education for medical trainees. We conducted this study in the following 2 steps.

Step 1: Translation and Cross-Cultural Adaptation

In accordance with an international guideline [15], we translated the original ATTARI-12 into Japanese. First, the first author (HF) asked the creator of the original ATTARI-12 and corresponding author of the article that reported it to allow us to develop the Japanese version. The creator willingly provided permission for our translation of the scale. Second, 2 translators (HF and KK) conducted forward translations independently. Both translators were familiar with Western and Japanese cultures and had rich experience in developing translated versions of scales in the field of health profession education. In particular, KK is a fluent speaker of Japanese and English. Third, the translators performed a synthesis of their translations. Discrepancies were resolved through discussion (version 1). Fourth, HF asked professional bilingual translators who were not involved in our study to translate the Japanese text back into English. HF and KK then compared the back-translated and original English versions item by item and then revised the Japanese version (version 2). Fifth, an expert review of version 2 was conducted by an AI expert (YY) and a health profession education expert (YN) at HF's request. These experts concluded that no amendment was required. Sixth, HF contacted the creator of the original scale again and asked him to check version 2. The creator concluded that no revision was necessary. Seventh, pilot testing was performed with 2 medical trainees, which indicated no problematic items. Finally, version 2 was confirmed as the finalized Japanese version of the ATTARI-12 (J-ATTARI-12).

Step 2: Investigation of Psychometric Properties

Participants

Between June 2025 and July 2025, we recruited study participants from 5 universities and 9 hospitals across Japan, each of which varied in type and location (Multimedia Appendix 1). We asked the medical education directors of the universities and residency training directors of the hospitals to distribute our anonymous online self-administered questionnaire. An email with a link to the online questionnaire was sent to medical students at the universities and resident physicians at the hospitals via the respective directors. The participants were provided with a brief explanation of the study and indicated their consent to take part by checking the consent box. They were then able to access and complete the questionnaire. The survey duration was 1 month. To maximize the response rate, we sent reminders several times during the survey period.

Measures

The original version of the ATTARI-12 has 12 items [10], with responses on a 5-point Likert scale (1="strongly disagree"; 5="strongly agree") [10]. Items 2, 4, 7, 8, 10, and 12 are reverse

XSL•FO
RenderX

items and, therefore, are reverse scored [10]. The sum of all item scores is averaged to create a total score ranging from 1 to 5, with higher scores indicating more positive attitudes toward AI.

Statistical Analysis

To investigate the structural validity of the instrument, we performed exploratory factor analysis (EFA) and confirmatory factor analysis (CFA). Because conducting these 2 types of factor analysis on the same dataset would potentially raise concerns [16], we adopted the split-half validation approach, in which the participants were randomly divided into 2 groups, half (group A) for EFA and the other half (group B) for CFA. As mentioned above, we aimed to develop the scale in a manner that was culturally adapted to the Japanese medical education context. Consequently, EFA was conducted first, followed by CFA.

To determine the appropriateness of EFA, we checked the Kaiser-Meyer-Olkin (KMO) coefficient and Bartlett sphericity test. Running EFA requires a KMO value over 0.80 and a significant result in the Bartlett test [17,18]. We applied EFA to the responses of group A to explore the factor structure of the J-ATTARI-12 using maximum likelihood estimation and Promax rotation. We determined the final factor solution using the results of parallel analysis and the factor loading values (cutoff value=0.30).

We subsequently conducted CFA on group B to confirm the model obtained in EFA. CFA was performed using the maximum likelihood estimator method. Model fitness of CFA is commonly conducted using indexes, including the comparative fit index (CFI), root mean square error of approximation (RMSEA), and standardized root mean square residual (SRMR), with higher CFI (>0.900), lower RMSEA (<0.080), and lower SRMR (<0.080) values indicating a good fit [19-21]. In this study, we compared these indexes between the 2 models, namely, the model suggested by the EFA (2-factor model; described in detail below) and the 1-factor model.

Convergent validity was evaluated through hypothesis testing. On the basis of a previous finding that ATTARI-12 scores were positively linked with specific attitudes toward robots [10], we investigated this relationship using Pearson correlation coefficients. With reference to previous studies [10,22], attitudes toward robots were assessed using 3 items (eg, “Robots are necessary as they can do jobs that are too hard or too dangerous for people”), each of which was answered on a 4-point Likert scale (Multimedia Appendix 2). The scores were computed by averaging the responses to the 3 items (Cronbach α=0.78; mean 3.25, SD 0.55 in our dataset), with higher scores indicating a more positive attitude toward robots. The Pearson correlation coefficients are deemed meaningful if they are greater than 0.30 [23].

We used Cronbach α values to evaluate the internal consistency reliability of the scale. In this study, we computed the values using the responses of the entire sample (N=326; described in detail below). Cronbach α values above 0.70 are acceptable [24].

The possible influence of participant gender and year group on ATTARI-12 scores was explored using descriptive statistics and comparison analyses (independent 2-tailed *t* test or 1-way ANOVA). With regard to gender, as outlined later, the “Others” category comprised only 4 participants, rendering it inadequate for inclusion in a 1-way ANOVA. Given the high likelihood of unstable estimates and heterogeneity of variance associated with such a small group, we compared scores between only the 2 larger groups (man and woman) via independent *t* test. The undergraduate medical curriculum in Japan has traditionally consisted of 2 phases: the first 4 years of preclinical education and the subsequent 2 years of clinical education (ie, clinical clerkship) [25]. Accordingly, medical trainees were divided into preclinical medical students (ie, first- to fourth-year students), clinical medical students (ie, fifth- and sixth-year students), and medical residents. All statistical analyses were performed using R (version 4.5.1; R Foundation for Statistical Computing) and SPSS (version 30.0; IBM Corp), with 2-sided *P* values of <.05 considered statistically significant.

Ethical Considerations

This study was conducted according to the ethical standards and principles of the Declaration of Helsinki and was approved by the ethics committee of Juntendo University Faculty of Medicine (E25-0028). All participants checked the consent box at the beginning of the questionnaire to indicate their informed consent to take part in the study. To ensure confidentiality, all participant data were anonymized before analysis. Participants were entered into a draw for 1 of 10 ¥5000 (approximately US $30) gift cards.

Results

Overview

In total, 9.2% (326/3551) of the eligible participants responded to the survey. There were no missing data. Table 1 shows the participants’ characteristics, and Table 2 shows the responses to each item.

XSL•FO
RenderX

Table . Characteristics of the participants (N=326).

Characteristic	Participants, n (%)
Gender	
Woman	142 (43.6)
Man	180 (55.2)
Nonbinary	4 (1.2)
Year group	
Medical students	
First	103 (31.6)
Second	19 (5.8)
Third	51 (15.6)
Fourth	60 (18.4)
Fifth	13 (4.0)
Sixth	20 (6.1)
Medical residents	
First	32 (9.8)
Second	28 (8.6)

XSL•FO
RenderX

Table . Responses to the 12-item Attitudes Towards Artificial Intelligence scale (N=326).

Item (as in the original English-language version)	Responses, n (%)[a]				
	1	2	3	4	5
Item 1: “AI will make this world a better place.”	1 (0.3)	11 (3.4)	72 (22.1)	201 (61.7)	41 (12.6)
Item 2: “I have strong negative emotions about AI.”[b]	49 (15.0)	163 (50.0)	71 (21.8)	37 (11.3)	6 (1.8)
Item 3: “I want to use technologies that rely on AI.”	0 (0.0)	3 (0.9)	36 (11.0)	200 (61.3)	87 (26.7)
Item 4: “AI has more disadvantages than advantages.”[b]	28 (8.6)	171 (52.5)	101 (31.0)	22 (6.7)	4 (1.2)
Item 5: “I look forward to future AI developments.”	2 (0.6)	4 (1.2)	28 (8.6)	157 (48.2)	135 (41.4)
Item 6: “AI offers solutions to many world problems.”	2 (0.6)	24 (7.4)	73 (22.4)	175 (53.7)	52 (16.0)
Item 7: “I prefer technologies that do not feature AI.”[b]	19 (5.8)	124 (38.0)	120 (36.8)	51 (15.6)	12 (3.7)
Item 8: “I am afraid of AI.”[b]	17 (5.2)	62 (19.0)	77 (23.6)	147 (45.1)	23 (7.1)
Item 9: “I would rather choose a technology with AI than one without it.”	5 (1.5)	16 (4.9)	111 (34.0)	154 (47.2)	40 (12.3)
Item 10: “AI creates problems rather than solving them.”[b]	19 (5.8)	115 (35.3)	137 (42.0)	47 (14.4)	8 (2.5)
Item 11: “When I think about AI, I have mostly positive feelings.”	2 (0.6)	45 (13.8)	89 (27.3)	157 (48.2)	22 (6.7)
Item 12: “I would rather avoid technologies that are based on AI.”[b]	38 (11.7)	177 (54.3)	84 (25.8)	23 (7.1)	4 (1.2)

[a]5-point Likert scale (1=“strongly disagree” to 5=“strongly agree”)

[b]These are reverse-scored items.

Structural Validity Analysis

We performed EFA on group A using maximum likelihood estimation with Promax rotation (154/326, 47.2% of the participants). The KMO value was 0.83, and the Bartlett test was significant (P<.001). Table 3 shows the results of the EFA. EFA suggested a 2-factor solution. After discussion among the team members, we named the factors as follows: factor 1 was AI anxiety and aversion, and factor 2 was AI optimism and acceptance.

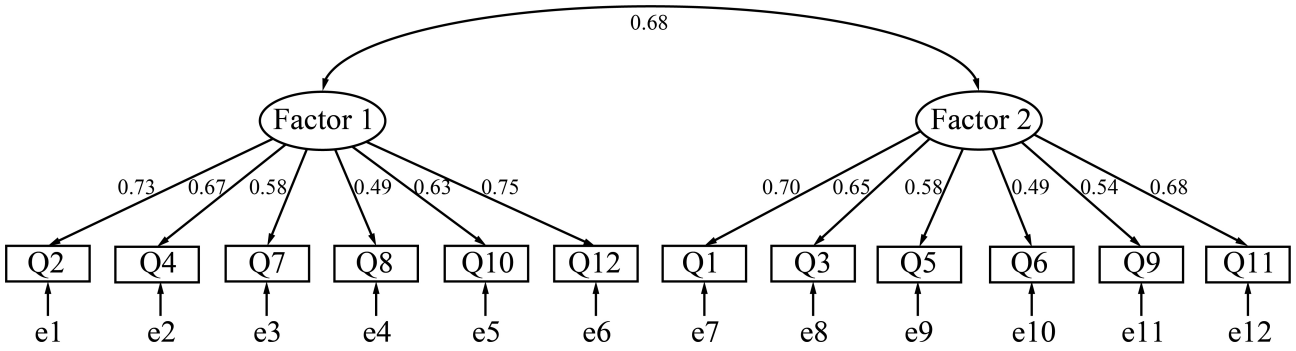
XSL•FO
RenderX

Table . Results of the exploratory factor analysis.

	Factor 1	Factor 2
Item 1 factor loading	−0.16	0.73
Item 2 factor loading	0.59	0.18
Item 3 factor loading	0.21	0.52
Item 4 factor loading	0.73	−0.07
Item 5 factor loading	−0.05	0.85
Item 6 factor loading	0.05	0.35
Item 7 factor loading	0.73	−0.06
Item 8 factor loading	0.30	0.12
Item 9 factor loading	−0.05	0.47
Item 10 factor loading	0.78	−0.15
Item 11 factor loading	0.12	0.44
Item 12 factor loading	0.64	0.07
Eigenvalue	3.79	0.76
Percentage of the variance explained	22	18

We then performed CFA on group B (the remaining 172/326, 52.8% of the participants) using the maximum likelihood estimator method. The 2-factor model suggested by the EFA yielded better goodness-of-fit results (CFI=0.914; RMSEA=0.075; SRMR=0.056) than the 1-factor model (CFI=0.804; RMSEA=0.113; SRMR=0.078). Multimedia Appendix 3 shows the CIs for the RMSEA. Figure 1 shows the path diagram. Therefore, we adopted the 2-factor model.

Figure 1. Factor structure of the Japanese version of the 12-item Attitudes Towards Artificial Intelligence scale (confirmatory factor analysis). Ellipses indicate latent variables (factors). Rectangles indicate observed variables (items). Values on single-headed arrows indicate standardized factor loadings. Values on double-headed arrows indicate correlation coefficients.



Convergent Validity Analysis

We computed the Pearson correlation coefficient between the ATTARI-12 scores and the attitude toward robots scores. The coefficient value was 0.52 (P<.001), which indicated a positive correlation between these 2 scores.

Internal Consistency Reliability Analysis and Descriptive Statistics

Table 4 shows the descriptive statistics with the Cronbach α values. The Cronbach α for all 12 items was 0.84. The Cronbach α values for factors 1 and 2 were 0.80 and 0.76, respectively. These values were above the cutoff (0.70).

Table . Internal consistency reliability analysis and descriptive statistics.

	Number of items	Mean (SD)	Observed range	Cronbach α
Total	12	3.61 (0.50)	2.08 - 5.00	0.84
Factor 1	6	3.36 (0.63)	1.00 - 5.00	0.80
Factor 2	6	3.87 (0.52)	2.17 - 5.00	0.76

Table 5 shows descriptive data and comparison by gender and year group. The comparison analyses did not show any significant differences in ATTARI-12 scores.

XSL•FO
RenderX

Table . Descriptive data and comparison by gender and year group.

	Score on ATTARI-12[a], mean (SD)	*P* value
Gender[b]		.71[c]
Woman	3.60 (0.47)	
Man	3.62 (0.53)	
Year group		.21[d]
First- to fourth-year medical students	3.58 (0.51)	
Fifth- and sixth-year medical students	3.67 (0.55)	
Medical residents	3.70 (0.42)	

[a]ATTARI-12: 12-item Attitudes Towards Artificial Intelligence scale.

[b]The "Nonbinary" group included only 4 participants, which is too small to be appropriately included in a 1-way ANOVA. Given the high likelihood of unstable estimates and heterogeneity of variance associated with such a small group, we compared scores between only the 2 larger groups (man and woman).

[c]*P* value from independent-sample *t* test.

[d]*P* value from 1-way ANOVA.

Discussion

Principal Findings

In this study, we translated the ATTARI-12, originally developed for the general population, into Japanese in accordance with an international guideline [15] and then validated its structural and convergent validity and internal consistency reliability for medical trainees. Applying this scale in the context of medical trainees in Japan has the potential to stimulate future research and educational interventions. Such initiatives would serve to effectively assess and enhance the integration of AI in clinical practice and the medical education system.

Our study found that the internal consistency reliability of the J-ATTARI-12 was good. The findings were consistent with those of the original ATTARI-12 developmental study [10]. The original study reported Cronbach α values of 0.93 for a US sample and 0.90 for a German sample [10], which suggested that the ATTARI-12 is likely a helpful measure with good internal consistency reliability across countries.

Factor analysis indicated that the J-ATTARI-12 had a 2-factor structure, in contrast to the original English-language version's unidimensional structure. The original ATTARI-12 was conceptualized as a unidimensional scale. The original article by Stein et al [10] suggested that the developers intentionally balanced positively and negatively worded items. This design makes it unsurprising that the Japanese version yielded 2 factors corresponding to negative (factor 1: AI anxiety and aversion) and positive (factor 2: AI optimism and acceptance) item valence. Therefore, we should acknowledge that, although 2 factors emerged, they simply reflect positive versus negative wording and that, for the purpose of international comparison, using a total score based on a 1-factor assumption may remain preferable. At the same time, the content of the 2 factors (ie, AI anxiety and aversion vs AI optimism and acceptance) may reflect attitudinal domains that are influenced by broader cultural characteristics, such as uncertainty avoidance, technophilia, or collectivism. To determine whether a 2-factor structure offers conceptual or psychometric advantages beyond a unidimensional model and whether this pattern is observed across different cultural contexts, future cross-cultural validation studies with larger and more diverse samples are required.

We note a couple of potential limitations of our study. First, the response rate was relatively low. In addition, the survey was conducted using convenience sampling and included only 14 institutions, and it is likely that participating institutions or individuals had a greater interest in AI. It is increasingly challenging to obtain high response rates to online surveys, and rates frequently drop to 10% [26,27]. Nevertheless, the literature suggests that the response rate to our survey may have been sufficient to provide reliable data [28,29]. Additionally, it should be noted that, despite the absence of accurate statistics on the demographic variables of medical trainees across Japan, recent reports have indicated that the proportion of female physicians aged ≤29 years stands at approximately 40%, with a steady upward trend observed in recent years [30]. Thus, the finding that 43.6% (142/326) of the respondents were women does not generate concerns regarding the representativeness of the sample. Second, in this study, although we evaluated the structural validity, convergent validity, and internal consistency reliability of the scale, other determinants of validity (eg, discriminant validity and predictive validity) and reliability (eg, test-retest reliability) have yet to be examined. Future studies should examine these psychometric properties. Third, the use of a gift card lottery can raise concerns about bias. Nevertheless, this method is commonly used and appears to be acceptable [31-34]. Fourth, our results may indicate a ceiling effect. For example, 41.4% (135/326) of the respondents selected the highest score (5) for item 5. However, according to the original ATTARI-12 paper, responses tend to be skewed toward a score of 4, suggesting that the scale inherently elicits generally positive attitudes toward AI. Therefore, these findings do not represent an unusual deviation but are consistent with the response patterns reported in the original study.

Despite these limitations, this study produced the first Japanese version of the ATTARI-12, which is likely to be used as a novel

XSL•FO
RenderX

measure for assessing AI attitudes among medical students and residents. Two outcomes are expected. First, by incorporating the J-ATTARI-12 into medical curricula, medical educators will be able to develop a more comprehensive understanding of the trainees' readiness for AI adoption. This will likely be advantageous in the design of customized educational interventions using AI based on the level of each trainee's attitude toward AI [35] and will support curriculum development by enabling medical educators to identify learners who may need more foundational AI exposure or targeted support. Second, the J-ATTARI-12 will facilitate medical education research in Japan. The scale allows researchers to examine the impact of specific educational interventions on attitude change concerning AI, and it is suitable for repeated administration in longitudinal studies to track how attitudes evolve over the course of training. In addition, the availability of the J-ATTARI-12 will promote cross-cultural comparison of attitudes toward AI and its possible outcomes internationally. For these reasons, we expect that the J-ATTARI-12 will play a pivotal role in facilitating data-informed curriculum development and contribute to the expanding body of medical education research in this era of AI.

Conclusions

In this study, we developed the J-ATTARI-12, originally developed for the general population, in accordance with an international guideline. A validation survey revealed that the structural and convergent validity, as well as the internal consistency reliability, were good for medical trainees in Japan. The developed measure can be used for customized educational initiatives using AI based on the level of each trainee's attitude toward AI. It will also provide helpful information to medical education researchers in Japan in this era of AI.

Acknowledgments

The authors would like to thank all the participants. They would also like to thank ChatGPT-5.1 by OpenAI for its valuable assistance in refining their academic writing.

Funding

This work was funded by the "Creating training hubs for advanced medical personnel (supporting the fostering of doctors with advanced clinical and research capabilities)" project and was supported by the Ministry of Education, Culture, Sports, Science, and Technology of Japan.

Data Availability

The datasets generated or analyzed during this study are available from the corresponding author on reasonable request.

Authors' Contributions

HF conceived the study with input from HM, KK, YN, YY, and TN. HF conducted the data analysis. HF wrote the first draft of the manuscript. All authors reviewed, discussed, proofread, and approved the final version of the manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Characteristics of the participating institutions.
[DOCX File, 18 KB - mededu_v12i1e81986_app1.docx]

Multimedia Appendix 2
Measure for attitudes toward robots.
[DOCX File, 18 KB - mededu_v12i1e81986_app2.docx]

Multimedia Appendix 3
CIs for the root mean square error of approximation.
[DOCX File, 18 KB - mededu_v12i1e81986_app3.docx]

References

1. Buch VH, Ahmed I, Maruthappu M. Artificial intelligence in medicine: current trends and future possibilities. Br J Gen Pract 2018 Mar;68(668):143-144. [doi: 10.3399/bjgp18X695213] [Medline: 29472224]
2. Wu K, Wu E, Theodorou B, et al. Characterizing the clinical adoption of medical AI devices through U.S. insurance claims. NEJM AI 2024 Nov 9;1(1). [doi: 10.1056/AIoa2300030]

XSL•FO
RenderX

3. Habli I, Lawton T, Porter Z. Artificial intelligence in health care: accountability and safety. Bull World Health Organ 2020 Feb 25;98(4):251-256. [doi: 10.2471/BLT.19.237487] [Medline: 32284648]
4. Bavli I, Ho A, Mahal R, McKeown MJ. Ethical concerns around privacy and data security in AI health monitoring for Parkinson's disease: insights from patients, family members, and healthcare professionals. AI Soc 2024 Jan 23;40:155-165. [doi: 10.1007/s00146-023-01843-6]
5. Katirai A. The ethics of advancing artificial intelligence in healthcare: analyzing ethical considerations for Japan's innovative AI hospital system. Front Public Health 2023 Jul 17;11:1142062. [doi: 10.3389/fpubh.2023.1142062] [Medline: 37529426]
6. Khan B, Fatima H, Qureshi A, et al. Drawbacks of artificial intelligence and their potential solutions in the healthcare sector. Biomed Mater Devices 2023 Feb 8:1-8. [doi: 10.1007/s44174-023-00063-2] [Medline: 36785697]
7. Andrade-Arenas L, Yactayo-Arias C. A bibliometric analysis of the advance of artificial intelligence in medicine. Int J Electr Comput Eng 2024;14(3):3350. [doi: 10.11591/ijece.v14i3.pp3350-3361]
8. Dave M, Patel N. Artificial intelligence in healthcare and education. Br Dent J 2023 May 26;234:761-764. [doi: 10.1038/s41415-023-5845-2]
9. Mir MM, Mir GM, Raina NT, et al. Application of artificial intelligence in medical education: current scenario and future perspectives. J Adv Med Educ Prof 2023 Jul;11(3):133-140. [doi: 10.30476/JAMP.2023.98655.1803] [Medline: 37469385]
10. Stein JP, Messingschlager T, Gnambs T, Hutmacher F, Appel M. Attitudes towards AI: measurement and associations with personality. Sci Rep 2024 Feb 5;14(1):2909. [doi: 10.1038/s41598-024-53335-2] [Medline: 38316898]
11. Horowitz MC, Kahn L. Leading in artificial intelligence through confidence building measures. Wash Q 2021 Oct 2;44(4):91-106. [doi: 10.1080/0163660X.2021.2018794]
12. Naiseh M, Babiker A, Al-Shakhsi S, et al. Attitudes towards AI: the interplay of self-efficacy, well-being, and competency. J Technol Behav Sci 2025. [doi: 10.1007/s41347-025-00486-2]
13. Sindermann C, Sha P, Zhou M, et al. Assessing the attitude towards artificial intelligence: introduction of a short measure in German, Chinese, and English language. Künstl Intell 2020;35:109-118. [doi: 10.1007/s13218-020-00689-0]
14. Hofstede GJ. Adoption of communication technologies and national culture. Syst Inf Manag 2001;6(3):55-69 [FREE Full text]
15. Beaton DE, Bombardier C, Guillemin F, Ferraz MB. Guidelines for the process of cross-cultural adaptation of self-report measures. Spine (Phila Pa 1976) 2000 Dec 15;25(24):3186-3191. [doi: 10.1097/00007632-200012150-00014] [Medline: 11124735]
16. Lorenzo-Seva U. SOLOMON: a method for splitting a sample into equivalent subsamples in factor analysis. Behav Res Methods 2022 Dec;54(6):2665-2677. [doi: 10.3758/s13428-021-01750-y] [Medline: 34918226]
17. Kaiser HF. An index of factorial simplicity. Psychometrika 1974;39(1):31-36. [doi: 10.1007/BF02291575]
18. Karimian Z, Chahartangi F. Development and validation of a questionnaire to measure educational agility: a psychometric assessment using exploratory factor analysis. BMC Med Educ 2024 Nov 9;24:1284. [doi: 10.1186/s12909-024-06307-z]
19. Brown TA. Confirmatory Factor Analysis for Applied Research: The Guilford Press; 2015.
20. Tabachnick BG, Fidell LS. Using Multivariate Statistics: Pearson Education; 2013.
21. Vandenberg RJ, Lance CE. A review and synthesis of the measurement invariance literature: suggestions, practices, and recommendations for organizational research. Organ Res Methods 2000;3(1):4-70. [doi: 10.1177/109442810031002]
22. Gnambs T, Appel M. Are robots becoming unpopular? Changes in attitudes towards autonomous robotic systems in Europe. Comput Human Behav 2019 Apr;93:53-61. [doi: 10.1016/j.chb.2018.11.045]
23. Cohen J. Statistical power analysis. Curr Dir Psychol Sci 1992;1(3):98-101. [doi: 10.1111/1467-8721.ep10768783]
24. DeVellis RF. Scale Development: Theory and Applications: Sage Publications; 2003.
25. Nishigori H. Medical education in Japan. Med Teach 2024 Sep;46(sup1):S4-S10. [doi: 10.1080/0142159X.2024.2372108] [Medline: 39545499]
26. Fan W, Yan Z. Factors affecting response rates of the web survey: a systematic review. Comput Human Behav 2010 Mar;26(2):132-139. [doi: 10.1016/j.chb.2009.10.015]
27. Van Mol C. Improving web survey efficiency: the impact of an extra reminder and reminder content on web survey response. Int J Soc Res Methodol 2017;20(4):317-327. [doi: 10.1080/13645579.2016.1185255]
28. Fosnacht K, Sarraf SA, Howe E, Peck LK. How important are high response rates for college surveys? Rev High Educ 2017;40(2):245-265. [doi: 10.1353/rhe.2017.0003]
29. Wu MJ, Zhao K, Fils-Aime F. Response rates of online surveys in published research: a meta-analysis. Comput Hum Behav Rep 2022 Aug;7:100206. [doi: 10.1016/j.chbr.2022.100206]
30. Overview of statistics on doctors, dentists, and pharmacists in 2022 [Web page in Japanese]. Ministry of Health, Labor and Welfare, Japan. 2022. URL: https://www.mhlw.go.jp/toukei/saikin/hw/ishi/22/index.html [accessed 2025-12-10]
31. Fujikawa H, Ando T, Kondo K, Hayashi M. Development and psychometric properties of the Japanese version of the capacity for wonder scale. Int J Med Educ 2025 May 29;16:100-106. [doi: 10.5116/ijme.6819.fcae] [Medline: 40449508]
32. Fujikawa H, Aoki T, Eto M. Associations between workplace social capital, well-being, and work engagement in medical residents: a multicenter cross-sectional study. BMC Med Educ 2024 Sep 28;24(1):1063. [doi: 10.1186/s12909-024-06055-0] [Medline: 39342158]

XSL•FO
RenderX

33. Fujikawa H, Hayashi M, Son D, Kondo K, Eto M. Translating, adapting, and validating the medical student version of the patient care ownership scale for use in Japan. BMC Med Educ 2024 Jun 28;24(1):706. [doi: 10.1186/s12909-024-05704-8] [Medline: 38943114]
34. Shinohara K, Aoki T, So R, et al. Influence of overstated abstract conclusions on clinicians: a web-based randomised controlled trial. BMJ Open 2017 Dec 14;7(12):e018355. [doi: 10.1136/bmjopen-2017-018355] [Medline: 29247102]
35. Sami A, Tanveer F, Sajwani K, et al. Medical students' attitudes toward AI in education: perception, effectiveness, and its credibility. BMC Med Educ 2025 Jan 17;25(1):82. [doi: 10.1186/s12909-025-06704-y] [Medline: 39833834]

Abbreviations

AI: artificial intelligence
ATTARI-12: 12-item Attitudes Towards Artificial Intelligence scale
CFA: confirmatory factor analysis
CFI: comparative fit index
EFA: exploratory factor analysis
J-ATTARI-12: Japanese version of the 12-item Attitudes Towards Artificial Intelligence scale
KMO: Kaiser-Meyer-Olkin
RMSEA: root mean square error of approximation
SRMR: standardized root mean square residual

Edited by Z Khlaif; submitted 07.Aug.2025; peer-reviewed by T Iwayama, T Yusuff; revised version received 11.Dec.2025; accepted 14.Dec.2025; published 14.Jan.2026.

Please cite as:
Fujikawa H, Mori H, Kondo K, Nishizaki Y, Yano Y, Naito T
Adaptation of the Japanese Version of the 12-Item Attitudes Towards Artificial Intelligence Scale for Medical Trainees: Multicenter Development and Validation Study
JMIR Med Educ 2026;12:e81986
URL: https://mededu.jmir.org/2026/1/e81986
doi:10.2196/81986

© Hirohisa Fujikawa, Hirotake Mori, Kayo Kondo, Yuji Nishizaki, Yuichiro Yano, Toshio Naito. Originally published in JMIR Medical Education (https://mededu.jmir.org), 14.Jan.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Education, is properly cited. The complete bibliographic information, a link to the original publication on https://mededu.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Evaluation of a Problem-Based Learning Program's Effect on Artificial Intelligence Ethics Among Japanese Medical Students: Mixed Methods Study

Yuma Ota[1], MSN, RN; Yoshikazu Asada[2], PhD; Saori Kubo[3], PhD; Takeshi Kanno[4,5], MD, PhD; Machiko Saeki Yagi[2], RN, CHSE, PhD; Yasushi Matsuyama[6], MD, MHPE, PhD

[1]Medical Education Center, Jichi Medical University Graduate School, Shimotsuke, Japan

[2]Medical Education Center, Jichi Medical University, Shimotsuke, Japan

[3]Graduate School of Education, Tohoku University, Sendai, Japan

[4]R & D Division of Career Education for Medical Professionals, Medical Education Center, Jichi Medical University, Shimotsuke, Japan

[5]Division of Gastroenterology, Graduate School of Medicine, Tohoku University, Sendai, Japan

[6]Medical Education Center, Jichi Medical University, Tochigi, Japan

Corresponding Author:
Yasushi Matsuyama, MD, MHPE, PhD
Medical Education Center
Jichi Medical University
3311-1 Yakushiji, Shimotsuke
Tochigi, 329-0498
Japan
Phone: 81 285 58 7067
Email: yasushim@jichi.ac.jp

Abstract

Background: The rapid advancement of artificial intelligence (AI) has had a substantial impact on medicine, necessitating the integration of AI education into medical school curricula. However, such integration remains limited. A key challenge is the discrepancy between medical students' positive perceptions of AI and their actual competencies, with research in Japan identifying specific gaps in the students' competencies in understanding regulations and discussing ethical issues.

Objective: This study evaluates the effectiveness of an educational program designed to improve medical students' competencies in understanding legal and ethical AI-related issues. It addresses the following research questions: (1) Does this educational program improve students' knowledge of AI and its legal and ethical issues, and what is each program element's contribution to this knowledge? (2) How does this educational program qualitatively change medical students' thoughts on these issues from an abstract understanding to a concrete and structured thought process?

Methods: This mixed methods study used a single-group pretest and posttest framework involving 118 fourth-year medical students. The 1-day intervention comprised a lecture and problem-based learning (PBL) session centered on a clinical case. A 24-item multiple-choice questionnaire (MCQ) was administered at 3 time points (pretest, midtest, and posttest), and descriptive essays were collected before and after the intervention. Data were analyzed using linear mixed-effects models, the Wilcoxon signed-rank test, and text mining, including comparative frequency analysis and cooccurrence network analysis with Jaccard coefficients. An optional survey on student perceptions based on the attention, relevance, confidence, and satisfaction model was conducted (n=76, 64.4%).

Results: Objective knowledge scores increased significantly from the pretest (median 17, IQR 15-18) to posttest (median 19, IQR 17-21; β=1.42; *P*<.001). No significant difference was observed between score gains during the lecture and PBL phases (*P*=.54). Qualitative text analysis revealed the significant transformation of cooccurrence network structures (Jaccard coefficients 0.116 and 0.121) from fragmented clusters to integrated networks. Students also used professional and ethical terminology more frequently. For instance, use of the term "bias" in patient explanations increased from 10 (8.5%) at pretest to 25 (21.2%) at posttest, while references to "personal information" in physician precautions increased from 36 (30.5%) to 50 (42.4%). The optional survey indicated that students' confidence (mean 3.78, SD 0.87) was significantly lower than their perception of the program's relevance (mean 4.20, SD 0.71; *P*<.001).

XSL•FO
RenderX

Conclusions: This PBL-based program was associated with the improvements in knowledge and, more importantly, a structural transformation in students' thinking about AI ethics from an abstract level to a concrete, clinically grounded reasoning. The discrepancy between quantitative and qualitative findings suggests limitations of MCQs in assessing higher-order skills fostered by PBL. Overall, this study indicates the potential of PBL as an effective pedagogical method for AI ethics education.

(JMIR Med Educ 2026;12:e84535) doi:10.2196/84535

KEYWORDS

artificial intelligence; undergraduate medical education; medical ethics; problem-based learning; mixed methods research; text mining

Introduction

The rapid advancement of artificial intelligence (AI) has significantly affected the medical domain, which uses applications ranging from diagnostic support to treatment selection [1-3]. This necessitates the integration of AI education into medical school curricula to ensure that future physicians have the necessary competencies to use these technologies effectively and ethically [4-6]. Both educators and students recognize this necessity, and many studies support the implementation of systematic AI education [7,8]. However, despite the establishment of international recommendations, the integration of AI education into medical school curricula remains limited [3,6]. One of the key challenges in AI education identified by earlier research is the discrepancy between medical students' perceptions of and competencies in AI [6,7,9]. Recent systematic reviews consistently indicate that despite generally maintaining a positive attitude toward AI's health care potential, students have a moderate to low level of objective knowledge and practical skills [7]. This gap between the positive perceptions and insufficient competency acquisition of students indicates the need to integrate systematic and effective educational programs into the medical curriculum [9-11].

Furthermore, AI education poses not only technical challenges but also legal and ethical ones [1,2,12]. The use of AI can give rise to new legal and ethical issues, such as the risk of algorithmic bias resulting in patient discrimination, challenges related to patient privacy, and problems associated with accountability for AI-driven decisions [2,3,13]. Further, expert consensus studies indicate that nontechnical competencies such as ethics, law, communication, and collaboration are considered more important than advanced technical skills such as programming for physicians [8,11]. However, practical, case-based ethics education and standardized frameworks to teach these important topics are currently lacking [3,12,13].

This global context is highly relevant in Japan. In response to these trends, Japan's Ministry of Education, Culture, Sports, Science, and Technology revised the national Model Core Curriculum for medical education in 2022, establishing "Competency in Utilizing Information Science and Technology" as a core competency for all future physicians [14,15]. This necessitated an investigation of the specific educational needs of Japanese medical students. A recent national cross-sectional study of sixth-year medical students in Japan clarified the acquisition status of specific competencies [16]. The study revealed particularly low self-assessment in three key learning objectives: (1) understanding the regulations, laws, and guidelines for information science and technology (IST) adoption in medicine; (2) discussing ethical issues, such as social disparities; and (3) understanding and discussing the potential applications of specific technologies, including AI, in medical care [16]. These results highlight the necessity of implementing focused educational interventions to address these gaps among Japanese medical students. To address these identified gaps, we chose a problem-based learning (PBL) framework for our educational intervention. This decision was directly informed by the findings of the aforementioned national study [16], which not only highlighted these specific areas of low confidence but also revealed a significant positive association between prior PBL experience and students' self-assessed competence in these very domains. This suggests that the PBL methodology—which promotes active, case-based, and collaborative problem-solving—is a particularly effective approach for navigating the complex, nontechnical challenges of AI ethics and law.

Accordingly, this study aims to evaluate the effectiveness of an educational program designed to improve medical students' competencies in addressing the legal and ethical issues associated with AI use. To this end, we developed and implemented a PBL program centered on a clinical case requiring fourth-year medical students to navigate AI-related ethical dilemmas. The use of the term "artificial intelligence" by Japan's Model Core Curriculum highlights AI's development period [14,15]. However, the educational program developed in this study considers recent technological advancements and the societal landscape to primarily address the ethical challenges posed by generative AI. Therefore, although this study uses the term AI in a broader sense in alignment with earlier research and the core curriculum, its content and examples mainly focus on generative AI. Finally, we address the following research questions (RQs):

RQ1: Does this educational program improve students' knowledge of AI and its legal and ethical issues? Furthermore, which program elements improve knowledge, and how?

RQ2: How does the educational program qualitatively change medical students' thoughts on the legal and ethical issues of AI? In particular, does their thinking shift from an abstract understanding to a more concrete and structured thought process that is based on the clinical context?

XSL•FO
RenderX

Methods

Study Design and Participants

This study adopted a mixed methods design using a single-group pretest and posttest framework to evaluate an educational intervention. Specifically, we used a convergent approach integrating quantitative assessment (multiple-choice questionnaire [MCQ] scores) and quantitative text analysis of qualitative data (student essays). The participants were fourth-year medical students (n=124) at Jichi Medical University who undertook the intervention program as a mandatory part of their curriculum during the 2025 academic year. The program was conducted immediately before the students commenced their full-time clinical clerkship; this time was selected because they were expected to soon encounter ethical and legal issues related to AI's adoption in clinical practice.

Educational Intervention

The intervention in this study was a 1-day (approximately 7 hours) program based on the PBL model [17]. The program's learning objectives were selected such that they address the 3 key competency gaps (understanding regulations, discussing ethical issues, and understanding AI applications) that had low self-assessment scores among Japanese medical students in an earlier national cross-sectional study. The program required students to achieve the following three competencies based on a rubric proposed by earlier research [18]:

1. For standard clinical settings and cases, be able to take action based on regulations (laws, guidelines, etc) associated with IST's use in medicine at a certain rate.
2. Be able to exemplify and discuss the ethical issues, such as the social disparities caused by the digital divide, potentially arising from the use of digital information and scientific technology in medicine and welfare.
3. Be able to understand the IST (AI) related to medicine and discuss it with the support of practitioners.

PBL Case and Program Structure

At the beginning of the PBL-based program, a video depicting a challenging clinical scenario was shown. A patient who is diagnosed with "muscle pain" for his left shoulder pain by a medical AI consultation app is convinced of the diagnosis. However, his anxious son is concerned about the app's reliability (since it lists acute myocardial infarction as the second possibility) and asks a medical student on clinical rotation to explain the AI diagnosis' credibility. The case was designed to encompass multiple key AI ethics themes, including physician accountability, potential AI bias and safety issues, and empathetic communication.

The program was conducted over 1 complete day with the following structure (Table 1). First, at the beginning of the program, the learning objectives and the case video were presented. The case involved a scenario in which a medical student on clinical rotation was asked by a patient's anxious son to explain the reliability of the medical AI consultation app that had been used to diagnose the patient's left shoulder pain as muscle pain, a diagnosis the patient firmly believed. This case encompassed multiple key themes in AI ethics, including physician accountability, potential AI bias and safety issues, and empathetic communication. Subsequently, a preprogram assessment (pretest: an MCQ and descriptive questions) was conducted. Then, the foundational knowledge of AI was detailed in lecture 1, which referenced internationally acknowledged AI education content for medical students [3]. This lecture supported the acquisition of sufficient foundational knowledge to address the PBL case. It provided an overview of the fundamentals of AI, including its definition, strengths, machine and deep learning mechanisms, and the benefits and challenges of AI's introduction into health care. Immediately after lecture 1, the second assessment (midtest: MCQ alone) was administered.

Table 1. Overview of the program components and schedule.

Content	Time	Activity format
Introduction: presentation of learning objectives and case	15 minutes	Individual
Assessment 1 (pretest): MCQ[a] and descriptive questions	35 minutes	Individual
Lecture 1: foundational knowledge of AI[b]	35 minutes	Individual
Assessment 2 (midtest): MCQ	15 minutes	Individual
PBL[c] session 1: confirmation of schedule and learning process	5 minutes	Group
PBL session 2: problem clarification	20 minutes	Group
PBL session 3: self-learning (including lunch break)	120 minutes	Individual
PBL session: problem-solving group work	70 minutes	Group
Lecture 2: case review and debriefing	40 minutes	Individual
Assessment 3 (posttest): MCQ and descriptive questions	35 minutes	Individual

[a]MCQ: multiple-choice questionnaire.

[b]AI: artificial intelligence.

[c]PBL: problem-based learning.

XSL•FO
RenderX

The PBL process proceeded in the following stages: (1) PBL learning process confirmation, (2) problem clarification, (3) self-directed learning (including a lunch break), and (4) a 70-minute problem-solving group work session. PBL was conducted in 16 small groups of 7-8 students each. The session comprised problem clarification, self-directed learning, and problem-solving phases and included a final 70-minute group work session where students used shared Google slides to collaboratively devise an explanation for the patient and his son. Five tutors facilitated the process using a standardized tutor guide and, to enhance discussion quality, key domestic AI guidelines [19,20] were distributed as group work materials.

At the end of the program, in lecture 2, an instructor reviewed the case to consolidate important concepts regarding AI's legal and ethical issues. Immediately thereafter, the final assessment (posttest: an MCQ and descriptive questions) was administered. A detailed description of the PBL program's development and theoretical underpinnings has been published [21].

Data Collection

To evaluate the program's effectiveness, we collected both quantitative and qualitative data using the university's learning management system.

Using the MCQ

A 24-item MCQ, with its content aligned with lecture 1 materials, was used to assess students' knowledge. Each item scored 1 point for a correct answer, the total possible score being 24 points. The questionnaire comprised 8 subscales, each with 3 items, covering the mechanisms of AI, the strengths of AI, machine learning, deep learning, overfitting, prompts, personal information protection, and challenges associated with AI use. To assess knowledge acquisition over time, the same set of questions was administered at 3 time points: before the intervention (pretest), after the lecture (midtest), and after the PBL program's completion (posttest). To minimize recall bias, the order of the questions and their options was randomized at each administration.

Descriptive Essay Questions

Two descriptive essay questions were administered before the intervention (pretest) and after the program (posttest): "A patient literally believes the results diagnosed by a personal AI consultation tool. As a medical student, how will you explain the dangers of believing the results diagnosed by AI?" (question 1) and "When using AI as a physician, what points do you think require caution?" (question 2). These questions were designed to evaluate changes in students' patient communication strategies and their understanding of the professional and ethical responsibilities associated with AI use.

Program Perception Survey

After the program, an optional survey was administered to assess students' perceptions of the program based on the attention, relevance, confidence, and satisfaction (ARCS) model [22]. Students rated their agreement with statements for each component on a 5-point Likert scale (ranging from 1=strongly disagree to 5=strongly agree).

Data Analysis

All statistical analyses were performed using R software (version 4.4.3; R Foundation for Statistical Computing) [23].

MCQ Data

The normality of data distribution for score gains was assessed using the Shapiro-Wilk test to determine the appropriate statistical test to compare educational phases. To assess the overall trajectory of the increase in score across the time points pretest, midtest, and posttest, a linear mixed-effects model was used. To compare the score gains between the lecture phase (midtest-pretest) and the PBL phase (posttest-midtest), a paired-samples 2 tailed *t* test was performed. Any violation of the assumption of normality was prespecified to result in the use of the nonparametric Wilcoxon signed-rank test as an alternative.

Descriptive Essay Data

To quantitatively and objectively analyze the content of free-text data, a quantitative text analysis (text mining) approach using several R packages was adopted. The analysis targeted the free-text responses to 2 questions at 2 time points: question 1 (patient explanation) at pretest and posttest and question 2 (physician precautions) at pretest and posttest. Further, data handling, aggregation, and visualization were performed using the *dplyr*, *tidyr*, *tidytext*, *widyr*, *igraph*, and *ggraph* R packages [24].

Morphological Analysis for Word Tokenization and Normalization

To segment Japanese free-text into analyzable word units, morphological analysis was performed using the *RMeCab* package (version 1.1.4) [25], which interfaces with the open-source engine MeCab, using the standard Information-technology Promotion Agency dictionary (ipadic). The analysis was limited to content words (nouns, verbs, and adjectives), with verbs and adjectives being normalized to nouns. Nouns alone are typically used as morphemes in network analysis [26].

Word Frequency Analysis

Extracted and normalized words were used to calculate the word frequencies for the responses to each question at each time point (question 1 at pretest, question 1 at posttest, question 2 at pretest, and question 2 at posttest). To assess the breadth of concept adoption across the cohort and mitigate the impact of outliers (eg, a single student repeating the same word), we treated word usage as a binary variable (presence or absence) for each student. Therefore, instead of raw counts, we calculated the number of students who used each word at least once to identify the most frequently used terms.

Comparative Word Frequency and Transition Analysis

To capture changes in descriptive content, the number of students using specific words was compared among time points for each question (pretest vs posttest for question 1 and pretest vs posttest for question 2). In addition to calculating the net difference in prevalence, we performed a transition analysis to examine individual-level changes. Specifically, we identified the number of students who newly adopted a word (nonuse at

XSL•FO
RenderX

pretest to use at posttest) and those who dropped a word (use at pretest to nonuse at posttest). This allowed us to distinguish whether the observed changes in prevalence were driven by the acquisition of new concepts or the abandonment of previous ones.

Cooccurrence Network Analysis

To structurally clarify the contextual relationships among words irrespective of their frequencies, a quantitative cooccurrence network analysis was performed. In this study, cooccurrence involves capturing the semantic proximity of a pair of words appearing adjacently within the same sentence. Once text was segmented into sentence units, adjacent word pairs were extracted. Similar to the word frequency analysis, we calculated the number of students who used each word pair (binary counting) to avoid overrepresentation by single individuals, using the *widyr* package. Subsequently, a network diagram was created using the words mentioned by 5 or more students as nodes and their cooccurrence relationships as edges, with edge thickness indicating the relationship's strength.

The Jaccard coefficient was used as a similarity metric to quantify the structural changes in these networks. The Jaccard coefficient measures the similarity between 2 sets by taking their intersection and dividing it by their union [26]. In this study, the coefficient was applied to compare entire network structures and was defined as the number of shared cooccurrence pairs (edges) divided by the total number of unique pairs across both pretest and posttest networks. We applied this metric because it assesses the number of common cooccurrence pairs without being affected by each pair's frequency. Accordingly, we could determine the extent to which the network structure changed over time, which indicated a shift in conceptual relationships, rather than a simple change in discourse volume. We identified the numbers of common pairs (present at both time points) and unique pairs (present at only one time point), as well.

Program Perception Survey Data

For the optional ARCS survey data, the normality of the data distribution was assessed using the Shapiro-Wilk test. We elected to use parametric statistics by prioritizing the robustness of these tests given the adequate sample size, which mitigates the impact of potential deviations from normality. Descriptive statistics (mean and SD) were calculated for each of the 4 components. To clarify the differences between components, repeated-measures ANOVA was conducted. Further, pairwise *t* tests with a Bonferroni correction were used for post hoc analysis.

Ethical Considerations

The study's description stated that participation was voluntary and would not affect students' grades. Only those medical students who agreed to participate accessed the questionnaire. The study was approved by the Ethics Review Committee of Jichi Medical University School of Medicine, Shimotsuke, Japan (approval number 24-137).

Informed consent was obtained electronically; students agreed to participate by accessing the questionnaire after reading the study description. To ensure privacy and confidentiality, all data were anonymized and stored securely with restricted access. No compensation was provided for participation.

Results

Participant Characteristics

In this study, a total of 118 fourth-year medical students participated. All participants provided complete data for both the MCQ and descriptive essay questions, which were the main aspects analyzed in this study. The participants were 71 (60.2%) male and 47 (39.8%) female students, all of whom were younger than 30 years.

MCQ Scores: Descriptive Statistics and Overall Effectiveness

The MCQ's internal consistency and reliability were assessed using the Cronbach α value at each time point. The α coefficients for the total score were 0.651 at pretest, 0.765 at midtest, and 0.782 at posttest. The α coefficients for the 8 subscales ranged from 0.631 to 0.760, indicating acceptable to good reliability. Table 2 depicts the detailed Cronbach α values for the total score and all subscales.

Table 2. Cronbach α coefficients for the multiple-choice questionnaire at each time point.

Factor	Pretest	Midtest	Posttest
Strengths of AI[a]	0.631	0.662	0.692
AI architecture	0.639	0.744	0.724
AI output issues	0.743	0.732	0.754
Prompts	0.657	0.707	0.696
Personal data leakage	0.76	0.708	0.77
Machine learning	0.666	0.751	0.726
Deep learning	0.704	0.729	0.757
Overfitting	0.716	0.733	0.743
Total score	0.651	0.765	0.782

[a]AI: artificial intelligence.

XSL•FO
RenderX

As per the Shapiro-Wilk test, the MCQ scores at none of the 3 time points were normally distributed (P<.001 for all values). Therefore, descriptive statistics were presented as median (IQR) values. The median total MCQ score increased across the time points, from 17 (IQR 15-18) at pretest to 18 (IQR 16-20) at midtest and 19 (IQR 17-21) at posttest (Table 3).

Table 3. Median (IQR) scores of the multiple-choice questionnaire (MCQ) at each time point. The total possible score range for the MCQ is 0-24.

Factor	Pretest, median (IQR)	Midtest, median (IQR)	Posttest, median (IQR)
Strengths of AI[a]	2 (2-3)	2 (2-3)	2 (2-3)
Mechanisms of AI	2 (2-3)	2 (2-3)	3 (2-3)
Challenges associated with AI use	2 (2-3)	2 (2-3)	2 (2-3)
Prompts	2 (1-2)	2 (1-3)	2 (1-3)
Personal information protection	3 (3-3)	3 (3-3)	3 (3-3)
Machine learning	2 (1-2)	2 (1-3)	2 (1-3)
Deep learning	1 (1-2)	2 (1-3)	2 (1-3)
Overfitting	2 (1-3)	2 (1-3)	2 (1-3)
Total score	17 (15-18)	18 (16-20)	19 (17-21)

[a]AI: artificial intelligence.

A linear mixed-effects model was used to assess the trajectory of this increase in scores. Results (Table 4) highlight the stepwise effect of the program. For instance, a statistically significant increase of 0.75 points from the baseline score at pretest was noted after the lecture phase at midtest (β=.75; P=.01). Furthermore, the end-of-program (posttest) score significantly exceeded the baseline score by 1.42 points (β=1.42; P<.001).

Table 4. Fixed effects from linear mixed - effects model predicting total score. Model: Score ~ Time + (1 | ID); n=354 observations from 118 students.

Effect	Estimate β (SE)	*t* test (*df*)	*P* value
Intercept (pretest)	16.86 (0.34)	49.55 (179)	<.001
T2 (midtest)	0.75 (0.27)	2.79 (234)	.006
T3 (posttest)	1.42 (0.27)	5.30 (234)	<.001

To address the quantitative RQ1 regarding the PBL phase's additional effect, the Wilcoxon signed-rank test was performed to compare score gains. The test found no statistically significant difference between the score gain during the lecture phase (midtest-Pretest) and that during the PBL phase (posttest-midtest) for either the total score (V=3084.5; P=.54) or any of the 8 subscales (Table 5).

Table 5. Wilcoxon signed-rank test comparing score gains between lecture (midtest-pretest) and PBL (posttest-midtest) phases.

Factor	*V* statistic	*P* value
Strengths of AI[a]	203.0	.058
Mechanisms of AI	1733.5	.21
Challenges associated with AI use	1347.0	.82
Prompts	1601.5	.76
Personal information protection	1044.0	.55
Machine learning	981.0	.61
Deep learning	120.0	.31
Overfitting	782.0	.90
Total score	3084.5	.54

[a]AI: artificial intelligence.

Qualitative RQ

To address RQ2, free-text responses to question 1 (patient explanation) and question 2 (physician precautions) were analyzed by comparing word frequencies and cooccurrence network structures between pretest and posttests.

XSL•FO
RenderX

Responses to Question 1

After the intervention, a significant increase in the number of students using words related to AI mechanisms, limitations, and applications was noted. In particular, a high number of students used "learning" (net increase of 18 students), "input" (+17), "bias" (+15), and "utilization" (+13). Transition analysis revealed that this increase was driven by new adoption; for instance, 37 (31.4%) students newly used the term "learning" and 20 (16.9%) students newly used "bias" in the posttest, indicating the acquisition of new perspectives. Contrastingly, the use of words describing the clinical case's initial facts decreased. Only a few students used words such as "symptoms" (decrease of 11 students), "examination" (–11), and "explanation" (–9). Notably, transition analysis indicated that 23 (19.5%) students who used the term "symptoms" at pretest dropped it at posttest, suggesting a shift away from merely describing clinical manifestations. Table 6 depicts the top 10 words with the largest changes and their transition details.

Table 6. Changes in word usage frequencies and transition analysis for "patient explanation" (question 1; N=118).

Word	Pretest, n (%)	Posttest, n (%)	Net change	New adoption[a], n (%)	Dropped[b], n (%)
Words with increased usage					
Learning	32 (27.1)	50 (42.4)	+18	37 (31.4)	19 (16.1)
Input	20 (16.9)	37 (31.4)	+17	29 (24.6)	12 (10.2)
Bias	10 (8.5)	25 (21.2)	+15	20 (16.9)	5 (4.2)
Utilization	12 (10.2)	25 (21.2)	+13	20 (16.9)	7 (5.9)
Misdiagnosis	7 (5.9)	20 (16.9)	+13	15 (12.7)	2 (1.7)
Information	24 (20.3)	35 (29.7)	+11	25 (21.2)	14 (11.9)
Case	31 (26.3)	40 (33.9)	+9	22 (18.6)	13 (11)
Hospital	9 (7.6)	17 (14.4)	+8	13 (11)	5 (4.2)
Convenient	4 (3.4)	10 (8.5)	+6	9 (7.6)	3 (2.5)
Judgment	32 (27.1)	38 (32.2)	+6	16 (13.6)	10 (8.5)
Words with decreased usage					
Symptoms	27 (22.9)	15 (12.7)	–12	11 (9.3)	23 (19.5)
Examination	18 (15.3)	7 (5.9)	–11	4 (3.4)	15 (12.7)
Explanation	22 (18.6)	13 (11)	–9	8 (6.8)	17 (14.4)
Presentation	12 (10.2)	4 (3.4)	–8	2 (1.7)	10 (8.5)
During	11 (9.3)	4 (3.4)	–7	2 (1.7)	9 (7.6)
Consultation	23 (19.5)	16 (13.6)	–7	8 (6.8)	15 (12.7)
View	8 (6.8)	2 (1.7)	–6	1 (0.8)	7 (5.9)
Result	29 (24.6)	24 (20.3)	–5	13 (11)	18 (15.3)
Certain	23 (19.5)	18 (15.3)	–5	8 (6.8)	13 (11)
Many	15 (12.7)	10 (8.5)	–5	5 (4.2)	10 (8.5)

[a]The term "new adoption" indicates students who did not use the word at pretest but used it at posttest, reflecting the acquisition of new concepts.

[b]"Dropped" indicates students who used the word at pretest but excluded it at posttest.

Subsequently, we assessed the structural changes in students' thinking by comparing the cooccurrence networks between pretest and posttests. The patient explanation network's Jaccard coefficient was 0.1156, indicating very low similarity between the 2 time points. This dramatic shift was driven by a large number of unique pairs at each time point. The network had only 264 shared cooccurrence pairs; however, 1030 unique pairs existed exclusively at pretest, and 989 unique pairs emerged at posttest. These quantitative findings are supported by a qualitative analysis of the network structures. In the pretest, the network was structured around the case's core elements (Figure 1). However, the network became highly complex and specific at posttest, with keywords like "learning" and "data" being directly linked to "AI" and new vocabulary, such as "bias" and "blind acceptance," explaining the limitations and risks of AI use (Figure 2).

XSL•FO
RenderX

Figure 1. The network visualizes the relationships between keywords in students' explanations to the patient before the program. Each node represents a word, and the size and color of the node indicate its degree (number of connections). The thickness and color of the edge (line) connecting 2 nodes represent the number of students who used those 2 words adjacently.

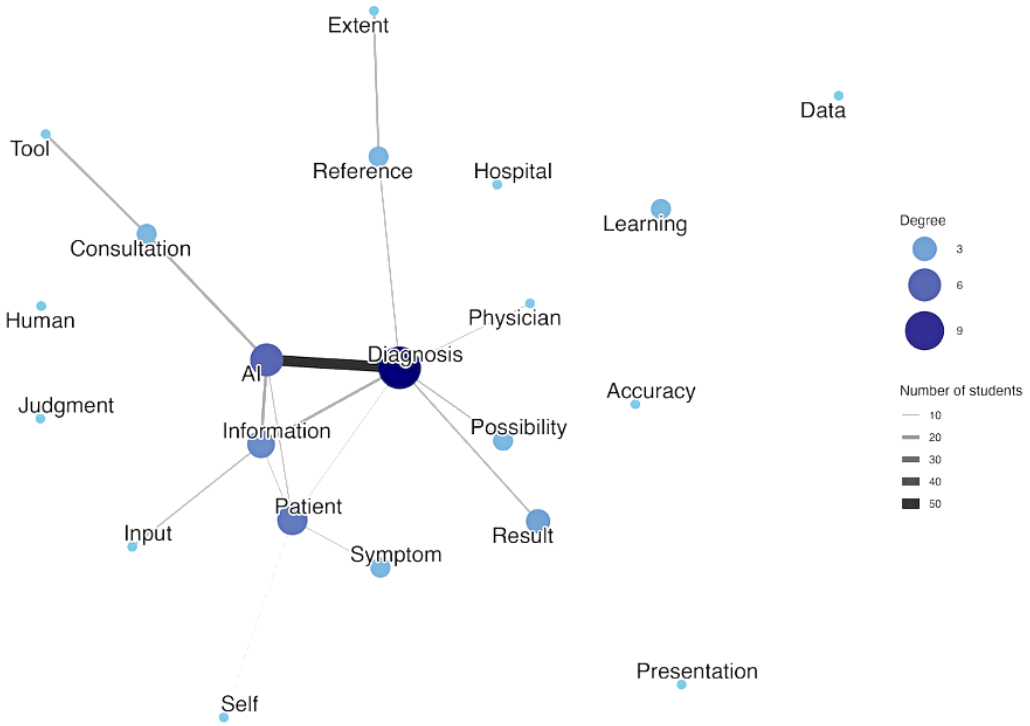
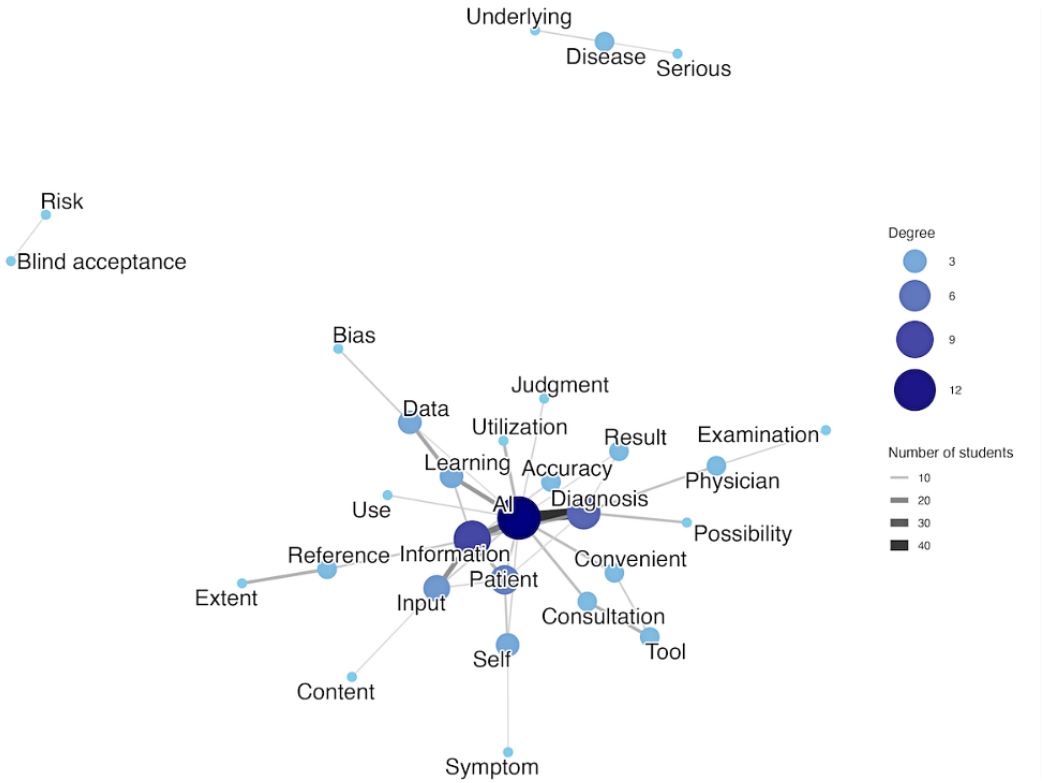


Figure 2. The network visualizes the relationships between keywords in students' explanations to the patient after the program.



Responses to Question 2

For question 2, a significant increase in the use of words related to professional and ethical responsibilities was observed. The number of students using key concepts such as "information" (net increase of 25 students), "personal info" (+14), "accuracy" (+13), and "bias" (+10) increased significantly, as well. Notably, 31 (26.3%) students newly adopted the term "personal info" in their posttest responses, suggesting a shift in their awareness of privacy issues. Contrastingly, the number of students using abstract or cautionary words, such as "possibility" (decrease of 9 students), "reference" (–5), and "keep in mind" (–5), decreased. Notably, transition analysis revealed that 30 (25.4%) students dropped the term "reference" and 24 (20.3%) students

XSL•FO
RenderX

dropped "possibility" in their posttest responses, suggesting a departure from vague expressions. Table 7 depicts the top 10 words with the largest changes and their transition details.

Table 7. Changes in word usage frequencies and transition analysis for "physician precautions" (question 2; N=118).

Word	Pretest, n (%)	Posttest, n (%)	Net change	New adoption[a], n (%)	Dropped[b], n (%)
Words with increased usage					
Information	56 (47.5)	81 (68.6)	+25	38 (32.2)	13 (11)
AI[c]	49 (41.5)	66 (55.9)	+17	29 (24.6)	12 (10.2)
Personal info	36 (30.5)	50 (42.4)	+14	31 (26.3)	17 (14.4)
Accuracy	22 (18.6)	35 (29.7)	+13	26 (22)	13 (11)
Input	2 (1.7)	12 (10.2)	+10	12 (10.2)	2 (1.7)
Bias	5 (4.2)	15 (12.7)	+10	13 (11)	3 (2.5)
Patient	18 (15.3)	26 (22)	+8	17 (14.4)	9 (7.6)
Assistance	11 (9.3)	18 (15.3)	+7	12 (10.2)	5 (4.2)
Final	14 (11.9)	20 (16.9)	+6	15 (12.7)	9 (7.6)
View	0 (0)	5 (4.2)	+5	5 (4.2)	0 (0)
Words with decreased usage					
Possibility	37 (31.4)	28 (23.7)	–9	15 (12.7)	24 (20.3)
Reference	67 (56.8)	61 (51.7)	–6	24 (20.3)	30 (25.4)
Oneself	17 (14.4)	11 (9.3)	–6	4 (3.4)	10 (8.5)
Keep in mind	15 (12.7)	10 (8.5)	–5	7 (5.9)	12 (10.2)
Presentation	12 (10.2)	7 (5.9)	–5	5 (4.2)	10 (8.5)
Knowledge	15 (12.7)	10 (8.5)	–5	5 (4.2)	10 (8.5)
Examination	8 (6.8)	4 (3.4)	–4	3 (2.5)	7 (5.9)
Caution	54 (45.8)	50 (42.4)	–4	24 (20.3)	28 (23.7)
Output	5 (4.2)	1 (0.8)	–4	1 (0.8)	5 (4.2)
Source	17 (14.4)	13 (11)	–4	6 (5.1)	10 (8.5)

[a]The term "new adoption" indicates students who did not use the word at pretest but used it at posttest, reflecting the acquisition of new concepts.

[b]"Dropped" indicates students who used the word at pretest but excluded it at posttest.

[c]AI: artificial intelligence.

The cooccurrence network for physician precautions showed a significant change, as well. The Jaccard coefficient was 0.1213, confirming the occurrence of a major structural transformation. The low score was driven by a large number of unique pairs. Only 191 pairs were shared, whereas 631 unique pairs were found exclusively at pretest, and 753 new pairs appeared at posttest. The higher number of unique pairs found at posttest compared to pretest suggests that students not only abandoned previously acquired cautionary topics but also actively explored a wider range of new professional and ethical concerns. These quantitative results were corroborated by a qualitative analysis of network structures. The pretest network comprised 3 separate, unconnected clusters (Figure 3), which transformed into a single, integrated network at posttest, where "AI" and "information" acted as hubs (Figure 4).

XSL•FO
RenderX

Figure 3. The network visualizes the relationships between keywords regarding precautions for physicians before the program.

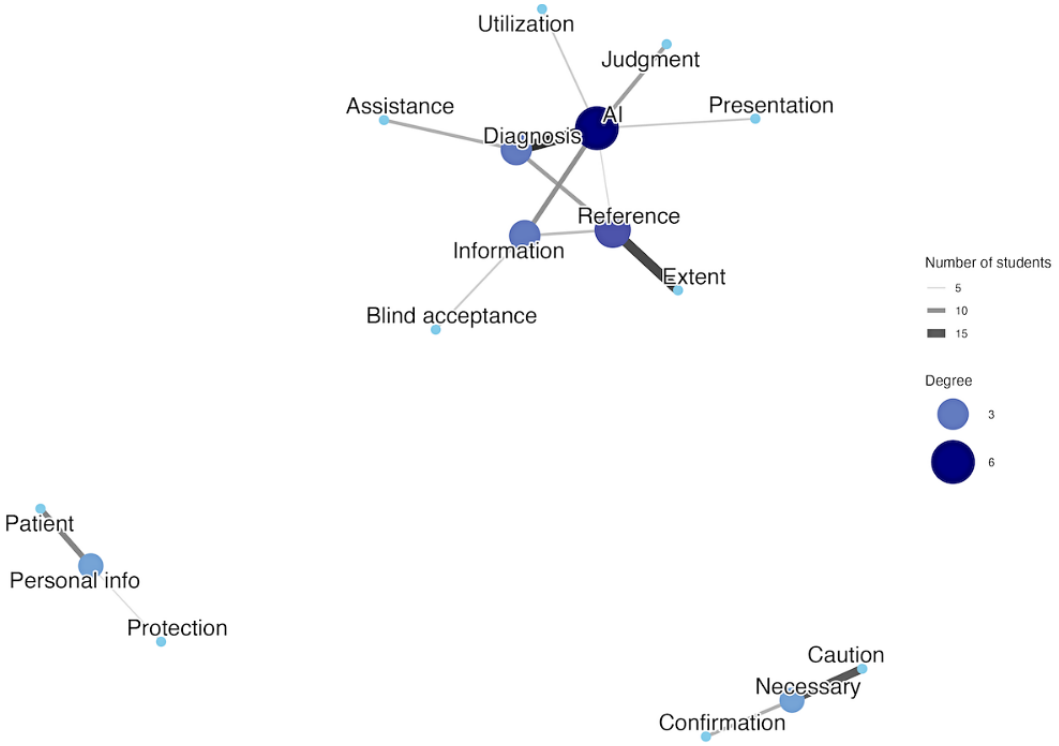
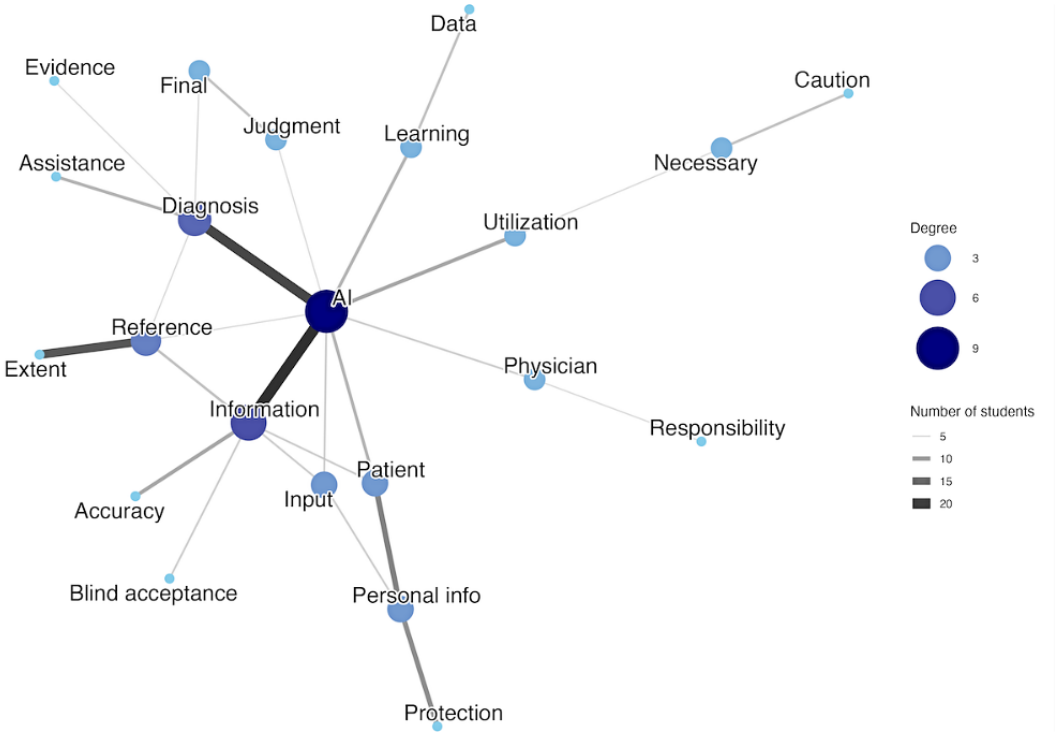


Figure 4. The network visualizes the relationships between keywords regarding precautions for physicians after the program.



Students' Perceptions of the Program

Among the 118 participants, 76 (64.4%) completed an optional postprogram survey based on the ARCS model. On a 5-point Likert scale, students rated the program highly for attention (mean 4.01, SD 0.62), relevance (mean 4.20, SD 0.71), confidence (mean 3.78, SD 0.87), and satisfaction (mean 4.03, SD 0.75). A repeated-measures ANOVA was conducted to compare and reveal a statistically significant difference between the mean scores of the 4 ARCS components ($F_{3,\ 225}$=13.15; P<.001). Additionally, post hoc tests with Bonferroni correction revealed that the mean score for confidence was significantly lower than the scores for attention, relevance, and satisfaction. Furthermore, relevance was rated significantly higher than satisfaction.

XSL•FO
RenderX

Discussion

Principal Findings

This study's primary objective was to evaluate the effectiveness of a newly developed PBL program. Further, this study addresses the following RQs:

RQ1: The educational program improves students' objective knowledge of AI. Accordingly, we hypothesized that foundational knowledge would increase after the lecture phase (from pretest to midtest) and the subsequent PBL phase (from midtest to posttest) would produce a significant additional increase in practical and ethical knowledge.

RQ2: The educational program qualitatively transforms students' thinking regarding AI-related ethical challenges from an abstract understanding to a concrete and structured thought process based on the clinical context.

Regarding RQ1, although a significant improvement in MCQ scores was observed for the program in its entirety, no statistically significant "additional effect" of the PBL phase was observed. Regarding RQ2, text mining results suggested that after the intervention, students' thinking underwent a qualitative transformation, which was characterized by the use of specific and professional terms, such as "bias," "personal information," and "guidelines," and the development of a structured thought process.

Interpretation of Findings

Regarding RQ1, a clear additional effect of PBL was not detected in MCQ scores due to 2 main reasons. First, the knowledge measured by MCQ was the foundational content that probably had been efficiently acquired during the lecture phase. The second reason is the limitation of the MCQ format itself. As argued by van der Vleuten and Schuwirth [27] in their paper on PBL assessment, traditional standardized knowledge tests cannot adequately assess abilities such as clinical reasoning and metacognition, which are fostered by PBL. Consequently, the lack of significant MCQ score improvement during the PBL phase does not imply a lack of learning; rather, it suggests that the unique contribution of PBL lies in the qualitative structural transformation of students' thinking—specifically, a shift from abstract knowledge to context-based application—which is difficult to capture with standard testing. Although MCQs are suitable to measure objective knowledge, they may have low sensitivity in capturing changes in the higher-order practical skills fostered by PBL. Indeed, reviews of earlier research on PBL's effectiveness indicate that whereas knowledge acquisition is comparable to or better than traditional lecture-based learning, PBL is more effective in improving higher-order skills, such as communication, problem-solving, self-learning, and metacognitive competencies [28,29]. Therefore, the reason a statistically significant additional effect of PBL was not detected in MCQ scores is that PBL's true effect was reflected in the qualitative changes in students' written responses, rather than the multiple-choice questions themselves.

This is clarified by the results for RQ2. The quantitative text analysis revealed that the program structurally transformed students' thinking. Among the responses to question 1 (patient explanation), the increased use of words related to AI's mechanisms and limitations was driven by new adoption. For instance, 37 (31.4%) students newly adopted the term "learning" and 20 (16.9%) newly adopted "bias." This suggests that students based their explanations of the bases and limitations of AI's judgments on the transparency principle (explainability) and accountability principle detailed in the AI Utilization Guideline and AI Business Guideline [19,20]. Conversely, 23 (19.5%) students dropped the term "symptoms," indicating a shift away from merely describing clinical manifestations to addressing the underlying logic of the AI. Similarly, among the responses to question 2 (physician precautions), the significant increase in the use of words related to ethical and professional responsibilities was also driven by new adoption. Notably, 31 (26.3%) students newly adopted "personal info," which can be interpreted as the result of students' efforts to learn and apply the privacy principle [19,20]. Simultaneously, we observed a decrease in abstract vocabulary. Transition analysis revealed that 30 (25.4%) students dropped the term "reference" and 24 (20.3%) dropped "possibility." This reduction in vague or passive expressions suggests that students adopted a more specific and responsible professional stance. The cooccurrence network analysis' quantitative results support the inference that the qualitative change in thinking was accompanied by the structuring of knowledge. The cooccurrence network for question 1 revealed a very low Jaccard coefficient (0.1156) between the pretest and postintervention networks. This indicates the significant transformation of the contextual relationships between words. For 264 common pairs, there were 1030 unique pairs before and 989 unique pairs after the intervention; accordingly, new knowledge and concepts were established by replacing traditional knowledge. Moreover, the Jaccard coefficient for the cooccurrence network of question 2 was low (0.1213), indicating a similar structural change. In the question 2 cooccurrence network, whereas "AI's function," "privacy," and "general caution" existed as separate clusters at pretest, a more integrated network was formed with the word "information" as a new hub, connecting "AI" and "diagnosis" with important concepts like "accuracy," at posttest. The study's quantitative and qualitative findings resonate with the argument of McCoy et al [12] that medical students must possess knowledge about AI. They argue that future physicians must be able to not only "use" AI but also "interpret" its results, recognize potential errors and biases, and "explain" the results and processes to patients and other health care professionals. In this study, the qualitative improvement in the postintervention descriptions highlight that students' abilities expanded from the "use" perspective to the "interpret and explain" perspective.

One reason for fostering this qualitative transformation of thinking is interpreted from the perspective of the 4 components of the ARCS model of motivation. First, the clinical case presented as a video at the beginning of the group work session effectively captured students' attention (mean rating 4.01, SD 0.62). A systematic review of emotions' role in medical education highlights the significance of emotional experiences in the learning process [30]. The video depicted a medical student facing an ethical dilemma and potentially fostered emotional engagement, thereby enhancing the students' motivation for subsequent learning. Second, their awareness of

XSL•FO
RenderX

the challenge involved in responding to a patient who is overreliant on AI probably enabled students to clearly understand the learning content's relevance to their future practice (mean rating 4.20, SD 0.71). This educational design aligns with the principles of "authentic learning," which involves the solving of complex, real-world contextual problems to enhance learners' competency in the practical application of knowledge [31]. Further, students perceived the case in this program as a personal and relevant task due to the effective functioning of the elements of authentic learning. On the other hand, the students' confidence in addressing a complex ethical issue resulted in a mean rating of 3.78 (SD 0.87), which was significantly lower than the values for attention, relevance, and satisfaction ($F_{3,\ 225}$=13.15; P<.001). Therefore, when the program prompted a qualitative transformation in the students' thinking, a single intervention was insufficient for them to acquire complete confidence in such a complex and uncertain topic as AI ethics. This program was limited to considering the explanation's content to the patient and his son and did not provide any opportunities for practical application, such as a role-play of the explanation. This lack of practical experience probably contributed to the students' low confidence in applying knowledge. An earlier study highlights the importance of maintaining a balance between confidence and humility and adopting a "not-knowing" stance as a key clinical leadership competency aligned with the Japanese health care culture [31]. The low confidence scores observed in our study were interpreted as not any failure to learn, but rather the development of the intellectual humility required by a professional, particularly those confronting complex and uncertain topics such as AI ethics.

However, the high ratings for components other than confidence, particularly attention and relevance, comprised a strong motivation for students to clearly understand the subject's importance and engage in self-directed learning [21]. The effective functioning of these elements enabled the students to complete a learning process involving deep thought, and the resulting insights probably led to the final Satisfaction (mean rating 4.03, SD 0.75) and structuring of thought observed in this study.

Research Significance and Limitations and Future Research Scope

The primary significance of this study is that it develops, implements, and comprehensively evaluates an educational program themed on a realistic and novel clinical challenge that may be faced by physicians in this era of widespread AI use: responding to a patient who relies heavily on AI. Presenting this ethical dilemma as a video teaching material has pedagogical validity because a systematic literature review indicates that video materials can effectively stimulate student discussion in ethics education [32]. Furthermore, the visualization of the program's educational effect as not only a change in objective knowledge but also a qualitative and structural change in thinking using text mining is a novel approach to evaluate the effects of constructivist learning on complex topics such as AI ethics.

This study has several limitations. First, this study used a single-group pretest and posttest design without a control group. Therefore, while we observed significant changes, we cannot fully rule out the influence of external factors or natural maturation. Second, because it was conducted at a single university in Japan, the generalizability of its results may be limited by the specific cultural and educational context. However, we emphasize that the core framework of this program—integrating technical AI knowledge with ethical PBL scenarios—is adaptable. Future multiinstitutional collaborative studies are needed to validate whether similar structural changes in thinking occur in different educational and cultural contexts. Third, this study considered only the program's short-term effects on learning. Therefore, the participating students must be subjected to a follow-up study when they proceed to their clinical clerkships to verify the learning effects' persistence. Clarifying how the qualitative transformation in thinking indicated by this study influences students' decision-making and behavior in clinical settings is important to evaluate the value of this type of educational programs. Fourth, as mentioned earlier, the MCQ format may not fully capture the effects of PBL. Therefore, there is an urgent need to develop and validate new assessment measures that can objectively assess the practical reasoning and communication skills that cannot be captured by applying MCQs alone.

Conclusions

The results of this study suggest that the PBL program developed and implemented in this study influenced both medical students' knowledge of AI and their thoughts on its ethical challenges. Although the program in its entirety resulted in a significant increase in objective knowledge, a statistically significant additional effect was not confirmed from the PBL phase alone. However, there was a clear change in the qualitative data revealed by text mining: Students' thinking underwent qualitative transformation from an abstract level to a concrete and structured thought process based on the clinical context. Therefore, this study indicates the potential of PBL as an effective pedagogical method to foster higher-order thinking skills in complex topics, such as AI ethics. Hence, this program's design and evaluation framework can serve as a practical model to develop and evaluate future AI-related curricula in medical education.

Acknowledgments

We extend our sincere gratitude to all the students who participated in this study. Our heartfelt thanks are also directed to Ms Akane Sumi for her cooperation in the video recording and to Ms Kazumi Kato for her professional administrative support at the graduate school.

XSL•FO
RenderX

Funding

This work was supported by research funding at the Medical Education Center of Jichi Medical University.

Data Availability

The datasets generated and/or analyzed during this study are not publicly available due to ethical constraints but are available from the corresponding author on reasonable request.

Authors' Contributions

YO was involved in all study aspects. YA contributed to the administration of the educational program and the analysis and interpretation of data. SK contributed to the creation of the evaluation instruments and the analysis and interpretation of data. YM contributed to the conceptualization and supervision of all study aspects and critically reviewed the manuscript. TK and MSY contributed significantly to the administration of problem-based learning sessions and the writing of the manuscript. All the authors reviewed and approved the final manuscript.

Conflicts of Interest

None declared.

References

1. Masters K. Artificial intelligence in medical education. Med Teach 2019;41(9):976-980. [doi: 10.1080/0142159X.2019.1595557] [Medline: 31007106]
2. Paranjape K, Schinkel M, Nannan Panday R, Car J, Nanayakkara P. Introducing artificial intelligence training in medical education. JMIR Med Educ 2019;5(2):e16048 [FREE Full text] [doi: 10.2196/16048] [Medline: 31793895]
3. Rincón EHH, Jimenez D, Aguilar L, Flórez JMP, Tapia AER, Peñuela CLJ. Mapping the use of artificial intelligence in medical education: a scoping review. BMC Med Educ 2025;25(1):526 [FREE Full text] [doi: 10.1186/s12909-025-07089-8] [Medline: 40221725]
4. Car J, Ong QC, Erlikh Fox T, Leightley D, Kemp SJ, Švab I, Digital Health Systems Collaborative, et al. The digital health competencies in medical education framework: an international consensus statement based on a Delphi study. JAMA Netw Open 2025;8(1):e2453131 [FREE Full text] [doi: 10.1001/jamanetworkopen.2024.53131] [Medline: 39888625]
5. Longhini J, Rossettini G, Palese A. Digital Health competencies among health care professionals: systematic review. J Med Internet Res 2022;24(8):e36414 [FREE Full text] [doi: 10.2196/36414] [Medline: 35980735]
6. Lee J, Wu AS, Li D, Kulasegaram KM. Artificial intelligence in undergraduate medical education: a scoping review. Acad Med 2021;96(11S):S62-S70. [doi: 10.1097/ACM.0000000000004291] [Medline: 34348374]
7. Amiri H, Peiravi S, Rezazadeh Shojaee SS, Rouhparvarzamin M, Nateghi MN, Etemadi MH, et al. Medical, dental, and nursing students' attitudes and knowledge towards artificial intelligence: a systematic review and meta-analysis. BMC Med Educ 2024;24(1):412 [FREE Full text] [doi: 10.1186/s12909-024-05406-1] [Medline: 38622577]
8. Çalışkan SA, Demir K, Karaca O. Artificial intelligence in medical education curriculum: an e-Delphi study for competencies. PLoS One 2022;17(7):e0271872 [FREE Full text] [doi: 10.1371/journal.pone.0271872] [Medline: 35862401]
9. Weidener L, Fischer M. Artificial intelligence in medicine: cross-sectional study among medical students on application, education, and ethical aspects. JMIR Med Educ 2024;10:e51247 [FREE Full text] [doi: 10.2196/51247] [Medline: 38180787]
10. Hale J, Alexander S, Wright ST, Gilliland K. Generative AI in undergraduate medical education: a rapid review. J Med Educ Curric Dev 2024;11. [doi: 10.1177/23821205241266697]
11. Singla R, Pupic N, Ghaffarizadeh S, Kim C, Hu R, Forster BB, et al. Developing a Canadian artificial intelligence medical curriculum using a Delphi study. NPJ Digit Med 2024;7(1):323 [FREE Full text] [doi: 10.1038/s41746-024-01307-1] [Medline: 39557985]
12. McCoy LG, Nagaraj S, Morgado F, Harish V, Das S, Celi LA. What do medical students actually need to know about artificial intelligence? NPJ Digit Med 2020;3:86 [FREE Full text] [doi: 10.1038/s41746-020-0294-7] [Medline: 32577533]
13. Weidener L, Fischer M. Proposing a principle-based approach for teaching AI ethics in medical education. JMIR Med Educ 2024;10:e55368 [FREE Full text] [doi: 10.2196/55368] [Medline: 38285931]
14. Medical Education Model Core Curriculum Expert Research Committee. The model core curriculum for medical education in Japan 2022 revision. Ministry of Education, Culture, Sports, Science and Technology. URL: https://www.mext.go.jp/content/20250411-mxt_igaku-000028108_00003-2.pdf [accessed 2025-12-23]
15. Onoue T, Asada Y, Imafuku R, Kou S, Takami H, Takahashi Y, et al. Developing competencies relating to information science and technology in Japanese undergraduate medical education. Med Teach 2024;46(sup1):S31-S37 [FREE Full text] [doi: 10.1080/0142159X.2024.2385199] [Medline: 39545502]
16. Ota Y, Asada Y, Mieno M, Matsuyama Y. Competencies required to make use of information science and technology among Japanese medical students: a cross-sectional study. BMC Med Educ 2024;24(1):840 [FREE Full text] [doi: 10.1186/s12909-024-05786-4] [Medline: 39107733]

XSL•FO
RenderX

17. Taylor D, Miflin B. Problem-based learning: where are we now? Med Teach 2008;30(8):742-763. [doi: 10.1080/01421590802217199] [Medline: 18946818]
18. Ota Y, Asada Y, Matsuyama Y. Proposal of a rubric for 'Competency in Utilizing IT: Information Science and Technology' education (in Japanese). Igaku Kyoiku 2025;56(1):40-41. [doi: 10.11307/mededjapan.56.1_40]
19. Ministry of Internal Affairs and Communications (MIC), Japan. AI utilization guidelines: practical reference for AI utilization. 2019 Presented at: The Conference toward AI Network Society; August 9, 2019; Mita Kyoyo Kaigisho URL: https://www.soumu.go.jp/main_content/000658284.pdf
20. Ministry of Internal Affairs and Communications, Ministry of Economy, Trade and Industry. AI guidelines for business (ver1.1). 2025. URL: https://www.meti.go.jp/shingikai/mono_info_service/ai_shakai_jisso/pdf/20240419_14.pdf [accessed 2025-12-23]
21. Ota Y, Asada Y, Matsuyama Y. Implementation of a PBL program on the credibility of AI-based clinical judgments: insights from the ARCS model (in Japanese). Igaku Kyoiku 2025;56(3):181-188. [doi: 10.11307/mededjapan.56.3_181]
22. Keller J. Motivational Design for Learning and Performance: The ARCS Model Approach. London: Springer; 2010.
23. R Core Team. R: a language and environment for statistical computing. 2023. URL: https://www.R-project.org/ [accessed 2025-12-23]
24. Silge J, Robinson D. tidytext: text mining and analysis using tidy data principles in R. J Open Source Softw 2016;1(3):37. [doi: 10.21105/joss.00037]
25. Ishida M. RMeCab: an R interface to MeCab, a Japanese morphological analysis engine. GitHub. 2008. URL: https://github.com/IshidaMotohiro/RMeCab [accessed 2025-12-23]
26. Kang K, Han SJ, Chun J, Kim H. Research trends related to childhood and adolescent cancer survivors in South Korea using word co-occurrence network analysis. Child Health Nurs Res 2021;27(3):201-210 [FREE Full text] [doi: 10.4094/chnr.2021.27.3.201] [Medline: 35004510]
27. van der Vleuten CPM, Schuwirth LWT. Assessment in the context of problem-based learning. Adv Health Sci Educ Theory Pract 2019;24(5):903-914 [FREE Full text] [doi: 10.1007/s10459-019-09909-1] [Medline: 31578642]
28. Trullàs JC, Blay C, Sarri E, Pujol R. Effectiveness of problem-based learning methodology in undergraduate medical education: a scoping review. BMC Med Educ 2022;22(1):104 [FREE Full text] [doi: 10.1186/s12909-022-03154-8] [Medline: 35177063]
29. Stentoft D. Problem-based projects in medical education: extending PBL practices and broadening learning perspectives. Adv Health Sci Educ Theory Pract 2019;24(5):959-969. [doi: 10.1007/s10459-019-09917-1] [Medline: 31641941]
30. Lee J, Campbell S, Choi M, Bae J. Authentic learning in healthcare education: a systematic review. Nurse Educ Today 2022;119:105596. [doi: 10.1016/j.nedt.2022.105596] [Medline: 36283196]
31. Shikama Y, Oikawa S, Stanyon M, Yasuda M, Otani K. Culturally-aligned clinical leadership competencies for effective teamwork in Japanese healthcare. BMC Med Educ 2024;24(1):1315 [FREE Full text] [doi: 10.1186/s12909-024-06272-7] [Medline: 39548540]
32. Rattani A, Kaakour D, Syed RH, Kaakour A. Changing the channel on medical ethics education: systematic review and qualitative analysis of didactic-icebreakers in medical ethics and professionalism teaching. Monash Bioeth Rev 2021;39(1):125-140. [doi: 10.1007/s40592-020-00120-2] [Medline: 33070300]

Abbreviations

AI: artificial intelligence
ARCS: attention, relevance, confidence, and satisfaction
IST: information science and technology
MCQ: multiple-choice questionnaire
PBL: problem-based learning
RQ: research question

Edited by D Chartash; submitted 22.Sep.2025; peer-reviewed by N Bievre, A Marušić; comments to author 07.Dec.2025; accepted 21.Dec.2025; published 14.Jan.2026.

Please cite as:
Ota Y, Asada Y, Kubo S, Kanno T, Yagi MS, Matsuyama Y
Evaluation of a Problem-Based Learning Program's Effect on Artificial Intelligence Ethics Among Japanese Medical Students: Mixed Methods Study
JMIR Med Educ 2026;12:e84535
URL: https://mededu.jmir.org/2026/1/e84535
doi:10.2196/84535
PMID:

XSL•FO
RenderX

©Yuma Ota, Yoshikazu Asada, Saori Kubo, Takeshi Kanno, Machiko Saeki Yagi, Yasushi Matsuyama. Originally published in JMIR Medical Education (https://mededu.jmir.org), 14.Jan.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Education, is properly cited. The complete bibliographic information, a link to the original publication on https://mededu.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

AI-Driven Objective Structured Clinical Examination Generation in Digital Health Education: Comparative Analysis of Three GPT-4o Configurations

Zineb Zouakia[1,2], MD; Emmanuel Logak[1,2], MD; Alan Szymczak[1,2], MD; Jean-Philippe Jais[3,4], MD, PhD; Anita Burgun[1,2], MD, PhD; Rosy Tsopra[1,2,5], MD, PhD

[1]Clinical Bioinformatics Laboratory, Imagine Institute, Université Paris Cité, INSERM UMR1163, Paris, France

[2]Department of Medical Informatics, Hôpital Européen Georges Pompidou, Hôpital Necker Enfants Malades, AP-HP, Paris, France

[3]Imagine Institute, Université Paris Cité, Paris, France

[4]Biostatistic Unit, Necker University Hospital, AP-HP, Paris, France

[5]UPPERS US 007 F-75006, Université Paris Cité, Paris, France

Corresponding Author:
Zineb Zouakia, MD
Clinical Bioinformatics Laboratory
Imagine Institute
Université Paris Cité, INSERM UMR1163
24 Bd du Montparnasse, 75015 Paris
Paris, F-75006
France
Phone: 33 668169844
Email: zineb.zouakia@gmail.com

Abstract

Background: Objective Structured Clinical Examinations (OSCEs) are used as an evaluation method in medical education, but require significant pedagogical expertise and investment, especially in emerging fields like digital health. Large language models (LLMs), such as ChatGPT (OpenAI), have shown potential in automating educational content generation. However, OSCE generation using LLMs remains underexplored.

Objective: This study aims to evaluate 3 GPT-4o configurations for generating OSCE stations in digital health: (1) standard GPT with a simple prompt and OSCE guidelines; (2) personalized GPT with a simple prompt, OSCE guidelines, and a reference book in digital health; and (3) simulated-agents GPT with a structured prompt simulating specialized OSCE agents and the digital health reference book.

Methods: Overall, 24 OSCE stations were generated across 8 digital health topics with each GPT-4o configuration. Format compliance was evaluated by one expert, while educational content was assessed independently by 2 digital health experts, blind to GPT-4o configurations, using a comprehensive assessment grid. Statistical analyses were performed using Kruskal-Wallis tests.

Results: Simulated-agents GPT performed best in format compliance and most content quality criteria, including accuracy (mean 4.47/5, SD 0.28; *P*=.01) and clarity (mean 4.46/5, SD 0.52; *P*=.004). It also had 88% (14/16) for usability without major revisions and first-place preference ranking, outperforming the other configurations. Personalized GPT showed the lowest format compliance, while standard GPT scored lowest for clarity and educational value.

Conclusions: Structured prompting strategies, particularly agents' simulation, enhance the reliability and usability of LLM-generated OSCE content. These results support the use of artificial intelligence in medical education, while confirming the need for expert validation.

(JMIR Med Educ 2026;12:e82116) doi:10.2196/82116

KEYWORDS

large language models; generative artificial intelligence; ChatGPT; GPT-4o; medical education; objective structured clinical examination; digital health; digital health education; medical informatics; prompt design

XSL•FO
RenderX

Introduction

Since ChatGPT was launched by OpenAI in 2022, large language models (LLMs) have undergone rapid development, leading to growing interest in medical education [1-6]. Their ability to produce diverse and context-specific texts from simple prompts offers new opportunities to create and improve educational resources.

In medical education, several studies explored the ability of LLMs to answer medical examination questions [7-12], generate multiple-choice questions [13-16], or simulate patients [17-26], with results considered promising by both educators and learners. However, using LLMs to generate clinical scenarios is still new and has only been explored in a few studies, across three situations:

1. Generation of clinical cases or vignettes (ie, scenarios given to students without interaction): LLMs demonstrated their ability to generate cases considered accurate and appropriate in pharmacology [27], but also useful [28] and usable (with moderate or minor edits) in general medicine [29]. However, the authors noticed some issues with realism [28] and precision [28,29].
2. Generation of clinical simulations (ie, scenarios given to students with interaction, such as simulated patients): LLMs demonstrated their ability to generate accurate cases in asthma [30], as well as realistic cases in the nursing field [31]. They also offered potential time savings, reducing development time by 2.6 hours compared to experts [32]. However, limitations remained, such as the omission of critical patient characteristics [31].
3. Generation of Objective Structured Clinical Examination (OSCE) stations: These scenarios, introduced by Harden and Gleeson in 1979 [33], are key assessment tools in medical education. OSCEs consist of simulated clinical situations conducted within a controlled and reproducible environment [34,35], where students interact with a standardized participant (SP), playing the role of a patient or health care professional. Students are observed and evaluated by an examiner using a standardized scoring rubric, based on predefined criteria. Previous studies explored the potential of LLMs to simulate standardized participants [36-38], examiners [39-43], or students [44-46]. Conceptual work has also discussed their potential to enhance OSCE development [47-49]. However, to our knowledge, no study has yet generated OSCE stations using LLMs.

Designing OSCE stations is a complex and resource-intensive process [50]. It requires considerable clinical and educational expertise to produce realistic, standardized scenarios that accurately assess targeted competencies [51]. The need for precision, clarity, and reproducibility, combined with time and human resource constraints in educational teams, makes OSCE development challenging and burdensome [52,53]. In domains like digital health, this challenge is amplified by a shortage of available experts. In this context, LLMs could help educators by assisting them in creating OSCE stations.

To explore this hypothesis, we aimed to design and assess 3 configurations for generating OSCE stations using ChatGPT in the area of digital health, an emerging field in medical education.

Methods

Background

The study follows the recommendations of the METRICS (Model, Evaluation, Timing, Range/Randomization, Individual factors, Count, Specificity) checklist, for designing and reporting studies involving generative language models in medical education [54] (Multimedia Appendix 1).

To analyze the ability of ChatGPT to generate OSCE stations in digital health, we tested 3 configurations of the ChatGPT-4o model (Table 1). These differed by prompt type (simple or advanced) and the documents provided to the LLM (reference books for OSCE or digital health or both).

The configurations of the ChatGPT-4o model are listed as follows:

1. Configuration 1: standard GPT, used the free version of ChatGPT-4o with a simple prompt and the reference book for OSCE construction (named “OSCE Vademecum”).
2. Configuration 2: personalized GPT, used the paid version of custom GPT, with a simple prompt, the OSCE Vademecum, and the reference book on digital health from Université Paris Cité (named “UPCité reference book”).
3. Configuration 3: simulated-agents GPT, used the paid version of custom GPT, with a prompt simulating a fictional multiagent system specialized in OSCE construction (aligned with the OSCE Vademecum), and the UPCité reference book.

XSL•FO
RenderX

Table 1. Description of the 3 GPT configurations analyzed.

Characteristics	Standard GPT	Personalized GPT	Simulated-agents GPT
Objective	Simulate minimal and spontaneous use of GPT	Simulate personalized use of GPT with a knowledge base specific to the medical field	Simulate advanced use of GPT with specialized assistants and a knowledge base specific to the medical field
Model	GPT-4o	GPT-4o	GPT-4o
Account type	Free	Paid	Paid
Prompt[a]	Simple	Simple	Simulated multiagent system
OSCE[b] Vademecum[c]	PDF file	PDF file	Instructions embedded in the prompt
UPCité reference book[d] (knowledge base)	—[e]	5 Microsoft Word documents	5 Microsoft Word documents

[a]In French (Multimedia Appendix 2).

[b]OSCE: Objective Structured Clinical Examination.

[c]National reference document for designing OSCEs (67-page PDF).

[d]The UPCité reference book was written by 25 educators from 7 health disciplines at Université Paris Cité. It is the reference book for all health students (medicine, pharmacy, nursing, rehabilitation, midwifery, and dentistry). It is divided into 5 booklets: health data, cybersecurity, communication, digital tools in health care, and telehealth (5 Word documents, a total of 222 pages).

[e]Not applicable.

To allow comparison, we provided the same sets of digital health competencies to all 3 configurations, following the process listed in Textbox 1.

Textbox 1. Digital health competencies for configurations.

- 8 digital health topics were first selected: 5 were single-theme ("Health data, Communication, Cybersecurity, Digital tools, Telehealth"), and 3 were multiple-theme ("Communication and Digital tools"; "Communication, Digital tools, and Cybersecurity"; and "Telehealth, Health data, and Communication").
- For each selected topic (eg, cybersecurity), 3 digital health competencies were then provided to each configuration (eg, responding to a cyberattack).

  Each configuration then generated an OSCE station related to these predefined competencies.
- In total, 24 OSCE stations were generated (8 topics × 3 configurations), all on the same day, June 1, 2025, during a single 4-hour session.
- Each generated OSCE was in French and included 4 components:
  - The primary and secondary learning domains,
  - The case vignette for the student,
  - The standardized observation checklist for examiners,
  - The script for standardized patients (SP script).

Description of the 3 Configurations Used to Generate OSCE Stations

Configuration 1: Standard GPT

This configuration aimed to assess ChatGPT's ability to generate digital health OSCE stations in a minimal configuration, involving a simple prompt and the OSCE Vademecum. It simulated a minimalist and spontaneous use case, akin to an educator working without technical assistance or a reference book.

The configuration description is as follows: (1) Model type: the free version of ChatGPT 4o [55]. (2) Prompt design: a simple prompt relied on a role-prompting strategy. The model was instructed to act as a "digital health expert and instructional designer" and create an OSCE based on 3 given digital health competencies. It was invited to review the OSCE Vademecum to choose learning domains, and then produced the OSCE's components. No sequential structure or examples were provided, making this a basic, minimalist prompt (Multimedia Appendix 2). (3) Documents used: only the OSCE Vademecum (in PDF format) was provided and served as the reference framework for structuring the OSCE station. (4) Technical configuration: ChatGPT interface, without customization and advanced features.

Configuration 2: Personalized GPT

This configuration aimed to assess whether providing a domain-specific knowledge base (here, the reference book in digital health) could improve the quality of the generated OSCE stations. It simulated the use of a personalized GPT by an educator equipped with a reference book.

XSL•FO
RenderX

The configuration description is as follows: (1) Model type: the paid version custom GPT-4o (cGPT) [56], which allows configuring GPT with a specific role using tailored instructions, example queries, and integrated knowledge bases. (2) Prompt design: the simple prompt was similar to Configuration 1, but added an instruction to review the UPCité reference book on digital health (Multimedia Appendix 2). (3) Documents used: both the OSCE Vademecum (in PDF format) and UPCité reference book (in Word format) were provided to cGPT. To optimize readability for the model, the UPCité digital health reference book was preprocessed by removing noneducational pages, simplifying formatting, and converting tables and figures to plain text. (4) Technical configuration:

1. assigning an icon and name to the cGPT
2. inserting instructions in the form of a prompt
3. adding a conversation starter with a user query, here: "I would like to design an OSCE station for assessment in digital health"
4. integrating a knowledge base, here the OSCE Vademecum and UPCité reference book.

None of the optional Custom GPT features (eg, web browsing, code interpreter, data analysis, and image generation) was used.

Configuration 3: Simulated-Agents GPT

This configuration aimed to assess whether a rigorous procedural structure, via the simulation of step-by-step simulated agents' reasoning, could improve the quality of the generated OSCE stations. It simulated the use of specialized assistants at each stage of OSCE construction by an educator equipped with a reference book.

Configuration description: (1) Model type: the same as in Configuration 2 (ie, the paid version of cGPT). (2) Prompt design: the prompt was built from instructions extracted from the OSCE Vademecum and structured to simulate a multiagent system. It followed a supervised, sequential, and specialized architecture, functioning like a processing chain in which each agent had a distinct pedagogical role in OSCE generation. The supervisor agent acted as the central coordinator. It collected the 3 competencies provided by the educator, determined the sequence of agent activation, passed contextual information between them, and ensured the overall coherence, without interfering with pedagogical content. After each step, it asked, "Would you like me to proceed to the next agent?" simulating a controlled, step-by-step process.

The following specialized agents were used: (1) The learning domain agent selected a primary and secondary domain from the predefined list included in the OSCE Vademecum, aligned with the learning objectives of the graduate medical curriculum (eg, "Education and Prevention"). (2) The vignette agent generated the case vignette for students, conforming to standardized formatting. (3) The checklist agent created the observation checklist for examiners, following the standardized format. (4) The SP script agent drafted the script for the standardized patient, following the standardized format.

Agents were activated sequentially, each receiving the outputs of the previous steps, and generated a formatted output according to OSCE guidelines and UPCité reference book (Multimedia Appendix 2).

The prompt design combined multiple prompting strategies [48,57]:

1. Role prompting by assigning a specific pedagogical role to each agent;
2. Instruction-based prompting by providing precise, structured instructions drawn from the OSCE Vademecum;
3. Chain prompting by structuring tasks into a sequenced, logical workflow;
4. Few-shot prompting by including illustrations of expected formats to improve consistency and reproducibility.

(3) Documents used: the UPCité reference book on digital health was provided in the cGPT's knowledge base. (4) Technical configuration: similar setup to Configuration 2 (provision of a name, icon, and prompt, same conversation starter, knowledge base, and no optional features).

Evaluation of the Quality of OSCE Stations Generated by the 3 Configurations

To compare the 3 configurations, OSCE stations were generated and assessed by experts. The evaluation was conducted blindly of the configuration used, with a comprehensive assessment grid based on a literature review [27-32,58-62].

Evaluation of the Format Compliance of the Generated OSCEs

A fifth-year medical student, with a background in computer science and trained in OSCE methodology, measured how well the generated OSCE adhered to the OSCE Vademecum in terms of format compliance.

The compliance checklist included 27 criteria covering (1) the validity of the learning domains, (2) the vignette format, (3) the structure of the observation checklist, and (4) the completeness of the SP script. Each criterion was assessed using a binary scale (Yes or No), focusing only on format (independently of pedagogical quality).

Evaluation of the Educational Quality of the Generated OSCEs

A duo of digital health experts, involving the head of the medical informatics ward of "Hôpital Européen Georges Pompidou and Hôpital Necker Enfants Malades AP-HP, and a resident in digital health with a computer science background, conducted the evaluation of the content quality, independently and blindly of the configuration used (each OSCE station was anonymized and randomized).

For each OSCE, 9 evaluation dimensions were assessed:

1. "Relevance of the learning domains," "Clarity," "Pedagogical validity," "Realism," "Feasibility," "Educational value," and "Originality," which were rated using a 5-point Likert scale of agreement (1: strongly disagree, 2: somewhat disagree, 3: neither agree nor disagree, 4: somewhat agree, and 5: strongly agree).

XSL•FO
RenderX

2. "Information accuracy", which was rated on a severity scale (1: major issue compromising usability, 2: major issue, 3: moderate issue, 4: minor issue, and 5: no issues identified).
3. "Overall usability," which was rated on a usability scale (1: not usable, 2: usable with major revisions, 3: usable with minor revisions, and 4: usable as is, without modification).

The criteria used were derived from the literature [27-32,58-61] and from the international OSCE guidelines [62], in order to align with international standards.

At the end of the evaluation process, OSCEs were grouped by similar topic (eg, Cybersecurity). For each group of 3 OSCEs (one per configuration), evaluators were asked to rank them, from first to third, without knowing which configuration had generated each one.

Statistical Analysis

For format-related criteria, binary variables (Yes or No) were described using frequencies and the percentage of "Yes" responses.

For educational criteria, expert ratings were averaged per OSCE and summarized using means and SD values (n=8 per configuration). Overall usability and ranking were categorical variables, analyzed without averaging expert ratings (n=16 per configuration), and described using frequencies and percentages. Due to the nonparametric distribution of data and the ordinal nature of variables, the Kruskal-Wallis test was used to compare the 3 configurations. For variables comprising multiple sub-criteria, an overall mean score per dimension was calculated and used in the Kruskal-Wallis test to determine statistical significance.

All statistical analyses were performed using the R statistical environment (R version 4.3.3) with RStudio GUI (version 2023.12.1+402; Posit PBC). The threshold for statistical significance was set at $P<.05$. Given the exploratory nature of the study, no correction for multiple testing was performed, with an emphasis on observed trends and effect sizes.

Ethical Considerations

This study did not involve human subjects. All generated outputs were fictional and intended for educational use. Therefore, in accordance with French Law n°2012-300 [63], no Institutional Review Board approval was required.

Results

Format Compliance of the OSCEs Generated

Overall, the simulated-agents GPT showed the highest format compliance, followed by the standard GPT, while the personalized GPT exhibited the lowest performance (Table 2).

Standard GPT OSCEs were fully compliant for the "learning domain" and "vignette" components. The "checklist" showed less compliance (between 3/8, 38% and 4/8, 50%) for the advanced criteria ("distinct items", "assessment of skill only", and "clear validation criteria for each item"). The "SP script" component was incomplete for the contextual clinical data of all generated cases, with information, such as symptoms, medications, and socioprofessional background, never present.

Personalized GPT OSCEs were poorly compliant for all components. The compliance was full only for 3 criteria of the "vignette" component (eg, "tasks to perform").

Simulated-agents GPT OSCEs were fully compliant for the "learning domain," "vignette," and "SP script" components. Only 4 items of the "checklist" ("observable items" and the 3 advanced criteria) were partially compliant, ranging from 12% (1/8) to 75% (6/8).

XSL•FO
RenderX

Table 2. Evaluation of the format compliance of the Objective Structured Clinical Examination generated by the configuration type.

OSCE[a] components and criteria	Standard GPT (n=8), n (%)	Personalized GPT (n=8), n (%)	Simulated-agents GPT (n=8), n (%)
Domain			
The selected primary learning domain is included in the list of the 11 domains related to the OSCE assessment	8 (100)	0 (0)	8 (100)
The selected secondary learning domain is included in the list of the 11 domains related to the OSCE assessment	8 (100)	0 (0)	8 (100)
Vignette			
States the candidate's role	8 (100)	6 (75)	8 (100)
Describes the setting	8 (100)	8 (100)	8 (100)
Introduces the SP's identity (name, gender, and age)	8 (100)	6 (75)	8 (100)
Presents the issue or reason for consultation	8 (100)	8 (100)	8 (100)
Provides tasks to be performed by the student	8 (100)	8 (100)	8 (100)
Specifies actions the student should not take	8 (100)	0 (0)	8 (100)
Checklist			
Contains 10 to 15 items	8 (100)	2 (25)	8 (100)
Each item begins with an action verb	8 (100)	5 (62)	8 (100)
Items are observable	7 (88)	6 (75)	6 (75)
Items are dichotomous (Yes or No)	8 (100)	5 (62)	8 (100)
Items are distinct (or grouped if necessary, with precise scoring instructions)	3 (38)	1 (12)	1 (12)
Assesses only skills (not attitudes or communication)	4 (50)	1 (12)	2 (25)
Provides clear validation criteria for each item	3 (38)	1 (12)	3 (38)
SP script components included			
Scenario summary	4 (50)	4 (50)	8 (100)
SP's mindset or behavior	6 (75)	6 (75)	8 (100)
Additional data	0 (0)	1 (12)	8 (100)
Opening sentence	2 (25)	0 (0)	8 (100)
Identity	7 (88)	6 (75)	8 (100)
Socioprofessional background – Hobbies	0 (0)	2 (25)	8 (100)
Personal medical history	1 (12)	1 (12)	8 (100)
Familial medical history	0 (0)	0 (0)	8 (100)
Current medications	0 (0)	0 (0)	8 (100)
Symptoms	0 (0)	0 (0)	8 (100)
Conditional disclosure of information	8 (100)	7 (88)	8 (100)
Answers to all the items present in the "checklist" are provided	6 (75)	2 (25)	8 (100)

[a]OSCE: Objective Structured Clinical Examination.

XSL•FO
RenderX

Educational Quality of the OSCEs Generated

Learning Domain Relevance

Primary domain relevance was satisfactory for all configurations, highest for the simulated-agents GPT (mean 4.38, SD 1.38). Secondary domain relevance was more variable, with the standard GPT scoring lowest (mean 2.62, SD 1.36). Differences were not statistically significant (*P*=.16; Tables 3 and 4; Figure 1).

XSL•FO
RenderX

Table 3. Evaluation of the content quality of the generated Objective Structured Clinical Examinations, by configuration type.

Evaluation domain, items, and section	Standard GPT (n=8), mean (SD)	Personalized GPT (n=8), mean (SD)	Simulated-agents GPT (n=8), mean (SD)
Learning domain relevance[a]			
Primary domain	3.50 (1.13)	3.94 (1.82)	4.38 (1.38)
Secondary domain	2.62 (1.36)	3.56 (1.76)	3.56 (1.18)
Information accuracy[b]			
No errors or inexact information			
Vignette	3.88 (0.88)	4.38 (0.88)	4.75 (0.46)
Checklist	4.06 (0.86)	4.19 (0.75)	4.88 (0.35)
SP[c] script	4.25 (0.85)	3.62 (0.69)	4.19 (0.92)
No missing information			
Vignette	4.19 (1.13)	4 (0.60)	4.81 (0.37)
Checklist	4.06 (0.82)	3.88 (0.52)	4.62 (0.44)
SP script	3.94 (1.02)	3.44 (0.86)	4.19 (0.92)
No irrelevant or unnecessary information			
Vignette	4.56 (0.73)	4.62 (0.58)	4.69 (0.37)
Checklist	4 (1)	4.31 (0.65)	4.81 (0.26)
SP script	4.62 (0.69)	4.38 (0.88)	4.56 (0.73)
No direct cues, leading formulations, or implicit answers			
Vignette	3 (1.51)	2.38 (1.75)	2.69 (1.19)
SP script	4.56 (0.82)	4.88 (0.35)	5 (0)
Clarity[a]			
Clear and comprehensible writing			
Vignette	3 (0.96)	3.62 (1.38)	4.44 (0.78)
Checklist	3.75 (0.65)	3.62 (0.64)	4.56 (0.56)
SP script	3 (0.85)	3.56 (0.98)	4.38 (0.88)
Pedagogical validity[a]			
Alignment with the competencies to be assessed			
All	3.88 (0.64)	4.06 (0.82)	4.56 (0.42)
Alignment with the expectations of a professional trained in digital health			
All	4.12 (0.52)	4.44 (0.68)	4.50 (0.46)
Realism[a]			
Realistic situation in a medical professional's practice			
All	3.62 (1.13)	3.25 (0.89)	4.12 (0.69)
Feasibility[a]			
Ease of implementation			
All	4.19 (0.80)	3.56 (1.35)	4.06 (0.98)
Feasible within an 8-minute timeframe			
All	4.19 (0.70)	4.06 (0.78)	4 (0.46)
Ease of recruiting an SP			
All	4.25 (0.85)	4.62 (0.88)	4.69 (0.59)
Educational value[a]			

XSL•FO
RenderX

Evaluation domain, items, and section	Standard GPT (n=8), mean (SD)	Personalized GPT (n=8), mean (SD)	Simulated-agents GPT (n=8), mean (SD)
Useful for promoting digital health learning			
All	3.62 (0.74)	3.94 (0.86)	4.12 (0.88)
Encourages analysis, reflection, and decision-making			
All	3.62 (0.88)	3.81 (1.07)	3.44 (0.73)
Originality[a]			
Creative potential			
All	4.06 (0.50)	3.88 (0.23)	3.50 (0.76)

[a]Criteria assessed with a Likert scale from 1: strongly disagree to 5: strongly agree.

[b]Criteria assessed with a severity scale from 1: major issues to 5: no issues.

[c]SP: standardized participant.

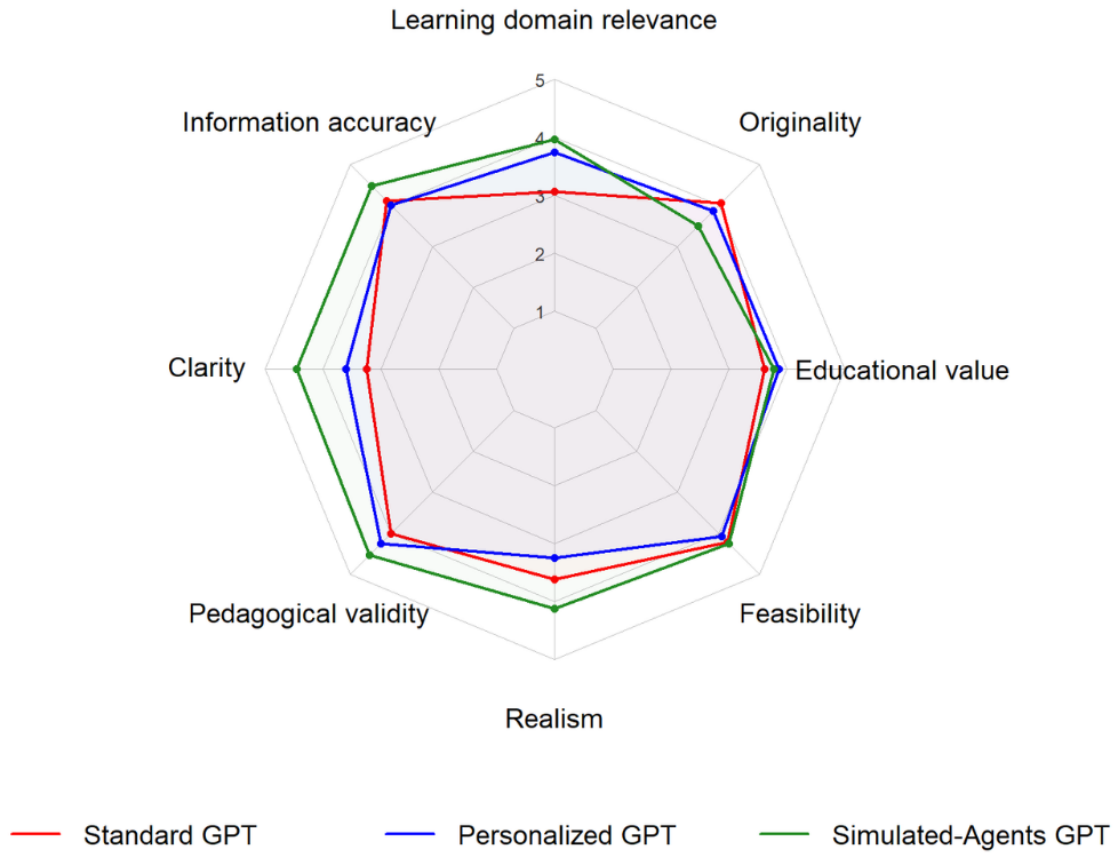
Table 4. Kruskal-Wallis test comparisons of the 3 GPT configurations by evaluation domain. To perform the statistical test, the mean score was calculated for each evaluation domain and configuration. Test results are presented as P values from the Kruskal-Wallis test.

Evaluation domain	Standard GPT, mean (SD)	Personalized GPT, mean (SD)	Simulated-agents GPT, mean (SD)	*P* value[a]
Learning domain relevance	3.06 (1.02)	3.75 (1.75)	3.97 (1.06)	.16
Information accuracy	4.10 (0.36)	4.01 (0.21)	4.47 (0.28)	.01[a]
Clarity	3.25 (0.45)	3.60 (0.83)	4.46 (0.52)	.004[a]
Pedagogical validity	4 (0.5)	4.25 (0.65)	4.53 (0.34)	.12
Realism	3.63 (1.13)	3.25 (0.89)	4.13 (0.70)	.19
Feasibility	4.21 (0.70)	4.08 (0.80)	4.25 (0.51)	.92
Educational value	3.63 (0.65)	3.88 (0.93)	3.78 (0.76)	.75
Originality	4.06 (0.49)	3.88 (0.23)	3.50 (0.76)	.32

[a]$P<.05$ (statistically significant).

XSL•FO
RenderX

Figure 1. Radar plot of evaluation scores (5-point scales) across the 3 GPT configurations. Agreement-based dimensions (eg, clarity, realism) were rated from 1 (strongly disagree) to 5 (strongly agree); information accuracy used a severity scale from 1 (major issues) to 5 (no issues). The simulated-agents GPT tended to score higher in information accuracy, clarity, realism, pedagogical validity, and learning domain relevance, while it lagged in terms of originality. All 3 configurations showed similar performance for feasibility and educational value.



Information Accuracy

Overall, there was a significant difference in information accuracy (P=.01), with the simulated-agents GPT providing the most accurate and comprehensive content (Tables 3 and 4; Figure 1). Specifically,

1. "Errors or inaccurate information" were low across configurations (mean>3.88), with a favorable trend toward the simulated-agents GPT, particularly in the "vignette" (mean 4.75, SD 0.46) and the "checklist" (mean 4.88, SD 0.35) components.
2. "Missing information" was rare in the simulated-agents GPT (mean≥4.19 across all sections), but more common in the personalized GPT, especially in the "script" (mean 3.44, SD 0.86).
3. "Irrelevant or unnecessary information" was rarely noted, with similarly high averages across configurations (mean ≥ 4).
4. "The presence of direct cues or leading formulations" was frequent in the "vignette" (2.38< mean< 3), but nearly absent in the "script", especially for the simulated-agents GPT (mean 5, SD 0).

Clarity

Clarity was significantly different among the 3 configurations (P=.004), with a trend in favor of the simulated-agents GPT, getting better results for all components (mean>4.38). The personalized GPT showed intermediate results, while the standard GPT scored lower, particularly for the "vignette" and the "script" (mean 3 for both, SD 0.96 and 0.85, respectively; Tables 3 and 4; Figure 1).

Pedagogical Validity

Pedagogical validity was high across all configurations (mean>3.88). Simulated-agents GPT scored highest (mean>4.5), followed by the personalized GPT (mean>4.06), and then the standard GPT (mean>3.88). Although differences were not statistically significant (P=.12), a favorable trend for simulated-agents GPT emerged (Tables 3 and 4; Figure 1).

Realism and Feasibility

OSCE's realism was satisfactory across configurations, with no statistically significant difference (P=.19). The simulated-agents GPT scored the highest (mean 4.12, SD 0.69), followed by the standard GPT (mean 3.62, SD 1.13) and the personalized GPT (mean 3.25, SD 0.89).

No significant difference was observed for the feasibility (P=.92). The standard GPT tended to propose OSCEs that were easiest to implement and more feasible to perform within an 8-minute timeframe (mean 4.19, SD 0.70). Ease of recruiting an SP was rated high for all, with a slight advantage for the

XSL•FO
RenderX

simulated-agents GPT (mean 4.69, SD 0.59; Tables 3 and 4; Figure 1).

Educational Value and Originality

Educational value was comparable across configurations, with no statistically significant differences (*P*=.75). The simulated-agents GPT had the highest score for the use to promote digital health learning (mean 4.12, SD 0.88), while it had a slightly lower score than personalized GPT for encouraging analysis and decision-making (mean 3.44 SD 0.73 vs mean 3.81 SD 1.07 for personalized GPT; Tables 3 and 4; Figure 1).

Regarding originality, there was no significant difference observed (*P*=.32). The standard GPT had the highest average (mean 4.06, SD 0.50), followed by the personalized GPT (mean 3.88, SD 0.23) and the simulated-agents GPT (mean 3.50, SD 0.76; Tables 3 and 4; Figure 1).

Overall Assessment

Overall, usability differed significantly across configurations (*P*=.02). For simulated-agents GPT, 88% (14/16) of OSCEs were usable without major revisions, while for standard and personalized GPT, 31% (5/16) of OSCEs required major revisions, and 12% (2/16) were unusable (Table 5).

Overall, generated case rankings differed significantly as well (*P*<.001). The simulated-agents GPT stood out, with 88% (14/16) of OSCEs ranked first and none ranked third. Conversely, the standard GPT ranked third (9/16, 56%) or second (7/16, 44%) but never ranked first. The personalized GPT received a few first-place rankings (2/16, 12%) and was mainly split between second and third places (7/16, 44% each; Table 5).

Table 5. Overall assessment of the OSCEs generated by the type of configuration. Test results are presented as P values from the Kruskal-Wallis test.

Assessment domain and items	Standard GPT (n=16), n (%)	Personalized GPT (n=16), n (%)	Simulated-agents GPT (n=16), n (%)	*P* value[a]
Usability				.02[a]
Not usable	2 (12)	2 (12)	1 (6)	
Usable with major revisions	5 (31)	5 (31)	1 (6)	
Usable with minor revisions	7 (44)	6 (38)	9 (56)	
Usable as is, without modification	2 (12)	3 (19)	5 (31)	
Ranking				<.001[a]
First	0 (0)	2 (12)	14 (88)	
Second	7 (44)	7 (44)	2 (12)	
Third	9 (56)	7 (44)	0 (0)	

[a]*P*<.05 (statistically significant).

Qualitative Feedback

In addition to quantitative ratings, qualitative feedback from experts revealed consistent issues across configurations. All tended to include excessive vignette details, often with leading cues that overly guided students (eg, “Identify and articulate the main risks associated with the use of telemedicine (technical, organizational, legal)”). Inconsistencies were noted between the vignette and SP script, with details missing or contradictory (eg, in the vignette: “Mrs. D […] has also received a message on a secure messaging platform”; in the script: “You do not own a smartphone […] You do not have a computer at home.”). Finally, even when highly rated, OSCEs were often considered too theoretical, likely due to reliance on the UPCité digital health booklet, reducing their practical, hands-on applicability for OSCEs.

Discussion

Principal Findings

This study compared 3 configurations of GPT for generating OSCE stations in the field of digital health: the standard GPT (simple prompt, OSCE Vademecum); the personalized GPT (simple prompt, OSCE Vademecum, a reference book in digital health), and the simulated-agents GPT (a prompt simulating a fictional multiagent system specialized in OSCEs, a reference book in digital health). Regarding format, the simulated-agents GPT achieved the highest compliance, with minor gaps in advanced checklist criteria. The standard GPT ranked second, with weaker checklist performance and consistently incomplete SP scripts, while the personalized GPT showed the lowest compliance across all components. Regarding educational content, the simulated-agents GPT achieved significantly higher ratings for information accuracy (*P*=.01) and clarity (*P*=.004). It also outperformed in educational validity and realism, although these differences weren’t statistically significant. Additionally, it also had significantly better usability, with 88% (14/16) of the OSCEs usable (*P*=.02). Overall, the simulated-agents GPT significantly outperformed the others, ranking first in 88% (14/16) of the OSCEs and never placing third.

Interpretation of Results

Differences between the 3 configurations may be explained by several reasons. Regarding format compliance, the differences might reflect how each configuration processed and prioritized

XSL•FO
RenderX

information. The standard GPT, which used only the OSCE Vademecum, followed the guidelines more closely than the personalized GPT that also included the UPCité reference book. This may be explained by cognitive overload, as adding an extra knowledge source increased retrieval complexity and reduced effective prioritization of key instructions [64]. These issues are amplified by the reliance on a proprietary Retrieval-Augmented Generation (RAG) system that offers only limited transparency and user control [65,66]. The simulated-agents GPT, which used a structured prompt with explicit OSCE guidelines, appeared to improve format adherence. Although it did not fully meet all checklist-format requirements, this suggests that carefully designed prompts are crucial for ensuring format accuracy.

Regarding educational content. The simulated-agents GPT, despite the 8000-character limit, produced high-quality outputs, showing that structured, modular prompts work well even with technical constraints [67]. Interestingly, the standard and personalized GPTs were perceived as more original and more effective at fostering analytical thinking. This suggests a trade-off: while structured approaches improve consistency and adherence to standards, more flexible, broader configurations can boost creativity and originality, albeit with reduced accuracy and lower consistency and control.

Finally, these findings should be interpreted with caution because technology is rapidly evolving [68,69]. The evaluation was based on a dataset generated at a specific time, and since generative models can produce different outputs with the same prompts, reproducibility may be affected. Additionally, the algorithms powering systems like OpenAI's GPT are constantly improving, which could impact future results. However, these considerations do not undermine the validity of our findings. The main goal of the study was to compare 3 OSCE generation strategies, offering insights into how generative models can support medical education rather than advocating for a fixed implementation.

Limitations

Regarding the LLM used. We focused on ChatGPT, whereas other LLMs have already been used to generate educational resources. For example, Google Bard (now called Gemini Google DeepMind) and Microsoft Bing were used to create multiple-choice questions in medical physiology [70], and LLaMA (Meta AI) was used in radiology [71]. However, prior to designing this study, we conducted preliminary tests with open-source models such as Mixtral-8x7B (Mistral AI) and Llama-3.3-70B combined with custom RAG pipelines. These initial tests were quite disappointing as they had suboptimal performances, particularly for OSCE-style tasks. Notably, recent comparative studies confirm GPT-4's superior performance on RAG-enhanced clinical reasoning tasks [72,73], which reinforces our choice of model despite the broader range of available LLMs.

Regarding the evaluation. It was conducted by 2 experts, which may introduce subjective bias and limit the diversity of evaluative perspectives. However, this limitation is mitigated by their good expertise in digital health and the rigorous evaluation process. A total of 24 OSCEs were reviewed, covering many topics, providing a diverse and valuable dataset. Each OSCE was evaluated with detailed criteria for both format and content, making the results more reliable. A key strength of this study is its strict use of the METRICS checklist [54], a recent guide for studies involving generative models in medical education. This makes the results clearer, easier to reproduce, and comparable. Additionally, while the focus on digital health allowed for targeted exploration in an emerging educational field, it may limit the generalizability of the results to other medical specialties. However, since digital health is growing across many areas, the insight from this work, especially about prompt design and model behavior, is likely useful in other competency-based assessments too.

Comparison With Other Studies

In the context of OSCEs, other studies used LLMs for tasks other than the generation of OSCE stations:

1. To simulate standardized patients [36-38]. For example, Yamamoto et al [36] used artificial intelligence (AI)–powered chatbots to simulate standardized patients, helping students improve their clinical interview skills. Students working with the chatbots scored significantly higher than those who did not.
2. To automate the OSCE assessment [39-43]. For example, Jamieson et al [41] developed an AI-based grading system to grade post-encounter notes, matching expert scores with 90% agreement while reducing manual grading effort by 91%.
3. To simulate OSCE candidates [44-46]. For instance, Huang et al [44] found that ChatGPT-4.0 performed as well as or better than junior emergency residents in history-taking and record-writing, though humans still demonstrated higher overall consultation quality.

Beyond the context of OSCEs, LLMs have been used to create clinical cases or vignettes for medical education. Like the simulated-agents GPT, studies showed high clarity of the generated content. For example, Coşkun et al [59] reported a 4.11/5 rating for the comprehensibility of the generated scenarios. Additionally, similar to our results, studies also showed strong usability, with Yanagita et al [29] indicating a 97% usability rate for GPT-generated general medicine cases. However, issues with “accuracy” and “missing information” were also reported. For example, Takahashi et al [28] reported 68% accuracy, Yanagita et al [29] reported only 58%, and Scherr et al [30] achieved 100% accuracy, but their evaluation was limited to a single and narrow area (acute asthma) Vaughn et al [31] found that 88% of nursing simulations lacked critical clinical details. In contrast, our simulated-agents GPT consistently yielded complete and context-rich OSCE stations. However, these comparisons should be made carefully, as methodology, design, and metrics differ considerably between studies.

In summary, previous works have largely focused on the use of LLMs to simulate patients, students, or evaluators, or to generate isolated clinical cases. In contrast, to our knowledge, our study is the first to use LLMs to generate OSCE stations, including the vignette, checklist, and SP script.

XSL•FO
RenderX

Future Directions

This work is part of the DigiHealth Paris Cité project [74], which aims to integrate digital health training into health curricula through innovative methods such as simulation and immersive learning. Here, the objective was to support educators in designing reliable, standardized OSCE stations for student assessment. Based on our findings, the simulated-agents GPT configuration appeared best suited as an assistant for educators, helping to quickly create structured OSCEs that still require validation before introduction into official evaluation sessions.

Regarding technical and pedagogical improvements. The simulated-agents GPT configuration could be further optimized through the refinement of the prompting strategy to enhance output quality and alignment with educational objectives. One effective approach will be iterative development, where expert feedback is continuously integrated into the LLM's generation cycle. For instance, Coşkun et al [59] used expert prompts to revise and improve 15 clinical cases directly in ChatGPT without manual editing. Similarly, Yanagita et al [29] demonstrated that feeding expert-reviewed vignettes back into the system can help guide and standardize future generations. Therefore, our next steps will include developing an interactive OSCE-generation assistant for educators using the simulated-agents GPT framework, integrating continuous expert feedback to improve content quality, and creating an internal repository of validated AI-generated OSCEs for educators at our faculty, promoting transparency, reproducibility, and shared pedagogical development.

Regarding implementation. The next phase of our work will include a 2-step validation process. First, a pilot testing of expert-validated AI-generated OSCEs with a panel of medical students will assess the realism, usability, and educational value of the generated OSCEs. Second, scaling up to entire student cohorts will enable evaluation of student satisfaction, performance, and learning outcomes at a larger scale. For example, Başaranoğlu et al [58] showed that AI-generated clinical scenario–based questions administered during a urology rotation significantly improved student performance, supporting the educational value of such approaches. Finally, a multisite expansion across other medical faculties is planned to assess transferability. These steps will help translate our findings into practical educational tools that responsibly integrate LLMs into the assessment and training of future medical professionals.

Conclusions

This study demonstrated the promising potential of LLMs, particularly through a structured, simulated-agents prompting strategy, to create high-quality, usable OSCE stations in the emerging field of digital health. The results highlighted the importance of carefully designed prompts and structured workflows to effectively use LLMs for educational purposes. Finally, our findings suggest that LLMs could be valuable assistants for educators, but expert oversight is still crucial to ensure content quality and relevance of the generated content. Future research should explore the practical application of AI-assisted OSCEs in various educational settings and their integration into medical curricula.

Acknowledgments

This work was supported by a French government grant managed by the Agence Nationale de la Recherche under the France 2030 program, reference DigiHealth Paris Cité ANR-23-CMAS-0034 (coordinator: RT).

Funding

No external financial support or grants were received from any public, commercial, or not-for-profit entities for the research, authorship, or publication of this article.

Data Availability

The prompts used to generate OSCEs are available in Multimedia Appendix 2. The OSCEs generated during this study are available from the corresponding author on reasonable request.

Conflicts of Interest

None declared.

Multimedia Appendix 1
METRICS checklist.
[DOCX File , 16 KB - mededu_v12i1e82116_app1.docx]

Multimedia Appendix 2
Prompts used to generate the OSCEs for each configuration.
[DOCX File , 21 KB - mededu_v12i1e82116_app2.docx]

References

XSL•FO
RenderX

1. Preiksaitis C, Rose C. Opportunities, Challenges, and Future Directions of Generative Artificial Intelligence in Medical Education: Scoping Review. JMIR Med Educ 2023 Oct 20;9:e48785 [FREE Full text] [doi: 10.2196/48785] [Medline: 37862079]
2. Aster A, Laupichler MC, Rockwell-Kollmann T, Masala G, Bala E, Raupach T. ChatGPT and other large language models in medical education - scoping literature review. Med Sci Educ 2025;35(1):555-567. [doi: 10.1007/s40670-024-02206-6] [Medline: 40144083]
3. Gordon M, Daniel M, Ajiboye A, Uraiby H, Xu N, Bartlett R, et al. A scoping review of artificial intelligence in medical education: BEME guide No. 84. Med Teach 2024;46(4):446-470 [FREE Full text] [doi: 10.1080/0142159X.2024.2314198] [Medline: 38423127]
4. Lucas HC, Upperman JS, Robinson JR. A systematic review of large language models and their implications in medical education. Med Educ 2024;58(11):1276-1285. [doi: 10.1111/medu.15402] [Medline: 38639098]
5. Wu Y, Zheng Y, Feng B, Yang Y, Kang K, Zhao A. Embracing ChatGPT for medical education: exploring its impact on doctors and medical students. JMIR Med Educ 2024;10:e52483 [FREE Full text] [doi: 10.2196/52483] [Medline: 38598263]
6. Zhui L, Yhap N, Liping L, Zhengjie W, Zhonghao X, Xiaoshu Y, et al. Impact of large language models on medical education and teaching adaptations. JMIR Med Inform 2024;12:e55933 [FREE Full text] [doi: 10.2196/55933] [Medline: 39087590]
7. Liu M, Okuhara T, Chang X, Shirabe R, Nishiie Y, Okada H, et al. Performance of ChatGPT across different versions in medical licensing examinations worldwide: systematic review and meta-analysis. J Med Internet Res 2024;26:e60807 [FREE Full text] [doi: 10.2196/60807] [Medline: 39052324]
8. Gilson A, Safranek CW, Huang T, Socrates V, Chi L, Taylor RA, et al. How Does ChatGPT perform on the United States medical licensing examination (USMLE)? The implications of large language models for medical education and knowledge assessment. JMIR Med Educ 2023;9:e45312 [FREE Full text] [doi: 10.2196/45312] [Medline: 36753318]
9. Kung TH, Cheatham M, Medenilla A, Sillos C, De Leon L, Elepaño C, et al. Performance of ChatGPT on USMLE: potential for AI-assisted medical education using large language models. PLOS Digit Health 2023;2(2):e0000198 [FREE Full text] [doi: 10.1371/journal.pdig.0000198] [Medline: 36812645]
10. Nakao T, Miki S, Nakamura Y, Kikuchi T, Nomura Y, Hanaoka S, et al. Capability of GPT-4V(ision) in the Japanese national medical licensing examination: evaluation study. JMIR Med Educ 2024;10:e54393 [FREE Full text] [doi: 10.2196/54393] [Medline: 38470459]
11. Zong H, Wu R, Cha J, Wang J, Wu E, Li J, et al. Large language models in worldwide medical exams: platform development and comprehensive analysis. J Med Internet Res 2024;26:e66114. [doi: 10.2196/66114] [Medline: 39729356]
12. Jaleel A, Aziz U, Farid G, Zahid Bashir M, Mirza TR, Khizar Abbas SM, et al. Evaluating the potential and accuracy of ChatGPT-3.5 and 4.0 in medical licensing and in-training examinations: systematic review and meta-analysis. JMIR Med Educ 2025;11:e68070 [FREE Full text] [doi: 10.2196/68070] [Medline: 40973108]
13. Artsi Y, Sorin V, Konen E, Glicksberg BS, Nadkarni G, Klang E. Large language models for generating medical examinations: systematic review. BMC Med Educ 2024;24(1):354 [FREE Full text] [doi: 10.1186/s12909-024-05239-y] [Medline: 38553693]
14. Kıyak YS, Emekli E. ChatGPT prompts for generating multiple-choice questions in medical education and evidence on their validity: a literature review. Postgrad Med J 2024;100(1189):858-865. [doi: 10.1093/postmj/qgae065] [Medline: 38840505]
15. Cheung BHH, Lau GKK, Wong GTC, Lee EYP, Kulkarni D, Seow CS, et al. ChatGPT versus human in generating medical graduate exam multiple choice questions-a multinational prospective study (Hong Kong S.A.R., Singapore, Ireland, and the United Kingdom). PLoS One 2023;18(8):e0290691 [FREE Full text] [doi: 10.1371/journal.pone.0290691] [Medline: 37643186]
16. Abouzeid E, Wassef R, Jawwad A, Harris P. Chatbots' role in generating single best answer questions for undergraduate medical student assessment: comparative analysis. JMIR Med Educ 2025;11:e69521 [FREE Full text] [doi: 10.2196/69521] [Medline: 40446330]
17. Holderried F, Stegemann-Philipps C, Herschbach L, Moldt J, Nevins A, Griewatz J, et al. A generative pretrained transformer (GPT)-powered chatbot as a simulated patient to practice history taking: prospective, mixed methods study. JMIR Med Educ 2024;10:e53961 [FREE Full text] [doi: 10.2196/53961] [Medline: 38227363]
18. Holderried F, Stegemann-Philipps C, Herrmann-Werner A, Festl-Wietek T, Holderried M, Eickhoff C, et al. A language model-powered simulated patient with automated feedback for history taking: prospective study. JMIR Med Educ 2024;10:e59213 [FREE Full text] [doi: 10.2196/59213] [Medline: 39150749]
19. Öncü S, Torun F, Ülkü HH. AI-powered standardised patients: evaluating ChatGPT-4o's impact on clinical case management in intern physicians. BMC Med Educ 2025;25(1):278 [FREE Full text] [doi: 10.1186/s12909-025-06877-6] [Medline: 39979969]
20. Yu H, Zhou J, Li L, Chen S, Gallifant J, Shi A, et al. AIPatient: simulating patients with EHRs and LLM powered agentic workflow. : arXiv [FREE Full text] Preprint posted online on 27 Nov 2025

XSL•FO
RenderX

21. Wang C, Li S, Lin N, Zhang X, Han Y, Wang X. Application of large language models in medical training evaluation-using ChatGPT as a standardized patient: multimetric assessment. J Med Internet Res 2025;27:e59435 [FREE Full text] [doi: 10.2196/59435] [Medline: 39742453]
22. Stamer T, Steinhäuser J, Flägel K. Artificial intelligence supporting the training of communication skills in the education of health care professions: scoping review. J Med Internet Res 2023;25:e43311 [FREE Full text] [doi: 10.2196/43311] [Medline: 37335593]
23. Li Y, Zeng C, Zhong J, Zhang R, Zhang M, Zou L. Leveraging large language model as simulated patients for clinical education. : arXiv [FREE Full text] Preprint posted online on 25 Apr 2024
24. Liu Y, Shi C, Wu L, Lin X, Chen X, Zhu Y, et al. Development and validation of a large language model-based system for medical history-taking training: prospective multicase study on evaluation stability, human-ai consistency, and transparency. JMIR Med Educ 2025;11:e73419 [FREE Full text] [doi: 10.2196/73419] [Medline: 40882613]
25. Cook DA, Overgaard J, Pankratz VS, Del Fiol G, Aakre CA. Virtual patients using large language models: scalable, contextualized simulation of clinician-patient dialogue with feedback. J Med Internet Res 2025;27:e68486. [doi: 10.2196/68486] [Medline: 39854611]
26. Cross J, Kayalackakom T, Robinson RE, Vaughans A, Sebastian R, Hood R, et al. Assessing ChatGPT's capability as a new age standardized patient: qualitative study. JMIR Med Educ 2025;11:e63353 [FREE Full text] [doi: 10.2196/63353] [Medline: 40393017]
27. Sridharan K, Sequeira RP. BMC Med Educ 2024 Apr 22;24(1):431 [FREE Full text] [doi: 10.1186/s12909-024-05365-7] [Medline: 38649959]
28. Takahashi H, Shikino K, Kondo T, Komori A, Yamada Y, Saita M, et al. Educational utility of clinical vignettes generated in Japanese by chatgpt-4: mixed methods study. JMIR Med Educ 2024;10:e59133 [FREE Full text] [doi: 10.2196/59133] [Medline: 39137031]
29. Yanagita Y, Yokokawa D, Uchida S, Li Y, Uehara T, Ikusaka M. Can AI-generated clinical vignettes in Japanese be used medically and linguistically? J Gen Intern Med 2024;39(16):3282-3289. [doi: 10.1007/s11606-024-09031-y] [Medline: 39313665]
30. Scherr R, Spina A, Dao A, Andalib S, Halaseh FF, Blair S, et al. Novel evaluation metric and quantified performance of ChatGPT-4 patient management simulations for early clinical education: experimental study. JMIR Form Res 2025;9:e66478 [FREE Full text] [doi: 10.2196/66478] [Medline: 40013991]
31. Vaughn J, Ford SH, Scott M, Jones C, Lewinski A. Enhancing healthcare education: leveraging chatgpt for innovative simulation scenarios. Clinical Simulation in Nursing 2024;87:101487. [doi: 10.1016/j.ecns.2023.101487]
32. Violato E, Corbett C, Rose B, Rauschning B, Witschen B. The effectiveness and efficiency of using ChatGPT for writing health care simulations. ijohs 2023. [doi: 10.54531/wjgb5594]
33. Harden RM, Gleeson FA. Assessment of clinical competence using an objective structured clinical examination (OSCE). Med Educ 1979;13(1):41-54. [Medline: 763183]
34. Patrício MF, Julião M, Fareleira F, Carneiro AV. Is the OSCE a feasible tool to assess competencies in undergraduate medical education? Med Teach 2013;35(6):503-514. [doi: 10.3109/0142159X.2013.774330] [Medline: 23521582]
35. Majumder MAA, Kumar A, Krishnamurthy K, Ojeh N, Adams OP, Sa B. An evaluative study of objective structured clinical examination (OSCE): students and examiners perspectives. Adv Med Educ Pract 2019;10:387-397 [FREE Full text] [doi: 10.2147/AMEP.S197275] [Medline: 31239801]
36. Yamamoto A, Koda M, Ogawa H, Miyoshi T, Maeda Y, Otsuka F, et al. Enhancing medical interview skills through ai-simulated patient interactions: nonrandomized controlled trial. JMIR Med Educ 2024;10:e58753 [FREE Full text] [doi: 10.2196/58753] [Medline: 39312284]
37. Pereira DSM, Falcão F, Nunes A, Santos N, Costa P, Pêgo JM. Designing and building OSCEBot ® for virtual OSCE - performance evaluation. Med Educ Online 2023;28(1):2228550 [FREE Full text] [doi: 10.1080/10872981.2023.2228550] [Medline: 37347808]
38. Hirosawa T, Yokose M, Sakamoto T, Harada Y, Tokumasu K, Mizuta K, et al. Utility of Generative Artificial Intelligence for Japanese Medical Interview Training: Randomized Crossover Pilot Study. JMIR Med Educ 2025 Aug 01;11:e77332 [FREE Full text] [doi: 10.2196/77332] [Medline: 40749190]
39. Luordo D, Torres Arrese M, Tristán Calvo C, Shani Shani KD, Rodríguez Cruz LM, García Sánchez FJ, et al. Application of Artificial Intelligence as an Aid for the Correction of the Objective Structured Clinical Examination (OSCE). Applied Sciences 2025 Jan 23;15(3):1153. [doi: 10.3390/app15031153]
40. Geathers J, Hicke Y, Chan C, Rajashekar N, Sewell J, Cornes S, et al. Benchmarking generative AI for scoring medical student interviews in objective structured clinical examinations (OSCEs). : ArXiv [FREE Full text] Preprint posted online on May 15, 2025
41. Jamieson AR, Holcomb MJ, Dalton TO, Campbell KK, Vedovato S, Shakur AH, et al. Rubrics to Prompts: Assessing Medical Student Post-Encounter Notes with AI. NEJM AI 2024 Nov 27;1(12). [doi: 10.1056/aics2400631]
42. Shakur AH, Holcomb MJ, Hein D, Kang S, Dalton TO, Campbell KK, et al. Large language models for medical OSCE assessment: a novel approach to transcript analysis. : ArXiv [FREE Full text] Preprint posted online on 11 Oct 2024

XSL•FO
RenderX

43. Tekin M, Yurdal MO, Toraman Ç, Korkmaz G, Uysal İ. Is AI the future of evaluation in medical education?? AI vs. human evaluation in objective structured clinical examination. BMC Med Educ 2025;25(1):641 [FREE Full text] [doi: 10.1186/s12909-025-07241-4] [Medline: 40312328]
44. Huang T, Hsieh PH, Chang Y. Performance comparison of junior residents and ChatGPT in the objective structured clinical examination (OSCE) for medical history taking and documentation of medical records: development and usability study. JMIR Med Educ 2024;10:e59902 [FREE Full text] [doi: 10.2196/59902] [Medline: 39622713]
45. Li SW, Kemp MW, Logan SJS, Dimri PS, Singh N, Mattar CNZ, et al. ChatGPT outscored human candidates in a virtual objective structured clinical examination in obstetrics and gynecology. Am J Obstet Gynecol 2023;229(2):172.e1-172.e12 [FREE Full text] [doi: 10.1016/j.ajog.2023.04.020] [Medline: 37088277]
46. Aqavil-Jahromi S, Eftekhari M, Akbari H, Aligholi-Zahraie M. Evaluation of correctness and reliability of GPT, bard, and bing chatbots' responses in basic life support scenarios. Sci Rep 2025;15(1):11429 [FREE Full text] [doi: 10.1038/s41598-024-82948-w] [Medline: 40180985]
47. Tsang R. Practical applications of ChatGPT in undergraduate medical education. J Med Educ Curric Dev 2023;10:23821205231178449 [FREE Full text] [doi: 10.1177/23821205231178449] [Medline: 37255525]
48. Maaz S, Palaganas JC, Palaganas G, Bajwa M. A guide to prompt design: foundations and applications for healthcare simulationists. Front Med (Lausanne) 2024;11:1504532 [FREE Full text] [doi: 10.3389/fmed.2024.1504532] [Medline: 39980724]
49. Misra SM, Suresh S. Artificial intelligence and objective structured clinical examinations: using ChatGPT to revolutionize clinical skills assessment in medical education. J Med Educ Curric Dev 2024;11:23821205241263475 [FREE Full text] [doi: 10.1177/23821205241263475] [Medline: 39070287]
50. Bearman M, Ajjawi R, Bennett S, Boud D. The hidden labours of designing the objective structured clinical examination: a practice theory study. Adv Health Sci Educ Theory Pract 2021;26(2):637-651. [doi: 10.1007/s10459-020-10015-w] [Medline: 33196956]
51. Khan KZ, Ramachandran S, Gaunt K, Pushkar P. The objective structured clinical examination (OSCE): AMEE Guide No. 81. Part I: an historical and theoretical perspective. Med Teach 2013;35(9):e1437-e1446. [doi: 10.3109/0142159X.2013.818634] [Medline: 23968323]
52. Armijo-Rivera S, Fuenzalida-Muñoz B, Vicencio-Clarke S, Elbers-Arce A, Bozzo-Navarrete S, Kunakov N, et al. Advancing the assessment of clinical competence in Latin America: a scoping review of OSCE implementation and challenges in resource-limited settings. BMC Med Educ 2025;25(1):587 [FREE Full text] [doi: 10.1186/s12909-025-07151-5] [Medline: 40259375]
53. Ataro G, Worku S, Asaminew T. Experience and challenges of objective structured clinical examination (OSCE): perspective of students and examiners in a clinical department of ethiopian university. Ethiop J Health Sci 2020;30(3):417-426 [FREE Full text] [doi: 10.4314/ejhs.v30i3.13] [Medline: 32874085]
54. Sallam M, Barakat M, Sallam M. A preliminary checklist (METRICS) to standardize the design and reporting of studies on generative artificial intelligence-based models in health care education and practice: development study involving a literature review. Interact J Med Res 2024;13:e54704 [FREE Full text] [doi: 10.2196/54704] [Medline: 38276872]
55. ChatGPT. OpenAI. 2025. URL: https://chatgpt.com [accessed 2025-06-29]
56. ChatGPT: explore GPTs. OpenAI. 2025. URL: https://chatgpt.com/gpts [accessed 2025-06-29]
57. Meskó B. Prompt engineering as an important emerging skill for medical professionals: tutorial. J Med Internet Res 2023;25:e50638 [FREE Full text] [doi: 10.2196/50638] [Medline: 37792434]
58. Başaranoğlu M, Akbay E, Erdem E. AI-generated questions for urological competency assessment: a prospective educational study. BMC Med Educ 2025;25(1):611 [FREE Full text] [doi: 10.1186/s12909-025-07202-x] [Medline: 40281621]
59. Coşkun Ö, Kıyak YS, Budakoğlu I. ChatGPT to generate clinical vignettes for teaching and multiple-choice questions for assessment: a randomized controlled experiment. Med Teach 2025;47(2):268-274. [doi: 10.1080/0142159X.2024.2327477] [Medline: 38478902]
60. Ahmed A, Kerr E, O'Malley A. Quality assurance and validity of AI-generated single best answer questions. BMC Med Educ 2025;25(1):300 [FREE Full text] [doi: 10.1186/s12909-025-06881-w] [Medline: 40001164]
61. Arain SA, Akhund SA, Barakzai MA, Meo SA. Transforming medical education: leveraging large language models to enhance PBL-a proof-of-concept study. Adv Physiol Educ 2025;49(2):398-404 [FREE Full text] [doi: 10.1152/advan.00209.2024] [Medline: 39918742]
62. Khan KZ, Gaunt K, Ramachandran S, Pushkar P. Med Teach 2013 Sep;35(9):e1447-e1463. [doi: 10.3109/0142159X.2013.818635] [Medline: 23968324]
63. Titre II : Recherches impliquant la personne humaine (Articles L1121-1 à L1128-12) [Webpage in French]. Légifrance. 2025. URL: https://www.legifrance.gouv.fr/codes/section_lc/LEGITEXT000006072665/LEGISCTA000006154978/#LEGISCTA000025457387 [accessed 2025-08-06]
64. Upadhayay B, Behzadan V, Karbasi A. Cognitive overload attack:prompt injection for long context. : ArXiv [FREE Full text] Preprint posted online on 15 Oct 2024
65. Knowledge in GPTs. OpenAI Help Center. URL: https://help.openai.com/en/articles/8843948-knowledge-in-gpts [accessed 2025-06-29]

XSL•FO
RenderX

66. Retrieval augmented generation (RAG) and semantic search for GPTs. OpenAI Help Center. 2025. URL: https://help.openai.com/en/articles/8868588-retrieval-augmented-generation-rag-and-semantic-search-for-gpts [accessed 2025-06-29]
67. Key guidelines for writing instructions for custom GPTs. OpenAI Help Center. URL: https://help.openai.com/en/articles/9358033-key-guidelines-for-writing-instructions-for-custom-gpts [accessed 2025-06-29]
68. Clusmann J, Kolbinger FR, Muti HS, Carrero ZI, Eckardt J, Laleh NG, et al. The future landscape of large language models in medicine. Commun Med (Lond) 2023;3(1):141 [FREE Full text] [doi: 10.1038/s43856-023-00370-1] [Medline: 37816837]
69. Wang Z, Chu Z, Doan TV, Ni S, Yang M, Zhang W. History, development, and principles of large language models: an introductory survey. AI Ethics 2024;5(3):1955-1971. [doi: 10.1007/s43681-024-00583-7]
70. Agarwal M, Sharma P, Goswami A. Analysing the applicability of ChatGPT, bard, and bing to generate reasoning-based multiple-choice questions in medical physiology. Cureus 2023;15(6):e40977 [FREE Full text] [doi: 10.7759/cureus.40977] [Medline: 37519497]
71. Mistry NP, Saeed H, Rafique S, Le T, Obaid H, Adams SJ. Large language models as tools to generate radiology board-style multiple-choice questions. Acad Radiol 2024;31(9):3872-3878 [FREE Full text] [doi: 10.1016/j.acra.2024.06.046] [Medline: 39013736]
72. Vaid A, Lampert J, Lee J, Sawant A, Apakama D, Sakhuja A, et al. Natural language programming in medicine: administering evidence based clinical workflows with autonomous agents powered by generative large language models. : ArXiv [FREE Full text]
73. Ke YH, Jin L, Elangovan K, Abdullah HR, Liu N, Sia ATH, et al. Retrieval augmented generation for 10 large language models and its generalizability in assessing medical fitness. NPJ Digit Med 2025;8(1):187 [FREE Full text] [doi: 10.1038/s41746-025-01519-z] [Medline: 40185842]
74. DigiHealth Paris Cité : la révolution de l'enseignement de la santé numérique [webpage in French]. Faculté de Santé, Université Paris Cité. 2025. URL: https://u-paris.fr/sante/digihealth/ [accessed 2025-10-13]

Abbreviations

AI: artificial intelligence
cGPT: custom GPT-4o
LLM: large language model
METRICS: Model, Evaluation, Timing, Range/Randomization, Individual factors, Count, Specificity
OSCE: Objective Structured Clinical Examination
RAG: retrieval-augmented generation
SP: standardized participant

Edited by R Pellegrino; submitted 10.Aug.2025; peer-reviewed by E Emekli, A Hidki, J Wang; comments to author 09.Oct.2025; revised version received 15.Oct.2025; accepted 13.Dec.2025; published 15.Jan.2026.

Please cite as:
Zouakia Z, Logak E, Szymczak A, Jais JP, Burgun A, Tsopra R
AI-Driven Objective Structured Clinical Examination Generation in Digital Health Education: Comparative Analysis of Three GPT-4o Configurations
JMIR Med Educ 2026;12:e82116
URL: https://mededu.jmir.org/2026/1/e82116
doi:10.2196/82116
PMID:

©Zineb Zouakia, Emmanuel Logak, Alan Szymczak, Jean-Philippe Jais, Anita Burgun, Rosy Tsopra. Originally published in JMIR Medical Education (https://mededu.jmir.org), 15.Jan.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Education, is properly cited. The complete bibliographic information, a link to the original publication on https://mededu.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Using AI to Train Future Clinicians in Depression Assessment: Feasibility Study

Friederike Holderried[1], MME, MD; Alessandra Sonanini[1], MD; Annika Philipps[2], MD; Christian Stegemann-Philipps[1], Dr rer nat; Lea Herschbach[1], MSc; Teresa Festl-Wietek[1], Dr rer nat; Stephan Zipfel[2,3], Prof Dr Med; Rebecca Erschens[2], Dr rer nat; Anne Herrmann-Werner[1,3], MME, Prof Dr Med

[1]TIME - Tübingen Institute for Medical Education, University of Tübingen, Tübingen, Baden-Wurttemberg, Germany

[2]Department of Psychosomatic Medicine and Psychotherapy, University Hospital Tübingen, University of Tübingen, Tübingen, Germany

[3]German Center for Mental Health (DZPG), Partner Site Tübingen, Tübingen, Germany

Corresponding Author:
Lea Herschbach, MSc
TIME - Tübingen Institute for Medical Education
University of Tübingen
Elfriede-Aulhorn-Strasse 10
Tübingen, 72076
Germany
Phone: 49 70712973715
Email: lea.herschbach@med.uni-tuebingen.de

Abstract

Background: Depression is a major global health care challenge, causing significant individual distress but also contributing to a substantial global burden. Timely and accurate diagnosis is crucial. To help future clinicians develop these essential skills, we trained a generative pretrained transformer (GPT)–powered chatbot to simulate patients with varying degrees of depression and suicidality.

Objective: This study aims to evaluate the applicability and transferability of our GPT-4-powered chatbot for psychosomatic cases. Specifically, we aim to investigate how accurately the chatbot can simulate patients exhibiting various stages of depression and phases of suicidal ideation, while adhering to a predefined role script and maintaining a sufficient level of authenticity. Additionally, we want to analyze to what level the chatbot is suitable for practicing correctly diagnosing depressive disorders in patients, as well as assessing suicidality stages.

Methods: We developed 3 virtual patient role scripts depicting complex, realistic cases of depression and varying degrees of suicidality collaboratively with field experts and aligned with mental health assessment guidelines. These cases were integrated into a GPT-4–powered chatbot for practicing clinical history-taking. A total of 148 medical students, with an average age of 22.71 years and mostly in their sixth semester, interacted individually with one of the randomly assigned virtual patients through chat. Following this, they completed a questionnaire assessing their demographics and user experience. Chats were analyzed descriptively to assess diagnostic accuracy and suicidality assessments, as well as the role script adherence and authenticity of the artificial intelligence (AI). This was done to gain further insight into the chatbot's behavior and the students' diagnostic accuracy.

Results: In over 90% (725/778) of the answers, the chatbot maintained its assigned role. On average, students correctly identified the severity of depression in 60% (81/135) and the phase of suicidality in 67% (91/135) of the cases. Notably, the majority either failed to address or insufficiently explored the topic of suicidality despite explicit instructions beforehand.

Conclusions: This study demonstrates that a GPT-powered chatbot can simulate patients with depression fairly accurately. More than two-thirds of participants perceived the AI-simulated patients with depression as authentic, and nearly 80% (106/135) indicated they would like to use the application for further practice, highlighting its potential as a training tool. While a small proportion of students expressed reservations, and the overall diagnostic accuracy varied depending on the severity of the case, the findings overall support the feasibility and educational value of AI-based role-playing in clinical training. AI-supported virtual patients provide a highly flexible, standardized, and readily available training tool, independent of real-life constraints.

(JMIR Med Educ 2026;12:e87102) doi:10.2196/87102

XSL•FO
RenderX

KEYWORDS

depression diagnosis; generative pretrained transformer; GPT-powered chatbot; large language model; LLM; medical students; suicidality

Introduction

Depression affects an estimated 280 million people globally, representing one of the world's most prevalent mental health conditions [1,2]. This disorder, along with anxiety, contributes to a US $1 trillion annual economic burden [3]. Despite the availability of structured diagnostic criteria through the *ICD-10* (*International Statistical Classification of Diseases, Tenth Revision*) and the *DSM-5* (*Diagnostic and Statistical Manual of Mental Disorders* [Fifth Edition]) [4,5], accurately identifying depression remains challenging. Symptom interpretation is often complex, as presentations can vary significantly between individuals [6]. Additionally, factors such as stigma and limited access to care contribute to the high rate of undiagnosed cases [7,8]. Effective diagnosis, therefore, requires not only familiarity with the formal criteria but also clinical expertise to contextualize symptoms within the broader patient history and presentation [6]. Worldwide, organizations have respective guidelines on diagnosing and treating depression [9-11]. Thorough patient assessment is essential for classification and therapy of depression, but also suicidality—as currently over 700,000 suicide deaths occur annually [12]. However, studies show that diagnostic accuracy is inconsistent among physicians [13-16]. Respective training is crucial to close this gap and has proven to be successful [17-22].

Artificial intelligence (AI)–simulated patients allow learners to practice clinical interviews flexibly, repeatedly, and at scale, providing realistic interactions to train diagnostic accuracy [23-25]. Our research group has already demonstrated the effectiveness of a generative pretrained transformer (GPT)–powered chatbot used as an AI-driven virtual patient for clinical history-taking exercises of somatic diseases, with the capability to provide automated feedback [26,27]. Despite these advances, empirical research on AI-driven virtual patients in the field of mental health training is still limited [28]. In particular, little is known about the stability of large language model (LLM)–generated symptom presentations, their ability to represent nuanced depressive or suicidal symptoms, or their impact on students' diagnostic reasoning. Therefore, we developed a training module for medical students within their regular teaching in psychosomatic medicine and psychotherapy (PSM), using AI-driven virtual patients with depressive symptoms and varying degrees of suicidality. Virtual patients are well-established in medical education and have been shown to help students apply communication skills in clinical practice [29-31]. Recent advances in natural language processing have enhanced interactions with virtual patients to increase user engagement [32]. LLMs are particularly useful in mental health training and have been shown to accurately answer questions related to depression, even outperforming general practitioners in direct comparisons of treatment recommendations [17,33]. They have also been used as interviewee models for depression screening [34].

In this study, we aim to evaluate the applicability and transferability of our GPT-4–powered chatbot for psychosomatic cases, specifically addressing the following research questions: (1) How accurately can a GPT-4–powered chatbot simulate patients exhibiting various stages of depression and phases of suicidal ideation, while adhering to a predefined role script and maintaining a sufficient level of authenticity? What specific challenges and limitations can be observed? (2) To what level is the GPT-4–powered chatbot suitable for practicing correctly diagnosing depressive disorders in patients, as well as assessing suicidality stages?

Methods

Creation of AI-Simulated Patients

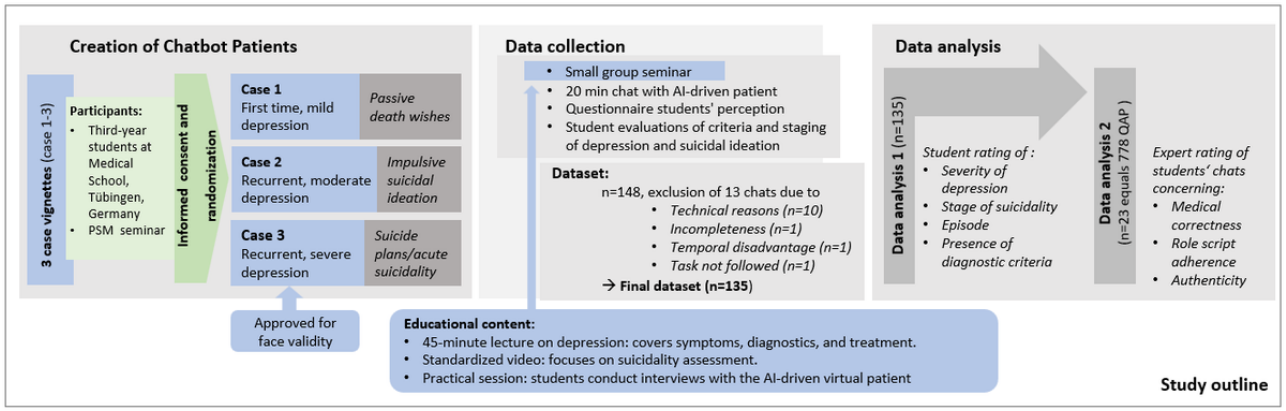
For this study, we created 3 different virtual patient role scripts, each describing a complex, realistic patient history and current symptomatology of depression or suicidality stage (Figure 1 provides further details). The case vignettes and role scripts were collaboratively developed and critically revised by field experts and aligned with guidelines for mental health assessment in virtual patients [35]. All case vignettes were created exclusively by clinical experts and were not generated or assisted by any AI system. These patient cases were integrated into our GPT-4–powered chatbot, which was prompted for the purpose of practicing clinical history taking [26] and adapted to the new challenges within the mental health context. ChatGPT (OpenAI) was selected due to its wide availability, established performance in health-related natural language processing tasks, and strong conversational capabilities. At the time of data collection, GPT-4 was one of the leading models and the most advanced OpenAI Model. We had prior experience with GPT-4 in complex role-playing settings and were thus able to ensure a high level of quality of responses. In addition, the OpenAI account we used to access the application programming interface had a sufficiently high token limit to allow usage in groups of 10-20 people simultaneously. Temperature for GPT-4 was set to 0.1 to minimize risk of diversion from the provided information. While slightly repetitive, the answers were usually regarded as natural enough, even with this low setting.

Patient cases, student interactions, and questionnaires were conducted in German, and all examples cited in this document were translated into English via DeepL [36]. Textbox S3 and Figure S3 in Multimedia Appendix 1 [9,37] provide an exemplary chat history and screenshots of the application.

To categorize the levels of depression, we used the *ICD-10* diagnostic criteria from the World Health Organization's classification system [5,38], which are also integrated into the curriculum at German universities (Table S1 in Multimedia Appendix 1). To classify suicidality, 4 stages have been defined, ranging from passive death wish (stage 1), suicidal ideation (stage 2), and suicide plans/preparations (stage 3) to suicide actions (stage 4) [37] (Table S2 in Multimedia Appendix 1).

XSL•FO
RenderX

Figure 1. Study outline. AI: artificial intelligence; PSM: psychosomatic medicine and psychotherapy; QAP: question-answer pair.



Experts

Case vignettes were developed by 2 experts in the field of PSM. Subsequently, 2 other experts evaluated the student chats, and 3 further experts with longstanding clinical experience in PSM and mental health reevaluated the vignettes diagnostically to check concordance. Finally, question-answer pairs (QAPs) were qualitatively coded for recurring errors by different experts in the field of PSM (Table S4 in Multimedia Appendix 1).

Data Collection

Figure 1 provides detailed information about the setting, participants, and study design. The data collection took place between April and July 2024; the AI component of the session lasted 1 hour. Students, equipped with laptops, were introduced to a preconfigured web interface. Students were randomly assigned to interact with virtual patients exhibiting varying levels of depression severity (mild, moderate, and severe). Provided with general case details (setting, gender, and age of the virtual patient), they engaged in a 20-minute individual interaction with the AI-driven patient suffering from depression; the medical conversation was simulated through chat. After the digital interaction, students supplied demographic information and completed a questionnaire. The first section of the questionnaire focused on students' subjective evaluation of the application, assessing the authenticity of the virtual patient and the likelihood of using the application regularly.

The second part of the questionnaire evaluated the students' diagnostic accuracy. This included identifying the relevant *ICD-10* diagnostic criteria for depression, assessing the severity and recurrence of the episode, and determining the degree of suicidality.

The session concluded with an interactive discussion with a teacher from the department of PSM present, which involved reviewing correct answers to the questions and exploring additional insights into the diagnosis and treatment of depression aside from the encountered AI case.

Prompt Engineering

Generally, we built on the techniques described in preceding publications [26,27]. Main aspects of the prompting are general descriptions of the patients' situation and character, as well as a customizable, case-dependent list of medical categories. For each category, information is provided in the form of clinical information or as example answers. For this study, we included depression criteria and phases of suicidal ideation as categories in order to accurately describe the patients' situation. We also expanded the general prompting to account for the required emotional depth of the simulated interviews. To increase authenticity, we adapted the prompts and the role scripts as described next.

We refined the virtual patient's scripts by adding explanatory notes for ambiguous terms and structuring responses progressively, revealing details only upon user inquiry to enhance authenticity. To improve realism in sensitive discussions, we instructed the model to simulate patient reluctance and provided example dialogues. Additionally, we implemented safeguards to prevent prompt extraction attacks in example inputs that make the model repeat the hidden system prompt in the running chat, exposing it to the user. In our case, this would have included the complete character definition. These safeguards were thus also meant to ensure the model maintains its role without disclosing unintended information (Textboxes S1 and S2 in Multimedia Appendix 1 provide prompt engineering techniques and full prompting).

Data Analysis

An overview of the data analysis process is presented in Figure 1. After completing data collection, all data were exported. Due to various reasons, 13 chats were excluded, leaving a total of 135 chats for inclusion in phase 1 of the data analysis.

Demographic Data

The dataset was analyzed descriptively in terms of demography and the students' assessment of the severity of depression, the episode, the presence of the individual diagnostic criteria, and the stage of suicidality. Furthermore, questions concerning the authenticity of simulated patients and the likelihood of using the application in the future were evaluated descriptively. Frequencies and distributions were calculated for all variables mentioned above to provide an overview of students' assessments and demographic characteristics.

Students´ Evaluation of Diagnosis and Suicidality

In data analysis phase 2, a total of 23 chats were selected out of the 135 chats for a more detailed exploration at the QAP level. To maximize issue detection, we selected chats in which students incorrectly assessed depression severity and suicidal

XSL•FO
RenderX

ideation based on the classification established during the development of the role scripts. From this group, a random sample was drawn, stratified by the observed error rates across severity levels: 11 chats for moderate depression (highest error rate), 7 for mild (lower error rate), and 5 for severe depression (lowest error rate). In the 23 exemplary chats, 778 QAPs were recorded, with a median of 29 per conversation. To assess interrater reliability, 2 experts independently evaluated 23 randomly selected chats in terms of the extent to which the criteria had been addressed, not addressed at all, partially addressed, or fully addressed. The criteria evaluated were authenticity, episode, severity, suicidality, as well as the 10 diagnostic criteria of depression. Prior work shows that users interpret authenticity as a cue for sincerity, interpersonal alignment, and emotional credibility in chatbot interactions [39], which aligns closely with the evaluative focus of this study. More specifically, factors that influenced the authenticity of the interaction in this study included instances in which the response did not correspond with either the question or the prompt, or in which the interaction was considered to be inauthentic for other reasons, such as repetitions or the use of unconventional phrasing. Since the sources of error already occurred repeatedly in the selected chats, it can be assumed that the findings can also be transferred to the entire dataset. The QAPs served as the units of analysis. Both experts examined them independently in a Microsoft Excel 2408 sheet (Microsoft Corporation). They achieved Cohen κ=0.794 (QAP level) and κ=0.846 (chat level). Due to the high level of agreement, 1 rating was selected randomly. Statistical analysis was performed using SPSS (version 28; IBM Corp).

First, all QAPs were analyzed to assess the extent to which students considered the diagnostic criteria for depression and suicidality. Subsequently, the 23 selected chats were analyzed to identify reasons for incorrect statements, including insufficient discussion of the criteria, misinterpretation of the AI answers by the students, or AI-driven conversations that led to incorrect diagnoses.

Expert Validation of Case Vignettes

As we identified cases in which misleading prompts led to incorrect assessments and were based on content ambiguities in the prompt, the prompts were retrospectively reviewed independently by 3 additional experts in the field of psychosomatic medicine with regard to the diagnostic criteria for all 3 case vignettes. The experts received the same chat information as the participants to assess whether they could correctly identify the severity of depression and suicidality; the prompts were not modified or retested as part of this study.

Error Analysis and Prompt Associations

To examine patterns in students' diagnostic thinking and deviations in the role of AI, we applied an inductive, theme-oriented coding approach based on the principles of data-driven analysis developed by Braun and Clarke [40]. Two expert raters independently used qualitative coding to identify recurring patterns in the QAPs, including errors in students' diagnostic reasoning and role deviations, nonresponsive answers, and inappropriate repetitions. After discussing discrepancies to reach consensus, findings were summarized descriptively to highlight common issues and inform recommendations for improving AI-driven training interactions.

Statistical Analysis

In phase 1, descriptive statistics (frequencies and percentages) were calculated for all variables to provide an overview of students' assessments and demographic characteristics. In phase 2, all QAPs were aggregated and stored in a Microsoft Excel 2408 sheet. Two raters separately assessed authenticity and diagnostic criteria of each QAP. Cohen κ was calculated to determine the interrater reliability between the 2 raters.

Statistical analysis was performed using SPSS (version 28). Mean values, associated SDs, frequencies, and percentages were calculated. Figure generation was performed using Excel 2408.

Ethical Considerations

The Ethics Committee of the Faculty of Medicine at University Hospital Tübingen approved the study (209/2024BO2). Participation was voluntary, with informed consent obtained; no compensation was offered for participation. All data were anonymized. All procedures were conducted in accordance with the Declaration of Helsinki and ethical guidelines for human participants. Informed consent for participation in this study was obtained verbally from all participants. Verbal consent was chosen over written consent because the study was conducted within an educational setting as part of the participants' course activities. Additionally, since no sensitive or personally identifiable information was collected, verbal consent was deemed appropriate and approved by the Ethics Committee. Furthermore, participants also confirmed their agreement to data collection and processing by ticking the respective box on the computer. Participants were fully informed about the study's purpose, procedures, and their right to withdraw at any time and without negative consequences before providing their consent.

Results

Demographic Data

A total of 148 chat datasets were collected, with 13 excluded (Figure 1), leaving 135 for analysis. Participants included 89 females, 45 males, and 1 diverse participant, with a mean age of 22.71 (SD 2.44; range 19-30) years. The majority (n=130) of the participants were in their sixth semester, 1 each in their fourth and seventh semesters, and 3 in their fifth semester. According to the curriculum, while most students have had some prior exposure to the topics of depression and suicidality, they have not yet studied them in a structured seminar format. Cohort sizes vary by topic and are clearly indicated to ensure accurate interpretation.

Authenticity

Based on the dataset shown above (n=135), Table 1 shows that more than two-thirds of the students rated the AI-controlled presentation of the patient with depression as convincing, very convincing, or humanlike. A total of 78.52% (106/135) of respondents said they would use this application again (sometimes, often, or whenever possible).

XSL•FO
RenderX

To assess the perceived authenticity, AI adherence to the role script, and students' chats in relation to the suspected diagnoses, 2 experts analyzed 23 selected chats in detail (Data Analysis section). Factors influencing authenticity, including frequent repetitions, inappropriate use of medical terminology, and unclear references to the prompt (65/778, 8.35%), were assessed. Based on these criteria, 87.5% (678/778) of QAPs were classified as authentic, while 12.85% (100/778) of QAPs were classified as inauthentic.

Table 1. Student´s opinion on the application.

Questions and responses	Frequency, n/N (%)
How convincing was the AI[a] in the patient role?	
Not convincing	2/135 (1.48)
Acceptable	40/135 (29.63)
Convincing	54/135 (40)
Very convincing	36/135 (26.67)
Human like	3/135 (2.22)
How likely would it be for you to use the application if it was available on a regular basis?	
Not at all	7/135 (5.19)
Rarely	22/135 (16.30)
Sometimes	70/135 (51.85)
Often	29/135 (21.48)
Whenever possible	7/135 (5.19)

[a]AI: artificial intelligence.

Role Script Adherence

For content categorization, QAPs were classified as "relevant" (diagnostic criteria or suicidality) or "not relevant" (other medical topics or general conversation elements). A total of 49.36% (384/778) QAPs were considered relevant for diagnosis. For role conformity, expert raters assessed whether GPT-4's responses (1) matched to the question asked and (2) remained consistent with the predefined role script. As shown in Table 2, GPT-4 deviated from the given role in 6.81% (53/778) of the QAPs.

Of the initial 6.81% (53/778) deviations, a total of 1.54% (12/778) stemmed from "not relevant" topics outside depression or suicidality (eg, other medical conditions) and were excluded from analysis. The remaining 5.27% (41/778) of "relevant' QAPs were distributed across the 5 key diagnostic criteria: "loss of interest," "loss of drive," "pessimistic future prospects," "changes in appetite," and "impaired concentration" (Table 2).

Table 2. Role adherence of the chatbot.

GPT-4 roleplay outcome and QAP[a] type	Frequency, n/N (%)
GPT[b] stays in role	725/778 (93.19)
GPT falls out of role, QAP concerning diagnostic criteria	41/778 (5.27)
Loss of interest	14/778 (1.80)
Loss of drive	5/778 (0.64)
Pessimistic future prospects	7/778 (0.90)
Changes in appetite	8/778 (1.03)
Impaired concentration	6/778 (0.77)
Other	1/778 (0.13)
GPT falls out of role, QAP concerning other medical information	12/778 (1.54)

[a]QAP: question-answer pair.

[b]GPT: generative pretrained transformer.

Qualitative Analysis

Identified erroneous criteria were qualitatively coded by the 2 expert raters to detect recurring patterns. The qualitative analysis of role deviations revealed 3 key issues (Table S4 in Multimedia Appendix 1 provides detailed examples). First, GPT-4 struggled with temporal consistency; time jumps (patient memories) or inconsistent timelines in patient history ("she has gradually

XSL•FO
RenderX

started engaging in her previous leisure activities") led to role-deviant behavior. This highlights the need to clearly separate current symptoms from previous history and to provide explicit timelines in the model. Second, inconsistency within the role scripts showed that overemphasis on isolated case details could lead GPT-4 to disproportionately amplify these aspects, resulting in misleading responses ("[My Concentration] has gotten a bit better since I started taking the new medication (…) But it's still not as good as it used to be"). Third, we observed a tendency toward symptom aggravation, especially in mild cases, likely influenced by common training data patterns and negatively phrased diagnostic criteria. In 97% (33/34) of the analyzed misleading answers, there was a negative shift, resulting in a depiction of greater depression than intended in the prompt (Table S4 in Multimedia Appendix 1 summarizes the findings). These findings underscore the importance of neutral, precisely structured prompts to ensure stable and realistic virtual patient behavior.

Students' Diagnostic Accuracy

After the patient interviews, students provided a suspected diagnosis, including episode type, depression severity, and suicidality stage. Comparing their assessments (n=135) with the predefined case diagnoses showed a 94.07% (127/135) agreement on episode classification, with minor case-related deviations. Agreement on depression severity, particularly relevant for therapy decisions, averaged 60.00% (81/135). The distribution across the subgroups is particularly noteworthy: the case with severe depression reached an agreement of 100%, while the other cases reached only 40.82% (20/49; moderate depression) and 47.92% (23/48; mild depression) agreement. Regarding the phase of suicidality, agreement across all subgroups was 67.41% (91/135) on average, with a tendency toward better assessment of passive death wishes. To examine discrepancies in severity assessment, we reviewed the diagnostic criteria identified by students. Figure S1 in Multimedia Appendix 1 shows correctly recognized criteria (blue) and those incorrectly selected despite not being part of the role script (gray). For mild depression, students struggled to recognize "feelings of guilt" as present, while "reduced self-esteem" and "loss of drive" were mistakenly classified as present. In contrast, 100% correctly identified "sleep disorders" as absent. For severe depression, all main and secondary criteria were present, preventing false positives. The main challenge was identifying "impaired concentration" (28/38, 73.68% correct), while all other criteria were correctly assessed by at least 86.84% (33/38). Figure S1 in Multimedia Appendix 1 visualizes these findings.

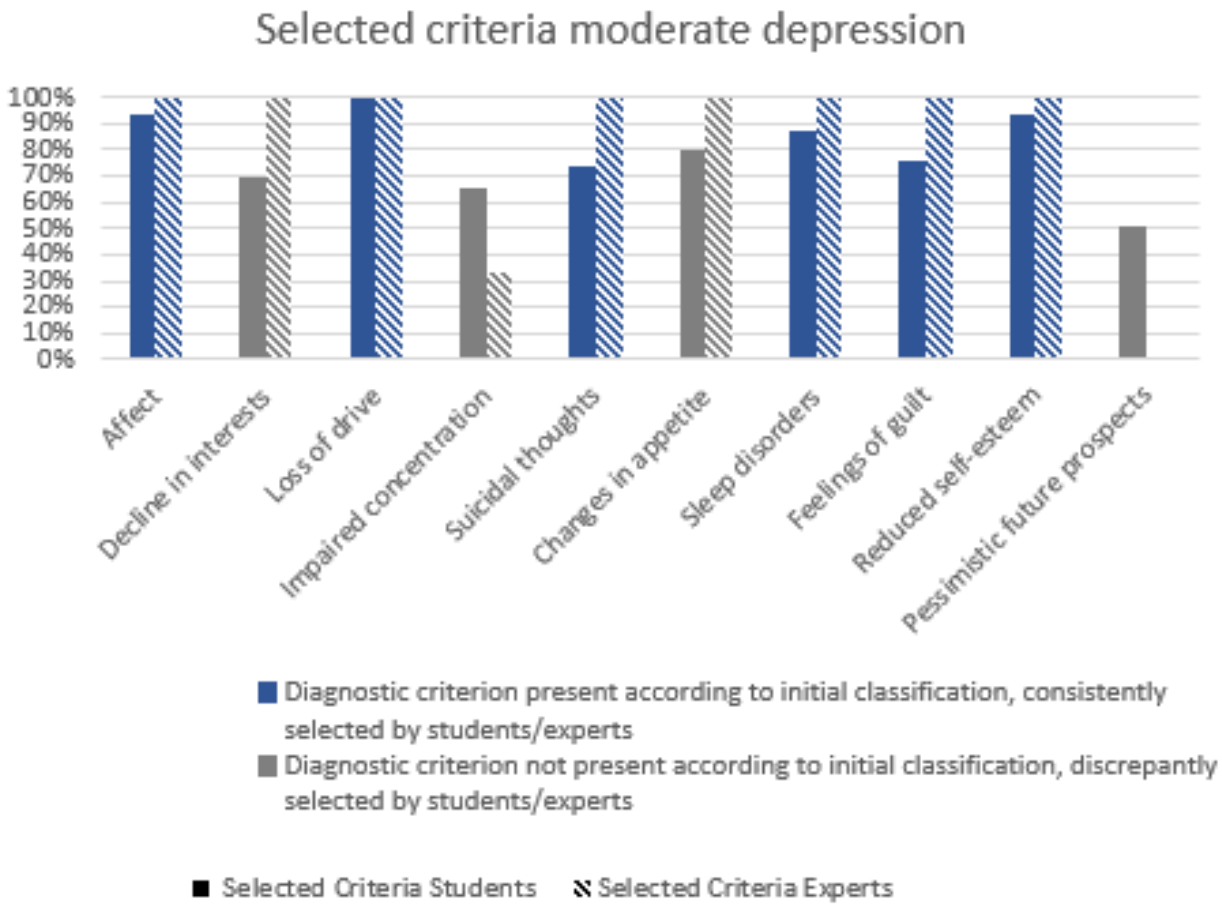
Expert Validation of Case Vignettes

Figure 2 shows the diagnostic criteria for the case developed as moderate depression. Notably, students frequently assessed "loss of interest," "concentration disorders," "changes in appetite," and "pessimistic future prospects" as present, although these were not defined as present by 2 mental health field experts who had created the case vignettes. To verify this discrepancy, all 3 vignettes were reevaluated by 3 additional field experts. These experts received the same prompt as GPT-4 and assessed the virtual patient according to the diagnostic criteria. For the cases of mild depression and severe depression, there were no relevant deviations between the initial developing expert group and the reassessing expert group. For moderate depression, significant deviations occurred in individual diagnostic criteria (Figure 2). Student selections (monochrome) are contrasted with expert assessments (striped). Blue indicates agreement with the initial role script classification, while gray indicates criteria not originally included, highlighting discrepancies. Two criteria—"decline in interests" (core symptom) and "changes in appetite" (additional symptom)—were consistently rated as present by all 3 additional experts. In contrast, "impaired concentration" and "pessimistic future prospects" were selected as present by more than half of the students, while the majority of experts did not consider them to be present.

If the additional experts' classification is used as the reference, the case originally labeled as moderate depression would instead be classified as severe. Based on this revised classification, 59% (29/49) of students correctly assessed the severity level.

XSL•FO
RenderX

Figure 2. Students' and experts' assessment of a moderate depression case vignette. Percentages of students (monochrome bars) and experts (striped bars) who identified the displayed International Classification of Diseases, Tenth Revision (ICD-10) diagnostic criteria for depression as present in the case vignette are shown. Blue bars indicate agreement with the initial role-script classification, whereas gray bars represent criteria not originally included, highlighting discrepancies between the vignette and the standard diagnostic criteria.



Error Analysis and Prompt Associations

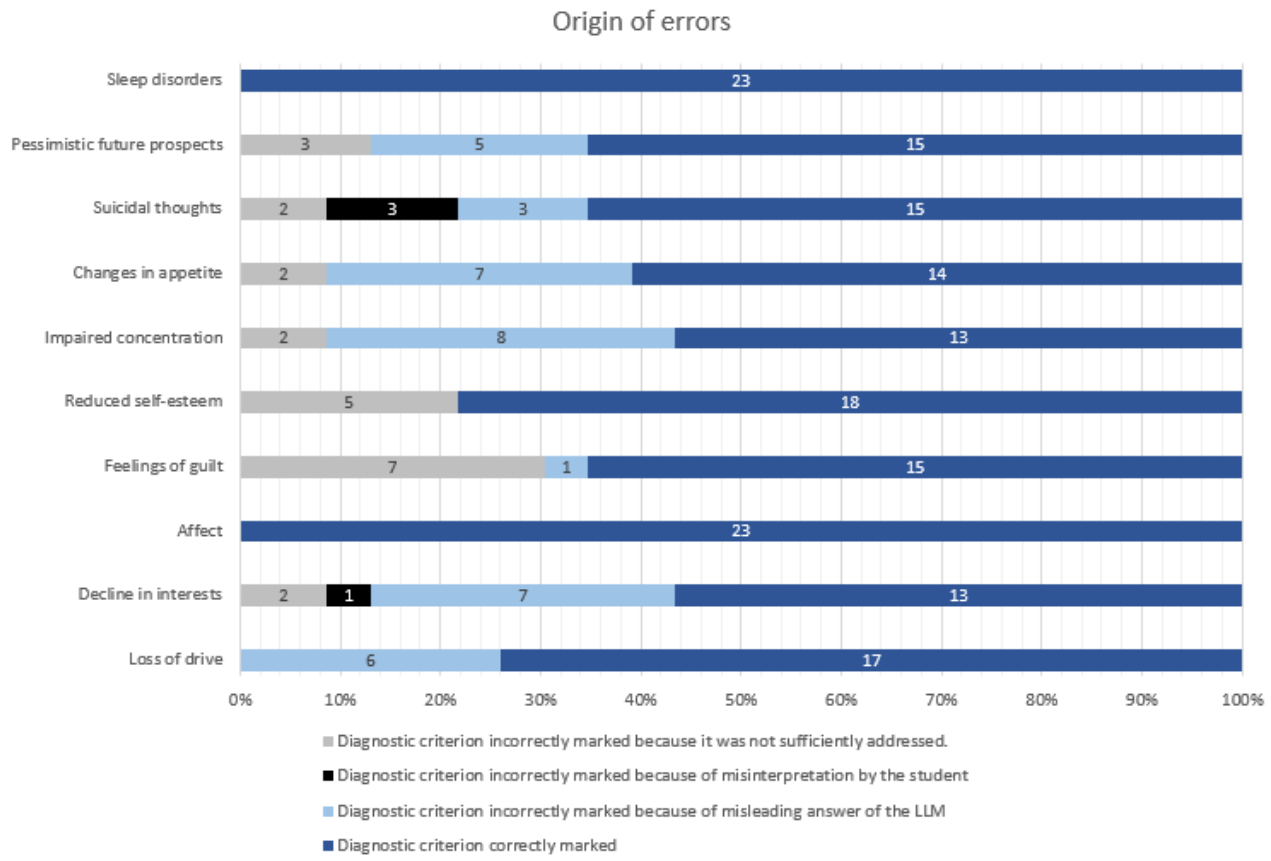
Due to discrepancies in the diagnostic criteria, a detailed analysis was conducted to verify the intended diagnostic combinations for each case. Figure S2 in Multimedia Appendix 1 shows the overall and case-related results. Even with a margin of error of ±1 criterion, only for severe depression did a significant number of students (32/38) select the correct combination. For mild and revised severe (formerly moderate) depression, fewer than one-third (13/48 mild and 12/49 revised major) achieved this goal.

To identify error sources unrelated to case vignette creation, 2 independent experts analyzed all QAPs from 23 chats across 3 cases. Three main causes were identified (1) misleading AI responses, (2) misinterpretation of AI responses by students, and (3) insufficient exploration of the criterion by students (Figure 3).

Diagnostic criteria were assessed with varying accuracy. "Depressed mood" and "sleep disorders" were correctly identified by 100% of students, while "impaired concentration" and "decline in interests" (13/23, 56.52% each) posed the greatest challenges, mainly due to misleading AI responses. Only 1.74% (4/230) of errors resulted from misinterpretation by students, while 10.00% (23/230) resulted from insufficient follow-up questions. Correct selections also included cases in which students guessed without a solid chat-based assessment.

XSL•FO
RenderX

Figure 3. Detailed analysis of errors at the chat level. Errors were categorized according to their origin (1) errors due to insufficient addressing of the user's query, (2) errors arising from misinterpretation of the input, and (3) errors resulting from misleading or incorrect responses generated by the large language model (LLM).



Students' Evaluation of Suicidality Stages

To analyze the assessment of suicidality in more detail, a differentiated evaluation of this part of the anamnesis interviews was carried out. This involved categorizing the information based on how the topic was addressed and an assessment by the experts regarding the degree to which the students evaluated this symptomatology. It showed that 21.74% (5/23) of the students did not address the issue of suicidality at all. Another 39.13% (9/23) did not address it sufficiently enough for a proper assessment, and only 39.13% (9/23) addressed it comprehensively enough.

Discussion

Overview

This study shows GPT-4's capability to accurately simulate patients exhibiting different levels of depression and stages of suicidality in a virtual patient chat format while adhering to the role script and maintaining authenticity. Additionally, the detailed analysis of student anamnesis and their interpretation of collected patient information in relation to diagnostic criteria and suspected diagnoses reveals learning challenges inherent to the diagnostic process of depression and suicidal ideation, which this training approach helps to make more transparent and analyzable.

To ensure quality and enable further optimization, evaluating GPT-4's role stability—especially with regard to different severity levels and stages of suicidality—is essential. Our findings suggest that the current prompting approach already provides a high degree of role stability. While in some cases inauthentic behavior was observed, these cases were rare and mostly due to structural deficiencies in the prompt design rather than inherent AI inconsistencies. Given the importance of role stability in medical applications, we propose specific recommendations for further improvements.

Our findings are consistent with previous studies [26,27] indicating that observed deviations are mainly due to structural issues in prompt design rather than from inconsistent AI responses. This suggests that case presentations can be further refined by addressing known sources of error. Other studies confirm that effective prompt engineering significantly enhances chatbot performance [41-43]. However, existing recommendations remain general, and to our knowledge, no medical education application uses a patient chatbot at this level of detail. Standard prompts like "avoid technical terms" often fail to produce consistent results. Instead, our approach relies on detailed, structured example responses, requiring careful case development but yielding significantly more precise and authentic AI-generated answers. Given the limited peer-reviewed literature, we also considered open-source online resources that document prompt engineering strategies [44].

XSL•FO
RenderX

In-Depth Analysis of Role-Playing

Our qualitative analysis of the chats that deviated from the predefined role led to several detailed recommendations for prompt design, particularly with regard to example responses (Textbox 1). A comprehensive list is provided in Table S4 in Multimedia Appendix 1.

Textbox 1. In-depth analysis of role-playing

Temporal structure

- Temporal changes in symptoms, especially in chronic, psychiatric, or recurrent illnesses, must be clearly structured in case vignettes. Since LLMs have difficulties with the consistent presentation of time courses, we recommend a subdivision into "current state" (eg, last 2 weeks) and "previous findings" (course of the disease), as well as the addition of a timeline to emphasize diagnostically relevant time periods.

Formal consistency

- Formal consistency is crucial for complex cases that require a detailed role script. Placing more emphasis on individual aspects can lead GPT-4 to overemphasize them, thereby misleading users.

Symptom aggravation

- We observed a tendency toward symptom aggravation, especially in mild depression. Large language models generally tend to shift toward normal variants because they reflect frequent patterns from the training data [45,46]. Possible causes are that the model associates "typical depression" with more symptoms or tends toward moderate levels, thus creating deviations from the role script. Additionally, negatively formulated diagnostic criteria (eg, "decreased energy" in *International Classification of Diseases, Tenth Revision*) may have influenced the responses. Future prompts should be neutrally formulated ("energy") and clearly distinguish between the presence and absence of a symptom.

Overall, the AI performed convincingly, aligning with previous findings [26,27]. The suggested optimizations concern fine details, and the cost–benefit ratio depends on the application context. In virtual training, absolute accuracy is not required because real patients can also provide incorrect information due to memory errors, misunderstandings, or shame [47], and medical students need to be trained for these situations. Standardized patients, who are frequently used in medical education, also show variability, although their authenticity is often discussed as a limitation [48,49]. However, for unsupervised student feedback or exam assessments, a high degree of role stability is an essential requirement.

AI-supported virtual patients offer even greater availability, flexibility, and standardization than simulated patients [50]. A key factor for acceptance in skill learning is perceived authenticity [51,52]. In this study, over two-thirds of students rated the tool as at least "convincing," with a further 29.62% (32/108) rating it as "acceptable," indicating great potential for future use.

While our results confirm the high role stability of GPT-4, certain discrepancies were found in the interpretation of the severity classifications. Due to divergent assessments by our experts—particularly regarding the diagnostic criteria "loss of interest" and "appetite changes"—the original classification of the case as "moderate depression" could not be reliably maintained. To ensure clarity and accuracy, the results were reevaluated, with this case reclassified as "revised severe (formerly moderate) depression" and the results carefully contextualized in the discussion. Where reevaluation was not feasible, these data were excluded from further interpretation.

AI Simulated Patients to Improve Students' Diagnostic Accuracy

We found that students' diagnostic accuracy in assessing the severity of depression depends heavily on its actual severity. While severe cases were consistently recognized, the fact that fewer than half (23/48, 47.92%) correctly diagnosed mild depression indicates deficits in the recognition of mild but clinically relevant cases. This highlights the need for more targeted training to ensure that future doctors can identify depression at all levels of severity at an early stage—especially those that would otherwise be overlooked and thus delay timely treatment.

This becomes particularly evident when examining the selection of diagnostic criteria used in forming a suspected diagnosis. Figure S2 in Multimedia Appendix 1 reveals a low success rate in correctly identifying all relevant criteria per case. Even with a margin of error of 1 criterion, only 27.08% (13/48) of students achieved this goal for mild depression and 24.49%% (12/49) for formerly moderate/revised severe depression. The results were best for severe depression: two-thirds of students selected the criteria correctly, presumably due to the higher number of symptoms present. It is important to note that students' performance was affected by inconsistent case vignettes and AI role-playing errors, which must be taken into account when interpreting the results. In addition, they had limited time to reach a diagnosis and only little prior training on depression and suicidality. All of this should be examined more closely in future research. The difficulty of distinguishing between mild and moderate depression contributes to the lower success rates—a problem that has also been observed in general practitioners [53,54]. According to Kroenke [55], primary care physicians correctly diagnose depression only about 50% of the time, and are more likely to over-diagnose (n=15 per 100 patients) than under-diagnose (n=10) or correctly diagnose (n=10). These findings highlight that even experienced clinicians struggle with accurate classification. This underscores the need for improved diagnostic training at all levels. *ICD-10* makes this distinction even more difficult, as it often requires only a single additional minor criterion to distinguish between moderate and mild depression [38]. Mental Health cases pose an additional layer of complexity. In our experience, such cases require a higher face validity than somatic illnesses in order to compensate

XSL•FO
RenderX

for potential uncertainty or overlap between individual diagnostic criteria [26,31]. A potential key improvement could be the future use of *ICD-11* (*International Classification of Diseases, 11th Revision*), which is intended to provide a less rigid discrimination of the severity levels of depression [56,57]. This was not used in this study, as it was not part of the curriculum at the time of data collection; however, it could be seamlessly integrated into the application in the future.

To further evaluate students' diagnostic accuracy, we conducted a detailed analysis at the QAP level (n=778). While the criteria "sleep disorder" and "depressive mood" were consistently well recognized, errors mainly occurred when students did not explore key aspects of the medical history in sufficient depth. This was particularly evident in the diagnostic criteria "feelings of guilt" and "reduced self-esteem." It is noteworthy that misinterpretations of the AI-generated answers were rare (less than 2%) and mainly related to suicidal ideation. This suggests that diagnostic difficulties were due more to insufficient follow-up questions than to a misunderstanding of the AI answers.

Neglecting Suicidal Tendencies

Even more concerning is that the majority of students (14/23, 60.87%) either did not address the topic of potential suicidality at all or did so only insufficiently—despite explicit instruction during the introduction. Our qualitative analysis shows that the LLM sometimes responded very avoidantly to questions about suicidality, which may have given students a false impression of the actual severity of the topic. However, the model never denied existing suicidality but was simply reluctant to talk about it, which adequately mirrors some real patients' behavior. Future prompts should, nonetheless, contain clearer risk indicators to guide students toward more targeted questions and avoid making the scenarios too challenging, especially at such early career training stages. Additionally, although students knew that asking explicit questions about suicidal thoughts is a standard procedure in mental health assessment, many still found it difficult to address, causing considerable discomfort and uncertainty. Literature shows that suicidality is often a shameful topic for medical students, and they fear that addressing it will intensify suicidal thoughts [58,59]. However, the opposite is the case, which underscores the need for targeted training in how to deal with suicidality [58,60,61]. The integration of AI-supported virtual patients into mental health curricula, in combination with structured feedback sessions, could offer sustainable opportunities to practice assessing suicidality and reduce discomfort and uncertainty among students.

Implications for AI-Supported Medical Training

Our results show that AI-supported online training for the diagnosis of depression and suicidality is a feasible and effective method. Nevertheless, we recommend integrating it into face-to-face formats—either fully, as in this study, or in a blended learning approach.

While other studies have shown that online training for suicide prevention can be equally effective or even superior [62,63], our results suggest that the immersive nature of virtual patient encounters contributes to a high level of perceived realism. In the feedback sessions, some students reported feeling emotionally affected, particularly when dealing with suicidality, which underscores the importance of structured support in such training formats.

The extent to which these observations can be generalized beyond our specific context remains to be investigated. Nevertheless, the potential for broader application in various medical education contexts appears promising.

Limitations

This study has several limitations. First, it was conducted at a single medical school, limiting generalizability. Second, students had limited time to interact with the AI-supported patient, reflecting real-world time constraints but possibly insufficient for a practice scenario. Since the chatbot provides more information in a limited time than a real patient typically would, the diagnostic skills practiced may not fully transfer to real-world settings, and it must be assumed that performance in actual clinical encounters could be somewhat lower. Furthermore, the participants' prior experience with AI-based systems was not assessed, which may have influenced their assessment of the tool and their interaction with it. This variable should be systematically evaluated in future studies.

Additionally, only 23 of 135 chats were analyzed at the QAP level. This selection was made because we aimed to maximize issue detection by focusing on chats in which students incorrectly assessed depression severity or suicidal ideation. This procedure ensured that the analyzed subset was both representative and information-rich. Nevertheless, analyzing full chat transcripts of incorrect diagnoses might have provided further insights.

Despite these limitations, this study provides valuable insights into the use of AI-based virtual patients to assess depression and suicidality. Future research should further develop chatbot models to support diagnostic training more comprehensively—in particular, through more precise and reliable feedback, improved conversational techniques (eg, question formulation and depth of exploration), and targeted support for clinical reasoning processes to better understand underlying thought patterns. Furthermore, it would also be interesting to examine the students' reflections on the ethical aspects of interacting with chatbot patients.

Conclusion

In summary, the chatbot can successfully simulate patients with depression with various stages of suicidality with high authenticity and adherence to predefined role scripts. Students' diagnostic accuracy was influenced by the actual severity of depression, with moderate cases posing the greatest challenge. Notably, the majority either failed to address or insufficiently explored the topic of suicidality, which negatively influenced the correct classification of the phase of suicidality. AI-supported virtual patients provide a feasible and valuable tool for psychiatric history-taking training when cases are carefully developed and reviewed by experts. Potential risks, such as misinformation due to AI limitations and ethical considerations, should be addressed in future implementations.

XSL•FO
RenderX

Future medical education cannot be envisioned without integration of AI to practice realistic patient encounters—rigorous research on how these models should be programmed and used best is thus necessary. The AI tool presented here could benefit not only students and trainees but also general practitioners by increasing awareness of depression. Due to its online nature, it can be transferred to a variety of different settings and languages; nevertheless, it may require adaptation for less commonly used languages or specific cultural contexts. Ideally, this could contribute to improving the care of patients with depressive symptoms and thus to better management of this increasing societal challenge.

Acknowledgments

We acknowledge the support of the Open Access Publishing Funds of the University of Tuebingen. The authors declare the use of generative artificial intelligence (GAI) in the research and writing process. According to the GAIDeT taxonomy [64], the following tasks were delegated to GAI tools under full human supervision: literature search and systematization, adapting and adjusting emotional tone, and translation. The GAI tool used were Chat GPT-5 and DeepL Write. Responsibility for the final manuscript lies entirely with the authors. GAI tools are not listed as authors and do not bear responsibility for the final outcomes. This declaration was submitted by LH. After using these tools, the authors thoroughly reviewed, revised, and edited the content, and take full responsibility for the final version of the manuscript.

Funding

No external financial support or grants were received from any public, commercial, or not-for-profit entities for the research, authorship, or publication of this article.

Data Availability

The datasets generated or analyzed during this study are available from the corresponding author (LH) on reasonable request.

Authors' Contributions

FH and AHW conceptualized the study. FH and TFW were responsible for data curation. AP, RE and AS performed the formal analysis, and AP and RE conducted the investigation. LH, TFW, and AS were responsible for the methodology and interpretation. LH contributed to the visualization and CSP was responsible for the software. AHW did the project administration, while SZ and AHW provided resources and supervision. FH and AHW drafted the manuscript, and all authors participated in critically revising and editing it. All authors provided final approval of the version to be published.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Additional material and prompts.
[DOCX File , 271 KB - mededu_v12i1e87102_app1.docx]

References

1. GBD 2019 Diseases and Injuries Collaborators. Global burden of 369 diseases and injuries in 204 countries and territories, 1990-2019: a systematic analysis for the Global Burden of Disease Study 2019. Lancet 2020;396(10258):1204-1222 [FREE Full text] [doi: 10.1016/S0140-6736(20)30925-9] [Medline: 33069326]
2. Institute for Health Metrics and Evaluation. GBD Results Tool. 2023. URL: https://vizhub.healthdata.org/gbd-results [accessed 2024-08-18]
3. WHO Guidelines on Mental Health at Work. First edition. Geneva: World Health Organization; 2022.
4. Diagnostic and Statistical Manual of Mental Disorders. Fifth edition. Washington: American Psychiatric Association; 2013.
5. The ICD-10 Classification of Mental and Behavioural Disorders: Clinical Descriptions and Diagnostic Guidelines. Geneva: World Health Organization; 1993.
6. Lynch CJ, Gunning FM, Liston C. Causes and consequences of diagnostic heterogeneity in depression: paths to discovering novel biological depression subtypes. Biol Psychiatry 2020;88(1):83-94. [doi: 10.1016/j.biopsych.2020.01.012] [Medline: 32171465]
7. Handy A, Mangal R, Stead TS, Coffee RL, Ganti L. Prevalence and impact of diagnosed and undiagnosed depression in the United States. Cureus 2022;14(8):e28011 [FREE Full text] [doi: 10.7759/cureus.28011] [Medline: 36134073]
8. Pelletier L, O'Donnell S, Dykxhoorn J, McRae L, Patten SB. Under-diagnosis of mood disorders in Canada. Epidemiol Psychiatr Sci 2016;26(4):414-423. [doi: 10.1017/s2045796016000329]

XSL•FO
RenderX

9. Depression in adults: treatment and management (NICE guideline NG222). National Institute for Health and Care Excellence. 2022. URL: https://www.nice.org.uk/guidance/ng222 [accessed 2024-08-18]
10. American Psychological Association. Clinical Practice Guideline for the Treatment of Depression Across Three Age Cohorts. 2019. URL: https://www.apa.org/depression-guideline [accessed 2024-08-18]
11. Arbeitsgemeinschaft der Wissenschaftlichen Medizinischen Fachgesellschaften. Nationale VersorgungsLeitlinie Unipolare Depression – Langfassung, Version 3. 2023. URL: https://www.leitlinien.de/themen/depression/version-3 [accessed 2024-08-18]
12. Depressive disorder (depression). World Health Organization. 2023. URL: https://www.who.int/news-room/fact-sheets/detail/depression [accessed 2025-01-07]
13. Mojtabai R. Clinician-identified depression in community settings: concordance with structured-interview diagnoses. Psychother Psychosom 2013;82(3):161-169. [doi: 10.1159/000345968] [Medline: 23548817]
14. Cepoiu M, McCusker J, Cole MG, Sewitch M, Belzile E, Ciampi A. Recognition of depression by non-psychiatric physicians: a systematic literature review and meta-analysis. J Gen Intern Med 2008;23(1):25-36 [FREE Full text] [doi: 10.1007/s11606-007-0428-5] [Medline: 17968628]
15. Houston K, Haw C, Townsend E, Hawton K. General practitioner contacts with patients before and after deliberate self harm. Br J Gen Pract 2003;53(490):365-370 [FREE Full text] [Medline: 12830563]
16. Richards JC, Ryan P, McCabe MP, Groom G, Hickie IB. Barriers to the effective management of depression in general practice. Aust N Z J Psychiatry 2004;38(10):795-803. [doi: 10.1080/j.1440-1614.2004.01464.x] [Medline: 15369538]
17. Levkovich I, Elyoseph Z. Identifying depression and its determinants upon initiating treatment: ChatGPT versus primary care physicians. Fam Med Community Health 2023;11(4):e002391 [FREE Full text] [doi: 10.1136/fmch-2023-002391] [Medline: 37844967]
18. Coppens E, Van Audenhove C, Gusmão R, Purebl G, Székely A. Effectiveness of General Practitioner training to improve suicide awareness and knowledge and skills towards depressionEffectiveness of general practitioner training to improve suicide awareness and knowledge and skills towards depression. J Affect Disord 2018;227:17-23 [FREE Full text] [doi: 10.1016/j.jad.2017.09.039] [Medline: 29049931]
19. van Os TWDP, van den Brink RHS, Jenner JA, van der Meer K, Tiemens BG, Ormel J. Effects on depression pharmacotherapy of a Dutch general practitioner training program. J Affect Disord 2002;71(1-3):105-111. [doi: 10.1016/s0165-0327(01)00415-3] [Medline: 12167506]
20. Henriksson S, Isacsson G. Increased antidepressant use and fewer suicides in Jämtland county, Sweden, after a primary care educational programme on the treatment of depression. Acta Psychiatr Scand 2006;114(3):159-167. [doi: 10.1111/j.1600-0447.2006.00822.x] [Medline: 16889586]
21. Public Health Action for the Prevention of Suicide: A Framework. Geneva: World Health Organization; 2012. URL: https://iris.who.int/handle/10665/75166 [accessed 2024-12-18]
22. Preventing Suicide: A Global Imperative. Geneva: World Health Organization; 2014. URL: https://iris.who.int/handle/10665/131056 [accessed 2025-01-07]
23. Yamamoto A, Koda M, Ogawa H, Miyoshi T, Maeda Y, Otsuka F. Enhancing medical interview skills through AI-simulated patient interactions: nonrandomized controlled trial. JMIR Med Educ 2024;10:e58753 [FREE Full text] [doi: 10.2196/58753] [Medline: 39312284]
24. Cook DA. Creating virtual patients using large language models: scalable, global, and low cost. Med Teach 2025;47(1):40-42. [doi: 10.1080/0142159X.2024.2376879] [Medline: 38992981]
25. Potter L, Jefferies C. Enhancing communication and clinical reasoning in medical education: building virtual patients with generative AI. Future Healthc J 2024;11:100043. [doi: 10.1016/j.fhj.2024.100043]
26. Holderried F, Stegemann-Philipps C, Herschbach L, Moldt JA, Nevins A, Griewatz J. A generative pretrained transformer (GPT)-powered chatbot as a simulated patient to practice history taking: prospective, mixed methods study. JMIR Med Educ 2024;10:e53961 [FREE Full text] [doi: 10.2196/53961] [Medline: 38227363]
27. Holderried F, Stegemann-Philipps C, Herrmann-Werner A, Festl-Wietek T, Holderried M, Eickhoff C. A language model–powered simulated patient with automated feedback for history taking: prospective study. JMIR Med Educ 2024;10:e59213 [FREE Full text] [doi: 10.2196/59213] [Medline: 39150749]
28. Atapattu T, Thilakaratne M, Do DN, Herath M, Falkner KE. Exploring the role of mental health conversational agents in training medical students and professionals: a systematic literature review. In: Che W, Nabende J, Shutova E, Pilehvar MT, editors. Findings of the Association for Computational Linguistics: ACL 2025. Vienna: Association for Computational Linguistics; 2025:20785-20798.
29. Cook DA, Erwin PJ, Triola MM. Computerized virtual patients in health professions education: a systematic review and meta-analysis. Academic Medicine 2010;85(10):1589-1602. [doi: 10.1097/acm.0b013e3181edfe13]
30. Foster A, Chaudhary N, Kim T, Waller JL, Wong J, Borish M. Using virtual patients to teach empathy. Simul Healthc 2016;11(3):181-189. [doi: 10.1097/sih.0000000000000142]
31. Chaby L, Benamara A, Pino M, Prigent E, Ravenet B, Martin JC. Embodied virtual patients as a simulation-based framework for training clinician-patient communication skills: an overview of their use in psychiatric and geriatric care. Front Virtual Real 2022;3 [FREE Full text] [doi: 10.3389/frvir.2022.827312]

XSL•FO
RenderX

32. Campillos-Llanos L, Thomas C, Bilinski É, Neuraz A, Rosset S, Zweigenbaum P. Lessons learned from the usability evaluation of a simulated patient dialogue system. J Med Syst 2021;45(7):69 [FREE Full text] [doi: 10.1007/s10916-021-01737-4] [Medline: 33999302]
33. Sezgin E, Chekeni F, Lee J, Keim S. Clinical accuracy of large language models and Google search responses to postpartum depression questions: cross-sectional study. J Med Internet Res 2023;25:e49240 [FREE Full text] [doi: 10.2196/49240] [Medline: 37695668]
34. Rosenman G, Wolf L, Hendler T. LLM questionnaire completion for automatic psychiatric assessment. arXiv Preprint posted online on June 12, 2024. [doi: 10.48550/arXiv.2406.06636]
35. Dupuy L, de Sevin E, Cassoudesalle H, Ballot O, Dehail P, Aouizerate B, et al. Guidelines for the design of a virtual patient for psychiatric interview training. J Multimodal User Interfaces 2020;15(2):99-107. [doi: 10.1007/s12193-020-00338-8]
36. DeepL Translator. 2023. URL: https://www.deepl.com/de/translator [accessed 2025-01-07]
37. Althaus D, Hegerl U. Ursachen, Diagnose und Therapie von Suizidalität [Causes, diagnosis and treatment of suicidality]. Nervenarzt 2004;75(11):1123-1135. [doi: 10.1007/s00115-004-1824-2] [Medline: 15480527]
38. Müssigbrodt H, Michels R, Malchow CP, Dilling H, Munk-Jørgensen P, Bertelsen A. Use of the ICD-10 classification in psychiatry: an international survey. Psychopathology 2000;33(2):94-99. [doi: 10.1159/000029127] [Medline: 10705253]
39. Seitz L. Artificial empathy in healthcare chatbots: does it feel authentic? Comput Hum Behav Artif Hum 2024;2(1):100067. [doi: 10.1016/j.chbah.2024.100067]
40. Braun V, Clarke V. Using thematic analysis in psychology. Qualitative Research in Psychology 2008;3(2):77-101. [doi: 10.1191/1478088706qp063oa]
41. Chen S, Wu M, Zhu KQ, Lan K, Zhang Z, Cui L. LLM-empowered chatbots for psychiatrist and patient simulation: application and evaluation. arXiv Preprint posted online on May 24, 2023. [doi: 10.48550/arXiv.2305.13614]
42. Wang R, Milani S, Chiu JC, Zhi J, Eack SM, Labrum T. PATIENT-Ψ: using large language models to simulate patients for training mental health professionals. arXiv Preprint posted online on May 28, 2024. [doi: 10.48550/arXiv.2405.19660]
43. Wang C, Li S, Lin N, Zhang X, Han Y, Wang X. Application of large language models in medical training evaluation—using ChatGPT as a standardized patient: multimetric assessment. J Med Internet Res 2025;27:e59435 [FREE Full text] [doi: 10.2196/59435] [Medline: 39742453]
44. Character design. SillyTavern. 2025. URL: https://docs.sillytavern.app/usage/core-concepts/characterdesign/ [accessed 2026-01-24]
45. Salewski L, Alaniz S, Rio-Torto I, Schulz E, Akata Z. In-context impersonation reveals large language models' strengths and biases. arXiv Preprint posted online on May 24, 2023. [doi: 10.48550/arXiv.2305.14930]
46. Gupta S, Shrivastava V, Deshpande A, Kalyan A, Clark P, Sabharwal A. Bias runs deep: implicit reasoning biases in persona-assigned LLMs. arXiv Preprint posted online November 8, 2023. [doi: 10.48550/arXiv.2311.04892]
47. Kessels RPC. Patients' memory for medical information. J R Soc Med 2003;96(5):219-222 [FREE Full text] [doi: 10.1177/014107680309600504] [Medline: 12724430]
48. Kühne F, Ay DS, Otterbeck MJ, Weck F. Standardized patients in clinical psychology and psychotherapy: a scoping review of barriers and facilitators for implementation. Acad Psychiatry 2018;42(6):773-781. [doi: 10.1007/s40596-018-0886-6] [Medline: 29423828]
49. Kühne F, Maaß U, Weck F. Standardized patients in clinical psychology: from research to practice. Verhaltenstherapie 2021;32(Suppl 1):245-253. [doi: 10.1159/000510049]
50. Barrows HS. An overview of the uses of standardized patients for teaching and evaluating clinical skills. Acad Med 1993;68(6):443-451. [doi: 10.1097/00001888-199306000-00002] [Medline: 8507309]
51. Kneebone R, Nestel D, Wetzel C, Black S, Jacklin R, Aggarwal R. The human face of simulation: patient-focused simulation training. Acad Med 2006;81(10):919-924. [doi: 10.1097/01.acm.0000238323.73623.c2]
52. Lane C, Rollnick S. The use of simulated patients and role-play in communication skills training: a review of the literature to August 2005. Patient Educ Couns 2007;67(1-2):13-20. [doi: 10.1016/j.pec.2007.02.011] [Medline: 17493780]
53. Krupinski J, Tiller JWG. The identification and treatment of depression by general practitioners. Aust N Z J Psychiatry 2001;35(6):827-832. [doi: 10.1046/j.1440-1614.2001.00960.x] [Medline: 11990894]
54. Goldman LS, Nielsen NH, Champion HC. Awareness, diagnosis, and treatment of depression. J Gen Intern Med 1999;14(9):569-580 [FREE Full text] [doi: 10.1046/j.1525-1497.1999.03478.x] [Medline: 10491249]
55. Kroenke K. Review: GPs accurately diagnose about 50% of patients with depression and accurately classify 81% of nondepressed patients. Ann Intern Med 2010;152(8):JC4. [doi: 10.7326/0003-4819-152-8-201004200-02013] [Medline: 20404370]
56. Gaebel W, Stricker J, Kerst A. Changes from ICD-10 to ICD-11 and future directionsin psychiatric classification . Dialogues Clin Neurosci 2022;22(1):7-15. [doi: 10.31887/dcns.2020.22.1/wgaebel]
57. Gaebel W, Stricker J, Riesbeck M, Zielasek J, Kerst A, Meisenzahl-Lechner E. Accuracy of diagnostic classification and clinical utility assessment of ICD-11 compared to ICD-10 in 10 mental disorders: findings from a web-based field study. Eur Arch Psychiatry Clin Neurosci 2020;270(3):281-289. [doi: 10.1007/s00406-019-01076-z] [Medline: 31654119]

XSL•FO
RenderX

58. Schulz P, Zapata I, Huzij T. Examination of medical student and physician attitudes towards suicide reveals need for required training. Front Public Health 2024;12:1331208 [FREE Full text] [doi: 10.3389/fpubh.2024.1331208] [Medline: 38633234]
59. Jamison JM, Brady M, Fang A, Bùi Trà-My N, Wolk CB, Davis M. A qualitative examination of clinician anxiety about suicide prevention and its impact on clinical practice. Community Ment Health J 2025;61(3):568-575. [doi: 10.1007/s10597-024-01364-6] [Medline: 39361092]
60. Omerov P, Steineck G, Dyregrov K, Runeson B, Nyberg U. The ethics of doing nothing: suicide-bereavement and research—ethical and methodological considerations. Psychol Med 2014;44(16):3409-3420 [FREE Full text] [doi: 10.1017/S0033291713001670] [Medline: 23867073]
61. Dazzi T, Gribble R, Wessely S, Fear NT. Does asking about suicide and related behaviours induce suicidal ideation? What is the evidence? Psychol Med 2014;44(16):3361-3363. [doi: 10.1017/s0033291714001299]
62. Holmes G, Clacy A, Hamilton A, Kõlves K. Online versus in-person gatekeeper suicide prevention training: comparison in a community sample. J Ment Health 2024;33(5):605-612 [FREE Full text] [doi: 10.1080/09638237.2024.2332811] [Medline: 38602188]
63. Pechek AA, Vincenzes KA, Forziat-Pytel K, Nowakowski S, Romero-Lucero L. Teaching suicide assessment and intervention online: a model of practice. Prof Couns 2024;14(1):100-112 [FREE Full text]
64. Suchikova Y, Tsybuliak N, Teixeira da Silva JA, Nazarovets S. GAIDeT (Generative AI Delegation Taxonomy): a taxonomy for humans to delegate tasks to generative artificial intelligence in scientific research and publishing. Accountability Res 2025:1-27. [doi: 10.1080/08989621.2025.2544331] [Medline: 40781729]

Abbreviations

AI: artificial intelligence
DSM-5: Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition
GPT: generative pretrained transformer
ICD-10: International Statistical Classification of Diseases, Tenth Revision
ICD-11: International Classification of Diseases, 11th Revision
LLM: large language model
PSM: psychosomatic medicine and psychotherapy
QAP: question-answer pair

Edited by A Stone, T Leung; submitted 04.Nov.2025; peer-reviewed by S Schmidt, D Shin; comments to author 20.Nov.2025; accepted 29.Dec.2025; published 12.Feb.2026.

Please cite as:
Holderried F, Sonanini A, Philipps A, Stegemann-Philipps C, Herschbach L, Festl-Wietek T, Zipfel S, Erschens R, Herrmann-Werner A
Using AI to Train Future Clinicians in Depression Assessment: Feasibility Study
JMIR Med Educ 2026;12:e87102
URL: https://mededu.jmir.org/2026/1/e87102
doi:10.2196/87102
PMID:

©Friederike Holderried, Alessandra Sonanini, Annika Philipps, Christian Stegemann-Philipps, Lea Herschbach, Teresa Festl-Wietek, Stephan Zipfel, Rebecca Erschens, Anne Herrmann-Werner. Originally published in JMIR Medical Education (https://mededu.jmir.org), 12.Feb.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Education, is properly cited. The complete bibliographic information, a link to the original publication on https://mededu.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Susceptibility of Assessment Types to AI-Generated Content in Digital Health and Health Information Management Education: Quasi-Experimental Pilot Study

Tafheem Ahmad Wani[1], PhD; Michael Liem[1], PhD; Natasha Prasad[1], MHIM; Kerin Robinson[1], PhD; Abbey Nexhip[1], BHSc (Hons), BHSc; Melanie Tassos[1], MAppSc; Stephanie Gjorgioski[1], BHSc (Hons), BHIM; Urooj Raza Khan[1], PhD; James Boyd[1], PhD; Merilyn Riley[1], PhD

School of Psychology and Public Health, La Trobe University, Melbourne, Victoria, Australia

Corresponding Author:
Tafheem Ahmad Wani, PhD
School of Psychology and Public Health
La Trobe University
Health Sciences 2
Plenty Road, Bundoora
Melbourne, Victoria, 3086
Australia
Phone: 61 451906170
Email: t.wani@latrobe.edu.au

Abstract

Background: Generative artificial intelligence (AI) tools, such as ChatGPT, are increasingly used in higher education and have raised significant concerns about assessment validity and academic integrity. In Digital Health and Health Information Management (DIGHIM) programs, assessments are designed to evaluate a mix of technical skills, contextual reasoning, and professional judgment that underpin medical and health practice. Understanding how generative AI performs across different assessment types is, therefore, critical to identifying which formats are most susceptible to AI-generated content and how assessments may be redesigned to remain authentic and educationally meaningful.

Objective: This study aimed to evaluate ChatGPT's performance across diverse assessment types in DIGHIM education by examining how task complexity influences AI-generated output quality, and develop recommendations for ethical and effective AI integration in assessments.

Methods: A pilot quasi-experimental design compared ChatGPT-generated responses with deidentified student submissions across 5 assessment types: digital health solution design, business case analysis, reflective assessment, SQL health database programming, and a health classification quiz. For each task, multiple AI submissions were produced using different prompting strategies, including rubric integration and the use of ChatGPT (GPT-4 and o1 Preview model). Blinded academic markers evaluated all AI-generated submissions and previously submitted deidentified student assessments against standard rubrics, and descriptive statistics were used to compare performance.

Results: ChatGPT's performance varied considerably across assessment types. It achieved its highest accuracy scores in objective, rule-based tasks such as multiple-choice quiz items in health classification (mean 88%, SD 0%) and produced well-structured, coherent responses for reflective assessments (mean 69%, SD 12.8%), though these often lacked personalization and nuanced industry context. In descriptive analytical tasks, such as digital health business cases and solution designs, ChatGPT produced logically structured work with reasonable use of evidence but failed to provide deep contextualization, domain-specific insights, or visual elements expected in DIGHIM practice. Technical assessments revealed the greatest limitations: SQL programming tasks averaged 42% (SD 17.2%) with persistent schema errors, incomplete queries, and weak interpretation of health data outputs, while scenario-based clinical coding scored just 7% (SD 0%), reflecting a lack of precision in applying *ICD-10-AM* (*International Classification of Diseases, Tenth Revision, Australian Modification*) rules and coding conventions. Structured prompting and rubric integration improved results, particularly in descriptive and reflective tasks (up to 80%), but the advanced o1 Preview model did not consistently outperform earlier versions.

Conclusions: While ChatGPT performs well in structured, rule-based, and reflective tasks, it remains limited in technical accuracy, contextual reasoning, and applied DIGHIM competencies. To support academic integrity and workforce readiness,

XSL•FO
RenderX

assessment design should prioritize critical thinking, ethical reasoning, and scenario-based problem-solving aligned with health care practice. Using AI as a tool for critique and refinement, rather than a substitute for student work, may help educators prepare learners for responsible AI use in medical and health professional education.

(JMIR Med Educ 2026;12:e82988) doi:10.2196/82988

KEYWORDS

academic integrity; assessment design; ChatGPT performance; digital health education; generative artificial intelligence; health information management; quasi-experimental study

Introduction

Background

Artificial intelligence (AI) is a transformative technology that enables systems to generate outputs, such as content, recommendations, or decisions, based on human-defined objectives [1,2]. A key branch, generative AI (GenAI), produces human-like media (text, images, videos, and code) in response to prompts [3]. At the forefront are large language models (LLMs) such as ChatGPT (OpenAI), Microsoft Copilot, and Gemini (Google), which use deep learning and vast datasets to understand context, generate coherent responses, and adapt tone and style [4-8]. These models improve continuously through reinforcement learning with human feedback [9-11].

In higher education, LLMs have rapidly gained attention among academics and students [8,12-14], supporting assessment design, automated marking, and curriculum development [14-17], improving feedback timeliness and reducing grading errors [14,18,19]. They also help students refine language, generate ideas, and start research inquiries [20]. Overall, recent evidence in medical education highlights the rapid diversification of GenAI applications across learning resources, instructional methods, and assessment practices, offering tangible opportunities in areas such as documentation support, simulation-based training, and personalized learning [21-23].

Despite the benefits, the integration of GenAI technologies into higher education can present significant ethical challenges [24]. These include ChatGPT's use of inaccurate content (including fictitious reference material) [25], concerns regarding human teacher replacement [11], and negative repercussions on students' critical thinking and problem-solving skills [26]. Most significantly, the use of ChatGPT can pose a threat to academic integrity and ethics [27]. A recent 2024 scoping review has highlighted that traditional assessment methods do not operate effectively in GenAI-facilitated learning environments, prompting the need for innovative and refocused assessment designs that foster career-driven competencies and lifelong learning skills [28]. The ethical framework surrounding students' use of AI in educational assessments primarily involves transparency, accountability, and reliability [29]. Transparency requires that students acknowledge the use of AI appropriately in their submissions. Students must also be accountable for the content they submit, as "outputs of AI tools can include biased, inaccurate, or incorrect content that users should be aware of" [29]. In some instances, GenAI can fabricate information, which raises the question of the reliability of the outputs [7].

The accessibility of AI tools and the difficulty in detecting AI-generated content [7,16] may tempt students to engage in academic misconduct. This raises concerns about assessment fairness, equity, and the credibility of academic credentials earned through digital platforms. Educators, therefore, face challenges in maintaining the integrity of assessments and ensuring educational quality [16,30]. Recent studies also report learners' desire for clear institutional guidance on appropriate AI use, aligning with calls for explicit policy and training [31,32].

La Trobe University, Australia, offers a wide range of health programs. The Digital Health and Health Information Management (DIGHIM) programs provide formal education in disciplines that support contemporary health care and medical practice, aiming to develop students' technical, professional, and interpersonal skills [33,34]. Typical enrolments include undergraduate and master's-level students from clinical or quasi-clinical backgrounds such as nursing, allied health, health sciences, health administration, biomedical sciences, and related fields, as well as graduates seeking roles at the intersection of health care delivery, health data, and digital systems. The curriculum is aligned with medical education priorities, including clinical documentation, health classification, health data governance, digital health (DH) system design, and the ethical use of health information, all of which directly support contemporary medical practice and health service delivery. The DIGHIM programs emphasize discipline-specific and generic competencies, which are inclusive of problem-solving, critical thinking, and ethical decision-making that are typically assessed through academic essays, reports, multiple-choice quizzes, presentations, case studies, programming exercises, and practical simulations [35,36]. The proliferation of AI-generated content threatens the validity and reliability of assessments in most of these areas, potentially compromising the effectiveness of evaluating students' comprehension and competence [16]. In this context, the broader medical-education literature now differentiates by task type and competency domain: from automated short-answer scoring and case generation to simulation and reflective work, strengthening the rationale for DIGHIM-specific analysis of assessment susceptibility and redesign [21,37-39].

Previous research on the impact of GenAI tools in higher education has primarily focused on specific types of assessments, such as multiple-choice questions or essay evaluations, mainly within the medical [40-42] and business fields [4]. For example, Chaudhry et al [16] undertook an evaluation of AI-generated assessments in a Bachelor of Business Administration, including case analysis, empirical

XSL•FO
RenderX

study report, self-reflection, group work, and calculation-based assessments. These findings do not necessarily translate well to the more specialized fields of DH and health information management (HIM), each of which requires a distinct combination of technical, analytical, and decision-making skills. DIGHIM programs, for instance, demand expertise in areas such as health classification, epidemiology and biostatistics, DH solution design, health data governance, and interoperability [33,34]. These specialist areas of study necessitate a rigorous and diverse set of assessment methods, incorporating scenario-based problem-solving, data interpretation, and system implementation tasks that test both theoretical knowledge and practical application. Given the interdisciplinary nature of DH, which aims to bridge clinical practice, data science, and IT, there is a need for a separate evaluation of how GenAI interacts with these unique assessments [43]. This study responds to that gap by examining assessment type susceptibility in DIGHIM and aligning implications with contemporary redesign strategies and governance recommendations emerging from the 2024-2025 literature [21,31,32,44].

Aims

This study examined how GenAI, specifically ChatGPT, performed across different DIGHIM assessment types and how its outputs compared with student work, with the aim of identifying assessment formats most susceptible to AI-generated content. It further sought to inform assessment redesign by identifying approaches that promoted authenticity, higher-order reasoning, and responsible AI use, thereby strengthening academic integrity through design rather than detection.

Methods

Study Design

A quasi-experimental design was piloted to evaluate ChatGPT's performance on a range of assessment types, comparing its results with submitted assessments of past students. This exploratory approach was intended to provide preliminary insights and inform the design of larger-scale studies. While similar methods have been applied in ChatGPT response evaluation within the education sector, this study pilots the approach in the context of DIGHIM, where it has not yet been examined [11,16,41].

Assessment Materials and Markers (Participants)

Assessments from 5 DIGHIM subjects were purposively selected because of specific assessment characteristics, to maximize diversity and breadth in the type and nature of assessments chosen (Table 1). Past students' deidentified assessments, originally submitted for academic credit, were reevaluated for research purposes as a benchmark against which the ChatGPT-generated responses were compared. Markers, blinded to the origin of each assessment, were responsible for evaluating both the ChatGPT-generated assessments and the past students' assessments. Markers were academic staff, both full-time and casual, with postgraduate qualifications and advanced professional and teaching experience in DH or HIM relevant to the specific assessments they were assigned, and with demonstrated expertise in evaluating student work using standardized marking rubrics.

XSL•FO
RenderX

Table 1. Summary of assessment types/sample chosen for the study.

Assessment type	Level	Discipline	Assessment description	Experimental (AI[a]) group assessments marked, n	Control (student) group assessments marked, n
Digital health solution design	Masters	Digital health	Propose and justify a structured approach to addressing a specific health care challenge through a digital health solution, incorporating design principles, implementation frameworks, and the consideration of barriers and enablers.	4	3
Digital health business case proposal report	Masters	Digital health	Develop a proposal identifying gaps in a digital health system, analyzing barriers, and recommending evidence-based, innovative solutions.	4	3
Health information management reflective assessment	Fourth-year undergraduate	Health information management	Reflection on health information management students' professional practice (work-integrated learning) placement experience and learning	4	3
Health database SQL programming	Second-year undergraduate	Health information management	Use of SQL to query and analyze a health database, generating reports, and extracting relevant data	4	3
HIM[c] health classification online quiz examination	First-year undergraduate	Health information management	Online examination comprising long, short, and objective (MCQ[b], true/false, and 1-word answer) questions, assessing knowledge of health classification systems (*ICD-10-AM*[d]) and clinical coding principles	2	3

[a]AI: artificial intelligence.

[b]MCQ: multiple-choice question.

[c]HIM: health information management.

[d]*ICD-10-AM*: International Classification of Diseases, Tenth Revision, Australian Modification.

Sample

Control Group

Subject coordinators selected a sample of student assessments (3 per each assessment type) completed between July 1, 2023, and June 30, 2024, to form the control group. To ensure representativeness, assessments were chosen across 3 grade bands: high performance (>80%), medium performance (70%-79%), and low performance (50%-59%). This resulted in the inclusion of a total of 15 student assessments across the 5 subjects and 5 assessment types. All assessments were deidentified prior to analysis to maintain student anonymity.

Experimental Group

ChatGPT was used to complete the same assessment tasks as were selected for the control group. A total of 2-4 AI-generated assessments were created for each assessment type (Table 1; n=18) to reflect variations in how a typical student might approach the task. The criteria for these variations and their alignment with student input patterns are detailed below in the Experiment section.

In total, 33 assessments, across the 5 assessment types, were selected for the study: 18 were AI-generated (experimental group), and 15 were past students' assessments (control group). The 5 assessment types include both undergraduate and postgraduate assessments, reflecting core educational activities that develop competencies directly relevant to medical education, including clinical documentation, health classification, health data analysis, DH system design, and reflective practice in health care settings.

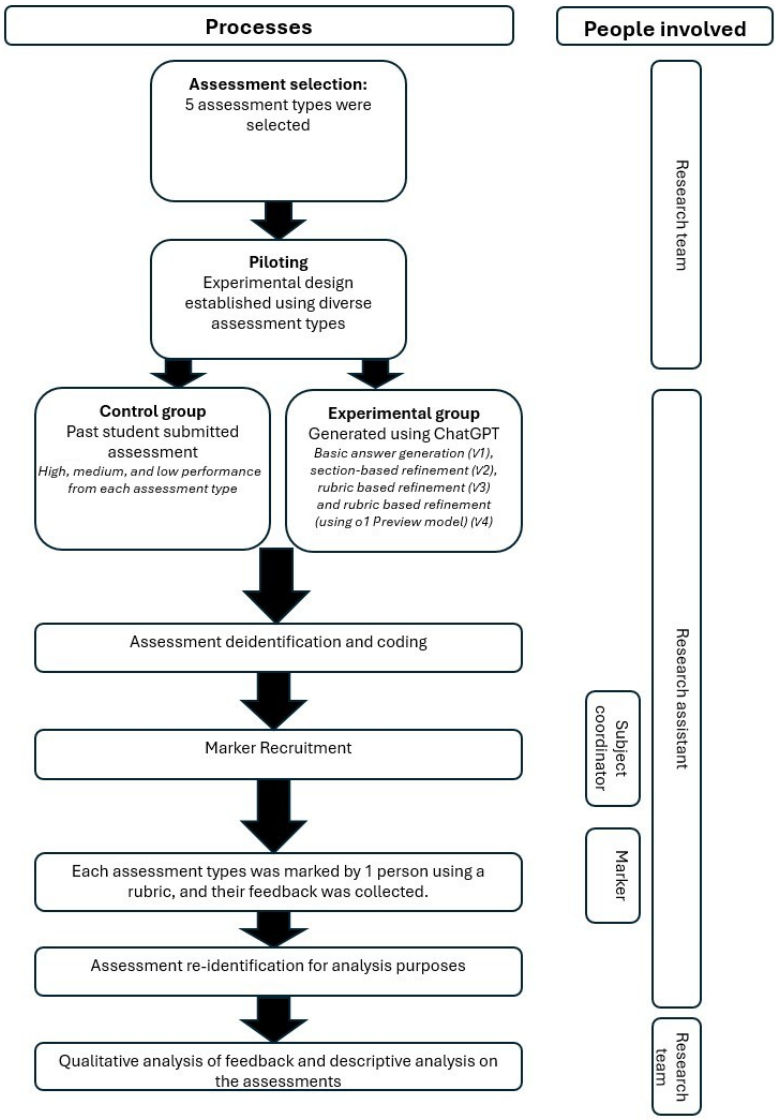
Experiment

Overview

An independent research assistant (RA) followed a structured process, established by the researchers, to gather ChatGPT assessment responses for the chosen sample of assessments (Figure 1).

XSL•FO
RenderX

Figure 1. Overview of methodology.



Generating ChatGPT Responses

To ensure reliability, completeness, and alignment with academic expectations, multiple submission versions (ie, basic answer generation [V1], section-based refinement [V2], rubric-based refinement [V3], and rubric-based refinement using o1 Preview model [V4]) of assessment instructions were developed for each assessment type for submission to ChatGPT and its versions (Table 2). Following the preliminary testing phase described in the Testing and Review subsection below, a refined and detailed protocol was created to guide the research assistant in generating all 4 structured versions (V1-V4; Multimedia Appendix 1).

It was expected that full-instruction (V1) and step-by-step approaches (V2 and V3) would allow for a nuanced analysis of AI-generated outputs, with V2 and V3 specifically designed to assess how incremental guidance impacted response quality. Furthermore, inclusion of rubrics (V3) was expected to further improve contextual alignment. The inclusion of o1 Preview (V4) used the o1 model, which was designed for advanced reasoning, and provided insights into GenAI's evolving capabilities. For SQL tasks, submission versions (V2 and V3) included the database schema and rubric to support contextual accuracy; however, the underlying health dataset itself was not provided to ChatGPT. In contrast, for objective-style questions in the health classification assessment, only 2 submission versions were needed due to the structured format of the questions and the straightforward nature of the required responses. It is important to note that in all versions, ChatGPT was provided with the same assessment instructions and marking rubrics that were originally given to students, ensuring comparability between human and AI-generated work.

XSL•FO
RenderX

Table 2. Description of artificial intelligence–generated submission versions and prompting strategies.

Assessment type and submission version	Description
Subjective and analytical assessments (reflective assessment, business case proposal/case study, and digital health solution)	
V1	Full assessment instructions entered into ChatGPT using the GPT-4 model for a single-step response generation.
V2	Assessment broken into parts as per instructions; ChatGPT using the GPT-4 model generated each part separately and compiled.
V3	Same as V2, with the rubric provided to guide response generation for better alignment with criteria.
V4	Same as V3, but generated using the o1 Preview model to evaluate differences in performance.
SQL/programming assessment	
V1	Query instructions provided without a schema for response generation using GPT-4.
V2	Query instructions with the schema provided to enhance contextual accuracy using GPT-4.
V3	Schema and rubric provided to ensure responses aligned with evaluation criteria using GPT-4.
V4	Same as V3, but generated using the o1 Preview model for comparative analysis.
HIM health classification/clinical coding online quiz examination (scenario-based and objective questions)	
V1	Objective questions generated using GPT-4.
V2	Objective questions generated using the o1 Preview model to compare differences in outputs.

Blinded Review Preparation

The research assistant formatted ChatGPT-generated responses to resemble student submissions. These AI-generated responses were then mixed with real, deidentified student assessments to maintain blinding. To ensure an unbiased review process, all assessments were anonymized and assigned coded identifiers.

Assessment Allocation

Based on their expertise in the subject, each subject coordinator suggested potential markers to the RA to assess and mark both ChatGPT responses and student assessments. The RA contacted potential markers individually via email to seek their participation and to shortlist 1 marker per assessment type. Markers were provided with participant information consent forms. To maintain impartiality, the nominated markers did not include the academics who originally marked the student assessments. Assessments were randomly assigned to the markers by the RA, acting as an independent coordinator to ensure a fair distribution and prevent any conflicts of interest.

Marking Process

Markers followed a standardized marking rubric to evaluate the assessments. To ensure consistency in grading, clear instructions were provided by the principal investigator and communicated by the independent RA. Additionally, support was available throughout the process to address any questions or uncertainties. In addition to numerical scores, markers provided written feedback aligned with the rubric criteria, along with overall free-form comments to justify their grading and offer insights into the quality of the work. To further ensure reliability, if a discrepancy of more than 10 marks was identified in any student assessment (between the original mark and remark), all assessments within that assessment type were independently reviewed by a second subject expert. Any differences were then discussed, and a consensus was reached to finalize the grades.

Data Tracking and Management

The RA maintained a Microsoft Excel spreadsheet to track the marking process, including details of assessments sent to staff members and their feedback.

Testing and Review

As part of the experimental process, an initial round of testing was undertaken to refine the methodology and confirm the feasibility of the planned research. This involved a smaller sample of assessments across various types (Table 2) to trial the generation of AI responses, the collection and deidentification of student work, and the blinded review process (see the sections “Blinded Review Preparation,” “Assessment Allocation,” “Marking Process,” and “Data Tracking and Management” above). Insights from this stage were used to make necessary adjustments prior to the main experiment. The assessments used during this testing phase were excluded from the final data analysis. During preliminary testing, we determined that generating multiple outputs from the same input was not suitable for this study’s design. When repeated prompts were issued within the same ChatGPT account, model-state carryover effects created dependency between responses due to latent conversational memory, thereby compromising independence across outputs. Conversely, when the same inputs were generated using different accounts, the resulting outputs were highly similar in structure, sequencing, and wording. Such similarity posed a risk to the blinded marking process, as a single marker evaluating near-identical responses could readily identify them as AI-generated, introducing potential confirmation bias. For these reasons, we adopted a structured versioning approach (V1-V4) instead of generating multiple outputs for identical inputs.

Synthesis and Analysis

Marks and qualitative feedback provided by academic markers were compiled for each assessment type. Descriptive statistics

XSL•FO
RenderX

(means, percentages, and score ranges) were calculated across all student grade bands (high, medium, and low) to support the blinded marking process and ensure natural variation within the control group. While ChatGPT outputs were compared against all remarked student assessments, the primary analytical focus was on the highest-graded submissions, as these offered a stable and meaningful benchmark for assessing whether AI-generated work could approximate high-quality human performance. Performance was examined both within individual assessment types and across assessment categories.

Comparative trend analysis was undertaken to evaluate progression across ChatGPT versions (V1-V4), focusing on improvements associated with structured prompting, schema inclusion, and rubric integration. In addition, rubric-level synthesis was conducted to identify recurring strengths and weaknesses.

Feedback comments for both student and AI-generated submissions were reviewed and compared, with particular focus on top-performing student work. This comparison helped identify where the evaluative feedback on AI outputs differed most noticeably from that of human-authored work.

Results were further synthesized into cross-assessment comparisons, with performance grouped into broader task categories (objective, reflective, descriptive analytical, scenario-based analytical, communication/referencing, and programming) to identify where AI demonstrated relative strengths versus persistent limitations.

GenAI Tools Used

The study used GPT-4 (free version/limited) and o1 Preview (pro/paid version) for the experimentation. The researchers opted out of ChatGPT's AI training model using an available feature, ensuring that the submitted assessment guidelines and instructions were not used to further train the AI. All necessary privacy settings were enabled to maintain data confidentiality throughout the research process.

Ethical Considerations

The research study was approved by La Trobe University's Human Research Ethics Committee (application number HEC24286). No student assessment submissions, student-generated content, or identifiable educational data were entered into any GenAI tools as part of this study. Only publicly available assessment instructions and marking rubrics were provided to ChatGPT to generate AI responses. All student work used for comparison was deidentified, stored securely, and assessed offline by academic markers.

Results

Performance by Assessment Type

Overview

This section provides a detailed breakdown of ChatGPT's performance across each assessment type, with comparisons to the highest-graded student assessments. Comprehensive data analysis and full qualitative feedback for all assessments are provided in Multimedia Appendix 2. This appendix includes comparisons with all types of student submissions. Results are presented by assessment type and task category in the same sequence as outlined in the Methods section, enabling direct comparison between study design and outcomes.

Assessment Type 1: DH Solution Design

This assessment required students to design and justify a DH solution using implementation frameworks and a critical analysis of barriers and enablers, assessed against rubric criteria for problem framing, contextual relevance, and professional communication.

AI-generated responses showed progressive improvements with structured prompting and rubric integration, most notably in GPT-4 (V3), which scored 69% compared to 56% for GPT-4 (V1; Table 3). Overall, even with these refinements, AI outputs lacked the depth, specificity, and contextual alignment observed in student submissions. Notably, the use of the more advanced o1 model (V4) did not improve performance over GPT-4 (V3), scoring only 63%, and continued to exhibit generic recommendations and limited tailoring to the scenario. In contrast, the top-performing student submission (81%) demonstrated nuanced justification, visual clarity, and strong alignment with project objectives.

XSL•FO
RenderX

Table 3. Comparison of outcomes from ChatGPT submission versions for digital health solution assessment.

GPT model and submission version	Score (%)	Key positive points noted by the marker	Key points for improvement noted by the marker
GPT-4 (V1[a])	56	• Design stages justification incomplete	• Vague barrier discussion • Unclear language • Misplaced citations
GPT-4 (V2[b])	60	• Improved clarity and referencing (than V1) • Better articulation of barriers	• Generic roadmap activities • Limited alignment with goals
GPT-4 (V3[c])	69	• Design approach supported with clear evidence • Better understanding of challenges • Strong structure • Clearer communication • Improved referencing	• Activities remained broad • Lacked visual aids and specificity
o1 Preview (V4[d])	63	• Consistent framework discussion • Professional tone	• Roadmap remained generic • Limited contextual alignment

[a]V1: basic answer generation.

[b]V2: section-based refinement.

[c]V3: rubric-based refinement.

[d]V4: rubric-based refinement using o1 Preview model.

Assessment Type 2: DH Business Case Proposal Report

This assessment required students to develop a DH business proposal identifying system gaps and recommending innovative technologies, with emphasis on evidence-based justification, analysis of implementation barriers, and professional communication demonstrating health care value and impact.

ChatGPT-generated submissions yielded similar scores across versions (Table 4). GPT-4 (V2) scored the highest (68%), showing improved referencing and clearer structure. However, despite incremental refinements, all AI outputs were critiqued for their generic content, lack of tailored insights, and absence of visual elements such as patient journey maps or system flow diagrams. Notably, the more advanced o1 Preview model (V4) did not improve performance (score: 64%) and continued to present directive rather than strategic framing. In contrast, the top-performing student submission (score: 82%) demonstrated personalized problem-solution alignment, stronger analytical depth, and enhanced clarity through well-integrated visual aids.

Table 4. Comparison of outcomes from ChatGPT submission versions (V1-V4)—digital health business proposal/case study assessment.

GPT model and submission version	Score (%)	Key positive points noted by the marker	Key points for improvement noted by the marker
GPT-4 (V1[a])	62	• Basic identification of system gaps • logical structure	• Weak justification of solutions • insufficient references • vague insights
GPT-4 (V2[b])	68	• Clearer articulation of barriers • improved referencing and layout	• Lacked personalization • limited strategic alignment with health care needs
GPT-4 (V3[c])	67	• Stronger evidence-based rationale • professional tone	• Solutions not well integrated with patient needs • absence of visual aids such as journey maps
o1 Preview (V4[d])	64	• Consistently identified issues and proposed digital interventions	• Lacked strategic framing • weak linkage between problems and proposed actions

[a]V1: basic answer generation.

[b]V2: section-based refinement.

[c]V3: rubric-based refinement.

[d]V4: rubric-based refinement using o1 Preview model.

Assessment Type 3: HIM-Reflective Assessment

This assessment required students to critically reflect on their HIM placement, using empirical evidence to evaluate professional competencies, identify strengths and weaknesses, set actionable development goals, and demonstrate professional communication.

ChatGPT-generated responses showed a wide range in performance (Table 5). GPT-4 (V3) achieved the highest AI score (80%), presenting a well-structured report with detailed

XSL•FO
RenderX

reflections and systematic use of empirical evidence, even surpassing the top student marked (70%). However, it exceeded the word limit and lacked visual enhancements, while some goals remained broad and misaligned with competencies expected in the health industry. Earlier versions showed key weaknesses: V1 (53%) was concise but underdeveloped and lacked depth in reflection and linkage to placement experience, and V2 (70%) improved on structure and relevance but included fabricated references and overly clinical goals. The most advanced version, o1 Preview (V4), scored only 55%, meeting word limits but offering shallow insights and repetitive, vague reflections.

While ChatGPT demonstrated fluency, structure, and effective use of reflective frameworks, it fell short in personalization, contextual relevance, and depth, elements that distinguished the top student submission, which provided clear placement-specific insights, nuanced analysis, and realistic strategies for professional development. This also aided in evaluating whether AI could produce realistic reflective narratives and to inform how such assessments might be strengthened in the future.

Table 5. Comparison of outcomes from ChatGPT submission versions (V1-V4)—HIM reflective assessment.

GPT model and submission version	Score (%)	Key positive points noted by the marker	Key points for improvement noted by the marker
GPT-4 (V1[a])	53	• Structured format • Identified gaps in reflective practice	• Omitted details • Weak placement linkage • Lacked depth and clarity
GPT-4 (V2[b])	70	• Strong organization • Good competency linkage, visual structure	• Overly clinical focus • Exceeded word count • Included fabricated reference
GPT-4 (V3[c])	80	• Comprehensive, well-referenced • Detailed placement reflection	• Wordy • Lacked visuals • Some goals too broad for HIM context
o1 Preview (V4[d])	55	• Concise and grammatically sound • Some relevant evidence	• Shallow reflection • Repetitive content • Weak critical analysis

[a]V1: basic answer generation.

[b]V2: section-based refinement.

[c]V3: rubric-based refinement.

[d]V4: rubric-based refinement using o1 Preview model.

Assessment Type 4: Health Database SQL Programming

This assessment evaluated students' ability to assess identification of health information needs, write accurate SQL queries to analyze health data, extract relevant information, and interpret results to support health care decision-making.

AI-generated responses demonstrated only modest improvement across iterations, with schema and rubric integration resulting in higher scores (Table 6). GPT-4 (V1), generated without schema input, scored the lowest (17%), with all queries based on incorrect table and column names. Schema inclusion in V2 improved structural accuracy (47%), but errors in data grouping and missing outputs persisted. V3 (49%) refined query logic slightly, though misspellings and omission of key fields remained problematic. The most advanced model, o1 Preview (V4), achieved the highest AI score (56%), correctly executing about half of the queries and demonstrating improved adherence to SQL conventions, but still suffered from inaccuracies in age group breakdowns, incorrect counts, and missing contextual details such as diagnosis descriptions.

In contrast, the top student submission (100%) showed precise application of SQL logic, full query completion, and context-aware interpretation of health trends. It demonstrated superior attention to schema structure, data accuracy, and professional communication in presenting findings, areas where ChatGPT consistently underperformed.

XSL•FO
RenderX

Table 6. Comparison of outcomes from ChatGPT submission versions (V1-V4)—health database SQL programming.

GPT model and submission version	Score (%)	Key positive points noted by the marker	Key points for improvement noted by the marker
GPT-4 (V1[a])	17	• Correct use of SQL structure • Basic understanding of syntax	• All table/column names incorrect • No valid outputs • Missing queries
GPT-4 (V2[b])	47	• Some correct queries • Schema improved the structural logic	• Wrong grouping • Spelling errors • Partial outputs • Inconsistent results
GPT-4 (V3[c])	49	• Better rubric alignment • Clearer structure	• Inaccurate grouping • Incomplete queries • Diagnosis data omitted
o1 Preview (V4[d])	56	• Improved accuracy • Half the queries returned correct results	• Ongoing errors in counts and terminology • Limited interpretation

[a]V1: basic answer generation.

[b]V2: section-based refinement.

[c]V3: rubric-based refinement.

[d]V4: rubric-based refinement using o1 Preview model.

Assessment Type 5: HIM Health Classification/Clinical Coding Online Quiz Examination

This quiz assessed students' application of *ICD-10-AM* (*International Classification of Diseases, Tenth Revision, Australian Modification*) health classification standards through scenario-based coding, line coding, and objective questions.

AI responses (GPT-4 [V1] and o1 Preview [V2]) performed best in the objective section (88%) of the quiz but scored only 7% in scenario-based coding tasks, with a 32% overall score, compared to 87% scored by the top student (Table 7). Key issues included incorrect block numbers, inaccurate sequencing, and vague or missing code justifications. In contrast, the top student demonstrated precision and contextual understanding aligned with national coding standards. These results reinforce that while AI handles structured recall effectively, it lacks the nuanced reasoning required for applied clinical coding.

Table 7. Comparison of outcomes from ChatGPT submission versions (V1-V4)—health classification quiz. Only 2 versions were tested, as rubric- and breakdown-based versions were not applicable to this quiz format.

GPT model and submission version	Score (%)	Key positive points noted by the marker	Key points for improvement noted by the marker
GPT-4 (V1[a])	32.4	• Strong performance in the objective section (14/16 correct)	• Incorrect block numbers • Poor sequencing • Inaccurate tabular usage
o1 Preview (V2[b])	32.4	• Consistent objective accuracy • Good understanding of conventions	• Weak in scenario coding • Missing procedural codes • Vague codes/justifications

[a]V1: basic answer generation.

[b]V2: section-based refinement using o1 Preview model.

Cumulative Performance Comparison Across Assessment Types and Tasks

Comparative Analysis Across Assessment Types

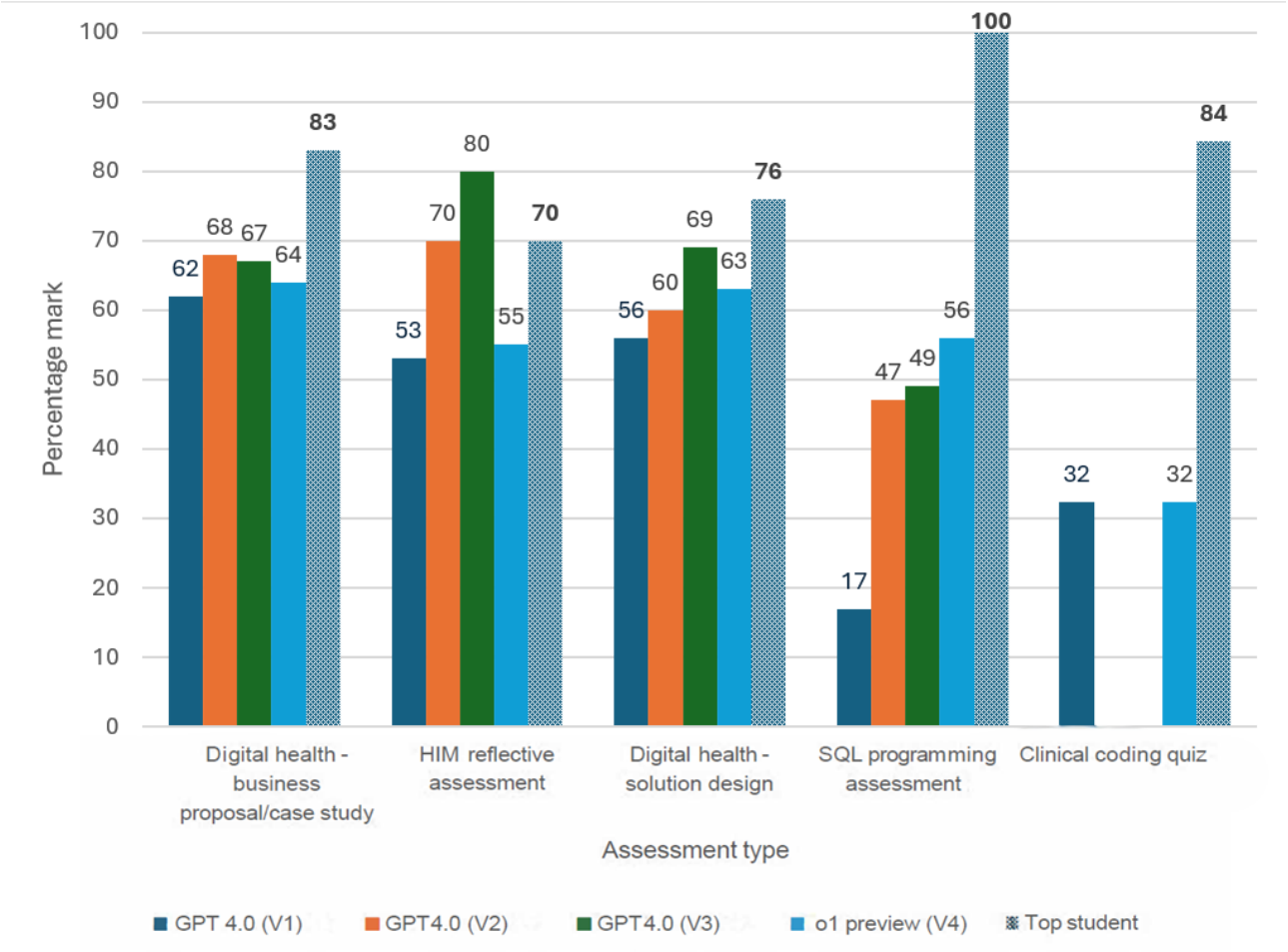
An overall performance comparison across different versions of ChatGPT responses, based on average scores, revealed noticeable variations in performance across assessments (Figure 2).

Descriptive and reflective assessments consistently achieved higher average scores in ChatGPT-generated responses, compared to technical assessments such as SQL programming and health classification quiz, which received substantially lower average scores. Lower averages in these tasks reflected the complexity of accurately interpreting technical prompts, adhering to database structures, and executing precise scripts. Progression from V1 to V4 in technical tasks demonstrated the importance of context, schema inclusion, and rubrics, which helped improve performance over iterations. However, even the best-performing versions did not match the precision and depth required for these tasks, underlining limitations in handling computational assessments.

XSL•FO
RenderX

Overall, V3 achieved the highest average scores across assessment types, benefiting from rubric-driven structuring and better alignment with task criteria. V2 showed notable success in descriptive tasks due to its structured approach, while V4 demonstrated improvements in technical assessments through contextual and schema support. In contrast, V1 consistently underperformed across most assessments, reflecting its reliance on single-step, unguided generation, which often lacked depth, accuracy, and alignment with task expectations.

Figure 2. Overall performance comparison across assessment types. V1: basic answer generation; V2: section-based refinement; V3: rubric-based refinement; V4: rubric-based refinement using o1 Preview model.

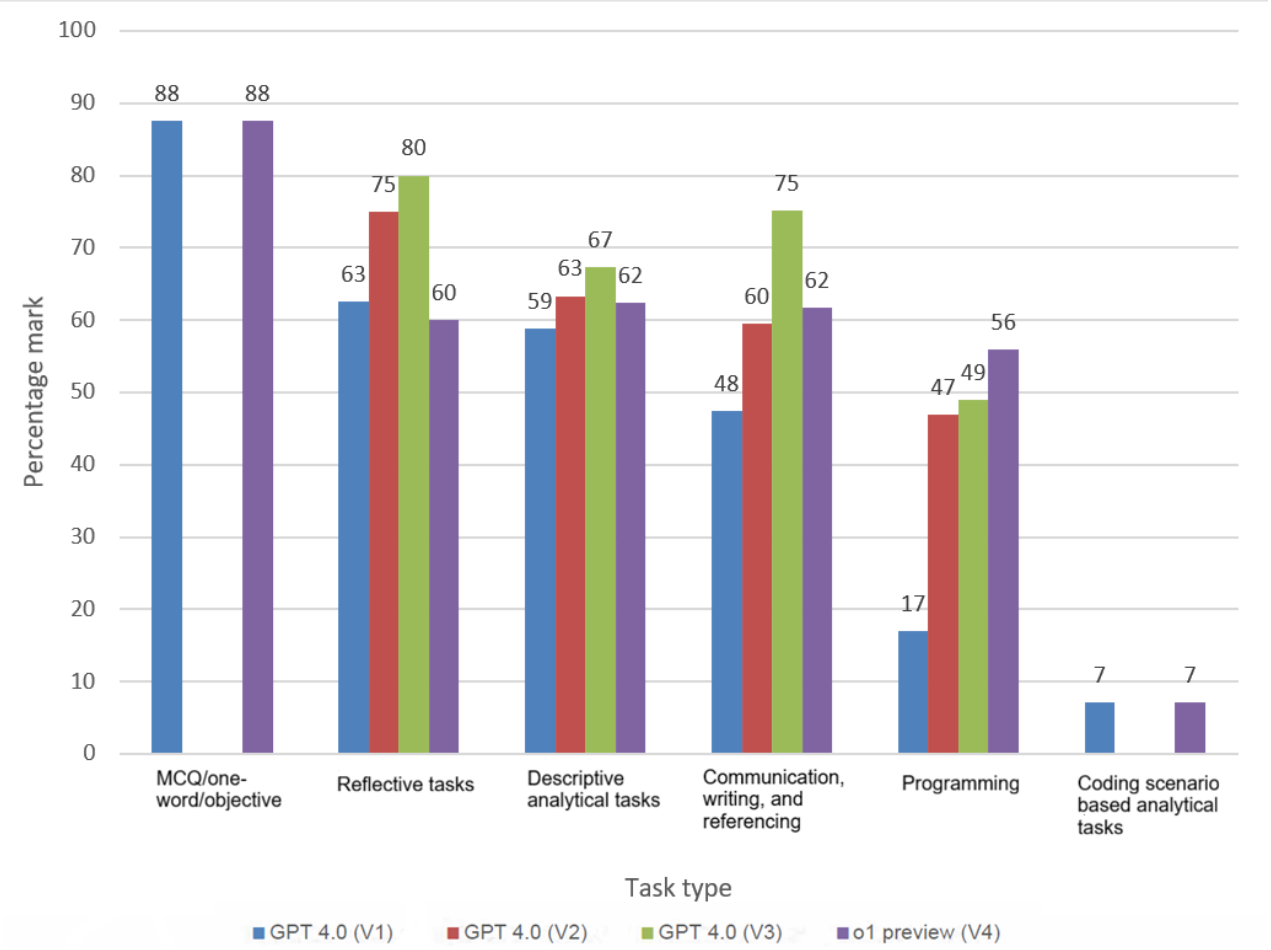


Comparative Analysis of ChatGPT Versions Across Task Types

Evaluating ChatGPT's performance across assessment task types, ChatGPT performed best on objective tasks, showing strong accuracy in factual, rule-based questions. Reflective tasks followed, with later versions demonstrating improved reasoning. Descriptive analytical tasks were next in terms of average performance, though gains plateaued in newer versions. Communication and referencing tasks also showed similar performance with a clearer structure in later iterations. Programming tasks improved over time but continued to face challenges in accuracy and contextual understanding. The lowest performance was in complex scenario-based health classification tasks, reflecting current limitations in contextual and analytical reasoning (Figure 3).

XSL•FO
RenderX

Figure 3. Overall performance comparison of ChatGPT versions across assessment task types/rubric criteria. V1: basic answer generation; V2: section-based refinement; V3: rubric-based refinement; V4: rubric-based refinement using o1 Preview model.



Discussion

Overview

This research aimed to pilot the evaluation of ChatGPT's performance across diverse DH and HIM assessment tasks, with a focus on promoting academic integrity. This analysis not only revealed the specific contexts in which ChatGPT excelled or fell short but also suggested broader implications for designing assessments that foster critical thinking, ethical practice, and real-world applicability. By situating ChatGPT's performance within the study's original goals, we were able to gain clearer insights into the potential and the limitations of GenAI in academic and professional environments. However, because this pilot examined only the final outputs generated by ChatGPT, it does not evaluate AI's role as a formative learning partner, such as when students iteratively critique, refine, or codevelop AI-generated drafts.

Principal Findings and Comparison With Prior Literature

The DH case study and solution design assessments revealed clear limitations in ChatGPT's capacity to address complex, context-specific problems. While AI-generated submissions consistently identified foundational frameworks and produced well-structured outputs, they lacked the depth and contextualization required to develop effective, patient-specific solutions. Weak alignment between identified gaps and proposed interventions reflects broader challenges in maintaining contextual coherence. This is consistent with prior findings across disciplines that report similar shortcomings despite fluent language, structural coherence, and high grammatical accuracy [16,44]. The absence of visual artifacts, such as patient journey maps or flowcharts, further constrained the communication of complex ideas, underscoring limitations in GenAI's ability to produce meaningful visual representations [45]. These findings reinforce the essential role of human expertise in refining and contextualizing solutions. For DIGHIM assessments, tasks should prioritize deep contextual engagement, patient-specific scenarios, and dynamic problem-solving. Requiring the integration of real-world data, graphical visualizations, and detailed case-based justification can better assess applied competence while reducing overreliance on AI tools. This approach aligns with constructivist learning theory, which emphasizes learning through authentic engagement rather than surface-level reproduction [46,47], and with higher-order cognitive processes in Bloom taxonomy that current GenAI systems cannot reliably replicate [48,49].

Reflective assessments demonstrated ChatGPT's strongest performance, with later versions, particularly GPT-4 (V3), achieving the highest score at 80% and surpassing student submissions in structure, coherence, and adherence to reflective frameworks. Similar patterns have been reported in dental

XSL•FO
RenderX

education, where ChatGPT performed strongly in structured-portfolio tasks [17]. However, despite polished language and systematic organization, AI-generated reflections often lacked genuine depth, contextual nuance, and personalized insight, which are central to HIM professional development. These results indicate a need to redesign reflective assessments to preserve authenticity. This aligns with social cognitive theory, which emphasizes authentic self-reflection, mastery experiences, and the development of self-efficacy as central to professional learning, all of which are elements that GenAI cannot meaningfully replicate [50]. Emphasizing personalized insights, unique contextual connections, and critical evaluation of lived experiences, alongside dynamic scenarios, peer interaction, or real-time contextual observations, can reduce AI dependence while reinforcing academic integrity and professional skill development [51,52].

Objective assessments, including multiple-choice and true or false questions, showed consistently high AI performance with minimal error rates. This reflects known strengths of AI in structured, rule-based tasks requiring factual recall and logical reasoning and aligns with prior evaluations of GenAI performance on objective assessments [53-56]. However, this raises concerns regarding the validity of such tasks in measuring individual learning, particularly in online contexts. To maintain assessment integrity, objective questions should be supplemented with higher-order components such as justification, explanation, or reasoning. Embedding real-world context and multistep problem-solving can further enhance authenticity and alignment with learning outcomes while limiting AI dominance [57].

Clinical coding scenario-based assessments revealed a clear performance gap between AI and students, particularly in tasks requiring nuanced contextual interpretation and application of coding rules. While AI performed comparably on objective components, it struggled with complex scenarios, resulting in significantly lower scores than high-performing students. This aligns with US-based findings showing that ChatGPT performs adequately for simple, single-diagnosis coding tasks but struggles with complex patient data [58,59]. Soroush et al [59] reported that GPT-4 achieved only a 34% exact match rate for *ICD-10-CM* (*International Classification of Diseases, Tenth Revision, Clinical Modification*)codes and concluded that LLMs lack a complete internal representation of medical coding rules, rendering them unsuitable for clinical coding tasks. These findings underscore the need for assessments that emphasize contextual reasoning and real-world coding application, reinforcing the importance of developing independent clinical coding proficiency.

SQL programming assessments highlighted further limitations in ChatGPT's ability to perform technically precise, context-dependent tasks. Although later versions showed incremental improvement, persistent issues included incorrect table and column usage, overreliance on prompts, limited understanding of health data conventions, and failure to execute complex multistep queries. Errors such as misspellings and the absence of real-time testing and debugging capabilities further undermined output reliability. These findings contrast with prior research reporting strong SQL generation performance by LLMs [60], suggesting that health database contexts demand deeper domain knowledge and contextual awareness. Previous work has emphasized the importance of iterative testing and contextual understanding in SQL code generation [61]. For educators, this highlights the value of assessments that extend beyond syntax to include debugging, real-world application, and analytical interpretation of health data, thereby fostering job-ready competencies that AI cannot fully replicate.

In communication, writing, and referencing tasks, ChatGPT demonstrated strong performance in structure, clarity, and adherence to academic conventions, particularly in later versions such as V3. Nonetheless, limitations in depth, contextual relevance, repetition, and critical engagement persisted, especially in longer writing tasks. Similar findings have been reported in business education, where AI-generated responses for extended academic writing scored between 40% and 77% due to insufficient depth and analysis despite high grammatical quality [16]. Issues such as hallucinated references and inconsistent citation formatting further undermined credibility, consistent with prior research [62,63]. These results indicate that while AI can support foundational academic writing, it struggles with nuanced argumentation and original critical thought. Assessment designs should therefore require critical engagement with sources, precise referencing, and personalized application of concepts to ensure authentic student learning.

Chain prompting, involving iterative refinement of prompts, significantly improved AI performance, particularly in descriptive and reflective tasks, with later versions benefiting from rubric-aligned guidance. Comparable improvements have been observed in pharmacy education, where structured prompting enhanced performance in knowledge recall tasks [56]. This demonstrates the effectiveness of multistep prompting in optimizing AI outputs for well-defined assessment formats.

Overall, o1 Preview (V4) underperformed in descriptive tasks relative to V3, although performance was comparable or improved in simpler objective assessments. This suggests that newer AI versions do not necessarily yield better outcomes for complex, context-driven tasks. While earlier studies reported GPT-4 superiority in medical licensing examinations [64] and basic sciences [65], such improvements were not consistently observed in this study. There was no clear evidence that o1 Preview enhanced performance in tasks requiring critical thinking or reflection. These findings highlight the importance of assessment designs that prioritize originality, contextual understanding, and higher-order cognition over tasks that can be easily completed using AI assistance.

Strengths and Limitations

The findings from this pilot study highlight the importance of designing assessments that align with the evolving roles in HIM and DH while fostering the ethical and practical use of GenAI. Health information managers play a crucial role in health data management, analytics, and health information and communication technology [33], and as DH increasingly integrates disciplines such as clinical care, data science, and IT [66], it is essential for educational assessments to reinforce interdisciplinary skills that align with the evolving demands of these domains [67,68]. Recent evidence highlights that

XSL•FO
RenderX

traditional assessment methods in higher education are increasingly ineffective in GenAI-facilitated learning environments, necessitating redesigned assessments that promote critical thinking, creativity, and lifelong learning skills [28]. For instance, practical redesign approaches, such as "flipped assessment," can be implemented, where students cocreate and then critically validate AI-generated items under educator supervision, improving confidence while maintaining expert oversight [31]. This type of redesign aligns with this study's recommendation for tasks that encourage students to critique, refine, and contextualize AI outputs rather than merely generate them.

Assessments should emphasize tasks that require contextual understanding, technical proficiency, and decision-making. Furthermore, these assessments should integrate AI as a tool to augment, rather than a substitute for critical thinking. For example, students could refine AI-generated clinical coding or SQL outputs to align with standards-based frameworks, critically evaluate AI-generated business proposals for depth and contextual relevance, or integrate visualizations such as patient journey maps into health information reports. These approaches not only build essential competencies such as critical analysis, programming, and communication but also encourage students to navigate AI's limitations and responsibly incorporate its outputs into their work, promoting the ethical use of GenAI to support, rather than replace, student efforts.

In DIGHIM education, AI use in assessments should be carefully structured to align with professional competencies while maintaining academic integrity. Some tasks should avoid reliance on AI, particularly those requiring personal reflections, ethical decision-making, and critical thinking in complex DH scenarios. Other tasks can adopt AI for efficiency, such as assisting in summarizing health policies, generating structured reports, or organizing data-driven insights. Finally, AI-generated content can be adapted in assessments that require students to critically refine, validate, and contextualize AI outputs, such as evaluating AI-generated clinical coding recommendations, refining SQL queries for health databases, or assessing the applicability of AI-generated DH solutions within industry frameworks. Additionally, educators and students must ensure privacy by refraining from supplying GenAI tools with copyrighted or sensitive health data and should receive structured training on the responsible and ethical use of AI in DH and HIM to support professional readiness.

Taken together, current 2024-2025 evidence converges on a dual imperative: redesign assessments to cultivate critique, verification, and reflection skills, and establish clear governance mechanisms, including policy guidance, training programs, and disclosure norms, proportionate to the evolving capabilities and risks of GenAI systems [31,32,38].

These approaches prepare students for real-world applications of GenAI in HIM and DH roles, where such tools may be used to assist with technical tasks such as data analysis, report generation, and strategic planning. By designing assessments that both test and build skills in these areas while fostering responsible AI use, educators can equip graduates to navigate the interdisciplinary and technology-driven landscape of modern DH care effectively.

Table 8 provides a summary of important recommendations.

This study had several limitations. First, as a pilot study, the findings are preliminary and intended to provide directional insights rather than definitive conclusions; however, the use of multiple assessment types and structured AI submission versions was intended to strengthen internal comparison. Larger-scale studies are needed to confirm these patterns. Second, the quasi-experimental design lacked random assignment, which may limit generalizability. This was partially mitigated by comparing AI-generated submissions with authentic, previously graded student work under blinded conditions. Future studies could strengthen external validity through multiinstitutional or randomized designs. Third, assessments were marked by discipline-specific academic staff, and individual grading variability may have influenced results despite mitigation strategies, including standardized rubrics, blinded review, and second-marker consensus when remarked scores differed by more than 10 marks. Future work could further enhance reliability by incorporating multiple independent markers per assessment. Fourth, the number of assessment samples within each category was limited, which may reduce the robustness of task-specific conclusions. This was addressed by selecting assessments across multiple levels and disciplines, though larger samples are required in future studies. Fifth, the study could not fully replicate real-time student interactions with GenAI, such as iterative prompting or blended human-AI workflows. The structured versioning approach (V1-V4) was used to approximate common usage patterns, but longitudinal studies are needed to capture authentic use over time. Sixth, while all student submissions underwent formal academic integrity checks and no concerns were identified, it was not possible to determine with absolute certainty whether any student work involved AI assistance, which may have influenced comparisons. Finally, the study did not systematically assess the detectability of AI-generated content. Although markers occasionally inferred AI authorship, these judgments were informal and sometimes inaccurate, consistent with prior reports of false positives [16,30,69] and highlighting the unreliability of human detection methods.

Future research should explicitly investigate the accuracy of markers in distinguishing AI- from student-generated work and consider how assessment design can emphasize originality, critical thinking, and personalized insights to reduce reliance on detection and strengthen academic integrity. Another area that future studies could explore is expanding the scope beyond ChatGPT to include other advanced LLMs (eg, Gemini, Claude, and AI-assisted coding tools such as GitHub Copilot), to allow broader comparison of capabilities and provide a more comprehensive understanding of how different AI systems perform across assessment types in DIGHIM.

XSL•FO
RenderX

Table 8. Recommendations for ethical assessment design in Digital Health and Health Information Management in the artificial intelligence (AI) era.

Theme	Key findings	Recommendations
Contextual and higher-order assessment design	• AI struggled with context-specific, patient-centered, and judgement-based tasks. • Underperformance in descriptive, analytical, and coding scenario questions.	• Use assessments requiring deep contextualization, scenario complexity, and justification of decisions. • Emphasize ethical reasoning, critique, and higher-order application rather than factual recall.
Reflective and authentic assessments	• AI produced a strong structure but fabricated plausible experiences. • Risk to authenticity in reflective tasks.	• Require personalized and context-verified reflections (eg, workplace events, supervisor-signed logs). • Use scenario-based reflections, peer engagement, or real-time journaling to ensure authenticity.
Objective and procedural tasks	• AI performed strongly on MCQs[a], true/false, and structured objective items. • Risk of undermining originality.	• Add reasoning, explanation, or linked vignettes to objective tasks. • Combine correctness with justification to assess understanding.
Technical and programming tasks (SQL/coding)	• AI struggled with schema use, debugging, and domain conventions. • Frequent inaccuracies despite rubric/schema input.	• Use live, hands-on coding tasks requiring testing, debugging, and scenario-based logic. • Assess understanding of health-data structures and domain-specific conventions.
Academic writing and referencing	• Strong structure but limited depth and contextual argumentation	• Require deeper analysis, original reasoning, and evidence validation • Include tasks where students examine or verify AI-generated claims.
Ethics, equity, privacy, and governance	• Prompting skill differences create inequities • AI detection unreliable • Privacy and rubric transparency concerns.	• Provide standardized AI literacy training • Avoid overreliance on detection tools; design for originality and personal insight • Train students not to enter sensitive data into AI tools. • Use rubrics that guide learning without enabling “AI gaming.” • Develop institutional policies for ethical AI use.

[a]MCQ: multiple-choice question.

Conclusions

This pilot study has highlighted the potential and limitations of ChatGPT in HIM and DH assessments. While GenAI demonstrates strengths in structured tasks and foundational content generation, it struggles with contextualization, depth, and the technical precision required for assessments requiring independent critical thinking. These findings emphasize the need for carefully designed assessments that integrate AI ethically and prioritize tasks that require human judgment and contextual understanding to ensure meaningful learning outcomes and academic integrity.

Acknowledgments

The authors sincerely thank Sheikh Hanan Bashir, Saja Sammour, and Kay Bonello for their valuable contributions to this study. The authors also acknowledge the staff of the Digital Health and Health Information Management cluster, the participating markers, and the academic leadership of the School of Psychology and Public Health at La Trobe University for their support.

Generative artificial intelligence tools such as Microsoft Copilot were used for limited language editing and refinement of the manuscript. All content was reviewed, verified, and finalized by the authors, who take full responsibility for the accuracy, originality, and integrity of the work.

Funding

The research was funded through La Trobe University’s internal artificial intelligence in teaching grant scheme. The funder had no role in the study design; data collection, analysis, or interpretation; manuscript preparation; or the decision to submit the manuscript for publication.

Data Availability

Summary data and detailed analyses supporting this study are provided in the manuscript and Multimedia Appendix 2. Access to underlying deidentified student assessment data is restricted due to ethical and privacy considerations; however, reasonable requests for additional deidentified data may be considered by the corresponding author, subject to institutional ethics approval.

XSL•FO
RenderX

Authors' Contributions

Conceptualization: TAW, ML, NP, MR
Data curation: TAW, ML, NP, AN, MR
Formal analysis: TAW, KR, AN, MT, SG, URK, JB, MR
Investigation: TAW, ML, NP, KR, AN, MT, SG
Methodology: TAW, ML, NP, MR
Validation: TAW, ML, MT, SG, URK, JB, MR
Project administration: TAW, ML, NP, MR
Resources: TAW, ML, NP, MR
Supervision: TAW, KR, MR
Visualization: TAW
Writing—original draft: TAW
Writing—review & editing: ML, NP, KR, AN, MT, SG, URK, JB, MR

Conflicts of Interest

None declared.

Multimedia Appendix 1
Protocol for research assistant.
[DOCX File , 356 KB - mededu_v12i1e82988_app1.docx]

Multimedia Appendix 2
Detailed performance analysis by assessment type.
[XLSX File (Microsoft Excel File), 84 KB - mededu_v12i1e82988_app2.xlsx]

References

1. Makridakis S. The forthcoming artificial intelligence (AI) revolution: its impact on society and firms. Futures 2017;90:46-60. [doi: 10.1016/j.futures.2017.03.006]
2. ISO/IEC 22989:2022(en), Information technology -- Artificial intelligence concepts and terminology. ISO. 2022. URL: https://www.iso.org/obp/ui/#iso:std:iso-iec:22989:ed-1:v1:en [accessed 2025-05-16]
3. Marr B. The 4 types Of generative AI transforming our world. Forbes. URL: https://www.forbes.com/sites/bernardmarr/2024/04/29/the-4-types-of-generative-ai-transforming-our-world/ [accessed 2025-05-16]
4. Hasanein AM, Sobaih AEE. Drivers and consequences of ChatGPT use in higher education: key stakeholder perspectives. Eur J Investig Health Psychol Educ 2023;13(11):2599-2614 [FREE Full text] [doi: 10.3390/ejihpe13110181] [Medline: 37998071]
5. Kasneci E, Sessler K, Küchemann S, Bannert M, Dementieva D, Fischer F, et al. ChatGPT for good? On opportunities and challenges of large language models for education. Learn Individ Differ 2023;103:102274. [doi: 10.1016/j.lindif.2023.102274]
6. Nazi ZA, Peng W. Large language models in healthcare and medical domain: a review. Informatics 2024;11(3):57. [doi: 10.3390/informatics11030057]
7. Gao CA, Howard FM, Markov NS, Dyer EC, Ramesh S, Luo Y, et al. Comparing scientific abstracts generated by ChatGPT to real abstracts with detectors and blinded human reviewers. NPJ Digit Med 2023;6(1):75 [FREE Full text] [doi: 10.1038/s41746-023-00819-6] [Medline: 37100871]
8. Lund BD, Wang T, Mannuru NR, Nie B, Shimray S, Wang Z. ChatGPT and a new academic reality: artificial Intelligence-written research papers and the ethics of the large language models in scholarly publishing. Asso Info Science Tech 2023;74(5):570-581. [doi: 10.1002/asi.24750]
9. Lin Z. Why and how to embrace AI such as ChatGPT in your academic life. R Soc Open Sci 2023;10(8):230658 [FREE Full text] [doi: 10.1098/rsos.230658] [Medline: 37621662]
10. Deng J, Lin Y. The benefits and challenges of ChatGPT: an overview. FCIS 2023;2(2):81-83. [doi: 10.54097/fcis.v2i2.4465]
11. Pradana M, Elisa HP, Syarifuddin S. Discussing ChatGPT in education: a literature review and bibliometric analysis. Cogent Educ 2023;10(2):2243134. [doi: 10.1080/2331186x.2023.2243134]
12. Baig MI, Yadegaridehkordi E. ChatGPT in the higher education: a systematic literature review and research challenges. Int J Educ Res 2024;127:102411. [doi: 10.1016/j.ijer.2024.102411]
13. Rasul T, Nair S, Kalendra D, Robin M, Santini FDO, Ladeira WJ. The role of ChatGPT in higher education: benefits, challenges, and future research directions. JALT 2023;6(1):41-56. [doi: 10.37074/jalt.2023.6.1.29]
14. Rawas S. ChatGPT: empowering lifelong learning in the digital age of higher education. Educ Inf Technol 2023;29(6):6895-6908. [doi: 10.1007/s10639-023-12114-8]

XSL•FO
RenderX

15. Nikolic S, Daniel, Haque R, Belkina M, Hassan GM, Grundy S, et al. ChatGPT versus engineering education assessment: a multidisciplinary and multi-institutional benchmarking and analysis of this generative artificial intelligence tool to investigate assessment integrity. Eur J Eng Educ 2023;48(4):559-614. [doi: 10.1080/03043797.2023.2213169]
16. Chaudhry IS, Sarwary SAM, El Refae GA, Chabchoub H. Time to revisit existing student's performance evaluation approach in higher education sector in a new era of ChatGPT — a case study. Cogent Educ 2023;10(1):2210461. [doi: 10.1080/2331186x.2023.2210461]
17. Ali K, Barhom N, Tamimi F, Duggal M. ChatGPT-A double-edged sword for healthcare education? Implications for assessments of dental students. Eur J Dent Educ 2024;28(1):206-211. [doi: 10.1111/eje.12937] [Medline: 37550893]
18. Hadi Mogavi R, Deng C, Juho Kim J, Zhou P, D. Kwon Y, Hosny Saleh Metwally A, et al. ChatGPT in education: a blessing or a curse? A qualitative study exploring early adopters' utilization and perceptions. Comput Hum Behav Artif Hum 2024;2(1):100027. [doi: 10.1016/j.chbah.2023.100027]
19. Sallam M, Salim N, Barakat M, Al-Tammemi A. ChatGPT applications in medical, dental, pharmacy, and public health education: a descriptive study highlighting the advantages and limitations. Narra J 2023;3(1):e103 [FREE Full text] [doi: 10.52225/narra.v3i1.103] [Medline: 38450035]
20. Akiba D, Fraboni MC. AI-supported academic advising: exploring ChatGPT's current state and future potential toward student empowerment. Educ Sci 2023;13(9):885. [doi: 10.3390/educsci13090885]
21. Lin Y, Luo Z, Ye Z, Zhong N, Zhao L, Zhang L, et al. Applications, challenges, and prospects of generative artificial intelligence empowering medical education: scoping review. JMIR Med Educ 2025;11:e71125 [FREE Full text] [doi: 10.2196/71125] [Medline: 41128430]
22. Janumpally R, Nanua S, Ngo A, Youens K. Generative artificial intelligence in graduate medical education. Front Med (Lausanne) 2024;11:1525604 [FREE Full text] [doi: 10.3389/fmed.2024.1525604] [Medline: 39867924]
23. Weightman MJ, Chur-Hansen A, Clark SR. AI in psychiatric education and training from 2016 to 2024: scoping review of trends. JMIR Med Educ 2025;11:e81517 [FREE Full text] [doi: 10.2196/81517] [Medline: 41474738]
24. Farhud DD, Zokaei S. Ethical issues of artificial intelligence in medicine and healthcare. Iran J Public Health 2021;50(11):i-v [FREE Full text] [doi: 10.18502/ijph.v50i11.7600] [Medline: 35223619]
25. Rahman MM, Watanobe Y. ChatGPT for education and research: opportunities, threats, and strategies. Appl Sci 2023;13(9):5783. [doi: 10.3390/app13095783]
26. Neumann M, Rauschenberger M, Schön EM. "We need to talk about ChatGPT": the future of AI and higher education. : IEEE; 2023 Presented at: 2023 IEEE/ACM 5th International Workshop on Software Engineering Education for the Next Generation (SEENG); 2023 May 16; Melbourne, Australia. [doi: 10.1109/seeng59157.2023.00010]
27. AlAfnan MA, Dishari S, Jovic M, Lomidze K. ChatGPT as an educational tool: opportunities, challenges, and recommendations for communication, business writing, and composition courses. JAIT 2023;3:60-68. [doi: 10.37965/jait.2023.0184]
28. Weng X, XIA Q, Gu M, Rajaram K, Chiu TKF. Assessment and learning outcomes for generative AI in higher education: a scoping review on current research status and trends. AJET 2024;40(6):37-55. [doi: 10.14742/ajet.9540]
29. Foltynek T, Bjelobaba S, Glendinning I, Khan ZR, Santos R, Pavletic P, et al. ENAI recommendations on the ethical use of artificial intelligence in education. Int J Educ Integr 2023;19(1):1-4. [doi: 10.1007/s40979-023-00133-4]
30. Elkhatat AM, Elsaid K, Almeer S. Evaluating the efficacy of AI content detection tools in differentiating between human and AI-generated text. Int J Educ Integr 2023;19(1):1-16. [doi: 10.1007/s40979-023-00140-5]
31. Birks S, Gray J, Darling-Pomranz C. Using artificial intelligence to provide a 'flipped assessment' approach to medical education learning opportunities. Med Teach 2025;47(8):1377-1384 [FREE Full text] [doi: 10.1080/0142159X.2024.2434101] [Medline: 39616548]
32. Ichikawa T, Olsen E, Vinod A, Glenn N, Hanna K, Lund GC, et al. Generative artificial intelligence in medical education-policies and training at us osteopathic medical schools: descriptive cross-sectional survey. JMIR Med Educ 2025;11:e58766 [FREE Full text] [doi: 10.2196/58766] [Medline: 39934984]
33. Gjorgioski S, Riley M, Lee J, Prasad N, Tassos M, Nexhip A, et al. Workforce survey of Australian health information management graduates, 2017-2021: a 5-year follow-on study. Health Inf Manag 2025;54(1):43-54 [FREE Full text] [doi: 10.1177/18333583231197936] [Medline: 37753774]
34. Riley M, Robinson K, Prasad N, Gleeson B, Barker E, Wollersheim D, et al. Workforce survey of Australian graduate health information managers: employability, employment, and knowledge and skills used in the workplace. Health Inf Manag 2020;49(2-3):88-98. [doi: 10.1177/1833358319839296] [Medline: 31006266]
35. Temsah MH, Aljamaan F, Malki KH, Alhasan K, Altamimi I, Aljarbou R, et al. ChatGPT and the future of digital health: a study on healthcare workers' perceptions and expectations. Healthcare (Basel) 2023;11(13):1812 [FREE Full text] [doi: 10.3390/healthcare11131812] [Medline: 37444647]
36. Tudor Car L, Kyaw BM, Nannan Panday RS, van der Kleij R, Chavannes N, Majeed A, et al. Digital health training programs for medical students: scoping review. JMIR Med Educ 2021;7(3):e28275 [FREE Full text] [doi: 10.2196/28275] [Medline: 34287206]

XSL•FO
RenderX

37. Yu P, Fang C, Liu X, Fu W, Ling J, Yan Z, et al. Performance of ChatGPT on the Chinese postgraduate examination for clinical medicine: survey study. JMIR Med Educ 2024;10:e48514 [FREE Full text] [doi: 10.2196/48514] [Medline: 38335017]
38. Lee QY, Chen M, Ong CW, Ho CSH. The role of generative artificial intelligence in psychiatric education- a scoping review. BMC Med Educ 2025;25(1):438 [FREE Full text] [doi: 10.1186/s12909-025-07026-9] [Medline: 40133891]
39. Cherif H, Moussa C, Missaoui AM, Salouage I, Mokaddem S, Dhahri B. Appraisal of ChatGPT's aptitude for medical education: comparative analysis with third-year medical students in a pulmonology examination. JMIR Med Educ 2024;10:e52818 [FREE Full text] [doi: 10.2196/52818] [Medline: 39042876]
40. Eysenbach G. The role of ChatGPT, generative language models, and artificial intelligence in medical education: a conversation with ChatGPT and a call for papers. JMIR Med Educ 2023;9:e46885 [FREE Full text] [doi: 10.2196/46885] [Medline: 36863937]
41. Meo SA, Al-Masri AA, Alotaibi M, Meo MZS, Meo MOS. ChatGPT knowledge evaluation in basic and clinical medical sciences: multiple choice question examination-based performance. Healthcare (Basel) 2023;11(14):2046 [FREE Full text] [doi: 10.3390/healthcare11142046] [Medline: 37510487]
42. Benito P, Isla-Jover M, González-Castro P, Fernández Esparcia PJ, Carpio M, Blay-Simón I, et al. GPT-4o and openAI o1 performance on the 2024 Spanish competitive medical specialty access examination: cross-sectional quantitative evaluation study. JMIR Med Educ 2026;12:e75452 [FREE Full text] [doi: 10.2196/75452] [Medline: 41525685]
43. Wani TA, Liem M, Boyd J, Wijesooriya H, Khan UR. Digital health education in Australian universities: trends, gaps, and future directions. Int J Med Inform 2026;205:106105 [FREE Full text] [doi: 10.1016/j.ijmedinf.2025.106105] [Medline: 40916275]
44. Kolade O, Owoseni A, Egbetokun A. Is AI changing learning and assessment as we know it? Evidence from a ChatGPT experiment and a conceptual framework. Heliyon 2024;10(4):e25953 [FREE Full text] [doi: 10.1016/j.heliyon.2024.e25953] [Medline: 38379960]
45. Moin KA, Nasir AA, Petroff DJ, Loveless BA, Moshirfar OA, Hoopes PC, et al. Assessment of generative artificial intelligence (AI) models in creating medical illustrations for various corneal transplant procedures. Cureus 2024;16(8):e67833. [doi: 10.7759/cureus.67833] [Medline: 39328681]
46. Vygotsky LS. Mind in Society: The Development of Higher Psychological Processes. Cambridge, Massachusetts: Harvard University Press; 1978.
47. Tran M, Balasooriya C, Semmler C, Rhee J. Generative artificial intelligence: the 'more knowledgeable other' in a social constructivist framework of medical education. NPJ Digit Med 2025;8(1):430 [FREE Full text] [doi: 10.1038/s41746-025-01823-8] [Medline: 40646156]
48. Anderson LW, Krathwohl DR. A taxonomy for learning, teaching, and assessing: a revision of Bloom's taxonomy of educational objectives: complete edition. EDUQ.info.: Addison Wesley Longman; 2001. URL: https://eduq.info/xmlui/handle/11515/18824 [accessed 2025-11-14]
49. Bloom BS, Engelhart MD, Furst EJ, Hill WH, Krathwohl DR. Taxonomy of Educational Objectives. New York: Longmans, Green; 1964.
50. Bandura A. Social Foundations of Thought and Action. Englewood Cliffs NJ: Prentice Hall; 1986.
51. Powell S, Forsyth R. Generative AI and the implications for authentic assessment. In: Using Gener AI Eff High Educ. New York: Routledge; 2024.
52. Salinas-Navarro DE, Vilalta-Perdomo E, Michel-Villarreal R, Montesinos L. Designing experiential learning activities with generative artificial intelligence tools for authentic assessment. ITSE 2024;21(4):708-734. [doi: 10.1108/itse-12-2023-0236]
53. Hersh W, Fultz Hollis K. Results and implications for generative AI in a large introductory biomedical and health informatics course. NPJ Digit Med 2024;7(1):247 [FREE Full text] [doi: 10.1038/s41746-024-01251-0] [Medline: 39271955]
54. Morjaria L, Burns L, Bracken K, Ngo QN, Lee M, Levinson AJ, et al. Examining the threat of ChatGPT to the validity of short answer assessments in an undergraduate medical program. J Med Educ Curric Dev 2023;10:23821205231204178 [FREE Full text] [doi: 10.1177/23821205231204178] [Medline: 37780034]
55. Newton P, Xiromeriti M. ChatGPT performance on multiple choice question examinations in higher education. A pragmatic scoping review. Assess Eval High Educ 2024;49(6):781-798. [doi: 10.1080/02602938.2023.2299059]
56. Yang H, Hu M, Most A, Hawkins WA, Murray B, Smith SE, et al. Evaluating accuracy and reproducibility of large language model performance on critical care assessments in pharmacy education. Front Artif Intell 2024;7:1514896 [FREE Full text] [doi: 10.3389/frai.2024.1514896] [Medline: 39850846]
57. Nguyen Thanh B, Vo DTH, Nguyen Nhat M, Pham TTT, Thai Trung H, Ha Xuan S. Race with the machines: assessing the capability of generative AI in solving authentic assessments. AJET 2023;39(5):59-81. [doi: 10.14742/ajet.8902]
58. Abdelgadir Y, Thongprayoon C, Miao J, Suppadungsuk S, Pham JH, Mao MA, et al. AI integration in nephrology: evaluating ChatGPT for accurate ICD-10 documentation and coding. Front Artif Intell 2024;7:1457586 [FREE Full text] [doi: 10.3389/frai.2024.1457586] [Medline: 39286549]
59. Soroush A, Glicksberg BS, Zimlichman E, Barash Y, Freeman R, Charney AW, et al. Large language models are poor medical coders — benchmarking of medical code querying. NEJM AI 2024;1(5):AIdbp2300040. [doi: 10.1056/aidbp2300040]

XSL•FO
RenderX

60. Pornphol P, Chittayasothorn S. Verification of relational database languages codes generated by ChatGPT. : Association for Computing Machinery; 2024 Presented at: ASSE '23: Proceedings of the 2023 4th Asia Service Sciences and Software Engineering Conference; 2023 October 27 - 29; Aizu-Wakamatsu City Japan p. 17-22. [doi: 10.1145/3634814.3634817]
61. Carr N, Shawon FR, Jamil HM. An experiment on leveraging ChatGPT for online teaching and assessment of database students. 2023 Presented at: 2023 IEEE Int Conf Teach Assess Learn Eng TALE; 2023 November 27 – 1 December; Auckland, New Zealand p. 1-8. [doi: 10.1109/tale56641.2023.10398239]
62. Walters WH, Wilder EI. Fabrication and errors in the bibliographic citations generated by ChatGPT. Sci Rep 2023;13(1):14045 [FREE Full text] [doi: 10.1038/s41598-023-41032-5] [Medline: 37679503]
63. Williams A. Comparison of generative AI performance on undergraduate and postgraduate written assessments in the biomedical sciences. Int J Educ Technol High Educ 2024;21(1):52. [doi: 10.1186/s41239-024-00485-y]
64. Liu M, Okuhara T, Chang X, Shirabe R, Nishiie Y, Okada H, et al. Performance of ChatGPT across different versions in medical licensing examinations worldwide: systematic review and meta-analysis. J Med Internet Res 2024;26:e60807 [FREE Full text] [doi: 10.2196/60807] [Medline: 39052324]
65. Jones N. 'In awe': scientists impressed by latest ChatGPT model o1. Nature 2024;634(8033):275-276. [doi: 10.1038/d41586-024-03169-9] [Medline: 39354139]
66. Meehan R. Health informatics workforce in the digital health ecosystem. Stud Health Technol Inform 2024;310:1226-1230. [doi: 10.3233/SHTI231160] [Medline: 38270010]
67. Butler-Henderson K, Gray K, Arabi S. Roles and responsibilities of the global specialist digital health workforce: analysis of global census data. JMIR Med Educ 2024;10:e54137 [FREE Full text] [doi: 10.2196/54137] [Medline: 39118468]
68. Keep M, Janssen A, McGregor D, Brunner M, Baysari MT, Quinn D, et al. Mapping eHealth education: review of eHealth content in health and medical degrees at a metropolitan tertiary institute in Australia. JMIR Med Educ 2021;7(3):e16440 [FREE Full text] [doi: 10.2196/16440] [Medline: 34420920]
69. Khalil M, Er E. Will ChatGPT get you caught? Rethinking of plagiarism detection. Cham: Springer Nature Switzerland; 2023 Presented at: 10th International Conference, LCT 2023, Held as Part of the 25th HCI International Conference, HCII 2023; 2023 July 23–28; Copenhagen, Denmark p. 475-487. [doi: 10.35542/osf.io/fnh48]

Abbreviations

AI: artificial intelligence
DH: digital health
DIGHIM: Digital Health and Health Information Management
GenAI: generative artificial intelligence
HIM: health information management
ICD-10-AM: International Classification of Diseases, Tenth Revision, Australian Modification
ICD-10-CM: International Classification of Diseases, Tenth Revision, Clinical Modification
LLM: large language model
RA: research assistant
V1: basic answer generation
V2: section-based refinement
V3: rubric-based refinement
V4: rubric-based refinement using o1 Preview model

Edited by T Leung, J Eriksen; submitted 26.Aug.2025; peer-reviewed by J Li, A Azizi; comments to author 27.Oct.2025; revised version received 21.Jan.2026; accepted 31.Jan.2026; published 30.Mar.2026.

Please cite as:
Wani TA, Liem M, Prasad N, Robinson K, Nexhip A, Tassos M, Gjorgioski S, Khan UR, Boyd J, Riley M
Susceptibility of Assessment Types to AI-Generated Content in Digital Health and Health Information Management Education: Quasi-Experimental Pilot Study
JMIR Med Educ 2026;12:e82988
URL: https://mededu.jmir.org/2026/1/e82988
doi:10.2196/82988
PMID:

©Tafheem Ahmad Wani, Michael Liem, Natasha Prasad, Kerin Robinson, Abbey Nexhip, Melanie Tassos, Stephanie Gjorgioski, Urooj Raza Khan, James Boyd, Merilyn Riley. Originally published in JMIR Medical Education (https://mededu.jmir.org), 30.Mar.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License

XSL•FO
RenderX

(https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Education, is properly cited. The complete bibliographic information, a link to the original publication on https://mededu.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Faculty Perspectives on AI Integration in Anatomy Education in the United Arab Emirates: Cross-Sectional Survey

Prince Last Mudenda Zilundu[1], BSc, MSc, PhD; Pedzisai Mazengenya[1], BSc, MSc, PhD; Jayaraj Kodangattil Narayanan[2], BSc, MSci, PhD; LiHua Zhou[3], MD, PhD

[1]Basic Medical Sciences, Ajman University, Ajman, United Arab Emirates

[2]Basic Medical and Dental Sciences, Ajman University, Ajman, United Arab Emirates

[3]Sun Yat-sen University, Shenzhen, China

Corresponding Author:
Prince Last Mudenda Zilundu, BSc, MSc, PhD
Basic Medical Sciences
Ajman University
P O Box 346
Ajman
United Arab Emirates
Phone: 1 971554804132
Email: p.zilundu@ajman.ac.ae

Abstract

Background: Artificial intelligence (AI) is reshaping medical and health professions education; yet, adoption in anatomy remains uneven and often ad hoc. Anatomy's spatial and visualization demands make it a compelling domain for AI, but discipline-specific opportunities and risks are not well characterized in the United Arab Emirates.

Objective: This study examines United Arab Emirates anatomy educators' AI use, attitudes, perceived barriers and enablers, and strategic perspectives on AI integration using a design informed by the Unified Theory of Acceptance and Use of Technology 2 (UTAUT2).

Methods: A cross-sectional survey of anatomy faculty at United Arab Emirates medical and health sciences colleges used 5-point Likert items to assess educational technology proficiency, AI use patterns, AI attitudes, perceived barriers and facilitators, and professional development needs. Quantitative data were summarized descriptively and explored with nonparametric tests. Open-ended strengths, weaknesses, opportunities, and threats questions were analyzed using reflexive thematic analysis, organized within the strengths, weaknesses, opportunities, and threats framework, and interpreted through UTAUT2 constructs. Quantitative and qualitative strands were integrated at interpretation through triangulation.

Results: In total, 30 anatomy faculty participated. Self-rated educational technology proficiency was high (mean 3.73 out of 5, SD 1.01), and overall attitudes toward AI in anatomy education were positive (mean 4.23, SD 0.73), with strong interest in AI-focused professional development (mean 4.50, SD 0.73). Most respondents reported using generative AI tools, predominantly ChatGPT, for content creation, quiz and examination item generation, summarization of complex material, and, to a lesser extent, visualization and workflow streamlining. Capacity-related barriers predominated: time and workload pressures (mean 3.27, SD 1.17) and training gaps (mean 3.13, SD 1.22) were rated as moderate obstacles, whereas budget or resource limitations (mean 2.63, SD 1.19) and academic integrity concerns (mean 2.80, SD 1.10) were minor obstacles. Student interest (mean 4.23, SD 0.86) and institutional encouragement (mean 4.00, SD 1.14) emerged as strong facilitators, with no statistically detectable differences by academic rank, age, or years of experience in this small, underpowered sample. Qualitatively, themes highlighted strong institutional support and digital readiness as strengths; training needs, workload, and policy gaps as weaknesses; visualization, personalization, and efficiency as opportunities; and overreliance, ethical risks, and erosion of hands-on anatomy pedagogy as threats. UTAUT2 interpretation indicated high performance expectancy and social influence (student and institutional support) but reduced effort expectancy and facilitating conditions due to time, training, and governance constraints, collectively tempering behavioral intention.

Conclusions: In this exploratory sample, United Arab Emirates anatomy educators were broadly receptive to generative AI and already experimenting and valuing the benefits for 3D visualization, adaptive practice, and feedback. However, workload, limited

XSL•FO
RenderX

training, and unclear governance (disclosure, assessment integrity, and cadaveric or patient images) constrain uptake, underscoring the need for protected time, workflow-aligned training, and discipline-specific policies to enable sustainable, ethical integration.

(JMIR Med Educ 2026;12:e87418) doi:10.2196/87418

KEYWORDS

artificial intelligence; anatomy education; faculty perspectives; SWOT; strengths, weaknesses, opportunities, and threats; medical education; United Arab Emirates; professional development; policy; assessment; extended reality; 3D visualization

Introduction

Artificial intelligence (AI) is rapidly transforming medical and health professions education. For instance, intelligent tutoring systems and adaptive learning platforms can diagnose knowledge gaps and deliver tailored remediation, supporting self-directed and mastery-based learning [1]. AI-driven simulation is enhancing experiential training through virtual and augmented reality environments as well as synthetic patient encounters, allowing learners to practice decision-making safely and repeatedly [2]. Assessment has similarly been reshaped by automated scoring, pattern recognition, and feedback algorithms that provide rapid, personalized performance insights [3,4]. These innovations extend beyond content delivery to student engagement, where conversational agents, game-based adaptive systems, and learning analytics have improved motivation and interactivity [5-7]. Yet, the educational impact of these tools ultimately depends on educators' willingness to adopt and integrate them into curricula.

Although much popular attention has focused on large generative models such as ChatGPT, the AI landscape in medical education is broader. Natural language processing, image analysis, and intelligent search underpin a growing range of educational tools from anatomy image annotation and question generation to personalized study schedules [8]. Global surveys show that while students adopt AI quickly, faculty uptake is uneven; educators express both interest and uncertainty about safe and pedagogically sound integration [9-11]. Recent scoping reviews and analyses emphasize that generative artificial intelligence (GenAI) adoption in formal medical curricula remains early and often unstructured [12-14]. Considering these patterns, this study contributes discipline-specific, United Arab Emirates–based evidence by detailing how anatomy educators use as well as perceive AI's benefits, risks, and practical enablers and by indicating where structured support is most needed next.

Anatomy, a fundamental subject in medical and health professions education, is particularly well-suited for AI-supported innovation. The subject's large knowledge base, reliance on spatial reasoning, and traditional use of cadaveric resources create long-standing challenges for visualization and cognitive load [15]. AI can augment 3D understanding through advanced segmentation and labeling of images, adaptive quizzing, and mixed-reality simulations [16,17]. Early work combining AI with augmented and virtual reality shows promise for improving anatomical comprehension and retention [18]. However, these opportunities bring discipline-specific questions: how do anatomy educators view AI's role or adoption alongside dissection and prosection, and what support do they need to use it responsibly?

The Unified Theory of Acceptance and Use of Technology (UTAUT) and its extension Unified Theory of Acceptance and Use of Technology 2 (UTAUT2) offer a robust framework to examine this adoption, highlighting how performance expectancy (improving teaching), effort expectancy (ease of use), social influence, and facilitating conditions (support from peers or institutions), as well as hedonic motivation and habits (personal enjoyment and prior experience), shape technology-related behavioral intentions in medical education settings [19,20]. Recent studies applying the UTAUT framework to faculty adoption of AI in higher education have shown that UTAUT constructs significantly influence educators' behavioral intentions to integrate AI into teaching and learning [21]. Applying this framework allows for systematic exploration of anatomy educators' perceptions, intentions, and contextual barriers to AI use in medical education. However, technology adoption theories [22], while theoretically robust, may not fully capture the discipline-specific, institutional, and context-grounded realities that emerge from educator experiences in specialized fields such as anatomy, where pedagogical traditions (cadaveric dissection, hands-on learning) are deeply embedded [23,24]. Consequently, this study triangulates the deductive, theory-driven UTAUT2 framework with an inductive strengths, weaknesses, opportunities, and threats (SWOT) analysis. This mixed methods approach provides both theoretical coherence, through UTAUT2's operationalization of adoption constructs, and practical, context-specific insights that illuminate the particular barriers and enablers relevant to anatomy education. The SWOT analysis captures anatomy educators' lived experiences, institutional constraints, and aspirations without forcing findings into predetermined theoretical categories, thereby enriching the applicability of adoption theory to a specialized educational context [25]. In the United Arab Emirates, where institutions are advancing national digital transformation agendas [26], examining faculty perspectives on AI integration is timely and essential to inform policy, professional development, and curriculum innovation.

Globally, faculty report persistent concerns about AI-related academic integrity, superficial learning, and uneven, early-stage uptake, underscoring the need for explicit institutional frameworks [27-29]. In anatomy education, these issues are amplified by governance over cadaveric and patient images as well as by threats to assessment validity [13]. In the Middle East and North Africa, the evidence base remains emergent and predominantly student-focused rather than faculty or discipline-specific [30,31]. For example, in Saudi Arabia, the

XSL•FO
RenderX

few faculty-oriented studies report moderate awareness but uneven competence and a need for structured training, alongside concerns about policy inconsistencies, academic integrity, and data governance issues [32-34]. Published United Arab Emirates data are particularly scarce and largely student-focused; to our knowledge, no peer-reviewed study has yet mapped how anatomy educators in the United Arab Emirates perceive AI in their teaching and assessment contexts [31,35].

This study sought to bridge this evidence gap by exploring the perspectives of anatomy educators in the United Arab Emirates regarding AI integration in medical and health professions education. Specifically, this study aimed (1) to examine anatomy educators' use and perceived benefits or risks of AI in teaching and assessment; (2) to assess their readiness and willingness to adopt AI-supported pedagogical tools in their discipline; and (3) to identify the availability and adequacy of institutional support, policy frameworks, and professional development needed to enable responsible and pedagogically sound AI integration. Understanding these views is essential to inform responsible, discipline-appropriate AI adoption in the region and to align global innovations with local educational realities.

Methods

Study Design and Reporting

This study adopted a cross-sectional design integrating quantitative survey data and qualitative thematic analysis to explore anatomy educators' use patterns and perspectives on AI integration in teaching and assessment. Reporting follows STROBE (Strengthening the Reporting of Observational Studies in Epidemiology) for observational cross-sectional studies and relevant CHERRIES (Checklist for Reporting Results of Internet E-Surveys) items for web surveys; qualitative analysis is described in line with SRQR (Standards for Reporting Qualitative Research) and reflexive thematic analysis guidance. Survey prompts and the full instrument are provided in Multimedia Appendix 1. Checklists (STROBE, CHERRIES, and SRQR) are present in Multimedia Appendix 2.

Setting and Recruitment

Data were collected at the 4th Annual Anatomy and Cellular Biology Conference (Dubai Medical College; May 10-11, 2025), an annual national forum drawing anatomy faculty (AF) from all United Arab Emirates institutions. The conference served as an ideal recruitment setting due to its disciplinary focus, diverse institutional representation, and concentration of target participants in a single venue. Conference organizers announced the study during the opening plenary, directing attendees to the printed QR codes at the venue. Research staff (in acknowledgements) provided a brief orientation and offered 2 completion options: email and on-site (via QR codes). Recruitment was extended 2 weeks (May 12-25) via institutional faculty email lists, United Arab Emirates Anatomy Society WhatsApp group (41 members), and snowball referrals. All invitations emphasized voluntary participation, anonymity, independent completion, and absence of institutional coercion. Surveys were delivered via Microsoft Forms.

Participants and Eligibility

Eligible participants were faculty actively teaching anatomy or closely related subdisciplines (eg, gross anatomy, histology, neuroanatomy, embryology, or anatomy and physiology) in undergraduate medical, dental, or allied-health programs in the United Arab Emirates. Inclusion required current teaching involvement and electronic consent. At the time of the study, anatomy courses were offered across 16 colleges of medicine, pharmacy, and health sciences in the country.

Quantitative Data Collection

The structured questionnaire was derived from contemporary literature on AI in health professions education and adapted to faculty contexts [7,12,32,36]. Item development was also informed by the UTAUT2 [19,20], which conceptualizes 6 key determinants of technology adoption: performance expectancy, effort expectancy, social influence, facilitating conditions, hedonic motivation, and habit. These constructs guided the formulation of items assessing faculty perceptions, institutional support, and behavioral intentions toward AI integration.

The quantitative component comprised 5-point Likert-type items (1=strongly disagree to 5=strongly agree), multiple-choice questions, and closed ended checklists organized across four domains: (1) demographics and teaching background, (2) familiarity with educational technology and current use of AI tools, (3) perceived benefits and challenges (including ethics, integrity, privacy, and governance), and (4) professional development interest and future adoption. Each quantitative item explicitly operationalized 1 or more UTAUT2 constructs: performance expectancy (benefit-related items), effort expectancy (ease-of-use items), social influence (peer and leadership support items), facilitating conditions (enabler items), hedonic motivation (enjoyment and satisfaction items), habit (prior experience items), and behavioral intention (stated willingness to adopt). This quantitative strand generated the primary and secondary outcomes described in the Outcomes and Measures section.

Qualitative Data Collection

To capture more nuanced faculty perspectives that could not be fully represented through closed-ended items, the survey incorporated qualitatively oriented open-ended questions organized using the SWOT framework. These SWOT prompts functioned as the primary mechanism for qualitative data collection, inviting free-text narrative responses to elicit detailed, context-rich accounts of faculty experiences and views. The Strengths domain asked educators to identify and describe perceived advantages of AI integration, recount or describe successful experiences, and provide examples of improved teaching or student engagement. The Weaknesses domain invited participants to reflect on and clearly describe observed challenges (technological, pedagogical, institutional, or other), the limitations encountered, and any documented effects on teaching practices or student outcomes. The Opportunities domain explored future applications, novel teaching and research areas, and potential contributions to faculty development and curriculum innovation. The Threats domain elicited concerns about risks, ethical and professional issues, and institutional

XSL•FO
RenderX

risk-mitigation strategies. Two additional meta-questions assessed educators' views on the utility of SWOT analysis for policy development and their recommendations for structured faculty workshops.

Qualitative responses to all SWOT and meta-questions were analyzed thematically using inductive coding to identify recurring themes, negative cases, exemplary quotations, and actionable recommendations. The resulting qualitative themes were then triangulated with the quantitative findings from the structured items to contextualize faculty attitudes, clarify apparent contradictions, and highlight areas of convergence and divergence across data sources. The full mixed methods instrument, including quantitative items and qualitative SWOT prompts, is provided in Multimedia Appendix 1.

Survey Administration and Data Management

The open survey was hosted on Microsoft Forms. IP addresses or institutional identifiers were not collected. To limit duplicate entries, the landing screen requested 1 response per educator; responses were screened for obvious duplicates (identical role or institution or time-stamp patterns, of which there were none) before analysis. Data were exported to CSV for analysis and stored on an encrypted drive accessible only to the research team. Missing responses were left as missing; denominators are reported per item where applicable.

Outcomes and Measures

Four primary outcomes were assessed: (1) attitudes toward AI's role in anatomy education (augment existing practices vs replace traditional methods), (2) perceived benefits of AI (visualization, adaptive learning, assessment efficiency, student engagement, and accessibility), (3) perceived barriers and enablers (technical, pedagogical, ethical or governance, and capacity-related factors), and (4) professional development needs (AI literacy, pedagogy, ethics, technical training, and mentorship). Secondary outcomes included (1) AI familiarity and current use patterns (5-point scale plus yes or no items on use in lectures, assessment, feedback, research, and productivity), (2) differences in perceptions by academic rank (junior, mid-career, and senior), (3) differences by years teaching anatomy (0-5, 6-15, 16-25, and >25 years), (4) differences by age group, and (5) differences by subdiscipline (gross anatomy, neuroanatomy, embryology, and histology). The structured questionnaire used 5-point Likert scales (1=strongly disagree to 5=strongly agree) for all primary outcome items. Multidomain subscales were created for benefits, barriers, and enablers. Composite scores were calculated as means. Open-ended items gathered qualitative elaboration on barriers and professional development needs. Likert items were treated as ordinal. Where applicable, survey items operationalized UTAUT2 constructs: performance expectancy (benefit items), effort expectancy (ease-of-use items), social influence (peer or leadership support), facilitating conditions (enabler items), hedonic motivation (enjoyment), habit (prior experience), and behavioral intention (stated willingness to adopt).

Quantitative Analysis

Five negatively framed attitude items were reverse-coded (6-original score) to ensure consistent directionality, where higher scores indicate more positive attitudes toward AI in anatomy education across all items. Given the small sample size (N=30) and subgroup imbalance (subgroup n ranges 1-18), all subgroup comparisons are framed as exploratory and descriptive rather than inferential. Likert items and variables failing normality were summarized using nonparametric descriptive statistics (medians and IQRs). To identify preliminary patterns, exploratory group comparisons were conducted using Kruskal-Wallis *H* for multigroup contrasts and Mann-Whitney *U* for 2-group comparisons, with effect sizes (Cliff δ) and 95% CIs reported as aids to interpretation. Where composite scales met normality and internal consistency criteria (α≥.70), Welch ANOVA was used as an alternative. However, readers are cautioned that (1) statistical power is limited for detecting real differences in subgroups; (2) null findings do not imply true group equivalence, only lack of detectable difference; (3) effect size CIs are wide due to small sample size, indicating imprecision; and (4) these exploratory findings should not be interpreted as definitive evidence of group differences and require confirmation in larger samples. No formal type I error control was applied to these exploratory comparisons. All tests were 2-tailed with an α of .05, and analyses were performed in Jamovi (version 2.7.6; The Jamovi Project).

Qualitative Analysis

Open-ended responses were analyzed using reflexive thematic analysis (Braun and Clarke's [37] 6 phases): familiarization, inductive coding, theme generation, review, definition or naming, and reporting. Themes were then mapped to the SWOT schema to organize perceptions salient to anatomy teaching and assessment. The coding output was systematically aligned with the 4 SWOT domains, such that each finalized theme and supporting quote were assigned to the quadrant that best represented its perceived valence or strategic implication (eg, internal facilitators under strengths and external challenges under threats). The lead analyst (PLMZ) (an anatomy educator) maintained reflexive memos and an audit trail of coding decisions; peer debriefing with a second anatomist (PM) was used to challenge interpretations and enhance credibility. In line with reflexive thematic analysis, we did not calculate intercoder agreement. Coding was conducted in Microsoft Excel, an approach commonly used in medical-education research, including SWOT-informed curriculum and innovation studies [38]. Representative quotes were selected to illustrate each finalized theme.

Integration of Quantitative and Qualitative Strands

We used a convergent approach: quantitative and qualitative findings were analyzed in parallel and then integrated at interpretation, using the SWOT map to contextualize descriptive statistics (eg, professional development needs, integrity concerns, and time constraints) within faculty narratives.

Sample Size and Precision

This was a census-style pragmatic sample of accessible United Arab Emirates anatomy educators at or after an annual national conference. Although no a priori power calculation was feasible, precision is communicated via effect sizes with CIs and transparent reporting of denominators, where necessary. The

XSL•FO
RenderX

modest sample (N=30) reflects the practical constraints of recruiting at a single national conference and brief postconference follow-up. Although underpowered for robust subgroup or parametric analyses, it is suitable for descriptive, exploratory work mapping educator perceptions in an understudied United Arab Emirates context. Given the paucity of regional, faculty-focused evidence on AI in anatomy education, a qualitative-led mixed methods approach prioritizes contextual depth over statistical generalizability. Convenience sampling limits representativeness; yet, it yields a diverse mix of anatomy educators across United Arab Emirates institutions, supporting thematic exploration with institutional relevance. Quantitative subgroup comparisons are, therefore, presented as descriptive, not inferential, and the sample size was insufficient for regression or structural equation modeling of UTAUT2 constructs. Further interpretive constraints are detailed in the Study Limitations section.

Ethical Considerations

The study received ethics approval from Ajman University Institutional Review Board (approval M-F-H-9-May 2025). Participation was voluntary with electronic informed consent. Participants were not compensated for their participation. Data were anonymous and stored securely. As requested by some participants, individual faculty viewpoints were anonymized and not linked to specific institutions; instead, they were aggregated into broader thematic categories.

Results

Participant Characteristics

A total of 30 AF members from medical and health sciences institutions across the United Arab Emirates participated in the survey, 3 were excluded due to foreign institutions (Table 1). Most respondents were 40- to 49-year-old assistant or associate professors with ≥11 years of teaching experience, and the majority taught gross anatomy, histology, embryology, and neuroanatomy, with half also teaching anatomy and physiology. Institutions were fully deidentified.

Table 1. Demographic and professional characteristics of anatomy faculty respondents who participated in the United Arab Emirates–wide survey on artificial intelligence integration in anatomy education (N=30).

Variable and category	Values, n (%)
Age group (years)	
30-39	6 (20)
40-49	18 (60)
50-59	4 (13.3)
≥60	2 (6.7)
Academic rank or position	
Instructor or tutor	3 (10)
Lecturer	5 (16.7)
Assistant professor	11 (36.7)
Associate professor	7 (23.3)
Professor	4 (13.3)
Years of teaching experience	
0-5	1 (3.3)
6-10	1 (3.3)
11-15	13 (43.3)
≥16	15 (50)
Teaching field[a] (multiple response)	
Gross anatomy	27 (90)
Histology or cell biology or cytology	23 (76.7)
Embryology or developmental biology	22 (73.3)
Neuroanatomy or neurobiology	22 (73.3)
Anatomy and physiology	15 (50)

[a]Multiple responses allowed; percentages for teaching fields use 30 as the denominator and therefore do not sum to 100%.

XSL•FO
RenderX

Faculty Proficiency With Educational Technology

Overall, 73.3% (22/30) of respondents expressed confidence in their technological skills (N=30; mean 3.73, SD 1.01; median 4.0, IQR 3.25-4.00; 95% CI 3.35-4.11), with 56.7% (17/30) agreeing and 16.7% (5/30) strongly agreeing that they were proficient in using educational technologies (Figure S1 and Table S1 in Multimedia Appendix 3). A smaller proportion reported neutral views (5/30, 16.7%), while only 6.7% (2/30) disagreed and 3.3% (1/30) strongly disagreed. Shapiro-Wilk tests confirmed nonnormal distributions for all variables (all $P<.05$). Exploratory Kruskal-Wallis comparisons revealed no detectable differences in proficiency levels across academic ranks or age groups (all $P>.05$). Given small subgroup sizes (n=2-11 per category), these null findings are exploratory and should not be treated as evidence of true group equivalence. These findings indicate a generally high level of digital competence among United Arab Emirates anatomy educators, providing a favorable foundation for adopting AI-enhanced teaching tools and digital learning innovations.

GenAI Use in Anatomy Education

Table 2 shows that most faculty used GenAI in teaching (28/30, 93.3%), led by ChatGPT (26/30, 86.7%), with smaller uptake of Gemini (11/30, 36.7%) and Bing Copilot (10/30, 33.3%). Fewer reported writing assistants (8/30, 26.7%), image generators (6/30, 20%), custom GPTs or plugins (5/30, 16.7%), or Claude (3/30, 10%); 6 of 30 (20%) listed other tools, and 2 of 30 (6.7%) reported no GenAI use. Use was typically weekly or daily, mainly for quiz generation, content summarization, and drafting teaching materials. Furthermore, survey responses to “In what ways have you used GenAI tools in your work as an anatomy educator?” in Table 3 indicated that use was concentrated in teaching content development, particularly generating quiz or examination questions (19/30, 63.3%), summarizing complex topics for lectures (16/30, 53.3%), and drafting instructional materials such as slides or handouts (15/30, 50%). Nearly half also reported research writing or manuscript support (13/30, 43.3%). Assessment-related uses were less frequent, including preparing student feedback or reflections (8/30, 26.7%) and grading assistance or rubric design (8/30, 26.7%), while visual content generation was reported by 7 of 30 (23.3%). A minority had not used GenAI yet (3/30, 10%).

Table 2. Self-reported uses of generative artificial intelligence (GenAI) tools among participating United Arab Emirates anatomy faculty (N=30)[a].

GenAI tool used	Faculty, n (%)
ChatGPT (OpenAI)	26 (86.7)
Google Gemini (formerly Bard)	11 (36.7)
Bing Copilot (Microsoft or OpenAI)	10 (33.3)
Notion AI[b] or Grammarly or writing assistants	8 (26.7)
DALL·E or Midjourney or image generators	6 (20)
Custom GPTs or plugins	5 (16.7)
Claude (Anthropic)	3 (10)
None	2 (6.7)

[a]GenAI tool platforms personally used for teaching. The faculty column shows the number of respondents selecting each platform; percentages are calculated out of 30. Multiple selections were permitted, so totals may exceed 100%.

[b]AI: artificial intelligence.

Table 3. Self-reported purposes for using generative artificial intelligence (GenAI) tools among participating United Arab Emirates anatomy faculty (N=30).

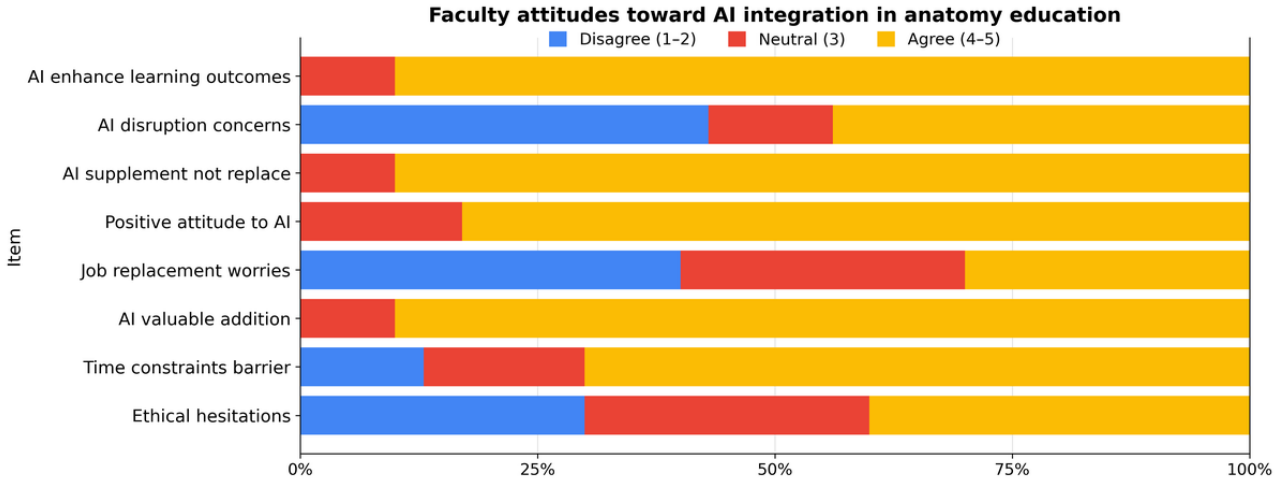
GenAI tool use	Faculty, n (%)
Generating quiz or examination questions	19 (63.3)
Summarizing complex topics for lectures	16 (53.3)
Drafting instructional materials	15 (50)
Research writing or manuscript support	13 (43.3)
Preparing student feedback or reflections	8 (26.7)
Grading assistance or rubric design	8 (26.7)
Generating anatomy images or diagrams	7 (23.3)
I have not used GenAI tools yet	2 (6.7)

XSL•FO
RenderX

Attitudes Toward AI in Medical Education

Figure 1 presents overall positive faculty attitudes toward integrating AI into anatomy education (mean 4.23, SD 0.73). Overall, the majority of respondents (83.3%) agreed or strongly agreed that AI can enhance learning outcomes (mean 4.30, SD 0.65) and represents a valuable addition to anatomy teaching (mean 4.33, SD 0.84). Nearly all participants endorsed the view that AI should supplement rather than replace human educators (mean 4.50, SD 0.57), reflecting a balanced approach to innovation. While ethical hesitations (30%; mean 2.97, SD 0.96) and job replacement worries (26.7%; mean 2.87, SD 1.04) persisted among some faculty, most respondents remained positive about AI's pedagogical role. Time constraints (63.3%; mean 3.60, SD 1.07) and perceived disruption to existing teaching routines (30%; mean 3.20, SD 1.22) emerged as moderate barriers to adoption. These findings indicate that United Arab Emirates anatomy educators exhibit strong acceptance of AI's educational potential, tempered by reasonable caution regarding its ethical and professional implications, an attitudinal profile consistent with early-stage adoption models.

Figure 1. Attitude response patterns toward AI in anatomy education. Response patterns across 8 AI attitude domains showing percentage distribution of agreement levels. Blue bars represent disagreement (scores 1-2; disagree+strongly disagree), reflecting reservations; red bars represent neutrality (score 3), illustrating ambivalent or undecided attitudes; and yellow bars represent agreement (scores 4-5; agree+strongly agree), representing positive attitudes. All items demonstrated significant departures from normality (Shapiro-Wilk W 0.710-0.840; all $P<.001$). Exploratory group comparisons (Kruskal-Wallis H; academic rank: H=10.0-12.6; $P=.01$-$.04$; age group: H=8.8-10.0; $P=.02$-$.03$) revealed preliminary patterns, suggesting that assistant professors endorsed AI somewhat more highly than other ranks, and younger faculty reported greater time constraints. However, these differences are exploratory, given small subgroup sizes (n=2-18 per category), and should be interpreted with caution. The substantive finding is more noteworthy: across all demographic subgroups, faculty maintained predominantly positive attitudes (mean 3.60, SD 1.07; mean 4.5, SD 0.57; median 3.69, IQR 3.50-4.00), with internal consistency high (Cronbach α 0.7-0.9; McDonald ω 0.8), indicating broad consensus on AI's educational role. The observed variation appears to reflect gradations of enthusiasm rather than fundamental disagreement. Power and precision limitations due to small subgroup sizes mean that these exploratory comparisons should be confirmed in larger, representative samples before treating them as definitive group differences (N=30). AI: artificial intelligence.



Barriers and Enablers to AI Integration in Anatomy Education

As shown in Table 4, the most prominent obstacles were capacity-related, notably time constraints, given the workload (mean 3.27, SD 1.17; 46.7% agreement) and training gaps (mean 3.13, SD 1.22; 36.7%), both rated as moderate barriers. Over a quarter of respondents (26.7%-30%) disagreed or strongly disagreed, highlighting variability in readiness across institutions. Budget or resource limitations (mean 2.63, SD 1.19; 53.3% disagreement) and academic-integrity concerns (mean 2.80, SD 1.10; 40% disagreement) were the least endorsed, indicating minor barriers overall. Conversely, motivation by student interest (mean 4.23, SD 0.86; 80% agreement) and institutional encouragement (mean 4.00, SD 1.14; 73.3%) emerged as strong facilitators, with very few dissenting responses (≤10%). Facilitators outweighed barriers, and the main obstacles were capacity-related (time or workload and training) rather than resource-related (budget), indicating that targeted professional development and workload support could capitalize on existing institutional encouragement and student-driven motivation to enable AI integration.

XSL•FO
RenderX

Table 4. Perceived barriers and facilitators to artificial intelligence integration in anatomy education among United Arab Emirates faculty[a].

Factor type and domain	Item	Mean (SD)	Median (IQR)	Agreement (%)	Disagreement (%)	Severity
Barriers						
Capacity constraints	Time or workload constraints	3.27 (1.17)	3.0 (3.0-4.0)	50	23.3	Moderate
Capacity constraints	Training gaps	3.13 (1.22)	3.0 (2.0-4.0)	46.7	30	Moderate
Quality concerns	Academic integrity	2.80 (1.09)	3.0 (2.0-4.0)	30	40	Minor
Resource constraints	Budget or resource limitations	2.63 (1.19)	2.0 (2.0-3.5)	26.7	53.3	Minor
Facilitators						
Extrinsic motivation	Motivation by student interest	4.23 (0.86)	4.0 (4.0-5.0)	80	3.3	Strong
Institutional support	Institutional encouragement	4.00 (1.15)	4.0 (4.0-5.0)	73.3	10	Strong

[a]Responses were rated on a 5-point Likert scale (1=strongly disagree, 5=strongly agree). Agreement (%)=percentage scoring 4-5; agree+strongly agree; Disagreement (%)=percentage scoring 1-2; disagree+strongly disagree; severity classifies barriers (minor: mean<3.0; moderate: mean≥3.0) and facilitators (moderate: mean<4.0; strong: mean≥4.0). Capacity constraints (time and training) emerged as primary barriers, while social and institutional support provided strong facilitation. All items nonnormally distributed (Shapiro-Wilk P<.05). Exploratory comparisons revealed no detectable differences by academic rank or age (Kruskal-Wallis P>.05) in perceived barriers or enablers. However, this null finding should not be interpreted as evidence of true group equivalence; the small sample size and subgroup heterogeneity limit statistical power to detect meaningful differences if they exist. A larger, stratified sample would be needed to definitively test whether barriers differ by seniority, age, or experience. Descriptively, all faculty emphasized time and training as primary obstacles regardless of demographic background (N=30).

Thematic SWOT Analysis

Overview

In the thematic SWOT analysis, coded segments were organized within a strengths-weaknesses-opportunities-threats structure, with weaknesses expressed more prominently and accompanied by recurrent support-needs statements. Themes clustered around infrastructure or resource readiness and training or professional development as core prerequisites for adoption, alongside concerns relating to time or workload, cost, ethics or privacy or integrity, and leadership or policy or governance. Smaller but salient strands captured pedagogical opportunities (including 3D personalization), caution about overreliance and erosion of critical thinking, and equity or access considerations. These themes were summarized within the SWOT framework (Textbox 1) and mapped to UTAUT2 constructs to contextualize faculty-perceived enablers and barriers to AI integration in anatomy education.

XSL•FO
RenderX

Textbox 1. Strengths, weaknesses, opportunities, and threats (SWOT) analysis of faculty perspectives on artificial intelligence (AI) integration in anatomy education, mapped to key constructs of the Unified Theory of Acceptance and Use of Technology 2 (UTAUT2).

Strengths (faculty or institutional strengths)

- Supportive institutional climate (social influence and facilitating conditions that foster behavioral intention to adopt AI tools)
- Robust digital infrastructure (facilitating conditions that underpin the strong performance expectancy)
- Emerging professional development and skill building (facilitating conditions+effort expectancy)
- Innovation culture and growth mindset (reduces perceived effort expectancy and supports hedonic motivation+behavioral intention)

Weaknesses (barriers and internal constraints)

- Insufficient training and awareness (reduced effort expectancy and limited facilitating conditions)
- Time and workload burden (reduced effort expectancy and limited facilitating conditions)
- Infrastructure and IT support gaps (reduced performance expectancy, doubt AI will work reliably at the point of need)
- Policy and governance ambiguity (undermine social influence and behavioral intention—policy ambiguity can make faculty reluctant to fully embrace AI if they are not given clear rules on how to use it)
- Student preparedness and time pressures (without strengthened facilitating conditions, effort expectancy, and governance, AI adoption risks remaining fragmented and superficial)

Opportunities (perceived benefits or future potential)

- Enhanced anatomy learning outcomes (high performance expectancy; AI can make anatomy more comprehensible and memorable)
- Workload efficiency and scalability (improving both performance expectancy and effort expectancy reduces preparation and workflow burden)
- Personalization and engagement in anatomy pedagogy (performance expectancy [better learning] and hedonic motivation [more engaging, enjoyable experiences]).

Threats (perceived risks and external pressures)

- AI overreliance and deskilling (unhelpful habit formation and diminished performance expectancy for deeper, independent learning)
- Ethical and integrity risks (undermine social influence and behavioral intention)
- Loss of traditional, hands-on anatomy teaching (diminished performance expectancy, hedonic motivation threat to learning outcomes, and valued experiential and relational aspects of anatomy education)

Note: Open-ended responses (N=145 coded segments) were analyzed using Braun and Clarke's reflexive thematic analysis and organized into 4 SWOT domains. Each quadrant with themes reflects patterned meaning across faculty narratives within the SWOT domains, illustrating how perceived strengths and opportunities (enablers) coexist with governance, ethical, and infrastructural vulnerabilities (weaknesses and threats) in AI-driven anatomy education. Each quadrant represents inductively derived subthemes linked deductively to UTAUT2 constructs: performance expectancy (perceived usefulness and teaching enhancement), effort expectancy (perceived ease of AI adoption), social influence (peer and student encouragement), facilitating conditions (institutional and technical supports), hedonic motivation (enjoyment and engagement), habit, and behavioral intention.

Current Strengths (Internal Assets)

Thematic analysis of the open-ended responses identified 4 interrelated themes reflecting strong internal capacities for AI integration in anatomy education (Table S1 in Multimedia Appendix 4). Collectively, these strengths map primarily onto the UTAUT2 constructs of facilitating conditions, performance expectancy, social influence, and hedonic motivation. Respondents described a generally supportive environment characterized by digital readiness, leadership encouragement, and a culture of experimentation.

Many educators credited their institutions with explicit endorsement of AI-enhanced teaching, describing "leadership promoting AI" and "supportive administrative, leadership drive toward technology adoption." Such "strong support" signals organizational commitment and establishes a positive normative climate for faculty adoption. Faculty perceive encouragement from both senior management and departmental peers as lowering the perceived risk of innovation, thereby strengthening behavioral intention to engage with AI tools. Typical remarks included:

> *supportive administrative, leadership drive towards AI-driven pedagogy and incorporation into everyday work* [AF30]
>
> *my institution is very supportive* [AF3]
>
> *AI policies in place* [AF14]

A recurrent strength was the presence of solid digital infrastructure, "access to digital infrastructure such as high-speed internet, modern computers, and interactive displays" and "Internet infrastructure." These assets provide essential technical foundations for integrating AI into anatomy teaching workflows. Respondents viewed the availability of virtual reality, simulation platforms, and medical holodecks as tangible evidence that AI integration is practical and sustainable. Representative statements include:

> *I use VR headsets and Medical Holodeck* [AF4]

XSL•FO
RenderX

> *we have high-speed internet, modern computers, and interactive displays* [AF17]
>
> *have immersive rooms, AI-backed simulators* [AF11]
>
> *The infrastructure (is) provided by the institutional Design Lab supports trying AI.* [AF5]

Several participants highlighted ongoing or planned professional development and skill-building initiatives: "faculty development sessions," "certificate courses on AI in education," and "formal training in integration." Such structured learning opportunities reduce the cognitive and technical burden associated with AI use, reinforcing confidence and ease of use. Faculty see these opportunities as a bridge from conceptual awareness to confident practice. Examples of supporting statements include:

> *There are faculty development session and even certificate courses on AI in education* [AF28]
>
> *general AI courses offered to all.* [AF20]

Multiple responses emphasized an innovative culture and growth mindset as depicted by positive, experimental attitude, "readiness to explore and experiment with the available tools," "curiosity to integrate AI," and "general willingness to learn." This intrinsic motivation reflects institutional cultures that reward innovation and experimentation. Enjoyment and curiosity act as important motivational drivers that complement the extrinsic facilitators of infrastructure and policy: "explore and experiment with available tools"(AF19) and "institutional encouragement" for technological integration (AF8).

Weaknesses (Internal Challenges)

The 61 coded weakness segments described a landscape of readiness gaps and resource strain that constrain otherwise positive attitudes toward AI in anatomy education (Table S2 in Multimedia Appendix 4). Inductively, 4 main themes emerged, which were then interpreted through the UTAUT2 lens.

The most widespread concern was the need for faculty upskilling for AI use and awareness gaps thereof, expressed in statements such as:

> *Many colleagues don't know how to use AI tools.* [AF21]
>
> *Lack of training or awareness,* [AF10]
>
> *insufficient professional development in AI,* [AF3]
>
> *requests for formal training in the integration of AI-related technologies.* [AF25]

These findings highlight systemic and capacity bottlenecks that constrain AI adoption. The majority of participants wanted "increased awareness of the available AI resources" (AF13) and "formal training and collaborative research (evidence-based) opportunities underlining a pressing need for institutional investment in capacity building" (AF7).

Second, respondents described time, workload, and resourcing constraints as key internal barriers. Several references to high existing workload and calls for "protected time to experiment" and "less workload and better financial support" illustrate that frontline faculty often lack the time, hardware, and licenses needed to experiment, redesign sessions, and troubleshoot. These constraints weaken facilitating conditions and lower performance expectancy (doubts that AI will work reliably in real teaching situations) and further depress effort expectancy. These were expressed through statements such as:

> *need time and workload relief, protected space to experiment with tools, and improved hardware and connectivity* [AF21]
>
> *no access to time to develop skill* [AF11]
>
> *not enough time to learn or implement AI tools due to workload.* [AF15]

Third, ethical anxieties featured prominently across responses. Participants anticipated "risk of cheating and plagiarism" and "inappropriate reliance on generative outputs," expressing fear that AI might compromise academic integrity if inadequately regulated. Such concerns influence social norms around AI use and erode behavioral intention, as educators hesitate to integrate tools that could foster academic misconduct or blur authorship accountability.

> *risk of cheating and plagiarism.*
>
> *Concerns about academic integrity*
>
> *Limited student preparedness* [AF1]
>
> *cadaver, xrays and patient case data will be uploaded to AI*
>
> *AI holds promise for anatomy, but its success depends on ethical use, proper training, and ensuring it complements, not replaces, human instruction.* [AF7]

A final prominent theme was fragility or shortage of reliable AI infrastructure and information technology support: "Sometimes the network and lab equipment are not dependable," "... limited AI enabled infrastructure or IT support," one respondent wrote, while another summarized simply, "lack of institutional AI policy." Even where digital resources existed, several educators worried about their uneven availability and sustainability across programs. The absence of clear institutional policies leaves uncertainty about acceptable practices, further weakening facilitating conditions and normative guidance. In contrast to the strong internal assets identified among strengths, weaknesses expose systemic shortfalls in readiness, particularly around faculty capability, protected time, and policy coherence.

Opportunities (External Enablers)

Thematic analysis of the opportunity statements revealed a forward-looking optimism toward AI's potential to transform anatomy education (Table S3 in Multimedia Appendix 4). These opportunities align closely with the UTAUT2 constructs of performance expectancy, effort expectancy, hedonic motivation, and facilitating conditions. The current participants envisioned equity and concrete pedagogical and operational gains, with gains including cost, time, and workload.

A dominant theme in the Opportunities responses was the view that AI can strengthen conceptual understanding by making complex 3D anatomy more vivid, supporting personalized learning pathways, and enabling rapid formative feedback. Some participants explicitly anticipated that:

XSL•FO
RenderX

> *AI in anatomy education offers personalized learning, interactive 3D visualization, instant feedback, and efficient assessment,* [AF5]
>
> *AI can help generate diagrams and 3D models for better understanding.* [AF18]
>
> *be used as a tool that complements the available teaching resources since it also has its own limitations, particularly for practical courses like anatomy.* [AF8]

Furthermore, efficiency and productivity gains featured with statements such as "less time to prepare for lectures," "efficient workflow with large groups," and "streamlining work." Others expected similar benefits through "long-term protection of time for faculty" and "AI will lessen workload," as routine content generation and assessment tasks will be automated. Several noted potentials for "enhancing student engagement," improved access for diverse and multilingual learners if AI was to be integrated thoughtfully as represented by:

> *"AI could make learning available to more students" in "different languages, forms and catering for different abilities."* [AF9]
>
> *Students may be able to grasp certain concepts better with AI enhanced tools or visualization* [AF14]
>
> *AI in anatomy education offers personalized learning, interactive 3D models, and adaptive quizzes.* [AF5]
>
> *will provide more means to support students' engagement and curiosity.* [AF17]

High performance expectancy and hedonic motivation dominate this quadrant, as educators see AI as a means to make anatomy learning more efficient, engaging, and accessible.

Threats (External Challenges)

The external risks and cautionary perspectives were also evident and were centered around educational quality and system reliability (open-ended threat responses; Table S4 in Multimedia Appendix 4). Although participants generally support AI integration, they cautioned against unintended consequences that could compromise learning depth, authenticity, and human engagement. These perceived risks align with the UTAUT2 constructs of habit, performance expectancy, social influence, and behavioral intention, reflecting a nuanced tension between enthusiasm for innovation and anxiety about its potential excesses.

The anatomy educators worried about potential overreliance on AI and erosion of critical thinking or clinical reasoning and practical skill development, warning that:

> *Over relying on AI tools may decrease students/educators ability to reasoning, imagination and self development.* [AF4]

Others noted:

> *Overreliance on AI could weaken hands-on skills* [AF13]
>
> *Decrease in students learning capacity and long-term memory, lack of hands-on training, lack of live social communication* [AF23]
>
> *unrealistic representation of anatomy* [AF20]
>
> *complete loss of traditional way of learning anatomy.* [AF30]
>
> *hard to know if assignments are genuine* [AF11]
>
> *AI in anatomy education may pose risks like data privacy issues, biased content, reduced human interaction, and high implementation costs.* [AF6]

Others flagged technical failures and infrastructural vulnerability, recognizing that system failures or licensing problems could disrupt classes: "If the system fails, teaching will be disrupted" (AF13). One participant remarked on the "lack of infrastructure, even pro-versions of the AI tools" (AF10), underscoring that the cost of subscriptions to commercially available teaching tools is currently being shouldered largely by anatomy educators. Plagiarism and academic integrity concerns were also raised repeatedly, reflecting anxiety about unmonitored AI use in assessments and scholarly work. More broadly, respondents noted that the growing integration of GenAI in anatomy education introduces additional threats, including data privacy risks, biased outputs, high implementation costs, and increasing difficulty in verifying the authenticity of students' work. Finally, a subset of participants report "None" or "Nothing I can think of, if used wisely," indicating conditional trust; AI is seen as safe if guided by thoughtful policies and pedagogy. Overall, threats describe a tension between the allure of innovation and a desire to protect core educational values, standards, and skills.

Quantitative-Qualitative Integration Summary

The qualitative SWOT findings complemented the survey data by contextualizing the statistically dominant barriers and facilitators identified in Table 4. Quantitative ratings of time constraints due to existing workload and training deficit (moderate barriers) were elaborated through rich descriptions of heavy workload and capacity challenges, confirming limited facilitating conditions under the UTAUT2 framework. Likewise, the strong quantitative endorsement of motivation by student interest and institutional encouragement as facilitators was echoed in the SWOT opportunities, where learner enthusiasm and organizational innovation as well as supportive culture emerged as primary drivers of AI adoption. The synthesis, thus, revealed a coherent pattern of high perceived usefulness and social motivation counterbalanced by systemic limitations imposed by governance deficits, infrastructural limitations, and workload. Collectively, these results depict United Arab Emirates anatomy educators as being in a transitional phase of AI acceptance, optimistic about pedagogical benefits, yet constrained by governance, training, and ethical infrastructure gaps that must be addressed for sustainable, responsible implementation.

Discussion

Principal Findings

This study provides one of the first systematic accounts of United Arab Emirates anatomy educators' patterns of use and perspectives on AI, showing strong but cautious uptake of generative tools in everyday work. Most faculty expressed

XSL•FO
RenderX

confidence in their educational technological skills and had already experimented with GenAI, mainly using ChatGPT and similar systems for content creation, assessment support, and research-related tasks, with smaller but notable use for visualization and workflow streamlining. Attitudes toward AI in anatomy education were broadly positive and oriented toward augmentation rather than replacement of traditional teaching and assessment, with no significant attitudinal differences by rank, age, or experience. These patterns merit confirmation in larger cohorts, given this modest (N=30), conference-recruited sample's power limitations for subgroup analysis (n=1-18). Quantitative barriers to AI use centered on time or workload pressures and training gaps, while student interest and institutional support or encouragement emerged as strong enablers, a pattern reinforced by SWOT themes of digital readiness, innovation or supportive culture, and concurrent concerns about ethics, governance, and deskilling. Interpreted through UTAUT2, these findings reflect high performance expectancy and supportive social influence but reduced effort expectancy under workload and training constraints, which together temper behavioral intention and keep AI use largely exploratory rather than routine. In this mixed methods interpretation, SWOT themes were mapped onto UTAUT2 constructs to illustrate how perceived benefits, barriers, and institutional supports cohere into an adoption profile for United Arab Emirates anatomy educators; however, this mapping is interpretive and should be understood as a heuristic rather than a definitive test of the theory.

Faculty Readiness and Professional Development

Most respondents were already using GenAI tools daily or weekly for quizzes, lectures, and content preparation and reported a strong interest in AI's educational value, yet only moderate confidence in their own skills. SWOT comments such as "many colleagues do not know how to use AI tools," "lack of training or awareness," and calls for "formal training in the integration of AI-related technologies" highlighted limited hands-on experience as a key weakness that constrains confident classroom use. In UTAUT2 terms, this combination of high perceived benefit and low practical confidence reflects reduced effort expectancy: faculty see AI as pedagogically useful (high performance expectancy) but technically demanding to learn and integrate, which dampens behavioral intention despite strong student interest and institutional encouragement. These findings point to the need for structured, workflow-aligned faculty development focused on concrete use cases, prompt design, and critical appraisal of AI outputs to convert enthusiasm into routine, pedagogically grounded integration in anatomy teaching. Similar readiness gaps are reported internationally, where educators endorse AI's potential but describe low competence and limited time for meaningful adoption [9,11,39-41]. Regional work from the Middle East and North Africa and South East Asia similarly describes faculty who recognize AI's potential, yet lack skills and call for structured capacity building [32,34,42]. Together with these reports, the present findings support targeted, context-sensitive upskilling in practical AI use, verification, and ethics as a key strategy to overcome low effort expectancy and enable sustainable adoption [43-45].

Institutional Climate: Support Present, Policy Lagging

Quantitative data showed strong institutional encouragement (mean 4.00, SD 1.14, 73.3% agree+strongly agree) and student interest (mean 4.23, SD 0.86, 80% agree+strongly agree) as leading facilitators, alongside 93.3% GenAI adoption primarily for anatomy content creation and research-related support. Within the UTAUT2 framework, this pattern suggests strong social influence from students and institutional leadership, alongside generally favorable facilitating conditions (eg, adequate digital infrastructure and/or a supportive culture). Collectively, these factors are likely to enhance performance expectancy and, in turn, strengthen behavioral intention to adopt AI [20]. However, SWOT responses revealed gaps in AI disclosure rules, cadaveric or patient image governance, and assessment integrity policies, leaving faculty uncertain about acceptable use for anatomy-specific materials. This policy ambiguity weakens facilitating conditions and generates risk-averse behavior, stalling progression from exploratory to routine use even in a supportive climate [19,46]. These tensions mirror regional deanery reports of pro-AI attitudes coexisting with "immature" formal frameworks [29,47] and align with anatomy-specific governance concerns around image provenance and technology-mediated assessments [13,36], underscoring the need for discipline-specific policy in the United Arab Emirates.

Ethics, Integrity, and Data Governance

Although mean scores for ethical worry were moderate, open-ended responses repeatedly cited risks of plagiarism and overreliance, the need to verify and document AI outputs, and sensitivities around patient or cadaveric data. Faculty worried that students might use AI to shortcut assignments, undermining hands-on skill development, deep spatial reasoning, and professionalism, threats that are particularly acute in anatomy's experiential, cadaver-based learning. In UTAUT2 terms, such uncertainties erode facilitating conditions, as a lack of clear governance and oversight make faculty cautious despite high performance expectancy for AI-supported visualization and feedback. These concerns are consistent with international scholarship calling for strong human oversight of AI in teaching and assessment and warning against uncritical automation [36]. Multisite qualitative work in Saudi Arabia, concordant with current findings, reports early enthusiasm tempered by uncertainty about trustworthiness, bias, and privacy, emphasizing the need for explicit use norms [48,49]. While AI-related policy appears to be lagging in the current United Arab Emirates study, medical schools are advised at the institutional level to implement governance frameworks that include explicit acceptable-use statements, disclosure requirements and/or audit trails, and staff development focused on bias detection and verification [50]. This aligns with scoping reviews that advocate for structured, ethically grounded implementation rather than ad-hoc experimentation. In the United Arab Emirates context, the absence of formal AI policies leaves AF reliant on personal judgment, which can inhibit broader adoption by increasing uncertainty and perceived risk. Clear institutional guidelines would strengthen facilitating conditions and support confident, ethical integration of AI tools in anatomy curricula.

XSL•FO
RenderX

Workload and Pragmatic Enablers

SWOT responses identified time and workload pressure as prominent barriers, with faculty citing heavy teaching loads, few anatomy-specific AI training opportunities, and a lack of protected time to experiment with tools. While some respondents anticipated that AI might eventually reduce effort (eg, by automating quiz generation or lesson planning), they consistently noted that these benefits would follow a "long learning curve" and initial workload increase due to prompt development and careful verification of outputs. This "time-savings paradox," upfront effort before efficiency gains, is echoed in the broader AI-in-education literature [9,51,52]. One participant stressed the need for "time and workload relief, protected space to experiment with tools," illustrating that adoption depends on capacity and support, not just access. Mapped onto UTAUT2, these findings show that high workload and limited support reduce effort expectancy and weaken facilitating conditions, thereby suppressing behavioral intention, whereas structured experimentation, repeated exposure, and organizational reinforcement can eventually consolidate AI use into a habit [20,53]. Editorial and policy guidance similarly recommend sandboxing, prompt-craft workshops, and verification checklists before expecting durable time savings at scale [54].

Strengths Unique to Anatomy

A distinctive contribution of this study is that faculty did not describe AI in generic "productivity" terms alone; rather, they articulated discipline-specific affordances that map directly onto core challenges in anatomy teaching. Respondents consistently highlighted the potential for enhanced visualization and spatial support, including automated labeling of anatomical structures, adaptive quizzing, and access to high-fidelity 3D models, capabilities they framed as particularly valuable for topics where learners must build robust 3D mental models. Importantly, faculty linked these affordances to learning-process benefits, noting that well-designed tools could support spatial reasoning and reduce extraneous cognitive load, especially when visualization, labeling, and assessment cues are integrated coherently. This anatomy-specific emphasis, visual-spatial scaffolding, structure identification, and 3D conceptualization represent a notable and somewhat unique pattern in our dataset relative to broader AI adoption discussions in medical education.

Equally salient in our findings was a discipline-grounded boundary condition: faculty were not "anti-AI," but they were explicit that AI must complement, not displace, laboratory-based, tactile, and relational forms of learning that are central to anatomy's epistemology and professional socialization. Participants cautioned against overreliance, potential erosion of hands-on competence, and diminished human interaction and situated learning, mirroring early evaluations and commentaries that position tools such as ChatGPT, simulations, and automated resources as adjuncts rather than substitutes for dissection or prosection and cadaveric context [55-57]. Recent syntheses likewise argue for explicit human oversight, verification practices, and alignment with learning-science principles (including cognitive load) to protect fidelity and assessment validity in anatomy [16,58,59]. Taken together, this study adds a nuanced, anatomy-specific account: educators recognize powerful visual-spatial advantages, but they also define clear pedagogical "red lines" around authenticity, embodiment, and professional formation.

On the basis of these discipline-specific insights, we recommend that United Arab Emirates and wider Gulf Cooperation Council institutions move beyond generic AI workshops and instead implement structured, anatomy-tailored faculty development, competency-aligned training in prompt design for anatomical content, bias or accuracy appraisal, and ethical verification workflows. In parallel, institutions should operationalize governance addressing AI disclosure, acceptable use of cadaveric and patient images, and safeguards for assessment validity. Finally, given the regional commonalities identified in this cohort, Gulf Cooperation Council–wide collaboration (shared faculty-development hubs, shared evaluation checklists, and interuniversity policy alignment) could accelerate safe innovation while supporting broader digital-transformation ambitions in higher education [44].

Study Limitations

This study has several limitations. This census-style pragmatic sample (N=30) of United Arab Emirates anatomy educators recruited primarily at a national conference (May 2025) plus brief follow-up reflects practical constraints but limits representativeness and statistical power. Convenience sampling underpowers robust subgroup analyses (n=1-18 per category; ~20%-35% power for medium effects) and precludes regression of UTAUT2 pathways; null results should therefore be interpreted as lack of detectable difference, not evidence of equivalence. Self-report measures may be influenced by social desirability or early-adopter bias, given the high GenAI use reported, although SWOT comments also captured critical perspectives and governance concerns. No a priori power calculation was feasible; consequently, the emphasis is on descriptive quantitative patterns and qualitative depth in an understudied context. Future research should use larger, stratified samples, longitudinal designs, and intervention trials to test whether AI-supported anatomy teaching improves learning outcomes and preserves assessment validity as technologies and institutional policies evolve. Triangulation of ordinal data and reflexive thematic SWOT across United Arab Emirates institutions as well as a discipline-specific focus are marked strengths.

Conclusions

In this national snapshot, United Arab Emirates anatomy educators reported broadly positive attitudes toward AI, oriented toward augmentation rather than replacement of established practices. Faculty perceptions converged across quantitative results and SWOT analysis: strong enablers like institutional encouragement and student interest outweighed barriers of time or workload, training gaps, and integrity concerns. Notably, 93.3% already used GenAI tools, primarily ChatGPT for anatomy content creation, visualization or 3D simulation, and assessment support. Educators perceived pedagogical value in 3D visualization and adaptive quizzing, while cautioning against overreliance eroding tactile dissection skills essential for spatial reasoning. Faculty-identified needs centered on development,

XSL•FO
RenderX

governance or disclosure policies, and assessment safeguards to maintain integrity amid GenAI use.

Acknowledgments

The authors thank anatomy educators across United Arab Emirates medical and health sciences colleges for participating and the organizers of the 4th Anatomy and Cellular Biology Conference (Dubai Medical College, May 2025) for facilitating initial data collection. The authors also acknowledge Drs Samson Chengetanai and Trust Nyirenda for their assistance with conference-based data collection. Generative artificial intelligence (ChatGPT [OpenAI] and Perplexity.ai) was used during manuscript preparation to support limited tasks such as language editing and drafting or refining text. All artificial intelligence–assisted content was critically reviewed, edited, and fact-checked by the authors, who take full responsibility for the manuscript.

Data Availability

The datasets generated or analyzed during this study are available from the corresponding author on reasonable request.

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors. The study was supported by institutional resources (including article processing charge) from the Ajman University.

Authors' Contributions

PLMZ conceived the study, designed the survey instrument, coordinated data collection, and led the writing of the manuscript. PM assisted in quantitative analysis, visualization, and drafting of the results section. JKN contributed to data validation, qualitative coding, and integration of mixed methods results. LZ supervised the study, provided critical conceptual and editorial input, and approved the final manuscript. All authors reviewed and approved the final version of the manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Survey prompts and the full instrument.
[DOCX File , 18 KB - mededu_v12i1e87418_app1.docx]

Multimedia Appendix 2
Completed reporting checklists for this mixed methods study: CHERRIES (web-based survey reporting), STROBE (observational reporting), and SRQR (qualitative reporting).
[PDF File (Adobe PDF File), 382 KB - mededu_v12i1e87418_app2.pdf]

Multimedia Appendix 3
United Arab Emirates anatomy faculty members' self-rated technology proficiency, familiarity with artificial intelligence and generative artificial intelligence tools, and the self-reported frequency of generative artificial intelligence use for teaching and other academic purposes (eg, lesson planning, assessment item writing, research writing, and administrative tasks).
[DOCX File , 193 KB - mededu_v12i1e87418_app3.docx]

Multimedia Appendix 4
Strengths, weaknesses, opportunities, and threats data.
[DOCX File , 27 KB - mededu_v12i1e87418_app4.docx]

References

1. Chan KS, Zary N. Applications and challenges of implementing artificial intelligence in medical education: integrative review. JMIR Med Educ 2019;5(1):e13930 [FREE Full text] [doi: 10.2196/13930] [Medline: 31199295]
2. Mergen M, Junga A, Risse B, Valkov D, Graf N, Marschall B, medical tr.AI.ning consortium. Immersive training of clinical decision making with AI driven virtual patients—a new VR platform called medical tr.AI.ning. GMS J Med Educ 2023;40(2):Doc18 [FREE Full text] [doi: 10.3205/zma001600] [Medline: 37361242]
3. Cianciolo AT, LaVoie N, Parker J. Machine scoring of medical students' written clinical reasoning: initial validity evidence. Acad Med 2021;96(7):1026-1035 [FREE Full text] [doi: 10.1097/ACM.0000000000004010] [Medline: 33637657]
4. Narayanan S, Ramakrishnan R, Durairaj E, Das A. Artificial intelligence revolutionizing the field of medical education. Cureus 2023;15(11):e49604 [FREE Full text] [doi: 10.7759/cureus.49604] [Medline: 38161821]

XSL•FO
RenderX

5. Ang ET, Chan JM, Gopal V, Li Shia N. Gamifying anatomy education. Clin Anat 2018;31(7):997-1005. [doi: 10.1002/ca.23249] [Medline: 30168609]
6. Okonkwo CW, Ade-Ibijola A. Chatbots applications in education: a systematic review. Comput Educ Artif Intell 2021;2:100033. [doi: 10.1016/j.caeai.2021.100033]
7. van Gaalen AEJ, Brouwer J, Schönrock-Adema J, Bouwkamp-Timmer T, Jaarsma ADC, Georgiadis JR. Gamification of health professions education: a systematic review. Adv Health Sci Educ Theory Pract 2021;26(2):683-711 [FREE Full text] [doi: 10.1007/s10459-020-10000-3] [Medline: 33128662]
8. Sriram A, Ramachandran K, Krishnamoorthy S. Artificial intelligence in medical education: transforming learning and practice. Cureus 2025;17(3):e80852 [FREE Full text] [doi: 10.7759/cureus.80852]
9. Blanco MA, Nelson SW, Ramesh S, Callahan CE, Josephs KA, Jacque B, et al. Integrating artificial intelligence into medical education: a roadmap informed by a survey of faculty and students. Med Educ Online 2025;30(1):2531177 [FREE Full text] [doi: 10.1080/10872981.2025.2531177] [Medline: 40660466]
10. Buele J, Llerena-Aguirre L. Transformations in academic work and faculty perceptions of artificial intelligence in higher education. Front Educ 2025;10:1603763. [doi: 10.3389/feduc.2025.1603763]
11. Saleh ZT, Rababa M, Elshatarat RA, Alharbi M, Alhumaidi BN, Al-Za'areer MS, et al. Exploring faculty perceptions and concerns regarding artificial intelligence chatbots in nursing education: potential benefits and limitations. BMC Nurs 2025;24(1):440 [FREE Full text] [doi: 10.1186/s12912-025-03082-0] [Medline: 40251532]
12. Gordon M, Daniel M, Ajiboye A, Uraiby H, Xu NY, Bartlett R, et al. A scoping review of artificial intelligence in medical education: BEME Guide No. 84. Med Teach 2024;46(4):446-470 [FREE Full text] [doi: 10.1080/0142159X.2024.2314198] [Medline: 38423127]
13. Lazarus MD, Truong M, Douglas P, Selwyn N. Artificial intelligence and clinical anatomical education: promises and perils. Anat Sci Educ 2024;17(2):249-262. [doi: 10.1002/ase.2221] [Medline: 36030525]
14. Preiksaitis C, Rose C. Opportunities, challenges, and future directions of generative artificial intelligence in medical education: scoping review. JMIR Med Educ 2023;9:e48785 [FREE Full text] [doi: 10.2196/48785] [Medline: 37862079]
15. Bölek KA, De Jong G, Henssen D. The effectiveness of the use of augmented reality in anatomy education: a systematic review and meta-analysis. Sci Rep 2021 Jul 27;11(1):15292. [doi: 10.1038/s41598-021-94721-4] [Medline: 34315955]
16. Joseph TS, Gowrie S, Montalbano MJ, Bandelow S, Clunes M, Dumont AS, et al. The roles of artificial intelligence in teaching anatomy: a systematic review. Clin Anat 2025;38(5):552-567. [doi: 10.1002/ca.24272] [Medline: 40269576]
17. Nakao E, Igeta M, Kobayashi N, Kumazu Y, Otani Y, Murakami M, et al. Effectiveness of artificial intelligence-based visualization for surgical anatomy education: a cluster quasirandomized controlled trial. Surgery 2025;188:109723. [doi: 10.1016/j.surg.2025.109723] [Medline: 41005002]
18. Moro C, Birt J, Stromberga Z, Phelps C, Clark J, Glasziou P, et al. Virtual and augmented reality enhancements to medical and science student physiology and anatomy test performance: a systematic review and meta-analysis. Anat Sci Educ 2021;14(3):368-376. [doi: 10.1002/ase.2049] [Medline: 33378557]
19. Venkatesh V, Morris M, Davis G, Davis F. User acceptance of information technology: toward a unified view. MIS Q 2003;27(3):425-478. [doi: 10.2307/30036540]
20. Venkatesh V, Thong JYL, Xu X. Consumer acceptance and use of information technology: extending the unified theory of acceptance and use of technology. MIS Q 2012;36(1):157-178. [doi: 10.2307/41410412]
21. Perez RCL. AI in higher education: Faculty perspective towards artificial intelligence through UTAUT approach. HCMCOUJS Soc Sci 2024 Apr 02;14(4):32-50. [doi: 10.46223/hcmcoujs.soci.en.14.4.2851.2024]
22. Venkatesh V, Thong JYL, Xu X. Unified theory of acceptance and use of technology: a synthesis and the road ahead. J Assoc Inf Syst 2016;17(5):328-376. [doi: 10.17705/1jais.00428]
23. Rizzolo LJ. Human dissection: an approach to interweaving the traditional and humanistic goals of medical education. Anat Rec 2002 Dec 15;269(6):242-248 [FREE Full text] [doi: 10.1002/ar.10188] [Medline: 12467081]
24. McBride JM, Drake RL. National survey on anatomical sciences in medical education. Anat Sci Educ 2018;11(1):7-14. [doi: 10.1002/ase.1760] [Medline: 29265741]
25. Phadermrod B, Crowder RM, Wills GB. Importance-performance analysis based SWOT analysis. Int J Inform Manag 2019;44:194-203. [doi: 10.1016/j.ijinfomgt.2016.03.009]
26. UAE National Strategy for Artificial Intelligence 2031. UAE AI Strategy. 2018. URL: https://ai.gov.ae/strategy/ [accessed 2026-02-03]
27. Kallunki V, Kinnunen P, Pyörälä E, Haarala-Muhonen A, Katajavuori N, Myyry L. Navigating the evolving landscape of teaching and learning: university faculty and staff perceptions of the artificial intelligence-altered terrain. Educ Sci 2024;14(7):727. [doi: 10.3390/educsci14070727]
28. Maluleke AF. AI adoption in African higher education: a systematic review of benefits and ethical implications. IJER 2025;7(2):a05. [doi: 10.38140/ijer-2025.vol7.2.05]
29. Ichikawa T, Olsen E, Vinod A, Glenn N, Hanna K, Lund GC, et al. Generative artificial intelligence in medical education-policies and training at US osteopathic medical schools: descriptive cross-sectional survey. JMIR Med Educ 2025;11:e58766 [FREE Full text] [doi: 10.2196/58766] [Medline: 39934984]

XSL•FO
RenderX

30. Almalki M, Alkhamis MA, Khairallah FM, Choukou MA. Perceived artificial intelligence readiness in medical and health sciences education: a survey study of students in Saudi Arabia. BMC Med Educ 2025;25(1):439 [FREE Full text] [doi: 10.1186/s12909-025-06995-1] [Medline: 40140763]
31. Issa WB, Shorbagi A, Al-Sharman A, Rababa M, Al-Majeed K, Radwan H, et al. Shaping the future: perspectives on the integration of artificial intelligence in health profession education: a multi-country survey. BMC Med Educ 2024;24(1):1166 [FREE Full text] [doi: 10.1186/s12909-024-06076-9] [Medline: 39425151]
32. Al Zahrani EM, Elsafi SH, Al Musallam LD, Alharbi AH, Aldossari HM, Alomar AM, et al. Faculty perspectives on artificial intelligence's adoption in the health sciences education: a multicentre survey. Front Med (Lausanne) 2025;12:1663741 [FREE Full text] [doi: 10.3389/fmed.2025.1663741] [Medline: 41090120]
33. Alsahafi Z, Baashar A. Knowledge, Attitudes, and practices related to artificial intelligence among medical students and academics in Saudi Arabia: a systematic review. Cureus 2025 May;17(5):e83437. [doi: 10.7759/cureus.83437] [Medline: 40462802]
34. Alshanberi AM, Mousa AH, Hashim SA, Almutairi RS, Alrehali S, Hamisu AM, et al. Knowledge and perception of artificial intelligence among faculty members and students at Batterjee Medical College. J Pharm Bioallied Sci 2024;16(Suppl 2):S1815-S1820 [FREE Full text] [doi: 10.4103/jpbs.jpbs_1162_23] [Medline: 38882896]
35. Alam Z, Abdullahi AS, Alnuaimi SNS, Al Shaka HA, Alderei SSS, Alhemeiri AAA, et al. eHealth literacy and attitudes towards use of artificial intelligence among university students in the United Arab Emirates, a cross-sectional study. Front Digit Health 2025;7:1574263 [FREE Full text] [doi: 10.3389/fdgth.2025.1574263] [Medline: 40978699]
36. Eva KW. Medical education and artificial intelligence: responsible and effective practice requires human oversight. Med Educ 2024;58(11):1260-1261. [doi: 10.1111/medu.15495] [Medline: 39180233]
37. Braun V, Clarke V. Using thematic analysis in psychology. Qual Res Psychol 2008;3(2):77-101. [doi: 10.1191/1478088706qp063oa]
38. Lertsakulbunlue S, Sertsuwankul P, Thammasoon K, Jongcherdchootrakul K, Sakboonyarat B, Kantiwong A. Student engagement in medical research curriculum development strategies: a comprehensive approach utilizing mixed-methods study and SWOT-TOWS-AHP analysis. Med Sci Educ 2025;35(2):741-754. [doi: 10.1007/s40670-024-02226-2] [Medline: 40352982]
39. Cervantes J, Smith B, Ramadoss T, D'Amario V, Shoja MM, Rajput V. Decoding medical educators' perceptions on generative artificial intelligence in medical education. J Investig Med 2024;72(7):633-639. [doi: 10.1177/10815589241257215] [Medline: 38785310]
40. Dong C, Aw DCW, Lee DWC, Low SC, Yan CC. Generative AI in healthcare: insights from health professions educators and students. Int Med Educ 2025;4(2):11. [doi: 10.3390/ime4020011]
41. McCoy L, Ganesan N, Rajagopalan V, McKell D, Niño DF, Swaim MC. A training needs analysis for AI and generative AI in medical education: perspectives of faculty and students. J Med Educ Curric Dev 2025;12:23821205251339226 [FREE Full text] [doi: 10.1177/23821205251339226] [Medline: 40376309]
42. Salem MA, Zakaria OM, Aldoughan EA, Khalil ZA, Zakaria HM. Bridging the AI gap in medical education: a study of competency, readiness, and ethical perspectives in developing nations. Computers 2025;14(6):238. [doi: 10.3390/computers14060238]
43. Ehmke SD, Bridges J, Patel SE. Self-perceived knowledge, skills, and attitude of nursing faculty on generative artificial intelligence in nursing education: a descriptive, cross-sectional study. Teach Learn Nurs 2025;20(3):222-227. [doi: 10.1016/j.teln.2025.01.029]
44. Fadlelmula FK, Qadhi SM. A systematic review of research on artificial intelligence in higher education: practice, gaps, and future directions in the GCC. J Univ Teach Learn Pract 2024;21(6) [FREE Full text] [doi: 10.53761/pswgbw82]
45. Heston T, Khun C. Prompt engineering in medical education. IME 2023;2(3):198-205. [doi: 10.3390/ime2030019]
46. Xu S, Chen P, Zhang G. Exploring Chinese university educators' acceptance and intention to use AI tools: an application of the UTAUT2 model. Sage Open 2024;14(4):21582440241290013. [doi: 10.1177/21582440241290013]
47. Mansour N, Msheik EF, Zaatari G, Harb M. The perspectives of deans of medical schools on introducing artificial intelligence and computer literacy to medical curricula in Arab countries. In: Badran A, Baydoun E, Hillman S, Mesmar J, editors. Higher Education in the Arab World. Cham: Springer; 2024:309-344.
48. Abouammoh N, Alhasan K, Aljamaan F, Raina R, Malki KH, Altamimi I, et al. Perceptions and earliest experiences of medical students and faculty with ChatGPT in medical education: qualitative study. JMIR Med Educ 2025;11:e63400 [FREE Full text] [doi: 10.2196/63400] [Medline: 39977012]
49. Salih SM. Perceptions of faculty and students about use of artificial intelligence in medical education: a qualitative study. Cureus 2024;16(4):e57605 [FREE Full text] [doi: 10.7759/cureus.57605] [Medline: 38707183]
50. Triola MM, Rodman A. Integrating generative artificial intelligence into medical education: curriculum, policy, and governance strategies. Acad Med 2025;100(4):413-418. [doi: 10.1097/ACM.0000000000005963] [Medline: 39705530]
51. Berg GVD. Generative AI and educators: partnering in using open digital content for transforming education. Open Praxis. 2024. URL: https://openpraxis.org/articles/10.55982/openpraxis.16.2.640 [accessed 2025-10-06]

XSL•FO
RenderX

52. Lee D, Arnold M, Srivastava A, Plastow K, Strelan P, Ploeckl F, et al. The impact of generative AI on higher education learning and teaching: a study of educators' perspectives. Comput Educ Artific Intell 2024;6:100221. [doi: 10.1016/j.caeai.2024.100221]
53. Kim J, Klopfer M, Grohs JR, Eldardiry H, Weichert J, Cox LA, et al. Examining faculty and student perceptions of generative AI in university courses. Innov High Educ 2025;50(4):1281-1313. [doi: 10.1007/s10755-024-09774-w]
54. Hodges CB, Kirschner PA. Innovation of instructional design and assessment in the age of generative artificial intelligence. TechTrends 2023;68(1):195-199. [doi: 10.1007/s11528-023-00926-x]
55. Cornwall J, Hildebrandt S, Champney TH, Goodman K. Ethical concerns surrounding artificial intelligence in anatomy education: should AI human body simulations replace donors in the dissection room? Anat Sci Educ 2024;17(5):937-943. [doi: 10.1002/ase.2335] [Medline: 37750493]
56. Leng L. Challenge, integration, and change: ChatGPT and future anatomical education. Med Educ Online 2024;29(1):2304973 [FREE Full text] [doi: 10.1080/10872981.2024.2304973] [Medline: 38217884]
57. Mogali SR. Initial impressions of ChatGPT for anatomy education. Anat Sci Educ 2024;17(2):444-447. [doi: 10.1002/ase.2261] [Medline: 36749034]
58. Pandurangam G, Gurajala S, Nagajyothi D. Artificial intelligence in anatomy teaching and learning: a literature review. Nat J Clin Anat 2024;13(3):158-163. [doi: 10.4103/njca.njca_103_24]
59. Sirasanagandla SR, Rajendran SS, Mogali SR, Bouchareb Y, Shaffi N, Al-Rahbi A. From cadavers to neural networks: a narrative review on artificial intelligence tools in anatomy teaching. Educ Sci 2025;15(3):283. [doi: 10.3390/educsci15030283]

Abbreviations

AF: anatomy faculty
AI: artificial intelligence
CHERRIES: Checklist for Reporting Results of Internet E-Surveys
GenAI: generative artificial intelligence
SRQR: Standards for Reporting Qualitative Research
STROBE: Strengthening the Reporting of Observational Studies in Epidemiology
SWOT: strengths, weaknesses, opportunities, and threats
UTAUT: Unified Theory of Acceptance and Use of Technology
UTAUT2: Unified Theory of Acceptance and Use of Technology 2

Edited by A Stone; submitted 08.Nov.2025; peer-reviewed by S Qazi, F Fukuzawa; comments to author 15.Dec.2025; revised version received 15.Jan.2026; accepted 20.Jan.2026; published 21.Apr.2026.

Please cite as:
Zilundu PLM, Mazengenya P, Narayanan JK, Zhou L
Faculty Perspectives on AI Integration in Anatomy Education in the United Arab Emirates: Cross-Sectional Survey
JMIR Med Educ 2026;12:e87418
URL: https://mededu.jmir.org/2026/1/e87418
doi:10.2196/87418

©Prince Last Mudenda Zilundu, Pedzisai Mazengenya, Jayaraj Kodangattil Narayanan, LiHua Zhou. Originally published in JMIR Medical Education (https://mededu.jmir.org), 21.Apr.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Education, is properly cited. The complete bibliographic information, a link to the original publication on https://mededu.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Performance of Large Language Models on the Brazilian National Medical Education Examination: Comparative Benchmark Study

Francys de Luca Fernandes da Silva[1*], MD; Eduardo Augusto Roeder[2*], MD; João Victor Bruneti Severino[1,2], MSc; Matheus Nespolo Berger[2]; Pedro Angelo Basei de Paula[2], MD; Davi Ferreira[3], MD, MSc; Maria Han Veiga[4], PhD; Thyago Proença de Moraes[1*], MD, PhD; Gustavo Lenci Marques[1,2*], MD, PhD

[1]R. Imac. Conceição, 1155 - Prado Velho, Pontifícia Universidade Católica do Paraná, Curitiba, Paraná, Brazil

[2]Universidade Federal do Paraná, Curitiba, Paraná, Brazil

[3]Instituto Tecnológico de Aeronáutica, São José dos Campos, São Paulo, Brazil

[4]The Ohio State University, Columbus, OH, United States

[*]these authors contributed equally

Corresponding Author:
Gustavo Lenci Marques, MD, PhD
Universidade Federal do Paraná
Rua General Carneiro 181, 10° andar
Curitiba, Paraná, 80.060-900
Brazil
Phone: 55 41 3360 1800
Email: Gustavomarques@ufpr.br

Abstract

Background: Large language models (LLMs) are rapidly incorporated into medical education and examination preparation; yet, most benchmarking evidence is derived from English-language material. Whether frontier commercial models and Brazilian Portuguese domain-specialized systems perform equivalently on high-stakes Brazilian medical examinations remains unclear.

Objective: This study aims to quantify and compare the performance of 9 frontier commercial LLMs and 1 Brazilian Portuguese domain-specialized system (Charcot, Voa Health) on 2026 Brazilian National Medical Education Examination (Exame Nacional de Avaliação da Formação Médica [ENAMED] 2026) and to describe the patterns of systematic between-model error as complementary quality signal.

Methods: All 100 items of ENAMED 2026 (99 valid after annulment) were administered to 10 frontier-panel models across 5 independent runs under identical Portuguese prompts (temperature=0; top-p=.95). Commercial models were accessed through a unified OpenRouter client layer (DeepSeek provider-pinned). The primary outcome was mean accuracy against the preliminary key; the secondary outcomes were convergence error (CE), normalized mean response time (NMRT), and intermodel agreement. Accuracy was analyzed with Shapiro-Wilk, Levene, Kruskal-Wallis (ε^2), Dunn-Holm post hoc, and a binomial generalized linear mixed model with question and run random intercepts. NMRT excluded Charcot (different stack) and Grok 4 (latency outlier). A total of 7 open-weight and small language models were assessed as a small language model (SLM) substudy.

Results: Frontier-panel accuracy ranged from 73.74% (365/495) for GPT-4o-mini to 96.97% (480/495) for Charcot. Accuracy was nonnormal (Shapiro-Wilk, W=0.82; P<.001) with homogeneous variance (Levene P=.26). Kruskal-Wallis showed large between-model differences (H_9=47.65; P<.001; ε^2=0.97). Dunn-Holm flagged 8 of 45 pairs: Charcot was separable from GPT-4o-mini, DeepSeek v3.2-exp, and Grok 4, but not from the top frontier cluster. The generalized linear mixed model preserved the ranking (all comparators odds ratio<1 vs Charcot; upper CI<1 except GPT-5). A total of 9 items met the default CE criterion; item 77 showed 10-of-10 convergence, later confirmed by Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira rectification, and sensitivity analysis preserved the CE set (5-20 items). Intermodel agreement was high (Fleiss κ=0.852; Krippendorff α=.852). Among 8 retained commercial models, NMRT correlated positively with accuracy (Spearman ρ=0.74; P=.04); these results are interpreted as descriptive of an orchestration-level latency–accuracy law. SLM accuracy ranged from 47.47% (Gemma 3 4B) to 82.22% (GPT-OSS 120B), with lower agreement (Fleiss κ=0.508) and 17 CE items. The best SLM-panel model lagged every frontier-panel model except GPT-4o-mini, with a 12-15 percentage-point gap against the top cluster.

XSL•FO
RenderX

Conclusions: On ENAMED 2026, a Brazilian Portuguese domain-specialized system ranked first, indistinguishable from frontier commercial cluster and above subfrontier and open-weight systems. Charcot's architecture is not publicly disclosed; these findings should be interpreted as comparative black-box evidence of performance and not as mechanistic evidence of specialization. CE was stable, and it prospectively flagged 1 rectified item, supporting its use as a quality assurance screen.

(JMIR Med Educ 2026;12:e89839) doi:10.2196/89839

KEYWORDS

artificial intelligence; benchmarking; Brazilian Portuguese; clinical reasoning; ENAMED; Exame Nacional de Avaliação da Formação Médica; large language models; medical education

Introduction

Contemporary clinical judgment requires rational analysis amid a growing body of scientific evidence. Large language models (LLMs) have emerged as candidate tools to support this work, owing to their ability to integrate multiple sources and maintain stable behavior across long interactions, free from the fatigue and working memory constraints that limit human performance. Understanding how reliably LLMs address complex clinical tasks is therefore a pivotal question for their safe adoption in real-world scenarios [1].

Licensing and certification examinations based on multiple-choice questions (MCQs) are widely used as benchmarks. Most published evidence, however, relies on Anglophone exams such as the United States Medical Licensing Examination, limiting extrapolation to health care systems with distinct terminology, epidemiology, and guidelines, including Brazil [2-5].

Recent work evaluating generalist models on Brazilian national examinations (the Revalida and the Brazilian Society of Cardiology board certification exam) has shown inconsistent performance, suggesting that language and local clinical context influence results. In parallel, domain-specific adaptations such as fine-tuning and retrieval-augmented generation over national guidelines have been associated with improvements in clinical reasoning [6-9].

The National Examination for the Evaluation of Medical Education (Exame Nacional de Avaliação da Formação Médica [ENAMED]), established by the Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (INEP), comprises MCQs spanning multiple specialties and is aligned with Brazilian national guidelines. ENAMED is therefore a standardized national testbed for both generalist and context-adapted LLMs [10].

In this study, we compared 9 frontier generalist LLMs and 1 Brazilian Portuguese domain-specialized system (Charcot, Voa Health) on ENAMED 2026. We evaluated accuracy, temporal behavior (wall clock), intermodel agreement, and patterns of shared error. Our a priori hypothesis was that regional/domain specialization would be associated with higher accuracy on context-dependent items and that between-model error convergence could identify both shared model blind spots and potential answer-key inconsistencies. We also conducted a secondary subanalysis on 7 open-weight and small language models (SLMs) to characterize the gap between frontier and subfrontier systems in Brazilian Portuguese board-level medical reasoning.

Methods

Ethical Considerations

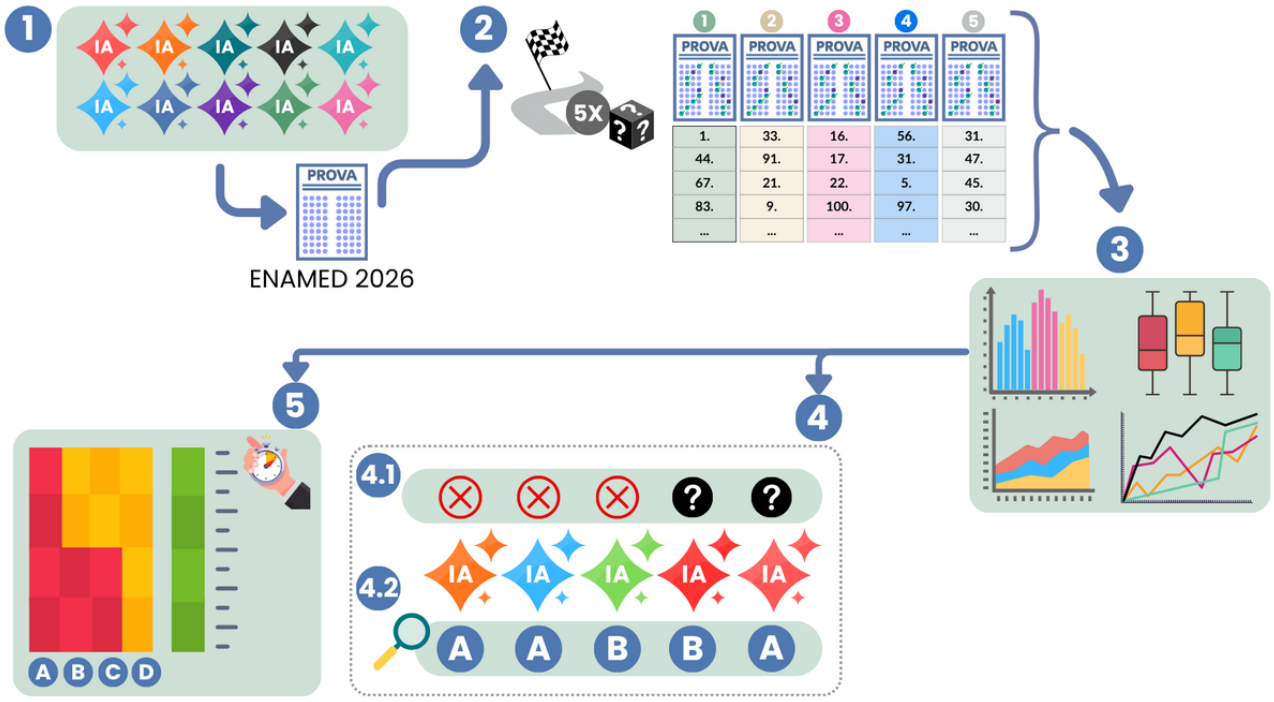
This study did not involve human participants or identifiable data; it used publicly available ENAMED 2026 examination items published by INEP. Institutional review board approval was therefore not applicable.

Study Design

This observational quantitative benchmark administered the ENAMED 2026 examination to 10 LLMs under identical prompting, inference parameters, and orchestration. The study did not involve human participants or identifiable data; it used publicly available ENAMED 2026 items released by INEP. Institutional review board approval was therefore not applicable. Figure 1 summarizes the construction of the study design.

XSL•FO
RenderX

Figure 1. Study workflow for Exame Nacional de Avaliação da Formação Médica (ENAMED) 2026. A total of 10 models (frontier panel) plus 7 small language model (SLM)–panel models were administered all 100 items (99 valid) over 5 runs each, under temperature=0.0 and top-p=0.95, with per-run randomization of item order and alternative order. Analyses: accuracy (primary); convergence error, NMRT, intermodel agreement, and per-question false discovery rate (FDR)–corrected chi-square (secondary); and generalized linear mixed model (GLMM) as sensitivity analysis.



Evaluated Models

We selected 10 models for the frontier panel (Table 1). Charcot is a proprietary Brazilian Portuguese domain-specialized system. The remaining 9 are frontier commercial application programming interfaces (APIs) accessed through a unified client layer. A total of 7 additional open-weight and SLMs were evaluated under the same protocol as a secondary subanalysis (Table 1 for panel assignment). Model identifiers, version strings, and access windows are reported in Multimedia Appendix 1.

XSL•FO
RenderX

Table 1. Evaluated models. Every model other than Charcot was accessed through a unified OpenRouter client layer resolving to the provider's canonical first-party endpoint (DeepSeek provider-pinned; details provided in the "Methods" section and Multimedia Appendix 1).

Model	Panel[a]
Charcot	Frontier[b]
GPT-5	Frontier
Claude Opus 4.1	Frontier
Claude Sonnet 4.5	Frontier
Gemini 2.5 Pro	Frontier
Grok 4	Frontier
GPT-4.1	Frontier
GPT-4o	Frontier
GPT-4o-mini	Frontier
DeepSeek v3.2-exp	Frontier
GPT-OSS 120B	SLM[c]
Llama 4 Scout	SLM
Qwen3 8B	SLM
Gemma 3 27B	SLM
Ministral 3B	SLM
Phi-4	SLM
Gemma 3 4B	SLM

[a]Indicates analytical panel assignment.

[b]Primary hypothesis panel.

[c]SLM: small language model (subanalysis of open-weight and SLMs).

Model Access and Reproducibility

All models other than Charcot were accessed through a unified API gateway (OpenRouter [11]) configured to resolve every request to a fixed canonical provider endpoint (OpenAI, Anthropic, Google, xAI, DeepSeek, Meta, Mistral, Microsoft, Alibaba) using a standardized client-side interface. OpenRouter was chosen to equalize client-side orchestration, parameter serialization (temperature, top-p, and seed where exposed), retry/timeout handling, and response parsing across providers, thereby reducing implementation variance between models.

For first-party proprietary models, OpenRouter is methodologically equivalent to a unified client library: the model weights, inference infrastructure, and version exposed are those of the provider itself and do not cache chat completions by default. Additionally, each of the 5 runs used independent seeds and randomized question and alternative order, making response-level cache hits improbable regardless of the caching configuration. Specific OpenRouter model identifiers, version pins, access windows, and the underlying provider routing for each model are listed in Multimedia Appendix 1.

Among the evaluated systems, DeepSeek v3.2-exp is the only model for which OpenRouter exposes multiple inference providers. To prevent uncontrolled provider rotation from introducing silent infrastructure variance between runs, the DeepSeek route was pinned to a single provider for the entire study window, with the resolved provider identifier, routing decision, and per-call headers logged for every request. Provider logs and configuration are retained in Multimedia Appendix 2. For all other commercial models in the frontier panel, OpenRouter resolves to the provider's canonical first-party end point, so provider pinning is not applicable.

Charcot is a proprietary system developed by Voa Health. For competitive and regulatory reasons, its architecture, training corpus composition, retrieval components, and inference pipeline cannot be disclosed at the level of granularity available for other models. From a methodological standpoint, Charcot can be interpreted as a black-box system under evaluation, subjected to the same operational conditions as the commercial APIs (single final answer plus rationale, identical Portuguese prompts, no external browsing, and deterministic inference parameters where exposed). This opacity is an inherent limitation of including proprietary systems in benchmarks and is properly discussed in the "Study Limitations" section.

All inference calls were executed on a single orchestrator host (specifications are available in Multimedia Appendix 1). External browsing and tool use were disabled for every model. Scoring was automated against the official preliminary key to eliminate assessor subjectivity.

ENAMED 2026 Examination

The official ENAMED 2026 examination (Caderno 01, exam type 1) was used. It comprises 100 MCQs, each with a single correct answer among 5 alternatives (A-E), spanning internal

XSL•FO
RenderX

medicine, surgery, primary care, public health, pediatrics, and obstetrics and gynecology. Items and alternatives were transcribed verbatim into structured text. The preliminary official answer key released by INEP was adopted as the gold standard. Item 10 was annulled by INEP after release, yielding 99 valid items. Item 77 was subsequently rectified by INEP (key D to B); primary accuracy was computed against the preliminary key used at analysis time, and the postrectification interpretation of item 77 is provided as a secondary contextual analysis in the "Convergence Error" section.

Examination Administration Protocol

Each model completed the full examination in 5 independent runs, producing 500 responses per model (100 items × 5 runs). In every run we applied (1) question-order randomization preserving content integrity, (2) internal randomization of alternatives (A-E) preserving the text bound to each option, (3) fixed pseudorandom seeds per run for reproducibility, and (4) standardized system and user prompts in Portuguese requesting a single final alternative plus a concise justification. For every item, we recorded the selected alternative, the generated justification, and the end-to-end wall-clock response time (from request submission to full response receipt).

Inference parameters were held constant across models wherever exposed by the API: temperature=0.0, top-p=0.95. Temperature=0.0 reduces output variability but does not guarantee full determinism, because of hardware nondeterminism (CUDA kernel reductions, atomic adds, and mixed precision) and tied token probabilities [12-14]. The 5-run protocol was therefore chosen to empirically capture residual stochastic variability rather than to force strict determinism; between-run SDs per model are reported in Results (Frontiers in Multimedia Appendix 3 and SLM in Multimedia Appendix 4). Prompt templates (including exploratory reasoning-oriented variants tested during piloting and not used in the primary analysis) are provided in Multimedia Appendix 1.

Outcomes

Primary Outcome: Accuracy

The primary outcome was mean accuracy per model, defined as the proportion of valid items answered correctly relative to the preliminary official answer key, averaged over the 5 runs.

Secondary Outcome: Convergence Error

Convergence error (CE) is a construct introduced in this work to capture systematic shared error between models. A question was flagged as CE-positive when at least 3 evaluated models converged on the same incorrect alternative in all 5 runs (default threshold: ≥3 models × 5 of 5 runs). A sensitivity grid spanning {2, 3, 4} model thresholds × {3, 4, 5} run thresholds was prespecified to assess CE set stability. CE is a descriptive signal; no inferential test was attached to it.

Secondary Outcome: Normalized Mean Response Time

Wall-clock time per item was averaged across the 5 runs for each model and question, then standardized as a *z* score against (1) the global mean and SD (Normalized Mean Response Time [NMRT]-global) and (2) the within-model mean and SD (NMRT-model). NMRT reflects end-to-end orchestration latency (routing, provider queueing, inference, and streaming), rather than pure inference time. Two prespecified exclusions were applied to NMRT. Charcot was excluded because it does not traverse the OpenRouter stack used for every other commercial model, so absolute wall-clock comparisons would be structurally nonequivalent. Grok 4 was excluded because its xAI-via-OpenRouter latency distribution showed an extremely long tail (mean 35.6, SD 33.6 s; median 22.6, IQR 18.2-37.7 s; occasional spikes exceeding 400 s) that dominated the variance of the global *z* score and distorted visual comparisons among the remaining, more homogeneously-routed models.

Secondary Outcome: Intermodel Agreement

Agreement on selected alternatives across the evaluated models was quantified with Fleiss κ and Krippendorff α (nominal) over 99 valid items, providing a complementary view of between-model convergence that is symmetric with respect to correctness (both shared right and shared wrong answers increase agreement).

Qualitative Analysis of Rationales

For CE-positive items and items of particular clinical interest, we conducted a descriptive qualitative analysis of the generated justifications aimed at (1) identifying correct reasoning leading to incorrect conclusions (or vice versa), (2) mapping alignment with Brazilian national guidelines, and (3) characterizing recurrent reasoning shortcuts or omissions. No formal faithfulness metric was applied; this analysis contextualizes the numerical findings rather than quantifying them.

Statistical Analysis

All analyses were conducted in Python 3.13 using scipy, statsmodels, scikit-posthocs, and purpose-built modules for CE, NMRT, and agreement. Accuracy distributions were checked for normality with the Shapiro-Wilk test and for variance homogeneity with the Levene test. Because normality was rejected, between-model differences were tested with the Kruskal-Wallis rank-sum test; effect sizes were quantified as epsilon squared and eta-squared-H, interpreted following Tomczak and Tomczak [15]: small, 0.01-0.06; moderate, 0.06-0.14; and large, ≥0.14. When the omnibus test was significant, pairwise comparisons followed with Dunn test under Holm correction, using rank-biserial correlation magnitude as the pair-level effect size with direction given by mean ordering. Per-question heterogeneity was assessed with chi-square tests pooled across runs under Benjamini-Hochberg false discovery rate (FDR) control at α=.05. As a prespecified sensitivity analysis, a binomial GLMM with logit link modeled per-item correctness as a function of model (fixed effect, Charcot as reference) with crossed random intercepts for question and run, producing item- and run-conditioned odds ratios (ORs) with 95% CIs. Between-run variability was summarized as SD [16], coefficient of variation, and range in percentage points (pp) per model. The Spearman correlation between mean NMRT and mean accuracy was computed over the subset of models retained for NMRT analysis. The significance threshold for all inferential tests was α=.05; CE and agreement analyses were descriptive.

XSL•FO
RenderX

Open-Weight and SLMs Substudy

A total of 7 additional systems were evaluated under the same protocol as an SLM substudy: 4 open-weight mid/large models (GPT-OSS 120B, Llama 4 Scout, Qwen3 8B, and Gemma 3 27B) and 3 SLMs (Ministral 3B, Phi-4, and Gemma 3 4B). Selection prioritized models that (1) were publicly released with open weights, (2) are representative of current open-weight frontiers at distinct capacity tiers (3B to 120B parameters), and (3) were accessible through the same OpenRouter client layer used for the frontier panel, ensuring identical prompting, decoding (temperature=0.0; top-p=0.95), 5-run structure, question and alternative randomization, and scoring. OpenRouter IDs, underlying providers, version pins, and access windows for each SLM-panel model are listed in Multimedia Appendix 1. This panel was analyzed independently from the frontier hypothesis test, with its own Kruskal-Wallis, CE, agreement, and run-variability outputs, to avoid diluting the primary contrast between frontier commercial models and the Brazilian Portuguese domain-specialized system while still allowing the 2 panels to be read side by side when characterizing the frontier × subfrontier gap.

Results

General Model Accuracy

The preliminary official key comprised 99 valid items (item 10 annulled). Across 5 independent runs with temperature=0.0, Charcot achieved the highest mean accuracy (480/495, 96.97%; SD 0.71 pp; 95% CI 96.36%-97.58%), followed by GPT-5 (467/495, 94.34%; SD 1.53), Gemini 2.5 Pro (465/495, 93.94%; SD 0.71), Claude Opus 4.1 (455/495, 91.92%; SD 0.00), Claude Sonnet 4.5 (452/495, 91.31%; SD 0.55), GPT-4o (450/495, 90.91%; SD 0.00), GPT-4.1 (445/495, 89.90%; SD 0.71), Grok 4 (436/495, 88.08%; SD 0.45), DeepSeek v3.2-exp (424/495, 85.66%; SD 1.11), and GPT-4o-mini (365/495, 73.74%; SD 1.01) (Table 2 and Figure 2A). Within-model SD stayed below 1.53 pp in every model (median 0.71, IQR 0.48-0.94 pp), supporting the adequacy of the 5-run protocol for capturing residual stochastic variability under deterministic decoding (detailed per-model run variability and quantile-quantile diagnostic are available in Multimedia Appendix 3). Accuracy was not normally distributed (Shapiro-Wilk W=0.82; P<.001; n=50) but variances were homogeneous (Levene W=1.29; P=.26, k=10). Kruskal-Wallis therefore served as the primary between-model test: H_9=47.65; P<.001; ε^2=0.97; η^2_H=0.97. The unusually high ε^2 is a structural property of the design rather than an overstatement: with only 5 replicates per model under deterministic decoding, within-model variance is near 0, so the between-model component accounts for virtually all variance in the pooled rank distribution. We therefore interpret ε^2 here as a between-model separability index under controlled decoding, and report distribution-level (Dunn-Holm) and item-conditional (GLMM) evidence as complementary, less ceiling-sensitive views.

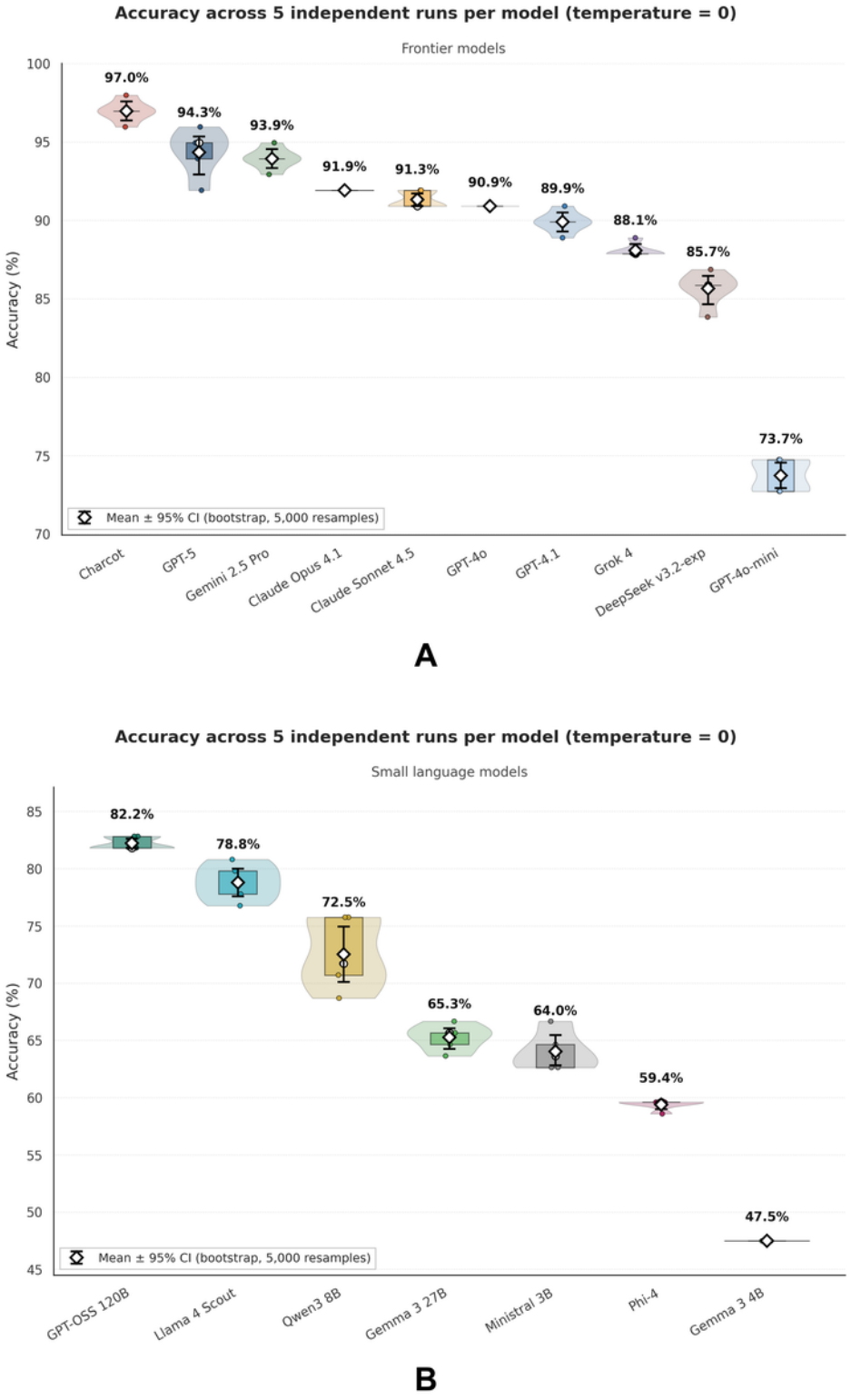
Table 2. Accuracy on Exame Nacional de Avaliação da Formação Médica (ENAMED) 2026, frontier panel (10 models × 5 runs × 99 valid items). We computed 95% CIs by 5000-draw bootstrap. Kruskal-Wallis H9=47.65; P<.001; ε2=0.97 (large, details in the “Methods” section).

Model	Accuracy, mean (SD)[a]	Median (IQR); % (pp)	Minimum-maximum (%)	95% CI	Correct/total, n/N
Charcot	96.97 (0.71)	96.97 (96.97-96.97); 0.00	95.96-98.99	96.36-97.58	480/495
GPT-5	94.34 (1.53)	94.95 (93.94-94.95); 1.01	91.92-95.96	92.93-95.35	467/495
Gemini 2.5 Pro	93.94 (0.71)	93.94 (93.94-93.94); 0	92.93-94.95	93.33-94.55	465/495
Claude Opus 4.1	91.92 (0.00)	91.92 (91.92-91.92); 0	91.92-91.92	91.92-91.92	455/495
Claude Sonnet 4.5	91.31 (0.55)	90.91 (90.91-91.92); 1.01	90.91-91.92	90.91-91.72	452/495
GPT-4o	90.91 (0.00)	90.91 (90.91-90.91); 0	90.91-90.91	90.91-90.91	450/495
GPT-4.1	89.90 (0.71)	89.90 (89.90-89.90); 0	88.89-90.91	89.29-90.51	445/495
Grok 4	88.08 (0.45)	87.88 (87.88-87.88); 0	87.88-88.89	87.88-88.48	436/495
DeepSeek v3.2-exp	85.66 (1.11)	85.86 (85.86-85.86); 0	83.84-86.87	84.65-86.46	424/495
GPT-4o-mini	73.74 (1.01)	73.74 (72.73-74.75); 2.02	72.73-74.75	72.93-74.55	365/495

[a]Mean accuracy is expressed in percentage (%), while SD is expressed in percentage points (pp).

XSL•FO
RenderX

Figure 2. Mean accuracy per model with 95% bootstrap CIs (5000 draws) on Exame Nacional de Avaliação da Formação Médica (ENAMED) 2026 (99 valid items × 5 runs). (A) Frontier panel: 10 models (Charcot and 9 frontier commercial application programming interfaces [APIs]). (B) Small language model (SLM) substudy: 7 open-weight and SLMs. Error bars reflect between-run variability. Between-run SD stayed ≤1.53 percentage points (pp) in every frontier-panel model (median 0.71, IQR 0.48-0.94 pp) and ≤3.15 pp in every SLM-panel model, supporting the 5-run protocol. Detailed run-variability plots and quantile-quantile diagnostics supporting the use of Kruskal-Wallis in both panels are reported in Multimedia Appendices 3 and 4.



Pairwise Post Hoc Comparisons (Dunn-Holm)

Dunn test with Holm correction identified 8 of 45 pairs as significant (P<.05; Table 3 and Figure 3). Charcot differed from GPT-4o-mini (P=<.001; $|r|$=1.00), DeepSeek v3.2-exp (P=<.001; $|r|$=1.00), and Grok 4 (P=.006; $|r|$=1.00). GPT-4o-mini was significantly lower than GPT-5 (P=.002), Gemini 2.5 Pro (P=.003), and Claude Opus 4.1 (P=.049). DeepSeek v3.2-exp differed from GPT-5 (P=.01) and Gemini 2.5 Pro (P=.02). All other pairwise comparisons (including Charcot vs GPT-5,

XSL•FO
RenderX

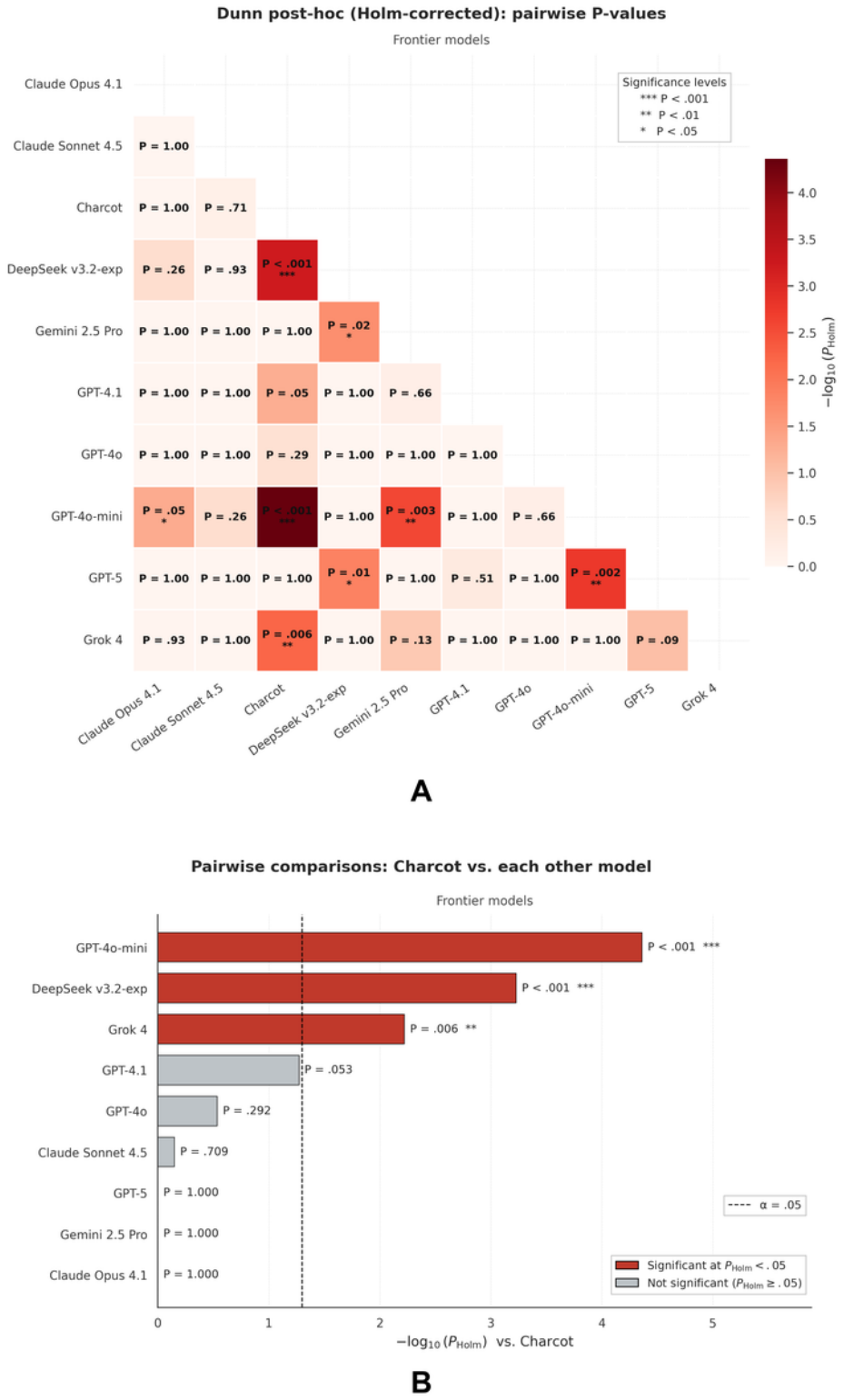
Gemini 2.5 Pro, Claude Opus 4.1, Claude Sonnet 4.5, GPT-4o, and GPT-4.1) did not reach significance after Holm adjustment ($P \geq .05$). Rank-biserial magnitudes are reported in Table 3; the direction of each pairwise comparison is given by the mean ordering in Table 2.

Table 3. Dunn pairwise post hoc (Holm correction); only pairs with $P<.05$ are shown. Effect sizes are reported as rank-biserial magnitudes (|r|); the direction of each comparison is given by the ordering of mean accuracy in The remaining 37 pairs were nonsignificant after Holm adjustment (P≥.05).

Model A	Model B	*P* (Holm)	Rank-biserial, |*r*|
Charcot	GPT-4o-mini	<.001	1.00
Charcot	DeepSeek v3.2-exp	<.001	1.00
GPT-5	GPT-4o-mini	.002	1.00
Gemini 2.5 Pro	GPT-4o-mini	.003	1.00
Charcot	Grok 4	.006	1.00
GPT-5	DeepSeek v3.2-exp	.01	1.00
Gemini 2.5 Pro	DeepSeek v3.2-exp	.02	1.00
Claude Opus 4.1	GPT-4o-mini	.049	1.00

XSL•FO
RenderX

Figure 3. Pairwise post hoc structure (Dunn-Holm). (A) Full Dunn-Holm heatmap of pairwise *P* values across the 10 frontier-panel models; cells below the Holm-corrected .05 threshold are highlighted. Charcot is separable from GPT-4o-mini, DeepSeek v3.2-exp, and Grok 4, but overlaps with the top frontier commercial cluster (GPT-5, Gemini 2.5 Pro, Claude Opus 4.1, Claude Sonnet 4.5, GPT-4o, and GPT-4.1). (B) Charcot-focused detail of the same matrix, restricted to pairs involving Charcot and ordered by adjusted *P* value, to facilitate interpretation of the primary contrast.



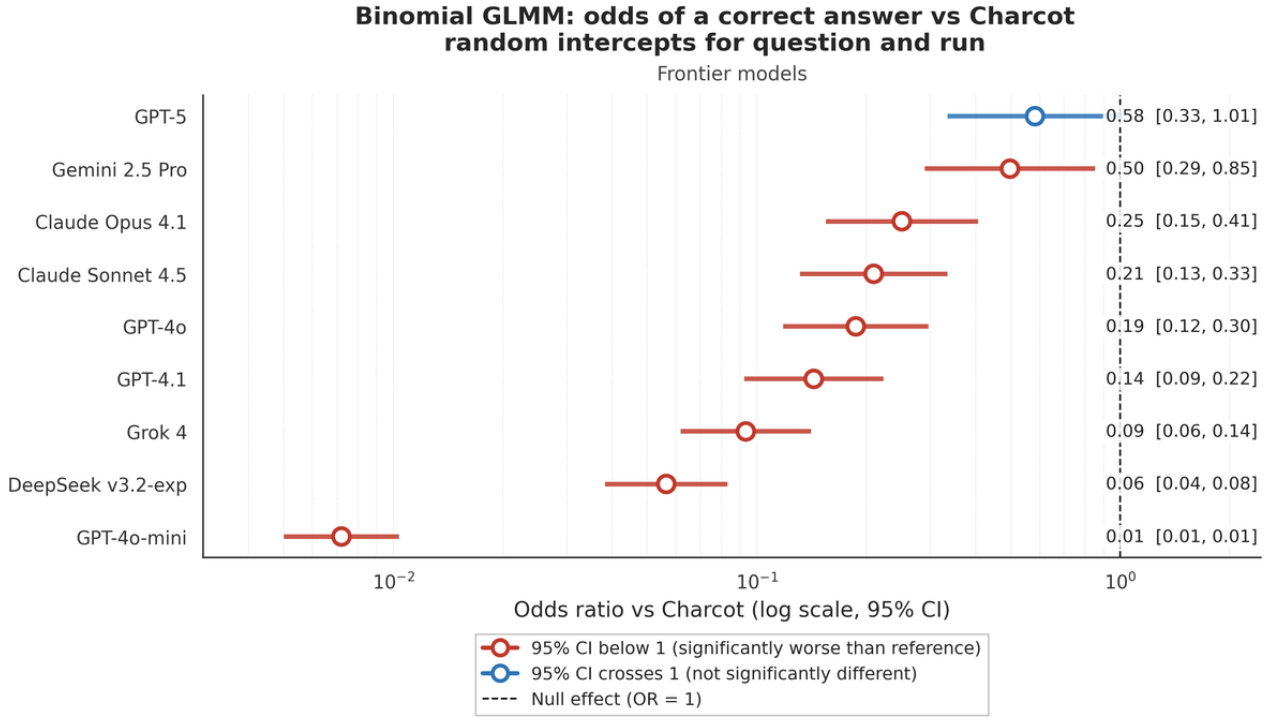
Item-Conditional Sensitivity Analysis (GLMM)

The binomial GLMM (details in "Methods" section) absorbed item-level and run-level heterogeneity. Every comparator model had an OR below 1 against Charcot. GPT-5 crossed unity by the upper CI (OR 0.58, 95% CI 0.33-1.01), which was not significant; all other models had upper 95% CIs below 1: Gemini 2.5 Pro OR 0.50 (95% CI 0.29-0.85), Claude Opus 4.1 OR 0.25 (95% CI 0.15-0.41), Claude Sonnet 4.5 OR 0.21 (95% CI 0.13-0.33), GPT-4o OR 0.19 (95% CI 0.12-0.30), GPT-4.1 OR 0.14 (95% CI 0.09-0.22), Grok 4 OR 0.09 (95% CI 0.06-0.14), DeepSeek v3.2-exp OR 0.06 (95% CI 0.04-0.08), GPT-4o-mini OR 0.007 (95% CI 0.005-0.010). The GLMM (Figure 4) therefore preserved the Kruskal-Wallis ranking after

XSL•FO
RenderX

conditioning on item difficulty and run effects (Figures 3A and B).

Figure 4. Generalized linear mixed model (GLMM) forest plot of frontier-panel odds ratios (ORs) vs Charcot (reference) with 95% CIs. Question and run were entered as crossed random intercepts. GPT-5 is the only comparator whose CI crosses unity.



Per-Question Between-Model Differences

Chi-square tests (model × correct, pooled over runs) followed by Benjamini-Hochberg FDR control at α=.05 identified 34.3% (34/99) of valid items as showing significant between-model differences, considerably above the 5% expected under the global null. The maximum $-\log_1$ (P) reached 6.97, corresponding to items in which model responses split in a near-binary fashion. This concentration of FDR-significant items supports the substantive rather than spurious nature of between-model differences. The full per-question $-\log_1$ (P) plot is provided in Multimedia Appendix 3.

Intermodel Agreement

Agreement on selected alternatives across the 10 frontier-panel models was high: Fleiss κ=0.852 and Krippendorff α=.852 over 99 valid items × 5 runs. The SLM panel (7 models) showed substantially lower agreement (κ=0.508; α=.509), quantifying the greater idiosyncrasy of open-weight and SLMs.

CE Criterion

Using the preregistered default threshold (≥3 models converging on the same incorrect alternative in all 5 runs), 9 of 99 items met the CE criterion in the frontier panel (Q33, Q48, Q66, Q77, Q83, Q87, Q94, Q97, Q100; Table 4 and Figures 5A and B). Item 77 was the strongest case: all 10 frontier-panel models selected alternative B, whereas the preliminary official key was D, consistent with the subsequent INEP rectification to B. Items 48, 33, 97, and 100 showed convergence involving 7 to 8 of 10 models on an alternative other than the preliminary key, and items 66, 83, 87, and 94 showed tighter convergence involving 3 of 10 models.

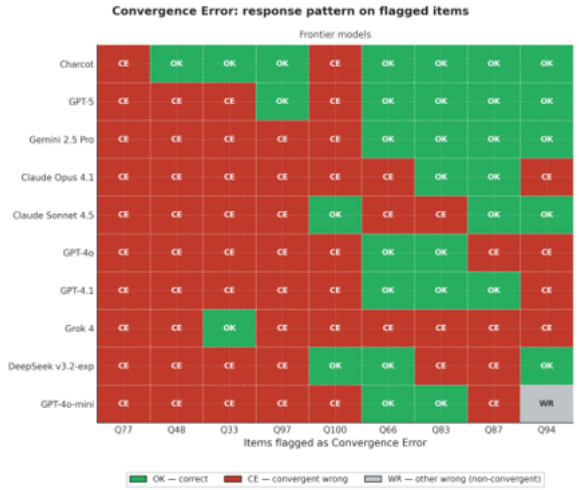
XSL•FO
RenderX

Table 4. Convergence error items at the default threshold (≥3 models × 5 of 5 runs) in the frontier panel.

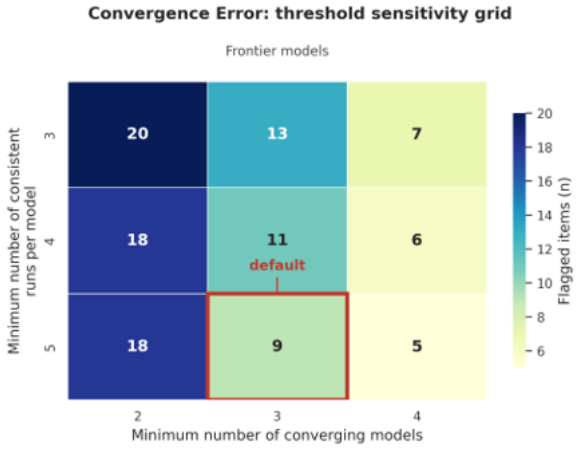
Item	Preliminary key	Convergent wrong	Models, n/N
77[a]	D[a]	B[a]	10/10
48	D	C	8/10
33	C	A	7/10
97	C	D	7/10
100	A	C	7/10
66	A	B	3/10
83	B	C	3/10
87	B	D	3/10
94	A	C	3/10

[a]Subsequently rectified by the Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (INEP; key D to B).

Figure 5. Convergence error (CE; frontier panel). (A) Heatmap across the 10 frontier-panel models for the 9 items flagged at the default threshold (≥3 models × 5 of 5 runs); each cell shows the alternative selected by a model across the 5 runs. Item 77 (all 10 models =B vs preliminary key =D) was subsequently rectified by the Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (INEP). (B) Threshold sensitivity grid: number of CE-positive items as a function of model threshold (x-axis) and run threshold (y-axis). The preregistered default is highlighted; the CE set is stable across adjacent configurations.



A



B

XSL•FO
RenderX

Sensitivity analysis across {2, 3, 4} model thresholds × {3, 4, 5} run thresholds confirmed that the CE set is stable (Table 5 and Figure 3B). The strictest configuration (≥4 models × 5 runs) retained 5 items, all contained within the default set; the most permissive configuration (≥2 models × 3 runs) expanded to 20 items; and the default 9-item set persisted across adjacent configurations.

Table 5. Convergence error sensitivity grid: number of flagged items by threshold combination.

Run threshold	Model threshold		
	≥2 models	≥3 models	≥4 models
≥3 of 5 runs	20[a]	13	7
≥4 of 5 runs	18	11	6
≥5 of 5 runs (default)[b]	18	9[c]	5[d]

[a]Corresponds to the most permissive configuration (≥2 models × 3 of 5 runs).

[b]All 5 items flagged under the strictest configuration are contained in the default set.

[c]Corresponds to the preregistered default (≥3 models × 5 of 5 runs).

[d]Corresponds to the strictest configuration (≥4 models × 5 of 5 runs).

NMRT Analyses

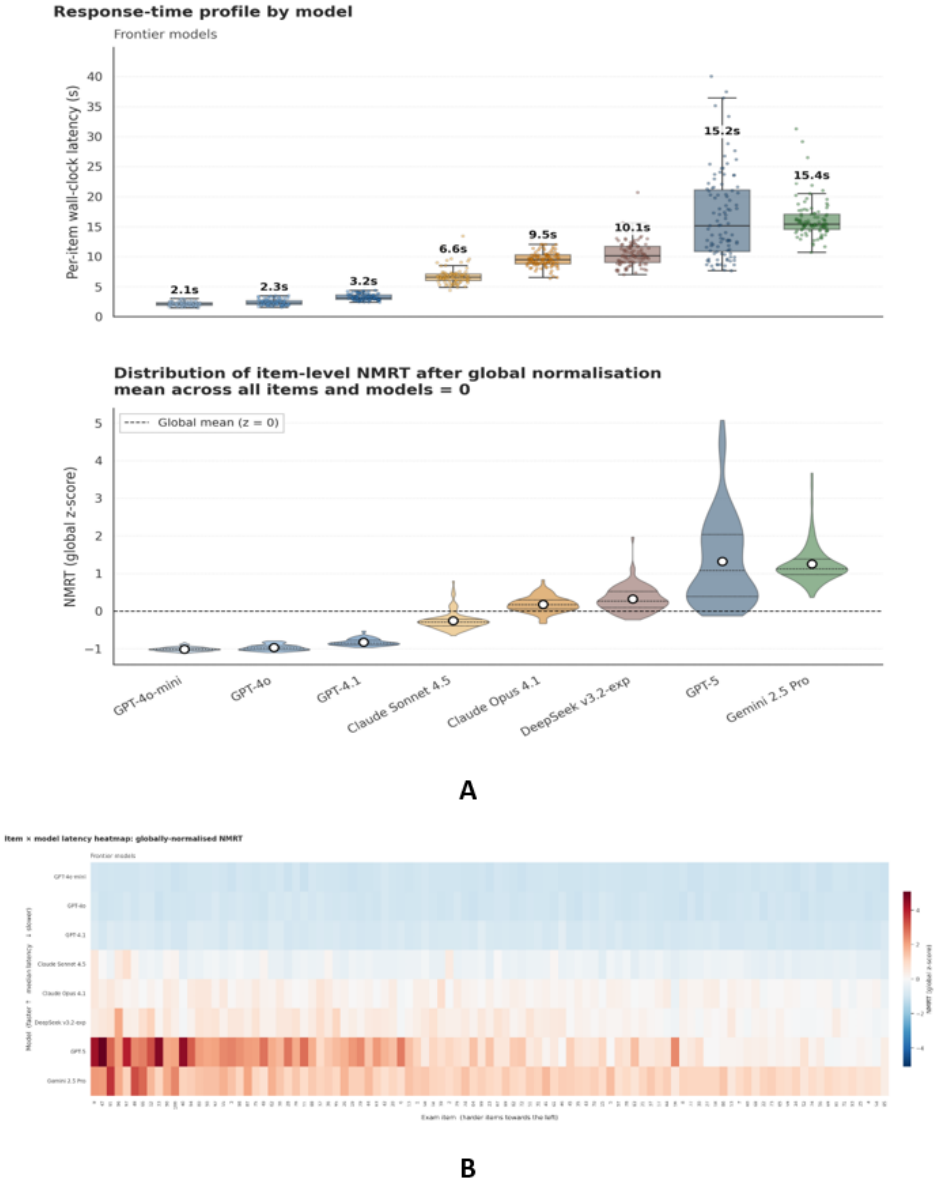
NMRT analyses used the 2 prespecified exclusions (Charcot: different stack; Grok 4: long-tail routing latency; details in the "Methods" section). Among the 8 OpenRouter-routed commercial models retained, mean per-question wall-clock time ranged from 2.11 s (GPT-4o-mini) to 16.64 s (GPT-5), with reasoning-native models (GPT-5 and Gemini 2.5 Pro) clustered near 16 s and nonthinking models (GPT-4o-mini, GPT-4o, and GPT-4.1) clustered near 2-3 s (Table 6 and Figure 6). Among the 8 comparably routed models, the reasoning-native systems (GPT-5 and Gemini 2.5 Pro) occupied the upper tails of both the latency and accuracy distributions. A positive Spearman rank correlation was observed between mean NMRT and mean accuracy (ρ=0.74; *P*=.04; n=8); we interpret this observation as descriptive of orchestration behavior (routing, provider queueing, and reasoning-token generation) under the included models rather than as evidence of a general latency–accuracy relationship, given the small n and the prespecified exclusions required to define the analytic set. The correlation also does not survive reinclusion of Grok 4, whose atypical xAI-via-OpenRouter tail compresses the rank statistic. Per-model internal NMRT *z* scores are reported to avoid confounding item difficulty with baseline model speed.

Table 6. Secondary outcome, normalized mean response time (NMRT; frontier panel; Charcot and Grok 4 excluded per prespecification). Spearman rank correlation (NMRT × accuracy) ρ=0.74; *P*=.04; n=8.

Model	Wall-clock (s), mean (SD)	Median (IQR) (s)	NMRT global *z* score	Questions timed, n
GPT-4o-mini	2.11 (SD 0.34)	2.09 (1.87-2.33)	−1.01	99
GPT-4o	2.37 (SD 0.49)	2.28 (2.00-2.68)	−0.97	99
GPT-4.1	3.26 (SD 0.55)	3.15 (2.87-3.59)	−0.83	99
Claude Sonnet 4.5	6.81 (SD 1.47)	6.59 (6.01-7.07)	−0.26	99
Claude Opus 4.1	9.56 (SD 1.31)	9.51 (8.78-10.31)	+0.18	99
DeepSeek v3.2-exp	10.41 (SD 2.07)	10.12 (9.04-11.71)	+0.32	99
Gemini 2.5 Pro	16.22 (SD 3.22)	15.41 (14.51-17.07)	+1.25	99
GPT-5	16.64 (SD 7.25)	15.15 (10.87-21.12)	+1.32	99

XSL•FO
RenderX

Figure 6. Normalized mean response time (NMRT) frontier panel. (A) NMRT distribution per retained model (violin plot). Charcot and Grok 4 are excluded per prespecification. Reasoning-native models (GPT-5 and Gemini 2.5 Pro) dominate the upper latency tail; nonthinking models cluster near 2-3 s. (B) Per-item NMRT delta heatmap, showing per-model deviation from the item-level NMRT mean, which complements panel A by exposing item-specific latency structure beyond the model-level summary.



Qualitative Analysis of Rationales

For the items with the highest CE or clinical salience signal (Q33, Q48, Q77, Q97, and Q100) we reviewed the generated justifications. In Q33 (postabandonment tuberculosis treatment), Charcot answered correctly in all runs, aligning with Ministry of Health recommendations regarding sensitivity testing and sputum culture in the presence of discordance between the rapid molecular test for tuberculosis and smear microscopy; several generalist models, notably Gemini 2.5 Pro, tended to prematurely restart the basic regimen or repeat the rapid molecular test for tuberculosis as primary management. In Q48 (pleural effusion of probable tuberculous etiology), Charcot prioritized bedside ultrasound and diagnostic thoracentesis in most runs, whereas GPT-5 and Gemini 2.5 Pro concentrated on computed tomography and chest drainage strategies that did not match the expected answer. In Q77 (cultural health knowledge in indigenous populations), all frontier-panel models selected alternative B, later confirmed by INEP rectification. For Q97 (familial short stature), Charcot answered correctly in all runs but occasionally conflated elements of familial short stature and constitutional delay of growth and puberty in its justifications, while Gemini 2.5 Pro and GPT-5 consistently classified the condition as constitutional delay. Finally, in Q100 (criteria for high-risk prenatal care), DeepSeek v3.2-exp and Claude Sonnet 4.5 correctly prioritized third-trimester malaria, while most of the other models (including Charcot in 4 of 5 runs) converged on dichorionic diamniotic twin pregnancy [17].

Open-Weight and SLMs

In the SLM substudy, accuracy spanned roughly 35 pp within a single panel: from 82.22% (407/495) in GPT-OSS 120B (SD 0.55 pp) down to 47.47% (235/495) in Gemma 3 4B (SD 0 pp; Table 7 and Figure 2B). The internal ranking tracked capacity

XSL•FO
RenderX

tier broadly but not monotonically: GPT-OSS 120B (465/495, 82.22%) and Llama 4 Scout (390/495, 78.79%) led the open-weight subgroup, followed by Qwen3 8B (359/495, 72.53%) and Gemma 3 27B (323/495, 65.25%); the 3 SLMs clustered lower, with Ministral 3B at 64.04% (317/495) outperforming the larger Gemma 3 27B, Phi-4 at 59.39% (294/495), and Gemma 3 4B at 47.47% (235/495), well above the 20% chance baseline for a 5-alternative MCQ but far below the frontier-panel floor. Between-run SD was ≤3.15 pp in every SLM-panel model, indicating that the 5-run protocol captured residual stochasticity adequately at this accuracy regime as well.

Table 7. Exploratory subanalysis: accuracy on Exame Nacional de Avaliação da Formação Médica (ENAMED) 2026 across 7 open-weight and small language models (5 runs × 99 valid items). Kruskal-Wallis H6=33.07; *P*<.001; ε2=0.97; Fleiss κ=0.508; Krippendorff α=.509.

Model	Mean (SD)[a]	95% CI	Correct/total, n/N
GPT-OSS 120B	82.22 (0.55)	81.82–82.63	407/495
Llama 4 Scout	78.79 (1.60)	77.58–80.00	390/495
Qwen3 8B	72.53 (3.15)	70.10–74.95	359/495
Gemma 3 27B	65.25 (1.15)	64.24–66.06	323/495
Ministral 3B	64.04 (1.69)	62.83–65.45	317/495
Phi-4	59.39 (0.45)	58.99–59.60	294/495
Gemma 3 4B	47.47 (0.00)	47.47–47.47	235/495

[a]Mean is expressed in percentage (%), while SD is expressed in percentage points (pp).

Normality and variance homogeneity were rejected for the SLM panel as well (Shapiro-Wilk W=0.92; *P*=.02; n=35; Levene W=3.64; *P*=.009; k=7), motivating the use of Kruskal-Wallis for consistency with the primary analysis. A separate Kruskal-Wallis test on the SLM panel yielded a large effect (H_6=33.07; *P*<.001; ε^2=0.97; n=35 runs), consistent with the wide between-model spread. Intermodel agreement was substantially lower than in the frontier panel (Fleiss κ=0.508; Krippendorff α=.509), and 17 items met the default CE criterion under the same ≥3 models × 5-of-5 runs threshold, almost twice the number seen in the frontier panel (Table 8). Importantly, the strongest SLM-panel model (GPT-OSS 120B; 407/495, 82.22%) still ranked below every frontier-panel model except GPT-4o-mini (365/495, 73.74%), and the gap between the best SLM-panel model and the top frontier-panel cluster (Charcot: 480/495, 96.97%; GPT-5: 467/495, 94.34%; Gemini 2.5 Pro: 465/495, 93.94%) was in the 12-15 pp range. Together, the SLM panel quantifies a frontier × subfrontier accuracy gap and a larger error idiosyncrasy (frontier panel κ=0.852 vs SLM κ=0.508) in Portuguese-language board-level medical reasoning.

XSL•FO
RenderX

Table 8. Convergence error (CE) items in the small language model (SLM) panel at the default threshold (≥3 models × 5 of 5 runs). Item 77 converged on alternative B across all 7 SLM-panel models, paralleling the 10-of-10 convergence observed in the frontier panel and the subsequent Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (INEP) rectification (D to B). A total of 5 of the 17 SLM CE items overlap with the frontier-panel CE set (Q33, Q48, Q77, Q83, and Q97), indicating partially shared failure modes between the 2 panels; the remaining 12 items are specific to the SLM panel and reflect the greater error idiosyncrasy of open-weight and SLMs.

Item	Preliminary key	Convergent wrong	Models, n/N
77	D	B	7/7
9	A	B	6/7
48	D	C	6/7
83	B	C	6/7
35	D	B	5/7
97	C	D	5/7
13	C	A	4/7
38	D	C	4/7
44	B	A	4/7
47	C	A	4/7
57	D	A	4/7
91	B	A	4/7
16	B	A	3/7
29	A	B	3/7
33	C	A	3/7
42	D	A	3/7
88	D	A	3/7

Discussion

Principal Findings

In this ENAMED 2026 benchmark, 9 of 10 frontier-panel models exceeded 85% mean accuracy and the Brazilian Portuguese domain-specialized system (Charcot) ranked first at 96.97% (480/495; 95% CI 96.36%-97.58%). Between-model differences were statistically significant (Kruskal-Wallis H_9=47.65; P<.001; ε^2=0.97) and the ranking was preserved in the item- and run-conditioned GLMM sensitivity analysis. Pairwise post hoc testing distinguished Charcot from the subfrontier cluster (GPT-4o-mini, DeepSeek v3.2-exp, and Grok 4) but not from the top frontier commercial cluster (GPT-5, Gemini 2.5 Pro, Claude Opus 4.1, Claude Sonnet 4.5, GPT-4o, and GPT-4.1). A total of 9 items met the preregistered CE criterion, with item 77 showing 10-of-10 convergence subsequently confirmed by INEP rectification. Intermodel agreement was high in the frontier panel (Fleiss κ=0.852) and substantially lower in the SLM open-weight panel (Fleiss κ=0.508), reflecting greater idiosyncrasy of subfrontier systems.

Interpretation

Because Charcot's architecture, training corpus, retrieval components, and inference pipeline are not publicly documented, the performance advantage observed in this study cannot be causally attributed to mechanisms such as Portuguese-language adaptation, domain fine-tuning, retrieval over Brazilian guidelines, prompt engineering, or any combination thereof. Our findings are consistent with prior evidence that domain and language adaptation may improve clinical reasoning in country-specific contexts [6-9,18]. Additionally, it is important to highlight that this study does not constitute mechanistic evidence for specialization but provides black-box comparative evidence under a standardized Portuguese-language protocol. The statistically honest reading is that Charcot is indistinguishable from the top frontier commercial cluster after Holm correction and clearly separable from subfrontier commercial and all open-weight models tested.

The CE construct proved stable across threshold choices (range, 5 to 20 items across a 3 × 3 sensitivity grid) and, in one prospectively flagged case (item 77), anticipated an official INEP rectification. This suggests that CE can serve as a practical item-level signal for examination quality assurance, providing a cheap descriptive screen that surfaces candidates for human review rather than a replacement for expert adjudication. Also, the intermodel agreement metrics (Fleiss κ and Krippendorff α) are complementary to CE, as they summarize the global convergence landscape symmetrically with respect to correctness, whereas CE isolates the asymmetric subset in which convergence occurs on an incorrect alternative.

The positive Spearman correlation between NMRT and accuracy in the 8 retained commercial models (ρ=0.74; P=.04) sits close to the critical value for n=8 and depends on 2 prespecified exclusions (Charcot and Grok 4). For this reason, and because NMRT captures end-to-end orchestration latency rather than inference cost, we interpret this association as descriptive of

XSL•FO
RenderX

orchestration behavior under the included models rather than as evidence of a general latency–accuracy relationship. It is primarily an artifact of reasoning-native systems (GPT-5 and Gemini 2.5 Pro) occupying both the upper latency and upper accuracy tails of the OpenRouter-routed panel, and it does not survive reinclusion of Grok 4, whose atypical xAI-via-OpenRouter queueing tail compresses the rank statistic.

Comparison With Prior Work

Prior Brazilian benchmarks of generalist LLMs on the Revalida and on the Brazilian Society of Cardiology certification exam have reported inconsistent performance, attributed partly to language and local clinical context [6-9]. Our results on ENAMED 2026 extend that evidence to a broader generalist curriculum and to a domain-specialized comparator, and add: (1) effect-size reporting alongside omnibus and pairwise tests, (2) a GLMM sensitivity analysis conditioning on item difficulty and run, (3) prespecified intermodel agreement metrics (Fleiss κ and Krippendorff α), and (4) the CE construct, whose prospective flagging of item 77 we interpret as a positive but anecdotal signal of practical utility.

Methodological Considerations

A unified client layer (OpenRouter) was used for all commercial models. For first-party proprietary models, this is methodologically equivalent to a standardized client library: OpenRouter forwards requests to the canonical provider end points and exposes the same model weights, inference infrastructure, and version pins reported by each provider, while equalizing client-side serialization, retry/timeout handling, and response parsing. Framing OpenRouter as a client-layer control (rather than as a third-party inference rehosting) is supported by documentation and log-level evidence retained in Multimedia Appendix 2. Wall-clock latency measured in this study therefore includes an OpenRouter routing overhead that is systematic across all routed models; relative intermodel NMRT comparisons remain internally valid, whereas absolute NMRT should not be compared with studies using direct provider software development kits.

Temperature=0.0 reduces variability but does not guarantee full determinism [12-14]. The 5-run protocol was chosen as an empirical characterization of residual stochasticity rather than a guarantee of determinism; within-model SD stayed ≤1.53 pp for every model and was 0.00-0.71 pp for 7 of 10 models, supporting the adequacy of 5 runs at this accuracy regime. Reasoning-oriented prompting (eg, chain-of-thought, CoT) was not applied in the primary protocol because (1) reasoning-native frontier models perform internal reasoning by default, making explicit chain-of-thought (CoT) [19] redundant or detrimental to consistency, (2) CoT is known to amplify positional and base-rate biases and to produce post hoc rationalizations in MCQ tasks; and (3) applying CoT selectively to a subset of models would break the standardized-prompt comparability that is the core control of this benchmark. A fully factorial model × prompt-strategy evaluation is outside the scope of this study [20-22].

Open-Weight and SLMs

The SLM substudy extends the frontier-panel picture to a disjoint capacity range (3B to 120B parameters, open weights) under identical prompting and scoring. Three observations follow. First, the strongest open-weight system tested (GPT-OSS 120B; 407/495, 82.22%) was statistically and substantively below every frontier-panel frontier commercial model except GPT-4o-mini, establishing a frontier × subfrontier accuracy gap of roughly 12-15 pp against the top cluster (Charcot, GPT-5, and Gemini 2.5 Pro). Second, this gap widens sharply at the SLM tier: Ministral 3B, Phi-4, and Gemma 3 4B, none of which exceeded 65% (eg, maximum of 323 out of 495), fell approximately 30-50 pp below Charcot. Third, error patterns in the SLM panel were markedly more idiosyncratic than in the frontier panel (Fleiss κ dropping from 0.852 to 0.508, and the number of CE items nearly doubling, 9 to 17), which is consistent with lower-capacity systems converging on different, less stable wrong alternatives rather than sharing blind spots with each other.

Practically, this has 2 implications: open-weight and SLMs are attractive for deployment under data sovereignty, cost, or offline constraints, but at ENAMED 2026 board-level difficulty in Portuguese they remain below both the Brazilian Portuguese domain-specialized system and the top frontier commercial cluster. Their larger error idiosyncrasy also makes ensemble-based audit signals such as CE noisier in this range. Reducing the frontier × subfrontier gap in Portuguese medical reasoning likely requires either scaling, targeted domain and language adaptation of open-weight systems, or a combination. This result is descriptive and specific to ENAMED 2026; generalization to other Brazilian examinations or to deployment-grade clinical tasks requires further benchmarking.

Implications for Educational Assessment

The results support cautious use of multimodel ensembles as an audit signal for item banks and answer keys, particularly when consensus conflicts with the official key, as in item 77 of ENAMED 2026. Model consensus should be treated as a cue for expert review rather than as an arbiter of correctness, and CE should be framed as a descriptive screen, not as an inferential test. Beyond item review, the high-accuracy frontier and specialized models demonstrated here may be useful as curated study aids and as components of simulated assessment environments, subject to the usual caveats about MCQ performance does not generalize to clinical communication or decision-making.

Study Limitations

This study has important limitations. First, it is based on a single examination administered in a single year. Although ENAMED 2026 covers a broad range of topics, it cannot represent all specialties, question formats, or longitudinal shifts in the medical curriculum. Therefore, these results may not generalize to other national examinations or to future ENAMED editions.

Second, performance was assessed exclusively using MCQs. MCQ performance is only a partial proxy for clinical reasoning and does not capture open-ended clinical communication,

XSL•FO
RenderX

multistep decision-making, or empathy. As a result, these scores may overestimate real-world clinical capability.

Third, the internal architecture and training characteristics of Charcot are not publicly documented because it is a proprietary model. Its architecture, training corpus, retrieval components, and inference pipeline are therefore not fully transparent. Consequently, the performance differences observed here cannot be attributed specifically to language adaptation, domain fine-tuning, retrieval augmentation, or prompt design.

Another important limitation relates to the OpenRouter routing layer. Accessing models through OpenRouter introduces an additional network hop compared with direct provider software development kits. While this does not affect accuracy, it adds systematic latency overhead. The unified client layer was used to minimize implementation variance across providers, but absolute NMRT values should not be compared directly with studies that used provider-native APIs.

Decoding-level nondeterminism should also be considered. Although temperature was set to 0.0, some output variability remains possible. The 5-run protocol was designed to characterize this residual stochasticity empirically. Additional runs would likely narrow CIs modestly but would be unlikely to substantially alter the overall ranking of models [23].

A further limitation concerns latency analysis. Grok 4 was excluded from NMRT analyses because of an extreme long-tail latency distribution when accessed via xAI through OpenRouter. This exclusion was prespecified and documented, but it means that NMRT comparisons are valid only for the 8 more homogeneously routed commercial models included in that analysis.

In addition, all experiments were conducted from a single orchestrator host in a single geographical region. Results may differ under alternative client-side infrastructure, network paths to provider regions, or rate-limit conditions.

Finally, rationale analysis was descriptive rather than formally validated. The qualitative review of model justifications did not use a formal faithfulness metric. Therefore, findings related to reasoning patterns should be interpreted as illustrative and hypothesis-generating rather than definitive.

Conclusion

A Brazilian Portuguese domain-specialized system achieved the highest rank on ENAMED 2026, proving statistically indistinguishable from the top frontier commercial cluster. While the lack of architectural disclosure means these results serve as black-box comparative evidence rather than mechanistic proof of specialization, the system's performance remained significantly above subfrontier and open-weight models. The stability of the CE criterion, which accurately flagged a rectified item (Q77), further supports its adoption as a robust tool for examination quality assurance.

Data Availability

All datasets (Exame Nacional de Avaliação da Formação Médica [ENAMED] 2026 items, the official Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira [INEP] answer keys, the OpenRouter provider logs, and the raw JSON outputs from all model runs) are provided in Multimedia Appendix 2, and the analysis code is openly available in the repository referenced in Multimedia Appendix 1. No access restrictions apply to these datasets; the proprietary nature of the Charcot model concerns its architecture and is disclosed in the Conflicts of Interest and Study Limitations sections.

Funding

This study was supported by Voa Health, which provided computational infrastructure and resources for data analysis. The funding source had no role in the study design, data collection, analysis, or interpretation of the data.

Authors' Contributions

FLFS conceptualized the study, developed the methodology, and prepared the original draft. JVBS, EAR, DF, and PABP conducted the experimental runs and collected the data. EAR and DF were responsible for software development, data curation, and resource acquisition, including application programming interface (API) access. DF also carried out the critical review and editing. MHV performed the formal statistical analysis and mathematical validation. GLM and TPM supervised the overall project, validated the findings, and completed the final editing.

Conflicts of Interest

Authors GLM, EAR, and DF hold equity (shares) in Voa Health, the company that developed the Charcot model evaluated in this study. EAR and DF are employed as artificial intelligence (AI) engineers at Voa Health. Author FLFS receives financial compensation from Voa Health for marketing and social media promotion of the Voa Health platform. These financial and professional relationships constitute a conflict of interest. The remaining authors declare no conflicts of interest.

Multimedia Appendix 1
The OpenRouter model identifiers, underlying provider versions, access windows, and inference parameters used for each model in the ENAMED 2026 benchmark.
[DOCX File , 33 KB - mededu_v12i1e89839_app1.docx]

XSL•FO
RenderX

Multimedia Appendix 2
Complete raw data and reproducibility artefacts for the ENAMED 2026 benchmark study. All artefacts are hosted in a persistent, read-only shared repository.
[DOCX File , 8 KB - mededu_v12i1e89839_app2.docx]

Multimedia Appendix 3
Additional statistical outputs: main panel.
[DOCX File , 798 KB - mededu_v12i1e89839_app3.docx]

Multimedia Appendix 4
Additional outputs: exploratory open-weight and small language model substudy.
[DOCX File , 2181 KB - mededu_v12i1e89839_app4.docx]

References

1. Thirunavukarasu AJ, Ting DSJ, Elangovan K, Gutierrez L, Tan TF, Ting DSW. Large language models in medicine. Nat Med 2023;29(8):1930-1940. [doi: 10.1038/s41591-023-02448-8] [Medline: 37460753]
2. Kung TH, Cheatham M, Medenilla A, Sillos C, De Leon L, Elepaño C. Performance of chatGPT on USMLE: potential for AI-assisted medical education using large language models. PLOS Digit Health 2023;2(2):e0000198 [FREE Full text] [doi: 10.1371/journal.pdig.0000198] [Medline: 36812645]
3. Moëll B, Sand Aronsson F. Harm reduction strategies for thoughtful use of large language models in the medical domain: perspectives for patients and clinicians. J Med Internet Res 2025;27:e75849 [FREE Full text] [doi: 10.2196/75849] [Medline: 40712151]
4. Chen Y, Huang X, Yang F, Lin H, Lin H, Zheng Z. Performance of ChatGPT and bard on the medical licensing examinations varies across different cultures: a comparison study. BMC Med Educ 2024;24(1):1372 [FREE Full text] [doi: 10.1186/s12909-024-06309-x] [Medline: 39593041]
5. Zong H, Wu R, Cha J, Wang J, Wu E, Li J. Large language models in worldwide medical exams: platform development and comprehensive analysis. J Med Internet Res 2024;26:e66114 [FREE Full text] [doi: 10.2196/66114] [Medline: 39729356]
6. Alonso I, Oronoz M, Agerri R. MedExpQA: multilingual benchmarking of large language models for medical question answering. Artif Intell Med 2024;155:102938 [FREE Full text] [doi: 10.1016/j.artmed.2024.102938] [Medline: 39121544]
7. Bruneti Severino JV, Basei de Paula PA, Berger MN, Loures FS, Todeschini SA, Roeder EA. Benchmarking open-source large language models on Portuguese revalida multiple-choice questions. BMJ Health Care Inform 2025;32(1):e101195 [FREE Full text] [doi: 10.1136/bmjhci-2024-101195] [Medline: 40000063]
8. Bessa RF, de Oliveira AC, Bessa RF, Sousa DL, Alves R, Barbosa A. Performance comparison of large language models on Brazil's medical revalidation exam for foreign-trained graduates. Appl Sci 2025;15(13):7134. [doi: 10.3390/app15137134]
9. Severino JVB, Berger MN, de Paula PAB, Loures FS, Todeschini SA, Roeder EA. Performance benchmarking of open-source large language models on the Brazilian society of cardiology's certification exam. Int J Cardiovasc Sci 2025;38:e20240231. [doi: 10.36660/ijcs.20240231]
10. Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (INEP). Exame Nacional de Avaliação da Formação Médica (Enamed). URL: https://www.gov.br/pt-br/servicos/fazer-o-exame-nacional-de-avaliacao-da-formacao-medica [accessed 2026-05-15]
11. OpenRouter. URL: https://openrouter.ai [accessed 2026-05-15]
12. Ouyang S, Zhang JM, Harman M, Wang M. An empirical study of the non-determinism of ChatGPT in code generation. ACM Trans Softw Eng Methodol 2025;34(2):1-28. [doi: 10.1145/3697010]
13. Atil B. LLM stability: a detailed analysis with some surprises. ArXiv. Preprint posted online on August 6, 2024 :1-11 [FREE Full text]
14. Song Y, Wang G, Li S, Lin BY. The good, the bad, and the greedy: evaluation of LLMs should not ignore non-determinism. ArXiv. Preprint posted online on July 15, 2024 :1-10 [FREE Full text]
15. Tomczak M, Tomczak E. The need to report effect size estimates revisited. An overview of some recommended measures of effect size. Trends Sport Sci 2014;1(21):19-25 [FREE Full text]
16. Miller E. Adding error bars to evals: a statistical approach to language model evaluations. ArXiv. Preprint posted online on November 1, 2024 [FREE Full text]
17. High-risk pregnancy manual. Brazil Ministry of Health. 2022. URL: https://bvsms.saude.gov.br/bvs/publicacoes/manual_gestacao_alto_risco.pdf [accessed 2026-05-25]
18. Wang Z, Wang H, Danek B, Li Y, Mack C, Arbuckle L. A perspective for adapting generalist AI to specialized medical AI applications and their challenges. NPJ Digit Med 2025;8(1):429 [FREE Full text] [doi: 10.1038/s41746-025-01789-7] [Medline: 40646157]

XSL•FO
RenderX

19. Wei J, Wang X, Schuurmans D, Bosma M, Ichter B. Chain-of-thought prompting elicits reasoning in large language models. 2022 Presented at: NIPS'22: 36th International Conference on Neural Information Processing Systems; November 28-December 9, 2022; New Orleans p. 24824-24837. [doi: 10.52202/068431-1800]
20. Moore K, Roberts J, Pham T, Fisher D. Reasoning beyond bias: a study on counterfactual prompting and chain of thought reasoning. ArXiv. Preprint posted online on September 6, 2024 :1-8 [FREE Full text]
21. Turpin M, Michael J, Perez E, Bowman SR. Language models don't always say what they think: unfaithful explanations in chain-of-thought prompting. 2023 Presented at: NIPS'23: 37th International Conference on Neural Information Processing Systems; December 10-16, 2023; New Orleans p. 74952-74965. [doi: 10.52202/075280-3275]
22. Arcuschin I, Janiak J, Krzyzanowski R, Rajamanoharan S, Nanda M, Conmy A. Chain-of-thought reasoning in the wild is not always faithful. ArXiv. Preprint posted online on June 17, 2025 [FREE Full text]
23. Blackwell RE, Barry J, Cohn AG. Towards reproducible LLM evaluation: quantifying uncertainty in LLM benchmark scores. ArXiv. Preprint posted online on October 4, 2024 :1-9 [FREE Full text]

Abbreviations

API: application programming interface
CE: convergence error
CoT: chain-of-thought
ENAMED: Exame Nacional de Avaliação da Formação Médica
FDR: false discovery rate
GLMM: generalized linear mixed model
INEP: Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira
LLM: large language model
MCQ: multiple-choice question
NMRT: normalized mean response time
OR: odds ratio
pp: percentage points
SLM: small language model

Edited by J Gentges; submitted 20.Dec.2025; peer-reviewed by N Aldahoul, E Abdulwahed; comments to author 31.Mar.2026; revised version received 28.Apr.2026; accepted 08.May.2026; published 29.May.2026.

Please cite as:
Fernandes da Silva FDL, Roeder EA, Bruneti Severino JV, Nespolo Berger M, Basei de Paula PA, Ferreira D, Han Veiga M, de Moraes TP, Lenci Marques G
Performance of Large Language Models on the Brazilian National Medical Education Examination: Comparative Benchmark Study
JMIR Med Educ 2026;12:e89839
URL: https://mededu.jmir.org/2026/1/e89839
doi:10.2196/89839
PMID:

©Francys de Luca Fernandes da Silva, Eduardo Augusto Roeder, João Victor Bruneti Severino, Matheus Nespolo Berger, Pedro Angelo Basei de Paula, Davi Ferreira, Maria Han Veiga, Thyago Proença de Moraes, Gustavo Lenci Marques. Originally published in JMIR Medical Education (https://mededu.jmir.org), 29.May.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Education, is properly cited. The complete bibliographic information, a link to the original publication on https://mededu.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Impact of an Ambient AI Scribe on Medical Student Objective Structured Clinical Examination Notes: Nonrandomized Clinical Trial

Jaideep S Talwalkar[1], MD; Donald S Wright[2], MD, MHS; Lee H Schwamm[3], MD; Gary Leydon[4], BS; Veronika Shabanova[5], PhD

[1]Departments of Internal Medicine and Pediatrics, School of Medicine, Yale University, New Haven, CT, United States

[2]Departments of Emergency Medicine and Biomedical Informatics and Data Science, School of Medicine, Yale University, New Haven, CT, United States

[3]Departments of Neurology and Biomedical Informatics and Data Science, School of Medicine, Yale University, New Haven, CT, United States

[4]School of Medicine, Yale University, New Haven, CT, United States

[5]Departments of Pediatrics and Biostatistics, School of Medicine, Yale University, New Haven, CT, United States

Corresponding Author:
Jaideep S Talwalkar, MD
Departments of Internal Medicine and Pediatrics
School of Medicine
Yale University
367 Cedar Street, Bldg D
New Haven, CT, 06510
United States
Phone: 1 203 737 4190
Fax: 1 203 737 4199
Email: jaideep.talwalkar@yale.edu

Abstract

Background: Ambient artificial intelligence (AI) scribes for chart documentation have seen rapid adoption in clinical practice, but their educational impact on medical students has not been described.

Objective: The purpose of this study was to determine the impact of an AI scribe on preclerkship medical student note writing.

Methods: In this prospective nonrandomized pretest-posttest study, all first-year medical students (N=104) at a single US medical school submitted "human-only" notes based on a summative objective structured clinical examination station in May 2025. An AI scribe generated independent AI notes after the objective structured clinical examination from recorded audio. A subgroup of students (n=47) consented to complete a second "hybrid" note by revisiting their human-only notes and incorporating AI notes as perceived necessary, followed by a brief survey about the AI notes. Trained, blinded fourth-year medical student raters were randomly assigned to score all notes on 10 elements using QNOTE acceptability criteria (0="unacceptable," 50="partially acceptable," and 100="fully acceptable"). A post hoc, exploratory element-level review was then conducted.

Results: Across all elements, median evaluation scores of human-only notes were high (range 81.3-100) and similar between students who submitted "hybrid" notes and those who did not. In paired analyses between "human-only" and "hybrid" notes, the only notable element-level change was a decline in "chief complaint" scores (P=.05). Symptom duration was mentioned in the chief complaint section in 17% (8/47) of the AI notes. No score differences were observed in QNOTE elements requiring documentation of pertinent findings and prioritized lists. Participants agreed that the AI note "was more concise than my note" (37/47, 78.7%) and would be "helpful as a first draft" (31/47, 66%); 55.3% (26/47) agreed that the AI note "left out important details," and 21.3% (10/47) agreed that the AI note "may reduce my ability to learn how to write a good note."

Conclusions: Interaction with AI notes among preclerkship medical students had little impact on the quality of "hybrid" notes. Chief complaint scores likely declined due to conciseness in AI notes that often omitted symptom duration. Our findings suggest that, among students who predominantly wrote close to fully acceptable "human-only" notes, there was no detriment to clinical reasoning, and students were discerning in balancing AI's conciseness and its omissions. The lack of impact on note quality may have been due to the workflow used in this study, in which students were required to generate independent judgments before exposure to AI-generated content. Future work must explore longitudinal use of such tools using standard workflows observed

XSL•FO
RenderX

in clinical settings, where AI notes serve as true first drafts. AI scribes could enhance students' own note writing, especially for lower-performing students, although educational safeguards are necessary given the potential for harm due to overreliance on automated systems.

(JMIR Med Educ 2026;12:e88264) doi:10.2196/88264

KEYWORDS

ambient scribe; artificial intelligence; AI; documentation; medical student education; note writing; objective structured clinical examination; postencounter note; standardized patient

Introduction

The use of ambient artificial intelligence (AI) scribes for chart documentation has seen rapid adoption among practicing physicians [1,2]. In clinical settings, AI scribes listen in the background during a patient visit and automatically create a structured medical note from the conversation [3]. Multiple recent studies have demonstrated that the use of AI scribes among practicing physicians is associated with outcomes such as reductions in time spent on documentation, cognitive load, task load, and burnout [1,4-8].

Given the degree to which documentation burden can negatively impact the graduate medical education training experience [9], Wright et al [10] conducted a pilot of AI scribes with resident physicians. Similar to the experiences of practicing physicians, residents found that AI scribes decreased task load and documentation burden.

Documentation skills are broadly taught and assessed in undergraduate medical education curricula [11-14], and the Association of American Medical Colleges lists the ability to document a clinical encounter as 1 of 13 core entrustable professional activities that medical students should be able to perform when entering residency [15]. Additionally, adult learning theory in medical education supports experiential learning, in which students engage in hands-on practice with the concepts and tools they will encounter in the workplace [16,17].

Despite the emphasis on teaching documentation skills, acceptance of the value of experiential learning in medical education, and rapid growth of AI in undergraduate medical education curricula overall [18], to our knowledge, the use of AI scribes with medical students has not been described previously.

Documentation in medical education serves not only as a communication tool but also as a cognitive exercise through which students organize data, generate differential diagnoses, and demonstrate clinical reasoning skills [19]. Introducing automation into this developmental process may have complex effects. Automation can reduce cognitive load, which may alleviate working memory constraints during demanding tasks, thereby facilitating learning and performance [20,21]. However, the cognitive load associated with tasks such as clinical reasoning is desirable in medical education for learners to develop essential skills [20].

The possibility of harm with AI scribes has led some experts to advise against early introduction of this technology in medical training [22], although the educational implications remain unclear. Shielding current medical students from technology that they will almost certainly encounter in practice may be pedagogically unsound. In the absence of empirical evidence, there is an urgent need to understand the educational effects of AI scribes. Therefore, the purpose of this study was to examine the impact of an ambient AI scribe on the quality of preclerkship medical student note writing and explore how a human-AI hybrid workflow influences note structure and content, specifically as relates to clinical reasoning.

Methods

Educational Setting

This prospective nonrandomized pretest-posttest study took place at a single US medical school (Yale School of Medicine, New Haven, Connecticut) in May 2025 as part of the required summative objective structured clinical examination (OSCE) that occurs at the end of the first-year Clinical Skills Course for all Yale medical students. During the OSCE, students conduct 2 complete histories and physical examinations in back-to-back 45-minute stations. Students then have 48 hours to submit a complete clinical note on their second case. All students participate in the same 2 cases in different orders, and all encounters are digitally recorded using LearningSpace (Elevate Healthcare). Passing is achieved by reaching predetermined thresholds on the history and physical examination and submitting a note. While submission of the note is required, the content and quality of the note do not factor into the OSCE score.

Prior to the study, as part of the Clinical Skills Course, students had 7 formative medical interviewing workshops with standardized patients (SPs), 10 formative physical examination workshops with peers and SPs, and 15 clinical sessions primarily with inpatients to apply these skills. Prior note-writing education and practice included a 9-minute video tutorial, a brief handout on note writing, and at least 2 previous notes on which students received narrative formative feedback from instructors.

For the May 2025 OSCE, an ambient AI scribe (Abridge Inc) was used for all students in preparation for a future classroom discussion that was not part of the study. To avoid disruption during a high-stakes assessment, the ambient AI scribe note (hereafter referred to as the "AI note") was created using the digital recording immediately after the OSCE. Abridge-enabled mobile devices were used to create a separate AI note for each encounter using the audio extracted from each student's original encounter. This process was previously checked for quality control in conjunction with the health system's technology team

XSL•FO
RenderX

and Abridge technical support. Quality control consisted of comparison of transcript content with source audio using multiple existing recordings within LearningSpace to ensure fidelity of transcription. Students had not used ambient AI scribes for note writing during any Clinical Skills Course sessions prior to the OSCE.

Participants

As part of standard curricular requirements, all first-year students must complete the OSCE, including submission of a complete note. The study was described to students during an in-class announcement followed by an emailed description. Students were invited to complete a second note (described further below) as part of the study. Submission of the second note was voluntary and had no impact on OSCE grading or advancement decisions. All people involved in OSCE scoring or student advancement decisions (eg, SPs, note raters, and course faculty) were blinded regarding student participation status.

Ethical Considerations

This study was granted approval by the Yale School of Medicine Committee to Review Student Participation in Research on January 14, 2025, and an exemption from full review due to the minimal-risk educational nature of the project by the Yale University Human Research Protection Program on January 23, 2025 (protocol ID 2000039478). Those interested in completing a second note were instructed to submit a consent form to a course administrator who was not involved in grading or advancement decisions. Participation status of students was unknown to the investigators, and data were deidentified prior to analysis. Student identity was known to the raters to facilitate assessment as the study was embedded within a curricular activity. Students who completed a second note received a US $25 meal voucher upon completion of the study.

Note Assessment Instrument and Rater Training

Five assessment instruments were reviewed [23-27], and QNOTE [23] was selected due to alignment with the note-writing expectations of first-year medical students (ie, sufficient detail to demonstrate understanding of note structure) and the task at hand (ie, initial visit with an adult patient in an outpatient setting). As described by Burke et al [23], QNOTE consists of 12 elements: “chief complaint,” history of present illness, problem list, past medical history, medications, adverse drug reactions, social and family history, review of systems, physical findings, assessment, plan of care, and follow-up information. Each element comprises 1 to 4 components. Each component is rated as “fully acceptable,” “partially acceptable,” or “unacceptable or missing,” which correspond to numerical values of 100, 50, and 0, respectively, for the purpose of scoring. Following the methodology described in the QNOTE development, element-level scores were calculated as the means across their respective components.

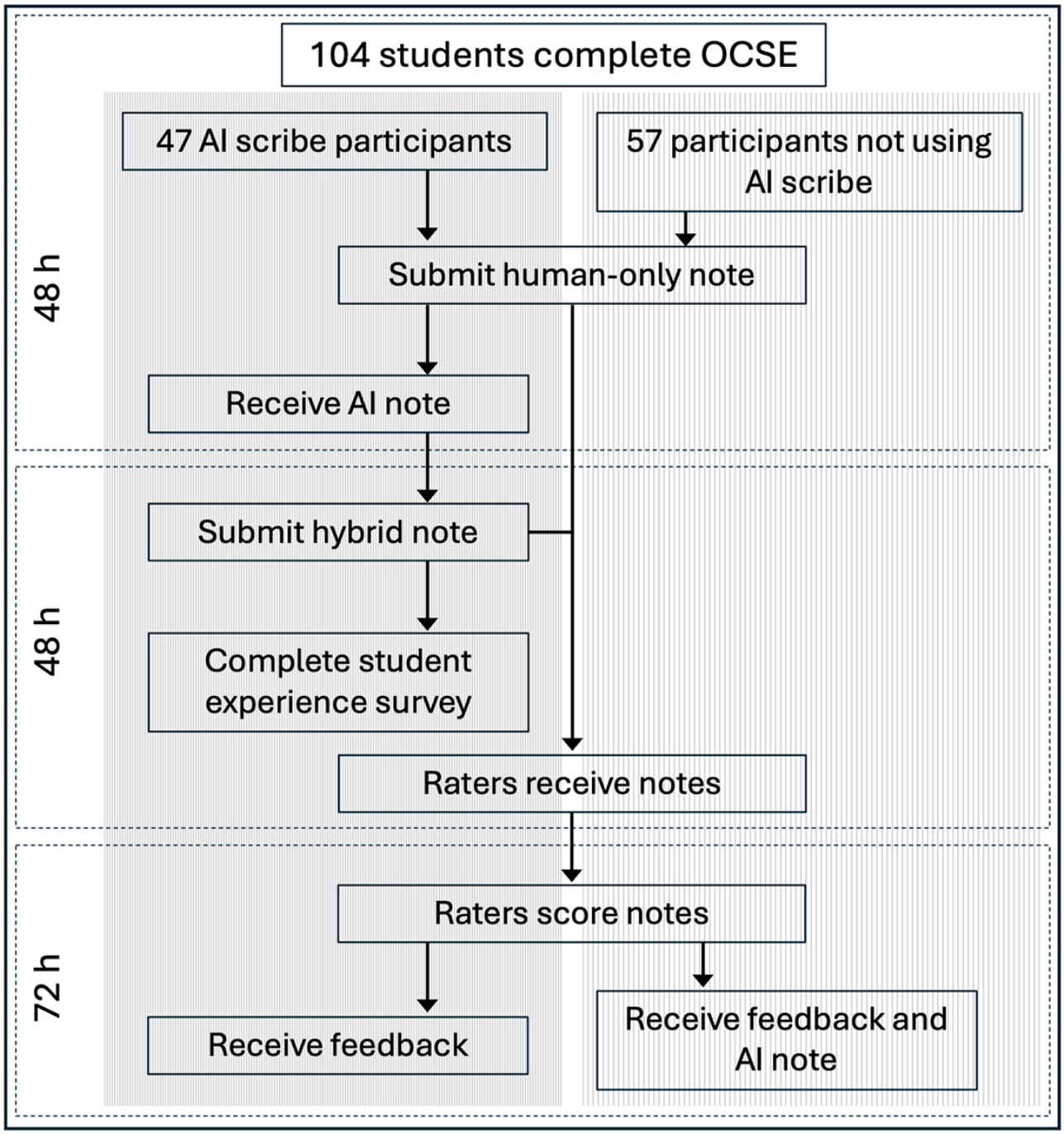
Eight fourth-year medical students were recruited as raters of notes via email announcement from a course administrator. Raters received detailed instructions via email that included teaching materials about note writing that had been provided to the first-year students, 3 sample notes from the prior year’s OSCEs, and the QNOTE assessment instrument. Sample notes were chosen by the course director as examples of note quality to reflect each of the 3 overall note score rating categories described by Burke et al [23]: “fully acceptable,” “partially acceptable,” or “unacceptable.” Each of the 8 raters independently scored the 3 sample notes using QNOTE, and then they met as a group for additional rater training. The goal of rater training was to achieve good interrater reliability, defined as an intraclass correlation coefficient of at least 0.80 [28]. During a 75-minute training meeting, raters viewed their scores relative to each other and discussed their rationale and approaches to interpreting QNOTE elements. Specifically, the raters negotiated a shared mental model for “physical findings” as the OSCE required a complete physical examination, whereas QNOTE prioritized the focused, hypothesis-driven examination more commonly used in practice. Additionally, the group decided to omit 2 QNOTE elements (problem list and follow-up information) for the purposes of this study because these elements were not explicitly included in the teaching materials and typically appeared in health system notes automatically through electronic health record macros. Their exclusion was determined during rater calibration to ensure alignment between curricular expectations and assessment criteria. Therefore, the modified version of QNOTE used for this study consisted of 10 elements.

Note Writing and Assessment Workflow

All notes were requested, submitted, and assessed through a Microsoft Power Automate workflow programmed by an administrator to ensure blinding of student participation status and note type to raters and course faculty (Figure 1). Only Microsoft Word documents were permitted in the workflow. We used a cognitive forcing strategy [29] requiring students to generate notes independently before exposure to AI-generated content. Immediately following the OSCE, all 104 students received a personalized email with a link to submit their required “human-only note” based on their second case from the OSCE. Students were permitted to use any resources available to them to complete this human-only note. Upon submission of the human-only note, AI scribe participants received another personalized email with the AI note for the same case and a link to submit a “hybrid” note within 48 hours. For this hybrid note, participants were instructed to review their human-only and AI notes and submit a new note based on these documents. Thus, in a hybrid note, students would decide whether and how to update their human-only note with the information in the AI note. Upon submission of the hybrid note, raters received separate emails prompting them to score the human-only and hybrid notes using QNOTE in an electronic form.

XSL•FO
RenderX

Figure 1. Note writing and assessment workflow. All 104 first-year medical students completed notes documenting a simulated clinical encounter. A subset of students engaged in an opportunity to modify their notes into a "hybrid note" after reviewing the output of an ambient artificial intelligence (AI) scribe. All notes were scored by blinded raters. All workflow steps were automated by an administrator to ensure blinding of student participation status and note type to raters and course faculty. OSCE: objective structured clinical examination.



Each note was scored once by 1 of 8 trained raters. For students who submitted both human-only and hybrid notes, the same rater scored both notes. Raters were blinded, meaning that they were unaware that the 2 notes were paired or that one represented a hybrid condition. Simultaneously, raters received the human-only notes from students who had not consented to create a hybrid note. These human-only notes were also rated via the same process. All notes were evenly distributed to raters through this automated workflow without identifiers indicating note type or participation status.

After primary scoring, we conducted a post hoc, exploratory, qualitative review of the chief complaint section of the AI notes (n=47). An unblinded investigator (JST) reviewed the text to identify recurrent patterns and to aid interpretation of the element-level results from the primary scoring. The review was descriptive and did not use a formal coding framework.

Survey Details

Following submission of the hybrid note, participants completed an anonymous 6-item student experience survey assessing their perceptions of the AI note. Items were rated on a 5-point Likert scale ranging from "strongly disagree" to "strongly agree." The survey was administered electronically through the previously described automated workflow.

XSL•FO
RenderX

Statistical Analysis

An intraclass correlation coefficient for interrater reliability across all 10 QNOTE elements was calculated using a 2-way random-effects model [30,31]. Descriptive summaries of QNOTE element scores consisted of means with SDs and medians with IQRs. Comparisons of QNOTE element scores between groups (AI scribe participants vs nonparticipants) were performed using the Wilcoxon rank-sum test for independent samples. Within-group (hybrid vs human-only notes among AI scribe participants) comparisons of QNOTE element scores were implemented using Wilcoxon signed-rank tests for paired samples. A 2-sided type I error of .05 was used in statistical comparisons. We did not adjust for multiple comparisons as the statistical hypotheses corresponding to each QNOTE element did not reflect disjunction testing for a joint null hypothesis [32]. Each QNOTE element represented a distinct, prespecified construct rather than components of a single joint null hypothesis. Separate composite QNOTE scores for human-only and hybrid notes were calculated for each AI scribe participant as the mean of all 10 QNOTE elements. Composite QNOTE scores for human-only notes were then categorized into quartiles based on their empirical distribution and also plotted against their paired composite QNOTE scores for hybrid notes. Percentages of students in the lowest quartile of human-only composite QNOTE scores who improved on their hybrid notes and percentages of students whose scores declined by at least 20 points were summarized. Students' responses to the student experience survey were presented as horizontal stacked bars of percentages for each Likert scale category. No a priori sample size or statistical power analyses were conducted as we only had access to the census of the first-year class and we did not have prior data to base our effect sizes on a power calculation. Analyses were conducted in R (version 4.5.0; R Foundation for Statistical Computing).

Results

All 104 students submitted the required human-only write-up. A total of 45.2% (n=47) of the students consented to write a hybrid note; all completed this assignment and the student experience survey (Figure 1). Across all QNOTE elements, median evaluation scores of human-only notes were high (range 81.3-100) and were similar between students who submitted hybrid notes (n=47, 45.2%) and those who did not (n=57, 54.8%; Table 1).

XSL•FO
RenderX

Table 1. Comparison of QNOTE element scores on human-only and hybrid notes.

QNOTE element (0-100)	Human-only note not using AI[a] scribe (n=57)	Participants using the AI scribe (n=47)		*P* value	
		Human-only note	Hybrid note	Unpaired samples[b]	Paired sample[c]
Chief complaint					
Mean (SD)	72.8 (34.2)	69.1 (30.5)	59.6 (34.0)	—[d]	—
Median (IQR)	100.0 (50.0-100)	50.0 (50.0-100)	50 (50.0-100)	.41	.05[e]
History of present illness					
Mean (SD)	82.5 (19.9)	87.5 (14.3)	87.8 (14.4)	—	—
Median (IQR)	87.5 (75.0-100)	87.5 (87.5-100)	87.5 (87.5-100)	.21	.86
Past medical history					
Mean (SD)	82.0 (27.1)	88.6 (21.0)	87.5 (21.3)	—	—
Median (IQR)	87.5 (75.0-100)	100.0 (87.5-100)	87.5 (87.5-100)	.24	.34
Medications					
Mean (SD)	83.0 (27.7)	91.1 (13.4)	88.7 (20.0)	—	—
Median (IQR)	100.0 (83.3-100)	100.0 (83.3-100)	100.0 (83.3-100)	.35	.67
Adverse drug reactions					
Mean (SD)	87.4 (25.4)	90.8 (21.6)	86.5 (28.6)	—	—
Median (IQR)	100.0 (83.3-100)	100.0 (83.3-100)	100.0 (83.3-100)	.73	.17
Social and family history					
Mean (SD)	85.8 (18.6)	92.3 (8.40)	90.4 (9.55)	—	—
Median (IQR)	90.0 (80.0-100)	90.0 (90.0-100)	90.0 (80.0-100)	.30	.14
Review of systems					
Mean (SD)	72.5 (34.0)	79.4 (27.0)	77.3 (32.1)	—	—
Median (IQR)	83.3 (66.7-100)	83.3 (83.3-100)	83.3 (66.7-100)	.58	.80
Physical findings					
Mean (SD)	88.9 (21.0)	92.2 (12.0)	88.7 (19.4)	—	—
Median (IQR)	100 (83.3-100)	100 (83.3-100)	100.0 (83.3-100)	.77	.29
Assessment					
Mean (SD)	73.2 (28.5)	79.3 (22.5)	78.7 (21.6)	—	—
Median (IQR)	75.0 (62.5-100)	87.5 (75.0-100)	87.5 (68.8-100)	.42	.97
Plan of care					
Mean (SD)	77.6 (24.7)	81.9 (22.7)	76.6 (25.6)	—	—
Median (IQR)	87.5 (75.0-100)	87.5 (75.0-100)	87.5 (75.0-93.8)	.39	.26

[a]AI: artificial intelligence.

[b]Comparisons of QNOTE element scores on all human-only (n=57, participants not using AI scribe) and (n=47, participants using AI scribe) write-ups using the Wilcoxon rank-sum test.

[c]Comparisons of QNOTE element scores on human-only and hybrid write-ups among AI scribe participants using the Wilcoxon signed-rank test.

[d]Not applicable as we did not compare means of individual domains due to the skewed distributions.

[e]The shift in distribution is due to fewer "fully acceptable" (16/47, 34% vs 21/47, 44.7%) and more "unacceptable" (7/47, 14.9% vs 3/47, 6.4%) hybrid notes as compared to human-only notes for the paired sample comparison of chief complaint element-level scores.

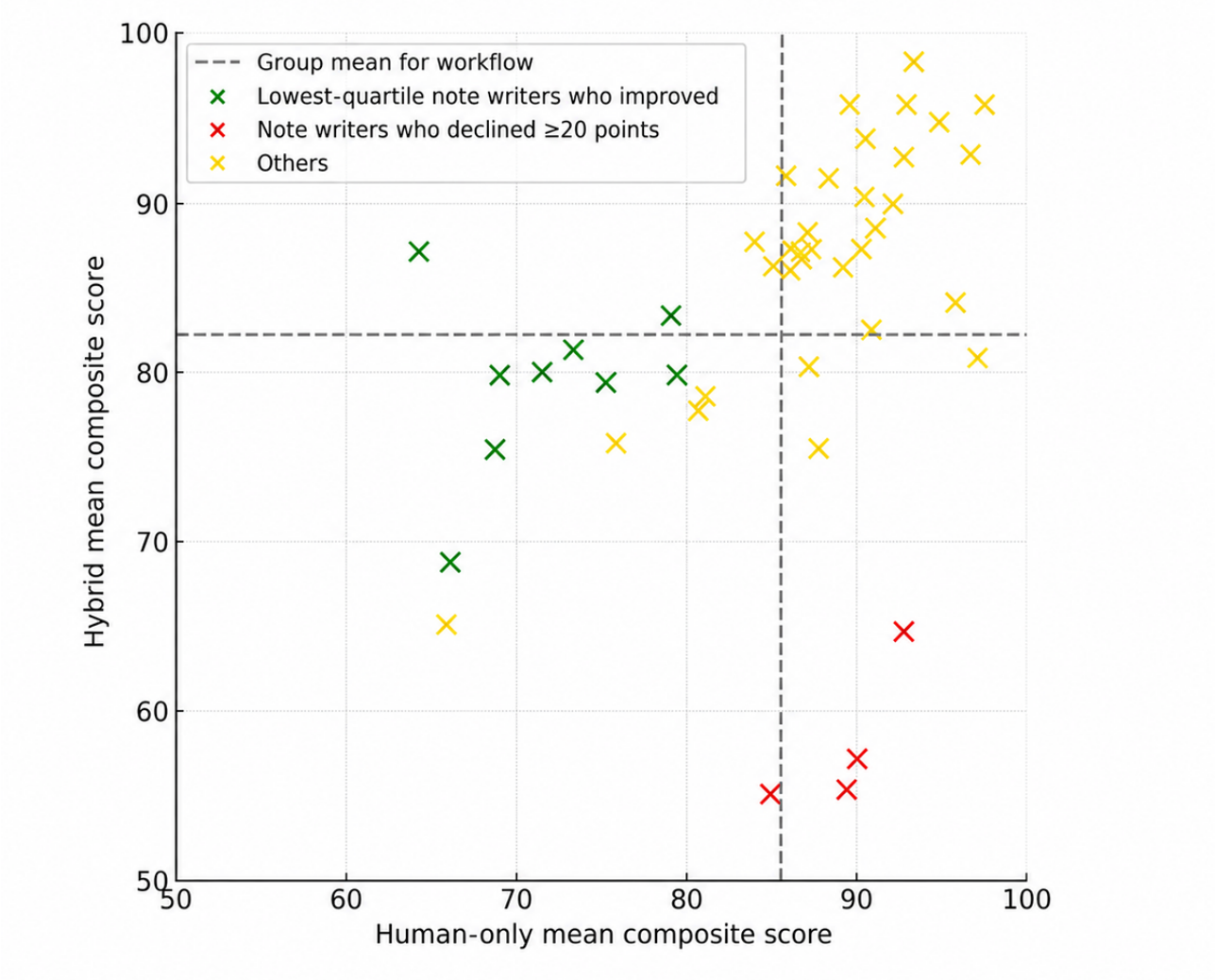
In paired analysis of QNOTE elements between human-only and hybrid notes among AI scribe participants, the largest element-level change was a decline in chief complaint scores (*P*=.05) driven by fewer "fully acceptable" (16/47, 34% vs 21/47, 44.7%) and more "unacceptable" (7/47, 14.9% vs 3/47, 6.4%) hybrid notes as compared to human-only notes, respectively. Scores on all other elements were similar (Table 1 and Multimedia Appendix 1). Post hoc analysis of AI notes

XSL•FO
RenderX

revealed that symptom duration, an explicit criterion required for full QNOTE credit in the chief complaint element, was mentioned in the chief complaint section in 17% (8/47) of the AI notes.

Of the 12 participants in the bottom quartile of human-only composite QNOTE scores, 9 (75%) exhibited improvement with hybrid scores, whereas 8 (66.7%) in the top quartile exhibited a score decline. Overall, 10.6% (5/47) exhibited score declines of at least 20 points between human-only and hybrid notes (Figure 2).

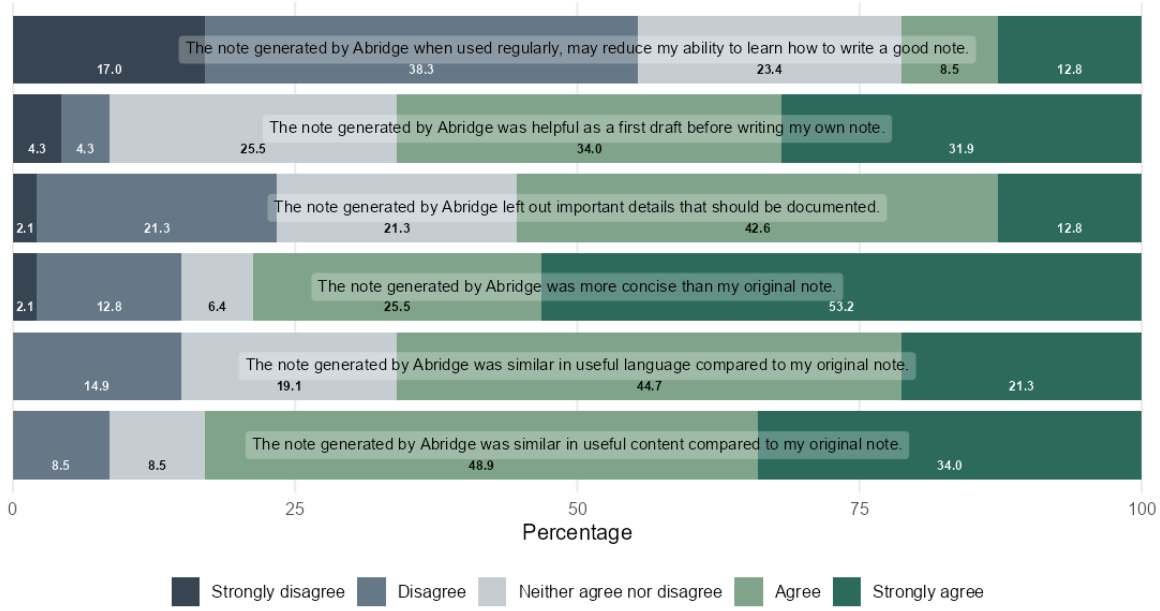
Figure 2. Human-only vs hybrid mean composite ordinal QNOTE scores in the paired samples (n=47). Each point represents 1 participant. Because QNOTE ratings are based on discrete ordinal categories, the plotted mean scores are quasi-continuous summary measures used for visual comparison but not for parametric inference. This figure was created with the assistance of OpenAI's ChatGPT (GPT-5 model). The authors verified the accuracy and originality and hold ownership of the content.



Participants agreed or strongly agreed that the AI write-up "was similar in useful content compared to my original note" (39/47, 83%), "was more concise than my note" (37/47, 78.7%), and would be "helpful as a first draft before writing my own note" (31/47, 66%). A total of 55.3% (26/47) agreed or strongly agreed that the AI write-up "left out important details that should be documented." In total, 21.3% (10/47) agreed or strongly agreed that the AI write-up "when used regularly, may reduce my ability to learn how to write a good note" (Figure 3).

XSL•FO
RenderX

Figure 3. Results of the student experience survey on using an ambient artificial intelligence (AI) scribe (n=47). Students completed this 6-item survey after review of a note created by an ambient AI scribe following a simulated clinical encounter.



Discussion

After writing high-quality "human-only" notes, interaction with notes generated by the AI scribe had little impact on the quality of the "hybrid" notes written by preclerkship medical students. Across 10 discrete note elements, the only element-level change between human-only and hybrid notes was a decline in chief complaint scores. To our knowledge, this study is the first to examine the use of ambient AI scribes for chart documentation among medical students.

Multiple prior studies have focused on the use of AI scribes among practicing physicians [1,4-8]. Tierney et al [33] and Cain et al [34] used different structured assessment instruments to demonstrate the high quality of notes after incorporation of an AI scribe in a large health system. Similarly, You et al [7] found general utility and satisfaction with AI notes in a qualitative analysis in 2 health systems. Our study suggests similar note quality with an AI workflow based on high QNOTE scores on hybrid notes and student survey responses indicating useful language and content in the AI note.

Despite these similarities with existing literature, the fundamental purpose of chart documentation for practicing physicians is different than that for medical students. In practice, the chart note serves as a communication tool between clinicians; the basis for billing, quality improvement, and performance evaluation; and a legal document [26]. While students learn about these features, they primarily write notes for educational purposes as they learn how to organize, prioritize, and synthesize information [15] and hone their ability to serve as narrators of the patient's story [35]. Outcomes of interest in studies of AI scribes among practicing physicians, such as efficiency, cognitive load, and burnout related to note writing, are not immediately relevant to medical students. As such, in this study, we focused on students' ability to structure a note using a conventional format while reasoning through a clinical scenario.

The QNOTE instrument emphasizes note structure with aspects of clinical reasoning embedded throughout the checklist. QNOTE elements reward clear, logical narratives that mirror a well-reasoned diagnostic process. Documentation of pertinent findings and prioritized lists are necessary for full credit on the "history of present illness" and "assessment" elements, respectively, among others [23]. As the AI scribe used in this study is designed to convert descriptions of findings into diagnostic considerations, one might expect an impact on students' clinical reasoning. However, students on average received high scores for all elements on human-only notes with little impact after reviewing AI notes, suggesting no detriment to or augmentation of clinical reasoning overall.

The lack of impact on clinical reasoning may be due to the note writing workflow used in this study. We used a cognitive forcing strategy requiring students to generate independent judgments before exposure to potentially biasing information [29]. Cognitive forcing has been proposed as a strategy to reduce overreliance on AI [36]. In our study, requiring students to produce a human-only note prior to reviewing AI output may have preserved analytic processing and mitigated anchoring effects. While this differs from real-world clinical workflows in which AI-generated notes are the starting point, cognitive forcing through deliberate sequencing within human-AI collaboration offers a potential solution to concerns about introducing ambient AI scribes early in medical training [22].

The only observed difference between human-only and hybrid notes was a decline in "fully acceptable" scores mirrored by an increase in "unacceptable" scores in the chief complaint element. The difference in chief complaint scores was likely due to conciseness in AI notes that often omitted symptom duration, an explicit criterion for full credit on QNOTE. While errors of omission are common among AI scribes [37,38], the consistency

XSL•FO
RenderX

with which duration was omitted from the chief complaint section suggests that this was due to prompt engineering [3]. In our study, over half (26/47, 55.3%) of students recognized that the AI note "left out important details that should be documented." These omissions had minimal impact on hybrid note quality for most QNOTE elements, with scores on average approaching "fully acceptable" when written without AI. However, chief complaint was most susceptible as, on average, the scores for this element were mostly in the "partially acceptable" range among human-only notes in the first place. This decline in quality serves as a reminder of the power of automation bias, the overreliance on automated systems among early learners with AI-generated content as it can result in undesired cognitive offloading [39].

While the risk of automation bias is not unique to students, given their lack of domain expertise, early learners may be particularly vulnerable [39]. Experience and confidence in one's own decisions decreases reliance on automated systems [40] and susceptibility to error in the face of imperfect automated outputs [41,42]. Automation bias may not be completely avoidable when confronted with complex tasks [43], but various educational safeguards have been proposed to mitigate its effect. Deliberate design of systems can help [40]. In this case, using prompt engineering better aligned with the curriculum rather than using an off-the-shelf commercial solution may have prevented the decline in chief complaint scores. However, such a solution may not be practical given resource constraints and is not a replacement for engaging students in training on the appropriate use of AI tools [22,40]. The finding that 21.3% (10/47) of the students recognized that the AI scribe "may reduce my ability to learn how to write a good note" suggests some awareness of this issue among students themselves. Our findings support the conjecture that AI tools should be released to trainees with appropriate educational safeguards and scaffolding [20,22].

The hybrid workflow may have been particularly beneficial for lower-performing students, most of whom demonstrated score increases on hybrid notes. Their improvement must be interpreted cautiously in light of potential regression to the mean as some of the highest performers saw score declines on hybrid notes. High-performing students' human-only note scores were near the maximum possible, so ceiling compression may have limited measurable improvement and could partially explain the shift in the top quartile. As the QNOTE composite score has an upper limit, clustering near this boundary may have amplified apparent declines in the highest-performing group. It is possible that these observations may indicate true signal as they are consistent with findings in economics in which generative AI improves the quality of output for low performers while potentially having the opposite effect on high performers [44]. Establishing a causal signal will require designs that mitigate regression and ceiling effects. Future research should incorporate randomized controlled trial designs with repeated baseline measurements, longitudinal follow-up of performance trajectories, and counterbalanced AI-first vs human-first workflows (with educational safeguards) [22] to isolate the effect of AI exposure.

There are limitations to our study. This study was conducted at a single medical school within a preclerkship curriculum, limiting generalizability. Replication across multiple institutions and clerkship settings and in graduate medical education is necessary to clarify applicability longitudinally across training levels. Next, students self-selected to write hybrid notes rather than being randomly selected. However, the similar evaluation scores on human-only notes between students who submitted hybrid notes and those who did not offsets some complaint for selection bias, although unmeasured differences in motivation or attitudes toward AI may remain. Additionally, we used educational safeguards as this study took place within a summative assessment. Given the high-stakes nature of the OSCE, it was critical that the OSCE itself was unaffected by the study. Therefore, students passively received AI notes after their encounters. This workflow, which differs from ambient scribe use in workplace settings, leaves us unable to comment on the user experience with the technology. To support development of foundational skills prior to AI-supported learning [22,45], we had students write human-only notes before receiving their AI notes. This decision upheld educational best practices but represents a deviation from standard clinical workflows, which limits generalizability beyond workflows that use cognitive forcing. Finally, we did not separately analyze the AI notes because prior work has demonstrated their quality [33,34] and we had no expectation that such a note would stand alone without human review. However, future work should incorporate systematic scoring of AI notes to better understand subsection-level strengths and weaknesses to guide educational efforts.

In conclusion, the use of an ambient AI scribe following a summative OSCE for preclerkship medical students had little impact on note quality. Students found that AI notes were concise and prone to omissions and, on average, submitted high-quality notes with and without AI assistance. For low-performing students, the ambient AI scribe may have helped improve note quality. Given these findings, AI scribes could educationally enhance students' own note writing by helping with brevity and affirming the quality of human-only notes. Enthusiasm for ambient AI scribes in medical student education should be balanced with the finding that omission of key information on an AI note can also yield lower-quality hybrid notes for elements of already lower quality without AI assistance. Our study offers a first snapshot of the impact of AI scribes on chart documentation for medical students, but future research must explore the longitudinal use of such tools. Earlier introduction could potentially jump-start skill development, although it could simultaneously trigger feelings of inadequacy [46]. Deployment later in training or with AI notes as a true first draft could increase efficiency at the expense of creating a cognitive debt due to overreliance on AI [10,22,38]. Given the rapid adoption of ambient AI scribes in clinical settings, it is crucial to further explore the educational opportunities, impact, and trade-offs of these tools.

XSL•FO
RenderX

Acknowledgments

The authors wish to acknowledge Tracy Yale for her administrative support. David Grandchamp, Phil Weiss, and other members of Yale New Haven Health digital technology services and Abridge, Inc, provided verification of technical workflows. The views expressed in this article are those of the authors and do not necessarily reflect the position or policy of the US Department of Veterans Affairs or the US government. The authors used large language models in preparation of the manuscript to assist with formatting references (GPT-4o; OpenAI), mock peer review (GPT-4o), and formatting of Figure 3 (GPT-5; OpenAI). GPT-4o was accessed using a generative artificial intelligence tool housed within Yale University's secure infrastructure. GPT-5 was accessed using Yale School of Medicine's enterprise version of OpenAI's ChatGPT. Both models function within walled gardens in which inputs are not added to the external training dataset. There is no plagiarism of text or images related to our use of artificial intelligence.

Data Availability

The datasets generated or analyzed during this study are not publicly available due to the need to protect student privacy, as the data were collected within a single class of students at a single medical school. The data are available from the corresponding author on reasonable request.

Funding

Funding was provided by the Yale School of Medicine Office of Curriculum.

Authors' Contributions

JST contributed to conceptualization, formal analysis, funding acquisition, investigation, methodology, project administration, resources, software, supervision, visualization, and writing—original draft. DSW contributed to data curation, formal analysis, methodology, validation, visualization, and writing—review and editing. LHS contributed to conceptualization, methodology, resources, and writing—review and editing. GL contributed to investigation, project administration, software, and writing—review and editing. VS contributed to data curation, formal analysis, methodology, validation, visualization, and writing—review and editing.

Conflicts of Interest

LHS reports consulting with Medtronic, LifeImage, Genentech, and Penumbra unrelated to this work; serving as a content advisor on digital health to the Stroke editorial board; holding an artificial intelligence–related patent pending (US2024031761); and representing Yale New Haven Health on a hospital advisory committee for Abridge, an ambient artificial intelligence clinical documentation company. All other authors declare no other conflicts of interest.

Multimedia Appendix 1

QNOTE element scores by workflow. First-year medical students (N=104) completed human-only notes after a simulated clinical encounter. A subset (n=47) then modified their notes into hybrid notes after reviewing the output of an ambient artificial intelligence scribe. Blinded raters scored 10 elements in standard clinical documentation via QNOTE. Percentages of scores in the "fully acceptable" range (76-100) on QNOTE are depicted by note type.

[PNG File , 59 KB - mededu_v12i1e88264_app1.png]

References

1. Tierney AA, Gayre G, Hoberman B, Mattern B, Ballesca M, Wilson Hannay SB, et al. Ambient artificial intelligence scribes: learnings after 1 year and over 2.5 million uses. NEJM Catal 2025 Apr 16;6(5). [doi: 10.1056/CAT.25.0040]
2. Poon EG, Lemak CH, Rojas JC, Guptill J, Classen D. Adoption of artificial intelligence in healthcare: survey of health system priorities, successes, and challenges. J Am Med Inform Assoc 2025 Jul 01;32(7):1093-1100. [doi: 10.1093/jamia/ocaf065] [Medline: 40323320]
3. Morse J, Gilbert K, Shin K, Cooke R, Rose P, Sullivan J, et al. A custom-built ambient scribe reduces cognitive load and documentation burden for telehealth clinicians. arXiv. Preprint posted online on May 2, 2025 . [doi: 10.48550/arXiv.2507.17754]
4. Shah SJ, Devon-Sand A, Ma SP, Jeong Y, Crowell T, Smith M, et al. Ambient artificial intelligence scribes: physician burnout and perspectives on usability and documentation burden. J Am Med Inform Assoc 2025 Feb 01;32(2):375-380. [doi: 10.1093/jamia/ocae295] [Medline: 39657021]
5. Shah SJ, Crowell T, Jeong Y, Devon-Sand A, Smith M, Yang B, et al. Physician perspectives on ambient AI scribes. JAMA Netw Open 2025 Mar 03;8(3):e251904 [FREE Full text] [doi: 10.1001/jamanetworkopen.2025.1904] [Medline: 40126477]
6. Duggan MJ, Gervase J, Schoenbaum A, Hanson W, Howell JT3, Sheinberg M, et al. Clinician experiences with ambient scribe technology to assist with documentation burden and efficiency. JAMA Netw Open 2025 Feb 03;8(2):e2460637 [FREE Full text] [doi: 10.1001/jamanetworkopen.2024.60637] [Medline: 39969880]

XSL•FO
RenderX

7. You JG, Dbouk RH, Landman A, Ting DY, Dutta S, Wang JC, et al. Ambient documentation technology in clinician experience of documentation burden and burnout. JAMA Netw Open 2025 Aug 01;8(8):e2528056 [FREE Full text] [doi: 10.1001/jamanetworkopen.2025.28056] [Medline: 40839265]
8. Olson KD, Meeker D, Troup M, Barker TD, Nguyen VH, Manders JB, et al. Use of ambient AI scribes to reduce administrative burden and professional burnout. JAMA Netw Open 2025 Oct 01;8(10):e2534976 [FREE Full text] [doi: 10.1001/jamanetworkopen.2025.34976] [Medline: 41037268]
9. Levy DR, Rossetti SC, Brandt CA, Melnick ER, Hamilton A, Rinne ST, et al. Interventions to mitigate EHR and documentation burden in health professions trainees: a scoping review. Appl Clin Inform 2025 Jan;16(1):111-127 [FREE Full text] [doi: 10.1055/a-2434-5177] [Medline: 39366661]
10. Wright DS, Kanaparthy NS, Melnick ER, Levy DR, Huot SJ, Hsiao A, et al. The effect of ambient artificial intelligence scribes on trainee documentation burden. Appl Clin Inform 2025 Aug;16(4):872-878. [doi: 10.1055/a-2647-1142] [Medline: 40602775]
11. Aulet TH, Moore JS, Callas PW, Nicholas C, Hulme M. (En)trust me: validating an assessment rubric for documenting clinical encounters during a surgery clerkship clinical skills exam. Am J Surg 2020 Feb;219(2):258-262. [doi: 10.1016/j.amjsurg.2018.12.055] [Medline: 30639130]
12. Gagliardi JP, Bonanno B, McPeek Hinz ER, Musser RC, Knudsen NW, Palko M, et al. Implementation of changes to medical student documentation at Duke University Health System: balancing education with service. Acad Med 2021 Jun 01;96(6):900-905. [doi: 10.1097/ACM.0000000000003729] [Medline: 32909999]
13. Brown DR, Moeller JJ, Grbic D, Andriole DA, Cutrer WB, Obeso VT, et al. Comparing entrustment decision-making outcomes of the core entrustable professional activities pilot, 2019-2020. JAMA Netw Open 2022 Sep 01;5(9):e2233342 [FREE Full text] [doi: 10.1001/jamanetworkopen.2022.33342] [Medline: 36156144]
14. Gallagher BD, Green ML, Talwalkar JS. Applying a validated scoring rubric to pre-clerkship medical students' standardized patient notes: a pilot study. BMC Med Educ 2023 Jul 13;23(1):504 [FREE Full text] [doi: 10.1186/s12909-023-04424-9] [Medline: 37438775]
15. Core entrustable professional activities for entering residency faculty and learners' guide. Association of American Medical Colleges. 2014. URL: https://store.aamc.org/downloadable/download/sample/sample_id/66/%20 [accessed 2025-10-22]
16. Rolfe IE, Sanson-Fisher RW. Translating learning principles into practice: a new strategy for learning clinical skills. Med Educ 2002 Apr;36(4):345-352. [doi: 10.1046/j.1365-2923.2002.01170.x] [Medline: 11940175]
17. Welcher CM, Hersh W, Takesue B, Stagg Elliott V, Hawkins RE. Barriers to medical students' electronic health record access can impede their preparedness for practice. Acad Med 2018 Jan;93(1):48-53 [FREE Full text] [doi: 10.1097/ACM.0000000000001829] [Medline: 28746069]
18. Farmakidis AL, Singh A, Leaf K, Rossi A. Artificial intelligence curricula in U.S. and Canadian medical schools. Association of American Medical Colleges. 2025 Jun. URL: https://www.aamc.org/media/84666/download [accessed 2025-10-22]
19. Bowker D, Torti J, Goldszmidt M. Documentation as composing: how medical students and residents use writing to think and learn. Adv Health Sci Educ Theory Pract 2023 May;28(2):453-475 [FREE Full text] [doi: 10.1007/s10459-022-10167-x] [Medline: 36319807]
20. Abdulnour RE, Gin B, Boscardin CK. Educational strategies for clinical supervision of artificial intelligence use. N Engl J Med 2025 Aug 21;393(8):786-797. [doi: 10.1056/NEJMra2503232] [Medline: 40834302]
21. Risko EF, Gilbert SJ. Cognitive offloading. Trends Cogn Sci 2016 Sep;20(9):676-688. [doi: 10.1016/j.tics.2016.07.002] [Medline: 27542527]
22. Abernethy J, Shah A, Chen B, Reynolds S, Wright SM, O'Rourke P. Integrating AI scribes into medical education: guardrails for preserving clinical reasoning. J Gen Intern Med 2026 Feb 02 (forthcoming). [doi: 10.1007/s11606-025-10149-w] [Medline: 41627656]
23. Burke HB, Hoang A, Becher D, Fontelo P, Liu F, Stephens M, et al. QNOTE: an instrument for measuring the quality of EHR clinical notes. J Am Med Inform Assoc 2014;21(5):910-916 [FREE Full text] [doi: 10.1136/amiajnl-2013-002321] [Medline: 24384231]
24. Stetson PD, Bakken S, Wrenn JO, Siegler EL. Assessing electronic note quality using the Physician Documentation Quality Instrument (PDQI-9). Appl Clin Inform 2012;3(2):164-174 [FREE Full text] [doi: 10.4338/aci-2011-11-ra-0070] [Medline: 22577483]
25. Bierman JA, Hufmeyer KK, Liss DT, Weaver AC, Heiman HL. Promoting responsible electronic documentation: validity evidence for a checklist to assess progress notes in the electronic health record. Teach Learn Med 2017;29(4):420-432. [doi: 10.1080/10401334.2017.1303385] [Medline: 28497983]
26. Talwalkar J, Ouellette J. A structured workshop to improve chart documentation among housestaff. MedEdPORTAL 2009;5095:5 [FREE Full text] [doi: 10.15766/mep_2374-8265.5095]
27. DeLeon S, Mothner B, Middleman A. Improving student documentation using a feedback tool. Clin Teach 2018 Feb;15(1):48-51. [doi: 10.1111/tct.12625] [Medline: 28225204]
28. Liljequist D, Elfving B, Skavberg Roaldsen K. Intraclass correlation - a discussion and demonstration of basic features. PLoS One 2019 Jul 22;14(7):e0219854 [FREE Full text] [doi: 10.1371/journal.pone.0219854] [Medline: 31329615]

XSL•FO
RenderX

29. Croskerry P. The importance of cognitive errors in diagnosis and strategies to minimize them. Acad Med 2003 Aug;78(8):775-780. [doi: 10.1097/00001888-200308000-00003] [Medline: 12915363]
30. Shrout PE, Fleiss JL. Intraclass correlations: uses in assessing rater reliability. Psychol Bull 1979;86(2):420-428. [doi: 10.1037/0033-2909.86.2.420]
31. Koo TK, Li MY. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. J Chiropr Med 2016 Jun;15(2):155-163 [FREE Full text] [doi: 10.1016/j.jcm.2016.02.012] [Medline: 27330520]
32. Rubin M. When to adjust alpha during multiple testing: a consideration of disjunction, conjunction, and individual testing. Synthese 2021 Jul 06;199:10969-11000. [doi: 10.1007/s11229-021-03276-4]
33. Tierney AA, Gayre G, Hoberman B, Mattern B, Ballesca M, Kipnis P, et al. Ambient artificial intelligence scribes to alleviate the burden of clinical documentation. NEJM Catal 2024 Feb 21;5(3). [doi: 10.1056/CAT.23.0404]
34. Cain CH, Davis AC, Broder B, Chu E, DeHaven AH, Domenigoni A, et al. Quality assurance during the rapid implementation of an AI-assisted clinical documentation support tool. NEJM AI 2025 Mar 27;2(4). [doi: 10.1056/AIcs2400977]
35. Brender TD, Celi LA, Cobert JM. Clinical notes as narratives: implications for large language models in healthcare. J Gen Intern Med 2025 Feb;40(3):687-689. [doi: 10.1007/s11606-024-09093-y] [Medline: 39367287]
36. Buçinca Z, Malaya MB, Gajos KZ. To trust or to think: cognitive forcing functions can reduce overreliance on AI in AI-assisted decision-making. Proc ACM Hum Comput Interact 2021 Apr 22;5(CSCW1):1-21. [doi: 10.1145/3449287]
37. Biro J, Handley JL, Cobb NK, Kottamasu V, Collins J, Krevat S, et al. Accuracy and safety of AI-enabled scribe technology: instrument validation study. J Med Internet Res 2025 Jan 27;27:e64993 [FREE Full text] [doi: 10.2196/64993] [Medline: 39869899]
38. Leung TI, Coristine AJ, Benis A. AI scribes in health care: balancing transformative potential with responsible integration. JMIR Med Inform 2025 Aug 01;13:e80898 [FREE Full text] [doi: 10.2196/80898] [Medline: 40749188]
39. Kirpalani A, Grimmer J, Wang PZ. Exploring medical student trust in generative artificial intelligence (ChatGPT) versus peers in team-based learning. Med Sci Educ 2025 Jan 24;35(2):671-674. [doi: 10.1007/s40670-025-02289-9] [Medline: 40353026]
40. Goddard K, Roudsari A, Wyatt JC. Automation bias: a systematic review of frequency, effect mediators, and mitigators. J Am Med Inform Assoc 2012;19(1):121-127 [FREE Full text] [doi: 10.1136/amiajnl-2011-000089] [Medline: 21685142]
41. Dratsch T, Chen X, Rezazade Mehrizi M, Kloeckner R, Mähringer-Kunz A, Püsken M, et al. Automation bias in mammography: the impact of artificial intelligence BI-RADS suggestions on reader performance. Radiology 2023 May;307(4):e222176. [doi: 10.1148/radiol.222176] [Medline: 37129490]
42. Gaube S, Suresh H, Raue M, Merritt A, Berkowitz SJ, Lermer E, et al. Do as AI say: susceptibility in deployment of clinical decision-aids. NPJ Digit Med 2021 Feb 19;4(1):31 [FREE Full text] [doi: 10.1038/s41746-021-00385-9] [Medline: 33608629]
43. Lyell D, Coiera E. Automation bias and verification complexity: a systematic review. J Am Med Inform Assoc 2017 Mar 01;24(2):423-431 [FREE Full text] [doi: 10.1093/jamia/ocw105] [Medline: 27516495]
44. Brynjolfsson E, Li D, Raymond L. Generative AI at work. Q J Econ 2025 May;140(2):889-942. [doi: 10.1093/qje/qjae044]
45. Gin BC, LaForge K, Burk-Rafel J, Boscardin CK. Macy Foundation innovation report part II: from hype to reality: innovators' visions for navigating AI integration challenges in medical education. Acad Med 2025 Sep 01;100(9S Suppl 1):S22-S29. [doi: 10.1097/ACM.0000000000006117] [Medline: 40479503]
46. Lee AH, Link CD, Hersh D, Angoff NR. The student physician–patient–AI relationship. NEJM AI 2025 Jul 16;2(8). [doi: 10.1056/AIp2500260]

Abbreviations

AI: artificial intelligence
OSCE: objective structured clinical examination
SP: standardized patient

Edited by A Stone; submitted 23.Nov.2025; peer-reviewed by B Rubinstein, K Johl, JL Irons; comments to author 26.Feb.2026; revised version received 09.May.2026; accepted 11.May.2026; published 02.Jun.2026.

Please cite as:
Talwalkar JS, Wright DS, Schwamm LH, Leydon G, Shabanova V
Impact of an Ambient AI Scribe on Medical Student Objective Structured Clinical Examination Notes: Nonrandomized Clinical Trial
JMIR Med Educ 2026;12:e88264
URL: https://mededu.jmir.org/2026/1/e88264
doi:10.2196/88264
PMID:42125891

XSL•FO
RenderX

©Jaideep S Talwalkar, Donald S Wright, Lee H Schwamm, Gary Leydon, Veronika Shabanova. Originally published in JMIR Medical Education (https://mededu.jmir.org), 02.Jun.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Education, is properly cited. The complete bibliographic information, a link to the original publication on https://mededu.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Training Empathetic Communication Skills in Medical Students With a Role-Prompted GPT-4o Chatbot: Quasi-Experimental Intervention Study

Yannik Müller[1]; Hans Lemberg[1], BSc; Paul Gellert[1], Prof Dr; Jan C Zoellick[1], PD, Dr rer med

Institute of Medical Sociology and Rehabilitation Science, Charité - Universitätsmedizin Berlin, Berlin, Berlin, Germany

Corresponding Author:
Yannik Müller
Institute of Medical Sociology and Rehabilitation Science
Charité - Universitätsmedizin Berlin
Charitéplatz 1
Berlin, Berlin, 10117
Germany
Phone: 49 30450529055
Email: yannik.mueller@charite.de

Abstract

Background: There is growing concern that artificial intelligence (AI) may diminish the quality of human relationships. However, in a context of widespread social importance (empathetic conversations between doctors and patients), AI can actually improve human conversational skills, potentially enhancing professional relationships. Recent advances in AI allow for realistically role-prompted counterparts for practicing professional conversations, enabling relational learning without the need for human counterparts.

Objective: This study aimed to show the effectiveness of AI chatbots for learning professional communicative skills in medical education. Specifically, we hypothesized that a single conversation with an AI chatbot improves communication skills in medical students across 4 different conversational competencies.

Methods: We conducted a quasi-experimental intervention study involving 4 distinct role-prompted scenarios (ie, shared decision-making, motivational interviewing, sexually transmitted diseases, and breaking bad news)—each designed to elicit in-depth empathic conversational skills aligned with key learning objectives in medical curricula. Students rated their competence for the 4 scenarios before and after a conversation with GPT-4o (OpenAI) using default settings, without fine-tuning. We expected higher perceived communication competence (PCC) in their conversation topic after the interaction compared with before the interaction in a 2-sided paired *t* test. Participants received AI-generated feedback, which they rated regarding adequacy. Post hoc analyses addressed gender and case effects, feedback adequacy, and prevalues in PCC.

Results: This study shows that a role-prompted GPT chatbot improves PCC in 162 medical students after a single conversation with mean of 13 (SD 4.8; 95% CI 12-14) prompt-response pairs. We found an increase in PCC with a mean difference of 0.94 (SD 1.64; 95% CI 0.69-1.20; Cohen *d*=0.58) from 5.89 (95% CI 5.55-6.23; scale 0-10) before the conversation to 6.83 (95% CI 6.55-7.12) after the conversation across 4 different patient role prompts. Furthermore, we found participants rating AI feedback of their conversation to be useful (mean 7.92, SD 1.61; 95% CI 7.67-8.17; scale 0-10), but feedback adequacy did not correspond to PCC increase (*r*=0.08; *P*=.32).

Conclusions: Our results demonstrate how role-prompted GPT increases self-assessed communication competencies, introducing a novel tool for teaching relational learning. Our results present a starting point for using AI in education, particularly teaching communication in professional roles. On the basis of our findings in medical education, we anticipate further studies to investigate conversational training between lawyers and clients, marketers and customers, or managers and employees. Our research thus has implications for any field with a need for conversational training and relational learning.

(JMIR Med Educ 2026;12:e88092) doi:10.2196/88092

KEYWORDS

artificial intelligence; AI; generative pretrained transformer; ChatGPT; chatbot; medical education; empathy; feedback; didactic method; educational tool; simulated patient

XSL•FO
RenderX

Introduction

Background

Communicating effectively and empathetically in professional roles is a key competence that builds trust, fosters empathy, and enhances outcomes for the parties involved. In medicine particularly, good communication leads to better adherence [1] and better health outcomes for patients [2,3]. Thus, teaching communication competencies is a core learning objective in contemporary medical curricula [4,5]. Conventional didactic methods to teach communication combine experiential and cognitive aspects and include (videotaped) observations, roleplays, actors as standardized or simulated patients, or real patient interactions [6-8]. However, those methods are resource-intensive and time-consuming. Evidence for the effectiveness of digital didactic formats continues to be weak [9].

Generative artificial intelligence (AI) chatbots have the potential to fundamentally change conversational patterns and human interpersonal relationships. Chatbots, described as "stochastic parrots" [10], might induce psychosis in vulnerable populations [11,12]: they "hallucinate" (ie, spread false information as credible) [13], reduce social interactions in professional relationships [14], or make established assessments such as essays insufficient in educational settings [15]. However, promising aspects have also been discussed. AI chatbots potentially reduce conspiracy beliefs [16] and climate change skepticism [17], provide correct medical information with high accuracy [18-20], and increase efficiency in administrative processes, such as discharging patients [21], writing medical reports [22,23], or analyzing medical notes [24].

In this study, we focus on AI chatbots as a novel didactic tool for communication competence. Previous proof-of-concept and feasibility studies have demonstrated that GPT technologies can be used as sophisticated role-prompted simulations in medical education [25-34]. However, these studies focused on single, confined tasks such as history taking or mammography, and their feedback—if provided at all—consisted of simple yes or no checklists of included topics diverging from behavior of actors, patients, or lecturers in educational settings. However, feedback is a central part of learning [35], particularly when presented in an adequate, resource-oriented fashion [36].

Building on these initial studies, we developed a tool that not only includes several simulated patient personas across various communicative techniques but also a feedback loop for performance on 8 different learning objectives as a didactic feature. The role prompts corresponded to core learning objectives in our university's medical curricula. We hypothesized that a single conversation with an AI chatbot improves communication skills in medical students in 4 different conversational situations: shared decision-making (SDM) for treatment plans [37], motivational interviewing (MI) [38], sexual anamnesis and education [39], and breaking bad news [40]. We further investigated whether these competencies were domain-specific (ie, training with an SDM case only enhances SDM skills) or generalized (ie, training with SDM also enhances skills regarding MI, sexual anamnesis, and breaking bad news).

The purpose of this paper was to show the effectiveness and potential of AI chatbots for learning professional communicative skills in higher education. We used the GPT-4o model from OpenAI for building our tool with medical students from our university at different stages of their study.

Key Innovations

Our approach complements and expands existing didactic tools in multiple ways. First, the tool offers a resource-efficient training tool that can be used by many students simultaneously anytime anywhere overcoming restrictions of time and space [41]. Second, the tool can be expanded easily by integrating other personas and use cases by inserting other system prompts. Third, while we focus on German language, the tool could easily adapt to any languages and cultural contexts included in the underlying GPT model. Such variability and expandability make our approach a viable tool that integrates easily into teaching and learning.

Methods

Study Design and Setting

We conducted a quasi-experimental intervention study assigning medical students to 1 of 4 intervention groups based on their semester of study to interact with an AI chatbot. For this purpose, we programmed a website for data collection and chatbot interaction called PythonAnywhere using Flask (version 3.1.0) in the backend [42]. The code is available upon reasonable request. Medical education in Germany consists of a 6-year program based on learning objectives that traditionally involves a preclinical (year 1 and 2), a clinical (years 3-5), and a practical phase (year 6) with state-regulated examinations [5,43]. Our medical faculty implements a reformed curriculum that integrates preclinical and clinical training with an emphasis on promoting communicative competencies. We developed 4 cases corresponding to core learning objectives in the German national competence-based learning objective catalog [5] that applies to all medical faculties: SDM, MI, sexual anamnesis, and breaking bad news. We mapped core learning objectives to these cases in Multimedia Appendix 1.

Procedure

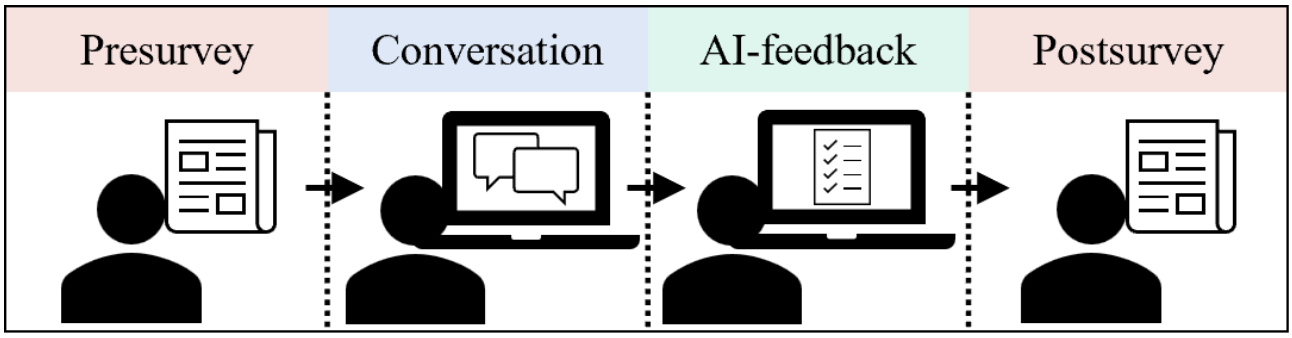
After encountering study information and providing informed consent, participants were asked questions on gender, semester of study, and the preinteraction outcome measure perceived communication competence (PCC) for all 4 topics to measure domain-specific or generalized skill development. On the basis of the semester of study, participants entered 1 of 4 chat dialogues: group 1 (semester 2 and 3) conversed with the chatbot prompted as a 58-year-old woman with a recent type 2 diabetes mellitus diagnosis regarding therapy options, group 2 (semester 4 and 5) used MI with a 58-year-old woman with chronic obstructive pulmonary disease indicating lifestyle changes, group 3 (semesters 6-8) performed a sexual anamnesis with a 25-year-old man with symptoms after practicing unprotected sex, and group 4 (semester 9 and higher) delivered the diagnosis lung cancer to a 58-year-old man, a heavy smoker with productive coughing who underwent computer tomography. All 4 system prompts can be found in Multimedia Appendix 1

XSL•FO
RenderX

alongside our prompt engineering process. For conversations with the chatbot, we used the application programming interface (API) for the GPT-4o model (default values for all parameters) [44]. We opted for GPT-4o as it represented the most advanced model during development. GPT-4o outperformed Anthropic Claude and Mistral AI Mixtral (we compared 4o to Claude Sonnet 4.0 and Mistral Large 2.1) in our initial tests. Participants had to give at least 4 inputs to ensure a minimum level of content for each conversation. We limited the number of messages sent via the API per conversation to 40 to cap resource use and avoid hour-long interactions that would be insufficient simulations of clinical practice. For feedback, we also used the API for the GPT-4o model (temperature=0.7; default values for all other parameters) [44].

After interacting with the chatbot, participants immediately received feedback from GPT-4o using the case-specific feedback prompt found in Multimedia Appendix 1. The feedback had a standardized structure with a grade between 0% (very poor) and 100% (very good) followed by textual assessments across 8 criteria and a conclusion. Four general criteria applied to all cases (empathy, clear structure of the conversation, patient-centered approach, and open communication with open-ended questions), and 4 criteria were case-specific—always matched to the case participants experienced (eg, strengthening self-efficacy in case 1 or using MI techniques in case 2). The textual assessments often included quotes from the conversation and specific suggestions for improvement on the respective criterion. After receiving feedback, participants answered the 4 outcome questions again (postinteraction collection), and they evaluated the feedback using the perceived adequacy of feedback (PAF) scale [36]. We then asked participants to answer scales on technology acceptance [45,46] and the personality trait of technology commitment (TC) [47]. Finally, participants provided free comments on their experience during the study with the item "Do you have any additional feedback?" The study design is displayed in Figure 1.

Figure 1. Study design. Participants displayed here as generic personas answered a presurvey, engaged in a conversation with an artificial intelligence (AI) chatbot, received AI feedback, and answered the postsurvey.



Outcomes and Assessments

We used the students' PCC measured via preparation for a medical consultation as primary outcome obtained via self-reports of participants before and after the interaction with the chatbot (repeated measure) with the question "How well do you feel prepared for a medical consultation about..." rated for 4 topics on an 11-point Likert scale ranging between not prepared at all (0) and very well prepared (10). The 4 topics were (1) SDM for treatment plans, (2) MI, (3) sexually transmitted diseases, and (4) breaking bad news. We calculated the outcome as within-person difference postvalue minus prevalue.

The secondary outcome assessed the adequacy of AI feedback. We used the PAF subscale from the Feedback Perceptions Questionnaire [36] with 9 items rated on an 11-point Likert scale ranging between not at all useful (0) and extremely useful (10). After performing principal component and reliability analyses, we calculated a PAF mean score across all 9 items.

Our questionnaire also contained items on gender (male or female or diverse) and current semester of study (semesters 1-10 and practical year). Finally, we included instruments on technology acceptance. The simple scale for acceptance was a semantic differential with 9 items on a 5-point Likert scale ranging between −2 and 2 [45]. Perceived usefulness (PU) and perceived ease of use (PEOU) from the technology acceptance model [46] each consisted of 4 items on a 5-point Likert scale ranging between completely disagree (score=1) and completely agree (score=5). Finally, TC encompassed 12 items on a 5-point Likert scale with 3 dimensions containing 4 items each ranging between completely disagree (score=1) and completely agree (score=5) [47]. After performing principal component and reliability analyses (acceptance: α=.87; PU: α=.93; PEOU: α=.94; TC [control beliefs]: α=.77; TC [competence]: α=.89; TC [acceptance]: α=.84), we calculated mean scores for each construct.

We obtained the length of the conversations by counting the number of messages and responses, and the number of characters written by the student and the chatbot, including spaces. We also obtained the AI-calculated grade between 0% (very poor) and 100% (very good) as part of the feedback.

Sample Size

In our sample size calculation with nQuery+nTerim 4.0, we expected an average increase in PCC from a mean of 6.5 (before) to a mean of 7.0 (after) with equal SDs of 1.0 corresponding to a medium effect size (Cohen d=0.50) using established values for power (1-β=.80) and statistical significance (α=.05). Expecting a dropout of 10%, we formulated a recruitment target of 156 participants (ie, n=39 participants per case).

XSL•FO
RenderX

Participants (Eligibility Criteria)

Our inclusion criteria were adult medical students enrolled at our university in study semesters 2 to 10. We excluded datasets that did not fulfill the inclusion criteria and those without responses to the primary outcome, multiple participations based on user and session IDs, or conversation topics outside of the intended medical case.

Sampling Method and Recruitment Strategies

We used convenience sampling by inviting participants using internal student mailing lists and chatgroups enabling participation between June 17, 2025, and September 14, 2025.

Statistical Analysis

We calculated descriptive statistics on primary and secondary outcomes. Regarding our primary hypothesis, we calculated a 2-sided paired *t* test between the prevalue and postvalue across all cases. We calculated additional 2-sided paired *t* tests between the prevalues and postvalues for each case individually with Bonferroni-Holm-adjusted significance levels. We additionally calculated a linear regression analysis with the difference between postvalue and prevalue as outcome and the prevalue in PCC as predictor to investigate the impact of prevalues on the learning increase. As secondary analyses, we calculated 2-sided independent *t* tests for gender differences in increased PCC, as well as correlations between the PCC scores, the AI feedback grade, the PAF score, and the acceptance measures. Finally, we identified the 5 AI feedbacks with the highest and lowest PAF scores for each case and contrasted their structural characteristics (ie, length and grade), as well as prevalues and pre-post changes for PCC in 2-sided independent *t* tests. We performed all data cleaning and analyses with MS Excel and SPSS (version 29.0; IBM Corp). All data can be found in Multimedia Appendix 2.

Ethical Considerations

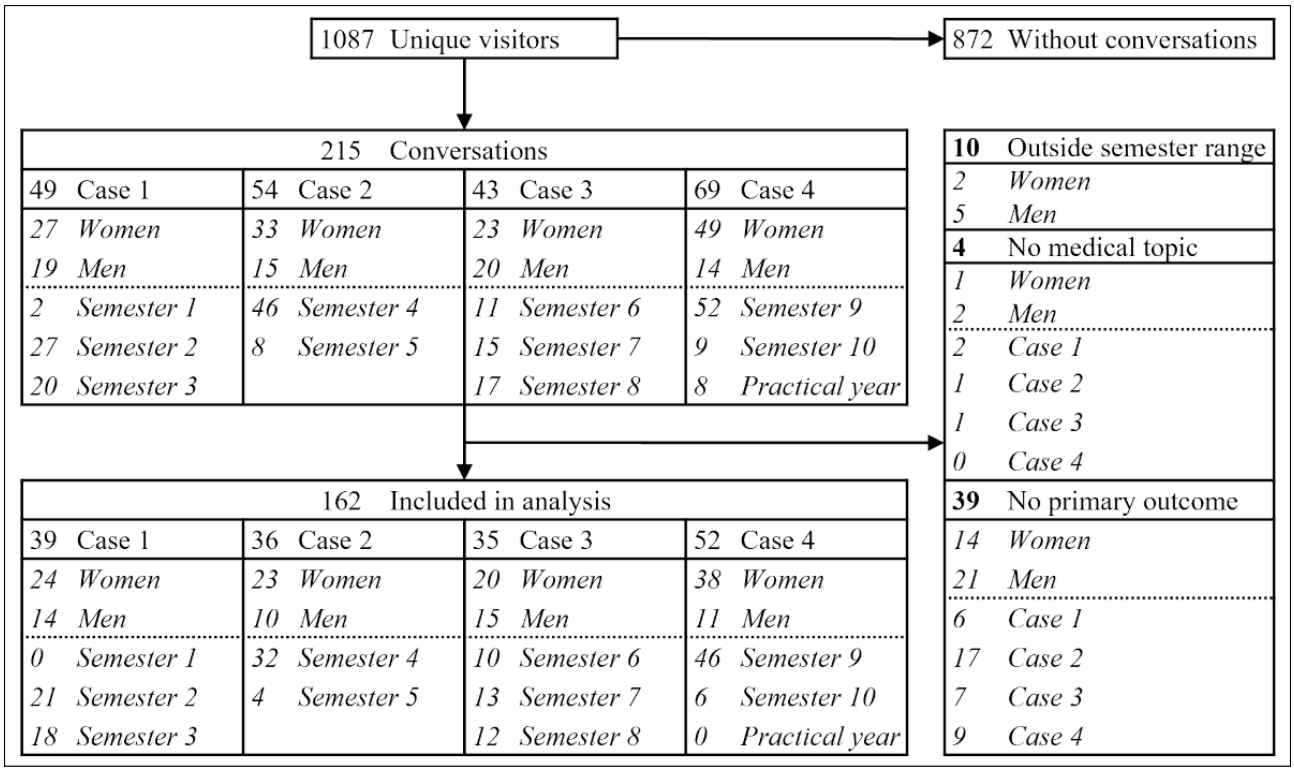
We preregistered our study design prior to data collection at AsPredicted (approval number 228674). The study was approved by the Charité Ethics Committee (EA1/038/24), and the study adhered to the Declaration of Helsinki. All participants gave informed consent.

Results

Sample

A total of 1087 unique users visited the starting page of our website. Of those, 215 unique users engaged in conversations with the chatbot. We excluded 10 datasets from students outside the semester range and 4 datasets because they conversed about unintended topics. Finally, we excluded 39 datasets without postvalues for PCC. After exclusion, 162 unique users remained included in the analysis (Figure 2). Participants were mostly female (105/162, 65%) corresponding to the relative ratio of female students in our faculty (Table 1). Case-specific prevalues in PCC were the lowest in case 1 and the highest in case 4 mirroring study progression from beginning to end of the taught phase in medical studies. Participants reported high TC competence compared with the other dimensions of TC. Sensitivity analyses between 162 included datasets and 39 excluded datasets did not reveal differences regarding prevalues for PCC, number and length of messages, and grade in AI feedback (Multimedia Appendix 1). A total of 28 unique users engaged in 36 additional conversations with the chatbot, which we also excluded from our analysis.

Figure 2. Study flowchart. For each case, we provided absolute numbers on gender and semester of participants. For each exclusion criteria, we provided absolute numbers for gender and assigned case. Semester 1 and practical year were outside the semester range. Missing data on gender account for differences between totals and sums.



XSL•FO
RenderX

Table 1. Sample characteristics (N=162).

Variables	Total	Case 1 (SDM[a]; n=39)	Case 2 (MI[b], n=36)	Case 3 (STD[c]; n=35)	Case 4 (bad news; n=52)
Gender (woman), n (%)	105 (65)	24 (62)	23 (64)	20 (57)	38 (73)
Respective PCC[d] prevalue, mean (SD)	5.89 (2.20)	4.64 (2.15)	5.78 (2.28)	5.91 (2.25)	6.88 (1.63)
TC[e] control beliefs, mean (SD)	3.46 (0.77)	3.27 (0.83)	3.50 (0.81)	3.65 (0.84)	3.47 (0.62)
TC competence, mean (SD)	4.22 (0.85)	4.15 (1.05)	4.15 (0.85)	4.30 (0.79)	4.26 (0.74)
TC acceptance, mean (SD)	3.47 (0.91)	3.39 (1.09)	3.44 (0.88)	3.52 (0.76)	3.51 (0.90)

[a]SDM: shared decision-making.

[b]MI: motivational interviewing.

[c]STD: sexually transmitted disease.

[d]PCC: perceived communication competence.

[e]TC: technology commitment.

Conversations

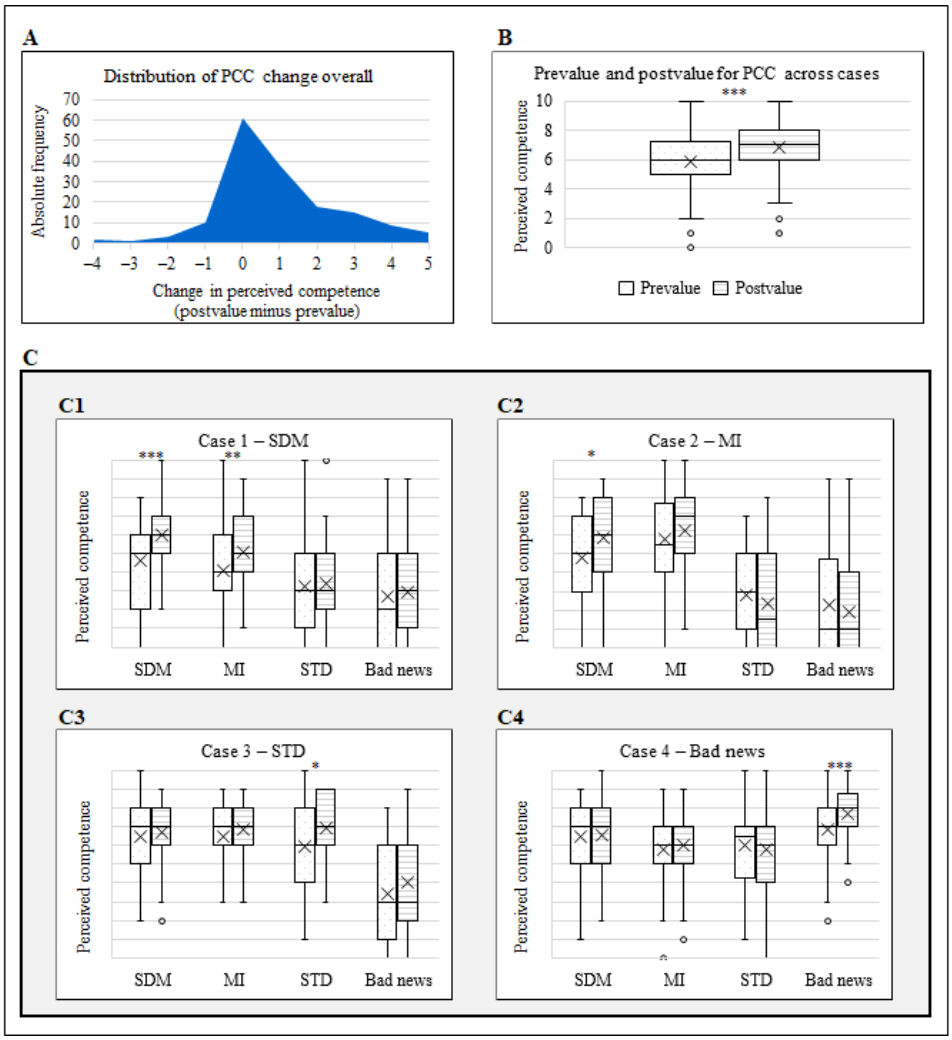
The conversations lasted for a median of 22 (IQR 16-30) minutes and contained on average 13 (SD 4.8) message pairs. Conversations in case 1 had few but long student prompts in contrast to the other 3 cases. Longer student prompts produced longer chatbot responses (r=0.59; P<.001). Descriptive statistics of the conversations and high-ranking and low-ranking exemplary conversations for each case are found in the Multimedia Appendix 1. We used around 5,400,000 input tokens and 370,000 output tokens for all conversations and feedback during data collection.

Perceived Communication Competence

Participants reported higher PCC for the interview situation in their group's topic after the conversation with the chatbot (mean 6.83, SD 1.84; 95% CI 6.55-7.12) than before the conversation (mean 5.89, SD 2.20; 95% CI 5.55-6.23; mean difference 0.94, SD 1.64; 95% CI 0.69-1.20; t_{161}=7.33; P<.001; Cohen d=0.58). We observed increased PCC for the respective topic in cases 1 (mean difference 1.41, SD 1.77; t_{38}=4.97; P<.001; Cohen d=0.80; n=39), 3 (mean difference 1.03, SD 1.89; t_{34}=3.23; P=.001; Cohen d=0.55; n=35), and 4 (mean difference 0.85, SD 1.27; t_{51}=4.79; P<.001; Cohen d=0.66; n=52), but not in case 2 (mean difference 0.50, SD 1.63; t_{35}=1.84; P=.07; n=36). Figure 3 shows distributions and boxplots with prevalue and postvalues overall and for each of the 4 topics in each condition. We observed generalized PCC development in cases 1 and 2 with participants reporting increased perceived MI competence after conversing about SDM and vice versa. In contrast, participants in groups 3 and 4 experienced domain-specific skill development without generalization to other situations. Participants with lower prevalues reported higher increases in PCC in a linear regression analysis with post-pre difference as outcome and ln(prevalue) as predictor (B=−0.10, SE 0.008; β=−0.67; t_{160}=−11.49; P<.001; R^2=0.45).

XSL•FO
RenderX

Figure 3. Graphical display of results for the primary outcome perceived communication competence (PCC): (A) distribution of change in PCC across all 4 cases, (B) boxplots for prevalue and postvalue PCC across all cases, and (C) case-specific boxplots for prevalue and postvalue PCC for all 4 interview situations. (C1) Prevalues and postvalues in PCC for all 4 interview situations of participants in case 1 (chatbot conversation about shared decision-making [SDM]), (C2) prevalues and postvalues in PCC for all 4 interview situations in case 2 (chatbot conversation about motivational interviewing [MI]), (C3) prevalue and postvalues in PCC for all 4 interview situations in case 3 (chatbot conversation about sexually transmitted diseases [STD]), and (C4) prevalues and postvalues in PCC for all 4 interview situations in case 4 (chatbot conversation about breaking bad news). Indicators for Bonferroni-Holm–corrected *P* values: **P*≤.05, ***P*≤.01, ****P*≤.001.



Feedback Adequacy

The feedback contained on average 3341 (SD 400) characters ranging in length between case 3 (mean 3139, SD 358) and case 4 (mean 3498, SD 473). Participants received an average grade of 83% (SD 5%) with no differences between cases. The feedback was generally perceived positively with an average PAF score of 7.92 (SD 1.61; scale 0-10). Participants rated feedback adequacy in case 1 the highest (mean 8.09, SD 1.66) compared with case 4 as the lowest (mean 7.67, SD 1.87). The PAF score did not correlate with the grade in the feedback (r=0.06; P=.44; n=156) or the length of the feedback (r=−0.01; P=.87; n=156).

Contrasting the highest- and lowest-ranked AI feedback of the 4 cases did not reveal differences regarding the length of the feedback (t_{38}=0.33; P=.75) or the AI-generated grade (t_{38}=−0.32; P=.75); participants did not differ regarding their prevalues (t_{38}=0.61; P=.55) or their PCC gains (t_{38}=−0.86; P=.40). Each of the 40 feedbacks followed the same structure, analyzing and highlighting improvements on the 8 categories.

Participants' responses to the final open question of our survey regarding additional comments addressed generic AI phrases or the wish for specific and detailed examples for improvements. Some participants found the grade too harsh and not corresponding to the positive content of the text. However, participants providing these comments did not rank among the lowest PAF scores.

Acceptance of Chatbot for Conversational Training

Participants reported high acceptance (mean 1.23, SD 0.55; scale −2 to 2), PU (mean 4.07, SD 0.89; scale 1-5), and PEOU (mean 4.56, SD 0.84; scale 1-5) after their interactions with the chatbot, without differences regarding cases, indicating similar usefulness and ease of use regardless of conversational situation. PCC increase correlated moderately with acceptance (r=0.29; P<.001) and PU (r=0.23; P=.004), but not with PEOU (r=0.03; P=.68). Other measures such as number of messages or characters as well as grade in the AI feedback did not correlate

XSL•FO
RenderX

with acceptance ($|r|\leq 0.11$; $P\geq .15$). Men and women did not differ regarding acceptance ($t_{151}=-1.05$; $P=.30$), PU ($t_{150}=0.26$; $P=.80$), or PEOU ($t_{150}=-0.56$; $P=.58$). Table 2 shows subjective assessment of the chatbot in the 4 cases. Participants' responses to the final open question of our survey regarding additional comments addressed the wish for spoken language instead of written text to include nonverbal cues, the unrealistic friendliness of the chatbot patient and its long responses, as well as technical barriers such as not being able to retract messages or sending messages too quickly by pressing enter.

Table 2. Descriptive statistics of acceptance, perceived usefulness (PU), and perceived ease of use (PEOU; n=160).

Variables	Total, mean (SD)	Case 1 (SDM[a]; n=38), mean (SD)	Case 2 (MI[b]; n=36), mean (SD)	Case 3 (STD[c]; n=35), mean (SD)	Case 4 (bad news; n=51), mean (SD)
Acceptance	1.23 (0.55)	1.28 (0.63)	1.07 (0.65)	1.34 (0.40)	1.23 (0.50)
PU	4.07 (0.89)	4.03 (0.98)	4.01 (0.92)	4.17 (0.95)	4.08 (0.76)
PEOU	4.56 (0.84)	4.46 (0.98)	4.42 (0.93)	4.69 (0.73)	4.65 (0.74)

[a]SDM: shared decision-making.

[b]MI: motivational interviewing.

[c]STD: sexually transmitted disease.

Discussion

Summary

Our study addresses the debate about harmful or beneficial effects of AI for interpersonal relations. Using highly emotive and professionally vital doctor-patient conversations, we demonstrate enhanced PCC across 4 scenarios requiring advanced skills and empathy. We thus provide evidence that an AI chatbot can be an effective didactic method to teach conversational skills. While we addressed medical students and their simulated patients, our tool targets high-level conversational skills also required in other professional relationships such as lawyers and clients, marketers and customers, or managers and employees. Specifically, we prompted GPT-4o as a simulated patient in 4 roles corresponding to core learning objectives in medical curricula with emotionally sensitive topics and a need for empathy (ie, SDM, MI, STDs, and breaking bad news). We combined this experiential component with structured feedback consisting of an overall grade for the student performance and text-based evaluations and improvements across 8 dimensions. The resulting tool is easily accessible, integrates seamlessly into the medical curricula, and complements existing didactic methods such as roleplays or actors simulating patients. Our aim was to increase students' self-assessed competence in these scenarios after a single conversation with the AI chatbot compared to before. Analyzing data from 162 medical students, we demonstrated increased PCC scores overall and specifically in 3 of the 4 cases. Participants in the MI case did not report significantly better PCC regarding MI after exposure to the chatbot. We attribute these null findings to the comparatively high level of complexity of MI as conversational techniques, which makes achieving gains in PCC more challenging. Moreover, MI typically involves patients who express ambivalence, conflicting cognitions, and resistance, which are elements that a fundamentally supportive AI chatbot may struggle to represent adequately. The AI-generated feedback received positive evaluations by students regarding adequacy using the validated PAF scale. Contrary to studies addressing concerns [10-14], we demonstrate how AI chatbots can enhance communication skills fostering interpersonal relationships.

Comparison With Prior Work

To the best of our knowledge, our tool is the first application of a GPT didactic tool for teaching professional communication skills in the broader sense. By that we mean an empathetic, patient-centered approach that uses open communication to develop specific, shared decisions for further medical procedures such as treatment or lifestyle interventions. Thus, we expand on use cases introduced in the study by Ning et al [41]. Accordingly, students in our study received detailed feedback that assessed the conversation across 8 dimensions providing improvements using specific examples from the interaction. Such a procedure enhances the learning experience of students because it contextualizes the assessment. This approach to communication and feedback sets our tool apart from previous feasibility or proof-of-concept studies that treated communication as a structured almost algorithmic task focusing solely on the information dimension [48]. For example, the most comparable study focused on history taking assessing the GPT conversations based on a list of topics included in the conversation (eg, main complaint, severity of symptoms, medication, or family history) [28]. Accordingly, their feedback consisted of a checklist of topics lacking a qualitative component assessing communication skills. Other studies did not include feedback at all but simply rated student performance in cases on chest and abdominal pain [27]. Finally, one other study used a similar approach to communication training as we did [32]. However, again, the focus of that study was on basic communication skills, such as history completeness and clinical reasoning, rather than social and emotional conversation competencies. We want to stress the importance of a holistic approach to communication as it encompasses those elements that build trust, create a shared understanding, and that are ultimately responsible for improving patient outcomes and experiences [1-3]. Indeed, the observed effects in PCC increase between case 1 (SDM) and case 2 (MI) demonstrate how empathy and goal-oriented guidance in medical conversations are generalized skills applicable to multiple situations expanding beyond a specific task.

XSL•FO
RenderX

Strengths and Limitations

We built a chatbot as a didactic tool that integrates seamlessly in German medical curricula as it is based on core learning objectives relevant for all medical faculties. We matched cases to study progression addressing relevant learning objectives for each participant. The 4 cases cover several medical disciplines, and they combine general and domain-specific communication competencies. The gender distribution in our sample corresponds with our student population. We recruited a large sample size of students across the stages of medical studies. Thus, we are confident that our results can be generalized to other medical faculties in these regards. Cases might be matched to other semesters in medical faculties using the traditional curriculum. Selection bias potentially applies to technology affinity as more affine students likely participated in our study. However, we do not have data on technology affinity of nonparticipants or the student population in general.

Limitations of our study include the lack of a control group without exposure to our chatbot. We also used written text as mode of communication, which excludes nonverbal cues such as mimic, gestures, and prosody found to be informative in face-to-face interaction [49-53]. We opted for written text because this feature was cost-effective regarding OpenAI's pricing model and it did not require participants to have audio input and output. We thus consider our findings a low bound estimate for training effects, and we expect studies with spoken language to increase PCC gains above the effect sizes we found in this study [32]. However, self-reports might also overstate short-term gains in PCC compared to external assessments. Conversations lasted 22 minutes on average exceeding time available for clinicians in practice. Furthermore, the grading system of AI feedback did not produce much variance. The AI feedback used only 17 grades out of 101 possibilities, and it provided the grade 85% in 92 of the 162 cases. We acknowledge that we recruited a homogeneous sample of students from the same university studying the same subject with the same curriculum, thus having a somewhat similar approach to patient interactions. However, we expected more variance in performance. In follow-up studies, we will thus continue to fine-tune the feedback prompt to create more variance in assessments. Finally, PCC as self-assessment provides relevant insights into student experiences and learning. Combining these self-assessments with external assessments in examinations with experienced lecturers expands the evaluation of AI chatbots.

Future Research and Teaching

Our study with its preregistered effects and large sample size marks a shift from pilot studies to more advanced designs and approaches regarding chatbots in medical education. We propose future studies to follow more rigorous designs with randomization and control groups combining self-assessment with external assessment of learning outcomes, and varying difficulty and symptom severity of cases as well as AI models. For more realistic simulations of patient encounters, we suggest integrating spoken language and potentially avatars in a next step. Limiting interaction times to 5 to 10 minutes would realistically simulate clinical practice.

For teaching, our tool offers different deployment models: exclusive self-learning, preclass training and reflection, small-group debrief, or structured affiliation in a skills laboratory. In line with the human-in-the-loop paradigm, we suggest integrating AI conversations and feedback into the classes to discuss the interactions, their realism, and potentially irritating experiences resulting from AI hallucinations. We also propose a feedback loop for reporting irritating experiences to the technical chatbot deployment team for prompt modifications. Data protection and AI literacy is most relevant for students (eg, avoiding real names or personal information) as our proposed tool does not involve patient data.

Conclusions

In conclusion, a single conversation with an AI chatbot role-prompted as a patient enhances students' PCC across a variety of cases that represent core learning objectives in medical curricula. Our tool presents a novel didactic method that is versatile in its use and can thus be easily integrated into existing curricula. Potential deployment models for practical use include exclusive self-learning, preclass training and reflection, small-group debrief, or structured affiliation in a skills laboratory. Expanding the mode to spoken language integrated into a larger training regime offers potential for higher education in professional communication beyond medicine.

Acknowledgments

The authors would like to thank Rolf Kienle, Peter Eberz, and Sylke Langenbeck from the Prodekanat für Studium und Lehre at Charité – University Medical Center Berlin for their comments on the case vignettes and feedback. The authors would also like to thank Andrea Stroux and Matthias Becher from the Institute of Biometry and Clinical Epidemiology at Charité – University Medical Center Berlin for their methodological advice. Furthermore, the authors would like to thank Josephine Konrad from the Institute of Medical Sociology and Rehabilitation Science at Charité – University Medical Center Berlin for her support in data cleaning. YM and JZ presented preliminary findings at the 6th TRAIL Summer Workshop on Artificial Intelligence at King's College London between August 25, 2025, and September 5, 2025. The authors would like to thank the organizers and all participants for their valuable feedback, particularly Nicolas Bioul, Luca La Fisca, Thomas Baldassarre, Corentin Warzée, Yihang Zhao, and Wenxin Zhang. Finally, the authors would like to thank Sophus zu Ermgassen from Oxford University for providing feedback on an earlier version of the manuscript draft.

XSL•FO
RenderX

Funding

No external financial support or grants were received from any public, commercial, or not-for-profit entities for the research, authorship, or publication of this paper.

Data Availability

All data generated or analyzed during this study are included with this submission (and its multimedia appendices).

Authors' Contributions

Conceptualization: YM, HL, JCZ
Data curation: YM, JCZ
Formal analysis: YM, JCZ
Investigation: YM, HL
Methodology: YM, JCZ
Project administration: PG, JCZ
Resources: JCZ
Software: HL
Supervision: PG, JCZ
Validation: JCZ
Visualization: YM, JCZ
Writing—original draft: YM, JCZ
Writing—review and editing: HL, PG
All authors had full access to all the data in the study, accept final responsibility for the decision to submit for publication, and take responsibility for the content of the paper.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Prompts used and example conversations of all cases as well as sensitivity analyses.
[DOCX File , 122 KB - mededu_v12i1e88092_app1.docx]

Multimedia Appendix 2
Data used in our analyses.
[XLSX File (Microsoft Excel File), 390 KB - mededu_v12i1e88092_app2.xlsx]

References

1. Zolnierek KB, Dimatteo MR. Physician communication and patient adherence to treatment: a meta-analysis. Med Care 2009 Aug;47(8):826-834 [FREE Full text] [doi: 10.1097/MLR.0b013e31819a5acc] [Medline: 19584762]
2. Howick J, Moscrop A, Mebius A, Fanshawe TR, Lewith G, Bishop FL, et al. Effects of empathic and positive communication in healthcare consultations: a systematic review and meta-analysis. J R Soc Med 2018 Jul;111(7):240-252 [FREE Full text] [doi: 10.1177/0141076818769477] [Medline: 29672201]
3. Riedl D, Schüßler G. The influence of doctor-patient communication on health outcomes: a systematic review. Z Psychosom Med Psychother 2017 Jun;63(2):131-150 [FREE Full text] [doi: 10.13109/zptm.2017.63.2.131] [Medline: 28585507]
4. Foundational competencies for undergraduate medical education. Association of American Medical Colleges, American Association of Colleges of Osteopathic Medicine, Accreditation Council for Graduate Medical Education. 2024. URL: https://image.email.aamc.org/lib/fe8e13727c63047f73/m/1/3d55a3aa-b3ba-4c14-a447-8c60e27a6f05.pdf [accessed 2025-09-18]
5. Learning opportunities, objectives and outcomes platform (LOOOP) Nationaler Kompetenzbasierter Lernzielkatalog Medizin. Wayback Machine. URL: https://web.archive.org/web/20240812213243/https://nklm.de/zend/objective/list/orderBy/@objectivePosition/studiengang/PF2/zeitsemester/2021/fachsemester/VIII.2.%20%C3%84rztliche%20Gespr%C3%A4chsf%C3%BChrung [accessed 2026-06-03]
6. Kienle R, Freytag J, Lück S, Eberz P, Langenbeck S, Sehy V, et al. Communication skills training in undergraduate medical education at Charité - Universitätsmedizin Berlin. GMS J Med Educ 2021 Mar 15;38(3):Doc56 [FREE Full text] [doi: 10.3205/zma001452] [Medline: 33824892]

XSL•FO
RenderX

7. Gysels M, Richardson A, Higginson IJ. Communication training for health professionals who care for patients with cancer: a systematic review of training methods. Support Care Cancer 2005 Jun;13(6):356-366. [doi: 10.1007/s00520-004-0732-0] [Medline: 15586302]
8. Lee J, Kim H, Kim KH, Jung D, Jowsey T, Webster CS. Effective virtual patient simulators for medical communication training: a systematic review. Med Educ 2020 Sep;54(9):786-795. [doi: 10.1111/medu.14152] [Medline: 32162355]
9. Kyaw BM, Posadzki P, Paddock S, Car J, Campbell J, Tudor Car L. Effectiveness of digital education on communication skills among medical students: systematic review and meta-analysis by the digital health education collaboration. J Med Internet Res 2019 Aug 27;21(8):e12967 [FREE Full text] [doi: 10.2196/12967] [Medline: 31456579]
10. Bender EM, Gebru T, McMillan-Major A, Shmitchell S. On the dangers of stochastic parrots: can language models be too big? In: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency. 2021 Presented at: FAccT '21; March 3-10, 2021; Virtual Event. [doi: 10.1145/3442188.3445922]
11. Østergaard SD. Will generative artificial intelligence chatbots generate delusions in individuals prone to psychosis? Schizophr Bull 2023 Nov 29;49(6):1418-1419. [doi: 10.1093/schbul/sbad128] [Medline: 37625027]
12. Looi JC, Allison S, Bastiampillai T, Reutens S, Looi RC. Illusions of intelligence, connection and reality: perils of large-language AI models for people with severe mental illness. Australas Psychiatry 2026 Feb;34(1):5-7. [doi: 10.1177/10398562251380544] [Medline: 40952393]
13. Wach K, Duong CD, Ejdys J, Kazlauskaitė R, Korzynski P, Mazurek G, et al. The dark side of generative artificial intelligence: a critical analysis of controversies and risks of ChatGPT. Entrep Bus Econ Rev 2023 Jun 30;11(2):7-30. [doi: 10.15678/EBER.2023.110201]
14. Millard J, Christie AP, Dicks LV, Isip JE, Johnson TF, Skinner G, et al. ChatGPT is likely reducing opportunity for support, friendship and learned kindness in research. Methods Ecol Evol 2024 Oct;15(10):1764-1766. [doi: 10.1111/2041-210X.14395]
15. Rudolph J, Tan S, Tan S. ChatGPT: bullshit spewer or the end of traditional assessments in higher education? J Appl Learn Teach 2023;6:342-363. [doi: 10.37074/JALT.2023.6.1.9]
16. Costello TH, Pennycook G, Rand DG. Durably reducing conspiracy beliefs through dialogues with AI. Science 2024 Sep 13;385(6714):eadq1814. [doi: 10.1126/science.adq1814] [Medline: 39264999]
17. Hornsey MJ, Pearson S, Bretter C, MacInnes S, Nylund JL, Rekker S. The promise and limitations of using GenAI to reduce climate scepticism. Nat Clim Change 2025 Sep 19;15:1183-1189. [doi: 10.1038/s41558-025-02425-8]
18. Johnson SB, King AJ, Warner EL, Aneja S, Kann BH, Bylund CL. Using ChatGPT to evaluate cancer myths and misconceptions: artificial intelligence and cancer information. JNCI Cancer Spectr 2023 Mar 01;7(2):pkad015 [FREE Full text] [doi: 10.1093/jncics/pkad015] [Medline: 36929393]
19. Potapenko I, Boberg-Ans LC, Stormly Hansen M, Klefter ON, van Dijk EH, Subhi Y. Artificial intelligence-based chatbot patient information on common retinal diseases using ChatGPT. Acta Ophthalmol 2023 Nov;101(7):829-831 [FREE Full text] [doi: 10.1111/aos.15661] [Medline: 36912780]
20. Levine DM, Tuwani R, Kompa B, Varma A, Finlayson SG, Mehrotra A, et al. The diagnostic and triage accuracy of the GPT-3 artificial intelligence model: an observational study. Lancet Digit Health 2024 Aug;6(8):e555-e561 [FREE Full text] [doi: 10.1016/S2589-7500(24)00097-9] [Medline: 39059888]
21. Sánchez-Rosenberg G, Magnéli M, Barle N, Kontakis MG, Müller AM, Wittauer M, et al. ChatGPT-4 generates orthopedic discharge documents faster than humans maintaining comparable quality: a pilot study of 6 cases. Acta Orthop 2024 Mar 21;95:152-156 [FREE Full text] [doi: 10.2340/17453674.2024.40182] [Medline: 38597205]
22. Zhou Z. Evaluation of ChatGPT's capabilities in medical report generation. Cureus 2023 Apr 14;15(4):e37589 [FREE Full text] [doi: 10.7759/cureus.37589] [Medline: 37197105]
23. Jeblick K, Schachtner B, Dexl J, Mittermeier A, Stüber AT, Topalis J, et al. ChatGPT makes medicine easy to swallow: an exploratory case study on simplified radiology reports. Eur Radiol 2024 May;34(5):2817-2825 [FREE Full text] [doi: 10.1007/s00330-023-10213-1] [Medline: 37794249]
24. Menezes MC, Hoffmann AF, Tan AL, Nalbandyan M, Omenn GS, Mazzotti DR, et al. The potential of Generative Pre-trained Transformer 4 (GPT-4) to analyse medical notes in three different languages: a retrospective model-evaluation study. Lancet Digit Health 2025 Jan;7(1):e35-e43. [doi: 10.1016/s2589-7500(24)00246-2]
25. Brügge E, Ricchizzi S, Arenbeck M, Keller MN, Schur L, Stummer W, et al. Large language models improve clinical decision making of medical students through patient simulation and structured feedback: a randomized controlled trial. BMC Med Educ 2024 Nov 28;24(1):1391 [FREE Full text] [doi: 10.1186/s12909-024-06399-7] [Medline: 39609823]
26. Cook DA, Overgaard J, Pankratz VS, Del Fiol G, Aakre CA. Virtual patients using large language models: scalable, contextualized simulation of clinician-patient dialogue with feedback. J Med Internet Res 2025 Apr 04;27:e68486 [FREE Full text] [doi: 10.2196/68486] [Medline: 39854611]
27. Hirosawa T, Yokose M, Sakamoto T, Harada Y, Tokumasu K, Mizuta K, et al. Utility of generative artificial intelligence for Japanese medical interview training: randomized crossover pilot study. JMIR Med Educ 2025 Aug 01;11:e77332 [FREE Full text] [doi: 10.2196/77332] [Medline: 40749190]
28. Holderried F, Stegemann-Philipps C, Herrmann-Werner A, Festl-Wietek T, Holderried M, Eickhoff C, et al. A language model-powered simulated patient with automated feedback for history taking: prospective study. JMIR Med Educ 2024 Aug 16;10:e59213 [FREE Full text] [doi: 10.2196/59213] [Medline: 39150749]

XSL•FO
RenderX

29. Holderried F, Stegemann-Philipps C, Herschbach L, Moldt JA, Nevins A, Griewatz J, et al. A generative pretrained transformer (GPT)-powered chatbot as a simulated patient to practice history taking: prospective, mixed methods study. JMIR Med Educ 2024 Jan 16;10:e53961 [FREE Full text] [doi: 10.2196/53961] [Medline: 38227363]
30. Park C, An MH, Hwang G, Park RW, An J. Clinical performance and communication skills of ChatGPT versus physicians in emergency medicine: simulated patient study. JMIR Med Inform 2025 Jul 17;13:e68409 [FREE Full text] [doi: 10.2196/68409] [Medline: 40674718]
31. Takahashi H, Shikino K, Kondo T, Yamada Y, Tomoda Y, Kishi M, et al. AI- vs human-based assessment of medical interview transcripts in a generative AI-simulated patient system: cross-sectional validation study. JMIR Med Educ 2026 Feb 17;12:e81673 [FREE Full text] [doi: 10.2196/81673] [Medline: 41701946]
32. Wang Z, Fan TT, Li ML, Zhu NJ, Wang XC. Feasibility study of using GPT for history-taking training in medical education: a randomized clinical trial. BMC Med Educ 2025 Jul 10;25(1):1030 [FREE Full text] [doi: 10.1186/s12909-025-07614-9] [Medline: 40640776]
33. Weisman D, Sugarman A, Huang YM, Gelberg L, Ganz PA, Comulada WS. Development of a GPT-4-powered virtual simulated patient and communication training platform for medical students to practice discussing abnormal mammogram results with patients: multiphase study. JMIR Form Res 2025 Apr 17;9:e65670 [FREE Full text] [doi: 10.2196/65670] [Medline: 40246299]
34. Yamamoto A, Koda M, Ogawa H, Miyoshi T, Maeda Y, Otsuka F, et al. Enhancing medical interview skills through AI-simulated patient interactions: nonrandomized controlled trial. JMIR Med Educ 2024 Sep 23;10:e58753 [FREE Full text] [doi: 10.2196/58753] [Medline: 39312284]
35. Kluger AN, DeNisi A. The effects of feedback interventions on performance: a historical review, a meta-analysis, and a preliminary feedback intervention theory. Psychol Bull 1996;119(2):254-284. [doi: 10.1037/0033-2909.119.2.254]
36. Strijbos JW, Narciss S, Dünnebier K. Peer feedback content and sender's competence level in academic writing revision tasks: are they critical for feedback perceptions and efficiency? Learn Instr 2010 Aug;20(4):291-303. [doi: 10.1016/j.learninstruc.2009.08.008]
37. Elwyn G, Frosch D, Thomson R, Joseph-Williams N, Lloyd A, Kinnersley P, et al. Shared decision making: a model for clinical practice. J Gen Intern Med 2012 Oct;27(10):1361-1367 [FREE Full text] [doi: 10.1007/s11606-012-2077-6] [Medline: 22618581]
38. Hettema J, Steele J, Miller WR. Motivational interviewing. Annu Rev Clin Psychol 2005;1:91-111. [doi: 10.1146/annurev.clinpsy.1.102803.143833] [Medline: 17716083]
39. Althof SE, Rosen RC, Perelman MA, Rubio-Aurioles E. Standard operating procedures for taking a sexual history. J Sex Med 2013 Jan;10(1):26-35. [doi: 10.1111/j.1743-6109.2012.02823.x] [Medline: 22970717]
40. Baile WF, Buckman R, Lenzi R, Glober G, Beale EA, Kudelka AP. SPIKES-a six-step protocol for delivering bad news: application to the patient with cancer. Oncologist 2000;5(4):302-311 [FREE Full text] [doi: 10.1634/theoncologist.5-4-302] [Medline: 10964998]
41. Ning Y, Ong JC, Cheng H, Wang H, Ting DS, Tham YC, et al. How can artificial intelligence transform the training of medical students and physicians? Lancet Digit Health 2025 Oct;7(10):100900 [FREE Full text] [doi: 10.1016/j.landig.2025.100900] [Medline: 41047321]
42. Flask. Pallets. URL: https://flask.palletsprojects.com [accessed 2025-06-14]
43. Voss A, Dettmer S, Tschorn M, Zoellick JC. Study demands and health status among medical students in two German universities. J Public Health (Berl.) 2024 Aug 27;34:999-1007. [doi: 10.1007/s10389-024-02336-6]
44. OpenAI Python API library. OpenAI. URL: https://github.com/openai/openai-python [accessed 2025-06-14]
45. Van Der Laan JD, Heino A, De Waard D. A simple procedure for the assessment of acceptance of advanced transport telematics. Transp Res C Emerg Technol 1997 Feb;5(1):1-10. [doi: 10.1016/s0968-090x(96)00025-3]
46. Davis FD, Bagozzi RP, Warshaw PR. User acceptance of computer technology: a comparison of two theoretical models. Manag Sci 1989 Aug 1;35(8):982-1003. [doi: 10.1287/mnsc.35.8.982]
47. Neyer FJ, Felber J, Gebhardt C. Entwicklung und validierung einer kurzskala zur erfassung von technikbereitschaft. Diagnostica 2012 Apr;58(2):87-99. [doi: 10.1026/0012-1924/a000067]
48. Schulz von Thun F. Miteinander reden 1 Störungen und Klärungen: Allgemeine Psychologie der Kommunikation. Hamburg, Germany: Rowolth; 1981.
49. Hall JA, Horgan TG, Murphy NA. Nonverbal communication. Annu Rev Psychol 2019 Jan 04;70(1):271-294. [doi: 10.1146/annurev-psych-010418-103145] [Medline: 30256720]
50. Henry SG, Fuhrel-Forbis A, Rogers MA, Eggly S. Association between nonverbal communication during clinical interactions and outcomes: a systematic review and meta-analysis. Patient Educ Couns 2012 Mar;86(3):297-315. [doi: 10.1016/j.pec.2011.07.006] [Medline: 21824738]
51. Hostetter AB. When do gestures communicate? A meta-analysis. Psychol Bull 2011 Mar;137(2):297-315. [doi: 10.1037/a0022128] [Medline: 21355631]
52. Laukka P, Elfenbein HA. Cross-cultural emotion recognition and in-group advantage in vocal expression: a meta-analysis. Emot Rev 2020 Feb 13;13(1):3-11. [doi: 10.1177/1754073919897295]

XSL•FO
RenderX

53. Leung Y, Oates J, Chan SP. Voice, articulation, and prosody contribute to listener perceptions of speaker gender: a systematic review and meta-analysis. J Speech Lang Hear Res 2018 Feb 15;61(2):266-297. [doi: 10.1044/2017_jslhr-s-17-0067]

Abbreviations

AI: artificial intelligence
API: application programming interface
MI: motivational interviewing
PAF: perceived adequacy of feedback
PCC: perceived communication competence
PEOU: perceived ease of use
PU: perceived usefulness
SDM: shared decision-making
TC: technology commitment

Edited by S Zelko, M Montagna; submitted 19.Nov.2025; peer-reviewed by L Gelberg, D Yang; comments to author 09.Feb.2026; revised version received 04.May.2026; accepted 15.May.2026; published 12.Jun.2026.

Please cite as:
Müller Y, Lemberg H, Gellert P, Zoellick JC
Training Empathetic Communication Skills in Medical Students With a Role-Prompted GPT-4o Chatbot: Quasi-Experimental Intervention Study
JMIR Med Educ 2026;12:e88092
URL: https://mededu.jmir.org/2026/1/e88092
doi:10.2196/88092
PMID:

©Yannik Müller, Hans Lemberg, Paul Gellert, Jan C Zoellick. Originally published in JMIR Medical Education (https://mededu.jmir.org), 12.Jun.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Education, is properly cited. The complete bibliographic information, a link to the original publication on https://mededu.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Voice Cloning Using AI vs Traditional Audio Recording for Prerecorded Courses in Medical Pedagogy: Randomized Controlled Trial

Antoine Gavoille[1,2], MD; Fabien Subtil[1,2], PhD; Mikaïl Nourredine[1,2], MD

[1]Laboratoire de Biométrie et Biologie Evolutive, Université Claude Bernard Lyon 1, Lyon, Rhône-Alpes, France

[2]Service de Biostatistiques-Bioinformatiques, Hospices Civils de Lyon, Lyon, France

Corresponding Author:
Antoine Gavoille, MD
Laboratoire de Biométrie et Biologie Evolutive
Université Claude Bernard Lyon 1
Lyon, Rhône-Alpes,
France
Phone: 33 4 72 11 57 49
Email: ant.gavoille@gmail.com

Abstract

Background: Prerecorded courses are increasingly used in medical education, and audio quality is known to influence learners' comprehension and engagement. Traditional audio recording, however, is time-consuming and may be uncomfortable for some educators. Advances in generative artificial intelligence (AI) now allow for realistic voice cloning, but its pedagogical value compared with conventional recording has not been assessed.

Objective: This study aimed to evaluate the usefulness and perception of AI-based voice cloning for prerecorded courses in medical pedagogy compared with traditional audio recording.

Methods: We conducted a randomized trial among fourth- and fifth-year medical students at a French university. Participants accessed four 10-minute prerecorded lectures on critical appraisal of medical research. The control group received lectures with audio recorded by the teacher, whereas the intervention group received audio generated from the teacher's cloned voice using a commercial AI text-to-speech system, with identical slides and scripts. The primary outcome was the total score on 2 online tests (17 multiple-choice questions on knowledge acquisition and 15 multiple-choice questions on knowledge application). Secondary outcomes included satisfaction ratings, course viewing metrics, and production time.

Results: A total of 88 students were randomized, and 64 (72.7%) watched at least 15 seconds of video and, thus, were included in the modified intention-to-treat population. Mean total test scores did not differ significantly between the AI voice cloning and audio recording groups (51.2, SD 4.9 vs 51.8, SD 6.9 out of 100; adjusted mean difference −0.9, 95% CI −2.7 to 4.4; *P*=.60). Satisfaction was high in both groups. Production time was shorter with AI (22.5, SD 6.45 vs 35, SD 7.07 minutes per video).

Conclusions: We did not detect a significant difference in learning outcomes or satisfaction between AI voice cloning and conventional recording, whereas AI voice cloning reduced production time, making it a practical alternative for prerecorded medical courses. Nevertheless, some students may perceive synthetic voices as less authentic, representing a potential barrier to widespread adoption.

(JMIR Med Educ 2026;12:e86569) doi:10.2196/86569

KEYWORDS

medical pedagogy; prerecorded course; generative artificial intelligence; voice cloning; audio quality; artificial intelligence; AI

Introduction

The growing integration of online resources into medical education has led to an increasing reliance on prerecorded video courses [1,2]. This evolution, accelerated by the COVID-19 pandemic, has necessitated a rapid transition to virtual learning environments. Consequently, it is essential to ensure the effectiveness and quality of online education to meet the needs of modern learners.

XSL•FO
RenderX

Audio quality is a key factor contributing to the success of online learning. Audio is the primary channel for conveying information in prerecorded courses, and audio quality can affect a student's ability to understand the material and stay engaged. Furthermore, the perceived audio quality could influence a student's unconscious assessment of the overall quality of a video [3]. Research on cognitive fluency, although not in an educational context, has shown that, when audio is difficult to process due to extraneous variables such as poor quality, it can lead to less favorable evaluations of the speaker and content and even influence the memorization of key facts [4,5]. This underscores the importance of considering audio quality as a critical variable in educational content.

The traditional method of recording audio for prerecorded courses presents several challenges for educators. It is a time-consuming process that requires access to studios and specialized equipment to achieve high quality. In practice, however, many teachers record on their PCs, using low-quality microphones and working in environments that are not always favorable to good sound capture. In addition, some educators may not be comfortable with the recording process, leading to diction or speech issues that can negatively impact the learning experience. Emerging technologies such as voice cloning using artificial intelligence (AI) offer a promising solution to these challenges [6]. This technology allows for the creation of a high-quality prerecorded audio track that retains the familiar voice of the educator, which students can recognize. With the recent advances in AI and deep learning, voice synthesis technology is evolving rapidly and can now produce voices so realistic that it is often impossible to distinguish them from actual human speech [7,8]. This approach has several advantages, including the ability to reduce production time and ensure consistent, high-quality sound without the need for a physical studio [9]. In addition, it allows for the correction of speech imperfections, improving clarity and intelligibility. However, no studies have specifically evaluated modern AI voice cloning for prerecorded lectures.

The aim of this study was to evaluate the usefulness and perception of AI voice cloning for generating audio in prerecorded video courses compared with traditional voice recording in medical training. The primary objective was to compare student learning outcomes between the 2 methods; secondary objectives included satisfaction, engagement, and production time. We hypothesized that traditional recording might perform better than AI-generated voice cloning in terms of student learning outcomes and satisfaction, whereas AI voice cloning would offer practical advantages, including reduced production time.

Methods

Study Design

This was a single-center randomized study conducted at a French university. Given the nature of the study, full blinding was not feasible, so students were aware that they were participating in a study but were not informed that the purpose of the study concerned the quality of the audio recordings to preserve ecological validity and avoid attention bias. Neither the participants nor the researchers responsible for analyzing the data were aware of the group assignments. All prerecorded lectures and assessments were hosted in a single online environment, the Moodle platform, an institutional learning management system used by the University of Lyon. Moodle provided a secure environment for delivering course materials, tracking user activity (total watch time and number of views), and collecting survey responses. Data were exported in anonymized form for analysis.

Participants

Fourth- and fifth-year medical students were recruited on a voluntary basis to participate in this study. Volunteer students were randomized into 2 groups at a 1:1 ratio at the start of the study stratified by year. Participants who did not have a stable internet connection could have one provided by the faculty if needed. The conditions under which participants viewed the prerecorded courses were not standardized, each student being free to watch the courses in their own uncontrolled environment.

Ethical Considerations

All participating students provided written informed consent prior to inclusion and were informed that they could withdraw their consent at any time without consequence. At the end of the study, the objective was disclosed to all participants during a group debriefing session conducted via videoconference, and an in-person debriefing was offered but was not requested by any participant. Voice cloning was performed exclusively by the teacher (AG) using a personal ElevenLabs account under his full informed consent. ElevenLabs requires voice identity verification prior to cloning, ensuring that only the voice owner can initiate the process. The voice recordings used for cloning were processed by ElevenLabs as a data processor under its Data Processing Addendum in compliance with the European Union General Data Protection Regulation. According to ElevenLabs' privacy policy, raw voice data are retained for up to 3 years after the last interaction; the teacher retained the right to request data deletion at any time. Beyond ElevenLabs as the processor, no other party had access to the voice data. The study protocol was approved by an ethics committee (*Comité d'Ethique de la Recherche* of the University of Lyon; reference: 2025 06-12-002). This report follows the CONSORT (Consolidated Standards of Reporting Trials) guidelines.

Intervention

Participants had online access to 4 prerecorded video courses, each lasting approximately 10 minutes, on the topic of critical appraisal skills in medical research and epidemiology. They could watch the courses at their convenience as many times as they wanted during a period of 2 months. The slides and text scripts were identical for both groups, with the only difference being the nature of the audio. The control group received lectures with traditional audio recordings produced under optimized nonprofessional conditions representative of typical academic settings: a lapel RODE Lavalier GO microphone was used, the instructor followed basic recording guidelines, and nonexpert postproduction editing (noise reduction and level normalization) was applied without access to a professional recording studio. For the intervention group, the teacher's voice

XSL•FO
RenderX

was cloned using the Professional Voice Cloning system from ElevenLabs [10] trained on 30 minutes of audio recorded using the same microphone that was distinct from the course content itself. Audio was then generated by an AI text-to-speech tool, the Turbo 2.5 model from ElevenLabs [11], using exactly the same text script. The following generation parameters were applied consistently across all AI-generated audio files: speech rate of 1.15, similarity of 80%, and stability of 40%. Audio samples generated using the 2 methods are provided in Multimedia Appendix 1.

Outcome

One month after access to the videos was granted, students were given access to 2 online tests: a knowledge acquisition test with 17 multiple-choice questions on the content of the prerecorded lectures and a knowledge application test with 15 multiple-choice questions on a medical research article. Each question included 5 items, with 1 point per correct answer. In addition, participants were asked to complete a satisfaction questionnaire immediately after viewing the course, evaluating overall satisfaction, audio quality, voice, and attention maintenance throughout the videos on a 4-point Likert scale. The time required for the teacher to produce the audio track and participants' viewing metrics (duration and number of views) were also collected for each video.

The primary outcome was the total score on the 2 tests, expressed as a score out of 100. Secondary outcomes were the score on the knowledge acquisition and application tests separately, each item of the satisfaction questionnaire (averaged across all the prerecorded course videos watched by the participant), and the number of prerecorded course videos watched entirely (>90%). Finally, we evaluated the mean time required to produce the audio recording for each video, defined as the total active time from slides without audio to final video with recorded sound. For the traditional recording method, this included script drafting, audio recording (including failed takes), and postproduction editing. For the AI voice cloning method, this included refining the script into a precise written text (with particular attention to punctuation, which influences text-to-speech prosody) and audio-video synchronization; the audio generation step (approximately 2-3 minutes per video, requiring no instructor involvement) was excluded.

Statistical Analysis

A power calculation was performed to determine the sample size required to detect a significant difference between the 2 groups. The calculation was based on a 2-sided hypothesis test with an α level of 5% and a power of 80%. For an expected score of 60 points in the control group with an SD of 15 points (based on an expected distribution of scores with a first quartile at 50 points and a third quartile at 70 points) to detect a minimum difference of 10 points between the control group and the intervention group and including a 20% loss to follow-up, a minimum of 45 participants per group was required for the study, for a total of 90 students.

Quantitative variables are presented as means and SDs, and categorical variables are presented as counts and percentages. Outcomes were analyzed in the modified intention-to-treat (mITT) population, corresponding to participants who watched at least 15 seconds of 1 prerecorded course video. This threshold was chosen based on viewing analytics showing a first sharp drop before this point, reflecting accidental or exploratory accesses without genuine viewing intent, and was also considered too short for a participant to consciously identify an AI-generated voice and disengage. For participants who responded to only 1 of the 2 knowledge tests (n=8), their total score was calculated from the available test alone, rescaled to 100 points using the same formula as for each individual test. This approach is equivalent to imputing the missing test score by the participant's score on the completed test and assumes that performance on the missing test would have been, on average, equivalent to that on the completed test. As a sensitivity analysis, the primary analysis was repeated in the subgroup of participants who completed both tests (n=42). Other missing data were not imputed, and only complete cases were analyzed for each outcome. Knowledge test scores were analyzed using a linear regression adjusted for the intervention or control group and year of study. Other categorical outcomes were compared between the intervention and control groups using a nonparametric Fisher exact test. *P* values below .05 were considered statistically significant. Analyses were performed using the R software (version 4.4.2; R Foundation for Statistical Computing) [12].

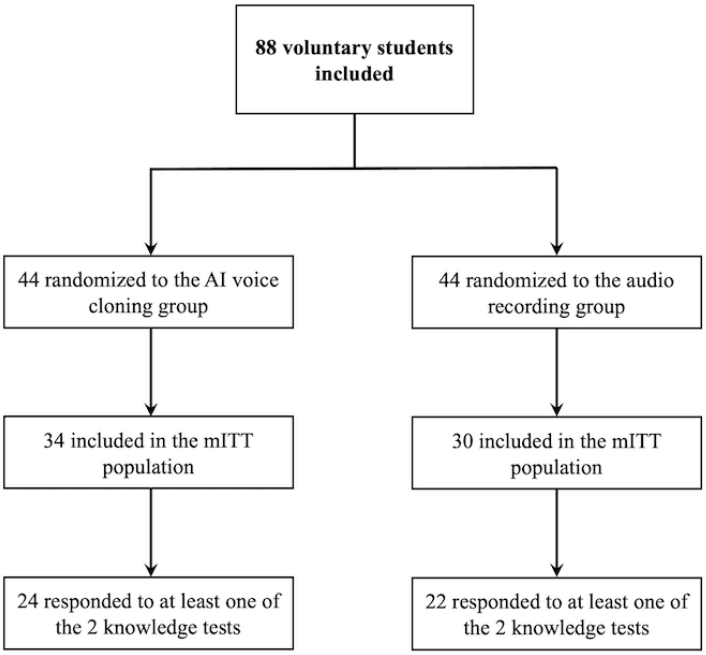
Results

Overview

A total of 88 participants took part in June 2025, of whom 44 (50%) were randomized to the audio recording group and 44 (50%) were randomized to the AI voice cloning group; the study flowchart is shown in Figure 1. Of these 88 participants, 64 (72.7%) watched at least 15 seconds of 1 prerecorded course video and, thus, were included in the mITT population: 34 (53.1%) participants in the AI voice cloning group and 30 (46.9%) in the audio recording group. Participant characteristics by study group are shown in Table 1. In total, 71.9% (46/64) of the participants were women, and 68.8% (44/64) were in fifth year. Year of study was well balanced in the initially randomized population, whereas there were more fourth-year students in the AI voice cloning group in the mITT population. Regarding missing data, 78.1% (50/64) of the participants responded to at least one of the knowledge tests, including 12.5% (8/64) who responded to only 1 of the 2 tests, based on which their total score was calculated, and 93.8% (60/64) who responded to at least one satisfaction questionnaire.

XSL•FO
RenderX

Figure 1. Study flowchart. The modified intention-to-treat (mITT) population was defined as the population who watched at least 15 seconds of a prerecorded course video. AI: artificial intelligence.



XSL•FO
RenderX

Table 1. Participant characteristics and outcomes in the modified intention-to-treat population (N=64).

	Total	Randomization group	
		AI[a] voice cloning (n=34)	Audio recording (n=30)
Participant characteristics, n (%)			
Men	18 (28.1)	9 (26.5)	9 (30)
Year of study			
Fourth	20 (31.3)	7 (20.6)	13 (43.3)
Fifth	44 (68.8)	27 (79.4)	17 (56.7)
Primary outcome			
Total score (out of 100)[b]			
Mean (SD)	51.5 (6.0)	51.2 (4.9)	51.8 (6.9)
Median (IQR)	51.3 (48.1-55.0)	51.3 (48.1-53.8)	51.3 (48.8-56.0)
Missing, n (%)	14 (21.9)	9 (26.5)	5 (16.7)
Secondary outcomes			
Knowledge acquisition test score (out of 100)			
Mean (SD)	53.5 (5.6)	54.3 (4.9)	52.6 (6.3)
Median (IQR)	54.1 (49.4-56.5)	54.7 (50.6-58.8)	51.8 (48.2-56.5)
Missing, n (%)	18 (28.1)	10 (29.4)	8 (26.7)
Knowledge application test score (out of 100)			
Mean (SD)	49.3 (8.7)	47.0 (6.0)	51.5 (10.2)
Median (IQR)	47.3 (44.0-54.7)	46.7 (42.7-50.7)	50.0 (44.7-57.3)
Missing, n (%)	18 (28.1)	12 (35.3)	6 (20)
Number of prerecorded course videos watched (>90%), n (%)			
0	2 (3.1)	1 (2.9)	1 (3.3)
1	3 (4.7)	2 (5.9)	1 (3.3)
2	5 (7.8)	2 (5.9)	3 (10)
3	5 (7.8)	3 (8.8)	2 (6.7)
4	49 (76.6)	26 (76.5)	23 (76.7)

[a]AI: artificial intelligence.

[b]Primary outcome: total score (out of 100) on the 2 knowledge tests (knowledge acquisition+knowledge application).

Participants' Results on the Knowledge Test Scores

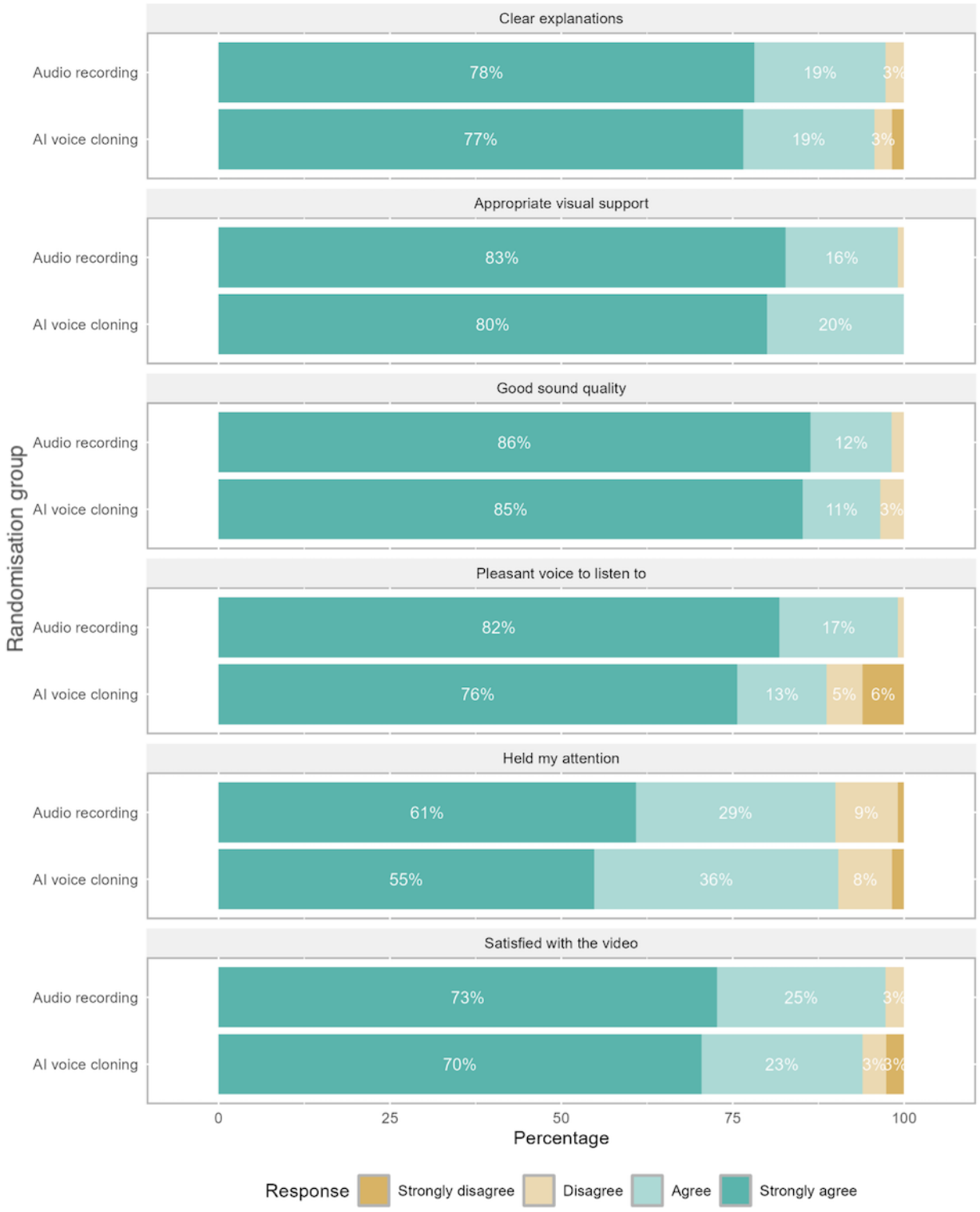
Primary and secondary outcome results are shown in Table 1. The mean total score on the 2 knowledge tests was 51.2 (SD 4.9) points in the AI voice cloning group and 51.8 (SD 6.9) in the audio recording group, with no significant difference between the 2 groups (adjusted mean difference −0.9, 95% CI −2.7 to 4.4; *P*=.60). On the knowledge acquisition test, the mean score was 54.3 (SD 4.9) in the AI voice cloning group and 52.6 (SD 6.3) in the audio recording group (adjusted mean difference 1.2, 95% CI −2.2 to 4.6; *P*=.50). On the knowledge application test, the mean score was 47.0 (SD 6.0) in the AI voice cloning group and 51.5 (SD 10.2) in the audio recording group (adjusted mean difference −4.9, 95% CI −7.7 to 0.4; *P*=.07).

Satisfaction Questionnaire

Over the 4 prerecorded course videos, participants had an overall high satisfaction level. Averaged across participants, a mean proportion of 98% agreed or strongly agreed that the prerecorded course videos had good sound quality, 96% agreed or strongly agreed that the explanations were clear, 94% agreed or strongly agreed that the voice was pleasant to listen to, 90% agreed or strongly agreed that the video held their attention, and 96% agreed or strongly agreed that they were globally satisfied with the videos. We found no significant difference in the proportion of "agree" or "strongly agree" responses to satisfaction items between the AI voice cloning group and the audio recording group (Figure 2). However, a mean proportion of 11% disagreed or strongly disagreed that the voice was pleasant to listen to in the AI voice cloning group compared to 1% in the audio recording group, but the difference was not significant (*P*=.24). Satisfaction results for each prerecorded course video were similar to the overall satisfaction results (Multimedia Appendix 2).

XSL•FO
RenderX

Figure 2. Satisfaction level by study group. AI: artificial intelligence.



Other Outcomes

Overall, 76.6% (49/64) of the participants watched the 4 prerecorded course videos entirely, with no significant difference between the AI voice cloning and audio recording groups (P=.98). The AI voice cloning method was faster for generating the audio, with a mean of 22.5 (SD 6.45) minutes per video for AI voice cloning and 35 (SD 7.07) minutes per video for audio recording, representing a 36% reduction.

Sensitivity Analysis

The results of the analysis restricted to participants who completed both tests (42/64, 65.6%) were consistent with those of the main analysis, with an adjusted mean difference of 2.3 in the total score between groups (95% CI −1.3 to 5.9; P=.20).

XSL•FO
RenderX

Discussion

In this randomized controlled study comparing AI-based voice cloning with traditional audio recording for prerecorded medical courses, we found no significant difference in students' knowledge acquisition and application, satisfaction, or engagement, whereas AI voice cloning allowed for a reduction in production time.

Recent advances in generative AI for text to speech have brought synthetic voices to a level of realism that can effectively compete with traditional human recordings. Modern systems are now capable of reproducing natural prosody and timbre with high fidelity, making it increasingly difficult for listeners to distinguish between cloned and real voices [8,13]. In this context, our results suggest that the substitution of human recordings through AI voice cloning did not compromise learning effectiveness. This stands in contrast to studies on degraded audio, which have shown that poor sound quality can impair comprehension, reduce trust in the speaker, and negatively affect performance [3,4]. In our study, both methods provided sufficiently clear audio to avoid such detrimental effects, indicating that the shift from traditional audio recording to AI synthesis did not perturb the educational process.

Student satisfaction was generally high across groups, confirming that audio quality, whether cloned or recorded, was adequate. It should be noted that traditional audio recordings were produced under carefully optimized real-world conditions, including the use of a lapel microphone and basic postproduction editing. The relative advantages of AI voice cloning are likely greater in settings with poorer recording conditions, such as built-in microphone use or no postproduction. Our results may therefore underestimate the practical benefit of voice cloning for educators with limited access to recording equipment or technical support. However, a nonsignificant trend suggested that a slightly lower proportion of students in the AI group found the voice pleasant. This aligns with prior observations that, despite advances in voice synthesis, subtle elements of natural prosody and emotional tone remain difficult to replicate [5]. Student perceptions of AI technology may also influence their satisfaction and, consequently, their attention and engagement if the "illusion" of naturalness is not fully achieved. In our study, while overall satisfaction remained high, one participant explicitly identified the voice as AI generated and expressed marked discontent. This reaction illustrates that, even when objective measures of learning are unaffected, individual opinions on AI can shape the learner's experience. Such perceptions may represent an obstacle to the broader acceptance of voice cloning technologies in medical education, especially if students feel that artificiality compromises authenticity [14].

From a practical perspective, the reduced production time associated with AI voice cloning could represent a meaningful advantage for educators. Traditional recording often requires repeated takes, technical adjustments, and postprocessing to obtain an acceptable result, which can be discouraging for teachers who are not comfortable with oral performance or technical aspects of recording. Diction problems, hesitations, or lack of confidence can further extend the process and reduce the overall quality of the material. In our study, the time required to produce a 10-minute video was reduced by approximately one-third using AI synthesis. Beyond raw recording time, preparation and rerecording can accumulate into several hours for a series of courses, whereas AI allows for rapid corrections and consistent results with minimal additional effort.

AI voice cloning can also have additional advantages for educators. It could allow for switching of the cloned voice if a teacher is no longer involved in the course, thereby maintaining a sense of continuity and social connection between students and their current instructors. Moreover, these systems could integrate flexible options, for instance, adjusting the pace of speech while preserving natural prosody for students who wish to watch the course at a different speed rather than simply speeding up or slowing down the audio, which results in a deterioration of quality [15]. Finally, AI voice cloning could also enhance accessibility by providing clearer diction for students with hearing difficulties, tailored pronunciation for nonnative speakers, or even multilingual courses in languages that the lecturer does not speak [16], thus widening the inclusiveness of digital pedagogy. However, a limitation remains with respect to video: while audio can now be convincingly synthesized, generative technologies are not yet mature enough to produce realistic video sequences of the teacher speaking, although such developments are likely forthcoming [14,17].

Further research should explore how AI voice cloning interacts with other aspects of digital pedagogy, such as visual presence, interactivity, and feedback dynamics [15]. Experimental designs combining AI-generated audio with avatars or virtual instructors could help understand how different modalities influence motivation, empathy, and cognitive load [14,17]. Longitudinal studies are also needed to assess whether repeated exposure to synthetic voices affects learners' engagement or recall. Finally, cost-effectiveness analyses could quantify the institutional benefits of AI-assisted production and inform decisions about large-scale adoption in medical schools.

The use of AI voice cloning in educational settings raises important ethical considerations regarding teachers' consent and responsibility. In our study, voice cloning was performed by the teacher himself, who retained full control over the process and the resulting synthetic voice. We believe that this model, where the teacher whose voice is cloned is also the one operating the tool, represents the most ethical approach as it preserves both consent and accountability. Delegating voice cloning to a third party would raise substantially more complex questions regarding ownership, storage, and the potential for misuse of a person's vocal identity. We therefore recommend that AI voice cloning in medical education be performed exclusively by the instructor themselves, who bears responsibility for both the content and the audio identity of the course material.

This study has several limitations. First, although the sample size calculation incorporated a 20% loss to follow-up margin, an additional 21.9% (14/64) of the mITT participants had missing data for the primary outcome, resulting in the analysis of only 78.1% (50/64) of the participants. This reduced the statistical power to detect small differences; however, the observed between-group difference was less than 1 point, well

XSL•FO
RenderX

below the minimum clinically relevant threshold of 10 points, suggesting that the null result is not solely attributable to insufficient power. Furthermore, 12.5% (8/64) of the participants completed only 1 of the 2 knowledge tests; their total score was derived from the available test alone. Although the assumption of interchangeability between the 2 tests is supported by their conceptual proximity, this imputation remains imperfect given that the tests assess distinct constructs (knowledge recall vs knowledge application) and that total scores derived from a single test have higher measurement variance, which may introduce minor heterogeneity in the primary outcome measure. No data were collected on the reasons for nonviewing or early dropout, but this likely reflects the real-world study conditions: video viewing was voluntary and took place during a period without formal classes, when students were engaged in hospital internships and, for fifth-year students, intensive examination preparation. Although the complete-case analysis would yield an unbiased estimate under a missing completely at random or missing at random mechanism conditional on baseline covariates, a missing not at random mechanism related to group allocation cannot be excluded, for instance, if students who disengaged from the AI voice cloning group due to perceived artificiality were also less likely to complete the tests. A sensitivity analysis restricted to participants who completed both tests yielded consistent results. Second, although participants were not informed that the study concerned audio quality, blinding cannot be guaranteed as one participant explicitly identified the voice as AI generated. The study protocol included a poststudy questionnaire assessing whether participants had guessed their group allocation; however, the response rate was insufficient (6/64, 9.4%), and these data could not be reported. This is unlikely to have affected knowledge test scores, which were automatically graded, but may have influenced self-reported satisfaction outcomes. Third, the study was conducted in an uncontrolled environment, reflecting real-world conditions but introducing variability in viewing practices. Fourth, no baseline knowledge assessment was performed before the intervention, which would have improved statistical precision. This design choice was made deliberately to limit participant burden in the context described above. Finally, our study was designed as a superiority trial, so the absence of a statistically significant difference does not demonstrate equivalence between methods. The 95% CI excluded the prespecified minimum clinically relevant difference, suggesting that AI voice cloning may represent a practical alternative for prerecorded medical courses. However, the knowledge application test showed a nonsignificant trend favoring traditional audio, which could suggest that subtle differences in prosody or intonation might influence higher-order reasoning; this warrants further investigation. Future studies should explore the impact of voice cloning in larger, more diverse populations and in different pedagogical contexts.

In conclusion, our trial did not detect a significant difference in learning outcomes or satisfaction between AI-based voice cloning and traditional recording for prerecorded course videos in medical pedagogy. While AI-based voice cloning does not improve learning outcomes compared with human audio, it offers a practical advantage in terms of efficiency without compromising student satisfaction or knowledge acquisition. However, its adoption may still face resistance as some students remain sensitive to the artificial nature of synthetic voices and may perceive them as less authentic. Addressing this challenge of global acceptance will be essential to integrate AI-generated audio into medical education.

Funding

This study received no funding.

Data Availability

Data are available from the corresponding author on reasonable request.

Authors' Contributions

Conceptualization: AG, FS, MN
Data curation: AG, MN
Formal analysis: AG, MN
Investigation: AG, MN
Methodology: AG, MN
Project administration: FS
Supervision: FS
Visualization: AG, MN
Writing—original draft: AG
Writing—review and editing: FS, MN
All authors approved the final manuscript and agree to be accountable for all aspects of the work.

Conflicts of Interest

None declared.

XSL•FO
RenderX

Multimedia Appendix 1
Samples of audio recording and artificial intelligence (AI) voice cloning. Short video presenting 2 consecutive audios from one of the prerecorded lectures: the traditional recording (0:01-0:07) followed by the AI voice cloning version (0:07-0:13).
[MP4 File (MP4 Video), 578 KB - mededu_v12i1e86569_app1.mp4]

Multimedia Appendix 2
Satisfaction level by study group for each prerecorded course video.
[DOCX File , 910 KB - mededu_v12i1e86569_app2.docx]

Multimedia Appendix 3
CONSORT 2025 checklist.
[PDF File (Adobe PDF File), 134 KB - mededu_v12i1e86569_app3.pdf]

References

1. Tate T, Warschauer M. Equity in online learning. Educ Psychol 2022 Jun 10;57(3):192-206. [doi: 10.1080/00461520.2022.2062597]
2. Cheung YY, Lam KF, Zhang H, Kwan CW, Wat KP, Zhang Z, et al. A randomized controlled experiment for comparing face-to-face and online teaching during COVID-19 pandemic. Front Educ 2023 Jun 1;8:1160430. [doi: 10.3389/feduc.2023.1160430]
3. Puyana-Romero V, Díaz-Márquez AM, Ciaburro G, Hernández-Molina R. The acoustic environment and university students' satisfaction with the online education method during the COVID-19 lockdown. Int J Environ Res Public Health 2022 Dec 30;20(1):709 [FREE Full text] [doi: 10.3390/ijerph20010709] [Medline: 36613032]
4. Newman EJ, Schwarz N. Good sound, good research: how audio quality influences perceptions of the research and researcher. Sci Commun 2018 Mar 20;40(2):246-257. [doi: 10.1177/1075547018759345]
5. Bild E, Redman A, Newman EJ, Muir BR, Tait D, Schwarz N. Sound and credibility in the virtual court: low audio quality leads to less favorable evaluations of witnesses and lower weighting of evidence. Law Hum Behav 2021 Oct;45(5):481-495. [doi: 10.1037/lhb0000466] [Medline: 34871019]
6. Arik SO, Chen J, Peng K, Ping W, Zhou Y. Neural voice cloning with a few samples. arXiv Preprint posted online on February 14, 2018. [doi: 10.48550/arXiv.1802.06006]
7. Hu W, Zhu X. A real-time voice cloning system with multiple algorithms for speech quality improvement. PLoS One 2023 Apr 3;18(4):e0283440 [FREE Full text] [doi: 10.1371/journal.pone.0283440] [Medline: 37011068]
8. Perrotin O, Stephenson B, Gerber S, Bailly G, King S. Refining the evaluation of speech synthesis: a summary of the Blizzard Challenge 2023. Comput Speech Lang 2025 Mar;90:101747. [doi: 10.1016/j.csl.2024.101747]
9. Mogali SR, Ng O, Tan JX, San TH, Ng KB. Voice-over anatomy lectures created by AI-voice cloning technology: a descriptive article. Anat Sci Educ 2024 Oct 22;17(9):1686-1693. [doi: 10.1002/ase.2524]
10. Voice cloning: how it works. ElevenLabs. URL: https://elevenlabs.io/docs/eleven-api/concepts/voice-cloning [accessed 2026-06-17]
11. Introducing Eleven Turbo v2.5. ElevenLabs. URL: https://elevenlabs.io/blog/introducing-turbo-v25 [accessed 2026-06-17]
12. R Core Team. R: a language and environment for statistical computing. R Foundation for Statistical Computing. 2022. URL: https://www.R-project.org/ [accessed 2026-06-17]
13. Bruder C, Breda P, Larrouy-Maestri P. Attractive synthetic voices. Comput Hum Behav Artif Hum 2025 Dec;6:100211. [doi: 10.1016/j.chbah.2025.100211]
14. Netland T, von Dzengelevski O, Tesch K, Kwasnitschka D. Comparing human-made and AI-generated teaching videos: an experimental study on learning effects. Comput Educ 2025 Jan;224:105164. [doi: 10.1016/j.compedu.2024.105164]
15. Ahn D, Chan JC. What drives student engagement and learning in video lectures? An investigation of instructor visibility, playback speed, and student preferences. Appl Cogn Psychol 2025 Feb 25;39(2):e70026. [doi: 10.1002/acp.70026]
16. Pérez A, Díaz-Munío GG, Giménez A, Silvestre-Cerdà JA, Sanchis A, Civera J, et al. Towards cross-lingual voice cloning in higher education. Eng Appl Artif Intell 2021 Oct;105:104413. [doi: 10.1016/j.engappai.2021.104413]
17. Xu T, Liu Y, Jin Y, Qu Y, Bai J, Zhang W, et al. From recorded to AI-generated instructional videos: a comparison of learning performance and experience. Br J Educ Technol 2024 Oct 28;56(4):1463-1487. [doi: 10.1111/bjet.13530]

Abbreviations

AI: artificial intelligence
CONSORT: Consolidated Standards of Reporting Trials
mITT: modified intention-to-treat

XSL•FO
RenderX

Edited by P Kanzow; submitted 27.Oct.2025; peer-reviewed by F Fukuzawa, J Wilkinson; comments to author 30.Mar.2026; revised version received 18.May.2026; accepted 19.May.2026; published 02.Jul.2026.

Please cite as:
Gavoille A, Subtil F, Nourredine M
Voice Cloning Using AI vs Traditional Audio Recording for Prerecorded Courses in Medical Pedagogy: Randomized Controlled Trial
JMIR Med Educ 2026;12:e86569
URL: https://mededu.jmir.org/2026/1/e86569
doi:10.2196/86569
PMID:42390914

©Antoine Gavoille, Fabien Subtil, Mikaïl Nourredine. Originally published in JMIR Medical Education (https://mededu.jmir.org), 02.Jul.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Education, is properly cited. The complete bibliographic information, a link to the original publication on https://mededu.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

AI-Generated Microlearning for Plastic Surgery Residency: Single-Arm Pre-Post Feasibility Study

Marius Drysch[1], MD; Sonja Verena Schmidt[1], MD; Felix Reinkemeier[1], MD; Flemming Puscz[1], MD; Alexander Fiedler[1], MD; Maria Fueth[1], MD; Marcus Lehnhardt[1], MD; Alexander Sogorski[1], MD; Christoph Wallner[1], MD, PhD

Department of Plastic Surgery, BG University Hospital Bergmannsheil Bochum, Bochum, North Rhine-Westphalia, Germany

Corresponding Author:
Marius Drysch, MD
Department of Plastic Surgery
BG University Hospital Bergmannsheil Bochum
Bürkle de la Camp-Platz 1
Bochum, North Rhine-Westphalia, 44789
Germany
Phone: 49 234 302 0
Email: marius.drysch@rub.de

Abstract

Background: Surgical residency training faces mounting pressure from expanding subspecialty knowledge requirements alongside compressed learning opportunities. Microlearning, defined as brief, focused educational episodes, offers a pedagogically sound approach suited to fragmented clinical schedules, while large language models present capabilities for scalable content generation. However, the feasibility, quality, and educational value of large language model–generated microlearning in surgical training remain unexplored.

Objective: This study aimed to assess the feasibility and acceptability of artificial intelligence (AI)-generated microlearning modules among plastic surgery residents, evaluate content quality through systematic faculty evaluation, and generate preliminary effect size estimates for self-perceived knowledge and confidence changes.

Methods: A single-arm pilot feasibility study with a pre-post design was conducted at an academic plastic surgery residency program (N=11 residents) over 12 weeks. Six modules covering core subspecialties were developed using Google Gemini 2.5 Pro. From 60 AI-generated multiple-choice questions, 42 were selected following independent evaluation by 2 board-certified plastic surgeons using a 5-dimension rubric (accuracy, relevance, clarity, pedagogical value, and safety). Modules were released biweekly via Google Forms. Outcomes included retention, acceptability (7-point Likert scales), faculty evaluation ratings, item-level psychometrics, and pre-post changes in self-perceived knowledge and confidence. Composite scores were calculated as the mean change across all 6 subspecialty topics. Wilcoxon signed-rank tests served as the primary analysis for matched pairs (n=9). Analyses were conducted using Python 3.11.

Results: All 11 residents completed postassessments. Acceptability was high across all 6 subitems (all medians ≥5 out of 7). Faculty evaluation confirmed high content quality (mean quality index 4.68/5); interrater reliability was poor (intraclass correlation coefficient (3,1)=0.231). Among 9 matched pairs, composite confidence improved significantly (median change 0.33; d=0.65; P=.047), while composite knowledge showed a medium effect (d=0.71; P=.09). Topic-specific gains were largest where baseline was lowest: hand surgery knowledge (d=1.48; P=.02) and confidence (d=1.41; P=.02). Resource use and sense-of-community measures remained stable.

Conclusions: AI-generated microlearning modules are feasible and acceptable for plastic surgery residency education. Faculty-evaluated content achieved high quality ratings (mean 4.68/5), and residents showed strong engagement with preliminary improvements in self-perceived confidence. The intervention was integrated into existing learning ecosystems without disrupting traditional resources or peer dynamics. These findings provide effect size estimates to support the design of adequately powered randomized trials.

(JMIR Med Educ 2026;12:e88695) doi:10.2196/88695

XSL•FO
RenderX

KEYWORDS

artificial intelligence; graduate medical education; plastic surgery; microlearning; large language models; feasibility studies; multiple-choice questions

Introduction

Surgical residency training faces growing tensions between expanding knowledge requirements and compressed learning opportunities [1,2]. Duty-hour restrictions, increasing clinical demands, and the shift toward competency-based medical education have collectively reduced protected didactic time while raising expectations for subspecialty knowledge acquisition [1,3,4]. In plastic surgery, trainees must achieve proficiency across diverse domains, from microsurgical reconstruction and hand surgery to burns, aesthetic procedures, and oncologic principles, within a finite training period [5-7]. Traditional educational models relying on scheduled lectures and self-directed textbook study struggle to accommodate the fragmented schedules characteristic of modern surgical training [8].

Microlearning, defined as brief, focused learning episodes delivered in digestible segments, offers a pedagogically sound approach to these constraints [9]. The format aligns with cognitive load theory, reducing excessive cognitive load through brief, focused content delivery [9,10]. Short modules can be completed during brief intervals between cases or during call nights without requiring sustained blocks of protected time. Multiple-choice questions (MCQs) with detailed explanations represent an effective microlearning modality, combining active retrieval practice with immediate feedback [11-13].

Concurrently, large language models (LLMs) have demonstrated the ability to generate medical MCQs that approximate the quality of expert-authored items [14-16], with specialty-specific investigations confirming acceptable psychometric properties for LLM-generated MCQs in radiology [17] and urology [18]. A randomized trial in orthopedic education further reported knowledge gains when students used an LLM as a study aid [19]. However, a systematic review noted that most work has assessed content quality rather than educational deployment [20]. No study has embedded artificial intelligence (AI)–generated MCQs within a structured microlearning curriculum for surgical residency training, leaving questions of feasibility, content safety, and efficacy in this high-stakes educational context unanswered [21,22].

To address these gaps, we conducted a pilot feasibility study implementing AI-generated microlearning modules covering 6 core plastic surgery subspecialties within an academic residency program. Module topics comprised the 4 classical pillars of the field (reconstructive surgery, aesthetic surgery, hand surgery, and burn surgery) complemented by soft tissue tumor surgery [23] and an integrative module on core surgical principles. Our primary objectives were to (1) assess the feasibility and acceptability of an LLM-generated microlearning intervention among plastic surgery residents, (2) evaluate the quality of AI-generated content through systematic expert faculty evaluation, and (3) generate preliminary effect size estimates for self-perceived knowledge and confidence changes to inform future trials. Secondary objectives examined whether the intervention integrated into existing learning ecosystems without disrupting study habits or peer learning dynamics.

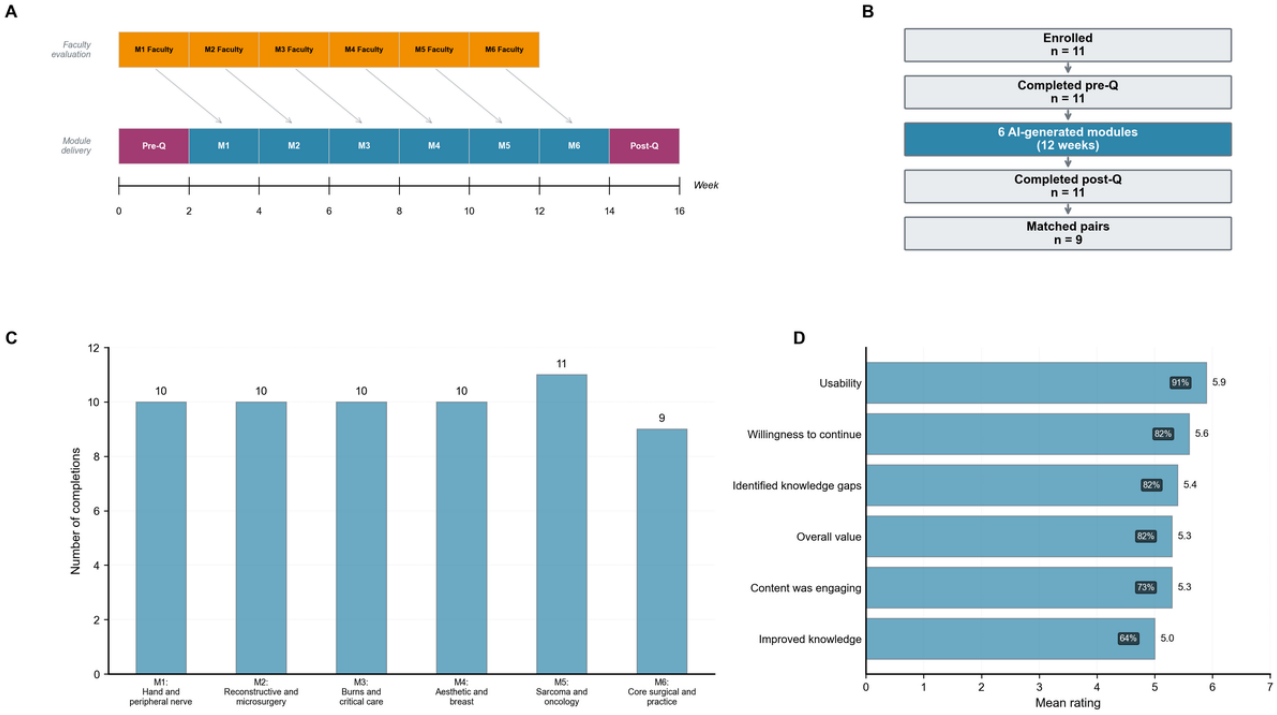
Methods

Study Design and Setting

A single-arm pilot feasibility study with a pre-post design was conducted at the Department of Plastic and Reconstructive Surgery, University Hospital BG Bergmannsheil Bochum, an academic medical center with an accredited plastic surgery residency program in Bochum, Germany. The study spanned a 12-week intervention period from July through September 2025. Reporting was guided by relevant elements of the CONSORT (Consolidated Standards of Reporting Trials) extension for pilot and feasibility studies [24]. The intervention followed a rolling production model in which faculty evaluation and selection of subsequent modules occurred concurrently with the biweekly delivery of active modules to residents, ensuring continuous content delivery throughout the study period (Figure 1A).

XSL•FO
RenderX

Figure 1. Pilot study design and feasibility outcomes. (A) Intervention timeline illustrating the rolling production model, where faculty evaluation (orange bars) occurred concurrently with the biweekly delivery of active modules (blue bars) over the 12-week period. (B) Participant flow diagram showing enrollment (n=11), postassessment completion (n=11), and the final matched cohort (n=9) used for paired analyses. (C) Module completion frequency by subspecialty topic. Bars indicate the number of module submissions per topic. (D) Postintervention acceptability ratings on a 7-point Likert scale (1=strongly disagree, 7=strongly agree). Numbers to the right of bars indicate mean ratings. Percentages inside bars indicate the proportion of residents scoring 5 or higher. AI: artificial intelligence.



Participants

All residents enrolled in the plastic surgery residency program were eligible for participation using a census sampling approach (N=11). No a priori sample size calculation was performed, consistent with the pilot feasibility design; the sample comprised the entire residency cohort to maximize enrollment and provide preliminary data across training levels.

All residents actively enrolled in the plastic surgery residency program at any postgraduate year (PGY) level were eligible; no exclusion criteria were applied.

The 11 enrolled residents were distributed across junior (PGY 1-2; n=4), mid-level (PGY 3-4; n=2), and senior (PGY 5-6; n=5) training years. Participation was voluntary, and no incentives were provided beyond access to the educational content. All 11 participants completed both pre- and postintervention assessments. Two postassessments could not be linked to their corresponding pre-assessments due to inconsistent pseudonym entry, yielding 9 matched pairs for paired analyses (Figure 1B).

AI-Generated Microlearning Modules

Module Topics

Six microlearning modules were developed covering core plastic surgery subspecialties: (1) hand and peripheral nerve surgery, (2) reconstructive principles and microsurgery, (3) burns and surgical critical care, (4) aesthetic and breast surgery, (5) sarcoma and oncologic surgery, and (6) core surgical and practice management. Topics were selected to align with the German plastic surgery curriculum (from the Accreditation Council for Graduate Medical Education, and the European Board of Plastic Reconstructive and Aesthetic Surgery; *Musterweiterbildungsordnung* [MWBO] in German) and to broadly correspond to international standards, such as those of the Accreditation Council for Graduate Medical Education and the European Board of Plastic Reconstructive and Aesthetic Surgery [25,26].

Item Generation

MCQs were generated using Google Gemini 2.5 Pro (Google; the first module used Gemini 2.5 Pro Preview, subsequent modules used the stable release). For each module, 10 initial questions were generated in separate LLM sessions using a standardized prompt (Multimedia Appendix 1). The prompt instructed the model to emulate board-style item writing as found in relevant standard textbooks (eg, *Essentials of Plastic Surgery*), target the Bloom's taxonomy "application" level or higher, adhere to evidence-based accuracy and patient safety principles, and integrate systematic knowledge such as clinical classification systems and pertinent anatomy. Each generated item was required to include a clinical vignette (2 to 4 sentences) or direct question stem, five answer options with plausible distractors, identification of the correct answer, and a structured explanation comprising: (1) rationale for the correct answer, (2) rationales for each incorrect option, (3) a systematic knowledge highlight identifying the core concept, and (4) broader educational links contextualizing the topic within plastic surgery training.

XSL•FO
RenderX

Item Selection

From the initial pool of 60 items (10 per module), 42 questions (7 per module) were selected for the pilot. Following independent evaluation (see the Faculty Quality Evaluation Protocol section), both faculty reviewers jointly reviewed the aggregated ratings and discussed each item, considering overall quality scores, content coverage across subtopics within each module, and alignment with residency training objectives. Items with lower ratings were retained when they addressed essential subtopics not covered by higher-scoring alternatives. Both reviewers reached consensus on all inclusion and exclusion decisions. The quality distributions of selected and excluded items are compared in Figure S3 in Multimedia Appendix 2.

Delivery

Modules were delivered via Google Forms, accessible on desktop and mobile devices. Residents received email and messenger notifications when new modules became available, released on a biweekly schedule (1 module every 2 weeks). Deadlines were set to the end of the study period, allowing self-paced engagement rather than enforcing a fixed completion window per module.

Faculty Quality Evaluation Protocol

All AI-generated content was independently evaluated by 2 attending plastic surgeons, each with more than 10 years of clinical experience and board certification in plastic surgery. Both faculty members were blinded to each other's assessments during independent evaluation. To ensure standardization, faculty were provided with a written grading guideline outlining the evaluation rubric, scoring anchors, and examples of acceptable versus unacceptable item features. Faculty evaluated each question and its accompanying explanation across five dimensions using standardized 5-point scales (1=poor, 5=excellent):

1. Accuracy and currency—factual correctness, alignment with current evidence-based practice guidelines, and absence of outdated information.
2. Relevance and objective alignment—clinical relevance to plastic surgery training and alignment with core competency frameworks (from the MWBO).
3. Clarity and quality of construction—question and explanation clarity, absence of ambiguity, quality of distractors, and logical organization.
4. Pedagogical value—educational effectiveness for the target learner level, appropriate difficulty calibration, and comprehensive answer justification.
5. Ethical and safety considerations—freedom from demographic or cultural biases, appropriate safety cautions, and absence of potentially harmful recommendations.

The overall quality index for each item was calculated as the mean across all 5 dimensions. The complete rubric with detailed scoring criteria is provided in Multimedia Appendix 1. Faculty reviewed all 60 items in a structured, module-by-module sequence before pilot implementation. Interrater reliability was assessed using intraclass correlation coefficients (ICC; two-way mixed-effects model, single measures, ICC[3,1]) [27].

Pre- and Postintervention Assessments

Prior to module release, all residents completed a baseline questionnaire. Following the 12-week intervention, residents completed an identical questionnaire with additional post-only measures. Table 1 lists all questionnaire items with verbatim wording, response formats, and assessment timing.

XSL•FO
RenderX

Table 1. Pre- and postintervention questionnaire items.

Item	Domain	Item wording	Response format	Timing
1	Demographics	Postgraduate year (PGY) level in plastic surgery	Single choice: PGY-1/2; PGY-3/4; PGY-5/6; Prefer not to disclose	Pre + post
2	Knowledge	"My current *knowledge* of this topic is:"[a]	7-point scale (1=very low to 7=very high); matrix with 6 subspecialty topics[b]	Pre + post
3	Confidence	"My current *confidence* in clinically applying knowledge related to this topic is:"[a]	7-point scale (1=very low to 7=very high); matrix with 6 subspecialty topics[b]	Pre + post
4	Study habits	"On average, how many hours per week do you dedicate to independent study specifically for Plastic Surgery topics?"	Hours per week (0 to >6 in 0.5-hour increments)	Pre + post
5	Study habits	"How often do you use the following resources?"[a]	5-point frequency scale (daily, weekly, monthly, rarely, never); matrix with 6 resource types[c]	Pre + post
6	Study habits	"If you use any important study resources not listed in the previous question, please specify them here."	Open text	Pre + post
7	Study habits	"I actively seek out new information or resources to deepen my understanding of complex plastic surgery topics."	7-point Likert (1=strongly disagree to 7=strongly agree)	Pre + post
8	Study habits	"How would you rate your familiarity with using AI[d] tools (eg, ChatGPT) for studying plastic surgery topics?"	7-point scale (1=very low to 7 = very high)	Pre + post
9	Study habits	"What are the primary barriers that limit your ability to study independently?"	Multiselect checklist (9 predefined options + Other)[e]	Pre + post
10	Community	"I feel that our residency program effectively addresses important learning needs that are common among Plastic Surgery trainees like myself."	7-point Likert (1=strongly disagree to 7=strongly agree)	Pre + post
11	Community	"I often discuss challenging clinical topics or new learnings with my fellow residents."	7-point Likert (1=strongly disagree to 7=strongly agree)	Pre + post
12	Community	"I feel a sense of connection with my fellow residents through our shared learning experience in plastic surgery."	7-point Likert (1=strongly disagree to 7=strongly agree)	Pre + post
13	Community	"I feel comfortable asking my fellow residents for help or admitting when I don't know something."	7-point Likert (1=strongly disagree to 7=strongly agree)	Pre + post
14	Community	"In our program, residents actively share knowledge and opportunities to help each other succeed."	7-point Likert (1=strongly disagree to 7=strongly agree)	Pre + post
15	Participation	"How many of the six modules did you complete?"	Single choice: 1 through 6	Post only
16	Participation	"How carefully did you engage with the module content?"	5-point ordinal scale[f]	Post only
17	Participation	"On average, how much time did you spend on each module?"	Categorical: <10 min; 10-20 min; 20-40 min; 40-60 min; >60 min	Post only
18	Acceptability	"This program was a valuable addition to my residency education."	7-point Likert (1=strongly disagree to 7=strongly agree)	Post only
19	Acceptability	"The AI-generated MCQs[g] helped me effectively identify my knowledge gaps."	7-point Likert (1=strongly disagree to 7=strongly agree)	Post only
20	Acceptability	"Participating in this pilot project has improved my knowledge base in the topics covered."	7-point Likert (1=strongly disagree to 7=strongly agree)	Post only
21	Acceptability	"The micro-learning format was an engaging way to study complex topics."	7-point Likert (1=strongly disagree to 7=strongly agree)	Post only
22	Acceptability	"The program's format was user-friendly and easy to navigate."	7-point Likert (1=strongly disagree to 7=strongly agree)	Post only
23	Open feedback	"What did you like most about this program?"	Open text	Post only
24	Open feedback	"What suggestions do you have for future iterations of this micro-learning format?"	Open text	Post only

XSL•FO
RenderX

Item	Domain	Item wording	Response format	Timing
25	Acceptability	"If this program were to continue, would you like to participate again?"	7-point scale (1=definitely not to 7=definitely yes)	Post only

[a]Matrix item: the stem was rated once for each row listed in the corresponding footnote.

[b]Subspecialty topics (rated as separate rows within one matrix): (1) hand and peripheral nerve surgery, (2) reconstructive and microsurgery, (3) burns and surgical critical care, (4) aesthetic and breast surgery, (5) sarcoma and oncologic surgery, (6) core surgical and practice management.

[c]Resource types (rated as separate rows within one matrix): (1) standard plastic surgery textbooks (eg, Essentials of Plastic Surgery), (2) journal articles (eg, Plastic and Reconstructive Surgery), (3) departmental resources, (4) online video platforms (eg, YouTube), (5) online databases (eg, Amboss, Microsurgeon), (6) AI-powered tools (eg, ChatGPT and Gemini).

[d]AI: artificial intelligence.

[e]Predefined barrier options: Limited time due to clinical duties; Fatigue or mental exhaustion from clinical work; Lack of a clear study plan or curriculum; Limited access to high-quality study materials; Financial constraints related to educational resources; Distracting study environment or difficulty maintaining focus; Lack of motivation or feelings of burnout; Family or personal commitments; Preference for hands-on learning.

[f]Content engagement response options: (1) I answered the questions quickly, (2) I answered the questions without deep reflection, (3) I answered the questions and reflected on the explanations, (4) I answered, reflected, and sometimes looked up related topics, (5) I answered, reflected deeply, and frequently consulted additional sources.

[g]MCQ: multiple-choice question.

Operational Definitions

Knowledge change was operationalized as the difference between postintervention and pre-intervention self-reported subspecialty knowledge scores on the 7-point Likert scale. Composite knowledge change was calculated as the mean of individual change scores across all 6 subspecialty topics. Confidence change was calculated identically for the confidence domain.

Explanation quality was operationalized as the mean faculty evaluation score across the 5 quality dimensions for the explanation component of each item, as rated independently by both evaluators. This metric was used for correlation analyses examining the relationship between content quality and learning outcomes at the module level.

Outcomes

Outcomes were organized into a hierarchical framework reflecting the pilot nature of this study.

Primary Outcomes (Feasibility)

Retention rate (proportion completing postintervention assessment), module completion rates, time spent (minutes per module and cumulative time across all modules), content engagement level, and acceptability ratings.

Secondary Outcomes (Content Quality)

Faculty evaluation ratings across the 5 quality dimensions and the overall quality index, interrater reliability (ICC), and item-level psychometric properties (item difficulty and item discrimination).

Exploratory Outcomes (Educational Efficacy)

Knowledge and confidence change scores, reported as both composite measures (mean across all 6 topics) and topic-specific measures for each subspecialty. Given the pilot sample size (n=9 matched pairs), these analyses were hypothesis-generating rather than confirmatory.

Exploratory Outcomes (Integration)

Changes in study habits (weekly study time and resource use patterns), AI tool familiarity, and sense of community dimensions. These analyses tested the preliminary hypothesis that asynchronous AI-based microlearning would integrate into existing learning structures without disrupting social learning dynamics or increasing perceived workload.

Statistical Analysis

As a pilot feasibility study designed to assess implementation feasibility and generate effect size estimates for future trials, formal a priori power calculations were not conducted. The sample included all eligible program residents (N=11).

Descriptive Statistics

Feasibility metrics were reported using proportions with 95% CIs for categorical outcomes. For ordinal outcomes (Likert-scale responses), medians (IQRs) were reported as the primary summary measure. Means (SDs) are reported where appropriate for comparison with existing literature. Faculty evaluation ratings were summarized using mean (SDs) and 95% CIs.

Primary Analysis

Pre-post changes in educational and integration outcomes were analyzed using Wilcoxon signed-rank tests for participants with complete matched data (n=9 matched pairs). Nonparametric testing was selected as the primary analysis given the ordinal nature of Likert-scale responses and the small sample size, where parametric assumptions may be violated. Normality of difference scores was assessed using Shapiro-Wilk tests.

Effect Sizes

Cohen *d* was calculated for descriptive comparison of pre-post changes, interpreted as small (*d*=0.2), medium (*d*=0.5), or large (*d*=0.8) effects [28]. CIs for Cohen *d* were derived from the t-distribution; as these are parametric approximations, minor discrepancies with Wilcoxon-based inference are expected for small samples. Rank-biserial correlations (r) were calculated as nonparametric effect size measures for Wilcoxon signed-rank tests.

XSL•FO
RenderX

Correlation Analyses

Associations between module-level outcomes (MCQ accuracy, perceived difficulty, quality, and relevance) and educational outcomes (knowledge change) were examined using Pearson correlation coefficients.

Item-Level Psychometrics

Item difficulty was calculated as the proportion of correct responses (*P* value). Item discrimination was assessed using point-biserial correlations between item scores and total test scores. Internal consistency was evaluated using Cronbach α at the module level. The

α level was set at $P<.05$ (2-tailed). Given the pilot nature and hypothesis-generating purpose, no adjustments for multiple comparisons were applied to exploratory analyses. Analyses with varying sample sizes due to item-level missing data report n per analysis. All analyses were conducted using Python (version 3.11) with *scipy*, *pandas*, *statsmodels*, *matplotlib*, and *seaborn* libraries. Complete statistical analysis code is available upon request.

Ethical Considerations

This study was reviewed by the Ethics Committee of the Medical Chamber of Westfalen-Lippe (Ärztekammer Westfalen-Lippe). The committee confirmed that formal ethical approval was not required, as the study exclusively involved anonymized data and therefore posed no risk to participant rights or safety, in accordance with the Declaration of Helsinki [29] and section 15 of the Professional Code of Conduct (reference 2025-605-f-N). All participants provided informed consent.

Results

Participant Characteristics and Study Flow

All 11 residents enrolled in the plastic surgery residency program consented to participate and completed both baseline and postintervention assessments. The cohort comprised junior (PGY 1-2; n=4), mid-level (PGY 3-4; n=2), and senior (PGY 5-6; n=5) trainees. Two postassessments could not be linked to their corresponding preassessments due to inconsistent pseudonym entry, yielding 9 matched pairs for paired analyses (Figure 1B). Module completion ranged from 9 to 11 submissions across the 6 modules, with modules 1-4 completed by 10 residents (91%), module 5 by 11 (100%), and module 6 by 9 (82%) (Figure 1C). Estimated cumulative time spent was 60 to 120 minutes over 12 weeks, equivalent to 10 to 20 minutes per module and comparable to 1-2 traditional didactic sessions.

Feasibility and Acceptability

Content engagement was moderate to high: 8 of 11 residents (73%) reported answering questions and reflecting on explanations, and 2 (18%) reported additionally consulting related topics (Table S1 in Multimedia Appendix 3). Postintervention acceptability ratings were uniformly positive, with all 6 items exceeding the scale midpoint of 4 on the 7-point scale (1 = strongly disagree, 7 = strongly agree) (Figure 1D). Usability received the highest rating (median 6, IQR 6-6; 10/11 scored 5 or higher). Willingness to continue was similarly high (median 6, IQR 5-7; 9/11 scored 5 or higher). Knowledge gap identification and overall training value each had 9 of 11 residents scoring 5 or higher. Engagement quality had a median of 6 (IQR 4.5-6.0; 8/11 scored 5 or higher). Perceived knowledge improvement received the lowest acceptability rating (median 5, IQR 4-6; 7/11 scored 5 or higher). Item-level response distributions are provided in Table S1 in Multimedia Appendix 3.

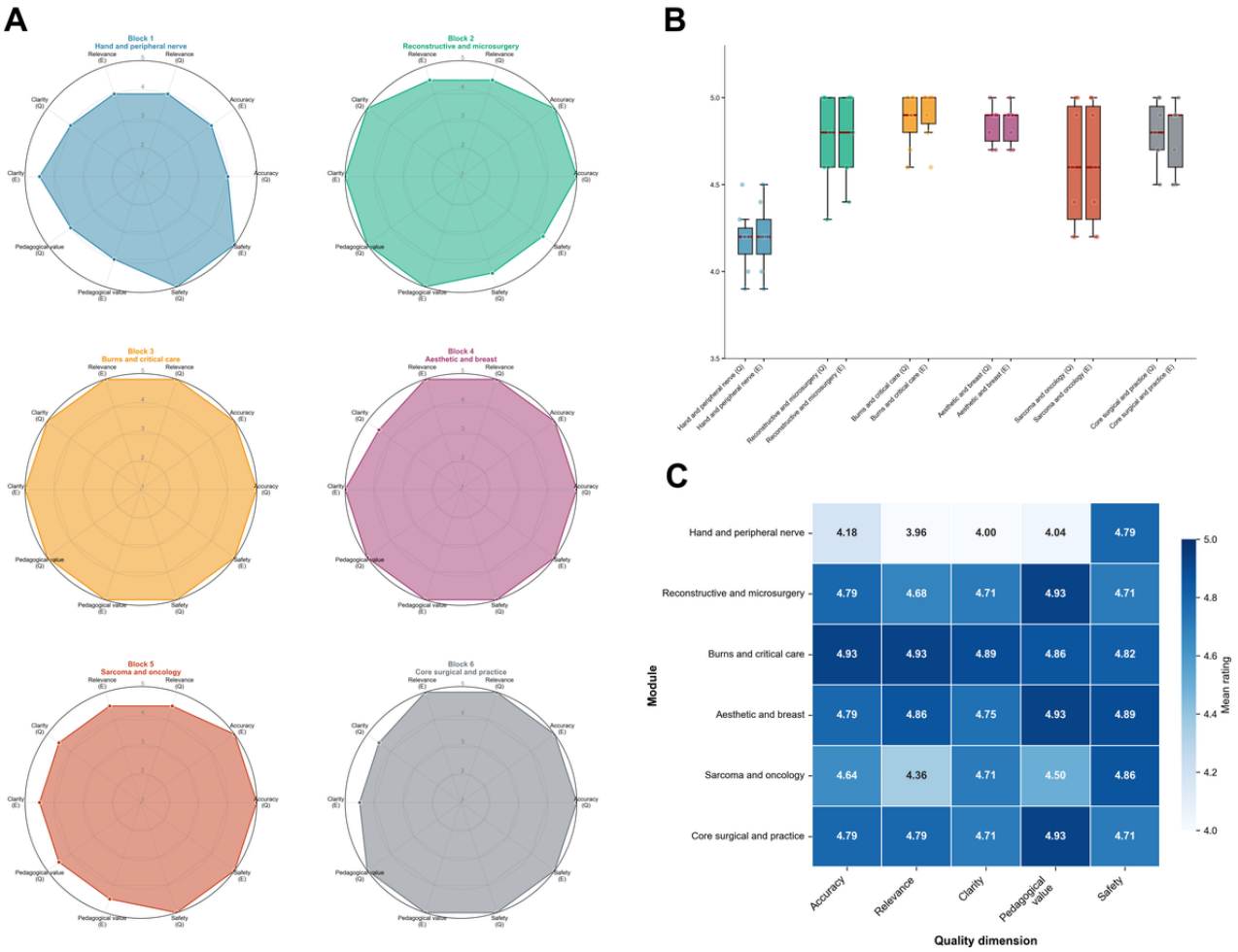
Content Quality and Expert Evaluation

Of the 60 AI-generated items evaluated by 2 board-certified faculty members, 42 were selected for the pilot (7 per module). The overall quality index for selected items was 4.68 out of 5 (SD 0.44), compared with 4.28 out of 5 for excluded items (Figure S3 in Multimedia Appendix 2). Question and explanation quality showed strong positive correlations across dimensions (pedagogical value $r=0.94$, accuracy $r=0.91$; Figure S2 in Multimedia Appendix 2), while the safety dimension showed negligible correlation with other metrics ($r=–0.05$ with relevance), indicating that clinical safety standards were maintained independent of variation in other quality dimensions. Score distributions for questions and explanations showed close alignment within each subspecialty module (Figure 2B). Among the 5 evaluation dimensions, safety received the highest mean rating (4.80), followed by pedagogical value (4.70), accuracy (4.68), clarity (4.63), and relevance (4.60) (Figure 2C). The Safety dimension remained consistently high across all modules (range 4.71-4.89), independent of variation in other dimensions. Module-level quality varied, with burns and critical care scoring highest (4.89) and hand and peripheral nerve surgery scoring lowest (4.19; notably, this was the only module generated with Gemini 2.5 Pro Preview rather than the stable release) (Figure 2A). Individual item ratings are provided in Table S2 in Multimedia Appendix 3. Faculty quality ratings for all 60 generated items and interdimensional correlations are shown in Figures S1 and S2 in Multimedia Appendix 2.

Interrater reliability for the overall quality evaluation was poor (ICC(3,1)=0.231, 95% CI 0.154-0.305; $P<.001$). Among individual dimensions, pedagogical value showed the highest agreement (ICC=0.314, 95% CI 0.143-0.466) and relevance the lowest (ICC=0.180, 95% CI 0.002-0.348). The low agreement is consistent with the subjective nature of content quality judgments between 2 independent raters and the limited precision afforded by a 2-rater design.

XSL•FO
RenderX

Figure 2. Faculty quality evaluation of artificial intelligence-generated items selected for the pilot (n = 42). (A) Radar plots displaying separate mean quality ratings for questions (Q) and explanations (E) per subspecialty module across five dimensions (Accuracy, Relevance, Clarity, Pedagogical Value, Safety). (B) Box plots showing the distribution of mean scores for Questions (Q) and Explanations (E). Center lines represent medians, box limits indicate the interquartile range, and whiskers represent the range. (C) Heatmap of mean ratings for each quality dimension across modules. Darker shading indicates higher ratings.



Item Characteristics and Module Performance

Overall person-level accuracy across the 42 selected items averaged 72.5% (SD 14.1%), falling within the recommended 65% to 75% zone for formative assessment (Figure 3A-C). Module-level accuracy ranged from 55.8% (sarcoma and oncology) to 92.1% (core surgical and practice) (Table 2).

Item-level psychometric properties varied across modules (Table 2). Mean item difficulty ranged from 0.56 (sarcoma and oncology) to 0.92 (core surgical and practice). mean item discrimination (point-biserial correlation) ranged from -0.34 (core surgical and practice) to 0.46 (burns and critical care; sarcoma and oncology). Internal consistency was low overall, with 3 of 6 modules yielding negative Cronbach α values (–0.82 to –0.20). Each module's 7 items spanned distinct subtopics rather than targeting a single construct, which likely accounts for this pattern. Burns (α=0.69) and sarcoma (α=0.74) were the only modules with positive internal consistency, and both also had the highest item discrimination.

At the module level, perceived difficulty showed a strong negative association with MCQ accuracy (r=–0.75; Figure 3A), indicating that modules rated as more challenging yielded lower scores. Perceived relevance (r=0.75; Figure 3C) and content quality (r=0.59; Figure 3B) showed positive associations with accuracy. Content quality correlated most strongly with knowledge change (r=0.79; Figure 3E), followed by perceived relevance (r=0.69; Figure 3F), while difficulty showed only a weak negative association (r=–0.31; Figure 3D). Conversely, core surgical and practice showed the highest accuracy (92.1%) but no knowledge change (d=0), suggesting a ceiling effect. Module-level perception outcomes and the complete correlation matrix are provided in Table S3 in Multimedia Appendix 3.

XSL•FO
RenderX

Figure 3. Association between resident perceptions and performance outcomes. (A-C) Scatter plots correlating mean multiple-choice question accuracy (%) with resident ratings of module difficulty (A), quality (B), and relevance (C). The shaded region represents the recommended optimal difficulty zone (65%-75%) for formative assessment. (D-F) Scatter plots correlating knowledge change (Cohen d) with resident ratings of module difficulty (D), quality (E), and relevance (F). The horizontal dotted lines indicate the thresholds for medium (d=0.5) and large (d=0.8) effect sizes. Pearson correlation coefficients (r) are displayed for each comparison. MCQ: multiple-choice questions.

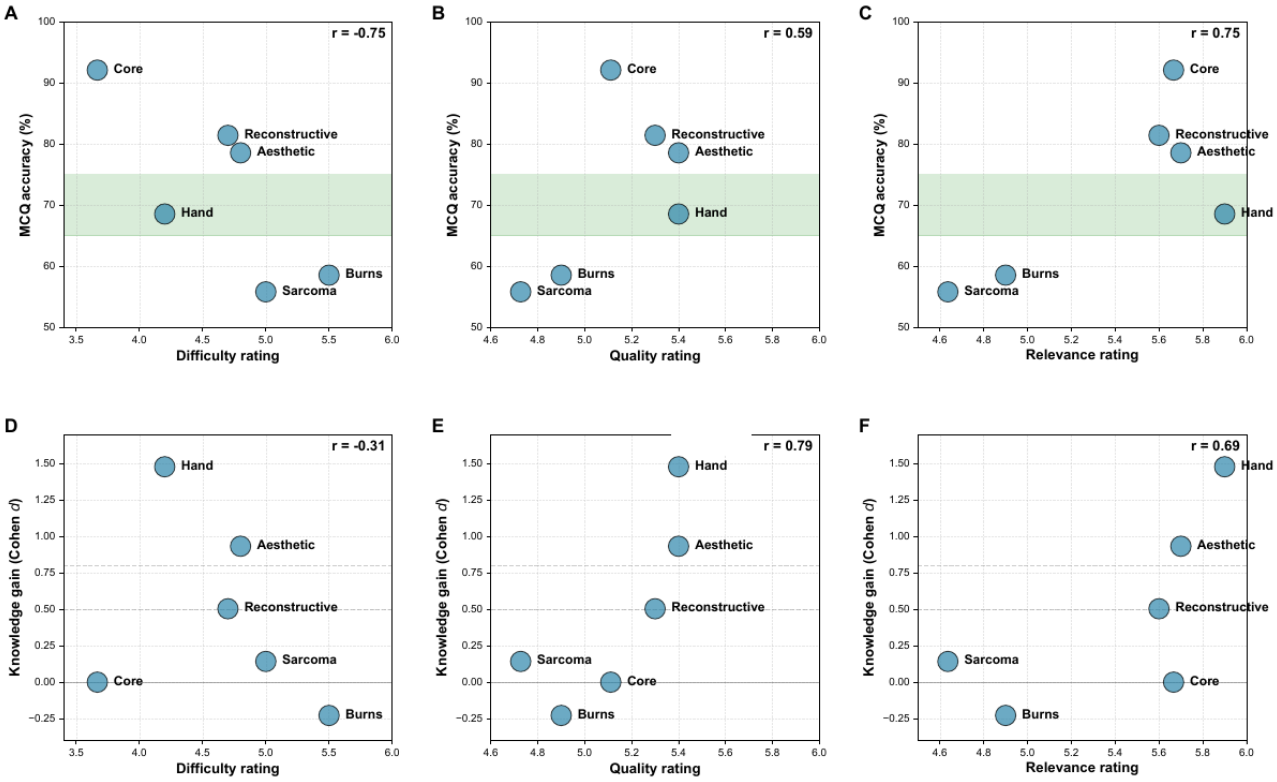


Table 2. Module performance and item-level psychometric properties.

Modules	n	MCQ[a] accuracy, % median (IQR)	Mean item difficulty (p)[b]	Mean item discrimination (r_{pb})	Cronbach α
Hand and peripheral nerve[c]	10	71.4 (60.7-71.4)	0.69	–0.12	–0.37
Reconstructive and microsurgery	10	85.7 (71.4-85.7)	0.81	–0.04	–0.20
Burns and critical care	10	50.0 (42.9-82.1)	0.59	0.46	0.69
Aesthetic and breast	10	78.6 (71.4-85.7)	0.79	0.07	0.09
Sarcoma and oncology	11	42.9 (28.6-85.7)	0.56	0.46	0.74
Core surgical and practice	9	85.7 (85.7-100)	0.92	–0.34	–0.82
Overall	10[d]	71.4 (57.1-85.7)	0.73	0.08	—[e]

[a]MCQ: multiple-choice question.

[b]Item difficulty (p) = proportion of correct responses; values approaching 0.50 indicate maximum discriminating power, values approaching 1.00 indicate easy items. Item discrimination (r_{pb}) = point-biserial correlation between item score and total module score. Cronbach α assesses internal consistency at the module level; negative values for three modules reflect the intentionally heterogeneous content within each module, where 7 items were designed to span distinct subtopics rather than measure a single construct. Individual item statistics are provided in the supplementary materials.

[c]Generated with Gemini 2.5 Pro Preview; all other modules used the stable release.

[d]Median module completion.

[e]Not applicable.

Educational Efficacy

Among the 9 matched pairs, composite self-perceived confidence improved significantly (median change 0.33; d=0.65; P=.047), while composite knowledge showed a medium effect that did not reach statistical significance (median change 0.17; d=0.71; P=.09) (Figure 4A; Table 3).

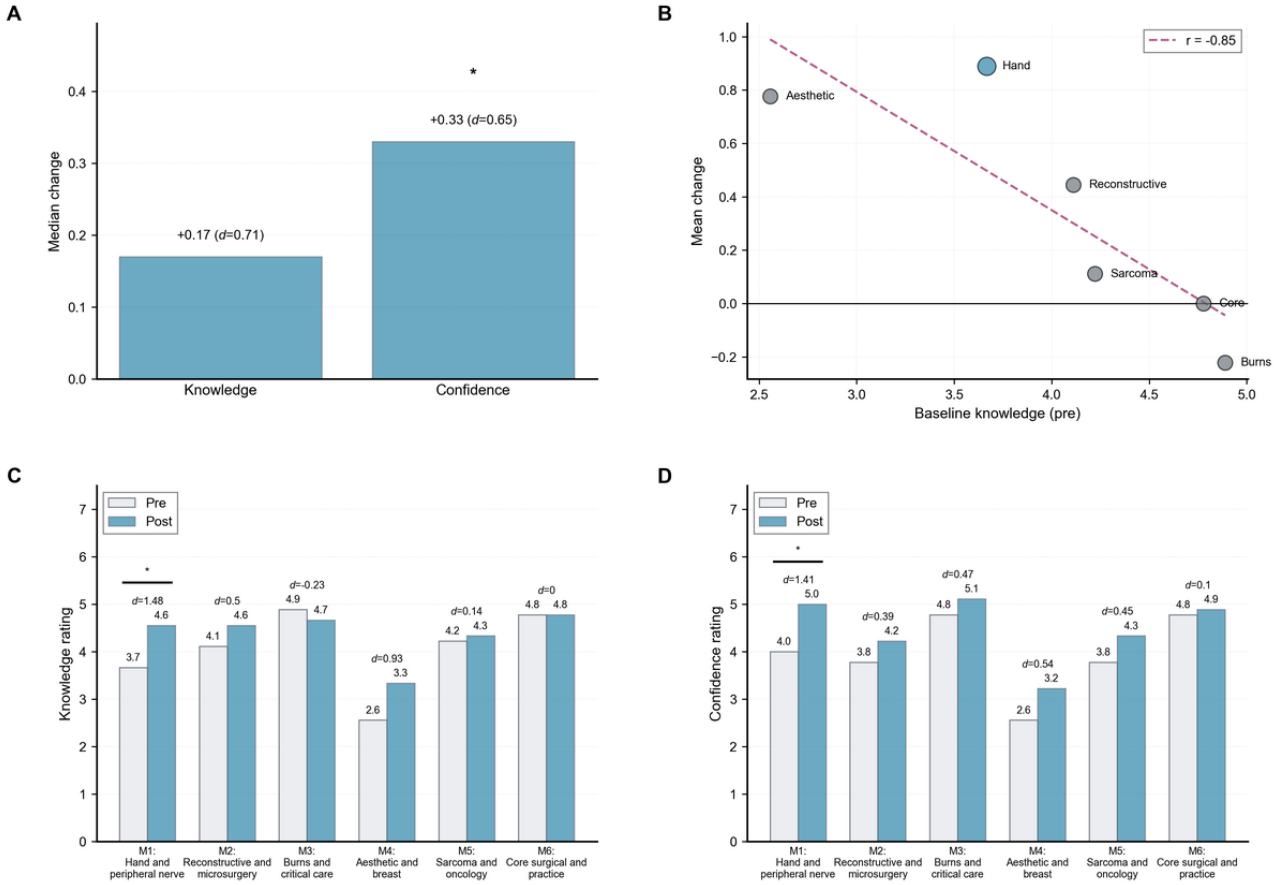
Topic-specific outcomes are summarized in Table 3. The largest effects emerged in hand and peripheral nerve surgery, with significant improvements in both knowledge (d=1.48; P=.02) and confidence (d=1.41; P=.02). Aesthetic and breast surgery

XSL•FO
RenderX

showed a large knowledge effect size (d=0.93) that did not reach significance (P=.06). Burns and critical care was the only module with a negative knowledge trajectory (d=–0.23), while core surgical and practice showed no knowledge change (d=0). A consistent inverse relationship between baseline scores and improvement was observed across modules (r=–0.85; Figure 4B), with the largest gains occurring in topics with the lowest baseline self-assessments.

A forest plot of all effect sizes is available in Figure S4 in Multimedia Appendix 2. Complete parametric statistics are provided in Table S4 in Multimedia Appendix 3 for methodological transparency.

Figure 4. Educational efficacy outcomes. (A) Composite change in self-reported knowledge and confidence scores (n = 9 matched pairs), displaying the median change and effect sizes (Cohen d). *P<.05 by Wilcoxon signed-rank test. (B) Relationship between baseline knowledge and knowledge change across modules (r=–0.85), indicating that topics with lower starting scores showed greater improvement. (C) Pre- and postintervention self-perceived knowledge by subspecialty module. (D) Pre- and postintervention self-perceived confidence by subspecialty module. Cohen d values indicate the effect size for each comparison; *P<.05 by Wilcoxon signed-rank test.



XSL•FO
RenderX

Table 3. Pre-post changes in self-perceived knowledge and confidence (n=9 matched pairs).

Outcomes	Pre median (IQR)	Post median (IQR)	Median diff	*W*	*P* value	*r*	*d* (95% CI)
Composite measures[a]							
Knowledge	4 (3.33-4.33)	4.17 (4.00-4.83)	0.17	7.5	.09	0.67	0.71 (–0.06 to 1.48)
Confidence	4.17 (3-4.33)	4.67 (4.17-5.17)	0.33	5.5	.047	0.76	0.65 (–0.11 to 1.42)
Topic-specific: knowledge							
Hand and peripheral nerve	4 (3-4)	5 (4-5)	1	1.5	.02	0.93	1.48 (0.71-2.25)
Reconstructive and microsurgery	4 (3-5)	5 (4-5)	0	11.5	.31	0.49	0.50 (–0.26 to 1.27)
Burns and critical care	4 (4-6)	4 (4-6)	0	18	.75	0.20	–0.23 (–1 to 0.54)
Aesthetic and breast	2 (2-3)	4 (3-4)	1	5	.06	0.78	0.93 (0.16-1.70)
Sarcoma and oncology	4 (4-4)	4 (4-5)	0	19	≥.99	0.16	0.14 (–0.63 to 0.91)
Core surgical and practice	5 (4-6)	5 (4-5)	0	20	.94	0.11	0 (–0.77 to 0.77)
Topic-specific: confidence							
Hand and peripheral nerve	4 (4-5)	5 (4-6)	1	1.5	.02	0.93	1.41 (0.65-2.18)
Reconstructive and microsurgery	4 (2-5)	5 (3-5)	0	14.5	.50	0.36	0.39 (–0.38 to 1.16)
Burns and critical care	5 (4-5)	5 (4-6)	0	12	.38	0.47	0.47 (–0.30 to 1.24)
Aesthetic and breast	2 (2-3)	3 (3-4)	0	11	.25	0.51	0.54 (–0.22 to 1.31)
Sarcoma and oncology	4 (3-4)	5 (3-5)	1	12.5	.36	0.44	0.45 (–0.32 to 1.22)
Core surgical and practice	5 (4-6)	5 (5-5)	0	19	≥.99	0.16	0.11 (–0.66 to 0.87)

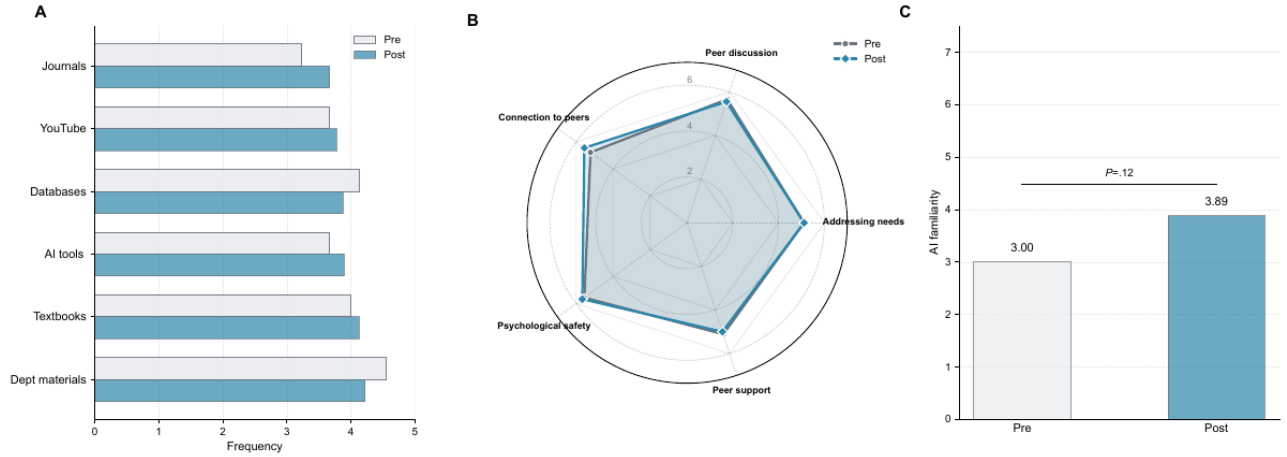
[a]Bold rows indicate statistical significance (*P*<.05). Medians and IQR are reported on the original 7-point Likert scale (1=strongly disagree/not at all confident to 7=strongly agree/very confident). Median diff=median of individual paired differences. *W*=Wilcoxon signed-rank test statistic (smaller values indicate greater positive shift); *r*=rank-biserial correlation (nonparametric effect size); *d* = Cohen *d* with 95% CI. Complete parametric statistics are provided in Table S4 in Multimedia Appendix 3.

Learning Integration

No significant changes in educational resource use were observed over the 12-week intervention period (textbooks *P*≥.99, journal articles *P*=.25, departmental materials *P*=.25, YouTube videos *P*≥.99, databases *P*=.62, and AI tools *P*=.62; Figure 5A; Table S5 in Multimedia Appendix 3). Weekly study time remained stable (pre median 2 hours, IQR 1-3; post median 2 hours, IQR 1.5-2; *P*=.45). Traditional resource use, including textbooks (*P*≥.99), journal articles (*P*=.25), and departmental materials (*P*=.25), did not change. This pattern suggests the intervention operated as an additive supplement rather than displacing established learning resources. All 5 sense-of-community dimensions remained stable (program addressing shared needs *P*≥.99, peer discussion *P*≥.99, connection to peers *P*=.50, psychological safety *P*≥.99, and peer support *P*≥.99; Figure 5B), with median ratings at or above 5 on the 7-point scale throughout. This stability provides preliminary evidence that asynchronous digital learning did not disrupt the apprenticeship model central to surgical training. AI tool familiarity showed the largest contextual change (pre median 3, IQR 2-5; post median 4, IQR 4-5; *d*=0.70; *P*=.12), though this did not reach statistical significance (Figure 5C).

XSL•FO
RenderX

Figure 5. Learning context and integration outcomes. (A) Frequency of educational resource use at baseline (pre) and postintervention (post). No significant changes were observed (textbooks *P*≥.99, journal articles *P*=.25, departmental materials *P*=.25, youtube videos *P*≥.99, databases *P*=.62, and artificial intelligence tools *P*=.62). (B) Radar plot of sense-of-community dimensions, showing stable ratings across all five measures (program addressing shared needs P≥.99, peer discussion P≥.99, connection to peers *P*=.50, psychological safety *P*≥.99, and peer support *P*≥.99). (C) Self-reported artificial intelligence familiarity scores pre- and postintervention (d=0.70; *P*=.12). AI: artificial intelligence.



Discussion

Principal Findings

This pilot study provides preliminary evidence that AI-generated microlearning modules can be feasibly integrated into surgical residency education. Over 12 weeks, all 11 enrolled residents completed postintervention assessments, and module submissions ranged from 9 to 11 across the 6 subspecialty topics. Cumulative time spent of 60 to 120 minutes was distributed across brief, self-paced sessions (10-20 minutes per module), suggesting compatibility with clinical schedules. Faculty evaluation confirmed high content quality across all modules, and postintervention acceptability ratings were uniformly positive. Among 9 matched pairs, composite self-perceived confidence improved significantly (*d*=0.65; *P*=.047), while composite knowledge showed a medium effect size (*d*=0.71) that did not reach statistical significance (*P*=.09). These preliminary findings support the feasibility of this approach and provide effect size estimates to inform the design of future confirmatory trials.

Positioning Within the Literature

Our findings extend the emerging literature on LLM-generated medical education content. Prior studies have established that AI systems can generate MCQs achieving quality ratings comparable to human-authored items when evaluated by content experts [15,16,22]. However, most investigations have examined question quality as an endpoint rather than deploying AI-generated content within authentic educational interventions [21,22]. Our study moves beyond quality benchmarking by embedding curated AI content into a 12-week residency curriculum and measuring educational outcomes, albeit in an exploratory pilot design with a small sample.

When positioned within the broader literature, our composite effect sizes (*d*=0.65-0.71) align with the upper range of effects reported in continuing medical education interventions, where effect sizes corresponding to Cohen *d* of approximately 0.4 to 0.7 are typical [30]. These are notably smaller than effects reported in undergraduate medical education meta-analyses (SMDs of 1.2-3.4 for case-based and simulation-based learning) [31,32], likely reflecting the greater baseline knowledge and clinical experience of postgraduate learners. Against this benchmark, our preliminary findings fall within the upper range of reported continuing medical education effect sizes, despite the minimal time burden, asynchronous format, and reliance on AI-generated content.

Predictors of Educational Effectiveness

Module-level analyses offered exploratory insight into content characteristics associated with educational outcomes. Perceived content quality (*r*=0.79) and relevance (*r*=0.69) showed strong positive associations with knowledge change, whereas perceived difficulty showed only a weak negative association (*r*=–0.31). These patterns are consistent with prior work demonstrating that perceived clinical relevance fosters knowledge retention [33]. If confirmed in larger samples, these associations carry practical implications: faculty reviewers and developers of AI-generated educational content might prioritize clinical relevance and pedagogical quality as primary selection criteria.

A consistent inverse relationship between baseline self-assessed knowledge and improvement emerged across modules (*r*=–0.85), with the largest gains in hand and peripheral nerve surgery (*d*=1.48; *P*=.02) and aesthetic and breast surgery (*d*=0.93; *P*=.06), where baseline scores were lowest. Conversely, modules with higher baseline scores showed negligible or no change (core surgical *d*=0; burns *d*=–0.23). This floor-ceiling pattern is consistent with targeted learning theory and suggests that AI-generated microlearning may be particularly suited for addressing identified knowledge gaps rather than reinforcing already-strong domains.

Content Quality and Expert Evaluation

The structured faculty evaluation process resulted in the selection of 42 items from an initial pool of 60, elevating the mean quality index from 4.28 to 4.68 on a 5-point scale. The safety dimension remained consistently high across all modules (range 4.71-4.89) and showed negligible correlation with other

XSL•FO
RenderX

quality dimensions, suggesting that LLM outputs can maintain appropriate clinical safety standards even when other aspects of content quality vary. This structural independence is noteworthy given concerns about AI-generated medical content potentially propagating harmful recommendations. Despite this independence, the expert curation process proved essential for overall content quality, suggesting that human expert oversight remains important for deploying AI-generated content in medical education contexts where content errors carry safety implications.

However, interrater reliability between the 2 faculty evaluators was poor (ICC(3,1)=0.231, 95% CI 0.154-0.305). This finding highlights the inherent subjectivity of content quality judgments and the limited precision afforded by a 2-rater design. Future implementations should consider larger evaluation panels and consensus-based approaches to improve reliability. Despite the low ICC, both raters independently selected similar items for inclusion, suggesting agreement on which items were clearly strong or clearly weak, even when absolute ratings diverged.

Internal consistency was low, with negative Cronbach α values in 3 of 6 modules, reflecting the heterogeneous content within each module where items spanned distinct subtopics. The 2 modules with positive internal consistency (burns, α=0.69; sarcoma, α=0.74) also showed the highest item discrimination, suggesting that future item development might benefit from greater subtopic coherence within modules.

Learning Integration

The stability of residents' learning ecosystems throughout the intervention provides preliminary evidence that asynchronous AI-based microlearning can supplement existing educational resources without displacement. Weekly study time, traditional resource use, and all 5 sense-of-community dimensions remained unchanged, suggesting the modules functioned as an additive intervention. This nondisruption is relevant given concerns that digital learning tools might undermine the apprenticeship model central to surgical training.

AI tool familiarity showed the largest contextual change (d=0.70), though this did not reach statistical significance (P=.12). Exposure to faculty-evaluated AI content may function as experiential onboarding, potentially preparing residents for the AI-augmented clinical environments they will increasingly encounter. Whether curated AI educational content can serve as an entry point for broader AI literacy warrants investigation in future studies.

Strengths and Limitations

This study has several methodological strengths. The rolling production model enabled continuous content delivery while maintaining rigorous faculty evaluation, demonstrating a workflow for AI-assisted curriculum development. The standardized prompting strategy and comprehensive evaluation rubric provide reproducible frameworks for other programs seeking to implement similar interventions. The multidimensional assessment captured not only efficacy outcomes but also feasibility, acceptability, and ecosystem integration metrics essential for translating pilot findings into sustainable educational practice.

Several limitations warrant consideration when interpreting these findings. First, the single-arm design without a control group precludes causal attribution; secular trends, Hawthorne effects, or concurrent clinical experiences may have contributed to observed changes. Second, the small sample size (n=9 matched pairs) limited statistical power, as reflected in wide confidence intervals, and some outcomes that showed medium-to-large effect sizes did not reach statistical significance. Third, self-reported knowledge and confidence outcomes may not reflect objective performance on standardized examinations or clinical competency assessments. Fourth, interrater reliability for the faculty evaluation was poor (ICC=0.231), limiting the precision of content quality judgments. Fifth, the single-institution setting within a German academic plastic surgery program, the 12-week follow-up period, and the reliance on a single LLM (Google Gemini 2.5 Pro) with a specific prompting strategy all constrain generalizability. However, the evaluation framework and curriculum structure remain applicable regardless of which AI system generates initial content.

Conclusions

This pilot study provides evidence that AI-generated microlearning modules are feasible and acceptable for plastic surgery residency education. Faculty-evaluated content achieved high-quality ratings, and residents showed strong engagement with meaningful preliminary improvements in self-perceived confidence. The intervention was integrated into existing learning ecosystems without disrupting traditional resources or peer learning dynamics. The methodology described here, combining structured prompt engineering with subspecialty-specific content generation, systematic faculty evaluation, and learner-centered outcome assessment, may provide a framework for programs exploring similar AI-assisted educational interventions, though replication across institutions and specialties is needed. These findings, while preliminary, support proceeding to adequately powered randomized trials evaluating AI-generated microlearning against standard educational approaches, with attention to objective outcome measures, longer follow-up, and multi-institutional settings.

Acknowledgments

Use of generative artificial intelligence: all educational content (60 multiple-choice questions with explanations) used as the study intervention was generated using Gemini 2.5 Pro (Google), as described in the AI-Generated Microlearning Modules section. Claude Code (Anthropic) was used to assist with data analysis automation (eg, template code generation and figure scaffolding), statistical programming, and copyediting. All outputs were reviewed, edited, and verified by the authors, who retain full responsibility for the analyses and their interpretation.

XSL•FO
RenderX

Funding

This study received no external funding.

Data Availability

All data analysis code as well as deidentified participant data are available upon reasonable request from the corresponding author. Data access may be subject to institutional review board approval.

Authors' Contributions

Conceptualization: MD, SVS, CW
Methodology: MD, SVS, AS, CW
Software: MD, FR, FP
Validation: AS, CW
Formal analysis: MD, SVS
Investigation: SVS, AF, MF
Data curation: SVS, FP
Writing – original draft: MD
Writing – review and editing: SVS, CW, FR, ML, AF, MF, AS
Visualization: MD, FR
Supervision: ML, AS, CW

Conflicts of Interest

None declared.

Multimedia Appendix 1
Artificial intelligence content generation prompt, faculty evaluation rubric with scoring criteria, and example multiple-choice questions with explanations.
[DOCX File , 53 KB - mededu_v12i1e88695_app1.docx]

Multimedia Appendix 2
Additional figures supporting the study findings.
[DOCX File , 1179 KB - mededu_v12i1e88695_app2.docx]

Multimedia Appendix 3
Additional tables supporting the study findings.
[DOCX File , 74 KB - mededu_v12i1e88695_app3.docx]

References

1. Ahmed N, Devitt KS, Keshet I, Spicer J, Imrie K, Feldman L, et al. A systematic review of the effects of resident duty hour restrictions in surgery: impact on resident wellness, training, and patient outcomes. Ann Surg 2014;259(6):1041-1053 [FREE Full text] [doi: 10.1097/SLA.0000000000000595] [Medline: 24662409]
2. Awan M, Zagales I, McKenney M, Kinslow K, Elkbuli A. ACGME 2011 duty hours restrictions and their effects on surgical residency training and patients outcomes: a systematic review. J Surg Educ 2021;78(6):e35-e46. [doi: 10.1016/j.jsurg.2021.06.001] [Medline: 34183278]
3. Desai SV, Asch DA, Bellini LM, Chaiyachati KH, Liu M, Sternberg AL, et al. Education outcomes in a duty-hour flexibility trial in internal medicine. N Engl J Med 2018;378(16):1494-1508 [FREE Full text] [doi: 10.1056/NEJMoa1800965] [Medline: 29557719]
4. Moeller A, Webber J, Epstein I. Resident duty hour modification affects perceptions in medical education, general wellness, and ability to provide patient care. BMC Med Educ 2016;16:175 [FREE Full text] [doi: 10.1186/s12909-016-0703-4] [Medline: 27411835]
5. Knox ADC, Gilardino MS, Kasten SJ, Warren RJ, Anastakis DJ. Competency-based medical education for plastic surgery: where do we begin? Plast Reconstr Surg 2014;133(5):702e-710e. [doi: 10.1097/PRS.0000000000000082] [Medline: 24776573]
6. Kania K, Chang DK, Abu-Ghname A, Reece EM, Chu CK, Maricevich M, et al. Microsurgery training in plastic surgery. Plast Reconstr Surg Glob Open 2020;8(7):e2898 [FREE Full text] [doi: 10.1097/GOX.0000000000002898] [Medline: 32802641]

XSL•FO
RenderX

7. Morris MP, Toyoda Y, Christopher AN, Broach RB, Percec I. A systematic review of aesthetic surgery training within plastic surgery training programs in the USA: an in-depth analysis and practical reference. Aesthetic Plast Surg 2022;46(1):513-523. [doi: 10.1007/s00266-021-02557-0] [Medline: 34467421]
8. Sawatsky AP, Ratelle JT, Bonnes SL, Egginton JS, Beckman TJ. A model of self-directed learning in internal medicine residency: a qualitative study using grounded theory. BMC Med Educ 2017;17(1):31 [FREE Full text] [doi: 10.1186/s12909-017-0869-4] [Medline: 28148247]
9. De Gagne JC, Park HK, Hall K, Woodward A, Yamane S, Kim SS. Microlearning in health professions education: scoping review. JMIR Med Educ 2019;5(2):e13997 [FREE Full text] [doi: 10.2196/13997] [Medline: 31339105]
10. Lopez S. The impact of cognitive load theory on the effectiveness of microlearning modules. EJEDU 2024;5(2):29-35. [doi: 10.24018/ejedu.2024.5.2.799]
11. Lee YM, Jahnke I, Austin L. Mobile microlearning design and effects on learning efficacy and learner experience. Education Tech Research Dev 2021;69(2):885-915. [doi: 10.1007/s11423-020-09931-w]
12. Riggs CD, Kang S, Rennie O. Positive impact of multiple-choice question authoring and regular quiz participation on student learning. CBE Life Sci Educ 2020;19(2):ar16 [FREE Full text] [doi: 10.1187/cbe.19-09-0189] [Medline: 32357094]
13. Kirubarajan A, Young D, Khan S, Crasto N, Sobel M, Sussman D. Artificial intelligence and surgical education: a systematic scoping review of interventions. J Surg Educ 2022;79(2):500-515. [doi: 10.1016/j.jsurg.2021.09.012] [Medline: 34756807]
14. Abouzeid E, Wassef R, Jawwad A, Harris P. Chatbots' role in generating single best answer questions for undergraduate medical student assessment: comparative analysis. JMIR Med Educ 2025;11:e69521 [FREE Full text] [doi: 10.2196/69521] [Medline: 40446330]
15. Laupichler MC, Rother JF, Grunwald Kadow IC, Ahmadi S, Raupach T. Large language models in medical education: comparing ChatGPT- to human-generated exam questions. Acad Med 2024;99(5):508-512 [FREE Full text] [doi: 10.1097/ACM.0000000000005626] [Medline: 38166323]
16. Law AK, So J, Lui CT, Choi YF, Cheung KH, Kei-Ching Hung K, et al. AI versus human-generated multiple-choice questions for medical education: a cohort study in a high-stakes examination. BMC Med Educ 2025;25(1):208 [FREE Full text] [doi: 10.1186/s12909-025-06796-6] [Medline: 39923067]
17. Mistry NP, Saeed H, Rafique S, Le T, Obaid H, Adams SJ. Large language models as tools to generate radiology board-style multiple-choice questions. Acad Radiol 2024;31(9):3872-3878 [FREE Full text] [doi: 10.1016/j.acra.2024.06.046] [Medline: 39013736]
18. Kim JK, Chua ME, Lorenzo AJ, Rickard M, Andreacchi L, Kim M, et al. Use of artificial intelligence-generated multiple-choice questions for the examination of surgical subspecialty residents report of feasibility and psychometric analysis. Can Urol Assoc J 2025;19(6):182-187 [FREE Full text] [doi: 10.5489/cuaj.9020] [Medline: 40031953]
19. Gan W, Ouyang J, Li H, Xue Z, Zhang Y, Dong Q, et al. Integrating ChatGPT in orthopedic education for medical undergraduates: randomized controlled trial. J Med Internet Res 2024;26:e57037 [FREE Full text] [doi: 10.2196/57037] [Medline: 39163598]
20. Artsi Y, Sorin V, Konen E, Glicksberg BS, Nadkarni G, Klang E. Large language models for generating medical examinations: systematic review. BMC Med Educ 2024;24(1):354 [FREE Full text] [doi: 10.1186/s12909-024-05239-y] [Medline: 38553693]
21. Feigerlova E, Hani H, Hothersall-Davies E. A systematic review of the impact of artificial intelligence on educational outcomes in health professions education. BMC Med Educ 2025;25(1):129 [FREE Full text] [doi: 10.1186/s12909-025-06719-5] [Medline: 39871336]
22. Gordon M, Daniel M, Ajiboye A, Uraiby H, Xu NY, Bartlett R, et al. A scoping review of artificial intelligence in medical education: BEME guide no. 84. Med Teach 2024;46(4):446-470 [FREE Full text] [doi: 10.1080/0142159X.2024.2314198] [Medline: 38423127]
23. Götzl R, Sterzinger S, Arkudas A, Boos AM, Semrau S, Vassos N, et al. The role of plastic reconstructive surgery in surgical therapy of soft tissue sarcomas. Cancers (Basel) 2020;12(12):3534 [FREE Full text] [doi: 10.3390/cancers12123534] [Medline: 33256182]
24. Eldridge SM, Chan CL, Campbell MJ, Bond CM, Hopewell S, Thabane L, et al. CONSORT 2010 statement: extension to randomised pilot and feasibility trials. BMJ 2016;355:i5239 [FREE Full text] [doi: 10.1136/bmj.i5239] [Medline: 27777223]
25. Accreditation Council for Graduate Medical Education. ACGME program requirements for graduate medical education in plastic surgery. ACGME. 2022. URL: https://www.acgme.org/specialties/plastic-surgery/program-requirements-and-faqs-and-applications/ [accessed 2026-04-10]
26. Goodacre TEA, Olariu R. European training requirements for the specialty of plastic, reconstructive and aesthetic surgery: European standards of postgraduate medical specialist training. European Union of Medical Specialists (UEMS), Section and Board of Plastic, Reconstructive and Aesthetic Surgery. Brussels; 2016. URL: https://ebopras.eu/media/1/UEMS-2024.31-Revised-ETR-Plastic-Reconstructive-and-Aesthetic-Surgery.pdf# [accessed 2026-04-10]
27. Koo TK, Li MY. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. J Chiropr Med 2016;15(2):155-163 [FREE Full text] [doi: 10.1016/j.jcm.2016.02.012] [Medline: 27330520]
28. Cohen J. Statistical Power Analysis for the Behavioral Sciences. Hillsdale, NJ: Lawrence Erlbaum Associates; 1988.

XSL•FO
RenderX

29. WMA declaration of Helsinki - ethical principles for medical research involving human participants. World Medical Association. URL: https://www.wma.net/policies-post/wma-declaration-of-helsinki/ [accessed 2025-11-25]
30. Mansouri M, Lockyer J. A meta-analysis of continuing medical education effectiveness. J Contin Educ Health Prof 2007;27(1):6-15. [doi: 10.1002/chp.88] [Medline: 17385735]
31. Zhang SL, Ren SJ, Zhu DM, Liu TY, Wang L, Zhao JH, et al. Which novel teaching strategy is most recommended in medical education? A systematic review and network meta-analysis. BMC Med Educ 2024;24(1):1342 [FREE Full text] [doi: 10.1186/s12909-024-06291-4] [Medline: 39574112]
32. Maia D, Andrade R, Afonso J, Costa P, Valente C, Espregueira-Mendes J. Academic performance and perceptions of undergraduate medical students in case-based learning compared to other teaching strategies: a systematic review with meta-analysis. Education Sciences 2023;13(3):238. [doi: 10.3390/educsci13030238]
33. Malau-Aduli BS, Lee AY, Cooling N, Catchpole M, Jose M, Turner R. Retention of knowledge and perceived relevance of basic sciences in an integrated case-based learning (CBL) curriculum. BMC Med Educ 2013;13:139 [FREE Full text] [doi: 10.1186/1472-6920-13-139] [Medline: 24099045]

Abbreviations

AI: artificial intelligence
CONSORT: Consolidated Standards of Reporting Trials
ICC: intraclass correlation coefficient
LLM: large language model
MCQ: multiple-choice question
MWBO: Musterweiterbildungsordnung
PGY: postgraduate year

Edited by J Gentges; submitted 30.Nov.2025; peer-reviewed by M Kalem, S Mallick, D Chen, WMM Ahmed; comments to author 13.Apr.2026; accepted 20.May.2026; published 14.Jul.2026.

Please cite as:
Drysch M, Schmidt SV, Reinkemeier F, Puscz F, Fiedler A, Fueth M, Lehnhardt M, Sogorski A, Wallner C
AI-Generated Microlearning for Plastic Surgery Residency: Single-Arm Pre-Post Feasibility Study
JMIR Med Educ 2026;12:e88695
URL: https://mededu.jmir.org/2026/1/e88695
doi:10.2196/88695
PMID:

©Marius Drysch, Sonja Verena Schmidt, Felix Reinkemeier, Flemming Puscz, Alexander Fiedler, Maria Fueth, Marcus Lehnhardt, Alexander Sogorski, Christoph Wallner. Originally published in JMIR Medical Education (https://mededu.jmir.org), 14.Jul.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Education, is properly cited. The complete bibliographic information, a link to the original publication on https://mededu.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

AI-Simulated Patients for Training Shared Decision-Making: Feasibility Study in Medical Education

Lea Herschbach[1], MSc; Teresa Festl-Wietek[1], Dr rer nat; Alessandra Sonanini[1], MD; Friederike Holderried[1], MME, MD; Tobias Albrecht[1,2], MME, MD; Bernd Herrmann[1,3], Dipl; Anne Herrmann-Werner[1,3], MME, Prof Dr Med

[1]TIME - Tuebingen Institute for Medical Education, University of Tübingen, Tuebingen, Baden-Wurttemberg, Germany

[2]Department of Otolaryngology, Head and Neck Surgery, University Hospital Tübingen, Tuebingen, Baden-Wurttemberg, Germany

[3]Deutsches Zentrum für Psychische Gesundheit, Partner site Tübingen, Tuebingen, Baden-Wurttemberg, Germany

Corresponding Author:
Lea Herschbach, MSc
TIME - Tuebingen Institute for Medical Education
University of Tübingen
Elfriede-Aulhorn-Str. 10
Tuebingen, Baden-Wurttemberg, 72076
Germany
Phone: 49 70712973715
Email: lea.herschbach@med.uni-tuebingen.de

Abstract

Background: Shared decision-making (SDM) is a key element of patient-centered care; however, opportunities for structured and scalable SDM training remain limited in both medical education and clinical practice. Advances in AI have enabled chatbot-based simulations that may support repeated practice and provide automated feedback on SDM-related communication behaviors.

Objective: This study aimed to (1) evaluate the agreement between AI-generated and human ratings of SDM performance, (2) compare SDM-related performance and attitudes between medical students and physicians, and (3) explore changes in communicative self-efficacy (CSE) following chatbot-based SDM practice.

Methods: We conducted a feasibility study using a pre-post mixed methods design. Medical students and physicians practiced SDM in a 20-minute chat-based interaction with an AI-simulated patient, followed by automated feedback on SDM performance. Prior to the interaction, participants' attitudes toward SDM (IcanSDM) and their CSE (Self-Efficacy in Patient-Centeredness Questionnaire [SEPCQ-24-GER]) were assessed; CSE was reevaluated after the intervention. SDM performance was measured using the validated Observing Patient Involvement in Decision Making (OPTION-12) scale, with the AI evaluating participants' performance from the patient's perspective. These AI-generated ratings were then compared with human ratings to evaluate agreement and rating quality. To further investigate discrepancies between AI and human ratings, a qualitative analysis of selected cases was conducted. Additional measures included demographics, perceived authenticity, and benefits of the interaction. Quantitative analyses included group comparisons, pre-post analyses, and psychometric evaluation of the OPTION-12 scale in an AI-mediated setting. The study was preregistered on the Open Science Framework.

Results: A total of 58 participants were included in the analysis, 22 (37.9%) of whom were students, 34 (58.2%) were physicians, and 2 did not declare. Both groups reported similarly positive attitudes toward SDM, but medical students (n=22; mean 28.67, SD 4.32) achieved higher OPTION-12 performance scores than physicians (n=34; mean 22.79, SD 8.14). CSE showed a slight decrease following the intervention. Agreement between AI-generated and human OPTION-12 total scores was high (intraclass correlation coefficient=0.86, df=2,1; P<.001), although AI ratings were systematically higher. The OPTION-12 scale in the AI setting demonstrated good internal consistency (Cronbach α=0.88; mean interitem correlation r=0.43). Participants perceived the AI-simulated interaction as authentic and potentially useful for SDM training.

Conclusions: This feasibility study provides preliminary and exploratory insights into the use of AI-simulated patients for SDM training in medical education. The findings contribute to the contextual validation of the OPTION-12 scale in AI-mediated consultations and provide valuable insights into participants' attitudes toward SDM. However, while findings suggest potential for supporting communication skills development and structured feedback, conclusions are limited by the study design and sample size. Further research with larger samples and controlled designs is needed to establish effectiveness and generalizability.

XSL•FO
RenderX

(JMIR Med Educ 2026;12:e100467) doi:10.2196/100467

KEYWORDS

artificial intelligence; chatbots; communication; medical education; patient participation; professional competence; SDM; self-efficacy; shared decision-making; simulation training

Introduction

Background and Relevance of Shared Decision-Making

Shared decision-making (SDM) is a collaborative, patient-centered approach in which health care professionals and patients jointly make medical decisions based on evidence and patients' preferences [1]. While SDM is often regarded as a modern model of care, its conceptual roots can be traced back to early discussions in the 1950s emphasizing patient involvement as a core element of the "modern doctor" [2]. Charles et al [3] later formalized SDM into a significant framework comprising 4 key elements: the involvement of at least 2 people, reciprocal information exchange, joint deliberation regarding treatment options, and agreement on a specific decision. SDM is widely recognized as a strategy to reduce informational and power asymmetries in clinical encounters by strengthening patient autonomy and encouraging active participation [4,5]. Empirical evidence indicates that SDM strengthens the therapeutic alliance and is associated with higher patient satisfaction, improved adherence, and better health outcomes [6-10] without increasing consultation time or health care costs [11]. Additionally, when patients experience decision-making as participatory, they report enhanced affective and cognitive outcomes, including greater perceived control, improved understanding, and increased emotional security [12]. SDM is particularly relevant in the management of chronic or complex conditions, where sustainable treatment agreements are essential [13]. Overall, SDM contributes to the quality, sustainability, and acceptance of medical decisions.

Implementation Challenges in Clinical Practice

Despite its strong ethical and policy endorsement, SDM is not yet consistently implemented in routine clinical practice [14]. Although physicians often report favorable attitudes toward SDM, their actual engagement varies depending on the clinical context, the type of treatment decision, and patient characteristics [15]. Patient-related factors, including cultural beliefs, language barriers, and socioeconomic status, may further limit the consistent application of SDM in everyday care [6]. Moreover, effective participation requires both adequate knowledge and a sense of agency, which not all patients are equally equipped with [16]. Consequently, SDM remains widely endorsed in principle yet difficult to implement consistently in practice due to structural, contextual, and individual barriers [17].

AI in Medical Education and Communication Training

Training programs aimed at improving SDM and communication skills vary considerably in content, structure, and delivery format, and evidence regarding their effectiveness remains limited [18]. While educational interventions for medical students and health care professionals can improve communication skills in the short term [19], robust evidence on long-term effects, participant satisfaction, and sustained behavioral change remains limited [20]. While patient-reported experience measures exist, they are subject to bias, exhibit low response rates, and are not widely implemented [21]. Observer-based measures rely on human raters and are therefore resource-intensive, prone to lapses in concentration, and yield only modest reliability [22-24]. In parallel, AI is increasingly being integrated into medical education, particularly in the domains of communication training, simulation-based learning, formative feedback, and reflective practice [25-27]. AI-driven systems, including large language models, enable scalable and standardized simulated patient encounters that allow learners to repeatedly practice clinical communication in a safe environment [28-30]. Such systems can also provide immediate, structured feedback, potentially supporting deliberate practice and self-regulated learning [31].

Recent studies suggest that large language models (eg, generative pretrained transformer [GPT]–based systems) can reliably evaluate communication quality in simulated interactions and distinguish between different levels of performance [24]. For example, Selvaraj et al [24] demonstrated that AI-generated assessments of SDM-related communication show strong correlations with human ratings, indicating potential for more scalable and efficient evaluation processes. However, despite these promising developments, empirical evidence on the educational impact of AI-based SDM training—particularly regarding learner outcomes such as communicative self-efficacy (CSE), attitudes, and behavioral change—remains limited. Furthermore, CSE as a key factor for effective communication and the implementation of SDM has not been addressed in this context so far. However, evidence shows that communication skills training can improve both performance and self-efficacy [32].

Research Gaps and Study Rationale

Several important gaps remain in the current literature. First, there is limited evidence on the use of AI-simulated patients specifically for SDM training in medical education. Second, although AI systems have shown promise in evaluating communication behavior, few studies have systematically validated AI-generated feedback against human expert ratings in SDM contexts. Third, there is limited evidence on how AI-based SDM training influences CSE as an important learner-centered outcome in communication skills development. Finally, little is known about how learners perceive AI-generated feedback in SDM training and whether such feedback is experienced as useful and credible for supporting reflective learning.

To address these gaps, this study investigates a chatbot-based SDM training intervention using an AI-simulated patient combined with automated feedback. In addition, the accuracy and perceived usefulness of AI-generated feedback in SDM

XSL•FO
RenderX

training are assessed. Furthermore, the study examines changes in CSE and compares SDM attitudes and performance between medical students and physicians.

To evaluate the effectiveness of the proposed intervention, the following research questions are examined:

- Research question 1: to what extent can the AI chatbot provide accurate feedback on SDM using the Observing Patient Involvement in Decision Making (OPTION-12) scale compared with human ratings?
- Research question 2: how do medical students and physicians differ in SDM attitudes and performance?
- Research question 3: how does CSE change after practicing SDM with an AI chatbot?

Methods

Study Design and Overview

This feasibility study used a mixed methods pre-post design to evaluate an AI-supported SDM training intervention. The study consisted of a single-session interaction with an AI-simulated patient, followed by automated feedback and postintervention assessment. In addition, a group comparison was conducted between licensed physicians and medical students. Quantitative data were collected at baseline and immediately after the intervention, complemented by qualitative reflections.

Setting and Participants

Participants included medical students and licensed physicians. Eligibility criteria comprised sufficient proficiency in the German language and current enrollment in medical training after passing the second state examination or active clinical practice. Participants were recruited either via a study link distributed through mailing lists and professional networks or in person during psychosomatic training sessions, where they completed the study on site. Participation was voluntary, and no compensation was provided. Datasets with incomplete data and chats comprising fewer than 5 messages were excluded from analyses.

Intervention and Procedure

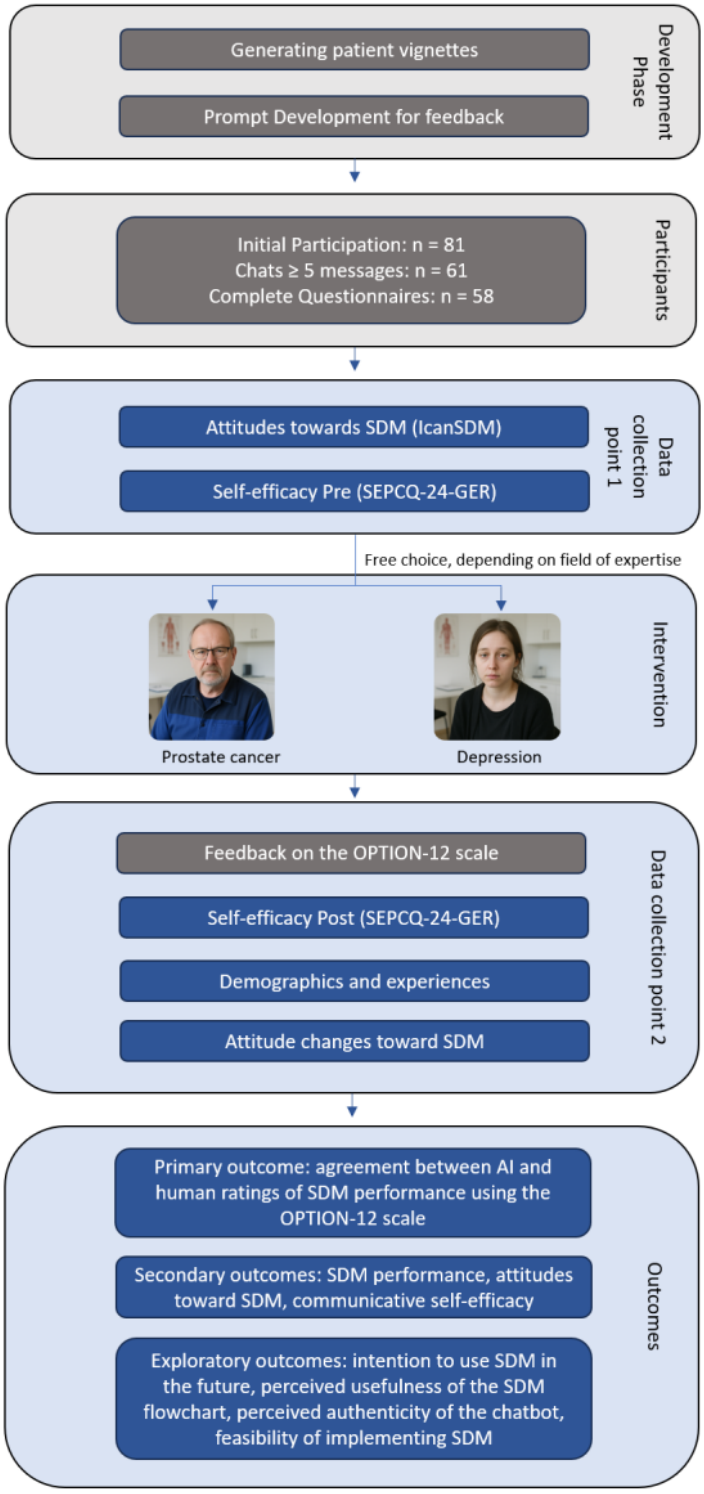
The data collection took place from November 2025 until February 2026. Initially, participants were introduced to an explanatory flowchart [33] illustrating 6 steps of SDM to ensure a basic level of SDM knowledge across all participants (Multimedia Appendix 1). Participants then engaged in a standardized 20-minute text-based consultation with an AI-simulated patient designed to elicit SDM behaviors. The AI system was configured to assume the role of a patient with a predefined clinical case and responded dynamically to participant input while maintaining consistency across interactions through structured prompting. Following the consultation, participants received AI-generated structured feedback on their SDM performance. The entire process, consisting of chat and questionnaires, took about 30 to 40 minutes. Data collection occurred at 2 time points:

- Preintervention: attitudes toward SDM (IcanSDM) and CSE (Self-Efficacy in Patient-Centeredness Questionnaire [SEPCQ-24-GER])
- Postintervention: CSE, perceived authenticity, perceived usefulness of feedback, and demographic variables

A detailed flowchart illustrating the study procedure is presented in Figure 1.

XSL•FO
RenderX

Figure 1. Flowchart illustrating the design of the experiment. OPTION-12: Observing Patient Involvement in Decision Making; SDM: shared decision-making; IcanSDM: measure of attitudes toward SDM; SEPCQ-24-GER: Self-Efficacy in Patient-Centeredness Questionnaire.



AI-Simulated Patient and Feedback System

This application is described in detail in previous studies [34-36]. As validated in [35], this setup produces stable, medically accurate conversations. The model GPT-4-0613 was used via the OpenAI API with the temperature set to 0.1 and other parameters left at default. Details about models and parameters are provided in Multimedia Appendix 2. The feedback system was set to request feedback on buckets of 4 items during each model call. A general description of the 5 response levels of the OPTION-12 scale was included in the prompt, as well as the complete chat. The model GPT-5 was used via the OpenAI API as it was the frontier model at the time of evaluation. Multimedia Appendix 2 provides complete prompts, as well as models and parameters. Screenshots, exemplary interactions with the AI-simulated patient, and

XSL•FO
RenderX

AI-generated feedback are provided in Multimedia Appendices 1 and 3 to illustrate the training experience and feedback structure.

To enable AI-generated feedback, a training dataset consisting of 9 expert-evaluated chats assessed using the OPTION-12 scale was developed in a separate pretest phase. These pretest data were collected prior to the main study with different participants and were fully independent of the final evaluation dataset. This strict separation ensured that no overlap occurred between data used for model refinement and data used for evaluation, thereby minimizing the risk of data leakage or circular evaluation. The absolute agreement between the human and AI ratings was moderate (intraclass correlation coefficient [ICC]=0.633, *df*=2,1, 95% CI 0.039-0.902; *P*=.02) but considered sufficient for the study's purpose. The pretest annotations were used to support prompt development and calibration of the AI feedback system. The resulting prompts were iteratively refined through a structured manual optimization process, including the integration of illustrative examples and clarification of category boundaries to improve output consistency and interpretability. The final set of prompts is provided in Multimedia Appendix 2.

Clinical Case Vignettes

Two clinical scenarios (prostate cancer and depression) were selected to represent distinct SDM contexts. The patient vignettes were specifically developed by experts. The first case describes a female student aged 28 years with recurrent depressive disorder. She is afraid of medication but is open to psychotherapy. The second case depicts a craftsman aged 62 years who presented with prostate cancer. He "just wants the thing out" and is skeptical of other procedures. The diagnosis is already known to the patients; the following conversation should therefore focus on discussing treatment options. The detailed patient vignettes are included in Multimedia Appendix 4.

Both cases involve distinct but comparable SDM challenges, including emotional burden, treatment uncertainty, and the need to balance medical evidence with patient preferences. The scenarios were also chosen to allow participants to engage with a case that was broadly aligned with their clinical background or expertise. Both cases were developed based on established clinical guidelines and structured to ensure comparability across participants. To ensure clinical validity and relevance, the cases were subsequently reviewed by additional experts in psychiatry and oncology. These experts assessed the medical accuracy, internal consistency, and appropriateness of the scenarios for SDM training. Feedback from this review process was incorporated to refine and finalize the case vignettes. The cases were integrated into the chatbot and comprehensively tested.

Measures

Attitudes Toward SDM

Attitudes toward SDM were collected using the German version of the IcanSDM scale [37]. The scale consists of 8 items and examines the perceived barriers and attitudes of health care professionals toward SDM. The response scale is based on an 11-point Likert scale ranging from 0 (strongly disagree) to 10 (strongly agree).

CSE

CSE was measured using the subscale "sharing information and power" of the SEPCQ-24-GER [38]. The subscale assesses the extent to which respondents feel able to share information and power with patients. The subscale consists of 10 items and is answered on a Likert scale from 1 (to a very small extent) to 5 (to a very large extent). In the study, participants answered the scale twice, once before and once after the chat.

SDM Performance

To assess SDM performance, the OPTION-12 scale [22] was used. The OPTION-12 scale is an observer-based instrument consisting of 12 items that capture key elements of SDM during clinical encounters. Each item is rated on a Likert scale from 0 (competence not observed) to 4 (competence meets a very high standard), resulting in a total score reflecting the extent to which clinicians involve patients in decision-making. In this context, the AI adopted the role of the patient and evaluated the extent to which the patient was involved in the treatment decision using this instrument. The OPTION-12 scale is widely regarded as a valid and established instrument for assessing SDM in clinical encounters. The interrater ICC (0.77) and Cronbach α (0.68) are above acceptable thresholds. However, its applicability in AI-mediated consultation settings has not yet been sufficiently evaluated, as assessed in the present study. Our AI-generated OPTION-12 ratings demonstrated similar reliability, with Cronbach α=0.88 and single-rater absolute ICC=0.86. In addition to AI-generated ratings, an expert in clinical communication independently rated all anonymized transcripts. The expert was familiar with observer-based SDM assessment and received standardized instruction on the use of the OPTION-12 scale. Both AI and human ratings were conducted blinded to each other's scores. No adjudication of discrepancies was performed. Agreement was assessed statistically using ICC and correlation analyses.

Perceived Feedback Quality and Authenticity

Participants rated the authenticity of the AI-simulated patient, the applicability of SDM in clinical practice, and their intention to use SDM more frequently. All items were assessed using a 5-point Likert scale ranging from "strongly disagree" to "strongly agree."

Outcome Definitions

Primary Outcome

The primary outcome was the level of agreement between AI-generated and human ratings of SDM performance as assessed using the OPTION-12 scale. Agreement was quantified using ICCs and interpreted as an indicator of the validity of AI-generated feedback.

Secondary Outcomes

Secondary outcomes included SDM performance as measured by the OPTION-12 total score, attitudes toward SDM assessed with the IcanSDM questionnaire, and CSE measured using the SEPCQ-24-GER. CSE was assessed preintervention and postintervention to evaluate changes following the AI-supported SDM training. Additionally, differences between medical

XSL•FO
RenderX

students and physicians were analyzed for SDM performance and attitudes.

Exploratory Outcomes

As exploratory outcomes, participants rated their intention to use SDM more frequently in future clinical practice, as well as the perceived usefulness of the SDM flowchart for supporting future application. In addition, they assessed the perceived authenticity of the chatbot interaction and the feasibility of implementing SDM in everyday clinical practice. To further investigate discrepancies between AI and human ratings, a qualitative analysis of selected cases was conducted.

Statistical Analysis

Chat scripts and questionnaire results were collected anonymously via the application. In total, 61 chats and 58 questionnaires were included in the analysis. Chats were segmented into question-answer pairs and compiled in a Microsoft Excel 2408 sheet [39]. Statistical analyses were performed in R (version 4.4.1; R Core Team, 2024 [40]), with significance set at $P<.05$. Figures were created using Microsoft Excel and R. All statistical tests were conducted using a 2-sided significance level of $\alpha=.05$. Descriptive statistics were calculated for all variables. Reported metrics include means, SDs, percentages, and Pearson correlations.

SDM performance was operationalized using the total score of the OPTION-12 scale. To evaluate the applicability of the OPTION-12 scale in the AI setting, internal consistency was assessed using Cronbach α, and item-level relationships were examined using interitem correlations. To assess the quality of AI-generated feedback, OPTION-12 ratings produced by the AI were compared with human ratings of the same interactions. Agreement between raters was evaluated using ICCs and Pearson correlation. ICCs were interpreted according to established guidelines [41]. Based on the 95% CI of the ICC estimate, values below 0.50, between 0.50 and 0.75, between 0.75 and 0.90, and above 0.90 indicate poor, moderate, good, and excellent reliability, respectively. These thresholds were used to interpret the level of agreement between AI-generated and human ratings.

Normality of continuous variables was assessed prior to analysis. Nonparametric Wilcoxon signed-rank tests were used when normality assumptions were violated. Homogeneity of variance between groups was assessed using variance ratio tests (*F* test). Depending on the outcome, either Student *t* test for independent samples (in cases of equal variances) or Welch *t* test (in cases of unequal variances) was applied to calculate group differences. Changes in CSE (pre-post) were analyzed using the Wilcoxon signed-rank test, as the distribution of paired differences deviated from normality. In addition to inferential statistics, the effect size was estimated using Cliff delta, a nonparametric measure that quantifies the magnitude and direction of differences between groups by estimating the probability of stochastic dominance without assuming a normal distribution.

Qualitative Analysis

A qualitative analysis of selected cases was conducted to investigate differences between AI and human OPTION-12 scores. Chats with the largest absolute differences in the OPTION-12 total scores were identified and subjected to in-depth examination. For each selected chat, item-level ratings were compared, and those with the greatest divergences were analyzed in detail. The analysis was guided by the OPTION-12 coding manual [42], with particular attention to how rating criteria were applied. This approach allowed for a systematic exploration of underlying reasons for rating differences between AI and human raters.

Preregistration

The study design, hypotheses, and analysis plan were preregistered on October 23, 2025, prior to data collection on the Open Science Framework [43].

Ethical Considerations

The study was approved by the Ethics Committee of the Faculty of Medicine at University Hospital Tübingen (375/2025BO2). Participation was voluntary, and no compensation was offered for participation. Informed consent was obtained from all participants prior to participation. Participants confirmed their agreement to the data collection and processing by ticking the respective box on the computer. Participants were fully informed about the study's purpose and procedures. All data were anonymized. Appropriate measures were taken to ensure the confidentiality and security of participant information. All procedures were conducted in accordance with the Declaration of Helsinki and ethical guidelines for human participants.

Results

Overview

Results are presented according to participant-level and encounter-level analyses and structured in line with the 3 research questions. All analyses were conducted in accordance with the prespecified primary, secondary, and exploratory outcomes.

Participant Characteristics and Outcomes

The following results address research questions 2 and 3 at the participant level.

Demographics

Of the 81 individuals who accessed the study in the first place, datasets with incomplete responses or fewer than 5 chat messages were excluded, as they were considered to reflect insufficient engagement. The final sample consisted of 58 participants, of whom 36 identified as female, 20 as male, and 2 did not specify. A total of 22 participants were students in their practical year (mean age 28, SD 3.08 years), 34 participants were certified physicians (mean age 35.79, SD 6.51 years), and 2 did not declare. Further, 27 participants reported working in a hospital setting, while 15 were working in private practice; the rest did not specify. Regarding medical specialty, 20 participants reported a background in general medicine, 4 in gynecology, 1 in psychosomatic medicine, and 1 in pediatrics; the remaining participants selected "other."

XSL•FO
RenderX

Differences Between Medical Students and Physicians

In line with the secondary outcomes, a comparison of SDM attitudes between students (n=22) and physicians (n=34) was conducted using the IcanSDM mean score. Only participants who completed the IcanSDM questionnaire and specified their background (student or physician) were included in the analysis (n=56). As SDM scores were normally distributed for both students (W=0.974; *P*=.79) and physicians (W=0.940; *P*=.06) and no significant difference in variances ($F_{21,33}$=0.645; *P*=.29) was found, a standard independent-samples 2-tailed *t* test was conducted. The results indicated no significant difference in SDM attitudes between students (n=22; mean 3.96, SD 1.05) and physicians (n=34; mean 4.13, SD 1.31; t_{54}=–0.507; *P*=.61; 95% CI –0.835 to 0.498).

In addition, encounter-level SDM performance, measured by the OPTION-12 total score, was compared between students (n=22) and physicians (n=34). In total, 56 participants were included in this analysis. The OPTION scores were normally distributed for both students (W=0.955; *P*=.40) and physicians (W=0.962; *P*=.28). The assumption of equal variances was violated ($F_{21,33}$=0.290; *P*=.004); therefore, a Welch 2-tailed 2-sample *t* test was conducted. Results showed that students (n=22; mean 28.41, SD 4.38) performed significantly better than physicians (n=34; mean 22.79, SD 8.14; $t_{52.60}$=3.49; *P*=.002; 95% CI 2.24-8.99).

Changes in CSE Following SDM Practice

In the next step, also in line with the secondary outcomes, changes in CSE before and after practicing SDM in the chat were examined using the SEPCQ-24-GER scores (n=58 participants). A Wilcoxon signed-rank test was conducted as the differences between pretraining and posttraining scores were not normally distributed (W=0.865; *P*<.001). Results showed a statistically significant change in CSE (V=216; *P*=.003) with higher scores before training (mean 4.02, SD 0.49) than after (mean 3.77, SD 0.79). In an additional post hoc directional analysis using a one-sided test (alternative="less"), the decrease in scores remained statistically significant (V=216; *P*=.002). A small negative effect was confirmed by Cliff delta (Δ=–0.186; 95% CI –0.384 to 0.028).

Evaluation and Benefits

As exploratory outcomes, participants' perceptions of the AI-supported SDM training were assessed. The analysis included 57 participants who completed the exploratory questions. Participants generally reported positive perceptions of SDM. For the intention to use SDM more often from now on, 52.63% (30/57) of participants agreed or strongly agreed, and 36.84% (21/57) were neutral. A total of 66.67% (38/57) of participants agreed or totally agreed with the statement that the flowchart depicting SDM shown in the study would help them apply SDM more often in the future. The chatbot was considered authentic by 54.39% (31/57) of participants, though 29.83% (17/57) were neutral. A total of 87.72% (50/57) of participants agreed or strongly agreed that SDM is feasible in everyday clinical practice.

Encounter-Level Analysis of SDM Performance

Overview

The final section reports validation analyses of AI-generated OPTION-12 ratings (research question 1). For the validation analysis, 61 chats meeting predefined completeness criteria (minimum of 5 messages exchanged) were included to ensure meaningful participation.

Descriptive Results of OPTION-12 Scores

Overall, the mean OPTION-12 score across encounters was 25.23 (SD 7.34; n=61) out of a maximum of 48 possible points. The median score was 26 (IQR 21-30). Observed scores ranged from 7 to 37. At the item level, the lowest mean score was obtained on item 3 ("The clinician assesses the patient's preferred approach to receiving information to assist decision-making"; mean 0.87, SD 0.74), and the highest mean score on item 2 ("The clinician states that there is more than one way to deal with the identified problem"; mean 3.23, SD 0.96).

Psychometric Properties of AI-Generated OPTION-12 Ratings

The AI-generated ratings (n=61) demonstrated good internal consistency across the 12 items (Cronbach α=0.88, 95% CI 0.83-0.92), with an average interitem correlation of *r*=0.43. The corrected item-total correlation (r.drop) ranged from 0.440 (item 12: "The clinician indicates the need to review the decision [or deferment]") to 0.733 (item 2: "The clinician states that there is more than one way to deal with the identified problem"). The corrected item-total correlation (r.drop) reflects the correlation between each item and the total score excluding that item, indicating that all items positively contributed to internal consistency. More psychometric details for human and AI ratings are presented in Table 1.

XSL•FO
RenderX

Table 1. Observing Patient Involvement in Decision Making (OPTION-12) item-level statistics.

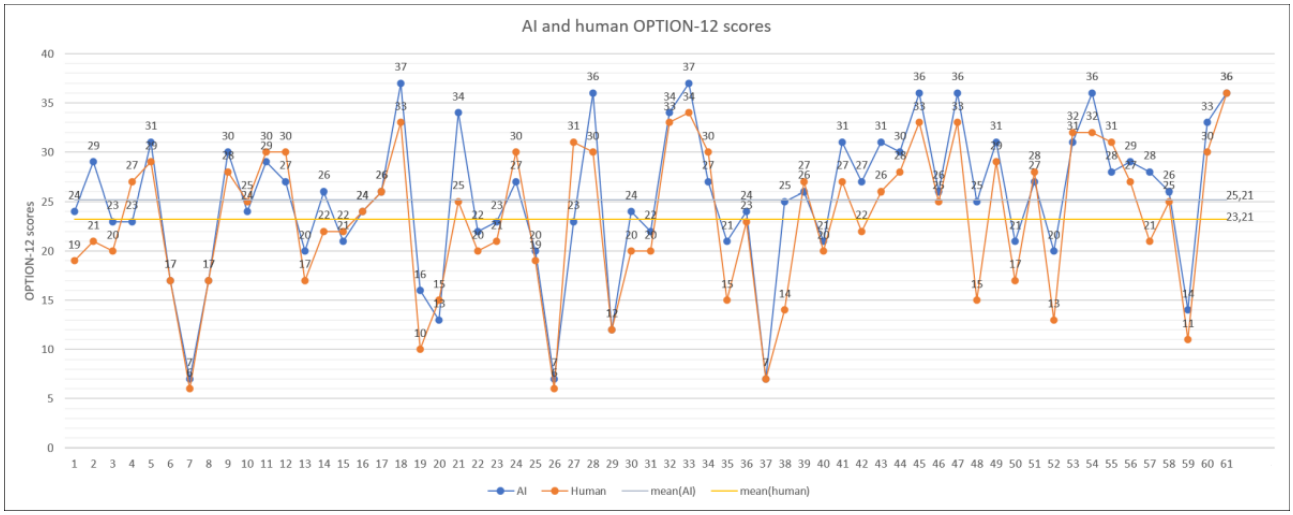
OPTION-12 items	Human rating, mean (SD)	AI rating, mean (SD)	Corrected item-total correlation (r.drop)	Interrater agreement (ICC)
Item1	2.77 (0.80)	2.97 (0.58)	0.717	0.578
Item2	2.61 (1.05)	3.23 (0.96)	0.733	0.645
Item3	0.41 (0.74)	0.87 (0.74)	0.560	0.520
Item4	2.49 (0.96)	2.49 (0.79)	0.713	0.787
Item5	2.05 (0.96)	1.90 (0.83)	0.713	0.727
Item6	2.44 (0.99)	2.43 (0.74)	0.445	0.689
Item7	2.08 (1.20)	2.16 (1.02)	0.521	0.769
Item8	1.46 (1.29)	1.33 (0.72)	0.663	0.609
Item9	1.74 (1.22)	2.11 (1.32)	0.488	0.839
Item10	1.59 (0.92)	1.69 (0.90)	0.714	0.724
Item11	1.92 (0.88)	2.07 (0.81)	0.614	0.536
Item12	1.64 (1.17)	1.98 (1.43)	0.440	0.791
Total	23.2 (7.52)	25.2 (7.34)	N/A[a]	0.859

[a]N/A: not applicable.

Agreement Between AI and Human OPTION-12 Ratings

In line with the primary outcome, agreement between AI and human ratings (n=61) was analyzed. There was strong agreement between the AI and the human rater for the total score (ICC=0.86, *df*=2,1; *P*<.001), although the AI systematically assigned slightly higher scores (mean 25.21, SD 7.34) than the human rater (mean 23.21, SD 7.51). The ICC quantifies the level of agreement between human and AI ratings. ICC (*df*=2,1) values per item, reflecting absolute agreement between AI and human ratings, ranged from 0.520 (item 3: "The clinician assesses the patient's preferred approach to receiving information to assist decision-making") to 0.839 (item 9: "The clinician offers the patient explicit opportunities to ask questions during the decision-making process"), suggesting moderate to strong agreement across items. Table 1 provides more details. The ratings of the 2 raters were strongly positively correlated (*r*=0.88), indicating a high level of consistency in relative scoring across cases. Exact agreement between raters occurred in 9.84% (6/61) of cases. In 21.31% (13/61) of cases, the human rater assigned a higher score, whereas in 68.85% (42/61) of cases, the AI assigned a higher score. Figure 2 provides more details.

Figure 2. Comparison of human and AI ratings on the Observing Patient Involvement in Decision Making (OPTION-12) scale, including mean values.



Qualitative Analysis of Individual Chats

As can be seen in Figure 2, the ratings that AI and human raters allocated differ considerably in some cases. To understand these large discrepancies, the 3 chats with the highest differences in the total OPTION-12 score (chats 21, 38 and 48 with differences of 9, 11, and 10, respectively) were analyzed qualitatively. For each chat, item-level differences between AI and human ratings were examined, and items with the greatest discrepancies were analyzed in detail and compared against the coding manual definitions [42]. In chat 21, item 5 ("The clinician explains the pros and cons of options to the patient"), the physician focused

XSL•FO
RenderX

exclusively on explaining the effectiveness and side effects of medications rather than outlining the advantages and disadvantages of all available treatment options. While the human rater assessed the item as a minimal attempt, the AI rated it as exhibiting a good standard. In chat 38, item 7 ("The clinician explores the patient's concerns [fears] about how problem[s] are to be managed"), AI ratings credited items as fulfilled even when patients introduced relevant information without being explicitly prompted by the physician. Similarly, for item 10 ("The clinician elicits the patient's preferred level of involvement in decision-making"), the physician only implicitly inquired about patients' preferred level of involvement in decision-making; nevertheless, the patient provided minimal indications of their preferences, which were again recognized as sufficient by the AI but not by the human rater. In chat 48, item 2 ("The clinician states that there is more than one way to deal with the identified problem"), the physician acknowledged the existence of multiple valid treatment approaches without providing a comprehensive or detailed overview. The AI rated this as "The behavior is observed and performed at a high level," while the human rater rated it as "The clinician gives the impression that the options are valid and need to be examined more closely." Overall, discrepancies appear to stem from differences in how implicit communication and patient-initiated contributions are interpreted and evaluated. More details are available in Multimedia Appendix 5.

Discussion

Principal Findings

This study examined a chat-based SDM training intervention using an AI-simulated patient combined with automated feedback on SDM performance. The principal findings indicate that participants were able to engage in structured SDM interactions within the AI-simulated patient and that the system was generally perceived as feasible for communication training. Preliminary results indicate good to strong agreement between AI-generated and human ratings of SDM performance based on the OPTION-12 scale. In addition, medical students and physicians reported similarly positive attitudes toward SDM, but students achieved higher performance scores on the OPTION-12 scale. CSE slightly decreased following the SDM practice intervention and feedback. The observed engagement with the AI chatbot in SDM tasks, as well as the alignment between AI and human ratings, suggests that large language models can support structured communication training. However, given the exploratory nature of the study, the limited sample size, and the fact that the results do not provide sufficient evidence for the validity of AI-generated assessments for SDM performance, the results should be interpreted as preliminary evidence for future research.

Interpretation of the Findings

Both groups of medical students and licensed physicians reported similarly positive attitudes toward SDM. This is in line with the consistent endorsement of SDM principles across levels of professional training described in the literature [15]. In contrast, differences emerged in observed SDM performance, with medical students achieving significantly higher OPTION-12 scores than physicians, indicating greater adherence to structured SDM approaches in recently trained students [44]. These findings are in line with previous research reporting an inverse association between SDM-related competencies and age [45]. One possible explanation is that students may adhere more closely to structured SDM models due to recent formal training, whereas physicians may rely more on routinized communication patterns shaped by clinical experience [46-48]. The discrepancy between attitudes and observable behavior highlights the potential gap between endorsement of SDM and its practical implementation [14,15].

The observed decrease in CSE following the intervention may reflect a possible recalibration effect, as has been suggested in previous educational research [49]. This effect describes a phenomenon in which engagement in structured learning and feedback increases awareness of the complexity of the task and potentially lowers self-perceived competence. However, in the absence of a control group, this interpretation remains speculative, and causal conclusions cannot be drawn.

Descriptive analyses of OPTION-12 values showed moderate overall performance and values similar to those reported in the literature [50]. Certain SDM components appeared to be implemented more consistently (eg, "The clinician states that there is more than one way to deal with the identified problem"), while others were more challenging (eg, "The clinician assesses the patient's preferred approach to receiving information to assist decision-making").

Regarding the comparison of AI-generated and human ratings, the 2 measures showed a strong positive correlation and good interrater reliability, indicating a high level of consistency, as previously described in similar contexts [24]. Nevertheless, the AI consistently assigned higher scores than the human rater. This tendency to agree and give socially desirable answers is a pattern that is already recognized [51,52]. In the exemplary qualitative analysis conducted as part of our study, it became clear that discrepancies appear to stem mainly from differences in how implicit communication and patient-initiated contributions are interpreted. Overall, these findings should be interpreted cautiously given the exploratory nature of the study and the absence of a fully established validation framework.

Practical Implications

The results suggest that AI-powered chatbots may be a viable and scalable tool for teaching SDM skills in medical education and professional development. Feedback on real clinical conversations is often difficult to implement, critical to the protection of personal data, and resource-intensive to obtain, making AI a less demanding alternative. This study shows that AI can apply the OPTION-12 scale in a similar way to a human rater and can adopt the patient's perspective during the assessment. Without having surveyed actual patients, we were therefore able to gain an initial impression of the patient's perspective, as suggested in the literature [20,21]. Such tools can complement existing communication training formats by providing low-threshold, time-independent practice opportunities and individualized feedback. In terms of training practice, integrating AI-supported SDM training into medical curricula

XSL•FO
RenderX

could help standardize the teaching of SDM principles and establish SDM on a broad scale.

Limitations and Recommendations for Future Research

Several limitations should be considered when interpreting the results of this study. First, the simulated consultations were text-based and conducted with an AI-simulated patient, which may not have fully captured the complexity of real-world clinical encounters, particularly nonverbal aspects of communication. Future research should therefore examine the transferability of SDM skills acquired in chatbot-based training to real-world patient encounters. Second, the study was based on a pre-post design without a control group, which limits causal inferences about the effects of chatbot-based SDM training. Future studies should focus on randomized controlled designs compared with traditional teaching methods. Furthermore, longitudinal follow-up assessments are necessary to evaluate sustained impact. Third, the study was conducted with a relatively small and self-selected sample, which may introduce selection bias and limit representativeness. Furthermore, the study did not assess participants' prior familiarity with AI tools, digital literacy, or typing speed, all of which may have influenced their interaction with the chatbot. Participants with higher levels of digital competence or faster typing skills may have been better able to engage with the system, potentially confounding performance outcomes. Future studies should consider measuring and controlling for these factors to better isolate the effects of the intervention. The fairly high dropout rate could be attributed to a lack of time and competing clinical or academic obligations, since this was a voluntary study with no incentives. Participants may have clicked on the link out of initial interest but lost interest shortly thereafter, perhaps due to a lack of time. Future studies should therefore aim to recruit larger and more diverse samples across multiple institutions to improve generalizability and reduce selection bias. Fourth, the evaluation of the AI-generated feedback was conducted by only 1 rater. This introduces a potential risk of subjectivity and limits the robustness of the validation process, reducing the methodological rigor of the findings. Future research should therefore incorporate multiple raters and report interrater reliability metrics to ensure a more reliable and comprehensive validation of AI-generated feedback. Finally, performance bias cannot be excluded, as participants were not blinded to the study purpose and may have adapted their behavior based on perceived expectations. Similarly, assessment bias may have been introduced due to the lack of blinding in human ratings. Future research should implement blinding procedures for both participants and raters wherever feasible to minimize expectancy effects and rating bias.

Conclusion

This exploratory study provides preliminary evidence that AI-simulated patients can be integrated into communication training to support SDM. Participants were able to engage in structured SDM interactions within the AI environment, and initial results suggest a good level of agreement between AI-generated and human ratings of SDM performance. However, these findings should be interpreted cautiously due to the exploratory design, limited sample size, and methodological constraints. While AI-simulated patients may have potential as a supportive tool for SDM training, further research is required to establish validity, reliability, and generalizability using larger samples and standardized, rigorously validated AI implementations.

Acknowledgments

The authors acknowledge the support of the Open Access Publishing Funds of the University of Tuebingen. Generative AI tools (ChatGPT and DeepL Write) were used in the preparation of this manuscript to improve writing quality. After using these tools, the authors thoroughly reviewed, revised, and edited the content and take full responsibility for the final version of the manuscript.

Funding

The study itself received no specific funding. The authors acknowledge support from the Open Access Publishing Fund of the University of Tübingen.

Data Availability

The data that support the findings of this study are available from the corresponding author (LH) upon reasonable request.

Authors' Contributions

LH and AH-W conceptualized and designed the study. LH and AS were responsible for the data collection. LH and TF-W performed the data analysis and interpretation. LH contributed to the creation of figures and illustrations for the manuscript. FH and TA provided resources and supervision. BH was responsible for the software. LH drafted the manuscript, and all authors participated in revising it critically for intellectual content. All authors provided final approval of the version to be published.

Conflicts of Interest

None declared.

XSL•FO
RenderX

Multimedia Appendix 1
Screenshots of the AI-based Application.
[DOCX File , 282 KB - mededu_v12i1e100467_app1.docx]

Multimedia Appendix 2
Model and prompting details.
[DOCX File , 33 KB - mededu_v12i1e100467_app2.docx]

Multimedia Appendix 3
Example of an interaction with the AI-simulated patient (translated into English).
[DOCX File , 16 KB - mededu_v12i1e100467_app3.docx]

Multimedia Appendix 4
AI-simulated patient vignettes.
[DOCX File , 76 KB - mededu_v12i1e100467_app4.docx]

Multimedia Appendix 5
Qualitative analysis.
[DOCX File , 24 KB - mededu_v12i1e100467_app5.docx]

References

1. Elwyn G, Laitner S, Coulter A, Walker E, Watson P, Thomson R. Implementing shared decision making in the NHS. BMJ 2010;341:c5146 [FREE Full text] [doi: 10.1136/bmj.c5146] [Medline: 20947577]
2. Menzel H, Coleman J, Katz E. Dimensions of being "modern" in medical practice. J Chronic Dis 1959;9(1):20-40. [doi: 10.1016/0021-9681(59)90135-3]
3. Charles C, Gafni A, Whelan T. Shared decision-making in the medical encounter: what does it mean? (or it takes at least two to tango). Soc Sci Med 1997;44(5):681-692. [doi: 10.1016/s0277-9536(96)00221-3] [Medline: 9032835]
4. Emanuel EJ, Emanuel LL. Four models of the physician-patient relationship. JAMA 1992;267(16):2221. [doi: 10.1001/jama.1992.03480160079038]
5. Eddy DM. Anatomy of a Decision. JAMA 1990;263(3):441. [doi: 10.1001/jama.1990.03440030128037]
6. Grad R, Sandhu A, Majdpour D, Kitner S, Rodriguez C, Elwyn G. Perspectives of family physician educators on shared decision making in preventive health care: a qualitative descriptive inquiry. Patient Educ Couns 2025;134:108681 [FREE Full text] [doi: 10.1016/j.pec.2025.108681] [Medline: 39889418]
7. Hughes TM, Merath K, Chen Q, Sun S, Palmer E, Idrees JJ, et al. Association of shared decision-making on patient-reported health outcomes and healthcare utilization. Am J Surg 2018;216(1):7-12. [doi: 10.1016/j.amjsurg.2018.01.011] [Medline: 29395026]
8. Niburski K, Guadagno E, Abbasgholizadeh-Rahimi S, Poenaru D. Shared decision making in surgery: a meta-analysis of existing literature. Patient 2020;13(6):667-681. [doi: 10.1007/s40271-020-00443-6] [Medline: 32880820]
9. Suh WS, Lee CK. Impact of shared-decision making on patient satisfaction. J Prev Med Public Health 2010;43(1):26-34 [FREE Full text] [doi: 10.3961/jpmph.2010.43.1.26] [Medline: 20185980]
10. Turabian JL. Strategies to improve therapeutic compliance in general medicine. Community Med Health Educ Res 2019;1(1):31-39 [FREE Full text]
11. Bruch JD, Khazen M, Mahmic-Kaknjo M, Légaré F, Ellen ME. The effects of shared decision making on health outcomes, health care quality, cost, and consultation time: an umbrella review. Patient Educ Couns 2024;129:108408. [doi: 10.1016/j.pec.2024.108408] [Medline: 39214045]
12. Shay LA, Lafata JE. Where is the evidence? A systematic review of shared decision making and patient outcomes. Med Decis Making 2014;35(1):114-131. [doi: 10.1177/0272989x14551638]
13. Joosten EAG, DeFuentes-Merillas L, de Weert GH, Sensky T, van der Staak CPF, de Jong CAJ. Systematic review of the effects of shared decision-making on patient satisfaction, treatment adherence and health status. Psychother Psychosom 2008;77(4):219-226. [doi: 10.1159/000126073] [Medline: 18418028]
14. Alsulamy N, Lee A, Thokala P, Alessa T. What influences the implementation of shared decision making: an umbrella review. Patient Educ Couns 2020;103(12):2400-2407 [FREE Full text] [doi: 10.1016/j.pec.2020.08.009] [Medline: 32839047]
15. Pollard S, Bansback N, Bryan S. Physician attitudes toward shared decision making: a systematic review. Patient Educ Couns 2015;98(9):1046-1057. [doi: 10.1016/j.pec.2015.05.004] [Medline: 26138158]
16. Joseph-Williams N, Elwyn G, Edwards A. Knowledge is not power for patients: a systematic review and thematic synthesis of patient-reported barriers and facilitators to shared decision making. Patient Educ Couns 2014;94(3):291-309. [doi: 10.1016/j.pec.2013.10.031] [Medline: 24305642]

XSL•FO
RenderX

17. Elwyn G, Gulbrandsen P, Leavitt H, Abukmail E, Clayman ML, Edwards A, et al. Shared decision-making. A primer for clinicians. J Gen Intern Med 2025;40(16):3889-3899. [doi: 10.1007/s11606-025-09707-z] [Medline: 41055684]
18. Légaré F, Politi MC, Drolet R, Desroches S, Stacey D, Bekker H, SDM-CPD Team. Training health professionals in shared decision-making: an international environmental scan. Patient Educ Couns 2012;88(2):159-169. [doi: 10.1016/j.pec.2012.01.002] [Medline: 22305195]
19. Xiao L, Huang C, Bai Y, Ding J. Shared decision-making training embedded in undergraduate and postgraduate medical education: a scoping review. Patient Educ Couns 2024;123:108186. [doi: 10.1016/j.pec.2024.108186] [Medline: 38331626]
20. Singh Ospina N, Toloza FJK, Barrera F, Bylund CL, Erwin PJ, Montori V. Educational programs to teach shared decision making to medical trainees: a systematic review. Patient Educ Couns 2020;103(6):1082-1094. [doi: 10.1016/j.pec.2019.12.016] [Medline: 32005556]
21. Gärtner FR, Bomhof-Roordink H, Smith IP, Scholl I, Stiggelbout AM, Pieterse AH. The quality of instruments to assess the process of shared decision making: a systematic review. PLoS One 2018;13(2):e0191747 [FREE Full text] [doi: 10.1371/journal.pone.0191747] [Medline: 29447193]
22. Elwyn G, Hutchings H, Edwards A, Rapport F, Wensing M, Cheung W, et al. The OPTION scale: measuring the extent that clinicians involve patients in decision-making tasks. Health Expect 2005;8(1):34-42 [FREE Full text] [doi: 10.1111/j.1369-7625.2004.00311.x] [Medline: 15713169]
23. Elwyn G, Frosch DL, Kobrin S. Implementing shared decision-making: consider all the consequences. Implement Sci 2016;11(1):114 [FREE Full text] [doi: 10.1186/s13012-016-0480-9] [Medline: 27502770]
24. Selvaraj SP, Yen RW, Forcino R, Elwyn G. Automating the observer OPTION-5 measure of shared decision making: assessing validity by comparing large language models to human ratings. Patient Educ Couns 2026;142:109362 [FREE Full text] [doi: 10.1016/j.pec.2025.109362] [Medline: 41016196]
25. Stamer T, Steinhäuser J, Flägel K. Artificial intelligence supporting the training of communication skills in the education of health care professions: scoping review. J Med Internet Res 2023;25:e43311 [FREE Full text] [doi: 10.2196/43311] [Medline: 37335593]
26. Jacobs C, Johnson H, Tan N, Brownlie K, Joiner R, Thompson T. Application of AI communication training tools in medical undergraduate education: mixed methods feasibility study within a primary care context. JMIR Med Educ 2025;11:e70766-e70766 [FREE Full text] [doi: 10.2196/70766] [Medline: 41135055]
27. Chan KS, Zary N. Applications and challenges of implementing artificial intelligence in medical education: integrative review. JMIR Med Educ 2019;5(1):e13930 [FREE Full text] [doi: 10.2196/13930] [Medline: 31199295]
28. Sriram A, Ramachandran K, Krishnamoorthy S. Artificial intelligence in medical education: transforming learning and practice. Cureus 2025;17(3):e80852. [doi: 10.7759/cureus.80852] [Medline: 40255837]
29. Müller Y, Lemberg H, Gellert P, Zoellick J. Training empathetic communication skills in medical students with a role-prompted GPT-4o chatbot: quasi-experimental intervention study. JMIR Med Educ 2026;12:e88092 [FREE Full text] [doi: 10.2196/88092] [Medline: 42284570]
30. Bowers P, Graydon K, Ryan T, Lau JH, Tomlin D. Artificial intelligence-driven virtual patients for communication skill development in healthcare students: a scoping review. Australas J Educ Technol 2024:39-57. [doi: 10.14742/ajet.9307]
31. Yu S, Liu J. Mapping the landscape of AI-assisted formative feedback in medical education: a bibliometric analysis. Medicine (Baltimore) 2026;105(5):e47489. [doi: 10.1097/md.0000000000047489]
32. Mata NDS, de Azevedo KPM, Braga LP, de Medeiros GCBS, de Oliveira Segundo VH, Bezerra INM, et al. Training in communication skills for self-efficacy of health professionals: a systematic review. Hum Resour Health 2021;19(1):30 [FREE Full text] [doi: 10.1186/s12960-021-00574-3] [Medline: 33676515]
33. Was ist SHARE TO CARE? SHARE TO CARE. URL: https://share-to-care.de/programm [accessed 2026-04-07]
34. Bodonhelyi A, Stegemann-Philipps C, Sonanini A, Herschbach L, Szep M, Herrmann-Werner A, et al. Modeling challenging patient interactions: LLMs for medical communication training. ArXiv Preprint posted online on March 28, 2025 [FREE Full text]
35. Holderried F, Stegemann-Philipps C, Herschbach L, Moldt J, Nevins A, Griewatz J, et al. A generative pretrained transformer (GPT)-powered chatbot as a simulated patient to practice history taking: prospective, mixed methods study. JMIR Med Educ 2024;10:e53961 [FREE Full text] [doi: 10.2196/53961] [Medline: 38227363]
36. Herschbach L, Festl-Wietek T, Stegemann-Philipps C, Sonanini A, Herrmann B, Erschens R, et al. Evaluation of an AI-Based Chatbot providing real-time feedback in communication training for mental health care professionals: proof-of-concept observational study. J Med Internet Res 2025;27:e82818 [FREE Full text] [doi: 10.2196/82818] [Medline: 41314643]
37. Giguere AM, Bogza LM, Coudert L, Carmichael PH, Renaud JS, Légaré F, et al. Development of the IcanSDM scale to assess primary care clinicians' ability to adopt shared decision-making. medRxiv 2014:2020.07.01.20144204. [doi: 10.1101/2020.07.01.20144204]
38. Heuser C, Ernstmann N, Karger A, Petermann-Meyer A, Geiser F, Vitinius F, et al. Measuring self-efficacy in patient-centeredness of physicians in oncology—translation, adaptation and evaluation of psychometric properties of the SEPCQ-27 German version (SEPCQ-24-GER). Ann Med 2025;57(1):2468255 [FREE Full text] [doi: 10.1080/07853890.2025.2468255] [Medline: 40029046]

XSL•FO
RenderX

39. Microsoft Excel. Microsoft Corporation. 2023. URL: https://www.microsoft.com/en-in/microsoft-365/excel [accessed 2026-07-12]
40. The R Project for Statistical Computing. The R Foundation. 2024. URL: https://www.r-project.org/ [accessed 2026-07-12]
41. Koo TK, Li MY. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. J Chiropr Med 2016;15(2):155-163 [FREE Full text] [doi: 10.1016/j.jcm.2016.02.012] [Medline: 27330520]
42. OPTION rater manual: observing patient involvement: evaluating the extent that clinicians involve patients in decisions. glynelwyn.com. 2005. URL: https://www.glynelwyn.com/uploads/2/4/0/4/24040341/option_12_rater_manual.pdf [accessed 2026-07-12]
43. Open Science Framework. URL: https://osf.io/935rd/ [accessed 2026-07-08]
44. Yen RW, Barr PJ, Cochran N, Aarts JW, Légaré F, Reed M, et al. Medical students' knowledge and attitudes toward shared decision making: results from a multinational, cross-sectional survey. MDM Policy Pract 2019;4(2):2381468319885871 [FREE Full text] [doi: 10.1177/2381468319885871] [Medline: 31742232]
45. Elwyn G, Edwards A, Wensing M, Hood K, Atwell C, Grol R. Shared decision making: developing the OPTION scale for measuring patient involvement. Qual Saf Health Care 2003;12(2):93-99 [FREE Full text] [doi: 10.1136/qhc.12.2.93] [Medline: 12679504]
46. Sehlbach C, Teunissen PW, Driessen EW, Mitchell S, Rohde GGU, Smeenk FWJM, et al. Learning in the workplace: use of informal feedback cues in doctor-patient communication. Med Educ 2020;54(9):811-820 [FREE Full text] [doi: 10.1111/medu.14148] [Medline: 32150761]
47. Shing YL, Brod G. Effects of prior knowledge on memory: implications for education. Mind Brain Educ 2016;10(3):153-161. [doi: 10.1111/mbe.12110]
48. Caddick ZA, Fraundorf SH, Rottman BM, Nokes-Malach TJ. Cognitive perspectives on maintaining physicians' medical expertise: II. Acquiring, maintaining, and updating cognitive skills. Cogn Res Princ Implic 2023;8(1):47 [FREE Full text] [doi: 10.1186/s41235-023-00497-8] [Medline: 37488460]
49. Nuutila K, Tapola A, Tuominen H, Molnár G, Niemivirta M. Mutual relationships between the levels of and changes in interest, self-efficacy, and perceived difficulty during task engagement. Learn Individ Differ 2021;92:102090. [doi: 10.1016/j.lindif.2021.102090]
50. Stubenrouch FE, Pieterse AH, Falkenberg R, Santema TKB, Stiggelbout AM, van der Weijden T, et al. OPTION(5) versus OPTION(12) instruments to appreciate the extent to which healthcare providers involve patients in decision-making. Patient Educ Couns 2016;99(6):1062-1068. [doi: 10.1016/j.pec.2015.12.019] [Medline: 26776490]
51. Tekin M, Yurdal MO, Toraman Ç, Korkmaz G, Uysal Ö. Is AI the future of evaluation in medical education? AI vs human evaluation in objective structured clinical examination. BMC Med Educ 2025;25(1):641 [FREE Full text] [doi: 10.1186/s12909-025-07241-4] [Medline: 40312328]
52. Wetzler EL, Cassidy KS, Jones MJ, Frazier CR, Korbut NA, Sims CM, et al. Grading the graders: comparing generative AI and human assessment in essay evaluation. Teach Psychol 2024;52(3):298-304. [doi: 10.1177/00986283241282696]

Abbreviations

CSE: communicative self-efficacy
GPT: generative pretrained transformer
ICC: intraclass correlation coefficient
OPTION-12: Observing Patient Involvement in Decision Making
SDM: shared decision-making
SEPCQ-24-GER: Self-Efficacy in Patient-Centeredness Questionnaire

Edited by A Stone; submitted 08.May.2026; peer-reviewed by L Mosca, N Delvaux; comments to author 04.Jun.2026; revised version received 02.Jul.2026; accepted 02.Jul.2026; published 16.Jul.2026.

Please cite as:
Herschbach L, Festl-Wietek T, Sonanini A, Holderried F, Albrecht T, Herrmann B, Herrmann-Werner A
AI-Simulated Patients for Training Shared Decision-Making: Feasibility Study in Medical Education
JMIR Med Educ 2026;12:e100467
URL: https://mededu.jmir.org/2026/1/e100467
doi:10.2196/100467
PMID:

©Lea Herschbach, Teresa Festl-Wietek, Alessandra Sonanini, Friederike Holderried, Tobias Albrecht, Bernd Herrmann, Anne Herrmann-Werner. Originally published in JMIR Medical Education (https://mededu.jmir.org), 16.Jul.2026. This is an open-access

XSL•FO
RenderX

article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Education, is properly cited. The complete bibliographic information, a link to the original publication on https://mededu.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Estimating the Prevalence of Generative AI Use in Medical School Application Essays: Cross-Sectional Study

Nicholas C Spies[1], MD; Valerie S Ratts[2], MD; Ian S Hagemann[3], MD, PhD

[1]Department of Pathology, University of Utah, Salt Lake City, UT, United States

[2]Department of Obstetrics and Gynecology, School of Medicine, Washington University in St. Louis, St. Louis, MO, United States

[3]Department of Pathology and Immunology, School of Medicine, Washington University in St. Louis, St. Louis, MO, United States

Corresponding Author:
Ian S Hagemann, MD, PhD
Department of Pathology and Immunology
School of Medicine
Washington University in St. Louis
425 S. Euclid Ave.
St. Louis, MO, 63110
United States
Phone: 1 3147478924
Email: hagemani@wustl.edu

Abstract

Background: Generative AI tools became widely available to the public in November 2022. The extent to which these tools have been used by medical school applicants during the admissions process is unknown.

Objective: We aimed to estimate the extent of generative AI use among cohorts of applicants spanning the rollout of these tools.

Methods: We retrospectively analyzed 6000 essays from 2364 applicants submitted to a US medical school in 2021 to 2022 (baseline, before the wide availability of AI) and 2023 to 2024 (test year) to estimate the prevalence of AI use and its relation to other application data. We used GPTZero, a commercially available detection tool, to generate a metric (P_{human}) reflecting the predicted probability that each essay was completely human generated, ranging from 0 (the essay appears to be entirely AI generated) to 1 (the essay appears to be entirely human generated).

Results: Fully human-generated negative controls demonstrated a median P_{human} of 0.93 (range 0.89-0.97), while fully AI-generated positive controls demonstrated a median P_{human} of 0.01 (range 0.00-0.01). The "Personal Comments" essays submitted in the 2023 to 2024 application cycle had a median P_{human} of 0.77 (95% CI 0.76-0.78) compared with 0.83 (95% CI 0.82-0.85) during the 2021 to 2022 cycle. Approximately 12.3% and 2.7% of essays were evaluated as having P_{human} <0.5 in the test and baseline years, respectively. Essays submitted as part of the secondary application demonstrated lower P_{human} values than those of the American Medical College Application Service (AMCAS) "Personal Comments" essays. In applicant-clustered, multivariable generalized estimating equation analyses, supplementary essay type and younger age were significantly associated with lower P_{human}. Application completion date, self-reported gender, program type (MD vs MD-PhD), grade point average (GPA), Medical College Admission Test (MCAT) score, socioeconomic status, and undergraduate major were not significant predictors after false discovery rate correction. P_{human} was not predictive of interview invitation or acceptance in adjusted applicant-level logistic regression analyses.

Conclusions: An AI detection algorithm identified signs of increased use of generative AI in 2023 to 2024 medical school admission applications compared to those in the 2021 to 2022 baseline period, before AI was widely available. AI use did not appear to confer an admissions advantage. Although these results provide information about the applicant pool as a whole, AI detection is imperfect. We do not recommend deploying AI detection for individual applications in live admissions cycles.

(JMIR Med Educ 2026;12:e96673) doi:10.2196/96673

KEYWORDS

medical schools; artificial intelligence; AI; medical students; student selection

XSL•FO
RenderX

Introduction

Generative AI has exploded in popularity and availability in recent years, with "chatbots" driven by large language models (LLMs) providing a means to rapidly accomplish text-oriented tasks. These tools have a myriad of potential applications in medical education, including information retrieval [1] and the generation of practice questions [2], clinical vignettes, and simulations [3]. However, they also threaten the pedagogical value of writing assignments by allowing students to produce responses with minimal effort or understanding [4]. AI's role in medical education remains a rapidly evolving field.

One setting where AI chatbots may have far-reaching implications is the admissions process. Medical school application essays are intended to give admissions officers information about applicants' interests, experiences, attributes, and motivations. Moreover, the essays can approximate the work quality that can be expected of students if they ultimately matriculate. In November 2022, the first widely available chatbot, ChatGPT 3.5, was released to the public, introducing the possibility that medical school applicants could use this tool either to write their essays outright or for subtasks such as drafting or editing.

There is no consensus on the extent to which AI use by medical applicants is acceptable. Applicants have long relied on aids such as spelling and grammar checkers, formal editing services, and feedback from friends, family, and advisers. Receiving help from a chatbot could level the playing field for applicants with less access to conventional writing aids. On the other hand, AI-generated essays provide less insight into the applicant's authentic self and work performance and are therefore less fit for purpose as components of a school application. For the 2023 to 2024 application cycle, the American Medical College Application Service (AMCAS) adopted a policy stating that essays must "not be written, in part or in whole, by another author and...not [be] the product of artificial intelligence" [5]. Although all applicants attested to this certification statement, it is uncertain to what degree they complied.

The popularity, potential impact, and ethical considerations of AI-driven chatbots have motivated the development of methods to distinguish AI-generated text from human-generated text. One reported method relies on the extent to which each word is predictable based on those that came before it ("perplexity") and the degree of variation in sentence length and structure ("burstiness") [6]. Human-generated writing tends to have higher perplexity and burstiness than the output of current LLMs, reflecting the greater variety and spontaneity of natural human expression. In this model, AI use is not detected directly; instead, detection of an AI-related style signal is used as a surrogate, with caveats.

An effective AI detector must be both sensitive and specific, similar to other diagnostic tests. In a benchmarking study, GPTZero attained an area under the receiver operating characteristic (AUROC) curve of greater than 96% on diverse writing samples, including product reviews, blogs, news stories, fiction, and, perhaps most relevant to this study, résumés written by a variety of LLMs [7]. We thus considered it fit for purpose for our study.

To explore the extent to which LLMs may be used in medical school applications, we conducted a retrospective study of AMCAS writing samples from before and after the surge in popularity of AI-driven chatbots in 2022. We analyzed these essays using GPTZero for AI detection [6] and compared the detector's outputs across relevant demographic and essay-related metadata.

Methods

Data Retrieval

We conducted this study at a US medical school that participates in the AMCAS Data Exchange Service. Applicants were included in the study if they selected the school as part of their AMCAS application, regardless of whether they completed a secondary application. We retrieved data from the school's applicant tracking system, identified only by a coded identifier. An honest broker who was not a member of the study team provided the data to eliminate the possibility of reidentification. We coded an application as "complete" if the student submitted an AMCAS file, letters of recommendation meeting the school's requirement, and a secondary application. We retrieved the following data from the AMCAS application: age, self-reported gender, program type (MD vs MD-PhD), visa status, socioeconomic status indicator, undergraduate majors, undergraduate grade point average (GPA), highest Medical College Admission Test (MCAT) 3-digit score, and the applicant's personal statement. We coded socioeconomic status using the AMCAS indicator EO1 and/or EO2 (indicating that neither parent has a college degree or holds an executive, managerial, or professional occupation) vs other. We manually coded each applicant's undergraduate majors as follows: science, technology, engineering, and mathematics (STEM); non-STEM; or both (applicable only to applicants declaring multiple majors; Table S1 in Multimedia Appendix 1). Visa status was retrieved as an imperfect surrogate for native English proficiency. We retrieved the following data from the school-specific secondary application: a supplementary essay in which the student was asked to describe a time in life when they were unsuccessful ("Failure" essay, optional in 2021-2022 but required in 2023-2024), a free-text box where they could enter any additional information they wished to share ("Anything Else" essay, optional), and the date of application submission. Some applicants lacked a submission date, indicating that they failed to assemble a complete file (usually due to not submitting their secondary application). Each student's interview decision (interview vs no interview) and admission decision (accepted vs rejected) were extracted from our database. We considered placement on the Alternate List as rejection for this study.

Within each application cycle, we constructed separate extracts for the "Personal Comments," "Failure," and "Anything Else" essays. A random sample of 1000 "Personal Comments" essays from each year was extracted. Then, any "Failure" or "Anything Else" essay for those applicant IDs was specifically included. Given the optional nature of the secondary essays and the fact that some students did not submit their secondary applications,

XSL•FO
RenderX

the extraction set was supplemented with another random sampling of secondary essays to complete the remaining cohort to 1000 essays of each type. This sampling approach enabled us to perform within- and across-applicant analyses to better characterize the association among year, applicant, and essay type with the likelihood of AI use.

Detecting Evidence of AI Use in Applicant Essays

We used the GPTZero API (version 2.0.0; Superhuman), a commercially available web-based service, to detect AI use. We sent the entirety of each writing sample to the API for analysis, which returned a vector (class_probabilities) of 3 elements: the model's estimated probabilities that the document was fully generated by a human, fully generated by AI, or mixed [8]. For downstream analysis, we used the first of these outputs, denoted as P_{human}. This measure is not the proportion of the essay predicted to be human-written, although there is likely to be a correlation. In essays written with heavy use of AI, many sentences will be flagged as AI-written or mixed; these sentences will have a low detected probability of being human-written, and the overall P_{human} will be low. Essays with mainly human-written sentences will, conversely, have a high sentence-wise and document-wise predicted probability of being written by a human and, therefore, a high P_{human}. A P_{human} of <0.5 denotes an input that was more likely than not written with the help of AI.

Given the reported difficulties in accurately discriminating between human and AI-generated text [9], we first performed a preliminary analysis using negative and positive controls. We defined continuation criteria for the real-world analysis as the detection of a difference in predicted probability of greater than 0.5 with 95% confidence (eg, a median human probability of >0.75 in negative controls and <0.25 in positive controls). For negative controls, we subjected 5 essays known to be completely human generated, obtained from the authors' personal files, to the workflow described below. For positive controls, we submitted essay prompts to GPT 3.5 (OpenAI) and Claude 3 Sonnet (Anthropic Inc) through their respective online user interfaces. The 52 responses from each chatbot (104 total) were completely generated by the LLM, with no subsequent human editing. The prompts and resulting positive-control essays are available in Table S2 in Multimedia Appendix 2.

Statistical Analysis

We performed all analyses in R (version 4.5.1; R Foundation for Statistical Computing) using the *tidyverse* [10] framework and generated tables using *gtsummary* [11]. The code is available at [12].

We used Gaussian linear models for univariate analyses of application factors. To account for repeated essays in the application cycle analysis, we fit a Gaussian identity-link generalized estimating equation (GEE) containing application cycle, essay type, and their interaction, with robust sandwich SEs clustered by coded AMCAS identifier. The independence working correlation retained applicants represented by 1, 2, or 3 essays, and overall application cycle means were standardized with equal weight across the 3 essay types. We also fit an exchangeable-working-correlation GEE as a sensitivity analysis.

We also conducted 2 additional clustered multivariable sensitivity analyses of the 2023 to 2024 essays: one restricted to applicants classified as having a completed application using the application completion indicator (a nonmissing application submission date), and one restricted to applicants who submitted all 3 essay types. These models used the same GEE specification and covariates as the primary model, except that application completion was omitted because it was fixed in the completed application subset and nearly invariant in the all-3-essay subset. Full specifications are provided in Multimedia Appendix 3.

Ethical Considerations

The Human Research Protection Office at our institution determined that this study did not constitute human subjects research (202401024, issued on January 16, 2024). Moreover, permission was obtained from the Association of American Medical Colleges to use AMCAS data for this study. Privacy and confidentiality were maintained by labeling data only with a coded identifier; an honest broker who was not a study author held the key and removed other personally identifiable information. There was no compensation to participants.

Results

Selection of Essays for Analysis

In the 2021 to 2022 application cycle, the school received 6137 applications before the November 15, 2021, deadline; in the 2023 to 2024 application cycle, it received 5055 applications before the November 15, 2023, deadline. We analyzed 1000 selected "Personal Comments," 1000 "Failure," and 1000 "Anything Else" essays from each cycle, for a total of 3000 essays per cycle and 6000 overall. The selected essays represented 1184 applicants in the 2021 to 2022 application cycle and 1189 in the 2023 to 2024 application cycle (Table 1). Applicants contributed a mean of 2.53 (SD 0.75) and 2.52 (SD 0.76) selected essays, respectively; 819 (69.2%) and 818 (68.8%) contributed all 3 essay types.

XSL•FO
RenderX

Table 1. Characteristics of applicants represented in the essay sample.

Characteristics	2021 to 2022 (n=1184)	2023 to 2024 (n=1189)	*P* value[a]
Age (years), median (IQR)	22.6 (21.6-23.7)	22.7 (21.8-23.9)	*.03*
Self-reported gender, n (%)			*.002*
Female	632 (53)	592 (50)	
Male	550 (46)	582 (49)	
Other	2 (0.2)	15 (1.3)	
Program, n (%)			.17
Regular MD	1057 (89)	1040 (87)	
Combined MD-PhD	127 (11)	149 (13)	
Visa status, n (%)			.83
US citizen	1100 (93)	1102 (93)	
Other visa status	84 (7.1)	87 (7.3)	
American Medical College Application Service socioeconomic status, n (%)			.58
EO1 and/or EO2	159 (13)	169 (14)	
Other	1025 (87)	1020 (86)	
Grade point average, median (IQR)	3.89 (3.72-3.97)	3.91 (3.77-3.98)	*<.001*
Medical College Admission Test, median (IQR)	518 (514-521)	518 (514-521)	.96
Completed application, n (%)	1098 (93)	1105 (93)	.85
Invited to interview, n (%)	300 (25)	297 (25)	.84
Accepted, n (%)	100 (8.4)	93 (7.8)	.58
Essays included per student, mean (SD)	2.53 (0.75)	2.52 (0.76)	.78

[a]*P* values were calculated using the Wilcoxon rank-sum test or the Pearson chi-square test, as appropriate. Italics indicate $P \leq .05$.

We hypothesized that AI use would be rare or absent in the baseline year (2021-2022), leading the detector to produce P_{human} probabilities near 1.00. Conversely, in the first application cycle after the widespread availability of generative AI tools (2023-2024), we expected a proportion of students to have written some or all of their essays with these tools, leading to lower P_{human} estimates.

Labeling of Human-Generated Negative Controls and AI-Generated Positive Controls by GPTZero

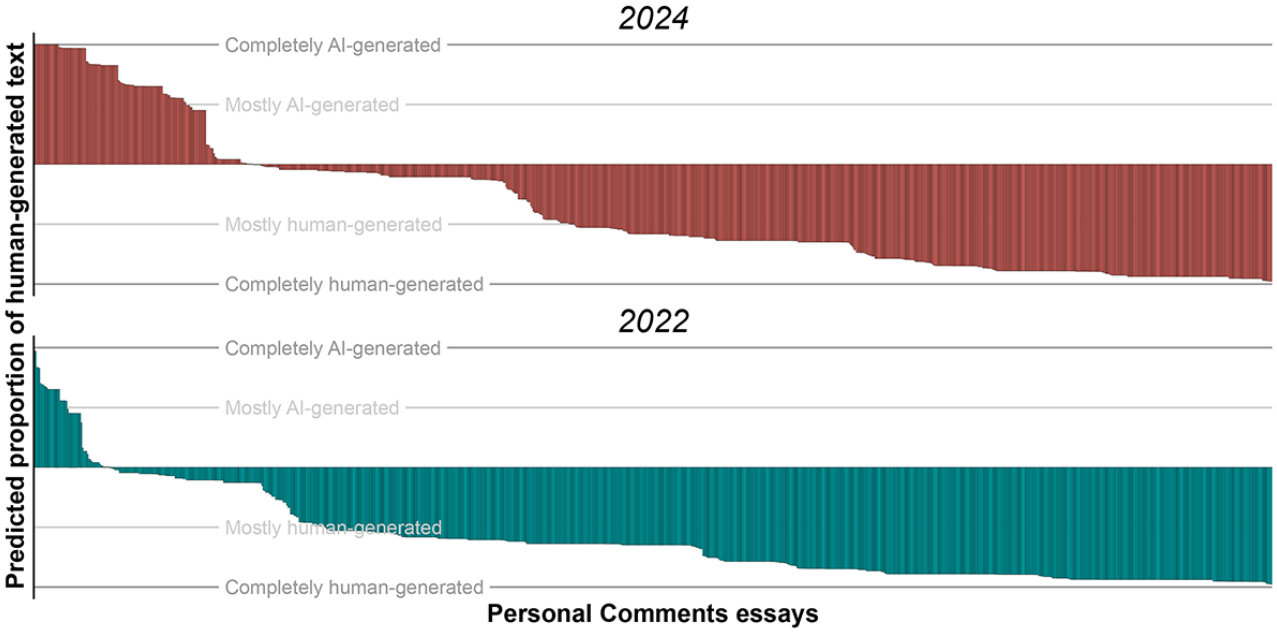
A series of negative controls known to be completely human generated demonstrated a median P_{human} of 0.93 (95% CI 0.86-1), with a representative essay yielding human, AI, and mixed scores of 0.945, 0.054, and 0.001, respectively. Conversely, a series of completely AI-generated positive controls, for which no downstream editing was performed, demonstrated a median P_{human} of 0.01 (95% CI 0-0.05), with a representative essay yielding human, AI, and mixed scores of 0.001, 0.997, and 0.002, respectively. These results met our continuation criteria, and we proceeded with our planned study.

Comparison of Estimated Human-Generated Probabilities Between the 2023-2024 and 2021-2022 Application Cycles

We compared P_{human} values for 1000 selected "Personal Comments" essays from the 2023 to 2024 application cycle to those from the pre-ChatGPT 2021 to 2022 application cycle (Figure 1). The median P_{human} was 0.77 in the 2023 to 2024 cycle (95% CI 0.76-0.78) compared to 0.83 in the 2021 to 2022 baseline (95% CI 0.82-0.85). Overall, 12.3% of essays had $P_{human}<0.5$ in the 2023 to 2024 cycle. In contrast, only 2.7% of essays had $P_{human}<0.5$ in the baseline cycle, significantly fewer than that in the first ChatGPT year (Fisher exact test: $P<.001$).

XSL•FO
RenderX

Figure 1. Waterfall plot of GPTZero-estimated probabilities that 1000 selected "Personal Comments" essays were AI or human generated. The 2024 application cycle (red, top) is compared with the pre-ChatGPT baseline 2022 cycle (tan, bottom).



Across all 3 essay types, the applicant-clustered GEE estimated a difference in P_{human} of −14.5% between the 2023 to 2024 and 2021 to 2022 application cycles, standardized with equal weight across essay types (robust SE 0.0099, 95% CI −16.4% to −12.5%; *P*<.001). In other words, after accounting for the correlation among essays from the same applicant, the probability that an essay was classified as fully human generated was 14.5 percentage points lower in the 2023 to 2024 cohort than in the pre-ChatGPT baseline.

Analysis of Factors Associated With GPTZero-Estimated Human-Generated Probabilities

We analyzed the 1000 selected "Personal Comments," "Failure," and "Anything Else" essays from the 2023 to 2024 cycle using GPTZero's AI detection algorithm. These 3000 essays represented 1189 unique applicants, as some applicants did not contribute every secondary essay type. We used the predicted probability that each essay was entirely human generated (P_{human}) as the outcome for univariate linear models and the primary applicant-clustered multivariable GEE. Predictors were essay type, application completion, age, self-reported gender, program type, visa status, socioeconomic status, GPA, and MCAT score.

In the univariate analysis (Table 2), the essays in the secondary application had lower P_{human} than the universal "Personal Comments" essay by an average of 5% to 9% (*P*<.001). For every 1-year increase in applicant age, there was an average of a 1% increase in P_{human} (*P*<.001). Essays submitted by applicants requiring a visa to study in the United States had P_{human} values that were 8% lower than those submitted by US citizens (*P*<.001). For each 0.1-point increase in GPA, P_{human} was 1% lower (*P*<.001). Incomplete applications had P_{human} values that were 9% higher than those of complete applications (*P*=.01). No significant differences were observed based on program type, self-reported gender, socioeconomic status, or MCAT score.

XSL•FO
RenderX

Table 2. Univariate analysis of application-related factors as predictors of P_{human}, the GPTZero-predicted probability that a document was fully human generated (N=3000).

Characteristics	Essays, n	Effect size[a]	95% CI	*P* value	q value[b]
Essay type					
Personal Comments	1000	—[c]	—	—	—
Failure	1000	0.91	0.88-0.93	<.001	*<.001*
Anything Else	1000	0.95	0.92-0.97	<.001	*<.001*
Age (years)	3000	1.01	1.01-1.01	<.001	*<.001*
Self-reported gender					
Female	1494	—	—	—	—
Male	1474	1.00	0.98-1.03	.73	.73
Other	32	1.11	0.99-1.24	.08	.11
Program					
Regular MD	2630	—	—	—	—
Combined MD-PhD	370	1.02	0.98-1.05	.40	.48
Visa status					
US citizen	2771	—	—	—	—
Other visa status	229	0.92	0.88-0.96	<.001	*<.001*
Medical College Admission Test	2998	1.00	1.00-1.00	.04	.06
Grade point average	3000	0.99	0.98-1.00	<.001	*<.001*
Application completion					
Completed	2911	—	—	—	—
Not completed	89	1.08	1.02-1.14	.01	*.03*

[a]Effect=e^{β} in the regression model, representing the average fold change in estimated P_{human} associated with a unit increment in each feature. For age, effect size gives the increase associated with a unit increase of 1 year; for Medical College Admission Test, a 1-point increase in 3-digit score; and for GPA, a 0.1-point increase in undergraduate GPA.

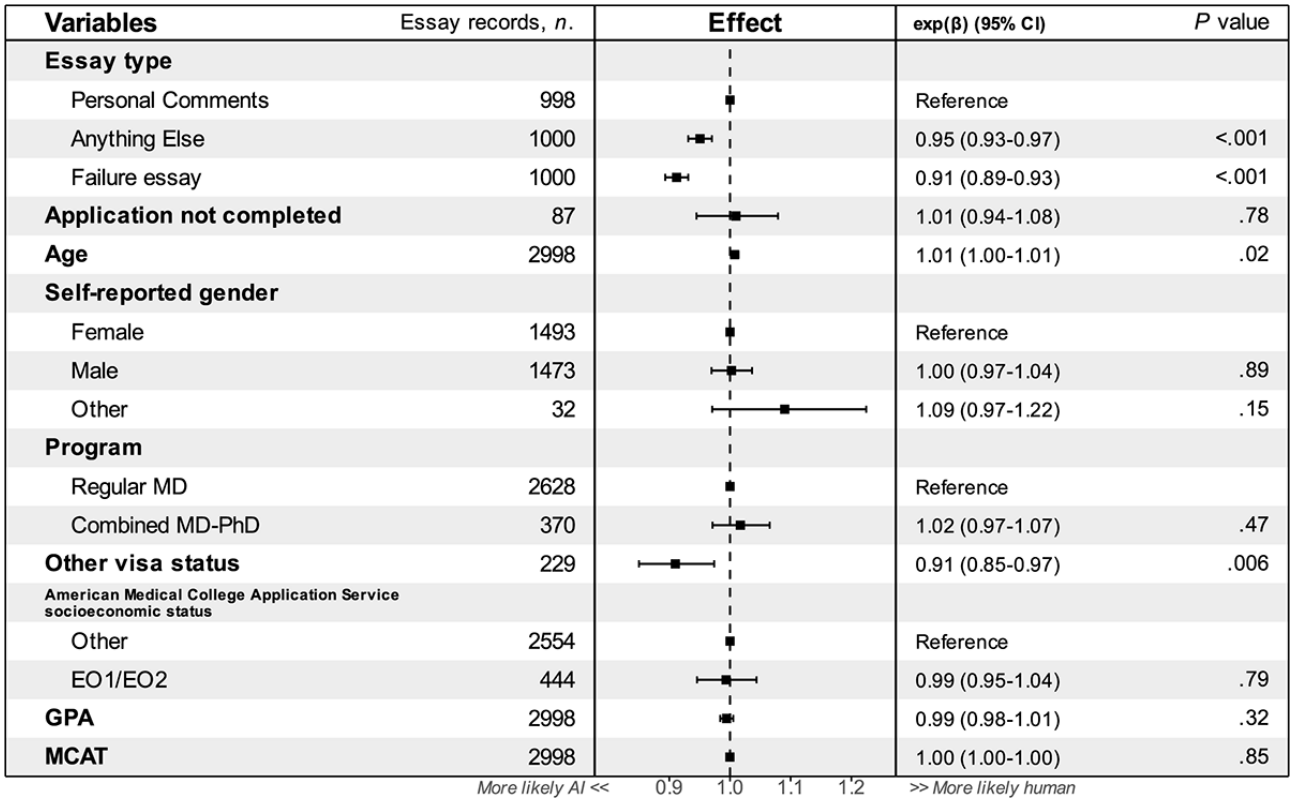
[b]False discovery rate–adjusted *P* values for multiple testing. Italics indicate $q \leq 0.05$.

[c]Not applicable.

The applicant-clustered multivariable GEE included 2998 essays from 1187 applicants. “Failure” essays (β=−.092, 95% CI −0.113 to −0.071; q<0.001) and “Anything Else” essays (β=−.051, 95% CI −0.071 to −0.03; q<0.001) had lower P_{human} values than “Personal Comments” essays (Figure 2). Older applicant age was associated with higher P_{human} (β=.0077 per year, 95% CI 0.0015-0.0139; q=0.043) and having other visa status was associated with lower P_{human} (β=−.094, 95% CI −0.162 to −0.027; q=0.024). Application completion, self-reported gender, program type, socioeconomic status, GPA, and MCAT score were not significant after false discovery rate correction.

XSL•FO
RenderX

Figure 2. Applicant-clustered multivariable analysis of application-related factors associated with GPTZero-predicted P_{human} in the 2023 to 2024 application cycle. Estimates were obtained from a Gaussian identity-link generalized estimating equation with an independence working correlation and robust sandwich SEs clustered by coded American Medical College Application Service identifier. Black squares and horizontal lines show exp(β) and robust 95% CIs, respectively; these transformed identity-link coefficients are not odds ratios or risk ratios. The displayed *P* values are unadjusted, whereas the conclusions in the text are based on Benjamini-Hochberg–adjusted q values. Age was modeled per 1-year increase, Medical College Admission Test (MCAT) per 1-point increase, and grade point average (GPA) per 0.1-point increase.



In the completed application sensitivity analysis (2911 essays from 1105 applicants), both secondary essay types, age, and visa status remained significant after false discovery rate correction (age q=0.033; visa status q=0.021; Figure S2 in Multimedia Appendix 3). In the analysis restricted to applicants represented by all 3 selected essay types (2454 essays from 818 applicants), both secondary essay effects remained significant (both q<0.001), whereas the age and visa effect estimates remained in the same direction but were no longer significant after correction (age q=0.055; visa status q=0.133; Figure S3 in Multimedia Appendix 3). Complete clustered estimates are reported in the Supplementary Methods and Results sections in Multimedia Appendix 3.

We hypothesized that students who applied late might have been more likely to use AI to meet the deadline. Moreover, we hypothesized that applicants declaring only majors in STEM might be less comfortable with written expression and more likely to use AI. However, in exploratory analyses (data not shown), we found no correlation between P_{human} and the date of application submission (r=0.01) or the declaration of a STEM major (P=.50).

Association Between Estimated AI Use and Admissions Outcomes

At the study school, application essays are used at every stage of review and could affect admissions decisions. We tested whether P_{human} was associated with interview invitation or medical school acceptance using adjusted applicant-level logistic regression models based on 1 “Personal Comments” essay per applicant (Table 3). Interview invitation was associated with male gender (odds ratio [OR] 0.51, 95% CI 0.37-0.72; q<0.001), EO1 and/or EO2 socioeconomic status (OR 3.31, 95% CI 2.08-5.29; q<0.001), MCAT score (OR 1.19 per point, 95% CI 1.14-1.24; q<0.001), and GPA (OR 1.81 per 0.1-point increase; q<0.001). Acceptance was associated with EO1 and/or EO2 socioeconomic status (OR 3.36; q<0.001), age (OR 1.17 per year; q=0.019), MCAT score (OR 1.18 per point; q<0.001), and GPA (OR 1.61 per 0.1-point increase, 95% CI 1.24-2.19 ; q=0.002). P_{human} was not associated with interview invitation (OR 1.61, 95% CI 0.87-3.02; q=0.2) or acceptance (OR 1.14, 95% CI 0.46-2.96; q=0.8).

XSL•FO
RenderX

Table 3. Adjusted applicant-level logistic regression analyses of factors associated with admissions outcomes in applicants with complete data from the 2023 to 2024 admissions cycle (N=815).

Characteristics	Interviewed applicants (n=142, 17%)			Accepted applicants (n=79, 10%)		
	OR[a] (95% CI)	*P* value	q value[b]	OR (95% CI)	*P* value	q value
P_{human}[c]	1.61 (0.87-3.02)	.13	0.19	1.14 (0.46-2.96)	.80	0.81
Age (years)	1.09 (1.00-1.18)	.04	0.061	1.17 (1.03-1.3)	.01	*0.019*
Self-reported gender						
Female	__[d]	—	—	—	—	—
Male	0.51 (0.37-0.72)	<.001	*<0.001*	0.59 (0.35-0.97)	.04	0.064
Program						
Regular MD	—	—	—	—	—	—
Combined MD-PhD	0.76 (0.45-1.25)	.30	0.30	1.79 (0.95-3.27)	.06	0.084
Visa status						
US citizen	—	—	—	—	—	—
Other visa status	1.95 (0.99-3.75)	.049	0.066	1.4 (0.5-3.41)	.50	0.58
American Medical College Application Service socioeconomic status						
Other	—	—	—	—	—	—
EO1 and/or EO2	3.31 (2.08-5.29)	<.001	*<0.001*	3.36 (1.8-6.15)	<.001	*<0.001*
Medical College Admission Test	1.19 (1.14-1.24)	<.001	*<0.001*	1.18 (1.11-1.26)	<.001	*<0.001*
Grade point average	1.81 (1.50-2.22)	<.001	*<0.001*	1.61 (1.24-2.19)	<.001	*0.002*

[a]OR: odds ratio. For age, the OR gives the change associated with a 1-year increase; for the Medical College Admission Test, a 1-point increase in the 3-digit score; and for grade point average, a 0.1-point increase in undergraduate grade point average.

[b]False discovery rate–adjusted *P* values for multiple testing. Italics indicate $q \leq 0.05$.

[c]P_{human} refers to GPTZero's determination for the "Personal Comments" essay.

[d]Not applicable.

Discussion

Principal Findings

We report our analysis of a natural experiment, comparing a cohort of medical school applicants who applied in 2021 to 2022, before LLM-driven chatbots were widely available, with the cohort who applied in 2023 to 2024, after the release of ChatGPT and GPT-3/4, when chatbots had become easily accessible, free, and frequently discussed in mainstream media and academic circles.

In the baseline year, 2.7% of essays had $P_{human}<0.5$ and therefore showed a style signal suggesting that AI was more likely than not to have been used in some way during preparation. Although some essays might have used earlier generative AI tools, this finding more likely reflects the population-wide lower limit of the detector due to false-positive determinations. Our major finding was a 9.6-percentage-point higher proportion of essays below this threshold in the 2023 to 2024 cycle. In the applicant-clustered multivariable GEE, lower P_{human} was independently associated with secondary essay type, younger age, and other visa status; GPA and application completion were not associated after false discovery rate correction. There was no evidence of an association with socioeconomic status or undergraduate major, and we did not find evidence that late applicants were more likely to have lower P_{human}. P_{human} was not significantly associated with interview invitations or acceptance decisions.

Strengths

Strengths of this paper include the large dataset, the use of multiple time points (bracketing the introduction of widely available generative AI), and the use of control inputs. Applicants to the study school represented approximately 10% of the 49,570 individuals who applied to US medical schools through AMCAS in 2023 to 2024. The 2021 to 2022 group had, at most, rare and sporadic access to AI and therefore provides a benchmark against which the 2023 to 2024 group can be compared. Additionally, the highly confident positive predictions for the positive controls and negative predictions for the negative controls support the validity of the methods.

An additional strength was the explicit modeling of within-applicant correlation among repeated essays. The similar essay-type estimates in the full, completed application and all-3-essay clustered models support the robustness of the primary essay-type findings to variation in an applicant's propensity to use generative AI for any essay, as well as to the unequal numbers of selected essays per applicant.

XSL•FO
RenderX

Limitations

Limitations of the paper include its single-site nature and the fact that the applicants to the study school have, on average, higher academic achievement than the AMCAS applicant pool as a whole and may otherwise be nonrepresentative. In our analysis of AI use in relation to application outcomes, we had information only on the admissions actions at the study school; some applicants rejected by the study school were undoubtedly accepted at other schools.

For practical reasons, we used only one AI detector, although several are available. GPTZero was chosen due to its wide adoption and the availability of an application programming interface. Moreover, the accuracy of a similar algorithm was 91% in a study that compared human-authored personal statements to known synthetic ones [13], and to be at least 96% in a larger benchmarking study of diverse writing samples [7]. Subsequently, new tools have been released with incrementally better performance and could be applied in future studies. There may also be other approaches, such as applicant surveys, to learn about AI use. These approaches could provide a direct assessment of AI use rather than an inferred assessment but would be limited by the honesty of the respondents.

A methodological limitation of this project is that the distinction between human-written and AI-written text is excessively dualistic. Although some applicants may blithely copy AI output directly into their application materials, it seems more plausible that they will adapt the AI text to their own situation, thus moving some or all sentences away from being purely AI generated. Such edited text could pass for human-written although AI would have played a part in producing it. For the present study, we assumed that at least some AI-derived text remains detectable after human editing. We used controls to understand how GPTZero scores known human and known AI essays, but we did not study its behavior in less dualistic scenarios (eg, asking AI to tweak known human essays, using humans to tweak AI essays, or manually blending human and AI inputs).

An important additional caveat is that AI detection is known to be imperfect [14,15]. The sensitivity and specificity of AI detectors can be tuned and generally are set so that the specificity is high at the expense of sensitivity. For example, in one small-scale study involving 50 writing samples, GPTZero was benchmarked at a sensitivity of 65% and a specificity of 90%, against an overall accuracy of 80% [14]. A larger study found a false-negative rate of between 0.2% and 3%, depending on the model used to generate the input text, and a false-positive rate of less than 1% [7]. The tradeoff is biased toward higher specificity because the consequences of a false-positive error (false detection of AI use, potentially leading to wrongful accusations of academic malfeasance or policy noncompliance) are less palatable than the consequences of a false-negative error. The result, however, is that some AI-generated text will be classified as human generated. Moreover, AI detectors may be more effective in identifying earlier and less advanced iterations of AI chatbots [9], whereas newer models generate more humanlike output and, in some cases have been specifically designed to evade detection. Prompt engineering can be used to direct chatbots to write more like humans, and the resulting text is less readily detected [16,17]. As some chatbots are marketed on a freemium model (free tier or paid tier), students with greater financial resources may have access to more sophisticated versions. Anecdotally, some types of inputs, such as lists, may be erroneously flagged as AI generated. This failure mode could lead to inaccurate detection in medical school application materials, which sometimes include lists of students' activities or publications. Together, these factors could lead to unjust outcomes if AI detection were deployed with real-world consequences (eg, disqualifying applicants for detected use of AI).

AI detectors may misclassify the authentic work of nonnative English writers as AI generated [16] due to their more restricted vocabulary and syntax. Indeed, we found that applicants requiring a visa to study in the United States had a lower probability of human-generated text, which may reflect higher AI use but could also reflect this failure mode of AI detection. English-language learners may also be more likely to use AI to proofread and correct authentic human-written essays. We do not know how AI-assisted "proofreading" affects P_{human}, compared with outright composition of writing samples by AI, for which we have benchmarked P_{human}. Because of these factors, reliance on AI detection could unintentionally penalize English-language learners.

As a final caveat, the importance placed on application essays may vary from one school to another. The impact of AI use on admissions outcomes would vary accordingly. There may be variation between schools in the propensity of their respective applicant pools to use AI. The structure of the application (including the number and length of essays and the time permitted to complete the secondary application) could be an underlying factor. Our findings suggest that lengthy secondary applications are more likely to be completed with AI assistance. Students applying to more or less selective schools might be more or less likely to use AI. The magnitude and even the direction of this effect are not known.

Comparison With Prior Work

To the best of our knowledge, this is the first real-world study of generative AI use by medical school applicants. A recent study studied a smaller set of 32 essays in the style of medical school applicant essays, written explicitly for the purposes of the study using varying degrees of AI assistance. The authors found no association between AI use and raters' subjective assessment of these essays [18], suggesting that applicants gain little advantage through AI use, similar to our findings.

Another study by Vaccaro et al [13] compared a baseline set of essays written before the advent of generative chatbots to a second set of essays generated using AI. Human readers were unable to distinguish between human-written and AI-written essays and assigned marginally higher scores to the AI-written essays. The study by Vaccaro et al [13] provides validity evidence for our use of linguistic-pattern–based AI detection. The authors concluded that AI use threatens the use of personal statements in admissions, whereas our assessment is more sanguine given the lack of an impact on admissions outcomes.

XSL•FO
RenderX

There is evidence that writers who are not medical school applicants are using generative AI as well. The editors of the journal *Organization Science* used a methodology similar to ours, but with a different AI detector, and found evidence of a marked increase in AI-written submissions to their journal, accompanied by a decrease in quality, after the public launch of ChatGPT in 2022 [19]. Although we found that AI use conferred no admissions advantage (or disadvantage), the authors found that AI-written papers were more likely to be rejected. The authors concluded that AI-written submissions are placing stress on the peer review ecosystem. It remains to be seen what stressors AI will place on medical school admissions.

Conclusions

Our results suggest that approximately 10% of medical school application essays were written with the help of generative AI in the first application cycle after these tools became widely available. Based on data from a single medical school, there appeared to be no correlation between AI use and admissions outcomes.

Given the difficulty of definitively identifying AI use in applicant essays and the lack of clarity regarding the appropriate response, we recommend continuing to study such data only in aggregate and on an informational basis. If there were evidence to suggest widespread AI use in answering a specific item, for example, the appropriate institutional response might be to design a new item that is less amenable to AI rather than to penalize applicants who appear to be providing AI-generated answers. Similarly, we do not recommend that schools adopt software that would automatically flag AI use at the level of an individual applicant, although such software is likely to become available in the future. Playing a cat-and-mouse game around AI can only induce cynicism and erode overall trust in the application process, with the potential for unjust outcomes or the exacerbation of disparities.

It is unlikely that AI use can or should be entirely eliminated from medical school applications. Indeed, the 2025 AMCAS application contains a revised certification statement in which the applicant must agree that although AI use is permitted, the final product must be “a true reflection of [their] own work and represents [their] experiences” [20]. Applicants’ reliance on AI could be a symptom of the heavy cognitive burden associated with preparing a complete medical school application. As secondary applications showed more evidence of AI use than the common AMCAS application, schools should consider whether there is a benefit in including additional writing samples in their school-specific secondary applications. These one-off tasks are numerous and are completed on a shorter timeline than the AMCAS essay, which could put applicants under pressure to take shortcuts. In turn, prehealth advisers should inform applicants that using AI deprives them of the opportunity to tell their own story and highlight the unique contribution they will make to medicine.

Acknowledgments

The authors thank Christina Twist for serving as the honest broker for data retrieval. Generative AI tools (Claude Code [Anthropic Inc], Codex [OpenAI], and GPT-5.4 [OpenAI]) were used to develop the analysis code and provide copyediting assistance for the manuscript. The authors assume full responsibility for the content of this manuscript.

Data Availability

Due to the nature of the study, the data are not directly available but may be made available upon reasonable request.

Funding

The study was funded using the senior author’s discretionary departmental funds.

Authors' Contributions

Conceptualization: NCS, VSR, ISH
Data curation: ISH, NCS
Formal analysis: ISH, NCS
Funding acquisition: VSR
Methodology: ISH, NCS
Project administration: ISH, NCS
Resources: ISH, VSR
Software: NCS
Supervision: VSR, ISH
Visualization: NCS
Writing—original draft: NCS
Writing—review and editing: NCS, VSR, ISH

Conflicts of Interest

None declared.

XSL•FO
RenderX

Multimedia Appendix 1
Classification of undergraduate major subjects as science, technology, engineering, and mathematics (STEM) vs non-STEM.
[DOCX File , 100 KB - mededu_v12i1e96673_app1.docx]

Multimedia Appendix 2
Positive-control essays generated using AI chatbots.
[DOCX File , 71 KB - mededu_v12i1e96673_app2.docx]

Multimedia Appendix 3
Supplemental methods and results.
[DOCX File , 1014 KB - mededu_v12i1e96673_app3.docx]

References

1. Singhal K, Tu T, Gottweis J, Sayres R, Wulczyn E, Amin M, et al. Toward expert-level medical question answering with large language models. Nat Med 2025 Mar;31(3):943-950. [doi: 10.1038/s41591-024-03423-7] [Medline: 39779926]
2. Qiu J, Xiong D. Generating highly relevant questions. arXiv. Preprint posted online on October 8, 2019 [FREE Full text] [doi: 10.48550/arXiv.1910.03401]
3. Li J, Wang S, Zhang M, Li W, Lai Y, Kang X, et al. Agent hospital: a simulacrum of hospital with evolvable medical agents. arXiv. Preprint posted online on May 5, 2024 [FREE Full text] [doi: 10.48550/ARXIV.2405.02957]
4. Cotton DR, Cotton PA, Shipway JR. Chatting and cheating: ensuring academic integrity in the era of ChatGPT. Innov Educ Teach Int 2023 Mar 13;61(2):228-239. [doi: 10.1080/14703297.2023.2190148]
5. 2024 AMCAS® application workbook. American Medical College Application Service. 2024. URL: https://web.archive.org/web/20230521071233/https://students-residents.aamc.org/media/14376/download [accessed 2026-08-05]
6. GPTZero's AI detection technology. GPTZero. URL: https://gptzero.me/technology [accessed 2024-06-17]
7. Jabarian B, Imas A. Artificial writing and automated detection. National Bureau of Economic Research. 2025. URL: https://www.nber.org/papers/w34223 [accessed 2026-08-05]
8. Document predictions. GPTZero API. URL: https://gptzero.stoplight.io/docs/gptzero-api/707838f7e089d-document-predictions [accessed 2026-08-04]
9. Elkhatat AM, Elsaid K, Almeer S. Evaluating the efficacy of AI content detection tools in differentiating between human and AI-generated text. Int J Educ Integr 2023 Sep 01;19:17. [doi: 10.1007/s40979-023-00140-5]
10. Wickham H, Averick M, Bryan J, Chang W, D'Agostino LM, François R, et al. Welcome to the Tidyverse. J Open Source Software 2019 Nov 21;4(43):1-6 [FREE Full text] [doi: 10.21105/joss.01686]
11. Sjoberg DD, Whiting K, Curry M, Lavery JA, Larmarange J. Reproducible Summary Tables with the gtsummary Package. The R Journal 2021 Jun 21;13:570-580 [FREE Full text] [doi: 10.32614/RJ-2021-053]
12. nspies13 / llm_use_in_medical_school_applications. GitHub. URL: https://github.com/nspies13/llm_use_in_medical_school_applications [accessed 2026-08-17]
13. Vaccaro MJ, Sharma I, Espina-Rey AP, Lyman N, Palacios C, Zhang Y, et al. Death of the personal statement: qualitative comparison between human-authored and artificial intelligence-generated medical school admissions essays. J Am Coll Surg 2026 Jan 01;242(1):47-52. [doi: 10.1097/XCS.0000000000001602] [Medline: 41051105]
14. Habibzadeh F. GPTZero performance in identifying artificial intelligence-generated medical texts: a preliminary study. J Korean Med Sci 2023 Sep 25;38(38):e319 [FREE Full text] [doi: 10.3346/jkms.2023.38.e319] [Medline: 37750374]
15. Rashidi HH, Fennell BD, Albahra S, Hu B, Gorbett T. The ChatGPT conundrum: human-generated scientific manuscripts misidentified as AI creations by AI text detection tool. J Pathol Inform 2023 Oct 17;14:100342 [FREE Full text] [doi: 10.1016/j.jpi.2023.100342] [Medline: 38116171]
16. Liang W, Yuksekgonul M, Mao Y, Wu E, Zou J. GPT detectors are biased against non-native English writers. Patterns (N Y) 2023 Jul 10;4(7):100779 [FREE Full text] [doi: 10.1016/j.patter.2023.100779] [Medline: 37521038]
17. Perkins M, Roe J, Vu BH, Postma D, Hickerson D, McGaughran J, et al. Simple techniques to bypass GenAI text detectors: implications for inclusive education. Int J Educ Technol High Educ 2024 Sep 09;21(1) [FREE Full text] [doi: 10.1186/s41239-024-00487-w]
18. Choi J, Zhao J, Ngo TA, Grierson L. Investigating the threat of AI to undergraduate medical school admissions: a study of its potential impact on the rating of applicant essays. Can Med Educ J 2025 Dec 22;16(6):15-20. [doi: 10.36834/cmej.79690] [Medline: 41584942]
19. Gartenberg C, Hasan S, Murray A, Pierce L. More versus better: artificial intelligence, incentives, and the emerging crisis in peer review. Organ Sci 2026 May;37(3):795-812. [doi: 10.1287/orsc.2026.ed.v37.n3]
20. 2025 AMCAS® application workbook. Association of American Medical Colleges. 2024. URL: https://web.archive.org/web/20241123001043/https://students-residents.aamc.org/media/14376/download [accessed 2026-08-02]

XSL•FO
RenderX

Abbreviations

AMCAS: American Medical College Application Service
AUROC: area under the receiver operating characteristic
GEE: generalized estimating equation
GPA: grade point average
LLM: large language model
MCAT: Medical College Admission Test
OR: odds ratio
STEM: science, technology, engineering, and mathematics

Edited by B Lesselroth; submitted 30.Mar.2026; peer-reviewed by R Kravitz, M Chakit; comments to author 20.May.2026; revised version received 21.Jul.2026; accepted 28.Jul.2026; published 03.Sep.2026.

Please cite as:
Spies NC, Ratts VS, Hagemann IS
Estimating the Prevalence of Generative AI Use in Medical School Application Essays: Cross-Sectional Study
JMIR Med Educ 2026;12:e96673
URL: https://mededu.jmir.org/2026/1/e96673
doi:10.2196/96673
PMID:

©Nicholas C Spies, Valerie S Ratts, Ian S Hagemann. Originally published in JMIR Medical Education (https://mededu.jmir.org), 03.Sep.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Education, is properly cited. The complete bibliographic information, a link to the original publication on https://mededu.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Large Language Model Performance on Multistep Clinical Cases: Comparative Study Across Question and Case Levels

Jiaxue Cha[1,2], PhD; Yan Zhao[3], MS; Hui Zong[1], PhD

[1]Centre for Medical Big Data and Artificial Intelligence, The First Affiliated Hospital (Southwest Hospital) of Army Medical University (Third Military Medical University), Chongqing, China

[2]School of Life Sciences and Technology, Tongji University, Shanghai, China

[3]Department of Prevention Healthcare, The First Affiliated Hospital (Southwest Hospital) of Army Medical University (Third Military Medical University), Chongqing, China

Corresponding Author:
Hui Zong, PhD
Centre for Medical Big Data and Artificial Intelligence
The First Affiliated Hospital (Southwest Hospital) of Army Medical University (Third Military Medical University)
No. 30 Gaotanyan Main Street
Chongqing, 400038
China
Phone: 86 23 68766993
Email: zonghui0228@163.com

Abstract

Background: Most large language models (LLMs) have achieved passing scores on medical licensing examinations. However, most evaluations focus on single-question accuracy, overlooking performance on multistep patient management scenarios, such as making a diagnosis followed by a treatment plan. It is unclear if LLMs can maintain high performance on complete clinical cases.

Objective: This study aimed to evaluate the performance of LLMs on multistep clinical cases and to investigate the impact of model size scaling and case complexity on performance stability.

Methods: We curated a dataset of 189 unique clinical cases comprising 473 individual questions from the Chinese National Medical Licensing Examination. The number of questions related to each case ranged from 2 to 4. A physician with more than 8 years of clinical experience annotated the cases, confirming that 82.5% (156/189) contained sequentially dependent questions. We tested 4 representative LLMs: DeepSeek-R1, GPT-4o, Gemini-2.5-Flash, and Qwen2.5. We measured 2 metrics: the question pass rate (QPR) and the case pass rate (CPR), where a case was considered correct only if the model answered all its questions correctly. The consistency gap was calculated as the difference between QPR and CPR. Additionally, we tested Qwen2.5 at different parameter sizes (3B, 7B, 14B, and 32B) to examine the effect of model size. We also calculated the expected CPR and used the McNemar test with Bonferroni correction to compare model performance.

Results: All LLMs achieved a QPR exceeding 83%. However, the CPR was lower for all models. DeepSeek-R1 achieved the highest QPR at 89.9% (425/473), with a CPR of 79.9% (151/189), corresponding to the smallest consistency gap at 10%. In contrast, GPT-4o exhibited a QPR of 83.5% (395/473) and a CPR of 65.6% (124/189), resulting in the largest consistency gap at 17.9%. DeepSeek-R1 and Qwen2.5-32B significantly outperformed GPT-4o in both QPR and CPR (P<.05). Observed and expected CPR values were closely aligned across models. Increasing the model size from 3B to 32B reduced the consistency gap by approximately 50%. As the number of questions per case increased, the consistency gap tended to increase across all evaluated LLMs.

Conclusions: Current LLMs exhibit high accuracy in answering individual questions, but their ability to correctly answer all questions within a complete clinical case is substantially lower. This study design evaluated independent question answering, with responses aggregated at the case level. The performance depends heavily on model scale and case complexity. LLMs should be used as support tools rather than independent decision-makers in medical education and clinical practice.

(JMIR Med Educ 2026;12:e95342) doi:10.2196/95342

XSL•FO
RenderX

KEYWORDS

large language models; clinical cases; medical education; ChatGPT; DeepSeek

Introduction

Medical licensing examinations serve as the gold standard for assessing the professional competence of physicians, ensuring that they possess the requisite knowledge for safe clinical practice [1]. With the rapid evolution of AI, large language models (LLMs) have demonstrated remarkable capability in these standardized assessments [2]. Recent benchmarks indicate that state-of-the-art models, such as GPT-4 and other related LLMs, have achieved passing scores on the United States Medical Licensing Examination and the Chinese National Medical Licensing Examination (CNMLE), in some cases exceeding the performance of average human examinees [3-5]. These achievements have prompted growing interest in the potential application of LLMs as clinical decision support systems [6]. However, the current evaluation methodologies largely conceptualize clinical questions as isolated data points, focusing primarily on the accuracy of answering individual medical questions rather than performance on complete clinical cases.

In real-world clinical practice, physicians do not answer independent questions in isolation. They need to deal with complex, multistep clinical scenarios. A typical clinical encounter requires a logically structured sequence of decisions, progressing from symptom analysis and diagnostic testing to treatment planning and long-term follow-up [7]. In this context, accuracy metrics calculated solely on a per-question basis can be systematically overestimated [8]. For instance, a model might correctly diagnose a patient based on symptoms (question A) but subsequently recommend a clinically inappropriate or contraindicated treatment (question B) for the same patient. While a traditional per-item accuracy metric might score this as 50% accuracy, from a clinical perspective, this inconsistency renders the model unreliable and potentially unusable [9]. Therefore, the lack of assessment metrics focusing on the case-level performance, where success is defined by the correct resolution of the entire clinical trajectory, represents a significant methodological limitation in current AI medical research.

Prior studies have reported rapid improvements in performance of LLMs on the CNMLE, primarily focusing on question-level accuracy metrics. A synthesis of performance data from 8 independent studies published between 2023 and 2025 [10-17], collectively covering 28 distinct evaluation records on official CNMLE datasets (Multimedia Appendix 1), revealed a clear temporal pattern. Early-generation models, such as GPT-3.5, consistently failed to meet the passing threshold (60%), with accuracy scores ranging from 36.5% to 54.7%. In contrast, more recent models evaluated in 2024 and 2025, including GPT-4o, DeepSeek-R1, and Qwen3, generally achieved substantially higher performance, with several evaluations exceeding 90% accuracy. However, it is important to note that these published results predominantly rely on question-level accuracy measures, leaving the model performance on multistep clinical cases largely unexplored.

To address this gap between standardized testing scores and performance on multistep clinical tasks, this study uses multistep case-based questions from the CNMLE to systematically assess whether high question-level accuracy translates to success on complete clinical cases. We selected 4 representative models, including GPT-4o, Gemini-2.5-Flash, DeepSeek-R1, and Qwen2.5, to represent both internationally developed and domestically developed state-of-the-art architectures. Beyond comparing question-level pass rates with case-level pass rates, we further investigate the impact of model scaling on performance by analyzing the Qwen2.5 series across parameter sizes ranging from 3B to 32B. By shifting the focus from isolated accuracy measures to case-level outcome validity, this study aims to provide a more clinically grounded evaluation of the readiness of LLMs for integration into complex clinical workflows.

Methods

Data Collection

We collected a raw dataset from the CNMLE, covering a 5-year period from 2017 to 2021 [12]. The initial corpus consisted of 3000 multiple-choice questions (600 questions per year). These examinations are designed to assess the clinical competency of physicians and include both independent knowledge-based questions and case-based questions. To evaluate performance on complete clinical cases comprising multiple questions, our study focused exclusively on sequential case-based questions. These questions present a shared clinical vignette (eg, demographic characteristics, medical history, and symptoms), followed by a series of questions (ranging from 2 to 4 questions per case) that collectively cover different stages of clinical management (eg, from diagnosis to treatment).

We applied a rigorous 3-step screening process. First, we manually extracted all sequential case-based question sets from the raw dataset, resulting in 196 clinical cases comprising 494 individual questions. Second, we identified and removed 1 duplicate case (containing 4 questions) that appeared across different examination years, reducing the dataset to 195 cases with 490 questions. Third, to ensure the validity of the clinical context, we excluded cases lacking specific patient demographic or clinical information. This step removed 6 cases with 17 questions.

To clarify the extent to which subquestions within each case are interdependent, we invited a physician with more than 8 years of clinical experience to annotate all 189 cases. Each case was classified as having sequentially dependent questions if later questions (eg, about examinations or treatments) directly referred to the clinical context or decisions established in earlier questions (eg, a specific diagnosis). On the basis of this annotation, 82.5% (156/189) of cases were identified as containing such dependent question sequences. Representative examples of cases containing 2, 3, and 4 questions are presented in Multimedia Appendix 2.

XSL•FO
RenderX

LLMs

To ensure a comprehensive evaluation of generative AI capabilities, we selected 4 representative LLMs that reflect the state of the art in both international and Chinese technological landscapes. The study included GPT-4o (OpenAI; released in May 2024) [18], Gemini-2.5-Flash (Google; released in June 2025) [19], DeepSeek-R1 (DeepSeek-AI Team; released in January 2025) [20], and Qwen2.5 (Alibaba Cloud; released in September 2024) [21]. These models were chosen for their widespread adoption and advanced linguistic proficiency.

Furthermore, to investigate the "scaling law," specifically, how model parameter size influences case-level performance stability, we used the Qwen2.5 family across 4 distinct sizes: 3B, 7B, 14B, and 32B. This gradation allows a granular analysis of the relationship between computational scale and performance consistency across sequential questions.

The deployment methods varied according to model accessibility and architecture. We accessed GPT-4o, Gemini-2.5-Flash, and DeepSeek-R1 via their official APIs, whereas the Qwen2.5 series models were deployed locally using the Ollama framework. The specific API model identifiers were GPT-4o (gpt-4o), Gemini-2.5-Flash (gemini-2.5-flash), and DeepSeek-R1 (deepseek-reasoner). Qwen2.5 models (3B, 7B, 14B, and 32B parameters) were deployed locally using Ollama with the corresponding model tags (qwen2.5:3b, qwen2.5:7b, qwen2.5:14b, and qwen2.5:32b). The models were run using the default Ollama quantization configuration (Q4_K_M).

To rigorously test the models' inherent knowledge and question-answering abilities without providing specific examples, a zero-shot prompting strategy was used for all queries. A unified prompt template was applied across all models (Multimedia Appendix 3). No manual tuning of decoding parameters was performed. Parameters were kept at the default settings provided by each model provider or inference framework. Each question was evaluated once. For clinical cases with multiple questions, individual questions were submitted independently in separate conversations without retaining conversation history from previous questions. These experiments were conducted in September 2025.

For the computational infrastructure, local inference was performed on a server running Ubuntu 22.04, using a hardware configuration that included an NVIDIA RTX 4090 (24 GB) graphics processing unit (GPU) and an NVIDIA Virtual 48 GB GPU to accommodate varying memory requirements. The software environment leveraged PyTorch (version 2.1), Ollama (version 0.9.0), and Transformers (version 4.25.4) to manage model loading and execution. The experimental workflows were managed through Jupyter Lab, and all model outputs were automatically logged into an SQL database for subsequent postprocessing and analysis.

Metrics Definition

To evaluate the utility of LLMs in multistep clinical cases, we established a multidimensional metric framework that went beyond simple accuracy. We first calculated the question pass rate (QPR), which serves as the traditional baseline metric representing global accuracy. This indicator is calculated as follows: QPR=total number of correctly answered questions/total number of questions.

While QPR reflects general knowledge retrieval, it often fails to capture the holistic performance required for complex clinical scenarios where a single patient case involves multiple interdependent questions. To address this limitation, we introduced the case pass rate (CPR) as a stricter case-level accuracy metric. In this metric, a clinical case is considered successfully solved only if the model answers all subquestions associated with that specific case correctly; partial correctness is treated as a failure. Accordingly, CPR is calculated as follows: CPR=number of cases with all subquestions correct/total number of cases.

We estimated an expected CPR under the assumption that questions within a case are conditionally independent, using a stratified accuracy approach that does not rely on the correctness pattern of any individual case. Specifically, we first grouped all clinical cases by their length, that is, cases containing 2, 3, or 4 questions. For each length group, we calculated the overall question accuracy across all cases within that group: P_k=correct questions in group k/total questions in group k.

This group-level accuracy reflects the model's average performance on questions from cases of a given length, and crucially, it is estimated independently of the correctness pattern of any specific case. For any individual case, we then used the accuracy of its corresponding length group, rather than the case's own accuracy, to estimate the expected probability that all questions in that case would be answered correctly under independence. The overall expected CPR was then obtained by averaging these case-specific expected probabilities across all 189 cases:



This stratified approach provides a valid independence-based benchmark for comparing against the observed CPR, free from the circularity of using a case's own outcomes to predict its own performance.

Finally, to quantify the disparity between question-level and case-level performance, we computed the consistency gap. This metric represents the arithmetic difference between the QPR and CPR, defined as follows: consistency gap=QPR–CPR.

A smaller gap indicates better preservation of performance when moving from isolated questions to complete cases, whereas a larger gap suggests that the model's ability to answer individual questions does not reliably translate to success on multistep case-level tasks. It is important to note that this gap is a descriptive measure of performance degradation from question-level to case-level aggregation, rather than a direct indicator of how consistently a model performs across dependent questions.

Statistical Analysis

To assess whether the observed differences in QPR and CPR between models were statistically significant, we performed pairwise comparisons using the McNemar test. This test is appropriate for paired binary outcomes, as each question or case

XSL•FO
RenderX

was evaluated by all models under identical conditions. For QPR comparisons, each of the 473 questions served as a paired unit; for CPR comparisons, each of the 189 cases served as a paired unit. Given that 6 pairwise comparisons were made across the 4 main models (DeepSeek-R1, Qwen2.5-32B, Gemini-2.5-Flash, and GPT-4o), we applied Bonferroni correction, with statistical significance defined as $P<.05$. For the Qwen2.5 scaling analysis, 6 pairwise comparisons were also made across the 4 parameter sizes (3B, 7B, 14B, and 32B), with the same significance threshold. All statistical analyses were performed using Python (version 3.13.13; Python Software Foundation) with the *statsmodels* library (version 0.14.6).

Ethical Considerations

All data used in this study were obtained from publicly accessible sources. The research design did not include any intervention involving human participants or animals, and no sensitive or private information was collected or processed. Therefore, according to the ethical guidelines of the First Affiliated Hospital of Army Medical University, this study did not require ethical review or approval from the institutional ethics committee.

Results

Dataset Characteristics

The final curated dataset consisted of 189 unique clinical cases comprising 473 individual questions (Table 1). The structure of the 189 cases varied in complexity: 53.4% (n=101) cases contained 2 questions, 42.9% (n=81) cases contained 3 questions, and 3.7% (n=7) cases contained 4 questions. The dataset covers a diverse range of patient demographics representative of general medical practice. Among the patients described in the case vignettes, 48.7% (92/189) were male and 49.2% (93/189) were female. The minimum age of patients was 3 days, and the maximum age was 75 years. The clinical fields encompassed internal medicine, surgery, pediatrics, obstetrics and gynecology, and other specialties.

Table 1. Composition and characteristics of the curated clinical case dataset derived from the Chinese National Medical Licensing Examination (N=189).

Feature	Patients
Total questions, n	473
Case structure (number of questions per case), n (%)	
2	101 (53.4)
3	81 (42.9)
4	7 (3.7)
Patient sex, n (%)	
Male	92 (48.7)
Female	93 (49.2)
Unspecified	4 (2.1)
Patient age, range	3 days to 75 years

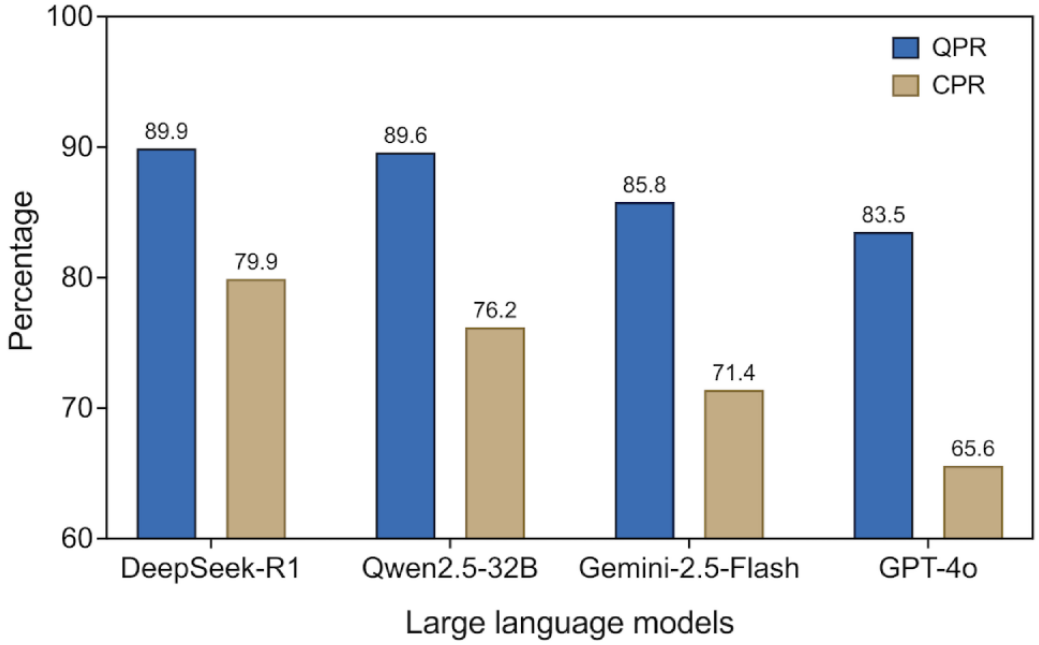
Overall Performance

All 4 LLMs achieved QPR above 83% (Figure 1). DeepSeek-R1 showed the highest QPR at 89.9% (425/473), closely followed by Qwen2.5-32B at 89.6% (424/473), with no significant difference between them. DeepSeek-R1 significantly outperformed GPT-4o (395/473, 83.5%; $P=.001$) and Gemini-2.5-Flash (406/473, 85.8%; $P=.02$). Qwen2.5-32B also significantly outperformed GPT-4o ($P=.003$). Differences between Qwen2.5-32B and Gemini-2.5-Flash, as well as between Gemini-2.5-Flash and GPT-4o, were not significant. At the case level, DeepSeek-R1 achieved the highest CPR of 79.9% (151/189), significantly higher than GPT-4o (124/189, 65.6%; $P<.001$) and Gemini-2.5-Flash (135/189, 71.4%; $P=.02$). Qwen2.5-32B (144/189, 76.2%) also significantly outperformed GPT-4o (124/189, 65.6%; $P=.03$). No significant differences were found between DeepSeek-R1 and Qwen2.5-32B, Qwen2.5-32B and Gemini-2.5-Flash, or Gemini-2.5-Flash and GPT-4o. The consistency gaps (QPR minus CPR) were 10% for DeepSeek-R1, 13.5% for Qwen2.5-32B, 14.4% for Gemini-2.5-Flash, and 17.9% for GPT-4o. These results show that although all models performed well on individual questions, their performance drops consistently when handling complete cases, with GPT-4o showing the largest decline.

We further compared the observed CPR with the expected CPR for each model. The expected CPRs were 76.6% for DeepSeek-R1, 76.3% for Qwen2.5-32B, 68.6% for Gemini-2.5-Flash, and 64.4% for GPT-4o. When compared with the observed CPRs (DeepSeek-R1: 151/189, 79.9%; Qwen2.5-32B: 144/189, 76.2%; Gemini-2.5-Flash: 135/189, 71.4%; and GPT-4o: 124/189, 65.6%), the observed and expected values were closely aligned.

XSL•FO
RenderX

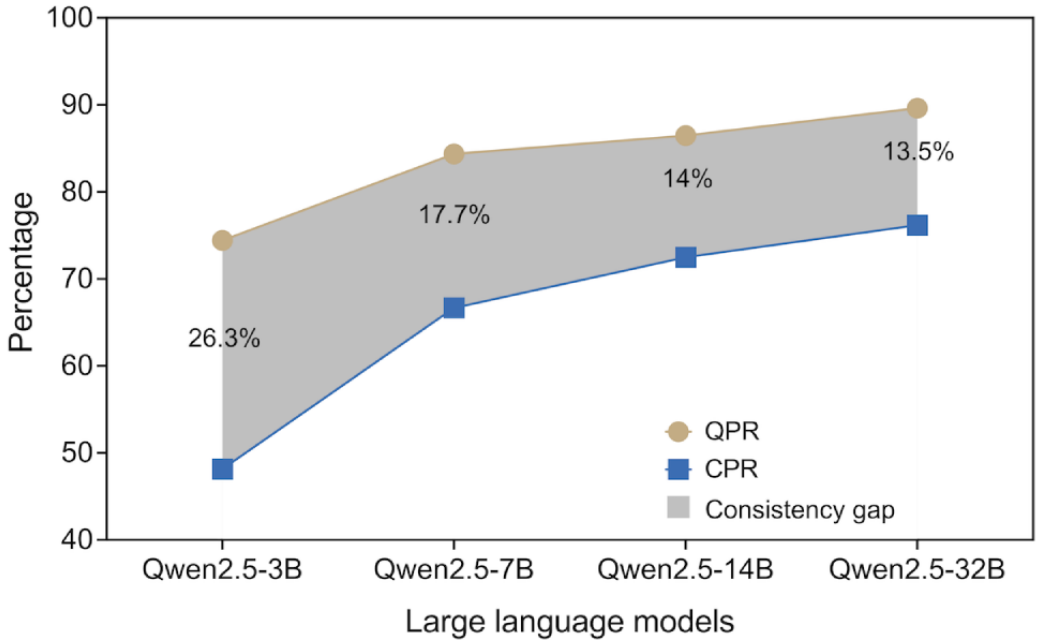
Figure 1. Comparative analysis of overall performance metrics across 4 large language models. The grouped bar chart displays the question pass rate (QPR) and case pass rate (CPR) for DeepSeek-R1, Qwen2.5-32B, Gemini-2.5-Flash, and GPT-4o. While all models achieved a high QPR exceeding 83%, a consistent performance degradation was observed in CPR, highlighting the challenge between isolated question answering and complete case-level clinical tasks.



Performance Impact of Model Scaling

We observed a clear positive correlation between model parameter size and overall performance within the Qwen2.5 family. As illustrated in Figure 2, increasing the model size from 3B to 32B resulted in a consistent improvement in both metrics, with QPR increasing from 74.4% (352/473) to 89.6% (424/473) and CPR increasing from 48.1% (91/189) to 76.2% (144/189). All comparisons between the 3B model and larger models (7B, 14B, and 32B) showed significant differences (all $P<.001$). The 32B model also significantly outperformed the 7B model in QPR ($P=.02$), but differences between 7B vs 14B and 14B vs 32B in CPR were not significant. More importantly, the consistency gap showed a substantial decline as model size increased. The gap reduced by approximately 50%, from 26.3% in the 3B model to 13.5% in the 32B model. This pattern suggests that while smaller models may correctly answer individual questions, larger models are more likely to achieve success on complete patient cases.

Figure 2. Impact of model scaling on performance across question and case levels. The line chart illustrates the performance of the Qwen2.5 series across different parameter sizes (3B, 7B, 14B, and 32B). The upper line represents the question pass rate (QPR), while the lower line indicates the case pass rate (CPR). The shaded region between the 2 curves highlights the consistency gap, which quantifies the disparity between question-level accuracy and case-level accuracy. As the model scale increases, both metrics improve, and the consistency gap decreases from 26.3% (3B) to 13.5% (32B), indicating performance degradation when moving from question-level accuracy to case-level aggregate accuracy.

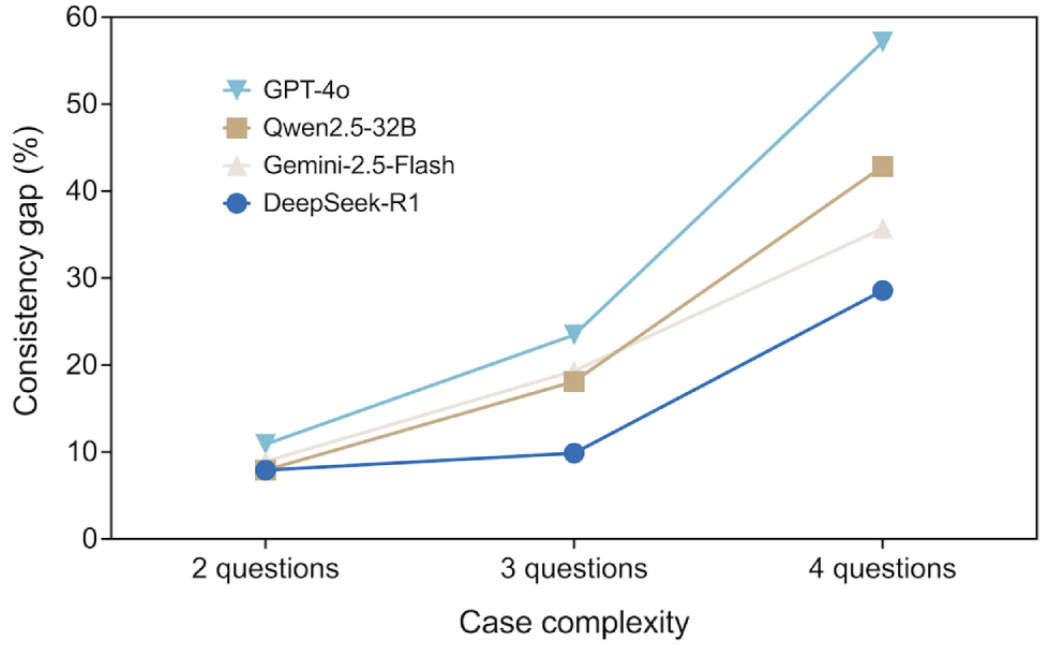


XSL•FO
RenderX

Impact of Case Complexity on Consistency Gap

To examine the relationship between task complexity and the consistency gap, we grouped the clinical cases according to the number of questions (2, 3, or 4 questions per case) and evaluated the consistency gap within each subgroup. As shown in Figure 3, the consistency gap increased with case complexity across all LLMs. This suggests that longer case sequences may be associated with greater performance degradation when moving from question-level to case-level evaluation. In lower-complexity settings (2 questions per case), the models showed similar levels of stability, with consistency gaps ranging from 7.9% (DeepSeek-R1 and Qwen2.5-32B) to 10.9% (GPT-4o). As complexity increased, however, the performance patterns differed. For cases with 3 questions, DeepSeek-R1 maintained a relatively small gap of 9.9%, whereas the gaps for Gemini-2.5-Flash and GPT-4o increased to 19.3% and 23.5%, respectively. The difference was largest in higher-complexity settings (4 questions per case). However, it is important to note that this subgroup contained only 7 cases, which limits the generalizability of this specific observation. DeepSeek-R1 limited the gap to 28.6%, while GPT-4o showed a substantial reduction in consistency, with the gap rising to 57.1%.

Figure 3. Impact of case complexity on consistency gap. The line chart shows changes in the consistency gap as the number of questions per clinical case increases from 2 to 4. A larger gap represents a greater difference between question-level accuracy and case-level accuracy. Across all models, the gap increases with case complexity.



Discussion

Principal Findings

This study provides a multidimensional evaluation of LLMs on the CNMLE, shifting the focus from traditional question-level accuracy to case-level performance on multistep clinical cases. Our results revealed a key finding. Although current models such as DeepSeek-R1 and GPT-4o achieved impressive question-level accuracy, their ability to solve complete clinical cases was significantly lower, revealing a consistency gap ranging from 10% to 17.9%. This indicates that current LLMs are excellent knowledge retrievers but show substantial performance decline when required to answer multiple interdependent questions correctly, highlighting a limitation in their practical applicability to complete clinical tasks. Furthermore, our model scaling analysis on the Qwen2.5 series (from 3B to 32B) demonstrated a strong negative correlation between model parameter size and the consistency gap. The consistency gap narrowed by nearly 50% as the model scaled up, suggesting that larger models are better at preserving performance when moving from question-level to case-level evaluation. Notably, our analysis of case complexity revealed a consistent pattern across models: the consistency gap between question-level and case-level performance tended to increase with the number of questions per case. As the number of questions within a clinical case increased, the consistency gap also tended to increase across the evaluated models, indicating greater performance degradation from question-level to case-level evaluation in longer cases.

Comparison With Prior Work

Our results align with the general pattern reported in recent literature, confirming that state-of-the-art LLMs have largely surpassed the passing threshold for medical licensing examinations [2]. Previous studies have documented the impressive performance of models such as GPT-4 on the CNMLE, often citing accuracy rates exceeding 80% [15,17]. However, a critical limitation of these prior evaluations is their predominant focus on question-level accuracy. In these studies, clinical questions are typically treated as isolated data points. This approach simplifies the complexity of medical practice and may overestimate the reliability of LLMs in real-world settings.

In actual clinical practice, patient management is not a discrete event but rather a longitudinal process involving consecutive clinical actions [22]. A standard clinical workflow typically

XSL•FO
RenderX

progresses from history taking and physical examination to the selection of diagnostic investigations, diagnostic confirmation, treatment strategy development, and subsequent prognosis assessment or follow-up planning [23,24]. These steps are intrinsically linked. A correct diagnosis is clinically insufficient if followed by an inappropriate or incompatible treatment recommendation. Therefore, evaluating models based on a single-item accuracy metric fails to capture the case-level continuity required for safe and reliable medical decision-making.

Our study addresses this methodological gap. By leveraging the case-based questions from the CNMLE, we evaluated model performance using groups of questions that represent different stages of clinical management within a shared patient vignette. In our assessment, a clinical case consisted of multiple logically interdependent questions covering prevention, examination, diagnosis, treatment, and rehabilitation. We propose the CPR, which defines success only when the model answers all related questions for a given patient correctly. This shift from question-level performance measures to a case-level evaluation paradigm provides a more stringent and meaningful benchmark, thereby revealing limitations of current LLMs that are masked by high accuracy scores in traditional assessments.

Educational Implications

This study makes 3 main contributions for application of LLMs in medical education. First, while prior evaluations of LLMs on medical licensing examinations have focused almost exclusively on question-level accuracy [25], we introduced a case-level evaluation framework that better reflects the multistep nature of clinical cases. By requiring all questions within a case to be answered correctly, the CPR provides a more stringent and clinically meaningful benchmark. Second, we empirically demonstrated a substantial gap between question-level and case-level performance across multiple state-of-the-art models, showing that high accuracy on isolated questions did not reliably translate to success on complete clinical cases. Third, we quantified the effect of model scaling and case complexity on this gap, revealing that while larger models perform better, the performance degradation from multistep cases persisted across all model sizes.

This study has important implications for how LLMs should be integrated into medical education. There is growing interest in using LLMs as virtual teaching tools for medical students [26-29]. However, our results indicate that such use requires caution. First, model size had a clear impact on performance. Smaller LLMs, such as those with fewer than 7B parameters, showed large consistency gaps and low CPR. As a result, they are not suitable for independent teaching tasks. These models may encourage fragmented learning, where students learn to answer individual questions correctly but fail to understand the overall logic of patient management. Second, with respect to overreliance on AI, students need to be made aware of the false sense of understanding that AI systems may create [30]. Even more advanced models, such as GPT-4o, may answer some questions within a clinical case incorrectly despite strong overall question-level performance. Medical education programs should therefore train students to critically evaluate AI-generated answers across the full clinical context rather than accepting individual answers without verification. Overall, the use of LLMs in medical education should be carefully guided [31]. AI systems should be viewed not as authoritative instructors but as support tools whose outputs require continuous human review and judgment.

Future Directions

Future research should focus on reducing the consistency gap identified in this study. From a methodological perspective, improving prompting strategies may help models achieve more stable performance in multistep clinical cases. For example, future studies could explicitly preserve conversation history across questions and evaluate sequential prompting strategies, including step-by-step prompts, to determine whether such approaches improve performance across successive stages of a case. In addition, real-world clinical information is not limited to text alone. Future evaluations should therefore include medical images and clinical records to better reflect actual clinical settings. Testing models with radiology images and narrative medical notes may provide a more realistic assessment of clinical performance. Finally, training open-source LLMs on well-curated clinical case data may further improve their performance. LLMs could benefit from task-specific training focused on clinical workflows. This approach may support the development of specialized medical systems that perform more reliably than general-purpose models in complex clinical tasks.

Limitations

Our study has several limitations. First, the evaluation was limited to text-based questions. Cases requiring interpretation of medical images, such as electrocardiograms or radiographs, were not included, even though these are essential components of routine clinical practice. Second, our analysis of case complexity was constrained by an imbalanced dataset. Most cases contained 2 or 3 questions, while only a few cases had 4 questions. Consequently, the observed performance drop in the most complex cases should be viewed as a preliminary finding. The generalizability of this result needs to be validated with a more balanced set of clinical cases. Third, the study was conducted only in Chinese. While this offers useful non-English evidence for medical AI research, the applicability of the results to other languages or medical systems requires further validation. Finally, although official licensing examination questions were used, the multiple-choice format does not fully reflect the complexity of real-world clinical practice. The CPR measures complete correctness across all questions within a case, but it does not directly assess whether answers are logically compatible with one another. In real practice, clinicians face open-ended problems with many possible actions, rather than a fixed set of predefined options. Additionally, because each question was submitted independently without retaining prior conversation history, our results reflect performance on independently answered questions aggregated at the case level rather than true sequential performance across clinical steps.

Conclusions

This study provides a critical reassessment of LLMs in CNMLE by distinguishing between question-level performance and

XSL•FO
RenderX

case-level performance. We found that while state-of-the-art models, such as DeepSeek-R1 and GPT-4o, demonstrated exceptional knowledge retrieval with high QPR, they exhibited a lower performance in CPR when navigating multistep clinical scenarios. This consistency gap indicates that current models frequently fail to achieve the same level of success on complete clinical cases as they do on individual questions, highlighting a practical limitation in their application to multistep clinical tasks. Furthermore, our analysis confirmed that while increasing model scale reduced the consistency gap, the challenge of handling complex cases remained a universal bottleneck. Consequently, we conclude that traditional single-question benchmarks overestimate the clinical capabilities of LLMs. Future research should further investigate model performance using explicitly sequential clinical evaluation designs. Until LLMs can reliably handle complete clinical cases with the same proficiency as individual questions, they should be deployed as support assistants in medical education and health care environments.

Acknowledgments

The authors declare the use of generative AI (GenAI) in the research and writing process. According to the Generative AI Delegation Taxonomy (2025), the following tasks were delegated to GenAI tools under full human supervision: translation, proofreading, and editing. The GenAI tool used was DeepSeek-V3. Responsibility for the final manuscript lies entirely with the authors. GenAI tools are not listed as authors and do not bear responsibility for the final outcomes.

Funding

This work was supported by Southwest Hospital (2025CXJS26), Technological Innovation and Application Development Major Project of Chongqing Municipal Science and Technology Bureau (CSTB2025TIAD-STX0029), and Technological Innovation and Application Development Project of Chongqing Municipal Science and Technology Bureau (CSTB2022TIAD-KPX0167).

Data Availability

The data that support the findings of this study are available from the corresponding author on reasonable request.

Authors' Contributions

JC: conceptualization, data curation, formal analysis, visualization, and writing—original draft. YZ: conceptualization and data curation. HZ: conceptualization, data curation, formal analysis, visualization, writing—original draft, project administration, and supervision. All authors read and approved the final manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Summary of large language model performance on the Chinese National Medical Licensing Examination from 8 peer-reviewed studies.
[DOCX File , 16 KB - mededu_v12i1e95342_app1.docx]

Multimedia Appendix 2
Representative examples of 2-, 3-, and 4-part clinical cases with annotated question dependency structure.
[DOCX File , 16 KB - mededu_v12i1e95342_app2.docx]

Multimedia Appendix 3
Zero-shot prompting template used for large language model evaluation.
[DOCX File , 13 KB - mededu_v12i1e95342_app3.docx]

References

1. Gilson A, Safranek CW, Huang T, Socrates V, Chi L, Taylor RA, et al. How does ChatGPT perform on the United States Medical Licensing Examination (USMLE)? The implications of large language models for medical education and knowledge assessment. JMIR Med Educ 2023 Feb 08;9:e45312 [FREE Full text] [doi: 10.2196/45312] [Medline: 36753318]
2. Zong H, Wu R, Cha J, Wang J, Wu E, Li J, et al. Large language models in worldwide medical exams: platform development and comprehensive analysis. J Med Internet Res 2024 Dec 27;26:e66114 [FREE Full text] [doi: 10.2196/66114] [Medline: 39729356]

XSL•FO
RenderX

3. Yang Z, Yao Z, Tasmin M, Vashisht P, Jang WS, Ouyang F, et al. Unveiling GPT-4V's hidden challenges behind high accuracy on USMLE questions: observational study. J Med Internet Res 2025 Feb 07;27:e65146 [FREE Full text] [doi: 10.2196/65146] [Medline: 39919278]
4. Liu M, Okuhara T, Chang X, Shirabe R, Nishiie Y, Okada H, et al. Performance of ChatGPT across different versions in medical licensing examinations worldwide: systematic review and meta-analysis. J Med Internet Res 2024 Jul 25;26:e60807 [FREE Full text] [doi: 10.2196/60807] [Medline: 39052324]
5. Chen Y, Huang X, Yang F, Lin H, Lin H, Zheng Z, et al. Performance of ChatGPT and Bard on the medical licensing examinations varies across different cultures: a comparison study. BMC Med Educ 2024 Nov 26;24(1):1372. [doi: 10.1186/s12909-024-06309-x] [Medline: 39593041]
6. Sandmann S, Hegselmann S, Fujarski M, Bickmann L, Wild B, Eils R, et al. Benchmark evaluation of DeepSeek large language models in clinical decision-making. Nat Med 2025 Aug;31(8):2546-2549. [doi: 10.1038/s41591-025-03727-2] [Medline: 40267970]
7. Qiu P, Wu C, Liu S, Fan Y, Zhao W, Chen Z, et al. Quantifying the reasoning abilities of LLMs on clinical cases. Nat Commun 2025 Nov 06;16(1):9799 [FREE Full text] [doi: 10.1038/s41467-025-64769-1] [Medline: 41198657]
8. Jin Q, Chen F, Zhou Y, Xu Z, Cheung JM, Chen R, et al. Hidden flaws behind expert-level accuracy of multimodal GPT-4 vision in medicine. NPJ Digit Med 2024 Jul 23;7(1):190 [FREE Full text] [doi: 10.1038/s41746-024-01185-7] [Medline: 39043988]
9. Yang Y, Jin Q, Zhu Q, Wang Z, Erramuspe Álvarez F, Wan N, et al. Beyond multiple-choice accuracy: real-world challenges of implementing large language models in healthcare. Annu Rev Biomed Data Sci 2025 Aug;8(1):305-316 [FREE Full text] [doi: 10.1146/annurev-biodatasci-103123-094851] [Medline: 40198845]
10. Wang X, Gong Z, Wang G, Jia J, Xu Y, Zhao J, et al. ChatGPT performs on the Chinese National Medical Licensing Examination. J Med Syst 2023 Aug 15;47(1):86. [doi: 10.1007/s10916-023-01961-0] [Medline: 37581690]
11. Fang C, Wu Y, Fu W, Ling J, Wang Y, Liu X, et al. How does ChatGPT-4 preform on non-English national medical licensing examination? An evaluation in Chinese language. PLOS Digit Health 2023 Dec 01;2(12):e0000397 [FREE Full text] [doi: 10.1371/journal.pdig.0000397] [Medline: 38039286]
12. Zong H, Li J, Wu E, Wu R, Lu J, Shen B. Performance of ChatGPT on Chinese National Medical Licensing Examinations: a five-year examination evaluation study for physicians, pharmacists and nurses. BMC Med Educ 2024 Feb 14;24(1):143 [FREE Full text] [doi: 10.1186/s12909-024-05125-7] [Medline: 38355517]
13. Ming S, Guo Q, Cheng W, Lei B. Influence of model evolution and system roles on ChatGPT's performance in Chinese Medical Licensing Exams: comparative study. JMIR Med Educ 2024 Aug 13;10:e52784 [FREE Full text] [doi: 10.2196/52784] [Medline: 39140269]
14. Zhang S, Chu Q, Li Y, Liu J, Wang J, Yan C, et al. Evaluation of large language models under different training background in Chinese medical examination: a comparative study. Front Artif Intell 2024 Dec 4;7:1442975 [FREE Full text] [doi: 10.3389/frai.2024.1442975] [Medline: 39697797]
15. Wu J, Wang Z, Qin Y. Performance of DeepSeek-R1 and ChatGPT-4o on the Chinese National Medical Licensing Examination: a comparative study. J Med Syst 2025 Jun 03;49(1):74. [doi: 10.1007/s10916-025-02213-z] [Medline: 40459679]
16. Luo D, Liu M, Yu R, Liu Y, Jiang W, Fan Q, et al. Evaluating the performance of GPT-3.5, GPT-4, and GPT-4o in the Chinese National Medical Licensing Examination. Sci Rep 2025 Apr 23;15(1):14119 [FREE Full text] [doi: 10.1038/s41598-025-98949-2] [Medline: 40269046]
17. Wang W, Zhou Y, Fu J, Hu K. Evaluating the performance of DeepSeek-R1 and DeepSeek-V3 versus OpenAI models in the Chinese National Medical Licensing Examination: cross-sectional comparative study. JMIR Med Educ 2025 Nov 14;11:e73469 [FREE Full text] [doi: 10.2196/73469] [Medline: 41237388]
18. OpenAI. GPT-4 technical report. ArXiv Preprint posted online on March 15, 2023 [FREE Full text]
19. Gemini Team. Gemini 2.5: pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. ArXiv Preprint posted online on July 7, 2025 [FREE Full text]
20. Guo D, Yang D, Zhang H, Song J, Wang P, Zhu Q, et al. DeepSeek-R1 incentivizes reasoning in LLMs through reinforcement learning. Nature 2025 Sep;645(8081):633-638. [doi: 10.1038/s41586-025-09422-z] [Medline: 40962978]
21. Bai J, Bai S, Chu Y, Cui Z, Dang K, Deng X, et al. Qwen technical report. ArXiv Preprint posted online on September 28, 2023 [FREE Full text]
22. Rajkomar A, Oren E, Chen K, Dai AM, Hajaj N, Hardt M, et al. Scalable and accurate deep learning with electronic health records. NPJ Digit Med 2018 May 8;1:18 [FREE Full text] [doi: 10.1038/s41746-018-0029-1] [Medline: 31304302]
23. Swinckels L, Bennis FC, Ziesemer KA, Scheerman JF, Bijwaard H, de Keijzer A, et al. The use of deep learning and machine learning on longitudinal electronic health records for the early detection and prevention of diseases: scoping review. J Med Internet Res 2024 Aug 20;26:e48320 [FREE Full text] [doi: 10.2196/48320] [Medline: 39163096]
24. Zong H, Wu R, Cha J, Feng W, Wu E, Li J, et al. Advancing Chinese biomedical text mining with community challenges. J Biomed Inform 2024 Sep;157:104716 [FREE Full text] [doi: 10.1016/j.jbi.2024.104716] [Medline: 39197732]

XSL•FO
RenderX

25. Zong H, Cha J, Wang J, Song Y, Zhao Y, Shi M, et al. A dataset for evaluating large language models on Chinese National Medical Licensing Examinations. Sci Data 2026 Apr 17;13(1):898 [FREE Full text] [doi: 10.1038/s41597-026-07261-9] [Medline: 41998016]
26. Li J, Zong H, Wu E, Wu R, Peng Z, Zhao J, et al. Exploring the potential of artificial intelligence to enhance the writing of English academic papers by non-native English-speaking medical students - the educational application of ChatGPT. BMC Med Educ 2024 Jul 09;24(1):736 [FREE Full text] [doi: 10.1186/s12909-024-05738-y] [Medline: 38982429]
27. Abd-Alrazaq A, AlSaad R, Alhuwail D, Ahmed A, Healy PM, Latifi S, et al. Large language models in medical education: opportunities, challenges, and future directions. JMIR Med Educ 2023 Jun 01;9:e48291 [FREE Full text] [doi: 10.2196/48291] [Medline: 37261894]
28. Preiksaitis C, Rose C. Opportunities, challenges, and future directions of generative artificial intelligence in medical education: scoping review. JMIR Med Educ 2023 Oct 20;9:e48785 [FREE Full text] [doi: 10.2196/48785] [Medline: 37862079]
29. Wang H, Shan W, Liu R, Wang Z. Can large language models serve as digital assistants for medical undergraduates? - A bibliometric mapping and scoping analysis of the medical-education literature. Digit Health 2025 Oct 17;11:20552076251390280 [FREE Full text] [doi: 10.1177/20552076251390280] [Medline: 41143096]
30. Tran M, Balasooriya C, Jonnagaddala J, Leung GK, Mahboobani N, Ramani S, et al. Situating governance and regulatory concerns for generative artificial intelligence and large language models in medical education. NPJ Digit Med 2025 May 27;8(1):315 [FREE Full text] [doi: 10.1038/s41746-025-01721-z] [Medline: 40425695]
31. Cione NJ, Pillow MT. Creating more reliable large language models in medical education. Acad Med 2025 Sep 01;100(9):1001 [FREE Full text] [doi: 10.1097/ACM.0000000000006058] [Medline: 40168555]

Abbreviations

CNMLE: Chinese National Medical Licensing Examination
CPR: case pass rate
GPU: graphics processing unit
LLM: large language model
QPR: question pass rate

Edited by A Stone; submitted 15.Mar.2026; peer-reviewed by K Pushpanathan, F Fukuzawa; comments to author 20.Jul.2026; revised version received 04.Sep.2026; accepted 08.Sep.2026; published 22.Sep.2026.

Please cite as:
Cha J, Zhao Y, Zong H
Large Language Model Performance on Multistep Clinical Cases: Comparative Study Across Question and Case Levels
JMIR Med Educ 2026;12:e95342
URL: https://mededu.jmir.org/2026/1/e95342
doi:10.2196/95342
PMID:

©Jiaxue Cha, Yan Zhao, Hui Zong. Originally published in JMIR Medical Education (https://mededu.jmir.org), 22.Sep.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Education, is properly cited. The complete bibliographic information, a link to the original publication on https://mededu.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Generative AI Factors, Cognitive Engagement, and Patient Safety Learning in Medical Education: Mixed Methods Study

Gilbert Golahi[1,2], MD; Ferdi Antonio[1], MD, MBA, MHM, PhD; Devina Nur Annisa[1,3], MD

[1]Department of Hospital Administration, Graduate School of Management, Pelita Harapan University, South Jakarta, Jakarta, Indonesia

[2]Siloam Hospitals TB Simatupang, South Jakarta, Jakarta, Indonesia

[3]JEC Eye Hospitals & Clinics, West Jakarta, Jakarta, Indonesia

Corresponding Author:
Ferdi Antonio, MD, MBA, MHM, PhD
Department of Hospital Administration
Graduate School of Management
Pelita Harapan University
Jl. Garnisun Dalam No. 8, RT 5, RW 4, Karet Semanggi
South Jakarta, Jakarta, 12930
Indonesia
Phone: 62 8121010676
Email: ferdi.antonio@gmail.com

Abstract

Background: Generative AI (GenAI) is increasingly used during medical training, including clerkships in which learning takes place alongside supervised patient care. Its educational value may depend on how students interpret, verify, and regulate AI-generated output.

Objective: This study explored medical clerkship students' accounts of GenAI use and examined how GenAI-related factors were associated with self-reported ward-based infection-prevention practice through self-efficacy in AI-generated usage and perceived AI self-control balance.

Methods: We used an exploratory sequential mixed methods design. Eight medical clerkship students with experience using GenAI participated in semistructured interviews analyzed using reflexive thematic analysis. The qualitative findings were considered alongside theory and prior evidence to specify a theory-based model. This model was subsequently examined through an online cross-sectional survey of 214 eligible students recruited from 5 private universities in the Jakarta Metropolitan Area, Indonesia. Quantitative associations were analyzed using partial least squares structural equation modeling with bootstrapping.

Results: The interviews yielded 16 themes concerning appraisal of AI-generated output, confidence in using GenAI, interaction control, patient-safety awareness, and infection-prevention learning. The structural paths were unchanged, although 3 low-loading measurement indicators were removed before hypothesis testing. Global model fit was limited (estimated-model standardized root mean square residual 0.089; normed fit index 0.691), and pairwise residual fit was not used as a retention criterion. Explainability quality exposure and perceived algorithmic fairness were positively associated with both self-efficacy and perceived AI self-control balance. Trust calibration literacy was associated with self-efficacy but not perceived AI self-control balance. Privacy concern was positively associated with self-efficacy (β=0.294, P<.001), contrary to the hypothesized direction, but not with perceived AI self-control balance. Self-efficacy (β=0.477, P<.001) and perceived AI self-control balance (β=0.407, P<.001) were positively associated with self-reported ward-based infection-prevention practice; the model explained 51.9% of its variance. Specific indirect effects through self-efficacy were observed for all 4 antecedents and through perceived AI self-control balance only for explainability and fairness. A small positive interaction was observed between promotion of patient safety and self-efficacy (β=0.122, P=.04); the corresponding interaction involving perceived AI self-control balance was not observed.

Conclusions: Self-efficacy was the more consistent pathway linking students' appraisals of GenAI with self-reported ward-based infection-prevention practice. By integrating students' accounts with quantitative testing of a theory-based model, this study extends research that has largely focused on GenAI acceptance and general educational use by examining a safety-oriented outcome within clinical training. The findings add evidence on the human factors through which GenAI use may relate to patient-safety learning, although the cross-sectional, self-reported data do not establish causal or clinical effects. Medical schools and teaching hospitals could use these findings to inform clerkship guidance on supervised verification of AI-generated output, patient-data protection, alignment with local protocols, and continued human responsibility for clinical decisions.

XSL•FO
RenderX

(JMIR Med Educ 2026;12:e100828) doi:10.2196/100828

KEYWORDS

patient safety competencies; generative artificial intelligence; medical clerkship education; self-efficacy; human-AI interaction

Introduction

Generative AI (GenAI) is increasingly embedded in medical education, changing how students seek explanations, organize information, practice reasoning, and manage cognitive effort during demanding learning tasks [1,2]. In medical clerkship education, this development is particularly important because learning occurs in safety-critical clinical environments where students' documentation, communication, and clinical reasoning are closely connected to the development of patient safety competencies [3,4]. Recent evidence indicates that medical learners already use GenAI for study support, clinical learning, and academic work, while also expressing concerns about accuracy, bias, and overreliance [4-6]. These concerns are relevant because GenAI may support reflective learning when used appropriately, but may also become a convenient answer source if learners do not remain cognitively engaged [2,4,7].

The educational value of GenAI is not automatic. Its contribution depends on task design, AI literacy, supervision, and the extent to which students remain active in interpreting and verifying AI-generated outputs [2,4,7,8]. This issue is especially salient in clerkship settings, where learners participate in patient-facing routines under supervision and gradually develop professional judgment through workplace-based learning [3,4]. Although GenAI has been discussed as a tool for feedback, clinical reasoning support, documentation, and learning-resource development, many institutions are still developing formal policies and training structures to guide its use in medical education [2,8,9]. This gap creates a need to understand how students use, experience, and regulate GenAI during clinical learning.

Patient safety is a central dimension of health care quality and concerns the prevention of avoidable harm during the delivery of care [10,11]. Health care–associated infections (HAIs), or nosocomial infections, remain a major patient safety challenge despite advances in infection prevention and control practices [11,12]. Systematic reviews and meta-analyses suggest that the burden of HAIs is particularly significant in low- and middle-income countries, where health system capacity, infection-control resources, and implementation consistency may vary across settings [13,14]. AI-supported systems are also increasingly being explored for HAI surveillance, outbreak monitoring, risk prediction, and infection-prevention monitoring, illustrating the growing role of digital tools in safety-critical hospital functions [12,15].

However, the potential contribution of AI to infection prevention does not depend on algorithms alone. Prior work emphasizes that human factors influence how AI outputs are interpreted, trusted, questioned, and translated into practice [15]. This is highly relevant for medical clerkship students, who occupy a transitional position between formal education and supervised clinical participation. Clerkship students contribute to ward routines such as documentation, communication, patient preparation, and infection-prevention behaviors, even though they remain under supervision [3]. As many of these students are familiar with digital technologies and increasingly exposed to GenAI tools, they represent an important learner group for understanding how AI-supported learning may relate to patient safety–oriented practice [6,16-19].

In this context, critical thinking involves questioning AI-generated output, comparing it with clinical evidence and supervisor guidance, and deciding whether the output should be accepted, revised, or rejected. Earlier studies on health technology adoption have often been informed by the technology acceptance model (TAM), which explains technology use through perceived usefulness, perceived ease of use, attitudes, and behavioral intention [20]. The TAM remains useful for examining adoption, including when AI systems are customized or embedded in nonstandard workflows; however, it does not by itself explain safe GenAI use in clinical learning environments, where patient safety, professional accountability, and supervision are central concerns [21-23]. GenAI systems can generate fluent and plausible outputs even when the content is incomplete or incorrect, creating risks of automation bias, misplaced confidence, and uncritical acceptance of AI-generated suggestions [22,23]. Medical education research therefore requires models that move beyond acceptance and examine the human factors that shape how learners engage with AI responsibly.

Existing studies have examined GenAI use in medical education, including students' perceptions, learning engagement, academic use, and institutional preparedness [2,4-6,9,17,18]. However, less is known about how GenAI-related factors are associated with patient-safety learning in clerkship settings, particularly in relation to ward-based infection-prevention practice. Much of the literature still emphasizes acceptance, satisfaction, intention to use, or general educational utility, with less attention to how students' confidence, perceived control, and risk awareness may relate to safety-oriented learning in clinical environments [2,4,21]. This distinction is relevant because clerkship students learn through supervised participation in clinical work, where the use of GenAI may intersect with documentation, infection-prevention routines, and patient care.

To address this gap, the study focused on medical clerkship students as learners positioned between classroom-based education and supervised clinical practice. It linked students' perceptions of GenAI with 2 human-centered mechanisms, which are self-efficacy in AI-generated usage and perceived AI self-control balance (PAISCB), and examined self-reported ward-based infection-prevention practice as a safety-oriented learning outcome. After the qualitative phase was completed, the themes were considered alongside existing theory to develop the conceptual framework and directional hypotheses. This work was completed before quantitative data collection and analysis. The themes informed the selection and arrangement of

XSL•FO
RenderX

theory-based constructs but were not used to create new psychometric constructs or questionnaire items. Consistent with social cognitive theory, the resulting paths were treated as connected parts of a single model [24].

The 12 directional hypotheses were organized into 3 related groups. H1-H8 concerned the associations between 4 appraisals of AI-generated output and the 2 human-centered mechanisms. H9 and H10 concerned the associations of these mechanisms with self-reported ward-based infection-prevention practice. H11 and H12 examined whether promotion of patient safety moderated the 2 mechanism-outcome associations. The individual paths were retained because they represented distinct theoretical relationships, but the findings were interpreted within these 3 groups rather than as unrelated hypothesis tests.

The first group considered how students' appraisals of AI-generated output might relate to their confidence and perceived control when using GenAI. Perceived algorithmic fairness may influence whether students regard AI-generated output as legitimate and appropriate for use in supervised learning [25,26]. Trust calibration literacy is relevant because students need to determine when AI output can be relied upon, when it should be questioned, and how it compares with clinical reasoning and supervisor guidance [27-29]. Explainability quality exposure may support learning by allowing students to understand the reasoning behind an output instead of accepting a recommendation at face value [30-32]. Privacy concern toward AI monitoring may have the opposite effect if concern about data use, surveillance, or unintended disclosure discourages students from engaging confidently with GenAI [33,34].

These appraisals were examined in relation to 2 aspects of student agency. Self-efficacy in AI-generated usage refers to students' belief in their ability to use AI effectively for learning and clinical tasks [35,36]. PAISCB concerns whether students feel able to retain autonomy and control over how they initiate, guide, and end an interaction with GenAI [37,38]. Fairness, trust calibration, and explainability were expected to be positively associated with both mechanisms. Privacy concern was expected to be negatively associated with them. This reasoning informed H1-H8:

- H1: Perceived algorithmic fairness is positively associated with self-efficacy in AI-generated usage.
- H2: Perceived algorithmic fairness is positively associated with PAISCB.
- H3: Trust calibration literacy is positively associated with self-efficacy in AI-generated usage.
- H4: Trust calibration literacy is positively associated with PAISCB.
- H5: Explainability quality exposure is positively associated with self-efficacy in AI-generated usage.
- H6: Explainability quality exposure is positively associated with PAISCB.
- H7: Privacy concern toward AI monitoring is negatively associated with self-efficacy in AI-generated usage.
- H8: Privacy concern toward AI monitoring is negatively associated with PAISCB.

The second group addressed whether the 2 human-centered mechanisms were associated with ward-based infection-prevention practice. Social cognitive theory proposes that contextual conditions influence action partly through people's beliefs about their own capabilities [24]. In clerkship settings, students apply what they learn through repeated ward activities, including documentation, communication, and infection-prevention behavior [3,39]. Confidence in using GenAI may help students use its output as a provisional learning aid, while perceived control over the interaction may allow them to decide when and how that output should inform their work. Neither mechanism implies that clinical judgment can be delegated to the system. AI-supported learning remains dependent on verification, supervision, and attention to the risks of inappropriate trust or loss of human authority [22,23,40]. Ward-based infection-prevention practice was selected as the outcome because it represents a safety-oriented aspect of clinical learning that is directly relevant to clerkship activities [41]. On this basis, H9 and H10 concerned the 2 mechanism-outcome paths:

- H9: Self-efficacy in AI-generated usage is positively associated with ward-based infection-prevention practice.
- H10: PAISCB is positively associated with ward-based infection-prevention practice.

The third group examined whether patient-safety orientation influenced the strength of these relationships. Students who place greater emphasis on patient safety may be more likely to translate confidence and control during GenAI use into cautious, verified practice. Promotion of patient safety was therefore examined as a moderator of the associations between the two human-centered mechanisms and ward-based infection-prevention practice [42]. This reasoning informed H11 and H12:

- H11: Promotion of patient safety positively moderates the association between self-efficacy in AI-generated usage and ward-based infection-prevention practice.
- H12: Promotion of patient safety positively moderates the association between PAISCB and ward-based infection-prevention practice.

The conceptual model and its component paths are presented in Multimedia Appendix 1.

This study aimed to explore how medical clerkship students described their use of GenAI in clinical learning and to examine, in a subsequent quantitative phase, the associations proposed in the qualitatively informed, theory-based model. The model linked students' appraisals of AI-generated output and 2 human-centered mechanisms with self-reported ward-based infection-prevention practice.

The study addressed 3 research questions. First, what GenAI-related factors did medical clerkship students identify as relevant to clinical learning and patient-safety practice? Second, to what extent were the associations specified in the quantitative model observed in relation to self-reported ward-based infection-prevention practice through self-efficacy in AI-generated usage and PAISCB? Third, how did the qualitative themes and quantitative findings converge, expand upon, or diverge from one another in explaining GenAI-supported patient-safety learning?

XSL•FO
RenderX

Methods

Research Design Overview

This study used a sequential exploratory mixed methods design comprising an initial qualitative phase followed by a cross-sectional quantitative phase. In this study, exploratory refers to the sequence and purpose of the inquiry rather than to the development of new psychometric constructs or questionnaire items. The qualitative phase used semistructured interviews and reflexive thematic analysis to examine how medical clerkship students described GenAI use in clinical learning, including their experiences of trust, fairness, explainability, privacy, human control, and patient safety–related practice. The quantitative phase subsequently examined the associations among the corresponding theory-based constructs, self-efficacy in AI-generated usage, PAISCB, promotion of patient safety, and self-reported ward-based infection-prevention practice [43-45].

This design was selected because GenAI use in clerkship education involves both context-dependent experiences and relationships that can be examined quantitatively. The qualitative phase provided an account of how students understood and used GenAI within their clerkship setting, whereas the quantitative phase examined whether the specified relationships were observed in a larger sample. Reporting of the qualitative and mixed methods components was informed by the relevant Journal Article Reporting Standards (JARS) recommendations [46]. Construct measurement remained grounded in the source instruments described in the “Measures and Covariates” section.

Integration followed a sequential building strategy with a theory-informed abductive logic. After the qualitative themes had been developed inductively, they were considered alongside existing theory and prior evidence to develop the conceptual framework and directional hypotheses for the quantitative phase. This work was completed before quantitative data collection and analysis. The themes informed the selection and arrangement of theory-based constructs within the structural model, while the direction of the proposed relationships was supported by theory and previous empirical findings [47,48]. The quantitative phase subsequently examined the resulting model, and findings from both phases were considered together during interpretation [45,49].

Study Setting and Recruitment Period

This study was conducted among medical clerkship students from 5 private universities in the Jakarta Metropolitan Area, Indonesia. Data collection was conducted sequentially. Qualitative interviews took place from January to February 2026, followed by recruitment for the online quantitative survey from February to March 2026. The qualitative participants represented major clinical departments, including internal medicine, surgery, pediatrics, and obstetrics, whereas the quantitative survey was distributed through the faculties of medicine of the participating universities.

Ethical Considerations

The study was approved by the Pelita Harapan University Institutional Ethics Committee (approval number 060/MARS/EC/XII/2025). Participation in both phases was voluntary. Written informed consent was obtained before the qualitative interviews, while quantitative respondents provided electronic informed consent before beginning the survey. Before providing consent, participants were informed about the study purpose and procedures, potential risks and benefits, confidentiality safeguards, and their right to withdraw without penalty. No directly identifying information was collected through the quantitative survey. Qualitative interview transcripts were deidentified, and pseudonyms were used in place of participants’ names. Electronic study data were stored on encrypted drives and used only for academic research purposes. Participants received no financial or material compensation. No identifiable participant information is presented in this paper or supplementary materials.

Phase 1: Qualitative Methods

The first phase used semistructured, in-depth interviews to examine students’ accounts of GenAI use in clinical and academic learning. The transcripts were analyzed using reflexive thematic analysis, with themes developed inductively before mixed methods integration [43,50,51].

Study Participants or Data Sources

Researcher Description

The qualitative phase was conducted by a 3-member research team with prior experience in qualitative and mixed methods research. GG, a medical doctor and graduate student in hospital administration at the time of the study, conducted all interviews. He had received training in qualitative research methods and led the coding and initial development of themes. FA provided methodological supervision and contributed to validation of the analysis, while DNA supported data curation and validation.

The research team recognized that GG’s clinical background and interest in GenAI could shape both the interview process and the interpretation of participants’ accounts. Contextual observations were documented in field notes, and preliminary interpretations were discussed with FA and DNA and checked against the interview transcripts before the themes were finalized. These discussions were used to examine whether the interpretations remained grounded in participants’ accounts rather than reflecting the research team’s initial expectations.

Participants or Other Data Sources

Qualitative data were obtained from 8 active medical clerkship students who had experience using GenAI for clinical or academic learning. Their ages ranged from 22 to 27 years; 6 were women and 2 were men. The participants represented 4 major clinical departments: internal medicine, surgery, pediatrics, and obstetrics. Each participant contributed one semistructured interview, and the thematic analysis included all 8 verbatim transcripts.

Researcher-Participant Relationship

The interviewer (GG) had no prior relationship with any of the participants and did not serve as their lecturer, clinical supervisor, assessor, or academic administrator. Information about the study was distributed through faculty representatives. Before the interviews, participants were informed that GG was

XSL•FO
RenderX

a medical doctor and graduate student in hospital administration conducting the research. They were also told that participation was voluntary and that neither their decision to participate nor their responses would affect their academic grades, clinical rotations, or relationship with the faculty.

Participant Recruitment

Recruitment Process

Recruitment took place from January to February 2026. Faculty representatives at the participating universities distributed information about the study to medical clerkship students but were not involved in determining eligibility. With the students' permission, the contact details of those who expressed interest were forwarded to GG, who conducted the eligibility screening and provided further information about the interview. Eleven students were approached. Three declined without providing a reason, while 8 met the eligibility criteria, provided written informed consent, and completed the interviews. The participants were recruited from several of the participating universities. No financial or material compensation was offered.

Participant Selection

Participants were selected purposively based on their relevance to the qualitative aim. Eligible participants were active medical clerkship students with direct experience using GenAI for clinical or academic learning. Selection also sought variation across major clinical departments, including internal medicine, surgery, pediatrics, and obstetrics. Sample adequacy was assessed using the concept of information power, considering the focused study aim, the specificity of the participant group, the relevance of their experience, the use of an established theoretical framework, and the depth of the interviews [52]. After 8 interviews, the dataset was judged to provide sufficient contextual range and depth to address the purpose of the qualitative phase, and recruitment was closed. The later interviews largely reinforced the patterns already identified and did not add substantively different contextual insights.

Data Collection

Interview Procedures

Qualitative data were collected from January to February 2026 through individual, face-to-face, semistructured interviews conducted in Indonesian by GG. Each participant completed one interview in a private consultation room. Only the participant and the interviewer were present. The interviews lasted between 45 and 60 minutes, with a mean duration of 53.8 (SD 4.7) minutes. No repeat interviews were conducted.

The interview guide was developed from the study objectives, relevant theory and literature, and discussions within the research team. FA and DNA reviewed the guide before data collection began. The interviews covered participants' use of GenAI in clinical and academic learning, their interpretation of AI-generated outputs, and their experiences related to trust, fairness, explainability, privacy, human control, and patient safety–related practice. The order of questions was adapted to the direction of each interview, and follow-up questions were used when clarification or further detail was needed. Field notes were made during and immediately after each interview. The same interview guide was used throughout the qualitative phase, with no substantive changes made during data collection. The interview guide is available from the corresponding author upon reasonable request.

Recording and Data Transformation

All interviews were audio-recorded using a password-protected digital recording device. GG transcribed the recordings manually and verbatim in Indonesian. Each transcript was checked in full against the corresponding audio recording, and any transcription errors were corrected before analysis. Names and other potentially identifying details were removed, and each participant was assigned a pseudonym.

The original Indonesian transcripts were used for coding and thematic analysis. After the analysis, quotations selected for publication were translated into English by GG and reviewed by DNA to ensure that their meaning and context were retained. Any differences in interpretation were resolved through discussion. Full transcripts were not returned to participants; participant review was limited to thematic summaries and is described in the "Methodological Integrity" section. Audio recordings, deidentified transcripts, and field notes were stored on encrypted drives accessible only to the research team.

Analysis

Data-Analytic Strategy

The interview transcripts were examined using Braun and Clarke's 6-phase reflexive thematic analysis [50,51]. This analysis was intended to develop a contextual account of how clerkship students experienced and interpreted GenAI use in clinical learning. GG conducted coding and theme development manually without dedicated qualitative data-analysis software. GG read each transcript several times before coding began. The full transcript served as the primary unit of analysis, with codes applied to meaningful segments of text. Initial codes were developed across the dataset without a predetermined coding framework. The theory-based constructs included in the quantitative model were not used as coding categories at this stage.

Codes were compared within and across interviews to identify recurring patterns as well as differences in participants' accounts. Related codes were brought together into candidate themes, which were then reviewed against the coded extracts and the full dataset. Themes were combined, separated, or refined when their boundaries were unclear or their supporting data were insufficiently coherent. The central idea and scope of each theme were subsequently defined before the final names were assigned. These phases were recursive rather than strictly sequential, allowing movement between transcripts, codes, and candidate themes throughout the analysis. No separate formal discrepant-case analysis was conducted; however, contrasting accounts were retained when they clarified variation within a theme or qualified a broader pattern.

Theme development remained inductive and data-driven [50,51]. After the themes had been finalized, they were considered alongside existing theory during the abductive integration described in the "Mixed Methods Analysis" section. The

XSL•FO
RenderX

direction and theoretical rationale of the proposed relationships were established from prior theory and empirical evidence before quantitative data collection and analysis.

Methodological Integrity

Credibility and confirmability were supported through comparison across participants, peer debriefing, and participant feedback. During theme development, GG compared coded extracts within and across transcripts to determine whether candidate themes were adequately supported while retaining variation in participants' accounts. Developing codes and themes were discussed with FA and DNA through peer debriefing. These discussions were used to question assumptions, consider alternative readings of the data, and refine the scope of the themes rather than to establish intercoder agreement.

Three participants reviewed concise summaries of the themes and commented on whether the summaries reasonably reflected the substance of their accounts. This process involved thematic summaries rather than the return or correction of full transcripts. Final themes were checked against the source transcripts and presented with deidentified participant quotations to show how the interpretations were grounded in the interview data.

The a priori constructs were not used as initial coding categories. No formal discrepant-case analysis was undertaken, although contrasting accounts were retained when they qualified a broader pattern. These boundaries informed the subsequent mixed methods interpretation.

Phase 2: Quantitative Methods

Inclusion and Exclusion

The quantitative phase included medical clerkship students aged 18 to 27 years from 5 private universities in the Jakarta Metropolitan Area. Respondents were eligible if they had completed or were currently undergoing a major clinical rotation and reported frequent or very frequent use of GenAI, including ChatGPT, Google Gemini, Grok, BLACKBOX AI, or another platform, in clinical learning. A submission was not included if the screening responses did not confirm eligibility or if the main questionnaire was incomplete.

Participant Characteristics

The final quantitative analytic sample comprised 214 medical clerkship students. Table 1 summarizes their sex and age distribution, frequency of GenAI use in clinical learning, and the platforms they reported using. Respondents could select more than one platform; therefore, the percentages reported for GenAI platforms do not sum to 100%.

Table 1. Demographic characteristics and generative AI (Gen AI) use profile of 214 medical clerkship students recruited in the quantitative phase from 5 private universities in the Jakarta Metropolitan Area, Indonesia.

Description and category	Value, n (%)
Sex	
Male	56 (26)
Female	158 (74)
Age (years)	
18-24	194 (91)
25-27	20 (9)
GenAI usage frequency in the clinical study setting	
Frequent	162 (76)
Very frequent	52 (24)
GenAI platforms used	
ChatGPT	193 (90)
DeepSeek	57 (27)
Google Gemini	135 (63)
Meta AI	89 (42)
Microsoft Copilot	29 (14)
Claude.AI	40 (19)
Consensus	28 (13)
Grok	21 (10)
BLACKBOX AI	80 (37)
Others	2 (1)

XSL•FO
RenderX

Sampling Procedures

The quantitative phase used purposive, nonprobability sampling. From February to March 2026, faculty representatives at the 5 participating universities distributed the study information and a link to the cloud-based survey. Their role was limited to distributing the invitation; they did not determine eligibility or have access to individual responses. Students decided independently whether to open the survey and provide electronic informed consent. Mandatory screening questions were presented before the main questionnaire.

Faculty representatives sent the survey invitation directly to medical clerkship students identified as potential participants. Eligibility was not assumed from receipt of the invitation but was determined through the mandatory screening questions presented before the main questionnaire. The quantitative participant flow is presented in Multimedia Appendix 2.

Sample Size, Power, and Precision

The required sample size was estimated using an a priori power analysis in G*Power 3.1. (Franz Faul, Edgar Erdfelder, Axel Buchner, and Albert-Georg Lang). Assuming a medium effect size (f^2=0.15), statistical power of 0.80, an alpha level of .05, and 7 predictor variables, the analysis indicated that at least 103 participants were required [53]. The final analytic sample comprised 214 respondents and therefore exceeded the minimum sample size determined by the power analysis.

The calculation addressed statistical power for the structural analysis rather than precision for estimating a population prevalence. A margin-of-error calculation was not used because the study applied purposive, nonprobability sampling.

Measures and Covariates

The quantitative survey used a structured questionnaire consisting of closed-ended items. Four GenAI-related constructs were examined as antecedents: AI-perceived algorithmic fairness [25], AI-trust calibration literacy [54], AI-explainability quality exposure [30], and AI-privacy concern toward AI monitoring [34]. The model included self-efficacy in AI-generated usage [36] and PAISCB [38] as mediating constructs. Promotion of patient safety was examined as a moderating variable [42], while ward-based infection-prevention practice was treated as the dependent variable [41].

PAISCB referred to students' perceived autonomy and control over their interaction with GenAI. In this construct, "balance" denotes the student's sense of retaining control while using AI rather than a formal division of authority between the learner and the system. The measure was not intended to assess clinical oversight, accountable decision-making, or compliance with clinical governance procedures [38].

Sex, age, GenAI usage frequency, and the platforms used were collected to describe the sample but were not entered as covariates in the structural model. No additional covariates were included.

Data Collection

Quantitative data were collected from February to March 2026 through an individually accessed cloud-based survey. After completing the electronic consent and eligibility screening steps described above, eligible respondents proceeded to the main questionnaire. They were instructed to answer based on their own experiences of using GenAI during medical clerkship learning. Instructions were provided at the beginning of each questionnaire section, and completion typically required approximately 8 to 12 minutes. No IP addresses or other directly identifying information were collected. After data collection closed, the survey data were downloaded for response screening and statistical analysis.

Quality of Measurements

All questionnaire items were translated into Indonesian and reviewed by a linguistic expert for semantic accuracy, conceptual equivalence, and relevance to the study context. The adapted questionnaire was then assessed by a 5-member expert panel comprising 3 academics, 1 information technology professional, and 1 infection prevention and control professional. The panel reviewed each item for clarity, relevance, and appropriateness for medical clerkship students. Minor wording revisions were made in response to their assessment.

Before the main survey, 40 medical clerkship students were approached for a preliminary pilot and readability assessment, of whom 30 participated. They were drawn from the same target population but were not included in the main analytic sample. The assessment examined whether the instructions and item wording were understandable and appropriate to the intended clinical-learning context. The questionnaire was finalized after this assessment.

Several procedural measures were also used to limit common method bias [55]. The survey instructions stated that there were no right or wrong answers and asked respondents to answer according to their own experiences with GenAI. Items representing the antecedent, mediating, moderating, and outcome constructs were arranged to reduce obvious response patterns. These measures were procedural; no marker variable or other dedicated statistical measure of common method bias was included.

Instrumentation

The questionnaire comprised items adapted from previously validated instruments and adjusted to the medical clerkship and GenAI-supported learning context. No ad hoc instrument or newly written questionnaire item was introduced. The response-scale length of each source instrument was retained: AI-trust calibration literacy used a 7-point Likert scale, PAISCB used a 6-point Likert scale, and the remaining constructs used 5-point Likert scales. These response scales were not converted to a common number of categories during data collection. Apart from the planned reverse coding of the negatively worded ward-based infection-prevention items, the original response formats were retained. Construct definitions and response-scale lengths are provided in Multimedia Appendix 3, while the adapted item wording and reverse-coding status are provided in Multimedia Appendix 4.

The questionnaire used in the main survey contained 35 indicators adapted from 8 source measures. The indicators were analyzed as individual items rather than as parcels or summed

XSL•FO
RenderX

scale scores. All 8 measurement models were specified as reflective, and the Likert-type responses were treated as metric indicators in the partial least squares structural equation modeling (PLS-SEM) analysis.

Masking

Masking did not apply to the quantitative phase because the study used a nonexperimental cross-sectional survey with no intervention, experimental manipulation, random allocation, or condition assignment. All respondents completed the same questionnaire, and the outcomes were self-reported. There were therefore no intervention providers or independent outcome assessors to be masked.

Psychometrics

The expert assessment yielded Aiken V values above 0.40 and item-level content validity index values above 0.80 for all questionnaire items after minor wording revisions. In the preliminary assessment involving 30 medical clerkship students, Cronbach α exceeded 0.70 and the average variance extracted (AVE) exceeded 0.50 for each construct. These findings were considered sufficient to proceed with the main survey.

The psychometric properties of the scores obtained from the main sample were examined as part of the PLS-SEM measurement model assessment [56]. Indicator reliability was evaluated using outer loadings. Indicators with loadings between 0.40 and 0.70 were considered in relation to their effect on construct reliability, convergent validity, and conceptual coverage rather than being removed solely based on the loading value. Internal consistency was examined using Cronbach α, rho_A, and composite reliability, with values of 0.70 or higher considered acceptable. Convergent validity was assessed using the AVE, with values above 0.50 indicating adequate convergence.

The initial measurement model contained 35 indicators. Three indicators from the promotion of patient safety measure had low outer loadings. They were removed after their loadings were considered together with internal consistency, convergent validity, and the conceptual coverage of the construct. The model was then reestimated with 32 indicators. Each construct retained at least 3 indicators, and the removal did not change the definition of the promotion of patient safety construct. The retained indicators and sample-specific estimates are reported in the "Measurement Model Estimates" section and in Multimedia Appendix 4.

Discriminant validity was evaluated using the heterotrait-monotrait ratio of correlations (HTMT) point estimates and bootstrapped 95% CIs. A point estimate below 0.90 was used as a general guideline, while the CIs were examined to determine whether they included 1.00. Estimates approaching 0.90 or intervals with upper limits above 0.90 were interpreted cautiously rather than treated as unequivocal evidence of construct separation [56]. The sample-specific psychometric estimates are reported in the "Measurement Model Estimates" section and in Multimedia Appendices 4-6.

Conditions and Designs

The quantitative phase used a nonexperimental, cross-sectional observational design. All constructs were measured once through the same online questionnaire. There was no intervention, manipulation of study variables, random allocation, comparison group, or repeated measurement. The conceptual framework and the direct, indirect, and moderating relationships were finalized after the qualitative phase and before quantitative data collection and analysis. No interim analysis or data-dependent stopping rule was used; recruitment ended at the close of the prespecified data-collection period. The estimated relationships were interpreted as cross-sectional associations and not as evidence of temporal or causal effects.

Data Diagnostics

Before analysis, submissions were screened for eligibility and completion of the main questionnaire. Item responses were then checked for missing values, values outside the permitted response ranges, and coding inconsistencies. Response-pattern checks were descriptive and were not used as an automatic exclusion rule; no response was removed solely because the same category had been selected across multiple items. No additional cases were excluded after the analytic dataset had been formed.

The final dataset of 214 respondents contained no item-level missing data; therefore, missing completely at random testing and data imputation were not performed. Observed ranges and item distributions were inspected before model estimation. Skewness and kurtosis were used descriptively, while standardized construct scores with an absolute value greater than 3.29 were treated as flags for review rather than automatic grounds for exclusion. No case was removed solely because of an extreme score or departure from normality. The negatively worded ward-based infection-prevention practice (NIFRDT) indicators were reverse-coded so that higher values represented better self-reported ward-based infection-prevention practice and lower agreement with noncompliant behaviors. No transformation was applied to the remaining item responses. The reverse-coded indicators are identified in Multimedia Appendix 4.

Analytic Strategy

Specification

The quantitative analysis used observed responses from 214 participants rather than simulated data. The structural model specified AI-perceived algorithmic fairness, AI-trust calibration literacy, AI-explainability quality exposure, and AI-privacy concern toward AI monitoring as antecedent constructs. Self-efficacy in AI-generated usage and PAISCB were positioned as intervening constructs, ward-based infection-prevention practice was the endogenous outcome, and promotion of patient safety moderated the 2 intervening-construct-to-outcome relationships. The hypothesized model and its directional paths are presented in Multimedia Appendix 1.

PLS-SEM was selected because the analysis combined explanatory and predictive aims and included several latent

XSL•FO
RenderX

constructs, specific indirect effects, interaction effects, and importance-performance analysis [56]. The qualitatively informed, theory-based model was evaluated as specified, without fitting alternative structural models. The analysis followed a 2-stage procedure, beginning with measurement-model evaluation and continuing with structural-model assessment [56]. The criteria used to evaluate the measurement model are described in the "Psychometrics" section.

The final model contained 8 reflective constructs measured by 32 retained indicators. The indicators were questionnaire items adapted from several previously validated instruments and administered together in one questionnaire. Individual items were entered directly, without parceling or preliminary aggregation, and were treated as metric indicators. Each construct had at least 3 retained indicators. The structural model was recursive and contained no reciprocal or feedback paths.

PLS-SEM was estimated as a recursive component-based path model. Conventional covariance-based structural equation modeling degrees of freedom and identification tests were therefore not used because the model was not estimated by reproducing the sample covariance matrix [56]. All constructs were specified using reflective measurement models with distinct sets of indicators; each construct retained at least 3 indicators, and no reciprocal or feedback paths were included. These specifications allowed the measurement and structural parameters to be estimated in SmartPLS (SmartPLS GmbH).

The model included specific indirect pathways through self-efficacy in AI-generated usage and PAISCB. Direct paths from the 4 antecedent constructs to ward-based infection-prevention practice were not specified; accordingly, the indirect effects were not classified as indirect-only or complementary mediation. Moderation was specified through interaction terms between promotion of patient safety and each intervening construct. Promotion of patient safety and the corresponding intervening constructs were retained as constituent main effects in the outcome equation.

Participants were recruited from 5 universities. Because the respondent-level dataset did not retain a university identifier, institution-level clustering could not be estimated or adjusted for. The quantitative results therefore represent a pooled analysis across the participating universities.

Estimation

PLS-SEM was performed in SmartPLS version 4.1.1.2. Estimation used the path weighting scheme, initial weights of 1.0, a maximum of 3000 iterations, and a stop criterion of 10 . Both unstandardized and standardized estimates were obtained for the structural paths, while standardized estimates were used for model evaluation.

The structural model was assessed using inner variance inflation factors, coefficients of determination (R^2), and the magnitude and precision of the estimated path coefficients. Statistical inference was based on 10,000 bootstrap subsamples with bias-corrected (BC) CIs. Because the directional hypotheses were finalized before quantitative data collection and analysis, 1-tailed tests were conducted at α=.05, with 90% BC CIs. A hypothesis was considered supported only when the coefficient was in the proposed direction and met the prespecified statistical criterion. A statistically significant coefficient in the opposite direction was classified as unsupported.

The 12 directional paths were treated as related components of a single theory-informed model. They were considered in 3 groups: paths from the 4 antecedent constructs to the 2 intervening constructs, paths from the intervening constructs to ward-based infection-prevention practice, and the 2 interaction paths. No formal multiplicity adjustment was applied because each path represented a distinct theory-based association and the study conclusions were not based on whether at least one path reached statistical significance. All estimated paths were reported regardless of their direction or statistical significance. Interpretation considered the magnitude and direction of the coefficients, their CIs, and the consistency of the findings within each hypothesis group. The reported *P* values are therefore path-specific and unadjusted, and findings close to the nominal significance threshold were interpreted cautiously. Specific indirect effects, predictive assessment, and importance-performance analysis were treated as secondary analyses and were not used to determine support for H1 to H12.

Specific indirect effects were evaluated using the same BC bootstrap procedure. The interaction terms were estimated using the 2-stage approach in SmartPLS. Statistically significant interactions were examined using simple-slope plots to describe the associations at different levels of promotion of patient safety.

Out-of-sample predictive performance was assessed using Q^2predict and Cross-Validated Predictive Ability Test (CVPAT) [57,58]. CVPAT compared the prediction errors of the PLS-SEM model with indicator average and linear model benchmarks at the overall-model and construct levels [57]. PLSpredict and CVPAT used 10-fold cross-validation repeated 10 times with a fixed seed. Predictive performance was examined at both the indicator and construct levels rather than from an overall statistic alone.

Importance-performance map analysis (IPMA) was conducted at the indicator level to compare total effects with performance scores [58]. Performance was rescaled from 0 to 100 using the theoretical endpoints of each instrument: 1-7 for AI-trust calibration literacy, 1-6 for PAISCB, and 1-5 for the remaining constructs. The original item responses were not converted to a common response scale for the PLS-SEM analysis; rescaling was applied only when calculating IPMA performance. Indicators with comparatively high importance and lower performance were considered possible priorities for educational improvement.

Mixed Methods Analysis

Integration occurred at 2 stages. First, after the qualitative analysis had been completed, the inductively developed themes were examined alongside relevant theory and prior research. This process was used to determine which established constructs were relevant to the participants' accounts and how those constructs should be arranged in the quantitative model. The conceptual framework and directional hypotheses were finalized before quantitative data collection and analysis. This

XSL•FO
RenderX

between-phase process represented sequential building and was guided by theory-informed abductive reasoning [47,48].

An analytic matrix was used during this process to relate the qualitative themes to relevant theoretical concepts and previously validated measures. The themes informed the selection and arrangement of 4 GenAI-related appraisals, 2 human-centered mechanisms, promotion of patient safety as a moderator, and self-reported ward-based infection-prevention practice as the outcome. No new psychometric constructs or questionnaire items were developed from the interviews. The qualitative codes and themes were also not revised to make them conform to the resulting quantitative model.

The second stage of integration took place after the quantitative analysis. The qualitative findings and structural-model results were considered together to identify areas of convergence, contextual expansion, and divergence. Integration was conducted through narrative comparison, supported by a tabular presentation of the qualitative findings and their links with the theory-based constructs in the "Results" section. The qualitative and quantitative phases involved separate participant groups; the 8 interview participants were not included in the quantitative survey sample. Qualitative data were not converted into numerical variables, and no participant-level data were merged across phases. Integration therefore informed the interpretation of the findings but did not alter the quantitative estimates [45,49].

Because the qualitative findings contributed to the development of the quantitative framework, agreement between the two phases was not treated as independent confirmation. Participant accounts that did not align with the model, as well as quantitative paths that were unsupported or opposite to the hypothesized direction, were retained in the interpretation. No formal discrepant-case analysis was conducted. The integrated conclusions therefore reflected convergence, expansion, and divergence across the 2 phases rather than agreement alone.

Validity, Reliability, and Methodological Integrity

Procedures supporting methodological integrity in the qualitative phase are described in the "Participant Selection" and "Methodological Integrity" sections. In brief, the analysis included comparisons within and across interviews, peer debriefing within the research team, and feedback from 3 participants on concise thematic summaries. Candidate themes were checked against the source transcripts, and contrasting accounts were retained when they qualified a broader pattern. Deidentified quotations were used in reporting to show the basis of the interpretations. These procedures were intended to keep the analysis grounded in participants' accounts rather than to establish intercoder agreement [59].

Quantitative validity and reliability were addressed during questionnaire adaptation and again in the main sample analysis. The study used measures adapted from previously validated instruments, followed by linguistic review, expert assessment, and readability testing before the survey was administered. In the main sample, measurement quality was examined through indicator reliability, internal consistency, convergent validity, and discriminant validity. Heterotrait-monotrait ratio estimates were considered together with their bootstrapped CIs, and results close to the stated criterion were interpreted cautiously rather than treated as unequivocal evidence of construct separation [56]. The assessment procedures are described in the "Quality of Measurements" and "Psychometrics" sections, whereas the sample-specific estimates are reported in the "Results" section and Multimedia Appendices 1-7.

Mixed methods legitimacy was considered in relation to the sequence of the study and the role of each phase. Qualitative themes were developed inductively before they were considered alongside theory and prior evidence. The conceptual framework and directional hypotheses were then finalized before quantitative data collection and analysis. The qualitative findings informed the selection and arrangement of established constructs but were not used to create new psychometric constructs or questionnaire items. After the quantitative analysis, findings from the 2 phases were considered for convergence, contextual expansion, and divergence [45,49]. Because the qualitative findings contributed to the development of the quantitative framework, agreement across phases was interpreted as coherence rather than independent confirmation. Contrasting qualitative accounts and quantitative paths that were unsupported or opposite to the proposed direction were retained in the interpretation. The integrated conclusions were therefore limited to what could reasonably be inferred from the qualitative accounts, the cross-sectional quantitative associations, and their points of intersection.

Results

Phase 1: Qualitative Findings

Analysis of the 8 interview transcripts yielded 16 themes. Four domains described how students appraised AI-generated output: perceived algorithmic fairness, trust calibration literacy, explainability quality exposure, and privacy concern toward AI monitoring. The remaining themes concerned students' confidence in using GenAI, their perceived control over the interaction, patient-safety awareness, and ward-based infection-prevention practice. Table 2 presents the themes, representative quotations, analytic interpretations, and the theory-based constructs considered during mixed methods integration.

Participants generally described GenAI as useful for obtaining explanations and supporting clinical learning, but not as sufficient on its own. They valued its apparent objectivity while recognizing that its training data might not reflect local patients or uncommon clinical presentations. Trust was conditional: students compared AI-generated output with supervisors' views, their own clinical reasoning, or cases encountered in practice. When an output was incorrect, comparison with the clinical situation sometimes became a learning opportunity.

Explainability influenced whether students felt comfortable using the output. Reasoned explanations supported engagement, whereas answers without justification prompted hesitation. Privacy concerns focused on possible disclosure or reuse of patient information and, for some participants, the sense of being monitored. These concerns coexisted with continued

XSL•FO
RenderX

GenAI use and shaped how cautiously students approached the technology.

Confidence developed through repeated use, but participants continued to place final judgment with the learner and clinical supervisor. Discussing AI-generated output with supervisors helped students compare the system's reasoning with clinical expectations. Perceived control was therefore described as an active part of using GenAI rather than a simple decision to accept or reject its output.

Table 2. Qualitative themes derived from interviews with medical clerkship students, representative quotations, conceptual interpretations, and corresponding theory-based quantitative constructs.

Theme	Respondent quotes	Conceptual interpretation	Related variable
T1: Perceived fairness in AI decision-making	"AI seems fairer than my supervisor, but I worry the dataset doesn't represent our local patients."	Reflects concern about algorithmic objectivity and contextual fairness.	AI-perceived algorithmic fairness
T2: Algorithmic bias awareness	"Sometimes AI misses rare cases; I think it's trained on typical Western data."	Awareness of cultural and contextual limitations in AI datasets.	AI-perceived algorithmic fairness
T3: Trusting but verifying AI output	"I trust AI only after checking with my consultant's opinion."	Demonstrates calibrated trust and critical engagement.	AI-trust calibration literacy
T4: Learning from AI mistakes	"When AI gave a wrong result, I learned why by comparing with real cases."	Reflective practice promotes deeper literacy in trust calibration.	AI-trust calibration literacy
T5: Understanding AI logic	"It helps when AI explains the reasoning behind its diagnosis."	Perceived transparency enhances learning confidence.	AI-explainability quality exposure
T6: Struggling with nontransparent AI	"AI gives results without justification; I hesitate to rely on it."	Lack of interpretability reduces engagement and self-efficacy.	AI-explainability quality exposure
T7: Privacy concerns about patient data	"I'm afraid our patients' data could leak through the AI system."	Highlights ethical awareness and sensitivity to data governance.	AI-privacy concern toward AI monitoring
T8: Feeling monitored by AI systems	"It feels like I'm constantly being observed by the system."	AI surveillance may reduce autonomy and comfort.	AI-privacy concern toward AI monitoring
T9: Ethical reflection on data use	"We must ensure data isn't reused without consent."	Reinforces professional accountability in digital contexts.	AI-privacy concern toward AI monitoring
T10: Building confidence through practice	"At first, I was nervous, but after using AI daily, I got confident."	Self-efficacy improves through repetitive, supportive use.	Self-efficacy in AI-generated usage
T11: AI as a learning partner	"AI helped me learn diagnostic reasoning faster."	AI functions as a co-educator, enhancing learning autonomy.	Self-efficacy in AI-generated usage
T12: Balancing AI and human judgment	"AI assists with screening, but final judgment must be mine."	Reflects perceived equilibrium between automation and human control.	Perceived AI self-control balance
T13: Human-AI collaboration in clinical tasks	"Discussing AI results with my supervisor made me understand both sides."	Collaborative reasoning fosters confidence and ethical control.	Perceived AI self-control balance
T14: Patient safety awareness through AI	"AI reminders reduced small medication errors."	Demonstrates how AI integration links to safer clinical practices.	Promotion of patient safety
T15: AI-driven infection monitoring	"AI flags unusual infection trends before we notice."	Links AI insights to infection prevention strategies.	Ward-based infection-prevention practice
T16: Institutional support for safe AI use	"When the hospital trains us on AI ethics, we feel more secure."	Organizational facilitation reinforces trust and responsible use.	Promotion of patient safety

Patient-safety themes appeared in accounts of medication-error reminders, recognition of unusual infection trends, and institutional training on safe AI use. These examples did not suggest that GenAI independently improved patient safety. Rather, participants understood its value as conditional on verification, supervision, and clear responsibility for clinical decisions. Across the themes, GenAI was experienced as a learning aid whose usefulness remained tied to explainability, critical review, privacy awareness, and continued human judgment.

Phase 2: Quantitative Results

Participant Flow

Survey invitations were sent individually to 258 medical clerkship students through faculty representatives. Of these, 238 submitted the online questionnaire, corresponding to a response rate of 92.3% among the invited students. Twenty-four submissions were excluded because the screening responses did not confirm eligibility or the main questionnaire was incomplete. The remaining 214 responses were included in the PLS-SEM analysis, representing 89.9% (218/238) of the submitted questionnaires and 82.9% (214/258) of the students

XSL•FO
RenderX

invited. The quantitative participant flow is shown in Multimedia Appendix 2, and the characteristics of the final analytic sample are presented in Table 1.

Recruitment

Recruitment and online survey data collection for the quantitative phase took place from February to March 2026. Because the survey was cross-sectional, no follow-up assessment was conducted.

Statistics and Data Analysis

Data Diagnostics

Missing Data

The SmartPLS execution log identified the input as a complete dataset. All 214 observations were used for each of the 32 indicators retained in the final measurement model. No item-level missing data were reported; therefore, missing completely at random (MCAR) testing, imputation, and other missing-data treatments were not required.

Distributions

The observed item values reported in the SmartPLS multivariate descriptives remained within the response range of their respective measures. Across the 32 retained indicators, skewness ranged from −0.938 to 0.434, and excess kurtosis ranged from −0.701 to 3.282. The Cramér-von Mises tests were statistically significant for all indicators (all $P<.001$), indicating departures from univariate normality. No normalizing transformation was applied. Model estimation was conducted using PLS-SEM, with statistical inference based on bootstrapping with 10,000 subsamples [56]. All 214 observations were retained in the analysis.

Data Summary

Descriptive statistics and correlations among the 8 construct scores are shown in Table 3. Complete construct scores were available for all 214 respondents. Because the measures used different response ranges, mean scores were interpreted within their respective scales and were not compared directly across constructs. The correlations are provided as descriptive information and do not replace the HTMT assessment reported in the “Measurement Model Estimates” section.

Table 3. Descriptive statistics and correlations among construct scores in the quantitative sample (n=214)[a].

Constructs	Mean	SD	AI-PAF[b]	AI-TCL[c]	AI-EQE[d]	AI-PCAM[e]	SEAIGU[f]	PAISCB[g]	PPS[h]	NIFRDT[i]
AI-PAF	4.109	0.418	—	0.663	0.557	0.506	0.590	0.480	0.401	0.484
AI-TCL	3.994	0.447	0.663	—	0.684	0.693	0.677	0.454	0.410	0.616
AI-EQE	4.134	0.495	0.557	0.684	—	0.658	0.686	0.547	0.327	0.711
AI-PCAM	4.079	0.457	0.506	0.693	0.658	—	0.682	0.396	0.369	0.591
SEAIGU	4.186	0.449	0.590	0.677	0.686	0.682	—	0.364	0.429	0.606
PAISCB	3.662	0.621	0.480	0.454	0.547	0.396	0.364	—	0.077	0.569
PPS	4.127	0.647	0.401	0.410	0.327	0.369	0.429	0.077	—	0.196
NIFRDT	3.924	0.502	0.484	0.616	0.711	0.591	0.606	0.569	0.196	—

[a]AI-TCL was measured on a 7-point scale, PAISCB on a 6-point scale, and the remaining constructs on 5-point scales. Ward-based infection-prevention items (NIFRDT) were reverse-coded as specified in Multimedia Appendix 4; higher scores indicate better self-reported practice. Off-diagonal values are Pearson correlations among unstandardized partial least squares construct scores. The matrix is symmetric.

[b]AI-PAF: AI-perceived algorithmic fairness.

[c]AI-TCL: AI-trust calibration literacy.

[d]AI-EQE: AI-explainability quality exposure.

[e]AI-PCAM: AI-privacy concern toward AI monitoring.

[f]SEAIGU: self-efficacy in AI-generated usage.

[g]PAISCB: perceived AI self-control balance.

[h]PPS: promotion of patient safety.

[i]NIFRDT is the construct code used for ward-based infection-prevention practice.

Model Fit

For the standardized solution, the saturated model had a standardized root mean square residual (SRMR) of 0.082 and a normed fit index (NFI) of 0.702, while the estimated model had an SRMR of 0.089 and an NFI of 0.691. The corresponding d_ULS values were 3.539 and 4.223, and the d_G values were 1.283 and 1.348, respectively. The observed SRMR, d_ULS, and d_G values were above their 99% bootstrap quantiles for both models. Thus, the bootstrap results did not support exact model fit.

Pairwise correlation residuals were not used as a separate criterion for model retention. Consistent with the predictive orientation of PLS-SEM, the model was evaluated primarily through outer loadings, construct reliability and validity, collinearity, explained variance, structural path estimates, and out-of-sample predictive performance [56]. Indicator- and construct-level measurement results, including SEAIGU5, which assessed students’ perceived ability to keep up with changes in

XSL•FO
RenderX

GenAI service features, and the borderline HTMT CIs for 3 construct pairs, are reported in the "Measurement Model Estimates" section. All findings reported below are based on the final 32-indicator model described in the "Respecification" section.

Respecification

The initial measurement model contained 35 indicators, including 6 indicators for promotion of patient safety. During the initial assessment, PPS4, PPS5, and PPS6 had low outer loadings. These indicators were removed after their contribution to indicator reliability, construct reliability, convergent validity, and conceptual coverage had been considered [56]. The removal did not materially narrow the construct because promotion of patient safety remained represented by PPS1-PPS3. In the reestimated model, this construct had a Cronbach α of 0.841, composite reliability of 0.899, and AVE of 0.748.

The final measurement model contained 32 indicators. SEAIGU5 had an outer loading of 0.543 but was retained because its loading remained above 0.40; the self-efficacy construct continued to meet the reliability and convergent-validity criteria, and the item captured students' perceived ability to keep up with changes in GenAI features [56].

No construct, hypothesis, or structural path was added, removed, or redirected during respecification. The change was confined to the measurement model and was completed before hypothesis testing, predictive assessment, moderation analysis, and importance-performance map analysis. All subsequent analyses used the final 32-indicator model. The retained indicators and final measurement-model estimates are reported in Multimedia Appendix 6.

Estimates

Measurement Model Estimates

The final measurement model was evaluated using standardized outer loadings, Cronbach α, rho_A, composite reliability, AVE, and the HTMT [56]. All retained indicators except SEAIGU5 had outer loadings above 0.70, ranging from 0.711 to 0.913. SEAIGU5 had an outer loading of 0.543 and was retained for the reasons described in the "Respecification" section.

All 8 constructs met the stated criteria for internal consistency and convergent validity. Cronbach α values ranged from 0.747 to 0.887, composite reliability values ranged from 0.840 to 0.926, and AVE ranged from 0.567 to 0.806. For promotion of patient safety, rho_A was 0.972 and exceeded its composite reliability of 0.899. Because this construct contained only 3 retained indicators, its internal consistency was interpreted mainly from Cronbach α (0.841) and composite reliability (0.899), together with an AVE of 0.748. The indicator loadings and construct-level estimates are reported in Multimedia Appendices 4 and 6.

Discriminant validity was examined using HTMT point estimates and bootstrapped 95% CIs. All point estimates were below 0.90, and none of the CIs included 1.00. The highest estimate was found between AI-trust calibration literacy and AI-perceived algorithmic fairness (HTMT=0.861, 95% CI 0.765-0.940). The upper confidence limit also exceeded 0.90 for AI-trust calibration literacy and AI-privacy concern toward AI monitoring (HTMT=0.830, 95% CI 0.736-0.909) and for ward-based infection-prevention practice and AI-explainability quality exposure (HTMT=0.845, 95% CI 0.770-0.909). These 3 pairs were therefore interpreted cautiously rather than treated as unequivocally distinct. The complete HTMT results are presented in Multimedia Appendix 5.

Structural Model Estimates

The inner variance inflation factor values ranged from 1.246 to 2.786, indicating that collinearity was not a concern in the structural model. The 4 GenAI-related antecedents explained 61.6% of the variance in self-efficacy in AI-generated usage (R^2=0.616, adjusted R^2=0.609) and 34.1% of the variance in PAISCB (R^2=0.341, adjusted R^2=0.328). The complete model explained 51.9% of the variance in ward-based infection-prevention practice (R^2=0.519, adjusted R^2=0.507).

Statistical inference was based on 10,000 bootstrap subsamples, one-tailed tests at α=.05, and 90% BC CIs. Table 4 presents the unstandardized and standardized path estimates, bootstrap SEs and CIs, exact *P* values, and decisions for the individual paths. The findings varied across the 3 hypothesis groups. Among the appraisal-mechanism paths specified in H1-H8, 5 met the prespecified directional and statistical criteria, while H7 was statistically significant in the opposite direction. Both mechanism-outcome paths specified in H9 and H10 met the prespecified criteria. Among the 2 interaction paths, H11 met the nominal criterion, whereas H12 did not.

Self-efficacy in AI-generated usage had the largest association with ward-based infection-prevention practice (β=0.477, 90% BC CI 0.355-0.604; f^2=0.327; *P*<.001), followed by PAISCB (β=0.407, 90% BC CI 0.270-0.518; f^2=0.264; *P*<.001).

XSL•FO
RenderX

Table 4. Structural model hypothesis testing results[a].

Hypothesis	Structural path	b (SE)	90% BC[b] CI for b	β (SE)	90% BC CI for β	*P* value	Decision
H1	AI-PAF[c] → SEAIGU[d]	0.188 (0.076)	0.062 to 0.312	0.185 (0.070)	0.062 to 0.294	.004	Supported
H2	AI-PAF → PAISCB[e]	0.364 (0.151)	0.137 to 0.625	0.236 (0.098)	0.066 to 0.391	.008	Supported
H3	AI-TCL[f] → SEAIGU	0.165 (0.092)	0.022 to 0.323	0.165 (0.085)	0.021 to 0.298	.03	Supported
H4	AI-TCL → PAISCB	0.032 (0.166)	−0.221 to 0.313	0.033 (0.119)	−0.169 to 0.223	.39	Not supported
H5	AI-EQE[g] → SEAIGU	0.256 (0.083)	0.124 to 0.402	0.279 (0.084)	0.136 to 0.412	<.001	Supported
H6	AI-EQE → PAISCB	0.500 (0.124)	0.305 to 0.704	0.401 (0.096)	0.242 to 0.559	<.001	Supported
H7	AI-PCAM[h] → SEAIGU	0.289 (0.083)	0.148 to 0.419	0.294 (0.081)	0.162 to 0.429	<.001	Not supported[i]
H8	AI-PCAM → PAISCB	−0.008 (0.155)	−0.305 to 0.212	−0.013 (0.108)	−0.187 to 0.168	.45	Not supported
H9	SEAIGU → NIFRDT[j]	0.543 (0.073)	0.426 to 0.663	0.477 (0.075)	0.355 to 0.604	<.001	Supported
H10	PAISCB → NIFRDT	0.329 (0.065)	0.213 to 0.422	0.407 (0.077)	0.270 to 0.518	<.001	Supported
H11	SEAIGU × PPS[k] → NIFRDT	0.214 (0.104)	0.058 to 0.376	0.122 (0.071)	0.011 to 0.237	.04	Supported
H12	PAISCB × PPS → NIFRDT	−0.118 (0.096)	−0.261 to 0.035	−0.097 (0.085)	−0.216 to 0.056	.13	Not supported

[a]SEs and 90% bias-corrected CIs were derived from 10,000 bootstrap subsamples. H9 and H10 estimates are conditional direct effects evaluated at the mean of promotion of patient safety. *P* values and decisions were based on standardized, 1-tailed bootstrap tests at α=.05. The *P* values are path-specific and unadjusted; no formal multiplicity adjustment was applied. Decisions concern individual directional hypotheses, not an omnibus test of the model, and findings near the nominal threshold should be interpreted cautiously. b is the unstandardized coefficient and β is the standardized coefficient.

[b]BC: bias-corrected.

[c]AI-PAF: AI-perceived algorithmic fairness.

[d]SEAIGU: self-efficacy in AI-generated usage.

[e]PAISCB: perceived AI self-control balance.

[f]AI-TCL: AI-trust calibration literacy.

[g]AI-EQE: AI-explainability quality exposure.

[h]AI-PCAM: AI-privacy concern toward AI monitoring.

[i]H7 was classified as not supported because the observed association was positive rather than negative as hypothesized.

[j]NIFRDT is the construct code used for ward-based infection-prevention practice.

[k]PPS: promotion of patient safety.

Specific Indirect Effects

The structural model did not include direct paths from the 4 GenAI-related antecedents to ward-based infection-prevention practice. The results are therefore reported as specific conditional indirect effects through self-efficacy in AI-generated usage and PAISCB, without assigning a mediation type. Because promotion of patient safety moderated both mediator-outcome paths, the indirect effects were evaluated at the mean level of the moderator. Inference was based on 10,000 bootstrap subsamples, 1-tailed tests at α=.05, and 90% BC CIs. Bootstrapping was used because an indirect effect is a product of path coefficients and its sampling distribution need not be normal. The *P* values reported below are from the standardized bootstrap estimates.

At the mean level of promotion of patient safety, all 4 conditional indirect effects through self-efficacy in AI-generated usage were positive and statistically significant: AI-explainability quality exposure (b=0.139, SE=0.053, 90% BC CI 0.069-0.242; β=0.133, SE=0.052, 90% BC CI 0.058-0.228; *P*=.006), AI-perceived algorithmic fairness (b=0.102, SE=0.045, 90% BC CI 0.034-0.182; β=0.088, SE=0.038, 90% BC CI 0.030-0.155; *P*=.01), AI-privacy concern toward AI monitoring (b=0.157, SE=0.046, 90% BC CI 0.088-0.242; β=0.140, SE=0.042, 90% BC CI 0.075-0.214; *P*<.001), and AI-trust calibration literacy (b=0.090, SE=0.053, 90% BC CI 0.016-0.186; β=0.079, SE=0.042, 90% BC CI 0.012-0.147; *P*=.03).

The conditional indirect effects through PAISCB were positive and statistically significant for AI-explainability quality exposure (b=0.164, SE=0.059, 90% BC CI 0.085-0.277; β=0.163, SE=0.057, 90% BC CI 0.084-0.275; *P*=.002) and AI-perceived algorithmic fairness (b=0.120, SE=0.052, 90% BC CI 0.049-0.228; β=0.096, SE=0.042, 90% BC CI 0.036-0.177; *P*=.01). The corresponding effects were not statistically significant for AI-trust calibration literacy (b=0.011, SE=0.054, 90% BC CI −0.070 to 0.097; β=0.014, SE=0.047, 90% BC CI −0.067 to 0.088; *P*=.39) or AI-privacy concern toward AI monitoring (b=−0.003, SE=0.050, 90% BC CI −0.100 to 0.070; β=−0.005, SE=0.043, 90% BC CI −0.076 to 0.065; *P*=.45).

XSL•FO
RenderX

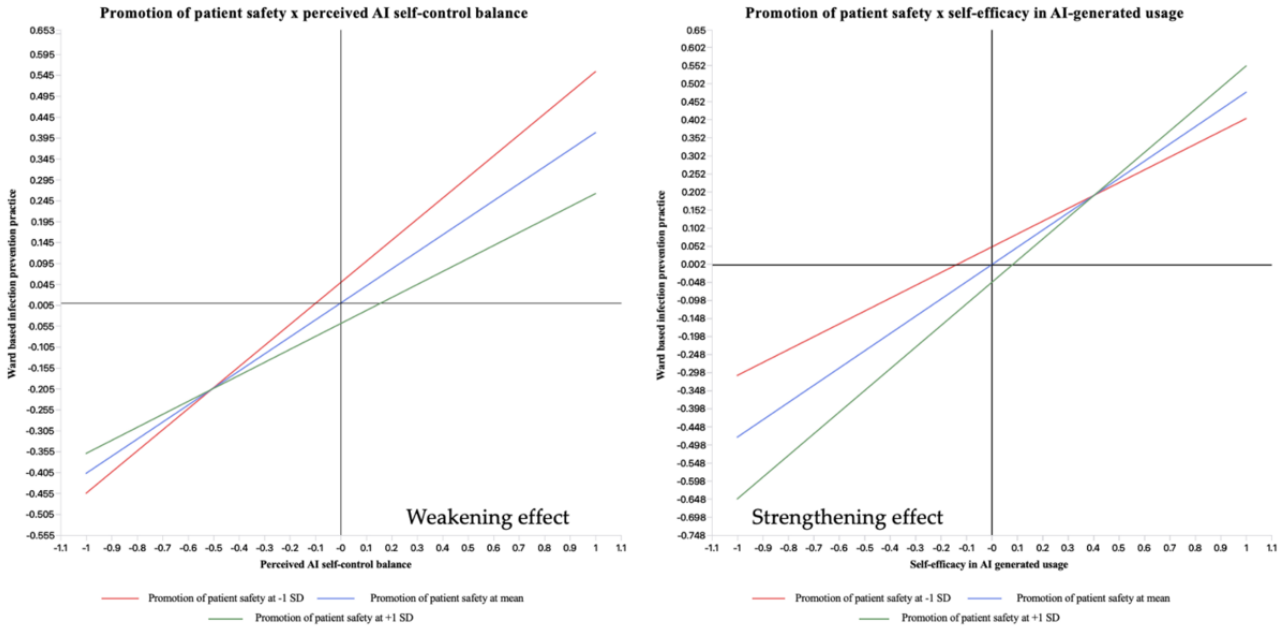
Interaction Effects

Promotion of patient safety was examined as a moderator of the associations between the 2 human-centered constructs and ward-based infection-prevention practice. Its constituent main effect was retained in the outcome equation but was not statistically significant (β=−0.049, 90% BC CI −0.149 to 0.041; *P*=.20).

The interaction between self-efficacy in AI-generated usage and promotion of patient safety was small and positive (β=0.122, 90% BC CI 0.011-0.237; f^2=0.028; *P*=.04), meeting the prespecified nominal criterion for H11. Given its proximity to the significance threshold and small effect size, this result was interpreted cautiously. The conditional association between self-efficacy and ward-based infection-prevention practice increased from β=0.355 at a low level of promotion of patient safety (−1 SD; 90% BC CI 0.167-0.562), to β=0.477 at its mean (90% BC CI 0.355-0.604), and β=0.599 at a high level (+1 SD; 90% BC CI 0.460-0.723). The simple slopes therefore followed the direction of the positive interaction.

The interaction between PAISCB and promotion of patient safety was negative but not statistically significant (β=−0.097, 90% BC CI −0.216 to 0.056; f^2=0.018; *P*=.13), and H12 was not supported. The corresponding conditional coefficients were β=0.504 at −1 SD (90% BC CI 0.229-0.675), β=0.407 at the mean (90% BC CI 0.270-0.518), and β=0.311 at +1 SD (90% BC CI 0.185-0.443). Although the coefficients became smaller as promotion of patient safety increased, this pattern should not be interpreted as evidence of moderation because the interaction term was not statistically significant. The interaction estimates are included in Table 4, and both simple-slope patterns are shown in Figure 1.

Figure 1. Simple slope plots showing the moderating role of promotion of patient safety in the associations between self-efficacy in AI-generated usage, perceived AI self-control balance, and ward-based infection-prevention practice.



Predictive Assessment

All 3 endogenous constructs had positive Q^2predict values. The value was 0.588 for self-efficacy in AI-generated usage, 0.298 for PAISCB, and 0.486 for ward-based infection-prevention practice. These results indicated that the PLS-SEM predictions performed better than the corresponding naïve prediction benchmark for each construct [56].

Predictive performance was further examined using the CVPAT. Compared with the indicator average benchmark, the PLS-SEM model produced lower loss values for ward-based infection-prevention practice (average loss difference=−0.123, *P*<.001), PAISCB (average loss difference=−0.117, *P*=.001), and self-efficacy in AI-generated usage (average loss difference=−0.118, *P*<.001). The overall average loss difference was −0.119 (*P*<.001), indicating better predictive performance than the indicator average benchmark.

The results were less consistent when the PLS-SEM model was compared with the linear model benchmark. PLS-SEM produced lower loss values for ward-based infection-prevention practice (average loss difference=−0.038, *P*=.02) and self-efficacy in AI-generated usage (average loss difference=−0.021, *P*=.02). For PAISCB, however, the PLS-SEM loss was higher than the linear model loss (average loss difference=0.075, *P*=.004). At the overall-model level, the loss difference was 0.009 (*P*=.37), indicating no statistically significant difference between PLS-SEM and the linear model benchmark.

Taken together, the model showed predictive usefulness relative to the indicator average benchmark but did not consistently outperform the linear model benchmark. The construct-specific result for PAISCB was therefore considered when interpreting findings involving this construct [56,57,60]. The complete CVPAT results are presented in Table 5.

XSL•FO
RenderX

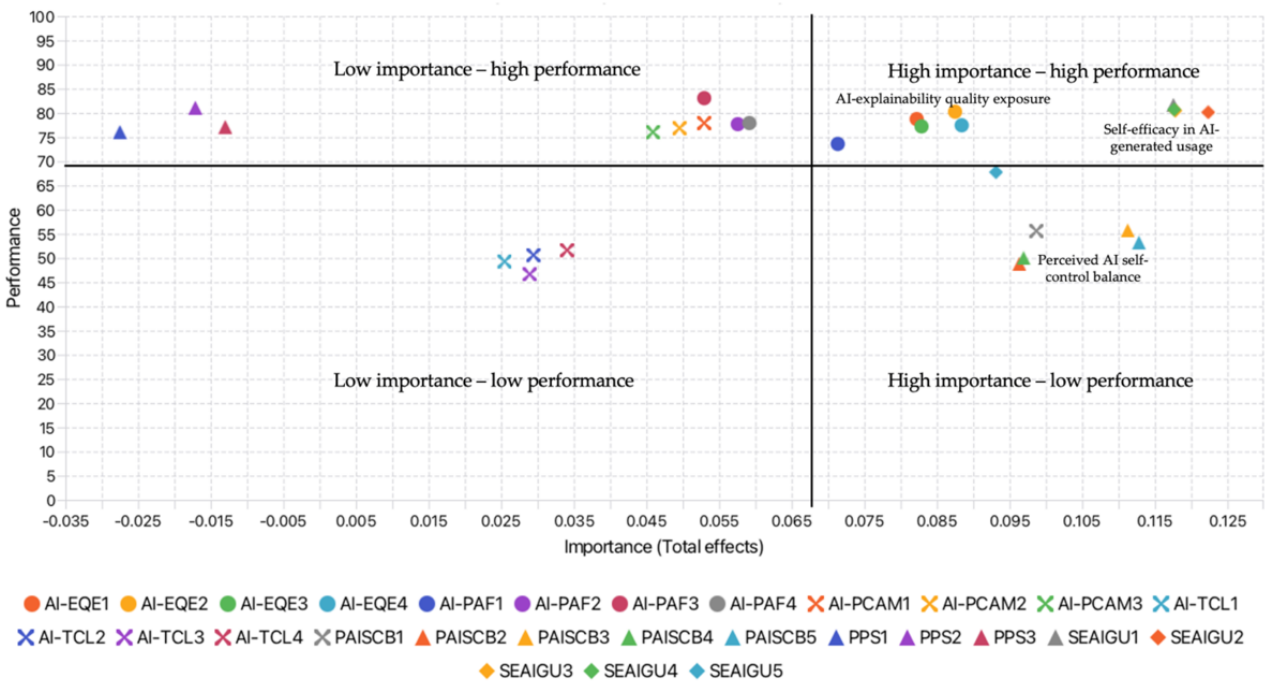
Table 5. Construct-specific and overall predictive performance of the partial least squares structural equation modeling (PLS-SEM) model compared with indicator average (IA) and linear model (LM) benchmarks using the Cross-Validated Predictive Ability Test (CVPAT).

Variable	PLS-SEM vs IA				PLS-SEM vs LM			
	PLS loss	IA loss	Average loss difference	*P* value	PLS loss	LM loss	Average loss difference	*P* value
Ward-based infection-prevention practice	0.365	0.488	–0.123	<.001	0.365	0.403	–0.038	.02
Perceived AI self-control balance	0.542	0.658	–0.117	.001	0.542	0.466	0.075	.004
Self-efficacy in AI-generated usage	0.202	0.320	–0.118	<.001	0.202	0.223	–0.021	.02
Overall	0.370	0.489	–0.119	<.001	0.370	0.361	0.009	.37

Importance-Performance Map Analysis

Importance-performance map analysis was conducted to identify indicators that were both relevant to ward-based infection-prevention practice and actionable for clerkship education. Following Ringle and Sarstedt [58], importance was interpreted from the total effects in the PLS-SEM model, while performance was interpreted from rescaled indicator mean scores. The response scale length for each construct is reported in Multimedia Appendix 3 to support transparency in performance rescaling. The IPMA results are shown in Figure 2.

Figure 2. Importance-performance map analysis of indicators associated with ward-based infection-prevention practice among medical clerkship students from 5 private universities in the Jakarta Metropolitan Area, Indonesia. AI-EQE: AI-explainability quality exposure; AI-PAF: AI-perceived algorithmic fairness; AI-PCAM: AI-privacy concern toward AI monitoring; AI-TCL: AI-trust calibration literacy; PAISCB: perceived AI self-control balance; PPS: promotion of patient safety; SEAIGU: self-efficacy in AI-generated usage.



The indicator-level IPMA results are shown in Figure 2. All 4 indicators of AI-explainability quality exposure were located in the high-importance and high-performance area. Four of the 5 self-efficacy indicators (SEAIGU1-SEAIGU4) showed the same pattern, whereas SEAIGU5 had high importance but performance slightly below the reference value. These findings suggest that explainability quality exposure and most aspects of self-efficacy were relative strengths among the respondents.

All 5 indicators of PAISCB were located in the high-importance and lower-performance area, making this construct the most consistent priority area in the map. The AI-trust calibration literacy indicators showed comparatively low importance and low performance. The AI-privacy concern indicators and 3 of the 4 AI-perceived algorithmic fairness indicators had lower importance but relatively high performance, while AI-PAF1 was located in the high-importance and high-performance area.

The promotion of patient safety indicators showed high performance but low or negative importance. Because promotion of patient safety was specified as a moderator rather than a direct predictor, these indicator positions were not interpreted as direct targets for educational intervention. The priority identified for PAISCB should also be interpreted cautiously because the

XSL•FO
RenderX

construct-specific CVPAT showed that the PLS-SEM model did not outperform the linear benchmark for this construct.

Integration of Qualitative and Quantitative Findings

Convergence, Expansion, and Divergence

At the second stage of integration, following the quantitative analysis, the qualitative themes in Table 2 were compared with the structural model results in Table 4. This comparison examined whether the interview accounts converged with, added context to, or differed from the quantitative findings.

The clearest convergence concerned explainability. Participants were more confident using GenAI when its responses included understandable reasoning and were more hesitant when answers were presented without justification. In the quantitative phase, explainability quality exposure was positively associated with self-efficacy in AI-generated usage (β=0.279, *P*<.001) and PAISCB (β=0.401, *P*<.001). Fairness showed a similar pattern. Participants valued the apparent objectivity of GenAI but also questioned whether its outputs adequately represented local patients and clinical settings. Perceived algorithmic fairness was positively associated with both self-efficacy (β=0.185, *P*=.004) and PAISCB (β=0.236, *P*=.008).

The qualitative findings added context to the more selective result for trust calibration literacy. Participants described checking AI-generated output against supervisors' opinions, clinical reasoning, and actual patient cases. Some also reported learning from incorrect AI responses. In the quantitative phase, trust calibration literacy was associated with self-efficacy (β=0.165, *P*=.03), but not with PAISCB (β=0.033, *P*=.39). The ability to question and verify AI output therefore corresponded with greater confidence in using GenAI, but not necessarily with a stronger perception of control over the interaction.

Privacy concern showed the clearest divergence from the hypothesized direction. Participants expressed concern about patient-data leakage, unauthorized data reuse, and monitoring, but some continued to use GenAI because it remained useful for clinical and academic learning. Privacy concern was positively associated with self-efficacy (β=0.294, *P*<.001), contrary to the negative association proposed in H7, and was not associated with PAISCB (β=−0.013, *P*=.45). Thus, the two phases were not contradictory. Both showed that privacy concern could coexist with continued and confident GenAI use; the divergence concerned the direction proposed in H7.

The findings concerning human agency also showed convergence with some qualification. Participants described gaining confidence through repeated use while maintaining that final judgment should remain with the learner and supervisor. Quantitatively, self-efficacy (β=0.477, *P*<.001) and PAISCB (β=0.407, *P*<.001) were positively associated with self-reported ward-based infection-prevention practice. The specific indirect effects further showed that self-efficacy carried the associations of all 4 GenAI-related antecedents, whereas PAISCB carried only the associations of explainability and fairness. A small positive interaction was observed between promotion of patient safety and self-efficacy, whereas the corresponding interaction involving PAISCB was not observed, as shown in Figure 1. The interview findings provided context for the importance of safety orientation and supervision, but were not treated as independent evidence of statistical moderation.

Discussion

Principal Findings

This study examined how medical clerkship students used and interpreted GenAI in a safety-critical learning environment. The qualitative phase identified 4 GenAI-related domains that shaped students' learning experiences: perceived algorithmic fairness, trust calibration literacy, explainability quality exposure, and privacy concern toward AI monitoring. These domains reflected how students appraised GenAI when using it alongside clinical learning, documentation, and patient safety–related tasks during ward rotations. In the quantitative phase, these domains were examined as antecedents of 2 human-centered mechanisms, namely self-efficacy in AI-generated usage and PAISCB, and their associations with self-reported ward-based infection-prevention practice. Self-efficacy carried significant indirect effects from all 4 antecedents, whereas PAISCB carried only those of explainability and fairness. This asymmetry indicates that self-efficacy was the more consistent pathway linking students' appraisals of GenAI to self-reported infection-prevention practice. The two phases follow a single social cognitive logic, in which students' appraisals of GenAI relate to safety-oriented practice mainly through their beliefs about their own capabilities rather than through acceptance alone [24]. GenAI-supported patient-safety learning in clerkship depends on access to AI tools and on how students understand, verify, and regulate AI-generated output within supervised workplace-based learning [2,4,7,8,39].

Among the GenAI-related antecedents, explainability quality exposure showed the most consistent association with both self-efficacy in AI-generated usage and PAISCB. Perceived algorithmic fairness was also positively associated with both mediators. The qualitative interviews showed the same pattern. Several participants described GenAI as useful when they could understand how an answer was produced, especially when preparing subjective-objective-assessment-plan (SOAP) notes, revising documentation, or organizing oral case presentations during ward learning. In this context, explainability functioned as a practical learning support that helped students compare AI-generated output with clinical expectations instead of accepting it as a final answer [30-32]. Fairness also appeared to support engagement with GenAI. When AI outputs were perceived as reasonable, consistent, and procedurally acceptable, students were more willing to use them during learning tasks related to clinical reasoning and patient safety. However, perceived fairness should not be equated with actual freedom from bias, particularly because AI-generated outputs may remain sensitive to training data, task context, and automation-related overconfidence [22,23,25,26].

The findings for trust calibration literacy were more selective. Trust calibration literacy was associated with self-efficacy in AI-generated usage but was not meaningfully associated with PAISCB. This distinction indicates that knowing when to question AI may strengthen students' confidence in using GenAI, but does not automatically create a stronger sense of

XSL•FO
RenderX

control over the interaction. In clerkship settings, students may understand the need to verify AI output while still having limited time, feedback, or structured opportunities to practice verification during ward routines. Research on automation bias and cognitive forcing supports this interpretation. Reflective engagement requires task design, supervision, and feedback that help learners decide whether AI output should be accepted, revised, or rejected [22,23,27-29,61].

The privacy concern finding also requires careful interpretation. Privacy concern toward AI monitoring was positively associated with self-efficacy in AI-generated usage, although the hypothesized direction was negative. The finding was therefore directionally reversed rather than statistically null. One possible explanation is risk-aware engagement. Students with greater privacy concerns may have paid closer attention to the information they entered, how prompts were framed, and when GenAI use was inappropriate. The interviews provide some context for this interpretation. Several students described feeling monitored by AI systems. They expressed concern about the responsible use of patient-related data (Table 2), yet these concerns did not lead them to discontinue GenAI use. Instead, they described using it more carefully. Privacy paradox research similarly suggests that people may continue to use digital tools despite privacy concerns when the perceived benefits are immediate and practical [33,34,62,63]. Other explanations cannot be ruled out. Because privacy concern and self-efficacy were measured in the same self-report survey, the observed association may partly reflect shared method variance or a general tendency to endorse questionnaire items. The AI-PCAM items may also have reflected privacy awareness among students who were already more confident in using GenAI, rather than concern that discouraged its use. The cross-sectional design cannot distinguish among these possibilities. For clerkship education, privacy concern may become educationally useful when it is connected to data governance, professional accountability, and supervised guidance on safe AI use [4,8,9].

The findings involving PAISCB should be interpreted within the scope of the construct and its measurement. PAISCB was positively associated with self-reported ward-based infection-prevention practice. However, the construct-specific CVPAT showed that the PLS-SEM model did not outperform the linear benchmark for this construct [56,57,60]. The items captured students' perceived control over GenAI interaction [38]. This construct was associated with self-reported infection-prevention practice but did not measure clinical oversight or accountability; safe use still depends on supervision, protocol alignment, and clear boundaries of responsibility [4,8,22,23,39].

Across phases, the findings converged for explainability and fairness, expanded the interpretation of trust calibration, and helped explain the positive association between privacy concern and self-efficacy. Because the qualitative findings informed development of the quantitative framework, these correspondences were interpreted as cross-phase coherence rather than independent confirmation [45]. Together, the findings position GenAI as a supervised educational practice linked to patient-safety learning rather than technology acceptance alone [2,4,6,9,64].

Theoretical Implications

This study contributes to the theoretical understanding of GenAI-supported learning in medical clerkship by positioning self-efficacy and PAISCB as human-centered mechanisms, rather than as general indicators of technology acceptance. The findings are organized around a single social cognitive frame, in which students' appraisals of AI-generated output relate to learning behavior mainly through their beliefs about their own capabilities rather than through acceptance alone [24]. Within this frame, self-efficacy reflects students' belief in their capacity to organize and perform learning tasks with AI support [24], and it emerged as the mechanism most clearly linked to self-reported infection-prevention practice [65,66].

The reasoning behind this link can be stated more precisely. Infection-prevention practice on the ward is not a single decision but a set of repeated, effortful routines carried out under supervision and time pressure, including hand hygiene, use of protective equipment, and documentation [3,39]. Social cognitive theory holds that self-efficacy influences whether learners initiate such effortful actions, how much effort they sustain, and how they persist when a task is demanding [24]. A student may use AI to clarify a precaution, check a protocol, or prepare a case. When the output is verified against clinical findings and supervisor guidance, that support is more likely to inform consistent practice. Avoiding AI or relying on it without checking offers less support for such practice. In theoretical terms, the contribution of self-efficacy operates less through the technology itself than through the learner's willingness to translate AI support into deliberate, verified action within supervised work. This extends self-efficacy theory into a clinical education context in which learners use AI assistance while still working under supervision, clinical uncertainty, and patient-safety expectations.

Within the same agentic view, the findings refine the theoretical role of autonomy in GenAI-supported learning. Self-determination theory serves here as a complementary lens, emphasizing autonomy and competence as psychological conditions for learning [67]. PAISCB can be viewed as a learning-related form of agency in human-AI interaction [38]. It does not imply independent control over patient-facing decisions, which remain dependent on supervision, verification, and institutional safeguards.

These individual-level mechanisms operate within a wider educational context, which workplace-based learning and competency-based medical education help to specify. Workplace-based learning views clinical education as participation in real clinical work, where supervision, feedback, and graded responsibility shape how students develop professional competence [39,68,69]. Competency-based medical education similarly emphasizes progression toward defined competencies and readiness for safe practice rather than learning based only on time spent in training [70]. By examining self-reported ward-based infection-prevention practice as a safety-oriented learning outcome, this study links GenAI-related perceptions to the practical routines of clerkship learning. GenAI is positioned as a tool whose educational value depends on how

XSL•FO
RenderX

it is embedded within ward supervision, feedback, protocol use, and patient-safety routines.

Complementing this individual-level account, the findings also support a situated view of GenAI in medical education. From this perspective, learning is shaped by individual cognition and by interactions among learners, supervisors, digital tools, institutional protocols, and clinical tasks [71]. Explainability and perceived fairness are not solely technical attributes of AI systems. In clerkship learning, they also function as sociocognitive resources that influence whether students are willing to bring AI-generated output into documentation, case preparation, and infection-prevention reasoning. This framing helps explain why the same AI tool may support learning in one setting but encourage superficial use in another, depending on how the tool is situated within supervision and local clinical practice.

The pattern of supported and unsupported paths provides a boundary condition for theories of human-AI learning. Trust calibration literacy was associated with self-efficacy but not with PAISCB, and general promotion of patient safety did not strengthen the association between PAISCB and ward-based infection-prevention practice. The results suggest that awareness of AI limitations and broad safety messages may be insufficient unless they are translated into supervised verification practices. Research on automation bias, cognitive forcing, and the cognitive atrophy paradox supports this interpretation. Intelligent systems may improve efficiency while reducing active cognitive engagement when users offload reasoning to AI [22,23,61,72]. In clerkship education, accounts of GenAI-supported learning should include the educational conditions that preserve students' reasoning, verification, and professional responsibility.

Overall, GenAI is best understood as part of a supervised learning ecology whose educational value depends on verification, feedback, and institutional safeguards.

Educational and Patient-Safety Implications

The IPMA placed PAISCB, rather than self-efficacy as a whole, as the main area with room for improvement. All PAISCB indicators combined high importance with lower performance, whereas 4 of the 5 self-efficacy indicators and all AI-explainability quality exposure indicators showed high importance and high performance. The more immediate educational need may therefore be to help students retain control over their interaction with GenAI—for example, by setting clear goals, deciding how prompts are used, and checking AI-generated outputs—rather than focusing only on confidence in using the technology. This interpretation remains cautious because the PLS-SEM model did not outperform the linear benchmark for perceived AI self-control balance in the construct-specific CVPAT [56,58,60].

GenAI should be introduced as a provisional learning support, not as a substitute for clinical judgment. Training in verification, patient-data protection, and appropriate GenAI use should begin early in medical education and be reinforced during clerkship. Students can use GenAI to organize infection prevention and control guidance, summarize isolation criteria, clarify the rationale behind precautions, or prepare an initial interpretation before ward discussion. The educational value of these activities depends on verification against patient findings, institutional protocols, and supervisor feedback [70,73]. Case discussions provide a practical setting for this process. Supervisors should ask how students used GenAI. They should also ask which outputs were accepted or rejected and how the final interpretation was checked against clinical and infection-prevention standards [70]. Educational activities can translate this form of control into observable practices, including prompt formulation, source checking, comparison with institutional protocols, supervisor confirmation, and reflection on uncertainty [70,73].

The comparatively strong performance of explainability quality exposure indicates that many students already value AI outputs that are understandable and usable for learning. Explainability is insufficient when it remains disconnected from supervised clinical behavior [30,58]. Medical schools and teaching hospitals should develop clerkship guidance for GenAI use, including patient-data protection rules, examples of acceptable and unacceptable prompts, structured verification steps, and supervisor-led reflection on AI-generated outputs. This guidance would embed GenAI-supported learning in patient-safety education and connect infection-prevention learning with clinical accountability and institutional standards [73-76].

Limitations and Future Research

This study should be interpreted in light of several limitations. First, the quantitative phase used a cross-sectional design; therefore, the observed associations cannot be interpreted as causal relationships. The sequential mixed methods design allowed the findings from both phases to be considered together; however, the survey data captured students' perceptions and self-reported practices at one point in time. The reported model was retained as the primary model because its structure was developed through the sequential integration of the qualitative findings with social cognitive theory and prior evidence and was finalized before quantitative data collection and analysis. It was therefore not selected after comparing alternative arrangements against the observed quantitative results. Alternative arrangements of the constructs were not estimated. The findings support the reported associations within the specified model but do not establish that it was the only plausible representation of the observed relationships. In addition, the structural hypotheses were evaluated as directional paths using *P* values and CIs, without a formal adjustment for the number of paths tested; results close to the significance threshold warrant particular caution. Future studies should use longitudinal or interventional designs to examine whether structured GenAI training, supervised verification exercises, or infection-prevention learning modules produce sustained changes in students' AI-related self-efficacy and ward-based patient-safety learning.

Second, the study relied on self-reported ward-based infection-prevention practice rather than direct observation, supervisor ratings, or clinical infection indicators. This approach was appropriate for examining students' perceived learning and reported practice during clerkship, but it remains subject to recall bias and social desirability bias. Procedural steps were

XSL•FO
RenderX

taken to reduce common method bias, including anonymity, screening procedures, and clear response instructions; however, residual common method bias cannot be fully excluded in self-report survey research [55]. Future research should combine student-reported outcomes with observed behaviors, supervisor assessments, simulation-based performance, or audit-based indicators of infection-prevention practice.

Third, the study was conducted among medical clerkship students from 5 private universities in the Jakarta Metropolitan Area using purposive sampling. This sampling strategy ensured that respondents had direct experience using GenAI in clinical learning, but it also limits the generalizability of the findings to other medical schools, public universities, rural training sites, or clerkship systems with different supervision models and institutional policies. The qualitative phase also included 8 informants, which was appropriate for contextual exploration based on information power and the contextual range and depth of the interviews, but was not intended to represent the full diversity of medical clerkship experiences. Future studies should include larger and more diverse samples across public and private institutions, multiple regions, and different types of teaching hospitals. In addition, because participation required prior frequent or very frequent GenAI use, self-selection bias may have occurred, as students who were more digitally engaged or more interested in GenAI-supported learning may have been more likely to participate. Accordingly, the quantitative estimates should be read as conditional on this group of frequent GenAI users rather than as estimates for the broader clerkship population. Restricting the sample in this way may also affect the magnitude of the observed associations, although the direction of this effect, whether attenuation or strengthening, cannot be determined from the present data. Although 238 of the 258 individually invited students submitted the survey, the available data did not permit a comparison between respondents and nonrespondents; some nonresponse bias therefore remains possible.

Fourth, perceived AI self-control balance did not measure supervised verification, human accountability, protocol alignment, or clinical responsibility [38]. Its construct-specific CVPAT also showed less consistent predictive performance than that of self-efficacy [56,57,60]. Future studies should refine this construct by developing or adapting items that more directly capture supervised verification, human accountability, protocol alignment, and clinical responsibility in AI-supported patient-safety learning.

Fifth, the mixed methods integration has limitations that qualify the interpretation of the correspondences between the two phases. Although the themes were developed inductively, the interview guide was informed by the study objectives, theory, and prior literature, and the subsequent abductive building stage considered the themes alongside established theoretical constructs. This process may have made the analysis more attentive to patterns that corresponded with the selected theories and measures. A formal discrepant-case analysis was not conducted, although contrasting participant accounts and quantitative paths that were unsupported or opposite to the hypothesized direction were retained in the interpretation. Because the qualitative findings contributed to the development of the quantitative framework, agreement across phases should be understood as theory-informed coherence rather than independent confirmation. Future studies could strengthen mixed methods legitimacy through explicit joint displays, formal discrepant-case analysis, and further examination of explanations that do not fit the initial theoretical framework [45,49].

Sixth, although none of the bootstrapped HTMT CIs included 1.00, the upper limits exceeded 0.90 for 3 construct pairs. This finding indicates some empirical overlap in the present sample, particularly between AI-trust calibration literacy and AI-perceived algorithmic fairness. These constructs were retained because they address different aspects of students' appraisal of GenAI, but findings involving the borderline pairs should be interpreted cautiously. Further studies should reassess their empirical separation in more diverse clerkship populations. AI-trust calibration literacy concerns how students evaluate and calibrate their reliance on AI-generated conclusions, whereas AI-perceived algorithmic fairness concerns judgments of procedural fairness and nondiscrimination in AI-based services.

Finally, heterogeneity in clerkship environments may have influenced the observed relationships. Students rotated through different hospitals, clinical departments, patient case mixes, supervision styles, and infection-prevention policy environments. This variation reflects the real-world nature of workplace-based medical education, but it also introduces contextual differences that were not fully accounted for in the present model. Because university identifiers were not retained in the respondent-level dataset, clustering by institution and differences among the 5 participating universities could not be examined. Future research should retain nonidentifying institutional codes and examine whether institutional support, supervisor readiness, clinical workload, and department-level safety culture influence the educational value of GenAI during clerkship. Such studies may help clarify how GenAI can be integrated into patient-safety learning without weakening students' clinical reasoning, verification habits, or professional responsibility.

Conclusion

In the interviews, students valued GenAI when its responses were understandable and could be checked against clinical knowledge and supervisor guidance, while also raising concerns about bias, privacy, and the need to retain human judgment. In the quantitative phase, explainability quality exposure and perceived algorithmic fairness were associated with students' self-efficacy and perceived control in using GenAI. Self-efficacy was the most consistent pathway linking all 4 antecedents to self-reported ward-based infection-prevention practice, whereas perceived AI self-control balance carried only the effects of explainability and fairness. The positive association between privacy concern and self-efficacy suggests risk-aware engagement, although the cross-sectional, self-reported findings do not establish causal or clinical effects. The focus on medical clerkship students and ward-based infection-prevention practice extends research that has largely examined GenAI adoption, acceptance, academic support, learning engagement, or clinical reasoning [2,4,6,9,64]. Medical schools and teaching hospitals should therefore treat GenAI as a supervised learning support,

XSL•FO
RenderX

with patient-data protection, protocol alignment, verification of AI-generated outputs, and supervisor-led reflection incorporated into clerkship guidance. Perceived control over AI interaction should not be interpreted as evidence of clinical oversight or accountability. Future longitudinal, interventional, and multisite studies should assess observed behavior and supervisor-rated performance.

Acknowledgments

The authors would like to thank all participating medical clerkship students for their time and contribution to this study. The authors also acknowledge the support of colleagues and faculty representatives who assisted in distributing the study information and questionnaire.

The authors declare the use of generative AI (GenAI) in the research and writing process. According to the GAIDeT (Generative AI Delegation taxonomy; 2025), the following tasks were delegated to GenAI under full human supervision: proofreading and editing, and adapting and adjusting emotional tone. The GenAI tool used was ChatGPT. Responsibility for the final manuscript lies entirely with the authors. GenAI tools are not listed as authors and do not bear responsibility for the final outcomes. The declaration was submitted by all authors.

Data Availability

The data generated and analyzed during this study are not publicly available because they contain participant responses from medical clerkship students and were collected under institutional ethics approval with confidentiality safeguards. Deidentified data may be made available from the corresponding author upon reasonable request and subject to ethics approval and institutional data-sharing requirements.

Funding

The authors declared that no financial support was received for this work.

Authors' Contributions

GG contributed to conceptualization, methodology, investigation, formal analysis, project administration, resources, visualization, writing of the original draft, and review and editing of the manuscript. FA contributed to conceptualization, methodology, project administration, supervision, validation, and review and editing of the manuscript. DNA contributed to data curation, validation, visualization, and supporting the writing of the original draft. All authors reviewed and approved the final manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
The conceptual framework of the hypothesized associations among generative AI–related perceptions, self-efficacy in AI-generated usage, perceived AI self-control balance, promotion of patient safety, and self-reported ward-based infection-prevention practice.
[PNG File , 253 KB - mededu_v12i1e100828_app1.png]

Multimedia Appendix 2
Participant flow through the sequential exploratory mixed methods study. Eight students completed the qualitative phase, 238 responses were received in the quantitative phase, and 214 responses were included in the final partial least squares structural equation modeling analysis.
[PNG File , 135 KB - mededu_v12i1e100828_app2.png]

Multimedia Appendix 3
Definitions and response scales for the 8 constructs included in the quantitative model.
[DOCX File , 46 KB - mededu_v12i1e100828_app3.docx]

Multimedia Appendix 4
Measurement items, outer loadings, descriptive statistics, internal consistency estimates, and average variance extracted for the eight constructs examined in the partial least squares structural equation modeling analysis.
[DOCX File , 29 KB - mededu_v12i1e100828_app4.docx]

Multimedia Appendix 5

XSL•FO
RenderX

Each cell reports the heterotrait-monotrait ratio of correlations point estimate followed by its bootstrapped 95% CI. None of the CIs included 1.00. Upper confidence limits exceeded 0.90 for AI-trust calibration literacy (AI-TCL)–AI-perceived algorithmic fairness and reached 0.909 for AI-TCL–AI-privacy concern toward AI monitoring and NIFRDT–AI-explainability quality exposure; these pairs are discussed as borderline in the Results.
[DOCX File , 16 KB - mededu_v12i1e100828_app5.docx]

Multimedia Appendix 6
Partial least squares structural equation modeling measurement model of generative AI–related factors, human-centered mediators, promotion of patient safety, and ward-based infection-prevention practice. Values beside the indicators are outer loadings, and values within the constructs are Cronbach α coefficients.
[PNG File , 282 KB - mededu_v12i1e100828_app6.png]

Multimedia Appendix 7
Structural model of generative AI–related factors, human-centered mediators, promotion of patient safety, and ward-based infection-prevention practice among medical clerkship students in Jakarta Metropolitan Area, Indonesia.
[PNG File , 315 KB - mededu_v12i1e100828_app7.png]

References

1. Liu X, Zhong B. Integrating generative artificial intelligence into student learning: a systematic review from a TPACK perspective. Educ Res Rev 2025 Nov;49:100741. [doi: 10.1016/j.edurev.2025.100741]
2. Lin Y, Luo Z, Ye Z, Zhong N, Zhao L, Zhang L, et al. Applications, challenges, and prospects of generative artificial intelligence empowering medical education: scoping review. JMIR Med Educ 2025 Oct 23;11:e71125 [FREE Full text] [doi: 10.2196/71125] [Medline: 41128430]
3. Yang X, Wu Q, Yang H, Deng Y, Chen F, An Y, et al. Comparative evaluation of competency-based and real-patient education in undergraduate clinical clerkships: insights from student feedback and performance. BMC Med Educ 2025 Nov 26;25(1):1761 [FREE Full text] [doi: 10.1186/s12909-025-08323-z] [Medline: 41299399]
4. Rodger D, Mann SP, Earp B, Savulescu J, Bobier C, Blackshaw BP. Generative AI in healthcare education: how AI literacy gaps could compromise learning and patient safety. Nurse Educ Pract 2025 Aug;87:104461 [FREE Full text] [doi: 10.1016/j.nepr.2025.104461] [Medline: 40633198]
5. Mondal H, Karri JKK, Ramasubramanian S, Mondal S, Juhi A, Gupta P. A qualitative survey on perception of medical students on the use of large language models for educational purposes. Adv Physiol Educ 2025 Mar 01;49(1):27-36 [FREE Full text] [doi: 10.1152/advan.00088.2024] [Medline: 39447120]
6. Tran C, Hryciw BN, Moore SW, Chaput A, Seely AJE. Perceptions and use of generative artificial intelligence in medical students: a multicenter survey. J Med Educ Curric Dev 2025;12:23821205251391969 [FREE Full text] [doi: 10.1177/23821205251391969] [Medline: 41181167]
7. Pallant JL, Blijlevens J, Campbell A, Jopp R. Mastering knowledge: the impact of generative AI on student learning outcomes. Stud High Educ 2025 May 06;51(4):714-735. [doi: 10.1080/03075079.2025.2487570]
8. Mehta N, Mehta S, Rubenstein A, Wood SK. Not replaced, but reinvented: AI education pathways to prepare future physicians to lead healthcare transformation. Perspect Med Educ 2025;14(1):849-859 [FREE Full text] [doi: 10.5334/pme.2233] [Medline: 41278578]
9. Ichikawa T, Olsen E, Vinod A, Glenn N, Hanna K, Lund GC, et al. Generative artificial intelligence in medical education-policies and training at US osteopathic medical schools: descriptive cross-sectional survey. JMIR Med Educ 2025;11:e58766 [FREE Full text] [doi: 10.2196/58766] [Medline: 39934984]
10. Handbook for National Quality Policy and Strategy: A Practical Approach for Developing Policy and Strategy to Improve Quality of Care. Geneva: World Health Organization; 2018.
11. Global Patient Safety Action Plan 2021-2030: Towards Eliminating Avoidable Harm in Health Care. Geneva: World Health Organization; 2021.
12. El Arab RA, Almoosa Z, Alkhunaizi M, Abuadas FH, Somerville J. Artificial intelligence in hospital infection prevention: an integrative review. Front Public Health 2025;13:1547450. [doi: 10.3389/fpubh.2025.1547450] [Medline: 40241963]
13. Odoom A, Tetteh-Quarcoo PB, Donkor ES. Prevalence of hospital-acquired infections in low- and middle-income countries: systematic review and meta-analysis. Asia Pac J Public Health 2025;37(5):448-466. [doi: 10.1177/10105395251338002] [Medline: 40401334]
14. Raoofi S, Pashazadeh Kan F, Rafiei S, Hosseinipalangi Z, Noorani Mejareh Z, Khani S, et al. Global prevalence of nosocomial infection: a systematic review and meta-analysis. PLoS One 2023;18(1):e0274248 [FREE Full text] [doi: 10.1371/journal.pone.0274248] [Medline: 36706112]
15. Gastaldi S, Tartari E, Satta G, Allegranzi B. Advancing infection prevention and control through artificial intelligence: a scoping review of applications, barriers, and a decision-support checklist. Antimicrob Steward Healthc Epidemiol 2025;5(1):e317. [doi: 10.1017/ash.2025.10191] [Medline: 41446414]

XSL•FO
RenderX

16. Kim KJ. Medical student needs for e-learning: perspectives of the generation Z. Korean J Med Educ 2024;36(4):389-399 [FREE Full text] [doi: 10.3946/kjme.2024.312] [Medline: 39639634]
17. Abdelhafiz AS, Farghly MI, Sultan EA, Abouelmagd ME, Ashmawy Y, Elsebaie EH. Medical students and ChatGPT: analyzing attitudes, practices, and academic perceptions. BMC Med Educ 2025;25(1):187 [FREE Full text] [doi: 10.1186/s12909-025-06731-9] [Medline: 39910593]
18. Tajima H, Kasai H, Shikino K, Shimizu I, Ito S. Perceptions and intentions to use generative AI among first-year medical students in Japan: cross-sectional survey study. JMIR Med Educ 2025;11:e77552 [FREE Full text] [doi: 10.2196/77552] [Medline: 41232013]
19. Shorey S, Chan V, Rajendran P, Ang E. Learning styles, preferences and needs of generation Z healthcare students: scoping review. Nurse Educ Pract 2021;57:103247. [doi: 10.1016/j.nepr.2021.103247] [Medline: 34768214]
20. Davis FD. Perceived usefulness, perceived ease of use, and user acceptance of information technology. MIS Quarterly 1989;13(3):319. [doi: 10.2307/249008]
21. Lee JWY, Tan JY, Bello F. Technology acceptance model in medical education: systematic review. JMIR Med Educ 2025;11:e67873 [FREE Full text] [doi: 10.2196/67873] [Medline: 40668663]
22. Abdelwanis M, Alarafati HK, Tammam MMS, Simsekler MCE. Exploring the risks of automation bias in healthcare artificial intelligence applications: a bowtie analysis. J Saf Sci Resil 2024;5(4):460-469. [doi: 10.1016/j.jnlssr.2024.06.001]
23. Romeo G, Conti D. Exploring automation bias in human–AI collaboration: a review and implications for explainable AI. AI Soc 2025;41(1):259-278. [doi: 10.1007/s00146-025-02422-7]
24. Bandura A. Self-efficacy: toward a unifying theory of behavioral change. Psychol Rev 1977;84(2):191-215.
25. Yang Q, Lee Y. Ethical AI in financial inclusion: the role of algorithmic fairness on user satisfaction and recommendation. Big Data Cogn Comput 2024;8(9):105. [doi: 10.3390/bdcc8090105]
26. Choung H, Seberger JS, David P. When AI is perceived to be fairer than a human: understanding perceptions of algorithmic decisions in a job application context. Int J Hum-Comput Interact 2023 Oct 15;40(22):7451-7468. [doi: 10.1080/10447318.2023.2266244]
27. de BDR, Abreu MC, Campos J, Paiva A. The amplifying effect of explainability in AI-assisted decision-making in groups. 2025 Presented at: CHI '25: Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems; April 26 to May 1, 2025; Yokohama, Japan. [doi: 10.1145/3706598.3713534]
28. Naiseh M, Al-Thani D, Jiang N, Ali R. How the different explanation classes impact trust calibration: the case of clinical decision support systems. Int J Hum Comput Stud 2023;169:102941. [doi: 10.1016/j.ijhcs.2022.102941]
29. Naiseh M, Babiker A, Al-Shakhsi S, Cemiloglu D, Al-Thani D, Montag C, et al. Attitudes towards AI: the interplay of self-efficacy, well-being, and competency. J Technol Behav Sci 2025;10(4):705-718. [doi: 10.1007/s41347-025-00486-2]
30. Holzinger A, Carrington A, Müller H. Measuring the quality of explanations: the System Causability Scale (SCS): comparing human and machine explanations. Kunstliche Intell (Oldenbourg) 2020;34(2):193-198 [FREE Full text] [doi: 10.1007/s13218-020-00636-z] [Medline: 32549653]
31. Kahr PK, Rooks G, Willemsen MC, Snijders CCP. Understanding trust and reliance development in AI advice: assessing model accuracy, model explanations, and experiences from previous interactions. ACM Trans Interact Intell Syst 2024;14(4):1-30. [doi: 10.1145/3686164]
32. Rieger T, Schindler H, Onnasch L, Roesler E. Explaining AI weaknesses improves human–AI performance in a dynamic control task. Int J Hum Comput Stud 2025;199:103505. [doi: 10.1016/j.ijhcs.2025.103505]
33. Menard P, Bott GJ. Artificial intelligence misuse and concern for information privacy: new construct validation and future directions. Inf Syst J 2024;35(1):322-367. [doi: 10.1111/isj.12544]
34. Schepman A, Rodway P. Initial validation of the general attitudes towards artificial intelligence scale. Comput Hum Behav Rep 2020;1:100014 [FREE Full text] [doi: 10.1016/j.chbr.2020.100014] [Medline: 34235291]
35. Wang YY, Chuang YW. Artificial intelligence self-efficacy: scale development and validation. Educ Inf Technol 2023;29(4):4785-4808. [doi: 10.1007/s10639-023-12015-w]
36. Carolus A, Koch MJ, Straka S, Latoschik ME, Wienrich C. MAILS - meta AI literacy scale: development and testing of an AI literacy questionnaire based on well-founded competency models and psychological change- and meta-competencies. Comput Hum Behav Artif Hum 2023;1(2):100014. [doi: 10.1016/j.chbah.2023.100014]
37. Van den Bosch K, Van Diggelen J, Verdult S, Haije T, Van der Waa J. Measuring meaningful human control in human–AI teaming: effects of team design in AI-assisted pandemic triage. AI Ethics 2025;5(3):3329-3353. [doi: 10.1007/s43681-024-00647-8]
38. Wang F, Cheung ACK, Chai CS, Liu J. Development and validation of the perceived interactivity of learner-AI interaction scale. Educ Inf Technol 2024;30(4):4607-4638. [doi: 10.1007/s10639-024-12963-x]
39. Dornan T, Boshuizen H, King N, Scherpbier A. Experience-based learning: a model linking the processes and outcomes of medical students' workplace learning. Med Educ 2007;41(1):84-91. [doi: 10.1111/j.1365-2929.2006.02652.x] [Medline: 17209896]
40. Tun HM, Rahman HA, Naing L, Malik OA. Trust in artificial intelligence-based clinical decision support systems among health care workers: systematic review. J Med Internet Res 2025;27:e69678 [FREE Full text] [doi: 10.2196/69678] [Medline: 40772775]

XSL•FO
RenderX

41. Aktas E, Çövener Özçelik Ç, ferda Ocakçi A. Reliability and validity of nurses? Experiences of infection prevention and control questionnaire. Clin Exp Health Sci 2020;10(2):1371-1141. [doi: 10.33808/clinexphealthsci.646357]
42. Oldland E, Hutchinson AM, Redley B, Mohebbi M, Botti M. Evaluation of the validity and reliability of the nurses' responsibility in healthcare quality questionnaire: an instrument design study. Nurs Health Sci 2021;23(2):525-537. [doi: 10.1111/nhs.12844] [Medline: 33899312]
43. Creswell JW, Poth CN. Qualitative Inquiry and Research Design Choosing among Five Approaches. Thousand Oaks, California: SAGE Publications, Inc; 2018.
44. Creswell JW, Plano Clark VL. Designing and Conducting Mixed Methods Research. Thousand Oaks, California: SAGE Publications; 2017.
45. Fetters MD, Curry LA, Creswell JW. Achieving integration in mixed methods designs-principles and practices. Health Serv Res 2013;48(6 Pt 2):2134-2156 [FREE Full text] [doi: 10.1111/1475-6773.12117] [Medline: 24279835]
46. Levitt HM, Bamberg M, Creswell JW, Frost DM, Josselson R, Suárez-Orozco C. Journal article reporting standards for qualitative primary, qualitative meta-analytic, and mixed methods research in psychology: the APA Publications and Communications Board Task Force Report. Am Psychol 2018;73(1):26-46 [FREE Full text] [doi: 10.1037/amp0000151] [Medline: 29345485]
47. Timmermans S, Tavory I. Theory construction in qualitative research: from grounded theory to abductive analysis. Sociol Theory 2012;30(3):167-186. [doi: 10.1177/0735275112457914]
48. Lipscomb M. Abductive reasoning and qualitative research. Nurs Philos 2012;13(4):244-256. [doi: 10.1111/j.1466-769X.2011.00532.x] [Medline: 22950728]
49. Guetterman TC, Fetters MD, Creswell JW. Integrating quantitative and qualitative results in health science mixed methods research through joint displays. Ann Fam Med 2015;13(6):554-561 [FREE Full text] [doi: 10.1370/afm.1865] [Medline: 26553895]
50. Braun V, Clarke V. Conceptual and design thinking for thematic analysis. Qual Psychol 2022;9(1):3-26. [doi: 10.1037/qup0000196]
51. Braun V, Clarke V. Using thematic analysis in psychology. Qual Res Psychol 2008;3(2):77-101. [doi: 10.1191/1478088706qp063oa]
52. Malterud K, Siersma VD, Guassora AD. Sample size in qualitative interview studies: guided by information power. Qual Health Res 2016;26(13):1753-1760. [doi: 10.1177/1049732315617444] [Medline: 26613970]
53. Faul F, Erdfelder E, Buchner A, Lang A. Statistical power analyses using G*Power 3.1: tests for correlation and regression analyses. Behav Res Methods 2009;41(4):1149-1160. [doi: 10.3758/BRM.41.4.1149] [Medline: 19897823]
54. Sidra S, Mason C. Generative AI in human-AI collaboration: validation of the collaborative AI literacy and collaborative AI metacognition scales for effective use. Int J Hum-Comput Interact 2025;42(7):5084-5108. [doi: 10.1080/10447318.2025.2543997]
55. Podsakoff PM, MacKenzie SB, Lee J, Podsakoff NP. Common method biases in behavioral research: a critical review of the literature and recommended remedies. J Appl Psychol 2003;88(5):879-903. [doi: 10.1037/0021-9010.88.5.879] [Medline: 14516251]
56. Hair JF, Hult GTM, Ringle CM, Sarstedt M. A Primer on Partial Least Squares Structural Equation Modeling (PLS-SEM). Thousand Oaks, CA: SAGE Publications; 2022.
57. Liengaard BD, Sharma PN, Hult GTM, Jensen MB, Sarstedt M, Hair JF, et al. Prediction: coveted, yet forsaken? Introducing a cross‐validated predictive ability test in partial least squares path modeling. Decis Sci 2020;52(2):362-392. [doi: 10.1111/deci.12445]
58. Ringle CM, Sarstedt M. Gain more insight from your PLS-SEM results. IMDS 2016;116(9):1865-1886. [doi: 10.1108/imds-10-2015-0449]
59. Nowell LS, Norris JM, White DE, Moules NJ. Thematic analysis: striving to meet the trustworthiness criteria. Int J Qual Methods 2017;16(1):160940691773384. [doi: 10.1177/1609406917733847]
60. Sarstedt M, Hair JF, Pick M, Liengaard BD, Radomir L, Ringle CM. Progress in partial least squares structural equation modeling use in marketing research in the last decade. Psychol Mark 2022;39(5):1035-1064. [doi: 10.1002/mar.21640]
61. Buçinca Z, Malaya MB, Gajos KZ. To Trust or to think: cognitive forcing functions can reduce overreliance on AI in AI-assisted decision-making. Proc ACM Hum-Comput Interact 2021;5(CSCW1):1-21. [doi: 10.1145/3449287]
62. Alexandrou A, Chen L. Perceived security of BYOD devices in medical institutions. Int J Med Inform 2022;168:104882. [doi: 10.1016/j.ijmedinf.2022.104882] [Medline: 36242855]
63. Schroeder T, Haug M, Gewald H. Data privacy concerns using mHealth apps and smart speakers: comparative interview study among mature adults. JMIR Form Res 2022;6(6):e28025 [FREE Full text] [doi: 10.2196/28025] [Medline: 35699993]
64. Utomo PS, Yan J. The use of GenAI in medical education in China and Indonesia: a comparative literature review. Educ Health 2025;38(4):347-360. [doi: 10.62694/efh.2025.404]
65. Kwak Y, Ahn J, Seo YH. Influence of AI ethics awareness, attitude, anxiety, and self-efficacy on nursing students' behavioral intentions. BMC Nurs 2022;21(1):267 [FREE Full text] [doi: 10.1186/s12912-022-01048-0] [Medline: 36180902]

XSL•FO
RenderX

66. Li X, Zong Q, Cheng M. The impact of medical explainable artificial intelligence on nurses' innovation behaviour: a structural equation modelling approach. J Nurs Manag 2024;2024:8885760. [doi: 10.1155/2024/8885760] [Medline: 40224836]
67. Ryan RM, Deci EL. Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being. Am Psychol 2000;55(1):68-78. [doi: 10.1037//0003-066x.55.1.68] [Medline: 11392867]
68. Morris C. Work-based Learning. In: Swanwick T, Forrest K, O'Brien BC, editors. Understanding Medical Education: Evidence, Theory, and Practice. Chichester, West Sussex, England: John Wiley & Sons, Ltd; 2018:163-177.
69. Minouei MS, Omid A, Yamani N, Nasri P, Ahmady S. Enhancing clinical faculties' knowledge, attitudes, and performance in clinical supervision: a workplace-based faculty development program using proctor's model. BMC Med Educ 2025;25(1):529 [FREE Full text] [doi: 10.1186/s12909-025-06769-9] [Medline: 40229799]
70. Ten Cate O. Competency-based postgraduate medical education: past, present and future. GMS J Med Educ 2017;34(5):Doc69 [FREE Full text] [doi: 10.3205/zma001146] [Medline: 29226237]
71. Orlikowski WJ. Sociomaterial practices: exploring technology at work. Organ Stud 2007;28(9):1435-1448. [doi: 10.1177/0170840607081138]
72. Kabashkin I. Cognitive atrophy paradox of AI–human interaction: from cognitive growth and atrophy to balance. Information 2025;16(11):1009. [doi: 10.3390/info16111009]
73. Cen Y, Lao C, Li Z, Zhao H, Wang T, Fan C, et al. Association between infection prevention and control safety culture and healthcare workers' compliance with infection control measures: a cross-sectional study. Front Public Health 2025;13:1668493 [FREE Full text] [doi: 10.3389/fpubh.2025.1668493] [Medline: 41179756]
74. Chakma SK, Hossen S, Rakib TM, Hoque S, Islam R, Biswas T, et al. Effectiveness of a hand hygiene training intervention in improving knowledge and compliance rate among healthcare workers in a respiratory disease hospital. Heliyon 2024;10(5):e27286 [FREE Full text] [doi: 10.1016/j.heliyon.2024.e27286] [Medline: 38486731]
75. Goh E, Gallo R, Hom J, Strong E, Weng Y, Kerman H, et al. Large language model influence on diagnostic reasoning: a randomized clinical trial. JAMA Netw Open 2024;7(10):e2440969 [FREE Full text] [doi: 10.1001/jamanetworkopen.2024.40969] [Medline: 39466245]
76. Nagendran M, Festor P, Komorowski M, Gordon AC, Faisal AA. Eye tracking insights into physician behaviour with safe and unsafe explainable AI recommendations. NPJ Digit Med 2024;7(1):202 [FREE Full text] [doi: 10.1038/s41746-024-01200-x] [Medline: 39095449]

Abbreviations

AVE: average variance extracted
BC: bias-corrected
CVPAT: Cross-Validated Predictive Ability Test
GenAI: generative AI
HAI: health care–associated infection
HTMT: heterotrait-monotrait ratio of correlations
JARS: Journal Article Reporting Standards
IPMA: importance-performance map analysis
MCAR: missing completely at random
NFI: normed fit index
PAISCB: perceived AI self-control balance
PLS-SEM: partial least squares structural equation modeling
SOAP: subjective-objective-assessment-plan
SRMR: standardized root mean square residual
TAM: technology acceptance model

Edited by S Brini; submitted 09.May.2026; peer-reviewed by R Marshall, AIA Rahim; comments to author 10.Jun.2026; revised version received 27.Sep.2026; accepted 30.Sep.2026; published 27.Oct.2026.

Please cite as:
Golahi G, Antonio F, Annisa DN
Generative AI Factors, Cognitive Engagement, and Patient Safety Learning in Medical Education: Mixed Methods Study
JMIR Med Educ 2026;12:e100828
URL: https://mededu.jmir.org/2026/1/e100828
doi:10.2196/100828
PMID:

XSL•FO
RenderX

©Gilbert Golahi, Ferdi Antonio, Devina Nur Annisa. Originally published in JMIR Medical Education (https://mededu.jmir.org), 27.Oct.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Education, is properly cited. The complete bibliographic information, a link to the original publication on https://mededu.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Use of Large Language Models for Automated Short-Answer Question Grading in Second-Cycle Medical Education: Performance and Consistency Evaluation Study

Sabrina Sidali[1,2*], MD, PhD; Pierre-Antoine Soret[3,4*], MD, PhD; Mathilde Wagner[5,6], MD, PhD; Thibault Voron[7], MD, PhD; Quentin Vanderbecq[5,6*], MD, PhD

[1]Cordeliers Research Center, Sorbonne University, Inserm, University of Paris, Functional Genomics of Solid Tumors Team, French National Cancer League–Accredited Team, Labex OncoImmunology, F-75006 Paris, Île-de-France, France

[2]University of Paris, Greater Paris University Hospitals (AP-HP), DMU DIGEST, Beaujon Hospital, Reference Center for Vascular Liver Diseases, FILFOIE, ERN RARE-LIVER, Inflammation Research Center, Inserm UMR 1149, Paris, Île-de-France, France

[3]Sorbonne University, Inserm, Saint-Antoine Research Center (CRSA), Paris, Île-de-France, France

[4]Saint-Antoine Hospital, Greater Paris University Hospitals (AP-HP), Paris, Île-de-France, France

[5]Department of Radiology, Saint-Antoine Hospital, Greater Paris University Hospitals (AP-HP), Paris, Île-de-France, France

[6]LIB, UMR 7371, Sorbonne University, CNRS, Inserm U1146, Paris, Île-de-France, France

[7]Department of General and Digestive Surgery, Saint-Antoine Hospital, Greater Paris University Hospitals (AP-HP), Sorbonne University, Paris, Île-de-France, France

[*]these authors contributed equally

Corresponding Author:
Quentin Vanderbecq, MD, PhD
Department of Radiology, Saint-Antoine Hospital, Greater Paris University Hospitals (AP-HP)
184 Rue du Faubourg Saint-Antoine
Paris, Île-de-France, 75012
France
Phone: 33 149282000
Email: q.vanderbecq@gmail.com

Abstract

Background: Short-answer questions (SAQs) effectively assess higher-order reasoning skills in medical education but face challenges in scalability due to time-intensive manual grading and concerns about interrater reliability. Recent advances in large language models (LLMs) may offer a solution through automated assessment.

Objective: This study evaluated the performance and consistency of LLMs—ChatGPT, Claude Sonnet 4, and Mistral Pro—compared with those of human graders for SAQ grading in hepatology and gastroenterology.

Methods: We retrospectively analyzed SAQs from hepatology and gastroenterology examinations at Sorbonne University, Paris, France (2023-2025). Three human evaluators (2 hepatogastroenterologists and 1 radiologist) independently graded all responses twice, with the majority consensus score serving as the reference standard. Six LLMs were evaluated: ChatGPT (GPT-4o, GPT-4.5, GPT-4o3, and GPT-5), Claude Sonnet 4, and Mistral Pro. Each LLM graded all responses twice using standardized prompts, including questions, model answers, grading guidelines, and scoring rubrics. Agreement was assessed using quadratic weighted kappa (QWK), Cohen κ, the intraclass correlation coefficient (ICC), and the mean absolute error (MAE) for interhuman, intrahuman, LLM-versus-human, and intramodel comparisons.

Results: A total of 902 student responses from 12 SAQs were included and analyzed. Interhuman grader agreement was almost perfect (QWK 0.90-0.92; ICC 0.90-0.92; MAE 0.05-0.07). Claude Sonnet 4 and GPT-4o3 demonstrated near-human–level performance when compared with the majority vote of human evaluators (Claude Sonnet 4: QWK 0.88-0.90; ICC 0.88-0.90; MAE 0.06-0.07 and GPT-4o3: QWK 0.82-0.89; ICC 0.82-0.89; MAE 0.07-0.09). GPT-5 showed substantial agreement (QWK 0.77-0.79). Mistral Pro and older ChatGPT versions performed poorly (QWK <0.63). Intramodel consistency was the highest for Claude Sonnet 4 (QWK 0.93), approaching the best human intrarater reliability. Agreement did not correlate with prompt design choices (number of examples or minimum word count).

Conclusions: Advanced LLMs, particularly Claude Sonnet 4 and GPT-4o3, demonstrated near-human–level grading performance, supporting their potential use as assistive tools in SAQ grading. However, performance varied by model, and agreement alone

XSL•FO
RenderX

does not ensure educational validity. Until broader multicenter validation is conducted across diverse domains, stakes, and question formats—along with safeguards for bias, privacy, and model drift—LLMs should augment rather than replace human faculty graders.

(JMIR Med Educ 2026;12:e89445) doi:10.2196/89445

KEYWORDS

large language models; short-answer grading; artificial intelligence; AI; medical education

Introduction

The assessment of learning and competencies is a central pedagogical approach in initial health sciences training programs, whether during the preclinical curriculum, in the context of clinical rotations in the undergraduate program, or in the context of rotations in the postgraduate program [1]. Assessments are used, particularly in the preclinical curriculum, to provide an initial checkpoint for student learning, that is, to verify that students possess adequate knowledge and the necessary skills to enter the clinical curriculum [1]. Among the various assessment formats used in medical education, short-answer questions (SAQs) have emerged as a valuable tool for evaluating student knowledge and clinical reasoning capabilities. Indeed, SAQs constitute a widely recognized assessment format for evaluating the knowledge and cognitive skills of medical students [2-4]. In contrast to multiple-choice questions, which primarily assess the recall of discrete facts and may encourage superficial memorization strategies, SAQs offer the opportunity to evaluate a broader spectrum of competencies, including theoretical understanding, analytic reasoning, and reflective thinking [5]. Furthermore, research in cognitive psychology suggests that the retrieval practice inherent in SAQ formats promotes superior long-term retention compared with recognition-based assessments [6]. This capacity to capture higher-order cognitive processes makes SAQs particularly valuable in the context of competency-based medical education.

However, despite these advantages, the large-scale implementation of SAQs in medical school examinations remains constrained by several factors. Chief among these is the significant administrative and faculty burden imposed on institutions with large student cohorts. Specifically, the time-intensive nature of manual grading represents a significant practical constraint, particularly for large student cohorts. Moreover, human grading is inherently susceptible to variability and potential bias, thereby raising concerns about interrater reliability and the objectivity of assessment outcomes [7]. These limitations have historically restricted the use of SAQs in high-stakes examinations, despite their pedagogical advantages.

Recent advances in AI, particularly the development of large language models (LLMs), offer a promising avenue for addressing these challenges. LLMs have demonstrated the capacity to process and evaluate natural language responses at scale, potentially enabling automated grading of SAQs with accuracy comparable to that of human evaluators. If validated, such systems could reduce the logistical and temporal barriers associated with SAQ assessment, thereby facilitating their broader adoption in high-stakes medical examinations [7-9].

Therefore, in this study we aimed to evaluate the performance, of 6 models of 3 LLMs—ChatGPT (GPT-4o, GPT-4.5, GPT-4o3, and GPT-5), Claude Sonnet 4, and Mistral Pro—in the automated grading of SAQs from second-cycle examinations in hepatology and gastroenterology.

Methods

Study Design and Data Selection

We analyzed retrospective data from the SAQs of second-cycle hepatology and gastroenterology examinations conducted in the 2023 to 2024 and 2024 to 2025 academic years at Sorbonne University Medical School, Paris, France.

The dataset included the questions administered to students, the students' responses, the grading guidelines provided to evaluators, and the scores awarded to students.

Assessment Procedure

Each student response underwent dual grading: human grading and LLM grading. Human grading was performed twice at 3-month intervals by 3 independent evaluators (2 hepato-gastroenterologist (PAS [evaluator 1] and SS [evaluator 2]), and 1 radiologist (QV [evaluator 3]). The 3 evaluators assigned a score of 0, 0.5, or 1 to each student response, based on the grading guidelines set by the teacher who wrote the question. The final score for human grading corresponded to the majority score assigned by at least 2 of the 3 evaluators from their first evaluation. A median score of 0.5 was assigned to 3 different responses.

LLM grading was performed twice by the 6 selected LLMs: 4 versions of ChatGPT (GPT-4o [July 2025 version], GPT-4.5 [July 2025 version], GPT-4o3 [July 2025 version], and GPT-5 with advanced reasoning ["Thinking"; September 2025 version]; OpenAI), Claude Sonnet 4 (September 2025 version; Anthropic), and Mistral Pro (July 2025 version; Mistral AI). Multiple ChatGPT versions were included to assess performance evolution across model iterations, whereas only the most recent versions of Claude Sonnet 4 and Mistral Pro were tested as these represented the state-of-the-art models available from these providers at the time of the study. All models were queried with default temperature settings unless otherwise specified. For each question, the LLMs received the following information: the question prompt, several examples of correct answers, grading guidelines provided to human evaluators, if available, the list of student responses, and the scoring rubric. Multiple responses from the same SAQ were provided within the same prompt rather than graded individually in separate prompts. To reduce redundancy, identical student responses were included only once. Each LLM assigned a score of 0, 0.5, or 1 to each

XSL•FO
RenderX

student response. An example of a typical prompt used is provided in Multimedia Appendix 1.

Ethical Considerations

This retrospective study was conducted in accordance with institutional guidelines and approved by Sorbonne University Medical School. The study was approved by the institutional review board under approval number CRM-2607-587. All examination data were fully anonymized by the university prior to analysis to ensure student confidentiality, and no individual student identifiers were accessible to the research team. As this study involved only the secondary analysis of existing anonymized educational data, individual informed consent was not required.

Statistical Analysis

Data are presented as median (IQR) or number (proportion) as appropriate. Comparisons of quantitative and qualitative variables were made using the Mann-Whitney *U* test and chi-square or Fisher exact test, respectively.

Comparison of Human Grading

Scores assigned by 3 independent evaluators were compared with each other using quadratic weighted kappa (QWK), Cohen κ, intraclass correlation coefficient (ICC; absolute agreement) and mean absolute error (MAE) to assess consistency among human evaluators. Global interrater agreement among the 3 evaluators was assessed using Fleiss κ. To assess intrarater agreement, each evaluator reannotated all responses after a minimum of a 3-month interval, allowing sufficient time to minimize recall bias. Agreement between the first and second annotations was assessed using QWK, Cohen κ, ICC, and MAE. Interevaluator agreement was scored according to the Landis and Koch criteria [10].

Comparison Between LLMs and Human Grading

The performance of each LLM was compared with reference human grading. For each SAQ, the score assigned by each LLM was compared with the score given by the human grader. Agreement between each LLM and the human grader by majority voting was assessed using QWK, Cohen κ coefficient, ICC, and MAE.

Intramodal Agreement

To assess the stability of LLM grading over time, each model was prompted 2 times on the same set of SAQs. Agreement across repeated inferences from the same model was quantified using QWK, Cohen κ, ICC, and MAE, analogous to intrarater agreement in human evaluators.

Correlation Analysis

A correlation analysis (Spearman rank correlation coefficient) was performed to examine the association between QWK and 2 prompt design variables: the number of examples included in the prompt and the minimum number of words required in the response. The minimum threshold was set with the understanding that a correct response may naturally be longer, depending on the use of synonyms or alternative phrasing.

All tests were 2-sided, and $P<.05$ was considered significant. Data handling and analysis were performed using Python (version 3.11; Python Software Foundation) and scikit-learn (version 1.7) [11].

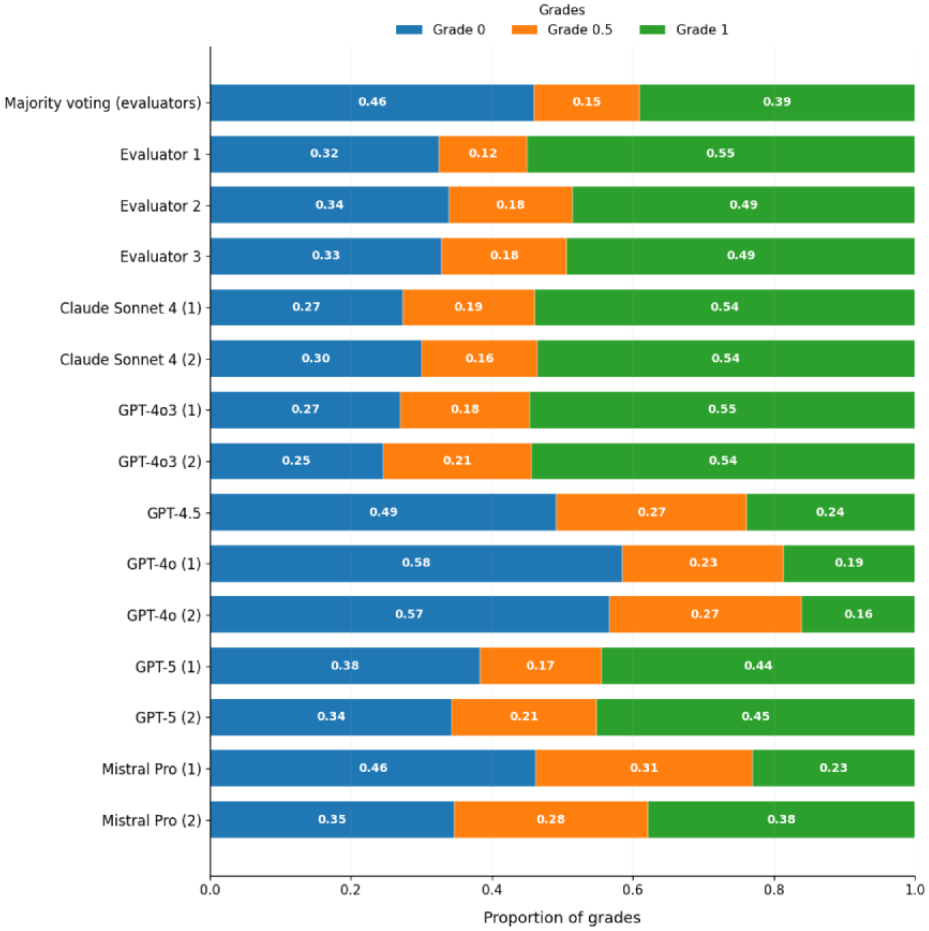
Results

Overview

A total of 902 student responses from 12 SAQs across 6 second-cycle hepatology and gastroenterology examinations were analyzed. The median number of student responses per SAQ was 53 (IQR 26-116; range 16-208). The median number of examples provided to the LLMs was 3.5 (IQR 2.0-4.5; range 1-10). The median prompt length (excluding spaces) was 1213 (IQR 954-1262; range: 672-1518) characters. The distribution of grades across models and human evaluators is represented in Figure 1.

XSL•FO
RenderX

Figure 1. Distribution of grades across models or evaluators.



Comparison of Human Grading

Interrater agreement was high across all evaluator pairs, with QWK values ranging from 0.90 to 0.92, ICCs from 0.90 to 0.92, and MAEs between 0.05 and 0.07. Pairwise Cohen's kappa values ranged from 0.81 to 0.87, confirming strong consistency among graders. Intrarater agreement was also robust: evaluator 2 demonstrated almost perfect reliability (QWK and ICC 0.98), while evaluators 1 and 3 achieved substantial to almost perfect agreement (QWK 0.85 and 0.90, respectively) according to Landis and Koch criteria [10] (Table 1 and Figures 2-3).

XSL•FO
RenderX

Table 1. Interrater and intrarater agreement among human evaluators.

	QWK[a] (95% CI)	Cohen κ (95% CI)	ICC[b] (95% CI)	MAE[c] (95% CI)
Interrater agreement[d]				
Evaluators 1 and 2	0.91 (0.88-0.93)	0.86 (0.83-0.88)	0.91 (0.88-0.93)	0.05 (0.03-0.07)
Evaluators 2 and 3	0.90 (0.87-0.92)	0.81 (0.78-0.84)	0.90 (0.87-0.92)	0.07 (0.05-0.08)
Evaluators 3 and 1	0.92 (0.90-0.94)	0.87 (0.84-0.89)	0.92 (0.90-0.94)	0.05 (0.04-0.06)
Intrarater agreement				
Evaluator 1	0.85 (0.82-0.88)	0.84 (0.81-0.87)	0.85 (0.82-0.88)	0.07 (0.06-0.09)
Evaluator 2	0.98 (0.96-0.99)	0.98 (0.97-0.99)	0.98 (0.96-0.99)	0.009 (0.003-0.017)
Evaluator 3	0.90 (0.88-0.91)	0.76 (0.72-0.79)	0.90 (0.88-0.91)	0.08 (0.07-0.09)
Agreement with majority voting[d]				
Evaluator 1	0.97 (0.96-0.98)	0.95 (0.93-0.97)	0.97 (0.96-0.98)	0.02 (0.01-0.02)
Evaluator 2	0.95 (0.94-0.97)	0.90 (0.88-0.92)	0.95 (0.94-0.97)	0.03 (0.02-0.04)
Evaluator 3	0.95 (0.93-0.97)	0.91 (0.89-0.93)	0.95 (0.93-0.97)	0.03 (0.02-0.04)

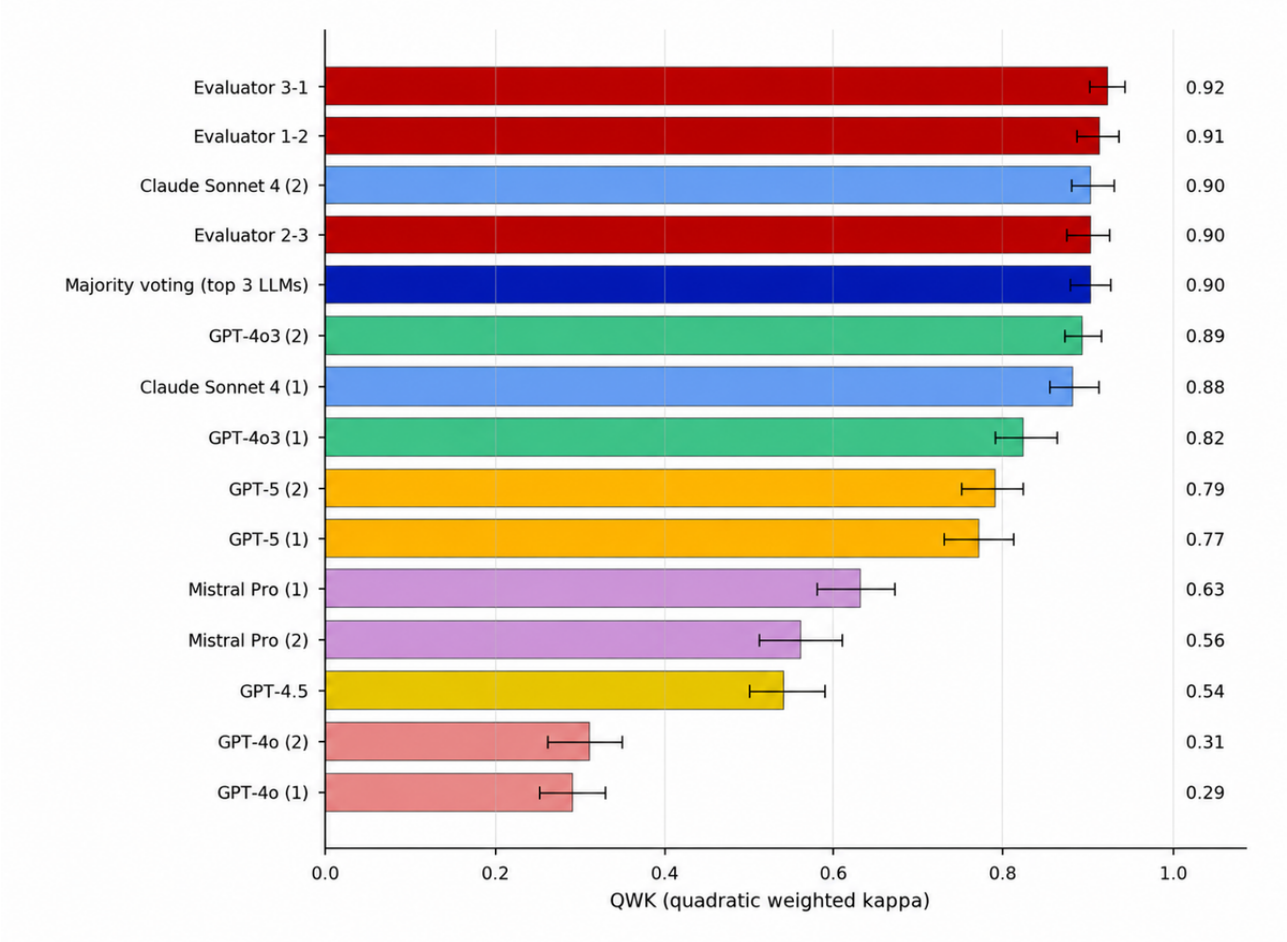
[a]QWK: quadratic weighted kappa.

[b]ICC: intraclass correlation coefficient.

[c]MAE: mean absolute error.

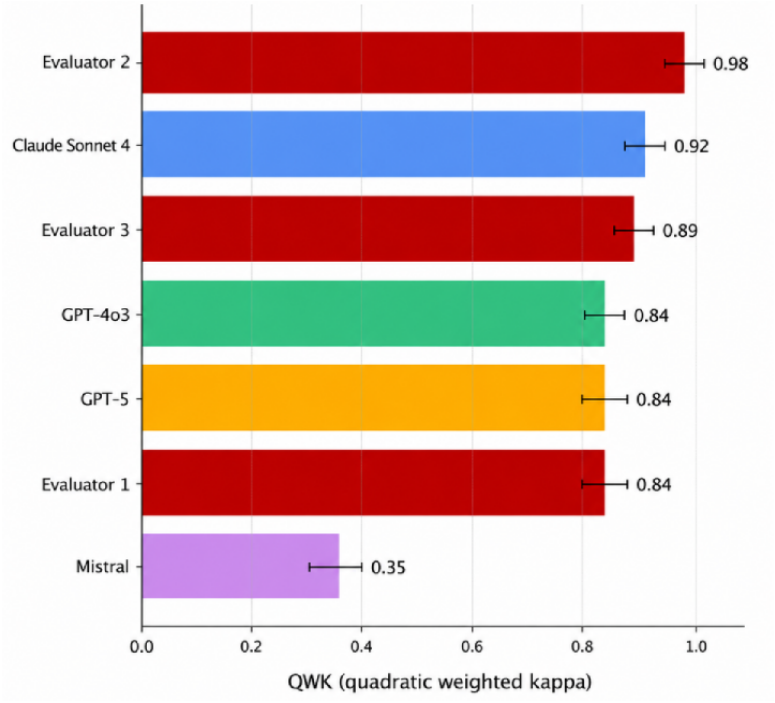
[d]Interrater agreement and agreement against the majority vote were calculated only on the first evaluation of each evaluator.

Figure 2. Quadratic weighted kappa (QWK) score for humans and large language models (LLMs). QWK scores were calculated between pairs of human evaluators based on their initial annotations. For LLMs, QWK scores were calculated against the majority vote of the 3 human evaluators using the initial (1) or repeated (2) LLM assessments. CIs were estimated using bootstrapping with 3000 samples and plotted accordingly. MV: majority voting.



XSL•FO
RenderX

Figure 3. Intrarater and intramodel agreement for evaluators and the best-performing large language models (LLMs). Quadratic weighted kappa (QWK) scores were calculated between pairs of evaluations. CIs were estimated using bootstrapping with 3000 samples and plotted accordingly.



Comparison Between LLMs and Human Grading

A comprehensive comparison of all metrics across models and human evaluators is presented in Table 2. When compared with the majority vote of human evaluators (Table 2 and Figure 2), the best-performing models were Claude Sonnet 4 (QWK 0.88-0.90; ICC 0.88-0.90; MAE 0.06-0.07) and GPT-4o3 (QWK 0.82-0.89; ICC 0.82-0.89; MAE 0.07-0.09), both achieving almost perfect agreement according to the Landis and Koch criteria. These values approached those observed between human evaluator pairs (QWK 0.90-0.92; ICC 0.90-0.92; MAE 0.05-0.07), indicating near-human–level performance. GPT-5 demonstrated substantial agreement (QWK 0.77-0.79; ICC 0.77-0.79; MAE 0.12-0.13), though lower than any human evaluator pair. Mistral Pro showed moderate alignment (QWK 0.56-0.63; ICC 0.56-0.63; MAE 0.22-0.25), while older versions of ChatGPT (GPT-4o and GPT-4.5) exhibited fair to moderate agreement (QWK <0.55 and MAE >0.25) clearly below human consistency levels.

In terms of reproducibility (Table 3 and Figure 3), intramodel consistency was the highest for Claude Sonnet 4 (QWK 0.93; ICC 0.93; MAE 0.04), approaching but not exceeding the best human intrarater agreement (evaluator 2: QWK 0.98; ICC 0.98; MAE 0.009) and surpassing evaluators 1 and 3 (QWK 0.85-0.90; ICC 0.85-0.90; MAE 0.07-0.08). GPT-4o3 and GPT-5 also demonstrated almost perfect stability (QWK 0.85; ICC 0.85; MAE≈0.08), comparable to human evaluators 1 and 3. By contrast, Mistral Pro showed fair intramodel reliability (QWK 0.36; ICC 0.36; MAE 0.28), substantially below any human evaluator's test-retest performance.

Question-level analysis revealed a strong correlation between interhuman agreement and LLM performance. Among the 5 SAQs with the lowest interhuman agreement, 4 also yielded the poorest performance for GPT-4o3 and Claude Sonnet 4. These questions were neither those with the greatest diversity of student responses nor those requiring the longest average response length (number of words).

XSL•FO
RenderX

Table 2. Agreement between large language models (LLMs) versus majority voting of evaluators[a].

LLM vs majority voting	QWK[b] (95% CI)	Cohen κ (95% CI)	ICC[c] (95% CI)	MAE[d] (95% CI)
Claude Sonnet 4 (2)	0.90 (0.88-0.93)	0.83 (0.80-0.86)	0.90 (0.88-0.93)	0.06 (0.05-0.07)
Claude Sonnet 4 (1)	0.88 (0.85-0.91)	0.81 (0.77-0.84)	0.88 (0.85-0.91)	0.07 (0.06-0.08)
GPT-4o3 (2)	0.89 (0.87-0.91)	0.79 (0.76-0.82)	0.89 (0.87-0.91)	0.07 (0.05-0.08)
GPT-4o3 (1)	0.82 (0.79-0.86)	0.78 (0.74-0.81)	0.82 (0.79-0.86)	0.09 (0.07-0.11)
GPT-5 (2)	0.79 (0.75-0.82)	0.69 (0.65-0.72)	0.79 (0.75-0.82)	0.12 (0.11-0.14)
GPT-5 (1)	0.77 (0.73-0.81)	0.67 (0.63-0.71)	0.77 (0.73-0.81)	0.13 (0.11-0.15)
GPT-4.5	0.54 (0.50-0.59)	0.40 (0.36-0.44)	0.54 (0.50-0.59)	0.26 (0.24-0.29)
GPT-4o (2)	0.31 (0.26-0.35)	0.31 (0.27-0.35)	0.31 (0.26-0.35)	0.37 (0.34-0.40)
GPT-4o (1)	0.29 (0.25-0.33)	0.23 (0.19-0.26)	0.29 (0.25-0.33)	0.39 (0.36-0.42)
Mistral Pro (1)	0.63 (0.58-0.67)	0.43 (0.38-0.47)	0.63 (0.58-0.67)	0.22 (0.20-0.24)
Mistral Pro (2)	0.56 (0.51-0.61)	0.38 (0.33-0.43)	0.56 (0.51-0.61)	0.25 (0.23-0.28)

[a]The numbers (1) and (2) following model names indicate the initial and repeated grading assessments, respectively.

[b]QWK: quadratic weighted kappa.

[c]ICC: intraclass correlation coefficient.

[d]MAE: mean absolute error.

Table 3. Intramodel agreement among large language models.

	QWK[a] (95% CI)	Cohen κ (95% CI)	ICC[b] (95% CI)	MAE[c] (95% CI)
GPT-4o3	0.85 (0.82-0.88)	0.77 (0.74-0.80)	0.85 (0.82-0.88)	0.08 (0.07-0.09)
GPT-5	0.85 (0.83-0.88)	0.79 (0.76-0.82)	0.85 (0.83-0.88)	0.08 (0.07-0.10)
Mistral Pro	0.36 (0.30-0.42)	0.38 (0.34-0.43)	0.36 (0.30-0.42)	0.28 (0.25-0.30)
Claude Sonnet 4	0.93 (0.91-0.95)	0.90 (0.87-0.92)	0.93 (0.92-0.95)	0.04 (0.03-0.05)

[a]QWK: quadratic weighted kappa.

[b]ICC: intraclass correlation coefficient.

[c]MAE: mean absolute error.

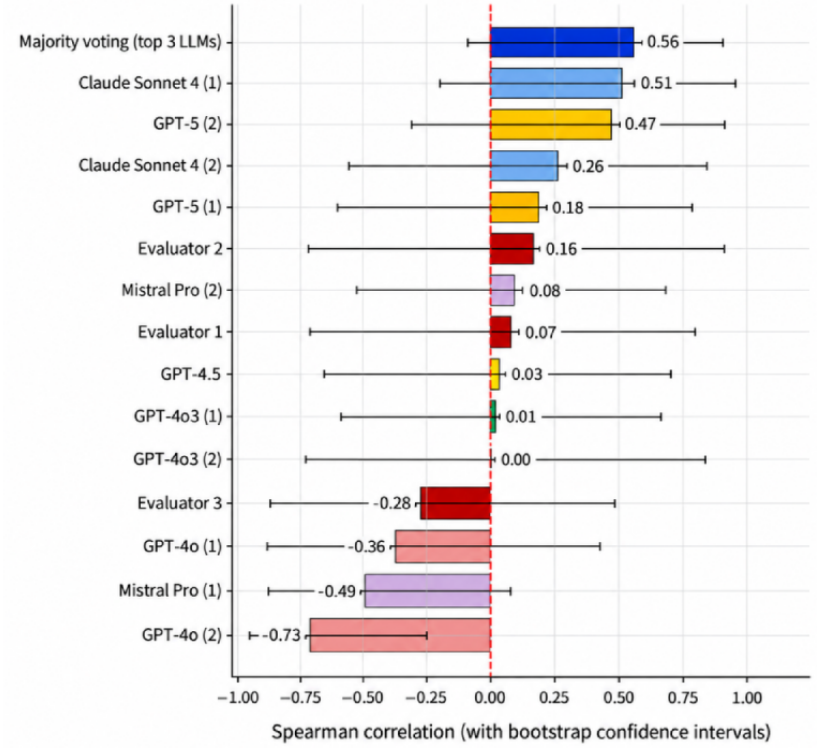
Correlation Analysis: Association Between Grading Agreement (QWK) and Prompt Design

Correlations between QWK and the number of examples provided were generally weak and not statistically significant across all models and human evaluators (Figure 4 and Table 4). Most coefficients were close to 0, with wide CIs spanning both negative and positive values. The strongest positive trends were observed for the consensus score of the 3 best-performing LLMs ("LLM (top3)" (ρ [rho]=0.56, 95% CI −0.09 to 0.92; *P*=.09) and Claude Sonnet 4 at first annotation ("Claude Sonnet 4 (1)" (ρ [rho]=0.51, 95% CI −0.18 to 0.90; *P*=.13), but neither reached statistical significance.

Correlations between QWK and the minimum number of words required for a full score were also low overall (Table 4). No consistent or significant association was observed across models or evaluators. The strongest negative trend appeared for Mistral Pro at first annotation ("Mistral Pro (1)" ρ [rho]=−0.59, 95% CI −0.89 to 0.02; *P*=.07), suggesting a possible inverse relationship, though this did not achieve statistical significance. Human evaluators showed comparable weak correlations, with values ranging from −0.16 to 0.50 and wide CIs.

XSL•FO
RenderX

Figure 4. Correlation between quadratic weighted kappa scores for student responses and the number of examples provided. The numbers (1) and (2) indicate the initial and repeated assessments by the large language models (LLMs), respectively.



XSL•FO
RenderX

Table 4. Correlation between response length and number of examples provided[a].

Correlation characteristic and model or evaluator	ρ (rho) (95% CI)	*P* value[b]
Number of examples provided		
GPT-4.5	0.03 (−0.67 to 0.71)	.94
LLM[c] MV[d] (top3)	0.56 (−0.09 to 0.92)	.09
GPT-5 (1)	0.20 (−0.60 to 0.75)	.58
GPT-5 (2)	0.49 (−0.30 to 0.87)	.15
Mistral pro (1)	−0.49 (−0.86 to 0.06)	.15
Mistral pro (2)	0.08 (−0.53 to 0.68)	.84
GPT-4o3 (1)	0.01 (−0.59 to 0.67)	.97
GPT-4o3 (2)	0.00 (−0.74 to 0.81)	>.99
Claude Sonnet 4 (1)	0.51 (−0.18 to 0.90)	.13
Claude Sonnet 4 (2)	0.26 (−0.53 to 0.82)	.46
Evaluator 1	0.07 (−0.71 to 0.76)	.85
Evaluator 2	0.18 (−0.71 to 0.89)	.63
Evaluator 3	−0.34 (−0.87 to 0.51)	.34
Minimum number of words required for a full score		
GPT-4.5	−0.34 (−0.97 to 0.51)	.34
LLM MV (top3)	0.21 (−0.58 to 0.95)	.56
GPT-5 (1)	−0.23 (−0.86 to 0.57)	.51
GPT-5 (2)	−0.16 (−0.82 to 0.61)	.66
Mistral Pro (1)	−0.59 (−0.89 to 0.02)	.07
Mistral Pro (2)	−0.02 (−0.74 to 0.75)	.96
GPT-4o3 (1)	0.08 (−0.59 to 0.91)	.83
GPT-4o3 (2)	−0.28 (−0.82 to 0.59)	.44
Claude Sonnet 4 (1)	0.36 (−0.42 to 0.95)	.30
Claude Sonnet 4 (2)	0.31 (−0.39 to 0.83)	.38
Evaluator 1	0.36 (−0.40 to 0.86)	.30
Evaluator 2	0.50 (−0.17 to 0.90)	.14
Evaluator 3	−0.16 (−0.72 to 0.78)	.65

[a]Spearman correlations were computed between the number of examples provided or the minimum words required for a full score and QWK scores of models or evaluators against the majority vote of the 3 human evaluators. The numbers (1) and (2) indicate the initial and repeated assessments by the large language models, respectively. 95% CIs were estimated by bootstrapping.

[b]*P* values were obtained from Spearman rank test.

[c]LLM: large language model.

[d]MV: majority voting.

Discussion

In our study, Claude Sonnet 4 and GPT-4o3 showed the closest alignment with human grading, with stable reproducibility across repeated inferences. None of the models surpassed human evaluators in either agreement or reproducibility; the best systems reached human‐level ranges of agreement and reproducibility. Correlation analyses did not reveal clear associations between the prompt design features examined and model performance. These findings suggest that LLM-based grading systems hold promise for enhancing the scalability, efficiency, and potential objectivity of medical education assessment. However, it should be noted that human and LLM biases differ fundamentally: while human graders may be affected by fatigue, familiarity with students, or inconsistent rubric interpretation, LLMs may exhibit systematic biases inherited from pretraining data, sensitivity to prompt framing, or tendencies toward certain response patterns—none of which are yet fully characterized in medical education assessment contexts.

Our results are consistent with existing data. Several studies suggest that LLMs can approximate, but not consistently exceed,

XSL•FO
RenderX

human grading in medical education. Seneviratne and Manathunga [12] reported a strong correlation between an AI-assisted SAQ scoring tool and human examiner markings, supporting the feasibility of automated SAQ scoring when aligned with examiner rubrics. Similarly, Rajan et al [13] reported that GPT-4 scoring could be statistically equivalent to faculty grading at the aggregate level, while discrepancies persisted on items requiring more complex reasoning, supporting the feasibility of LLM-assisted scoring without establishing validity across all unseen assessment contexts. Beyond SAQs, Burke et al [14] showed that ChatGPT could score structured free-text clinical notes with fewer scoring errors than human graders, suggesting that LLMs may perform particularly well when evaluation criteria are explicit and fact-based.

At the same time, published findings remain heterogeneous regarding the LLM grading performance in medical education. For example, Grevisse [7] observed moderate agreement between GPT-4 or Gemini and SAQ instructors and noted that GPT-4 tended to be conservative in awarding full credit, highlighting the need for human oversight. Similarly, Bolgova et al [15] showed that expert-expert agreement remained higher than LLM-expert agreement, with performance varying across questions and depending on rubric structure, highlighting the value of expert-expert agreement for identifying settings in which human oversight remains important. In a medical education setting on mechanical ventilation, ChatGPT demonstrated weak agreement with human graders, and the authors cautioned against using LLMs as sole graders in high-stakes contexts [16].

Emerging evidence also suggests that agreement between LLMs and faculty may not be uniform across question types: in a case-based short-answer exam, LLM grading was comparable to faculty overall, but discrepancies increased for harder questions and for higher Bloom's levels (eg, understanding and evaluating), implying that LLM performance may depend on the cognitive processes targeted by each SAQ rather than reflecting a stable "global" grading ability [13]. In other domains, GPT-4 achieved near-human consistency for lower-stakes short-answer scoring in K-12 education, reinforcing the idea that task difficulty, stakes, and rubric clarity likely influence observed agreement [17].

In our dataset, some questions yielded lower interrater consistency among human graders than among certain LLMs. Rather than indicating that humans "underperformed," this pattern more plausibly suggests that a subset of SAQs may have been insufficiently precise or supported by incomplete grading guidance, increasing interpretive latitude. This interpretation aligns with works showing that question framing and prompt design can strongly influence LLM outputs [18,19]. Clearer and more specific prompts consistently yield more accurate and relevant outputs because prompt engineering acts as a "steering mechanism" that guides the model's reasoning. A notable example is chain-of-thought prompting: simply adding a phrase like "Let's think step by step" can encourage systematic reasoning and dramatically improve accuracy on complex tasks; for instance, boosting GPT-3's performance on math word problems from about 17% to 78% [20]. Numerous studies likewise show that prompts providing context, explicit instructions, or stepwise guidance enhance LLM outputs across domains, including education, where well-designed prompts have been found to significantly improve the usefulness of responses [21]. Together, these findings emphasize that observed LLM-human agreement is not solely a property of the model but also of assessment and implementation design, including question construction, rubric specificity, and prompting strategy (instruction clarity, context provided, and scoring constraints).

This study has several limitations. First, the limited sample size (12 SAQs from a single institution and specialty) constrained statistical power, generalizability, and the robustness of question-level exploratory analyses. Moreover, strong performance on this specific dataset does not guarantee correct assessment in other medical domains, as LLM outputs depend not only on pretraining data but also on rubric quality, question framing, and contextual factors specific to each assessment setting. Second, grading guidelines were not uniformly available, which may have introduced variability in both human and model scoring. Third, multiple student responses were presented within a single prompt for each SAQ rather than being graded in separate prompts. Although identical responses were included only once to reduce redundancy, potential effects of context or order cannot be excluded. Fourth, we evaluated specific LLM versions (July-September 2025) using default parameters without optimization; rapidly evolving models may demonstrate different capabilities. Specifically, LLMs were not trained using previous questions to improve scoring performance. Fifth, practical implementation challenges, including cost, latency, privacy, and student acceptance, were not examined. Finally, our analysis focused on binary and partial-credit scoring for SAQs; performance may differ for other assessment formats (eg, extended responses, oral exams, and objective structured clinical examination [OSCE] narratives) and more granular scales. In addition, although the majority vote of human graders served as a practical reference standard, high agreement with this standard and strong reproducibility should not be interpreted as equivalent to educational validity, particularly in high-stakes assessment contexts. Multicenter studies across specialties, languages, and SAQ formats are needed to better characterize reliability, reproducibility, and external validity in real-world settings.

In conclusion, Claude Sonnet 4 and GPT-4o3 demonstrated the strongest alignment with human grading in our second-cycle medical education SAQ context, potentially enabling more scalable use of SAQs while preserving near-human consistency. However, models have not yet surpassed human evaluators in reproducibility or accuracy across all question types, and performance appears sensitive to rubric quality and cognitive complexity. Approaches such as retrieval-augmented generation and parameter-efficient fine-tuning may further improve task-specific reliability. At present, LLMs should be considered supportive tools integrated into grading workflows with transparent validation and ongoing monitoring, rather than autonomous replacements for faculty graders, particularly for high-stakes assessments where interpretability, fairness, and educational validity are essential.

XSL•FO
RenderX

Acknowledgments

The authors would like to acknowledge the contributions of Professor Fleur Cohen and the AI Sorbonne University committee for providing the authors with the medical school short-answer examination and for supporting this project.

Funding

The authors declared no financial support was received for this work.

Data Availability

The datasets generated or analyzed during this study are available from the corresponding author on reasonable request.

Authors' Contributions

Conceptualization: SS, PAS, QV, MW
Data curation: SS, PAS, QV, TV
Formal analysis: SS, PAS, QV
Investigation: SS, PAS, QV
Writing—original draft: SS, PAS, QV, MW, TV
Writing—review and editing: SS, PAS, QV, MW, TV

Conflicts of Interest

SS received funding from Roche and AstraZeneca for invited talks, and from Elsevier for writing educational content. PAS received funding from AbbVie, Gilead, Ipsen, and Mirum for invited talks and participation on advisory boards. MW received travel support from Guerbet. The other authors declare no conflicts of interest.

Multimedia Appendix 1
Example of a standardized prompt used to grade short-answer questions with large language models.
[DOCX File , 13 KB - mededu_v12i1e89445_app1.docx]

References

1. van der Vleuten C. Validity of final examinations in undergraduate medical training. BMJ 2000 Nov 11;321(7270):1217-1219 [FREE Full text] [doi: 10.1136/bmj.321.7270.1217] [Medline: 11073517]
2. Mee J, Pandian R, Wolczynski J, Morales A, Paniagua M, Harik P, et al. An experimental comparison of multiple-choice and short-answer questions on a high-stakes test for medical students. Adv Health Sci Educ Theory Pract 2024 Jul;29(3):783-801 [FREE Full text] [doi: 10.1007/s10459-023-10266-3] [Medline: 37665413]
3. Sam AH, Field SM, Collares CF, van der Vleuten CP, Wass VJ, Melville C, et al. Very-short-answer questions: reliability, discrimination and acceptability. Med Educ 2018 Apr;52(4):447-455 [FREE Full text] [doi: 10.1111/medu.13504] [Medline: 29388317]
4. Puthiaparampil T, Rahman MM. Very short answer questions: a viable alternative to multiple choice questions. BMC Med Educ 2020 May 06;20(1):141 [FREE Full text] [doi: 10.1186/s12909-020-02057-w] [Medline: 32375739]
5. Burrows S, Gurevych I, Stein B. The eras and trends of automatic short answer grading. Int J Artif Intell Educ 2015 Mar 23;25(1):60-117. [doi: 10.1007/s40593-014-0026-8]
6. Greving S, Richter T. Practicing retrieval in university teaching: short-answer questions are beneficial, whereas multiple-choice questions are not. J Cogn Psychol 2022 Jun 08;34(5):657-674 [FREE Full text] [doi: 10.1080/20445911.2022.2085281]
7. Grévisse C. LLM-based automatic short answer grading in undergraduate medical education. BMC Med Educ 2024 Sep 27;24(1):1060 [FREE Full text] [doi: 10.1186/s12909-024-06026-5] [Medline: 39334087]
8. Abd-Alrazaq A, AlSaad R, Alhuwail D, Ahmed A, Healy PM, Latifi S, et al. Large language models in medical education: opportunities, challenges, and future directions. JMIR Med Educ 2023 Jun 01;9:e48291 [FREE Full text] [doi: 10.2196/48291] [Medline: 37261894]
9. Impey C, Wenger M, Garuda N, Golchin S, Stamer S. Using large language models for automated grading of student writing about science. Int J Artif Intell Educ 2025 Dec;35(4):1825-1859. [doi: 10.1007/s40593-024-00453-7]
10. Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics 1977 Mar;33(1):159-174. [Medline: 843571]
11. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: machine learning in Python. arXiv Preprint posted online on January 2, 2012. [doi: 10.48550/arXiv.1201.0490]

XSL•FO
RenderX

12. Seneviratne HM, Manathunga SS. Artificial intelligence assisted automated short answer question scoring tool shows high correlation with human examiner markings. BMC Med Educ 2025 Aug 05;25(1):1146 [FREE Full text] [doi: 10.1186/s12909-025-07718-2] [Medline: 40764994]
13. Rajan A, Alexander SM, Shenvi CL. Can AI grade like a professor? Comparing artificial intelligence and faculty scoring of medical student short-answer clinical reasoning exams. Adv Health Sci Educ Theory Pract 2026 Apr;31(2):607-617. [doi: 10.1007/s10459-025-10462-3] [Medline: 40767985]
14. Burke HB, Hoang A, Lopreiato JO, King H, Hemmer P, Montgomery M, et al. Assessing the ability of a large language model to score free-text medical student clinical notes: quantitative study. JMIR Med Educ 2024 Jul 25;10:e56342 [FREE Full text] [doi: 10.2196/56342] [Medline: 39118469]
15. Bolgova O, Ganguly P, Ikram MF, Mavrych V. Evaluating large language models as graders of medical short answer questions: a comparative analysis with expert human graders. Med Educ Online 2025 Dec;30(1):2550751 [FREE Full text] [doi: 10.1080/10872981.2025.2550751] [Medline: 40849930]
16. Jade T, Yartsev A. ChatGPT for automated grading of short answer questions in mechanical ventilation. arXiv Preprint posted online on May 5, 2025 [FREE Full text] [doi: 10.48550/arXiv.2505.04645]
17. Henkel O, Boxer A, Hills L, Roberts B, Levonian Z. Can large language models make the grade? An empirical study evaluating LLMs ability to mark short answer questions in K-12 education. In: L@S '24: Proceedings of the Eleventh ACM Conference on Learning @ Scale. New York, NY: Association for Computing Machinery; 2024:300-304.
18. Wu JH, Nishida T, Moghimi S, Weinreb RN. Effects of prompt engineering on large language model performance in response to questions on common ophthalmic conditions. Taiwan J Ophthalmol 2024 Aug 27;14(3):454-457. [doi: 10.4103/tjo.TJO-D-23-00193] [Medline: 39430350]
19. He J, Rungta M, Koleczek D, Sekhon A, Wang FX, Hasan S. Does prompt formatting have any impact on LLM performance? arXiv Preprint posted online on November 15, 2024. [doi: 10.48550/arXiv.2411.10541]
20. Kojima T, Gu SS, Reid M, Matsuo Y, Iwasawa Y. Large language models are zero-shot reasoners. arXiv Preprint posted online on May 24, 2022. [doi: 10.48550/arXiv.2205.11916]
21. Wang L, Chen X, Deng X, Wen H, You M, Liu W, et al. Prompt engineering in consistency and reliability with the evidence-based guideline for LLMs. NPJ Digit Med 2024 Feb 20;7(1):41 [FREE Full text] [doi: 10.1038/s41746-024-01029-4] [Medline: 38378899]

Abbreviations

ICC: intraclass correlation coefficient
LLM: large language model
MAE: mean absolute error
OSCE: objective structured clinical examination
QWK: quadratic weighted kappa
SAQ: short-answer question

Edited by P Kanzow; submitted 15.Dec.2025; peer-reviewed by ABL Narayanan, L Peng; comments to author 11.Apr.2026; revised version received 07.Jul.2026; accepted 11.Jul.2026; published 02.Nov.2026.

Please cite as:
Sidali S, Soret PA, Wagner M, Voron T, Vanderbecq Q
Use of Large Language Models for Automated Short-Answer Question Grading in Second-Cycle Medical Education: Performance and Consistency Evaluation Study
JMIR Med Educ 2026;12:e89445
URL: https://mededu.jmir.org/2026/1/e89445
doi:10.2196/89445
PMID:

©Sabrina Sidali, Pierre-Antoine Soret, Mathilde Wagner, Thibault Voron, Quentin Vanderbecq. Originally published in JMIR Medical Education (https://mededu.jmir.org), 02.Nov.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Education, is properly cited. The complete bibliographic information, a link to the original publication on https://mededu.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Curriculum-Aligned Conversational AI to Support Pathology Self-Directed Learning in Preclinical Medical Students: Crossover Randomized Pilot Study

Kasie Roark[1,2,3], MS; Yumin Choi[2], BS; Elizabeth Hoy[2], BS; Noel Tufts[2], MPH; Robert Amevor[4], MS, PhD; Rahul Ghosal[4], PhD; Brian Keisler[5], MD; Leonardo Bonilha[2,3,6], MD, PhD; Ioulia Chatzistamou[1], MD, PhD

[1]Department of Pathology, Microbiology and Immunology, Floyd School of Medicine, University of South Carolina, Columbia, SC, United States
[2]AI in Medicine Extracurricular Track, Floyd School of Medicine, University of South Carolina, Columbia, SC, United States
[3]Office of Research, University of South Carolina, Columbia, SC, United States
[4]Department of Epidemiology and Biostatistics, University of South Carolina, Columbia, SC, United States
[5]Department of Family and Preventive Medicine, Floyd School of Medicine, University of South Carolina, Columbia, SC, United States
[6]Department of Neurology, Floyd School of Medicine, University of South Carolina, Columbia, SC, United States

Corresponding Author:
Ioulia Chatzistamou, MD, PhD
Department of Pathology, Microbiology and Immunology
Floyd School of Medicine
University of South Carolina
6439 Garners Ferry Rd, Bldg #1, Room B33
Columbia, SC, 29209
United States
Phone: 1 803 2163467
Email: chatzistamou@uscmed.sc.edu

Abstract

Background: AI is reshaping the landscape of medical education, yet the widespread adoption of general-purpose AI tools in student learning raises substantial concerns about curricular alignment, information accuracy, and the absence of pedagogical scaffolding.

Objective: This crossover pilot study evaluated a curriculum-aligned custom GPT that was developed to support AI-scaffolded self-directed learning (SDL), preparation that remains student-initiated but is supported by structured Socratic prompting, compared with fully unguided traditional preparation in second-year medical students at the University of South Carolina, Floyd School of Medicine (USC Floyd SoM).

Methods: Forty-four second-year medical students voluntarily participated in the 2 faculty-led pathology small group discussions (Cushing disease and prostate cancer), using the custom GPT for one case and conventional resources for the other. Faculty assessed SDL documentation forms and small group discussion performance using standardized rubrics.

Results: Students using the curriculum-aligned custom GPT achieved significantly higher SDL rubric scores (10.3 vs 9.4 out of 12; mean difference 0.9, 95% CI 0.5-1.3; P<.001; Cohen d=0.71) and small group discussion scores (14.9 vs 13.7 out of 16; mean difference 1.2, 95% CI 0.7-1.7; P<.001; Cohen d=0.78), with moderate-to-large effect sizes. Delayed retention (end-of-block examination items) did not differ significantly (Wilcoxon signed-rank P=.052). Students also reported significantly greater gains in understanding and pathophysiological reasoning. A total of 34 out of 44 (77%) reported improved reasoning clarity, and 31 out of 44 (70%) improved conceptual understanding. Assessment of 5 prespecified AI-related concern domains did not identify major concerns.

Conclusions: In this crossover pilot study, use of a curriculum-aligned custom GPT was associated with higher faculty-assessed SDL documentation and small group discussion performance. Delayed knowledge retention did not differ significantly in the primary analysis. These findings support further investigation of curriculum-aligned AI tools in larger, multisite studies.

(JMIR Med Educ 2026;12:e102908) doi:10.2196/102908

XSL•FO
RenderX

KEYWORDS

artificial intelligence; medical education; self-directed learning; large language models; pathology; clinical competence

Introduction

The integration of AI into medical education has grown rapidly over the last few years. From adaptive question banks to AI-generated explanations and clinical decision support tools, medical schools are increasingly asking a key question: how can AI be used in ways that support the learning environment and enhance, rather than hinder, the development of clinical reasoning, self-directed learning (SDL), and professional judgment? This is not only a pedagogical issue, but it also has implications for accreditation, particularly with respect to Liaison Committee on Medical Education (LCME) Standard 6.3, which emphasizes that medical curricula should prepare students for lifelong SDL [1].

With the rise of general-purpose large language models such as ChatGPT (OpenAI), students now have on-demand access to medical information at an extraordinary scale [2,3]. However, while this access can be beneficial, it also introduces important challenges. These tools are not designed around a specific curriculum, learning objectives, or case structure [4,5]. They often provide direct answers rather than guiding students through the reasoning process, are not calibrated to the learner's level, and lack the intentional design that distinguishes active learning from passive information retrieval [6,7]. In a survey of undergraduate medical students in the United Kingdom, respondents reported that AI integration enhanced personalized learning but raised concerns about critical thinking and overreliance [8]. In practice, when students use a general-purpose AI to prepare for a case discussion, their learning is no longer aligned with the course's learning objectives or faculty expectations. In many ways, this is like studying from an unstructured, unvetted source rather than engaging with material that is intentionally designed to support learning. Ultimately, the issue is not just about accuracy; it is about alignment. Without clear boundaries around scope, depth, and pedagogical intent, general-purpose AI tools may not consistently align with the goals of competency-based medical education when used without curricular or pedagogical constraints.

Curriculum-aligned custom GPTs—user-configured versions of ChatGPT that follow author-defined instructions and draw on author-supplied reference materials—offer a more structured and pedagogically grounded way to approach this challenge. Instead of providing broad general medical information, a custom GPT can be intentionally designed around specific learning objectives, a case-specific knowledge base, and a Socratic interaction style that prompts students to explain, synthesize, and defend their reasoning rather than passively receive answers [6,7]. This design reconceptualizes AI from a passive information source into an active learning partner—serving as a "study buddy" that helps identify knowledge gaps, fosters deeper thinking and reasoning, and guides students through the clinical and pathophysiological logic embedded in the curriculum. Critically, this approach is designed to address many of the quality and alignment concerns associated with general AI use. By constraining the tool to curriculum-aligned content and clearly defined educational goals, it creates a more controlled and reliable learning environment.

Despite the growing interest in AI in medical education, there is still limited rigorous evidence supporting curriculum-aligned AI tools in SDL, particularly in relation to accreditation standards. Most existing studies relied on self-reported data, lacked appropriate control groups, or focused on general-purpose AI rather than purpose-built educational tools [9-12]. To our knowledge, no prior study has evaluated a curriculum-aligned AI chatbot using a crossover design that includes faculty assessment of both SDL documentation quality and small group discussion (SGD) performance, alongside delayed retention outcomes.

This study aims to address that gap. We present findings from a crossover pilot study (N=44) evaluating a curriculum-aligned custom GPT versus traditional resources for SDL preparation for pathology SGDs in the second-year Endocrine and Reproductive System block at the University of South Carolina, Floyd School of Medicine (USC Floyd SoM). We use the term AI-scaffolded SDL to describe the GPT condition: preparation that is student-initiated but guided by structured Socratic prompting aligned with the curriculum. The traditional condition reflects fully unguided, self-directed preparation using conventional resources at the student's discretion. We hypothesized that use of the custom GPT would lead to higher-quality SDL documentation, stronger SGD performance, and improved knowledge retention compared with traditional preparation methods, without raising major concerns in 5 prespecified AI-related domains.

Methods

Study Design and Setting

We conducted a randomized crossover pilot study during the spring 2026 semester in the Endocrine and Reproductive System block of the integrated second-year curriculum at USC Floyd SoM. This block includes 2 faculty-led pathology SGDs: one focused on Cushing disease and one on prostate cancer. Both cases required students to complete SDL preparation using a standardized SDL form consistent with LCME Standard 6.3 SDL competency domains. The complete study design is summarized in Multimedia Appendix 1. This study is reported in accordance with the CONSORT-EHEALTH (Consolidated Standards of Reporting Trials of Electronic and Mobile Health Applications and Online Telehealth) checklist (Multimedia Appendix 2).

Participants

Second-year medical students were invited to participate voluntarily. Forty-four students enrolled and were randomly assigned to 2 groups of 22 using a computer-generated random sequence; allocation was concealed from faculty evaluators until group assignments were finalized. Group 1 used the custom

XSL•FO
RenderX

GPT for the Cushing disease preparation and traditional resources for prostate cancer; Group 2 completed the reverse sequence. Each student served as their own control to enable pairwise testing. Participation was voluntary and did not affect grades. The 2 SGD sessions were separated by approximately 13 days. Students were instructed not to discuss their preparation method or custom GPT access with peers during the study period, consistent with the academic honor code; they were also informed that all enrolled students would receive access to the custom GPT tool at the conclusion of the study, removing any incentive to share access prematurely. As a pilot study, the sample size was determined by voluntary enrollment within the block rather than a formal a priori power calculation; the observed effect sizes (Cohen d=0.71-0.78) provide a basis for powering a future, larger trial. The study was granted an exemption by the University of South Carolina Institutional Review Board (IRB; see the "Ethical Considerations" section). Because FERPA (Family Educational Rights and Privacy Act) regulations precluded access to student academic records from prior blocks, randomization was unstratified. End-of-block final grades were comparable between groups (Group 1: mean 87.1%, SD 6.4%; Group 2: mean 86.8%, SD 6.7%); age, sex, and prior AI experience were not collected due to these access restrictions.

Custom GPT

We developed 2 curriculum-aligned custom GPTs, one for each case, using the GPTs feature of ChatGPT. The system prompts were shaped with input from faculty to ensure alignment with the learning goals, while still allowing flexibility in how students engaged with the material. Faculty also reviewed the case flow and embedded resources to ensure alignment with the depth and reasoning expectations of the pathology curriculum. They also refined the Socratic prompts to encourage clinical and pathophysiological reasoning rather than simple answer retrieval. The goal was to create a structured but interactive tool that could guide students through the cases in a way that supports reasoning rather than simply providing answers. The design focused on several key elements: (1) keeping the interaction tightly aligned with each case's learning objectives; (2) using a Socratic approach that required students to explain and justify their clinical and pathophysiological reasoning; (3) guiding students through the case stages, aligned with the structure of the SDL forms; (4) incorporating case-specific clinical vignettes, imaging descriptions, and laboratory data provided in the study guide students received before the discussion sessions; and (5) directing students to open-access, peer-reviewed resources. The tools were reviewed by pathology faculty with content expertise and piloted and refined with input from students participating in the AI in Medicine Extracurricular Track at USC Floyd SoM to ensure usability and alignment with the learner experience and needs. Students assigned to the GPT condition were instructed to use the custom GPT as part of their SDL preparation. Objective usage metrics were not available to independently verify the extent or duration of engagement. Students accessed each custom GPT through their institutional USC ChatGPT account via the link provided in an email at no cost, as part of required SDL preparation for each case; no separate account creation or payment was required. Students were not given a specific time or interaction quota and used the tool ad libitum before each SGD. No automated prompts or reminders were sent to encourage use beyond standard course scheduling communications. Students did not receive formal orientation or training on use of the custom GPT before the study, although the discussion identifies a brief orientation as a potential improvement for future implementations. The content and configuration of both custom GPTs were frozen for the duration of the spring 2026 data collection period, with no functional or content changes made once data collection began.

Traditional Resources

Students in the traditional condition completed SDL preparation using conventional resources: course lecture notes, pathology textbooks, peer-reviewed literature via PubMed, and other clinical resources at their discretion. Access to the custom GPT tool was not provided to students in this condition, as the tool link was distributed exclusively to students in the GPT condition.

Assessment Instruments

All students completed a standardized SDL form documenting learning goals, identified knowledge gaps, information sources, clinical reasoning synthesis, and self-assessment. SDL forms were scored by 1 of the 3 pathology small group facilitators using a grading rubric (range 6-12), assessing quality of gap identification, information synthesis, pathophysiological reasoning, and resource use. SGD performance was assessed by the faculty pathology facilitator using a grading rubric scored out of 16, evaluating diagnostic depth, clinical-pathological integration, communication, and professionalism. Both scoring rubrics are provided in Multimedia Appendices 3 and 4. Each SDL form and SGD performance was scored by a single facilitator; interrater reliability was not assessed. All faculty evaluators (IC and 2 additional pathology faculty facilitators) were blinded to study participation: they evaluated their entire small group cohort under routine conditions and were unaware of which students had enrolled in the study or which preparation method any student had used. The faculty facilitators rotated across all 9 small groups on a fixed schedule established at the start of the academic year by the pathology content leader; this schedule was followed for all pathology small groups throughout the year. The group assignment process was computer-generated and was not influenced by the principal investigator (IC), who was aware of the overall study enrollment but was blinded to each student's preparation condition for the case being assessed on that date. Every participant was assessed by a different facilitator for each of the 2 cases, and both study conditions were represented among each facilitator's groups in both sessions. Short-term retention was assessed using student performance on 2 in-house multiple-choice questions per quiz administered 2-4 days after each SGD, while delayed knowledge retention was assessed using 2 multiple-choice questions per case from the end-of-block examination, a customized National Board of Medical Examiners (NBME) examination (NBME Customized Assessment Services). The SDL documentation form was completed and submitted immediately before each SGD as a required deliverable, capturing the student's self-identified learning needs, resources used, information synthesis, and self-assessment of gaps; it preceded and directly

XSL•FO
RenderX

informed the discussion. Representative examples of student interactions with the custom GPT tool for each case, provided by consenting study participants, are available in Multimedia Appendices 5 and 6. Students completed surveys immediately after preparation, via a REDCap (Vanderbilt University) questionnaire accessed through a link distributed by email by KR, including 5-point Likert items on pre- and postpreparation understanding, confidence, knowledge impact, mechanism understanding improvement, and resource helpfulness, along with open-text reflections. Potential AI-related concerns were assessed across 5 prespecified domains informed by previous reports on AI use in educational settings: academic integrity, time burden, overreliance, content accuracy, and fairness across performance tiers. Other potential concerns, including hallucinations, data privacy, and inappropriate clinical guidance, were not formally evaluated.

Statistical Analysis

We analyzed the quantitative data using paired *t* tests for within-student (crossover) comparisons and Mann-Whitney *U* tests for between-group survey responses (GPT condition: n=44 survey observations; traditional condition: n=42 survey observations). For the delayed retention score (range 0-2, based on 2 binary items per case), the Wilcoxon signed-rank test was used as the primary analysis because of the discrete nature of the outcome. For the SDL rubric and SGD scores, paired *t* tests were used because these outcomes were treated as approximately continuous, given the number of possible score values and the within-subject crossover design. We acknowledge that treating rubric-based scores as continuous represents an analytical assumption and therefore a potential limitation [13]. Chi-squared tests were used for categorical variables, and effect sizes were reported using Cohen *d* for paired samples, based on within-subject differences, with 95% CIs for the mean differences (*t* distribution) and for Cohen *d* (noncentral *t* distribution). All tests were 2-tailed, with statistical significance set at α=.05. Faculty-assessed SDL and SGD rubric scores were complete for all 44 enrolled students (N=44 paired observations). Two students did not complete the postpreparation survey in the traditional condition, yielding n=42 survey observations for that condition; these were treated as missing data under a complete-case approach, and no imputation was performed.

We also conducted exploratory post hoc analyses by grouping students into mutually exclusive tertiles based on end-of-block academic performance (low: final grade <84.5%, n=15; middle: 84.5%-90.3%, n=15; high: >90.3%, n=14) to explore whether the impact of the GPT differed by academic performance tier. Because final block grade was calculated after the intervention period, it is a concurrent rather than preintervention measure. These analyses were not prespecified, were not powered for subgroup comparisons, and should be considered hypothesis-generating. Sequence effects were examined by comparing within-student GPT-minus-traditional differences between the 2 crossover groups using a Mann-Whitney *U* test.

Ethical Considerations

This study was reviewed by the University of South Carolina IRB and determined to be exempt from IRB review under Category 1 (research conducted in established or commonly accepted educational settings) on August 12, 2025 (study ID: STUDY00000007). As an exempt protocol, formal documented informed consent was not required by the IRB; all participants were informed of the study's purpose, voluntary nature, and that participation would not affect grades before enrolling.

This pilot study was not registered in a public trial registry. As a minimal-risk, exempt educational pilot study conducted within a single course block, formal registration was not pursued; the authors note this as a limitation and recommend prospective registration for future, larger-scale studies of this intervention.

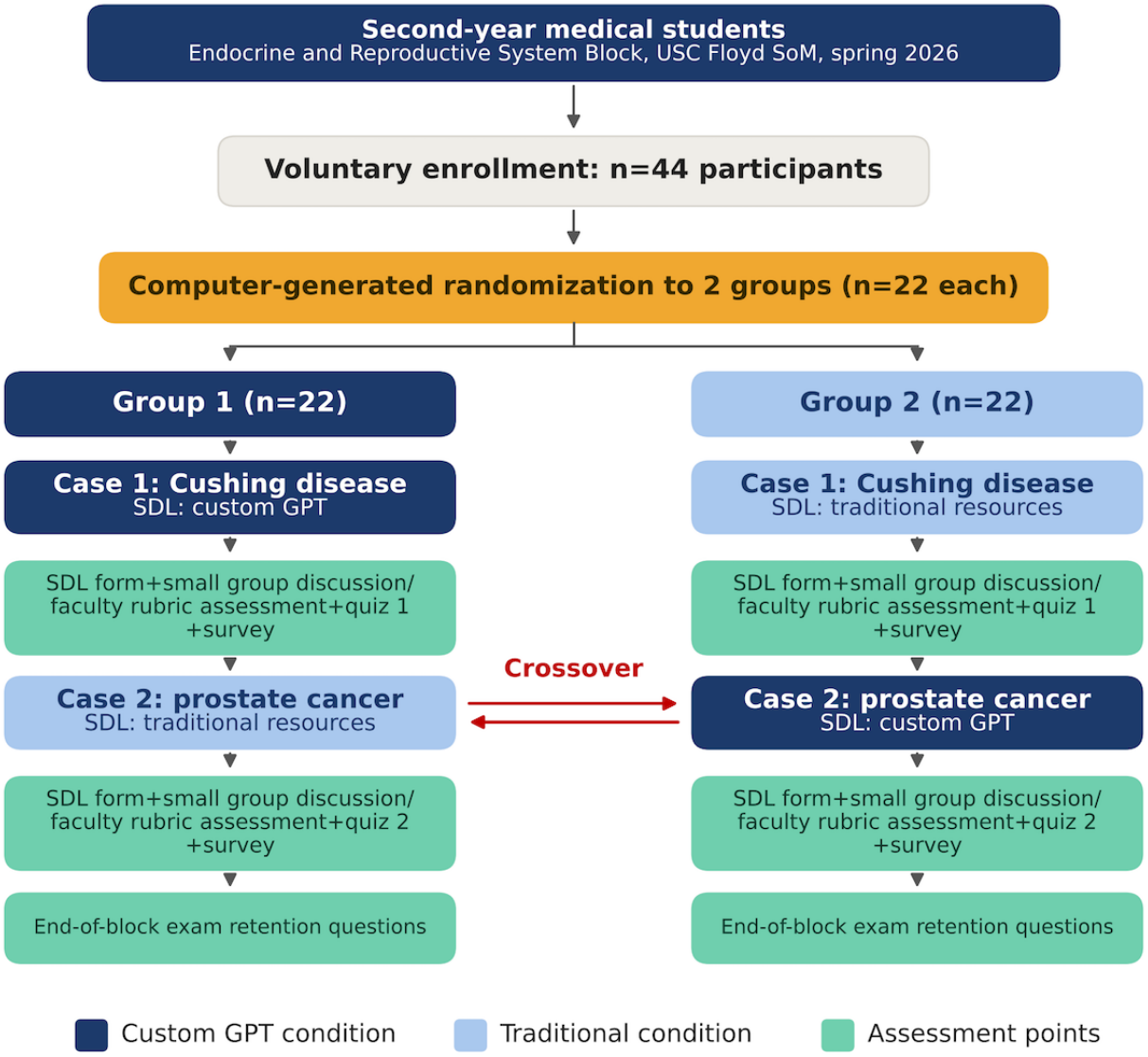
Results

Overview

All 44 enrolled students were randomly assigned (22 per group), completed both cases, and were included in all analyses of faculty-assessed outcomes and delayed retention; 2 students in the traditional condition did not complete the postpreparation survey (Figure 1).

XSL•FO
RenderX

Figure 1. Study design: crossover randomized pilot study. SDL: self-directed learning; USC Floyd SoM: University of South Carolina, Floyd School of Medicine.



SDL Documentation and SGD Performance

Students who used the custom GPT performed significantly better on both faculty-assessed outcomes, with moderate-to-large effect sizes (Table 1; Figure 2). SDL rubric scores were higher in the GPT group (mean 10.3, SD 1.4 vs mean 9.4, SD 1.5 out of 12; mean difference 0.9, 95% CI 0.5-1.3; P<.001; Cohen d=0.71, 95% CI 0.38-1.04). Small group discussion performance showed a similar pattern (mean 14.9, SD 0.8 vs mean 13.7, SD 1.5 out of 16; mean difference 1.2, 95% CI 0.7-1.7; P<.001; Cohen d=0.78, 95% CI 0.44-1.11). **Scores were higher in the GPT condition for both cases.** A significant sequence effect was observed for SGD scores (P=.007), with a larger GPT-associated difference for the prostate cancer case (+1.8 points) than for the Cushing disease case (+0.6 points). No significant sequence effects were identified for SDL scores (P=.09) or delayed retention (P=.98).

XSL•FO
RenderX

Table 1. Summary of quantitative outcomes: GPT versus traditional condition.

Outcome measure	GPT, mean (SD)	Traditional, mean (SD)	Δ (95% CI)[a]	*P* value	Cohen *d* (95% CI)[b]
SDL[c] rubric score (out of 12)[d]	10.3 (1.4)	9.4 (1.5)	+0.9 (0.5-1.3)	<.001	0.71 (0.38-1.04)
SGD[e] score (out of 16)[d]	14.9 (0.8)	13.7 (1.5)	+1.2 (0.7-1.7)	<.001	0.78 (0.44-1.11)
Delayed retention score (out of 2)[f]	1.7	1.5	+0.2	.052	—[g]
Postpreparation understanding (1-5)[h]	4.3	3.9	+0.4	<.001	—
Mechanism understanding improvement (1-5)[h]	3.6 (1.0)	3.1 (0.8)	+0.5	.003	—
In-case quiz score (out of 2)	1.8	1.9	−0.1	.32	—

[a]Mean difference (GPT minus traditional); 95% CIs are reported for the paired rubric outcomes.

[b]Calculated for the paired rubric outcomes only.

[c]SDL: self-directed learning.

[d]Within-student comparison (n=44 pairs); paired *t* test.

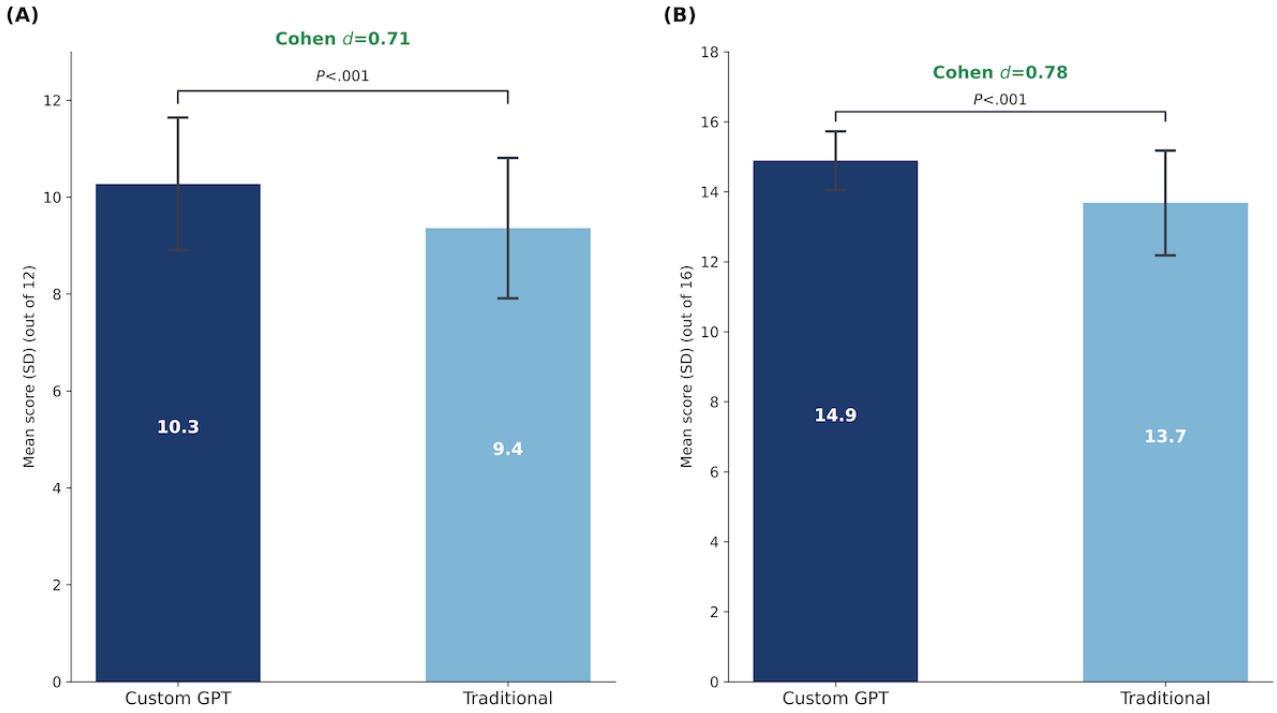
[e]SGD: small group discussion.

[f]Within-student comparison (n=44 pairs); Wilcoxon signed-rank test.

[g]Not applicable.

[h]Between-condition comparison of survey responses (GPT condition: n=44; traditional condition: n=42); Mann-Whitney *U* test.

Figure 2. Faculty-assessed performance scores: custom GPT versus traditional condition. (A) Mean (SD) self-directed learning (SDL) documentation rubric scores (out of 12). (B) Mean (SD) small group discussion (SGD) rubric scores (out of 16). Both comparisons used within-subject crossover data (n=44 paired observations; paired *t* test). Error bars represent ±1 SD. The custom GPT condition produced significantly higher scores on both measures (SDL: 10.3 vs 9.4, *P*<.001, Cohen *d*=0.71; SGD: 14.9 vs 13.7, *P*<.001, Cohen *d*=0.78), with moderate-to-large effect sizes.



Delayed Knowledge Retention

No significant differences were observed between groups in quiz performance 2-4 days after the SGD (1.8 vs 1.9 out of 2; *P*=.32). The primary analysis of delayed knowledge retention did not reach the conventional significance threshold (Wilcoxon signed-rank *P*=.052; Figure 3B). Mean scores were 1.7 versus 1.5 out of 2. Among the 44 paired observations, scores were identical for 26 out of 44 (59%) students, higher under the GPT condition for 13 out of 44 (30%), and lower for 5 out of 44 (11%). In an exploratory post hoc analysis, we divided students into 3 groups based on final block performance (low <84.5%, n=15; middle 84.5%-90.3%, n=15; high >90.3%, n=14). The difference was +0.4 in the lower-performing group, +0.1 in the middle group, and +0.1 in the higher-performing group (Figure 4A). These subgroup findings are exploratory and based on small numbers.

XSL•FO
RenderX

Figure 3. Mechanism understanding improvement and delayed knowledge retention. (A) Distribution of self-reported mechanism understanding improvement. A total of 28 out of 44 (64%) of GPT users reported significant or extreme improvement in pathophysiological reasoning versus 12 out of 42 (29%) traditional users (Mann-Whitney U, *P*=.003). (B) Delayed retention (end-of-block examination items), pooled across both cases. Overall, 84% versus 75% of answers were correct (1.7 vs 1.5 out of 2); the difference was not statistically significant (Wilcoxon signed-rank *P*=.052).

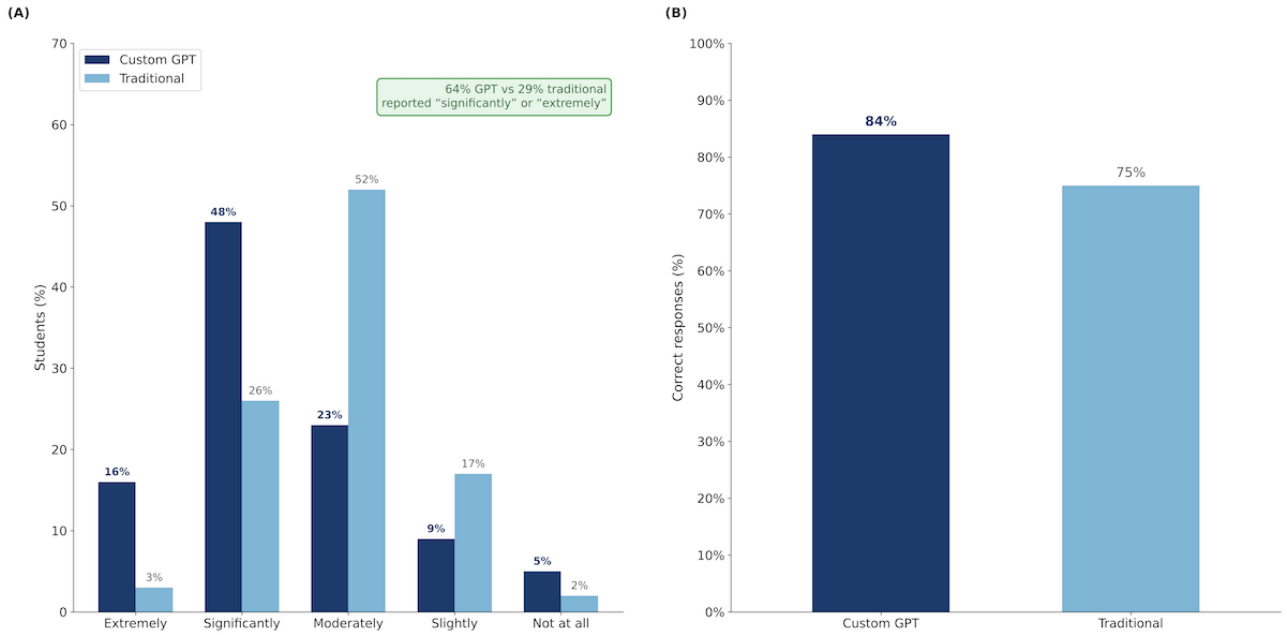
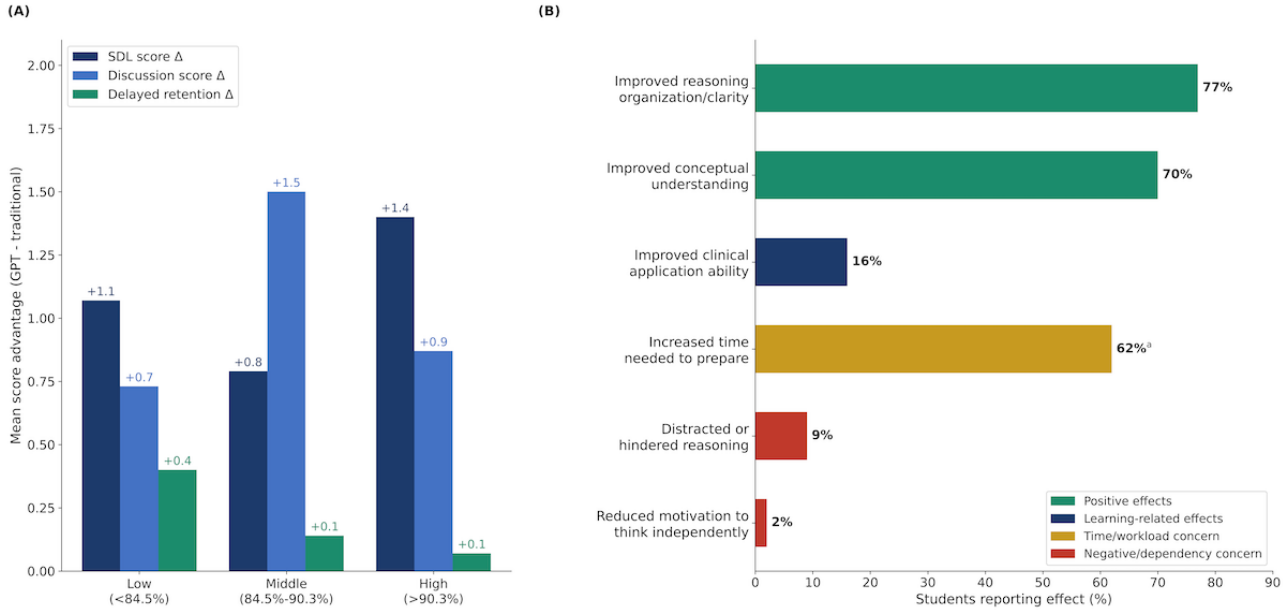


Figure 4. GPT benefits by academic performance tier and student-reported learning effects. (A) GPT score advantage (GPT minus traditional) on self-directed learning (SDL) documentation, small group discussion (SGD), and delayed retention, by post hoc academic performance tertile (n≈14-15 per tier). Low performers (<84.5%) showed the largest delayed retention difference (Δ=+0.4), while middle and high performers showed larger discussion score differences. These subgroup comparisons are exploratory. (B) Student-reported effects of the custom GPT (multiple-select, n=44). A total of 34 out of 44 (77%) reported improved reasoning clarity; 31 out of 44 (70%) improved conceptual understanding. The most frequently cited concern was a perceived increase in the time needed to prepare (16 out of 26 commenting students, 62%). [a]: Proportion of the 26 students who provided optional open-text comments.



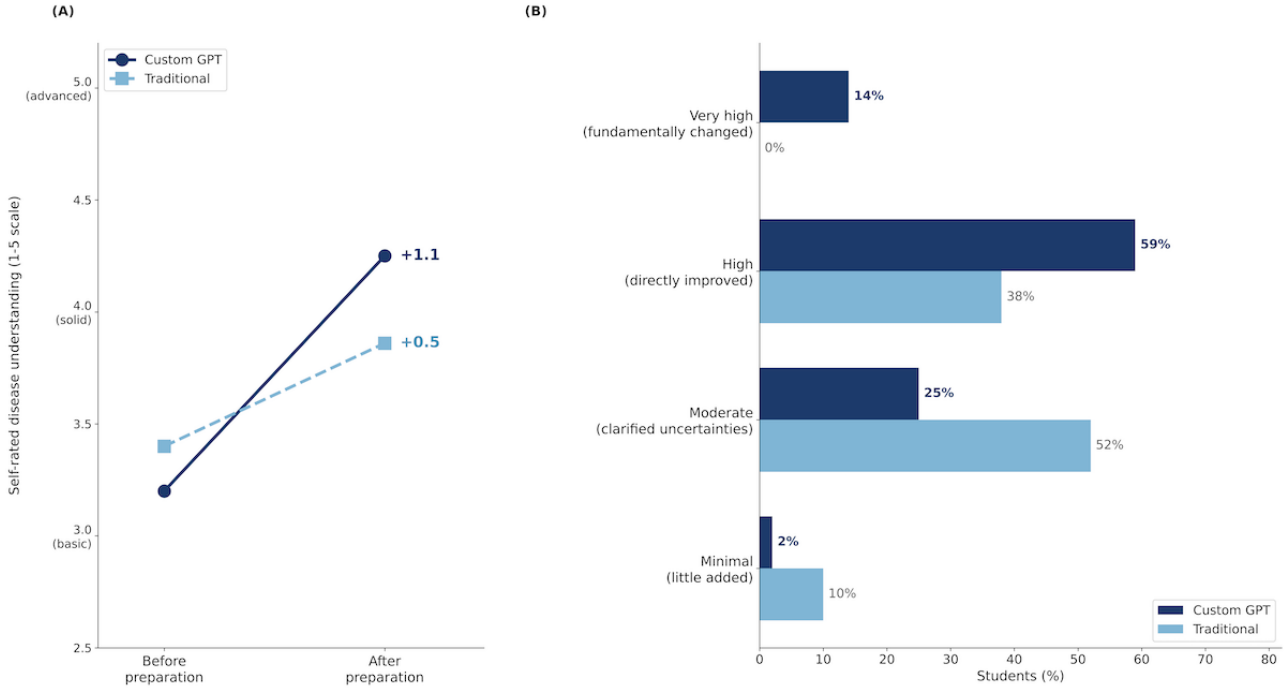
Self-Reported Learning Outcomes

Students using the GPT reported larger gains in their understanding of the disease from pre- to postpreparation (+1.1 vs +0.5; *P*<.001). While both groups improved significantly within-group (*P*<.001), the GPT group ended at a higher overall level of understanding (4.3 vs 3.9; Figure 5A). Perceived impact on knowledge also differed between groups (χ^2_1=9.1; *P*=.003; Figure 5B): 32 out of 44 (73%) GPT users reported a high or very high impact, compared with 16 out of 42 (38%) students using traditional resources. No students in the traditional group reported a "very high" impact (ie, a fundamental change in understanding), whereas 6 out of 44 (14%) GPT users did. A similar pattern was seen for mechanism understanding: 28 out of 44 (64%) GPT users reported significant or extreme

XSL•FO
RenderX

improvement, compared with 12 out of 42 (29%) in the traditional group (P=.003; Figure 3A).

Figure 5. Self-reported learning gains: pre-to-post understanding and knowledge impact. (A) Slope graph of self-rated disease understanding (1=none to 5=advanced) before and after preparation. GPT users improved by +1.1 points versus +0.5, finishing at a significantly higher postpreparation level (4.3 vs 3.9; Mann-Whitney U, P<.001). Within-group changes were significant (Wilcoxon signed-rank, P<.001). (B) Knowledge impact distribution. GPT users were significantly more likely to report high or very high impact (32 out of 44, 73% vs 16 out of 42, 38%; 4-category distribution: χ^2_3=13.8; P=.003). No traditional-condition student reported very high knowledge impact.



Patterns Consistent With LCME Standard 6.3 SDL Competency Domains

When we examined SDL form content in relation to the competency domains described in LCME Standard 6.3, differences emerged in how students approached SDL (Table 2). Students using the custom GPT were more likely to focus on disease mechanisms and pathophysiology as their main learning need (33/44, 75% vs 20/42, 48%), while students using traditional resources more often emphasized treatment and management algorithms (12/42, 29% vs 4/44, 9%). AI chatbot users also reported identifying knowledge gaps they were not previously aware of—a metacognitive skill closely aligned with the competency domains described in LCME Standard 6.3.

These differences were reflected in the SDL rubric scores as well, particularly in goal setting and the quality of learning documentation, which showed a moderate-to-large effect size (Cohen d=0.71) and were consistent across both cases.

Table 2. Type and depth of learning gaps identified: GPT versus traditional condition.

What students said they needed to learn[a]	GPT group (n=44), n (%)	Traditional group (n=42), n (%)	Difference, percentage points
Disease mechanism/pathophysiology	33 (75)	20 (48)	+27
Diagnostic criteria/tests	15 (34)	8 (19)	+15
Treatment/management	4 (9)	12 (29)	−20
Prognosis/outcomes	8 (18)	11 (26)	−8

[a]Students could select more than one category.

Student Satisfaction and Concerns

Overall satisfaction with the GPT was high (Figure 4B). Most students reported meaningful benefits, with 34 out of 44 (77%) noting improved clarity in their reasoning and 31 out of 44 (70%) reporting better conceptual understanding. Only 7 out of 44 (16%) students reported improved clinical application ability. Additionally, student reflections highlighted the value of the coaching-style interaction. One student shared that “this custom AI engine is extremely impressive; having it in a coaching mindset was phenomenal in terms of making sure I truly grasped the material,” while another noted that “the customized ChatGPT was extremely helpful in mastering lecture content and breaking down complex pathophysiology into manageable chunks.” The most common worry was the time required to use the tool, mentioned by 16 out of 26 (62%) GPT-condition students who provided optional comments. Students also suggested a few practical improvements, including fewer simultaneous questions per exchange, clearer signals for when they had reached the end of a task, and occasional support for recalling foundational facts

XSL•FO
RenderX

when needed. Findings across the 5 prespecified AI-related concern domains are summarized in Table 3.

Table 3. Assessment of prespecified AI-related concerns.

Concern domain[a]	What the data show	Interpretation
Academic integrity	Only 1 out of 44 (2%) students felt the tool reduced their independent thinking. Faculty scores were higher in the GPT group, suggesting students still did their own analytical work	No major concern identified in assessed measures
Time and effort	16 out of 26 (62%) students who provided comments said the tool took more time than usual. However, this extra effort was accompanied by higher faculty-assessed scores	Student-reported time burden identified
Overreliance on the tool	Students who used the GPT for one case performed similarly independently on the subsequent case; no evidence of increased reliance on the tool	No signal identified; assessment limited
Accuracy of content	2 out of 44 (5%) students expressed trust concerns; however, because the tool was built from course-specific material, it may be less prone to the inaccuracies seen in general-purpose AI tools	Not formally evaluated
Fairness across students	In an exploratory post hoc analysis, the largest retention difference was observed in students with lower academic performance (Δ=+0.4)	Exploratory finding; requires confirmation

[a]Potential AI-related concerns were reviewed across 5 prespecified domains using student surveys, faculty rubric scores, and written student reflections.

Discussion

Principal Findings

This crossover pilot study suggests that a curriculum-aligned custom GPT can improve second-year medical students' SDL preparation and small group discussion performance. The improvements observed across 2 different cases, in both faculty-assessed and student-reported outcomes, suggest that thoughtfully designed AI tools may have meaningful educational value in undergraduate medical education.

It is also important to distinguish between general-purpose AI tools and curriculum-aligned custom GPTs. General-purpose large language models are not specifically designed around individual medical curricula or the development of SDL competencies. While they may provide accurate information, that information is not necessarily aligned with course learning objectives, may not address specific reasoning gaps, and often lacks the structured, iterative engagement that supports deeper learning [4,5]. In contrast, the curriculum-aligned custom GPTs in this study were intentionally built around specific cases and learning objectives. They asked students to explain and justify their reasoning while directing them toward relevant, curriculum-aligned resources. The improvements observed, particularly in pathophysiological reasoning and SDL documentation quality, may reflect this more intentional and guided learning approach. These findings are consistent with recent controlled studies showing that structured, curriculum-aligned AI tools improve SDL outcomes compared with unassisted preparation [11,12] and add faculty-assessed objective measures that were absent from prior work.

The patterns observed in SDL content are especially interesting in the context of LCME Standard 6.3. Students using the GPT were more likely to identify disease mechanisms and pathophysiology as learning needs, a pattern consistent with the intervention's intended emphasis on mechanistic reasoning. Students also reported identifying previously unrecognized knowledge gaps, suggesting potential support for metacognitive aspects of SDL. These findings should be interpreted as patterns consistent with SDL-related competencies rather than evidence that LCME Standard 6.3 competencies were achieved. The qualitative finding that students were able to identify knowledge gaps they were not previously aware of is particularly noteworthy, as this level of metacognitive awareness can be difficult to achieve with unguided SDL and is essential for lifelong learning [14]. By contrast, only 7 out of 44 (16%) students reported improved clinical application ability, which may reflect the preclerkship context, in which students have limited clinical experience against which to judge such changes; this could be examined in clerkship-stage learners. In the post hoc analysis, the largest retention difference was observed in lower-performing students (Δ=+0.4), and GPT users in the lowest academic tier reached SGD scores of 14.7, compared with 13.8 among higher-performing students using traditional preparation. While these subgroup findings are exploratory and based on relatively small numbers (n≈14-15 per group), they suggest that the GPT may offer helpful structure for students who struggle more with self-directed learning. This potential equity-related benefit is worth exploring further [15].

The primary analysis of delayed knowledge retention did not reach statistical significance (Wilcoxon signed-rank P=.052). Interpretation was limited by the outcome measure, which was based on only 2 end-of-block examination items per case. This result is therefore inconclusive and warrants confirmation in larger studies using more comprehensive, multi-item measures of knowledge retention.

A main limitation was related to the time required to use the tool. Students perceived preparation with the GPT as taking longer than more traditional, reading-based approaches, as reported by 16 out of 26 (62%) students who provided comments. Another factor to consider is timing. The study was conducted in the middle of the spring semester, at a point when students had already established their approaches to preparing for pathology small group discussions. Adopting a new method at that stage may have added to the perceived time burden, as students were not only learning the content but also adjusting

XSL•FO
RenderX

to a different way of studying. Some targeted design changes, such as asking fewer questions at a time, providing clearer end points, and offering a brief orientation, could help reduce the time burden without compromising the educational value. Part of the observed educational benefit may reflect increased time-on-task rather than the GPT's design alone, as 16 out of 26 (62%) GPT students who provided comments reported spending more time than usual on preparation; objective time-on-task was not measured. Future studies should measure preparation time and engagement metrics to disentangle these effects.

Several limitations should be considered when interpreting these findings. First, this pilot study was conducted at a single institution using 2 pathology cases from one preclinical block, which limits the generalizability of the results. In addition, the study was not powered by a formal a priori sample-size calculation; the enrolled sample (n=44) reflected voluntary participation within the course. Although moderate-to-large effect sizes were observed (Cohen *d*=0.71-0.78), these should not be interpreted as evidence that the study was adequately powered, particularly given the width of their 95% CIs. The observed effect sizes may, however, provide a useful basis for sample-size calculations in future multisite studies. Subgroup (tier-based) analyses were exploratory and should be interpreted cautiously. The delayed knowledge retention outcome was also limited by using only 2 end-of-block examination items per case, reducing the precision and sensitivity of the measure to detect modest treatment effects. In the traditional preparation condition, 16 out of 42 (38%) survey respondents reported using general-purpose AI tools. Consequently, the comparison was not between AI and no AI, but rather between a curriculum-aligned custom GPT and usual self-directed preparation, during which some students independently used general AI tools. The direction and magnitude of any resulting bias cannot be determined. It is important to note that this study did not include a condition in which students used general-purpose AI tools without curriculum alignment. Accordingly, the findings do not permit conclusions about whether a curriculum-aligned custom GPT is more effective than unrestricted general-purpose AI use. The advantage demonstrated here is relative to unguided traditional preparation, a comparison that also involves some degree of general AI use in the control condition. Future studies using 3-arm designs (traditional preparation, general-purpose AI, and curriculum-aligned AI) would be needed to test the incremental benefit of curriculum alignment directly. Additionally, objective usage metrics for the custom GPT (eg, number of logins, message counts, or time-on-task) were not collected. Engagement with the custom GPT was inferred from completion of the required SDL form rather than measured directly, limiting our ability to distinguish the effects of the tool's design from the effects of increased time-on-task. Baseline participant characteristics from prior coursework were unavailable because of FERPA restrictions. Although end-of-block performance was comparable between the randomized groups (Group 1: mean 87.1%, SD 6.4%; Group 2: mean 86.8%, SD 6.7%), prior academic performance and prior AI experience could not be fully characterized. However, all students had institutional access to the ChatGPT platform and AI training resources provided by the University of South Carolina. A significant sequence effect was observed for SGD performance, reflecting a larger GPT-associated benefit for the prostate cancer case than for the Cushing disease case. Because case and intervention order are linked in the crossover design, case-specific and order effects cannot be fully distinguished. Notably, sequence effects were not significant for SDL documentation or delayed retention. Each SDL form and SGD performance was scored by a single facilitator, and interrater reliability was not assessed, although facilitators were blinded to preparation condition. Finally, participants could not be blinded to their assigned preparation method because of the nature of the intervention. In addition, multiple outcomes were evaluated without formal adjustment for multiplicity, increasing the possibility of a type I error. Accordingly, the findings should be interpreted in the context of the overall pattern of results rather than any single statistical test.

Conclusions

In this single-institution crossover pilot study of 44 second-year medical students, access to a curriculum-aligned custom GPT was associated with significantly higher faculty-assessed SDL documentation and small group discussion performance compared with unguided traditional preparation (Cohen *d*=0.71 and 0.78, respectively). These findings were consistent across both pathology cases. The primary analysis of delayed knowledge retention did not reach statistical significance. Assessment across 5 prespecified AI-related concern domains did not identify major concerns (Table 3). The main drawback was perceived increased preparation time, reported by 16 out of 26 students who provided comments (62%). These findings support further investigation of curriculum-aligned AI tools in undergraduate medical education in larger, multisite, prospectively registered studies.

Acknowledgments

The authors thank the second-year medical students in the class of 2028 at the University of South Carolina, Floyd School of Medicine (USC Floyd SoM) for their participation in this study. We also thank Drs Christine Wheeler and Avery Dingle for validating the tools and for their blinded assessment of self-directed learning documentation and small group discussion performance. We thank Dr James Catroppo for his contributions to the integrated curriculum of USC Floyd SoM, his role in organizing the pathology small group sessions, and for developing the student and faculty study guides.

XSL•FO
RenderX

Funding

This project was supported by an internal Medical Education Research Innovation (MEDRI) grant from the University of South Carolina School of Medicine, Greenville. The funder had no role in study design, data collection, analysis, interpretation, or the decision to submit this work for publication. RA and RG were supported by the National Institute of General Medical Sciences of the National Institutes of Health under Award Number P20GM155896 through the Center for Clinical & Translational Research at USC (ClinTrUSC). The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Data Availability

Datasets are available from the corresponding author upon reasonable request. The study protocol and the system prompts/configuration of the custom GPTs are likewise available from the corresponding author upon reasonable request.

Generative AI Use

ChatGPT (OpenAI) was used to assist with grammar and language refinement and figure preparation. All content was reviewed and verified by the authors.

Authors' Contributions

Conceptualization: IC, BK, LB
Data collection: IC, YC, EH, NT, BK
Formal analysis: IC, KR, YC, EH, NT, RA, RG
Methodology: IC, KR, BK, LB
Writing – original draft: IC
Writing – review & editing: IC, KR, YC, EH, NT, RA, RG, BK, LB

Conflicts of Interest

The study authors, including IC, developed, validated, and tested the custom GPTs evaluated in this study; no commercial or proprietary interest is held in the OpenAI custom GPT platform or in the specific tools beyond their role as study investigators and developers.

Multimedia Appendix 1
Study design summary.
[PDF File (Adobe PDF File), 32 KB - mededu_v12i1e102908_app1.pdf]

Multimedia Appendix 2
CONSORT-EHEALTH checklist.
[PDF File (Adobe PDF File), 165 KB - mededu_v12i1e102908_app2.pdf]

Multimedia Appendix 3
Self-directed learning documentation rubric.
[PDF File (Adobe PDF File), 55 KB - mededu_v12i1e102908_app3.pdf]

Multimedia Appendix 4
Small group discussion rubric.
[PDF File (Adobe PDF File), 37 KB - mededu_v12i1e102908_app4.pdf]

Multimedia Appendix 5
Example of interaction with custom GPT 1.
[PDF File (Adobe PDF File), 5 KB - mededu_v12i1e102908_app5.pdf]

Multimedia Appendix 6
Example of interaction with custom GPT 2.
[PDF File (Adobe PDF File), 7 KB - mededu_v12i1e102908_app6.pdf]

References

XSL•FO
RenderX

1. Liaison Committee on Medical Education. Functions and structure of a medical school: standards for accreditation of medical education programs leading to the MD degree. Liaison Committee on Medical Education. 2024. URL: https://lcme.org/publications/ [accessed 2026-10-31]
2. Triola MM, Rodman A. Integrating generative artificial intelligence into medical education: curriculum, policy, and governance strategies. Acad Med 2025;100(4):413-418 [FREE Full text] [doi: 10.1097/ACM.0000000000005963] [Medline: 39705530]
3. Rodman A, Mark NM, Artino AR, Lessing JN. Using generative artificial intelligence in medical education. Acad Med 2025;100(2):250 [FREE Full text] [doi: 10.1097/ACM.0000000000005937] [Medline: 39622003]
4. Cascella M, Montomoli J, Bellini V, Bignami E. Evaluating the feasibility of chatGPT in healthcare: an analysis of multiple clinical and research scenarios. J Med Syst 2023;47(1):33 [FREE Full text] [doi: 10.1007/s10916-023-01925-4] [Medline: 36869927]
5. Kung TH, Cheatham M, Medenilla A, Sillos C, De Leon L, Elepaño C, et al. Performance of ChatGPT on USMLE: potential for AI-assisted medical education using large language models. PLOS Digit Health 2023;2(2):e0000198 [FREE Full text] [doi: 10.1371/journal.pdig.0000198] [Medline: 36812645]
6. Hwang GJ, Chang CY. A review of opportunities and challenges of chatbots in education. Interact Learn Environ 2021;31(7):4099-4112 [FREE Full text] [doi: 10.1080/10494820.2021.1952615]
7. Kasneci E, Sessler K, Küchemann S, Bannert M, Dementieva D, Fischer F, et al. ChatGPT for good? On opportunities and challenges of large language models for education. Learn Individ Differ 2023;103:102274 [FREE Full text] [doi: 10.1016/j.lindif.2023.102274]
8. Sunmboye K, Strafford H, Noorestani S, Wilison-Pirie M. Exploring the influence of artificial intelligence integration on personalized learning: a cross-sectional study of undergraduate medical students in the United Kingdom. BMC Med Educ 2025;25(1):570 [FREE Full text] [doi: 10.1186/s12909-025-07084-z] [Medline: 40247306]
9. Aster A, Laupichler MC, Rockwell-Kollmann T, Masala G, Bala E, Raupach T. ChatGPT and other large language models in medical education - scoping literature review. Med Sci Educ 2025;35(1):555-567. [doi: 10.1007/s40670-024-02206-6] [Medline: 40144083]
10. Gilson A, Safranek CW, Huang T, Socrates V, Chi L, Taylor RA, et al. How does ChatGPT perform on the United States Medical Licensing Examination (USMLE)? The implications of large language models for medical education and knowledge assessment. JMIR Med Educ 2023;9:e45312 [FREE Full text] [doi: 10.2196/45312] [Medline: 36753318]
11. Shalong W, Yi Z, Bin Z, Ganglei L, Jinyu Z, Yanwen Z, et al. Enhancing self-directed learning with custom GPT AI facilitation among medical students: a randomized controlled trial. Med Teach 2025;47(7):1126-1133. [doi: 10.1080/0142159X.2024.2413023] [Medline: 39425996]
12. Kalam KA, Masoud FD, Muntaser A, Ranga R, Geng X, Goyal M. ChatGPT as a learning tool for medical students: results from a randomized controlled trial. Cureus 2025;17(6):e85767. [doi: 10.7759/cureus.85767] [Medline: 40656250]
13. Lakens D. Calculating and reporting effect sizes to facilitate cumulative science: a practical primer for t-tests and ANOVAs. Front Psychol 2013;4:863 [FREE Full text] [doi: 10.3389/fpsyg.2013.00863] [Medline: 24324449]
14. Zimmerman BJ. Becoming a self-regulated learner: an overview. Theory Into Practice 2002;41(2):64-70 [FREE Full text] [doi: 10.1207/s15430421tip4102_2]
15. Ricotta DN, Richards JB, Atkins KM, Hayes MM, McOwen K, Soffler MI, (on behalf of Millennium Conference 2019 writing group). Self-directed learning in medical education: training for a lifetime of discovery. Teach Learn Med 2022;34(5):530-540. [doi: 10.1080/10401334.2021.1938074] [Medline: 34279167]

Abbreviations

CONSORT-EHEALTH: Consolidated Standards of Reporting Trials of Electronic and Mobile Health Applications and Online Telehealth
FERPA: Family Educational Rights and Privacy Act
IRB: Institutional Review Board
LCME: Liaison Committee on Medical Education
NBME: National Board of Medical Examiners
SDL: self-directed learning
SGD: small group discussion
USC Floyd SoM: University of South Carolina, Floyd School of Medicine

XSL•FO
RenderX

Edited by P Kanzow; submitted 29.May.2026; peer-reviewed by J Wilkinson, L Corral-Gudino; comments to author 03.Aug.2026; revised version received 09.Oct.2026; accepted 13.Oct.2026; published 06.Nov.2026.

Please cite as:
Roark K, Choi Y, Hoy E, Tufts N, Amevor R, Ghosal R, Keisler B, Bonilha L, Chatzistamou I
Curriculum-Aligned Conversational AI to Support Pathology Self-Directed Learning in Preclinical Medical Students: Crossover Randomized Pilot Study
JMIR Med Educ 2026;12:e102908
URL: https://mededu.jmir.org/2026/1/e102908
doi:10.2196/102908
PMID:

©Kasie Roark, Yumin Choi, Elizabeth Hoy, Noel Tufts, Robert Amevor, Rahul Ghosal, Brian Keisler, Leonardo Bonilha, Ioulia Chatzistamou. Originally published in JMIR Medical Education (https://mededu.jmir.org), 06.Nov.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Education, is properly cited. The complete bibliographic information, a link to the original publication on https://mededu.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Medical Interview Performance Using an AI-Based Standardized Patient vs a Human Standardized Patient: Randomized Crossover Trial

Yasutaka Yanagita[1], MD, PhD; Daiki Yokokawa[1], MD, PhD; Fumitoshi Fukuzawa[2], MD; Shiichi Ihara[1], MD; Ryo Yoshida[1], MD; Rurika Sato[1], MD, PhD; Takanori Uehara[1], MD, PhD

[1]Department of General Medicine, Chiba University Hospital, Chiba, Japan

[2]Department of Medicine, Mount Sinai Morningside, Icahn School of Medicine at Mount Sinai, New York, NY, United States

Corresponding Author:
Yasutaka Yanagita, MD, PhD
Department of General Medicine
Chiba University Hospital
1-8-1 Inohana
Chiba, 260-8677
Japan
Phone: 81 43 222 7171
Email: y.yanagita@gmail.com

Abstract

Background: In Japan, standardized patients (SPs) have traditionally been recruited to teach medical interview competencies. Recent advances in large language models have enabled increasingly natural conversational interactions, raising the possibility of using AI-based SPs (AI-SPs) for medical interview training. However, medical interview performance during interactions with AI-SPs compared with human SPs remains insufficiently understood.

Objective: This randomized single-center crossover trial aimed to compare medical interview performance and learner-reported experiences between encounters with a large language model–based AI-SP and a human SP.

Methods: Participants were fifth-year medical students undertaking clinical clerkships at Chiba University Hospital. The AI-SP was a text-based simulated patient implemented using GPT-4o. Participants were randomized to AI-SP followed by human SP, or the reverse. A washout period of at least 3 days was set between conditions. Each participant completed 2 cases, depression and appendicitis, in both conditions. For the AI-SP interviews, chat logs were retained, and a physician provided feedback after completion of the 2 cases. Human SP interviews were video recorded, and the physician portraying the SP provided feedback immediately after each interview. Medical interview performance was assessed using selected items from the Mini-Clinical Evaluation Exercise (Mini-CEX), with ratings performed independently by 2 general medicine specialists. Participants completed a self-rated questionnaire (7-point Likert-type scale) regarding learning satisfaction, perceived rapport formation, and active engagement in history taking.

Results: Forty-eight participants were analyzed. The Mini-CEX–based mean total score was significantly higher for the human SP than for the AI-SP (23.9, SD 1.7 vs 22.7, SD 1.7, respectively; mean difference 1.24; t_{47}=3.7; P=.001). Scores for history taking and overall clinical competence were higher in the human SP condition. Interrater reliability (intraclass correlation coefficient of 2,1) was 0.76 (95% CI 0.66-0.83). No significant differences were found between conditions for learning satisfaction (human SP: mean 5.9, SD 1.1; AI-SP: mean 6.0, SD 0.8) or perceived rapport formation (human SP: mean 5.4, SD 1.3; AI-SP: mean 5.3, SD 1.3). Active engagement in history taking was significantly higher in the AI-SP condition (human SP: mean 5.8, SD 1.1; AI-SP: mean 6.2, SD 0.7; t_{47}=2.7; P=.009).

Conclusions: Medical interview performance differed between the 2 SP conditions, with higher history taking and overall clinical competence scores in the human SP condition. The text-based nature of the AI-SP and the absence of nonverbal cues may have contributed to these differences. Although no significant differences between conditions were observed in most other assessment outcomes, learning satisfaction, or perceived rapport, self-rated active engagement in history taking was higher with the AI-SP condition. AI-SPs may offer a complementary modality for medical interview practice; however, this study assessed performance during simulated encounters rather than subsequent learning gains or performance with real patients.

XSL•FO
RenderX

Trial Registration: UMIN Clinical Trials Registry UMIN000056228; https://center6.umin.ac.jp/cgi-open-bin/ctr_e/ctr_view.cgi?recptno=R000064153

(JMIR Med Educ 2026;12:e111030) doi:10.2196/111030

KEYWORDS

AI standardized patient; large language model; medical interview education; artificial intelligence; AI

Introduction

Medical interviewing is an essential clinical skill that enables health care professionals to accurately elicit the necessary information from patients to establish diagnosis, develop a good patient-clinician relationship, and guide subsequent processes such as treatment decision-making. It is important for professionals, especially novice learners, to know that medical interviewing goes beyond making predetermined questions, involving also the eliciting of patients' concerns, understanding the course of symptoms and their impacts on daily life, responding empathetically to patients' anxieties, and exploring patients' explanatory models. Some effective measures for improving medical interviewing skills among novice learners include appropriate feedback and opportunities for repeated practice in settings that closely resemble actual clinical encounters [1].

The Mini-Clinical Evaluation Exercise (Mini-CEX) has been widely used to observe and assess learners' clinical competence in contexts similar to real clinical practice and provide them with feedback [2,3]. In Japan, practical training using standardized patients (SPs) has traditionally played a central role in medical interview education [4,5]. Indeed, SPs allow learners to experience patient-clinician interactions in a safe environment and acquire clinical competencies (eg, history taking, communication, and professionalism) in a more integrated manner. However, many SP programs rely heavily on volunteers, and their implementation requires substantial effort both for SP recruitment and to train SPs to understand case scenarios and standardize their performances [6]. Additional burdens associated with these programs include securing faculty members, managing the program, and covering operational costs. Moreover, with the need to support practical examinations such as the objective structured clinical examination, which is conducted across Japanese medical schools, it has become increasingly difficult for many medical education institutions to continuously secure a sufficient number of SPs [7,8].

More recently, advances in AI, particularly in large language models, have generated a growing interest in the educational use of AI systems capable of engaging in natural dialogue with humans. Medical education has recently begun to see the emergence of AI-based medical interview simulations and educational support tools. Prior research has demonstrated that AI-based SPs (hereinafter referred to as AI-SPs) can understand learners' questioning intentions with high accuracy and provide clinically appropriate responses, to the point that preliminary evidence has accumulated regarding both the feasibility of implementing such systems as dialogue-based educational tools and the accuracy of their basic conversational capabilities [9,10].

The problem that remains, then, is that the bulk of the literature has focused primarily on technical validation, including speech recognition accuracy, question understanding, and response accuracy. Meanwhile, a void persists of studies directly investigating the impact of AI-SPs on learners' medical interviewing skills and learning experiences in authentic medical education settings [11]. Related investigations are important because AI-SPs offer several potential advantages, including enabling learners to practice without fear of failure in a psychologically safe environment; engage repeatedly with the same case; simultaneously access training across large numbers of learners; and experience configuration flexibility for patient characteristics such as disease, age, sex, social background, and chief concern [12,13].

In this study, we conducted a randomized crossover trial to compare medical interview performance and learner-reported experiences between encounters with a large language model–based AI-SP and a human SP. Medical interview performance was assessed using selected items from the Mini-CEX, and participants' experiences were assessed using self-rated questionnaires. Rather than evaluating subsequent learning gains or transfer to real patient encounters, this study focused on differences in performance and learner experience under the 2 simulated patient conditions. If effective, AI-SPs could provide a flexible and efficient approach to medical interview education, particularly in settings with limited human resources. This approach may also be applicable in the near future to interprofessional education involving medical students, nurses, pharmacists, and other health care professionals.

Methods

Study Design

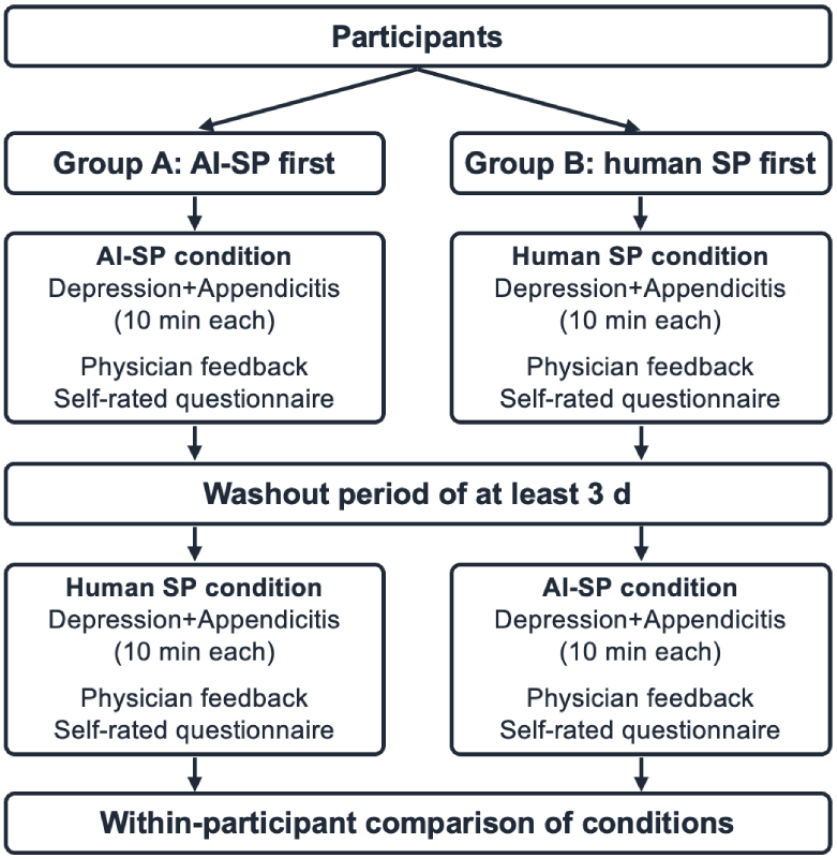
This was a single-center crossover trial. To examine whether AI-SPs could complement human SPs in medical interview education, we compared chat-based medical interviews using an AI-SP with face-to-face medical interviews using a human SP. Data were collected from December 2024 to September 2025. The participants were medical students from Chiba University School of Medicine who participated in a clinical clerkship at the Department of General Medicine, Chiba University Hospital.

Participants were randomly assigned to 1 of 2 groups. In the AI-SP–first group, participants first conducted chat-based medical interviews with the AI-SP; after a washout period of at least 3 days, they conducted face-to-face medical interviews with the human SP. In the human SP–first group, participants first conducted face-to-face medical interviews with the human SP; after the same washout period as in the other group, they conducted chat-based medical interviews with the AI-SP. Each

XSL•FO
RenderX

participant completed medical interviews for 2 cases, depression and appendicitis, under both the AI-SP and human SP conditions. This design enabled within-participant comparisons of medical interview performance between the AI-SP and human SP conditions. After each interview, a physician provided feedback to the participant. Medical interview performance was independently assessed by 2 general medicine specialists using selected items from the Mini-CEX. Participants then rated 3 aspects of their learning experience on a 7-point Likert scale. Figure 1 shows the flowchart of this study.

Figure 1. Study flow. AI-SP: AI-based standardized patient; SP: standardized patient.



Participants

A total of 76 fifth-year medical students from a single academic year at Chiba University School of Medicine participated in this study. Students were eligible if they were fifth-year medical students participating in the clinical clerkship at the Department of General Medicine during the study period and had completed the common preclinical examinations, including computer-based testing and the objective structured clinical examination. Eligible students were consecutively approached for participation during their clinical clerkship. The purpose and procedures of the study were explained to them, and students who agreed to participate were enrolled after providing verbal informed consent. Participation was voluntary and did not affect their clerkship evaluation. Because this study used a crossover design, participants were required to complete both medical interview sessions, with a washout period of at least 3 days between the sessions. The 76 participants were randomly assigned in equal numbers to the AI-SP–first sequence (n=38) or the human SP–first sequence (n=38). Participants who did not complete both medical interview sessions were excluded from the analysis. Of the 76 participants, 28 did not complete both medical interview sessions and were therefore excluded from the analysis, leaving 48 participants for the final analysis (AI-SP first: n=25; human SP first: n=23).

Case Scenarios

In this study, we selected 2 cases, depression and appendicitis, as conditions commonly encountered in primary care outpatient settings and that require multidimensional history taking. For each case, we developed a scenario that included the chief concern, age, sex, medical history, medication history, social history, circumstances at presentation, and symptom course. The complete case scenarios used in this study are provided in Multimedia Appendix 1.

Generation and Validation of the AI-SP

GPT-4o (OpenAI) was used to create the AI-SP. The AI-SP interviews were conducted using GPT-4o through the ChatGPT web interface with a research-specific account from December 2024 to September 2025. Temporary Chat mode was used, and a new chat session was initiated for each participant and each case. Thus, conversation content from previous participants or cases was not carried over to subsequent sessions. No additional ChatGPT tools were used during the interviews. The model available during the study was identified as GPT-4o; a specific model snapshot or version identifier was not recorded. Prompts were developed to enable the AI to respond as an SP based on the case settings. The prompts included the patient's basic information, chief concern, symptom details, social history, attitude during responses, and information to be disclosed

XSL•FO
RenderX

depending on the participant's questions. Each case-specific prompt instructed the model to remain in the role of the simulated patient and disclose information according to the questions asked by the participant. The complete prompts used for the 2 cases are provided in Multimedia Appendix 2. The AI-SP was designed to avoid inappropriate expressions during the conversation and terminate the interview if the participant made inappropriate remarks. The AI-SP scenarios and responses were reviewed by 2 board-certified general internists, who evaluated their medical validity and educational appropriateness.

Procedure

For the AI-SP interviews, case-specific prompts were entered into GPT-4o in advance, and the model was instructed to behave like an SP. Participants conducted 10-minute chat-based medical interviews with the AI-SP for each of the 2 cases, depression and appendicitis. After both AI-SP cases were completed, a physician reviewed the chat logs with the participant and provided feedback on the interviews.

For the human SP interviews, participants conducted 10-minute face-to-face medical interviews for each of the same 2 cases, and the interviews were recorded. The human SPs were portrayed by 7 general physicians. Before participating in the study, all physicians reviewed the standardized case scenarios and received instructions regarding patient characteristics, clinical information, and the information to be disclosed in response to participants' questions to promote consistency across encounters. The physicians portraying the human SPs provided feedback immediately after each case. Medical interview performance was subsequently assessed by 2 general medicine specialists.

Evaluation

Medical interview performance was evaluated using the Mini-CEX, a tool used to assess clinical competence. Because the AI-SP condition did not involve any type of physical examination, the relevant item for this topic was excluded, and 6 items were adopted for the Mini-CEX–based assessment (each rated on a 9-point scale; maximum total score of 54 points). The assessment items included history taking, communication, diagnosis, management, professionalism, and overall clinical competence.

Medical interview performance was evaluated after participants completed both the depression and appendicitis cases in each condition (AI-SP and human SP). For the AI-SP condition, chat logs were used for assessment; for the human SP condition, recorded videos were used. Two specialist physicians engaged in general medicine who were not involved in the medical interviews or feedback sessions independently evaluated the participants' performance; they assessed each participant independently without consulting each other. For each condition, the depression and appendicitis cases were not scored separately. Instead, each evaluator reviewed the participant's performance across both cases and assigned a single score for each Mini-CEX item for that condition. The scores assigned by the 2 independent evaluators were then averaged to obtain the final score for each participant in each condition.

Participants used a 7-point Likert-type scale to provide self-evaluations on 3 items in the participant questionnaire, namely, satisfaction with the simulated interview, degree of rapport formation with the SP, and active engagement in history taking.

Statistical Analysis

The required sample size was calculated using G*Power (version 3.1) and was determined to be 34 participants while considering a 2-sided test, a significance level of .05, power of 0.8, and an effect size of 0.5. For each participant, medical interview performance was evaluated once in each condition after completion of the two cases, depression and appendicitis. Thus, each participant had 1 Mini-CEX–based assessment score for the AI-SP condition and 1 for the human SP condition. The scores were assigned by the 2 independent evaluators, and the mean score was used as the final score for each participant in each condition. The Mini-CEX–based assessment scores and questionnaire responses on the 7-point Likert scale were compared between the AI-SP and human SP conditions within the same participants using 2-tailed paired *t* tests. All statistical analyses were performed using SPSS Statistics for Windows (version 26.0; IBM Corp).

Ethical Considerations

This study was approved by the ethics committee of Chiba University Hospital (approval 545). The study protocol, including the use of verbal informed consent rather than written informed consent, was approved by the ethics committee. Verbal informed consent was obtained from all participants before enrollment in the study. Participants were instructed not to enter personally identifiable information, including their names or student identification numbers, during interactions with the AI-SP. A research-specific ChatGPT account was used for the AI-SP sessions. Chat logs used for study evaluation were stored on a university-managed computer, and access to the study data was restricted to the investigators involved in the study and assessment. No financial compensation was provided for participation. This study was registered with the University Hospital Medical Information Network Clinical Trials Registry on November 21, 2024 (unique trial number UMIN000056228).

Results

Participant Characteristics

Of the 76 students who participated in this study, 38 were randomly assigned to the AI-SP–first sequence, and 38 were assigned to the human SP–first sequence. Twenty-eight participants did not complete both medical interview sessions and were excluded from the analysis. Of these, 13 were in the AI-SP–first sequence, and 15 were in the human SP–first sequence. Reasons for noncompletion included clinical clerkship scheduling and personal circumstances, such as leave or illness. Consequently, 48 participants were included in the final analysis (AI-SP first: n=25; human SP first: n=23). Among these 48 fifth-year medical students, the median age was 22.0 years (IQR 22.0-23.0); 33 (68.8%) were male, and 15 (31.3%) were female. Participant characteristics are shown in Table 1 according to intervention sequence. No adverse events, privacy breaches,

XSL•FO
RenderX

major technical problems, or other unintended incidents related to either study condition were observed. The participant flow is shown in Figure 2.

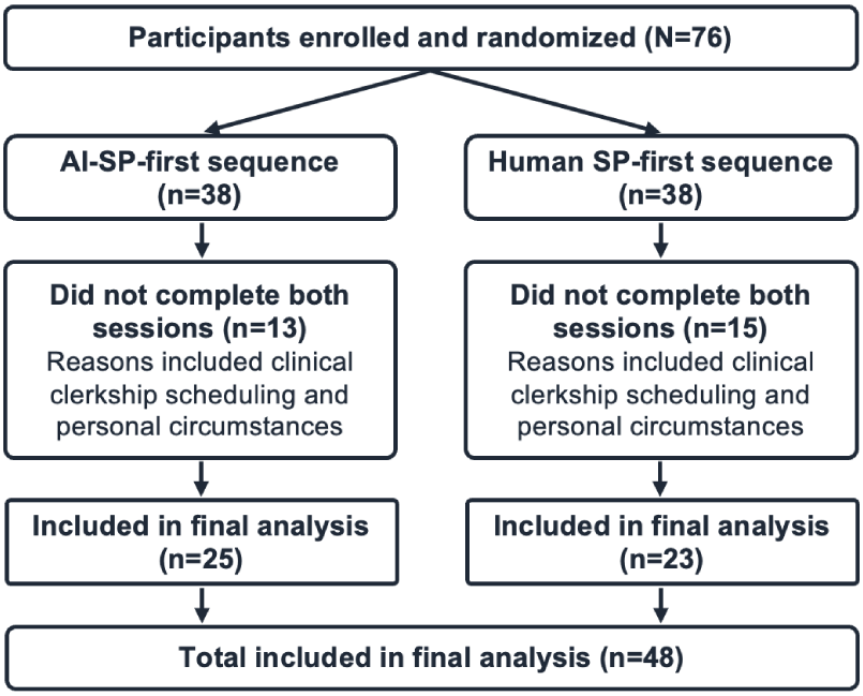
Table 1. Participant characteristics.

Characteristics	AI-SP[a] first (n=25)	Human SP[b] first (n=23)
Age (y), median (IQR)	22.0 (22.0-23.0)	22.0 (22.0-23.0)
Male, n (%)	16 (64)	17 (73.9)
Female, n (%)	9 (36)	6 (26.1)

[a]AI-SP: AI-based standardized patient.

[b]SP: standardized patient.

Figure 2. CONSORT (Consolidated Standards of Reporting Trials) flow diagram of study participants. AI-SP: AI-based standardized patient; SP: standardized patient.



Medical Interview Performance

Total scores were significantly higher in the human SP condition than in the AI-SP condition (mean 23.9, SD 1.7 vs mean 22.7, SD 1.7; t_{47}=3.7; *P*=.001), and the interrater reliability of the Mini-CEX–based assessment was good, with an intraclass correlation coefficient (2,1) of 0.76 (95% CI 0.66-0.83). Among the individual assessment items, history taking scores were significantly higher in the human SP condition than in the AI-SP condition (mean 4.3, SD 0.5 vs mean 3.5, SD 0.5; t_{47}=6.9; *P*<.001), and so were overall clinical competence scores (mean 3.9, SD 0.2 vs mean 3.7, SD 0.5; t_{47}=3.2; *P*=.002). No significant differences were observed between conditions for communication, diagnosis, professionalism, or management (Table 2).

XSL•FO
RenderX

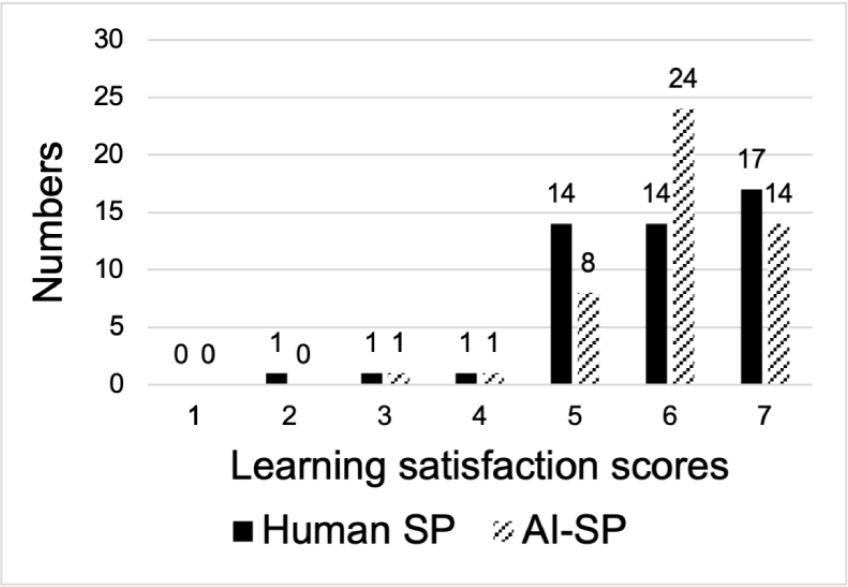
Table 2. Comparison of evaluation items between human standardized patients (SPs) and AI-based SPs (AI-SPs)[a].

	Human SP, mean (SD)	AI-SP, mean (SD)	Mean difference (95% CI)	Cohen *dz*	*t* test (*df*)	*P* value
History taking	4.3 (0.5)	3.5 (0.5)	0.81 (0.58 to 1.05)	1.00	6.9 (47)	<.001
Communication	3.9 (0.5)	3.9 (0.5)	0.00 (−0.19 to 0.19)	0.00	0.0 (47)	>.99
Diagnosis	4.0 (0.5)	3.9 (0.5)	0.11 (−0.02 to 0.25)	0.26	1.8 (47)	.09
Professionalism	3.9 (0.4)	3.9 (0.4)	−0.01 (−0.16 to 0.14)	0.01	0.1 (47)	.89
Management	3.9 (0.3)	3.9 (0.3)	0.08 (−0.03 to 0.20)	0.22	1.5 (47)	.15
Comprehensive evaluation	3.9 (0.2)	3.7 (0.5)	0.24 (0.09 to 0.39)	0.46	3.2 (47)	.002
Total	23.9 (1.7)	22.7 (1.7)	1.24 (0.56 to 1.92)	0.53	3.7 (47)	.001

[a]Scores for each Mini-Clinical Evaluation Exercise item were rated on a 9-point scale, with a maximum total score of 54. Mean differences were calculated as human SP minus AI-SP.

Questionnaire results showed no significant differences between the human SP and AI-SP conditions in learning satisfaction (mean 5.9, SD 1.1 vs mean 6.0, SD 0.8) or perceived rapport formation (mean 5.4, SD 1.3 vs mean 5.3, SD 1.3). However, participants reported significantly greater active engagement in history taking in the AI-SP condition than in the human SP condition (mean 6.2, SD 0.7 vs mean 5.8, SD 1.1; t_{47}=2.7; *P*=.009; Figures 3-5).

Figure 3. Learning satisfaction scores in the Human SP and AI-SP conditions. AI-SP: AI-based standardized patient; SP: standardized patient.



XSL•FO
RenderX

Figure 4. Degree of rapport formation scores in the Human SP and AI-SP conditions. AI-SP: AI-based standardized patient; SP: standardized patient.

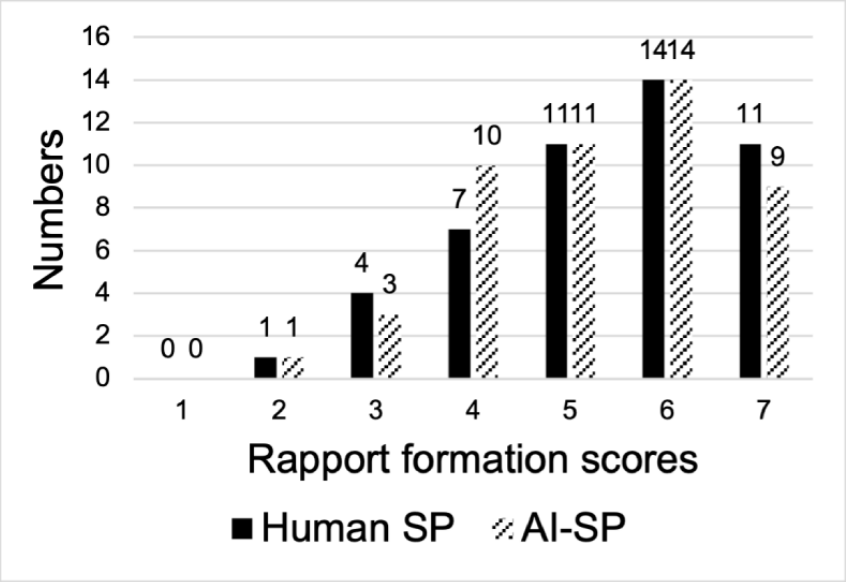
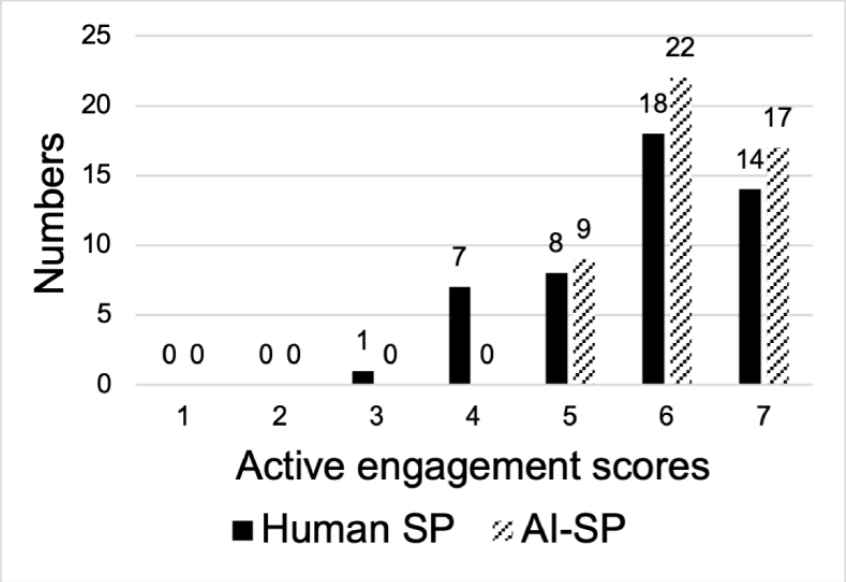


Figure 5. Active engagement in history taking scores in the Human SP and AI-SP conditions. AI-SP: AI-based standardized patient; SP: standardized patient.



Discussion

Principal Findings

This study compared the educational effects of the AI-SP and human SP conditions using the Mini-CEX and a self-rated questionnaire. Physician assessments based on the Mini-CEX showed significantly higher total scores in the human SP condition than in the AI-SP condition, with significantly higher scores observed for history taking and overall clinical competence; no significant differences between the 2 conditions were observed in communication, diagnosis, professionalism, or management. In the self-rated questionnaire, no significant differences were observed between conditions in learning satisfaction or perceived rapport formation, whereas active engagement in history taking was significantly higher in the AI-SP condition.

The higher history taking and overall clinical competence scores observed in the human SP condition should be interpreted as differences in performance under the two interaction conditions rather than as evidence of greater educational effectiveness of human SP training. Human SPs provide rich nonverbal cues, including facial expressions, vocal tone, pauses, and hesitation, which can prompt learners to explore and elicit clinically relevant information [14]. Moreover, nonverbal behaviors such as eye contact, laughter, and gestures are known to be critical for shaping the quality of physician-patient interactions during simulated medical interviews [15]. Advanced interpersonal skills, including waiting for patients to elaborate on their concerns, probing ambiguous concerns, and gathering information while responding sensitively to emotional distress [16], may be supported by interactions that incorporate such nonverbal information. In contrast, the AI-SP used in this study was limited to text-based communication. The absence of vocal

XSL•FO
RenderX

tone, facial expressions, pauses, and other nonverbal cues may have affected not only participants' information gathering but also their interpretation of the simulated patient's responses. In addition, the AI-generated responses may have been more coherent and logically structured, which may have reduced the extent to which emotional nuances and patient affect were conveyed. These differences in interactional characteristics may therefore have contributed to the observed differences in performance between the 2 conditions.

No significant differences were observed between the AI-SP and human SP conditions regarding communication, diagnosis, management, and professionalism. As this study was not designed as an equivalence or noninferiority trial, these findings should not be interpreted as evidence that AI-SPs are equivalent or noninferior to human SPs in these domains. These findings suggest that AI-SPs can provide specific learning opportunities for medical interview education but not to all learning domains. Further research is required to determine the extent of their educational effectiveness and roles relative to human SPs.

Another factor that should be considered is the potential for evaluator bias. Because blinding was not feasible, the evaluators were necessarily aware of whether they were assessing a face-to-face interaction with a human SP or a text-based chat with an AI-SP. This awareness may have implicitly influenced their judgment. In particular, the domain of overall clinical competence likely encompasses tacit and less formalized evaluation criteria such as interpersonal skills, situational awareness, and the ability to interpret the social and emotional context of the encounter. Evaluators may have perceived interactions with human SPs as being more "clinical" or "authentic," and this may have led them to assign higher ratings accordingly. Previous research has shown that when experts evaluate the outputs generated by humans and AI, they may exhibit a perception bias favoring content believed to be human generated even when objective output quality is equivalent between the 2 conditions [17]. Therefore, evaluator bias may have contributed, at least partially, to the observed differences between the AI-SP and human SP conditions in this study.

Participant self-assessments revealed no significant differences between the AI-SP and human SP conditions regarding learning satisfaction or degree of rapport formation. Notably, there was a similar degree of rapport formation with the AI-SP despite the interaction being conducted entirely through text. What this indicates is that a well-designed AI-SP can provide meaningful interactive experiences for learners even through a text-based format [18]. Furthermore, active engagement in history taking was significantly higher in the AI-SP condition. Although conventional interactions with SPs may be associated with psychological stress for some learners [19], psychological safety, anxiety, interpersonal tension, and fear of failure were not directly measured in the present study. Therefore, the higher self-rated active engagement observed in the AI-SP condition should not be interpreted as evidence of greater psychological safety. One possible hypothesis is that interaction with an AI-SP may reduce perceived interpersonal pressure and thereby facilitate active participation; however, this mechanism should be directly examined in future studies using validated measures of psychological safety and anxiety.

Additionally, AI-SPs may provide meaningful learning opportunities in domains evaluated within a structured framework of medical interviewing, such as communication and professionalism. In the context of communication training using AI-based chatbots, most of the sample in a previous study found the feedback useful and perceived improvements in physician-patient communication and recognition of mental health conditions [20]. As AI systems can deliver consistent responses according to predefined patterns, they may be particularly effective in helping learners acquire the fundamental structure and flow of medical interviews. As AI-SP technology continues to evolve, advances in multimodal systems that integrate voice, text, and visual information may enable more natural and realistic interactions that incorporate emotional expressions and nonverbal cues [21]. Such developments could help address some of the current limitations of AI-SPs in history taking and rapport building highlighted by the present study, further expanding their potential role in medical interview education.

Importantly, AI-SPs should not be viewed as direct replacements for human SPs but as complementary educational tools that can support the repeated practice of medical interviewing with a lower psychological burden on participants. One of the greatest advantages of AI-SPs is their ability to provide opportunities for repeated practice without time or location constraints. However, AI-SPs remain limited in their ability to reproduce communication through nonverbal cues such as facial expressions, vocal intonation, and silence [14,15]. In this respect, human SPs continue to play important and irreplaceable roles. Therefore, AI-SPs and human SPs serve distinct educational functions, and a staged educational approach that combines both modalities may represent an effective implementation strategy. Such an approach could involve the use of AI-SPs during initial practice stages followed by progression to interviews with human SPs as participants gain proficiency. In educational settings where securing human SPs is challenging and there is a lack of human resources, AI-SPs may also be actively used within the scope of their capabilities. The appropriate integration of these tools into medical education has the potential to expand educational opportunities and contribute to the training of future physicians.

Limitations

This study had several limitations. First, although the crossover design enabled within-participant comparisons between the AI-SP and human SP conditions and reduced the influence of individual differences in medical interviewing and communication skills, only 2 case scenarios were included. Therefore, it remains unclear whether similar findings would be observed across various diseases or more complex clinical presentations. In addition, although a washout period of at least 3 days was implemented, potential order effects and carryover effects resulting from prior interview experiences and feedback cannot be completely excluded. Moreover, 28 of the 76 randomized participants did not complete both study sessions and were excluded from the final analysis. Although the number of noncompleters was similar between the 2 sequences, the possibility of attrition bias cannot be excluded.

XSL•FO
RenderX

Second, this study included fifth-year medical students from a single medical school, which may limit the generalizability of the findings. The AI-SP used was limited to a text-based chat interface and did not provide nonverbal cues such as vocal characteristics, facial expressions, or other affective signals. Consequently, participants may have had difficulty assessing patients' understanding, anxiety, or degree of acceptance, limiting their opportunities to learn how to adapt explanations to patient needs and establish rapport through emotionally responsive communication. As nonverbal communication constitutes an important component of real-world clinical encounters, human SPs currently retain a substantial educational value that may not be fully replicated by chat-based AI systems. Results may differ when multimodal AI-SPs capable of incorporating voice, facial expressions, and other nonverbal cues are used.

Third, AI technology is evolving rapidly, and both the language model and prompt design used here may become outdated over time. Future AI implementation in medical education will require systematic approaches to prompt engineering as well as ongoing evaluation of the educational effectiveness and validity of virtual SPs.

Fourth, the Mini-CEX–based assessment was originally developed for face-to-face clinical encounters such that its construct validity in the context of AI-SP–based interviews is inherently limited. Domains that depend primarily on the content of the interaction, such as history taking, diagnosis, and management, may be more directly assessable in a text-based format. In contrast, domains that incorporate nonverbal and interpersonal behaviors, including communication, professionalism, and overall clinical competence, may not be assessed equivalently between text-based AI-SP and face-to-face human SP encounters. The present study was not designed to distinguish whether the observed differences in assessment scores reflected differences in participants' clinical competence or in the interactional characteristics of the 2 conditions and the interactional characteristics provided by the AI-SP. Future studies should consider developing and validating assessment frameworks specifically tailored to AI-mediated clinical communication environments.

Conclusions

This study shows how human SPs outperformed AI-SPs in the assessment of history taking and overall clinical competence, although there were no significant differences between the 2 conditions in communication, diagnosis, management, or professionalism. Meanwhile, participants demonstrated greater active engagement in history taking when interacting with AI-SPs. These findings suggest that human SPs remain superior in teaching history taking skills that require nonverbal interaction and advanced interpersonal communication. Thus, AI-SPs should not be viewed as direct replacements for human SPs, but they could serve as complementary educational tools that promote active learning and support early-stage practice in the stepwise acquisition of medical interviewing skills. Future research should focus on establishing assessment frameworks tailored to AI-SP–based learning environments and validating their educational effectiveness across multiple institutions to strengthen the evidence base for the implementation of AI-SPs in medical education.

Data Availability

The case questions and datasets related to the results of this study are available from the corresponding author (YY) on reasonable request.

Funding

This work was supported by Japan Society for the Promotion of Science Grant-in-Aid for Scientific Research JP25K20499. The article processing charge was funded by the Ministry of Health, Labour and Welfare, Japan, under its FY2026 physician training program.

Generative AI Use

During the revision of this manuscript, the authors used GPT-4o (OpenAI) to assist with language editing, restructuring, and refinement of the manuscript text. All AI-generated output was reviewed and revised by the authors, who take full responsibility for the final content.

Authors' Contributions

Conceptualization: YY, DY, TU
Formal analysis: YY, DY, FF
Investigation: YY, SI, RI, RS
Project administration: YY, DY, TU
Writing—original draft: YY
Writing—review and editing: DY, TU, FF
All authors have read and approved the final manuscript, agree to be accountable for all aspects of the work, and ensure that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved.

XSL•FO
RenderX

Conflicts of Interest

None declared.

Multimedia Appendix 1
Original and English-translated cases.
[DOCX File , 23 KB - mededu_v12i1e111030_app1.docx]

Multimedia Appendix 2
Original and English-translated prompts.
[DOCX File , 18 KB - mededu_v12i1e111030_app2.docx]

References

1. Abraham RM, Singaram VS. Using deliberate practice framework to assess the quality of feedback in undergraduate clinical skills training. BMC Med Educ 2019 Apr 11;19(1):105 [FREE Full text] [doi: 10.1186/s12909-019-1547-5] [Medline: 30975213]
2. Norcini JJ, Blank LL, Arnold GK, Kimball HR. The mini-CEX (clinical evaluation exercise): a preliminary investigation. Ann Intern Med 1995 Nov 15;123(10):795-799. [doi: 10.7326/0003-4819-123-10-199511150-00008] [Medline: 7574198]
3. Norcini JJ, Blank LL, Duffy FD, Fortna GS. The mini-CEX: a method for assessing clinical skills. Ann Intern Med 2003 Mar 18;138(6):476-481. [doi: 10.7326/0003-4819-138-6-200303180-00012] [Medline: 12639081]
4. Akatsu H, Shiima Y, Gomi H, Hegab AE, Kobayashi G, Naka T, et al. Teaching "medical interview and physical examination" from the very beginning of medical school and using "escape rooms" during the final assessment: achievements and educational impact in Japan. BMC Med Educ 2022 Jan 28;22(1):67 [FREE Full text] [doi: 10.1186/s12909-022-03130-2] [Medline: 35090459]
5. Harada Y, Kubota Y, Hirayama Y, Otaki J, Mitoma H. Medical school staff's standardized patient experience alters their understanding of student education. Tohoku J Exp Med 2022 Jan;256(1):63-71 [FREE Full text] [doi: 10.1620/tjem.256.63] [Medline: 35095029]
6. Hayashi M. The paradox caused by the standardized pre-clinical objective structured clinical examination in Japan. Korean J Med Educ 2025 Mar;37(1):85-87 [FREE Full text] [doi: 10.3946/kjme.2025.327] [Medline: 40049687]
7. Flanagan OL, Cummings KM. Standardized patients in medical education: a review of the literature. Cureus 2023 Jul 17;15(7):e42027 [FREE Full text] [doi: 10.7759/cureus.42027] [Medline: 37593270]
8. Bosse HM, Nickel M, Huwendiek S, Schultz JH, Nikendei C. Cost-effectiveness of peer role play and standardized patients in undergraduate communication training. BMC Med Educ 2015 Oct 24;15:183 [FREE Full text] [doi: 10.1186/s12909-015-0468-1] [Medline: 26498479]
9. Lucas HC, Upperman JS, Robinson JR. A systematic review of large language models and their implications in medical education. Med Educ 2024 Nov;58(11):1276-1285. [doi: 10.1111/medu.15402] [Medline: 38639098]
10. Jowsey T, Stokes-Parish J, Singleton R, Todorovic M. Medical education empowered by generative artificial intelligence large language models. Trends Mol Med 2023 Dec;29(12):971-973. [doi: 10.1016/j.molmed.2023.08.012] [Medline: 37718142]
11. Maicher KR, Stiff A, Scholl M, White M, Fosler-Lussier E, Schuler W, et al. Artificial intelligence in virtual standardized patients: combining natural language understanding and rule based dialogue management to improve conversational fidelity. Med Teach 2023 Mar;45(3):279-285. [doi: 10.1080/0142159X.2022.2130216] [Medline: 36346810]
12. Gray M, Baird A, Sawyer T, James J, DeBroux T, Bartlett M, et al. Increasing realism and variety of virtual patient dialogues for prenatal counseling education through a novel application of ChatGPT: exploratory observational study. JMIR Med Educ 2024 Feb 01;10:e50705 [FREE Full text] [doi: 10.2196/50705] [Medline: 38300696]
13. Aljedaani S, Sultan I, El-Tarhouny S. Knowledge-based chatbots in clinical teaching: AI-powered virtual patients to improve history-taking during internal medicine rotations. Int Med Educ 2026 May 21;5(2):51 [FREE Full text] [doi: 10.3390/ime5020051]
14. Park KH, Park SG. The effect of communication training using standardized patients on nonverbal behaviors in medical students. Korean J Med Educ 2018 Jun;30(2):153-159 [FREE Full text] [doi: 10.3946/kjme.2018.90] [Medline: 29860781]
15. D'Agostino TA, Bylund CL. Nonverbal accommodation in health care communication. Health Commun 2014;29(6):563-573 [FREE Full text] [doi: 10.1080/10410236.2013.783773] [Medline: 24138223]
16. Mast MS. On the importance of nonverbal communication in the physician-patient interaction. Patient Educ Couns 2007 Aug;67(3):315-318. [doi: 10.1016/j.pec.2007.03.005] [Medline: 17478072]
17. Franke Föyen L, Zapel E, Lekander M, Hedman-Lagerlöf E, Lindsäter E. Artificial intelligence vs. human expert: licensed mental health clinicians' blinded evaluation of AI-generated and expert psychological advice on quality, empathy, and perceived authorship. Internet Interv 2025 Jun 3;41:100841 [FREE Full text] [doi: 10.1016/j.invent.2025.100841] [Medline: 40525210]

XSL•FO
RenderX

18. Holderried F, Stegemann-Philipps C, Herschbach L, Moldt JA, Nevins A, Griewatz J, et al. A generative pretrained transformer (GPT)-powered chatbot as a simulated patient to practice history taking: prospective, mixed methods study. JMIR Med Educ 2024 Jan 16;10:e53961 [FREE Full text] [doi: 10.2196/53961] [Medline: 38227363]
19. Deinzer R, Kiupel S, Weik U. Endocrine and psychological stress response in simulated doctor-patient interactions in medical education. Psychoneuroendocrinology 2019 Jul;105:172-177. [doi: 10.1016/j.psyneuen.2018.09.028] [Medline: 30292650]
20. Herschbach L, Festl-Wietek T, Stegemann-Philipps C, Sonanini A, Herrmann B, Erschens R, et al. Evaluation of an AI-based chatbot providing real-time feedback in communication training for mental health care professionals: proof-of-concept observational study. J Med Internet Res 2025 Nov 28;27:e82818 [FREE Full text] [doi: 10.2196/82818] [Medline: 41314643]
21. Jacobs C, Johnson H, Brownlie K, Joiner R, Thompson T. Scaling multimodal agentic AI in medical education: multisite cross-sectional study of simulation effectiveness in primary care. JMIR Form Res 2026 Mar 23;10:e88905 [FREE Full text] [doi: 10.2196/88905] [Medline: 41871335]

Abbreviations

AI-SP: AI-based standardized patient
CONSORT-EHEALTH: Consolidated Standards of Reporting Trials of Electronic and Mobile Health Applications and Online Telehealth
Mini-CEX: Mini-Clinical Evaluation Exercise
SP: standardized patient

Edited by T de Azevedo Cardoso; submitted 02.Sep.2026; peer-reviewed by H Akatsu, X Liang, N Nguyen; comments to author 05.Oct.2026; revised version received 08.Oct.2026; accepted 26.Oct.2026; published 16.Nov.2026.

Please cite as:
Yanagita Y, Yokokawa D, Fukuzawa F, Ihara S, Yoshida R, Sato R, Uehara T
Medical Interview Performance Using an AI-Based Standardized Patient vs a Human Standardized Patient: Randomized Crossover Trial
JMIR Med Educ 2026;12:e111030
URL: https://mededu.jmir.org/2026/1/e111030
doi:10.2196/111030
PMID:

©Yasutaka Yanagita, Daiki Yokokawa, Fumitoshi Fukuzawa, Shiichi Ihara, Ryo Yoshida, Rurika Sato, Takanori Uehara. Originally published in JMIR Medical Education (https://mededu.jmir.org), 16.Nov.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Education, is properly cited. The complete bibliographic information, a link to the original publication on https://mededu.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Correction: Performance of GPT-4o and Claude in Medical Ethics Scenarios: Comparative Study

Karishma R Desai, BS; Anna L Gorsky, MD; Nicole A Zelenski, MD

Division of Hand Surgery, Department of Orthopaedic Surgery, Emory University, 80 Jesse Hill Jr Drive SE, Atlanta, GA, United States

Corresponding Author:
Karishma R Desai, BS
Division of Hand Surgery, Department of Orthopaedic Surgery, Emory University, 80 Jesse Hill Jr Drive SE, Atlanta, GA, United States

Related Article:

Correction of: https://mededu.jmir.org/2026/1/e70199

(JMIR Med Educ 2026;12:e113933) doi:10.2196/113933

In "Performance of GPT-4o and Claude in Medical Ethics Scenarios: Comparative Study" [1], the authors noted 3 errors. The below changes have been made.

First, Table 2 incorrectly listed the denominator for the GTP-4o UWorld Interpersonal-Specific questions as 11; the correct denominator is 10. The AMBOSS numerator and denominator for the GPT-4o questions were listed as 1/2 and have been corrected to 9/18. The revised table is presented as Table 1 below, with changes formatted in italics.

Table . Incorrectly answered questions per question set for which GPT-4o and Claude selected the incorrect answer most commonly chosen by human test takers.

Question set	GPT-4o[a], n/N (%)	Claude[b], n/N (%)
UWorld All Social Sciences	6/13 (46)	5/7 (71)
UWorld Interpersonal-Specific	*9/10 (90)*	4/4 (100)
AMBOSS	*9/18 (50)*	3/4 (75)
Orthobullets Ethics	6/7 (86)	4/4 (100)
Orthobullets Legal	7/12 (58)	1/1 (100)

[a]Total: 37/60 (62%).

[b]Total: 17/20 (85%).

Second, the comparison between incorrect responses was mistakenly listed as significant during the extensive recalculation of statistics; the χ^2 value of this test (P=.03) was incorrectly used instead of the previously used and more appropriate Fisher exact interpretation (P=.07). Therefore, a few manuscript sections were amended.

In the Results section of the Abstract, the following sentence was revised:

> *GPT-4o selected the incorrect response for 60 (30%) total questions and chose the incorrect response most commonly selected by humans significantly more frequently on UWorld interpersonal-specific questions (30/40, 75%) than on UWorld all social sciences (27/40, 68%; P=.03).*

The sentence now reads:

> *Neither GPT-4o or Claude showed a significant difference in the rate of most common incorrect response selection between question categories.*

The following sentence from the Results was revised:

> *Analysis did show a significant difference in the frequency that GPT-4o chose the most popular incorrect answer between both UWorld general social sciences and interpersonal-specific sets (P=.03), but no difference between the UWorld (general and interpersonal) and AMBOSS (legal medicine and ethics) sets (P=0.33), Orthobullets Ethics and Legal sets (P=.22), or medical student general and orthopedic resident–level (UWorld/AMBOSS and Orthobullets) sets (P=.46; Table 2).*

The sentence now reads:

> *Analysis did not show a significant difference in the frequency that GPT-4o chose the most popular incorrect answer between both UWorld general social sciences and interpersonal-specific sets (P=.07), UWorld (general and interpersonal) and AMBOSS (legal medicine and ethics) sets (P=.33), Orthobullets Ethics and Legal sets (P=.22), or medical student general and orthopedic resident–level*

XSL•FO
RenderX

(UWorld/AMBOSS and Orthobullets) sets (P=.46; Table 2).

In the Implications and Conclusions section of the Discussion, the following sentences were revised:

Further studies designed to explore variations in training data and their effects on ethical-based scenario model output would be helpful in making this distinction. For instance, it is not clear why GPT-4o chose the most popular incorrect answer significantly more frequently on the UWorld interpersonal questions than on UWorld Social Sciences, but Claude, which is designed to be more conversational, did not.

The revised sentences now read:

Further studies designed to explore variations in training data and their effects on ethical-based scenario model output would be helpful in making this distinction.

Lastly, the editor requested that a statement be provided as to whether any limitations to the scope of the article are created due to the public unavailability of the question bank benchmarks. The following paragraph was added to the end of the Strengths and Limitations section of the Discussion:

The public unavailability of the question bank benchmarks limit the statistical interpretation of the human test taker comparisons in this study to purely descriptive, as described in the Results and Discussion. The benchmarks provide a helpful contextual reference point for readers who are unfamiliar with question bank–specific performance averages. However, the limitation created by the lack of publicly available data is that the human test taker performance in our study cannot be interpreted as representative of human performance. The main aim of our study was to observe and document potential differences in large language model–specific performance. Therefore, the findings of this study should not be used to make conclusions about model versus human performance.

Fortunately, these corrections have only a minor effect on the results as reported and do not alter the overall interpretation or conclusions of the manuscript. We appreciate the efforts of the editorial office in identifying these errors, allowing us the opportunity for correction, and confirming that the findings of the article are not affected.

The correction will appear in the online version of the paper on the JMIR Publications website together with the publication of this correction notice. Because this was made after submission to PubMed, PubMed Central, and other full-text repositories, the corrected article has also been resubmitted to those repositories.

Reference

1. Desai KR, Gorsky AL, Zelenski NA. Performance of GPT-4o and Claude in medical ethics scenarios: comparative study. JMIR Med Educ 2026 Jul 23;12:e70199. [doi: 10.2196/70199] [Medline: 42492056]

Submitted 08.Oct.2026; this is a non–peer-reviewed article;accepted 14.Oct.2026; published 11.Nov.2026.

Please cite as:
Desai KR, Gorsky AL, Zelenski NA
Correction: Performance of GPT-4o and Claude in Medical Ethics Scenarios: Comparative Study
JMIR Med Educ 2026;12:e113933
URL: https://mededu.jmir.org/2026/1/e113933
doi:10.2196/113933

© Karishma R Desai, Anna L Gorsky, Nicole A Zelenski. Originally published in JMIR Medical Education (https://mededu.jmir.org), 11.Nov.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Education, is properly cited. The complete bibliographic information, a link to the original publication on https://mededu.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Correction: Upskilling in Healthy Longevity Medicine and Its Association With Physicians' Implementation Intent and Self-Reported Clinical Confidence: Cross-Sectional Observational Study

Evelyne Bischof[1,2], Dr med, MD, PhD; Dominika Wilczok[3,4]; James L Kirkland[5], MD, PhD, Dr Med; Bhirau Wilaksono[6,7], MD; Christine Yuan Huang[8], MD; Suwanna Suwannaphong[9], MD; Wanviput Sanphasitvong[9], MD; Dalila Čamdžić[10], MD; Carolina Hernandez[11], BSc; Yoko Madea[12], MD; Hidekazu Yamada[13], MD; Melissa Alexandre Fernandes[14], MD; Ricardo Gaminha Pacheco[15], MD, MBA; Fabiano M Serfaty[16], MD, PhD; Fernanda Calvo-Fortes[17], MSc; Amit Goldman[18], EMBA; Andrea B Maier[19,20,21], MD; Alexey Moskalev[22], PhD; Morten Scheibye-Knudsen[23], MD, PhD; Alex Zhavoronkov[24], PhD

[1]Sheba Longevity Center, Sheba Medical Center, Gray Faculty of Medical & Health Sciences, Tel Aviv University, Sheba Road 2, Ramat Gan, Tel Aviv, Israel

[10]Sarajevo School of Science and Technology, Sarajevo, Bosnia and Herzegovina

[11]Departent of Biomedicine, University of Barcelona, Barcelona, Spain

[12]Madea Clinic, Okayama, Japan

[13]Anti-Aging Center, Kindai University, Osaka, Japan

[14]Novartis Institutes for BioMedical Research, Basel, Switzerland

[15]Insilico Medicine Hong Kong Ltd, Hong Kong, China (Hong Kong)

[16]Departent of Biomedicine, Universidade do Estado do Rio de Janeiro, Rio de Janeiro, Brazil

[17]School of Nursing, Federal University of Minas Gerais, Belo Horizonte, Brazil

[18]Saïd Business School, University of Oxford, Oxford, United Kingdom

[19]Healthy Longevity Translational Research Program, Yong Loo Lin School of Medicine, National University of Singapore, Singapore, Singapore

[2]Shanghai University of Medicine and Health Sciences, Shanghai, China

[20]NUS Academy for Healthy Longevity, Yong Loo Lin School of Medicine, National University of Singapore, Singapore, Singapore

[21]Department of Human Movement Sciences, Faculty of Behavioral and Movement Sciences, Amsterdam Movement Sciences, Vrije Universiteit Amsterdam, Amsterdam, The Netherlands

[22]Petrovsky Russian Research Center for Surgery, Moscow, Russian Federation

[23]Center for Healthy Aging, Department of Cellular and Molecular Medicine, University of Copenhagen, Copenhagen, Denmark

[24]Insilico Medicine US Inc, Cambridge, MA, United States

[3]Duke University, Durham, NC, United States

[4]Duke Kunshan University, Kunshan, Jiangsu, China

[5]Center for Advanced Gerotherapeutics, Division of Endocrinology and Metabolism, Department of Medicine, Cedars Sinai Health Sciences University, Los Angeles, CA, United States

[6]Center for Healthy Longevity, Faculty of Medicine, Universitas Negeri Makassar, Makassar, Indonesia

[7]Department of Biochemistry, Faculty of Medicine, Universitas Negeri Makassar, Makassar, Indonesia

[8]Quantum Life Inc, Hong Kong, China (Hong Kong)

[9]VitalLife, Bumrungrad International Hospital, Bangkok, Thailand

Corresponding Author:
Evelyne Bischof, Dr med, MD, PhD
Sheba Longevity Center, Sheba Medical Center, Gray Faculty of Medical & Health Sciences, Tel Aviv University, Sheba Road 2, Ramat Gan, Tel Aviv, Israel

Related Article:

Correction of: https://mededu.jmir.org/2026/1/e83779

(JMIR Med Educ 2026;12:e114811) doi:10.2196/114811

In "Upskilling in Healthy Longevity Medicine and Its Association With Physicians' Implementation Intent and

XSL•FO
RenderX

Self-Reported Clinical Confidence: Cross-Sectional Observational Study" [1], the authors have identified the need to provide further details in the Conflicts of Interest section.

The original disclosure identified several authors as volunteers of Longevity Education Hub but did not specify their individual responsibilities in the development, delivery, administration, and promotion of the educational resources evaluated in the study. It also omitted author ABM's position as chief medical officer of NU. The following correction provides these additional disclosures in the interest of transparency.

In the originally published article, the Conflicts of Interest section read as follows:

> *EB, DW, BW, CYH, SS, WS, DČ, CH, YM, HY, MAF, RGP, MS-K, AM, AG, FC-F, FS, and AZ are volunteers of the Longevity Education Hub—the nonprofit online education platform for Healthy Longevity Medicine Courses. AZ is the founder and shareholder of Insilico Medicine, a generative artificial intelligence and quantum computing–enabled longevity biotechnology company. All other authors declare no competing interests. AM is Director of the NUS Academy for Healthy Longevity (disclosure, not a conflict of interest).*

The Conflicts of Interest section has been revised to read as follows:

> *EB, DW, BW, CYH, SS, WS, DČ, CH, YM, HY, MAF, RGP, MS-K, AM, AG, FC-F, FMS, and AZ are volunteers of the Longevity Education Hub (LEH), a volunteer-run, noncommercial educational initiative offering free, online Healthy Longevity Medicine courses. None of the authors receives remuneration or holds an ownership interest in LEH. The following authors have additional nonfinancial roles in the initiative and its educational resources evaluated in this study: AZ and EB cofounded the initiative and conceptualized, developed, and authored course curricula and educational materials; MS-K and AM contributed scientific expertise and coauthored educational content; DW and DČ were responsible for communications, social media outreach, website coordination, and operational support, including the dissemination and accessibility of the courses. DČ additionally undertook course management and administrative coordination.*
>
> *AZ is the founder and shareholder of Insilico Medicine, a generative AI and quantum computing–enabled longevity biotechnology company. ABM is chief medical officer of NU, a software company that provides analysis of health data and does not offer educational programs. ABM is also director of the NUS Academy for Healthy Longevity (disclosure, not a conflict of interest). All other authors declare no competing interests.*

This correction clarifies the authors' disclosed relationships and does not alter the study results, analyses, or conclusions.

The correction will appear in the online version of the paper on the JMIR Publications website together with the publication of this correction notice. Because this was made after submission to PubMed, PubMed Central, and other full-text repositories, the corrected article has also been resubmitted to those repositories.

Reference

1. Bischof E, Wilczok D, Kirkland JL, et al. Upskilling in healthy longevity medicine and its association with physicians' implementation intent and self-reported clinical confidence: cross-sectional observational study. JMIR Med Educ 2026 Mar 19;12:e83779. [doi: 10.2196/83779] [Medline: 41855480]

Submitted 21.Oct.2026; this is a non–peer-reviewed article;accepted 30.Oct.2026; published 10.Nov.2026.

Please cite as:
Bischof E, Wilczok D, Kirkland JL, Wilaksono B, Huang CY, Suwannaphong S, Sanphasitvong W, Čamdžić D, Hernandez C, Madea Y, Yamada H, Fernandes MA, Pacheco RG, Serfaty FM, Calvo-Fortes F, Goldman A, Maier AB, Moskalev A, Scheibye-Knudsen M, Zhavoronkov A
Correction: Upskilling in Healthy Longevity Medicine and Its Association With Physicians' Implementation Intent and Self-Reported Clinical Confidence: Cross-Sectional Observational Study
JMIR Med Educ 2026;12:e114811
URL: https://mededu.jmir.org/2026/1/e114811
doi:10.2196/114811

© Evelyne Bischof, Dominika Wilczok, James L Kirkland, Bhirau Wilaksono, Christine Yuan Huang, Suwanna Suwannaphong, Wanviput Sanphasitvong, Dalila Čamdžić, Carolina Hernandez, Yoko Madea, Hidekazu Yamada, Melissa Alexandre Fernandes, Ricardo Gaminha Pacheco, Fabiano M Serfaty, Fernanda Calvo-Fortes, Amit Goldman, Andrea B Maier, Alexey Moskalev, Morten Scheibye-Knudsen, Alex Zhavoronkov. Originally published in JMIR Medical Education (https://mededu.jmir.org), 10.Nov.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License

XSL•FO
RenderX

(https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Education, is properly cited. The complete bibliographic information, a link to the original publication on https://mededu.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Correction: Performance of ChatGPT on Nursing Licensure Examinations in the United States and China: Cross-Sectional Study

Zelin Wu[1*], PhD; Wenyi Gan[2*], PhD; Zhaowen Xue[1*], PhD; Zhengxin Ni[3*], PhD; Xiaofei Zheng[1], PhD; Yiyi Zhang[1], PhD

[1]Department of Bone and Joint Surgery and Sports Medicine Center, The First Affiliated Hospital of Jinan University, No.613, Huangpu Avenue West, Tianhe District, Guangzhou, Guangdong, China

[2]Department of Joint Surgery and Sports Medicine, Zhuhai People's Hospital (Zhuhai Clinical Medical College of Jinan University), Zhuhai, Guangdong, China

[3]School of Nursing, Yangzhou University, Yangzhou, Jiangsu, China

[*]these authors contributed equally

Corresponding Author:
Yiyi Zhang, PhD
Department of Bone and Joint Surgery and Sports Medicine Center, The First Affiliated Hospital of Jinan University, No.613, Huangpu Avenue West, Tianhe District, Guangzhou, Guangdong, China

Related Article:

Correction of: https://mededu.jmir.org/2024/1/e52746

(JMIR Med Educ 2026;12:e105550) doi:10.2196/105550

In "Performance of ChatGPT on Nursing Licensure Examinations in the United States and China: Cross-Sectional Study" [1], an error was identified.

The affiliation of authors ZW, ZX, XZ, and YZ was incorrectly listed as:

> *Department of Bone and Joint Surgery and Sports Medicine Center, The First Affiliated Hospital, Guangzhou, China*

This has been corrected to the following:

> *Department of Bone and Joint Surgery and Sports Medicine Center, The First Affiliated Hospital of Jinan University, Guangzhou, China*

This correction appeared in the online version of the article on the JMIR Publications website, together with this correction notice. Because this correction was made after submission to PubMed, PubMed Central, and other full-text repositories, the corrected article was resubmitted to these repositories.

Reference

1. Wu Z, Gan W, Xue Z, Ni Z, Zheng X, Zhang Y. Performance of ChatGPT on nursing licensure examinations in the United States and China: cross-sectional study. JMIR Med Educ 2024 Oct 3;10:e52746. [doi: 10.2196/52746] [Medline: 39363539]

Submitted 25.Jun.2026; this is a non–peer-reviewed article;accepted 14.Jul.2026; published 24.Jul.2026.

Please cite as:
Wu Z, Gan W, Xue Z, Ni Z, Zheng X, Zhang Y
Correction: Performance of ChatGPT on Nursing Licensure Examinations in the United States and China: Cross-Sectional Study
JMIR Med Educ 2026;12:e105550
URL: https://mededu.jmir.org/2026/1/e105550
doi:10.2196/105550

© Zelin Wu, Wenyi Gan, Zhaowen Xue, Zhengxin Ni, Xiaofei Zheng, Yiyi Zhang. Originally published in JMIR Medical Education (https://mededu.jmir.org), 24.Jul.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Education, is properly cited. The

XSL•FO
RenderX

complete bibliographic information, a link to the original publication on https://mededu.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Correction: Augmented Reality in Surgical Training: Systematic Review of Its Impact on Technical Performance in Surgical Trainees

Mahmoud El Ashry[1*], MBChB; Ahmed El Ashry[1*], MBChB, MSc; Hamza Khalique[1], MBChB; Yahye Abdalle[1], MBChB; Thomas Yeung[2], MBBCh

[1]Bristol Medical School, University of Bristol, Beacon House, Queens Road, Bristol, United Kingdom

[2]ENT/Otolaryngology Department, University Hospitals of North Midlands NHS Trust, Stoke-on-Trent, United Kingdom

[*]these authors contributed equally

Corresponding Author:
Mahmoud El Ashry, MBChB
Bristol Medical School, University of Bristol, Beacon House, Queens Road, Bristol, United Kingdom

Related Article:

Correction of: https://mededu.jmir.org/2026/1/e71572

(JMIR Med Educ 2026;12:e106576) doi:10.2196/106576

In "Augmented Reality in Surgical Training: Systematic Review of Its Impact on Technical Performance in Surgical Trainees" [1], the authors made one correction.

The fourth author's name was amended to read from the following:

Yahyea Abdalle

The author's name now reads:

Yahye Abdalle

The correction will appear in the online version of the paper on the JMIR Publications website, together with the publication of this correction notice. Because this was made after submission to PubMed, PubMed Central, and other full-text repositories, the corrected article has also been resubmitted to those repositories.

Reference

1. El Ashry M, El Ashry A, Khalique H, Abdalle Y, Yeung T. Augmented reality in surgical training: systematic review of its impact on technical performance in surgical trainees. JMIR Med Educ 2026 Jun 3;12:e71572. [doi: 10.2196/71572] [Medline: 42234977]

Submitted 08.Jul.2026; this is a non–peer-reviewed article;accepted 08.Jul.2026; published 23.Jul.2026.

Please cite as:
El Ashry M, El Ashry A, Khalique H, Abdalle Y, Yeung T
Correction: Augmented Reality in Surgical Training: Systematic Review of Its Impact on Technical Performance in Surgical Trainees
JMIR Med Educ 2026;12:e106576
URL: https://mededu.jmir.org/2026/1/e106576
doi:10.2196/106576

© Mahmoud El Ashry, Ahmed El Ashry, Hamza Khalique, Yahye Abdalle, Thomas Yeung. Originally published in JMIR Medical Education (https://mededu.jmir.org), 23.Jul.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Education, is properly cited. The complete bibliographic information, a link to the original publication on https://mededu.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Correction: Collaborative Simulation-Based Learning to Develop Competencies in Physical Therapy Students: Randomized Controlled Trial

Iván Cuyul-Vásquez[1,2*], PT, MSc; Hilda Aravena-Páez[3*], PT, MSc; Felipe Araya-Quintanilla[4], PT, PhD; Sergio Guinez-Molinos[5], PhD; Rodrigo Núñez-Cortés[6], PT, PhD; Héctor Gutiérrez Espinoza[7,8], PT, PhD

[1]Departamento de Procesos Terapéuticos, Facultad de Ciencias de la Salud, Universidad Católica de Temuco, Temuco, Chile

[2]Facultad de Ciencias de la Salud, Universidad Arturo Prat, Sede Victoria, Chile

[3]Facultad de Medicina, Departamento de Clínicas, Kinesiología, Universidad Católica del Norte, Coquimbo, Chile

[4]Escuela de Kinesiología, Facultad de Ciencias de la Rehabilitación y Calidad de Vida, Universidad San Sebastián, Santiago, Chile

[5]Laboratorio de Simulación e Informática Biomédica, Facultad de Medicina, Universidad de Talca, Talca, Chile

[6]Department of Physical Therapy, Faculty of Medicine, University of Chile, Santiago, Chile

[7]Faculty of Education, Universidad Autónoma de Chile, Av. Pedro de Valdivia 425, Santiago, Chile

[8]Universidad Espíritu Santo, Samborondón, Ecuador

[*]these authors contributed equally

Corresponding Author:
Héctor Gutiérrez Espinoza, PT, PhD
Faculty of Education, Universidad Autónoma de Chile, Av. Pedro de Valdivia 425, Santiago, Chile

Related Article:

Correction of: https://mededu.jmir.org/2026/1/e85442

(JMIR Med Educ 2026;12:e112897) doi:10.2196/112897

In "Collaborative Simulation-Based Learning to Develop Competencies in Physical Therapy Students: Randomized Controlled Trial" [1], a conflicts of interest declaration was omitted.

The Conflicts of Interest statement has been added as follows:

> *SG-M serves on the JMIR Publications Editorial Board as an associate editor of JMIR Medical Education. However, he was not involved in the review process for this manuscript.*

The correction will appear in the online version of the paper on the JMIR Publications website together with the publication of this correction notice. Because this was made after submission to PubMed, PubMed Central, and other full-text repositories, the corrected article has also been resubmitted to those repositories.

Reference

1. Cuyul-Vásquez I, Aravena-Páez H, Araya-Quintanilla F, Guinez-Molinos S, Núñez-Cortés R, Gutiérrez Espinoza H. Collaborative simulation-based learning to develop competencies in physical therapy students: randomized controlled trial. JMIR Med Educ 2026 Sep 21;12:e85442. [doi: 10.2196/85442] [Medline: 42766711]

Submitted 24.Sep.2026; this is a non–peer-reviewed article;accepted 24.Sep.2026; published 30.Oct.2026.

Please cite as:
Cuyul-Vásquez I, Aravena-Páez H, Araya-Quintanilla F, Guinez-Molinos S, Núñez-Cortés R, Gutiérrez Espinoza H
Correction: Collaborative Simulation-Based Learning to Develop Competencies in Physical Therapy Students: Randomized Controlled Trial
JMIR Med Educ 2026;12:e112897
URL: https://mededu.jmir.org/2026/1/e112897
doi:10.2196/112897

XSL•FO
RenderX

© Iván Cuyul-Vásquez, Hilda Aravena-Páez, Felipe Araya-Quintanilla, Sergio Guinez-Molinos, Rodrigo Núñez-Cortés, Héctor Gutiérrez Espinoza. Originally published in JMIR Medical Education (https://mededu.jmir.org), 30.Oct.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Education, is properly cited. The complete bibliographic information, a link to the original publication on https://mededu.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Correction: Attitudes Toward Digital Planetary Health Before and After a Lecture Series on Environmental Awareness and Future Visions: Repeated Cross-Sectional Survey Study Regarding Twin Transformation

Julia Nitsche[1*], Dr rer medic; Theresa Sophie Busse[2*], Prof Dr rer medic; Sonja Schmalen[3], MBA; Ina Otte[4], Prof Dr sc med; Jan Peter Ehlers[1], Prof Dr med vet

[1]Department of Didactics and Educational Research in Health Science, Faculty of Health, Witten/Herdecke University, Alfred-Herrhausen-Straße 50, Witten, North Rhine-Westphalia, Germany

[2]Junior Professorship for Digital Health, Faculty of Health, Witten/Herdecke University, Witten, North Rhine-Westphalia, Germany

[3]Health for Future, Deutsche Allianz für Klimawandel und Gesundheit, Berlin, Germany

[4]Department for Health Care Research, Institute for Diversity Medicine, Ruhr University Bochum, Bochum, North Rhine-Westphalia, Germany

[*]these authors contributed equally

Corresponding Author:
Julia Nitsche, Dr rer medic
Department of Didactics and Educational Research in Health Science, Faculty of Health, Witten/Herdecke University, Alfred-Herrhausen-Straße 50, Witten, North Rhine-Westphalia, Germany

Related Article:

Correction of: https://mededu.jmir.org/2026/1/e89259/

(JMIR Med Educ 2026;12:e111087) doi:10.2196/111087

In "Attitudes Toward Digital Planetary Health Before and After a Lecture Series on Environmental Awareness and Future Visions: Repeated Cross-Sectional Survey Study Regarding Twin Transformation" [1], the authors noted three errors.

The degree for author TSB was changed from the following:

Dr rer medic

The degree now reads:

Prof Dr rer medic

In Table 1, the following text in the first row has been revised:

JN, TSB, SSchS

The text now reads:

JN, TSB, SSch

Finally, in the Background section, the following sentence was revised:

Lecturers from Witten/Herdecke University (UW/H; JN, TSB, and JPE), German Climate Change and Health Alliance (SSch), and Ruhr University Bochum (RUB; IO)

The sentence now reads:

Lecturers from Witten/Herdecke University (UW/H; JN and JPE), German Climate Change and Health Alliance (SSch), and Ruhr University Bochum (RUB; TSB)

The corrections will appear in the online version of the paper on the JMIR Publications website together with the publication of this correction notice. Because this was made after submission to PubMed, PubMed Central, and other full-text repositories, the corrected article has also been resubmitted to those repositories.

Reference

1. Nitsche J, Busse TS, Schmalen S, Otte I, Ehlers JP. Attitudes toward digital planetary health before and after a lecture series on environmental awareness and future visions: repeated cross-sectional survey study regarding twin transformation. JMIR Med Educ 2026 Aug 4;12:e89259. [doi: 10.2196/89259] [Medline: 42550983]

XSL•FO
RenderX

Submitted 02.Sep.2026; this is a non–peer-reviewed article;accepted 02.Sep.2026; published 25.Sep.2026.

Please cite as:
Nitsche J, Busse TS, Schmalen S, Otte I, Ehlers JP
Correction: Attitudes Toward Digital Planetary Health Before and After a Lecture Series on Environmental Awareness and Future Visions: Repeated Cross-Sectional Survey Study Regarding Twin Transformation
JMIR Med Educ 2026;12:e111087
URL: https://mededu.jmir.org/2026/1/e111087
doi:10.2196/111087

© Julia Nitsche, Theresa Sophie Busse, Sonja Schmalen, Ina Otte, Jan Peter Ehlers. Originally published in JMIR Medical Education (https://mededu.jmir.org), 25.Sep.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Education, is properly cited. The complete bibliographic information, a link to the original publication on https://mededu.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Re-Centering Clinical Documentation in the Age of AI Scribes: Four Aims of the Patient Chart Note

Chwen-Yuen Angie Chen[1*], MD; Tiffany I Leung[2,3*], MPH, MD; Rika Bajra[1], MD

[1]Division of Primary Care Population Health, Department of Medicine, Stanford University, 211 Quarry Road, 4th Floor, Palo Alto, CA, United States

[2]JMIR Publications, Toronto, ON, Canada

[3]Department of Internal Medicine, Southern Illinois University School of Medicine, Springfield, IL, United States

[*]these authors contributed equally

Corresponding Author:
Chwen-Yuen Angie Chen, MD
Division of Primary Care Population Health, Department of Medicine, Stanford University, 211 Quarry Road, 4th Floor, Palo Alto, CA, United States

Abstract

Clinical documentation is a foundational skill in medicine, developed during training and required in everyday practice. Historically, the chart note functioned as a clinician-centered cognitive tool for reasoning, teaching, and communication but has evolved into a multipurpose document shaped by administrative, regulatory, and financial demands, and is increasingly experienced as burdensome. The electronic health record, intended to improve efficiency, has introduced additional complexity and workflow strain, contributing to clinician burnout. Ambient artificial intelligence (AI) scribe technologies are rapidly being adopted to address these challenges, yet their implementation has outpaced evidence regarding their impact on learning, cognition, and clinical reasoning. We raise questions regarding the underexplored consequences of AI-assisted documentation, particularly cognitive off-loading and the potential for de-skilling, echoing historical concerns surrounding earlier cognitive technologies that externalized thought. We propose a practical framework that re-centers clinical documentation around four core aims: supporting clinical reasoning ("note to self"), facilitating communication ("note to others"), meeting medicolegal and billing requirements, and enhancing patient education in the era of open notes. Incorporating this framework into training may promote more intentional documentation practices before routine reliance on AI. We advocate for reframing the chart note to support clinician development and preserve its role in high-quality, patient-centered care.

(JMIR Med Educ 2026;12:e95218) doi:10.2196/95218

KEYWORDS

medical education; clinical training; artificial intelligence; ambient scribe; medical documentation

Clinical documentation has evolved beyond its original, singular purpose of supporting patient care. As health care systems expanded to include the goals of prevention, population health, and data portability, documentation requirements grew in scope and complexity [1]. Secondary uses of the electronic health record (EHR) data, including billing, quality reporting, and regulatory compliance, have further shaped documentation practices and system design [2,3]. What was intended to improve efficiency and care coordination also introduced a significant administrative burden and workflow strain, contributing to clinician burnout and lost revenue from unclosed charts [4-6]. In their systematic review, using a rigorous search strategy across multiple databases assessing EHR documentation, Colicchio and Cimino [7] elaborate on how EHR documentation now supports multiple overlapping and at times competing purposes that have transformed clinical chart notes into a multipurpose artifact, often diluting its primary role as a tool for clinical reasoning and communication. Their review suggests that EHR documentation often fails to adequately capture clinicians' integrated understanding of the patient: how presentation, problems, interventions, and goals are meaningfully connected as well as what clinicians are "thinking" about the patients' problems.

Considering these challenges, a re-centering of the purpose of clinical documentation is needed. In this context, we propose four core aims of the clinical chart note as a pragmatic, clinician-centered framework that selectively prioritizes key functions most essential to patient care and professional development. This "four aims" framework can be viewed as a normative, pedagogical approach designed to help clinicians and trainees navigate these tensions of overlapping purposes and increasing external pressures of clinical charting, while reflecting both the historical functions and enduring purpose of clinical documentation, especially in the era of artificial intelligence (AI)–assisted ambient documentation. The work of Colicchio and Cimino [7] also invites renewed focus on the most used section of the note—the assessment and plan—as an opportunity to clearly articulate clinical reasoning and synthesize the patient's situation, including social determinants of health, in a language accessible to both clinicians and patients.

XSL•FO
RenderX

The four aims are ordered to mirror the historical development of medical documentation. The first two purposes, delivering good patient care, remain unchanged; the third purpose arose with malpractice liability; and as open notes have become widespread, the fourth purpose of patient education has evolved:

1. Reminder to self: The chart note serves as a reminder of clinical reasoning, salient findings, and interventions, supporting continuity of care.
2. Communication with other professionals: Documentation communicates pertinent information and rationale to colleagues for care continuity.
3. Medicolegal and administrative record: Accurate documentation supports legal protection, quality control, and financial requirements; activity that one forgets to document essentially did not occur, and activity that did not occur should not be carried over in a note.
4. Patient education: Sharing information in lay terms educates patients and their caregivers; open notes support transparency and ownership of one's own health information and allow patients into clinical reasoning.

Examining the history of clinical documentation reveals that the burden of charting is not inherent to the act itself, but rather a consequence of how its purpose has expanded over time to accommodate administrative, regulatory, and financial demands. Understanding this evolution helps contextualize modern burnout and creates space to re-center documentation around its most meaningful functions. Looking back, medical documentation is as old as cave illustrations of injury, healing rituals, and death [8]. With the advent of written records, ancient civilizations documented illness and treatment to transfer knowledge across generations. Over time, these ancient scripts evolved from narrative accounts into more structured approaches to catalogs of patient cases for education and communication [8]. In 1968, Dr Lawrence Weed introduced the subjective, objective, assessment, and plan (SOAP) note format [9], which was followed by the problem-oriented medical record (POMR) [10] and later adaptations such as the assessment, plan, subjective, and objective (APSO) format [11]. These frameworks were eventually embedded within the EHR.

The rise of human medical scribes over the past decade reflects an effort to off-load documentation tasks in response to increasing administrative burden [12]. However, the impact of scribes, whether human or AI, on documentation quality and clinician workload remains mixed [13]. Clinicians' individual narrative styles may support recall and understanding, including an emotional connection with the patient, thus creating a cognitive imprint that may be diminished when documentation is delegated. Yet the widespread and rapid adoption of AI-assisted documentation has outpaced evidence regarding its effect on cognition and clinical reasoning [14,15]. Emerging evidence suggests that the use of large language models in writing tasks may reduce memory recall, perceived ownership, and the ability to reproduce one's own work [16]. More importantly, it is now widely discussed how reliance on AI scribes introduces cognitive off-loading that may reduce clinicians' active engagement in synthesizing and documenting clinical encounters, processes essential to critical thinking and clinical reasoning [17,18].

Concerns about cognitive off-loading are not new. Early critics of writing warned that externalizing knowledge would weaken memory and internal understanding [8]. The evolution of documentation reflects an ongoing balance between cognitive work, efficiency, and standardization. Early narrative records were an expression of the individual clinician's interpretation and memory of a clinical encounter or experience, while the introduction of structured formats such as SOAP and POMR reflected a shift toward formalizing clinical reasoning for teaching and reproducibility. Each transition introduced trade-offs: increased standardization improved communication and scalability but risked constraining narrative nuance and individual cognitive processes. Educators adapted by incorporating these formats into training, using documentation as a tool to externalize and teach clinical reasoning. Dictation devices, EHR templates, copy-and-paste practices, and now AI-assisted documentation all represent a continuation of these shifts. However, unlike prior changes, AI introduces the possibility of not only delegating documentation tasks but also off-loading the process of cognitive work itself. This distinction raises important questions about how current innovations compare with prior disruptions and whether they may fundamentally alter the role of documentation as a tool to teach clinical reasoning in medical education.

Educators have raised concerns that cognitive off-loading from AI tools may diminish the "desirable difficulties" necessary for developing deep understanding and critical thinking [19-21]. Historically, trainees have grappled with writing patient encounter notes, first by hand and later digitally, a productive struggle that supports the development of essential clinical reasoning skills. Although the longitudinal impact of ambient AI scribes remains unclear, educators caution that introducing these tools before trainees establish foundational competencies may lead to "never-skilling," in which essential skills fail to develop due to premature cognitive off-loading [17,18,22]. Even for advanced learners, passive reliance on AI-generated drafts risks automation bias [17], allowing errors or embedded biases to go unrecognized and potentially negatively impacting patient-provider relationships and the quality of care. A competency-based approach [23-27] would introduce ambient AI scribes only after learners demonstrate readiness through observable documentation behaviors, supported by entrustment decisions indicating their ability to independently achieve the four aims of clinical documentation. This transition should be scaffolded by a longitudinal AI curriculum that assesses competencies in critical appraisal of AI-generated notes (identifying errors of omissions, hallucinations, and inaccuracies) [28-31], bias recognition and mitigation (eg, stereotyped language or speech recognition errors in patients with limited English proficiency) [32-39], and independent verification of the accuracy of AI-assisted notes as a core component of professional accountability. Drawing on their internal medicine residency pilot at Johns Hopkins, Abernethy et al [27] advocate that residents demonstrate foundational note-writing competence before using AI scribes and that residents receive structured AI teaching to monitor for hallucinations and recognize algorithmic bias that may be embedded in the training data, including stigmatizing language, misattribution of pronouns, and omission of social determinants

XSL•FO
RenderX

of health [27]. Educational strategies could include the discussion, evidence, feedback, and training (DEFT) AI framework, applied during clinical supervision to explore trainees' clinical reasoning and reinforce a "verify, then trust" approach that promotes vigilance against overreliance and automation bias, while applying the four aims framework.

While we are optimistic about the tangible benefits of ambient AI scribe use, we remain concerned that its implementation is outpacing robust evaluation of its impact on the cognitive development and practices of physicians, trainees, and other health professionals. Do we have the skills to discern accurate and safe use of ambient AI scribes? What are the ramifications of patients verifying the accuracy of our open documentation, especially if we are not in the habit of reviewing AI-scribed notes? Current evaluations of ambient AI scribes primarily focus on note quality, productivity, burnout, and user satisfaction, but have yet to adequately address these other questions. AI will inevitably bring new ways of learning and integrating knowledge, just as past advances have done. In the meantime, we advocate that the four main aims of the chart note, whether handwritten or digitized, serve as practical guidance and a rubric for trainee feedback regardless of what is gained and lost in this evolution of medical knowledge processing.

Ultimately, the essence of the chart note is to preserve accuracy, communicate with clarity, provide transparency, maintain patient dignity, and most importantly, exhibit that which is innately human: our personal connections to each other.

Funding

The authors declared no financial support was received for this work.

Authors' Contributions

Conceptualization: CYAC, TIL
Writing – original draft: CYAC, TIL
Writing – review & editing: CYAC, TIL, RB
Writing – content expertise: RB
CYAC and TIL are co–first authors.

Conflicts of Interest

TIL is the scientific editorial director at JMIR Publications. She had no involvement in the editorial review and processing of this manuscript. TIL is also a volunteer director on the Board of Directors, American Medical Informatics Association. CYAC is an independent contractor for Anonymous Health, a digital health company that uses artificial intelligence technology.

References

1. Blumenthal D, Tavenner M. The "meaningful use" regulation for electronic health records. N Engl J Med 2010 Aug 5;363(6):501-504. [doi: 10.1056/NEJMp1006114] [Medline: 20647183]
2. Tomines A, Readhead H, Readhead A, Teutsch S. Applications of electronic health information in public health: uses, opportunities & barriers. EGEMS (Wash DC) 2013 Oct 28;1(2):1019. [doi: 10.13063/2327-9214.1019] [Medline: 25848571]
3. Mishuris RG, Linder JA. Electronic health records and the increasing complexity of medical practice: "It Never Gets Easier, You Just Go Faster". J Gen Intern Med 2013 Apr;28(4):490-492. [doi: 10.1007/s11606-012-2304-1] [Medline: 23247583]
4. Shanafelt TD, Dyrbye LN, Sinsky C, et al. Relationship between clerical burden and characteristics of the electronic environment with physician burnout and professional satisfaction. Mayo Clin Proc 2016 Jul;91(7):836-848. [doi: 10.1016/j.mayocp.2016.05.007] [Medline: 27313121]
5. Overhage JM, McCallie D Jr. Physician time spent using the electronic health record during outpatient encounters: a descriptive study. Ann Intern Med 2020 Feb 4;172(3):169-174. [doi: 10.7326/M18-3684] [Medline: 31931523]
6. Shah M, De Arrigunaga S, Forman LS, West M, Rowe SG, Mishuris RG. Cumulated time to chart closure: a novel electronic health record-derived metric associated with clinician burnout. JAMIA Open 2024 Feb 8;7(1):ooae009. [doi: 10.1093/jamiaopen/ooae009] [Medline: 38333109]
7. Colicchio TK, Cimino JJ. Clinicians' reasoning as reflected in electronic clinical note-entry and reading/retrieval: a systematic review and qualitative synthesis. J Am Med Inform Assoc 2019 Feb 1;26(2):172-184. [doi: 10.1093/jamia/ocy155] [Medline: 30576561]
8. Wong CJ, Jackson SL, editors. The Patient-Centered Approach to Medical Note-Writing: Springer; 2024. [doi: 10.1007/978-3-031-43633-8]
9. Weed LL. Medical records that guide and teach. N Engl J Med 1968 Mar 14;278(11):593-600. [doi: 10.1056/NEJM196803142781105] [Medline: 5637758]
10. Jacobs L. Interview with Lawrence Weed, MD- the father of the problem-oriented medical record looks ahead. Perm J 2009;13(3):84-89. [doi: 10.7812/TPP/09-068] [Medline: 20740095]

XSL•FO
RenderX

11. Shoolin J, Ozeran L, Hamann C, Bria W. Association of Medical Directors of Information Systems consensus on inpatient electronic health record documentation. Appl Clin Inform 2013 Jun 26;4(2):293-303. [doi: 10.4338/ACI-2013-02-R-0012] [Medline: 23874365]
12. Gidwani R, Nguyen C, Kofoed A, et al. Impact of scribes on physician satisfaction, patient satisfaction, and charting efficiency: a randomized controlled trial. Ann Fam Med 2017 Sep;15(5):427-433. [doi: 10.1370/afm.2122] [Medline: 28893812]
13. Kanaparthy NS, Villuendas-Rey Y, Bakare T, et al. Real-world evidence synthesis of digital scribes using ambient listening and generative artificial intelligence for clinician documentation workflows: rapid review. JMIR AI 2025 Oct 10;4:e76743. [doi: 10.2196/76743] [Medline: 41071988]
14. Everson J, Nong P, Richwine C. Uptake of generative AI integrated with electronic health records in US hospitals. JAMA Netw Open 2025 Dec 1;8(12):e2549463. [doi: 10.1001/jamanetworkopen.2025.49463] [Medline: 41385223]
15. Atiku S, Olakotan O, Owolanke K. Usability-related barriers and facilitators influencing the adoption and use of AI scribes in healthcare: a scoping review. J Eval Clin Pract 2026 Feb;32(1):e70365. [doi: 10.1111/jep.70365] [Medline: 41533034]
16. Kosmyna N, Hauptmann E, Yuan YT, et al. Your brain on ChatGPT: accumulation of cognitive debt when using an AI assistant for essay writing task. arXiv. Preprint posted online on Jun 10, 2025. [doi: 10.48550/arXiv.2506.08872]
17. Abdulnour REE, Gin B, Boscardin CK. Educational strategies for clinical supervision of artificial intelligence use. N Engl J Med 2025 Aug 21;393(8):786-797. [doi: 10.1056/NEJMra2503232] [Medline: 40834302]
18. Gin BC, LaForge K, Burk-Rafel J, Boscardin CK. Macy Foundation innovation report part II: from hype to reality: innovators' visions for navigating AI integration challenges in medical education. Acad Med 2025 Sep 1;100(9S Suppl 1):S22-S29. [doi: 10.1097/ACM.0000000000006117] [Medline: 40479503]
19. Nelson A, Eliasz KL. Desirable difficulty: theory and application of intentionally challenging learning. Med Educ 2023 Feb;57(2):123-130. [doi: 10.1111/medu.14916] [Medline: 35950522]
20. Gerlich M. AI tools in society: impacts on cognitive offloading and the future of critical thinking. Societies (Basel) 2025;15(1):6. [doi: 10.3390/soc15010006]
21. Bastani H, Bastani O, Sungu A, Ge H, Kabakcı Ö, Mariman R. Generative AI without guardrails can harm learning: evidence from high school mathematics. Proc Natl Acad Sci U S A 2025 Jul;122(26):e2422633122. [doi: 10.1073/pnas.2422633122] [Medline: 40560616]
22. Izquierdo-Condoy JS, Arias-Intriago M, Tello-De-la-Torre A, Busch F, Ortiz-Prado E. Generative artificial intelligence in medical education: enhancing critical thinking or undermining cognitive autonomy? J Med Internet Res 2025 Nov 3;27:e76340. [doi: 10.2196/76340] [Medline: 41183320]
23. Russell RG, Lovett Novak L, Patel M, et al. Competencies for the use of artificial intelligence-based tools by health care professionals. Acad Med 2023 Mar 1;98(3):348-356. [doi: 10.1097/ACM.0000000000004963] [Medline: 36731054]
24. Ma Y, Song Y, Balch JA, et al. Promoting AI competencies for medical students: a scoping review on frameworks, programs, and tools. arXiv. Preprint posted online on Jul 10, 2024. [doi: 10.48550/arXiv.2407.18939]
25. Car J, Ong QC, Erlikh Fox T, et al. The digital health competencies in medical education framework: an international consensus statement based on a Delphi study. JAMA Netw Open 2025 Jan 2;8(1):e2453131. [doi: 10.1001/jamanetworkopen.2024.53131] [Medline: 39888625]
26. Gazquez-Garcia J, Sánchez-Bocanegra CL, Sevillano JL. AI in the health sector: systematic review of key skills for future health professionals. JMIR Med Educ 2025 Feb 5;11:e58161. [doi: 10.2196/58161] [Medline: 39912237]
27. Abernethy J, Shah A, Chen B, Reynolds S, Wright SM, O'Rourke P. Integrating AI scribes into medical education: guardrails for preserving clinical reasoning. J Gen Intern Med 2026 Feb 2. [doi: 10.1007/s11606-025-10149-w] [Medline: 41627656]
28. Lukac PJ, Turner W, Vangala S, et al. Ambient AI scribes in clinical practice: a randomized trial. NEJM AI 2025 Dec;2(12). [doi: 10.1056/aioa2501000] [Medline: 41497288]
29. Balloch J, Sridharan S, Oldham G, et al. Use of an ambient artificial intelligence tool to improve quality of clinical documentation. Future Healthc J 2024 Jun 26;11(3):100157. [doi: 10.1016/j.fhj.2024.100157] [Medline: 39371531]
30. Bracken A, Reilly C, Feeley A, Sheehan E, Merghani K, Feeley I. Artificial intelligence (AI) - powered documentation systems in healthcare: a systematic review. J Med Syst 2025 Feb 18;49(1):28. [doi: 10.1007/s10916-025-02157-4] [Medline: 39966286]
31. Biro J, Handley JL, Cobb NK, et al. Accuracy and safety of AI-enabled scribe technology: instrument validation study. J Med Internet Res 2025 Jan 27;27:e64993. [doi: 10.2196/64993] [Medline: 39869899]
32. Goddu AP, O'Conor KJ, Lanzkron S, et al. Do words matter? Stigmatizing language and the transmission of bias in the medical record. J Gen Intern Med 2018 May;33(5):685-691. [doi: 10.1007/s11606-017-4289-2] [Medline: 29374357]
33. Park J, Saha S, Chee B, Taylor J, Beach MC. Physician use of stigmatizing language in patient medical records. JAMA Netw Open 2021 Jul 1;4(7):e2117052. [doi: 10.1001/jamanetworkopen.2021.17052] [Medline: 34259849]
34. Himmelstein G, Bates D, Zhou L. Examination of stigmatizing language in the electronic health record. JAMA Netw Open 2022 Jan 4;5(1):e2144967. [doi: 10.1001/jamanetworkopen.2021.44967] [Medline: 35084481]
35. Alpert AB, Mehringer JE, Orta SJ, et al. Experiences of transgender people reviewing their electronic health records, a qualitative study. J Gen Intern Med 2023 Mar;38(4):970-977. [doi: 10.1007/s11606-022-07671-6] [Medline: 35641720]

XSL•FO
RenderX

36. Bilotta I, Tonidandel S, Liaw WR, et al. Examining linguistic differences in electronic health records for diverse patients with diabetes: natural language processing analysis. JMIR Med Inform 2024 May 23;12:e50428. [doi: 10.2196/50428] [Medline: 38787295]
37. Hirshman R, Hamilton S, Walker M, Ellis AR, Ivey N, Clifton D. Stigmatizing and affirming provider language in medical records on hospitalized patients with opioid use disorder. J Hosp Med 2025 Jan;20(1):26-32. [doi: 10.1002/jhm.13472] [Medline: 39080856]
38. Topaz M, Peltonen LM, Zhang Z. Beyond human ears: navigating the uncharted risks of AI scribes in clinical practice. NPJ Digit Med 2025 Sep 24;8(1):569. [doi: 10.1038/s41746-025-01895-6] [Medline: 40993221]
39. Tate S. Generative artificial intelligence tools in medicine will amplify stigmatizing language. J Addict Med 2024;18(1):90. [doi: 10.1097/ADM.0000000000001237] [Medline: 37862115]

Abbreviations

AI: artificial intelligence
APSO: assessment, plan, subjective, and objective
DEFT: discussion, evidence, feedback, and training
EHR: electronic health record
POMR: problem-oriented medical record
SOAP: subjective, objective, assessment, and plan

Edited by B Lesselroth; submitted 12.Mar.2026; this is a non–peer-reviewed article;accepted 15.May.2026; published 08.Jun.2026.

Please cite as:
Chen CYA, Leung TI, Bajra R
Re-Centering Clinical Documentation in the Age of AI Scribes: Four Aims of the Patient Chart Note
JMIR Med Educ 2026;12:e95218
URL: https://mededu.jmir.org/2026/1/e95218
doi:10.2196/95218

© Chwen-Yuen Angie Chen, Tiffany I Leung, Rika Bajra. Originally published in JMIR Medical Education (https://mededu.jmir.org), 8.Jun.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Education, is properly cited. The complete bibliographic information, a link to the original publication on https://mededu.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

AI Competency: Current State and Challenges

Sian Tsuei[1,2], MD, PhD, CCFP

[1]Department of Family Practice, University of British Columbia , David Strangway Bldg, 5950 University Blvd 3rd Floor, Vancouver, BC, Canada

[2]Department of Global Health and Population, T.H. Chan School of Public Health, Harvard University, 677 Huntington Avenue, Boston, MA, United States

Corresponding Author:
Sian Tsuei, MD, PhD, CCFP
Department of Family Practice, University of British Columbia , David Strangway Bldg, 5950 University Blvd 3rd Floor, Vancouver, BC, Canada

Abstract

As artificial intelligence (AI) develops, the medical education community has begun defining the relevant forms of competency. Many experts emphasize the importance of optimizing AI tools' output or understanding the relevant technical and normative considerations around using AI tools. A recent publication in this journal showed that optimizing instructions for large language models may yield diminishing returns as such tools improve. This suggests the need for a new competency—one that focuses on choosing the appropriate AI tools. I briefly summarize the current competency domains and examples to contextualize the current state of AI competency development, highlighting the need for further synthesis. I then introduce a hierarchical framework of competencies that might assist with priority-setting around subsequent competency development work. It consists of cognitive, operational, and meta-AI domains, which respectively correspond with the knowledge around understanding, using, and choosing AI tools. The final section describes the potential challenges associated with the development of AI competency. These include traditional concerns around competency-based medical education: deciding whether and which competencies are meaningful for measuring the targets of interest; adjusting the relevant measurements to reflect the necessary temporal and institutional context; and setting up the relevant organizational support to encourage measurement of competency. This section also discusses the challenges of developing the relevant performance indicators for AI tools across different clinical contexts. Such indicators will be necessary for guiding the choice of AI tools for the clinical context, but medical educators may not have the skills to develop them. In addition to identifying potential sources for relevant indicators, the medical education community may shape physicians' norms of practice to drive the AI industry into producing the relevant indicators. The potential for physicians to incur higher medical liability from poor choice of AI may lead them to demand more nuanced performance indicators from AI suppliers. Physicians are also in a position to do so, since the competitive AI market may provide them more bargaining power.

(JMIR Med Educ 2026;12:e86686) doi:10.2196/86686

KEYWORDS

artificial intelligence; medical education; AI

Background

Given the increasing popularity and capability of modern artificial intelligence (AI) tools [1], helping students develop AI competency is becoming increasingly crucial. Although recommendations vary, the core skills generally focus on helping trainees understand the various AI tools' technological approaches, uses, and risks, as well as the relevant ethical, social, cultural, and legal context around AI tool use [2-6].

The work from Hsieh et al [7] suggests the need for a potentially novel type of AI competency—one that focuses on the choice of AI tool. Their article showed that even though providing better instructions for large language models (LLMs) can improve their performance, the margin of improvement diminished for more advanced LLMs. This suggests that as LLMs improve, optimizing the prompt may become less important. Instead, choosing an appropriate LLM may become more critical. This implication calls into question the extent to which current medical education literature has considered the choice of LLM as a key competency.

To help medical educators and trainees critically engage with the future development of AI-related competencies, this editorial article therefore aims to (1) illustrate the state of AI competency development so that (2) it can consider worthwhile directions for this field. In what follows, I will briefly describe the current stage of AI competency development, drawing on up-to-date literature, followed by the introduction of a new taxonomy framework that can help priority-setting around future competency development work. I follow with a brief reflection on the challenges of AI competency development, first summarizing from literature the traditional concerns affecting competency development, and propose additional considerations that might uniquely affect competencies around the choice of LLM.

XSL•FO
RenderX

Domains of AI Competencies Considered

Typically, the development of medical education competencies relies on four steps: "listing competencies, devising training programmes to fit these, utilising appropriate assessment methods and determining pass levels" [8]. Much of the current discourse around AI competence has focused on the first step, seeming to largely focus on the following domains: (1) AI fundamentals, that is, understanding the mechanics underpinning AI tool performance; (2) ethical and legal considerations, understanding the ethical and legal implications of using AI tools; (3) data analysis and management, understanding how data drives AI tools' function and appropriate data handling practices; (4) evaluation of AI tools, assessing AI tools' performance to ensure that they meet necessary standards; and (5) use of AI tools, understanding whether, why, when, and how to use AI tools effectively for clinical care [3,9-13]. Currently, the American Association of Medical Colleges is synthesizing AI competencies [14], and other countries' medical associations may also wish to undertake their own adaptations to ensure local appropriateness.

Beyond competency description, some medical educators are already developing training programs and evaluation methods [15,16]. Further systematic and critical synthesis of the current AI curricular directions, evaluation methods, and impacts would be helpful for guiding subsequent efforts. Ideally, this synthesis would be able to benchmark AI competencies against established competency frameworks.

As medical educators progress along the AI-related competency-based medical education (CBME) journey, identifying which types of competencies deserve priority may be helpful. Some of the competencies may be more foundational, while others are more advanced. Table 1 describes a taxonomy that synthesizes the range of competency domains, arranging them from foundational to the most advanced.

Table . Hierarchical progression of artificial intelligence (AI) and large language model (LLM) competency.

Level of competency	Type of competency	Competency domains implicated	Goals of competency for trainees	Examples of trainee competencies
Foundational	Cognitive competency	• AI fundamentals	Understand the theoretical and operational approaches, benefits, and limitations of LLMs	Explain how deep learning works
Advanced	Operational competency	• Ethical and legal considerations • Data analysis and management • Use of AI tools	Use LLMs appropriately, recognizing ethical, social, cultural, and legal implications and limitations	Understand best practices for data storage
Most advanced	Meta-AI competency	• Evaluation of AI tools • Use of AI tools	Choose appropriate LLM for the clinical context	Understand relevant indicators of LLM performance

The first level—cognitive competency—helps individuals understand the development and operations of LLMs. Medical educators might illustrate the technical intuitions behind deep learning, modeling approaches, and data transformation. This can then help trainees understand why LLMs might suffer from technical limitations (eg, hallucinations, sycophancy bias, and algorithmic biases) [3,11-13].

The second level—operational competency—helps trainees use LLMs more effectively. Medical educators may teach trainees to identify the relevant clinical features of a case and refine the prompt and context to optimize the usefulness of an LLM's response [3,12,13]. Educators may also help trainees appreciate the relevant ethical, social, cultural, and legal requirements so that data are ethically processed.

The third level—meta-AI competency—ensures that trainees can pick the most appropriate AI models. The trainees will need to identify the clinical problem, discern the markers that show that an AI model is context-appropriate, and select a suitable LLM [12]. This requires that the trainees first understand how LLMs work and how to use them, so that they can meaningfully identify the best choice of LLM for the target scenarios. Of note, only two of the competencies appear to touch on this level of competency (Table 1's "Use of AI tools" and "Evaluation of AI tools" domains). Further development of the competencies at this level may be necessary.

Challenges Related to the Development of AI Competency

The development of AI competency may face multiple challenges. Some of them pertain to CBME in general, and some are uniquely related to aspects of AI competency. In what follows, I will first highlight the more prominent and recent critiques of the general topic of CBME, guided by the timely and thorough review of Hamza et al [17]. I connect these concerns with AI competency where possible. I will subsequently consider the unique challenges of meta-AI competency. Its development may draw in the AI development industry, transcending the usual considerations related to medical education competency development.

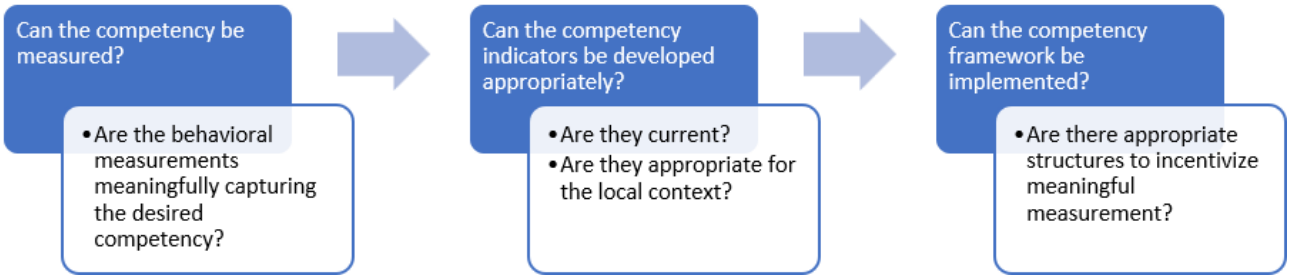
First Type of Challenge: General Critique of CBME

Figure 1 summarizes the major veins of critiques of CBME. They largely question (1) whether the relevant competency can be meaningfully measured; (2) assuming that such measurement

XSL•FO
RenderX

is possible, whether the measurement can be appropriately adapted; and (3) assuming that adapting the measurement is possible, whether the appropriate measurements can be practically implemented.

Figure 1. Conventional concerns of CBME.



First, the critics challenge the appropriateness of CBME as a paradigm. Fundamentally, CBME relies on assessing observable indicators, but sometimes the target competency cannot be measured. As a result, the assessment may target "things that may be more easily measured, instead of asking the more difficult questions" [8]. This risks distracting learners from meaningfully understanding the content or reflecting on the nuances of the situation [18], which may be necessary for doctors to reason from first principles and adapt to circumstances beyond the original medical training content. For example, if trainees fail to understand that AI models can continuously evolve based on live use data, they may fail to grasp the potential for the performance of the models to drift over time. If they continue to rely on AI tools to the same extent even as the tools' performance drops, this can undermine care quality. As the medical education community proposes and synthesizes AI competency, carefully considering which competencies are truly meaningful representations of the underlying cognitive, attitudinal, and behavioral targets will be key.

Second, critics are concerned that competency measurements may not be adequately adjusted so that the measurement actually captures the relevant competency. From a temporal standpoint, competencies need to be regularly updated to keep up with the times [19]. This is especially relevant for AI competencies. AI tools are developing rapidly, and the way they are used may drastically change in a few years. What are the indicators that should trigger the necessary updates? Should competencies simply be reviewed at a prespecified interval? These are key questions for AI-related CBME. Furthermore, from an institutional standpoint, different health care systems may have unique cultures, workflows, and resources, requiring local adaptation [20,21]. Although the competencies may be stable, the demonstration—and measurement—of those competencies may need to be adapted to the context [17,19,22]. In the context of AI, local resource constraints, population demographics, and patient preferences may further complicate these decisions, highlighting the importance for medical educators of developing local expertise regarding AI and health systems so that the measurement of the competencies can be adapted appropriately.

Third, organizational structures may not provide appropriate incentives for teachers and learners to check competence [17]. This can be problematic if modern AI accelerates the pace of service delivery and the health care system fails to concomitantly introduce appropriate incentives to encourage in-depth assessment and feedback. Health system and change management expertise are necessary complements for meaningful AI-related CBME.

Second Type of Challenge: AI Governance Challenges Undermining Meta-AI Competency Development

Aside from these conventional critiques of competency, developing meta-AI competency may face unique challenges. If physicians are to choose AI tools for specific contexts, they need to know how the AI tool performs within various clinical contexts (eg, different populations, clinical specialties, and care providers). Without such performance indicators, choosing the AI tools for the relevant patient encounters may be challenging. These competencies are therefore predicated on the presence of accurate and meaningful performance indicators of various AI tools in a range of contexts.

Meta-AI competency may be challenging for medical educators to develop. Whereas conventional medical competencies can draw directly from medical knowledge, AI performance indicators require deep AI expertise, which may be outside current medical educators' skill sets. Depending on the AI industry to provide useful performance metrics requires at least some reliance on effective governance of the AI industry.

Approaches to such governance seem to be emerging along a spectrum. The European Union (EU) represents the more stringent end: Article 10, Paragraph 3 of the EU AI Act specifies that "[AI] providers must ensure that the training, validation, and testing datasets shall be relevant, sufficiently representative, and, to the best extent possible, free of errors and complete according to the intended purpose." The implication is that if an AI tool is intended for use in dermatology, for example, but its dataset only contains images from a specific demographic or body part, a developer would need to document this limitation to comply with the data governance requirements. The "intended purpose" must clearly define the scope, which can include medical specialties and specific contexts of use.

On the other end of the spectrum, the United States seems to be much friendlier to the AI industry, focusing on the economic opportunities associated with AI proliferation and providing minimal effective guardrails around AI use. In fact, even though the Food and Drug Administration (FDA) is empowered to regulate AI tools, regulation is severely lacking, and entire columns are sometimes missing in the FDA database [23]. AI

XSL•FO
RenderX

tools may therefore have little incentive to provide meaningful evidence of their performance in different contexts, which would undermine the development of relevant meta-AI competency.

Implicitly, each approach reflects the value society places on stringent regulation of an emerging technology and on safeguarding patients' health. For medical educators, however, what might matter more is considering how teachers and learners ought to adapt when the relevant governance approaches fail to drive the disclosure of necessary AI performance indicators.

Educators may highlight to students the possibility of identifying the necessary information elsewhere. Performance indicators from the EU AI Act's database might be transferable to the population of interest, because the AI supplier validated the AI tool using a population of similar demographic composition. Such adaptation should note, however, that different health care systems deviate from each other in various ways [24], so careful consideration of whether to use the tool and how to do so appropriately will be key.

Even more critically, the medical education community can use their educational power to drive market change. Rather than awaiting regulatory saviors, educators can step in to drive the trajectory of competency assessment. Medical educators have the power to institutionalize behavioral norms among medical trainees. Medical training is rigorous and formalized so that the trainees graduate as a group of individuals who are homogeneous in skills, aptitudes, and values [25]. If medical trainees learn to consistently review relevant performance indicators before using AI tools, this behavioral norm can be a meaningful bargaining force in the AI industry, potentially pushing AI suppliers to change the relevant information for their AI products. AI suppliers may realize that they can capture more physicians as customers if they provide more indicators of their tools' performance. Essentially, medical educators' power to influence curricula can trickle down and affect doctors' behavioral norms, which can then change industry norms.

However, critics may argue that past efforts from medical education failed to influence industry standards. A good example is medical education's influence on the medical documentation process and electronic medical record (EMR) industry. First, despite medical educators' efforts to teach proper charting, medical doctors' charting standards are still problematic [26]. This suggests that practicing doctors may shed their training if it goes against their style of practice. If so, doctors may never end up pushing the industry to provide relevant metrics, because they might not care about academic lessons during training. The attending physician might reach for the LLM most commonly used without vetting its metrics for the specific context.

Second, even if physicians' practices change and they push against the industry, there is no promise that the vendors will make the needed changes. For example, physicians' concerns over many EMR platforms' poor usability have been well documented since 2004 [27], but complaints about this have persisted into 2025 [28,29].

The unique context of the current medical AI industry may refute these concerns. First, the potential for legal risks related to using medical AI tools for clinical services may drive physicians to more closely adhere to training recommendations in carefully vetting AI tools. Administrative capabilities aside, AI tools' capacity to support medical diagnoses can seriously alter the course of treatment for patients. Improper vetting of AI tools can significantly increase physicians' liability [30], suggesting that physicians may have much interest in adhering to training recommendations for AI tool vetting.

Second, the current AI market structure appears to provide buyers with more power relative to suppliers, suggesting that AI vendors may wish to cater to doctors' concerns. Foundational LLMs are becoming substitutable commodities [31], as many LLMs provide similar levels of performance. Furthermore, these foundational models help lower the entry cost for secondary AI tool developers to release similar AI tools that compete intensely for consumers. The boom in AI scribes is one such example, and aggressive attempts to improve physicians' user experiences by offering ways to integrate their products into different EMR platforms is an unsurprising consequence.

Conclusion

Medical education experts have proposed several domains and competencies related to use of AI, and further synthesis of such work can be beneficial. I proposed a hierarchical framework of AI competency that can help medical educators discern the order of priority for competency development. The meta-AI competency, which is at the top of the hierarchy, may be particularly challenging to measure and implement. In addition to conventional challenges related to CBME, medical educators require effective AI governance to identify AI performance indicators that would most support trainees in selecting appropriate AI tools for the clinical context. Given the current market structure, the community of medical educators may have an opportunity to leverage their power in influencing the norms of practice to drive further AI performance indicators.

Acknowledgments

I am tremendously grateful for the input from Dr Blake Lesselroth and Dr Tiffany Leung. Their careful review and thoughtful comments prompted deeper thinking and improved the article significantly.

Funding

This work received no funding.

Data Availability

This work used no data, and no data will be made available.

XSL•FO
RenderX

Authors' Contributions

ST was the sole author and is responsible for all facets of this work.

Conflicts of Interest

ST serves as a member on the Artificial Intelligence Advisory Group for the College of Family Physicians of Canada. ST serves as an associate editor for *JMIR Medical Education*.

References

1. Preiksaitis C, Rose C. Opportunities, challenges, and future directions of generative artificial intelligence in medical education: scoping review. JMIR Med Educ 2023 Oct 20;9(1):e48785. [doi: 10.2196/48785] [Medline: 37862079]
2. Cai C, Duell J, Minghui Chen D, Kin Ho W, Kwong Lee BT, Li F, et al. Advancing AI literacy in medical education: a medical AI competency framework development. In: Cristea AI, Walker E, Lu Y, Santos OC, Isotani S, editors. Artificial Intelligence in Education: Springer Nature; 2025:116-123. [doi: 10.1007/978-3-031-98462-4_15]
3. Çalışkan SA, Demir K, Karaca O. Artificial intelligence in medical education curriculum: an e-Delphi study for competencies. PLOS ONE 2022;17(7):e0271872. [doi: 10.1371/journal.pone.0271872] [Medline: 35862401]
4. Chiu TKF, Ahmad Z, Ismailov M, Sanusi IT. What are artificial intelligence literacy and competency? A comprehensive framework to support them. Comput Educ Open 2024 Jun;6:100171. [doi: 10.1016/j.caeo.2024.100171]
5. Long D, Magerko B. What is AI literacy? competencies and design considerations. In: Bernhaupt R, Mueller FF, editors. CHI '20: CHI Conference on Human Factors in Computing Systems: Association for Computing Machinery; 2020:1-16. [doi: 10.1145/3313831.3376727]
6. Ng DTK, Leung JKL, Chu SKW, Qiao MS. Conceptualizing AI literacy: an exploratory review. Comput Educ Artif Intell 2021;2:100041. [doi: 10.1016/j.caeai.2021.100041]
7. Hsieh MY, Wang TL, Su PH, Chou MC. Impact of prompt engineering on the performance of ChatGPT variants across different question types in medical student examinations: cross-sectional study. JMIR Med Educ 2025 Oct 1;11:e78320. [doi: 10.2196/78320] [Medline: 41032724]
8. Talbot M. Monkey see, monkey do: a critique of the competency model in graduate medical education. Med Educ 2004 Jun;38(6):587-592. [doi: 10.1046/j.1365-2923.2004.01794.x] [Medline: 15189254]
9. Scott IA, Shaw T, Slade C, et al. Proposing core competencies for physicians in using artificial intelligence tools in clinical practice. Intern Med J 2025 Aug;55(8):1403-1409. [doi: 10.1111/imj.70112] [Medline: 40576330]
10. Car J, Ong QC, Erlikh Fox T, et al. The digital health competencies in medical education framework: an international consensus statement based on a Delphi study. JAMA Netw Open 2025 Jan 2;8(1):e2453131. [doi: 10.1001/jamanetworkopen.2024.53131] [Medline: 39888625]
11. Russell RG, Lovett Novak L, Patel M, et al. Competencies for the use of artificial intelligence-based tools by health care professionals. Acad Med 2023 Mar 1;98(3):348-356. [doi: 10.1097/ACM.0000000000004963] [Medline: 36731054]
12. Gazquez-Garcia J, Sánchez-Bocanegra CL, Sevillano JL. AI in the health sector: systematic review of key skills for future health professionals. JMIR Med Educ 2025 Feb 5;11(1):e58161. [doi: 10.2196/58161] [Medline: 39912237]
13. Moldt JA, Festl-Wietek T, Fuhl W, et al. Assessing AI awareness and identifying essential competencies: insights from key stakeholders in integrating AI into medical education. JMIR Med Educ 2024 Jun 12;10(1):e58355. [doi: 10.2196/58355] [Medline: 38989834]
14. CGEA Artificial Intelligence Framework for Educators. Association of American Medical Colleges. 2025. URL: https://www.aamc.org/about-us/mission-areas/medical-education/advancing-ai-resource-collection/artificial-intelligence-competencies-medical-educators [accessed 2025-12-12]
15. Lee J, Wu AS, Li D, Kulasegaram KM. Artificial intelligence in undergraduate medical education: a scoping review. Acad Med 2021 Nov 1;96(11S):S62-S70. [doi: 10.1097/ACM.0000000000004291] [Medline: 34348374]
16. Kim DH, Kang YJ, Lee YM. Twelve tips for developing and implementing AI curriculum for undergraduate medical education. Med Educ Online 2025 Dec 31;30(1):2585637. [doi: 10.1080/10872981.2025.2585637] [Medline: 41327942]
17. Hamza DM, Hauer KE, Oswald A, et al. Making sense of competency-based medical education (CBME) literary conversations: a BEME scoping review: BEME Guide No. 78. Med Teach 2023 Aug;45(8):802-815. [doi: 10.1080/0142159X.2023.2168525] [Medline: 36668992]
18. Leung WC. Competency based medical training: review. BMJ 2002 Sep 28;325(7366):693-696. [Medline: 12351364]
19. Nazeha N, Pavagadhi D, Kyaw BM, Car J, Jimenez G, Tudor Car L. A digitally competent health workforce: scoping review of educational frameworks. J Med Internet Res 2020 Nov 5;22(11):e22706. [doi: 10.2196/22706] [Medline: 33151152]
20. Chou FC, Hsiao CT, Yang CW, Frank JR. "Glocalization" in medical education: a framework underlying implementing CBME in a local context. J Formos Med Assoc 2022 Aug;121(8):1523-1531. [doi: 10.1016/j.jfma.2021.10.024] [Medline: 34782195]
21. Ott MC, Dengler L, Hibbert K, Ott M. Fixing disconnects: exploring the emergence of principled adaptations in a competency-based curriculum. Med Educ 2025 Apr;59(4):428-438. [doi: 10.1111/medu.15475] [Medline: 39105665]

XSL•FO
RenderX

22. Salem MA, Zakaria OM, Aldoughan EA, Khalil ZA, Zakaria HM. Bridging the AI gap in medical education: a study of competency, readiness, and ethical perspectives in developing nations. Computers 2025 Jun;14(6):238. [doi: 10.3390/computers14060238]
23. Babic B, Glenn Cohen I, Stern AD, Li Y, Ouellet M. A general framework for governing marketed AI/ML medical devices. NPJ Digit Med 2025 May 31;8(1):328. [doi: 10.1038/s41746-025-01717-9] [Medline: 40450160]
24. Sahiner B, Chen W, Samala RK, Petrick N. Data drift in medical machine learning: implications and potential remedies. Br J Radiol 2023 Oct;96(1150):20220878. [doi: 10.1259/bjr.20220878] [Medline: 36971405]
25. DiMaggio PJ, Powell WW. The iron cage revisited: institutional isomorphism and collective rationality in organizational fields. Am Sociol Rev 1983 Apr;48(2):147. [doi: 10.2307/2095101]
26. Shahbodaghi A, Moghaddasi H, Asadi F, Hosseini A. Documentation errors and deficiencies in medical records: a systematic review. J Health Manag 2024 Apr;26(2):351-368. [doi: 10.1177/09720634241229545]
27. Miller RH, Sim I. Physicians' use of electronic medical records: barriers and solutions. Health Aff (Millwood) 2004;23(2):116-126. [doi: 10.1377/hlthaff.23.2.116] [Medline: 15046136]
28. Melnick ER, Dyrbye LN, Sinsky CA, et al. The association between perceived electronic health record usability and professional burnout among US physicians. Mayo Clin Proc 2020 Mar;95(3):476-487. [doi: 10.1016/j.mayocp.2019.09.024] [Medline: 31735343]
29. Schwappach D, Hautz W, Krummrey G, Pfeiffer Y, Ratwani RM. EMR usability and patient safety: a national survey of physicians. NPJ Digit Med 2025 May 15;8(1):282. [doi: 10.1038/s41746-025-01657-4] [Medline: 40374898]
30. Cohen IG. Informed consent and medical artificial intelligence: what to tell the patient? Georgetown Law J 2020;108:1425-1469 [FREE Full text]
31. McKendrick J. Forbes. As AI rapidly becomes a commodity, time to consider the next step. 2024. URL: https://www.forbes.com/sites/joemckendrick/2024/02/07/as-ai-rapidly-becomes-a-commodity-time-to-consider-the-next-step/ [accessed 2026-02-20]

Abbreviations

AI: artificial intelligence
CBME: competency-based medical education
EMR: electronic medical record
EU: European Union
FDA: Food and Drug Administration
LLM: large language model

Edited by B Lesselroth, T Leung; submitted 28.Oct.2025; this is a non–peer-reviewed article;accepted 10.Feb.2026; published 03.Mar.2026.

Please cite as:
Tsuei S
AI Competency: Current State and Challenges
JMIR Med Educ 2026;12:e86686
URL: https://mededu.jmir.org/2026/1/e86686
doi:10.2196/86686

© Sian Tsuei. Originally published in JMIR Medical Education (https://mededu.jmir.org), 3.Mar.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Education, is properly cited. The complete bibliographic information, a link to the original publication on https://mededu.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Research Letter

Evaluating Microlearning for Faculty Development in Medical Education: Mixed Methods Pilot Study

Darci L Lammers[1], PhD; Jeffrey B Geske[2], MD; Jane A Linderbaum[2], APRN, CNP; Michael W Cullen[2], MD

[1]Cardiovascular Education, Mayo Clinic, Rochester, MN, United States

[2]Cardiovascular Diseases, Mayo Clinic, Rochester, MN, United States

Corresponding Author:
Darci L Lammers, PhD
Cardiovascular Education
Mayo Clinic
200 1st St SW
Rochester, MN, 55905
United States
Phone: 1 507 284 2511
Email: lammers.darci@mayo.edu

Abstract

This mixed methods pilot study evaluates the feasibility and effectiveness of microlearning for faculty development in cardiovascular education. Microlearning appears feasible and well-received for faculty development, offering a scalable, flexible approach.

(JMIR Med Educ 2026;12:e87980) doi:10.2196/87980

KEYWORDS

faculty development; medical education; continuing medical education; knowledge transfer; behavioral change; instructional design; digital learning; scalability; sustainability; cardiovascular diseases; mixed methods; knowledge retention; asynchronous learning; professional development

Introduction

Traditional faculty development relies on time-intensive, in-person sessions, limiting participation for clinician-educators balancing clinical and administrative duties [1,2].

Microlearning delivers brief, focused segments (≤15 min) aligned with objectives [3,4]. Grounded in cognitive science, it reduces cognitive load and enhances retention by presenting information in small units [5]. For busy clinician-educators, microlearning offers flexible, asynchronous learning that addresses time constraints and supports targeted, on-demand modules for skill application [3,4,6], and it improves engagement and aligns with digital learning preferences [6,7].

However, research has focused only on the short-term outcomes of microlearning, leaving gaps in sustained knowledge transfer and behavioral change [3,4,6]. Without accessible and time-efficient faculty development approaches, clinician-educators may continue to rely on informal or inconsistent training, potentially compromising the quality of educational assessments, learner outcomes, and the validity of continuing medical education (CME) activities.

We selected CME multiple-choice question (MCQ) development as our intervention topic based on an internal needs analysis indicating that faculty responsible for creating CME assessment items receive little or no structured guidance in item writing. This task was identified locally as a high-frequency responsibility in which faculty desired more support. MCQ development serves as a practical context to evaluate microlearning for clinician-educators.

This mixed methods pilot evaluates microlearning feasibility and effectiveness in cardiovascular faculty development by assessing learning transfer, satisfaction, and 4-month knowledge application.

Methods

We conducted a sequential explanatory study guided by Kirkpatrick's framework [8] to assess a microlearning module for faculty development in cardiovascular education. The Mayo Clinic Institutional Review Board reviewed the study and determined it to be exempt. Participants provided informed consent and could withdraw any time. Those completing all components received US $200 remuneration. All data were deidentified. Quantitative analysis was prioritized, with

XSL•FO
RenderX

qualitative interviews used to explain and contextualize test results.

Cardiology clinician-educators responsible for CME MCQs were recruited via email. Eligibility required current responsibility for board-style MCQ development and no formal training or related coursework within 6 months. Of the 75 identified faculty, 34 met the criteria; 8 enrolled and completed all the components.

The pretest (18 items across 4 sections; total 400 points, scored in the learning management system) preceded each corresponding microlearning segment (Multimedia Appendix 1). The module comprised 4 short videos, a quick-reference guide, and an MCQ-writing template, delivered asynchronously via the cardiology CME learning management system for 3 months. After a 15-item satisfaction survey (Multimedia Appendix 2), access was discontinued. An identical posttest occurred 4 months after completion to assess retention; 2 weeks later, semistructured Teams interviews explored application and perceptions (Multimedia Appendix 3).

We compared pre/post scores using the 2-sided Wilcoxon signed-rank test (reporting Pratt method including ties), calculated matched-pairs rank-biserial effect sizes, and summarized distributions with medians (IQRs).

Survey responses were summarized descriptively and used to inform preliminary qualitative themes. We conducted a structured hybrid deductive–inductive analysis. The lead investigator completed line-by-line coding by using an a priori codebook derived from the interview guide, microlearning constructs, and Kirkpatrick's model, with inductive codes added as needed. An education researcher not involved in the study independently reviewed all the coded transcripts. Discrepancies were resolved through consensus, and the codebook was refined iteratively.

Quantitative and qualitative data were integrated by the lead and co-lead investigators, who jointly analyzed test scores, survey responses, and interview transcripts. Quantitative results informed qualitative coding, enabling exploration of trends and outliers. Both strands were interpreted together to provide a comprehensive understanding of the module's impact.

Results

Among the 8 completers, the median pretest score was 366.07 (IQR 338.93-389.28), and the median posttest score was 400.00 (IQR 361.96-400.00). Of the 8 paired scores, 5 improved, 2 were unchanged (ties), and 1 decreased; thus, informative pairs were n=6.

A Wilcoxon signed-rank test (excluding ties) yielded W=5.0, *P*=.25; using the Pratt method (including ties), W=7.0, *P*=.18. The matched-pairs rank-biserial correlation was 0.52, indicating a moderate positive effect of the intervention.

Satisfaction survey responses showed strong agreement across all dimensions, with participants endorsing the module's relevance, clarity, and flexibility (Table 1).

Table 1. Quantitative summary of the satisfaction survey responses (N=8).

Survey item	Strongly agree, n (%)	Agree, n (%)
Videos were engaging	6 (75)	2 (25)
Video length was appropriate	8 (100)	0 (0)
Modular format was effective	7 (88)	1 (12)
Prepared to write board-style review questions	6 (75)	2 (25)
Quick reference guides were valuable	8 (100)	0 (0)
Pretest helped gauge prior knowledge	5 (63)	3 (37)
Pretest guided learning	5 (63)	3 (37)
Posttest reinforced learning	5 (63)	3 (37)

Qualitative analysis revealed three main themes: (1) appreciation for the concise, flexible format; (2) direct application of learned principles in educational practice; and (3) high perceived value and satisfaction (Table 2). Time constraints remained a barrier to engaging in faculty development, but microlearning was widely endorsed as effective and scalable.

XSL•FO
RenderX

Table 2. Key themes from qualitative interviews.

Theme	Subthemes	Representative findings
Value of microlearning format	• Flexibility and time efficiency • Reduced cognitive load • Point-of-need learning	Microlearning was praised for fitting into busy schedules and enabling learning in short, focused bursts
Knowledge application and transfer	• Direct application to CME[a] • Immediate use of quick-reference codes and template • Knowledge shared with colleagues	Learned principles were immediately used in CME MCQ[b] development and shared with colleagues. Modules will serve as an ongoing reference
Perceived value and satisfaction	• High ratings (excellent/very good) • Clear, concise, relevant content • Multimedia quality	Participants appreciated the clarity, relevance, and production quality of the modules and resources
Barriers to faculty development	• Time constraints • Administrative duties • Lack of protected time from clinical responsibilities	Common barriers included competing clinical/administrative demands and lack of dedicated time for learning
Suggestions for improvement	• Familiar platforms • Time estimates for modules • Follow-up opportunities • Resource access	Recommendations included hosting on familiar organizational platforms, adding time estimates, and offering structured follow-up with an expert

[a]CME: continuing medical education.

[b]MCQ: multiple choice question.

Discussion

We demonstrated that a microlearning intervention for cardiovascular faculty development was feasible, well-received, and associated with improved knowledge scores and perceived application of skills at 4 months, aligning with the objectives to assess learning transfer, satisfaction, and sustained knowledge application.

Findings align with prior evidence that microlearning enhances engagement, skill acquisition, and retention in health education [1,3]. The module's brevity, relevance, and asynchronous access likely contributed to its effectiveness. The 4-month follow-up addresses a gap in prior faculty-focused studies, which often lack longer-term data [2]. Microlearning can improve learning outcomes and self-efficacy, further supporting its value in CME [9].

By integrating quantitative data from pretests and posttests and satisfaction surveys with qualitative insights from interviews, investigators contextualized statistical findings with participant perspectives. Quantitative results informed preliminary qualitative coding, enabling exploration of trends and outliers. This approach provided a more comprehensive understanding of the module's impact and feasibility, revealing factors and barriers not apparent from a single data type.

Mixed methods revealed both measurable gains and contextual insights, including barriers such as time constraints. Participant recommendations such as hosting modules on familiar platforms and providing time estimates for each section may further improve engagement.

Limitations include the single-institution setting, small sample size, exclusion of tied pairs in the primary analysis, lack of a control group, and potential response bias. The 4-month follow-up may not capture durable behavior change. Integrating quantitative and qualitative methods enhanced understanding but may introduce interpretive complexity.

Overall, microlearning appears to be a scalable, flexible approach to faculty development, well-suited to clinical educators. Institutions should consider implementing microlearning modules with structured follow-up to reinforce learning. Future research should use larger, more diverse samples, include control groups, and extend follow-up to evaluate long-term behavioral and organizational outcomes.

Acknowledgments

The authors would like to thank Metta A Kuehntopp, MEd, and Jeffrey C Williams for their part in the creation of the intervention used in this study. We thank Patricia K Guthrie for her valuable contributions to developing content within our learning management system. No generative artificial intelligence was used in the preparation of this manuscript.

XSL•FO
RenderX

Data Availability

The datasets generated and analyzed during this study are not publicly available because participant consent for data sharing was not obtained. Data may be available from the corresponding author upon reasonable request and with appropriate institutional approvals, provided such sharing complies with participant privacy and ethical guidelines.

Funding

This research was funded by the Mayo Clinic College of Medicine and Science Office of Applied Scholarship and Education Science Endowment for Education Research Award. The study was funded for US $12,000 for a one-year period starting January 2024.

Authors' Contributions

Conceptualization: DLL (lead), MWC (supporting)
Data curation: DLL
Formal analysis: DLL
Funding acquisition: DLL
Investigation: DLL
Methodology: DLL (lead), MWC (supporting)
Project administration: DLL (lead), JBG (supporting), MWC (supporting)
Resources: DLL (lead), JBG (supporting)
Supervision: MWC
Validation: DLL
Visualization: DLL (lead), MWC (supporting)
Writing – original draft: DLL (lead), MWC (supporting)
Writing – review & editing: DLL (lead), MWC (supporting), JBG (supporting), JAL (supporting)

Conflicts of Interest

None declared.

Multimedia Appendix 1
Microlearning pretest and posttest.
[DOCX File , 499 KB - mededu_v12i1e87980_app1.docx]

Multimedia Appendix 2
Post course evaluation.
[DOCX File , 31 KB - mededu_v12i1e87980_app2.docx]

Multimedia Appendix 3
Follow-up interview guide.
[DOCX File , 29 KB - mededu_v12i1e87980_app3.docx]

References

1. Dyrbye L, Bergene A, Leep HA, Billings H. Reimagining faculty development deployment: a multipronged, pragmatic approach to improve engagement. Acad Med 2022 Sep;97(9):1322-1330. [doi: 10.1097/acm.0000000000004688]
2. Cook DA, Steinert Y. Online learning for faculty development: a review of the literature. Med Teach 2013 Nov;35(11):930-937. [doi: 10.3109/0142159X.2013.827328] [Medline: 24006931]
3. De Gagne JC, Park HK, Hall K, Woodward A, Yamane S, Kim SS. Microlearning in health professions education: scoping review. JMIR Med Educ 2019 Jul 23;5(2):e13997 [FREE Full text] [doi: 10.2196/13997] [Medline: 31339105]
4. Bowler C, Foshee C, Haggar F, Simpson D, Schroedl C, Billings H. Got 15? try faculty development on the Fly: a snippets workshop for microlearning. MedEdPORTAL 2021 Jun 14;17:11161 [FREE Full text] [doi: 10.15766/mep_2374-8265.11161] [Medline: 34189261]
5. Taylor A, Hung W. The effects of microlearning: a scoping review. Education Tech Research Dev 2022 Jan 26;70(2):363-395. [doi: 10.1007/s11423-022-10084-1]
6. Monib WK, Qazi A, Apong RA. Microlearning beyond boundaries: a systematic review and a novel framework for improving learning outcomes. Heliyon 2025 Jan 30;11(2):e41413 [FREE Full text] [doi: 10.1016/j.heliyon.2024.e41413] [Medline: 39882484]
7. Moore R, Hwang W. A systematic review of mobile-based microlearning in adult learner contexts. Educ Technol Soc 2024:27-146. [doi: 10.30191/ETS.202401_27(1).SP02]

XSL•FO
RenderX

8. Kirkpatrick J, Kirkpatrick W. Kirkpatrick's Four Levels of Training Evaluation. Alexandria, VA: Association for Talent Development; 2016.
9. Zarshenas L, Mehrabi M, Karamdar L, Keshavarzi MH, Keshtkaran Z. The effect of micro-learning on learning and self-efficacy of nursing students: an interventional study. BMC Med Educ 2022 Sep 07;22(1):664 [FREE Full text] [doi: 10.1186/s12909-022-03726-8] [Medline: 36071456]

Abbreviations

CME: continuing medical education
MCQ: multiple choice question

Edited by A Stone; submitted 17.Nov.2025; peer-reviewed by BS Chisholm, S Otero; comments to author 15.Dec.2025; revised version received 18.Feb.2026; accepted 19.Feb.2026; published 11.Mar.2026.

Please cite as:
Lammers DL, Geske JB, Linderbaum JA, Cullen MW
Evaluating Microlearning for Faculty Development in Medical Education: Mixed Methods Pilot Study
JMIR Med Educ 2026;12:e87980
URL: https://mededu.jmir.org/2026/1/e87980
doi:10.2196/87980
PMID:41711588

©Darci L Lammers, Jeffrey B Geske, Jane A Linderbaum, Michael W Cullen. Originally published in JMIR Medical Education (https://mededu.jmir.org), 11.Mar.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Education, is properly cited. The complete bibliographic information, a link to the original publication on https://mededu.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Interprofessional Training in Virtual Reality for Health Care: Experimental Study on Procedural Knowledge and Willingness to Collaborate

Melanie Bauer, BSc; Matthias J Witti, Dr rer biol hum; Matthias Stadler, Prof Dr; Anna Weiß; Martin R Fischer, Prof Dr Med; Constanze Richters, PhD

Institute of Medical Education, LMU University Hospital, LMU Munich, Pettenkoferstr. 8a, Munich, Germany

Corresponding Author:
Constanze Richters, PhD
Institute of Medical Education, LMU University Hospital, LMU Munich, Pettenkoferstr. 8a, Munich, Germany

Abstract

Background: High-quality wound care requires early and effective interprofessional collaboration between medical, nursing, and pharmacy professionals. However, interprofessional education (IPE) in this context remains limited in higher education. Immersive virtual reality (iVR) seems to be a promising IPE tool, enabling a standardized, realistic, and safe learning environment that allows multiple learners from different professions to train together. However, its educational effectiveness likely depends on instructional design that supports learning while managing cognitive demands.

Objective: This study examined whether a newly developed interprofessional iVR wound-care training improves (1) procedural knowledge and (2) willingness to collaborate among medical, nursing, and pharmacy students, and how cognitive load relates to these outcomes.

Methods: A within-subjects design with a pre- and posttest was implemented with 116 students from medicine, nursing, and pharmacy. Students completed 2 iVR sessions (≈25 and 15 min) in interprofessional triads, addressing a pressure ulcer case. The training integrated step-by-step scaffolding for the wound care task and collaboration scripts to guide teamwork. Procedural knowledge and willingness to collaborate were assessed before and after the sessions, and cognitive load was measured after the sessions. Data were analyzed using repeated-measures analysis of covariances and a mediation model to test the preregistered effects.

Results: Procedural knowledge increased significantly from pre- to posttest ($F_{1, 107}$=26.19, P<.001, η^2=.08). Cognitive load showed no significant effect on this gain. Willingness to collaborate did not change after the first session ($F_{1, 80}$=3.55, P=.063, η^2=.01), and was unaffected by cognitive load. Exploratory analyses showed that willingness to collaborate was significantly higher after the second session (t_{64}=3.16, P=.007, mean difference=0.202). Effects on procedural knowledge and willingness to collaborate did not depend on the learner's profession.

Conclusions: These findings suggest that the iVR training effectively supported learning by providing a clear structure and managing cognitive demands, enabling students from different professions to acquire procedural knowledge. The absence of cognitive load effects may suggest that the instructional design helped balance task complexity and guidance. The delayed increase in collaboration willingness further suggests that attitudinal change requires sustained, repeated engagement in interprofessional contexts rather than a single exposure. Notably, no profession-related differences emerged in either procedural knowledge or willingness to collaborate, indicating that the iVR training supported learners equally across professional backgrounds. This study highlights the potential of iVR as a scalable, theory-based approach to IPE that can support interprofessional learning and provide a structured environment for collaborative skill development. Future research should examine sustained effects and comparative effectiveness when iVR is implemented in routine curricular IPE settings beyond controlled study conditions.

(JMIR Med Educ 2026;12:e85139) doi:10.2196/85139

KEYWORDS

virtual reality; immersive; interprofessional education; wound care; cognitive load; procedural knowledge; interprofessional collaboration; medical education; nursing education; pharmacy education

XSL•FO
RenderX

Introduction

Background

Complex care situations involving emergencies or chronic illnesses with functional impairments present a significant challenge to health care systems. Chronic wounds, for example, are a growing global health problem for countries with an aging population that presents challenges not only for affected patients but for the health care system [1-3]. Providing high-quality wound care requires early and well-coordinated collaboration between clinicians from different professions, such as medicine, nursing, and pharmacy [4]. This is referred to as interprofessional collaboration practice [5].

To prepare future health care professionals for these demands, interprofessional education (IPE) aims to foster shared understanding, role clarity, and joint decision-making among learners from different disciplines [5,6]. Despite its importance, IPE remains challenging to implement in practice. Differences in curricula, institutional separation, and mismatched schedules or experience levels frequently stand in the way of meaningful interprofessional learning opportunities. As a result, the development of interprofessional competencies is frequently inconsistent across curricula [7,8].

Thus, there is a need for innovative interprofessional training opportunities. Immersive virtual reality (iVR) could be a promising approach to enable such an innovative interprofessional training opportunity [9-12]. First, iVR provides a standardized, scalable learning environment [13]. This ensures comparability and consistency across training contexts. By simulating realistic health care scenarios and team interactions [14], iVR eliminates the variability of human instructors, such as fatigue or bias, from the training process. iVR enables repeated practice in a safe environment, protecting patient safety and reducing reliance on the limited availability of real patients [10]. Second, in the context of IPE, iVR eliminates the need for physical co-location of different professional groups [13]. It allows multiple professional perspectives to be integrated within a shared virtual space, creating innovative opportunities for interprofessional learning [15].

Immersion refers to the user's sense of being present in that environment [16]. The sensory richness and complexity of iVR can impose a cognitive load. Cognitive load is defined as the mental effort required to process and act on information [17]. Thoughtful instructional design can help manage these demands and support learning.

Prior studies suggest that iVR-based training can enhance procedural knowledge [18], foster interprofessional understanding (eg, by strengthening awareness of professional roles and perspectives), and increase learners' willingness to collaborate [10,19]. However, current IPE approaches still rarely integrate immersive technologies in a structured and theory-based way. Recent reviews call for more rigorous evaluations of digital IPE formats using validated outcomes and theoretically grounded instructional frameworks [20].

Study Aim: Interprofessional iVR Training

To address this gap, this study investigates the effects of a newly developed interprofessional iVR training focusing on the topic of wound care. Chronic wound care offers a suitable context as it is not yet sufficiently anchored in the medical curriculum [21]. The training engages students from medicine, nursing, and pharmacy, each bringing distinct but complementary knowledge to the collaborative clinical scenario. The theory-based instructional design [22,23] helps learners complete a complex procedure while coordinating effectively across professional roles. We investigate whether this iVR training improves learners' procedural knowledge (RQ1) and willingness to collaborate (RQ2), and how these outcomes are influenced by learners' cognitive load experience.

Cognitive Load and Instructional Design in iVR

Learning outcomes are shaped by the amount and type of cognitive load learners experience, that is, the amount of mental effort required to process information and perform a task according to cognitive load theory [17]. A key distinction is made between extraneous cognitive load (ECL), which results from suboptimal or distracting details and negatively affects learning outcomes. For example, instructions for the wound care process may be displayed in varying positions within the virtual room. Intrinsic cognitive load (ICL) depends on the complexity of the learning task relative to the learner's prior knowledge and is most beneficial when it reaches a moderate level of challenge. If a task is too simple, learners may disengage, and if it is too difficult, they may become overwhelmed. This suggests an inverted U-shaped relationship between ICL and performance (Seufert [24] explains the effect between ICL and self-regulation). Consequently, cognitive load is an important factor to consider when investigating learning outcomes, particularly in iVR. ECL in iVR can be reduced by adapting the instructional design using computer-based scaffolding [25]. Scaffolding provides temporary support to help learners accomplish tasks beyond their current abilities [25], for example, by breaking down authentic tasks into smaller, clearly identifiable steps, such as a step-by-step guide to wound care, and providing structured instructions. In addition to managing CL, instructional design can also support and structure collaboration through collaboration scripts [22,23]. Such scripts guide specific collaborative activities, for example, discussing wound dressing, and prompt learners to build on each other's contributions to deepen elaboration. An instructional design for iVR that combines computer-based scaffolding in the form of a step-by-step guide breaking down the wound care process, alongside collaboration scripts, appears particularly promising for improving learning for adequate interprofessional wound care.

Research Questions and Hypotheses

With respect to RQ1, we hypothesize that students from medicine, nursing, and pharmacy will gain procedural knowledge from the training (H1.1), but that this gain will be negatively moderated by ECL (H1.2) and negatively mediated by ICL, following an inverse U-shaped pattern (H1.3).

XSL•FO
RenderX

In addition, learners may differ in how they benefit from the training due to the different health professions they come from [26]. Prior work suggests that students in medicine, nursing, and pharmacy vary in their clinical experience, communication practices, and expectations regarding interprofessional collaboration [26-28]. These differences may influence how learners engage with the iVR scenario and what they learn from it. Thus, we explore whether the expected knowledge gains differ by learners' profession (H1.4; exploratory).

With respect to RQ2, we expect students from medicine, nursing, and pharmacy to report increased willingness to collaborate after the training (H2.1), with this increase being negatively moderated by ECL (H2.2). Since learners' professional backgrounds also shape their collaborative role expectations and prior exposure to teamwork, we explore whether the training affects collaboration willingness differently across professions (H2.3; exploratory). H1.1 to H1.3 and H2.1 to H2.2 were preregistered prior to data collection and are available via Open Science Framework (Multimedia Appendix 1). Findings from this study may inform the development of scalable, theory-based instructional strategies for integrating iVR into IPE.

Methods

Sample and Design

The study uses data from an iVR-based training based on a within-subjects design with a pre- and posttest comparison. An a priori power analysis for the preregistered hypotheses, assuming a medium effect size (*f*) of 0.15, with α=.05, and 1-β=.80, yielded a minimum required sample size of n=103. Accordingly, a total of 122 students participated, of whom 6 were excluded due to technical issues with the iVR headsets at the outset, resulting in a final sample of n=116 (N_{female}=96, $N_{diverse}$=1; mean age 24.8, SD 4.9 years) of whom 63 were medical students, 31 nursing students, and 22 pharmacy students. Intermediate medical students were included from their third year onwards (mean 4.7, SD 1.1; clinical phase) of a 6-year study program. Nursing and pharmacy students were included from their first study year onwards of a 3.5-year study program (mean 2.1, SD 0.5), respectively, of a 4-year study program (mean 2.1, SD 0.8 years). Participants were scheduled to train in interprofessional triads. In total, training was organized into 30 planned triads; due to anticipated but absent participants and the need to merge remaining attendees to keep sessions feasible, the realized training groups comprised 9 dyads and 2 tetrads. Previous experience with wound care was reported by 23 medical students, 4 nursing students, and 3 pharmacy students. Previous experience with VR was reported by 16 medical students, 4 nursing students, and 3 pharmacy students.

iVR Environment/learners' Task

The authors with backgrounds in medicine, psychology, learning sciences, and nursing education developed an iVR training in collaboration with medverse GmbH. The contents and procedures had been developed for the iVR environment based on established clinical standards [29] and were agreed upon through a consensus among wound care experts from medicine, nursing, and pharmacy, incorporating profession-specific perspectives. The iVR is divided into 3 sections: preparation, patient contact, and follow-up.

In the first section, preparation, participants find themselves in front of the patient's room. There, the senior physician presents the case using the SBAR (Situation-Background-Assessment-Recommendation) framework, which is a well-established standard for structured patient handovers [30]. SBAR improves nurse–physician communication and reduces unexpected death. The patient is an 81-year-old man with a category 3 pressure ulcer (EPUAP classification [31]) on his right heel. The ulcer developed following surgical treatment of a femoral neck fracture with total hip arthroplasty, after which the patient experienced severe postoperative pain and markedly limited mobility, leading to prolonged pressure exposure. At the time of the scenario, the wound measured 5.5 cm × 3 cm, was painful, moderately exudative, and showed scattered fibrin deposits. The clinical situation required coordinated wound care, appropriate pain management, pressure relief, and mobilization planning. Multimedia Appendix 2 provides a detailed description of the case materials used. The participants then gather for an initial team meeting to discuss the latest wound documentation. This is presented in written form and includes a detailed description of the wound (eg, size), previous treatment (eg, dressing), and the patient's reported pain level.

In the second section, patient contact, participants enter the patient's room, introduce themselves, explain their purpose, and ask about the patient's pain level. After, learners are guided to discuss pain management, review current therapy, and possible therapy adjustments. Further, they consider potential changes to the patient's position, as well as mobilization strategies. An animation demonstrates bandage removal. Participants examine the removed dressing and its documentation (eg, moisture), inspect the wound, and select a cleaning method and product from 2 options. They then receive feedback on whether their choice was correct. An animation illustrates the cleaning process. They photograph the wound, review it after cleaning with the written documentation, and decide on wound edge protection and dressings, again from 2 options, with feedback. Another animation shows the application of the bandage. Finally, they discuss patient instruction, considering the information needed for wound self-care, including infection risks and preventive measures.

In the final section, follow-up, participants leave the patient's room and gather for a final team meeting. They discuss the treatment provided and any further care required, such as intervals for changing the dressing, and the involvement of other professional groups. The iVR training concludes at this point. Figures 1 and 2 display screenshots from the iVR environment at different time points. The training covers adherence to German hygiene standards [32] in the wound care process. Participants learn when to disinfect their hands and surfaces, when to put on and change protective clothing, including gloves, and how to position equipment correctly.

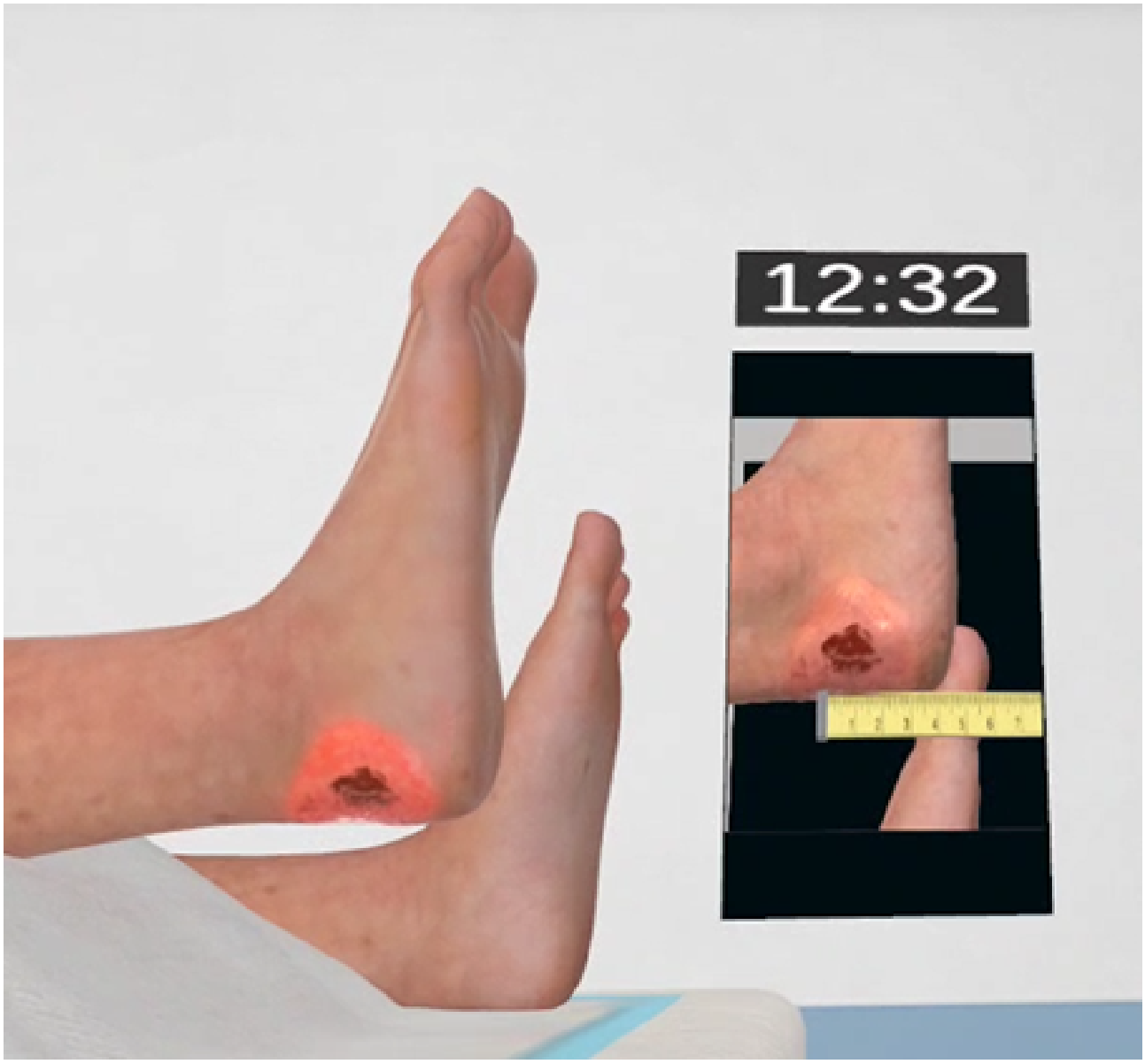
XSL•FO
RenderX

Figure 1. Participant interacting with the patient in the virtual patient room.



XSL•FO
RenderX

Figure 2. Detailed view of the pressure ulcer on the patient's right heel.



Procedure

The study was conducted as part of a larger data collection in June 2025. Recruitment took place via semester kick-off events, lectures, student learning platforms, messaging groups, and campus notices. Learners registered individually and were assigned to interprofessional triads (at least 2 different professions per group). Participation was voluntary; written informed consent was obtained prior to participation. Each participant received €20 (US $23.22) via bank transfer after participation, and compensation was fixed and not performance-based. Each session lasted approximately 2 hours and comprised a pretest, a learning phase (iVR training unit), and a posttest. Participants began individually with a tablet-based pretest covering demographics, prior experience with iVR and wound care, procedural knowledge, and willingness to collaborate. After standardized equipment instructions, triad members sat in the same room in a semicircle, put on their iVR headsets, took their controllers, and completed the iVR training unit (average duration: 25 min), communicating verbally but without facial cues. This was followed by a posttest assessing cybersickness, procedural knowledge, cognitive load, and willingness to collaborate. In addition, a second iVR training unit (average duration: 15 min) was administered, followed by a second posttest (posttest 2). The iVR training was developed based on shared learning objectives targeting both procedural wound care and interprofessional collaboration. The instructional design comprised 2 consecutive levels with constant procedural content but reduced collaborative guidance in the second level. Procedural knowledge was assessed at pretest and posttest 1, as the stepwise wound-care sequence remained unchanged across levels and the sequencing task was susceptible to retest effects. Willingness to collaborate was assessed at all 3 time points to allow examination of potential changes across repeated interprofessional engagement and varying degrees of collaborative support. Analyses involving posttest 2 are reported as exploratory.

Missing data resulted from technical issues (eg, temporary internet connectivity problems leading to headset malfunctions and occasional questionnaire interruptions) and time constraints,

XSL•FO
RenderX

particularly affecting posttest 2. No systematic dropout related to study variables was observed.

Instructional Support

Throughout the iVR training, participants were supported with scaffolding and collaboration scripts implemented as prompts. Scaffolding was provided via step-by-step tooltips aligned with the wound care process (eg, "Adjust the bed to working height"), displayed as pop-ups at hand level. Collaboration scripts were implemented at 7 predefined points identified with external wound experts from medicine, nursing, and pharmacy. For example, participants received the meta-prompt "Please come together for a brief team discussion on pain management." Discussions were structured by a checklist (eg, "Discuss previous pain therapy and possible adjustments"), both displayed above the patient's bed in the iVR environment (Figure 3).

Figure 3. Example of a meta-prompt with a checklist in the immersive virtual reality (iVR) to guide participants' collaboration.



In the second iVR training unit, only the meta-prompt was provided to encourage more self-structured discussion. All prompts and checklists are listed in Multimedia Appendix 3. Progression required completing all actions; some steps were triggered automatically (eg, tapping the bed), while checklist items required active confirmation via a "Next" button.

Measures

Procedural knowledge was assessed based on the expert-based wound care steps in the pre- and post-tests using a drag-and-drop sequencing task (Multimedia Appendix 4). The test covered all 12 steps of the wound care process. Each step was described in concise, clear wording (eg, "Assess the wound and change gloves," "Select wound edge protection and dressing, disinfect hands, put on gloves, dress the wound"). Responses were scored using Kendall τ, a rank correlation coefficient that quantifies the correspondence between the participant's ordering and the correct reference ordering. Higher values indicate better procedural knowledge, ranging from −1 (perfectly reversed order) to +1 (perfectly correct order), with values around 0 reflecting largely unsystematic or random ordering. Thus, positive shifts in Kendall τ reflect increasingly correct sequencing of clinically relevant procedural steps. Because the

XSL•FO
RenderX

task required ordering the complete procedure as a whole, individual steps were not scored independently, and step-level error rates were not separately coded.

Willingness to collaborate was measured using a current German translation (validation study ongoing) of the Interprofessional Socialization and Valuing Scale (ISVS) [33]. Willingness to collaborate is considered a key construct in interprofessional contexts, where it is critical for team effectiveness. This ability develops through a process of interprofessional socialization [33], during which clinicians adopt professional and interprofessional identities, recognize the value of other disciplines' contributions, and function confidently as collaborative team members. This socialization lays the groundwork for the delivery of high-quality, integrated care in complex clinical settings. Willingness to collaborate was assessed at 3 measurement timepoints (pretest, posttest 1, and posttest 2). Accordingly, 3 corresponding versions of the ISVS were used: ISVS-9A (9 items) at pretest, ISVS-9B (9 items) at posttest 1, and ISVS-21 (21 items) at posttest 2. These are validated equivalent forms of the same underlying construct. The two 9-item versions were specifically developed for repeated-measures designs and have demonstrated strong agreement and measurement invariance with each other. The use of different but equivalent forms was a deliberate methodological choice to reduce response burden and potential repetition effects across repeated measurements while maintaining comparability across timepoints. An example item from set A was "I am able to share and exchange ideas in a team discussion" and "I feel comfortable in describing my professional role to another team member." All items were rated on a 7-point Likert scale (1=does not apply at all, 7=fully applies; 0=no answer). Negatively worded items were reverse-coded so that higher scores indicate greater willingness to collaborate. For each participant, we calculated the mean score across all nonmissing items. Internal consistency was good (ω_{pre}=0.77; ω_{post1}=0.77).

Cognitive load was assessed using the scale by Klepsch et al [34]. Participants rated ICL and ECL on a 7-point Likert scale ranging from absolutely wrong to absolutely right. ICL was measured with 2 items (eg, "For this task, many things needed to be kept in mind simultaneously"; "This task was very complex"), with ω=0.63. ECL was measured with 3 items (eg, "The instructions and explanations during this task were very unclear"; "During this task, it was difficult to recognize and link the crucial information"), with ω_{post1}=0.67. For each participant, we calculated the mean score across all nonmissing items.

Cybersickness was assessed as a control variable using the German version of the Virtual Reality Symptom Questionnaire (VRSQ-G) [35]. The questionnaire contains 9 symptoms (eg, fatigue), each rated on a 4-point scale (0=not at all to 3=very much). Internal consistencies were good (ω_{post1}=0.87).

Analyses

To address RQ1 and RQ2, we examined changes in the primary outcomes across measurement points and then tested whether these changes were associated with cognitive load while controlling for cybersickness. For RQ1, we examined whether procedural knowledge increased from pretest to posttest 1 (H1.1) and whether this change was moderated by ECL (H1.2). In addition, we tested the proposed nonlinear association between ICL and procedural knowledge (H1.3). Profession-related differences were explored (H1.4).

For RQ2, willingness to collaborate was assessed at pretest, posttest 1, and posttest 2. We examined whether willingness to collaborate increased over time (H2.1) and tested whether these changes were moderated by ECL while controlling for cybersickness (H2.2). Analyses including posttest 2 were conducted exploratorily, as this third measurement point was not part of the preregistered hypotheses. Profession-related differences were explored (H2.3).

Change and moderation hypotheses were tested using repeated-measures analyses of covariance (ANCOVAs). For RQ1 (procedural knowledge), analyses included 2 measurement points (pretest and posttest 1, corresponding to the first iVR training level). For RQ2 (willingness to collaborate), repeated-measures ANCOVAs included 3 measurement points (pretest, posttest 1, and posttest 2). For moderation analyses, ECL was included as a covariate and cybersickness as a control variable. The proposed nonlinear effect of ICL was examined using a mediation model including a quadratic term (ICL^2). Exploratory profession-related differences were examined using ANCOVAs.

All analyses were based on the preregistered analysis plan described above. While the theoretical assumptions and target constructs remained unchanged, minor refinements were made to align the analytic implementation with the final study design and improve interpretability.

Because participants were trained in interprofessional triads, potential clustering effects were examined using linear mixed models with random intercepts for participants and triads. The necessity of multilevel modeling was evaluated using the design effect [36]. When the design effect was smaller than two, single-level models were retained as primary analyses, and mixed models were estimated as robustness checks.

Missing data were handled according to the estimation method used in each analysis. Repeated-measures ANCOVAs were estimated using listwise deletion within each model. Linear mixed models were estimated using restricted maximum likelihood and included all available observations. Mediation analyses were conducted using complete cases as implemented in Jamovi. All analyses were conducted in Jamovi (version 2.0; The Jamovi project, 2024) and can be retrieved from the Open Science Framework (Multimedia Appendix 1).

Ethical Considerations

The Ethics Committee of the Medical Faculty at LMU Munich confirmed that no formal ethics review was required for this study (number 25-0754-KB). The study was conducted in accordance with the ethical principles of the Declaration of Helsinki (WMA). All participants received information about the study procedures, the voluntary nature of participation, and their right to withdraw from the study at any time without disadvantage. Informed consent was obtained from all

XSL•FO
RenderX

participants prior to participation. Privacy and confidentiality were maintained throughout the study. Data were collected and analyzed in pseudonymized form, and no directly identifying information is reported in this manuscript. Each participant received €20 (US $23.22) via bank transfer after participation. Compensation was fixed and not performance-based.

Results

Overview

Professions were distributed across the sample. Table 1 contains descriptives of the study variables with the respective correlations in Table 2.

Table . Descriptive statistics of the study variables.

Variable	N	Missing	Mean	SD	Min	Max	*P* (Shapiro-Wilk)
Procedural knowledge pre	116	0	−0.111	0.0653	−0.333	0.0606	<.001
Procedural knowledge post	110	6	0.0501	0.0571	−0.212	0.242	<.001
Willingness to collaborate pre	100	16	5.6	0.631	3.67	6.78	.23
Willingness to collaborate post	97	19	5.58	0.666	4.11	7.0	.38
Willingness to collaborate post 2	90	26	5.74	0.608	3.78	6.89	.055
ECL[a] post	110	6	2.53	1.09	1.0	5.67	<.001
ICL[b] post	110	6	3.26	1.2	1.0	6.0	<.001
Cybersickness post	110	6	1.84	0.66	1.0	4.0	<.001

[a]ECL: extraneous cognitive load.

[b]ICL: intrinsic cognitive load.

Table . Correlations between the study variables. Pearson correlations are reported.

Variable	1	2	3	4	5	6	7	8
Procedural knowledge pre								
Procedural knowledge post	−.045							
Willingness to collaborate pre	−.161	.020						
Willingness to collaborate post	.006	.104	.527[a]					
Willingness to collaborate post 2	−.133	−.108	.567[a]	.632[a]				
ICL post	.000	−.063	.158	.165	.181			
ECL post	.029	.080	−.180	−.272[b]	−.180	.113		
Cybersickness post	−.018	.126	−.213[c]	−.202[c]	−.040	−.017	.343[a]	

[a]*P*<.001.

[b]*P*<.01.

[c]*P*<.05.

Because participants were trained in interprofessional triads, we examined whether clustering at the triad level affected the estimated effects. Linear mixed models including random intercepts for participants and triads were estimated for both

XSL•FO
RenderX

outcomes. For procedural knowledge, the intraclass correlation at the triad level was small (intraclass correlation coefficient [ICC]=0.035). For willingness to collaborate, the triad-level variance component was estimated at zero (ICC=0.000). The resulting design effects were below the commonly discussed threshold of two. As the pattern of fixed effects remained unchanged in the mixed models, the preregistered ANCOVAs are reported as the primary analyses.

RQ1: Effects of iVR Training on Procedural Knowledge and the Role of Cognitive Load

To address RQ1, we first examined whether procedural knowledge changed from pretest to posttest 1. Mean values increased from −0.11 at pretest to 0.05 at posttest, indicating that the ordering of steps changed from a systematic error to a level slightly above random. This change was tested using a repeated-measures ANCOVA (n=110) with measurement point (pretest, posttest) as a within-subjects factor, ECL as a covariate, and cybersickness as a control variable. The analysis revealed a significant main effect of measurement point, $F_{1,\ 107}$=26.19, P<.001, η^2=.08, indicating a gain in procedural knowledge. Thus, H1.1 is supported. On average, students moved from a slightly systematically incorrect ordering at pretest (negative Kendall τ) to a predominantly correct ordering at posttest (positive Kendall τ), reflecting a meaningful reorganization of the clinically relevant wound-care sequence.

A post-hoc comparison confirmed that procedural knowledge was significantly higher at posttest compared with pretest (t_{107}=19.60, P<.001), with a mean difference of 0.163. Neither the interaction between measurement point and ECL ($F_{1,\ 107}$=.00, P=.99, η^2=.00) nor the interaction between measurement point and cybersickness ($F_{1,\ 107}$=0.89, P=.35, η^2=.00) was significant. Thus, H1.2 was not supported. H1.3 posited a quadratic mediating effect of ICL on procedural knowledge. The mediation model (n=110), including ICL^2 showed no statistically significant effects. There was no indirect effect (b=0.00, SE=0.00, Z=0.014, P=.99), no direct effect (b=–0.04, SE=0.09, Z=–0.472, P=.64), and no total effect (b=–0.04, SE=0.09, Z=–0.470, P=.64). Thus, H1.3 was not supported. Further, the exploratory analysis (n=104) revealed no significant differences between professions ($F_{2,\ 101}$=0.07, P=.93, η^2=.00). Thus, H1.4 was not supported.

RQ2: Effects of iVR Training on the Willingness to Collaborate and the Role of Cognitive Load

To address RQ2, we examined whether willingness to collaborate changed from pre to posttest. Mean willingness to collaborate was 5.6 at pre and 5.58 at posttest. A repeated measures ANCOVA (n=83) with measurement point (pre, posttest) as a within-subjects factor, ECL as a covariate, and cybersickness as a control variable revealed no significant main effect of measurement point ($F_{1,\ 80}$=3.55, P=.06, η^2=.01), indicating no measurable change in willingness to collaborate from pre- to posttest and thus not supporting H2.1. A post hoc comparison confirmed that willingness to collaborate did not differ significantly between pre- and posttest (t_{80}=–0.54, P=.59), with a mean difference of –0.036. Neither the interaction between measurement point and ECL was significant ($F_{1,\ 80}$=3.71, P=.058, η^2=.01), nor the interaction between measurement point and cybersickness ($F_{1,\ 80}$=0.51, P=.477, η^2=.001). Thus, H2.2 was not supported. Exploratory analyses (n=83) including the third measurement point (posttest 2) showed that willingness to collaborate was significantly higher at posttest 2 than at posttest 1 (t_{64}=3.16, P=.007, mean difference=0.202) and significantly higher at posttest 2 than at pretest (t_{64}=2.47, P=.043, mean difference=0.167). No significant difference emerged between pretest and posttest 1 (t_{64}=–0.48, P=.882, mean difference=–0.034). Further, no significant differences were found between professions ($F_{2,\ 75}$=0.23, P=.794, η^2=.00). Thus, H2.3 was not supported.

Discussion

Principal Findings

This study investigated the effects of a newly developed iVR training on wound care, completed in interprofessional groups of students from medicine, nursing, and pharmacy. Specifically, we examined whether the training improves learners' procedural knowledge and willingness to collaborate, and how these outcomes are shaped by cognitive load.

With respect to RQ1, learners demonstrated a large and statistically significant gain in procedural knowledge from the pre to posttest, thus confirming H1.1. Importantly, this shift reflects a move from slightly systematically incorrect sequencing at baseline to predominantly correct procedural ordering after the training.

This finding is consistent with previous studies showing that iVR can promote procedural knowledge [18] by simulating realistic clinical processes and enabling repeated, safe practice [10]. The effect was observed across medical, nursing, and pharmacy students who all started with low levels of prior knowledge and showed marked gains after the training, independent of their professional background. Contrary to H1.2, ECL did not moderate this knowledge gain. This may have been because ECL was managed effectively through the scaffolding [25] and collaboration scripts [22,23] implemented in the iVR. More precisely, we implemented scaffolding by providing a step-by-step structure that broke wound care down into smaller, more manageable steps. This guided learners through the entire process, ensuring that they did not overlook essential actions, since they had to confirm or interact in order to proceed. In addition, animations of complex handling tasks (eg, unwrapping a bandage) reduced task complexity, helping learners avoid overload and focus on the essential learning content. Similarly, some hygiene steps (eg, hand disinfection) required only confirmation rather than haptic execution at a dispenser, further lowering unnecessary demands. Our structured design may have enabled learners with limited prior knowledge to concentrate on the essential steps, possibly without becoming cognitively overloaded. In line with cognitive load theory [17], such instructional support may contribute to a reduction in ECL, enabling learners to focus on the task itself. Nevertheless, caution is required when interpreting this. As outlined in the Limitations section, modest subscale reliability and self-report

XSL•FO
RenderX

constraints may have limited sensitivity. Thus, the absence of cognitive load effects should not be attributed solely to instructional load control through the instructional design.

According to H1.3, ICL was expected to show an inverted U-shaped effect, but this was not supported. One possible explanation is that all learners experienced a moderate level of ICL, enabling them to benefit from the learning environment and achieve overall learning gains. This interpretation is consistent with previous theoretical models' [37] findings on the relationship between ICL and self-regulation. In line with this pattern, no profession-related differences were found in terms of gains in procedural knowledge either. Contrary to H1.4, this indicates that the iVR training effectively fostered procedural knowledge across different professional backgrounds, providing comparable benefits for medical, nursing, and pharmacy students.

With respect to RQ2, there was no significant increase in willingness to collaborate between the pre- and posttest, which provides no support for H2.1. However, the exploratory analysis revealed an improvement in learners' willingness to collaborate after the second iVR session. This suggests that a single 25-minute iVR session may be insufficient to change learners' attitudes towards collaboration, as these attitudes are developed over time through an interprofessional socialization process [33,38]. An attitudinal change may require more time to unfold. This pattern also aligns with the idea that interprofessional values, role understandings, communication skills, and teamwork competencies develop gradually over an extended period of training and interaction within interprofessional teams [6,33]. A single exposure, even within an immersive and structured environment, may not be sufficient to induce measurable change on a self-report scale. To increase sensitivity for this outcome, future studies could incorporate repeated or longitudinal interprofessional VR sessions, structured debriefings, or explicit perspective-taking tasks to deepen reflection on professional roles. In addition, combining self-report measures with behavioral indicators of collaboration quality may provide a more fine-grained assessment of interprofessional development.

Contrary to H2.2, ECL did not moderate changes in willingness to collaborate. One possible explanation might be that learners were able to effectively manage the task with the help of instructional support [22,23]. In our study, we implemented collaboration scripts by structuring the team process, indicating when discussions should take place, and providing checklists detailing the issues to be addressed in each meeting. Learners therefore received individual prompts such as "Please come together for a brief team discussion on pain management," and structured checklists highlighted the topics that they needed to discuss. Learners could progress to the next step only after confirming that they had addressed all points. However, again, caution is required when interpreting the absence of ECL-related effects on willingness to collaborate as evidence of effective instructional load control. Furthermore, contrary to H2.3, no differences in willingness to collaborate based on profession were observed. This indicates that nursing, pharmacy, and medical students increased their willingness to collaborate similarly after the iVR training. Overall, iVR training led to an immediate and substantial increase in procedural knowledge, whereas willingness to collaborate improved only after repeated exposure, indicating that attitudinal outcomes require a longer time frame to develop. While the instructional design may have contributed to structuring task demands, this finding should be interpreted in light of the modest reliability of the cognitive load subscales and the limitations of self-report measures in immersive and collaborative VR environments.

Limitations and Further Research

Several limitations should be considered when interpreting the findings. The study used a single-group pretest–posttest design without a comparison condition. This design was chosen to evaluate the feasibility and initial learning effects of a newly developed interprofessional VR training under controlled conditions. While causal attribution to the instructional format itself cannot be established within this design, and alternative explanations such as testing or maturation effects cannot be fully excluded, it allows a focused examination of within-group development during early-stage implementation. Comparative and noninferiority VR trials exist in other domains [39]; however, such designs are not always required at early stages of instructional development, particularly when the primary aim is to examine feasibility, acceptability, and initial learning processes within a new interprofessional context. This study, therefore, addressed a distinct research question focused on interprofessional implementation and within-group development. Future research should build on these findings using controlled designs to determine under which conditions iVR provides added value within IPE.

The collaborative interaction relied exclusively on verbal communication without nonverbal cues, which may limit ecological validity compared with in-person interprofessional teamwork. Future studies should examine how integrating nonverbal communication elements, for example, through augmented or mixed-reality environments, influences interprofessional collaboration processes.

Moreover, this study focused on learning-related outcomes and did not systematically evaluate structural or economic aspects of implementation. At the same time, a key structural advantage of iVR is that, once developed and technically implemented, it can be deployed independently of dedicated simulation centers, potentially lowering access barriers for institutions with limited simulation infrastructure. Future research should therefore examine how such VR-based formats can be sustainably integrated into IPE and evaluate their scalability and comparative effectiveness under routine conditions.

Measurement-related constraints should also be acknowledged. Internal consistency for cognitive load was moderate, and self-report measures may only partially capture the multidimensional demands of collaborative iVR environments, which combine sensory immersion, interface navigation, clinical decision-making, and real-time interprofessional coordination. In such settings, learners may experience overlapping cognitive, social, and perceptual demands that are difficult to fully disentangle through brief retrospective ratings. Willingness to collaborate was assessed using validated equivalent ISVS forms across time points, which reduces repetition effects and

XSL•FO
RenderX

participant burden but limits strict item-level longitudinal comparability, particularly for the exploratory third measurement.

In addition, procedural knowledge was assessed at pretest and posttest 1 to avoid retest effects of the sequencing task and to align with the preregistered focus on the first training level. This design does not permit conclusions about retention or additional gains after repeated exposure. The study also did not include process-based indicators of collaboration quality, as the primary focus was on attitudinal and knowledge-related outcomes. Future research could complement these measures with longitudinal assessments, parallel or alternative knowledge indicators across multiple training iterations, and process-oriented metrics of interprofessional interaction.

Further, the sample consisted of students whose professional identities are still developing. While this limits direct transferability to experienced clinicians, early professional socialization is a core objective of IPE. This suggests that iVR training should be delivered as a series of sessions rather than as a one-off event, as repeated exposure may help learners build trust and appreciation for other professions and gradually increase openness and readiness to collaborate. Future research should therefore examine how collaborative attitudes and skills develop longitudinally across repeated IPE experiences and at more advanced stages of professional training. While our training already includes an in-built debriefing component, this could be complemented by facilitated interprofessional reflection outside the VR environment. In addition, collaboration quality could be captured via observable and codable behavioral indicators (eg, closed-loop communication) instead of being assessed solely through self-report measures.

Finally, the training focused on a single wound-care case with structured decision pathways to maintain manageable cognitive demands within the immersive environment. This design enabled a focused examination of interprofessional coordination in a clearly defined clinical scenario. At the same time, clinical wound management is inherently context-dependent, shaped by patient-specific factors, institutional routines, and practical constraints, and different constellations may warrant alternative decisions.

Future work should therefore extend immersive interprofessional simulations to diverse cases and decision pathways in order to examine transfer across clinical contexts.

Implications

The findings of this study have several implications for the design and implementation of IPE. First, the substantial gain in procedural knowledge among all learners indicates that iVR training can be used effectively. Integrating such training early on into the curriculum could help to ensure that students of medicine, nursing, and pharmacy acquire a shared foundation precisely in domains where their prior knowledge is typically low but clinical relevance is high, such as wound care.

Second, the results suggest that fostering interprofessional attitudes needs repeated, well-structured learning opportunities. In our study, willingness to collaborate only increased after the second training session, demonstrating that such changes do not occur immediately. This suggests that iVR training should be delivered as a series of sessions rather than as a one-off event. Repeated exposure allows learners to build trust and appreciation for other professions and fosters a gradual shift towards openness and readiness to collaborate. Third, the instructional design features used, such as step-by-step scaffolding throughout the case and structured collaboration prompts via meta-prompt plus checklist, proved to be feasible and effective. These features can serve as instructional blueprints for other interprofessional learning contexts. These elements reduce unnecessary demands and provide clear guidance for interaction, enabling learners to focus on essential steps and collaborative decision-making. These elements can be flexibly applied to iVR scenarios, which is more difficult to achieve in real-world settings with simulated patients.

Finally, the study highlights the curricular and organizational potential of iVR. Beyond acute scenarios, iVR enables systematic training in areas of practice that are otherwise difficult to integrate into curricula and where learners typically have little prior exposure. Its standardization and relatively low resource requirements make it a scalable approach to IPE that can be implemented even in institutions without large simulation centers, thereby broadening access to authentic interprofessional learning opportunities.

Conclusions

This study provides initial evidence that iVR can support IPE by improving procedural knowledge in a standardized, theory-based learning environment. Students from medicine, nursing, and pharmacy showed significant gains in procedural knowledge after the iVR wound care training, suggesting that immersive, scaffolded, and script-based instructional designs may facilitate complex skill acquisition across professional groups. In contrast, willingness to collaborate increased only after repeated exposure, suggesting that attitudinal and interprofessional competencies require sustained, iterative learning experiences rather than single training sessions. These training effects were independent of the learners' professions. Overall, the findings highlight the potential of iVR as a scalable approach for integrating interprofessional training into health care curricula. Repeated and structured iVR sessions could complement existing IPE programs, particularly in domains where task coordination and shared decision-making are central to patient care. Future research should examine long-term iVR effects on collaboration attitudes and behavior, extend iVR evaluations to various practicing professionals, and develop refined measurement tools for cognitive and collaborative demands in iVR for IPE.

XSL•FO
RenderX

Acknowledgments

The authors thank the students who participated in the study, as well as the student assistants Feyza Arslan and Raisa Kiriakidou for their valuable support with data collection. The authors also thank Raisa Kiriakidou for her assistance with language checks. We further acknowledge the Department of General Practice and Health Services Research, University Hospital Heidelberg, Heidelberg, Germany, for providing the German version of the Interprofessional Socialization and Valuing Scale (ISVS) used in this study. ChatGPT and DeepL Write were used to improve readability and language. All generated content was reviewed and edited by the authors, who take full responsibility for the final manuscript.

Funding

Stiftung Innovation in der Hochschullehre, Treuhandstiftung in Trägerschaft der Toepfer Stiftung GmbH (Freiraum 2022; FR-153/2023).

Data Availability

The data sets generated and analyzed during this study are available in the Open Science Framework (OSF) repository in Multimedia Appendix 1.

Authors' Contributions

Data curation: CR (lead), MB (supporting), MJW (supporting)
Formal analysis: CR (lead), MB (equal)
Funding acquisition: MJW (equal), MS (equal), MRF (equal)
Investigation: CR (lead), MB (equal), AW (supporting)
Methodology: CR (lead), MB (supporting), MJW (supporting), MS (supporting)
Project administration: CR (lead), MB (equal), MJW (equal), MS (supporting)
Resources: CR (lead), MB (equal), MJW (equal), AW (supporting), MS (supporting)
Supervision: CR (lead), MJW (supporting), MS (supporting)
Visualization: MB (lead), CR (supporting)
Writing – original draft: MB (lead), CR (equal)
Writing – review and editing: MS (equal), MJW (equal), AW (supporting), MRF (supporting), MB (supporting), CR (supporting)

Conflicts of Interest

None declared.

Multimedia Appendix 1
Preregistration and analyses.
[PDF File, 28 KB - mededu_v12i1e85139_app1.pdf]

Multimedia Appendix 2
Patient presentation and wound documentation.
[PDF File, 124 KB - mededu_v12i1e85139_app2.pdf]

Multimedia Appendix 3
Instructional support.
[PDF File, 222 KB - mededu_v12i1e85139_app3.pdf]

Multimedia Appendix 4
Procedural knowledge test.
[PDF File, 310 KB - mededu_v12i1e85139_app4.pdf]

References

1. Falanga V, Isseroff RR, Soulika AM, et al. Chronic wounds. Nat Rev Dis Primers 2022 Jul 21;8(1):50. [doi: 10.1038/s41572-022-00377-3] [Medline: 35864102]
2. Graves N, Phillips CJ, Harding K. A narrative review of the epidemiology and economics of chronic wounds. Br J Dermatol 2022 Aug;187(2):141-148. [doi: 10.1111/bjd.20692] [Medline: 34549421]
3. Olsson M, Järbrink K, Divakar U, et al. The humanistic and economic burden of chronic wounds: a systematic review. Wound Repair Regen 2019 Jan;27(1):114-125. [doi: 10.1111/wrr.12683] [Medline: 30362646]

XSL•FO
RenderX

4. Boersema GC, Smart H, Giaquinto-Cilliers MGC, et al. Management of nonhealable and maintenance wounds: a systematic integrative review and referral pathway. Adv Skin Wound Care 2021 Jan;34(1):11-22. [doi: 10.1097/01.ASW.0000722740.93179.9f] [Medline: 33323798]
5. WHO. Framework for Action on Interprofessional Education & Collaborative Practice: WHO; 2010.
6. Witti MJ, Zottmann JM, Wershofen B, Thistlethwaite JE, Fischer F, Fischer MR. FINCA - a conceptual framework to improve interprofessional collaboration in health education and care. Front Med (Lausanne) 2023;10:1213300. [doi: 10.3389/fmed.2023.1213300] [Medline: 37849484]
7. Kaap-Fröhlich S, Ulrich G, Wershofen B, et al. Position paper of the GMA committee interprofessional education in the health professions - current status and outlook. GMS J Med Educ 2022;39(2):Doc17. [doi: 10.3205/zma001538] [Medline: 35692364]
8. Maddock B, Dārziņš P, Kent F. Realist review of interprofessional education for health care students: what works for whom and why. J Interprof Care 2023;37(2):173-186. [doi: 10.1080/13561820.2022.2039105] [Medline: 35403557]
9. Djukic M, Adams J, Fulmer T, et al. E-learning with virtual teammates: a novel approach to interprofessional education. J Interprof Care 2015;29(5):476-482. [doi: 10.3109/13561820.2015.1030068] [Medline: 26120894]
10. Qiao J, Xu J, Li L, Ouyang YQ. The integration of immersive virtual reality simulation in interprofessional education: a scoping review. Nurse Educ Today 2021 Mar;98:104773. [doi: 10.1016/j.nedt.2021.104773] [Medline: 33497990]
11. Tschannen D, Dorn R, Tedesco C. Improving knowledge and behavior of leadership and followership among the interprofessional team. Int J Med Educ 2018 Jun 29;9:182-188. [doi: 10.5116/ijme.5b30.9a84] [Medline: 29965798]
12. Williams D, Stephen LA, Causton P. Teaching interprofessional competencies using virtual simulation: a descriptive exploratory research study. Nurse Educ Today 2020 Oct;93:104535. [doi: 10.1016/j.nedt.2020.104535] [Medline: 32717697]
13. Lewis KO, Popov V, Fatima SS. From static web to metaverse: reinventing medical education in the post-pandemic era. Ann Med 2024 Dec;56(1):2305694. [doi: 10.1080/07853890.2024.2305694] [Medline: 38261592]
14. Abbas JR, O'Connor A, Ganapathy E, et al. What is virtual reality? a healthcare-focused systematic review of definitions. Health Policy Technol 2023 Jun;12(2):100741. [doi: 10.1016/j.hlpt.2023.100741]
15. Neher AN, Bühlmann F, Müller M, Berendonk C, Sauter TC, Birrenbach T. Virtual reality for assessment in undergraduate nursing and medical education - a systematic review. BMC Med Educ 2025 Feb 22;25(1):292. [doi: 10.1186/s12909-025-06867-8] [Medline: 39987099]
16. Jennett C, Cox AL, Cairns P, et al. Measuring and defining the experience of immersion in games. Int J Hum Comput Stud 2008 Sep;66(9):641-661. [doi: 10.1016/j.ijhcs.2008.04.004]
17. Sweller J. Element interactivity and intrinsic, extraneous, and germane cognitive load. Educ Psychol Rev 2010 Jun;22(2):123-138. [doi: 10.1007/s10648-010-9128-5]
18. Jongbloed J, Chaker R, Lavoué E. Immersive procedural training in virtual reality: a systematic literature review. Comput Educ 2024 Nov;221:105124. [doi: 10.1016/j.compedu.2024.105124]
19. Makransky G, Lilleholt L. A structural equation modeling investigation of the emotional value of immersive virtual reality in education. Educ Technol Res Dev 2018 Oct;66(5):1141-1164. [doi: 10.1007/s11423-018-9581-2]
20. Marion-Martins AD, Pinho DLM. Interprofessional simulation effects for healthcare students: a systematic review and meta-analysis. Nurse Educ Today 2020 Nov;94:104568. [doi: 10.1016/j.nedt.2020.104568] [Medline: 32932058]
21. Mahler L, Conrad PV, Busch D, Erfurt-Berge C. Experts teaching wound management-a survey of current practices in Germany. Int Wound J 2025 Jun;22(6):e70694. [doi: 10.1111/iwj.70694] [Medline: 40419292]
22. Fischer F, Kollar I, Stegmann K, Wecker C. Toward a script theory of guidance in computer-supported collaborative learning. Educ Psychol 2013 Jan;48(1):56-66. [doi: 10.1080/00461520.2012.748005] [Medline: 23378679]
23. Vogel F, Wecker C, Kollar I, Fischer F. Socio-cognitive scaffolding with computer-supported collaboration scripts: a meta-analysis. Educ Psychol Rev 2017 Sep;29(3):477-511. [doi: 10.1007/s10648-016-9361-7]
24. Seufert T. The interplay between self-regulation in learning and cognitive load. Educ Res Rev 2018 Jun;24:116-129. [doi: 10.1016/j.edurev.2018.03.004]
25. Belland BR, Walker AE, Kim NJ, Lefler M. Synthesizing results from empirical research on computer-based scaffolding in STEM education: a meta-analysis. Rev Educ Res 2017 Apr;87(2):309-344. [doi: 10.3102/0034654316670999] [Medline: 28344365]
26. Reeves S, Zwarenstein M, Goldman J, et al. The effectiveness of interprofessional education: key findings from a new systematic review. J Interprof Care 2010 May;24(3):230-241. [doi: 10.3109/13561820903163405] [Medline: 20178425]
27. Mertens JF, Hessel MHM, Kempen TGH, et al. Evaluation of an interprofessional education program involving medical and pharmacy students: a mixed-method study. BMC Med Educ 2025 Jan 11;25(1):48. [doi: 10.1186/s12909-024-06574-w] [Medline: 39799314]
28. Wong AKC, Wong FKY, Chan LK, Chan N, Ganotice FA, Ho J. The effect of interprofessional team-based learning among nursing students: a quasi-experimental study. Nurse Educ Today 2017 Jun;53:13-18. [doi: 10.1016/j.nedt.2017.03.004] [Medline: 28340482]
29. Deutsches Netzwerk für Qualitätsentwicklung in der Pflege (DNQP). Expertenstandard Pflege von Menschen Mit Chronischen Wunden – Einschließlich Kommentierung Und Literaturstudie 2 Aktualisierung: Hochschule Osnabrück, DNQP; 2025.

XSL•FO
RenderX

30. De Meester K, Verspuy M, Monsieurs KG, Van Bogaert P. SBAR improves nurse-physician communication and reduces unexpected death: a pre and post intervention study. Resuscitation 2013 Sep;84(9):1192-1196. [doi: 10.1016/j.resuscitation.2013.03.016] [Medline: 23537699]
31. Kottner J, Cuddigan J, Carville K, et al. Pressure ulcer/injury classification today: an international perspective. J Tissue Viability 2020 Aug;29(3):197-203. [doi: 10.1016/j.jtv.2020.04.003] [Medline: 32414554]
32. Händehygiene in Einrichtungen des Gesundheitswesens: Empfehlung der Kommission für Krankenhaushygiene und Infektionsprävention (KRINKO) beim Robert Koch-Institut (RKI). Bundesgesundheitsblatt Gesundheitsforschung Gesundheitsschutz 2016;59(9):1189-1220. [doi: 10.1007/s00103-016-2416-6] [Medline: 27558147]
33. King G, Orchard C, Khalili H, Avery L. Refinement of the interprofessional socialization and valuing scale (ISVS-21) and development of 9-item equivalent versions. J Contin Educ Health Prof 2016;36(3):171-177. [doi: 10.1097/CEH.0000000000000082] [Medline: 27583993]
34. Klepsch M, Schmitz F, Seufert T. Development and validation of two instruments measuring intrinsic, extraneous, and germane cognitive load. Front Psychol 2017;8:1997. [doi: 10.3389/fpsyg.2017.01997] [Medline: 29201011]
35. Biniok M, Forbrig TA, Gellert P, Gräske J. Analysis of cybersickness in virtual nursing simulation: a German longitudinal study. BMC Nurs 2024 Mar 20;23(1):187. [doi: 10.1186/s12912-024-01833-z] [Medline: 38509512]
36. Lai MHC, Kwok OM. Examining the rule of thumb of not using multilevel modeling: the “Design Effect Smaller Than Two” rule. J Exp Educ 2015 Jul 3;83(3):423-438. [doi: 10.1080/00220973.2014.907229]
37. Seufert T, Hamm V, Vogt A, Riemer V. The interplay of cognitive load, learners’ resources and self-regulation. Educ Psychol Rev 2024 Jun;36(2):50. [doi: 10.1007/s10648-024-09890-1]
38. Zottmann JM, Dieckmann P, Taraszow T, Rall M, Fischer F. Just watching is not enough: fostering simulation-based learning with collaboration scripts. GMS J Med Educ 2018;35(3):Doc35. [doi: 10.3205/zma001181] [Medline: 30186945]
39. Popov V, Mateju N, Jeske C, Lewis KO. Metaverse-based simulation: a scoping review of charting medical education over the last two decades in the lens of the “marvelous medical education machine”. Ann Med 2024 Dec;56(1):2424450. [doi: 10.1080/07853890.2024.2424450] [Medline: 39535116]

Abbreviations

ANCOVA: analysis of covariance
ECL: extraneous cognitive load
ICC: intraclass correlation coefficient
ICL: intrinsic cognitive load
IPE: interprofessional education
ISVS: Interprofessional Socialization and Valuing Scale
iVR: immersive virtual reality
SBAR: Situation-Background-Assessment-Recommendation
VRSQ-G: Virtual Reality Symptom Questionnaire - German version

Edited by A Hasan Sapci; submitted 03.Nov.2025; peer-reviewed by Q Zhou, T Dege; revised version received 26.Feb.2026; accepted 05.Apr.2026; published 27.May.2026.

Please cite as:
Bauer M, Witti MJ, Stadler M, Weiß A, Fischer MR, Richters C
Interprofessional Training in Virtual Reality for Health Care: Experimental Study on Procedural Knowledge and Willingness to Collaborate
JMIR Med Educ 2026;12:e85139
URL: https://mededu.jmir.org/2026/1/e85139
doi:10.2196/85139

© Melanie Bauer, Matthias J Witti, Matthias Stadler, Anna Weiß, Martin R Fischer, Constanze Richters. Originally published in JMIR Medical Education (https://mededu.jmir.org), 27.May.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Education, is properly cited. The complete bibliographic information, a link to the original publication on https://mededu.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Development and Validation of a Scale for Measuring Leadership and Managerial Competencies of Middle Managers in Health Care and Medical Education in the Gulf Region: Cross-Sectional Study

Ahmed Mohamed Al Ansari[1,2*], MD, PhD; Elizabeth L Holloway[3*], PhD; Reginald Paul Sequeira[4*], PhD

[1]National Health Regulatory Authority, 1011, Manama, Bahrain

[2]Royal College of Surgeons in Ireland - Bahrain, Muharraq, Bahrain

[3]Graduate School of Leadership & Change, Antioch University, Yellow Springs, Ohio, United States

[4]Department of Pharmacology & Therapeutics, College of Medicine and Health Sciences, Arabian Gulf University, Manama, Bahrain

[*]all authors contributed equally

Corresponding Author:
Ahmed Mohamed Al Ansari, MD, PhD
National Health Regulatory Authority, 1011, Manama, Bahrain

Abstract

Background: The Kingdom of Saudi Arabia and the Kingdom of Bahrain are transforming their health care systems toward more self-sustained, autonomous systems. Effective leadership at all management levels, particularly middle management, is critical for operational success.

Objective: The study aimed to develop a feasible, reliable, and valid scale for measuring leadership and managerial competencies of middle managers in health care and medical education in the Gulf region.

Methods: A mixed methods approach was adopted. Semistructured interviews were conducted with top management, middle management, and lower management staff (N=27). Reflexivity of coders and thematic saturation were considered to ensure the reliability and transparency of qualitative findings. Thematic analysis informed scale item creation. A total of 202 participants from medical education and health care sectors completed the scale. Cronbach α, exploratory factor analysis (principal axis factoring), and confirmatory factor analysis validated internal consistency and factor structure. Responses marked as "6=unable to assess" were treated as missing data during quantitative analysis.

Results: Seven themes emerged: personality, agility, attitude, managerial skills, work ethics, mental ability, and interaction. Three constructs were identified: professionalism and problem-solving, team management and adaptation, and time management and expertise. A validated scale with 17 competency measure points and 16 characteristic items was finalized. Only competency measure points underwent exploratory factor analysis and confirmatory factor analysis, while characteristic items were retained for descriptive evaluation.

Conclusions: The Leadership and Managerial Competency Scale for Middle Managers in the Gulf Region reflects perceived competencies valued at all management levels. The scale measures perceived importance, not actual performance; the conclusions align with the validated scope.

(JMIR Med Educ 2026;12:e77476) doi:10.2196/77476

KEYWORDS

leadership; managerial competency; medical education; health care; management; middle management; lower management; scale development; middle managers

Introduction

Globally, the health care sector is experiencing significant transformation as more effective and efficient models of care delivery are developed in response to pressing challenges. Health care systems in the Gulf region face the challenge of delivering effective responses to issues such as communicable diseases, rapid population growth, a rising older population, socioeconomic reforms, liberal urbanization, and changes in lifestyle, which reflect on the health determinants of the population [1-3]. The Kingdom of Saudi Arabia and the Kingdom of Bahrain are currently transforming their health care systems from a government system to a more self-sustained autonomous system [1,2]. This goal requires effective leadership at all management levels, especially at the middle management level, which is vital for effective planning and operational implementation [4,5].

XSL•FO
RenderX

Leadership has been discussed from diverse perspectives such as philosophical, managerial, and psychological [3,6,7]. Key themes identified across leadership definitions include the idea that leaders work toward attaining a shared objective [8,9], leaders unify and influence rather than dictate, and the results of leadership constitute an improved current state [10,11]. Leaders are involved in the selection, equipping, training, and influencing of followers, who themselves possess a collection of diverse skills, talents, and abilities [12-15]. Managerial perspectives discuss leadership roles in critical organizational functions such as strategy execution and resource use [15-17]. In the health care sector, efficient leadership is central to improvements as organizations seek to improve quality and efficiency in health care [18,19], improve the management of human resources, achieve turnover and cost reductions, as well as fulfill specific strategic priorities [13,20-22]. Leadership in health care is also tasked with functions such as addressing the disempowerment of marginalized groups, improving accountability, fostering transparency and cross-communication, encouraging innovations, and expanding awareness of pertinent clinical and administrative issues [13,23,24]. Preparation for health care leadership is currently embedded in the training of physicians [21,25,26]. The leadership components prepare doctors to handle complex leadership and management scenarios, the evolving needs of stakeholders in the health care sector, and various external factors that they might encounter in the workplace and throughout their careers [9,14,20,27-30].

Mid-level management is in the middle of hierarchical managerial systems, situated between top-level managers and managers of the lowest (first) level. Mid-level managers are responsible for the implementation of top management decisions; they interpret the strategy set by top-level management and oversee its implementation [3,31,32]. They are authorized to manage the organizational units and are responsible for the results of these units. The role of the middle manager includes directing personnel, administering resources, operations control, and solving complex issues within organizations [4,33-35]. Effective middle management is critical for the success of the organizations, as the agency is responsible for translating strategies into successful operations and achieving high-performance cultures [5,24,36,37]. Middle managers play a central role in implementing health care innovation, particularly by addressing the informational gaps between senior management and frontline clinicians, mediating between strategy and day-to-day activities, and building momentum for innovation implementation [33,36,38]. Successful leaders and middle managers in the health care sector must therefore possess a complex set of skills across several areas, including clinical, leadership, managerial, social, and personal aspects [8,26,39-41].

Several studies emphasize the connection between transformational leadership and managerial behavior, in general, and middle management, in particular [42]. Such leaders communicate a well-articulated vision, create a sense of belonging, encourage employees to adjust positively to changes, and provide coaching to subordinates [23,42-45]. Other intersecting competencies identified for middle managers in the literature include cultural competency, the ability to effectively manage a workforce comprised of individuals from multiple cultural and ethnic groups and adapt to ways of doing business in new cultures and environments [40,46], learning-oriented leadership and at-work learning [41], and high emotional intelligence [47].

Arab leadership forms the underlying framework for health care leadership in the Middle East. One of the main differences between Arab leadership and other kinds of leadership (Western leadership) is that the Arab leadership is based on the Quran and accepted Hadiths [48,49]. In the Quran, God provides people with principles, tools, and skills that are required in leading their lives, relating with others, and in realizing their fullest potential. Arab leaders are not free to act as they choose or must submit to the wishes of others but must only act to implement the laws of Allah on earth [48,50]. Thus, the motivational agencies in this style of leadership derive from spiritual sources. Arab leadership is characterized by spiritual and religious behavior and is oriented toward people rather than products [1,2,48,50]. Arab leadership (Islamic leadership) is transformational, as leaders actively work to facilitate and enable transformation in their followers in tandem with Islamic values [49]. As in other cultural contexts, in Arab contexts, middle managers often oversee the implementation of quality improvement and innovative initiatives within their facilities, including innovations in health care [32,36,39], all while working within the tenets of Islamic values. Developments, such as increasing complexities in health financing, new institutional models, changes in regulation, insurance models, challenges in assuring patient safety and high-quality care, as well as challenges in meeting strategic priorities, have now made it urgent to consider the key competencies that characterize successful second-line leaders in the health care sector [39,51-53]. This study evaluated the competencies required for effective middle management in health care institutions and medical educational institutions in the kingdoms of Saudi Arabia and Bahrain. The leadership and managerial competencies identified in the study were then applied to construct a scale for required competencies for middle managers in health care and medical education in the Gulf region.

Methods

Study Design

The study was implemented using a mixed methods approach involving both a qualitative and a quantitative phase. The qualitative phase involved 3 distinct steps: the collection of primary data using semistructured interviews; the analysis of the data using emergent thematic analysis; and the conversion of the outputs from data analysis into scale items. Scale development involved developing a reliable and validated measure of the construct such that an attribute of interest was measured using an appropriate numerical dimension [54]. This study followed the 5-step mixed method proposed by Zhou [55] as the conceptual framework for scale development.

Step 1: Qualitative Exploration

Semistructured interviews, using a validated interview guide, were conducted to develop items for the leadership and managerial competencies scale. The interview guide was developed by the research team and is provided as Multimedia

XSL•FO
RenderX

Appendix 1. Participants were selected using a simple, modified random sampling technique. The inclusion criteria were that participants must be 30 years of age or older, work in an educational or health care organization with distinct management roles, and work in an educational or health care organization in the Kingdom of Bahrain or in the Kingdom of Saudi Arabia.

The sample (N=27) comprised 7 top leaders (CEO, dean, or vice dean), 10 middle managers (heads of the clinical departments in a hospital setting and heads of departments in an academic setting), and 10 employees (working under middle managers). In the study, 37% (n=10) of the participants were in the 31 to 40 years age group, 51% (n=14) were in the 41 to 50 years age group, 7% (n=2) were in the 51 to 60 years age group, and 3% (n=1) were in the 61 to 70 years age group. Therefore, the data obtained largely reflect the perceptions of managers who are between 41 and 50 years of age. The majority of middle managers were within the 35 to 50 years age group (n=4, 15%). A total of 62% (n=17) of the sample were male participants, and 37% (n=10) were female participants. The mean years of experience in the present position of the sample related to top leadership, middle, and lower management in health care was 6 (SD 4.2) years.

Step 2: Thematic Analysis and Converting Qualitative Data to Scale Items

The data from step 1 were analyzed using an emergent thematic analysis approach to identify competency and characteristic scale items. The competency scale items were referred to as competency measure points (CMPs). A team of researchers, each with expertise in qualitative research and leadership competencies, independently coded the interview transcripts to identify initial themes and subthemes. Each researcher was assigned a distinct portion of the data to code, ensuring various perspectives and reducing the risk of bias in the analysis. The coding process followed a structured procedure, with each researcher highlighting the key excerpts from the interviews that represented relevant themes or subthemes.

After initial coding, the researchers met to compare and discuss their findings. A series of team discussions was held to review the codes and the themes that emerged from them. Any discrepancies or disagreements between the coders were resolved through team discussions, in which the team aimed to reach a consensus on the themes and subthemes. This process was critical to ensure the reliability and consistency of the themes identified. Thematic saturation was monitored throughout the qualitative analysis. Saturation was considered to be reached when no new themes or meaningful variations in existing themes emerged from subsequent interviews. During coding, the research team engaged in regular discussions to reflect on how their professional backgrounds in health care leadership and medical education might influence the interpretation of the data, ensuring themes remained grounded in participant perspectives.

The final themes were refined through iterative discussions and re-examination of the data until all researchers agreed on the final list of themes. These final refined themes and subthemes were then converted into scale items. The scale items were framed as clear, concise questions that reflected the competencies identified in the thematic analysis.

The scale items were developed using a 5-point Likert scale to measure the level of importance (1=not at all important, 2=less important, 3=neutral, 4=important, and 5=very important) [7]. To ensure validity, the scale items were reviewed for clarity and relevance, and any ambiguous items were modified.

Step 3: Content Validation

The face validity and content validity of the scale items were assessed by sending the newly developed scale to 6 experts for an independent review. These experts were senior professionals in health care management and leadership, with extensive experience in research and scale development. Their expertise included organizational leadership, hospital administration, medical education, and human resource management. The items were presented as a binomial scale with 2 options to denote whether the corresponding item could be included in the scale or not: (1) “Yes” or “No” and (2) “Agree” or “Disagree.” Any disagreements among the experts were resolved through a follow-up discussion, during which the team of researchers reviewed the feedback provided and addressed any concerns raised by the experts. A consensus approach was used to determine which scale items to retain; items were selected for inclusion in the scale if 4 experts out of 5 (80%) or 5 experts out of 6 (>83%) agreed on their inclusion [9]. In cases where further clarification was needed, additional rounds of discussion were held to ensure the items’ appropriateness. The newly developed and validated scale is presented in Multimedia Appendices 2 and 3, respectively. A total of 19 CMPs and 16 characteristic scale items were developed.

Step 4: New Scale Administration

In the Kingdom of Bahrain, 2 medical universities and 3 main public hospitals were approached by the principal investigator through email and telephonic conversations. Similarly, 10 of the largest universities and hospitals were approached through email and telephonic conversations in the Kingdom of Saudi Arabia. After obtaining consent from the universities and hospitals, the developed scale items (Multimedia Appendices 2 and 3, respectively) were administered to willing participants. To ensure that the scale could capture diverse perspectives, demographic variables such as age, gender, nationality, and origin were included in the survey. These variables were collected to explore potential variations in responses based on different participant characteristics. Specifically, we analyzed whether there were any significant differences in the important ratings of leadership and managerial competencies across various demographic groups. The survey was sent to 500 individuals, with 202 responding to the survey. This response rate was adequate as the target number of respondents was 150. Responses marked as “6=unable to assess” were treated as missing values and excluded from statistical analyses. The data from the survey were analyzed using SPSS version 19 [11].

Step 5: Quantitative Validation (Reliability and Construct Validity)

Exploratory factor analysis (EFA) and confirmatory factor analysis (CFA) were conducted only on the CMPs, which

XSL•FO
RenderX

represent core competency constructs. Characteristic items, which reflect personal and interpersonal qualities, were retained for descriptive evaluation but were not included in the factor analysis. Reliability is a measure of the internal consistency of the measuring instrument [4]. Cronbach α was used to check the internal consistency of the instrument by calculating the α coefficient [10]. Construct validity was confirmed using EFA and CFA, ensuring that the identified CMPs accurately reflected distinct competency dimensions. EFA using principal components extraction with varimax rotation was conducted to identify the underlying competency dimensions represented in the scale.

Ethical Considerations

The research protocol, including both qualitative and quantitative data collection, was approved by the review boards of the participating hospitals and medical educational settings in the Kingdoms of Saudi Arabia (REC-HSD-B-1 - 2020) and Bahrain (E025-PI-12/20). All ethical principles required for studies involving human participants were upheld. All participants signed informed consent before participating in the study. All procedures followed were in accordance with the ethical standards of the responsible committees on human experimentation (institutional and national) and with the Declaration of Helsinki of the World Medical Association.

Informed consent was obtained from all individual participants involved in the study. Participants were fully informed of the nature and possible consequences of the research before providing their consent.

To protect participant privacy and confidentiality, all data were anonymized and deidentified prior to analysis. No identifying information (eg, names, hospital numbers, or direct identifiers) was collected or reported. Access to raw data was restricted to the research team and stored on secure servers in compliance with local data protection regulations. No compensation was provided to participants for taking part in this study.

Results

Participant Demographics

A total of 27 leaders and managers were interviewed in the step 1 survey, and 202 participants were involved in the step 4 survey. Most of the qualitative data obtained were from middle and lower management staff, reflecting their key role in the research focus.

Qualitative Findings: Key Themes and Subthemes

Seven key themes representing competencies and skills that are critical for middle managers emerged from the thematic analysis (step 2) of the qualitative data: personality, agility, attitude, managerial skills, work ethics, mental ability, and interaction. Each of these key themes was further refined into subthemes, providing a deeper understanding of the competencies required.

- Personality: self-confidence, emotional stability, and adaptability
- Agility: decision-making under pressure, responsiveness to change, and strategic flexibility
- Attitude: proactiveness, resilience, and willingness to learn
- Managerial skills: delegation, leadership style, and communication effectiveness
- Work ethics: accountability, integrity, and commitment
- Mental ability: critical thinking, problem-solving, and analytical reasoning
- Interaction: team collaboration, negotiation skills, and conflict resolution

Through EFA (step 5), 3 interrelated constructs were identified that encapsulate the core competencies required for middle management in the Gulf region:

- Team management and adaptation
- Professionalism and problem-solving
- Time management and expertise

Quantitative Findings

Based on the findings from the qualitative analysis and subsequent review by the expert panel, a set of survey items was developed for measuring leadership and managerial competencies of middle managers in the Gulf region (Kingdom of Bahrain and Kingdom of Saudi Arabia), as shown in Multimedia Appendices 2 and 3. The survey was distributed to 500 individuals, and 202 responses were received, resulting in a 40.4% response rate, which exceeded the target of 150 respondents. Descriptive statistics, including the percentage of responses, mean and SD, and measures of skewness and kurtosis, were computed for each item in the scale.

Exploratory Factor Analysis

A principal component analysis with varimax rotation for the CMPs was carried out using the original 19 items in Multimedia Appendix 2. The determination of the number of factors to be extracted depends on how strongly and cleanly the variables load on the factors. The variable will load strongly on a particular factor if the loading is greater than 0.40, and it is considered clean if the absolute difference between the loadings is greater than 0.20. Factor analysis results for the competency scale demonstrated that CMP11 (“Should be an effective communicator”) and CMP13 (“Should be organized in thoughts, words, and deeds”) were removed due to low factor loadings (<0.40) and poor alignment with other items, resulting in a more robust and interpretable factor structure. Therefore, CMP11 and CMP13 were deleted from the model (Table 1). Varimax rotation was applied to achieve a clearer factor structure by maximizing the variance of factor loadings. Although moderate correlations between factors were observed, the orthogonal rotation facilitated the identification of distinct competency domains during the exploratory stage of analysis.

XSL•FO
RenderX

Table . Exploratory factor analysis of competency scale item[a].

Serial number	Scale	Item	Factor 1	Factor 2	Factor 3	Commonalities
1	Should be able to work in a team	CMP1[b]	—[c]	0.776	—	0.64
2	Should be able to work with multiple teams	CMP10	0.569	—	—	0.64
3	Should be an effective listener	CMP12	0.618	—	—	0.69
4	Should demonstrate integrity	CMP14	0.841	—	—	0.75
5	Should display effective organizational skills	CMP15	0.761	—	—	0.71
6	Should follow democratic ways	CMP16	0.545	—	—	0.57
7	Should have long-term vision for the organization	CMP17	0.754	—	—	0.59
8	Should possess leadership skills	CMP18	0.788	—	—	0.64
9	Should possess problem-solving skills	CMP19	0.681	—	—	0.72
10	Able to adapt to changes	CMP2	—	0.694	—	0.71
11	Should be able to achieve tasks as per timeline	CMP3	—	—	0.770	0.63
12	Possess up-to-date knowledge in the respective field	CMP4	—	—	0.590	0.63
13	Should address any conflict of interest effectively	CMP5	—	—	0.573	0.61
14	Should be able to balance various roles (multitasking)	CMP6	—	0.550	—	0.69
15	Should be able to encourage teamwork	CMP7	—	0.773	—	0.68
16	Should be able to manage resources (money or manpower)	CMP8	—	0.728	—	0.72
17	Should be able to take appropriate decisions	CMP9	0.662	—	—	0.72

[a]Kaiser-Meyer-Olkin measure and Bartlett's test of sphericity. Kaiser-Meyer-Olkin measure of sampling adequacy is 0.96 in this study. Approximate $^2_{136}$=2905.11 and significance at 0.

[b]CMP: competency measure point.

[c]Not applicable.

The Bartlett test of sphericity in EFA evaluated all the factors together and each factor separately against a hypothesis stating that there are no factors. The Bartlett test of sphericity in this study was significant (P<.001), indicating that enough shared variance is present. The Kaiser-Meyer-Olkin (KMO) is a measure that provides an approach to comparing the zero-order

XSL•FO
RenderX

correlations with the partial correlations between pairs of variables. The KMO in the study model was 0.96. Kaiser (1974) stated that if the KMO is greater than 0.50, it is acceptable. The closer the KMO value is to 1, the better the correlations between the pairs of variables that can be explained by the other variables.

Analysis of moment structure (AMOS) implements the general approach to data analysis known as structural equation modeling. The multivariate normality of the data was examined by conducting normality checks using the AMOS (version 24.0) software. The distribution of the 17 items in this study is acceptable because none of them deviates from normality (Table 1). Cronbach α coefficient for the internal consistency of the total sample was 0.93. The Cronbach α value was found to be 0.94 for factor 1, 0.93 for factor 2, and 0.86 for factor 3.

A structural model was identified for the scale based on 3 interrelated constructs—professionalism and problem-solving, team management and adaptation, and time management and expertise. As stated previously, these 3 factors emerged from the EFA analysis. The level of correlation between the variables was determined, and then the statistical model fit was evaluated for alignment with the actual dataset using a number of "goodness-of-fit" statistics, as presented in Table 2.

Table . Correlation matrix.

Factor	Professionalism and problem-solving	Team management and adaptation	Time management and expertise
Professionalism and problem-solving	1.000	0.791	0.759
Team management and adaptation	0.791	1.000	0.773
Time management and expertise	0.759	0.773	1.000

Confirmatory Factor Analysis

The CFA was carried out on the variance-covariance matrix for the 3-factor model through the AMOS (version 24.0) statistical package. The estimation method used was maximum likelihood. To achieve model identification, the regression coefficients of the error terms over the endogenous variables were fixed to 1. The CFA was carried out to determine whether the hypothesized statistical model fit the actual dataset, and a number of "goodness-of-fit" statistics were used on the 3-factor models derived by means of EFA. The model fits the data adequately, with a goodness of fit index (0.90), comparative fit index (0.90), Tucker-Lewis index (0.96), and root mean square of error approximation (0.06). The raw χ^2 is 116, and the χ^2/df is 1.76, with a P value less than .001.

The final validated scale developed through these processes is termed the Leadership and Managerial Competency Scale for Middle Managers in the Gulf Region (LMCS-MM Gulf Region; Multimedia Appendix 4) and reflects what people in the 3 levels of management (lower, middle, and top) value.

Discussion

Main Findings

The objective of this study was to develop a valid scale for measuring the competencies for successful middle managers in health care and medical education in the Gulf region, particularly in the Kingdom of Saudi Arabia and the Kingdom of Bahrain. Despite the recognized importance of middle management in health care, there has been a notable gap in the availability of validated instruments specifically tailored to the cultural and operational contexts of the Gulf region. This study aimed to fill that void.

Scholars proffer that interviews can facilitate the collection of key information regarding a variable of interest, and that the findings from the analysis of such data in turn inform the survey design and final scale development [52,56]. This study followed the above principle and adopted the 5-step mixed method by Zhou [55] as the conceptual framework for scale development. The framework involves qualitative exploration, scale item identification, content validation, scale data collection, scale administration, and scale validation [55].

In step 1, qualitative interviews examined the scale construct of interest, which is the competency required for successful middle management. Data were collected based on the personal and social contextual and developmental characteristics that are relevant for successful middle managers, providing the foundation for scale item identification and the evidence to support the content validity of the new scale. The competencies explored in the interviews were extracted from the empirical literature and were consistent with the previous frameworks, such as Goleman emotional intelligence model and Bass transformational leadership theory, indicating theoretical convergence. A number of studies identified transformational leadership attributes such as a well-articulated vision, the ability to create a sense of belonging, providing coaching to subordinates, and inspiring or motivating behavioral change in followers as being essential for success as a middle manager [23,42-45]. Competencies such as emotional intelligence, cultural competency, promoting learning cultures, and high emotional intelligence [24] were also identified in the literature and align with the findings from the qualitative interviews conducted in this study. Numerous authors have observed that one of the ways through which strategic priorities and improvements in operational results can be obtained in the health care sector is through leveraging leadership development [6,13,38,57]. Skills and qualities such as leadership, empathy, effective decision-making, communication, collaboration, and community leadership are emphasized in medical education with a view to instilling these values into future medical professionals [27,53,56,58]. These findings reinforce the

XSL•FO
RenderX

construct validity of the scale and reflect cross-contextual alignment with widely accepted competency frameworks.

In step 2, themes were converted to codes and then to scales, transforming the qualitative data into measurable items on ordinal scales that would allow the use of CFA for data analysis. In step 3, a review of the scale items for content validity was done using debriefing and panel review. The expert panel selected 19 items that should be included in the survey. In step 4, the new instrument was administered to a large representative sample as a quantitative survey.

In step 5, evidence of construct-based validity of the scale items was established through a series of statistical tests measuring scale validity and reliability. Scale validity involves testing the new scale to ascertain that it indeed measures the construct that it is designed to measure (valid scale), and scale reliability means that the scale measures the construct in a precise and consistent manner [54,59]. A scale can be valid but not reliable, and vice versa. Therefore, both measures of quality are important aspects in the scale development process to ensure the accuracy and adequacy of the scale [60,61].

Reliable scales yield the same results every time the underlying condition is the same. Thus, reliability indicates that the scale is consistent [55,62]. Internal consistency reliability was measured for the new scale and focuses on the consistency between different items. It indicates the extent to which participants in the study rated the multiple-item construct measure administered to them in a similar manner [51,60]. For this scale, reliability was estimated in terms of average item-to-total correlation, and Cronbach α was calculated. Cronbach α is a measure of reliability that measures the strength of consistency between different items. The resulting reliability coefficient α has a value ranging from 0 to 1. If scale items are not related or correlated, then α=0.

The factors that emerged from the EFA using principal axis factoring were professionalism and problem-solving, team management and adaptation, and time management and expertise. These resulting constructs from the EFA were validated using CFA. The internal consistency of the factors was measured using composite reliability. Good reliability was obtained as the coefficient values were greater than 0.70.

Assessment of validity using empirical methods yields a type of validity known as criterion-related validity [58,63]. Thus, criterion-related validity was measured for the new scale. In this study, the convergent and discriminant validity of the new scale were also measured. Convergent validity indicates how closely a measure converges or relates to the construct being measured, while discriminant validity shows the opposite effect or discrimination.

Bivariate correlations were computed for this 3-item measure of competencies for the successful middle manager—professionalism and problem-solving, team management and adaptation, and time management and expertise. Adequate discriminant validity was obtained for the scale on the competencies of middle managers, as the average variance extracted was greater than the maximum shared squared variance. For convergent validity, the average variance extracted was 0.50 or higher and lower than the composite reliability. In other words, the variance explained by the construct was greater than the measurement error and greater than cross-loadings. Thus, a validated scale for measuring the competencies required for successful middle managers in the Gulf region emerged from EFA and CFA, with statements related to the characteristics and competencies required for middle managers in the Gulf region.

It is important to highlight that the scale developed under this study (LMCS-MM Gulf Region scale) did not measure the skills and attributes of potential candidates for middle management positions. Rather, the measures reflect what people in the 3 levels of management (lower, middle, and top) value in a (hypothetical) middle manager. Many of the items found in the scale, such as emotional intelligence, transformational leadership (intellectual stimulation and individualized consideration), high-level cognitive skills, and a constellation of other skills (decision-making, communication, problem-solving), already have existing, reliable measurement tools, further strengthening both the reliability and the use of the scale for measuring leadership and managerial competencies. The competencies identified should also be interpreted within the cultural and organizational context of the Gulf region. Middle managers frequently serve as key intermediaries by translating strategic decisions into operational practice within the relatively hierarchical institutional structures. Workplace cultures in the region often emphasize collaboration, interpersonal respect, and relationship-based leadership. These contextual characteristics may have contributed to the prominence of competencies such as teamwork, integrity, adaptability, and effective interpersonal interaction observed in the present scale. The identified competency domains also align with the global health care leadership frameworks, such as the NHS Healthcare Leadership Model, which emphasizes teamwork, professional credibility, ethical leadership, and effective decision-making. This alignment suggests that while leadership competencies may be influenced by local cultural context, many core competencies remain consistent across international health care systems.

Nevertheless, the study has limitations. The sample was restricted to the 2 Gulf countries, which may limit generalizability across the entire Middle East and North Africa region. Additionally, while expert validation was conducted, end-user feedback from actual middle managers in practice was not captured in the final validation stage. These factors may implicate how the scale proves to be universally applicable or context-sensitive. Future research should focus on applying this scale across different health care systems and cultural settings in the wider Arab world to evaluate its external validity. Longitudinal studies could also examine how scores on the scale predict actual performance outcomes or leadership development trajectories in middle managers.

From a practical perspective, organizations can use the LMCS-MM Gulf Region scale for structured recruitment, leadership development, and performance evaluation. Embedding this scale into managerial training programs could help in aligning individual growth with institutional leadership expectations and foster sustainable health care leadership pipelines. As the data were collected using a single self-report

XSL•FO
RenderX

survey instrument, the possibility of common method variance cannot be entirely excluded. Future studies may employ multisource assessments or longitudinal designs to reduce this potential bias and further validate the scale.

Conclusion

This study developed and validated the LMCS-MM Gulf Region scale, offering a robust tool to assess essential competencies in health care and medical education. The scale demonstrates strong psychometric properties and addresses a critical gap in leadership assessment within the Gulf context. The scale measures the perceived importance of leadership and managerial competencies rather than directly assessing managerial performance. As such, it provides a structured framework that may support leadership development, training, and competency-based management initiatives within health care and medical education institutions. Its application can enhance recruitment, training, and performance evaluation of middle managers, thereby supporting the ongoing health system transformations and leadership development initiatives in the region.

Acknowledgments

The authors would like to sincerely thank all the participants of this study. No generative artificial intelligence tools (including but not limited to ChatGPT, DeepSeek, Claude, or similar) were used at any stage in the preparation of this manuscript, including for writing, editing, data analysis, figure creation, or idea generation. All content is original work by the named authors.

Funding

No external funding was received for this study.

Data Availability

The datasets used or analyzed during this study are available from the corresponding author upon reasonable request.

Authors' Contributions

AMAA and ELH conceptualized the study. AMAA did the data collection, data entry, and data analysis. RPS provided intellectual input throughout the study. All authors contributed immensely to the writing and editing of the manuscript. All authors read and approved the final manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
The use of semistructured interviews.
[DOCX File, 13 KB - mededu_v12i1e77476_app1.docx]

Multimedia Appendix 2
Competency scale items developed after content validation.
[DOCX File, 17 KB - mededu_v12i1e77476_app2.docx]

Multimedia Appendix 3
Characteristic scale items developed after content validation.
[DOCX File, 16 KB - mededu_v12i1e77476_app3.docx]

Multimedia Appendix 4
Final scale developed: Leadership and Managerial Competency Scale for Middle Managers in Gulf Region (LMCS-MM Gulf Region).
[DOCX File, 16 KB - mededu_v12i1e77476_app4.docx]

References

1. AL-Eitan L, Sendyani S, Alnemri M. Applications of the one health concept: current status in the Middle East. J Biosaf Biosecur 2023 Mar;5(1):21-31. [doi: 10.1016/j.jobb.2023.01.001]
2. Gurajala S. Healthcare system in the Kingdom of Saudi Arabia: an expat doctor's perspective. Cureus 2023 May;15(5):e38806. [doi: 10.7759/cureus.38806] [Medline: 37303448]

XSL•FO
RenderX

3. Sheikh JI, Cheema S, Chaabna K, Lowenfels AB, Mamtani R. Capacity building in health care professions within the Gulf cooperation council countries: paving the way forward. BMC Med Educ 2019 Mar 14;19(1):83. [doi: 10.1186/s12909-019-1513-2] [Medline: 30871521]
4. Funke A, Wilden R, Gudergan S. Only senior managers lead business model innovation, or do they? Levels of management and dynamic capability deployment. Ind Mark Manag 2023 Oct;114:181-195. [doi: 10.1016/j.indmarman.2023.08.011]
5. Tinline G, Cooper C. The Outstanding Middle Manager: How to Be a Healthy, Happy, High-Performing Mid-Level Manager: Kogan Page Publishers; 2016. URL: https://tinyurl.com/33hy89sb [accessed 2026-04-30]
6. Takyi Mensah E. Changes in Ghana's technical and vocational education and training since its independence: from the perspective of educational policy. Int J Vocat Tech Educ Res 2023 Feb 26;9(2):40-49. [doi: 10.37745/ijvter.15/vol9n24049]
7. Adegbola AE, Adegbola MD, Amajuoyi P, et al. Fostering product development efficiency through cross-functional team leadership: insights and strategies from industry experts. Int J Manag Entrep Res 2024 May 21;6(5):1733-1753. [doi: 10.51594/ijmer.v6i5.1144]
8. Kilag OKT, Manguilimotan AMG, Maraño JC, Jordan RP, Columna PAF, Camaso MFA. A conceptual framework: a systematic literature review on educational leadership and management. Sci Educ 2023;4(9):262-273 [FREE Full text]
9. Schmidt DH, van Dierendonck D, Weber U. The data-driven leader: developing a big data analytics leadership competency framework. J Manag Dev 2023 Jul 19;42(4):297-326. [doi: 10.1108/JMD-12-2022-0306]
10. Kongensgaard R, Skriver C, Gregersen M, Bæksted Jørgensen BK. Can physiotherapists and occupational therapists help mitigate understaffing in nursing care? A qualitative explorative study from Denmark. Nord Sygeplejeforsk 2024 Aug 30;14(4):1-17. [doi: 10.18261/nsf.14.4.2]
11. Kim BY, Park JH, Jo HY, Koo SK, Park MH. Optimized detection of insertions/deletions (INDELs) in whole-exome sequencing data. PLoS ONE 2017;12(8):e0182272. [doi: 10.1371/journal.pone.0182272] [Medline: 28792971]
12. Bolden R, Gosling J, Hawkins B. Exploring Leadership, 2nd edition: Oxford University Press; 2023.
13. Joseph-Richard P, McCray J. Evaluating leadership development in a changing world? Alternative models and approaches for healthcare organisations. Hum Resour Dev Int 2023 Mar 15;26(2):114-150. [doi: 10.1080/13678868.2022.2043085]
14. Lee SY, Matthews M, Jung JY, Kim J, Park H. A comparative study of teachers' beliefs about gifted students' leadership talent. High Ability Studies 2023 Jan 2;34(1):131-158. [doi: 10.1080/13598139.2022.2081964]
15. Kim KN, Yoon SW. Plotting the blank space among leadership, job crafting, and career development: an integrative review and future agendas for HRD. Hum Resour Dev Rev 2025 Mar;24(1):70-92. [doi: 10.1177/15344843241280999]
16. Uhl-Bien M. Complexity leadership and followership: changed leadership in a changed world. J Change Manag 2021 Apr 3;21(2):144-162. [doi: 10.1080/14697017.2021.1917490]
17. Stephen C, Stemshorn B. Leadership, governance and partnerships are essential One Health competencies. One Health 2016 Dec;2:161-163. [doi: 10.1016/j.onehlt.2016.10.002] [Medline: 28616493]
18. Gunawan J, Aungsuroch Y. Managerial competence of first‐line nurse managers: a concept analysis. Int J of Nurs Practice 2017 Feb;23(1):e12502. [doi: 10.1111/ijn.12502]
19. Klingensmith ME. Leadership and followership in surgical education. Am J Surg 2017 Feb;213(2):207-211. [doi: 10.1016/j.amjsurg.2016.07.032] [Medline: 27751527]
20. Islam Pia N, Huuskonen A, Kunnas K, et al. DigiCare model: digitalized healthcare and coaching of patients in an Asian context. : Tampere Universities of Applied Sciences; 2023 URL: https://www.theseus.fi/bitstream/handle/10024/817904/Low-Fidelity_Simulation.pdf?sequence=1 [accessed 2026-05-13]
21. Gill C, Gardner W, Claeys J, Vangronsvelt K. Using theory on authentic leadership to build a strong human resource management system. Hum Resour Manag Rev 2018 Sep;28(3):304-318. [doi: 10.1016/j.hrmr.2018.02.006]
22. Sohn JY, Jang D. Leadership and reciprocity in a leading-by-example game. SSRN. Preprint posted online on May 8, 2023. [doi: 10.2139/ssrn.4441639]
23. Lumban Gaol NT. School leadership in Indonesia: a systematic literature review. Educ Manag Adm Leader 2023 Jul;51(4):831-848. [doi: 10.1177/17411432211010811]
24. Li G, Li L, Xie L, Lopez OS. The effects of ethical leadership on creativity: a conservation of resources perspective. Curr Psychol 2024 Feb;43(6):1-11. [doi: 10.1007/s12144-023-04703-0]
25. Nightingale S, Spiby H, Sheen K, Slade P. The impact of emotional intelligence in health care professionals on caring behaviour towards patients in clinical and long-term care settings: findings from an integrative review. Int J Nurs Stud 2018 Apr;80:106-117. [doi: 10.1016/j.ijnurstu.2018.01.006] [Medline: 29407344]
26. Pesut DJ, Thompson SA. Nursing leadership in academic nursing: the wisdom of development and the development of wisdom. J Prof Nurs 2018;34(2):122-127. [doi: 10.1016/j.profnurs.2017.11.004] [Medline: 29703314]
27. McKimm J, Ramani S, Forrest K, et al. Adaptive leadership during challenging times: effective strategies for health professions educators: AMEE Guide No. 148. Med Teach 2023 Feb;45(2):128-138. [doi: 10.1080/0142159X.2022.2057288] [Medline: 35543323]
28. Berghout MA, Fabbricotti IN, Buljac-Samardžić M, Hilders C. Medical leaders or masters? A systematic review of medical leadership in hospital settings. PLoS ONE 2017;12(9):e0184522. [doi: 10.1371/journal.pone.0184522] [Medline: 28910335]
29. Kouzes JM, Posner BZ. The Student Leadership Challenge: Five Practices for Becoming an Exemplary Leader: John Wiley & Sons; 2024. URL: https://tinyurl.com/furtwfp6 [accessed 2025-03-26]

XSL•FO
RenderX

30. Linville MD, Bates JE. Graduate medical education-accelerated change. Am J Med Sci 2017 Feb;353(2):126-131. [doi: 10.1016/j.amjms.2016.12.001] [Medline: 28183412]
31. Bhat S, Gijo EV, Antony J, Cross J. Strategies for successful deployment and sustainment of Lean Six Sigma in healthcare sector in India: a multi-level perspective. TQM J 2023 Jan 26;35(2):414-445. [doi: 10.1108/TQM-10-2021-0302]
32. Engle RL, Lopez ER, Gormley KE, Chan JA, Charns MP, Lukas CV. What roles do middle managers play in implementation of innovative practices? Health Care Manage Rev 2017;42(1):14-27. [doi: 10.1097/HMR.0000000000000090] [Medline: 26488239]
33. Wilden R, Lin N, Hohberger J, Randhawa K. Selecting innovation projects: do middle and senior managers differ when it comes to radical innovation? J Manage Stud 2023 Nov;60(7):1720-1751. [doi: 10.1111/joms.12874]
34. Farver AR. Value of coaching in building leadership capacity of principals in urban schools: a case study [Master's thesis]. : University of Arkansas; 2014 URL: https://scholarworks.uark.edu/etd/2326/ [accessed 2025-03-26]
35. Chun E, Evans A. The Department Chair as Transformative Diversity Leader: Building Inclusive Learning Environments in Higher Education, 1st edition: Routledge; 2023.
36. Nannings ERH. Middle managers' roles and behaviours during organizational change: a qualitative study among managers and employees [Master's thesis]. : University of Twente; 2017 Nov 5 URL: https://essay.utwente.nl/fileshare/file/72977/Thesis%20Elise%20NanningsFinal_2.pdf [accessed 2025-03-26]
37. Xu JH. Leadership theory in clinical practice. Chin Nurs Res 2017 Dec;4(4):155-157. [doi: 10.1016/j.cnre.2017.10.001]
38. Harris A, Jones M. Middle leaders matter: reflections, recognition, and renaissance. Sch Leadersh Manag 2017 May 27;37(3):213-216. [doi: 10.1080/13632434.2017.1323398]
39. Greilich PE, Phelps ME, Daniel W. Diffusing innovation and best practice in health care. Anesthesiol Clin 2018 Mar;36(1):127-141. [doi: 10.1016/j.anclin.2017.10.009] [Medline: 29425595]
40. Pardazi AM. Mid-level hospital leader descriptions of sensitivity to operations and commitment to resilience [Dissertation]. : Grand Canyon University; 2024 Oct 9 URL: https://search.proquest.com/openview/0f32c0e84334c27e5781b36eadd09ffc/1?pq-origsite=gscholar&cbl=18750&diss=y [accessed 2025-03-26]
41. Döös M, Johansson P, Wilhelmson L. Beyond being present: learning-oriented leadership in the daily work of middle managers. J Workplace Learn 2015 Aug 10;27(6):408-425. [doi: 10.1108/JWL-10-2014-0077]
42. Ratiu L, David OA, Baban A. Developing managerial skills through coaching: efficacy of a cognitive-behavioral coaching program. J Rat-Emo Cognitive-Behav Ther 2017 Mar;35(1):88-110. [doi: 10.1007/s10942-015-0225-8]
43. Antonakis J, House RJ. Instrumental leadership: measurement and extension of transformational–transactional leadership theory. Leadersh Q 2014 Aug;25(4):746-771. [doi: 10.1016/j.leaqua.2014.04.005]
44. Hartley K. Untangling approaches to management and leadership across systems of medical education. BMC Health Serv Res 2016 May 24;16(Suppl 2):180. [doi: 10.1186/s12913-016-1391-9] [Medline: 27230432]
45. Rachmawati AW, Lantu DC. Servant leadership theory development & measurement. Procedia Soc Behav Sci 2014 Feb;115:387-393. [doi: 10.1016/j.sbspro.2014.02.445]
46. Wirtz J, Hofmeister J, Chew PYP, Ding X. Digital service technologies, service robots, AI, and the strategic pathways to cost-effective service excellence. Serv Ind J 2023 Dec 10;43(15-16):1173-1196. [doi: 10.1080/02642069.2023.2226596]
47. Tognazzo A, Gubitta P, Gerli F. Fostering performance through leaders' behavioral competencies: an Italian multi-level mixed-method study. Int J Organ Anal 2017 May 8;25(2):295-311. [doi: 10.1108/IJOA-07-2016-1044]
48. Mohammad J. Towards developing a conceptual framework of Islamic leadership: the role of Taqwa as a moderator. Int J Innovation Business Strategy 2015;3(1) [FREE Full text]
49. Khan A. Islamic leadership principles: a successful model for everyone and all times. : Islamic Research Foundation International, Inc; 2008 URL: https://www.irfi.org/articles/articles_1401_1450/islamic_leadership_principles.htm [accessed 2026-04-30]
50. Ezzani MD, Brooks MC, Yang L, Bloom A. Islamic school leadership and social justice: an international review of the literature. Int J Leadersh Educ 2023 Sep 3;26(5):745-777. [doi: 10.1080/13603124.2021.2009037]
51. Creswell JW, Clark VLP. Designing and Conducting Mixed Methods Research: Sage Publications; 2017. URL: https://collegepublishing.sagepub.com/products/designing-and-conducting-mixed-methods-research-3-241842 [accessed 2026-04-30]
52. Chambliss DF, Schutt RK. Making Sense of the Social World: Methods of Investigation: Sage Publications; 2024. URL: https://tinyurl.com/j6d8py9 [accessed 2026-05-13]
53. Fraser MR, Castrucci BC. Building Strategic Skills for Better Health: A Primer for Public Health Professionals, 1st edition: Oxford University Press; 2023.
54. DeVellis RF, Thorpe CT. Scale Development: Theory and Applications: Sage Publications; 2021. URL: https://tinyurl.com/48fvfxdp [accessed 2026-04-30]
55. Zhou Y. A mixed methods model of scale development and validation analysis. Meas Interdiscip Res Perspect 2019 Jan 2;17(1):38-47. [doi: 10.1080/15366367.2018.1479088]
56. Streiner DL, Norman GR, Cairney J. Health Measurement Scales: A Practical Guide to Their Development and Use: Oxford University Press; 2024. URL: https://academic.oup.com/book/56061 [accessed 2026-04-30]

XSL•FO
RenderX

57. Liu Y, Watson S. Whose leadership role is more substantial for teacher professional collaboration, job satisfaction and organizational commitment: a lens of distributed leadership. Int J Leadersh Educ 2023 Nov 2;26(6):1082-1110. [doi: 10.1080/13603124.2020.1820580]
58. Miller SR, Moore F, Eden L. Ethics and international business research: considerations and best practices. Int Bus Rev 2024 Feb;33(1):102207. [doi: 10.1016/j.ibusrev.2023.102207]
59. Gavrilas L, Kotsis KT. Development and validation of a survey instrument towards attitude, knowledge, and application of educational robotics (AKAER). Int J Res Method Educ 2025 Jan;48(1):44-66. [doi: 10.1080/1743727X.2024.2358780]
60. Matović N, Ovesni K. Interaction of quantitative and qualitative methodology in mixed methods research: integration and/or combination. Int J Soc Res Methodol 2023 Jan 2;26(1):51-65. [doi: 10.1080/13645579.2021.1964857]
61. Love HR, Fettig A, Steed EA. Putting the “mix” in mixed methods: how to integrate quantitative and qualitative research in early childhood special education research. Topics Early Child Spec Educ 2023 Nov;43(3):174-186. [doi: 10.1177/02711214231199268]
62. Denzin NK, Giardina MD, editors. Qualitative Inquiry in Transition—Pasts, Presents, & Futures: A Critical Reader, 1st edition: Routledge; 2024. URL: https://www.routledge.com/Qualitative-Inquiry-in-Transition--Pasts-Presents--Futures-A-Critical-Reader/Denzin-Giardina/p/book/9781032657615 [accessed 2026-04-30] [doi: 10.4324/9781032676067-1]
63. Braun V, Clarke V. (Mis)conceptualising themes, thematic analysis, and other problems with Fugard and Potts’ (2015) sample-size tool for thematic analysis. Int J Soc Res Methodol 2016 Nov;19(6):739-743. [doi: 10.1080/13645579.2016.1195588]

Abbreviations

AMOS: analysis of moment structure
CFA: confirmatory factor analysis
CMP: competency measure point
EFA: exploratory factor analysis
KMO: Kaiser-Meyer-Olkin
LMCS-MM Gulf Region: Leadership and Managerial Competency Scale for Middle Managers in Gulf Region

Edited by LT Car; submitted 14.May.2025; peer-reviewed by MA Zoubi, M Mahboubi, S Karkada, YH Mughal; revised version received 25.Mar.2026; accepted 21.Apr.2026; published 04.Jun.2026.

Please cite as:
Al Ansari AM, Holloway EL, Sequeira RP
Development and Validation of a Scale for Measuring Leadership and Managerial Competencies of Middle Managers in Health Care and Medical Education in the Gulf Region: Cross-Sectional Study
JMIR Med Educ 2026;12:e77476
URL: https://mededu.jmir.org/2026/1/e77476
doi:10.2196/77476

© Ahmed Mohamed Al Ansari, Elizabeth L Holloway, Reginald Paul Sequeira. Originally published in JMIR Medical Education (https://mededu.jmir.org), 4.Jun.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Education, is properly cited. The complete bibliographic information, a link to the original publication on https://mededu.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Profiling Human-AI Regulatory Support in Technology-Enhanced Interprofessional Education Among Health Professions Students: Person-Centered Exploratory Study

Fraide A Ganotice Jr, PhD; John Ian Wilzon T Dizon, MSc; Xiaoai Shen, MSc

Bau Institute of Medical and Health Sciences Education, Li Ka Shing Faculty of Medicine, The University of Hong Kong, 5/F William MW Mong Block, 21 Sassoon Road, Pokfulam, Hong Kong SAR, Hong Kong, Hong Kong SAR, China (Hong Kong)

Corresponding Author:
Fraide A Ganotice Jr, PhD
Bau Institute of Medical and Health Sciences Education, Li Ka Shing Faculty of Medicine, The University of Hong Kong, 5/F William MW Mong Block, 21 Sassoon Road, Pokfulam, Hong Kong SAR, Hong Kong, Hong Kong SAR, China (Hong Kong)

Abstract

Background: Technology-enhanced health care interprofessional education (IPE) places high demands on students' self-regulated learning (SRL) and their ability to work productively with others to prepare them for collaborative practice in health care settings. Yet, little is known about how health professions students perceive and combine their own SRL with coregulation from human and AI-based support in such environments.

Objective: This exploratory and observational study adopted a person-centered approach to identify regulatory profiles based on self-reported SRL and perceived coregulation with near-peer teachers (NPTs) and generative AI (GenAI), and to examine how these profiles were associated with interprofessional learning outcomes.

Methods: Health professions students (N=136) enrolled in a technology-enhanced IPE completed an SRL questionnaire at the beginning of the program. They rated coregulation from NPTs at midprogram and coregulation from GenAI at the end, together with communication, collaboration, and learning satisfaction. A 2-stage exploratory clustering approach using SRL and coregulation scores as indicators was used to identify distinct regulatory profiles. Profile differences in interprofessional communication, collaboration, and learning satisfaction were examined using independent samples 2-tailed *t* tests.

Results: A 2-profile solution provided the best fit to the data. The positive human-NPT–GenAI regulation profile (51/136, 37.5%) was characterized by moderately high SRL, high coregulation with NPTs, and above-average coregulation with GenAI. The negative human-NPT–GenAI regulation profile (85/136, 62.5%) showed the opposite pattern, with moderately low SRL, low coregulation with NPTs, and below-average coregulation with GenAI. Students in the positive profile reported significantly higher interprofessional communication (P=.005) and collaboration (P=.006) and greater learning satisfaction (P<.001) than those in the negative profile, with medium to medium-large effect sizes.

Conclusions: These findings provide person-centered evidence that self-regulation and coregulation from human and AI agents form distinct perceived regulatory configurations in technology-enhanced IPE and that a richer "regulatory ecology," combining stronger SRL with greater perceived coregulation from near-peer and GenAI support, was associated with more favorable interprofessional outcomes. The study highlights the importance of deliberately designing near-peer teaching and GenAI-supported activities as complementary regulatory scaffolds in health professions education.

(JMIR Med Educ 2026;12:e95772) doi:10.2196/95772

KEYWORDS

self-regulated learning; coregulation; generative AI; cluster analysis; interprofessional education; health professions

Introduction

Background

To set the stage for recognizing the importance of self-regulation and coregulation in technology-enhanced interprofessional learning, consider the following scenario. Before a near-peer–led, case-based tutorial, Maria, a health professions student, reads the patient scenario, reviews the learning objectives, and sets personal goals for the session. During the tutorial, she and her peers work with a near-peer teacher (NPT) who helps them clarify priorities, prompts them to explain their reasoning, and suggests strategies for approaching similar cases. Maria uses a checklist to follow the discussion, asks focused questions, and briefly consults a generative AI (GenAI) tool to explore alternative diagnoses and management plans. After the session, she reviews her notes and the GenAI-generated suggestions, identifies any remaining points of confusion, and adjusts her study plan.

XSL•FO
RenderX

This pattern of planning, strategy use, monitoring, and reflection is termed self-regulated learning (SRL): learner's proactive, cyclical control over their cognition, motivation, behavior, and environment to achieve learning goals [1,2]. Yet Maria does not regulate alone. Her NPT and the GenAI tool each serve as coregulatory agents that temporarily scaffold her goal-setting, monitoring, and strategy adaptation [3]. Although interprofessional education (IPE) is widely regarded as essential for preparing students to collaborate across disciplinary boundaries [4], it remains unclear how students combine their own SRL with coregulation from these human and AI-based supports, or whether different combinations carry implications for learning outcomes.

In health professions education, higher levels of SRL have been linked not only to better academic performance but also to more effective clinical reasoning and stronger orientations toward lifelong learning [5-7]. Technology-enhanced learning environments, including blended formats, simulation platforms, and GenAI tools, are now increasingly central to health professions curricula [8-10], and research has shown that students' SRL is a critical determinant of success in such settings [11,12]. Theoretical advances have extended individual SRL models to encompass coregulation, the transitory process by which a more knowledgeable agent guides a learner's planning, monitoring, or reflection, and have situated both constructs within collaborative learning environments [3,13]. Near-peer teaching has demonstrated cognitive and affective benefits for health professions learners [4,14], and GenAI tools now offer an always-available source of guidance that can explain concepts, suggest plans, and critique answers on demand [9,15].

Despite these advances, 3 gaps persist. First, most SRL research uses variable-centered methods that estimate average effects across samples, potentially obscuring distinct subgroups of learners who configure their regulatory resources differently [16,17]; person-centered methods have rarely been applied in IPE settings (confer [18,19]). Second, coregulation has been examined almost exclusively in relation to human agents; how students engage with GenAI as a coregulatory support remains empirically unexamined. Third, no study has jointly profiled SRL alongside coregulation from both a human and an AI-based source to determine whether distinct configurations exist and whether they relate to educational outcomes.

To address these gaps, the present study adopted a person-centered approach to identify distinct regulatory profiles among health professions students enrolled in a technology-enhanced IPE program that incorporated both near-peer teaching and GenAI tools. Profiles were derived from students' self-reported SRL and their perceived coregulation from NPTs and GenAI across the program, and were then compared on interprofessional communication, collaboration, and learning satisfaction. In doing so, this study extends the near-peer teaching literature by reframing NPTs as coregulatory agents within an SRL framework and advances the emerging GenAI literature by examining GenAI not in isolation but as part of a coordinated human-AI regulatory ecology in technology-enhanced IPE.

SRL-Coregulation Framework

We draw on multilevel regulation of learning perspectives [3,13], which distinguish between regulation by the individual learner (self-regulation) and regulation that is supported by others or tools (coregulation). Although these perspectives also describe socially shared regulation of learning (group-level regulation enacted jointly by a team), socially shared regulation is not the focus of the present study and was not measured.

Coregulation describes episodes in which another person or tool deliberately scaffolds a learner's regulation, for example, by clarifying goals, proposing strategies, prompting monitoring, or supporting emotion regulation. In this study, we focus on NPTs and GenAI as 2 complementary coregulatory agents within the same regulatory cycle: near-peers provide human scaffolding in small-group-learning, while GenAI offers on-demand, tool-based guidance.

This SRL-coregulation perspective is grounded in Winne and Hadwin [20] cyclical model of SRL, in which learners move through the phases of task definition, goal setting and planning, enactment, and adaptation. Across these phases, regulation can target cognition (eg, strategy choice), motivation (eg, sustaining effort), emotions (eg, managing frustration), and behavior (eg, time management). Coregulation operates within the same cycle: effective coregulation provides temporary scaffolds that can strengthen learners' SRL over time.

Within this framework, baseline SRL is conceptualized as a relatively stable individual resource that students bring into the IPE program. Coregulation from NPTs (mid of the program) and from GenAI (end of the program) represents external regulatory supports embedded in the same instructional context. This lens allows these to be treated as components of a single regulatory system, and the resulting profiles (eg, high SRL or high coregulation; low SRL or low coregulation) to be interpreted as different configurations of individual and distributed regulatory capacity.

From Variable-Centered to Person-Centered Regulation

Most empirical work on SRL, coregulation, and technology has adopted variable-centered approaches, correlating average SRL scores or perceived instructional support with outcomes such as grades, satisfaction, or perceived competence [5,7]. These studies show that, on average, stronger SRL and higher-quality support are beneficial. Yet, variable-centered analyses assume that the same relationships hold for all students and cannot reveal whether there are qualitatively distinct subgroups who combine SRL and coregulation in different ways, for example, highly self-regulated students who seldom seek help vs low SRL students who rely heavily on external scaffolding.

Person-centered approaches, such as cluster analysis and latent profile analysis, offer a complementary lens. They group individuals based on patterns of scores across multiple variables, thereby identifying naturally occurring profiles rather than relying solely on averages. In motivational and emotional research, person-centered analyses have revealed profiles such as "high positive or low negative emotions" or "high autonomous or low controlled motivation," which are

XSL•FO
RenderX

differentially associated with engagement and achievement [16,17,21]. In digital learning contexts, students' SRL strategies and behaviors in technology-rich environments also cluster into distinct profiles linked to persistence and achievement [11,22]. However, very few studies have used cluster analysis to examine regulatory profiles that explicitly combine SRL with coregulation, and, to our knowledge, none has derived SRL-coregulation profiles where coregulation is provided by both human NPTs and GenAI within an authentic educational setting.

The Present Study

This study had 2 primary goals. First, we sought to identify regulatory profiles of health professions students in a technology-enhanced IPE, based on their baseline SRL, perceived coregulation from NPTs, and perceived coregulation from GenAI. Because this is one of the earliest person-centered investigations that jointly model human and AI-based coregulation within an SRL-coregulation framework, we did not specify a hypothesis about the exact number of profiles. Conceptually, however, extreme combinations can be described in terms of more vs less SRL and more versus less coregulation (eg, "positive" vs "negative" regulatory ecologies).

Importantly, the present study was exploratory and observational. Although the data were collected across 3 time points, students were not randomly assigned to different regulatory supports, and the focal constructs were measured through self-report. Therefore, the study does not test whether NPT or GenAI support causes improvements in communication, collaboration, or satisfaction. Rather, it examines whether distinct perceived self-regulation and coregulation profiles can be identified and whether membership in these profiles is associated with study outcomes.

The first research question (RQ) is as follows: what self-regulation and coregulation profiles emerge among health professions students in a technology-enhanced IPE program, based on SRL and coregulation with NPTs and GenAI?

Second, we examined how profile membership related to interprofessional learning outcomes, operationalized as students' end-of-program ratings of communication and collaboration, and of their learning satisfaction. Prior work has shown that, on average, higher SRL and higher quality instructional support are associated with better collaborative competence and satisfaction. Extending this to a person-centered context, we expected the more "resource-rich" regulatory profiles to show more adaptive outcomes.

The second RQ is as follows: do self-regulation and coregulation profiles differ in interprofessional communication, collaboration, and learning satisfaction at the end of the IPE program?

The hypotheses are as follows:

- H2.1: students in a profile characterized by relatively higher SRL and stronger coregulation with both NPTs and GenAI would report higher interprofessional communication and collaboration than those in a profile characterized by lower SRL and weaker coregulation.
- H2.2: students in the higher regulation profile would also report greater learning satisfaction than those in the lower regulation profile.

Significance

Theoretically, this study extends SRL-coregulation perspectives on multilevel regulation of learning to a technology-enhanced IPE context by conceptualizing NPTs and GenAI as complementary coregulatory agents within a single distributed regulatory system [3,13]. For the educational technology community, this study illustrates how internal SRL and external, digitally mediated supports can be modeled jointly as regulatory ecologies, aligning with ongoing work on learner profiling, learning analytics, and AI-supported regulation of learning (eg, [11,12,22]). The study also has relevance beyond IPE. Many contemporary educational settings, including team-based learning, project-based learning, clinical simulation, online collaborative learning, and AI-supported higher education, require learners to coordinate their own regulation with support from peers, facilitators, digital tools, and, increasingly, GenAI systems.

Practically, it provides an evidence-based way to identify distinct regulatory profiles among health professions students and to link these profiles to interprofessional communication, collaboration, and learning satisfaction, thereby informing the design of targeted supports for vulnerable learners and the deliberate use of near-peer teaching and GenAI as regulatory scaffolds. Methodologically, it introduces a person-centered, multiwave-design using a 2-stage cluster analysis with pre-post outcome measurement and, to our knowledge, is among the first to integrate SRL with coregulation from both human near-peers and GenAI within one IPE program.

Methods

Participants and Context

Participants (N=136) were health professions students from 2 universities in Hong Kong and 1 in Mainland China, all enrolled in a compulsory, cross-institutional IPE stroke rehabilitation simulation hosted by a government-subsidized university in Hong Kong (institution blinded for review). The analytic sample comprised only those with complete data at all 3 time points and included students from Chinese medicine (n=4), medicine (n=13), nursing (n=46), pharmacy (n=18), social work (n=10), physiotherapy (n=30), and speech and hearing sciences (n=15), who worked in mixed-profession teams on technology-enhanced, case-based learning activities. Of these, 80 (58.8%) were female participants, 55 (40.4%) were male participants and one's sex was not reported (0.7%); most were undergraduates in year 2 (n=61, 44.9%), year 3 (n=35, 25.7%), year 4 (n=30, 22.1%), or year 5 (n=1, 0.7%), with an additional 9 (6.6%) master's-level physiotherapy students.

Ethical Considerations

The study formed part of a broader evaluation of the IPE program and was reviewed and approved by the Human Research Ethics Committee of the University of Hong Kong (approval number: EA260174). Participation was voluntary and had no bearing on students' course grades or progression. Data

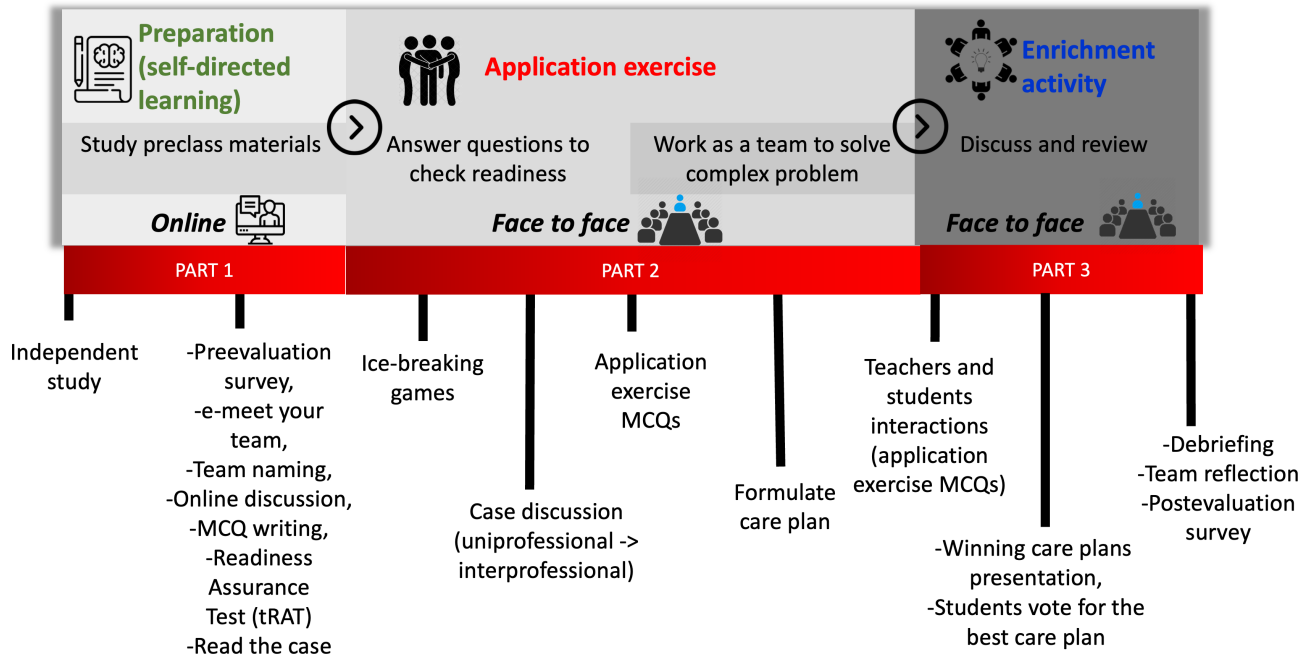
XSL•FO
RenderX

were collected and stored in accordance with institutional data protection policies and the Declaration of Helsinki. All students in the targeted cohorts were invited to participate, and only those who provided informed consent and completed all 3 survey waves were included in the analyses.

IPE Program

The IPE program centered on a stroke rehabilitation simulation designed to support interprofessional communication, role clarification, collaborative care planning, and team-based decision-making. The program ran over 2 months and included 3 parts (Figure 1): part 1 was preparation, part 2 was application exercise, and part 3 was enrichment activity. These activities were intended to help students practice communication and collaboration in a simulated but authentic interprofessional context. Students were allocated to mixed-profession teams so that each team included learners from multiple health and social care disciplines. Across the program, teams worked through a shared patient case, discussed discipline-specific and interprofessional priorities, developed collaborative management plans, and reflected on team processes. The technology-enhanced format included online survey administration, hybrid team activities, and structured opportunities to use GenAI during selected learning tasks.

Figure 1. Learning activities in the interprofessional education (IPE) program. MCQ: multiple choice question.



Procedure

Data were collected at 3 time points aligned with key phases of the IPE program. At the beginning of the program (time 1; before part 1), before any IPE activities, students completed an online survey assessing baseline SRL and demographic or background information. At midprogram (time 2; at the end of part 2), after a month-long block of near-peer-facilitated IPE activities delivered in a hybrid format (both in-person and online), students completed a second survey. During this block, interprofessional teams engaged in core IPE tasks such as team-based care planning and team readiness assurance tests designed to socialize students into interprofessional collaborative practice, particularly by strengthening interprofessional communication and teamwork. Each team was randomly assigned a trained NPT. Each NPT was a senior student (eg, undergraduate year 4 or year 5) who had completed or was completing structured preparation in the Peer Teaching Certificate Programme, which required NPTs to view 4 training videos, participate in peer-teaching activities, and submit a final reflective essay to obtain the certificate. Their role was facilitative rather than primarily didactic: they helped teams clarify task goals, prompted students to explain their reasoning, encouraged participation across professions, monitored group progress, and supported reflection on communication and collaboration. In the time-2 survey, students rated the extent to which their NPT coregulated their learning in these activities.

At the end of the program (time 3; at the end of part 3), following IPE activities that explicitly incorporated GenAI tools, students completed a third survey assessing perceived coregulation from GenAI, interprofessional collaborative competencies, and overall learning satisfaction.

GenAI was introduced as a learning support rather than as an authoritative clinical decision-maker. A university-provided GenAI platform, integrating ChatGPT (OpenAI), DeepSeek (Hangzhou DeepSeek Artificial Intelligence Co, Ltd), and Gemini, was available during these activities, although students were also permitted to use other GenAI tools, where possible, to select tools that they considered capable and appropriate for the learning task. This decision was intended to reflect authentic student use of GenAI in contemporary learning environments, where learners often interact with a range of publicly available or personally accessible tools.

Students were encouraged to use GenAI to clarify concepts in the stroke-rehabilitation case, compare possible care planning options, check the completeness of team reasoning, and support posttask reflection. Students also received general guidance on

XSL•FO
RenderX

the responsible use of GenAI, including how to frame clear prompts, critically evaluate GenAI-generated responses, and compare outputs with course materials, professional knowledge, and team reasoning. However, no fixed or standardized prompt templates were provided. They were also reminded to evaluate GenAI outputs critically and not to treat them as authoritative clinical advice. Accordingly, the GenAI measure in this study captures students' perceived regulatory support from self-selected GenAI tools, rather than exposure to a standardized platform. Actual GenAI use was not objectively tracked; therefore, we did not collect usage logs, record which tools students used, or measure the frequency, duration, or content of their GenAI interactions.

All 3 surveys were administered online via Qualtrics, took approximately 15 to 20 minutes to complete, and included at least one instructed-response item (for example, "Please select 'agree' for this item") as well as system-logged completion time to enhance data quality. Responses were anonymous, and only participants who provided complete data across all 3 time points were included in the present analyses.

Measures

Regulation Measures

All regulation constructs were assessed with 9 items adapted from the Metacognitive Self-Regulation subscale of the Motivated Strategies for Learning Questionnaire (MSLQ [23]). Wording was modified so that items referred either to students' self-regulation, by NPTs, or by GenAI. In all cases, responses were given on a 7-point Likert scale (1="strongly disagree" to 7="strongly agree") and averaged, with higher scores indicating stronger self-regulation or coregulation.

SRL (Time 1)

SRL at time 1 was measured with 9 items capturing goal setting, strategic planning, monitoring of understanding, and reflection (eg, "I ask myself questions to make sure I know the material I have been studying"). Higher scores indicate stronger SRL. Confirmatory factor analysis indicated an acceptable model fit for this scale: incremental fit index (IFI) of 0.980, goodness of fit index (GFI) of 0.956, normed fit index (NFI) of 0.925, root mean square error of approximation (RMSEA) of 0.050, and standardized root mean square residual (SRMR) of 0.054.

Coregulation From NPTs (Time 2)

Perceived coregulation by NPTs at time 2 was assessed with an adapted version of the same item set, reworded so that the NPT was the regulator. Items tapped four domains: (1) task structuring and goal clarification, (2) strategic guidance, (3) monitoring and feedback, and (4) motivational or emotional support (eg, "Our near-peer teacher helped clarify goals for each interprofessional task"). Higher scores reflect stronger perceived coregulation by NPTs. Model fit indices for this scale were IFI of 0.980, GFI of 0.931, NFI of 0.964, RMSEA of 0.096, and SRMR of 0.079.

Coregulation From GenAI (Time 3)

Perceived coregulation by GenAI was measured with parallel items in which the regulator was the AI system rather than the NPT. The items covered the same 4 domains: task structuring and goal clarification, strategic guidance, monitoring and feedback, and motivational or emotional support (eg, "The AI tool helped me plan how to approach each interprofessional task"). Higher scores indicate stronger perceived coregulation by GenAI. Model fit indices for this scale were IFI of 0.986, GFI of 0.948, NFI of 0.965, RMSEA of 0.070, and SRMR of 0.036.

Communication and Collaboration (Time 3)

Communication and collaboration were measured at time 3 using 2 subscales from the Interprofessional Collaborative Competencies Attainment Survey [24]. The communication subscale captured participants' self-assessed abilities in facilitating effective interprofessional dialogue, active listening, expressing ideas constructively, and providing feedback (eg, "Actively listen to IP team members' ideas and concerns"). Responses were given on a 7-point scale (1="strongly disagree" to 7="strongly agree"; 5 items) and were averaged; higher scores indicate greater communication competencies. The collaboration subscale assessed participants' perceived competence in proactively engaging with team members, working cooperatively to enhance care, and learning from interprofessional interactions (eg, "Work effectively with IP team members to enhance care"). Responses were given on a 7-point scale (1="strongly disagree" to 7="strongly agree"; 3 items) and were averaged; higher scores indicate greater collaboration competencies.

Learning Satisfaction (Time 3)

Postprogram learning satisfaction was measured at time 3 using the Learning Satisfaction Questionnaire [25]. Items captured overall satisfaction with the IPE experience, perceived usefulness for future professional practice, and satisfaction with opportunities to develop interprofessional skills (eg, "Overall, I am satisfied with this interprofessional education programme"). Responses were given on a 5-point scale (1="strongly disagree" to 5="strongly agree"; 4 items) and were averaged; higher scores indicate greater satisfaction.

Demographic and Background Variables (Time 1)

At time 1, students reported their health profession program, year of study, sex, and prior IPE experience (eg, none vs some). These variables were used to describe the sample.

Data Analysis

Data analyses were conducted in SPSS (version 28; IBM). We first screened the data for missingness, outliers, and assumptions of normality. Cases with missing data on all 3 profile indicators (SRL, near-peer coregulation, and GenAI coregulation) were removed. Additional cluster-solution robustness checks were conducted in R (version 4.5.3; R Core Team [26]) using the cluster package.

We computed descriptive statistics (means and SDs) for all variables and examined internal consistency reliability (Cronbach α) for each multi-item scale. Confirmatory factor analyses were conducted to evaluate the measurement properties of the SRL, coregulation with NPT, and coregulation with GenAI. Model fit was judged using standard indices (eg, IFI, GFI, RMSEA, and SRMR). Pearson correlations among SRL,

XSL•FO
RenderX

coregulation indicators, ICCAS scores, satisfaction, and demographic variables were then inspected.

To address the first research question, we conducted a 2-stage exploratory clustering approach, which is appropriate to handle continuous data and ensure statistical robustness [27]. We used time 1 SRL, time 2 coregulation with NPT, and time 3 coregulation with GenAI as continuous profile indicators. Indicators were standardized (as *z* scores) prior to analysis. Initially, to extract the number of clusters from the data, we performed a hierarchical cluster analysis using Ward method and squared Euclidean distance, and we identified 2 clusters based on the resulting dendrogram. Consequently, we ran a k-means cluster analysis to identify the final cluster centers and classify the participant's cluster membership.

We additionally compared the retained 2-cluster solution with 3-cluster and 4-cluster alternatives using the same standardized profile indicators and k-means clustering procedure. Because no single index is definitive in determining the optimal number of clusters, cluster quality was evaluated using average silhouette values, with Calinski-Harabasz values, cluster sizes, interpretability, and parsimony inspected as supplementary considerations.

To address the second research question, we ran an independent 2-tailed *t* test with 1000 bootstrap samples and examined whether there would be significant differences in participants' collaboration and communication outcomes and learning satisfaction based on the 2 cluster profiles yielded. A *P* value less than .05 was considered significant, and Cohen *d* of 0.2 (small effect size), 0.5 (medium effect size), and 0.8 (large effect size) were used as reference points for interpretation.

Results

The descriptive statistics and correlational analyses are presented in Tables 1 and 2. The reliability coefficients (Cronbach α) were acceptable for all variables. Students' self-regulation was positively associated with coregulation from NPTs (r=0.358; P<.001), postintervention communication competencies (r=0.275; P=.001), and postintervention collaboration competencies (r=0.230; P=.007), but was not significantly related to coregulation from GenAI (r=−0.066; P=.44) or learning satisfaction (r=0.127; P=.14). Coregulation from NPTs was positively related to postintervention communication (r=0.175; P=.04), postintervention collaboration (r=0.213; P=.01), and learning satisfaction (r=0.271; P=.001). Similarly, coregulation from GenAI was positively associated with communication (r=0.178; P=.04), collaboration (r=0.175; P=.04), and learning satisfaction (r=0.353; P<.001). Notably, postintervention communication and collaboration competencies were highly intercorrelated (r=0.927; P<.001), and both were strongly associated with learning satisfaction (communication: r=0.624, P<.001; collaboration: r=0.613, P<.001).

Table . Descriptive statistics (N=136).

Variable	Score, mean (SD; range)	Skewness (SE)	Kurtosis (SE)	Cronbach α
Self-regulation	4.90 (0.71; 1 - 7)	0.18 (0.20)	0.95 (0.41)	0.72
NPT[a] coregulation	5.13 (0.80; 1 - 7)	0.38 (0.20)	−0.08 (0.41)	0.73
Gen-AI coregulation	4.44 (0.86; 1 - 7)	−0.27 (0.20)	0.42 (0.41)	0.79
Communication	5.32 (0.98; 1 - 7)	−0.62 (0.20)	2.16 (0.41)	0.96
Collaboration	5.36 (1.03; 1 - 7)	−0.78 (0.20)	1.99 (0.41)	0.94
Learning satisfaction	3.83 (0.65; 1 - 5)	−0.21 (0.20)	0.62 (0.41)	0.88

[a]NPT: near-peer teacher.

XSL•FO
RenderX

Table . Correlational analyses.

Variable	SR[a]	NPTCR[b]	GenAICR[c]	Communication	Collaboration	LS[d]
SR, *r*	—[e]	.36[f]				
NPTCR, *r*	.36[f]	—				
GenAICR, *r*	−.07	.097	—			
Communication, *r*	.28[g]	.18[h]	.18[h]	—		
Collaboration, *r*	.23[g]	.21[h]	.18[h]	.93[f]	—	
LS, *r*	.13	.27[g]	.35[f]	.62[f]	.61[f]	—

[a]SR: self-regulation.

[b]NPTCR: near-peer teacher coregulation.

[c]GenAICR: generative AI coregulation.

[d]LS: learning satisfaction.

[e]Not applicable.

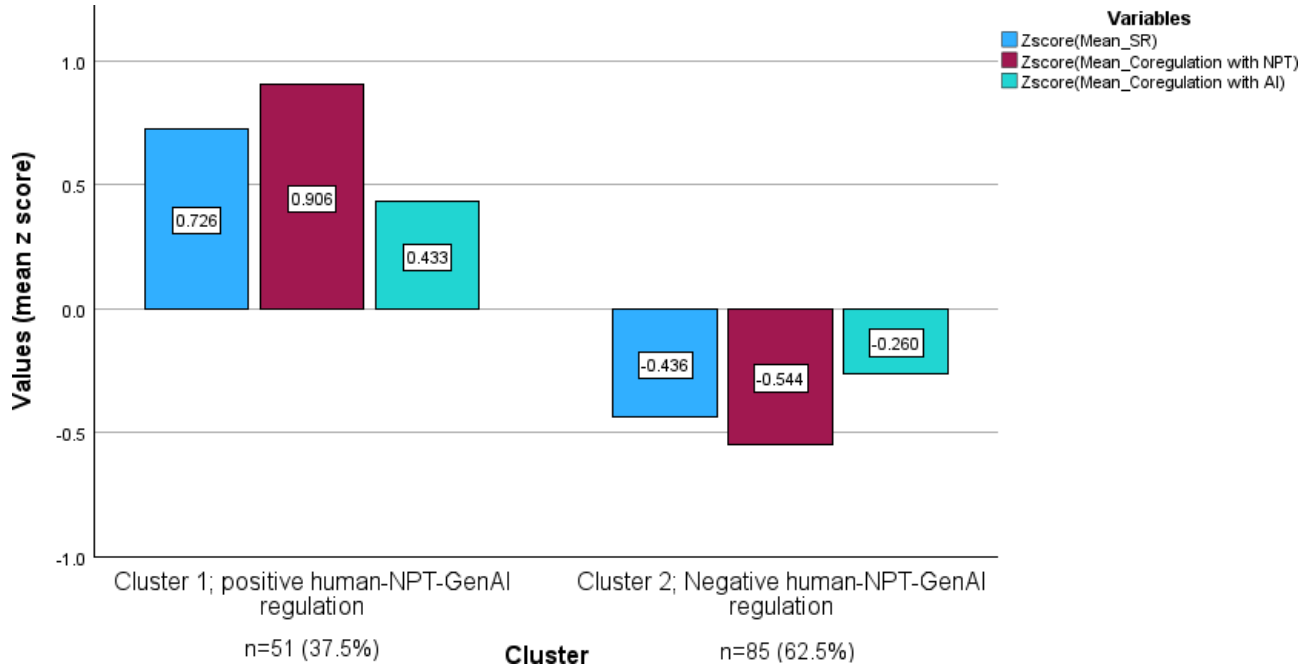
[f]*P*<.001.

[g]*P*<.01.

[h]*P*<.05.

Figure 2 depicts the standardized scores for self-regulation and coregulation with NPTs and GenAI across the 2 profiles. The first profile, labeled positive human-NPT–GenAI regulation (cluster 1; 51/136, 37.5% of participants), was characterized by moderately high self-regulation, high coregulation with NPTs, and above-average coregulation with GenAI. Students in this profile, therefore, appeared to combine stronger personal self-regulatory resources with active engagement with both human and AI-based coregulatory scaffolds.

Figure 2. Final cluster centers for self-regulation and coregulation profiles by human–near-peer teacher (NPT)–generative AI (GenAI) regulation cluster (N=136).



The second profile, labeled negative human-NPT–GenAI regulation (cluster 2; 85/136, 62.5% of participants), showed the opposite pattern. These students reported below-average self-regulation, low coregulation with NPTs, and slightly below-average coregulation with GenAI. This profile thus reflected learners who reported relatively weak internal regulatory resources alongside limited use of both human and AI-based coregulatory supports.

Alternative cluster solutions were compared using the same 3 standardized indicators. Table 3 summarizes the cluster sizes and quality indicators for the 2-cluster solutions, 3-cluster solutions, and 4-cluster solutions. The 2-cluster solution yielded the highest average silhouette value and the strongest Calinski-Harabasz value among the compared solutions. The 3-cluster and 4-cluster alternatives also produced smaller or imbalanced clusters and did not offer a clearer substantive interpretation. Taken together, the quality indicators,

XSL•FO
RenderX

interpretability, and parsimony of the profiles supported retaining the 2-profile solution.

Table . Comparison of alternative cluster solutions.[a]

Solution	Cluster sizes	Average silhouette	Calinski-Harabasz	Interpretation
k=2	85, 51	0.297	60.10	Best quality indicators; retained solution
k=3	75, 51, 10	0.262	43.73	Lower silhouette and Calinski-Harabasz values
k=4	60, 45, 22, 9	0.215	39.61	Lowest silhouette and 1 small cluster

[a]Higher average silhouette and Calinski-Harabasz values indicate stronger relative support among the compared solutions.

We also explored the distribution of the yielded cluster profiles across the participants' sex, discipline, and year level. Table 4 shows that regarding sex, female participants were more represented overall (80/136, 58.8%) and were distributed similarly across both clusters (cluster 1 [positive]: 29/51, 56.9%; cluster 2 [negative]: 51/85, 60%), as were male participants (cluster 1: 21/51, 41.2%; cluster 2: 34/85, 40%). Across disciplines, nursing constituted the largest group in both clusters (cluster 1: 16/51, 31.4%; cluster 2: 30/85, 35.3%), followed by physiotherapy (cluster 1: 12/51, 23.5%; cluster 2: 18/85, 21.2%) and pharmacology and pharmacy (cluster 1: 6/51, 11.8%; cluster 2: 12/85, 14.1%). Notably, all Chinese medicine students were classified in the positive regulation group (cluster 1: n=4/51, 7.8%), while speech and hearing sciences students were more concentrated in the negative regulation group (cluster 2: 12/85, 14.1%) relative to the positive regulation group (cluster 1: 3/51, 5.9%). With respect to year level, year 2 students (combining the half-class and new curriculum cohorts, total n=61) were the largest year-level group overall, though year 2 new curriculum students were notably more represented in the negative regulation group (cluster 2: 24/85, 28.2%) than in the positive regulation group (cluster 1: 7/51, 13.7%). Year 4 students were similarly more prevalent in the negative regulation group (cluster 2: 21/85, 24.7%) than the positive regulation group (cluster 1: 9/51, 17.6%). Overall, the distribution of participants across clusters was broadly comparable by sex, discipline, and year level.

XSL•FO
RenderX

Table . Composition of extracted human–near-peer teacher (NPT)–generative AI (GenAI) regulation clusters by sex, discipline, and grade (year level).

Variable	Cluster 1: positive human-NPT–GenAI regulation (n=51), n (%)	Cluster 2: negative human-NPT–GenAI regulation (n=85), n (%)	Total, N
Sex[a]			
Female	29 (56.9)	51 (60)	80
Male	21 (41.2)	34 (40)	55
Not reported	1 (1.9)	0	1
Total	51 (100)	85 (100)	136
Discipline[b]			
Chinese medicine (Mainland China)	4 (7.8)	0 (0)	4
MBBS	5 (9.8)	8 (9.4)	13
Nursing	16 (31.4)	30 (35.3)	46
Pharmacology and pharmacy	6 (11.8)	12 (14.1)	18
Social work	5 (9.8)	5 (5.9)	10
Physiotherapy	12 (23.5)	18 (21.2)	30
Speech and hearing sciences	3 (5.9)	12 (14.1)	15
Total	51 (100)	85 (100)	136
Grade (year level)			
Year 5	1 (1.9)	0 (0)	1
Year 4	9 (17.6)	21 (24.7)	30
Year 3	15 (29.4)	20 (23.5)	35
Year 2 (half class)	14 (27.5)	16 (18.8)	30
Year 2 (new curriculum)	7 (13.7)	24 (28.2)	31
Masters year 1 (part-time)	5 (9.8)	4 (4.7)	9
Total	51 (100)	85 (100)	136

[a]Sex data were available for 135 participants; 1 participant did not report sex.

[b]All disciplines are based in Hong Kong, unless stated otherwise.

We conducted independent samples 2-tailed *t* tests to examine differences between the positive and negative human-NPT–GenAI regulation groups on 3 interprofessional learning outcomes: communication, collaboration, and learning satisfaction. Note that all means and SDs are reported as standardized (*z* score) values. Across all 3 outcomes, the positive regulation group (cluster 1) scored significantly higher than the negative regulation group (cluster 2; Table 5 and Figure 3). For communication, the positive regulation group (mean 0.31, SD 1.10) outscored the negative regulation group (mean −0.18, SD 0.89; t_{134}=2.85; *P*=.005; mean difference [MD]=0.49, 95% CI 0.12-0.81, *d*=0.51, 95% CI 0.15-0.86). A similar pattern emerged for collaboration, with the positive regulation group (mean 0.30, SD 1.07) scoring higher than the negative regulation group (mean −0.18, SD 0.91; t_{134}=2.81; *P*=.006; MD=0.49, 95% CI 0.13-0.83; *d*=0.50, 95% CI 0.15-0.85). The largest difference was observed for learning satisfaction, where the positive regulation group (mean 0.44, SD 1.04) substantially outscored the negative regulation group (mean −0.27, SD 0.88; t_{134}=4.26; *P*<.001; MD=0.71, 95% CI 0.34-1.04; *d*=0.75, 95% CI 0.39-1.11). Effect sizes ranged from medium (*d*=0.50-0.51) to medium-large (*d*=0.75), and Levene test confirmed equal variances across groups for all variables (all Fs<1.00, all ps>0.30). Bootstrap 95% CI based on 1000 samples corroborated these findings. These results indicate that profile membership was associated with differences in self-reported interprofessional outcomes; however, the observational design does not allow conclusions about whether regulatory profile membership caused these differences.

XSL•FO
RenderX

Table . Independent samples 2-tailed *t* tests comparing cluster 1 (positive human-NPT–GenAI regulation) and cluster 2 (negative human-NPT–GenAI regulation) on communication, collaboration, and learning satisfaction (N=136).[a]

Variable	Cluster 1: positive human-NPT–GenAI regulation (n=51, 37.5%), mean (SD)	Cluster 2: negative human-NPT–GenAI regulation (n=85, 62.5%), mean (SD)	*t* test (*df*)	*P* value	MD[b]	Bootstrap 95% CI[c]	Cohen *d*[d]
Communication	0.31 (1.10)	−0.18 (0.89)	2.85 (134)	.005	0.49	(0.12-0.81)	0.51
Collaboration	0.30 (1.07)	−0.18 (0.91)	2.81 (134)	.006	0.49	(0.13-0.83)	0.50
Learning satisfaction	0.44 (1.04)	−0.27 (0.88)	4.26 (134)	<.001	0.71	(0.34-1.04)	0.75

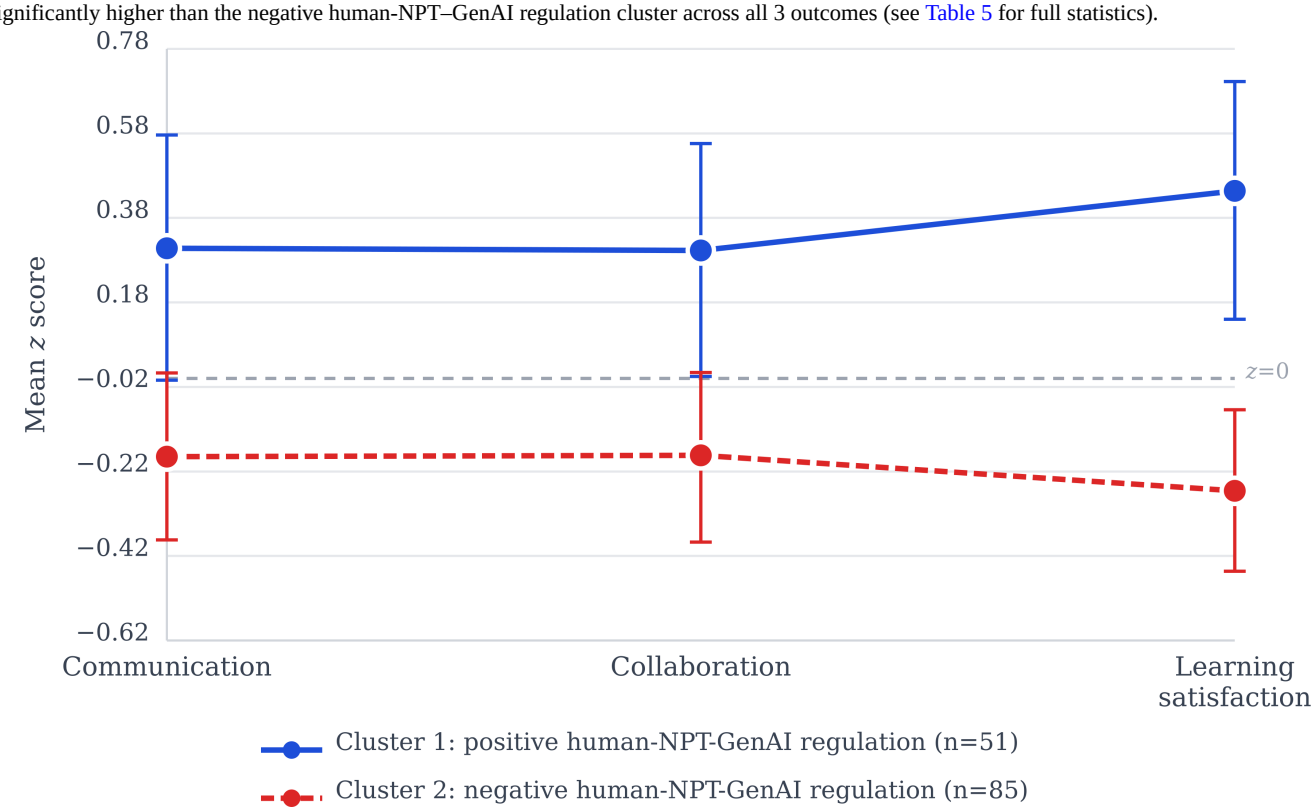
[a]Mean and SD values are standardized (*z* scored). Equal variances were assumed for all variables (Levene test: all $F<1.00$, all $P>.30$). Outcome values are standardized *z* scores and should be interpreted as differences in self-reported outcomes, not as causal effects of profile membership.

[b]MD: mean difference (cluster 1-cluster 2).

[c]Bootstrap 95% CIs are based on 1000 bootstrap samples.

[d]Cohen *d* uses the pooled SD. Benchmarks: |0.2|=small, |0.5|=medium, |0.8|=large [28].

Figure 3. Standardized mean scores (*z* scores) for communication, collaboration, and learning satisfaction by cluster. Values are standardized *z* scores (mean 0, SD 1) computed across the full sample (N=136). Error bars represent bootstrap 95% CI based on 1000 bootstrap samples. The dashed horizontal reference line marks the overall sample mean (*z*=0). The positive human–near-peer teacher (NPT)–generative AI (GenAI) regulation cluster scored significantly higher than the negative human-NPT–GenAI regulation cluster across all 3 outcomes (see Table 5 for full statistics).



Discussion

Principal Findings

This person-centered study examined how health professions students reported combining SRL with perceived coregulation from NPTs and GenAI in a technology-enhanced IPE. Two perceived regulatory profiles emerged and were associated with students' end-of-program self-reported communication, collaboration, and learning satisfaction. In interpreting these findings, it is important to emphasize that the study identifies associations among self-reported constructs rather than causal effects of NPTs or GenAI support.

Regulatory Profiles as Distributed "Regulatory Ecologies"

Addressing RQ1, the 2-stage cluster analysis yielded a positive human-NPT–GenAI regulation profile characterized by moderately high SRL, strong coregulation from NPTs, and above-average coregulation from GenAI, as well as a negative human-NPT–GenAI regulation profile characterized by the opposite pattern. Rather than mixed configurations (eg, high

XSL•FO
RenderX

SRL paired with low coregulation), we observed a "low-on-all vs high-on-all" pattern, suggesting that internal and external regulatory resources tended to accumulate rather than compensate for each other: students who regulated more effectively also reported greater perceived coregulation from human and AI supports, whereas those with weaker SRL reported limited uptake of these supports.

This accumulation pattern aligns with SRL-Coregulation perspectives that view regulation as distributed across individuals, social partners, and tools [3,13], and with person-centered work showing that adaptive resources often cluster into coherent "high-high" and "low-low" configurations [16,17]. Psychologically, the positive profile reflects a productive balance between agency and reliance: students maintain ownership of their goals and decisions yet intentionally draw on NPTs and GenAI to extend their regulatory capacity. Their use of external support is therefore not a sign of dependency but of adaptive help-seeking and willingness to distribute regulatory work when tasks are complex. NPTs and GenAI function as external "mirrors" for thinking, helping students test ideas, surface misconceptions, and maintain a coherent understanding of their team's goals and roles. By contrast, students in the negative profile adopt a more constrained stance, combining weaker SRL with reluctance or difficulty in mobilizing available human and AI supports, leaving them more reactive and potentially more vulnerable during collaboration.

Theoretically, these findings extend SRL-coregulation perspectives by modeling NPTs and GenAI as complementary coregulatory agents within a single distributed regulatory system. Rather than treating regulation as an individual trait, the profiles illustrate how internal strategies and external supports form coherent regulatory ecologies, some richer, some poorer, with differential associations with interprofessional learning.

Profile Differences in Interprofessional Outcomes

Addressing RQ2, students in the positive profile reported significantly higher interprofessional communication and collaboration, and greater learning satisfaction than those in the negative profile, with effect sizes ranging from medium (communication and collaboration, d=0.50-0.51) to medium-large (satisfaction, d=0.75). These findings support H2.1 and H2.2 and suggest that students' self-regulatory and coregulatory configurations are associated with educationally meaningful differences in self-reported outcomes.

The advantage of the positive profile is consistent with evidence that higher SRL is associated with better performance and readiness for lifelong learning in health professions education [5-7] and that high-quality instructional and social supports are associated with collaborative processes and satisfaction [12,29]. Our findings add nuance by showing that it is the combination of stronger SRL with active engagement with NPT and GenAI supports, a richer regulatory ecology, that is associated with more favorable interprofessional outcomes, rather than any single resource in isolation.

Equally important are the weaker outcomes for the negative profile. These students entered with lower SRL and reported limited engagement with both NPT and GenAI coregulation, despite having equivalent access to these supports. Their lower communication, collaboration, and satisfaction echo work showing that learners with weaker SRL and less use of external supports are at risk in complex, self-directed tasks [8]. For IPE specifically, this suggests that regulatory vulnerabilities may undermine the development of collaborative competence even when institutional resources are available.

The pattern for GenAI deserves particular attention. Students in the positive profile reported stronger perceived coregulation from GenAI across planning, conceptual clarification, monitoring, and reflection, and this was associated with better outcomes. This offers early empirical support for the idea that large language models may be perceived by students as coregulatory partners when students engage with them productively [9,15]. At the same time, the sizable negative profile indicates that simply providing access to GenAI and NPT facilitators is insufficient; students with weaker SRL may not spontaneously perceive or use these supports as regulatory resources.

Implications for Designing Human-AI Coregulation in and Beyond IPE

Theoretically, the findings support a distributed view of regulation in which learners' SRL is situated within a broader ecology of human and technological supports. In this ecology, NPTs and GenAI may serve different but complementary regulatory functions. NPTs can provide socially responsive scaffolding, monitor group interaction, prompt participation across professions, and support emotional or motivational regulation. GenAI, by contrast, can provide on-demand prompts, alternative explanations, planning support, and opportunities for reflection. The value of a human-AI regulatory ecology may therefore lie not in replacing human facilitation with AI but in coordinating the strengths of both forms of support.

Practically, SRL and coregulation should be treated as explicit design targets in technology-enhanced IPE. Educators should design NPTs and GenAI supports as explicit regulatory scaffolds, rather than assuming students will use them productively on their own. NPTs and GenAI can be positioned as "regulation partners" whose roles are aligned with key SRL processes. For example, near-peers can guide planning and monitoring, while GenAI prompts can be structured to support checking assumptions, exploring alternative management plans, and promoting reflective practice. Making these roles explicit in instructions and debriefings may help students, especially those in the negative profile, use available supports more effectively.

The implications also extend beyond IPE. In project-based learning, team-based Science, Technology, Engineering, and Mathematics courses, teacher education, clinical simulation, and online collaborative learning, students increasingly work with both human facilitators and AI tools. The present findings suggest that simply providing access to these supports may be insufficient, particularly for learners who report weaker SRL. Learners may need explicit instruction in how to seek help,

XSL•FO
RenderX

evaluate AI responses, use prompts strategically, and integrate feedback from human and AI sources. Thus, human-AI learning environments should include not only access to technology but also regulatory routines, prompt guidance, reflection templates, and opportunities for facilitated debriefing.

The identification of a vulnerable negative profile also suggests potential opportunities for targeted intervention. These students may benefit from early SRL workshops (eg, on goal setting, strategic planning, and structured reflection) combined with more structured requirements to engage with near-peers and GenAI (eg, mandatory planning templates cocompleted with NPTs, guided GenAI prompts embedded in cases). Conversely, students in the positive profile may require less basic scaffolding and could be challenged to take on leadership roles or to critically interrogate GenAI suggestions.

Brief profile screening early in an IPE program, using short SRL questionnaires and early perceived-support or help-seeking questionnaires, could help educators identify students who may be at risk of remaining in a low-regulation, low-support pattern. In resource-constrained settings, such information could guide the allocation of mentoring and technological resources to those most likely to benefit.

Limitations and Future Directions

Several limitations should be noted. First, the modest sample size (N=136) may limit statistical power, the stability of the person-centered profiles, and transferability to other IPE contexts. Although the 2-profile solution was interpretable and supported by cluster diagnostics, person-centered results can be sensitive to sample composition and analytic decisions. Second, all focal constructs were assessed via self-report. This is a substantive limitation because SRL, coregulation from NPTs, coregulation from GenAI, communication, collaboration, and satisfaction were all reported by the same respondents. As a result, the observed associations may partly reflect shared method variance, general positive response tendencies, or students' overall satisfaction with the program rather than distinct regulatory processes. The very high correlation between communication and collaboration also suggests that students may not have clearly differentiated these competencies in their self-assessments. Future work should incorporate observational data, peer and NPT ratings, performance-based assessments, video-coded team interaction, and learning analytics to capture coregulation processes more directly. Third, GenAI use was self-selected and was not objectively tracked. We did not record which tools students used, how often they used them, or the content and quality of their prompts and GenAI-generated responses. Therefore, low GenAI coregulation scores may reflect nonuse, limited access, weaker prompting, or low perceived usefulness, and the findings cannot be attributed to any particular GenAI platform. Fourth, the design was not experimental, and the person-centered analysis involved analytic choices (eg, number of clusters) that may influence results. Experimental or quasi-experimental designs that manipulate the availability or structure of near-peer and GenAI support, combined with longitudinal profiling, would provide stronger evidence about how regulatory ecologies develop and can be reshaped.

Conclusions

This study suggests that health professions students in a technology-enhanced IPE program formed distinct perceived regulatory configurations, differing jointly in SRL and coregulation from NPTs and GenAI, and that these configurations were associated with different interprofessional outcomes. These findings suggest that students' perceived regulatory ecologies are associated with self-reported interprofessional learning outcomes. For educators, the key message is that NPT facilitation and GenAI tools should be designed as coordinated regulatory supports rather than offered as isolated enhancements, with particular attention to scaffolding students who do not spontaneously engage with these resources. For researchers, the findings point to the value of person-centered, multimethod, and longitudinal designs that trace how human-AI regulatory ecologies evolve and how they can be reshaped to foster more adaptive self-regulation and coregulation in collaborative professional learning.

Acknowledgments

The authors declare the use of GPT-5.5 (OpenAI) for overall grammar-checking of the manuscript and for reference management cross-checking purposes.

Funding

The authors declared no financial support was received for this work.

Conflicts of Interest

None declared.

References

1. Panadero E. A review of self-regulated learning: six models and four directions for research. Front Psychol 2017;8:422. [doi: 10.3389/fpsyg.2017.00422] [Medline: 28503157]
2. Zimmerman BJ. Becoming a self-regulated learner: an overview. Theory Pract 2002 May;41(2):64-70. [doi: 10.1207/s15430421tip4102_2]

XSL•FO
RenderX

3. Hadwin AF, Järvelä S, Miller M. Self-regulated, co-regulated, and socially shared regulation of learning. In: Zimmerman BJ, Schunk DH, editors. Handbook of Self-Regulation of Learning and Performance: Routledge; 2011:65-84.
4. Reeves S, Fletcher S, Barr H, et al. A BEME systematic review of the effects of interprofessional education: BEME Guide No. 39. Med Teach 2016 Jul;38(7):656-668. [doi: 10.3109/0142159X.2016.1173663] [Medline: 27146438]
5. Artino AR Jr, Stephens JM. Academic motivation and self-regulation: a comparative analysis of undergraduate and graduate students learning online. Internet High Educ 2009 Dec;12(3-4):146-151. [doi: 10.1016/j.iheduc.2009.02.001]
6. Brydges R, Butler D. A reflective analysis of medical education research on self-regulation in learning and practice. Med Educ 2012 Jan;46(1):71-79. [doi: 10.1111/j.1365-2923.2011.04100.x] [Medline: 22150198]
7. van Houten-Schat MA, Berkhout JJ, van Dijk N, Endedijk MD, Jaarsma ADC, Diemers AD. Self-regulated learning in the clinical context: a systematic review. Med Educ 2018 Oct;52(10):1008-1015. [doi: 10.1111/medu.13615] [Medline: 29943415]
8. Broadbent J, Poon WL. Self-regulated learning strategies & academic achievement in online higher education learning environments: a systematic review. Internet High Educ 2015 Oct;27:1-13. [doi: 10.1016/j.iheduc.2015.04.007]
9. Kasneci E, Sessler K, Küchemann S, et al. ChatGPT for good? On opportunities and challenges of large language models for education. Learn Individ Differ 2023 Apr;103:102274. [doi: 10.1016/j.lindif.2023.102274]
10. Zawacki-Richter O, Marín VI, Bond M, Gouverneur F. Systematic review of research on artificial intelligence applications in higher education – where are the educators? Int J Educ Technol High Educ 2019 Dec;16(1):39. [doi: 10.1186/s41239-019-0171-0]
11. Kizilcec RF, Pérez-Sanagustín M, Maldonado JJ. Self-regulated learning strategies predict learner behavior and goal attainment in Massive Open Online Courses. Comput Educ 2017 Jan;104:18-33. [doi: 10.1016/j.compedu.2016.10.001]
12. Roll I, Winne PH. Understanding, evaluating, and supporting self-regulated learning using learning analytics. J Learn Anal 2015;2(1):7-12. [doi: 10.18608/jla.2015.21.2]
13. Järvelä S, Hadwin AF. New frontiers: regulating learning in CSCL. Educ Psychol 2013 Jan;48(1):25-39. [doi: 10.1080/00461520.2012.748006]
14. Burgess A, McGregor D, Mellis C. Medical students as peer tutors: a systematic review. BMC Med Educ 2014 Jun 9;14:115. [doi: 10.1186/1472-6920-14-115] [Medline: 24912500]
15. Kung TH, Cheatham M, Medenilla A, et al. Performance of ChatGPT on USMLE: potential for AI-assisted medical education using large language models. PLOS Digit Health 2023 Feb;2(2):e0000198. [doi: 10.1371/journal.pdig.0000198] [Medline: 36812645]
16. Ganotice FA, Datu JAD, King RB. Which emotional profiles exhibit the best learning outcomes? A person-centered analysis of students' academic emotions. Sch Psychol Int 2016 Oct;37(5):498-518. [doi: 10.1177/0143034316660147]
17. Nurmi JE, Aunola K. Task-motivation during the first school years: a person-oriented approach to longitudinal data. Learn Instr 2005 Apr;15(2):103-122. [doi: 10.1016/j.learninstruc.2005.04.009]
18. Ganotice FA Jr, Dizon J, Shen X, et al. Exploring health and allied science students' achievement emotions profiles: a person-centered analysis. BMC Med Educ 2025 Oct 29;25(1):1518. [doi: 10.1186/s12909-025-08088-5] [Medline: 41163164]
19. Spada F, Caruso R, Notarnicola I, et al. Analyzing readiness for interprofessional education among health program students using hierarchical clustering. J Interprof Care 2025;39(3):391-398. [doi: 10.1080/13561820.2025.2452973] [Medline: 39835532]
20. Winne PH, Hadwin AF. Studying as self-regulated learning. In: Hacker DJ, Dunlosky J, Graesser AC, editors. Metacognition in Educational Theory and Practice: Lawrence Erlbaum Associates; 1998:277-304.
21. Roeser RW, Galloway M. Studying motivation to learn in early adolescence: a holistic perspective. In: Pajares F, Urdan T, editors. Academic Motivation of Adolescents: Information Age Publishing; 2002:331-372. [doi: 10.1108/978-1-60752-554-720251013]
22. Pardo A, Han F, Ellis RA. Combining university student self-regulated learning indicators and engagement with online learning events to predict academic performance. IEEE Trans Learning Technol 2017;10(1):82-92. [doi: 10.1109/TLT.2016.2639508]
23. Pintrich PR, De Groot EV. Motivational and self-regulated learning components of classroom academic performance. J Educ Psychol 1990;82(1):33-40. [doi: 10.1037/0022-0663.82.1.33]
24. MacDonald CJ, Archibald D, Trumpower DL, Casimiro L, Cragg B, Jelley W. Designing and operationalizing a toolkit of bilingual interprofessional education assessment instruments. J Res Interprof Pract Educ 2010;1(3):304-316. [doi: 10.22230/jripe.2010v1n3a36]
25. Kuo YC, Walker AE, Belland BR, Schroder KEE. A predictive study of student satisfaction in online education programs. Int Rev Res Open Distrib Learn 2013;14(1):16-39. [doi: 10.19173/irrodl.v14i1.1338]
26. R Core Team. The R Project for Statistical Computing. 2024. URL: https://www.R-project.org/ [accessed 2026-08-15]
27. Hair JF, Black WC, Babin BJ, Anderson RE. Multivariate Data Analysis, 8th edition: Cengage Learning; 2019.
28. Cohen J. Statistical Power Analysis for the Behavioral Sciences, 2nd edition: Lawrence Erlbaum Associates; 1988. URL: https://www.taylorfrancis.com/books/mono/10.4324/9780203771587/statistical-power-analysis-behavioral-sciences-jacob-cohen [accessed 2026-08-17]

XSL•FO
RenderX

29. Aleven V, Roll I, McLaren BM, Koedinger KR. Help helps, but only so much: research on help seeking with intelligent tutoring systems. Int J Artif Intell Educ 2016 Mar;26(1):205-223. [doi: 10.1007/s40593-015-0089-1]

Abbreviations

GenAI: generative AI
GFI: goodness of fit index
IFI: incremental fit index
IPE: interprofessional education
MD: mean difference
MSLQ: Motivated Strategies for Learning Questionnaire
NFI: normed fit index
NPT: near-peer teacher
RMSEA: root mean square error of approximation
SRL: self-regulated learning
SRMR: standardized root mean square residual

Edited by A Hasan Sapci; submitted 20.Mar.2026; peer-reviewed by G Kolcu, M Elbattah, SSS Yeung; revised version received 10.Jul.2026; accepted 01.Aug.2026; published 10.Sep.2026.

Please cite as:
Ganotice Jr FA, Dizon JIWT, Shen X
Profiling Human-AI Regulatory Support in Technology-Enhanced Interprofessional Education Among Health Professions Students: Person-Centered Exploratory Study
JMIR Med Educ 2026;12:e95772
URL: https://mededu.jmir.org/2026/1/e95772
doi:10.2196/95772

© Fraide A Ganotice Jr, John Ian Wilzon T Dizon, Xiaoai Shen. Originally published in JMIR Medical Education (https://mededu.jmir.org), 10.Sep.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Education, is properly cited. The complete bibliographic information, a link to the original publication on https://mededu.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

A Dual Hospital–University Specialist Training Reform to Address Workforce Shortage and Geographic Maldistribution in Indonesia: Integrative Qualitative Policy Process Review

Asnawi Abdullah[1,2], BSc PH, MHSM, MSc HPPF, DLSHTM, PhD; Yuli Farianti[3], MD, MEpid; Lupi Trilaksono[3], SF, Apt, MM; Arianti Anaya[4], MKM; Anna Kurniati[3], SKM, MA, PhD; Oos Fatimah Rosyati[3], Dra, MKes; Etik Retno Wiyati[3], MD, MARS, MH; Mimi Sumiarsih[3], SKM, MKM; Christa Gumanti Manik[3], SKep, NS, MPH; Nyiayu H A Sonia[3], drg, MKES; Farizal Rizky Muharram[5], MD, MMSc; Kevin Tadeus Simanjuntak[5], SKed; Sydney Tjandra[5], SKed; Akemat Akemat[3], SKep, MKes, DNS; Else Mutiara Sihotang[3], MD, SpPK; Leni Kuswandari[3], drg, MKM; Siti Yunianti[3], MD, MKM; Fialisa Asriwhardani[3], NERS, MSN

[1]Health Policy Agency, Ministry of Health, Jl. Percetakan Negara No.29 lantai 2, RT.23/RW.7, Johar Baru, Central Jakarta City, Jakarta, Indonesia

[2]University Muhammadiyah Aceh, Banda Aceh, Indonesia

[3]Ministry of Health, Jakarta, Jakarta, Indonesia

[4]Indonesian Health Council, Jakarta, Indonesia

[5]ARC Institute Indonesia, Surabaya, Indonesia

Corresponding Author:
Asnawi Abdullah, BSc PH, MHSM, MSc HPPF, DLSHTM, PhD
Health Policy Agency, Ministry of Health, Jl. Percetakan Negara No.29 lantai 2, RT.23/RW.7, Johar Baru, Central Jakarta City, Jakarta, Indonesia

Abstract

Background: Indonesia faces 3 interlocking medical workforce crises: an absolute specialist deficit projected to reach 70,000 by 2032 (national density of 0.18 per 1000 population vs the Ministry of National Development Planning [Bappenas] target of 0.28), severe maldistribution (with nearly 59% of specialists concentrated in Java), and a structural anomaly in which residents pay tuition while performing essential clinical work. The 2023 Health Law (Law 17/2023) authorized a transformative reform: a hospital-based residency pathway (Rumah Sakit Pendidikan Penyelenggara Utama [primary teaching hospital; RSPPU]) operating in parallel with the long-established university-based system.

Objective: This study aimed to examine the rationale, policy design, and early implementation of Indonesia's dual hospital–university specialist medical education reform, interpreted through an 8-step change management framework developed by Kotter, and to identify transferable lessons for low- and middle-income countries.

Methods: We conducted an integrative qualitative policy process review combining two evidence streams: (1) systematic documentary analysis of 19 source documents (17 primary legal, regulatory, and policy instruments plus 2 interministerial joint monitoring site-visit reports) and (2) engagement of 34 key informants through semistructured interviews and focus group discussions (45 - 120 min), comprising policymakers, collegium representatives, hospital leaders, and residents across all 6 pilot sites, recruited purposively until thematic saturation. Interview and focus group data were analyzed using a hybrid deductive-inductive thematic approach with an 8-step change management framework developed by Kotter as an a priori coding frame, and member checking was completed with 7 of the 34 informants.

Results: The reform designated 6 top-tier national referral hospitals as RSPPUs and enrolled 52 residents from 412 applicants (an acceptance rate of 12.6%) across 6 high-need specialties (ophthalmology, cardiology, pediatrics, orthopedics, neurology, and oncology). All 6 pilot sites established functional education units and designated institutional officials, adopted dual accreditation, and operationalized an integrated e-logbook for competency tracking, real-time monitoring of 80-hour duty limits, and anonymous bullying reporting. An interministerial joint monitoring team visited all 6 sites and scored each site as satisfactory or better across governance, curriculum, faculty, infrastructure, and learner support. Four cross-cutting themes emerged: financial-barrier removal, dual-governance pragmatism, accreditation strain, and equity-anchored deployment. Persistent tensions include variable educator compensation across hospitals; however, a standardized national framework remains under development. Mapping to the framework developed by Kotter demonstrated strong evidence for steps 1 to 6 and early evidence for steps 7 to 8.

Conclusions: Indonesia's dual hospital–university residency model is a scalable, equity-oriented, and competency-based reform that is operationally feasible and globally aligned in its early implementation. While long-term effectiveness and sustainability

XSL•FO
RenderX

await longitudinal evaluation, the design offers a transferable, not yet definitively replicable, template for low- and middle-income countries confronting parallel workforce crises.

(JMIR Med Educ 2026;12:e90270) doi:10.2196/90270

KEYWORDS

internship and residency; medical education; health workforce; health policy; health care disparities; universal health coverage; Indonesia

Introduction

The Specialist Workforce Crisis

Indonesia faces a persistent crisis in the production and equitable distribution of medical specialists, a challenge that directly undermines the quality, accessibility, and equity of health services for its 287 million people [1]. As of January 2025, national data report only 51,443 specialists, equating to a density of 0.18 specialists per 1000 population [2], significantly below the Ministry of National Development Planning (Bappenas) target of 0.28 per 1000 [3]. This overall deficit is projected to widen to 70,000 specialists by 2032 [4]. Compounding this quantitative shortage is severe maldistribution: nearly 59% of specialists are concentrated in Java and major urban centers [5], leaving vast areas of the archipelago, particularly underserved regions (Daerah Tertinggal, Perbatasan, dan Kepulauan [underdeveloped, border, and island regions; DTPK]), critically underserved [4].

This specialist deficit translates directly into significant public health vulnerability and contributes to avoidable mortality and morbidity across national priorities [6,7]. Delays in accessing obstetric and pediatric specialists contribute to persistently high maternal and neonatal mortality, particularly in regions with the lowest specialist availability [8]. The inadequacy of specialist coverage critically undermines core initiatives, including Maternal and Child Health (Kesehatan Ibu dan Anak [KIA]) programs [9]. The growing burden of noncommunicable diseases, which account for up to 73% of all deaths in Indonesia [10], necessitates timely access to specialized cardiology, oncology, and neurology services—most of which remain out of reach for populations outside major cities [9]. Low specialist density is consistently associated with adverse public health outcomes, including higher hospitalization rates for ambulatory care–sensitive conditions, longer diagnostic delays, and reduced survival for time-sensitive emergencies [11,12]. Given that the current university-based system produces only about 2700 specialists per year, eliminating the existing deficit of roughly 30,000 specialists would take at least a decade, even before accounting for retirements or the rising demand driven by population aging [5].

Generalist and Specialist Pathways in Indonesia

To situate the present reform, we briefly describe the Indonesian medical workforce hierarchy. Generalists (dokter umum or general practitioners) complete a 5.5‐year to 6-year undergraduate medical program comprising preclinical and clinical training, followed by a 1-year mandatory national internship (Internship Dokter Indonesia) before obtaining independent licensure. As of January 2025, Indonesia had approximately 200,000 active generalists staffing community health centers (Puskesmas), district hospitals, and family practice settings [13]. Generalist training and the internship are governed by separate regulatory cascades and have been described elsewhere [14].

Specialists complete an additional residency program (Program Pendidikan Dokter Spesialis [specialist medical education program; PPDS]) lasting 2 to 6 years, in one of 39 specialties recognized by the Indonesian Health Council (Konsil Kesehatan Indonesia [KKI]) and governed by specialty collegia. Until 2023, all PPDS programs were exclusively university-based, with residents officially registered as tuition-paying postgraduate students. A consultant subspecialist tier requires a further 2 to 3 years. This paper concerns the specialist (PPDS) tier; the hospital-based RSPPU pathway introduced in 2024 sits within this tier and is the focus of analysis.

Indonesia's existing residency framework presents a structural anomaly when benchmarked against global standards [15]. In the United States, the United Kingdom, the Netherlands, Japan, and Australia, residents are recognized as licensed members of the clinical workforce who receive salaries, benefits, and formal employment status [16-20]. This holds in Malaysia as well, which operates both hospital-based and university-based systems but pays all residents a government salary [21]. In Indonesia's traditional university-based model, by contrast, residents are classified as tuition-paying students who pay substantial education fees while performing full clinical duties without remuneration [22]. This arrangement creates substantial equity barriers to entry, limiting the pool of aspiring physicians able to pursue specialization and constraining the size of the specialist workforce [23,24].

The 2023 Reform: A Dual Hospital–University Pathway

To urgently address the triple crisis of insufficient production, inequitable distribution, and the structural anomaly of unpaid residency, the Indonesian government mandated a transformative reform under the 2023 Health Law [25]. This landmark legislation introduced a dual-track specialist education system that formally integrated the Hospital-Based Specialist Program (Rumah Sakit Pendidikan Penyelenggara Utama [primary teaching hospital; RSPPU]) with existing university-based programs. The new model positions designated national referral hospitals as co-organizers of residency training, ensures that residents receive living stipends and pay no tuition, and explicitly links specialist training to equity goals by binding deployment to underserved regions upon graduation. This dual approach is a comprehensive, scalable, and competency-based

XSL•FO
RenderX

strategy to accelerate specialist output and drastically improve equitable access to specialized care [26].

Aim and Research Questions

The aim of this paper is threefold: (1) to document the rationale, policy design, and early implementation of Indonesia's dual hospital–university specialist medical education reform; (2) to interpret these processes through an 8-step change management framework developed by Kotter; and (3) to extract transferable policy lessons for low- and middle-income countries (LMICs) facing analogous workforce crises. We address three guiding research questions: (RQ1) What policy logic and political processes catalyzed the dual-track reform? (RQ2) How is the hospital-based pathway being implemented during the first 18 months and with what early outcomes? (RQ3) What systemic tensions and policy implications emerge for domestic scale-up and international adaptation?

Methods

Study Design and Theoretical Framework

We used an integrative qualitative policy process review design to examine the design and early operational dynamics of Indonesia's dual-track specialist education reform. An 8-step change management framework developed by Kotter [27-29] served as both an a-priori interpretive lens and a deductive coding frame. Each step was mapped to observable indicators in the RSPPU rollout: (1) establishing urgency: quantification of the specialist deficit and high-stakes political articulation; (2) forming a guiding coalition: cross-ministerial governance arrangements involving the Ministry of Health (MoH), Ministry of Higher Education, KKI, Indonesian Accreditation Agency for Higher Education in Health (Lembaga Akreditasi Mandiri Pendidikan Tinggi Kesehatan Indonesia [LAM-PTKes]; IAAHEH), and the specialty collegia; (3) developing a change vision: Law 17/2023 and its successor regulations; (4) communicating the vision: subsequent public communication; (5) empowering broad-based action: Government Regulation 28/2024 [30], Permenkes 14/2024 [31], and Lembaga Pengelola Dana Pendidikan (Indonesia Endowment Fund for Education; LPDP) financing; (6) generating short-term wins: pilot cohort enrollment and the unified digital platform; (7) consolidating gains: joint monitoring, dual accreditation, and expansion to additional RSPPUs; and (8) anchoring change in culture: embedding mechanisms including the integrity pact, Pangkalan Data Pendidikan Tinggi (Higher Education Database; PDDikti) recognition, and the Badan Penyelenggara Jaminan Sosial (National Social Security Agency; BPJS)–linked financing pipeline.

Study Setting and Timing of Data Collection

Data collection ran from October to November 2025, covering the first 18 months of program operation. At that point, all 52 inaugural residents were in their fourth or fifth semester; none had yet graduated, as the shortest pilot specialty (ophthalmology) requires a minimum of 7 semesters before the national board examination. The study therefore examines design and early implementation fidelity, not graduate outcomes.

Data Sources and Triangulation

Documentary Analysis

The documentary corpus comprised 19 source documents: 17 primary legal, regulatory, policy, and standards documents and 2 official interministerial joint monitoring site-visit reports from December 2024 and June 2025. The 17 primary documents included Law 17/2023 on Health; Government Regulation 28/2024; Ministerial Regulations Permenkes 31/2022 and 14/2024; operational guidelines for the RSPPU program; regulatory materials from the KKI; and global and national graduate medical education standards (Accreditation Council for Graduate Medical Education [ACGME]–International [16] and IAAHEH/LAM-PTKes). Documents were retrieved between September and November 2025 from official ministerial portals. A standardized data-extraction template captured legal authority, effective date, governance roles, financing provisions, and accreditation requirements. In addition to these primary legal and regulatory instruments, we obtained and analyzed the 2 official site-visit reports produced by the interministerial joint monitoring team (December 2024 and June 2025), which applied a standardized site-assessment instrument across all 6 pilot sites; these monitoring reports formed part of the documentary evidence stream and are the source of the findings reported under "Joint Monitoring Team Findings" in the *Results* section.

Semistructured Key Informant Interviews and Focus Group Discussions

We purposively recruited 34 key informants spanning all institutional roles relevant to the reform (Table 1): 11 national policymakers (coded P-01 to P-11) from the MoH and the Indonesian Health Council; 7 collegium representatives (C-01 to C-07; 1-2 per specialty); 8 hospital leaders (H-01 to H-08) drawn from designated institutional officials (DIOs), Program Directors, and Clinical Competency Committees (CCCs) across the 6 RSPPU sites; and 8 enrolled residents (R-01 to R-08; 1-2 per site). Recruitment proceeded iteratively until sufficient thematic saturation was achieved across stakeholder groups.

XSL•FO
RenderX

Table . Key informant sample composition (N=34).

Stakeholder role	Informants, n	Coding scheme
National policymakers (MoH[a], Indonesian Health Council)	11	P-01 to P-11
Collegium (1 - 2 per specialty)	7	C-01 to C-07
Hospital leaders (DIOs[b], Program Directors, and/or CCCs[c], 1 - 2 per RSPPU[d] site)	8	H-01 to H-08
Residents (1 - 2 per RSPPU site)	8	R-01 to R-08
Total	34	—[e]

[a]MoH: Ministry of Health.

[b]DIO: designated institutional official.

[c]CCC: Clinical Competency Committee.

[d]RSPPU: Rumah Sakit Pendidikan Penyelenggara Utama (primary teaching hospital).

[e]Not applicable.

Data were collected through semistructured key informant interviews (KIIs) and focus group discussions (FGDs). FGDs were held separately by institutional role (residents, collegium representatives, hospital leaders, and MoH technical teams) to minimize hierarchical constraints and encourage candid exchange. Several higher-level decision-makers participated in individual KIIs instead, given that their roles were less suited to group discussion. The FGD format gave participants room to respond to, clarify, and build on each other's accounts of implementation experiences and institutional dynamics. All sessions were conducted in Bahasa Indonesia by trained qualitative researchers with prior experience in health policy and health systems research. Each session was staffed by 2 interviewers: one MoH-affiliated researcher and one external researcher. This pairing was intended to balance contextual familiarity and institutional legitimacy against independence from formal supervisory or evaluative relationships. A semistructured topic guide, structured around an 8-step framework developed by Kotter, was used throughout, supplemented by probes specific to financial mechanisms, dual accreditation, and equity-oriented deployment. Sessions lasted approximately 120 minutes, were audio-recorded with verbal informed consent, transcribed verbatim, and anonymized using a role-and-number coding scheme. Field notes were maintained throughout. Transcripts and documentary extracts were analyzed using a hybrid deductive-inductive thematic approach [32], with an 8-step framework developed by Kotter as an a-priori deductive coding frame and inductive codes capturing emergent operational tensions. Rigor was reinforced through triangulation between the documentary and interview streams, member checking of preliminary themes with 7 of the 34 informants, and an audit trail linking each theme to its source segments; reporting follows the Standards for Reporting Qualitative Research [33].

Ethical Considerations

Verbal informed consent was obtained from all interviewees prior to recording. Data were anonymized, and any identifiable information was excluded to protect participant privacy. Formal ethical approval was not required for this study, as it involved a policy evaluation using nonclinical, nonidentifiable data [34]. No patient-level data were used.

Results

Overview of Findings

The integrative policy review confirms that Indonesia's reform establishes a dual-track specialist training system that is operationally feasible and addresses critical shortages and maldistribution. Across the 2 evidence streams, four cross-cutting themes structure the findings: (1) financial barrier removal as an equity lever; (2) dual governance pragmatism; (3) accreditation strain and capacity-building; and (4) equity-anchored deployment. Triangulated evidence supports strong fidelity to Kotter's steps 1 to 6 and early evidence for steps 7 to 8 (Table 2). Key successes include eliminating financial barriers through stipends and waived tuition, ensuring equitable access via a centralized digital selection system, and strengthening quality assurance and resident well-being through digital tools (an e-logbook with anonymous reporting). Persistent operational tensions are detailed in the dedicated subsection below.

XSL•FO
RenderX

Table . Triangulated evidence mapped to an 8-step change management framework developed by Kotter.

Kotter step	Indonesian RSPPU[a] evidence (illustrative)	Strength of evidence
Step 1—Establish urgency	Specialist deficit 70,000 by 2032; presidential and ministerial articulation	Strong[b]
Step 2—Form a guiding coalition	Joint interministerial governance (MoH[c], MoHE[d], KKI[e], IAAHEH[f], and collegia)	Strong
Step 3—Develop vision and strategy	Law 17/2023; Government Regulation 28/2024; Permenkes 14/2024	Strong
Step 4—Communicate the vision	May 2024 Presidential launch; sustained ministerial public communication	Strong
Step 5—Empower broad-based action	LPDP[g]-financed positions; e-logbook digital platform	Strong
Step 6—Generate short-term wins	52 residents enrolled across 6 RSPPUs; 100% slots filled	Strong
Step 7—Consolidate gains	Joint monitoring (Dec 2024+Jun 2025); dual accreditation; scale-up announced	Early evidence[h]
Step 8—Anchor change in culture	Integrity pact, PDDikti[i] integration, BPJS[j]-linked financing—all in progress	Early evidence

[a]RSPPU: Rumah Sakit Pendidikan Penyelenggara Utama (primary teaching hospital).

[b]"Strong" indicates triangulated convergence across both evidence streams.

[c]MoH: Ministry of Health.

[d]MoHE: Ministry of Higher Education.

[e]KKI: Konsil Kesehatan Indonesia (Indonesian Health Council).

[f]IAAHEH: Indonesian Accreditation Agency for Higher Education in Health (Lembaga Akreditasi Mandiri Pendidikan Tinggi Kesehatan Indonesia [LAM-PTKes]).

[g]LPDP: Lembaga Pengelola Dana Pendidikan (Indonesia Endowment Fund for Education).

[h]"Early evidence" indicates partial documentation, with mechanisms still being institutionalized.

[i]PDDikti: Pangkalan Data Pendidikan Tinggi (Higher Education Database).

[j]BPJS: Badan Penyelenggara Jaminan Sosial (National Social Security Agency).

Documentary Findings

The Legal and Regulatory Cascade

Indonesia's Prereform Specialist Training System

Indonesia's specialist medical education system was historically rooted in hospital-based apprenticeship. It originated in the colonial era at the Geneeskundige Hogeschool (Medical College), where postgraduate training produced "assistant specialists" without conferring a formal specialist degree. Following independence, specialist training was re-established in the early 1950s by Universitas Indonesia and Universitas Gadjah Mada, maintaining a hospital-centered model. Admissions during this period were largely informal and managed directly by teaching hospitals [35].

A major institutional transition occurred in 1979 when the Ministry of Education and Culture assumed authority over postgraduate medical education through Decree 024/DJ/Kep/1979, shifting responsibility from the Indonesian Medical Association (Ikatan Dokter Indonesia [IDI]). In the same year, the Consortium of Health Sciences was formed to develop national guidelines and assess institutional readiness for launching residency programs [14]. Entry into residency was subsequently standardized, requiring completion of a compulsory internship and successful passage of a national competency examination, the precursor to today's Uji Kompetensi Nasional Peserta Didik Profesi Dokter (UKNPDPD) [35].

Further refinements followed the 2004 Medical Practice Law, which established the KKI and specialty-specific collegia [35,36]. These bodies were mandated to set national training standards, oversee program accreditation, and administer board certification examinations. The system was formally codified under the 2013 Medical Education Law, which legally entrenched a university-based residency model, positioning higher-education institutions as the official organizers of specialist training in collaboration with professional associations [14]. While this framework strengthened academic oversight, it also distanced program governance from the clinical environments where most specialist training occurs, a structural tension that would later motivate the 2023 reform introducing the hospital-based RSPPU pathway.

From Bottleneck to Breakthrough: The 2023 Reform

Following the 2012 Higher Education Law [37] and the 2013 Medical Education Law [38], universities assumed full responsibility for curriculum design, certification, and program accreditation. While this framework was intended to safeguard academic quality, it inadvertently constrained the system's adaptability. Annual specialist output plateaued at approximately

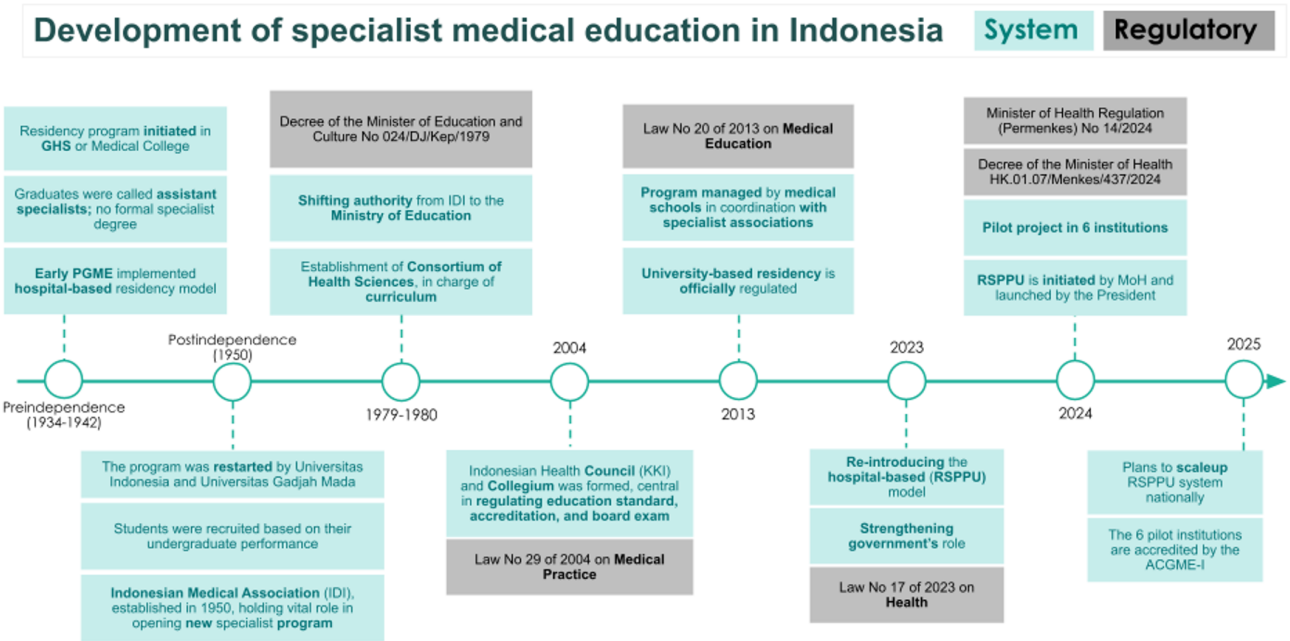
XSL•FO
RenderX

2700 graduates, which was insufficient to close Indonesia's growing deficit. Stringent accreditation requirements, while essential for quality, created bottlenecks in expanding program capacity. More critically, the system classified residents as tuition-paying students despite their performance of full clinical duties [22], imposing financial hardship and contradicting global norms in which residents are recognized as supervised professionals and compensated accordingly.

Prior to the introduction of the RSPPU model, the MoH pursued multiple strategies to alleviate Indonesia's chronic specialist shortage: (1) expanding enrollment quotas in university-based residency programs; (2) implementing the Wajib Kerja Dokter Spesialis (mandatory specialist service program; WKDS) mandatory rural service program, later invalidated by the Supreme Court [33]; (3) establishing academic health system networks to integrate universities, teaching hospitals, and local governments; and (4) sponsoring physicians for specialist training abroad. While each measure yielded partial gains, none addressed the systemic constraints of a rigid, university-centric training architecture. Both the scale-up of training capacity and the retention of graduates in underserved regions remained elusive.

By 2022, Indonesia's chronic specialist shortage—exacerbated by a double burden of disease and a rapidly aging population—had reached a tipping point. Policymakers proposed a structural recalibration: revival of a hospital-based residency model, informed by international experience in the United Kingdom, the United States, South Korea, and Singapore. The approach did not seek to replace universities but to complement them, recognizing that the vast majority of specialist training already occurs in clinical settings. With nearly 500 hospitals formally affiliated with medical schools as teaching sites and over 3000 hospitals nationwide, the untapped institutional capacity was substantial. This vision was codified in Health Law 17/2023, which formally introduced the RSPPU, primary teaching hospitals authorized to serve as co-organizers of specialist training. The proposal triggered significant interinstitutional tension. Universities expressed concern about diminished academic authority, while the Ministry of Higher Education raised questions about degree legitimacy and quality assurance. After months of deliberation, a pragmatic compromise was forged: RSPPUs would assume primary responsibility for training delivery, mentorship, and day-to-day program management, while universities would retain cosignatory authority over professional degrees and academic certification. The operational parameters were detailed in Government Regulation No 28/2024 and in MoH Regulation 14/2024. The development of specialist medical education in Indonesia is summarized chronologically in Figure 1.

Figure 1. Chronological development of specialist medical education in Indonesia, from the colonial-era Geneeskundige Hogeschool through the 2023 hospital-based Rumah Sakit Pendidikan Penyelenggara Utama (primary teaching hospital; RSPPU) pathway. The lower text block now reads: "Hospital-based specialist training was re-established ('restarted') under the RSPPU model in 2024 after a 45-year hiatus during which all specialist training had been university-led." [14,35]. ACGME-I: Accreditation Council for Graduate Medical Education–International; GHS: Geneeskundige Hogeschool (now Faculty of Medicine, Universitas Indonesia); MoH: Ministry of Health; PGME: Postgraduate Medical Education.



Pilot Cohort Profile and Selection Outcomes

The inaugural cohort comprised 52 residents selected from 412 (overall acceptance rate 12.6%) eligible applicants nationally. All 52 announced positions were filled. Distribution across pilot specialties was as follows: ophthalmology (Cicendo Eye Hospital, n=8 slots filled, from 96 applicants); cardiology (National Cardiovascular Center Hospital Harapan Kita, n=10, from 87 applicants); pediatrics (National Women and Children Health Center Hospital Harapan Kita, n=8, from 79 applicants); orthopedics (Prof Dr Soeharso Orthopedic Hospital, n=10, from 64 applicants); neurology (National Brain Center Hospital, n=10, from 45 applicants); and oncology (Dharmais Cancer Hospital, n=6, from 41 applicants). As of December 2025, no resident had withdrawn; all 52 remained enrolled.

XSL•FO
RenderX

Joint Monitoring Team Findings

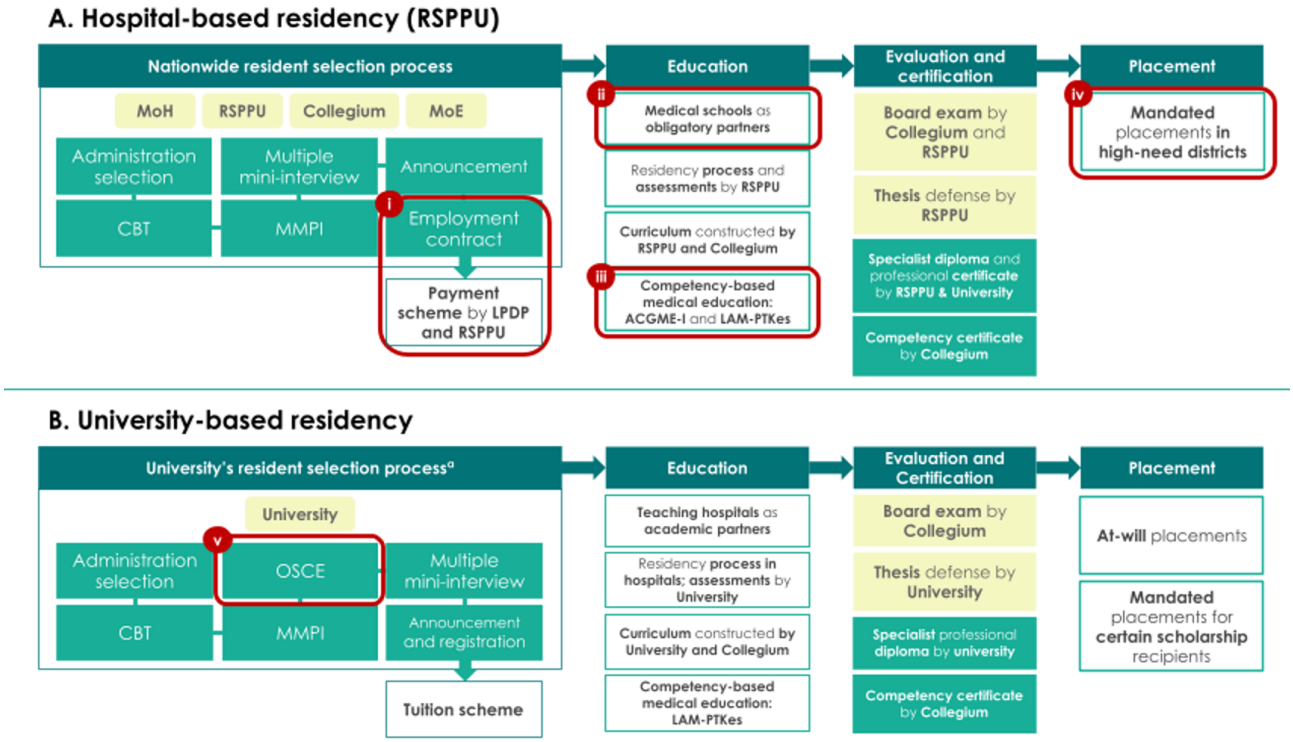
The joint monitoring exercise was documented in 2 official interministerial site-visit reports that were analyzed as part of the documentary evidence stream (see *Methods*). A joint monitoring team composed of representatives from the MoH (Directorate of Health Human Resources Provision), the Ministry of Higher Education, IAAHEH/LAM-PTKes, and the relevant specialty collegia conducted 2 structured site visits: the first in December 2024 (approximately 7 mo after the May 2024 presidential launch) and the second in June 2025 (approximately 13 mo post-launch). At each visit, a standardized 23-item instrument adapted from the ACGME-International Self-Study framework assessed five governance domains: (1) governance and DIO leadership; (2) curriculum implementation and milestone tracking; (3) faculty development and supervision ratios; (4) infrastructure (simulation, library, and e-logbook uptime); and (5) learner support (well-being, anonymous reporting use, duty-hour compliance). Each domain was scored on a 4-point ordinal scale (excellent, satisfactory, needs improvement, or unsatisfactory). Across both visits, all 6 pilot sites were rated satisfactory or better on all 5 domains, with no domain at any site scored needs improvement or unsatisfactory (Table 2, Kotter step 7).

Implementation: Hospital Selection, Governance, and Training Processes

Hospital Designation and Functional Education Units

By mid-2024, the reform transitioned from policy to practice, with the designation of 6 national referral hospitals as pilot RSPPUs (Figure 2). Hospitals were selected based on demonstrable institutional commitment, top-tier (Paripurna) accreditation, high clinical caseloads, and robust infrastructure. Although all 6 had previously served as affiliate training sites or hosted fellowship programs, none functioned as primary teaching hospitals under the university-based model. This historical positioning afforded them greater autonomy to design residency curricula fully embedded within their clinical ecosystems. Each RSPPU established a functional education unit responsible for curriculum delivery, mentorship, and quality assurance, with duty hours capped at 80 hours per week and night shifts integrated into a structured schedule. Crucially, residents are now formally recognized not as tuition-paying students but as supervised health professionals entitled to social protection, welfare safeguards, and employment-based rights—a fundamental shift codified in Permenkes 14/2024.

Figure 2. Comparative process flow for university-based vs hospital-based specialist training. Red highlight boxes mark distinctive Rumah Sakit Pendidikan Penyelenggara Utama (primary teaching hospital; RSPPU)–specific elements: (A) the binding preadmission integrity pact for return-of-service; (B) the LPDP-financed full tuition and stipend coverage; and (C) the dual Accreditation Council for Graduate Medical Education–International (ACGME-I)/Indonesian Accreditation Agency for Higher Education in Health (Lembaga Akreditasi Mandiri Pendidikan Tinggi Kesehatan Indonesia [LAM-PTKes]; IAAHEH) accreditation requirement. *Process may vary by university regulation and autonomy. CBT: Computer-Based Test; CCC: Clinical Competency Committee; DIO: designated institutional official; LPDP: Lembaga Pengelola Dana Pendidikan (Indonesia Endowment Fund for Education); MMPI: Minnesota Multiphasic Personality Inventory; MoE: Ministry of Education; MoH: Ministry of Health; OSCE: Objective Structured Clinical Examination; PDDikti: Pangkalan Data Pendidikan Tinggi (Higher Education Database); PEC: Program Evaluation Committee.



Dual Accreditation and Curriculum Codevelopment

To align with global best practices, the MoH adopted core elements of the ACGME-International framework. Each RSPPU appointed a DIO responsible for graduate medical education oversight, governance, and regulatory compliance. While hospitals initially struggled to meet the ACGME-I expectation that DIOs dedicate 80% of their time to educational duties, given their concurrent clinical responsibilities, ACGME-led training workshops proved instrumental in clarifying

XSL•FO
RenderX

foundational governance concepts, including the CCC and Program Evaluation Committee (PEC). Concurrently, all RSPPU programs remain subject to national accreditation by IAAHEH/LAM-PTKes. To ensure parity with university-based residencies, both tracks adhere to the same core curriculum, as defined by the National Specialist Education Standard issued by KKI. Program-specific adaptations (muatan lokal) are permitted to reflect institutional strengths; for example, Cicendo Eye Hospital integrates community ophthalmology as a distinguishing component.

Selection Process, Financing, and Graduation Pathway

Registration was conducted via a unified online information system accessible nationwide. Every applicant signed an integrity pact pledging to return to their region of origin upon graduation. The selection protocol incorporated administrative screening, computer-based cognitive testing, psychological assessment (including the Minnesota Multiphasic Personality Inventory [MMPI]), and a structured interview. Successful candidates received an employment contract rather than a tuition invoice. The e-logbook supported structured documentation of clinical procedures, milestone-based competencies, real-time duty-hour monitoring, and anonymous reporting of bullying or mistreatment. All program-related costs were fully covered by the LPDP. In 2024, the Ministry of Finance allocated 1000 funded positions with a target of 5000 specialists within 2 years. Upon completion, residents undergo a dual evaluation: a national board examination administered by the collegium and the RSPPU and a thesis defense overseen by the hospital. Graduates receive a joint specialist diploma cosigned by the RSPPU and the partner university and a national competency certificate issued by the collegium. The final phase—mandatory deployment to districts with critical shortages of specialists—operationalizes the reform's redistributive mission.

Stakeholder Themes From KIIs

Theme 1: Financial Barrier Removal as an Equity Lever

All 8 residents interviewed (R01 to R08) emphasized that the elimination of tuition fees and the introduction of a monthly stipend were the single decisive enabler of their decision to pursue specialization, particularly for those from outside Java:

> *At that time, I was not financially able to pay for specialist training on my own, so when support from the Ministry of Health became available, especially because it was structured as a scholarship, it further strengthened my motivation to pursue specialist training.* [R-08, Pediatrics]

Theme 2: Dual Governance: Pragmatism Over Purism

National policymakers (P-01 through P-04) and a representative of the Indonesian Health Council framed the reform as a functional realignment rather than an institutional displacement:

> *Until today, some still see RSPPU as competition. It is not. University-based and hospital-based training must coexist. We share the same vision: to ensure Papua and Maluku have their own specialists.* [P-07, Indonesian Health Council]

Theme 3: Accreditation Strain and Capacity-Building

Hospital leaders and collegium representatives described initial challenges in reconciling ACGME-International expectations with national IAAHEH standards, mainly highlighting documentation requirements and navigating interim governance structures. Meanwhile, the complementarity in quality was also valued; intensive capacity-building measures were implemented to ensure quality clinical teaching.

> *But if we look closely, both (IAAHEH and ACGME) essentially emphasize quality aspects, and ACGME further reinforces this.* [C-05, Collegium representative]

> *We examined 2 accreditation frameworks: the ACGME, and, like it or not, LAM-PTKes as well.* [H-03, RSPPU pilot]

Theme 4: Equity-Anchored Deployment

Policymakers and university administrators emphasized the structural placement of equity at the point of entry, not exit:

> *Selection in the hospital-based program is national, transparent, and system-based. All stages—written test, psychological test, interview—are computerized and involve the Ministry of Health, Higher Education, hospitals, and the professional collegia. Subjectivity is minimal.* [P-06, then Directorate of Health Human Resources Provision]

Faculty and University Administrator Perspectives

> *The DIO is expected to dedicate eighty percent of their work to education and only twenty percent to clinical service. In reality, however, our primary role is fundamentally service-oriented, as we are functionally appointed for clinical care. This has become one of the practical challenges we face in the field: how to divide our time appropriately.* [H-07, Senior Consultant Faculty, Radiation Oncology]

Early Implementation Tensions and Resident Experience

Three implementation tensions remained unresolved at 18 months. First, residents reported uncertainty about their formal recognition in the PDDikti national student database, a status that affects access to certain academic benefits such as government scholarships for subspecialization, library privileges, and postgraduate rail and air fare discounts. As of February 2026, an interministerial working group had agreed in principle on a dual learner-employee status entry, but technical integration was still pending. Second, the first selection round (May 2024) exposed logistical inequities: of 412 applicants, 301 (73.1%) came from outside Java, yet all centralized written and computer-based examinations required travel and accommodation that applicants estimated at IDR 4 to 9 million (US $250 - $570). Third, no national framework exists to compensate clinical educators for the additional teaching workload. One pilot RSPPU introduced interim honoraria arrangements, but the variation across sites raises equity concerns and is not sustainable at scale.

XSL•FO
RenderX

Mapping Findings to an 8-Step Framework Developed by Kotter

Triangulated evidence mapped to an 8-step framework developed by Kotter showed strong fidelity for steps 1 to 6 and early, partial evidence for steps 7 to 8 (Table 2). Step 1 (urgency) was established through the 70,000-specialist deficit projection and high-level political articulation. Step 2 (guiding coalition) is operationalized through the joint interministerial monitoring team. Step 3 (vision) is codified in Law 17/2023. Step 4 (communication) was executed through the nationally broadcast May 2024 presidential launch and sustained public communication by the MoH. Step 5 (empowerment) is reflected in successor regulations (Government Regulation 28/2024, Permenkes 14/2024) and LPDP financing. Step 6 (short-term wins) is evidenced by the enrollment of 52 residents and the operational unified digital platform. Step 7 (consolidating gains) is partially evidenced through dual accreditation and announced expansion to additional RSPPUs in 2025 to 2026. Step 8 (anchoring in culture) remains incomplete: the integrity pact, PDDikti integration, and BPJS-linked financing each represent unfinished embedding mechanisms.

Discussion

Principal Findings

This integrative policy process review of Indonesia's dual-track specialist training reform yields 3 principal findings. First, the RSPPU model is operationally feasible at 18 months: all 6 pilot hospitals established the requisite educational infrastructure, all 52 selected residents remain enrolled, and a multiagency joint monitoring framework documented satisfactory or better performance across governance, curriculum, faculty, infrastructure, and learner support. Second, the reform's defining innovation is structural rather than pedagogical. By abolishing resident tuition, reclassifying residents as supervised health professionals, and binding selection to an equity-anchored integrity pact, the model recasts specialist training as a public investment in spatial health equity rather than a private financial transaction. Third, the persistent implementation tensions, including dual-accreditation adaptation, the absence of a national clinical-educator compensation framework, PDDikti integration delays, and Java-centric examination logistics, are operational and tractable rather than conceptual and define a clear policy agenda for scale-up.

Comparison With Prior Work

The RSPPU model corrects a longstanding misalignment: specialist training has always taken place predominantly in hospitals, yet formal authority resided with universities. The 2023 Health Law resolves this not through institutional substitution but through functional realignment. Hospitals now lead curriculum delivery, mentorship, and day-to-day program management; universities retain cosignatory authority over degrees; and professional collegia ensure that both pathways adhere to the National Specialist Education Standard [39]. Internationally, Indonesia's design sits between fully hospital-employer models such as the United Kingdom NHS and the United States ACGME, and university-mandated systems found in parts of Continental Europe and Japan. Malaysia's parallel dual system [21] is the closest structural comparator, but Indonesia's model is distinctive in 3 respects: the explicit binding of admission to a return-of-service integrity pact; LPDP-financed full coverage of resident fees; and a dual ACGME-I/IAAHEH accreditation requirement applied to every pilot site. Unlike the US model, Indonesia's pathway emerged from negotiated tripartite compromise rather than a top-down mandate. Unlike the UK's centrally employed model, it preserves academic coauthority. This hybridity may be what makes the model relevant to other LMICs whose health and higher-education ministries operate in parallel rather than in aligned hierarchies.

Operationalization of Equity

The reform's equity imperative is operationalized through 3 reinforcing mechanisms. First, a unified national selection platform removes geographic barriers, allowing physicians from Aceh to Papua to compete on equal terms. Second, a binding integrity pact at the point of entry obliges graduates to serve in their regions of origin, a commitment built into admission rather than imposed after graduation. Third, residents are now salaried professionals rather than fee-paying students. With tuition fully covered by LPDP and monthly stipends provided, the model removes the largest financial barrier that historically excluded talented but underresourced candidates, bringing Indonesia into line with the global norm of recognizing residents as members of the clinical workforce. The equity mechanism operates at entry, not exit. Prior postgraduation deployment schemes such as the WKDS were judicially invalidated [40], and exit-based coercion has failed in multiple LMICs. By embedding the return-of-service obligation in the admission contract and financing the residency with public funds, the reform converts deployment from a contested instrument of coercion into a consensual condition of participation.

Early Implementation Considerations

Early implementation across 6 national referral hospitals confirms operational feasibility. The joint monitoring framework, itself a Kotter step 7 consolidation mechanism, produced satisfactory or better ratings in both the 2024 and 2025 site visits. Several tensions identified in the *Results* nonetheless require interpretation. The variability in clinical-educator compensation across pilot sites is not a temporary inconvenience but a fundamental design gap. Without a national framework, scale-up risks producing a 2-tier system in which residents at well-resourced flagship hospitals receive structured mentorship while those at expanded RSPPUs encounter underincentivized faculty. The PDDikti integration delay signals a related institutional question: are RSPPU residents learners, employees, or both? Resolving this status ambiguity will require coordinated action by the MoH and the Ministry of Higher Education, ideally through a dual-status entry in PDDikti that preserves academic benefits while reflecting the residents' employment-grade compensation. From a Kotter perspective, these tensions are characteristic of steps 7 and 8. The reform has generated visible wins and structural commitments, but cultural and bureaucratic embedding will require sustained leadership attention through at least the first graduating cohort in 2028.

XSL•FO
RenderX

Policy Implications

The reform's domestic durability hinges on five concrete actions: (1) issuance of a national clinical-educator compensation framework anchored in BPJS Health case-mix reimbursement, with a minimum hourly teaching credit codified in a Permenkes successor regulation; (2) full PDDikti integration with an explicit dual learner-employee status to resolve resident benefit ambiguity; (3) regional decentralization of selection examination hubs to neutralize Java-centric travel costs; (4) statutory consolidation of the integrity-pact return-of-service obligation into the new specialist license cycle so that it is enforceable through licensure renewal; and (5) a national tracking platform for RSPPU graduates' geographic deployment, retention beyond the initial bond, and downstream clinical outcomes in destination districts.

Four design lessons are transferable to LMICs facing the same triple workforce challenge of undersupply, maldistribution, and structurally unpaid residency, though direct replication is not implied. (1) Anchor reform in primary legislation, not subordinate regulation alone. Indonesia's Law 17/2023 created the political foundation under which negotiated compromises could be operationalized through successor regulations. Lower legal instruments are reversible by a single minister. (2) Resolve the resident-status anomaly first. Removing financial barriers is the largest single equity lever and the most politically defensible against reversal. (3) Build dual accreditation deliberately. Treating ACGME-I and a strong national accreditor as complementary rather than competing reduces foreign-perception risk while preserving sovereignty. (4) Engineer equity at entry, not at graduation. A binding integrity pact at admission is administratively simpler and politically more durable than postgraduation deployment schemes, which have failed in multiple LMICs including Indonesia's own invalidated WKDS [40] and earlier Malaysian iterations [21].

Strengths and Limitations

Strengths

The integrative design captures policy-process complexity unavailable in single-method studies; member checking with 7 of 34 informants and triangulation across documentary and interview streams reduce interpretive bias; Kotter operationalization provides an auditable analytic frame; and reporting follows the Standards for Reporting Qualitative Research [33].

Limitations

Four limitations warrant acknowledgment. First, the study covers only the first 18 months of program operation; effectiveness on graduate output, geographic retention, and clinical outcomes will require longitudinal cohort tracking. A separate study has been initiated and is expected to report in 2028. Second, residents' lived experience is captured at a single early point; structured longitudinal follow-up is planned. Third, this is a single-country case, and cross-LMIC comparative evaluation is needed to test transferability claims empirically. Fourth, while triangulation was applied, some members of the authoring team are professionally affiliated with the reform's implementation.

Conclusions

Indonesia's dual hospital-university residency model addresses the nation's specialist shortage and geographic maldistribution through structural rather than incremental reform. By designating national referral hospitals as co-organizers of training under collegium-led standards, the RSPPU pathway eliminates tuition barriers, provides living stipends, and directs deployment to underserved regions through a binding integrity pact. Early implementation across 6 pilot sites demonstrates operational feasibility: residents are salaried professionals trained in high-volume clinical settings, supported by a digital platform for transparent selection, competency tracking, workload monitoring, and anonymous well-being reporting. Remaining challenges, including accreditation adaptation, educator incentives, learner-database integration, and centralized examination logistics, are operational and tractable rather than conceptual.

Pending longitudinal evaluation of graduate output, geographic retention, and clinical outcomes, this model offers a transferable design template for LMICs confronting parallel workforce crises, though it is not yet a definitively replicable blueprint. Whether the reform delivers on its equity promise will be answered not by its design but by whether the first 52 residents and the cohorts that follow practice in the districts where Indonesia's health system most needs them.

Acknowledgments

The authors acknowledge Mr Budi Gunadi Sadikin, the Minister of Health, for his political stewardship of the reform; the Accreditation Council for Graduate Medical Education (ACGME)–International team for technical workshops on graduate medical education governance; the Human Resources for Health Policy and Strategy Team of the Center for Health Resource System Policy; the Indonesian Health Council and the national specialty collegia in orthopedics, ophthalmology, oncology, pediatrics, cardiology, and neurology for their insights into early-phase implementation; and the leadership, faculty, and inaugural cohort of 52 residents at the 6 pilot Rumah Sakit Pendidikan Penyelenggara Utama (primary teaching hospital; RSPPUs; Cicendo Eye Hospital, National Cardiovascular Center Hospital Harapan Kita, National Women and Children Health Center Hospital Harapan Kita, Prof Dr Soeharso Orthopedic Hospital, National Brain Center Hospital Mahar Mardjono, and Dharmais Cancer Hospital) for their institutional and personal contributions to this evaluation. No AI tools were used to generate scientific content; DeepSeek AI-assisted language editing was limited to grammatical refinement of author-drafted text.

XSL•FO
RenderX

Funding

The authors declared no financial support was received for this work.

Data Availability

The legal and regulatory documents analyzed in this study are publicly available from official Government of Indonesia portals, as cited in the reference list. Interview and focus group transcripts cannot be shared publicly because participants consented only to anonymized use, and the material contains potentially identifying institutional information; deidentified thematic summaries are available from the corresponding author on reasonable request.

Authors' Contributions

A Abdullah, MS, and NHAS conceived and designed the study. FRM, LT, ERW, CGM, AK, and FA led the drafting of the manuscript. YF, A Anaya, OFR, A Akemat, EMS, LK, SY, KTS, and ST contributed to data collection and processing. All authors critically reviewed, revised, and approved the final manuscript and agree to be accountable for all aspects of the work.

Conflicts of Interest

A Abdullah, YF, LT, AK, OFR, ERW, MS, CGM, NHAS, AA, EMS, LK, SY, and FA are employed by the Indonesian Ministry of Health (MoH) and were involved in the policy design described in this manuscript. ARC Institute Indonesia (independent of the policy implementation chain) was embedded in the analytic process. No financial conflicts of interest are declared.

References

1. Kurniati A, Efendi F, Widowati AR, et al. What medical specialists want to stay in remote areas of Indonesia: discrete choice experiments. PLOS One 2024;19(8):e0308225. [doi: 10.1371/journal.pone.0308225] [Medline: 39146287]
2. Direktorat Jenderal SDM Kesehatan, Kementerian Kesehatan Indonesia. Peta rasio dokter spesialis terhadap populasi [Article in Indonesian]. Indikator SDM Kesehatan. 2025. URL: https://eplanning-ditjennakes.kemkes.go.id/indikator/user/rasio_populasi_sps [accessed 2025-09-30]
3. Pendidikan Dokter Spesialis Berbasis Rumah Sakit Resmi Diluncurkan Presiden Joko Widodo [Article in Indonesian]. Kementerian Kesehatan Republik Indonesia. 2024. URL: https://kemkes.go.id/id/pendidikan-dokter-spesialis-berbasis-rumah-sakit-resmi-diluncurkan-presiden-joko-widodo [accessed 2024-05-16]
4. Perencanaan Nasional Tenaga Medis Dan Tenaga Kesehatan 2023-2032: Health Workforce National Planning [Book in Indonesian]: Ministry of Health Indonesia; 2024. URL: https://repositori-ditjen-nakes.kemkes.go.id/785/ [accessed 2026-08-11]
5. Setiawan A. Revolusi Pelayanan Kesehatan Indonesia dengan Program PPDS hospital based [Article in Indonesian]. iNDONESIAGOID. URL: https://indonesia.go.id/kategori/editorial/8397/revolusi-pelayanan-kesehatan-indonesia-dengan-program-ppds-hospital-based [accessed 2025-09-29]
6. Liu J, Eggleston K. The association between health workforce and health outcomes: a cross-country econometric study. Soc Indic Res 2022;163(2):609-632. [doi: 10.1007/s11205-022-02910-z] [Medline: 35310535]
7. Muharram FR, Sulistya HA, Swannjo JB, et al. Adequacy and distribution of the health workforce in Indonesia. WHO South East Asia J Public Health 2024 Jul 1;13(2):45-55. [doi: 10.4103/WHO-SEAJPH.WHO-SEAJPH_28_24] [Medline: 39995001]
8. Tjokroprawiro BA, Sulistya HA, Muharram FR, et al. Geospatial access to emergency obstetric surgery in Indonesia: is travel time for access too long? Ann Glob Health 2024;90(1):82. [doi: 10.5334/aogh.4598] [Medline: 39758809]
9. Indonesia - health systems strengthening project. : World Bank Group; 2023 URL: https://documents.worldbank.org/en/publication/documents-reports/documentdetail/099111723114093413 [accessed 2026-08-11]
10. Noncommunicable diseases Indonesia 2018 country profile. : World Health Organization; 2018 URL: https://cdn.who.int/media/docs/default-source/country-profiles/ncds/idn_en.pdf [accessed 2025-12-16]
11. Johnston KJ, Wen H, Joynt Maddox KE. Lack of access to specialists associated with mortality and preventable hospitalizations of rural Medicare beneficiaries. Health Aff (Millwood) 2019 Dec 1;38(12):1993-2002. [doi: 10.1377/hlthaff.2019.00838] [Medline: 31794307]
12. Laditka JN, Laditka SB, Probst JC. Health care access in rural areas: evidence that hospitalization for ambulatory care-sensitive conditions in the United States may increase with the level of rurality. Health Place 2009 Sep;15(3):731-740. [doi: 10.1016/j.healthplace.2008.12.007] [Medline: 19211295]
13. Konsil Kesehatan Indonesia [Article in Indonesian]. 2025. URL: https://www.kki.go.id/ [accessed 2026-02-18]
14. Mustika R, Nishigori H, Ronokusumo S, Scherpbier A. The odyssey of medical education in Indonesia. TAPS 2019 Jan 2;4(1):4-8. [doi: 10.29060/TAPS.2019-4-1/GP1077]
15. Program baru PPDS: pendidikan spesialis digaji dan dibiayai pemerintah [Article in Indonesian]. Kementerian Kesehatan Republik Indonesia. 2025. URL: https://kemkes.go.id/id/program-baru-ppds-pendidikan-spesialis-digaji-dan-dibiayai-pemerintah [accessed 2025-12-16]

XSL•FO
RenderX

16. Guide to the common program requirements (residency). Version 5.2. : Accreditation Council for Graduate Medical Education (ACGME); 2025 URL: https://www.acgme.org/globalassets/pdfs/guide-to-the-common-program-requirements-residency.pdf [accessed 2026-08-11]
17. Pathways in medicine. Australian Medical Association (AMA). 2020. URL: https://www.ama.com.au/pathways-in-medicine [accessed 2025-12-16]
18. Contracts. British Medical Association (BMA). URL: https://www.bma.org.uk/pay-and-contracts/contracts [accessed 2025-12-16]
19. Enari T, Hashimoto H. Does salary affect the choice of residency in non-university teaching hospitals? A panel analysis of Japan Residency Matching Programme data. Hum Resour Health 2013 Mar 8;11:12. [doi: 10.1186/1478-4491-11-12] [Medline: 23496935]
20. Hoff RG, Frenkel J, Imhof SM, Ten Cate O. Flexibility in postgraduate medical training in the Netherlands. Acad Med 2018 Mar;93(Supplement_1):S32-S36. [doi: 10.1097/ACM.0000000000002078] [Medline: 29485485]
21. Arif MT. Postgraduate medical education: future needs and challenges [Article in Malay]. Academy of Medicine of Malaysia (AMM). 2002. URL: https://www.acadmed.org.my/newsmaster.cfm?action=view&menuid=3&retrieveid=2 [accessed 2025-12-16]
22. Lukman K. Establishment of hospital-based surgical residency programs as a health policy for surgical needs provision: an indonesian perspective. Ann Med Surg (Lond) 2023 Sep;85(9):4643-4645. [doi: 10.1097/MS9.0000000000001130] [Medline: 37663682]
23. Wakaba M, Mbindyo P, Ochieng J, et al. The public sector nursing workforce in Kenya: a county-level analysis. Hum Resour Health 2014 Jan 27;12:6. [doi: 10.1186/1478-4491-12-6] [Medline: 24467776]
24. Putra ST. Dealing with shortage of medical doctors in Indonesia. Kompas.id. 2022. URL: https://www.kompas.id/baca/english/2022/07/20/dealing-with-shortage-of-medical-doctors-in-indonesia [accessed 2025-12-16]
25. Undang-Undang Republik Indonesia Nomor 17 Tahun 2023 Tentang Kesehatan [Article in Indonesian]. 2023. URL: https://www.kemkes.go.id/id/undang-undang-republik-indonesia-nomor-17-tahun-2023-tentang-kesehatan [accessed 2026-08-11]
26. Arlinta D. First batch of hospital-based PPDS participants begin educational orientation [Article in Indonesian]. Kompas.id. 2025. URL: https://www.kompas.id/artikel/en-peserta-ppds-berbasis-rumah-sakit-angkatan-pertama-mulai-ikuti-orientasi-pendidikan [accessed 2025-12-16]
27. Kotter JP. Leading change: why transformation efforts fail. IEEE Eng Manag Rev 2009;37(3):42-48. [doi: 10.1109/EMR.2009.5235501]
28. Haas MRC, Munzer BW, Santen SA, et al. #DidacticsRevolution: applying Kotter's 8-step change management model to residency didactics. West J Emerg Med 2019 Dec 19;21(1):65-70. [doi: 10.5811/westjem.2019.11.44510] [Medline: 31913821]
29. Dagnone JD, Taylor D, Acker A, et al. Seven ways to get a grip on implementing Competency-Based Medical Education at the program level. Can Med Educ J 2020 Sep;11(5):e92-e96. [doi: 10.36834/cmej.68221] [Medline: 33062098]
30. Peraturan Pemerintah Nomor 28 Tahun 2024 [Article in Indonesian]. Kemenkes. URL: https://jdih.kemkes.go.id/documents/peraturan-pemerintah-nomor-28-tahun-2024 [accessed 2026-08-17]
31. Peraturan Menteri Kesehatan Nomor 14 Tahun 2024 [Article in Indonesian]. Kemenkes. URL: https://jdih.kemkes.go.id/documents/peraturan-menteri-kesehatan-nomor-14-tahun-2024 [accessed 2026-08-17]
32. Fereday J, Muir-Cochrane E. Demonstrating rigor using thematic analysis: a hybrid approach of inductive and deductive coding and theme development. Int J Qual Methods 2006 Mar;5(1):80-92. [doi: 10.1177/160940690600500107]
33. O'Brien BC, Harris IB, Beckman TJ, Reed DA, Cook DA. Standards for Reporting Qualitative Research: a synthesis of recommendations. Acad Med 2014 Sep;89(9):1245-1251. [doi: 10.1097/ACM.0000000000000388] [Medline: 24979285]
34. Pedoman dan standar etik penelitian dan pengembangan kesehatan nasional. Kementerian Kesehatan Republik Indonesia. URL: https://admkep.unpad.ac.id/referensietik/BUKU-PEDOMAN-DAN-STANDAR-ETIK-PENELITIAN-DAN-PENGEMBANGAN-KESEHATAN-NASIONAL.pdf [accessed 2026-09-08]
35. Sejarah Kebijakan Pendidikan Dokter Umum dan Dokter Spesialis: Apakah terjadi "History in Making" karena adanya UU Kesehatan 2023? [Article in Indonesian]. Universitas Gadjah Mada, Pusat Kebijakan dan Manajemen Kesehatan, Divisi Pelatihan. URL: https://diklatkesehatan.net/2025/01/sejarah-kebijakan-pendidikan-dokter-umum-dan-dokter-spesialis-apakah-terjadi-history-in-making-karena-adanya-uu-kesehatan-2023/
36. Undang-Undang Nomor 29 Tahun 2004 tentang Praktik Kedokteran [Report in Indonesian]. : Government of the Republic of Indonesia; 2004 URL: https://farmalkes.kemkes.go.id/unduh/uu-29-2004/ [accessed 2026-08-17]
37. Undang-Undang Nomor 12 Tahun 2012 [Article in Indonesian]. Hukum Online.com. URL: https://www.hukumonline.com/pusatdata/detail/lt504716e4cc717/undang-undang-nomor-12-tahun-2012/document/ [accessed 2026-08-17]
38. Undang-Undang Nomor 20 Tahun 2013 [Article in Indonesian]. Hukum Online.com. URL: https://www.hukumonline.com/pusatdata/detail/lt522d44ca9e9a5/undang-undang-nomor-20-tahun-2013/ [accessed 2026-08-17]
39. Herbert CP, Busing N, Nasmith L. Collaborative governance of postgraduate medical education: can it be achieved? Med Teach 2021 Dec;43(12):1413-1418. [doi: 10.1080/0142159X.2021.1951691] [Medline: 34292796]

XSL•FO
RenderX

40. Putusan nomor 62 p/HUM/2018 [Article in Indonesian]. Direktori Putusan Mahkamah Agung Republik Indonesia. 2018. URL: https://putusan3.mahkamahagung.go.id/direktori/putusan/87225be24412b5807284c8011e1e80ee.html [accessed 2025-12-16]

Abbreviations

ACGME-I: Accreditation Council for Graduate Medical Education–International
BPJS: Badan Penyelenggara Jaminan Sosial (National Social Security Agency)
CCC: Clinical Competency Committee
DIO: designated institutional official
DTPK: Daerah Tertinggal, Perbatasan, dan Kepulauan (underdeveloped, border, and island regions)
FGD: focus group discussion
IAAHEH: Indonesian Accreditation Agency for Higher Education in Health (Lembaga Akreditasi Mandiri Pendidikan Tinggi Kesehatan Indonesia [LAM-PTKes])
IDI: Ikatan Dokter Indonesia
KII: key informant interview
KKI: Konsil Kesehatan Indonesia (Indonesian Health Council)
LMIC: low- and middle-income country
LPDP: Lembaga Pengelola Dana Pendidikan (Indonesia Endowment Fund for Education)
MMPI: Minnesota Multiphasic Personality Inventory
MoH: Ministry of Health
PDDikti: Pangkalan Data Pendidikan Tinggi (Higher Education Database)
PEC: Program Evaluation Committee
PPDS: Program Pendidikan Dokter Spesialis (specialist medical education program)
RSPPU: Rumah Sakit Pendidikan Penyelenggara Utama (primary teaching hospital)
UKNPDPD: Uji Kompetensi Nasional Peserta Didik Profesi Dokter (national competency examination)
WKDS: Wajib Kerja Dokter Spesialis (mandatory specialist service program)

Edited by X Catherine Tong; submitted 24.Dec.2025; peer-reviewed by KF Hollis, RD Kirkcaldy; revised version received 12.Jul.2026; accepted 14.Jul.2026; published 09.Sep.2026.

Please cite as:
Abdullah A, Farianti Y, Trilaksono L, Anaya A, Kurniati A, Rosyati OF, Wiyati ER, Sumiarsih M, Manik CG, Sonia NHA, Muharram FR, Simanjuntak KT, Tjandra S, Akemat A, Sihotang EM, Kuswandari L, Yunianti S, Asriwhardani F
A Dual Hospital–University Specialist Training Reform to Address Workforce Shortage and Geographic Maldistribution in Indonesia: Integrative Qualitative Policy Process Review
JMIR Med Educ 2026;12:e90270
URL: https://mededu.jmir.org/2026/1/e90270
doi:10.2196/90270

© Asnawi Abdullah, Yuli Farianti, Lupi Trilaksono, Arianti Anaya, Anna Kurniati, Oos Fatimah Rosyati, Etik Retno Wiyati, Mimi Sumiarsih, Christa Gumanti Manik, Nyiayu H A Sonia, Farizal Rizky Muharram, Kevin Tadeus Simanjuntak, Sydney Tjandra, Akemat Akemat, Else Mutiara Sihotang, Leni Kuswandari, Siti Yunianti, Fialisa Asriwhardani. Originally published in JMIR Medical Education (https://mededu.jmir.org), 9.Sep.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Education, is properly cited. The complete bibliographic information, a link to the original publication on https://mededu.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Inclusive Medical Spanish Education to Address Bias in Clinical Communication: Tutorial on Patient-Informed Module Development

Alexandra Lopez Vera, MPH, PhD

California University of Science and Medicine, 1501 Violet St, Colton, CA, United States

Corresponding Author:
Alexandra Lopez Vera, MPH, PhD
California University of Science and Medicine, 1501 Violet St, Colton, CA, United States

Related Articles:

Comment in: https://mededu.jmir.org/2026/1/e110709

Comment in: https://mededu.jmir.org/2026/1/e110302

Abstract

Spanish-speaking patients who are lesbian, gay, bisexual, transgender, queer, and other minority sexual orientations and gender identities (LGBTQ+) may encounter overlapping barriers in clinical care related to language discordance, misgendering, heteronormative questioning, and culturally unresponsive communication. These barriers can undermine trust, limit disclosure, and negatively affect care experiences and continuity of care. At the same time, medical Spanish curricula in the United States remain highly variable in content, structure, and assessment, and inclusive, bias-aware communication has not been consistently integrated into language training for health professions learners. Our prior qualitative work with Latinx, Hispanic, and Spanish-origin LGBTQ+ adults identified communication priorities directly relevant to educational design, including respectful forms of address, open-ended relationship language, broader recognition of family structures, and reduced reliance on binary or heteronormative assumptions. Our prior curricular work has also shown that structured medical Spanish instruction using self-study, faculty-led teaching, peer support, standardized patient encounters, and performance-based assessment can be implemented in undergraduate medical education. This tutorial presents a patient-informed framework for developing an inclusive medical Spanish module to address bias in clinical communication. Rather than reporting a new intervention study, it synthesizes prior qualitative findings and prior curricular experience to provide practical guidance for educators. The framework describes how to translate patient-reported communication priorities into curricular design principles, learning objectives, technology-enhanced preparatory activities, standardized patient and live-practice components, feedback and assessment strategies, and implementation planning. It argues that inclusive communication should not be treated as optional cultural content or as a narrow language add-on but as a core component of competent clinical performance in Spanish-language care. This tutorial is intended for medical educators, language educators, course directors, and simulation teams seeking to strengthen language-concordant and bias-aware communication training in medical Spanish. By linking patient-informed findings to concrete educational design choices, the framework offers a practical model for integrating inclusive communication throughout instruction, simulation, and assessment. A patient-informed approach of this kind may help health professions programs move beyond general statements about inclusion and toward more deliberate preparation of learners to communicate respectfully and effectively with Spanish-speaking LGBTQ+ patients.

(JMIR Med Educ 2026;12:e98398) doi:10.2196/98398

KEYWORDS

medical Spanish; inclusive communication; LGBTQ+ health; bias; clinical communication; language concordance; standardized patients; simulation training; digital education; curriculum development

Introduction

Bias in health care interactions can be reinforced through gendered assumptions and norms that shape how patients are perceived and treated [1]. For lesbian, gay, bisexual, transgender, queer, and other minority sexual orientations and gender identities (LGBTQ+) patients, noninclusive communication, misgendering, and other forms of exclusion may negatively affect disclosure, trust, and overall care experiences [2-6]. These challenges may become even more complex when patients must also navigate sensitive clinical conversations across language barriers [7].

XSL•FO
RenderX

This issue is especially relevant in the context of Spanish-speaking patient care. Medical Spanish education has expanded across US medical schools in response to the growing need for language-concordant care, yet programs remain highly variable in content, structure, and assessment [8-11]. Prior scholarship has shown that medical Spanish curricula can improve learner confidence, vocabulary, and communication skills through approaches such as didactic instruction, role-play, standardized patient encounters, immersion, and systems-based teaching [8-13]. However, the integration of inclusive, bias-aware communication into these curricula remains limited, despite increasing recognition that many trainees feel underprepared to care for transgender and gender-diverse populations and that this area remains insufficiently addressed in medical education [14-17]. National practice-level data reinforce this gap: in a 2025 survey of 1245 US primary care practices, most practices collected patient data on gender identity (77%) and sexual orientation (76%), yet only 34% provided LGBTQ+-focused training for clinicians, suggesting that data collection has outpaced workforce preparation for affirming communication [18].

For Spanish-speaking LGBTQ+ patients, communication barriers are not solely linguistic. They are also shaped by the intersection of gender identity, cultural context, and clinical assumptions. In our prior qualitative work, Latinx, Hispanic, and Spanish-origin LGBTQ+ adults described several recurring concerns in health care interactions, including misgendering, binary assumptions in routine questioning, noninclusive language related to sexual practices and relationships, and a lack of culturally responsive, language-concordant care [2]. The qualitative study informing this tutorial included 28 Latinx, Hispanic, and Spanish-origin LGBTQ+ adults between the ages of 19 and 43 years who participated in brief, semistructured interviews about gender-affirming language in Spanish-language clinical communication [2]. Participants were recruited through community organizations, health care providers, and word-of-mouth within Latinx, Hispanic, and Spanish-origin communities, and interviews were conducted in person, by phone, or by videoconferencing according to participant preference [2]. The study examined participants' experiences with pronoun use, binary and heteronormative language, relationship and family terminology, sexual health communication, and culturally responsive language-concordant care [2]. Participants also described clear preferences for affirming pronoun use, open-ended relationship language, broader recognition of family structures, and communication practices that reduce the burden of repeatedly explaining identity within the medical encounter [2].

In this tutorial, open-ended relationship language refers to Spanish phrasing that allows patients to describe their relationships, partners, and support systems in their own terms rather than requiring them to fit into presumed categories such as *esposo*/*esposa*, *novio*/*novia*, or *madre*/*padre*. In clinical communication, this includes using broader terms such as *pareja*, *persona de apoyo*, or *familia/red de apoyo* when the patient's relationship structure or preferred terminology has not yet been established. This approach is intended to reduce heteronormative and family-structure assumptions while preserving clinical clarity. These findings suggest that medical Spanish training should move beyond vocabulary acquisition alone and explicitly address how language choices can either reinforce or reduce exclusion in clinical care.

At the same time, our prior curricular work has demonstrated the feasibility of incorporating this content into structured medical Spanish instruction. Our previously published endocrine system module combined self-study, faculty-led sessions, standardized patient practice, peer tutor support, and performance-based assessment to help first-year medical students strengthen Spanish-language clinical communication while incorporating inclusive care principles [12]. That experience showed that blended, systems-based teaching models can provide a useful foundation for more intentional educational design in this area [12].

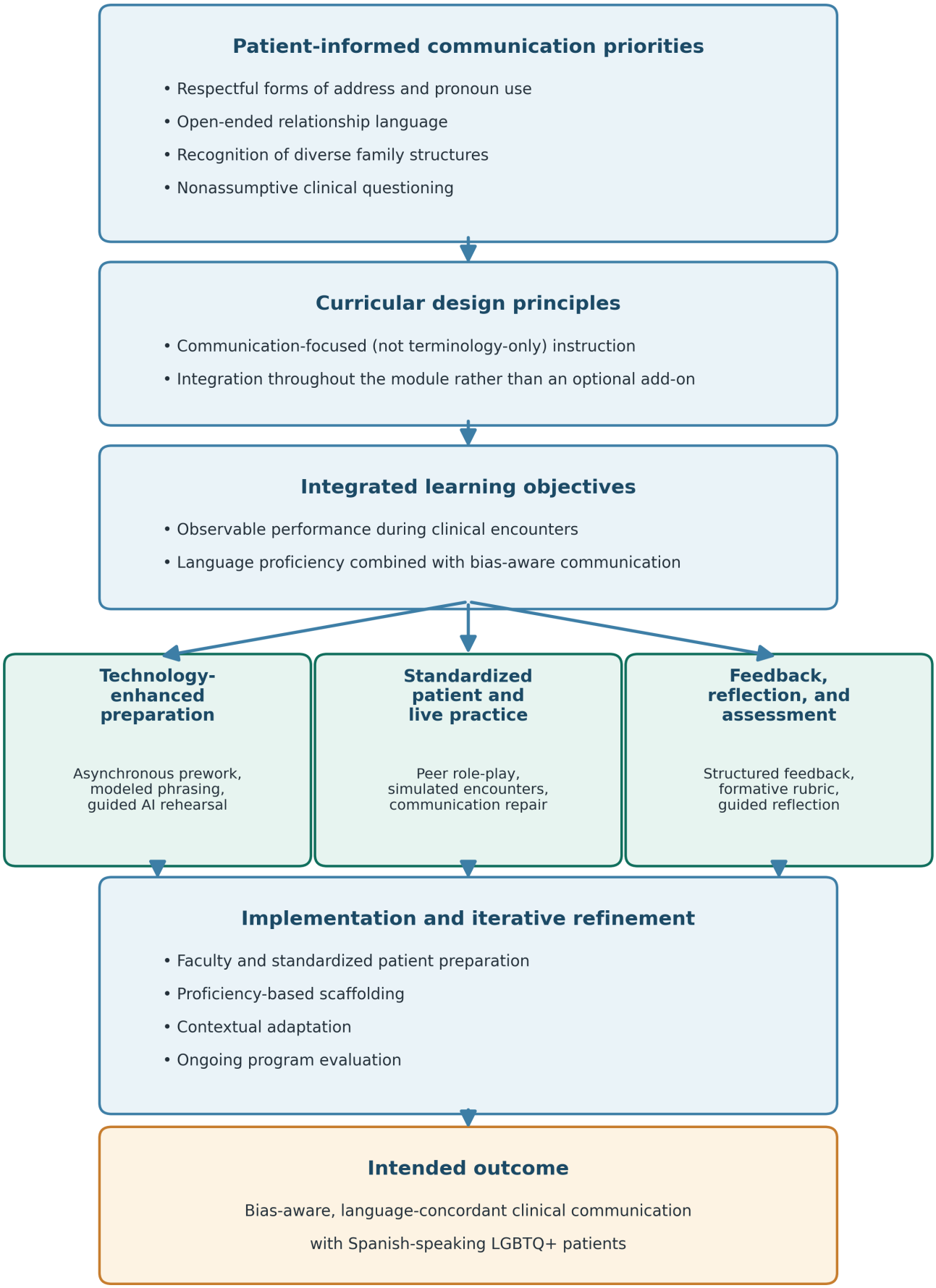
Despite these advances, educators still lack practical guidance for translating patient-informed findings on inclusive and bias-aware communication into a teachable and adaptable medical Spanish module. This gap is particularly relevant to current discussions in health professions education regarding bias, intersectionality, and the need for more intentional training approaches for inclusive care [10,14,16]. The purpose of this paper is therefore to present a patient-informed framework for developing a medical Spanish educational module that addresses bias in clinical communication with Spanish-speaking LGBTQ+ patients. Rather than reporting a new intervention study, this paper synthesizes prior qualitative findings and prior curricular experience to offer an educator-facing model for module design, implementation, and adaptation across training settings.

Visual Framework Summary

The framework presented in this tutorial can be summarized as a sequence of five linked components: (1) patient-informed communication priorities identified in prior qualitative work; (2) curricular design principles that treat inclusive communication as a clinical skill rather than an optional add-on; (3) integrated learning objectives that combine linguistic performance with bias-aware communication; (4) instructional delivery through technology-enhanced preparation, standardized patient and live practice, and structured feedback, reflection, and assessment; and (5) implementation planning with iterative refinement. Figure 1 presents this framework as a logic model linking patient-informed evidence to instructional design and to the intended outcome of bias-aware, language-concordant clinical communication.

XSL•FO
RenderX

Figure 1. Visual summary of the patient-informed framework for developing an inclusive medical Spanish module. LGBTQ+: lesbian, gay, bisexual, transgender/transsexual, queer, and other minority sexual orientations and gender identities.



XSL•FO
RenderX

From Patient-Informed Findings to Curricular Design Principles

An inclusive medical Spanish module should begin with the recognition that communication failures in this area are not limited to vocabulary gaps. They are also shaped by assumptions embedded in routine clinical language, including assumptions about gender identity, pronoun use, sexual practices, relationships, and family structure [2-6]. For Spanish-speaking LGBTQ+ patients, these challenges may be amplified when language discordance intersects with culturally unresponsive care, making it insufficient to teach terminology or memorized question patterns alone [2].

Patient-informed findings should therefore serve as a central foundation for curricular design rather than as supplemental cultural content. Our prior qualitative work identified several recurring priorities that can be translated directly into educational goals [2]. First, learners must be prepared to avoid defaulting to binary language and to ask respectfully how patients would like to be addressed. Because Spanish is a highly gendered language, inclusive medical Spanish instruction must address more than pronoun use alone. In this tutorial, "defaulting to binary language" refers to the automatic use of gendered forms of address, relationship terms, family terms, grammatical agreement, and clinical categories that presume a patient is male or female, heterosexual, cisgender, or situated within a conventional family structure. Examples include relying on *señor*/*señora* before confirming how a patient wishes to be addressed, using *esposo*/*esposa* or *novio*/*novia* when pareja would allow broader self-description, or framing reproductive and anatomical questions in ways that equate gender identity with specific organs or reproductive capacity. This issue is especially complex in Spanish because gender marking is embedded across nouns, adjectives, articles, and many routine clinical phrases. Therefore, inclusive Spanish-language clinical communication should not be reduced to substituting one term for another. Instead, educators should teach learners to ask respectfully about preferred forms of address, use neutral or open-ended terms when patient preferences are unknown, apply anatomy-specific language when clinically relevant, and adapt phrasing based on the patient's own terminology.

Second, they must learn to conduct relationship, sexual, and social history taking without presuming heterosexuality, monogamy, or conventional reproductive expectations [2-4]. Third, they must be able to use language that acknowledges broader family structures and diverse lived experiences while maintaining cultural and linguistic responsiveness [2]. These priorities are central to whether a clinical encounter feels respectful, accurate, and safe.

Educators should also anticipate regional and generational variation in Spanish itself. Gender-neutral innovations such as the pronoun *elle* and -e or -x word endings (eg, *Latine*, *Latinx*) are still evolving, and they are not universally adopted, understood, or welcomed across Spanish-speaking regions, dialects, and generations; some patients experience these forms as affirming, whereas others find them unfamiliar or externally imposed. Instructional guidance should therefore be explicit about how to navigate this variation in a clinical setting. Learners should be taught to follow the patient's own terminology rather than to impose any particular form, to ask respectfully when uncertain, and to use rephrasing strategies that avoid gender marking altogether when preferences are unknown (eg, selecting verbs and sentence structures that do not require gendered agreement). Widely understood terms such as *pareja* or *persona de apoyo* provide a practical starting point across regions. At the program level, facilitator orientation should include a brief discussion of the regional and generational variation most relevant to the local patient population, and standardized patients, community advisors, and local language experts can help adapt phrase banks to the linguistic profiles of the communities each program serves [2,12].

Translating these findings into curricular design requires a shift from terminology-focused instruction to communication-focused instruction. Medical Spanish education has demonstrated value in improving learner confidence, vocabulary, and clinical communication through approaches such as didactic teaching, role-play, systems-based instruction, and standardized patient encounters [8-12]. At the same time, the literature shows that training related to transgender and gender-diverse care remains limited across medical education, and many learners continue to feel underprepared in this area [14-17]. In this context, an inclusive medical Spanish module should not be treated as a narrow language add-on. Rather, it should be framed as a structured response to an educational gap at the intersection of language, identity, and clinical bias.

In practical terms, this means that curricular design should be organized around communication tasks that commonly become sites of exclusion in clinical encounters. These include opening the visit, asking about names and pronouns, taking a sexual history, discussing partners, exploring family context, and eliciting social history. Each of these moments should be treated as a core teaching opportunity rather than as incidental cultural content. Cases and scenarios should require learners to apply inclusive language in realistic clinical interactions rather than limiting this content to isolated vocabulary lists or brief awareness discussions [2,12].

This approach also fits well within existing medical Spanish educational models. Learners benefit from structured sequencing, repeated exposure, and multiple opportunities to practice clinical communication through self-study, faculty-led teaching, peer-supported learning, and standardized patient encounters [8-12]. Building on that foundation, an inclusive module should be intentionally sequenced so that learners first develop the language needed for respectful and nonassumptive communication and then apply that language in progressively more complex clinical contexts. In this way, patient-informed curricular design becomes the bridge between language preparation and bias-aware clinical performance.

A brief clinical vignette illustrates how these design principles converge within a single, clinically ordinary encounter in which identity-related communication is integrated into a standard presenting complaint rather than being treated as the subject of the visit (Textbox 1).

XSL•FO
RenderX

Textbox 1. Illustrative clinical vignette: integrating inclusive communication into a routine asthma follow-up visit.

A second-year medical student conducts a Spanish-language follow-up visit for asthma. The student opens with "Buenos días, soy estudiante de medicina. ¿Cómo prefiere que le llame?" and learns that the patient goes by Alex. The visit proceeds as an ordinary follow-up: inhaler technique, symptom frequency, and nighttime awakenings. When reviewing triggers in the home environment, the student asks "¿Quién vive con usted?" rather than assuming "¿Está casada?," and the patient describes a household that includes their partner and a cousin who smokes on the balcony—clinically relevant information for trigger reduction that a narrower, assumption-laden question might have missed. While updating the medication list, the student learns that the patient recently began gender-affirming hormone therapy at another clinic and acknowledges it briefly and respectfully—"Gracias por decirme; lo anoto para revisar sus medicamentos"—without making identity the focus of the visit. The encounter closes with an updated asthma action plan. Identity is never the chief complaint; it surfaces naturally, is handled respectfully, and improves the clinical quality of the visit. Full standardized patient cases built on this structure, including expected communication behaviors and common learner errors, are provided in Multimedia Appendices 1 and 2.

Learning Objectives for an Inclusive Medical Spanish Module

Learning objectives for an inclusive medical Spanish module should reflect both linguistic performance and clinical communication. If the objectives focus only on vocabulary or grammatical accuracy, they risk overlooking the very forms of exclusion the curriculum is intended to address. Communication barriers related to misgendering, heteronormative assumptions, and limited recognition of diverse family and relationship structures are well documented [2], suggesting that these areas should be treated as core instructional targets rather than as secondary cultural considerations. These objectives should also remain consistent with the broader medical Spanish literature, which shows that communication skills are most effectively developed when language instruction is embedded in clinically meaningful contexts [8-12].

One principle is especially important: objectives should be written around what learners must be able to do during an encounter, not simply what they must know. In this context, knowledge of inclusive terminology is important, but it is not sufficient. Learners must be able to introduce themselves respectfully, ask how patients would like to be addressed, obtain histories without relying on heterosexual or binary assumptions, and navigate identity-sensitive topics in Spanish with clarity and professionalism [2,12]. Framing objectives in this way keeps the curriculum grounded in observable performance rather than abstract awareness alone.

A second principle is that objectives should integrate language and bias-responsive care rather than separating them into different parts of the module. In many educational settings, language instruction and equity-oriented content are delivered on parallel tracks. For this topic, such separation is counterproductive. The aim is not simply to teach learners about inclusion and then separately teach them Spanish. It is to help them use Spanish in ways that reduce exclusion and improve the quality of clinical communication. For that reason, objectives should consistently reflect both the communicative function and the relational intent [10,12].

The following objectives may serve as a practical starting point for module design:

1. Use respectful Spanish introductions and ask patients how they would like to be addressed, including a discussion of name and pronoun preferences when clinically appropriate [2-4].
2. Conduct relationship, sexual, and social history taking in Spanish without assuming heterosexuality, monogamy, a binary gender identity, or conventional reproductive goals [2-4,12].
3. Use open-ended and inclusive Spanish language to discuss partners, family structures, and lived experiences relevant to patient care [2,4].
4. Recognize how routine clinical phrasing can communicate exclusion and revise questions in ways that are more affirming, accurate, and patient centered [2].
5. Apply inclusive communication principles in simulated clinical encounters with Spanish-speaking patients in a manner that is linguistically appropriate and culturally responsive [2,12,13].
6. Demonstrate increasing comfort with identity-sensitive communication in Spanish through repeated practice, feedback, and guided reflection [8-12].

Programs may adapt these objectives depending on learner proficiency, available instructional time, and the clinical focus of the module. For novice learners, emphasis may be placed on introductions, names and pronouns, basic nonassumptive history questions, and simple communication repair when a linguistic or interpersonal misstep occurs. For more advanced learners, the module can extend to more complex discussions of sexual health, family context, counseling, and more nuanced repair strategies. Communication repair should therefore be introduced across proficiency levels rather than reserved for advanced learners, with expectations scaffolded according to linguistic complexity. In either case, the objectives should remain closely tied to the communication priorities identified by patients and to the practical demands of real clinical encounters [2,12].

Well-designed objectives also support alignment across the rest of the curriculum. Once the objectives are clearly defined, they can guide the selection of digital prework, live teaching content, standardized patient scenarios, peer tutor activities, and performance assessments. In this way, the objectives become more than a formal requirement; they become the organizing framework for a module that is both educationally coherent and responsive to documented gaps in clinical communication training [12].

XSL•FO
RenderX

Technology-Enhanced and Preparatory Components of an Inclusive Medical Spanish Module

An inclusive medical Spanish module is well suited to a technology-enhanced format because learners benefit from repeated exposure to terminology, phrasing, and clinical scenarios before they are asked to perform in real time. In our prior curricular work, structured self-study paired with faculty-led teaching and simulation-based instruction supported vocabulary development, learner confidence, and clinical communication [11,12]. In this context, preparatory activities should do more than preview content. They should help learners arrive with sufficient familiarity to focus during live sessions on respectful language use, nonassumptive questioning, and communication repair.

Preparatory components should, therefore, be designed around high-yield communicative tasks rather than broad content review alone. Before a live session, learners may complete short asynchronous activities introducing inclusive greetings, questions about names and forms of address, open-ended relationship language, and sexual history phrasing that does not presume binary identity or a heterosexual partnership [12]. These materials can also introduce broader family language and examples of how small wording changes may shape the tone and inclusivity of a clinical encounter [2]. At this stage, the goal is not mastery. It is to provide enough baseline familiarity so that live instruction can focus on application rather than first exposure.

Technology-enhanced preparation can take multiple forms depending on institutional resources. Practical options include narrated slide presentations, short instructional videos, online glossaries, digital case prompts, learning management system modules, and recorded demonstrations of inclusive and noninclusive communication. These formats are consistent with the broader direction of medical Spanish education, which has incorporated multiple instructional modalities to support communication training [11-13]. In a module focused on inclusive and bias-aware communication, digital content should be selected not only for convenience but also for its ability to model language in context. Learners should be able to hear and see how questions are framed, how clarifying language can be used respectfully, and how patient-centered phrasing differs from more exclusionary defaults. Table 1 provides concrete examples of technology-enhanced preparatory activities, the communication focus of each, and practical implementation notes.

Table . Examples of technology-enhanced preparatory activities for an inclusive medical Spanish module.

Preparatory activity	Example	Communication focus
Narrated microlessons or short instructional videos	About 5- to 10-minute recordings that model respectful openings and preference-checking questions (eg, “¿Cómo prefiere que le llame?")	Forms of address; avoiding automatic señor/señora defaults
Recorded contrast demonstrations	Paired clips of the same intake performed with assumptive versus nonassumptive phrasing, followed by brief comparison questions	Recognizing how routine phrasing communicates inclusion or exclusion
Digital phrase banks and glossaries	Searchable phrase lists organized by communicative function (forms of address, relationships, family and support systems, sexual history, communication repair) rather than alphabetically	Open-ended relationship, family, and support-system language
Learning management system case prompts	Short written vignettes that ask learners to draft or select the next question in Spanish, with automated or faculty feedback	Nonassumptive history taking with clinical specificity
Faculty-guided AI conversational rehearsal	Structured prompts for rehearsing Spanish-language history questions with a large language model, with explicit safeguards (no real patient information, faculty-reviewed prompt sets, and reminders that AI feedback may be inaccurate or culturally limited)	Low-stakes rehearsal of phrasing, tone, and communication repair

Emerging AI tools, including large language models and conversational agents, may also support preparatory practice when used with clear instructional boundaries. For example, learners may use faculty-guided AI prompts to rehearse Spanish-language clinical questions, receive formative feedback on tone and phrasing, compare assumptive and nonassumptive wording, or practice communication repair after a linguistic or interpersonal misstep. However, AI-based practice should not replace faculty instruction, standardized patient encounters, or expert feedback. Programs should establish safeguards related to accuracy, privacy, bias, and oversight, including prohibiting the entry of real patient information, reviewing AI-generated language for clinical and linguistic appropriateness, and making clear to learners that AI feedback may be incomplete, culturally limited, or inaccurate.

Preparatory work should also be sequenced intentionally. One practical approach is to begin with foundational terminology and short phrase sets, then move to brief clinical exchanges, and finally introduce more complex scenario-based prompts. For example, learners may first review Spanish expressions for introductions and patient comfort, then practice asking about partners or family in more open-ended ways, and only afterward work through a short clinical vignette that requires them to

XSL•FO
RenderX

integrate several elements into a single interaction [2,12]. This progression is particularly useful for learners managing the combined demands of language production and clinical reasoning.

Importantly, preparatory materials should not isolate inclusive communication as a purely cultural add-on. If this content appears only in a side module or optional reading, learners may interpret it as secondary rather than clinically essential. Instead, inclusive and bias-aware communication should be embedded directly into the same materials learners use to prepare for history taking, counseling, and standardized patient encounters [2,12,13]. This helps frame inclusive language not as a special topic separate from clinical skill but as part of competent clinical communication itself.

When possible, preparatory activities should also create continuity with later in-person or synchronous instruction. The same terminology, scenarios, and communicative goals introduced asynchronously should reappear during faculty-led teaching, peer-supported practice, and simulated encounters. This alignment reinforces learning and helps ensure that digital preparation is not perceived as disconnected homework but as the first stage of a coherent instructional sequence [11-13]. In a patient-informed module, technology-enhanced preparation is most effective when it serves as a bridge between foundational language review and the more demanding interpersonal work of live clinical communication.

Standardized Patient and Live Practice Components of the Module

Live practice is essential in an inclusive medical Spanish module because many of the communication challenges identified in prior work arise during real-time interaction rather than in isolated vocabulary recall [12,13]. Misgendering, heteronormative questioning, and noninclusive assumptions about family or sexual practices do not usually occur because learners lack a single term. Rather, they emerge when learners must make rapid linguistic and interpersonal choices during a clinical encounter [2-4]. For this reason, the live portion of the module should be designed to help learners apply inclusive language within realistic clinical tasks rather than simply repeat prepared phrases.

Standardized patient encounters are particularly useful for this purpose. Our prior curricular experience demonstrated the value of structured practice with standardized patients and peer-supported rehearsal as part of systems-based instruction [11-13]. In an inclusive module, these encounters should be designed so that bias-aware communication is not peripheral to the case but integral to it. Learners should need to ask how a patient would like to be addressed, avoid assumptions about partner gender or relationship structure, and navigate identity-sensitive history taking in Spanish as part of the encounter itself [2,12]. This helps ensure that inclusive communication is taught as a component of competent clinical practice rather than as an optional addition after the main interaction.

Case design is therefore critical. A well-designed standardized patient scenario should create authentic opportunities for learners to practice inclusive language in context. This may occur during the opening of the encounter, the social history, the sexual history, or discussions involving family, support systems, or prior experiences with health care. Cases should be written carefully so that learners must gather clinically relevant information without relying on binary or heteronormative defaults [2-4]. At the same time, the case should remain clinically coherent and should not reduce LGBTQ+ identity to the sole focus of the encounter. The goal is not to create “identity cases” but to normalize inclusive communication within ordinary patient care.

Small-group live practice can further strengthen this part of the module. Learners benefit from opportunities to rehearse language with faculty, peer tutors, or both before and after standardized patient encounters [12,13]. These sessions allow learners to test phrasing, receive immediate correction, and build confidence before higher-stakes performance. In an inclusive medical Spanish module, small-group practice should focus not only on whether a sentence is grammatically correct but also on whether it is respectful, open-ended, and clinically appropriate. This distinction matters because a grammatically correct question can still communicate exclusion if it embeds the wrong assumptions.

Feedback during live practice should be structured and specific. Learners should receive guidance on forms of address, question framing, avoidance of assumptions, communication repair, and overall patient-centeredness [2,12]. Feedback is especially important when learners make errors that may not be obvious to them, such as shifting into binary phrasing after an inclusive introduction or asking follow-up questions that quietly reintroduce heterosexual assumptions. Correcting these patterns in real time can help learners understand that inclusive communication depends not only on isolated terminology but also on consistency across the encounter.

Whenever possible, the standardized patient and live practice components should be sequenced from lower-stakes to higher-stakes interactions. Learners may first practice short exchanges in pairs or small groups, then move into longer simulated encounters that require greater fluency, flexibility, and clinical reasoning [11-13]. This progression supports skill development while helping learners manage the communicative demands of the encounter more effectively. In a patient-informed module, live practice is most effective when it gives learners repeated opportunities to use inclusive Spanish in realistic settings, make mistakes, receive correction, and try again under guided conditions.

To make the tutorial more directly usable for educators, supplementary teaching materials that translate the framework into sample instructional tools are provided. Multimedia Appendix 1 includes a sample patient case focused on forms of address, pronoun-related communication, and inclusive intake language. Multimedia Appendix 2 includes a sample patient case focused on relationship language, sexual history-taking, and avoidance of heteronormative assumptions. Multimedia Appendix 3 includes a facilitator guide and sample assessment

XSL•FO
RenderX

rubric that can be adapted for peer role-play, standardized patient encounters, or formative assessment. These materials are intended as adaptable examples rather than fixed scripts, allowing programs to modify the level of linguistic complexity, clinical focus, and assessment expectations according to learner proficiency, curricular time, and local implementation needs.

Feedback, Reflection, and Assessment Strategies

Assessment in an inclusive medical Spanish module should extend beyond whether learners can recall vocabulary or complete a checklist of phrases. If the goal is to prepare learners for more respectful and bias-aware clinical communication, assessment must capture how language is used in interaction, how questions are framed, and whether learners can sustain inclusive communication across the encounter [2-4]. This is especially important because some of the most consequential communication failures in this area do not stem from the absence of Spanish proficiency but from the persistence of binary, heteronormative, or culturally unresponsive assumptions within otherwise fluent interactions [2].

For that reason, feedback should be built into the module as an ongoing instructional component rather than reserved for a final evaluation. Structured practice; performance-based assessment; and feedback from faculty, peer tutors, and standardized patients have all demonstrated value in medical Spanish education, including in our own curricular experience [11-13]. In an inclusive medical Spanish module, feedback should address both linguistic and interpersonal dimensions of performance. Learners may need correction not only on vocabulary choice or grammar but also on how they introduce themselves, whether they ask about names and forms of address appropriately, whether they default to binary or heteronormative assumptions, and whether they adjust their phrasing respectfully when clarification is needed [2,12].

Reflection can further strengthen this process. Because bias in communication is often reproduced through habitual phrasing, learners may not immediately recognize where exclusion has occurred in their own language use. Short reflective activities after simulated encounters can help learners identify moments in which their wording narrowed the encounter, introduced assumptions, or failed to align with patient preferences [2-4]. Reflection should not be treated as a substitute for skill practice, but it can help learners connect specific language choices with broader relational and clinical consequences. In this way, reflection supports the module's purpose of moving learners from awareness to more intentional communication.

Because learners at all proficiency levels may make errors in gendered or relationship language, feedback should explicitly include how learners recognize, acknowledge, and repair missteps in Spanish. At earlier levels, this may involve brief acknowledgments and clarification requests; at more advanced levels, learners can be expected to repair the interaction while maintaining the flow of clinical history taking and patient-centered communication.

Assessment strategies should be aligned directly with the module's learning objectives. If the objectives include asking respectfully about names and pronouns, conducting nonassumptive histories, and using more inclusive language around relationships and family, then those elements should appear explicitly in assessment tools [2,12]. Programs may choose to assess confidence, knowledge, and performance at different points in the module. For example, confidence surveys can help identify perceived growth in comfort with inclusive Spanish communication, whereas written or digital exercises can assess comprehension of key terminology and phrasing [12]. Performance-based assessments, however, are especially important because they allow educators to evaluate whether learners can apply these skills during interactions rather than only recognize them in theory.

Standardized patient encounters, observed role-play, and oral assessments are particularly well suited to this purpose [8-13]. These formats allow educators to evaluate not only what learners ask, but how they ask it and whether their communication remains respectful and open-ended throughout the encounter. In an inclusive medical Spanish module, assessment rubrics should therefore include criteria related to inclusive introductions, avoidance of unsupported assumptions, appropriateness of follow-up questioning, consistency in address language, and overall patient-centeredness [2,12]. Table 2 presents a simplified formative assessment rubric organized around these domains; a complete facilitator guide and expanded rubric are provided in Multimedia Appendix 3. Even when formal grading is limited, these domains can be incorporated into formative rubrics to guide feedback and make expectations explicit.

XSL•FO
RenderX

Table . Simplified formative assessment rubric for inclusive Spanish-language clinical communication.

Domain	Developing	Competent	Strong
Respectful opening and form of address	Relies on visual assumptions or inconsistent forms of address	Asks or confirms how the patient wants to be addressed and generally follows the preference	Consistently uses patient-centered forms of address from the start of the encounter
Nonassumptive relationship and family language	Uses gendered or heteronormative terms without confirming relevance	Uses broader terms (eg, pareja, red de apoyo) when preferences are unknown	Uses open-ended language naturally and follows the patient's own terminology throughout
Sexual history and clinical specificity	Avoids sensitive questions or embeds assumptions about partners or practices	Asks relevant questions with mostly inclusive, clinically clear phrasing	Balances inclusive language with precise behavior-specific and anatomy-specific questions
Communication repair	Does not recognize or repair a misstep or becomes defensive	Acknowledges correction and makes a basic repair	Repairs briefly and respectfully, then continues without centering the mistake
Spanish language effectiveness	Meaning is frequently unclear or dependent on English	Spanish is understandable and adequate for the task	Spanish is clear, organized, and appropriate to the clinical task and learner level
Patient-centeredness	Interaction feels checklist-driven or assumptive	Maintains a respectful tone; patient can provide relevant information	Creates an interaction in which the patient can self-describe and participate comfortably

Importantly, assessment in this context should aim to support growth rather than punish hesitation. Many learners will be practicing language that feels unfamiliar while also trying to avoid deeply normalized clinical defaults. A well-designed module should therefore create opportunities for low-stakes feedback before high-stakes performance. This allows learners to revise phrasing, improve consistency, and build confidence over time [8-13]. In a patient-informed and bias-responsive curriculum, the purpose of assessment is not only to document performance, but to help learners develop more respectful and clinically effective communication practices through repeated practice, targeted correction, and reflection.

Implementation Considerations and Faculty Preparation

One major consideration is faculty preparation. Instructors, peer tutors, and standardized patients should be oriented not only to the linguistic content of the module but also to its communication goals. Medical Spanish instruction often relies on multiple teaching formats and facilitators, which can strengthen learning but also create variation in how content is presented [8-12]. In an inclusive module, facilitator preparation is particularly important because learners are likely to notice inconsistencies quickly. If inclusive language is emphasized in one part of the module but absent or handled hesitantly in another, the curriculum may inadvertently signal that bias-aware communication is optional or secondary. Faculty development should therefore include guidance on core terminology, question framing, case goals, and approaches to correcting learner language in ways that are clear, respectful, and educationally constructive [2,12].

To help programs budget administrative resources, approximate time requirements can be estimated as follows. Faculty and peer-tutor orientation can usually be accomplished in a 1- to 2-hour workshop covering module goals, core phrase categories, anticipated learner errors, and rubric use; a half-day session is preferable when facilitators are new to LGBTQ+ health content or when multiple new cases will be piloted. Standardized patient training typically requires an additional 2 to 3 hours per case, including case familiarization, rehearsal of responses to assumptive versus nonassumptive questioning, and calibration of feedback. Initial development or local adaptation of preparatory materials and cases commonly requires 10 to 20 hours of faculty time, which can be substantially reduced by adapting existing materials such as those provided in Multimedia Appendices 1-3.

Learner discomfort with identity-sensitive communication deserves particular attention because topics involving gender identity and sexual orientation can encounter personal, cultural, or political resistance from some learners. Several concrete strategies can help facilitators mitigate this discomfort. First, framing matters: the module should be introduced explicitly in terms of patient safety, rapport, diagnostic accuracy, and clinical evidence—learners are being trained to obtain complete and accurate histories and to avoid communication failures that reduce disclosure—rather than as ideological content. Second, facilitators should normalize discomfort openly, acknowledging that practicing unfamiliar phrasing in a second language feels awkward for nearly all learners and should model making and repairing errors themselves. Third, when defensive reactions arise during discussion, facilitators can redirect them to the clinical task (eg, "How would you ask this so the patient can give you accurate information?") rather than debating personal beliefs, keeping attention on observable communication behaviors. Fourth, low-stakes sequencing—pair practice before group performance, and formative feedback before any graded assessment—reduces the social risk of experimentation. Finally, ground rules established at the outset, including respectful language, confidentiality of practice-room errors, and feedback directed at behavior rather than character, create the predictability that supports engagement. These strategies allow programs to preserve the module's goals while meeting learners where they are [2,12,13].

XSL•FO
RenderX

Because institutions differ in faculty expertise, learner proficiency, curricular time, simulation resources, and community partnerships, implementation should be planned with anticipated barriers in mind. Table 3 summarizes common challenges and practical strategies for adapting the module across educational settings. These strategies are intended to help programs preserve the core communication goals of the module while adjusting the instructional format to local resources and constraints.

Table . Anticipated implementation challenges and practical strategies.

Implementation challenge	Practical strategy
Limited faculty expertise in both medical Spanish and LGBTQ+[a] health	Use a coteaching model when possible, pairing medical Spanish educators with faculty or staff who have expertise in LGBTQ+ health. If co-teaching is not feasible, provide facilitator scripts, phrase banks, and a brief orientation guide before implementation [2,8-12].
Wide variation in learner Spanish proficiency	Use scaffolded materials that address the same communication goal at different proficiency levels. Novice learners can focus on respectful introductions and basic open-ended phrasing, whereas advanced learners can practice more complex sexual history, counseling, and communication repair [8-13].
Limited curricular time	Move terminology, short explanations, and preparatory examples to asynchronous prework. Reserve synchronous time for role-play, standardized patient practice, feedback, and reflection [11-13].
Lack of trained standardized patients	Begin with peer role-play, scripted cases, or facilitator-led practice. Standardized patients can be added later once case goals, feedback expectations, and assessment criteria are established [12,13].
Learner discomfort or fear of making mistakes	Normalize error and repair as part of language learning. Introduce brief repair phrases early and create low-stakes practice opportunities before formal assessment [2,12,13].
Inconsistent facilitator feedback	Use a shared facilitator guide and rubric so that instructors, peer tutors, and standardized patients evaluate the same domains, including forms of address, nonassumptive questioning, relationship language, and patient-centeredness [2,12,13].
Institutional discomfort with inclusive language content	Frame the module as patient-centered clinical communication, accuracy, trust-building, and reduction of assumptions in care, rather than as optional cultural content [2-4,10,14-17].
Limited community partnerships	Use patient-informed literature, published cases, and faculty-reviewed materials initially, while developing partnerships with community advisors or LGBTQ+ health organizations over time [2,12,13].
Risk of treating LGBTQ+ identity as the entire case	Design cases in which identity is relevant to communication but not the only clinical issue. This helps normalize inclusive communication as part of routine care rather than presenting it as a separate diversity exercise [2-4,12].
Sustainability over time	Collect learner feedback, facilitator feedback, and performance data after each implementation cycle. Revise the module iteratively instead of treating it as a one-time workshop [8-13].

[a]LGBTQ+: lesbian, gay, bisexual, transgender, queer, and other minority sexual orientations and gender identities.

A second consideration is learner proficiency. Medical Spanish courses often include students with varying levels of comfort and fluency, and this variation can influence how ambitious a module can be at different stages [8-11]. For novice learners, overly complex discussions of identity-sensitive care may become overwhelming if they are introduced before students can manage basic introductions and history-taking structures. For more advanced learners, overly simplified content may feel artificial or insufficiently clinical. Module design should therefore be calibrated carefully so that inclusive communication is taught at an appropriate linguistic level while still remaining authentic to clinical practice [12,13]. This may require scaffolded phrasing, optional enrichment material, or parallel practice tasks that allow learners to engage with the same communication goals at different levels of language complexity.

Standardized patient preparation is another critical component. Because the module depends on realistic and respectful live practice, standardized patients should understand not only the clinical details of the case but also the communication priorities being assessed. Structured standardized patient training is a key ingredient of effective medical Spanish education, as our prior curricular work has shown [12,13]. In this setting, preparation should include attention to names and pronouns, anticipated learner errors, how to respond to assumptive versus nonassumptive questioning, and how to provide useful feedback when appropriate. Cases should also avoid reducing LGBTQ+ identity to a single teaching point. Instead, identity should be

XSL•FO
RenderX

integrated into a broader clinical and social context so that the encounter reflects ordinary patient care rather than a narrowly staged diversity exercise [2,12].

Programs should also consider involving Spanish-speaking LGBTQ+ community members directly in the module rather than only teaching about them. Concrete options include recruiting community members to serve as standardized patients, or training existing standardized patients with structured community input; inviting community advisors to review case vignettes, phrase banks, and rubric language for authenticity and regional appropriateness; including guest facilitators or panelists from LGBTQ+ community organizations during live practice or debriefing sessions; and compensating community partners appropriately for their time and expertise. Beyond improving the realism of cases, this participation shifts the curriculum from studying a marginalized community to actively partnering with it, builds durable institutional relationships, and creates a mechanism for ongoing feedback as language and community norms evolve [2,12,13].

Programs should also consider where the module fits within the broader curriculum. An inclusive medical Spanish module can be delivered as a stand-alone workshop, embedded within a systems-based block, or incorporated longitudinally across a larger medical Spanish program [11,12]. The best placement depends on available curricular space and institutional priorities. Embedding the module within existing clinical content may be especially effective because it helps position inclusive communication as part of routine patient care rather than as an isolated topic. This can also make it easier to align language instruction with concurrent clinical skills teaching and reinforce the relevance of the material to patient encounters learners are already studying [11,12].

Finally, implementation should be understood as iterative. Even when a module is grounded in patient-informed findings and supported by prior curricular experience, it will still require refinement based on learner performance, facilitator feedback, and practical constraints [2,12,13]. Challenges may include limited time, uneven attendance, variation in facilitator experience, or learner discomfort with identity-sensitive communication [8-12]. These challenges do not argue against implementation. Rather, they underscore the importance of building in opportunities for revision so that the module remains responsive to both educational realities and patient-informed communication goals. In this context, implementation is most successful when it is treated not as a one-time curricular addition, but as an ongoing process of alignment between patient needs, educational design, and clinical communication training.

Discussion

Principal Considerations

This tutorial offers a patient-informed framework for developing an inclusive medical Spanish module to address bias in clinical communication with Spanish-speaking LGBTQ+ patients. Rather than treating this content as a narrow language supplement or a generic diversity add-on, the framework positions it as a structured educational response to a documented communication gap at the intersection of language, identity, culture, and clinical assumptions [2,7,12]. A central implication is that inclusive communication in medical Spanish should be taught as a clinical skill rather than as isolated terminology. Learners need repeated opportunities to apply respectful and nonassumptive language during authentic clinical tasks, including opening an encounter, asking about names and forms of address, taking social and sexual histories, discussing partners and family, and responding to patient cues in real time [2-4,12].

Comparison With Prior Work

This framework builds on 2 areas of prior scholarship. The first is the medical Spanish education literature, which has shown the value of structured language training, systems-based approaches, and performance-based assessment in preparing learners for Spanish-language patient care [8-13]. The second is the literature on transgender and gender-diverse health education, which has identified substantial gaps in learner preparation and called for more deliberate integration of inclusive content into health professions curricula [14-17]. National survey data further indicate that although most US primary care practices now collect sexual orientation and gender identity information, far fewer provide training in LGBTQ+-affirming care, underscoring the need for educational models that prepare learners before they enter practice [18]. What has been less developed is the integration of these 2 areas in a way that is explicitly informed by the communication priorities identified by Spanish-speaking patients themselves. This tutorial addresses that gap by linking patient-reported findings on exclusion and inclusive language preferences to concrete decisions about module objectives, preparatory work, live practice, feedback, and assessment [2,12]. This approach also aligns with broader calls to strengthen language-concordant care and to recognize communication as a core component of equitable health care delivery [7,10].

Practical Implications for Educators

For educators, one important implication is that inclusive communication should be embedded throughout the curriculum rather than confined to a brief cultural segment or optional add-on. Integrating this content into preparation, simulation, and assessment helps frame it as part of competent clinical performance rather than as secondary material [2,12]. This tutorial also suggests that technology-enhanced preparation, standardized patient practice, and structured feedback are particularly useful when teaching communication that requires both linguistic precision and relational awareness [8-13]. Beyond assessing individual learners, educators should also plan to evaluate the module itself over successive implementations. Established evaluation frameworks such as the Kirkpatrick model provide a practical structure for this purpose, moving from learner reaction and measured learning to observed behavior change in simulated or clinical settings and, ultimately, patient-level outcomes [19].

Limitations

This paper is a tutorial, not a report of a new intervention study. It does not present new learner outcomes or multisite implementation data. Instead, it synthesizes prior qualitative

XSL•FO
RenderX

findings and prior curricular experience to offer a practical framework for educators [2,12,13]. As a result, institutions adopting this model will still need to adapt it to their own learner population, faculty expertise, curricular structure, and available resources. In addition, because medical Spanish programs vary widely across settings, the specific balance of asynchronous preparation, live teaching, simulation, and assessment may need to be modified in different contexts. Another limitation is that no single module can fully address the broader structural and institutional forces that shape bias in health care and medical education. However, the communication patterns targeted in this framework remain important because they directly affect whether patients feel recognized, respected, and safe during clinical interactions [2].

Two further limitations deserve note. First, the assessment strategies described in this tutorial emphasize immediate formative evaluation; future implementations of this framework would benefit from longitudinal evaluation, such as reassessment of learners' inclusive communication skills 6 to 12 months after module completion, to measure skill retention and transfer to clinical practice. Second, Spanish-speaking LGBTQ+ patients are not a monolithic population. Patients hold intersecting identities related to race, ethnicity, socioeconomic status, disability, and immigration experience, and for some patients—including members of many Indigenous communities in Mexico and Central America—Spanish is itself a second language, which can add further complexity to the negotiation of gender-neutral terminology. Educators should therefore treat this framework as a starting point to be adapted in partnership with local communities rather than as a uniform template.

Conclusions

Developing an inclusive medical Spanish module to address bias in clinical communication requires more than adding terminology to an existing lesson. It requires patient-informed curricular design that recognizes how exclusion is often reproduced through routine clinical language and that prepares learners to communicate more respectfully and effectively with Spanish-speaking LGBTQ+ patients. By drawing on prior qualitative findings and prior curricular implementation experience, this tutorial offers a practical framework for educators seeking to integrate inclusive communication, technology-enhanced preparation, simulation-based practice, and performance-oriented assessment into medical Spanish instruction [2,12,13]. A patient-informed approach of this kind may help health professions programs move beyond general statements about inclusion and toward more concrete training in bias-aware clinical communication.

Acknowledgments

Generative AI (Claude; Anthropic) was used to assist with language editing, formatting, and preparation of the revision materials. All contents, including all references and citations, were reviewed and verified by the author, who takes full responsibility for the accuracy, originality, and integrity of the manuscript. No patient data were entered into any AI tool. This disclosure was also included in the cover letter accompanying the revised submission.

Funding

The authors declared no financial support was received for this work.

Data Availability

No datasets were generated or analyzed in the preparation of this tutorial.

Authors' Contributions

Conceptualization: ALV
Investigation: ALV
Methodology: ALV
Project administration: ALV
Writing – original draft: ALV
Writing – review & editing: ALV

Conflicts of Interest

None declared.

Multimedia Appendix 1
Sample patient case: inclusive intake.
[DOCX File, 13 KB - mededu_v12i1e98398_app1.docx]

Multimedia Appendix 2
Sample patient case: relationship and sexual history.
[DOCX File, 12 KB - mededu_v12i1e98398_app2.docx]

XSL•FO
RenderX

Multimedia Appendix 3
Facilitator guide and assessment rubric.
[DOCX File, 14 KB - mededu_v12i1e98398_app3.docx]

References

1. Samulowitz A, Gremyr I, Eriksson E, Hensing G. “Brave men” and “emotional women”: a theory-guided literature review on gender bias in health care and gendered norms towards patients with chronic pain. Pain Res Manag 2018;2018:6358624. [doi: 10.1155/2018/6358624] [Medline: 29682130]
2. Lopez Vera A. Addressing linguistic gaps in gender-affirming healthcare for Spanish-speaking LGBTQ+ patients: insights from a qualitative study. Hispania 2026 Mar;109(1):31-37. [doi: 10.1353/hpn.2026.a984947]
3. Goldhammer H, Malina S, Keuroghlian AS. Communicating with patients who have nonbinary gender identities. Ann Fam Med 2018 Nov;16(6):559-562. [doi: 10.1370/afm.2321] [Medline: 30420374]
4. Hagen DB, Galupo MP. Trans* Individuals’ experiences of gendered language with health care providers: recommendations for practitioners. Int J Transgenderism 2014 Jan 2;15(1):16-34. [doi: 10.1080/15532739.2014.890560]
5. Winter S, Diamond M, Green J, et al. Transgender people: health at the margins of society. Lancet 2016 Jul 23;388(10042):390-400. [doi: 10.1016/S0140-6736(16)00683-8] [Medline: 27323925]
6. Jefferson K, Neilands TB, Sevelius J. Transgender women of color: discrimination and depression symptoms. Ethn Inequal Health Soc Care 2013;6(4):121-136. [doi: 10.1108/EIHSC-08-2013-0013] [Medline: 25346778]
7. Lopez Vera A, Thomas K, Trinh C, Nausheen F. A case study of the impact of language concordance on patient care, satisfaction, and comfort with sharing sensitive information during medical care. J Immigr Minor Health 2023 Dec;25(6):1261-1269. [doi: 10.1007/s10903-023-01463-8] [Medline: 36840903]
8. Morales R, Rodriguez L, Singh A, et al. National survey of medical Spanish curriculum in U.S. medical schools. J Gen Intern Med 2015 Oct;30(10):1434-1439. [doi: 10.1007/s11606-015-3309-3] [Medline: 25862190]
9. Hardin KJ, Hardin DM Jr. Medical Spanish programs in the United States: a critical review of published studies and a proposal of best practices. Teach Learn Med 2013;25(4):306-311. [doi: 10.1080/10401334.2013.827974] [Medline: 24112199]
10. Ortega P, Pérez N, Robles B, Turmelle Y, Acosta D. Teaching medical Spanish to improve population health: evidence for incorporating language education and assessment in U.S. medical schools. Health Equity 2019;3(1):557-566. [doi: 10.1089/heq.2019.0028] [Medline: 31701080]
11. Chang KE, Lewis J, Lopez Vera A. A comprehensive medical Spanish curriculum model: the Vida Medical Spanish Curriculum. BMC Med Educ 2023 Jun 30;23(1):488. [doi: 10.1186/s12909-023-04473-0] [Medline: 37391778]
12. Lopez Vera A, Ahmad J, Ramos C, Kao K, Flores A. Integrating gender-affirming care in a medical Spanish endocrine system curriculum. MedEdPORTAL 2024;20:11456. [doi: 10.15766/mep_2374-8265.11456] [Medline: 39445117]
13. Lopez Vera A, Ahmad J, Ramos C. Enhancing clinical Spanish proficiency through standardized patient interactions: a case study on medical students’ performance. BMC Med Educ 2025 Mar 10;25(1):358. [doi: 10.1186/s12909-025-06966-6] [Medline: 40065334]
14. Hana T, Butler K, Young LT, Zamora G, Lam JSH. Transgender health in medical education. Bull World Health Organ 2021 Apr 1;99(4):296-303. [doi: 10.2471/BLT.19.249086] [Medline: 33953447]
15. Bhatt N, Cannella J, Gentile JP. Gender-affirming care for transgender patients. Innov Clin Neurosci 2022;19(4-6):23-32. [Medline: 35958971]
16. Lee JY, Rosenthal SM. Gender-affirming care of transgender and gender-diverse youth: current concepts. Annu Rev Med 2023 Jan 27;74:107-116. [doi: 10.1146/annurev-med-043021-032007] [Medline: 36260812]
17. Juarez PD, Ramesh A, Reuben JS, et al. Transforming medical education to provide gender-affirming care for transgender and gender-diverse patients: a policy brief. Ann Fam Med 2023 Feb;21(Suppl 2):S92-S94. [doi: 10.1370/afm.2926] [Medline: 36849474]
18. Akre ERL, Yang CWW, Bauer GR, et al. Sexual orientation- and gender identity-affirming activities provided in primary care. JAMA Netw Open 2025 Mar 3;8(3):e250392. [doi: 10.1001/jamanetworkopen.2025.0392] [Medline: 40063021]
19. Kirkpatrick DL, Kirkpatrick JD. Evaluating Training Programs: The Four Levels, 3rd edition: Berrett-Koehler Publishers; 2006.

Abbreviations

LGBTQ+: lesbian, gay, bisexual, transgender/transsexual, queer, and other minority sexual orientations and gender identities

XSL•FO
RenderX

Edited by M Molina; submitted 15.Apr.2026; peer-reviewed by H Rodriguez, TK Isleyici; revised version received 09.Jul.2026; accepted 25.Jul.2026; published 21.Aug.2026.

Please cite as:
Lopez Vera A
Inclusive Medical Spanish Education to Address Bias in Clinical Communication: Tutorial on Patient-Informed Module Development
JMIR Med Educ 2026;12:e98398
URL: https://mededu.jmir.org/2026/1/e98398
doi:10.2196/98398

© Alexandra Lopez Vera. Originally published in JMIR Medical Education (https://mededu.jmir.org), 21.Aug.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Education, is properly cited. The complete bibliographic information, a link to the original publication on https://mededu.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

A National Exploration of Medical Students' Views Surrounding Skin Color Terminology in Undergraduate Education

Niamh Theresa McSwiney[1,2*]; Eliza Hutchison[3,4*]; Stephanie Lax[5]; Paul Leighton[5]; Hannah Wainman[3,6]

[1]Bristol Medical School, University of Bristol, 5 Tyndall Avenue, Bristol, United Kingdom

[2]Department of Acute Medicine, University Hospitals Dorset NHS Foundation Trust, Castle Lane East, Bournemouth, Dorset, England, United Kingdom

[3]Centre for Applied Excellence in Skin and Allergy Research (CAESAR), University of Bristol, Bristol, United Kingdom

[4]Department of Dermatology, Royal United Hospital Bath NHS Trust, Bath, United Kingdom

[5]Lifespan and Population Health, School of Medicine, University of Nottingham, Nottingham, United Kingdom

[6]Department of Dermatology, University Hospitals Bristol and Weston NHS Foundation Trust, Bristol, United Kingdom

[*]these authors contributed equally

Corresponding Author:
Niamh Theresa McSwiney
Bristol Medical School, University of Bristol, 5 Tyndall Avenue, Bristol, United Kingdom

Abstract

We aimed to capture UK medical students' preferences regarding skin color terminology in medical education and in relation to how they describe their own skin tone, to better understand how we can develop more inclusive, diverse medical language for both educational and clinical settings.

(JMIR Med Educ 2026;12:e92420) doi:10.2196/92420

KEYWORDS

skin of color; medical education; dermatology; medical student; terminology

Introduction

Many clinicians lack confidence recognizing and managing dermatological disease in patients with skin of color [1]. The clinical imagery and terminology surrounding skin color often present a narrow spectrum of individual experiences. For example, the Fitzpatrick Scale was developed in 1975 to classify reactivity of White skin to phototherapy but has now been adopted as a surrogate for describing skin color [2]. A public perspective on this topic has been considered, concluding that skin color scales should better reflect the lived experience and diversity of skin tones, particularly for darker skin, and descriptors should be meaningful to the public and not directly linked to race [3]. However, medical students' voices remain lacking from these discussions. Medical students are the future of health care, medical education, and public health contributions. Therefore, their input can help shape more accurate, inclusive, and effective clinical training.

We aimed to capture UK medical students' preferences regarding skin color terminology in medical education and in relation to how they describe their own skin tone, to better understand how we can develop more inclusive, diverse medical language for both educational and clinical settings.

Methods

Overview

A 16-question (multiple-choice and free-text) e-survey was created on Jisc Online Surveys and distributed to all UK medical schools with assistance from the Medical Schools Council, informed by a previous public survey [3]. The questions focused on views regarding the Fitzpatrick scale, descriptors used to describe respondents' own skin, preferences for terminology when ethnicity is unknown, examples of inappropriate terminology in the educational setting, and subjective preparedness for clinical practice.

Survey distribution was standardized through a recruitment email, containing the study's description, a participant information sheet, the author's contact details, and a hyperlink to the online survey, disseminated to all academic year groups by each university, and remained open from October 1 to December 20, 2024. As dermatology teaching and clinical exposure vary between UK institutions in both timing and format, the authors included medical students from all academic year groups in survey participation to capture a broad, representative range of experiences. Free-text responses were not mandatory to promote survey completion. The survey was not piloted prior to nationwide distribution.

Quantitative responses were analyzed using descriptive statistics. Free-text answers were coded, categorized into themes, and

XSL•FO
RenderX

analyzed manually using Microsoft Excel (version 26.108). All 367 responses were included in the final analysis.

Ethical Considerations

A favorable ethnical opinion was granted by the University of Bristol Faculty of Health Sciences Research Ethics Committee (Ref 17622). No financial incentive was offered to the participants.

Results

There were 367 responses from medical students studying at 22 medical schools (Table 1). Of 6 free-text questions, 1503 written responses were obtained (average response rate=68%, range 32% - 100%; response length: mean 6, range 1-127 words). Most respondents were aged 19 - 24 years (69%) and female (71%)—slightly higher than the proportion of female medical school entrants nationwide (approximately two-thirds of entrants) [4]. Respondents were from diverse ethnic groups, with the ethnic distribution broadly similar to nationwide medical school demographics [4]. Participants were recruited from medical schools across the United Kingdom, with the largest proportions studying in the South West of England (22%), North East England (16%), and Scotland (13%).

Responses demonstrated that neither the pictorial nor text versions of the Fitzpatrick scale adequately represent how medical students consider their own skin, with 56% of them positioning themselves between or outside of the 6 skin tone images. Comments on the scale highlighted the need for greater representation of darker skin tones and consideration of other factors influencing skin color such as body site, ambient lighting, and sun exposure. On asking participants to describe their own skin color, the most frequent descriptors were brown (125 mentions), pale (n=123), and White (n=116). To describe their skin when inflamed or irritated, the most frequent terms were red (337 mentions), pink (n=97), brown (n=27), angry (n=22), and blotchy (n=20). The most common alternative descriptors for inflamed skin were bumpy/raised (95 mentions), dry/flaky/rough (n=91), and hot/warm (n=86). In situations when medical images are used without the individual's ethnicity being known, the most frequently suggested alternative terminology was the objective skin color (28 mentions), such as "brown," "Black," or "White." Other suggestions included the country of origin, descent, or heritage (n=19) and "light" or "dark" skin (n=7) of the person undergoing imaging. Medical students were asked whether they had perceived inappropriate terminology having been used in medical education, resulting in 48 comments. Thematic analysis generated 5 themes, which are summarized alongside representative quotes in Table 2. Many commented on the use of offensive or outdated terminology, such as the term "colored," historical labels such as "Caucasian" or "Oriental" and the use of racial or ethnic terms as nouns such as "Blacks."' Several respondents commented on a lack of variation in descriptors and highlighted skin color as a spectrum rather than just "White or Black." Generalizations such as "BAME" (Black, Asian and Minority Ethnic), "brown," and "skin of color" were perceived as restrictive and unable to capture important differences. Many reported "White skin" and "redness" predominating in their education with limited exposure to diverse imagery and language. When skin tone was discussed by educators, some felt it was delivered in a "disrespectful" and "awkward" manner, and many still "struggle to find the right terminology with darker skin."

When asked to rank their subjective preparedness for clinical practice using a 5-point Likert scale (1=not at all prepared and 5=extremely well-prepared), 265 (72%) responders reported feeling neutral to unprepared in relation to seeing patients with diverse skin tones and 274 (74%) reporting feeling neutral to underprepared in relation to discussing differences in the presentation of dermatoses in skin of color with patients.

XSL•FO
RenderX

Table . Characteristics of respondents (N=367).

Characteristics	Respondents, n (%)
Age (years)	
<18	43 (12)
19 - 24	253 (69)
25 - 34	60 (16)
≥35	8 (2)
Prefer not to say	3 (1)
Gender	
Female	261 (71)
Male	89 (24)
Other/prefer not to say	17 (5)
Ethnicity	
Asian or Asian British	110 (30)
Black, African, Caribbean or Black British	38 (10)
Mixed/Multiple Ethnic Groups	28 (8)
White	174 (47)
Other Ethnic Group	17 (5)
Location of the medical school	
England	
East of England	32 (9)
East Midlands	12 (3)
London	8 (2)
North East	57 (16)
North West	32 (9)
South East	41 (11)
South West	82 (22)
West Midlands	31 (8)
Scotland	49 (13)
Wales	23 (6)

XSL•FO
RenderX

Table . Key themes and representative quotes generated from survey.

Theme	Representative quote
Offensive and outdated terminology	"Colored - this has a dark history in relation to when the word came about and is unnecessary." [Female, 19 - 24, Black Caribbean]
	"I don't like the term 'Caucasian' being used to describe people with white skin, like myself. It is outdated and comes from a racist, hierarchical naming convention and I would much rather be called white." [Female, 25 - 34, White]
	"Depends on the situation - some individuals dislike being called 'colored' or 'Black,' and other individuals feel that these terms are acceptable. I think it may depend on the individual's experience and the behavior of the person who is referring to them (social environment, power imbalance)." [Male, 19 - 24, Mixed White and Asian]
Lack of variation in descriptors	"Sometimes skin has been referred to as black or white (such as the current general terms for race). I think feel that medically these terms should not be used to describe skin color as this is factually incorrect. No skin is either white or black. There are various shades of browns, pink and cream." [Female, 35 - 44, Mixed White and Black African]
	"Explicitly white or black skin - neither of the skin colors are true representatives; paler or darker skin color would be more acceptable as skin color is a spectrum rather than direct category" [Female, 19 - 24, Any other White background]
Generalizations	"I think the term BAME is outdated because it groups together all non-white patients when the differences between them in terms of treatment and outcomes can massively vary." [Male, 19 - 24, Mixed White and Asian]
	"Asian being used to cover East Asian, South Asian, and West Asian skin colors as a catch all, even though they are very different and not descriptive at all." [Male, 19 - 24, Asian Indian]
Representation in medical education	"Usually we are not exposed to many skin type images for conversations to even occur." [Female, 19 - 24, Asian Bangladeshi]
	"I find the use of redness to be a relic of white hegemony in medicine given that most people in the world are non-white. I have seen textbooks say conditions are harder to recognize in darker skin which is lazy and irresponsible given how many doctors abroad rely on them for medical training." [Male, 25 - 34, Black African]
Insensitive delivery	"I do think people struggle to find the right terminology with darker skin, often clinicians just want to be correct (politically and clinically)." [Female, 19 - 24, White]
	"In another lecture black scientists were randomly discussed as if to show that there are a couple of black people who have contributed to science. It is well intentioned (supposedly) but awkward and passive aggressive." [Female, 25 - 34, Asian Indian]

Discussion

This nationwide survey highlights the need for more inclusive, representative terminology for skin color in medical education. Tools such as the Fitzpatrick scale fail to reflect the full spectrum of skin tones, which is acknowledged in the literature [2]. More inclusive models are being developed, including The Eumelanin Human Skin Color Scale and the Monk Skin Tone Scale, and it is essential that these scales are adapted for and incorporated into medical education [5-7]. Aside from ethnicity, suggestions for preferred terminology were heterogeneous with objective color-based descriptors predominating, although these terms can underrepresent the full spectrum of skin tone variation. Despite recent efforts to diversify medical curricula, White-centric imagery and language still predominate, translating to many students still feeling underprepared seeing patients with diverse skin tones. When educators do discuss skin color, there are still instances of inappropriate, outdated terminology being used.

Limitations of our study include the small sample size and potential selection and recall bias. This study was not funded; hence, there was no financial incentive for medical students to participate. Incentives or more targeted recruitment strategies may have improved participation. However, to our knowledge, this is the first study exploring medical students' opinions on this topic, with UK-wide experiences captured.

We highlight the importance of developing clearer, more descriptive skin color terminology with greater representation of diverse skin tones in medical curricula. Educators should consider the language they use to describe skin color and be open to conversations with students about appropriate terminology. Future work to develop a consensus about how

XSL•FO
RenderX

we describe skin color in dermatology is essential, and it is important to consider the views of medical students as well as patients and clinicians [7]. This will not only support students' learning but also foster more equitable dermatological care for patients with diverse skin tones.

Acknowledgments

We would like to thank the Medical Schools Council for assisting in dissemination of this survey and all the medical students who completed the survey.

Funding

The authors declared no financial support was received for this work.

Conflicts of Interest

NM, EH, SL, and PL have no conflicts of interest. HW is in receipt of a doctoral fellowship from the National Institute for Health and Care Research. She has been on advisory boards for UCB Pharma and received honoraria from Leo, L''Oréal, AbbVie, Dermal, Novartis, UCB Pharma, Incyte, and Eli Lilly. She received course fees and/or travel expenses from UCB Pharma and Novartis. She is a consultant for AbbVie, Novartis, Incyte, Johnson and Jonson, and Boehringer Ingelheim.

References

1. Hutchison E, Yoseph R, Wainman H. Skin of colour: essentials for the non-dermatologist. Clin Med 2023 Jan;23(1):2-8. [doi: 10.7861/clinmed.2022-0335]
2. Okoji UK, Taylor SC, Lipoff JB. Equity in skin typing: why it is time to replace the Fitzpatrick scale. Br J Dermatol 2021 Jul;185(1):198-199. [doi: 10.1111/bjd.19932] [Medline: 33666245]
3. Lax SJ, Roberts A, Cliffe E, et al. Public views are needed for skin colour scales. Br J Dermatol 2023 Oct 25;189(5):623-624. [doi: 10.1093/bjd/ljad189] [Medline: 37287218]
4. Medical Schools Council. Fostering potential: 10 years on from selecting for excellence. Medical Schools Council. 2024. URL: https://www.medschools.ac.uk/wp-content/uploads/2025/05/fostering-potential-10-years-on-from-selecting-for-excellence.pdf [accessed 2026-07-06]
5. Dadzie OE, Sturm RA, Fajuyigbe D, Petit A, Jablonski NG. The Eumelanin Human Skin Colour Scale: a proof-of-concept study. Br J Dermatol 2022 Jul;187(1):99-104. [doi: 10.1111/bjd.21277] [Medline: 35349165]
6. Ulrich P, Zink A, Biedermann T, Sitaru S. Beyond Fitzpatrick: automated artificial intelligence-based skin tone analysis in dermatological patients. NPJ Digit Med 2025 Jun 20;8(1):378. [doi: 10.1038/s41746-025-01770-4] [Medline: 40537526]
7. Oyesiku L, Dlova N, Ahmed S, et al. Inclusive and equitable language for skin colour in the British Journal of Dermatology: acknowledging our past and laying the foundations to move forward. Br J Dermatol 2024 Apr 17;190(5):605-607. [doi: 10.1093/bjd/ljae064] [Medline: 38477822]

Abbreviations

BAME: Black, Asian and Minority Ethnic

Edited by M Molina; submitted 29.Jan.2026; peer-reviewed by J Sorensen, L Sanders, N Gulati, S Nakae; revised version received 26.May.2026; accepted 20.Jun.2026; published 14.Jul.2026.

Please cite as:
McSwiney NT, Hutchison E, Lax S, Leighton P, Wainman H
A National Exploration of Medical Students' Views Surrounding Skin Color Terminology in Undergraduate Education
JMIR Med Educ 2026;12:e92420
URL: https://mededu.jmir.org/2026/1/e92420
doi:10.2196/92420

© Niamh Theresa McSwiney, Eliza Hutchison, Stephanie Lax, Paul Leighton, Hannah Wainman. Originally published in JMIR Medical Education (https://mededu.jmir.org), 14.Jul.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Education, is properly cited. The

XSL•FO
RenderX

complete bibliographic information, a link to the original publication on https://mededu.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Gender Differences in Medical Students' Self-Assessment: Longitudinal Multi-Cohort Study

Lennart Handke, MA; Hendrik Friederichs, MME, Prof Dr Med

Medical Education Research Group, Medical School OWL, Bielefeld University, Universitätsstraße 25, Bielefeld, Germany

Corresponding Author:
Lennart Handke, MA
Medical Education Research Group, Medical School OWL, Bielefeld University, Universitätsstraße 25, Bielefeld, Germany

Abstract

Background: Self-assessment is a key requirement for lifelong learning in medicine. Evidence from gender-related research indicates that important moderators affecting self-assessment are influenced by gender. Therefore, systematic gender differences in the accuracy of self-assessment may be assumed.

Objective: The present study aimed to examine gender differences in medical students' self-assessment. Specifically, this study addressed two research questions: (1) Are there systematic gender differences in medical students' self-assessment accuracy? (2) What is the magnitude of these gender differences when accounting for academic progress and knowledge?

Methods: Medical students from 3 cohorts at Medical School OWL were surveyed in 3 waves between April 2023 and April 2024 during the Progress Test Medicine. Prior to taking the test, students were asked to indicate the percentage of Progress Test Medicine questions that they expected to answer correctly in 5 knowledge areas. Self-assessment accuracy was calculated as the difference between the subjective self-assessment and the objective test score. Linear mixed models were used to analyze the influence of gender on students' self-assessment accuracy while accounting for academic progress and knowledge.

Results: A total of 165 students with 404 data points participated in this study (269/404, 66.6% women; 135/404, 33.4% men; mean age 21.96, SD 3.61 years). Across all models, women rated themselves significantly less accurately than men (adjusted *P* values ranged from <.001 to .01). The observed gender effect ranged from –6.08 to –3.74 percentage points.

Conclusions: The results indicated systematic gender differences in medical students' self-assessment in favor of men, with a magnitude comparable to the average knowledge acquired in an entire semester of study. In view of the potentially negative consequences of inaccurate self-assessment, targeted support for realistic self-assessment during medical studies may be particularly beneficial for women.

(JMIR Med Educ 2026;12:e91071) doi:10.2196/91071

KEYWORDS

self-assessment; gender differences; lifelong learning; progress testing; medical education

Introduction

Continuing professional development requires physicians to regularly update their knowledge and acquire new clinical skills. However, this challenge begins already during undergraduate medical education, where students must also reflect on their knowledge and skills and regulate their learning accordingly (self-regulated learning) [1,2]. Self-regulated learning is defined as a "cyclical process" in which learners analyze their tasks, define goals, plan learning strategies, monitor their performance, and critically reflect on their results [3]. To meet this demand, medical students must be able to assess their knowledge and skills accurately, which is known as self-assessment [4,5]. Good self-assessment is not only essential for self-regulated learning but is also associated with higher academic performance, student motivation, engagement, and self-efficacy [6-8].

Self-assessment is a complex process influenced not only by the observation of one's own performance, knowledge, and skills but also by self-image and attitudes. Important factors influencing self-assessment include self-confidence and self-esteem, which in turn are influenced by a person's socialization and demographic variables such as gender [7,9-11]. Evidence from gender-related research indicates that women often have lower self-esteem and less confidence in their own abilities than men. In addition, women tend to attribute their successes to external factors such as favorable circumstances or luck, whereas they attribute failures to personal shortcomings [11]. Considering these gender-specific differences in central moderators of self-assessment, systematic gender differences in the accuracy of self-assessment may be assumed.

A review of the literature on gender differences in medical students' self-assessment reveals that the topic remains underexplored. For more than two decades, meta-analyses have

XSL•FO
RenderX

concluded that gender differences have not been sufficiently investigated to allow for reliable conclusions [8,12,13]. The current evidence on gender differences in self-assessment can be broadly categorized into two central questions: first, whether women and men systematically under- or overestimate themselves and, second, to what extent their self-assessments deviate from external assessments. Regarding the first question, the literature shows that women systematically underestimate their abilities. For men, in contrast, no consistent pattern emerges. The only consistent finding is that men underestimate themselves less often and overestimate themselves more frequently than women [10,11,13-15]. Regarding the second question, it is difficult to draw clear conclusions from the existing literature. Results vary considerably for both women and men, making it impossible to identify a "typical" deviation. Nevertheless, most studies indicate that the differences are overall lower for men, indicating a higher accuracy in their self-assessments [15-17]. Despite the indications of gender-specific differences in self-assessment, the overall evidence remains limited and is characterized by methodological weaknesses.

The present study aimed to examine gender differences in medical students' self-assessment and, thereby, contributes to closing the research gap discussed above. Specifically, this study addresses the following research questions:

- Are there systematic gender differences in medical students' self-assessment accuracy?
- What is the magnitude of these gender differences when accounting for academic progress and knowledge?

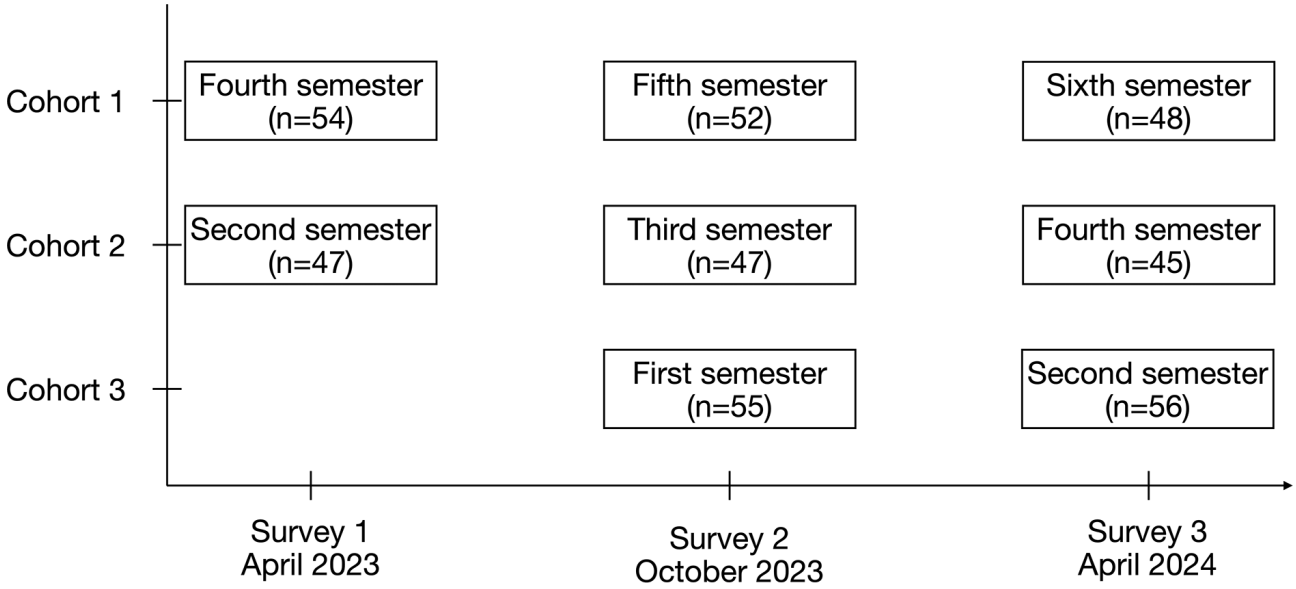
Using an objective standard of comparison and calculating standardized effect sizes, this study addresses central methodological limitations of previous research.

Methods

Study Design

To investigate a possible gender bias in medical students' self-assessment, students from 3 cohorts at Medical School OWL of Bielefeld University were surveyed between April 2023 and April 2024. The duration of undergraduate medical studies in Germany is 6 years. The last year is a practical year in clinical rotations. A total of 3 surveys were conducted during this period. Only students who were enrolled and participated in the required courses were included in this study. To obtain assessments across the first 6 semesters within a single year, an accelerated longitudinal multi-cohort design was implemented. The first cohort took part 3 times (fourth to sixth semester), the second cohort also took part 3 times (second to fourth semester), and the third cohort took part 2 times (first and second semester). The accelerated longitudinal multi-cohort design is illustrated in Figure 1. A priori power analysis was conducted using PowerLMM.js to determine the required sample size for the planned 3-wave measurement structure. We used an α level of .05, a power level of 0.80, an expected effect size of Cohen *d*=0.40, and an anticipated intraclass correlation coefficient of 0.15. On the basis of these assumptions, the required sample size was 336.

Figure 1. Accelerated longitudinal multi-cohort study design.



Data Collection and Outcome Measures

The surveys were conducted as part of the Progress Test Medicine (PTM)—developed by Charité in Berlin—a formative assessment consisting of 200 multiple-choice questions in single-best-answer format. The PTM covers the entire medical curriculum at the level of the medical state exam and is administered every 6 months to give the students a regular overview of their progress. The questions are classified by Charité according to subjects, organ systems, domains, and skills, among other aspects.

Prior to taking the test, students were asked to indicate the percentage of the PTM questions that they expected to answer correctly in 5 knowledge areas. The original wording of the self-assessment items (in German) is provided in Multimedia Appendix 1. For this study, a division into subjects (surgical and nonsurgical) and domains (diagnosis, pathogenesis, and

XSL•FO
RenderX

therapy) was used. For the differentiation between surgical and nonsurgical subjects, the medical subjects used by Charité to classify the PTM questions were grouped into these 2 superordinate categories (eg, orthopedics, surgery, and urology were classified as surgical and anatomy, internal medicine, and neurology were classified as nonsurgical). The domains were selected from the Charité classification. This division was made to reflect an established differentiation of knowledge in medicine and to have enough PTM items per area to enable statistically valid conclusions. As these are different classification criteria, there are overlaps between the knowledge areas. However, the classification reflects established distinctions in medical education and clinical practice and is therefore familiar to students. An overview of the knowledge areas and the number of associated PTM questions is provided in Multimedia Appendix 1. For this study, the students' test results were also summarized according to the 5 knowledge areas. As a result, 10 percentage values were calculated for each student: 5 for their self-assessment and 5 reflecting their actual performance in the knowledge areas. Missing data were left as "NA" and handled via listwise exclusion.

To analyze a possible gender bias, the self-assessment accuracy was calculated as the difference between the subjective self-assessment and the objective test score in the various knowledge areas. Positive values indicated an overestimation, and negative values indicated an underestimation. In addition, the absolute difference was used to map the extent of the misjudgments regardless of their direction. To avoid loss of information, self-assessment accuracy was calculated at the individual level, following the methodological recommendations of Ward et al [18].

Statistical Analysis

A linear mixed model (LMM) was used for the analysis. Gender, academic progress (operationalized as the students' semester of study), and students' medical knowledge (total test score) were used as fixed effects (independent variables), and the absolute value of self-assessment accuracy served as the dependent variable. Following the self-assessment literature, academic progress (semester of study) was used as a moderator as it is considered a more appropriate indicator of student experience than age [7,12,19]. Total score was included in the model as previous research has identified student performance as a strong predictor of self-assessment [13,20]. Gender was recorded as "woman," "man," and "nonbinary/other." To account for dependencies within repeated measures, a random intercept for each participant was included in the model. In this way, interindividual differences in the average baseline level of the dependent variable were modeled. Model diagnostics indicated no notable assumption violations. Graphical analysis of residuals and random effects suggested only minor deviations from normality, against which LMMs are considered robust [21]. No indications of heteroscedasticity were observed. The intraclass correlation coefficient exceeded the common threshold of 0.15 and, therefore, supported the use of LMMs over simple regression models [22]. Tests for multicollinearity based on variance inflation factors yielded values of 2.18 or lower, which is clearly below the common cutoff of 5 [23]. The chosen model structure allowed for the examination of both the effects of time-invariant predictors (gender) and time-variant predictors (semester and test score). In addition, LMMs are robust to unbalanced panel structures, such as those present in this dataset [24]. A separate LMM was calculated for each of the 5 knowledge areas. To control for multiple testing, Holm-Bonferroni corrections [25] were applied separately for each independent variable (sex, academic progress, and knowledge level) across the 5 models. This approach reflected the analytical structure of the analysis in which each independent factor represented a statistical family and resulted in 3 statistical families with 5 tests each [26]. All analyses were conducted in R [27] (version 4.5.2; R Foundation for Statistical Computing). The LMMs were created using *lmerTest* [28] and supplemented with the *parameters* [29] and *performance* [30] packages to calculate standardized regression coefficients and model fit indexes.

Ethical Considerations

Students were informed about the study before completing the test. Participation was voluntary, and students could opt out at any time by skipping individual questions or the self-assessment entirely. Before analysis, all data were anonymized by an independent data trustee at Bielefeld University. For this purpose, personal identifiers in the data were removed and replaced with a randomly generated 10-digit code. The research team only received the anonymized data and had no access to identifying information. This study was approved by the Medical Ethics Committee Westphalia-Lippe (2024-156-f-S) and conducted in accordance with the Declaration of Helsinki.

Results

Table 1 presents an overview of the sample composition. A total of 90.2% (165/183) of the eligible students volunteered to take part in the study: 33.9% (56/165) from the first cohort, 29.7% (49/165) from the second cohort, and 36.4% (60/165) from the third cohort. This resulted in a total of 404 data points with complete information. The average age of the students at the time of taking the test was 21.96 (SD 3.61) years. In total, 66.6% (269/404) of the participants were women; no one identified as "nonbinary/other." The gender distribution was typical for medical studies in Germany. According to information from the Federal Statistical Office, 65.5% of medical students across Germany in 2024 were women [31].

For an initial examination of whether there were gender-specific differences in students' self-assessment, the mean values for self-assessment accuracy are presented separately for women and men in Table 2. Figure 2 provides a visual representation of these results to facilitate comparisons across semesters, knowledge areas, and gender groups.

XSL•FO
RenderX

Table . Age and gender distribution of the study participants.

	All (n=404), n (%)	Men (n=135), n (%)	Women (n=269), n (%)	Age (y), mean (SD)
Semester 1	55 (13.6)	19 (14.1)	36 (13.4)	21.44 (5.12)
Semester 2	103 (25.5)	33 (24.4)	70 (26)	21.15 (3.77)
Semester 3	47 (11.6)	16 (11.9)	31 (11.5)	21.36 (2.25)
Semester 4	99 (24.5)	35 (25.9)	64 (23.8)	22.24 (2.95)
Semester 5	52 (12.9)	17 (12.6)	35 (13)	22.92 (3.3)
Semester 6	48 (11.9)	15 (11.1)	33 (12.3)	23.29 (3.29)
Total	404 (100)	135 (100)	269 (100)	21.96 (3.61)

Table . Mean value of self-assessment accuracy in the 5 knowledge areas by semester. Negative values indicate underestimation.

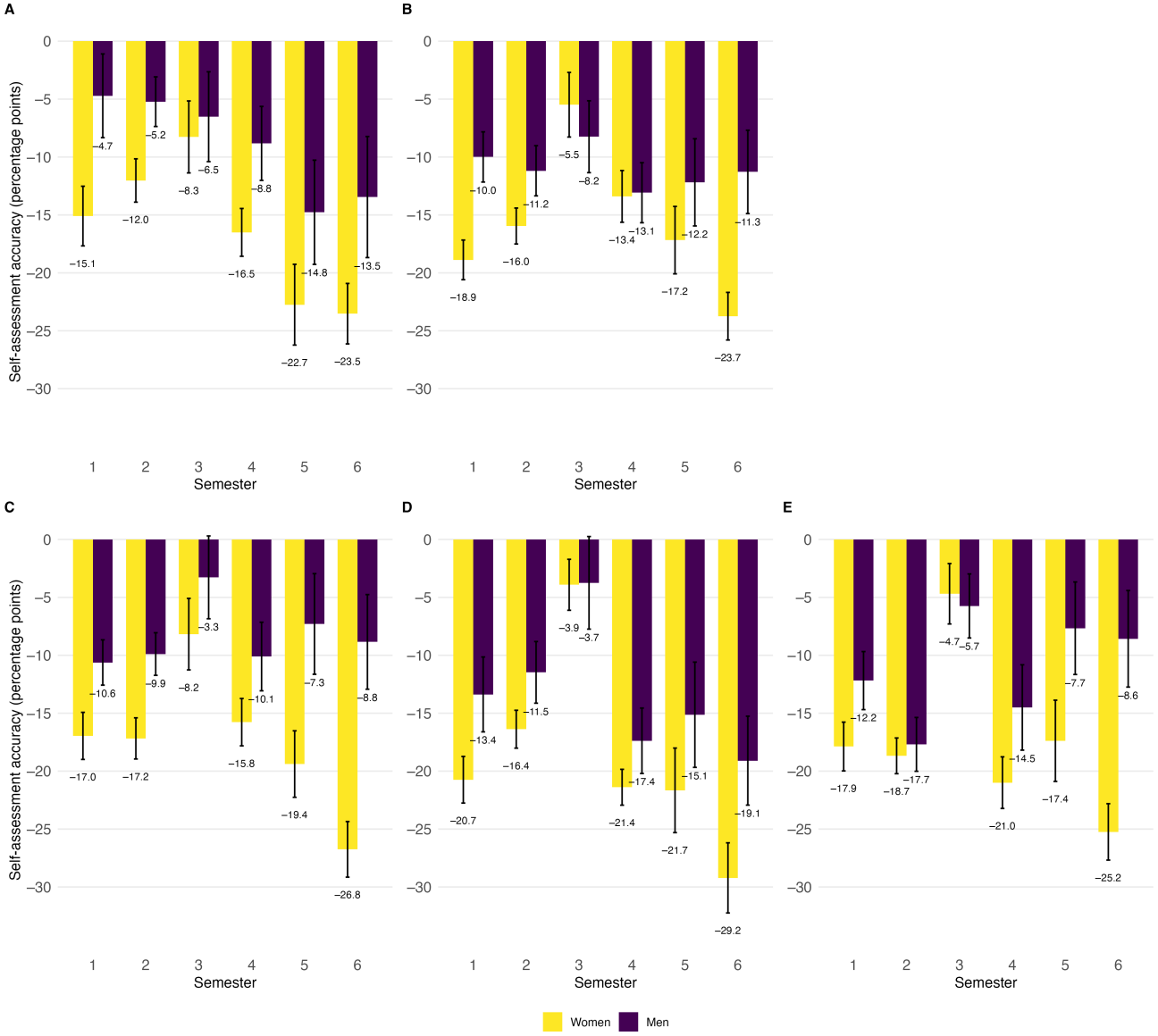
Semester	Subjects, percentage points		Domains, percentage points		
	Surgical	Nonsurgical	Diagnosis	Pathogenesis	Therapy
Women					
1	–15.1	–18.88	–16.96	–17.88	–20.75
2	–12.03	–15.96	–17.17	–18.68	–16.38
3	–8.27	–5.49	–8.17	–4.69	–3.91
4	–16.51	–13.4	–15.77	–20.99	–21.39
5	–22.75	–17.18	–19.39	–17.38	–21.66
6	–23.53	–23.74	–26.75	–25.25	–29.22
Men					
1	–4.71	–9.99	–10.62	–12.18	–13.38
2	–5.22	–11.19	–9.89	–17.69	–11.47
3	–6.52	–8.25	–3.27	–5.74	–3.74
4	–8.82	–13.08	–10.1	–14.5	–17.38
5	–14.77	–12.19	–7.29	–7.66	–15.13
6	–13.45	–11.28	–8.84	–8.58	–19.1

Both women and men underestimated themselves in all knowledge areas, but in all except 2 of the 30 subject-by-semester comparisons, women underestimated themselves more than men. The difference between the subjective self-assessment and the objective test score ranged from –29.22 to –3.91 percentage points for women and from –19.1 to –3.27 percentage points for men. Underestimation among women was greatest in the sixth semester (from –29.22 to –23.53 percentage points). The largest difference between men and women was observed in the sixth semester in the area of diagnosis, where women underestimated themselves by 17.91 percentage points more than men. The largest difference in the other direction was in the nonsurgical area in the third semester, where men underestimated themselves by 2.76 percentage points more than women. Among women, underestimation increased over the course of their studies, whereas among men, this pattern was observed only in surgical subjects.

To analyze the extent of self-assessment accuracy in more detail, Table 3 reports absolute differences. In this way, a distortion caused by the offsetting of positive and negative differences is prevented.

XSL•FO
RenderX

Figure 2. Mean self-assessment accuracy by semester, gender, and knowledge area. Panels show surgical subjects (A), nonsurgical subjects (B), diagnosis (C), therapy (D), and pathogenesis (E). Negative values indicate underestimation; error bars represent SEs.



XSL•FO
RenderX

Table . Mean of the absolute self-assessment accuracy in the 5 knowledge areas.

Semester	Subjects, percentage points		Domains, percentage points		
	Surgical	Nonsurgical	Diagnosis	Pathogenesis	Therapy
Women					
1	17.72	19.01	17.11	18.74	20.97
2	16.08	17.53	18.56	19.39	17.6
3	14.65	13.26	15.21	11.24	10.5
4	20.24	19.16	19.56	23.83	22.21
5	27.47	20.1	21.39	22.28	26.26
6	25.55	23.93	27.03	25.8	30.26
Men					
1	13.34	11.75	10.91	14.11	15.69
2	10.17	13.8	12.46	18.1	15.28
3	14.51	12.16	10.89	10.68	12.04
4	16.2	16.22	15.79	20.7	20.49
5	17.33	14.35	13.29	13.95	16.63
6	17.64	12.35	14.65	14.45	19.1

Compared with Table 2, the differences between women and men are smaller in Table 3, mainly because the absolute differences increased more strongly among men. The absolute differences ranged from 10.5 to 30.26 percentage points for women and from 10.17 to 20.7 percentage points for men. Among women, absolute differences also increased over the course of their studies. For men, the absolute differences remained more stable.

The observations from the descriptive analyses were checked using LMMs and examined with regard to significance and effect sizes. The results of the 5 LMMs are shown in Table 4. For the sake of clarity, only the standardized and unstandardized coefficients and R^2 values are reported. The complete model outputs are provided in Multimedia Appendix 1.

Table . Summary of the fixed effects of the 5 linear mixed models. The reference category for gender is "woman."

Knowledge area	Gender			Gender, β[a]	Semester			Semester, β	Test score			Test score, β	Marginal R^2	Conditional R^2
	B [b]	Holm-Bonferroni–adjusted *P* value	*P* value		*B*	Holm-Bonferroni–adjusted *P* value	*P* value		*B*	Holm-Bonferroni–adjusted *P* value	*P* value			
Surgical	–4.88	.002	<.001	–0.18	–1.64	.005	.003	–0.21	45.66	<.001	<.001	0.60	0.26	0.452
Nonsurgical	–4.72	.002	<.001	–0.21	–2.52	<.001	<.001	–0.35	41.44	<.001	<.001	0.61	0.216	0.490
Diagnosis	–6.08	<.001	<.001	–0.23	–1.24	.03	.03	–0.16	39.63	<.001	<.001	0.53	0.225	0.493
Pathogenesis	–4.04	.01	.006	–0.15	–2.90	<.001	<.001	–0.35	43.54	<.001	<.001	0.56	0.168	0.351
Therapy	–3.74	.01	.01	–0.13	–2.84	<.001	<.001	–0.34	61.02	<.001	<.001	0.76	0.322	0.527

[a]Standardized coefficient.

[b]Unstandardized coefficient.

The marginal R^2, which indicates the variance explained by the fixed effects, ranged from 0.168 (pathogenesis) to 0.322 (therapy). Following Cohen [32], these values are to be regarded as medium (R^2>0.13) to large (R^2>0.26) effect sizes. The conditional R^2, which also takes into account the variance explanation of the random intercept, ranged from 0.351 to 0.527.

XSL•FO
RenderX

The difference between the R^2 values shows that a substantial proportion of the variance was due to interindividual differences in the initial level of self-assessment accuracy. Considering standardized coefficients, gender and semester had small (>0.1) to medium (>0.3) effects, and the actual test score had a large effect (>0.5) [32]. However, unstandardized coefficients were more suitable for interpreting the results. Across all models, gender and semester each had negative coefficients, whereas the test score had a positive coefficient. A negative coefficient here indicates a higher accuracy with increasing values of the independent variable. This is consistent with the descriptive analysis, which showed that men assessed themselves more accurately (*B*=–6.08 to –3.74). In the "diagnosis" model, for example, self-assessment and actual test score were 6.08 percentage points further apart for women than for men. In addition, self-assessment became more accurate over the course of medical studies (*B*=–2.90 to –1.24). Students' actual knowledge showed the largest coefficients across the models (*B*=39.63-61.02), whereby higher test scores were associated with lower self-assessment accuracy. For example, an increase in the test score from 40% to 50% was linked to an average decrease in self-assessment accuracy from 3.96 to 6.1 percentage points. This finding is examined in detail in the Discussion section.

Discussion

The results of this study show systematic gender differences in medical students' self-assessment. A consistent pattern emerged: in all LMMs, women rated themselves significantly less accurately than men (adjusted *P* values ranged from <.001 to .01). The difference ranged from –6.08 to –3.74 percentage points (*B*), and the effect sizes ranged from –0.23 to –0.13 (β). Notably, nearly all students underestimated themselves, indicating that these gender differences do not primarily lie in the direction of self-assessment but in the extent of inaccuracy. Thus, the results are generally consistent with the current state of research [11,13,15]. However, it should be emphasized that previous studies have only reported tendencies toward more accurate self-assessment by men. In contrast, the present results showed consistent and significant differences across all models. The magnitude of the gender effects highlights their practical relevance. Considering that medical students typically gain approximately 5% of knowledge in the PTM per semester, the observed difference in self-assessment accuracy between women and men is roughly equivalent to the knowledge acquired in an entire semester [33,34]. Although the effect size of the gender effect is statistically small, it is nevertheless meaningful from an educational perspective.

Another notable finding is the influence of students' knowledge on their self-assessment accuracy. Contrary to what the current state of research suggests, higher knowledge in this study was associated with lower self-assessment accuracy. This contrasts with previous research, which has generally suggested that self-assessment accuracy improves with increasing knowledge. One possible explanation is that students do not perceive the knowledge they have already acquired as a substantial part of the overall medical curriculum, which may lead to an underestimation of their knowledge level. In addition, as students' knowledge increases, they may also become more aware of the complexity of the medical curriculum and the limits of their own knowledge. This could lead to a situation in which the standard against which students evaluate their knowledge increases more than their perceived gain in knowledge. This, in turn, may contribute to the observed tendency to underestimate their own performance more strongly despite objectively higher levels of knowledge. In addition to this possible increase in the standard against which students evaluate their knowledge, high-achieving students may hold themselves to an even higher, potentially idealized internal standard. As a result, they may judge their own performance more strictly than lower-achieving students. This could also indicate that students make limited use of available information on their performance (eg, exams and PTM results) in their self-assessment and are therefore unable to adequately assess the extent of their knowledge. According to the model by Yan and Brown [35], independently obtaining and using feedback and information are central elements of self-assessment. Furthermore, ceiling effects may limit further increases in self-assessment for high-achieving students even when their knowledge continues to grow. Moreover, statistical phenomena such as regression toward the mean or other measurement-related artifacts may contribute to the observed counterintuitive influence of students' knowledge, although the study design was chosen to minimize these risks. However, the present data do not allow for analytical conclusions about the underlying causes, and further research is needed.

Another noteworthy finding of the descriptive analyses is that students' self-assessment accuracy initially appeared to decrease over the course of their studies (Tables 2 and 3). However, the multivariable results of the LMMs (Table 4) show that academic progress was associated with higher accuracy when knowledge and gender effects were taken into account simultaneously. At the descriptive level, the effect of academic progress was masked by the stronger opposing effect of knowledge. As students' knowledge increases over the course of their studies and knowledge has a greater effect on self-assessment accuracy than academic progress, the descriptive results suggest that self-assessment becomes less accurate over time. The comparatively small effect of academic progress may also indicate that increasing experience contributes only modestly to improvements in self-assessment and that self-assessment is not systematically developed throughout the curriculum. The observed effect of academic progress is consistent with previous research indicating that self-assessment accuracy tends to improve over time and with increasing student experience.

It is also noteworthy that the medical students assessed their knowledge very imprecisely overall. One possible explanation for this may lie in the design of this study, which uses the entire medical knowledge as a reference standard. Some authors, such as Eva and Regehr [36], argue that people generally have difficulties with realistic self-assessment, especially if the benchmark is abstract and general. The medical curriculum as a whole represents such an abstract benchmark and could therefore partly explain the observed inaccuracy.

Nevertheless, from an analytical standpoint, the use of the PTM is a strength of this study, avoiding some methodological

XSL•FO
RenderX

weaknesses common in this field of research. For example, the PTM allows for a longitudinal study as the same benchmark was used for each survey, and the results can therefore be directly compared across time points. As both the self-assessment and the objective comparative measurement were based on the percentage of correctly answered PTM questions, both values could be compared statistically without conversion. As a metric scale, it also allowed for the calculation of effect sizes that can be compared with those of other studies. Another strength of the scale lies in its unambiguous content, which leaves little room for subjective interpretation. Finally, the PTM score is not a subjective evaluation of student performance by teachers but an objective, validated measurement tool. This avoids the difficulties that are often discussed under the term "gold standard" [18,37,38].

A key limitation of this study is that it was a single-center study, so results may differ at other medical faculties (eg, due to differences in curricula or program organization). Generalizations can therefore only be made with caution. However, as progress tests are carried out at many medical faculties, comparable or multicenter studies would be desirable to verify generalizability. In addition, the assignment of PTM items to specific subjects and domains cannot always be considered strictly distinct. Some questions or answer options may address content from multiple disciplines, which may lead students to conceptually associate certain items with a different subject or domain than that specified by the Charité classification. Furthermore, the number of PTM items assigned to each subject or domain varies slightly between test administrations. As a consequence, percentage scores are not equally stable across all subjects because individual items carry greater weight in subjects with fewer questions. To mitigate these limitations, individual subjects were aggregated into superordinate knowledge areas, thereby reducing the influence of single items and buffering minor ambiguities in content classification. Furthermore, this study did not assess additional variables that could help explain the observed gender differences in self-assessment. The underlying mechanisms of these differences could therefore not be examined directly.

In view of the potentially negative consequences of inaccurate self-assessment (as described in the Introduction section), targeted support for realistic self-assessment during medical studies may be particularly beneficial for women. Implications for curriculum design could include, for example, targeted feedback, training to improve self-assessment accuracy, and gender-sensitive teaching interventions. However, as the self-assessment was inaccurate not only among women but across all students, additional measures should be implemented to help students situate their performance within the broader context of their studies. Possible approaches include regular feedback formats, transparent performance evaluations, and guided opportunities for reflection. To sustainably reduce gender disparities in self-assessment, however, it is not sufficient to implement measures exclusively at the student level. Educators, evaluators, and faculty members should also be made aware of potential gender differences and trained in how to address them. For this purpose, it would be beneficial to regularly evaluate existing gender differences at the faculty level and make them transparent to faculty members and relevant committees. Furthermore, potential causes of gender disparities, including university-level factors such as teaching, examination, and feedback formats, as well as specific strategies for reducing them, should be addressed in structured professional development programs. Such measures should be institutionally anchored at the faculty level and integrated into the curriculum to ensure that their implementation does not depend solely on the commitment of individual educators.

For future research, it would be worthwhile to analyze the mechanisms underlying the observed gender differences in more detail. In particular, studies should examine how women and men differ in their interpretation and use of performance-related feedback and how these differences contribute to the observed gender bias in students' self-assessment. In addition, future research should examine how teaching, examination, and feedback practices may unconsciously contribute to the emergence and maintenance of the observed gender differences. At the institutional level, future research should also investigate whether gender-related stereotypes or expectations are conveyed in medical studies in ways that may contribute to the development of such differences. With regard to the counterintuitive influence of students' knowledge on self-assessment accuracy, future research should investigate how students perceive the medical curriculum and conceptualize their own knowledge within it. Furthermore, future studies should explicitly model potential sociocultural factors such as achievement motivation and internalized gender roles to examine their contribution to the observed gender bias.

Acknowledgments

The authors would like to thank Yu-Mei Li for her valuable feedback and helpful suggestions during the development of this manuscript. AI-based tools (DeepL [DeepL SE] and GPT-4o [OpenAI]) were used solely for linguistic refinement. No content, analyses, or interpretations were generated by these tools.

Funding

No external funding was received for this study.

Data Availability

The datasets used and analyzed during this study are available from the corresponding author on reasonable request.

XSL•FO
RenderX

Authors' Contributions

Conceptualization: LH (lead), HF (supporting)
Data curation: LH
Formal analysis: LH (lead), HF (supporting)
Investigation: LH
Methodology: LH
Project administration: HF
Supervision: HF
Validation: LH
Visualization: LH (lead), HF (supporting)
Writing—original draft: LH
Writing—review and editing: HF

Conflicts of Interest

None declared.

Multimedia Appendix 1
Supplementary tables detailing the English translation of the self-assessment items, the composition of the knowledge areas based on the subject and domain classification used by Charité, the number of questions per area, and the complete linear mixed model results.
[DOCX File, 44 KB - mededu_v12i1e91071_app1.docx]

References

1. Langendyk V. Not knowing that they do not know: self-assessment accuracy of third-year medical students. Med Educ 2006 Feb;40(2):173-179. [doi: 10.1111/j.1365-2929.2005.02372.x] [Medline: 16451246]
2. Bußenius L, Harendza S, van den Bussche H, Selch S. Final-year medical students' self-assessment of facets of competence for beginning residents. BMC Med Educ 2022 Feb 7;22(1):82. [doi: 10.1186/s12909-021-03039-2] [Medline: 35130891]
3. Panadero E. A review of self-regulated learning: six models and four directions for research. Front Psychol 2017;8:422. [doi: 10.3389/fpsyg.2017.00422] [Medline: 28503157]
4. Colbert-Getz JM, Fleishman C, Jung J, Shilkofski N. How do gender and anxiety affect students' self-assessment and actual performance on a high-stakes clinical skills examination? Acad Med 2013 Jan;88(1):44-48. [doi: 10.1097/ACM.0b013e318276bcc4] [Medline: 23165273]
5. Sargeant J, Armson H, Chesluk B, et al. The processes and dimensions of informed self-assessment: a conceptual model. Acad Med 2010 Jul;85(7):1212-1220. [doi: 10.1097/ACM.0b013e3181d85a4e] [Medline: 20375832]
6. Yan Z, Wang X, Boud D, Lao H. The effect of self-assessment on academic performance and the role of explicitness: a meta-analysis. Assess Eval High Educ 2023;48(1):1-15. [doi: 10.1080/02602938.2021.2012644]
7. Brown GT, Harris LR. Student self-assessment. In: McMillan JH, editor. SAGE Handbook of Research on Classroom Assessment: SAGE Publications; 2013:367-393. [doi: 10.4135/9781452218649.n21]
8. Panadero E, Jonsson A, Botella J. Effects of self-assessment on self-regulated learning and self-efficacy: four meta-analyses. Educ Res Rev 2017 Nov;22:74-98. [doi: 10.1016/j.edurev.2017.08.004]
9. Kim YH, Kwon H, Lee J, Chiu CY. Why do people overestimate or underestimate their abilities? A cross-culturally valid model of cognitive and motivational processes in self-assessment biases. J Cross Cult Psychol 2016;47(9):1201-1216. [doi: 10.1177/0022022116661243]
10. Lind DS, Rekkas S, Bui V, Lam T, Beierle E, Copeland EM. Competency-based student self-assessment on a surgery rotation. J Surg Res 2002 Jun 1;105(1):31-34. [doi: 10.1006/jsre.2002.6442] [Medline: 12069498]
11. Torres-Guijarro S, Bengoechea M. Gender differential in self-assessment: a fact neglected in higher education peer and self-assessment techniques. High Educ Res Dev 2017;36(5):1072-1084. [doi: 10.1080/07294360.2016.1264372]
12. Falchikov N, Boud D. Student self-assessment in higher education: a meta-analysis. Rev Educ Res 1989;59(4):395-430. [doi: 10.3102/00346543059004395]
13. Blanch-Hartigan D. Medical students' self-assessment of performance: results from three meta-analyses. Patient Educ Couns 2011 Jul;84(1):3-9. [doi: 10.1016/j.pec.2010.06.037] [Medline: 20708898]
14. González-Betancor SM, Bolívar-Cruz A, Verano-Tacoronte D. Self-assessment accuracy in higher education: the influence of gender and performance of university students. Act Learn High Educ 2019;20(2):101-114. [doi: 10.1177/1469787417735604]
15. Madrazo L, Lee CB, McConnell M, Khamisa K. Self-assessment differences between genders in a low-stakes objective structured clinical examination (OSCE). BMC Res Notes 2018 Jun 15;11(1):393. [doi: 10.1186/s13104-018-3494-3] [Medline: 29903050]

XSL•FO
RenderX

16. Cooney CM, Aravind P, Lifchez SD, et al. Differences in operative self-assessment between male and female plastic surgery residents: a survey of 8,149 cases. Am J Surg 2021 Apr;221(4):799-803. [doi: 10.1016/j.amjsurg.2020.04.009] [Medline: 32414498]
17. Lynn DJ, Holzer C, O'Neill P. Relationships between self-assessment skills, test performance, and demographic variables in psychiatry residents. Adv Health Sci Educ Theory Pract 2006 Feb;11(1):51-60. [doi: 10.1007/s10459-005-5473-4] [Medline: 16583284]
18. Ward M, Gruppen L, Regehr G. Measuring self-assessment: current state of the art. Adv Health Sci Educ Theory Pract 2002;7(1):63-80. [doi: 10.1023/a:1014585522084] [Medline: 11912336]
19. Yan Z. Student self-assessment practices: the role of gender, school level and goal orientation. Assess Educ Princ Policy Pract 2018;25(2):183-199. [doi: 10.1080/0969594X.2016.1218324]
20. Brown GT, Andrade HL, Chen F. Accuracy in student self-assessment: directions and cautions for research. Assess Educ Princ Policy Pract 2015;22(4):444-457. [doi: 10.1080/0969594X.2014.996523]
21. McNeish D. A practical guide to selecting and blending approaches for clustered data: clustered errors, multilevel models, and fixed-effect models. Psychol Methods 2026 Apr;31(2):225-251. [doi: 10.1037/met0000620] [Medline: 37956085]
22. Jäckle S, Schärdel J. Mehrebenenanalyse. In: Jäckle S, editor. Neue Trends in den Sozialwissenschaften: Springer; 2017:147-175. [doi: 10.1007/978-3-658-17189-6_6]
23. Yu H, Jiang S, Land KC. Multicollinearity in hierarchical linear models. Soc Sci Res 2015 Sep;53:118-136. [doi: 10.1016/j.ssresearch.2015.04.008] [Medline: 26188442]
24. Cnaan A, Laird NM, Slasor P. Using the general linear mixed model to analyse unbalanced repeated measures and longitudinal data. Stat Med 1997;16(20):2349-2380. [doi: 10.1002/(sici)1097-0258(19971030)16:20<2349::aid-sim667>3.0.co;2-e] [Medline: 9351170]
25. Holm S. A simple sequentially rejective multiple test procedure. Scand J Stat 1979;6(2):65-70 [FREE Full text]
26. Anderson SF. Multiplicity in multiple regression: defining the issue, evaluating solutions, and integrating perspectives. Psychol Methods 2023 Dec;28(6):1223-1241. [doi: 10.1037/met0000457] [Medline: 35588079]
27. R Core Team. R: a language and environment for statistical computing. R Foundation for Statistical Computing. 2025. URL: https://www.r-project.org [accessed 2026-07-09]
28. Kuznetsova A, Brockhoff PB, Christensen RH. lmerTest package: tests in linear mixed effects models. J Stat Softw 2017;82(13):1-26. [doi: 10.18637/jss.v082.i13]
29. Lüdecke D, Ben-Shachar MS, Patil I, Makowski D. Extracting, computing and exploring the parameters of statistical models using R. J Open Source Softw 2020;5(53):2445. [doi: 10.21105/joss.02445]
30. Lüdecke D, Ben-Shachar MS, Patil I, Waggoner P, Makowski D. performance: an R package for assessment, comparison and testing of statistical models. J Open Source Softw 2021;6(60):3139. [doi: 10.21105/joss.03139]
31. Konjunkturindikatoren im studienfach medizin. Statistisches Bundesamt (Destatis). URL: https://www.destatis.de/DE/Themen/Wirtschaft/Konjunkturindikatoren/Lange-Reihen/Bildung/lrbil05a.html?view=main [accessed 2026-01-08]
32. Cohen J. Statistical Power Analysis for the Behavioral Sciences, 2nd edition: Routledge; 1988. [doi: 10.4324/9780203771587]
33. Nouns Z, Schauber S, Witt C, Kingreen H, Schüttpelz-Brauns K. Development of knowledge in basic sciences: a comparison of two medical curricula. Med Educ 2012 Dec;46(12):1206-1214. [doi: 10.1111/medu.12047] [Medline: 23171263]
34. Görlich D, Friederichs H. Using longitudinal progress test data to determine the effect size of learning in undergraduate medical education - a retrospective, single-center, mixed model analysis of progress testing results. Med Educ Online 2021 Dec;26(1):1972505. [doi: 10.1080/10872981.2021.1972505] [Medline: 34459724]
35. Yan Z, Brown GT. A cyclical self-assessment process: towards a model of how students engage in self-assessment. Assess Eval High Educ 2017;42(8):1247-1262. [doi: 10.1080/02602938.2016.1260091]
36. Eva KW, Regehr G. "I'll never play professional football" and other fallacies of self-assessment. J Contin Educ Health Prof 2008;28(1):14-19. [doi: 10.1002/chp.150] [Medline: 18366120]
37. Andrade HL. A critical review of research on student self-assessment. Front Educ 2019;4. [doi: 10.3389/feduc.2019.00087]
38. Karay Y, Schauber SK. A validity argument for progress testing: examining the relation between growth trajectories obtained by progress tests and national licensing examinations using a latent growth curve approach. Med Teach 2018 Nov;40(11):1123-1129. [doi: 10.1080/0142159X.2018.1472370] [Medline: 29950124]

Abbreviations

LMM: linear mixed model
PTM: Progress Test Medicine

XSL•FO
RenderX

Edited by AL Marshall; submitted 12.Jan.2026; peer-reviewed by R Wassef, R Basu; revised version received 19.May.2026; accepted 19.Jun.2026; published 17.Jul.2026.

Please cite as:
Handke L, Friederichs H
Gender Differences in Medical Students' Self-Assessment: Longitudinal Multi-Cohort Study
JMIR Med Educ 2026;12:e91071
URL: https://mededu.jmir.org/2026/1/e91071
doi:10.2196/91071

© Lennart Handke, Hendrik Friederichs. Originally published in JMIR Medical Education (https://mededu.jmir.org), 17.Jul.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Education, is properly cited. The complete bibliographic information, a link to the original publication on https://mededu.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Parenthood in US Medical Training Across Specialty Groups: Scoping Review

Zuleica M Santiago Delgado[1*], MD; Elizabeth M Vaughan[2,3], DO, MPH; Mary E Jones[4], MD; Forrest Jones[4], MD; Dylan Bridwell[5], DO, MPH; Laura R Porterfield[1*], MD, MPH

[1]Department of Family Medicine and Community Health, The University of Texas Medical Branch at Galveston, 301 University Boulevard, Galveston, TX, United States

[2]Department of Internal Medicine, The University of Texas Medical Branch at Galveston, Galveston, TX, United States

[3]Department of Medicine, Baylor College of Medicine, Houston, TX, United States

[4]Family Medicine Residency Program, Waco Family Medicine Institute, Waco, TX, United States

[5]Family Medicine and Obstetrics Fellowship, Medicos Medical Center, Memphis, TN, United States

[*]these authors contributed equally

Corresponding Author:
Laura R Porterfield, MD, MPH
Department of Family Medicine and Community Health, The University of Texas Medical Branch at Galveston, 301 University Boulevard, Galveston, TX, United States

Abstract

Background: Residency and fellowship are demanding phases characterized by intense schedules, limited autonomy, sleep deprivation, and hierarchical environments. Training years often coincide with peak reproductive age, presenting trainees with the dilemma of delaying parenthood or managing training and parenthood concurrently.

Objective: This scoping review examines recent literature on factors influencing the experiences of male and female physician trainee parents across various specialties.

Methods: Four databases were searched for studies published from January 2014 through January 2025 examining pregnancy, parenthood, or lactation outcomes among physician trainees and distinguishing between specialties. Eligible studies were screened, data extracted using predefined criteria, and findings thematically coded. Chi-square and Fisher exact tests assessed specialty-related differences for each theme.

Results: The authors identified 15,861 records, removed 6440 duplicates, and excluded 8989 records. Further, 432 full-text papers were reviewed, with 105 papers included in the final analysis: surgical (n=59), medical subspecialty (n=30), primary care (n=11), and multiple specialties (n=5). Surgical and medical subspecialty studies more often identified interpersonal and policy challenges than primary care. Surgical studies highlighted issues such as pregnancy-related bias, lack of support, rigid schedules, and training-parenthood conflict. Primary care studies more often identified conflicts between parenthood and patient care responsibilities (primary care: 3/11, 27% vs surgical: 3/59, 5% and medical subspecialty studies 0/30, 0%, *P*=.01). Negative well-being impacts from training and parenthood were reported more often in primary care (3/11, 27%) and surgical studies (13/59, 22%) than in medical subspecialties (1/30, 3%, *P*=.03).

Conclusions: Trainees face physician parenthood challenges across social ecological levels, with surgical and subspecialty trainees experiencing more systemic barriers and primary care trainees more patient care–related concerns.

Trial Registration: OSF Registries 3ufcn; https://osf.io/3ufcn

(JMIR Med Educ 2026;12:e87284) doi:10.2196/87284

KEYWORDS

parenthood; graduate medical education; residency; fellowship; lactation; pregnancy; parental leave; social ecological model; childcare; surgical training

Introduction

Graduate medical education training (residency and fellowship) represents one of the most demanding phases of a medical career, characterized by long work hours, limited autonomy, inflexible schedules, rigorous requirements, sleep deprivation, and hierarchical workplace structures [1,2]. These years often coincide with peak reproductive age, particularly for women [3]. For physicians, who already work 10 hours more per week than the average US worker, the additional responsibilities of parenthood can be difficult to integrate [4-7]. This strain is even greater for trainees, whose weekly work hours may be more

XSL•FO
RenderX

than double that of the typical US worker. Consequently, many physician trainees delay or choose against parenthood [8,9].

Decisions about timing and pursuit of parenthood are influenced by a range of factors, including individual concerns (eg, pregnancy complications, postpartum depression, and perinatal recovery), interpersonal dynamics (eg, peer resentment, supervisor bias, and guilt about neglecting family), and institutional or policy-related barriers (eg, limited parental leave and inadequate lactation accommodations) [10,11]. These challenges are shaped by the specialty context. Surgical, primary care, and nonsurgical medical subspecialties (referred to here as "medical subspecialties") have varying training demands and practice requirements, with length of training ranging from 3 - 5 years in primary care and medical subspecialties to 5 - 7 years in surgical specialties [12]. Medical students applying for residency report that their parenthood status significantly influences their specialty choice based on how "family-friendly" a specialty is; however, other factors also strongly influence specialty choice [13].

Prior reviews of physician parenthood have not compared experiences across specialty groups and have included limited male perspectives. A deeper understanding of how parenthood-related challenges differ by training environment could inform cultural, institutional, and policy strategies to more effectively support physician parents. Thus, the objective of this scoping review is to map the literature around experiences, challenges, and interactions of parenthood with training for male and female physicians in residency and fellowship as they relate to individual, interpersonal, organizational, and policy-related factors and their variation across specialty groups.

Methods

Study Design

This scoping review was conducted using the PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews) guideline [14,15]. The protocol was registered and can be found at Open Science Framework Registries [16]. Before initiating the review, a preliminary search of PROSPERO and the Cochrane Database of Systematic Reviews was conducted in January 2025 to ensure no duplicate reviews were underway. Detailed key search terms and a sample search strategy, developed in collaboration with a research librarian, are included in Multimedia Appendix 1; the Medline strategy illustrates the conceptual approach applied across databases, with database-specific adaptations to syntax and indexing. Searches were conducted for studies in English published between January 1, 2014, and January 23, 2025, in Ovid Medline, APA PsycInfo, Scopus, and CINAHL. In addition, references of relevant papers were reviewed for additional potential papers to include.

Inclusion and Exclusion Criteria

This review was structured using the population, intervention/exposure, comparator, outcome, and study design model: participants (US physician residents and fellows), intervention/exposure (parenthood, pregnancy, and lactation), comparison (surgical, primary care, and medical subspecialty), outcomes (factors influencing experiences including individual, interpersonal, organizational, policy, and interactions between training and parenthood), and study design (peer-reviewed primary studies or systematic/scoping reviews) [17]. Additional inclusion criteria were studies conducted in the United States, published January 2014 through January 2025, and having a primary outcome involving parenthood. Studies were excluded if they did not specifically address parenthood, had a nonphysician population, were conducted outside the United States, or combined outcomes of specialty groups without the ability to distinguish the outcomes. Gray literature, opinion pieces, and conference abstracts were excluded to maintain a focus on empirical findings and avoid the influence of opinion or unconfirmed data.

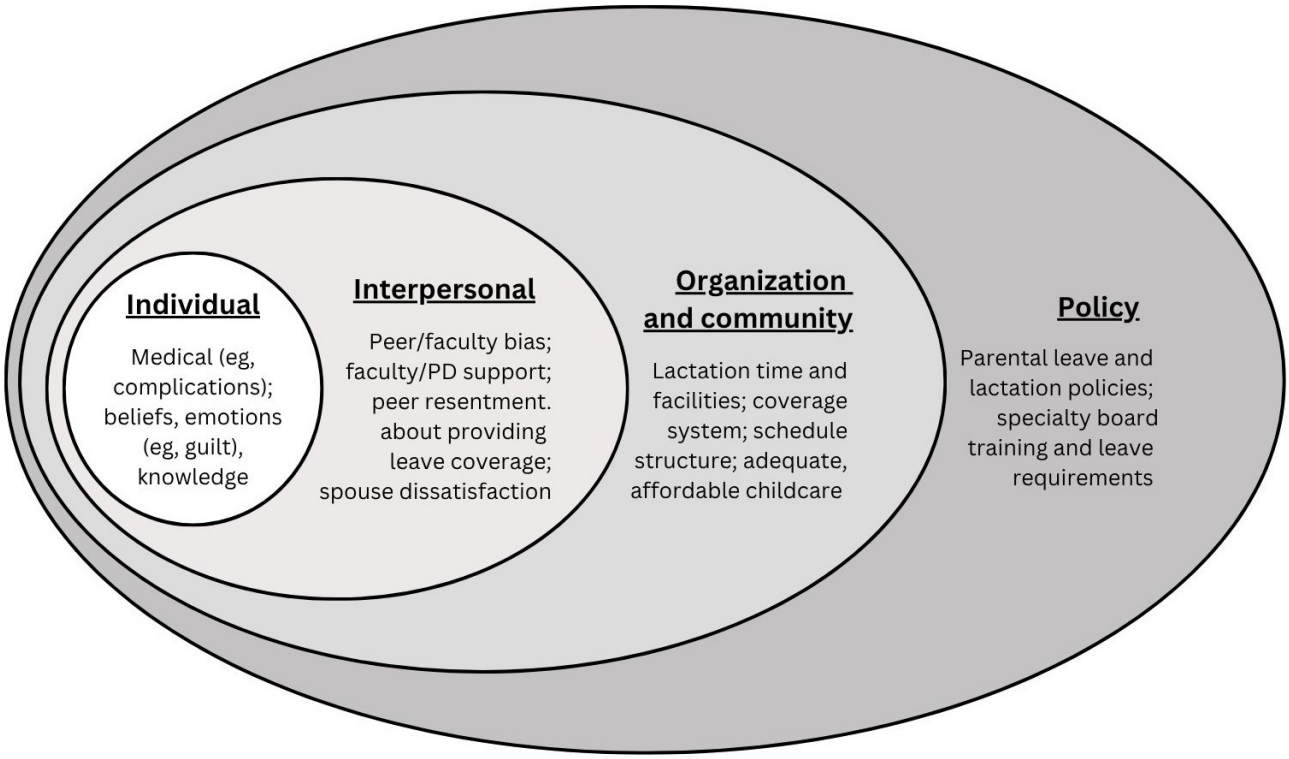
Social Ecological Model

Overview

The social ecological model (SEM) was used as a framework for understanding the factors affecting decisions about and experiences with physician parenthood. SEM organizes the complex social and environmental factors that shape behaviors and outcomes into levels of influence and explores interactions between levels [18]. For physicians experiencing or considering pregnancy and parenthood, individual factors (eg, self-doubt about ability to cope with both family and work duties and pregnancy complications), interpersonal factors (eg, the perceptions of peers and guilt about peers covering duties during parental leave), community and organizational factors (eg, requirements to make up call shifts that would have taken place during parental leave and availability of adequate childcare services and lactation facilities), and policy factors (eg, parental leave policies) can interact to affect a physician's experience with pregnancy/parenthood or a physician's decisions to postpone or avoid parenthood (see Figure 1).

XSL•FO
RenderX

Figure 1. A social ecological model of factors affecting physician parenthood. This figure illustrates a conceptual framework applying the social ecological model to experiences of physician trainee parents. It maps multilevel influences on physician parenthood across four domains: individual, interpersonal, institutional/community, and policy. PD: program director.



Study Selection

Studies were imported into Rayyan, and duplicates were excluded [19]. Two authors independently reviewed the first 500 abstracts and titles for relevance and excluded studies that did not meet the criteria. Interrater reliability for this stage was high (Cohen κ=0.79; 95% CI 0.63 - 0.96), with 98.8% agreement. Inclusion of the remaining abstracts was determined by one reviewer. Two authors then independently reviewed the full text of a subset of papers (n=50, 11.6%) for eligibility. Interrater reliability at the full-text stage was lower but remained within the range of substantial agreement (κ=0.61; 95% CI 0.36 - 0.86), with 86% agreement. Initial disagreement at the full-text stage most commonly arose in studies with complex or heterogeneous designs requiring interpretive application of prespecified inclusion criteria. These included studies in which outcomes were combined across specialty groups where it required detailed examination of data to determine whether there were separable specialty-specific results (eg, primary care trainees grouped with anesthesiology residents and medical and surgical oncology trainees analyzed together) [20,21]; studies combining United States and international trainee populations where outcomes of interest did not include stratified reporting [22]; studies in which parenthood-related variables were secondary or indirect rather than a primary focus of the study (eg, burnout studies including parental status only as a covariate) [23]; and studies focused on medical students applying to residency (as opposed to residents) [24].

In each discrepant case, reviewers reassessed eligibility using the original prespecified inclusion and exclusion criteria, with particular attention to whether physician trainee parenthood outcomes, specialty-specific analyses, and US-specific data were clearly distinguishable. No post hoc modifications to screening criteria were made during the review process. All discrepancies at both stages were resolved through structured discussion and consensus. The remaining full texts were assessed for eligibility by one author.

Data Extraction, Synthesis, and Statistical Calculations

Extracted data were incorporated into an Excel (Microsoft Corp) spreadsheet and included study details (author, year, and title); participant characteristics, demographics, and numbers; studied specialty; and study aim, design, methods, limitations, outcomes, and findings. During data extraction, study results were assigned to a corresponding level of SEM or to an overarching theme of interactions between parenthood and training or career. Data were extracted twice for 15 studies, with the remainder extracted by one author.

A convergent qualitative synthesis approach was used, with SEM serving as a framework to organize and map findings. One author conducted thematic coding to identify recurrent themes, after which themes were reviewed and verified by a second author. To explore specialty-based differences (surgical, primary care, and medical subspecialty), chi-square and/or Fisher exact tests were applied to compare the frequency of each theme by group. For themes in which positive influences on training-parenthood experiences were identified in addition to negative influences, analyses for those themes were repeated while excluding studies with exclusively positive results to highlight differences in areas of challenge across specialty groups.

XSL•FO
RenderX

Classification of Specialties

Included studies were categorized into three groups based on specialty type: surgical specialties, medical subspecialties, and primary care. This approach was selected to provide a pragmatic framework that reflects commonly used distinctions in physician workforce and residency experience research [25,26]. We further distinguished medical subspecialties from primary care to preserve clinically meaningful differences in training structure, continuity of care, and patient population. While this classification approach enhances interpretability across heterogeneous studies, it necessarily introduces within-group variability. In particular, some procedural specialties (eg, anesthesiology, interventional cardiology, and interventional gastroenterology) are grouped with primarily cognitive or consultative specialties (eg, psychiatry, neurology, and nephrology), reflecting the absence of a universally accepted classification system in the literature.

To assess the robustness of findings to alternative plausible groupings, we conducted sensitivity analyses reclassifying anesthesiology within the surgical cohort. We compared thematic and statistical outcomes under the original classification scheme (surgical, medical subspecialty, and primary care) with those generated using the alternative classification. This analysis was performed to assess the robustness of findings to plausible alternative specialty groupings and to evaluate whether anesthesiology materially influenced observed between-group differences.

Ethical Considerations and Reflexivity

This study was a scoping review of published literature and did not involve human participants or identifiable data. Therefore, institutional review board approval was not required.

The multidisciplinary author team included two medical students without lived experience of parenthood at the time of this study, one family medicine resident father, and three academic faculty (family and internal medicine) who had experienced pregnancy and/or parenthood during residency or fellowship training. Faculty authors also brought experience in residency leadership, core faculty roles, and departmental clinical leadership. These varied perspectives informed the interpretation of findings and encouraged consideration of trainee, parent, educator, and leadership viewpoints throughout the review process.

Although this study did not include a formal stakeholder consultation phase, interpretation of findings was informed through iterative discussions with family medicine program leaders and resident and former resident physician-parents across multiple specialties. Preliminary findings were also presented at the Society of Teachers of Family Medicine Annual Meeting, where feedback from educators and clinicians informed refinement of this paper and interpretation of results.

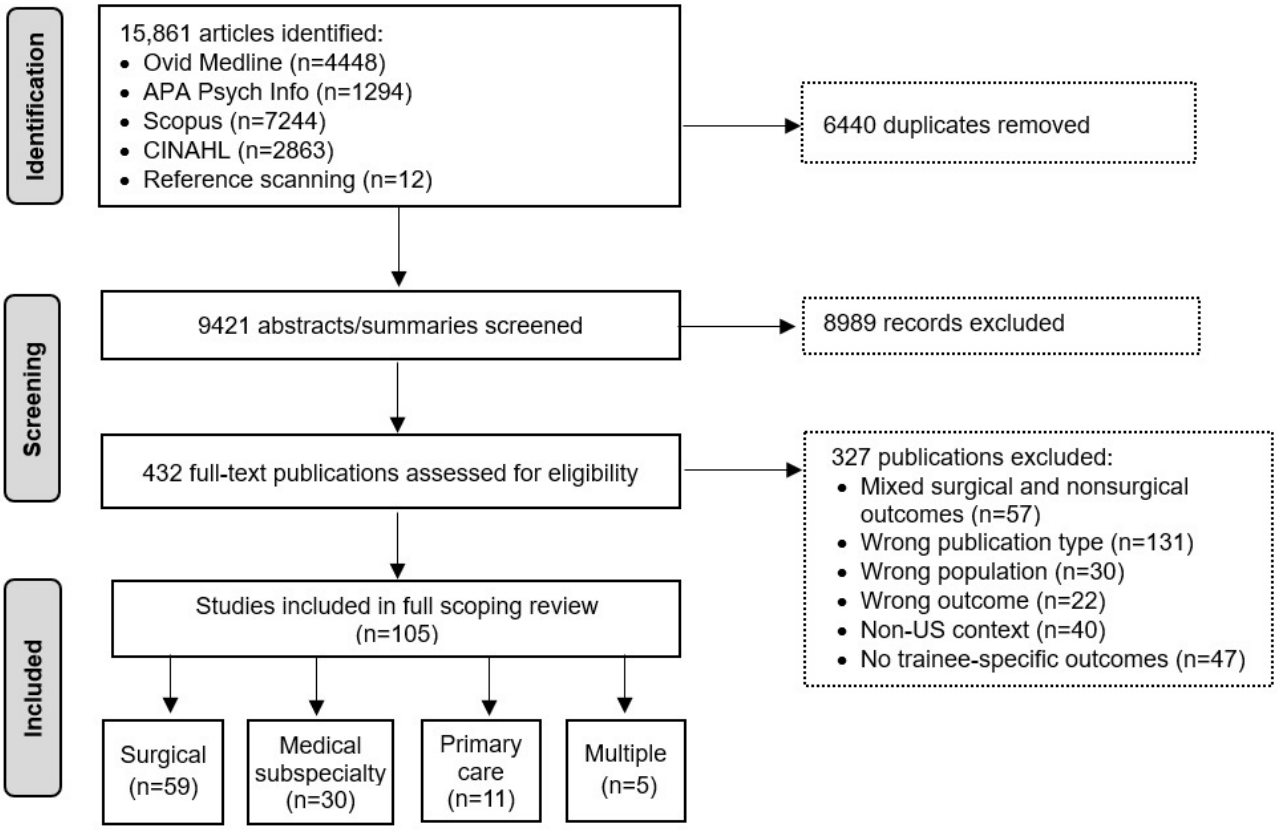
Results

Study Selection and Characteristics

A total of 15,861 papers were identified. After the elimination of 6440 duplicates, 9421 abstracts and titles were assessed for eligibility. Of these, 8989 abstracts did not meet the inclusion criteria, and 432 advanced to the full-text review phase. Of these, 105 were included, with the remaining 327 excluded for the following reasons: wrong population (n=77), lack of a parenthood-related outcome (n=22), non-US context (n=40), lack of separate specialty outcomes (n=57), and wrong publication type (eg, opinion piece; n=131). Included studies were further subdivided by surgical (n=59), medical subspecialty (n=30), primary care (n=11), or multiple specialty groups (n=5). The PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) flowchart is included in Figure 2.

XSL•FO
RenderX

Figure 2. PRISMA flow diagram of paper inclusion and exclusion process summarizes the screening and selection process for the scoping review on physician parenthood. APA: American Psychological Association; CINAHL: Cumulative Index to Nursing and Allied Health Literature; PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses.



The specialties represented in each specialty group are detailed in Table 1. Participants of included studies were predominantly trainees (n=37,986), with former trainees (n=3622) and program directors (PDs; n=3025) also represented. Trainee studies generally included both males and females (n=41 studies), but 16 studies included only females, and 2 only males. Most studies (99/105, 88%) used quantitative methods, particularly surveys (77/105, 73% of studies). Qualitative methods were used in 10/105 (10%) studies, and mixed methods in 3/105 (3%) studies. Primary parenthood-related outcomes included parental leave (38/105, 36% studies); program or colleague support of trainee parenthood or pregnancy (18/105, 17%); health or fertility outcomes (6/105, 6%); trainee parent performance (6/105, 6%); and lactation (4/105, 4%). Almost one-third of studies (33/105, 31%) explored parenthood or pregnancy outcomes across multiple domains.

XSL•FO
RenderX

Table . Characteristics of included studies.

	All studies	Surgical[a]	Primary care[b]	Medical specialty[c]	Multiple[d]
Number of participants (number of studies)	43,521 (105)	25,529 (59)	2669 (11)	15,948 (30)	375 (5)
Mean (SD)	414 (1279)	416 (1000)	242 (309)	532 (1912)	N/A[e]
Median (IQR)	153 (56-283)	163 (64.5-238.5)	64 (21.5-295.5)	121 (79-205)	N/A
Range	6 - 10,803	15 - 5692	6 - 1021	23 - 10,803	N/A
Trainees (studies)	37,986 (59)	22,417 (34)	1781 (8)	13,510 (17)	278 (1)
Female only	16	12	1	3	0
Male only	2	2	0	0	0
Female trainee range (%)	0 - 100	0 - 100	50 - 100	14 - 100	64
Graduates (studies)	3622 (18)	2441 (11)	N/A	1181 (7)	N/A
PD[f] (studies)	3025 (33)	1230 (15)	888 (4)	907 (13)	N/A
Website or policy (studies)	1530 (12)	1084 (6)	N/A	349 (2)	97 (4)
Review studies (studies)	44 (2)	44 (2)	N/A	N/A	N/A
Study methods					
Qualitative, n (%)	10/105 (10)	7	3	N/A	N/A
Focus group	3	2	1	N/A	N/A
Interview	8	5	3	N/A	N/A
Quantitative, n (%)	92/105 (88)	50	7	30	5
Survey	77	46	246	1	1
Quasi-experimental	4	N/A	2	2	N/A
Website, policy, etc	14	6	N/A	4	4
Review	2	2	N/A	N/A	N/A
Mixed, n (%)	3/105 (3)	2	1	N/A	N/A
Parenthood-related outcomes, n (%)					
Parental leave	38/105 (36)	20/59 (34)	2/11 (18)	13/30 (43)	N/A
Trainee support	18/105 (17)	9/59 (17)	2/11 (18)	6/30 (20)	N/A
Health/fertility	6/105 (6)	5/59 (8)	0/11	1/30 (3)	N/A
Performance	6/105 (6)	9/59 (15)	2/11 (18)	13/30 (43)	N/A
Lactation	4/105 (4)	1/59 (2)	1/11 (9)	1/30 (3)	N/A
Multiple	33/105 (31)	18/59 (31)	6/11 (55)	8/30 (27)	N/A

[a]Surgical (n=59): general surgery (n=18), gynecologic oncology (n=1), multiple (n=13), neurosurgery (n=4), obstetrics and gynecology (n=4), ophthalmology (n=4), orthopedic surgery (n=7), oral & maxillofacial (n=1), otolaryngology (n=1), plastic surgery (n=3), urology (n=2), and vascular surgery (n=1).

[b]Primary care (n=11): pediatrics (n=7), family medicine (n=3), and internal medicine (n=1).

[c]Medical subspecialty (n=30): anesthesiology (n=5), cardiology (n=3), dermatology (n=2), emergency medicine (n=2), gastroenterology (n=2), infectious disease (n=1), nephrology (n=1), neurology (n=2), oncology (n=1), physical medicine & rehabilitation (n=1), psychiatry (n=2), radiation oncology (n=4), and radiology (n=4).

[d]Multiple specialty groups (n=5).

[e]N/A: not applicable.

[f]PD: program director.

XSL•FO
RenderX

Details of each study's population, number of participants, percent female trainees, study design, methodology, and primary outcomes studied are included in Multimedia Appendix 2 [11,21,27-129].

Interventions

Five studies examined the effect of an intervention for trainee parents using pre- and postintervention data without a control group (Table 2). Interventions included schedule modifications, parental leave policies, parenthood electives, and lactation support. One study of emergency medicine residents instituted schedule changes, including the removal of night shifts for a pregnant resident and flexible scheduling for three expectant fathers, and found no notable negative impact on peer resident schedules [27]. Two studies evaluated the implementation of parental leave policies. An updated parental leave policy in a neurology residency program was associated with higher perceived leadership support of parental leave [28], while a comprehensive parental leave policy in a general surgery program improved perceptions of support and equity surrounding parenthood and training [29]. In a pediatrics residency program, implementation of an at-home neonatal care elective increased the minimum time taken by both child-bearing and nonchildbearing residents without affecting on-time graduation rates [30]. Another pediatrics study installed a hospital-grade pump in an equipped lactation room and reported improved resident engagement in patient care, increased comfort, and decreased guilt and anxiety among lactating residents [31].

Table 3 includes a summary of themes organized by specialty group and SEM level.

Table . Studies of interventions for trainee parents.

Study	Population (n)	Intervention	Results
Chernoby et al (2021) [27]	Emergency medicine residents (n=79)	Flexible scheduling policy (pregnant and new parent trainees, n=4 in intervention)	86% resident support for pregnancy and 91% for postpartum changes. High satisfaction from the 4 participating residents.
Conway et al (2022) [28]	Neurology residents (n=221)	Revision of parental leave policy: 12-week paid; no extension	Significant increase in perceived policy transparency, consistency, and program leadership support of parental leave.
Corbisiero et al (2024) [29]	General surgery residents (n=65)	Revision of parental leave policy, process, schedules	Significant increase in perception of policy meeting trainee needs, supporting postpartum self-care, policy fairness, and support of childbearing during residency.
Cree-Green et al (2020) [30]	Pediatrics residents (n=64)	At-home neonatal care elective	Increase in minimum time taken by childbearing residents (2 to 6 wk). A total of 79% of nonchildbearing residents took 4+ weeks compared to 0% preintervention. Decrease in unpaid leave.
Creo et al (2018) [31]	Pediatrics (n=6)	Improved lactation room with a hospital-grade pump	Pumping time decreased 8.5 min (95% CI 3.8 - 12.2, *P*=.45). Pumping volume increased 2.8 oz (95% CI 1.2 - 4.3). Increased patient care engagement, task completion, comfort; decreased anxiety/guilt.

XSL•FO
RenderX

Table . Summary of themes of included studies.

	Total (N=105), n (%)	Surgical (N=59), n (%)	PC[a] (N=11), n (%)	MS[b] (N=30), n (%)	All (N=5)	Values		Fisher exact test (*P* value)
						Chi-square (*df*)	*P* value	
Individual	4946 (46)	31 (53)	3 (27)	15 (50)	0	2.39 (2)	.30	.34
Knowledge/belief	20 (19)	14 (24)	0	6 (20)	0	3.26 (2)	.20	.18
Health	26 (25)	16 (27)	2 (18)	8 (27)	0	0.39 (2)	.82	.89
Fertility	18 (17)	11 (19)	0	7 (23)	0	3.01 (2)	.22	.22
Financial constraints	4 (4)	2 (3)	1 (9)	1 (3)	0	N/A[c]	N/A	.56
Interpersonal	68 (65)	42 (71)	6 (55)	18 (60)	2 (40)	1.83 (2)	.4	.39
Positive n (%)	12 (11)	5 (8)	3 (27)	3 (10)	0	N/A	N/A	N/A
Positive only excluded	62 (59)	41 (69)	3 (27)	16 (53)	0	7.68 (2)[d]	.02	.02[d]
Colleague burden	45 (43)	30 (51)	5 (45)	8 (27)	2 (40)	4.77 (2)	.09	.09
Positive only excluded	44 (42)	30 (51)	4 (36)	8 (27)	2 (40)	4.93 (2)	.09	.09
Bias/stigma	37 (35)	26 (43)	2 (18)	7 (23)	2 (40)	5.29 (2)	.07	.07
Positive only excluded	36 (34)	26 (43)	1 (9)	7 (23)	2 (40)	7.23 (2)[d]	.03	.03[d]
Faculty support	37 (35)	24 (41)	4 (36)	7 (23)17	2 (40)	2.64 (2)	.27	.26
Positive only excluded	32 (30)	24 (41)	1 (9)	5 (17)	2 (40)	8.03 (2)[d]	.02	.02[d]
Peer support	39 (37)	26 (44)	4 (36)	7 (23)	2 (40)	3.67 (2)	.16	.17
Positive only excluded	34 (32)	26 (44)	2 (18)	5 (17)	2 (40)	7.98 (2)[d]	.02[d]	.02[d]
Perceived performance	28 (27)	19 (32)	2 (18)	6 (20)	1 (20)	1.99 (2)	.37	.35
Positive only excluded	25 (24)	19 (32)	1 (9)	4 (13)	1 (20)	5.39 (2)	.07	.09
Patient relationships	7 (7)	3 (5)	4 (36)	0	0	N/A	N/A	.003[d]
Positive only excluded	6 (6)	3 (5)	3 (27)	0	0	N/A	N/A	.01[d]
Family relationships	10 (10)	6 (10)	3 (27)	1 (3)	0	N/A	N/A	.08
Positive only excluded	9 (9)	6 (10)	2 (18)	1 (3)	0	N/A	N/A	.28
Role model/mentors	10 (10)	7 (12)	2 (18)	1 (3)	0	N/A	N/A	.20
Positive only excluded	9 (9)	6 (10)	2 (18)	1 (3)	0	N/A	N/A	.28
Joy of parenthood	2 (2)	0	2 (18)	0	0	N/A	N/A	.01[d]

XSL•FO
RenderX

	Total (N=105), n (%)	Surgical (N=59), n (%)	PC[a] (N=11), n (%)	MS[b] (N=30), n (%)	All (N=5)	Values		Fisher exact test (*P* value)
						Chi-square (*df*)	*P* value	
Organization/community	87 (83)	48 (81)	11 (100)	26 (87)	2 (40)	N/A	N/A	.36
Positive n (%)	20 (19)	5 (8)	8 (73)	7 (23)	0	N/A	N/A	N/A
Positive only excluded	73 (70)	46 (78)	5 (45)	20 (67)	2 (40)	5.15 (2)	.08	.09
Childcare	26 (25)	16 (27)	4 (36)	6 (20)	0	1.21 (2)	.55	.56
Culture of support for pregnancy, parenthood	40 (38)	27 (46)	4 (36)	7 (23)	2 (40)	4.26 (2)	.11	.12
Positive only excluded	38 (36)	27 (46)	2 (18)	7 (23)	2 (40)	6.05 (2)[d]	.049	.05[d]
Culture of support PL[e]	55 (52)	31 (53)	7 (64)	16 (53)	1 (20)	0.47 (2)	.79	.80
Positive only excluded	46 (44)	30 (51)	3 (27)	14 (46)	1 (20)	2.07 (2)	.36	.36
Financial barriers	7 (7)	5 (8)	0	2 (7)	0	N/A	N/A	.86
PL coverage system	37 (35)	17 (29)	5 (45)	15 (50)	0	4.21 (2)	.12	.13
Lactation space	22 (21)	10 (17)	5 (45)	7 (23)	0	4.43 (2)	.11	.11
Positive only excluded	19 (18)	10 (17)	2 (18)	7 (23)	0	0.53 (2)	.77	.81
Lactation time	18 (17)	8 (14)	3 (27)	7 (23)	0	2.01 (2)	.37	.35
Schedule rigor or flexibility	42 (40)	25 (42)	4 (36)	12 (40)	1 (20)	1.34 (2)	.51	.54
Positive only excluded	34 (32)	24 (44)	1 (9)	8 (27)	1 (20)[d]	8.09 (2)[d]	.02[d]	.02[d]
PD[f] knowledge	10 (10)	4 (7)	1 (9)	5 (17)	0	N/A	N/A	.33
Gender of chair or PD	10 (10)	3 (5)	2 (18)	5 (17)	0	N/A	N/A	.10
Policy	83 (79)	47 (80)	6 (55)	26 (87)	4 (80)	5.04 (2)	.08	.09
Positive (study n)	5 (5)	2 (3)	2 (18)	1 (3)	0	N/A	N/A	N/A
Positive only excluded	79 (75)	46 (78)	4 (36)	25 (83)	4 (80)	19.1 (2)[d]	<.001	<.001[d]
PL policy existence	61 (58)	33 (56)	4 (36)	21 (70)	3 (60)	4.37 (2)	.11	.13
Positive only excluded	58 (55)	32 (54)	3 (27)	20 (67)	3 (60)	6.05 (2)[d]	.05	.06
PL length	52 (49)	28 (47)	4 (36)	16 (53)	4 (80)	0.95 (2)	.62	.64
Positive only excluded	50 (48)	28 (47)	3 (27)	15 (50)	4 (80)	1.8 (2)	.41	.41

XSL•FO
RenderX

	Total (N=105), n (%)	Surgical (N=59), n (%)	PC[a] (N=11), n (%)	MS[b] (N=30), n (%)	All (N=5)	Values		Fisher exact test (*P* value)
						Chi-square (*df*)	*P* value	
Childbearing PL	51 (49)	30 (51)	4 (36)	16 (53)	1 (20)	0.97 (2)	.62	.67
Nonchildbearing PL	37 (35)	21 (36)	3 (27)	12 (40)	1 (20)	0.58 (2)	.75	.83
Nontraditional PL	5 (5)	2 (3)	0	3 (10)	0	N/A	N/A	.84
Paid PL	38 (36)	19 (32)	6 (55)	11 (37)	2 (40)	1.61 (2)	.45	.50
Underuse of PL	17 (16)	8 (14)	3 (27)	6 (20)	0	1.51 (2)	.47	.42
Lactation policy	9 (9)	4 (7)	0	4 (13)	1 (20)	N/A	N/A	.46
Board policy	20 (19)	10 (17)	1 (9)	6 (20)	3 (60)	0.68 (2)	.71	.79
Institutional policy	25 (24)	13 (22)	1 (9)	10 (33)	1 (20)	2.9 (2)	.23	.29
Consistency of policies	13 (12)	9 (15)	1 (9)	3 (10)	0	6.89 (2)[d]	.03	.04[d]
Interactions	78 (74)	44 (75)	7 (64)	24 (80)	3 (60)	1.16 (2)	.56	.55
Positive (study n)	13 (12)	5 (8)	4 (36)	3 (10)	1 (20)	N/A	N/A	N/A
Positive only excluded	71 (68)	40 (68)	6 (55)	22 (73)	3 (60)	1.21 (2)	.52	.49
Training extension	37 (35)	17 (29)	4 (36)	13 (43)	3 (60)	0.32 (2)	.85	.87
Parenthood postponement	28 (27)	18 (31)	1 (9)	8 (27)	1 (20)	2.16 (2)	.34	.43
Training and parenthood incompatibility	12 (11)	11 (19)	1 (9)	0	0	N/A	N/A	.02[d]
Training experience	29 (28)	17 (29)	4 (36)	8 (27)	0	0.37 (2)	.83	.85
Positive only excluded	27 (26)	15 (29)	4 (36)	8 (27)	0	0.57 (2)	.75	.76
Specialty choice	16 (15)	8 (14)	1 (9)	6 (20)	1 (20)	N/A	N/A	.70
Career	28 (27)	15 (25)	3 (27)	10 (33)	0	0.93 (2)	.63	.64
Positive only excluded	21 (20)	11 (19)	2 (18)	8 (27)	0	0.83 (2)	.66	.63
Breastfeeding success	14 (13)	6 (10)	3 (27)	5 (17)	0	N/A	N/A	.22
Positive only excluded	13 (12)	6 (10)	2 (18)	5 (17)	0	N/A	N/A	.61
Combination and well-being/stress	18 (17)	13 (22)	3 (27)	2 (7)	0	3.9 (2)	.14	.10

XSL•FO
RenderX

	Total (N=105), n (%)	Surgical (N=59), n (%)	PC[a] (N=11), n (%)	MS[b] (N=30), n (%)	All (N=5)	Values		Fisher exact test (*P* value)
						Chi-square (*df*)	*P* value	
Positive only excluded	17 (16)	13 (22)	3 (27)	1 (3)	0	5.85 (2)[d]	.05	.03[d]

[a]PC: primary care.

[b]MS: medical subspecialties.

[c]N/A: not applicable.

[d]Statistically significant at $P<.05$.

[e]PL: parental leave.

[f]PD: program director.

Parenthood-Training Interactions

Three-quarters of included studies (78/105, 74%) described interactions in which training and/or parenthood impacted the other in some way [11,21,27,29-103]. The most common was extension of training as a result of parental leave (37/105, 35%) [30,35,36,38,40,42,44,46,47,49,51,53,57,62-68,70,73,74,76,78,79,81,84,87,90,91,95,96,98,99]. This interaction was particularly concerning for residents planning to pursue a fellowship. Both trainees and PDs described pressure to limit parental leave to avoid extending training as a result of some fellowships not accepting off-cycle residents or viewing off-cycle applicants as less desirable [35,40,46,63,91]. Some studies explored the influence of training on lactation goal achievement (14/105, 13%) [27,29,31,37,39,46,47,54,56,76,85,86,94,99].

Many studies also described postponement or avoidance of parenthood as a result of training (28/105, 27%) [21,29,33,34,37,39,47,48,50,51,54,56,58-60,69,72,75,79-81,87,88,94,95,98,101,102]. For instance, 3 studies of orthopedic trainees found many female trainees (range 48.4% - 68.6%) deliberately postponed childbearing until after training completion, most often to avoid a negative perception by male faculty and cotrainees and to avoid a detrimental career impact [75,87,88]. Studies also described parenthood's impact on quality of training experience (eg, challenges attaining adequate procedural numbers; 29/105, 28%) [11,31,34,38-40,42,44,45,47,49,54,57,60,71,76,79,81-83,86-88,91,92,95,98,99,103], specialty choice (16/105, 15%) [11,34,43,45,58,60,69,74,79,83,85,88,90,91,99,101], and career success (28/105, 26%) [34,38,41,42,44,46-49,54,57,58,60,69,71,76,77,79,81-83,87,88,91,95,99,102,103]. For example, a study of pediatric residents found female residents with children were significantly more likely to feel limited in their postresidency options because of family considerations compared to both female residents without children and male residents with children [82].

A theme of incompatibility of training and parenthood (12/105, 11%) was more common in surgical studies (11/59, 19%) compared to primary care (1/11, 9%) or medical subspecialties (0/30, 0%, $P=.02$) studies. In 1 study, 32% of female surgical trainees reported parenthood was strongly discouraged: "We were basically told that it would be forbidden to have a family" [80]. In 2 other surgical studies, over a third of female respondents who were pregnant during training (39% - 40%) strongly considered leaving residency due to their pregnancy experience, and a similar number (30% - 33%) would discourage female medical students from pursuing a surgical career because of the challenges of motherhood during surgical training [85,88]. The negative impact on well-being of the combination of parenthood and training (18/105, 17%) was more commonly identified in primary care studies (3/11, 27%) and surgical studies (13/59, 22%) compared to medical subspecialty studies (1/30, 3%, $P=.03$, statistically significant) [11,41,45,49,52,59,60,69,83,88,90,100].

While most interactions were negative, some positives emerged, such as 2 studies in which psychiatry PDs and pediatric trainees described heightened clinical and patient care skills in trainees with parenthood experience [44,99].

Individual Factors

Individual factors affecting trainee experiences of pregnancy and parenthood were identified in 46% (49/105) of studies [11,21,27,29,31-35,37,39,41,43,45,46,48,52,54,56,58,60,62,64,68,70,75,76,78,81,83,85,88,90,94,95,97,101,102,104-108]. Common themes included health concerns (n=26) such as miscarriage, pregnancy, postpartum, and neonatal complications [11,31,35,37,39,43,46-48,52,54,56,60,64,69,76,78,81,85,88,94,95,97,101,106,107]. For instance, in 1 study, almost two-thirds (63.6%) of general surgery trainees with pregnancies during training believed that their work schedules had harmed their own and/or their child's health [85]. In another, a majority of female orthopedic residents (60%) cited an inability to ensure optimal prenatal and postnatal care for herself and her child as a barrier to parenthood during residency [88]. Concerns about infertility (n=18) were found in surgical (n=11) and medical subspecialty studies (n=7) but not in primary care studies [21,32-35,37,39,43,58,60,64,81,90,94,95,102,105,108]. Factors involving trainee knowledge (eg, awareness of institutional parental leave policies and knowledge about pregnancy risks of radiation exposure) and beliefs (eg, trainee struggles should be hidden from peers and faculty) were present in 20 studies [21,27,29,33,41,45,46,49,54,56,59,62,68,70,75,87,98,102,104,105]. No statistically significant relationship was found between themes and specialty groups on an individual level.

Interpersonal Factors

Approximately two-thirds of studies (68/105, 65%) identified interpersonal factors influencing trainee parenthood experiences

XSL•FO
RenderX

[11,21,28,31,33-37,39-42,44-49,54-56,58-60,65,67,69,70,72-76,78,81,83,85-91,94,95,97-101,103-105,107-115]. When positive interpersonal studies were excluded, surgical (41/59, 69%) and medical subspecialty (16/30, 53%) studies were significantly more likely to identify interpersonal challenges than primary care (3/11, 27%; *P*=.02). The most frequently identified interpersonal theme was the extra burden on peers, who are often expected to cover duties for trainee parents on parental leave (45/105, 43%) [31,33,35,39-42,45-48,54-56,60,62,67,69,70,72,73,75,76,78,81,83,87,89,91,97-99,103-105,107,109-113,115]. In qualitative descriptions, this burden often took the form of peers assuming additional unpaid call or clinical duties to cover parental leave, which trainees described as fostering guilt among trainee parents, and in some cases, resentment or strained relationships among coresidents. Several studies noted that these dynamics discouraged trainees from taking the full parental leave available to them.

Experiences of bias toward trainee parents and experience of being stigmatized because of pregnancy were also common (37/105, 35%), particularly in surgical studies, in which 43% (26/59) identified negative experiences, compared to only 9% (1/11) of primary care studies and 23% (7/30) of medical subspecialty studies (*P*=.03) [11,21,31,33-37,39-41,46,49,54,55,58,60,62,64,69,70,72,74-76,81,83,85,87,88,90,94,95,98,99,103,109]. Reported manifestations included negative or disparaging comments about pregnant trainees and assumptions of reduced skill or commitment. For example, in 1 study, 72.9% of female surgery residents reported witnessing negative comments about pregnant colleagues, and over half perceived a negative stigma associated with pregnancy during training [85].

Other themes included the level of support from a trainee's faculty (37/105, 35%) [28,31,33,35-37,40,41,44-46,49,54,60,62,67,72,74-76,79,81,83,85-87,92,94,95,99,100,103,104,108,109,112] and fellow trainees (39/105, 37%) [28,31,33,35-37,39-42,44-46,49,55,62,67,70,72,74-76,79,81,83,85-88,94,95,99,100,103,104,108,109,112]. Excluding positive studies, surgical studies were significantly more likely to report issues for these themes (see Table 3).

Additional themes included others' perceptions about trainee parent performance as a result of their parenthood or pregnancy (28/105, 27%) [11,21,31,34,39-41,44,45,49,54,55,63,65,69,70,72,73,75,85,87-90,98,103,107,113]; the impact of relationships with family members (either support during residency, or strain on the relationship from partners having to shoulder extra home responsibilities; 10/105, 10%) [41,45,61,81,90,99-101,112,114]; the importance of experienced role models and mentors able to help navigate parenthood during training (10/105, 10%) [33,35,41,69,83,85,88,99,108,112]; and, uniquely in primary care studies, the joy of parenthood (2/105, 2%) [99,112]. Primary care studies (3/11, 27%) were significantly more likely than surgical (3/59, 5%) and medical subspecialty (0/30, 0%) studies to highlight challenges around patient relationships as a result of parenthood (*P*=.01) [41,45,61,81,90,99-101,112,114]. For instance, in 1 study, family medicine residents described guilt about having to choose between "going the extra mile" for their patients and leaving for home in time to see their children [112]. However, the number of primary care studies was relatively small (n=11), limiting confidence in specialty-based comparisons and raising the possibility that observed differences may partly reflect sampling variation or publication trends.

Organizational Factors

The organizational level was the SEM level with the most identified factors influencing trainee parenthood (87/105, 83%) [11,21,27-37,39-42,44-56,58-65,68-70,72-85,87-91,94,95,97-100,102-105,107-123]. Surgical studies were significantly more likely to identify challenges around having a supportive culture for pregnancy and parenthood (27/59, 46% vs 2/11, 18% primary care and 7/30, 23% medical subspecialty, *P*=.05) [11,35,37,39,45-50,52,55,59-61,64,69,72-76,78-81,87-89,94,95,97-100,103,105,107,108,121] and around the rigor and inflexibility of schedules (24/59, 41% vs 1/11, 9% primary care and 8/30, 27% medical subspecialty, *P*=.02) [11,27,33,35,37,39,42,45,47,48,51,53-55,59,61,63,66,67,70,74,76,78,79,81,84,87-90,93,98,100,103,105,107,108,110-112,114,119,122]. For example, surgical program expectations that trainees adhere to demanding call schedules and prolonged work hours up until delivery were noted to be associated with an increased risk of pregnancy complications (eg, preeclampsia and preterm labor) [24].

Themes without significant differences between specialty groups included affordability and adequacy of childcare services and hours (26/105, 25%) [11,32,35,37,45,46,48,58,75,78-80,82-85,87-89,95,99,100,102,108,112,114]; institutional-level financial barriers to supporting paid parental leaves (7/105, 7%) [40,44,48,55,59,65,114]; the system for covering duties of trainees taking parental leave (37/105, 35%) [28,34,41,42,44,46,47,50-55,62-65,68,70,77,79,81,82,90,97,99,103-105,111-115,121-123]; the culture around support for parental leave (55/105, 52%) [21,28,31,33,36,40,41,44-47,49,55,62,64,66,68,69,73,81,84,87-89,94,95,97-99,103-105,109,111,113,120-123]; lactation time (18/105, 17%) [11,35,37,39,47,50,54,56,64,68,75,78,79,89,99,103,108,112], lactation space (22/105, 21%) [31,35,37,39,45,47,50,54,56,64,68,78,79,82-85,88,89,99,108,112]; PD knowledge of board or institutional policies or guidelines (10/105; 10%) [21,51,54,55,65,70,91,99,110,115]; and whether sex of department chair or PD influenced support for trainee parenthood (10/105; 10%) [63,64,73,84,110,111,116-119].

Evidence on how the gender of leadership influences trainee parent experiences was inconsistent across studies. In 1 study, male anesthesiology fellowship directors were more likely to have a negative view of off-cycle applicants [63]. Similarly, 3 studies found that departments with female radiology chairs [116] and residency programs with female anesthesiology [117] or general surgery PDs [84] had higher rates of advertised paid parental leave than those led by males. However, 6 other studies found no significant difference in parental leave support or schedule accommodations based on the gender of department chairs in urology [110], ophthalmology [118], or anesthesiology [64], or of PDs in emergency medicine [111], family medicine [73,119], or anesthesiology [64].

Policy Factors

Policy factors were the second most common SEM level identified (83/105, 79%) [15,21,28-30,32-37,40-47,49,60,62-68,70,72,75,77-79,81-92,96-99,102,104-111,113,115-119,121-129]. When studies with purely positive findings were excluded,

XSL•FO
RenderX

surgical (46/59, 78%) and medical subspecialty (25/30, 83%) studies were significantly more likely than primary care to identify policy-level problems (4/11, 36%, $P<.001$) and were significantly more likely to highlight challenges around the lack of an institutional or program policy (surgical 32/59, 54%, primary care 3/11, 27%; medical subspecialty 20/30, 67%, $P = .05$) [2,3,33,34,40,46,48,52,56,58,64,66,67,75,77,83,85,89,99,102,105,107,109,111,113,115,118,121-123,127-129]. Surgical studies were more likely to highlight a lack of alignment of parental leave policies across different levels (eg, specialty board, ACGME [Accreditation Council for Graduate Medical Education], and institutional; surgical 9/59, 15%; primary care 1/11, 9%; medical subspecialty 3/30, 10%; $P=.03$) [33,35,41,43,54,55,87,98,110,113,123]. For instance, the American College of Obstetricians and Gynecologists' (ACOG) policy recommends at least 8 weeks of paid parental leave in addition to vacation and sick leave [130]. However, in 1 study, 71% of obstetrics and gynecology PDs were unaware of ACOG's parental leave policy, and local policies were not aligned with ACOG's policy, resulting in shorter than recommended parental leaves [55]. Similarly, specialty board training requirements may be an obstacle to optimizing institutional policies. For instance, an overwhelming majority (82.2%) of surgery trainees and graduates reported that an American Board of Surgery policy requiring 48 weeks of full-time clinical activity in each of the five years of surgical residency [131] was a barrier to obtaining their desired parental leave length, with over three-quarters (78.4%) receiving a maternity leave of 6 weeks or less [11].

Other policy-related challenges without significant differences between groups included the length of parental leave (52/105, 49%) [2,3,5,6,8,9,14,15,17,18,20,22,25,27,29,31,33,34,36,37,39,40,42,43,45,46,48,50,52,53,55,56,58,60,62,64,67,69,72,73,75,77,79,81,83,84,86,88,89,91,93,96,99,100,102,104,106,108,109,111,113,115,118,119,121,122,124,126] underuse of parental leave (17/105, 16%) [30,34,36,37,40-42,47,51-54,57,65,109,119,122]; and whether parental leave is paid (38/105, 36%) [30,34,40,42,47,52,53,56,62,65,67,70,72,77,78,81,82,84,86,87,89,91,100,102,104,109,110,113,115,119,121-123,127-129]. Almost half of the studies discussed parental leave for the birthing parent (52/105, 49%) [2,3,5,7,9,14,15,17,18,20,23,25,27,29,31,33,34,36,37,39,40,42,43,45,46,48,50,52,53,55,56,58,60,62,64,67,69,72,73,75,77,79,81,83,84,86,88,89,91,93,96,99,100,102,104,106,108,110,111,113,115,116,118,119,121-124,126] and just over a third discussed parental leave for the nonchildbearing parent (37/105, 35%) [34,37,41,42,45,47,51,53,57,62,63,66,67,70,72,82,84,86,87,89,91,93,103-110,113,115,116,118,119,121-124]. Only 5/105 (5%) studies discussed adoptive or nontraditional leave [53,65,121,122,124], and only 9/105 (9%) studies discussed lactation policies [37,53,58,68,86,89,115,121,125]. Institutional-level policies (25/105, 24%) [34,44,46,55,57,58,63-65,72,75,79,85,88,105,110,111,113,115,118,122-124,128,129] were more commonly addressed than specialty board-level policies (20/105, 19%) [33,34,44,46,49,52,54,55,67,82,85,88,91,96,98,110,111,113,125,128].

Sensitivity Analysis Findings

The sensitivity analysis reclassifying anesthesiology within the surgical specialties demonstrated overall stability in the majority of findings. Differences between specialty groups remained significant for several themes, including bias and stigma (excluding positive-only studies), faculty support, patient relationship, joy of parenthood, and the policy SEM level (excluding positive-only studies).

However, several themes were no longer statistically significant following reclassification. These included the interpersonal SEM level, schedule rigidity, consistency of policy implementation (excluding positive-only studies), perceived compatibility of training and parenthood, and the impact of trainee parenthood on well-being and stress (excluding positive-only studies).

Full results of the sensitivity analysis, including statistical comparisons, are presented in Multimedia Appendix 3.

Discussion

Principal Findings

This study applied SEM to map experiences of trainee parenthood across surgical, primary care, and medical subspecialty training. We found that organizational, policy, and interpersonal challenges are more frequently identified in the literature than individual-level barriers. While many challenges were common across specialties, interpersonal and policy-related issues were more prevalent in studies on surgical and medical subspecialty trainees compared to primary care studies. Surgical studies most frequently described stigma surrounding pregnancy and parenthood, a lack of faculty and peer support, rigid schedules, an unsupportive culture of parenthood, and perceived incompatibility between parenthood and training.

The challenges identified in this study align with prior literature, which has consistently described issues such as medical complications of pregnancy, concerns about faculty and peer bias, heavy workloads, rigid schedules, unclear or inadequate parental leave, insufficient lactation facilities, inadequate childcare support, and potential impact on training and career plans [3,8,9,112,132-137]. While no reported prior reviews have compared parenthood experiences across specialty groups, a 2019 systematic review of surgical trainees found that female surgical trainees gave birth to their first child later in life, had fewer children, and were more likely to experience infertility than women in the general population, and were also less likely to have children than male surgical trainees [138]. Although pregnant surgical trainees performed comparably on examinations and case numbers and did not have higher attrition rates, they faced stigma and negative attitudes from colleagues [138]. These results are consistent with our findings that surgical trainees may be more likely than other trainees to experience issues such as stigma, negative attitudes of individuals in the program, and an unsupportive culture for pregnancy and parenthood.

While surgical studies were more likely to report interpersonal and organizational challenges from trainee parenthood, both primary care and surgical studies were more likely than medical subspecialty studies to report a negative impact on well-being from the combination of parenthood and training. Primary care studies were also the most likely to report conflicts between patient care and family life. The included studies did not explicitly explain this finding, which should be interpreted cautiously given the small number of primary care studies.

XSL•FO
RenderX

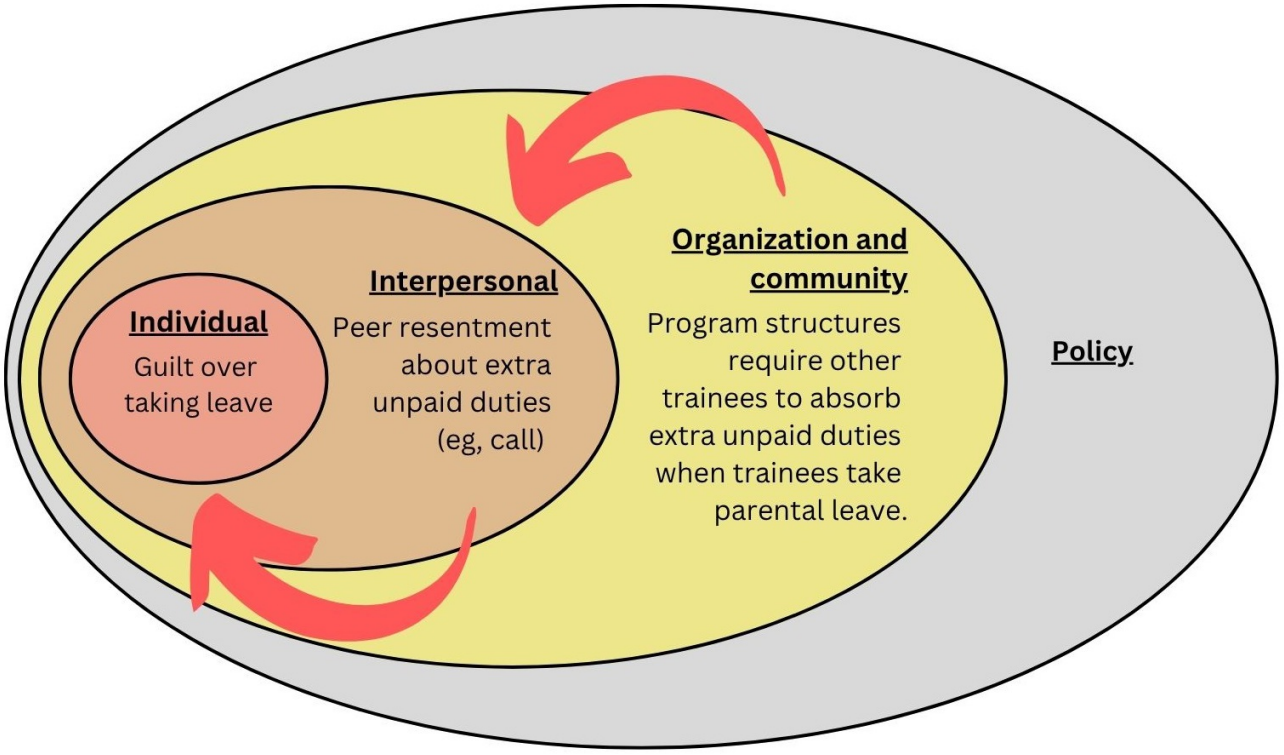
Several hypotheses may warrant future investigation. Relational continuity is a defining feature of primary care, and ongoing therapeutic relationships with patients may increase trainees' sense of responsibility toward patients when competing family demands arise [139,140]. In addition, primary care work frequently extends beyond scheduled clinical encounters through activities such as inbox management, medication refills, laboratory follow-up, and care coordination [141]. These responsibilities have been associated with after-hours work and erosion of work-life boundaries, which may contribute to work-family conflict [142]. However, it is also possible that differences in study focus, measurement, or reporting practices contributed to this finding rather than true differences between specialty groups.

Interactions Between SEM Levels

A key strength of the SEM is its ability to illustrate how policy, organizational, interpersonal, and individual factors interact to shape trainee experiences, outcomes, and behaviors. Across the included studies, these levels rarely operated independently; instead, policies were enacted through institutional structures that shaped interpersonal dynamics, which in turn influenced individual decision-making and well-being.

For example, parental leave policies and specialty board training requirements at the policy level determine how leave is operationalized within residency programs. Where coverage structures are not formally built into training systems, organizational implementation often shifts the burden of parental leave onto cotrainees, who absorb additional clinical duties without compensation or workload adjustment. This structural arrangement produces predictable interpersonal consequences, including resentment or "colleague burden" among peers covering additional call shifts, and guilt or moral distress among trainee parents taking leave. Over time, these interpersonal dynamics contribute to hesitancy to use available leave, even when policies formally permit it, thereby limiting the practical impact of supportive policy structures (Figure 3). Further, 1 study of female trainees found that even when parental leave was formally available, over half (52.1%) reported not feeling able to take leave, and among those who did, the majority (76% - 88%) reported the leave duration was insufficient [138]. Shortened parental leaves may have downstream consequences; 1 study found that female trainees who took shorter leaves (4 - 6 wk instead of 6 - 8 wk) were more likely to experience postpartum depression [143].

Figure 3. Program structure, interpersonal resentment, and trainee guilt. This figure illustrates how organizational-level factors (eg, training program schedules and coverage systems) interact with interpersonal dynamics (eg, peer resentment), which in turn shape individual-level outcomes such as trainee guilt related to taking parental leave.



Conversely, these cross-level interactions may also operate in a protective direction. Although many PDs are unaware of ACGME parental leave requirements, limiting consistent implementation (Figure 4), knowledgeable faculty mentors at the interpersonal level may help translate policy into actionable guidance for trainees, reinforce entitlement to leave, and advocate for equitable implementation at the organizational level. In these contexts, supportive interpersonal relationships can reduce stigma and mitigate the effects of organizational ambiguity (Figure 5). However, the literature suggests that this translation function is inconsistent, resulting in variable trainee experiences despite similar formal policies.

XSL•FO
RenderX

Figure 4. Program director policy awareness. This figure illustrates how, despite the existence of protective parental leave policies, organizational-level barriers (eg, limited program director awareness of these policies) may undermine their implementation and result in negative trainee outcomes, including inability to take desired parental leave. ACGME: Accreditation Council for Graduate Medical Education.

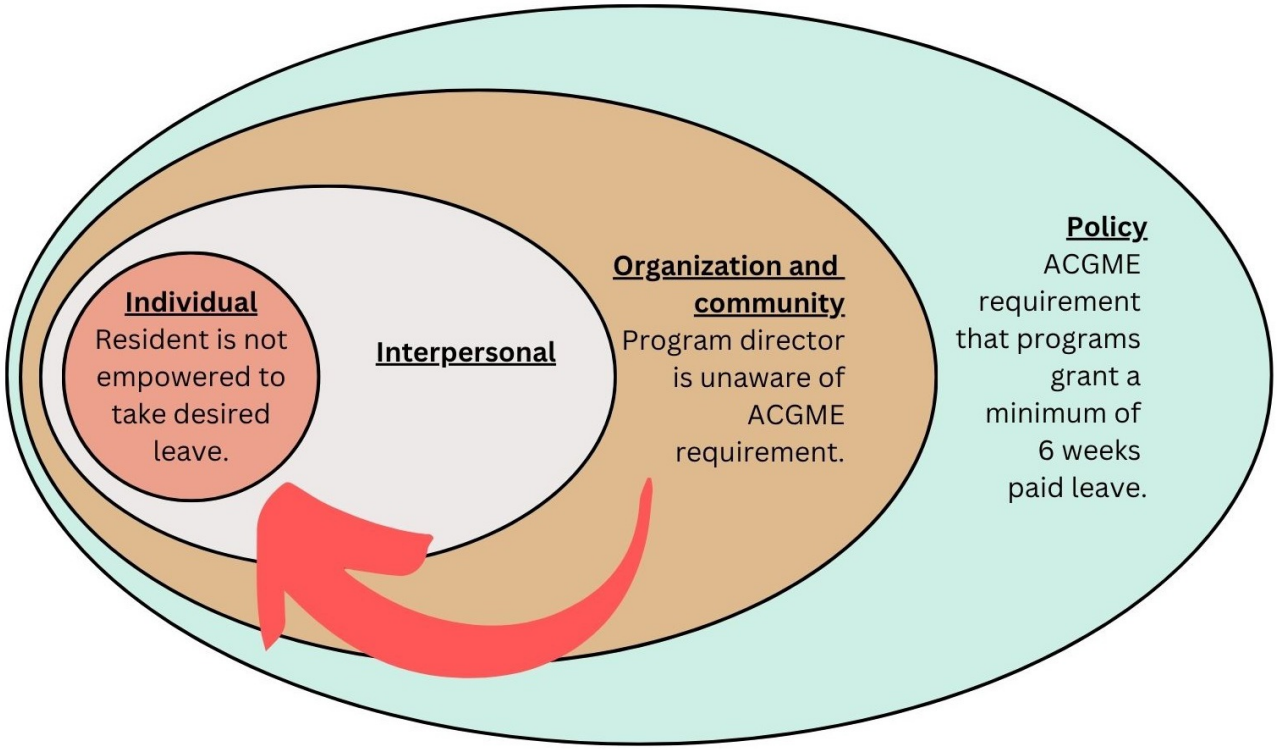
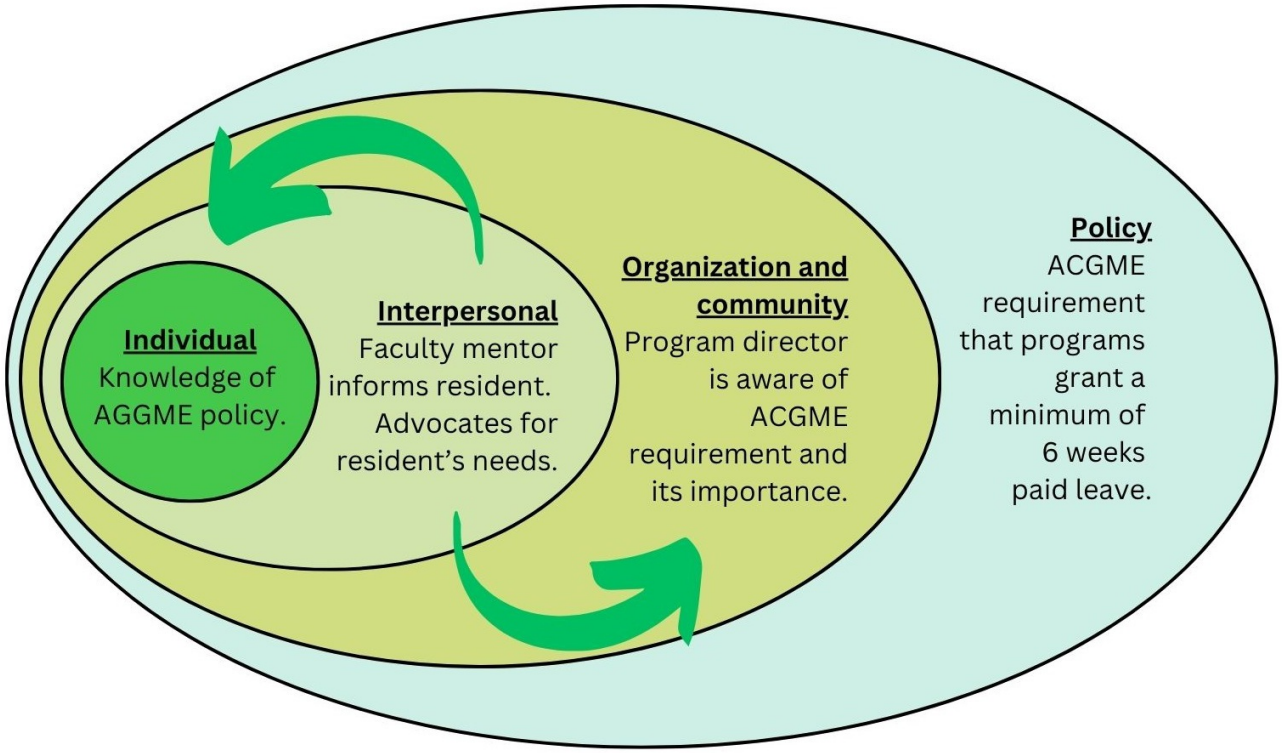


Figure 5. Faculty mentor influence on resident knowledge and organizational support. This figure illustrates how a faculty mentor (an interpersonal-level factor) can function as a positive catalyst for trainee parental leave experiences. Faculty mentors may promote awareness of relevant policies (eg, ACGME requirements for a minimum of 6 weeks of paid parental leave) among both trainees and program leadership, thereby enhancing trainee knowledge and empowerment as well as strengthening organizational support for parental leave. ACGME: Accreditation Council for Graduate Medical Education.



Taken together, these findings suggest that misalignment across SEM levels, particularly between policy intent and organizational implementation, may be a key mechanism through which structural conditions produce interpersonal strain and stigma, which in turn shape trainee behavior regarding leave usage and perceived legitimacy of parenthood during training.

XSL•FO
RenderX

Interventions

Only 5 uncontrolled studies evaluated interventions to improve trainee parenthood experiences. The interventions are clustered into three broad operational approaches. First, several programs modified scheduling and coverage structures. These included removal of overnight call during pregnancy, flexible scheduling for expectant parents, cross-coverage systems, and dedicated parenthood electives that allowed residents to remain engaged in educational activities while spending time at home with a newborn. The available studies suggest that accommodations such as schedule redesign, modified call structures, and parenthood electives may be feasible within existing residency training structures. Notably, the studies that evaluated operational outcomes did not identify adverse effects on resident scheduling or progression, although these outcomes were assessed in only a small number of programs.

Second, programs implemented formal parental leave policies and transparent leave processes. Studies evaluating these interventions reported improvements in perceived leadership support, fairness, and equity surrounding parenthood and training. These findings suggest that clear, standardized parental leave policies may influence trainee perceptions of institutional support, fairness, and equity surrounding parenthood during training.

Third, interventions addressing lactation support focused on reducing logistical barriers to breastfeeding through dedicated spaces and equipment. These interventions were associated with reduced guilt and anxiety and improved ability to remain engaged in clinical responsibilities while meeting lactation needs.

Overall, the limited interventional literature suggests that schedule flexibility, formalized leave policies, mentorship, and lactation infrastructure are the most consistently tested program-level strategies, though evidence remains preliminary.

Our findings underscore the need for multilevel interventions to support trainee parents. Interpersonal dynamics remain critical, especially for females, who are more likely to fear that parenthood will damage their career and professional image and burden peers [34,46-49,54,60,69,70,81,82,85,87,102,103,138]. Organizational and policy factors such as schedule flexibility and paid parental leave also play a vital role. Numerous studies highlight the importance of physician-parent mentorship in balancing family and professional duties [33,35,41,69,83,85,88,99,108,112]. Subsidized childcare can reduce financial burden and stress, especially for female trainees, for whom childcare often emerged as a more significant challenge than for male trainees [10,32,46,48,56,75,79,85,95,114]. Institutions should support lactating trainees through measures such as clear policies, accessible spaces, and schedule accommodations [20,144]. In addition, board policies and training requirements must evolve to protect rather than hinder parenthood during training. For example, if supported by factors at other social ecological levels, ACGME's 2022 parental leave policy can help ensure trainees are offered at least six weeks of paid leave [145].

We found that 33 of 105 (31.4%) studies reported at least one positive finding related to trainee parenthood experiences, while 103 of 105 (98.1%) studies included at least one negative outcome or challenge. The predominance of studies emphasizing negative outcomes may reflect publication bias, selective outcome framing, or a scholarly tendency to prioritize documenting barriers and strain in trainee parenthood. This imbalance is important because it may shape perceptions of parenthood during medical training. If the published literature disproportionately highlights stress, conflict, and barriers, it may inadvertently reinforce a narrative that parenthood and residency are inherently incompatible. Such framing could influence trainee specialty selection, discourage disclosure of parenting needs, and shape institutional policy discussions toward mitigation of harm rather than optimization of supportive environments. At the same time, the presence of positive findings in nearly one-third of studies suggests that supportive or adaptive experiences are not uncommon, even if they are less frequently reported than challenges.

Strengths and Limitations

Strengths of this study include the use of a structured framework (SEM), inclusion of male trainee perspectives, and comparison of physician parenthood experiences across specialty groups. However, several limitations reflect gaps in existing literature. Included studies varied widely in design, populations, and quality, with a predominance of survey-based studies increasing the risk of response and selection bias. Notably, studies directly comparing PD and trainee perspectives highlighted marked differences, with PDs reporting substantially more favorable perceptions of program-level support than trainees [21,54]. This finding highlights limitations of stakeholder self-report and raises concerns about measurement validity in the current evidence base.

Only 5 of the included studies evaluated explicit interventions, with the remaining literature largely descriptive, limiting the ability to draw conclusions about effective strategies to support trainee parents. Additionally, exclusion of gray literature may have limited capture of relevant policy-level data, including specialty board guidance and institutional policy statements. Although male trainees were included, the literature remains heavily skewed toward female trainee perspectives, and the positive aspects of parenthood during training were infrequently reported, suggesting possible negativity bias in the existing literature. Some specialties were underrepresented, and within-group heterogeneity (eg, differences between general cardiology and interventional cardiology) may obscure the influence of highly specific training environments on parenthood experiences.

Sensitivity analyses reclassifying anesthesiology within a surgical cohort demonstrated overall stability in the majority of findings, with attenuation of several interpersonal and well-being–related domains. These results suggest that while core thematic differences are robust to alternative classification schemes, certain constructs may be partially sensitive to how procedural specialties are grouped, underscoring the methodological trade-offs inherent in specialty categorization across heterogeneous training environments.

XSL•FO
RenderX

Despite these limitations, this scoping review provides valuable insights into how physicians experience parenthood across individual, interpersonal, organizational, and policy levels during training, and how this may be similar and different across specialty groups.

Future Research

Future research should prioritize rigorous evaluation of interventions designed to improve trainee parenthood experiences, moving beyond descriptive studies to assess effectiveness. In particular, intervention-focused work should explore which modifiable features of training (eg, specific forms of schedule flexibility, lactation support infrastructure, mentorship models, and childcare resources) are associated with improved trainee well-being, training outcomes, and retention. Incorporating objective institutional metrics will be essential to complement self-reported perceptions and to better understand discrepancies between trainee and program leadership perspectives.

Additional studies are needed to explore subspecialty-specific training environments, including nonsurgical procedural and high-acuity fields such as anesthesiology, where unique occupational exposures and demands may shape parenthood experiences and decisions. Greater inclusion of male trainees and a more balanced examination of both challenges and positive dimensions of parenthood during training would further strengthen the evidence base.

Finally, multi-institutional, international, and longitudinal studies—including posttraining follow-up—may clarify how training-era policies influence long-term physician career trajectories, family decisions, and workforce sustainability [146].

Conclusions

Trainee parents in the United States face significant individual, interpersonal, organizational, and policy challenges. While many of these are shared across specialties, surgical trainees more frequently report stigma, rigid schedules, and cultural barriers, while primary care trainees often struggle with the tension between patient care and parenting. Effective support for trainee parents requires systemic change. Systemic interventions such as equitable parental leave, childcare support, and cultural shifts in program parenting support are essential to enable all trainees to thrive as both physicians and parents.

Acknowledgments

The authors would like to acknowledge Tara N Atkins, MLIS, and Julie Trumble, MLIS, for their assistance in the development of a search strategy and database searches, and Christen Walcher, MPAff, for editing and formatting this paper. Preliminary findings from this study were presented at the Society for Teachers of Family Medicine Annual Meeting in Salt Lake City, Utah, May 2025. After drafting the manuscript, ChatGPT (version 5.0; OpenAI) was used to assist with proofreading and readability. All suggestions were reviewed and selectively incorporated by the authors, who remain fully responsible for the accuracy, originality, and integrity of all content in this paper, including all references and citations.

Funding

The authors declared no financial support was received for this work.

Data Availability

Raw data are available upon request to the corresponding author.

Authors' Contributions

Conceptualization: ZMSD (lead), LRP (equal), EMV (supporting)
Data curation: LRP (lead), ZMSD (supporting)
Formal analysis: LRP (lead), ZMSD (supporting), EMV (supporting), DB (supporting), FJ (supporting), MEJ (supporting)
Investigation: ZMSD (lead), LRP (equal), EMV (supporting), DB (supporting), FJ (supporting), MEJ (supporting)
Methodology: LRP (lead), ZMSD (equal)
Project administration: LRP (lead), ZMSD (equal)
Supervision: LRP (lead), ZMSD (equal)
Visualization: LRP
Writing - original draft: ZMSD (lead), LRP (equal)
Writing - review & editing: ZMSD (lead), LRP (equal), EMV (supporting), DB (supporting), FJ (supporting), MEJ (supporting)

Conflicts of Interest

None declared.

Multimedia Appendix 1
Supplement: search strategies.
[DOCX File, 26 KB - mededu_v12i1e87284_app1.docx]

XSL•FO
RenderX

Multimedia Appendix 2
Summary of individual studies' characteristics.
[DOCX File, 88 KB - mededu_v12i1e87284_app2.docx]

Multimedia Appendix 3
Sensitivity analysis of specialty-based themes with anesthesiology reclassified as a surgical specialty.
[DOCX File, 64 KB - mededu_v12i1e87284_app3.docx]

Checklist 1
PRISMA-ScR checklist.
[PDF File, 659 KB - mededu_v12i1e87284_app4.pdf]

References

1. Dyrbye LN, West CP, Satele D, et al. Burnout among U.S. medical students, residents, and early career physicians relative to the general U.S. population. Acad Med 2014 Mar;89(3):443-451. [doi: 10.1097/ACM.0000000000000134] [Medline: 24448053]
2. West CP, Dyrbye LN, Erwin PJ, Shanafelt TD. Interventions to prevent and reduce physician burnout: a systematic review and meta-analysis. Lancet 2016 Nov 5;388(10057):2272-2281. [doi: 10.1016/S0140-6736(16)31279-X] [Medline: 27692469]
3. Stack SW, Jagsi R, Biermann JS, et al. Childbearing decisions in residency: a multicenter survey of female residents. Acad Med 2020 Oct;95(10):1550-1557. [doi: 10.1097/ACM.0000000000003549] [Medline: 32568852]
4. Goldman AL, Barnett ML. Changes in physician work hours and implications for workforce capacity and work-life balance, 2001-2021. JAMA Intern Med 2023 Feb 1;183(2):106-114. [doi: 10.1001/jamainternmed.2022.5792] [Medline: 36534376]
5. Statistics on working time. International Labor Organization. 2024. URL: https://ilostat.ilo.org/topics/working-time/ [accessed 2026-06-10]
6. Ganguli I, Sheridan B, Gray J, Chernew M, Rosenthal MB, Neprash H. Physician work hours and the gender pay gap - evidence from primary care. N Engl J Med 2020 Oct 1;383(14):1349-1357. [doi: 10.1056/NEJMsa2013804] [Medline: 32997909]
7. Harry E, Sinsky C, Dyrbye LN, et al. Physician task load and the risk of burnout among US physicians in a national survey. Jt Comm J Qual Patient Saf 2021 Feb;47(2):76-85. [doi: 10.1016/j.jcjq.2020.09.011] [Medline: 33168367]
8. Sharp EA, Nicholson KJ, Taft M, Glaser DH, Gibson B, Kazmerski TM. Parent-in-training: resident and fellow experiences from pregnancy to parenthood. J Hosp Med 2022 Aug;17(8):609-623. [doi: 10.1002/jhm.12914] [Medline: 35855539]
9. Schwartz RW. Pregnancy in physicians: characteristics and complications. Obstet Gynecol 1985 Nov;66(5):672-676. [Medline: 4058826]
10. Freeman G, Bharwani A, Brown A, Ruzycki SM. Challenges to navigating pregnancy and parenthood for physician parents: a framework analysis of qualitative data. J Gen Intern Med 2021 Dec;36(12):3697-3703. [doi: 10.1007/s11606-021-06835-0] [Medline: 33959880]
11. Rangel EL, Smink DS, Castillo-Angeles M, et al. Pregnancy and motherhood during surgical training. JAMA Surg 2018 Jul 1;153(7):644-652. [doi: 10.1001/jamasurg.2018.0153] [Medline: 29562068]
12. Murphy B. Medical specialty choice: should residency training length matter? AMA. 2020 Nov 19. URL: https://www.ama-assn.org/medical-students/specialty-profiles/medical-specialty-choice-should-residency-training-length [accessed 2026-06-10]
13. Morrison GM, Di Cocco BL, Goldberg R, et al. Impact of parental status on US medical student specialty selection. Arch Womens Ment Health 2023 Dec;26(6):785-791. [doi: 10.1007/s00737-023-01366-5]
14. Tricco AC, Lillie E, Zarin W, et al. PRISMA Extension for scoping reviews (PRISMA-ScR): checklist and explanation. Ann Intern Med 2018 Oct 2;169(7):467-473. [doi: 10.7326/M18-0850] [Medline: 30178033]
15. Moher D, Liberati A, Tetzlaff J, Altman DG, The PRISMA Group. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. PLoS Med 2009 Jul 21;6(7):e1000097. [doi: 10.1371/journal.pmed.1000097]
16. Porterfield L, Delgado ZMS, Bridwell D, Jones F, Jones M, Vaughan EM. Resident physician parenthood: a scoping review protocol of physician experiences of parenthood in surgical, medical subspecialty, and primary care contexts. OSF Registries. 2025 Nov 7. URL: https://osf.io/pzk6c/overview [accessed 2026-06-10]
17. Eden J, Levit L, Berg A, Morton S, editors. Finding what works in health care: standards for systematic reviews. : National Academies Press (US); 2011. [doi: 10.17226/13059]
18. McLeroy KR, Bibeau D, Steckler A, Glanz K. An ecological perspective on health promotion programs. Health Educ Q 1988;15(4):351-377. [doi: 10.1177/109019818801500401] [Medline: 3068205]
19. Ouzzani M, Hammady H, Fedorowicz Z, Elmagarmid A. Rayyan-a web and mobile app for systematic reviews. Syst Rev 2016 Dec 5;5(1):210. [doi: 10.1186/s13643-016-0384-4] [Medline: 27919275]
20. Ames EG, Burrows HL. Differing experiences with breastfeeding in residency between mothers and coresidents. Breastfeed Med 2019 Oct;14(8):575-579. [doi: 10.1089/bfm.2019.0001] [Medline: 31219317]

XSL•FO
RenderX

21. Beltrán Ponce S, Jagsi R, Florez N, et al. Can I leave? Perspectives on parental leave and parenthood in medical training among program directors and trainees in oncologic specialties. J Womens Health 2024 Feb;33(2):218-227. [doi: 10.1089/jwh.2023.0200]
22. Dixit A, Feldman-Winter L, Szucs KA. "Frustrated," "depressed," and "devastated" pediatric trainees: US academic medical centers fail to provide adequate workplace breastfeeding support. J Hum Lact 2015 May;31(2):240-248. [doi: 10.1177/0890334414568119] [Medline: 25588382]
23. Chan MK, Chew QH, Sim K. Burnout and associated factors in psychiatry residents: a systematic review. Int J Med Educ 2019 Jul 30;10(149–60):149-160. [doi: 10.5116/ijme.5d21.b621] [Medline: 31381505]
24. Sealy institute for vaccine sciences. UTMB. URL: https://www.utmb.edu/sivs/home [accessed 2026-06-10]
25. Shawahna R, Maqboul I, Ahmad O, Al-Issawy A, Abed B. Prevalence of burnout syndrome among unmatched trainees and residents in surgical and nonsurgical specialties: a cross-sectional study from different training centers in Palestine. BMC Med Educ 2022 Apr 26;22(1):322. [doi: 10.1186/s12909-022-03386-8] [Medline: 35473599]
26. Serenari M, Cucchetti A, Russo PM, et al. Burnout and psychological distress between surgical and non-surgical residents. Updates Surg 2019 Jun;71(2):323-330. [doi: 10.1007/s13304-019-00653-0]
27. Chernoby KA, Pettit KE, Jansen JH, Welch JL. Flexible scheduling policy for pregnant and new parent residents: a descriptive pilot study. AEM Educ Train 2021 Apr;5(2):e10504. [doi: 10.1002/aet2.10504] [Medline: 33898908]
28. Conway SE, Vaswani PA, Budhu JA, et al. Development and impact of a progressive parental leave policy in a neurology residency. Neurology 2022 Jun 7;98(23):973-979. [doi: 10.1212/WNL.0000000000200729] [Medline: 35418451]
29. Corbisiero MF, Acker SN, Bothwell S, Christian N. Transforming perceptions: the impact of a formal parental leave policy on surgical trainees. J Surg Educ 2024 Jun;81(6):816-822. [doi: 10.1016/j.jsurg.2024.03.007] [Medline: 38677898]
30. Cree-Green M, Cree J, Urban K, Bunik M, Sass AE, Rosenberg A. A structured neonatal parenting elective: an approach for parenting leave during residency. Acad Pediatr 2020 Jul;20(5):595-599. [doi: 10.1016/j.acap.2020.02.008] [Medline: 32045680]
31. Creo AL, Anderson HN, Homme JH. Productive pumping: a pilot study to help postpartum residents increase clinical time. J Grad Med Educ 2018 Apr;10(2):223-225. [doi: 10.4300/JGME-D-17-00501.1] [Medline: 29686765]
32. Advani R, Saeed N, Arjonilla M, et al. Gender, marital, and parental barriers to pursuing gastroenterology fellowship: a multicenter analysis of a national survey and actionable solutions. Dig Dis Sci 2025 Feb;70(2):504-515. [doi: 10.1007/s10620-024-08783-1] [Medline: 39762595]
33. Altieri MS, Salles A, Bevilacqua LA, et al. Perceptions of surgery residents about parental leave during training. JAMA Surg 2019 Oct 1;154(10):952-958. [doi: 10.1001/jamasurg.2019.2985] [Medline: 31389989]
34. Baniel CC, Qu V, Ponce SB, et al. SUPPORT: Survey of Parental Leave Policies of Radiation oncology programs and residency applicants. Adv Radiat Oncol 2023;8(4):101207. [doi: 10.1016/j.adro.2023.101207] [Medline: 37124316]
35. Bernal IC, Moon SL, Hotta M, Newman MI. Residents' perspectives of pregnancy and growing a family during surgical training: a review of the literature. Cureus 2024 Apr;16(4):e58335. [doi: 10.7759/cureus.58335] [Medline: 38752085]
36. Bostock IC, Bayley EM, Antonoff MB. Barriers to parental leave during training: time for change. Am Surg 2023 Mar;89(3):452-456. [doi: 10.1177/00031348211031840] [Medline: 34250836]
37. Bourne DA, Chen W, Schilling BK, Littleton EB, Washington KM, De La Cruz C. The impact of plastic surgery training on family planning and prenatal health. Plast Reconstr Surg 2019 Nov;144(5):1227-1236. [doi: 10.1097/PRS.0000000000006100] [Medline: 31688771]
38. Brown EG, Galante JM, Keller BA, Braxton J, Farmer DL. Pregnancy-related attrition in general surgery. JAMA Surg 2014 Sep 1;149(9):893. [doi: 10.1001/jamasurg.2014.1227]
39. Castillo-Angeles M, Atkinson RB, Easter SR, et al. Pregnancy during surgical training: are residency programs truly supporting their trainees? J Surg Educ 2022;79(6):e92-e102. [doi: 10.1016/j.jsurg.2022.06.011] [Medline: 35842402]
40. Castillo-Angeles M, Smink DS, Rangel EL. Perspectives of US general surgery program directors on cultural and fiscal barriers to maternity leave and postpartum support during surgical training. JAMA Surg 2021 Jul 1;156(7):647-653. [doi: 10.1001/jamasurg.2021.1807] [Medline: 34009280]
41. Castillo-Angeles M, Stucke RS, Rosenkranz KM, Smink DS, Rangel EL. Paternity leave during surgical training: perspectives of male residents. J Surg Educ 2022 Nov;79(6):e85-e91. [doi: 10.1016/j.jsurg.2022.04.012]
42. Conway SE, Wang W, Prasad S. Barriers to increasing paid parental leave in U.S. neurology residencies: a survey of program directors. BMC Med Educ 2024 Apr 9;24(1):387. [doi: 10.1186/s12909-024-05333-1] [Medline: 38594709]
43. David YN, Dixon RE, Kakked G, et al. Pregnancy and the working gastroenterologist: perceptions, realities, and systemic challenges. Gastroenterology 2021 Sep;161(3):756-760. [doi: 10.1053/j.gastro.2021.05.053] [Medline: 34089733]
44. Dillinger RL. From requisite to right: assessing and addressing paid maternity leave in US psychiatry residency programs. Acad Psychiatry 2022 Apr;46(2):167-171. [doi: 10.1007/s40596-021-01523-x] [Medline: 34559391]
45. Diaz D, Freburg-Hoffmeister DL, Austin TM, Nyshadham S, Abramowicz S. Parental policy in oral and maxillofacial surgery residency programs is necessary but not available: a cross-sectional survey of oral and maxillofacial surgery residents' attitudes toward parental leave. J Oral Maxillofac Surg 2021 Dec;79(12):2404-2410. [doi: 10.1016/j.joms.2021.08.150] [Medline: 34547262]

XSL•FO
RenderX

46. Dixon A, Bansal N, Nicholas SB, Ostrow A, Kendrick J. A national survey of pregnancy and parenthood among nephrology trainees: a focus on nephrology fellowship. Clin J Am Soc Nephrol 2024 Aug 1;19(8):984-994. [doi: 10.2215/CJN.0000000000000486] [Medline: 38728092]
47. Dundon KMW, Powell WT, Wilder JL, et al. Parenthood and parental leave decisions in pediatric residency. Pediatrics 2021 Oct 1;148(4):e2021050107. [doi: 10.1542/peds.2021-050107]
48. Esfandiari N, Litzky J, Sayler J, Zagadailov P, George K, DeMars L. Egg freezing for fertility preservation and family planning: a nationwide survey of US obstetrics and gynecology residents. Reprod Biol Endocrinol 2019 Jan 29;17(1):16. [doi: 10.1186/s12958-019-0459-x] [Medline: 30696433]
49. Freudenberger DC, Riner AN, Herremans KM, et al. A gravid situation: general surgery faculty support for pregnant surgical residents. J Surg Res 2024 Jul;299:9-16. [doi: 10.1016/j.jss.2024.03.002] [Medline: 38677003]
50. Garza RM, Weston JS, Furnas HJ. Pregnancy and the plastic surgery resident. Plast Reconstr Surg 2017 Jan;139(1):245-252. [doi: 10.1097/PRS.0000000000002861] [Medline: 28027263]
51. Gardiner CP, Desrochers L, Finn K, et al. Paying for parenthood: understanding parental leave policies in infectious disease fellowship. Open Forum Infect Dis 2024 Feb;11(2):ofad685. [doi: 10.1093/ofid/ofad685] [Medline: 38390462]
52. Gaffley M, Hernandez S, Riera KM, Anzola S. Survey on the perceptions of pregnancy and parenthood in trainees: advances, obstacles, and growth opportunities. J Surg Res 2024 Mar;295:477-486. [doi: 10.1016/j.jss.2023.10.019] [Medline: 38070262]
53. Sherbaf FG, Lin DDM, Yousem DM. Parental leave policy in radiology residency programs: current status. J Am Coll Radiol 2020 Sep;17(9):1163-1171. [doi: 10.1016/j.jacr.2019.12.032] [Medline: 32275902]
54. Gracey LE, Cronin M, Shinkai K, Mathes EF. Program director and resident perspectives on new parent leave in dermatology residency. JAMA Dermatol 2018 Oct 1;154(10):1222-1225. [doi: 10.1001/jamadermatol.2018.2878] [Medline: 30167642]
55. Hariton E, Matthews B, Burns A, Akileswaran C, Berkowitz LR. Pregnancy and parental leave among obstetrics and gynecology residents: results of a nationwide survey of program directors. Am J Obstet Gynecol 2018 Aug;219(2):199. [doi: 10.1016/j.ajog.2018.04.017] [Medline: 29673570]
56. Holliday EB, Ahmed AA, Jagsi R, et al. Pregnancy and Parenthood in Radiation Oncology, Views and Experiences Survey (PROVES): results of a blinded prospective trainee parenting and career development assessment. Int J Radiat Oncol Biol Phys 2015 Jul;92(3):516-524. [doi: 10.1016/j.ijrobp.2015.02.024]
57. Huh DD, Wang J, Fliotsos MJ, et al. Association between parental leave and ophthalmology resident physician performance. JAMA Ophthalmol 2022 Nov 1;140(11):1066-1075. [doi: 10.1001/jamaophthalmol.2022.3778] [Medline: 36173610]
58. Humphrey VS, Wyant WA, Brag KO. Family planning influences and perceptions among dermatology residents from ACGME-accredited training programs: a survey-based study. Arch Dermatol Res 2024 May 11;316(5):159. [doi: 10.1007/s00403-024-02882-9] [Medline: 38734865]
59. Huynh M, Wang A, Ho J, Herndon CN, Aghajanova L. Fertility preservation and infertility treatment in medical training: an assessment of residency and fellowship program directors' attitudes. Womens Health Rep (New Rochelle) 2021;2(1):576-585. [doi: 10.1089/whr.2021.0044] [Medline: 35141706]
60. Li RD, Janczewski LM, Eng JS, et al. Pregnancy and parenthood among US surgical residents. JAMA Surg 2024 Oct 1;159(10):1127-1137. [doi: 10.1001/jamasurg.2024.2399] [Medline: 39018050]
61. Janczewski LM, Buchheit JT, Golisch KB, et al. Contemporary evaluation of work-life integration and well-being in US surgical residents: a national mixed-methods study. J Am Coll Surg 2024 Dec 1;239(6):515-526. [doi: 10.1097/XCS.0000000000001135] [Medline: 38920301]
62. Kling SM, Slashinski MJ, Green RL, Taylor GA, Dunham P, Kuo LE. Parental leave experiences for the non-childbearing general surgery resident parent: a qualitative analysis. Surgery 2024 Nov;176(5):1320-1326. [doi: 10.1016/j.surg.2024.04.035] [Medline: 38910045]
63. Kraus MB, Malinzak EB, Chandrabose R, et al. A nationwide cross-sectional survey of anesthesiology fellowship program directors: attitudes on parental leave in residency and fellowship training. Womens Health Rep (New Rochelle) 2022;3(1):395-404. [doi: 10.1089/whr.2021.0130] [Medline: 35652001]
64. Kraus MB, Thomson HM, Dexter F, et al. Pregnancy and motherhood for trainees in anesthesiology: a survey of the American Society of Anesthesiologists. J Educ Perioper Med 2021;23(1):E656. [doi: 10.46374/volxxiii_issue1_kraus] [Medline: 33778101]
65. Leandre FM, Sudak DM, Ginory A. Are psychiatry programs providing adequate parental leave to their residents? Acad Psychiatry 2022 Apr;46(2):162-166. [doi: 10.1007/s40596-021-01558-0] [Medline: 34751936]
66. Lumpkin ST, Klein MK, Battarbee AN, Strassle PD, Scarlet S, Duke MC. Fellowship or family? A comparison of residency leave policies with the Family and Medical Leave Act. J Surg Res 2019 Sep;241:302-307. [doi: 10.1016/j.jss.2019.03.004] [Medline: 31048221]
67. Magudia K, Campbell SR, Rangel EL, et al. Medical specialty board parental, caregiver, and medical leave policy updates after 2021 American Board of Medical Specialties Mandate. JAMA 2021 Nov 9;326(18):1867-1870. [doi: 10.1001/jama.2021.15871] [Medline: 34751719]
68. Malapati SJ, Idossa D, Singh SRK, et al. Parent penalty: parental leave experiences of trainees and early-career faculty in oncology subspecialties. In: Chino F, editor. JCO Oncol Pract 2023 Oct;19(10):899-906. [doi: 10.1200/OP.23.00242] [Medline: 37708434]

XSL•FO
RenderX

69. Maloni K, Calligaro K, Reed A, Vani K, Dougherty M, Troutman D. Attitudes regarding pregnancy among vascular surgery trainees and program directors based on a survey of vascular residents, vascular fellows, and the Association of Program Directors in Vascular Surgery. Vasc Endovascular Surg 2022 Aug;56(6):561-565. [doi: 10.1177/15385744221091384] [Medline: 35488400]
70. Mattei LH, Polan RM. Contemporary parental leave policies and practice among obstetrics and gynecology trainees and faculty. Clin Obstet Gynecol 2024 Sep 1;67(3):524-530. [doi: 10.1097/GRF.0000000000000876] [Medline: 38832715]
71. McDeavitt JT, Appelbaum NP, Raddatz MM, Driscoll SW, Kinney CL. Taking leave during residency: types of absences and subsequent delays and variations in physical medicine and rehabilitation medical board pass rates. Am J Phys Med Rehabil 2022 Jul 1;101(7 Suppl 1):S30-S34. [doi: 10.1097/PHM.0000000000002004] [Medline: 35706116]
72. Morgan JC, Owens T, Carmack MC, Braverman A, Class QA. Policy and perceptions of pregnancy during training among residents of various subspecialties. Postgrad Med J 2025 Apr 22;101(1195):441-446. [doi: 10.1093/postmj/qgae164] [Medline: 39574162]
73. Morris LE, Lindbloom E, Kruse RL, Washington KT, Cronk NJ, Paladine HL. Perceptions of parenting residents among family medicine residency directors. Fam Med 2018 Nov;50(10):756-762. [doi: 10.22454/FamMed.2018.978635] [Medline: 30428104]
74. Mundschenk MB, Krauss EM, Poppler LH, et al. Resident perceptions on pregnancy during training: 2008 to 2015. Am J Surg 2016 Oct;212(4):649-659. [doi: 10.1016/j.amjsurg.2016.06.018] [Medline: 27575602]
75. Mulcahey MK, Nemeth C, Trojan JD, O Connor MI. The perception of pregnancy and parenthood among female orthopaedic surgery residents. J Am Acad Orthop Surg 2019 Jul 15;27(14):527-532. [doi: 10.5435/JAAOS-D-18-00216] [Medline: 30499893]
76. Mwakyanjala EJ, Cowart JB, Hayes SN, Blair JE, Maniaci MJ. Pregnancy and parenting during cardiology fellowship. J Am Heart Assoc 2019 Jul 16;8(14):e012137. [doi: 10.1161/JAHA.119.012137] [Medline: 31286816]
77. Nguyen CV, Luong M, Weiss JM, Hardesty C, Karamitopoulos M, Poon S. The cost of maternity leave for the orthopaedic surgeon. J Am Acad Orthop Surg 2020 Nov 15;28(22):e1001-e1005. [doi: 10.5435/JAAOS-D-19-00337] [Medline: 32079849]
78. Oliveros E, Burgess S, Nadella N, et al. Becoming a parent during cardiovascular training. J Am Coll Cardiol 2022 May 31;79(21):2119-2126. [doi: 10.1016/j.jacc.2022.03.371] [Medline: 35618349]
79. Osborn VW, Doke K, Griffith KA, et al. A survey study of female radiation oncology residents' experiences to inform change. Int J Radiat Oncol Biol Phys 2019 Aug 1;104(5):999-1008. [doi: 10.1016/j.ijrobp.2019.05.013] [Medline: 31108141]
80. Ouyang K, Huang IA, Wagner JP, et al. Persistence of gender bias over four decades of surgical training. J Surg Educ 2021;78(6):1868-1877. [doi: 10.1016/j.jsurg.2021.06.008] [Medline: 34294569]
81. Parker SL, Conner CR, Mata DA, et al. Factors associated with pregnancy and perinatal outcomes in female neurosurgeons: a cross-sectional study. Neurosurg 2021;88(4):884-889. [doi: 10.1093/neuros/nyaa533]
82. Powell WT, Dundon KMW, Frintner MP, Kornfeind K, Haftel HM. Parenthood, parental benefits, and career goals among pediatric residents: 2008 and 2019. Pediatrics 2021 Dec 1;148(6):57-64. [doi: 10.1542/peds.2021-052931]
83. Rangel EL, Lyu H, Haider AH, Castillo-Angeles M, Doherty GM, Smink DS. Factors associated with residency and career dissatisfaction in childbearing surgical residents. JAMA Surg 2018 Nov 1;153(11):1004-1011. [doi: 10.1001/jamasurg.2018.2571] [Medline: 30073246]
84. Rangel EL, Castillo-Angeles M, Smith PM, Terhune KP. Parental support policies and benefits: a surgical training program report card of transparency. Ann Surg 2023 Jun 1;277(6):938-943. [doi: 10.1097/SLA.0000000000005566] [Medline: 35837953]
85. Rangel EL, Castillo-Angeles M, Changala M, Haider AH, Doherty GM, Smink DS. Perspectives of pregnancy and motherhood among general surgery residents: a qualitative analysis. Am J Surg 2018 Oct;216(4):754-759. [doi: 10.1016/j.amjsurg.2018.07.036] [Medline: 30072028]
86. Reilly G, Tipton C, Liberman P, Berkenstock M. Attitudes toward parental leave and breastfeeding during ophthalmology residency. Can J Ophthalmol 2022 Jun;57(3):175-187. [doi: 10.1016/j.jcjo.2021.02.039] [Medline: 33789088]
87. Reid DBC, Shah KN, Lama CJ, Kosinski LR, Daniels AH, Eberson CP. Parenthood among orthopedic surgery residents: assessment of resident and program director perceptions on training. Orthopedics 2021;44(2):98-104. [doi: 10.3928/01477447-20210201-08] [Medline: 33561867]
88. Ruse S, Bergman R, Crawford E. Pregnancy in orthopaedic residents: peripartum barriers identified. JB JS Open Access 2022;7(4):e22.00098. [doi: 10.2106/JBJS.OA.22.00098] [Medline: 36601291]
89. Sandler BJ, Tackett JJ, Longo WE, Yoo PS. Pregnancy and parenthood among surgery residents: results of the first nationwide survey of general surgery residency program directors. J Am Coll Surg 2016 Jun;222(6):1090-1096. [doi: 10.1016/j.jamcollsurg.2015.12.004] [Medline: 26776357]
90. Sanusi O, Kalu RU, Obayashi J, et al. Parental leave in neurosurgery: a US cross-sectional study. Neurosurg Pract 2024 Dec;5(4):e00116. [doi: 10.1227/neuprac.0000000000000116] [Medline: 39959550]
91. Sharpe EE, Ku C, Malinzak EB, et al. A cross-sectional survey study of United States residency program directors' perceptions of parental leave and pregnancy among anesthesiology trainees. Can J Anaesth 2021 Oct;68(10):1485-1496. [doi: 10.1007/s12630-021-02044-9] [Medline: 34159567]

XSL•FO
RenderX

92. Shifflette V, Hambright S, Amos JD, Dunn E, Allo M. The pregnant female surgical resident. Adv Med Educ Pract 2018;9(365–9):365-369. [doi: 10.2147/AMEP.S140738] [Medline: 29785149]
93. Szender JB, Grzankowski KS, Eng KH, Odunsi K, Frederick PJ. Evaluation of satisfaction with work-life balance among U.S. gynecologic oncology fellows: a cross-sectional study. Gynecol Oncol Rep 2016 Apr;16:17-20. [doi: 10.1016/j.gore.2016.03.001] [Medline: 27331129]
94. Thum JA, Chang D, Tata N, Liau LM. Neurosurgeons in 2020: the impact of gender on neurosurgical training, family planning, and workplace culture. Neurosurg Focus 2021 Mar 1;50(3):E11. [doi: 10.3171/2020.12.FOCUS20965]
95. Todd AR, Cawthorn TR, Temple-Oberle C. Pregnancy and parenthood remain challenging during surgical residency: a systematic review. Acad Med 2020 Oct;95(10):1607-1615. [doi: 10.1097/ACM.0000000000003351] [Medline: 32271231]
96. Varda BK, Glover M 4th. Specialty board leave policies for resident physicians requesting parental leave. JAMA 2018 Dec 11;320(22):2374-2377. [doi: 10.1001/jama.2018.15889] [Medline: 30535211]
97. Warner L, Guevara LH, Watson A, Farmer S, Mehta R, Homme J. Changes in attitudes, beliefs, and experiences related to pregnancy during graduate medical education training from 2005 to 2021. Biomol Biomed 2024 May 2;24(3):606-611. [doi: 10.17305/bb.2023.9865] [Medline: 38149830]
98. Wang KM, Lee B, Woreta FA, et al. Parental leave policy for ophthalmology residents: results of a nationwide cross-sectional study of program directors. J Surg Educ 2021 May;78(3):785-794. [doi: 10.1016/j.jsurg.2020.08.034]
99. Wilder JL, Pingree EW, Hark CM, et al. Pediatric trainees as parents: perspectives on parenthood from pediatric resident parents. Acad Pediatr 2021 Aug;21(6):934-942. [doi: 10.1016/j.acap.2021.04.009] [Medline: 33878479]
100. Wilder JL, Hark CM, Marcus CH, et al. Pediatric trainees as parents: perspectives from a pandemic. Acad Pediatr 2021 Aug;21(6):927-933. [doi: 10.1016/j.acap.2021.04.006] [Medline: 33872822]
101. Yong CM, Abnousi F, Rzeszut AK, et al. Sex differences in the pursuit of interventional cardiology as a subspecialty among cardiovascular fellows-in-training. JACC Cardiovasc Interv 2019 Feb 11;12(3):219-228. [doi: 10.1016/j.jcin.2018.09.036] [Medline: 30660463]
102. Zhang C, Garcia-Neuer M, Jogerst KM, Van Der Walt C, Chang Y, Nguyen MC. Reproductive planning and infertility: training the next generation of surgeons. Am J Surg 2024 Dec;238:115886. [doi: 10.1016/j.amjsurg.2024.115886] [Medline: 39128441]
103. Zmijewski P, Aleman C, Panzica N, et al. Does your gender impact resident operative experience? A multi- institutional qualitative study. J Surg Educ 2025 Feb;82(2):103368. [doi: 10.1016/j.jsurg.2024.103368] [Medline: 39709934]
104. Acker SN, Corbisiero MF, Romano J, et al. Attitudes of surgical trainees and faculty towards parental leave during surgical training. J Surg Educ 2024 Sep;81(9):1239-1248. [doi: 10.1016/j.jsurg.2024.06.012] [Medline: 38971678]
105. Mann H, Glazer T. Current state of safe pregnancy policies for the US surgical trainee. OTO Open 2024;8(3):e172. [doi: 10.1002/oto2.172] [Medline: 39036338]
106. Martin CE, Hipp HS, Kottke M, Haddad LB, Kawwass JF. Fertility, pregnancy, and postpartum: a survey of practicing Georgia obstetrician gynecologists. Matern Child Health J 2019 Oct;23(10):1299-1307. [doi: 10.1007/s10995-019-02801-9] [Medline: 31236824]
107. Tatarian T, Anderson B, Pucci MJ, et al. Impact of surgeon workload on pregnancy outcomes in surgical trainees and faculty. Ann Surg 2026 Jan 1;283(1):108-114. [doi: 10.1097/SLA.0000000000006426] [Medline: 38939968]
108. McGough E, Martin TW, White P, Garvan C, Esfandiary L, Morey TE. Perceptions of anesthesiology residents concerning parental leave during residency. J Clin Anesth 2022 Oct;81:110910. [doi: 10.1016/j.jclinane.2022.110910] [Medline: 35738027]
109. Castillo-Angeles M, Smink DS, Rangel EL. Perspectives of general surgery program directors on paternity leave during surgical training. JAMA Surg 2022 Feb 1;157(2):105-111. [doi: 10.1001/jamasurg.2021.6223] [Medline: 34851404]
110. Kenyon LE, Malik R, Rodriguez D, Carmel ME. Urology program directors' perception of pregnancy during residency. Urology 2021 Jul;153:75-80. [doi: 10.1016/j.urology.2020.10.072] [Medline: 33412219]
111. MacVane CZ, Puissant M, Fix M, et al. Scheduling practices for pregnant emergency medicine residents. AEM Educ Train 2022 Dec;6(6):e10813. [doi: 10.1002/aet2.10813] [Medline: 36425789]
112. Morris L, Cronk NJ, Washington KT. Parenting during residency: providing support for Dr mom and Dr dad. Fam Med 2016 Feb;48(2):140-144. [Medline: 26950787]
113. Tang AL, Miller A, Hauff S, et al. Maternity and paternity leave in otolaryngology residency training in the United States. Laryngoscope 2019 May;129(5):1093-1099. [doi: 10.1002/lary.27328] [Medline: 30315567]
114. Wallace CC, Edmunds RW, Bourne D, Wong L. Parenting in plastic surgery residency. Plast Reconstr Surg 2022;149(6):1465-1469. [doi: 10.1097/PRS.0000000000009134]
115. MacDonald SM, Raman JD. Widely variable parental leave practices for urology residency programs in the United States. Urology 2021 Jul;153:81-86. [doi: 10.1016/j.urology.2020.12.049]
116. Hui DHF, Yakub M, Tiwana S, et al. Gender of department chair and paid parental leave benefits in academic radiology residency programs. Curr Probl Diagn Radiol 2022;51(2):162-165. [doi: 10.1067/j.cpradiol.2021.08.004] [Medline: 34949474]
117. Morah O, Khosa F. Impact of department chair gender on paid parental leave across American anaesthesiology residencies. Leader 2025 Dec;9(4):376-380. [doi: 10.1136/leader-2024-001063]

XSL•FO
RenderX

118. Tao BKL, Hehar HK, Tien CW, Shunmugam M, Noureddin G, Khosa F. Gender of faculty leadership and online advertised parental leave policies for American ophthalmology residency programs. JFO Open Ophthalmol 2024 Jun;6:100106. [doi: 10.1016/j.jfop.2024.100106]
119. Wendling A, Paladine HL, Hustedde C, Kovar-Gough I, Tarn DM, Phillips JP. Parental leave policies and practices of US family medicine residency programs. Fam Med 2019 Oct 4;51(9):742-749. [doi: 10.22454/FamMed.2019.773836] [Medline: 31465110]
120. Ben-Zion S, Lehmann A, Price L, Burnett HQ, Michelson CD. The use of parenting electives in pediatric residency. Acad Pediatr 2022;22(4):513-517. [doi: 10.1016/j.acap.2021.11.017] [Medline: 34864134]
121. Sherbaf FG, Lin DDM, Yousem DM. Parental leave and neuroradiology fellowships. AJNR Am J Neuroradiol 2020 Aug;41(8):1348-1354. [doi: 10.3174/ajnr.A6648] [Medline: 32646942]
122. Weiss J, Teuscher D. What provisions do orthopaedic programs make for maternity, paternity, and adoption leave? Clin Orthop Relat Res 2016 Sep;474(9):1945-1949. [doi: 10.1007/s11999-016-4828-x] [Medline: 27075331]
123. Stack SW, Finn KM, Kisielewski M, Law KL, Milne CK, Best JA. Parental leave policies in residency: a national survey of internal medicine program directors. Acad Med 2022 Jul 1;97(7):1021-1028. [doi: 10.1097/ACM.0000000000004593] [Medline: 35020617]
124. Engelbrecht-Wiggans EA, Sundel MH, Newland JJ, Seyoum N, Brown RF. Parental leave policies in general surgery residencies. Am J Surg 2024 Jul;233:25-28. [doi: 10.1016/j.amjsurg.2023.12.017] [Medline: 38160066]
125. Kasemodel K, Hemal K, Chen W, Wallace C, Bourne DA. Evolution in parental leave policies: plastic surgery and obstetrics/gynecology lead the way. Plast Reconstr Surg 2022 Nov 1;150(5):1160-1168. [doi: 10.1097/PRS.0000000000009610] [Medline: 36067473]
126. Magudia K, Ng TSC, Bick AG, et al. Parenting while in training: a comprehensive needs assessment of residents and fellows. J Grad Med Educ 2020 Apr;12(2):162-167. [doi: 10.4300/JGME-D-19-00563.1] [Medline: 32322349]
127. Mercurio AM, Lynch OL, Stein BES, et al. Very few ACGME-accredited orthopaedic surgery residency programs have web-accessible leave policies dedicated to parental leave for residents, despite ACGME requirements. Clin Orthop Relat Res 2024 Oct 1;482(10):1779-1785. [doi: 10.1097/CORR.0000000000003091] [Medline: 38662935]
128. Siljander BR, Van Nortwick SS, Flakne JC, Van Heest AE, Bohn DC. What proportion of orthopaedic surgery residency programs have accessible parental leave policies, and how generous are they? Clin Orthop Relat Res 2020;478(7):1506-1511. [doi: 10.1097/CORR.0000000000001041]
129. Wong TY, Huang JJ, Hoffmann JC, Flug JA, Cooke EA, Donnelly EF. Resident wellness in radiology as portrayed by departmental websites. Acad Radiol 2022 Aug;29(8):1259-1265. [doi: 10.1016/j.acra.2021.07.016] [Medline: 34400076]
130. American College of Obstetricians & Gynecologists. ACOG 75. URL: https://www.acog.org [accessed 2026-06-10]
131. General surgery leave policy and considerations on workplace support of pregnant surgeons. The American Board of Surgery. URL: https://www.absurgery.org/resources/abs-policies/policy-leave [accessed 2026-06-10]
132. Judge-Golden CP, Dotters-Katz SK, Weber JM, Pieper CF, Gray BA. Parenthood and medical training: challenges and experiences of physician moms in the US. Teach Learn Med 2024;36(1):43-52. [doi: 10.1080/10401334.2022.2141750] [Medline: 36370040]
133. Shreffler J, Petrey J, Huecker M. The impact of COVID-19 on healthcare worker wellness: a scoping review. West J Emerg Med 2020 Aug 17;21(5):1059-1066. [doi: 10.5811/westjem.2020.7.48684] [Medline: 32970555]
134. Finch SJ. Pregnancy during residency: a literature review. Acad Med 2003 Apr;78(4):418-428. [doi: 10.1097/00001888-200304000-00021] [Medline: 12691977]
135. Moyer CA, Compton SD, Kaselitz E, Muzik M. Pregnancy-related anxiety during COVID-19: a nationwide survey of 2740 pregnant women. Arch Womens Ment Health 2020 Dec;23(6):757-765. [doi: 10.1007/s00737-020-01073-5] [Medline: 32989598]
136. Bakkensen JB, Smith KS, Cheung EO, et al. Childbearing, infertility, and career trajectories among women in medicine. JAMA Netw Open 2023 Jul 3;6(7):e2326192. [doi: 10.1001/jamanetworkopen.2023.26192] [Medline: 37498595]
137. Pendleton KM, Salles A. Maternity leave: not just a challenge for trainees. Acad Med 2020 Apr;95(4):489. [doi: 10.1097/ACM.0000000000003164] [Medline: 32209845]
138. Kin C, Yang R, Desai P, Mueller C, Girod S. Female trainees believe that having children will negatively impact their careers: results of a quantitative survey of trainees at an academic medical center. BMC Med Educ 2018 Dec;18(1):260. [doi: 10.1186/s12909-018-1373-1]
139. Garrison GM, Meunier MR, Boswell CL, Greenwood JD, Nordin T, Angstman KB. Continuity of care: a primer for family medicine residencies. Fam Med 2024 Feb;56(2):76-83. [doi: 10.22454/FamMed.2023.913197] [Medline: 38055847]
140. Love MM, Mainous AG 3rd, Talbert JC, Hager GL. Continuity of care and the physician-patient relationship: the importance of continuity for adult patients with asthma. J Fam Pract 2000 Nov;49(11):998-1004. [Medline: 11093565]
141. Porter J, Boyd C, Skandari MR, Laiteerapong N. Revisiting the time needed to provide adult primary care. J Gen Intern Med 2023 Jan;38(1):147-155. [doi: 10.1007/s11606-022-07707-x] [Medline: 35776372]
142. Porterfield L, Collazo A, Schlag KE, Rodríguez JE, Campbell KM. More technology, fewer boundaries, and physician burnout in academic primary care physicians. South Med J 2024 Oct;117(10):587-590. [doi: 10.14423/SMJ.0000000000001735] [Medline: 39366683]

XSL•FO
RenderX

143. Bye E, Leval R, Sayles H, Doyle M, Mathes M, Cudzilo-Kelsey L. Parental postpartum depression among medical residents. Arch Womens Ment Health 2022 Dec;25(6):1129-1135. [doi: 10.1007/s00737-022-01271-3] [Medline: 36434278]
144. Worthington RO, Adams DR, Fritz CDL, Tusken M, Volerman A. Supporting breastfeeding physicians across the educational and professional continuum: a call to action. Acad Med 2023 Jan 1;98(1):21-28. [doi: 10.1097/ACM.0000000000004898] [Medline: 35921171]
145. Tobin-Tyler E, Adashi EY. The ACGME's new paid family and medical leave policy: just the beginning. J Am Board Fam Med 2023 Feb 8;36(1):190-192. [doi: 10.3122/jabfm.2022.220246R1] [Medline: 36759135]
146. Nandi A, Jahagirdar D, Dimitris MC, et al. The impact of parental and medical leave policies on socioeconomic and health outcomes in OECD countries: a systematic review of the empirical literature. Milbank Q 2018 Sep;96(3):434-471. [doi: 10.1111/1468-0009.12340] [Medline: 30277601]

Abbreviations

ACGME: Accreditation Council for Graduate Medical Education
ACOG: American College of Obstetricians and Gynecologists
PD: program director
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
PRISMA-ScR: Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews
SEM: social ecological model

Edited by AL Marshall; submitted 08.Nov.2025; peer-reviewed by ERD Santos, V Minville; revised version received 03.Jun.2026; accepted 03.Jun.2026; published 02.Jul.2026.

Please cite as:
Santiago Delgado ZM, Vaughan EM, Jones ME, Jones F, Bridwell D, Porterfield LR
Parenthood in US Medical Training Across Specialty Groups: Scoping Review
JMIR Med Educ 2026;12:e87284
URL: https://mededu.jmir.org/2026/1/e87284
doi:10.2196/87284

© Zuleica M Santiago Delgado, Elizabeth M Vaughan, Mary E Jones, Forrest Jones, Dylan Bridwell, Laura R Porterfield. Originally published in JMIR Medical Education (https://mededu.jmir.org), 2.Jul.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Education, is properly cited. The complete bibliographic information, a link to the original publication on https://mededu.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Physical Attractiveness and Chances of Being Invited to Interview With a Medical Residency Program: Retrospective Cohort Study

Daria Hunter[1], MD; Luke Hunter[2], MA, MD; David Kerner[3], MD; Aidan Mullan[1], MA; Derek Vanmeter[1], BS; Ivan Khapov[4], MD; Alexander Finch[1], MD; Sara Hevesi[1], MD; Ivan Porter II[5], MD; Colin West[1], MD, PhD; James Homme[1], MD

[1]Mayo Clinic, 200 1st St SW, Rochester, MN, United States

[2]Olmsted Medical Center, Rochester, MN, United States

[3]Spartanburg Regional Medical Center, Spartanburg, SC, United States

[4]Geisinger Medical Center, Denvile, PA, United States

[5]Mayo Clinic, Jacksonville, FL, United States

Corresponding Author:
Daria Hunter, MD
Mayo Clinic, 200 1st St SW, Rochester, MN, United States

Abstract

Background: Applicants participating in the Residency Match generally submit a photograph through the Electronic Residency Application Service (ERAS). Studies demonstrate that subjectively more attractive applicants are more likely to succeed during job recruitment, including a paper related to the Residency Match.

Objective: This study further investigates the relationship between an applicant's attractiveness and the likelihood that they are invited to interview with a residency program to explore if more attractive applicants are more likely to be invited to interview when controlled for demographic and academic variables. If there are enough data suggesting that an ERAS photograph being visible prior to the interview gives an unfair advantage to more attractive applicants, this practice might be reconsidered by some residency programs or by ERAS itself.

Methods: Residency directors were surveyed on application review practices. Programs that viewed ERAS photographs prior to deciding whether to invite an applicant to interview were asked to share ERAS files of all reviewed applicants of the 2022 Match. A machine learning model was used to determine attractiveness scores for ERAS photographs. The scores ranged from 1 to 10, where 1 represents the least attractive and 10 represents the most attractive. Multivariable logistic regression analysis was performed considering attractiveness scores, demographics, and professional characteristics. The primary outcome of interest was an invitation to an interview with a residency program.

Results: The residency program response rate was 47.5% (29/61). Among 2681 unique applications to 10 specialties in a single academic health system, the median attractiveness score for all applicants was 6.02 (IQR 5.54-6.55). The univariable analysis indicated a 19% higher invitation likelihood with a 1-point increase in attractiveness. After adjusting for demographics and professional experiences, the association lost statistical significance. Additional adjustment for United States Medical Licensing Examination scores further attenuated the association.

Conclusions: While higher attractiveness scores correlated with an increased likelihood of securing an interview, this correlation was not statistically significant after adjusting for other variables.

(JMIR Med Educ 2026;12:e81052) doi:10.2196/81052

KEYWORDS

medical residency; retrospective cohort study; medical residents; attractiveness; residency; interview

Introduction

The practice of recruitment to medical residencies varies significantly between countries. It ranges from being mostly based on test scores to a series of in-person interviews. However, in most cases, the recruitment of new residents consists of an initial application that may include a professional photograph, followed by an interview in some cases, culminating in a decision on whether an applicant would be a good match. This study focuses on graduate medical education recruitment in the United States due to the location of our institution and US processes being well established and centralized. However, the results of the study may be relevant for other countries and institutions that review applicants' photographs as part of the residency recruitment process.

XSL•FO
RenderX

Most medical residency positions in the United States are filled via the National Residency Matching Program (NRMP), a system that helps to "match" residency applicants and programs utilizing a proprietary "matching algorithm" that is based on the stable marriage problem—a method of finding an optimal match between 2 equally sized cohorts, given an ordering of preferences for each element of both cohorts [1]. The algorithm takes into account the ranked preferences of both parties and then, through an iterative process, strives to find the match that maximizes overall satisfaction. In 2022, there were 48,156 registered NRMP applicants competing for 37,425 first-year and 2950 second-year positions [2]. Residency applicants create Electronic Residency Application Service (ERAS) applications to apply directly to individual programs. Residency programs use a variety of methodologies to review ERAS applications to select applicants to invite for a formal interview. After interviews are completed, both applicants and programs create Rank Order Lists using the NRMP software. Based on the Rank Order Lists, the matching algorithm pairs applicants and programs. Applicants are notified of the results on Match Day in cases of a successful match. This process is often referred to as "The Match" [2].

Applicants participating in The Match are advised to submit a professional photograph through their online application, along with United States Medical Licensing Examination (USMLE) scores, medical school information, Medical Student Performance Evaluation, letters of recommendation, and curriculum vitae. Public data on what percentage of applicants choose to submit a photograph in their applications are unknown. Based on the data available from our institution, most applicants include a photograph in their application. Programs may voluntarily blind themselves to these photographs through any part of application review, though there are currently no available national data on the utilization of this practice.

Physical attractiveness is generally known as a person's physical features that are considered esthetically pleasing or beautiful. Robust literature describes the relationship between candidate attractiveness and competitive advantage in a variety of domains, with examples including improved self-image and psychological well-being [3], better treatment by other people [4], higher income [5,6], and even differences in judgment and recommended penalty assessment in a simulated court of law [7]. Studies demonstrate the positive effect of attractiveness on academic performance for trainees [8] and, similarly, on instructor performance evaluations [9,10]. The job market is no exception, as there is evidence that more attractive applicants may be more likely to receive an interview offer [11,12]. Job candidate appearance, including physical attractiveness, attire, posture, and nonverbal communication, plays a significant role in shaping first impressions and can set a competitive advantage affecting hiring outcomes. Available data suggest that physical attractiveness may influence the employment process across multiple fields [11,13-16]. There is substantial empirical evidence that physical attractiveness increases the likelihood of being hired [11,16,17] and might positively affect perception of a candidate's achievements [18]. This bias, known in psychology as the "beautiful is good" phenomenon or the "halo effect," has been demonstrated in numerous studies and reviews and found in a variety of different cultures [19-22]. Some evidence suggests that for certain positions traditionally dominated by either male or female candidates, attractive applicants might be evaluated less favorably [23,24].

Hiring in the medical field might differ from hiring in most other professional domains due to the high-stakes nature of clinical work, strict regulatory requirements, and the ethical obligation to prioritize patient safety and competence. NRMP and ERAS platforms were designed to emphasize more objective criteria, including academic and examination performance, evaluations, and extracurricular activities, thereby reducing the influence of superficial characteristics such as physical appearance. Having said that, program directors are often required to review thousands of applications, many of which reflect similar academic metrics, test scores, and clinical experiences. In this context of applicant homogeneity and cognitive overload, subjective factors may regain influence. While the literature on physical attractiveness and its effect on job and academic prospects is abundant, limited data are available in the setting of medical residency. This is despite the fact that the medical residency selection process, unlike many other fields, almost universally utilizes professional photographs as a part of initial applications.

A study published in 2019 by Maxfield et al [25] explored this question in relation to the residency match process. The authors created mock radiology residency match applications utilizing open-access professional photographs. Academic characteristics were randomly assigned to the mock applicants. The authors found that more attractive applicants were likely to receive a higher score from volunteer core faculty members across multiple radiology residency programs than less attractive applicants when controlling for other variables [25]. To our knowledge, this is the only paper investigating the question of physical appearance in medical education and recruitment. There have been no studies of this question utilizing native applications.

While having a professional photograph as part of a residency application serves a clear purpose of making the applications and interviews easier to remember, as residency recruitment committees may go through thousands of applications and interview hundreds of applicants during a single Match season, it is unclear if having a photograph available for review prior to deciding whether to invite an applicant to interview may introduce a bias, as suggested by the results of the publication by Maxfield et al [25]. The goal of this study is to further evaluate the issue of applicants' physical attractiveness in the residency match interview selection process utilizing native Match data as opposed to mock applications.

Methods

Ethical Considerations

This study was approved by the Mayo Clinic Educational Research Committee and the Mayo Clinic Institutional Review Board (IRB), IRB ID 23 - 002393. For responding to residency recruitment committee members, a waiver of informed consent was utilized with IRB approval, as only those willing to

XSL•FO
RenderX

participate filled out the survey and provided requested data. All participants were informed that they were free to withdraw from the study at any time. For residency applicants, informed consent was obtained by ERAS at the time of registration. Data for both program leadership members and ERAS applicants were anonymized for privacy purposes. Our study complied with ERAS data policy, including only conducting research after the respective Match Day. For the manuscript, we utilized the recommendations for reporting machine learning (ML) analyses in clinical research guidelines [26]. There was no compensation for participating in the study.

Study Question and Intended Use of the Results

Our goal was to correlate candidate attractiveness as ranked by an ML algorithm, with the likelihood of receiving an interview invitation using primary residency applications (rather than simulated applications). To achieve this goal, we first surveyed 61 program directors from all residency programs in our institution, comprising 3 main academic sites in Minnesota, Arizona, and Florida, and multiple smaller community sites across these states. Twenty-nine program directors volunteered to participate in the survey. Questions in the survey were related to the residency interview selection process and included questions clarifying if the program was blinded to candidate photographs and if candidates' physical attractiveness was considered when extending interview invitations. As the focus of our research was to evaluate the relationship between physical appearance and chances to be invited to interview, programs that reported waiving their right to view ERAS photographs were not included in the study. Twenty programs that stated that they were not blinded to the photographs were asked to submit ERAS applications, including the photographs of all applicants who were reviewed for interview consideration. Of note, during the study period, the electronic application system used by emergency medicine was the ERAS. During the 2025 to 2026 residency recruitment cycle, this changed to residency centralized application service. Specifically, we only solicited ERAS applications of those applicants who passed initial "filters" and were reviewed holistically. For the study respondents who were surveyed and asked to submit program data, informed consent was implied by their participation in the survey and sharing of ERAS applications. Program leadership teams surveyed were provided with a summary of the study design, including information on the risks and benefits of participating in the research and the right to withdraw from participation at any time. While we provided a general description of the study and research questions, we did not emphasize the main research question. This was done to prevent potential self-selection bias.

For included participants, informed consent for research uses of application materials was obtained by ERAS and NRMP at the beginning of the Match season, which 100% of the applicants who participated in the Match through ERAS signed. We also contacted ERAS directly and confirmed that our specific study was compliant with their data policy. The intended use of the results was to raise the question of whether professional photographs should be available for programs to see prior to the decision of whether to invite a candidate to interview being made.

Available Data

Ten of the 20 programs provided complete data. From the 10 programs that provided complete data, we received 2694 unique ERAS applications across diverse programs, including diagnostic radiology, emergency medicine, family medicine, obstetrics and gynecology, pediatrics, physical medicine and rehabilitation, radiation oncology, and vascular surgery. There was 1 program in Arizona and 9 at the main academic site in Rochester, MN. From this pool, 13 total applications (<0.005%) were excluded (8 due to the absence of an uploaded photograph and 5 for having uploaded extremely low-resolution images [<50 kB] that prevented any meaningful human and/or computer assessment of attractiveness). The variables of interest, including photographs, USMLE scores, medical school, and demographic data (race, gender, age, visa status), were extracted from the ERAS applications.

ML Method and Rationale

Since attractiveness is a highly subjective measure, there is no defined gold standard. In addition, individuals of varying demographics and backgrounds may value different features in terms of attractiveness. Therefore, instead of using a relatively small focus group of individuals to manually rank applicants' attractiveness, we employed an ML algorithm that was trained on ratings by hundreds of individuals with distinct backgrounds, incorporating the opinions of a much larger and more diverse group of human raters.

We utilized the MEBeauty Python software package to create a race-sensitive ML model (meaning that the model determines the race of the person in each image and compares it within its race rather than mixing it with all races, given different facial structures and beauty standards between racial groups) that could autonomously estimate attractiveness scores for human faces in photographs. The software package includes a database of 2550 images of Black, Asian, Caucasian, Hispanic, Indian, Mideastern, male, and female faces, scored from 1 (least beautiful) to 10 (most beautiful) by approximately 300 individuals with diverse cultural and social backgrounds. According to the original paper, the scores for each image assigned by different raters fell between 1.5 and 2 SDs [27].

Evaluation Measures, Training Protocols, and Validation

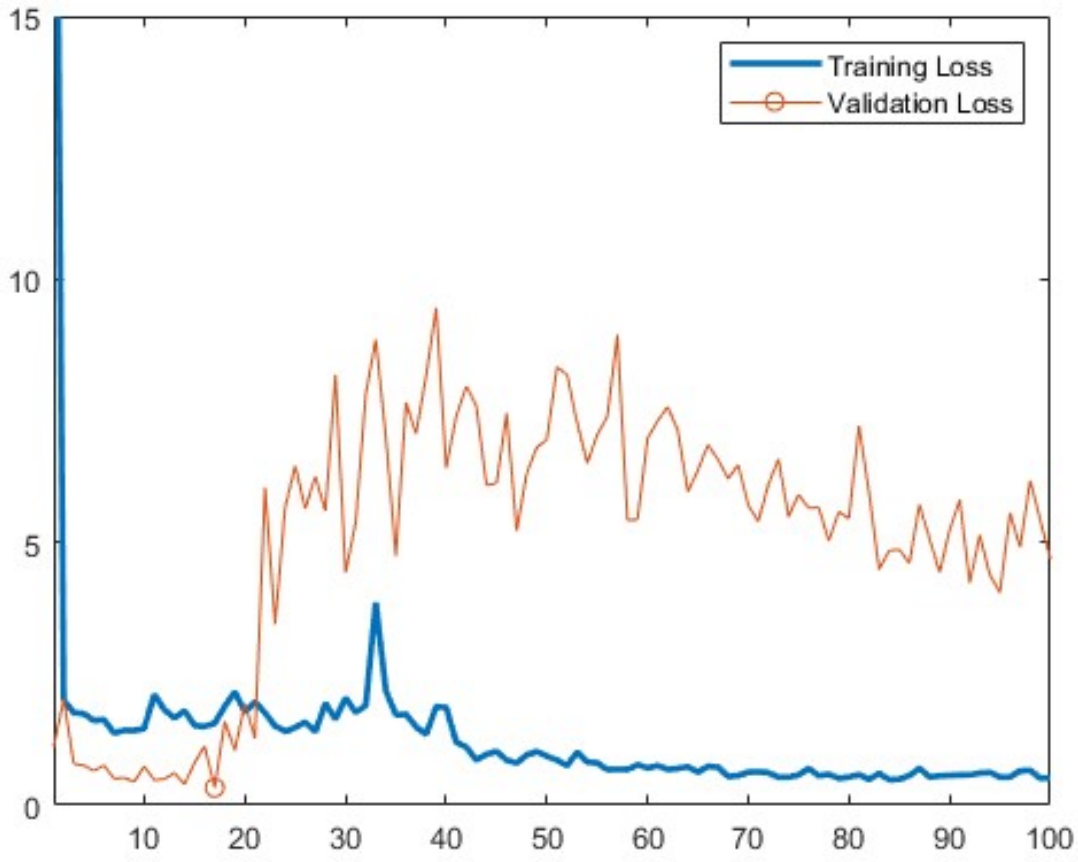
Using default MEBeauty parameters, 2681 images were preprocessed by a multitask cascaded convolutional neural network that standardized the images by vertically aligning and cropping faces from the full image.

Model training and validation were performed to identify the best prediction model and to avoid overtraining or undertraining. Native MEBeauty code was used to perform this task. All models processed 16 images at a time and iteratively refined the prediction of attractiveness using a maximum of 100 passes on the training data. The use of each database image as either a training or a validation image was predetermined by the MEBeauty database. Therefore, the model was developed using an existing database of images and then applied to the applicants' photographs. Several different model types were assessed: VGG-16, MobileNet, DenseNet, and AlexNet. All

XSL•FO
RenderX

models are open-access convolutional neural networks that are commonly used for image classification, but not exclusively for facial images. Therefore, several different models were trained and tested to select the best-performing network for the purposes of this paper. There was broad ranking agreement between the different candidate model types, with the VGG-16 yielding the best performance on validation images (validation loss 0.329, Figure 1) when compared with the other model types (MobileNet=0.940, DenseNet=0.510, and AlexNet=0.581). Therefore, this VGG-16 model was chosen as the facial beauty score predictor for the novel residency applicant images assessed that we acquired and assessed. Physical attractiveness scores were assigned to each applicant based on model prediction, and these data were combined with the rest of the variables extracted from the ERAS applications.

Figure 1. Training and validation loss for each epoch using the VGG-16 model.



Measurements and Outcomes

The primary outcome of interest was whether a residency applicant was invited to an interview with a program. Applicant photographs were obtained from the ERAS applications, and attractiveness was determined using the trained ML model. Applicant characteristics, including age, gender, race, residency region (Minnesota, Midwest United States, and non-Midwest United States), visa status, USMLE scores, and relevant research or volunteer experience, were also retrieved from ERAS applications.

All applicants were divided into 2 cohorts: those who were invited to interview with one of the residency programs participating in this research and those who were not. We did not consider whether an applicant accepted the interview offer. In our sample, less than 1% of applicants interviewed with more than one participating residency program.

Statistical Analysis

Numeric features were summarized with medians and IQRs; categorical features were summarized with frequency counts and percentages. For USMLE score analysis, we used national averages as a cutoff. Exploratory analysis compared applicant physical attractiveness based on applicant education, research or volunteer experience, visa status, and USMLE scores using 2-sided Kruskal-Wallis tests.

The association between artificial intelligence (AI)–determined physical attractiveness and invitation to interview was assessed using logistic regression. Models were unadjusted, adjusted only for the applicants' demographics (age, gender, race, and address location categorized as Minnesota, Midwest US, and non-Midwest US), and fully adjusted for demographics, medical credentials (Medical Doctor [MD] degree, Doctor of Osteopathic Medicine [DO] degree, or International Medical Graduate), relevant experience (number of research opportunities, number of volunteer opportunities), and USMLE step 1 and step 2 test

XSL•FO
RenderX

scores. Model results were reported using odds ratios (ORs) with 95% CIs. The association between invitation to interview and all applicant characteristics included in the fully adjusted model was reported in addition to the association with applicant attractiveness.

A minority of applicants did not report USMLE step 1 (16.9%) or USMLE step 2 (11.1%) or both scores and were excluded from this model. Sensitivity analysis included the applicants with missing USMLE scores as a standalone "Not Reported" group.

Secondary analysis considered each residency specialty individually when assessing the impact of AI-attractiveness on being invited to interview. Only candidates reviewed for each specialty were included. All specialties were considered during univariable analysis. To ensure that the adjusted regression models had sufficient statistical power, only specialties with at least 300 reviewed candidates were included for multivariable analysis.

Statistical significance was set at a threshold of α=.05. All analyses were conducted using R (version 4.2.2; R Foundation for Statistical Computing).

Results

General Results

Twenty-nine program directors out of 61 (47.5%) volunteered to participate in the survey. Of the 29 respondents, 20 indicated that their recruitment committee was not blinded to the applicants' photographs. All 29 respondents indicated that they did not consider applicants' physical appearance when deciding whether to invite them to interview. Ten programs provided complete applicant data, which were included in the final analysis.

A total of 2681 candidates were reviewed for at least one residency program. Across all specialties that participated in the study, the median applicant age was 27 (IQR 26 - 29) years, 1351 (50.4%) applicants were women, 1346 (50.2%) self-identified as White, and 184 (6.9%) required visa sponsorship. The median step 2 score was 249 (IQR 239 - 259), which is consistent with the national average [28]. For other academic characteristics, study applicants were similar to the national averages as reported by the NRMP [29-31] and are summarized in Table 1.

XSL•FO
RenderX

Table . Summary of applicant demographics, United States Medical Licensing Examination scores, and experiences (N=2681).

Demographics	Values
Age (y), median (IQR)	27 (26-29)
Gender, n (%)	
Women	1351 (50.4)
Men	1322 (49.3)
Other	7 (0.3)
Unknown	1 (0.0)
Race, n (%)	
American Indian	31 (1.2)
Asian	462 (17.2)
Asian Indian	283 (10.6)
Black or African American	169 (6.3)
Hispanic	208 (7.8)
White	1346 (50.2)
Other	182 (6.8)
Applicant type, n (%)	
MD[a]	1950 (72.7)
DO[b]	407 (15.2)
Non-US IMG[c]	237 (8.8)
US IMG	85 (3.2)
Other	2 (0.1)
Address location, n (%)	
Minnesota	143 (5.3)
Midwest (MN[d], IA[e], WI[f], ND[g], SD[h])	343 (12.8)
Home institution affiliation, n (%)	
Main institutional location	15 (0.6)
Across all sites of the institution	26 (1.0)
Authorized to work in the United States, n (%)	
No	106 (4.0)
Yes	2575 (96.0)
Visa sponsorship needed, n (%)	
No	24 (0.9)
Yes	184 (6.9)
Unknown/not applicable	2473 (92.2)
AI[i]-determined physical attractiveness, 1 - 10, median (IQR)	6.02 (5.54-6.55)
USMLE[j] scores and relevant experience	
Step 1 attempted, n (%)	
No	274 (10.2)
Yes	2407 (89.8)
Step 1 score, median (IQR)	238 (224-248)
Step 2 CK[k] attempted, n (%)	
No	299 (11.2)

XSL•FO
RenderX

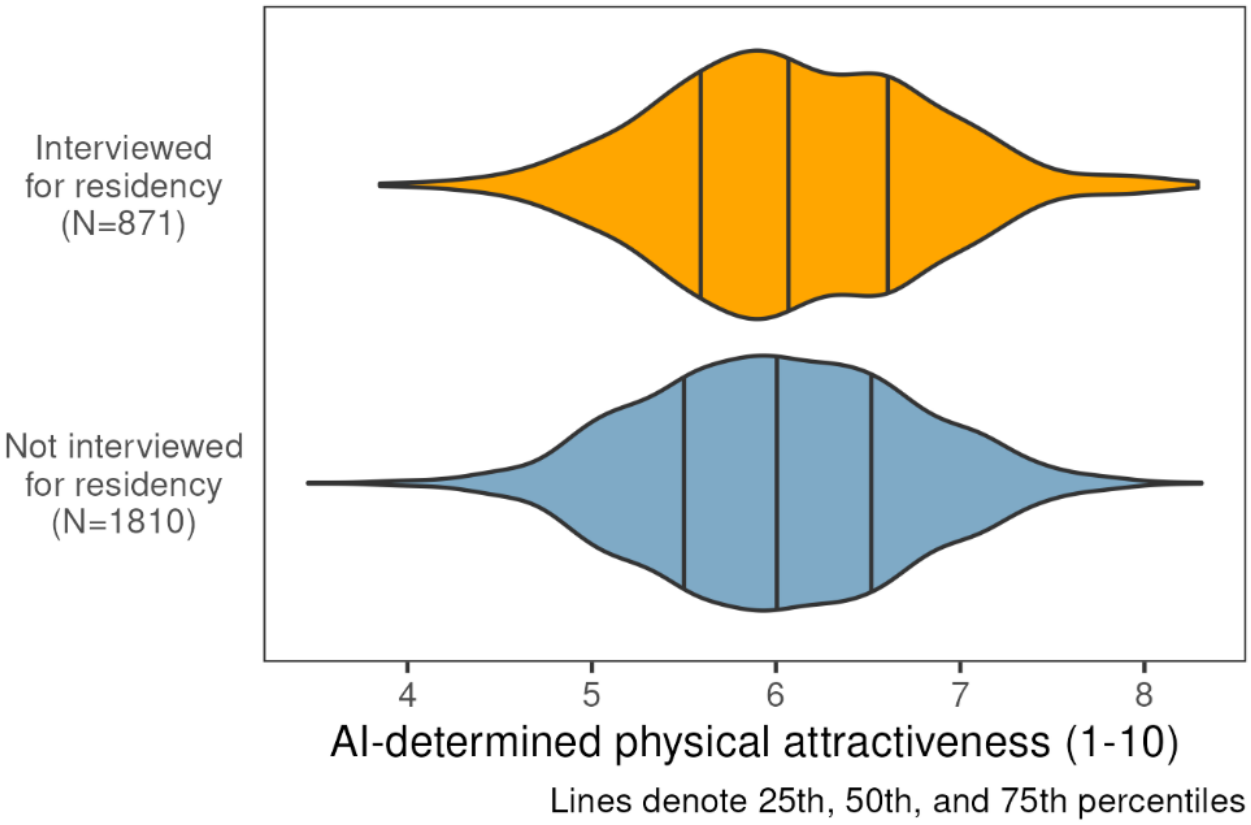
Demographics	Values
Yes	2382 (88.8)
Step 2 CK score, median (IQR)	249 (239-259)
Step 2 CS[l] attempted, n (%)	
No	2589 (96.6)
Yes	92 (3.4)
Step 2 CS result, n (%)	
Pass	92 (100)
Fail	0 (0)
Step 3 attempted, n (%)	
No	2555 (95.3)
Yes	126 (4.7)
Step 3 score, median (IQR)	226 (212-236)
Number of research experiences, median (IQR)	3 (2-5)
Number of volunteer experiences, median (IQR)	8 (5-11)

[a]MD: Doctor of Medicine.
[b]DO: Doctor of Osteopathic Medicine.
[c]IMG: International Medical Graduate.
[d]MN: Minnesota.
[e]IA: Iowa.
[f]WI: Wisconsin.
[g]ND: North Dakota.
[h]SD: South Dakota.
[i]AI: artificial intelligence.
[j]USMLE: United States Medical Licensing Examination.
[k]CK: clinical knowledge.
[l]CS: clinical skills.

Of the 2681 reviewed applications, 871 candidates were invited to an interview. The median AI-determined attractiveness score among all candidates was 6.02 (IQR 5.54 - 6.55). AI-determined physical attractiveness score distribution among candidates interviewed and not interviewed for medical residency is available in Figure 2.

XSL•FO
RenderX

Figure 2. Artificial intelligence (AI)–determined physical attractiveness among candidates interviewed and not interviewed for medical residency.



ML Attractiveness Grading

MEBeauty AI is subject to the same human biases as the people who trained it, as described in the original paper [27]. In our data, male candidates were assigned lower attractiveness scores than female candidates, with the median score for men being 5.67 (IQR 5.27-6.02) and for women being 6.48 (IQR 6.0-6.84), *P*<.001. In addition, White applicants received slightly higher scores than non-White applicants (median score for White candidates was 6.07, IQR 5.61-6.58, and for non-White candidates was 5.98, IQR 5.48-6.50; *P*<.001), although the difference was very small.

Association Between Attractiveness and Interview Invitation

Univariable logistic regression found that a 1-point increase in attractiveness was associated with 19% higher odds of being invited to interview for residency (OR 1.19, 95% CI 1.06 - 1.33; *P*=.003). However, when only adjusting for applicant demographics (age, gender, race, and address), there was no significant association between attractiveness and being selected to interview (OR 1.06, 95% CI 0.92 - 1.22; *P*=.45). Further adjusting for applicant certification, research and volunteer experience, and USMLE scores similarly found no association between attractiveness score and interview invitation (OR 1.01, 95% CI 0.86 - 1.20; *P*=.87; Table 2).

XSL•FO
RenderX

Table . Multivariable associations between applicant characteristics and being invited to interview for any specialty.

Applicant characteristic	Odds ratio (95% CI)	*P* value
Age, per year	1.06 (1.02-1.10)	.003
Gender (reference: women)		
Men	0.73 (0.57-0.95)	.02
Race (reference: White)		
Not White	1.02 (0.83-1.26)	.86
Applicant credentials (reference: no MD[a])		
MD	2.74 (2.03-3.70)	<.001
Address (reference: Minnesota)		
In the Midwest, non-Minnesota	0.31 (0.17-0.56)	<.001
Outside the Midwest	0.13 (0.08-0.20)	<.001
Authorized to work	1.22 (0.66-2.25)	.52
Research experience, per experience	1.05 (1.01-1.10)	.02
Volunteer experience, per experience	1.05 (1.02-1.07)	<.001
USMLE[b] step 1 score, per 10 points	1.04 (0.95-1.14)	.42
USMLE step 2 CK[c] score, per 10 points	1.30 (1.17-1.45)	<.001
AI[d] attractiveness, per 1 point	1.01 (0.86-1.20)	.87

[a]MD: Doctor of Medicine.

[b]USMLE: United States Medical Licensing Examination.

[c]CK: clinical knowledge.

[d]AI: artificial intelligence.

A sensitivity analysis including applicants with missing USMLE scores as a separate "not reported" group found similar results (OR 1.05, 95% CI 0.91 - 1.23; *P*=.49). When only adjusting for applicant academic and visa credentials other than USMLE scores (being an MD as opposed to DO, authorized to work in the United States, number of research experiences, and number of volunteer experiences), there was similarly no statistically significant association between attractiveness and getting an interview (OR 1.10, 95% CI 0.98 - 1.25; *P*=.11).

Association Between Applicant Characteristics and Interview Invitation

On the multivariable analysis including USMLE scores, we found that male candidates were less likely to be invited to interview than female candidates (OR 0.73, 95% CI 0.57 - 0.95; *P*=.02) and MD candidates were more likely to be interviewed than DOs and International Medical Graduates (OR 2.74, 95% CI 2.03 - 3.70; *P*<.001). The vast majority of analyzed ERAS applications came from Minnesota (9/10 residency programs), and applicants from the Midwest excluding Minnesota (where our institution is located) were less likely to be selected to interview (OR 0.31, 95% CI 0.17 - 0.56; *P*<.001) compared to candidates from Minnesota, as were applicants outside of the Midwest (OR 0.13, 95% CI 0.08 - 0.20; *P*<.001). There was a 6% increase in the odds of receiving an interview invitation per additional year of age (OR 1.06, 95% CI 1.02 - 1.10; *P*=.003), a 5% increase per additional research experience (OR 1.05, 95% CI 1.01 - 1.10; *P*=.02), a 5% increase per additional volunteer experience (OR 1.05, 95% CI 1.02 - 1.07; *P*<.001), and a 30% increase per 10 additional points on the USMLE step 2 clinical knowledge (CK) score (OR 1.30, 95% CI 1.17 - 1.45; *P*<.001). Race, work authorization in the United States, and USMLE step 1 score did not demonstrate statistically significant associations with being selected to interview.

Association Between Attractiveness and Other Variables

Examining attractiveness across applicant credentials, we found that MD candidates were rated as more attractive (median attractiveness 6.06 vs 5.94; Kruskal-Wallis *P*<.001), candidates with more research were more attractive (0 - 3 vs 4+ research experiences: median attractiveness 6.08 vs 5.98; *P*<.001), and candidates with more volunteering were more attractive (9+ vs 0 - 8 volunteer experiences: median attractiveness 6.19 vs 5.91; *P*<.001). There was no correlation between attractiveness and being authorized to work in the United States (median attractiveness 6.02 vs 6.01; *P*=.61). There was a small negative association between AI-attractiveness and USMLE step 1 score (step 1 score 239+ vs <239; median attractiveness 5.95 vs 6.10; *P*<.001), but there was no association between AI-attractiveness and USMLE step 2 CK score (step 2 CK score 250+ vs <250; median attractiveness 6.05 vs 6.00; *P*=.10) or step 3 score (step 3 score 227+ vs <227; median attractiveness 5.95 vs 5.79; *P*=.52).

When assessing program-specific data, some programs were demonstrated to be more likely to invite more attractive applicants to interview, but after adjusting for demographic and

XSL•FO
RenderX

objective academic parameters, no association of physical attractiveness with invitation to interview was found across programs.

Discussion

General Discussion

This study of native ERAS applications across 10 residency programs in an academic institution with locations in Minnesota and Arizona demonstrates an association between physical appearance and chances of being invited to interview with a residency program that disappears when adjusted for demographic and academic characteristics. While our research demonstrated a statistically significant unadjusted correlation between physical attractiveness and chances of being invited to a medical residency interview, this association was small and disappeared after adjusting for academic and demographic variables.

Our work builds on the contributions of others in this field. Maxfield et al [25] randomly assigned academic characteristics to open-access professional photographs and demonstrated a bias in favor of more attractive applicants when controlled for objective variables such as USMLE step scores. Regarding demographic variables, Maxfield et al [25] considered race or ethnicity and gender but not the age of the applicants. They demonstrated that reviewers preferred Black and Hispanic applicants relative to White or Asian applicants. There were no significant interactions between race or gender and attractiveness. Although a different methodology was utilized, our data demonstrate a relatively small difference in attractiveness between applicants invited and not invited to interview, while in the study by Maxfield et al [25], physical appearance was the second strongest predictor of being invited to interview after the USMLE step 1 score. This could be related to the fact that Maxfield et al [25] utilized stock images of mock applicants while we used native applications, or it could be attributed to the AI's tendency to regress to the mean. Of note, another study by Steiner et al [32] demonstrated opposite findings in terms of race-related outcomes, finding that White program directors were more likely to match White applicants.

We believe that our data extend this prior work. While our research shows an apparent initial association between physical appearance and the chances of being invited to interview with a residency program, it demonstrates that this association disappears when adjusted for demographic and academic characteristics. This result indicates that attractiveness is likely a confounder linked with other factors such as demographics and gender relating to interview selection rather than a main variable independently contributing to interview selection.

The findings that MD candidates and candidates with more research and volunteer experiences are assigned higher attractiveness scores also suggest that, while physical appearance may not be a major factor in residency interview selection, it may provide an advantage through applicants' career development. Interestingly, USMLE scores either did not correlate or negatively correlated with attractiveness, which is thought-provoking as, unlike getting accepted to an MD-granting medical school and securing volunteer or research experiences, USMLE performance does not require interaction with another human via interview or collaboration. These results further point to a potential advantage that more attractive people have when it comes to academic achievements that rely on interpersonal communication rather than just medical knowledge alone. It is also aligned with other studies focused on the mostly positive impact of physical attractiveness on academic and career success [5,6,8,10-12,17,18].

The fact that we observed no difference in interview offers when controlled for demographic and academic characteristics suggests that the ERAS application process is unlikely to directly contribute to biased interview selection if a professional photograph is available to residency selection committees prior to deciding to invite an applicant to interview. Similarly, for residency programs outside of the United States, based on our data, use of professional photographs in the initial application is unlikely to give an unfair advantage to more attractive applicants. Put another way, our results do not support a causative link between photographic attractiveness and propensity to receive an interview invite. However, our data, in the context of the multidisciplinary work referenced previously, continue to raise questions as to the benefits of increased attractiveness compounded over a career, which may yield multifactorial benefits resulting in improved academic and career prospects.

To our knowledge, ours is the first work that combines ML methodology in combination with native applications in a real-life application cycle to explore this important topic. One study has explored algorithms incorporating traditional residency application metrics and applicant demographics, demonstrating high correlation with match success, although that study did not focus on the photograph [33]. Traditionally, studies focusing on the effect of physical appearance have utilized a group of individuals to grade attractiveness. We suggest that utilization of an AI algorithm may be a more objective way to achieve this goal. Since AI tools are often trained on hundreds or thousands of people, they may be less prone to individual biases. Nonetheless, because AI machines are trained to replicate real people's opinions, there have been multiple reports demonstrating biases in AI's grading of attractiveness, such as preferring Caucasian race and female sex [34,35]. This can be minimized, although not eliminated, by using culturally diverse AIs such as MEBeauty that we selected for this project. We reviewed multiple AI algorithms and training datasets and chose MEBeauty due to its diverse range of ethnicities, ages, genders, and unrestricted facial expressions. In addition, not only did the training dataset include thousands of pictures of diverse people, but hundreds of individuals who graded the images to train the AI were also specifically selected to represent various gender, ethnic, and age backgrounds [27].

Limitations

This is a single-center study with a limited number of residency program participants. While the number of ERAS applications is more than 2000, it constitutes only a small percentage of the total number of annual residency match participants. In addition, all programs were from only 2 geographic locations, which may

XSL•FO
RenderX

represent local practice and not be representative of other programs in the United States. Our sample objective data, such as gender, USMLE scores, number of volunteer experiences, and number of research experiences, are nearly identical to national averages. This suggests, but cannot guarantee, representativeness in utilizing our data to provide some prediction of the impact of physical attractiveness on interview invitation within the larger applicant pool for the 2022 residency match. In addition, participating programs self-reported on blinding to ERAS photographs and the role of physical attractiveness in their interview selection process. These reported data may not represent actual practice; however, all programs were provided with a general description of the project without specific details on the main research hypothesis and were reported in aggregate to encourage accurate reporting. We intentionally did not report program-specific data to preserve program and applicant confidentiality.

Unfortunately, some of the original raw data are no longer available. This includes code that was used to train and validate the model, the specific loss function used, correlation metrics between predicted and ground-truth scores on the validation set, mean absolute error in interpretable units, the specific version of MEBeauty/PyTorch code used, the random seed for reproducibility, the preprocessing parameters applied to ERAS photographs, and whether the pretrained VGG-16 weights were ImageNet-initialized or trained from scratch.

We used professional photograph image files as the main metrics for quality of the photograph. While ERAS photographs are highly standardized in terms of applicant position and attire, there are some differences in facial expression, eyewear use, background color, and hairstyle that we were not able to assess given the large amount of data. No validation was performed on a subset of ERAS photos to confirm that the model generalizes appropriately to this new domain.

We utilized race-sensitive AI to reduce bias, as program committee members who responded to the survey were representatives of a variety of racial backgrounds. However, we did not specifically tailor the AI to assess applicants based on a specific program director's race. This was due to the fact that the selection process often involves multiple committee members, such as program directors, associate program directors, and program coordinators, and there is no reliable method to distinguish which applicant was invited according to the input of any individual faculty member. Finally, while we used race-sensitive AI, ML models are notorious for being biased, and it is possible that the algorithm we utilized introduced its own bias to this study.

Conclusion

In our study of a diverse sample of applicants to multiple programs within a large graduate medical education institution, there was no association between a higher AI-generated physical attractiveness score of an applicant and the likelihood of receiving an invitation to interview when controlled for other variables. This suggests that more attractive applicants do not have an advantage in the residency interview selection process simply due to their ERAS photograph. There could, however, be an advantage of being overall more successful academically for more attractive applicants, which indirectly leads to better chances of being interviewed for a residency position, because based on the results of this study, more attractive candidates were more likely to be US MDs and have a higher number of research and volunteer experiences, which are traditionally correlated with higher chances of being invited to interview as per NRMP data [2]. Future areas of research should seek to further evaluate the correlation of attractiveness and residency match while attempting to evaluate potential benefits of attractiveness compounded over a career and include multicenter, geographically diverse settings.

Acknowledgments

No generative artificial intelligence was used in preparation of this manuscript.

Funding

The authors declared that no financial support was received for this work.

Data Availability

The datasets generated or analyzed during this study are not publicly available due to concerns for applicant and program confidentiality and Electronic Residency Application Service (ERAS) policy but are available from the corresponding author on reasonable request if ERAS permission is obtained.

Conflicts of Interest

None declared.

References

1. McVitie DG, Wilson LB. The stable marriage problem. Commun ACM 1971 Jul;14(7):486-490. [doi: 10.1145/362619.362631]
2. Results and data: 2023 main Residency Match®. : National Resident Matching Program; 2023 URL: https://www.nrmp.org/wp-content/uploads/2023/05/2023-Main-Match-Results-and-Data-Book-FINAL.pdf [accessed 2026-05-27]

XSL•FO
RenderX

3. Datta Gupta N, Etcoff NL, Jaeger MM. Beauty in mind: the effects of physical attractiveness on psychological well-being and distress. J Happiness Stud 2016 Jun;17(3):1313-1325. [doi: 10.1007/s10902-015-9644-6]
4. Snyder M, Tanke ED, Berscheid E. Social perception and interpersonal behavior: on the self-fulfilling nature of social stereotypes. J Pers Soc Psychol 1977;35(9):656-666. [doi: 10.1037/0022-3514.35.9.656]
5. Scholz JK, Sicinski K. Facial attractiveness and lifetime earnings: evidence from a cohort study. Rev Econ Stat 2015 Mar;97(1):14-28. [doi: 10.1162/REST_a_00435] [Medline: 30505018]
6. Parrett M. Beauty and the feast: examining the effect of beauty on earnings using restaurant tipping data. J Econ Psychol 2015 Aug;49:34-46. [doi: 10.1016/j.joep.2015.04.002]
7. Efran MG. The effect of physical appearance on the judgment of guilt, interpersonal attraction, and severity of recommended punishment in a simulated jury task. J Res Pers 1974 Jun;8(1):45-54. [doi: 10.1016/0092-6566(74)90044-0]
8. Hansen K. The relationship between teacher perceptions of pupil attractiveness and academic ability. Br Educ Res J 2016 Jun;42(3):376-398. [doi: 10.1002/berj.3227]
9. Westfall R, Millar M, Walsh M. Effects of instructor attractiveness on learning. J Gen Psychol 2016;143(3):161-171. [doi: 10.1080/00221309.2016.1200529] [Medline: 27410051]
10. Riniolo TC, Johnson KC, Sherman TR, Misso JA. Hot or not: do professors perceived as physically attractive receive higher student evaluations? J Gen Psychol 2006 Jan;133(1):19-35. [doi: 10.3200/GENP.133.1.19-35] [Medline: 16475667]
11. Watkins LM, Johnston L. Screening job applicants: the impact of physical attractiveness and application quality. Int J Sel Assess 2000 Jun;8(2):76-84. [doi: 10.1111/1468-2389.00135]
12. Agthe M, Spörrle M, Maner JK. Does being attractive always help? Positive and negative effects of attractiveness on social decision making. Pers Soc Psychol Bull 2011 Aug;37(8):1042-1054. [doi: 10.1177/0146167211410355] [Medline: 21636731]
13. McElroy JC, DeCarlo TE. Physical attractiveness on cognitive evaluations of saleswomen's performance. J Mark Theory Pract 1999 Jan;7(1):84-100. [doi: 10.1080/10696679.1999.11501823]
14. Wu PPC, Hwang IS. The influence of gender dyads and physical appearance on the strength of the customer-provider relationship in the Taiwanese hairdressing industry. Serv Mark Q 2012 Apr;33(2):138-154. [doi: 10.1080/15332969.2012.662459]
15. Jin N, Merkebu J. The role of employee attractiveness and positive emotion in upscale restaurants. Anatolia 2015 Apr 3;26(2):284-297. [doi: 10.1080/13032917.2014.948895]
16. Dipboye RL, Fromkin HL, Wiback K. Relative importance of applicant sex, attractiveness, and scholastic standing in evaluation of job applicant resumes. J Appl Psychol 1975;60(1):39-43. [doi: 10.1037/h0076352]
17. Johnson SK, Podratz KE, Dipboye RL, Gibbons E. Physical attractiveness biases in ratings of employment suitability: tracking down the “beauty is beastly” effect. J Soc Psychol 2010;150(3):301-318. [doi: 10.1080/00224540903365414] [Medline: 20575336]
18. Chia RC, Allred LJ, Grossnickle WF, Lee GW. Effects of attractiveness and gender on the perception of achievement-related variables. J Soc Psychol 1998 Aug;138(4):471-477. [doi: 10.1080/00224549809600401] [Medline: 9664863]
19. Dion K, Berscheid E, Walster E. What is beautiful is good. J Pers Soc Psychol 1972 Dec;24(3):285-290. [doi: 10.1037/h0033731] [Medline: 4655540]
20. Langlois JH, Kalakanis L, Rubenstein AJ, Larson A, Hallam M, Smoot M. Maxims or myths of beauty? A meta-analytic and theoretical review. Psychol Bull 2000 May;126(3):390-423. [doi: 10.1037/0033-2909.126.3.390] [Medline: 10825783]
21. Zebrowitz LA, Franklin RG Jr. The attractiveness halo effect and the babyface stereotype in older and younger adults: similarities, own-age accentuation, and older adult positivity effects. Exp Aging Res 2014;40(3):375-393. [doi: 10.1080/0361073X.2014.897151] [Medline: 24785596]
22. Dion KK. Cultural perspectives on facial attractiveness. In: Rhodes G, Zebrowitz LA, editors. Facial Attractiveness: Evolutionary, Cognitive, and Social Perspectives: Ablex Publishing; 2002:239-259 URL: https://scholarworks.brandeis.edu/esploro/outputs/9924057161601921 [accessed 2026-05-28]
23. Cash TF, Gillen B, Burns DS. Sexism and beautyism in personnel consultant decision making. J Appl Psychol 1977;62(3):301-310. [doi: 10.1037/0021-9010.62.3.301]
24. Cash TF, Kilcullen RN. The eye of the beholder: susceptibility to sexism and beautyism in the evaluation of managerial applicants. J Appl Soc Psychol 1985 Jun;15(4):591-605. [doi: 10.1111/j.1559-1816.1985.tb00903.x]
25. Maxfield CM, Thorpe MP, Desser TS, et al. Bias in radiology resident selection: do we discriminate against the obese and unattractive? Acad Med 2019 Nov;94(11):1774-1780. [doi: 10.1097/ACM.0000000000002813] [Medline: 31149924]
26. Stevens LM, Mortazavi BJ, Deo RC, Curtis L, Kao DP. Recommendations for reporting machine learning analyses in clinical research. Circ Cardiovasc Qual Outcomes 2020 Oct;13(10):e006556. [doi: 10.1161/CIRCOUTCOMES.120.006556] [Medline: 33079589]
27. Lebedeva I, Guo Y, Ying F. MEBeauty: a multi-ethnic facial beauty dataset in-the-wild. Neural Comput Applic 2022 Sep;34(17):14169-14183. [doi: 10.1007/s00521-021-06535-0]
28. USMLE score interpretation guidelines. : United States Medical Licensing Examination (USMLE); 2022 URL: https://www.usmle.org/sites/default/files/2022-05/USMLE%20Step%20Examination%20Score%20Interpretation%20Guidelines_5_24_22_0.pdf [accessed 2026-05-27]

XSL•FO
RenderX

29. Charting outcomes in the match: senior students of U.S. MD medical schools: characteristics of U.S. MD seniors who matched to their preferred specialty in the 2022 main Residency Match. : National Resident Matching Program; 2022 URL: https://www.nrmp.org/wp-content/uploads/2026/03/Charting-Outcomes-MD-Seniors-2022_Final_revised-citation.pdf [accessed 2026-05-27]
30. Charting outcomes in the match: senior students of U.S. DO medical schools characteristics of senior students of U.S. DO medical schools who matched to their preferred specialty in the 2022 main Residency Match. : National Resident Matching Program; 2022 URL: https://www.nrmp.org/wp-content/uploads/2022/07/Charting-Outcomes-DO-Seniors-2022_Final.pdf [accessed 2026-05-27]
31. Charting outcomes in the match: International Medical Graduates: characteristics of International Medical Graduates who matched to their preferred specialty in the 2022 main Residency Match. : National Resident Matching Program; 2022 URL: https://www.nrmp.org/wp-content/uploads/2026/03/Charting-Outcomes-IMG-2022_Final_revised-citation.pdf [accessed 2026-05-27]
32. Steiner Q, Edalatpour A, Seitz AJ, Bentz ML, Afifi AM. Diversity in the plastic surgery match: the effect of program chair, program director, and faculties' race and sex on matched applicants. J Craniofac Surg 2023;34(5):1376-1380. [doi: 10.1097/SCS.0000000000009440] [Medline: 37259193]
33. Shaffrey EC, Moura SP, Wirth PJ, et al. Objective residency applicant assessment using a linear rank model. J Surg Educ 2023 Jun;80(6):776-785. [doi: 10.1016/j.jsurg.2023.03.002] [Medline: 37012141]
34. Makhortykh M, Urman A, Ulloa R. Detecting race and gender bias in visual representation of AI on web search engines. In: Advances in Bias and Fairness in Information Retrieval: Springer; 2021:36-50. [doi: 10.1007/978-3-030-78818-6_5]
35. Fuchs DJ. The dangers of human-like bias in machine-learning algorithms. Missouri S&T'S Peer to Peer 2018;2(1) [FREE Full text]

Abbreviations

AI: artificial intelligence
CK: clinical knowledge
DO: Doctor of Osteopathic Medicine
ERAS: Electronic Residency Application Service
IRB: institutional review board
MD: Doctor of Medicine
ML: machine learning
NRMP: National Residency Matching Program
OR: odds ratio
USMLE: United States Medical Licensing Examination

Edited by A Stone; submitted 21.Jul.2025; peer-reviewed by A Edalatpour, Y Yang; revised version received 21.Apr.2026; accepted 27.Apr.2026; published 04.Jun.2026.

Please cite as:
Hunter D, Hunter L, Kerner D, Mullan A, Vanmeter D, Khapov I, Finch A, Hevesi S, Porter II I, West C, Homme J
Physical Attractiveness and Chances of Being Invited to Interview With a Medical Residency Program: Retrospective Cohort Study
JMIR Med Educ 2026;12:e81052
URL: https://mededu.jmir.org/2026/1/e81052
doi:10.2196/81052

© Daria Hunter, Luke Hunter, David Kerner, Aidan Mullan, Derek Vanmeter, Ivan Khapov, Alexander Finch, Sara Hevesi, Ivan Porter II, Colin West, James Homme. Originally published in JMIR Medical Education (https://mededu.jmir.org), 4.Jun.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Education, is properly cited. The complete bibliographic information, a link to the original publication on https://mededu.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Gender and Racial Discrimination During Residency Training: Scoping Review

Ida John[1], BSc; Liz Dennett[2], MLIS; Julie Nguyen[1], MA; Marghalara Rashid[1], PhD

[1]Department of Pediatrics, Faculty of Medicine and Dentistry, University of Alberta, 3-591 Dianne and Irving Kipnes Health Research Academy, 11405 87 Avenue NW, Edmonton, AB, Canada

[2]Geoffrey and Robyn Sperber Health Sciences Library, University of Alberta, Edmonton, AB, Canada

Corresponding Author:
Marghalara Rashid, PhD
Department of Pediatrics, Faculty of Medicine and Dentistry, University of Alberta, 3-591 Dianne and Irving Kipnes Health Research Academy, 11405 87 Avenue NW, Edmonton, AB, Canada

Abstract

Background: Women and visible minorities (VMs) continue to face discrimination while working in health care. These instances of discrimination can range from those perpetrated by individuals, such as misidentification, to prejudices held by institutions, such as a lack of commitment to hiring VMs. Furthermore, many residents face unique experiences of discrimination due to the intersection of gender and race. Although numerous studies have been published on the experiences of physicians with discrimination, there is a limited literature specifically on the experiences of residents.

Objective: This study aimed to explore and analyze the impact of gender and racial discrimination on medical residents during their residency training.

Methods: This scoping review was conducted in accordance with the PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses Extension for Scoping Reviews) guidelines. A health sciences librarian searched 5 databases: Ovid Medline(R) ALL, Embase (via Ovid), APA PsycInfo (via Ovid), CINAHL Plus with Full Text (via EBSCOhost), and Scopus. No study design, country, or date restrictions were applied. Studies that explored the impact of both gender and racial discrimination on residents during their residency training were included. Two reviewers conducted title and abstract screening, followed by full-text screening, and any discrepancies were resolved through group consensus. This study's content was extracted using Microsoft Word to create tables for organizing and managing the information from the included studies.

Results: After removing duplicates, the literature search revealed 2435 papers for title and abstract screening. A total of 340 papers were selected for full-text review, and ultimately, 26 papers met the inclusion criteria. Included papers were published between 1995 and 2024 in the United States (n=16), Canada (n=4), Australia and New Zealand (n=2), Saudi Arabia (n=1), Mexico (n=1), South Africa (n=1), and the United Kingdom (n=1). Our analysis revealed four themes: (1) forms of discrimination faced by VMs, (2) sources of discrimination, (3) ramifications of discrimination, and (4) ways for individuals and institutions to reduce gender and racial discrimination.

Conclusions: This scoping review identified that discrimination primarily manifests as microaggressions against residents and revealed the negative impact it continues to have on their careers. Raising awareness about these issues can help programs and institutions develop tailored solutions to tackle these problems and provide a safe, inclusive training environment for all residents.

(JMIR Med Educ 2026;12:e87524) doi:10.2196/87524

KEYWORDS

discrimination; gender; racial; residents; scoping review

Introduction

Background

Gender and racial discrimination continue to be prevalent across society, especially within occupational settings, where most people report experiencing discrimination [1]. Discrimination refers to an "action, behavior, decision, or omission that treats a person or a group of people unfairly" due to traits such as race or gender [2]. These biases can be explicit or implicit but are more often implicit, as reported in the current literature [3,4]. Additionally, these instances of discrimination can range from occurring within individual interactions to institutional policies that marginalize groups of people, such as women and visible minorities (VMs) [5]. VMs refer to "persons, other than Aboriginal peoples, who are non-White in color" [6]. Within health care, many studies surveying physicians have found that women physicians and physicians of VM status report higher rates of patients refusing their care due to their race and/or gender [7,8]. Both women and VMs are also more likely to be

XSL•FO
RenderX

mistaken for nurses [8] or as maintenance and housekeeping staff [9]. In addition to these forms of discrimination, some individuals face prejudice stemming from the intersection of their gender and race, known as intersectionality [10]. Intersectionality is defined as "the interaction between gender, race, and other categories of difference in individual lives, social practices, institutional arrangements, and cultural ideologies, and the outcomes of these interactions in terms of power" [11]. An example of what intersectional discrimination could look like within health care can be seen in a study conducted by Pololi and Jones [12], where they interviewed women who were medical faculty regarding their experiences. Women of VM status reported how their double minority status impacted them in the workplace, especially how isolating it felt to be the only woman and/or person of VM status on the faculty [12].

Medical residents hold a unique position within health care, as they are simultaneously trainees and practitioners interacting with and caring for patients under supervision. Recent research conducted by Mocanu et al [13] surveyed general surgery residents on the effects that their gender and/or VM status had on their resilience and their residency training experience. Some experiences of discrimination reported were similar to those reported by physicians, such as how the majority of women residents and residents of VM status faced dismissal of their medical advice by patients due to their gender or race [13]. Residents also believed that they received fewer training opportunities due to their gender and/or race [13]. Another study conducted by Bhatt [14] found that women or men of VM status were rarely selected as chief residents; often, they were White men, even though residents are a diverse group. Residents also described being doubly discriminated against due to their gender and race [14].

Although individual research studies have been conducted on the experiences of residents facing discrimination, to the best of our knowledge, this is the first paper that compiles resident experiences of discrimination into 1 paper. This scoping review aims to explore and analyze existing literature on the impact of gender and racial discrimination on residents.

Positionality Statement

We acknowledge that our perspectives and identities may have influenced our scoping review. However, through reflexivity and debriefings as a team, we have examined potential preconceived assumptions and strived for an accurate and rigorous representation of the existing literature. The team includes diverse members. IJ is a woman from a VM group and a graduate research assistant on the project. LD is a White female librarian who conducted a comprehensive search for this scoping review. JN is from a VM background and is a research assistant. MR is a VM woman and a scholar in medical education whose work focuses on the experiences of structurally marginalized learners.

Methods

Overview

We used the Arksey and O'Malley [15] framework, which consists of 5 steps: identifying the research question, finding relevant reports, selecting reports, extracting data, and summarizing the results. Furthermore, for our data reporting, the PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses Extension for Scoping Reviews) guidelines were used to ensure transparent and clear reporting of the included studies in our scoping review [16]. Additionally, the protocol for this scoping review was developed in accordance with these guidelines [16]. The protocol has not been registered.

Study Eligibility Criteria

Included studies focused on medical residents who faced both gender and racial discrimination during their residency training. We included peer-reviewed studies, books, book chapters, peer-reviewed abstracts, dissertations, and papers written in English. Papers that explored the experiences of other populations, such as undergraduate medical students, fellows, clinical and academic faculty, health care professionals, or allied health care professionals, were excluded. Papers that did not align with our study objectives were excluded. Studies that focused solely on collecting data on gender or racial discrimination were not included, as the purpose of this scoping review was to explore the experiences of residents who faced discrimination based on both gender and race. Literature surrounding the impact of discrimination on resident assessments and evaluations was excluded, as there are abundant studies and reviews regarding this specific topic. We also removed studies focused on sexual orientation, as it is not the focus of this review and is a large topic in its own right. Studies that were not conducted in a residency setting were excluded. We also excluded the following: policy documents, vlogs, websites, working papers, newsletters, government publications, videos, study protocols, and papers written in other languages. There were no limitations imposed on the study design, the time frame, or the countries where the research was conducted.

Search Strategy

A health sciences librarian (LD) and other research team members collaborated to develop an initial search strategy to identify studies on the experiences of medical residents with intersectional gender and racial discrimination. The health sciences librarian then ran the search in the following databases: Ovid Medline(R) ALL, Embase (via Ovid), APA PsycInfo (via Ovid), CINAHL Plus with Full Text (via EBSCOhost), and Scopus from inception until May 2, 2025. The strategy combined keyword and subject heading search terms for four concepts: (1) medical residents; (2) gender or gender identity; (3) race, ethnicity, or foreign; and (4) bias or discrimination. The concepts were combined with Boolean AND, so all 4 concepts had to be present for a paper to be retrieved. Conference abstracts were removed. The full search strategy is provided in Multimedia Appendix 1. All search results were loaded into Covidence (Veritas Health Information) for deduplication and screening. Reference lists from all included studies were reviewed to identify additional relevant studies.

Study Screening

Reports were screened through a multiphase process. Two researchers reviewed the first hundred papers, and the kappa

XSL•FO
RenderX

coefficient was calculated to assess their agreement. We obtained a kappa of 0.72, indicating substantial agreement. Subsequently, all papers were screened based on title and abstract by 2 independent reviewers (IJ and JN), who either included or excluded each paper. A third reviewer (MR) resolved any discrepancies. During the full-text review phase, the 2 reviewers (IJ and JN) made the final decision to include or exclude each paper from the scoping review. Any discrepancies were resolved through group consensus.

Data Extraction

Study content was extracted using Microsoft Word to create tables for organizing and managing the information from the included studies. We extracted authors' details, the country where the studies were conducted, publication dates, sample sizes, research methods or study designs, and disciplines (Table 1). Additionally, we extracted study aims or objectives and findings. Two reviewers (IJ and JN) extracted data from all included studies, and a third reviewer (MR) double-checked the content and entries for accuracy, ensuring no data was overlooked. Any inconsistencies were resolved through group consensus. A quality appraisal of these reports was not performed, as it is not a requirement for scoping reviews [16].

XSL•FO
RenderX

Table . Characteristics of studies included in the literature review.

Studies	Country	Title	Sample (n)	Level of training	Disciplines	Study design
Zaeem et al [17]	Canada	Workplace discrimination and harassment among Alberta postgraduate medical trainees: a cross-sectional survey	195	PGY1-PGY4[a]	Anesthesiology, emergency medicine, family medicine, public health, internal medicine, dermatology, neurology, pediatrics, psychiatry, radiology, laboratory medicine specialties, and surgical specialties	Mixed methods
Kim et al [18]	United States	Evaluating the impact of gender, race, and training year on internal medicine residents' experiences across the United States	176	PGY1-PGY4+	Internal medicine	Quantitative
Lodha et al [19]	United States	Evaluating the impact of gender and race on otolaryngology resident experiences across the United States	61	PGY1-PGY5	Otolaryngology	Quantitative
Chiraroekmongkon [20]	United States	The invisibility cloak: narrative from an Asian American woman physician-in-training	N/A[b]	N/A	Psychiatry	Personal account
Koech et al [21]	United States	Minority resident physicians' perspectives on the role of race, ethnicity, culture, and gender in their surgical training experiences	23	PGY1-PGY4	Surgical specialties: orthopedic surgery, neurosurgery, ENT[c], general surgery, plastic surgery, and urology	Qualitative
Hussain et al [22]	United Kingdom	The impact of race and gender-related discrimination on the psychological distress experienced by junior doctors in the UK: a qualitative secondary data analysis	14	PGY1-PGY5	General pediatrics, obstetrics and gynecology, medicine, and emergency	Qualitative
Tyler [23]	United States	CORR[d] insights: the majority of Black orthopedic surgeons report experiencing racial microaggressions during their residency training	N/A	N/A	Orthopedic surgery	Commentary

XSL•FO
RenderX

Studies	Country	Title	Sample (n)	Level of training	Disciplines	Study design
Brooks et al [24]	United States	The majority of Black orthopedic surgeons report experiencing racial microaggressions during their residency training	310	PGY1-PGY4	Orthopedic surgery	Quantitative
Patel et al [25]	United States	Diversity, equity, and inclusion among anesthesiology trainees	135	CA[e] 0-CA 3	Anesthesiology	Quantitative
Sandoval-Bonilla et al [26]	Mexico	Discrimination of residents during neurosurgical training in Mexico: results of a survey before SARS-CoV-2	135	PGY1-PGY5	Neurosurgery	Quantitative
Lall et al [27]	United States	Prevalence of discrimination, abuse, and harassment in emergency medicine residency training in the US	7680	PGY1-PGY4	Emergency medicine	Quantitative
Villanueva et al [28]	Australia and New Zealand	"The odds were stacked against me": a qualitative study of underrepresented minorities in surgical training	8	PGY1-PGY5	General surgery, otolaryngology head and neck surgery, vascular surgery, plastic and reconstructive surgery	Qualitative
Zhuo et al [29]	United States	Facilitators and barriers to allyship in academic surgery: a qualitative study	15	PGY1-PGY4	Surgery	Qualitative
Mocanu et al [13]	Canada	Intersectionality of gender and visible minority status among general surgery residents in Canada	210	PGY1-PGY5	General surgery	Quantitative
Mgbako [30]	United States	Bonds in every color	N/A	PGY1-PGY4	Internal medicine	Personal account
Yuce et al [31]	United States	National evaluation of ethnic or racial discrimination in US surgical residency programs	5679	PGY1-PGY4	Surgery	Quantitative
Pearce et al [32]	Australia and New Zealand	Gender effects in anesthesia training in Australia and New Zealand	356	PGY1- provisional fellowship	Anesthesiology	Quantitative
Hu et al [33]	United States	Discrimination, abuse, harassment, and burnout in surgical residency training	7409	PGY1-PGY5	Surgery	Quantitative

XSL•FO
RenderX

Studies	Country	Title	Sample (n)	Level of training	Disciplines	Study design
Osseo-Asare et al [34]	United States	Minority resident physicians' views on the role of race or ethnicity in their training experiences in the workplace	27	PGY1-PGY4	Anesthesiology, emergency medicine, family medicine, internal medicine, neurology, obstetrics and gynecology, pediatrics, psychiatry, radiology, surgery, and urology	Qualitative
Fitzgerald et al [35]	United States	Screening for harassment, abuse, and discrimination among surgery residents: an EAST[f] multicenter trial	76	PGY1-PGY5	Surgery	Quantitative
Bhatt [14]	United States	The little brown woman: gender discrimination in American medicine	108	PGY1-PGY4	Dermatology, family medicine, internal medicine, neurology, pathology, pediatrics, physical medicine and rehabilitation, psychiatry, radiology, and surgical specialties	Qualitative
Fnais et al [36]	Saudi Arabia	Prevalence of harassment and discrimination among residents in three training hospitals in Saudi Arabia	213	PGY1-PGY5	All specialties	Quantitative
Crutcher et al [37]	Canada	Family medicine graduates' perceptions of intimidation, harassment, and discrimination during residency training	242	PGY1- PGY2	Family medicine	Mixed methods
London et al [38]	South Africa	A survey of trainee specialists' experiences at the University of Cape Town (UCT): impacts of race and gender	91	PGY1-PGY5	Medicine, pediatrics, and psychiatry	Quantitative
vanInevald et al [39]	Canada	Discrimination and abuse in internal medicine residency	543	PGY1-PGY4	Internal medicine	Quantitative
McNamara et al [40]	United States	The extent and effects of abuse and harassment of emergency medicine residents.	1774	PGY1-PGY5	Emergency medicine	Mixed methods

[a]PGY: postgraduate year.

[b]N/A: not applicable.

[c]ENT: ear, nose, and throat.

[d]CORR: Clinical Orthopaedics and Related Research.

[e]CA: clinical anesthesiology.

[f]EAST: Eastern Association for the Surgery of Trauma.

XSL•FO
RenderX

Synthesis of Results

Data were analyzed using conventional content analysis, a suitable approach given the lack of predefined categories. Content analysis allowed us to identify the main categories, relationships between them, and gaps in the literature. One team member (IJ) created data abstraction tables before data abstraction began. These abstraction forms included citation details, such as author, year of publication, title, and country of study, as well as study characteristics, such as study design, research objective, and sample size. Two team members (IJ and JN) coded the data and organized it into 4 themes. Any discrepancies at this stage were resolved through consensus.

Ethical Considerations

This is a scoping review that synthesizes published papers; therefore, no human ethics approval was required.

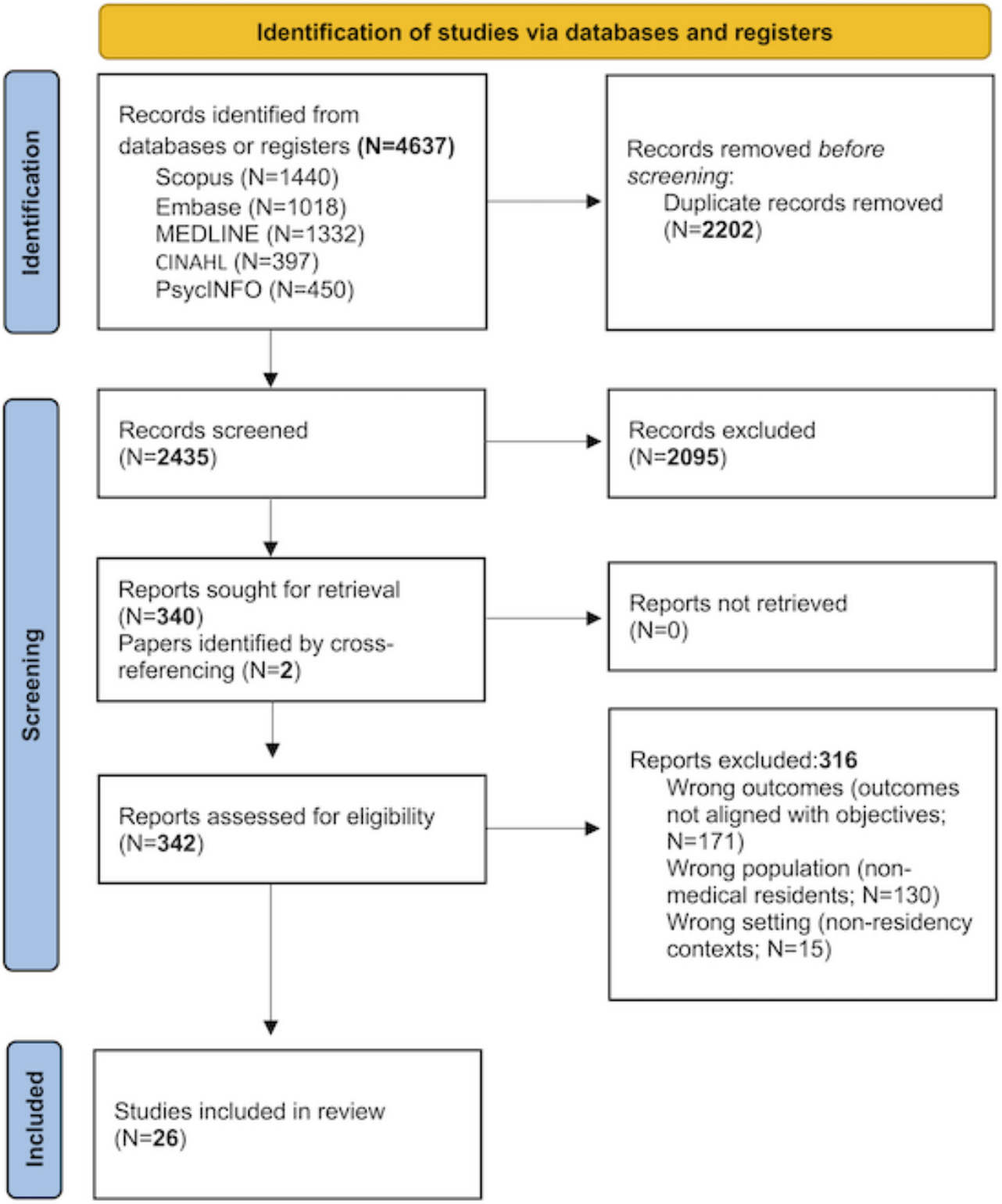
Results

Description of Included Studies

The initial search identified 4637 potential studies, of which 2435 remained for title and abstract screening after duplicates were removed. A total of 340 papers were selected for full-text review, and ultimately, 24 papers met the inclusion criteria. Reviewing the reference lists of the included papers uncovered 2 additional relevant reports, bringing the total to 26 studies included in the scoping review (Figure 1).

XSL•FO
RenderX

Figure 1. PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses Extension for Scoping Reviews) flow diagram for the papers included in this study.



An overview of the characteristics of the included studies is provided in Table 1. These papers were published between 1995 and 2024. Most studies (n=16) were conducted in the United States, followed by Canada (n=4), Australia and New Zealand (n=2), and 1 study each from Saudi Arabia, Mexico, South Africa, and the United Kingdom. Fourteen papers used a quantitative study design, followed by 6 qualitative studies, 3 mixed-methods studies, 2 personal accounts, and 1 commentary. Of the 23 research studies, sample sizes ranged from 8 to 7409. Residents came from a variety of specialties, including mixed specialties (n=8), surgery (n=5), internal medicine (n=3), anesthesiology (n=2), emergency medicine (n=2), orthopedic surgery (n=2), otolaryngology (n=1), family medicine (n=1), neurosurgery (n=1), and psychiatry (n=1).

Our data analysis revealed four themes: (1) forms of discrimination faced by VMs, (2) sources of discrimination, (3)

XSL•FO
RenderX

ramifications of discrimination, and (4) ways to reduce gender and racial discrimination (Table 2).

XSL•FO
RenderX

Table . Key findings of included studies.

Studies	Objectives	Key findings	Themes reported
Zaeem et al [17]	The goal of this study was to characterize the demographics and experiences of discrimination faced by residents to improve their learning environment.	More women than men have frequently experienced a form of discrimination, harassment, or mistreatment. No differences in the frequency of discrimination experienced were found between residents who are Black, Indigenous, or People of Color (BIPOC) and White residents; however, BIPOC residents reported hearing more racial slurs, either about them or in general. They also provided examples of how attendings often negatively commented on their work, leading to differential treatment of women and BIPOC residents compared to other groups.	Theme 1: the forms of discrimination faced by VM and Theme 2: sources of discrimination
Kim et al [18]	This study aimed to characterize how resident well-being and discrimination experienced by residents in internal medicine were impacted by their race, gender, and training year.	More women than men reported experiencing differential treatment from faculty and receiving fewer opportunities due to their gender. Residents of ethnic backgrounds described being treated differently by faculty, with mostly Black residents reporting facing discrimination and receiving fewer opportunities due to their race. The majority of Black residents and female residents reported being misidentified for another role, which contributed to feelings of burnout reported by female residents. Finally, the authors suggested that residency programs provide explicit and implicit bias training to residents and staff to reduce discrimination.	Theme 1: the forms of discrimination faced by VM, Theme 3: ramifications of discrimination, and Theme 4: ways for individuals and institutions to reduce discrimination
Lodha et al [19]	The purpose of this study was to determine if otolaryngology residents' experiences with discrimination and burnout varied by gender and race.	Many women and underrepresented minorities (URMs) reported experiencing misidentification as a nonphysician and perceived receiving fewer opportunities in research and training compared to men. Twenty-seven percent of women described experiencing burnout, and 40.5% of women felt less confident in their ability to independently practice. The authors recommended raising awareness among residents regarding these experiences and training them on how they can support others who may face discrimination.	Theme 1: the forms of discrimination faced by VM, Theme 3: ramifications of discrimination, and Theme 4: ways for individuals and institutions to reduce discrimination

XSL•FO
RenderX

Studies	Objectives	Key findings	Themes reported
Chiraroekmongkon [20]	This narrative captures the experiences of an Asian American woman with discrimination during residency.	The author described facing gender and racial discrimination from patients, including verbal abuse and other inappropriate comments. She also faced discrimination from colleagues through her ideas being credited to someone else and being confused with other Asian colleagues. Discrimination also exists institutionally, as Asian Americans are not considered underrepresented in medicine since they make up a representative portion in health care. However, this does not remove the discrimination they continue to face. Experiencing discrimination can result in internalization of these negative messages and silencing of one's identity.	Theme 1: the forms of discrimination faced by VM, Theme 2: sources of discrimination, and Theme 3: ramifications of discrimination
Koech et al [21]	The goal of this study was to determine how surgical resident training experiences were impacted by race, gender, and cultural background.	The majority of URMs described being treated differently, having difficulties fitting in, and experiencing microaggressions from various sources. They were also more likely to face greater criticism during training compared to their non-URM coresidents. Many residents combated these issues by attempting to assimilate, isolate, or seek assistance from outside sources. Participants suggested educational interventions for reducing discrimination. This included formal sessions, such as URM sharing their experiences, and informal sessions, where a senior resident or attending addresses an instance of discrimination as it occurs on the ward.	Theme 1: the forms of discrimination faced by VM, Theme 3: ramifications of discrimination, and Theme 4: ways for individuals and institutions to reduce discrimination
Hussain et al [22]	The objective of this qualitative study was to examine the experiences of racial and gender discrimination perceived by residents and its effects on their psychological distress (PD).	Many participants experienced different forms of discrimination, including microaggressions and explicit racism. Black, Asian, and minority ethnic (BAME) women experienced more discrimination than White women and BAME men. Patients were a common source of discrimination, with some requesting to see White physicians instead. Senior physicians and midwives were also described as perpetrators of discrimination. Participants reported feeling higher levels of PD, which was positively correlated with experiencing discrimination. Participants proposed improving the reporting system, providing educational sessions regarding discrimination, and making residents aware of the policy that allows them to refuse noncritical care to prejudiced patients.	Theme 1: the forms of discrimination faced by VM, Theme 2: sources of discrimination, Theme 3: ramifications of discrimination, and Theme 4: ways for individuals and institutions to reduce discrimination

XSL•FO
RenderX

Studies	Objectives	Key findings	Themes reported
Tyler [23]	The author of this commentary aimed to provide suggestions for improving the recruitment and retention of women and URMs into orthopedic residency.	One major suggestion described raising awareness among program directors and faculty by providing training on microaggressions and giving nonexclusionary feedback. Other suggestions included residents being provided with ways to report without fear of retribution, providing early support to Black women or those who face intersectional experiences, and acknowledging that these incidents occur.	Theme 1: the forms of discrimination faced by VM and Theme 4: ways for individuals and institutions to reduce discrimination
Brooks et al [24]	This study aimed to determine the degree to which Black residents experienced discrimination, the different types of microaggressions they faced, and whether there were differences in the discrimination faced by Black women and men.	The majority of participants reported experiencing either a lot of discrimination (34%) or some discrimination (44%) at their workplace from patients and faculty. Around 87% of participants described being misidentified as a nonphysician, and 81% of participants reported being misidentified for a nonmedical position. Fifty percent of participants also received devaluing feedback, with 87% reporting it to be based on race. Additionally, when comparing for gender, more Black women reported experiencing discrimination through microaggressions and receiving feedback than Black men. These forms of feedback often caused feelings of "othering." The authors listed some common forms of negative feedback to avoid and also recommended that leadership provide safe spaces for residents to report.	Theme 1: the forms of discrimination faced by VM, Theme 2: sources of discrimination, Theme 3: ramifications of discrimination, and Theme 4: ways for individuals and institutions to reduce discrimination
Patel et al [25]	This study aimed to determine residents' intentions in choosing anesthesiology, evaluate the extent of discrimination experienced by residents, and provide a foundation for implementing change in the future.	The majority of respondents identified mentors and role models as the most significant reason for choosing anesthesiology. More women than men also reported that women and diverse faculty were another important factor in their decision. More non-White residents reported facing racial discrimination during the residency application process and during residency training. Around 35.1% of women also faced gender discrimination during their training. Consequently, increasing the number of underrepresented minorities and women in leadership positions may assist in reducing discrimination.	Theme 4: ways for individuals and institutions to reduce discrimination

XSL•FO
RenderX

Studies	Objectives	Key findings	Themes reported
Sandoval-Bonilla et al [26]	The purpose of this study was to describe the discrimination experienced by neurosurgical residents.	Twenty-seven percent of participants experienced a type of discrimination, with the majority of experiences being due to place of origin, gender, or physical appearance. More women than men reported experiencing gender discrimination, and more foreign residents than Mexican residents reported facing discrimination due to place of origin. Participants who did not experience discrimination were more likely to report being satisfied with their education.	Theme 3: ramifications of discrimination
Lall et al [27]	This study aimed to evaluate the prevalence, forms, and causes of discrimination, abuse, and harassment experienced by residents in emergency medicine. They also sought to determine the association between these forms of mistreatment and suicidality.	Around 45.1% of participants reported experiencing 1 form of mistreatment (discrimination, abuse, or harassment) during the recent year, and many identified patients and patients' families as the main cause. More women than men described facing gender discrimination and sexual harassment, while residents of ethnic backgrounds described facing more racial discrimination than their White peers. Suicidality was equally prevalent among male and female residents, although adjusting for mistreatment significantly decreased suicidality among female residents. Educating residents and staff through cultural competency training may help with creating an inclusive environment for residents.	Theme 1: the forms of discrimination faced by VM, Theme 2: sources of discrimination, Theme 3: ramifications of discrimination, and Theme 4: ways for individuals and institutions to reduce discrimination
Villanueva et al [28]	The purpose of this study was to determine the barriers and examine the experiences of URM residents in choosing and completing the Surgical Education and Training (SET) program.	Women residents reported being discouraged from applying for surgery due to the possibility of raising a family in the future. International medical graduates (IMGs) reported experiencing racial discrimination, especially if they had an accent. Women also experienced sexual harassment and felt their opinions were undervalued by their male peers. As a result, many women reported feeling lonely or alienated. Recruiting more underrepresented minorities in leadership positions would provide additional role models for trainees.	Theme 1: the forms of discrimination faced by VM, Theme 3: ramifications of discrimination, and Theme 4: ways for individuals and institutions to reduce discrimination

XSL•FO
RenderX

Studies	Objectives	Key findings	Themes reported
Zhuo et al [29]	The goal of this qualitative study was to describe the covert types of biases that residents experienced and to explore the different approaches that residents took when addressing them.	Most participants reported experiencing biases based on their gender and race. These included patients requesting a physician of a different race and individuals in senior positions making subtle, derogatory comments about residents. Residents struggled with recognizing discrimination when it occurred, which affected when and how they responded. They were less likely to address bias if it was caused by someone in a position of power and if they believed that responding would take up a lot of energy. Participants mentioned that representation and administrative support influenced whether they responded to incidents or not.	Theme 1: the forms of discrimination faced by VM, Theme 2: sources of discrimination, Theme 3: ramifications of discrimination, and Theme 4: ways for individuals and institutions to reduce discrimination
Mocanu et al [13]	The objective of this study was to explore the effects of visible minority (VM) status and gender on the experiences of surgical residents.	Around 81.2% of women and 14.3% of those of VM status described being dismissed for their medical expertise due to their gender and race, respectively. Women of VM status and women in general are perceived to receive fewer opportunities due to their gender. Women of VM status were also more likely to report challenges with fitting in, forming good relationships with staff, and being appreciated. Institutions and individuals must be aware that these biases exist, and increasing the number of mentors and role models may assist with achieving a culture of equality.	Theme 1: the forms of discrimination faced by VM, Theme 3: ramifications of discrimination, and Theme 4: ways for individuals and institutions to reduce discrimination
Mgbako [30]	This commentary recounts a meeting where physicians and residents of color shared their experiences of discrimination.	Residents shared times when patients questioned their abilities due to their race, whether through comments on their accent or asking to be seen by another physician. During these incidents, residents reported suppressing their feelings and experiencing shame for not being able to defend themselves.	Theme 1: the forms of discrimination faced by VM and Theme 3: ramifications of discrimination

XSL•FO
RenderX

Studies	Objectives	Key findings	Themes reported
Yuce et al [31]	This study aimed to discover the prevalence, sources, and factors of racial discrimination and its impact on residents' well-being.	Non-White participants reported facing more discrimination due to race or ethnicity than White participants, with a larger portion (70.7%) of Black residents reporting racial discrimination than any other racial group. Most Black residents described being confused with other Black residents, experiencing role misidentification, being evaluated differently, and experiencing racial slurs. Patients were the most reported source of racist comments, while nurses and staff were most likely to confuse residents of the same race. Burnout and suicidal thoughts were more likely to be reported by residents who experienced discrimination. The authors recommended training residents and faculty to recognize and intervene during incidents and to encourage residents to report their experiences.	Theme 1: the forms of discrimination faced by VM, Theme 2: sources of discrimination, Theme 3: ramifications of discrimination, and Theme 4: ways for individuals and institutions to reduce discrimination
Pearce et al [32]	The objective of this survey was to evaluate the differences between male and female anesthesiology residents regarding the number of procedures performed, their confidence in their abilities, and whether they face gender and ethnic discrimination.	Male residents reported conducting procedures >10 times as frequently compared to female residents. They were also more likely to rate their abilities higher than their training level and to report greater competence with procedures to receive more opportunities. More female residents reported experiencing gender discrimination than male residents, which was more pronounced in female residents of VM status. Consequently, the authors suggested making faculty aware of these implicit biases and ensuring that women residents receive the same opportunities as male residents.	Theme 1: the forms of discrimination faced by VM and Theme 4: ways for individuals and institutions to reduce discrimination
Hu et al [33]	This survey aimed to identify the frequency and prevalence of experiences of mistreatment, burnout, suicidal thoughts, and their relationship with each other.	Women reported experiencing more gender discrimination and sexual harassment than male residents. Both gender and racial discrimination were reported to be caused mostly by patients. Around 38.5% of residents experienced burnout weekly, while 4.5% reported experiencing suicidal thoughts. Women who experienced mistreatment were more likely to experience burnout.	Theme 1: the forms of discrimination faced by VM, Theme 2: sources of discrimination, and Theme 3: ramifications of discrimination

XSL•FO
RenderX

Studies	Objectives	Key findings	Themes reported
Osseo-Asare et al [34]	The objective of this study was to identify and describe how race or ethnicity impacts resident experiences.	Residents reported experiencing microaggressions, including being asked questions regarding their ethnicity, being misidentified for another role by patients and staff, experiencing exoticization or being mistaken for another resident of color, and experiencing explicit racism. Other themes included being burdened with tasks related to diversity and struggling to balance professional and personal identities at work. These experiences could be improved upon by providing residents with forums to discuss these issues and by hiring more VMs to provide mentorship.	Theme 1: the forms of discrimination faced by VM, Theme 2: sources of discrimination, Theme 3: ramifications of discrimination, and Theme 4: ways for individuals and institutions to reduce discrimination
Fitzgerald et al [35]	This study sought to characterize the prevalence of harassment, abuse, and discrimination experienced by surgical residents using the HITS (Hurt, Insulted, Threatened with harm, or Screamed at) tool.	About 77.6% of participants experienced abuse and harassment, mostly receiving insults or being screamed at. The majority of residents reported experiencing sexual harassment, gender discrimination, and racial discrimination. Patients and supervising physicians were the most reported sources of discrimination. More women experienced gender discrimination, which has negatively impacted their environment and performance. Researchers also found that residents who reported gender discrimination were also more likely to report experiencing racial discrimination. They recommended training residents on how to report harassment and making them aware of accessible resources.	Theme 1: the forms of discrimination faced by VM, Theme 2: sources of discrimination, Theme 3: ramifications of discrimination, and Theme 4: ways for individuals and institutions to reduce discrimination
Bhatt [14]	The objective of this study was to explore the experiences of gender discrimination faced by residents of Indian origin and examine the interaction of gender with race in their careers.	Many participants described facing various forms of gender discrimination, starting at the residency application process, where women were asked about their plans for having children. Women were also more likely to be guided toward less labor-intensive specialties, such as pediatrics and family medicine. Race and gender together elevated their risk of facing discrimination, which was further exacerbated if residents were trained in a foreign country. Being a woman and of VM status also affected their chances of reaching leadership positions.	Theme 1: the forms of discrimination faced by VM and Theme 3: ramifications of discrimination

XSL•FO
RenderX

Studies	Objectives	Key findings	Themes reported
Fnais et al [36]	The purpose of this study was to determine how prevalent harassment and discrimination are during residency training.	Most participants experienced negative verbal comments during residency training, with consultants being the primary source. Rates of gender discrimination were reported equally by female and male residents; however, more female residents reported experiencing sexual harassment. Many participants reported wanting a different career. The authors recommended providing a confidential reporting system to allow residents to report their experiences.	Theme 1: the forms of discrimination faced by VM, Theme 2: sources of discrimination, Theme 3: ramifications of discrimination, and Theme 4: ways for individuals and institutions to reduce discrimination
Crutcher et al [37]	This study aimed to explore how intimidation, harassment, and discrimination (IHD) affected the experiences of family medicine residents.	About 44.7% of respondents experienced a form of IHD as a resident, with 94.3% experiencing negative verbal comments. The form of gender discrimination varied based on gender, with more female residents perceiving that they received fewer opportunities, while male residents perceived receiving more work as punishment. Additionally, IMGs were perceived to experience more discrimination based on their ethnicity and language than Canadian medical graduates. Observing resident well-being may be one way to measure experiences of IHD.	Theme 2: sources of discrimination, Theme 3: ramifications of discrimination, and Theme 4: ways for individuals and institutions to reduce discrimination
London et al [38]	The goal of this study was to identify the demographics of residents enrolled from 1999 to 2006 and to determine factors that impacted the recruitment and retention of Black female residents.	The number of VMs and women increased from 1999 to 2006; however, some participants reported knowing others who chose to go to other schools due to reports of racial discrimination. More Black residents than White residents reported feeling unwelcome. Fourteen participants experienced discrimination, but only half reported racial discrimination.	Theme 1: the forms of discrimination faced by VM and Theme 3: ramifications of discrimination
vanInevald et al [39]	This study aimed to explore the experiences of gender and racial discrimination and homophobia among residents in internal medicine.	Female residents revealed higher rates of gender discrimination, with patients (88%) and attendings (70%) being indicated as major sources. More female residents also reported experiencing sexual harassment, with 56% reporting patients and 35% reporting attendings as the main problem. Patients continued to be a major source for racial discrimination (67%), followed by other health care staff (52%), peers (50%), and attendings (49%). The authors suggested training residents on how to deal with patients causing discrimination.	Theme 1: the forms of discrimination faced by VM, Theme 2: sources of discrimination, and Theme 4: ways for individuals and institutions to reduce discrimination

XSL•FO
RenderX

Studies	Objectives	Key findings	Themes reported
McNamara et al [40]	The objective of this study was to assess the degree to which harassment and racial discrimination are experienced by residents in emergency medicine and how these experiences impact their well-being.	Ninety-eight percent of participants reported experiencing abuse or harassment, with women reporting higher rates of sexual harassment and differential treatment due to gender. Non-Whites reported experiencing racial discrimination more than White residents. Patients were the major cause of most discriminatory experiences. Encounters of discrimination left participants questioning their choice of becoming a physician and choosing emergency medicine. The researchers suggested providing educational sessions for residents, establishing reporting systems, and providing support and counseling.	Theme 1: the forms of discrimination faced by VM, Theme 2: sources of discrimination, Theme 3: ramifications of discrimination, and Theme 4: ways for individuals and institutions to reduce discrimination

Theme 1: Forms of Discrimination Faced by VMs

Twenty-three (88%) studies highlighted the different ways discrimination has manifested against VMs [13,14,17-24,27-36,38-40]. Eighteen papers described instances of microaggressions that were experienced by women and VM residents [13,14,17-22,24,28-32,34,35,38,40]. Participants in 11 studies reported experiencing misidentification of their role due to their gender and/or race [13,17-20,22,24,29,31,32,34]. Most residents who were assumed to be a nonphysician or a non–health care worker were women [13,17-20,22,29,32], of VM status [18,19,24,31,34], or both [18,24,32]. Women were more likely to be assumed to be a nurse or secretary [20,22,29], while VMs were more likely to be assumed to be custodial staff [24,29]. Moreover, residents experienced misidentification even after introducing themselves and wearing a white coat and stethoscope [34]. However, residents were aware that misidentification likely occurs due to stereotypes [29]. In addition to misidentification of role, VMs also reported being mistaken for another resident of minority status [20,31,34] and having patients ask to see a physician of another race [22,29,30].

VM residents in 10 papers also reported experiencing verbal microaggressions [17,20-22,24,29,31,34,35,40], including hearing racial slurs [17,31,40], receiving questions about their ethnicity [22,34], and astonished comments on their English proficiency [17,22], such as “Well, at least you speak good English!” [17]. The majority of these microaggressions originated from patients; however, attendings and other health care workers were a close second [21,24,29,31,34,40]. VMs were also discriminated against for their accents, often being treated with less respect as a result [14,28,30].

Another frequently reported form of microaggression across 5 studies was that VM residents were assessed using different standards than their non-VM coresidents [14,17,21,24,31]. This disparity manifested as harsher evaluations [17,21] and more critical feedback on performance [24].

Most gender-based microaggressions occurred against women. For example, women residents in 3 papers were more likely to report facing discrimination due to pregnancy and childcare than men residents [14,27,33]. These included women being asked about their family plans during residency interviews, being treated differently while pregnant during residency, and facing a lower chance of receiving a promotion [14]. Additionally, many faculties discouraged women from entering competitive disciplines, such as surgery, due to these specialties being considered labor-intensive and incompatible with family responsibilities [14,22,28]. Women were also more likely to face discrimination from other women, attending physicians, and nurses [14,22].

Furthermore, many studies recognized that microaggressions were compounded and worse for women of VM status [13,14,18,22,24,31,32,35]; however, only 3 studies investigated resident experiences with an intersectional lens [13,14,22]. One study found that VM women were less likely to feel that they had a good relationship with their staff [13]. Another study found that Black, Asian, and minority ethnic women often felt sidelined by men and found it overwhelming to respond to every instance of discrimination [22].

Another form of discrimination that was reported less frequently but consistently over the years is sexual harassment, which was primarily reported by women in 10 papers [20,22,27-29,33,35,36,39,40]. Inappropriate sexual comments [20,22,28,29,35,36] and unwanted sexual jokes [22,35,40] were the most commonly described, followed by sexual advances [36,40]. Other instances of sexual harassment include unwanted attention and disrespectful body language [36].

Theme 2: Sources of Discrimination

Fourteen (53%) papers explored who contributed to the discrimination experienced by VMs and women [17,20,22,24,27,29,31,33-37,39,40]. Our literature review found that VMs experienced discrimination from a multitude of perpetrators. These include, but are not limited to, patients, families, attendings, peers, nurses, and other health care professionals [17,20,22,24,27,29,31,33-37,39,40].

Participants in 12 studies reported patients and patients’ families as either the main or the second most common perpetrators of discrimination [20,22,24,27,29,31,33-36,39,40]. Of those who

XSL•FO
RenderX

reported patients as the main source, these experiences of discrimination were both racial [20,22,24,27,29,31,33,34,39,40] and gender-based [27,29,33,39]. Patients were also the most common cause of sexual harassment against women residents, followed by staff, including attending physicians and nurses [20,27,29,33,39,40].

Eight studies reported that attending physicians were another major contributor to the racial [17,24,29,33-35,39] and gender [33,35,36,39] discrimination experienced by VM residents. Attending physicians were also found to be a major cause of sexual harassment [33,35,39], with 1 paper reporting them to be the main perpetrators [35]. They highlighted the negative impact this can have on residents, as these attendings are often their mentors who serve as guides for their education and, therefore, can influence resident learning environments [35]. Eight papers described how VM residents often experienced discrimination from nursing staff and other health care workers [17,22,27,31,33,34,37,39]. Only 4 papers mentioned peers and colleagues being a cause of discrimination [17,33,34,39].

Theme 3: Ramifications of Discrimination

Twenty-one (80%) papers found that participants experienced the ramifications of discrimination in a variety of ways, with many papers exploring the effects of discrimination and how they manifest [13,14,18-22,24,26-31,33-38,40]. A major consequence of discrimination was residents being reluctant to report their experiences [13,14,22,28,29,34,35,40]. Many participants described being afraid that reporting these instances could have repercussions for their professional career [22,29,34,35,40], especially for those who experienced discrimination from senior colleagues and leadership [22,29,34]. Participants also avoided reporting incidents because they did not have the time and energy to do so [14,22,29,34]. Some residents mentioned being worried that their efforts to report would be wasted, as their experiences would just be swept under the rug [13,34,40]. Residents also felt that their experiences were not significant enough to report [22,35,40]. One paper mentioned residents being worried that their report would be perceived as playing the “race card” [34].

Five papers found that learners experienced burnout due to discrimination [18,19,29,31,33], with 1 paper reporting that learners felt more fatigued and exhausted on top of their current program workload and expectations [29]. Four studies spoke about the feelings of loneliness [21,24,28,34], with extreme cases of learners experiencing feelings of being “othered” [21,24]. Discrimination also caused some residents to have suicidal thoughts [23,27,33], with studies finding an association between mistreatment exposures and the number of suicidal thoughts [27,33]. Learners also reported that discrimination caused them to feel the need to suppress their feelings, resulting in residents ignoring the incident altogether [29,30]. A few residents mentioned internalizing messages of discrimination [20,22], making them feel less competent in their clinical abilities [22] or feeling the need to hide their culture [20]. Residents who experienced discrimination reported poorer learning environments [35,38]. Similarly, in 1 study, residents who did not experience discrimination were more likely to rate their academic satisfaction [26] highly. As a result, many women and/or VM residents mentioned contemplating withdrawing from their residency program or quitting entirely [13,31,35,36,40].

Theme 4: Ways for Individuals and Institutions to Reduce Gender and Racial Discrimination

Nineteen (73%) studies discussed and provided potential solutions for reducing gender and racial discrimination experienced by residents [13,18,19,21-25,27-29,31,32,34-37,39,40]. These recommendations ranged from educational interventions that could support residents to changes that could be implemented at the institutional level.

Many papers recommended various curricular and training changes to reduce discrimination [18,19,21-23,27,29,31,32,35,39,40]. These included providing training on what explicit and implicit biases are and how to recognize them [18,19,21-23,27-29,32,35,40], as well as incorporating training on how to respond and intervene during instances of discrimination [18,19,22,27,29,31,39]. In addition to training residents, some papers suggested providing faculty and other health care professionals with these educational opportunities as well [18,21-23,29,32,35].

Several studies also proposed increasing the number of women and underrepresented minorities in faculty and leadership positions as a way for institutions to foster an inclusive environment [13,21,25,28,34]. Many participants believed that hiring more minorities into positions of power would allow more residents access to mentors who could provide the support and resources they need while facing challenges in residency [13,21,25,28,34].

Some papers suggested that institutions should set policies to improve working conditions for women and VMs. These included encouraging residents to report incidents when they occur [22-24,31,35,36,40] and keeping them safe from retribution [23,24,36,40] by using a confidential reporting system [36]. Another suggestion was to implement policies that required residents to refer offending patients to other physicians, thereby allowing residents to work in a safe environment while ensuring that patients received the care they needed [22].

Discussion

Principal Findings

The purpose of this scoping review was to explore and highlight resident experiences of gender and racial discrimination to raise awareness of the impact of these experiences on residency training programs. The findings revealed that residents predominantly reported experiencing microaggressions, misidentification, and sexual harassment from multiple sources, including attendings and other health care workers. These experiences frequently had negative repercussions, including adverse effects on residents’ feelings of competency.

Educational Impact of Microaggressions

Discrimination experienced by VMs is an ongoing issue that has plagued the health care system for a very long time. Microaggressions are becoming an increasingly common form

XSL•FO
RenderX

of discrimination experienced by residents and physicians of VM status [2,34]. Several studies allude to discrimination toward medical trainees as part of their training due to factors such as hierarchy [41,42]. A recent study by Gianakos et al [43] found that the majority of incidents reported by residents in the study were perpetrated by supervising attendings and their senior coresidents. Similarly, our analysis found that attendings were the most common source of discrimination, second only to patient interactions. Microaggressions from attendings and other senior colleagues can have serious and lasting consequences for residents, particularly with respect to career trajectory. The impacts of these microaggressions on future careers can start well before residency; for example, many women reported being discouraged by faculty from entering competitive specialties in residency, such as surgery, preventing these disciplines from becoming diverse [14,22,28]. Furthermore, microaggressions that occur during residency also impact resident careers. For instance, our findings indicate that residents from marginalized communities are commonly subjected to higher standards and are more likely to receive harsher evaluations than their non-VM peers [21]. Such discriminatory feedback can negatively affect residents' progress during training and may also discourage them from pursuing careers in academia [44]. Additionally, several papers reported that residents expressed concern about receiving fewer opportunities based on their gender and/or race [13,18,19,22,35,37]. Residents may also lack sponsorship and general support from attending physicians, which can limit access to career-advancing opportunities, including employment at prestigious academic institutions [44]. Consequently, fewer women and VMs attain leadership positions, resulting in reduced availability of mentorship and advocacy for marginalized residents [13,21].

Similar patterns have been observed among faculty physicians. Faculty identifying as VMs report feeling pressured to meet higher performance standards than their non-VM colleagues to maintain their positions or advance professionally [45]. This differential treatment may further contribute to the underrepresentation of VMs in senior academic and leadership roles [45].

Intersectionality

The compounded effects of discrimination faced by residents with intersectional identities were acknowledged in many papers; however, they were not explored in depth. Learning about the unique experiences of residents who face discrimination due to both their gender and race is essential for developing solutions, since these issues cannot be resolved by addressing gender or racial discrimination alone [13]. The intersectional experiences of physicians described in the literature are similar to those reported by residents. A study that interviewed Black female physicians found that many of them felt isolated from their colleagues and patients, similar to how residents were less likely to feel connected to staff [13,46]. Physicians also described feeling burdened with high expectations due to their race and gender, while many did not expect them to succeed [46]. More qualitative studies on the intersectional experiences of residents are necessary to reveal how they may differ from or relate to physician experiences.

Strategies for Reducing Discrimination and Microaggressions

Many papers within this scoping review proposed strategies for reducing gender and racial discrimination against residents. Although many papers recommended incorporating training on discrimination and intervening in situations of microaggressions, none of the papers mentioned implementing these suggestions within their institutions or evaluating the impact of these solutions on resident experiences. It is important to highlight that this is a notable gap in the literature, as there are many studies describing discrimination and possible solutions for reducing these experiences, but very few studies that assess the efficacy of these recommendations. A study conducted by Mullett et al [47] focusing on equity, diversity, and inclusion curriculum development for pediatric residents revealed that when residents were exposed to training sessions on microaggressions, their knowledge and awareness about microaggressions in a clinical setting increased. Additionally, the residents were able to propose effective strategies to deal with microaggressions.

In addition to providing educational opportunities for residents, several papers suggested training faculty and leadership about discrimination [18,21-23,29,32,35] and how to respond and intervene during these situations [22,27,31]. Including staff and faculty in these educational sessions could motivate witnesses to intervene when an incident related to discrimination occurs [22,27,31]. Similar to the literature on residents, there is limited research evaluating the impact of such interventions on faculty. Gleeson et al [48] described the development and outcomes of a diversity, equity, and anti-racism curriculum implemented at their medical center. The curriculum aimed to raise awareness and equip physicians and trainees with skills to recognize racism through didactic sessions and interactive workshops. Although participants reported increased knowledge and rated the sessions as impactful, survey data collected longitudinally showed no change in the institutional discrimination experienced by participants or by individuals they knew.

Another strategy for decreasing discrimination was to increase the number of women and VMs in leadership positions, particularly for expanding the number of mentors and role models available for residents [13,21,25,28,34]. Mentorship is considered to be one of the most significant aspects of residency that can greatly impact a resident's experience, and studies show that women of VM status are the least likely to have a mentor during residency [13]. In addition to mentorship, VMs recruited into positions of leadership would be more likely to be advocates for residents who identify as women and VMs [21].

Implications

Findings from this scoping review generated new perspectives and knowledge to better understand how race and gender can combine to impact resident experiences. This scoping review illuminated areas of postgraduate medical education that need change and highlighted additional actions required to enhance resident experiences. Understanding how gender and racial discrimination intersect and manifest at the postgraduate level can aid programs in developing a tailored response to these challenges. Improving resident experiences and supporting them

XSL•FO
RenderX

in addressing these issues can positively influence their interactions with colleagues and patients, leading to improved patient care. Additionally, raising awareness of this topic can educate residents, physicians, and other staff, fostering better treatment of colleagues based on their gender and race. The results of this scoping review will contribute to postgraduate medical programs evaluating their curricula and refining or incorporating training on discrimination within existing programs.

Strengths and Limitations

This scoping review was based on a rigorous search and methodological process. The search strategy included many synonyms and was adapted for 5 databases. We followed the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) reporting checklist, which added to the rigor of our scoping review. However, some limitations should be acknowledged. We did not impose date restrictions because we aimed to examine how residents' experiences with discrimination may have changed over time. However, it is important to note that the composition of the resident workforce has evolved substantially during this period, particularly with an increase in the proportion of women in medical training. Although insufficient historical data precluded direct comparisons across earlier periods, our findings suggest that residents' experiences of discrimination have remained largely similar over time. Some studies may have been overlooked due to our search being limited to papers in English. Additionally, although we examined studies from different countries, our findings may not be generalizable to all residency programs due to differing societal contexts, varying norms, and possibly different training structures.

Conclusion

Findings from this scoping review describe the different ways women and VM residents often face discrimination, who perpetrates it, the consequences it may have on learners, and steps that individuals and institutions can take to reduce discrimination. Even though residents have been reporting experiences of gender and racial discrimination for many years, discrimination continues to be a persistent and ongoing issue in many residency programs. Learning about and raising awareness of resident experiences can assist institutions and residency programs in developing tailored solutions, especially for those who face intersectional discrimination.

Acknowledgments

We would like to thank the editor and the reviewers of our paper for their valuable comments.

Funding

No external financial support or grants were received for this work.

Data Availability

The datasets generated or analyzed during this study are available from the corresponding author on reasonable request.

Authors' Contributions

Conceptualization was conducted by IJ and MR
Data curation was conducted by IJ, JN, and MR
Methodology, formal analysis, and validation were conducted by IJ, LD, JN, and MR
Project administration was conducted by IJ
Visualization was conducted by IJ and LD
The original draft was written by IJ, LD, JN, and MR
Review and editing were conducted by IJ, LD, JN, and MR

Conflicts of Interest

None declared.

Multimedia Appendix 1
Search strategy.
[DOCX File, 28 KB - mededu_v12i1e87524_app1.docx]

Checklist 1
PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses Extension for Scoping Reviews) checklist.
[PDF File, 331 KB - mededu_v12i1e87524_app2.pdf]

References

1. Half of racialized people have experienced discrimination or unfair treatment in the past five years. Statistics Canada. URL: https://www150.statcan.gc.ca/n1/daily-quotidien/240516/dq240516b-eng.htm [accessed 2026-01-20]

XSL•FO
RenderX

2. About discrimination. The Canadian Human Rights Commission. URL: https://www.chrc-ccdp.gc.ca/individuals/human-rights/about-discrimination [accessed 2026-01-20]
3. Phillips NA, Tannan SC, Kalliainen LK. Understanding and overcoming implicit gender bias in plastic surgery. Plast Reconstr Surg 2016 Nov;138(5):1111-1116. [doi: 10.1097/PRS.0000000000002668] [Medline: 27391836]
4. Turner J, Higgins R, Childs E. Microaggression and implicit bias. Am Surg 2021 Nov;87(11):1727-1731. [doi: 10.1177/00031348211023418] [Medline: 34078131]
5. Filut A, Alvarez M, Carnes M. Discrimination toward physicians of color: a systematic review. J Natl Med Assoc 2020 Apr;112(2):117-140. [doi: 10.1016/j.jnma.2020.02.008] [Medline: 32197899]
6. Visible minority of person. Statistics Canada. URL: https://www23.statcan.gc.ca/imdb/p3Var.pl?Function=DEC&Id=45152 [accessed 2026-01-20]
7. Dyrbye LN, West CP, Sinsky CA, et al. Physicians' experiences with mistreatment and discrimination by patients, families, and visitors and association with burnout. JAMA Netw Open 2022 May 2;5(5):e2213080. [doi: 10.1001/jamanetworkopen.2022.13080] [Medline: 35587344]
8. Wheeler M, de Bourmont S, Paul-Emile K, et al. Physician and trainee experiences with patient bias. JAMA Intern Med 2019 Dec 1;179(12):1678-1685. [doi: 10.1001/jamainternmed.2019.4122] [Medline: 31657839]
9. Nunez-Smith M, Curry LA, Bigby J, Berg D, Krumholz HM, Bradley EH. Impact of race on the professional lives of physicians of African descent. Ann Intern Med 2007 Jan 2;146(1):45-51. [doi: 10.7326/0003-4819-146-1-200701020-00008] [Medline: 17200221]
10. Crenshaw K. Demarginalizing the intersection of race and sex: a Black feminist critique of antidiscrimination doctrine, feminist theory and antiracist politics. Univ Chic Leg Forum 1989(1):139-167 [FREE Full text]
11. Davis K. Intersectionality as buzzword: a sociology of science perspective on what makes a feminist theory successful. Fem Theory 2008;9(1):67-85. [doi: 10.1177/1464700108086364]
12. Pololi LH, Jones SJ. Women faculty: an analysis of their experiences in academic medicine and their coping strategies. Gend Med 2010 Oct;7(5):438-450. [doi: 10.1016/j.genm.2010.09.006] [Medline: 21056870]
13. Mocanu V, Kuper TM, Marini W, et al. Intersectionality of gender and visible minority status among general surgery residents in Canada. JAMA Surg 2020 Oct 1;155(10):e202828. [doi: 10.1001/jamasurg.2020.2828] [Medline: 32804994]
14. Bhatt W. The little Brown woman: gender discrimination in American medicine. Gender & Society 2013;27(5):659-680. [doi: 10.1177/0891243213491140]
15. Arksey H, O'Malley L. Scoping studies: towards a methodological framework. Int J Soc Res Methodol 2005 Feb;8(1):19-32. [doi: 10.1080/1364557032000119616]
16. Tricco AC, Lillie E, Zarin W, et al. PRISMA Extension for Scoping Reviews (PRISMA-ScR): checklist and explanation. Ann Intern Med 2018 Oct 2;169(7):467-473. [doi: 10.7326/M18-0850] [Medline: 30178033]
17. Zaeem Z, Smyth P, Lewis M, Ruzycki S. Workplace discrimination and harassment among Alberta postgraduate medical trainees: a cross-sectional survey. BMC Med Educ 2024 Nov 26;24(1):1376. [doi: 10.1186/s12909-024-06363-5] [Medline: 39593015]
18. Kim G, Lodha S, Wein L, et al. Evaluating the impact of gender, race, and training year on internal medicine residents' experiences across the United States. Perm J 2024 Sep 16;28(3):107-116. [doi: 10.7812/TPP/24.085] [Medline: 39192722]
19. Lodha S, Luzum N, Washabaugh C, et al. Evaluating the impact of gender and race on otolaryngology resident experiences across the United States. Otolaryngol Head Neck Surg 2025 Jan;172(1):144-151. [doi: 10.1002/ohn.966] [Medline: 39224039]
20. Chiraroekmongkon S. The invisibility cloak: narrative from an Asian American woman physician-in-training. Psychiatr Serv 2024 Jan 1;75(1):92-93. [doi: 10.1176/appi.ps.23074022] [Medline: 38161316]
21. Koech H, Albanese J, Saeks D, et al. Minority resident physicians' perspectives on the role of race/ethnicity, culture, and gender in their surgical training experiences. J Surg Educ 2023 Jun;80(6):833-845. [doi: 10.1016/j.jsurg.2023.03.009] [Medline: 37121866]
22. Hussain NM, Spiers J, Kobab F, Riley R. The impact of race and gender-related discrimination on the psychological distress experienced by junior doctors in the UK: a qualitative secondary data analysis. Healthcare (Basel) 2023 Mar 12;11(6):834. [doi: 10.3390/healthcare11060834] [Medline: 36981491]
23. Tyler WK. CORR Insights®: the majority of Black orthopaedic surgeons report experiencing racial microaggressions during their residency training. Clin Orthop Relat Res 2023 Apr 1;481(4):687-689. [doi: 10.1097/CORR.0000000000002517] [Medline: 36752747]
24. Brooks JT, Porter SE, Middleton KK, Carson EW, Ode GE. The majority of Black orthopaedic surgeons report experiencing racial microaggressions during their residency training. Clin Orthop Relat Res 2023 Apr 1;481(4):675-686. [doi: 10.1097/CORR.0000000000002455] [Medline: 36342502]
25. Patel S, Lin KK, Milam AJ, et al. Diversity, equity, and inclusion among anesthesiology trainees. Womens Health Rep (New Rochelle) 2022;3(1):414-419. [doi: 10.1089/whr.2021.0123] [Medline: 35559355]
26. Sandoval-Bonilla BA, la Cerda-Vargas MFD, Stienen MN, et al. Discrimination of residents during neurosurgical training in Mexico: results of a survey prior to SARS-CoV-2. Surg Neurol Int 2021;12:618. [doi: 10.25259/SNI_813_2021] [Medline: 34992934]

XSL•FO
RenderX

27. Lall MD, Bilimoria KY, Lu DW, et al. Prevalence of discrimination, abuse, and harassment in emergency medicine residency training in the US. JAMA Netw Open 2021 Aug 2;4(8):e2121706. [doi: 10.1001/jamanetworkopen.2021.21706] [Medline: 34410392]
28. Villanueva C, Cain J, Greenhill J, Nestel D. "The odds were stacked against me": a qualitative study of underrepresented minorities in surgical training. ANZ J Surg 2021 Oct;91(10):2026-2031. [doi: 10.1111/ans.17168] [Medline: 34476888]
29. Zhuo L, Ju V, Wakam G, Antunez A, Dossett LA. Facilitators and barriers to allyship in academic surgery: a qualitative study. Am J Surg 2021 May;221(5):950-955. [doi: 10.1016/j.amjsurg.2020.08.051] [Medline: 32928541]
30. Mgbako O. Bonds in every color. Ann Intern Med 2020 Sep 15;173(6):498. [doi: 10.7326/M20-0856] [Medline: 32598219]
31. Yuce TK, Turner PL, Glass C, et al. National evaluation of racial/ethnic discrimination in US surgical residency programs. JAMA Surg 2020 Jun 1;155(6):526-528. [doi: 10.1001/jamasurg.2020.0260] [Medline: 32293655]
32. Pearce G, Sidhu N, Cavadino A, Shrivathsa A, Seglenieks R. Gender effects in anaesthesia training in Australia and New Zealand. Br J Anaesth 2020 Mar;124(3):e70-e76. [doi: 10.1016/j.bja.2019.12.020] [Medline: 31982114]
33. Hu YY, Ellis RJ, Hewitt DB, et al. Discrimination, abuse, harassment, and burnout in surgical residency training. N Engl J Med 2019 Oct 31;381(18):1741-1752. [doi: 10.1056/NEJMsa1903759] [Medline: 31657887]
34. Osseo-Asare A, Balasuriya L, Huot SJ, et al. Minority resident physicians' views on the role of race/ethnicity in their training experiences in the workplace. JAMA Netw Open 2018 Sep 7;1(5):e182723. [doi: 10.1001/jamanetworkopen.2018.2723] [Medline: 30646179]
35. Fitzgerald CA, Smith RN, Luo-Owen X, et al. Screening for harassment, abuse, and discrimination among surgery residents: an EAST multicenter trial. Am Surg 2019 May 1;85(5):456-461. [doi: 10.1177/0003134819085005] [Medline: 31126355]
36. Fnais N, al-Nasser M, Zamakhshary M, et al. Prevalence of harassment and discrimination among residents in three training hospitals in Saudi Arabia. Ann Saudi Med 2013;33(2):134-139. [doi: 10.5144/0256-4947.2013.134] [Medline: 23563000]
37. Crutcher RA, Szafran O, Woloschuk W, Chatur F, Hansen C. Family medicine graduates' perceptions of intimidation, harassment, and discrimination during residency training. BMC Med Educ 2011 Oct 24;11:88. [doi: 10.1186/1472-6920-11-88] [Medline: 22018090]
38. London L, Kalula S, Xaba B. A survey of trainee specialists experiences at the University of Cape Town (UCT): impacts of race and gender. BMC Med Educ 2009 May 27;9(1):26. [doi: 10.1186/1472-6920-9-26] [Medline: 19473495]
39. vanIneveld CH, Cook DJ, Kane SL, King D. Discrimination and abuse in internal medicine residency. The Internal Medicine Program Directors of Canada. J Gen Intern Med 1996 Jul;11(7):401-405. [doi: 10.1007/BF02600186] [Medline: 8842931]
40. McNamara RM, Whitley TW, Sanders AB, Andrew LB. The extent and effects of abuse and harassment of emergency medicine residents. Acad Emerg Med 1995 Apr;2(4):293-301 [FREE Full text] [doi: 10.1111/j.1553-2712.1995.tb03225.x]
41. Fnais N, Soobiah C, Chen MH, et al. Harassment and discrimination in medical training: a systematic review and meta-analysis. Acad Med 2014 May;89(5):817-827. [doi: 10.1097/ACM.0000000000000200] [Medline: 24667512]
42. Camargo A, Radmard M, Dabiri M, Carson KA, Hsu L, Yousem DM. Discrimination faced by radiology residents: an analysis of experiences and mitigation strategies. Curr Probl Diagn Radiol 2024;53(5):600-605. [doi: 10.1067/j.cpradiol.2024.05.013] [Medline: 38714392]
43. Gianakos AL, Freischlag JA, Mercurio AM, et al. Bullying, discrimination, harassment, sexual harassment, and the fear of retaliation during surgical residency training: a systematic review. World J Surg 2022 Jul;46(7):1587-1599. [doi: 10.1007/s00268-021-06432-6] [Medline: 35006329]
44. Ibrahim H, Juve AM, Amin A, Railey K, Andolsek KM. Expanding the study of bias in medical education assessment. J Grad Med Educ 2023 Dec;15(6):623-626. [doi: 10.4300/JGME-D-23-00027.1] [Medline: 38045936]
45. Hassouneh D, Lutz KF, Beckett AK, Junkins EP, Horton LL. The experiences of underrepresented minority faculty in schools of medicine. Med Educ Online 2014;19(1):24768. [doi: 10.3402/meo.v19.24768] [Medline: 25472784]
46. Chilakala A, Camacho-Rivera M, Frye V. Experiences of race- and gender-based discrimination among Black female physicians. J Natl Med Assoc 2022 Feb;114(1):104-113. [doi: 10.1016/j.jnma.2021.12.008] [Medline: 35058066]
47. Mullett TA, Rooholamini SN, Gilliam C, McPhillips H, Grow HM. Description of a novel curriculum on equity, diversity and inclusion for pediatric residents. J Natl Med Assoc 2022 Jan;113(6):616-625. [doi: 10.1016/j.jnma.2021.05.014] [Medline: 34172296]
48. Gleeson SE, Zapata H, Bathgate ME, et al. Building an Infectious Disease Diversity, Equity, and Antiracism (ID2EA) curriculum: a single center's experience and reflections. Clin Infect Dis 2023 Sep 11;77(5):703-710. [doi: 10.1093/cid/ciad236] [Medline: 37078888]

Abbreviations

PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
PRISMA-ScR: Preferred Reporting Items for Systematic Reviews and Meta-Analyses Extension for Scoping Reviews
VM: visible minority

XSL•FO
RenderX

Edited by M Molina; submitted 10.Nov.2025; peer-reviewed by HEK Randell-Moon, I Menchetti, J Moll; accepted 27.Feb.2026; published 02.Apr.2026.

Please cite as:
John I, Dennett L, Nguyen J, Rashid M
Gender and Racial Discrimination During Residency Training: Scoping Review
JMIR Med Educ 2026;12:e87524
URL: https://mededu.jmir.org/2026/1/e87524
doi:10.2196/87524

© Ida John, Liz Dennett, Julie Nguyen, Marghalara Rashid. Originally published in JMIR Medical Education (https://mededu.jmir.org), 2.Apr.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Education, is properly cited. The complete bibliographic information, a link to the original publication on https://mededu.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Health and Humanities and Social Sciences Professionals' Perceptions Regarding the Teaching of the Effects of Racism in Medicine: Semistructured Interview Study

Mathilde Lambert[1,2], MD; Rahmeth Radjack[1,2,3], MD, PhD; Adil Mouhab[1,2,3], MD; Malika Mansouri[4], Prof Dr, PhD; Marie Rose Moro[1,2,3], Prof Dr Med, MD, PhD

[1]Université Paris Cité, 71 Avenue Edouard Vaillant, Boulogne-Billancourt (Hauts-de-Seine), France

[2]CESP-UVSQ, DevPsy, INSERM, Université Paris-Saclay, Villejuif, France

[3]Cochin Hospital, APHP, Maison de Solenn, Paris, France

[4]UTRPP Laboratory, Université Sorbonne Paris Nord, Villetaneuse, France

Corresponding Author:
Mathilde Lambert, MD
Université Paris Cité, 71 Avenue Edouard Vaillant, Boulogne-Billancourt (Hauts-de-Seine), France

Abstract

Background: In France, as in other countries, health disparities arise from multiple interacting factors, among which racism plays a significant role. Racism affects health through exposure to discrimination-related stress, environmental conditions, differential access to and quality of care, as well as representations and behaviors, some of which are rooted in the historical legacy of colonial medicine.

Objective: This study aimed to explore the perspectives of health care professionals and researchers in the humanities and social sciences (HSS) on the core content areas to be defined and the corresponding learning objectives to be established for the development of a curriculum addressing racial discrimination in medicine for undergraduate medical education.

Methods: Physicians, psychologists, and researchers in the HSS who had published on the medical care of people from racialized minorities and who had organized or participated in educational initiatives on this topic were recruited. Semistructured interviews were conducted and analyzed using reflexive thematic analysis.

Results: This study is the first French study to bring together the experiences of French-speaking experts in this field to inform the development of educational content integrating medicine, psychology, and the HSS. A total of 20 participants were interviewed. Three main themes emerged: (1) teaching how to identify and dismantle stereotypes in clinical practice; (2) teaching the psychological effects of racism and its overall impact on health; and (3) teaching the risks associated with racialized care and so-called colorblind approaches. Participants emphasized the importance of addressing racial stereotypes specific to the French context, such as the "Mediterranean syndrome," situating them within their historical background, and understanding the effects of discrimination on both physical and mental health. A key pedagogical challenge identified was finding a balance between acknowledging geographical or sociological specificities and avoiding culturalist interpretations that overemphasize context. All participants highlighted the necessity of such comprehensive education to ensure patient safety for all.

Conclusions: Teaching about the effects of racial discrimination and racism on health should be framed as a core public health issue. Although racial discrimination is influenced by economic, political, and sociological factors that extend beyond the medical sphere, its impact on health is neither inevitable nor immutable and therefore warrants explicit educational attention. From a future perspective, the ongoing French debate regarding the feasibility of conducting epidemiological studies on racial health disparities needs to be addressed to further inform education, research, and policy.

(JMIR Med Educ 2026;12:e90084) doi:10.2196/90084

KEYWORDS

racial discrimination; racism; Mediterranean syndrome; health inequalities; medical pedagogy; qualitative methods; health professionals; humanities and social sciences

Introduction

The term racism, as defined by Camara Jones in a seminal article, refers to a conceptual framework encompassing 3 levels: institutionalized racism, personally mediated racism, and internalized racism [1]. Race is understood as a social construct that accurately captures the impacts of racism, given that racial classification has no legitimate biological basis [2]. Racial

XSL•FO
RenderX

discrimination, as the effect or outcome of prejudice, refers to any unfavorable treatment of an individual based on these grounds, as well as any action that limits the opportunities of an individual or group due to characteristics such as race. In France, as in other countries where these phenomena have been documented, health disparities result from multiple interacting factors, with racism constituting a significant determinant.

First, the effects of racism and racial discrimination on health in general—and on mental health in particular—have been documented for several decades [3-8].

Second, challenging the genetic basis of racial categories and acknowledging that socioeconomic factors alone do not fully account for health disparities, a 2022 *Lancet* Commission report [9] examined the multiple pathways leading to poorer health outcomes among racialized minority populations. Stress responses associated with discrimination—including potential epigenetic changes that may be transmitted to subsequent generations—inequitable access to education, employment, healthy housing free from air pollution, green spaces, nutritious food, as well as the disproportionate health effects of climate change on minoritized populations, increase vulnerability to disease, coupled with higher occupational exposure risks and more limited access to timely diagnosis and appropriate treatment. These latter dimensions are particularly crucial for medical students to understand, as they will ultimately be key actors in addressing and reducing these inequities.

Such understanding requires multiple levels of analysis and dedicated educational content, especially since barriers to diagnosis and treatment are historically embedded, including within France's colonial past. In France, the medical field has historically produced and perpetuated racial and racist stereotypes, traces of which remain observable in contemporary clinical practice. The French Defender of Rights has highlighted persistent obstacles in patient-provider communication, clinical management, and access to medications [10].

As early as the 20th century, Foucault [11] emphasized that the use of race introduces differential management of populations. Paradoxically, even if the French context is characterized by an official rejection of racial or ethnic categories, Fassin [12] has described the use of substitute categories that allow differences to be discussed without being explicitly named, a practice that may reinforce stigmatization and contribute to the construction of social identities. The resulting differentiated care—whether related to access to health care facilities [13], implicit biases leading to overdiagnosis or underdiagnosis of certain conditions [14], or the use of clinical formulas incorporating racial characteristics without scientific basis [15,16]—remains present in France and requires dismantling through multilevel interventions. A key challenge is ensuring the appropriate use of geographical and sociological specificities to enhance patient safety without reinforcing racialized assumptions.

Despite these challenges, only a limited number of courses addressing these issues are currently offered to French medical students [17], and they largely stem from individual initiatives, notably those developed over the past 3 decades by specialists in transcultural psychiatry in France. Their work has focused on access to professional interpretation services, shared cultural representations of care, consideration of migration trajectories when relevant, and the identity construction of children and adolescents navigating multiple spheres of belonging—processes that can be profoundly affected by racial discrimination.

With a view to improving this pedagogical approach in France, the objective of this study is therefore to explore the perspectives of health care professionals and scholars in the humanities and social sciences (HSS) on the core content areas to be defined and the corresponding learning objectives to be established for the development of a curriculum addressing racial discrimination in medicine for undergraduate medical education.

Methods

Study Design

A qualitative study was conducted from January to June 2024 using semistructured interviews and reflexive thematic analysis (TA) and was designed and reported in accordance with the COREQ (Consolidated Criteria for Reporting Qualitative Research) guidelines. This approach was appropriate for exploring how health professionals and experts in the HSS understand and teach the effects of racism in medicine.

Participant Characteristics

Inclusion criteria were being a French-speaking professional specializing in medical issues and having worked on matters of racial discrimination in medicine, that is, having published on health care for racial minority populations and having organized or participated in related training or courses.

The study focuses on the French context by interviewing individuals with in-depth knowledge of this setting. Most participants were therefore French. However, the still-emerging nature of medical education in France in this area supported the inclusion of several francophone professionals from other countries, with the aim of fostering the exchange of best practices. Purposive sampling ensured diversity across profession, epistemology, practice setting, country, age, and gender to maximize qualitative validity. Participants included professors of medicine or psychology, teacher-researchers in the HSS with PhDs, and health care professionals with advanced academic qualifications in HSS. Several had established university programs on transcultural issues and health inequalities, led courses on discrimination, or coordinated medical curricula integrating these topics. Medical education on these topics must be conceived as inherently interdisciplinary and draw on multiple methodological approaches. This rationale motivated the inclusion, in this study, of professionals from diverse disciplinary backgrounds.

Data Generation

Participants were contacted via email by the primary investigator (ML). Semistructured interviews were conducted by ML from March to July 2024, either in person or remotely for those outside the Île-de-France region, lasting approximately 60 - 90 minutes. Verbal consent was audio-recorded. Out of 31 contacted individuals, 3 declined due to availability, 8 did not respond, and 20 agreed to participate.

XSL•FO
RenderX

Interview guides (Multimedia Appendix 1) were tailored to the participants' professional profiles, grouped into 3 categories: physicians, psychologists, and teacher-researchers in HSS. Topics included racial discrimination in health care, educational practices and formats, physical and psychological impacts of racism, awareness-raising strategies, and the contribution of HSS to teaching. Audio recordings were transcribed verbatim.

Data Analysis

A reflexive TA was conducted using the 6-step framework by Braun and Clarke [18-20]: familiarization with data, generation of initial codes, development and review of subthemes and themes, refinement of themes, and reporting. Transcripts were analyzed using NVivo 14 (Lumivero). Initial inductive coding was performed sequentially by ML. The study was based on a logic of interpretive sufficiency, with the corpus being deemed adequate when it supported an in-depth and coherent reflexive analysis of the phenomenon under study.

Codes were iteratively grouped into subthemes and themes through multiple revisions, resulting in an initial thematic framework. A cross-sectional analysis was conducted independently by ML and MM, that is, a precise and systematic comparison and analysis of themes and subthemes in relation to codes and significant verbatim quotes. It was followed by iterative discussions with ML, MM, RR, and MRM to ensure triangulation and reflexivity. Disagreements were pooled and discussed each time they arose, with back-and-forth exchanges of arguments and a collective decision made. This process led to consensus on final themes, supporting rigor and credibility. The analysis was data-driven, with themes identified at a semantic level and developed iteratively from description to interpretation [21]. Reflexive TA procedures reflect the values of a qualitative paradigm, centering researcher subjectivity, organic and recursive coding processes, and the importance of deep reflection on, and engagement with, data [20].

Team Reflexivity

The positionality of the research team may have influenced data collection and analysis. The interviewer (ML), a physician perceived as belonging to the French majority group, could have shaped interview dynamics and participants' responses. In accordance with French regulations, no systematic data on participants' ethnic or self-identified group affiliations were collected, though some participants voluntarily shared aspects of their background. This limitation may affect the interpretation regarding participants' positioning relative to the majority group.

The authors of this article have academic and professional training in France, with diverse cultural backgrounds and family histories marked by migration, which adds value and richness to the perspective. The authors are trained in transcultural psychiatry, which involves the complementary use of data from humanities disciplines such as anthropology, history, and sociology. One of the key concepts of transcultural psychiatry is decentering, which allows one to put oneself in another's shoes. Qualitative data triangulation spaces were regularly used to enable participants to become aware of their own feelings and biases toward the research topic.

The authors' engagement with antiracism issues may have influenced the interpretation of findings as tools for improving teaching practices. Reflexive triangulation was used among team members with diverse professional backgrounds, experiences, and age groups. Collective discussions, including a monthly doctoral seminar, supported ongoing reflexivity. These reflective discussions led to the renaming of themes that were potentially too descriptive and insufficiently analyzed, refocusing them on their pedagogical implications in relation to the study's objective. For example, the first theme, "Pain and Mind and Mind-Body Dualism: The Invocation of Stereotypes," was renamed "Teaching How to Identify and Deconstruct Stereotypes in Clinical Practice." Similarly, the theme "Inequalities in Access, Epidemiology, and Differentiated Care" was removed after discussion and merged with the other 3 themes because it did not, on its own, address the objective.

Cultural countertransference [22] was explicitly considered as an analytic resource, allowing engagement with participants' experiences of alterity while avoiding culturalist or exoticizing interpretations.

Ethical Considerations

Participants were contacted by email and informed about the study objectives, interview procedures, and confidentiality measures. Oral consent was obtained for both participation and the use of interview data in publications. Each participant was assigned a random alphanumeric code to ensure confidentiality. Interviews were analyzed by the research team, and data were securely stored on a password-protected computer managed by the primary investigator (ML). Data were retained for 1 year and subsequently destroyed. The study was approved by the INSERM (French National Institute for Health and Medical Research) ethics committee in March 2024 (CEEI/IRB No 24 - 1068) and conducted in accordance with the ethical principles of the Declaration of Helsinki.

Results

Overview

The study included 20 participants from diverse professional backgrounds, including medical doctors, psychologists, and teachers in the HSS (Table 1).

XSL•FO
RenderX

Table . Characteristics of the study participants (N=20).

Participants	Gender	Profession	Country
ID01, M	Women	Child and adolescent psychiatrist	France
ID02, nurse and HSS	Men	Nurse and Anthropologist	France
ID03, M	Women	Psychiatrist	France
ID04, M	Men	Dermatologist	France
ID05, M	Women	Pediatrician	France
ID06, HSS	Women	Sociologist	France and United States
ID07, M	Women	Gynecologist-Obstetrician	France
ID08, M	Men	Child and adolescent psychiatrist	Canada
ID09, P	Men	Psychologist	Canada
ID10, HSS	Women	Historian	France
ID11, M	Men	General practitioner	France
ID12, M	Women	Child and adolescent psychiatrist	France
ID13, M	Women	Psychiatrist	Canada
ID14, M	Men	Internist	Switzerland
ID15, HSS	Men	Anthropologist	United States
ID16, HSS	Men	Sociologist	Belgium
ID17, midwife and HSS	Women	Midwife and Sociologist	France
ID18, nurse and HSS	Women	Nurse and Sociologist	France
ID19, P	Women	Psychologist	France
ID20, P	Women	Psychologist	France

This multidisciplinary sample enabled the inclusion of complementary perspectives on the topic. Health care professionals contributed insights grounded in clinical and hospital practice, as well as in their interactions with students. These perspectives both complemented and contrasted with psychological approaches, which tend to emphasize how students receive information and the potential psychological effects of discrimination. The inclusion of the HSS was also essential, as these disciplines provide conceptual frameworks for understanding the social and historical dimensions of medicine, thereby supporting students' comprehension of the broader context in which their training takes place.

For clarity, participants' professional roles are identified in verbatim quotations as follows: medical doctor (M), psychologist (P), and humanities and social sciences teacher (HSS).

Three main themes emerged (Table 2): (1) teaching how to identify and dismantle stereotypes in clinical practice; (2) teaching the psychological effects of racism and its overall impact on health; and (3) teaching the risks associated with racialized care and so-called colorblind approaches.

Regarding the terms used in the results, it is worth recalling that although race is widely recognized as a social construct rather than a biological category, observable phenotypic traits and patterns of genetic ancestry may carry clinically relevant information in specific contexts.

Phenotype can function as a provisional clinical cue—particularly in contexts where genetic ancestry, environmental exposures, or structural inequities correlate with differential disease prevalence or treatment response. However, its use must remain probabilistic, context-dependent, and critically examined. Treating race as a biological essence risks reifying socially constructed categories and perpetuating bias; conversely, ignoring how socially mediated patterns of ancestry and embodiment affect health outcomes may obscure clinically relevant risk stratification.

Therefore, the pedagogical task is to train clinicians to distinguish among race as a social category, ancestry as a population-level genetic construct, and phenotype as an observable set of traits—while foregrounding the role of structural determinants of health. Responsible clinical reasoning requires acknowledging human variation without naturalizing racial categories or reinforcing deterministic assumptions.

XSL•FO
RenderX

Table . Themes and subthemes found.

Themes	Subthemes
Teaching how to identify and dismantle stereotypes in clinical practice	• Mediterranean syndrome and the hierarchy of pain • Ignoring context or overemphasizing it: two risks of misdiagnosis • Learning about illness-related behaviors through history to move beyond stereotypes
Teaching the psychological effects of racism and its overall impact on health	• Understanding the impact of racial discrimination on health and ensuring patient safety • Understanding the consequences of assimilation to support patients' processes of subjectivation
Teaching the risks associated with racialized care and so-called colorblind approaches	• The pitfall of culturalism • The question of differentiated formulas or protocols • Explaining the concept of sociological race and accounting for health-relevant ethnic differences

Teaching How to Identify and Dismantle Stereotypes in Clinical Practice

Mediterranean syndrome and representations of pain crystallize racist stereotypes in care, leading to diagnostic errors, including either a culturalist consideration of the context or ignorance of this context and rejection of the associated psychological causes. All of this creates a risk of misdiagnosis and medical error.

Mediterranean Syndrome and the Hierarchy of Pain

Racial stereotypes were highlighted by most participants, with a particular focus on pain, both in France, including in La Réunion, as well as in Switzerland and Canada:

> *Yes, Comorians and Mahorais are compared to other inhabitants. There's a syndrome, the fatigue syndrome. We'll tell you, the Mahorais are tired.* [ID02, nurse and HSS]

Mediterranean syndrome was central to the discussions and was described as an archetypal racist medical stereotype in Europe. This racist pseudosyndrome refers to the inadequate clinical management of symptoms perceived as exaggerated, based on the essentialization of individuals according to assigned racial identities and associated prejudices, often rooted in colonial medicine or the history of slavery. It has been identified as a source of medical error and potentially preventable harm. In France, this phenomenon has primarily affected individuals identified as Arab, Black, or even Roma, while in Switzerland, similar mechanisms have historically targeted Italian patients under the label of "Transalpine syndrome."

In contrast, in North America, participants referred to stereotypes portraying Black patients as having a higher resistance to pain. Referring to the example of sick leave duration being adjusted according to racial prejudice, one participant emphasized the need to avoid differential treatment, which otherwise leads to racial and social discrimination.

> *The doctor doesn't have to ask himself that question, he just has to hand out a piece of paper. The people who live in Montréal-Nord can barely read or don't know the system. Above all, they are already ashamed and don't want to ask for more. And so these are people who are going to go to hospital, receive care, and will never have time off or rest. So these are racial issues, it's racism, it's racial discrimination.* [ID09, P]

A participant from the United States highlighted that the sexist dimension of historical stereotypes related to pain is often compounded, reinforcing the dehumanization and hierarchization of patients. He referred to the notion of white vulnerability as a mirror image of prejudice that leads health care professionals to prioritize comfort and pain management for White patients.

> *Deirdre Cooper Owens, a historian, has clearly shown that there is a whole historical tradition that can be traced back to the beginning of the 19th century, around black women who are capable of giving birth without pain and who are naturally very fertile. There's also the whole issue of dehumanization behind it, white vulnerability, the fact that more attention is paid to treating pain.* [ID06, HSS]

Ignoring Context or Overemphasizing It: Two Risks of Misdiagnosis

Two types of diagnostic error related to patient context were identified. Participants described, on the one hand, the risk of ignoring contextual factors, including a lack of awareness of certain geographical or cultural medical specificities. On the other hand, they highlighted the risk of overemphasizing context, leading to culturalism and diagnostic error, described by one participant as "the overshadowing" (ID04, M)

Overemphasis on context was associated with an alleged knowledge of culturally specific signs that obscured standard diagnostic reasoning.

> *The psychiatrist tells me, this is a rite of passage. All teenage girls go through some sort of djinn crisis during adolescence. She came back a few days later, a young girl who was really decompensating. There was nothing cultural about it, especially as nothing could lead us to think of a cultural aspect.* [ID02, nurse and HSS]

Another participant provided the example of a patient whose traumatic experience had resulted in severe physical suffering.

XSL•FO
RenderX

She expressed indignation at the failure to consider the psychological context and at the immediate attribution of symptoms to "Mediterranean syndrome" by physicians who, in the absence of organic findings, resorted to racist prejudices.

Learning About Illness-Related Behaviors Through History to Move Beyond Stereotypes

Participants emphasized the relevance of teaching the history of science to contextualize the emergence of certain data rooted in the racist views of scientists, particularly during the colonial period. One participant referred to the notion of a North-South gradient in pain expression.

> *There was a period when it was politically correct, and there were even sociologists and anthropologists to corroborate that there was indeed a gradient in the expression of pain from North to South, that people shouted louder and louder when they were in pain. Then we realized that these were sociologists who had simply endorsed the racism of the doctors they were observing, that they were simply in the process of scientifically sociologising racist comments.* [ID11, M]

For this participant, focusing on illness-related behaviors allows clinicians to address individual characteristics in care delivery, thereby avoiding stereotypical reasoning and reducing the risk of misdiagnosis.

> *Is it the way in which, when we perceive an internal disturbance, we approach the medical profession? And what words do we use, how do we look, what questions do we ask? This approach to illness gives rise to listening skills. Because doctors need to be trained in how to deal with illness in order to react and avoid stereotypes.* [ID11, M]

Teaching the Psychological Effects of Racism and Its Overall Impact on Health

The effects of racism on the psyche were described as multiple, ranging from impacts on subjectivation to persistent doubt regarding the legitimacy of experiences perceived as discrimination, leading to ongoing cultural insecurity.

Understanding the Impact of Racial Discrimination on Health and Ensuring Patient Safety

The psychological impact of racial discrimination was described as well documented in American studies, particularly with regard to microaggressions and more acute traumatic experiences.

> *The way I teach this is that I talk about cumulative trauma. The example I give is the glass of water you hold. You hold it for 10 seconds without worrying. After a minute, you start to feel a bit unwell. After 5 - 10 minutes, it becomes very unpleasant. And then I mention stress, which has biological consequences.* [ID13, M]

Several participants referred to epidemiological studies, both North American and French, to describe the effects of racism on health, particularly mental health. These effects were sometimes described as being insufficiently recognized, including by those who experience them. The notion of cultural safety in care was emphasized, with participants warning of serious consequences if it is not ensured.

> *It means providing healthcare that is culturally safe and respectful of communities.* [ID13, M]

One participant addressed racism mediated through the gaze, referring to the concept of the racial epidermal schema developed by Caribbean psychiatrist Frantz Fanon, and emphasized the need to educate the clinical gaze in order to improve therapeutic relationships and counter racial prejudice.

> *He talks about the racial epidermal pattern, which means that in the lived experience of black people, they are victims of racism through the gaze. And that's something we don't work on enough in medicine, the gaze is never neutral, it's part of a personal, medical and colonial history. And Fanon says, we have to educate our gaze, listen to classical music, operas, we have to read...* [ID03, M]

Capturing changes in the gaze before and after training was described as a potential tool for evaluating shifts in students' subjectivity during teaching. Several participants also highlighted racial bias within the therapeutic relationship, including difficulties being taken seriously, the perceived obligation to dress or behave in certain ways to be heard, and the tendency for clinicians to rely on stereotypical interpretive frameworks in contexts of hospital crisis.

> *There's such a crisis that there's not much guarantee that we'll be well looked after. But there's still a bias at play, a kind of tacit reconduction, an acceptance of this kind of imposition of a line of reading that continues to work.* [ID20, P]

Understanding the Consequences of Assimilation to Support Patients' Processes of Subjectivation

Regardless of their profession, some professionals seemed very aware of the psychological risks posed by assimilation and of the need to teach it to medical students through the humanities and psychology, as well as to raise awareness of it in practice.

One participant, who was both a health care professional and a researcher in the humanities, recounted an experience in which a North African health care worker laughed at racist jokes as a form of internalized racism and conformity to the norms of certain team environments.

> *I remember a surgeon who liked to say that patients from the Maghreb inevitably had a lot more pain than others because they were wimps and had Mediterranean syndrome. They used to say it to the staff, and even the North Africans laughed. When you're a North African carer and you're laughing at that, I understand that it doesn't make sense to you in the end. And who were convinced that it was something that existed. That means we're almost assimilating colonial thinking.* [ID02, nurse and HSS]

Frantz Fanon was again referenced, particularly his distinction between assimilation and integration, and his description of

XSL•FO
RenderX

assimilation as a deprivation of self in order to resemble the other.

> *Fanon said that to be assimilated is the prefix A which is privative. He said that to be assimilated is to deprive oneself of a part of oneself or one's entirety in order to become similar to the other. And he contrasted assimilation with integration. Integration means coming with what you are, taking the best of the other and incorporating it into yourself. That's interculturality.* [ID02, nurse and HSS]

Assimilation was also discussed as a way of being in the world. One participant referred to the concept of double consciousness theorized by WEB Du Bois in 1903 in *The Souls of Black Folk*, describing the tension between imposed stereotypes and the constant effort to counter them.

> *This question of adaptability, psychologically, there's a charge of always adapting to the other. You're constantly thinking, consciously or unconsciously, about the stereotypes held about the group to which you're perceived to belong. And there's the action of constantly trying to thwart them in everyday life or, on the contrary, to overplay them. And once again, it's a mask to try and hide behind, because in the end, showing yourself is perceived by these people as a source of danger. So it's always black skin, white mask.* [ID19, P]

The contradictory injunctions related to belonging to the French collective were described as contributing to identification disorders that may become extreme and lead to forms of loss of identity anchors.

> *The effects are permanent illegitimisation of various kinds. Racialised people in French society are constantly reassigned to a specific position. This produces a form of identity crisis, let's say, a disturbance in identification that can be extremely intense, an intimate experience of an omnipresent contradictory double injunction.* [ID20, P]

The effects of racism on the psyche were further described as phenomenological, particularly with regard to whether individuals' identities are recognized within the therapeutic relationship. Nonrecognition was described as a form of violence and disqualification that is difficult to overcome.

> *Not taking the other person's words into account in their own language, and therefore saying to them, your words are symmetrical, I'm going to treat you, you're my fellow human being and your language is as good as mine. Really, when everything is done to put you in a position where the other person is less legitimate, it has a major impact on people's self-esteem.* [ID12, M]

This participant further discussed, particularly in relation to children, the dissociation between cognitive and affective dimensions that such experiences may require, leading to increased vulnerability. Drawing on Georges Devereux's concept from *La renonciation à l'identité. Défense contre l'anéantissement* (*The Renunciation of Identity: Defense Against Annihilation*), she emphasized the importance of recognition in enabling individuals to mobilize their full resources.

> *When all your affiliations are disqualified, it's only the cognitive that can be put forward and that's very vulnerable, not just for children. These are things we might call intersubjective. The more symmetrical the relationship, the more we can talk to each other, the more we respect the principle of psychological universality, the more likely it is that the other person will be able to use all his or her resources, and not put him or herself in a position of effacement or powerlessness, all things that do a lot of harm to the person, of renouncing his or her identity.* [ID12, M]

Teaching the Risks Associated With Racialized Care and So-Called Colorblind Approaches

The issue of racialized care was described along 2 main lines. First, participants highlighted the persistent misuse of racial criteria based on stereotypes inherited from colonial medicine and slavery. Second, they emphasized the underrepresentation, in medical textbooks, research, and training curricula, of what they referred to as geographical minorities, as well as of specific clinical features that require consideration in order to provide effective care.

The Pitfall of Culturalism

Several participants, mainly HSS professionals, questioned the use of the term culture, noting that its meaning varied across contexts and countries. On the one hand, culture could be understood as a useful concept for engaging with patients' representations of illness and ways of thinking about disease. On the other hand, participants warned that it could also be mobilized in a fixist and categorical manner that limits clinical reasoning and reinforces stereotypes. This latter use, referred to as culturalism, was described as an obstacle to care and a source of racial discrimination.

> *How do we sometimes refer to certain individuals, reduce their identity to so-called cultural stereotypes, on their way of thinking, their way of behaving. There are a certain number of representations that are attached to these individuals and which then condition the organization of a whole care system around these characteristics that do not correspond to the personality of the individual.* [ID06, HSS]

Referring to fieldwork conducted in the United States, the same participant described differences in care structures, including community-based services, in which health care providers were not always aware of their own stereotypes associated with particular communities. These representations were described as having direct consequences for the organization of care.

> *All this cultural organization also sometimes referred to very fixed categories, which we tried to develop in a different way and to differentiate from biologism, and which also referred to a whole strategy of representation which, itself, was very fixist and which again reduced the individual to boxes. We need to understand the social condition, the trajectories and*

XSL•FO
RenderX

> *the representations that these patients have of themselves and their families in order to reach out to them and improve care.* [ID06, HSS]

One participant described warning students against the overuse of the notion of culture, encouraging them to reflect on their own cultural frameworks and to mobilize cultural considerations only when sufficient time is available to explore patients' representations. Otherwise, there is a risk of reinforcing culturalism.

> *I'd say be careful with the notion of culture, because when you say African culture, North African culture, it doesn't really mean anything. Of course, you have to be taught to talk about yourself, talk about your body, express your feelings or not express them, hide them, and so on. If you don't have the time to touch it, don't touch it at all, forbid yourself to consider it in cultural terms. If you have the time, go and ask each of them, in fact, how do you want to give birth? Where do you get the urge to give birth like that, etc.? So go ahead, but really go ahead and take your time. And sometimes expect to be very disconcerted by the answers, because you've stereotyped something for yourself.* [ID17, midwife and HSS]

The Question of Differentiated Formulas or Protocols

The issue of differentiated clinical formulas or protocols was described as particularly salient in the United States. Several participants, combining transdisciplinary perspectives with physicians and HSS teachers, denounced the racist nature of the use of racial criteria without scientific basis, emphasizing the serious consequences for patient care.

> *There's the story of glomerular filtration rate in nephrology. They noticed that from time to time African-American dialysis patients had high creatinine, and rather than blaming it on renal failure they said it was because they had a large muscle mass. They stacked up a series of clichés like that and ended up with this. We realized that taking racial criteria into account when calculating the formula made a difference to the management of early renal failure in African-Americans. I tried to dig deeper, I tried to go back to, shall we say, original sin. It's impossible to find, really. There aren't really any studies, any sort of concrete data, just a posteriori studies.* [ID04, M]

The participant further described how post hoc analyses based on racial categories were used to generate generalized conclusions and algorithms, despite their methodological limitations. In addition to nephrology, similar practices were reported in other specialties, including algorithms related to cesarean sections, fever, and other routine clinical decisions.

Another participant with research experience in the United States highlighted the routine use of ethnicity categories in patient forms through what was described as heterodefinition (ID06, HSS), with a limited number of categories primarily designed for administrative purposes. These categories were then applied mechanically in clinical contexts, often under the assumption of improving efficiency.

> *In the United States, we are so close to these ethno-racial categories that they are no longer questioned at all. And so we slide from a category that sometimes means something cultural, that sometimes means something social, as I was saying about the proxy, the social race, and that sometimes also means something biological.* [ID06, HSS]

The same participant noted that drug adaptations and dosage adjustments were often based on hearsay inherited from broader systems of representation related to stereotypes of hypermasculinity and physical strength, rather than on formal medical training.

> *I would refer you to a whole body of literature on the sociology of the body, the history of the body, to understand how, in medicine and psychiatry, we have been able to define care protocols that are more or less informal and different.* [ID06, HSS]

A critical approach to complementary examinations, particularly equations incorporating ethnic correction factors, was recommended to students by another participant, who emphasized the importance of documenting patient histories, noting that subsequent clinical decisions depend heavily on how cases are described.

> *The next colleague's attitude will depend on the terms used.* [ID14, M]

In parallel, one participant working in Canada described efforts to take skin color into account in order to document discrimination in health care and identify the increased prevalence of certain diseases. The objective was to improve public health by drawing on community experiences and improving outreach and care.

Explaining the concept of sociological race and accounting for health-relevant ethnic differences

Training was described as addressing, first, the different meanings of the term race in medicine: historically as a biologized category, and sociologically as an assigned race whose effects must be recognized in order to be dismantled. Second, participants emphasized the need to address the specific medical needs of certain communities within the same geographical area.

One participant highlighted the mismatch between fixed phenotypical categories and the diversity of lived situations, which can lead to erroneous causal reasoning in diagnosis and treatment.

> *It's true that it's complicated for students. That's why, when we talk about essentialism, we tend to associate a racial category, which is a reduction of what we observe, with a phenotype without understanding the complexity of the phenomenon. And it's when we naturalize the category with reality, when we imbricate the two, that we create slippages and that these slippages can potentially have the effect of*

XSL•FO
RenderX

> *over-investing and over-producing causalities that in fact don't exist.* [ID06,HSS]

Dermatology was cited as an example of the type of knowledge required to provide care that accounts for individual characteristics. One dermatologist participant challenged the idea that dermatology on Black skin is inherently difficult, noting that while specific features exist, they are easy to learn when training includes representations of diverse skin types.

The example of ethnic pseudoneutropenia among individuals from Africa or the Middle East was also discussed, particularly regarding its implications for the prescription of medications such as clozapine in psychiatry or for chemotherapy dosing.

> *To be on the safe side, we're going to lower doses. In the United States, for example, it's well known that breast cancer has a poor prognosis in African-American women, mainly because of lower chemotherapy doses.* [ID04, M]

Several participants pointed to insufficient training on health issues specific to certain communities, including in dermatology and psychiatry. This lack of training was associated with the overrepresentation of racialized communities in psychiatric services, particularly for psychotic disorders. Participants emphasized that Western medical training was historically designed to treat White populations, a limitation sometimes summarized by the expression "racism, not race," highlighting the health consequences of racism itself.

> *We are ill-equipped to understand how certain psychiatric pathologies manifest themselves in certain communities. Because medical training has been based on treating white people. In psychiatry, we apply the DSM as if it had been made for everyone, whereas in Montreal, as I say, people from racialized communities are over-represented in our institutions, particularly when it comes to psychotic disorders.* [ID13, M]

Discussion

Principal Findings

The stated objective of this study was to identify the core content areas to be defined and the corresponding learning objectives to be established for the development of a curriculum addressing racial discrimination in medicine for undergraduate medical education.

A recurring challenge in medical education on these issues is avoiding culturalism, and our findings offer several insights in this regard. Specifically, the results underscore the importance of helping students recognize racial stereotypes and learn how to challenge them, understand the psychological burden of racial discrimination, and remain attentive to relevant geographical and sociological factors without essentializing patients or increasing the risk of misdiagnosis.

Several findings are consistent with existing literature, including the emphasis on hierarchies of pain perception [23,24], the risk of diagnostic error or misdiagnosis [25], and the dangers associated with clinical formulas or protocols differentiated on the basis of presumed ethnicity [26,27]. However, our study also highlights specific features that merit further discussion. Notably, likely due to the high proportion of psychologists and psychiatrists among participants, the psychological impact of racism on care and on identity emerged as a central theme. This psychological dimension is closely linked to the specificity of the "Mediterranean syndrome" in France, which is rooted in colonial history and requires critical examination to be effectively dismantled in both clinical practice and professional representations. Another key finding concerns the pedagogical challenge of enabling students to distinguish between medically relevant contextual knowledge and the continuation of culturalist reasoning that leads to unsafe, racialized care.

The aim is to provide future physicians with a form of symbolic authorization to introduce nuance into their clinical reasoning. In line with our previous work [17], which highlighted the importance of teaching students to reflect on countertransference in caregiving relationships, these results suggest that once such reflexive work is initiated, information about racism as a risk factor for health is more readily integrated into everyday clinical practice.

Pain, Mediterranean Syndrome, and Their Historical and Psychological Dimensions

Racialization can be defined as the process by which race is socially constructed to enable the domination of one group over another, leading societies to view races as real, different, and unequal across social, economic, and political domains [1]. According to the definition of the Human Rights Commission of Ontario in 2005, "Racialized people" refers to people seen as belonging to racialized minorities and are people who could be perceived as being socially different from, for example, the racial or ethnic majority. The word "racialized" stresses the fact that race is neither biological nor objective but is a concept that is societal in origin [28].

The use of racialized categories in medicine reinforces the risk of culturalism in care—that is, attributing fixed and naturalized characteristics to presumed patient cultures, homogenizing groups while obscuring social, historical, political, and structural determinants of health. One consequence of racialization, and a paradigmatic example of how stereotypes can bias the therapeutic relationship, concerns pain management, a symptom that is both physical and psychological. Racism, sexism, and stigma toward mental health conditions intersect in the so-called "Mediterranean syndrome," a term still used or implicitly invoked in French hospitals [29]. A growing body of French academic literature and media reports [29-31] has documented the seriousness of this prejudice, which often operates implicitly and is inherited from colonial medical traditions [32]. This issue has also been formally highlighted by the French Defender of Rights [10]. Such institutional recognition supports framing this phenomenon as a public health concern.

Our findings reveal the pervasive anxiety among patients about whether they will be taken seriously in emergency settings and suggest that this stereotype intersects with broader prejudices toward patients presenting physical symptoms without identifiable lesions or clear diagnoses, commonly categorized as functional or somatic symptom disorders [33]. These findings

XSL•FO
RenderX

resonate with the work of Frantz Fanon in 1952 in Le syndrome nord-africain (The North African syndrome) [34,35], who emphasized the role of psychological suffering in physical symptoms among North African workers and highlighted how racialization of the body produces specific forms of alienation and distress. Participants expressed strong concern about clinical situations in which the psychological context was ignored and symptoms were immediately attributed to "Mediterranean syndrome," reflecting the activation of racist assumptions in the absence of organic findings.

Drawing on the work of the phenomenological philosopher Merleau-Ponty [36], who emphasized that individuals engage with the world through their bodies, Frantz Fanon described the impact of the "White gaze," which objectifies Black individuals by reducing them to their bodies—and ultimately to their skin—what he termed the "racial epidermal schema" [37]. According to Fanon, this imposed separation between body and subjectivity, developed as a response to racism, can generate both physical and psychological distress, including feelings of alienation and disruption of one's body schema. These concepts resonate with our findings on the dissociation between affective and cognitive dimensions of care that can arise when patients' identities and lived experiences are denied. Such dissociation may negatively affect psychological development and health by prioritizing cognitive explanations while fostering mistrust toward the expression of emotions. This highlights an often-overlooked link between 2 forms of bias in health care: the belief that suffering is only legitimate when it is visible, measurable, or objectifiable. When pain does not fit these criteria, it may be dismissed as exaggerated, feigned, or "merely psychological." Framing symptoms as "cultural" or "all in the patient's head" effectively invalidates suffering, including suffering related to experiences of racism. Even when overt accusations of simulation are absent, the systematic relativization of physical symptoms—particularly among racialized patients—risks not only serious diagnostic errors but also the erasure of the psychological meaning of these symptoms. Building on Fanon work and that of Gayatri Spivak, Maude Ludot, a child psychiatrist and researcher on somatoform disorders in adolescents, argues that somatic symptoms may function as expressions of experiences that cannot otherwise be articulated, particularly experiences of discrimination [38]. When such experiences are not heard or acknowledged, they may instead be expressed through physical symptoms later described as "medically unexplained."

Research increasingly confirms the existence of a close link between experiences of racial discrimination and psychosomatic disorders, with some data also suggesting that these effects may be amplified in women [39]. From an educational and clinical perspective, it would be appropriate to develop a module on somatic and somatoform disorders linked to racial discrimination, in line with the concept of "embodiment of discrimination" [40], which aims to highlight how social processes (colonial heritage, microaggressions, denial of competence, etc) translate into concrete alterations of the body and the relationship to the body, involving psychological effects as well.

"Mediterranean syndrome" also stems from a misunderstanding between the legitimate need to consider so-called cultural factors when appropriate—particularly in the context of mental health care—and a form of culturalism that turns these factors into a generalized, essentialized, and ready-made explanatory framework, applied indiscriminately to patients without any knowledge of their individual affiliations.

Between Culturalism and Colorblindness: Identifying Medically Relevant Factors

One key concept emerging from both Fanon work and our findings is that of situational diagnosis, which involves systematically considering environmental factors in clinical assessment. While contemporary medicine increasingly acknowledges social determinants of health, racism itself is rarely treated as a risk factor in medical histories, despite substantial evidence suggesting that racism—not race—is the relevant determinant of adverse health outcomes [41,42]. This distinction is essential for medical education.

Students often struggle to differentiate between the legitimate use of certain geographical, sociological, or phenotypic information and the pitfalls of culturalism and racialized care. Our findings suggest addressing this challenge through 3 complementary approaches.

First, students should be taught a limited number of clinically relevant contextual features. Dermatology education and the importance of representing all skin tones [43-45]; so-called ethnic pseudoneutropenia whose benign nature, if not recognized, may lead to inappropriate dose reductions, particularly for cancer chemotherapy or clozapine in the treatment of schizophrenia [46]. Invisibilization also occurs at the research level, where studies frequently exclude minority populations that do not fit the profile of the "ideal" patient, despite the fact that clinical reality more closely reflects these populations—for example, through the systematic exclusion of allophone patients in studies of migrant groups. From a transcultural medicine perspective, often mobilized as a second-line approach, attention to patients' specific lived experiences may be clinically relevant, but only when it meaningfully contributes to the therapeutic relationship or the delivery of care, and when clinicians have the time and skills to do so [47,48]. Exploring patients' representations of illness, treatment, and the health care system requires decentering and a deliberate avoidance of stereotyping, essentialization, or fascination, all of which constitute racist biases that are both harmful and detrimental to quality of care.

Second, medical education must explicitly deconstruct the use of racialized formulas or protocols based on weak or biased evidence [49], many of which originate in colonial or slave-based medical histories [50]. Across medical specialties, so-called race-based clinical adjustments highlight the need for stronger critical training on the historical and educational roots of medical knowledge [51-57] to prevent misdiagnosis, improve clinical reasoning, and emphasize that observed disparities are more plausibly related to chronic exposure to stress and trauma than to biological racial differences [58,59].

XSL•FO
RenderX

Finally, reframing racism as a public health risk factor—distinct from any notion of biological race [60,61]—"racism not race," allows medical education to move beyond both culturalism and colorblindness [6]. This approach emphasizes structural, environmental, and psychosocial determinants of health [9], such as perceived discrimination, which has been shown to affect sleep [62], BMI [63], and engagement in high-risk behaviors [39], for example.

Patient safety is a term used repeatedly by participants and is an integral part of the explicit goal of combating health inequalities and discrimination in health care. The study includes some findings on how to achieve this, such as the need to include specific knowledge in teaching, while avoiding any cultural bias, and to teach about the psychological consequences of discrimination, which impact overall health. The more precise term "cultural safety," raised by several participants, evokes the need to address the patient's affiliations, namely their language, history, and perceptions. This term, prevalent in the literature [64,65] and tending to replace "cultural competence," [66] is presented as a way to legitimize and value cultural differences in order to prevent harm in interactions between health care professionals and patients.

Clarifying the concept of race is therefore essential. As emphasized by scholars in political philosophy and psychiatry [37], race is not a matter of biological essence but of social assignment, stigmatization, and lived experience. Cultural affiliations are dynamic and shaped by shared histories, social positioning, and, in many cases, exposure to discrimination or trauma. According to researcher and psychiatrist Blanc [67], strengthening resilience and advancing health equity thus require holistic, multisectoral approaches that integrate historical, social, political, and educational dimensions alongside clinical care.

Strengths and Limitations

A major strength of this study—uncommon in the French context—is the inclusion of diverse health professions educators' perspectives on racism in health care within French-speaking settings. These findings identify concrete opportunities to develop structured educational interventions for medical students and to extend them to other health care professionals through a national network of educators sharing resources and pedagogical approaches.

This study focuses exclusively on educators' perspectives and does not include those of students or patients. Although participants represented a range of professional backgrounds, regions, and countries of practice, future research should incorporate students' and patients' viewpoints to better assess the relevance, acceptability, and impact of these educational initiatives.

Another limitation is that because our inclusion criteria required prior publications and documented involvement in teaching initiatives, the sample overrepresents already-engaged educators and underrepresents those who are reluctant, resistant, or institutionally constrained, which may limit the diversity of perspectives captured. Also, nonspecialist teachers represent the majority of health professions educators. Future research should examine their roles and training needs, as their involvement will be essential to generalize this educational content.

Finally, this study may be influenced by social desirability and topic-engagement biases, as participants might have presented their views in ways they perceived as socially acceptable and were already invested in education on racial discrimination, potentially overrepresenting more positive or elaborated perspectives. In addition, given the small expert community and the detailed participant descriptors (profession and country), a risk of deductive disclosure cannot be fully excluded.

Conclusion

Teaching about the effects of racial discrimination and racism on health aims to establish these issues as core public health concerns. Beyond the sense of medical powerlessness that may lead to the reliance on stereotypes within health care systems characterized by time constraints and pressure for rapid diagnosis and treatment, students must be supported in understanding how clinical practices are shaped by their historical and social embeddedness. In such contexts, patients from racialized minorities are at greater risk of reduced access to safe and equitable care.

The impact of racial discrimination on health must be explicitly addressed in medical education, as it is neither inevitable nor immutable, even though it is deeply intertwined with economic, political, and sociological processes that extend beyond the medical sphere. From a future perspective, the ongoing French debate regarding the feasibility of conducting robust epidemiological studies on racial health disparities remains unresolved. Rigorous epidemiological research on differential care represents a critical tool for making health disparities visible, identifying causal pathways, informing prevention strategies, and mobilizing policymakers. Such research also plays a key educational role by enhancing understanding of how discrimination can operate at every stage of care, from initial patient access to the full clinical encounter.

Raised by some participants, the issue of the place and experiences of health care workers who themselves come from ethnic minorities should be considered in future work.

Acknowledgments

The authors thank all those who contributed to this study. Artificial intelligence (AI) tools (ChatGPT-5.3) were used to assist in the translation of an author's original work into a secondary language for inclusion in the manuscript.

XSL•FO
RenderX

Funding

The authors declared no financial support was received for this work.

Data Availability

All data generated or analyzed during this study are included in this published article.

Authors' Contributions

Conceptualization: ML, MRM (equal), MM, RR, AM (supporting)
Data curation: ML
Formal analysis: ML (lead), MM, RR (supporting)
Methodology: ML, AM
Resources: MRM
Supervision: MM, RR, MRM
Validation: MRM, RR
Writing – original draft: ML
Writing – review & editing: ML, MRM, MM, RR, AM

Conflicts of Interest

None declared.

Multimedia Appendix 1
Interview guides.
[DOCX File, 15 KB - mededu_v12i1e90084_app1.docx]

Checklist 1
COREQ reporting guidance.
[PDF File, 363 KB - mededu_v12i1e90084_app2.pdf]

References

1. Jones CP. Levels of racism: a theoretic framework and a gardener's tale. Am J Public Health 2000 Aug;90(8):1212-1215. [doi: 10.2105/ajph.90.8.1212] [Medline: 10936998]
2. Miles R, Brown M. Racism: Psychology Press; 2003.
3. Williams DR, Lawrence JA, Davis BA. Racism and health: evidence and needed research. Annu Rev Public Health 2019 Apr 1;40(105-25):105-125. [doi: 10.1146/annurev-publhealth-040218-043750] [Medline: 30601726]
4. Williams DR. How racism makes us sick. TED Talk. 2017. URL: https://www.ted.com/talks/david_r_williams_how_racism_makes_us_sick [accessed 2026-04-01]
5. Paradies Y, Ben J, Denson N, et al. Racism as a determinant of health: a systematic review and meta-analysis. PLoS One 2015;10(9):e0138511. [doi: 10.1371/journal.pone.0138511] [Medline: 26398658]
6. Medlock M, Shtasel D, Trinh N, Williams DR. Racism and Psychiatry: Contemporary Issues and Interventions: Humana Press; 2019.
7. Emmer C, Dorn J, Mata J. The immediate effect of discrimination on mental health: a meta-analytic review of the causal evidence. Psychol Bull 2024 Mar;150(3):215-252. [doi: 10.1037/bul0000419] [Medline: 38330346]
8. D'Almeida Y. Impact Des Microagressions et de La Discrimination Raciale Sur La Santé Mentale Des Personnes Racisées: Editions L'Harmattan; 2022.
9. Selvarajah S, Corona Maioli S, Deivanayagam TA, et al. Racism, xenophobia, and discrimination: mapping pathways to health outcomes. Lancet 2022 Dec 10;400(10368):2109-2124. [doi: 10.1016/S0140-6736(22)02484-9] [Medline: 36502849]
10. Report by the Human Rights Defender: Preventing Discrimination in Healthcare Pathways: A Challenge for Equality 2025. URL: https://www.defenseurdesdroits.fr/sites/default/files/2025-05/ddd_rapport_discriminations-parcours-de-soins_20250430.pdf [accessed 2026-04-13]
11. Foucault M. Society Must Be Defended: Lectures at the Collège de France, 1975–1976: Picador; 2003. URL: https://us.macmillan.com/books/9780312422660/societymustbedefended/ [accessed 2026-04-08]
12. Fassin D. Public Health in Action: Presses de l'EHESP; 2008. [doi: 10.3917/ehesp.fass.2008.01]
13. Carde E. Discrimination based on origin in access to healthcare. Santé Publique 2007 Jun 1;Vol. 19(2):99-109. [doi: 10.3917/spub.072.0099]
14. Roberts DE. The problem with race-based medicine. Am J Phys Med Rehabil 2023 Jul 1;102(7):567-570. [doi: 10.1097/PHM.0000000000002251] [Medline: 37317023]

XSL•FO
RenderX

15. Gama RM, Kalyesubula R, Fabian J, Mahalingasivam V. NICE takes ethnicity out of estimating kidney function. BMJ 2021 Sep 10;374:n2159. [doi: 10.1136/bmj.n2159] [Medline: 34507980]
16. Shema M. Health Is a Political Issue: Belfond; 2025.
17. Lambert M, Mansouri M, Mouhab A, Radjack R, Moro MR. Perspectives of health, and human and social sciences professionals on student transformation regarding racial discrimination in healthcare. BMC Med Educ 2025 Jul 1;25(1):905. [doi: 10.1186/s12909-025-07513-z] [Medline: 40598367]
18. Braun V, Clarke V. Using thematic analysis in psychology. Qual Res Psychol 2006 Jan;3(2):77-101. [doi: 10.1191/1478088706qp063oa]
19. Braun V, Clarke V. Successful Qualitative Research: A Practical Guide for Beginners: Sage; 2013.
20. Braun V, Clarke V. Reflecting on reflexive thematic analysis. Qual Res Sport Exerc Health 2019 Aug 8;11(4):589-597. [doi: 10.1080/2159676X.2019.1628806]
21. Patton MQ. Qualitative Evaluation and Research Methods, 2nd edition: Sage; 1990.
22. Devereux G. De l'angoisse à La Méthode Dans Les Sciences Du Comportement: Editions Flammarion; 1980.
23. Hoffman KM, Trawalter S, Axt JR, Oliver MN. Racial bias in pain assessment and treatment recommendations, and false beliefs about biological differences between blacks and whites. Proc Natl Acad Sci USA 2016 Apr 19;113(16):4296-4301. [doi: 10.1073/pnas.1516047113] [Medline: 27044069]
24. Booker SQ, Merriwether EN, Powell-Roach K, Jackson S. From stepping stones to scaling mountains: overcoming racialized disparities in pain management. Pain Manag 2024 Jan;14(1):5-12. [doi: 10.2217/pmt-2023-0098] [Medline: 38193345]
25. Jegarl AM, Jegede O, Isom J, Ciarleglio N, Black C. Psychotic misdiagnosis of racially minoritized patients: a case-based ethics, equity, and educational exploration. Harv Rev Psychiatry 2023;31(1):28-36. [doi: 10.1097/HRP.0000000000000353] [Medline: 36608081]
26. Cerdeña JP, Plaisime MV, Tsai J. From race-based to race-conscious medicine: how anti-racist uprisings call us to act. Lancet 2020 Oct 10;396(10257):1125-1128. [doi: 10.1016/S0140-6736(20)32076-6] [Medline: 33038972]
27. Roberts DE. Abolish race correction. Lancet 2021 Jan 2;397(10268):17-18. [doi: 10.1016/S0140-6736(20)32716-1] [Medline: 33388099]
28. Souissi T. Racialized minority. In: The Canadian Encyclopedia 2024. URL: https://www.thecanadianencyclopedia.ca/en/article/racialized-minorities [accessed 2026-04-01]
29. Lambert M, Lachal J, Mansouri M, Radjack R, Moro MR. Mediterranean syndrome and the French medical world, a racist prejudice still active. A parallel with Frantz Fanon's article about the "North African syndrome". Rev Med Interne 2022 Jul;43(7):399-401. [doi: 10.1016/j.revmed.2022.04.020] [Medline: 35623924]
30. Labainville I. The 'Mediterranean syndrome': the origins of a racialized category in the hospital. Genèses 2025 Apr 24;n° 138(1):14-36. [doi: 10.3917/gen.138.0014]
31. Dergham M, Charles R. The "Mediterranean syndrome": stigmatization through the categorization of disease patterns. Médecine 2020;16(10):460-464 [FREE Full text] [doi: 10.1684/med.2020.606]
32. Lambert M, Lachal J, Mansouri M, Moro MR. Expérience subjective du rapport aux soins pédopsychiatriques chez des adolescents franco-algériens: une perspective individuelle, et historique. Neuropsychiatr Enfance Adolesc 2022 Dec;70(8):441-447. [doi: 10.1016/j.neurenf.2022.09.001]
33. American Psychiatric Association. DSM-5: Manuel Diagnostique et Statistique Des Troubles Mentaux, 5th edition: Elsevier Masson; 2015.
34. Fanon F. The "North African syndrome". Esprit 1952;187(2):237-248 [FREE Full text]
35. Dhanvantari S. The concept of unlivability: a reading of Frantz Fanon's "The North African Syndrome" (1952). South J Philos 2024 Mar;62(1):45-64. [doi: 10.1111/sjp.12554]
36. Merleau-Ponty M. Phenomenology of Perception: Routledge; 2012. [doi: 10.4324/9780203720714]
37. Bessone M. Racism has numerous physical effects. Le Monde. 2021. URL: https://www.lemonde.fr/series-d-ete/article/2021/07/13/le-racisme-a-des-effets-corporels-multiples_6088076_3451060.html [accessed 2026-04-01]
38. Salmon M, Sibeoni J, Harf A, Moro MR, Ludot-Grégoire M. Systematic review on somatization in a transcultural context among teenagers and young adults: Focus on the nosography blur. Front Psychiatry 2022;13:897002. [doi: 10.3389/fpsyt.2022.897002] [Medline: 35958663]
39. Cénat JM, Kouamou LN, Moshirian Farahi SMM, et al. Perceived racial discrimination, psychosomatic symptoms, and resilience among Black individuals in Canada: a moderated mediation model. J Psychosom Res 2022 Dec;163(111053):111053. [doi: 10.1016/j.jpsychores.2022.111053] [Medline: 36244137]
40. Meziani Y. Discrimination: A Public Health Issue in France: The Hidden Cost: Le Bord de l'eau; 2025.
41. Hauspurg A, Jeyabalan A. The risk factor is racism, not race. Am J Obstet Gynecol 2022 Aug;227(2):358-359. [doi: 10.1016/j.ajog.2022.03.011] [Medline: 35278377]
42. Garvey A, Lynch G, Mansour M, Coyle A, Gard S, Truglio J. From race to racism: teaching a tool to critically appraise the use of race in medical research. MedEdPORTAL 2022;18:11210. [doi: 10.15766/mep_2374-8265.11210] [Medline: 35128047]
43. Mahé A, Faye O. Dermatologie de La Diversité: Elsevier Masson; 2022.

XSL•FO
RenderX

44. Brady J, Kashlan R, Ruterbusch J, Farshchian M, Moossavi M. Racial disparities in patients with melanoma: a multivariate survival analysis. Clin Cosmet Investig Dermatol 2021;14:547-550. [doi: 10.2147/CCID.S311694] [Medline: 34079319]
45. Ibe C. How diversity in medical illustrations can improve healthcare outcomes. TED Talk. 2022. URL: https://www.ted.com/talks/chidiebere_ibe_how_diversity_in_medical_illustrations_can_improve_healthcare_outcomes [accessed 2026-04-01]
46. Andreou A, Jayaram J, Walker A, Tek C, Williams JC. Re-examining the utility and validity of benign ethnic neutropenia: a narrative literature review. Schizophr Res 2023 Mar;253(48-53):48-53. [doi: 10.1016/j.schres.2022.02.009] [Medline: 35241313]
47. Radjack R, Moro MR. More cultural humility for the children of immigrants and their families. Vie sociale 2025 Apr 14;48(4):27-42. [doi: 10.3917/vsoc.048.0027]
48. Kleinman A, Benson P. Anthropology in the clinic: the problem of cultural competency and how to fix it. PLoS Med 2006 Oct;3(10):e294. [doi: 10.1371/journal.pmed.0030294] [Medline: 17076546]
49. Vyas DA, Eisenstein LG, Jones DS. Hidden in plain sight - reconsidering the use of race correction in clinical algorithms. N Engl J Med 2020 Aug 27;383(9):874-882. [doi: 10.1056/NEJMms2004740] [Medline: 32853499]
50. Peiretti-Courtis D. Black Bodies and White Doctors: La Découverte; 2021.
51. Inker LA, Eneanya ND, Coresh J, et al. New creatinine- and cystatin C-based equations to estimate GFR without race. N Engl J Med 2021 Nov 4;385(19):1737-1749. [doi: 10.1056/NEJMoa2102953] [Medline: 34554658]
52. Braun L. Breathing Race into the Machine: The Surprising Career of the Spirometer from Plantation to Genetics: University of Minnesota Press; 2014.
53. Cohen W. Medical apartheid: the dark history of medical experimentation on Black Americans from colonial times to the present. Appl Psychoanal Studies 2009 Dec;6(4):356-360 [FREE Full text] [doi: 10.1002/aps.223]
54. Tobin MJ. Fiftieth anniversary of uncovering the Tuskegee syphilis study: the story and timeless lessons. Am J Respir Crit Care Med 2022 May 15;205(10):1145-1158. [doi: 10.1164/rccm.202201-0136SO] [Medline: 35500908]
55. Skloot R. The Immortal Life of Henrietta Lacks: Crown; 2010. [doi: 10.5915/43-2-8609]
56. Sheffer E. Asperger's Children: The Origins of Autism in Nazi Vienna: WW Norton & Co; 2018. [doi: 10.3917/ess.046.0191]
57. Metzl J. How schizophrenia became a black disease. J Afr Am Hist 2010;97(4). [doi: 10.5323/jafriamerhist.97.4.0499]
58. Gasnier M, Aouizerat A, Chappell K, Baubet T, Corruble E. Psychotic and somatic symptoms are frequent in refugees with posttraumatic stress disorder: a narrative review. J Psychiatr Pract 2024 Mar 1;30(2):104-118. [doi: 10.1097/PRA.0000000000000772] [Medline: 38526398]
59. DeLuca JS, Novacek DM, Adery LH, et al. Equity in mental health services for youth at clinical high risk for psychosis: considering marginalized identities and stressors. Evid Based Pract Child Adolesc Ment Health 2022;7(2):176-197. [doi: 10.1080/23794925.2022.2042874] [Medline: 35815004]
60. Chadha N, Lim B, Kane M, Rowland B. Toward the abolition of biological race in medicine. Othering & Belonging Institute UC Berkeley. 2020. URL: https://escholarship.org/uc/item/4gt3n0dd [accessed 2026-04-01]
61. Yudell M, Roberts D, DeSalle R, Tishkoff S. Taking race out of human genetics. Science 2016 Feb 5;351(6273):564-565. [doi: 10.1126/science.aac4951]
62. Slopen N, Lewis TT, Williams DR. Discrimination and sleep: a systematic review. Sleep Med 2016 Feb;18(88-95):88-95. [doi: 10.1016/j.sleep.2015.01.012] [Medline: 25770043]
63. Bernardo CDO, Bastos JL, González-Chica DA, Peres MA, Paradies YC. Interpersonal discrimination and markers of adiposity in longitudinal studies: a systematic review. Obes Rev 2017 Sep;18(9):1040-1049. [doi: 10.1111/obr.12564] [Medline: 28569010]
64. Curtis E, Jones R, Tipene-Leach D, et al. Why cultural safety rather than cultural competency is required to achieve health equity: a literature review and recommended definition. Int J Equity Health 2019 Nov 14;18(1):174. [doi: 10.1186/s12939-019-1082-3] [Medline: 31727076]
65. Phiri J, Dietsch E, Bonner A. Cultural safety and its importance for Australian midwifery practice. Collegian 2010;17(3):105-111. [doi: 10.1016/j.colegn.2009.11.001] [Medline: 21046963]
66. Kirmayer LJ. Rethinking cultural competence. Transcult Psychiatry 2012 Apr;49(2):149-164. [doi: 10.1177/1363461512444673] [Medline: 22508634]
67. Blanc J. Racism and health disparities: the imperative of health equity. Revue L'Autre 2021;22(3):270-273. [doi: 10.3917/lautr.066.0270]

Abbreviations

COREQ: Consolidated Criteria for Reporting Qualitative Research
HSS: humanities and social sciences
INSERM : French National Institute for Health and Medical Research
TA: thematic analysis

XSL•FO
RenderX

Edited by XC Tong; submitted 02.Jan.2026; peer-reviewed by B Rezet, ED Bu; revised version received 08.Mar.2026; accepted 08.Mar.2026; published 17.Apr.2026.

Please cite as:
Lambert M, Radjack R, Mouhab A, Mansouri M, Moro MR
Health and Humanities and Social Sciences Professionals' Perceptions Regarding the Teaching of the Effects of Racism in Medicine: Semistructured Interview Study
JMIR Med Educ 2026;12:e90084
URL: https://mededu.jmir.org/2026/1/e90084
doi:10.2196/90084

© Mathilde Lambert, Rahmeth Radjack, Adil Mouhab, Malika Mansouri, Marie Rose Moro. Originally published in JMIR Medical Education (https://mededu.jmir.org), 17.Apr.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Education, is properly cited. The complete bibliographic information, a link to the original publication on https://mededu.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Publisher:
JMIR Publications
130 Queens Quay East.
Toronto, ON, M5A 3Y5
Phone: (+1) 416-583-2040
Email: support@jmir.org

XSL•FO
RenderX